**BIOINFORMATICS NOTE**

- To sort your contigs you can use some tools like Contiguator2. But you need a close related reference genome for this tool align your contigs properly, otherwise it will place your contigs in the wrong position.

An option to perform the ***gap filling*** is use tools like FGAP or GFinisher that will search for regions in the contigs ends that overlaps and merges these contigs.

- ***A scaffold*** is a portion of the genome sequence reconstructed from end-sequenced whole-genome shotgun clones. Scaffolds are composed of contigs and gaps. A contig is a contiguous length of genomic sequence in which the order of bases is known to a high confidence level.

- ***Primer walking*** (or Directed Sequencing) is a sequencing method of choice for sequencing DNA fragments between 1.3 and 7 kilobases. Such fragments are too long to be sequenced in a single sequence read using the chain termination method. This method works by dividing the long sequence into several consecutive short ones. The DNA of interest may be a plasmid insert, a PCR product or a fragment representing a gap when sequencing a genome. The term "primer walking" is used where the main aim is to sequence the genome. The term "chromosome walking" is used instead when the sequence is known but there is no clone of a gene. For example, the gene for a disease may be located near a specific marker such as an RFLP on the sequence.

The fragment is first sequenced as if it were a shorter fragment. Sequencing is performed from each end using either universal primers or specifically designed ones. This should identify the first 1000 or so bases. In order to completely sequence the region of interest, design and synthesis of new primers (complementary to the final 20 bases of the known sequence) is necessary to obtain contiguous sequence information

- ***Trim Ends*** removes misleading data from the ends of sequencing fragments. Trim Vector removes sequence-specific data contaminating the ends of your sequences. Trim to Reference eliminates the ends of sequences that extend beyond an assembled Reference sequence.

- ***Minimap2*** is a general-purpose alignment program to map DNA or long mRNA sequences against a large reference database. It works with accurate short reads of ≥100 bp in length, ≥1 kb genomic reads at error rate ∼15%, full-length noisy Direct RNA or cDNA reads and assembly contigs or closely related full chromosomes of hundreds of megabases in length. Minimap2 does split-read alignment, employs concave gap cost for long insertions and deletions and introduces new heuristics to reduce spurious alignments

- A gene family is a set of homologous genes within one organism. A ***gene cluster*** is a group of two or more genes found within an organism's DNA that encode similar polypeptides, or proteins, which collectively share a generalized function and are often located within a few thousand base pairs of each other. The size of gene clusters can vary significantly, from a few genes to several hundred genes.[1] Portions of the DNA sequence of each gene within a gene cluster are found to be identical; however, the resulting protein of each gene is distinctive from the resulting protein of another gene within the cluster. Genes found in a gene cluster may be observed near one another on the same chromosome or on different, but homologous chromosomes. An example of a gene cluster is the Hox gene, which is made up of eight genes and is part of the Homeobox gene family.

- ***Gene clusters*** are found to be close to one another when observed on the same chromosome. They are dispersed randomly; however, gene clusters are normally within, at most, a few thousand bases of each other. The distance between each gene in the gene cluster can vary. The DNA found between each repeated gene in the gene cluster is non-conserved.[10] Portions of the DNA sequence of a gene is found to be identical in genes contained in a gene cluster.

- ***FLAG-tag***, or FLAG octapeptide, or FLAG epitope, is a polypeptide protein tag that can be added to a protein using recombinant DNA technology, having the sequence motif DYKDDDDK (where D=aspartic acid, Y=tyrosine, and K=lysine). It is one of the most specific tags and it is an artificial antigen to which specific, high affinity monoclonal antibodies have been developed and hence can be used for protein purification by affinity chromatography and also can be used for locating proteins within living cells. It has been used to separate recombinant, overexpressed protein from wild-type protein expressed by the host organism. A FLAG-tag can be used in many different assays that require recognition by an antibody. If there is no antibody against a given protein, adding a FLAG-tag to a protein allows the protein to be studied with an antibody against the FLAG sequence. Examples are cellular localization studies by immunofluorescence, immunoprecipitation or detection by SDS PAGE protein electrophoresis and Western blotting.

