BROAD INSTITUTE

1000G August Release, whole genome call set

Ryan Poplin (rpoplin@broadinstitute.org)
Genome Sequencing and Analysis
Medical and Population Genetics
October 28, 2010

BROAD INSTITUTE

Data and Definitions -- Samples

- 1000G August release, whole genome
- Samples selected by the group for August analysis exercise
- 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

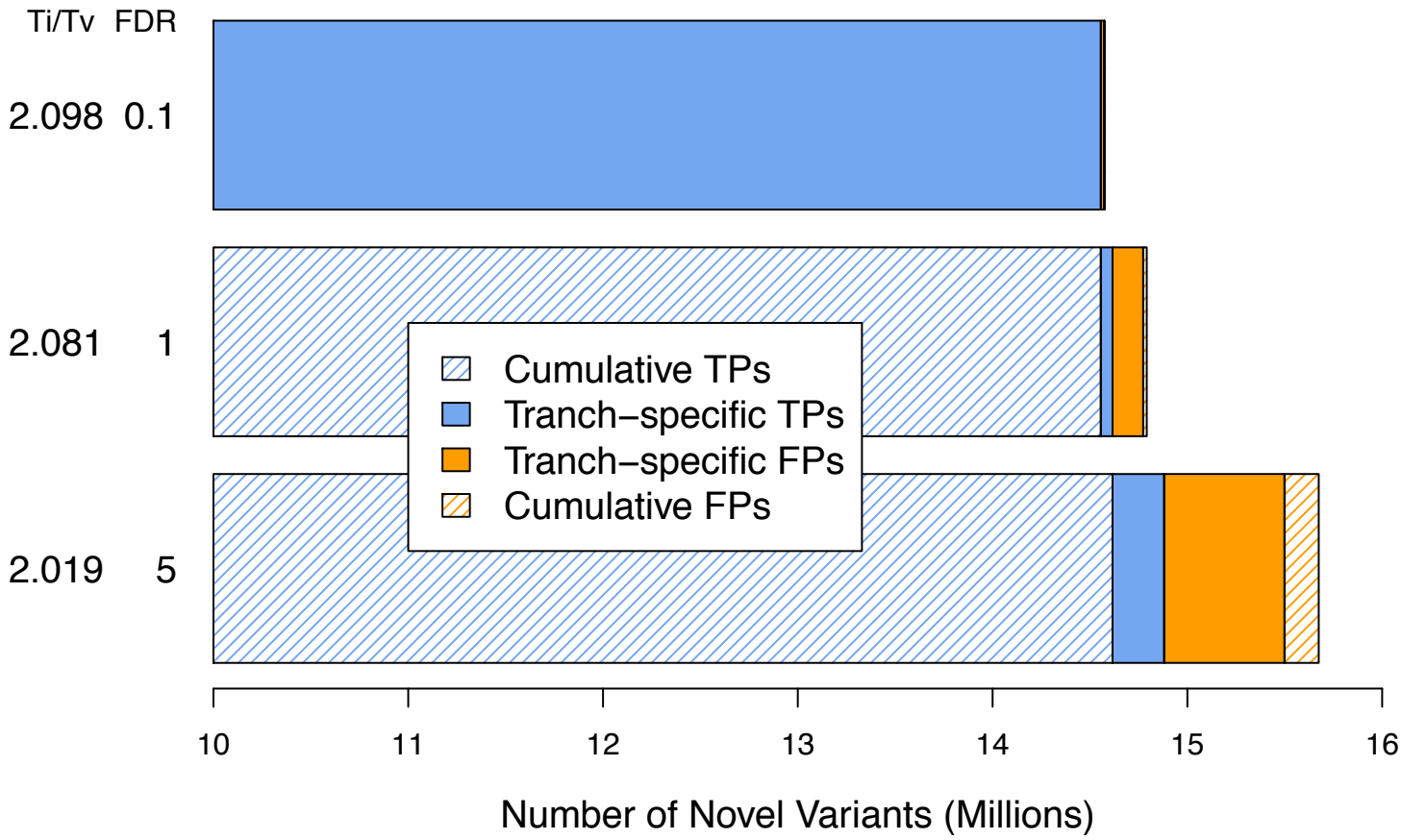
Data and Definitions -- Pipeline

- Sample-level cleaning at known indels
- BAQ calculation from Heng Li
- Called all samples simultaneously (629 samples)
- Filtering SNPs within 10bp of pilot Dindel calls
- Variant quality score recalibration by chromosome
  - Quality score cut chosen to give 0.1% implied novel FDR based on ti/tv from pilot2 NA19240 calls
- Genotype refinement via Beagle by analysis panel

