Pathways Project: Reference Glossary

Katie M. Sandlin; Modified from the "Glossary for Understanding Eukaryotic Genomes" which was created during the 2011 National GEP Faculty Workshop.

Jump to letter: A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Term	Definition
3' ("three prime")	The 3' (three prime) end of a **DNA** molecule is where the sugar's third carbon is unattached to an adjacent **nucleotide**. Since DNA is synthesized in the 5'-to-3' direction, nucleotides are added to the 3' end of the growing strand (cf. 5').
5' ("five prime")	Just like letters and words in the English language are read from left to right, **DNA** is 'read' in a specific direction 5'-to-3'. The 5' (five prime) end of a DNA molecule is where the sugar's fifth carbon phosphate group is unattached to an adjacent **nucleotide** (cf. 3').
ab initio	In computing, *ab initio* is a term used to define computations based solely on theory or using only fundamental constants. In computational biology the term refers to algorithms that use only sequence information rather than including experimental observations to make predictions about **gene** structure.
accession number	A unique identification number given to every **DNA**, **RNA**, and **protein** sequence submitted to the National Center for Biotechnology Information (NCBI) or equivalent **databases**. For example, the accession number for *Rheb* in *D. yakuba* is NC_052530 in the NCBI database.
adenine (A)	Adenine (A) is one of the four nucleotide **bases** in **DNA**, with the other three being **cytosine** (C), **guanine** (G) and **thymine** (T). Within a double-stranded DNA molecule, adenine bases on one strand pair with thymine bases on the opposite strand. The sequence of the four **nucleotide** bases encodes DNA's information[1].
algorithm	A sequence of instructions used to solve a mathematical or computational problem. An algorithm typically transforms the input value into the output value via a sequence of steps that include computations, iterations, and conditional branches.
alignment	In **bioinformatics**, a sequence alignment is a way of arranging two or more sequences of **DNA**, **RNA**, or **protein** to identify regions of similarity; such similarity may be a consequence of functional, structural, or **evolutionary** relationships between the sequences.

alignment score	An alignment score is a numerical value used in computational biology to quantify the level of similarity between two aligned sequences. Generally, the higher the score, the more similar the two sequences.
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[3].
allele frequency	The proportion of a specific gene **variant** (i.e., an **allele**) among all copies of the **gene** in the population[3].
alternative splicing	The inclusion or exclusion of certain **exons** in the **splicing** reactions that determine the sequences included in the final **mRNA** product of a **gene**. This mechanism is utilized to generate a series of closely related protein **isoforms**, which differ by the inclusion or exclusion of the **protein** domains encoded by those exons. Alternative splicing is directed by RNA-binding proteins that block or stimulate utilization of a particular splice site.
amino acid (aa)	Basic building block molecules of **peptides** and **proteins**. The sequence of amino acids in a protein is determined by the gene's **codon** sequence[3]. The basic building block of proteins. There are 20 different amino acids. A protein consists of one or more chains of amino acids (called **polypeptides**) whose sequence is encoded in a **gene**[1].
annotation	Gene annotation is the process of indicating the location, structure, and identity of **genes** in a **genome**. As this may be based on incomplete information, gene annotations are constantly changing with improved knowledge. Gene annotation **databases** change regularly, and different databases may refer to the same gene/protein by different names, reflecting a changing understanding of **protein** function. Gene Annotation assigns information to A's, T's, G's, and C's in **DNA** sequences; annotations can be divided into two types[2]: 1. Structural annotation – finding genes in the sequence 2. Functional annotation – assigning function to genes in the sequence
Annotation Files Merger	The Annotation Files Merger is designed to help you combine the individual files generated by the **Gene Model Checker** into a single project file suitable for project submission. This tool also performs additional checks to verify that all the **isoforms** have been **annotated** and it allows you to view all the annotated **gene models** on the **GEP UCSC Genome Browser**. You will only need to merge the files if you

	have more than one **unique isoform**. See Part 7.3 of the Pathways Project: Annotation Walkthrough for directions on using the Annotation Files Merger.
assembly	The genome assembly is the **genome** sequence produced after **chromosomes** have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together. A genome assembly is updated when **DNA** has been sequenced that allows **gaps** to be filled. It may also be updated when a new assembling algorithm is released and applied to the data, resulting in an assembly that's different from the previous version.
Augustus	Augustus is an ***ab initio*** single **isoform** gene finder.
base	Often used as a synonym for **adenine** (A), **cytosine** (C), **guanine** (G) or **thymine** (T) in **DNA**.
base pair (bp)	A base pair consists of two complementary DNA **nucleotide** bases that pair together to form a "rung of the **DNA** ladder." DNA is made of two linked strands that wind around each other to resemble a twisted ladder — a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four **bases**: **adenine** (A), **cytosine** (C), **guanine** (G) or **thymine** (T). The two strands are held together by hydrogen bonds between pairs of bases: adenine pairs with thymine (A-T), and cytosine pairs with guanine (C-G) [1].
bioinformatics	Bioinformatics, as related to genetics and **genomics**, is a scientific interdisciplinary field that involves using computer technology to collect, store, analyze, and disseminate biological data and information, such as **DNA** and **amino acid** sequences or **annotations** about those sequences. Scientists and clinicians use **databases** that organize and index such biological information to increase our understanding of health and disease and, in certain cases, as part of medical care[1]. 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[3].
***bl2seq* (BLAST 2 sequences)**	**BLAST** version in which two sequences can be compared to each other. The sequences could either be **nucleic acid** or **protein**; select "Align two or more sequences" within a particular BLAST search to use this feature.
BLAST	Basic Local Alignment Search Tool is a sequence comparison algorithm that is used to search sequence **databases** for optimal **local alignments** to a **query**[3]. BLAST finds regions of local similarity between **nucleotide** or **protein** sequences by comparing nucleotide or protein sequences to sequence databases (or to an individual nucleotide or protein sequence) and calculates the statistical significance of each match.

