READING NOTES: AN INTEGRATED MAP OF GENETIC VARIATION FROM 1,092 HUMAN GENOMES

MICHAEL DEWITT

Caveat Emptor

Objective

Charactertize rare variants in the human genome defined as those SNPs at 1% of the population.

1. Data sources

The 1092 human genomes data set is comprised of a mixed of low coverage whole-genome sequence (WGS) data, which is lower coverage as there are fewer reads per sequence, targeted deep exome sequence data, and dense SNP(13) data. These data were gathered from 14 different populations drawing from a mixture of existing and newly collected data. The fourteen different populations were:

- ASW: people living with African ancestry in Southwest United States (AFR)
- CEU: Utah residents with ancestry from North and Western Europe (EUR)
- CHB: Han Chinese in Beijing, China (EAS)
- CHS: Han Chinese is South, China (EAS)
- CLM: Colombians in Medellin, Columbia (AMR)
- FIN: Finnish in Finland (EUR)
- GBR: British from England and Scotland, UK (EUR)
- IBS: Iberian populations in Spain (EUR)
- LWK: Luhya in Webuye, Kenya (AFR)
- JPT: Japanese in Tokyo, Japan (EAS)
- MXL: people with Mexican ancestry in Los Angeles, Califormia (AMR)
- PUR: Puerto Ricans in Puerto Rica, USA (AMR)
- TSI: Toscani in Italia (EUR)
- YRI Yoruba in Ibadan, Nigeria (AFR)

There is some obvious selection bias in these data as many are from European or European descendents and these groupers are a bit vague. Central Asia, India, Australia, Northern Africa and the Middle East, South Central Africa, and many Pacific islands are completely missing. Similarly, the groupings fail to capture known unique populations (Basque in Spain) while grouping some cosmopolitans together with more murky migration history.

1.1. **Focus for mutations.**

Due to the challenges of identifying complex and large variants and shorter indels in regions of low complexity the researchers focused on:

- Biallelic(1) indels
- Large deletions

1.2. **Power the analysis for rare variants.**

They calculate that their analysis is sufficiently powered to detect an SNP(13) present at 1% of the study population at 99.3% (Figure 1). Similarly, they find that they can detect SNPs at 0.1% at over 90% power (except in the WGS data). This is due to lower read coverage in the WGS data as there

are fewer reads and thus less opportunity to separate noise/error from the processing at these lower frequencies.

The researchers used information from inferred haplotypes to enrich their data (using known familial pairings from mother-father-offspring trios). These results were used to enrich their data and improve their power (through the use of the linkage disequilibrium, LD(10) , as measured from these familial triads).

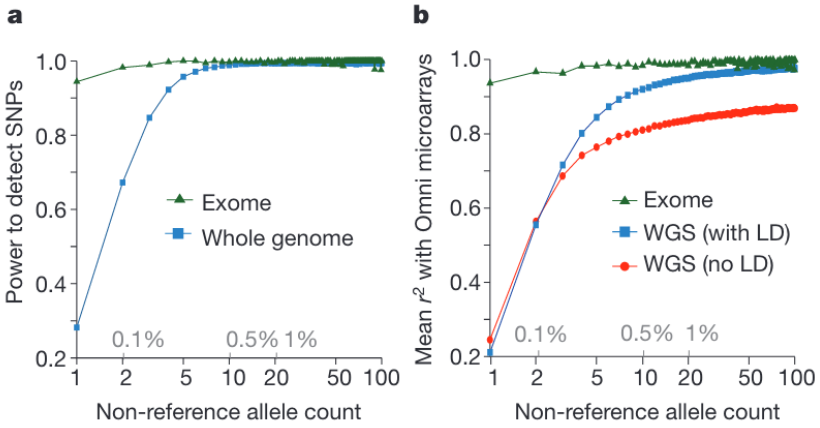

FIGURE 1. a, Power to detect SNPs as a function of variant count (and proportion) across the entire set of samples, estimated by comparison to independent SNP(13) array data in the exome (green) and whole genome (blue). b, Genotype accuracy compared with the same SNP(13) array data as a function of variant frequency, summarized by the r2 between true and inferred genotype (coded as 0, 1 and 2) within the exome (green), whole genome after haplotype integration (blue), and whole genome without haplotype integration (red). LD(10) , linkage disequilibrium; WGS, whole-genome sequencing.

