

# I: Genome Annotation II: Functional Annotation

Wednesday, September 22, 2010

1

- I: Genome Annotation
- Theory of gene prediction algorithms, ab initio and homology-based methods
- Repeat sequences in genomes
- workshop questions I

Wednesday, September 22, 2010

2

## What do you do with a genome sequence?

```
>scaffold_1
gaggataatagaaacatcaagataggatttgccctctctatgaGAGAGATtttCTGG
gCCCcCGATAgGTGAGACTGAGAAATGCATGCCGTGCTCAAATGAATGATAGTGTGA
TCAATATCATTtGAATTTATCAGTCTACTTAgAGATCGAGAAATCATAAGTTtGAACATT
GTAAGTTTGACATTGACCATGACTTATGTTCAAAC TAGATAGTGTACAAAAGGATTA
ATTCATGCTCATTTAAATAGGCTTATCGGACTATTGATCCTCATATTAGTTGGATAGTC
ATAATGCTTGCTAGAGGCCACTTATGACTTATAGGCCTTGTGGACTGGCCATTGGCC
AATTAATGTGAGCCTATTGGGTCACACACAAGAACGTTGTGATGCTATATTAAATGAG
CTAAAAGCCCTTAGTATAAATAATAAATAAAGTGTATTATTATAATTATTATTA
TTATAAGATAATAATTTAAAAGACTCTGAGCTATCTGAACTGTTACAAATAAAGG
ACTCTATCACATAAAGATTTTGAAGTATCTGAGAGATTGTAGTGGTCTGAAACACA
ACTCTTTGGGAAAAGAGTAGTAGGAAAACAAAAGACTGAAACATATAAATCTCTTC
TAGGAAGAGAGTGGTGACGACGGTTTTGAGAAAACAGTCTAAAAGATTGCAGATTGTG
GAGCACATAATCTATCATCATTGAGAGAGCTAGCACTTAGAAGAATAGAAAACGTTCCG
TGTGGATACCATTGAGGTTATTCGTTGAAAAGGCAACACCATCAGTCCGGCATTATATC
TTCTTGATGAGATTAAACAGTTAGTCTCATATCCATCTTTGAATTAACGAAGCACTCA
GATTCGGGAAAGGAAATCAACAGAGATTTATTTTCCGCTGCGCAACTGGGTGCAAC
AATCTCGGGATTTCCCAACACATATATGGCATTCTAGTGCATGAACATGCTAAAAAAT
TTCAATTTACTTTAAAACATAGTTCAAGTTCTACTAGCTCTGGCTGACTGAAACAGCTA
GCACTGCTGGCCTCTCGGTTCTCGGTCGATCTACACAGTGGACTCAAATGAGGG
ACAAAACAACCTAGCAAGACTCTAAACATCTCCCAAAAACCCCTAAAACATATAAA
ACATTCATAGAAAACATGCAAAGGAGGCTGAACAGGAGACTTCGGCGGCAGGTTCCGC
GGCCGAAGTCCCTCCAGAGCCGAAAGTCAAGCACTTCGGGGCAGGGTTCAGCGGCCG
AAAGTCCCTCAGAGCCGAAAGTCAACACTTCGGAGGCAAGTTCGGCGGCCGAAACTCC
CCTCTAGAGTCAAAGTCAAACCTTCGGGGCAGTTAGCGGCCAAACTGCCTCAC
AAGTTCGGCGGCCGAAACTGGCTTCGGCGGCCGAACTGGGCTCTCCAAAGGCGAAG
CCGATTCGCTTGCATCTAGCCACCCACAACAACACTGCTCGGGTTGTGGAAGCCGG
TAAAAAATAACCTTCTGGTTATCAACAATTTACAACATGCAGATAGTGGTGAAGGATCGA
ATCCACAGGAATTGATTACTATTATTTTCTCAATCAAGATCAATCAGAATCAATTG
AAAGCAAGTAGAGCAAAAGTAAAGTGAATAAAGTAAAAGAGGGTTTTGAAAATGATTT
GACTAACTATTGAAAAGCAATTAACAACGCAAGTAAATAAAATAAAGAGAAAATCAA
TACGGGAAAGTCTAGTTGAAGATAGGATCCACTTTGGTTGTTGGGTTGATCATTGAA
ACTTATGTTTTCTTGATTGATTCATAGATTAGTTATGGGGATTGAAAACGCTCTCACCC
ACCATGCTCTCTTATTAATCAATAGGGAACGTCCTAATCAATTAATAATTA
ACAAATTGCCAAGGAATGCTCTTGGGCTTAGGCATCAAAACAATTTGCAATTGCATAAA
GAATGAGAGAGATCAAACCTAGCTCAACCGCATGAGATGTTGATGATCATGCAAT
TCCTTAGTTTTTACCAAGTGTCTTAGGTCAGAAATATTTTACGCAATTACGGACTAAC
...
```

