Introduction to Epigenome Analysis and Genboree

Aleksandar Milosavljevic
Bioinformatics Research Laboratory (BRL)
Baylor College of Medicine

6th Genboree Workshop on Epigenome Informatics

March 4th, 2013

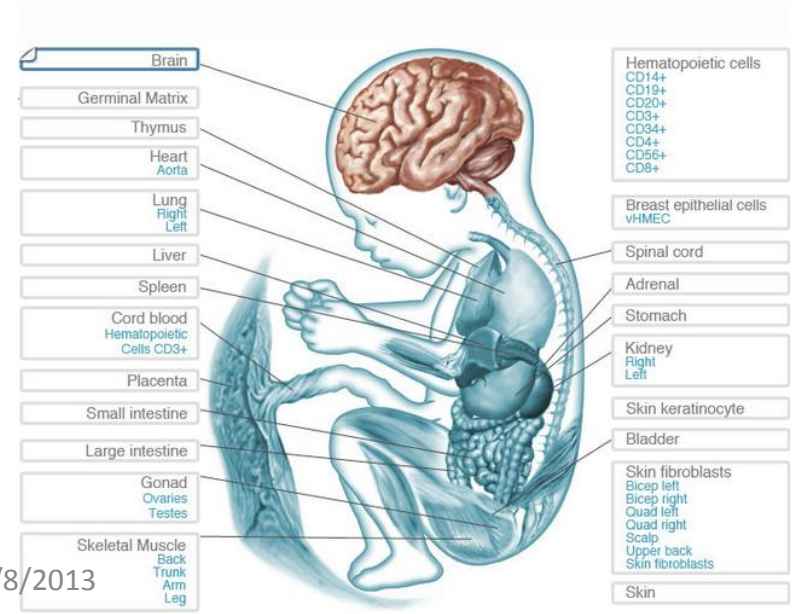
- NIH Roadmap Epigenomics Project
- Epigenome Analysis
- Genboree Workbench
- Genboree Network

- NIH Roadmap Epigenomics Project
- Epigenome Analysis
- Genboree Workbench
- Genboree Network

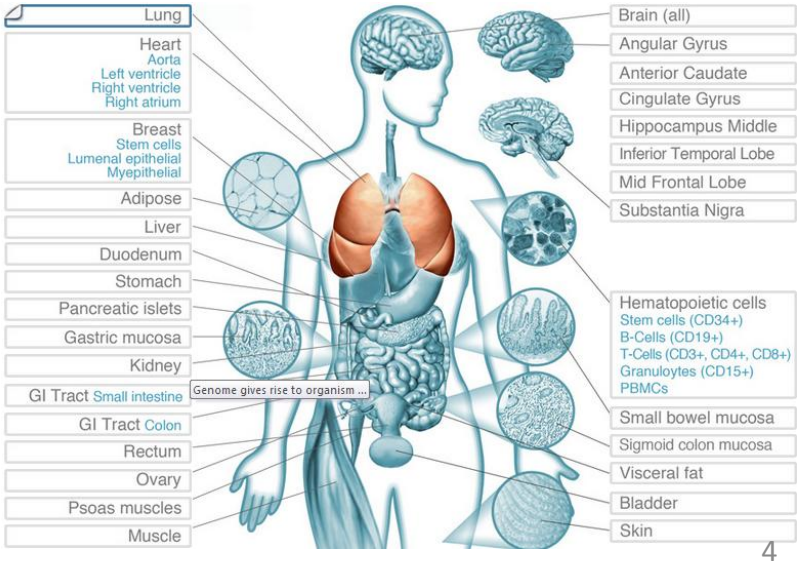
Epigenomic profiling of human cell-lines, primary cells, and tissues

Stem Cells

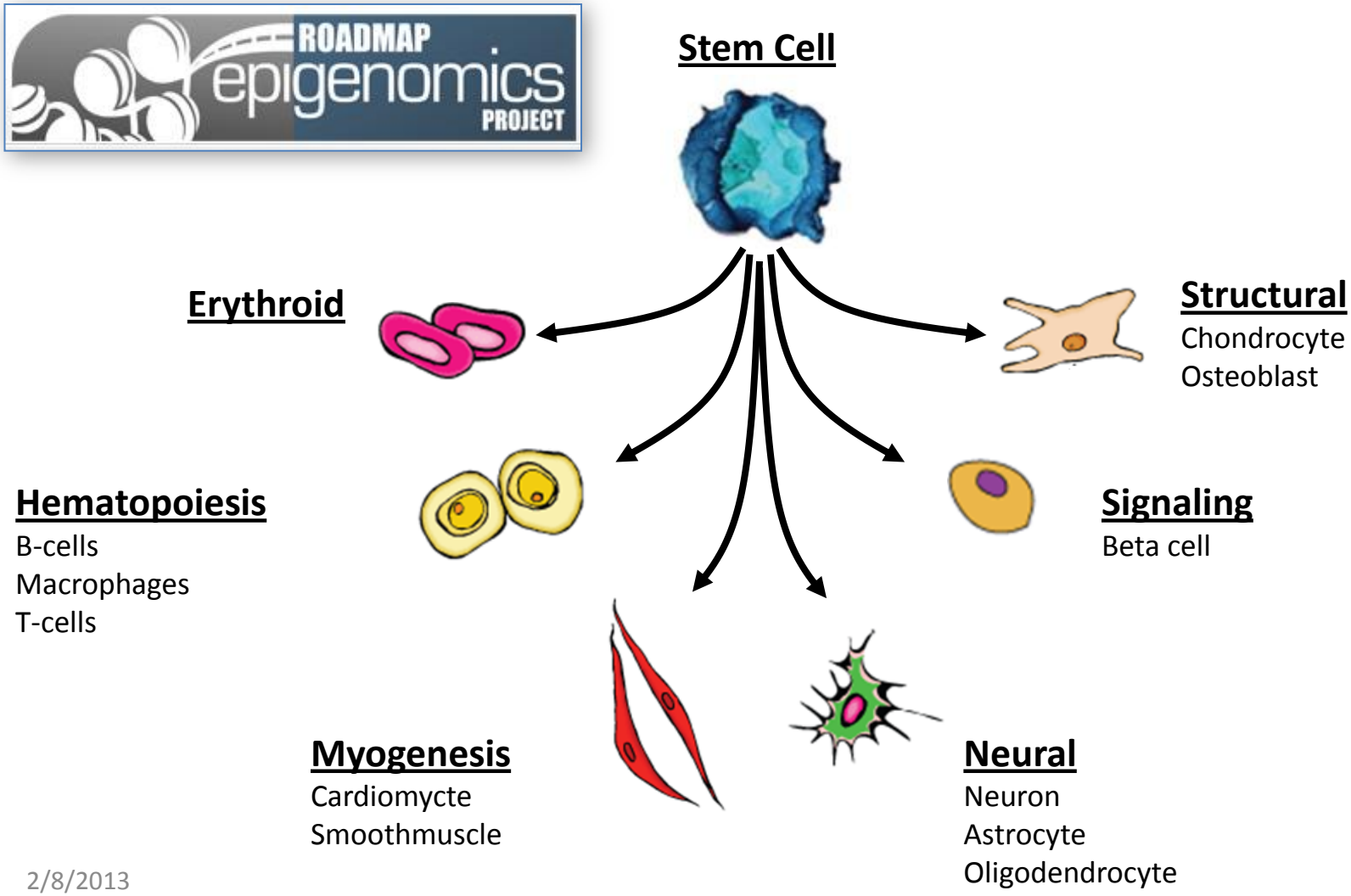
Fetal



Adult



Mapping epigenomic differences between cell types



Epigenomic profiling of human cell-lines, primary cells, and tissues

Methylomes
- Whole-genome bisulfite sequencing
- RRBS
- MeDIP-seq
- MRE-seq

Core histone marks
- H3K4me1
- H3K4me3
- H3K27me3
- H4K36me3
- H3K9me2
- H3K9ac / H3K27ac

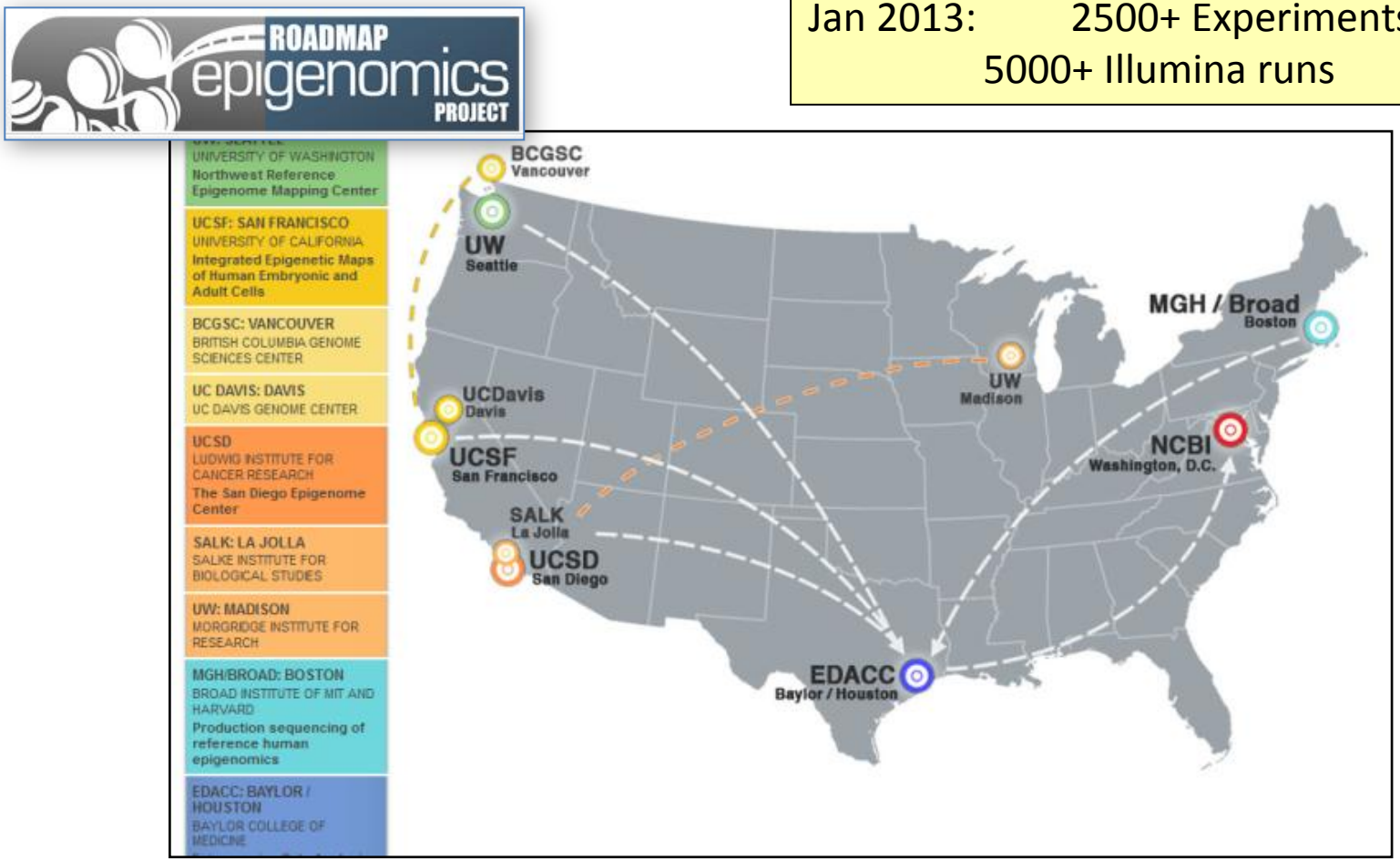
Chromatin accessibility
- Dnase hypersensitivity
- Digital genomic footprinting

Transcriptomes
- RNA-seq
- Expression array
- smallRNA

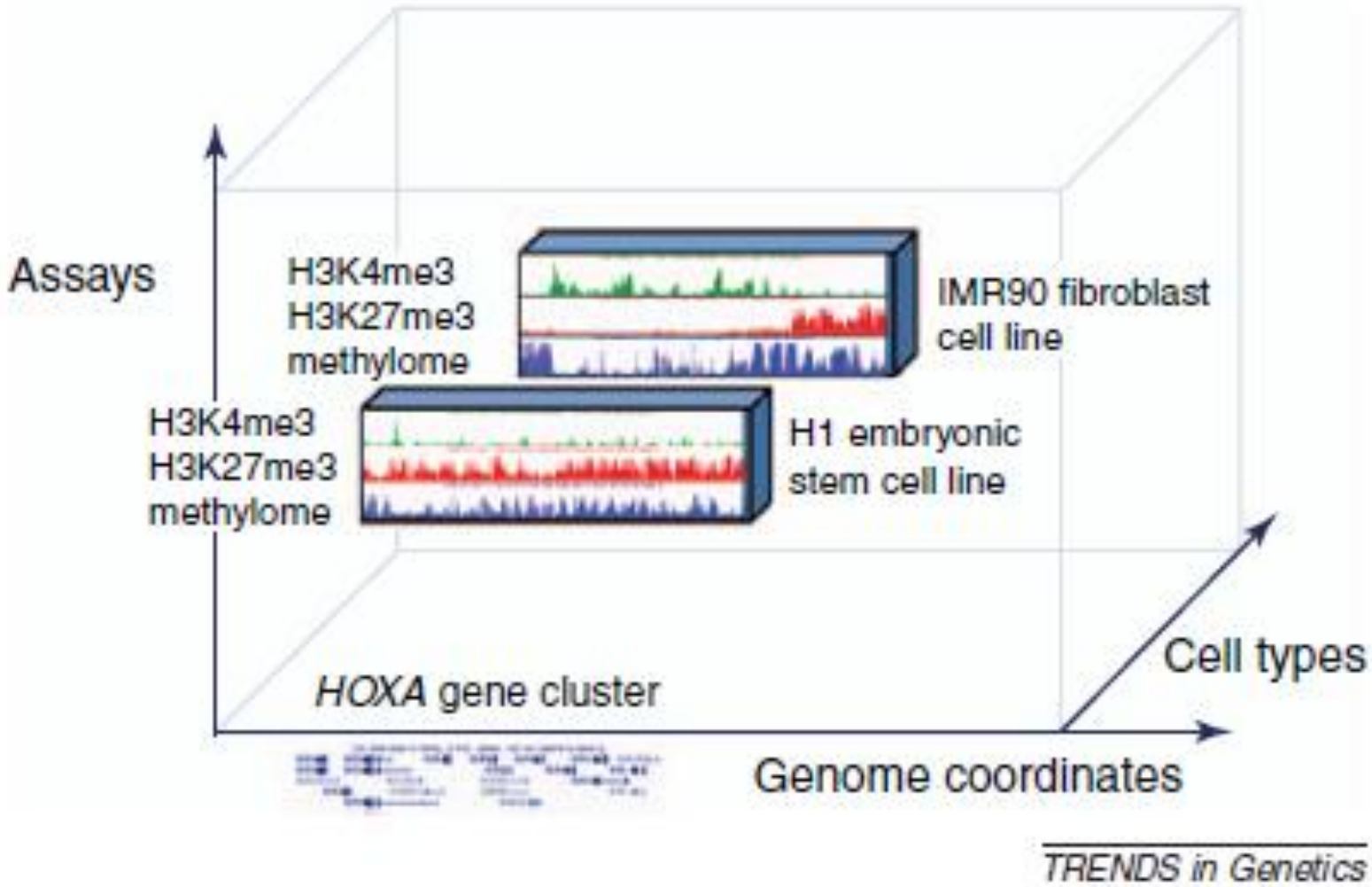
Genome
- SNP genotyping
- Resequencing from WGBS reads

Epigenomic profiling of human cell-lines, primary cells, and tissues

Jan 2013: 2500+ Experiments
5000+ Illumina runs



Space of Epigenomic Variation



Human Epigenome Atlas
www.epigenomeatlas.org

Assay types →

Sample types ↓

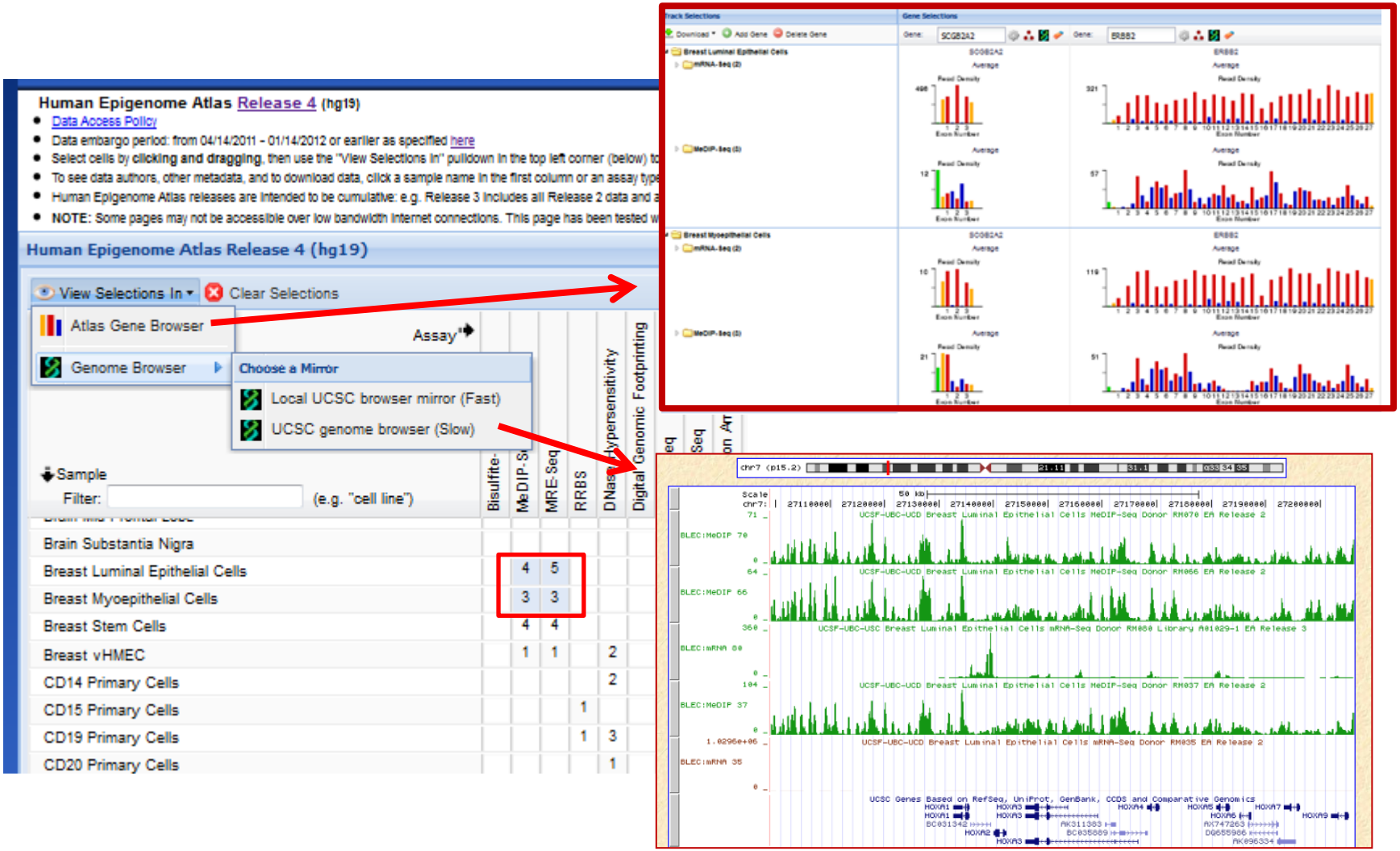
eaSampleType \ eaAssayType	Bisulfite-Seq	DNase Hypersensitivity	mRNA-Seq	Histone H3K27me3	Histone H3K36me3	Histone H3K4me1	Histone H3K4me3	Histone H3K9ac	Histone H3K9me3	Histone H3K27ac
Adult Liver	1		1	2	3	3	3	2	3	
Aorta				1	1				1	1
Bladder										
Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells										
Bone Marrow Derived Mesenchymal Stem Cells				4	4	4	4	4	4	
Brain Angular Gyrus				1	2	2	2	1	2	1
Brain Anterior Caudate				2	2	2	2	1	2	1
Brain Cingulate Gyrus				1	2	2	2	1	2	1
Brain Germinal Matrix	1		1	2	2	2	2		2	
Brain Hippocampus Middle	1		1	2	3	3	3	1	3	2
Brain Inferior Temporal Lobe				2	2	2	2	1	2	1
Brain Mid Frontal Lobe				1	2	2	2	1	2	1
Brain Substantia Nigra				2	2	2	2	1	2	
Breast Luminal Epithelial Cells			2	1	1	1			1	
Breast Myoepithelial Cells			2	2	2	2	2	2	2	
Breast Stem Cells			1							
Breast vHMEC		2	1	1	1	2	1		1	
CD14 Primary Cells		3								
CD15 Primary Cells				1	1	1	1		1	
CD19 Primary Cells		3		2	2	1	2		2	

