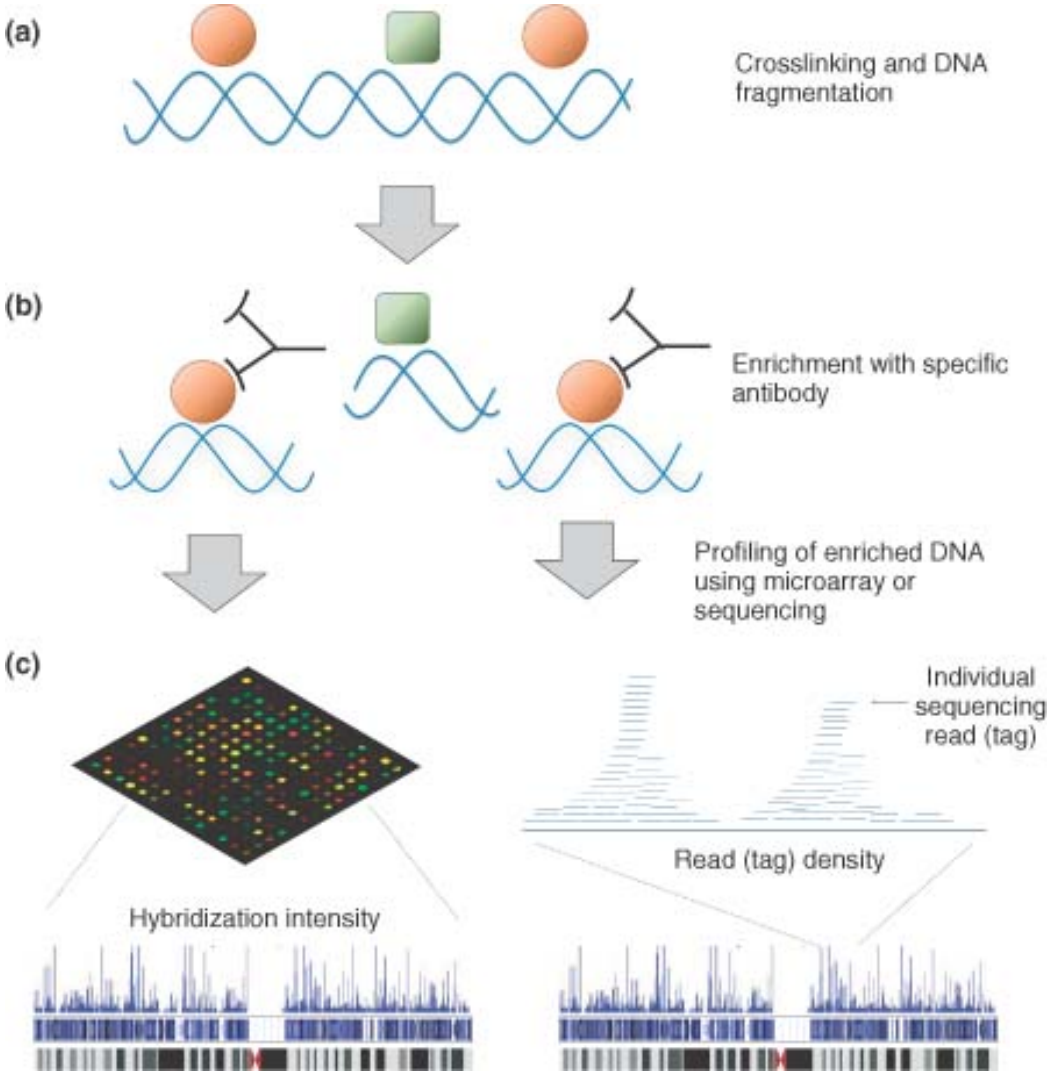
Use Case 14: ChIP-Seq Analysis Using MACS via the Genboree Workbench

Epigenome Informatics Workshop
Bioinformatics Research Laboratory

Peak Calling

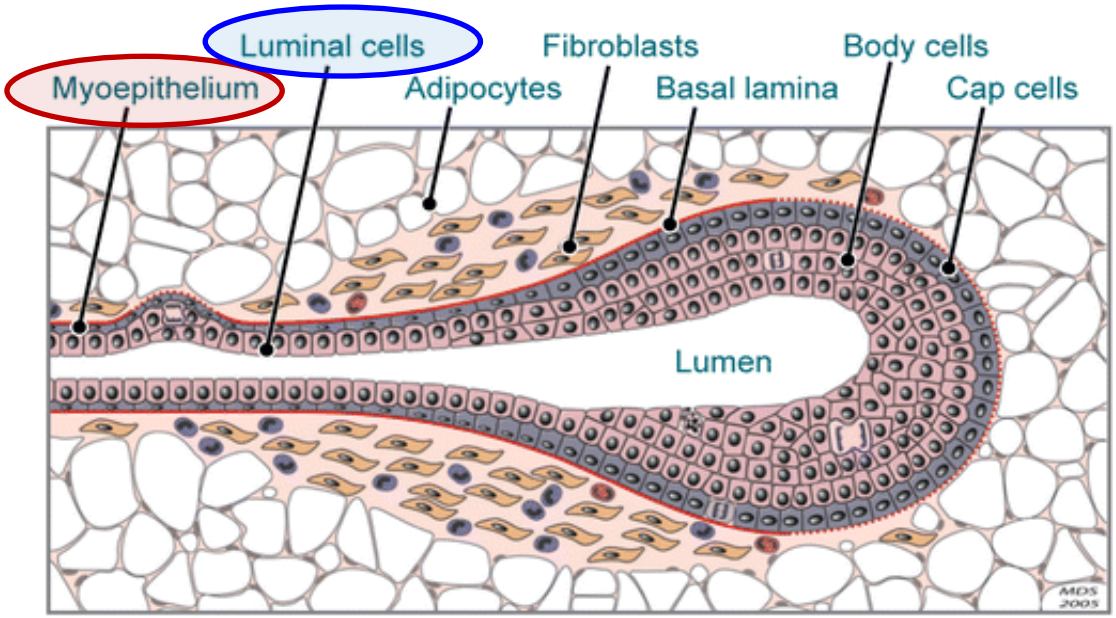
- Transcription Factors
- Histone modifications
  - H3K4me3, H3K27me3, H3K36me3, H3K9me3, H3K4me1, H3K27ac
- DNAse hypersensitivity
- Digital Genomic Footprinting
- MeDIP-Seq, MRE-Seq
- Many others

Chromatin **I**mmunoprecI**P**itation **Seq**uencing – ChIP-Seq



Breast Tissue Data from Epigenome Atlas

- Breast Luminal Epithelial Cell (BLEC): H3K4me1
  - ❑ H3K4me1 → mono-methylation of lysine 4 of H3 histone protein; associated with enhancers and DNA regions downstream of transcription starts
  - ❑ BRCA1 basal-like breast cancers originate from luminal epithelial progenitors
- Breast Myoepithelial Cells (BMC): H3K4me3
  - ❑ H3K4me3 → tri-methylation of lysine 4 of H3 histone protein; associated with promoters which are active or ready to be activated
  - ❑ Myoepithelial cells stabilize normal structure and help limit cancer growth



