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

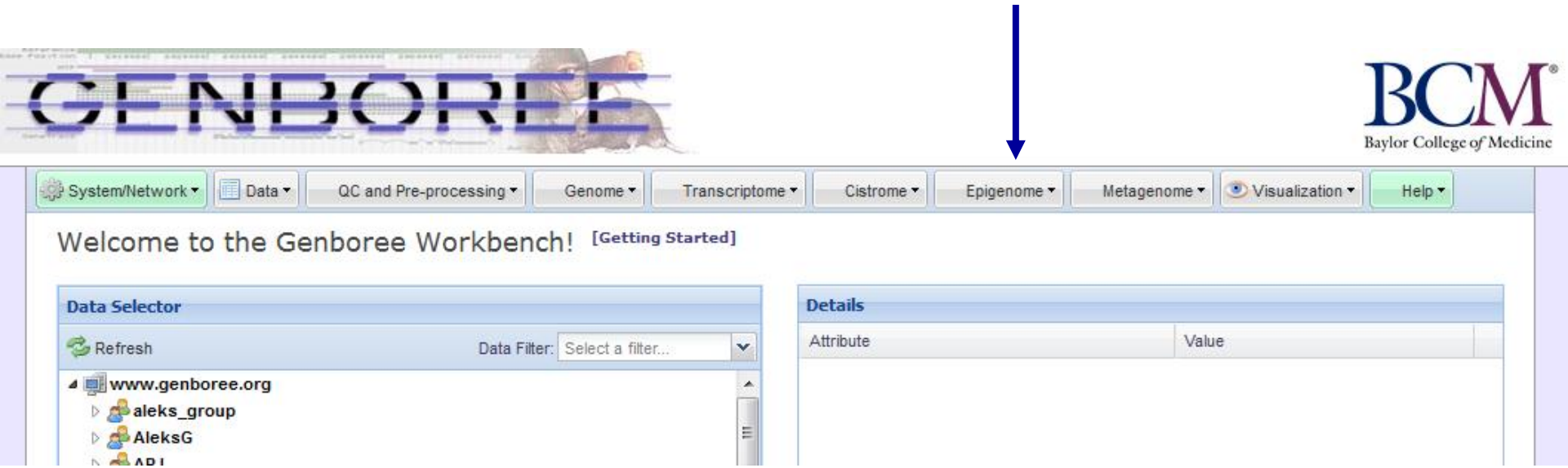
Two key integration problems

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

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

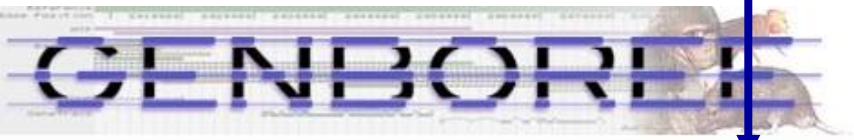
Epigenomic Toolset integrates Spark developed by the British Columbia Genome Center in Vancouver

