Integrative Genomics Viewer

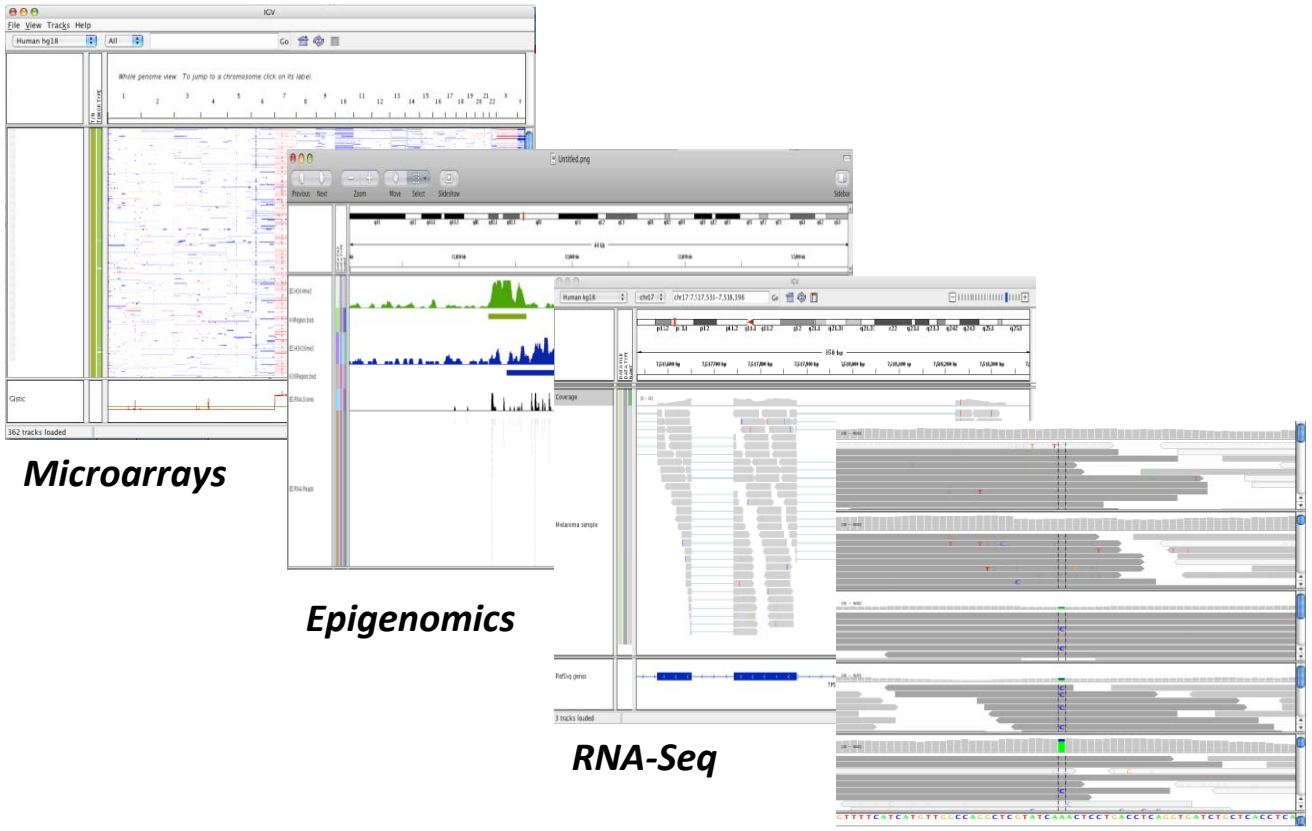
Prat Thiru

Outline

- Overview
- User Interface Basics
- Browsing the Data
- Data Formats
- IGV Tools
- Demo

Why IGV?

- IGV is an integrated visualization tool of large data types



Microarrays

Epigenomics

RNA-Seq

NGS alignments

Why IGV?

- Integrate different data types simultaneously
- View large datasets easily
- Faster navigation or browsing
- Run it locally on your desktop
- Easy to use interface

Large-Scale Projects using IGV

- The Cancer Genome Atlas (TCGA)
- Multiple Myeloma Research Consortium
- 1000 Genomes Project

