Computational and Experimental Analysis of Genetic Variants

Jeremy W. Prokop,[*1,2] *Vladislav Jdanov,*[1] *Lane Savage,*[1] *Michele Morris,*[3] *Neil Lamb,*[3] *Elizabeth VanSickle,*[4] *Cynthia L. Stenger,*[5] *Surender Rajasekaran,*[1,6,7] *and Caleb P. Bupp*[1,4]

ABSTRACT

Genomics has grown exponentially over the last decade. Common variants are associated with physiological changes through statistical strategies such as Genome-Wide Association Studies (GWAS) and quantitative trail loci (QTL). Rare variants are associated with diseases through extensive filtering tools, including population genomics and trio-based sequencing (parents and probands). However, the genomic associations require follow-up analyses to narrow causal variants, identify genes that are influenced, and to determine the physiological changes. Large quantities of data exist that can be used to connect variants to gene changes, cell types, protein pathways, clinical phenotypes, and animal models that establish physiological genomics. This data combined with bioinformatics including evolutionary analysis, structural insights, and gene regulation can yield testable hypotheses for mechanisms of genomic variants. Molecular biology, biochemistry, cell culture, CRISPR editing, and animal models can test the hypotheses to give molecular variant mechanisms. Variant characterizations can be a significant component of educating future professionals at the undergraduate, graduate, or medical training programs through teaching the basic concepts and terminology of genetics while learning independent research hypothesis design. This article goes through the computational and experimental analysis strategies of variant characterization and provides examples of these tools applied in publications. © 2022 American Physiological Society. *Compr Physiol* 12:3303-3336, 2022.

Didactic Synopsis

Major teaching points

- Common variants greater than 1% allele frequency nominated by GWAS and fine-mapping are characterized by screening of variants for protein changes, gene regulation, and colocalization of signals for other phenotypes or traits.
- Rare variants below 0.01% allele frequency are connected to phenotypes through variant filtering and trio-based sequencing followed by assessing how variants impact protein function.
- Known data filtering provides ultra-rapid insights on variant connections to genes, cells, protein networks, epigenetics, and animal phenotypes.
- Data compiled with bioinformatics including gene evolution, protein structures, posttranscriptional modifications, splicing, nonsense-mediated decay (NMD), and transcription factor binding can generate hypotheses for how a variant influences physiology.
- Molecular biology, biochemistry, cell culture, and animal modeling can then be used to test variant hypotheses.
- There is growing support for environmental components including hypoxia and viruses to activate responses that are modulated by genetic variants.
- Variant analysis can be integrated into independent research or classroom activities to expose students to genetics and data analysis.

Introduction

While most students are exposed to genomics in their training, there is a rapid growth in the past few years of databases and tools that allow for the translation of genomic knowledge into physiological mechanisms. Large-scale sequencing of

*Correspondence to jprokop54@gmail.com

[1]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, Michigan, USA

[2]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, Michigan, USA

[3]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA

[4]Medical Genetics, Spectrum Health, Grand Rapids, Michigan, USA

[5]Department of Mathematics, University of North Alabama, Florence, Alabama, USA

[6]Pediatric Intensive Care Unit, Helen DeVos Children's Hospital, Grand Rapids, Michigan, USA

[7]Office of Research, Spectrum Health, Grand Rapids, Michigan, USA

Published online, April 2022 (*comprehensivephysiology.com*)

DOI:10.1002/cphy.c210012

DNA and RNA has elevated genetics (targeting a set of genes) into genome wide insights for biology. The terminology of classroom genetics has changed from Punnett squares and pedigrees into advanced concepts of genome-wide association studies (GWAS), quantitative trait loci (QTL), and complex consomic/congenic animal models for physiology. Many of the current genomic terminologies, strategies, and insights are not readily available to the classically trained physiologist. Many of the tools and the databases that are available are unknown to many physiologists. Within a few mouse clicks, physiologists have thousands of datasets publicly available to them. This article serves as a resource for a physiologist to gain a working knowledge of the terminology, tools, datasets, and strategies in genomics that can be applied to nearly any genetic variant that has connections to phenotypes and physiology. Therefore, with the insights gained from this article, we anticipate any physiologist to be able to analyze genetic variants for their projects.

Genetic variants have a significant impact on human medicine. Every individual has millions of variants relative to the reference genome. Variants include single nucleotide variants (SNVs)/single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and copy number variants (CNVs). Dissecting each variant's contribution to biology is currently impossible. Therefore, we rely on statistical strategies, data filtering, and high-throughput approaches to nominate variants to study within laboratories. More importantly, these tools to identify functional variants also elucidate many insights into the physiological genomics, such that a variant can be associated to gene and cell/tissue changes that impact physiological processes. Two main areas of genetic filtering have emerged for variants, common allele frequency that is less penetrant for phenotypes vs. rare allele frequency with highly penetrant phenotypes. These two strategies use inverse statistical approaches to associate variants to biological traits.

Traditionally, the most intuitive understanding of genetics on human biology has been seen in pediatric populations, as infants and children have had little environmental exposure that modulate phenotypes compared to adults. This allows for the determination of highly penetrant genotype-to-phenotype associations. Rare variants below 0.1% allele frequency and ultrarare variants unique in one or a few patients (proband) can be connected to disease by comparing inheritance from parents using trio-based sequencing, determining the variant absence in healthy individuals (Figure 1). Common variants (>1% allele frequency in the population) instead of require statistical analysis such as Genome-Wide Association Studies (GWAS), where variants from either SNP arrays or whole-exome sequencing (WES)/whole-genome sequencing (WGS) are connected to phenotype by comparison of large populations of cases with phenotype relative to controls without phenotype (Figure 1). Variants with allele frequencies between 1% and 0.01% are mainly void of statistical strategies to associate variants to phenotypes and remain one of the most challenging genomic variant classes. Once an association of a variant is established, multiple tools are used to move from associations into causal mechanisms for variants. Throughout this educational review article, we explore the background of genomic variants and sequencing, followed by laying forth the tools used for establishing causal mechanisms for common to rare variants in phenotypes and physiology.

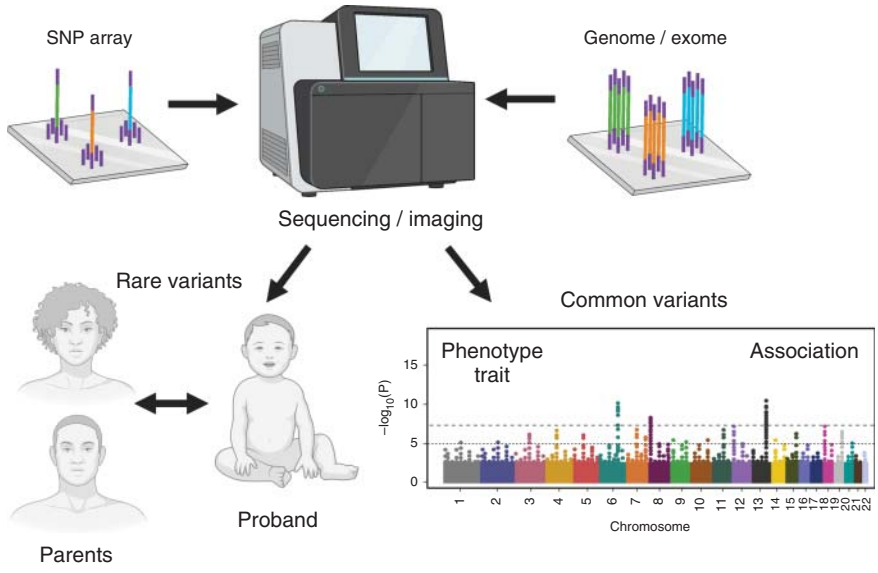


Figure 1 **Sequencing to association studies for rare and common variants**. Created with BioRender. Genomics can be performed by SNP arrays or genome/exome sequencing chemistries, followed by various statistical strategies for rare or common variants.

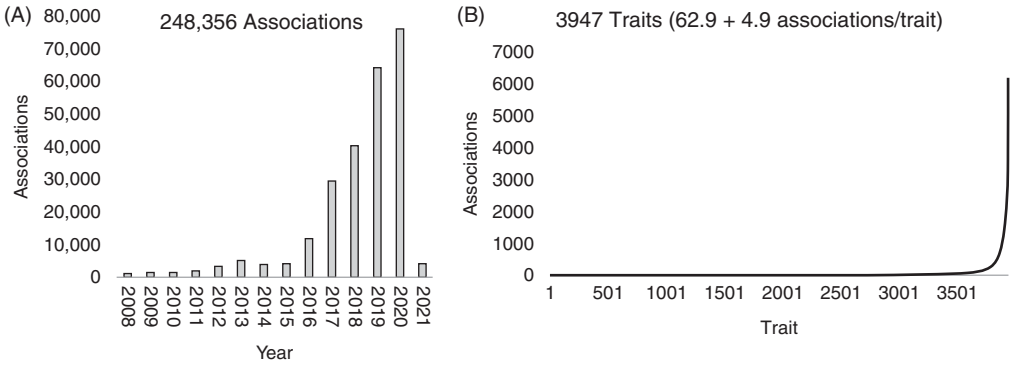


Figure 2 **EBI/NHGRI GWAS catalog**. (A) Associations per year. (B) The number of associations per trait is ranked based on the number of associations. Data was pulled from the GWAS catalog on March 6, 2021.

Common Variants (>1% Allele Frequency)

Human genomics had its most significant breakthrough in the early 2000s with the completion of the compiled draft genome (89). As the genome became confirmed for high-quality insights (141), the initial genome advanced into 1092 (1) and then 2504 individual genomes (2). These 2504 genomes were from 26 populations, allowing for better insights into how variants are shared across individuals of diverse geographical areas that have diverse physiology. In total, 84.7 million SNPs were observed in the first 2504 individuals sequenced. Additionally, 3.6 million indels and 60,000 structural variants were detected. Now, in 2021, with over a million genomes having been sequenced, we know that there are hundreds of millions of variants in the genome and that any site within the genome can vary due to *de novo* mutagenesis. The only way to identify every variant in an individual's genome is through whole-genome sequencing. Yet the cost of a human genome sequence still remains approximately $1000. Using the thousands of genomes sequenced/genotyped, variants that are coinherited together can be identified in linkage disequilibrium (LD), and variant blocks can be imputed through the detection of fewer base calls. A smaller set of variants can be developed into probes, known as a SNP array. A narrowed variant list can be generated, removing redundancy of linked variants, such that variants assessed have the power to impute the tens of millions of variants seen in the initial haplotypes. Common variants are the focus of SNP arrays. These strategies became widely adapted for GWAS and population genomics, including commercial entities of genomics like 23andMe.

GWAS/PheWAS

With these lower-cost strategies of genetic insights, the numbers of individuals with common variant maps grew into the millions. That allows for statistical strategies to identify when variants are associated with biological traits. Focusing on a single trait, GWAS can establish associations of a variant to that trait, yielding statistical p-values for each variant. These studies have been scaled to many traits over the past decade, with a highly curated list present in the EBI/NHGRI GWAS catalog (97). As of March 2021, 248,356 independent associations have been observed for 3947 different traits, yielding 62.9 ± 4.9 associations per trait (Figure 2). The rapid expansion of variant-to-trait associations has been driven by pairing genotypes of individuals to more extensive phenotype collections, such as surveys of traits within 23andMe or medical records of Phenome Wide Association Studies (PheWAS). For example, the UK Biobank has linked 4203 phenotypes to 361,194 individuals with genotype data (155). As the sample sizes continue to increase in these studies, these tools' power for lower allele frequency increases. The use of burden testing in disease association can also allow for variants on the genotyping platform with lower allele frequency to be associated with biological traits (61). To go from an association in GWAS or PheWAS to mechanisms that drive physiology, additional statistical approaches can narrow likely causal variants and potential mechanisms. As with many GWAS and PheWAS, the study design can influence statistical modeling (128). These can be normalized using multi-study integrations or through building insights of the overlap of physiological and variant mechanisms.

Fine-mapping associations with functional data

For each of the variants with trait associations, functionality can be narrowed using statistical approaches or existing data. The use of the GWAS data structure relative to imputed variants, in addition to multiple study overlap of signal allows for a statistical narrowing of likely variants, which can be further filtered with known data for protein and gene biology. These insights can then be used to formulate hypotheses that can be tested with biochemical, molecular, cellular, or animal physiological experiments.

Statistical fine-mapping

As noted, variants connected to biological traits using GWAS or PheWAS are only associations. This is primarily due

to LD, or the co-occurrence of variants inherited together in haplotypes. Any of the variants within LD can be the causal variant for the biological trait. Therefore, finding variants associated with a trait takes additional imputation of variants followed by statistical analysis and integration of datasets for causal variant mechanisms. The base of statistical fine-mapping takes the strong SNP-associated *p*-value sites through additional computation of genotype imputations and cross-study narrowing (140). The initial variant summary statistics can be combined with larger population LD inheritance structures (such as the TOPMed imputation server (37)), additional independent association signals, and a variety of statistical modeling to create a minimum list of likely causal variants (147). These statistical modeling can include heuristic methods, penalized regression, or Bayesian approaches (140). In some cases, based on limited LD variant coinheritance, fine-mapping minimum variant list can nominate a single potential causal variant. However, most of the time, fine-mapping narrows associations to a smaller list of multiple variants that might be causal for a trait. With these narrowed SNPs, screening can assess missense variants that impact protein biology or by identifying variants that impact splicing, expression, or more complex outcomes of genes. These rapid screening of variant insights can be done using tools such as the Variant Effect Predictor (VEP) and ANNOVAR (102, 161).

For example, the SNP rs12126142 is highly associated with interleukin 6 receptor subunit alpha (IL6R) protein measurement (*p*-value 4E-5877), ranking as one of the highest associations in the GWAS catalog. The variant is present at different allele frequencies in different populations and has greater than 20 variants in LD greater than 0.8 R^2 (Figure 3). The fine mapping can narrow a list of causal variants, which then when filtered reveal the rs2228145 SNP (R^2 of 1 with rs12126142) as a missense variant that falls near a splicing site of an alternatively used exon that can serve as a proteolytic cleavage site in activating the soluble receptor (56, 80).

Cross trait associations

As the associations' catalog grows, it becomes increasingly probable that a variant LD block has associations with more than one trait. Thus, by integrating multiple associated traits for a region, one can identify correlations between multiple disease associations into potential gene/protein outcomes. For example, the LD block of rs12126142, as shown in Figure 3, has 18 distinct traits associated within the GWAS catalog (Table 1). The most significant association is to interleukin 6 receptor subunit alpha measurement. It is also connected to diseases (Alzheimer's, asthma, ankylosing spondylitis, abdominal aortic aneurysm, coronary artery disease, rheumatoid arthritis, respiratory system disease), cell level insights (monocyte count, red blood cell distribution width, mean platelet volume), biomarkers (C-reactive protein, fibrinogen, alkaline phosphatase), and environmental responses (mosquito bite reaction itch intensity measurement). These trait integrations give an incredible insight into the mechanisms of a genome location to phenotypic outcomes.

Quantitative trait loci

As suggested from Table 1, it is possible to build associations from genomic loci to proteins or gene changes. Targeted analysis, for example, the measurement of cytokines like IL-6 in blood, can utilize GWAS to determine sites in the genome that contribute to variable expression levels. In the initial assessment screening of variants from GWAS, most loci do not have variants that directly change proteins (missense or nonsense). Therefore, it became increasingly important for new tools to assess how common genetic variants impact gene expression (135, 170). More extensive omic technologies such as transcriptomics or RNA-Sequencing (RNA-Seq) can derive associations between genomic loci and gene expression of each gene in the genome. To make this a reality, the Genotype-Tissue Expression (GTEx) project

rs12126142
(1_154452980_G_A)

Population	Minor allele frequency
African/African-American	0.136
Latino/Admixed American	0.509
Ashkenazi Jewish	0.455
East Asian	0.384
Finnish	0.302
Non-Finnish European	0.377
Non-Finnish European Estonian	0.342
Non-Finnish European North-Western European	0.390
Non-Finnish European Southern European	0.528
Other (population not assigned)	0.331

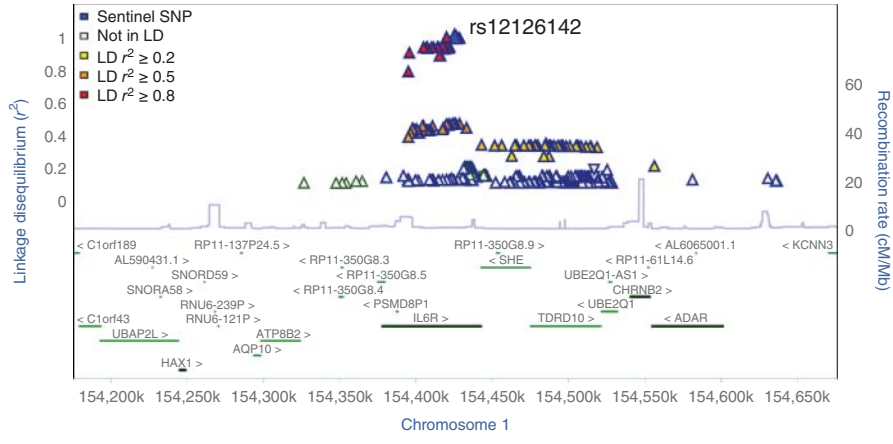


Figure 3 **Population variants and linkage disequilibrium for rs12126142.** The population allele frequency is listed to the left in diverse populations based on gnomAD. Below rs12126142 is listed the chromosome annotation (chromosome_location_wildtype_minor allele). LD was imputed from the 1000 genomes project phase 3 using European imputation.

Table 1 Traits Associated with the rs12126142 LD Block Pulled from the GWAS Catalog (97) on March 16, 2021

Mapped_trait	*p*-Value
Interleukin 6 receptor subunit alpha measurement	4E-5877
C-reactive protein measurement	3E-436
Alzheimer's disease	6.00E-63
Fibrinogen measurement	3.00E-36
Alkaline phosphatase measurement	2.00E-31
Monocyte count	1.00E-23
Asthma	1.00E-17
Mean corpuscular volume	8.00E-17
Red blood cell distribution width	1.00E-16
Ankylosing spondylitis	2.00E-15
Mean corpuscular hemoglobin	5.00E-15
Mean platelet volume	7.00E-15
Abdominal aortic aneurysm	5.00E-13
Coronary artery disease	3.00E-11
Rheumatoid arthritis	1.00E-10
Platelet count	2.00E-10
Mosquito bite reaction itch intensity measurement	9.00E-09
Respiratory system disease	9.00E-08

was launched (59, 60). In approximately 1000 individuals, a total of 54 different tissues were collected, followed by RNA-Seq that were combined with the individual's genomic data.

The RNA-Seq is paired to genotyping information for each individual, allowing for transcriptome-wide analysis of variants' influence on genes, known as a quantitative trait loci (QTL). The current GTEx release (v8) focuses on both the outcomes of variants to expression (eQTL) and splicing (sQTL). The number of genes in all tissues with eQTLs greatly outnumbers sQTLs (Figure 4A). For eQTLs, the relative change in the mapping of RNA reads to a gene for different genotypes (fold change) relative to the significance (*p*-value) shows that thousands of sites in the genome impact expression (Figure 4B). The sQTLs similarly have thousands of associations and represent a broad array of minor allele frequencies (Figure 4C).

To detail eQTL biology, we selected one of the most significant eQTLs, that at rs6593279 for the *PSPHP1* gene (*p*-value of 8.3e-252). The violin plot of normalized expression of *PSPHP1* when individuals are homozygous for G at rs6593279 (chr7 55,736,277) shows marked lower expression than either the heterozygous (GA) or homozygous A (AA) individuals (Figure 5A). This tiered expression of the three genotypes at the loci shows the additive effects of homozygous individuals' expression. It should be noted that the rs6593279 variant is in high LD with multiple additional SNPs (Figure 5B). Thus, the causal SNP is not determined by eQTL, requiring further data analysis to determine causal variants on gene expression.

Similar to the eQTL example, sQTLs show changes in splicing and exon usage that are associated with genetic variants. As an example, we show rs56105022 impact on splicing for the *CNIH4* gene. The variant is one of the most significant sQTLs within the genome (*p*-value <1e-300), with the signal in tissues including muscle and cultured fibroblasts (Figure 6A). The heterozygous individuals with CA have marked elevation (Figure 6A) of alternative splicing of exon 1 to exon 4, resulting in an altered CNIH4 splicing (Figure 6B-6C) and a resulting isoform that does not code for a protein (Figure 6D). It should also be noted that no homozygous A (AA) individuals are part of the GTEx analysis. In the case of this variant, the LD structure (Figure 6E), sQTL, and isolated presence of rs56105022 over the *CNIH4* gene suggest the narrowing of a causal variant to a single SNP, showing how fine-mapping can at times suggest single causal variants.

Building overlap of eQTL/sQTL signals for any two loci associations requires some careful consideration, performed through a colocalization of signal analysis (Figure 7). In these analyses, if the two associations' peaks are located on the curve's exact position, then the causal variants are likely to be the same for the two traits (Figure 7A). However, it is possible that the two peaks do not overlap, yet both have SNPs that reach significance due to LD (Figure 7B). In the latter case, the causal variants are likely to be different even though the same SNPs have shared significance for both traits. Accounting for colocalization of signals decreases the probability of false-positive trait overlap annotations.

ENCODE and epigenomics data

Narrowing from associations to causality for an individual variant, even with knowledge of cellular or gene-level mechanisms using the above approaches, requires screening gene regulation mechanisms at the variant resolution. The ENCODE project was launched and has completed three phases to provide a starting point of resolution for base-pair regulation (45–47). Using various cell lines and tissues, 18,905 different assays have been performed to give insights into transcription factor-DNA binding (ChIP-Seq), gene expression (RNA-Seq), DNA accessibility (DNase hypersensitivity, footprinting, ATAC-Seq), DNA methylation (bisulfite sequencing, RRBS), histone modifications (ChIP-Seq), and 3D chromatin structure (Hi-C). Integrated insights of the epigenetics data for individual variant sites such as CADD (130), state models of regulation like ChromHMM (136), or RegulomeDB (21) give a detailed map of overlapping data to prioritize the minimal fine-mapped variants of an association region (Figure 8). Additional tools such as GeneHancer (50) can link the regulation datasets to genes.