2. GENETIC VARIATION WITHIN AND BETWEEN POPULATIONS

Many of the more *common variants* identified in this study have been *previously described* (94% of variants with $\geq 5\%$). This study described some additional, less well-described variants.

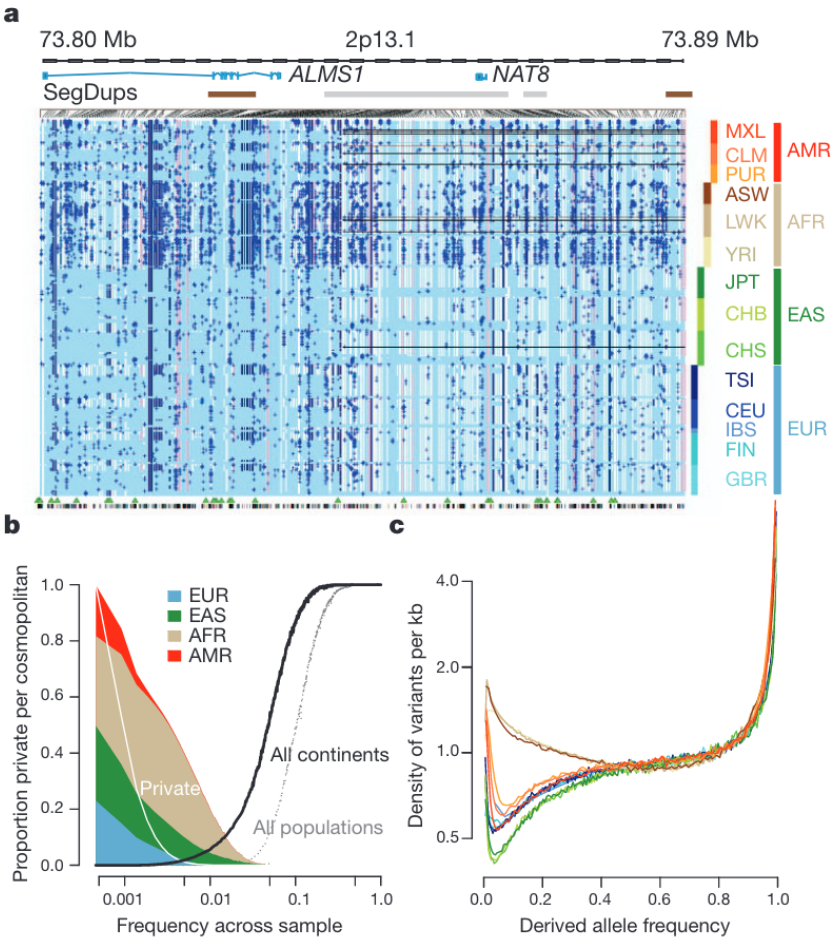

FIGURE 2. The distribution of rare and common variants. a, Summary of inferred haplotypes across a 100-kb region of chromosome 2 spanning the genes ALMS1 and NAT8, variation in which has been associated with kidney disease45. Each row represents an estimated haplotype, with the population of origin indicated on the right. Reference alleles are indicated by the light blue background. Variants (non-reference alleles) above 0.5% frequency are indicated by pink (typed on the high-density SNP(13) array), white (previously known) and dark blue (not previously known). Low frequency variants (,0.5%) are indicated by blue crosses. Indels are indicated by green triangles and novel variants by dashes below. A large, low-frequency deletion (black line) spanning NAT8 is present in some populations. Multiple structural haplotypes mediated by segmental duplications are present at this locus, including copy number gains, which were not genotyped for this study. Within each population, haplotypes are ordered by total variant count across the region. b, The fraction of variants identified across the project that are found in only one population (white line), are restricted to a single ancestry-based group (defined as in a, solid colour), are found in all groups (solid black line) and all populations (dotted black line). c, The density of the expected number of variants per kilobase carried by a genome drawn from each population, as a function of variant frequency (see Supplementary Information). Colours as in a. Under a model of constant population size, the expected density is constant across the frequency spectrum.

