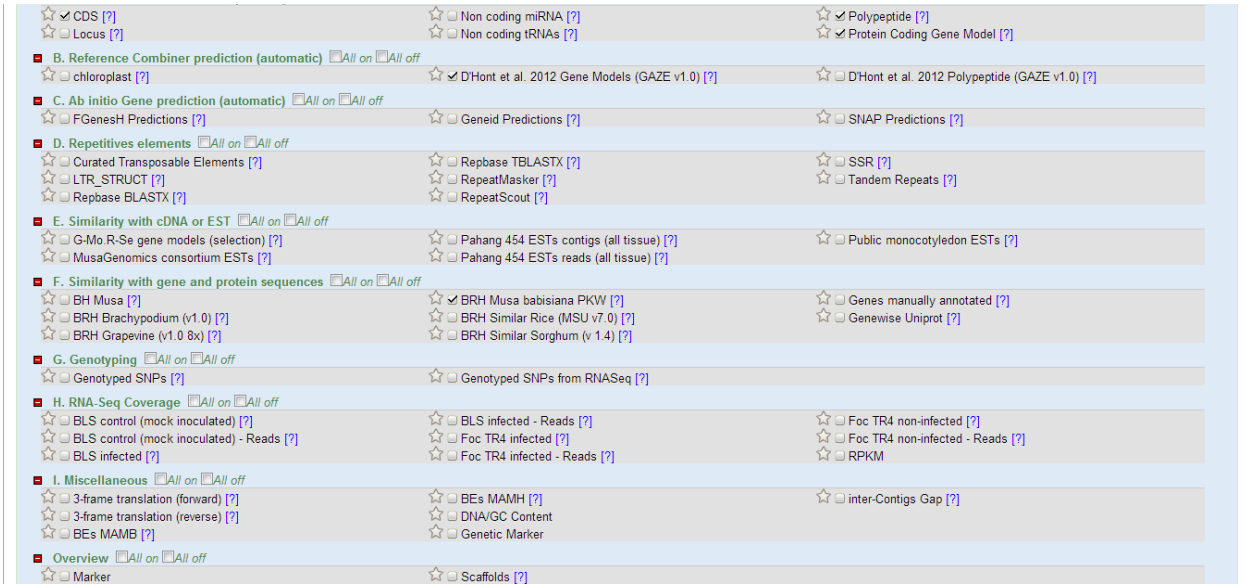
Practical session 1:

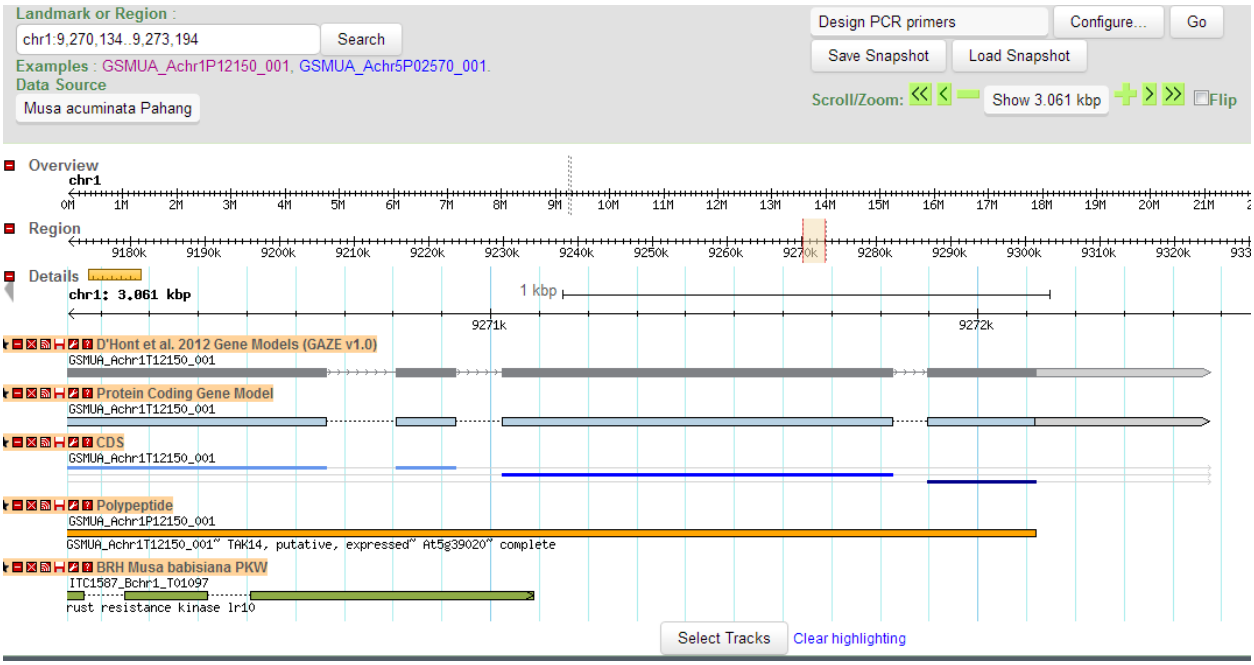
Banana Genome Hub and GreenPhyl

Exercise 1: Introduction to Gbrowse

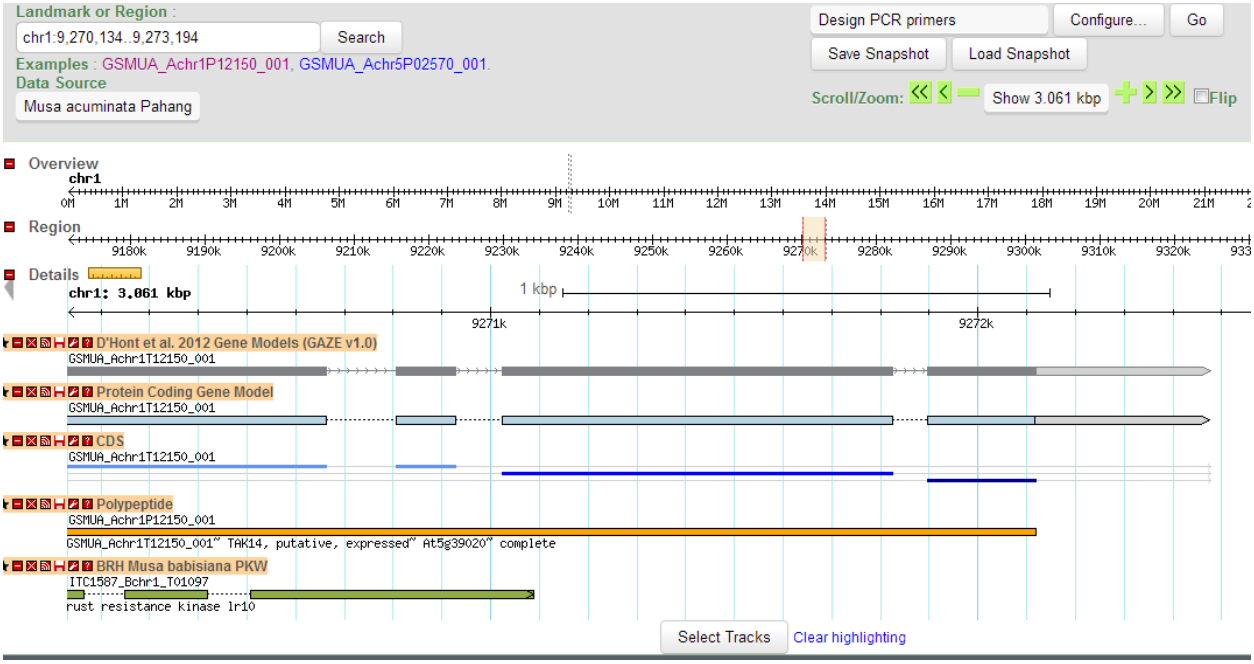
1. Go to the Genome browser http://banana-genome.cirad.fr/cgi-bin/gbrowse/musa_acuminata/
2. Look at the “select tracks” tab and look at all the categories.
3. Select the following tracks
    a. protein Coding Gene Model
    b. CDS
    c. polypeptide
    d. D'Hont et al. 2012 Gene Models (GAZE v1.0)
    e. BRH Musa babisiana PKW



