Glossary of common genetic and bioinformatic terms

Genetic terms

Allele

A variant form of a genetic sequence. Typically, a person may carry up to two unique alleles for a given gene, one on each paired chromosome.

Related terms: Gene, SNP

Antisense strand

The complement strand to the sense strand of the DNA. This is the strand that is transcribed by RNA polymerase.

Related genetic terms: Sense strand, Reverse complement

Related bioinformatic terms: + strand, – strand

Base pair

Two complementary nucleotides on different strands of DNA, whose bases are bound by hydrogen bonds.

Related terms: Nucleotide, Codon, Start codon, Stop codon

Central dogma of molecular biology

DNA is transcribed to RNA; messenger RNA is translated into protein.

Coding sequence (CDS)

The series of DNA or RNA nucleotides that codes for a protein (therefore, no introns).

Related terms: Open reading frame (ORF)

Coding strand

See Sense strand.

Codon

Three nucleotides that encode one amino acid.

Related terms: Nucleotide, Base pair, Start codon, Stop codon

Complex variant

A genetic variation in which several mutation types co-occur. For example, deletion of AA and insertion of CTG.

Related terms: Mutation, Insertion, Deletion, SNP, MNP, Indel

Deletion

Loss of one or more bases from a DNA sequence.

Related terms: Mutation, Insertion, Deletion, SNP, MNP, Indel

Exon

The section(s) of a transcript that will become the mature RNA transcript after RNA splicing has occurred.

Related terms: Gene, Transcript, Intron, UTR

Gene

A unit of heredity encoded in DNA with defined start and stop coordinates. All genes code for transcript RNAs, and some encoded RNAs will be translated into proteins.

Related terms: Transcript, Exon, Intron, UTR, Allele

HDR donor template

A single- or double-stranded DNA sequence containing a desired insert sequence flanked by homology arms complementary to the adjacent sequence of a planned break in genomic DNA. A donor template is included in CRISPR-Cas9 HDR experiments, allowing scientists to create desired point mutations, incorporate tags, or add other functional units into a specific genomic location.

Related terms: Homology-directed repair (HDR)

Homology-directed repair (HDR)

A cellular mechanism for repair of double-stranded breaks in genomic DNA involving homologous recombination of a donor DNA sequence into the genome. Scientists have taken advantage of this cellular mechanism to insert desired sequences into specific genomic locations after a double-stranded break (DSB) is generated by CRISPR-Cas9 cleavage.

Related terms: HDR donor template

Indel

Sequence that has been inserted or deleted in one genome relative to another. A deletion in one genome corresponds to an insertion in the other.

Related terms: Mutation, Deletion, Insertion, SNP, MNP, Complex variant

Insertion

Addition of one or more extra bases into a DNA sequence.

Related terms: Mutation, Deletion, SNP, MNP, Complex variant, Indel

Intron

The segment(s) of a gene that are transcribed to RNA but are not retained after RNA splicing has occurred.

Related terms: Gene, Transcript, Exon, UTR

MNP

Multi Nucleotide Polymorphism—multiple nucleotides are replaced with the same number of different nucleotides at a genetic locus.

Related terms: Mutation, Insertion, Deletion, SNP, Complex variant, Indel

Mutation

Any change in a DNA sequence. Examples of mutations include insertion, deletion, single nucleotide polymorphism (SNP), multi nucleotide polymorphism (MNP), and complex variant.

Non-coding strand

See Antisense strand.

Nucleotide

The basic structural unit of DNA. A nucleotide contains a nucleoside plus a phosphate. (A nucleoside contains a ribose or deoxyribose sugar and a pyrimidine or purine base).

Related terms: Base pair, Codon, Start codon, Stop codon

Oligo

Oligonucleotide; a short strand of nucleotides.

Related terms: ssODN, ssDNA

Open reading frame (ORF)

The translatable part of an RNA sequence beginning with the start codon and ending with the stop codon. An ORF may be predicted to be translated based on sequence alone (e.g., based on presence of start and stop codons) and is typically the longest sequence within a transcript.

