

Genomic Analysis on Hadoop

John J. Farrell

IT Manager

Biomedical Genetics

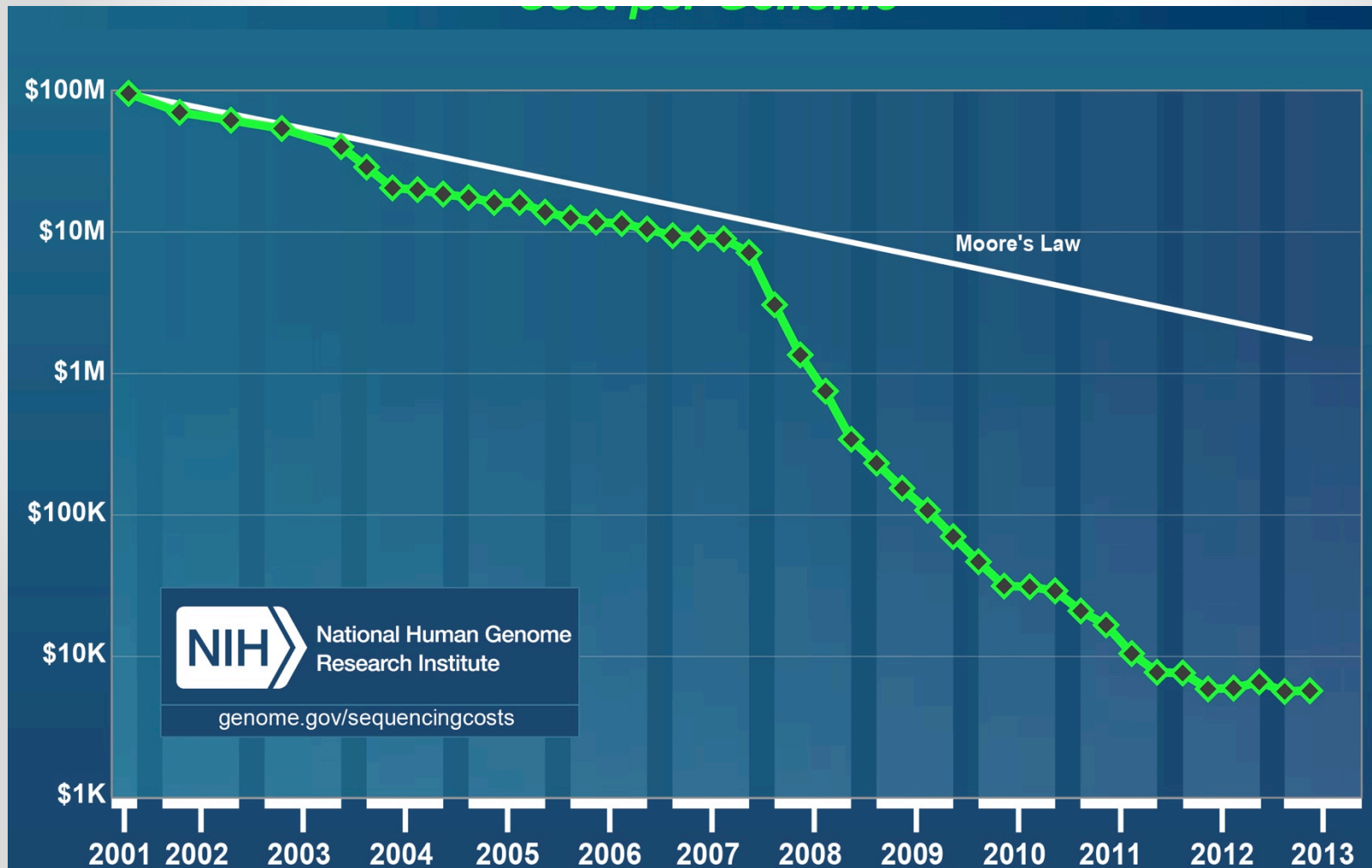


Genomics has a Big Data Problem

- Storage requirements are increasing rapidly due advances in genotyping and sequencing technologies.
- GWAS chips with 2.5+ million variants are available
- Whole exome sequencing(WES) is <\$1000 per individual
- Whole genome sequencing(WGS) is <\$7,000 per individual and rapidly approaching the milestone of \$1000 per individual. (Not much more than a MRI)
- Presently Illumina HiSeq 2500 generates 600GB (6 Billion PE Reads) in 11 Days.
- Often repeated claim is that **“90% of the data in the world today has been created in the last two years alone”**, IBM web site.

