PAGE
www.pagestudy.org

pegasus
http://pegasus.isi.edu

Conducting Large-Scale Imputation Studies on the Cloud

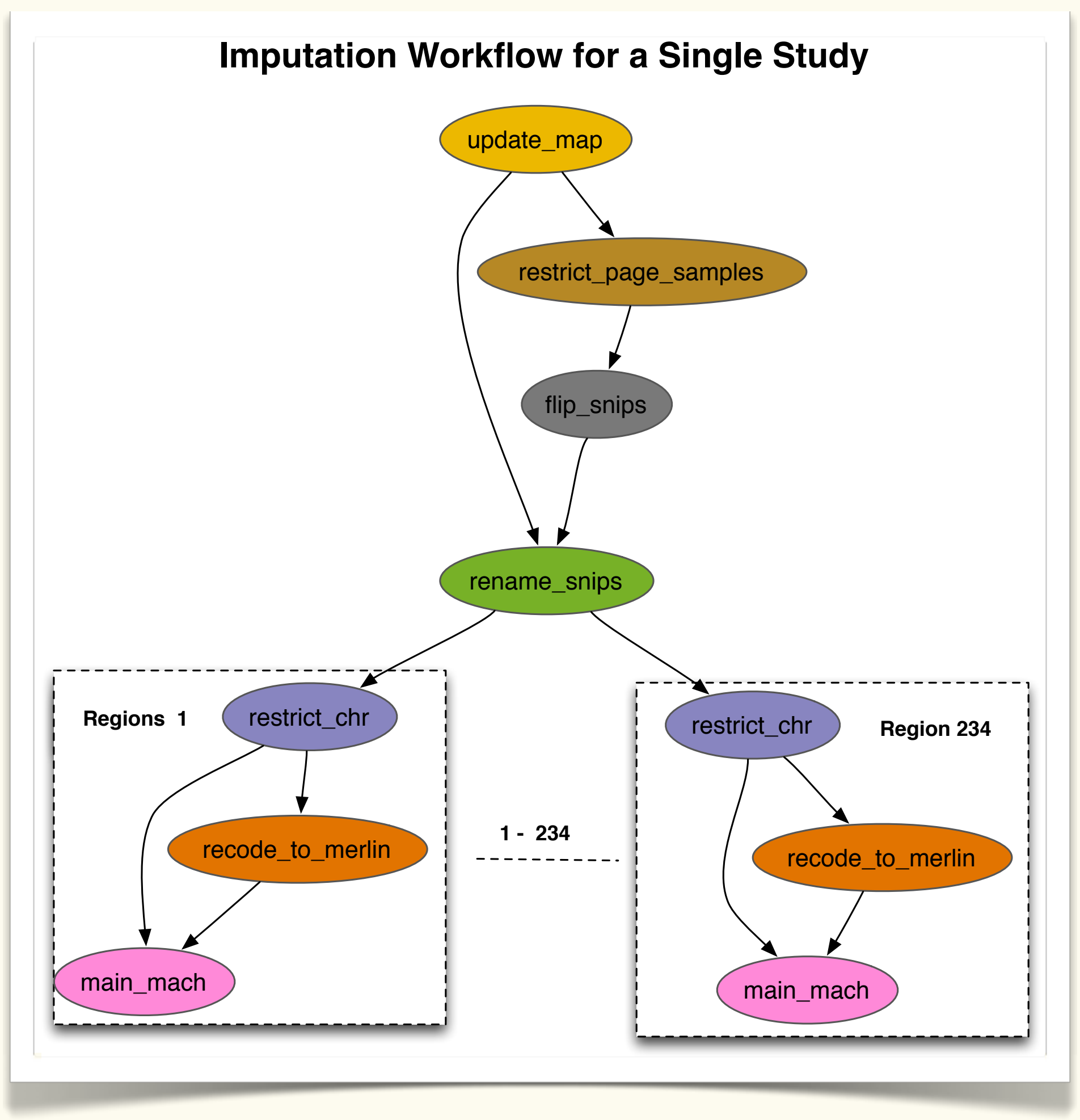
Steven Buyske[1,2], Karan Vahi[3], Ewa Deelman[3], Ulrike Peters[4], & Tara Matise[2]
[1]Department of Statistics & Biostatistics and [2]Department of Genetics, Rutgers University, Piscataway, NJ
[3]USC Information Sciences Institute, Marina del Rey, CA
[4]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA

Background

- PAGE has multiple studies and multiple ancestry groups genotyped on the Metabochip
- We wanted to impute the 235 fine-mapping Metabochip regions to the 1KGP cosmopolitan panel using MaCH-Admix
- Lot of data subsets makes this a logistic headache
- We decided to use the Pegasus Workflow Management System with the Amazon Elastic Compute Cloud (EC2)

A Scientific Workflow

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


Imputation Workflow for a Single Study

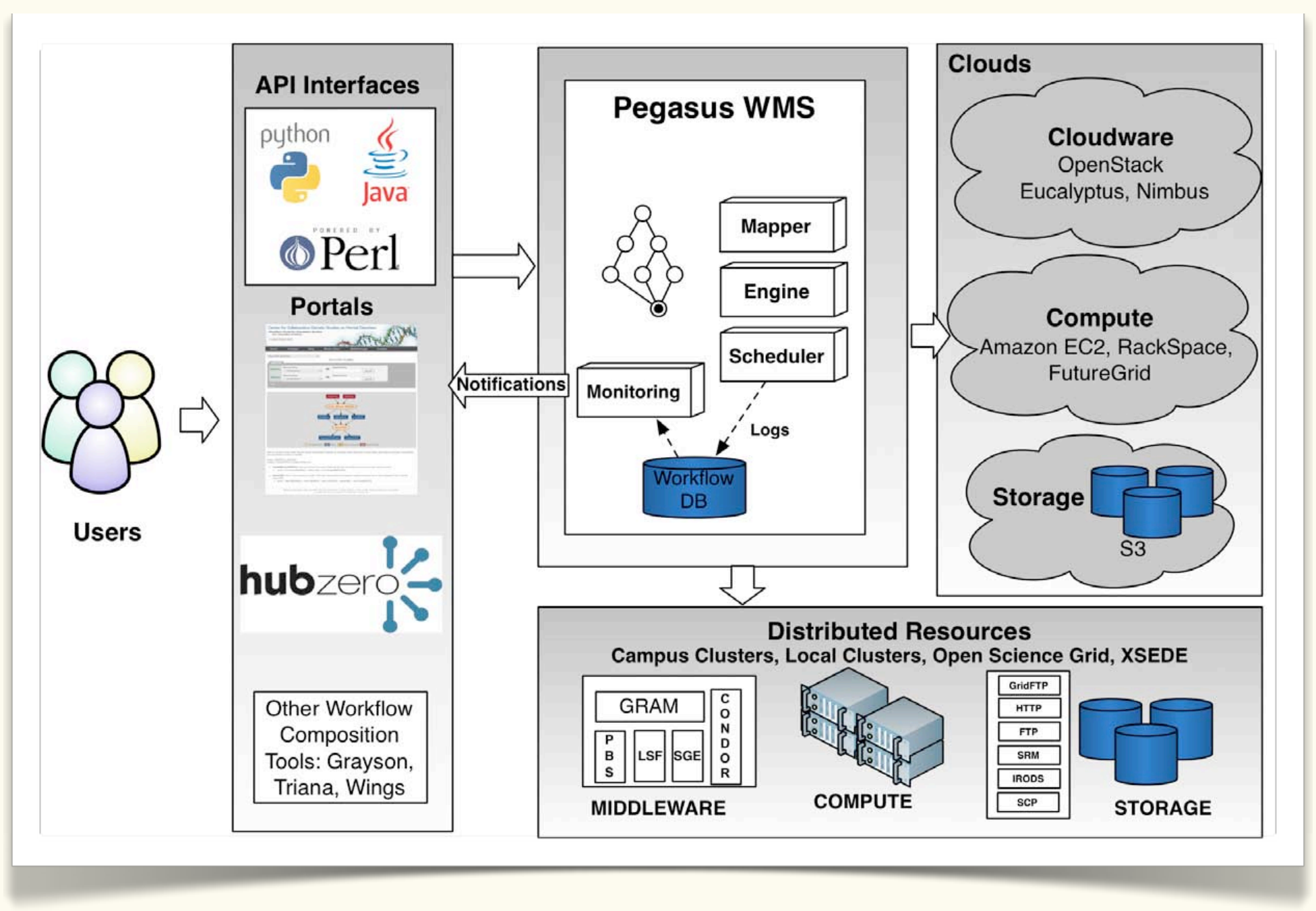
A Workflow Management System

- Automates tasks that user could perform manually … but the workflow management system takes care of automatically
- Includes features such as retries in the case of failures—avoids the need for user intervention
- Can itself include error checking
- Utilizes many resources from one user action while maintaining complex job interdependencies and data flows
- Efficiently uses computing resources / human time

Pegasus Workflow Management System

- Collaboration between USC and the Condor Team at UW Madison
- Maps a resource-independent "abstract" workflow onto resources and executes the "executable" workflow
- Provides reliability—can retry computations from the point of failure (with email notification)
- Provides scalability—can handle large data and many computations
- Structures workflows for performance
- Automatically captures provenance information
- Can run on resources distributed among institutions, laptop, campus cluster, grid, cloud
- pegasus.isi.edu

