Topic	Definition
3'	Refers to the third carbon of the nucleic acid sugar moiety to which additional nucleotides may be added by polymerase, often used to refer to that end of a single-stranded DNA or RNA molecule where the 3' carbon is unattached to an adjacent nucleotide; *cf.* 5'.
454 sequencing	A large-scale parallel pyrosequencing system capable of sequencing roughly 400-600 megabases of DNA per 10-hour run. The technology is known for its relatively unbiased sample preparation and moderately long, highly accurate sequence reads (~400 pairs in length).
5'	Refers to the fifth carbon of the nucleic acid sugar moiety, to which the triphosphate is attached in a nucleotide triphosphate, often used to refer to that end of a single-stranded DNA or RNA molecule where the 5' carbon's phosphate group is unattached to an adjacent nucleotide; *cf.* 3'.
ab-initio	In computing, *ab initio* is a term used to define computations based soley on theory or using only fundemental 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.
accesion number	A unique identification number given to every DNA, RNA, and protein sequence submitted to NCBI or equivalent databases. For example, the human leptin receptor's accession number is P48357 in the SwissProt database.
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.
alpha-satellite sequence	A family of tandemly repeated DNA sequences present in the chromosomes of many higher eukaryotes, believed to help maintain the structure and function of the centromere.
alternative splicing	The inclusion or exclusion of certain exons in the splicing reactions that determine the sequences included in the final mRNA product. This mechanism is utilized to generate a series of closely related protein isoforms, which differ by the inclusion or exclusion of the particular protein domains encoded by those exons. Alternative splicing is directed by RNA-binding proteins that block or stimulate utilization of a particular splice site.
ALU sequences	A set of dispersed, highly abundant related sequences, each about 300 bp long, in primate genomes; each Alu element is a retrotransposon. The individual members have a characteristic AluI restriction enzyme cleavage site
amino acid	The basic building block of proteins, a small molecule with a -C-C- core, an amino group at one end and a carboxylic acid group at the other end. The basic structure can be represented as NH2-CHR-COOH, where R can be any of 20 different moieties, including acidic, basic, or hydrophobic groups.
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.

antisense strand	Also called the negative, template, or non-coding strand. This strand of the DNA sequence of a slngle gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.
Augustus	A particular program used in molecular biology; Augustus is an ab-initio single isoform gene finder
BAC	Bacterial Artificial Chromosome, a cloning vector which can accept a large insert of foreign DNA (1000,000 to 2000,000 bp) and can be propagated as a chromosome in bacteria. In addition to the cloning site, the BAC contains selectable markers and a bacterial origin of replication which maintains one copy of the BAC per cell.
base	Although formally incorrect (the nitrogenous base which gives each nucleotide its name is only part of the nucleotide), this is often used as a synonym for "nucleotide."
base pair (base pairing)	The hydrogen bonding of one of the bases (A, C, G, T, U) with another, as dictated by the optimization of hydrogen bond formation in DNA (A-T and C-G) or in RNA (A-U and C-G). Two polynucleotide strands, or regions thereof, in which all the nucleotides form such base pairs are said to be complementary. In achieving complementarity, each strand of DNA can serve as a template for synthesis of its partner strand- the secret of DNA replication's extremely high accuracy and thereby of inheritance.
blast	Basic Local Alignment Search Tool, an algorithm used to detect local similarity between biological sequences (either nucleic acid or protein), typically used to search a large database of sequences with a single query sequence. For one implementation, see http://www.ncbi.nlm.nih.gov/BLAST.
blast2	BLAST version in which two sequences can be compared to each other. The sequences could either be nucleic acid or protein. To do this at NCBI, go to the BLAST web page and select "Align two or more sequences."
blastn	BLAST version in which the query and subject are both nucleotide sequences. Typically used to search a nucleotide database with a nucleotide sequence. Use the link http://www.ncbi.nlm.nih.gov/BLAST and select "blastn."
blastp	BLAST version in which both the query and subject are amino acid sequences. Typically used to search a protein database with a protein sequence. Use the link http://www.ncbi.nlm.nih.gov/BLAST and select "blastp."
blastx	BLAST version in which the query is nucleotide sequence and all 6 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. Use the link http://www.ncbi.nlm.nih.gov/BLAST and select "blastx."
