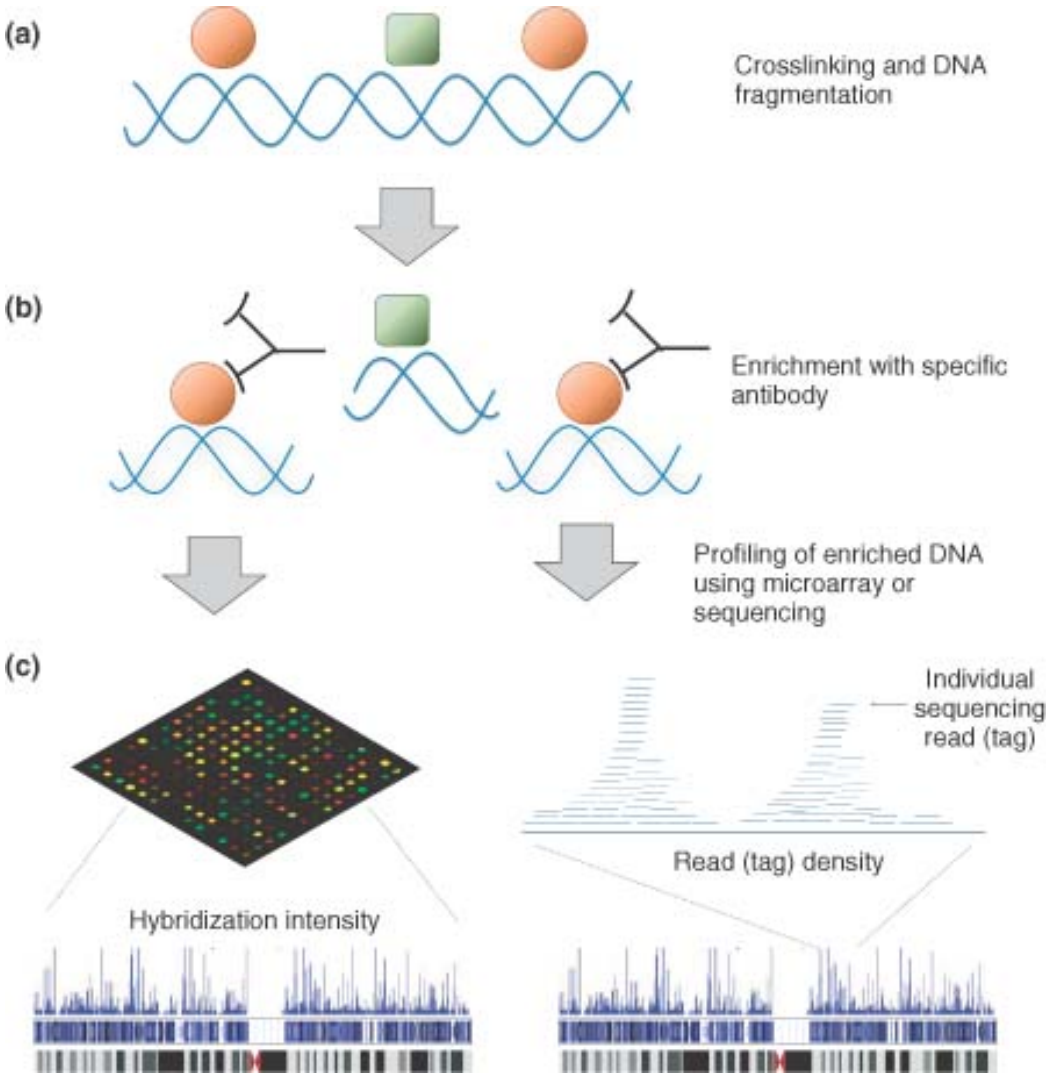
ChIP-Seq, mRNA-Seq, & Resequencing via the Genboree Workench

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



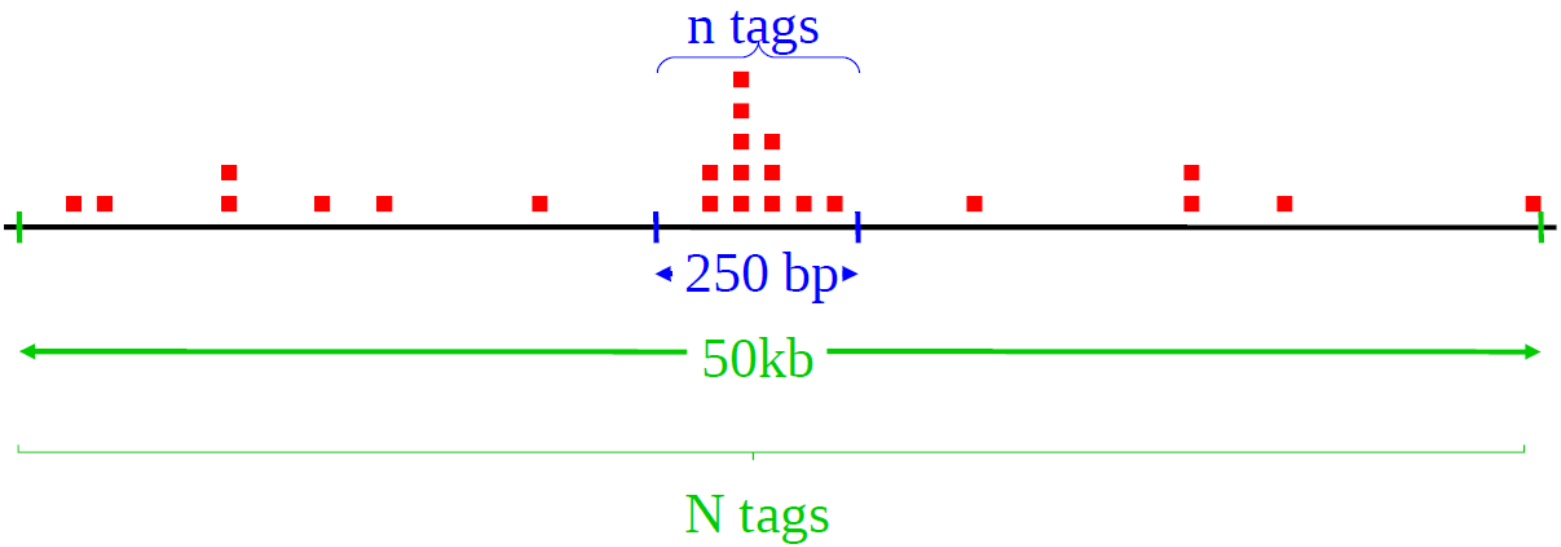
Peak Calling

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

Peak Calling

- Input from
  - NIH Epigenomic Roadmap mapping centers
  - ENCODE analysis group
- Deployed at EDACC/BRL
  - MACS (Harvard)
  - FindPeaks (BCGSC/UCSF)
  - HotSpot (UW)
- Tools under consideration
  - Scripture and PostProcessing (BI)
- Future tools
  - PeakSeq
  - MRE/MeDIP integration tools from WashU/UCSF

HotSpot



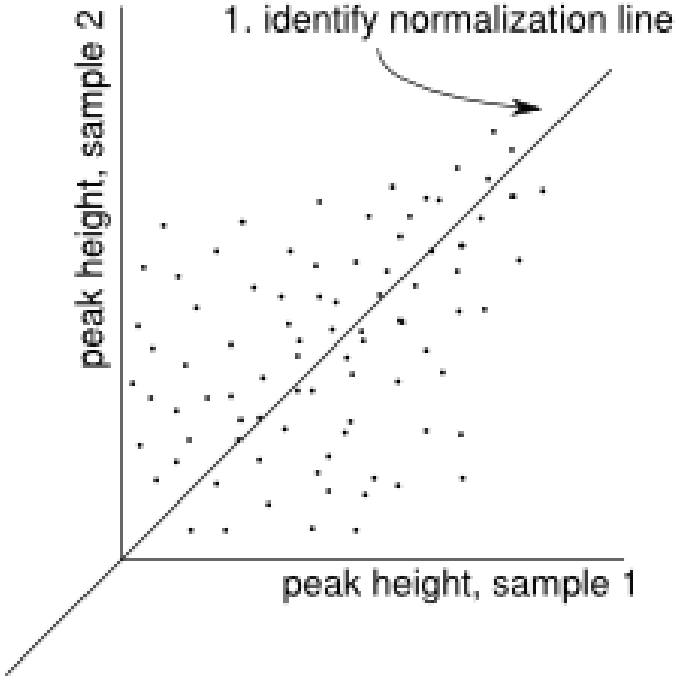
- Scan statistic gauging enrichment
  - z-score based on the binomial distribution.
- Binomial distribution
  - probability of n tags in small window given N tags large window.
  - adjust for local background fluctuations

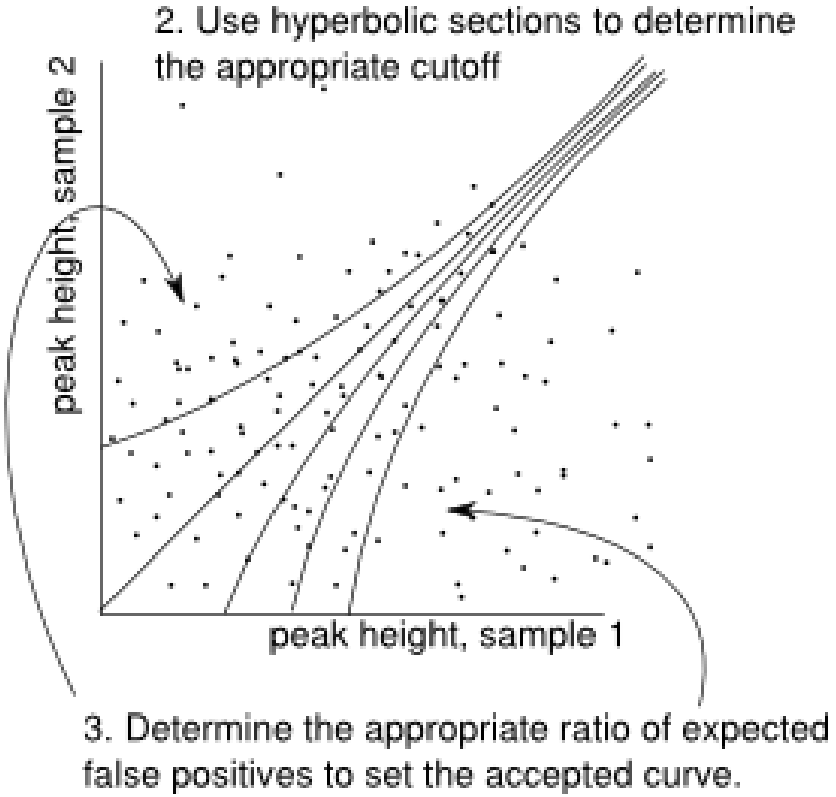
FindPeaks

- In the absence of input
  - Determine areas of enrichment
  - Simulate coverage of genome to determine FDR
  - Subset Peaks at FDR=0.01

