Introduction to the 6th Genboree Epigenome Informatics Workshop

Bioinformatics Research Laboratory

March 4-5, 2013
Houston, Texas

Workshop Objective:

Catalyze Conversion of Epigenomic Profiling Data into Biological Insights through Integrative Analysis

- Introduction to Workshop, Epigenome Informatics, Genboree
- Methods (Assays, Data Processing)
- Standards (Metadata, Interoperability)
- Data Resources (Human Epigenome Atlas, ENCODE)
- Tools (Epigenomic Toolset, Genboree Workbench, WashU Browser)
- Use Cases
- Collaborative Opportunities / Networking / Exchange of Experience

Monday, Mar 4th, 2013

8:00 – 8:45 am Badge pickup & continental breakfast (8th floor, Hilton)

Session 1

8:45 – 9:00 am Introduction to the Workshop – Matt Roth

9:00 – 9:45 am *Introduction to Epigenome Analysis & Genboree – Aleks Milosavljevic*

Session 2

9:45 – 10:45 am *Use case preparation – Setting up projects, databases, groups, accessing files, user privileges, navigating Genboree, toolsets, submitting jobs, etc. (BRL staff)*

10:45 – 11:00 am Break

11:00 – 1:00 pm Hands-on Case Studies: Epigenomic variation between tissues, individuals, and in Cancer (BRL staff)

1:00 – 2:00 pm Lunch

Monday, Mar 4th, 2013

Session 3

2:00 – 3:00 pm	Analysis of Infinium Methylation Arrays (BRL staff)
3:00 – 3:45 pm	Visualizing Human Epigenomic Data via WashU Genome Browser (Xin Zhou, Wash U)
4:00 – 5:30 pm	Track A: Analysis of individual data (for those who uploaded data), or continuation of use cases (attendees & BRL staff)
4:00 – 5:30	Track B: Discussion of the "programmable web", data and tool integration, and REST APIs (Aleks Milosavljevic)
6:00 pm	**Depart for dinner together (place TBD) *or* dinner on your own**

Tuesday, Mar 5th, 2013

8:30 – 9:00 am	Continental breakfast (outside meeting room)

Session 4

9:00 – 9:15 am	Review of Day 1 and preview of Day 2 – Matt Roth

Tuesday, Mar 5th, 2013

9:15 – 10:00 am	*Quantitative profiling of histone modifications, peak calling and segmentation of epigenomic signals, Chip-Seq, RNA-Seq (BRL staff)*
10:00 – 12:00 pm	Hands-on Case Study: Chip-Seq & RNA-Seq analysis (attendees & BRL staff)
12:00 – 1:00 pm	**Boxed lunch (outside auditorium)**

Session 5

1:00 – 3:00 pm	Track A: Analysis of individual data (for those who uploaded data), or continuation of use cases (*attendees & BRL staff*)
1:00 – 3:00 pm	Track B: Discussion of the "programmable web", data and tool integration, and REST APIs (Aleks Milosavljevic)
3:00 – 4:30 pm	Wrap-up data analysis: semi-structured time for completing individualized data analysis, case studies, informal discussions (attendees & BRL staff)
4:30 – 5:00 pm	**Open discussion and wrap-up, adjourn**

BRL Hosted Genboree Epigenome Informatics Workshops (on-site - Houston, TX)

March, 2012: Genboree end-user focused workshop, introduction to epigenome data analysis
-people wanted more hands-on with tools

May, 2012: Launch of RNA-Seq & Chip-Seq tools

October, 2012: Attendees could upload own data. Introduced Spark (epigenomic data visualization, *Nielsen et al.* Genome Research) & "programmable web" session

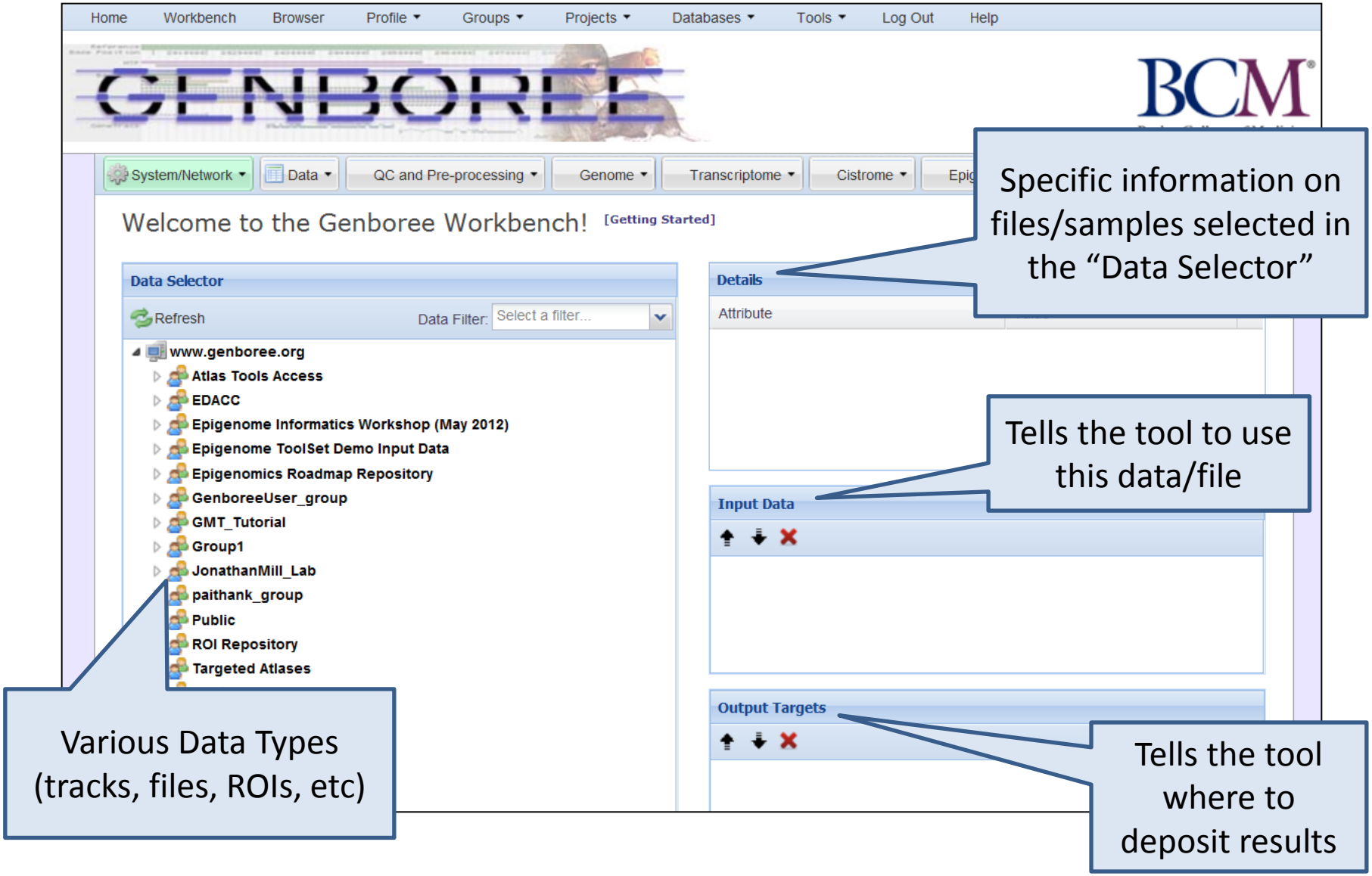
March, 2013: Attendees could upload own data. Expanded Spark demo, WashU demo

Workshop Participants:

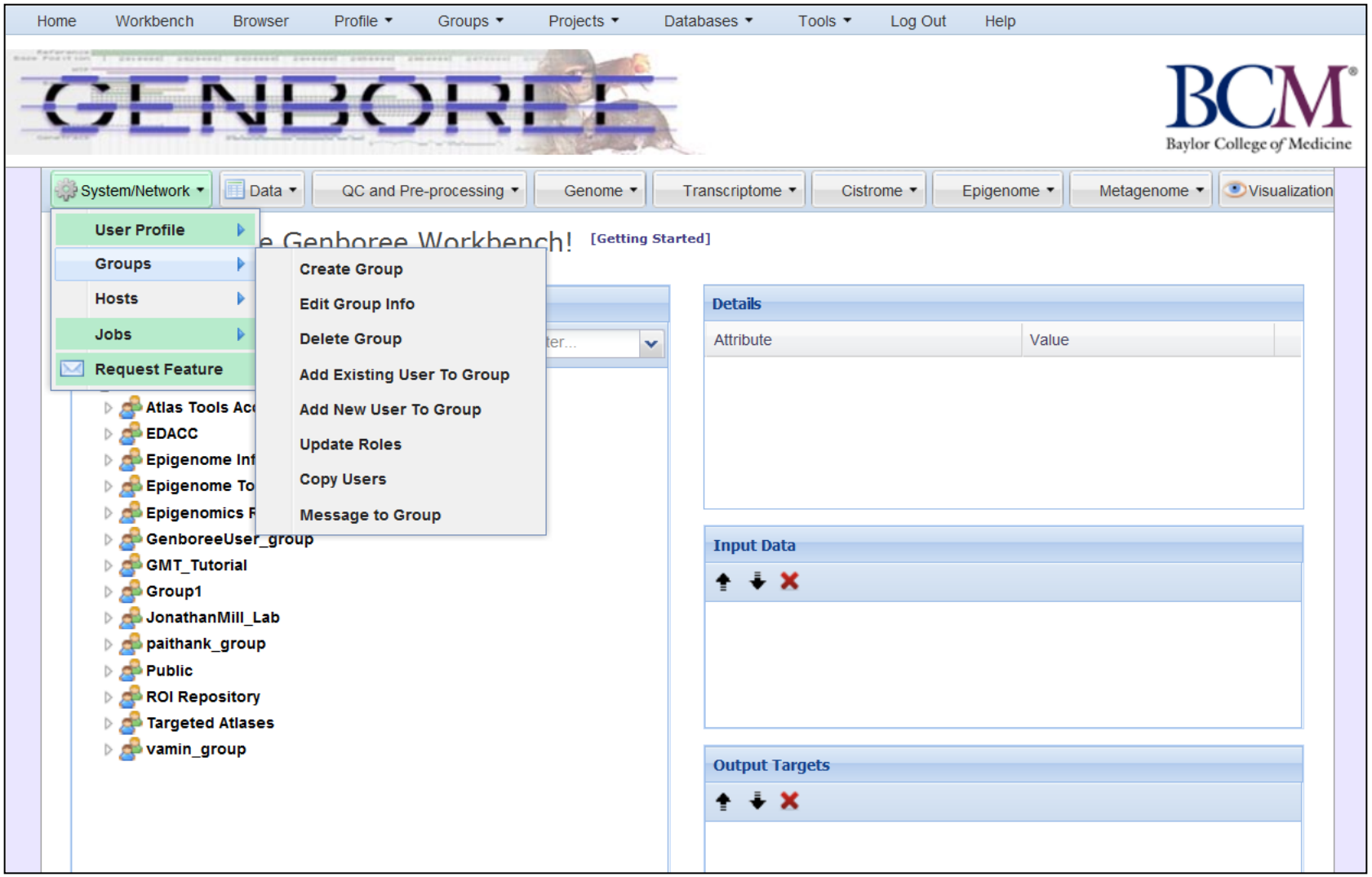
Brief introductions to facilitate networking;