⋮

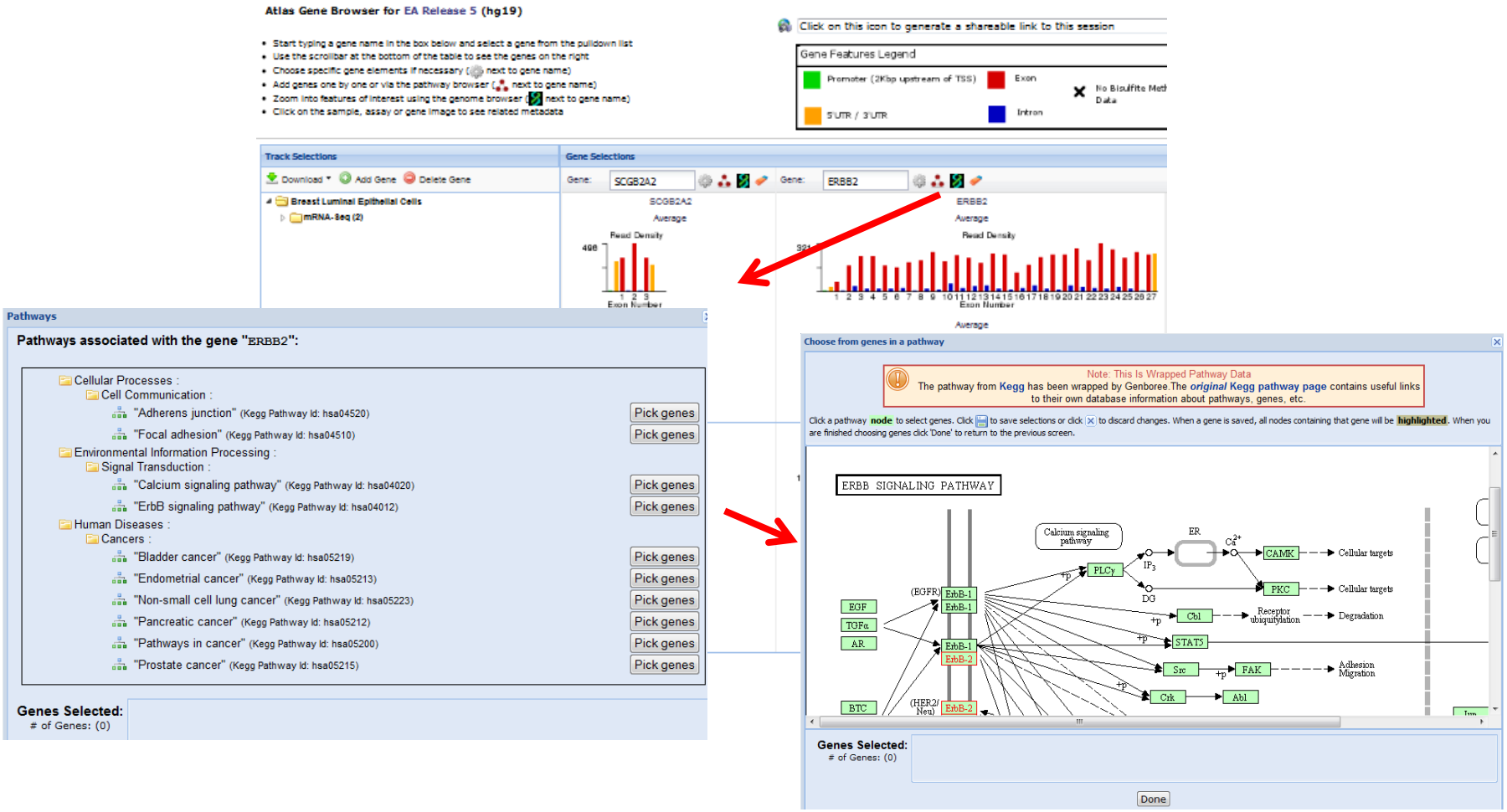
Genomic / Epigenomic Data

	“chip” data	**“seq” data**
Level 0	image	reads
Level 1	extracted features	mapped reads
Level 2	normalized Intensities (e.g., beta values)	read density maps (e.g., WIG file)
Level 3	Epigenomic state (e.g., quantitated peak calls)	
Level 4	Comparative analysis results (e.g., cell-type-specific marks)	

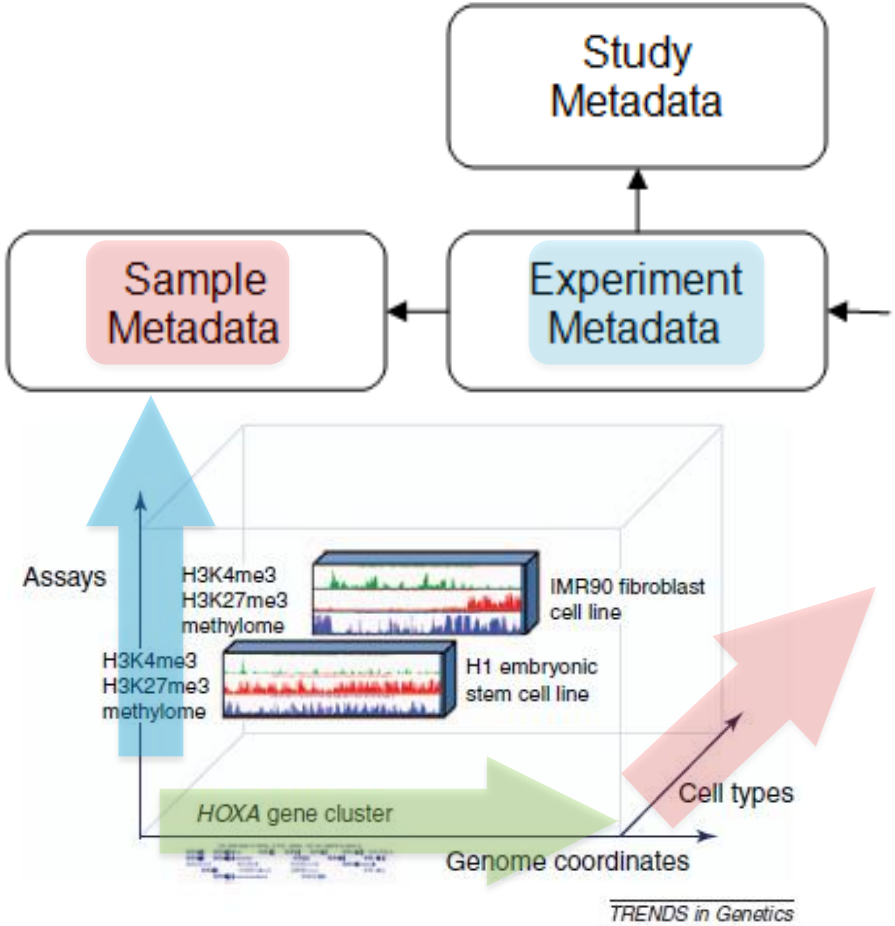
Viewing selections



Including More Genes in the Same Pathway



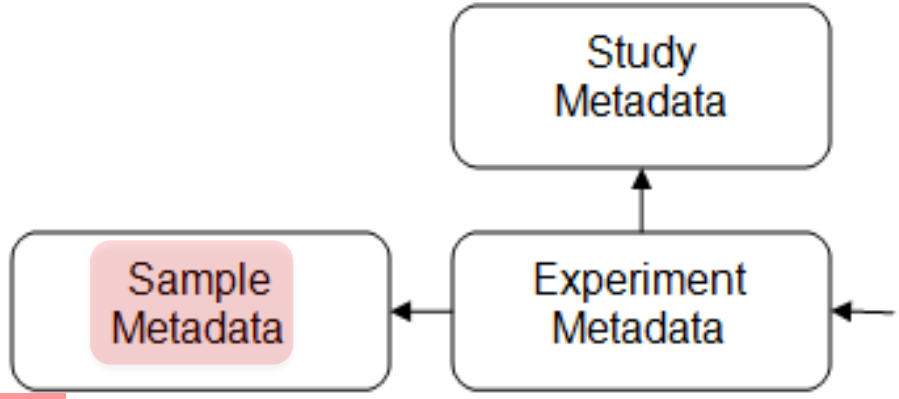
Epigenomic Metadata: SRA XML + Epigenomic Data Element Extensions



Level	Data	Metadata
0	Fastq / SRF	Run

Level	Analysis	Metadata
1	BED / BAM	Software Parameters QC
2	Wig / Counts	Software Parameters QC
3	Peaks / GFF3 / LFF	Software Parameters QC
4	Comparative Analysis / GFF3 / LFF	Software Parameters QC

Sample Metadata



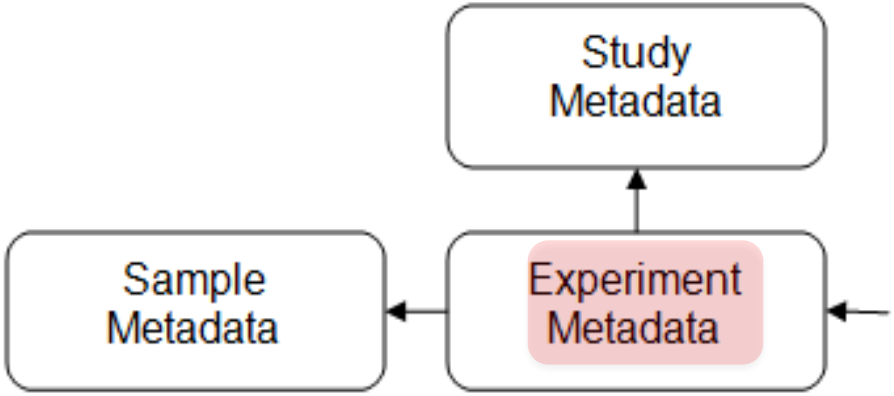
Level	Data	Metadata
0	Fastq / BAM	Run

Level	Analysis	Metadata
1	BED / BAM	Software Parameters QC
2	Wig / Counts	Software Parameters QC
3	Peaks / GFF3 / LFF	Software Parameters QC
4	Comparative Analysis / GFF3 / LFF	Software Parameters QC

Example:

<u>Primary Cell</u>
MOLECULE
DISEASE
BIOMATERIAL_PROVIDER
BIOMATERIAL_TYPE
CELL_TYPE
MARKERS
DONOR_ID
DONOR_AGE
DONOR_HEALTH_STATUS
DONOR_SEX
DONOR_ETHNICITY
PASSAGE_IF_EXPANDED

Experiment Metadata



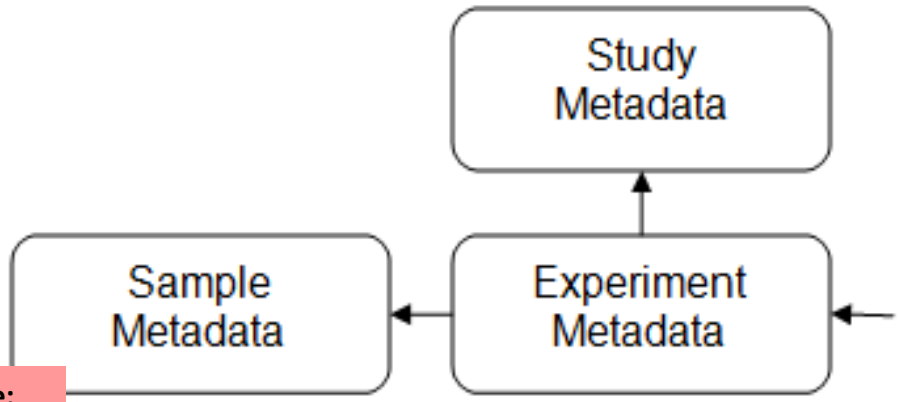
Level	Data	Metadata
0	Fastq / SRF	Run

Level	Analysis	Metadata
1	BED / BAM	Software Parameters QC
2	Wig / Counts	Software Parameters QC
3	Peaks / GFF3 / LFF	Software Parameters QC
4	Comparative Analysis / GFF3 / LFF	Software Parameters QC

Example:

Chip-Seq
EXPERIMENT_TYPE
EXTRACTION_PROTOCOL
EXTRACTION_PROTOCOL_TYPE_OF_SONICATOR
EXTRACTION_PROTOCOL_SONICATION_CYCLES
CHIP_PROTOCOL
CHIP_PROTOCOL_CHROMATIN_AMOUNT
CHIP_PROTOCOL_BEAD_TYPE
CHIP_PROTOCOL_BEAD_AMOUNT
CHIP_PROTOCOL_ANTIBODY_AMOUNT
CHIP_ANTIBODY
CHIP_ANTIBODY_PROVIDER
CHIP_ANTIBODY_CATALOG
CHIP_ANTIBODY_LOT

Ensuring Reproducibility: Metadata for Level 1 Analysis



Level	Data	Metadata
0	Fastq / SRF	Run

Level	Analysis	Metadata
1	BED / BAM	Software Parameters QC
2	Wig / Counts	Software Parameters QC
3	Peaks / GFF3 / LFF	Software Parameters QC
4	Comparative Analysis / GFF3 / LFF	Software Parameters QC

Example:

DATA_ANALYSIS_LEVEL - 1

EXPERIMENT_TYPE
GENOME_ASSEMBLY
SOFTWARE
SOFTWARE_VERSION
MAXIMUM_ALIGNMENT_LENGTH
MISMATCHES_ALLOWED
ALIGNMENTS_ALLOWED
TREATMENT_OF_MULTIPLE_ALIGNMENTS
TREATMENT_OF_IDENTICAL_ALIGNMENTS_OF_MULTIPLE_READS
ALIGNMENT_POSTPROCESSING
NUMBER_OF_MAPPED_READS
Quality Control

Implementation at the NIH Roadmap Data Coordination Center

Metadata can be submitted in two forms:

1. XML Document
   - automatic process
   - used for larger data submissions
   - includes format validators and QC steps
   - available for all assays

2. SpreadSheet
   - available for some assays
   - used by some smaller projects

Metadata for Illumina 450K Arrays (Excel SpreadSheet)

(“Study”)

(“Samples”)

(“Experiment”)

Use this template for Illumina submissions if you used an array already represented in GEO (for example, a commercial array).
Accompanying 'Matrix processed' and 'Matrix signal intensities' examples are included in other worksheets, click the tabs below.
 # Most fields in this template must be completed. Incomplete submissions will be returned.
 # Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines or,
 # CLICK HERE for Field Content Guidelines Web page.

SERIES

This section describes the overall experiment.

title	DNA methylation in {sampleType} from individuals with {diseaseState} {and normal controls}
summary	Genome wide DNA methylation profiling of {diseseState} {and normal} {sampleType} samples. The Illumina Infinium HumanMethylation450 BeadCh
overall design	Bisulfite converted DNA from the {number of samples} samples were hybridized to the Illumina Infinium HumanMethylation450 BeadChip v1.1
contributor	{contributor name}
contributor	{contributor name}

SAMPLES

 # The Sample names in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheets).
CLICK HERE to find the platform accession number (GPLxxxx).
Note: As of 5/17/2010, the GEO database has two Platform records for Illumina methylation BeadChips:
GPL8490 Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)
GPL9183 Illumina GoldenGate Methylation Cancer Panel I

Sample name	title	source name	organism	characteristics: gen	characteristics: a	characte
{sampleID}	Methylation in {sampleType} {sampleID}	{sampleType} {sampleID}	Homo sapiens	{Male, Female, or NA}	{Age}	{sampleTy
{sampleID}	Methylation in {sampleType} {sampleID}	{sampleType} {sampleID}	Homo sapiens	{Male, Female, or NA}	{Age}	{sampleTy

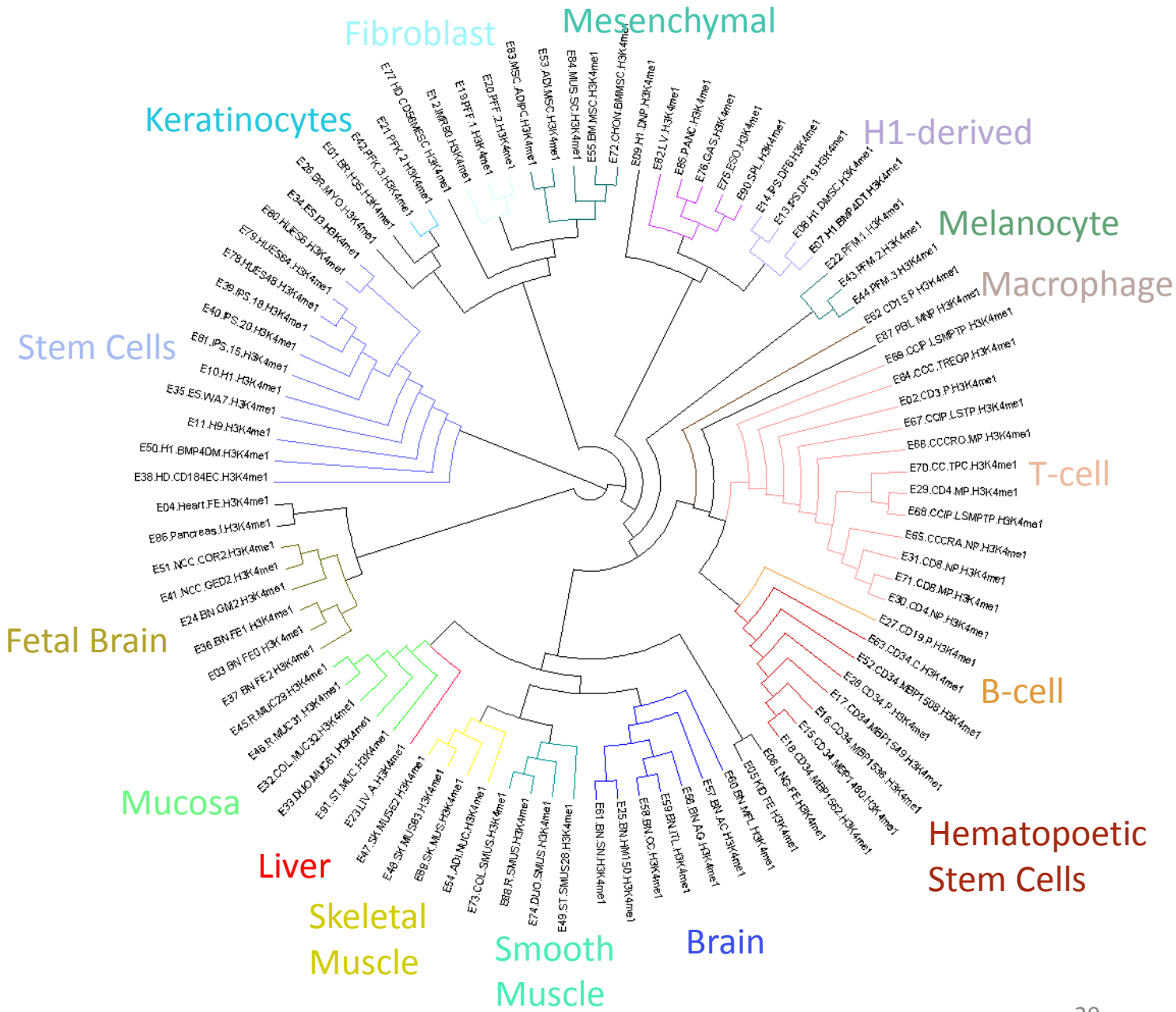
PROTOCOLS

This section includes protocols and fields which are common to all Samples.
Protocols which are applicable to specific Samples or specific channels should be included in additional columns of the SAMPLES section instead.