Blosum matrix	This is a matrix which gives a numerical value to the substitution of one amino acid for another when doing protein comparisons. It is based on the observed levels of amino acid substitution when comparing closely related but divergent blocks of protein sequences. The BLOSUM number refers to the percentage of identical amino acids found within the blocks used to generate the matrix. For example, BLOSUM62 uses blocks of protein sequences with no more than 62% identity.
C-terminal	Refers to the end of a protein that contains the carboxylic group -COOH, corresponding to the 3' end of the encoding gene. Also called “carboxy-terminal.”
canonical site	In molecular genetics/genomics, this typically refers to intron splicing sites. The vast majority of introns have the nucleotides GT (“splice donor site”) at their 5’ ends and AG (“splice acceptor site”) at their 3’ ends. These are the canonical sites. Variants are rare, and are called “non-canonical sites.”

cDNA	"complementary DNA," a double-stranded DNA molecule prepared *in vitro* by copying an RNA molecule back into DNA using reverse transcriptase. The RNA component of the resulting RNA-DNA hybrid is then destroyed by alkali, and the complementary strand to the remaining DNA strand synthesized by DNA polymerase. The resulting double-stranded DNA can be used for cloning and analysis.
CDS	Coding sequence, that part of the DNA sequence of a gene which is translated into protein.
chromosome	One molecule of double-stranded DNA, carrying an arrangements of genes interspersed with other sequences. In prokaryotes the chromosome is often a circle of DNA, while in eukaryotes chromosomes are typically linear, extending from one end, a telomere, through the chromosome center to the other telomere.
cleavage	An enzymatic or chemical breakage of the covalent bond that joins two nucleotides or two amino acids in their respective polymers
CLUSTALW	A program that produces global sequences aligments, typically with multiple related sequences.
coding exon	In a gene, any exon which contains some part of the CDS; in contrast, an exon which has no part translated into protein is called a "non-coding exon."
coding 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.
codon	The sequence of three nucleotides in DNA or RNA that specifies a particular amino acid.
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.
Consed	A program used to view and edit the DNA sequence data assembled by Phred/Phrad or other base-calling and assembly programs.
consensus	When comparing multiple sequences, whether by Phred/Phrap during sequence assembly or algorithms such as ClustalW, the sequence which reflects the most commonly seen base at each position.
contig	A contiguous assembly, without gaps, of overlapping sequences based on regions of shared sequence
coordinates	Numerical position within a biological sequence, e.g. the first base in a DNA sequence would have the coordinate 1.
cosmid	A type of hybrid plasmid (often used as a cloning vector) that contains cos sequences, DNA sequences originally from the Lambda phage. Cosmids are maintained in E. coli as multi-copy episomes, and are typically used to build genomic libraries *cf.* fosmid.
CpG island	A region of nucleotides, typically 300-3000 bp in length, which has a higher than expected frequency of the sequence C followed by G, usually found in or near promoter sequences in organisims with significant amounts of DNA methylation. Exact mathematically-based definitions vary among published studies.
cryptic site	A site which does not match an authentic splice site, whether donor or acceptor, but which upon mutation to a consensus splice site, is used by the cell's splicing machinery and thereby causes incorrectly-spliced products to be made.
***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.
degenerate sequence	A sequence in which one symbol can represent multiple possibilities. The genetic code is said to be degenerate because most amino acids are encoded by multiple codons. In DNA sequence a degenerate code allows for a single symbol to designate more than one possible base, e.g. B stands for C, G or T.

deletion	The removal of some part of a biological sequence.
dideoxynucleotide	[I agree with Chris's version, adding a period at the end and making the initial term singular: "A nucleotide"]
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.