Please describe your goals in attending workshop

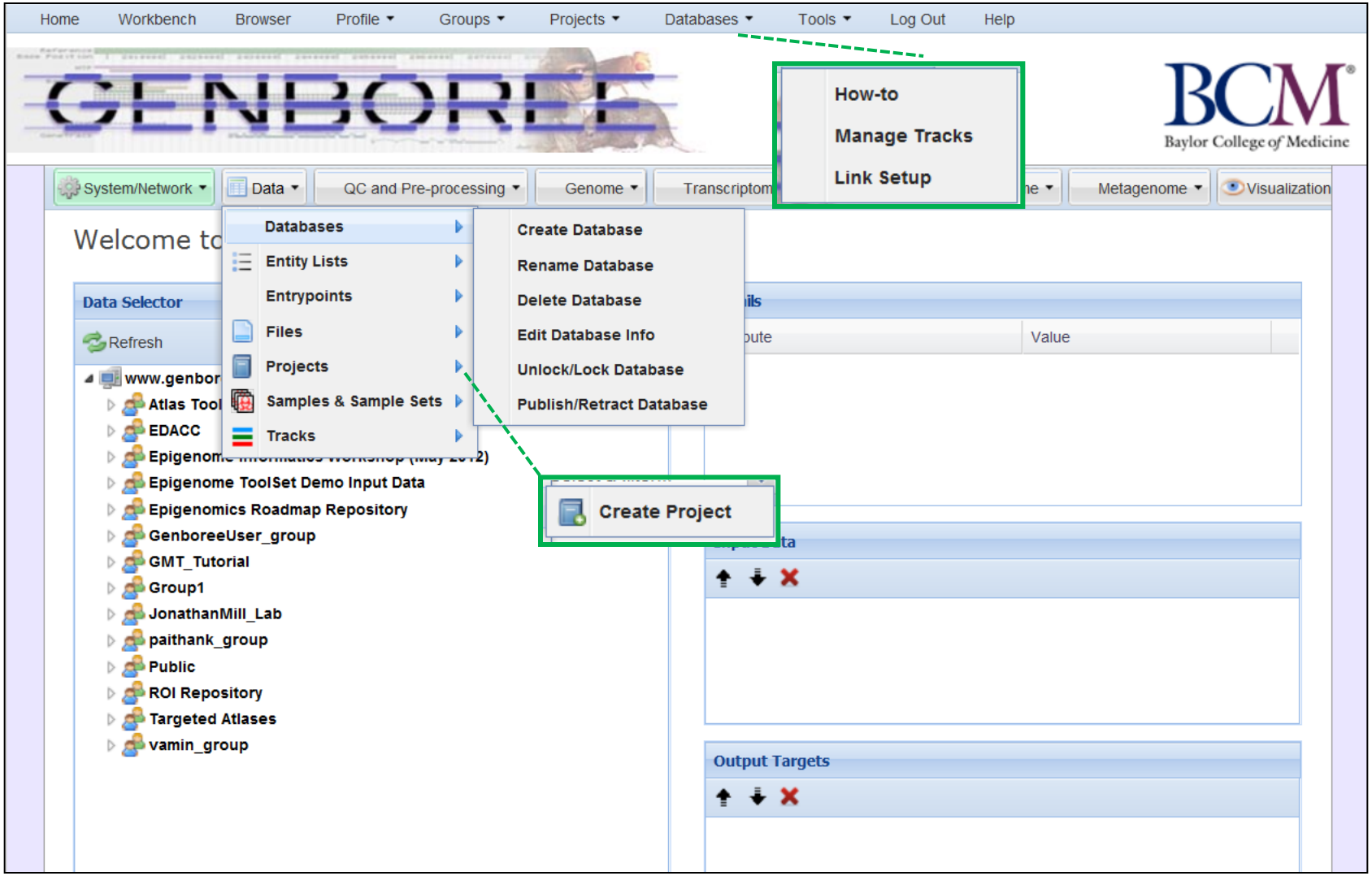
Welcome to The Genboree Workbench!



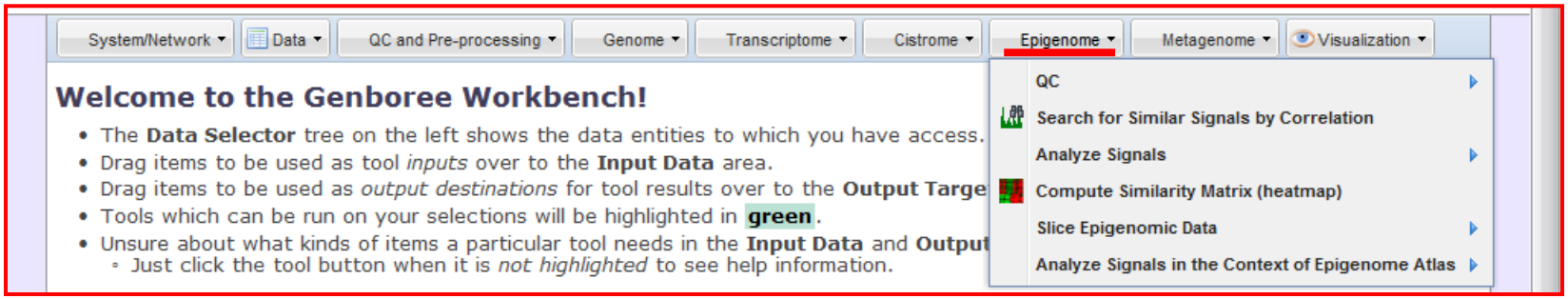
Create & Manage Groups

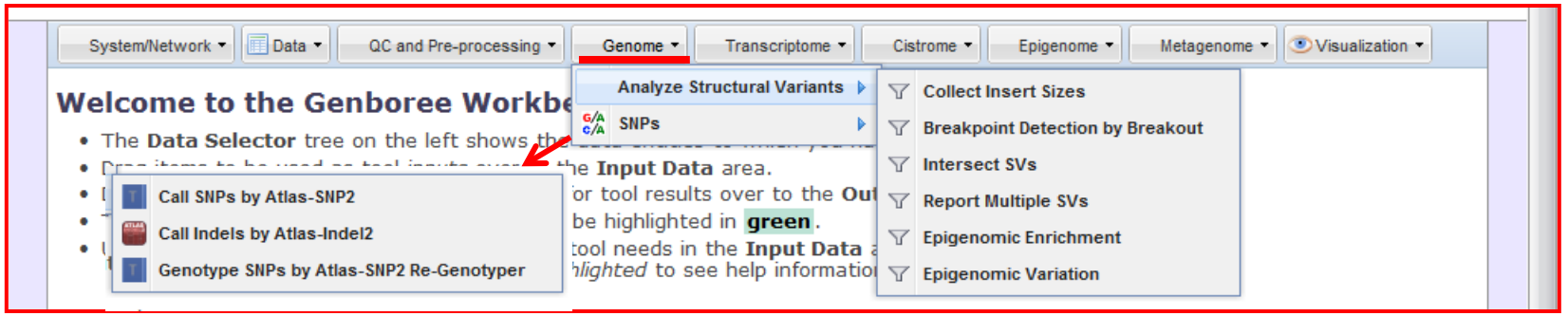


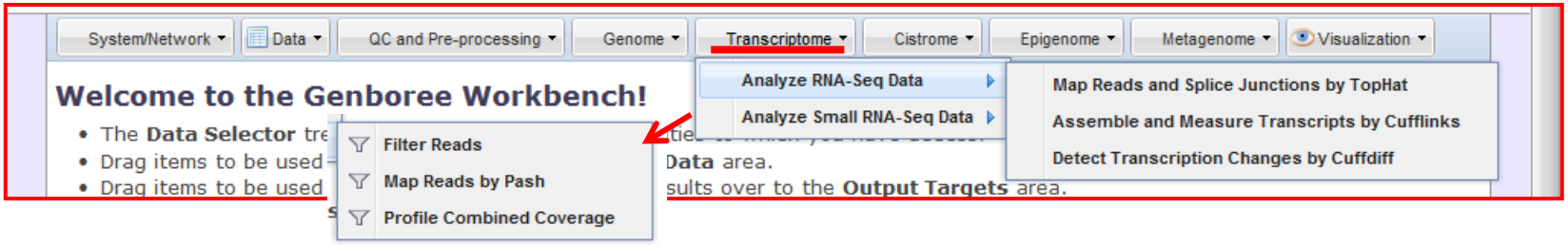
Create & Manage Databases & Projects

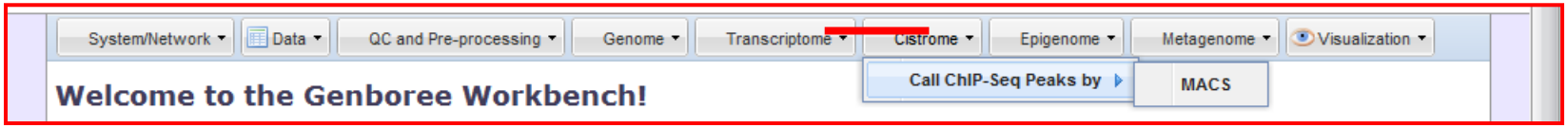


Welcome to the Genboree Workbench!



