Computational Challenges for Epigenomics in the Massively Parallel Sequencing Era

Andrew McLellan, PhD

Center for Epigenomics, Albert Einstein College of Medicine,
Price Center for Genetics and Translational Medicine,
1301 Morris Park Avenue, Bronx, NY 10461, USA.

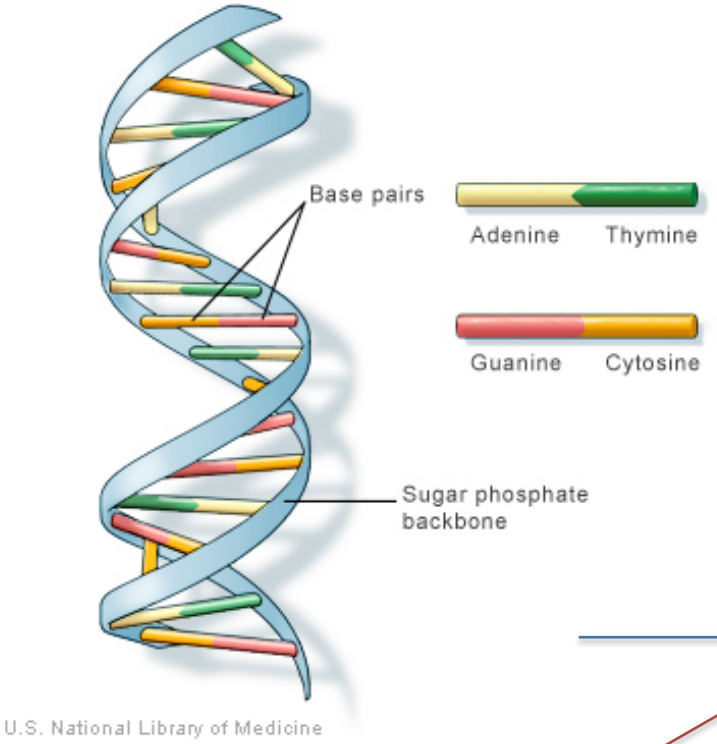
center for epigenomics | Einstein

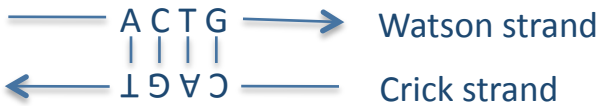
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Overview of Presentation

- Epigenetics: why we study it and how we study it at Einstein
- Massively Parallel DNA sequencing technology in use at Einstein
- Genome-wide epigenomic assays and computational epigenomics
- WASP: simplifying data management and analysis using HPC
- Data growth / current and future 'Big Data' problems /data integration challenges

Genetics 101





Watson strand

Crick strand

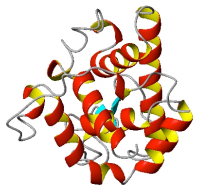
blueprint for making a protein

Exons

Promoter

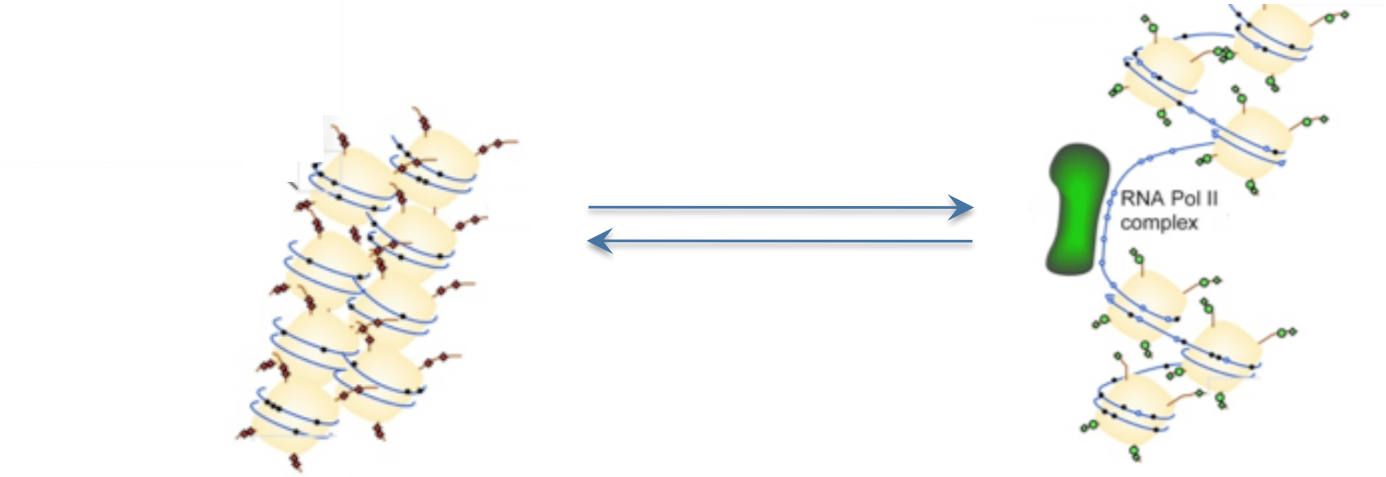
Contains DNA sequence patterns recognized by proteins that regulate / actuate gene expression or 'reading' of the gene (a sort of genetic switchboard)

Protein product



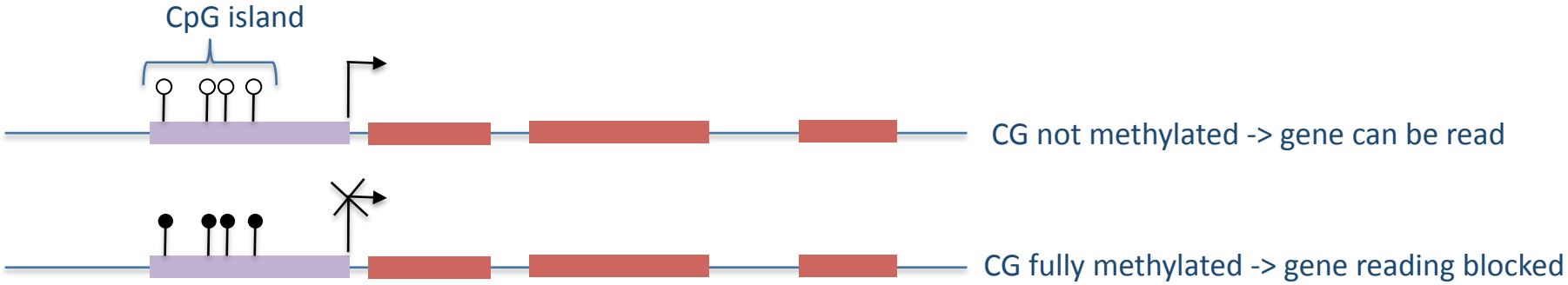
Epigenetics 101

- Gene regulation without change of underlying DNA sequence



Silenced State
- DNA tightly wrapped (compressed)
- ALL Genes in region not accessible

Active State
- DNA accessible to transcription factors
- Genes in region accessible and can be read



Epigenetics 101

- Gene regulation without change of underlying DNA sequence
- DNA sequence identical in all cells of an individual
  - reading of which genes when determines function (gene regulation)
- Heritable and Reversible Changes
  - modifications maintained during cell division
  - possible incomplete erasure during reprogramming ('factory reset') in early embryo
- Modifications accumulate throughout life
  - DNA methylation increases
  - lifestyle impacts
- Disease implications – when things go wrong!
  - Cancer, mental illness, autoimmune disease and ageing

Epigenomic Dysregulation Studies at Einstein

- **Center for Epigenomics**

  - Human disease:
    - cancer epigenomics, neuroepigenomics, epigenomics of infectious disease, ageing, renal disease, diabetes

    - epigenetic drug development

  - Basic science projects:
    - mechanisms of epigenomics

    - functional epigenomics program

    - computational epigenomics

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Einstein Shared Core Facility Assets for MPS

Illumina GA IIx

x2

August 2010

Illumina HiSeq 2000

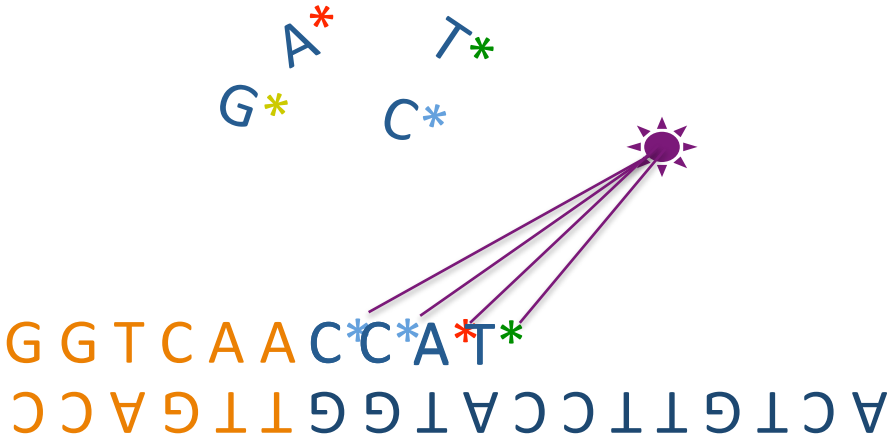
- Up to 50 Gbases per run (5Gbases per day)
- ~ll days from raw sample to aligned sequence
- up to ~30 GB sequence data / ~5 TB images

- Up to 200 Gbases per run (25 Gbases per day)
- ~ll days from raw sample to aligned sequence
- up to ~120 GB sequence data / ~32 TB images
- ~30x coverage of two human genomes in a single run for under $10,000

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

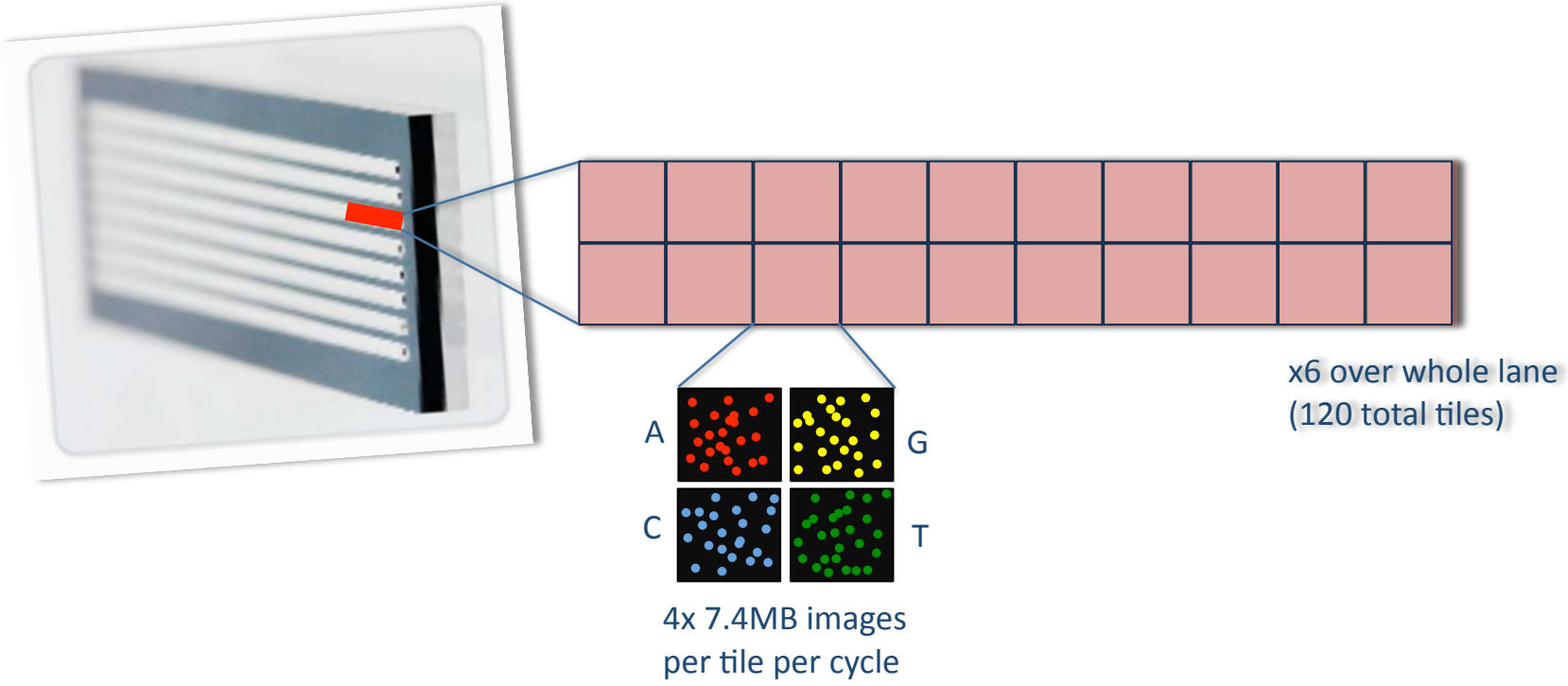
Sequencing By Synthesis



Cycle: 1 2 3 4

Base Calls: C C A T

Illumina GA IIx



8 lanes x 120 tiles x 200 cycles (100 base paired-end) x 4 images (A/C/T/G) x 7.4 MB = 5.4 TB images

