

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**

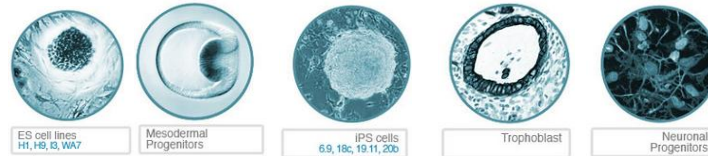
- NIH Roadmap Epigenomics Project
- Epigenome Analysis
- Genboree Workbench
- Genboree Network

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

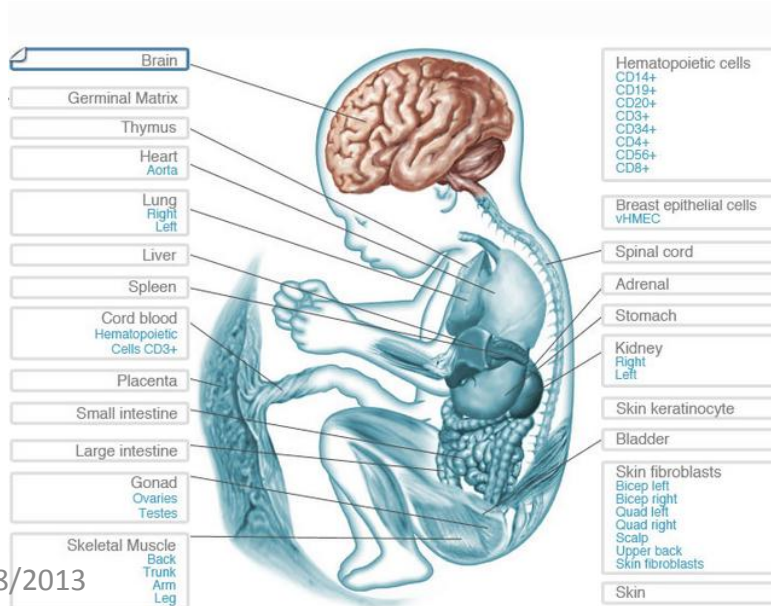
Epigenomic profiling of human cell-lines, primary cells, and tissues



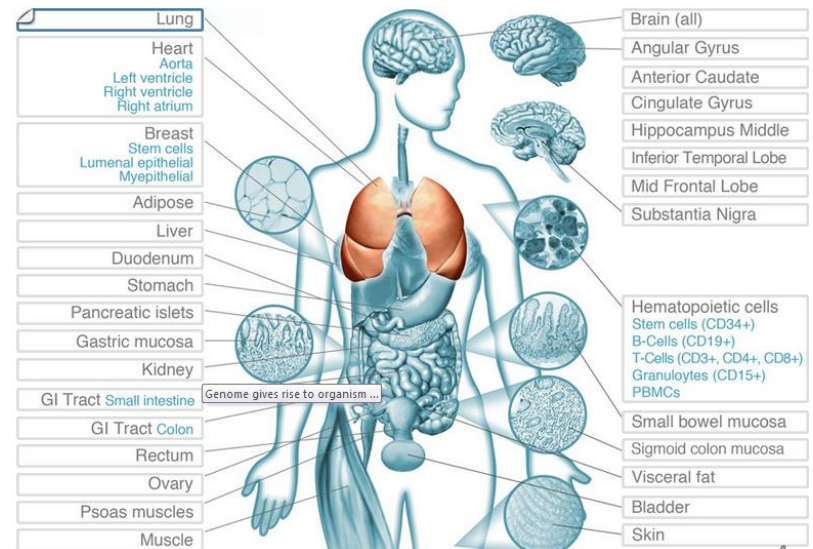
Stem Cells



Fetal



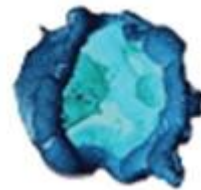
Adult



Mapping epigenomic differences between cell types



Stem Cell



Erythroid



Structural



Chondrocyte
Osteoblast

Hematopoiesis

B-cells
Macrophages
T-cells



Signaling

Beta cell



Myogenesis

Cardiomyocyte
Smoothmuscle



Neural

Neuron
Astrocyte
Oligodendrocyte



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

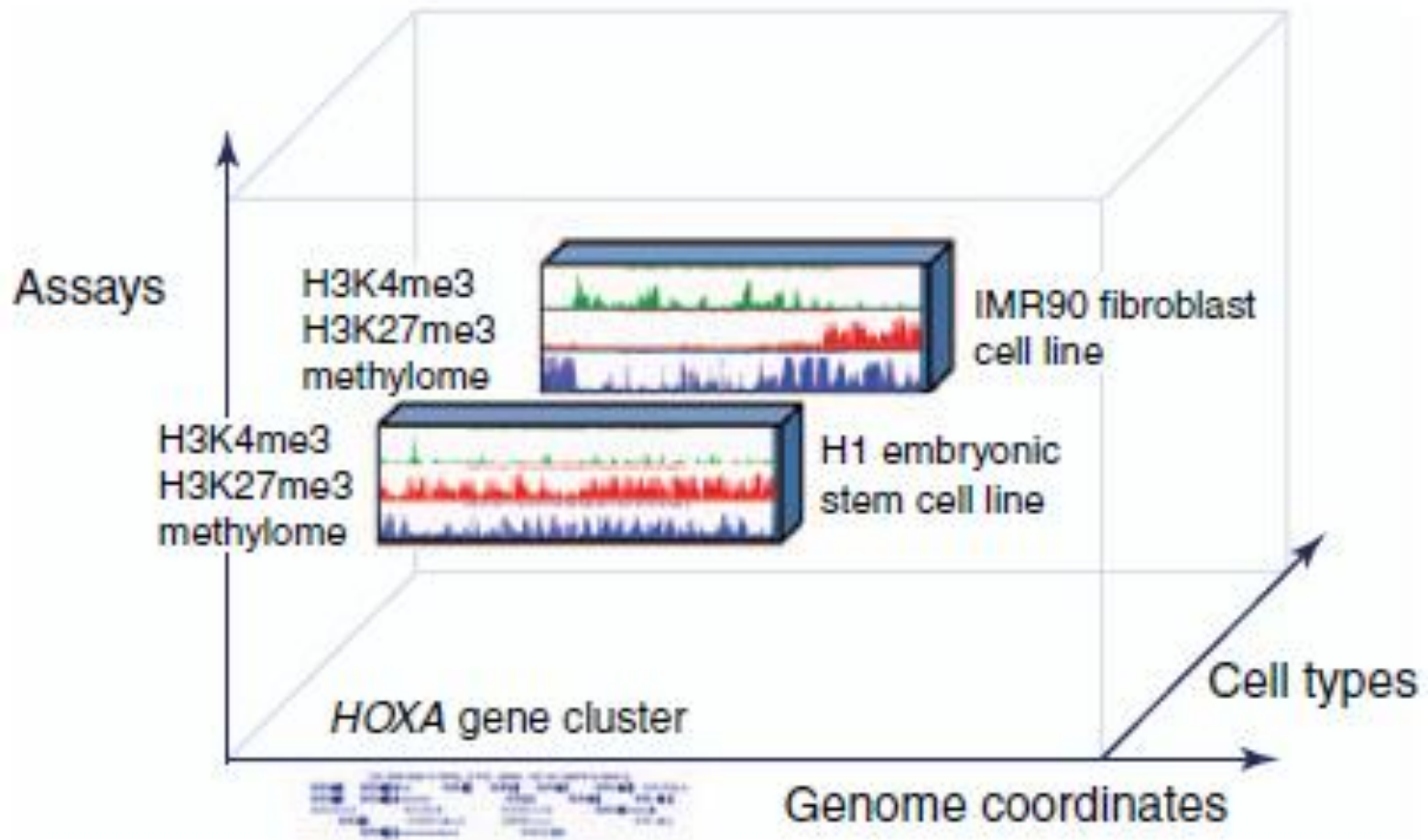
- SNP genotyping
- Resequencing from WGBS reads

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	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

Human Epigenome Atlas [Release 4](#) (hg19)

- [Data Access Policy](#)
- Data embargo period: from 04/14/2011 - 01/14/2012 or earlier as specified [here](#)
- Select cells by clicking and dragging, then use the "View Selections in" pulldown in the top left corner (below) to
- To see data authors, other metadata, and to download data, click a sample name in the first column or an assay type
- Human Epigenome Atlas releases are intended to be cumulative: e.g. Release 3 Includes all Release 2 data and a
- **NOTE:** Some pages may not be accessible over low bandwidth Internet connections. This page has been tested w

Human Epigenome Atlas Release 4 (hg19)

View Selections In

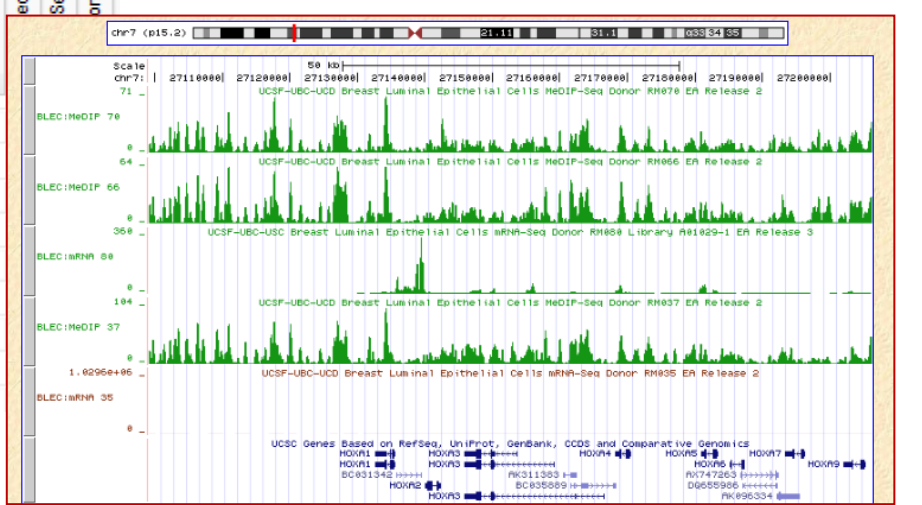
Atlas Gene Browser

Genome Browser

- Local UCSC browser mirror (Fast)
- UCSC genome browser (Slow)

Sample Filter: (e.g. "cell line")

Sample	Bisulfite-Seq	MeDIP-Seq	MRE-Seq	RRBS	DNAse-seq	Hypersensitivity	Digital Genomic Footprinting
Brain and Fetal EBEC							
Brain Substantia Nigra							
Breast Luminal Epithelial Cells		4	5				
Breast Myoepithelial Cells		3	3				
Breast Stem Cells		4	4				
Breast vHMEC		1	1				2
CD14 Primary Cells							2
CD15 Primary Cells				1			
CD19 Primary Cells				1	3		
CD20 Primary Cells					1		



Including More Genes in the Same Pathway

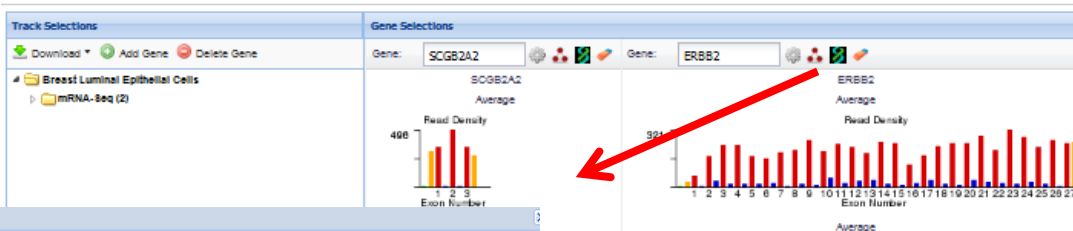
Atlas Gene Browser for EA Release 5 (hg19)

- Start typing a gene name in the box below and select a gene from the pulldown list
- Use the scrollbar at the bottom of the table to see the genes on the right
- Choose specific gene elements if necessary (🔍 next to gene name)
- Add genes one by one or via the pathway browser (🔗 next to gene name)
- Zoom into features of interest using the genome browser (🔍 next to gene name)
- Click on the sample, assay or gene image to see related metadata

Click on this icon to generate a shareable link to this session

Gene Features Legend

- Promoter (2Kbp upstream of TSS)
- Exon
- S'UTR / 3'UTR
- Intron
- ✗ No BioAffix Med Data



Pathways associated with the gene "ERBB2":

- Cellular Processes :
 - Cell Communication :
 - "Adherens junction" (Kegg Pathway Id: hsa04520) Pick genes
 - "Focal adhesion" (Kegg Pathway Id: hsa04510) Pick genes
 - Environmental Information Processing :
 - Signal Transduction :
 - "Calcium signaling pathway" (Kegg Pathway Id: hsa04020) Pick genes
 - "ErbB signaling pathway" (Kegg Pathway Id: hsa04012) Pick genes
 - Human Diseases :
 - Cancers :
 - "Bladder cancer" (Kegg Pathway Id: hsa05219) Pick genes
 - "Endometrial cancer" (Kegg Pathway Id: hsa05213) Pick genes
 - "Non-small cell lung cancer" (Kegg Pathway Id: hsa05223) Pick genes
 - "Pancreatic cancer" (Kegg Pathway Id: hsa05212) Pick genes
 - "Pathways in cancer" (Kegg Pathway Id: hsa05200) Pick genes
 - "Prostate cancer" (Kegg Pathway Id: hsa05215) Pick genes

Genes Selected:
of Genes: (0)

Choose from genes in a pathway

Note: This Is Wrapped Pathway Data
The pathway from Kegg has been wrapped by Genboree. The original Kegg pathway page contains useful links to their own database information about pathways, genes, etc.