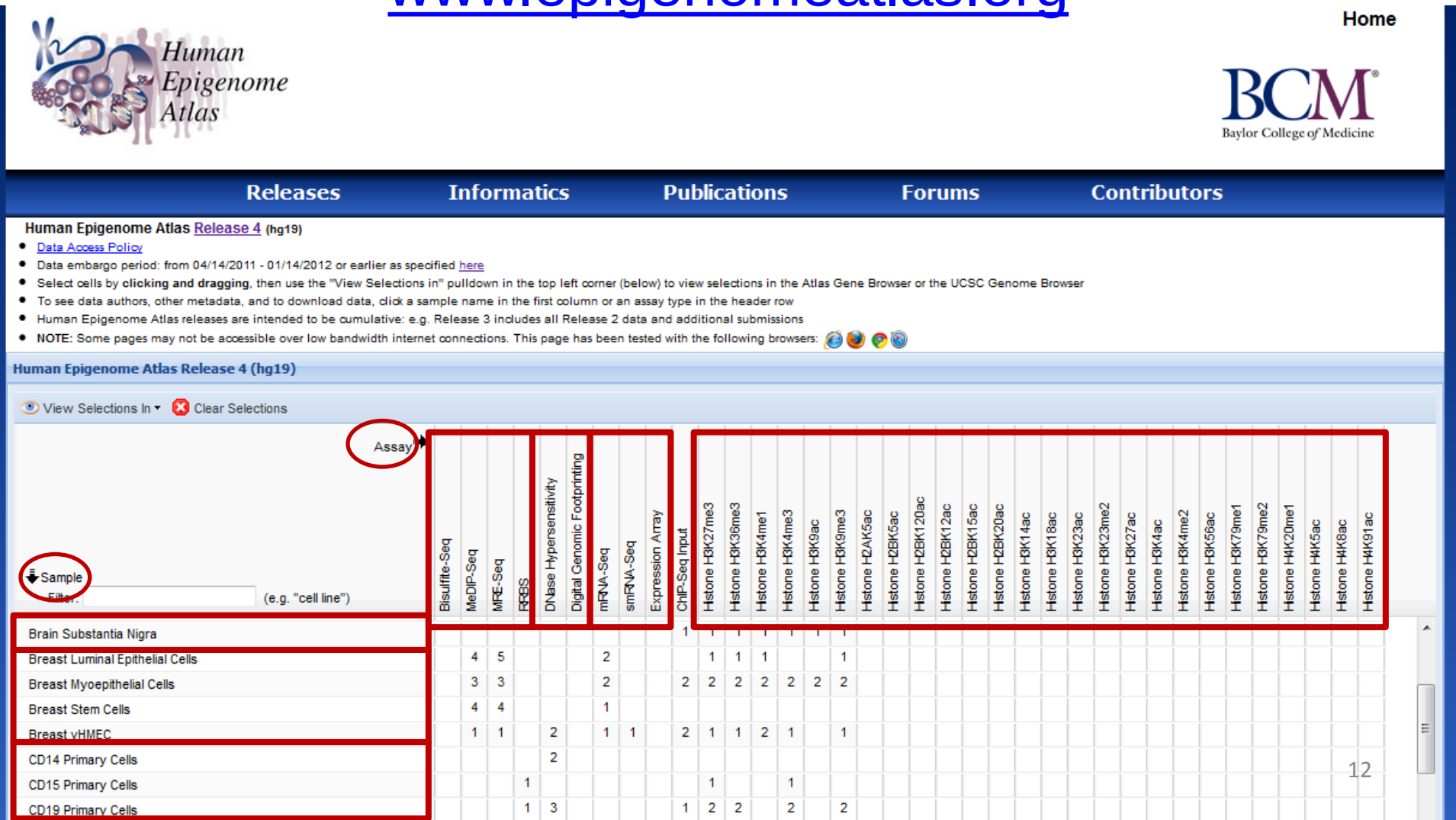




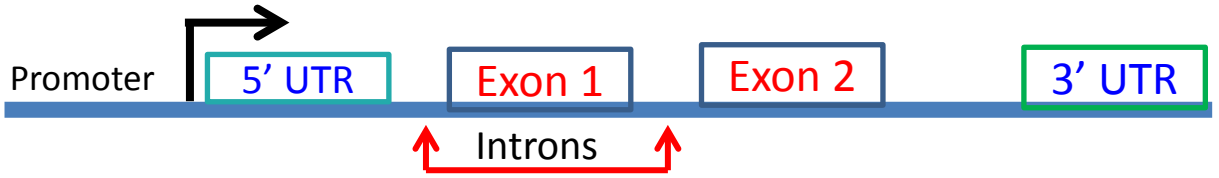


Epigenome Atlas Release 8
over 2000 experiments

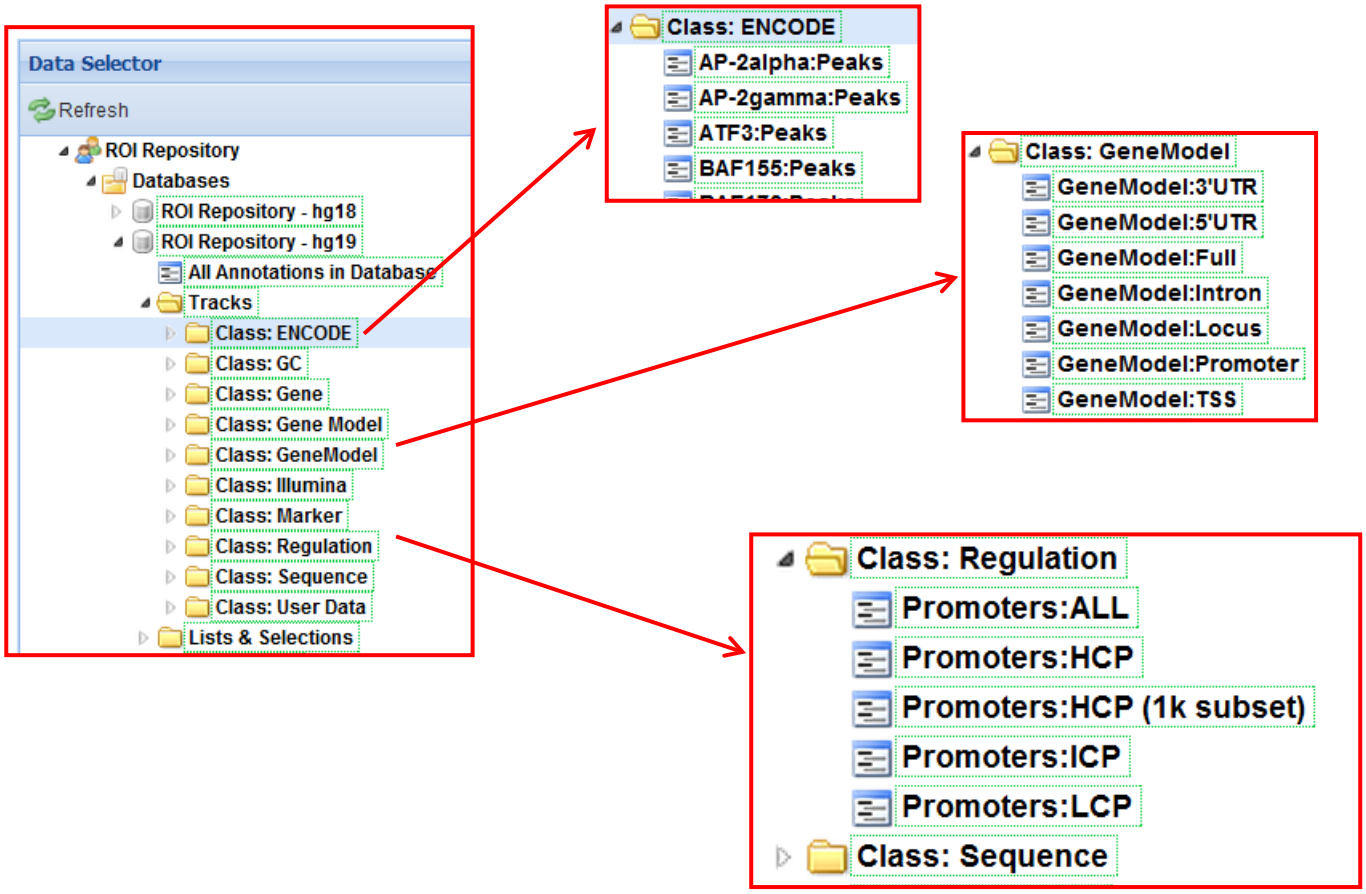
www.epigenomeatlas.org



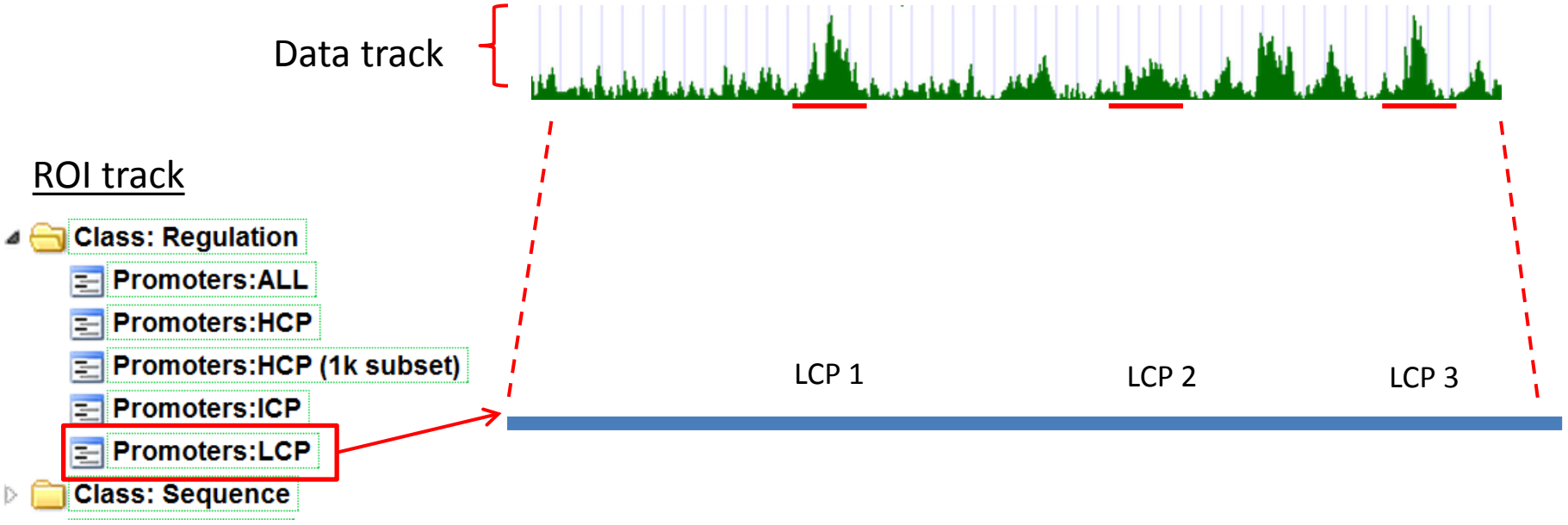
Biology Across Key Genetic Elements (promoters, exons, UTR, etc): Many ROI (i.e. annotation) Tracks Available