blastn **(nucleotide BLAST)**	**BLAST** version in which the **query** and **subject** are both **nucleotide** sequences. Typically used to search a nucleotide **database** with a nucleotide sequence. Query: nucleotide Database/Subject: nucleotide Function: searching with shorter queries, cross-species comparison Common Use Cases: map mRNAs against genomic assemblies
blastp **(protein BLAST)**	**BLAST** version in which both the **query** and **subject** are **amino acid** sequences. Typically used to search a protein **database** with a **protein** sequence. Query: protein Database/Subject: protein Function: general sequence identification and similarity searches Common Use Cases: search for proteins similar to predicted genes
blastx	**BLAST** version in which the **query** is a **nucleotide** sequence, and all 6 **reading frames** are translated and compared to the **subject**, which is an **amino acid** sequence. Typically used to search a protein **database** with a nucleotide sequence. Query: nucleotide translated to protein Database/Subject: protein Function: identify potential protein products encoded by a nucleotide query Common Use Cases: map proteins/CDS against genomic sequence
canonical	In agreement with existing principles and standards generated from data and evidence. For example, the canonical **splice donor** sequence is GT. In rare instances, the sequence GC is used instead and since GC is not the "standard," it would be referred to as **non-canonical**.
cDNA (copy or complementary DNA)	A **DNA** sequence obtained by reverse **transcription** of a messenger RNA (**mRNA**) sequence[3]. cDNA (short for copy DNA; also called complementary DNA) is synthetic DNA that has been transcribed from a specific mRNA through a reaction using the enzyme reverse transcriptase. While DNA is composed of both coding and **non-coding** sequences, cDNA contains only coding sequences. Scientists often synthesize and use cDNA as a tool in gene cloning and other research experiments[1].
chromosome	A chromosome is a molecule of double-stranded **DNA**, carrying an arrangement of **genes** interspersed with other sequences (e.g., **regulatory elements**). Chromosomes in eukaryotes are typically linear and extend from one end (a telomere) through the chromosome center (centromere) to the other telomere end.
coding DNA sequence (CDS)	The portion of the **DNA** sequence of a **gene** which is translated into **protein**. Successful **translation** of a CDS results in the synthesis of a protein[3]. CDS and **coding exon** are often used interchangeably.

coding exon	In a **gene**, any **exon** which contains some part of the **coding DNA sequence** (CDS). Coding exon and **CDS** are often used interchangeably.
coding region	The sequence of **DNA** that is translated into **protein** and includes an initiation **(start) codon** and a termination **(stop) codon**[3].
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[3].
collinear	In Euclidean geometry, collinearity refers to a set of points that lie in a straight line. In the context of the ***tblastn*** **alignments** between a **protein** sequence (**query**) and a **nucleotide** sequence (**subject**), a set of alignments are collinear if the order of the alignment blocks is consistent with the protein sequence. For a protein located on the **positive strand** of the nucleotide sequence, the subject **coordinates** for the collinear set of alignment blocks will be in ascending order with respect to the protein sequence. For a protein located on the minus strand of the nucleotide sequence, the subject coordinates for the collinear set of alignment blocks will be in descending order with respect to the protein sequence.
complement	The **nucleotide** sequence of the **nucleic acid** strand which would form a double-stranded molecule with the nucleic acid strand in question, using standard **base-pairing** rules.
computational biology	A field of study which develops and applies computational **algorithms** and statistical techniques to analyze biological datasets and solve biological problems. The field encompasses many different areas, including image processing, simulations, modeling of biological systems and proteins, pair-wise and multiple sequence alignments, constructing genome assemblies, and processing of genomic (e.g., variant calling), epigenomic (e.g., ChIP-Seq), and transcriptomic (e.g., RNA-Seq) datasets.
consensus sequence	When comparing multiple sequences, the consensus sequence reflects the most commonly seen **base** at each position.
conserved domain	A domain (a distinct functional and/or structural unit of a **protein**) that has been maintained in the **homologous genes** of different species 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 amino acid residues, and hence the structural and/or functional properties of that region of the protein[3].
conserved sequence	**Homologous** DNA or amino acid sequence with a high degree of similarity between two species.
contig	A contig (as related to **genomic** studies; derived from the word “contiguous”) is a set of **DNA** segments or sequences that overlap in a way that provides a contiguous representation of a genomic region (cf. scaffold)[1].

coordinate	Numerical position within a biological sequence, e.g., the first **base** in a **DNA** sequence would have the coordinate 1.
cytosine (C)	Cytosine (C) is one of the four nucleotide **bases** in **DNA**, with the other three being **adenine** (A), **guanine** (G) and **thymine** (T). Within a double-stranded DNA molecule, cytosine bases on one strand pair with guanine bases on the opposite strand. The sequence of the four **nucleotide** bases encodes DNA's information[1].
database	Store of a set of logically related data or collection of files amenable to retrieval using scripts or a computer[3].
data science	Data science involves the study of large, complex data sets that arise from various types of research projects. With respect to **genomic** studies, such work requires expertise in quantitative scientific disciplines such as **bioinformatics**, computational biology, and biostatistics[1].
deletion	The removal of some part of a biological sequence. A deletion, as related to **genomics**, is a type of **mutation** that involves the loss of one or more contiguous **nucleotides** from a segment of **DNA**. A deletion can involve the loss of any number of nucleotides, from a single nucleotide to an entire piece of a **chromosome**[1].
***de novo* prediction**	Analysis of a **DNA** sequence to predict the location of **genes** (**exons** and **introns**) using only the sequence itself and known characteristics of genes (e.g., consensus splice site sequences in eukaryotic genomes). No knowledge of experimentally confirmed structure or function is used in *de novo* prediction.
differential gene expression	Pattern of **gene expression** in multi-cellular organisms, where distinct patterns of **transcription** are shown by different cell types, or at different times during development, or under different environmental conditions.
DNA	Deoxyribonucleic acid (abbreviated DNA) is the molecule that carries genetic information for the development and functioning of an organism. DNA is made of two linked strands that wind around each other to resemble a twisted ladder — a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four **bases**: **adenine** (A), **cytosine** (C), **guanine** (G) or **thymine** (T). The two strands are connected by chemical bonds between the bases: adenine bonds with thymine, and cytosine bonds with guanine. The sequence of the bases along DNA's backbone encodes biological information, such as the instructions for making a **protein** or **RNA** molecule[1].
DNA sequencing	DNA sequencing refers to the general laboratory technique for determining the exact sequence of **nucleotides**, or **bases**, in a **DNA** molecule. The sequence of the bases (often referred to by the first letters of their chemical names: A, T, C, and G) encodes the biological information that cells use to develop and operate. Establishing the sequence of DNA is key to understanding the function of **genes** and other parts of the **genome**. There are now several different methods available for DNA sequencing, each with its own

