REVIEW

Human genome variations: databases and bioinformatics resources

Mamoon Rashid

Department of Biostatistics and Bioinformatics, King Abdullah International Medical Research Center, King Saud bin Abdulaziz University for Health Sciences, Ministry of the National Guard-Health Affairs, Riyadh 11426, Saudi Arabia

*Corresponding author e-mail:rashidmamoon@gmail.com

ABSTRACT

The advent of next-generation sequencing technology enabled population scale human genome projects. Contemporary development of high-throughput genotyping arrays contributed to deep characterization of each base of 3.2 billion bases of human genome. To host this vast amount of genetic variant data, large scalable and fast searchable databases and bioinformatics tools for variant annotations and analyses have been developed in the last decade. The databases and computational resources for genetic variants facilitated novel discoveries and observations across many disciplines of biology and medicine, such as the pattern of evolution, speciation, pharmacology, and last but not the least, genetic bases of human health and diseases. This article puts forth the databases and computational resources for human genome variation analysis.

KEYWORDS: human genome, bioinformatics resources, computational biology, human health

Citation: Rashid M. Human genome variations: databases and bioinformatics resources. Polymorphism. 2020;4:21-30.

INTRODUCTION

Recent advancements in genomic techniques and technologies have discovered genetic variants in the human genome at an unprecedented rate that has never been achieved in the history of mankind. Largely attributed to high-throughput next generation DNA sequencing technologies, the human genome is being characterized at a single base resolution, integrating the genotype and phenotype using much shorter time and resources. The emergence of genetic variation databases, such as i) dbSNP and HGV for short genetic variations, ii) dbVar and DGV for structural variations, iii) dbGaP for genotype/phenotype interaction studies, and iv) ClinVar and ClinGen for human variations of clinical significance, facilitates the contemporary identification/discovery of i) known or novel polymorphisms, ii) phenotype to genotype associations, and iii) clinically important human genetic variations. In a rough estimate, the number of Single Nucleotide Polymorphisms (SNPs) in dbSNP database has increased from ~61 million to >695 million (~11 fold) in the past few years (Figure 1A). To drive the functional implication of SNPs, enormous attempts have been made to associate the human diseases to these genetic variations, thus resulting in the growth of ClinVar database from ~64 thousand to ~888 thousand (~14 fold) entries in the last seven years (Figure 1B). Parallel to the above-mentioned growth of genetic variant databases, bioinformatics algorithms or tools have been developed to predict the impact of clinically important variants into different categories like "damaging", "pathogenic", "probable pathogenic" in a very short time. Together, these resources are becoming essential for functional analysis of genetic variants. The author must acknowledge that many of the databases and computational resources of similar theme could not be included due to the limitation of space in this article.

Database of short genetic variations (dbSNP)

Sequence variations in human genome may affect an individual's phenotype. These genetic polymorphisms from different genetic loci might be implicated in defining the susceptibility or propensity of a person towards complex diseases, such as cardiovascular diseases and cancer. In contrast, polymorphism in a single gene or locus can cause a disease that is inherited in Mendelian fashion, called monogenic diseases. The sequence variations or mutations in the genome have been utilized as tools for (i) physical mapping of genomic loci, (ii) functional interpretation, (iii) evolutionary studies, and (iv) large-scale genome-wide association studies.

The dbSNP is a public repository of all types of simple genetic variations from different species. Genetic variants or SNPs could be classified as germline and somatic on the basis of origin of the variants. The database includes single base substitutions (also known as SNPs), multi-base deletions or insertions (called INDELS), retrotransposable element insertions, and microsatellite repeat variations (also called as short tandem repeats or STRs). Each entry in the dbSNP database present the sequence context of the polymorphism (i.e. the surrounding nucleotide sequences), occurrence frequency of the polymorphism in an individual or population, experimental method used to determine that polymorphism.

Physical and genetic mapping

Genetic variations are used as positional markers in the physical map of the genome. Since the polymorphisms have sequence context, the variations mapping uniquely to the genome can serve as stable landmarks in the genome. Genetic maps can be created by identifying positions of different genetic markers, such as i) gene marker, and DNA markers including, ii) RFLP (Restriction Fragment Length Polymorphism), iii) SSLP (Simple

Sequence Length Polymorphism), and iv) SNP (Single Nucleotide polymorphism).