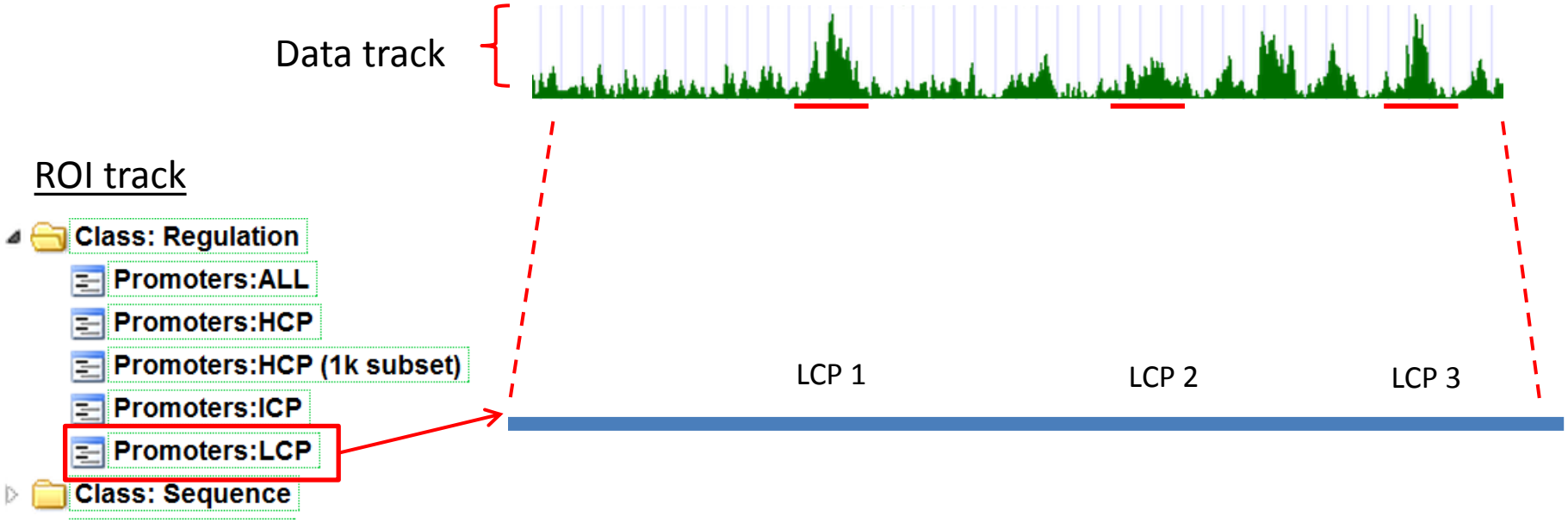
extract protocol	Genomic DNA was extracted and purified from peripheral blood samples using {kitName} according to standard instructions
label protocol	Standard Illumina protocol
hyb protocol	Bisulfite converted DNA was amplified, fragmented and hybridized to the Illumina Infinium HumanMethylation450 BeadChip using standard Illumina
scan protocol	Arrays were imaged using BeadArray Reader using standard recommended Illumina scanner settings
data processing	GenomeStudio software version 2010.3.0.30128
matrix processed value definition	normalized Average Beta
# matrix signal intensities value defi	Unmethylated and methylated signal intensities

- NIH Roadmap Epigenomics Project
- Epigenome Analysis
- Genboree Workbench
- Genboree Network

Integrative Analysis Paper: Cellular Differentiation

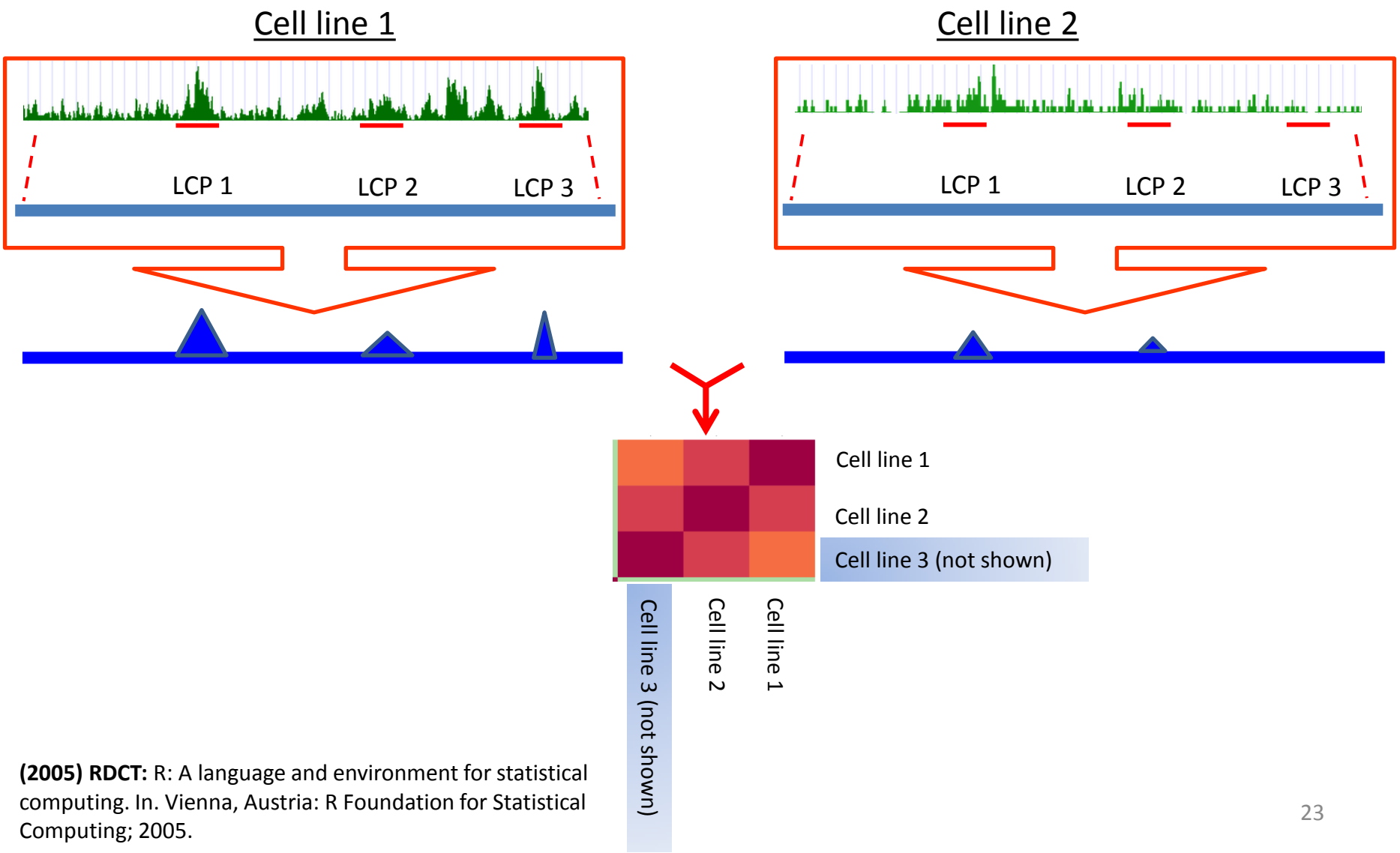


Regions of Interest (ROIs)



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

Similarity Matrix / Heatmap



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

The Epigenomic Tolset within the Genboree Workbench will have sufficient tools to reproduce key analyses from the integrative analysis paper

<u>The Toolset enables analyses of the Epigenome Atlas and of private data</u>

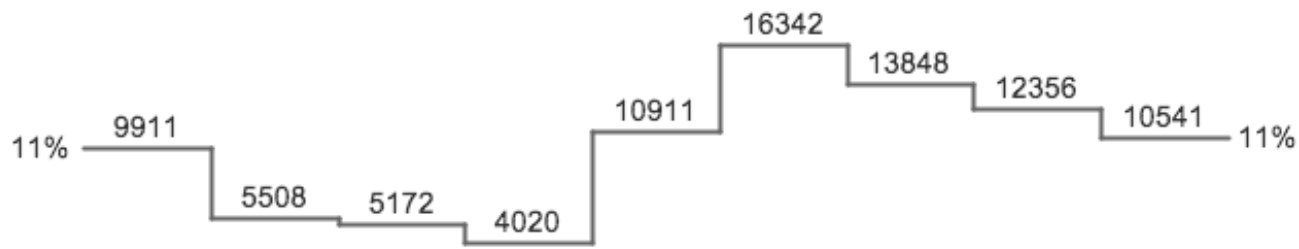
Generate heatmap clusterings and trees with different combinations of ROIs / marks / samples

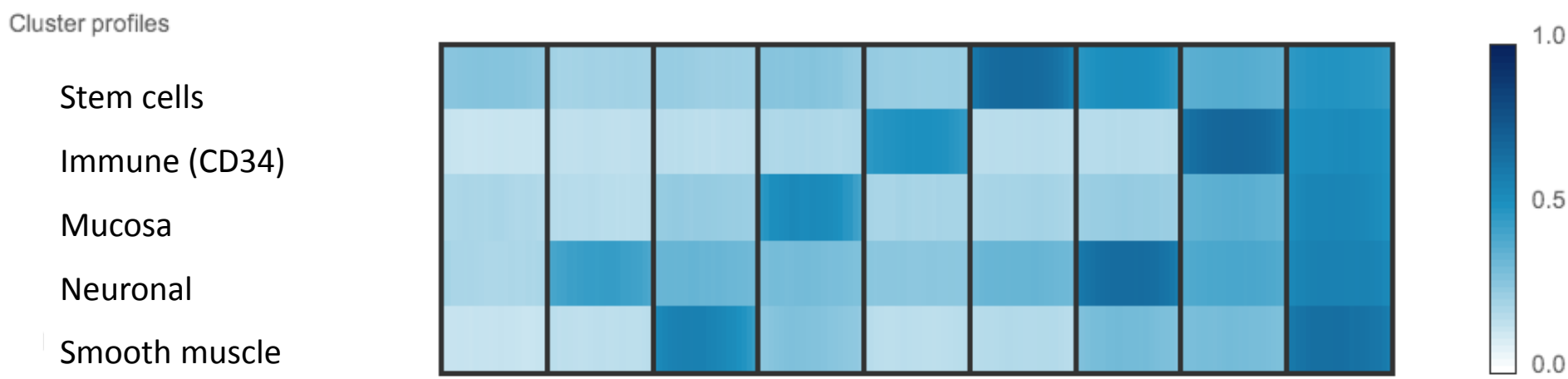
Identify branch-specific epigenomic changes

- Analyze oordinated epigenomic changes over enhancers
- Identify genes and pathways regulated by the enhancers
- Identify transcription Factors (TFs) involved in regulating specific branches
- Map patterns of epigenomic changes over regulated genes

Identify enhancers with branch-specific activation patterns (H3K4me1↑)

88,609 enhancers with branch-specific H3K4me1 signals clustered by Spark

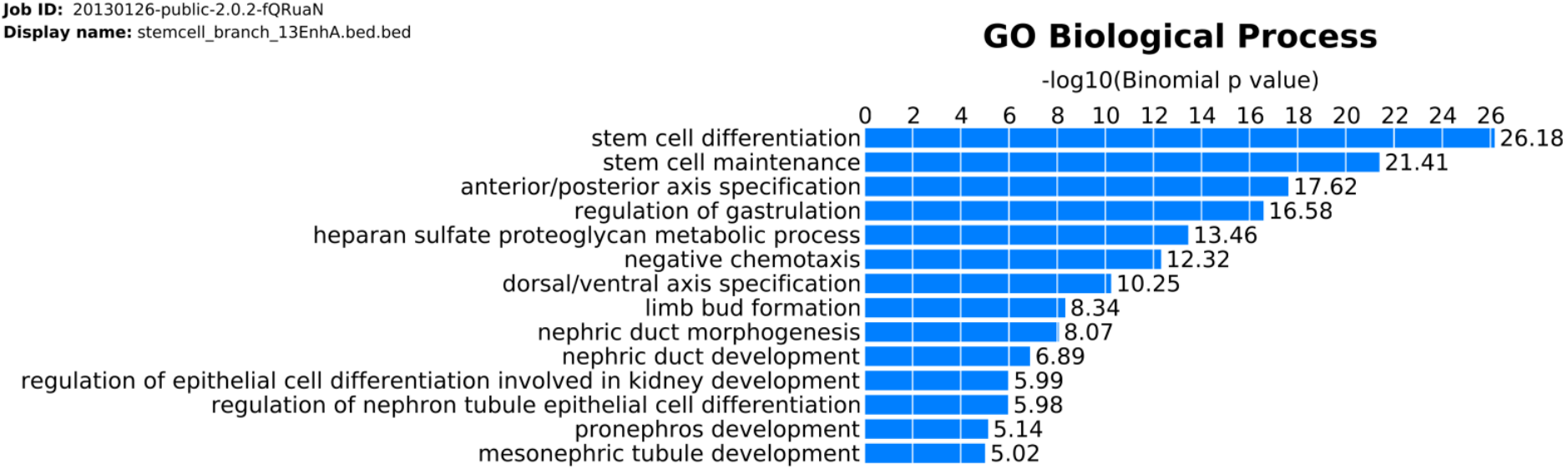




Spark : *Nielsen CB et al. Genome Res. (11):2262-9 2012*

GO analysis of genes associated with enhancers activated in the stem cell branch

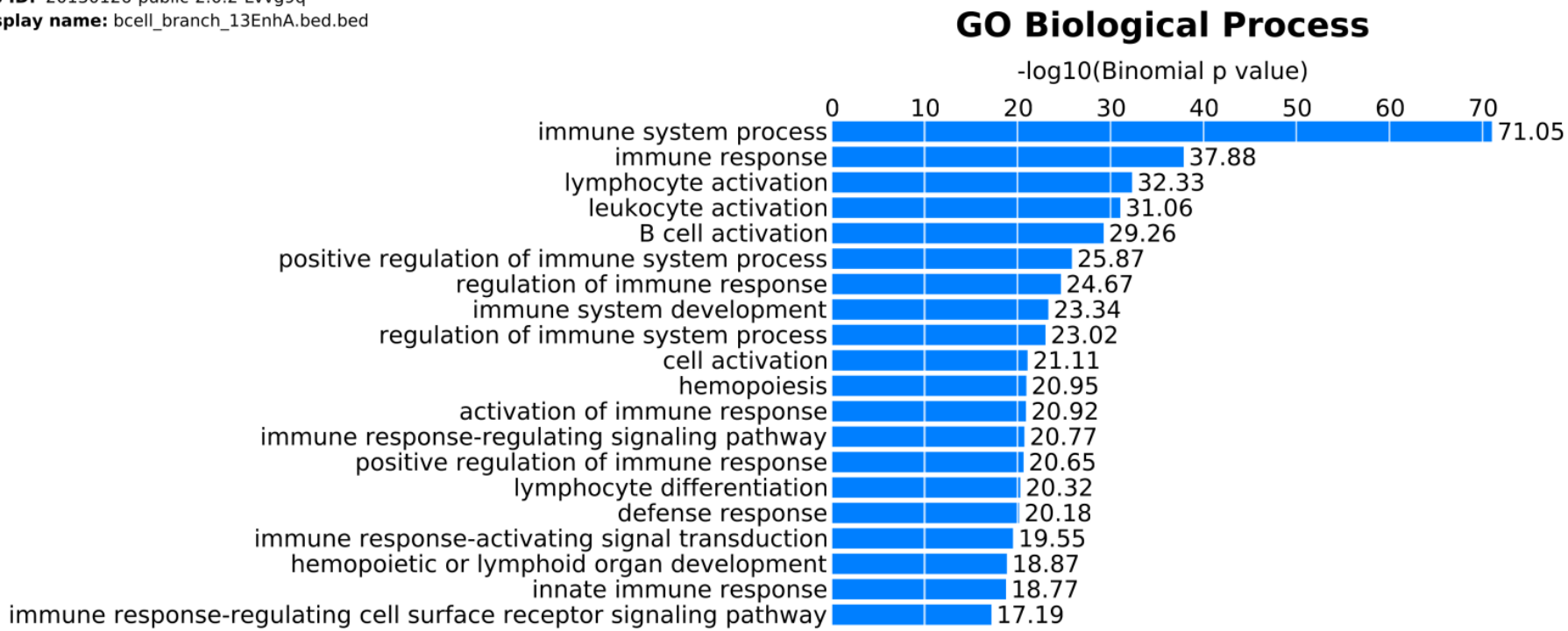
Job ID: 20130126-public-2.0.2-fQRuaN
Display name: stemcell_branch_13EnhA.bed.bed



GREAT tool from the Bejarano Lab, Stanford

GO analysis of genes associated with enhancers activated in the B-cell branch

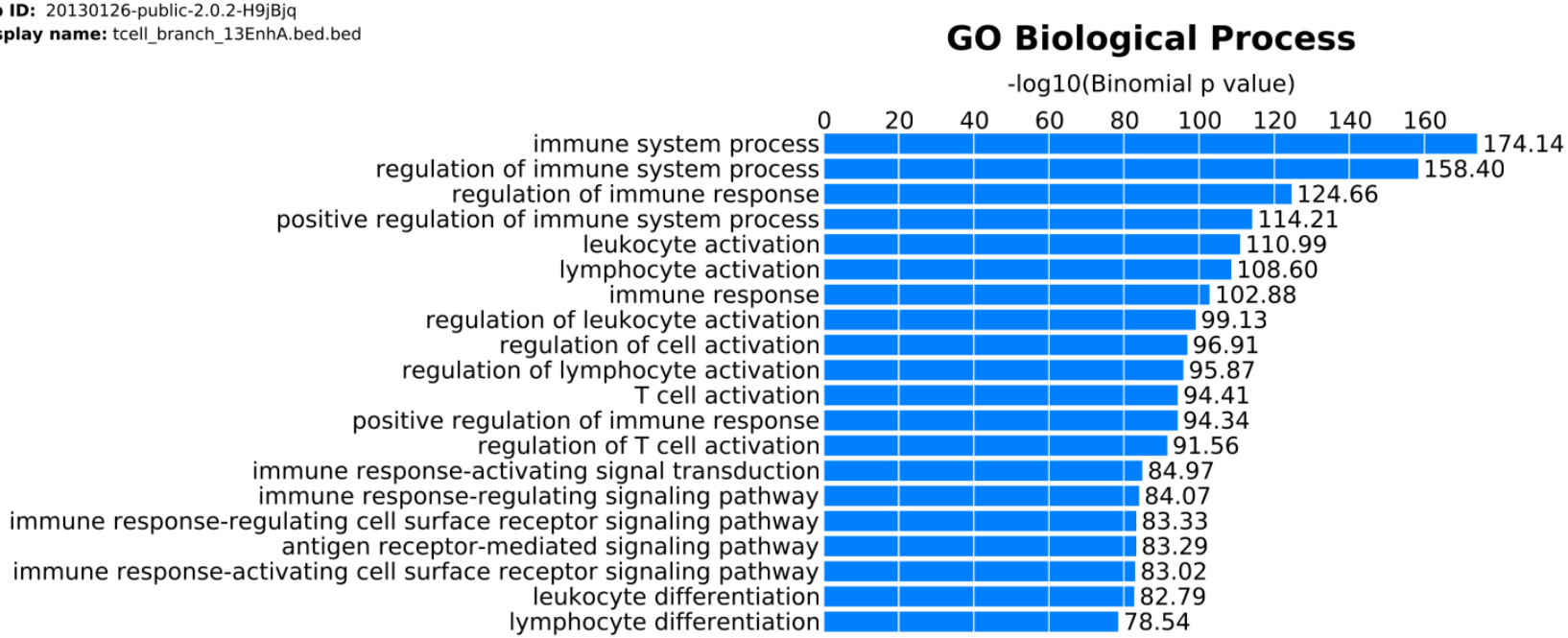
Job ID: 20130126-public-2.0.2-Lvvg9q
Display name: bcell_branch_13EnhA.bed.bed



GREAT tool from the Bejarano Lab, Stanford

GO analysis of genes associated with enhancers activated in the T-cell branch

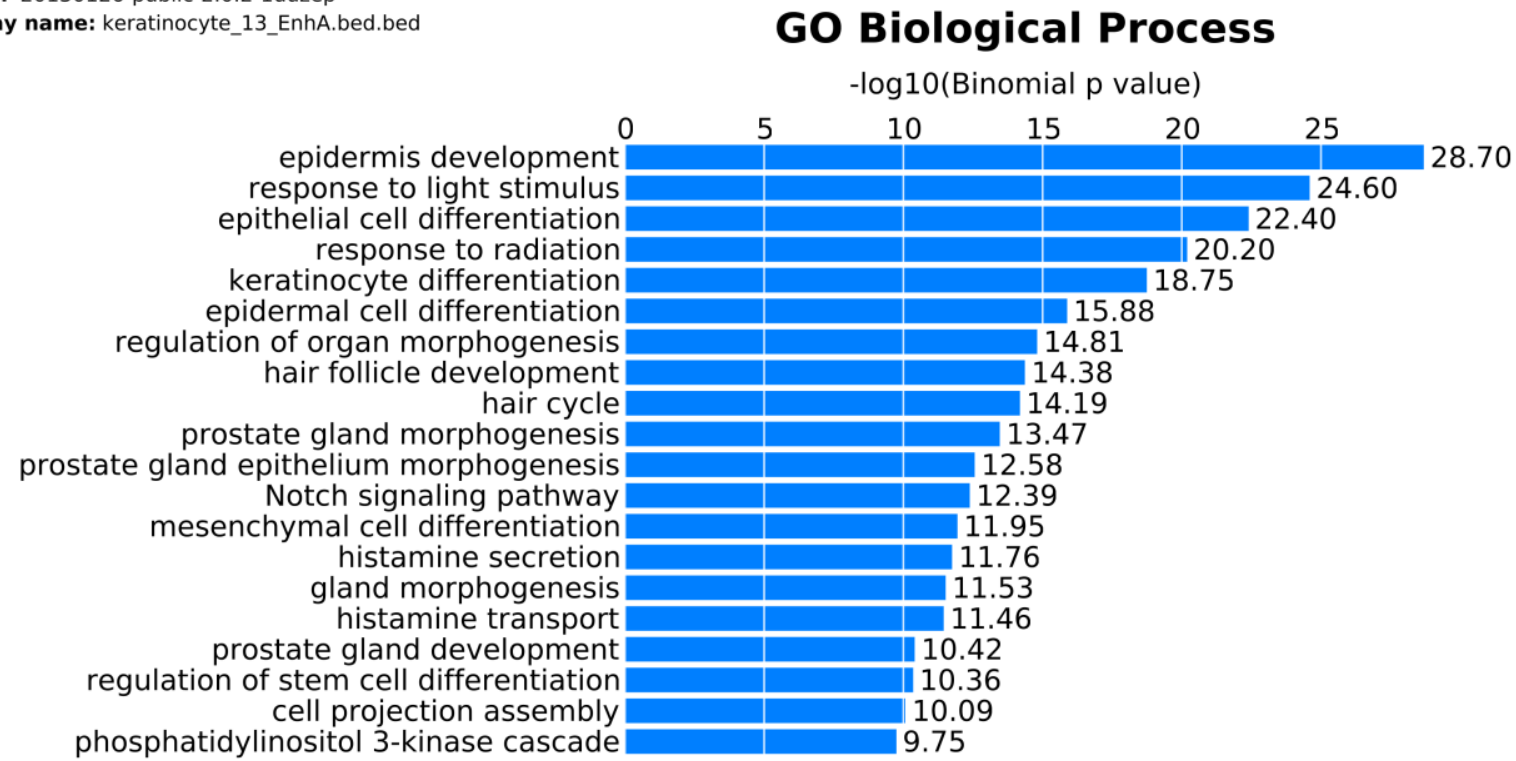
Job ID: 20130126-public-2.0.2-H9jBjq
Display name: tcell_branch_13EnhA.bed.bed