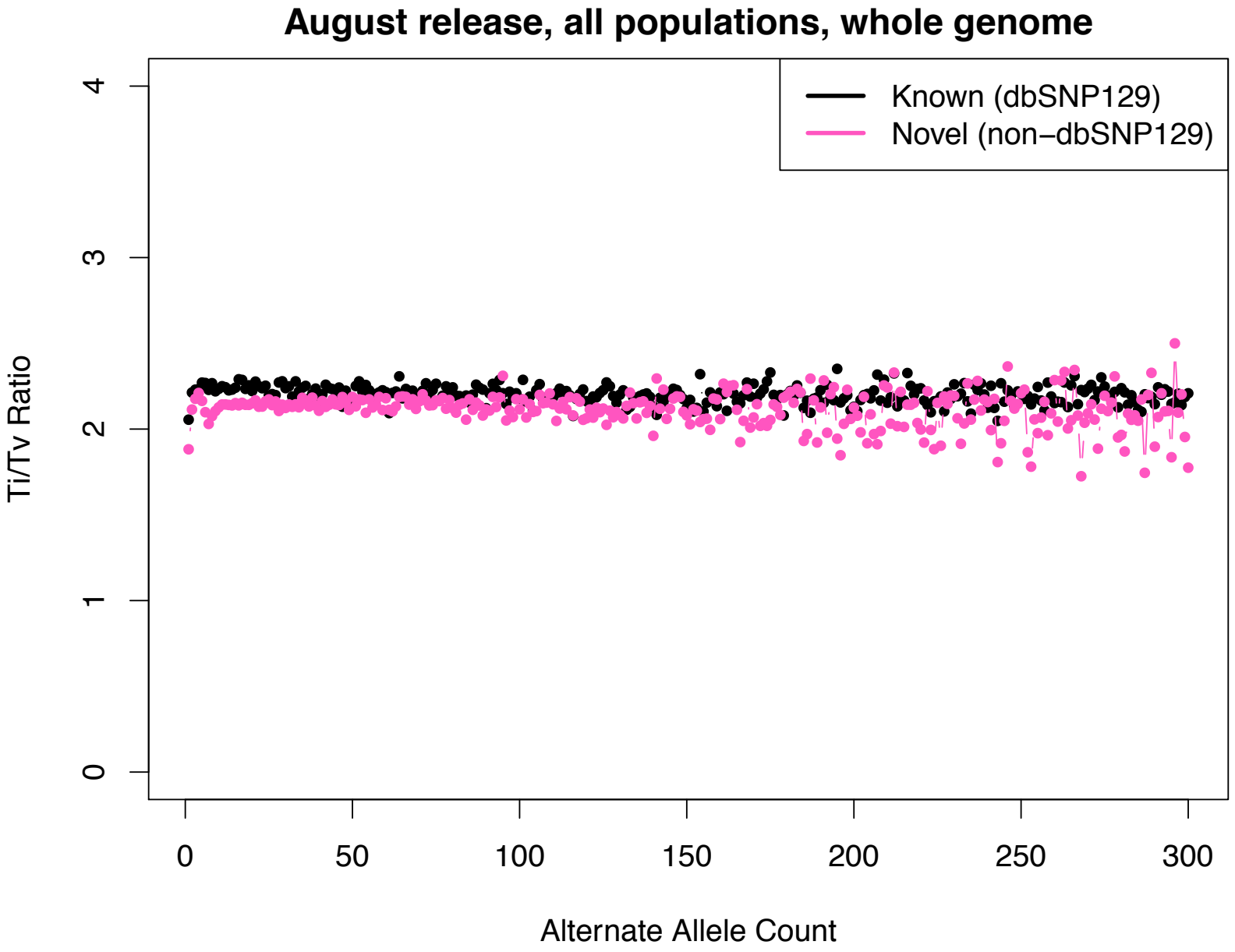
Calling all populations simultaneously yields large number of high-quality variants