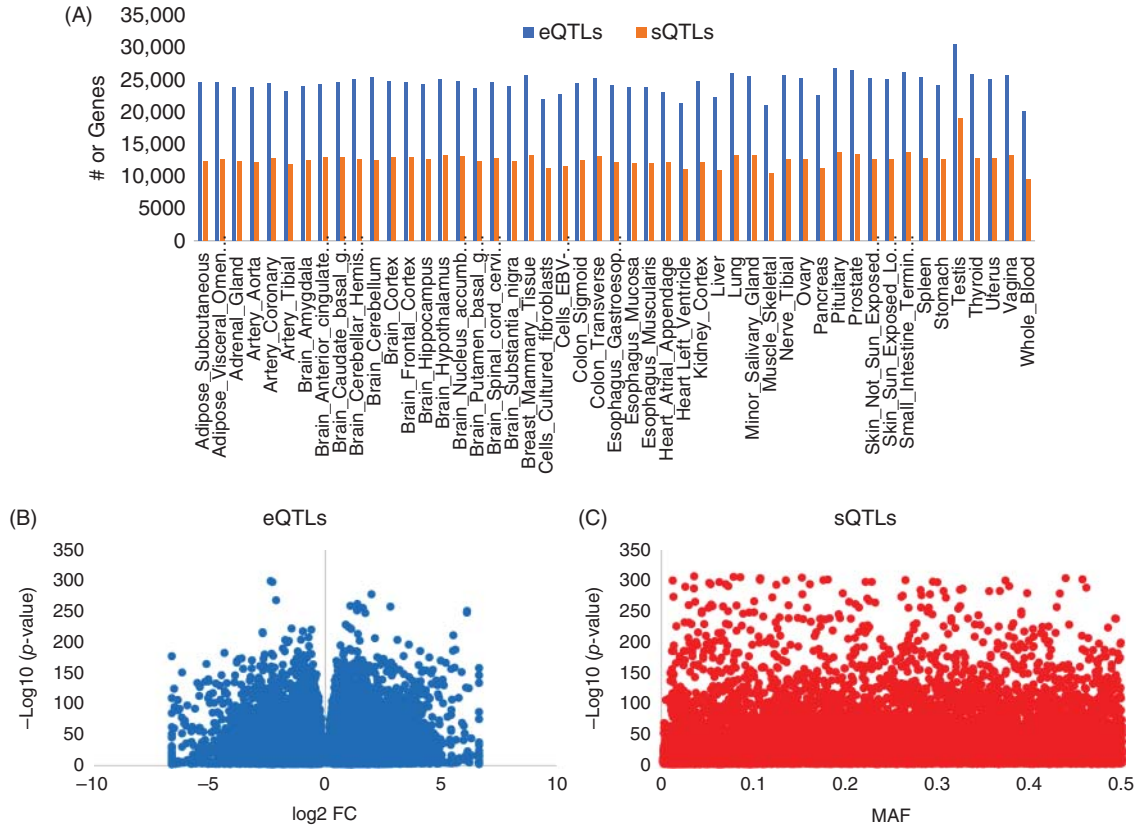


Figure 4 **GTEx v8 eQTL and sQTL analysis.** (A) The number of genes with GTEx annotated significant eQTLs (blue) or sQTLs (red) in each of 49 different tissues. (B) A volcano plot of log2 fold change (*x*-axis) relative to the −log10 *p*-value (*y*-axis) for all significant eQTLs. (C) A plot of mean allele frequency (MAF) relative to the -Log10 *p*-value (*y*-axis) for all significant sQTLs.

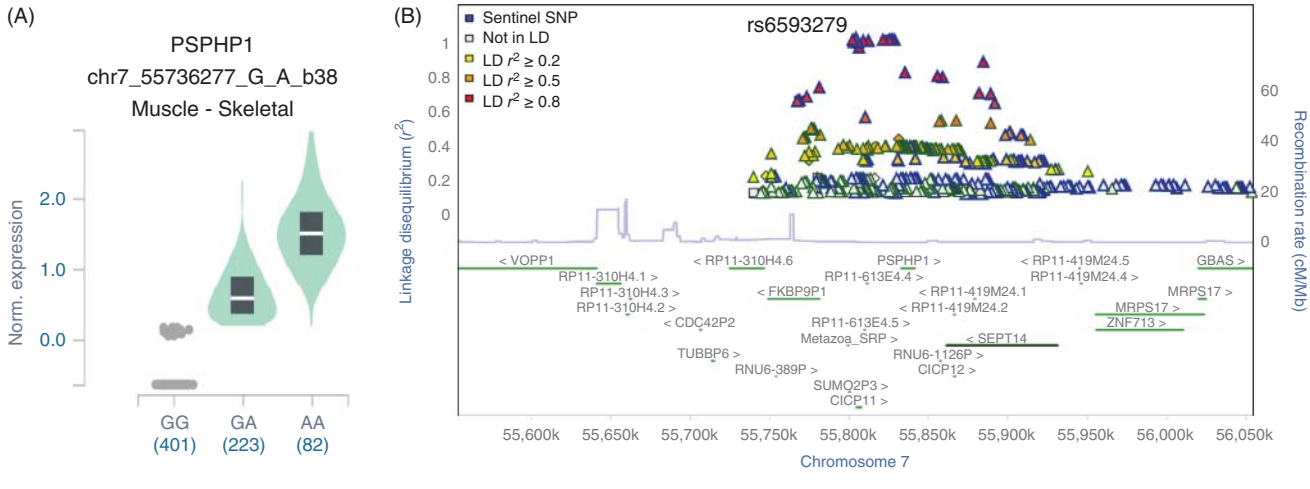
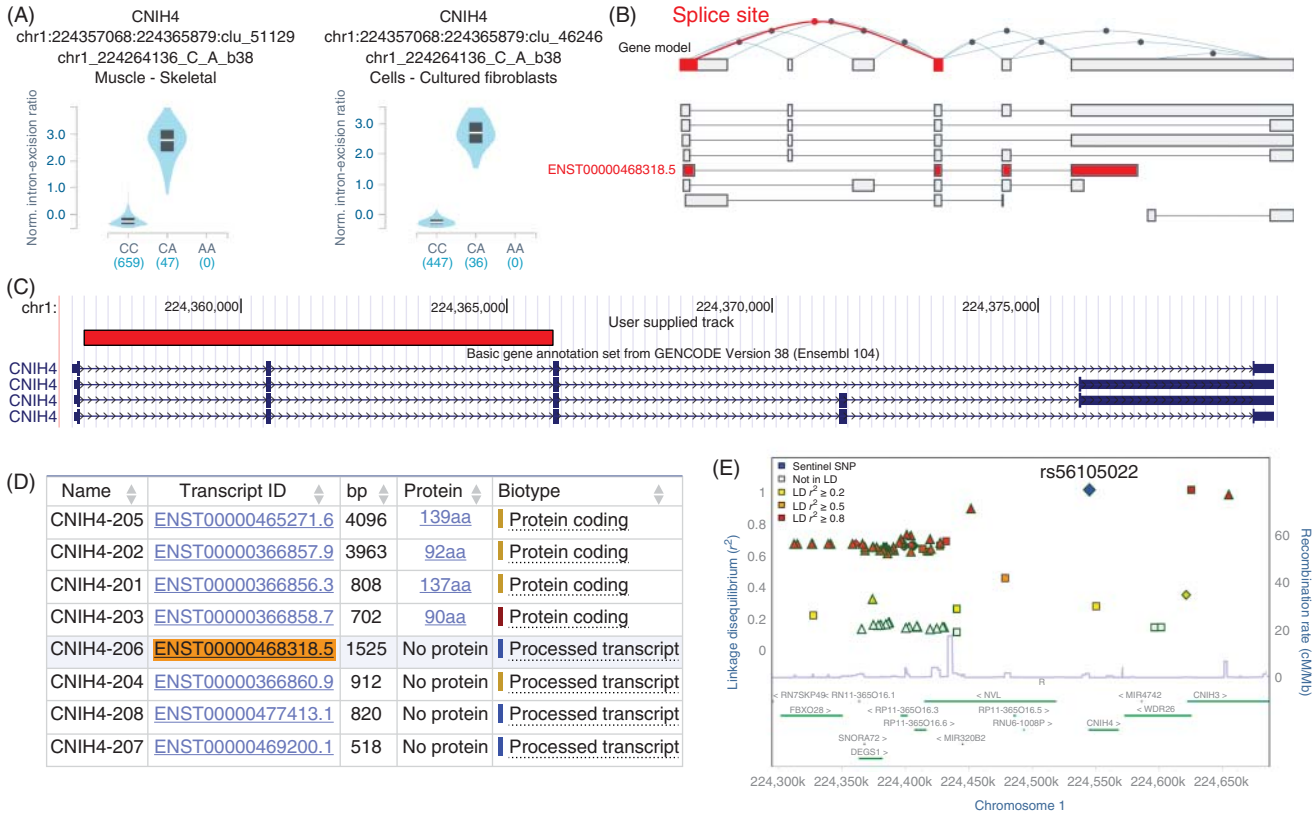


Figure 5 **The eQTL for PSPHP1 on chromosome 7**. (A) Violin plot rs6593279 in skeletal muscle tissue for each genotype. (B) LD analysis for rs6593279 using 1000 genomes project phase 3 East Asian imputation.

Noncoding RNA

As sequencing technologies for transcripts increase, our knowledge of RNA classes outside of those that code for proteins has advanced. Variants can fall within these noncoding RNA molecules and impact cellular function, such as gene regulation. The current human Gencode38 database (https://www.gencodegenes.org/human/) contains 236,186 known transcripts (Figure 9A), of which 63% are not protein-coding annotation (53). The long noncoding RNA (lncRNA)



Name	Transcript ID	bp	Protein	Biotype
CNIH4-205	ENST00000465271.6	4096	139aa	Protein coding
CNIH4-202	ENST00000366857.9	3963	92aa	Protein coding
CNIH4-201	ENST00000366856.3	808	137aa	Protein coding
CNIH4-203	ENST00000366858.7	702	90aa	Protein coding
CNIH4-206	ENST00000468318.5	1525	No protein	Processed transcript
CNIH4-204	ENST00000366860.9	912	No protein	Processed transcript
CNIH4-208	ENST00000477413.1	820	No protein	Processed transcript
CNIH4-207	ENST00000469200.1	518	No protein	Processed transcript

Figure 6 **GTEx sQTLs for rs56105022 on CNIH4.** (A) Violin plots for variants in skeletal muscle and cultured fibroblasts. (B) Exons of CNIH4 with splicing shown. In red is the alternative splicing event of panel A with the resulting splicing isoform shown below in red. (C) The genome browser view of the splicing site (red) in respect to the exon structure and chromosome positions. (D) The Ensembl table of isoforms with the resulting splice site highlighted, which does not code for a protein. (E) LD imputed from the 1000 genomes project phase 3 using European imputation for rs56105022.

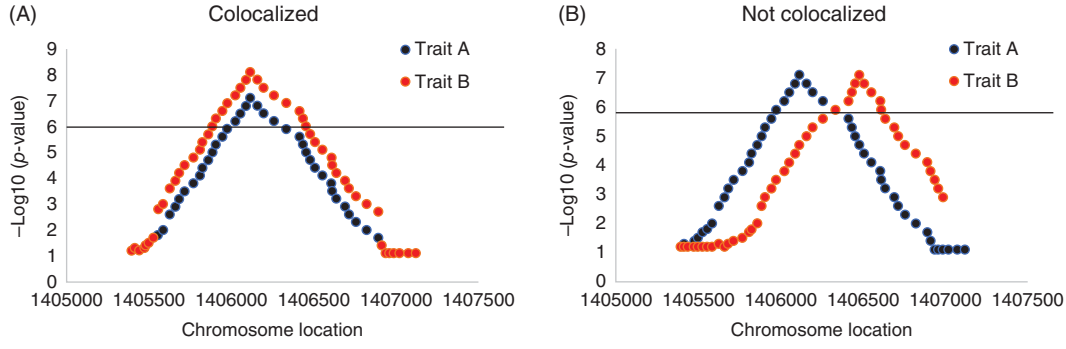


Figure 7 **Representative plot of colocalized (A) or not colocalized (B) associations between two traits.** The black line represents the significance cutoff.

transcripts represent 20% of all known transcripts, with 46,963 annotated, ranking as the largest class of the non-coding RNA (Figure 9B). Using the Gencode38 annotation statistics, these noncoding RNA are vastly under published relative to protein-coding transcripts (Figure 9C), making it incredibly difficult to interpret functional consequences of variants in these transcripts. Many noncoding RNA is not conserved between species and thus evolutionary analysis cannot be applied to defining the functional nucleotides of the noncoding RNA (42). Even when noncoding RNA are found between species, their expression can vastly differ, such that animal modeling of the human noncoding RNA has been a robust challenge in defining their roles in human diseases (94). In addition, many of these lncRNA transcripts are quite large, averaging 1318 ± 2247 bases per transcript (Figure 9D) with the largest that of XACT-203 (ENST00000674361 = 347,561

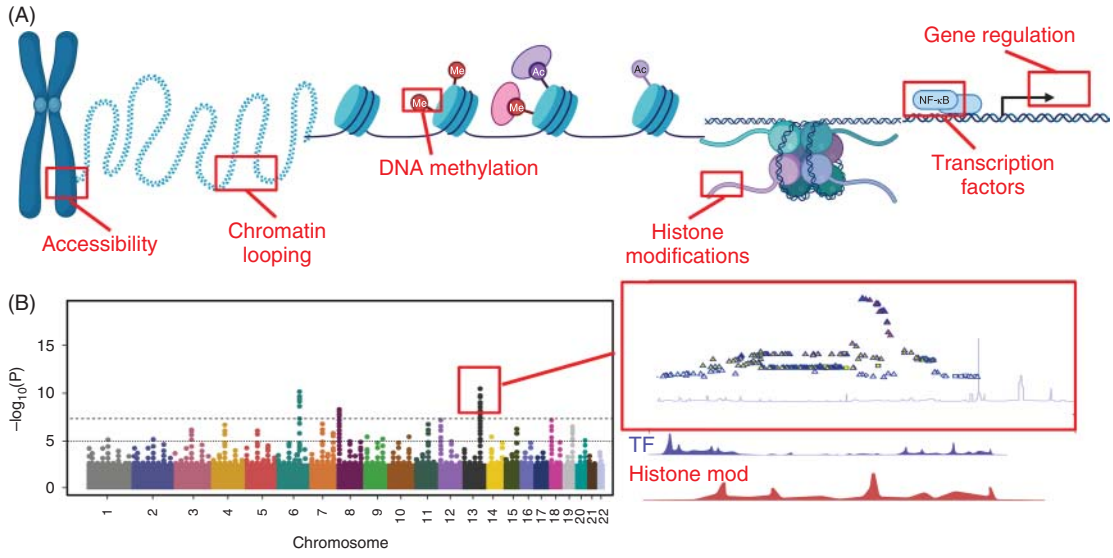


Figure 8 **Epigenetic regulation dataset overlap to association regions.** (A) Schematic of various epigenetic insights measured in ENCODE. (B) Representative GWAS hit in red box looking at the overlap of peaks for a transcription factor (TF, blue) or histone modification (red) relative to LD block for the site. Created with BioRender.

bases). Yet, multiple examples do exist for functional variants within or contributing to noncoding RNA. The most common mechanistic studies for noncoding RNAs involve changes in their expression, likely due to eQTLs, as these can be readily measured through total RNA-Seq strategies. Various GWAS for cardiovascular diseases (myocardial infarction, coronary artery disease, type-2 diabetes, blood pressure, etc.) have suggested disease involvement and expression changes of lncRNAs such as *ANRIL, CDKN2B-AS1, MIAT, H19*, and *LOC157273* (27, 58, 167). In autism spectrum disorder (ASD) and Development Delay/Intellectual Disability (DD/ID), the lncRNAs of *MSNP1AS, ASFMR1, ATXN8OS, BACE1-AS, BC200, MALAT1*, and *SOX2OT* are all known to contribute to disease progression (81, 151, 162). The ncRNAVar database (www.liwzlab.cn/ncrnavar/ncrnavar.html) has a curated list of 3112 variants that contribute to 711 different human diseases through noncoding RNA from common to rare disorders/phenotypes (171). Most of these variants are found associated with lncRNA (3203 associations, Figure 9B) followed by miRNA (622 associations). One of the more striking observations from this database is that most variants connected to noncoding RNAs are those that influence expression and are found intergenic, downstream, upstream, or intronic of transcript annotations (Figure 9F). Advancing resources in noncoding RNA show growth but also the need to better understand the mechanisms and physiology associated through noncoding RNA.

Pharmacogenomics

One area of common genomics that is actively growing is that of pharmacogenomics. The PharmGKB database (67) represents the most extensive integrated knowledgebase of pharmacogenomics, giving the ability to search any disease, gene, or variant and get to drug label annotation, clinical guideline, gene pathway, or annotated drugs. Pharmacogenomics is primarily linked to the genome's common variants, where GWAS strategies can be applied to populations in clinical trials to stratify drug delivery based on genotypes. A smaller subset of pharmacogenomics focuses on rare variants and rare diseases, such as cystic fibrosis.

Example variants in broad phenotypes

To show examples of how association data can be taken to mechanisms using the discussed statistical filtering and annotation tools, we have provided five examples of GWAS loci (Figures 10–14). Logically, many of the loci with causal variants determined are those sites where a pronounced missense variant exists. Computation tools such as PolyPhen2 (5), Provean (32), and SIFT (108) that use evolutionary data and amino acid functionality can narrow missense variants that likely cause a change to protein function. A total of 12,242 variant associations are seen for 3333 missense variants within the GWAS catalog, of which 102 missense variants (Figure 10A) are predicted to damage protein function using PolyPhen2 (5), Provean (32), and SIFT (108) filtering tools. There are only a total of 31 missense variant annotations that are predicted functional in the three tools and have a MLOG >20 (Table 2).

The missense variant with the lowest *p*-value within the entire GWAS catalog is ALDH2 E504K (rs671, *p*-value 1E-4740). rs671 is found listed in the GWAS catalog 37 times, mostly connected to alcohol drinking, consumption, or response (99, 125, 149). It is also linked to various cardiovascular outcomes such as gout, coronary artery disease,

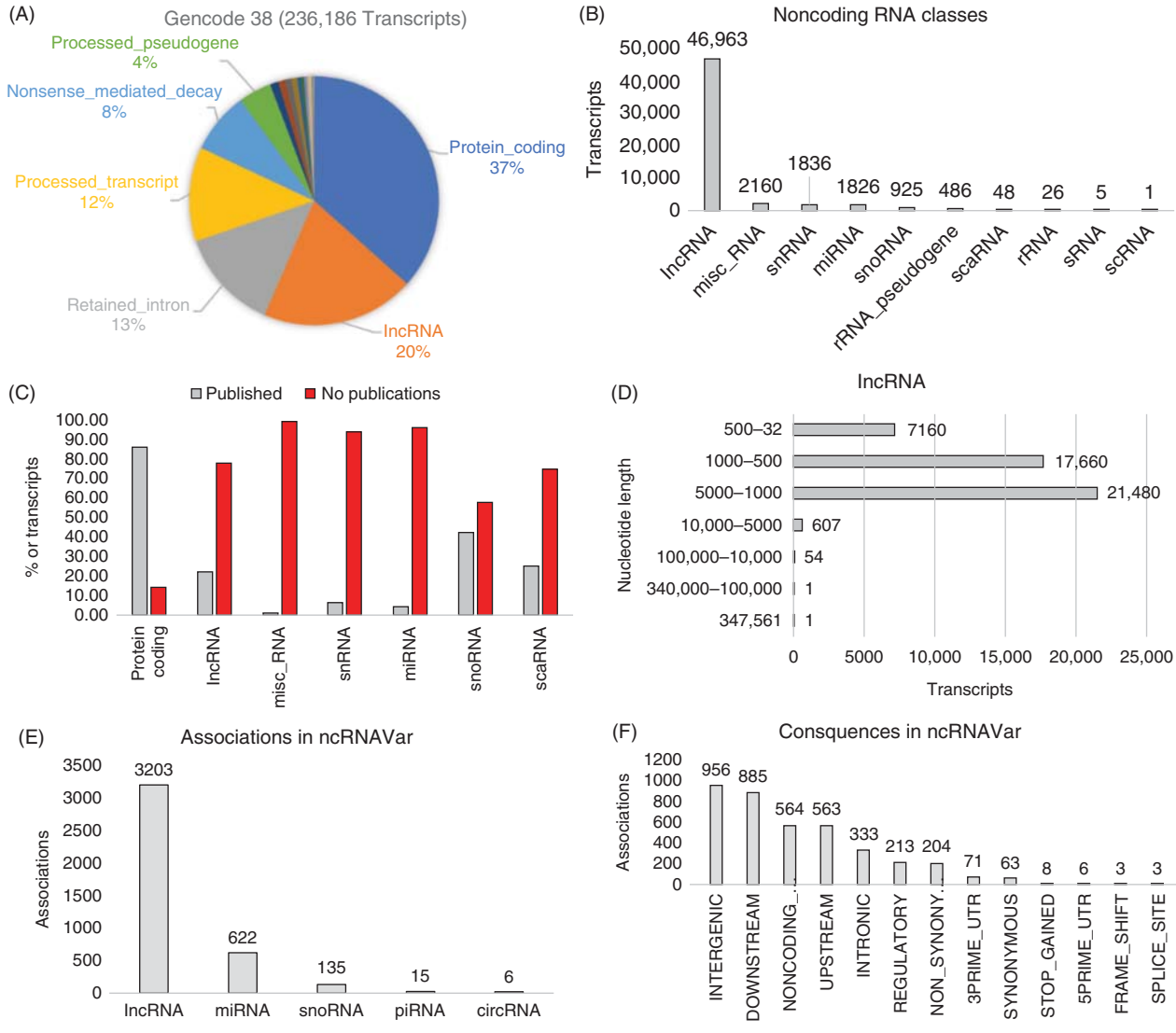


Figure 9 **Noncoding RNA.** (A) A breakdown of the biotype annotations within the Gencode38 database. The top 6 groups are labeled with the percentage of total transcripts. (B) The number of transcripts for several of the noncoding RNA groups. (C) The percent of transcripts in each biotype that have publications (gray) or do not have publications (red). (D) The size of long noncoding RNA (lncRNA) transcripts. (E-F) The number of disease associations (E) or VEP (F) annotated consequences for various noncoding RNA found within the ncRNAVar database (www.liwzlab.cn/ncrnavar/ncrnavar.html).

metabolic syndrome, and myocardial infarction (106, 150, 165, 172). The E504K variant is found near the enzyme's active site (Figure 10B) and is known to significantly reduce enzyme activity as determined by biochemical experiments (78). Homozygous knockout of *Aldh2* in mice results in multiple phenotypes connected to cardiovascular (https://www.mousephenotype.org/data/genes/MGI:99600#phenotypesTab) and ethanol processes (96).