Hebner C, et al. 2008.
Annu. Rev. Pathol. Mech. Dis. 3:313–39

MACS

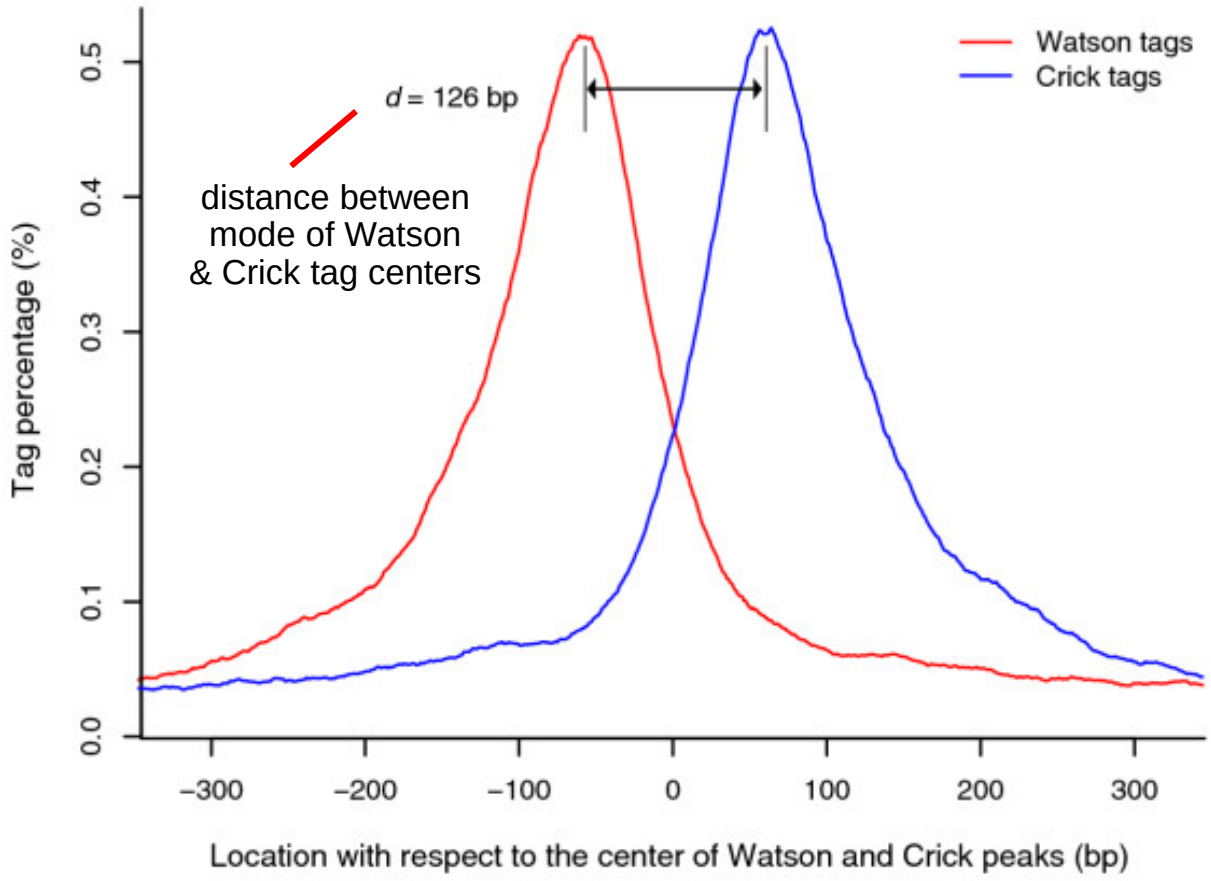
- Model Based Analysis of ChIP-Seq
- Call peaks on both strands
  - Empirical modeling of bandwidth (sonication size)
- Infer local distribution parameters
- Chip-Seq fragments equally likely to be sequenced from both ends (bimodal distribution around binding site)
- Account for local biases in sequencing
- Options
  - Run w/ control data
  - Run w/o control data

Zhang et al, Model-based Analysis of Chip-Seq (MACS) *Genome Biology* 2008, 9:R137

Feng et al, Using MACS to Identify Peaks from Chip-Seq Data, *Current Protocols in Bioinformatics* 2011, 34:2.14.1-2.14.14; John Wiley & Sons, Inc.