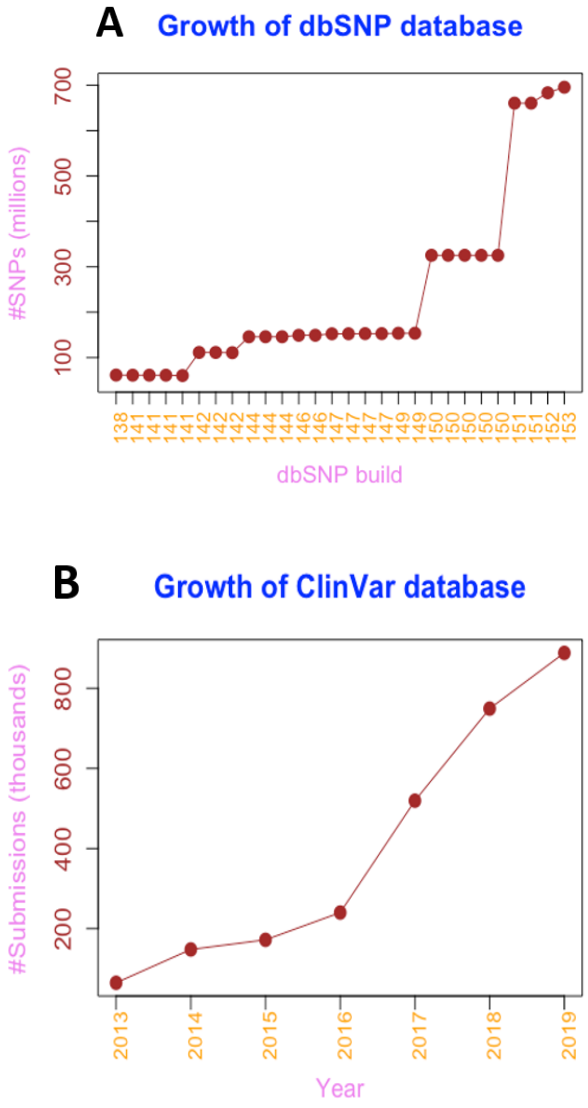

Figure 1: A) Total number of entries in dbSNP database as a function of database build. B) The similar statistics for ClinVar database but as a function of the past few years. These statistics were obtained in October 2019.

The resolving power of the genetic map depends on the cross-over mechanism. In the lower organisms, it is easy to study a large number of cross-overs (i.e. one can have millions of bacteria divided in a short time), but it is very difficult in eukaryotic organisms where the number of progenies is limited in a family. Thus, we need an alternate map of the genome called physical map to aid the genome sequencing project of complex organism like humans. The tools to create physical map could be i) restriction mapping, ii) FISH (Fluorescent in situ hybridization), and iii) STS (Sequence Tagged Sites). Genetic as well as physical mappings have played very crucial roles in human genome sequencing project. For smaller genome likes, bacterial and viral genomes, the assembly of the sequenced contigs is not very difficult due to the presence of high-resolution genetic maps. However, in the higher eukaryotes, the genome sequencing projects need a prior genome map in order to circumvent the issues of i) large amount of DNA repeats in the genome and ii) large size of the genome.

Functional analysis

Polymorphism in functional (coding) or regulatory (non-coding) regions of the genome might cause changes in the pattern of transcribed genes. The variations in the coding region of the gene might impact the final protein product depending on the nature of mutation (i.e. synonymous or non-synonymous). The non-synonymous SNP (nsSNP) leads to i) an amino acid change in the protein (called missense mutation), ii) stop-gain or chain-terminating codon, leading to premature truncated protein (nonsense mutation), and iii) stop-loss (mutation in the original stop codon) leading to abnormal extension to the carboxyl terminal of protein. In contrast, the SNPs in the regulatory regions (rSNP) of the gene might affect the transcription factor binding efficiency, altering the transcript levels.

The point of concern is the impact of nsSNPs on the normal function of the individual. Some nsSNPs can be tolerated and others might cause the disease. Bioinformatic prediction of tolerance for a specific nsSNP helps prioritize the potential disease-causing variants in humans. Such predictions are based on protein or amino acid features (at variant sites), such as sequence features, biophysical properties, structural

properties, and evolutionary properties derived from sequence alignment. Supervised machine learning methods are largely used to develop such prediction models. Most important methods include SIFT(Vaser et al. 2016), PolyPhen (Adzhubei et al. 2010), MutDB (Singh et al. 2008), SNPeffect (De Baets et al. 2012), MutationAssessor (Reva, Antipin, and Sander 2011), which predict the impact of non-synonymous mutations on protein function in different categorical classes such as "damaging", "benign", and "probable damaging".

