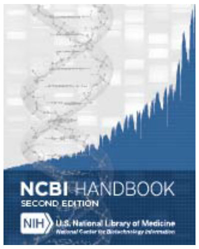
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Glossary.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Glossary

accession number — The accession number is a unique identifier assigned to a record in sequence databases such as GenBank. Several NCBI databases use the format [alphabetical prefix][series of digits]. A change in the record in some databases (e.g. GenBank) is tracked by an integer extension of the accession number, an Accession.version identifier. The initial version of a sequence has the extension ".1". When a change is made to a sequence in a GenBank record, the version extension of the Accession.version identifier is incremented. For the sequence NM_000245.3, ".3" indicates that the record has been updated twice. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the original Accession.version identifiers.

AGP file — AGP ('A Golden Path') file is used to describe the instructions for building a contig, scaffold, or chromosome sequence. This file specifies the order, orientation, and switch points for each genomic sequence.

alignment — An alignment is a representation of the similarity between 2 nucleotide or protein sequences. In the case of protein sequences, the amino acids derived from ancestral sequences are taken into consideration in the alignment to account for conserved sequence. A pairwise alignment involves 2 sequences and a multiple alignment involves 3 or more sequences. A global alignment involves aligning the entire sequence whereas a local alignment involves aligning subsequences. The optimum alignment is determined by highest score for a given system. In a structural alignment, 3-dimensional structures of proteins under consideration are superimposed (Koonin and Galperin 2003a).

allele — One of the variant forms of a gene at a particular locus on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one). When "genes" are considered simply as segments of a nucleotide sequence, allele refers to each of the possible alternative nucleotides at a specific position in the sequence. For example, a CT polymorphism such as CCT[C/T]CCAT would have two alleles: C and T.

allele frequency — The proportion of a specific gene variant among all copies of the gene in the population.

alternate locus — A sequence that provides an alternate representation of a locus. Alternate loci are collected into additional assembly units (i.e. not in the primary assembly).

Alu — The *Alu* repeat family comprises short interspersed elements (SINES) present in multiple copies in the genomes of humans and other primates. The *Alu* sequence is approximately 300 bp in length and is found commonly in introns, 3′ untranslated regions of genes, and intergenic genomic regions. They are mobile elements and are present in the human genome in extremely high copy number. Almost one million copies of the *Alu* sequence are estimated to be present, making it the most abundant mobile element. The *Alu* sequence is so named because of the presence of a recognition site for the *Alu*I endonuclease in the middle of the *Alu* sequence. Because of the widespread occurrence of the *Alu* repeat in the genome, the *Alu* sequence is used as a universal primer for PCR in animal cell lines; it binds in both forward and reverse directions.

amino acid — Basic building block molecules of peptides and proteins. The sequence of amino acids in a protein is determined by the RNA codon sequence.

anchor sequence — Anchor sequences are molecular markers that are unique loci in genetic linkage maps of multiple species and are used in comparative genomics for cross-species mapping along co-linear genomic regions.

API

Application Programming Interface. An API is a set of routines, data structures, variables, constants and/or classes for building software applications. APIs define how software components communicate with one another. For instance, for computers running a graphical user interface, an API manages an application's windows, icons, menus, and dialog boxes.

ASN.1

Abstract Syntax Notation 1 is an international standard data-representation format used to achieve interoperability between computer platforms. It allows for the reliable exchange of data in terms of structure and content by computer and software systems of all types.

assembly

A set of chromosomes, unlocalized and unplaced (random) sequences and alternate loci used to represent an organism's genome. Assemblies are constructed from one or more assembly units. Most current assemblies are a haploid representation of an organism's genome, although some loci may be represented more than once (see alternate locus). This representation may be obtained from a single individual (e.g. chimp or mouse) or multiple individuals (e.g. human reference assembly). Except in cases of organisms which have been bred to homozygosity, the haploid assembly does not typically represent a single haplotype, but rather a mixture of haplotypes. A diploid genome assembly is a chromosome assembly that is available for both sets of an individual's chromosomes. It is anticipated that a diploid genome assembly is representing the genome of an individual. Therefore it is not anticipated that alternate loci will be defined for this assembly, although it is possible that unlocalized or unplaced sequences could be part of the assembly. An assembly is constructed from one or more assembly units.

assembly release

Release of a genome assembly. A major release is any update that changes the sequence and/or changes the chromosome coordinate system defined in the primary assembly. A minor release is an update that does not change the coordinate system, but may add or modify information. Such events include addition of genome patches, assignment of unplaced sequences to a chromosome, or new placements for alternate loci.

assembly unit

The collection of sequences used to define discrete parts of an assembly. All assemblies must contain one assembly unit that represents the "primary assembly".

asserted position

A statement (assertion) based on experimental evidence that a variant is located at a particular position. Since asserted positions are based on experimental evidence, they cannot be seen as a conformation of the variant's position even if there are multiple claims by different submitters of a specific position for a particular variant.

NCBI does not independently verify assertions and cannot endorse their accuracy.