2.1. **Many common variants.**

This study found that variants present at $\geq 10\%$ were found in all of the population groups. 53% of the rare variants at 0.5% were found in a single population. They found that derived allele(22) frequency distribution diverged below 40% such that those individuals from African backgrounds carry three times as many low-frequency mutations. This may reflect ancestral bottlenecks in non-African populations. All populations shown rare variants (< 0.5 % frequency) which likely reflects the growing population sizes.

2.2. **Infering history.**

The researchers then examined some patterns sharing of variants (they refer to these as f_2 mutations). These reflects some "between population" sharing of mutations such as:

- Spanish population mutations are more likely to appear in those persons from Americas rather than other European grounds
- Within East Asian populations those from Han Chinese South, China are more likely shared with Han Chinese in Beijing rather than Japan. However, Japanese mutations are more likely to be shared with those from Beijing.
- Those persons with African descent in the American Southwest have mutations shared with those from Yoruba in Nigeria than Luhya in Kenya.

They also saw interesting dynamics with the Finnish have mutations more closely related to the African populations tested (which makes sense given the relative isolation of the Finnish language).

They found a negative correlation between the variant frequency and median length of the shared haplotype (i.e., longer mutations were less likely to be shared or were less predominant).

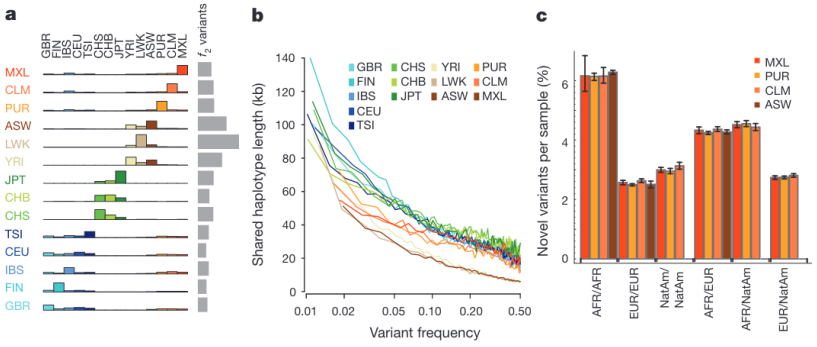

FIGURE 3. a, Sharing of f2 variants, those found exactly twice across the entire sample, within and between populations. Each row represents the distribution across populations for the origin of samples sharing an f2 variant with the target population (indicated by the left-hand side). The grey bars represent the average number of f2 variants carried by a randomly chosen genome in each population. b, Median length of haplotype identity (excluding cryptically related samples and singleton variants, and allowing for up to two genotype errors) between two chromosomes that share variants of a given frequency in each population. Estimates are from 200 randomly sampled regions of 1 Mb each and up to 15 pairs of individuals for each variant. c, The average proportion of variants that are new (compared with the pilot phase of the project) among those found in regions inferred to have different ancestries within ASW, PUR, CLM and MXL populations. Error bars represent 95% bootstrap confidence intervals. NatAm, Native American.

2.3. **Admixture(21) .**

Admixture(21) was examined between the cosmopolitans. On average MXL groups had the greatest proportion of Native American ancestry, but the individual variance was very high (3% to 92% despite an average of 47%).