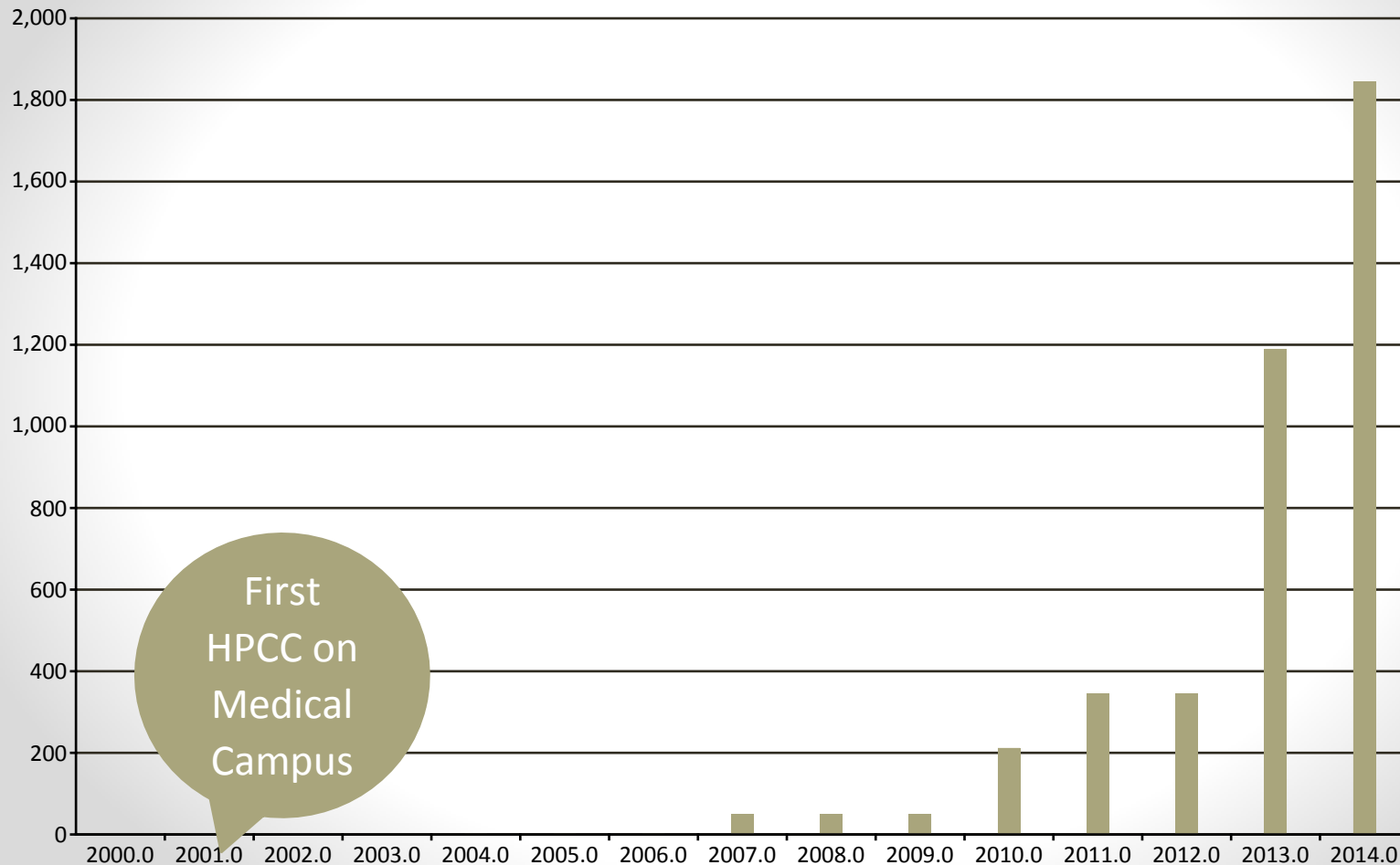


Sequencing Costs



BU HPCCC Parallel Storage

2 Proprietary Parallel Filesystem Used : IBM GPFS and HP IBRIX



Big Data in Genomic Research

- GWAS Data sets
 - Large consortiums are genotyping/imputing 37 million genotypes per subject for 10,000-50,000 individuals
- Next-Gen Sequencing
 - Whole Genome Sequencing
 - Coverage 30x will generate ~100 GB bam file per individual.
 - Whole Exome Sequencing
 - 40x Coverage will generate ~7 GB bam file per individual
 - RNA-Seq bam files on 1000 genome samples -Paired-End 75bp mRNA-seq with 50 million reads (3-5 GB) (Lappalainen, Sept. 2013, Nature Genetics)
- VCF (Variant Call File) will contain detailed information about the variant positions detected in a genome or exome. For an individual, there are typically 3-4 million variants. In recent release of 1000 genomes, there are 37 million variants (137 GB compressed) in 1092 individuals.