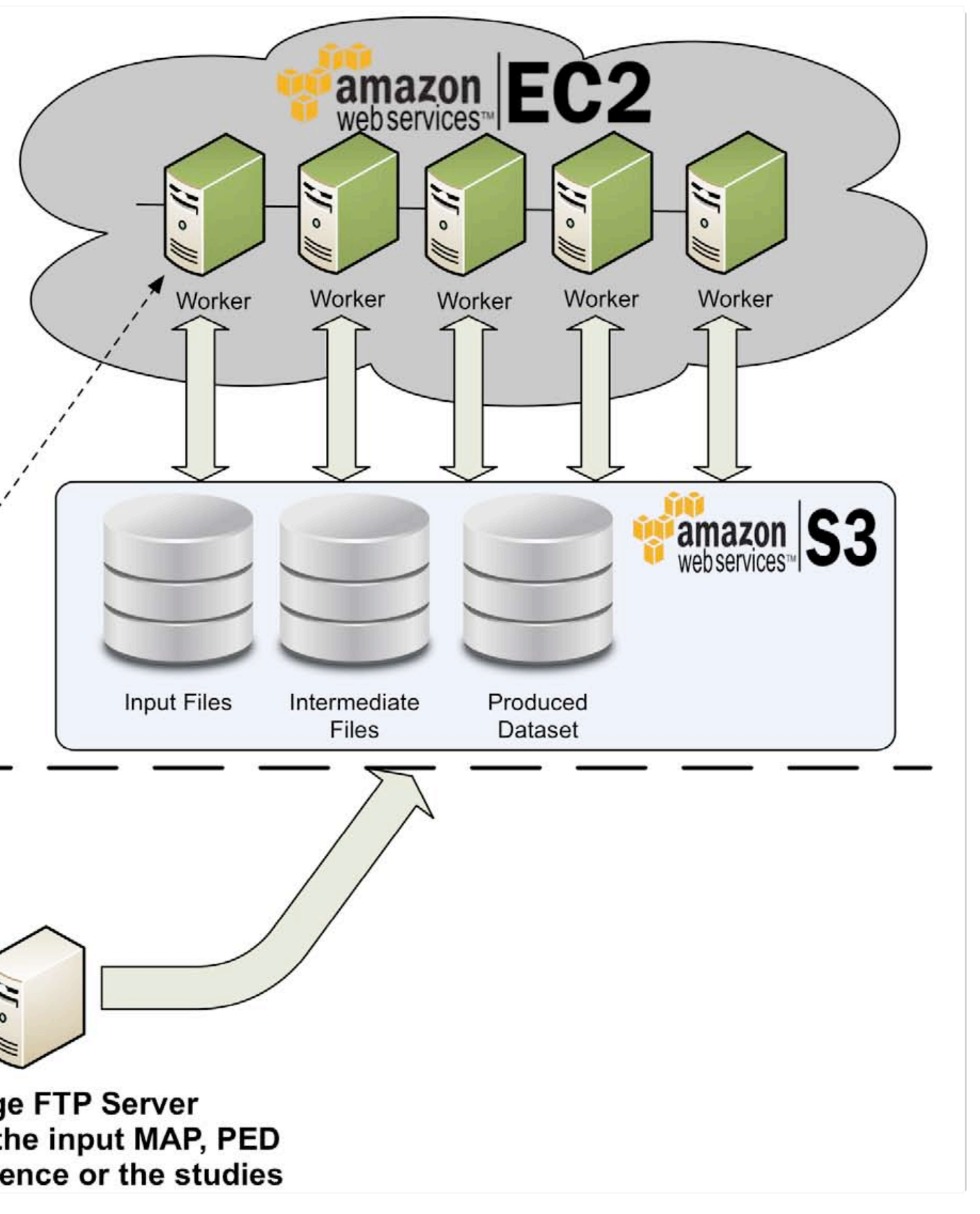
Pegasus Environment



Project Inputs & Outputs

- Workflow description (built from an example for a single run)
- EC2 machine image with Linux, PLINK, MaCH-Admix, etc.
- Locations of the input files on the PAGE FTP server
- Study-specific input files
  - Transferred on demand by Pegasus first to S3 (Amazon cloud storage system)
  - Grabbed by worker nodes
- Large reference files used for all runs (e.g., 1KGP reference panel)
  - transferred once to S3
  - grabbed directly from S3 by the jobs on the worker nodes
- Output files produced by worker nodes are pushed back to S3
  - Intermediate files can be stored for debugging or deleted
  - Final outputs can be bundled up for distribution or processed for further QC

Project Architecture



Missteps

- Test runs on small samples led to underestimation of required memory—8.5 GB/core, or an Amazon m2.xlarge instance
- Wrappers were needed to detect silent failures in executables
- Amazon normally limits users to 20 instances at a time—our ultimate use of m2.4xlarge instances, with 8 vCPUs, kept us away from the limit