Click a pathway **node** to select genes. Click 🔍 to save selections or click ✗ to discard changes. When a gene is saved, all nodes containing that gene will be **highlighted**. When you are finished choosing genes click 'Done' to return to the previous screen.

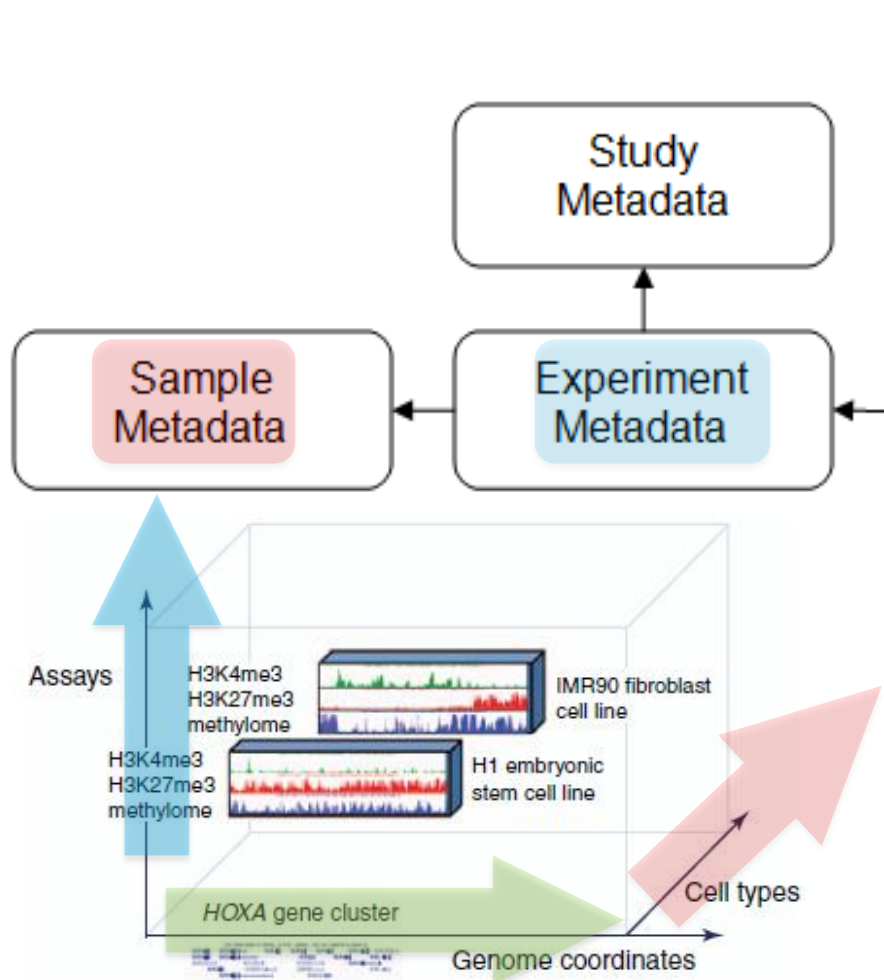
ERBB SIGNALING PATHWAY

The diagram shows the ERBB signaling pathway. Key nodes include ErbB-1 and ErbB-2 (highlighted in red), PLCγ, PKC, Cbl, FAK, and Abl. Ligands like EGF, TOFα, AR, and HER2/Neu are shown binding to ErbB-1 and ErbB-2. Downstream signaling involves PLCγ, PKC, Cbl, FAK, and Abl. Cellular targets and adhesion migration are also indicated.

Genes Selected:
of Genes: (0)

Done

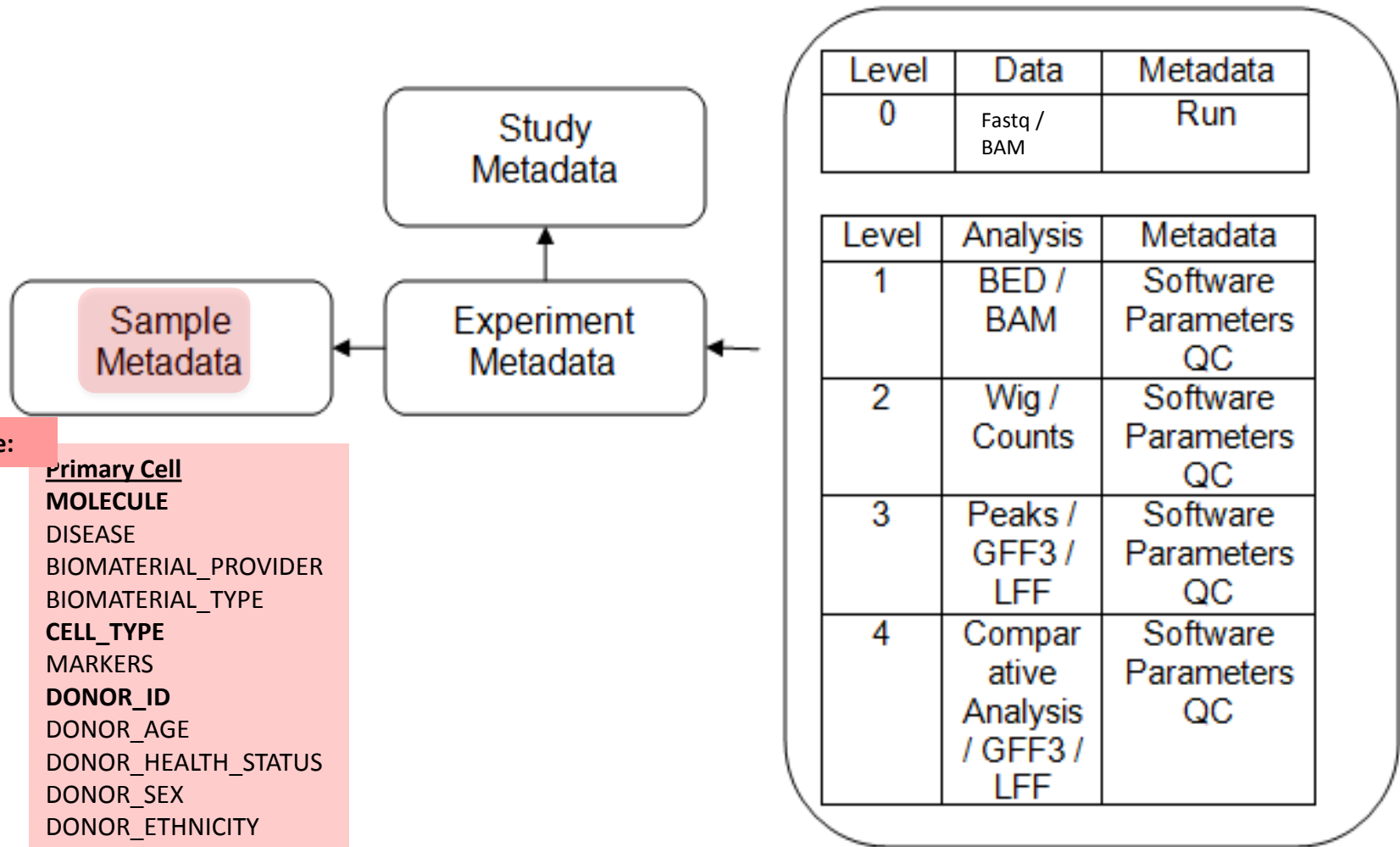
Epigenomic Metadata: SRA XML + Epigenomic Data Element Extensions



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

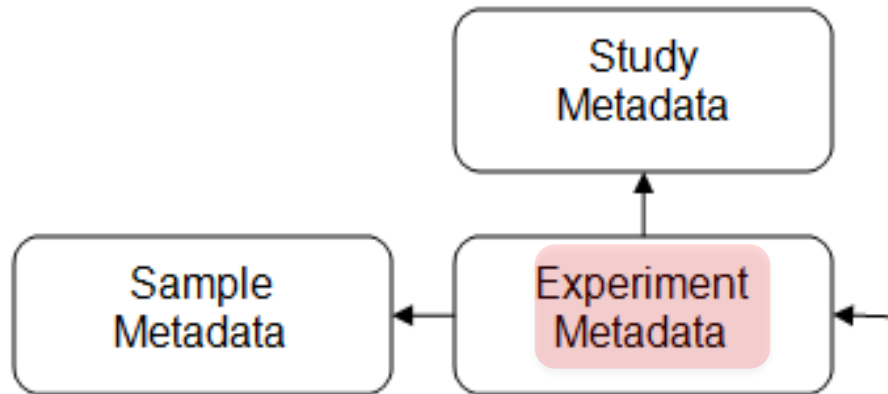
Sample Metadata



Example:

Primary Cell
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

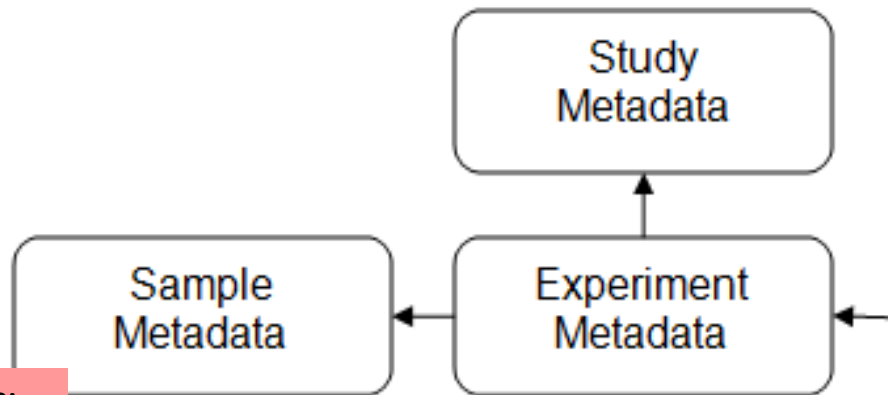
Level	Analysis	Metadata
1	BED / BAM	Software Parameters QC
2	Wig / Counts	Software Parameters QC
3	Peaks / GFF3 / LFF	Software Parameters QC
4	Comparative Analysis / GFF3 / LFF	Software Parameters QC

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)

Use this template for Illumina submissions if you used an array already represented in LitU (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.

("Study")

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 {diseaseState} {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")

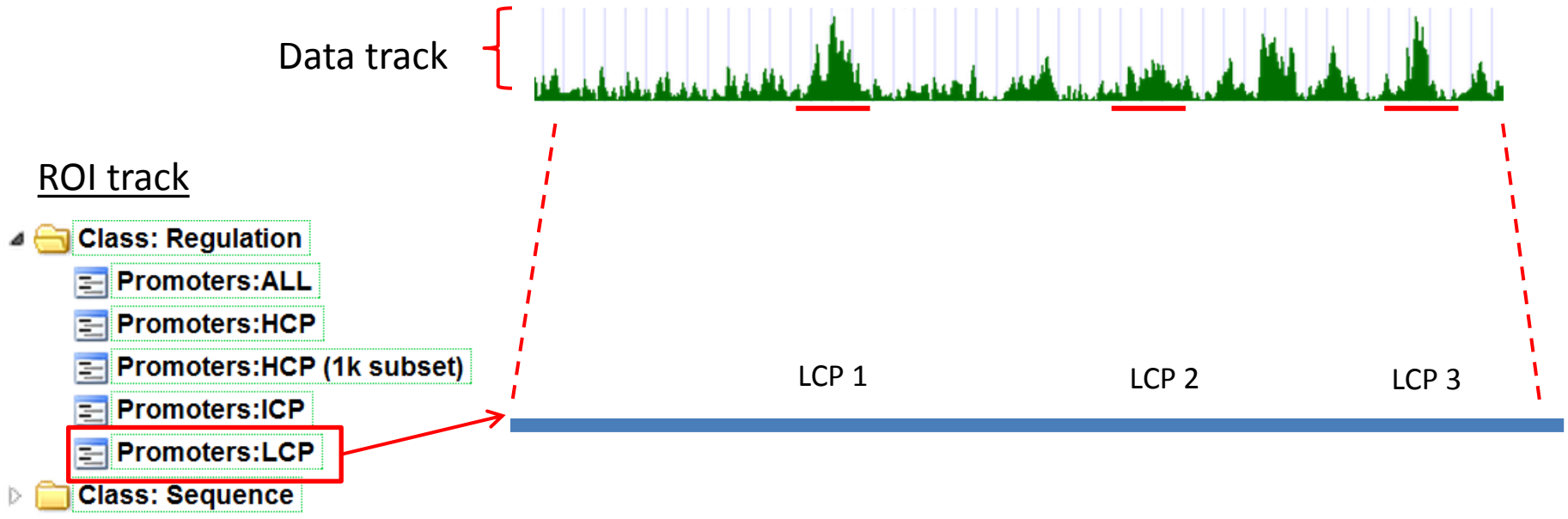
SAMPLES						
# The Sample names in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheets).						
# CLICK HERE to find the platform accession number (GPL:xxxx).						
# Note: As of 5/17/2010, the GEO database has two Platform records for Illumina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

("Experiment")