HPCC Scaling Issues

- HPCC systems like SCC and LinGA are designed for compute intensive applications but are not optimal for disk intensive tasks
- Jobs on HPCC compute nodes are not robust to hardware failures or software glitches on compute nodes.
- When submitting 1000s of jobs to test one model across millions of variants, it can become a major housekeeping headache to keep track of which jobs have failed due to node reboot or other hardware related failures.
- Parallel storage systems on HPCC clusters are proprietary and licensing and maintenance costs are expensive. IBM's GPFS is presently used on the BU Shared Computer Cluster (SCC). Large numbers of small files can create bottleneck due to metadata access (particular for backups/replication). Presently ~75% of the SCC GPFS storage is not replicated.
- Sorting large files (eg. BAM and VCF) is a relatively slow process.



Benefits of Apache Hadoop

- Scales well for processing Big Data- Hadoop Framework permits the analysis of large data sets across clusters of computers using very simple programming.
- Naively parallel jobs are easy to set up with Hadoop Streaming. Hadoop handles the splitting of jobs into 1000s of tasks rather than using leaving that to the researcher.
- Designed to work with commodity servers and be robust to hardware failures.
 - The HDFS storage has 3x replication. On drive failure, the data is simply replicated from the remaining two copies automatically. On node failure, tasks are rescheduled on remaining nodes.
- Data locality: Tasks are run where the data is stored rather than copying data across relatively slow network to the CPU. Eliminates data movement and maximizing throughput.
- Speed records for sorting TB size files are held on Hadoop Clusters
- No licensing fee-Active and well supported open source project (Yahoo, Facebook, Twitter, Hortonworks, Cloudera, NSA). There is extensive development and testing completed on large-scale Hadoop clusters by these companies before release.



Apache Hadoop 2.2.0 Features of Interest

- Released on October 15th, 2013. Cloudera and Hortonworks both have Virtual Machine Images of Hadoop 2.2.0 available for download to test out.
- YARN-Permits multi-tenancy and equitable sharing of Hadoop cluster resources.
- NFSv3 access to data in HDFS. Allows tight integration with HPC system.
- Snapshots now available for HDFS. Creating and deleting large files in a pipeline can be problematic on a snapshot file system. Often the snapshots will stealthily fill storage systems quota. *This implementation allows end-user to turned it on or off at the directory level.*
- Apache Pig 0.12: Easy to program language (Pig Latin) for Map-Reduce data flows.
- Apache Hive 0.12 (Highly Scalable Database with SQL support) has seen major speed improvement (50-100x) due to Stinger Initiative.



Hadoop-specific Data Mining and Analytic Toolkits

- Apache Mahout (mahout.apache.org) is a machine learning and data mining tool kit that scales to large data sets on the Hadoop Platform. Among the present algorithms that may be of interest for biomedical genetic research are:
 - K-Means Clustering
 - Random Forest
 - Logistic Regression
 - Naïve Bayes Classifier
- Revolution R Enterprise is a scalable/parallelized version of R for Hadoop (Free licensing for Academic Users).