GREAT tool from the Bejarano Lab, Stanford

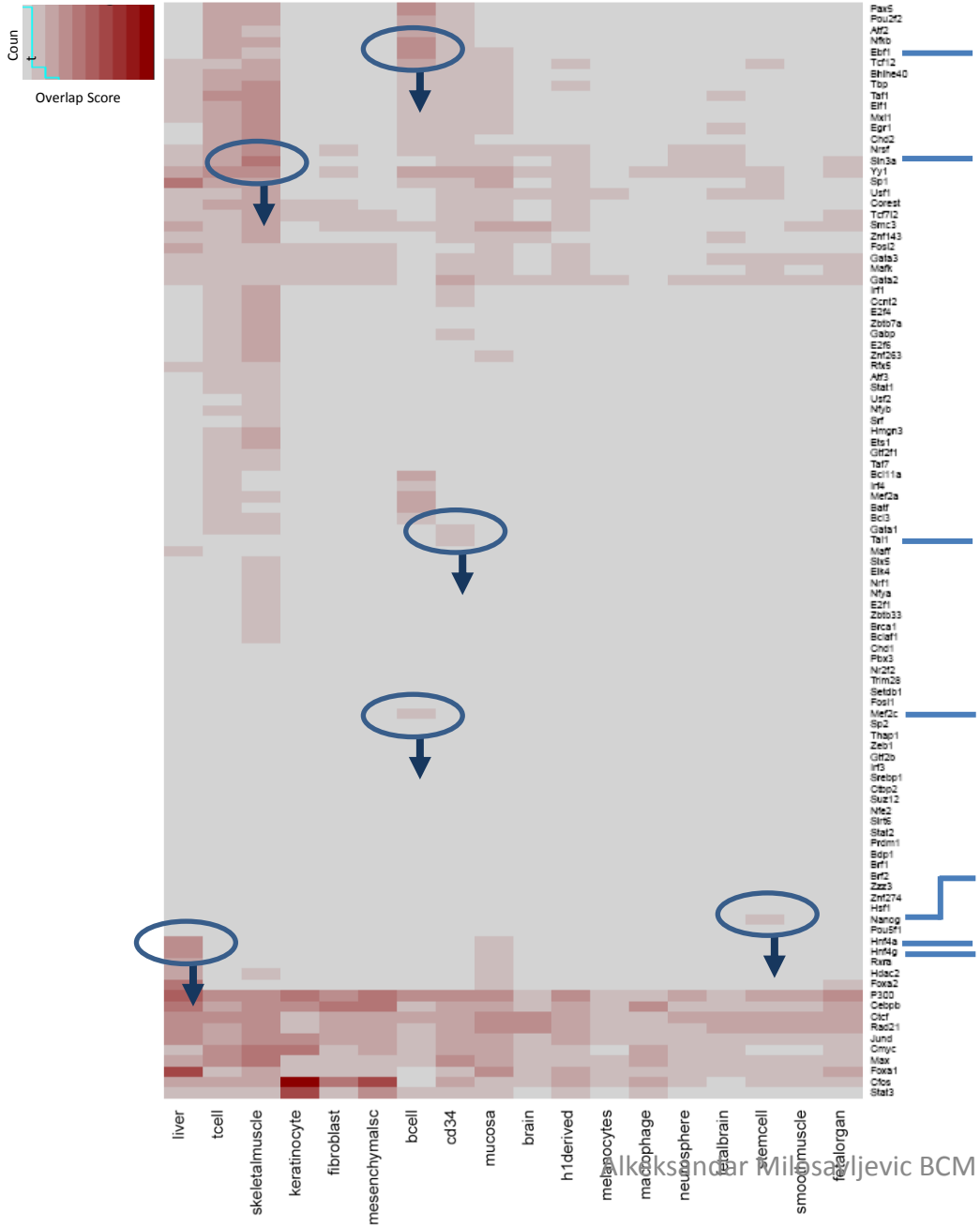
GO analysis of genes associated with enhancers activated in the keratinocyte branch

Job ID: 20130126-public-2.0.2-1uuzep
Display name: keratinocyte_13_EnhA.bed.bed



GREAT tool from the Bejarano Lab, Stanford

TF overlap with branch-specific enhancers



Efb1 is essential for **B cell** development

Sin3a is known to play role in **muscle** differentiation.

Tal1 plays key role in **HSCs** self renewal and engraftment

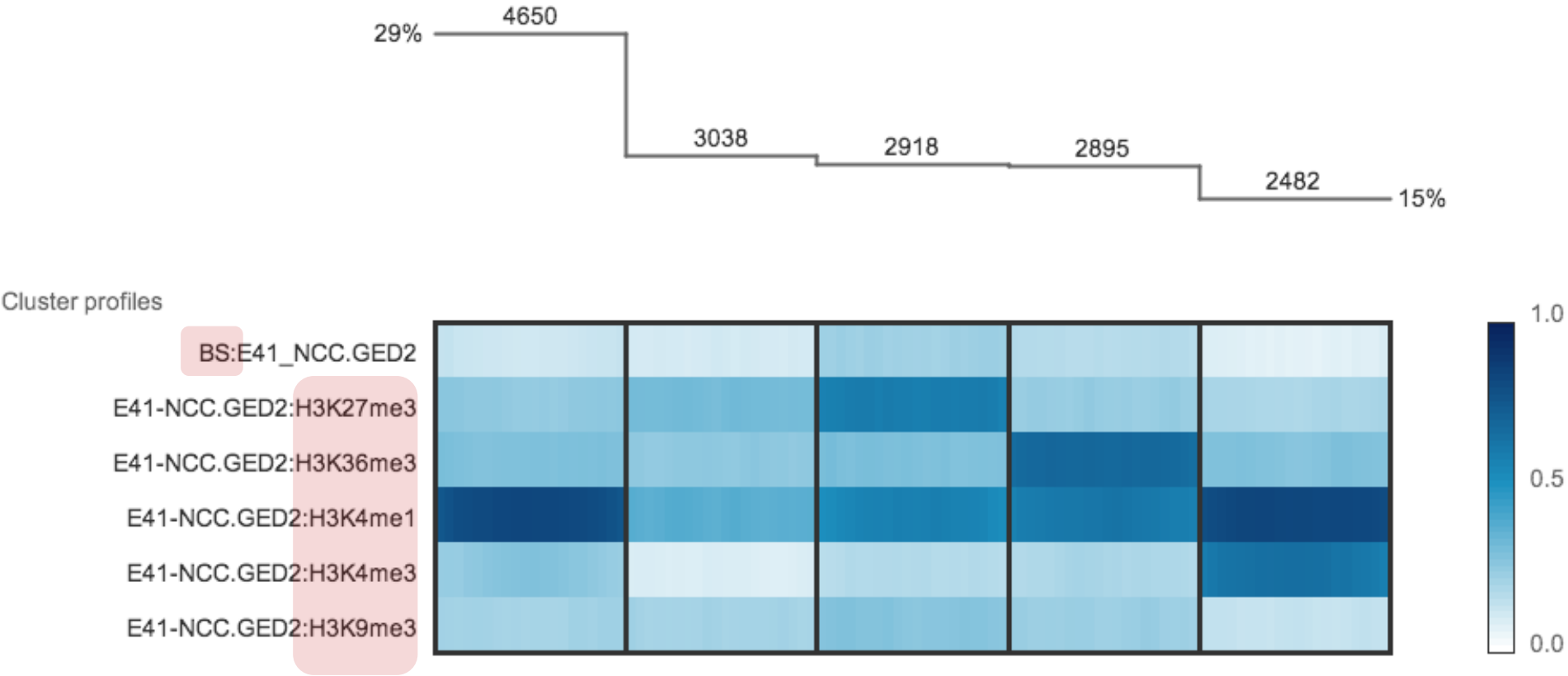
Mef2c is required for **B cell** proliferation

Nanog important for **ESCs** self-renewal

Hnf4a and Hnf4g regulate **hepatic** genes

Coordinated changes of epigenomic marks in the neuronal branch

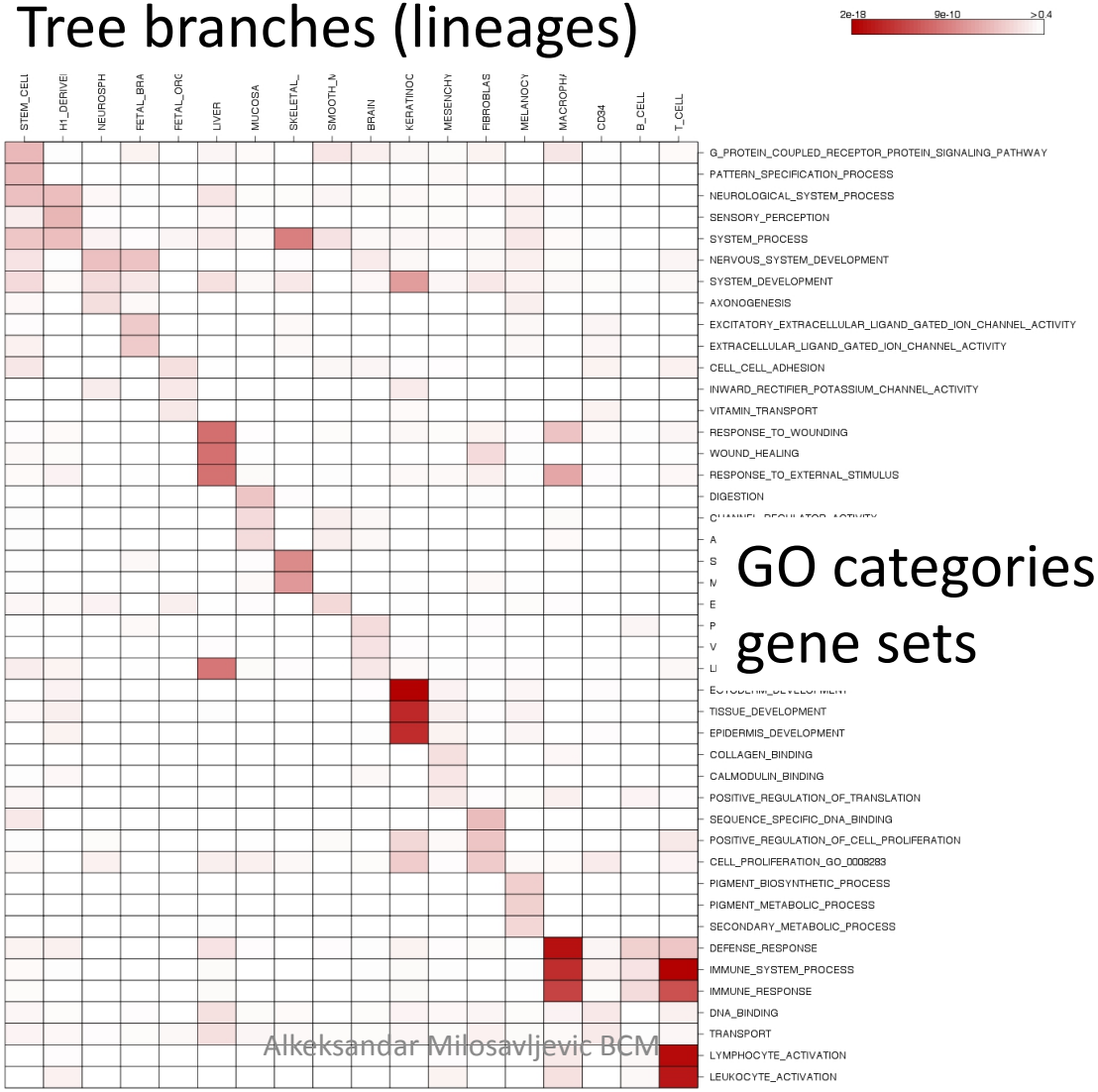
15,983 enhancers activated specifically in the neuronal branch



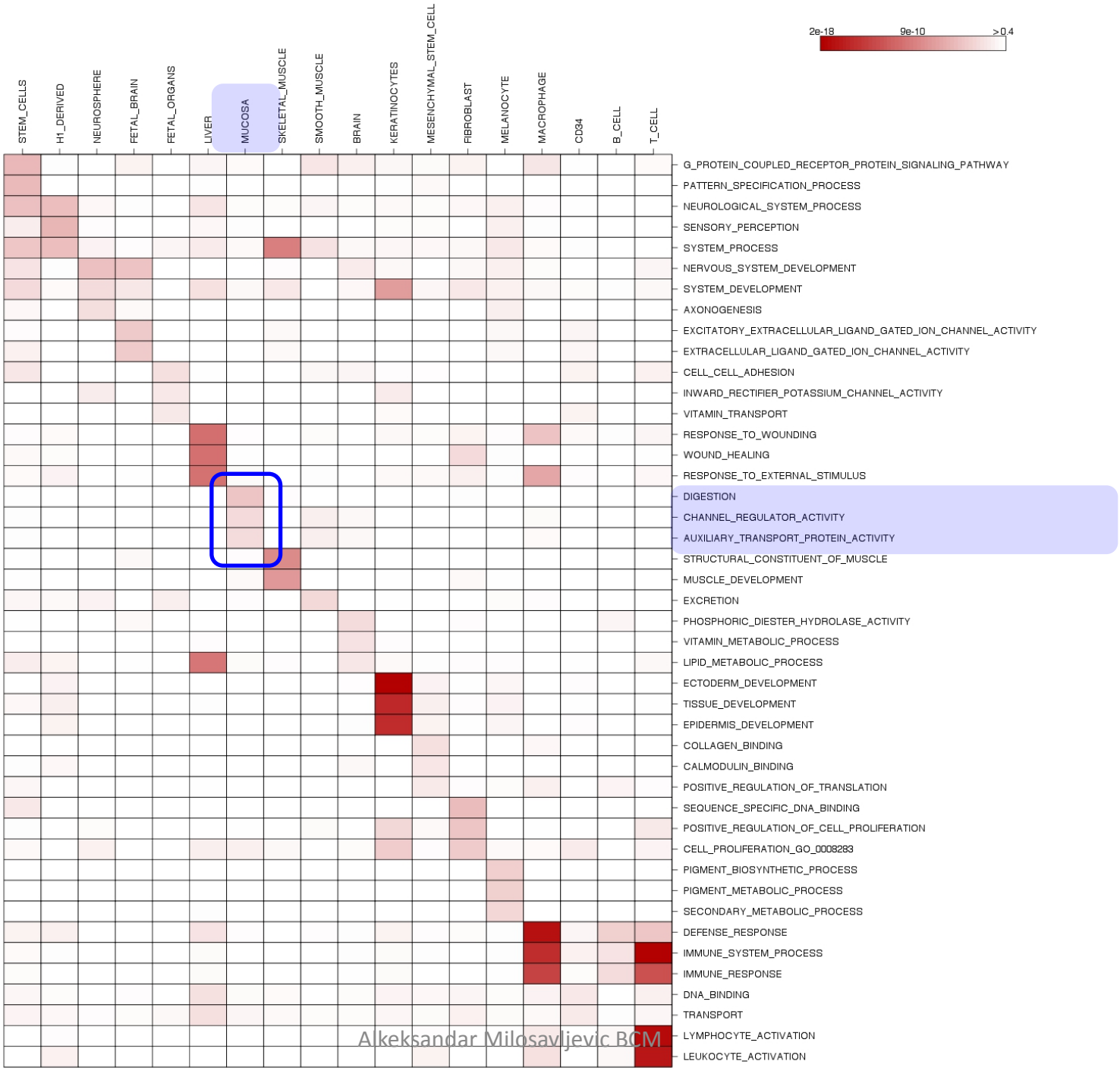
Spark : *Nielsen CB et al. Genome Res. (11):2262-9 2012*

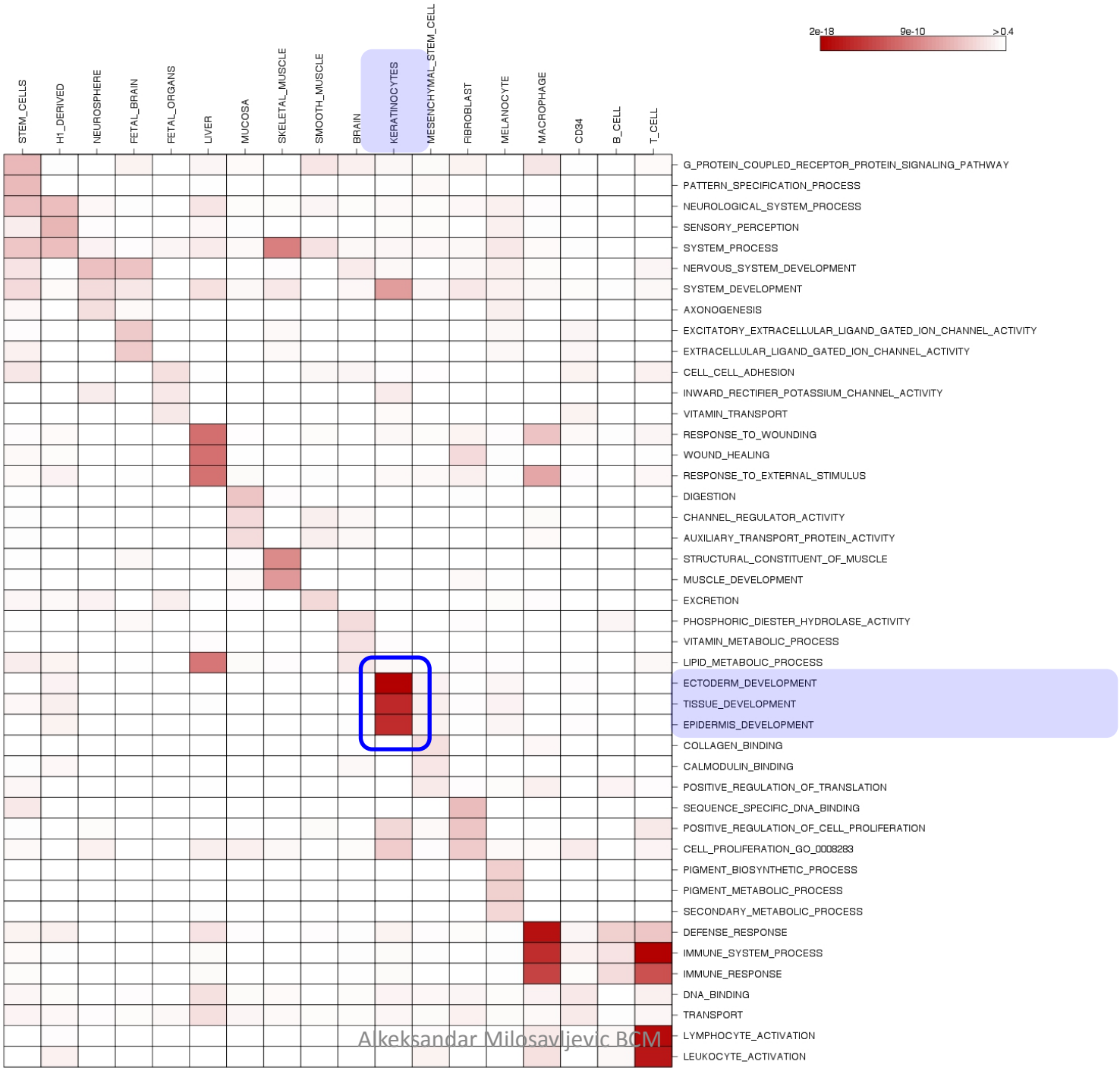
Branch-specific changes in H3K4me3 signals over promoters