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

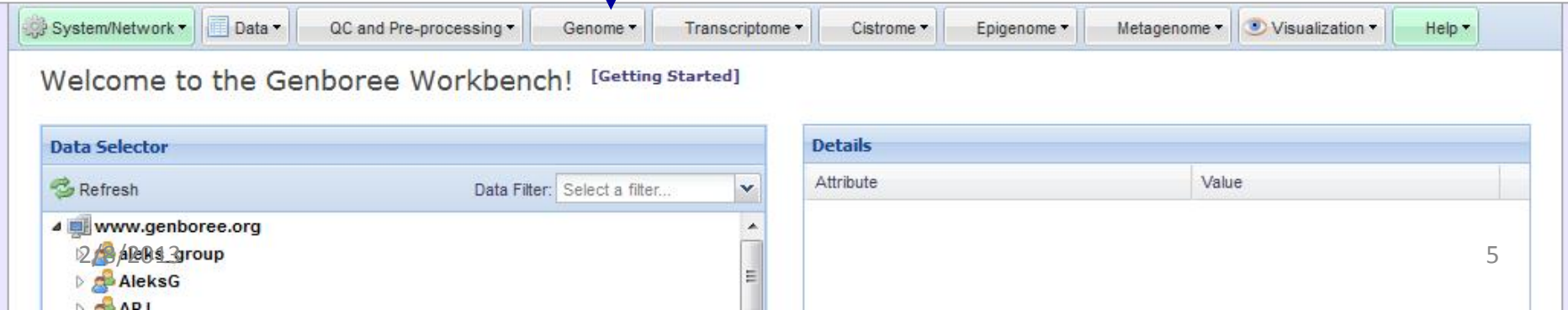


Genomic Toolset developed in collaboration with the Baylor Genome Center

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

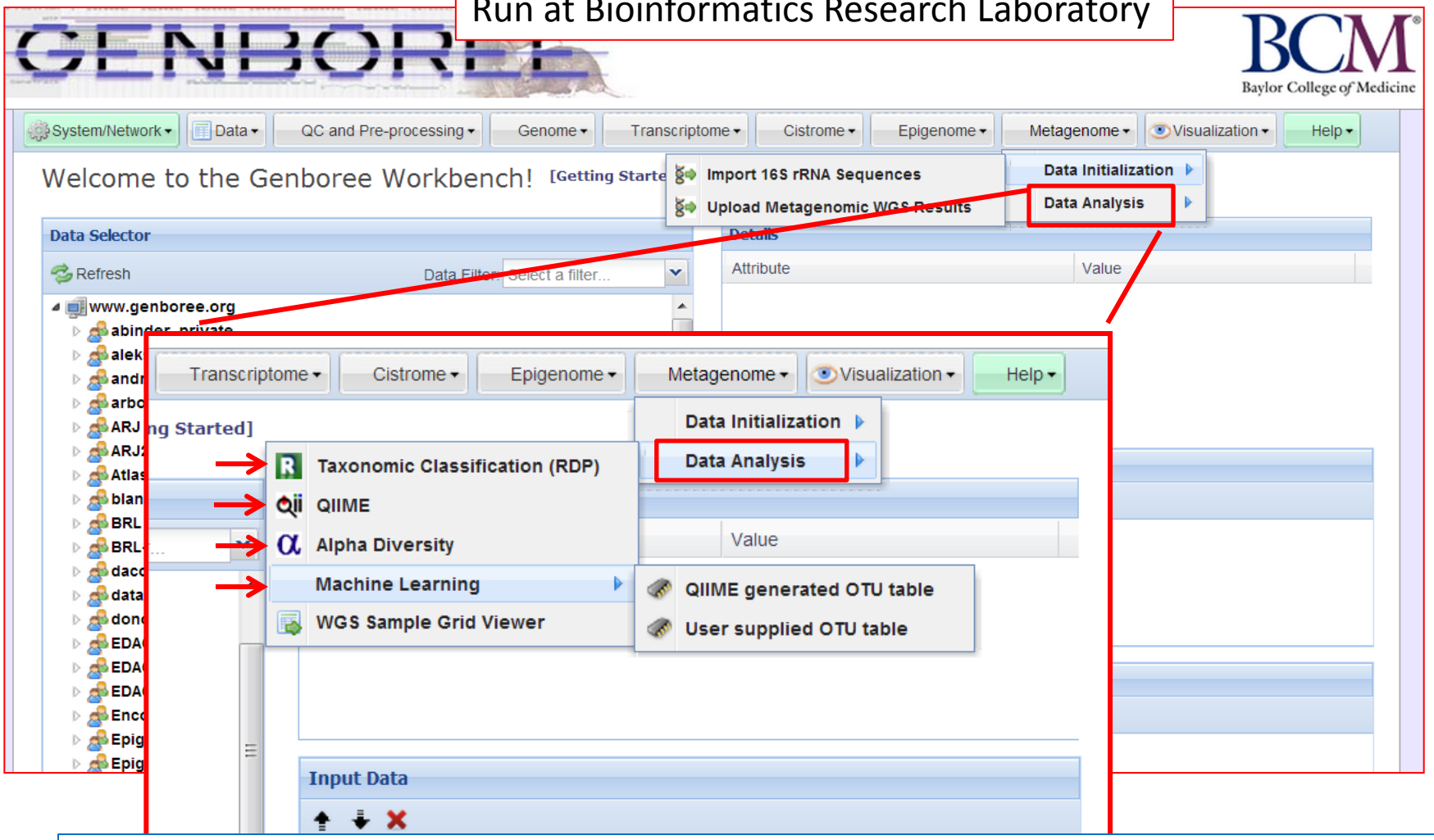


BCM
Baylor College of Medicine



Microbiome Toolset includes QIIME (U. Colorado)

Run at Bioinformatics Research Laboratory



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

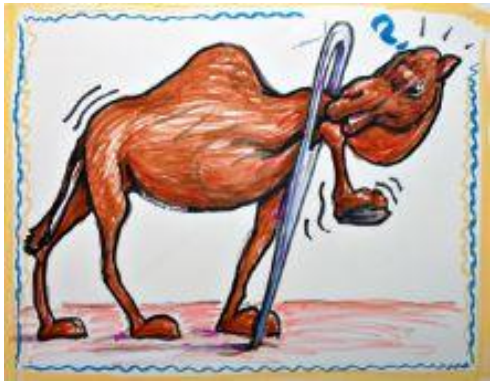
Other widely used tools are integrated within the Genboree Workench

- 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.