- ***Gene fragments*** are pieces of genes containing only the exons (those parts of the gene which actually encode the protein sequence). They are composed of cDNA.

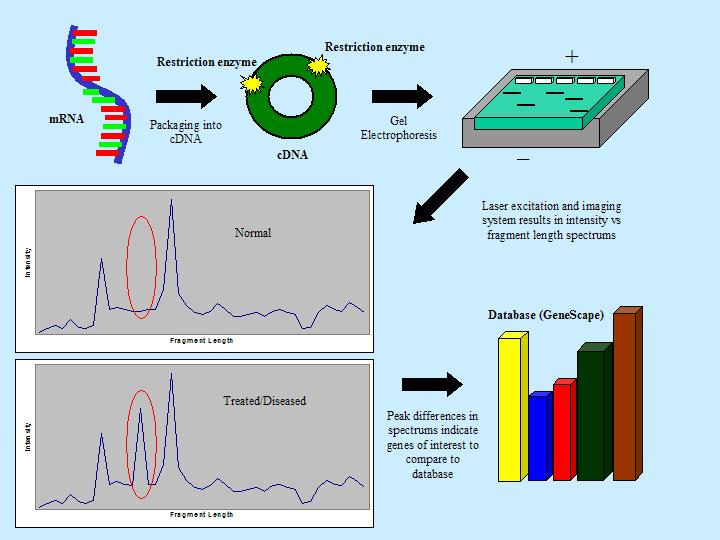
- ***cDNA:*** In genetics, complementary DNA (**cDNA**) is DNA synthesized from a single-stranded RNA (e.g., messenger RNA (mRNA) or microRNA (miRNA)) template in a reaction catalyzed by the enzyme reverse transcriptase. **cDNA** is often used to clone eukaryotic genes in prokaryotes.

- An **interrupted gene** (also called a [split gene](https://en.wikipedia.org/wiki/Split_gene_theory)) is a gene that contains expressed regions of DNA called [exons](https://en.wikipedia.org/wiki/Exon), split with unexpressed regions called [introns](https://en.wikipedia.org/wiki/Intron) (also called intervening regions). Exons provide instructions for coding proteins, which create [mRNA](https://en.wikipedia.org/wiki/MRNA) necessary for the synthesis of [proteins](https://en.wikipedia.org/wiki/Protein). Introns are removed by recognition of the donor site (5' end) and the splice acceptor site (3' end).[[1]](https://en.wikipedia.org/wiki/Interrupted_gene#cite_note-:0-1) The architecture of the interrupted gene allows for the process of [alternative splicing](https://en.wikipedia.org/wiki/Alternative_splicing), where various mRNA products can be produced from a single gene.[[2]](https://en.wikipedia.org/wiki/Interrupted_gene#cite_note-:1-2) The function of introns are still not fully understood and are called [noncoding or junk DNA](https://en.wikipedia.org/wiki/Noncoding_DNA).

- ***MLST (multi-locus sequence typing)*** is a classic technique for genotyping bacteria, widely applied for pathogen outbreak surveillance. Traditionally, MLST is based on identifying sequence types from a small number of housekeeping genes.

- In [molecular biology](https://en.wikipedia.org/wiki/Molecular_biology), **housekeeping genes** are typically [constitutive genes](https://en.wikipedia.org/wiki/Constitutive_gene) that are required for the maintenance of basic cellular function, and are [expressed](https://en.wikipedia.org/wiki/Gene_expression) in all cells of an organism under normal and patho-physiological conditions.[[1]](https://en.wikipedia.org/wiki/Housekeeping_gene#cite_note-:0-1)[[2]](https://en.wikipedia.org/wiki/Housekeeping_gene#cite_note-housekeeping-2)[[3]](https://en.wikipedia.org/wiki/Housekeeping_gene#cite_note-Butte-3)[[4]](https://en.wikipedia.org/wiki/Housekeeping_gene#cite_note-Jiang-4) Although some housekeeping genes are expressed at relatively constant rates in most non-pathological situations, the expression of other housekeeping genes may vary depending on experimental conditions.