3. Understanding purifying selection

The purpose of figure 4 from *An integreated map of genetic variation from 1,092 human genomes* [1] is to illustrate the role of purifying selecting within and between populations. Critically, this first requires a discussion of purifying selection. **Purifying selection** or negative selection(31) is a type of background selection resulting in lower genetic diversity [2]. When mutations occur in the genome which are highly deleterious, offspring do not survive long enough to pass on these mutations to subsequent generations, at least on longer timescales [2]. When this background selection occurs, the observed genetic diversity is lower than what would be expected under neutral substitition. However, these dynamics exist within the broader population level (longer) timescales and periodic deleterious mutations do appear to exist on shorter term time scales. Additionally, as described by Vitti et al, "selection operates at the level of the phenotype, alleles showing evidence of selection are likely to be of funtional relevance" [3]. Thus we would anticipate purifying selection to act on those alleles with functional relevance. Figure 4 of the human genomes paper then seeks to test this hypothesis by examing the rate of rare mutations at sites with different levels of conservation and assessing the conservation of sites associated with particular functionality.

4. Quantifying human variation

4.1. **Genomic Evolutionary Rate Profiling Score.**

The Genomic Evolutionary Rate Profiling Score (GERP(6)) scores provide a way of measuring which sites likely lead to delertious mutations. This statistical framework provides a way of estimating the difference between the expected number of mutations under a neutral substitution model (which assumes no impact on fitness) and the observed variation. Positive GERP(6) scores thus represent fewer mutations than would be expected, likely indicating a more conserved site, while negative scores would indicate more substitutions than expected. Put another way, we would expect that high GERP(6) scores will occur in regions which are important for survival to reproductive age and are largely conserved in the population (i.e., at higher levels in the population with fewer mutations in these regions).

4.2. **Derived Allele(22) Frequency.**

We can examine the derived allele(22) frequency (DAP) to assess the overall distribution of alleles within a population. As a reminder, a derived allele(22) are variants which have arisen since the last common ancestor. The derived allele(22) frequency is then a summarization of the pattern and frequency that these variant alleles appear. Taking the population sampled as a whole we can calculate the frequency with which each allele(22) variant appears.

5. Examination of Figure 3

5.1. **Panel A.**

In Figure 4 we see the following:

- X-axis: the GERP(6) score representing the evolutionary conservation where higher scores are more conserved.
- Y-axis: the proportion of variants with a DAF(4) $< 0.5\%$ where higher values indicate lower frequencies in the studied population
- Colored lines: the different functional elements
- cross on the x and y axes representing the average values for GERP(6) score and proportion of variants with a DAF(4) $< 0.5\%$, respectively.

From this figure we can conclude that:

- More generally, there are fewer mutation observed in more highly conserved sites (i.e., higher GERP(6) scores and higher proportions).

- Specifically, we see that additional Stop(16) codons, Splice(15) mutations, and non-synonymous (Nonsyn(11)) mutations appear very rare and are likely associated with deleterious effects. Later in the paper these are identified as "loss of function" mutations.
- The addition of Stop(16) codons continues to be relatively rare at most sites across GERP(6) scores. This implies that the addition of the codons likely results in a severe loss of function. As stop(16) codons stop(16) transcription(34) prematurely, these additions will result in macromolecules not being transcribed more generally. This patern is similar amongst the splice(15) mutations which impact the assembly of said macromolecules.
- We see an interesting phenomena with splice(15) variants having higher mutation in less conserved locations.
- The authors note that rare variant loads are similar for synomynous and nonsynomynous locations suggesting weak selective constraints