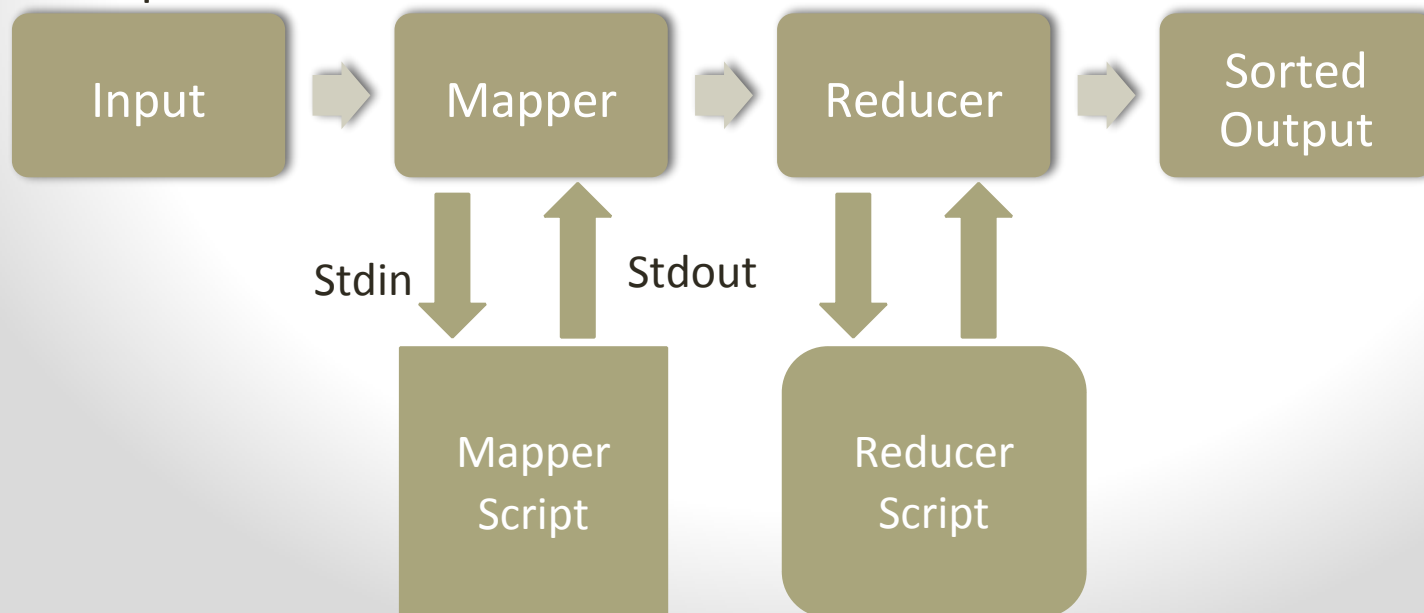
Bioinformatics Software designed for Hadoop

- Cloudburst: Highly sensitive read mapping with Map Reduce - Schatz, Bioinformatics, April 8, 2009
- BioPig: A Hadoop-based Analytic Toolkit for Large Scale Sequence data- Nordberg et al, Bioinformatics 9-10-2013
- Seal: A distributed short read mapping and duplicate removal tool, Pireddu et al, Bioinformatics March June 22, 2011
- Hadoop-Bam: directly manipulating next generation sequencing in the cloud, Neimenmaa et al, Bioinformatics, Feb 2 , 2012
- SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop, Schumacher et al, Bioinformatics, Oct 22, 2013
- BlueSNP: R Package for highly scalable genome-wide association studies using Hadoop Cluster, Huang et al, Bioinformatics, Nov 29, 2012
- Cloudbreak: Accurate and Scalable Genomic Structural Variation Detection in the Cloud with MapReduce, Whelan et al.



Hadoop Streaming

- Many Linux Bioinformatics software tools have been already been developed on HPC Linux Clusters. These tools are written in a variety of languages and scripts (R, Python, C, Perl, Java, etc).
- Hadoop Streaming provides a simple framework to parallelize those tools automatically.
- Hadoop Streaming feeds data into stdin of process and collects output from stdout.



2 Hadoop Streaming Examples

- Example Job 1 : Counting variants in a VCF file
 - Count the number of variants per chromosome in 1000 genomes Variant Call File (VCF)
 - VCF has detailed information for variants in 1090 individuals
 - More I/O intensive job
- Example Job 2 : GEE modeling for 4 phenotypes with genome variants
 - Run GEE models on 4600 individuals with 37 million imputed genotype dosages for 4 different phenotypes
 - Screen for Minor Allele Frequency >- 3% then run gee model:
Pheno~dose+gender+age+pc1+pc2+pc3
 - Compute intensive job

Example 1: Count variants in 1000 Genomes VCF file by chromosome

- The 1000 genomes VCF file contains 44+ million annotated variants for 1090 individuals (76 GB compressed/771 GB uncompressed). Uncompressed it will fill up your free project quota on SCC
- A simple linux command line to count the variants by chromosome (located in column 1 of the VCF file):
 - `cat 1000g.vcf | grep -v ^# | cut -f1 | uniq -c`
 - `grep -v ^#` skips comment lines in header
 - `cut -f1` cuts out column 1 (chromosome #) from vcf file
 - `uniq -c` counts by unique chromosome #