MACS

FoxA1 Chip-Seq (3.9 M uniquely mapped tags), Zheng et al



MACS in the Genboree Workbench

- Operates on BED input files
- 6 field format

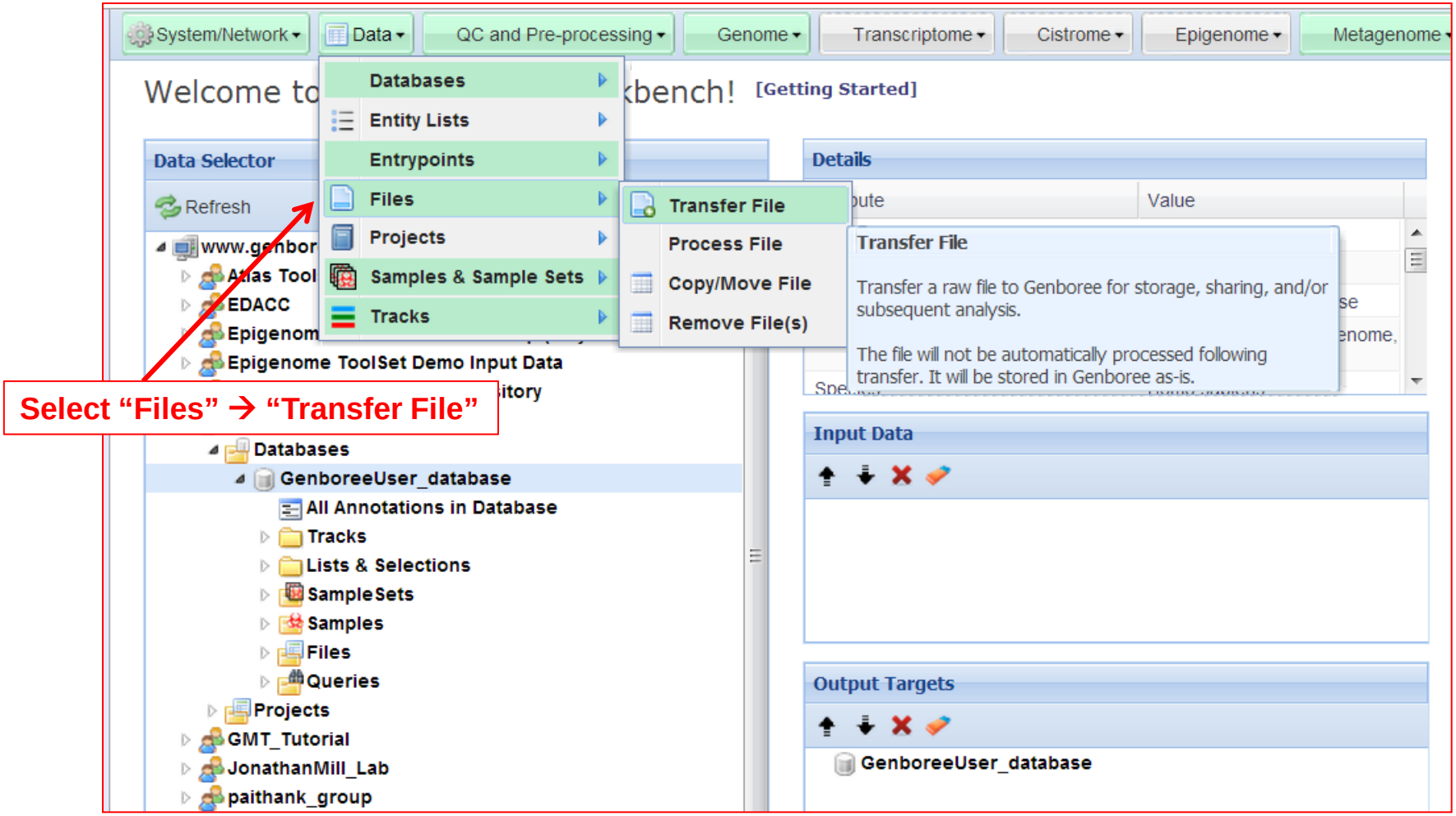
  Chr start stop name score strand

Chromosome	Start	Stop	Read Name	Score	Strand

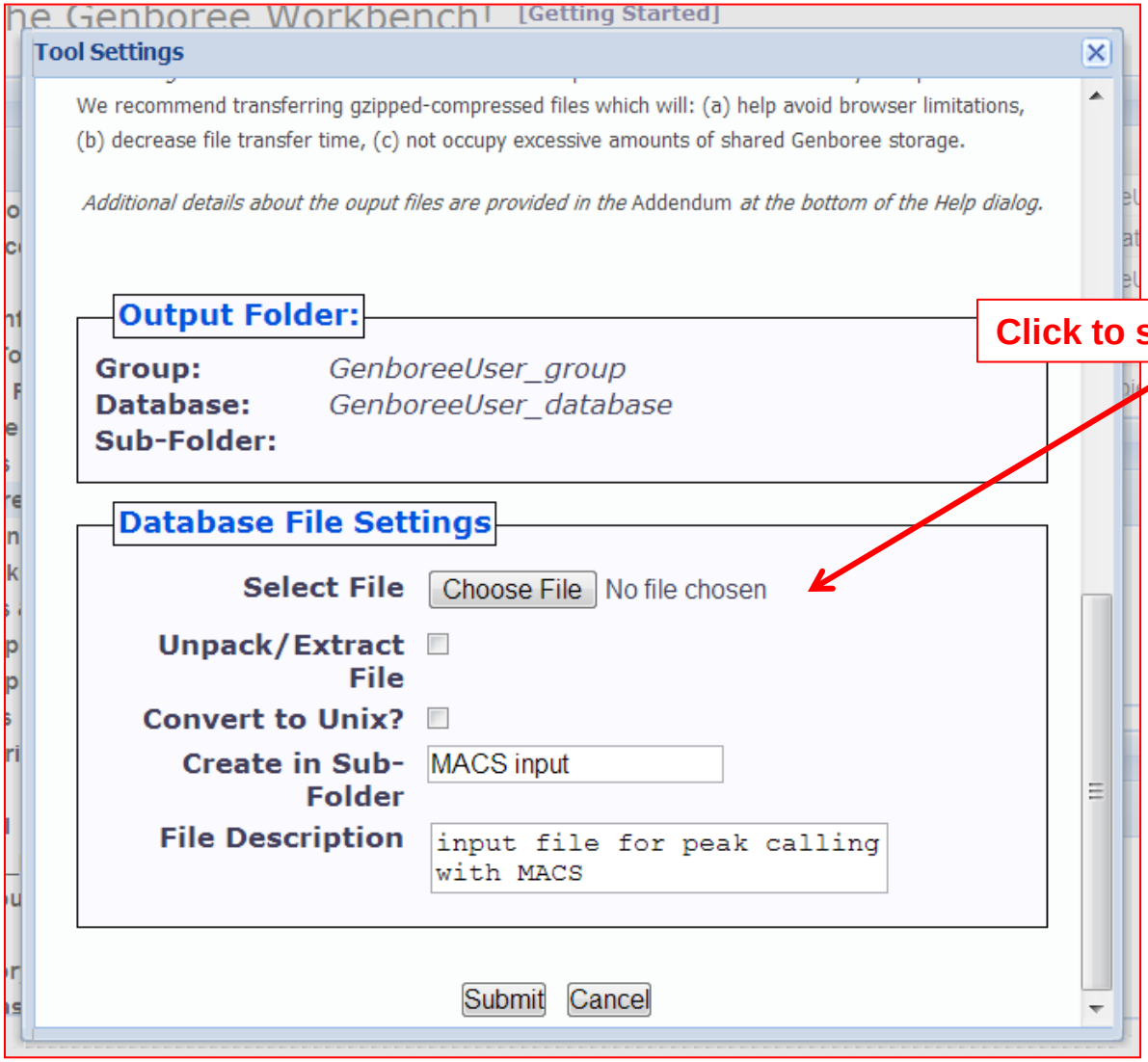
- Upload BED files using the Workbench UI
  - Data/Files/Transfer File

Upload Your BED File

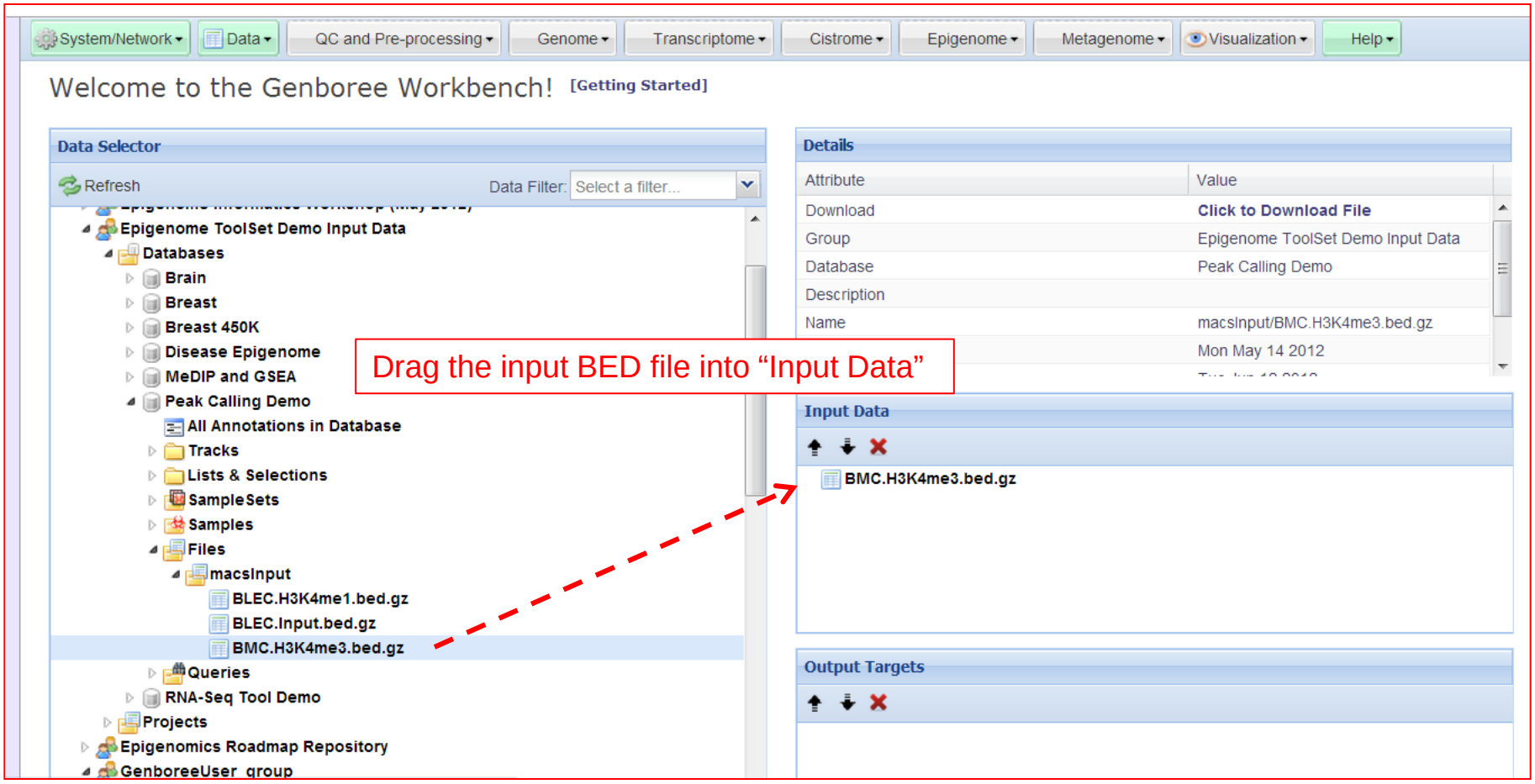
Please note: we have provided the BED file for you, this slide and the next show you how to upload your own BED files for future reference