Genetics and variants play a substantial role in neurological diseases, particularly in Alzheimer's, the most common form of dementia (173). Heritability is estimated to be up to 79% based on twin and family studies (57, 134). Alzheimer's can roughly be divided into early-onset (EOAD) and late-onset (LOAD), both of which GWAS, linkage studies, as well as other imaging modalities have identified autosomal dominant and sporadic genes associated with these conditions (173). Of note, *APP*, *PSEN1*, and *PSEN2* have been genes of interest that have been implicated in the pathogenesis of EOAD. For LOAD, *APOE* is a well-studied risk gene, with newer implicated in GWAS studies including *ABCA7*, *CLU*, *CR1*, and *DRB1*. There are currently 1361 association loci for Alzheimer's disease (Figure 11A). The variant rs429358, also known as APOE*4 allele, is found elevated in African/African American ancestry (Figure 11B), has independent inheritance without highly correlated LD (Figure 11C), and results in a missense variant APOE C130R (Figure 11D). The APOE star 4 allele is one of the most penetrant adult genotype-phenotype common variants, increasing Alzheimer's disease risk from 20% to 90% (35). This missense variant results in changes to the APOE protein structure and interaction with

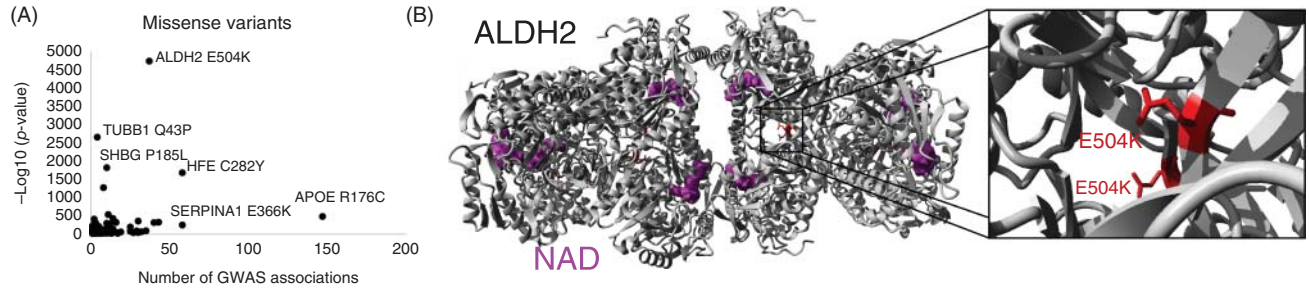


Figure 10 **Missense variants in GWAS LD blocks.** (A) Plot of GWAS associations (*x*-axis) from the GWAS catalog and the top-ranked *p*-value (*y*-axis) for that variant. Outlier missense variants are labeled. (B) The ALDH2 missense variant E504K.

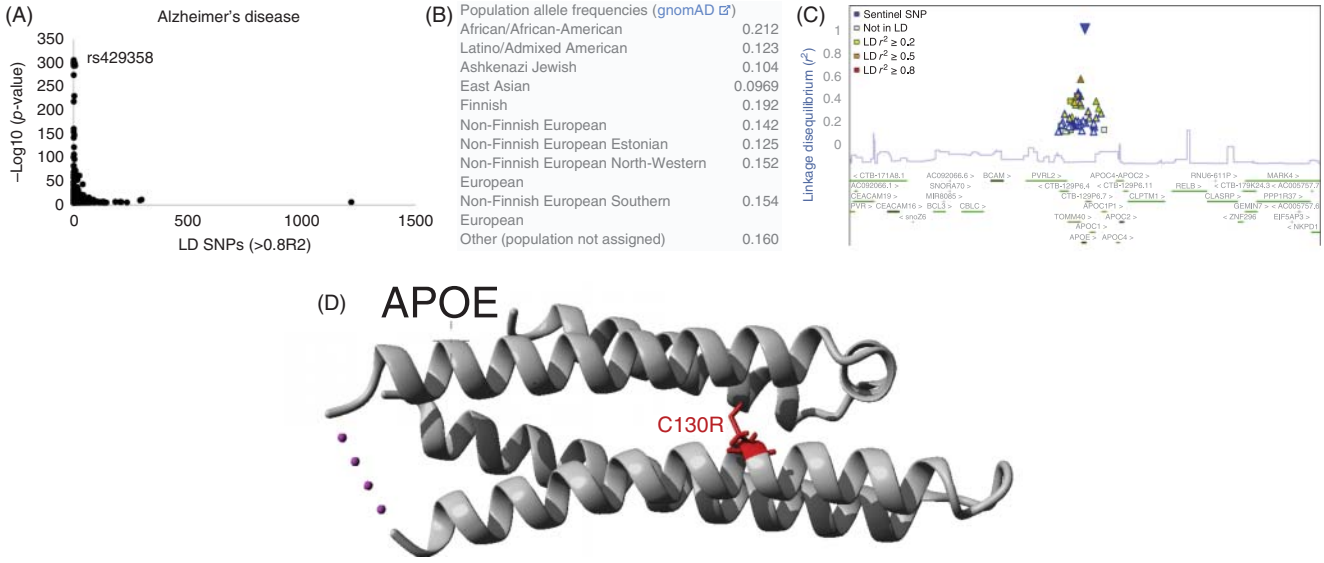


Figure 11 **Alzheimer's Disease GWAS.** (A) All SNPs in GWAS database for Alzheimer's Disease showing the number of LD SNPs (*x*-axis) and the significance (*y*-axis). (B) rs429358 allele frequencies in gnomAD populations. (C) LD imputed from the 1000 genomes project phase 3 using African imputation for rs429358. (D) rs429358 results in the missense variant APOE C130R.

APP/A4 amyloid-beta peptide that results in loss-of-function to the protein in neurons and is associated with elevated plasma cholesterol and triglyceride levels (39). The variant is amongst the most published variants, with over 100 current publications in PubMed (https://pubmed.ncbi.nlm.nih.gov/?term=rs429358&sort=pubdate), many of which detail mechanisms of physiology due to the variants. This example represents where a causal missense variant can be confidently mapped with fine-mapping strategies.

Stroke risk factors fall under modifiable and nonmodifiable risks, with numerous suggested genetic factors (158). These mutations can be rare single-gene disorders (such as cerebral autosomal dominant arteriopathy), single genes causing multisystem effects (such as sickle cell anemia), to common polygenetic variants. Common variants have been associated with increased stroke risk through genes including *TSPAN2*, *FOXF2*, *ABO*, and *PITX* (20). These loci, such as polymorphisms in 9q21, play a modest role in stroke development in patients. There are 353 loci for stroke in the GWAS database (Figure 12A). The rs6025 variant is amongst the most significant associations (70), with the minor allele found highest in Non-Finnish European ancestry (Figure 12B) with few SNPs in LD (Figure 12C). rs6025 results in a missense mutation F5 R534Q (Figure 12D), which is also associated with Budd-Chiari syndrome and is a driver for thrombosis complications (16, 29).

In total, 523 loci have an association with CKD (Figure 13A). The rs2147896 variant is the most significant CKD locus (*p*-value 3E-917), is found elevated in African/African American ancestry (Figure 13B), has a complex LD block of variants (Figure 13C), and has a pronounced eQTL for *PYROXD2* (Figure 13D). While this association is strong, there remains a surprisingly low number of publications on the gene. Thus, physiological validation using diverse techniques are needed to determine what the outcomes of changed *PYROXD2* expression can have on humans or model organisms.

Throughout many of the top LD blocks for traits within the GWAS database, causal variants are relatively thin, outside

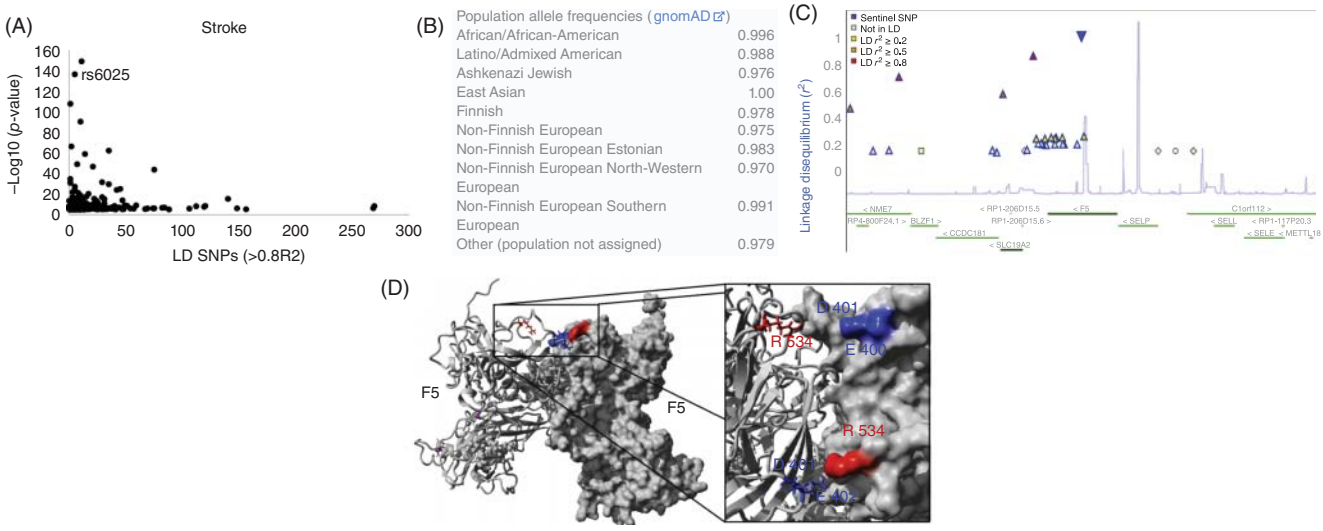


Figure 12 **Stroke GWAS.** (A) All SNPs in GWAS database for stroke showing the number of LD SNPs (*x*-axis) and the significance (*y*-axis). (B) rs6025 allele frequencies in gnomAD populations. (C) LD imputed from the 1000 genomes project phase 3 using American imputation for rs6025. (D) rs6025 results in the missense variant F5 R534Q.

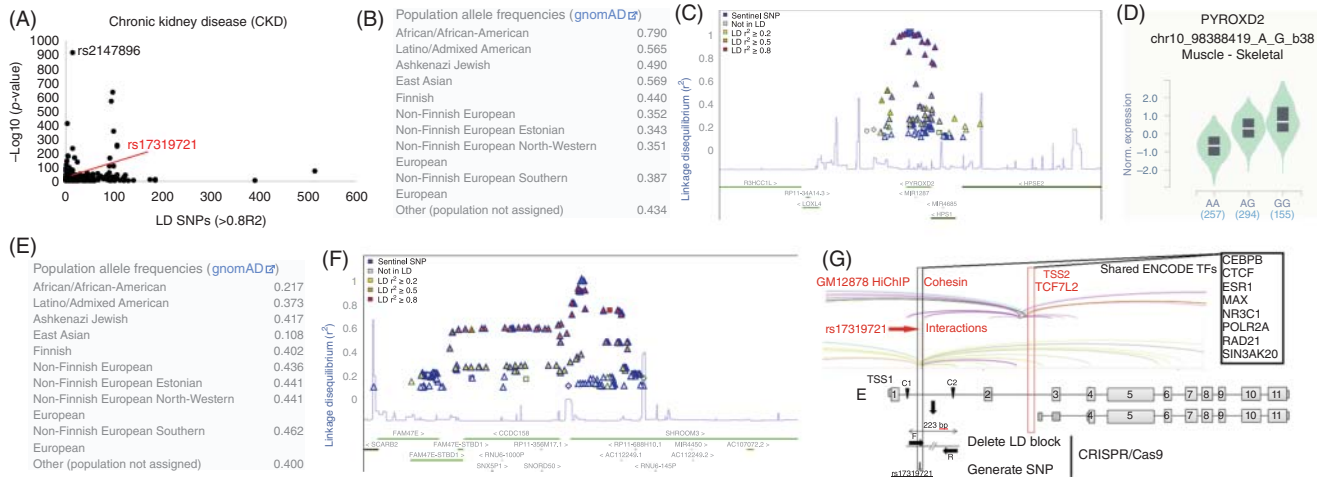


Figure 13 **Chronic Kidney Disease (CKD).** (A) All SNPs in GWAS database for Chronic Kidney Disease (CKD) showing the number of LD SNPs (*x*-axis) and the significance (*y*-axis). (B) rs2147896 allele frequencies in gnomAD populations. (C) LD imputed from the 1000 genomes project phase 3 using African imputation for rs2147896. (D) rs2147896 results in an eQTL for PYROXD2. (E) rs17319721 allele frequencies in gnomAD populations. (F) LD imputed from the 1000 genomes project phase 3 using American imputation for rs17319721. (G) rs17319721 results in altered gene regulation of a shortened SHROOM3 isoform through disruption of TCF7L2 binding and looping between an enhancer and promoter.

of missense variants. Our group defined one specific LD block to determine the causality of a single change on gene regulation, focusing on rs17319721 impact on *SHROOM3* in chronic kidney disease (CKD). rs17319721 is found elevated highest in Non-Finnish Southern European (Figure 13E) and has multiple variants in LD (Figure 13F). Using data filtering, electrophoresis mobility shift assays (EMSA), CRISPR modifications, and a zebrafish nephrology model, our group showed that the single variant disrupts TCF7L2 transcription factor binding, impacting structural looping of an enhancer to a secondary transcriptional start site that is used in podocytes (Figure 13G). The disruption of this enhancer by the variant directly impacts the expression of *SHROOM3* (124). That work represents a strategy for going from association to causal variant mechanisms, where many future sites must be analyzed for the GWAS catalog to take association into mechanism.

Rare Variants (<0.01% Allele Frequency)

Rare variants are often connected to rare Mendelian diseases through high penetrance. Rare variants can also have impact

Table 2 Missense Variants in GWAS Catalog Annotated as Damaging to Protein Function Using PolyPhen2, Provean, and SIFT from the GWAS Catalog (97) on March 16, 2021

SNP	UniProt	*Gene*	Variant	GWAS traits	Top mapped trait	Top MLOG
rs671	P05091	*ALDH2*	E504K	37	Alcohol drinking	4740.00
rs463312	Q9H4B7	*TUBB1*	Q43P	4	Platelet component distribution width	2658.40
rs6258	P04278	*SHBG*	P185L	10	Sex hormone-binding globulin measurement	1823.00
rs1800562	Q30201	*HFE*	C282Y	58	Mean corpuscular hemoglobin	1685.00
rs12975366	O75023	*LILRB5*	D247G	8	Blood protein measurement	1275.52
rs10490924	P0C7Q2	*ARMS2*	A69S	11	Age-related macular degeneration	539.40
rs7412	P02649	*APOE*	R176C	147	Blood protein measurement	483.52
rs1801690	P02749	*APOH*	W335S	14	Blood protein measurement	430.15
rs1043657	O43488	*AKR7A2*	A142T	2	Chronic kidney disease, urinary metabolite measurement	411.70
rs2228467	O00590	*ACKR2*	V41A	30	Monocyte count	360.40
rs4149056	Q9Y6L6	*SLCO1B1*	V174A	43	Blood metabolite measurement	327.22
rs678	P19827	*ITIH1*	E585V	16	Blood protein measurement	320.52
rs16891982	Q9UMX9	*SLC45A2*	L374F	40	Sunburn	319.52
rs11547464	Q01726	*MC1R*	R142H	3	Hair color	307.70
rs1805006	Q01726	*MC1R*	D84E	3	Hair color	307.70
rs6025	P12259	*F5*	R534Q	10	Venous thromboembolism	300.00
rs28385609	Q92484	*SMPDL3A*	P161S	2	Blood protein measurement	281.70
rs1042602	P14679	*TYR*	S192Y	12	Hair color measurement	278.52
rs1048328	Q9UBX7	*KLK11*	R166C	3	Blood protein measurement	261.00
rs28929474	P01009	*SERPINA1*	E366K	58	Sex hormone-binding globulin measurement	252.00
rs676210	P04114	*APOB*	P2739L	25	Triglyceride measurement	196.22
rs34210653	P16050	*ALOX15*	T560M	14	Eosinophil count	139.10
rs34557412	O14836	*TNFRSF13B*	C104R	35	Monocyte count	93.40
rs17580	P01009	*SERPINA1*	E288V	29	Blood protein measurement	89.00
rs1801689	P02749	*APOH*	C325G	25	Platelet count	80.70
rs9379084	Q92766	*RREB1*	D1171N	32	Body height	64.40
rs167479	Q3MIN7	*RGL3*	P162H	19	Systolic blood pressure	62.00
rs12210538	Q86VW1	*SLC22A16*	M409T	16	Reticulocyte measurement	46.70
rs2277339	P49642	*PRIM1*	D5A	30	Mean corpuscular volume	42.22
rs34536443	P29597	*TYK2*	P1104A	27	Platelet count	38.00
rs2229742	P48552	*NRIP1*	R448G	17	Erythrocyte count	20.30

The list includes all missense and damaging variants with an MLOG > 20 (*p*-value 10e-20).

on common diseases, potentially explaining much of the unknown disease inheritance not explained by common variants (33). A disease is defined as rare if it affects fewer than 200,000 people in the United States. Since most rare diseases primarily affect children and have a strong penetrant genetic component, this has promoted pediatric genetics and genomics as a growing medical specialty (163). The development of rapid genetic sequencing has revolutionized the medical management of critically ill children, with the potential of finding a diagnosis in as little as 19.5 h (34). This rapid turnover allows for more timely and efficient medical management, leading to better health outcomes and comfort for families (83). It is crucial that this improved turnaround time also be translated to smaller community hospitals that typically do not have the access to sequencing technology at the scale of larger academic hospitals, which

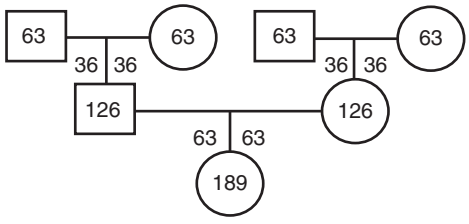


Figure 14 **Inheritance and *de novo* rare variants of the genome**. 63 is a representative number of *de novo* variants, with each individual of the pedigree have added variants in subsequent generations.

have demonstrated the ability to utilize this technology with great utility, while still maintaining cost-effectiveness (34).

The 1000 genomes project was the first to conclusively show that each individual has 42 to 82 unique variants not found in either parent (4). These changes are known as *de novo* variants. These *de novo* variants compound over generations, where half of one's germline *de novo* variants are passed on to children. This results in a pool of variants inherited in local family structures (Figure 14). With an average of 63 changes between an individual's parents and their own *de novo* variants, approximately 126 variants are found. Extending back generations of *de novo* accumulation would suggest >1000 variants arising in this way looking at the 16th generation of inheritance. These accumulated *de novo* variants represent the bulk of analysis for rare variants.

Our current statistical tools for identifying rare variants with phenotypic associations are based on comparing an individual with a phenotype to a large population of genomics without the phenotype. As rare variant analysis currently focuses on highly penetrant variant detection, assessing variants is limited by the number of "healthy" control genomes. Over the past few years, science has contributed to a boom in genomes sequenced. Projects such as gnomAD contain 76,156 genomes of diverse inheritance, and the TOPMed program has grown to 168,220 genomes (Figure 15). This large base allows for a patient's genome to be compared to remove all variants seen in the larger asymptomatic population. That analysis still leaves hundreds of potential damaging variants. In this case, it is preferred to compare the patient to their parents to determine *de novo* variants and inheritance structure. However, much of genomic sequencing has required identifying candidate regions of the genome for phenotypes, starting from the larger structural genomics field that has been around longer than sequencing, and now moving into clustering patients with overlapping gene signatures.

Structural genomics to disease

Large chromosomal structural changes, known as copy number variation (CNV), were amongst the first clinical genetic changes studied (26). Using karyotyping and multicolor spectral karyotyping, it is possible to visualize entire chromosome replications (such as trisomy 21) and substantial changes (142). Advancing microarray-based platforms allowed for scaling of CNV detections clinically (28, 117). CNVs account for genetic contributions for many diseases (14, 48), where targeted analysis is often performed for suspected diseases. With most CNVs, the focus has been on the gain or loss of one or more copies of a gene and how it impacts physiology, known as gene dosage sensitivity. The ClinGen database (129) has become the leading curated database for dosage sensitivity information at the gene level. As of March 8, 2021, there are 1457 dosage-sensitive genes curated in ClinGen. The removal of one copy of the gene to result in disease, haploinsufficiency, has sufficient evidence for 353 genes. In contrast, the gain of a copy of the gene, triplosensitivity, is only sufficient in 24 genes. The ClinGen database also includes curated information for clinical actionability and pathogenicity for genes.

The use of dosage sensitivity is significant for determining loss vs gain of function for a gene association to the disease. If CNVs establish haploinsufficiency for a gene to a specific disease, a single variant site (missense) in an individual with the same disease can be used as a filter to identify potential loss-of-function causal variants in the individual's genome. Similarly, if triplosensitivity has been observed for a disease, single variants can be screened for gain-of-function outcomes on the protein. Therefore, the curation of dosage sensitivity information lends important tools for assessing single variants to physiological outcomes. This is incredibly important recently with the growth of long-read sequencing of Pacific Biosciences circular consensus sequencing, able to determine smaller CNVs not assessed in karyotyping or array technologies, vastly improving the clinical detection of CNVs in diseases (68).

Clinical variant database (ClinVar)

The gain or removal of a copy of a gene is logically easier to understand proteins' outcome. More subtle changes, however, account for most genotype-phenotype relationships. Therefore, focusing on single base changes, databases such as ClinVar (90) have grown lists of hundreds of thousands of observed variants seen in patients with broad diseases. It should be noted that ClinVar is not curated and thus is based on submitters' annotation of variants. Care must be taken when extracting variant annotations into clinical or research interpretation. Dissection of the nearly million variants within ClinVar shows that variants of uncertain significance (VUS) account for 43% of variants, while pathogenic and likely pathogenic account for only 16% (Figure 16A). Variants are broken down into the following common groups within ClinVar (Figure 16B).

Missense (44.8%): Results in an altered amino acid of the coded protein. Example = NM_152486.3(SAMD11): c.106G>A (p.Ala36Thr), where c. represents the coding transcript position, and p. represents the protein position.

Synonymous (19.2%): Does not alter the sequence of the protein. These variants typically fall in codon wobble

gnomAD

Population	Overall
African/African-American	20,744
Amish	456
Latino/Admixed American	7647
Ashkenazi Jewish	1736
East Asian	2604
European (Finnish)	5316
Middle Eastern	158
European (non-Finnish)	34,029
South Asian	2419
Other	1047
Total	**76,156**

TOPMed

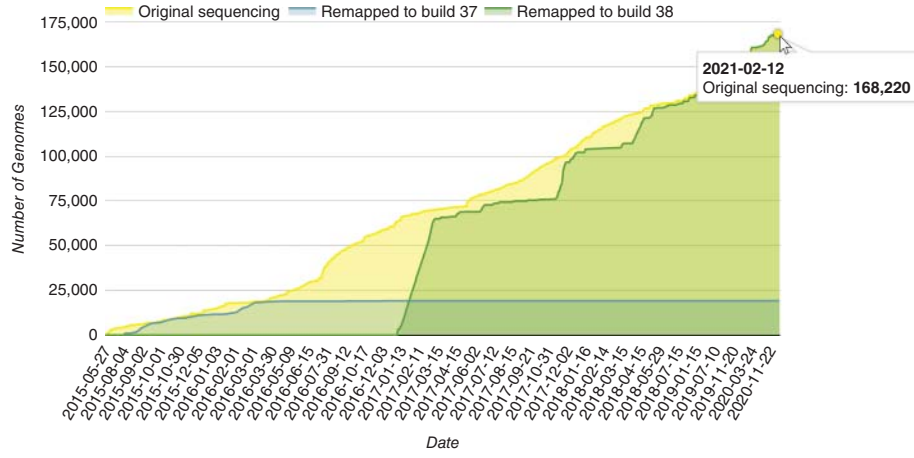


Figure 15 **Large-scale population whole genomes completed in gnomAD and TOPMed.** Data extracted on March 8, 2021.

