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
    e. BRH Musa balbisiana 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 protein Coding Gene Model
    a. open gene history
    b. open gene report
        i. 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? [

3. What is its functional annotation?

4. Was it manually curated? What has been done? visualize the modifications in Gbrowse (compare automatic and manual gene models track)

5. Is there any allelic variant? (SNP genotyping track)

pick up one

   a. In how many of the cultivars
   b. What is the allele of reference? Which position?
   c. Download track data

6. Is this gene expressed?

7. Is there any marker close to the gene? (use Markers and/or genetic marker tracks and zoom out)
   a. If yes, do they belong to any genetic map? What is name?
   b. Check type of marker and primers availability in TropGeneDB.
   c. Which scaffold of the physical does it belong to?

8. What is the reciprocal best hit (RBH) in the PKW genome (B genome)?

9. Which family gene does the sequence belong to?

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”
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?
6. What are the predicted orthologs for the Musa sequence?
7. Look at the phylogenetic results and visualize the gene tree in both viewer (archaeopteryx and IntreeGreat)
   a. Highlight the sequence with the viewers.
   b. What topology do you notice?

    c. Does it look consistent with the species tree? what are the possible explanations?
8. Open the Musa sequence page. Look at the number of exon.
9. Go the Banana Genome Hub (cross-references link) and check the status of the sequence (gene history)

Exercise 4: Advanced searches

1. Using search in the top banner, search for P37271 corresponding to the UniProt identifier of the phytoene syntase in Arabidopsis involved in the carotenoid biosynthesis pathway.
    a. What is the gene family identifier?
    b. How many homologs in Musa?
2. Search gene families with at least one banana gene
3. Search gene families specific of the monocotyledons (commelinids)
4. Search gene families with at least one banana gene and one rice gene

Exercise 5: Application for RNAseq

Let's say that you performed a run of 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
2. Check their Gene Family distribution using Toolbox 'sequence to families' on GreenPhyl

    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
4. Export sequences at fasta format

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?
2. How many Jumonji sequences in Musa? Compare wth other genomes? What do you observe?
3. Do the Musa genes all belong to the same gene family?