Practical Session 8

Sequence-based Association, Interpretation, Visualization using EPACTS

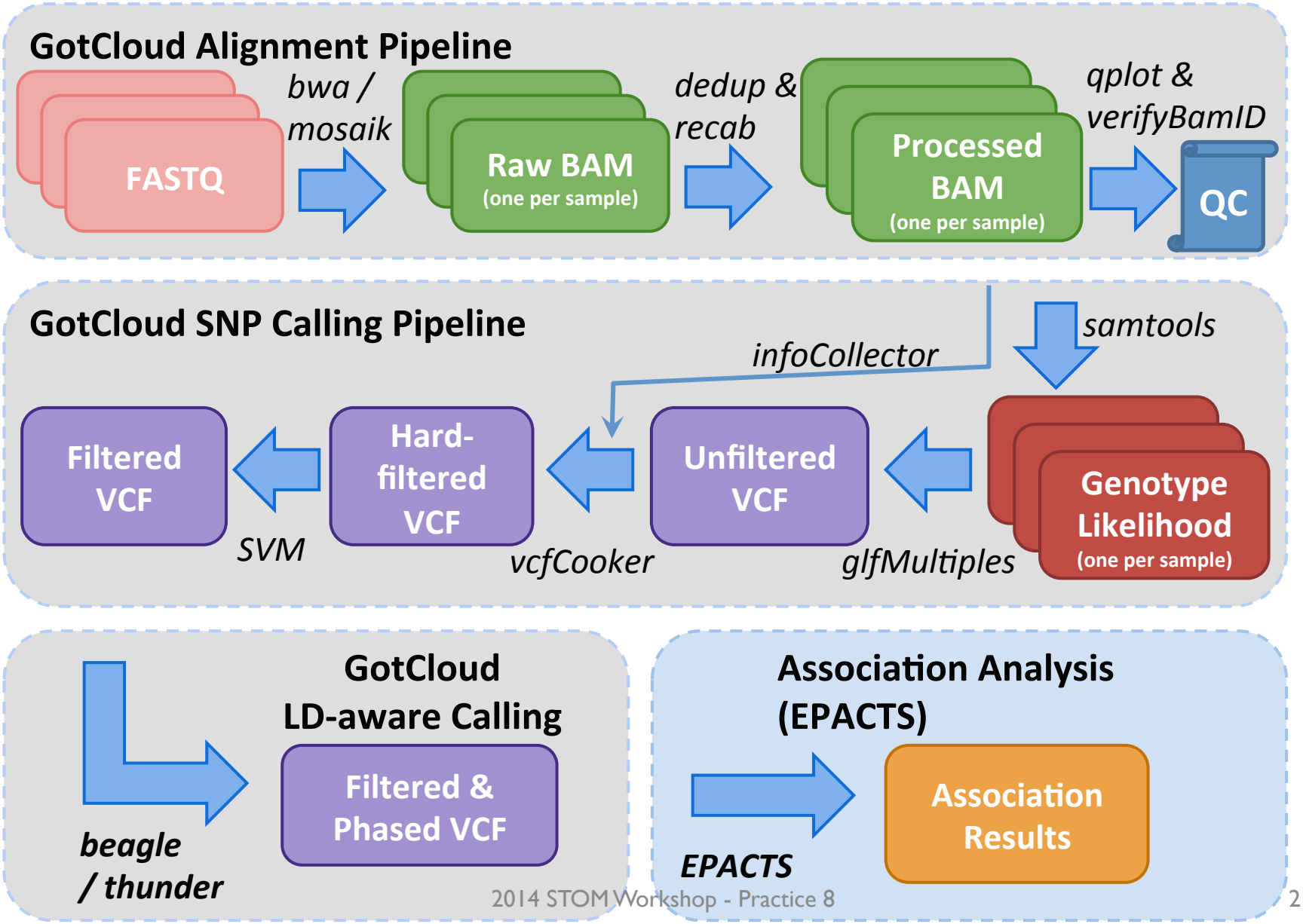