Note: Submissions based on asserted positions should reference a sequence accession that is part of an assembly represented in the NCBI Assembly Resource. If no assembly is available however, the reference sequence can be an INSDC sequence accession. If a submission asserts a position for a variation using an accession that cannot be aligned to an assembly, the rs for that variation cannot be annotated to the human assembly, and therefore will not appear on maps or graphic representations of the assembly.

BAC

Bacterial Artificial Chromosome. The BAC cloning system is based on a bacterial plasmid vector which is capable of carrying a large segment of genomic DNA (100–300 bp) for cloning in bacteria. BACs are used in the construction of complex genomic libraries because of their high cloning efficiency and stability of the cloned DNA.

backend

With reference to web applications, the backend refers to the components (server, application, and database) not directly accessed by the user.

BAF

B-Allele Frequency

base sequence

The sequence of purines and pyrimidines in nucleic acids and polynucleotides. It is also called nucleotide sequence.

bioinformatics

Bioinformatics is an interdisciplinary field that applies computational approaches for the collection, storage, manipulation, and analysis of biological data including large datasets, to make biological discoveries or predictions. At a minimum, it encompasses computer science, biology, genetics, genomics, statistics, mathematics and engineering to interpret biological data. It is closely related to computational biology.

BLAST

Basic Local Alignment Search Tool (Altschul et al. 1990). A sequence comparison algorithm that is used to search sequence databases for optimal local alignments to a query. See the BLAST chapter.

BLASTN

nucleotide–nucleotide BLAST. BLASTN takes nucleotide sequences and compares them against the NCBI nucleotide databases.

BLASTP

protein–protein BLAST. BLASTP takes protein sequences and compares them against the NCBI Protein databases.

BLASTX

BLASTX is an application that searches a nucleotide query against a protein database, dynamically translating the query in all six frames.

BLOB

Binary Large OBject. BLOB refers to a large piece of data, such as a bitmap. A BLOB is characterized by large field values, an unpredictable table size, and data that are formless from the perspective of a program. It is also a keyword designating the BLOB structure, which contains information about a block of data.

Boolean

This term refers to binary algebra that uses the logical operators AND, OR, XOR, and NOT; the outcomes consist of logical values (either TRUE or FALSE).

byte

In computer terms, a unit of storage that is equal to 8 bits.

CD

Conserved Domain. CD refers to a domain (a distinct functional and/or structural unit of a protein) that has been conserved during evolution. During evolution, changes at specific positions of an amino acid sequence in the protein may have occurred in a way that preserve the physico-chemical properties of the original residues, and hence the structural and/or functional properties of that region of the protein.

CDD

Conserved Domain Database. This database is a collection of sequence alignments and profiles representing protein domains conserved during molecular evolution.

cDNA

complementary DNA. A DNA sequence obtained by reverse transcription of a messenger RNA (mRNA) sequence.

CDS

Coding region, coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the sequence (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

CGI

Common Gateway Interface. A mechanism that allows a Web server to run a program or script on the server and send the output to a Web browser.

chip

See DNA chip.

chromosome

The threadlike structure comprised of DNA and protein contained with the nucleus of eukaryotic cells and containing the hereditary material or genes; in prokaryotes, the circular DNA that carries the genetic information.

clinical assertion

A statement (assertion) based on experimental evidence that a variant has a clinical phenotype. Clinical assertions submitted with a variant may or may not be specific as to the nature of the associated phenotype. Since clinical assertions are based on experimental evidence, they cannot be seen as a conformation of a clinical phenotype, even if there are multiple claims by different submitters of a specific clinical phenotype for a particular variant.

Clinical assertions can fall into one of the following categories:

- Pathogenic
- Probably Pathogenic
- Probably Non-pathogenic
- Non-pathogenic [benign]
- Affecting Drug Response
- Affecting Histocompatability
- Unknown
- Untested
- Other

As assertion categories may change, see ClinVar for up-to-date assertion definitions.

Example: For rs report for rs328, the asserted clinical significance for the cluster is clearly stated at the top of the report as well as in the "Allele" subsection.

For more information regarding clinical assertions, see the clinvar.vcf.gz section of "Human Variation Sets in VCF Format" or the FAQ for NCBI Variation Resources.

Note: NCBI does not independently verify assertions and cannot endorse their accuracy. Information obtained through this resource is not a substitute for professional genetic counseling and is not intended for use as the basis of medical decision making.

CLOB

Character Large OBject

clone

A clone can be considered a self-replicating system containing a DNA fragment of interest.

cloning vector

A small DNA molecule which is capable of autonomous replication within a host cell and is used to carry a fragment of genomic DNA or cDNA to be cloned; usually a bacterial plasmid or modified bacteriophage genome.

cluster

A group that is created based on certain criteria. For example, a gene cluster may include a set of genes whose expression profiles are found to be similar according to certain criteria, or a cluster may refer to a group of clones that are related to each other by homology.