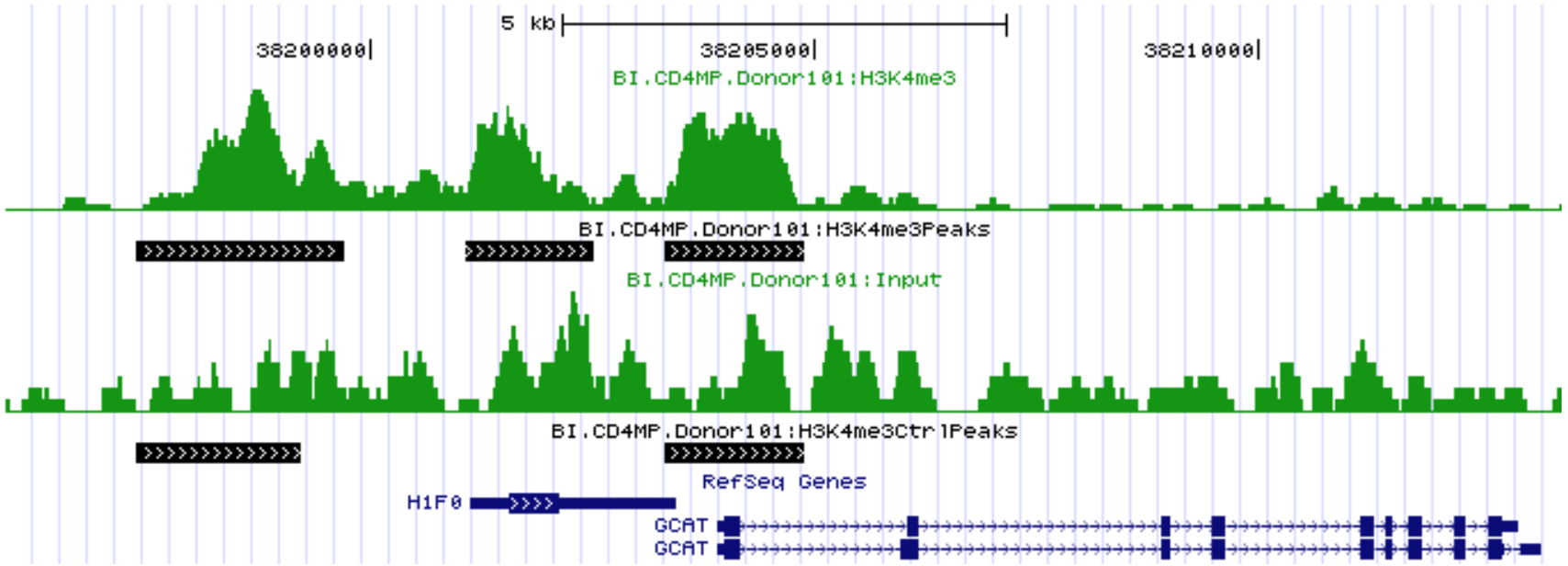
FindPeaks

In the presence of input

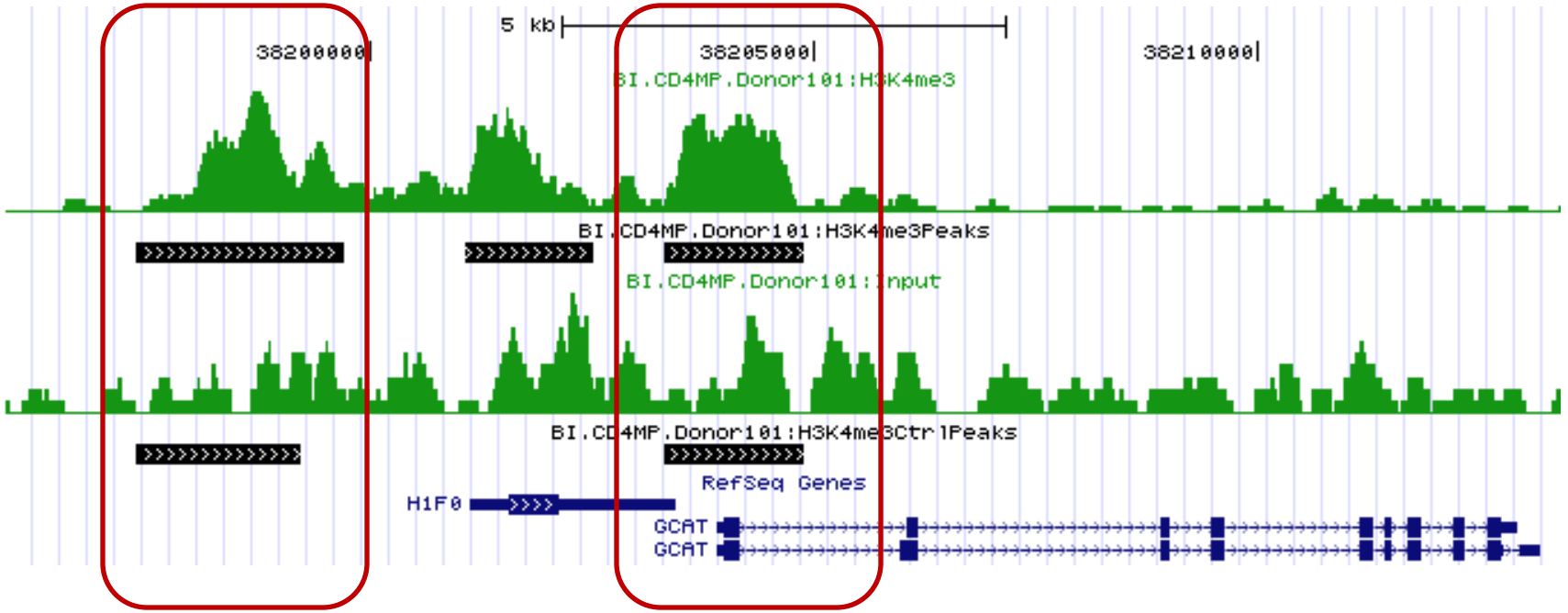




FindPeaks


5 kb
38200000
38205000
38210000
BI.CD4MP.Donor101:H3K4me3
BI.CD4MP.Donor101:H3K4me3Peaks
BI.CD4MP.Donor101:Input
BI.CD4MP.Donor101:H3K4me3CtrlPeaks
RefSeq Genes
H1F0
GCAT
GCAT

FindPeaks

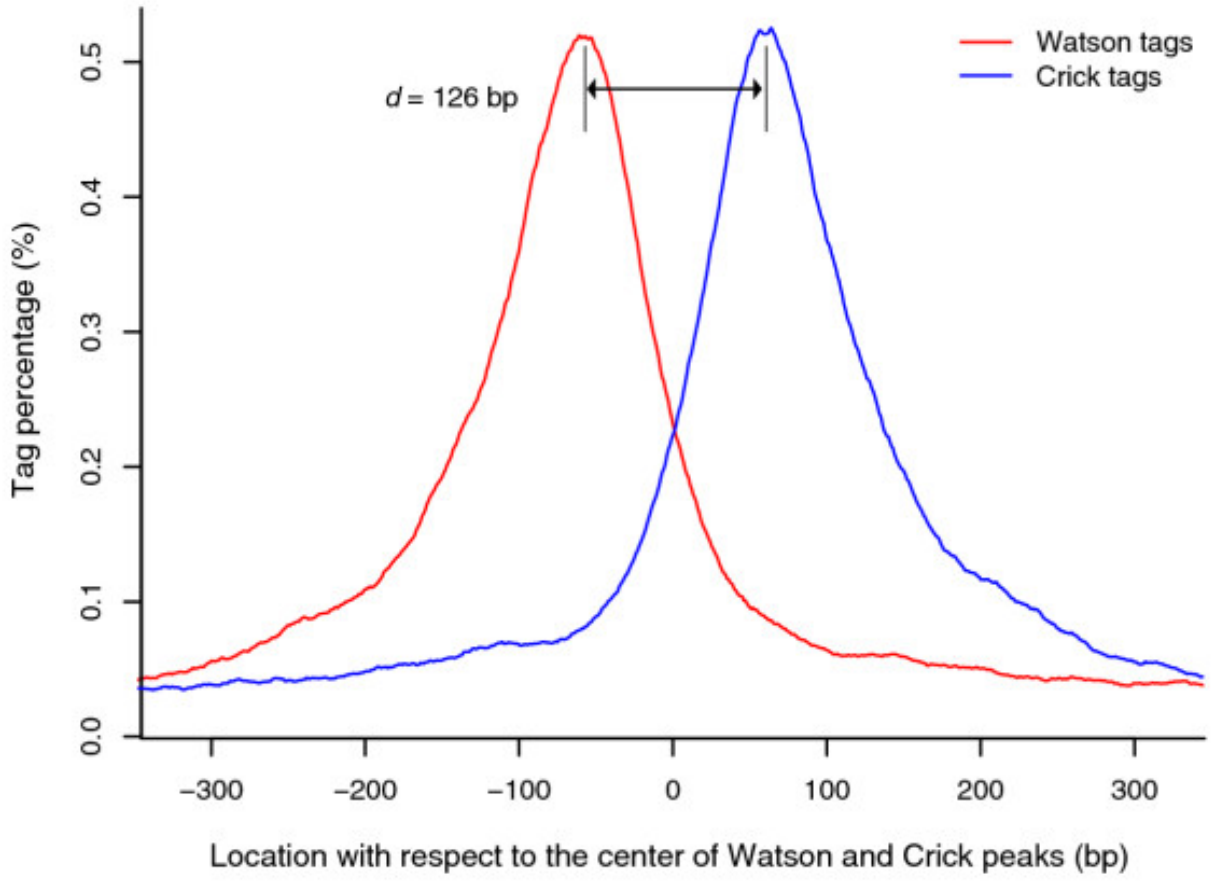


Only peaks above background are reported

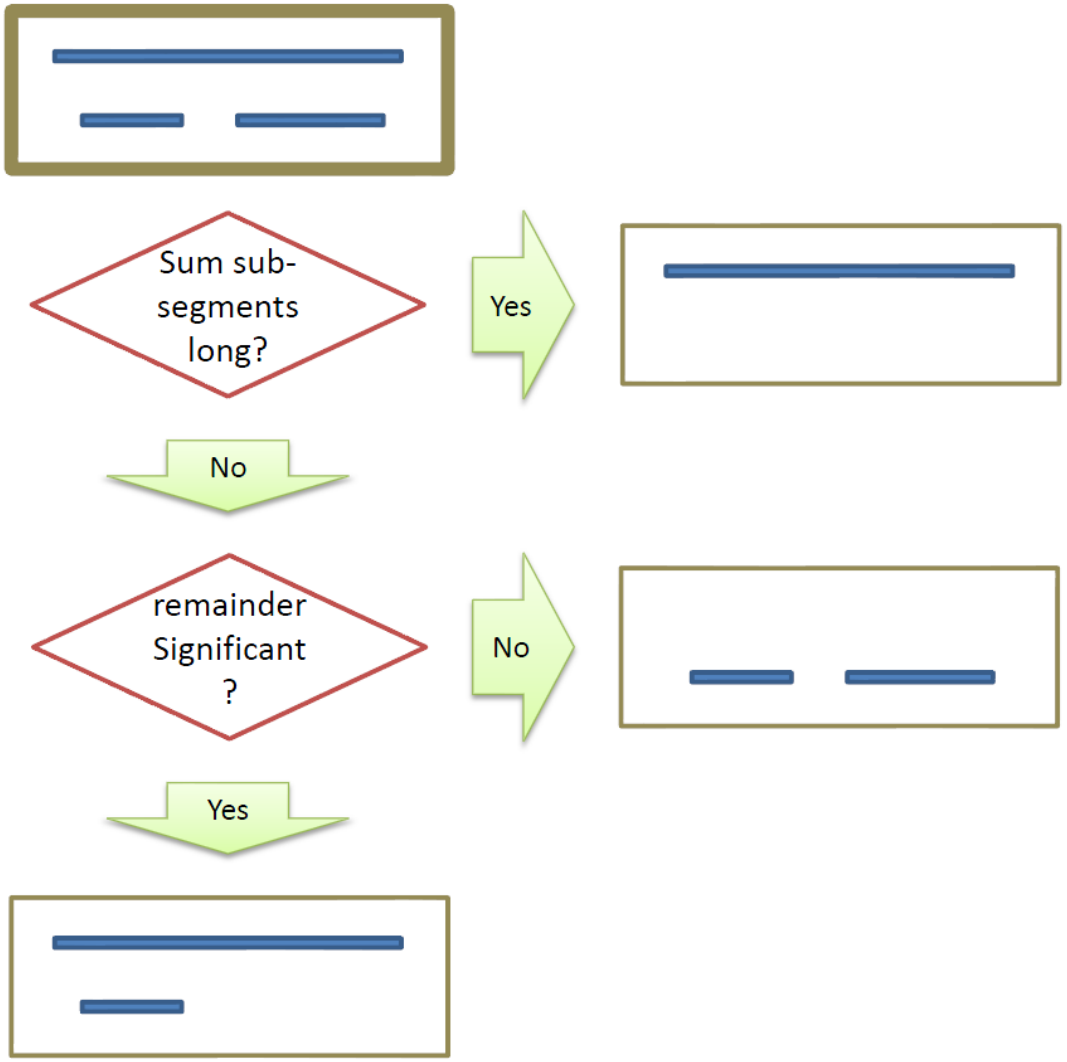
MACS

- Model Based Analysis of ChIP-Seq
- Call peaks on both strands
  - Infer bandwidth length
- Infer local distribution parameters
- Account for local biases in sequencing