Evolutionary studies

Genetic variations have been extensively used in studying the evolution of the genetic lineages of human population locally or globally (Chaubey et al. 2008; Romualdi et al. 2002). There has been a big initiative to identify the set of nearby SNPs on the same chromosome inherited in blocks (called "haplotype") in human genome. The term haplotype may be defined as a group of alleles in an organism that are inherited from a single parent. Some of the SNPs in the haplotype are unique enough to identify that haplotype and are thus called tag SNPs. Tag SNP may be defined as a representative SNP in the region of genome with high linkage disequilibrium. International HapMap project [https://www.genome.gov/10001688/international-hapmap-project] was an effort to create a map of such haplotypes on the human genome. Moreover, another big initiative was 1000 genome project (Altshuler et al. 2010; Genomes Project et al. 2012; Auton et al. 2015) [https://www.internationalgenome.org/] that seems to overshadow the utility of HapMap project. The resources created by 1000 genome project are valuable and are serving as a global reference for human genetic variations. A few of the most common applications (Zheng-Bradley and Flicek 2017) of this project are i) genotype imputation (supporting GWAS studies), ii) mapping eQTL (expression Quantitative Trait Loci), iii) prioritization of variants for pathogenicity, iv) supporting whole genome and cancer genome sequencing projects, and v) studying population structure and molecular evolution. Nevertheless, the quality assessment of 1000 genome dataset reflects that imputation and phasing for rare variants are unreliable (Belsare et al. 2019).

The genome similarity of two randomly selected human individuals is maximum 99.9%. Thus, 0.1% of their DNA is dissimilar, giving rise to individuality. Moreover, these variations in human genomes define diversity, susceptibility to diseases, and precision of the response to medicines (Shastry 2002). Positive selection has been an important evolutionary tool to shape the modern humans. Through the advent of the high-throughput genome sequencing technologies and the robust statistical models, various population scale evolutionary measurements have become possible to compute, such as heterozygosity, F_{ST} (population differentiation) (Beaumont 2005), and Tajima's D (allele frequency spectrum) (Tajima 1989) to identify positive signals (Cheng et al. 2009).

Genome-wide association studies (GWAS)

Genetic variations are well known to affect human phenotype, including health and diseases. Differences in SNPs between two individuals may lead to different phenotypes. With the availability of large genome-wide genetic variants data across populations, one can attempt to identify variant alleles linked to a particular phenotype such as ageing (Melzer, Pilling, and Ferrucci 2019), adipocity (Rask-Andersen et al. 2019), cancer (Couch et al. 2016), cardiovascular(Yao et al. 2018), or diabetes (Fuchsberger et al. 2016). This area named Genome-wide Association Studies (GWAS) dominated the genomics field for about a decade. The underlying assumption in GWAS is that the causal allele frequency is higher in cases (individuals with the trait under study) than in controls (individuals without the trait under study). Moreover, GWAS has its own benefits and limitations discussed in this excellent review article

(Tam et al. 2019). A recent article summarized the GWAS discoveries in the context of diseases biology and its application towards new therapeutics (Visscher et al. 2017).

There are several databases hosting the GWAS datasets and their output for public use. Few such databases are, i) NHGRI-EBI catalog of published genome-wide association studies called "GWAS Catalog" [https://www.ebi.ac.uk/gwas/], ii) GWAS central [https://www.gwascentral.org/], iii) GWASdb v2 [http://jjwanglab.org/gwasdb], iv) Case-control GWAS database [https://gwas.biosciencedbc.jp/cgi-bin/gwasdb/gwas_top.cgi], v) Human Genome Variation Database [https://gwas.biosciencedbc.jp/], vi) GWAS Atlas [https://atlas.ctglab.nl/].

The database of structural variations (dbVar)

The database of structural variation (dbVar) [https://www.ncbi.nlm.nih.gov/dbvar/] is a part of the NCBI databases. Structural variations (SV) comprise insertion, deletion, inversion, translocation, and duplication. By definition, SV is a region of DNA larger than 1000 bp (1kb) including genomic imbalances and commonly referred to as the copy number variation (CNV). The CNVs have also become notable for their contribution in genetic diversity and diseases (Redon et al. 2006; Freeman et al. 2006). Currently, the non-redundant dbVar database houses more than 2.2 million deletions, 1.1 million insertions, and 300 thousand duplications. Few other databases are available to mine the human genome structural variation data such as, i) Database of Genomic Variants (MacDonald et al. 2014) [http://dgv.tcag.ca/dgv/app/home], ii) An open resource of structural variations (Collins et al. 2019), iii) JVar-SV [https://www.ddbj.nig.ac.jp/jvar-sv/index-e.html]. A recent article describes the integrated set of eight structural variation classes constructed using short-read sequencing data and statistically phased haplotype blocks in 26 human populations (Sudmant et al. 2015). The major finding of this study was that the structural variations are enriched on haplotypes identified using GWAS studies and exhibit enrichment for expression quantitative trait loci (eQTL) (Sudmant et al. 2015).