# samples	Call set	Total # variants	dbSNP %	# knowns	Known ti/tv	# novels	Novel ti/tv
629	1000G August (EUR+AFR+ASN)	22,242,654	34.50	7,673,979	2.21	14,568,675	2.10
179	Pilot1 Nature Paper	14,894,361	54.00	8,042,955		6,851,406	

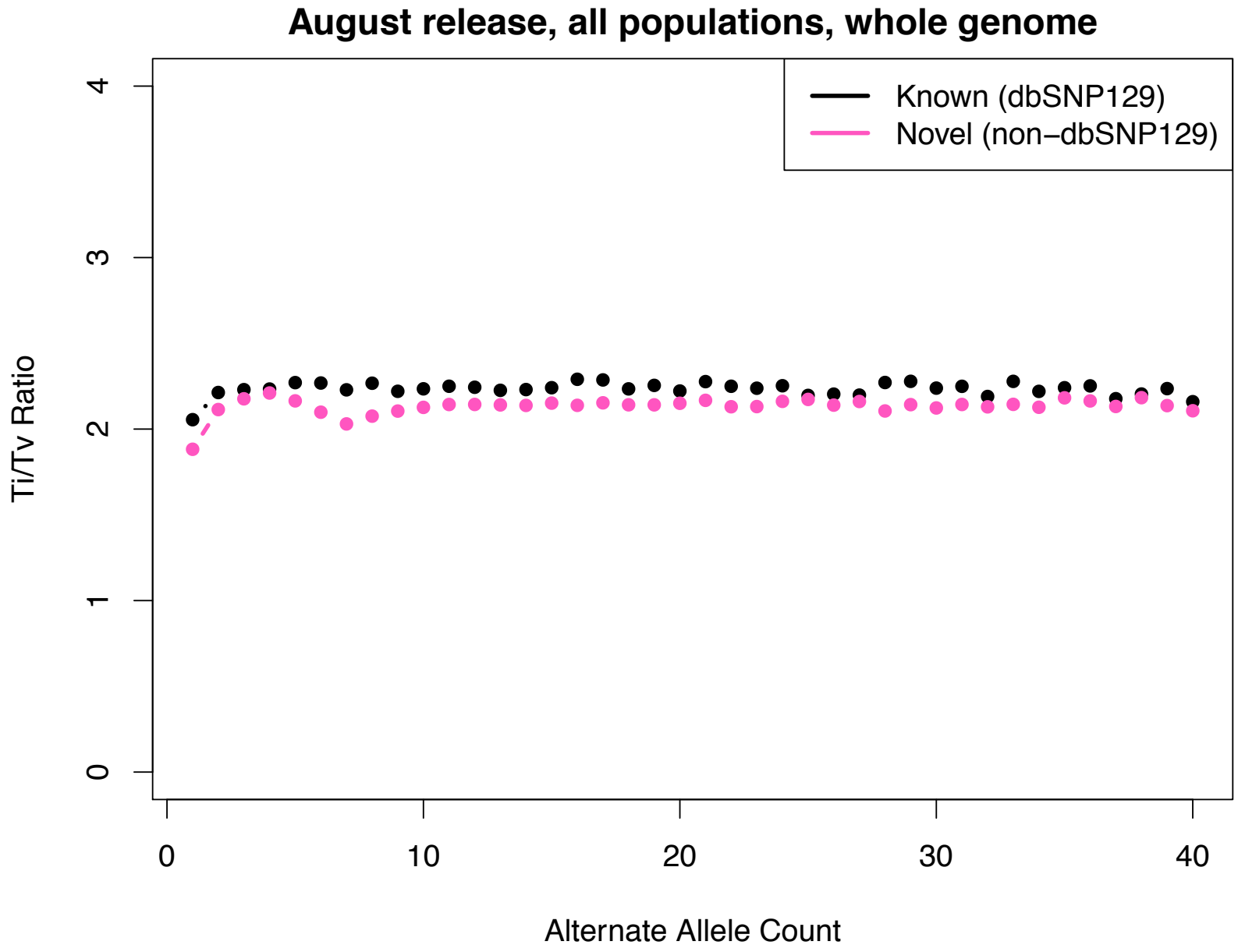
Variant Quality Score Recalibration provides call sets at several FDR levels



