PAGE
www.pagestudy.org

Automated Genotypic Imputation of PAGE II Data using Scientific Workflows

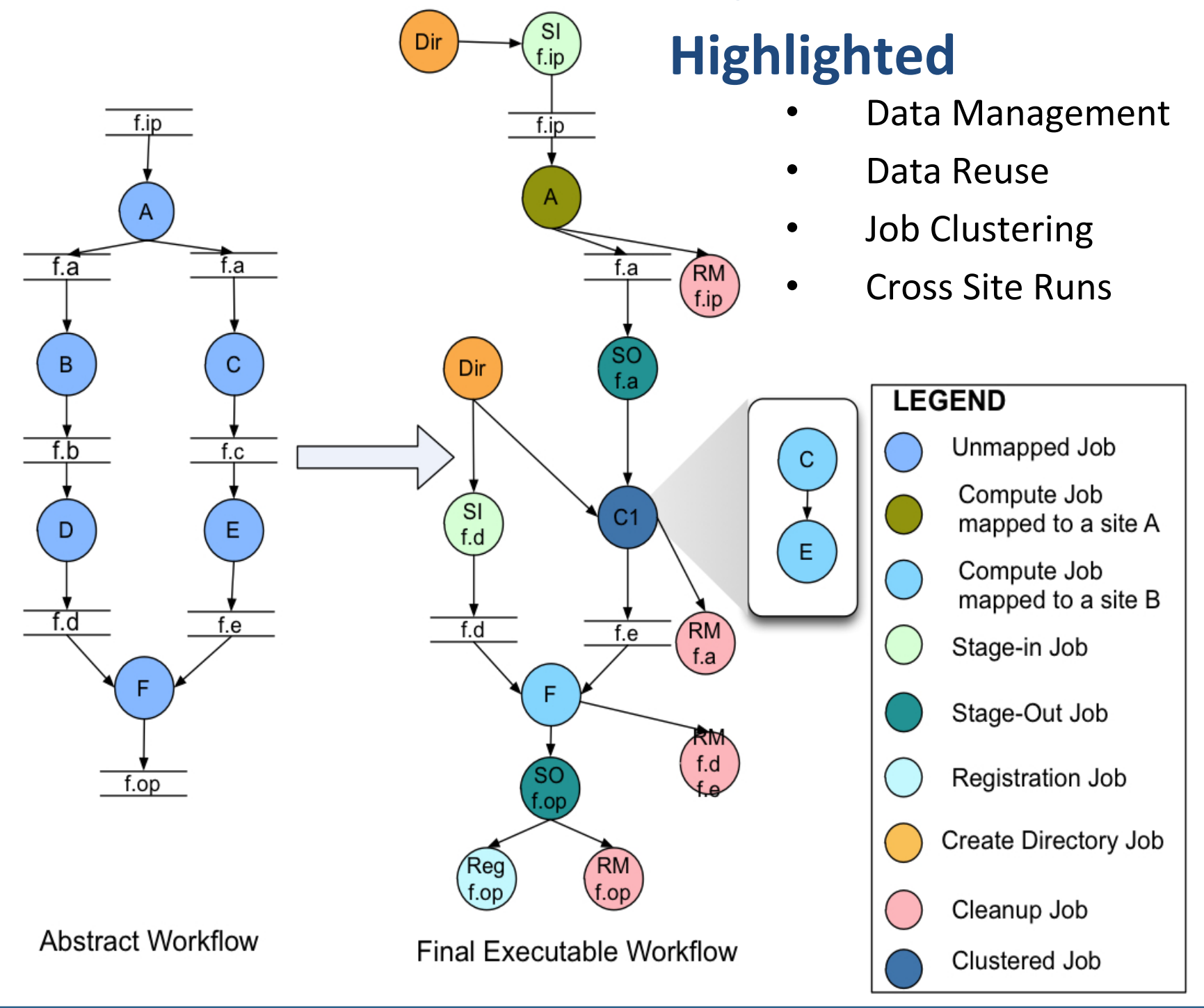
Karan Vahi[1], Steve Buyske[2], Lisheng Zhou[3], Tara Matise[3], Ewa Deelman[1]
[1]Science Automation Technologies, USC Information Science Institute
[2]Department of Statistics & Biostatistics, Rutgers University [3]Department of Genetics, Rutgers University

Background

- PAGE II (Population Architecture using Genomics and Epidemiology) study has genotyped 50,000 samples using MEGA, an Illumina high density consortium-built array array.
- We are also imputing an additional 50,000 subjects genotyped in 20 different GWAS studies.
- Imputation is done using SHAPEIT and IMPUTE2 with 1000 Genomes Project Reference Panel.
- Imputation can be parallelized by chromosome to reduce total processing time.
- We decided to use the Pegasus Workflow Management System to do runs on the Rutgers Genetics Cluster.