Wednesday, September 22, 2010

3

## Overview of ab initio Gene prediction algorithm

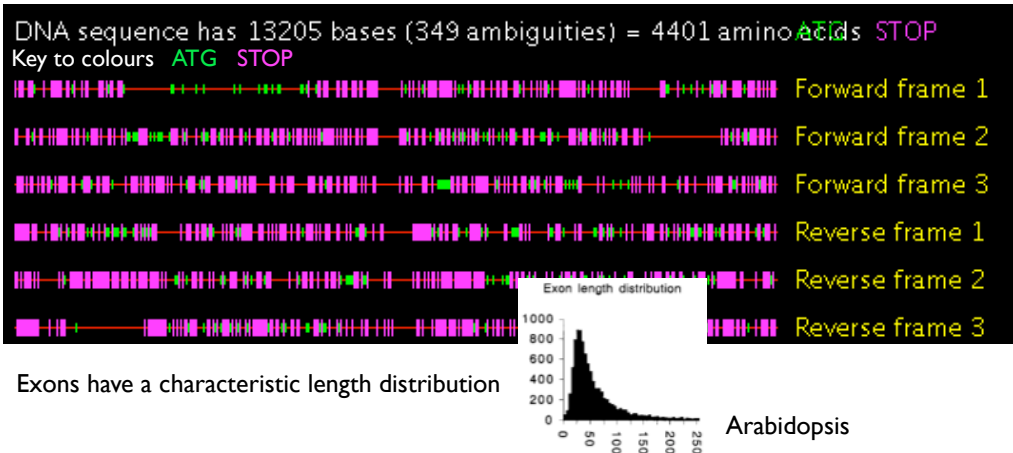
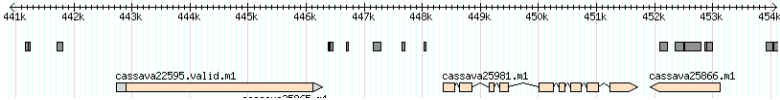
- open reading frames
- splice site models
- scoring schemes
- inaccuracies e.g. missing small exons, adjacent splice sites, genome sequence composition: GC-rich regions are reduced in stop sites.