Genome-Wide Epigenomic Assays

- Map modification (acetylation / methylation) of histones
  - Chromatin immunoprecipitation (ChIP) – seq modifications to genome
- Can examine / quantify methylation marks on DNA
  - HELP-tagging: measuring methylation at CCGG sites
  - Bisulphite: measure methylation of all C -> currently an expensive option
- Association studies
  - compare disease with normal
  - compare different tissue types or cell types

Computational Epigenomics

- Primary analysis and visualization of data -> assay specific applications
- Compare disease / normal epigenomes -> statistical analysis
- Machine learning tools for classification (SVM / ANN / SOM etc)
- Correlate data with results from other experiments and existing genome annotations to make functional predictions

HELP-Tagging Assay

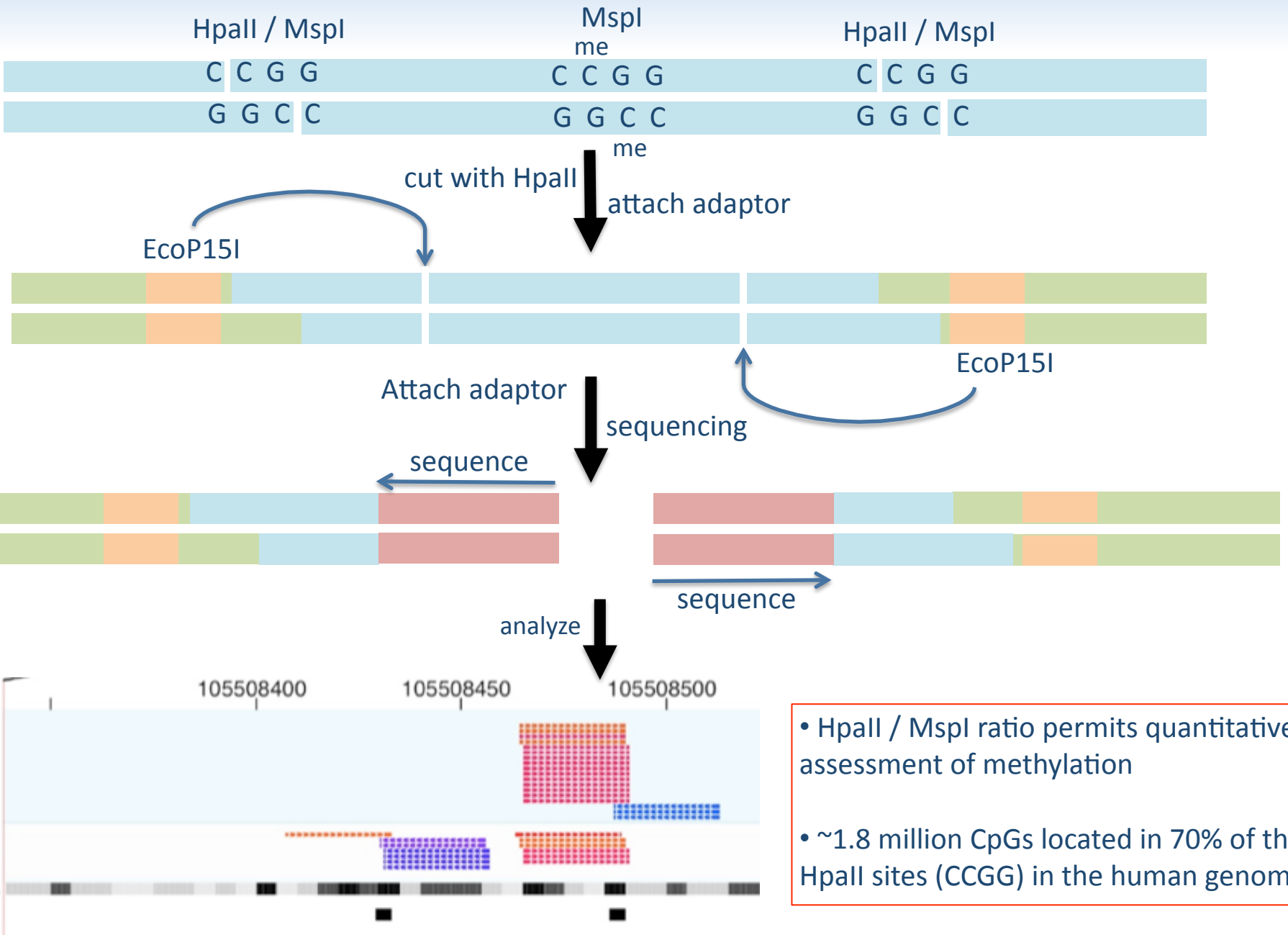
Suzuki et al. Genome Biology 2010, 11:R36
http://genomebiology.com/2010/11/4/R36

METHOD Open Access

Optimized design and data analysis of tag-based cytosine methylation assays

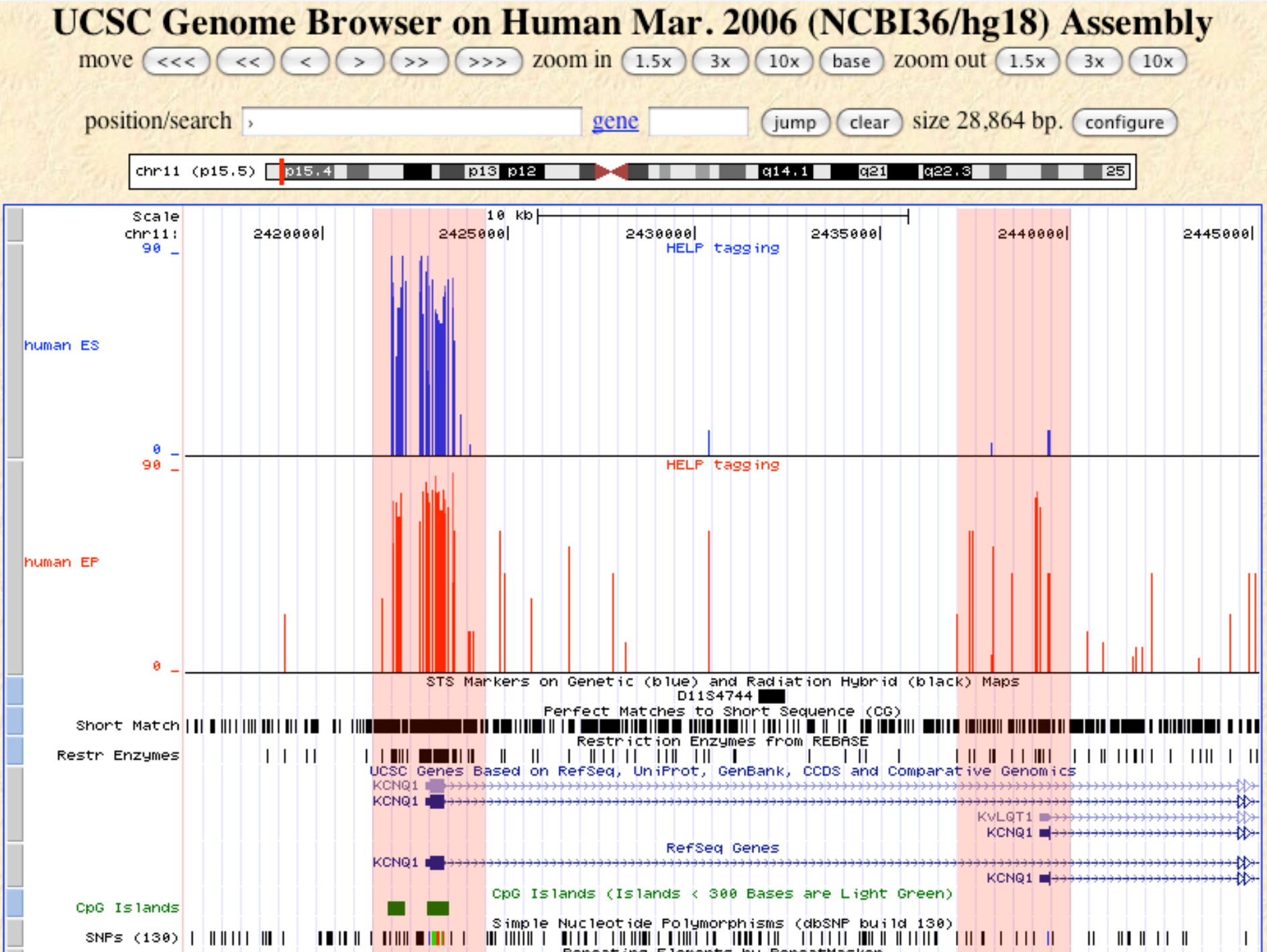
Masako Suzuki, Qiang Jing, Daniel Lia, Marién Pascual, Andrew McLellan and John M Greally*

HELP-Tagging Assay



- HpaII / MspI ratio permits quantitative assessment of methylation
- ~1.8 million CpGs located in 70% of the HpaII sites (CCGG) in the human genome

Genome Browser: HELP-tag



Integrating Data Management and Analysis

- HELP-tag analysis -> large computational overhead.
- Primary analysis consistent –> automation / HPC.
- Each lane of uncompressed Illumina sequences is > 2Gb and growing!!
  - Data transfer across network – keep analysis close to data source
  - Storage overhead
  - Processing overhead:- alignment, trimming, transforming
- Analyses and visualization of results
  - User friendly interpretation – graphics on genome browser
  - Dimensionality reduction (filter noise / select informative)
  - Statistical analysis (fit models / assess thresholds / assess significance)
  - Integration with other data sets / genome annotations

NATURE|Vol 464|1 April 2010 HUMAN GENOME AT TEN **NEWS FEATURE**

HUMAN GENOME

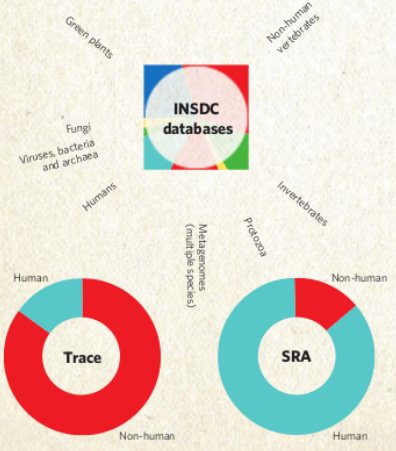
THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