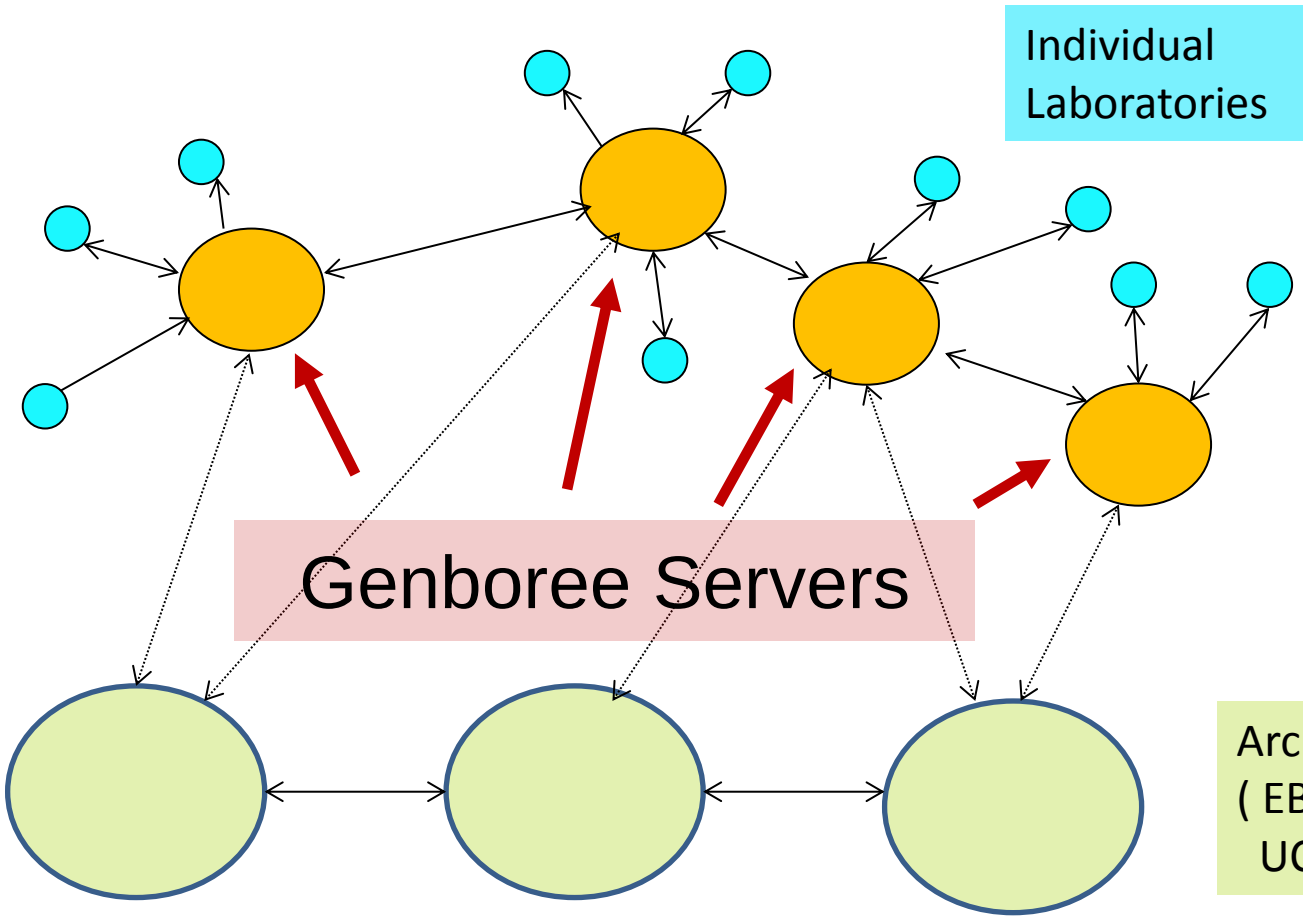


As the data volume explodes, ***physical*** data integration will become impossible

