

# A quick guide through bio-informatics for molecular genetics

Leonard Blaschek, William Gardi

November 27, 2019

## Introduction

Gene and genome duplication is a crucial cornerstone of plants' ability to adapt and evolve. In duplicating genes and forming multi-genic families, the plant creates redundancy in the machinery necessary to maintain its physiological functions. The new, redundant gene can either remain as it is and function as a backup against metabolic disruptions, or evolve to fulfil a subset of the original gene's functions – a process called *subfunctionalisation*. It can even develop a completely new function, undergoing *neofunctionalisation*. These new gene copies – called paralogs – allow the plant to adapt to new environments, outcompete rivals or defend against new enemies. Because they share a common origin, paralogs are *homologous* (as opposed to genes that serve similar functions as a result of convergent evolution, which are *analogous*). To increase the variability of functions within a gene family even further, the expression of the same paralog can result in different transcripts through a mechanism called *alternative splicing*. Consequentially, studying gene families and their variable functions is fundamental to our understanding of both plant physiology and adaptive evolution.

Blue text can be clicked and links to sections or web resources.

In this workshop, you will:

1. [Identify](#) all paralogs of a gene family in *Arabidopsis thaliana* and its *orthologs* in other species
2. [Analyse sequence similarity](#) between the paralogs and orthologs and visualise it by drawing a phylogenetic tree
3. Map the [gene structure](#) of the *A. thaliana* paralogs
4. Design paralog specific [RT-qPCR primers](#)
5. Design primers to identify a [T-DNA insertion](#) and a single nucleotide polymorphism in one of the paralogs

## Software and web tools

There are numerous tools available that have the necessary functions to complete the tasks of this workshop. Listed below is a short list of online based tools and suggested downloadable software.

Table 1: Suggested software for this workshop and beyond.

Software	Functionality	Link
SnapGene Viewer	sequence visualisation, primer design	<a href="http://snapgene.com">snapgene.com</a>
Jalview	sequence alignment and analysis	<a href="http://jalview.org">jalview.org</a>
TAIR	<i>Arabidopsis</i> databank (traffic limited)	<a href="http://arabidopsis.org">arabidopsis.org</a>
ThaleMine	unlimited alternative to TAIR	<a href="http://araport.org">araport.org</a>
Phytozome	general plant database	<a href="http://phytozome.jgi.doe.gov">phytozome.jgi.doe.gov</a>
Primer blast	primer design and testing	<a href="http://ncbi.nlm.nih.gov">ncbi.nlm.nih.gov</a>

Jalview is also available through [bioconda](#), if you are into that sorta thing.

## 1 Identifying paralogs

Paralogs, being products of relatively recent duplication of the same gene, are characterised by a high degree of sequence similarity. The most popular tool to identify sequences that are similar to the entered query is the Basic Local Alignment Search Tool, or BLAST. The degree of similarity is most succinctly summarised in the *E-value*, which essentially represents the likelihood of the query and result being as similar as they are by pure chance instead of common origin. A simplified way of defining a gene family is as the group of sequences sharing the highest degree of similarity within the genome of a given species.

**What sequence to use?** What type of sequence should you use? Think about what characterises nucleotide and protein sequences, and which one would be better suited to reliably identify homologous sequences. Theoretically, would the time that has passed since the duplication events play a role?

**Paralogs, duplicate entries & splice variants.** Depending on your gene family and the sequence type, your BLAST results will probably include all three of those. Genes are unambiguously identified by their genomic position, or *locus*. Multiple BLAST results that have the same locus are just duplicate sequence submissions of the same gene. Protein or mRNA sequences whose loci only vary in the trailing decimal (*e.g.* AT1G27920.1 and AT1G27920.2) are splice variants of the same gene.

**Suggested tools:** [NCBI BLAST](#)  
[TAIR BLAST](#)  
[Phytozome BLAST](#)

## 2 Analysing sequence similarity

**Alignments and scores.** Once you have identified your genes of interest, you can quantify and compare their degree of sequence similarity. In order to do so, the first step is to calculate a multiple alignment (Jalview: **Web Service** → **Alignment**). To compare the sequence similarity between two paralogs or orthologs numerically, the most easily interpretable measure is percent similarity. Put simply, if two sequences share the same amino acid at a given position, the similarity for that position is 100%. When the amino acids differ, the similarity is scored according to the chemical differences (charge, hydrophobicity, *etc.*) between the two amino acids based on the used substitution matrix (BLOSUM62, PAM250, *etc.*). The overall percent similarity of two sequences is then simply the average similarity score of all positions. A short macro to obtain a percent similarity matrix of all your sequences in Jalview you can copy from [here](#). After computing your alignment, simply open the console (**Tools** → **Groovy Console...**), paste the script into the box and execute it.

**Phylogenetic trees.** To get a good overview of the similarities between many sequences at once, we draw phylogenetic trees. The distance matrices underlying phylogenetic trees are, put simply, more sophisticated versions of the percent similarity matrix above. In order to *root* the phylogeny, you will need to include an *outgroup*. An outgroup is any sequence that is closely related to your paralogs, but not as closely as the paralogs to each other. To choose an outgroup sequence you can simply select a sequence from further down the list in your blast results with a higher E-value than your paralogs. Only when a tree is rooted can you make inferences about the *direction* of the tree, *i.e.* which sequences are ancestral and which are more recently diverged. To create a phylogenetic tree in Jalview, select **Calculate** → **Calculate Tree or PCA...**

**Suggested tools:** Jalview

## 3 Mapping gene structure

Mapping exons, introns and UTRs onto a genomic DNA sequence is a fundamental part of gene annotation. The easiest way to do it through simple sequence alignment. Just remember which sequences (gDNA, mRNA, CDS) contain which parts of the gene.

**Suggested tools:** Jalview  
SnapGene Viewer

## 4 Designing RT-qPCR primers

Multigenic families pose additional challenges when analysing their expression. Because their sequences are so similar, it can be difficult to find primers that are specific to only one paralog. It is often a good idea to design primers in the UTRs of the gene, since those are usually less conserved between paralogs. Even more difficult is the design of splice variant specific primers, which you need to design in the specific regions that are differently spliced.

RT-qPCR primers need to be designed to rather strict specifications to maximise the chance for successful experiments, as summarised in table 2.

Table 2: Restraints for the design of qPCR primers

Parameter	Ideal	Acceptable
Primer length	20 bases	18–22 bases
G/C content	> 50%	> 40%
$T_m$	60°C	58–63°C
$\Delta T_m$	< 0.5°C	< 1°C
off-target BLAST hits	none	mismatch at 3'-end
Amplicon length	180–200 bp	140–220 bp

**Suggested tools:** SnapGene Viewer  
[Primer BLAST](#)  
[NCBI BLAST](#)

## 5 Detecting mutations