Wednesday, September 22, 2010

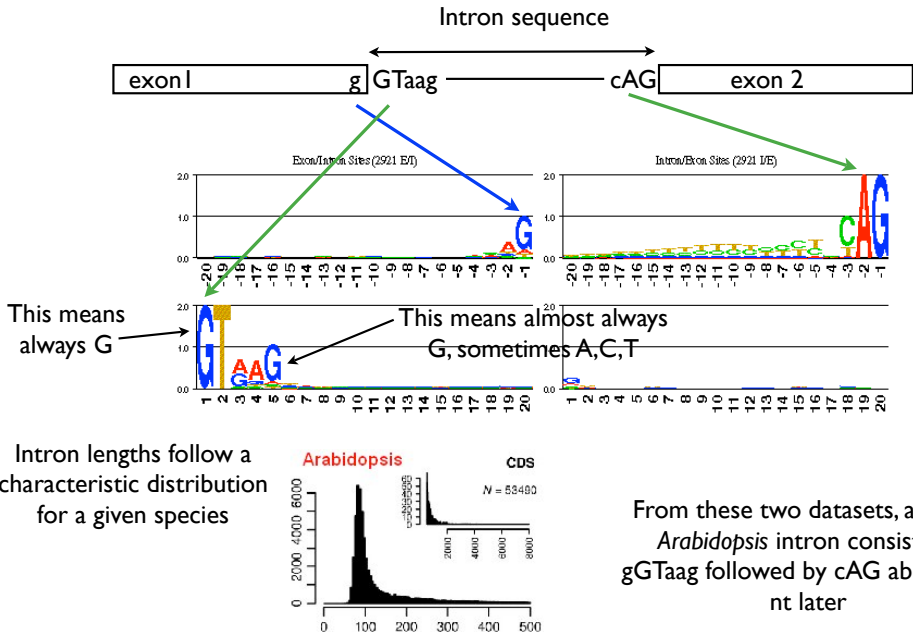
4

# Open reading frames

STOP codons occur commonly in random sequence. Selection acts on regions that encode protein to prevent STOP codons from evolving. Genes consist of alternating exons and introns separated by splice sites at characteristic distances. The exons make an ORF when spliced together

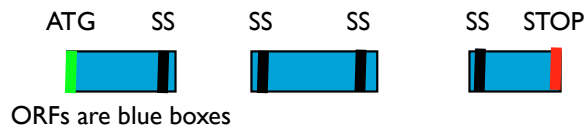


# Splice sites can be expressed statistically



## A simple gene prediction program

- Scan the genome for ORFs in all three reading frames that include splice sites and score them according to the expected lengths of exons as well as introns.
- Combine these two with information on how many introns are found in an average gene to make gene models.
- Make sure the complete gene ORF starts with Met and ends with STOP.
- Build models on the reverse strand too and pick non-overlapping models with the highest scores.
- This approach works but ok, not very well.
- You can write this code in your spare time!
- Examples include genie, genscan



Wednesday, September 22, 2010

7

## Homology-based gene prediction

- Use external evidence from protein homology and/or expression
- Add this extra information to ORF + splice site scoring system of ab initio prediction algorithm.
- Including homology and or EST evidence increases accuracy, decreases false positives.
- Examples include fgenesh, genomescan, augustus (very nice prediction program).

Wednesday, September 22, 2010

8

## The cassava gene prediction pipeline

- The cassava genes were predicted by a complex computer pipeline that runs in about 2-3 days using a cluster with 320 CPUs.
- The pipeline defines gene loci as regions that have EST expression evidence and/or protein homology with other organisms.
- These regions are given to fgenesh and genomescan with the homology evidence to generate gene predictions.
- We pick the “best” gene prediction at each locus (based on evidence supporting the model) and the use PASA to improve/extend the gene model using EST evidence.
- Lastly we filter gene models that are short or have homology to transposable elements.
- The gene models are very accurate for genes with EST support either over their entire length or at the ends and for genes with good homology in other species. Gene models are less accurate without good evidence from proteins and ESTs. Gene models are completely missing if there is no expression or homology evidence.

Wednesday, September 22, 2010

9

## Repeats

- plant genomes are very rich (sometimes over 60%) in repeat sequences. These are hard to sequence and assemble.
- simple repeats consist of a short sequence repeated many times, e.g. AAAAAAA or ATATATAT
- low complexity sequence regions do not include all nucleotides e.g. GCGGCGCGGCGCGGCG
- microsatellite ~50-~200 nt repeat units often found in e.g. telomeres
- transposable elements (TEs) belong to families and are ~1-10kb sequences that can move around the genome (transposition) either autonomously (larger TEs encode genes needed for transposition) or non-autonomously (smaller TEs have lost the genes that encode transposition proteins, but use the proteins encoded by other autonomous TEs of the same class) and this can lead to replication
- masking repeats temporarily hides sequences so you can predict genes accurately and the repeat sequences don't get in the way of gene model sequences.