DINE	A particular family of transposable elements found in Drosophila genomes.
dissociation	The separation of double stranded DNA into single strands by raising the temperature or the pH.
DNA affinity chromatography	A procedure in which a protein is separated from a mixture of other proteins by its ability to bind specifically to a particular sequence of DNA that has been immobilized on a matrix in a separation column.
DNA binding domain	The region of a protein that can specifically bind to DNA. Several motifs that can bind to DNA have been characterized (e.g. helix-loop-helix, leucine zipper, and zinc-finger domains). In most cases, the structure of the domain has evolved so that a portion of it interacts with a specific sequence of DNA by binding in the major groove of the DNA molecule.
DNA hypersensitive site	A region along the chromatin fiber that is nucleosome-free, and hence much more accessible to cleavage by DNase I or other nucleases. Such sites are usually found at the promoters and enhancers of active or inducible genes.
DNA ligase	An enzyme that joins pieces of DNA by catalyzing the formation of a covalent phosphodiester bond between the 5' phosphate end of one nucleotide and the 3'-OH group of the adjacent nucleotide.
DNA Pol (polymerase)	A group of enzymes capable of extending a strand of DNA by adding successive deoxyribonucleotides at the 3' end in the order dictated by an associated template strand; the basic biochemical reaction it catalyzes is: (dNMP)n + dNTP ----> (dNMP)n+1 + PPi
DNAase I footprinting	A means of determining the exact binding site of a protein on a specific DNA fragment. When a protein-DNA complex is digested with DNase I, the bound protein will protect the DNA from cleavage at the site of the protein-DNA complex. A comparison of the cleavage fragments from DNA with and without the protein bound to it will indicate the region of protein binding by the absence of cleavage.
dot chromosome	The fourth chromosome of Drosophila melanogaster which is comprised principally of heterochromatin; any very small chromosome which appears as a "dot" during mitosis.
Dot Matrix View	The comparison of two sequences on an X-Y plot where points ("dots") on the graph ("matrix") 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. "Upstream" similarly refers to something closer to the 5' end.
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.
E-value	Expected value, a numerical indication of the statistical significance of an alignment. Describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The lower the E value, the more significant the alignment. For small E-values, will be nearly identical to the probability of seeing an alignment with this quality purely by chance. See http://www.ncbi.nlm.nih.gov/books/NBK21106/

electrophoresis	A procedure to separate molecules in an electric field. DNA migration takes place through a gel matrix (agarose or polyacrylamide) that acts as a molecular sieve. Since DNA is negatively charged, when placed in an electric field, it is attracted toward the positive electrode. Because the negative charges are uniformly distributed along the DNA, molecules separate in the field based on their sizes.
end labeling	The addition of a radioactively labeled group to one end (5' or 3') of a DNA strand.
endonuclease	An enzyme that cleaves phosphodiester bonds within a nucleic acid chain. A particular endonuclease may be specific for RNA or for single-stranded or double-stranded DNA. Restriction enzymes are endonucleases that cut double-stranded DNA at or near a specific target sequence, such as CGCG.
enhancer	A eukaryotic DNA sequence located outside of the promoter region, where an activator of transcription (protein) may bind.
ENSEMBL	Joint genome browser maintained by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. Contains searchable genomic information for select model organisms.
environmental containment	Process for ensuring that an organism (such as a host cell) cannot survive long in the environment, based on its requirements for survival and growth, to prevent its uncontrolled proliferation. It is used as a safety precaution when growing recombinant DNA clones with potentially hazardous properties (for example, when cloning a gene encoding a neurotoxin).
enzyme	Enzymes are proteins that serve as biological catalysts. Once thought to be the exclusive domain of proteins there are now known to be catalytic RNAs called ribozymes.