Tree branches (lineages)



GO categories for gene sets


2e-18
9e-10
>0.4
STEM_CELLS
H1_DERIVED
NEUROSPHERE
FETAL_BRAIN
FETAL_ORGANS
LIVER
MUCOSA
SKELETAL_MUSCLE
SMOOTH_MUSCLE
BRAIN
KERATINOCYTES
MESENCHYMAL_STEM_CELL
FIBROBLAST
MELANOCYTE
MACROPHAGE
CD34
B_CELL
T_CELL
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY
PATTERN_SPECIFICATION_PROCESS
NEUROLOGICAL_SYSTEM_PROCESS
SENSORY_PERCEPTION
SYSTEM_PROCESS
NERVOUS_SYSTEM_DEVELOPMENT
SYSTEM_DEVELOPMENT
AXONOGENESIS
EXCITATORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
CELL_CELL_ADHESION
INWARD_RECTIFIER_POTASSIUM_CHANNEL_ACTIVITY
VITAMIN_TRANSPORT
RESPONSE_TO_WOUNDING
WOUND_HEALING
RESPONSE_TO_EXTERNAL_STIMULUS
DIGESTION
CHANNEL_REGULATOR_ACTIVITY
AUXILIARY_TRANSPORT_PROTEIN_ACTIVITY
STRUCTURAL_CONSTITUENT_OF_MUSCLE
MUSCLE_DEVELOPMENT
EXCRETION
PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY
VITAMIN_METABOLIC_PROCESS
LIPID_METABOLIC_PROCESS
ECTODERM_DEVELOPMENT
TISSUE_DEVELOPMENT
EPIDERMIS_DEVELOPMENT
COLLAGEN_BINDING
CALMODULIN_BINDING
POSITIVE_REGULATION_OF_TRANSLATION
SEQUENCE_SPECIFIC_DNA_BINDING
POSITIVE_REGULATION_OF_CELL_PROLIFERATION
CELL_PROLIFERATION_GO_0008283
PIGMENT_BIOSYNTHETIC_PROCESS
PIGMENT_METABOLIC_PROCESS
SECONDARY_METABOLIC_PROCESS
DEFENSE_RESPONSE
IMMUNE_SYSTEM_PROCESS
IMMUNE_RESPONSE
DNA_BINDING
TRANSPORT
LYMPHOCYTE_ACTIVATION
LEUKOCYTE_ACTIVATION

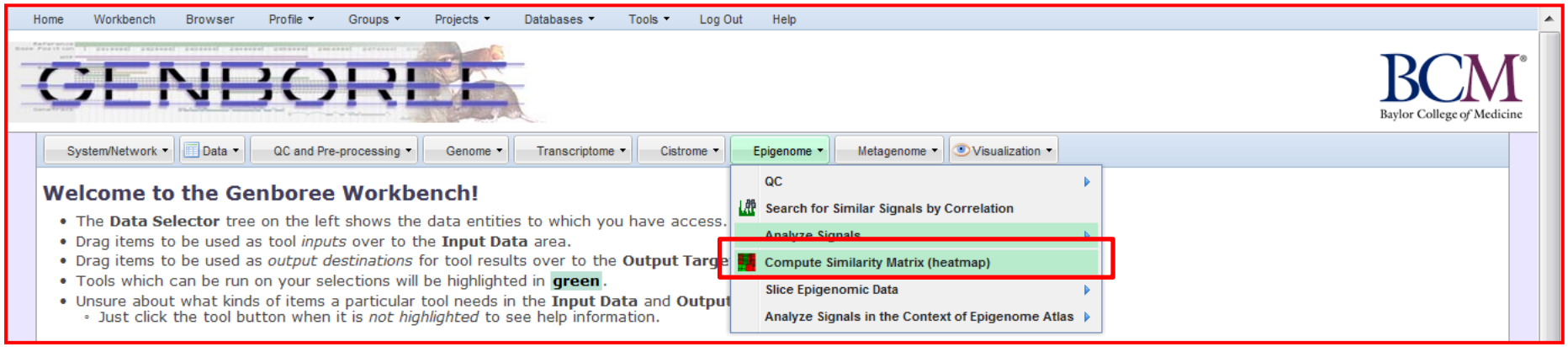

2e-18
9e-10
>0.4
STEM_CELLS
H1_DERIVED
NEUROSPHERE
FETAL_BRAIN
FETAL_ORGANS
LIVER
MUCOSA
SKELETAL_MUSCLE
SMOOTH_MUSCLE
BRAIN
KERATINOCYTES
MESENCHYMAL_STEM_CELL
FIBROBLAST
MELANOCYTE
MACROPHAGE
CD34
B_CELL
T_CELL
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY
PATTERN_SPECIFICATION_PROCESS
NEUROLOGICAL_SYSTEM_PROCESS
SENSORY_PERCEPTION
SYSTEM_PROCESS
NERVOUS_SYSTEM_DEVELOPMENT
SYSTEM_DEVELOPMENT
AXONOGENESIS
EXCITATORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
CELL_CELL_ADHESION
INWARD_RECTIFIER_POTASSIUM_CHANNEL_ACTIVITY
VITAMIN_TRANSPORT
RESPONSE_TO_WOUNDING
WOUND_HEALING
RESPONSE_TO_EXTERNAL_STIMULUS
DIGESTION
CHANNEL_REGULATOR_ACTIVITY
AUXILIARY_TRANSPORT_PROTEIN_ACTIVITY
STRUCTURAL_CONSTITUENT_OF_MUSCLE
MUSCLE_DEVELOPMENT
EXCRETION
PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY
VITAMIN_METABOLIC_PROCESS
LIPID_METABOLIC_PROCESS
ECTODERM_DEVELOPMENT
TISSUE_DEVELOPMENT
EPIDERMIS_DEVELOPMENT
COLLAGEN_BINDING
CALMODULIN_BINDING
POSITIVE_REGULATION_OF_TRANSLATION
SEQUENCE_SPECIFIC_DNA_BINDING
POSITIVE_REGULATION_OF_CELL_PROLIFERATION
CELL_PROLIFERATION_GO_0008283
PIGMENT_BIOSYNTHETIC_PROCESS
PIGMENT_METABOLIC_PROCESS
SECONDARY_METABOLIC_PROCESS
DEFENSE_RESPONSE
IMMUNE_SYSTEM_PROCESS
IMMUNE_RESPONSE
DNA_BINDING
TRANSPORT
LYMPHOCYTE_ACTIVATION
LEUKOCYTE_ACTIVATION


2e-18
9e-10
>0.4
STEM_CELLS
H1_DERIVED
NEUROSPHERE
FETAL_BRAIN
FETAL_ORGANS
LIVER
MUCOSA
SKELETAL_MUSCLE
SMOOTH_MUSCLE
BRAIN
KERATINOCYTES
MESENCHYMAL_STEM_CELL
FIBROBLAST
MELANOCYTE
MACROPHAGE
CD34
B_CELL
T_CELL
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY
PATTERN_SPECIFICATION_PROCESS
NEUROLOGICAL_SYSTEM_PROCESS
SENSORY_PERCEPTION
SYSTEM_PROCESS
NERVOUS_SYSTEM_DEVELOPMENT
SYSTEM_DEVELOPMENT
AXONOGENESIS
EXCITATORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY
CELL_CELL_ADHESION
INWARD_RECTIFIER_POTASSIUM_CHANNEL_ACTIVITY
VITAMIN_TRANSPORT
RESPONSE_TO_WOUNDING
WOUND_HEALING
RESPONSE_TO_EXTERNAL_STIMULUS
DIGESTION
CHANNEL_REGULATOR_ACTIVITY
AUXILIARY_TRANSPORT_PROTEIN_ACTIVITY
STRUCTURAL_CONSTITUENT_OF_MUSCLE
MUSCLE_DEVELOPMENT
EXCRETION
PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY
VITAMIN_METABOLIC_PROCESS
LIPID_METABOLIC_PROCESS
ECTODERM_DEVELOPMENT
TISSUE_DEVELOPMENT
EPIDERMIS_DEVELOPMENT
COLLAGEN_BINDING
CALMODULIN_BINDING
POSITIVE_REGULATION_OF_TRANSLATION
SEQUENCE_SPECIFIC_DNA_BINDING
POSITIVE_REGULATION_OF_CELL_PROLIFERATION
CELL_PROLIFERATION_GO_0008283
PIGMENT_BIOSYNTHETIC_PROCESS
PIGMENT_METABOLIC_PROCESS
SECONDARY_METABOLIC_PROCESS
DEFENSE_RESPONSE
IMMUNE_SYSTEM_PROCESS
IMMUNE_RESPONSE
DNA_BINDING
TRANSPORT
LYMPHOCYTE_ACTIVATION
LEUKOCYTE_ACTIVATION

The data for several of the workshop use cases is taken from this publication:

"Functional annotation of the human brain methylome across brain and blood". Matthew Davies[1], Manuela Volta[1], Abhishek Dixit[1], Simon Lovestone[1], Cristian Coarfa[2], R. Alan Harris[2], Aleksandar Milosavljevic[2], Claire Troakes[1], Safa Al-Sarraj[1], Richard Dobson[1], Leonard C. Schalkwyk[1], Jonathan Mill[1]*
Genome Biology, 12:R43, 2012

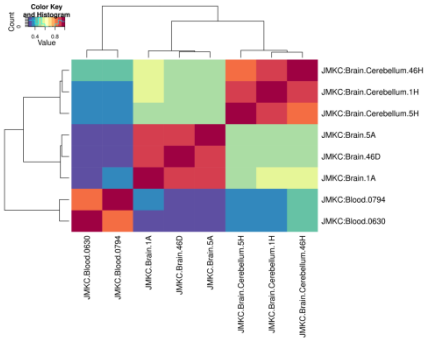
[1]Institute of Psychiatry, King's College London. UK. [2]Baylor College of Medicine, Houston, Texas. USA. *Corresponding Author: Dr. Jonathan Mill, Address: Institute of Psychiatry, SGDP Centre, De Crespigny Park, Denmark Hill, London.



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

Epigenomic Tracks:
- -Blood
- -Cerebellum
- -Cortex

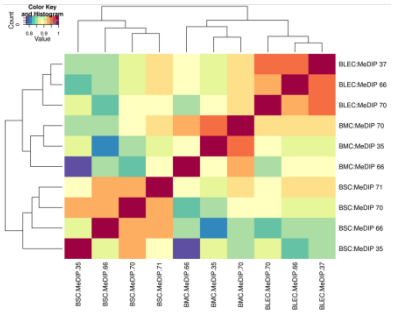
ROIs

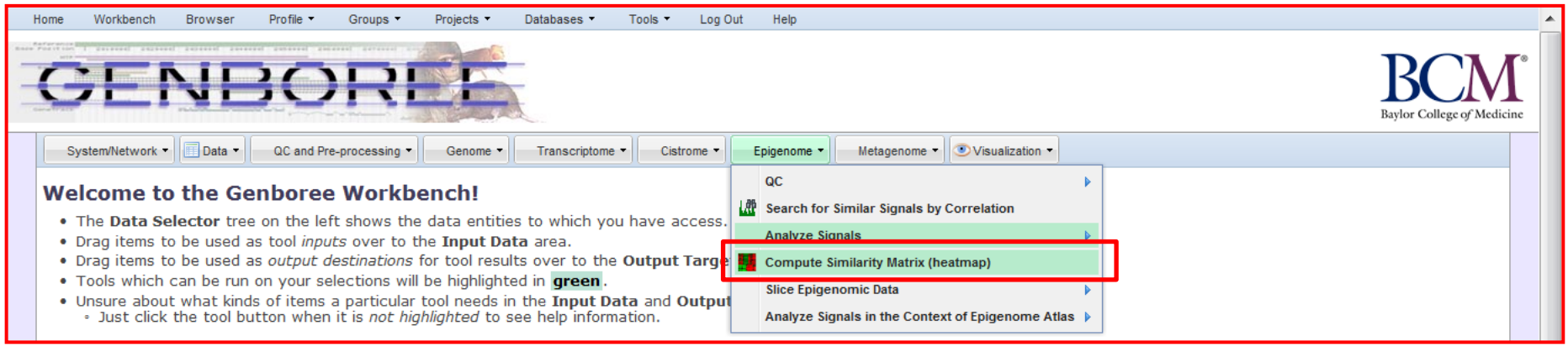


Use Case 2: Breast Cell Types Cluster Based on Their MeDIP-seq Profiles

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

ROIs

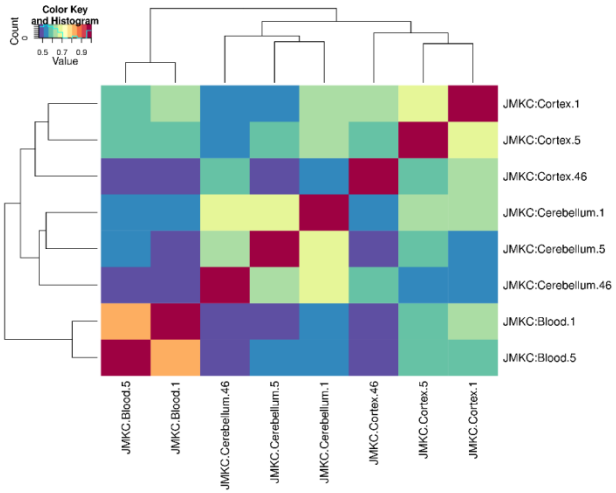




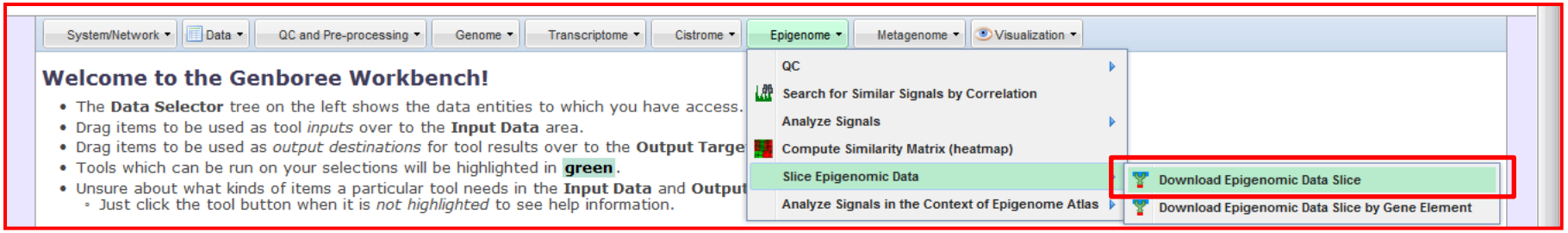
Use Case 5: Methylation of some features discriminate tissue type better than others

Epigenomic Tracks
- -Blood
- -Cerebellum
- -Cortex

ROIs



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



Epigenomic Tracks:
-H1 cell line
-IMR90 cell line

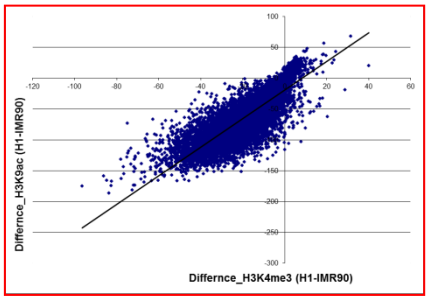
Collates score tracks into one data matrix, export to Excel

Bisulfite data **H1.H3K9ac** **IMR90.H3K9ac** **H1.H3K4me3**

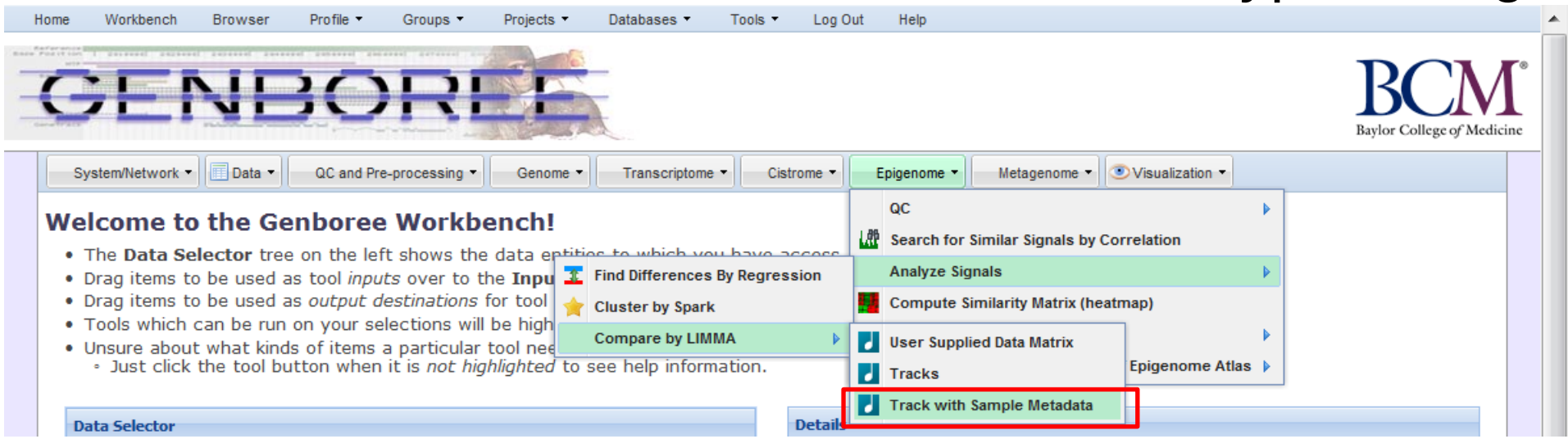
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Index	H1.BS.Con	IMR90.BS.Combined		H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac.68		IMR90.H3K	IMR90.H3K9ac.46		H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3K4m	H1.H3
2	HSAP04065	0	0.044444		0.314286	0.314286	0.314286	0	0.314286		0.314286	0.314286		0.628571	0.628571	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: Determine breast cancer cell type of origin



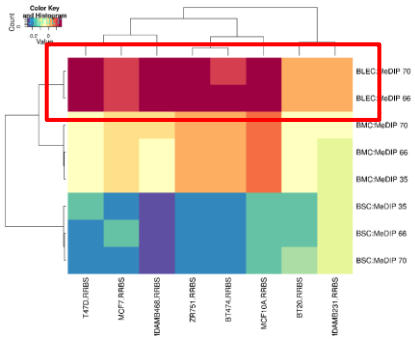
LIMMA: Smyth, G. K. Statistical Applications in Genetics and Molecular Biology (2005)

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

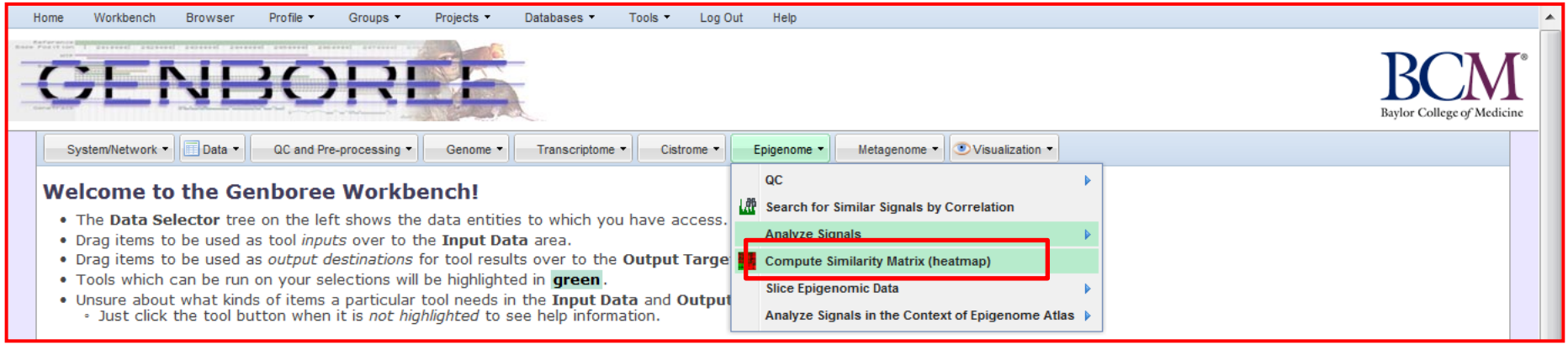
-Breast Luminal Epithelium
-Breast Myoepithelial
-Breast Stem Cell