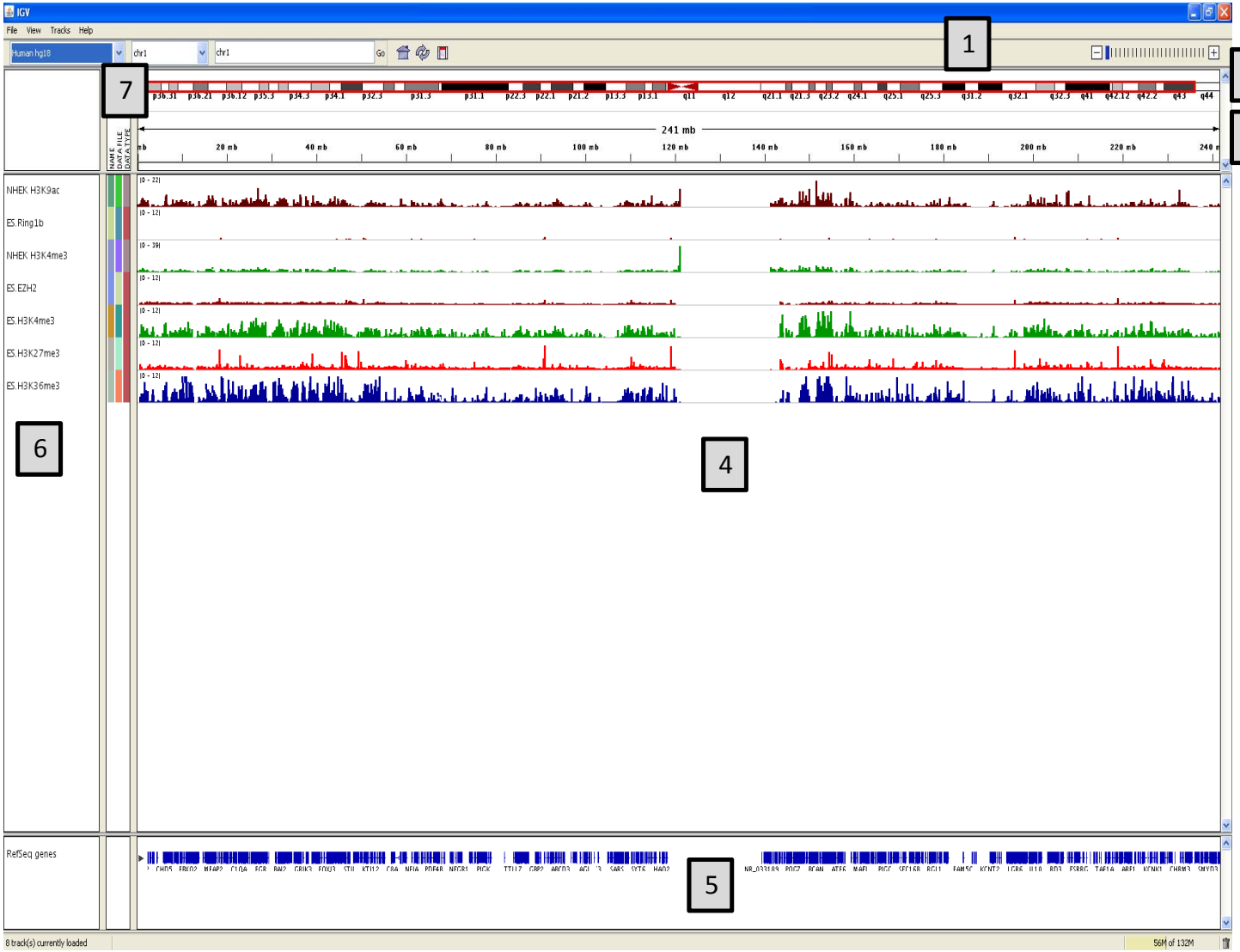
Getting IGV

http://www.broadinstitute.org/igv

History and Usage

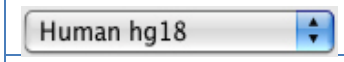
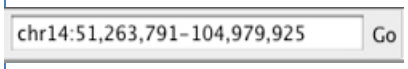
- First release was in Aug 2008
- Current version: 1.5
- More than 5500 registered users (registration is free)
- Open source and freely available

IGV Interface

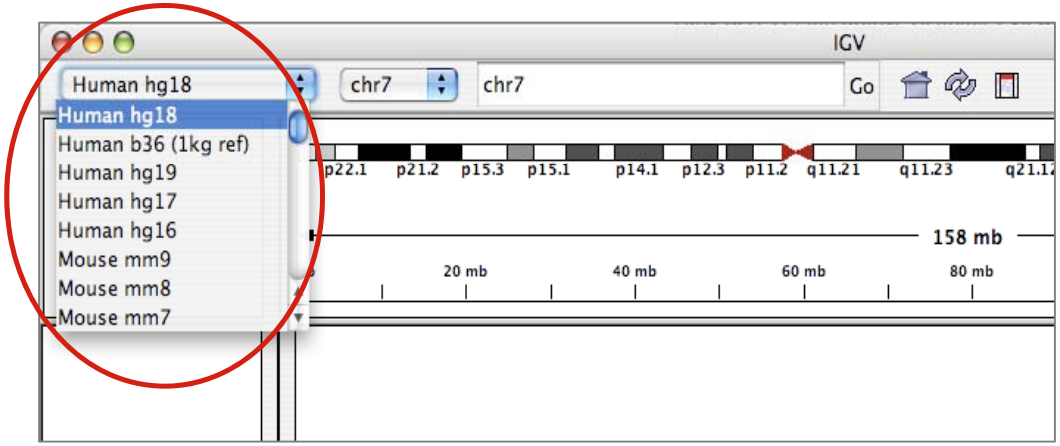


1. Tool Bar
2. Chromosome Ideogram
3. Ruler
4. Track Data
5. Features
6. Track Names
7. Attributes

Tool Bar

Genome drop-down box 	Loads a genome.
Chromosome drop-down box 	Zooms to a chromosome.
Search box 	Displays the chromosome location being shown. To scroll to a different location, enter the gene name, locus, or track name and click Go.
Whole genome view 	Zooms to whole genome view.
Refresh 	Refreshes the display.
Define a region 	Defines a region of interest on the chromosome.
Zoom slider 	Zooms in and out on a chromosome. Sometimes referred to as the "railroad track."