Example 1 on SCC

```
Time cat 1000G.integerated.phase1.vcf|grep -v ^#|cut -f1|uniq -c
```

```
3007194 1          1130553 15
3307588 2          1210617 16
2763451 3          1046732 17
2736764 4          1088820 18
2530215 5           816113 19
2424424 6           855166 20
2215228 7           518965 21
2183838 8           494328 22
1652386 9          1487477 X
1882662 10         2834 MT
1894908 11
1828006 12
1372998 13
1258252 14
```

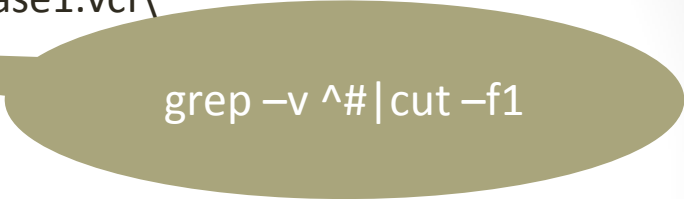
- Uggh-- The command completed in 52 minutes on the SCC



Example 1 on Hadoop

To parallelize on the Hadoop cluster, the following hadoop command is run:

```
hadoop jar $HADOOP_STREAMING \  
-input /user/farrell/vcf/1000G.integerated.phase1.vcf\  
-file `pwd`/map.cmd \  
-output /user/farrell/results/chr.count \  
-mapper 'map.cmd'\  
-reducer 'uniq -c'\  
-numReduceTasks 1  
hadoop fs -cat results/chr.count/p*
```



grep -v ^# | cut -f1

- Using Hadoop streaming, the count completes in 9 minutes on 8-node hadoop cluster.
- **Any program that reads data from stdin can be parallelized in this manner.**
- Using bwa-mem completed alignment of 40x whole genome in <3 hours (versus 48 hours).

Running bwa on Hadoop

```
hadoop jar $HADOOP_STREAMING \  
-D mapred.job.map.memory.mb=16000 \  
-input /user/farrell/fastq/LP6005113-  
DNA_A01.interleave.fastq.lzo \  
-file bwa \  
-file hs37d5.fasta.SA\  
-file hs37d5.fasta.amb\  
-file hs37d5.fasta.ann\  
-file hs37d5.fasta.pac\  
-file hs37d5.fasta.bwt\  
-file hs37d5.fasta.fai\  
-output /user/farrell/sam/A01 \  
-mapper 'bwa mem -p hs37d5.fasta -'\ \  
-reducer 'cat'\ \  
-numReduceTasks 1
```