Wednesday, September 22, 2010

10

## Workshop questions I

- What is an open reading frame?
- What are the two commonest splice site consensus sequences?
- Let's make our own gene predictions by running some genomic sequence through a genscan server.
- Open [http://128.32.19.162/cassava\\_genome\\_fragment\\_for\\_ORFs.fa.txt](http://128.32.19.162/cassava_genome_fragment_for_ORFs.fa.txt)
- Select the fasta genomic sequence and copy it.
- Go to <http://genes.mit.edu/GENSCAN.html>. Paste the genomic sequence into the sequence window. Select "Print CDS and peptides" from the print options menu. Select organism "Arabidopsis" and run Genscan, a gene prediction program.
- Results are organized by exons in the following table format

gene number and exon number		strand		length			statistics and scores						
Gn.	Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.06		PlyA	-	640	635	6							1.05
1.05		Term	-	2568	2224	345	0	0	72	43	153	0.447	7.71
1.04		Intr	-	3215	2934	282	2	0	-15	99	150	0.607	7.49
1.03		Intr	-	4070	3702	369	2	0	28	109	257	0.978	21.18
1.02		Intr	-	4523	4228	296	0	2	141	25	295	0.969	29.30
1.01		Init	-	4773	4713	61	0	1	34	100	75	0.994	7.77
1.00		Prom	-	6772	6733	40							-6.55

exon types are  
 PolyA site,  
 Terminal,  
 Internal,  
 Initial,  
 Promoter

Wednesday, September 22, 2010

11

- **Keep the output from genscan, you will need it for the next workshop questions. You can save the output as a text file or keep the browser window open.**
- How many genes were predicted and how many exons were in each?
- Below the exon coordinate table are the predicted peptide sequences and predicted CDS sequences.
- Let's work on the first gene that was predicted. It has GENSCAN\_predicted\_peptide\_1 in the title. What tools do you know about that you can use to learn more about this gene prediction? Try BLASTP at <http://blast.ncbi.nlm.nih.gov/Blast.cgi> against the nr database. Do you think you have a complete gene model? Look carefully at the coordinates of the query and hits in your BLAST results. Try blasting the protein sequence of GENSCAN\_predicted\_peptide\_2 as well.
- What did you learn about the proteins encoded by the gene models? You will need this information in the next workshop
- The genome sequence we predicted proteins on is part of the cassava genome sequence, scaffold00046. Now let's look at how the exons predicted by genscan map to the genome. TBLASTN the peptide sequence of GENSCAN\_predicted\_peptide\_1 against the masked cassava genome. How do you think the second best hit is related to the first? Click on a feature diagram to look at the region in Gbrowse.

Wednesday, September 22, 2010

12

- Zoom out by clicking on “Show 10 kb”. Your gene prediction is in the UserBlast track. The cassava gene in the genome prediction is in the transcript track. What’s missing from the genscan gene model you just predicted? Compare the genscan prediction to the transcript and the data in the orange peptide homology tracks below. Which prediction has better evidence?
- Turn on EST tracks if you haven’t already by scrolling down and turning on both PASA Aligned EST and PASA Assembled EST tracks and click on “Update Image” to redraw the region. What additional sequence is added to the gene model by evidence from the ESTs?
- If you want to compare tracks, you can click on the orange and green track name on the left and drag them up and down.
- The genscan gene model doesn’t have extra evidence from homology or ESTs, so is not as complete. Short coding exons are very hard to predict and there is no way to predict UTR regions (grey parts of the models in the transcript track) at all.
- The quality of the transcript is very high. It is probably completely accurate. This is because we have 1.5 million EST sequences from 454 sequencing. You can read all about how the transcripts were made here. <http://www.phytozome.net/cassava.php>