	characteristics, and the development of additional methods represents an active area of **genomics** research[1]. See Sequencing Technologies for more information.
domain	A portion of the protein that forms a distinct functional or structural unit. Protein domains can often fold and evolve independently from the rest of the protein sequence. Different proteins with similar functions often share the same domain architecture (see **conserved domain**). Multiple methods are used to represent a protein domain. For example, the NCBI Conserved Domains Database (CDD) uses Position-Specific Scoring Matrices (PSSMs) to capture sequence variations within the domain, while the Pfam database uses profile hidden Markov models to represent protein domains.
dot plot	The comparison of two sequences on an X-Y plot where points ("dots") on the graph ("plot") indicate sequence **identity** at the corresponding positions. A continuous line with slope 1 indicates high levels of sequence conservation in that region and provides confidence in the proposed **gene model**.
downstream	Toward the **3'** end of a single stranded **DNA** molecule or **gene** of interest (e.g., **target gene**). Downstream also refers to the genomic region that comes after the **feature** being examined (cf. upstream).
Drosophila	A genus of flies belonging to the Drosophilidae family in the Diptera order which includes more than 1,600 species (see Grady and DeSalle, 2018). *Drosophila* has been used in biological research for more than a century. For example, *Drosophila melanogaster* is a model organism that has been used in many studies focused on behavior, developmental biology, epigenetics, evolution, genetics, genomics, and as a model for studying human diseases (see the "Lay articles" section of the doso4public website for details).
duplication	Duplication, as related to **genomics**, refers to a type of **mutation** in which one or more copies of a **DNA** segment (which can be as small as a few **bases** or as large as a major **chromosomal** region) is produced. Duplications occur in all organisms. For example, they are especially prominent in plants, although they can also cause genetic diseases in humans. Duplications have been an important mechanism in the evolution of the **genomes** of humans and other organisms[1] (cf. gene duplication).
E-value (Expect Value)	The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a **database** of a particular size and complexity. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, you might expect to see one match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0," the higher is the "significance" of the match. However, it is important to note that searches with short sequences can be virtually identical and have a relatively high E-value. This is because the calculation of the

	E-value also considers the length of the **query** sequence and shorter sequences have a high probability of occurring in the database purely by chance[3].
Evidence track	An evidence track on the Genome Browser displays the computational analysis results (e.g., sequence alignments, gene predictions) and empirical data (e.g., RNA-Seq, ChIP-Seq data) for the genome of the target species. Evidence tracks can be used to identify genomic features within the target species that are of interest to researchers. For example, the "RefSeq Genes" evidence track shows the locations of the protein-coding genes and non-coding RNAs in the target genome predicted by the NCBI Eukaryotic Genome Annotation Pipeline. The "D. mel Proteins" evidence track shows the genomic regions in the target species that have significant sequence similarity to protein sequences in *Drosophila melanogaster*. The "RNA-Seq Coverage" evidence tracks are derived (primarily) from Illumina sequencing of processed mRNAs, and they demarcate the regions of the target species that are transcribed in the developmental stages and tissues sampled by RNA-Seq. The height of the histograms in the "RNA-Seq Coverage" evidence track is correlated with the relative expression level of the gene in the sample. Data from multiple evidence tracks can be integrated together to construct a hypothesis that is best supported by the available evidence. For example, annotators construct **gene models** by combining data from the sequence alignments, computational gene predictions, and RNA-Seq evidence tracks on the **GEP UCSC Genome Browser**.
evolution	Evolution, as related to **genomics**, refers to the process by which species change over time through changes in the **genome**. Such evolutionary changes result from **mutations** that produce **genomic variation**, giving rise to individuals whose biological functions or physical traits are altered. Those individuals who are best at adapting to their surroundings typically leave behind more offspring than less well-adapted individuals. Thus, over successive generations (in some cases spanning millions of years), one species may evolve to take on divergent functions or physical characteristics or may even evolve into a different species[1].
exon	An exon is a region of the **genome** whose **transcription** product ends up within a mature **mRNA** molecule. Some exons are **coding exons**, in that they contain information for making a **protein**, whereas others are **non-coding**. **Genes** in the genome consist of exons and **introns**[1]. Although introns are transcribed with the exons, the introns are **spliced** out and discarded during RNA processing.
Expressed Sequence Tag (EST)	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**[3].