Jan 7th, 2014

STOM 2014 Workshop

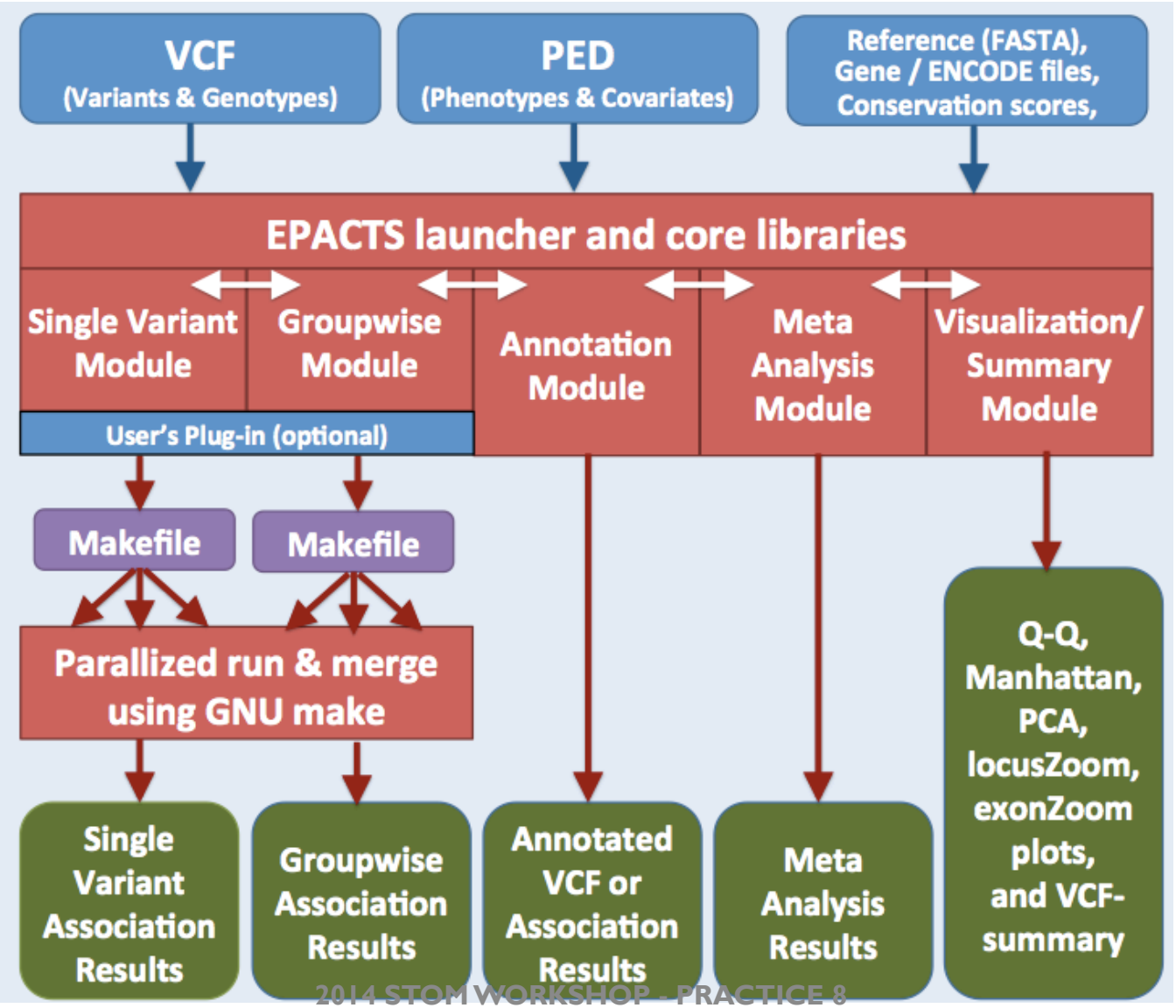
Hyun Min Kang

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”