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 softw are version 2010.3.0.30128					
matrix processed value definition	normalized Average Beta					
# matrix signal intensities value defi	Unmethylated and methylated signal intensities					

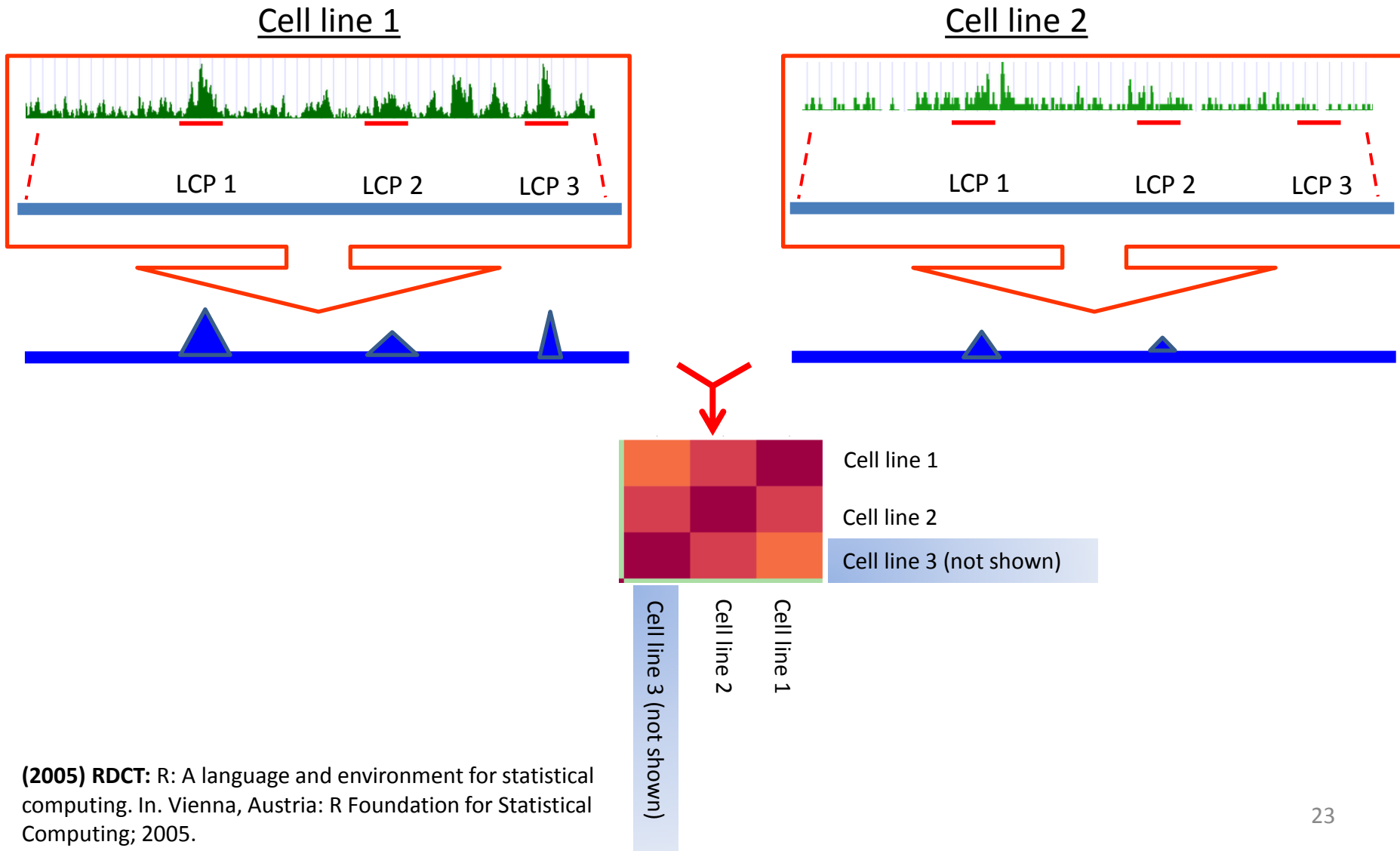
- NIH Roadmap Epigenomics Project
- **Epigenome Analysis**
- Genboree Workbench
- Genboree Network

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

The Toolset enables analyses of the Epigenome Atlas and of private data

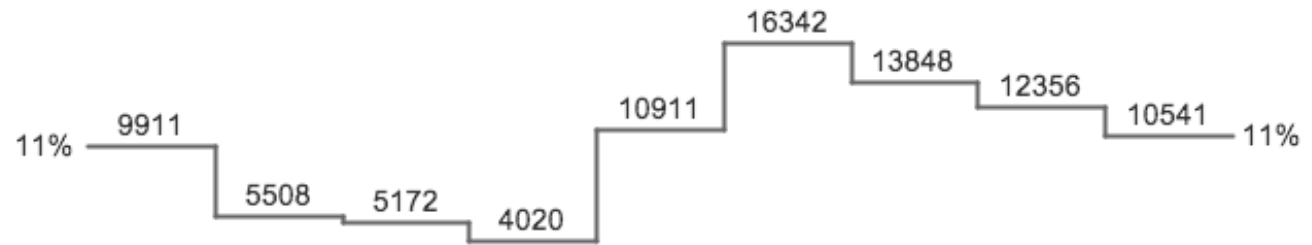
Generate heatmap clusterings and trees with different combinations of ROIs / marks / samples

Identify branch-specific epigenomic changes

- Analyze coordinated 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

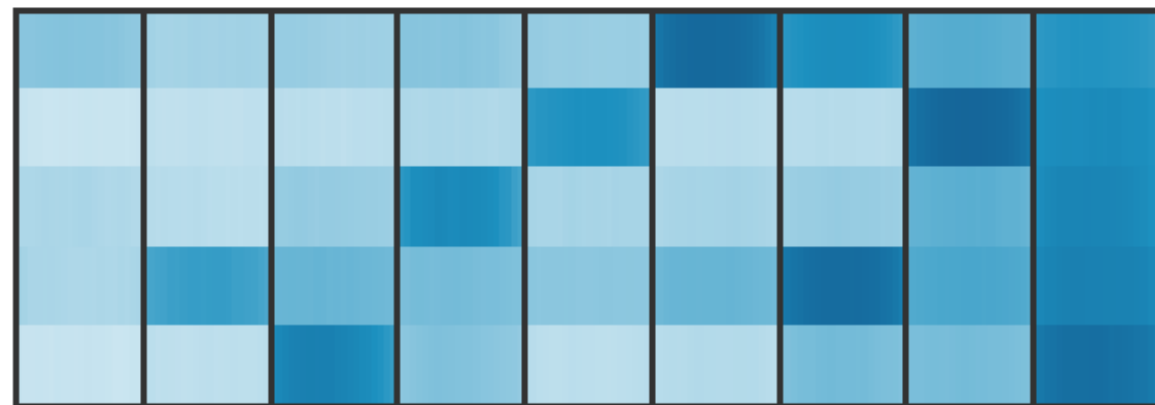
Identify enhancers with branch-specific activation patterns (H3K4me1 \uparrow)