The database of genotypes and phenotypes (dbGaP)

The dbGaP (Mailman et al. 2007; Tryka et al. 2014) is a collection of datasets and the corresponding results generated from studies investigating the association of genotypes and phenotypes [https://www.ncbi.nlm.nih.gov/gap/]. Broadly two classes of data are submitted to dbGaP, i) molecular data and ii) phenotypic data. Molecular data comprise genotype data, expression data, epigenomic data, genomic sequence data and somatic mutation data. Phenotype data are summarized as columns representing clinical, demographic, and exposure data for each row as a subject. Currently more than 2.4 million molecular assays have been documented in dbGaP studies. Each of such assays can have hundreds to millions of data points (e.g. SNPs, CNV, and nucleotides). The major objective of genotype-phenotype databases is to facilitate assigning pathogenicity to genetic variants. Currently, the quality and volume of phenotypic data compared to genotypic data in genotype-phenotype databases is lower owing to ethical, financial, legal and other challenges that must be overcome to produce large-scale phenotypic data (Brookes and Robinson 2015).

ClinVar database

ClinVar (Landrum and Kattman 2018; Landrum et al. 2018; 2016) [https://www.ncbi.nlm.nih.gov/clinvar/] is the public repository of the asserted or validated relationships among human genetic variants and the clinically significant phenotypes. These records

are supported by evidence. It also works as a collaboration among medical geneticists to identify the clinically significant genetic variations. ClinVar accepts submission of variants for any part of genome and for all types of conditions. Clinical testing, preclinical research, and published scientific literature form the basis of clinical assertion of variants in ClinVar database. Submissions are accepted at different complexity levels ranging from variant-level submission (representation of an allele and its interpretation) to case-level submission (experimental evidence about the effect of variants on phenotype). Currently this database contains about 888 thousand submissions.

Human Genome Variation (HGV) database

HGV [https://hgv.figshare.com/] is a collection of human genetic variants published in peer-reviewed "Data Reports" and other articles in *Human Genome Variation* journal. Data reports are short reports describing human genome variation and variability and their associated disease/phenotype. It has a standardized format of data collection for each variant. Data reports are linked to its HGV database links and vice versa. This database provides various filters such as population type, mutation type, and zygosity type, which can be used to compile useful datasets.

Database of Genomic Variants (DGV)

DGV(MacDonald et al. 2014) (http://dgv.tcag.ca/dgv/app/home) is a curated database of human genomic structural variation. The structural variation has been defined as any genetic alterations involving segment of DNA larger than 50 base pairs. This database contains comprehensive summary of human genome structural variations only from healthy control individuals. Currently DGV hosts more than 6.3 million CNVs and about 30446 inversions. This database may serve as a catalog of control data to be used in the studies aiming to correlate structural variations to phenotypic data.

ClinGen (Clinical Genome Resource)

ClinGen (https://www.clinicalgenome.org) is a resource for exploring clinical relevance of genes and variants. It is an NIH-funded program dedicated to developing a central resource that defines the clinical relevance of genes and variants. ClinGen presents important curation tools such as,i) Gene-disease validity tools, ii) Variant pathogenicity tools, iii) Dosage sensitivity tools, and iv) Clinical actionability tools to the scientific community. This resource is very useful in seeking answers to questions like,i) 'is this gene associated with a disease?', ii) 'is this variant causative?', iii) 'is this information actionable?'. ClinGen (Dolman et al. 2018) is facilitating to build a genomic knowledgebase together with ClinVar and other resources.