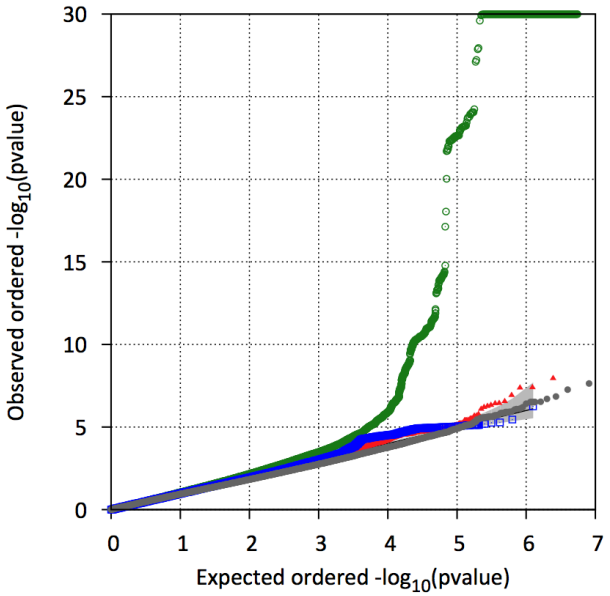
KEY FEATURES OF EPACTS

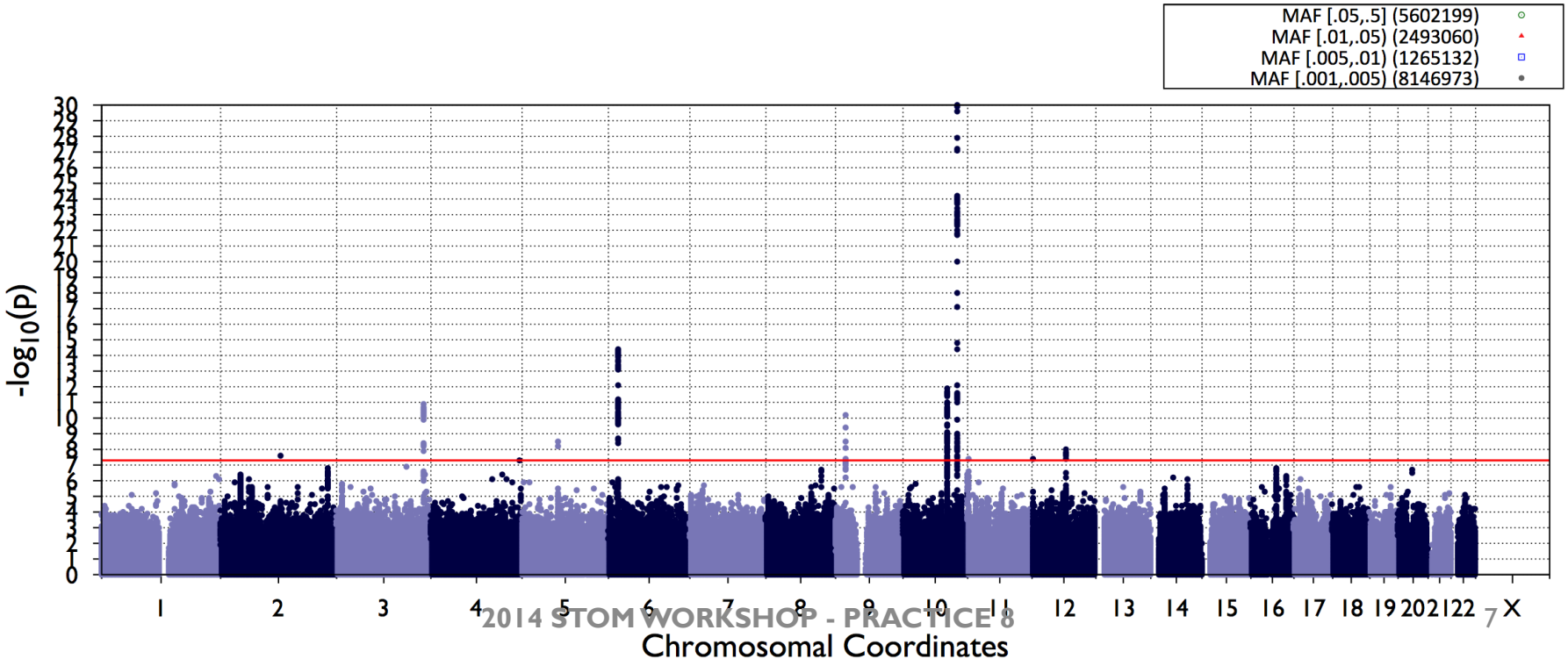
- Convenient and **dynamic plug-in** of user-defined statistical tests
  - Facilitate interaction between method developers and users
- **Efficient and parallel access of VCF** files
- **Fault-tolerant pipeline** structure based on GNU make
- Support of **a variety of single variant and groupwise tests**
- **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