CMS

Content Management System

codon

Sequence of three nucleotides in DNA or mRNA that specifies a particular amino acid during protein synthesis; also called a triplet. Of the 64 possible codons, 3 are stop codons, which do not specify amino acids.

coding region

It is the sequence of DNA that is translated into protein and includes an initiation codon and a termination codon.

complementary DNA

See cDNA.

computational biology

Computational biology involves the development and application of data-analytical and theoretical methods, algorithms, mathematical modeling and computational simulation techniques to the understanding of biological systems and to make predictions and discoveries from biological data, including large datasets. The field has its origins in computer science, applied mathematics, statistics, biophysics, genomics, molecular biology, and many areas of biology. It is closely related to bioinformatics.

consensus sequence

A representative or most typical nucleotide or amino acid sequence in which each nucleotide or amino acid is most often found at its respective position in the group of related sequences.

conserved domains

A conserved domain of a protein is a discrete three-dimensional independently folding structure that is comprised of one or more protein sequence motifs. Protein sequence motifs are conserved amino acid sequences that are a combination of secondary structures (example, helix-loop-helix) which have been shown to be important for protein function (Koonin and Galperin 2003b).

contig

A contiguous sequence generated from determining the non-redundant path along an ordered set of component sequences. A contig should contain no gaps.

CSS

Cascading Style Sheets (CSS) specify the formatting details that control the presentation and layout of HTML and XML elements. CSS can be used for describing the formatting behavior and text decoration of simply structured XML documents but cannot display structure that varies from the structure of the source data.

Cubby

A tool of Entrez, the Cubby was used to store search strategies that could be updated as well as LinkOut preferences to specify which LinkOut providers should be displayed in PubMed, and change the default document delivery service. It has been superceded by MyNCBI.

CUI

Concept Unique Identifier

cytogenetics

A sub discipline of genetics that deals with cytological and molecular analysis of chromosomes—their cellular location, structure, function, and abnormalities.

DAC

Data Access Committee

daemon

A computer program that runs as a background process or service and is not controlled by the user.

DAR

Data Access Request

database

Store of a set of logically related data or collection of files amenable to retrieval by scripts or computer.

dataset

Permanent store of an organized collection of data, for sharing, redistribution, processing, and analysis.

dbGSS

Genome Survey Sequences Database, a division of GenBank for genome sequences.

DDBJ

DNA Data Bank of Japan, a DNA nucleotide sequence collection center and member of INSDC.

DDD

Digital Differential Display, a feature of Unigene that allows analysis of EST expression profiles.

deletion variant

Type of mutation involving the removal of a single nucleotide or segment of DNA.

deoxyribonucleic acid

See DNA.

digital differential display

See DDD.

DNA

Deoxyribonucleic acid is the chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms. DNA is composed of two anti-parallel strands, each a linear polymer of nucleotides. Each nucleotide has a phosphate group linked by a phosphoester bond to a pentose (a five-carbon sugar molecule, deoxyribose), that in turn is linked to one of four organic bases, adenine, guanine, cytosine, or thymine, abbreviated A, G, C, and T, respectively. The bases are of two types: purines, which have two rings and are slightly larger (A and G); and pyrimidines, which have only one ring (C and T). Each nucleotide is joined to the next nucleotide in the chain by a covalent phosphodiester bond between the 5′ carbon of one deoxyribose group and the 3′ carbon of the next. DNA is a helical molecule with the sugar–phosphate backbone on the outside and the nucleotides extending toward the central axis. There is specific base-pairing between the bases on opposite strands in such a way that A always pairs with T and G always pairs with C.

DNA chip

A DNA chip (also referred to as a DNA microarray) is an organized arrangement of DNA sequences on a solid surface in a 2-dimensional (2D) or 3D manner, either covalently or non-covalently bound to the surface. Arrays contain oligonucleotide probes or short nucleotide “known” sequences that can be used to hybridize to sequences in sample for various applications such as measuring the level of gene expression or identifying a particular mutation of interest.

DOI

Digital Object Identifier, an international standard for persistent, actionable, interoperable identifiers that can be applied to objects such as publications.

DTD

Document Type Definition. The DTD is an optional part of the prolog of an XML document that defines the rules of the document. It sets constraints for an XML document by specifying which elements are present in the document and the relationships between elements, e.g., which tags can contain other tags, the number and sequence of the tags, and attributes of the tags. The DTD helps to validate the data when the receiving application does not have a built-in description of the incoming data.

DUC

Data Use Certification

E-utilities

Structured interface to the NCBI Entrez query and database system via 9 server-side programs: EInfo (database statistics), ESearch (text searches), EPost (UID uploads), ESummary (document summary downloads), EFetch (data record downloads), ELink (Entrez links), EGQuery (global query), ESpell (spelling suggestions), ECitMatch (batch citation searching in PubMed).

EMBL — European Molecular Biology Laboratory

ENA — European Nucleotide Archive at European Molecular Biology Laboratory (EMBL)

end sequence — A sequence obtained from the unidirectional sequencing of a genomic clone insert. A set of paired end sequences can be generated if the insert is sequenced from either end.