Available Genomes



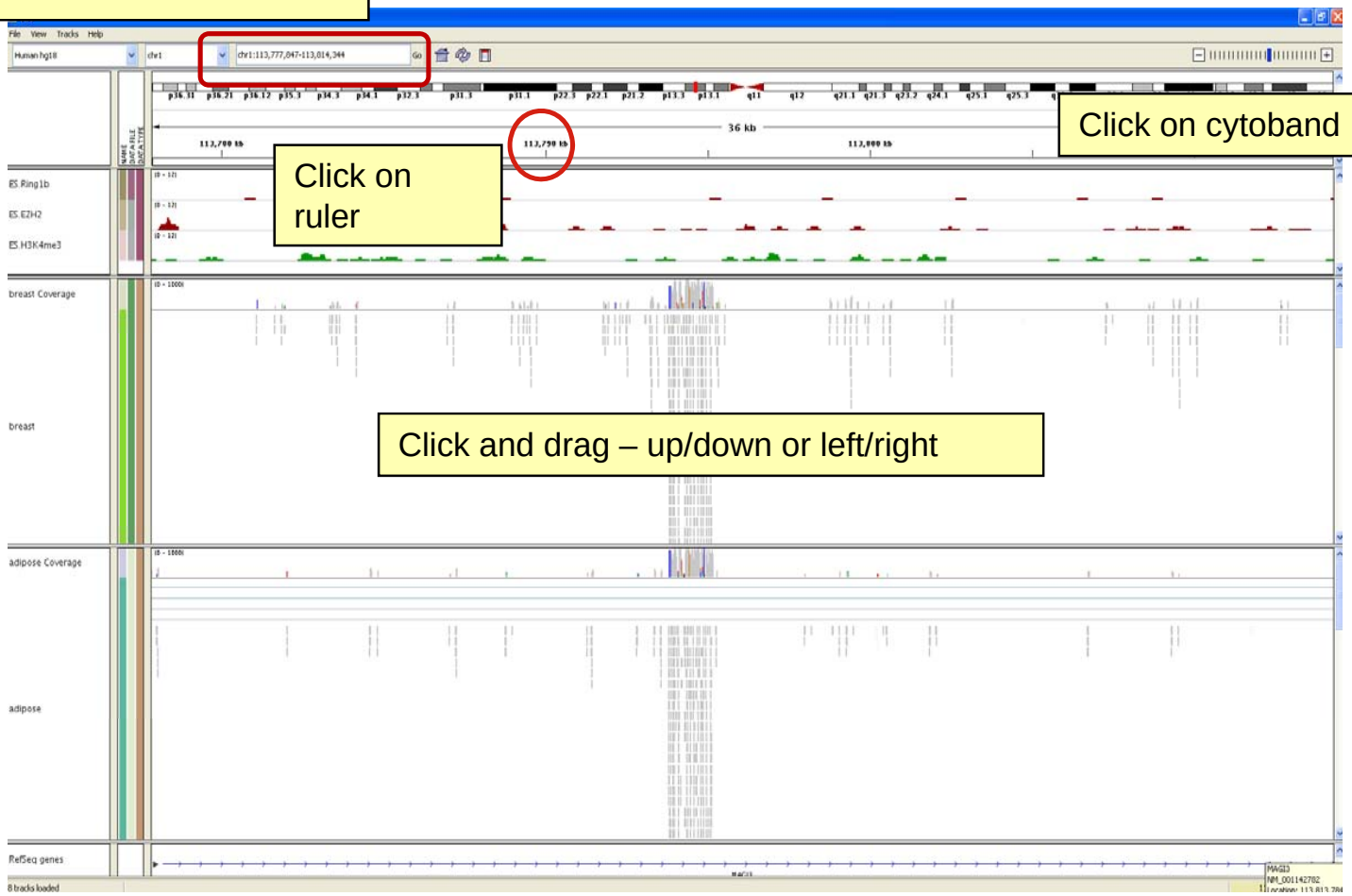
- Human, Mouse, S.cerevisiae, C.elegans, D.melanogaster, and many others…
- Import your favorite genome, if sequence is available

Loading the Data

- General Characteristics
  - Any data related to genome coordinates
  - Sample annotation/attributes
  - Genome annotations
- IGV supports multiple file formats