MACS


Watson tags
Crick tags
d = 126 bp
Tag percentage (%)
0.0
0.1
0.2
0.3
0.4
0.5
−300
−200
−100
0
100
200
300
Location with respect to the center of Watson and Crick peaks (bp)

Post Scripture
Multi Scale Segmentation



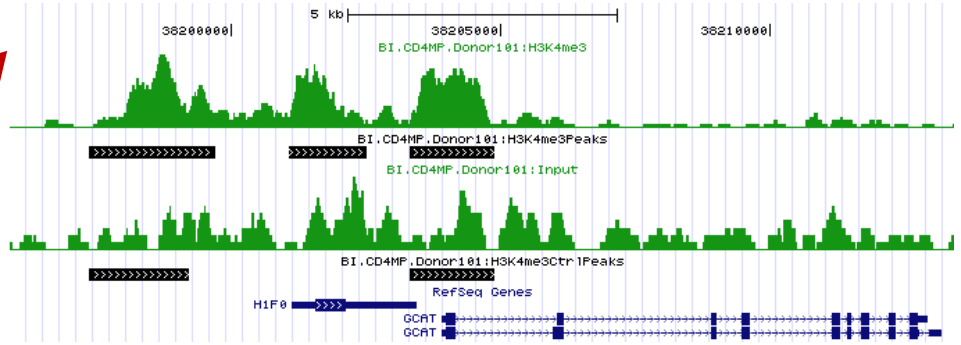
Deployed Tools

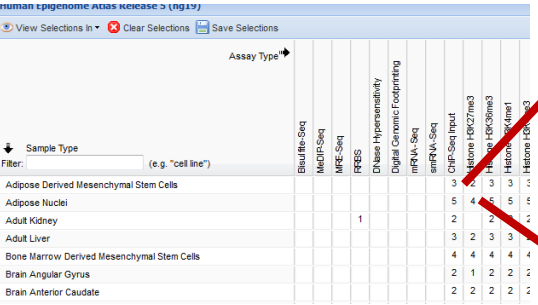
- MACS, FindPeaks
  - Histone modifications, transcription factors
- HotSpot
  - DNASE1/DGF

- Peaks called on all the experiments
  - Release 1 through Release 5
  - ChIP-Seq/DNAse1/Digital Genomic Footprinting
  - ~900 experiments (out of ~1500 total)

Exposing NIH Epigenomic Consortium Peak Calls

UCSC Visualization



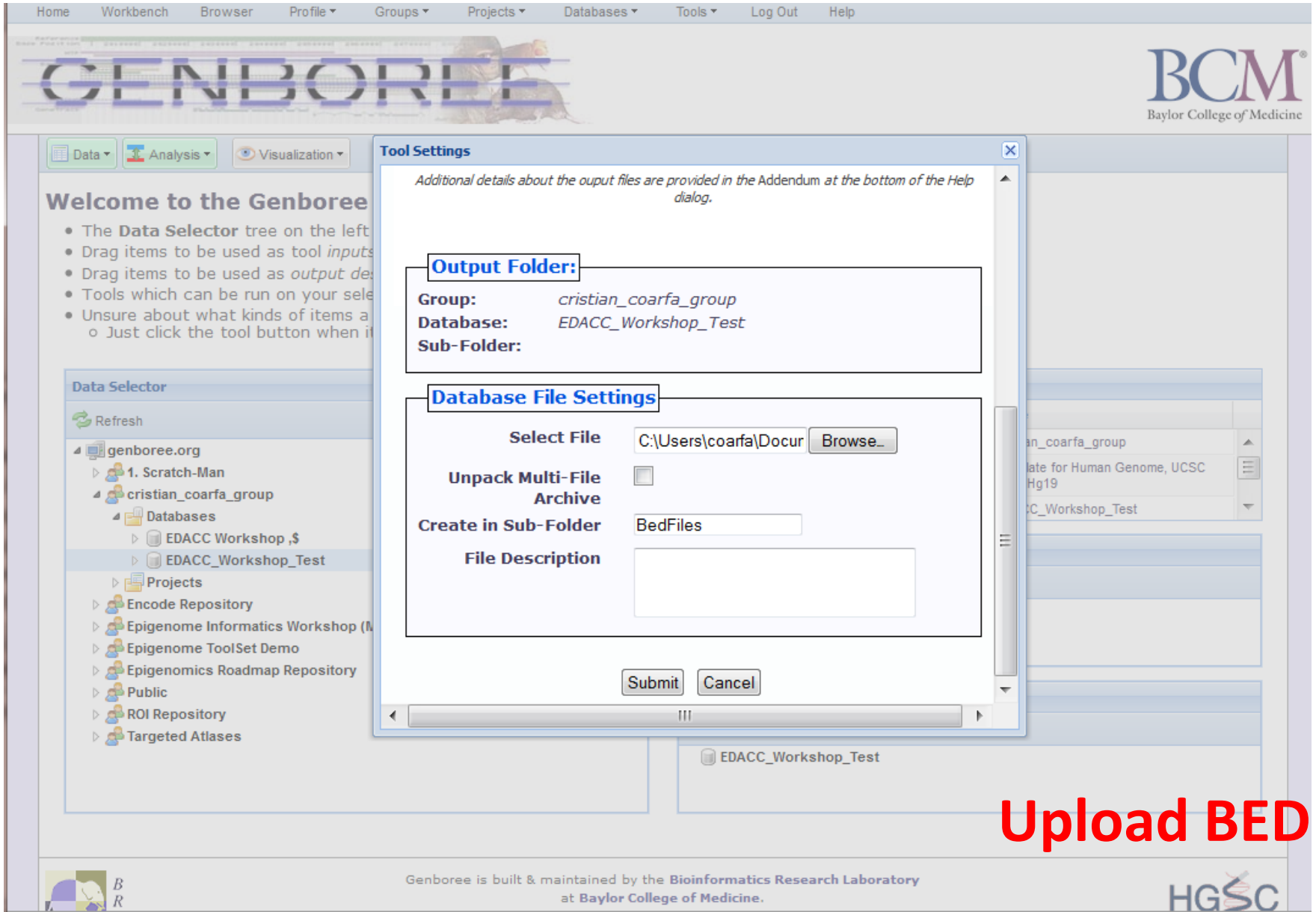


HTTP/FTP Download

Index of /EdaccData/Release-5/study-sample-experiment /BI/CD4_Memory_Primary_Cells/Histone_H3K4me3/

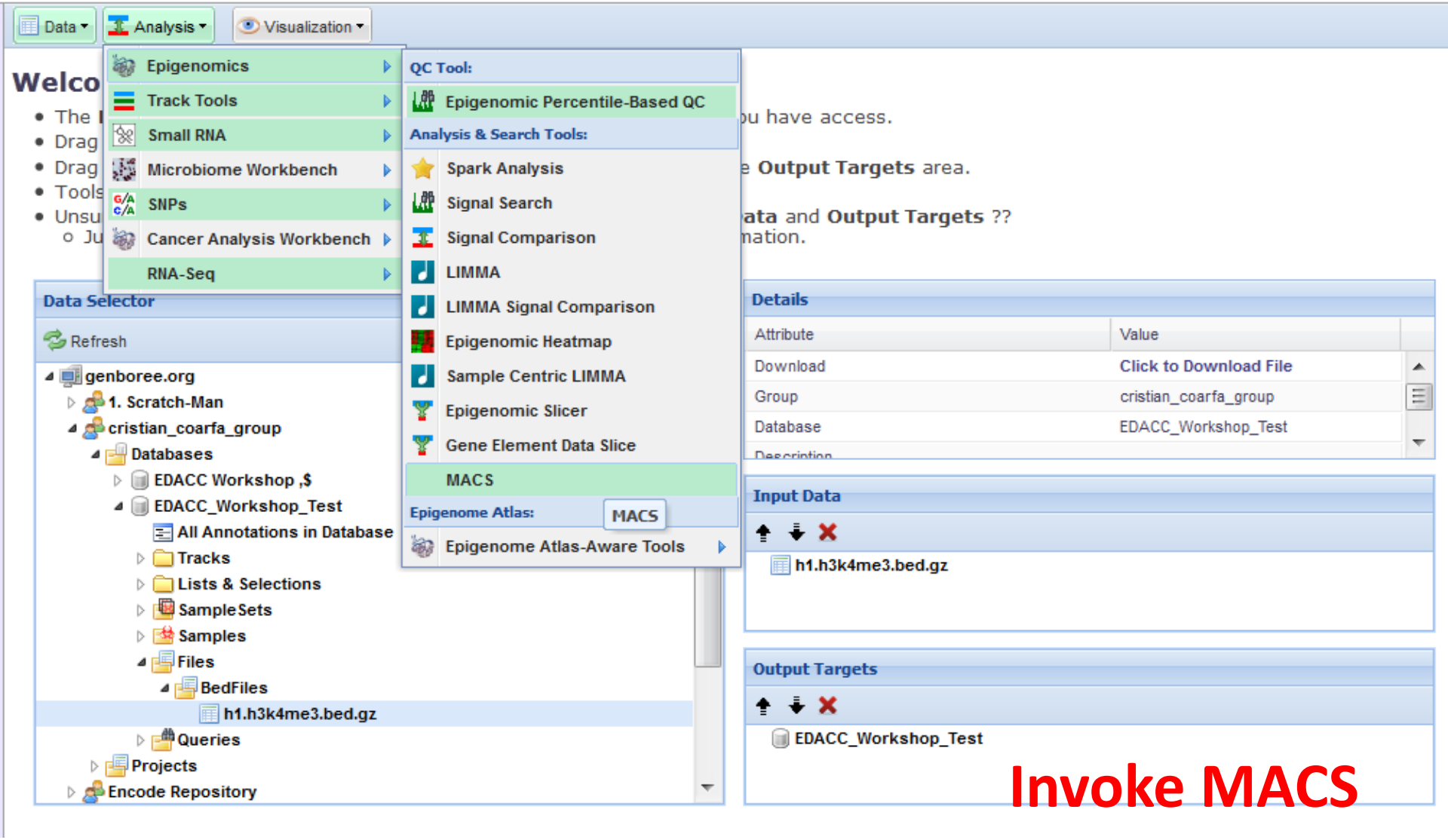
../		
BI.CD4_Memory_Primary_Cells.H3K4me3.20.bed.gz	28-Dec-2011 18:27	84M
BI.CD4_Memory_Primary_Cells.H3K4me3.20.wig.gz	21-Dec-2011 21:14	18M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_100_7..>	28-Dec-2011 18:27	418M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_100_7..>	21-Dec-2011 21:14	45M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_101_8..>	28-Dec-2011 18:27	312M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_101_8..>	21-Dec-2011 21:14	37M

