ESP6900 Call Set Overview

March 28th, 2012

Goo Jun, Hyun Min Kang and Gonçalo Abecasis
Mark Rieder and Mark dePristo & teams
Leslie Lange and Paul Auer

Overview of ESP6900 data set

- 6,916 samples included in calling
  - 5,407 QC-PASS samples from ESP5400
  - 1,509 new sampled added for this analysis
  - 70 TB of data coming in, 300GB VCF coming out

- ~100 TB of aligned sequence data
  - Avg. % consensus target bases covered per sample
    - ≥1x : 98%
    - ≥5x : 96%
    - ≥10x : 93%
    - ≥20x : 87%
    - ≥30x : 81%

- ~8 weeks of computational time for variant calling

ESP6900 Sample Distribution

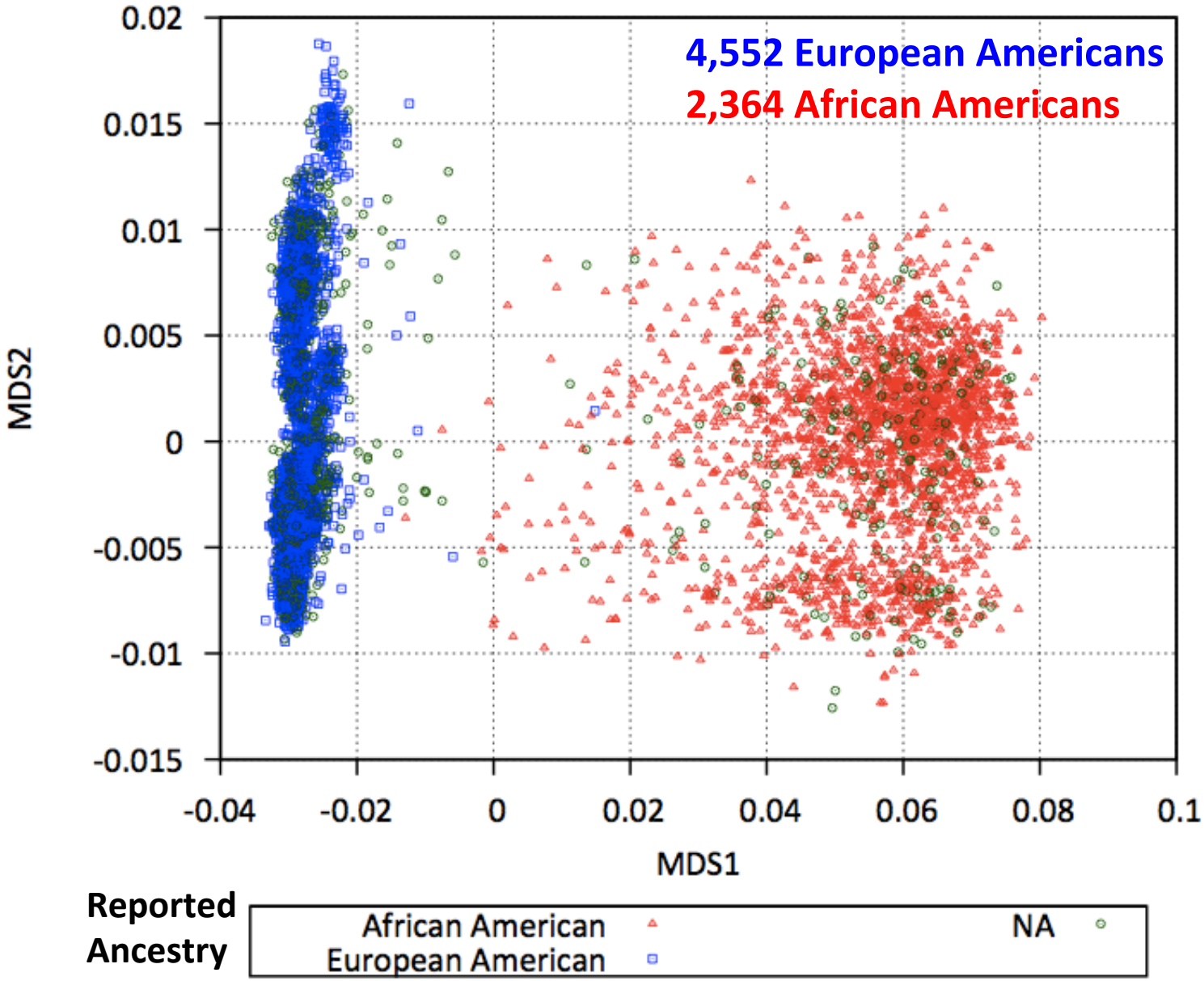
Category	Sample Distribution
By Target	BI (3,110), UW'08 (723), UW'09 (755), UW'10 (2,328)
By Cohort	ARIC(844), CARDIA(206), CHS (215), COPDGene(259), FHS(467), JHS(413), LHS(337), MESA(409), SARP(191), WHI(1,922), *UNKNOWN-TO-US (1,653)*
By Sex	Male (2,667), Female (4,249)
By Population	European (4,552) & African (2,364) Americans