EST	Expressed sequence tag: a short DNA sequence derived from a single read of a clone from a cDNA library. They are therefore by definition from genes which are expressed in that cell type under the growth conditions used. Typically ESTs have high levels of sequence error because they are from single reads, and not the consensus of multiple reads.
ethidium bromide	An intercalating chemical which is used to stain DNA; it is fluorescent under ultraviolet light. EtBr is a mutagen; use gloves and other appropriate protection when handling this material.
euchromatin	Those regions of the genome which do not remain condensed throughout the cell cycle. Euchromatic regions are typically enriched for genes, show significantly higher levels of recombination and lower levels of repeats than heterochromatic regions. *cf.* "heterochromatin."
eukaryote	The class of organisms composed of one or more cells, each of which contains a membrane-enclosed nucleus and packages its DNA with histones in a nucleosome array. Eukaryotic cells typically have other complex organelles, such as mitochondria.
exon	An exon is a contiguous segment of eukaryotic DNA that corresponds to a portion of the mature (processed) RNA product of that gene. Exons are found only in eukaryotic genomes, and are separated by introns. Although the introns are transcribed with the exons, the latter are spliced out and discarded during RNA processing.
exonuclease	An enzyme that cleave nucleotides one at a time from the end of a polynucleotide chain. A particular exonuclease may be specific for either the 5' or 3' end of either DNA or RNA. The 3'-to-5' exonuclease activity of DNA pol I and DNA pol III allows these enzymes to excise the nucleotide that they just added if it base pairs incorrectly ("editing").

FASTA	A text format used to represent nucleotide or protein sequences. FASTA-formatted sequences begin with a rightward-pointing angle bracket or chevron (>) after which follows information about the sequence on the same line. Line two begins the actual sequence of interest. Variations of FASTA format (e.g. Pearson FASTA and NCBI FASTA) exist, but may not be handled gracefully or even correctly by a given program, so care should be taken to use the desired version of FASTA as appropriate.
feature	Part of a GenBank entry containing information about a nucleotide sequence of interest. Features include but are not limited to the length of the sequence, predicted positions of promoters, ribosome binding sites, protein coding regions (CDS), and translation products.
FlyBase	A database of Drosophila genes and genomes, on the web at http:// flybase.org
fosmid	Cloning vectors based on bacterial F-plasmids. Fosmids are used in cloning genomic libraries, typically with insert sized of ~40 kb and are maintained in E. coli as a stable single copy episome; *cf.* "cosmid."
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 3 possible reading frames in an mRNA strand and six in a double stranded DNA molecule due to the two strands from which transcription is possible. Different computer programs number these frames differently, particularly for frames of the negative strand, so care should be taken when comparing designated frames from different programs.
gander	Server at GEP that houses a copy of the University of California San Diego genome browser filled with GEP-specific projects.
gaps	When comparing or aligning two or more protein or nucleic acid sequences, spaces ("gaps") may be introduced in the final alignment in order to maximize matches and minimize mismatches. Assignment of gaps is dependent on the alignment program and model parameters chosen.
GC rich region	A region of DNA with higher than expected fraction of CG basepairs. In some organisms GC-rich regions are found in or near promoters; *cf.* CpG island.
gel shift assay	Also known as electrophoretic mobility shift assay (EMSA), a gel shift assay is a means of detecting DNA-protein interactions. A complex between a fragment of DNA and a protein moves more slowly during gel electrophoresis than does the DNA fragment alone, resulting in a "shift" in the position of the DNA fragment in the presence of the protein. See Molecular Cell Biology, Lodish et.al. figure 10-7
Gene Model Checker	A program developed by GEP to test the validity of a proposed gene model after annotation. This programs 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 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.
GeneID	A number that identifies a protein with a known function at JCSG (Joint Center for Structural Genomics).
genome	All of the nucleotides, and their order, found in a single cell, or organelle. The entire complement of genetic material of an organism. "Haploid genome" refers to one copy of each chromosome in a diploid organism.
genscan	An ab initio gene prediction program developed at MIT (genes.mit.edu/GENSCAN.html).