- In the field of [genomics](https://en.wikipedia.org/wiki/Genomics), **GeneCalling** is an open-platform [mRNA](https://en.wikipedia.org/wiki/MRNA) transcriptional profiling technique.[[1]](https://en.wikipedia.org/wiki/GeneCalling#cite_note-Kirst2005-1) The GeneCalling protocol measures levels of [cDNA](https://en.wikipedia.org/wiki/CDNA), which are correlated with [gene expression](https://en.wikipedia.org/wiki/Gene_expression) levels of specific [transcripts](https://en.wikipedia.org/wiki/Transcription_(genetics)). Differences between gene expression in healthy tissues and disease or drug responsive tissues are examined and compared in this technology.[[2]](https://en.wikipedia.org/wiki/GeneCalling#cite_note-Green2001-2) The technique has been applied to the study of human tissues[[3]](https://en.wikipedia.org/wiki/GeneCalling#cite_note-3) and plant tissues.

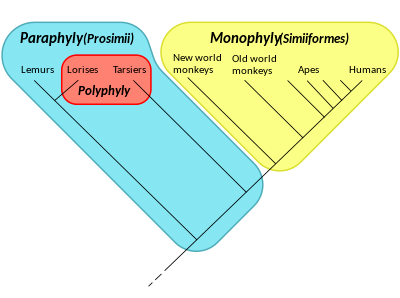
PHYLOGENETIC TREE

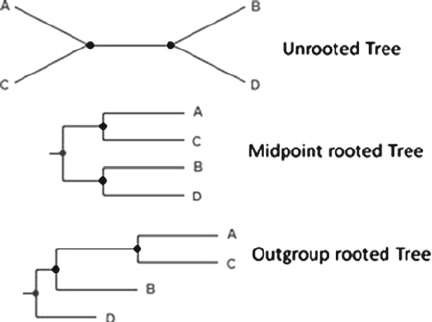
- In a phylogenetic tree, the relatedness of two species has a very specific meaning. Two species are more related if they have a more recent common ancestor, and less related if they have a less recent common ancestor.

- you may see trees with a polytomy (poly, many; tomy, cuts), meaning a branch point that has three or more different species coming off of it2. In general, a polytomy shows where we don't have enough information to determine branching order.