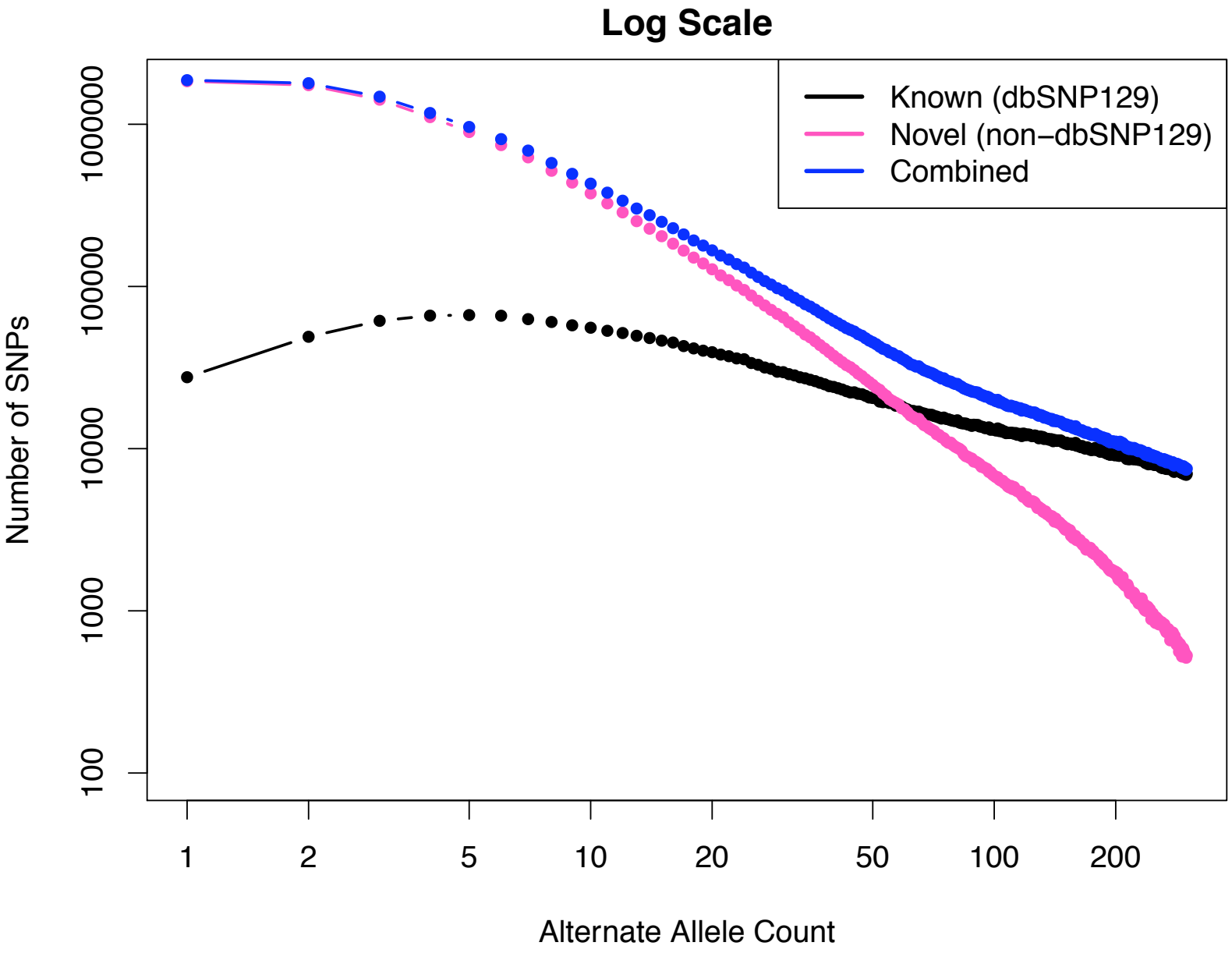
Ti/Tv ratio looks great across all allele counts