Solution: ***virtual*** data integration

Data Ecosystem Model proposed for IHEC

IHEC: International Human Epigenome Consortium

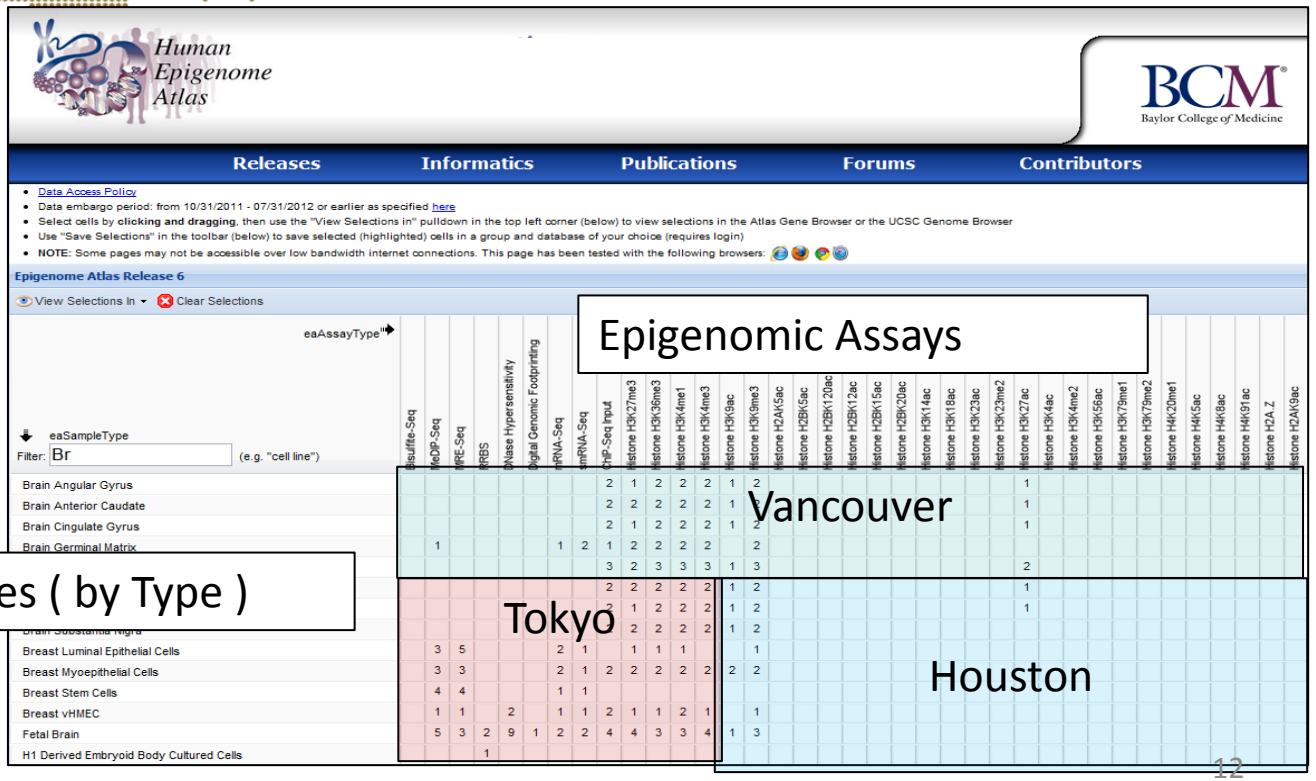


Virtual integration of reference epigenomes across three IHEC Data Centers

Vancouver
Steven Jones