Browsing the Data

Specify range or term in the search box



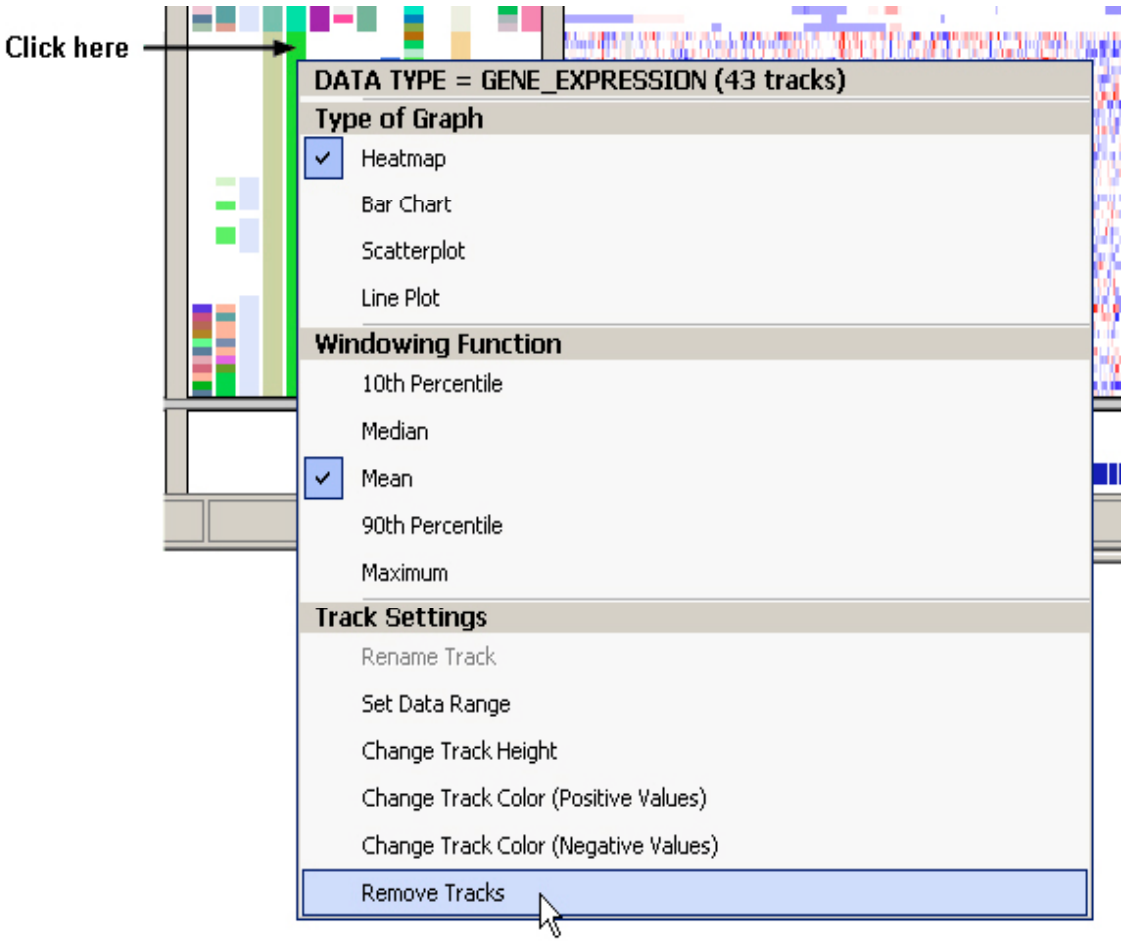
Use keyboard
(1) arrow keys
(2) Page Up, Page Down, Home, End

Tracks

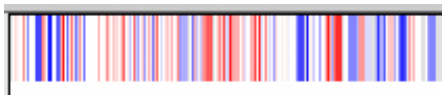
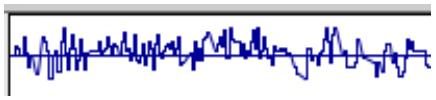
- Two generic types: data and annotation
- Defined by file format
- Set track display
  - Fit Data to Window, and other options

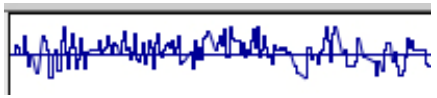
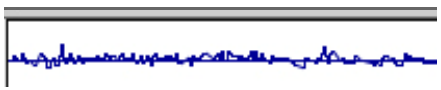
Tracks