See Editorial, page 649, and human genome special at www.nature.com/humangenome

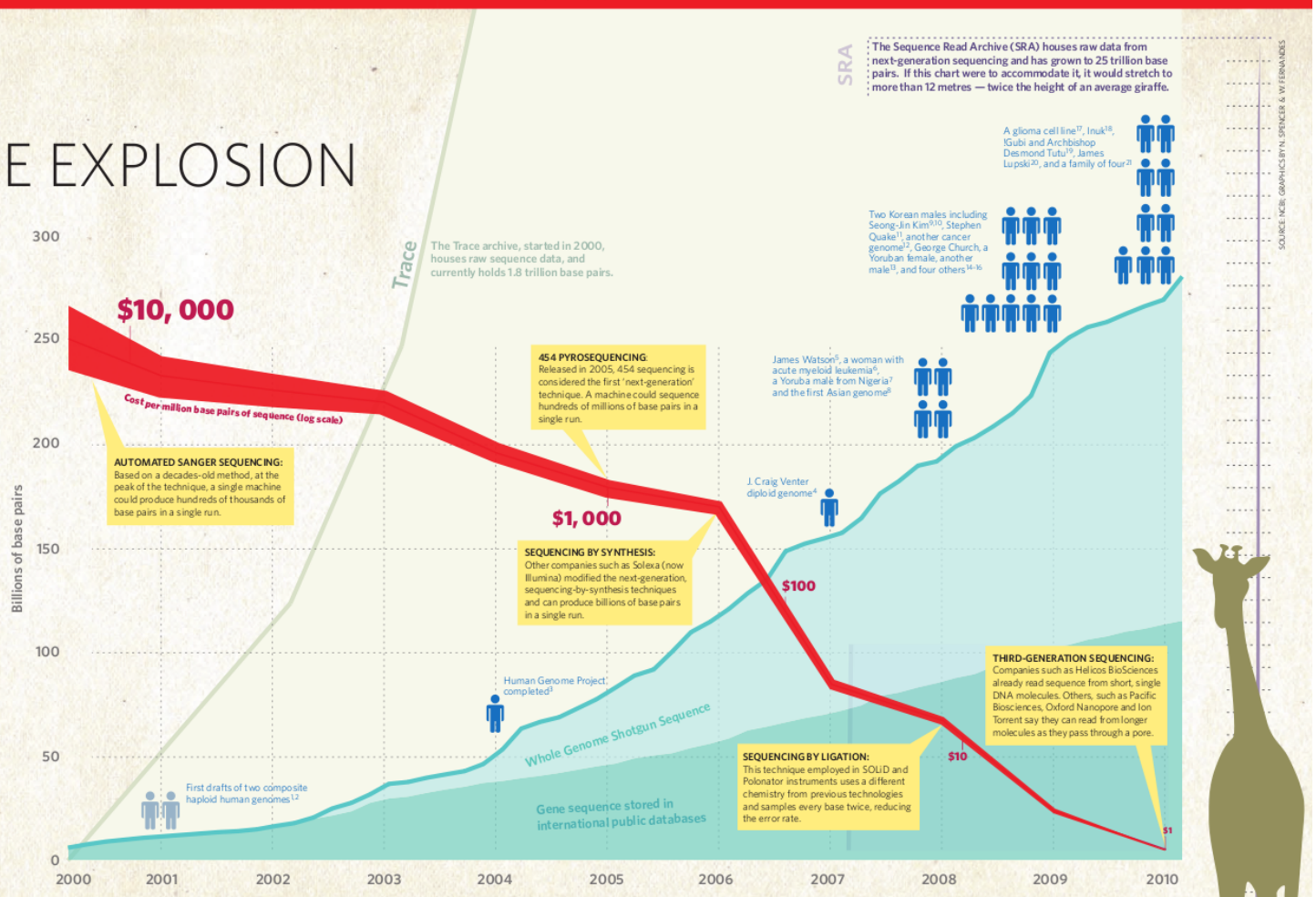
DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.



Trace Archive: Developed to house the raw output of high-throughput sequencers built in the late 1990s, the trace archive spans a wide range of taxa.

Sequence Read Archive: Houses raw data from next-generation sequencers. Dominated by human sequence, including multiple coverage for more than 170 people.



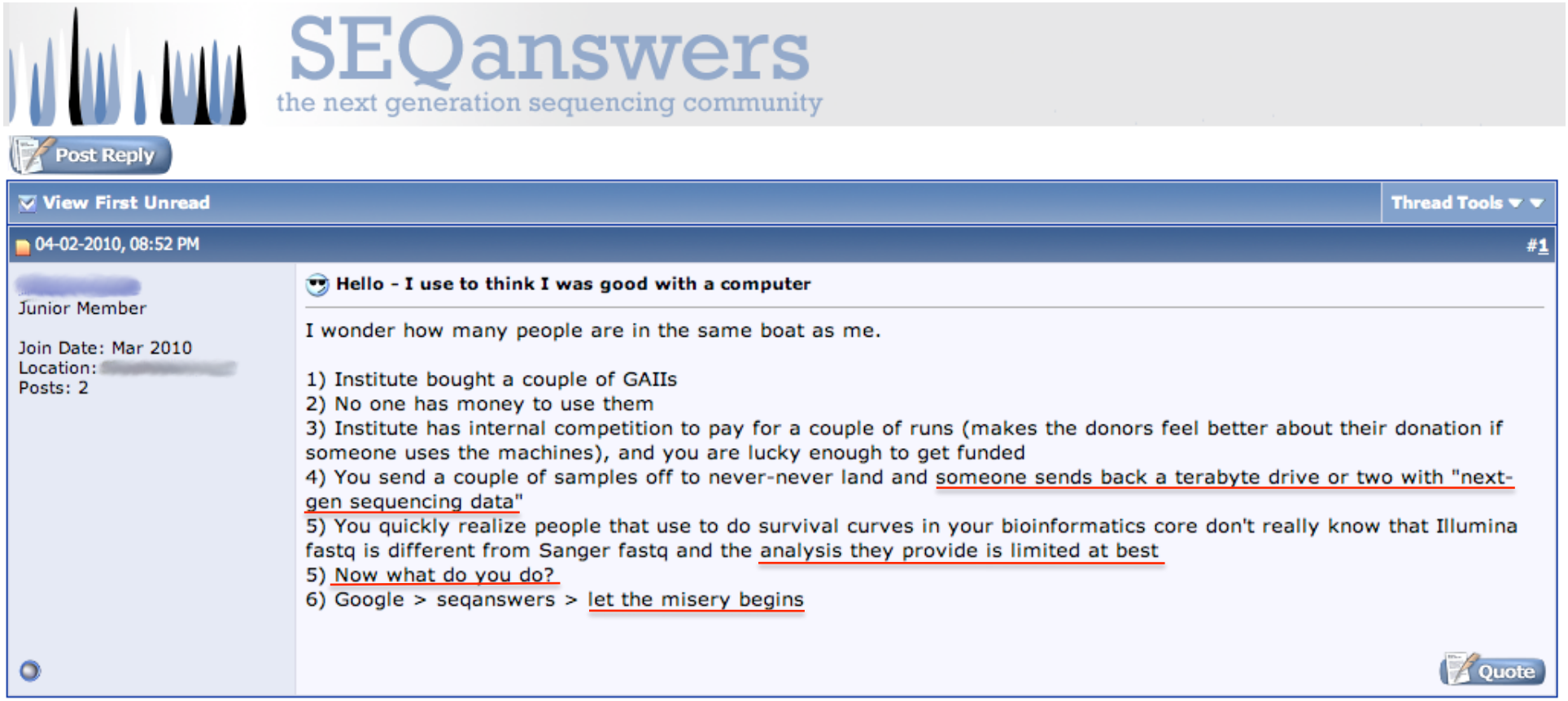
HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.

1. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
5. Wheeler, D. A. *et al. Nature* **452,** 872–876 (2008).
6. Ley, T. J. *et al. Nature* **456,** 66–72 (2008).
7. Bentley, D. R. *et al. Nature* **456,** 53–59 (2008).
8. Wang, J. *et al. Nature* **456,** 60–65 (2008).
9. Ahn, S.-M. *et al. Genome Res.* **19,** 1622–1629 (2009).
10. Kim, J.-I. *et al. Nature* **460,** 1011–1015 (2009).
11. Pushkarev, D., Neff, N. F. & Quake, S. R. *Nature Biotechnol.* **27,** 847–850 (2009).
12. Mardis, E. R. *et al. N. Engl. J. Med.* **10,** 1058–1066 (2009).
13. Drmanac, R. *et al. Science* **327,** 78–81 (2009).
14. McKernan, K. J. *et al. Genome Res.* **19,** 1527–1541 (2009).
15. Pleasance, E. D. *et al. Nature* **463,** 191–196 (2010).
16. Pleasance, E. D. *et al. Nature* **463,** 184–190 (2010).
17. Clark, M. J. *et al. PLoS Genet.* **6,** e1000832 (2010).
18. Rasmussen, M. *et al. Nature* **463,** 757–762 (2010).
19. Schuster, S. C. *et al. Nature* **463,** 943–947 (2010).
20. Lupski, J. R. *et al. N. Engl. J. Med.* doi:10.1056/NEJMoa0908094 (2010).
21. Roach, J. C. *et al. Science* doi:10.1126/science.1186802 (2010).

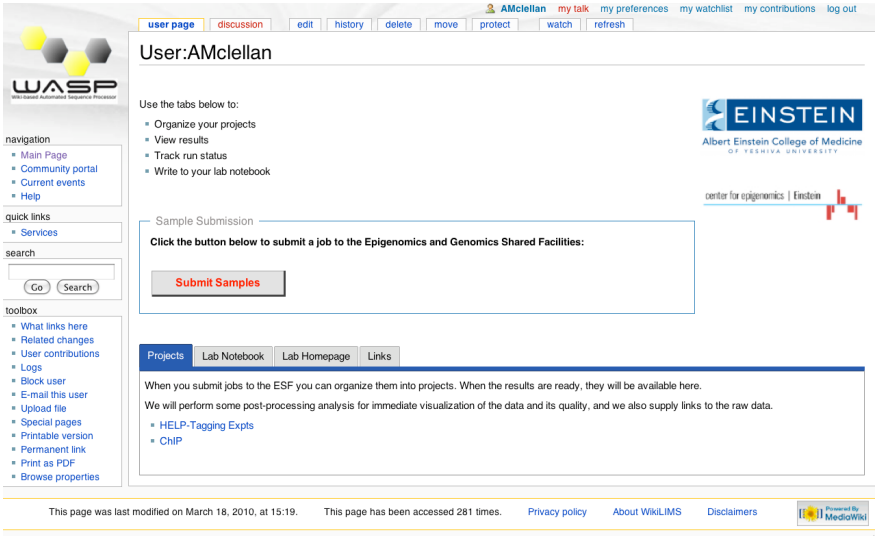
Page size by comparison

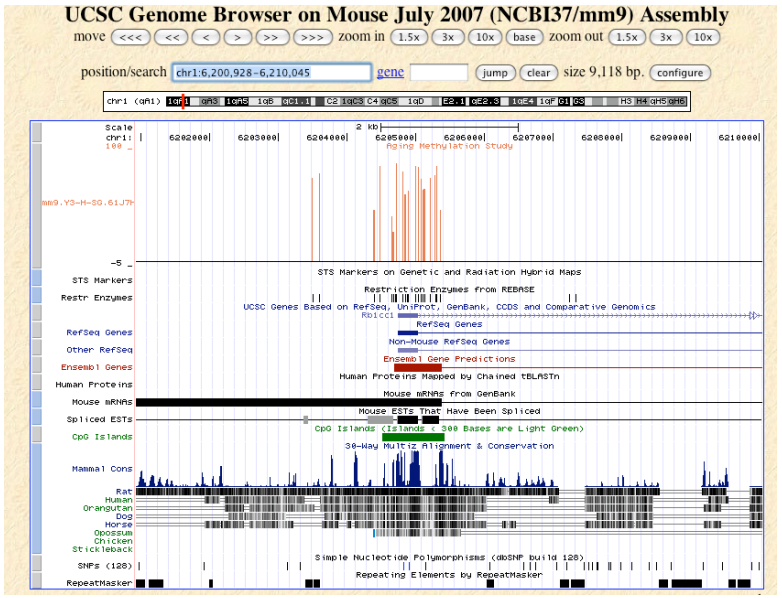
Bench Scientist's Dilemma

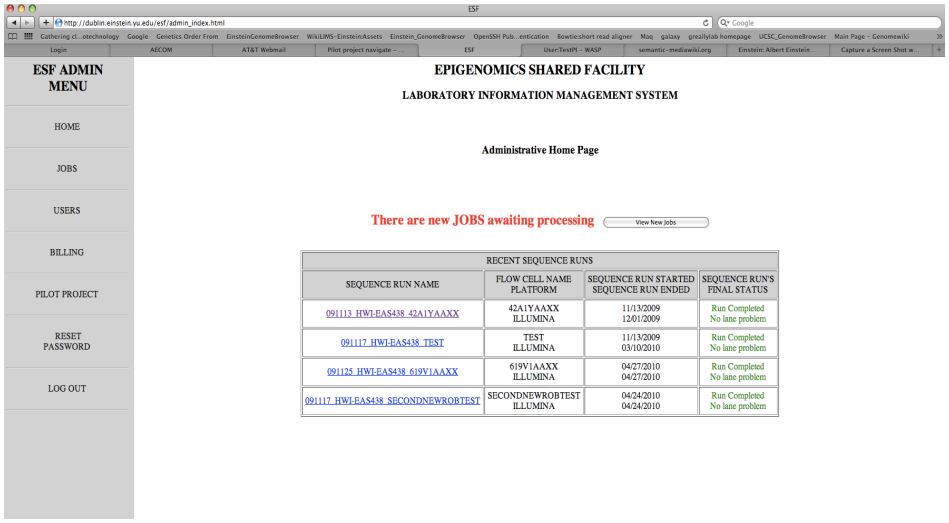


Getting easier and cheaper to generate lots of DNA sequence but what about handling and analyzing it??

