The Genome Center & The Open Science Grid

Gary Stiehr

THE Genome CENTER AT WASHINGTON UNIVERSITY

Washington University in St.Louis SCHOOL OF MEDICINE

The Genome Center

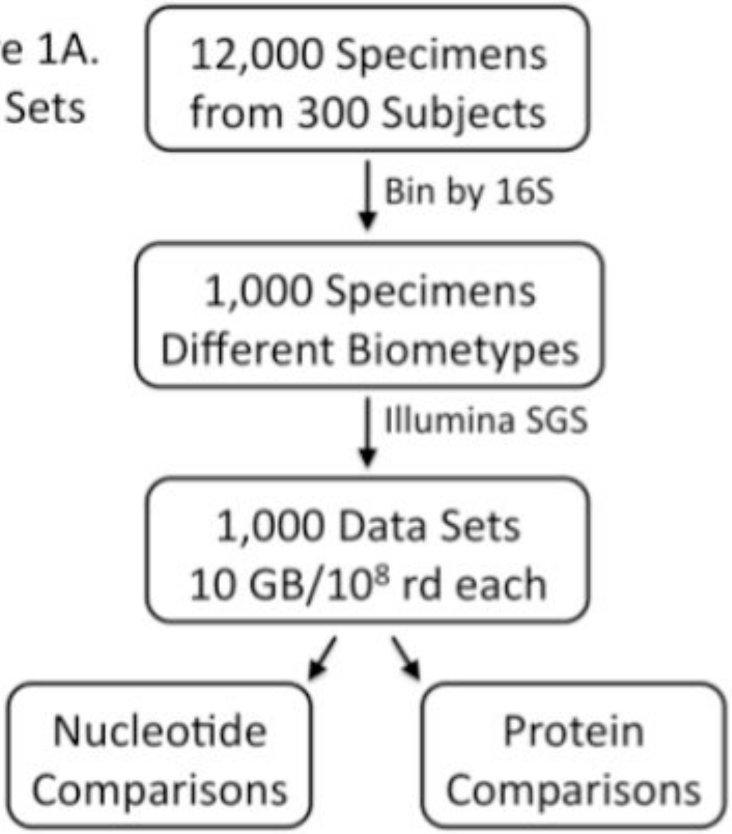
- The Genome Center is part of the School of Medicine at Washington University in St. Louis
- Contributed 25% of the finished sequence to The Human Genome Project
- Pioneers in analysis of cancer genomes; sequencing 600 pediatric cancer genomes in partnership with St. Jude Children’s Hospital
- Starting to sequence the Human Microbiome

Human Microbiome Project (HMP)

- Within the body of a healthy adult, microbial cells are estimated to outnumber human cells by a factor of ten to one.
- Broadly, the project has set the following goals:
  - Determining whether individuals share a core human microbiome
  - Understanding whether changes in the human microbiome can be correlated with changes in human health
  - Developing the new technological and bioinformatic tools needed to support these goals
  - Addressing the ethical, legal and social implications raised by human microbiome research.

