

Virtual Data Integration and The Genboree Network

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

**6th Genboree Workshop on Epigenome
Informatics
March 4th, 2013**

Two key integration problems

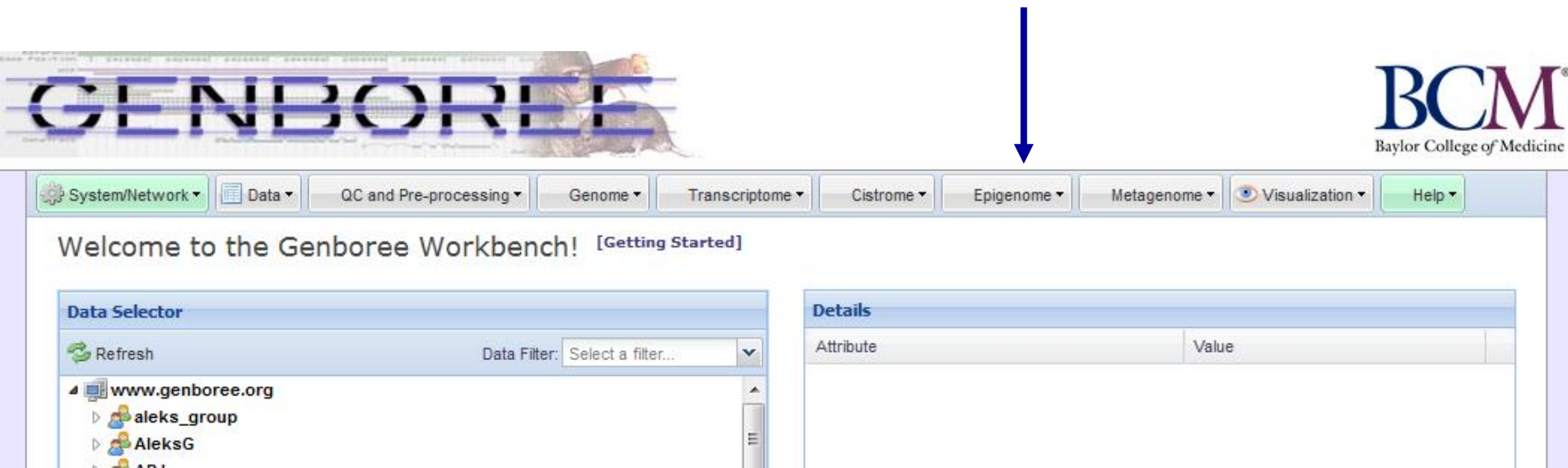
- Genboree Workbench: “Omic” **Toolset** Integration
- Genboree Network: Virtual **Data** Integration

- Genboree Workbench: “Omic” Toolset Integration
- Genboree Network: Virtual Data Integration

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 screenshot shows the Genboree Workbench interface. At the top, the 'GENBOREE' logo is displayed. Below it is a navigation bar with tabs: System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, Metagenome, Visualization, and Help. A blue arrow points to the 'Epigenome' tab. 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 'Refresh' button, a 'Data Filter' dropdown menu, and a tree view showing the directory structure: www.genboree.org, aleks_group, AleksG, and API. The right panel, titled 'Details', contains a table with two columns: 'Attribute' and 'Value'.