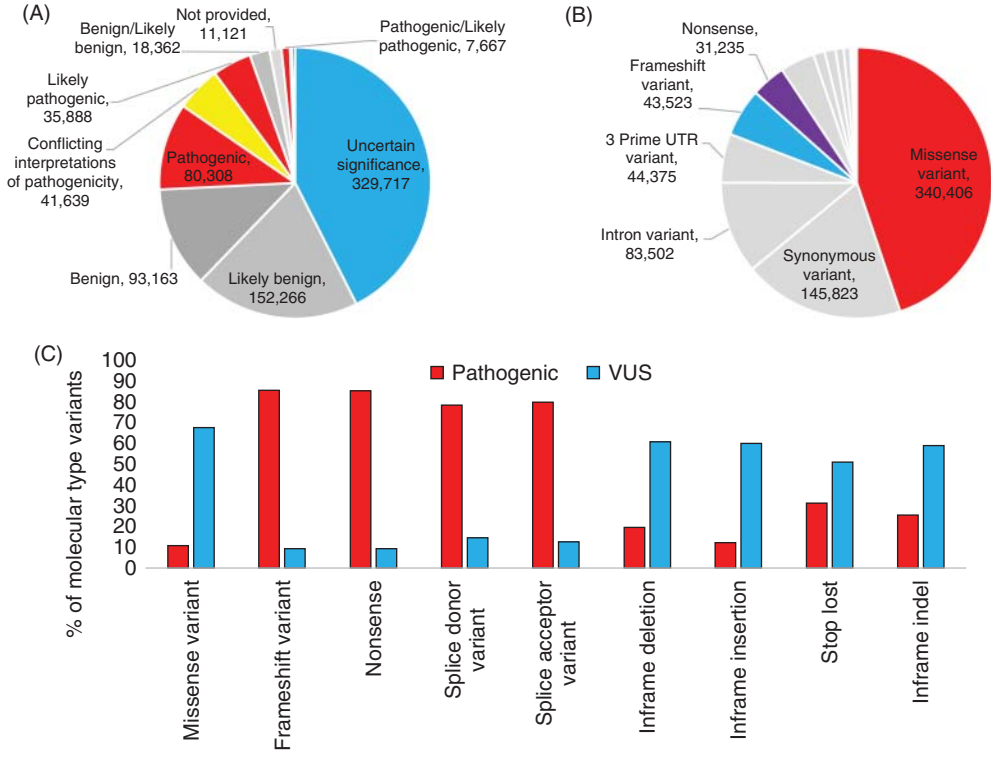


Figure 16 **ClinVar variants** (A) Clinical annotation of 774,966 ClinVar variants. (B) Molecular type of 774,966 ClinVar variants. (C) Percent of each molecular type that falls into pathogenic (red) or VUS (cyan) annotations. Data were extracted from the UCSC Genome Browser on October 14, 2020.

positions. Example = NM_152486.4(SAMD11):c.255A>G (p.Arg85=).

Intron (11.0%): A change in the DNA for regions spliced out of protein sequences. Example = NM_152486.4 (SAMD11):c.1565-3C>T, where the -3 represents the number of bases into the intron from the splice site at 1565 of the coding transcript.

3 prime UTR (5.8%): Found in a spliced gene sequence following the stop codon. Example = NM_198576.4(AGRN): c.*19C>T, where * represents the stop codon.

Frameshift (5.7%): A small indel (addition or removal of base/bases) disrupts the frame of codons. Example = NM_152486.3(SAMD11): c.1005dup (p.Ala336LysfsTer24), where the change results in a new amino acid (Lys instead of

Ala) followed by a new set of amino acid sequence from the frameshift (fs) that goes so many amino acids before a stop codon (Ter).

Nonsense (4.1%): A change of a base that results in a stop codon's insertion in the gene. Example = NM_152486.4 (SAMD11):c.1888C>T (p.Arg630Ter).

5 prime UTR (1.4%): Found in a spliced gene sequence before start codon. Example = NM_080605.4(B3GALT6):c.-38G>A, where the number is the bases before the start codon.

Splice donor (1.3%): A DNA change that might impact splicing due to its location near exon/intron boundaries. Example = NM_152486.3(SAMD11):c.869+1G>A.

Splice acceptor (1.0%): A DNA change that might impact splicing due to its location near exon/intron boundaries. Example = NM_002074.5(GNB1):c.700-1G>T.

Inframe deletion (0.8%): The deletion of three DNA bases that result in removing one or more amino acids. Example = NM_003036.4(SKI):c.280_291del (p.Ser94_Ser97del).

Inframe insertion (0.3%): The insertion of three DNA bases removes one or more amino acids. Example = NM_080605.4 (B3GALT6):c.22_36dup (p.Trp8_Ala12dup).

Several variant classes are more biased to pathogenic vs. VUS determination (Figure 16C). For example, if a variant is a frameshift, nonsense, or splice site and is within ClinVar, it is more likely to be pathogenic. Of the missense variants, 67.9% are annotated as VUS, while 10.8% are pathogenic. This suggests that there is currently a lag in the mechanistic characterization of missense variants, more so than any other molecular outcome group.

Not all genes are evenly divided in the observance of variants in ClinVar. Most genes within ClinVar have 1-9 variants listed, with a few genes having greater than 1000 listed variants (Figure 17A). Genes with high pathogenic or VUS levels are either large genes (such as TTN) or commonly studied genes that are focused on in targeted sequencing (Figure 17B).

When a variant result in loss-of-function (LoF) for a protein, this will often exert dosage issues. If approximately 50% of the gene function manifests disease phenotype, the variant is often dominant. Dominant-negative variants are those that not only decrease a protein's function but exert additional change on normal proteins, for example, multimer proteins where a variant at the dimer interface results in loss of interaction even within normal protein. If a gene requires both copies to be altered to result in disease, we annotate these as recessive variants. Some variants in the genome result in gain-of-function (GoF) changes, such as a missense variant in a degradation motif that results in the accumulation of protein. Most nonsense and frameshift mutations have probability of being GoF or dominant-negative if the resulting shortened protein were made. However, cells have a mechanism to prevent these proteins from being made, known as nonsense-mediated decay (NMD) (66). NMD results from the persistence of proteins on the extended 3′UTR that are not removed due to early release of the ribosome subunits by the nonsense variant, which initiates degradation of the entire mRNA (Figure 18). The new stop codon's location determines the likelihood for NMD to occur, such that the further towards the 5′ end of the mRNA the more likely NMD occurs. Thus, it is possible that in one gene, the location of NMD can result in one variant (further 5′) to undergo NMD and result in dosage changes of the protein and a second stop codon (further 3′) to make a shortened protein that drives dominant-negative or GoF outcomes, where the two variants have different phenotypes (71).

Clinical whole-genome sequencing

In clinical genomics for rare diseases, the gold standard has become trio-based sequencing (Figure 19). A proband's genome is first filtered against the large population genomics datasets such as gnomAD and TOPMed to remove common variants seen in many individuals without the patient phenotype. Then mom and dad's genomes are used in combination with known inheritance for the pathology to search for candidate variants. The list of candidate variants is

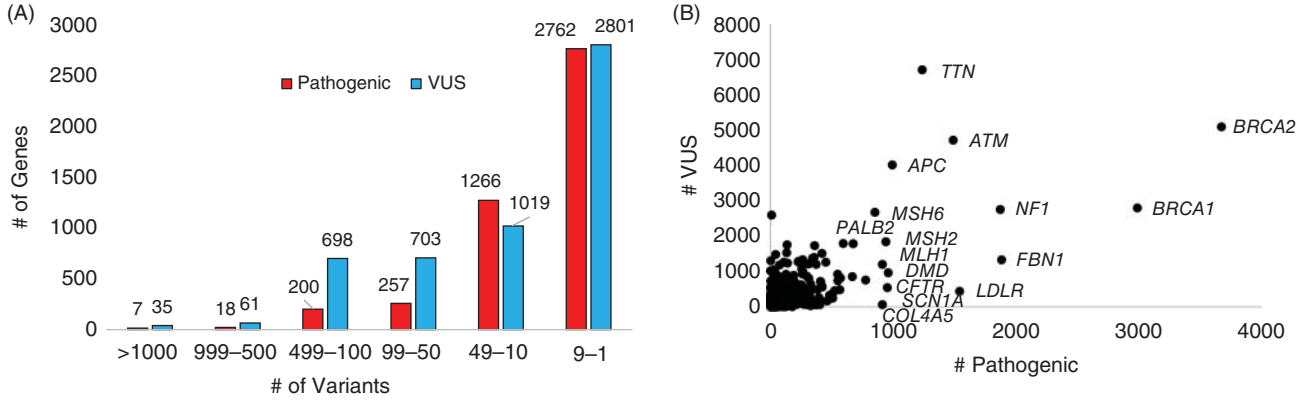


Figure 17 **ClinVar genes.** (A) Plot of the number of genes within the grouped pathogenic (red) or VUS (cyan) variants. (B) Scatter plot of the number of pathogenic (*x*-axis) vs. VUS (*y*-axis) with the top genes labeled.

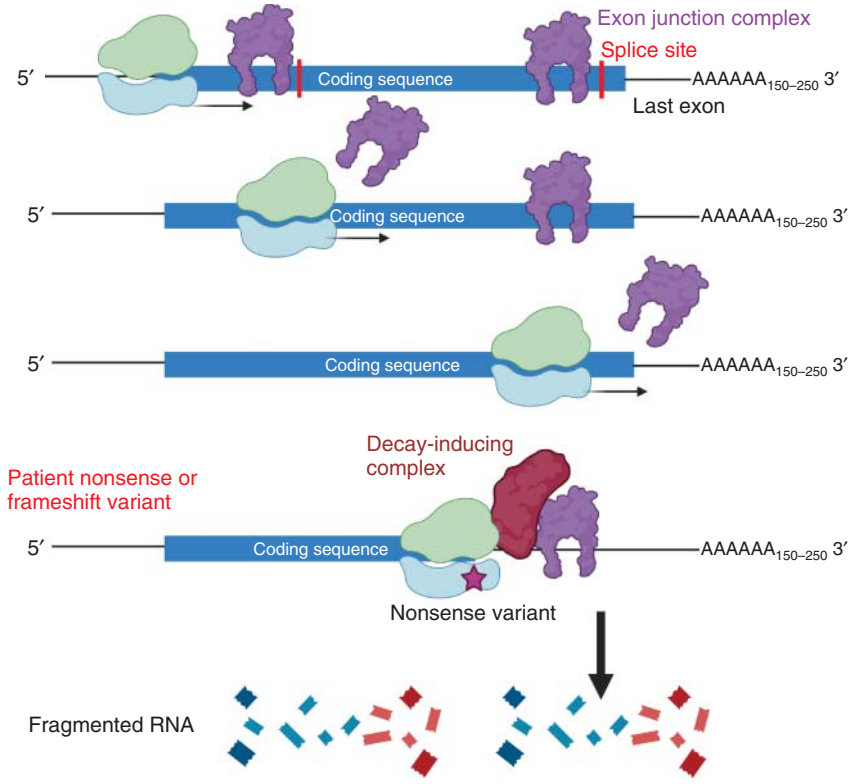


Figure 18 **Nonsense-mediated decay (NMD)**. As it reads mRNA, the ribosome (cyan/green) can remove RNA interacting proteins (magenta) that form exon junction complexes (EJC) near splice sites. When a nonsense variant arises, the 3′UTR is enlarged, with the accumulation of decay-inducing complex proteins (red) that initiate the mRNA's degradation to prevent partial protein production. Created with BioRender.

filtered through computational algorithms, publications, and genomic databases to address the variants' similarity in other patients and the predicted functional consequences of the variants. With the information, often a clinical analyst follows guidelines provided by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology to rank variants based on priority and insights (133). A certified pathologist then signs off on a report that annotates the top variant, annotation, condition/disease, and inheritance pattern. The annotations include:

Pathogenic: Variant must be conclusively connected to the condition. These often include nonsense and frameshift variants that without a doubt cause LoF in a gene where LoF is known for the condition.

Likely pathogenic: Variant that falls in a gene well connected to the condition, meets inheritance structure of the condition, and *in silico* analysis of the variant leave no room to doubt it will disrupt the protein.

Uncertain significance: Variants of uncertain significance (VUS) most often are those where a variant is found in a gene well connected to the condition, but the variant's role is still uncertain. In some occurrences, VUS can fall within a gene not connected to the condition before, but the variant's outcome meets inheritance and is predicted damaging, known as a Gene of Uncertain Disease Significance (GUDS). VUS and GUDS are not reported to families as they are still uncertain. The guidelines suggest follow-up experiments or waiting until another patient with a similar variant-condition combination is observed. As shown in Figure 16, most VUS are missense variants.

Likely benign: A variant that falls at a site that has never been connected to pathology or has no support for impacting the protein.

Benign: Without a doubt, the variant is not causal for pathology. These are often variants where population frequency is too high for it to drive a rare condition.

The application of whole-genome/exome sequencing in a large population of individuals with disease allows for detecting patients with variants in shared genes. This has enabled an accumulation of risk genes in data filtering and prioritization of variants. The most well-studied gene list is known as the ACMG59, a list of 59 medically actionable genes (76) connected to several phenotypes/physiology, including cancer and circulatory systems (Figure 20). When analyzing a genome, often findings can be made that do not directly relate to the condition being assessed, such as risk variants in the ACMG59 list that might put one at risk for a future medical condition such as cancer. These are known

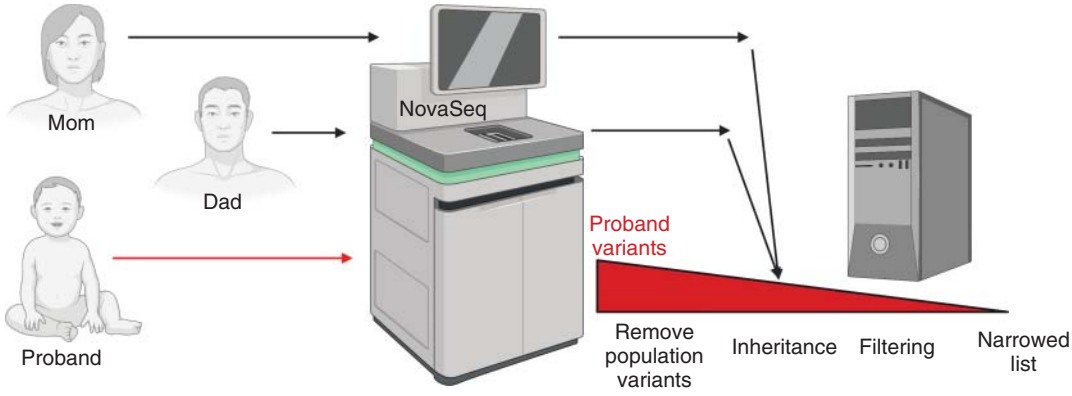


Figure 19 **Trio genomes: from variants to narrowed list.** Created with BioRender.

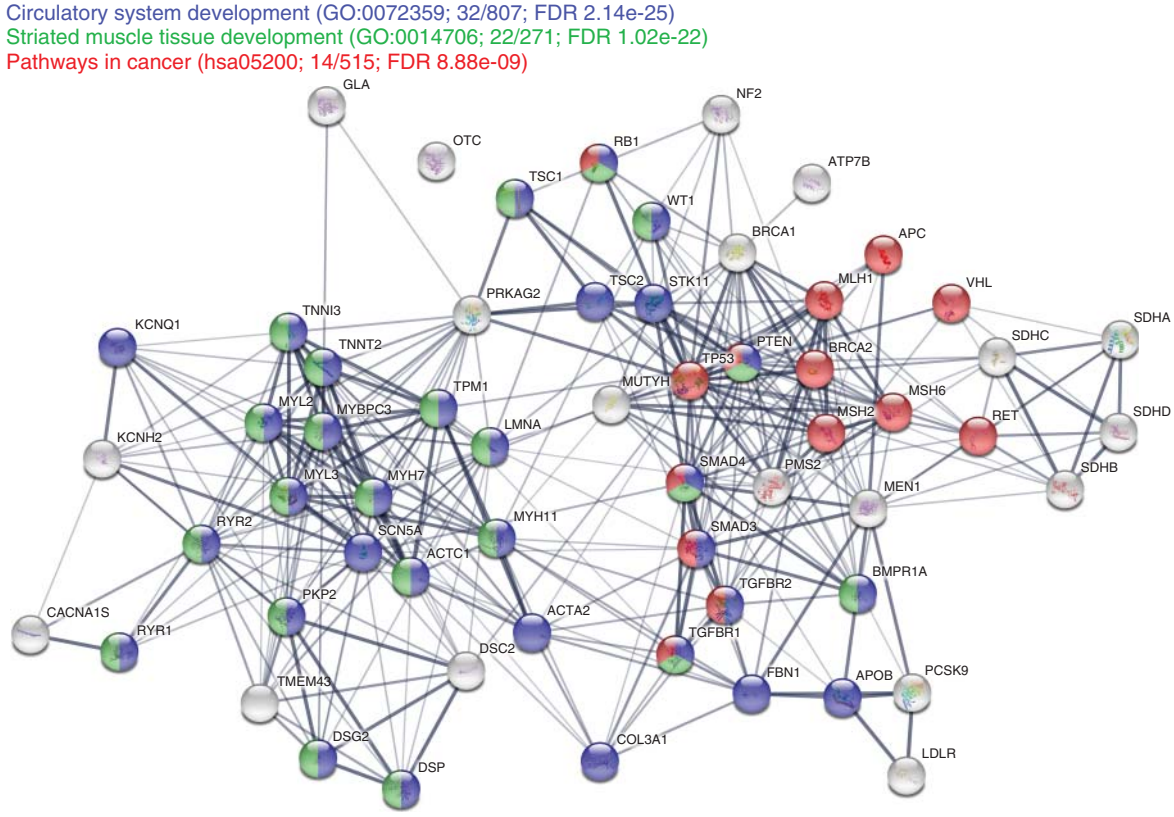


Figure 20 **ACMG59 STRING plot.** The 59 genes were processed with the STRING tools (https://string-db.org/) for known associations and enrichments in the genes. Enriched pathways are labeled at the top, and colors of gene nodes correspond to terms.

as secondary findings, and genetic counselors and clinicians are continually challenged on when and how to report these variants to families.

Clustering patients to genes

Historically, discovering a new genetic syndrome often required chance meeting or overlap of two or more patients with the same gene suggested. For example, Ellis and van Creveld's joint identification of the syndrome that bears their name took place after an incidental meeting on a train riding to a medical conference in 1939 (13). In the early pregenome era, these chance occurrences were often based more on phenotype/physiology matching for traits, conditions, or phenotypes that impact thousands of individuals, such as autism (138). Larger cohorts of sequencing databases allowed for the association of genes observed with variants in multiple individuals with these conditions. The advent of

the internet, social media, and instantaneous and straightforward worldwide connection now allow for more interaction and collaboration than ever before, particularly for phenotypes that are unique. This has directly translated to the improved matching of unique phenotypes, genes, and variants by clinicians, researchers, and even patients and families. This means that many physiology experiments are now actually performed by matching individuals across the world instead of needing animal models to confirm the causal genes. Tools such as GeneMatcher (146), Matchmaker Exchange (116), and MyGene2 allow profiles to be created and the sharing of genotype-phenotype insights between interested parties. Data sharing in these venues now spans thousands of contributors and cases across numerous countries. Other larger programs have recently emerged, which aim to drive genomic and physiological discoveries through the sharing of omics data, phenotypic details, and samples between collaborating institutions. The Genomics Research and Innovation Network, founded in 2019, combines the expertise at Boston Children's Hospital, Cincinnati Children's, and Children's Hospital of Philadelphia and standardizes data sharing between the institutions (98). This resource is needed to increase clinical sequencing volume, which translates to lower numbers of uncertain results, novel variants, and genetic variants in genes not currently associated with the disease. In many cases, the connection and discovery of overlap for VUS and GUDS in two or more individuals can elevate variants into likely pathogenic annotations and themselves provide critical physiological insights of humans.

From common to rare

It is possible to integrate both common and rare variant analyses so that one of the analysis strategies nominates a gene followed by variant screening for additional causal variants. This is particularly useful for finding variants in the 1% to 0.01% allele frequency range that lack the GWAS strategies (*p*-value rarely elevates to significance) and are too common for rare variant statistical analysis. For example, our group built on the GWAS analysis of *SHROOM3*, where we used the GWAS mechanisms that suggested *SHROOM3* as a causal gene for CKD (Figure 13) to screen missense variants below 1% allele frequency that are present in gnomAD (Figure 21). Computational tools as discussed below were used to score every missense variant within gnomAD for SHROOM3, followed by assessing population structure for inheritance and location of variants in the protein. The top variants were taken back to the CKD GWAS study groups to confirm high odds ratios (OR) where *p*-values do not reach significance due to low allele frequency, such as the P1244L variant (168). Using both animal models and molecular biochemistry, we showed that this variant functionally alters 14-3-3 protein interactions and changing kidney podocyte function (124, 168).

Tools for Association to Mechanisms

Population genomics and rare variant statistical analysis narrow down causal variants. Yet, the further narrowing of a single variant in LD or the causal role of rare variants such as VUS requires a different set of tools. Labs studying variants often start with known data, advancing the data using bioinformatics to generate hypotheses for how a variant drives the molecular outcome. These hypotheses can then be studied at the bench using biochemistry, molecular biology, cell culture, and ultimately animal or human models. Below we provide examples of tools commonly used for characterizing variants within each of these groups for both protein-coding and noncoding variants.

Known data

Gene/Protein level insights: Normally, the first step in understanding the potential outcomes of a variant is to connect the variant to a gene and establish the gene's role in biology and disease. The Online Mendelian Inheritance in Man (OMIM) database is the authoritative tool for gene-to-phenotype knowledge, including common inheritance and references for the historical insights on relationships (7). As mentioned above, ClinGen (129) and ClinVar (90) have become core resources in concatenated variants to gene insights. For simplified ClinVar analysis, the Broad Institute developed Simple ClinVar, a more user-friendly search to figure tool (114). The National Center for Biotechnology Information's (NCBI) Gene browser is a core of any analyst for extracting sequences, finding literature, and basic knowledge for any

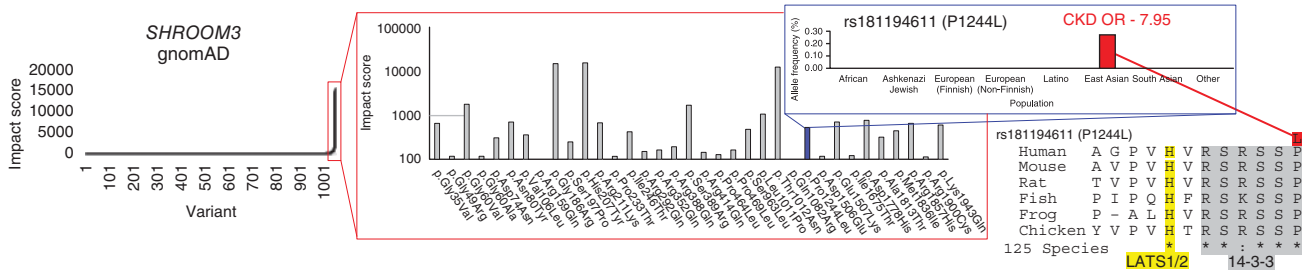


Figure 21 **Narrowing rare missense variants in SHROOM3 with pathogenic outcomes in Chronic Kidney Disease.** Modified, with permission, from Prokop JW, et al., 2018 (124).

gene in humans (23). UniProt represents the authoritative knowledgebase at the protein level, containing domain, motif, variant, sequence, and structure annotations for all human proteins (156).