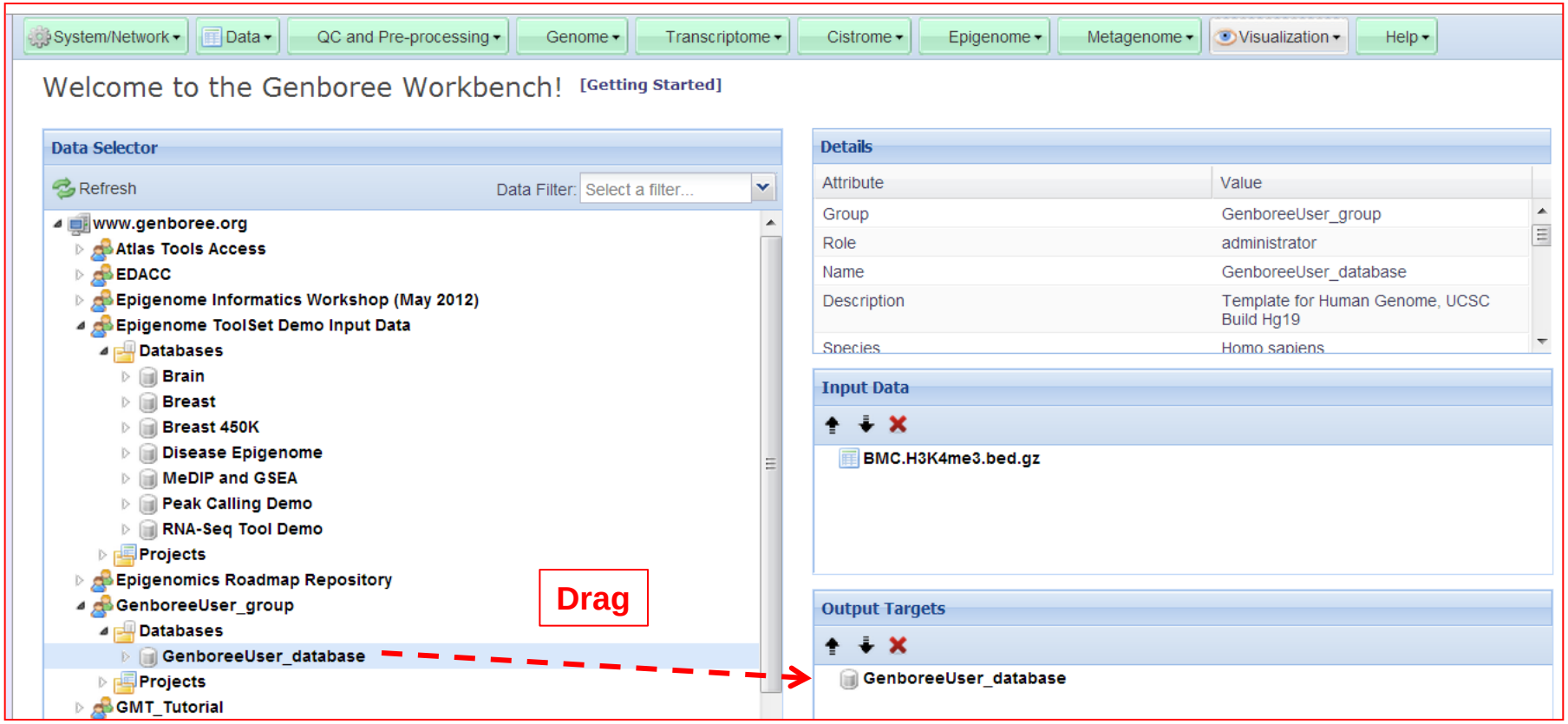
Select Your BED File



Prepare to Run MACS in the Genboree Workbench (populate Input Data)

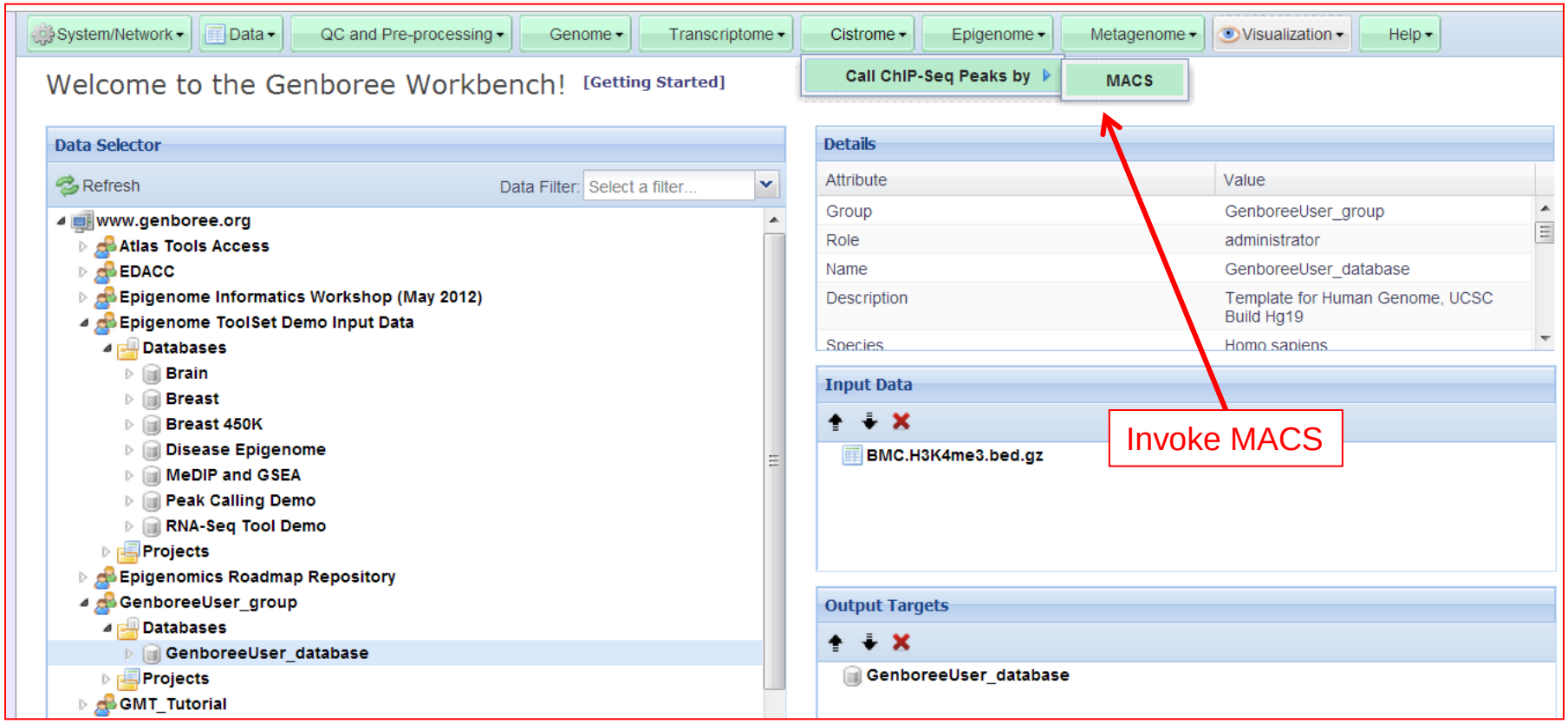


Prepare to Run MACS in the Genboree Workbench (populate Output Targets)

