BROAD INSTITUTE

1KG August Release, chr20 call sets

Ryan Poplin (rpoplin@broadinstitute.org)
Genome Sequencing and Analysis
Medical and Population Genetics
October 5, 2010

BROAD INSTITUTE

Data and Definitions -- Samples

- Production phase 1KG, August release, chr20
- Samples selected for August analysis
- 174 AFR = 78 YRI + 67 LWK + 24 ASW + 5 PUR
- 283 EUR = 90 CEU + 92 TSI + 43 GBR + 36 FIN + 17 MXL + 5 PUR
- 194 ASN = 68 CHB + 25 CHS + 84 JPT + 17 MXL

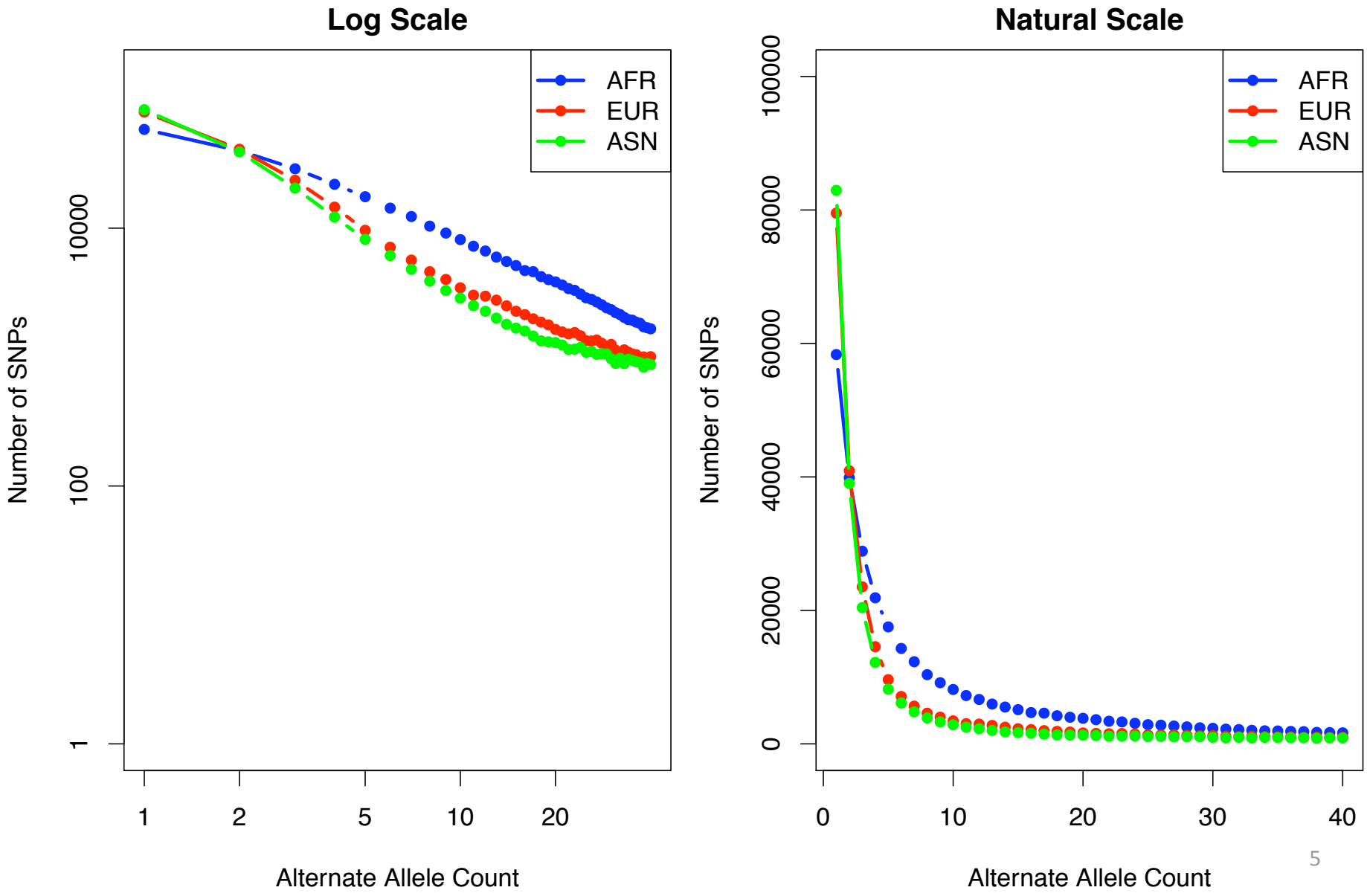
Data and Definitions -- Pipeline

- Sample-level cleaning at known indels
- Filtering SNPs within 10bp of pilot Dindel calls
- BAQ calculation from Heng Li
- Called all samples simultaneously (629)
- Variant quality score recalibration
- Subsetted to analysis panels, considering a site variant if at least one sample in the panel was non-reference
- No imputation or genotype refinement