feature	Any region of defined structure/sequence in a genomic fragment of **DNA**. Inherent features would include **genes**, **pseudogenes**, and **repetitive elements**. A feature may also be predicted by computational **algorithms**, such as those aimed at identifying **protein-coding genes**.
FlyBase	A **database** of ***Drosophila*** **genes** and **genomes**.
frame / reading frame	A frame is a single series of adjacent **nucleotide** triplets in **DNA** or RNA: one frame would have **bases** at positions 1, 4, 7, etc. as the first base of sequential **codons**. There are three possible reading frames in an **mRNA** strand and six in a double stranded DNA molecule because there are two strands from which **transcription** is possible. One common way is to refer to the three possible left-to-right reading frames as +1, +2, and +3, and the three possible right-to-left reading frames as -1, -2, and -3.
frameshift mutation	A frameshift mutation in a **gene** refers to the **insertion** or **deletion** of nucleotide **bases** in numbers that are not multiples of three. This is important because a cell reads a gene's code in groups of three bases when making a **protein**. Each of these "triplet **codons**" corresponds to one of 20 different **amino acids** used to build a protein. If a **mutation** disrupts this normal **reading frame**, then the entire gene sequence following the mutation will be incorrectly read. This can result in the addition of the wrong amino acids to the protein and/or the creation of a codon that stops the protein from growing longer[1].
gaps	When comparing or aligning two or more **protein** or **nucleic acid** sequences, spaces ("gaps") may be introduced in the final **alignment** to maximize matches and minimize mismatches. Assignment of gaps is dependent on the alignment program and model parameters chosen.
GeMoMa	GeMoMa uses conservation of **intron** positions across different species to improve the accuracy of **gene** predictions. It also uses sequence similarity to **proteins** in an **informant genome**, RNA-Seq **read** coverage, and **splice junction** predictions to improve the predictions of **protein-coding** genes in the **target genome**.
gene	basic unit of inheritance; contains the information needed to specify physical and biological traits Most genes code for specific **proteins**, or segments of proteins, which have differing functions within the body. Humans have approximately 20,000 **protein-coding genes**[1] and *Drosophila melanogaster* (fruit flies) have approximately 14,000.
gene duplication	The creation of a second copy of a sequence in a **genome**. A duplicate copy of a **gene** may be mutated without affecting the viability of the organism, so gene duplication is thought to be a significant factor in the **evolution** of genomic diversity (cf. duplication).

gene expression	Gene expression is the process by which the information encoded in a **gene** is used to either make RNA molecules that code for **proteins** or to make non-coding RNA molecules that serve other functions. Gene expression acts as an "on/off switch" to control when and where RNA molecules and proteins are made and as a "volume control" to determine how much of those products are made. The process of gene expression is carefully **regulated**, changing substantially under different conditions. The RNA and protein products of many genes serve to regulate the expression of other genes[1].
gene mapping	Gene mapping refers to the process of determining the location of **genes** on **chromosomes**. Today, the most efficient approach for gene mapping involves sequencing a **genome** and then using computer programs to analyze the sequence to identify the location of genes[1].
gene model	The description of the exact beginning and ending **coordinates** within the **genome** of every **exon** for a **gene** and how those exons are **assembled** into **mRNAs**; a gene model is the pattern and location of exons, **introns**, and other elements for an uncharacterized **target gene**.
Gene Model Checker	A program developed by **GEP** to test the validity of a proposed gene model after **annotation**. This program checks that the model complies with easily defined characteristics of a **gene** including the presence of start and **stop codons** and proper sequences at **intron**/exon **splice junctions**. It does not check for other important characteristics like overlap with evidence for **transcription** or degree of sequence conservation.
gene prediction	Gene prediction refers to the use of computational **algorithms** to identify regions of the genomic **sequence** that contains protein-coding **genes**. Many gene prediction algorithms are based on supervised machine learning algorithms such as hidden Markov models (HMMs) and Support Vector Machines (SVMs). *Ab initio* gene prediction algorithms such as Genscan use only the genomic sequence to predict the locations of protein-coding genes. More sophisticated gene prediction algorithms (e.g., Augustus, N-SCAN) incorporate extrinsic evidence (e.g., RNA-Seq data, protein sequence alignments, whole genome alignments) to improve the accuracy of the gene predictions and to predict multiple **isoforms**.
Gene Record Finder	A **GEP**-developed tool to provide information on *Drosophila melanogaster* **genes**, organized to describe how **isoforms** are related to each other (common and unique **exons**) and provide **protein** sequences corresponding to each exon.
gene regulation	Gene regulation is the process used to control the timing, location, and amount in which **genes** are expressed. The process can be complicated and is carried out by a variety of mechanisms, including through regulatory **proteins** and chemical modification of **DNA**. Gene regulation is key to the ability of an organism to respond to environmental changes[1].

genetic code	Genetic code refers to the instructions contained in a **gene** that tell a cell how to make a specific **protein**. Each gene's code uses the four nucleotide **bases** of **DNA**: **adenine** (A), **cytosine** (C), **guanine** (G) and **thymine** (T) — in various ways to spell out three-letter "**codons**" that specify which **amino acid** is needed at each position within a protein[1].
genetics	Genetics is the branch of biology concerned with the study of inheritance, including the interplay of genes, DNA variation and their interactions with environmental factors[1].
genome	A genome is all the genetic information an organism carries[4]. That is, all the nucleotides, and their order, found in a single cell, or organelle. The entire **complement** of genetic material of an organism. The genome is the entire set of **DNA** instructions found in a cell. A genome contains all the information needed for an individual to develop and function[1].
genome assembly	See "assembly"
genomic neighborhood	A local region of the **genome** containing a small number of **genes** (5-15) or other features (e.g., repeats). The genomic neighborhood, as related to the Pathways Project, is the region containing the **target gene** and its neighboring two closest **upstream** genes and two closest **downstream** genes. The genes flanking a particular target gene or other genomic element.
genomics	Genomics is a field of biology focused on studying all the **DNA** of an organism — that is, its **genome**. Such work includes identifying and characterizing all the **genes** and functional elements in an organism's genome as well as how they interact[1].
genomic variation / variants	Genomic variation refers to **DNA** sequence differences among individuals or populations. Some variants influence biological function (such as a **mutation** that causes a genetic disease), while others have no biological effects[1].
GEP	Genomics Education Partnership, an awesome (if underappreciated) group of biology educators dedicated to your successfully mastering the fundamentals of genomics!
GEP UCSC Genome Browser	A genome browser is a web-based visualization tool that allows users to examine the experimental data that provides evidence for a detailed **gene** structure. The GEP UCSC Genome Browser is a local mirror of the **UCSC** Genome Browser. It contains the reference sequence and working draft **assemblies** for many *Drosophila* **genomes** currently **annotated** by students participating in the **GEP**.