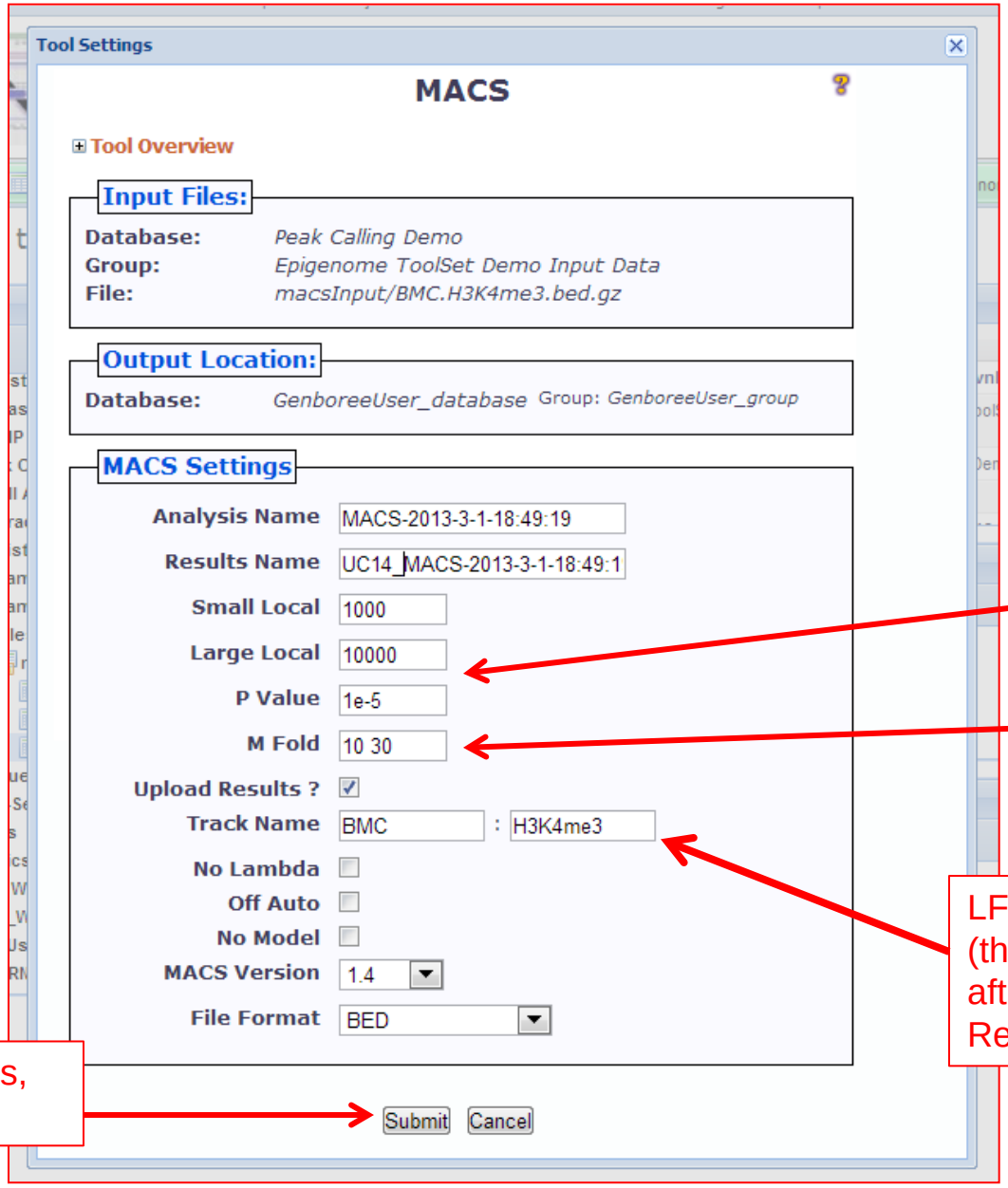


Designate the database where you want the MACS results to be deposited by dragging the it into “Output Targets”

Invoke MACS in the Genboree Workbench



Select Settings in the MACS Dialogue Window



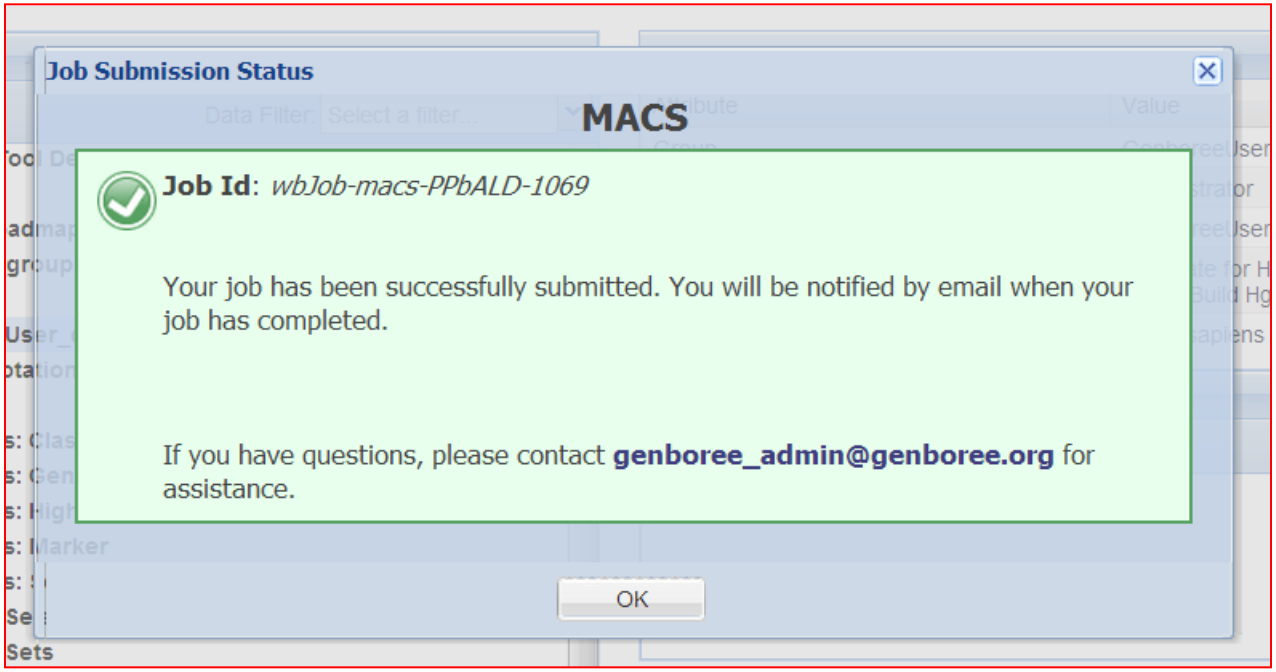
p-value cutoff to call peaks

min, max fold change

LFF track name (these boxes will appear after checking "Upload Results")

After selecting settings, click Submit.

MACS Job Submission Status



Genboree email Notifying You MACS Job is Complete

Hello Genboree User,

Your job completed successfully.

Job Summary:
JobID - wbJob-macs-PPbALD-1069

Additional Info:
Database: 'GenboreeUser_database'
Group: 'GenboreeUser_group'

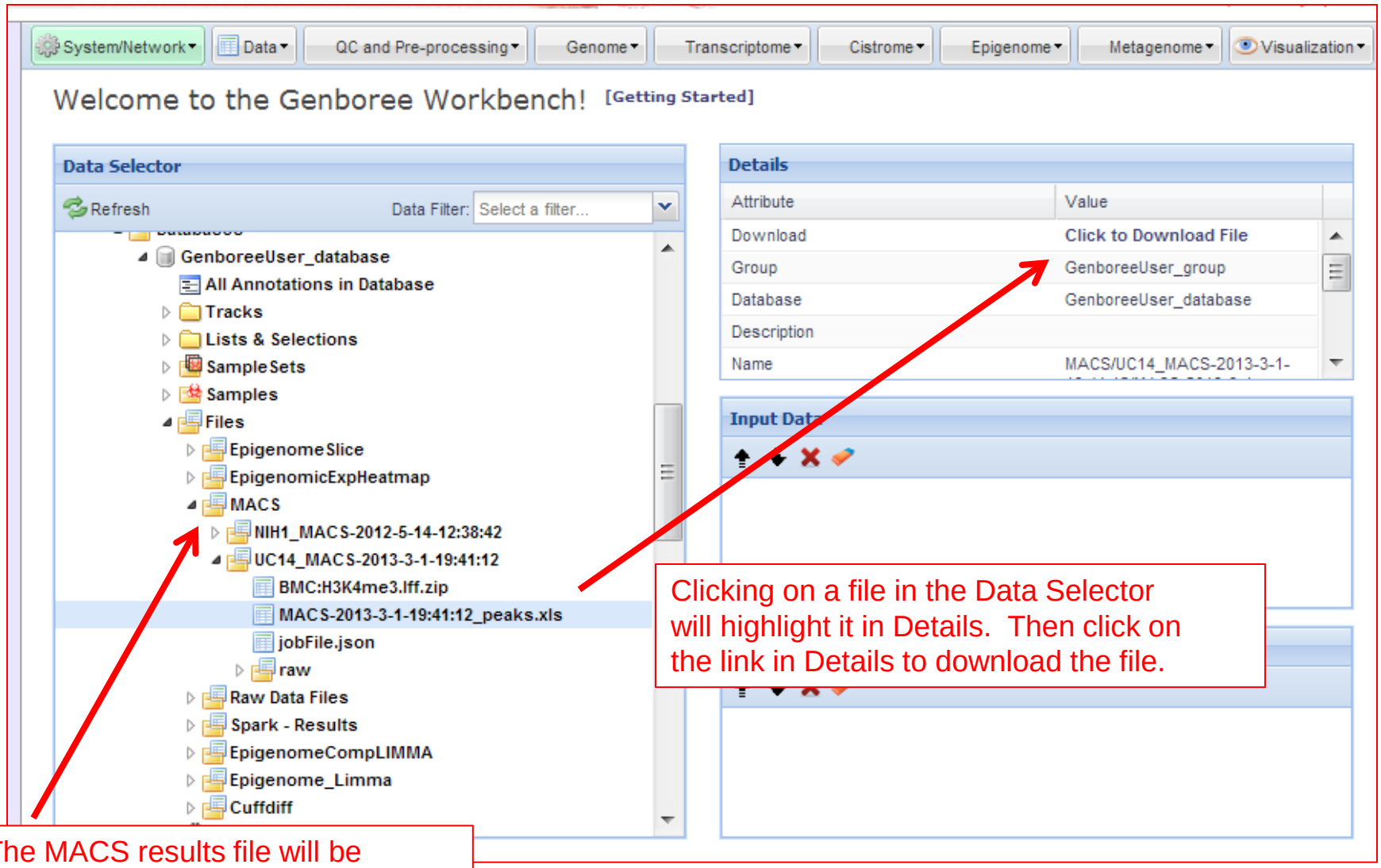
You can download result files from the 'MACS-2013-2-18-14:53:22' folder under the 'MACS' directory.