Related terms: Coding sequence (CDS)

Protein

A polymer of amino acids joined by peptide bonds, usually folded in a 3-dimensional functional structure. The amino acid sequence is defined by the transcript sequence, as described by the Central dogma of molecular biology.

Related terms: Central dogma of molecular biology

Reverse complement

The sequence of DNA that contains complementary bases in the reverse direction from the target molecule. The antisense strand is the "complement" strand and is called the "reverse complement" only when it is written in the 5′ to 3′ direction.

Example:

Sense strand: 5′-CCTGGAG-3′

Antisense (complement) strand: 3′-GGACCTC-5′

Reverse complement: 5′-CTCCAGG-3′

Related terms: Sense strand, Antisense strand

Sense strand

The strand of DNA with the same sequence as the transcribed mRNA, except DNA contains thymine, while RNA contains uracil.

Related genetic terms: Reverse complement, Antisense strand

Related bioinformatic terms: + strand, – strand

SNP

Single Nucleotide Polymorphism—a single nucleotide is replaced with a different single nucleotide at a genetic locus. To be considered a SNP, a variation must be present in more than 1% of the population. If the variation occurs in less than 1% of the population, it would be considered a rare mutation (abnormal change) instead of a SNP.

Related terms: Mutation, Synonymous SNP, Insertion, Deletion, MNP, Complex variant, Indel

Splicing

The removal of introns from a precursor mRNA to produce a mature mRNA.

ssDNA

Single-stranded DNA.

Related terms: ssODN, oligo

ssODN

Single-stranded oligodeoxynucleotide, often used as an HDR donor template in CRISPR genome editing experiments.

Related terms: ssDNA, oligo

Start codon

A codon that defines where on a transcript translation begins. The most common start codon is AUG.

Related terms: Nucleotide, Base pair, Codon, Stop codon

Stop codon

A codon that defines where on a transcript translation ends. UGA, UAA, and UAG are stop codons.

Related terms: Nucleotide, Base pair, Codon, Start codon

Synonymous SNP:

A SNP that does not change the protein sequence.

Related terms: Mutation, SNP

Transcript

A messenger RNA molecule which has been transcribed from DNA. Transcripts can contain untranslated regions (UTRs), exon, and intron segments. Some, but not all, transcripts encode proteins.

Related terms: Gene, Exon, Intron, UTR

UTR

Untranslated region—sections of RNA transcripts upstream (5′) or downstream (3′) of the coding sequence, retained after RNA splicing. UTR(s) may occupy part of the first or last exon within a transcript. Even though these are called untranslated regions, some 5′ UTRs contain small upstream open reading frames (uORFs) that help regulate gene expression.

Related terms: Gene, Transcript, Exon, Intron

Bioinformatic terms

+ strand

The strand of DNA in the reference file. This is a bioinformatic description that is distinct from sense or antisense strand designations.

Related bioinformatic terms: – strand

Related genetic terms: Sense strand, Antisense strand

– strand

The complement of the + strand in the reference file. This is a bioinformatic description that is distinct from sense or antisense strand designations.

Related bioinformatic terms: + strand

Related genetic terms: Sense strand, Antisense strand

1-based vs. 0-based coordinates

These are different ways of identifying nucleotides or amino acids within a sequence. In the 1-based system, the first unit of the polymer (e.g., the first nucleotide) is counted as number 1. In the 0-based system, the number 0 designates the location before the first nucleotide. Different databases and file types may use different systems. GFF, SAM, and VCF files use 1-based coordinates, while BED and BAM files use 0-based coordinates.

Example:

	Coordinates			
	1-based coordinate system	0-based coordinate system	Nucleotide(s)	Sequence*
Single nucleotide	5–5	4–5	C	AGTA**C**GTCAAG
Range of nucleotides	4–6	3–6	ACG	AGT**ACG**TCAAG
Single nucleotide variant	5–5	4–5	C/G	AGTA**G**GTCAAG
Deletion	5–5	4–5	C/–	AGTA–GTCAAG
Insertion	5–6	5–5	–/TAA	AGTAC**TAA**GTCAAG

* Variant sequence shown in bold green.