Wednesday, September 22, 2010

13

## Lecture II

- II: Functional Annotation
- The Gene Ontology (GO)
- Pfam, Panther and protein family construction
- annotating function
- workshop questions II

Wednesday, September 22, 2010

14

## Functional Annotation Systems

- These help us predict the function of a gene model using sequence similarity to a protein with a known function.
- Gene Ontology (GO)
- Sequence Ontology (SO)
- Plant ontology (PO)
- Pfam
- Panther
- EC numbers (Enzyme commission number based on the reaction the enzyme catalyzes)
- KEGG Orthology (KO)

Wednesday, September 22, 2010

15

## The Gene Ontology

- The most important organized list of biological terms.
- Started with [FlyBase](#) (*Drosophila*), the [Saccharomyces Genome Database](#) (SGD) and the [Mouse Genome Database](#) (MGD), in 1998. Now many collaborators.
- Motivation: saves time, increases efficiency for computer and human searching of data.
  - If you were searching for new targets for antibiotics, you might want to find all the gene products that are involved in bacterial protein synthesis, and that have significantly different sequences or structures from those in humans. If one database describes these molecules as being involved in 'translation', whereas another uses the phrase 'protein synthesis', it will be difficult for you - and even harder for a computer - to find functionally equivalent terms.
- Uses **controlled vocabularies** and organizes (structures) data
- The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, development of tools that facilitate the creation, maintenance and use of ontologies.

Wednesday, September 22, 2010

16

## The vocabularies

- Cellular component
  - A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
- Biological process
  - A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.
  - A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.
- Molecular function
  - Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.
  - It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word "activity". The documentation on the function ontology explains more about GO functions and the rules governing them.

Wednesday, September 22, 2010

17

## The biological process of “Response to water deprivation” (what we know as drought tolerance!)

- Here is the organization of the processes involved in “drought”.
- These are standardized in the Gene Ontology so that everyone uses the same terms.
- “Drought tolerance” is a response to stimulus, a response to water and a response to stress. In GO, there are 241 genes associated with this process.

- ▣ all : all [447732 gene products]
  - + ⓘ GO:0008150 : biological\_process [343626 gene products]
    - + ⓘ GO:0050896 : response to stimulus [36836 gene products]
      - + ⓘ GO:0009628 : response to abiotic stimulus [4512 gene products]
        - + ⓘ GO:0009415 : response to water [294 gene products]
          - + ⓘ GO:0009414 : response to water deprivation [241 gene products]
  - + ⓘ GO:0042221 : response to chemical stimulus [15714 gene products]
    - + ⓘ GO:0009415 : response to water [294 gene products]
      - + ⓘ GO:0009414 : response to water deprivation [241 gene products]
  - + ⓘ GO:0006950 : response to stress [20504 gene products]
    - + ⓘ GO:0009414 : response to water deprivation [241 gene products]

Wednesday, September 22, 2010

18

## The Sequence Ontology

- Most useful when someone is working with sequences computationally
- <http://www.sequenceontology.org/>
- The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence.
- SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process.
- Examples are binding\_site and exon. Biomaterial features are those which are intended for use in an experiment such as aptamer and PCR\_product.
- There are also experimental features which are the result of an experiment. SO also provides a rich set of attributes to describe these features such as "polycistronic" and "maternally imprinted".
- SO provides
  - structured controlled vocabulary for the description of primary annotations of nucleic acid sequence
  - To provide a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases.