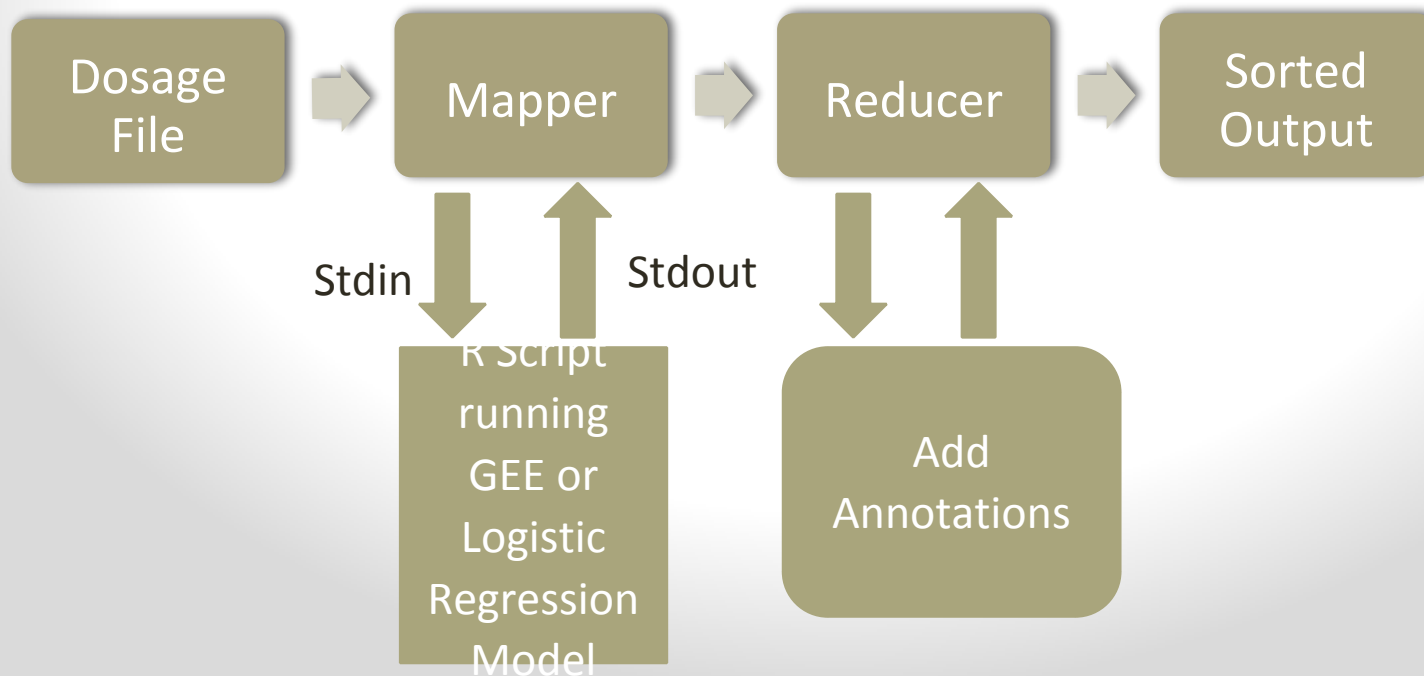


Example 2: Running GEE Models on Hadoop

- A common analysis in genomics is the modeling of phenotypes/diseases with the genomic variants/SNPs.
- The number of known variants has grown tremendously with about 37 million variants detected in 1000 genomes samples.
- A genomic study will run GEE or logistic regression models with multiple phenotypes and covariates for each variant above a minimum minor allele frequency. R is often the statistical package used on HPCC systems.
- To parallelize this on an HPCC system, the models will be split into 600 or more jobs to run tasks in parallel (600*50,000 variants=30 x 10⁶ models).
- However, HPCC scheduling system does not robustly deal with hardware failures during such runs.
- Splitting job into 1000s of tasks is not automatic must be coded/implemented by researcher.

Hadoop Streaming with R

- R is a widely used statistical software package on HPC that can also be readily used on Hadoop with Hadoop Streaming
- Hadoop + R gets around memory limitations and single core processing limitations of the base system
- Feeds data into stdin of R Script and collects output from stdout.

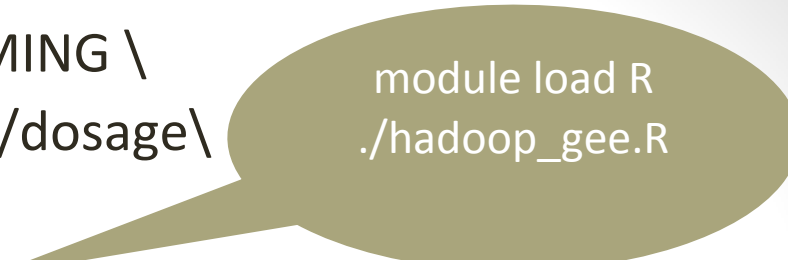


Benchmark of GEE on Hadoop

- Ran benchmark of GEE model on 3318 individuals with 37 million genotype doses (from Impute 2) on the 8-node FHS Hadoop Cluster
- If variant has MAF $\geq 3\%$ then run a GEE Model is run : $\text{Pheno} \sim \text{dose} + \text{gender} + \text{pc1} + \text{pc2} + \text{pc3}$)
- Hadoop automatically splits the submitted job into 2901 tasks
- Running 4 models completed in under 16 hours 38 min: about 4:10 per model
- Running just one model tools ran 6 hours (Estimated actually). To test robustness of system, one data node rebooted at 5 hours. Job continued to run, 32 tasks rescheduled automatically and job finished in 7 hours.
- On LinGA, 1 model took about 3 days. SCC estimate about 10 hours for one model.