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



Cluster profiles

Stem cells
Immune (CD34)
Mucosa
Neuronal
Smooth muscle

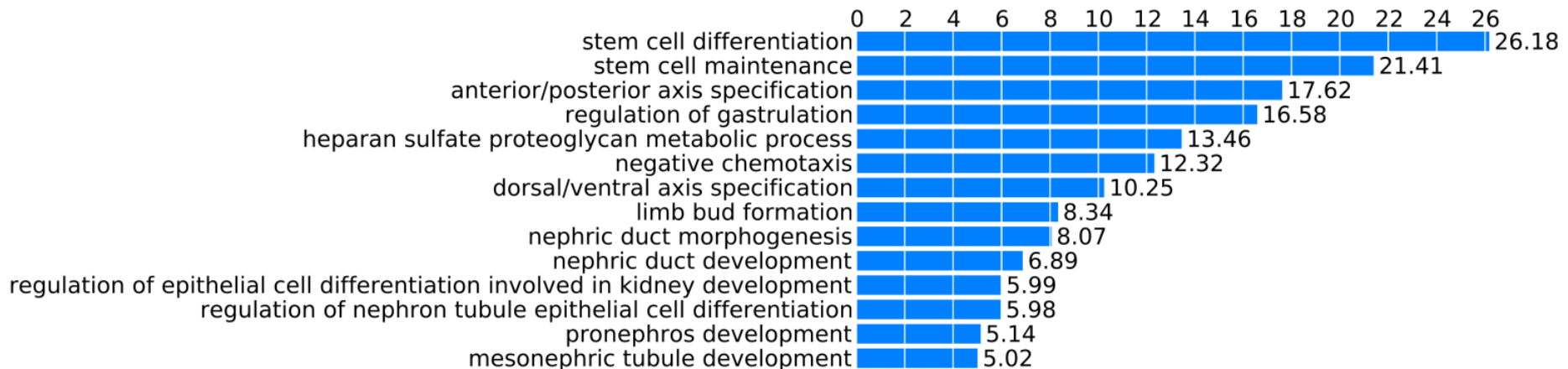


GO analysis of genes associated with enhancers activated in the stem cell branch

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

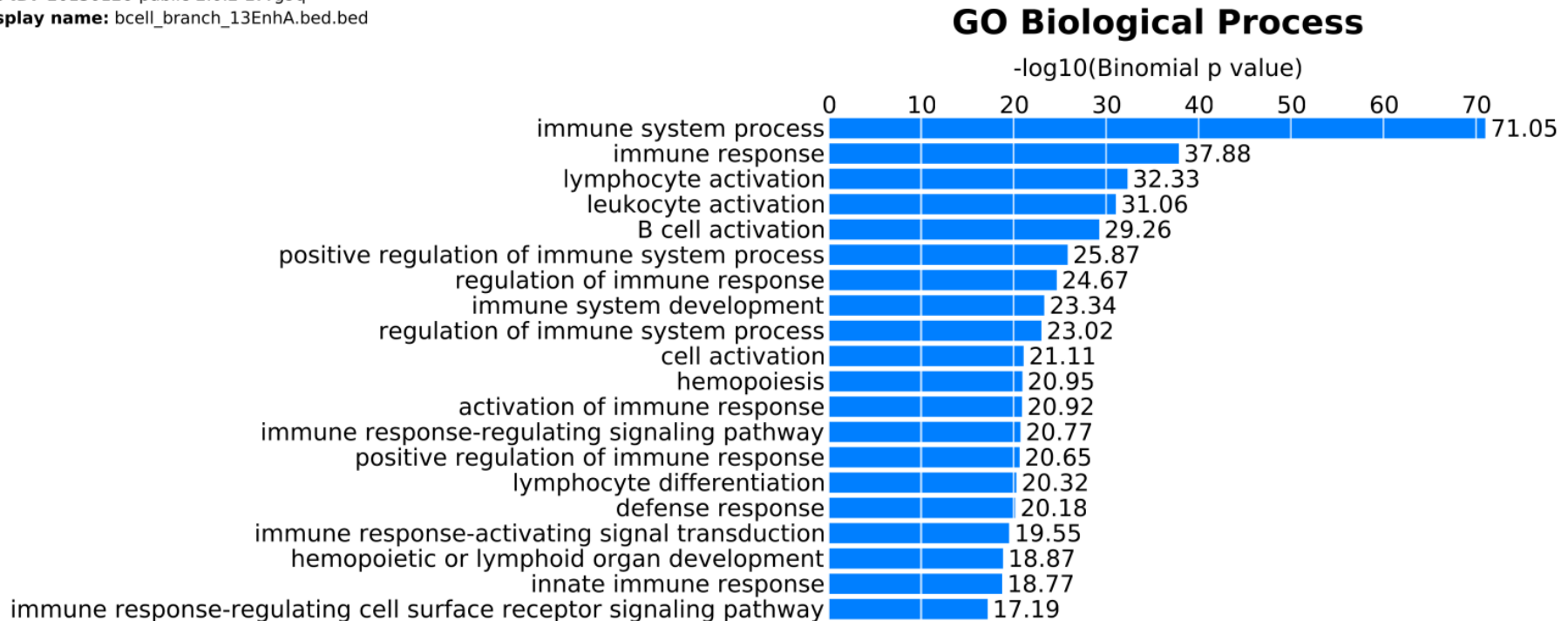
GO Biological Process

$-\log_{10}(\text{Binomial p value})$



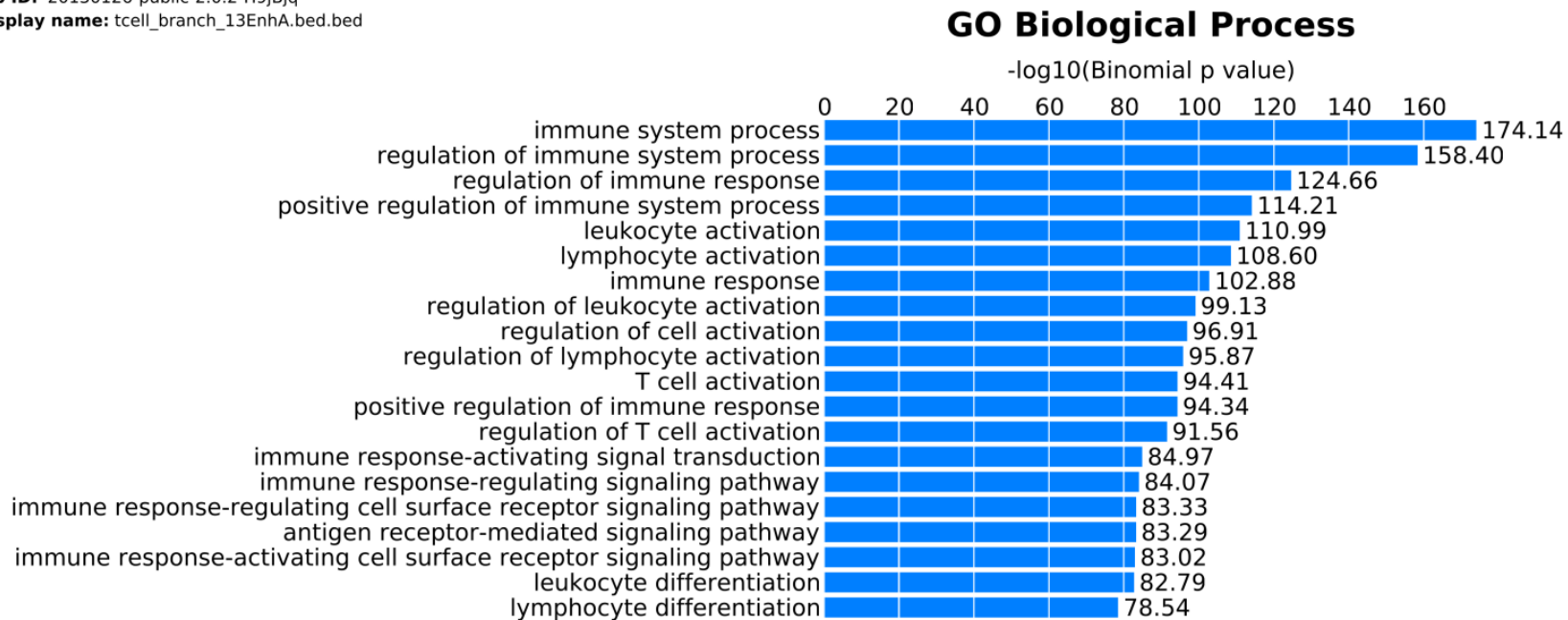
GO analysis of genes associated with enhancers activated in the B-cell branch

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



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

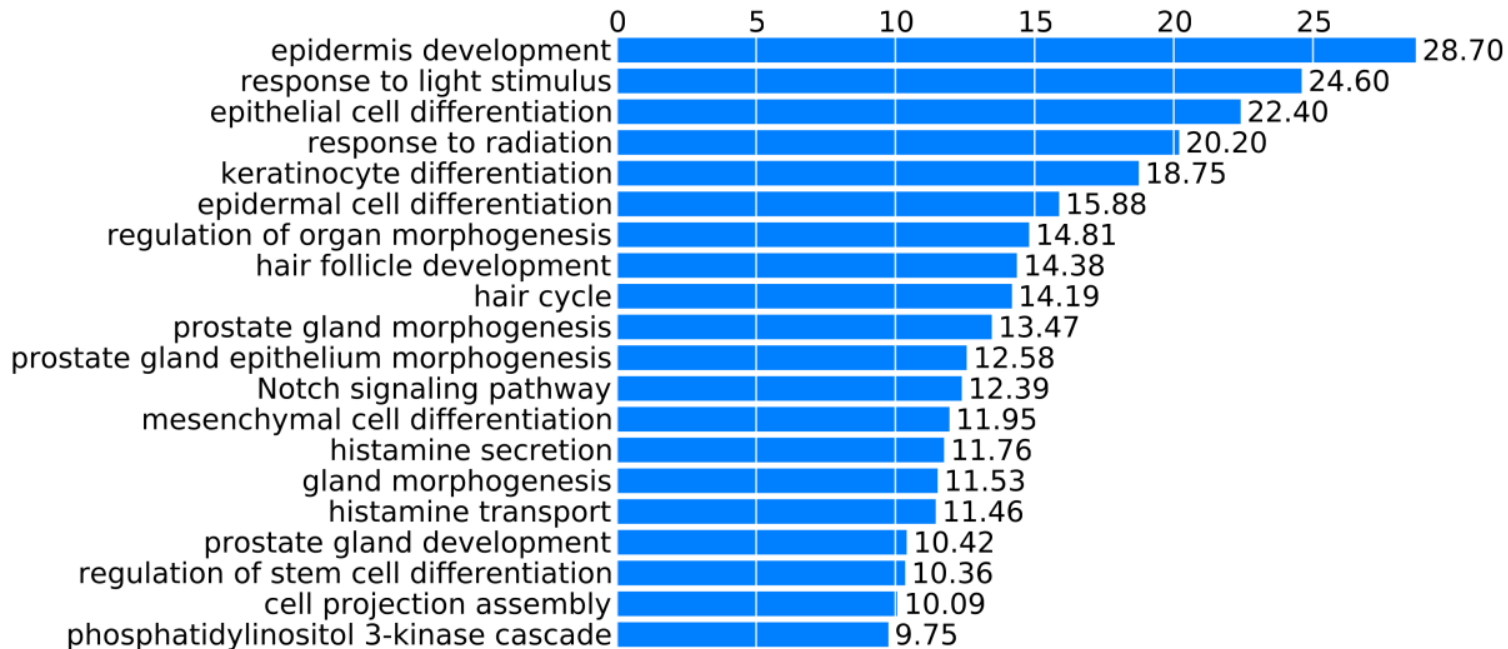


GO analysis of genes associated with enhancers activated in the keratinocyte branch

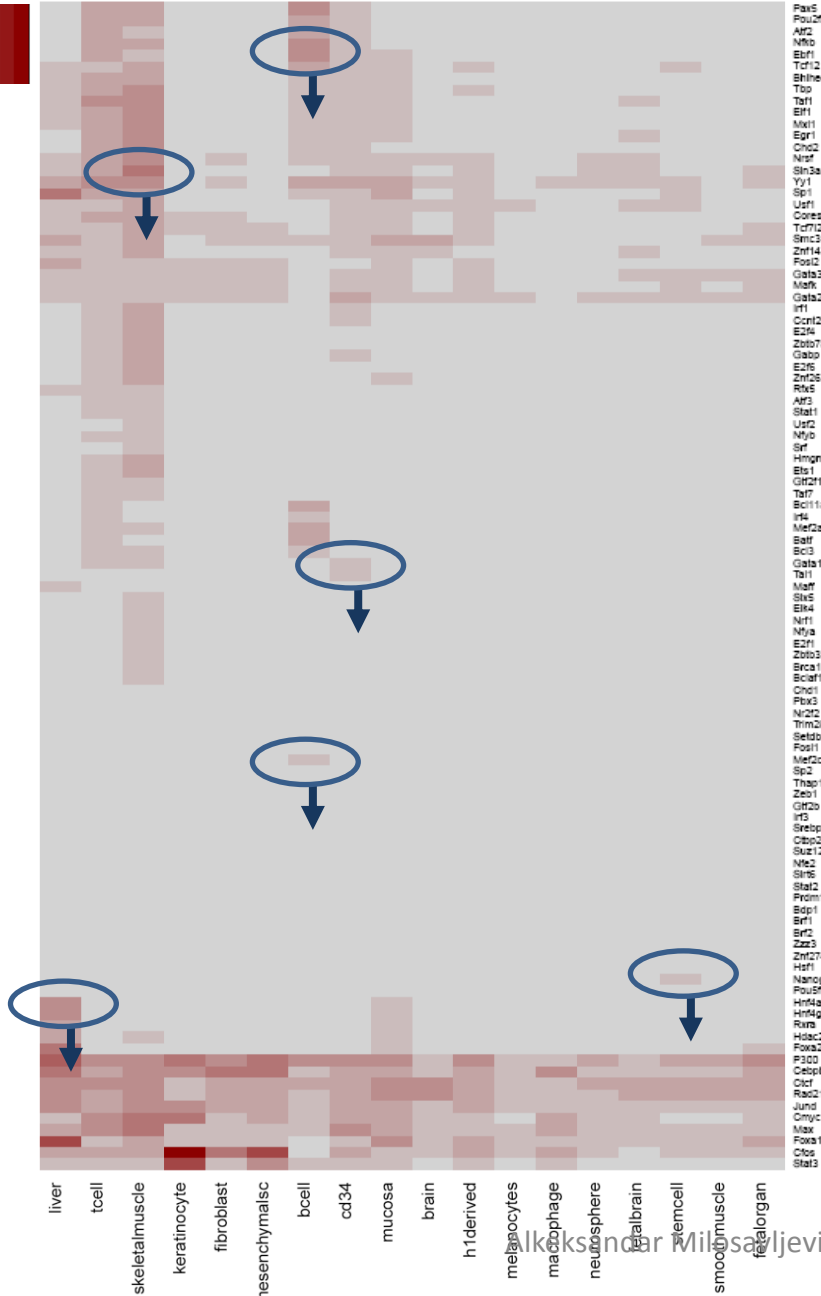
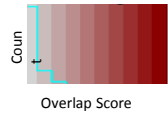
Job ID: 20130126-public-2.0.2-1uuzep
Display name: keratinocyte_13_EnhA.bed.bed

GO Biological Process

$-\log_{10}(\text{Binomial p value})$



TF overlap with branch-specific enhancers



Fox5
Pou2f2
Ar2
Nfya
Ebf1
Tcf12
Bhlhe40
Tcf7
Taf1
Ebf1
Mafk
Egr1
Chc2
Nrf2
Sin3a
Yy1
Sp1
Usf1
Coref1
Tcf12
Smc3
Znf143
Fos12
Gata3
Mafk
Gata2
Irf1
Ocm2
E2f4
Zbtb7a
Gata3
E2f6
Znf33
Rfx5
Ar3
Stat3
Usf2
Myo
Sp1
Hmgn3
Ets1
Gtf2i1
Taf7
Ectf1a
Irf4
Mef2a
Saf
Bcl3
Gata1
Tal1
Mafk
Six5
Ets4
Nrf1
Nfya
E2f1
Zbtb33
Sica1
Bclaf1
Chd1
Fox2
Naf2
Trim28
Selsb1
Foxf1
Mef2c
Sp2
Thap1
Zfp1
Gtf2b
Irf3
Stat3
Ctcf
Suz12
Mafk
Sin3b
Stat2
Prlm1
Bsp1
Ebf1
Erf2
Zfp3
Znf274
Hnf1
Nanog
Pou5f1
Hnf4a
Hnf4g
Rarb
Hnf2c
Foxa2
P300
Ctcf
Rad21
Jund
Cmyc
Max
Foxo1
Chc2
Stat3

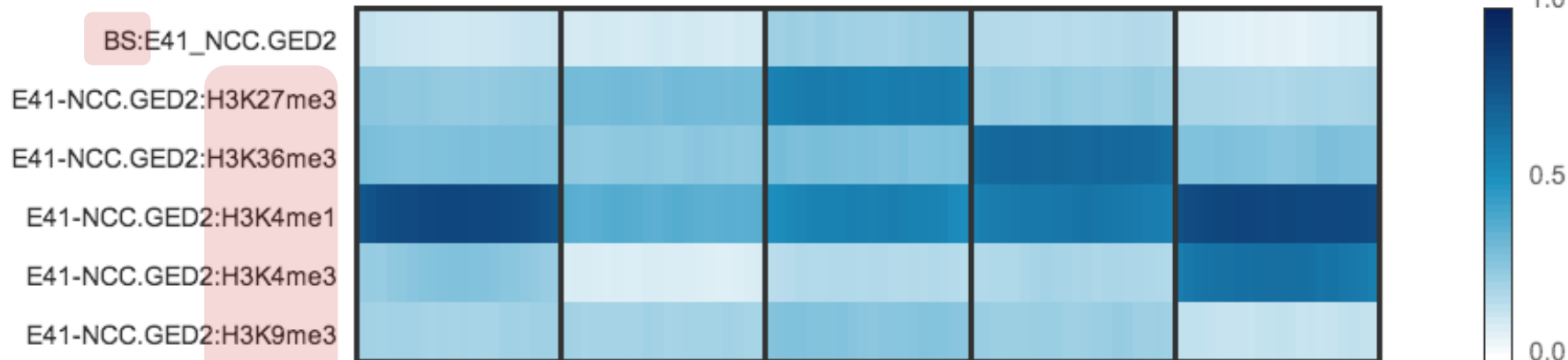
- Ebf1 is essential for **B cell** development
- Sin3a is known to play role in **muscle** differentiation.
- Tal1 plays key role in **HSCs** self renewal and engraftment
- Mef2c is required for **B cell** proliferation
- Nanog important for **ESCs** self-renewal
- Hnf4a and Hnf4g regulate **hepatic** genes

Coordinated changes of epigenomic marks in the neuronal branch

15,983 enhancers activated specifically in the neuronal branch

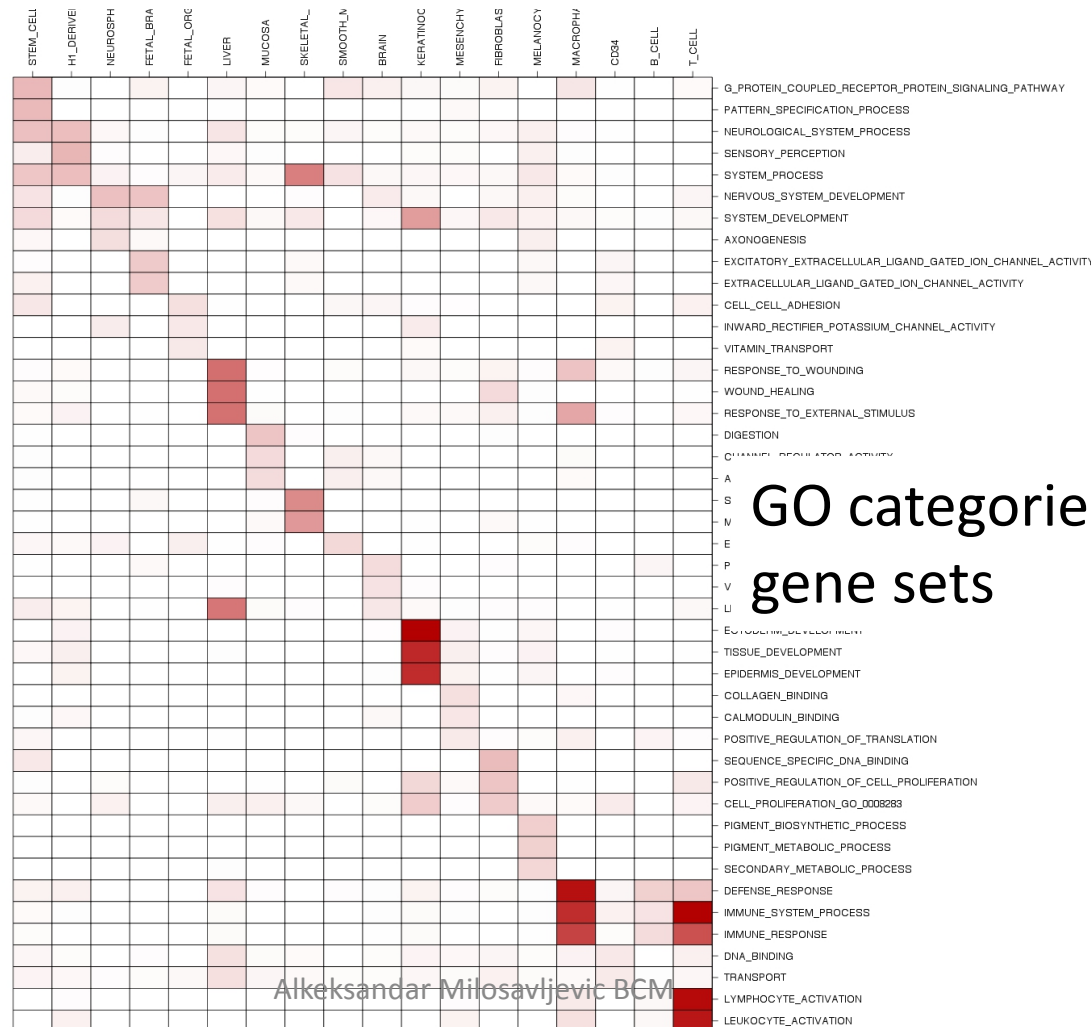
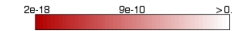


Cluster profiles



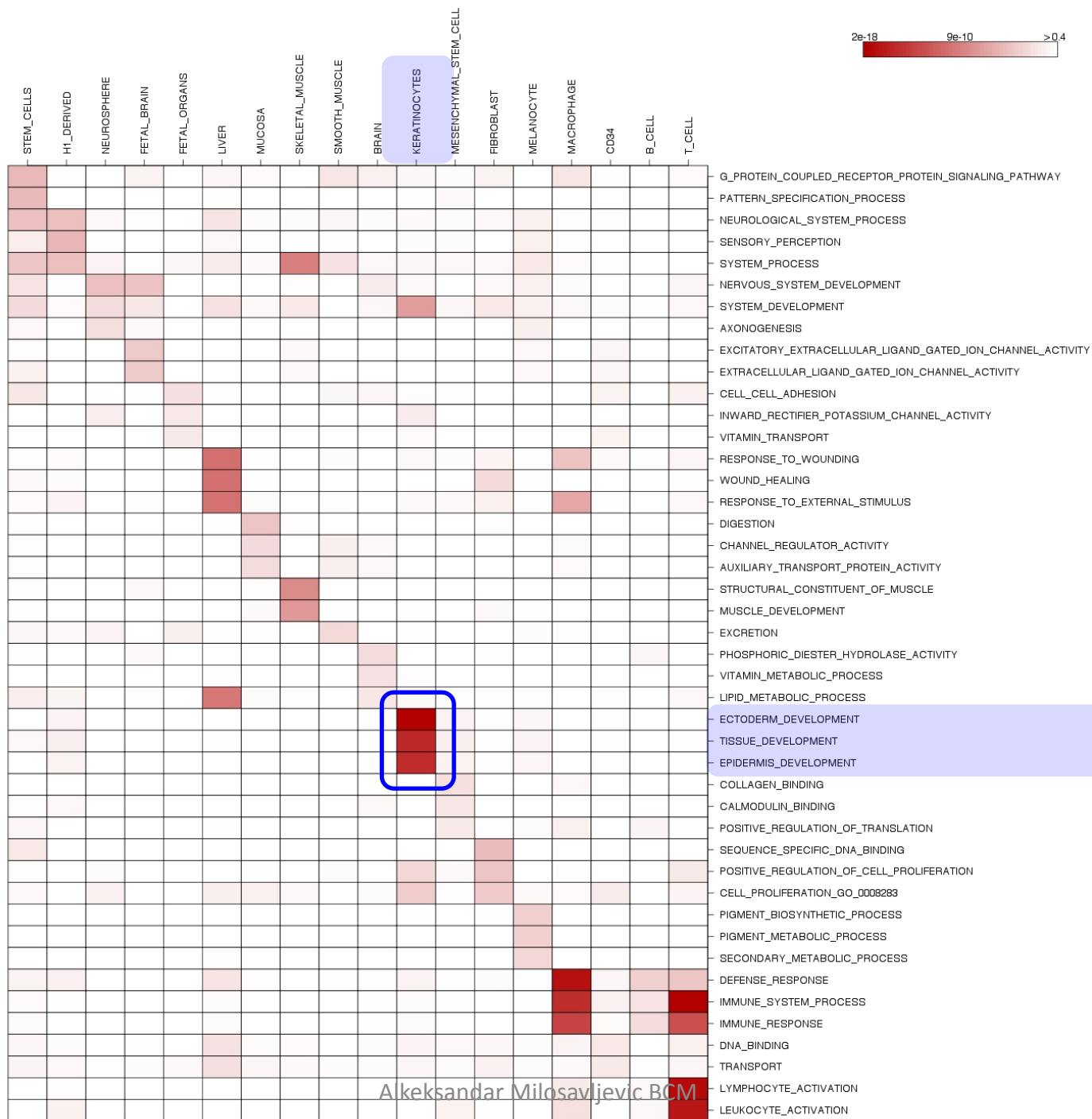
Branch-specific changes in H3K4me3 signals over promoters

Tree branches (lineages)



GO categories for gene sets



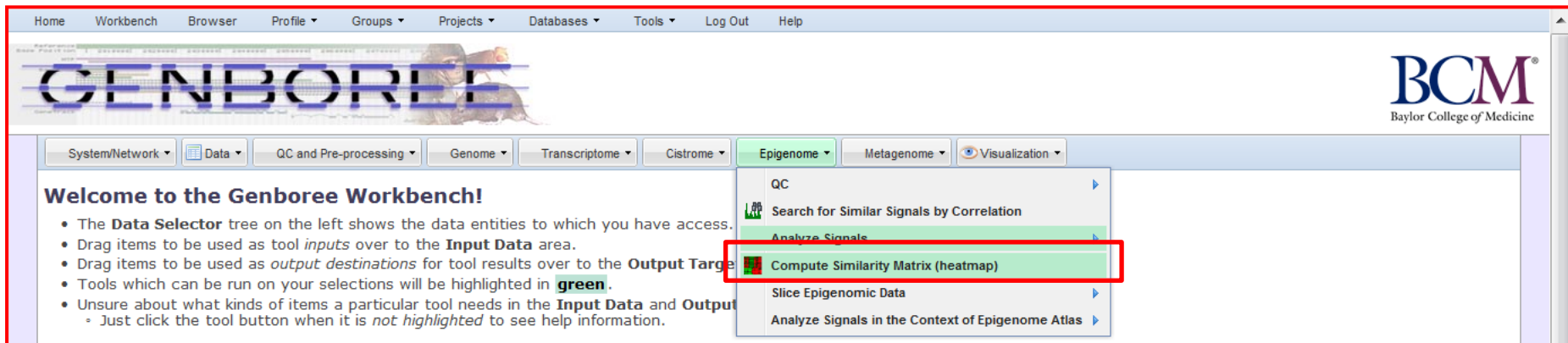




The data for several of the workshop use cases is taken from this publication:

“Functional annotation of the human brain methylome across brain and blood”. Matthew Davies¹, Manuela Volta¹, Abhishek Dixit¹, Simon Lovestone¹, Cristian Coarfa², R. Alan Harris², Aleksandar Milosavljevic², Claire Troakes¹, Safa Al-Sarraj¹, Richard Dobson¹, Leonard C. Schalkwyk¹, Jonathan Mill^{1*}
Genome Biology, 12:R43, 2012

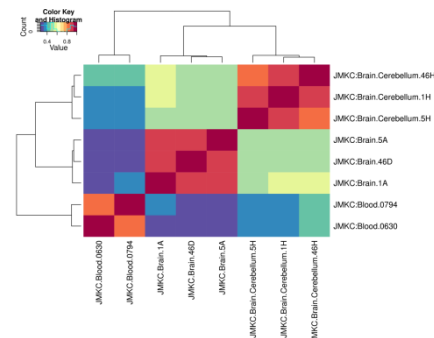
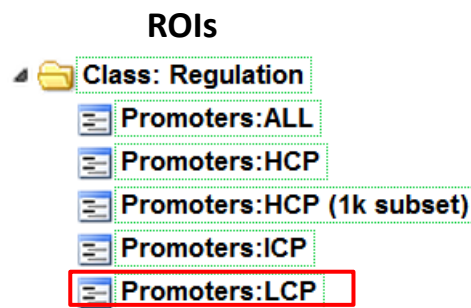
¹Institute of Psychiatry, King’s College London. UK. ²Baylor College of Medicine, Houston, Texas. USA. *Corresponding Author: Dr. Jonathan Mill, Address: Institute of Psychiatry, SGDP Centre, De Crespigny Park, Denmark Hill, London.



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

Epigenomic Tracks:

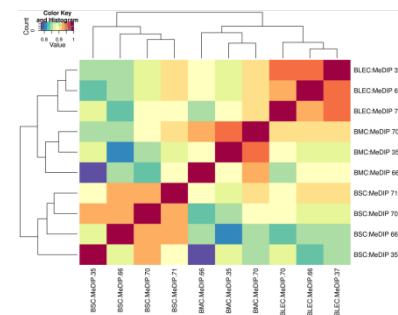
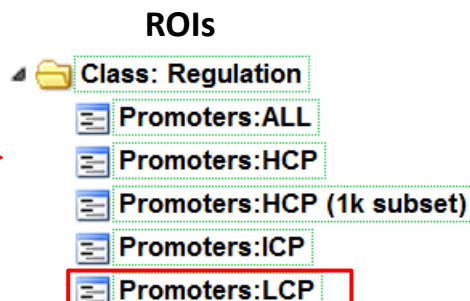
- Blood
- Cerebellum
- Cortex



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

Epigenomic Tracks:

- Breast Luminal Epithelium
- Breast Myoepithelial
- Breast Stem Cell



Home Workbench Browser Profile Groups Projects Databases Tools Log Out Help

GENBOREE BCM Baylor College of Medicine

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome **Epigenome** Metagenome Visualization

Welcome to the Genboree Workbench!

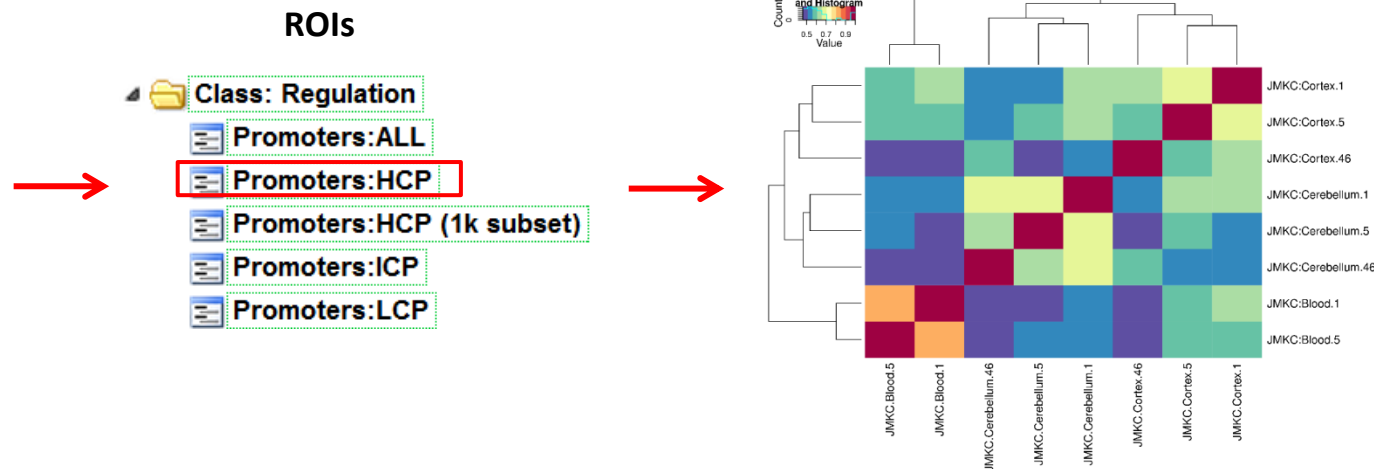
- The **Data Selector** tree on the left shows the data entities to which you have access.
- Drag items to be used as tool **inputs** over to the **Input Data** area.
- Drag items to be used as **output destinations** for tool results over to the **Output Target** area.
- Tools which can be run on your selections will be highlighted in **green**.
- Unsure about what kinds of items a particular tool needs in the **Input Data** and **Output Target** areas? Just click the tool button when it is *not highlighted* to see help information.

QC
 Search for Similar Signals by Correlation
Analyze Signals
Compute Similarity Matrix (heatmap)
 Slice Epigenomic Data
 Analyze Signals in the Context of Epigenome Atlas

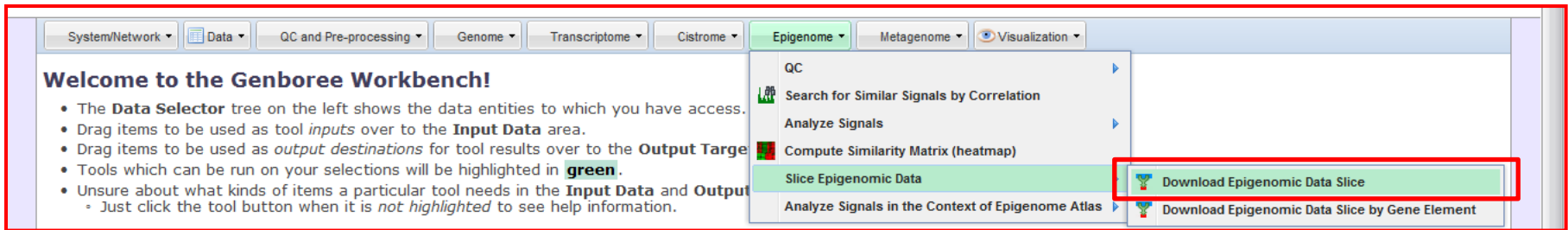
Use Case 5: Methylation of some features discriminate tissue type better than others

Epigenomic Tracks

- Blood
- Cerebellum
- Cortex



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



Epigenomic Tracks:

- H1 cell line
- IMR90 cell line



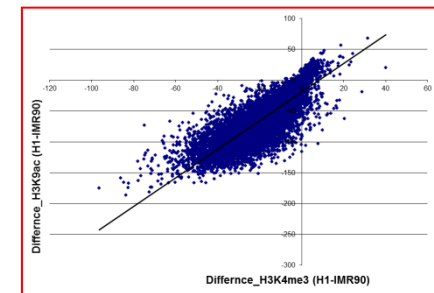
Collates score tracks into one data matrix, export to Excel

	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	BS.Combined	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	H1.H3K9ac	68	IMR90.H3K9ac	IMR90.H3K9ac	46	H1.H3K4me3	H1.H3K4me3	H1.H3K4me3	H1.H3K4me3	H1.H3K4me3
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

Home Workbench Browser Profile Groups Projects Databases Tools Log Out Help

GENBOREE

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome Metagenome Visualization

Welcome to the Genboree Workbench!

- The **Data Selector** tree on the left shows the data entities to which you have access.
- Drag items to be used as tool *inputs* over to the **Input** area.
- Drag items to be used as *output destinations* for tool outputs.
- Tools which can be run on your selections will be highlighted.
- Unsure about what kinds of items a particular tool needs? Just click the tool button when it is *not highlighted* to see help information.

Find Differences By Regression
Cluster by Spark
Compare by LIMMA
QC
Search for Similar Signals by Correlation
Analyze Signals
Compute Similarity Matrix (heatmap)
User Supplied Data Matrix
Tracks
Track with Sample Metadata

Data Selector

Details

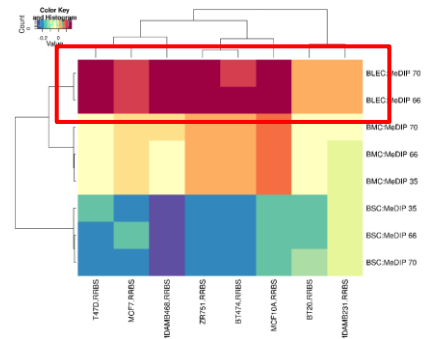
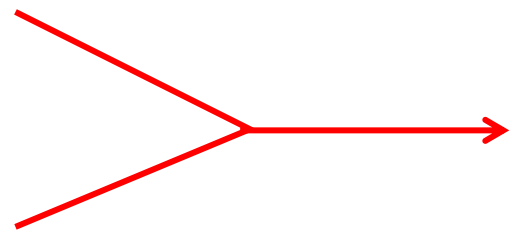
LIMMA: Smyth, G. K. Statistical Applications in Genetics and Molecular Biology (2005)

“Your” Epigenomic Tracks (RRBS):

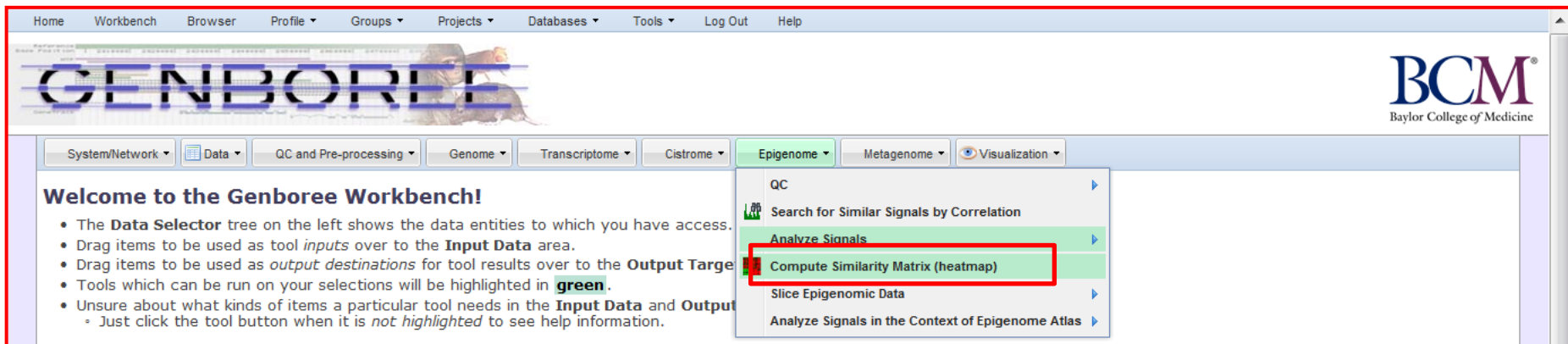
- Breast Luminal Epithelium
- Breast Myoepithelial
- Breast Stem Cell

Public Epigenomic Tracks (MeDIP):

- Breast Luminal Epithelium
- Breast Myoepithelial
- Breast Stem Cell



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



Use Case 13a: Cluster all 16 breast tissue samples

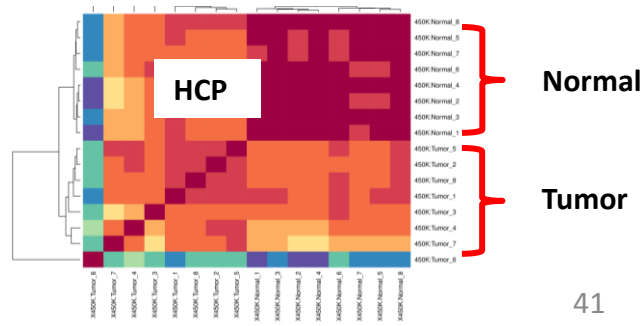
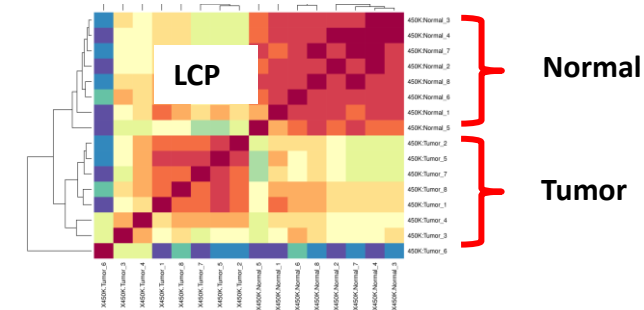
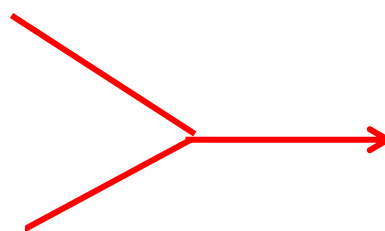
16 450 K Samples (Dedeurwaerder, S. et al. (2011))

-8 normal breast samples

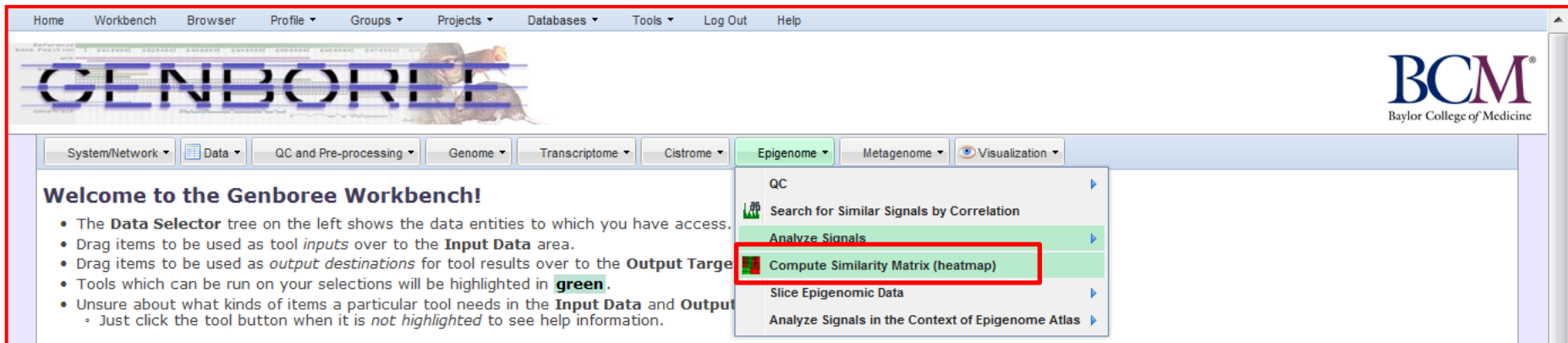
-8 cancerous breast samples

ROIs (HCP vs LCP)

- Class: Regulation
 - Promoters: ALL
 - Promoters: HCP**
 - Promoters: HCP (1k subset)
 - Promoters: ICP
 - Promoters: LCP**



Use Case 13: Epigenomic variation in breast tumors



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

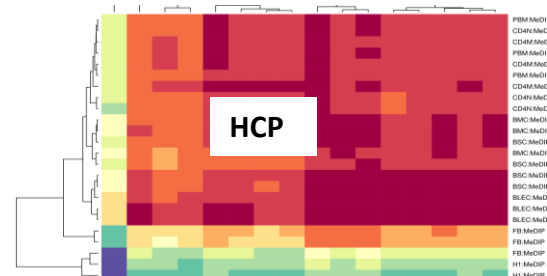
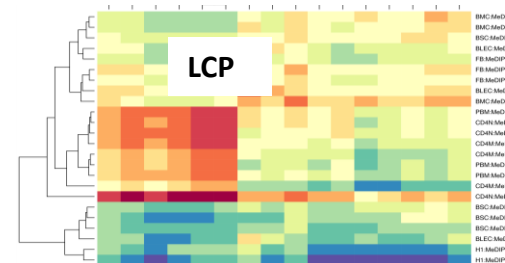
16 450 K Samples (Dedeurwaerder, S. et al. (2011))

-8 normal breast samples

-8 cancerous breast samples

ROIs

- Class: Regulation
 - Promoters:ALL
 - Promoters:HCP
 - Promoters:HCP (1k subset)
 - Promoters:ICP
 - Promoters:LCP



Use Case 13: Epigenomic variation in breast tumors

Welcome to the Genboree Workbench!

- The **Data Selector** tree on the left shows the data entities to which you have access.
- Drag items to be used as tool *inputs* over to the **Input** area.
- Drag items to be used as *output destinations* for tool outputs to the **Output** area.
- Tools which can be run on your selections will be highlighted.
- Unsure about what kinds of items a particular tool needs? Just click the tool button when it is *not highlighted* to see help information.

Epigenome menu items:

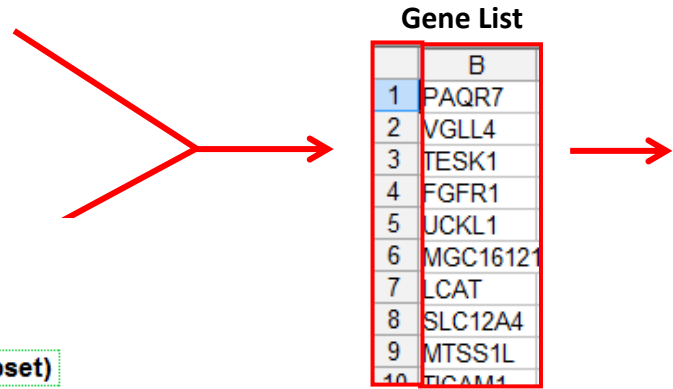
- QC
- Search for Similar Signals by Correlation
- Analyze Signals
- Compute Similarity Matrix (heatmap)
- User Supplied Data Matrix
- Tracks
- Track with Sample Metadata

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

- Class: Regulation**
- Promoters:ALL
- Promoters:HCP
- Promoters:HCP (1k subset)
- Promoters:ICP
- Promoters:LCP



Gene List

	B
1	PAQR7
2	VGLL4
3	TESK1
4	FGFR1
5	UCKL1
6	MGC16121
7	LCAT
8	SLC12A4
9	MTSS1L
10	TTCAM4

DAVID DATABASE

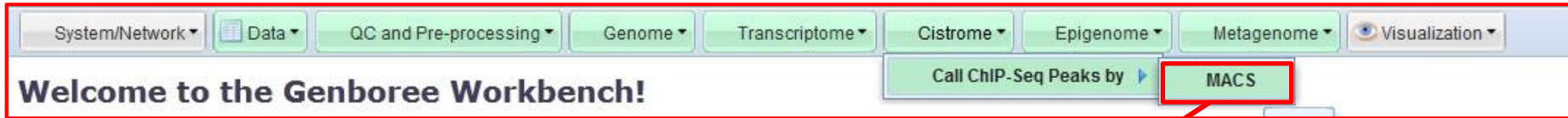
Home | Start Analysis | Shortcut to DAVID Tools

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease

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



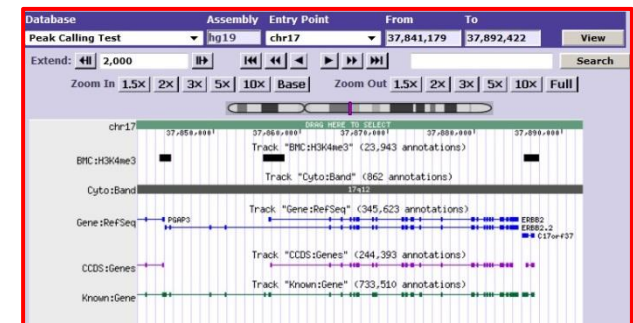
Zhang et al, *Genome Biology* (2008)

MACS results (file in Genboree)

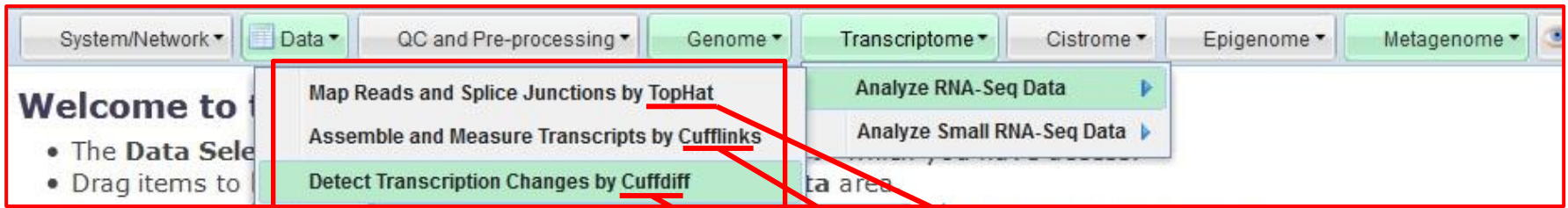
BED files

	chr	start	end	length	summit	tags	-10*LOG10 fold_enrichment	
20	chr1	9861	10677	817	360	292	2034.63	24.91
21	chr1	713307	715489	2183	1297	83	250.89	11.7
22	chr1	724760	727160	2401	2202	92	223.86	16.37
23	chr1	761468	763266	1799	865	133	463.34	14.42
24	chr1	833096	834002	907	577	20	96.47	11.54
25	chr1	839750	840203	454	262	15	77.67	11.4
26	chr1	859636	861418	1783	741	37	83.16	5.71
27	chr1							
28	chr1							7.95
29	chr1							2.67
30	chr1							4.84
31	chr1	911145	912090	952	598	25	123.98	13.95

Visualize in Genboree



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



Trapnell et al, *Bioinformatics* (2009)

Trapnell et al, *Nature Biotech* (2010)

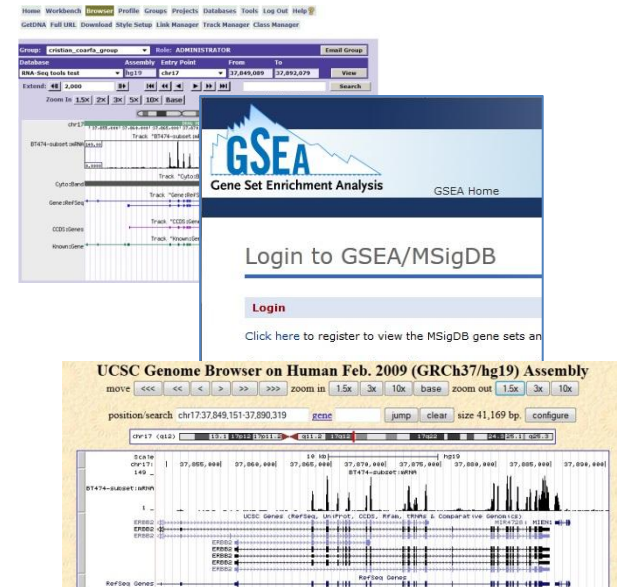
Trapnell et al, *Nature Biotech* (2010)

Gene expression diffs (file in Genboree)

FASTQ,
BAM
files

	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

GENBOREE

BCM
Baylor College of Medicine

System/Network ▾ Data ▾ QC and Pre-processing ▾ Genome ▾ Transcriptome ▾ Cistrome ▾ Epigenome ▾ Metagenome ▾ Visualization ▾ Help ▾

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter... ▾

- www.genboree.org
 - aleks_group
 - AleksG
 - API

Details

Attribute	Value
-----------	-------

The Genboree Workbench: Web-based Data Management & Analysis

The screenshot shows the Genboree Workbench interface. At the top is a navigation bar with links: Home, Workbench, Browser, Profile, Groups, Projects, Databases, Tools, Log Out, and Help. Below this is the Genboree logo and the BCM logo. A secondary navigation bar contains tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, and Epigenome. The main content area displays a welcome message and a 'Data Selector' panel on the left. The 'Data Selector' panel includes a 'Refresh' button and a 'Data Filter' dropdown. It lists various data sources under 'www.genboree.org', such as Atlas Tools Access, EDACC, Epigenome Informatics Workshop (May 2012), Epigenome ToolSet Demo Input Data, Epigenomics Roadmap Repository, GenboreeUser_group, GMT_Tutorial, Group1, JonathanMill_Lab, paithank_group, Public, ROI Repository, and Targeted Atlases. On the right, there are three panels: 'Details' (with an 'Attribute' section), 'Input Data' (with up, down, and delete icons), and 'Output Targets' (with up, down, and delete icons). Callout boxes provide explanations for these panels.

Home Workbench Browser Profile Groups Projects Databases Tools Log Out Help

GENBOREE

BCM

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- www.genboree.org
 - Atlas Tools Access
 - EDACC
 - Epigenome Informatics Workshop (May 2012)
 - Epigenome ToolSet Demo Input Data
 - Epigenomics Roadmap Repository
 - GenboreeUser_group
 - GMT_Tutorial
 - Group1
 - JonathanMill_Lab
 - paithank_group
 - Public
 - ROI Repository
 - Targeted Atlases

Details

Attribute

Input Data

↑ ↓ ×

Output Targets

↑ ↓ ×

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

Tells the tool to use this data/file

Tells the tool where to deposit results

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

Genboree Workbench: Create & Manage Collaborations

The screenshot displays the Genboree Workbench interface. At the top, there is a navigation bar with links: Home, Workbench, Browser, Profile, Groups, Projects, Databases, Tools, Log Out, and Help. Below this is the Genboree logo and the BCM Baylor College of Medicine logo. A secondary navigation bar contains tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, Metagenome, and Visualization. The main content area is titled "Welcome to the Genboree Workbench! [Getting Started]". On the left, a sidebar menu is open, showing "User Profile", "Groups", "Hosts", "Jobs", and "Request Feature". The "Groups" menu is expanded, revealing a list of groups: Atlas Tools Acc, EDACC, Epigenome Inf, Epigenome To, Epigenomics F, GenboreeUser_group, GMT_Tutorial, Group1, JonathanMill_Lab, paithank_group, Public, ROI Repository, Targeted Atlases, and vamin_group. A context menu is open over the "GenboreeUser_group" entry, listing actions: Create Group, Edit Group Info, Delete Group, Add Existing User To Group, Add New User To Group, Update Roles, Copy Users, and Message to Group. The main workspace contains three panels: "Details" (a table with columns "Attribute" and "Value"), "Input Data" (with up, down, and delete icons), and "Output Targets" (with up, down, and delete icons).

Create & Manage Databases & Projects – Share Data

The screenshot displays the GENBOREE web interface. At the top, there is a navigation bar with links: Home, Workbench, Browser, Profile, Groups, Projects, Databases, Tools, Log Out, and Help. Below this is the GENBOREE logo and the BCM Baylor College of Medicine logo. A secondary navigation bar contains tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Metagenome, and Visualization. The 'Data' tab is active, and its dropdown menu is open, showing options: Databases, Entity Lists, Entrypoints, Files, Projects, Samples & Sample Sets, and Tracks. The 'Projects' option is highlighted, and its sub-menu is also open, listing: Create Database, Rename Database, Delete Database, Edit Database Info, Unlock/Lock Database, and Publish/Retract Database. A green dashed line connects the 'Projects' option in the main menu to the 'Create Project' option in the sub-menu. The 'Create Project' option is highlighted with a green box. On the left, a 'Data Selector' panel shows a tree view of various data sources, including Atlas Tool, EDACC, Epigenomics Informatics Workshop (May 2012), Epigenome ToolSet Demo Input Data, Epigenomics Roadmap Repository, GenboreeUser_group, GMT_Tutorial, Group1, JonathanMill_Lab, paithank_group, Public, ROI Repository, Targeted Atlases, and vamin_group. The main content area is partially visible, showing a table with columns for 'Route' and 'Value', and an 'Output Targets' section with up, down, and delete icons.

Manage Databases

The image displays a database management interface. On the left, a vertical menu lists several actions: Create Database, Rename Database, Delete Database, Edit Database Info, Clone/Copy Database, Unlock/Lock Database, and Publish/Retract Database. Two red arrows originate from this menu: one points to the 'Rename Database' dialog box on the right, and the other points to the 'Edit Database Info' dialog box at the bottom. Both dialog boxes are titled 'Tool Settings' and contain sections for 'Tool Overview' and 'Settings'. The 'Rename Database' dialog shows the current database as 'GenboreeUser_database' and has a field for the 'New Database Name'. The 'Edit Database Info' dialog shows the same database and includes fields for 'New Database Name', 'Description', 'Species', and 'Version'. Both dialogs have 'Submit' and 'Cancel' buttons.

Menu Items:

- Create Database
- Rename Database
- Delete Database
- Edit Database Info
- Clone/Copy Database
- Unlock/Lock Database
- Publish/Retract Database

Rename Database Dialog:

Tool Settings

Rename Database

Tool Overview

Database to be renamed:

Database: *GenboreeUser_database* Group: *GenboreeUser_group*

Settings

New Database Name

Submit Cancel

Edit Database Info Dialog:

Tool Settings

Edit Database Info

Tool Overview

Database to update/edit:

Database: *GenboreeUser_database* Group: *GenboreeUser_group*

Settings

New Database Name

Description

Species

Version

Submit Cancel

Authorization: Decide How to Share Data

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

Tool Settings

Add New User To Group

Add user to group:

Group: *GenboreeUser_group*

Current Users:

The following users are members of this group:

Login	Name	Email Address	Role
andrewj	Andrew R Jackson	andrewj@bcm.edu	Administrator
paithank	Sameer Paithankar	paithank@bcm.tmc.edu	Administrator
raghuram	sriram raghuraman	raghuram@bcm.edu	Author
kevin_riehle	Kevin Riehle	riehle@bcm.edu	Administrator
mroth	matt roth	mattr@bcm.edu	Administrator
GenboreeUser	Genboree User	GenboreeUser@gmail.com	Administrator

Add User

First Name

Last Name

Email Address

Institution

Role

Genboree is a hosted service. Code is available **free for academic use.**

Assign roles

Data Sharing Option 1: Projects

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

GENBOREE
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

The screenshot displays the Genboree Workbench interface. At the top, there is a navigation bar with tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, Metagenome, Visualization, and Help. Below the navigation bar, a welcome message reads "Welcome to the Genboree Workbench! [Getting Started]".