- Right click on track

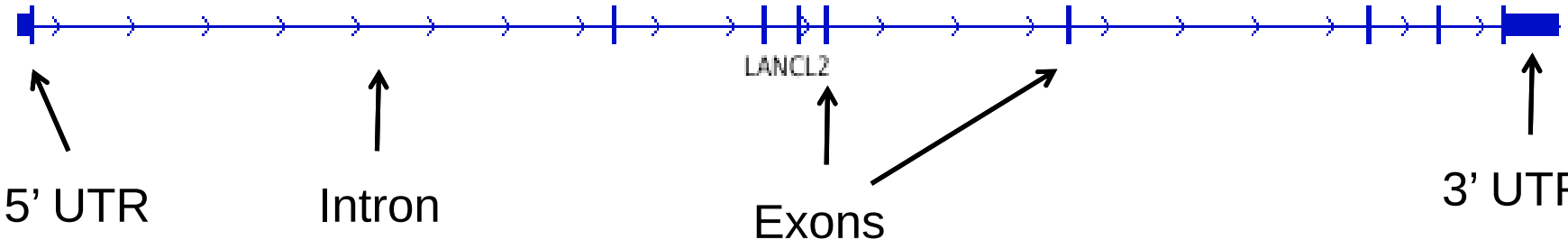


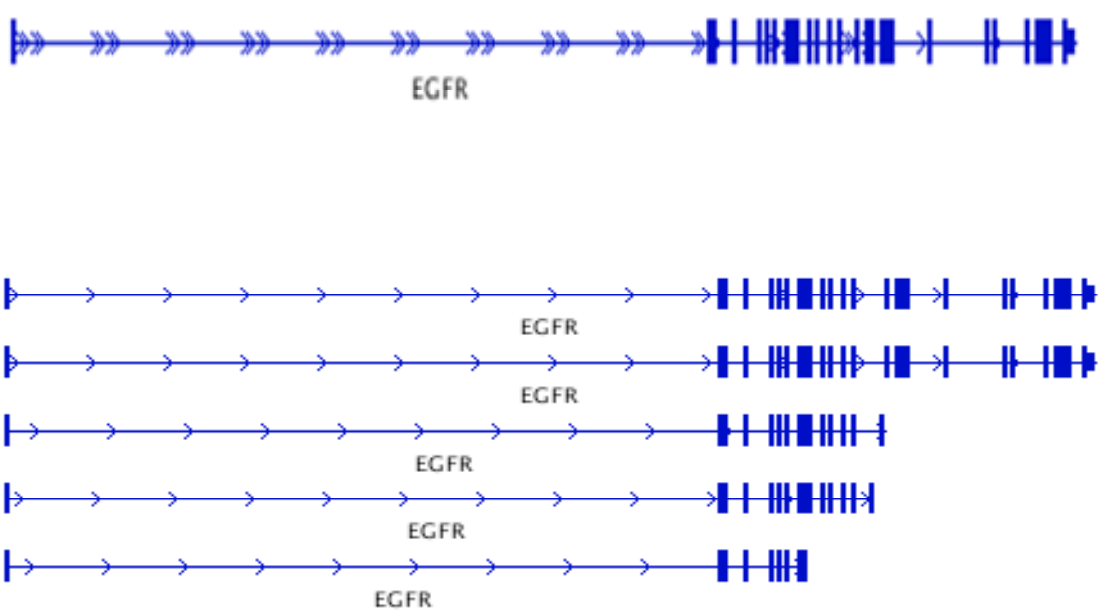
Tracks: Graph Type and Data Range

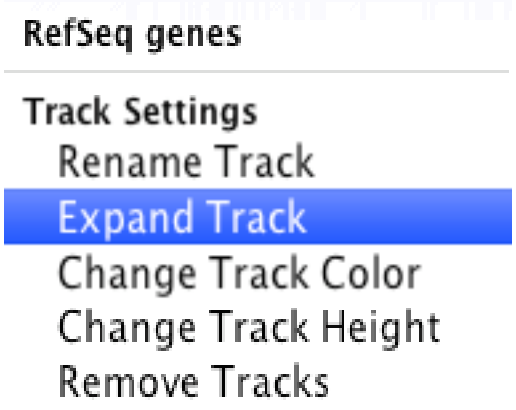
Heatmap	
Bar chart	
Scatter plot	
Line plot	

