C. VIRGINICA EXPANDED, CONTRACTED, SHARED GENES FROM CAFÉ ANALYSIS

Erin M. Roberts, PhD. Candidate URI

Marta Gomez-Chiarri Laboratory

May 13th, 2019

CAFÉ Data

- Downloaded proteins sequences predicted from these genomes from NCBI
  - *Crassostrea gigas*, *Crassostrea virginica*, *Mizuhopecten yessoensis*, *Biomphalaria glabrata* and *Octopus bimaculoides*
- Only those annotated in the NCBI pipeline were used

Conducted by Fábio Kuriki Mendes

CAFÉ Gene Family Generation

- Alignments created using BLASTP analyses, in an all-by-all fashion, and then grouping sequences (into gene families) based on their top BLASTP hits using mcl with an inflation parameter of 3 (Enright et al., 2002). Sequences within a group were aligned with MUSCLE (version 3.8.31; Edgar, 2004).

Conducted by Fábio Kuriki Mendes

Gene Families Used for CAFÉ Analysis

- Gene families excluded if only one species contributed to copy counts, or if one species had more than 1,000 gene copies
- Split gene families into two groups
  - one or more species had counts larger than 50 (146)
  - the remaining families (12,365)
- Group II families used to estimate λ to minimize the effects of saturation
- Group I families analyzed by fixing the λ value estimated from the smaller families.

Conducted by Fábio Kuriki Mendes

CAFÉ Output

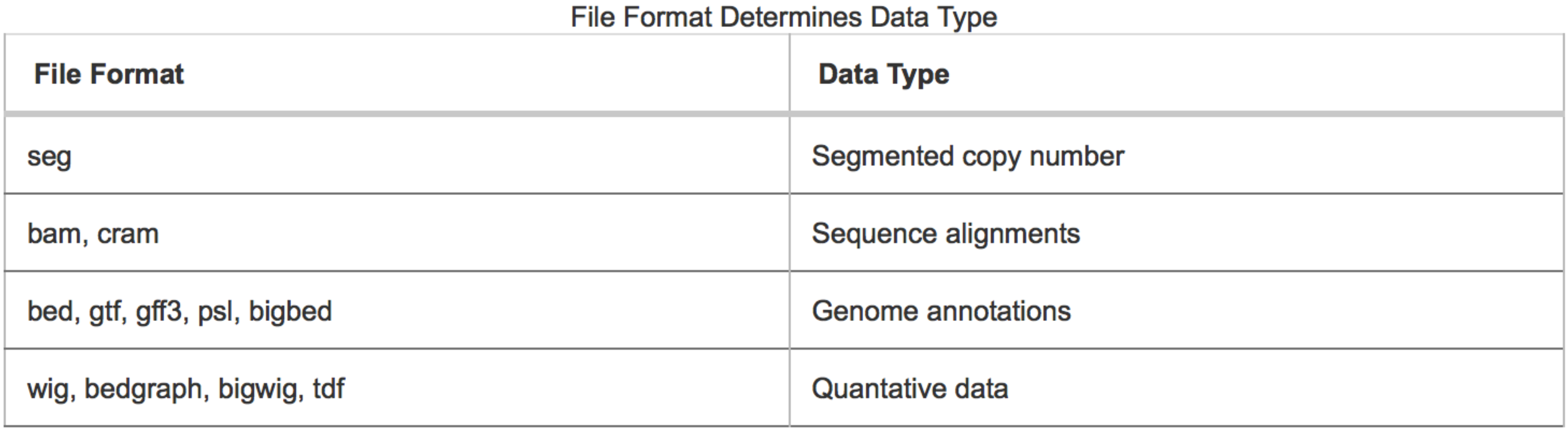
- Gene family fasta files with groups of unannotated protein sequences in a particular “gene family”
- Families grouped into expanded, contracted, ancestral, and shared folders

Conducted by Fábio Kuriki Mendes

BED FILES IN IGV

Tracks in IGV

- IGV uses file extension to determine file format and display settings



File Format Determines Data Type

File Format	Data Type
seg	Segmented copy number
bam, cram	Sequence alignments
bed, gtf, gff3, psl, bigbed	Genome annotations
wig, bedgraph, bigwig, tdf	Quantative data