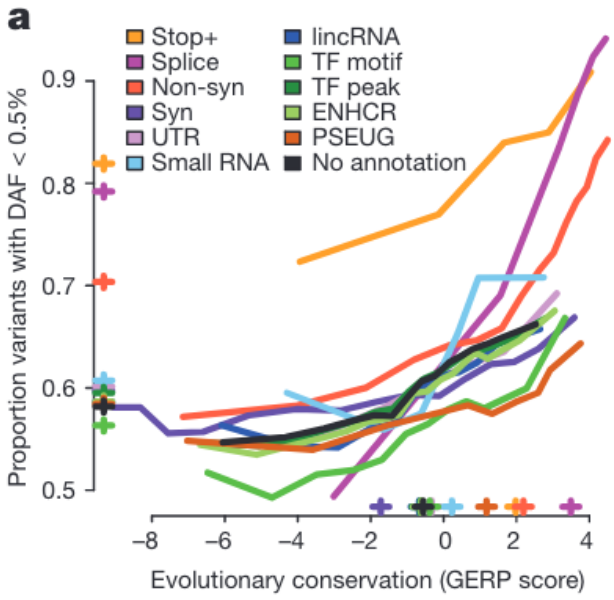

FIGURE 4. The relationaship between evolutionary conservation (measured by GERP(6) score) and rare variant proportion (fraction of all variants weith derived allele(22) frequency (DAF(4)) $< 5\%$)for variants occurring in different functional elements and with different coding consequences. Crosses indicate the average GERP(6) score at variant sites (x axis) and the proportion of rare variants (y axis) in each category.

5.2. **Panel B.**

In Figure 5 examines the CTCF-binding motif(30) within the CTCF-binding peaks. The transcription(34) repressor CTCF has been characterized as playing a vital role in transcription(34) regulation including the recombination of the antibody loci and the regulation of chromatin architecture [4, 5]. Intuitively, we would expect relatively low diversity in this gene as chromatin structural formation is vital for transcription(34) (and cellular generation more generally).

The binding motif(30) is shown in the picto-graphic (called the "logo plot") for the actual nucleotide. In all cases, red represents the "out peak" and blue represents the "in peak" from the Chip-seq (ChipSeq(3)) analysis which is used to map binding sites. Those sites that are located within the peak are likely related to binding and associated with function. The in peak is the mapped functional/ active site of the CTCF gene while the out peak represents the CTCF motif(30) , but not on the CTCF gene. This indicates

that the conservation and lower diversity rates are active site conserving (preserving functionality of the gene).

5.2.1. *Upper panel.*

The y axis again represents the GERP(6) score in the different regions. Higher values represent more likely to be a conserved region as it changes less than expected under a natural substitution model.

5.2.2. *Lower panel.*

The y axis represents the average diversity as defined as the per-nucleotide pairwise distance with higher values representing more differences (i.e., more distance and differences)

5.2.3. *Figure conclusions.*

As suspected, in an important gene we see that those sites associated with function ("in peak") vary less than expected under natural substition as shown by the GERP(6) scores and lower pair-wise nucleotide distances in the lower panel. This is not to say that there isn't a more complex story as there is a hint of degeneracy in position 8.

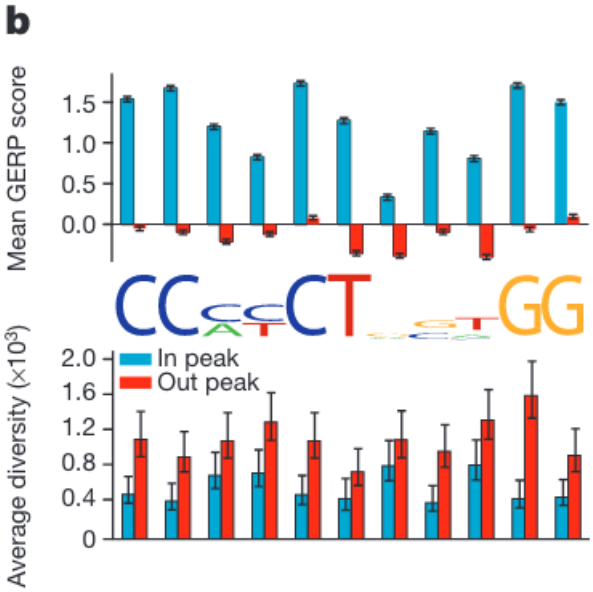

FIGURE 5. Levels of evolutionary conservation (mean GERP(6) score, top) and genetic diversity (per-nucleotide pairwise differences, bottom) for the sequences matching the CTCF-binding motif(30) within CTCF-binding peaks, as identified experimentally by ChIP-seq in the ENCODE project (blue) and in a matched set of motifs outside peaks (red). The logo plot shows the distribution of identified motifs within peaks.

