# Calling multiple populations simultaneously

Mark A. DePristo, Ryan Poplin, and Eric Banks

Manager, Medical and Population Genetics Analysis
Medical and Population Genetics Program
Broad Institute of Harvard and MIT
Sept 14, 2010

# Motivation

- Calling multiple population together should allow us to
  - Better discover low-frequency variants within a population that are shared among populations
  - Better distinguish true variant from machine error with multi-sample error covariates (strand bias) that we use in the GATK
- Some potential cost in very rare variants as base error rates approach variant frequency among chromosomes

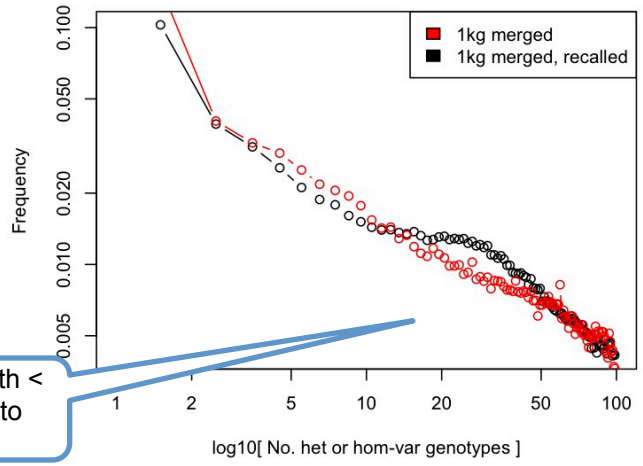
# Two parts to this presentation

- Part 1: recalling pilot variant sites in all populations simultaneously
  - How does this affect the AFS?
  - How many population-specific sites (CEU) have evidence for variation in the other populations?
- Part 2: calling all samples simultaneously in production phase
  - Contrasting sites called using only EUR samples with sites subsetted to EUR samples called in all samples simultaneously

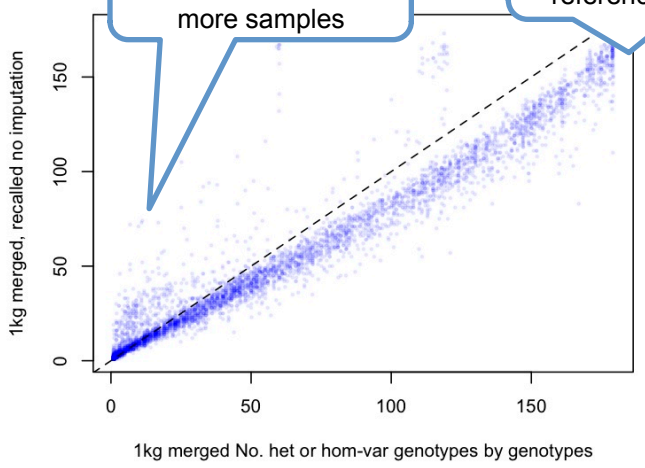
# Part 1

- Read-backed genotyping of pilot low-coverage sites
  - Took the released VCFs for CEU, YRI, and CHB/JPT and merged them into a single VCF file
    - No-call genotypes for sites present in one population and not the other
    - Updated AC,AN annotations
    - ~15M sites
  - Recalled at all ~15M sites with the GATK using merged pilot BAM files that have been locally realigned
    - No genotype refinement with Beagle
    - Genotypes assigned purely on basis of reads, no HW
    - GATK genotype likelihoods for each sample annotated in VCF
  - Data sets generated:
  - Three populations – all 179 samples called together
  - CEU only – using only CEU BAM, for comparison purposes
  - Analyzed chromosome 20, but genome-wide data set released
    - http://www.broadinstitute.org/gsa/wiki/index.php/GSA_FTP_Server
    - Directory pilot1GLAllPops