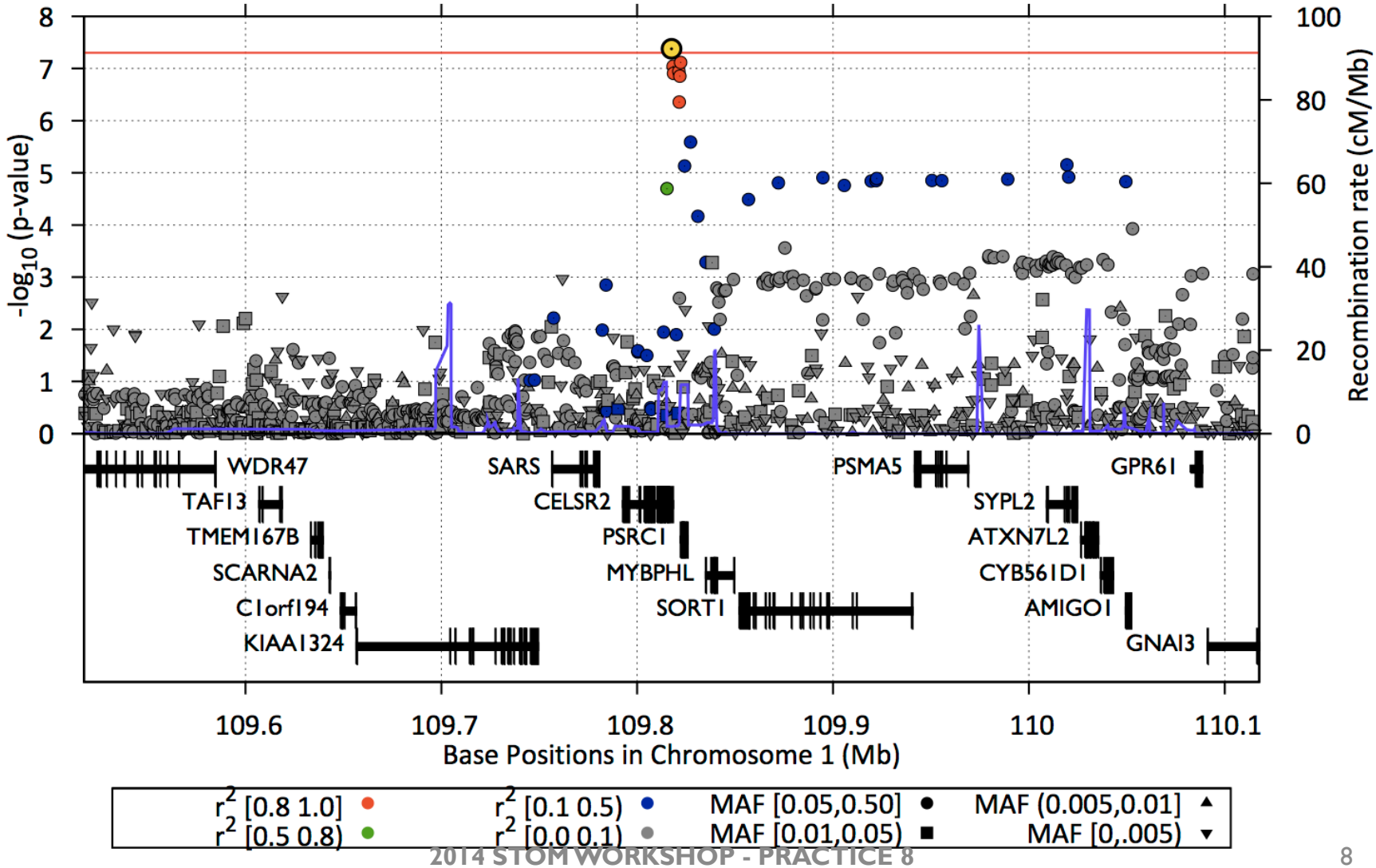
Single Variant Test	Groupwise Test
Wald Test	Collapsing
Score Test	Madsen-Browning*
Likelihood-ratio test	Reverse Regression
Firth bias-corrected LRT	SKAT / SKAT-O
Reverse Regression	VariableThreshold (VT)
Wilcoxon Rank Sum	EMMAX-Collapsing
EMMAX	EMMAX-VT