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

-Breast Luminal Epithelium
-Breast Myoepithelial
-Breast Stem Cell



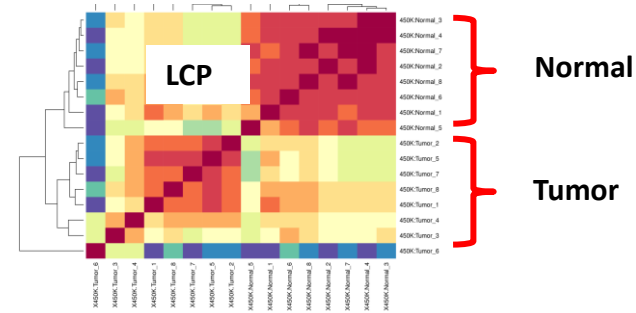
Use Case 13: Analysis of epigenomic variation in breast tumors (Illumina 450K)

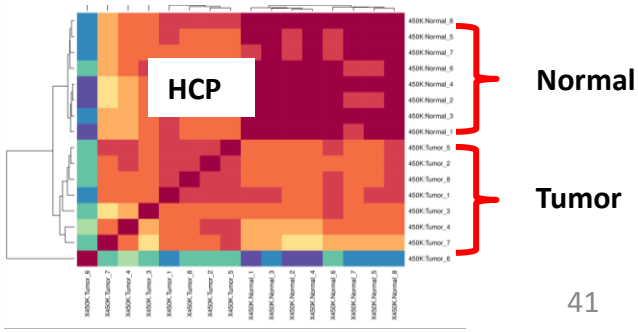


<u>Use Case 13a:</u> Cluster all 16 breast tissue samples

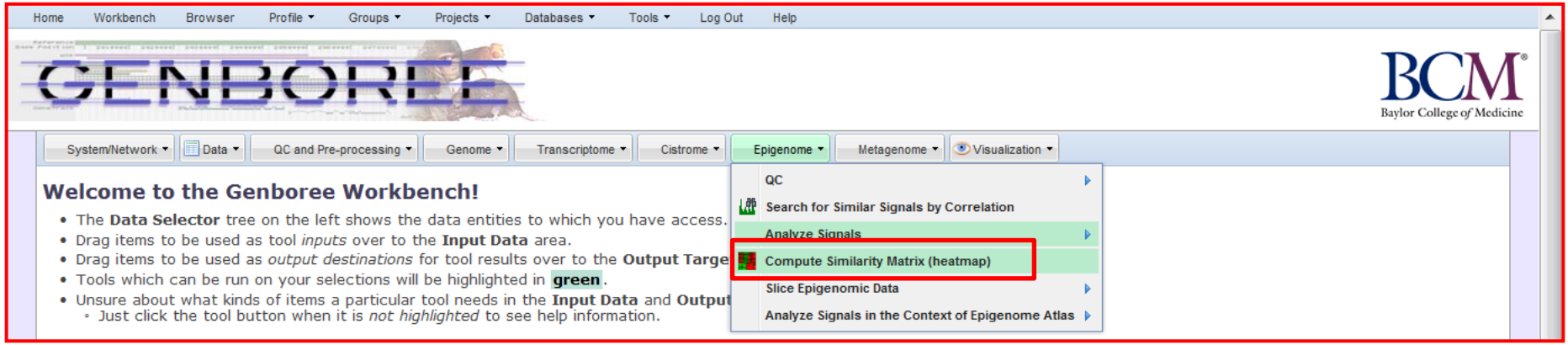
16 450 K Samples (Dedeurwaerder, S.et al. (2011)
-8 normal breast samples
-8 cancerous breast samples

ROIs (HCP vs LCP)





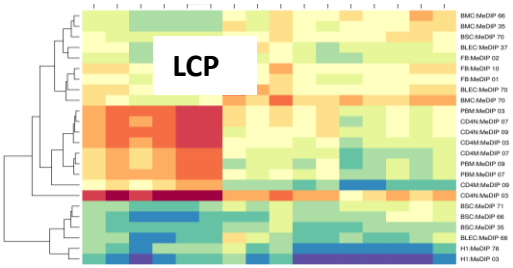
Use Case 13: Epigenomic variation in breast tumors

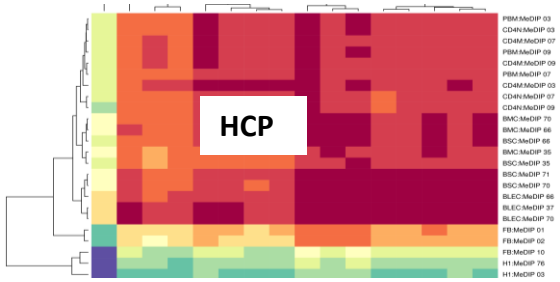


<u>Use Case 13b:</u> Compare 450K profiles (8 tumor, 8 normal) *against reference epigenomes* from the Epigenome Atlas

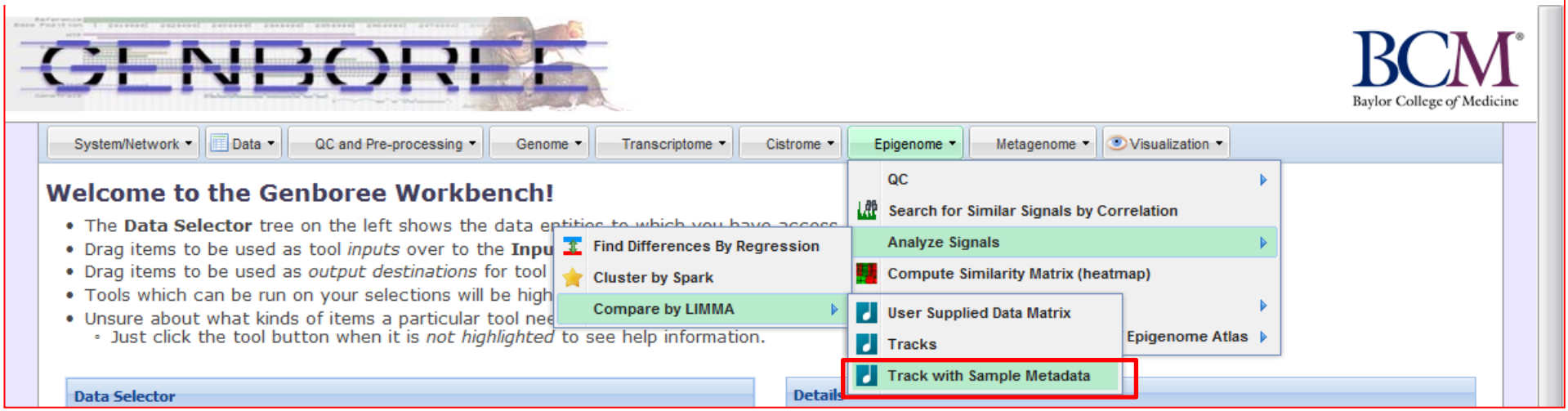
16 450 K Samples (Dedeurwaerder, S.et al. (2011)
-8 normal breast samples
-8 cancerous breast samples

ROIs





Use Case 13: Epigenomic variation in breast tumors



<u>Use Case 13c:</u> 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

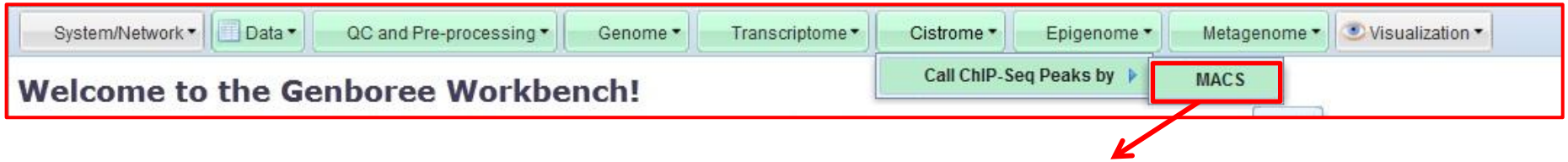
16 450 K Samples (Dedeurwaerder)
-8 normal breast samples
-8 cancerous breast samples

ROIs

Gene List

	B
1	PAQR7
2	VGLL4
3	TESK1
4	FGFR1
5	UCKL1
6	MGC16121
7	LCAT
8	SLC12A4
9	MTSS1L
10	TICAM1

Use Case 14: Chip-Seq and RNA-Seq Data Analysis



Zhang et al, *Genome Biology* (2008)

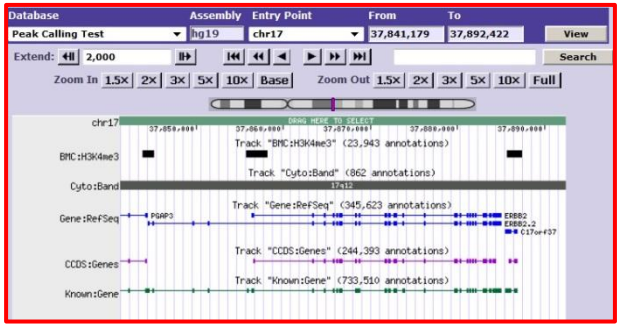
BED files →

MACS results (file in Genboree)

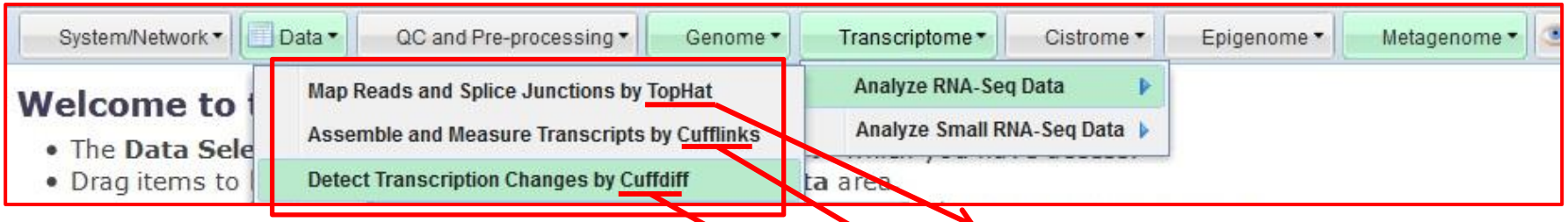
20	chr	start	end	length	summit	tags	-10*LOG1	fold_enrichment
21	chr1	9861	10677	817	360	292	2034.63	24.91
22	chr1	713307	715489	2183	1297	83	250.89	11.7
23	chr1	724760	727160	2401	2202	92	223.86	16.37
24	chr1	761468	763266	1799	865	133	463.34	14.42
25	chr1	833096	834002	907	577	20	96.47	11.54
26	chr1	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 14: Chip-Seq and RNA-Seq Data Analysis



Trapnell et al, *Bioinformatics* (2009)

Trapnell et al, *Nature Biotech* (2010)

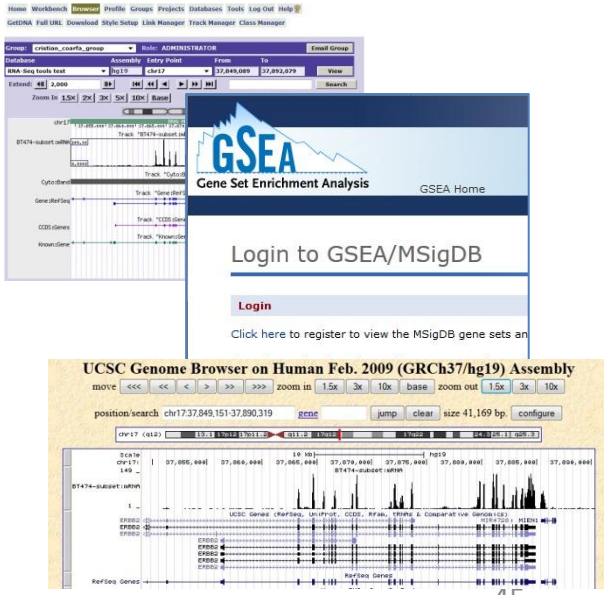
Trapnell et al, *Nature Biotech* (2010)

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

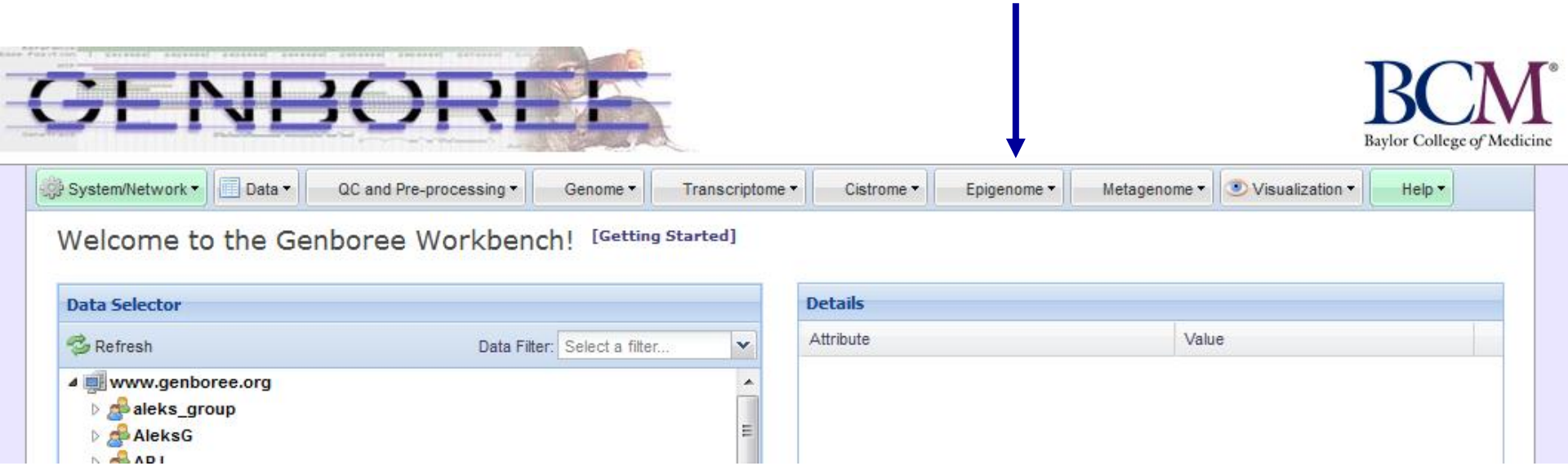
Visualization/pathway analysis



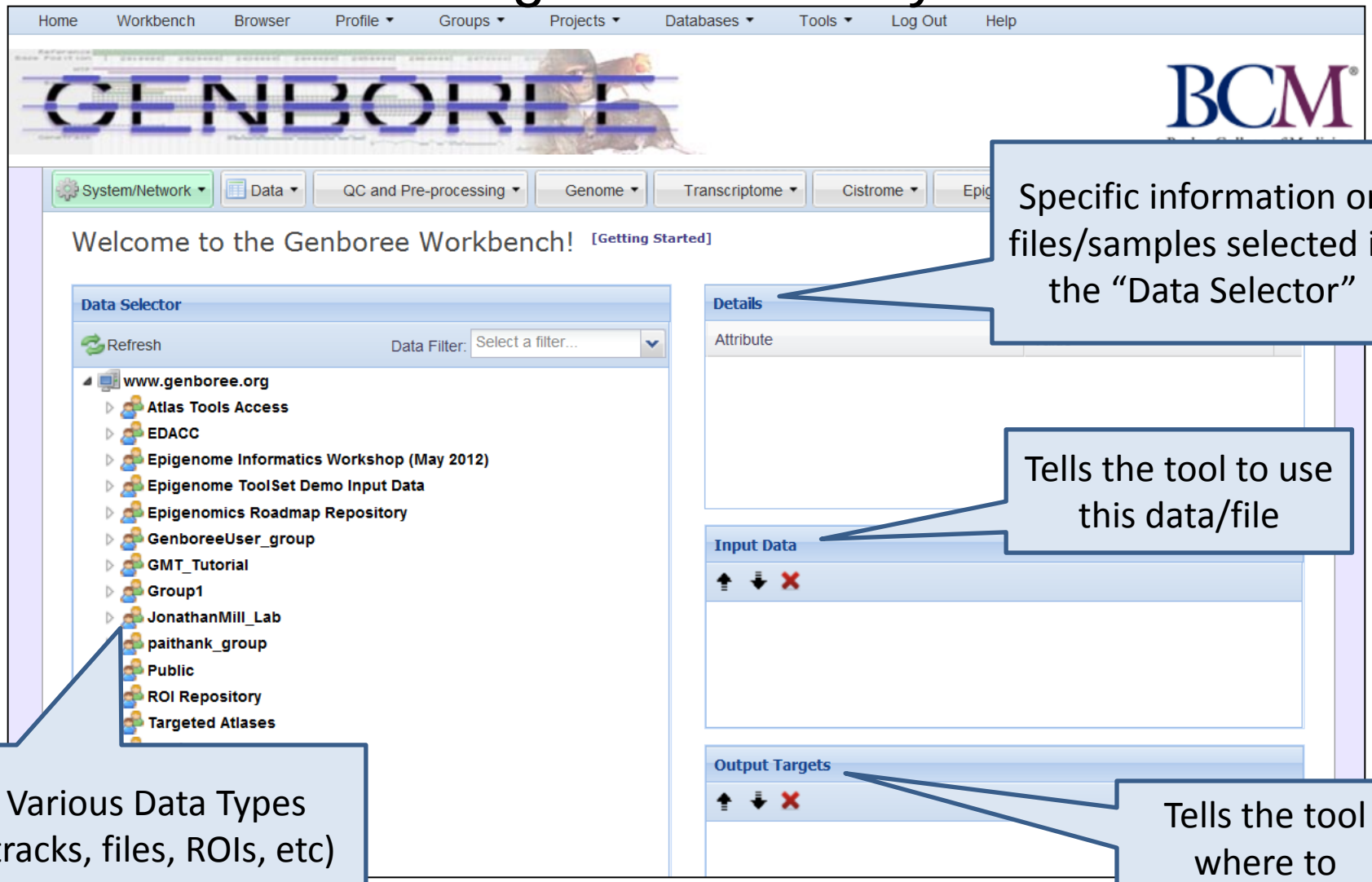
- NIH Roadmap Epigenomics Project
- Epigenome Atlas Analysis
- Genboree Workbench
- Genboree Network

Epigenomic Toolset integrates Spark developed by the British Columbia Genome Center in Vancouver

Spark:
Nielsen CB et al.Genome Res. (11):2262-9 2012



The Genboree Workbench: Web-based Data Management & Analysis



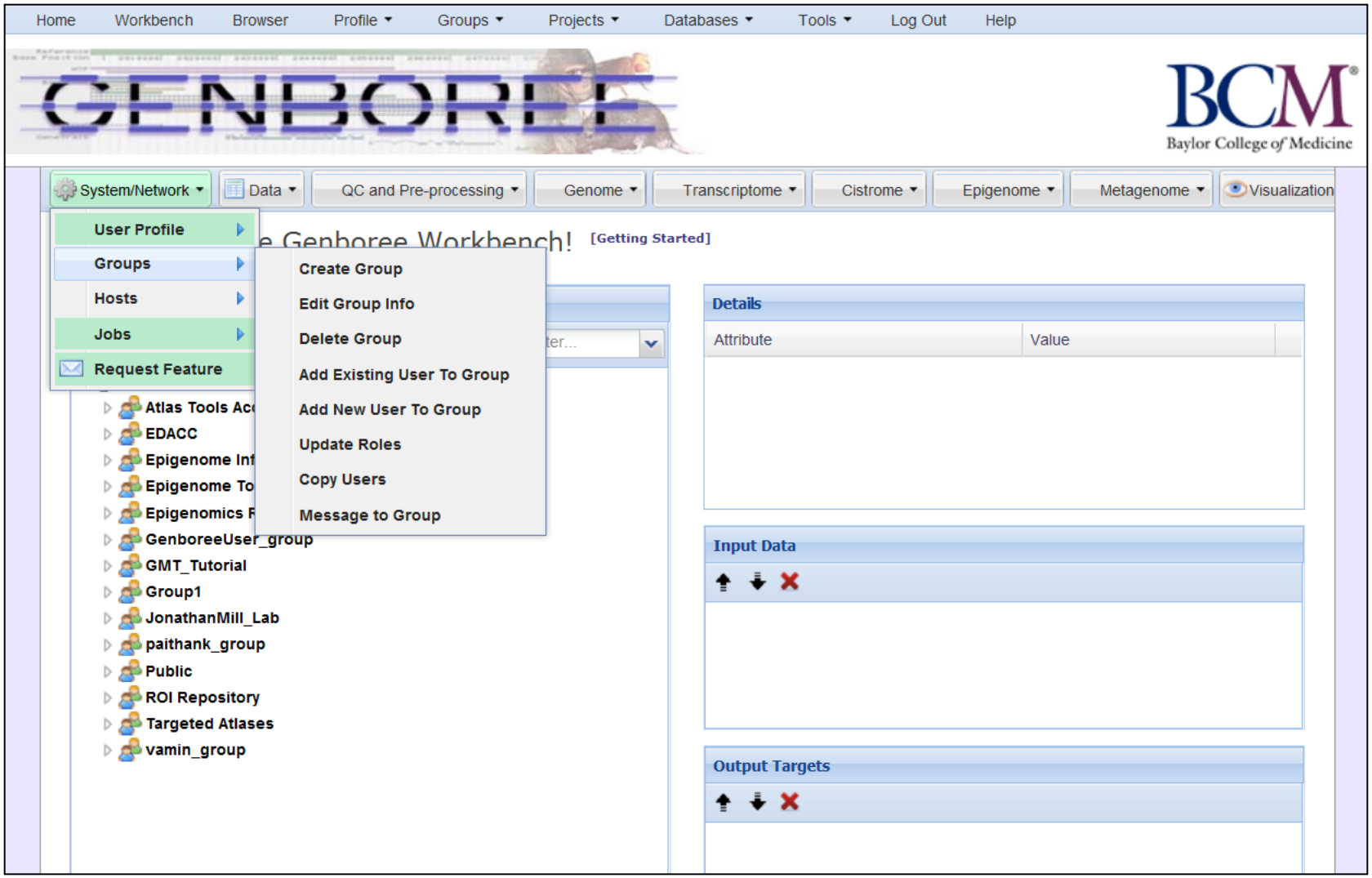
Specific information on files/samples selected in the "Data Selector"