ROI tracks available

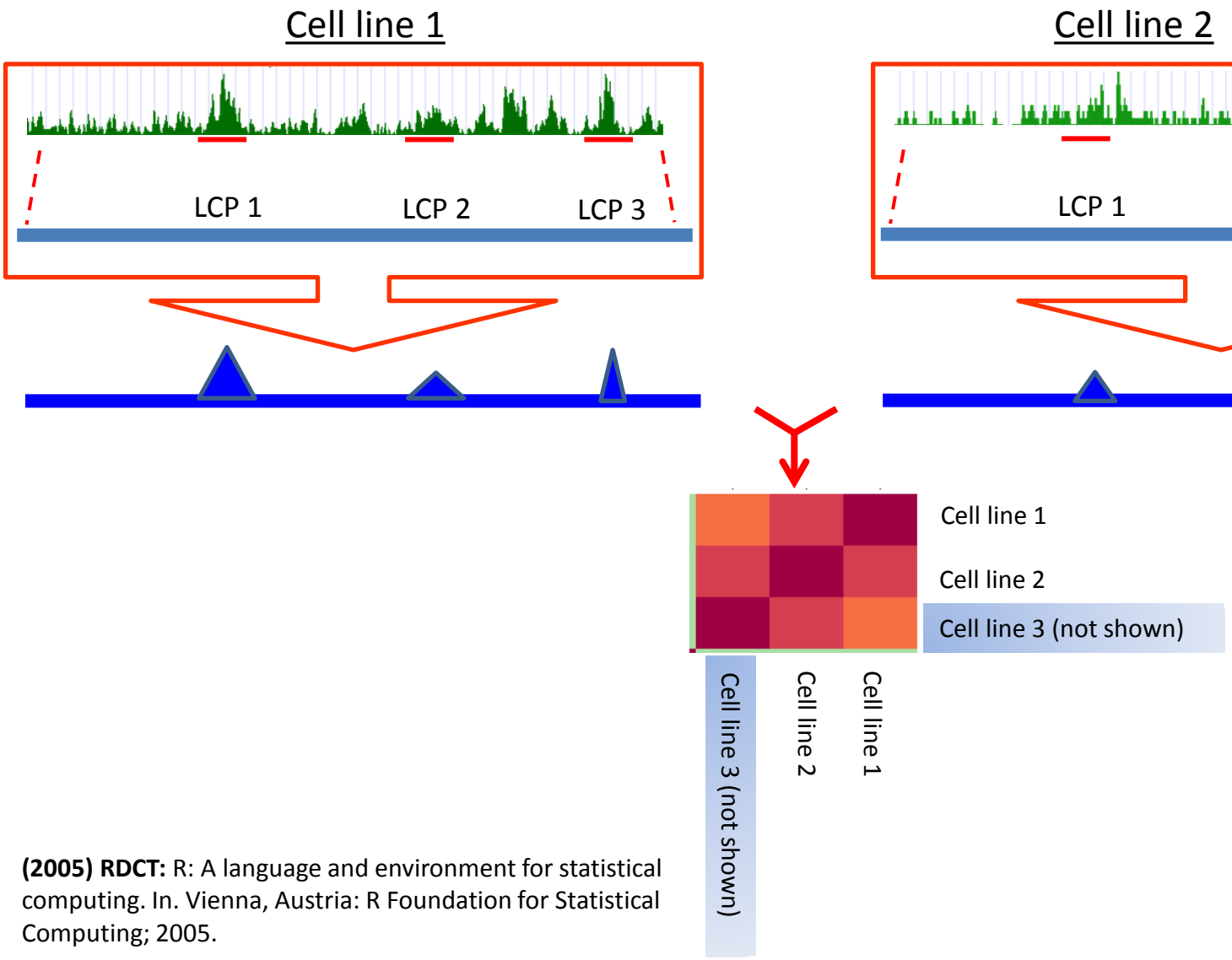


Data Tracks from Epigenomic Experiments Projected On To ROI (i.e. annotation) Tracks



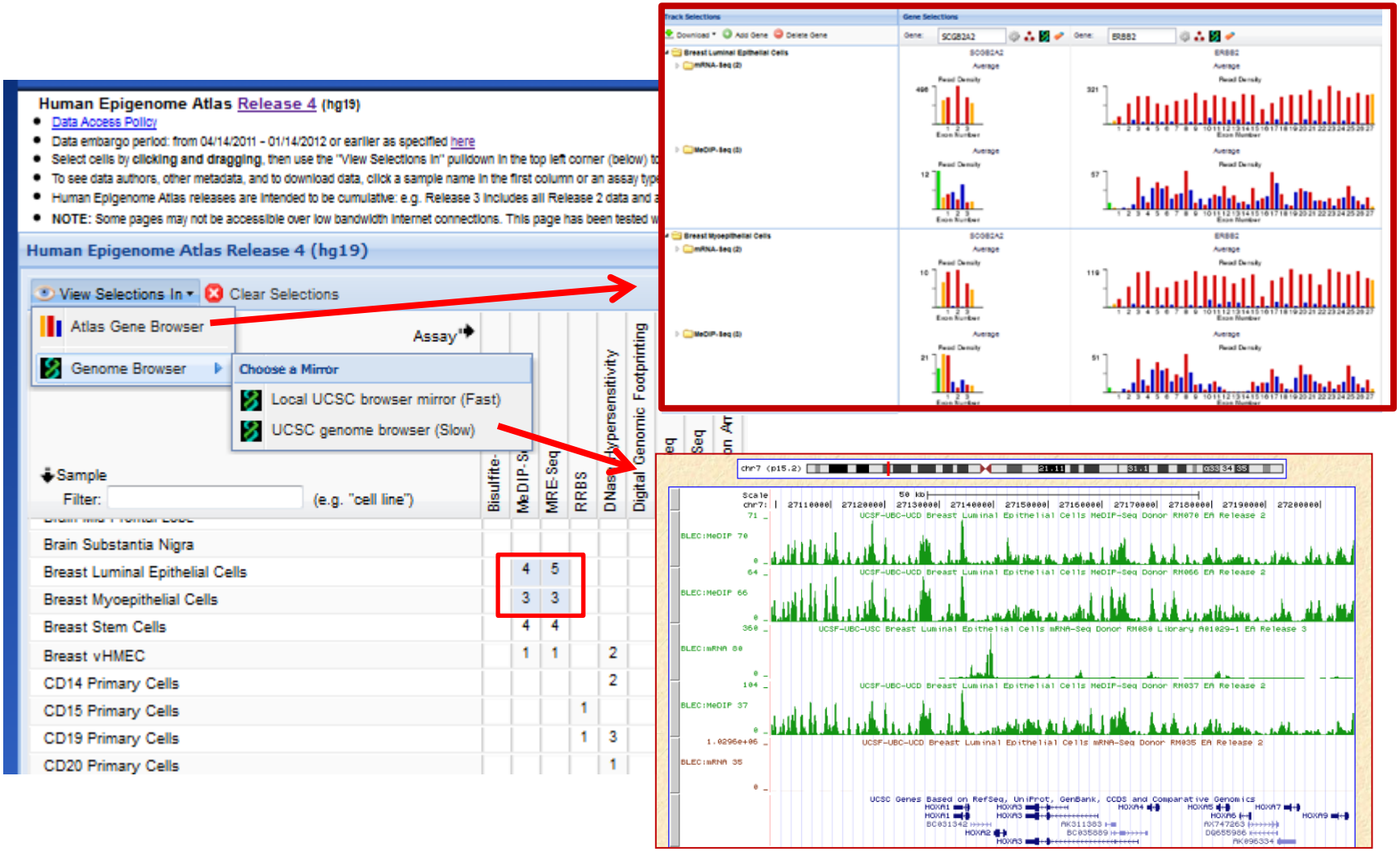
Promoters: LCP = Low-CpG promoters (as defined in Weber et al., *Nature Genetics* (2007)