Expression of gene: To build insight into the cells and tissues impacted by genetic variants, it becomes essential to know where in the human a gene is expressed. As discussed early, the GTEx database provides a robust expression of human genes in diverse individuals for 54 different tissues (60). The Human Protein Atlas (HPA) is the gold standard for assessing human gene and protein combined expression. The database consists of RNA-Seq for 37 tissues and is matched to 15,320 protein immunohistochemistry stains in each tissue (154). The combination allows for insights into gene expression and tracks the protein's histology, narrowing down cell types and cell location (nuclear, cytosol, membrane). The database has grown to include single-cell expression datasets and expression annotations for various cancer pathologies. For larger expression insights in more diverse samples, databases like FANTOM contain greater than 1000 samples of gene expressions (3). Scaling to a single-cell level, tools such as PanglaoDB provide millions of cells over 1368 datasets curated into enriched cell annotations for each gene of the genome (54). As many rare diseases are associated with neurological phenotypes, there is a high utility of additionally integrating the tools of the Allen Brain Atlas (74).

Protein interactions and network: Many protein–protein interactions are known in the literature. Tools such as BioGRID bring these interactions into searchable tools (31). Additional insight on gene interactions such as co-occurrence in publications is curated within the STRING database (52), allowing for the production of publication networks of proteins of interest (Figure 20). From the interaction network of a protein, the ontology terms annotated to each gene can be assessed for enrichment, a process known as Gene Ontology (GO) enrichment (103). For genes with little known, such as GUDS, GO enrichment of the protein network can provide strong insight into protein biology and association to overlap with phenotypes.

Epigenetics: When a variant is predicted to be noncoding based on gene regulation, the variant analysis focuses on understanding or generating epigenetics datasets. Thousands of datasets for any position in the genome are available from the UCSC Human Genome Browser (107). There are two common human genome releases used within these tools, hg19 and hg38. While hg38 is newer, it should be noted that many more datasets are found using hg19 coordinates. For extracting datasets built into or accessible within the genome browser, the table browser tools are more amenable to extracting data for a position within the genome (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1053662027_kl33ckiOAgtRf8J2QAAQQrk4oHj8). The NIH Roadmap Epigenomics Mapping Consortium provided integrative analysis (136) of datasets for positions in the chromosomes, allowing for the development of chromatic state annotations that can be easily assessed for a variant location (http://epigenomegateway.wustl.edu/legacy/?genome=hg19&datahub=https://egg2.wustl.edu/web_portal_cache/233165663.json). GeneHancer links many aspects of ENCODE datasets into readily available information for a variant region (50). More than a hundred chromatin interaction datasets (Hi-C, 4C, CIA-PET, HiChIP, PLAC-seq, Capture Hi-C) are easily visible in the Yue Lab online functional tools (http://3dgenome.fsm.northwestern.edu/index.html). Integrated insights of gene regulation are available from RegulomeDB (21). Suppose there is a desire to obtain overlapping data for a list of variants in the genome. In that case, tools like the Ensembl VEP and SNPnexus provide options for finding overlapping data of epigenetics and exporting them (102, 109).

Knockout and model animals: Multiple resources exist for animal phenotyping based on gene knockouts. The mouse genome database (MGD) and the rat genome database (RGD) contain many tools and resources for understanding genes in the individual species (24, 153). The GeneNetwork tools allow for assessing genes in diverse species, including the recombinant inbred and heterogenous stock model systems (105). The International Mouse Phenotyping Consortium (IMPC) aims to knock out every gene in the mouse genome followed by multiple phenotyping, building a database of genotype-to-phenotype information that overlaps with rare human diseases (41). As of March 10, 2021, there are 7970 phenotyped knockout mice in the database (https://www.mousephenotype.org/).

Paralog mapping: In some cases, genes get through all these tools and reveal very little insights. If this occurs, our group typically looks for paralogs of the gene of interest, those genes that share sequence similarity due to gene duplication in evolution. With the paralog genes, one can utilize all of the known data to generate hypotheses of function. In the past, these strategies worked well to help understand the *LIMD2* gene and its role in regulating cancer metastasis (113) or the integration of cancer genes *ASXL1/2* (112) into *ASXL3* biology in neural development and autism spectrum disorder (86). Moreover, the integration of knowledge for an entire gene family allows one to use variants in one gene to characterize paralogs in other genes, as previously done for the HMG-containing gene family (122).

Bioinformatics

Evolution and conservation

Information and knowledge can further be processed into usable insights for a variant location. One of the most valuable insights for a variant comes from evolutionary analysis and conservation. Tools such as PolyPhen2, Provean, and ConSurf (9) have already established protein alignments to determine a variant site's conservation. However, new genomes continue to be released, and there is additional value in the use of codon-level information (121). Rapid extraction

of protein or mRNA sequence can be obtained from NCBI Orthologs. Our group utilizes the mRNA sequence, which is then processed with Transdecoder to get open reading frames (62), aligned with ClustalW codon (91), and cleaned to remove sequences with ambiguity or missing exons. A manual step in alignment cleaning is always preferred to remove sequences with oddities and make sure alignments do not contain erroneous information. Codon selection metrics are calculated from the alignments, in addition to conservation. The calculation is placed on a sliding window to calculate the additive conservation of motifs and domains, a powerful approach in discovering regulation sites within proteins.

We highlight these additional conservation scores' power with an example from *ABCC8,* a transmembrane gene with many variants connected to familial hyperinsulinemic hypoglycemia and neonatal diabetes (11). Following alignment, a total of 221 species were assessed for ABCC8 open reading frames (Figure 22A). Conservation of each amino acid is quite challenging to visualize the regions conserved (Figure 22B), but with a 21-codon sliding window visibility of conserved motifs and domains becomes more understandable (Figure 22C). While the protein has many conserved transmembrane regions (Figure 22D), the protein's ATP binding sites are highly conserved. Detailing the conserved amino acids in the two ATP binding pockets shows that multiple familial hyperinsulinemic hypoglycemia variants cluster to conserved amino acids (Figure 22E), showing how motifs become enriched for disease variants.

While a gene like *ABCC8* is well known for functional sites, variants falling into conserved motifs often inform the research team on possible clinical/physiological variant analysis experiments. For example, the hypothesis of SHROOM3 P1244L role in 14-3-3 interaction (Figure 21) came from motif conservation analysis. The conservation of a motif can then be processed with tools such as the Eukaryotic Linear Motif tool (ELM, http://elm.eu.org/) to identify the potential role of short sequences of amino acids (43). In one case, the detection of the motif alone was able to help reclassify variants for a group of patients. A group of patients was identified with similar phenotypes with mutations in the *MED13* gene linked through social media platforms, where several patients had variants around a conserved motif. This motif was predicted to be a critical degradation box for the protein regulated by interaction with the SCF complex (111). A further literature review of the SCF complex revealed a paper that already showed a biochemical analysis of the amino acid sites (38), giving definitive biochemical support for variant outcomes. This was a lesson that review of literature for genes can be very tough to find the right papers that address a site and the impact on physiology when only searching a gene name and not the details of the complex interactions.

Protein domains and structures

Some variants fall within domains of proteins and contribute to complex structural changes. To assess these changes, one must start by identifying structures that can be used. Human structures for each protein are listed in the UniProt database. A BLAST against the Protein Data Bank (PDB) can be performed to find proteins including orthologs and paralogs. The PDB is the repository for all solved protein structures (15). In some cases, proteins can also have solved interaction partners ranging from chemicals, DNA, other proteins, or even large complexes as solved by cryo-EM techniques. These structures can be downloaded and used to generate insights into a variant position.

In some cases, multiple protein targets are available, or the targets have regions of the protein unresolved in the structure determination, such as dynamic loops. Therefore, to speed up analyses, many groups such as ours utilize modeling tools to screen through the structures, clean them, and merging them into a single model. We utilize the YASARA set of tools (85) for merging models and the integration of many other structural tools. These tools also allow the protein to be placed in physiological environments to relax crystal packing forces or mimic complex environments such as lipid membranes. With a protein structure, one can generate a high-resolution image of a variant or video for a qualitative assessment of its importance.

As with most science, qualitative analysis is not amenable to higher-throughput screening. Therefore, many groups turn to a set of tools for molecular dynamic simulations (mds), where atoms can move using biophysical approximation algorithms and tracked over a period of time, often in the nanosecond timescale. Different mathematical calculations can be used, referred to as force fields, including those of AMBER (160), CHARMM (22), and GROMACS (157). With atomic trajectories, any amino acid can be calculated for how it has moved throughout the mds. Using wild-type and mutant protein, a quantitative analysis can be performed on the variants' impact on protein movement. However, mds tools require a large amount of computing and take time to generate results. Thus, they are not a strong reactive tool for variant analysis when there is a clinical sensitivity to time for diagnosis. To bypass this limitation, mds can build a network insight of all amino acids and how they interact and move throughout time and space. Using a dynamics-cross correlation matrix (DCCM), a protein mds can be precomputed and a matrix generated for how every amino acid correlate in the movement to all other amino acids. This matrix becomes an asset in screening variants, allowing for quantitative insights into each amino acid of the protein.

To give an example of these tools, we highlight the ABCC8 protein. We begin with modeling ABCC8 from known structures (Figure 23A), embedding the protein within a lipid membrane (Figure 23B), followed by adding water and ions to physiological conditions (Figure 23C). The full simulation space was then run for extended >60 nanoseconds of molecular dynamics simulations that yielded equilibrium of movements throughout the simulation (Figure 23D), including for the secondary structure annotations (Figure 23E). We can calculate each amino acids average movement from the

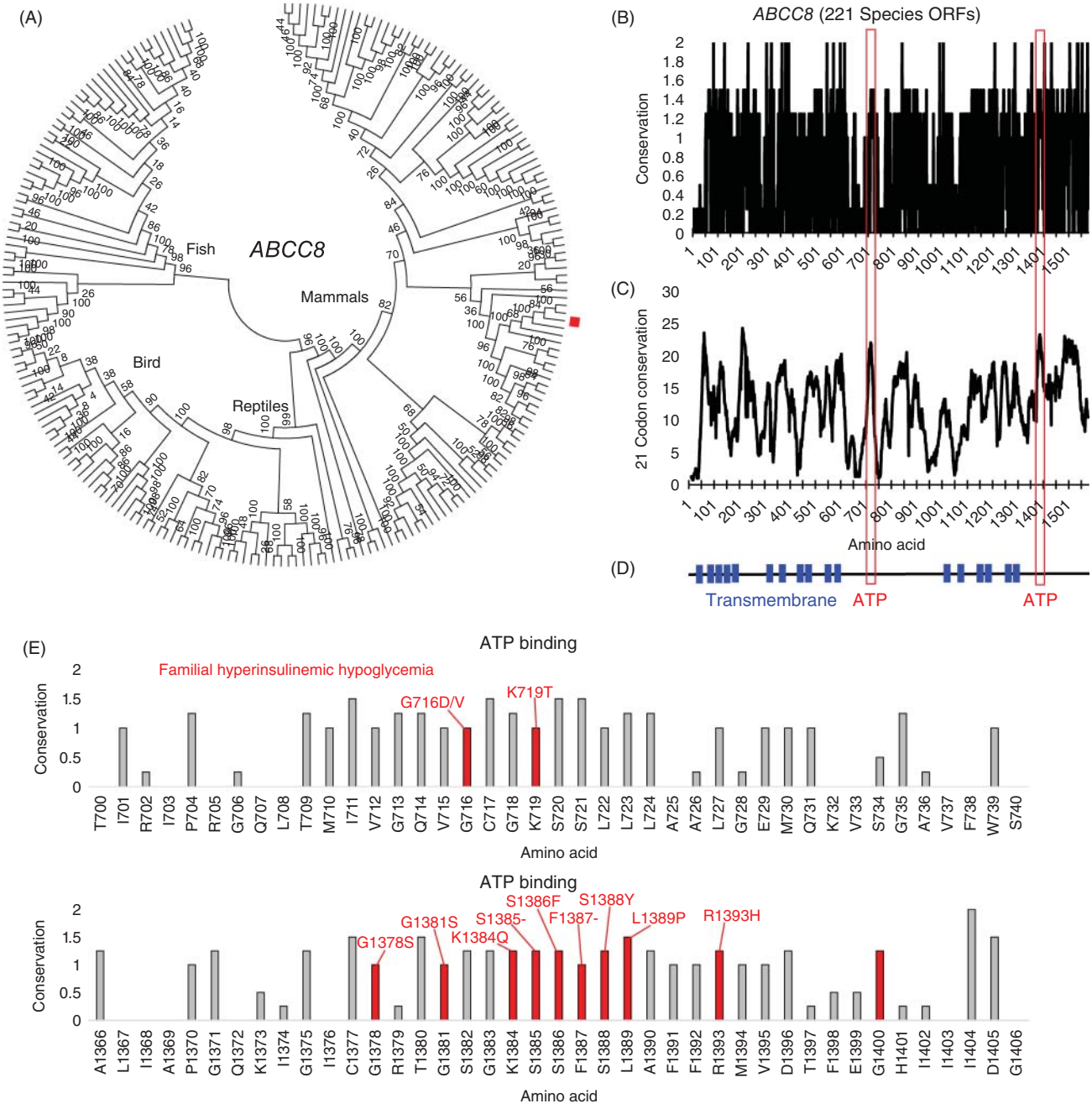


Figure 22 **ABCC8 evolution.** (A) Phylogenic tree of 221 sequences for ABCC8. Human is the red box. Values at the nodes represent the clustering of 50 bootstrapped trees. (B) Conservation score for each codon. (C) Conservation score on a 21-codon sliding window, where each site is added to the scores of 10 before and 10 codons after to smooth out enriched motifs/domains. (D) Pictoral of the transmembrane domains and the ATP binding sites. (E) Conservation of the two ATP binding site amino acids. Amino acids in red are those annotated in UniProt for association with familial hyperinsulinemic hypoglycemia.

trajectory data, giving quantitative assessments of folding space, with those sites having lower RMSF values to be well folded and those with higher values existing within disordered regions of the protein (Figure 23F). Tracking the movement of all amino acids, it is possible to calculate the correlation of movement for all amino acids relative to DCCM (Figure 23G). This DCCM can connect one amino acid to other sites within a protein structure through quantitative metrics (Figure 23H). With this structural dynamics data, combined with our evolutionary data (Figure 22), variants pulled from ClinVar, TOPMed, and gnomAD (Figure 23I), assessed with multiple variant prediction tools (Figure 23J), we can generate a complex impact score (Figure 23K) for each variant, accounting for domain and motif functionality for any variant within ABCC8. Several of the pathogenic variants and VUS from ABCC8 are present in diverse

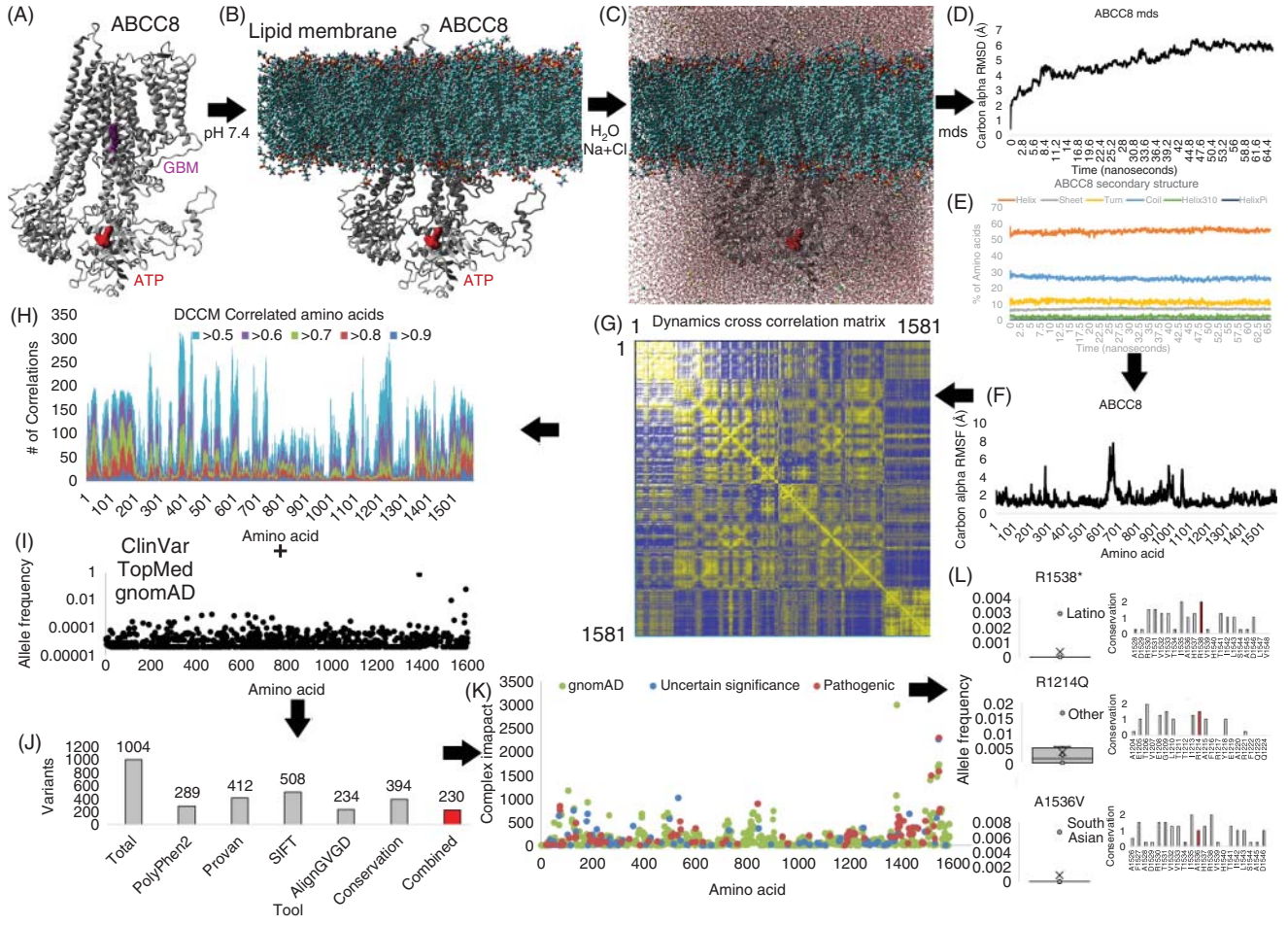


Figure 23 Screening variants for ABCC8 using protein model and variant analysis workflow.

ethnicities and within highly conserved motifs (Figure 23L), generating hypotheses for testing with lab-based tools.

Integrating conservation with structural biology is of high utility for screening variants for an entire protein or to create tools for rapid analysis of new variants (121). For example, the integration of data for the *CFTR* gene into a knowledgebase was able to identify variants in diverse ethnicities that contribute to cystic fibrosis (137). This knowledgebase and the precomputed matrix for each amino acid allow for *de novo* variant assessments of evolution and structure, connecting the new variant to other pathogenic variants through DCCM insights.

Posttranscriptional modifications

Each variant can be screened for potential posttranslational modification (ptm) alterations. Known ptm sites for genes are annotated in Uniprot (8) and HPRD (115). Predictions for modification can be generated using tools such as ProSite (73), NetPhos 2.0 (18), NetPhosK 1.0 (19), Phos3D (44), NetNGlyc 1.0 (19), UbPred (126), and SUMOsp (164). In the case of SHROOM3 P1244L, the variant falls flanking a highly predicted phosphorylation site that is critical for 14-3-3 interaction. This screening can provide valuable experiential hypotheses for variant outcomes on changing protein modifications and interactions.

Splicing

A variant's impact on splicing can be calculated using tools such as the Human Splicing Finder (40) or MutPred Splice (104). While the prediction of loss of splicing role can be easy to predict, the resulting exon splicing change is rather complex and hard to predict (148). Therefore, further laboratory analysis is often needed for the outcomes of the splicing change to determine if splicing results in nonsense or frameshift changes.

Nonsense-mediated decay

A nonsense or frameshift mutation that results in early truncation of the protein can result in RNA being broken down by NMD (Figure 18). This process can occur through either 3′ UTR exon junction complex (EJC)-dependent or independent mechanisms, but EJC-dependent accounts for most NMD (88). Databases such as SNP2NMD (63) and

NMD Classifier (72) can be used to guess if a change results in NMD. However, these predictions are not well-validated and usually require additional laboratory analysis, such as allelic bias expression analysis.

Transcription factor binding

Noncoding variants can result in disease by impacting transcription factor (TF) binding. Amongst the many tools for predicting if a variant alters TF is SNP2TFBS (87), giving metrics for loss or gain of TF binding due to a change. The tools also allow for screening large lists of variants, including all association regions' LD variants. The TF and DNA binding can be modeled as with protein modeling above, followed by mds of the typical DNA sequence and changes to the DNA sequence that mimic the variant. In the case of *SHROOM3*, rs17319721, and TCF7L2 binding (Figure 13), we were able to show changes in TF interaction with the DNA element using mds of the change (124). However, the difficulty in doing these analyses is that TF binding motifs are still an active area of research, and recruitment by additional flanking TFs contributes significantly to *in vivo* TF binding at motifs (111). An example of this is the degenerate Ebox binding motifs used by the TWIST protein, where dimers of dimers interact with larger motifs that can have genetic variants predicted to impact DNA binding. However, these are still preferred due to the higher order of multiple TF recruitments (30).

Molecular biology and biochemistry

With hypotheses for variant outcomes generated by known data or bioinformatics, it is possible to identify laboratory tools that can be used to test the hypothesis.