4. Get back to Browser tab and look at the 4 sections of the page
    a. Search
    b. Overview
    c. Region
    d. Details
5. Click on the example GSMUA_Achr1P12150_001
6. Drag and drop some tracks tracks. For instance, you can order them like
    a. D'Hont et al. 2012 Gene Models (GAZE v1.0)
    b. protein Coding Gene Model
    c. CDS
    d. polypeptide

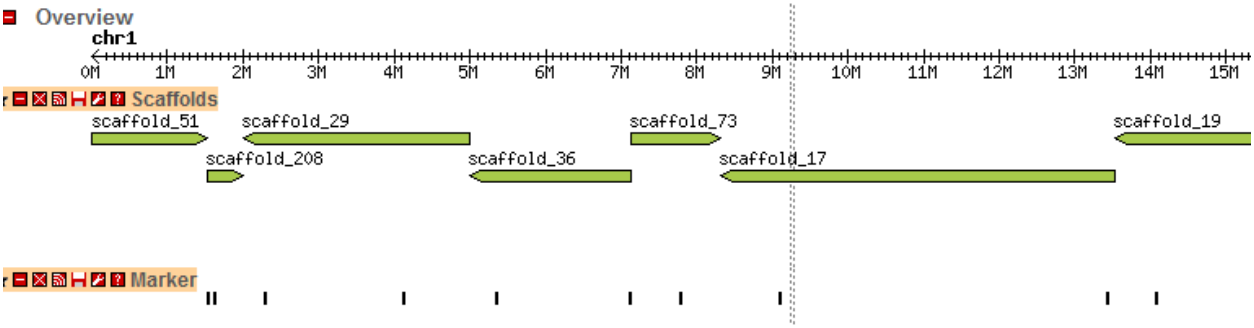


7. Click on the Protein Coding Gene Model



a. open GBrowse report
b. open Gene report and look at the menu on the right (cross-references, sequence, GO Assignments)

8. Back to Gbrowse, add tracks
    a. Scaffold
    b. Markers
    c. genetic markers



9. Zoom out to display a ~20kbp genomic region

Exercise 2: Hub overview

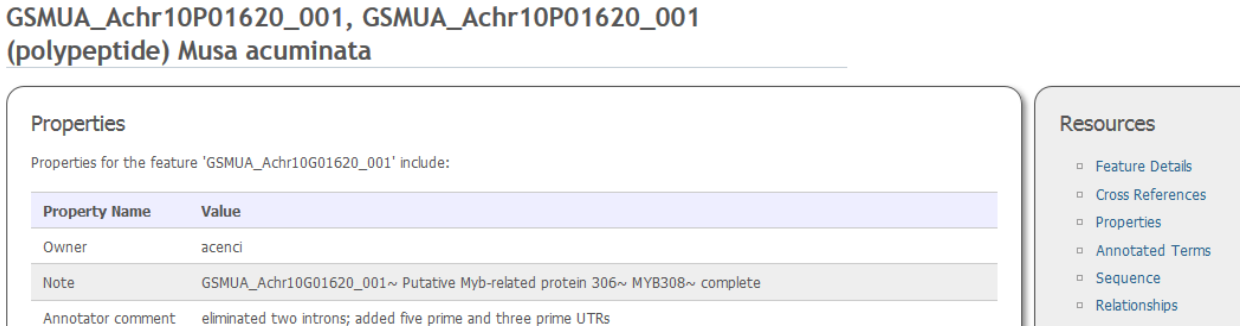
1. Search the best hit to the following sequence using Blast

>seq1
ATGGGAAGGCCTCCTTGCTGTGATAACATTGGCATCAAGAAAGGACCATGGACTCCTGAGGAGGAC
ATCGTCTTGGTCTCTTATATTCAGGAACATGGACCTGGAAACTGGAGATCAGTTCCCACAAGCACAG
GGTTGATGAGATGCAGTAAGAGCTGTAGATTGAGATGGACTAACTACCTCAGGCCTGGAATCAAACG
CGGCAACTTCACTCCGCATGAAGAACGAGTTATCATCCATCTCCAATCCTTGCTTGGCAACAGATGG
GCAGCCATTGCCTCTTACCTTCCCCAAAGAACCGACAATGATATCAAGAACTACTGGAACACACATCT
CAAGAAGAAGATCAACAAGATCCAGGGAGCTGCAGATGCAGATGGCAAGAAGCCCTCTTCTGATGC
TAGGCCTGATTGCCATGACTACGTGTTCCAAATCTACAAGATGATGGAATCAAGGAAGCAGGACCTC
GCCGCCACACTCCCCAGCTATCACCAGAAGTCGAGGTATGCCTCCAGCAGCGAGAACATCTCGAGG
CTCCTCCAGGGGTGGATGCAGTCATCGCCAACGGTCAACGCGCCAGGGAAGTTGAAAGAATCATGC
TCCACCGCCGACGATAACGACGATGAGAACAGCAACATCATCACCGCCCTTACAGCAGCGTCACTA
ATGGAGAACAGTCAAGCTGAAGGCGACCGAGGGAGCTGCGCCCCCATGACGCACGATGACTTCGA
CCTGCTGCATTCCTTCGAAAGCATGGACACTG

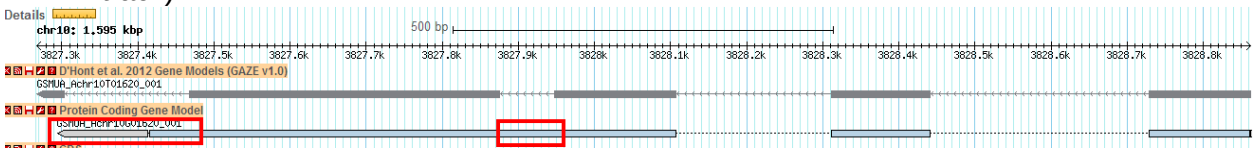
2. What is locus tag name of the best hit? which Chromosome? which positions?
    - GSMUA_Achr10G01620_001
    - Chromosome 10
    - start 3827413 and stop 3828861