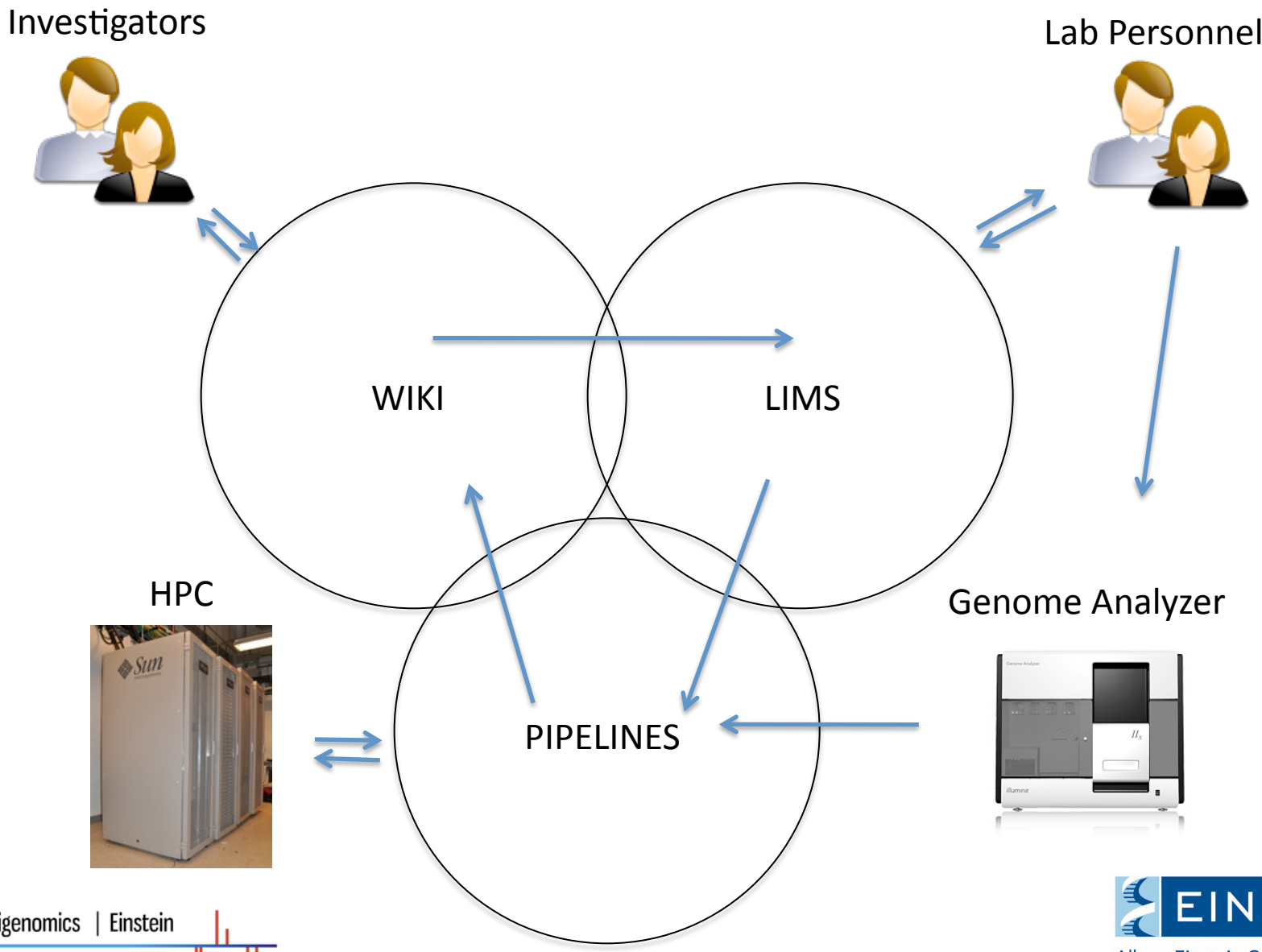

WASP
Wiki-based Automated Sequence Processor


AMclellan my talk my preferences my watchlist my contributions log out
user page discussion edit history delete move protect watch refresh
WASP
User:AMclellan
Use the tabs below to:
Organize your projects
View results
Track run status
Write to your lab notebook
EINSTEIN
Albert Einstein College of Medicine
center for epigenomics | Einstein
navigation
Main Page
Community portal
Current events
Help
quick links
Services
search
Go
Search
toolbox
What links here
Related changes
User contributions
Logs
Block user
E-mail this user
Upload file
Special pages
Printable version
Permanent link
Print as PDF
Browse properties
Sample Submission
Click the button below to submit a job to the Epigenomics and Genomics Shared Facilities:
Submit Samples
Projects
Lab Notebook
Lab Homepage
Links
When you submit jobs to the ESF you can organize them into projects. When the results are ready, they will be available here.
We will perform some post-processing analysis for immediate visualization of the data and its quality, and we also supply links to the raw data.
HELP-Tagging Expts
ChIP
This page was last modified on March 18, 2010, at 15:19.
This page has been accessed 281 times.
Privacy policy
About WikiLIMS
Disclaimers
Powered By MediaWiki


UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move
zoom in
zoom out
position/search
chr1:6,200,928-6,210,045
gene
jump
clear
size 9,118 bp.
configure
Aging Methylation Study
STS Markers on Genetic and Radiation Hybrid Maps
Restriction Enzymes from REBASE
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Non-Mouse RefSeq Genes
Ensembl Gene Predictions
Human Proteins Mapped by Chained tBLASTn
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced
CpG Islands (Islands < 300 Bases are Light Green)
30-Way Multiz Alignment & Conservation
Simple Nucleotide Polymorphisms (dbSNP build 128)
Repeating Elements by RepeatMasker


ESF ADMIN MENU
HOME
JOBS
USERS
BILLING
PILOT PROJECT
RESET PASSWORD
LOG OUT
EPIGENOMICS SHARED FACILITY
LABORATORY INFORMATION MANAGEMENT SYSTEM
Administrative Home Page
There are new JOBS awaiting processing
View New Jobs
RECENT SEQUENCE RUNS
SEQUENCE RUN NAME
FLOW CELL NAME PLATFORM
SEQUENCE RUN STARTED SEQUENCE RUN ENDED
SEQUENCE RUN'S FINAL STATUS
091113_HWI-EAS438_42A1YAAXX
42A1YAAXX ILLUMINA
11/13/2009 12/01/2009
Run Completed No lane problem
091117_HWI-EAS438_TEST
TEST ILLUMINA
11/13/2009 03/10/2010
Run Completed No lane problem
091125_HWI-EAS438_619V1AAXX
619V1AAXX ILLUMINA
04/27/2010 04/27/2010
Run Completed No lane problem
091117_HWI-EAS438_SECONDNEWROBTEST
SECONDNEWROBTEST ILLUMINA
04/24/2010 04/24/2010
Run Completed No lane problem

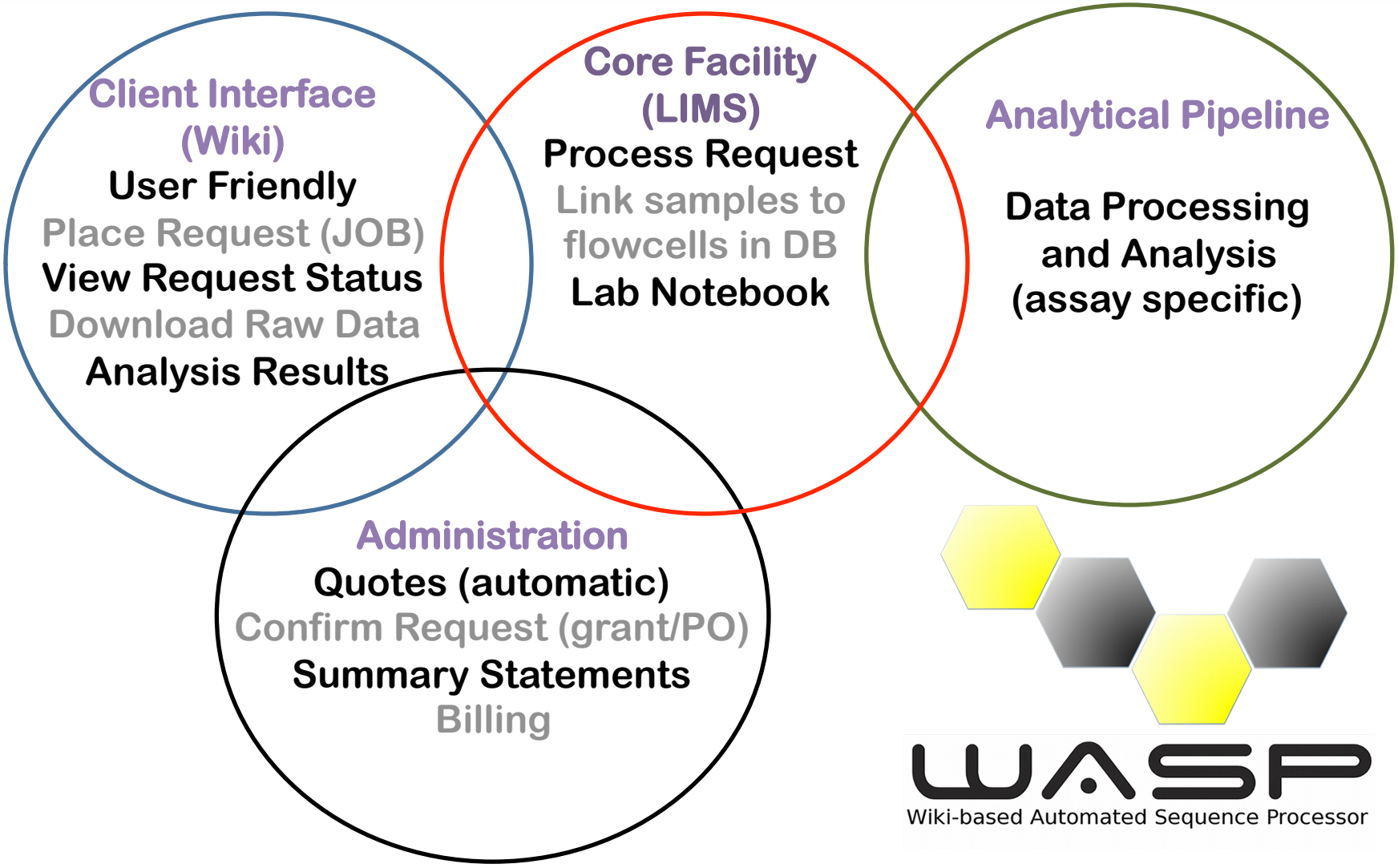
WASP: Component Interactions & Data Flow



center for epigenomics | Einstein

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Wiki-based Automated Sequence Processor