HMP Analysis Overview

Figure 1A.
Data Sets



12,000 Specimens
from 300 Subjects

Bin by 16S

1,000 Specimens
Different Biometypes

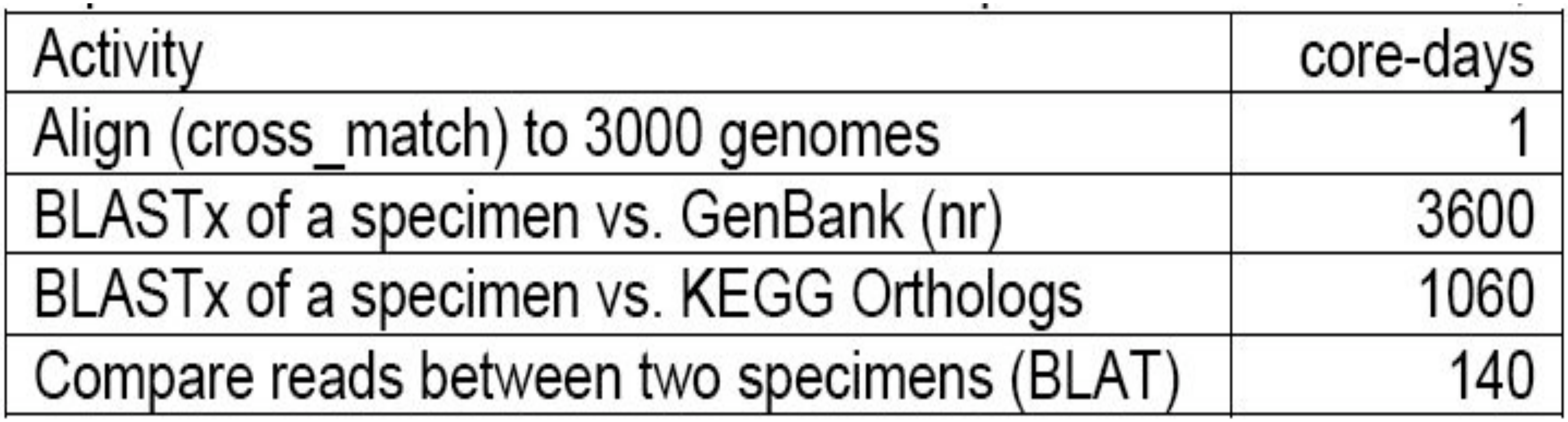
Illumina SGS

1,000 Data Sets
10 GB/10^8 rd each

Nucleotide
Comparisons

Protein
Comparisons

Some Anticipated HMP Runtimes



Activity	core-days
Align (cross_match) to 3000 genomes	1
BLASTx of a specimen vs. GenBank (nr)	3600
BLASTx of a specimen vs. KEGG Orthologs	1060
Compare reads between two specimens (BLAT)	140

- Above table assumes specimen data sets with 10^8 100-base reads
- At least 300 specimens need analysis at The Genome Center (will probably triple) within up to two years.
- BLASTx steps for 300 specimens may require 1,398,000 core-days.
- Over two years, need 1,915 cores to complete just one pass.
- It is anticipated that we will likely sequence and analyze around 1,000 specimens due to improving sequencing technology.

Speeding Up HMP Analysis

- Focus on protein comparisons (e.g. BLASTx) since they are currently the bottleneck.
- Alternative parameters to BLASTx
- Different algorithms to replace BLASTx
- Hardware acceleration (e.g., GPU) of BLASTx
- “Brute force” by using grid resources (e.g., OSG, Teragrid) or buying more cores.