guanine (G)	Guanine (G) is one of the four nucleotide **bases** in **DNA**, with the other three being **adenine** (A), **cytosine** (C) and **thymine** (T). Within a double-stranded DNA molecule, guanine bases on one strand pair with cytosine bases on the opposite strand. The sequence of the four **nucleotide** bases encodes DNA's information[1].
hit	See "match"
homolog	A specific member of a group of **homologous** sequences or molecules; approximately 75% of human disease **genes** have homologs in *Drosophila melanogaster*.
homologous	In genomics, **nucleic acids** and **proteins** are homologous if they evolved from a common ancestor.
homology	Homology is the state of being **homologous**. Algorithms such as **BLAST** identify similarity, which is evidence for, but not necessarily proof of, homology. **Orthologs** and **paralogs** are subcategories of homology.
hypothesis	A hypothesis is an explanation for an observation which has not been confirmed (or rejected).
identity	Two elements at comparable positions in an **alignment** (a **base** or an **amino acid**) that are the same are said to be identical; the fraction of two sequences which consist of such elements is expressed as "percent identity."
in-frame stop codon	A nonsense **mutation** changes the **codon** for an **amino acid** within an **open reading frame** into a **stop codon**. This in-frame stop codon will usually lead to premature termination of **translation**, which results in a truncated non-functional **protein** product.
informant species	Genome sequences that are aligned to the **target species**' genome and used as auxiliary information for annotating the **target gene**. *Drosophila melanogaster* is the informant species for the Pathways Project annotations.
insertion	Type of **mutation** that involves the addition of **DNA** within a given sequence; this may occur because of **duplication** or insertion of foreign sequences such as transposable elements or viral DNA.
intron	**Non-coding** sections of a eukaryotic **nucleic acid** sequence found between **exons**. An intron is a region that resides within a **gene** but does not remain in the final mature **mRNA** molecule following **transcription** of that gene and does not code for **amino acids** that make up the **protein** encoded by that gene[1].
inversion	An inversion in a **chromosome** occurs when a segment breaks off and reattaches within the same chromosome, but in reverse orientation. **DNA** may or may not be lost in the process[1].
isoform	Potentially different versions of a **protein** encoded by a single **gene**. Isoforms result from **alternative splicing** of a particular **pre-mRNA**, and/or the use of a different **transcription start site**.

local alignment	An alignment where short, highly similar sequences are displayed.
local similarity	Regions within the query and subject sequences that show identical or **similar** sequences as detected by a local alignment algorithm (e.g., BLAST).
local synteny	In comparative genomics, synteny refers to genes that are located on orthologous chromosomes in two different species. Local synteny is focused only on a small portion of the chromosome surrounding the gene of interests, and whether the genes within the genomic neighborhood are in the same relative order and orientations in the two species. For example, the Pathways Project typically examines the two genes upstream and the two genes downstream of the gene of interests as part of the local synteny analysis.
locus (pl. loci)	A locus, as related to genomics, is a physical site or location within a **genome** (such as a **gene** or another **DNA** segment of interest), somewhat like a street address[1].
mapping	Mapping refers to the process of determining the relative locations of landmarks or markers (such as **genes** (**gene mapping**), variants, and other **DNA** sequences of interest) within a **chromosome** or **genome**. Historically, there have been two approaches for mapping: physical mapping, which established maps based on physical distances between landmarks, and genetic mapping, which established maps based on the frequency with which two landmarks are inherited together. Today, the most efficient approach for mapping involves sequencing a genome and then using computer programs to analyze the sequence to identify the locations of landmarks[1].
match/hit	A sequence in the BLAST database that shows significant alignment to the query sequence. In BLAST searches, a match consists of one or more High-scoring Segment Pairs (HSPs).
mature mRNA	Mature messenger RNA that has been completely processed and is ready for **translation**; it has a 7-methylguanosine cap at its 5' end, a **poly-A tail** at its **3'** end, and has all its **introns** spliced out.
mutation	A mutation is a change in the **DNA** sequence of an organism. Mutations can result from errors in DNA replication during cell division, exposure to mutagens, or a viral infection. Germline mutations (that occur in eggs and sperm) can be passed on to offspring, while somatic mutations (that occur in body cells) are not passed on[1].
NCBI	National Center for Biotechnology Information
NCBI RefSeq Genes track	A subset of **RefSeq**, RefSeqGene defines genomic sequences to be used as reference standards for well-characterized **genes**. It provides a 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[3].