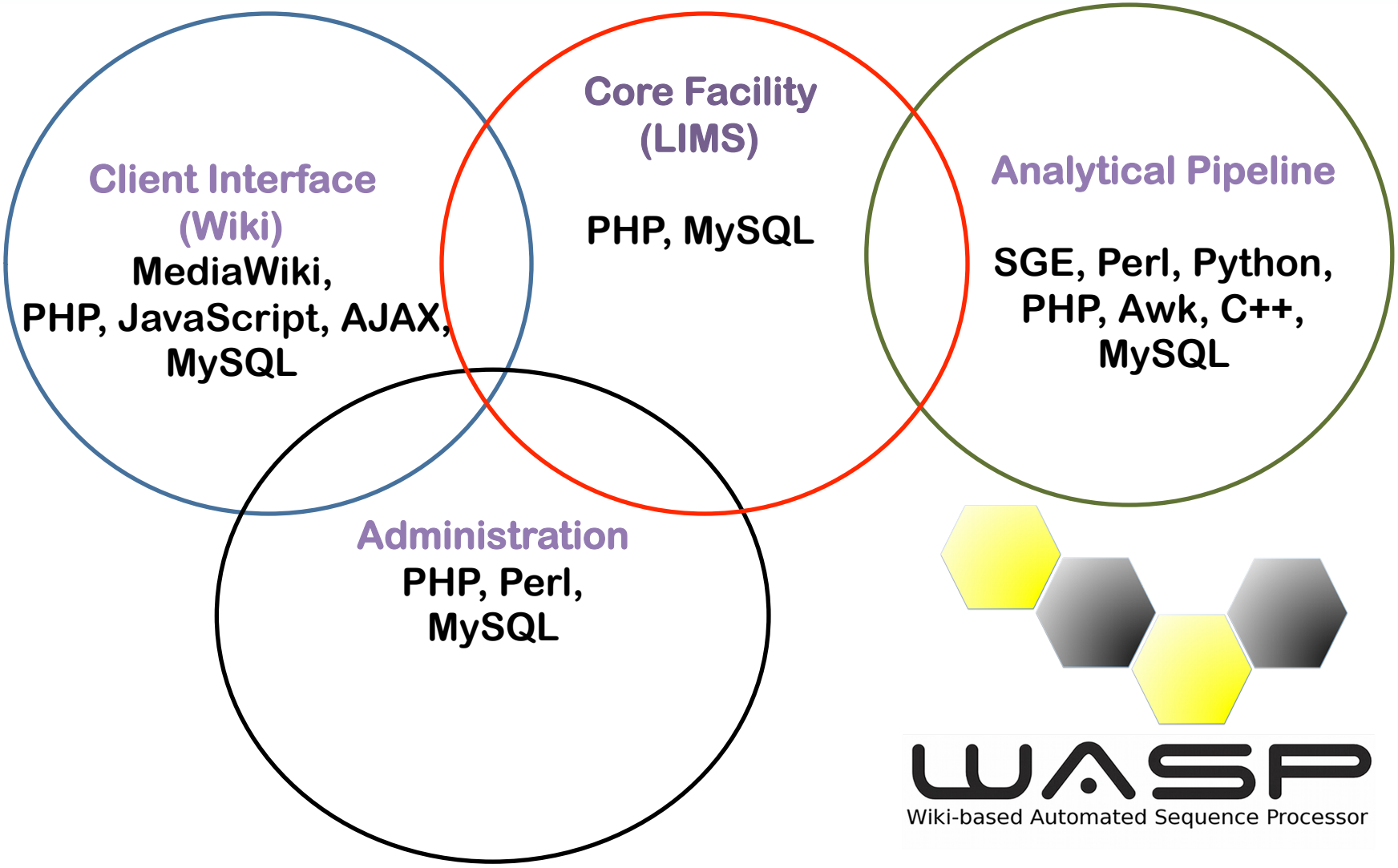


WASP
Wiki-based Automated Sequence Processor

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Basic Functionality

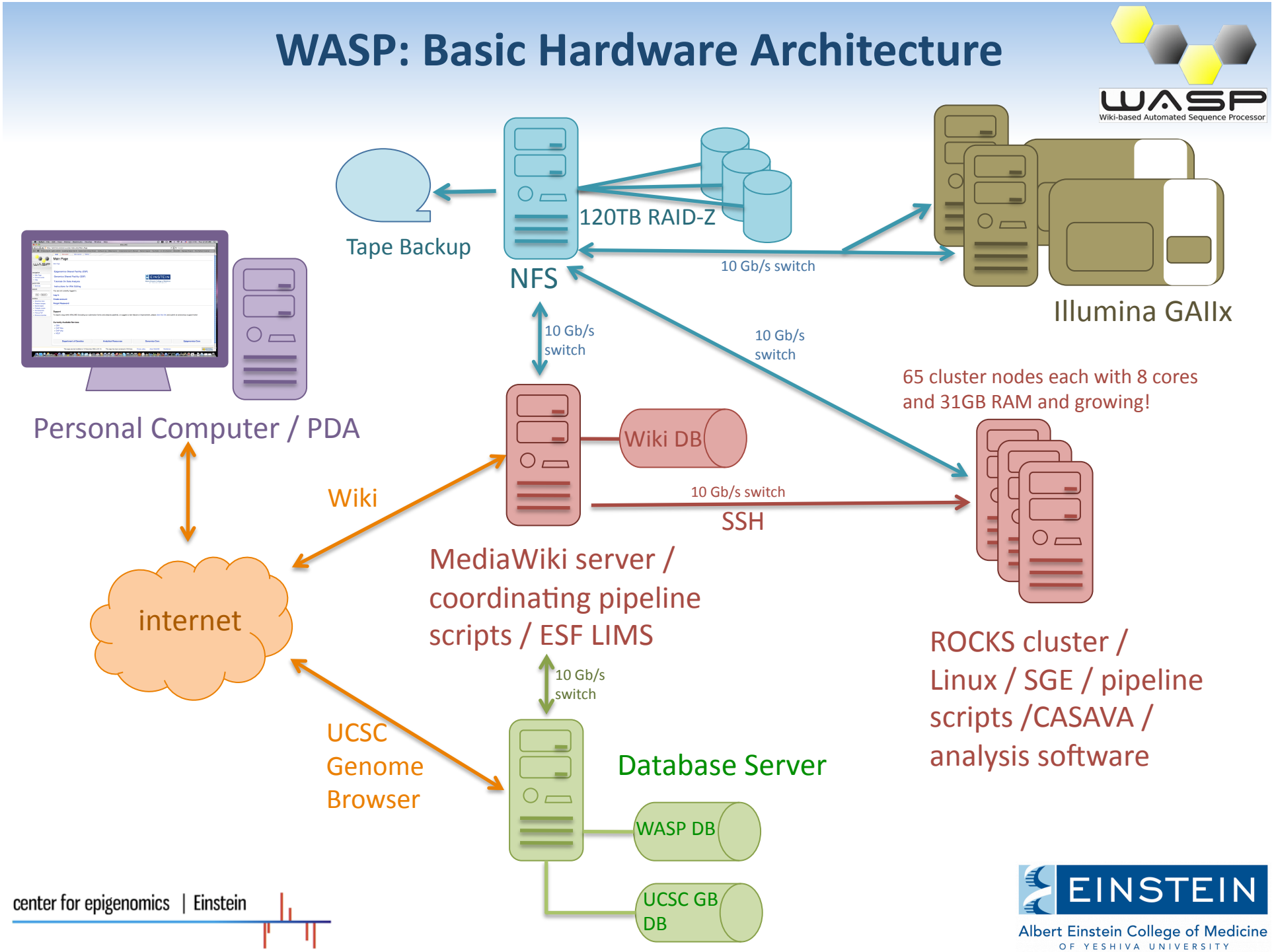


WASP
Wiki-based Automated Sequence Processor

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Basic Hardware Architecture



Server Room

Cluster

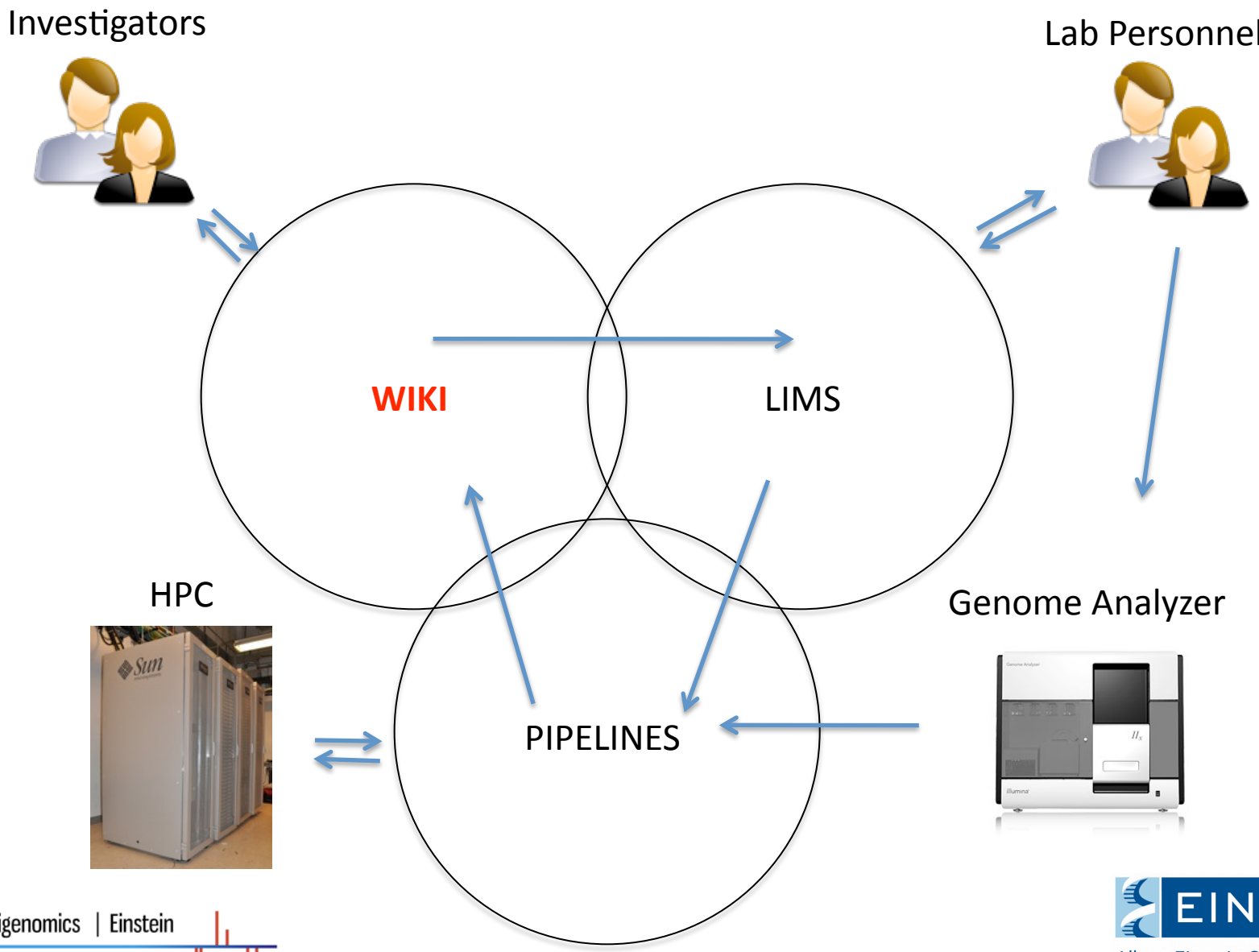
Linux Servers

RAID Storage

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

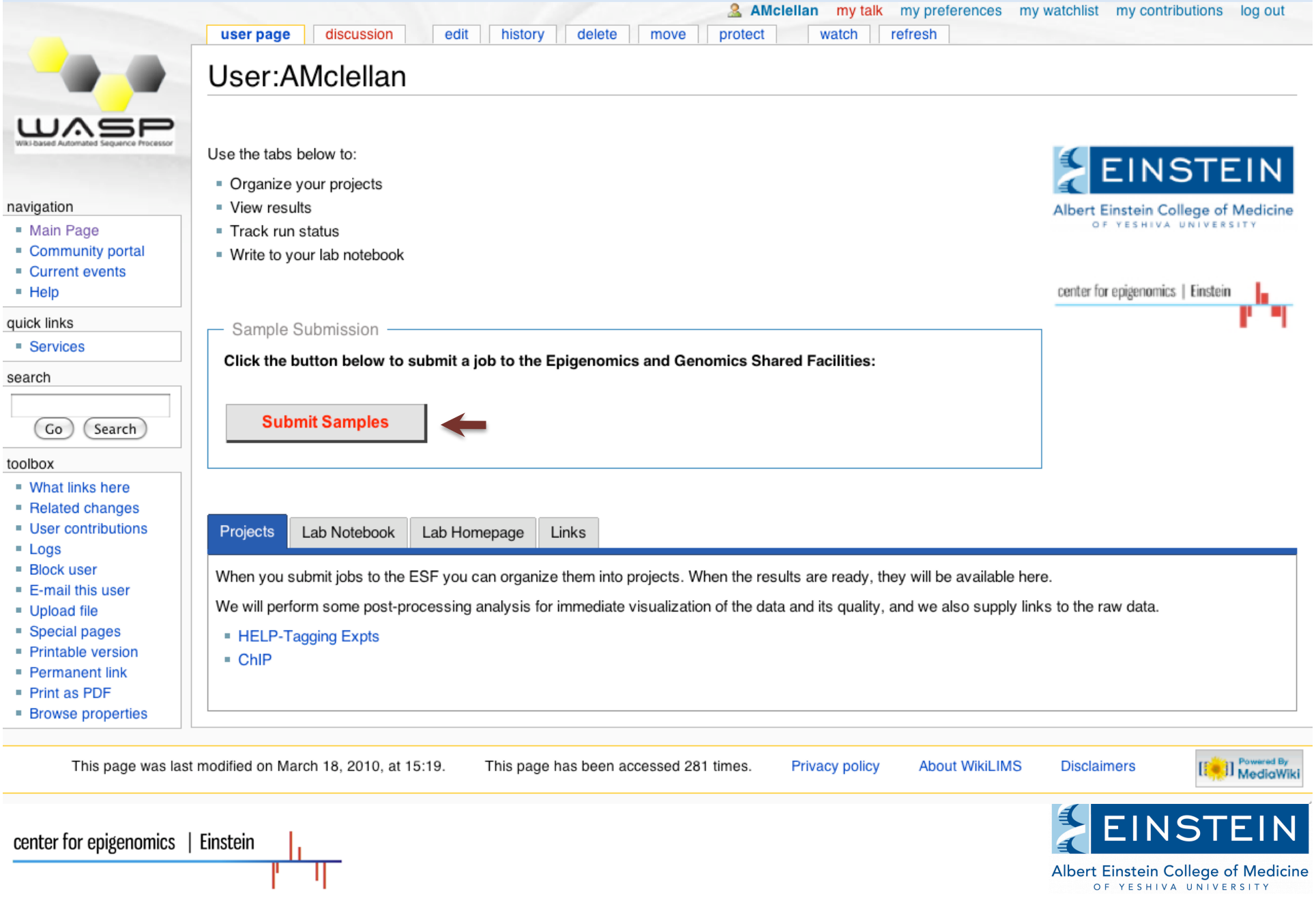
WASP: Component Interactions & Data Flow



