Tommy Conway

Activity

For today’s activity, you will be exploring a genome and its gene annotation of your choice from Phytozome. First select a genome and load the annotation in a genome browser (JBrowse). You can also navigate Phytozome to get some basic statistics on your chosen genome.

1. What Genome did you choose?

Solanum Lycopersicum iTAG4.0 --tomato

1. Explore Phytozome to learn about the genome. What sequence technology was used?
2. How many contigs and/or scaffolds are in the assembly? Do you think this is a good assembly? Why or why not?
3. How many gene models does this species have. Does it match your expectations given the ploidy or genome complexity?

Answer:

1. Solanum Lycopersicum iTAG4.0 --tomato
2. This genome Sl4.0 is PacBio long read sequencing.
3. 448 contigs and 13 scaffolds. Yes, at first glance I noticed that the website reports that this assembly is very contiguous with less gaps than previous genome builds. But how is this determined?

The number of scaffolds is also pretty close to the number of chromosomes (12). This suggests high continuity of the assembly with minimal gaps. I’m still learning how to read the N50 and L50 but I think these are also useful for assessing the assembly.

1. 34,075 protein encoding genes, which is only 6-7 thousand more genes than Arabidopsis (27-28k protein coding genes). It didn’t match my expectations because while the tomato plant is considered ‘diploid’, it’s genome is reported to be much larger and more complex than Arabidopsis (7x more base pairs!). Are the genes generally larger or are is there less gene density and more non-coding sequences in tomato?

Now, in JBrowse, explore a contig/chromosome from your plant genome and pick one gene to analyze further as described below.

1. What gene did you choose? Can you figure out the naming scheme for genes in this genome? (Note: this is sometimes not intuitive)
2. What is the structure of this gene? How many introns/exons? Are there multiple transcript isoforms? Do you think this is a ‘typical’ gene in the genome or is it smaller/larger than average?

A diagram of a solution

AI-generated content may be incorrect.

The gene is somewhat small for a tomato gene which appear to be around 3,281 bp, while this gene is only 1602 bp.

1. Are there any tracks on JBrowse that tell you more information about this gene? (Note: there might not be a lot of other tracks, this depends on the genome you choose!)

Answer:

1. Sollyc04g008300.1.1, which is a homolog of an atypical kinase gene ABC1K6 I’m studying in *Arabidopsis*.

The naming scheme is first 3 letters of genus,first 3 letters of species, chromosome number, g, gene unique ID.

The ID’s seem to be numerically in order as you go along the chromosome.

1. The gene is unusual in that it appears to be a single continuous CDS with no introns. So 1 exon with no introns. It is also reported in the literature that this gene has not introns (<https://www.researchgate.net/publication/282948169_Genome-wide_Identification_Phylogenetic_and_Expression_Analysis_of_ABC1K_Gene_Family_in_Tomato_Solanum_lycopersicum_L>)
2. There are not a lot of tracks for this gene yet aside from the functional annotation, protein homologs, and gene sequences.

Let’s look at the functional annotation of your gene to figure out what the gene might be doing. Maybe this gene is highly expressed in a specific tissue or stress, or maybe it is close to a QTL or GWAS peak from genetic mapping, and you want to determine ether further characterization of the gene is worthwhile.

1. Does your gene have any pfam domains? Click on them if so or look them up. What are they doing?
2. What about panther, interpro, or Kegg domains. If it has multiple, does these roles/annotations make sense?
3. What do you think your gene is doing in that genome?
4. Based on the protein homologs, do you think this gene is widely conserved across species? (Hint: a gene is likely widely conserved if it is present in many other plant genomes.) Does the level of conservation or species-specific uniqueness make sense in the context of its potential functn? !)

Answer:

1. Yes, PF00069 which is protein kinase domain. These are domains involved in phosphorylating proteins using ATP (or occasionally other NTP’s) and a metal ion cofactor.
2. Yes, panther reports AARF DOMAIN-CONTAINING PROTEIN KINASE 1, interpro reports protein kinase (and kinase-like) domain. Kegg reports ADCK or ABC1. These domain names are other names for the ABC1K kinase family that refer to other members of the genes in humans (ADCK) and yeast (ABC1).
3. I believe the gene is acting as a protein kinase like it’s homolog in Arabidopsis.
4. The sequence conservation appears to be 40-60%. Each of these ABC1K genes appear to relatively conserved across the plant kingdom and this exercise in tomato provides another example supporting this idea.

Despite sequence similarity of 40-60% when comparing homologs in other species, the presence of homologs across the plant kingdom suggests the genes are important for plants and arose early in the evolutionary history of plants. My PhD research involves deciphering the function of 6 ABC1K homologs that localize to the thylakoid plastoglobule. We think these homologs play a role in regulating photosynthesis for plant stress responses and senescence through protein phosphorylation. These would be very essential functions for all plants, so it isn’t a surprise these genes have persisted in so many species.