Entrez — Entrez is a retrieval system at NCBI for searching several linked databases, such as PubMed, GenBank, and PMC. See the Entrez chapter.

epigenomics — The study of changes in the expression or repression of genes by epigenetic mechanisms such as DNA methylation or histone modification that are not a result of changes in the DNA base sequence.

eQTL — expression Quantitative Trait Loci

EST — Expressed Sequence Tag. ESTs are short (usually approximately 300–500 base pairs), single-pass sequence reads from cDNA. Typically, they are produced in large batches. They represent the genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library. They are useful in identifying full-length genes and in mapping.

eukaryotic — Referring to organisms with cells having a true nucleus bounded by a nuclear membrane.

exon — Refers to the portion of a gene that encodes for a part of that gene's mRNA. A gene may comprise many exons, some of which may include only protein-coding sequence; however, an exon may also include 5' or 3' untranslated sequence. Each exon codes for a specific portion of the complete protein. In some species (including humans), a gene's exons are separated by long regions of DNA (called introns or sometimes “junk DNA”) that often have no apparent function but have been shown to encode small untranslated RNAs or regulatory information.

expressed sequence tag — See EST.

FASTA — The first widely used algorithm for similarity searching of protein and DNA sequence databases. The program looks for optimal local alignments by scanning the sequence for small matches called “words”. Initially, the scores of segments in which there are multiple word hits are calculated (“init1”). Later, the scores of several segments may be summed to generate an “initn” score. An optimized alignment that includes gaps is shown in the output as “opt”. The sensitivity and speed of the search are inversely related and controlled by the “k-tup” variable, which specifies the size of a “word” (Pearson and Lipman 1988). Also refers to a format for a nucleic acid or protein sequence.

FLAN — FLu Annotation

FlyBase — FlyBase is the primary database of genetic and genomic data for the insect family Drosophilidae.

frameshift — A mutation in which the number of nucleotides inserted or deleted from a protein coding sequence of DNA is not a multiple of 3, which results in a shift in the codon reading frame, creating an altered protein product.

frontend — With reference to web applications, the frontend refers to the interface which is directly accessible to the user through which other components such as databases and servers can be accessed.

FTP — File Transfer Protocol. A method of retrieving files over a network directly to the user's computer or to his/her home directory using a set of protocols that govern how the data are to be transported.

gap — A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acids is also penalized in the scoring of an alignment.

GB — Gigabytes; 10^9 bytes.

Gbps — Gigabits per second. Refers to the speed of data transfer.

gene expression — It is the process by which a gene is modulated for transcription to mRNA and translation to a protein. It can be measured by the levels of mRNA, protein, or observable phenotype of the cell.

gene frequency — See allele frequency.

gene trap — Gene trapping is a technique for determining gene function whereby specialized vector sequences are randomly inserted into the genome and can be used to tag genes, allowing for genetic or phenotypic approaches to study the effect of the mutation.

genetic code — The instructions in a gene that tell the cell how to make a specific protein. A, T, G, and C are the “letters” of the DNA code; they stand for the chemicals adenine, thymine, guanine, and cytosine, respectively, that make up the nucleotide bases of DNA. Each gene's code combines the four chemicals in various ways to spell out three-letter “words” that specify which amino acid is needed at every position for making a protein.

genetic recombination — Genetic recombination is the process by which DNA is broken and rejoined resulting in new arrangements, such as by crossing over of chromosomes during meiosis or by chromosomal exchange during genetic conjugation, transduction, or transformation.

genetic testing — The analysis of an individual’s genome for determining the presence of a mutation, carrier status of a mutation, disease risk, or relationship to other individuals.

genome — The genome is the complete genetic material of an organism. For eukaryotic organisms, it is the DNA in all chromosomes and in mitochondria or chloroplasts; for procaryotes, it includes the circular double-stranded DNA molecule. For viruses, it comprises DNA or RNA.

genome assembly — See assembly.

genome library — A DNA library which includes the complete sequences from the genome of an organism, i.e., introns and exons.

genomics — A field of study in genetics that applies molecular tools such as recombinant DNA technology and high-throughput sequencing, and bioinformatics approaches such as genome alignment and assembly towards the analysis of genome structure and function.

genotype — The genetic identity of an individual that does not show as outward characteristics. The genotype refers to the pair of alleles for a given region of the genome that an individual carries.

GFF — General Feature Format; it is used for the annotation of biological sequences

gnomon — Gene model prediction program. See http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml

haplotype — A haplotype is a set of DNA variants, SNPs, or other polymorphisms that are closely located on the same chromosome and are thus inherited together.

HapMap — Haplotype map, a now retired tool for finding genes and genetic variations that affect health and disease.