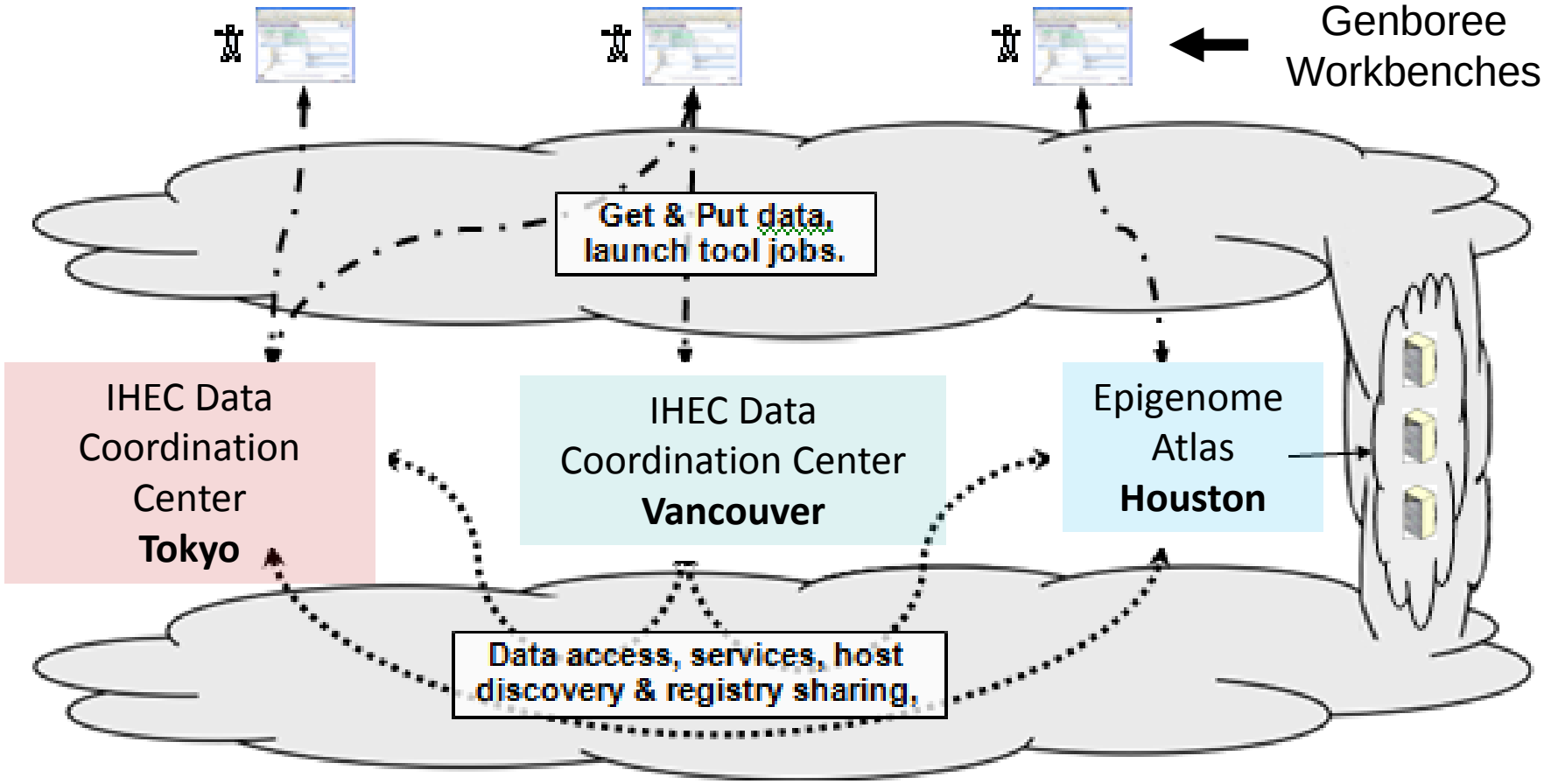
Houston
EDACC

Tokyo, Toutai Mituyama

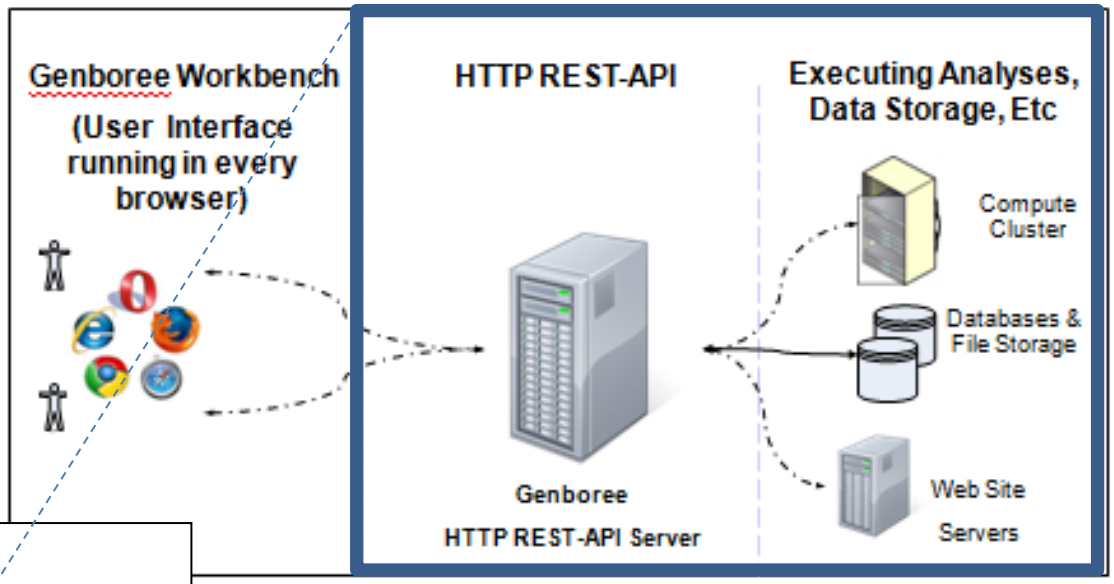


A pilot project in virtual data integration within IHEC using Genboree

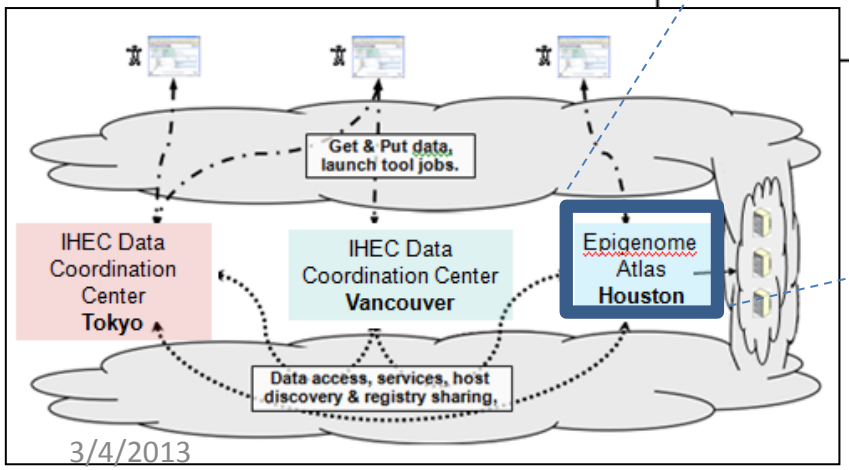
IHEC: International Human Epigenome Consortium



