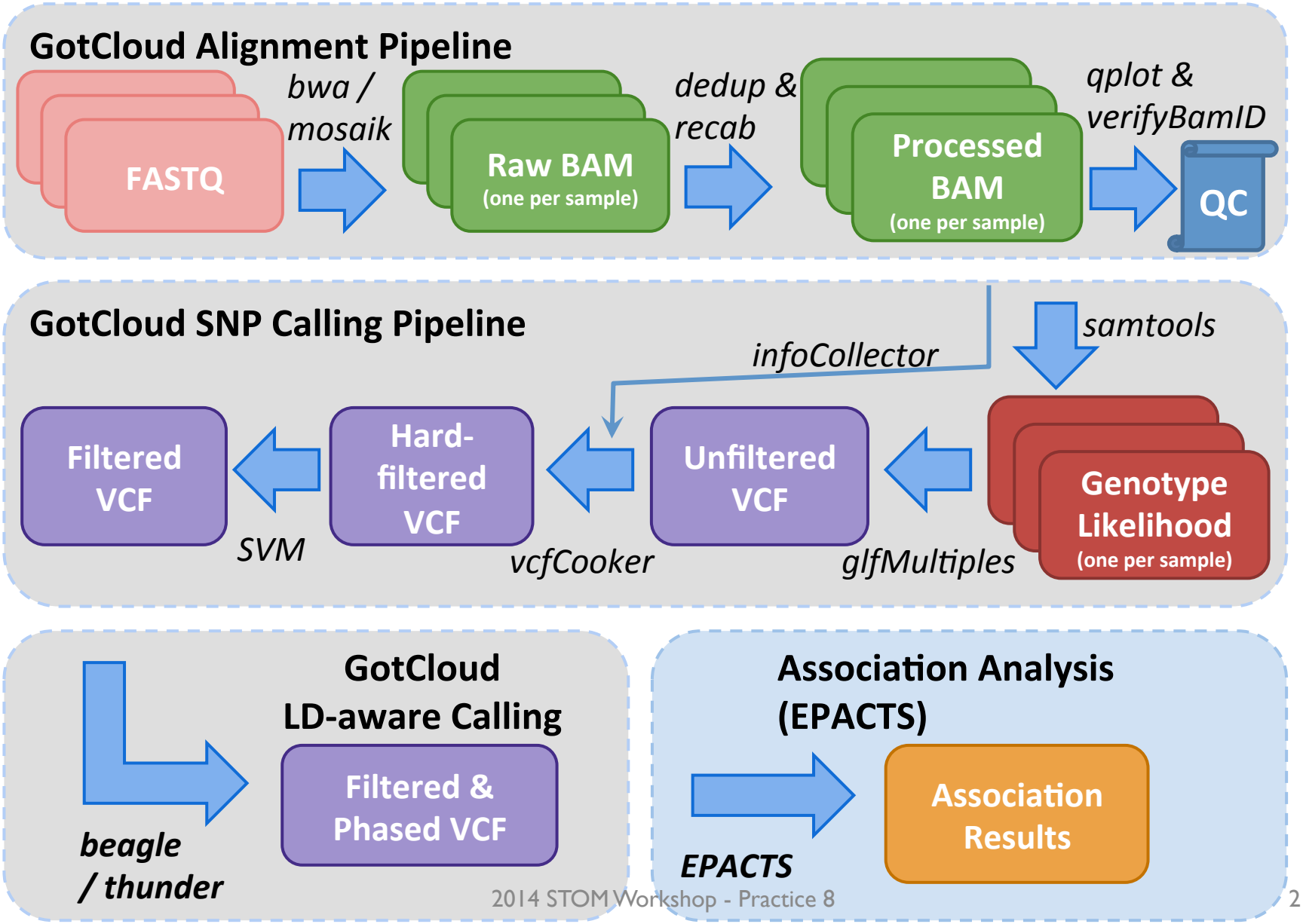
Sequence-based Association, Interpretation, Visualization using EPACTS