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

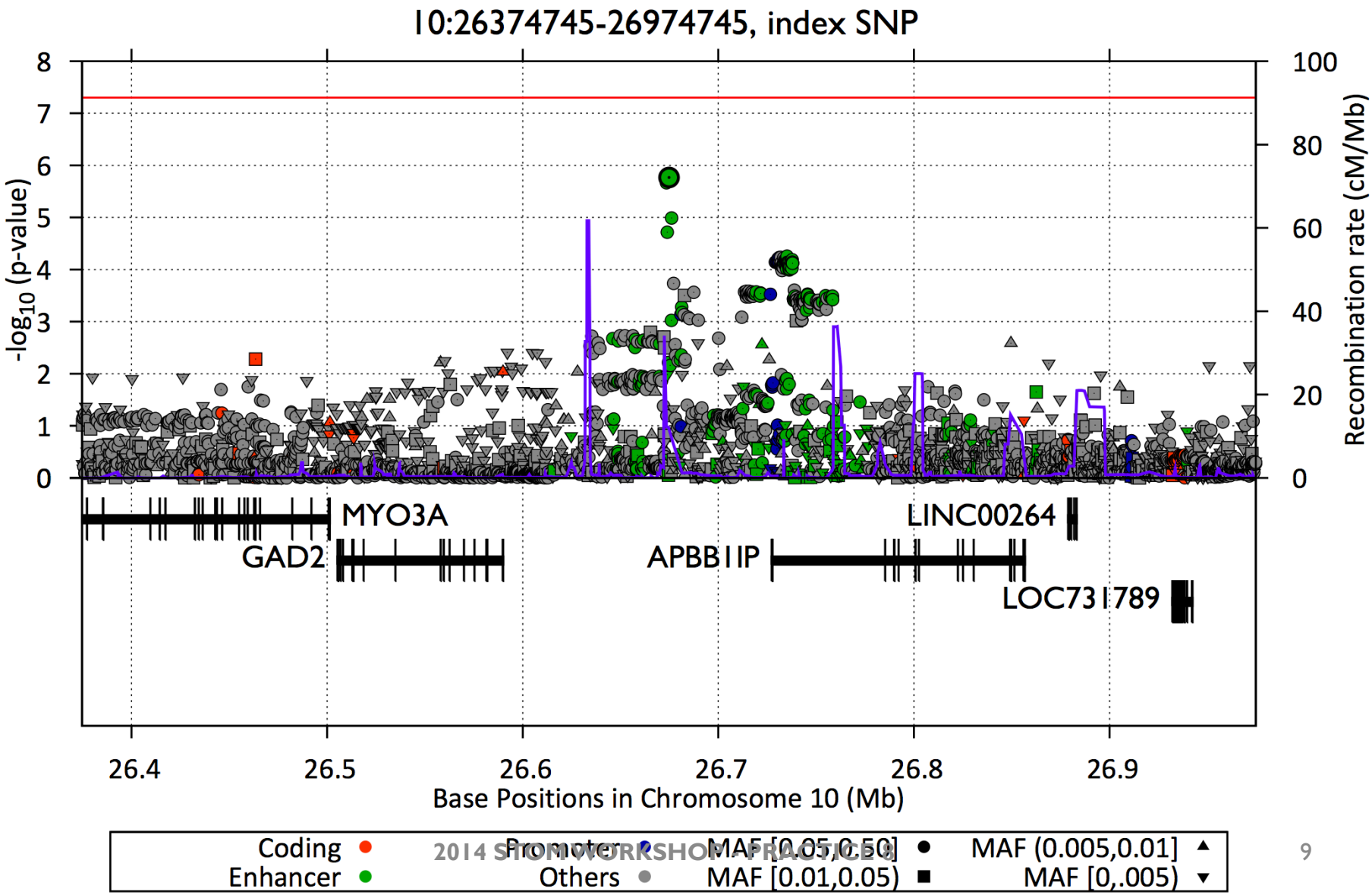




Zoom Plots for top associations



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

SETTING ENVIRONMENTAL VARIABLES AGAIN

- Check if the files are still accessible

  ```
  ls /data/stom2014/session5/
  ```

  – If you see any errors, please let me know now!

- For convenience, let's set some variables. ~/out does not have to be created again

  ```
  export S5=/data/stom2014/session5
  ```

Example of Input PED File

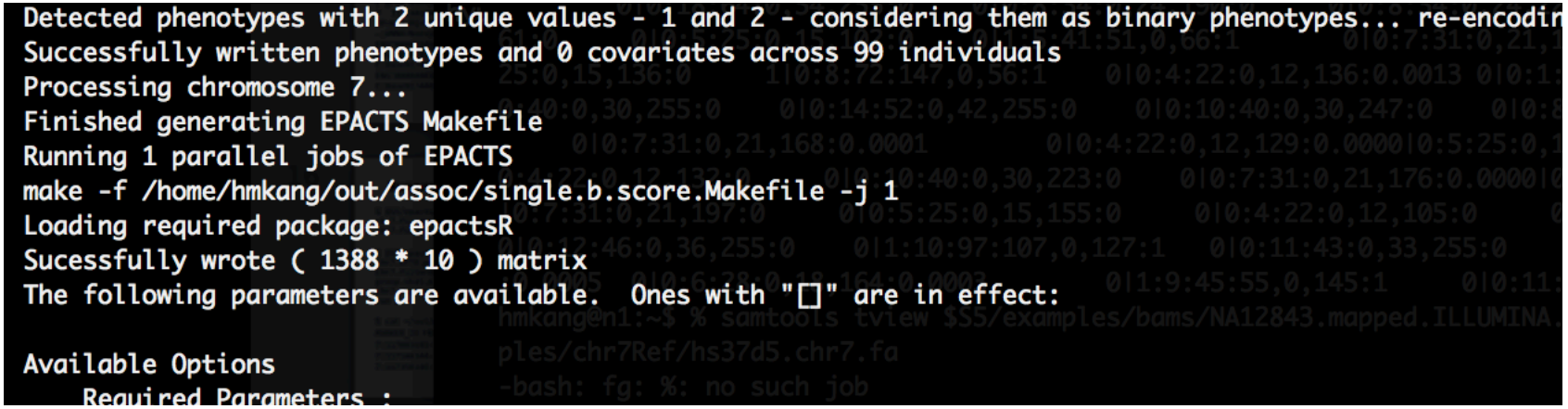
```
% less $S5/examples/index/chr7.CFTR.ped
```