Min, Baseline, Max	**Result**
0,0,3	
-1.5,0,1.5	
-5,0,5	

Tracks: Expanding

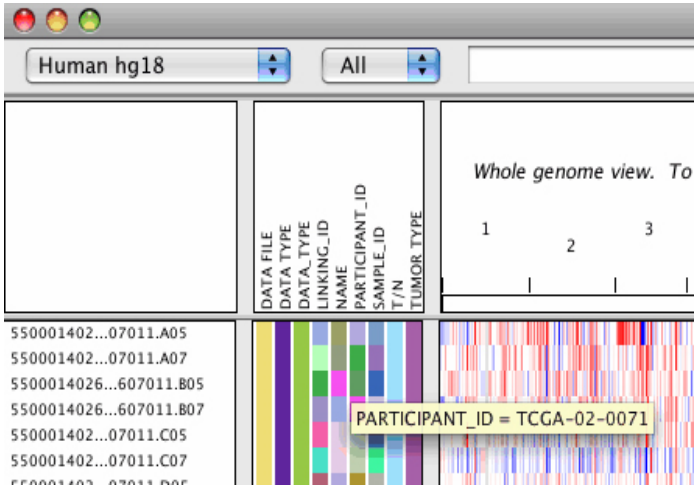






Attributes

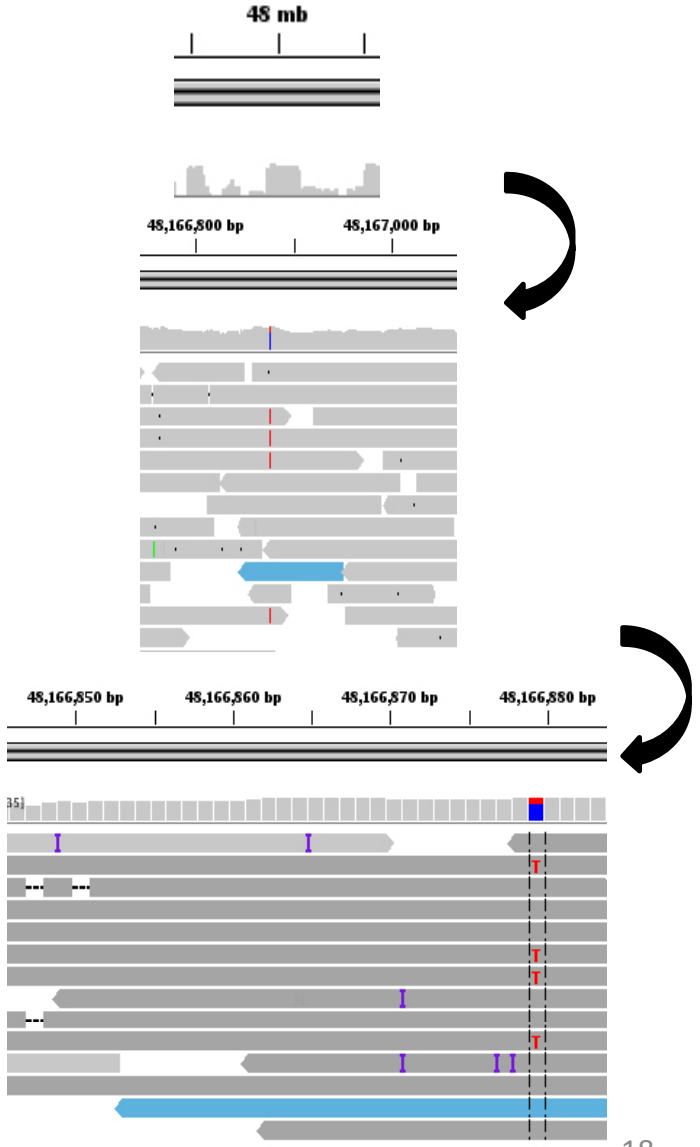
- Associated with tracks
- Used for filtering, sorting, and grouping data (Tracks → (Sort/Group/Filter…)
- File format example:



TRACK_ID	Data_Type	PARTICIPANT_ID	SAMPLE_ID	GENDER	T/N	Tumor_type		Treated Primary/Secondary		Hypermutated
EX-01-001	Expression	P-01-P001	P-01-S001	M	Tumor	GBM	Y	Primary Y		
CN-01-002	CopyNumber	P-01-P001	P-01-S001	M	Tumor	GBM	Y	Primary Y		
MU-01-003	Mutation	P-01-P001	P-01-S002	M	Tumor	GBM	Y	Primary Y		
EX-01-004	Expression	P-01-P002	P-01-S003	M	Normal	GBM	Y	Secondary	Y	
CN-01-005	CopyNumber	P-01-P002	P-01-S004	M	Tumor	GBM	Y	Secondary	N	
EX-01-006	Expression	P-01-P002	P-01-S004	M	Tumor	GBM	Y	Secondary	N	
ME-01-007	Methylation	P-01-P002	P-01-S004	M	Tumor	GBM	Y	Secondary	N	
EX-01-008	Expression	P-01-P003	P-01-S006	F	Tumor	GBM	N	Primary Y		
EX-01-009	Expression	P-01-P004	P-01-S009	F	Tumor	GBM	N	Primary Y		
EX-01-0010	Expression	P-01-P005	P-01-S0011	M	Control					

Viewing NGS Data

- BAM format recommended
- At low resolution only coverage is shown
- At higher resolution, reads are shown including where bases differ



File Formats

File Format	Data Type
GCT	Gene expression
RES	Gene expression
SAM, BAM	Sequence alignments
BED, GFF, GFF3	Genome annotations
IGV, WIG, HDF5 file not created with alignment processor	Other
Cytoband, FASTA	Not applicable. Cytoband and sequence files for an imported genome.