3. What is its functional annotation?
    - Myb transcription factor

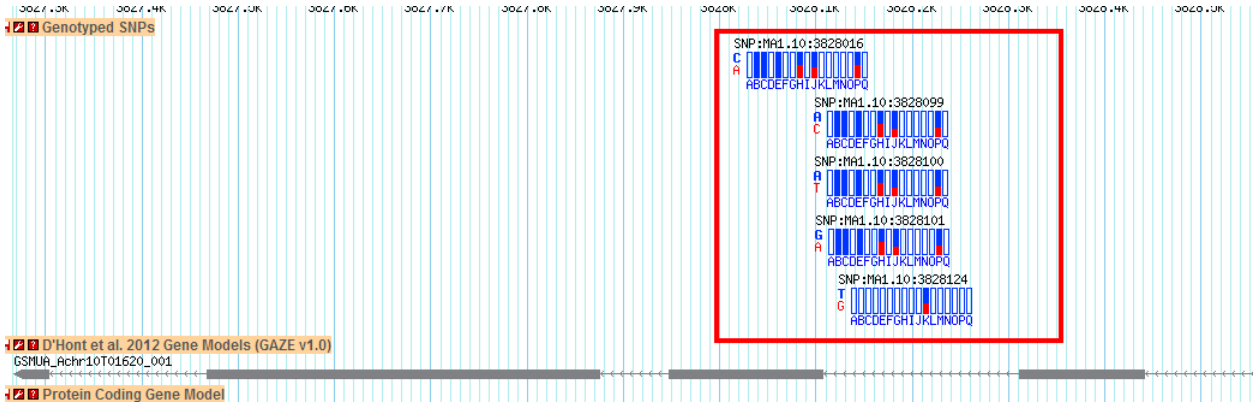
4. Was it manually curated? what has been done?
    - Curated by acenci (owner) n.b. when no curated the owner is musa
    - 2 introns deleted and UTR added



5. visualize the modifications in Gbrowse (compare automatic and manual gene models track)



6. is there any allelic variant? (SNP genotyping track)

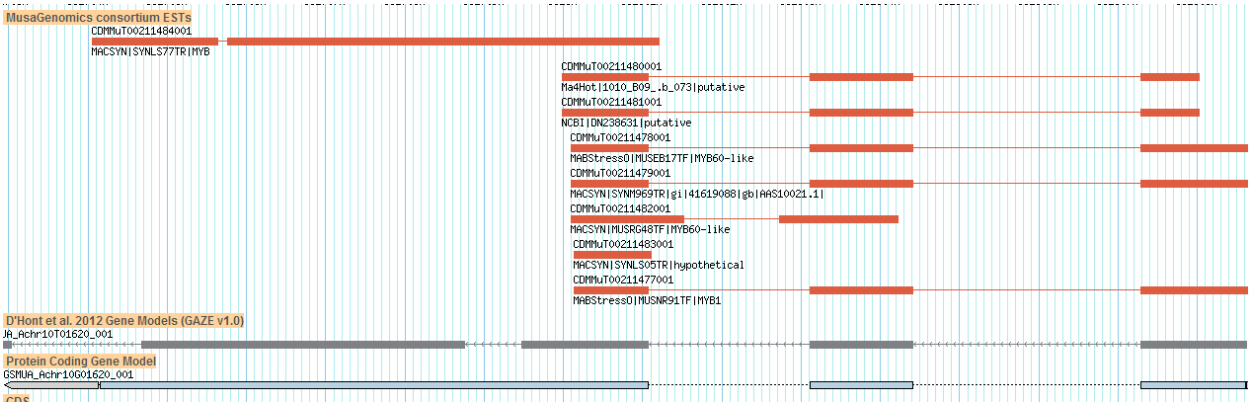


pick up one

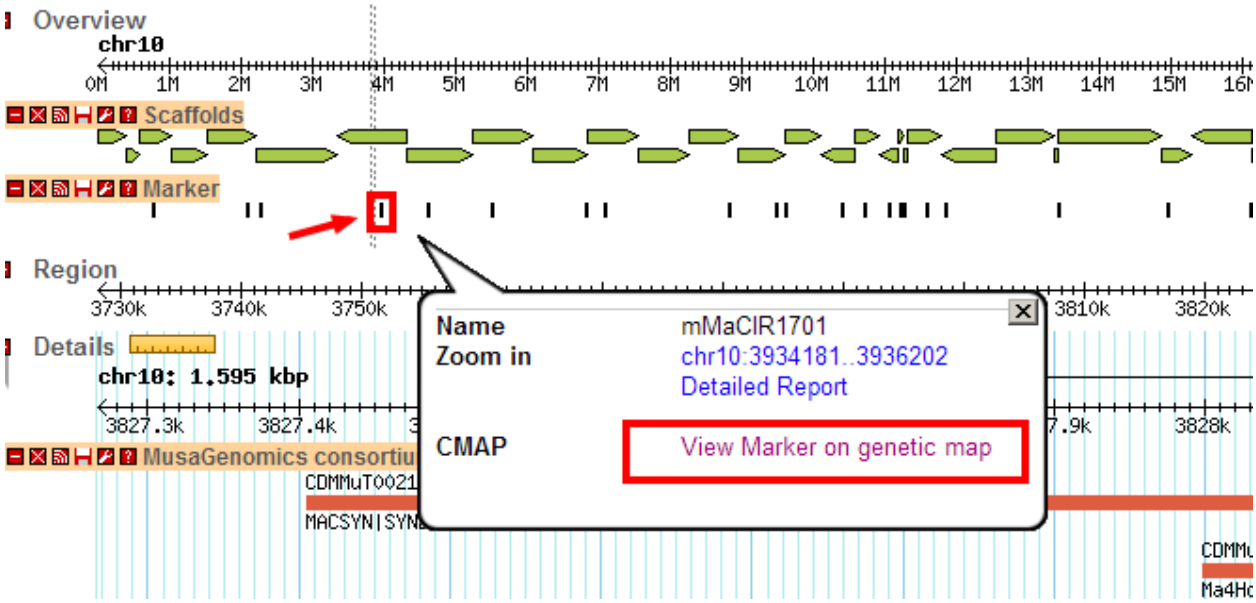
- In how many of the cutilvars [6]
- what is the allele of reference? Which position?
    i. C at position 3828016