MACS called 23943 peaks. The breakdown of the peaks with respect to the gene model follows:

Gene TSS +- 3K Peaks	13693	57.18%
Gene Exon Peaks	868	3.62%
Gene Intron Peaks	2707	11.3%
Intergenic Peaks	6675	27.87%

- The Genboree Team

Retrieve MACS Results from “Files” in Your Designated Database

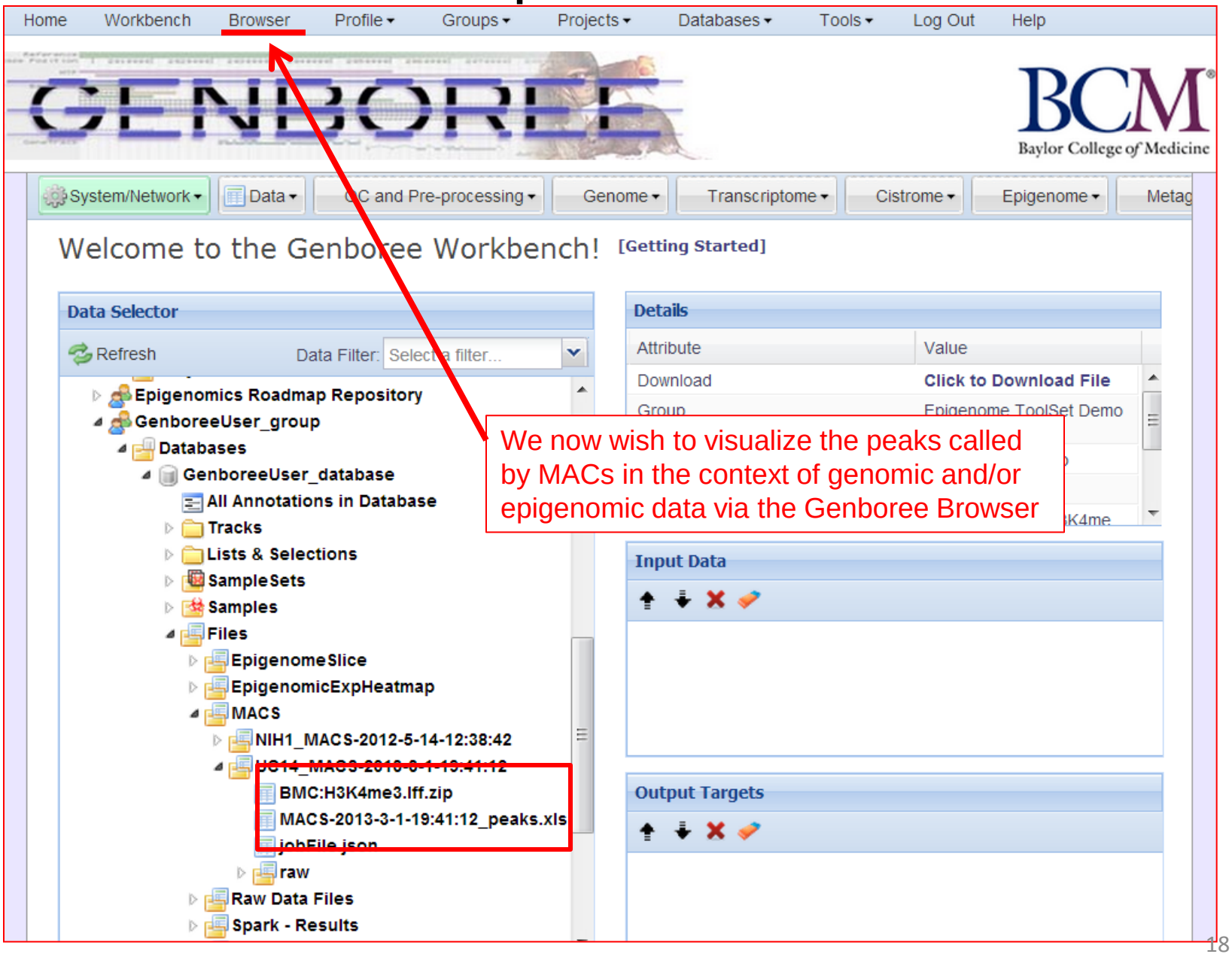


MACS Results Downloaded in Excel

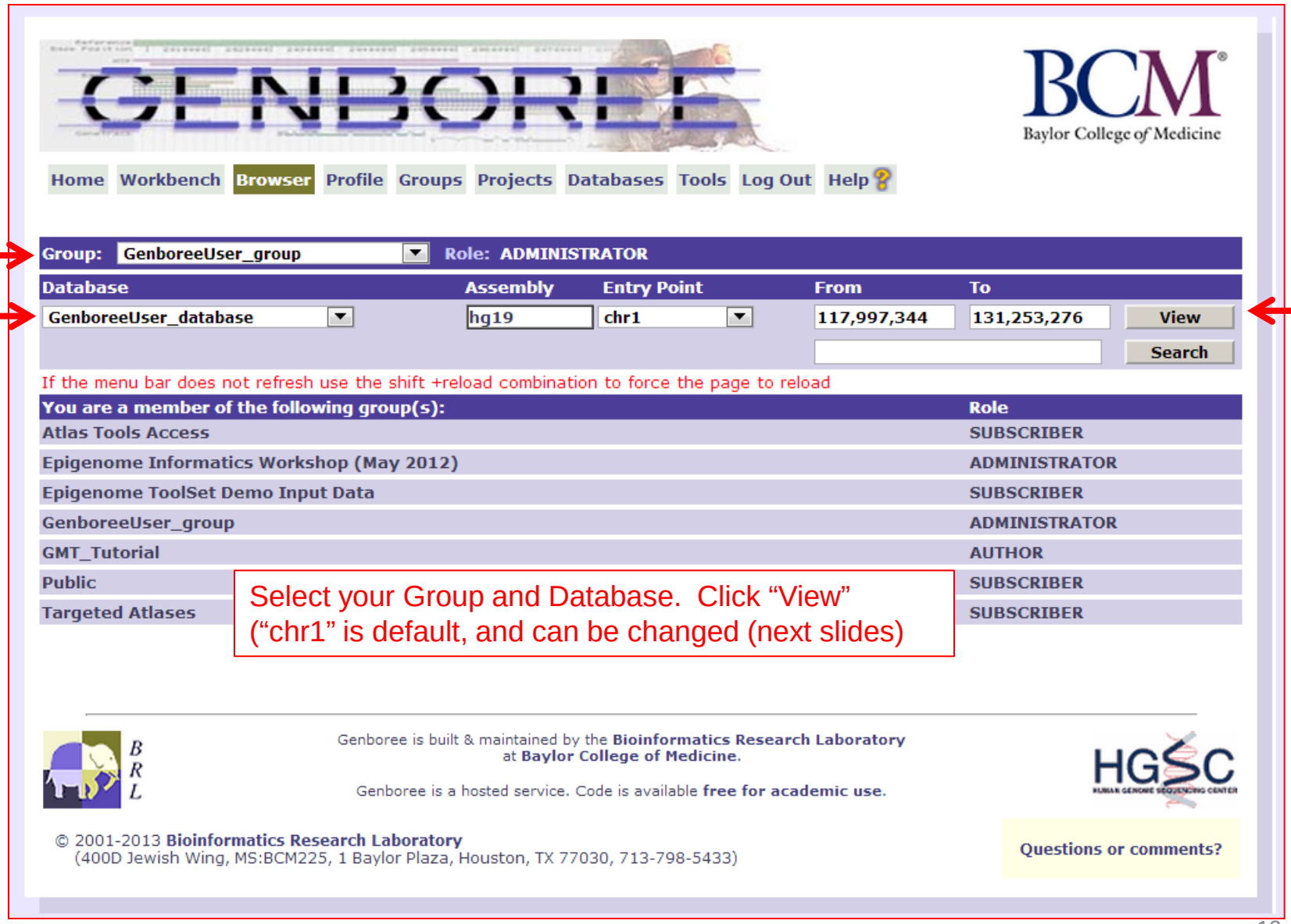
	A	B	C	D	E	F	G	H	I
1	# This file is generated by MACS version 1.4.1 20110627								
2	# ARGUMENTS LIST:								
3	# name = BMC_H3K4Me3_MACS-2013-2-18-14%3A53%3A22								
4	# format = AUTO								
5	# ChIP-seq file = ./9211e5b36b69c5ae404791db02a1d981af29e13e/file_0_macsInput%2FBMC.H3K4me3.bed								
6	# control file = None								
7	# effective genome size = 2.70e+09								
8	# band width = 300								
9	# model fold = 10,30								
10	# pvalue cutoff = 1.00e-05								
11	# Small dataset will be scaled towards larger dataset.								
12	# Range for calculating regional lambda is: 10000 bps								
13									
14	# tag size is determined as 49 bps								
15	# total tags in treatment: 8184971								
16	# tags after filtering in treatment: 8184971								
17	# maximum duplicate tags at the same position in treatment = 1								
18	# Redundant rate in treatment: 0.00								
19	# d = 200								
20	chr	start	end	length	summit	tags	-10*log10	fold_enrichment	
21	chr1	9861	10677	817	360	292	2034.63	24.91	
22	chr1	713307	715489	2183	1297	83	250.89	11.7	
23	chr1	724760	727160	2401	2202	92	223.86	16.37	
24	chr1	761468	763266	1799	865	133	463.34	14.42	
25	chr1	833096	834002	907	577	20	96.47	11.54	
26	chr1	839750	840203	454	262	15	77.67	11.4	
27	chr1	859636	861418	1783	741	37	83.16	5.71	
28	chr1	875788	876527	740	305	13	50.73	7.95	