Overview of ESP6900 SNP call sets

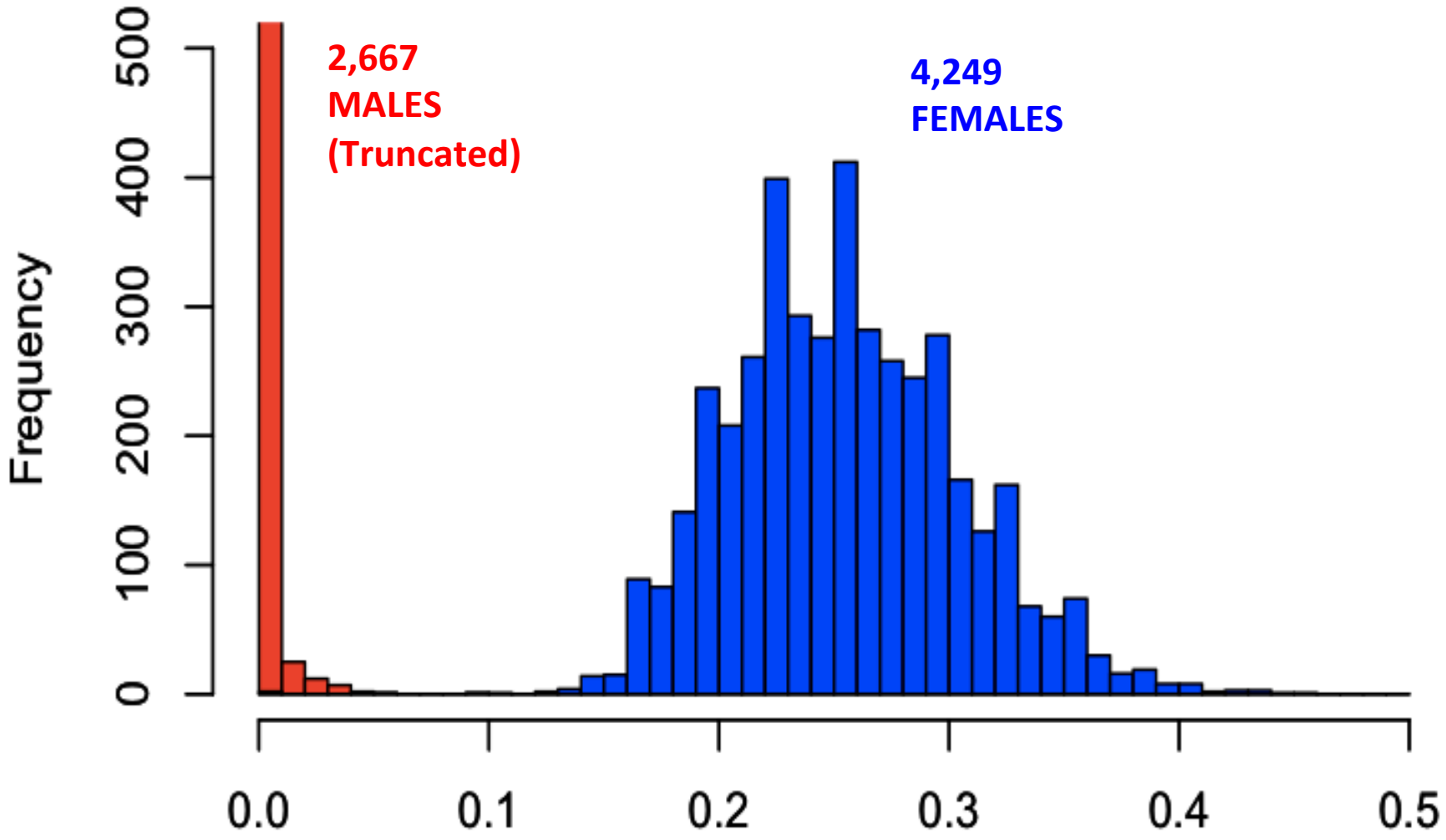
CATEGORY	#SNPs	% in dbSNP 129	% in 1000G Phase1	Known Ts/Tv	Novel Ts/Tv	%HM3 (ASW) Detected
UNFILTERED	2.10M	8.5	25.6	2.82	2.62	97.2
PASS	**1.92M**	**8.6**	**26.7**	**2.94**	**2.83**	**97.0**
SILENT	**433k**	**11.3**	**29.2**	**5.58**	**5.66**	**99.6**
MISSENSE	**729k**	**6.7**	**21.4**	**2.36**	**2.43**	**99.3**
NONSENSE	**19.6k**	**3.0**	**13.2**	**2.13**	**2.23**	**97.6**