# Non-reference samples: naïve merging vs. recalling

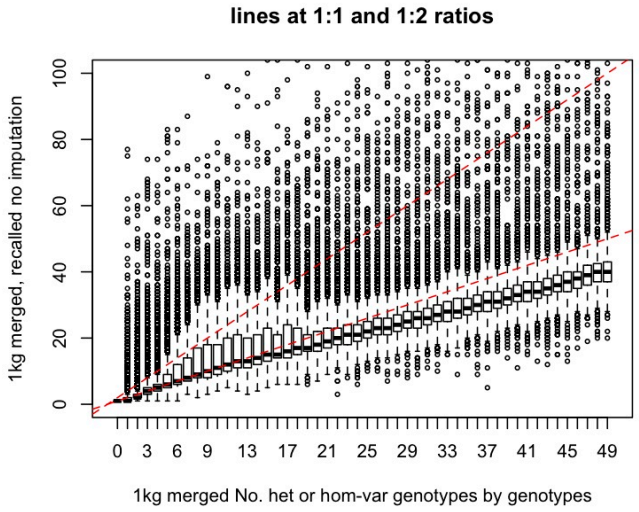


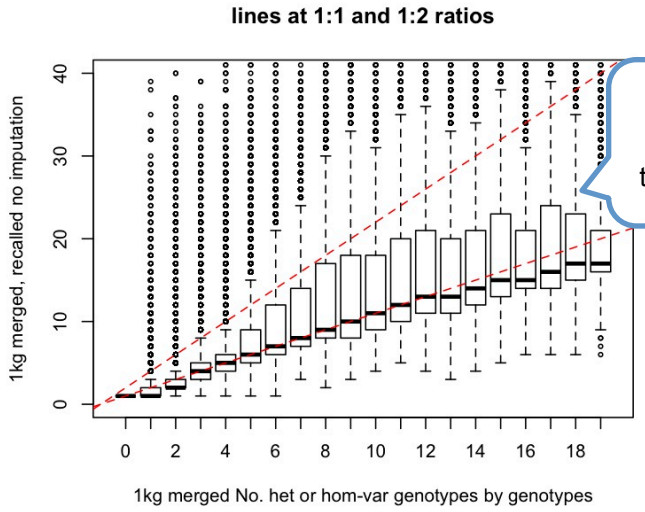

Shift from sites with < 10 NR samples to 20-50 range




Many low-frequency variants found in more samples

Raw reads and sampling underestimate non-reference genotype count

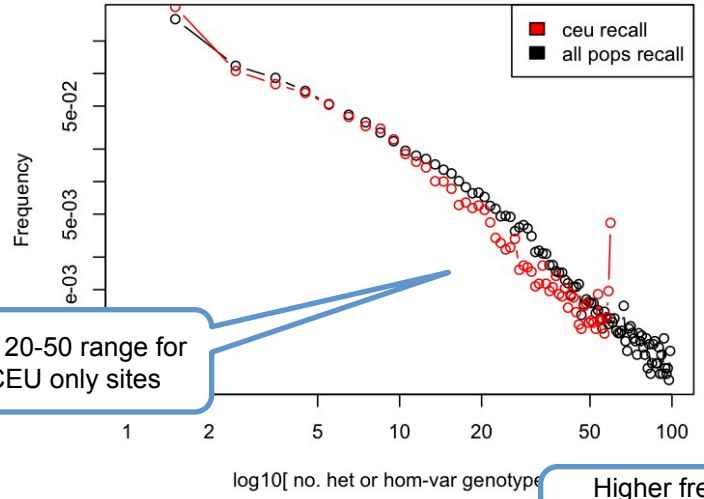


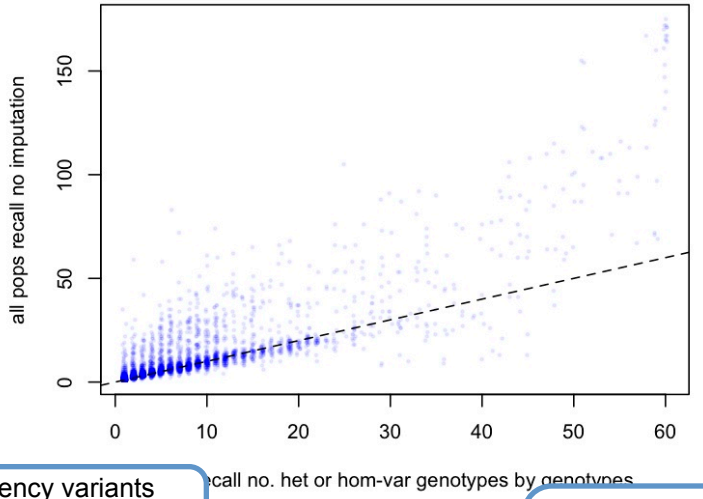

Systematic underestimate of true AF at lower AF

# Systematically finding more samples with non-reference alleles in other populations

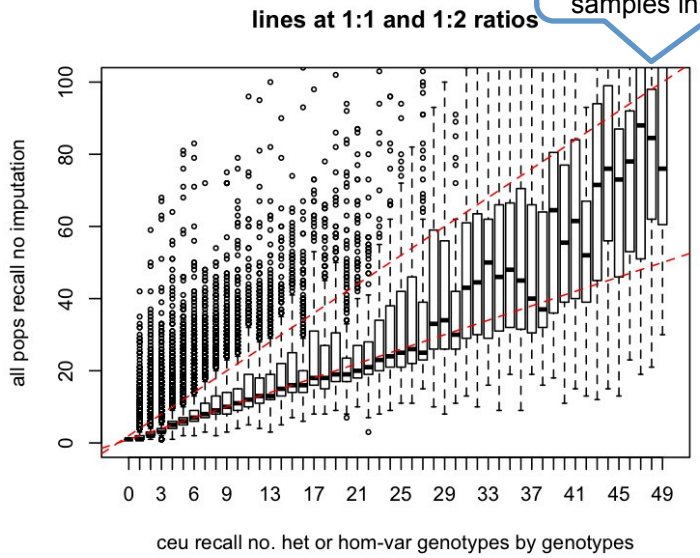
Only sites called in CEU pop. in the pilot (not in YRI or CHB/JPT)



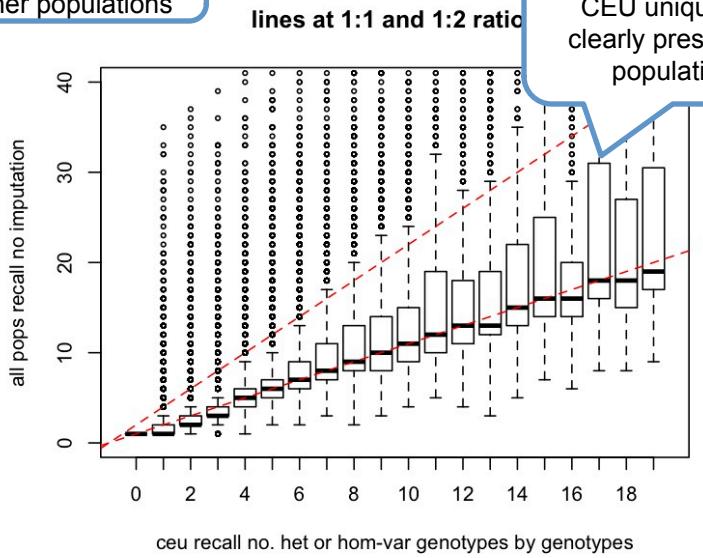

Similar shift to 20-50 range for at recalled CEU only sites




Higher frequency variants more likely to be found in samples in other populations




Many even low frequency CEU unique variants clearly present in other population data



# Part 1: Conclusions

- Many low frequency variants are also found in other populations
- Sites specific to one population often contain evidence for alternate alleles in the other population sequencing data
- Proposal: call all populations simultaneously
  - Without LD followed by global or population-specific LD-based genotype refinement
  - By population, then recall simultaneously in all samples, followed by LD refinement
- Suggestion: DCC should release merged BAMs by population by chromosome
  - Will improve performance of analysis