File Formats

- SAM/BAM
  - http://samtools.sourceforge.net
  - BAM files need to be indexed (use samtools). SAM files needs to be sorted by start position and indexed (use igvtools)
  - Example of SAM file:

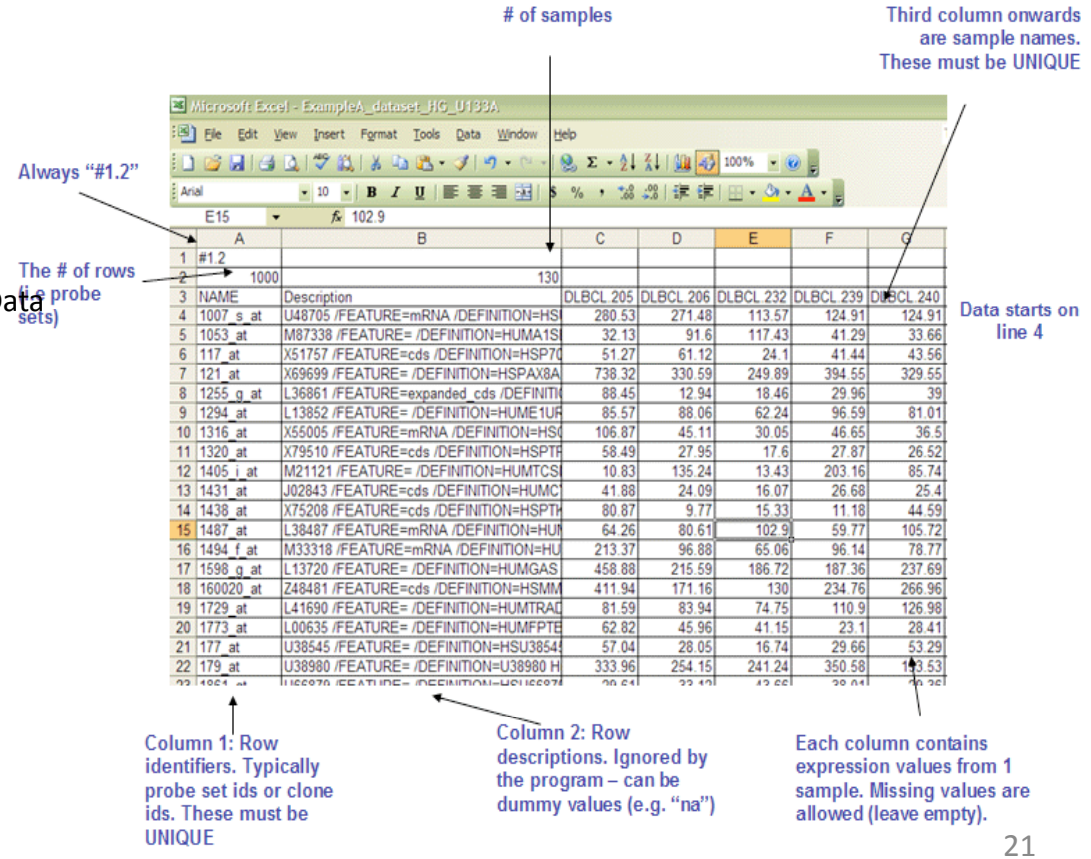
```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

- BED/GFF/WIG
  - http://genome.ucsc.edu/FAQ/FAQformat.html

File Formats: Expression Data

- GCT
  - http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats
  - http://www.broadinstitute.org/igv/GCT
  - If probe ids are used, specify where they map to in the second column, or IGV will try to map based on known ids (Affy, Agilent or Illumina)

http://www.broadinstitute.org/igv/ExpressionData



	A	B	C	D	E	F	G
1	#1.2						
2	1000	130					
3	NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
4	1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HS	280.53	271.48	113.57	124.91	124.91
5	1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
6	117_at	X51757 /FEATURE=cds /DEFINITION=HSP7	51.27	61.12	24.1	41.44	43.56
7	121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
8	1255_g_at	L36861 /FEATURE=expanded_cds /DEFINITI	88.45	12.94	18.46	29.96	39
9	1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
10	1316_at	X55005 /FEATURE=mRNA /DEFINITION=HS	106.87	45.11	30.05	46.65	36.5
11	1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTF	58.49	27.95	17.6	27.87	26.52
12	1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
13	1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
14	1438_at	X75208 /FEATURE=cds /DEFINITION=HSPTH	80.87	9.77	15.33	11.18	44.59
15	1487_at	L38487 /FEATURE=mRNA /DEFINITION=HU	64.26	80.61	102.9	59.77	105.72
16	1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
17	1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
18	160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
19	1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
20	1773_at	L00635 /FEATURE= /DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
21	177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22	179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	143.53

Choosing a File Format

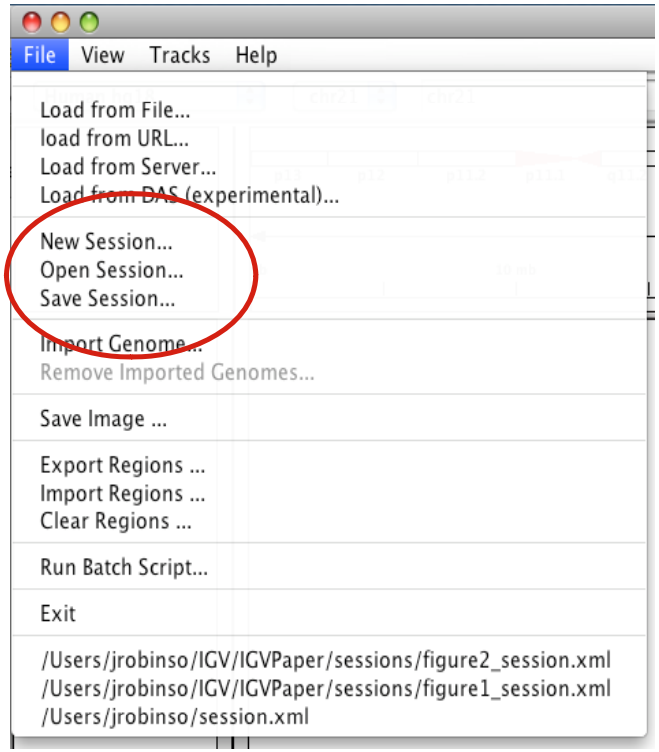
Source Data	Recommended File Formats