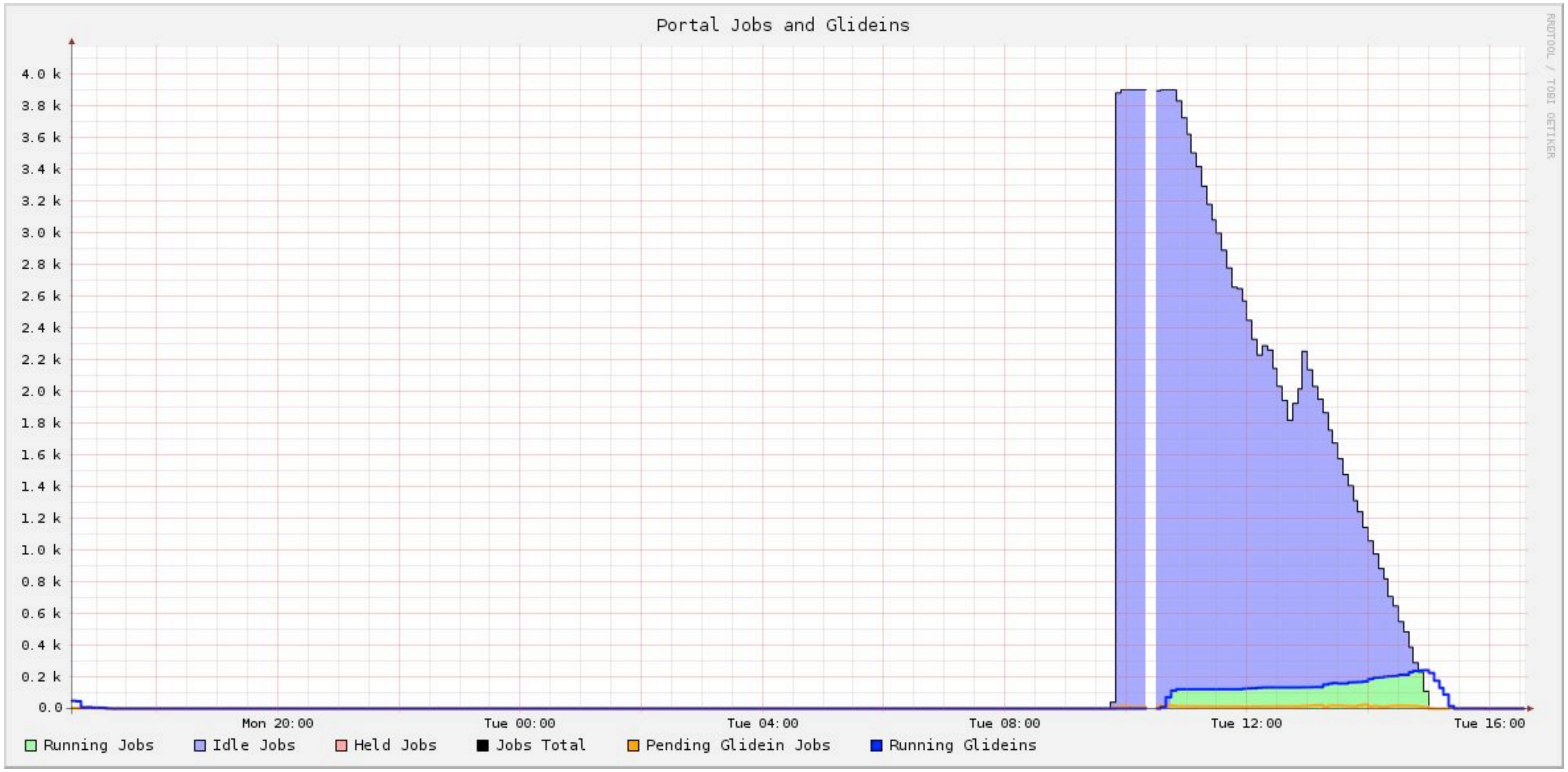
"Manual" OSG Job Submission

- We first setup a test CE with local resources and submitted jobs to it.
- We then submitted jobs to OSG as part of Engage VO.
- In our testing, we came up with these things to consider:
  - data movement overhead
  - remote hardware requirements of jobs
  - remote software requirements of jobs
  - availability of remote resources
  - reliability of remote resources
- Testing so far has been minimal and has not been tested to scale.

Running via RENCI Science Portal

- Used RENCI's BLASTMaster Science Desktop (Java-swing application).
- Some version of reference input data and BLASTx binary pre-staged at remote sites
- Used Glide-ins to better schedule jobs
- Automatically splits input queries into different jobs an distributes jobs to grid resources.

Running 5000 Reads/Jobs



garystiehr@wustl.edu

Plans for OSG

- Continue to learn about OSG job and data submission mechanism and issues.
- Work with RENCI to gain access to web services to run BLASTx via RENCI Science Portal (RSP) to automate job submission and potentially integrate into workflows.
- Determine how to pre-stage different versions of reference input data and BLASTx binaries to remote grid sites via RSP interfaces.
- Run a 100,000-read HMP data set via RSP
- Develop techniques to scale execution of 10,000,000-read HMP data set, including gathering of output.

Questions?

garystiehr@wustl.edu

THE Genome CENTER AT WASHINGTON UNIVERSITY