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

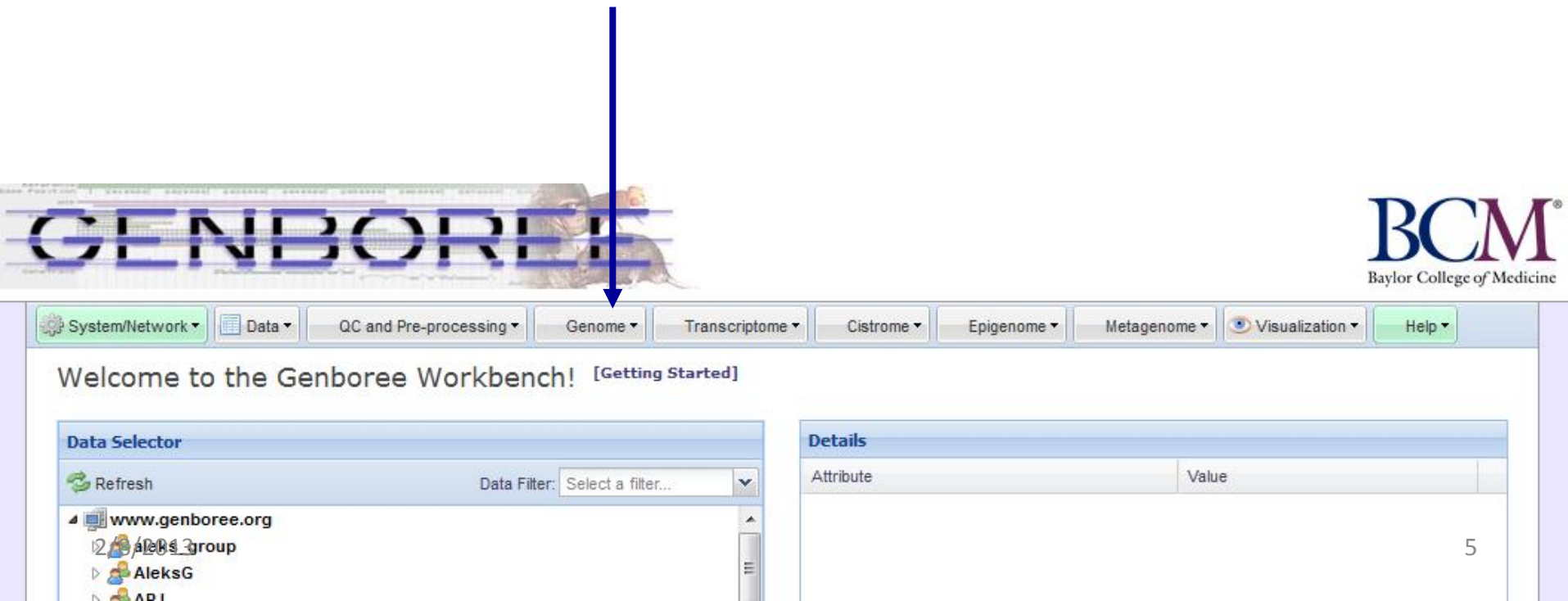
API

Details

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

Genomic Toolset developed in collaboration with the Baylor Genome Center

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



The screenshot displays the Genboree Workbench interface. At the top, the 'GENBOREE' logo is visible on the left, and the 'BCM Baylor College of Medicine' logo is on the right. Below the logos is a navigation bar with several menu items: 'System/Network', 'Data', 'QC and Pre-processing', 'Genome', 'Transcriptome', 'Cistrome', 'Epigenome', 'Metagenome', 'Visualization', and 'Help'. A blue arrow points to the 'Genome' menu. Below the navigation bar, the main content area is divided into two panels. The left panel, titled 'Data Selector', contains a 'Refresh' button, a 'Data Filter' dropdown menu, and a tree view showing a directory structure with 'www.genboree.org' at the top, followed by 'Aleks3 group', 'AleksG', and 'API'. The right panel, titled 'Details', contains a table with two columns: 'Attribute' and 'Value'. The table is currently empty. The page number '5' is visible in the bottom right corner.

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

Aleks3 group

AleksG

API

Details

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

5

Microbiome Toolset includes QIIME (U. Colorado)

Run at Bioinformatics Research Laboratory

The screenshot displays the Genboree Workbench interface. At the top, the 'GENBOREE' logo is on the left, and the 'BCM Baylor College of Medicine' logo is on the right. A navigation bar contains tabs for 'System/Network', 'Data', 'QC and Pre-processing', 'Genome', 'Transcriptome', 'Cistrome', 'Epigenome', 'Metagenome', 'Visualization', and 'Help'. Below this, a 'Welcome to the Genboree Workbench!' message is followed by a '[Getting Started]' link. A dropdown menu is open, showing 'Import 16S rRNA Sequences' and 'Upload Metagenomic WGS Results'. To the right, a 'Data Initialization' dropdown is open, showing 'Data Initialization' and 'Data Analysis'. The 'Data Selector' panel on the left shows a tree view of data sources, including 'www.genboree.org' and 'abinder_private'. A red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. Another red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A tenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A eleventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twelfth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fourteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventeenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A nineteenth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twentieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A twenty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirtieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A thirty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fortieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A forty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fiftieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A fifty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixtieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A sixty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A seventy-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eightieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. An eighty-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninetieth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-first red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-second red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-third red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-fourth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-fifth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-sixth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-seventh red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-eighth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A ninety-ninth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown. A hundredth red box highlights the 'Data Analysis' option in the 'Data Initialization' dropdown.