negative strand	The negative strand of the **DNA** sequence of a single **gene** is the **complement** of the **positive strand**. The term loses meaning for longer DNA sequences with genes on both strands. Also called the anti-sense, template, or non-coding strand (cf. positive strand).
nested/nesting gene	A nested gene, or gene-within-a-gene, refers to a **gene** that is contained within another external host gene. Most often we find nested genes completely within an **intron** of, and in the opposite orientation to (i.e., **positive** vs. **negative strand**) its host gene. See Kumar (2009) for more information on nested genes.
Next-Generation DNA Sequencing	**DNA sequencing** establishes the order of the **bases** that make up **DNA**. Next-generation DNA sequencing (abbreviated NGS) refers to the use of technologies for sequencing DNA that became available shortly after the completion of the Human Genome Project (which relied on the first-generation method of Sanger sequencing). Faster and cheaper than their predecessors, NGS technologies can sequence an entire human **genome** in a single day and for less than $1,000[1]. See Part 2 of The Evolution of DNA Sequencing Technology series.
non-canonical	In molecular genetics/genomics, non-canonical typically refers to **intron** splicing site sequences that are only very rarely used. The canonical **splice donor site** dinucleotide sequence is GT; in rare cases, the non-canonical sequence GC is used instead (cf. canonical).
non-coding DNA	Non-coding DNA corresponds to the portions of an organism's **genome** that do not code for **amino acids**, the building blocks of **proteins**. Some non-coding DNA sequences are known to serve functional roles, such as in the **regulation** of **gene expression**, while other areas of non-coding DNA have no known function[1].
non-consensus	A **base** or sequence which does not match the most common element found at a given position (cf. consensus sequence).
non-redundant	Refers to the absence of identical components; many **databases** have the same sequences present multiple times, but non-redundant versions are searched to save time and computing resources.
novel isoform	An **isoform** present in the **target species** (e.g., *D. yakuba*) that is not present in the **informant species** (e.g., *D. melanogaster*).
N-SCAN PASA-EST track	**Gene models** predicted by N-SCAN and **transcripts** assembled by TransDecoder based on RNA-Seq **reads** were analyzed by the Program to Assemble Spliced Alignments (PASA) pipeline to improve the accuracy of the predicted gene models. The output of the PASA pipeline consists of predictions for multiple **isoforms** of a **gene** with the locations of the **5'** and **3' untranslated regions (UTRs)**. The PASA pipeline was originally designed to incorporate transcriptome evidence from **Expressed Sequence Tags (ESTs)**, and the pipeline was modified to analyze **RNA-Seq** data.

nucleic acids	Nucleic acids are large biomolecules that play essential roles in all cells and viruses. A major function of nucleic acids involves the storage and expression of genomic information. Deoxyribonucleic acid, or **DNA**, encodes the information cells need to make **proteins**. A related type of nucleic acid, called ribonucleic acid (**RNA**), comes in different molecular forms that play multiple cellular roles, including protein synthesis[1].
nucleotide	A nucleotide is the basic building block of nucleic acids (**RNA** and **DNA**). A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing **base**. The bases used in DNA are **adenine** (A), **cytosine** (C), **guanine** (G) and **thymine** (T). In RNA, the base **uracil** (U) takes the place of thymine. DNA and RNA molecules are polymers made up of long chains of nucleotides[1].
Open Reading Frame (ORF)	An open reading frame, as related to genomics, is a **reading frame** that does not include a **stop codon** in a portion of a **DNA** sequence (e.g., **coding exon**). A long ORF is often part of a **gene** (that is, a sequence directly coding for a **protein**) [1].
ortholog (orthologous)	**Genes** in different species that derive from a single ancestral gene in the last common ancestor of the respective species. Orthologous genes have a shared ancestry of a character or genetic region due to recent **speciation**[3].
orthology	**Genes** in different species that derive from a common ancestor, i.e., they are direct evolutionary counterparts[3].
paralog (paralogous)	A paralog is one of a set of **homologous genes** that have diverged from each other because of **gene duplication**. For example, the mouse α-*globin* and β-*globin* genes are paralogs. The relationship between mouse α-*globin* and chick β-*globin* is also considered paralogous[3]. Paralogous genes are at different chromosomal locations in the *same* organism and have structural similarities indicating that they derived from a common ancestral gene, but they have since diverged from the parent copy by **mutation** and selection or drift.
paralogy	Describes the relationship of **homologous genes** that arose by **gene duplication**[3].
parsimony	Using the simplest assumptions to explain a result.
peptide	A peptide is a short chain of **amino acids** (typically 2 to 50) linked by chemical bonds (called peptide bonds). A longer chain of linked amino acids (51 or more) is a ***polypeptide***. The **proteins** manufactured inside cells are made from one or more polypeptides[1].
percent identity	See “identity”

phase	The phase describes the number of **bases** between the end of the **exon** (defined by the splice site) and the full **codon** nearest that splice site. The number of bases between the adjacent full codon and an exon/splice site can be 0, 1 or 2; this number is the phase. The phase of an **upstream** exon will determine which **frame** is translated in the **downstream** exon by indicating how many bases after the **splice acceptor site** are needed to create a full codon of 3 bases.
poly-A tail	About 250 **adenine** (A) **nucleotides** that are post-transcriptionally added by poly(A) polymerase to the **3'** end of eukaryotic **transcripts**, following cleavage of the newly synthesized **RNA** ~20 nucleotides **downstream** of an AAUAAA polyadenylation signal sequence.
polymorphism	The presence of two or more variant forms of a specific DNA sequence that can occur among different individuals or populations. The most common type of polymorphism involves variation at a single nucleotide (also called a single-nucleotide polymorphism, or SNP). Other polymorphisms can be much larger, involving longer stretches of DNA[1].
polypeptide	**Amino acid** chain containing more than 50 amino acids joined together by peptide (amide) bonds.
positive strand	In a **gene**, the **DNA** strand that has the sequence found in the **RNA** molecule; also called the sense, positive, or non-template strand.
pre-mRNA (primary transcript)	The initial transcript from a **protein-coding gene** that contains both **introns** and **exons**. Pre-mRNA requires the addition of a **5'** cap and **3' poly-A tail** and the removal of introns to produce the final mature **mRNA** molecule containing joined exons.
processed mRNA	See “mature mRNA”
protein	Proteins are large, complex molecules that play many important roles in the body. They are critical to most of the work done by cells and are required for the structure, function and regulation of the body’s tissues and organs. A protein is made up of one or more long, folded chains of **amino acids** (each called a **polypeptide**), whose sequences are determined by the **DNA** sequence of the **protein-coding gene**[1].
protein-coding gene	Any **gene** whose ultimate biologically functional product is a **protein**.
pseudogene	A pseudogene is a segment of **DNA** that structurally resembles a **gene** but is not capable of coding for a **protein**. Pseudogenes are most often derived from genes that have lost their **protein-coding** ability due to accumulated **mutations** that have occurred over the course of **evolution**[1].
putative	Something that may be predicted or inferred but that requires more evidence to confirm or refute.