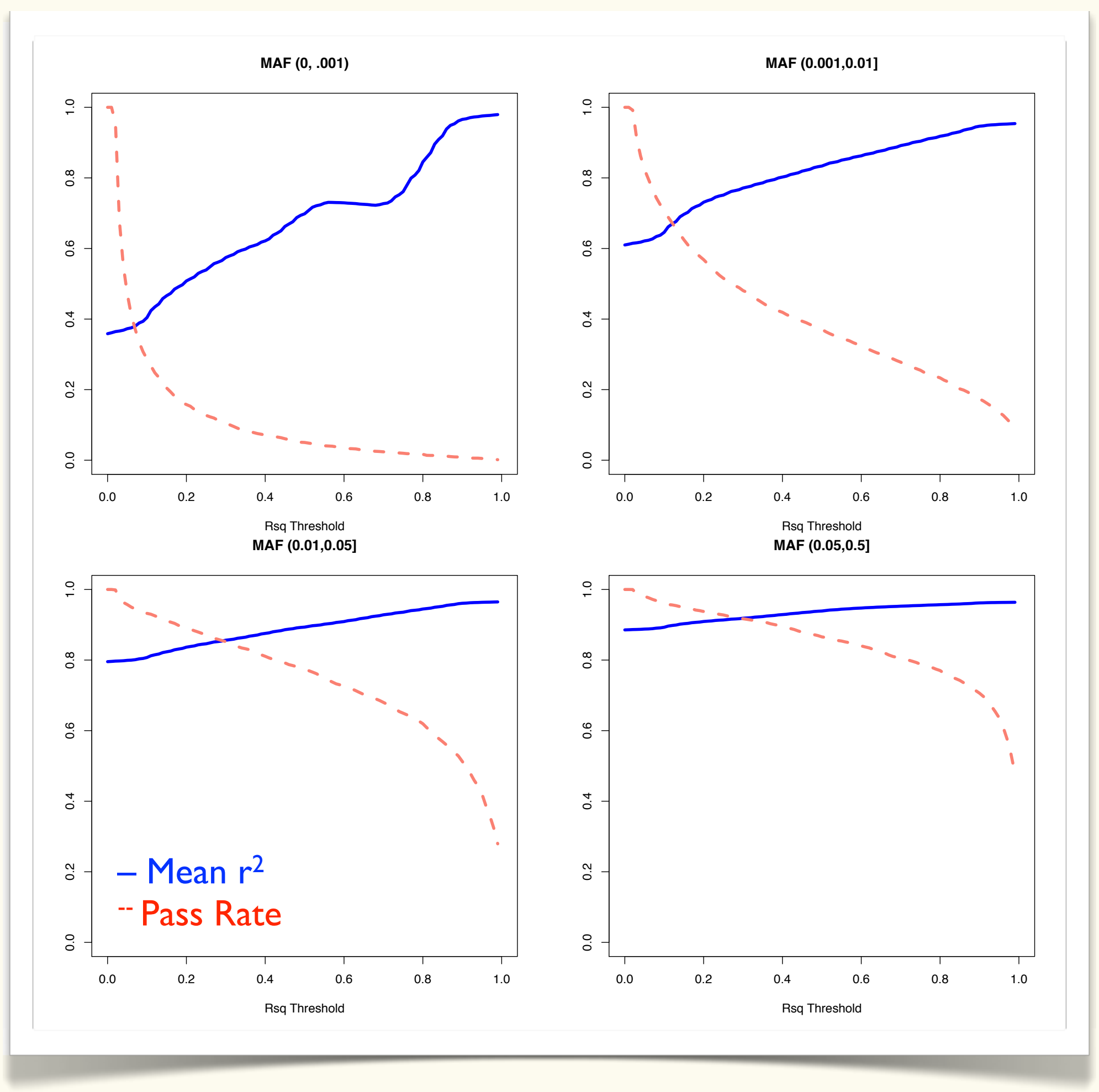
By the Numbers

- ~153 days computing time (m2.4xlarge instances)
  - Would require ~1 week using 20 instances
- ~$6,000 cost
- ~70,000 samples across 17 study/ancestry combinations in 235 regions