center for epigenomics | Einstein

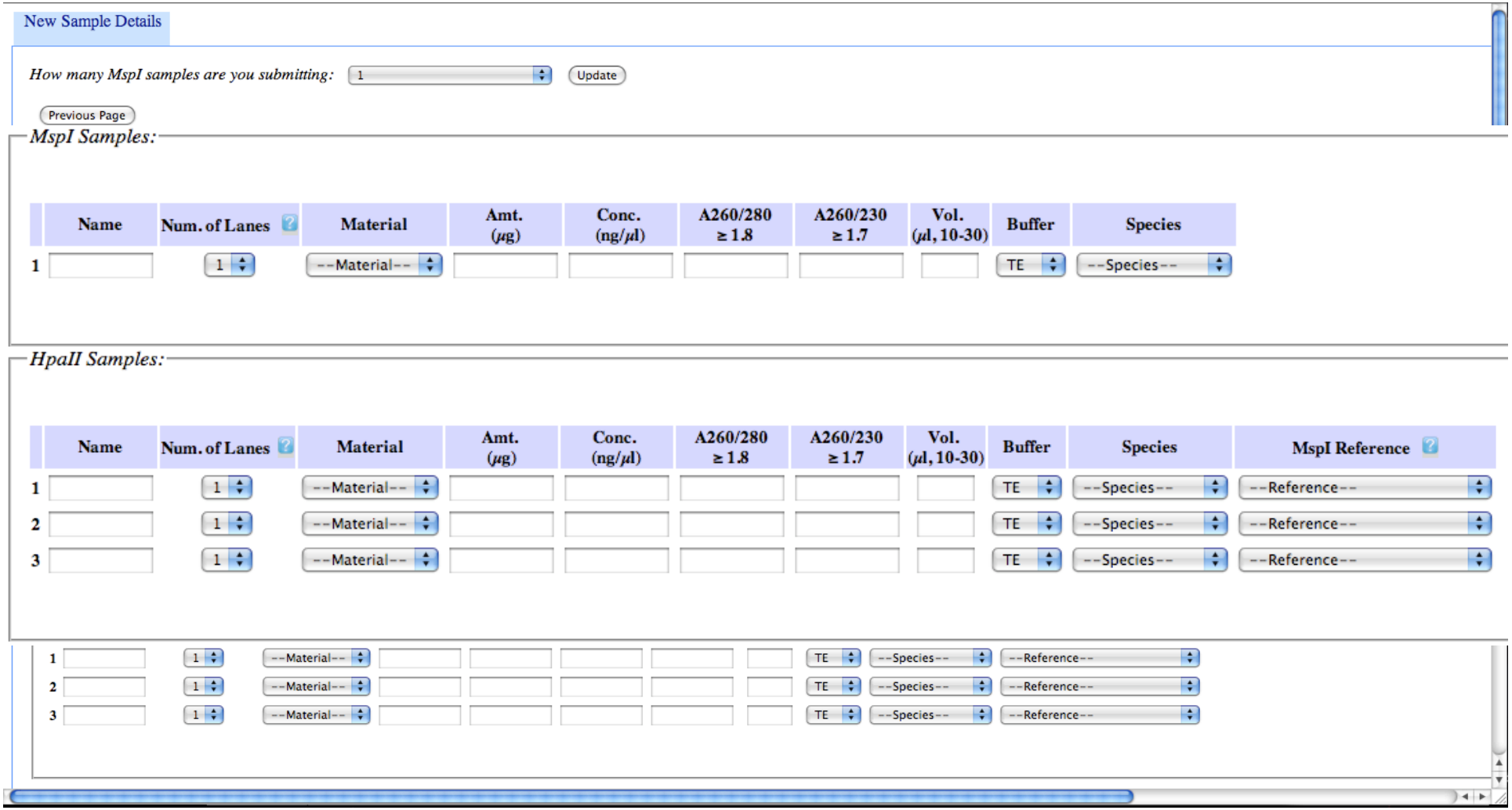
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: User personal Page



center for epigenomics | Einstein
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

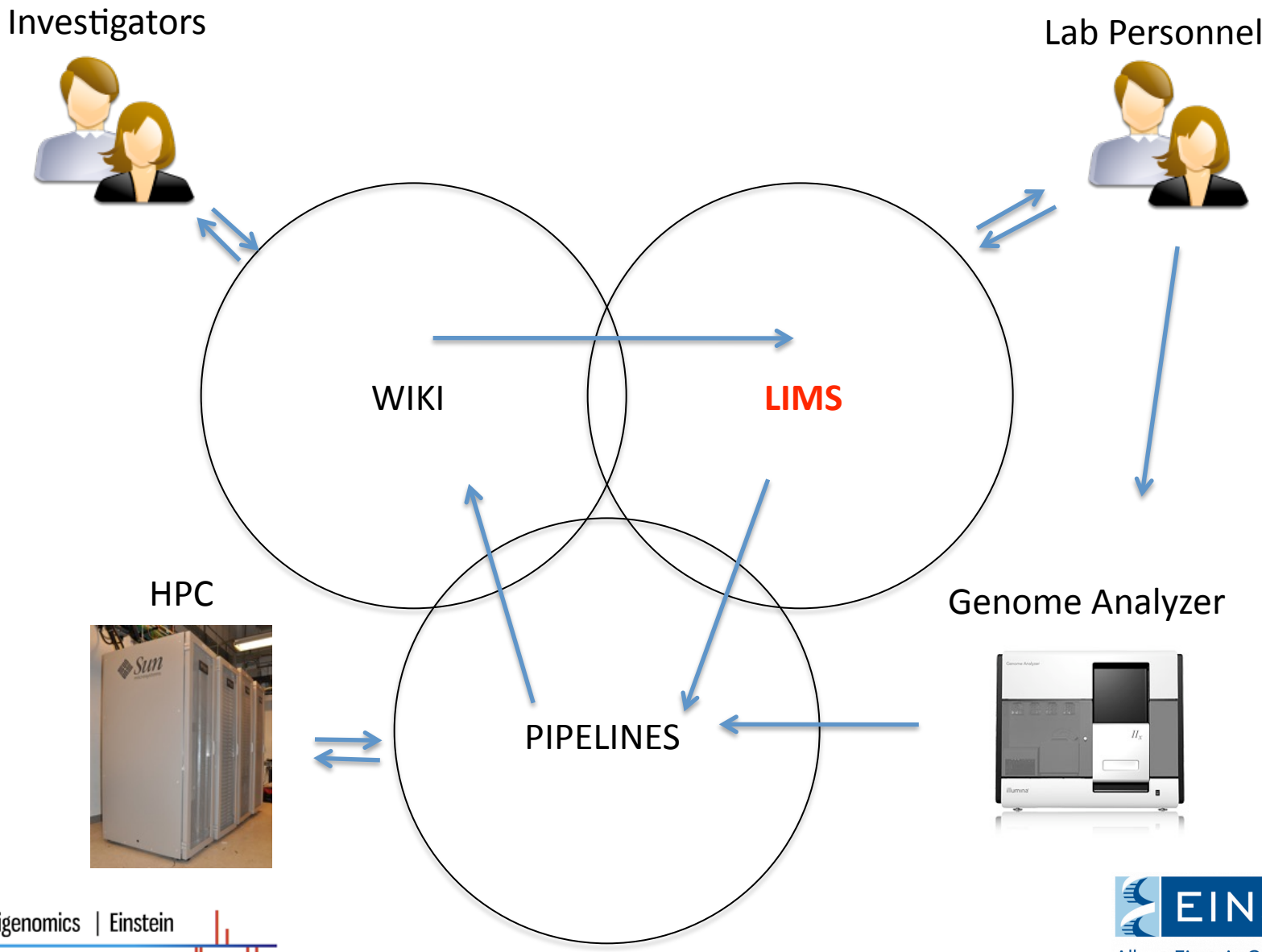
WASP: HELP-tagging Job Submission



center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

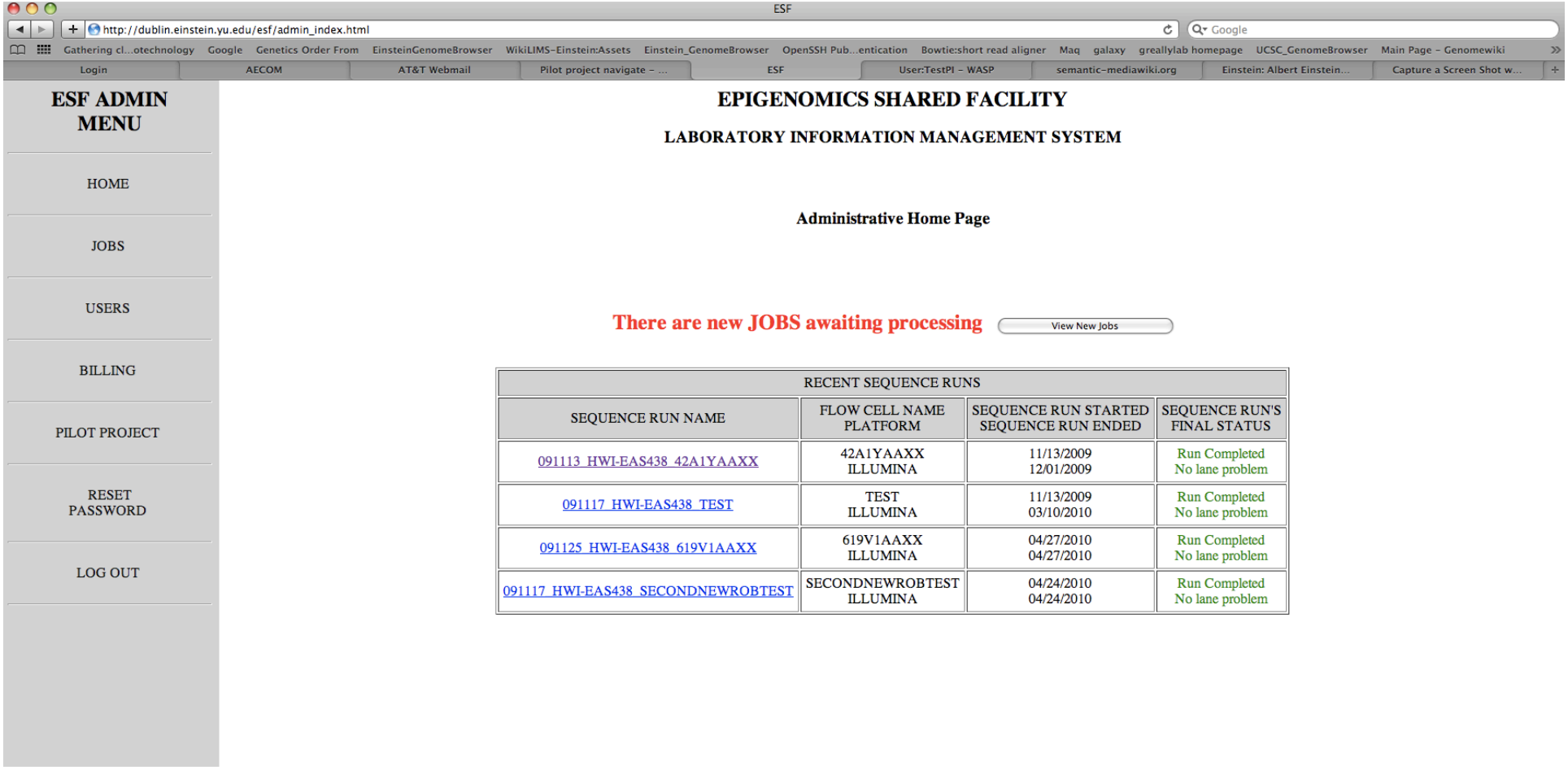
WASP: Component Interactions & Data Flow