GEP	Genomics Education Partnership, based at Washington University, St. Louis. See http://gep.wustl.edu/
gff	Gene Feature Format; a particular format for text files used to describe genomic annotations. For specifications see http://www.sanger.ac.uk/resources/software/gff/spec.html. Many genome browsers allow data to be imported using data encoded in GFF format.
global alignment	An alignment of two sequences over their entire lengths (note that the alignment need not be perfect).
golden tiling path	A method created by the Human Genome Project (HGP) sequencing labs which uses mapping markers to choose the minimum number of slightly overlapping clones that completely span a genomic region of interest.
handedness	A helix is characterized as right-handed if it is turning clockwise as it moves away from you; if it turns counter-clockwise, it is left-handed.
helicase	An enzyme that uses the energy of ATP hydrolysis to disrupt the hydrogen bonds that hold the two strands of DNA together, allowing the double helix to unwind.
heterochromatin	DNA that remains condensed throughout the cell cycle, heterochromatin is thought to be tightly bound to proteins and other molecules. Heterochromatic regions tend to have a high content of repetitious DNA (satellite DNA, middle repetitious sequences), are gene-poor, show little or no transcriptional activity, and replicate late in S-phase. Blocks of heterochromatin are generally found around the centromeres and telomeres; *cf.* euchromatin.
histones	The small, basic proteins used to package the DNA in chromatin in eukaryotes. The core histones (H2A, H2B, H3, and H4) are highly conserved throughout all eucharyotes, while histone H1 is more variable.
HLH (Helix-Loop-Helix or helix-turn-helix)	A ca. 20-amino acid protein motif that forms two alpha-helices that cross at an angle of ~120°. This motif occurs frequently in DNA binding proteins, with one of the helices (stabilized by the other) forming contacts with the DNA in the major groove.
homologous	Nucleic acids and proteins are homologous if they have evolved from a common ancestor.
homologue	A specific member of a group of homologous sequences or molecules.
homology	Homology is the state of being homologous. Algorithms such as BLAST identify similarity which is evidence for, but not necessarily proof of, homology.
host-vector system	The combination of a particular plasmid, phage, or virus and the bacterium or other host cell in which it is propagated. It is essential that the host cell be transformed by the vector DNA at a reasonable frequency, and that the vector have an appropriate origin of replication to function in the host cell.
hybridization or annealing	The process whereby complementary single strands of DNA come together to form a double helix. If one strand is labeled, it can be used to identify the second, unlabeled strand. For example, in filter hybridization, the labeled probe will hybridize to the unlabeled complementary DNA that is stuck to the filter.
hydrogen bond	A noncovalent bond between an electronegative atom (such as N or O) and a hydrogen atom covalently bonded to another electronegative atom. Hydrogen bonds are individually weak, but collectively contribute significantly to the stabilization of the DNA double helix. The pairing required to form optimal hydrogen bonds between DNA nucleotides underlies the principle of complimentarity.
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 silico	Performed on a computer.
In vitro	Performed in the absence of intact cells; “*in vitro*” literally means “in glass.”
In vivo	Occurring in living cells; *cf.* in vitro.

in-frame	Referring to something which does not alter the coding frame of a gene; *cf.* frame.
***in-situ* hybridization**	A technique performed by denaturing the DNA of cells or tissue sections and adding a single-stranded DNA probe. The probe is labeled so that the site of hybridization can be detected by autoradiography or other appropriate detection protocols.
indel	Shorthand term to designate a gap in an alignment which designates "either an insertion or deletion". Typically used when the historical event that created the difference between two sequences cannot be determined.
induced mutation	A change in a DNA sequence caused by exposure of the DNA to a mutagen.
initiation	The process in which DNA or RNA polymerase binds to a DNA strand to begin copying it.
initiation codon	The first codon of a coding sequence. In eukaryotes this is almost always ATG, which codes for Methionine.
initiator	A weak consensus sequence [PyPyAN(T/A)PyPyPy] in eukaryotes, found with the A at position +1 of the gene, which serves as a recognition sequence for RNA polymerase II.
insertion	The addition of DNA within a given sequence; this may occur as a result 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. Introns are removed (“spliced out”) of mRNA after transcription and before the molecule is exported to the cytoplasm for translation; *cf.* exon.
isoform	Alternate forms of a specific protein with slightly different amino acid sequences. Often different isoforms are produced by alternative splicing of a particular mRNA.