MACS in the Genboree Workbench

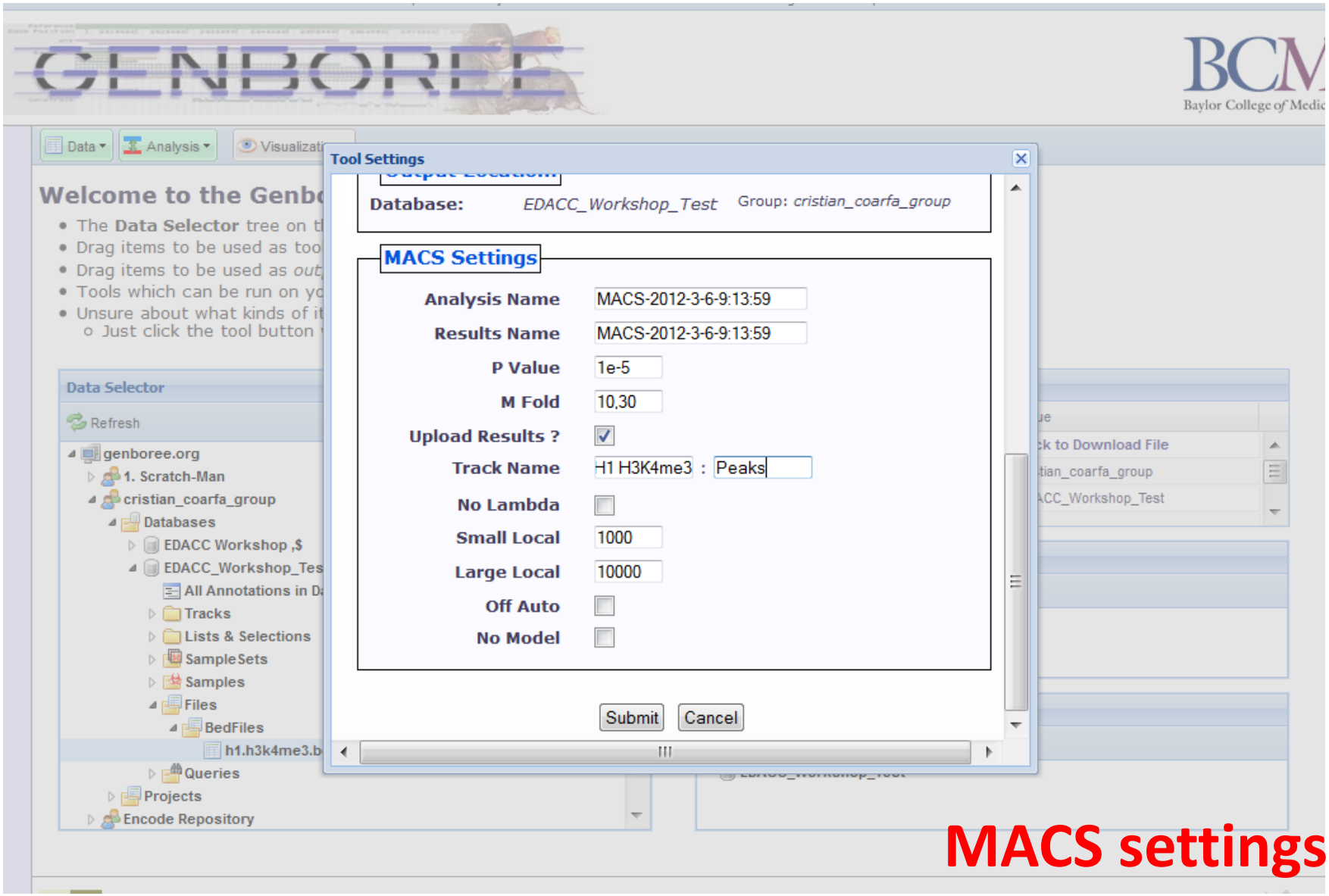


Upload BED file

MACS in the Genboree Workbench

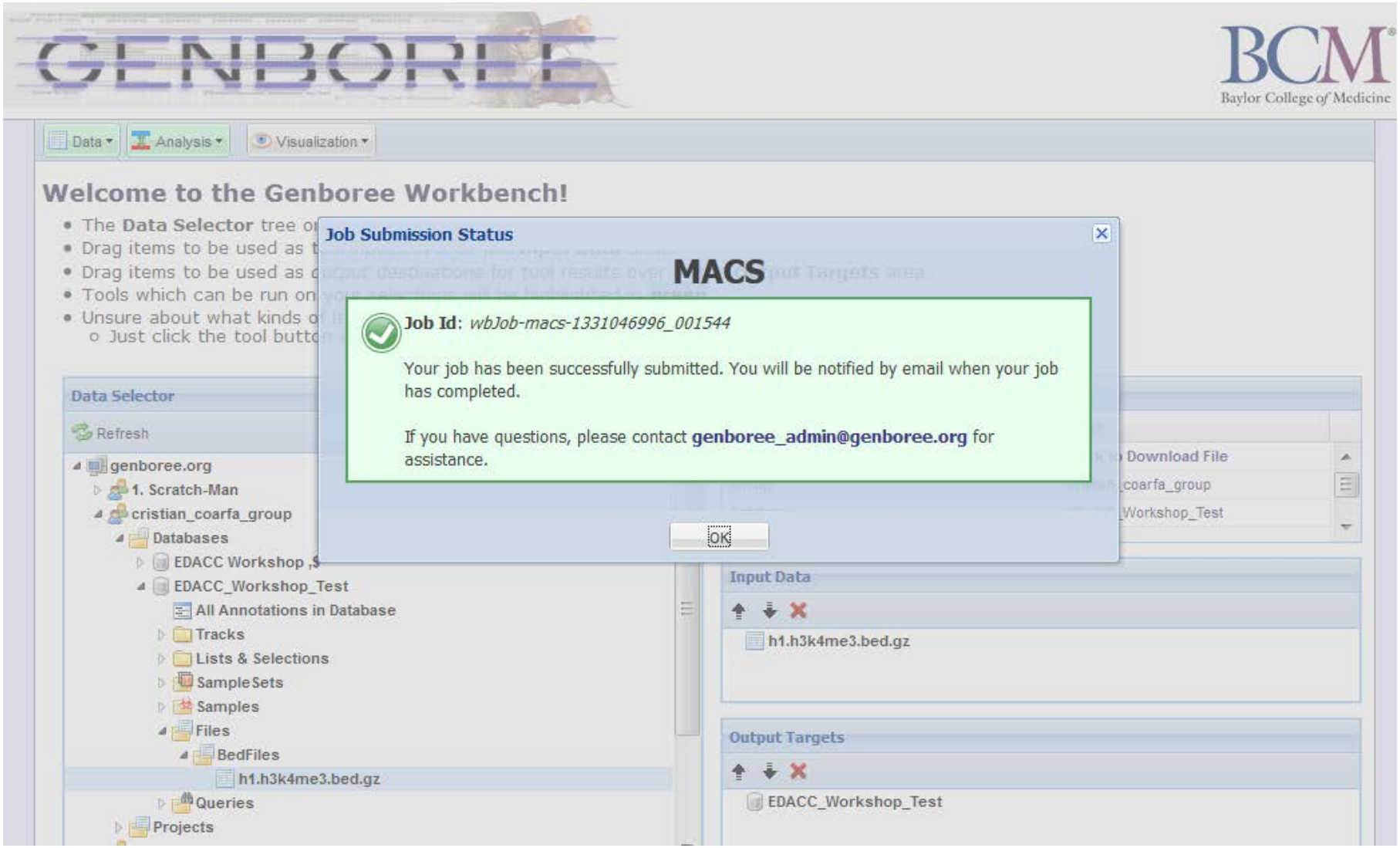


MACS in the Genboree Workbench

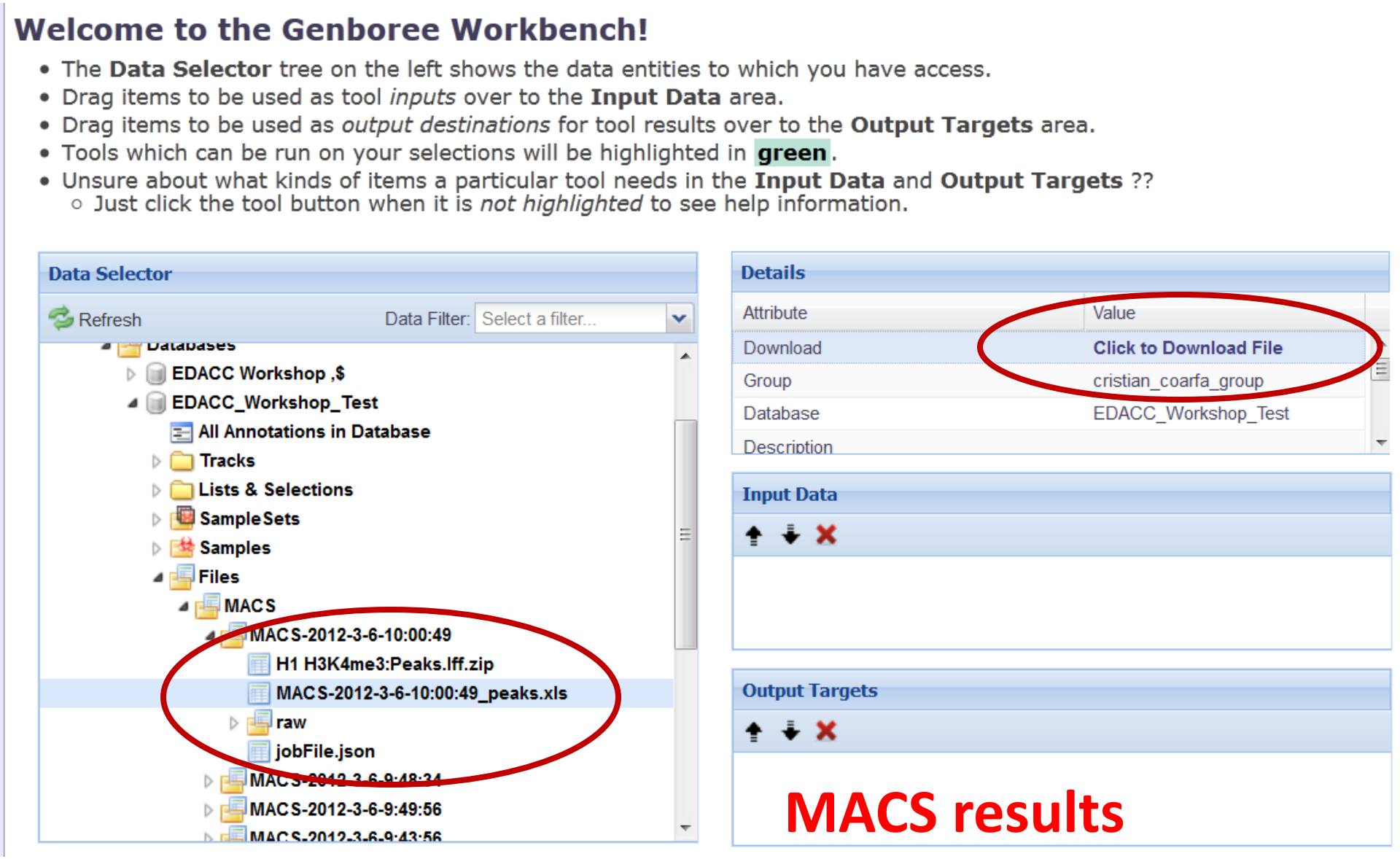


MACS settings

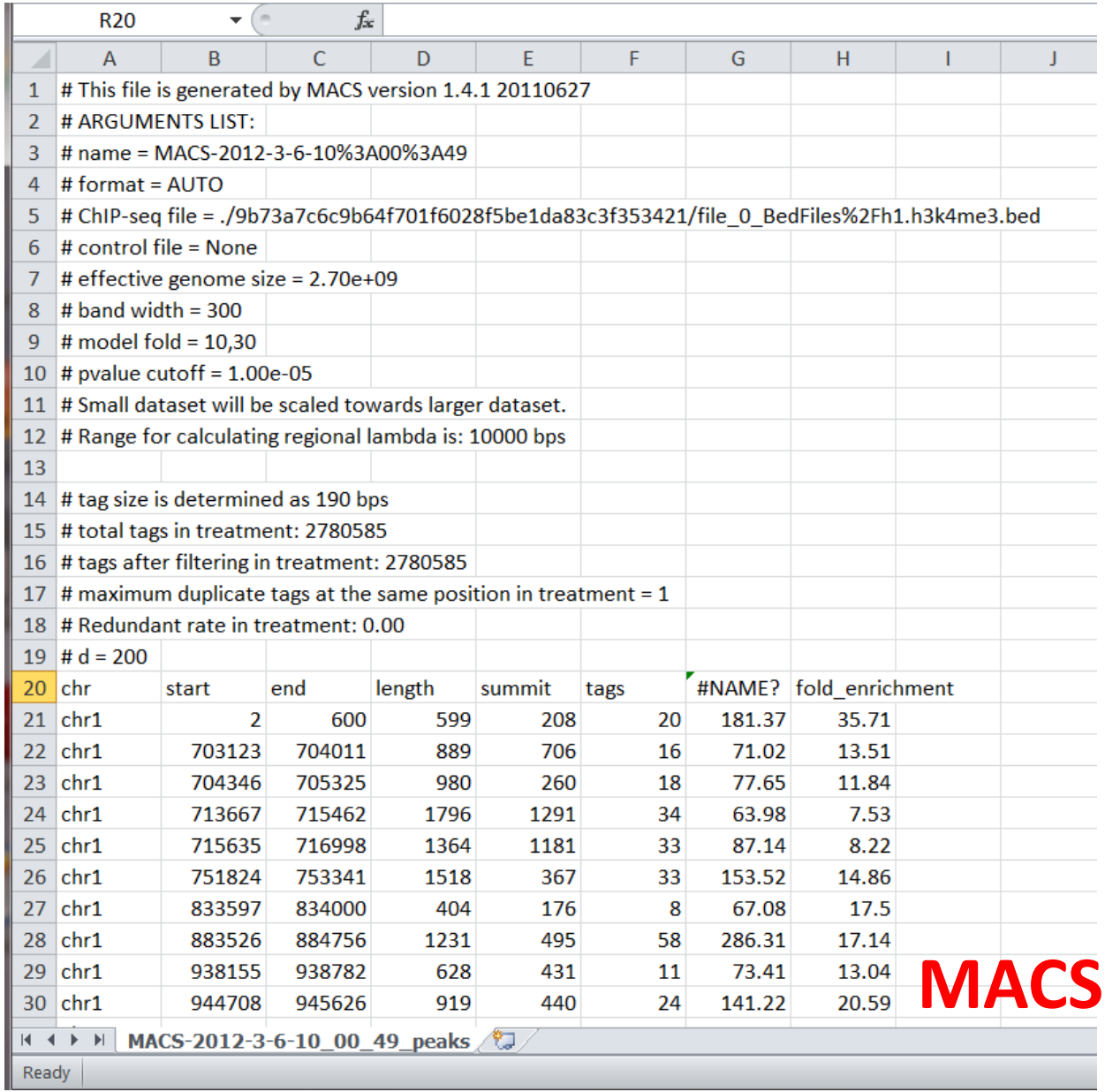
MACS in the Genboree Workbench



MACS in the Genboree Workbench



MACS in the Genboree Workbench



R20

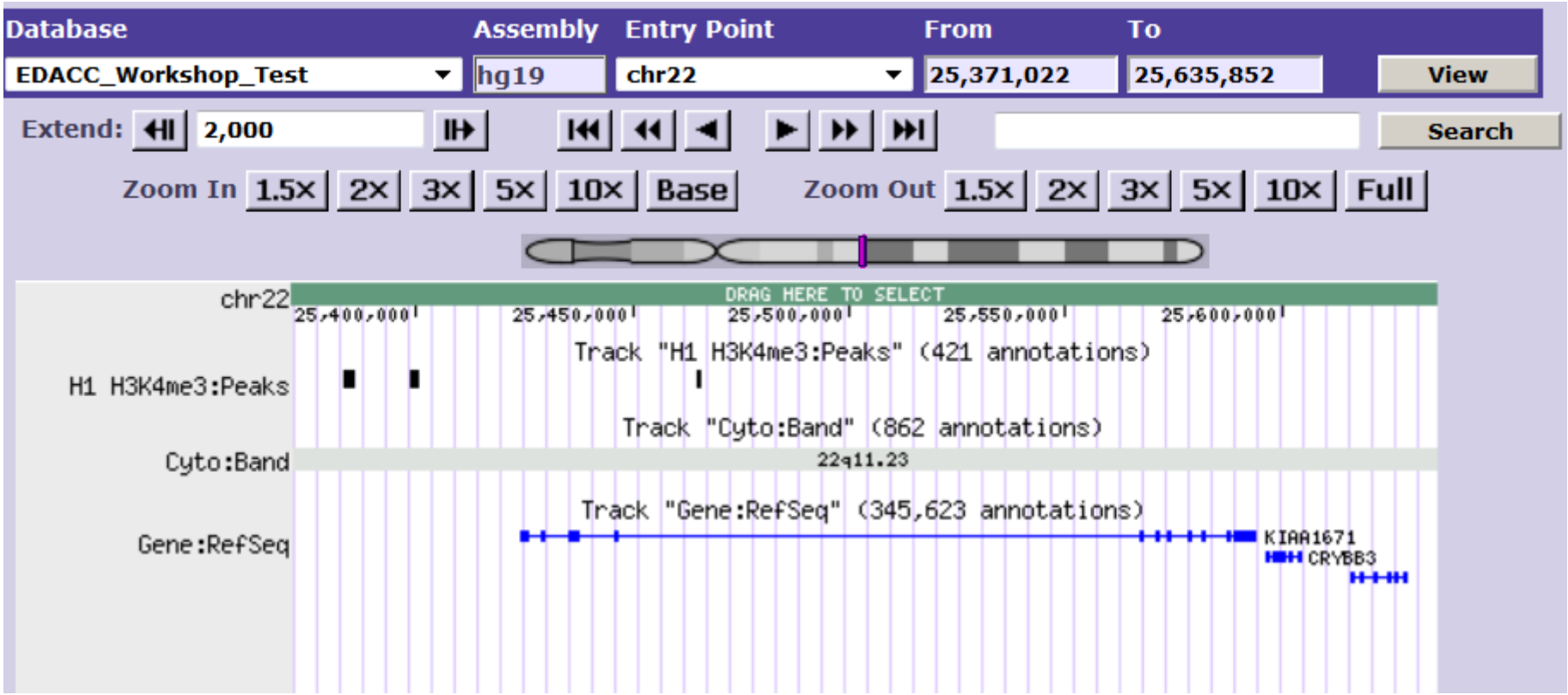
	A	B	C	D	E	F	G	H	I	J
1	# This file is generated by MACS version 1.4.1 20110627									
2	# ARGUMENTS LIST:									
3	# name = MACS-2012-3-6-10%3A00%3A49									
4	# format = AUTO									
5	# ChIP-seq file = ./9b73a7c6c9b64f701f6028f5be1da83c3f353421/file_0_BedFiles%2Fh1.h3k4me3.bed									
6	# control 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 190 bps									
15	# total tags in treatment: 2780585									