query	The input sequence (or other type of search term) with which all of the entries in a **database** are to be compared (i.e., sequence to match).
read	**RNA-Seq** reads (100–125 bp in length) are derived primarily from processed **mRNA** (i.e., after the **introns** have been removed). Hence, genomic regions with RNA-Seq read coverage usually correspond to transcribed **exons**, which include both the translated and **untranslated regions**.
reading frame	See "frame"
RefSeq	NCBI's Reference Sequence (RefSeq) project, is a **non-redundant**, **annotated** set of sequences that serve as reference standards. They are derived from the International Nucleotide Sequence Database Collaboration (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**[3]. See O'Leary et al. (2015) for a detailed overview.
regulation	See "gene regulation"
regulatory elements	**DNA** sequences that control **gene expression** by binding to **proteins** which increase or decrease synthesis of the **gene** product.
RepeatMasker track	A program that screens **DNA** sequences for interspersed repeats and low complexity DNA sequences.
repetitive element	A specific sequence of nucleotides that is repeated, sometimes many times, within our genetic material. See "Cartoon Explainer" for a more detailed overview.
RNA	Ribonucleic acid (abbreviated RNA) is a **nucleic acid** present in all living cells that has structural similarities to **DNA**. Unlike DNA, however, RNA is most often single-stranded. An RNA molecule has a backbone made of alternating phosphate groups and the sugar ribose, rather than the deoxyribose found in DNA. Attached to each sugar is one of four **bases**: **adenine** (A), **uracil** (U), **cytosine** (C) or **guanine** (G). Different types of RNA exist in cells: messenger RNA (**mRNA**), ribosomal RNA (rRNA) and transfer RNA (tRNA). In addition, some RNAs are involved in regulating **gene expression**. Certain viruses use RNA as their genomic material[1].
RNA-Seq	A technique which sequences RNA rather than **DNA**. It is most often used (in **annotation**) to identify genomic regions which are transcribed and present in **mRNA**, and thus possible exon **loci**.
scaffold	A scaffold is an ordered and oriented set of **contigs**[3].
sequence	For **DNA** and **RNA**, the sequence refers to the order of the chain of **nucleotides** in the molecule (A, T, G, C, U). For **proteins**, the sequence refers to the order of the **amino acids** in the polypeptide.
sequence alignment	See "alignment"

sequenced	For DNA and RNA, the use of sequencing techniques to determine the order of the nucleotides in the molecule. For proteins, the use of sequencing techniques to determine the order of the amino acids.
shotgun sequencing	Shotgun sequencing is a laboratory technique for determining the **DNA** sequence of an organism's **genome**. The method involves randomly breaking up the genome into small DNA fragments that are sequenced individually. A computer program looks for overlaps in the DNA sequences, using them to reassemble the fragments in their correct order to reconstitute the genome[1]. This strategy for sequencing whole genomes, pioneered by the for-profit company Celera, is prone to **assembly** errors.
significance	In the context of sequence alignments, statistical significance refers to the probability (p-value) of obtaining the observed alignment score or better under the null hypothesis that the query and subject sequences are unrelated to each other. If the p-value derived from the alignment score is smaller than the significance level (**α**) defined by the study, then the observed alignment is considered to be statistically significant. The significance level is defined as the probability of rejecting the null hypothesis when the null hypothesis is true (i.e., Type I error).
similarity	The definition of sequence "similarity" differs among the different alignment algorithms and the scoring systems. For example, the *blastp* program considers pair of aligned residues that have a positive score in the scoring matrix (e.g., BLOSUM62) to be "similar" (i.e., "positives").
simple repeat	A **nucleotide** repeat, with one or a small number of **bases**, such as AAAAAAAAAAA or CACACACACA.
SNP	Single-**nucleotide** polymorphism; a difference in **DNA** sequence at a single **base** between two sequences.
***Spaln* Alignment of *D. melanogaster* Proteins track**	**GEP UCSC Genome Browser** track showing the results of a *spaln* alignment, which here compares *D. melanogaster* **protein** sequences to the target **genome**; it can be thought of as similar to (though distinct from) ***tblastn***.
speciation	Speciation is a lineage-splitting event that produces two or more separate species. See "Defining speciation" for an example[4].
splice acceptor site	The **splicing** site at the **3'** end of an **intron**, at the boundary between an intron and the **exon** immediately **downstream** of the intron. The **canonical** splice acceptor site dinucleotide sequence is AG.
splice donor site	The **splicing** site at the **5'** end of an **intron**, at the boundary between an intron and the **exon** immediately **upstream** of the intron. The **canonical** splice donor site dinucleotide sequence is GT; in rare cases, the **non-canonical** sequence GC is used instead.
splice junction	Either a **splice acceptor site** or a **splice donor site**.
splicing	To cut **introns** out of an RNA **transcript** and rejoin the RNA molecule[4].