7. Is this gene expressed?

Select tracks for transcriptomics data such as ESTs or RNAseq



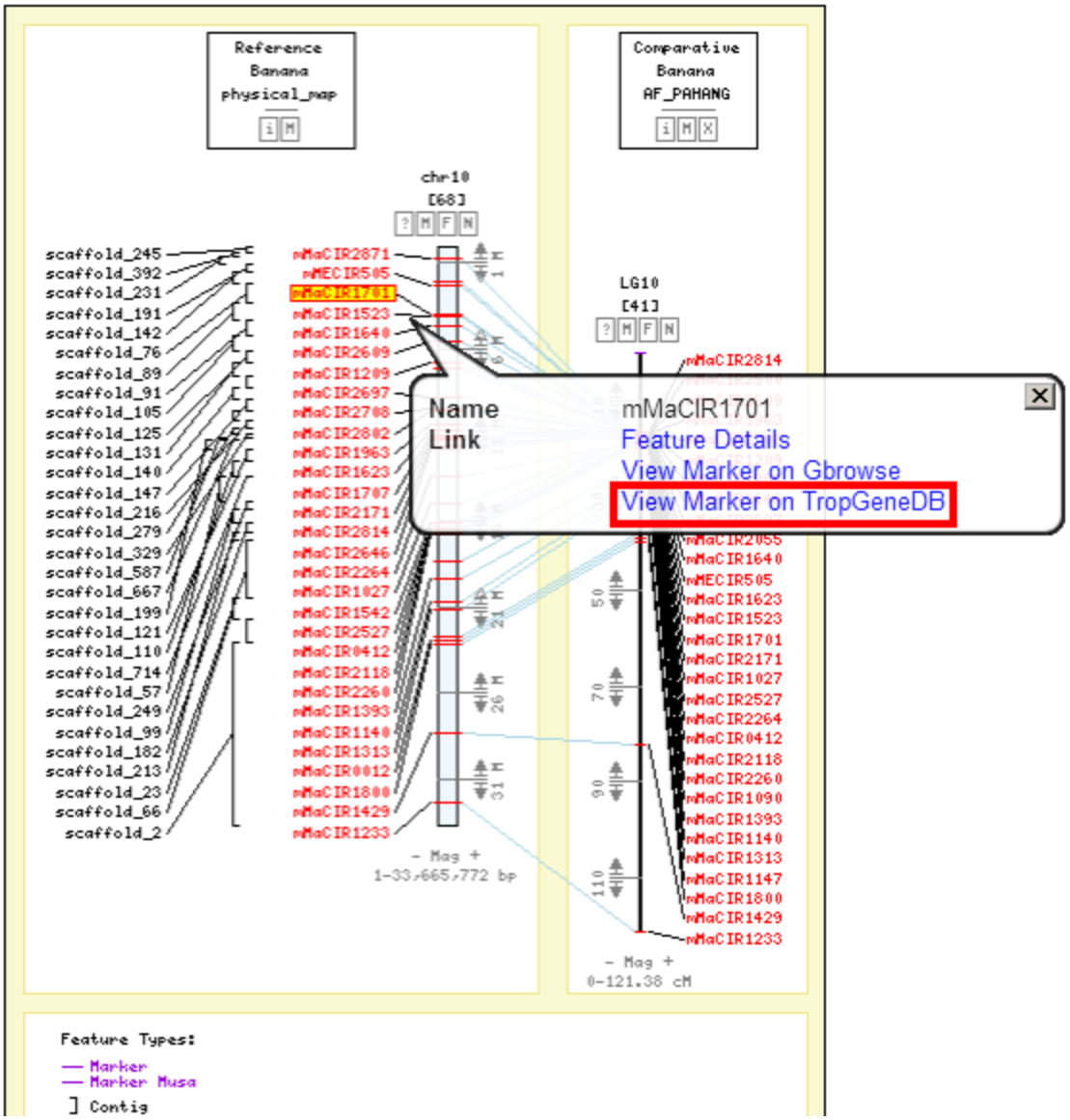
8. Is there any marker close to the gene? (use Markers and/or genetic marker tracks and zoom out)
   mMaCIR1701



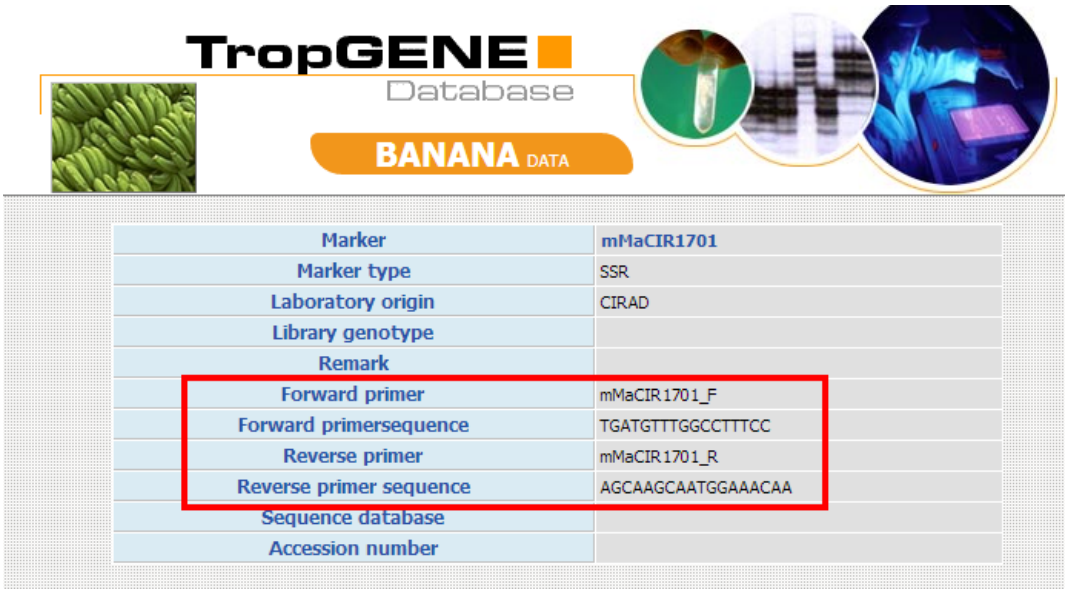
   - if yes, do they belong to any genetic map? what is name? [AF Pahang -LG10]
   - Which scaffold of the physical does it belong to? [scaffold_89]

cmap_viewer

Maps | Map Search | Feature Search | Matrix | Map Sets | Feature Types | Map Types | Evidence

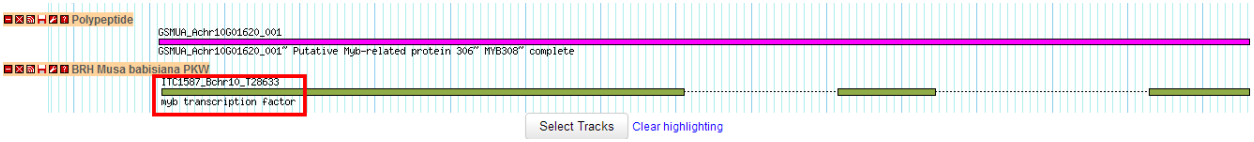


- check type of marker and primers availability in TropGeneDB. [SSR]

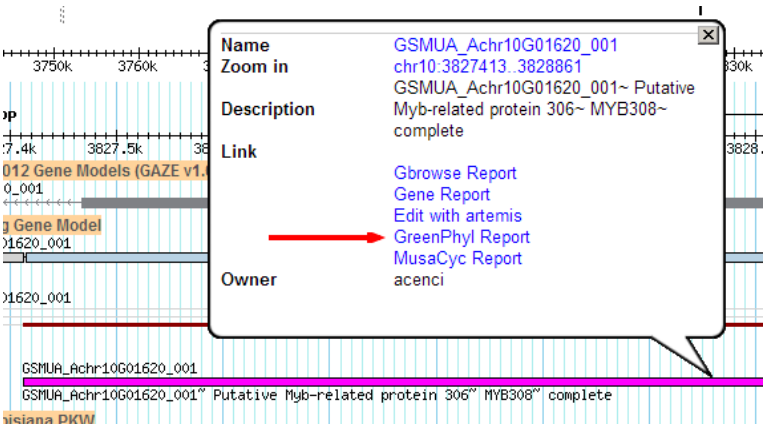


Marker	mMaCIR1701
Marker type	SSR
Laboratory origin	CIRAD
Library genotype	
Remark	
Forward primer	mMaCIR1701_F
Forward primersequence	TGATGTTTGGCCTTTCC
Reverse primer	mMaCIR1701_R
Reverse primer sequence	AGCAAGCAATGGAAACAA
Sequence database	
Accession number	