The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences.

Riehle K, Coarfa C, Jackson A, Ma J, Tandon A, Paithankar S, Raghuraman S, Mistretta TA, Saulnier D, Raza S, Diaz MA, Shulman R, Aagaard K, Versalovic J, Milosavljevic A.

BMC Bioinformatics. 2012;13 Suppl 13:S11. doi: 10.1186/1471-2105-13-S13-S11. Epub 2012 Aug 24. PMID = 23320832

Other widely used tools are integrated within the Genboree Workench

- MACS
- TopHat
- CuffLiks
- CuffDiff.
-

- Genboree Workbench: “Omic” Toolset Integration
- **Genboree Network: Virtual Data Integration**

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

- **Global proliferation of epigenomic profiling** will make it impossible to physically integrate data for analysis
- **IRB-mandated local physical custody** of potentially identifiable human subject data.
- Network bandwidth increases are not keeping up with the **rapidly escalating bandwidth** of sequencing machines.

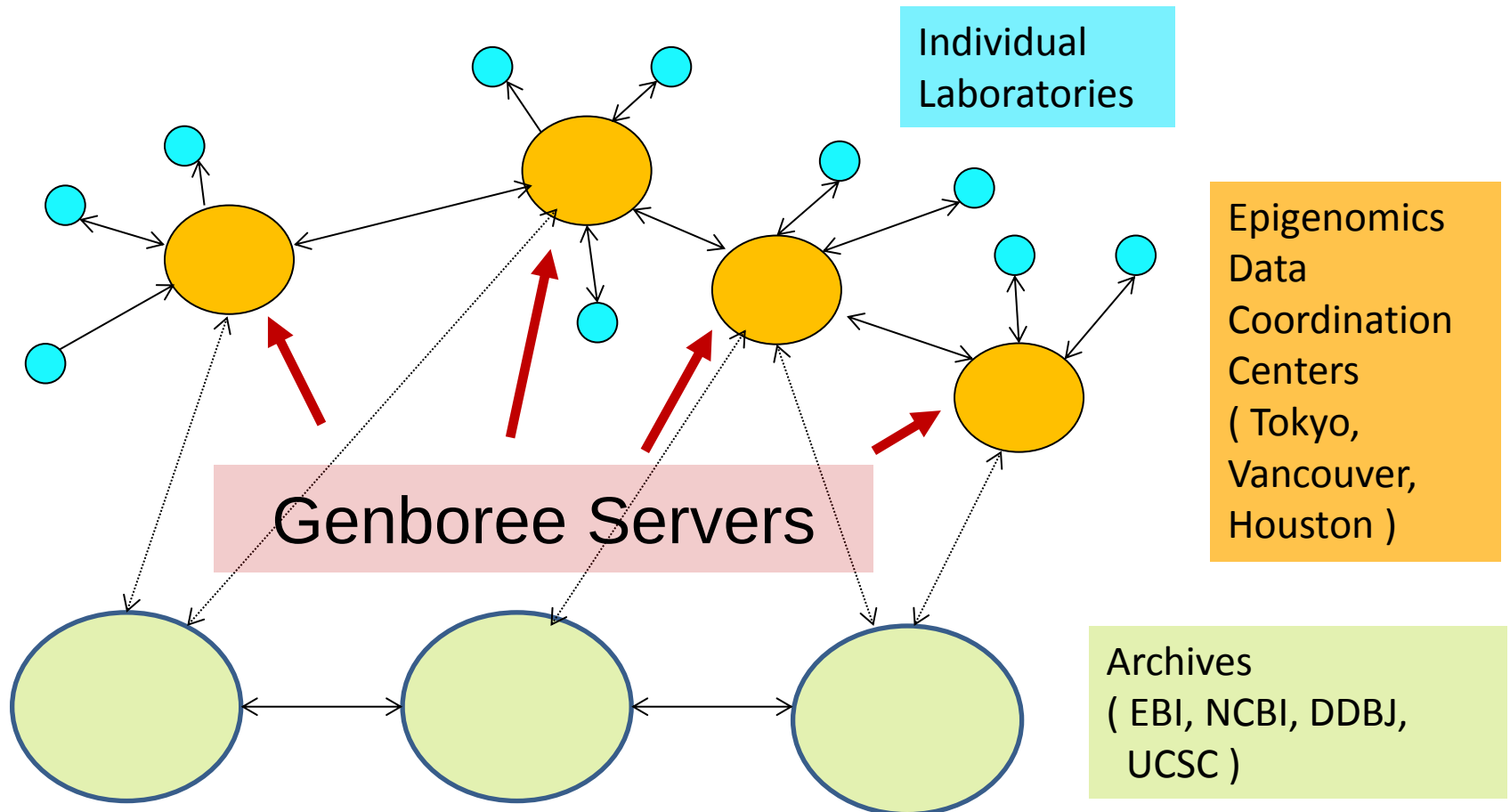


As the data volume explodes, *physical* data integration will become impossible

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" pulldown 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 Clear Selections

eaAssayType
eaSampleType
Filter: Br (e.g. "cell line")

Brain Angular Gyrus
Brain Anterior Caudate
Brain Cingulate Gyrus
Brain Germinal Matrix

Brain Substantia Nigra
Breast Luminal Epithelial Cells
Breast Myoepithelial Cells
Breast Stem Cells
Breast vHMEC
Fetal Brain
H1 Derived Embryoid Body Cultured Cells