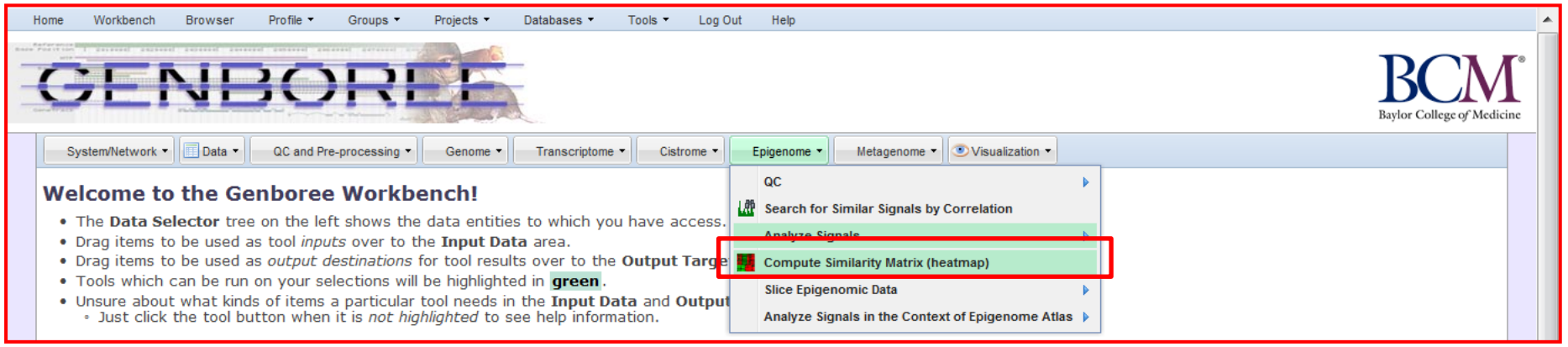
Compute Pearson Correlation Coefficient Between Experiments: Similarity Matrix is Output as Heatmap



(2005) RDCT: R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing; 2005.

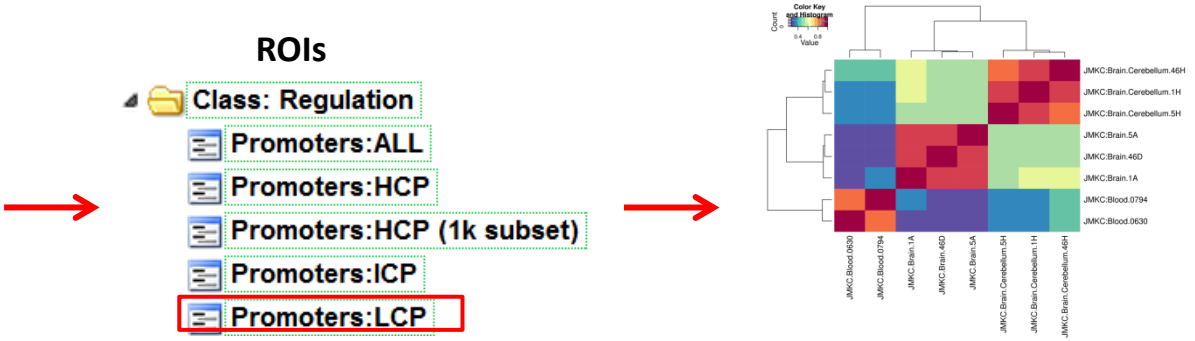
Viewing selections





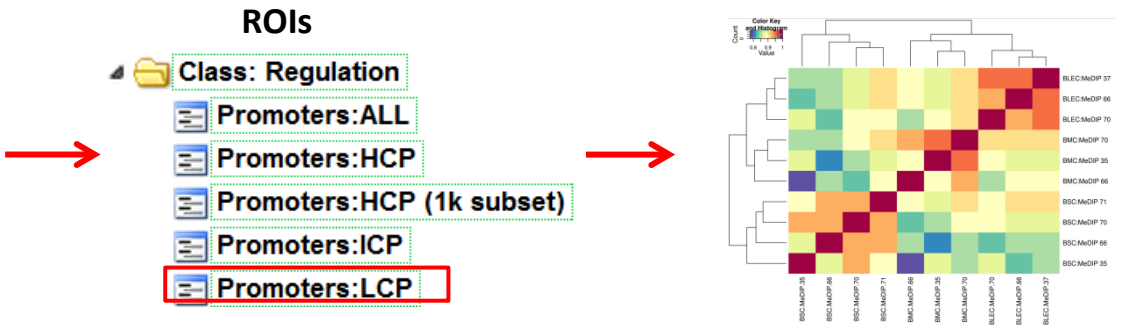
Use Case 1: Genomewide Patterns of Methylation can Distinguish Between Blood, Cerebellum, and Cortex

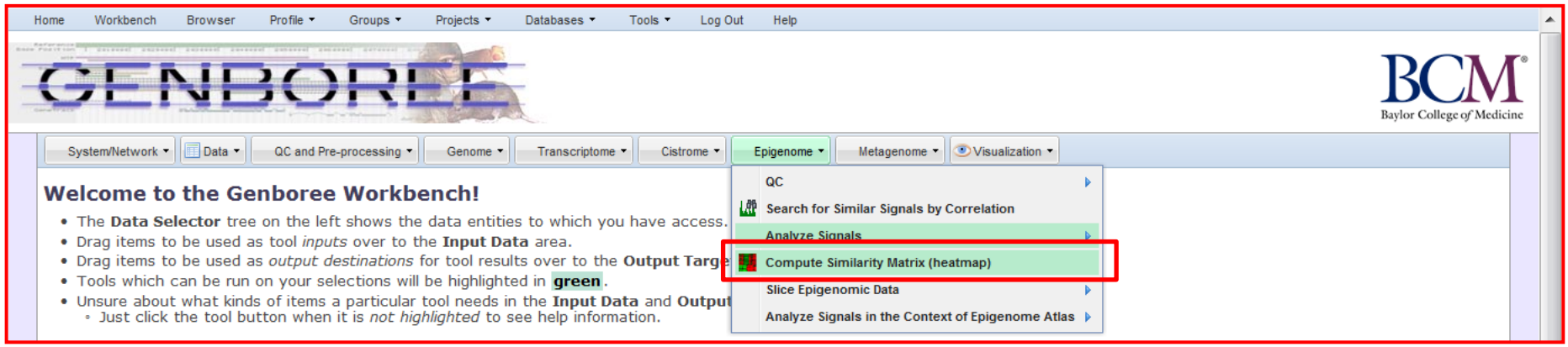
Epigenomic Tracks:
-Blood
-Cerebellum
-Cortex



Use Case 2: Breast Cell Types Cluster Based on Their MeDIP-seq Profiles (Epigenome Atlas and UCSF REMC data)

Epigenomic Tracks:
-Breast Luminal Epithelium
-Breast Myoepithelial
-Breast Stem Cell

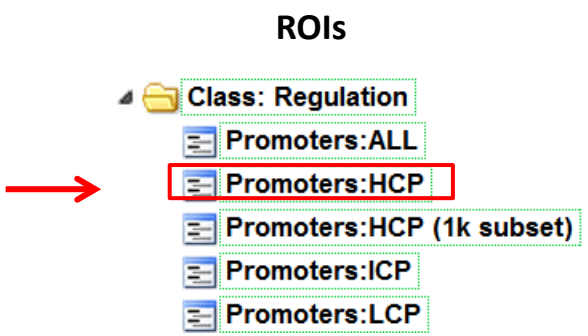


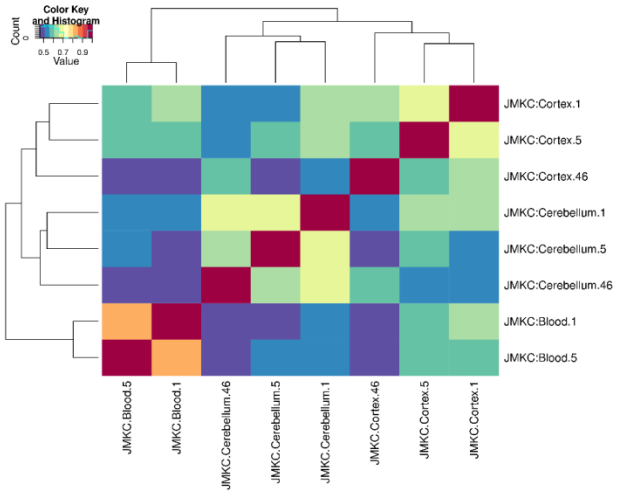


Use Case 5: Methylation of some features discriminate tissue type better than others