Example 2 Command Line

```
hadoop jar $HADOOP_STREAMING \  
-input /user/farrell/addiction/dosage\  
-file `pwd`/hadoop_gee.R \  
-file `pwd`/hadoop_gee.cmd \  
-file `pwd`/hwe.R \  
-file `pwd`/sample.gz \  
-file `pwd`/phenotype.gz\  
-output /user/farrell/addiction/gee.results \  
-mapper 'hadoop_gee.cmd\  
-reducer 'cat\  
-numReduceTasks 1  
hadoop fs -cat addiction/gee.results/*
```



module load R
./hadoop_gee.R

Example 2 Resource Usage

agement Console - Mozilla Firefox

File History Bookmarks Tools Help

b_201311190929_000... x Rocks Management Console x

alhost/rocksUI/

StackIQ

Monitor Discover Network Attributes

Monitoring
Category » Global

Edit Settings

FHS Hadoop Cluster Load last hour

| | | | | |
|-------|-----------|-----------|-----------|----------|
| 1-min | Now:274.7 | Min: 2.0 | Avg:275.4 | Max:599. |
| Nodes | Now: 10.0 | Min: 10.0 | Avg: 10.0 | Max: 10. |
| CPUs | Now:284.0 | Min:284.0 | Avg:284.0 | Max:284. |
| Procs | Now:288.5 | Min: 1.0 | Avg:275.4 | Max:731. |

FHS Hadoop Cluster Memory last hour

| | | | | |
|--------|-------------|-------------|-------------|-------------|
| Mem | Now: 155.8G | Min: 44.8G | Avg: 142.2G | Max: 211.8G |
| Heap | Now: 8.0 | Min: 0.0 | Avg: 3.0 | Max: 0.0 |
| Cache | Now: 828.7G | Min: 799.7G | Avg: 827.8G | Max: 966.8G |
| Buffer | Now: 4.5G | Min: 4.1G | Avg: 4.3G | Max: 4.5G |
| Swap | Now: 317.7M | Min: 311.7M | Avg: 315.2M | Max: 317.7M |
| Total | Now: 1.0T | Min: 1.0T | Avg: 1.0T | Max: 1.0T |

FHS Hadoop Cluster CPU last hour

| | | | | |
|--------|------------|------------|------------|------------|
| User | Now: 78.5% | Min: 0.7% | Avg: 72.4% | Max: 79.3% |
| Nice | Now: 0.0% | Min: 0.0% | Avg: 0.0% | Max: 0.0% |
| System | Now: 1.1% | Min: 0.4% | Avg: 2.0% | Max: 56.3% |
| Idle | Now: 0.1% | Min: 0.0% | Avg: 0.1% | Max: 0.2% |
| Total | Now: 20.4% | Min: 10.1% | Avg: 25.1% | Max: 98.9% |

FHS Hadoop Cluster Network last hour

| | | | | |
|-----|------------|------------|------------|-----------|
| In | Now:320.5k | Min: 55.3k | Avg:421.5k | Max: 5.6M |
| Out | Now:427.9k | Min: 67.6k | Avg:436.7k | Max: 6.1M |

FHS Hadoop Cluster Packet Report last hour

| | | | | |
|-----|-----------|-----------|-----------|-----------|
| In | Now:504.2 | Min:220.1 | Avg:624.3 | Max: 6.7k |
| Out | Now:554.0 | Min:340.0 | Avg:741.7 | Max: 7.0k |

Message

Time

Hadoop Job Status

Hadoop job_201311190929_0008 on name-0-1

User: farrell

Job Name: streamjob1502095097594272236.jar

Job File: hdfs://name-0-1.local:9020/user/farrell/.staging/job_201311190929_0008/job.xml

Submit Host: data-1-1.local

Submit Host Address: 192.168.30.121

Job-ACLs: All users are allowed

Job Setup: [Successful](#)