16	# tags after filtering in treatment: 2780585									
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	#NAME?	fold_enrichment		
21	chr1	2	600	599	208	20	181.37	35.71		
22	chr1	703123	704011	889	706	16	71.02	13.51		
23	chr1	704346	705325	980	260	18	77.65	11.84		
24	chr1	713667	715462	1796	1291	34	63.98	7.53		
25	chr1	715635	716998	1364	1181	33	87.14	8.22		
26	chr1	751824	753341	1518	367	33	153.52	14.86		
27	chr1	833597	834000	404	176	8	67.08	17.5		
28	chr1	883526	884756	1231	495	58	286.31	17.14		
29	chr1	938155	938782	628	431	11	73.41	13.04		
30	chr1	944708	945626	919	440	24	141.22	20.59		

MACS-2012-3-6-10_00_49_peaks

Ready

MACS results

Results uploaded to Genboree



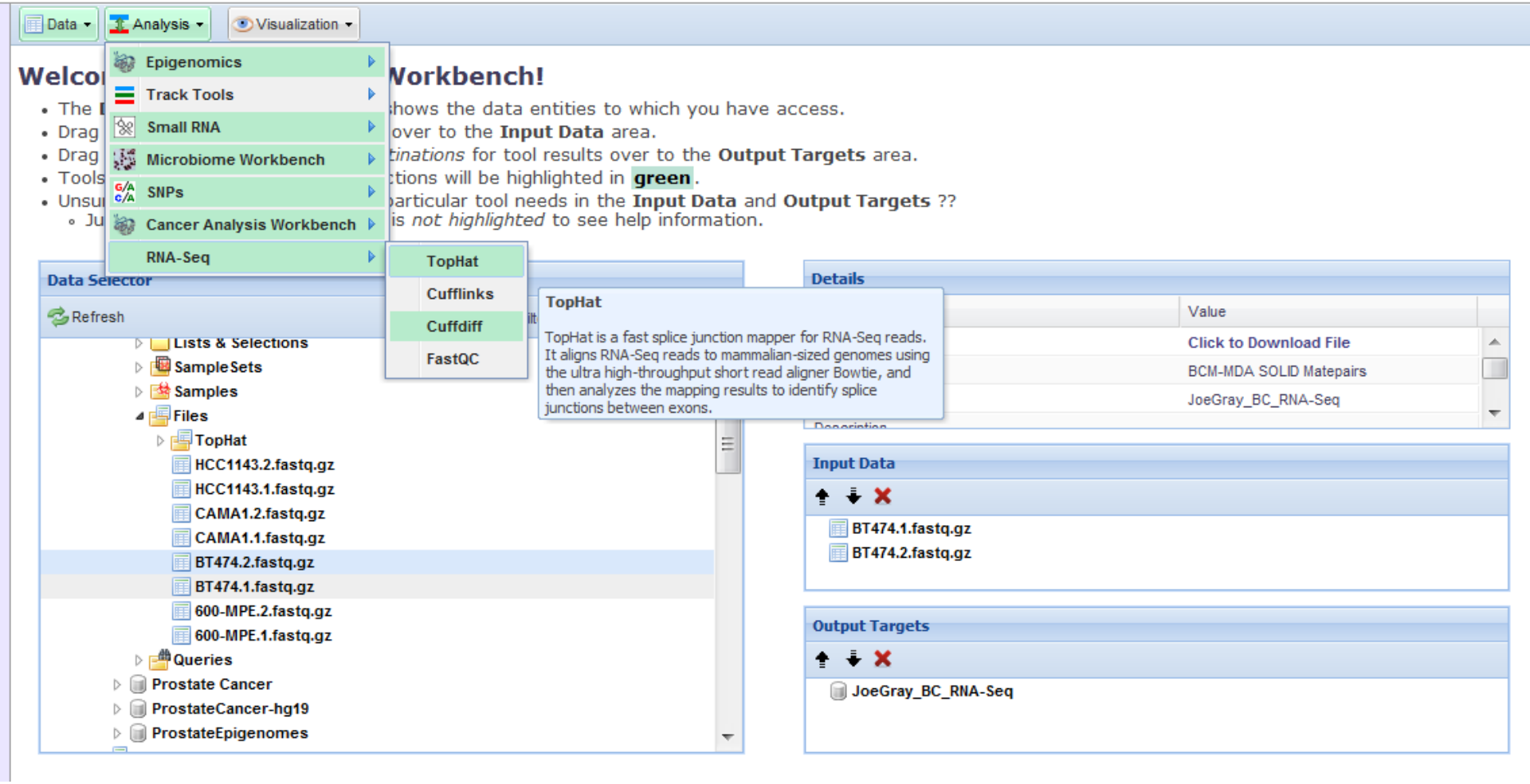
Coming attractions

- Peak interpretation
  - Gene promoters, miRNA promoters
  - Gene elements: introns, exons, 3’UTR, 5’UTR
  - Repeats
  - User defined