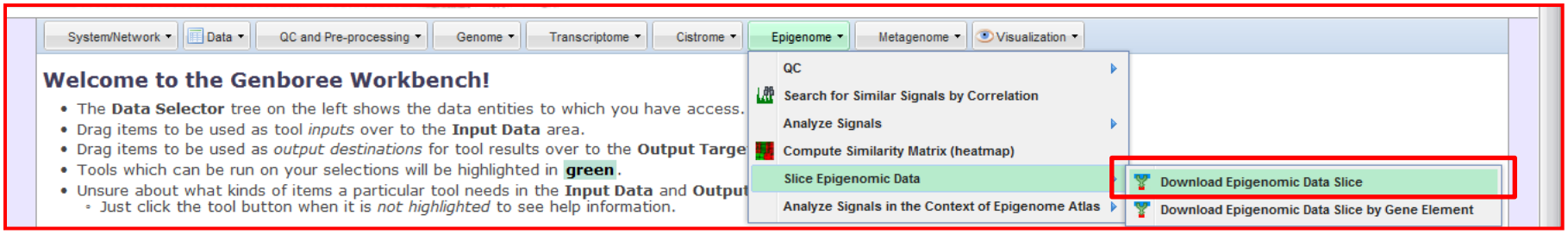
Similar to Use Case 1 & 2 but uses different ROIs to illustrate how different features produce different similarity matrices (heatmaps).

Epigenomic Tracks
-Blood
-Cerebellum
-Cortex





Use Case 9: Coordinated Changes of Epigenomic Marks Across Tissue Types



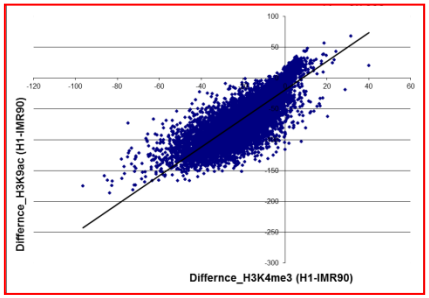
Epigenomic Tracks:
-H1 cell line
-IMR90 cell line

Collates score tracks into one data matrix, export to Excel

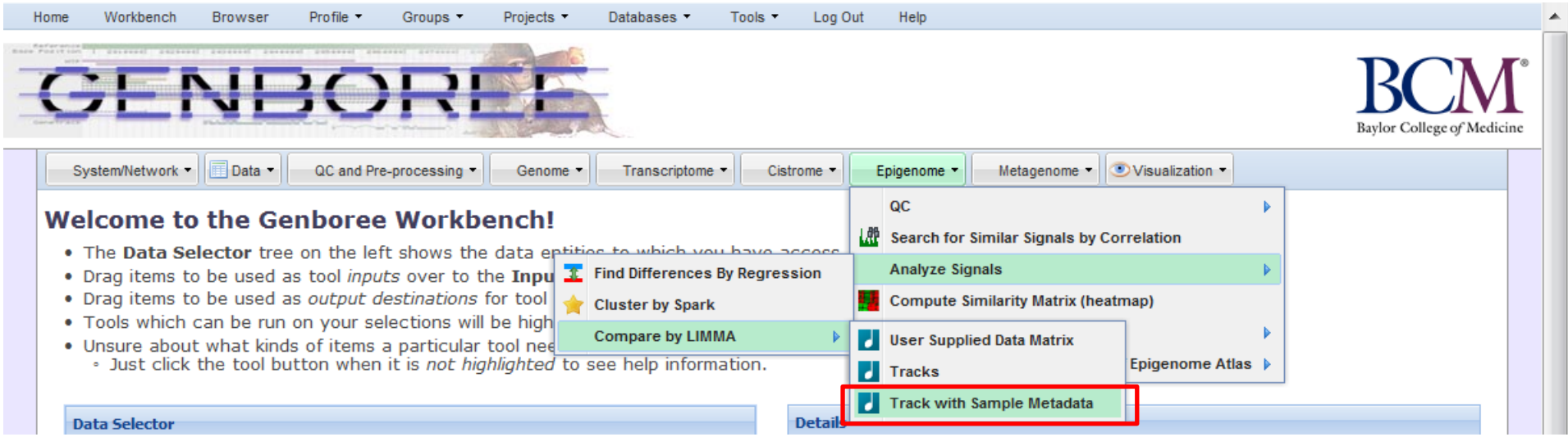
	Bisulfite data			H1.H3K9ac					IMR90.H3K9ac			H1.H3K4me3					
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Index	H1.BS.Con	IMR90.BS.Combined		H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac.68		IMR90.H3K	IMR90.H3K9ac.46		H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3
2	HSAP04065	0	0.044444		0.314286	0.314286	0.314286	0	0.314286		0.314286	0.314286		0.628571	0.628571	0.795714	0.314286	0.314286	
3	HSAP04065	0.047353	0.034789		4.51304	3.29043	2.35826	2.34435	2.73913		17.9513	16.5722		8.52	28.9374	23.92	11.84	2.3113	12.7
4	HSAP04065	0.208431	0.215174		5.79688	8.58438	5.85313	3.75312	8.56875		15.5813	22.7719		17.1375	39.5844	22.9281	14.525	5.03125	22.9
5	HSAP04065	0.209214	0.212334		1.07769	2.62314	1.87107	0.581818	1.20496		8.35537	9.14876		5.90248	14.443	27.2893	3.07769	0.363636	9.16

Column headers = experiments
Rows = ROIs

Scatter plots



Use Case 12: Assess breast cancer cell type of origin



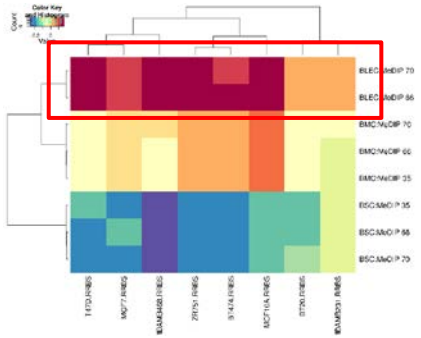
LIMMA: Smyth, G. K. Statistical Applications in Genetics and Molecular Biology (2005)

<u>"Your" Epigenomic Tracks (RRBS):</u>

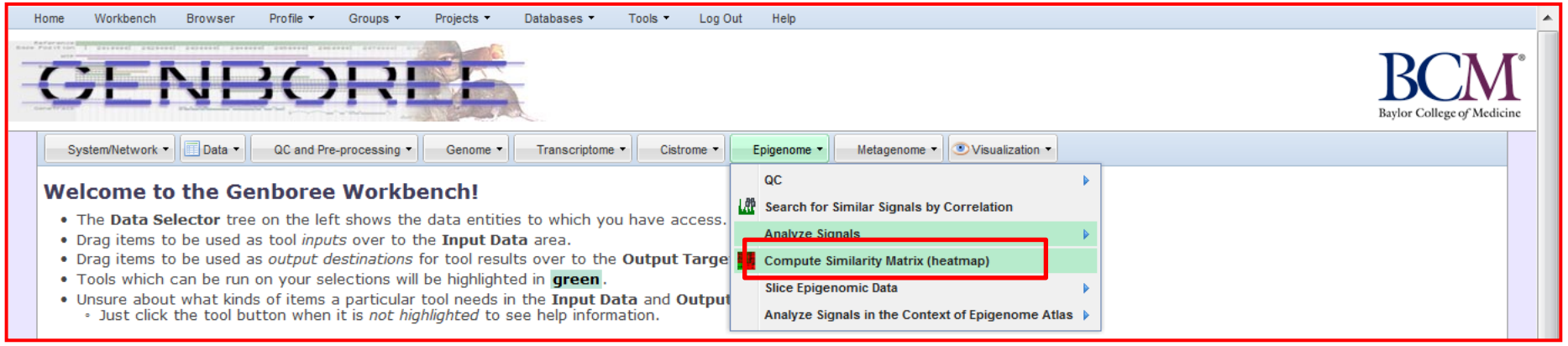
-Breast Luminal Epithelium
-Breast Myoepithelial
-Breast Stem Cell

<u>Public Epigenomic Tracks (MeDIP):</u>

-Breast Luminal Epithelium
-Breast Myoepithelial
-Breast Stem Cell



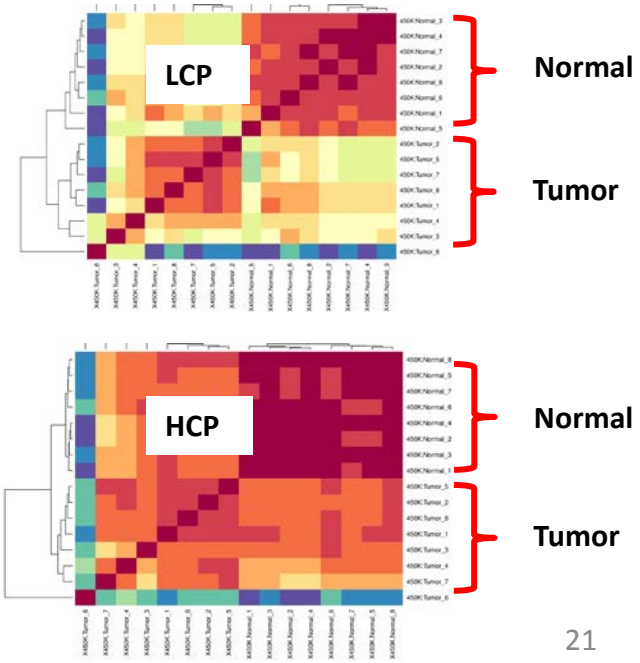
Use Case 13: Analysis of epigenomic variation in breast tumors (450K)