A Scientific Workflow

- Allows users to easily express multi-step computational tasks.
- Describes the dependencies among the tasks.
- Manages data flow.



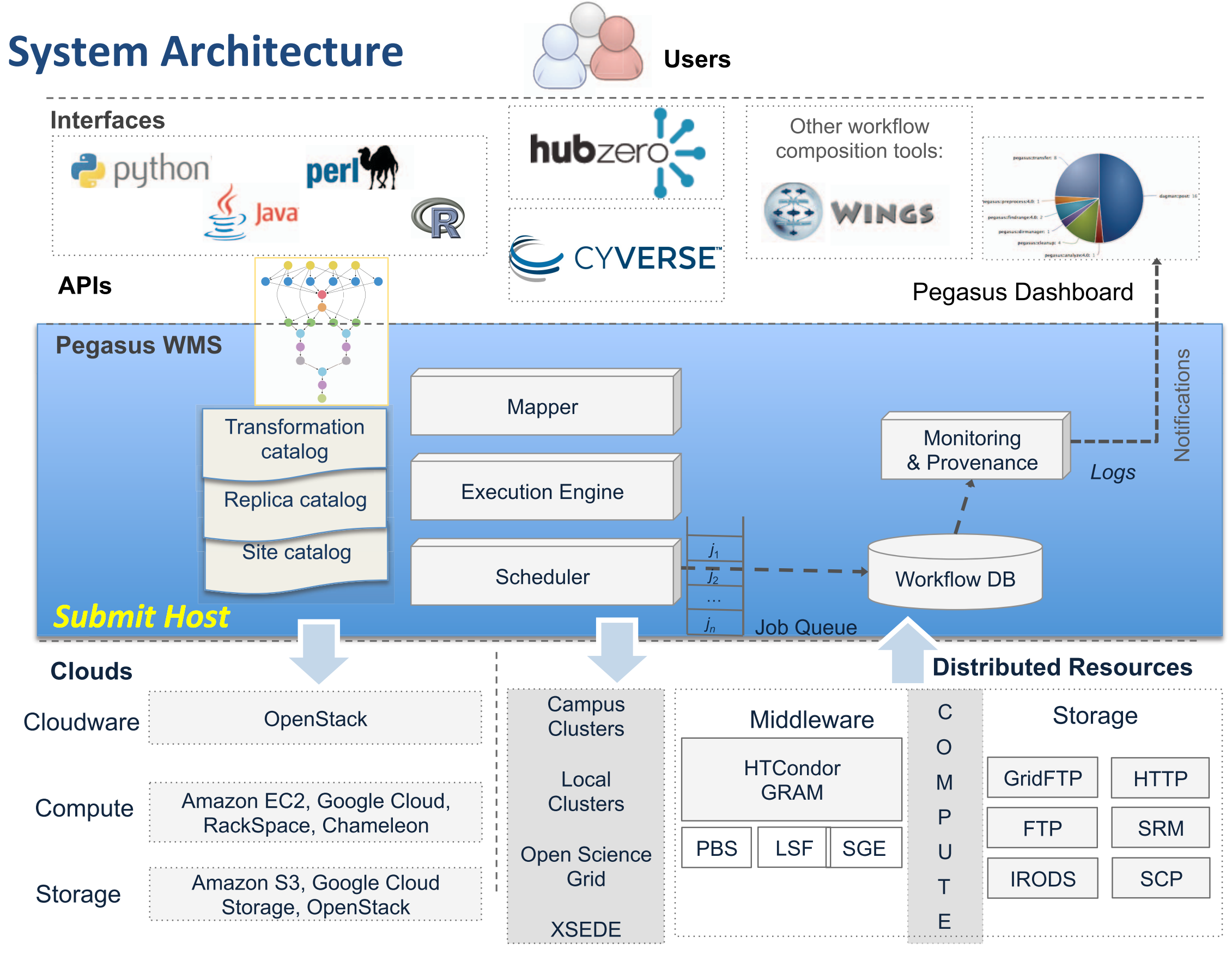
Capabilities Highlighted

- Data Management
- Data Reuse
- Job Clustering
- Cross Site Runs

Pegasus Workflow Management System

- General, open source solution for describing and executing workflows on laptops, clusters and clouds.
- Provides Python, Java, and Perl APIs for workflow creation.
- Provides portability, reliability, performance.