Wednesday, September 22, 2010

19

## Plant Ontology

- Very useful for standardizing description of plant phenotypes and traits and anatomy, so that the descriptions are easy to share with other scientists and for computers to understand.
- <http://www.plantontology.org/>
- Aim is to develop, curate and share **controlled vocabularies** (ontologies) that describe plant structures and growth and developmental stages, providing a semantic framework for meaningful cross-species queries across databases.
- Plant Structure
  - A controlled vocabulary of botanical terms describing morphological and anatomical structures representing organ, tissue and cell types and their relationships. Examples are gametophyte, parenchyma, guard cell, etc.
- Growth and developmental stages
  - A controlled vocabulary of terms describing (i) whole plant growth stages and (ii) plant structure developmental stages. Examples are seedling growth, rosette growth, leaf development stages, embryo development stages, flower development stages, etc.
- Plant ontology is not an extensive collection of botanical terms, but rather a complex hierarchical structure in which botanical concepts are described by their meaning and by relationship to each other. The main purpose of these vocabularies is to facilitate cross database querying and to foster consistent use of these vocabularies in the annotation of tissue and/or growth stage specific expression of genes, proteins and

Wednesday, September 22, 2010


20

- The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System is a unique resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are classified by expert biologists according to:
  - Gene families and subfamilies, including annotated phylogenetic trees
  - Gene Ontology classes: molecular function, biological process, cellular component
  - PANTHER Protein Classes
  - Pathways, including diagrams
  - PANTHER is part of the Gene Ontology Reference Genome Project.


Wednesday, September 22, 2010

21

## Pfam link from Protein page



[HOME](#) | [SEARCH](#) | [BROWSE ABOUT](#) | [FTP](#) | [HELP](#)



**Family: SNARE (PF05739)**

24 architectures
1747 sequences
2 interactions
165 species
31 structures

**Summary**

**Domain organisation**

**Alignments**

**HMM logo**

**Trees**

**Curation & models**

**Species**

**Interactions**

**Structures**

**Jump to...**

enter ID/acc

**Summary**

**SNARE domain**

Most if not all vesicular membrane fusion events in eukaryotic cells are believed to be mediated by a conserved fusion machinery, the SNARE [soluble N-ethylmaleimide-sensitive factor (NSF) attachment protein (SNAP) receptors] machinery. The SNARE domain is thought to act as a protein-protein interaction module in the assembly of a SNARE protein complex [1].

**Literature references**

1. Weimbs T, Low SH, Chapin SJ, Mostov KE, Bucher P, Hofmann K; , Proc Natl Acad Sci U S A 1997;94:3046-3051.: A conserved domain is present in different families of vesicular fusion proteins: a new superfamily. [PUBMED:9096343](#)

**InterPro entry IPR000727**

The process of vesicular fusion with target membranes depends on a set of SNAREs (SNAP-Receptors), which are associated with the fusing membranes [PUBMED:9239749](#), [PUBMED:9232812](#). Target SNAREs (t-SNAREs) are localised on the target membrane and belong to two different families, the syntaxin-like family and the SNAP-25 like family. One member of each family, together with a v-SNARE localised on the vesicular membrane, are required for fusion.

The Syntaxins are type-I transmembrane proteins that contain several regions with coiled-coil propensity in their cytosolic part, the SNARE motif. SNAP-25 () is a protein consisting of two coiled-coil regions, which is associated with the membrane by lipid anchors. SNARE motifs assemble into parallel four helix bundles stabilised by the burial of these hydrophobic helix faces in the bundle core. Monomeric SNARE motifs are disordered so this assembly reaction is accompanied by a dramatic increase in alpha-helical secondary structure [PUBMED:14570579](#). The parallel arrangement of SNARE motifs within complexes bring the transmembrane anchors, and the two membranes, into close proximity. Recently, it was shown that the two coiled-coil regions of SNAP-25 and one of the coiled-coil regions of the syntaxins are related [PUBMED:9096343](#). This domain is found in both Syntaxin and SNAP-25 families as well as in other proteins.