9. What is the reciprocial best hit (RBH) in the PKW genome (B genome)?



10. Which family gene does the sequence belong to? [GP000011 - Myb transcription factor family]



Exercise 3: Quick search and overview

1. Go to the GreenPhyl website http://www.greenphyl.org/
2. Go to the quick search and Search for the keyword Dehydrins

Quick Search results for Family Name

Search results for families with name corresponding to 'dehydrins'

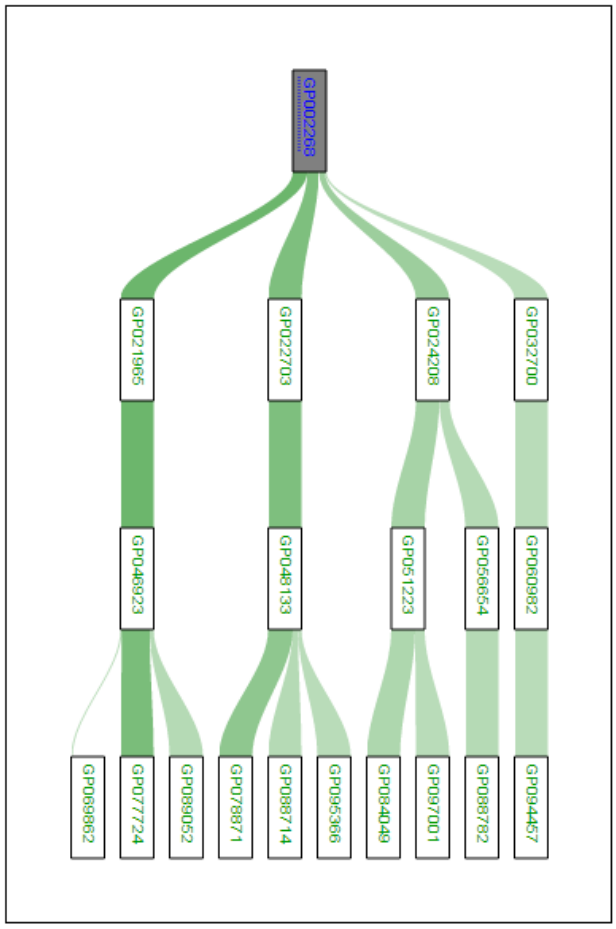
[Excel | CSV | XML]

Matching families:

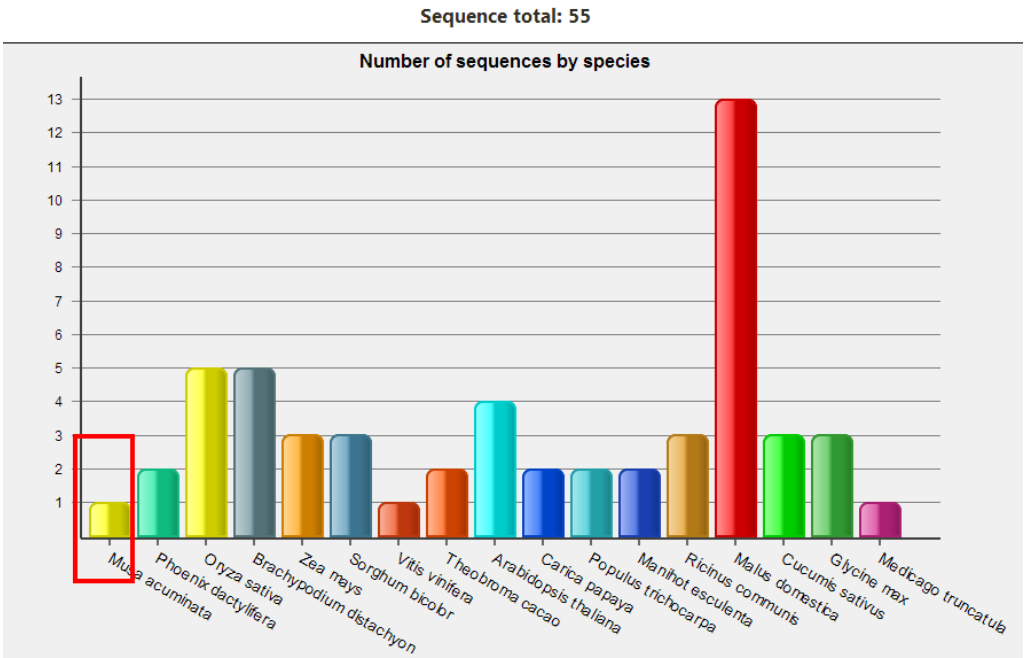
Family Id	Family Name	Number of sequences	Status	Analyzes
GP002268	Dehydrins Y2SK2	55		
GP003560	Dehydrins SK3	33		
GP005304	Dehydrins KS	14		
GP009271	Dehydrins KS	3		Not available
GP005537	Dehydrins KS	11		

Note: sort multiple columns simultaneously by holding down the shift key and clicking other column headers.

3. Open the Dehydrins Y2SK2 and look at the identity card of the Gene family
4. Display the advanced mode and to look at the flow of sequences.



5. Look at the species distribution on the bar chart. What do you notice?



The gene family is specific to angiosperms. No sequence in P. patens etc.

6. What are the predicted orthologs for the Musa sequence?

Display Ortholog for (default: all):

Species

Check/uncheck all

Arabidopsis thaliana	Brachypodium distachyon	Carica papaya	Cucumis sativus
Glycine max	Malus domestica	Manihot esculenta	Medicago truncatula
Musa acuminata	Oryza sativa	Phoenix dactylifera	Populus trichocarpa
Ricinus communis	Sorghum bicolor	Theobroma cacao	Vitis vinifera
Zea mays			