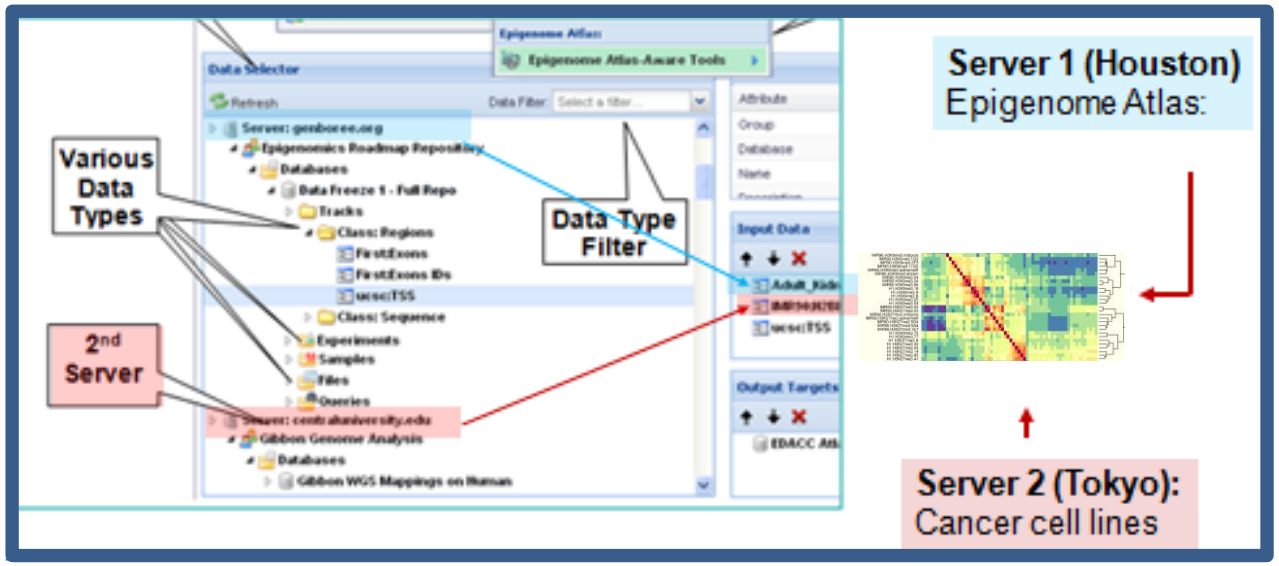
“Programmable Web”: REpresentational State Transfer (REST) Application Programming Interfaces (APIs)