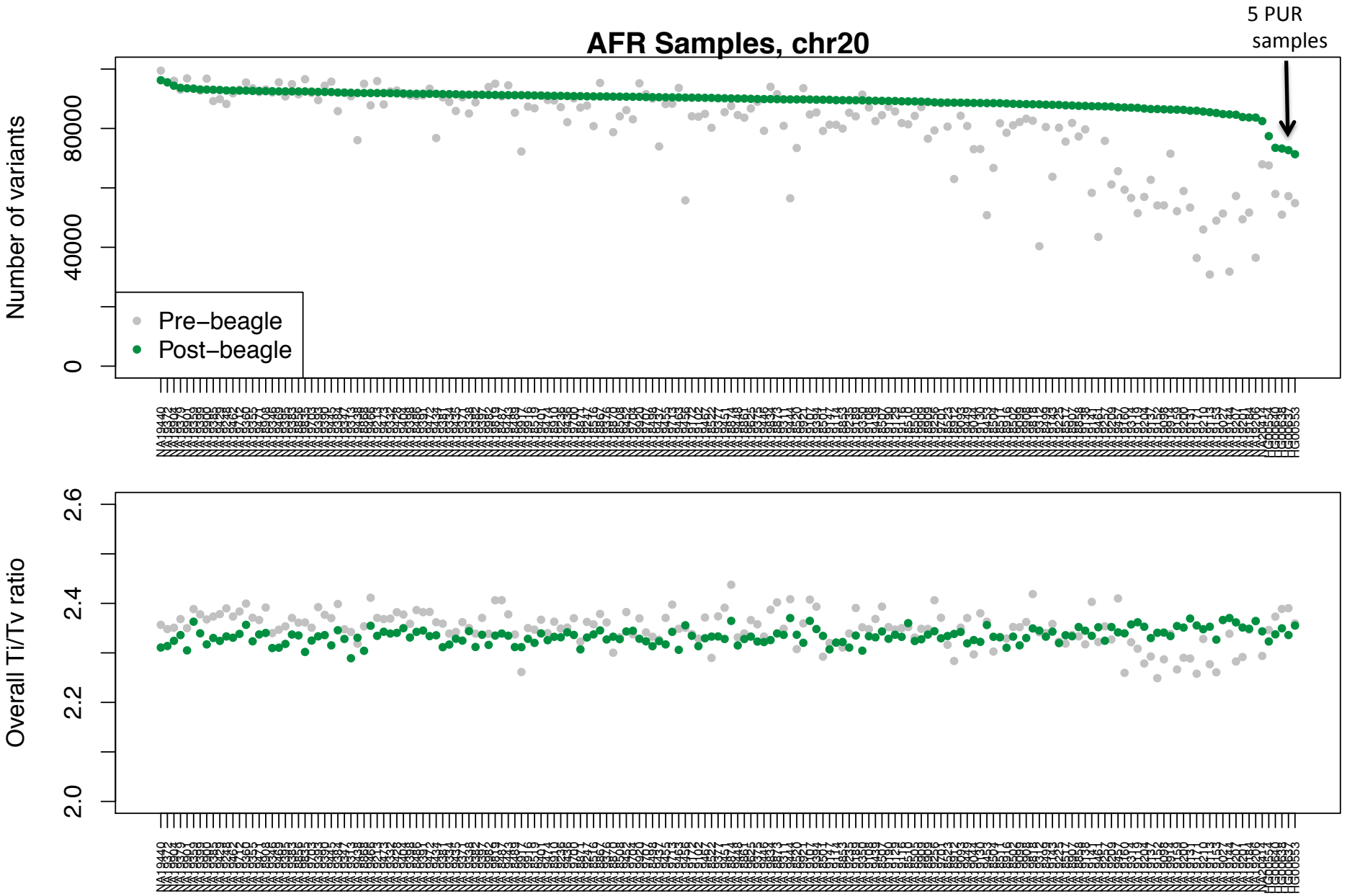
Ti/Tv ratio looks great across all allele counts