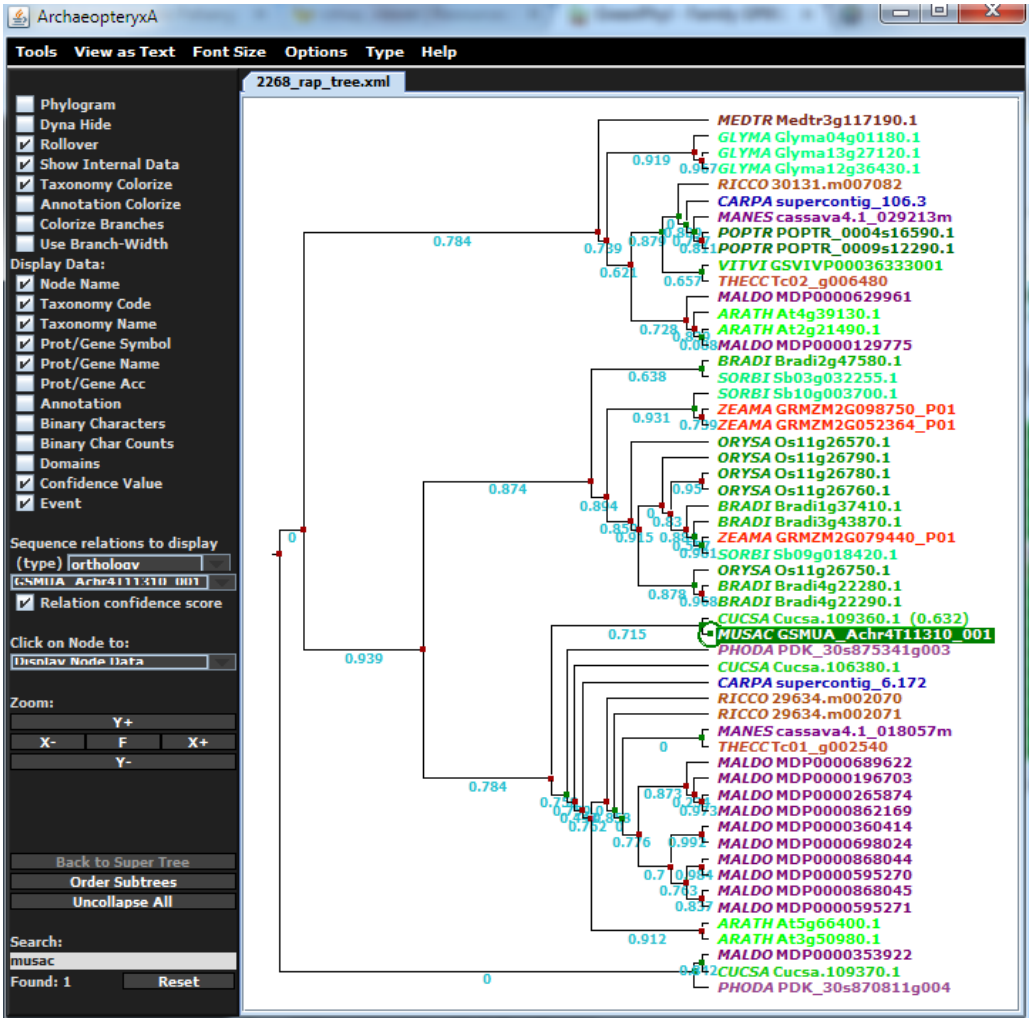
Display table

[Excel | CSV | XML | FASTA]

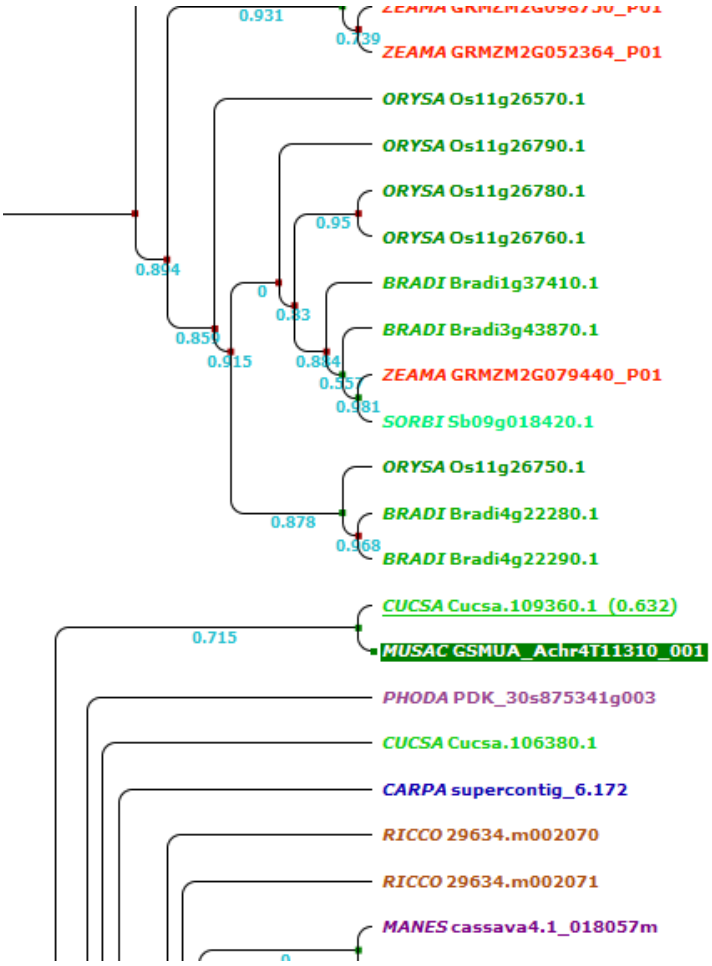
Family Sequence ID	Similar Sequence ID	Homology			BBMH	
		Type	Evolutionary distance	Node distance	score	e-value
Cucsa.109360.1	GSMUA_Achr4T11310_001	orthology	0.416	1	-	-

Note: sort multiple columns simultaneously by holding down the shift key and clicking other column headers.

7. Look at the phylogenetic results and visualize the gene tree in both viewer (Archeopteryx and IntreeGreat)
    a. Highlight the sequence with the viewers



b. What topology (pattern of branching) do you notice?



c. Does it look consitent with the species tree? what are the possible explanations?

Not really. The musa gene is clustered with cucumber and is located in a clade with dicots. It would be required to checked structural annotation and the divergence of this sequence.

8. Open the Musa sequence page.
9. Go the Banana Genome Hub (cross-references link) and check the status of the sequence (gene history)

Gene GSMUA_Achr4T11310_001

Sequence ID	GSMUA_Achr4T11310_001 add to my list
Species	Musa acuminata
Alias	No gene alias
Length	97aa
Cross-reference(s)	Hide
	Genome databases GBrowse
	Metabolic Pathway databases MusaCyc
Gene Annotation	Dehydrin Xero 1
Gene Ontology	Display term(s) (2)
Filter status	Not filtered

GreenPhyl contains automatic annotation published and according to the Hub. the sequences is partial and needed to be corrected.

GSMUA_Achr4P11310_001, GSMUA_Achr4P11310_001 (polypeptide) Musa acuminata

Properties

Properties for the feature 'GSMUA_Achr4G11310_001' include:

Property Name	Value
Owner	acenci
Note	GSMUA_Achr4G11310_001~ Dehydrin Xero 1~ ECP40~ complete
Annotator comment	extended coding sequence in five prime region; extended five prime and three prime UTRs
Color	2
Date	Tue Dec 7 21:17:29 CET 2010
Inference	{gaze}
Alternative splicing	to_fill
Original location	join(8106521..8106637,8106707..8106880)

Exercise 4: Advanced searches

1. Using search in the top banner, search for P37271 corresponding to the UniProt identifier of the phytoene syntase in Arabidopis involved in the carotenoid biosynthesis pathway.

Matching sequences:

Sequence	Species	Alias	Accession	Description	Annotation
AT5G17230.3	Arabidopsis thaliana	PSY1	P37271	Phytoene synthase, chloroplastic	| Symbols: | phytoene synthase (PSY) / geranylgeranyl-diphosphate geranylgeranyl transferase | chr5:5659839-5662087 REVERSE

Note: sort multiple columns simultaneously by holding down the shift key and clicking other column headers.

a. What is the gene family identifier?

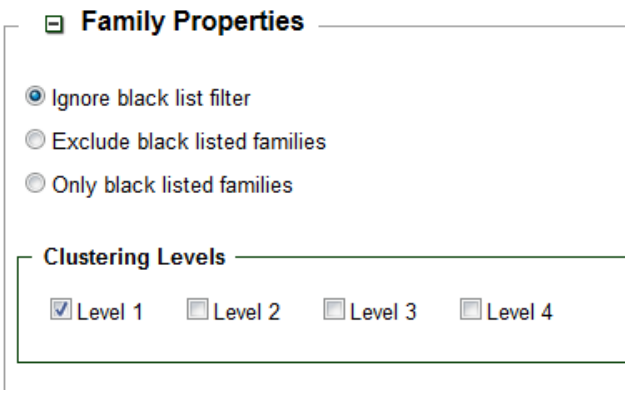
GP001432 phytoene synthase family

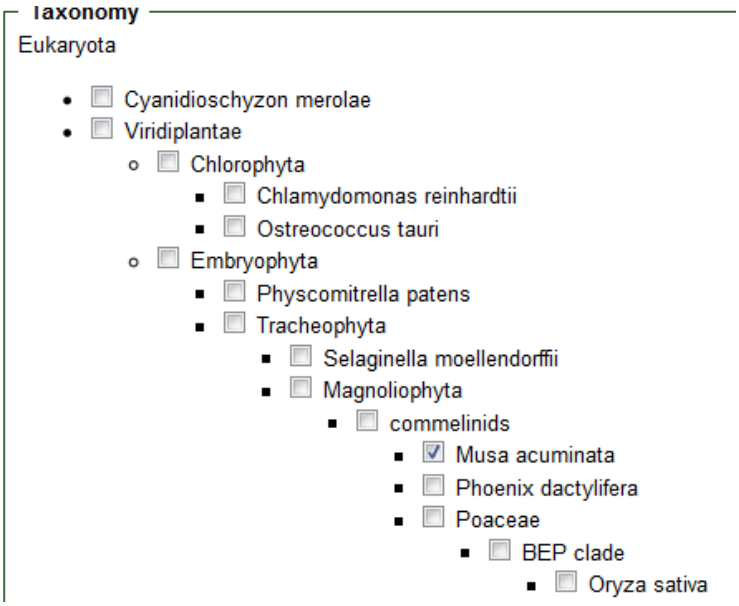
b. How many homologs in Musa?

The gene family contains 4 sequences of Musa. 2 only are predicted as orthologs.

2. Search gene families with at least one banana gene

At level 1

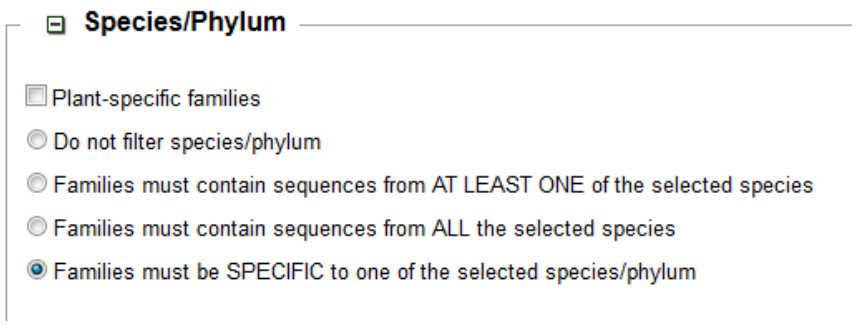


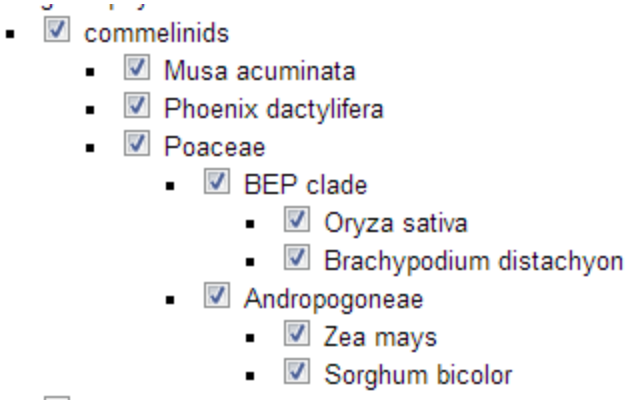


Number of clusters: 4686

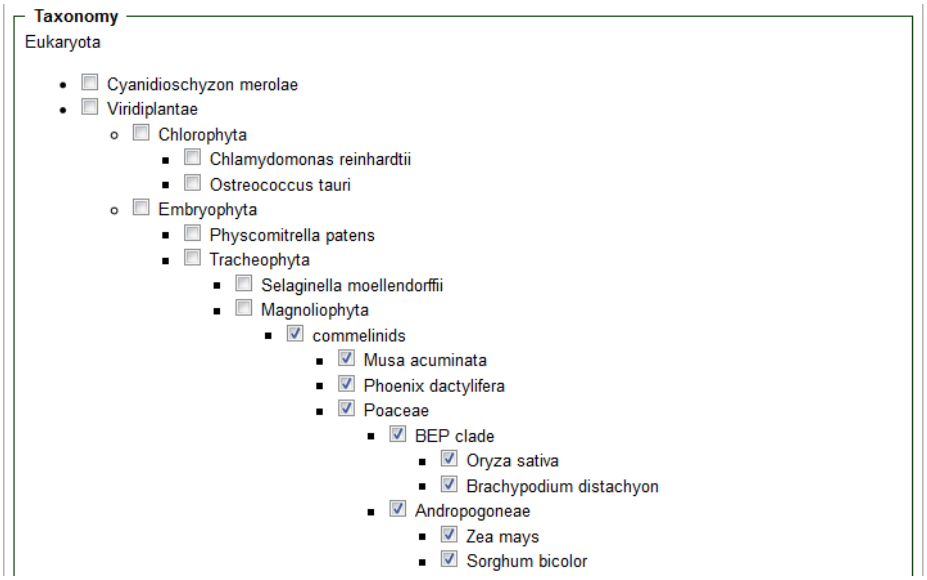
Family Id	Family Name	Number of sequences	Status	Analyzes
GP000001	Unannotated cluster	5885		Not available
GP000002	Unannotated cluster	9942		Not available
GP000003	Unannotated cluster	4662		Not available
GP000004	Kinase and/or LRR superfamily	6998		Not available

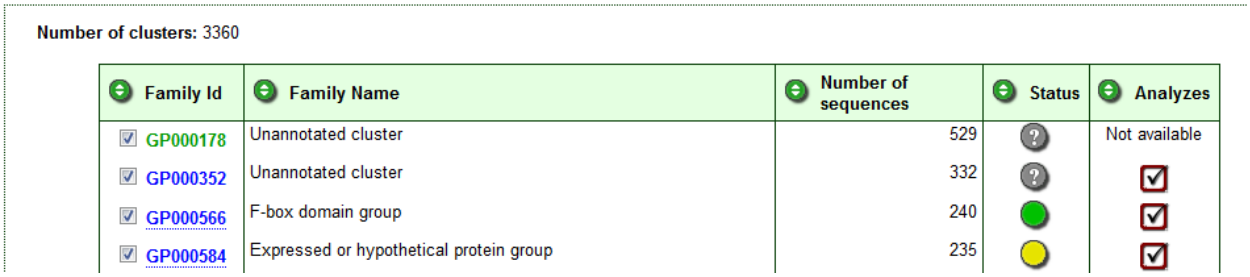
3. Search gene families specific of the monocotyledons (commelinids)





4. Search gene families with with at least one banana gene and one rice gene





Number of clusters: 3360

Family Id	Family Name	Number of sequences	Status	Analyzes
GP000178	Unannotated cluster	529	?	Not available
GP000352	Unannotated cluster	332	?	☑
GP000566	F-box domain group	240	●	☑
GP000584	Expressed or hypothetical protein group	235	●	☑

Exercise 5: Application for RNAseq

Let's say that you performed a run an illumina RNAseq for *Musa acuminata* cavendish cultivars (AAA). The resulting reads were mapped on the Musa acuminata DH Pahang genome and you obtained the following list of gene ids.