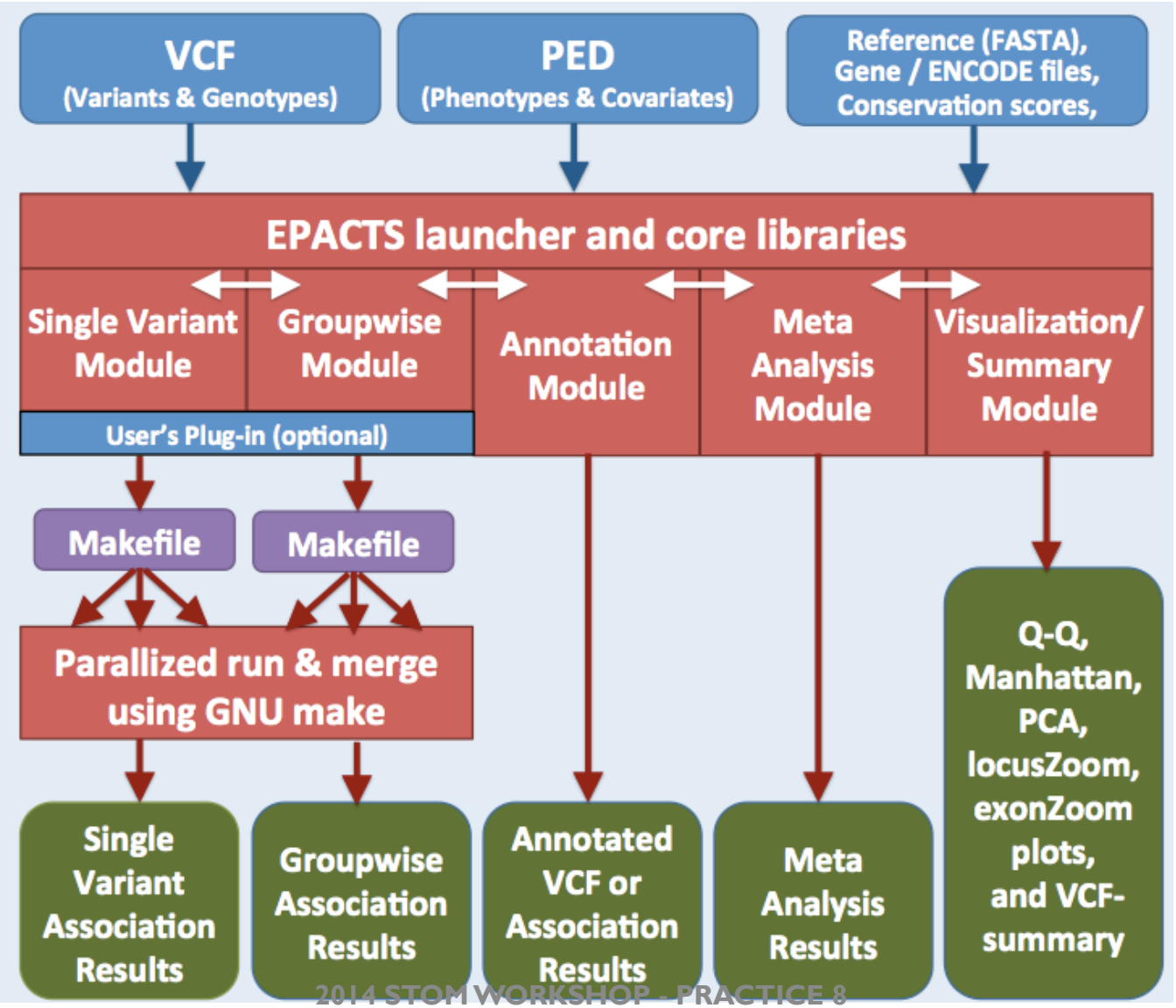
June 19th, 2014
Sequence Analysis Workshop

Hyun Min Kang
University of Michigan, Ann Arbor

EPACTS Association Analysis Pipeline



OVERVIEW OF EPACTS FRAMEWORK



CHALLENGES IN SEQUENCE-BASED ASSOCIATION

- Much larger (10~100x) data size
  - Efficient and parallel computation is important

- Complex representation of variants and genotypes
  - SNPs, Indels, structural variations with multi-allelic variants
  - Genotypes with uncertainty across different depth and quality
  - Efficient implementation VCF (Variant Call Format) files is not simple

- Many methods are published, but only a few are usefully implemented.
  - Software implementation is becoming a major bottleneck
  - Need tools to transform “methods” to “software”