Summit = location with highest fragment pileup, predicted as precise binding location

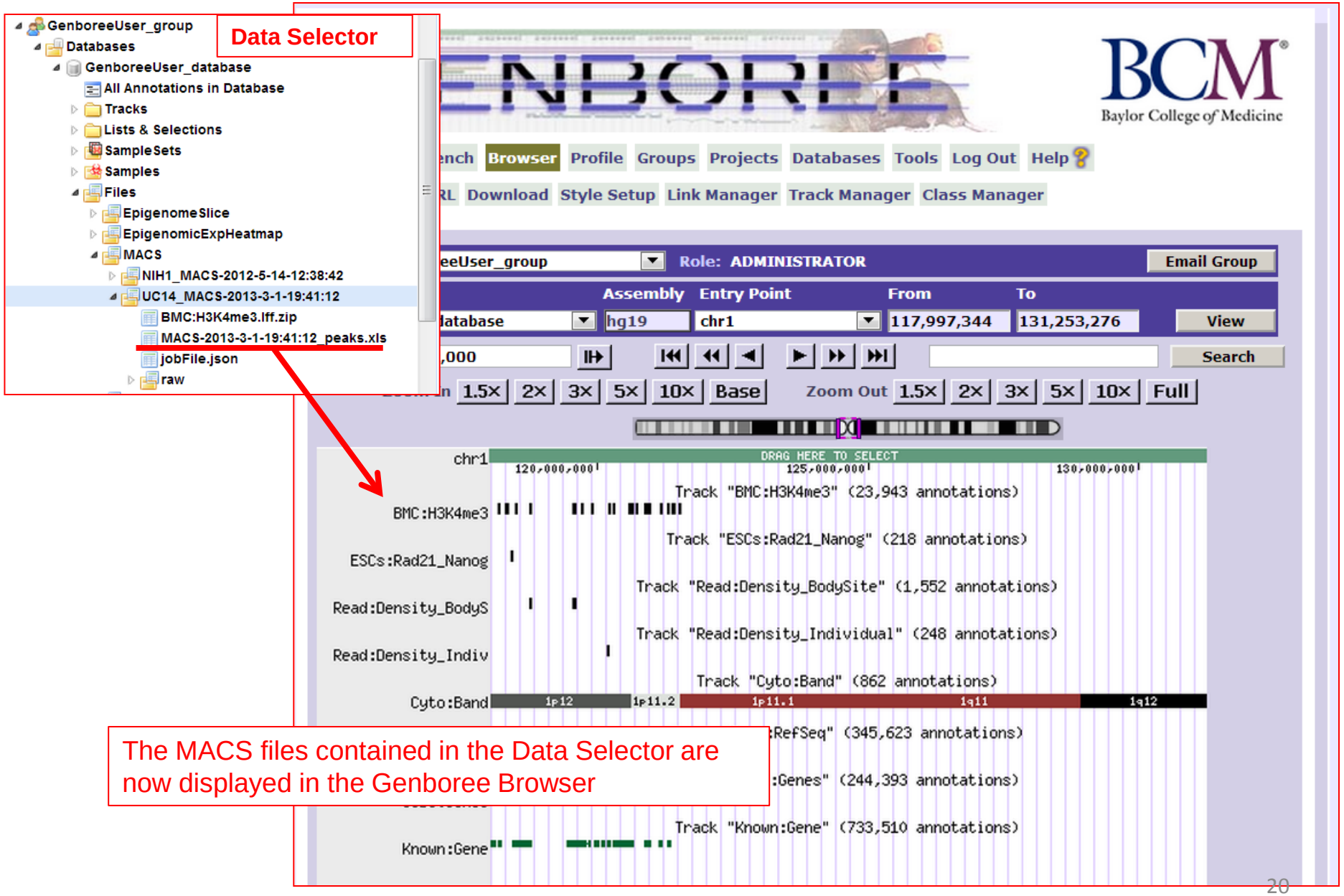
MAC Results Uploaded to Genboree



Tell the Genboree Browser What You Wish to View



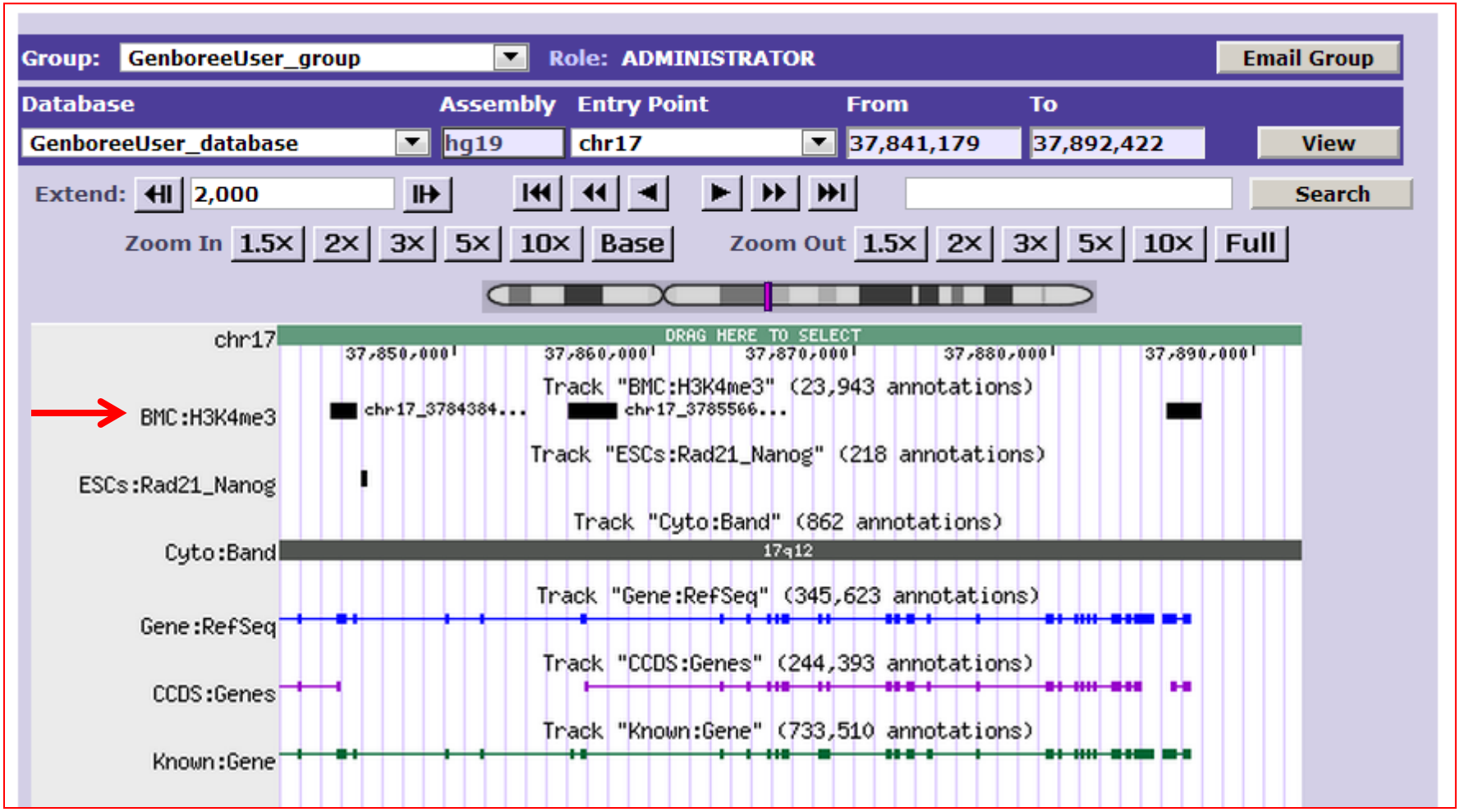
View MACS Results in the Genboree Browser



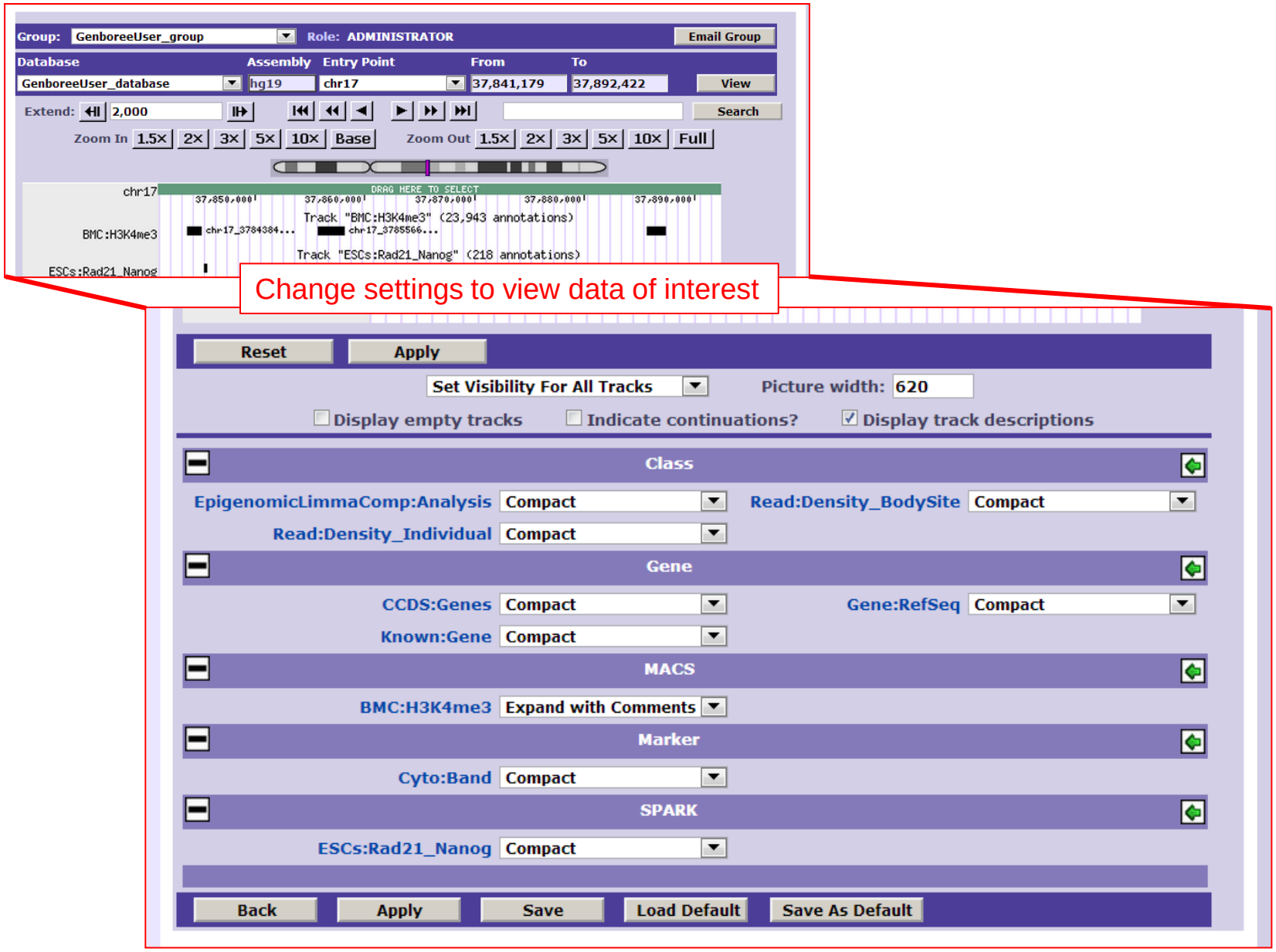
The MACS files contained in the Data Selector are now displayed in the Genboree Browser

Change View of BMC:H3Kme3 Data

Select chr17 and type in new “From” and “To” coordinates to generate this view



Explore Visualization Settings in Genboree Browser