	The process by which introns are removed and **exons** are joined to produce a mature, functional RNA (**mRNA**) from a **primary transcript**. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.
spurious	Likely occurred by chance alone and, therefore, is not evidence of real biological conservation.
start codon (initiation codon)	The first **codon** of a **CDS**. In eukaryotes this is almost always ATG, which codes for methionine (one of the 20 **amino acids**).
start site	The **nucleotide** at which **transcription** starts, usually denoted as position +1 in reference to the **gene** being transcribed.
stop codon (termination codon)	A stop codon is a sequence of three **nucleotides** (a trinucleotide) in **DNA** or messenger RNA (**mRNA**) that signals a halt to synthesis of that copy of a **(poly)peptide** in the cell. There are 64 different trinucleotide **codons**: 61 specify **amino acids** and 3 are stop codons (i.e., UAA, UAG and UGA)[1]. Since stop codons don't specify any amino acid, they are sometimes referred to as "nonsense codons."
subject	The sequence, typically retrieved from a **database**, to which the sequence of interest (the "**query**") is being compared (i.e., search subject for a match to the query).
synteny	The order and orientation of **genes** in a given chromosomal region; two regions are said to be syntenic if all genes are the same, in the same order, and in the same orientation; conservation of local **genomic neighborhood**.
target gene	The target gene, as related to the Pathways Project, is the **gene** (e.g., *Rheb*) your instructor assigned you to **annotate** in a specific **target species**.
target species	The target species, as related to the Pathways Project, is the species (e.g., *D. yakuba*) in which your instructor assigned you to **annotate** the **target gene**.
tblastn	**BLAST** search tool in which the **query** is an **amino acid** sequence, and the **subject** is the six amino acid sequences translated from the six **frames** found in double stranded **DNA**. Typically used when using a **protein** sequence to search a nucleotide **database**. Query: protein Database/Subject: nucleotide → protein Function: identifying database sequences encoding proteins similar to query Common Use Cases: map proteins against genomic assemblies
tblastx	**BLAST** version in which the **query** is all 6 possible **amino acid** sequences derived from **translation** of all 6 **frames** and the **subjects** are the 6 possible amino acid sequences derived from translation of all 6 frames of another **nucleotide** sequence. Not surprisingly, this is computationally very expensive. Query: nucleotide → protein

	Database/Subject: nucleotide → protein Function: identifying nucleotide sequences similar to the query based on their coding potential Common Use Cases: identify genes in unannotated sequences
telomere	A telomere is a region of repetitive DNA sequences at the end of a chromosome. Telomeres protect the ends of chromosomes from becoming frayed or tangled. Each time a cell divides, the telomeres become slightly shorter. Eventually, they become so short that the cell can no longer divide successfully, and the cell dies[1]. Unlike other eukaryotes, *Drosophila* telomeres consist of an array of HeT-A, TART, and TAHRE retrotransposons (HTT array) and a chromosome cap (reviewed in Mason et al., 2008).
thymine (T)	Thymine (T) is one of the four nucleotide **bases** in **DNA**, with the other three being **adenine** (A), **cytosine** (C) and **guanine** (G). Within a double-stranded DNA molecule, thymine bases on one strand pair with adenine bases on the opposite strand. The sequence of the four **nucleotide** bases encodes DNA's information[1].
track	See "evidence track"
transcript	The RNA molecule which is the immediate product of **transcription** of a **gene**, which is often modified before becoming fully functional.
transcription	The process of copying one strand of a **DNA** double helix by RNA polymerase, creating a complimentary strand of RNA called the **transcript**. Transcription, as related to **genomics**, is the process of making an RNA copy of a **gene's** DNA sequence. This copy, called messenger RNA (**mRNA**), carries the gene's **protein** information encoded in DNA. In humans and other complex organisms, mRNA moves from the cell nucleus to the cell cytoplasm (watery interior), where it is used for synthesizing the encoded protein[1].
translation	The process by which **codons** in an **mRNA** are used by the ribosome to direct **protein** synthesis. Translation, as related to **genomics**, is the process through which information encoded in messenger RNA (mRNA) directs the addition of **amino acids** during protein synthesis. Translation takes place on ribosomes in the cell cytoplasm, where mRNA is read and translated into the string of amino acid chains that make up the synthesized protein[1].

TSS (transcription start site)	The location in **DNA**, generally **upstream** of a **gene's** coding sequence, where RNA polymerase begins **transcription**.
UCSC	University of California Santa Cruz; host to a popular genome browser.
unique isoform	Multiple **isoforms** of a **gene** could differ only in their **untranslated regions** and encode the same **polypeptide** sequence. The term "unique isoforms" refers to the subset of isoforms that encode different polypeptide sequences of a gene. For example, if isoforms A and B use **coding exons** 1, 3, and 4, and isoform C uses coding exons 1, 2 and 4, then either the combination of A and C or the combination of B and C would constitute the set of unique isoforms for the gene.
upstream	Toward the **5'** end of a single stranded length of **DNA** or **gene** of interest (e.g., **target gene**). Upstream also refers to the genomic region prior to the **feature** being examined (cf. downstream).
uracil (U)	Uracil (U) is one of the four nucleotide **bases** in RNA, with the other three being **adenine** (A), **cytosine** (C) and **guanine** (G). In RNA, uracil pairs with adenine. In a **DNA** molecule, the nucleotide **thymine** (T) is used in place of uracil[1].
UTR (Untranslated Region)	"Untranslated region"; a segment of **DNA** (or RNA) that is transcribed and present in the mature **mRNA** but is not translated into **protein**. UTRs may be found at either or both the **5'** and **3'** ends of a **gene** or transcript.
variants / variation	See "genomic variation"

References:

1. Courtesy: National Human Genome Research Institute (NHGRI).
2. Anne Rosenwald, Gaurav Arora, Vinayak Mathur (2020). Lesson III - Annotation. Genome Solver, (Version 3.0). QUBES Educational Resources. doi:10.25334/JEY1-0927
3. The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK143764/
4. UC Museum of Paleontology Understanding Evolution, understandingevolution.org.