- ~120,000 Metabochip-genotyped SNPs in Metabochip's fine-mapping regions
- Average of (more than)* 340,000 additional SNPs imputed per study/ancestry group

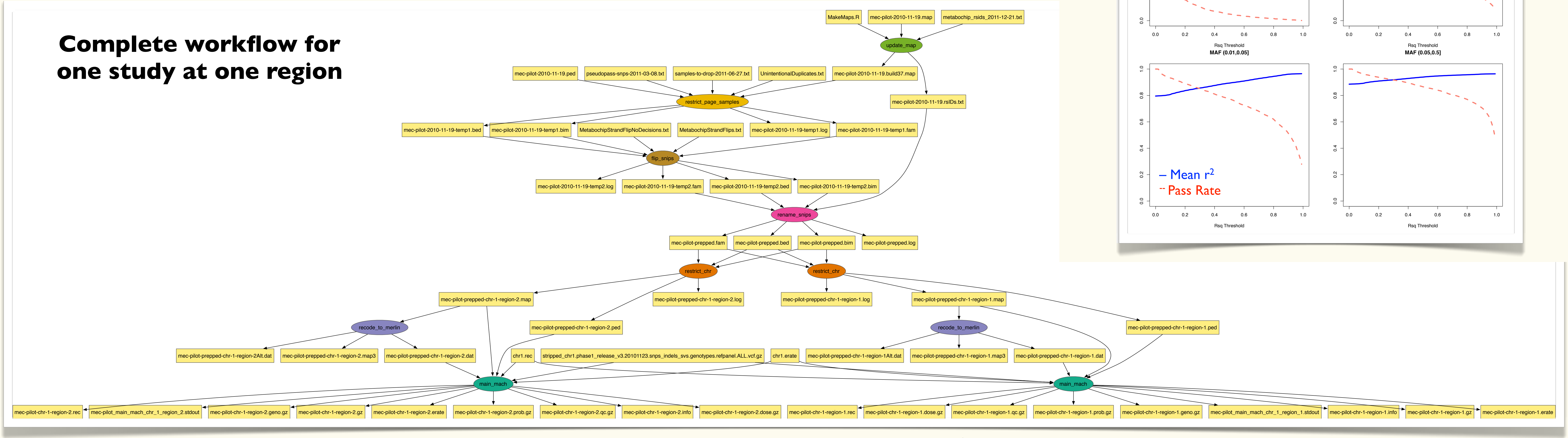
  *Some regions generated no SNPs (coding error) and have to be redone