LINE	Long Interspersed Nuclear Elements are a class of retrotransposon commonly found in eukaryotic genomes.
local alignment	An alignment where short, highly similar sequences are displayed; *cf.* global alignment.
low complexity DNA	DNA segments that have particularly simple sequences, such as mononucleotide runs (AAAAAAAA) or dinucleotide repeats (ATATATATATAT).
LTR	Long Terminal Repeats; DNA sequences present in many contiguous copies (up to several kilobases in total) found at the end of a class of retrotransposons.
mature mRNA	Messenger RNA that has been completely processed; 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 from it.
mRNA	Messenger RNA, a kind of RNA that is translated into a polypeptide.
N-terminal	Refers to the end of a protein that contains the amino group -NH2, corresponding to the 5' end of the encoding gene. Also called “amino-terminal.”
NCBI	National Center for Biotechnology Information. On the web at http://www.ncbi.nlm.nih.gov
non-cannonical	In molecular genetics/genomics, this typically refers to intron splicing site sequences that are only very rarely used and are never considered by gene prediction algorithms; *cf.* “canonical site”.
non-coding RNA	It is an RNA molecule that functions without being translated into protein for example, transfer RNA and ribosomal RNA; note that this is not the same thing as the “non-coding strand”.
non-coding strand	Also called the negative, template, or anti-sense strand. This strand of the DNA sequence of a slngle gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.
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.
nucleoside	A small molecule made up of either a purine or pyrimidine base linked to a pentose (sugar), generally either ribose or deoxyribose.
nucleotide	A nucleoside linked to one or more phosphate groups via an ester bond with the pentose. DNA and RNA are polymers of nucleotides, linked through the 5' and 3' carbons of the sugar.
ORF	Open Reading Frame, a long stretch of codons in the same reading frame uninterrupted by stop codons; an ORF may reflect the presence of a gene.
orthologous genes	Genes in different organisms that are direct evolutionary counterparts; that is, they are related by descent from a common ancestor. Orthologous genes normally have the same cellular function.
P-element	A Drosophila transposable element that has been used as a tool for insertion mutagenesis and for germline transformation.
paralogous genes	Genes at different chromosomal locations in the same organism that have structural similarities indicating that they derived from a common ancestral gene and have since diverged from the parent copy by mutation and selection or drift
PCR	Polymerase Chain Reaction, a method for creating billions of copies of a specific DNA segment from a complex mixture in vitro, using short oligonucleotide primers, that exploits certain features of DNA replication.
pep	The suffix of a protein/peptide sequence file.
phase	The phase describes the relationship between the translation frame of an exon and the position of a splice junction. In the GEP we define the term to describe 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 at an exon/site juction can be either 0, 1 or 2. The phase of an exon/splice-donor junction will determine which frame is translated in the downstream exon as it will indicate how many bases are used after the acceptor splice site to create a full codon of 3 bases.
phosphodiester bond	A covalent bond in which two hydroxyl groups form ester linkages to the same phosphate group; joins successive nucleotides in DNA or RNA.
Phred/Phrap	Phred is a base-calling algorithm and Phrap is an assembly algorithm used to align the output of multiple sequencing reactions.
plasmid	A small circular DNA molecule which carries genetic elements permitting its autonomous extra-chromosomal replication in bacteria or other single-cell organisms. A plasmid can be used as a recombinant DNA vector, to propagate foreign DNA in a bacterial cell. In addition to the essential origin for replication, plasmids generally carry a variety of marker genes, enabling easy identification of cells harboring the recombinant DNA.
