# A quick guide through bio-informatics for molecular genetics

Leonard Blaschek, William Gardi

July 4, 2020

## Introduction to multigenic families

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.

## The goal of this workshop

In this workshop you will characterise a gene family *in silico*. Specifically, 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

In the process, you will learn to use the strengths (and overcome the weaknesses) of core bioinformatic tools for molecular genetics. Beyond it's purpose for training, this type of analysis is a fundamental part of planning and executing experimental work in molecular genetics and biotechnology. The tasks in this project will require you to integrate multiple methods to reach a solution. As part of the exercise, you may have to seek out additional information, and decide on how to best apply the methods yourself. It is important that your strategy is not limited by your fundamental knowledge of genetics and molecular biology. If you do not understand the operating principles of the presented tools, re-read the respective chapters in your favourite textbook. Pay special attention that you understand the central dogma of molecular biology, as the processes are directly relevant for correctly applying the methods.

This document will point you in the right directions. Helpful considerations will be mentioned, although the specific implementations is for you to discover. We point you to towards tools and databases that are cross-platform (useable on Windows, Mac and Linux) and that we are familiar with. You are very welcome to experiment with other tools, but be aware that we might not be able to help if you run into problems with them.

## Sequence information and annotation

The standard way of digitally storing sequences is in the FASTA format. FASTA files are simple text, and can be opened in any text editor. If a program is having difficulties reading a FASTA file, ensure that the file extension is `.fasta`. The first line of a FASTA sequence always begins with a `>`, followed by reference information such as the gene identifier and chromosomal position. This is followed by the sequence denoted in either nucleotides, or amino acids depending on its type. Occasionally, you will encounter ambiguous codes such as `N` in DNA/RNA sequences, and `X` in protein sequences. These indicate low quality reads during sequencing. They are not necessarily problematic, but make sure that the tools you plan on using supports their presence. You can find sequences for genes, transcripts and proteins in several publicly available databases (refer to Table 1). On Phytozome, you can search for keywords in a given species, click on the gene view of a hit (`G`), and then go into the `Sequences` tab. On NCBI, search using the `Gene` database. Clicking on a hit will give you plenty of additional information, including cross-references to potential transcripts (generally denoted by `NM_xxxx`) and proteins (`NP_xxxx`). Find the `FASTA` link to fetch the sequence.

For a detailed explanation of the NCBI accession prefixes, see here.

## 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 | snapgene.com |
| Jalview | sequence alignment and analysis | jalview.org |
| TAIR | *Arabidopsis* databank (traffic limited) | arabidopsis.org |
| ThaleMine | unlimited alternative to TAIR | araport.org |
| Phytozome | general plant database | phytozome.jgi.doe.gov |
| Primer blast | primer design and testing | ncbi.nlm.nih.gov |

# 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, ou 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

Phenotyping mutants requires the zygosity of the mutation to be determined. Commonly, the initial step is to amplify the region surrounding the mutation. Genotyping can then be done through various ways depending on the type of mutation. SNPs can be characterized through CAPS (restriction enzyme digest), TILLING (heteroduplex formation) or direct sequencing. Insertions are easier to detect, since part of the inserted DNA sequence is known. Please refer to SIGnaL for the genotyping strategy of T-DNA lines. Mutant lines can be found on TAIR, in the `Polymorphism` section of a gene. It will report what kind of mutations there are, and their positions. For insertional mutations, the position is approximated by the Flanking Sequence Tag (FST). You can also find T-DNA lines directly via T-DNA Express. Simply supply the gene identifier in the `Query` field, and it will give you a map of associated T-DNA insertions. Primer design for genotyping follows similar rules as that for RT-qPCR (table 2), except that you aim for amplicon sizes between 500 and 1000 bp, and pay less attention to the $T_m$.