# Part 2

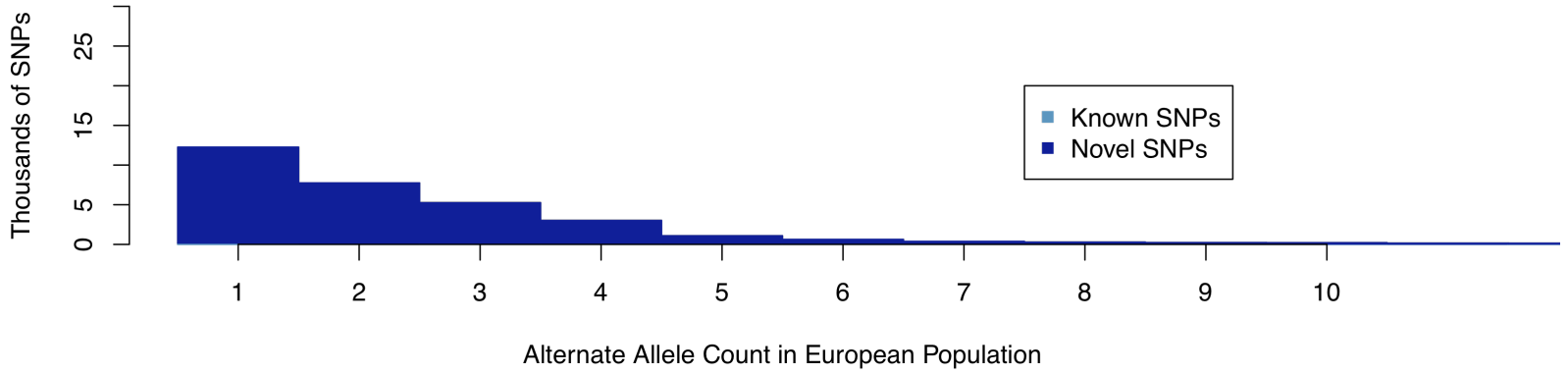
- Calling all populations simultaneously
  - Chr20
  - June BAM release from DCC, problematic SOLiD
  - GATK SNP calls with variant quality recalibration
  - No imputation for genotype refinement
- Called EUR samples (200) together
- Called all samples (538) together, and subsetted to EUR samples, considering sites variant if at least one EUR sample was non-reference
- Union of these two call sets
- Evaluating all three call sets

# Calling with all populations adds high quality novel variants

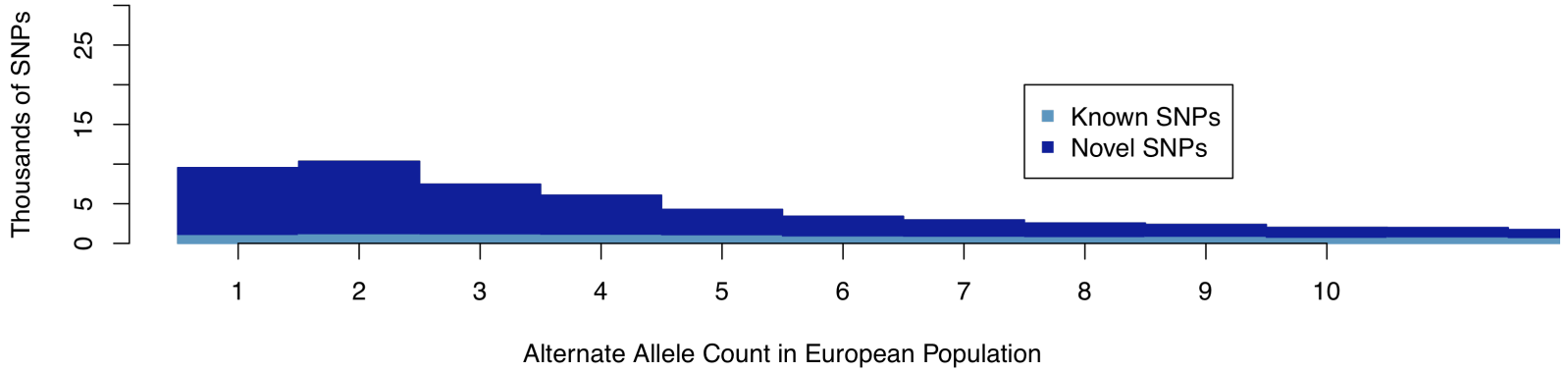
| No. | Callset (subset to EUR) | Total # variants | dbSNP % | # knowns | Known ti/tv | # novels | Novel ti/tv |
|---|---|---|---|---|---|---|---|
| 1 | EUR only | 260,441 | 52.13 | 135,759 | 2.36 | 124,682 | 2.07 |
| 2 | All populations | 251,251 | 41.30 | 103,769 | 2.36 | 147,482 | 2.12 |
| | Calls unique to 2 from 1 | | | 5,759 | 2.70 | 36,788 | 1.96 |
| 3 | EUR-only and all populations call sets merged | 348,372 | 41.56 | 144,790 | 2.36 | 203,582 | 2.02 |
| | Calls unique to 3 from 1 | | | 9,126 | 2.71 | 79,352 | 1.94 |