Tells the tool to use this data/file

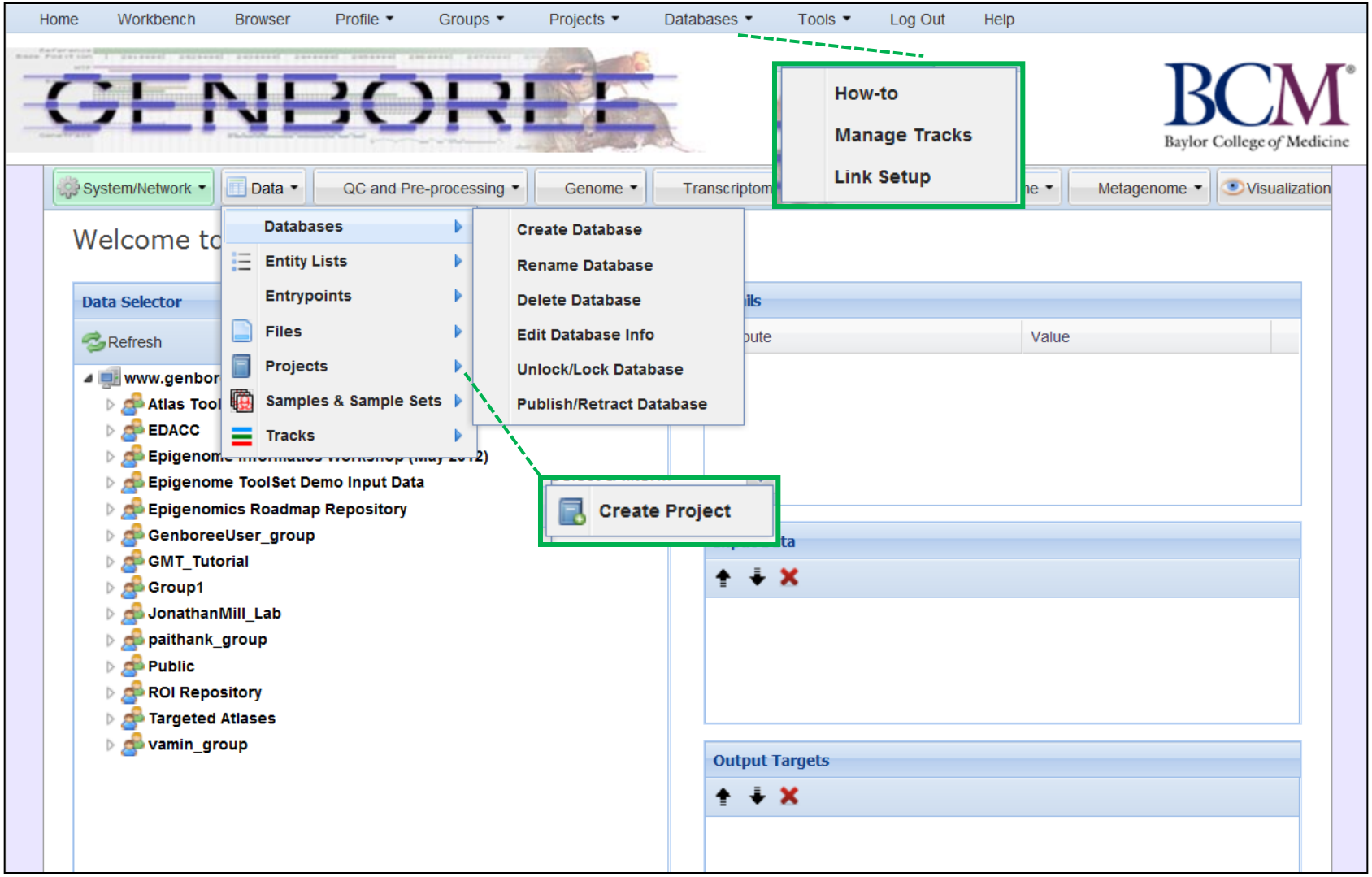
Various Data Types (tracks, files, ROIs, etc)

Tells the tool where to deposit results

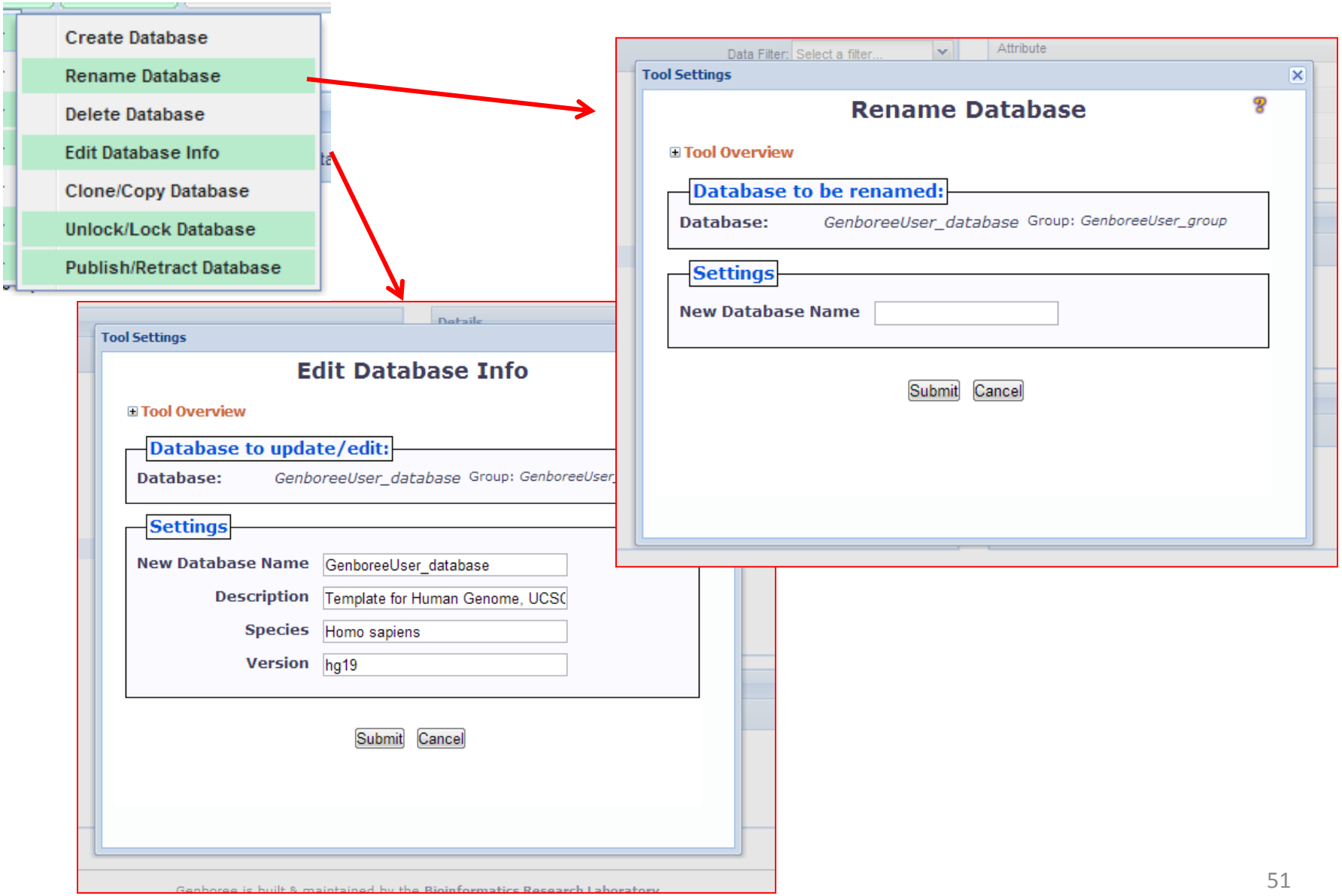
Genboree Workbench: Create & Manage Collaborations



Create & Manage Databases & Projects – Share Data

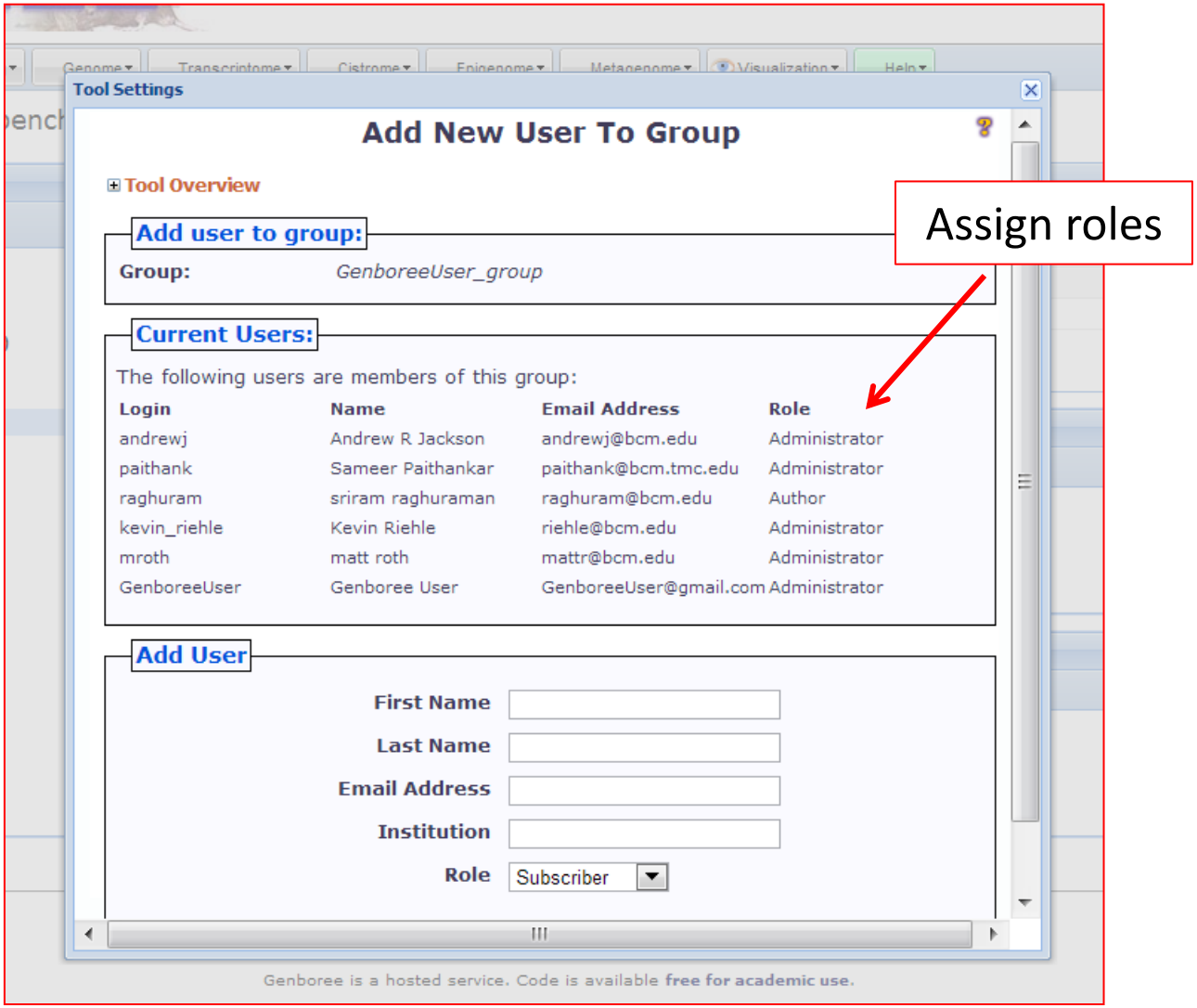


Manage Databases



Authorization: Decide How to Share Data

Share data with anyone with a browser; assign level of access



Data Sharing Option 1: Projects

Project Pages (below) and via the Genboree Workbench (next slide)

GENBOREE

BCM
Baylor College of Medicine

Home | Workbench | Browser | Profile | Groups | Projects | Databases | Tools | Log Out | Help

Edit Mode

Tutorial data

Project page for managing Genboree Microbiome Toolset results

Project News:

2013/2/24: Matt Roth ran a Alpha Diversity job (AD-Job-2013-02-24-15:35:21) and the results are available at the link below.

- **Study Name**: AlphaDiv_GMT_Tutorial
- **Job Name**: AD-Job-2013-02-24-15:35:21
- **Link to result plots** ← **access data**

2013/2/24: Matt Roth ran a Machine Learning job (ML-Job-2013-02-24-15:36:38) and the results are available at the link below.

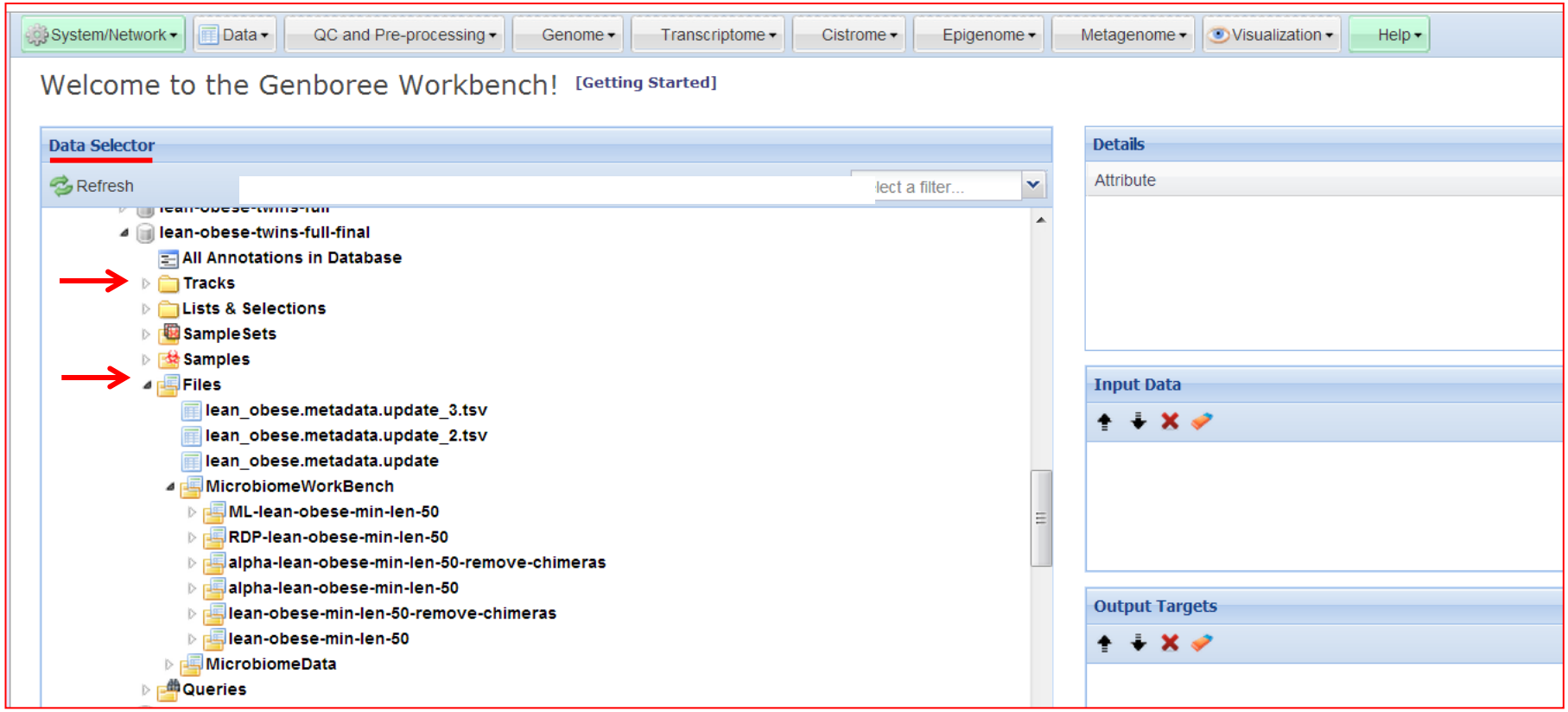
- **Study Name**: MachLearn_GMT_Tutorial
- **Job Name**: ML-Job-2013-02-24-15:36:38
- **Link to result plots** ←

2013/2/24: Matt Roth ran a QIIME job (Qiime-Job-2013-02-24-15:03:01) and the results are available at the links below.

- **Study Name**: QIIME_Tutorial
- **Job Name**: Qiime-Job-2013-02-24-15:03:01
- **Link to cdhit results**
- **Link to cdhit-normalized results** ←

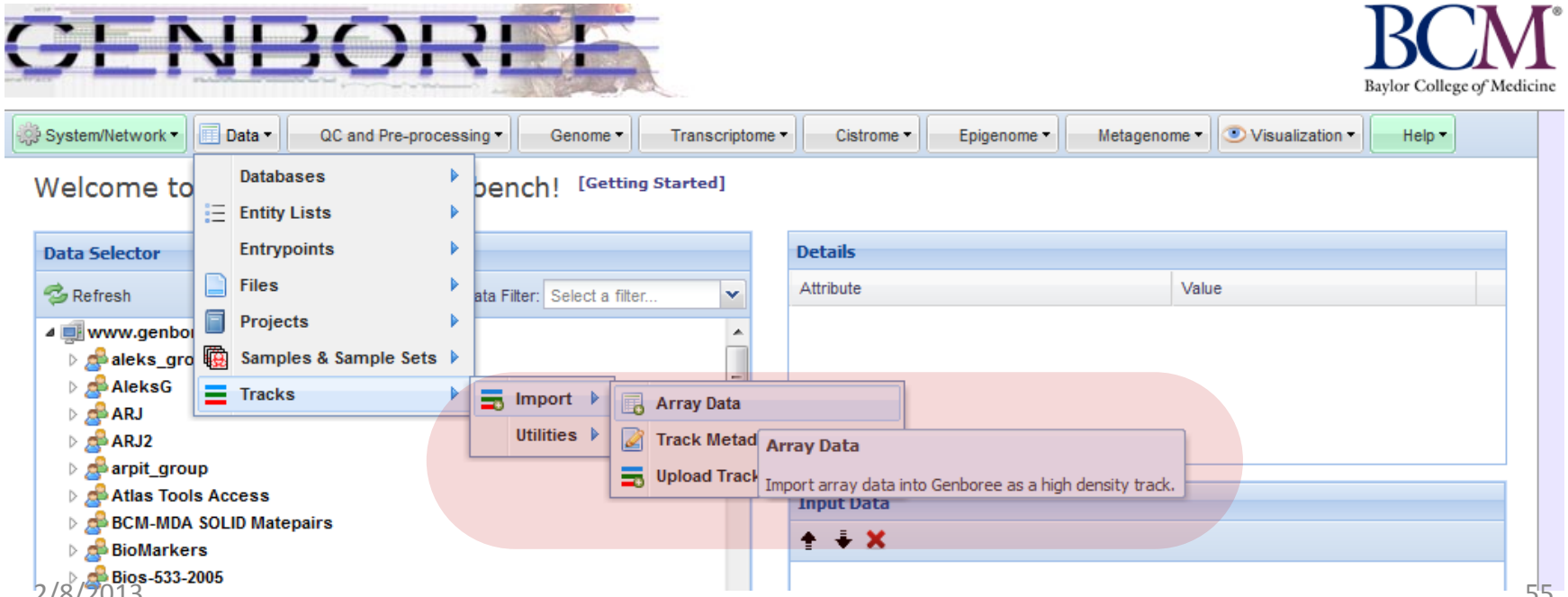
Data Sharing Option 2: Databases, tracks, files, etc.