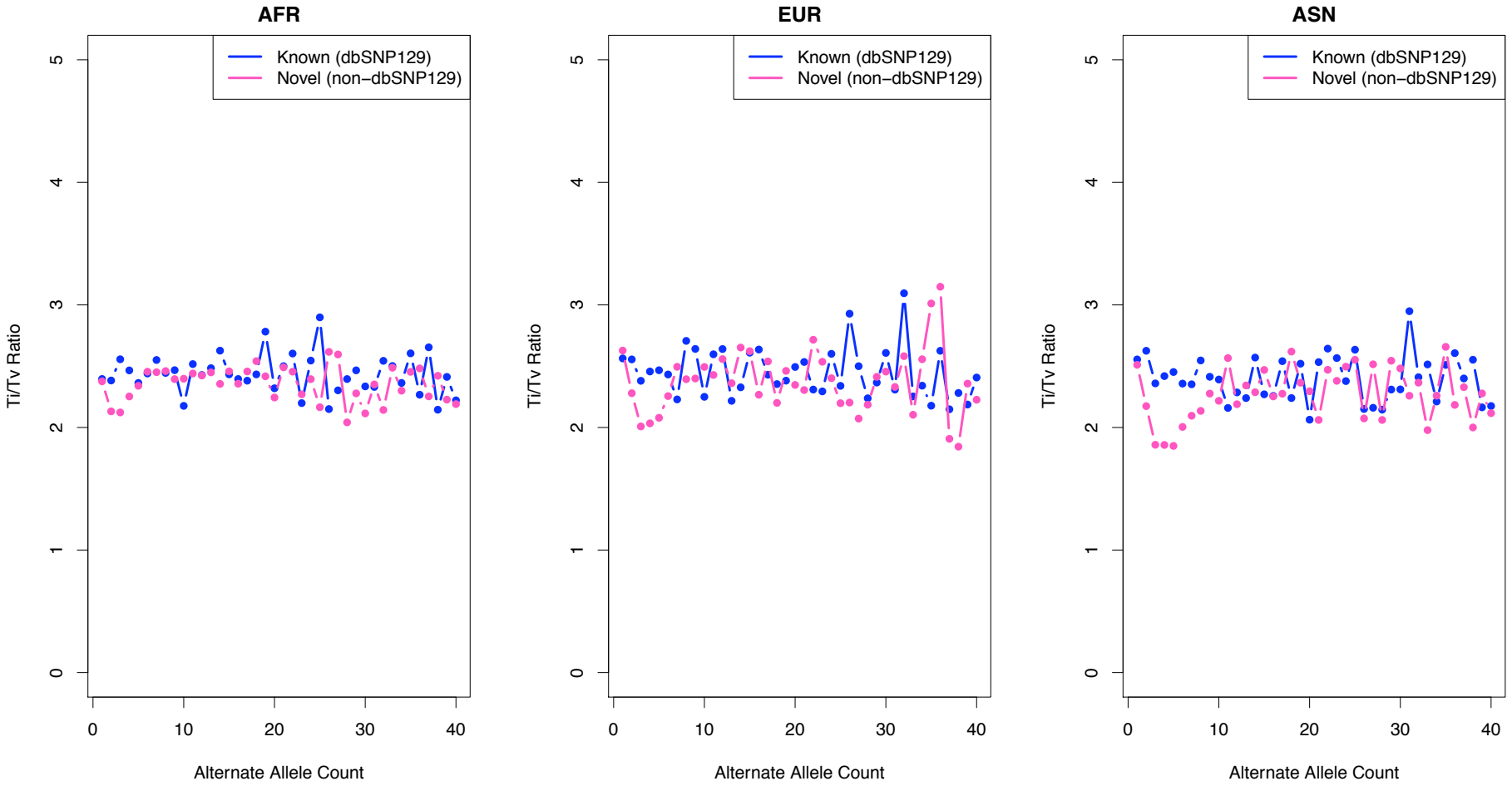
Calling all populations simultaneously yields large number of high-quality variants