Bioinformatic resources for functional interpretation of genetic variants

Genetic variants can be broadly classified into two groups, i) lying within the coding regions (cSNP), and ii) lying within the noncoding or regulatory regions (rSNP) of the genome. Interpretation of the cSNP is straightforward and relatively easy due to its mapping on the gene structure that is translated into a protein product. Therefore, one can know the exact location of cSNP in the frame of translation. cSNP may be synonymous or non-synonymous based on its location on the triplet codon and the degeneracy of codons. Synonymous SNP leads to no change in resulting amino acid in protein, but non-synonymous SNP do. Non-synonymous may be divided into different types such as missense, non-sense (stop gain), stop loss, etc. All this positional information along with great deal of functional annotations about SNPs can be obtained through softwares, such as ANNOVAR(Wang, Li, and Hakonarson 2010; Yang and Wang 2015), SnpEff (Cingolani et al. 2012), Variant Annotation Tools (San lucas et al. 2012), VarAFT (Desvignes et al. 2018), Vcfanno

(Pedersen, Layer, and Quinlan 2016). With this basic knowledge of non-synonymous mutations, one can use bioinformatics prediction methods to assess the impact of these mutations (Vaser et al. 2016; Adzhubei et al. 2010). The structural and functional interpretation of SNPs was also studied in the context of protein interaction network (Lu, Herrera Braga, and Fraternali 2016).

On the contrary, the functional interpretation of rSNP is relatively difficult for many reasons for e.g. i) it is not expressed as a protein product, ii) the regulatory mechanism in the human genome is complex and requires the knowledge of many layers of information to identify whether a rSNP is functional or not. Moreover, like the cSNPs, rSNPs have also been known to affect human health to a great deal. In cancer, the role of rSNPs is being studied deeply and this type of variations may disrupt transcription factor-binding sites, affecting the resulting gene expression or may affect non-coding RNA loci (Khurana et al. 2016).

There are various computational algorithms being developed to assess coding and non-coding variants in many different disease models. A method named Prioritization And Functional Assessment (PAFA) based on population differentiation measures was developed to assess the noncoding variants associated with complex diseases (Zhou and Zhao 2018). One can run an in-silico analysis on a set of SNPs (coding and noncoding) to identify the most impactful SNPs for a particular gene associated with a disorder, such as obesity and insulin resistance (Elkhattabi et al. 2019), retinal vasculature defect (Madelaine et al. 2018), Alzheimer's disease (Tey and Ng 2019). Noncoding risk variants can also be identified using disease-related gene regulatory networks (Gao et al. 2018). Another recent computational pipeline exploiting the conservation of the SNP alleles revealed a set of regulatory SNPs relevant to the peripheral nerve (Law et al. 2018). Moreover, the rSNPs at the promoter region of the *ERCC5* gene have been studied for their implication in transcription factor binding and thus affecting gene expression (Chen et al. 2016). To conclude this, I refer the readers to a few good reads about functional characterization of noncoding variants (Jin et al. 2018; Nishizaki and Boyle 2017).

Discussion

The evidences presented in this article suggest that there has been an unprecedented increase in the identification of human genetic variations (Figure 1A). Unsurprisingly, the functional studies of SNPs kept pace with the emergence of the huge SNP datasets as evidenced by the statistics of ClinVar database (Figure 1B). Along with the development of genetic variations databases, the evolution of bioinformatics tools and prediction algorithms to assess or interpret the functional significance of SNPs dataset is no longer lagging behind. Together these bioinformatics resources are serving the scientific community, saving their time and experimental complexity in order to predict the functional impact of a mutation.

The resources described in this manuscript have been heavily used by the evolutionary biologists to capture the genetic diversity within and across the human populations. Population-scale genome and exome sequencing revealed that individuals from different populations carry different profiles of rare and common variants and that low-frequency variants show substantial geographic differentiation (Genomes Project et al. 2012). The recent update from 1000 genome project provides statistics on over 88 million variants (84.7 million SNPs, 3.6 million short INDELS, and 60,000 structural variants) (Auton et al. 2015). This resource contains >99% of SNPs with an allele frequency of >1% derived from a variety of human ancestries. Together, the human genetic variation dataset provides insights into the evolutionary processes that shape genetic diversity. Moreover, the quality assessment of the large-scale dataset is also important to make use of the resources properly (Belsare et al. 2019).

The global reference of human genetic variations has great applications in identifying disease causing variants and developing diagnostics and therapeutics to treat various diseases (Lek et al. 2016). In context of rare genetic diseases (e.g. cystic fibrosis, phenylketonuria, sickle-cell anemia, Huntington's disease), the identification of causal variants (with low allele frequency) greatly depends on the availability of large human genome dataset along with computed allele frequency. Now we know that rare or low frequency variants not only cause rare genetic diseases but also cause common diseases (Bomba, Walter, and Soranzo 2017) or traits such as LDL cholesterol, low or high levels of cholesterol, triglyceride, type 2 diabetes, Alzheimer's disease, short stature, height, adiponectin issues etc.