#FAM_ID	IND_ID	DAD_ID	MOM_ID	SEX	PHENO	QT
NA06984	NA06984	0	0	0	1	-0.657149296617419
NA06985	NA06985	0	0	0	2	-0.449034856685281
NA06986	NA06986	0	0	0	1	0.0975849179986626
NA06989	NA06989	0	0	0	2	1.77095069670763
NA06994	NA06994	0	0	0	1	0.287475900193007
NA07000	NA07000	0	0	0	1	-1.36632872407691
NA07037	NA07037	0	0	0	2	-0.1278913321612
NA07048	NA07048	0	0	0	2	-1.45798868745693
NA07051	NA07051	0	0	0	1	-0.599618650565132

```
% ls $S5/out/snps/beagle/chr7/
chr7.filtered.PASS.beagled.vcf.gz
```

RUNNING SINGLE VARIANT ANALYSIS

```
% mkdir ~/out/assoc
% $S5/epacts/bin/epacts single --ped $S5/examples/index/
chr7.CFTR.ped --vcf $S5/out/snps/beagle/chr7/
chr7.filtered.PASS.beagled.vcf.gz --pheno PHENO --out ~/
out/assoc/single.b.score --test b.score --anno --ref $S5/
examples/chr7Ref/hs37d5.chr7.fa --region
7:117000000-117500000 --run 1
```



SINGLE VARIANT ASSOCIATION RESULTS

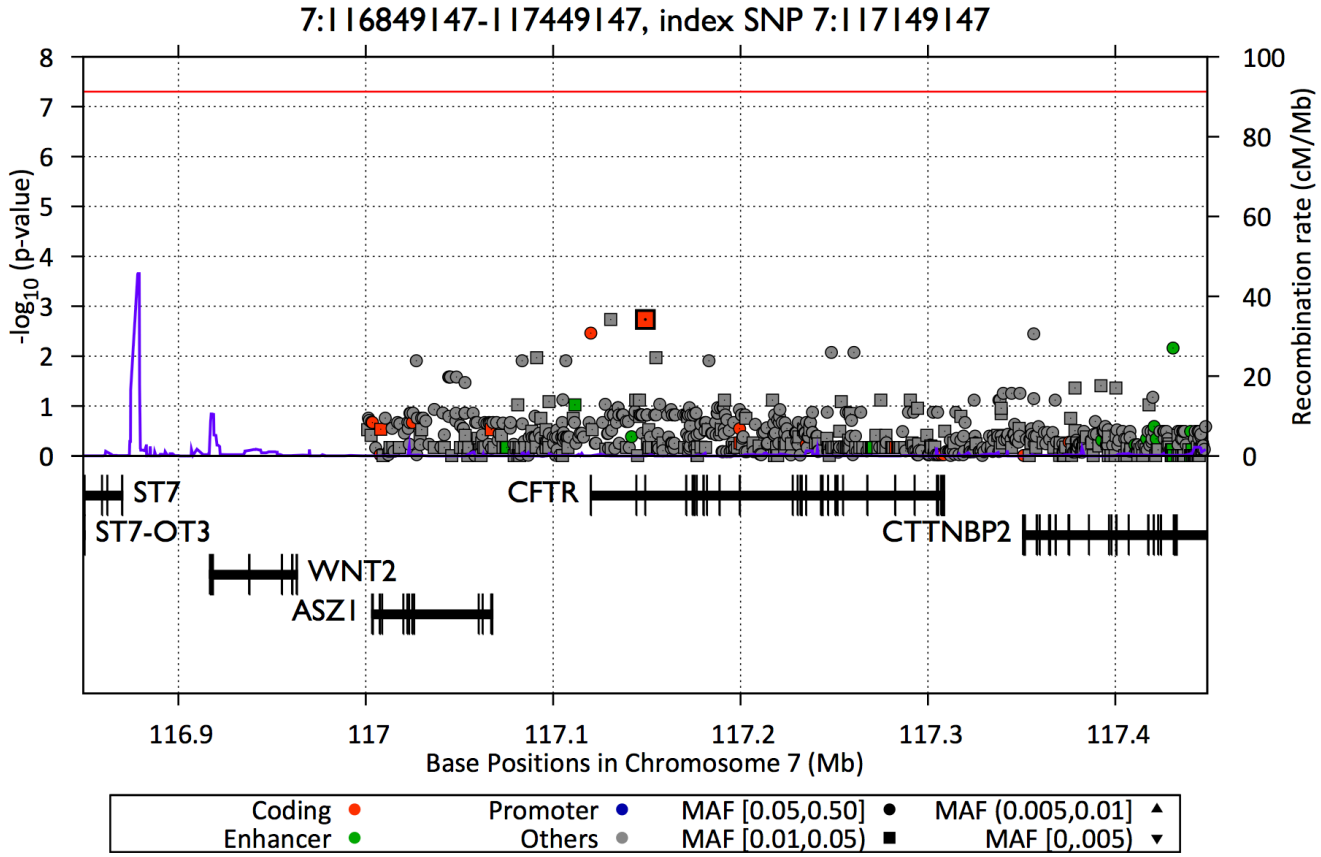
```
% head ~/out/assoc/single.b.score.epacts.top5000
```