HGNC — HUGO Gene Nomenclature Committee

HGP — Human Genome Project

HIPAA — Health Insurance Portability and Accountability Act of 1996

HLA — Human Leukocyte Antigen

HMM — Hidden Markov Model

HMP — Human Microbiome Project

HomoloGene — HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of several completely sequenced eukaryotic genomes.

homology — See homologous.

homologous — The term refers to similarity attributable to descent from a common ancestor. Homologous chromosomes are members of a pair of essentially identical chromosomes, each derived from one parent. They have the same or allelic genes with genetic loci arranged in the same order. Homologous chromosomes synapse during meiosis.

HPO — Human Phenotype Ontology

HUGO — Human Genome Organization

HUP — Hold-until-published

initiation codon — The canonical AUG codon or any alternative non-canonical start codon in the messenger RNA which serves as the recognition codon for binding of N-formylmethionyl transfer RNA ($\text{tRNA}^{\text{fmet}}$) and addition of the first methionine during the initiation of protein translation.

INSDC — International Nucleotide Sequence Database Collaboration

insert sequence — A piece of DNA inserted into a cloning vector by recombinant DNA techniques. For genomic cloning vectors, insert sequences typically range in size from 10's of kilobases (cosmids, fosmids), to 100's of kilobases (BAC, PACs) up to ~2MB (YACs).

intergenic — Regions of the genome that lie between genes and have no known function. Intergenic DNA is mainly made up of 2 types of repeated sequences: interspersed repeats and tandemly repeated DNA.

intervening sequence — See intron.

intron — Non-coding region of the DNA that gets transcribed in the primary messenger RNA but the sequence is removed from the mature RNA transcript when exons are spliced together.

IRB — Internal Review Board

ISSN — An ISSN is an 8-digit code used to identify newspapers, journals, magazines, periodicals, and electronic and print media. See http://www.issn.org

JATS — Journal Article Tag Suite, a NISO DTD Standard.

JPEG — Joint Photographic Experts Group; suffix for digital image file format.

locus — The position on a chromosome where a gene is located.

long repeat — See long terminal repeat.

long terminal repeat — Retroviruses integrate a reverse-transcribed double-stranded DNA copy of their RNA genome into host DNA. The human genome contains many copies of endogenous retroviral DNA sequences integrated in the genome. The viral DNA is flanked by long terminal repeat sequences (LTR) that contain regulatory elements, signal sequences, transcription factor binding sites, and polyadenylation signals that play a role in modulating gene expression.

LTR — Long Terminal Repeat

mapping — In genomics, mapping refers to the various techniques for determining the position and relative order of markers or genes (loci) on a chromosome and relative distance between them based on recombination frequency (genetic map), the absolute position of genes and the distance between them in nucleotide base pairs (physical map), or the position of markers or genes on the chromosome based on hybridization (cytogenetic map).

MapViewer — MapViewer is a genome browsing tool used to view and search an organism's genome and display chromosome maps.

MathML — Mathemathical Markup Language; it is used for handling mathemathical equations in XML.

MD5 — MD5 (Message-Digest Algorithm 5) is a database hash function used to validate data integrity.

MedGen — MedGen is an NCBI resource providing information related to human medical genetics, such as attributes of conditions with a genetic contribution.

MeSH — Medical Subject Headings is the National Library of Medicine's controlled vocabulary thesaurus used for indexing articles in PubMed.

metagenome — A metagenome is a collective genome representative of the community of organisms, for example, microorganisms, many of which cannot be cultivated outside of their environment.

MHC — Major Histocompatibility Complex

MIAME — Minimum Information About a Microarray Experiment

microarray — See DNA chip.

My NCBI — My NCBI is a tool that retains user information and database preferences to provide customized services for many NCBI databases.

N50 — The N50 value is the size of the smallest contig (or scaffold) such that 50% of the genome is contained in contigs of size N50 or larger. It is a statistic used in genome assemblies.

NCBI — National Center for Biotechnology Information

NIHMS — National Institutes of Health Manuscript Submission system

NISO — National Information Standards Organization

NLM — National Library of Medicine

OAI — Open Archives Initiative

OMIM — Online Mendelian Inheritance in Man

open access — Open access refers to online publications, often publicly funded research output, that have no restrictions on access (e.g., no user fees) and are not subject to restrictions on use (license and copyright).

open reading frame — Sequence of DNA that begins with an initiation codon and ends with a termination codon, specifying a gene sequence.

opposite strand — It is the DNA strand facing the template strand.

ORF — Open Reading Frame

ortholog — Orthologous genes are found in different species and arise from a single genomic locus in a common ancestor. Orthologs may not have a similar function.

overlapping gene — A gene that shares part or all of its nucleotide sequence with another gene which may include regulatory elements or intron sequence.

p-value — The p-value or probability value is a measure of statistical significance. If the null hypothesis is assumed to be true, the p-value indicates the probability of observing a particular outcome or result. The lower the p-value, the lower the probability of a result occurring by chance and therefore, greater the significance. A p-value of 0.05 indicates that there is a 5% probability of a chance outcome. For historical reasons, a p-value of 0.05 has been used as a cutoff for significance. Values less than 0.05 are considered significant. If the p value is 0.001 or less (less than a 0.1% probability that the results occurred by chance), the result is seen as highly significant.