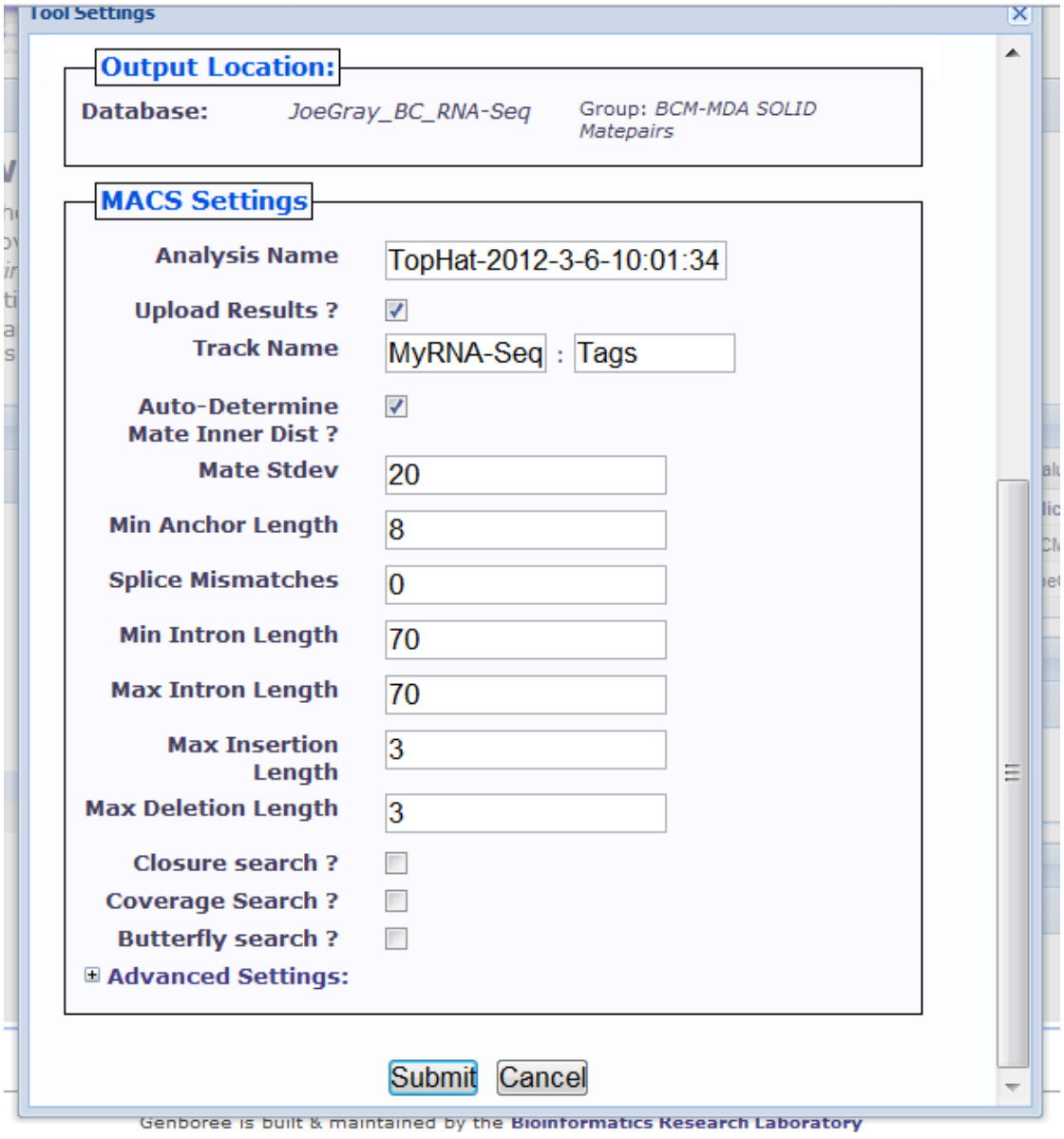
mRNA-Seq

- Deployed
  - TopHat
  - Cufflinks
  - Cuffdiff
  - FastQC
- Results usable in resequencing pipeline

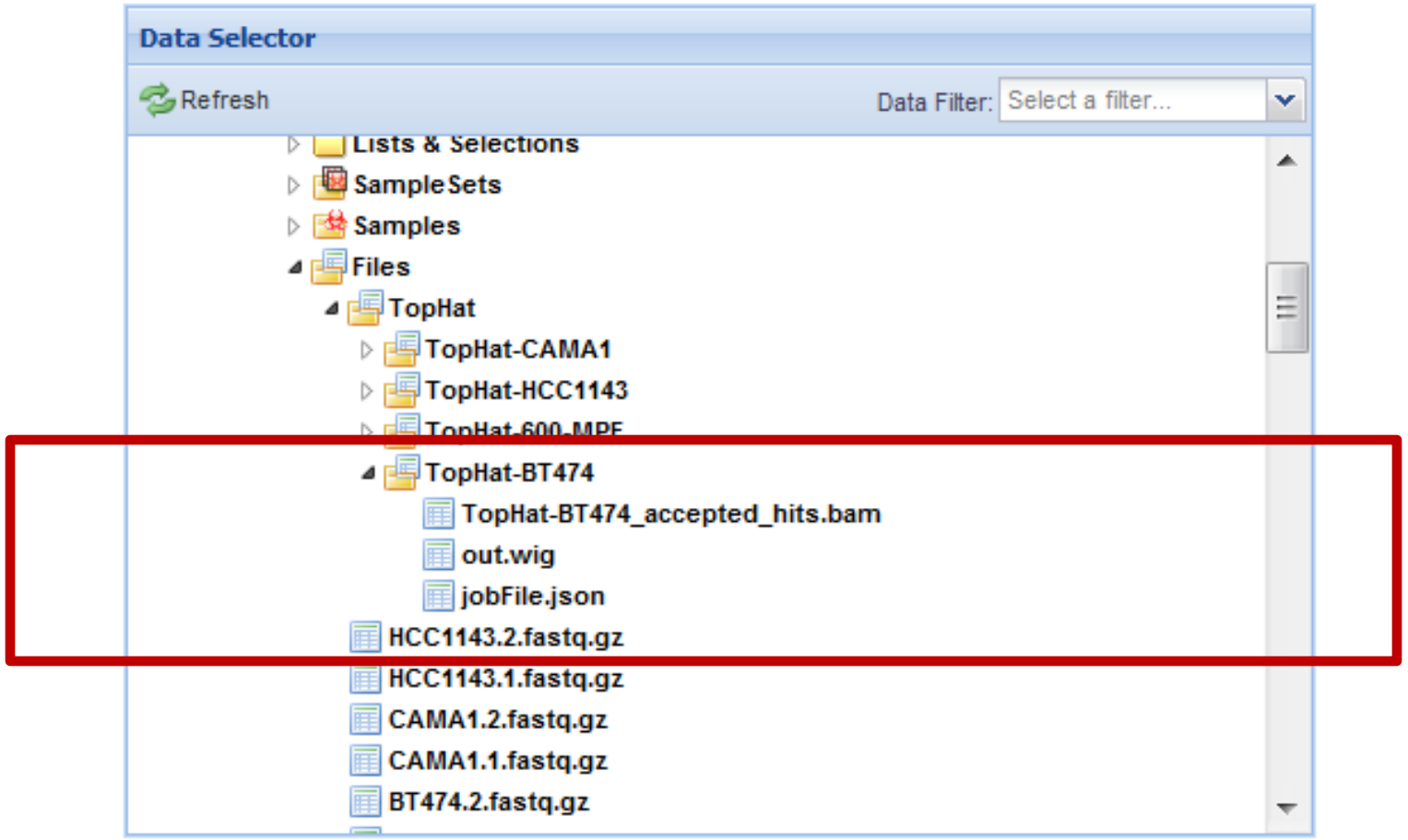
TopHat in the Genboree Workbench



TopHat in the Genboree Workbench



TopHat in the Genboree Workbench



TopHat in the Genboree Workbench

Coming attractions

- Alternative splicing
- Intron retention
- Gene fusions
- Current research collaborations
  - Ovarian cancer
  - Novel fusion with 20% recurrence in tumors

Resequencing

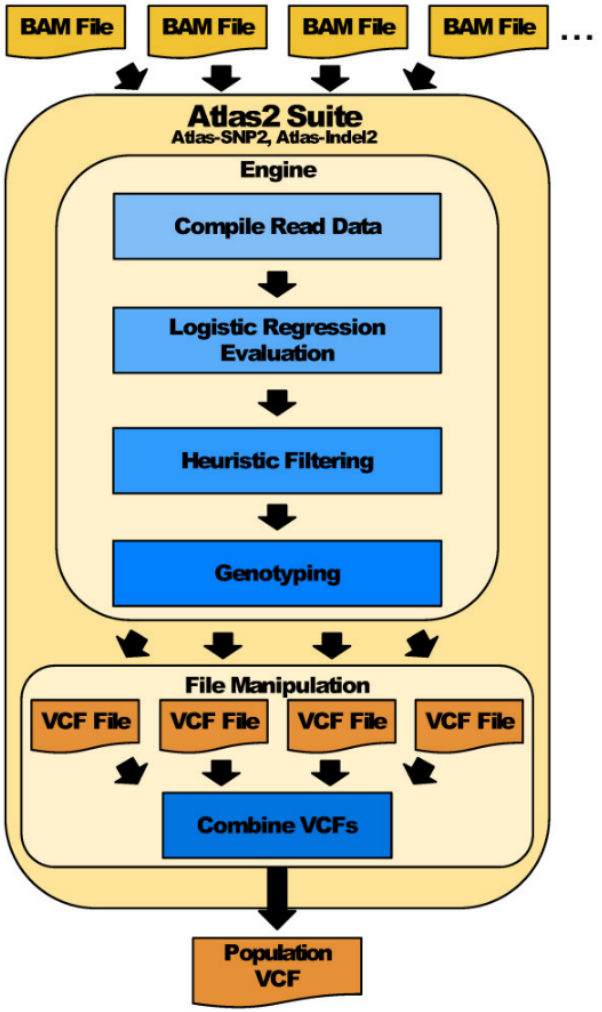
- “Genome in a day”
  - HiSeq 2500/1500
  - 40x coverage
- Whole genome: affordable
  - Less than $5000
- Whole exome: very attractive
  - Heterogeneous “tumor” samples
- SNP/Indel caller: Atlas-SNP2