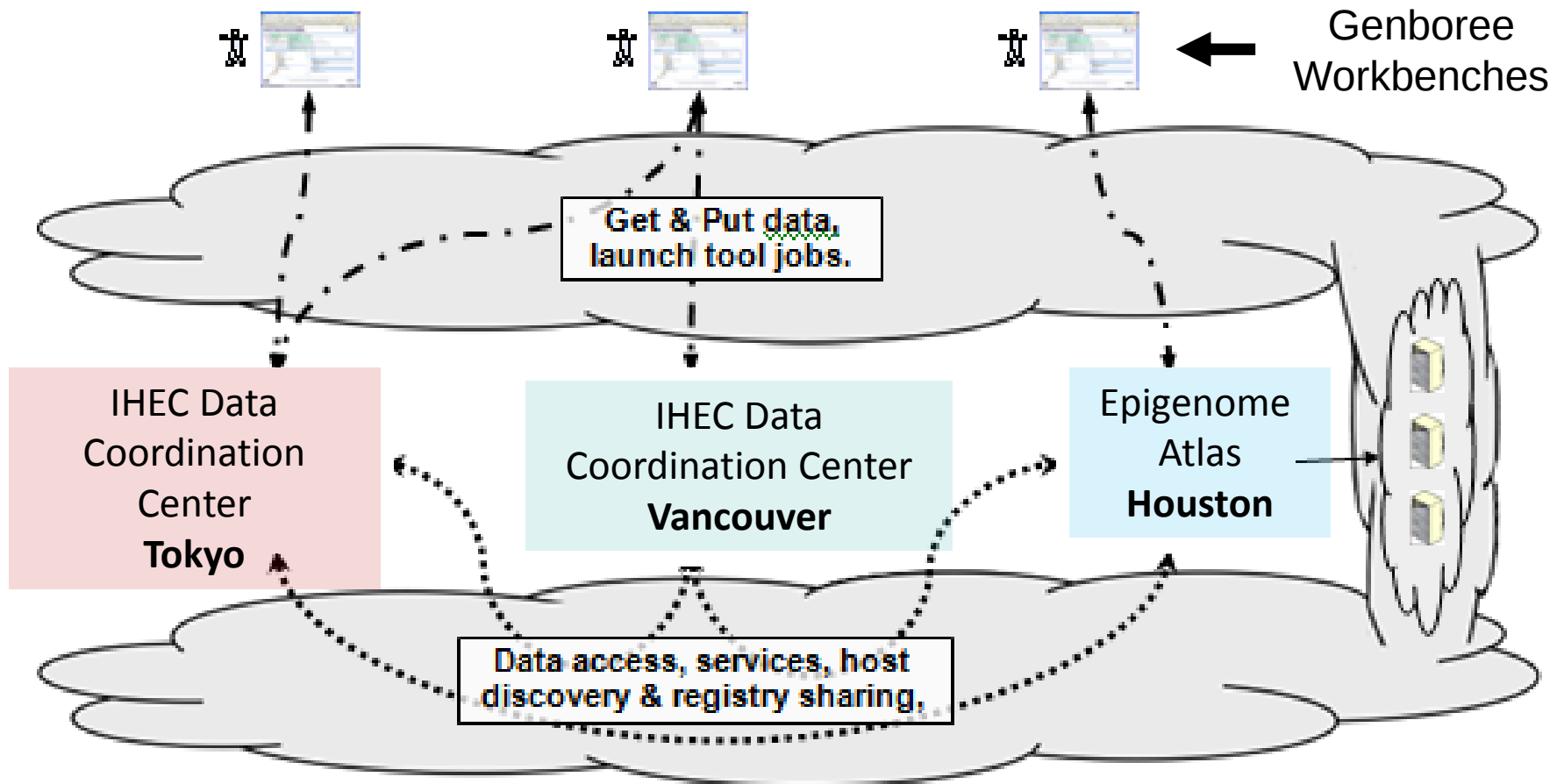
Epigenomic Assays

Assay	Vancouver	Tokyo	Houston
ChIP-Seq	1	1	1
H3K27me3	1	1	1
H3K36me3	1	1	1
H3K4me1	1	1	1
H3K4me3	1	1	1
H3K9ac	1	1	1
H3K9me3	1	1	1
H2A/Gac	1	1	1
H2BK12ac	1	1	1
H2BK15ac	1	1	1
H2BK12ac	1	1	1
H3K14ac	1	1	1