The main area is divided into two panels. The left panel, titled "Data Selector", contains a tree view of data. A "Refresh" button and a search filter "lect a filter..." are at the top of this panel. The tree view shows a folder structure under "lean-obese-twins-full-final":

- All Annotations in Database
- Tracks (indicated by a red arrow)
- Lists & Selections
- SampleSets
- Samples
- Files (indicated by a red arrow)
 - lean_obese.metadata.update_3.tsv
 - lean_obese.metadata.update_2.tsv
 - lean_obese.metadata.update
 - MicrobiomeWorkBench
 - ML-lean-obese-min-len-50
 - RDP-lean-obese-min-len-50
 - alpha-lean-obese-min-len-50-remove-chimeras
 - alpha-lean-obese-min-len-50
 - lean-obese-min-len-50-remove-chimeras
 - lean-obese-min-len-50
 - MicrobiomeData
 - Queries

The right panel contains three sections: "Details" (with an "Attribute" field), "Input Data" (with icons for up, down, delete, and refresh), and "Output Targets" (with the same icons).

Analysis of private data enabled by group-level access control

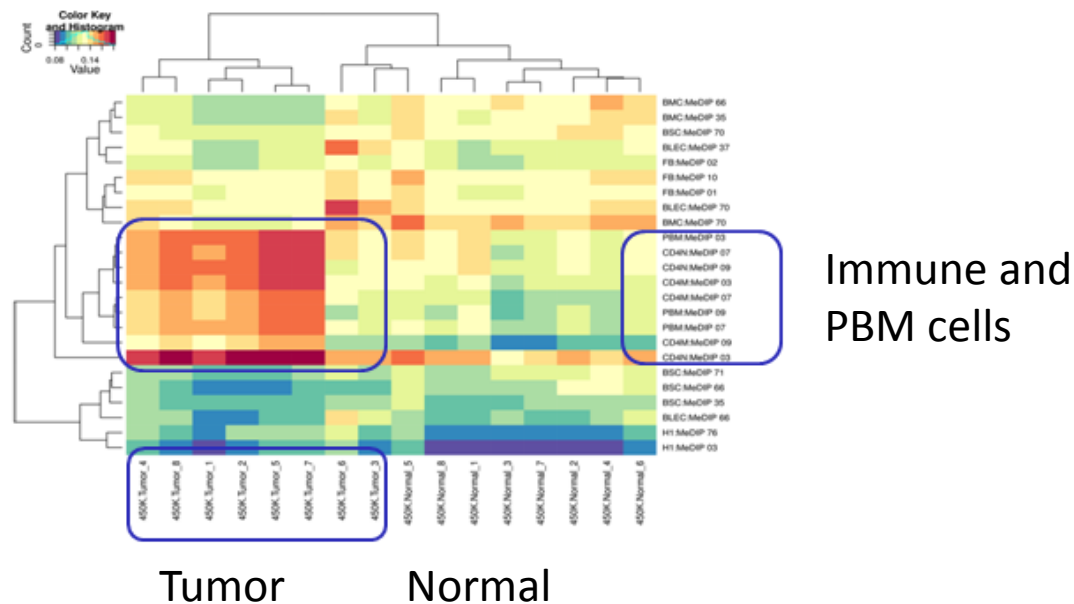
Illumina 450K array profiles can be compared against methyloms in the Epigenome Atlas

The screenshot displays the GENBORER web application interface. At the top left is the 'GENBORER' logo. At the top right is the 'BCM' logo with 'Baylor College of Medicine' underneath. Below the logo is a navigation bar with tabs for 'System/Network', 'Data', 'QC and Pre-processing', 'Genome', 'Transcriptome', 'Cistrome', 'Epigenome', 'Metagenome', 'Visualization', and 'Help'. The main content area shows a 'Welcome to' message and a 'Data Selector' sidebar on the left. The 'Data Selector' sidebar has a tree view with folders like 'aleks_group', 'AleksG', 'ARJ', 'ARJ2', 'arpit_group', 'Atlas Tools Access', 'BCM-MDA SOLID Matepairs', 'BioMarkers', and 'Bios-533-2005'. A 'Refresh' button is above the tree. A 'Data Filter' dropdown is set to 'Select a filter...'. The 'Import' menu is open, showing options: 'Array Data', 'Track Metadata', and 'Upload Track'. The 'Array Data' option is highlighted, and a tooltip is visible over it that reads: 'Array Data' and 'Import array data into Genboree as a high density track.' Below the tooltip is an 'Input Data' section with three icons: a green up arrow, a red down arrow, and a red X. The date '2/8/2013' is in the bottom left corner, and the page number '55' is in the bottom right corner.

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

GENBOREE

BCM
Baylor College of Medicine

System/Network Data QC and Pre-processing **Genome** Transcriptome Cistrome Epigenome Metagenome Visualization Help

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- www.genboree.org
 - AleksG group
 - AleksG
 - API

Details

Attribute	Value

57

Other widely used tools are integrated within the Genboree Workench

- MACS
- TopHat
- CuffLiks
- CuffDiff.
-

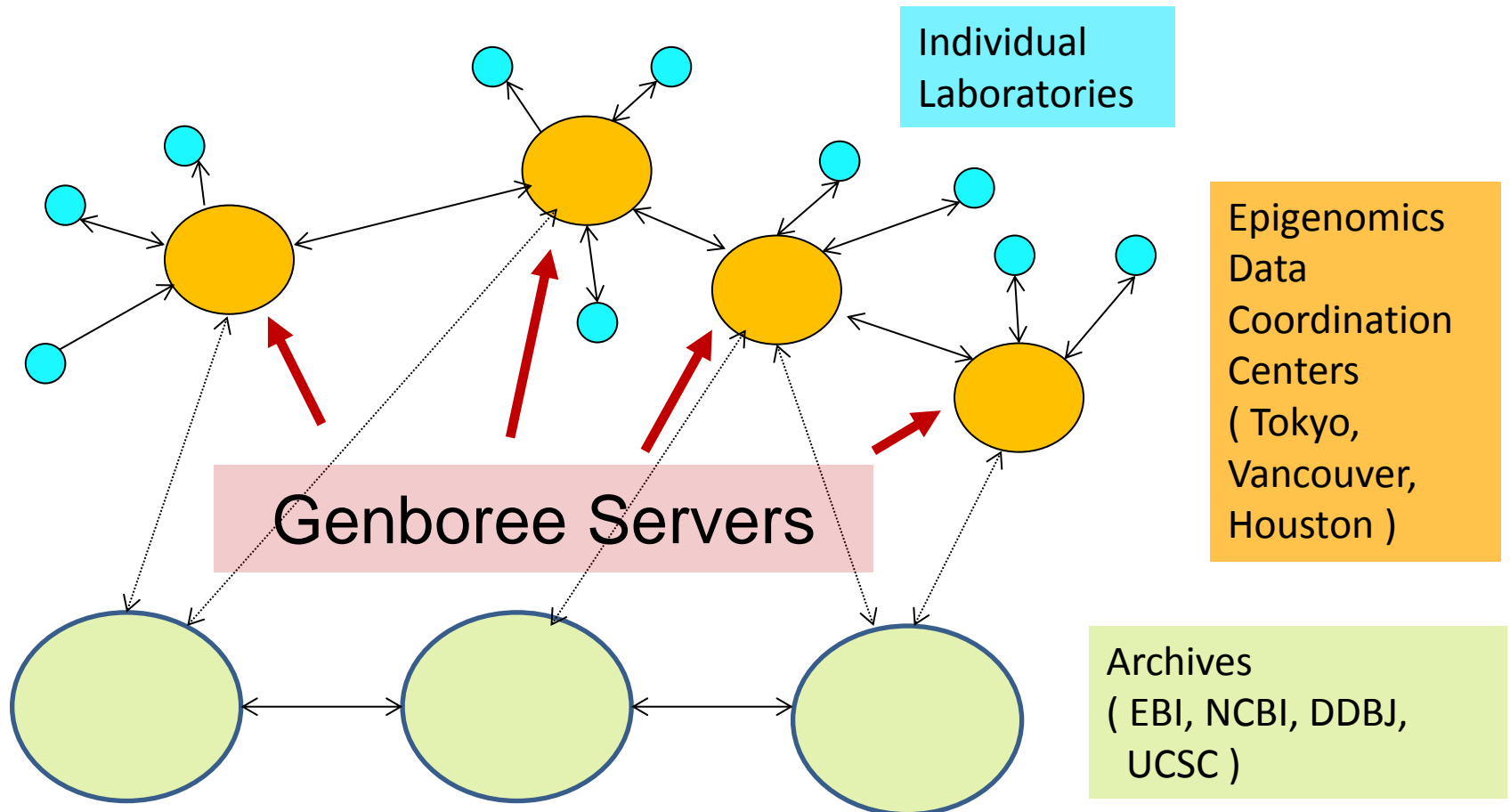
- NIH Roadmap Epigenomics Project
- Epigenome Atlas Analysis
- Genboree Workbench
- **Genboree Network**

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





Vancouver
Steven Jones


Houston
EDACC



Tokyo, Toutai Mituyama

Releases
Informatics
Publications
Forums
Contributors

- [Data Access Policy](#)
- Data embargo period: from 10/31/2011 - 07/31/2012 or earlier as specified [here](#)
- Select cells by **clicking and dragging**, then use the "View Selections in" pull-down in the top left corner (below) to view selections in the Atlas Gene Browser or the UCSC Genome Browser
- Use "Save Selections" in the toolbar (below) to save selected (highlighted) cells in a group and database of your choice (requires login)
- NOTE: Some pages may not be accessible over low bandwidth internet connections. This page has been tested with the following browsers: 

Epigenome Atlas Release 6

View Selections In x Clear Selections

eaAssayType*

eaSampleType

Filter: (e.g. "cell line")

Brain Angular Gyrus

Brain Anterior Caudate

Brain Cingulate Gyrus

Brain Germinal Matrix

	ChIP-Seq	Hi-C	ATAC-Seq	H3K4me1	H3K4me3	H3K9me3	H3K27me3	H3K9ac	H2A/Glac	H2BK120lac	H2BK12ac	H2BK14ac	H3K18ac	H3K23ac	H3K27ac	H3K4ac	H3K4me2	H3K79me1	H3K79me2	H4K20me1	H4K56ac	H4K91ac	H2A.Z	H2AK9ac
Brain Angular Gyrus		2	1	2	2	2	2	1	2															1
Brain Anterior Caudate		2	2	2	2	2	2	2	1															1
Brain Cingulate Gyrus		2	1	2	2	2	2	1	2															1
Brain Germinal Matrix	1			2	2	2	2	2																2
Brain Substantia nigra		2	2	2	2	2	2	2	1															1
Breast Luminal Epithelial Cells	3	5		2	1	1	1	1	1															2
Breast Myoepithelial Cells	3	3		2	1	2	2	2	2															1
Breast Stem Cells	4	4		1	1																			1
Breast vHMEC	1	1		2	1	1	2	1	1															3
Fetal Brain	5	3	2	9	1	2	2	4	4	3	3	4												1
H1 Derived Embryoid Body Cultured Cells			1																					

Samples (by Type)

Epigenomic Assays

Vancouver

Tokyo

Houston

Genboree Installations

Other locations that use Genboree



Vancouver, Steven Jones

Installation in-progress,
Targeted completion 2Q13



Toutai Mituyama

Installation completed,
Dec 2012



**Houston
EDACC**

Genboree on the Commercial Cloud (Rackspace)

System Data Analysis Visual

Welcome to the Genboree W

- The **Data Selector** tree on the left sh
- Drag items to be used as tool *inputs* o
- Drag items to be used as *output desti*
- Tools which can be run on your select
- Unsure about what kinds of items a pa
 - Just click the tool button when it is

Data Selector

Refresh

- ▶ shortTags
- ▶ SOLID-SV-CC-JR
- ▶ Spark Access
- ▶ Targeted Atlases
- ▶ TCGA
- ▶ TCGA-Reporting
- ▶ testAcgh
- ▶ testEDACC
- ▶ Tumor Sequencing Project (TSP)
- ▶ Universal Probes
- ▶ weilie_group
- ▶ Yue
- ▶ yxb4544_group
- ▶ zfranco_group
- ▶ zuozhouc_group

▶ www.brain-research-lab.org

▶ www.microbiome-center.org

Brain Research Lab #1

GENBOREE hosted site

You are currently not logged in.

[« 3rd Epigenome Informatics Workshop »](#)
We have prepared some resources for attendees of the March 2012 Workshop:
× [FAQs](#)
× [Use Cases](#)

[« Support Site »](#) There's now a **Genboree Community Support Site**. Registered users can access general Genboree forums and can make feature/fix requests. If you'd like a forum, issue tracker, wiki, etc, for your Genboree Project to facilitate

Login Name:
Password:
[Login](#) ([Forgot your password?](#))
[Guest/Public View](#)

New to Genboree? [Register here!](#)

Genboree access at www.genboree.org

Microbiome Center #1

GENBOREE hosted site

You are currently not logged in.

[« 3rd Epigenome Informatics Workshop »](#)
We have prepared some resources for attendees of the March 2012 Workshop:
× [FAQs](#)
× [Use Cases](#)

[« Support Site »](#) There's now a **Genboree Community Support Site**. Registered users can access general Genboree forums and can make feature/fix requests. If you'd like a forum, issue tracker, wiki, etc, for your Genboree Project to facilitate

Login Name:
Password:
[Login](#) ([Forgot your password?](#))
[Guest/Public View](#)

New to Genboree? [Register here!](#)

Genboree access at www.genboree.org

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)