center for epigenomics | Einstein

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

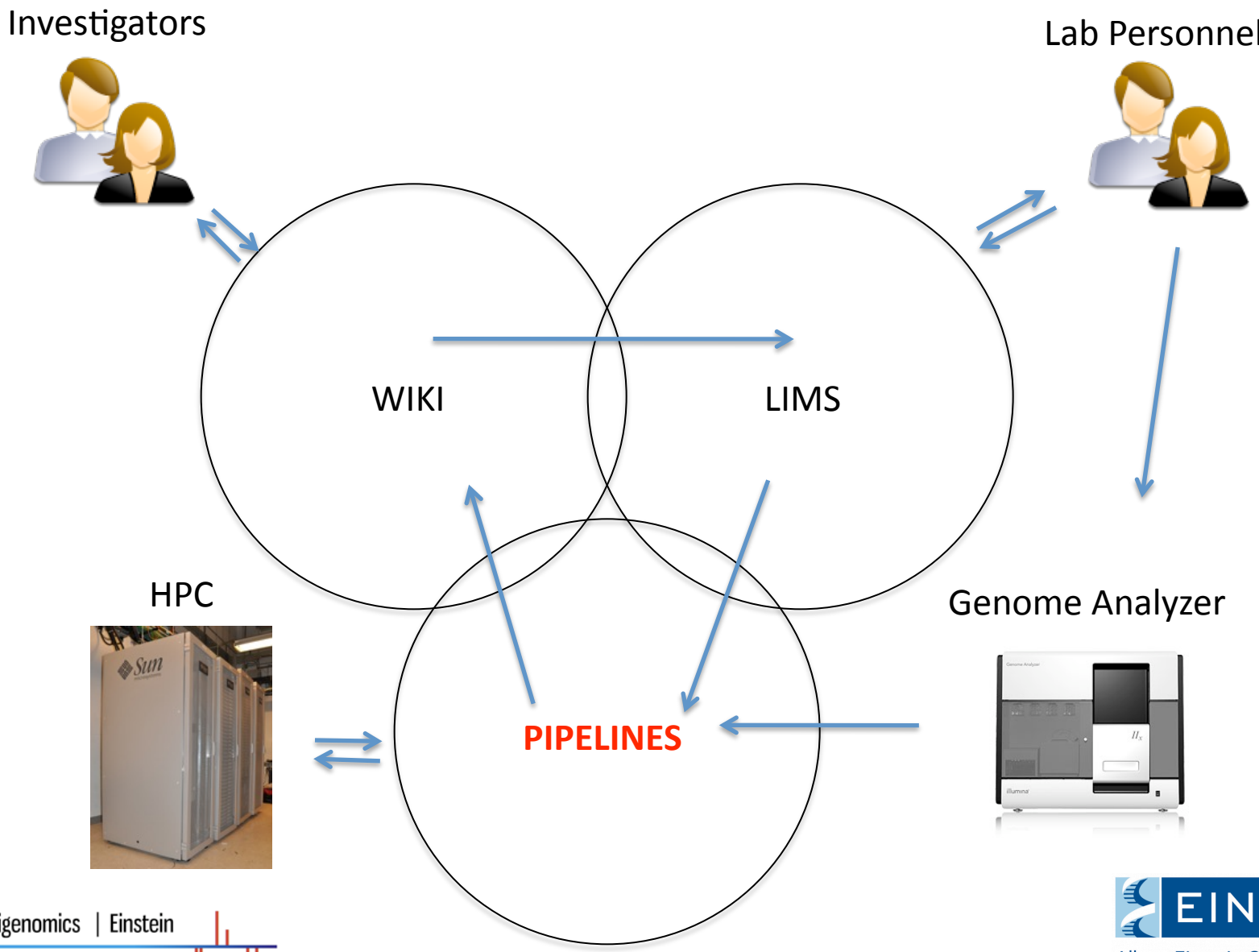
WASP: LIMS



center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

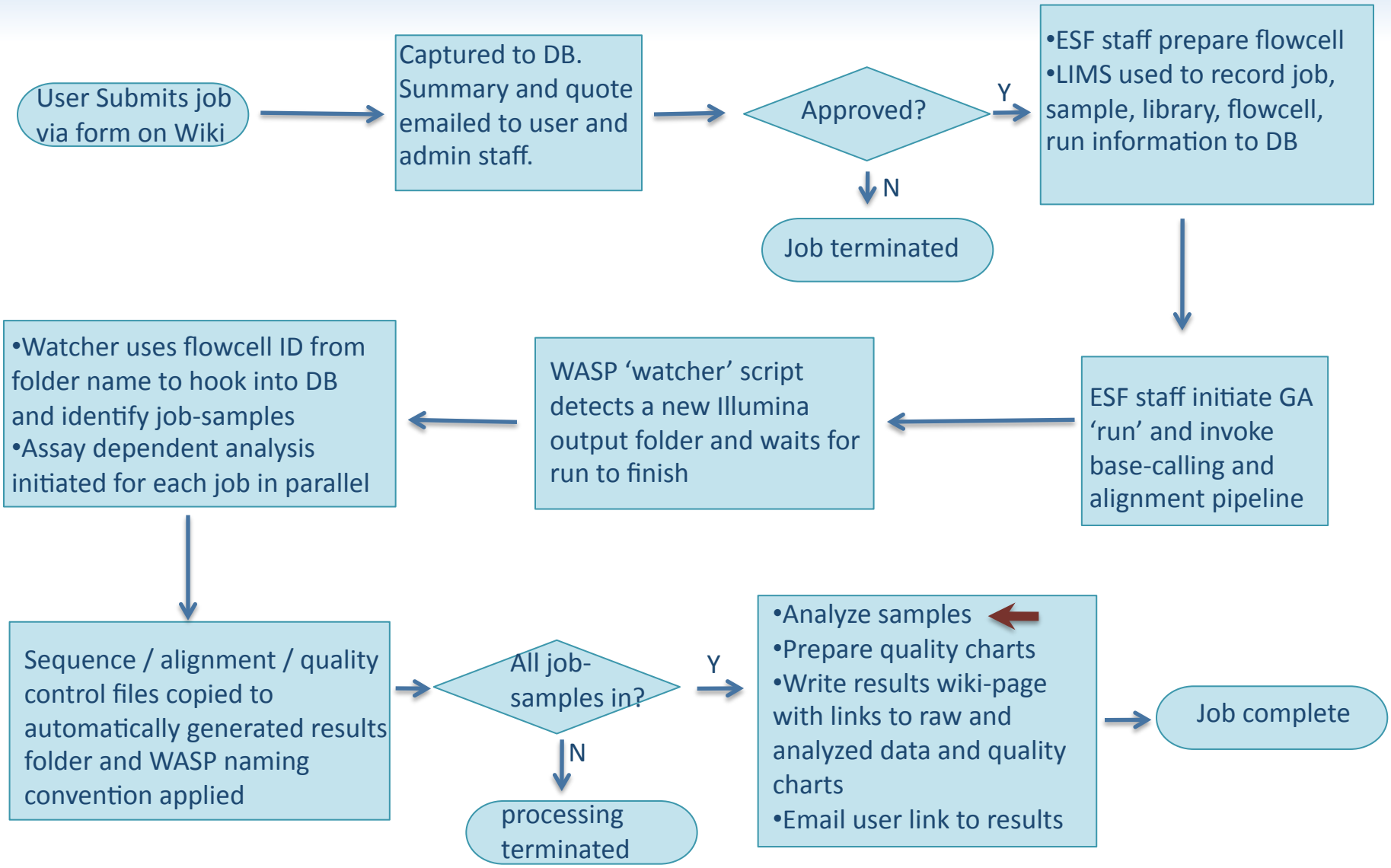
WASP: Component Interactions & Data Flow



center for epigenomics | Einstein

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Basic Workflow



center for epigenomics | Einstein

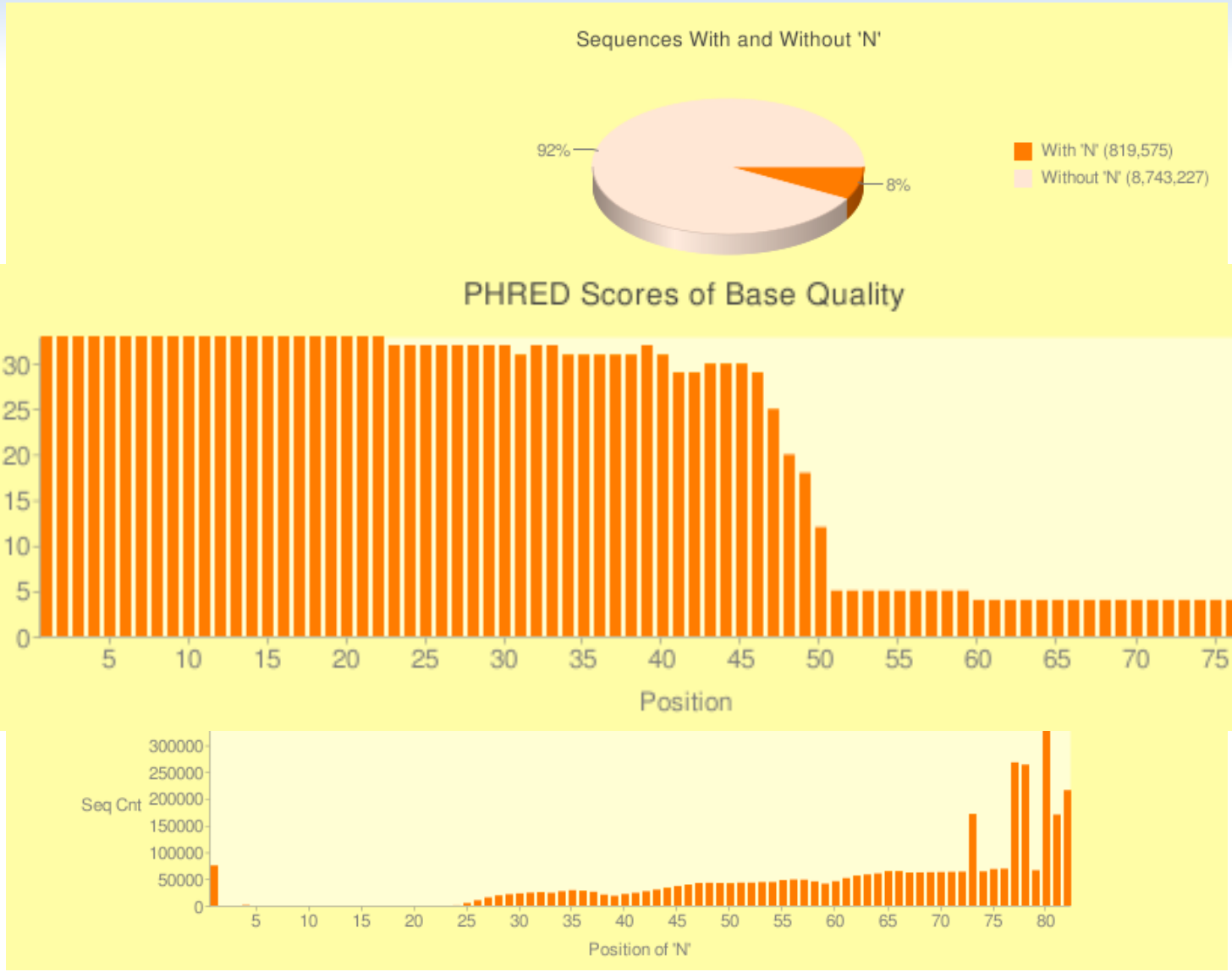
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Automated Writing of Results to Wiki