#CHROM	BEGIN	END	MARKER_ID	NS	AC	CALLRATE	MAF	PVALUE
7	117130755		117130755	7:117130755_C/G_Intron:CFTR			99	9
7	117149147		117149147	7:117149147_G/A_Nonsynonymous:CFTR			99	
7	117120141		117120141	7:117120141_G/C_Utr5:CFTR			99	14
7	117356483		117356483	7:117356483_C/T_Intron:CTTNBP2			99	11
7	117459023		117459023	7:117459023_A/G_Intron:CTTNBP2			99	121
7	117430909		117430909	7:117430909_T/C_Intron:CTTNBP2			99	10
7	117456904		117456904	7:117456904_C/T_Intron:CTTNBP2			99	115
7	117457141		117457141	7:117457141_G/C_Intron:CTTNBP2			99	115
7	117467003		117467003	7:117467003_T/A_Intron:CTTNBP2			99	122

SCORE	NS.CASE	NS.CTRL	AF.CASE	AF.CTRL				
1	0.045455	0.0018407	3.1148	50	49	0.09	0	
9	1	0.045455	0.0018407	3.1148	50	49	0.09	0
1	0.070707	0.0034417	-2.9253	50	49	0.02	0.12245	
1	0.055556	0.0035695	-2.9139	50	49	0.01	0.10204	
1	0.38889	0.0047674	2.8223	50	49	0.71	0.5102	
1	0.050505	0.0068926	-2.702	50	49	0.01	0.091837	
1	0.41919	0.0076959	2.6651	50	49	0.68	0.47959	
1	0.41919	0.0076959	2.6651	50	49	0.68	0.47959	
1	0.38384	0.0077344	2.6634	50	49	0.71	0.52041	

Creating Zoom Plot

```
$S5/epacts/bin/epacts-zoom --vcf ~/out/snps/
chr7.filtered.PASS.beagled.anno.vcf.gz --pos
7:117149147 --prefix ~/out/assoc/single.b.score --
cellType Gm12878
```



VARIANT ANNOTATION WITH EPACTS

```
% $S5/epacts/bin/epacts anno --in ~/out/snps/beagle/chr7/
chr7.filtered.PASS.beagled.vcf.gz --out ~/out/snps/
chr7.filtered.PASS.beagled.anno.vcf.gz --ref $S5/
examples/chr7Ref/hs37d5.chr7.fa
```

```
% zcat ~/out/snps/chr7.filtered.PASS.beagled.anno.vcf.gz
| grep Nonsynonymous | grep CFTR | cut -f 1-8 | head -1
```