PAC — P1-derived artificial chromosome; cloning system for large DNA inserts (100-300 kb) of genomic DNA which is based on a combination of the BAC cloning system and bacteriophage P1 vector.

pairwise alignment — Alignment of the two protein or nucleic acid sequences to determine regions of similarity that may reveal structural or functional relationships.

PDB — Protein Data Bank

paralog — Paralogs are homologous genes within a species that arose from a duplication event.

PDF — Portable Document Format

pedigree — In genetics, a pedigree is a multigenerational family hierarchy tree with symbol convention used to depict inheritance of normal or disease traits by individuals in the tree.

PheGenI — Phenotype Genotype Integrator, a resource for phenotypes compiled by merging NHGRI genome-wide association study (GWAS) catalog data with data from NCBI databases including Gene, dbGaP, OMIM, GTEx and dbSNP.

phenotype — The observable characteristics or features of a living organism.

phylogenetic tree — An evolutionary tree for organismal species or cellular macromolecules (example tRNA) that is built using inheritance or molecular sequence information.

PIR — Protein Information Resource. A free resource of protein databases and analysis tools.

PI — Principal Investigator

PLink — PLink is a free, open-source whole genome association analysis toolset.

PMCI — PubMed Central International

PMID — PubMed Identifier; a unique identifier for each PubMed record.

pNIHMS — Portable NIHMS

polyadenylation signal — Polyadenylation is a post-transcriptional modification to the 3' end of eukaryotic mRNA involving the addition of a sequence of ~250 adenosine nucleotides in a template-independent manner, and occurs about 30 base-pairs downstream from a short signal sequence in the primary transcript, typically AAUAAA.

polyprotein — Precursor protein that is enzymatically processed to form the mature protein.

ProtEST — A view in the Unigene browser for comparing proteins to the EST cDNA sequences.

pseudogene — An altered copy of the original gene, which may either arise from reverse transcription of mRNA followed by integration of the double-stranded cDNA into the chromosome at a break event (processed pseudogene) or by a gene duplication event (unprocessed pseudogene). For a long time, pseudogenes were thought to be non-functional transcriptionally and translationally, but new roles are emerging for pseudogenes as regulatory modulators.

PSI-BLAST — Position-Specific Iterative BLAST (PSI-BLAST) is an iterative search using the protein BLAST algorithm in which the amino acid frequency at each position determination built after the initial search is then used in subsequent searches. The process may be repeated, if desired, with new sequences found in each cycle used to refine the profile (Altschul et al. 1997).

PSSM — Position-Specific Scoring Matrix; a type of scoring matrix used in protein BLAST searches providing position-specific amino acid substitution scores for each position in a protein multiple-sequence alignment.

public access — See open access.

Pubreader — A viewer for reading books and journal articles at NCBI on tablet devices.

QA — See quality assurance.

QC — Quality Control

quality assessment — An assessment of quality is a part of quality assurance.

quality assurance — Standardized process designed and undertaken to avoid mistakes or errors in a released product.

quality control — A set of procedures that are performed to ensure that a product meets a specified standard.

query — Query refers to the term used in the search.

query translation — The full search expression including MeSH expansion and automatic term mapping, shown in the details box in Entrez search results.

reading frame — See open reading frame.

recombinant — Referring to the product of recombination, either DNA or gene or protein.

recombination — Recombination results from crossing-over events involving exchange of DNA sequences between structurally similar chromosomes during meiosis in the diploid cell. It is a process whereby new gene combinations are formed in the progeny. It can also be performed enzymatically on DNA *in vitro*.

reference sequence — See RefSeq.

RefSeq — RefSeq, NCBI's Reference Sequence project, is a non-redundant, annotated set of sequences that serve as reference standards. They are derived from the INSDC databases and include chromosomes, complete genomes (plasmids, organelles, viruses, archaea, bacteria, and eukaryotes), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.

RefSeqGene — A subset of RefSeq, RefSeqGene defines genomic sequences to be used as reference standards for well-characterized genes. It provides more stable gene-specific genomic sequence for each gene including upstream and downstream flanking regions, and versioning information for conversion of coordinates in case of updates.

refSNP — In dbSNP, variant information that results from aggregation of submission data by location on the genome and type of variation is assigned a refSNP (rs) identifier. The rs identifier is used as a reference for that variant location, but does not indicate the explicit sequence change at a location.

RepeatMasker — RepeatMasker is a program that analyzes a query sequence for repeat sequences, creating an output showing the annotation of the repeats as well as a modified query sequence that masks the annotated repeats.

RH map — A Radiation Hybrid (RH) map is obtained by fusing irradiated human cells with rodent cells to create hybrids. The radiation causes chromosomal breakage in a dose-dependent manner. Following fusion with the rodent cell line, the chromosomal fragments get integrated into the rodent chromosomes. The collection of hybrid cells forms a panel and can be used for mapping.