H3K18ac	1	1	1
H3K23ac	1	1	1
H3K27ac	1	1	1
H3K4ac	1	1	1
H3K4me2	1	1	1
H3K56ac	1	1	1
H3K79me1	1	1	1
H3K79me2	1	1	1
H4K20me1	1	1	1
H4K5ac	1	1	1
H4K8ac	1	1	1
H4K91ac	1	1	1
H2A.Z	1	1	1
H2A/Gac	1	1	1

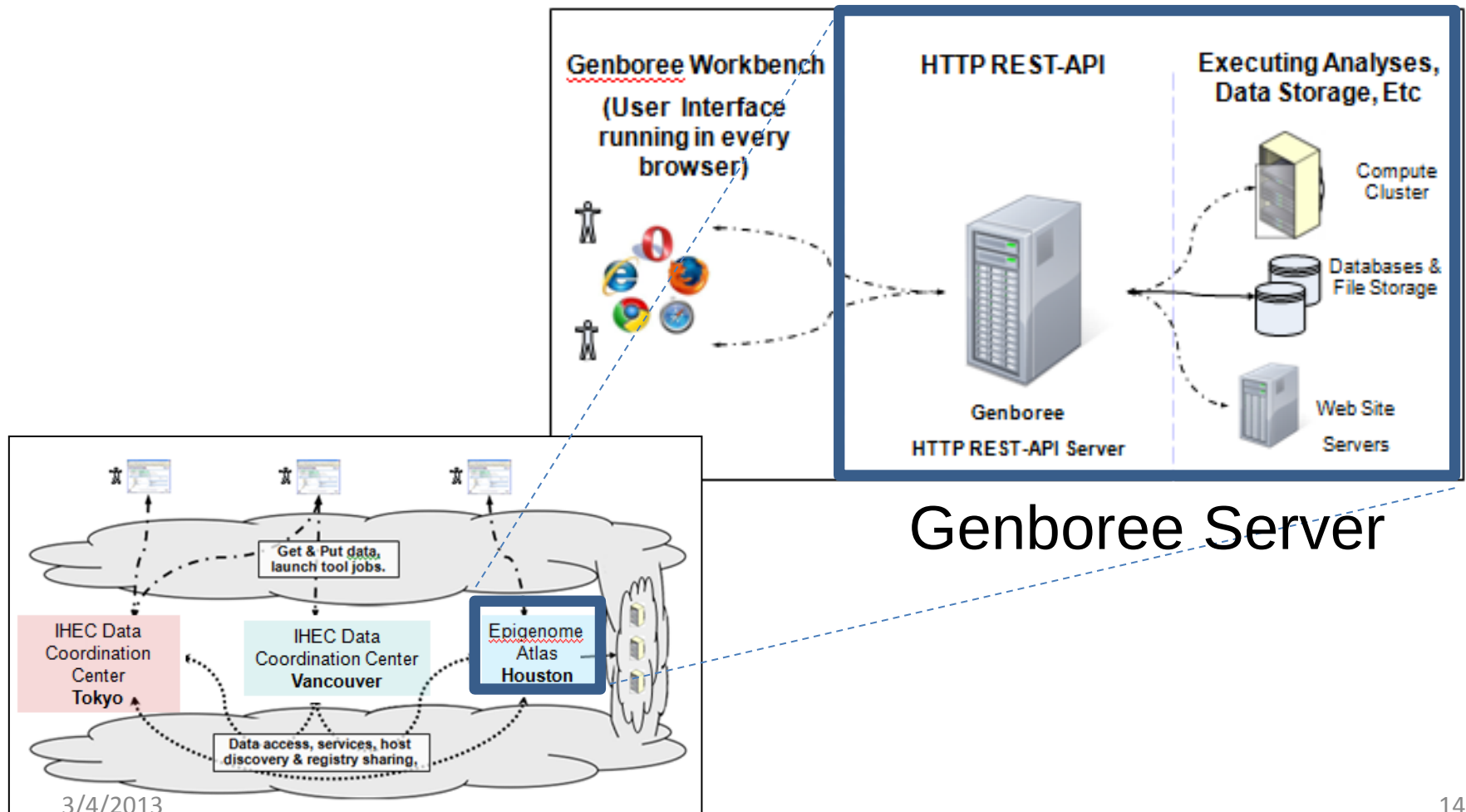
Samples (by Type)