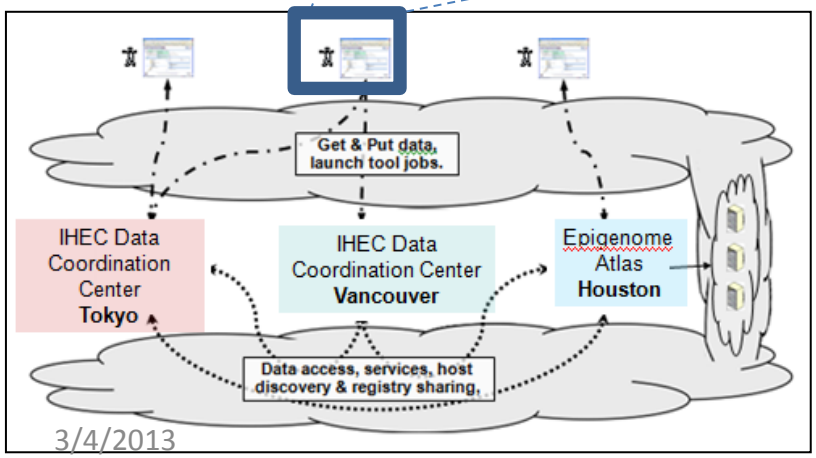


Genboree Server



Drag-and Drop comparison of epigenomes across different Genboree servers



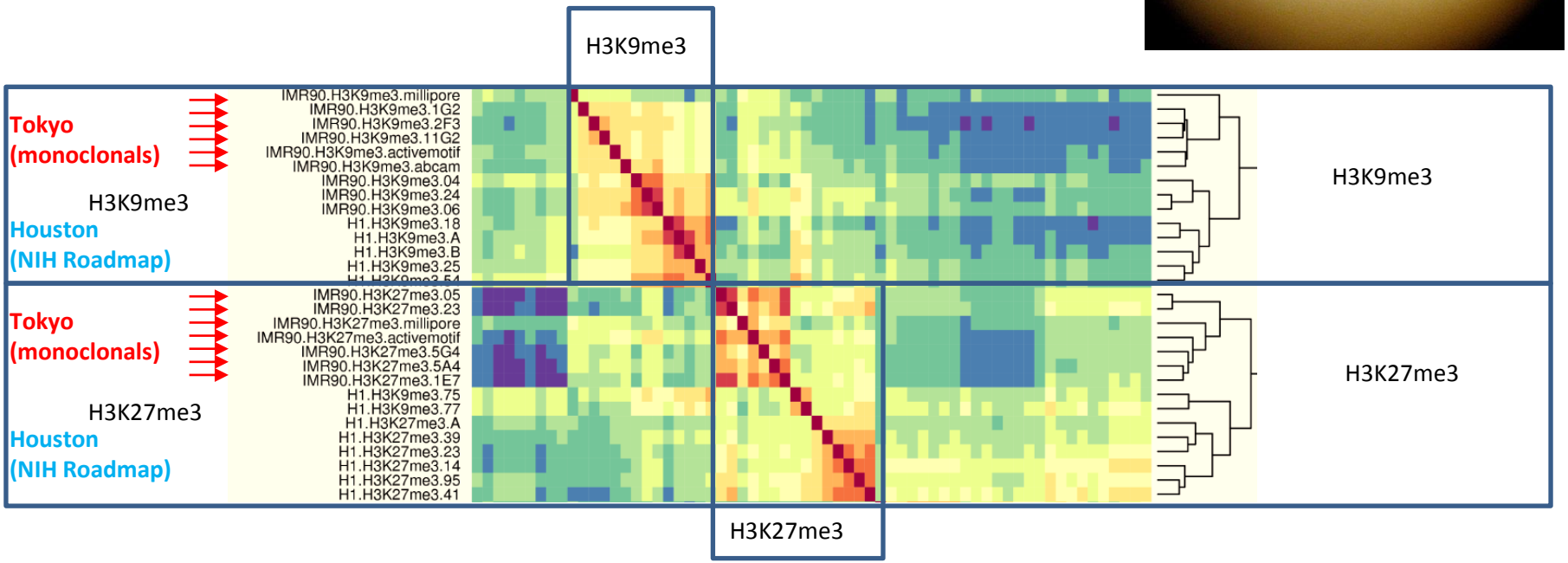


Genboree Workbench

(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.



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).

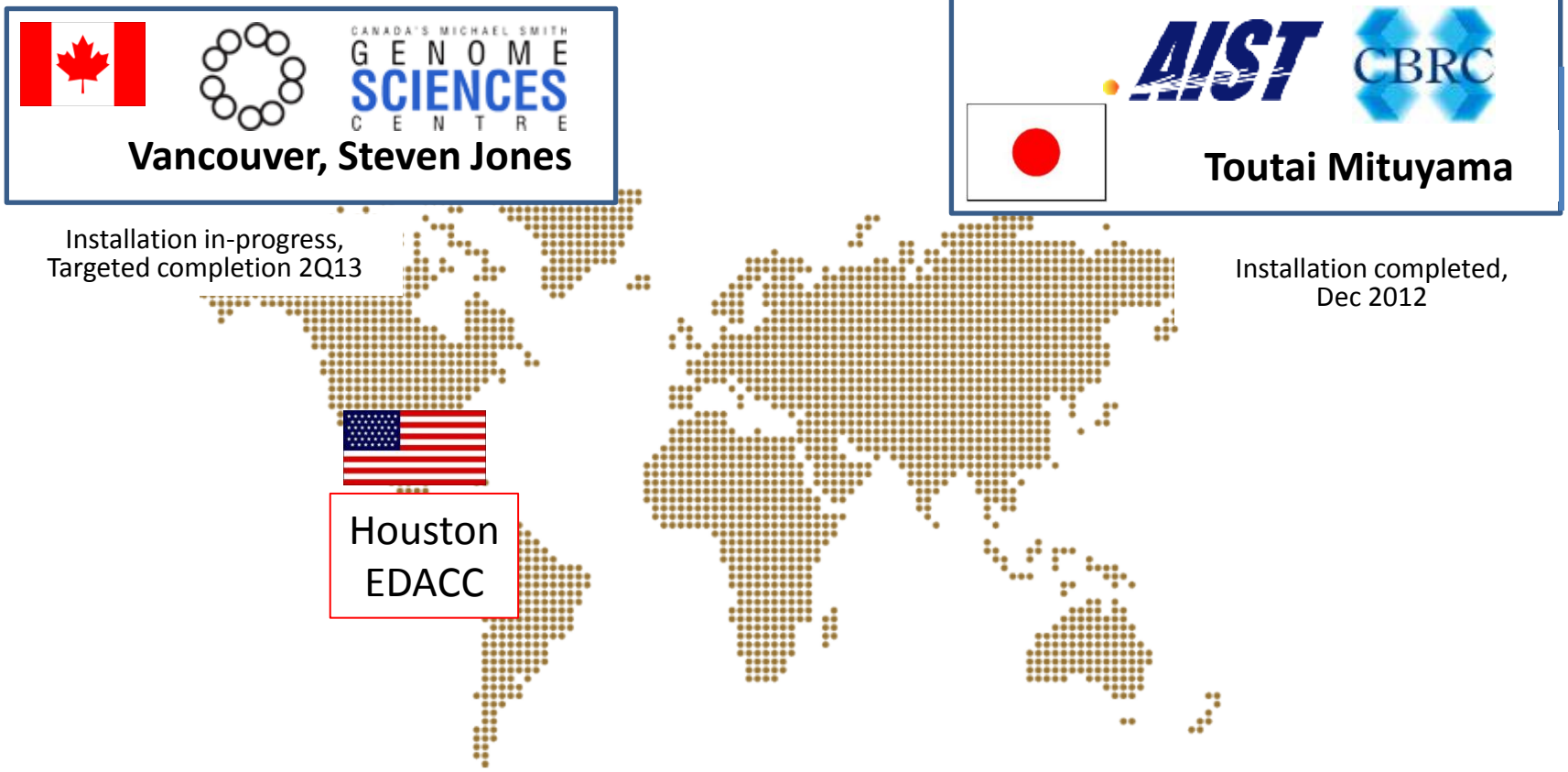
Requirements for virtual data integration