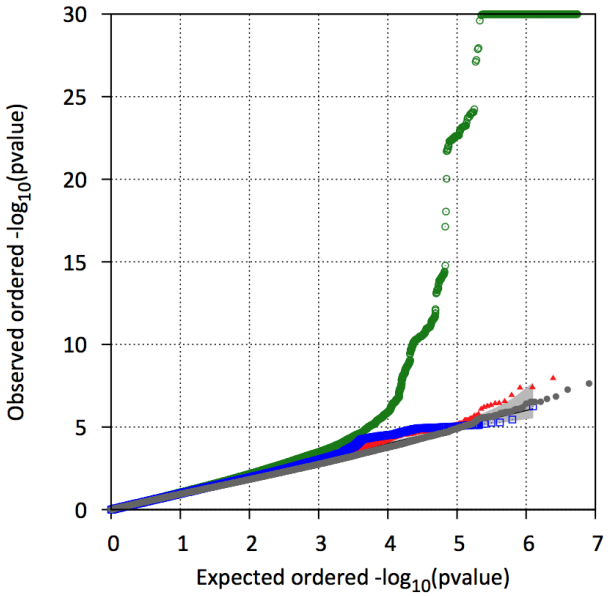
KEY FEATURES OF EPACTS

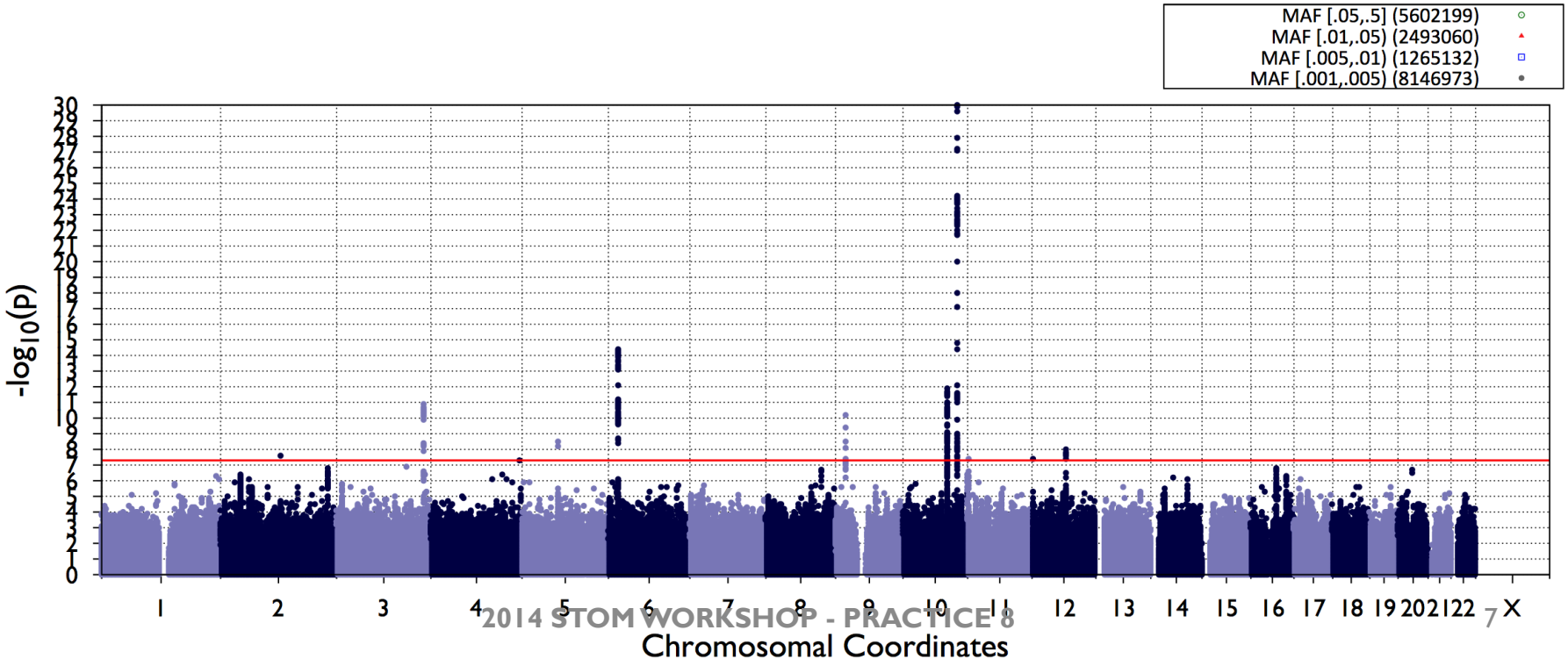
- Convenient and **dynamic plug-in** of user-defined statistical tests
  - Facilitate interaction between method developers and users
- **Efficient and parallel access of VCF** files
- **Fault-tolerant pipeline** structure based on GNU make
- Support of **a variety of single variant and groupwise tests**
- **Convenient** to run
  - All you need is just VCF and phenotype (PED) file
- **Automated visualization** of association signals and QC metrics
  - QQ-plot, Manhattan plot, PCA plot, LocusZoom plot
- Automated **annotation** of coding and noncoding variants
- Under **active development** more features are in progress (e.g. eQTL)