Nonetheless, the risk of genetic diseases can be misestimated across global population due to the fact that risk allele frequencies at known disease loci are significantly different for African populations compared to other continents (Kim et al. 2018). These continental differences in the risk allele frequencies can be moderately reduced by using whole genome sequences and hundreds of thousands of cases and controls across various populations. The availability of population specific genome databases in the future will help in precision diagnostic and medicine.

Acknowledgements

The author cordially thanks to Dr. Rajender Singh for his help in completing this manuscript and anonymous reviewers for their valuable time and effort.

Conflict of Interest

The author declares no competing interest exists.

REFERENCES

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4): 248–49. https://doi.org/10.1038/nmeth0410-248.

Altshuler, David L., Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, et al. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73. https://doi.org/10.1038/nature09534.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature*. https://doi.org/10.1038/nature15393.

Baets, Greet De, Joost Van Durme, Joke Reumers, Sebastian Maurer-Stroh, Peter Vanhee, Joaquin Dopazo, Joost Schymkowitz, and Frederic Rousseau. 2012. "SNPeffect 4.0: On-Line Prediction of Molecular and Structural Effects of Protein-Coding Variants." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr996.

Beaumont, Mark A. 2005. "Adaptation and Speciation: What Can Fst Tell Us?" *Trends in Ecology and Evolution* 20 (8): 435–40. https://doi.org/10.1016/j.tree.2005.05.017.

Belsare, Saurabh, Michal Levy-Sakin, Yulia Mostovoy, Steffen Durinck, Subhra Chaudhuri, Ming Xiao, Andrew S. Peterson, Pui Yan Kwok, Somasekar Seshagiri, and Jeffrey D. Wall. 2019. "Evaluating the Quality of the 1000 Genomes Project Data." *BMC Genomics*. https://doi.org/10.1186/s12864-019-5957-x.

Bomba, Lorenzo, Klaudia Walter, and Nicole Soranzo. 2017. "The Impact of Rare and Low-Frequency Genetic Variants in Common Disease." *Genome Biology*. https://doi.org/10.1186/s13059-017-1212-4.

Brookes, Anthony J., and Peter N. Robinson. 2015. "Human Genotype-Phenotype Databases: Aims, Challenges and Opportunities." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3932.

Chaubey, G, M Karmin, E Metspalu, M Metspalu, D Selvi-Rani, V K Singh, J Parik, et al. 2008. "Phylogeography of MtDNA Haplogroup R7 in the Indian Peninsula." *BMC Evol Biol* 8: 227. https://doi.org/1471-2148-8-227 [pii]10.1186/1471-2148-8-227.

Chen, Jianfang, Xi Luo, Ganfeng Xie, Keli Chen, Heng Jiang, Feng Pan, Jianjun Li, Zhihua Ruan, Xueli Pang, and Houjie Liang. 2016. "Functional Analysis of SNPs in the ERCC5 Promoter in Advanced Colorectal Cancer Patients Treated With Oxaliplatin-Based Chemotherapy." *Medicine* 95 (19): e3652. https://doi.org/10.1097/MD.0000000000003652.

Cheng, Feng, Wei Chen, Elliott Richards, Libin Deng, and Changqing Zeng. 2009. "SNP@Evolution: A Hierarchical Database of Positive Selection on the Human Genome." *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-9-221.

Cingolani, P, A Platts, L Wang le, M Coon, T Nguyen, L Wang, S J Land, X Lu, and D M Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3." *Fly (Austin)* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Amit V. Khera, Laurent C. Francioli, et al. 2019. "An Open Resource of Structural Variation for Medical and Population Genetics." *BioRxiv*. https://doi.org/10.1101/578674.

Couch, Fergus J., Karoline B. Kuchenbaecker, Kyriaki Michailidou, Gustavo A. Mendoza-Fandino, Silje Nord, Janna Lilyquist, Curtis Olswold, et al. 2016. "Identification of Four Novel Susceptibility Loci for Oestrogen Receptor Negative Breast Cancer." *Nature Communications*. https://doi.org/10.1038/ncomms11375.

