A quick guide through bio-informatics for molecular genetics

Leonard Blaschek, William Gardi

November 23, 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 an 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.

In this workshop, you will:

1. Identify all paralogs of a gene family in *Arabidopsis thaliana* and its *orthologs* in other species
2. Create a phylogenetic tree of paralogs and orthologs to visualise their similarity
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	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?Hint

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.

Software: SnapGene Viewer

Web tools: NCBI BLAST
TAIR BLAST
Phytozome BLAST

2 Drawing a phylogenetic tree

3 Mapping gene structure

4 Designing RT-qPCR primers

5 Detecting mutations