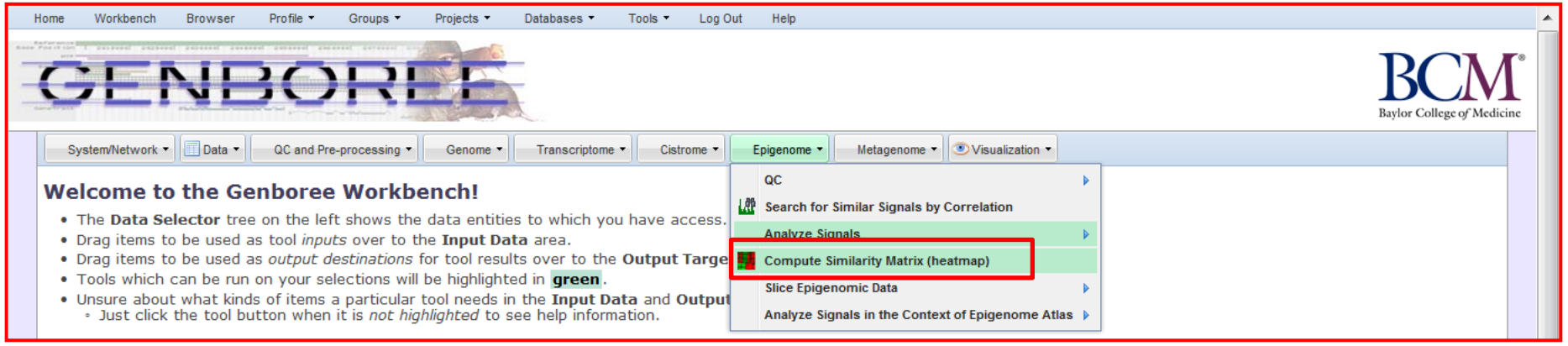
Use Case 13a: Cluster all 16 breast tissue samples

16 450 K Samples (Dedeurwaerder, S.et al. (2011)
-8 normal breast samples
-8 cancerous breast samples

ROIs (HCP vs LCP)



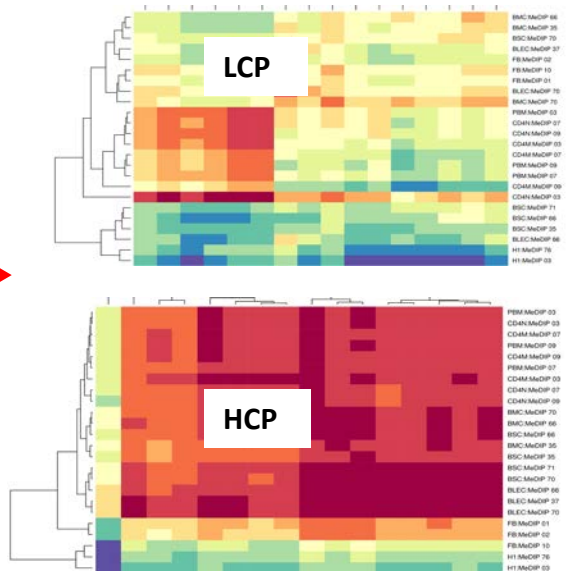
Use Case 13: Analysis of epigenomic variation in breast tumors



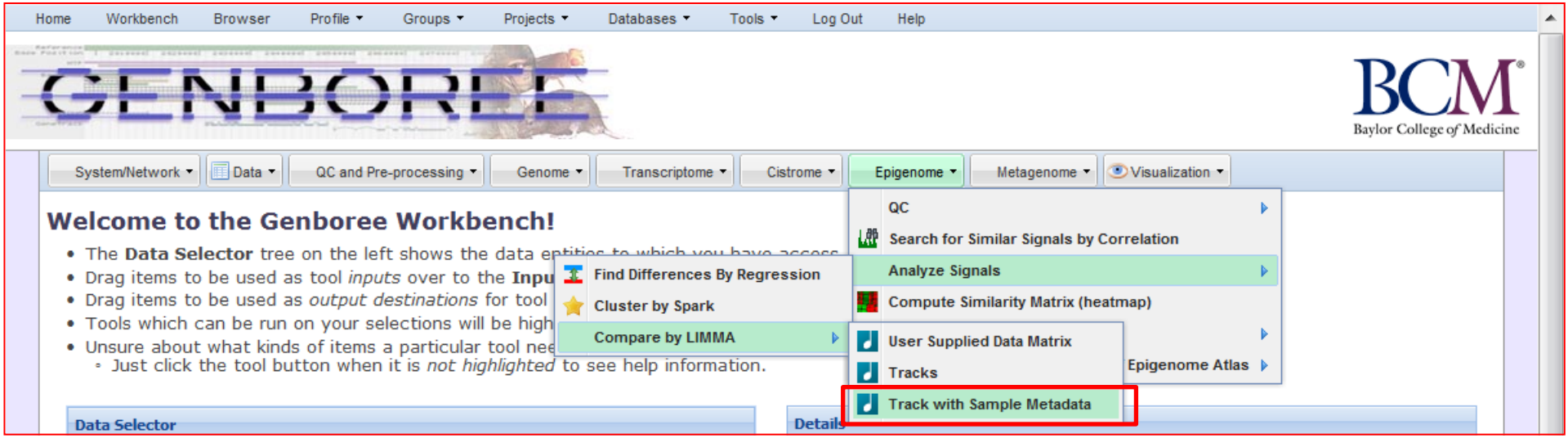
Use Case 13b: Compare 450K profiles (8 tumor, 8 normal) *against reference epigenomes* from the Epigenome Atlas

16 450 K Samples (Dedeurwaerder, S.et al. (2011)
-8 normal breast samples
-8 cancerous breast samples

ROIs



Use Case 13: Analysis of epigenomic variation in breast tumors



Use Case 13c: Since most breast tumor samples appear to contain excess of blood & immune cells, comparison of normal and tumor tissue may reveal differentially methylated genes (and corresponding pathways). Identify differentially methylated probes, genes, and pathways using LIMMA & online resources

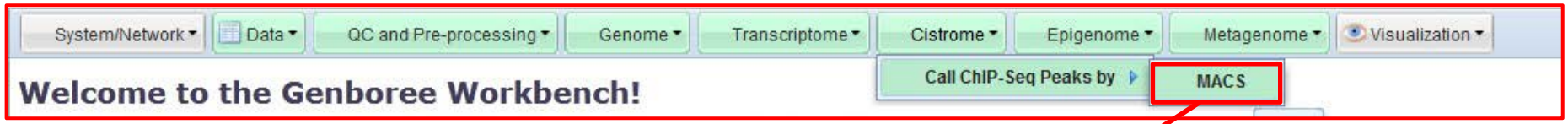
16 450 K Samples (Dedeurwaerder)
-8 normal breast samples
-8 cancerous breast samples

ROIs

Gene List

	B
1	PAQR7
2	VGLL4
3	TESK1
4	FGFR1
5	UCKL1
6	MGC1612
7	LCAT
8	SLC12A4
9	MTSS1L

Use Case 14: Chip-Seq Data Analysis



Zhang et al, *Genome Biology* (2008)

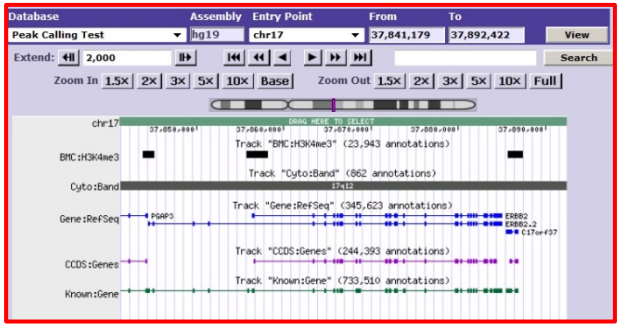
BED files

MACS results (file in Genboree)

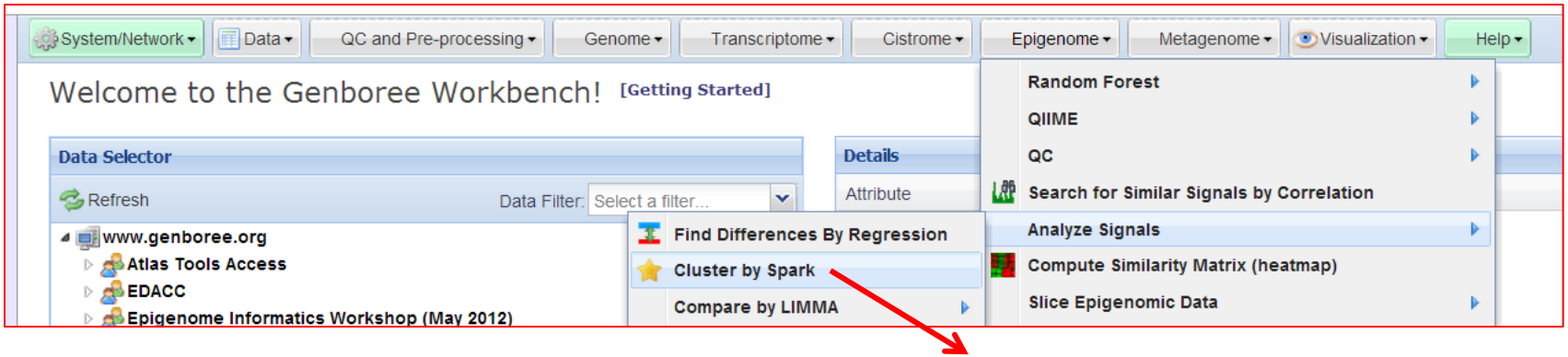
	chr	start	end	length	summit	tags	-10*LOG10	fold_enrichment
20	chr	start	end	length	summit	tags	-10*LOG1	fold_enrichment
21	chr1	9861	10677	817	360	292	2034.63	24.91
22	chr1	713307	715489	2183	1297	83	250.89	11.7
23	chr1	724760	727160	2401	2202	92	223.86	16.37
24	chr1	761468	763266	1799	865	133	463.34	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							7.95
29	chr1							2.67
30	chr1							4.84
31	chr1	911145	912096	952	598	25	123.98	13.95