# samples	Call set	Total # variants	dbSNP %	# knowns	Known ti/tv	# novels	Novel ti/tv
629	All Populations (EUR+AFR+ASN)	547,166	33.47	183,138	2.40	364,028	2.30
174	Subset to AFR (YRI+LWK+ASW+PUR)	436,830	40.03	174,854	2.40	261,976	2.31
283	Subset to EUR (CEU+TSI+GBR+FIN +MXL+PUR)	355,892	45.09	160,467	2.41	195,425	2.35
194	Subset to ASN (CHB+CHS+JPT+MXL)	321,450	47.20	151,742	2.41	169,708	2.23

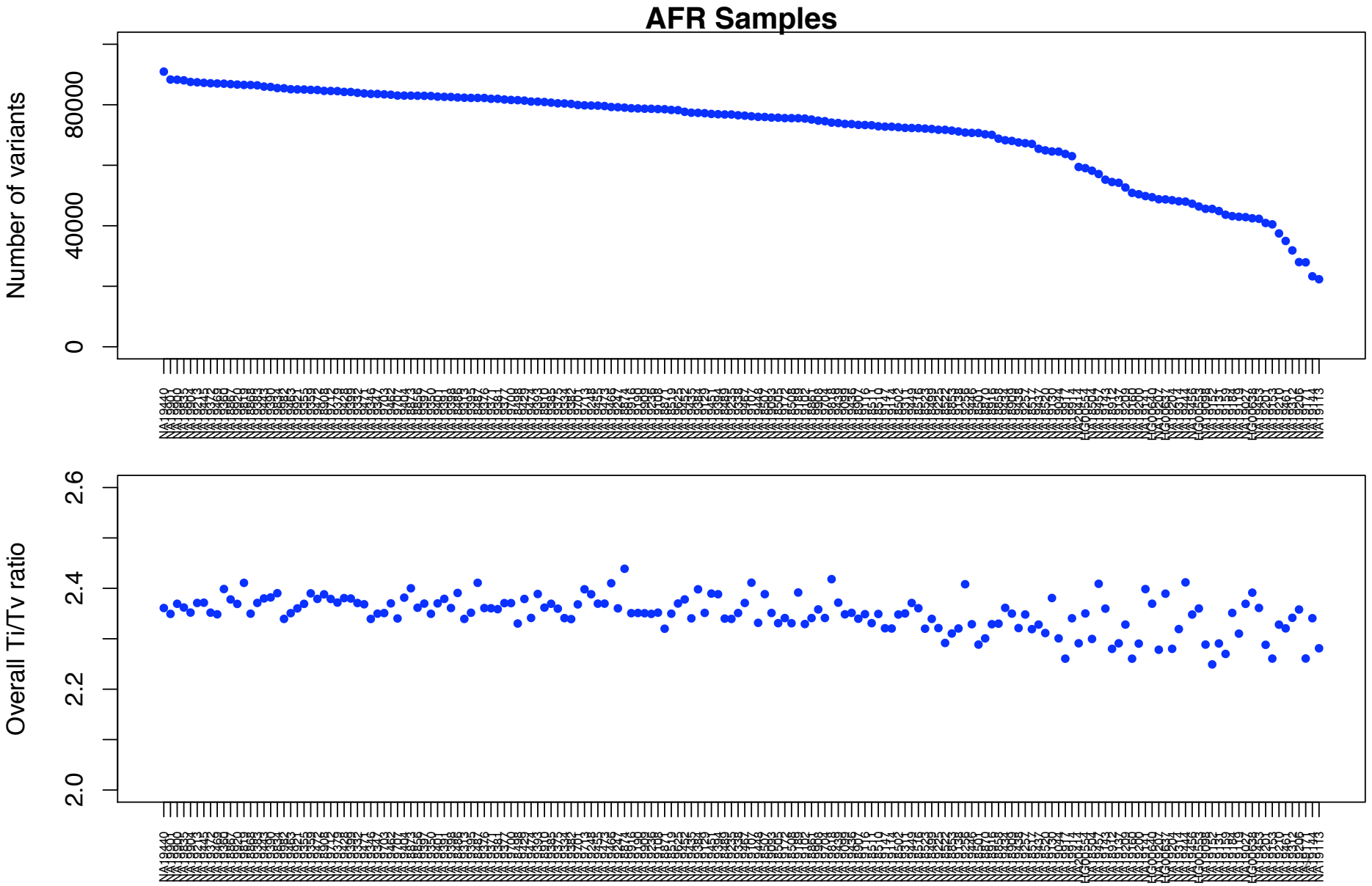
Allele frequency distribution follows expectation