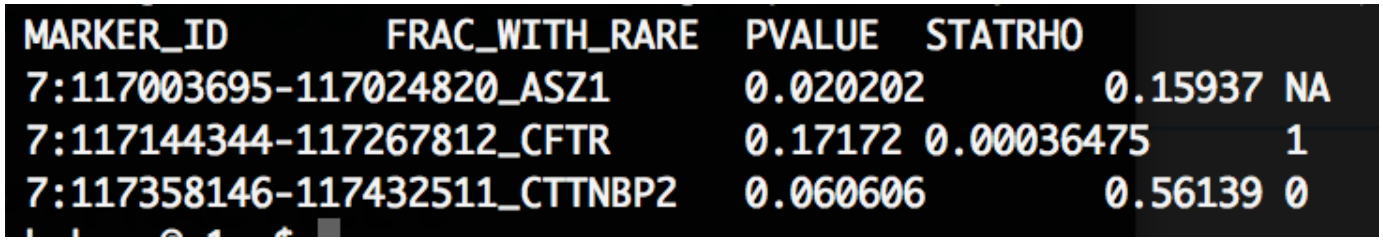
```
7       117144344       .       C       T       100     PASS    DP=689;MQ=58;NS=98;AN=198;AC=1;A
F=0.005323;AB=0.3278;AZ=-1.4260;FIC=-0.0056;SLRT=-0.0056;HWEAF=0.0053;HWDAF=0.0107,0.0000;LBS=0,
0,8,11,0,0,0,0;OBS=0,0,299,368,0,0,7,5;STR=-0.035;STZ=-0.921;CBR=-0.029;CBZ=-0.762;IOR=0.000;IOZ
=-1.018;AOI=-14.680;AOZ=-13.662;LQR=0.027;MQ0=0.000;MQ10=0.000;MQ20=0.000;MQ30=0.031;SVM=1.20661
;BAVGPOST=1.000;BRSQ=0.992;ANNO=Nonsynonymous:CFTR;ANNOFULL=CFTR/ENST00000546407.1:+:Exon:Noncod
ing|CFTR/ENST00000454343.1:+:Nonsynonymous(CGC/Arg/R->TGC/Cys/C:Base92/4260:Codon31/1420:Exon2/2
6):Exon|CFTR/ENST00000003084.6:+:Nonsynonymous(CGC/Arg/R->TGC/Cys/C:Base92/4443:Codon31/1481:Exo
n2/27):Exon
hmkang@n1:~$
```

RUNNING GENE-LEVEL GROUPWISE TEST

```
% $S5/epacts/bin/epacts make-group --vcf ~/out/snps/
chr7.filtered.PASS.beagled.anno.vcf.gz --out ~/out/snps/
chr7.filtered.PASS.beagled.anno.grp --nonsyn

% $S5/epacts/bin/epacts group --ped $S5/examples/index/
chr7.CFTR.ped --vcf ~/out/snps/
chr7.filtered.PASS.beagled.anno.vcf.gz --out ~/out/assoc/
group.skat.o --groupf ~/out/snps/
chr7.filtered.PASS.beagled.anno.grp --test skat --skat-o --
run 2

% cat ~/out/assoc/group.skat.o.epacts | cut -f 4,6,10,11
```



MARKER_ID	FRAC_WITH_RARE	PVALUE	STATRHO
7:117003695-117024820_ASZ1	0.020202	0.15937	NA
7:117144344-117267812_CFTR	0.17172	0.00036475	1
7:117358146-117432511_CTTNBP2	0.060606	0.56139	0

RUNNING GENE-LEVEL GROUPWISE TEST

```
% $S5/epacts/bin/epacts group --ped $S5/examples/
index/chr7.CFTR.ped --vcf ~/out/snps/
chr7.filtered.PASS.beagled.anno.vcf.gz --out ~/out/
assoc/group.VT --groupf ~/out/snps/
chr7.filtered.PASS.beagled.anno.grp --test VT --run 2
```

```
% cat ~/out/assoc/group.VT.epacts | cut -f 4,6,10,13
```

MARKER_ID	PASS_MARKERS	PVALUE	OPT_THRES_RAC
7:117003695-117024820_ASZ1	1	0.44	2
7:117144344-117267812_CFTR	7	2.7e-06	9
7:117358146-117432511_CTTNBP2	4	0.55	1

CREATING YOUR OWN TEST "mytest"

- For single variant test
  Create an R file : **`${EPACTS}/data/single.mytest.R`**

- For groupwise test
  Create an R file : **`${EPACTS}/data/group.mytest.R`**

- When running EPACTS association (single or groupwise), use **`--test mytest`** instead of other known test name

- Refer to existing code to learn the example
  `less $S5/epacts/share/EPACTS/single.q.lm.R`

SUMMARY : EPACTS

- Useful tool for association analysis for sequence data
- Widely used association tests are implemented
  - New tests can be easily plugged in
- Groupwise tests are also available
- Many automated tools available for easier use
  - Manhattan plot, QQ plot, zoom plot, annotation
- New features (partially available)
  - Functional enrichment analysis
  - eQTL analysis module
  - Meta-analysis

THANK YOU FOR LISTENING!