IGV User Guide, Broad Institute

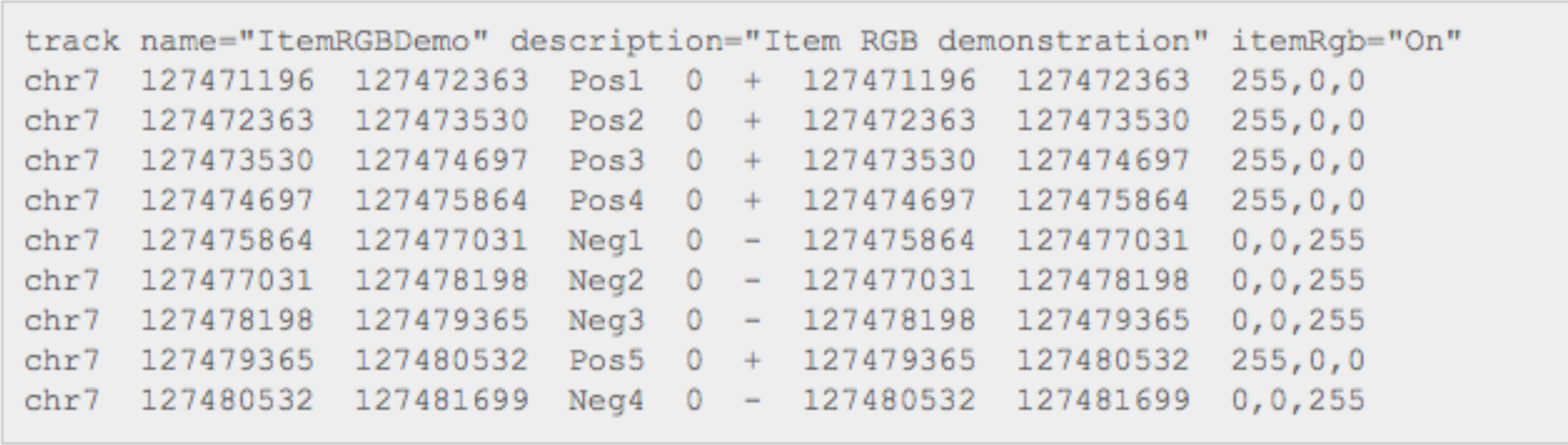
Track Lines: Changing How Data is Viewed

- Track lines can be added at the top of a data file to change how IGV track is displayed
- File formats that allow track lines in IGV:
    - BED, WIG, PSL
    - Track line must begin with # symbol: IGV, SNP, GFF, GFF3,SEG, LOH, CN
- Should be placed at the beginning of the list of features they are to affect

[1] Ensembl BED file format:https://useast.ensembl.org/info/website/upload/bed.html#tracklines

Track line options

- Track line starts with words “track name” and followed by space-separated key value pairs
  - Name: unique name to identify track
  - description: label under track
  - priority: describe which order to display tracks
  - color: RGB or hexidecimal
  - useScore: set 1 to render track in grayscale
  - itemRgb: if “on” specified color values will be used
  - #gffTags: will show your name column in Gff file format

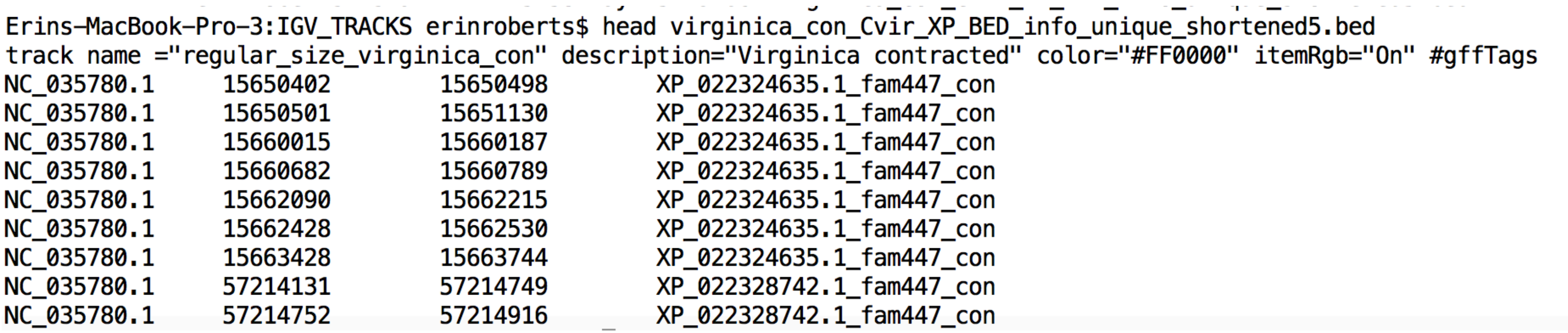


```
track name="ItemRGBDemo" description="Item RGB demonstration" itemRgb="On"
chr7  127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7  127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7  127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7  127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7  127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7  127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7  127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7  127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7  127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

[1] Ensembl BED file format:https://useast.ensembl.org/info/website/upload/bed.html#tracklines

BED file format

- Used for genome annotations[1]
- Required:
    - Chrom
    - chromStart
    - chromEnd
- Select Optional fields:
    - name: label displayed under feature