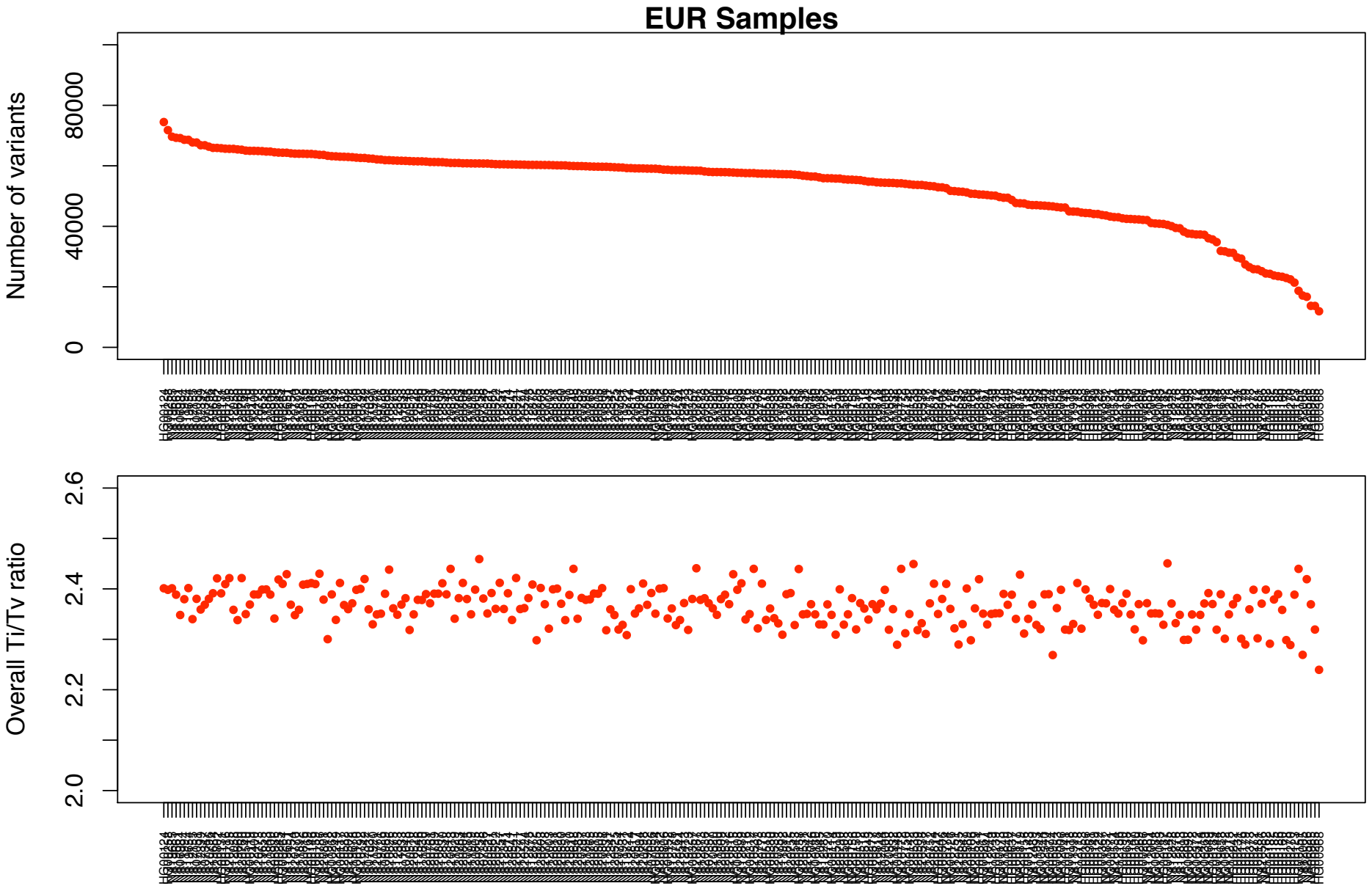
Ti/Tv ratio looks great across all allele counts