STATISTICAL TESTS AVAILABLE

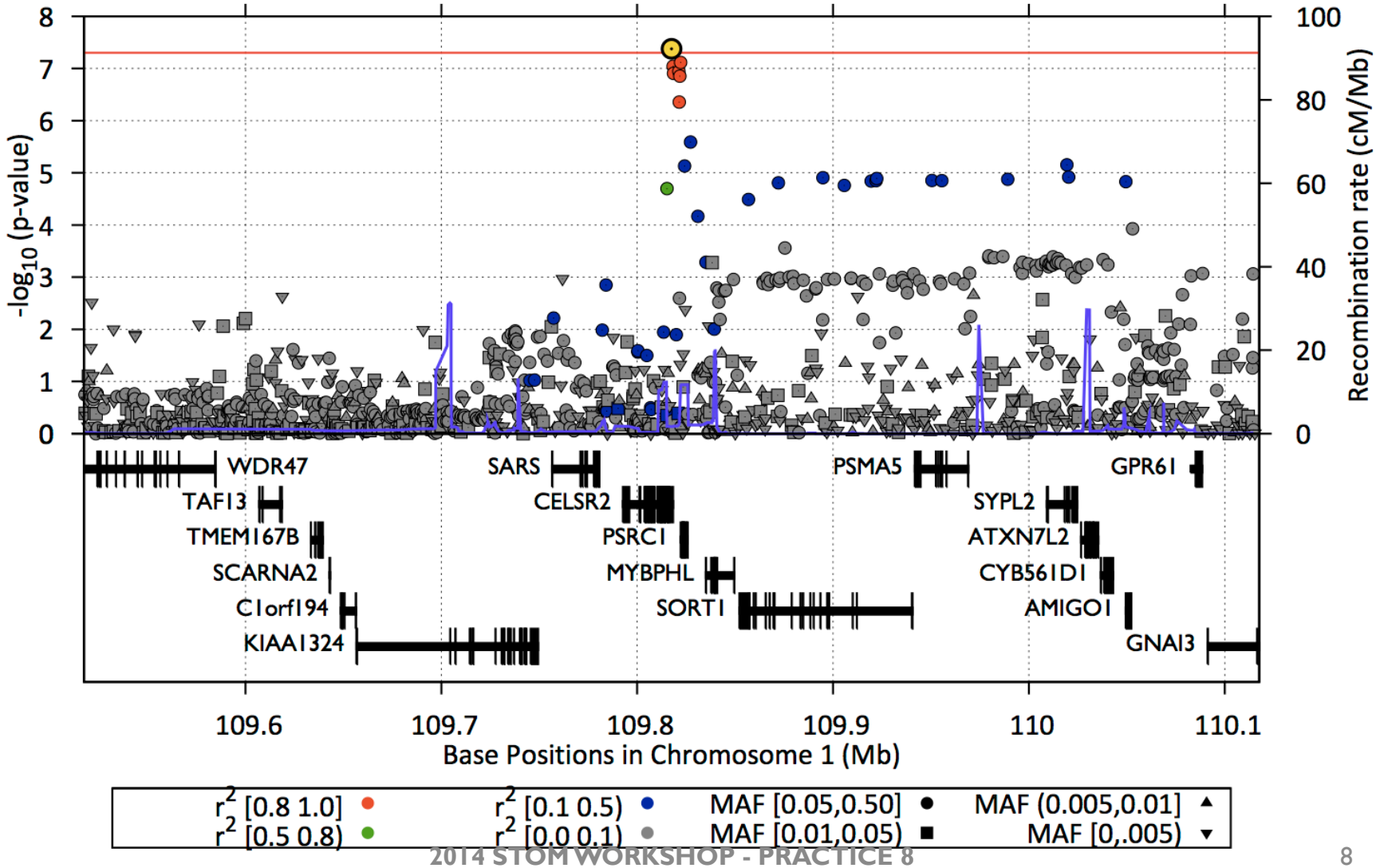
Single Variant Test	Groupwise Test
Wald Test	Collapsing
Score Test	Madsen-Browning*
Likelihood-ratio test	Reverse Regression
Firth bias-corrected LRT	SKAT / SKAT-O
Reverse Regression	VariableThreshold (VT)
Wilcoxon Rank Sum	EMMAX-Collapsing
EMMAX	EMMAX-VT

Example of Manhattan & QQ plots Automatically generated by EPACTS using a genome-wide data