```
Erins-MacBook-Pro-3:IGV_TRACKS erinroberts$ head virginica_con_Cvir_XP_BED_info_unique_shortened5.bed
track name ="regular_size_virginica_con" description="Virginica contracted" color="#FF0000" itemRgb="On" #gffTags
NC_035780.1     15650402        15650498        XP_022324635.1_fam447_con
NC_035780.1     15650501        15651130        XP_022324635.1_fam447_con
NC_035780.1     15660015        15660187        XP_022324635.1_fam447_con
NC_035780.1     15660682        15660789        XP_022324635.1_fam447_con
NC_035780.1     15662090        15662215        XP_022324635.1_fam447_con
NC_035780.1     15662428        15662530        XP_022324635.1_fam447_con
NC_035780.1     15663428        15663744        XP_022324635.1_fam447_con
NC_035780.1     57214131        57214749        XP_022328742.1_fam447_con
NC_035780.1     57214752        57214916        XP_022328742.1_fam447_con
```

[1] Ensembl BED file format:https://useast.ensembl.org/info/website/upload/bed.html#tracklines

Making a BED file

- Just a tab separated file that has the three required columns in correct order, no column names, and added track line added at top!
- To make from a .csv file, open in excel, add track line as the first line, save as tab delimited file
- Change ending file extension to .bed
- IGV currently doesn't support having multiple tracks in a single BED file[1]