A pilot project in virtual data integration within IHEC using Genboree

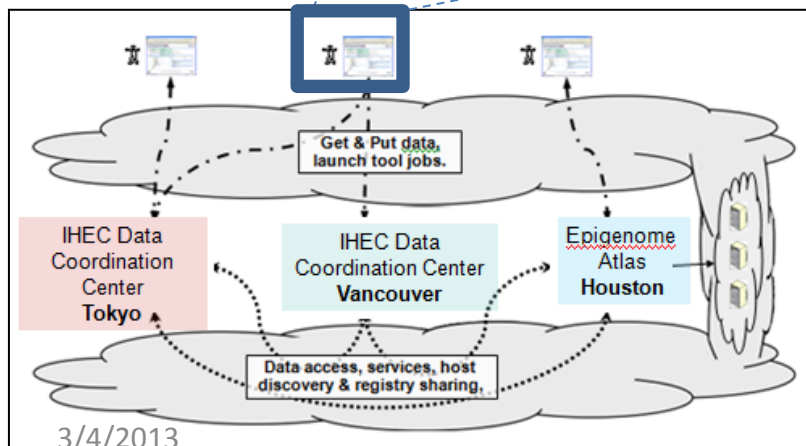
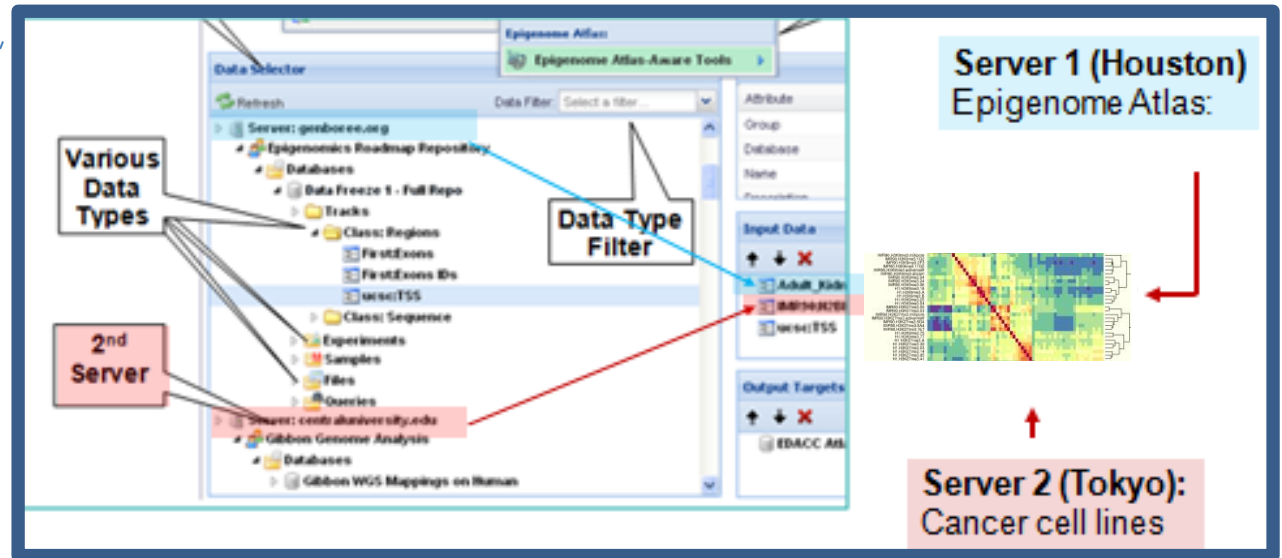
IHEC: International Human Epigenome Consortium