**Example structure**

**PDB entry 1xtg**: Crystal structure of NEUROTOXIN BONT/A complexed with Synaptosomal-associated protein 25

View a different structure:

Wednesday, September 22, 2010

22

# Annotations in Phytozome

- Annotations are displayed on protein pages
- Functional annotations Pfam, Panther,
- Ortholog and functional predictions: KOG (clusters of orthologous groups), KEGGORTH (KO, or Kyoto Encyclopedia of Genes and Genomes Orthology)

**Manihot esculenta gene cassava22596.valid.m1 :**

About this gene Sequences Peptide Homologs Gene Ancestry

**Info:**

<b>Locus name</b>	cassava22596.valid.m1
<b>Transcript name</b>	cassava22596.valid.m1
<b>Description</b>	
<b>Links to external DBs</b>	

**Functional annotations:**

<b>Pfam:05739</b>	SNARE domain
<b>Panther:19957</b>	SYNTAXIN
<b>KOG:0810</b>	SNARE protein Syntaxin 1 and related proteins
<b>KEGGORTH:08486</b>	STX3S; syntaxin 1B/2/3/4

[Links to descriptions of domains](#)

**Protein domain view:**

# Annotations on cluster pages

- Annotations are also displayed on protein family pages under DOMAINS tab

**Hypothetical Rosid Post-Duplication gene**  
Cluster 2239271. 12 members:  
*Mes Rco Ptr Gma Csa Ath Aly Cpa Vvi*  
2 1 1 1 1 2 2 1 1

Classification Find related families Align family members Get Data Display options

**KOG Class:**  
CELLULAR PROCESSES AND SIGNALING [ U ] : Intracellular trafficking, secretion, and vesicular transport  
**KEGG Class(es):**

Genes in this family Functional Annotation Multiple Sequence Alignment Family History

	ORG	DBXREF	SYMBOL	DEFINITION	DOMAINS	SYNTENY	EXONS
<input type="checkbox"/>			<i>Mes</i> cassava22596.valid.m1				305
<input type="checkbox"/>			<i>Mes</i> cassava23955.valid.m1				305
<input type="checkbox"/>			<i>Ptr</i> POPTR_0019s05240.1	Pl-SYP131.2			305
<input type="checkbox"/>			<i>Ath</i> AT3G03800.1	SYP131			307
<input type="checkbox"/>			<i>Ath</i> AT5G08080.1	SYP132			305
<input type="checkbox"/>			<i>Aly</i> 317280	fgenes11_pm.C_scaffold_3000276			307
<input type="checkbox"/>			<i>Aly</i> 487641	fgenes2_kg.6_775_AT5G08080.1			305

## workshop questions

- Remind yourself what you learnt about GENSAN\_predicted\_peptide\_2 from doing your BLAST searches in the last workshop.
- Copy the peptide sequence from GENSAN\_predicted\_peptide\_2 and paste it into InterproScan, a server which looks for many different protein domains including Pfam and Panther.
- Go to <http://www.ebi.ac.uk/Tools/InterProScan/>
- Click on Clear All then select just HMMPfam and HMMPanther to run these two domain searches. Paste GENSAN\_predicted\_peptide\_2 into the sequence window and Submit the Job.
- In a few seconds you should get results on your screen. What domain(s) are predicted in this protein and what's the most obvious difference between the Panther and Pfam domains? What does the presence of these domains tell you about the function of this protein? You can get more information about the domains by clicking on the Panther IDs (these look like this PTHR12345) or Pfam IDs (these look like this PF12345). Note that Panther domains link you to the Gene Ontology (GO) terms associated with each domain.
- The Phytozome protein page for this gene is here <http://www.phytozome.net/genePage.php?search=1&detail=1&crowd&method=0&searchText=transcriptid%3A17528957>
- The Adenylate cyclase-associated CAP, N-terminal domains is missing from the protein page because Phytozome uses a simpler threshold than InterProScan for predicting domains. This will be corrected in the future.