5.3. **Key conclusions.**

There are some highly conserved regions and likely purifying selection is driven by the addition of stop(16) codons, splice(15) mutations, and non-synomynous mutations at these conserved sites due to their lower observed frequencies of occuring.

6. Use of 1000 Genomes Project data in medical genetics

The authors argue that these data can serve as reference data for future GWAS(7) studies. As these data provide a "null model" for rare, low frequency, and common variants, they can provide a background for what to expect in a random sample of the population. This null model is vital in being able to detect the relationship between phenotypic and genotypic differences.

As they hint, the focus likely needs to be on the functionally important polymorphisms (e.g., those sites with high GERP(6) scores) which are likely tied to function.

"Because many variants contribution to disease risk are likely to be segregating at low frequency, we recommend that variant frequency be considered when using the resource to identify pathological candidates."

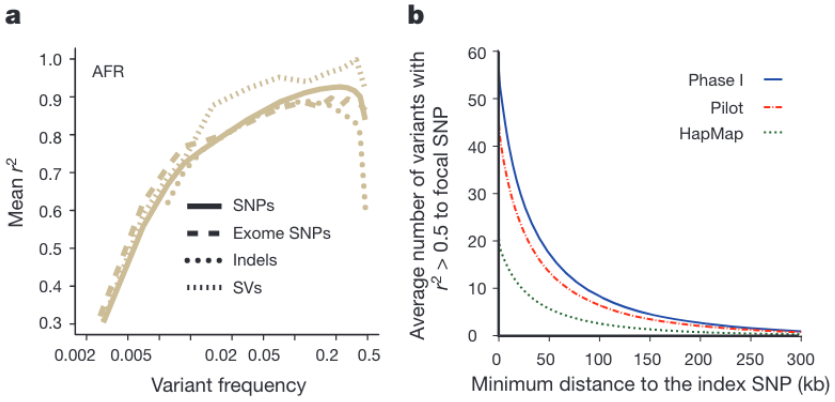

FIGURE 6. a, Accuracy of imputation of genome-wide SNPs, exome SNPs and indels (using sites on the Illumina 1 M array) into ten individuals of African ancestry (three LWK, four Masaai from Kinyawa, Kenya (MKK), two YRI), sequenced to high coverage by an independent technology3. Only indels in regions of high sequence complexity with frequency .1% are analysed. Deletion imputation accuracy estimated by comparison to array data46 (note that this is for a different set of individuals, although with a similar ancestry, but included on the same plot for clarity). Accuracy measured by squared Pearson correlation coefficient between imputed and true dosage across all sites in a frequency range estimated from the 1000 Genomes data. Lines represent whole-genome SNPs (solid), exome SNPs (long dashes), short indels (dotted) and large deletions (short dashes). SV, structural variants. b, The average number of variants in linkage disequilibrium (r2 . 0.5 among EUR) to focal SNPs identified in GWAS47 as a function of distance from the index SNP(13) . Lines indicate the number of HapMap (green), pilot (red) and phase I (blue) variants.

7. Author's discussion

7.1. **Rare variation is likely associated with complex diseases.**

The authors argue that understanding rare variation is important to understanding complexes diseases. A reminder here, rare is not private, rather it is the 1% of the variants.

7.2. **Cost-effective use of combining data from multiple sources.**

They authors argue that their incorportation of multiple types of data has allowed for a cost-effective way of reconstructing haplotypes. They didn't present any evidence of this directly, but rather show that WGS with LD(10) data can allow for reconstruction. They found many variants not in the dense reads of exons (40% not from exon(24) reads). They mention use of CHG array(2) data.

7.3. **Methodological advances.**

They mention that while they did some interesting things (combining reads and variant calls from multiple groups), there are still limitations with long reads, duplications, etc.

7.4. **Local differentiation through purifying selection despite metrics.**

The authors mention that purfying slection at functionally relevant sites can lead to substantial local differentiation despite low F_{ST}. Rare variants tend to be recent and restricted due to geography (or

ancestry). This suggests that local context (and inheritance) may be important for understanding particular disease phenotypes

Glossary

Biallelic(1) Affecting both alleles at a given a location. One allele(22) is taken as a reference allowing for the other allele(22) to be considered the variant allele(22) .
→

CHG array(2) A molecular method for analyzing copy number variations. You can compare quickly DNA from two sources to detect either gains or losses. This is often used in tumor identification and rapid identification of genetic anomalies.
→

ChipSeq(3) Chromatin immunoprecipitation(9) followed by sequencing (ChIP-seq) is extensively used to determine transcription(34) factor binding sites (TFBSs). DNA binding proteins can be cross-linked to the DNA that they are binding and by using an antibody against said proteins, the entire DNA/protein complex can be precipitated out of solution.
→ https://academic.oup.com/bioinformatics/article/29/21/2705/195767

DAF(4) Derived allele(22) frequency which represents the distribution of alleles at a given loci.
→ https://en.wikipedia.org/wiki/Allele_frequency_spectrum

ENHCR(5) Enhancers regulate the transcription(34) of genes by increasing the levels than without them.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445073/

GERP(6) The Genomic Evolutionary Rate Profiling is a statistical approach to quantify the deficit of substitutions in base pairs based on neutral the neutral rate of substitution. The neutral rate of substitution is an estimate of the number of new mutations per each generation multiplied by the probability that the substitution reaches fixations. In cases where neutral mutations exist, the rate of substitution will be equal to the rate of mutations. The GERP(6) Score thus indicates the deviation from the neutral rate of subsitution indicating that positive scores will indicate that fewer substitutions have taken place than would be expected under neutral substitution while negative scores indicate that more substitutions have taken place than would be expected under the neutral model.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1172034/

GWAS(7) Genome-wide association study. These studies examine the association between certain phenotypic outcomes and the genomic variations (generally just SNPs) to understand how the mutations are associated with particular outcomes.
→

IBD(8) Any pair of alleles present in a single inbred individual, we call these alleles inbreeding by descent, if they both derived by the DNA replication of a single allele(22) present in some ancestral population. If the ancestral population has no inbreeding then F = 0.
→

Immunoprecipitation(9) Use of an antibody to precipitate out a particular protein antigen.
→ https://en.wikipedia.org/wiki/Immunoprecipitation(9)

LD(10) Linkage disequilibrium is a measure of nonrandom association between loci. When D=0 the gametic frequencies equal the products of the relevant allele(22) frequencies and are said to be in linkage equilibrium.
→

Nonsyn(11) A non-synonymous mutation is a change in the DNA sequence which results in a change in the encoded amino acid.
→ https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5572-x

PSEUG(12) Pseudogenes are nonfunctional segments of DNA which resemble genes. These could be artifacts from mutations of prior functioning genes.
→ https://en.wikipedia.org/wiki/Pseudogene

SNP(13) Single nucleotide polymorphism(32) , i.e., a single nucleotide substitution at a specific position. Generally associated with a sufficiently large proportion of the population (over 1%). Multiple

SNPs at the same location are alleles.
→ https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

SmallRNA(14) Small RNA sequences are short, normally non-coding regions with a nucleotide length less than 200 bases. RNA silencing has been associated with these sequences.
→ https://en.wikipedia.org/wiki/Small_RNA

Splice(15) Often associated with post-translational changes in either the N-terminal region or a change in the availability of glycosylationsites through a change in the signal peptide. This can lead to secondary changes due to changes in the structure and is associated with a loss of function.
→ https://onlinelibrary.wiley.com/doi/10.1002/prot.10568

Stop(16) A nucleotide substitution leading to the addition of a stop(16) codon (UAA, UAG, UGA) and leads to premature truncation of a protein during transcription(34) . These are associated with a loss of function.
→ https://illustrated-glossary.nejm.org/term/stop(16) -gain_variant