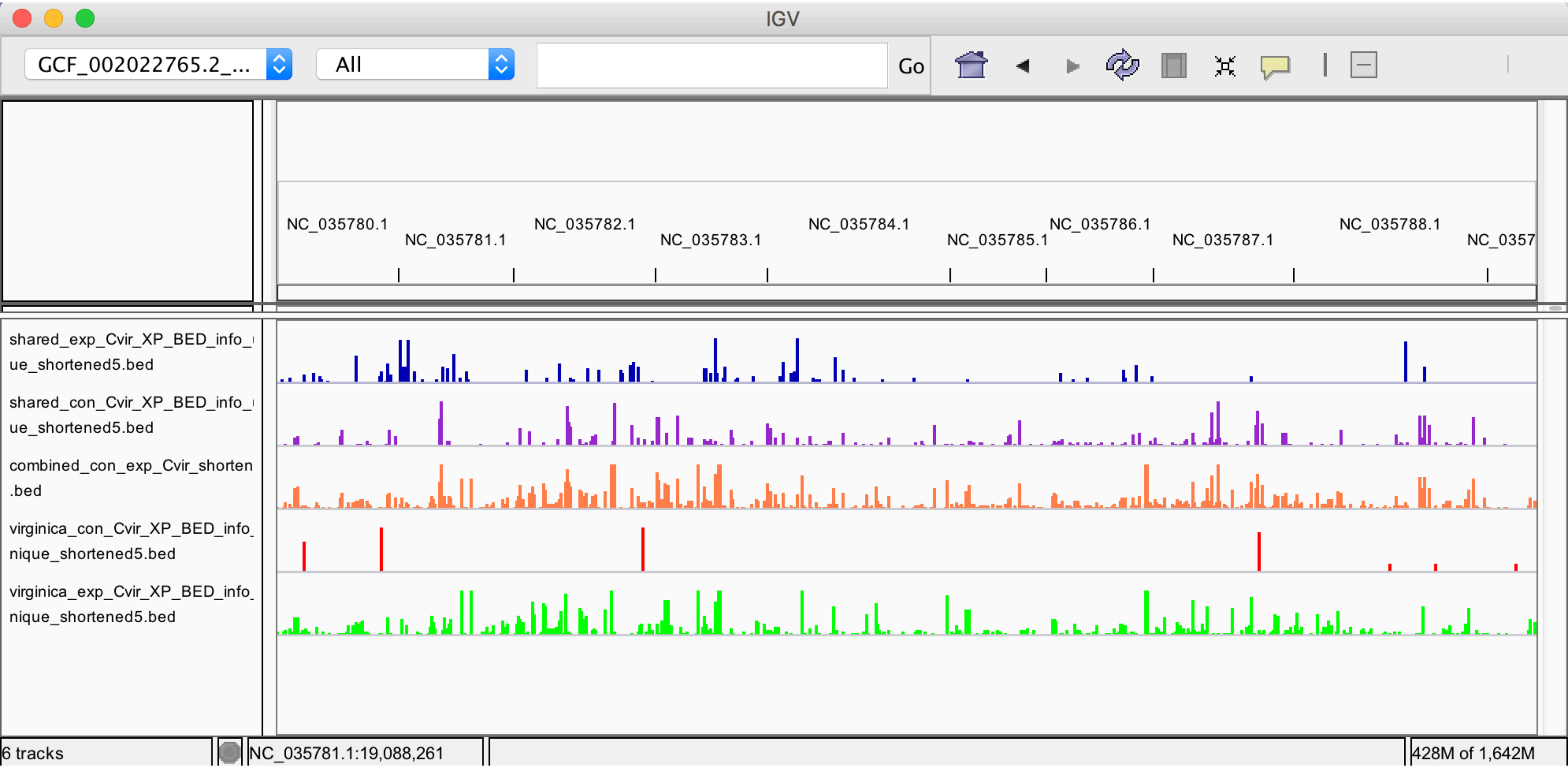
[1] https://software.broadinstitute.org/software/igv/BED

CAFÉ Analysis BED Files and Colors

5 tracks were constructed with *C. virginica* gene families:

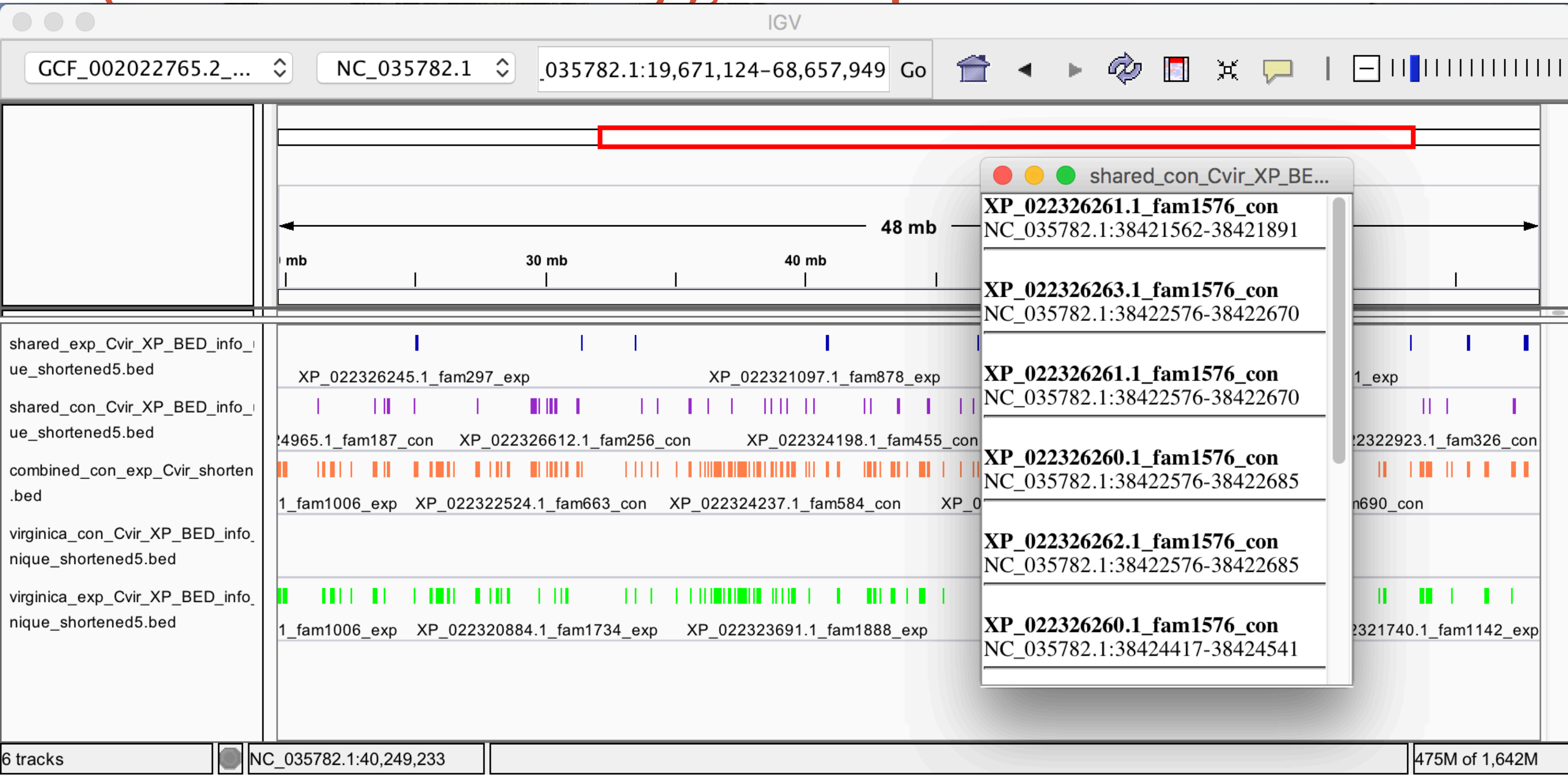
DATA	TRACK COLOR	FILENAME
All *C. virginica* expanded, contracted and shared gene families combined	**ORANGE**	combined_con_exp_Cvir_shortened5.bed
Just the *C. virginica* contracted	**RED**	virginica_con_Cvir_XP_BED_info_unique_shortened5.bed
Just the *C. virginica* expanded	**GREEN**	virginica_exp_Cvir_XP_BED_info_unique_shortened5.bed
C. virginica contracted sequences shared with other genomes	**PURPLE**	shared_con_Cvir_XP_BED_info_unique_shortened5.bed
C. virginica expanded sequences shared with other genomes	**LIGHT BLUE**	shared_exp_Cvir_XP_BED_info_unique_shortened5.bed

Open the Genome and then BED files in IGV



IGV
GCF_002022765.2_...
All
Go
NC_035780.1
NC_035781.1
NC_035782.1
NC_035783.1
NC_035784.1
NC_035785.1
NC_035786.1
NC_035787.1
NC_035788.1
NC_0357
shared_exp_Cvir_XP_BED_info_ue_shortened5.bed
shared_con_Cvir_XP_BED_info_ue_shortened5.bed
combined_con_exp_Cvir_shorten .bed
virginica_con_Cvir_XP_BED_info_nique_shortened5.bed
virginica_exp_Cvir_XP_BED_info_nique_shortened5.bed
6 tracks
NC_035781.1:19,088,261
428M of 1,642M