Desvignes, Jean Pierre, Marc Bartoli, Valérie Delague, Martin Krahn, Morgane Miltgen, Christophe Béroud, and David Salgado. 2018. "VarAFT: A Variant Annotation and Filtration System for Human next Generation Sequencing Data." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky471.

Dolman, Lena, Angela Page, Lawrence Babb, Robert R. Freimuth, Harindra Arachchi, Chris Bizon, Matthew Brush, et al. 2018. "ClinGen Advancing Genomic Data-Sharing Standards as a GA4GH Driver Project." *Human Mutation*. https://doi.org/10.1002/humu.23625.

Elkhattabi, Lamiae, Imane Morjane, Hicham Charoute, Soumaya Amghar, Hind Bouafi, Zouhair Elkarhat, Rachid Saile, Hassan Rouba, and Abdelhamid Barakat. 2019. "In Silico Analysis of Coding/Noncoding SNPs of Human RETN Gene and Characterization of Their Impact on Resistin Stability and Structure." *Journal of Diabetes Research* 2019. https://doi.org/10.1155/2019/4951627.

Freeman, Jennifer L., George H. Perry, Lars Feuk, Richard Redon, Steven A. McCarroll, David M. Altshuler, Hiroyuki Aburatani, et al. 2006. "Copy Number Variation: New Insights in Genome Diversity." *Genome Research*. https://doi.org/10.1101/gr.3677206.

Fuchsberger, Christian, Jason Flannick, Tanya M. Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J. Gaulton, Clement Ma, et al. 2016. "The Genetic Architecture of Type 2 Diabetes." *Nature*. https://doi.org/10.1038/nature18642.

Gao, Long, Yasin Uzun, Peng Gao, Bing He, Xiaoke Ma, Jiahui Wang, Shizhong Han, and Kai Tan. 2018. "Identifying Noncoding Risk Variants Using Disease-Relevant Gene Regulatory Networks." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-03133-y.

Genomes Project, Consortium, G R Abecasis, A Auton, L D Brooks, M A DePristo, R M Durbin, R E Handsaker, H M Kang, G T Marth, and G A McVean. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422): 56–65. https://doi.org/10.1038/nature11632.

Jin, Yu, Jingbo Wang, Maulana Bachtiar, Samuel S. Chong, and Caroline G. L. Lee. 2018. "Architecture of Polymorphisms in the Human Genome Reveals Functionally Important and Positively Selected Variants in Immune Response and Drug Transporter Genes." *Human Genomics* 12 (1): 43. https://doi.org/10.1186/s40246-018-0175-1.

Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. 2016. "Role of Non-Coding Sequence Variants in Cancer." *Nature Reviews Genetics* 17 (2): 93–108. https://doi.org/10.1038/nrg.2015.17.

Kim, Michelle S., Kane P. Patel, Andrew K. Teng, Ali J. Berens, and Joseph Lachance. 2018. "Genetic Disease Risks Can Be Misestimated across Global Populations." *Genome Biology*. https://doi.org/10.1186/s13059-018-1561-7.

Landrum, Melissa J., and Brandi L. Kattman. 2018. "ClinVar at Five Years: Delivering on the Promise." *Human Mutation*. https://doi.org/10.1002/humu.23641.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1222.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1153.

Law, William D., Elizabeth A. Fogarty, Aimée Vester, and Anthony Antonellis. 2018. "A Genome-Wide Assessment of Conserved SNP Alleles Reveals a Panel of Regulatory SNPs Relevant to the Peripheral Nerve." *BMC Genomics* 19 (1): 311. https://doi.org/10.1186/s12864-018-4692-z.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature*. https://doi.org/10.1038/nature19057.

Lu, Hui-Chun, Julián Herrera Braga, and Franca Fraternali. 2016. "PinSnps: Structural and Functional Analysis of SNPs in the Context of Protein Interaction Networks." *Bioinformatics (Oxford, England)* 32 (16): 2534–36. https://doi.org/10.1093/bioinformatics/btw153.

MacDonald, Jeffrey R., Robert Ziman, Ryan K.C. Yuen, Lars Feuk, and Stephen W. Scherer. 2014. "The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt958.

Madelaine, Romain, James H. Notwell, Gemini Skariah, Caroline Halluin, Charles C. Chen, Gill Bejerano, and Philippe Mourrain. 2018. "A Screen for Deeply

Conserved Non-Coding GWAS SNPs Uncovers a MIR-9-2 Functional Mutation Associated to Retinal Vasculature Defects in Human." *Nucleic Acids Research* 46 (7): 3517–31. https://doi.org/10.1093/nar/gky166.