Status: Succeeded

Started at: Tue Nov 19 09:41:20 EST 2013

Finished at: Wed Nov 20 02:19:35 EST 2013

Finished in: 16hrs, 38mins, 15sec

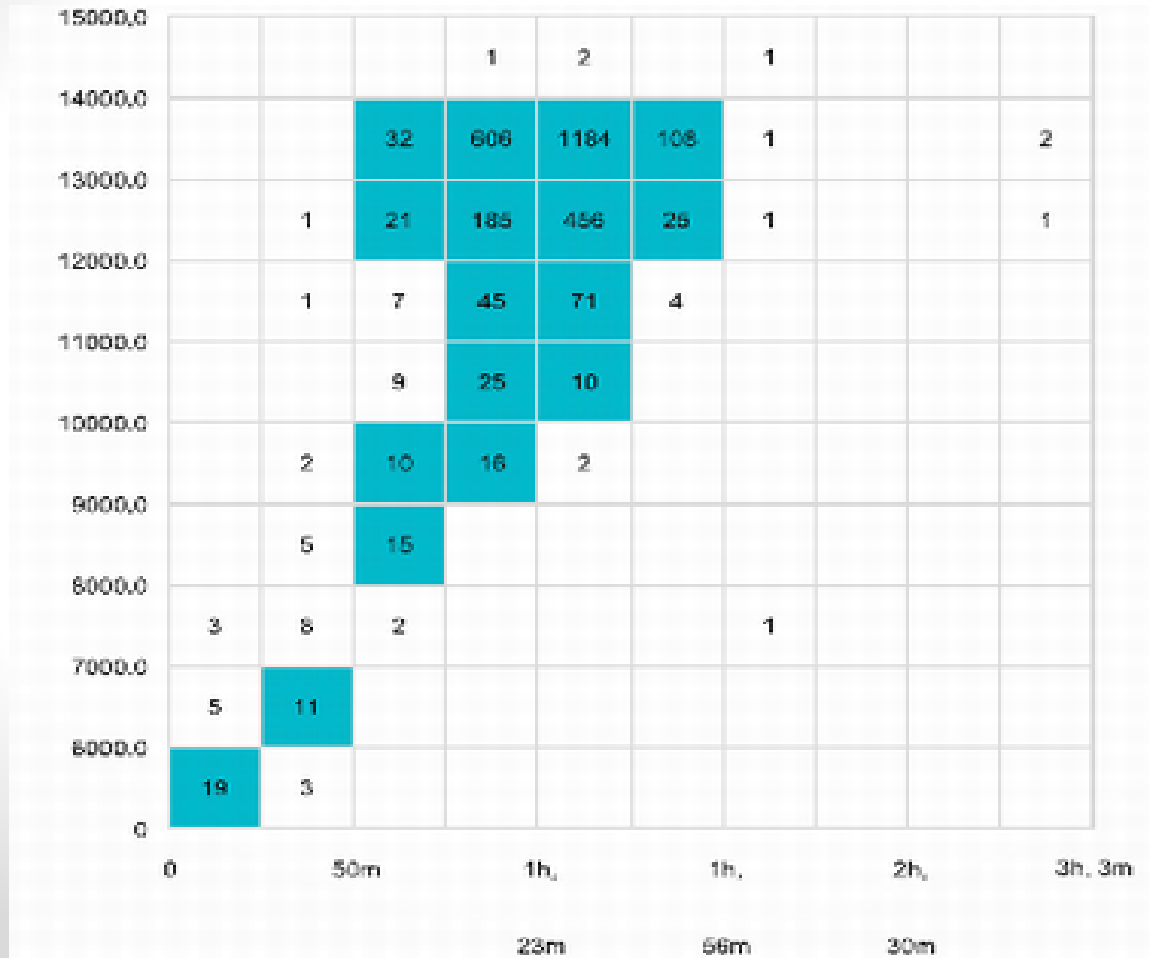
Job Cleanup: [Successful](#)

| Kind | % Complete | Num Tasks | Pending | Running | Complete | Killed | Failed/Killed Task Attempts |
|--------|----------------|-----------|---------|---------|----------|--------|-----------------------------|
| map | <u>100.00%</u> | 2901 | 0 | 0 | 2901 | 0 | 0 / 0 |
| reduce | <u>100.00%</u> | 1 | 0 | 0 | 1 | 0 | 0 / 0 |



Example 2: Map Input Records versus Duration

Mapped
Input
Records



Hadoop on Campus and in the Cloud

- Framingham Heart Study : 8-node Hadoop Cluster
- BU Computer Science Dept :12 node Hadoop Cluster
- Biomedical Genetics: 4 node cluster for development/testing.
- In the future ---a 24 node Hadoop Cluster has been proposed for FY2015 for Campus Wide Use with an opportunity for buy-in.
- Amazon Elastic Map Reduce (EMR) is the popular web service to distribute and process data on a resize-able Hadoop virtual cluster
- StarCluster (developed at MIT) is an open source cluster computing tool kit for deploying and managing clusters (HPCC and Hadoop) on Amazon's Elastic Compute Cloud.



Summary

- The Genomic Big Data continues to grow rapidly due to advances in sequencing technologies
- The Hadoop Framework permits the analysis of large genomic data sets across clusters of computers using very simple programming approaches
- Hadoop Streaming is a simple approach to scale analysis by using software already developed in R, Python, C , Java, and Perl.
- Apache Hadoop Platform's highly scalable framework will be an important platform to manage and analyze the avalanche of oncoming genomic data as whole genome sequencing becomes more widespread in research and clinical settings



- Thank you for your attention!!