Files/data/tracks accessible via the Data Selector in the Genboree Workbench



Analysis of private data enabled by group-level access control

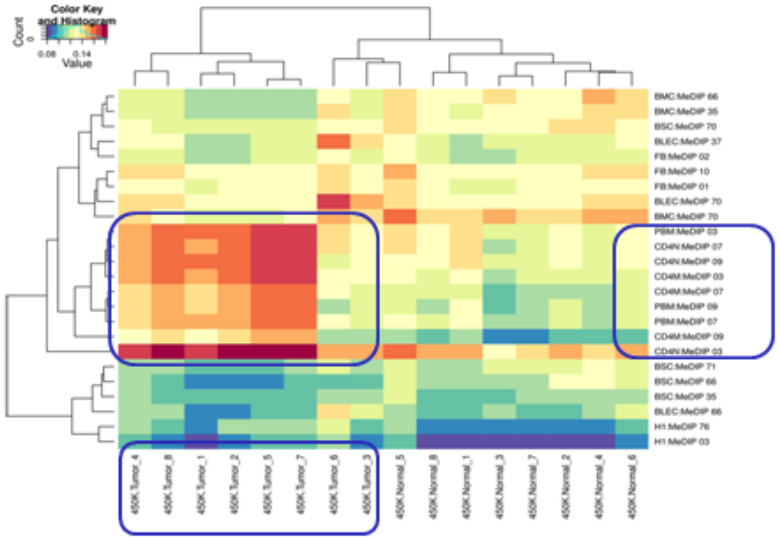
Illumina 450K array profiles can be compared against methyloms in the Epigenome Atlas



Identifying cell type composition of tumors

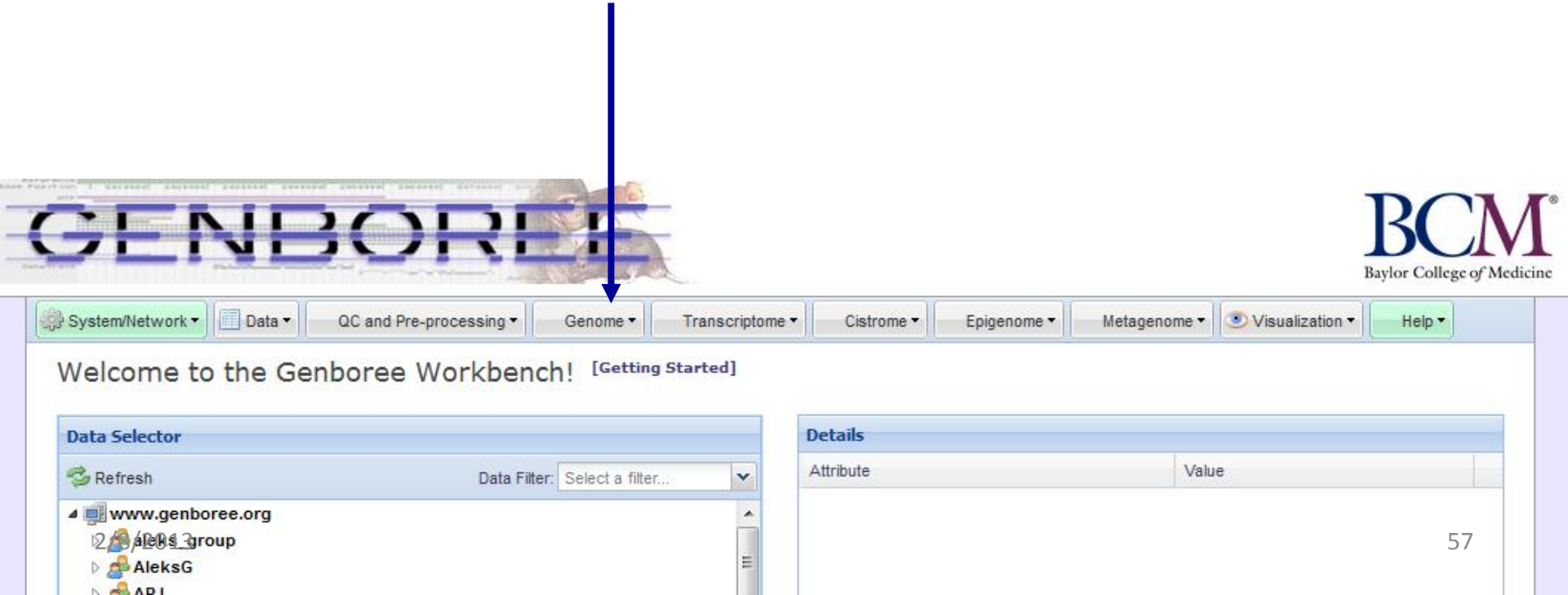
450K array profiles of breast tumors and adjacent normal tissue by Dedeurwaerder, S.et al. (2011) Epigenomics 3(6)

Reference epigenomes: MeDIP-seq data from Human Epigenome Atlas, contributed by the UCSF-UBC REMC



Genomic Toolset developed in collaboration with the Baylor Genome Center

Atlas2 genome resequencing:
Evani US et al. BMC Genomics 6:S19 2012



Other widely used tools are integrated within the Genboree Workench

- MACS
- TopHat
- CuffLiks
- CuffDiff.
- ….

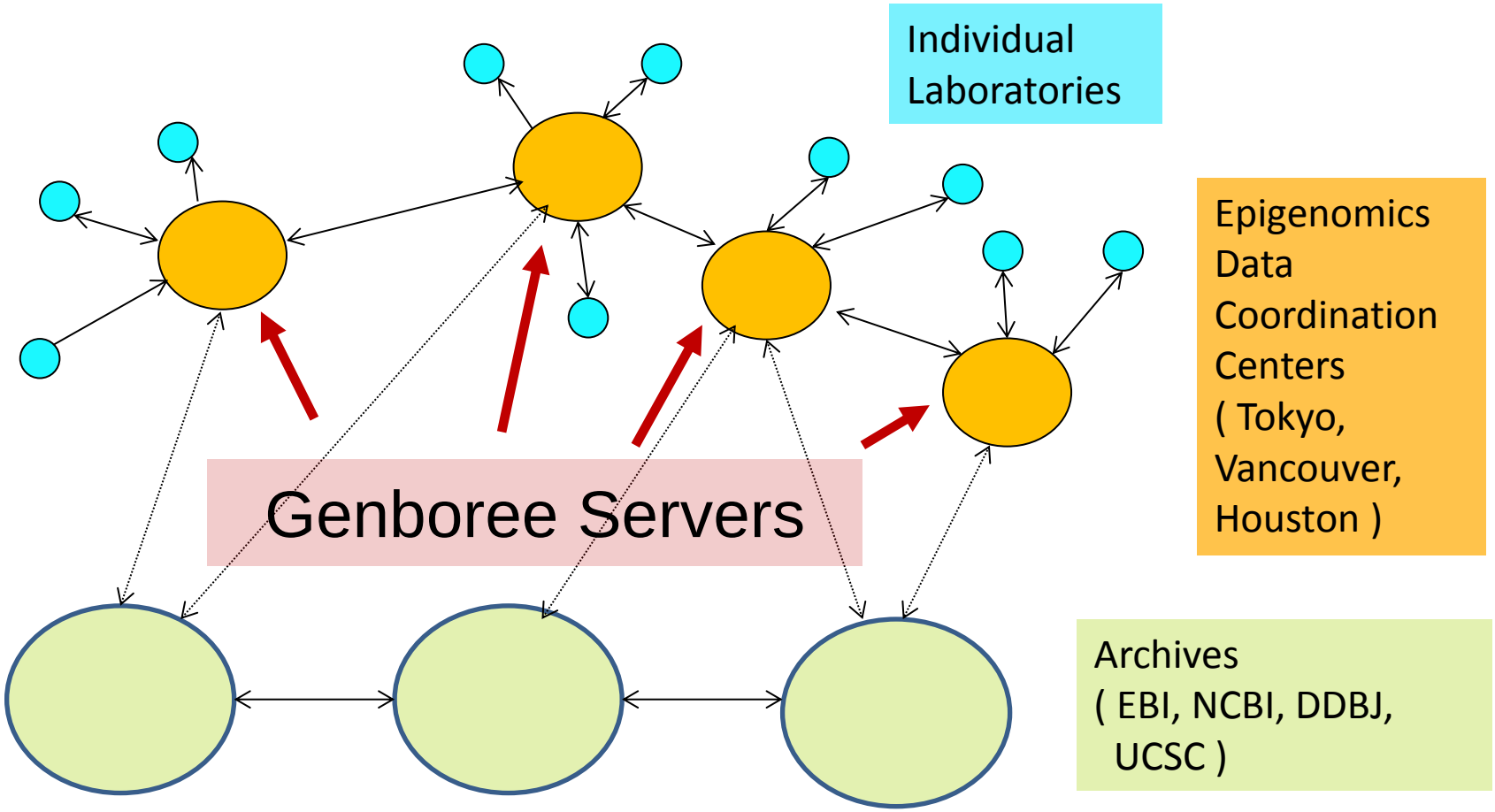
- NIH Roadmap Epigenomics Project
- Epigenome Atlas Analysis
- Genboree Workbench
- Genboree Network

As the data volume explodes, ***physical*** data integration will become impossible

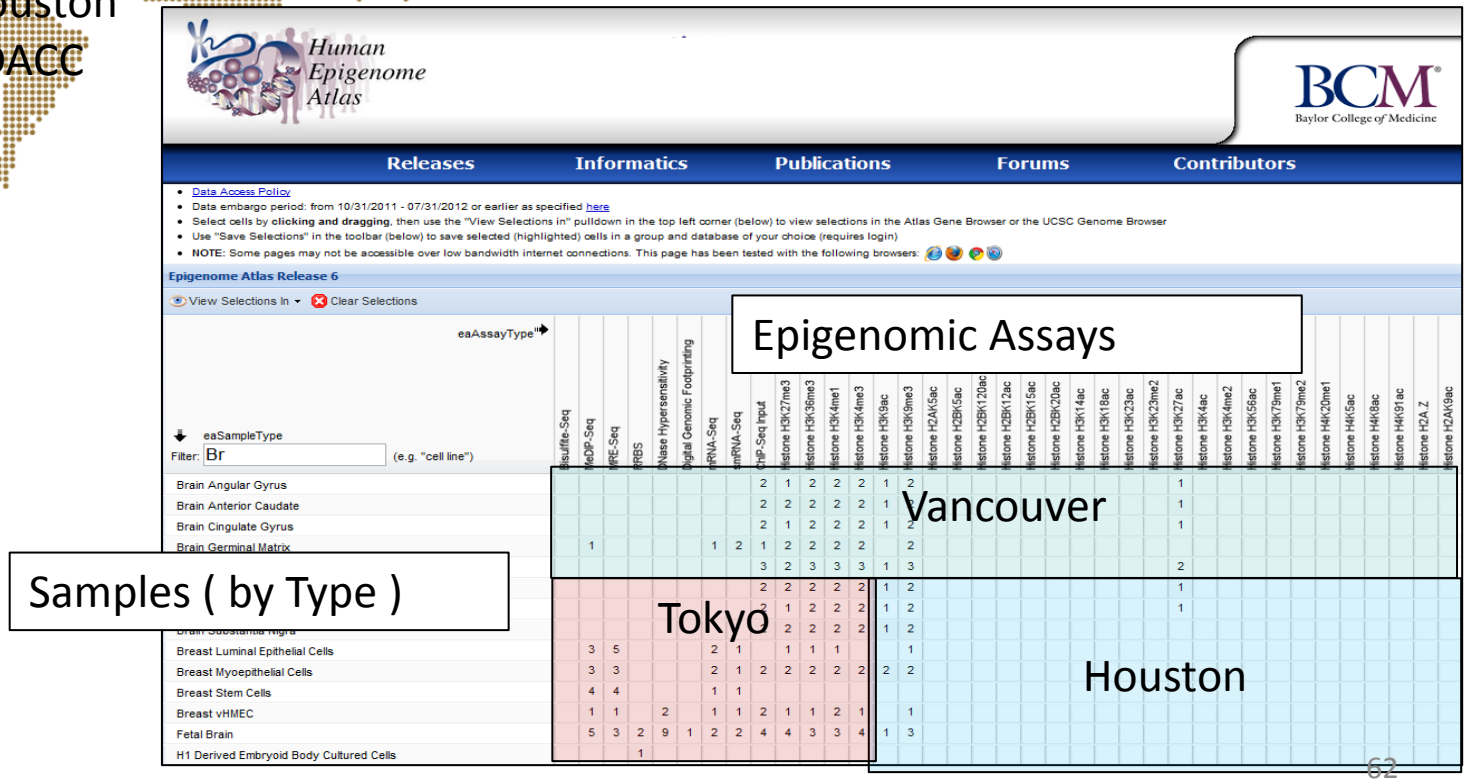
Solution: ***virtual*** data integration

Data Ecosystem Model proposed for IHEC

IHEC: International Human Epigenome Consortium

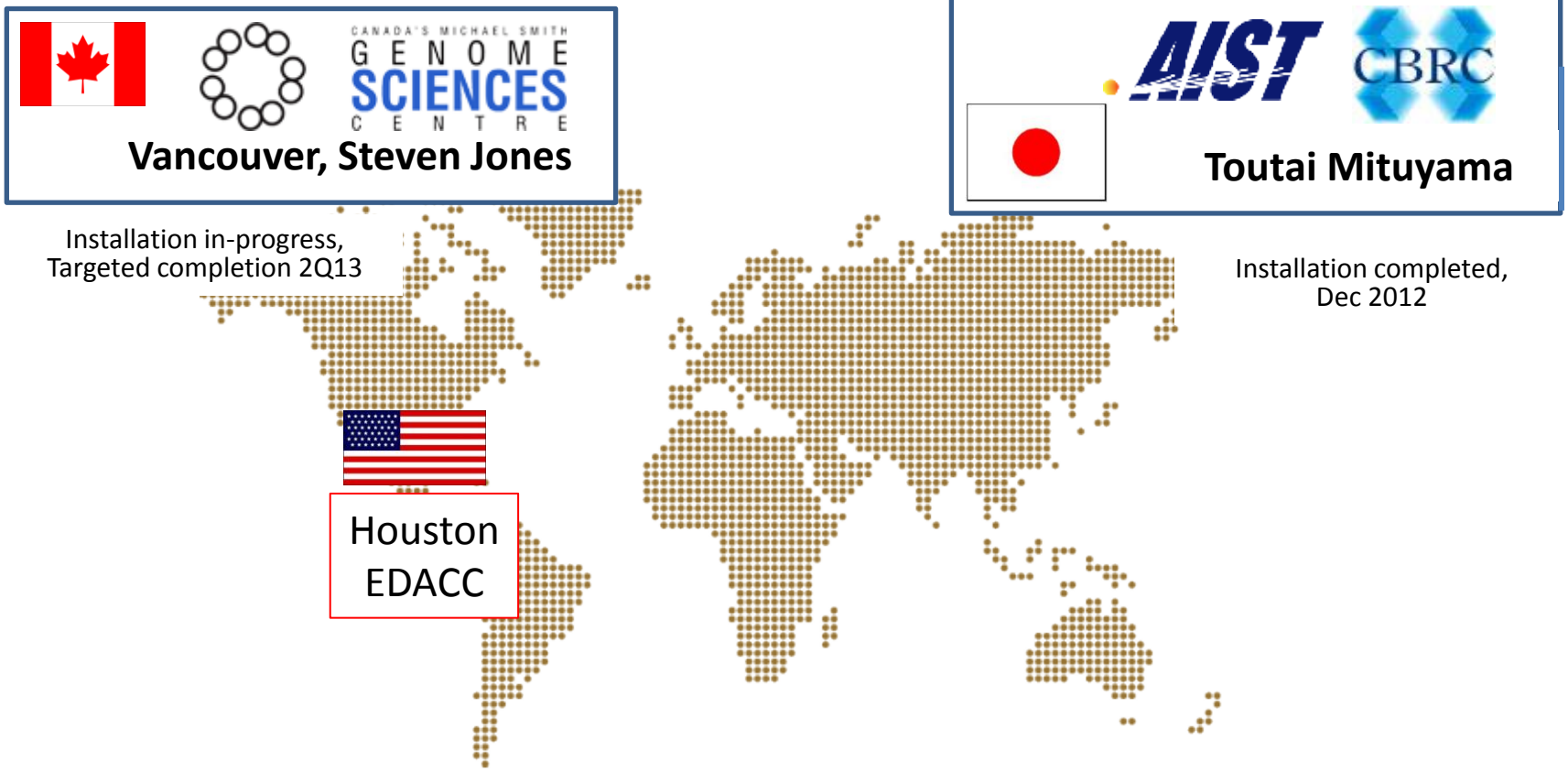


Virtual integration of reference epigenomes across three IHEC Data Centers

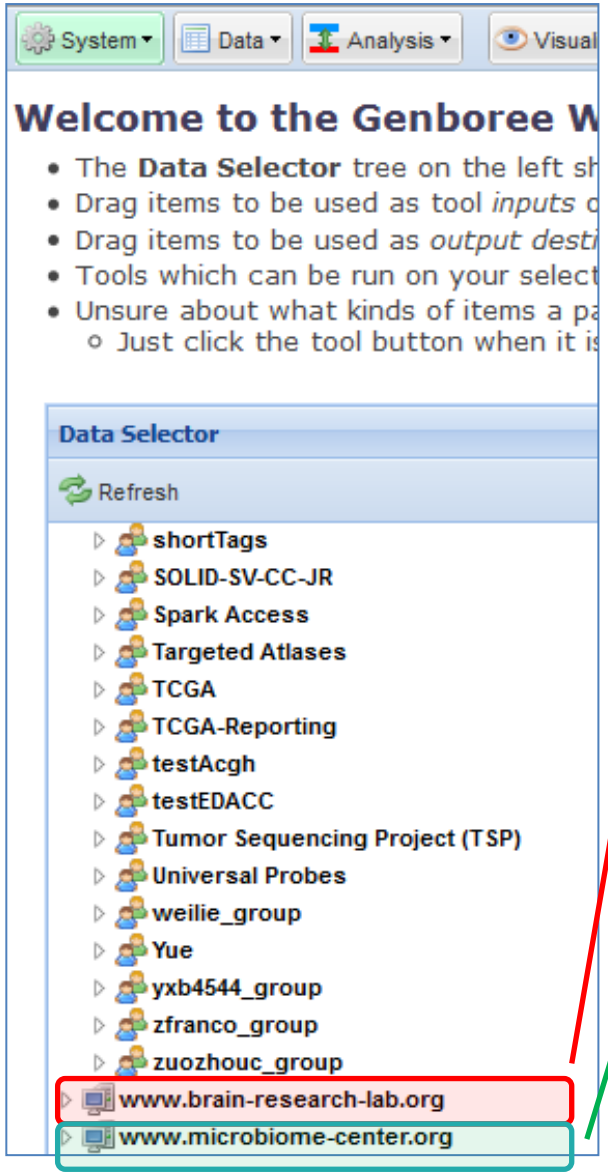


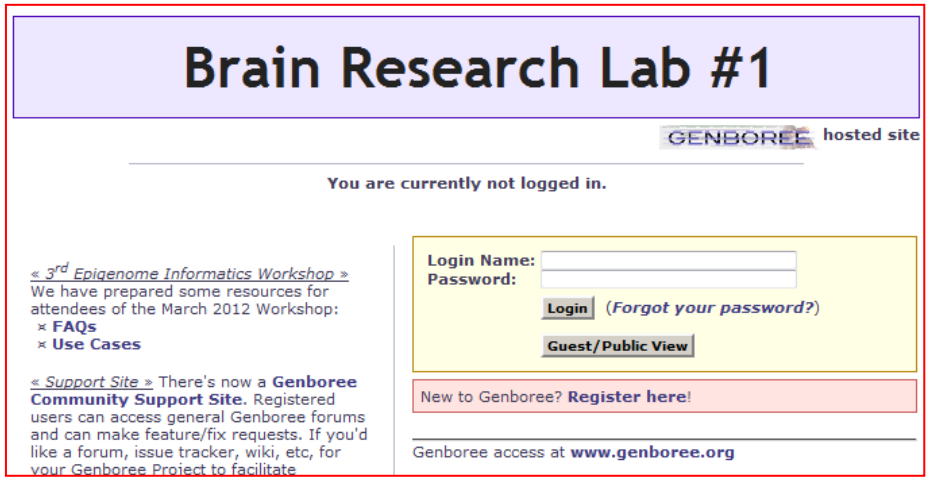
Genboree Installations

Other locations that use Genboree



Genboree on the Commercial Cloud (Rackspace)





A combination of dedicated hosting and elastic cloud computing accessible via the Genboree Workbench

- NIH Roadmap Epigenomics Project
- Epigenome Analysis
- Genboree Workbench
- Genboree Network

Workshop Evaluation (link)