RPSBLAST — Reverse-Position-Specific BLAST (RPSBLAST) is a tool that is used to search a protein query against a database of PSSMs that were usually produced by PSI-BLAST.

scaffold — A scaffold is an ordered and oriented set of contigs. It can contain gaps, but there is typically some evidence to support the contig order, orientation, and gap size estimates.

Schema — Representation

SEF — Serials Extract File, required for Medline indexing.

sequence masking — In determining similarities between homologous sequences, it is sometimes necessary to exclude non-specific or non-homologous similarities. This is done using standard techniques by which sequences which are of low complexity or short-period tandem repeat sequences are masked.

sequence pair — Two aligned component sequences used in the generation of a contig.

sequence tagged site — See STS

SGD — Saccharomyces Genome Database. A database for the molecular biology and genetics of *Saccharomyces cerevisceae*, also known as baker's yeast.

SGML — Standard Generalized Markup Language. The international standard for specifying the structure and content of electronic documents. SGML is used for the markup of data in a way that is self-describing. SGML is not a language but a way of defining languages that are developed along its general principles. A subset of SGML called XML is more widely used for the markup of data. HTML (Hypertext Markup Language) is based on SGML and uses some of its concepts to provide a universal markup language for the display of information and the linking of different pieces of that information.

similarity score — Quantitative measure of the similarity between two sequences.

single nucleotide polymorphism — See SNP

small RNA — Refer to snRNA? [CHECKING WITH thibaudf]

small nuclear RNA — See snRNA

snRNA — snRNAs (small nuclear RNAs) are small non-coding RNAs that are localized to the nucleus and that play a role in splicing of introns from premature transcripts.

SNOMED–CT — Systematized Nomenclature of Medicine–Clinical Terms. Vocabulary of clinical terminology, maintained by International Health Terminology Standards Development Organisation (IHTSDO).

SNP — Single Nucleotide Polymorphism. Common, but minute, variations that occur in human DNA at a frequency of 1 every 1,000 bases. An SNP is a single base-pair site within the genome at which more than one of the four possible base pairs is commonly found in natural populations. Over 10 million SNP sites have been identified and mapped on the sequence of the genome, providing the densest possible map of genetic differences. SNP is pronounced "snip".

source file — A source file refers to the original file or the file provided by the submitter.

Spidey — A software program used to align cDNA (spliced mRNA sequences) to genomic sequences (Wheelan et al. 2001).

splice form — See splice site

splice signal — See splice site.

splice site — Refers to the location of the exon-intron junctions in a pre-mRNA (i.e., the primary transcript that must undergo additional processing to become a mature RNA for translation into a protein). Splice sites can be determined by comparing the sequence of genomic DNA with that of the cDNA sequence. In mRNA, introns (non-protein coding regions) are removed by the splicing machinery; however, exons can also be removed. Depending on which exons (or parts of exons) are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are "splice variants" or "splice forms" or "alternatively spliced".

Splign — A software program comprising a set of algorithms for computing cDNA-to-Genome alignments (Kapustin et al. 2008).

Splitd — NCBI queuing system for processing BLAST requests.

SQL — Structured Query Language; a programming language for managing and processing data in relational databases.

ssRNA — single-stranded RNA; viruses such as Ebola virus and Marburg virus of the Family Filoviridae and Dengue virus and Yellow fever virus of the Family Flaviviridae have a single stranded RNA genome.

structural variation — Genomic structural variation includes insertions, deletions, duplications, inversions, or chromosomal translocations longer than 50 bp. These variants can occur in coding or noncoding DNA and they can be inherited or arise sporadically in the germline or somatic cells. Some of these variants may be benign, with or without phenotypic manifestations whereas others result in disease, for example, 22q11.2 Deletion Syndrome.

structured output — Results written to a file in a way that they conform to an external schema or set of rules, permitting validation and machine-readability.

STS — Sequence Tagged Site. A short DNA segment that occurs only once in the human genome, the exact location and order of bases of which are known. Because each is unique, STSs are helpful for chromosome placement of mapping and sequencing data from many different laboratories. STSs serve as landmarks on the physical map of the human genome.

style sheet — A style sheet is a mechanism to separate content from presentation and contains information necessary for formatting the text. For example, Cascading Style Sheets (CSS) are used to format HTML pages.

Swiss-Prot — Swiss-Prot is a curated protein sequence database which is a part of the UniProt knowledgebase, UniProtKB. It provides a high level of annotation (such as the description of protein function, nomenclature and taxonomy, structure, domains, sequence, post-translational modifications, variants, publications, etc.) and integration with other databases.

switch point — The switch point is the base at which a contig sequence stops being generated from one component sequence and switches to being generated from the next component sequence. There must be at least one switch point between adjacent component sequences in a contig.

tagserver — The TagServer is a database in PMC that is used for storing metadata about the journal article based on information mined from the article, such as Gene and protein names and identifiers.