- Variants are called within target±50bp
- Filters applied : SVM + nearby indels + strand bias filter for rare variants
- %HM3 sensitivity were evaluated in consensus target ± 50bp
- INDEL calling is in-progress : See Mark DePristo's presentation

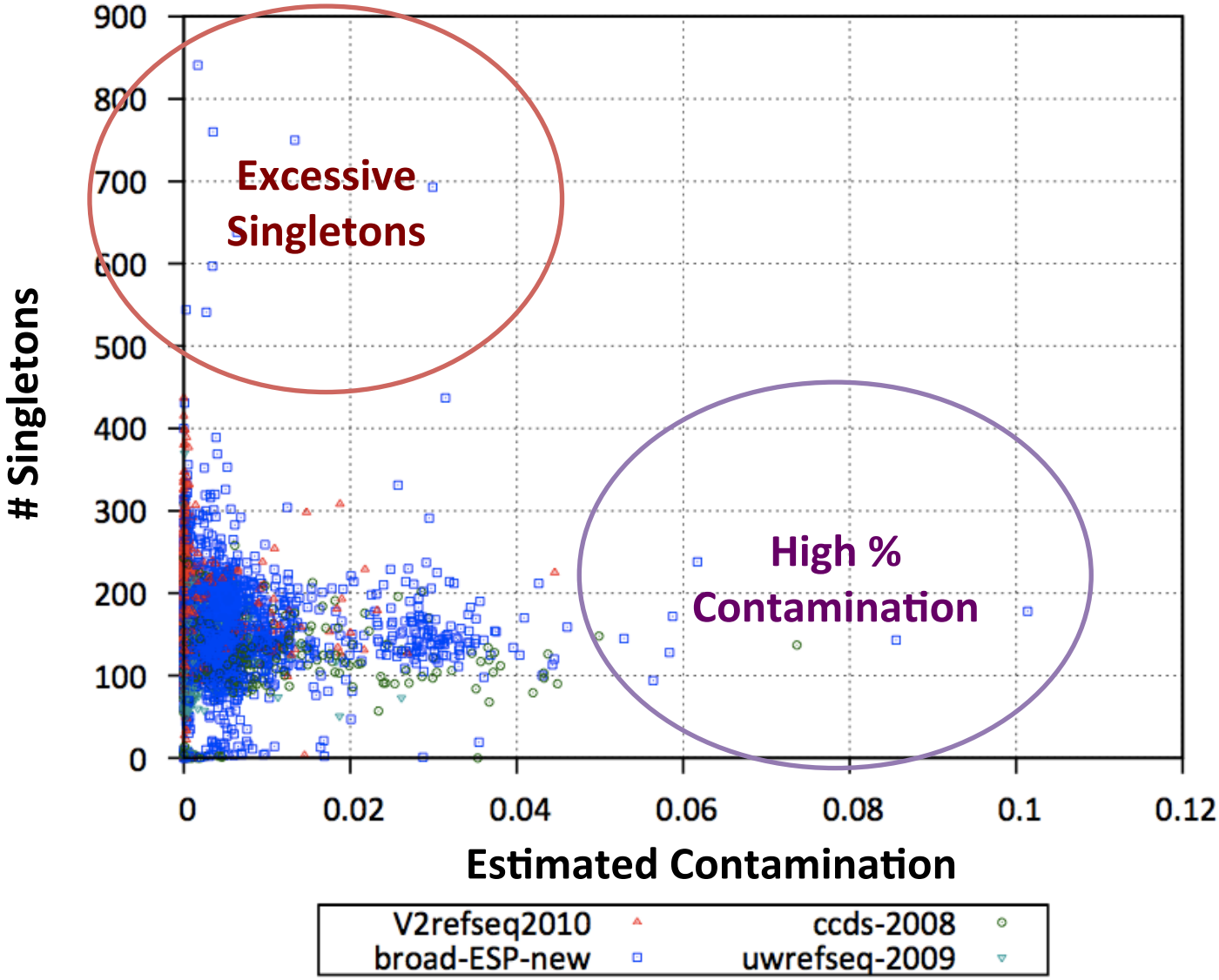
MDS plot shows expected population clusters



Sex is verified by chrX genotype likelihoods



Outliers detected by various QC metrics



Summary of sample level QC

# BEFORE QC	QC Applied	# AFTER QC
6,916	Sex Mismatch (7)	6,909
6,909	Excessive # of singletons (12)	6,897
6,896	>5% contamination (7)	6,890
6,890	>10% missing genotypes (7)	**6,883**
# samples included in the VCF release = 6,883		
6,883	Duplicated Samples (48)	6,883
6,835	Related Samples (95)	**6,740**
Preliminary # samples available for trait analysis = 6,740		