poly (A)tail	The segment of adenylate residues that is posttranscriptionally added to the 3' end of eukaryotic mRNA. About 250 nucleotides of (A) are added by poly (A) polymerase following cleavage of the newly synthesized RNA about 20 nucleotides downstream of an AAUAAA signal sequence.
polypeptide	Amino acid chain containing hundreds to thousands of amino acids joined together by peptide (amide) bonds.
pre-mRNA	The initial transcript from a protein-coding gene is often called a pre-mRNA and contains both introns and exons. Pre-mRNA requires processing (addition of 5' cap and 3' poly (A) tail, removal of introns} to produce the final mRNA molecule containing only exons.
primary transcript	The immediate product of transcription of a gene, which is often modified before becoming fully functional

prokaryotes	The class of single-cell organisms, including the eubacteria and archaea, which lack membrane-limited organelles, including a nucleus.
promoter	A segment of DNA to which RNA polymerase binds to initiate transcription of the downstream gene(s).
protein	A molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs; and each protein has unique functions. Examples are hormones, enzymes, and antibodies; c.f peptide.
protein coding gene	Any gene whose ultimate biologically functional product is a protein, as opposed to an RNA molecule such as tRNA or rRNA.
pseudogene	A sequence of DNA similar to a gene but nonfunctional, probably the remnant of a once functional gene that accumulated mutations.
purine (Pu)	Adenine (A) and guanine (G) are purines, two of the four nitrogenous bases found in DNA.
pyrimidine (Py)	Cytosine (C) and thymine (T) are pyrimidines, two of the four nitrogenous bases found in DNA. In RNA, thymine is replaced by uracil (U).
pyrosequencing	A sequencing technology based on sequencing by synthesis; bases are identified on the basis of the release of pyrophosphate as they are incorporated into the growing DNA chain.
query	The input sequence (or other type of search term) with which all of the entries in a database are to be compared
raw sequence	Sequence that has been neither finished nor curated, and therefore not ready for annotation.
reading frame	See “frame.”
regulatory elements	DNA sequences that control expression of a gene by binding to proteins which increase or decrease synthesis of the gene product.
RepeatMasker	A program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. Use the link, http://www.repeatmasker.org
repetitious DNA	See “repetitive DNA sequence.”
repetitive DNA sequence	A sequence which is repeated multiple times in the genome; such sequences can vary considerably in length and number of copies per genome.
replication	The process of producing two DNA molecules from one. During replication, the two strands of the parent helix separate and DNA polymerase synthesizes a new, complementary strand for each parental strand, following the rules of base pairing (A-T and G-C).
retrotransposon	These are transposable DNA elements (transposons) that employ retroviral-like reverse transcription during the process of transposition: retrotransposon DNA is first transcribed into an RNA template which is then reverse-transcribed into a DNA copy which is inserted into a new genomic site.
reverse strand	See “forward strand.”
RNA polymerase	The enzyme which synthesizes a strand of RNA by adding successive ribonucleotides in the order dictated by a template strand of DNA.
rRNA	Ribosomal RNA, RNA molecules that are components of the ribosome. rRNA forms the structural scaffold for assembly of the ribosome, and plays a critical role in catalyzing peptide bond formation.
satellite DNA / simple sequence DNA	Highly repetitious DNA; generally based on a short sequence (7-20 nucleotides) repeated up to a million times in the haploid genome. Usually found in heterochromatic regions, often associated with the centromere.
sense strand	See “forward strand.”

shotgun sequencing	A strategy for sequencing whole genomes, it was pioneered by the for-profit company Celera. Genomes are cut into very small pieces, cloned into plasmids, sequenced, and then assembled into whole chromosomes or genomes. This method is faster than hierarchical shotgun sequencing but more prone to assembly errors.
simple repeat	A nucleotide repeat with one or a small number of bases, such as AAAAAAAAAAA or CACACACACA.
SINE	Short Interspersed Nuclear Elements are a class of DNA segments derived from reverse-transcribed genes and commonly found in eukaryotic genomes.