System Architecture



Documentation
http://pegasus.isi.edu

Applications
https://pegasus.isi.edu/application-showcase/

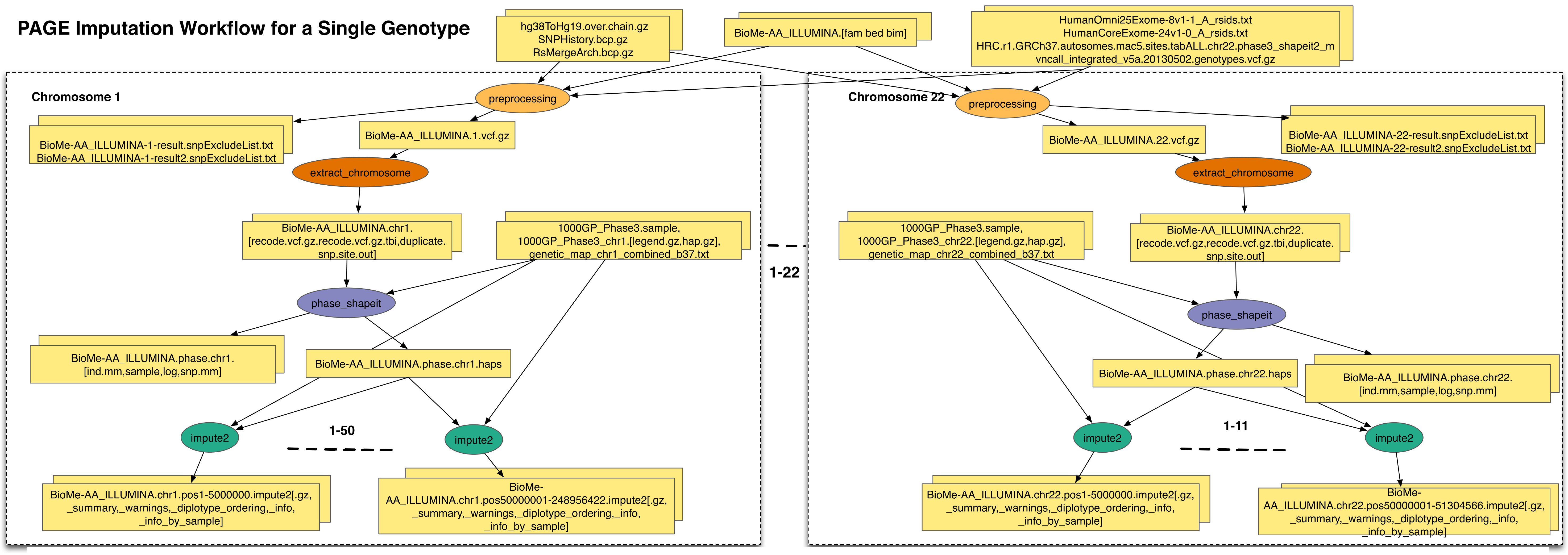
User Support
pegasus-users@isi.edu

Can handle workflows with millions of tasks, TB's of data.

- Can optimize the workflow from the point of view of performance, can handle data management across local and wide area networks, leveraging parallel file systems and object stores.
- Used in a number of domains: astronomy, bioinformatics, earthquake science, helioseismology, gravitational-wave physics, seismology, etc.