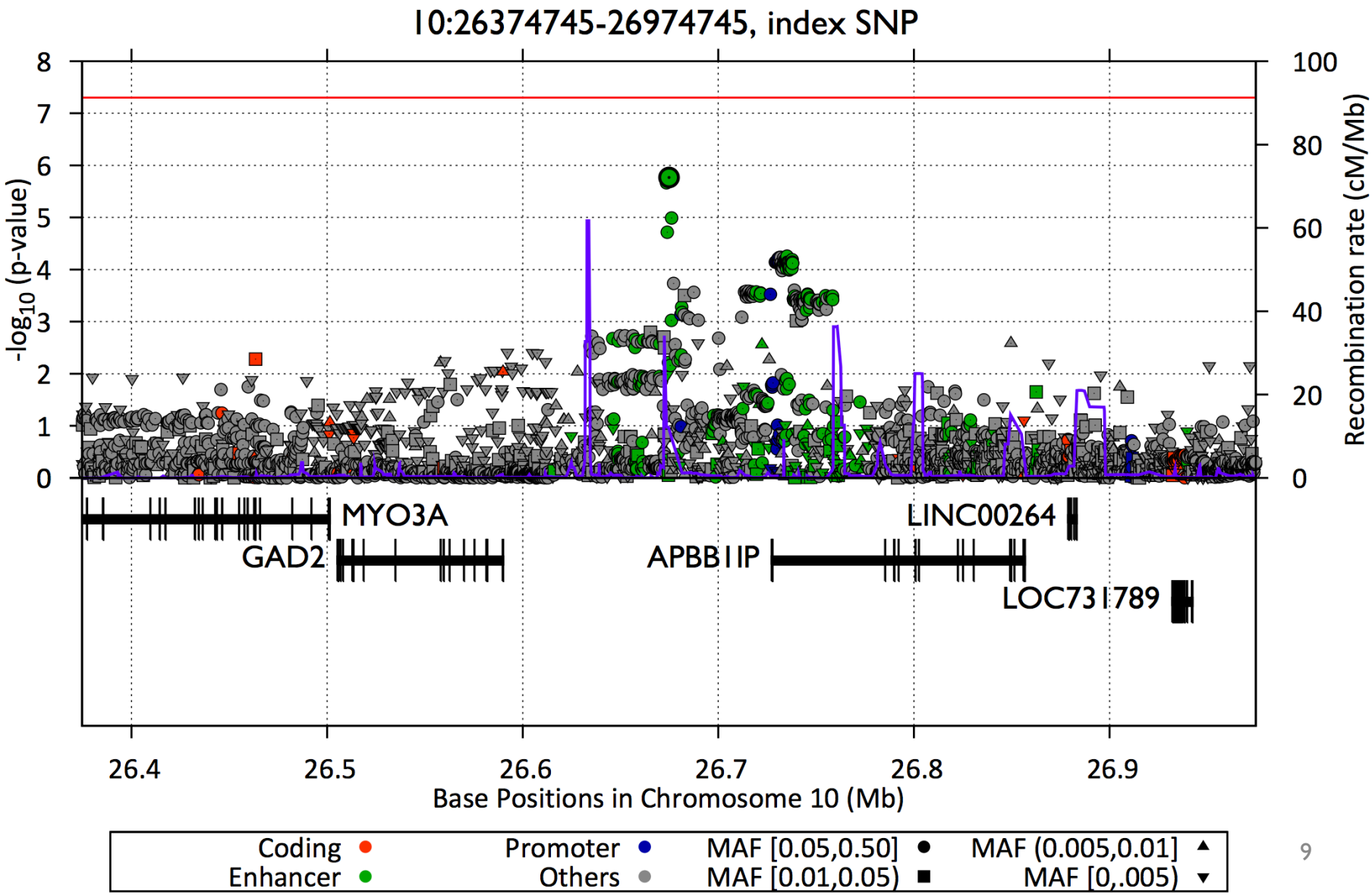




Zoom Plots for top associations



Zoom Plots by Regulatory Regions



GETTING STARTED WITH EPACTS

- Input Files – What should we provide?
  - VCF : genotype data (bgzipped and tabixed)
    - [prefix].vcf.gz and [prefix].vcf.gz.tbi should exist
  - PED : phenotype & covariate data
    - Header can be in a separate file (.dat) or in the first line (starting with #)
- Additional Input Files (Optional)
  - Marker group data (for groupwise test)
  - Reference genome sequence (for annotation)
  - Gene annotation files (in UCSC format)
  - ENCODE chromatin state predictions