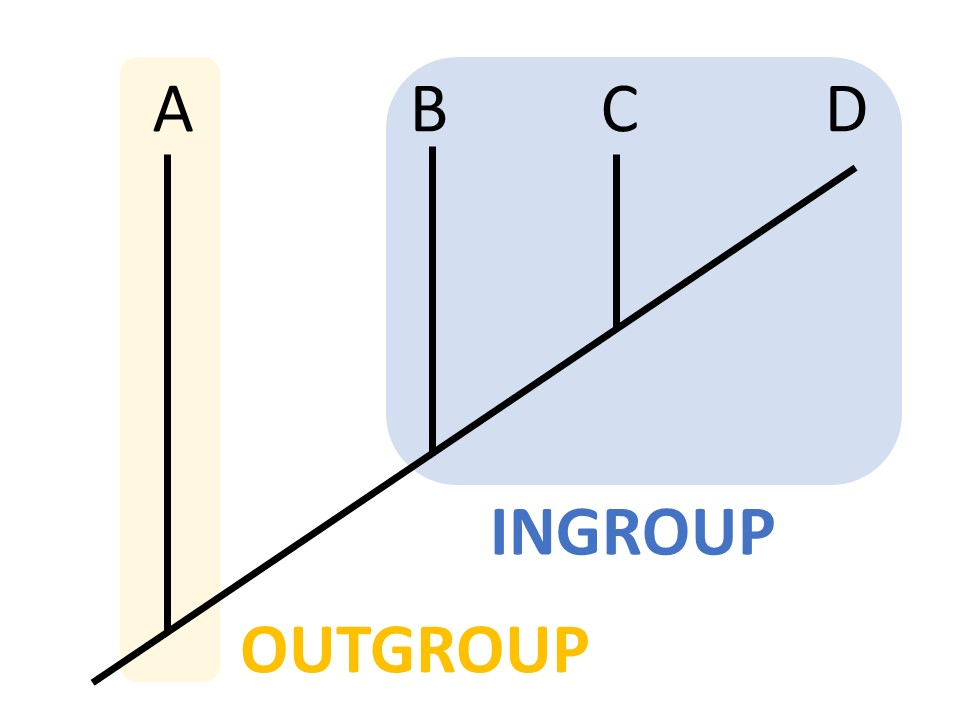
- In building a tree, we organize species into nested groups based on shared derived traits (traits different from those of the group's ancestor). The sequences of genes or proteins can be compared among species and used to build phylogenetic trees. Closely related species typically have few sequence differences, while less related species tend to have more. When we are building phylogenetic trees, traits that arise during the evolution of a group and differ from the traits of the ancestor of the group are called derived traits. In our example, a fuzzy tail, big ears, and whiskers are derived traits, while a skinny tail, small ears, and lack of whiskers are ancestral traits.

- orthologous: evolutionarily related genes or proteins. Homologous sequences originate from the same ancestors (homolog e.g all globin protein), which are separated from each other after a speciation event, e.g. human beta and chimp beta globin. An orthologous gene is a gene in different species that evolved from a common ancestor by speciation

polyphyly: convergent evolution



In [cladistics](https://en.wikipedia.org/wiki/Cladistics) or [phylogenetics](https://en.wikipedia.org/wiki/Phylogenetics), an **outgroup**[[1]](https://en.wikipedia.org/wiki/Outgroup_(cladistics)#cite_note-1) is a more distantly related group of organisms that serves as a reference group when determining the [evolutionary](https://en.wikipedia.org/wiki/Evolution) relationships of the ingroup, the set of organisms under study, and is distinct from [sociological outgroups](https://en.wikipedia.org/wiki/Ingroups_and_outgroups). The outgroup is used as a point of comparison for the ingroup and specifically allows for the [phylogeny](https://en.wikipedia.org/wiki/Phylogenetic_tree) to be rooted. Because the polarity (direction) of character change can be determined only on a rooted phylogeny, the choice of outgroup is essential for understanding the evolution of traits along a phylogeny

To qualify as an outgroup, a taxon must satisfy the following two characteristics:

* It must not be a member of the ingroup.
* It must be related to the ingroup, closely enough for meaningful comparisons to the ingroup.

ROARY:

## Input files from GenBank

On NCBI's website, GFF3 files only contain annotation and not the nucleotide sequence so cannot be used. You need to download the GenBank files plus nucleotide sequence and convert them. When downloading, click on the show sequence option, Update View then Send to a File of type GenBank. You can then use the Bio::Perl script bp\_genbank2gff3.pl to convert to GFF3. Just be aware that mixing different gene prediction methods and annotation pipelines can give noisier results.

Alternatively you can use [ncbi-genome-download](https://github.com/kblin/ncbi-genome-download) to pull down the FASTA files and convert them to GFF3 with Prokka.

SNP-SITES: use core\_gene\_alignment.aln

If you pass in the -e parameter to roary, a multi-FASTA alignment of all of the core genes is created. This can then be used as input to build a phylogenetic tree. To reduce the memory and run time, you can pre filter the alignment using [snp\_sites](https://github.com/sanger-pathogens/snp_sites).