Atlas-SNP2

Over 4000 accesses since publication

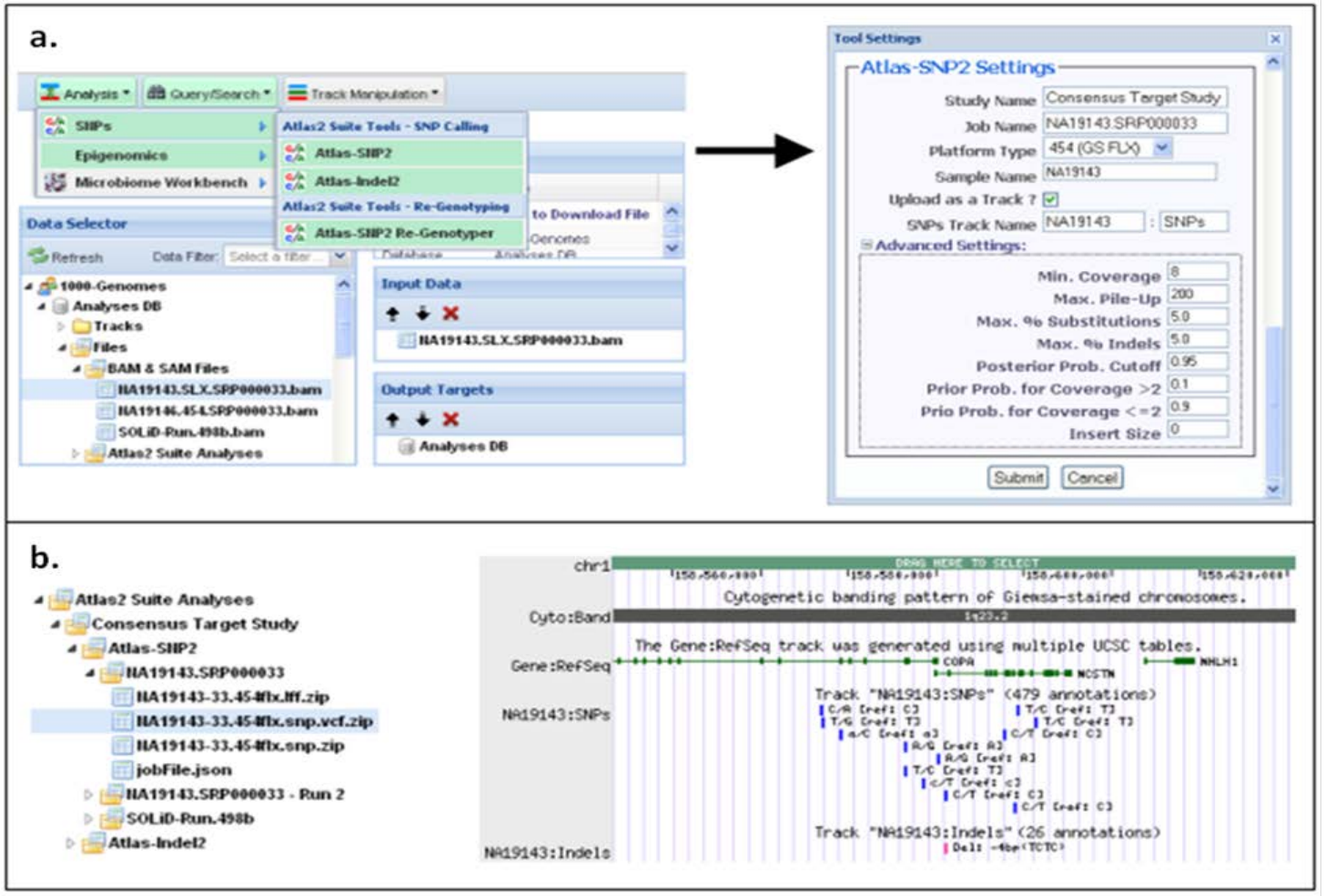
Atlas-SNP2

- **Linear regression model accounting for sources of sequencing biases and errors**
- **Trained for individual sequencing platforms and chemistry**
- **Models for 454, Illumina, SOLID**
- **Constantly updated**

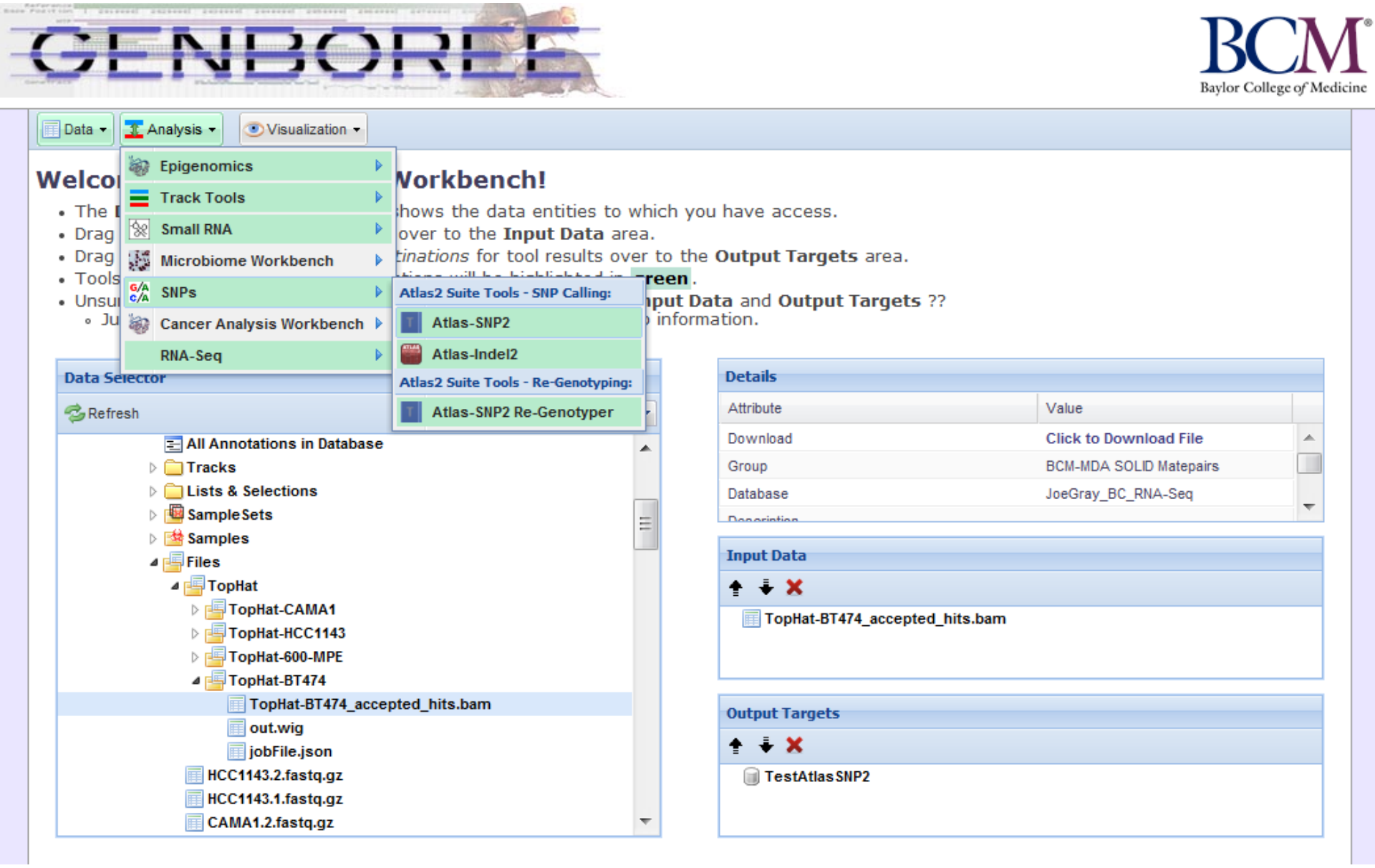


$$\text{logit}(p) = -19.18 - 0.4779 \times ref_var_ratio + 0.2368 \times mean_NBQ + 4.520 \times strand_dir$$
$$+ 0.5436 \times mean_dist3' + 0.08243 \times mean_VBQ - 0.00796 \times mean_NBQ \times mean_dist_3'$$
$$- 0.1122 \times strand_dir \times mean_dist_3'$$