PAGE II Imputation Pipeline

PAGE Imputation Workflow for a Single Genotype


hg38ToHg19.over.chain.gz
SNPHistory.bcp.gz
RsMergeArch.bcp.gz
BioMe-AA_ILLUMINA.[fam bed bim]
HumanOmni25Exome-8v1-1_A_rsids.txt
HumanCoreExome-24v1-0_A_rsids.txt
HRC.r1.GRCh37.autosomes.mac5.sites.tabALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
Chromosome 1
preprocessing
BioMe-AA_ILLUMINA-1-result.snpExcludeList.txt
BioMe-AA_ILLUMINA-1-result2.snpExcludeList.txt
BioMe-AA_ILLUMINA.1.vcf.gz
extract_chromosome
BioMe-AA_ILLUMINA.chr1.[recode.vcf.gz,recode.vcf.gz.tbi,duplicate.snp.site.out]
1000GP_Phase3.sample, 1000GP_Phase3_chr1.[legend.gz,hap.gz], genetic_map_chr1_combined_b37.txt
phase_shapeit
BioMe-AA_ILLUMINA.phase.chr1.[ind.mm,sample,log,snp.mm]
BioMe-AA_ILLUMINA.phase.chr1.haps
impute2
1-50
impute2
BioMe-AA_ILLUMINA.chr1.pos1-5000000.impute2[.gz,_summary,_warnings,_diplotype_ordering,_info,_info_by_sample]
BioMe-AA_ILLUMINA.chr1.pos50000001-248956422.impute2[.gz,_summary,_warnings,_diplotype_ordering,_info,_info_by_sample]
1-22
Chromosome 22
preprocessing
BioMe-AA_ILLUMINA.22.vcf.gz
BioMe-AA_ILLUMINA-22-result.snpExcludeList.txt
BioMe-AA_ILLUMINA-22-result2.snpExcludeList.txt
extract_chromosome
1000GP_Phase3.sample, 1000GP_Phase3_chr22.[legend.gz,hap.gz], genetic_map_chr22_combined_b37.txt
BioMe-AA_ILLUMINA.chr22.[recode.vcf.gz,recode.vcf.gz.tbi,duplicate.snp.site.out]
phase_shapeit
BioMe-AA_ILLUMINA.phase.chr22.haps
BioMe-AA_ILLUMINA.phase.chr22.[ind.mm,sample,log,snp.mm]
impute2
1-11
impute2
BioMe-AA_ILLUMINA.chr22.pos1-5000000.impute2[.gz,_summary,_warnings,_diplotype_ordering,_info,_info_by_sample]
BioMe-AA_ILLUMINA.chr22.pos50000001-51304566.impute2[.gz,_summary,_warnings,_diplotype_ordering,_info,_info_by_sample]

Pipeline available at **https://github.com/pegasus-isi/page-imputation**

- Many resources have been put into dense genotyping - we want to maximize the utility of that data.
- Multi-center consortia have heterogenous collections of GWAS array data.

Tools used

- Plink
- liftOver
- LiftRsNumber.py
- HRC-check-bim.pl
- vcftools
- SHAPEIT
- IMPUTE2

Inputs

- 53536 samples across 17 studies (71MB to 2.6GB)

Outputs

- Range from 21GB to 204GB
- Made available via the PAGE FTP Server

Benefits of Pegasus Workflows for Imputation

- Imputation results offer a common data framework, but the heterogeneity of the inputs can present difficulties.
- Future updates to imputation reference panels mean that the entire process may need to be repeated.
- This workflow is agnostic as to array, strand, and build of the input data
- If components of the imputation fail (in our case, usually because of our heterogeneous compute cluster), PEGASUS can automatically resume the required components while not repeating the completed components.
- Pegasus keeps the many component files distinct and prevents filename collisions.
- Once constructed, the workflow requires almost no analyst effort - literally a few minutes per dataset.
- To re-impute with a new panel therefore also requires almost no analyst effort.

Workflow Tracking and Error Reporting

- Workflow progress can be monitored through the pegasus dashboard, or the user can wait for workflow completion email notification.
- Generates an error report when things go wrong.
- Error reports indicate the source of error and what tasks failed.

Failure Reasons

- Jobs failed because of incorrect memory estimates or low memory available on nodes.
- Input data not clean. Some allele names are recorded as 0, scripts were modified to use PLINK to generate cleaned datasets from the bim files.

Recovery Semantics

- Ability to recover and resubmit pipeline from point of last failure.
- Automatic job retries.
- If everything else failed, memory requirements were increased and workflow was submitted again, pruning the jobs that were successfully executed part of the previous failed run.
- **For the largest study, impute jobs took about 42GB of memory.**

ASHG2016
VANCOUVER • OCTOBER 18-22, 2016
SHARING DISCOVERIES. SHAPING OUR FUTURE.

PAGE II
www.pagestudy.org
Population Architecture Using Genomics and Epidemiology

WHI
FRED HUTCHINSON CANCER RESEARCH CENTER

Coordinating Center
RUTGERS
USC INFORMATION SCIENCES INSTITUTE

MEC
The Multiethnic Cohort Study

CALiCo
ARIC
CARDIA

NHGRI