Patient RNA analysis

One of the growing tools in doing this is RNA abundance analyses done either in a targeted approach such as real-time quantitative PCR, targeted PCR sequencing, or using global transcriptomics of RNA-Seq. By sequencing RNA for readily obtainable material from patients, such as blood, if a gene of interest is expressed, nonsense and frameshift variants can be studied for NMD regulation. Also, sequencing the RNA can inform on splicing variant outcomes and the resulting change that is very hard to predict computationally. Trio-based genomics is ideal to do such a task, giving insight into variants on the two alleles to separate them in RNA reads, yielding a phased insight onto the gene level variants. Then RNA reads can be assessed for the variants to determine if the two alleles have even expression. Our work on a patient with an *RNASEH2B* splicing variant with hemophagocytic lymphohistiocytosis (HLH) showed that the allele inherited from the patient's mother was actively being suppressed when the patient was healthy (123). In that case, we were actively attempting to use the patient RNA-Seq to determine the outcome of the splicing variant inherited from the mother, observing that in 3/4 of blood samples collected on the patient that there were no changes in splicing. Using other variants in the gene, we showed that the RNA was only found in the blood for the father's inherited allele. At one time point, at the height of a viral infection, the mom's allele was observed. The splicing variant was determined to result in a frameshift variant that was likely inhibited in the cells through NMD making it challenging to observe the variant. This serves as a constant reminder that just because splicing, frameshift, or nonsense variant cannot be observed in a sample, does not mean it does not have functional outcomes. In the *RNASEH2B* case, compared to other RNA-Seq samples showed that the *RNASEH2B* gene was about 50% lower in the patient, further supporting NMD. In the cases of dosage-sensitive genes, that NMD process can still result in dominant diseases. In the cases of recessive diseases, the dosage levels are not likely to result in disease unless something is perturbed, such as dominant-negative partial proteins being made, as will be discussed later in the environmental genetics section.

In the case of missense or noncoding variants, RNA-Seq or targeted sequencing can be fruitful to determine the patient's dyshomeostasis. The RNA-Seq can tell overlapping genes to pathways, using GO enrichment, altered in a patient. This serves as a biomarker assessment tool for the phenotype that can confirm molecular level physiology.

Biomarker assays

Like RNA, many biomarkers can be measured in individual samples to determine the disease's molecular etiology. Mass spectrometry and/or HPLC can be used to determine metabolites (metabolomics), lipids (lipidomics), or small molecules altered in an individual. Proteomics can be used to determine if protein changes result from a variant. For many biochemical pathways altered in diseases, these biomarker assays serve as a critical tool to confirm etiology, which allows for connecting the genes into the known pathways. An example of these tools can be seen in the recent determination of the *ODC1*-related gain-of-function disorder. Patient samples show an alteration of the polyamine pathway determined by targeted metabolite analyses (25, 143). These techniques can be critically used in treatment as done with a patient with an ODC variant, where the metabolites were normalized with the drug eflornithine (127).

Protein-protein interactions

If a missense variant is predicted to fall at a protein-protein interface and likely impacts the interaction, then the logical follow-up experiment is an interaction assay. Large repositories of already cloned plasmids exist from places such as DNASU or Addgene, which can be used for site-directed mutagenesis followed by producing the protein in bacteria (or other system) with a tag (such as 6xHis), purifying it (FPLC with affinity columns), and conducting binding

assays. Obtaining these plasmids and performing mutations can take weeks. When characterization is needed quickly, an alternative is to custom synthesized plasmids. Companies such as ATUM (formerly DNA2.0) can generate wild type and mutation in less than a week, codon-optimized for the species desired, and adding the exact tagging system for purification. Interaction assays can be performed by either affinity capture or by immunoprecipitation (IP). In affinity capture, recombinantly purifying protein of interest with or without a variant is linked to a magnetic bead or column, baiting additional proteins from cellular/tissue lysates, and determining with western blots or mass spec the interaction partners. In IP experiments, the protein of interest with or without variant is expressed in a mammalian cell or tissue followed by capturing the protein and its native interactions with antibody-coated beads followed by similar detection platforms. In both cases, these can lend themselves to determining if a variant alters a known (western blot) or unknown (mass spec) interaction partner. An example of this use of interaction assays can be seen by our groups strategy for rapid characterization of NAA10 altered interaction with NAA15 in Ogden syndrome (6).

In some cases, variants fall into linear motifs that interact with proteins. In these cases, peptide synthesis works well. The wild type and variant peptides of the motif can be produced with biotin (or other) tags to allow other proteins' affinity capture. For SHROOM3 P1244L interaction with 14-3-3 (Figure 21), we used this strategy (124). In some cases, these changes can be taken into structure determination work to show advanced biophysical insights.

Protein–DNA interactions

In the case of a variant on a protein or DNA that is predicted to impact protein(TF)-DNA interactions, several binding strategies can be used. DNA oligos can be produced with major or minor alleles and labeled with either biotin or a fluorescence probe. The probes can be mixed with recombinant TFs or with nuclear protein extracts from cell lines or tissues followed by affinity capture, electrophoresis mobility shift assays (EMSA), isothermal titration calorimetry (ITC), or fluorescence anisotropy/polarization (FP). We have shown how it is possible to synthesize a dozen probes from LD SNPs of GWAS and rapidly screen alterations of TF-DNA binding from nuclear extracts of various cell types specified in the kidney (124). In the case of missense variants in TFs, the same strategies can be used but focusing on the wild type to variant TF impacts.

Protein enzymatic assays

If the protein with mutations is an enzyme or at a PTM site, functional biochemical assays can be used. Enzymes can be produced as wild type or variant and assessed for changes in a functional enzyme assay. If a variant falls on or near a PTM and is predicted to impact the modification, such as SHROOM3 P1244L (Figure 21), the protein or peptides can be generated and mixed into an enzyme assay. For phosphorylation, an ATP assay can now replace the need for phosphorus radioisotope work. If a variant is found in the enzyme, the recombinant enzyme can be produced followed by enzymatic assay. An example of this can be found in the characterization of ODC G84R, a 1% minor allele frequency variant that is associated with neurodevelopmental disorders (119).

Cell culture

Moving from the macromolecular alteration of variants into physiology requires using human or animal model systems. The quickest of these systems is to use human cells culture. Cells can be isolated from patients or controls and grown in the lab, known as primary cultures. Cells are often from accessible material, including blood isolated PBMCs (peripheral blood mononuclear cell) or skin isolated fibroblasts. These primary cultures are limited in the number of cell doublings that can occur in the lab. To get around these issues, cells can be immortalized using gene or viral processes or by identifying natural cells that can proliferate indefinitely. These cell lines can then be grown for years in the lab. Standard cell lines include EBV transformed blood cells, cancer cells from patients, TERT1 induced immortalization, embryonic stem cells, or gene transformed pluripotent stem cells. With these lines, many functional assays can be performed for characterizing variants.

Immunofluorescence: Cells allow for tracking how a variant influences cellular localization using immunofluorescence. These include variants that might impact nuclear, mitochondrial, membrane, or other cellular locations.

Overexpression: Genes cloned into mammalian expression vectors can be transfected, electroporated, or delivered using packaged lenti particles into cells. The plasmids can contain a broad array of tagging systems and mutations for studies. If an antibiotic selection marker is added to the plasmid, the successful delivery to the cells can be selected with antibiotic inclusion into the media. If the cells are left on antibiotics for an extended time (>1 month), cells that have stabilized the plasmid into the genome can be selected, known as a stable cell line. The overexpression of a wild-type allele into a cell line or tissue with the mutation and a measurable phenotype can be used as a rescue assay of the phenotype.

Knockdown: To study the outcomes of disrupting a gene, especially to understand the role of dosage sensitivity to cellular outcomes, gene knockdown systems can be used, known as RNA interference (RNAi). Small 21 to 23 nucleotides can be delivered, resulting in double-stranded RNA that is degraded (169). A short hairpin RNA (shRNA) can be cloned into mammalian plasmids that can be selected and stabilized into a cell line. Following knockdown, cell assays can determine the extent of gene knockdown and outcome on cell biology. For speed of knockdown, we often utilize the Sigma prepackaged lenti MISSION shRNA system.

Following knockdown and validation of an altered cell assay, additional delivery of an overexpression plasmid (or mRNA) can test if the wild type or variant protein can rescue the cell assay changes.

CRISPR-Cas9 mutations: Direct genome modifications are possible using CRISPR/Cas9 (12). Delivery of the Cas9 enzyme with gRNA targeted to a gene results in cutting the DNA until nonhomologous end joining occurs with a PAM site disruption. Cells can be selected through single-cell expansion for variants that result in frameshift changes that disrupt the gene. Delivery of an additional donor sequence containing a variant with homologous overlapping regions near the gRNA PAM site, allow homologous recombination, and can generate a cell with a variant. In the case of LD block regions, it is also possible to use two gRNAs and remove the entire LD block to determine if a region influences any genes' expression. Unlike with animal experiments where off-target editing can be removed with selected back crossing, cell line gene editing is more susceptible to off-target and need to be screened more carefully. An example of using both complete LD block removal and single SNP replacement with CRISPR/Cas9 can be seen for the noncoding variant regulation of SHROOM3 for CKD (124).

CRISPRi: For gene regulation variants, CRISPR mutations can take a long time and are not high throughput. Therefore, CRISPR strategies have been developed where the Cas9 is nonfunctional and contains gene inhibitor function, being driven to a target site by specific gRNAs, a technique known as CRISPR interference (CRISPRi) (92). This allows for rapid screening of sites of noncoding regulation such as GWAS and eQTLs, by suppressing the region and observing what genes change and how they change through various epigenetic assays.

Human-induced pluripotent stem cells (HiPSCs): A rapidly growing platform for understanding patient variants is iPSCs (166). Patient fibroblasts or PBMCs can be converted into iPSCs and differentiated into diverse tissue or cell types of the body. This is one of the few ways that complex cell types of a patient can be derived with the variant of interest to study cellular biology changes.

Animal models

There are many examples of animal models for characterizing genomic variants of common or rare diseases (75, 82). Mice and rats are commonly used for modeling human genes. As mentioned earlier, the IMPC within the mouse community has shown that many gene knockouts can mimic human disease (41). Cardiovascular diseases from congestive heart failure to myocardial infarction have been modeled, where the rat has been particularly useful due to the increased size of organs (51, 65). Zebrafish has been a robust model of early developmental genetics, mainly because of the fish's transparency, allowing for easy observation of phenotypic changes of internal organs (93). Animal models need to be used carefully, especially in the detailed phenotypic overlap and studying the role of drug pharmacokinetic/pharmacodynamic combined with disease genetics (101, 159). CRISPR, especially point mutations of patients, use in animal models has been invaluable to variant characterizations (110).

Genetics by Environment (GxE)

As our knowledge of genomics has advanced, so has our understanding of genetics' complex interactions with environmental stimuli. For most common disease genetics, the penetrance is not always high, and therefore we must always keep in mind how physiological responses interplay with genetic variants. For example, much of cardiovascular genetics are connected to how the immune system responds to damage (49). Hypoxia response in tissues and cells can be modulated by genetics (77, 78, 118), including cancer (95). One area this has become increasingly important as of late is COVID-19 response, where GWAS have suggested variants for the severity of response that do not overlap with other pathologies, suggesting that the virus and genomic elements interact (36, 64, 120, 144). Many other infections are connected to genetic changes, including cytokine responses (79).

As briefly mentioned in the section on RNA-Seq, the case of HLH and a heterozygous splicing variant that results in a frameshift to RNASEH2B, it is possible for the environment, including viral infections and hypoxia, to impact genes (123). The RNA-Seq of multiple time points for the patient suggested that the frameshift variant was cleared from the cell by NMD (Figure 24A), except in the height of EBV infection. Heterozygous mutations in RNASEH2B are not associated with any diseases. RNASEH2B is associated with recessive Aicardi-Goutières syndrome (132), a disease of altered interferon response and encephalopathy (131). Our patient with a single RNASEH2B frameshift variant was healthy for 16 years. An EBV infection brought the patient into the hospital with severe multi-organ failure, and at the height of the EBV infection, the frameshift variant appeared at high levels in the patient's blood. It turns out that EBV and many additional viruses can inhibit NMD pathways (Figure 24B), where the inhibition of NMD allows for the survival of viral RNA typically degraded by NMD (100). In the case of our RNASEH2B variant patient, at the height of EBV infection, NMD was inhibited, which results in the expression of a dominant-negative protein of RNASEH2B that inhibits the RNASEH2A/C and PCNA complex formation (Figure 24C) driving a rare disease that only manifests if NMD is altered in the cell. This represents an incredible new challenge for physiological genetics to begin understanding the interaction of environmental stimuli and dyshomeostasis onto genetic influenced changes within the cells and tissues. Going from a proteinuria loci nominated in the heterogenous stock rats to a novel cell culture system, it has been shown how kidney tubule cell response is altered when variants exist in hypoxia response genes (78). In neural development, the influence

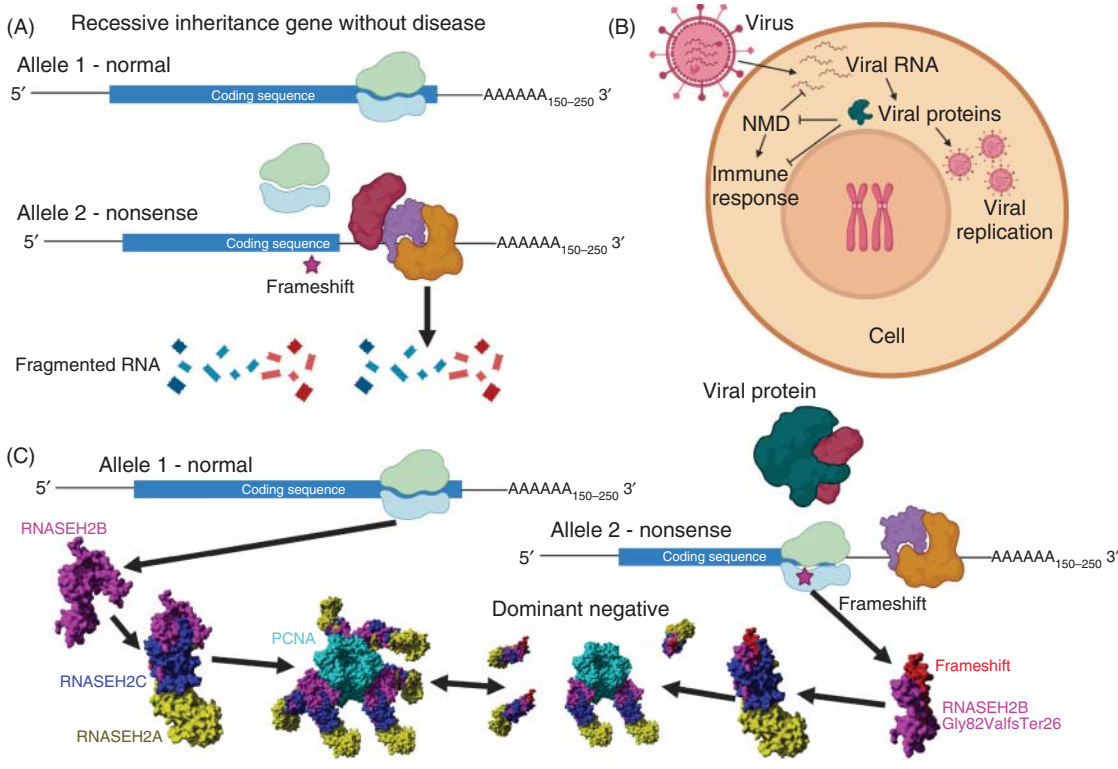


Figure 24 **Viral Induced Genetics and RNASEH2B.** (A) NMD of RNASEH2B alleles. (B) Pictorial of viral inhibition of NMD. (C) Viral activation of dominant-negative genetics of patriation RNASEH2B protein inhibiting the protein complex. It was created with BioRender.

of chemicals, such as diesel particulate matter, has been shown to modulate the expression of critical genes involved in ASD and DD/ID (17), where iPSCs were combined with advanced sequencing of single-cell RNA-seq and direct sequencing. Direct RNA-seq allows the transcriptome-wide detection of base-pair changes that occur in RNA molecules (55), a technique with a high potential to discover many new physiological genomic mechanisms in the future.

Variants in the Classroom

The characterization of genomic variants in the clinical setting, particularly variants identified as VUS, is imperative for patient care. Clinical geneticists and genetic counselors have taken on a large role in this effort. Multiple education resources aimed at genetics professionals explain how VUSs are classified, describe tools that predict pathogenicity, and seek to inform clinicians when VUSs have been reclassified (10). Building variant characterizations into future scientists' training is expedient as variant insights have become decisive for medical professionals.

Informational videos are available to explain the clinical interpretation of sequence variants from sources such as The Broad's Medical and Population Genetics Primer series of lectures (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/primers/primer-medical-and-pop). At the general population level, patients with genetic sequencing finding of a VUS can participate in research studies that educate and seek to reclassify the variant, such as the Family Variant Classification study, centered at the University of Washington. The current project builds on the FindMyVariant Research Study, which uses strategies to improve variant classification probability using familial segregation (152). The study aims to educate individuals about their VUS results and learn more about their unique variants through family data and DNA sequencing of relatives. The ultimate goal is to collect enough information to reclassify the VUS.

While the tools often used to predict variant impact are robust, such as SIFT, PolyPhen2, Provean, and AlignGVGD, they do not provide data on how variants can alter a protein's structure and potentially its cellular physiology. Improving the knowledgebase of academic researchers in computational variant analysis techniques will develop a powerful resource for evidence production surrounding a large number of VUS currently identified, while simultaneously exposing future professionals to the value and challenges of genetic mechanisms. The HudsonAlpha Institute for Biotechnology has developed an initial effort to move this type of research to the undergraduate level. The Characterizing Our DNA Exceptions (CODE) program (https://hudsonalphacode.org/) seeks to expand opportunities for authentic research by students at non-research-intensive colleges and universities,

while advancing student knowledge of genetic variants. Much of the CODE program is aimed at the undergraduate level with independent research projects. Faculty facilitators are trained in a VUS characterization workflow using open-source databases/tools and the YASARA molecular graphics, modeling, and simulation program. Student researchers examine clinical variants from the HudsonAlpha genomic sequencing projects to help explain complex VUS such as MED13 (145) and RALA (69). This work also suggests an incredible potential to bring this work into the training of future physiologists, providing lower-cost experimental hypothesis design in physiology departments without budgets for expensive variant characterizations.

The steps toward a research project that identifies a variant and explores how it might affect the structure and function of the encoded protein describe an immersive experience, where genetics is taught in the context of the specific disease/phenotype of interest. The first step is to choose a disease or disorder. A student drives the selection based on something that interests them, driving independent project design that does not have to align with a research mentors' interest. After selecting a disease of interest, students begin an exhaustive search and compilation of known data from databases such as gnomAD, TOPMed, ClinVar, and UniProt to determine what genes are associated with the disease, to select a gene for study, and gather data on the variants associated with the gene or a disease. Students synergize this information into slides and figures to discuss with their research mentors.

Students then begin characterizing variants by building an *in silico* model of the protein of interest, analyzing evolutionary conservation and molecular dynamic simulations. This includes the development of educational materials such as videos, 3D models, and publications that can be used by clinical staff to explain VUS. Students compare VUS to all known variants by collecting, analyzing, and interpreting data from other sources and ClinVar datasets. Molecular dynamic simulations provide additional information about the variant impact on a protein's movement in a computationally derived cellular environment by analyzing the movement trajectories. We provide students with a standardized macro for simulating proteins consisting of only two lines of code to initiate a simulation and analysis. These platforms are flexible to PC, Mac, or Linux environments based on schools computing resources. With the information gleaned from their investigations, students can develop a hypothesis about variant impacts, allowing them to share their findings with the scientific community through presentations and publications. With the required resources, students can expand their projects to functional assays or build collaborations with other scientists.

Independent research projects often lose some efficiency. Many students are challenged for time outside of their classes and extracurricular activities. Mentors often reteach the same content multiple times if they have several students. To expand variant characterizations into a classroom setting to account for these issues, schools such as the University of North Alabama (UNA) have made extensive efforts within the CODE program. The UNA CODE class is offered for upper-division mathematics credit and counts toward a major or minor in mathematics, but it often includes student cohorts of diverse majors. This creates a classroom environment where the students bring diverse integrative science backgrounds to variant characterizations, where chemistry, biology, statistics, medicine, and computer science backgrounds combine to tackle physiological problems. These courses are carefully designed to be employable by undergraduate students at all levels, to prepare students to participate in research through a series of skill-building activities, gathering known data and performing simulations, culminating in one or more research presentations at regional science or undergraduate research conferences.

Consequently, the course results in a decisive improvement in written and oral communication skills. Students receive instruction on literature searches, project design, implementation, data analysis, and scientific writing/presentation. Students are immersed in scientific research and incrementally learn skills and techniques as they are needed in this approach.

One group where variant characterizations fit well is medical (MD/DO) students. Students, especially those interested in medical genetics, can work with a clinician and a researcher to serve as the primary analyst for a case study. The MD student works with the lead clinician to go over the clinical phenotype/physiology, medical records, and disease overlaps for a patient with a newly reported clinical VUS. Then they process the variant similar to undergraduate students above. This positions the student to be the first author of a case study, such as those case studies authored by MD students for NAA10 (6), HSD17B4 (139), and IL11RA (84). With the next generation of professionals, clinicians, and researchers' experiences in variant characterizations, we continue advancing genomics knowledge and how it integrates with physiology.

Conclusions

The number of variants requiring further research to establish physiological mechanisms continues to grow. The majority of genetic associations from GWAS or PheWAS are at an association level only, with few examples of mechanisms for variants that cause the phenotype. Rare variants and their role in rare diseases have many more examples of mechanistic roles on phenotype, but the number of genomes being sequenced increases. For variants in 1% to 0.01% of individuals, we lack statistical tools to find causal variants to biology changes. The rate-limiting step for figuring out the intervention of genotype-to-phenotype often requires mechanisms. Therefore, the determination of these mechanisms using the techniques and strategies laid out is critically important. Additionally, having newly trained professionals that are focused on variant characterizations, or even know

the terminology of genetics so they can communicate with collaborators, has growing importance.

As our tools and strategies improve for rare and common variants, many areas will need continued growth. We need to start focusing on intermediate variant roles in physiology, developing new tools for statistics and characterization. Genome sequencing also needs to be expanded to more diverse individuals, including those from countries (such as Africa) where genomic diversity is high. For rare variants, the majority of the focus is still on protein-coding changes. As most common variants impact noncoding gene regulation, it is highly probable that rare noncoding variants also have a significant physiological role, notably for common disorders. While we lack statistical modeling for these rare variants that might impact gene expression, using the common variant GWAS and PheWAS along with larger epigenomics insights to narrow functional gene regulatory regions, these sites can then be screened for rare variant influences. This may one day open the door to characterizing these rare variants impact in physiology, moving to a patient/individual physiological genomics insight. While we have much growth in genomics, the ongoing work in computational and experimental variants' mechanisms is an exciting time for physiological genomics.

Related Articles

Genome–Wide Maps of Transcription Regulatory Elements and Transcription Enhancers in Development and Disease
Long Noncoding RNA: Genomics and Relevance to Physiology
Clinical and Molecular Genetic Features of Hereditary Pulmonary Arterial Hypertension
Genomics and Proteomics of Pulmonary Vascular Disease
Genetic Models of Diabetes Insipidus

Acknowledgments

The authors and much of the presented work from their labs have been funded by the National Institutes of Health Big Data to Knowledge program (K01ES025435 to JWP), Michigan State University, Alabama Power, HudsonAlpha Institute for Biotechnology, and Spectrum Health.