H3K4me3, H3K4me1

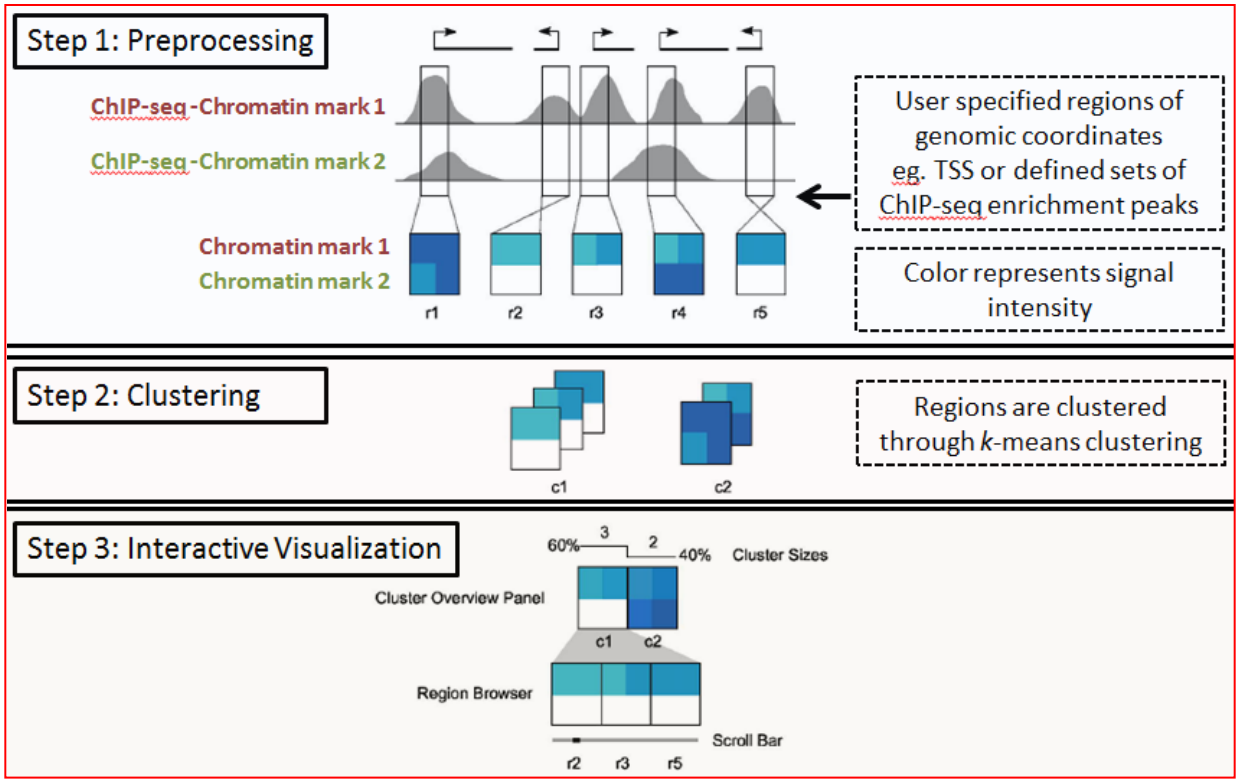
Visualize in Genboree



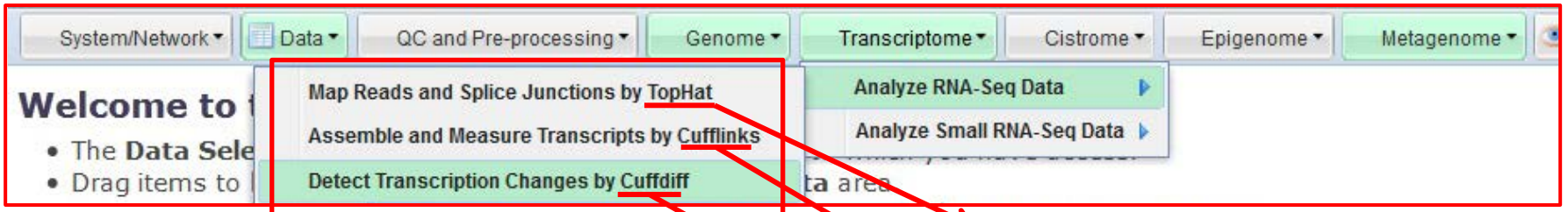
Use Case 15: Spark Analysis



Nielsen et al, *Genome Biology* (2012)



Use Case 16: RNA-Seq Cuffdiff Data Analysis
(Use Case 17: Cufflinks & TopHat)



Trapnell et al, *Bioinformatics* (2009)

Trapnell et al, *Nature Biotech* (2010)

Trapnell et al, *Nature Biotech* (2010)

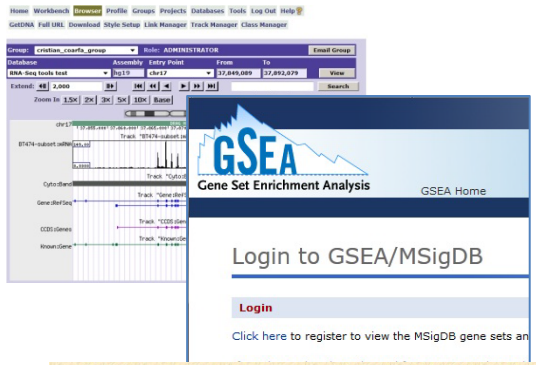
FASTQ, BAM files →

Gene expression diffs (file in Genboree)

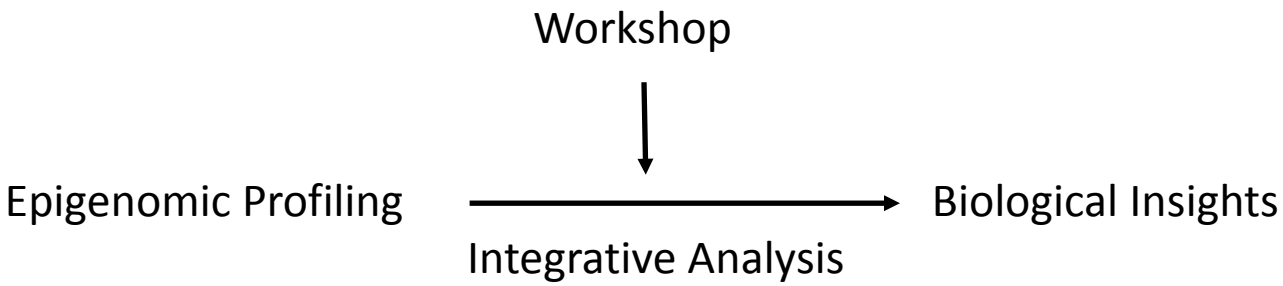
	A	B	C	D	E	F	G
1	test_id	gene_Nar	gene_id	gene	locus	sample_	sample_
5	NM_0000:	ACADS	NM_0000:	-	chr12:121:	Luminal	BasalA
10	NM_0000:	ADA	NM_0000:	-	chr20:432-	Luminal	BasalA
32	NM_0000-	AR	NM_0000-	-	chrX:6676:	Luminal	BasalA
41	NM_0000:	ATP7B	NM_0000:	-	chr13:525(	Luminal	BasalA
51	NM_0000(	C3	NM_0000(	-	chr19:667:	Luminal	BasalA
88	NM_0001(	CYBA	NM_0001(	-	chr16:887(	Luminal	BasalA
91	NM_0001(	CYP1B1	NM_0001(	-	chr2:3829-	Luminal	BasalA
195	NM_0002:	ITGA6	NM_0002:	-	chr2:1732:	Luminal	BasalA
199	NM_0002:	JAG1	NM_0002:	-	chr20:106:	Luminal	BasalA
254	NM_0002:	NPC1	NM_0002:	-	chr18:210:	Luminal	BasalA
290	NM_0003(	CTSA	NM_0003(	-	chr20:445:	Luminal	BasalA
327	NM_0003-	SOX9	NM_0003-	-	chr17:701:	Luminal	BasalA

→

Visualization/pathway analysis



Workshop Evaluation (link)



- Introduction to Workshop, Epigenome Informatics, Genboree
- Methods (Assays, Data Processing)
- Standards (Metadata, Interoperability)
- Data Resources (Human Epigenome Atlas)
- Tools (Epigenomic Toolset, Genboree Workbench)
- Use Cases / Case Studies
- Collaborative Opportunities / Networking / Exchange of Experience

Acknowledgments

- BRL
  - Aleksandar Milosavljevic
  - Cristian Coarfa, Alan R Harris, Viren Amin
  - Elke Norwig-Eastaugh
- BRL core
  - Matt Roth, Kevin Riehle, Tim Charnecki
- Genboree programming team
  - Andrew Jackson, Sameer Paithankar, Sriram Raghuram
- Former BRL members
  - Chia-Chin Wu, Arpit Tandon
- WashU
  - Xin Zhou
  - Ting Wang
- 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
  - Sean McGuire, Christine Pichot