Related terms: Genomic coordinate or genomic interval

Annotation file

A text file that has specific formatting requirements and contains information describing exon, mRNA, coding, intergenic, UTR, and regulatory sequence locations within a reference genome.

Example file types: BED, GTF, and GFF

FASTA file

A text file containing a nucleotide or protein sequence in a specific format. A unique description line, beginning with ">", is followed by lines of sequence information.

Example:

>IDT – Example sequence (DNA 1)

ACGCTGCTCGATGTTTAGCTAAAGCTAGCTAGCTAGCTAGGCATGCTAGCTAGCTAACTAGCTAGTATATTATATAGCGG
GCGATCGATCGATAGCATGGATAGCTAGCTAGATCTATCGATTATATATAGGCGCTAGGTTGAATATTCCCTAGGTCTAT
GCTAGCTAGACTAGCTAGCTAGCTAG

GenBank database

A database maintained by the US National Institutes of Health (NIH) containing all publicly available DNA sequences. Sequence entries are owned by the original submitter (e.g., laboratory or sequencing project), so some loci contain multiple or redundant entries. Entries may contain associated protein information.

Related terms: RefSeq (Reference Sequence database)

Gene symbol

An abbreviation for a gene name. Several species have nomenclature committees to define rules for or approve gene symbols to standardize nomenclature and promote clear communication. For example, HUGO (Human Genome Organisation) Gene Nomenclature Committee approves symbols and names for human genes. They have designated *HPRT1* as the gene symbol for the *hypoxanthine phosphoribosyltransferase 1* gene.

Related terms: NCBI accession number

Genomic coordinate or genomic interval

Consists of chromosome name and integers that together define a location (position or series of nucleotides) within a reference genome. The information specified typically includes chromosome name, start position, end position, and chromosome strand.

Related terms: 1-based vs. 0-based coordinates

HGVS nomenclature

Nomenclature system that provides standard recommendations for how to describe sequence variants. Details can be found at http://varnomen.hgvs.org/. (HGVS is the Human Genome Variation Society).

Transcript example: NM_001129765.1:c.696_698del

Related terms: Variant files, Variant call format (VCF) file

NCBI accession number

An alphanumeric identifier referring to a specific nucleotide or protein sequence in an NCBI database (e.g., GenBank, RefSeq). Not all databases use the same accession number for the same gene, and the format of NCBI accession numbers varies by database. For example, for the HPRT1 gene, the RefSeq accession number is NM_000194.3, and one of the GenBank accession numbers is CR407645.1.

Related terms: Gene symbol

Reference genome

A collection of sequences that is considered a representative of a species. Also known as a reference assembly.

Related terms: Variant file, Variant Call Format (VCF) file, Annotation file

RefSeq (Reference Sequence database)

A separate subset database of GenBank that is curated to eliminate sequence duplications. This database contains well-annotated, complete, organized sets of records for DNA, RNA, and protein sequences.

Related terms: GenBank database

Variant call format (VCF) file

Most often used as a file type, the nomenclature describes the positions of variants in a genome using the 1-based coordinate system. Reference and alternate allele sequences are listed. Details can be found at http://www.internationalgenome.org/wiki/Analysis/vcf4.0/.

Example (from VCF file):

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chrX	152867579	.	AGAA	A	.	.	.

Related terms: Variant file, Reference genome, HGVS nomenclature

Variant files

A text file that has specific formatting requirements and that contains sequence, location, and experimental detail (e.g., genotype quality, noise level, strand bias, variant frequency) information about variants in the genome.

Example file types: VCF and BCF

Related terms: Variant call format (VCF) file

Integrated DNA Technologies, Inc. (IDT) is your Advocate for the Genomics Age. For more than 30 years, IDT's innovative tools and solutions for genomics applications have been driving advances that inspire scientists to dream big and achieve their next breakthroughs. IDT develops, manufactures, and markets nucleic acid products that support the life sciences industry in the areas of academic and commercial research, agriculture, medical diagnostics, and pharmaceutical development. We have a global reach with personalized customer service. See what more we can do for **you** at www.idtdna.com.

Technical support:

applicationsupport@idtdna.com

CRS-10141-ED 05/19

INTEGRATED DNA TECHNOLOGIES