Imputation Quality

- Not the focus of the poster, but …
- Figure shows quality in an African-American subsample



Complete workflow for one study at one region


MakeMaps.R
mec-pilot-2010-11-19.map
metabochip_rsids_2011-12-21.txt
update_map
mec-pilot-2010-11-19.ped
pseudopass-snps-2011-03-08.txt
samples-to-drop-2011-06-27.txt
UnintentionalDuplicates.txt
mec-pilot-2010-11-19.build37.map
restrict_page_samples
mec-pilot-2010-11-19.rsIDs.txt
mec-pilot-2010-11-19-temp1.bed
mec-pilot-2010-11-19-temp1.bim
MetabochipStrandFlipNoDecisions.txt
MetabochipStrandFlips.txt
mec-pilot-2010-11-19-temp1.log
mec-pilot-2010-11-19-temp1.fam
flip_snips
mec-pilot-2010-11-19-temp2.log
mec-pilot-2010-11-19-temp2.fam
mec-pilot-2010-11-19-temp2.bed
mec-pilot-2010-11-19-temp2.bim
rename_snips
mec-pilot-prepped.fam
mec-pilot-prepped.bed
mec-pilot-prepped.bim
mec-pilot-prepped.log
restrict_chr
restrict_chr
mec-pilot-prepped-chr-1-region-2.map
mec-pilot-prepped-chr-1-region-2.log
mec-pilot-prepped-chr-1-region-1.log
mec-pilot-prepped-chr-1-region-1.map
recode_to_merlin
mec-pilot-prepped-chr-1-region-2.ped
recode_to_merlin
mec-pilot-prepped-chr-1-region-1.ped
mec-pilot-prepped-chr-1-region-2Alt.dat
mec-pilot-prepped-chr-1-region-2.map3
mec-pilot-prepped-chr-1-region-2.dat
chr1.rec
stripped_chr1.phase1_release_v3.20101123.snps_indels_svs.genotypes.refpanel.ALL.vcf.gz
chr1.erate
mec-pilot-prepped-chr-1-region-1Alt.dat
mec-pilot-prepped-chr-1-region-1.map3
mec-pilot-prepped-chr-1-region-1.dat
main_mach
main_mach
mec-pilot-chr-1-region-2.rec
mec-pilot_main_mach_chr_1_region_2.stdout
mec-pilot-chr-1-region-2.geno.gz
mec-pilot-chr-1-region-2.gz
mec-pilot-chr-1-region-2.erate
mec-pilot-chr-1-region-2.prob.gz
mec-pilot-chr-1-region-2.qc.gz
mec-pilot-chr-1-region-2.info
mec-pilot-chr-1-region-2.dose.gz
mec-pilot-chr-1-region-1.rec
mec-pilot-chr-1-region-1.dose.gz
mec-pilot-chr-1-region-1.qc.gz
mec-pilot-chr-1-region-1.prob.gz
mec-pilot-chr-1-region-1.geno.gz
mec-pilot_main_mach_chr_1_region_1.stdout
mec-pilot-chr-1-region-1.info
mec-pilot-chr-1-region-1.gz
mec-pilot-chr-1-region-1.erate

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

EAGLE

WHI

Coordinating Center

MEC

CALiCo

NHGRI