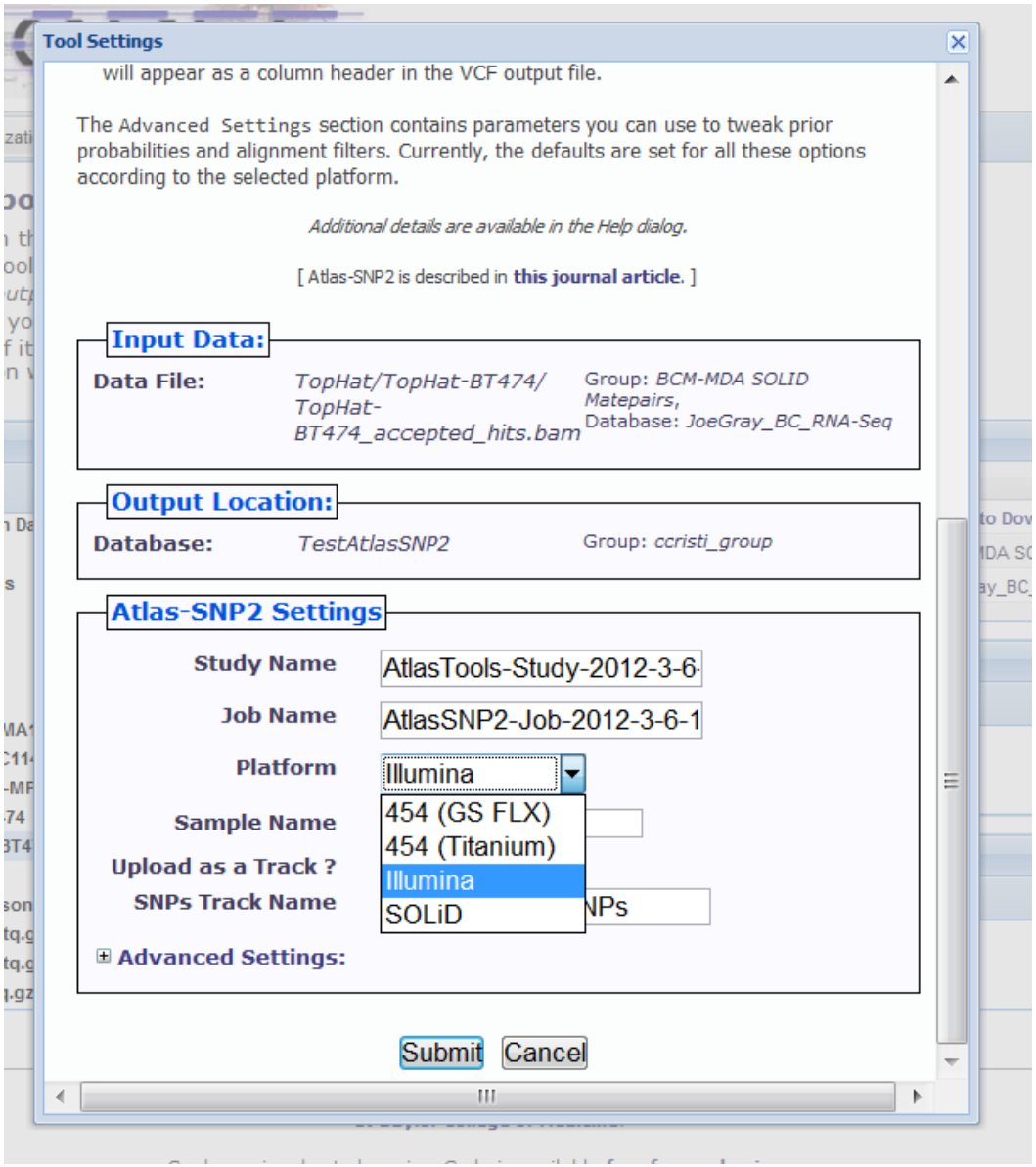
Atlas-SNP2 and Genboree



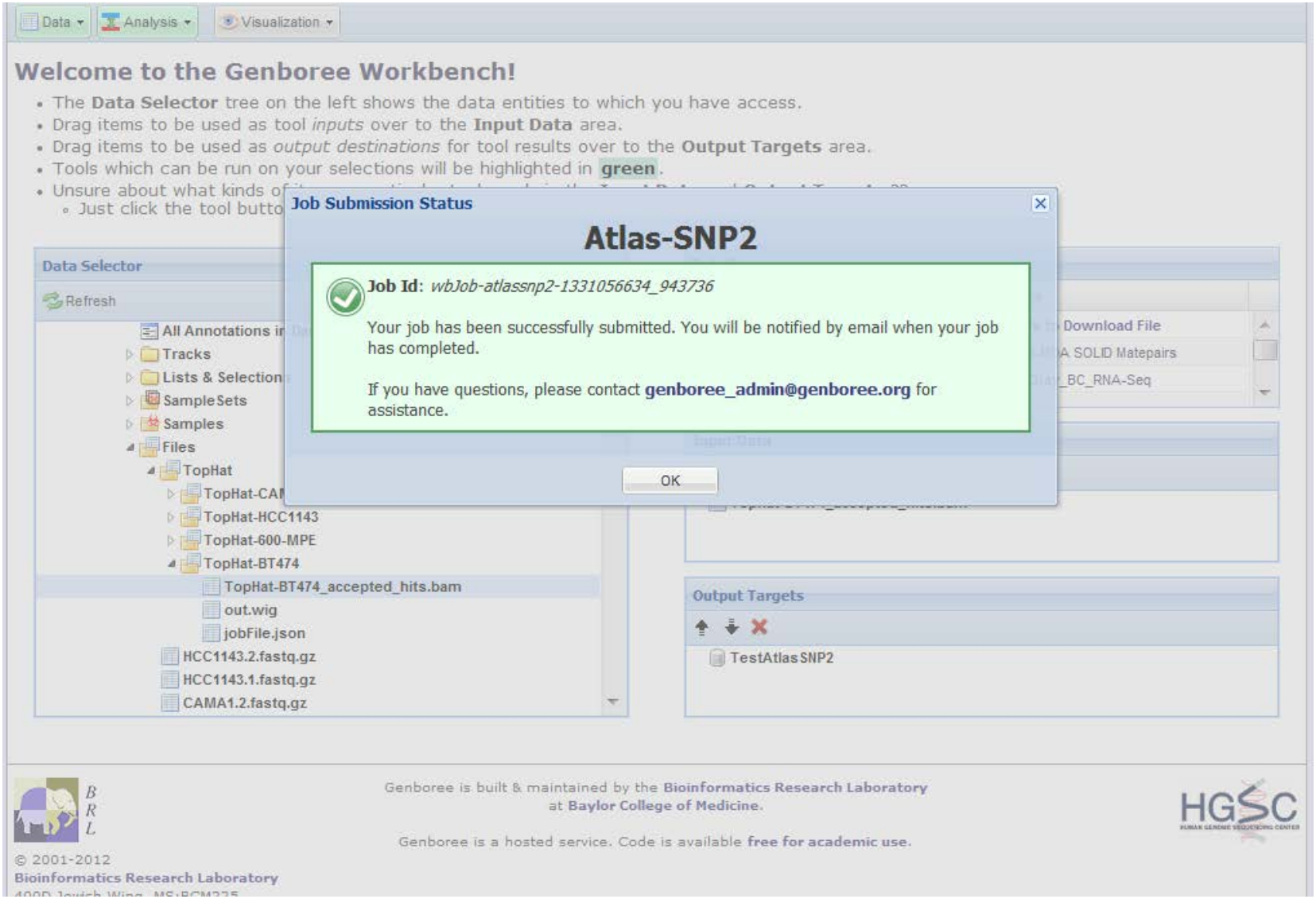
Atlas-SNP2 in the Genboree Workbench



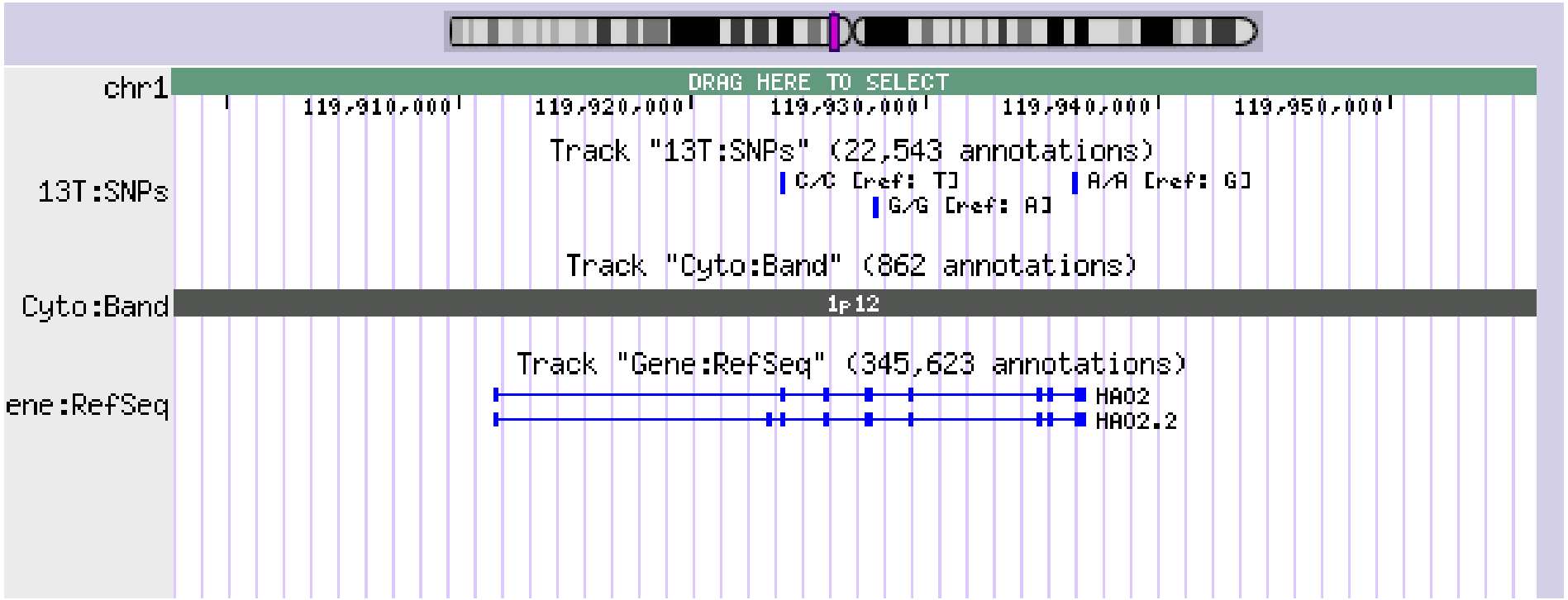
Atlas-SNP2 in the Genboree Workbench



Atlas-SNP2 in the Genboree Workbench



Atlas-SNP2 in the Genboree Workbench



Atlas-SNP2 in the Genboree Workbench

G1 | FILTER

	A	B	C	D	E	F	G	H	I	J
1	#CHROI	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMA	My RNA-Seq sample
26	chr1	14653	.	C	T	60	PASS	.	GT:VR:RR:	0/1:37:71:108:.
31	chr1	14677	.	G	A	60	PASS	.	GT:VR:RR:	0/1:37:97:134:.
45	chr1	14907	.	A	G	60	PASS	.	GT:VR:RR:	0/1:4:6:10:.
46	chr1	14930	.	A	G	60	PASS	.	GT:VR:RR:	0/1:5:1:6:.
59	chr1	15211	.	T	G	60	PASS	.	GT:VR:RR:	0/1:8:3:11:.
60	chr1	15274	.	a	T	60	PASS	.	GT:VR:RR:	1/1:9:0:9:.
64	chr1	15696	.	A	T	60	PASS	.	GT:VR:RR:	0/1:3:12:15:.
68	chr1	16378	.	T	C	60	PASS	.	GT:VR:RR:	0/1:4:2:6:.
71	chr1	16606	.	C	A	60	PASS	.	GT:VR:RR:	0/1:3:5:8:.
85	chr1	16963	.	G	A	60	PASS	.	GT:VR:RR:	0/1:34:131:165:.
89	chr1	16996	.	T	C	60	PASS	.	GT:VR:RR:	0/1:67:110:177:.
119	chr1	18064	.	G	T	60	PASS	.	GT:VR:RR:	0/1:6:44:50:.
122	chr1	18200	.	A	G	60	PASS	.	GT:VR:RR:	0/1:3:7:10:.
130	chr1	18643	.	G	A	60	PASS	.	GT:VR:RR:	0/1:3:13:16:.
153	chr1	20250	.	t	C	60	PASS	.	GT:VR:RR:	0/1:4:3:7:.
163	chr1	20813	.	A	G	60	PASS	.	GT:VR:RR:	0/1:3:25:28:.
199	chr1	23647	.	T	G	60	PASS	.	GT:VR:RR:	0/1:4:8:12:.
237	chr1	136781	.	T	C	60	PASS	.	GT:VR:RR:	1/1:4:0:4:.
238	chr1	136830	.	T	C	60	PASS	.	GT:VR:RR:	0/1:4:3:7:.

Coming Attraction

- SNP annotation
  - Annovar
  - Known SNPs : dbSNP, 1000 genomes
  - Genic/intergenic
  - Coding/non-coding variants
  - GWAS
  - Regulatory regions

Acknowledgments

- NIH Epigenomic Roadmap
  - Lisa Chadwick, Fred Tyson, Joni Rutter, John Satterlee, Astrid Haugen, Kim McAllister
- Peak Calling
  - Anshul Kundaje, Bob Thurman (UW), Noam Shoresh (BI), Martin Hirst (BCGSC), Lee Daniels (NIH), Wei Li (BCM)
- RNA-Seq
  - Adrian Lee, Joe Gray, Laising Yen, Kalpana Kannan
- SNP calling, resequencing
  - Fuli Yu, Richard Gibbs, Danny Challis, Jin Yu, Uday Evani
- BRL
  - Cristian Coarfa, Alan R Harris, Chia-Chin Wu, Aleksandar Milosavljevic
- BRL core
  - Matt Roth, Kevin Riehle
- Genboree programming team
  - Andrew Jackson, Sameer Paithankar, Sriram Raghuram,

Questions ?

Comments ?

Suggestions ?

amilosav@bcm.edu

coarfa@bcm.edu