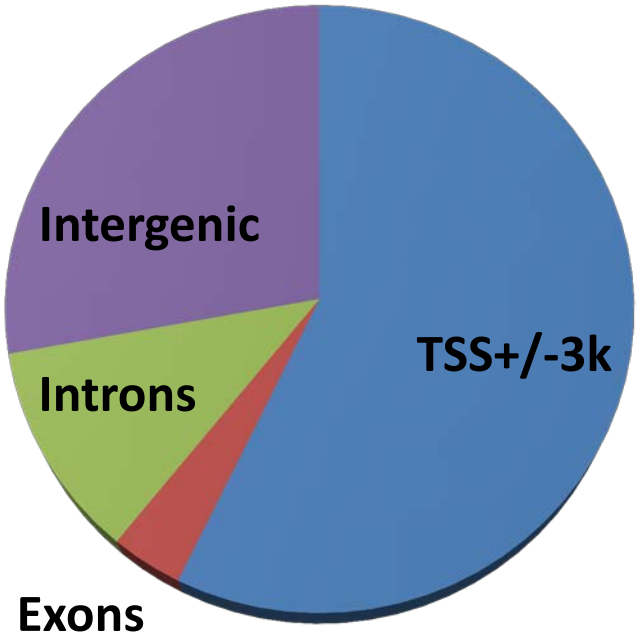
Peak evaluation

- H3K4me3
  - Actively transcribed genes
  - Poised genes
    - co-occuring with H3K27me3
  - Overlap with gene promoters
- H3K4me1
  - Mark of distal enhancers
- Expect different distribution H3K4me3
- Simple check
  - Summary of peak distribution
  - With respect to gene elements

	H3K4me3		H3K4me1	
TSS+/-3k	13693	57.2 %	5381	17.7%
Exons	868	3.6%	2154	7.1%
Introns	2707	11.3%	8416	27.7%
Intergenic	6675	27.9 %	14429	47.5%
Total	*23943*		*30380*	

Please note: Genboree does not generate these visuals (data is exported and pie charts are generated in Excel)

H3K4me3



H3K4me1