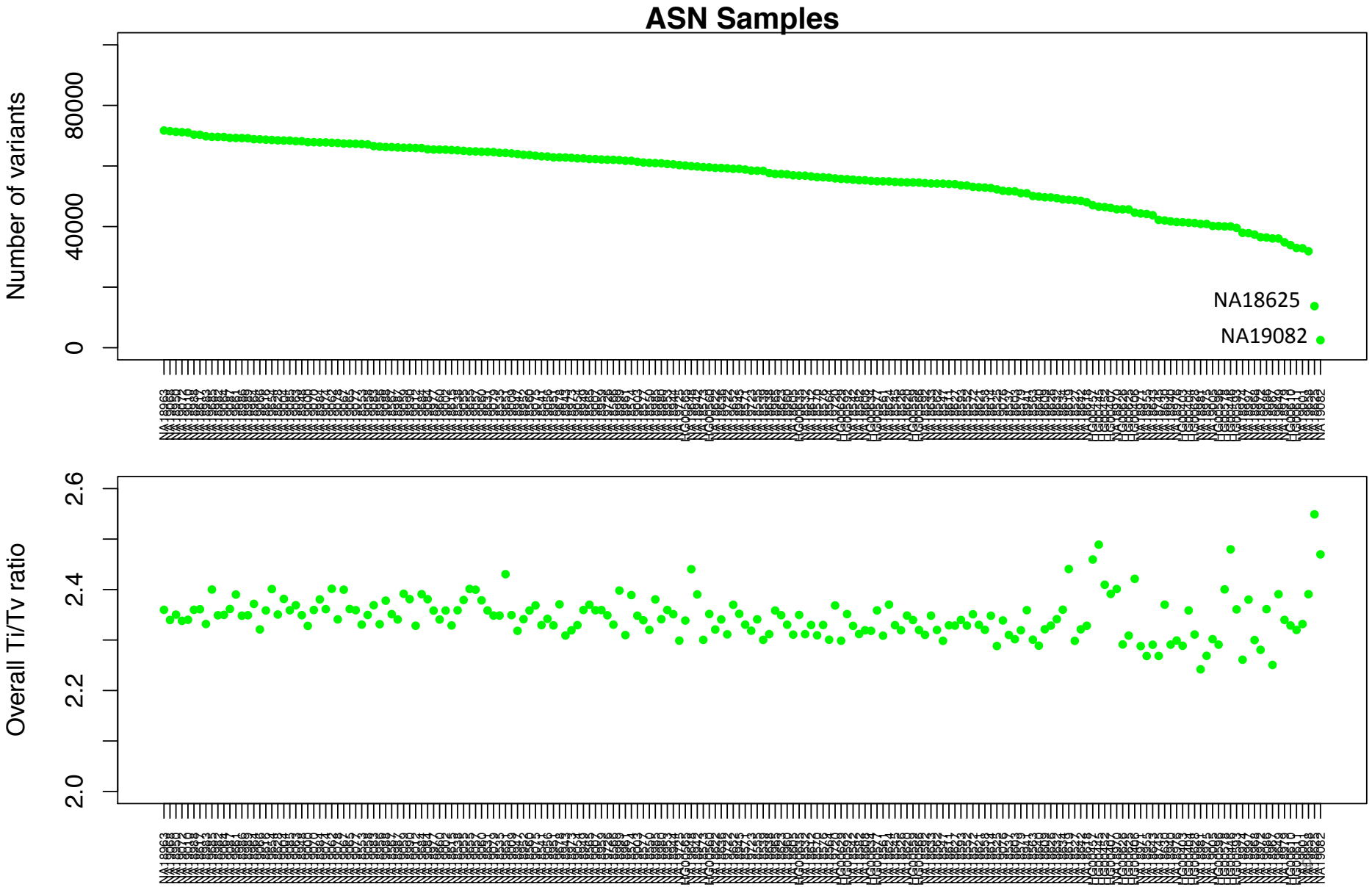
Ti/Tv ratio looks great across all samples



Ti/Tv ratio looks great across all samples



Ti/Tv ratio looks great across all samples



Conclusions

- Calling all populations simultaneously discovers many high-quality variants in each population
- BAQ calculation is immensely useful
- Call set metrics look quite good