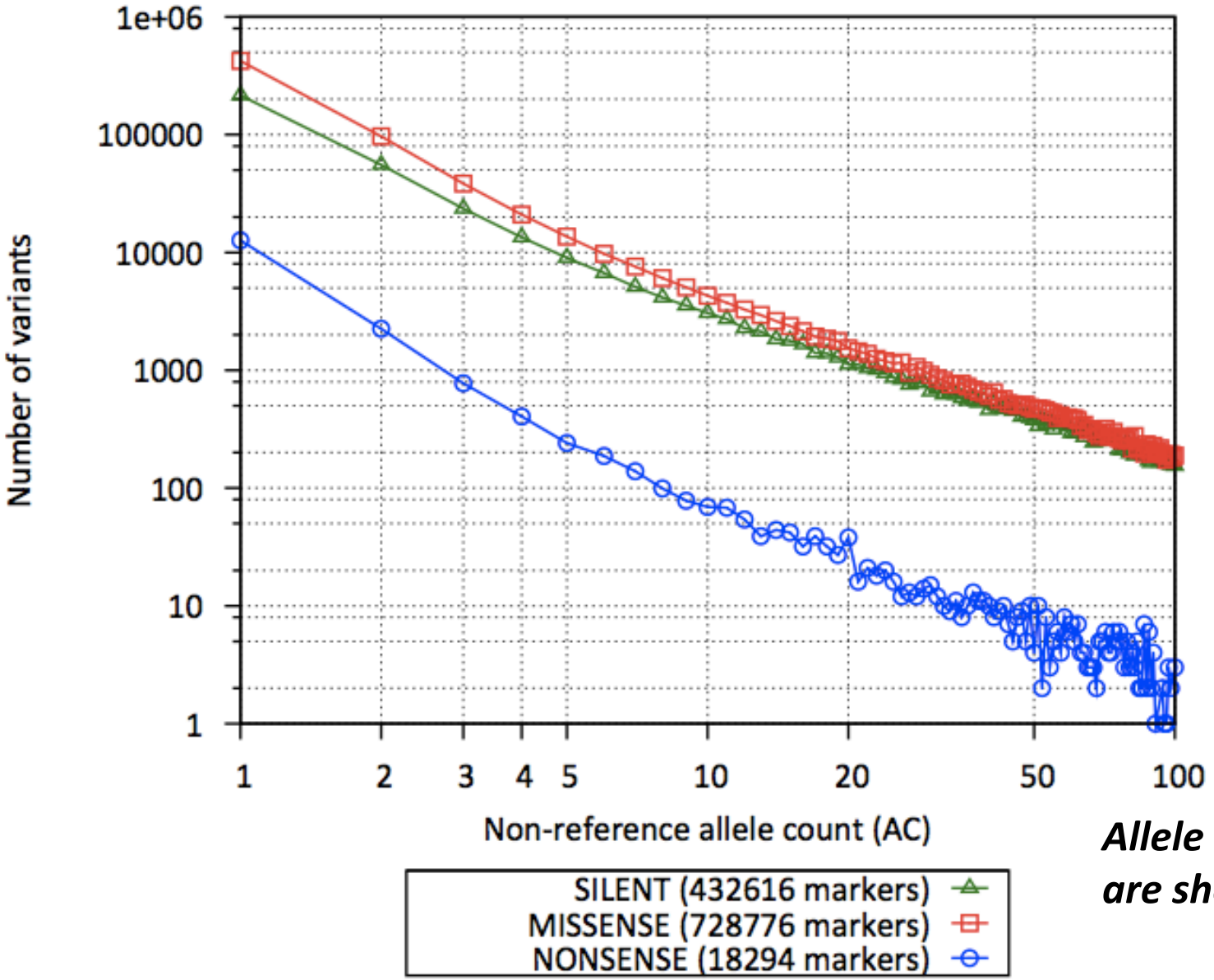
Most Variants Are Rare

CATEGORY	#SNPs	SINGLE-TONs	DOUBLE-TONs	AF<0.1%	AF<0.5%	AF<5%
PASS	1.92M	53%	13%	82%	89%	96%
SILENT	433k	50%	13%	80%	88%	95%
MISSENSE	729k	58%	13%	87%	93%	97%
NONSENSE	19.6k	69%	12%	93%	97%	99%

 # Singletons per individual :

29 SILENT, 59 MISSENSE, 2 NONSENSE

Allele Frequency Spectrum



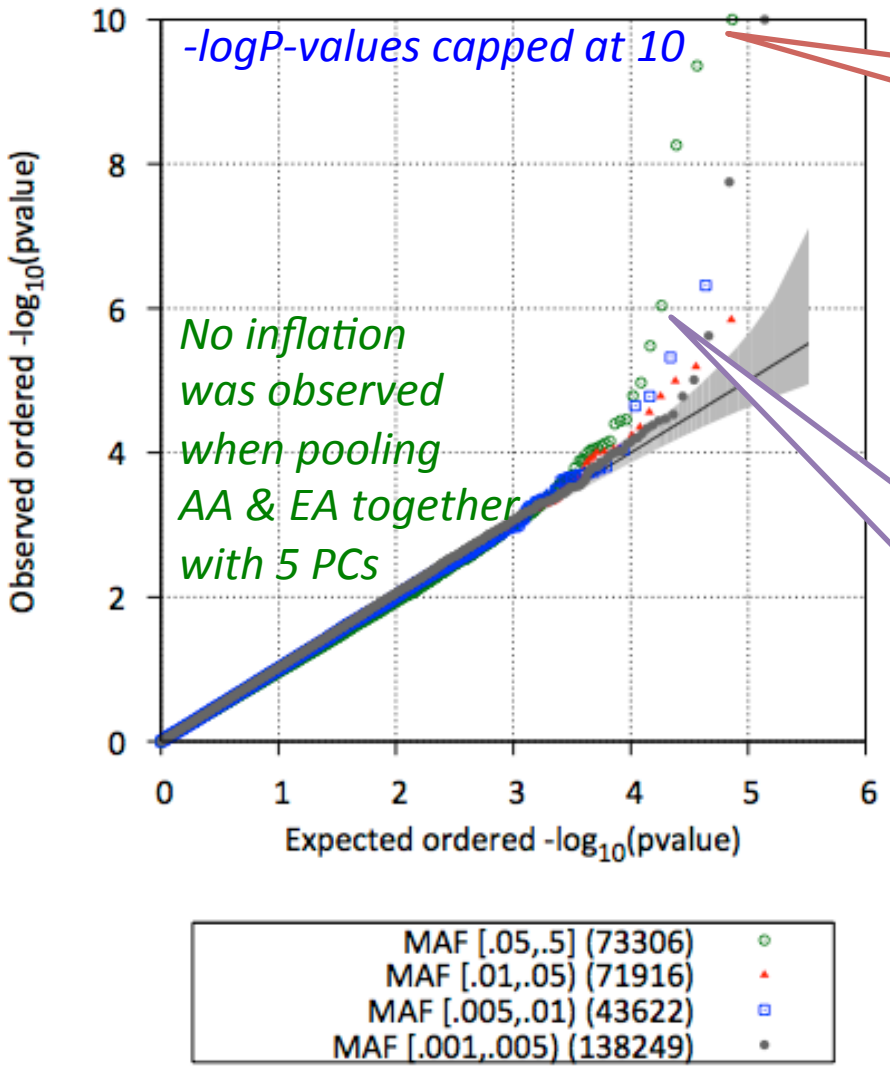
Allele count ≤ 100 are shown

How many variants per gene?

CATEGORY	SILENT	MISSENSE	NONSENSE
AVG # SNPs per GENE	23.5	39.3	2.0
% GENES with ≥ 1 SNPs	98.9%	99.6%	48.7%
% GENES with ≥ 5 SNPs	91.5%	95.3%	3.1%
% GENES with ≥ 10 SNPs	74.7%	87.0%	0.1%
% GENES with ≥ 20 SNPs	42.6%	67.0%	0.0%
% GENES with ≥ 50 SNPs	9.0%	24.5%	0.0%
% GENES with ≥ 100 SNPs	1.7%	5.7%	0.0%

- Gene length and mutation rate plays major role in statistical power given effect size
- Results are based on GenCodeV7 annotations

Preliminary single variant analysis on LDL

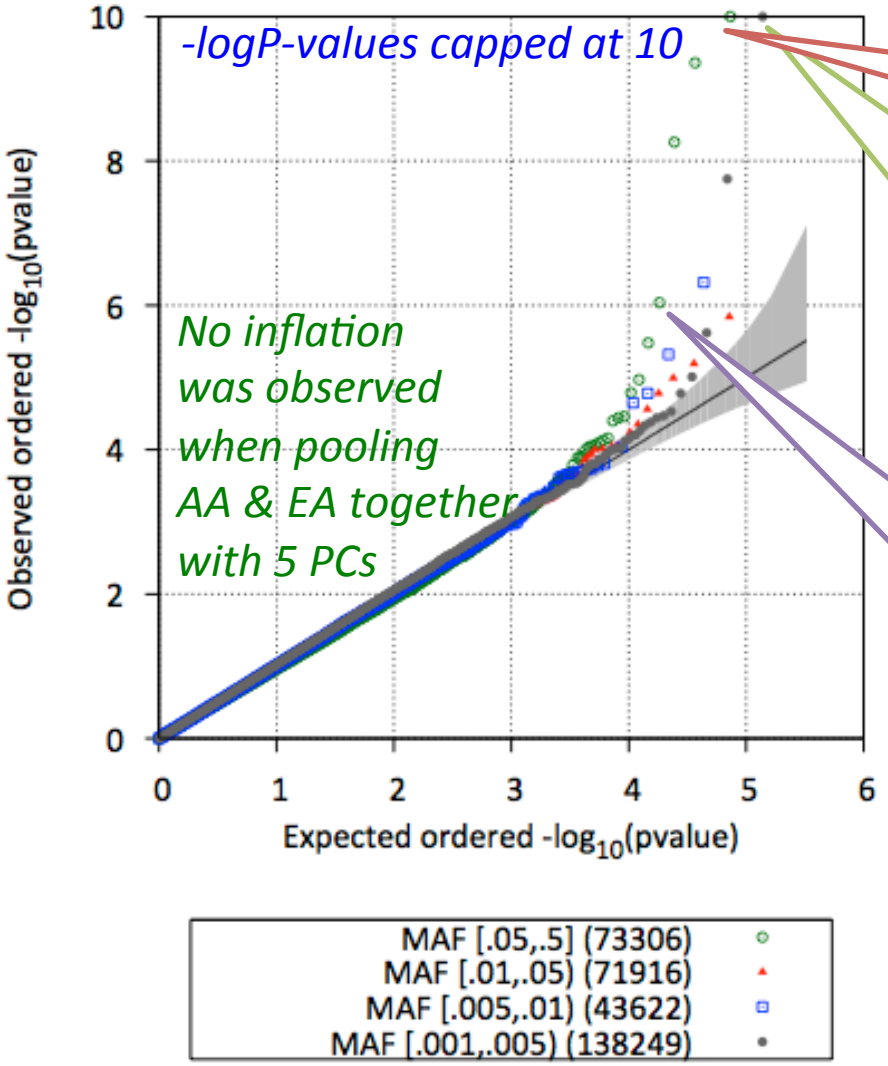


TOMM40 (near APOE)
Synonymous SNP
MAF 0.065
p-value : 4.2×10^{-24}

APOB
Missense SNP
MAF 0.21
p-value : 9.0×10^{-7}

N=3,342 individuals with phenotypes

Preliminary single variant analysis on LDL



TOMM40 (near APOE)
Synonymous SNP
MAF 0.065
p-value : 4.2×10^{-24}

PCSK9

Nonsense SNP 1	**Nonsense SNP 2**
MAF 0.005	**MAF 0.002**
p-value : 7.2×10^{-11}	**p-value : 1.8×10^{-8}**

APOB
Missense SNP
MAF 0.21
p-value : 9.0×10^{-7}

N=3,342 individuals with phenotypes

Summary

- ESP6900 SNP call sets are produced
  - Harmonized across different centers
  - Vast majority of them are novel rare variants
  - ~730k missense, ~430k silent, ~20k nonsense SNPs

- Rare variant association with larger sample size
  - Signals for known SNPs increases by orders of magnitude
  - Nonsense SNPs at MAF<.5% are genome-wide significant
  - Integrated and harmonized phenotype analyses likely improve statistical power

Acknowledgements

- @Michigan
  - Goo Jun
  - Hyun Min Kang
  - Gonçalo Abecasis
  - Youna Hu
  - Cristen Willer
  - Chenyi Xue
- @Broad
  - Mark DePristo
  - Stacey Gabriel
- @UW
  - Mark Rieder
  - Debbie Nickerson
- Leslie Lange @ UNC
- Paul Auer @ FHCRC
- Suzanne Leal @ BCM