Mailman, Matthew D., Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, et al. 2007. "The NCBI DbGaP Database of Genotypes and Phenotypes." *Nature Genetics*. https://doi.org/10.1038/ng1007-1181.

Melzer, David, Luke C. Pilling, and Luigi Ferrucci. 2019. "The Genetics of Human Ageing." *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0183-6.

Nishizaki, Sierra S, and Alan P Boyle. 2017. "Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms." *Trends Genet* 33 (1). https://doi.org/10.1016/j.tig.2016.10.008.

Pedersen, Brent S., Ryan M. Layer, and Aaron R. Quinlan. 2016. "Vcfanno: Fast, Flexible Annotation of Genetic Variants." *Genome Biology*. https://doi.org/10.1186/s13059-016-0973-5.

Rask-Andersen, Mathias, Torgny Karlsson, Weronica E. Ek, and Åsa Johansson. 2019. "Genome-Wide Association Study of Body Fat Distribution Identifies Adiposity Loci and Sex-Specific Genetic Effects." *Nature Communications*. https://doi.org/10.1038/s41467-018-08000-4.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." *Nature*. https://doi.org/10.1038/nature05329.

Reva, B, Y Antipin, and C Sander. 2011. "Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics." *Nucleic Acids Res* 39 (17): e118. https://doi.org/10.1093/nar/gkr407.

Romualdi, Chiara, David Balding, Ivane S Nasidze, Gregory Risch, Myles Robichaux, Stephen T Sherry, Mark Stoneking, Mark A Batzer, and Guido Barbujani. 2002. "Patterns of Human Diversity, within and among Continents, Inferred from Biallelic DNA Polymorphisms." *Genome Research* 12 (4): 602–12. https://doi.org/10.1101/gr.214902.

San lucas, F. Anthony, Gao Wang, Paul Scheet, and Bo Peng. 2012. "Integrated Annotation and Analysis of Genetic Variants from Next-Generation Sequencing Studies with Variant Tools." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr667.

Shastry, Barkur S. 2002. "SNP Alleles in Human Disease and Evolution." *Journal of Human Genetics*. https://doi.org/10.1007/s100380200086.

Singh, Arti, Adebayo Olowoyeye, Peter H. Baenziger, Jessica Dantzer, Maricel G. Kann, Predrag Radivojac, Randy Heiland, and Sean D. Mooney. 2008. "MutDB: Update on Development of Tools for the Biochemical Analysis of Genetic Variation." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm659.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature*. https://doi.org/10.1038/nature15394.

Tajima, F. 1989. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics*.

Tam, Vivian, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. 2019. "Benefits and Limitations of Genome-Wide Association Studies." *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0127-1.

Tey, Han Jieh, and Chong Han Ng. 2019. "Computational Analysis of Functional SNPs in Alzheimer's Disease-Associated Endocytosis Genes." *PeerJ* 7 (September): e7667. https://doi.org/10.7717/peerj.7667.

Tryka, Kimberly A., Luning Hao, Anne Sturcke, Yumi Jin, Zhen Y. Wang, Lora Ziyabari, Moira Lee, et al. 2014. "NCBI's Database of Genotypes and Phenotypes: DbGaP." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1211.

Vaser, R, S Adusumalli, S N Leng, M Sikic, and P C Ng. 2016. "SIFT Missense Predictions for Genomes." *Nat Protoc* 11 (1): 1–9. https://doi.org/10.1038/nprot.2015.123.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq603.

Yang, H, and K Wang. 2015. "Genomic Variant Annotation and Prioritization with ANNOVAR and WANNOVAR." *Nat Protoc* 10 (10): 1556–66. https://doi.org/10.1038/nprot.2015.105.

Yao, Chen, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B. Sun, et al. 2018. "Genome-wide Mapping of Plasma Protein QTLs Identifies Putatively Causal Genes and Pathways for Cardiovascular Disease." *Nature Communications*. https://doi.org/10.1038/s41467-018-05512-x.

Zheng-Bradley, Xiangqun, and Paul Flicek. 2017. "Applications of the 1000 Genomes Project Resources." *Briefings in Functional Genomics*. https://doi.org/10.1093/bfgp/elw027.

Zhou, Lin, and Fangqing Zhao. 2018. "Prioritization and Functional Assessment of Noncoding Variants Associated with Complex Diseases." *Genome Medicine* 10 (1): 53. https://doi.org/10.1186/s13073-018-0565-y.