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: Sequence Run Quality



center for epigenomics | Einstein

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

WASP: HELP-Tagging Data Return

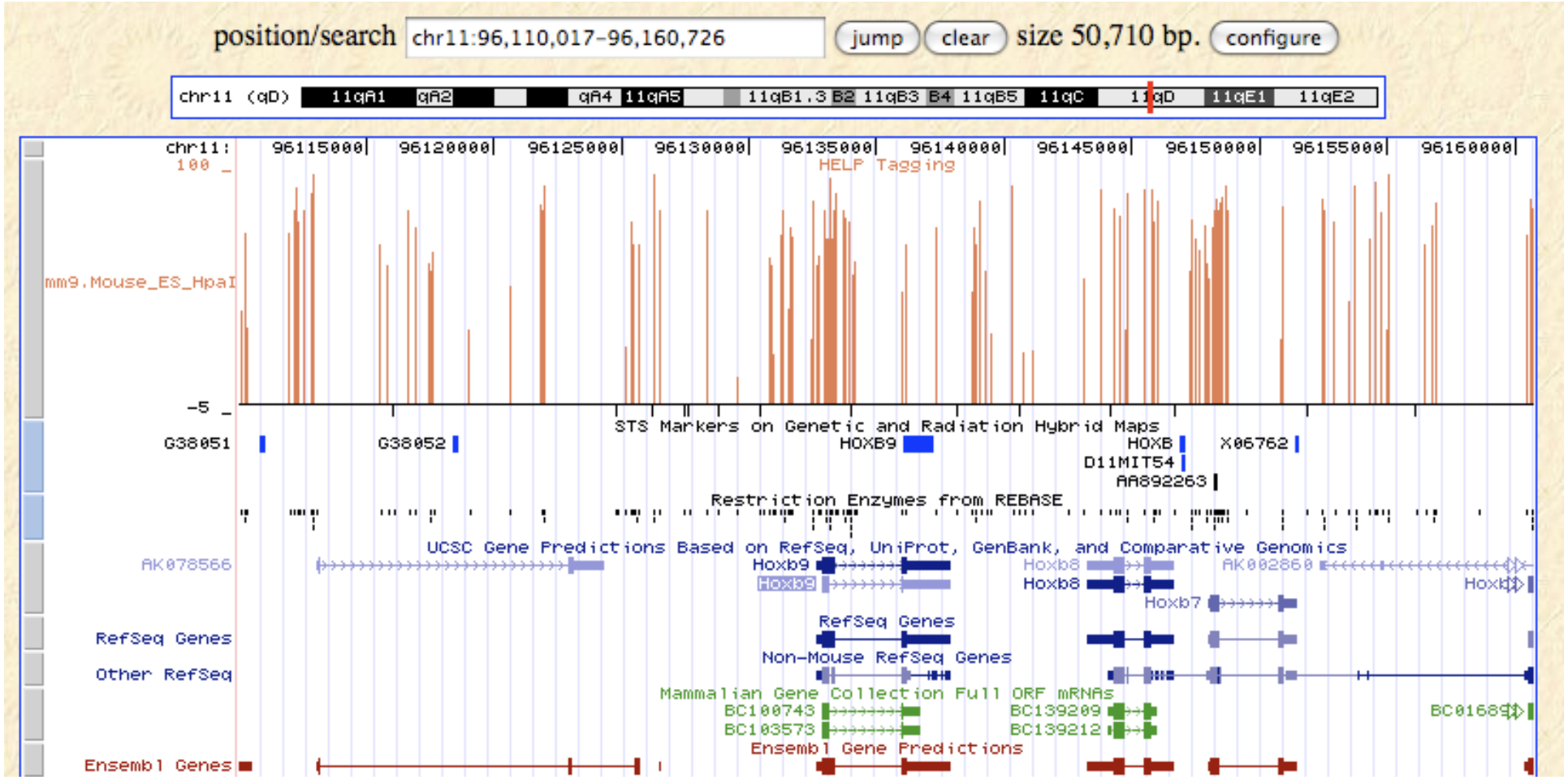
MD5 checksums for files (often requested when submitting raw data to repositories e.g.GEO)

Flowcell ID	Lane	Index	Sample Name	Pair	Raw Data Files	Alignment Result
61J65AAXX	1	0	03-H-SG	0	Download Fastq Sequence File Download ELAND-extended Alignment File	Download BAM File Download BAM index File Download SAM File Display reads in Genome Browser
	2	0	03-M-SG	0	Download Fastq Sequence File Download ELAND-extended Alignment File	Download BAM File Download BAM index File Download SAM File Display reads in Genome Browser
61J7HAAXX	1	0	Y3-H-SG	0	Download Fastq Sequence File Download ELAND-extended Alignment File	Download BAM File Download BAM index File Download SAM File Display reads in Genome Browser
	2	0	Y3-M-SG	0	Download Fastq Sequence File Download ELAND-extended Alignment File	Download BAM File Download BAM index File Download SAM File Display reads in Genome Browser

Methylation Status Analysis Results

Sample Name	Sample Type	Flowcell ID	Lane	Multi-Alignment Result	MspI/HpaII Comparison Result
Y3-H-SG	HpaII	61J7HAAXX	lane_1	Display HpaII BED Track in Genome Browser	Download Wiggle Track and HpaII Count File
Y3-M-SG	MspI	61J7HAAXX	lane_2	Display MspI BED Track in Genome Browser	Display Wiggle Track in Genome Browser
03-H-SG	HpaII	61J65AAXX	lane_1	Display HpaII BED Track in Genome Browser	Download Wiggle Track and HpaII Count File
03-M-SG	MspI	61J65AAXX	lane_2	Display MspI BED Track in Genome Browser	Display Wiggle Track in Genome Browser

WASP: HELP-Tagging Analysis



center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

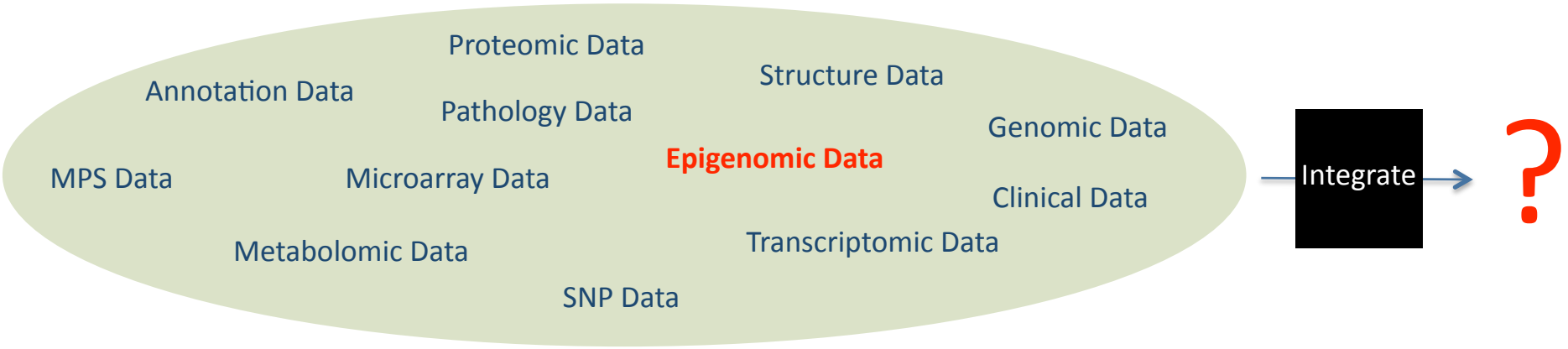
Parallelization in WASP

- Current:
  - up to 8 analysis instances initiated per run (8 lanes). Low resource shell execs.
  - each pipeline instance may submit many jobs to SGE to do heavy duty work
    * parallelize asynchronous tasks e.g. generate sample stats & perform analysis
    * parallelize sample analysis (individual / pairs)
- To do:
  - using MPI, parallelize counting tasks such as generating sequence statistics
  - identify or develop analysis tools optimized for cluster analysis
  - Get an HPC for dummies book and learn how to maximize our resources

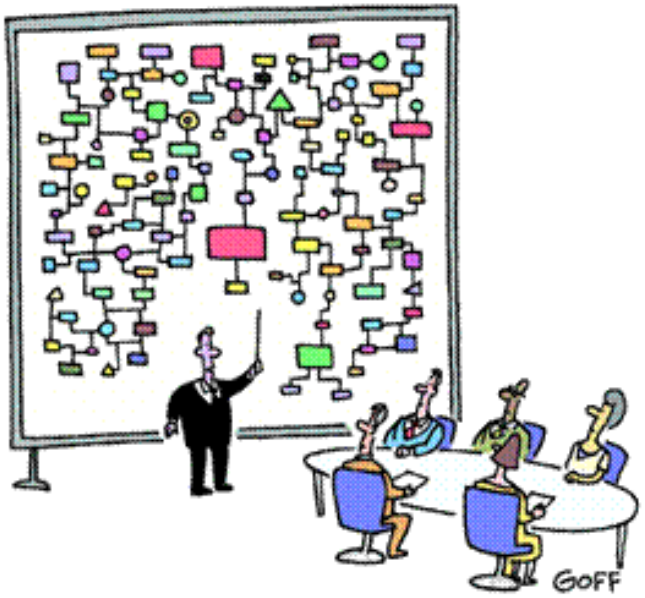
WASP: Future Plans

- Support more assays
  - e.g. Bisulphite-seq
- Improve efficiency
  - Optimize use of SGE for parallelizing tasks (in use but needs improvement)
  - Develop and utilize more software using MPI / MapReduce
- Improve metadata capture
  - Use ontology databases to enforce use of standard terms
- Allow custom secondary analyses
  - re-run an analysis with different parameters or a different software package

Data Integration Challenges



- Large search space to explore
- Powerful statistical / AI / pattern recognition algorithms – learn from other fields
- Understanding:
  - normal cell function
  - disease
  - potential drug targets



"And that's why we need a computer."

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Data Integration Challenges

- Data from many sources => heterogeneity and fragmentation
  - experiments performed on different samples under different conditions
  - many public data repositories (which ones to use?)
- Standards / conventions
  - use of ontology databases
  - look to major collaborative projects e.g. Encode / CaBig / GEO for emerging standards (e.g. BAM/SAM DNA alignment file format developed for the 1000 genomes project)
  - still waiting for microarray MIAME – like standards for MPS.
- Efficient data access
  - interface to data / HPC applications: API development / web services
  - transfer between sites. Local (ideal) or remote (cloud)
  - MapReduce. e.g Genome Analysis Toolkit (GATK)

Epigenetics Databases

- MethDB
  - 19,905 DNA methylation content data entries
- PubMeth
  - over 5,000 records on methylated genes in various cancer types
- REBASE
  - over 22,000 DNA methyltransferases genes
- MeInfoText
  - gene methylation information across 205 human cancer types
- MethPrimerDB
  - 259 primer sets from human, mouse and rat for DNA methylation analysis
- The Histone Database
  - over 2000 histone sequences
- ChromDB
  - 9,341 chromatin-associated proteins, including RNAi-associated proteins
- CREMOFAC
  - 1725 redundant and 720 non-redundant chromatin-remodeling factor sequences

Epigenetic Mapping Projects

- Encyclopedia of DNA elements (ENCODE)
  - aims to map all functional elements of human genome – including epigenetic modifications

- Human Epigenome Project
  - map methylation of 43 unrelated individuals

- Alliance for Human Epigenomics and Disease (AHEAD)
  - epigenome mapping project

- High-throughput Epigenetic Regulatory Organization in Chromatin (HEROIC)

HPC Software for MPS Applications

- Not a lot! … Yet!
- CloudBurst
  - short-read mapping to reference genome
  - single base mutation detection (SNP)
  - MapReduce / Hadoop
- Crossbow
  - short-read mapping to reference genome
  - single base mutation detection (SNP)
  - MapReduce / Hadoop
- Contrail
  - genome assembly
  - MapReduce / Hadoop
- GATK
  - structured programming framework for MPS analysis tool creation
  - MapReduce

Summary and a Look to The Future

- Development of new epigenomic assays for EWAS critically dependent on analytical capacity (HPC) and efficient data management systems e.g WASP

- Need to standardize data formats and interfaces to data
  - generally lacking for MPS data

- Consolidate / synchronize repositories

- Centralized data vs distributed data debate. Privacy / ownership /ethics.

- Moving data about will become a bigger problem in the future:
  - growth in data size exceeds Moore's Law.
  - how to move data from site to site if centralized or in the cloud?

- Need for more analysis software optimized for HPC / cloud to meet future challenges.

Acknowledgements

Einstein

Computational Epigenomics
Dr. Robert Dubin
Qiang Jing (A.J.)
Pilib Ó Broin

Computational Genomics
Brent Calder
David Moskowitz

Biostatistics
Dr. Melissa J. Fazzari

Epigenomics Shared Facility
Dr. John Greally (Faculty Advisor)
Dr. Shahina Maqbool (Director)
Raul Olea
Gael Westby

Greally Lab
Dr. John Greally
Dr. Masako Suzuki
Dr. Niki Athanasiadou
Dr. Niru Narayanan
Edyta Stasiek
Marién Pascual
Maria-Paz Ramos
Esther Berko
Andrew Ramnauth
Kevin Lau

Former lab personnel
Dr. Mayumi Oda
Dr. Khulan Batbayar
Dr. Jacob Glass
Dr. Priti Tewari
Dr. Reid Thompson

center for epigenomics | Einstein

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY