Bringing the Wasp System genomics cyberegocsystem to the grid
Acknowledgements

Andy McLellan
Brent Calder
Rob Dubin
AJ Jing
Natalia Volnova

Aaron Golden

Joseph Hargitai

@waspsystem
Introduction

Today’s challenge

Disruptive genomics technologies

A cyberecosystem solution

Wasp System

Remote resources

Cloud and grid

New challenges

Implementation and adoption

Aspirations

Nurturing the cyberecosystem
Introduction

Today’s challenge
Disruptive genomics technologies

A cyberecosystem solution
Wasp System

Remote resources
Cloud and grid

New challenges
Implementation and adoption

Aspirations
Nurturing the cyberecosystem

1990
Medical degree
Paediatrics residency

2000
PhD degree
K08 award: genomic imprinting
Clinical Genetics Fellowship
Faculty appointment
DNA methylation microarrays
Epigenetics research

2010
Massively-parallel sequencing
Center for Epigenomics
New York Genome Center

Monday, July 29, 13
Today’s challenge

The human genome

3 x $10^9$ base pairs of DNA sequence

Genes copied to RNA, communicate genomic information, template for protein production

http://genome.ucsc.edu/

PMID: 19339662
Today’s challenge

The human genome

$3 \times 10^9$ base pairs of DNA sequence

Packaged for function as chromatin

Heterochromatin and euchromatin now characterised in terms of molecular events
Today’s challenge

The human genome

3 x 10^9 base pairs of DNA sequence

Most mediators of molecular organisation in nucleus can now be defined using DNA sequencing technologies
Today’s challenge

Disruptive genomics technologies

DNA sequencing (1977-1985)

<10 samples per run
<200 bp per sample

2 x 10^3 bp per experiment

http://www.the-scientist.com
Today’s challenge

Disruptive genomics technologies

DNA sequencing (1985-2005)

384 samples per run
300-1,000 bp per sample

$4 \times 10^5$ bp per experiment

Human Genome Project technology

http://www.mun.ca
Today’s challenge

Disruptive genomics technologies

DNA sequencing (today)

16 lanes
~160 x 10^6 sequence clusters per lane
~200 bp per sample

~5 x 10^{11} bp per experiment

Illumina Sequencing Technology

Human Genome Project technology

http://www.illumina.com
Today’s challenge

Cost of sequencing

Flattening out after years of steep declines

Human genome ~$3,600 (NYGC)

Decreased costs leading to explosive growth of use of sequencing

http://www.genome.gov/sequencingcosts/
Today’s challenge

Diversity of assays

Sequencing now being used to test every aspect of genomic physiology

Cancer genomics a good example of information heterogeneity and complexity
Today’s challenge

Diversity of information

Human genome project
  Generated a static reference genome

Many more users now than ever before, many places at which sequencing performed

Increased use of DNA sequencing technology
  Diverse assays
  Multiple sequencing platforms
  Multiple analytical approaches
  High volume data sets
  Highly complex data landscape

Geneticists, physicians, other molecular biologists mostly lack programming skills
  Rely on colleagues for insights
Today’s challenge

Diversity of information

Human genome project
  Generated a static reference genome

Many more users now than ever before, many places at which sequencing performed

Increased use of DNA sequencing technology
  Diverse assays
  Multiple sequencing platforms
  Multiple analytical approaches
  High volume data sets

  Highly complex data landscape

Geneticists, physicians, other molecular biologists mostly lack programming skills
  Rely on colleagues for insights
Today’s challenge

Extreme foetal growth

U-shaped risk for disease decades later in life

Francine Einstein
NIH (NICHD) R01 HD063791
Today’s challenge

Extreme foetal growth

Studying epigenetic dysregulation in stem cells

DNA methylation assay at $2 \times 10^6$ loci throughout human genome

DNA methylation is an epigenetic regulator of gene expression, helps to switch genes on/off
Today’s challenge

Addressing the informatics challenge

Reducing the size of the dataset
3 whole genome sequences returned from Complete Genomics on a 3 TB hard drive

Providing something that the investigator can use

Capturing metadata systematically
The Wasp System

An institutional Center for Epigenomics resource

Automated workflows, metadata capture, visualisation tools

Running 4 years at Einstein
Serving ~350 registered users on 4 continents

Most users unaware of back end architecture, analysis abstracted
Today’s challenge

The challenges are now downstream

Making sense of data requires integration of multiple datasets

Data need to sit somewhere where adequate computational resources exist
Today’s challenge

Sequencing is dispersing

Sequencing had been restricted to institutional, center or core facility services

Now personal, lower cost machines being installed in individual labs

Multiple sequencing technologies

Illumina market leader, but also Life Technologies, Pacific Biosciences, Roche 454, others, each using distinct technologies

Sequencing becoming a commodity

Outsourcing increasingly an option

Data analytical approaches diverse

>300 applications for genomics analyses at Einstein
A cyberecosystem solution
Our guiding principles

One person, group or institution should not attempt to fix the entire problem

Amass the collective insights of the genomics community

Assume the open source and open development paradigm is the best approach

Create an environment rewarding data interoperability and innovation

Use enterprise-level programming

Make contributions easy by having system comprised of linked components
A cyberecosystem solution

Wasp System components

Valuable for end user, for administrator and as middleware

Addresses laboratory information management system (LIMS) needs while also hosting workflows and capturing metadata

Such LIMS interaction generally ignored

Provides middleware for remote grid/cloud computing resources
A cyberecosystem solution

Wasp System design

Uses Spring framework for Java
Open source, enterprise level
Nurtured open source model

Our goal is to facilitate community development of this cyberecosystem

All components plug-ins, swappable

Plug-ins being developed hosted at Einstein Github resource

Developer retains copyright of plug-in, but encouraged to use open source license

When plug-in complete, sent to Wasp team for formal testing. Advice on debugging provided (JIRA resource) as long as developer agrees to make plug-in universally available to Wasp user community

Plug-in manager to allow choices to be made when downloading or updating Wasp
A cyberecosystem solution

Active open development

If it's not used, it's not valuable

Twitter account @waspsystem

Administrator Olga Vitkovsky marketing software

Translation of interface to Russian, French

Partnership with SEQanswers web resource
  http://seqanswers.com/

@waspsystem

PMID: 22419780

Monday, July 29, 13
A cyberecosystem solution

**Wasp collaboration with Taverna**

http://www.taverna.org.uk

Expertise with workflow management

Will integrate systems and develop web-based, drag and drop workflow development

Goal to make this easy enough that it can be used by any biologist or physician exploring their data
Remote resources

Wasp System as middleware

As sequencing moves beyond central cores into individual labs, an increasing need for data management and analytical solutions

Essential to address needs from individual to national scales

Scalability necessary, but also ability to perform analyses without local HPC resources

*Alex Maslov*: performs technology development using a PGM sequencer (Life Technologies)

*NYGC*: member institutions awarded 5.4% of NIH budget in 2012
Remote resources

Wasp System as middleware

See talk by David Rhee at 11.30
‘Spring through the Gateway’: Deploying Genomic Workflows with XSEDE
Marina Ballroom G
Remote resources

Value of grid over cloud resources

- Clearer security
- Easier tracking and auditing (HIPAA)
- Defined machines
- Identifiable experts

Suitable for highly repetitive production jobs

Success stories:
ALLPATHS assembly software (Broad/Pittsburgh)
Novoalign/BWA use of infiniband low latency network on Lonestar

Problems:
- Checkpoints/restarts not common in biology
- Need for more parallel tools
- Long compute runs needed
- Need for **reference cache** for integrative analyses
Remote resources

Reference cache

Allows your data to be put into context

Our data

Reference data

Event at gene regulatory region
Remote resources

Reference cache

Reference data too big to load on demand
e.g. the Cancer Genome Atlas (TCGA): 2.5 PB

Need to house, version and update on remote resources

Using remote processing makes sense
New need for scientific service providers

Centralised data, federalised opportunities
Great minds...

National Cancer Institute cloud

July 10 2013 announcement

Cancer Genome Atlas (TCGA) 2.5 PB data set driving the conversation

“This model ... involves the creation of a set of data repositories with co-located computational capacity and an Application Programming Interface (API) that provides security, data and resource access for developers of analytic tools.

In this model, applications are brought to the data, rather than bringing the data to the applications.

Such a "Cancer Genomics Cloud" has the capability to democratize access to NCI-generated genomic data and provide a more cost-effective way to provide computational support to the cancer research community.”
National Cancer Institute cloud

Similar model of remote reference data cache

“Key anticipated outcomes and benefits include the ability to:

1. Enhance access to community-generated cancer data and tools;
2. Take advantage of increases in computational efficiency and scalability;
3. Accelerate bioinformatics tool development;
4. Serve the individual researcher possessing limited or complete lack of high-performance computing access;
5. Better assess and optimize hardware need and costs for future production system(s).”
Remote resources

National Cancer Institute cloud

Estimated costs
• Design and implementation phase: $3-5 million per cloud pilot
• Evaluation period: about $500,000 per pilot
• Operational phase (if pilots are successful): $3-5 million per cloud.

Our model largely consistent
Diverges by facilitating interoperability from the data generating step, using a cyberecosystem

The same cyberecosystem provides a gateway to the remote resources for data exploration, with the ability to develop and plug in new analytical modules

A cyberecosystem distributes the tasks and the opportunities
New challenges

Biology diverges from grid traditions

Remote caches, long runs

Disruptive technologies

Punctuated evolution of DNA sequencing technologies

Data standards

Not developed, and likely to be frequently updated

Funding community, not researcher

‘Big science’ pushback
Aspirations

Widespread local cyberecosystem use

Wasp/Taverna as paradigms
Collaborative platform
Medium for standards imposition
Facilitates data interoperability
Data reduction prior to remote analysis

Continuous analysis system

Automatically updates with new data

Reproducible research platform

Capture analysis histories (Toby Bloom, NYGC)
Allows reviewer and reader to explore data

Facilitation of teaching

Embedding lessons into Wasp, remote deployment of sample algorithms
Data visualisation

Daniel Kohn (Broad)

Art:Science interactions
http://www.artscienceobservatory.org

Effective and intuitive data representations
Taking cue from radiology

Embed resulting algorithms into Wasp
Aspirations

Inspiration from astrophysics

Virtual observatory model

- Type of observation
  - Gamma ray
  - X-ray
  - Ultra violet
  - Visible
  - Infrared
  - Microwave
  - Radio

- Observed object
  - Sky

- Observation equipment
  - Space telescopes

- Observation flow
  - Primary data analysis
    - Raw data to processed information
  - Integrative data analysis
    - Contextualised information to generate insights

- Data management

- Metadata acquisition

- Intuitive visualisation interface

- Computing hardware resources
  - HPC
  - Grid

- Virtual Observatory

- Investigator
Aspirations

Inspiration from astrophysics

Virtual observatory model

Wasp as engine for complete cyberecosystem
Aspirations

Inspiration from astrophysics

- Adoption of virtual observatory tools
- Aladin example shown
- Multiple wavelengths
- Multiple genomic assays

http://aladin.u-strasbg.fr/
Aspirations

Inspiration from astrophysics

Adoption of virtual observatory tools

Aladin with Hilbert curve representation of genome, chromatin component data, self-organising map and genome browser representations
Who would have thought around 1900 that in fifty years time we would know so much more and understand so much less.


Questions