“Programmable Web”: REpresentational State Transfer (REST) Application Programming Interfaces (APIs)



Drag-and Drop comparison of epigenomes across different Genboree servers

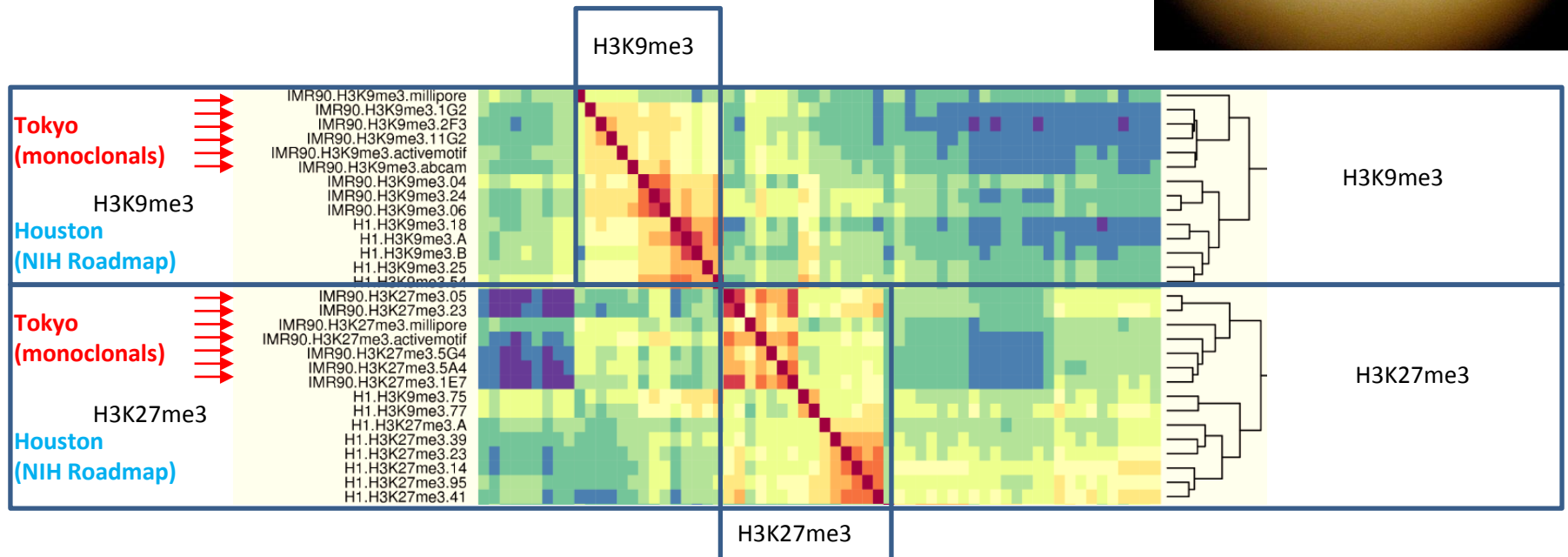


Genboree Workbench
(30+ Tools 50+ Utilities)

Virtual integration across IHEC DCCs

Only relevant portions of H3K9me3 and H3K27me3 signals (averages over promoters) were transferred between Tokyo and Houston.