GSMUA_Achr10T18460_001	GSMUA_Achr3T02890_001
GSMUA_Achr10T24860_001	GSMUA_Achr3T26090_001
GSMUA_Achr10T27580_001	GSMUA_Achr3T28450_001
GSMUA_Achr11T13360_001	GSMUA_Achr3T30550_001
GSMUA_Achr11T18690_001	GSMUA_Achr3T32220_001
GSMUA_Achr11T18740_001	GSMUA_Achr4T02660_001
GSMUA_Achr11T22990_001	GSMUA_Achr4T02930_001
GSMUA_Achr1T14320_001	GSMUA_Achr4T10090_001
GSMUA_Achr1T24730_001	GSMUA_Achr6T15150_001
GSMUA_Achr2T07990_001	GSMUA_Achr6T22330_001
GSMUA_Achr2T22340_001	GSMUA_Achr6T27210_001
GSMUA_Achr3T01960_001	GSMUA_Achr7T11920_001
GSMUA_Achr8T18780_001	GSMUA_Achr7T22540_001
GSMUA_AchrUn_randomT02840_001	GSMUA_Achr8T02150_001

1. Check their annotation and locations on the chromosomes using the Locus search on the Banana genome Hub

Copy and paste the sequences in this locus search
http://banana-genome.cirad.fr/advanced

2. Check their Gene Family distribution using Toolbox ‘sequence to families’ on GreenPhyl

copy and paste sequence identifier
http://www.greenphyl.org/cgi-bin/seq2families.cgi

   a. explore some of the genes families
   b. what type of functional classes did you see?

3. Search for the ortholog genes in the other species using Toolbox ‘Homolog sequences’ on GreenPhyl

copy and paste sequence identifiers
http://www.greenphyl.org/cgi-bin/get_homologs.cgi

4. Export sequences at fasta format

copy and paste sequence identifiers
www.greenphyl.org/cgi-bin/export_sequences.cgi

Exercise 6: InterPro Domain Distribution (ipr2genomes)

You want to identify the Jumonji transcription factor (TFs) in Plants. According to Lang et al, 2011, Jumonji are characterized by a combination of protein domains. All sequences must have the domains:

- JmjC - IPR003347
- JmjN - IPR003349

but should not contain:

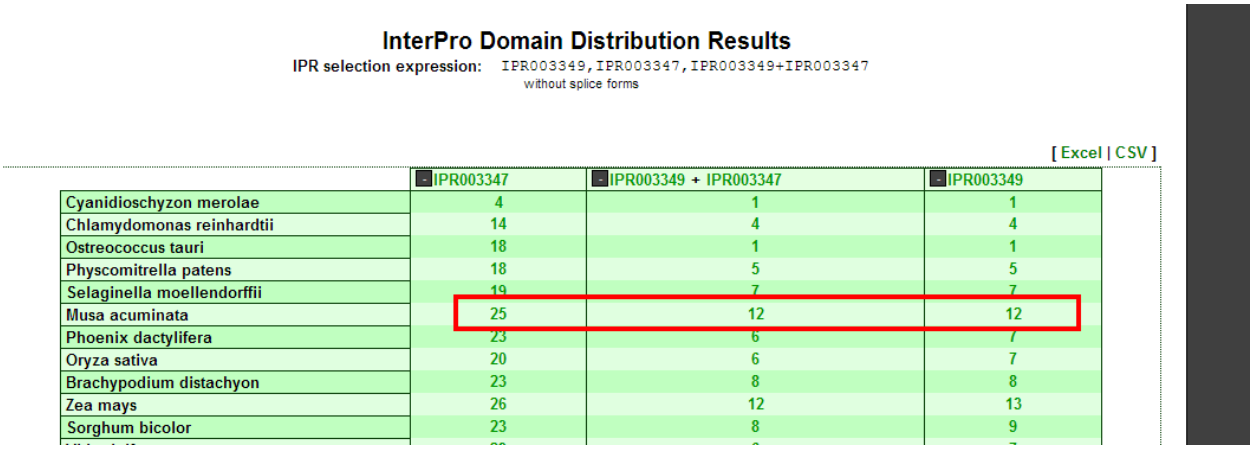
- ARID - IPR001606
- GATA - IPR000679
- zf-C2H2 - IPR007087
- Alfin-like - IPR021998

1. How many sequences have the JmjC domain? the JmjN domain? shared? [25,12,12]

Go to http://www.greenphyl.org/cgi-bin/ipr2genomes.cgi

Enter IPR003349,IPR003347,IPR003349+IPR003347

You will find 25,12 and 12 respectively.



InterPro Domain Distribution Results

IPR selection expression: IPR003349,IPR003347,IPR003349+IPR003347
without splice forms

[Excel | CSV]

	IPR003347	IPR003349 + IPR003347	IPR003349
Cyanidioschyzon merolae	4	1	1
Chlamydomonas reinhardtii	14	4	4
Ostreococcus tauri	18	1	1
Physcomitrella patens	18	5	5
Selaginella moellendorffii	19	7	7
Musa acuminata	25	12	12
Phoenix dactylifera	23	6	7
Oryza sativa	20	6	7
Brachypodium distachyon	23	8	8
Zea mays	26	12	13
Sorghum bicolor	23	8	9

2. How many Jumonji sequences in Musa?

IPR003349+IPR003347-IPR001606-IPR000679-IPR007087-IPR021998
The result is 8

InterPro Domain Details

IPR selection expression: IPR003349+IPR003347-IPR001606-IPR000679-IPR007087-IPR021998
without splice forms

Check classification [FASTA

Sequence name	Full IPR details
GSMUA_Achr1T05670_001	IPR003888 IPR003889 IPR003347 IPR003349 IPR004198 IPR013129 IPR018516 IPR018518
GSMUA_Achr1T06290_001	IPR003347 IPR003349 IPR004198 IPR013129
GSMUA_Achr1T09250_001	IPR003888 IPR003889 IPR003347 IPR003349 IPR004198 IPR013129 IPR018516 IPR018518
GSMUA_Achr2T18080_001	IPR003347 IPR003349 IPR004198 IPR013129
GSMUA_Achr5T18050_001	IPR003347 IPR003349 IPR004198 IPR013129
GSMUA_Achr6T19910_001	IPR003888 IPR003889 IPR003347 IPR003349 IPR004198 IPR013129 IPR018516 IPR018518
GSMUA_Achr9T20510_001	IPR003347 IPR003349 IPR013129
GSMUA_Achr9T30480_001	IPR003888 IPR003889 IPR003347 IPR003349 IPR004198 IPR013129 IPR018516 IPR018518

3. Compare wth other genomes? What do you observe?

Except Soybean (Glycine Max), Musa genes numbers are the more important among the other genomes.

4. Do the Musa genes all belong to the same gene family?

click on the check classification button.