# Glossary of Genomics Terms

THE FOLLOWING GLOSSARY REPRESENTS A COMPILATION of definitions based in part on published information from *Archives of Ophthalmology*, *JAMA*, *New England Journal of Medicine*, and Ontology for Biomedical Investigations.[1-4]

**Allele:** Alternative form of a gene or DNA sequence. Variations in clinical traits and phenotypes are allelic if they arise from the same gene sequence or locus and nonallelic if they arise from different gene sequences of different loci.

**Alternative splicing:** Use of different exons in formation of mRNA from initially identical transcripts. This results in the generation of related proteins from one gene, often in a tissue or developmental stage–specific manner.

**Analytical validity:** The likelihood that a test result is correct, ie, a specific variant said to be present is present or said to be absent is absent.

**Array:** See **Microarray**.

**Attributable risk:** The difference in incidence of disease in an exposed vs unexposed population; in genetics the exposure can be the presence of a specific genetic variation in the genome.

**Bacteriophage:** Viruses whose hosts are bacterial cells.

**Base pair (bp):** Two complementary nucleotides that are paired in double-stranded DNA. Adenosine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). A bp is also used as a physical distance of length of a sequence of nucleotides, eg, 20 bp is a chain of DNA composed of 20 nucleotides.

**Call rate:** The rate at which assignment of a specific nucleotide base (A,G,C,T) is made at specific positions in the genome during genotyping or sequencing.

**Candidate gene:** A gene believed to influence expression of complex phenotypes due to known biological and/or physiological properties of its products, or to its location near a region of association or linkage.

**cDNA (complementary DNA):** A DNA copy of the messenger RNA (mRNA) transcribed from a gene. The cDNA is made from the mRNA using the enzyme reverse transcriptase.

**Clinical utility:** The degree to which a test result guides clinical management and improves patient outcomes.

**Clinical validity:** The likelihood that a test result correctly predicts presence or absence of disease or risk of disease.

**Codon:** Three bases in a DNA or RNA sequence that specify a single amino acid.

**Copy number variants:** Stretches of genomic sequence of roughly 1000 base pairs (1kb) to 3 million base pairs (3 Mb) in size that are deleted or are duplicated in varying numbers.

**Comparative genomic hybridization (CGH)** *also* **array CGH:** Technology wherein a DNA test sample is competitively hybridized with a reference sample of DNA of known sequence to a DNA microarray, used to detect copy number changes in the test sample.

**Complex trait:** A trait that has a genetic component that does not follow strict mendelian inheritance. It may involve the interaction of 2 or more genes or gene-environment interactions.

**Coverage:** The number of times a portion of the genome is sequenced in a sequencing reaction. Often expressed as "depth of coverage" and numerically as 1X, 2X, 3X, etc.

**Cytogenomic analysis:** Technologies that assess the presence of copy number variants at locations throughout the genome, one example of which is comparative genomic hybridization.

**Denaturing high-performance liquid chromatography (DHPLC):** A high-performance liquid chromatography instrument uses temperature-dependent separation of DNA containing mismatched base pairs from PCR-amplified DNA fragments for chromatographic mutation analysis.

**DNA barcoding:** A method that uses a short genetic marker in a DNA sequence to identify it as belonging to a particular species or group of otherwise-related sequences.

**Environmental gene tag:** Short sequences of DNA that contain bacterial genes in whole or part that can be used to aid in identification of related genetic material.

**Exome:** The entire portion of the genome consisting of protein-coding sequences (as opposed to introns or noncoding DNA between genes).

**Exon:** Any segment of a gene that is represented in the mature messenger RNA (mRNA) product.

**Frame shift mutation:** Any mutation that disrupts the normal sequence of triplets causing a new sequence to be created that codes for different amino acids. Frame shift mutations are usually caused by an insertion or deletion of DNA and typically eventually produce a premature stop codon.

**GC content:** The percentage of nucleotides in a DNA sequence that are either guanine (G) or cytosine (C).

**Genetic heterogeneity:** A common phenotype caused by more than 1 gene.

**Genetic Information Nondiscrimination Act (GINA):** US federal law passed in 2008 prohibiting the use of genetic information for decisions regarding employment or health insurance.

**Genome:** The sum total of the genetic material of a cell or an organism.

**Genome annotation:** Attachment of biological information to DNA sequence data.

©2013 American Medical Association. All rights reserved.



Downloaded From: http://jama.jamanetwork.com/ on 07/24/2017

**Genome-wide analysis:** A genetic study evaluating the potential linkage of genetic markers located throughout the genome to a specific trait. This approach has been used for mendelian disorders as well as complex traits (genome-wide association study [GWAS]).

**Genomic inflation factor:** A mathematical term from genetic epidemiology used to control for population stratification in GWAS.

**Genomic medicine:** A term used to describe medical advances and approaches based on human genomic information, sometimes referred to as personalized or precision medicine.

**Genomics:** The study of genes and their function.

**Genotype:** The specific set of 2 alleles inherited at a genetic locus.

**Haplotype:** The combination of linked marker alleles (may be polymorphisms or mutations) for a given region of DNA on a single chromosome.

**HapMap:** The International HapMap Project developed a haplotype map of the human genome, the HapMap, that describes the common patterns of human DNA sequence variation. The HapMap is a key resource for finding genes affecting health, disease, and responses to drugs and environmental factors. The first release of the HapMap was made in 2005.

**Heterologous expression:** A research technique that causes a protein to be produced in a cell that does not normally make (ie, express) that protein.

**Heterotetrameric, homotetrameric, and heteromultimeric ion channels:** Ion channels made up of different combinations of protein subunits; 4 different subunits (heterotetrameric), 4 of the same subunit (homotetrameric), and 2 or more different subunits (heteromultimeric).

**Heterozygous (heterozyosity):** Having 2 unlike alleles at a particular locus.

**Homozygous (homozygosity):** Having 2 like or identical alleles at a particular locus in a diploid genome.

**Human Genome Project:** Collective name for several projects begun in 1986 by the US Department of Energy (DOE) to create an ordered set of DNA segments from known chromosomal locations, develop new computational methods for analyzing genetic map and DNA sequence data, and develop new techniques and instruments for detecting and analyzing DNA. The joint national effort, led by DOE and the National Institutes of Health, was known as the Human Genome Project. The first draft of the human genome DNA sequence, produced by the efforts of the Human Genome Project, was completed in 2001. The Human Genome Project officially ended in April 2003.

**Hybridization:** The bonding of single-stranded DNA or RNA into double-stranded DNA or RNA. The ability of complementary stretches of DNA or RNA to hybridize with each other is dependent on the base-pair sequence.

**Identity by descent (IBD):** The property of 2 or more alleles that are identical to an ancestral allele, used in gene association studies.

**Imputation:** A statistical method for inferring genotypes that are not directly measured.

**Intron:** A segment of DNA that is transcribed into RNA but is ultimately removed from the transcript by splicing together the sequences on either side (exons) to produce messenger RNA (mRNA).

**Kilobase (kb):** One thousand base pairs of DNA or RNA.

**Library:** A complete set of clones that contains all the genetic material from an organism, tissue, or specific cell type at a specific stage of development.

**Linkage:** Two loci (genes or other designated DNA sequence) that reside close enough to each other that recombination (crossing over) rarely occurs between them. Alleles at the 2 loci do not assort independently at meiosis but are likely to be inherited together.

**Linkage disequilibrium (LD):** Refers to alleles at loci close enough together that they remain inherited together through many generations because their extreme close proximity makes recombination (crossing over) between them highly unlikely.

**Locus (*plural* loci):** The physical site on a chromosome occupied by a particular gene or other identifiable DNA sequence characteristic.

**Megabase:** One million base pairs.

**Mendelian disorder (single-gene disorder):** A trait or disease that follows the patterns of inheritance that suggest the trait or disease is determined by a gene at a single locus.

**Metagenomics:** Study of a collection of genetic material (genomes) from a mixed community of organisms. Metagenomics usually refers to the study of microbial communities.

**Methylation:** Covalent attachment of methyl groups to DNA, usually at cytosine bases. Methylation can reduce transcription from a gene and is a mechanism in X-chromosome inactivation and imprinting.

**Microarray:** A technology used to study many genes simultaneously, usually consisting of an ordered microscopic pattern of known nucleic acid sequences on a glass slide. In a common type of microarray, a sample of DNA or RNA is added to the slide and sequence-dependent binding is measured using sensitive fluorescent detection methods.

**Minor allele:** The allele of a biallelic polymorphism that is less frequent in the study population. Minor allele frequency refers to the proportion of the less common of 2 alleles in a population (with 2 alleles carried by each person at each autosomal locus) ranging from less than 1% to less than 50%.

**Missense mutation:** A mutation that is typically the change of a single nucleotide that results in the substitution of one amino acid for another in the final gene product.

**Mutation:** Any alteration of a gene or genetic material from its natural state. Generally, mutations refer to changes that alter the gene in a negative sense causing the protein product of the gene to have an altered function.

**Next generation/high-throughput sequencing:** DNA sequencing technology that permits rapid sequencing of large portions of the genome; so called because it vastly increases the throughput over classic Sanger sequencing.

**Nonsense mutation:** Any mutation that results directly in the formation of a stop codon.



**©2013 American Medical Association. All rights reserved.**

**Downloaded From: http://jama.jamanetwork.com/ on 07/24/2017**

**Nonsynonymous variant:** A polymorphism that results in a change in the amino acid sequence of a protein (and therefore may affect the function of the protein).
**Nucleotide:** The combination of a nitrogen-containing base, a 5-carbon sugar, and phosphate group forming the A, G, C, T of the sequence of DNA (DNA), for example.
**Oncogene:** A gene, 1 or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.
**Patch-clamp technique:** A laboratory technique in electrophysiology that allows the study of single or multiple ion channels in cells.
**Penetrance:** The proportion of individuals of a given genotype who show any evidence of the associated phenotype.
**Pharmacogenetic polymorphism:** Genetic variants that alter the way an individual metabolizes or responds to a specific medication.
**Pharmacogenomics:** Study of genes related to genetic controlled variation in drug responses.
**Phenotype:** The total observable nature of an individual, resulting from interaction of the genotype with the environment.
**Plasmid:** Circular extrachromosomal DNA molecules in bacteria that can independently reproduce. Plasmids can be used as vectors in recombinant DNA research, and they can contain genes important to bacterial virulence such as antibiotic resistance in nature.
**Polymerase chain reaction (PCR):** A procedure in which segments of DNA (including DNA copies of RNA) can be amplified using flanking oligonucleotides called primers and repeated cycles of replication by DNA polymerase.
**Polymorphism:** Difference in DNA sequence among individuals that may underlie differences in health. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic linkage analysis. The vast majority of DNA polymorphisms are benign and not associated with a detectable phenotype.
**Population stratification (*also* population structure):** A form of confounding in genetic association studies caused by genetic differences between cases and controls unrelated to disease but due to sampling them from populations of different ancestries.
**Proband:** The affected person whose disorder, or concern about a disorder, brings a family or pedigree to be genetically evaluated.
**Promoter:** The sequence of nucleotides located 5' to the coding sequence of a gene that determines the site for binding of RNA polymerase and the initiation of transcription. More than 1 promoter may be present in a gene and may give rise to different versions of the protein.
**Prophage:** The genome of a bacteriophage when it is integrated into the host bacterial genome or a plasmid.
**Pyrosequencing:** A method of determining the ordering of nucleotide bases in a DNA molecule by measuring the synthesis of the complementary DNA strand.
**Quantitative PCR:** A PCR-based laboratory technique that allows the accurate measurement of the amount of specific nucleic acids (usually RNA) in a sample.
**Read:** A discrete segment of sequence information generated by a sequencing instrument; read length refers to the number of nucleotides in the segment.
**Recombination:** The formation of a new set of alleles on a single chromosome that is not the same as either parent owing to a crossover during meiosis.
**Restriction fragment-length polymorphism (RFLP):** A type of polymorphism that results from variation in the DNA sequence recognized by restriction enzymes. RFLPs can be used in linkage and gene association studies of traits and diseases.
**Single-nucleotide polymorphism (SNP):** DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. SNPs are the most abundant variant in the human genome and are the most common source of genetic variation, with more than 10 million SNPs present in the human genome, representing a density of 1 SNP for approximately every 100 bases.
**Stop codon (termination codon):** The DNA triplet that causes translation to end when it is found in messenger RNA (mRNA). The DNA stop codons are TAG, TAA, and TGA.
**Tag SNP:** A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms.
**Translocation:** A chromosomal segment that has been broken off and reinserted in a different place in the genome.
**Transversion:** The substitution of a purine for a pyrimidine nucleotide or vice versa (eg, an A for a C or T) in a DNA sequence.
**Uniparental disomy:** The inheritance of both parental copies of a chromosome from one parent and no homologous chromosome from the other parent. The resulting offspring could be affected with a recessive disease if the parent contributing both copies is a carrier.
**Variant of unknown significance (VUS):** Genetic variant that cannot be definitively determined to be associated with a specific phenotype.

## REFERENCES

**1.** Wiggs JL, Nemesure B. Glossary of genetic terms. *Arch Ophthalmol*. 2007; 125:E1-E7.
**2.** Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299(11):1335-1344.
**3.** Feero WG, Guttmacher AE, Collins FS. Genomic medicine: an updated primer. *N Engl J Med*. 2010;362(21):2001-2011.
**4.** Ontology for Biomedical Investigations. http://bioportal.bioontology.org/ontologies/44899/?p=terms&conceptid=obi%3AOBI_0001133.

## OTHER RESOURCES

**National Human Genome Research Institute (National Institutes of Health):** http://www.genome.gov/Glossary/index.cfm?showall=true&textonly=true

**Genetics Home Reference:** http://ghr.nlm.nih.gov/glossary

**Human Genome Project Information:** http://www.ornl.gov/sci/techresources/Human_Genome/glossary/index.shtml

©2013 American Medical Association. All rights reserved.



Downloaded From: http://jama.jamanetwork.com/ on 07/24/2017