Reduced data transfer
From 10 GB to 10 MB



Collaboration with

Drs. **T. Ushijima**, National Cancer Center, Tokyo, **T. Mituyama**, AIST, Tokyo, **Y. Kanai**, National Cancer Center, Tokyo, **H. Aburatani**, Univ. of Tokyo, **K. Shirahige**, Univ. of Tokyo, **Y. Wada**, Univ. of Tokyo, **H. Kimura**, Osaka Univ., **Y. Suzuki**, Univ. of Tokyo

Virtual integration principles

- Data integrated **just in time** for analysis. No data hoarding!
- Seamless integration **by “apps” accessing multiple sites** on behalf of the user.
- Data integrated **at the highest level of processing (most compressed)**.
- **Only the data relevant** for particular analysis (e.g., epigenomic marks over enhancers or promoters, not the whole genome).

Requirements for virtual data integration

- **Data standards**
(NIH Roadmap, IHEC)
- **Metadata standards**
(NIH Roadmap, IHEC)
- **Web-based interoperability**
(Genboree REST APIs)

Epigenomic data and metadata standards

www.ihec-epigenomes.org

IHEC
International Human Epigenome Consortium

Home Areas of focus **Standard Operating Procedures** Tools / Useful Information Policies and Guidelines IHEC Structure Outreach and Training Feedback Intranet

Links

The EPIGENOME Network of Excellence
The focal point for the European epigenetics research community

Centre for Epigenetics
Centre of Excellence funded by The Danish National Research Foundation.

IHEC Recommendations for Epigenomic Analysis (DRAFT)

Download

Data and Metadata Models Developed by the NIH Roadmap Epigenomics Project

(Under Consideration by the International Human Epigenome Consortium (IHEC) as a Basis for IHEC Recommendation)

These assay standards developed/in use by the NIH Roadmap Reference Epigenome Mapping Centers are being considered for potential use by IHEC.

RNA-seq Standards

ChIP-seq Standards

MethylC-seq Standards

Data and Metadata Models Developed by the NIH Roadmap Epigenomics Project

(Under Consideration by the International Human Epigenome Consortium (IHEC) as a Basis for IHEC Recommendation)

R. Alan Harris¹, Robert A. Waterland^{1,2}, Ryan Lister³, Martin Hirst⁴, Aleksandar Milosavljevic¹

¹Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

²USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA.

³Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California, USA.

⁴Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada.

Genboree Installations

Other locations that use Genboree



Installation in-progress,
Targeted completion 2Q13



Installation completed,
Dec 2012



**Houston
EDACC**

Cloud: Elastic Computing Plus Dedicated Hosting

The screenshot shows the Genboree Workbench interface. At the top, there are tabs for System, Data, Analysis, and Visual. Below the tabs, a welcome message is followed by a list of instructions: 'The Data Selector tree on the left side', 'Drag items to be used as tool inputs', 'Drag items to be used as output destinations', 'Tools which can be run on your selected data', and 'Unsure about what kinds of items a particular tool can take? Just click the tool button when it is selected'. The Data Selector tree on the left contains a 'Refresh' button and a list of projects: 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, and zuozhouc_group. At the bottom, two links are highlighted: 'www.brain-research-lab.org' and 'www.microbiome-center.org'.

The screenshot shows the 'Brain Research Lab #1' Genboree hosted site. The header is 'Brain Research Lab #1' with the Genboree logo and 'hosted site' text. Below the header, it says 'You are currently not logged in.' There is a login section with fields for 'Login Name:' and 'Password:', a 'Login' button, and a link '(Forgot your password?)'. Below the login section, there is a 'Guest/Public View' button. To the left of the login section, there are two links: '< 3rd Epigenome Informatics Workshop >' and '< Support Site >'. The workshop link has a description: 'We have prepared some resources for attendees of the March 2012 Workshop: * FAQs * Use Cases'. The support site link has a description: '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'. At the bottom, there is a link 'New to Genboree? Register here!' and 'Genboree access at www.genboree.org'.

The screenshot shows the 'Microbiome Center #1' Genboree hosted site. The header is 'Microbiome Center #1' with the Genboree logo and 'hosted site' text. Below the header, it says 'You are currently not logged in.' There is a login section with fields for 'Login Name:' and 'Password:', a 'Login' button, and a link '(Forgot your password?)'. Below the login section, there is a 'Guest/Public View' button. To the left of the login section, there are two links: '< 3rd Epigenome Informatics Workshop >' and '< Support Site >'. The workshop link has a description: 'We have prepared some resources for attendees of the March 2012 Workshop: * FAQs * Use Cases'. The support site link has a description: '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'. At the bottom, there is a link 'New to Genboree? Register here!' and 'Genboree access at www.genboree.org'.

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