Syn(17) A synonymous mutation is a change in the DNA sequence which does not result in a change in the encoded amino acid.
→ https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5572-x

TFmotif(18) Transcription(34) factor motif(30) is the sequence associated with the proteins which are involved in the transcription(34) of DNA to mRNA.
→ https://en.wikipedia.org/wiki/Transcription_factor

TFpeak(19) Transcription(34) factor motif(30)
→ https://en.wikipedia.org/wiki/Transcription_factor

UTR(20) The untranslated region. These appear on either side of the coding region with the 5' untranslated region called the leader sequence and the 3' untranslated region called the trailer sequence. The role of the untranslated region is not well described and may have a role in transcription(34) regulation and mutations in this region may be associated with increased risk of certain cancers.
→ https://en.wikipedia.org/wiki/Untranslated_region

admixture(21) when two distinct, isolated, or previously isolated genetic lineages mix. This results in the introduction of a new genetic lineage to a population.
→

allele(22) A variation in the same sequence of nucleotides at the same place on a long DNA molecule.
→ https://en.wikipedia.org/wiki/Allele(22)

codominance(23) the condition in which multiple alleles are dominant; the heterozygote expresses phenotypes associated with both alleles.
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

exon(24) The regions between the introns that remain in the fully process RNA.
→

frequency dependent selection(25) A trend in which the fitness of a given genotype is correlated with its prevalence in the population (e.g., if an allele(22) is advantageous when it is rare)
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

heterozygote advantage(26) A trend in which the fitness of a heterozygote is greater than that of either homozygote. Also referred to as overdominance.
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

indel(27) an insertion/ deletion polymorphism(32) . In most cases it is unclear if there was an insertion or a deletion in the ancestral sequence hence the terminology.
→

intron(28) The segments which are eliminated in the RNA transcript.
→

lincRNA(29) Large intergenic non-coding RNA are associated with remodelling of chromatin, genome architecture, RNA stabilisation, and transcription(34) regulation.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5889127/

motif(30) A motif(30) is a nucleotide or amino acid sequence associated with a specific structure and are associated with functionally important sites.
→ https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect08_Bind_Motifs.pdf

negative selection(31) in which an allele(22) is disfavored (also called purifying selection). Random mutations are more likely to be deleterious and immediately removed from the gene pool before they reach appreciable levels. This can also be referred to as background selection.
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

polymorphism(32) genetic differences that are common among organisms in the same species
→

positive selection(33) in which an allele(22) is favoured and is propagated. Positive selection(33) is understood to be the primary mechanism of adaptation (and is often more conspicuous).
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

transcription(34) the process by which a sequence of nucleotides present in one DNA strand of a gene is copied into the nucleotides of an RNA molecule.
→

References

[1] T. 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, Nov. 2021, doi: 10.1038/nature11632.

[2] I. Cvijović, B. H. Good, and M. M. Desai, "The Effect of Strong Purifying Selection on Genetic Diversity," *Genetics*, vol. 209, no. 4, Aug. 2018, doi: 10.1534/genetics.118.301058.

[3] J. J. Vitti, S. R. Grossman, and P. C. Sabeti, "Detecting natural selection in genomic data," *Annu. Rev. Genetics*, vol. 47, 2013, doi: 10.1146/annurev-genet-111212-133526.

[4] G. N. Filippova, S. Fagerlie, et al., "An Exceptionally Conserved Transcriptional Repressor, CTCF, Employs Different Combinations of Zinc Fingers To Bind Diverged Promoter Sequences of Avian and Mammalian c-myc Oncogenes," *Mol. Cellular Biol.*, vol. 16, no. 6, Jun. 1996, doi: 10.1128/MCB.16.6.2802.

[5] G. M. Cooper, E. A. Stone, et al., "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res.*, vol. 15, no. 7, Jul. 2005, doi: 10.1101/gr.3577405.

Department of Biology, Wake Forest University, Winston Salem, NC 27101
Email address: dewime23@wfu.edu
URL: www.michaeldewittjr.com