References

1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1092 human genomes. *Nature* 491: 56-65, 2012. DOI: 10.1038/nature11632.
2. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 526: 68-74, 2015. DOI: 10.1038/nature15393.
3. Abugessaisa I, Noguchi S, Carninci P, Kasukawa T. The FANTOM5 computation ecosystem: Genomic information hub for promoters and active enhancers. *Methods Mol Biol* 1611: 199-217, 2017. DOI: 10.1007/978-1-4939-7015-5_15.
4. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol* 17: 241, 2016. DOI: 10.1186/s13059-016-1110-1.
5. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 7: 248-249, 2010. DOI: 10.1038/nmeth0410-248.
6. Afrin A, Prokop JW, Underwood A, Uhl KL, VanSickle EA, Baruwal R, Wajda M, Rajasekaran S, Bupp C. NAA10 variant in 38-week-gestation male patient: A case study. *Cold Spring Harb Mol Case Stud* 6, 2020. DOI: 10.1101/mcs.a005868.
7. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 43: D789-D798, 2015. DOI: 10.1093/nar/gku1205.
8. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL. UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 32: D115-D119, 2004. DOI: 10.1093/nar/gkh131.
9. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38: W529-W533, 2010. DOI: 10.1093/nar/gkq399.
10. Augusto BM, Lake P, Scherr CL, Couch FJ, Lindor NM, Vadaparampil ST. From the laboratory to the clinic: Sharing BRCA VUS reclassification tools with practicing genetics professionals. *J Community Genet* 9: 209-215, 2018. DOI: 10.1007/s12687-017-0343-3.
11. Babenko AP, Polak M, Cavé H, Busiah K, Czernichow P, Scharfmann R, Bryan J, Aguilar-Bryan L, Vaxillaire M, Froguel P. Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N Engl J Med* 355: 456-466, 2006. DOI: 10.1056/NEJMoa055068.
12. Barrangou R, Doudna JA. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* 34: 933-941, 2016. DOI: 10.1038/nbt.3659.
13. Baujat G, Le Merrer M. Ellis-van Creveld syndrome. *Orphanet J Rare Dis* 2: 27, 2007. DOI: 10.1186/1750-1172-2-27.
14. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* 8: 639-646, 2007. DOI: 10.1038/nrg2149.
15. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35: D301-D303, 2007. DOI: 10.1093/nar/gkl971.
16. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 369: 64-67, 1994. DOI: 10.1038/369064a0.
17. Bilinovich SM, Uhl KL, Lewis K, Soehnlen X, Williams M, Vogt D, Prokop JW, Campbell DB. Integrated RNA sequencing reveals epigenetic impacts of diesel particulate matter exposure in human cerebral organoids. *Dev Neurosci* 42: 195-207, 2020. DOI: 10.1159/000513536.
18. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294: 1351-1362, 1999. DOI: 10.1006/jmbi.1999.3310.
19. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4: 1633-1649, 2004. DOI: 10.1002/pmic.200300771.
20. Boehme AK, Esenwa C, Elkind MSV. Stroke risk factors, genetics, and prevention. *Circ Res* 120: 472-495, 2017. DOI: 10.1161/CIRCRESAHA.116.308398.
21. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790-1797, 2012. DOI: 10.1101/gr.137323.112.
22. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. *J Comput Chem* 30: 1545-1614, 2009. DOI: 10.1002/jcc.21287.
23. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res* 43: D36-D42, 2015. DOI: 10.1093/nar/gku1055.
24. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, Mouse Genome Database Group. The Mouse Genome Database (MGD): Mouse biology and model systems. *Nucleic Acids Res* 36: D724-D728, 2008. DOI: 10.1093/nar/gkm961.

25. Bupp CP, Schultz CR, Uhl KL, Rajasekaran S, Bachmann AS. Novel de novo pathogenic variant in the ODC1 gene in a girl with developmental delay, alopecia, and dysmorphic features. *Am J Med Genet A* 176: 2548-2553, 2018. DOI: 10.1002/ajmg.a.40523.
26. Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, Mortier G, Speleman F, Menten B. Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. *Eur J Med Genet* 52: 398-403, 2009. DOI: 10.1016/j.ejmg.2009.09.002.
27. Calore M, De Windt LJ, Rampazzo A. Genetics meets epigenetics: Genetic variants that modulate noncoding RNA in cardiovascular diseases. *J Mol Cell Cardiol* 89: 27-34, 2015. DOI: 10.1016/j.yjmcc.2015.10.028.
28. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 39: S16-S21, 2007. DOI: 10.1038/ng2028.
29. Castoldi E, Simioni P, Kalafatis M, Lunghi B, Tormene D, Girelli D, Girolami A, Bernardi F. Combinations of 4 mutations (FV R506Q, FV H1299R, FV Y1702C, PT 20210G/A) affecting the prothrombinase complex in a thrombophilic family. *Blood* 96: 1443-1448, 2000.
30. Chang AT, Liu Y, Ayyanathan K, Benner C, Jiang Y, Prokop JW, Paz H, Wang D, Li H-R, Fu X-D, Rauscher FJ, Yang J. An evolutionarily conserved DNA architecture determines target specificity of the TWIST family bHLH transcription factors. *Genes Dev* 29: 603-616, 2015. DOI: 10.1101/gad.242842.114.
31. Chatr-Aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41: D816-D823, 2013. DOI: 10.1093/nar/gks1158.
32. Choi Y, Chan AP. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31: 2745-2747, 2015. DOI: 10.1093/bioinformatics/btv195.
33. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11: 415-425, 2010. DOI: 10.1038/nrg2779.
34. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, Ellsworth K, Camp B, Kint CI, Yacoubian C, Farnaes L, Bainbridge MN, Beebe C, Braun JJA, Bray M, Carroll J, Cakici JA, Caylor SA, Clarke C, Creed MP, Friedman J, Frith A, Gain R, Gaughran M, George S, Gilmer S, Gleeson J, Gore J, Grunenwald H, Hovey RL, Janes ML, Lin K, McDonagh PD, McBride K, Mulrooney P, Nahas S, Oh D, Oriol A, Puckett L, Rady Z, Reese MG, Ryu J, Salz L, Sanford E, Stewart L, Sweeney N, Tokita M, Van Der Kraan L, White S, Wigby K, Williams B, Wong T, Wright MS, Yamada C, Schols P, Reynders J, Hall K, Dimmock D, Veeraraghavan N, Defay T, Kingsmore SF. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med* 11, 2019. DOI: 10.1126/scitranslmed.aat6177.
35. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261: 921-923, 1993. DOI: 10.1126/science.8346443.
36. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet* 28: 715-718, 2020. DOI: 10.1038/s41431-020-0636-6.
37. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. *Nat Genet* 48: 1284-1287, 2016. DOI: 10.1038/ng.3656.
38. Davis MA, Larimore EA, Fissel BM, Swanger J, Taatjes DJ, Clurman BE. The SCF-Fbw7 ubiquitin ligase degrades MED13 and MED13L and regulates CDK8 module association with mediator. *Genes Dev* 27: 151-156, 2013. DOI: 10.1101/gad.207720.112.
39. de Knijff P, van den Maagdenberg AM, Frants RR, Havekes LM. Genetic heterogeneity of apolipoprotein E and its influence on plasma lipid and lipoprotein levels. *Hum Mutat* 4: 178-194, 1994. DOI: 10.1002/humu.1380040303.
40. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37: e67, 2009. DOI: 10.1093/nar/gkp215.
41. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H, Baker CN, Bower L, Brown JM, Caddle LB, Chiani F, Clary D, Cleak J, Daly MJ, Denegre JM, Doe B, Dolan ME, Edie SM, Fuchs H, Gailus-Durner V, Galli A, Gambadoro A, Gallegos J, Guo S, Horner NR, Hsu C-W, Johnson SJ, Kalaga S, Keith LC, Lanoue L, Lawson TN, Lek M, Mark M, Marschall S, Mason J, McElwee ML, Newbigging S, Nutter LMJ, Peterson KA, Ramirez-Solis R, Rowland DJ, Ryder E, Samocha KE, Seavitt JR, Selloum M, Szoke-Kovacs Z, Tamura M, Trainor AG, Tudose I, Wakana S, Warren J, Wendling O, West DB, Wong L, Yoshiki A, International Mouse Phenotyping Consortium, Jackson Laboratory, Infrastructure Nationale PHENOMIN, Institut Clinique de la Souris (ICS), Charles River Laboratories, MRC Harwell, Toronto Centre for Phenogenomics, Wellcome Trust Sanger Institute, RIKEN BioResource Center, MacArthur DG, Tocchini-Valentini GP, Gao X, Flicek P, Bradley A, Skarnes WC, Justice MJ, Parkinson HE, Moore M, Wells S, Braun RE, Svenson KL, de Angelis MH, Herault Y, Mohun T, Mallon A-M, Henkelman RM, Brown SDM, Adams DJ, Lloyd KCK, McKerlie C, Beaudet AL, Bućan M, Murray SA. High-throughput discovery of novel developmental phenotypes. *Nature* 537: 508-514, 2016. DOI: 10.1038/nature19356.
42. Diederichs S. The four dimensions of noncoding RNA conservation. *Trends Genet* 30: 121-123, 2014. DOI: 10.1016/j.tig.2014.01.004.
43. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jödicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K, Via A, Chatr-Aryamontri A, Haslam N, Grebnev G, Edwards RJ, Steinmetz MO, Meiselbach H, Diella F, Gibson TJ. ELM--the database of eukaryotic linear motifs. *Nucleic Acids Res* 40: D242-D251, 2012. DOI: 10.1093/nar/gkr1064.
44. Durek P, Schudoma C, Weckwerth W, Selbig J, Walther D. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 10: 117, 2009. DOI: 10.1186/1471-2105-10-117.
45. ENCODE. Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74, 2012. DOI: 10.1038/nature11247.
46. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung W-K, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei C-L, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CWH, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JNS, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VVB,

Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816, 2007. DOI: 10.1038/nature05874.

47. ENCODE Project Consortium, Snyder MP, Gingeras TR, Moore JE, Weng Z, Gerstein MB, Ren B, Hardison RC, Stamatoyannopoulos JA, Graveley BR, Feingold EA, Pazin MJ, Pagan M, Gilchrist DA, Hitz BC, Cherry JM, Bernstein BE, Mendenhall EM, Zerbino DR, Frankish A, Flicek P, Myers RM. Perspectives on ENCODE. *Nature* 583: 693-698, 2020. DOI: 10.1038/s41586-020-2449-8.
48. Estivill X, Armengol L. Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 3: 1787-1799, 2007. DOI: 10.1371/journal.pgen.0030190.
49. Fernández-Ruiz I. Immune system and cardiovascular disease. *Nat Rev Cardiol* 13: 503, 2016. DOI: 10.1038/nrcardio.2016.127.
50. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, Lancet D, Cohen D. GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017, 2017. DOI: 10.1093/database/bax028.
51. Flister MJ, Prokop JW, Lazar J, Shimoyama M, Dwinell M, Geurts A. 2015 Guidelines for Establishing Genetically Modified Rat Models for Cardiovascular Research. *J Cardiovasc Transl Res* 8 (4): 269-277, 2015.
52. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41: D808-D815, 2013. DOI: 10.1093/nar/gks1094.
53. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 47: D766-D773, 2019. DOI: 10.1093/nar/gky955.
54. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019, 2019. DOI: 10.1093/database/baz046.
55. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 15: 201-206, 2018. DOI: 10.1038/nmeth.4577.
56. Garbers C, Monhasery N, Aparicio-Siegmund S, Lokau J, Baran P, Nowell MA, Jones SA, Rose-John S, Scheller J. The interleukin-6 receptor Asp358Ala single nucleotide polymorphism rs2228145 confers increased proteolytic conversion rates by ADAM proteases. *Biochim Biophys Acta* 1842: 1485-1494, 2014. DOI: 10.1016/j.bbadis.2014.05.018.
57. Gatz M, Pedersen NL, Berg S, Johansson B, Johansson K, Mortimer JA, Posner SF, Viitanen M, Winblad B, Ahlbom A. Heritability for Alzheimer's disease: The study of dementia in Swedish twins. *J Gerontol A Biol Sci Med Sci* 52: M117-M125, 1997. DOI: 10.1093/gerona/52a.2.m117.
58. Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med* 5: 181, 2018. DOI: 10.3389/fcvm.2018.00181.
59. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369: 1318-1330, 2020. DOI: 10.1126/science.aaz1776.
60. GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts, Laboratory, Data Analysis & Coordinating Center (LDACC), NIH program management, Biospecimen collection, Pathology, eQTL manuscript working group, Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression across human tissues. *Nature* 550: 204-213, 2017. DOI: 10.1038/nature24277.
61. Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am J Hum Genet* 103: 522-534, 2018. DOI: 10.1016/j.ajhg.2018.08.016.
62. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8: 1494-1512, 2013. DOI: 10.1038/nprot.2013.084.
63. Han A, Kim W-Y, Park S-M. SNP2NMD: A database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. *Bioinformatics* 23: 397-399, 2007. DOI: 10.1093/bioinformatics/btl593.
64. Hartog N, Faber W, Frisch A, Bauss J, Bupp CP, Rajasekaran S, Prokop JW. SARS-CoV-2 infection: Molecular mechanisms of severe outcomes to suggest therapeutics. *Expert Rev Proteomics* 18 (2): 105-118, 2021.
65. Hasenfuss G. Animal models of human cardiovascular disease, heart failure and hypertrophy. *Cardiovasc Res* 39: 60-76, 1998. DOI: 10.1016/s0008-6363(98)00110-2.
66. Hentze MW, Kulozik AE. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* 96: 307-310, 1999. DOI: 10.1016/s0092-8674(00)80542-5.
67. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Res* 30: 163-165, 2002. DOI: 10.1093/nar/30.1.163.
68. Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Plott CB, Jenkins J, Gray DE, Holt JM, Bowling KM, Bebin EM, Grimwood J, Schmutz J, Cooper GM. Long-read genome sequencing for the diagnosis of neurodevelopmental disorders. *HGG Adv.* 2 (2): 100023.
69. Hiatt SM, Neu MB, Ramaker RC, Hardigan AA, Prokop JW, Hancarova M, Prchalova D, Havlovicova M, Prchal J, Stranecky V, Yim DKC, Powis Z, Keren B, Nava C, Mignot C, Rio M, Revah-Politi A, Hemati P, Stong N, Iglesias AD, Suchy SF, Willaert R, Wentzensen IM, Wheeler PG, Brick L, Kozenko M, Hurst ACE, Wheless JW, Lacassie Y, Myers RM, Barsh GS, Sedlacek Z, Cooper GM. De novo mutations in the GTP/GDP-binding region of RALA, a RAS-like small GTPase, cause intellectual disability and developmental delay. *PLoS Genet* 14, 2018. DOI: 10.1371/journal.pgen.1007671.
70. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, Germain M, Mälarstig A, Brown A, Soria JM, Dichgans M, Bing N, Franco-Cereceda A, Souto JC, Dermitzakis ET, Hamsten A, Worrall BB, Tung JY, METASTROKE Consortium, INVENT Consortium, Sabater-Lleal M. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet* 25: 1867-1874, 2016. DOI: 10.1093/hmg/ddw037.
71. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. Nonsense-mediated decay approaches the clinic. *Nat Genet* 36: 801-808, 2004. DOI: 10.1038/ng1403.
72. Hsu M-K, Lin H-Y, Chen F-C. NMD Classifier: A reliable and systematic classification tool for nonsense-mediated decay events. *PLoS One* 12: e0174798, 2017. DOI: 10.1371/journal.pone.0174798.
73. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJA. The 20 years of PROSITE. *Nucleic Acids Res* 36: D245-D249, 2008. DOI: 10.1093/nar/gkm977.
74. Jones AR, Overly CC, Sunkin SM. The Allen Brain Atlas: 5 years and beyond. *Nat Rev Neurosci* 10: 821-828, 2009. DOI: 10.1038/nrn2722.
75. Jucker M. The benefits and limitations of animal models for translational research in neurodegenerative diseases. *Nat Med* 16: 1210-1214, 2010. DOI: 10.1038/nm.2224.
76. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 19: 249-255, 2017. DOI: 10.1038/gim.2016.190.
77. Keele GR, Prokop JW, He H, Holl K, Littrell J, Deal AW, Kim Y, Kyle PB, Attipoe E, Johnson AC, Uhl KL. Sept8/SEPTIN8 involvement in cellular structure and kidney damage is identified by genetic mapping and a novel human tubule hypoxic model. *Sci Rep* 11 (1): 1-15, 2021.
78. Keele GR, Prokop JW, He H, Holl K, Littrell J, Deal AW, Kim Y, Kyle PB, Attipoe E, Johnson AC, Uhl KL, Sirpilla OL, Jahanbakhsh S, Robinson M, Levy S, Valdar W, Garrett MR, Solberg Woods LC. Sept8/SEPTIN8 involvement in cellular structure and kidney damage

is identified by genetic mapping and a novel human tubule hypoxic model. *Sci Rep* 11: 2071, 2021. DOI: 10.1038/s41598-021-81550-8.
79. Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, Fine J, Krichevsky A, Delude RL, Angus DC. GenIMS investigators. Understanding the inflammatory cytokine response in pneumonia and sepsis: Results of the genetic and inflammatory markers of sepsis (GenIMS) Study. *Arch Intern Med* 167: 1655-1663, 2007. DOI: 10.1001/archinte.167.15.1655.
80. Kelly KM, Smith JA, Mezuk B. Depression and interleukin-6 signaling: A Mendelian Randomization study. *Brain Behav Immun* 95: 106-114, 2021. DOI: 10.1016/j.bbi.2021.02.019.
81. Kerin T, Ramanathan A, Rivas K, Grepo N, Coetzee GA, Campbell DB. A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci Transl Med* 4: 128ra40, 2012. DOI: 10.1126/scitranslmed.3003479.
82. Kilk K. Metabolomics for Animal Models of Rare Human Diseases: An expert review and lessons learned. *OMICS* 23: 300-307, 2019. DOI: 10.1089/omi.2019.0065.
83. Kingsmore SF, Cakici JA, Clark MM, Gaughran M, Feddock M, Batalov S, Bainbridge MN, Carroll J, Caylor SA, Clarke C, Ding Y, Ellsworth K, Farnaes L, Hildreth A, Hobbs C, James K, Kint CI, Lenberg J, Nahas S, Prince L, Reyes I, Salz L, Sanford E, Schols P, Sweeney N, Tokita M, Veeraraghavan N, Watkins K, Wigby K, Wong T, Chowdhury S, Wright MS, Dimmock D. A randomized, controlled trial of the analytic and diagnostic performance of Singleton and Trio, rapid genome and exome sequencing in ill infants. *Am J Hum Genet* 105: 719-733, 2019. DOI: 10.1016/j.ajhg.2019.08.009.
84. Korakavi N, Prokop JW, Seaver LH. Evolution of the phenotype of craniosynostosis with dental anomalies syndrome and report of IL11RA variant population frequencies in a Crouzon-like autosomal recessive syndrome. *Am J Med Genet A* 179 (4): 668-673, 2019.
85. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 77 (Suppl 9): 114-122, 2009. DOI: 10.1002/prot.22570.
86. Kuechler A, Czeschik JC, Graf E, Grasshoff U, Hüffmeier U, Busa T, Beck-Woedl S, Faivre L, Rivière J-B, Bader I, Koch J, Reis A, Hehr U, Rittinger O, Sperl W, Haack TB, Wieland T, Engels H, Prokisch H, Strom TM, Lüdecke H-J, Wieczorek D. Bainbridge-Ropers syndrome caused by loss-of-function variants in ASXL3: A recognizable condition. *Eur J Hum Genet* 25: 183-191, 2017. DOI: 10.1038/ejhg.2016.165.
87. Kumar S, Ambrosini G, Bucher P. SNP2TFBS – A database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* 45: D139-D144, 2017. DOI: 10.1093/nar/gkw1064.
88. Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* 20: 406-420, 2019. DOI: 10.1038/s41580-019-0126-2.
89. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921, 2001. DOI: 10.1038/35057062.
90. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42 (D1): D980-D985.
91. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948, 2007. DOI: 10.1093/bioinformatics/btm404.
92. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* 8: 2180-2196, 2013. DOI: 10.1038/nprot.2013.132.
93. Lieschke GJ, Currie PD. Animal models of human disease: Zebrafish swim into view. *Nat Rev Genet* 8: 353-367, 2007. DOI: 10.1038/nrg2091.
94. Lorenzi L, Avila Cobos F, Decock A, Everaert C, Helsmoortel H, Lefever S, Verboom K, Volders P-J, Speleman F, Vandesompele J, Mestdagh P. Long noncoding RNA expression profiling in cancer: Challenges and opportunities. *Genes Chromosomes Cancer* 58: 191-199, 2019. DOI: 10.1002/gcc.22709.
95. Luoto KR, Kumareswaran R, Bristow RG. Tumor hypoxia as a driving force in genetic instability. *Genome Integr* 4: 5, 2013. DOI: 10.1186/2041-9414-4-5.
96. Ma H, Yu L, Byra EA, Hu N, Kitagawa K, Nakayama KI, Kawamoto T, Ren J. Aldehyde dehydrogenase 2 knockout accentuates ethanol-induced cardiac depression: Role of protein phosphatases. *J Mol Cell Cardiol* 49: 322-329, 2010. DOI: 10.1016/j.yjmcc.2010.03.017.
97. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 45: D896-D901, 2017. DOI: 10.1093/nar/gkw1133.
98. Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH, Biswas S, Bourgeois FT, Corsmo J, Dauber A, Devkota B, Fleisher GR, Heath AP, Helbig I, Hirschhorn JN, Kilbourn J, Kong SW, Kornetsky S, Majzoub JA, Marsolo K, Martin LJ, Nix J, Schwarzhoff A, Stedman J, Strauss A, Sund KL, Taylor DM, White PS, Marsh E, Grimberg A, Hawkes C, Genomics Research and InnovationNetwork. The Genomics Research and Innovation Network: Creating an interoperable, federated, genomics learning system. *Genet Med* 22: 371-380, 2020. DOI: 10.1038/s41436-019-0646-3.
99. Matoba N, Akiyama M, Ishigaki K, Kanai M, Takahashi A, Momozawa Y, Ikegawa S, Ikeda M, Iwata N, Hirata M, Matsuda K, Murakami Y, Kubo M, Kamatani Y, Okada Y. GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nat Hum Behav* 4: 308-316, 2020. DOI: 10.1038/s41562-019-0805-1.
100. May JP, Yuan X, Sawicki E, Simon AE. RNA virus evasion of nonsense-mediated decay. *PLoS Pathog* 14: e1007459, 2018. DOI: 10.1371/journal.ppat.1007459.
101. McGonigle P, Ruggeri B. Animal models of human disease: Challenges in enabling translation. *Biochem Pharmacol* 87: 162-171, 2014. DOI: 10.1016/j.bcp.2013.08.006.
102. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. *Genome Biol* 17: 122, 2016. DOI: 10.1186/s13059-016-0974-4.
103. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 45: D183-D189, 2017. DOI: 10.1093/nar/gkw1138.
104. Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Mooney SD. MutPred Splice: Machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol* 15: R19, 2014. DOI: 10.1186/gb-2014-15-1-r19.
105. Mulligan MK, Mozhui K, Prins P, Williams RW. GeneNetwork: A toolbox for systems genetics. *Methods Mol Biol* 1488: 75-120, 2017. DOI: 10.1007/978-1-4939-6427-7_4.
106. Nakayama A, Nakatochi M, Kawamura Y, Yamamoto K, Nakaoka H, Shimizu S, Higashino T, Koyama T, Hishida A, Kuriki K, Watanabe M, Shimizu T, Ooyama K, Ooyama H, Nagase M, Hidaka Y, Matsui D, Tamura T, Nishiyama T, Shimanoe C, Katsuura-Kamano S, Takashima N, Shirai Y, Kawaguchi M, Takao M, Sugiyama R, Takada

Y, Nakamura T, Nakashima H, Tsunoda M, Danjoh I, Hozawa A, Hosomichi K, Toyoda Y, Kubota Y, Takada T, Suzuki H, Stiburkova B, Major TJ, Merriman TR, Kuriyama N, Mikami H, Takezaki T, Matsuo K, Suzuki S, Hosoya T, Kamatani Y, Kubo M, Ichida K, Wakai K, Inoue I, Okada Y, Shinomiya N, Matsuo H, Japan Gout Genomics Consortium (Japan Gout). Subtype-specific gout susceptibility loci and enrichment of selection pressure on ABCG2 and ALDH2 identified by subtype genome-wide meta-analyses of clinically defined gout patients. *Ann Rheum Dis* 79: 657-665, 2020. DOI: 10.1136/annrheumdis-2019-216644.

107. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, Lee BT, Hinrichs AS, Fyfe AC, Fernandes JD, Diekhans M, Clawson H, Casper J, Benet-Pagès A, Barber GP, Haussler D, Kuhn RM, Haeussler M, Kent WJ. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* 49: D1046-D1057, 2021. DOI: 10.1093/nar/gkaa1070.
108. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31: 3812-3814, 2003.
109. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res* 48: W185-W192, 2020. DOI: 10.1093/nar/gkaa420.
110. Papasavva P, Kleanthous M, Lederer CW. Rare Opportunities: CRISPR/Cas-based therapy development for rare genetic diseases. *Mol Diagn Ther* 23: 201-222, 2019. DOI: 10.1007/s40291-019-00392-3.
111. Partridge EC, Chhetri SB, Prokop JW, Ramaker RC, Jansen CS, Goh S-T, Mackiewicz M, Newberry KM, Brandsmeier LA, Meadows SK, Messer CL, Hardigan AA, Coppola CJ, Dean EC, Jiang S, Savic D, Mortazavi A, Wold BJ, Myers RM, Mendenhall EM. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* 583: 720-728, 2020. DOI: 10.1038/s41586-020-2023-4.
112. Peng H, Prokop J, Karar J, Park K, Cao L, Harbour JW, Bowcock AM, Malkowicz SB, Cheung M, Testa JR, Rauscher FJ. Familial and somatic BAP1 mutations inactivate ASXL1/2-mediated allosteric regulation of BAP1 deubiquitinase by targeting multiple independent domains. *Cancer Res* 78: 1200-1213, 2018. DOI: 10.1158/0008-5472.CAN-17-2876.
113. Peng H, Talebzadeh-Farrooji M, Osborne MJ, Prokop JW, McDonald PC, Karar J, Hou Z, He M, Kebebew E, Orntoft T, Herlyn M, Caton AJ, Fredericks W, Malkowicz B, Paterno CS, Carolin AS, Speicher DW, Skordalakes E, Huang Q, Dedhar S, Borden KLB, Rauscher FJ. LIMD2 is a small LIM-only protein overexpressed in metastatic lesions that regulates cell motility and tumor progression by directly binding to and activating the integrin-linked kinase. *Cancer Res* 74: 1390-1403, 2014. DOI: 10.1158/0008-5472.CAN-13-1275.
114. Pérez-Palma E, Gramm M, Nürnberg P, May P, Lal D. Simple ClinVar: An interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. *Nucleic Acids Res* 47: W99-W105, 2019. DOI: 10.1093/nar/gkz411.
115. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13: 2363-2371, 2003. DOI: 10.1101/gr.1680803.
116. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, Dumitriu S, Dyke SOM, den Dunnen JT, Firth HV, Gibbs RA, Girdea M, Gonzalez M, Haendel MA, Hamosh A, Holm IA, Huang L, Hurles ME, Hutton B, Krier JB, Misyura A, Mungall CJ, Paschall J, Paten B, Robinson PN, Schiettecatte F, Sobreira NL, Swaminathan GJ, Taschner PE, Terry SF, Washington NL, Züchner S, Boycott KM, Rehm HL. The Matchmaker Exchange: A platform for rare disease gene discovery. *Hum Mutat* 36: 915-921, 2015. DOI: 10.1002/humu.22858.
117. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29: 512-520, 2011. DOI: 10.1038/nbt.1852.
118. Prabhakar NR. Sensing hypoxia: Physiology, genetics and epigenetics. *J Physiol* 591: 2245-2257, 2013. DOI: 10.1113/jphysiol.2012.247759.
119. Prokop JW, Bupp CP, Frisch A, Bilinovich SM, Campbell DB, Vogt D, Schultz CR, Uhl KL, VanSickle E, Rajasekaran S, Bachmann AS. Emerging role of ODC1 in neurodevelopmental disorders and brain development. *Genes (Basel)* 12: 470, 2021. DOI: 10.3390/genes12040470.
120. Prokop JW, Hartog NL, Chesla D, Faber W, Love CP, Karam R, Abualkheir N, Feldmann B, Teng L, McBride T, Leimanis ML, English BK, Holsworth A, Frisch A, Bauss J, Kalpage N, Derbedrossian A, Pinti RM, Hale N, Mills J, Eby A, VanSickle EA, Pageau SC, Shankar R, Chen B, Carcillo JA, Sanfilippo D, Olivero R, Bupp CP, Rajasekaran S. High-density blood transcriptomics reveals precision immune signatures of SARS-CoV-2 infection in hospitalized individuals. *Front Immunol*, 2021. DOI: 10.3389/fimmu.2021.694243.
121. Prokop JW, Lazar J, Crapitto G, Smith DC, Worthey EA, Jacob HJ. Molecular modeling in the age of clinical genomics, the enterprise of the next generation. *J Mol Model* 23: 75, 2017. DOI: 10.1007/s00894-017-3258-3.
122. Prokop JW, Leeper TC, Duan Z-H, Milsted A. Amino acid function and docking site prediction through combining disease variants, structure alignments, sequence alignments, and molecular dynamics: A study of the HMG domain. *BMC Bioinformatics* 13, S3 (Suppl 2), 2012. DOI: 10.1186/1471-2105-13-S2-S3.
123. Prokop JW, Shankar R, Gupta R, Leimanis ML, Nedveck D, Uhl K, Chen B, Hartog NL, Van Veen J, Sisco JS, Sirpilla O, Lydic T, Boville B, Hernandez A, Braunreiter C, Kuk CC, Singh V, Mills J, Wegener M, Adams M, Rhodes M, Bachmann AS, Pan W, Byrne-Steele ML, Smith DC, Depinet M, Brown BE, Eisenhower M, Han J, Haw M, Madura C, Sanfilippo DJ, Seaver LH, Bupp C, Rajasekaran S. Virus-induced genetics revealed by multidimensional precision medicine transcriptional workflow applicable to COVID-19. *Physiol Genomics* 52: 255-268, 2020. DOI: 10.1152/physiolgenomics.00045.2020.
124. Prokop JW, Yeo NC, Ottmann C, Chhetri SB, Florus KL, Ross EJ, Sosonkina N, Link BA, Freedman BI, Coppola CJ, McDermott-Roe C, Leysen S, Milroy L-G, Meijer FA, Geurts AM, Rauscher FJ, Ramaker R, Flister MJ, Jacob HJ, Mendenhall EM, Lazar J. Characterization of coding/noncoding variants forSHROOM3in patients with CKD. *J Am Soc Nephrol* 29 (5): 1525-1535, 2018.
125. Quillen EE, Chen X-D, Almasy L, Yang F, He H, Li X, Wang X-Y, Liu T-Q, Hao W, Deng H-W, Kranzler HR, Gelernter J. ALDH2 is associated to alcohol dependence and is the major genetic determinant of "daily maximum drinks" in a GWAS study of an isolated rural Chinese sample. *Am J Med Genet B Neuropsychiatr Genet* 165B: 103-110, 2014. DOI: 10.1002/ajmg.b.32213.
126. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 78: 365-380, 2010. DOI: 10.1002/prot.22555.
127. Rajasekaran S, Bupp CP, Leimanis-Laurens M, Shukla A, Russell C, Junewick J, Gleason E, VanSickle EA, Edgerly Y, Wittmann BM, Prokop JW, Bachmann AS. Repurposing eflornithine to treat a patient with a rare ODC1 gain-of-function variant disease. *Lifestyles* 10: e67097, 2021. DOI: 10.7554/eLife.67097.
128. Ray D, Boehnke M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet Epidemiol* 42: 134-145, 2018. DOI: 10.1002/gepi.22105.
129. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS. ClinGen. ClinGen–The clinical genome resource. *N Engl J Med* 372: 2235-2242, 2015. DOI: 10.1056/NEJMsr1406261.
130. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 47: D886-D894, 2019. DOI: 10.1093/nar/gky1016.
131. Rice G, Patrick T, Parmar R, Taylor CF, Aeby A, Aicardi J, Artuch R, Montalto SA, Bacino CA, Barroso B, Baxter P, Benko WS, Bergmann C, Bertini E, Biancheri R, Blair EM, Blau N, Bonthron DT, Briggs T, Brueton LA, Brunner HG, Burke CJ, Carr IM, Carvalho DR, Chandler KE, Christen H-J, Corry PC, Cowan FM, Cox H, D'Arrigo S, Dean J, De Laet C, De Praeter C, Dery C, Ferrie CD, Flintoff K, Frints SGM, Garcia-Cazorla A, Gener B, Goizet C, Goutieres F, Green AJ, Guet A, Hamel BCJ, Hayward BE, Heiberg A, Hennekam RC, Husson M, Jackson AP, Jayatunga R, Jiang Y-H, Kant SG, Kao A, King MD, Kingston HM, Klepper J, van der Knaap MS, Kornberg AJ, Kotzot D, Kratzer W, Lacombe D, Lagae L, Landrieu PG, Lanzi G, Leitch A, Lim MJ, Livingston JH, Lourenco CM, Lyall EGH, Lynch SA, Lyons MJ, Marom D, Mcclure JP, Mcwilliam R, Melancon SB, Mewasingh LD, Moutard M-L, Nischal KK, Ostergaard JR, Prendiville J, Rasmussen M, Rogers RC, Roland D, Rosser EM, Rostasy K, Roubertie A, Sanchis A, Schiffmann R, Scholl-Burgi S, Seal S, Shalev SA, Corcoles CS, Sinha GP, Soler D, Spiegel R, Stephenson JBP, Tacke U, Tan TY, Till M, Tolmie JL, Tomlin P, Vagnarelli F, Valente EM, Van Coster RNA, Van der Aa N, Vanderver A, Vles JSH, Voit T, Wassmer E, Weschke B, Whiteford ML, Willemsen MAA, Zankl A, Zuberi SM,

Orcesi S, Fazzi E, Lebon P, Crow YJ. Clinical and molecular phenotype of Aicardi-Goutieres syndrome. *Am J Hum Genet* 81: 713-725, 2007. DOI: 10.1086/521373.
132. Rice GI, Forte GMA, Szynkiewicz M, Chase DS, Aeby A, Abdel-Hamid MS, Ackroyd S, Allcock R, Bailey KM, Balottin U, Barnerias C, Bernard G, Bodemer C, Botella MP, Cereda C, Chandler KE, Dabydeen L, Dale RC, De Laet C, De Goede CGEL, Del Toro M, Effat L, Enamorado NN, Fazzi E, Gener B, Haldre M, Lin J-P-S-M, Livingston JH, Lourenco CM, Marques W, Oades P, Peterson P, Rasmussen M, Roubertie A, Schmidt JL, Shalev SA, Simon R, Spiegel R, Swoboda KJ, Temtamy SA, Vassallo G, Vilain CN, Vogt J, Wermenbol V, Whitehouse WP, Soler D, Olivieri I, Orcesi S, Aglan MS, Zaki MS, Abdel-Salam GMH, Vanderver A, Kisand K, Rozenberg F, Lebon P, Crow YJ. Assessment of interferon-related biomarkers in Aicardi-Goutières syndrome associated with mutations in TREX1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, and ADAR: A case-control study. *Lancet Neurol* 12: 1159-1169, 2013. DOI: 10.1016/S1474-4422(13)70258-8.
133. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. 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. *Genet Med* 17: 405-424, 2015. DOI: 10.1038/gim.2015.30.
134. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK. Alzheimer's Disease Genetics Consortium. Alzheimer's disease: Analyzing the missing heritability. *PLoS One* 8: e79771, 2013. DOI: 10.1371/journal.pone.0079771.
135. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods* 11: 294-296, 2014. DOI: 10.1038/nmeth.2832.
136. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317-330, 2015. DOI: 10.1038/nature14248.
137. Sanders M, Lawlor JMJ, Li X, Schuen JN, Millard SL, Zhang X, Buck L, Grysko B, Uhl KL, Hinds D, Stenger CL, Morris M, Lamb N, Levy H, Bupp C, Prokop JW. Genomic, transcriptomic, and protein landscape profile of CFTR and cystic fibrosis. *Hum Genet* 140: 423-439, 2021. DOI: 10.1007/s00439-020-02211-w.
138. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, Peng M, Collins R, Grove J, Klei L, Stevens C, Reichert J, Mulhern MS, Artomov M, Gerges S, Sheppard B, Xu X, Bhaduri A, Norman U, Brand H, Schwartz G, Nguyen R, Guerrero EE, Dias C, Autism Sequencing Consortium, iPSYCH-Broad Consortium, Betancur C, Cook EH, Gallagher L, Gill M, Sutcliffe JS, Thurm A, Zwick ME, Børglum AD, State MW, Cicek AE, Talkowski ME, Cutler DJ, Devlin B, Sanders SJ, Roeder K, Daly MJ, Buxbaum JD. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* 180: 568-584.e23, 2020. DOI: 10.1016/j.cell.2019.12.036.
139. Savage LT, Adams SD, James KN, Chowdhury S, Rajasekaran S, Prokop JW, Bupp CP. Rapid whole-genome sequencing identifies a homozygous novel variant, His540Arg, in HSD17B4 resulting in D-bifunctional protein deficiency disorder diagnosis. *Mol Case Stud* 6 (6): a005496, 2020.
140. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19: 491-504, 2018. DOI: 10.1038/s41576-018-0016-z.
141. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM. Quality assessment of the human genome sequence. *Nature* 429: 365-368, 2004. DOI: 10.1038/nature02390.
142. Schröck E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, Garini Y, Ried T. Multicolor spectral karyotyping of human chromosomes. *Science* 273: 494-497, 1996. DOI: 10.1126/science.273.5274.494.
143. Schultz CR, Bupp CP, Rajasekaran S, Bachmann AS. Biochemical features of primary cells from a pediatric patient with a gain-of-function ODC1 genetic mutation. *Biochem J* 476: 2047-2057, 2019. DOI: 10.1042/BCJ20190294.
144. Sirpilla O, Bauss J, Gupta R, Underwood A, Qutob D, Freeland T, Bupp C, Carcillo J, Hartog N, Rajasekaran S, Prokop JW. SARS-CoV-2-encoded proteome and human genetics: From interaction-based to ribosomal biology impact on disease and risk processes. *J Proteome Res* 19: 4275-4290, 2020. DOI: 10.1021/acs.jproteome.0c00421.
145. Snijders Blok L, Hiatt SM, Bowling KM, Prokop JW, Engel KL, Cochran JN, Bebin EM, Bijlsma EK, Ruivenkamp CAL, Terhal P, Simon MEH, Smith R, Hurst JA, McLaughlin H, Person R, Crunk A, Wangler MF, Streff H, Symonds JD, Zuberi SM, Elliott KS, Sanders VR, Masunga A, Hopkin RJ, Dubbs HA, Ortiz-Gonzalez XR, Pfundt R, Brunner HG, Fisher SE, Kleefstra T, Cooper GM. De novo mutations in MED13, a component of the mediator complex, are associated with a novel neurodevelopmental disorder. *Hum Genet* 137: 375-388, 2018. DOI: 10.1007/s00439-018-1887-y.
146. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* 36: 928-930, 2015. DOI: 10.1002/humu.22844.
147. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. *Hum Mol Genet* 24: R111-R119, 2015. DOI: 10.1093/hmg/ddv260.
148. Spurdle AB, Couch FJ, Hogervorst FBL, Radice P, Sinilnikova OM, IARC Unclassified Genetic Variants Working Group. Prediction and assessment of splicing alterations: Implications for clinical testing. *Hum Mutat* 29: 1304-1313, 2008. DOI: 10.1002/humu.20901.
149. Takeuchi F, Isono M, Nabika T, Katsuya T, Sugiyama T, Yamaguchi S, Kobayashi S, Ogihara T, Yamori Y, Fujioka A, Kato N. Confirmation of ALDH2 as a Major locus of drinking behavior and of its variants regulating multiple metabolic phenotypes in a Japanese population. *Circ J* 75: 911-918, 2011. DOI: 10.1253/circj.cj-10-0774.
150. Takeuchi F, Yokota M, Yamamoto K, Nakashima E, Katsuya T, Asano H, Isono M, Nabika T, Sugiyama T, Fujioka A, Awata N, Ohnaka K, Nakatochi M, Kitajima H, Rakugi H, Nakamura J, Ohkubo T, Imai Y, Shimamoto K, Yamori Y, Yamaguchi S, Kobayashi S, Takayanagi R, Ogihara T, Kato N. Genome-wide association study of coronary artery disease in the Japanese. *Eur J Hum Genet* 20: 333-340, 2012. DOI: 10.1038/ejhg.2011.184.
151. Tang J, Yu Y, Yang W. Long noncoding RNA and its contribution to autism spectrum disorders. *CNS Neurosci Ther* 23: 645-656, 2017. DOI: 10.1111/cns.12710.
152. Tsai GJ, Rañola JMO, Smith C, Garrett LT, Bergquist T, Casadei S, Bowen DJ, Shirts BH. Outcomes of 92 patient-driven family studies for reclassification of variants of uncertain significance. *Genet Med* 21: 1435-1442, 2019. DOI: 10.1038/s41436-018-0335-7.
153. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen C-F, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ. Rat Genome Database (RGD): Mapping disease onto the genome. *Nucleic Acids Res* 30: 125-128, 2002. DOI: 10.1093/nar/30.1.125.
154. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28: 1248-1250, 2010. DOI: 10.1038/nbt1210-1248.
155. UK Biobank [Online]. 2021. *Neale lab*: [date unknown]. http://www.nealelab.is/uk-biobank [March 6, 2021].
156. UniProt Consortium. UniProt: A hub for protein information. *Nucleic Acids Res* 43: D204-D212, 2015. DOI: 10.1093/nar/gku989.
157. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *J Comput Chem* 26: 1701-1718, 2005. DOI: 10.1002/jcc.20291.
158. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, UKA S, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, LB VW, Tsao CW, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2020 Update: A report from the American Heart Association. *Circulation* 141: e139-e596, 2020. DOI: 10.1161/CIR.0000000000000757.
159. Wall RJ, Shani M. Are animal models as good as we think? *Theriogenology* 69: 2-9, 2008. DOI: 10.1016/j.theriogenology.2007.09.030.
160. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *J Comput Chem* 25: 1157-1174, 2004. DOI: 10.1002/jcc.20035.

161. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164, 2010. DOI: 10.1093/nar/gkq603.
162. Wilkinson B, Campbell DB. Contribution of long noncoding RNAs to autism spectrum disorder risk. *Int Rev Neurobiol* 113: 35-59, 2013. DOI: 10.1016/B978-0-12-418700-9.00002-2.
163. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: Diagnosing rare disease in children. *Nat Rev Genet* 19: 253-268, 2018. DOI: 10.1038/nrg.2017.116.
164. Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: A web server for sumoylation site prediction. *Nucleic Acids Res* 34: W254-W257, 2006. DOI: 10.1093/nar/gkl207.
165. Yamada Y, Kato K, Oguri M, Horibe H, Fujimaki T, Yasukochi Y, Takeuchi I, Sakuma J. Identification of 13 novel susceptibility loci for early-onset myocardial infarction, hypertension, or chronic kidney disease. *Int J Mol Med* 42: 2415-2436, 2018. DOI: 10.3892/ijmm.2018.3852.
166. Yamanaka S. Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell* 10: 678-684, 2012. DOI: 10.1016/j.stem.2012.05.005.
167. Yari M, Bitarafan S, Broumand MA, Fazeli Z, Rahimi M, Ghaderian SMH, Mirfakhraie R, Omrani MD. Association between long noncoding RNA ANRIL expression variants and susceptibility to coronary artery disease. *Int J Mol Cell Med* 7: 1-7, 2018. DOI: 10.22088/IJMCM.BUMS.7.1.1.
168. Yeo NC, O'Meara CC, Bonomo JA, Veth KN, Tomar R, Flister MJ, Drummond IA, Bowden DW, Freedman BI, Lazar J, Link BA, Jacob HJ. Shroom3 contributes to the maintenance of the glomerular filtration barrier integrity. *Genome Res* 25: 57-65, 2015. DOI: 10.1101/gr.182881.114.
169. Zamore PD, Tuschl T, Sharp PA, Bartel DP. RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101: 25-33, 2000. DOI: 10.1016/S0092-8674(00)80620-0.
170. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet* 24: R102-R110, 2015. DOI: 10.1093/hmg/ddv259.
171. Zhang W, Zeng B, Yang M, Yang H, Wang J, Deng Y, Zhang H, Yao G, Wu S, Li W. ncRNAVar: A manually curated database for identification of noncoding RNA variants associated with human diseases. *J Mol Biol* 433: 166727, 2021. DOI: 10.1016/j.jmb.2020.166727.
172. Zhu Y, Zhang D, Zhou D, Li Z, Li Z, Fang L, Yang M, Shan Z, Li H, Chen J, Zhou X, Ye W, Yu S, Li H, Cai L, Liu C, Zhang J, Wang L, Lai Y, Ruan L, Sun Z, Zhang S, Wang H, Liu Y, Xu Y, Ling J, Xu C, Zhang Y, Lv D, Yuan Z, Zhang J, Zhang Y, Shi Y, Lai M. Susceptibility loci for metabolic syndrome and metabolic components identified in Han Chinese: A multi-stage genome-wide association study. *J Cell Mol Med* 21: 1106-1116, 2017. DOI: 10.1111/jcmm.13042.
173. Zou Z, Liu C, Che C, Huang H. Clinical genetics of Alzheimer's disease. *Biomed Res Int* 2014: 291862, 2014. DOI: 10.1155/2014/291862.