# Calling all populations discovers very low-frequency variants but at the cost of variants

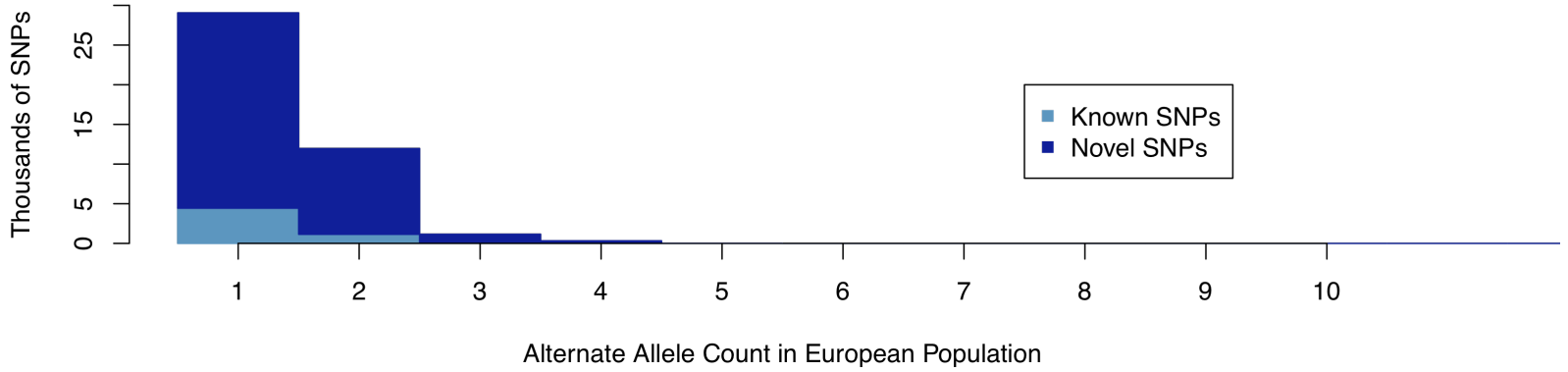
Euro. only disjoint set




Intersection




All pops. disjoint set



# Part 2 conclusions

- Calling multiple populations simultaneously discovers many more high-quality variants in each population
  - On chr20, finds 80K more novel variants at 1.94 Ti/Tv in addition to the 125K found by calling the EUR samples alone
  - Negative impact on population-specific calls can be ameliorated by merging global calls with population-specific calls
- Additional benefits
  - More straightforward Fst calculations
  - Cleaner resolution of admixed populations
- Preliminary results that can be improved
- Impact on imputation for genotype refinement unclear

# Appendix

# Side-note: merged BAMs by population

- At BI we keep merged BAM files by chromosome for each population
  - 59/60 samples in a single BAM
  - 120Gb for chr1 of CEU
- Reduces I/O burden for analysis
  - Recalling with the GATK took ~24 hours total running each chromosome in parallel, dynamically merging three population BAMs
- Can easily call all samples simultaneously merging 25 BAM files, 1 for each of the 25 populations