SNP	Single-nucleotide polymorphism; a difference in DNA sequence at a single base between two sequences.
splicing	The process by which introns are removed and exons are joined to produce a mature, functional RNA from a primary transcript. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.
splicing acceptor site	The boundary between an intron and the exon immediately downstream (i.e. on the 3' side of the intron).
splicing donor site	The boundary between an intron and the exon immediately upstream (i.e. on the 5' side of the intron).
splicing junction	Either a splicing acceptor site or a splicing donor site.
splicing transesterification mechanism	A chemical reaction that joins the 5' phosphate of the first nucleotide located at the 5' end of the downstream exon with the 3' hydroxyl group of the last nucleotide of the upstream exon forming a phosphodiester bond.
start codon	See "initiation codon."
start site	The nucleotide at which transcription starts, usually denoted as position +1 in reference to the gene being transcribed.
stop codon	A codon that specifies the termination of peptide synthesis; sometimes called "nonsense codons," since they do not specify any amino acid.
strand plus/minus	See "forward strand."
STRs	Short tandem repeats. At many places in genomes, there are short sequences (~5-35 bp) of bases which are not transcribed and which are repeated several times in a row (a tandem array). Different individuals will often have a different number of repeats and populations usually have a wide range of copy numbers at a given site. The number of repeats can therefore be convenient genetic markers for determining genetic relationships.
subject	The sequence, typically retrieved from a database, to which the sequence of interest (the "query") is being compared.
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.
tandem array	The same sequence, repeated multiple times, where each copy of the repeat immediately follows the previous copy. Genes encoding rRNA and the histones are usually in tandem arrays. Repetitious sequences that are NOT in a tandem array are referred to as "dispersed".
tblastn	Blast search tool in which the query is an amino acid sequence and the subject are 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. Use the link, http://www.ncbi.nlm.nih.gov/BLAST and select tblastn
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.

template	The starting material in a PCR reaction. When referring to a DNA strand, it is also called the negative, anti-sense, or non-coding strand. This strand of the DNA sequence of a slngle gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.
termination codon	See “stop codon.”
tiling path	A set of overlapping clones which cover (ideally) the entire sequence being assembled. See also “golden tiling path.”
transcript	See “transcription.”
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 terminator	Also called simply a terminator, it is a section of genomic DNA that marks the end of gene or operon, where transcription should stop.
translation	The process by which codons in an mRNA are used by the ribosome to direct protein synthesis.
translational start	See “start codon.”
translocation	Literally "a change in location". It usually refers to genetic translocations, in which part of a chromosome is transferred to another position in the genome.
transposable genetic element	A genetic element that can insert into (and exit from) a chromosome, and may therefore relocate in the genome; this class includes insertion sequences, transposons, retrotransposons, some phages, and controlling elements. Much of the middle repetitious DNA in eukaryotic genomes is made up of damaged transposable elements.
transposons	Segments of DNA that can move around to different positions in the genome of a single cell. In the process, they may cause mutations or increase (or decrease) the amount of DNA in the genome
tRNA	Transfer RNA, small (ca. 75 bp) L-shaped RNA molecules that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA. The proper tRNA is selected through the complementary base pairing of its three-nucleotide anticodon with the mRNA's codon, and its amino acid group is transferred to the growing polypeptide.
TwinScan	An gene prediction algorithm that uses conservation to a second "informat" genome to assist in the prediction of genes.
UCSC	Univ. of California Santa Cruz; host to a popular genome browser. See: http://genome.ucsc.edu/
upstream	Toward the 5' end of a single stranded lenth of DNA or gene of interest; *cf.* Downstream.
UTR	Untranslated region; a segment of DNA (or RNA) which is transcribed and present in the mature mRNA, but not translated into protein. UTRs may occur at either or both the 5’ and 3’ ends of a gene or transcript.
vector	A plasmid, phage, or other DNA that is used to maintain and propagate inserted foreign DNA in a host cell.
VNTRs	Variable Number of Tandem Repeats. See “STRs.”