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

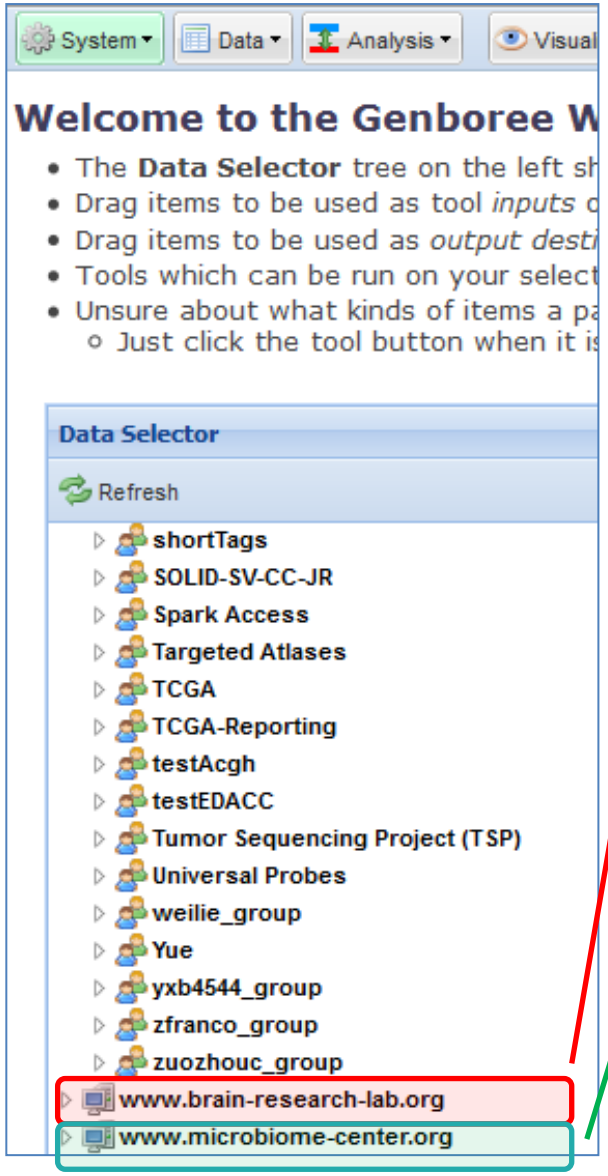
Epigenomic data and metadata standards
www.ihec-epigenomes.org

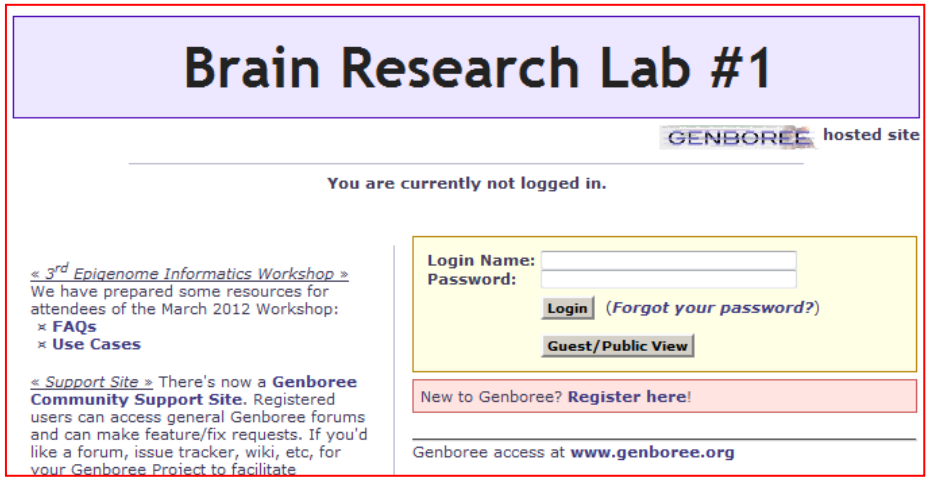
Genboree Installations

Other locations that use Genboree



Cloud: Elastic Computing Plus Dedicated Hosting





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