TBLASTN — TBLASTN is an application which searches a protein query against a nucleotide database, dynamically translating the database.

term mapping — Untagged terms that are entered in the search box are matched against a series of translation tables (example, for Medical Subject Headings [MeSH], journals, author names etc.) and indexes in a defined order until a match is found. Once the match is found, the mapping process is complete.

termination codon — Canonical or non-canonical codon at which the ribosome is released from the RNA and translation of protein synthesis ends.

termination signal — See termination codon.

tiling path — An ordered list or map that defines a set of overlapping clones that covers a chromosome or other extended segment of DNA.

TPF — Tiling Path Format. A table format used to specify the set of clones that will provide the best possible sequence coverage for a particular chromosome, the order of the clones along the chromosome, and the location of any gaps in the clone tiling path. Also used to refer to a file (Tiling Path File) in which the minimal tiling path of clones covering a chromosome is specified in Tiling Path Format.

traceback — The traceback is the process which converts the High Scoring Segment Pairs (HSPs) and generates alignments in BLAST.

transcriptome — The transcriptome refers to the full set of transcripts in a cell assembled by a method called RNA-seq in which RNA from cells is collected, sampled, and sequenced. It includes alternative splice variants, variants created by alternative transcription initiation and alternative transcription termination, and noncoding RNA genes.

transfer RNA — See tRNA.

translation initiation site — See translation start site.

translation start site — The position within an mRNA at which synthesis of a protein begins. The translation start site is usually an AUG codon, but occasionally, GUG or CUG codons are used to initiate protein synthesis.

translation stop signal — See termination codon.

tree viewer — Graphical representation of a tree hierarchy showing the root, branches, and leaves (nodes).

tRNA — Transfer RNA (tRNA) is a small RNA molecule that plays a role in protein synthesis. Typically tRNAs have highly conserved sequences and four-armed cloverleaf secondary structures formed by base pairing within the tRNA resulting in the formation of hairpin loops. Exceptions to this cloverleaf structure occur in mitochondrial and nematode tRNAs. There are two parts to the tRNA that are involved in protein synthesis: the anticodon that recognizes and binds to the complementary codon in the mRNA transcript; and the amino-acid binding site. The aminoacyl-tRNA synthetase are involved in pairing the amino acid with its cognate tRNA. Amino acids are transferred from the amino-acyl tRNA to the growing peptide chain via the formation of a peptide bond.

tRNAscan-SE — tRNAscan-SE is a program for identifying tRNA genes in DNA sequence.

UID — Identifier for a public record (e.g., publication or sequence) in the NCBI Entrez system.

UMLS — Unified Medical Language System

UniGene — UniGene is a computational system for analyzing the transcriptome, expression of transcripts, and the libraries from which they were derived, allowing evaluation of expression pattern by various parameters such as tissue or health status.

UniProt — Universal Protein Resource (UniProt) is a resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR).

URI — Uniform Resource Identifier (URI)

UTR — Untranslated region (UTR) is the region of mRNA at the 5' or 3' end that does not code for protein and typically includes regulatory sequence.

WAM — Weight Array Method (WAM), a method for sequence analysis (Zhang and Mar3 1993).

WGS — Whole Genome Shotgun; refers to sequencing approach in which the whole genome is fragmented and sequenced using a shotgun approach, and then the sequence of the fragments are reassembled for genome assembly. [CHECK – IMPROVE]

WMM — Weight Matrix Method (WMM), [CHECK, NEED REFERENCE]

XHTML — Extensible Hypertext Markup Language (XHTML)

XML — Extensible Markup Language

XQuery — XML Query; a query language for structured data such as XML.

XSL — Extensible Stylesheet Language

YAC — Yeast artificial chromosomes (YACs) were developed for cloning large fragments of genomic DNA into yeast. YACs can carry large, megabase-size inserts of genomic DNA. BACs or PACs have advantages over YACs. Cloned DNA in the YAC system is more difficult to manipulate and is often chimeric. Also, recombination events in yeast can lead to deletions or rearrangements of the insert DNA, thus YACs are less stable than BACs or PACs.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10. PubMed PMID: 2231712.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402. Review. PubMed PMID: 9254694.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091 doi: 10.1371/journal.pbio.1001091. PubMed PMID: 21750661.

Genome Reference Consortium (GRC) Assembly Terminology. Available from http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/info/definitions.shtml [Accessed October 12, 2017].

Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct. 2008 May 21;3:20 doi: 10.1186/1745-6150-3-20. PubMed PMID: 18495041.

Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003a. Chapter 4, Principles and Methods of Sequence Analysis. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20261/

Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003b. Chapter 3, Information Sources for Genomics. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20256/

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444–8. PubMed PMID: 3162770.

Wheelan SJ, Church DM, Ostell JM. Spidey: a tool for mRNA-to-genomic alignments. Genome Res. 2001 Nov;11(11):1952–7. PubMed PMID: 11691860.

Zhang MQ, Marr TG. A weight array method for splicing signal analysis. Comput Appl Biosci. 1993 Oct;9(5):499–509. PubMed PMID: 8293321.