ChIP-Seq, RNA-Seq	TDF format. Use the igvtools package (count command) to generate a binary read count density file in TDF format. Load the resulting TDF file into IGV.
Gene expression data	GCT format, RES format
Genome annotations	GFF or GFF3 format, BED format
Sequence alignment data	SAM format (must be sorted/indexed), BAM format (must be indexed)
Any numeric data	IGV format, TAB format, WIG format

Default Display Options by Data Types

Data Type	Default Graph Type	Default Data Range	Default Colors
Copy number	Heatmap	-1.5 to 1.5	Blue to red
Gene expression	Heatmap	-1.5 to 1.5	Blue to red
Chip	Bar chart	None, data is autoscaled	Blue
DNA methylation	Heatmap	0 to 1 (methylation score)	Green
Other	Bar chart	None, data is autoscaled	Blue

Sessions

- Your current session can be saved
  - Restore
  - Share

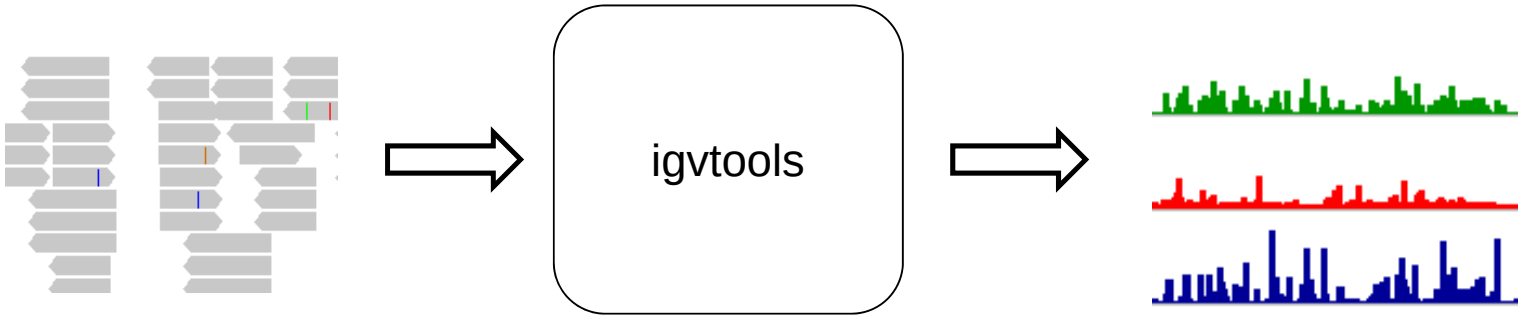


IGV Tools

- **tile**
  Converts a sorted data input file to a binary tiled data (.tdf) file.
  Supported input file formats: .wig, .cn, .snp, .igv, and .gct
- **count**
  Computes average alignment or feature density for over a specified window size across the genome.
  Supported input file formats: .sam, .bam, .aligned, .psl, .pslx, and .bed
- **index**
  Creates an index file for an ASCII alignment or feature file.
  Index files are required for loading alignment files into IGV, and can significantly improve performance for large feature files.
  Supported input file formats: .sam, .aligned, .vcf, .psl, and .bed
- **sort**
  Sorts the input file by start position.
  Supported input file formats: .cn, .igv, .sam, .aligned, .psl, .bed, and .vcf

IGV Tools: Count

- The count command is used to transform alignment files to read density TDF files.



Alignments
Alignments in bam/sam, .aligned, or bed format.

Read density
Tiled Data File or .tdf format

IGV Tools

- Interface
  - File → Run igvtools…
- Command line
  - Download scripts/utilities from the Downloads page
  - Run the command

    (eg. igvtools sort [options] [inputFile] [outputFile])

More Information…

- IGV Help
  - http://www.broadinstitute.org/igv/UserGuide
  - From IGV Menu Bar: Help → Help or Tutorial
- Neilsen, C.B., et al. *Visualizing Genomes: techniques and challenges* Nature Methods 7:S5-S15 (2010)