Click on a Gene to see Protein Name (with CAFÉ family), and position



GCF_002022765.2_... | NC_035782.1 | _035782.1:19,671,124-68,657,949 Go

shared_con_Cvir_XP_BE...

XP_022326261.1_fam1576_con
NC_035782.1:38421562-38421891

XP_022326263.1_fam1576_con
NC_035782.1:38422576-38422670

XP_022326261.1_fam1576_con
NC_035782.1:38422576-38422670

XP_022326260.1_fam1576_con
NC_035782.1:38422576-38422685

XP_022326262.1_fam1576_con
NC_035782.1:38422576-38422685

XP_022326260.1_fam1576_con
NC_035782.1:38424417-38424541

6 tracks | NC_035782.1:40,249,233 | 475M of 1,642M

IGV
48 mb
mb
30 mb
40 mb
shared_exp_Cvir_XP_BED_info_ue_shortened5.bed
shared_con_Cvir_XP_BED_info_ue_shortened5.bed
combined_con_exp_Cvir_shorten.bed
virginica_con_Cvir_XP_BED_info_nique_shortened5.bed
virginica_exp_Cvir_XP_BED_info_nique_shortened5.bed
XP_022326245.1_fam297_exp
XP_022321097.1_fam878_exp
1_exp
24965.1_fam187_con
XP_022326612.1_fam256_con
XP_022324198.1_fam455_con
2322923.1_fam326_con
1_fam1006_exp
XP_022322524.1_fam663_con
XP_022324237.1_fam584_con
XP_0
n690_con
1_fam1006_exp
XP_022320884.1_fam1734_exp
XP_022323691.1_fam1888_exp
2321740.1_fam1142_exp

Questions

- Email Erin Roberts at erin_roberts@my.uri.edu with any questions
- See erin's github repository for the BED file tracks https://github.com/erinroberts/EOGC-CAFE-Gene-Family-analysis/tree/master/IGV_TRACKS