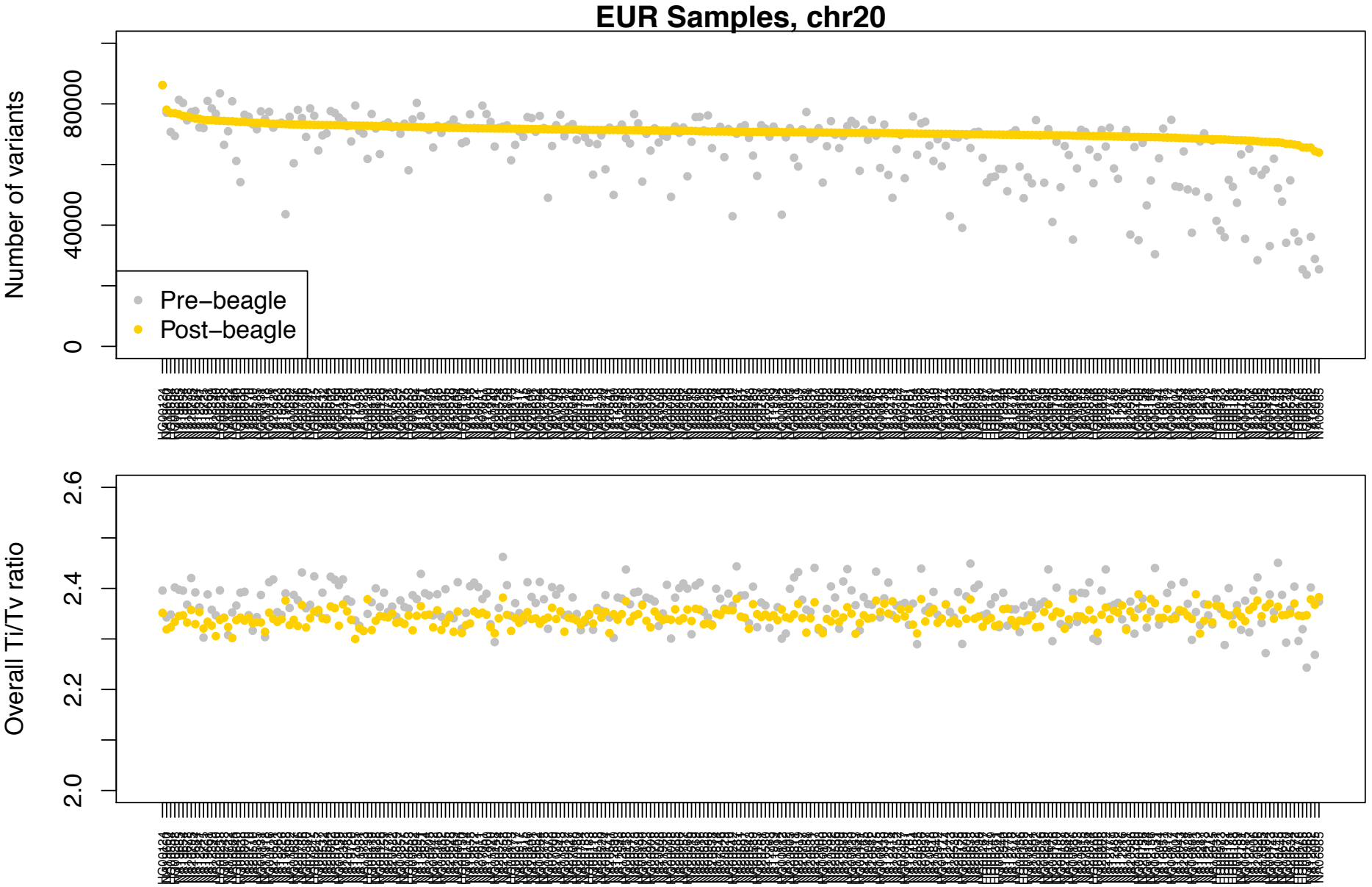
Allele frequency distribution



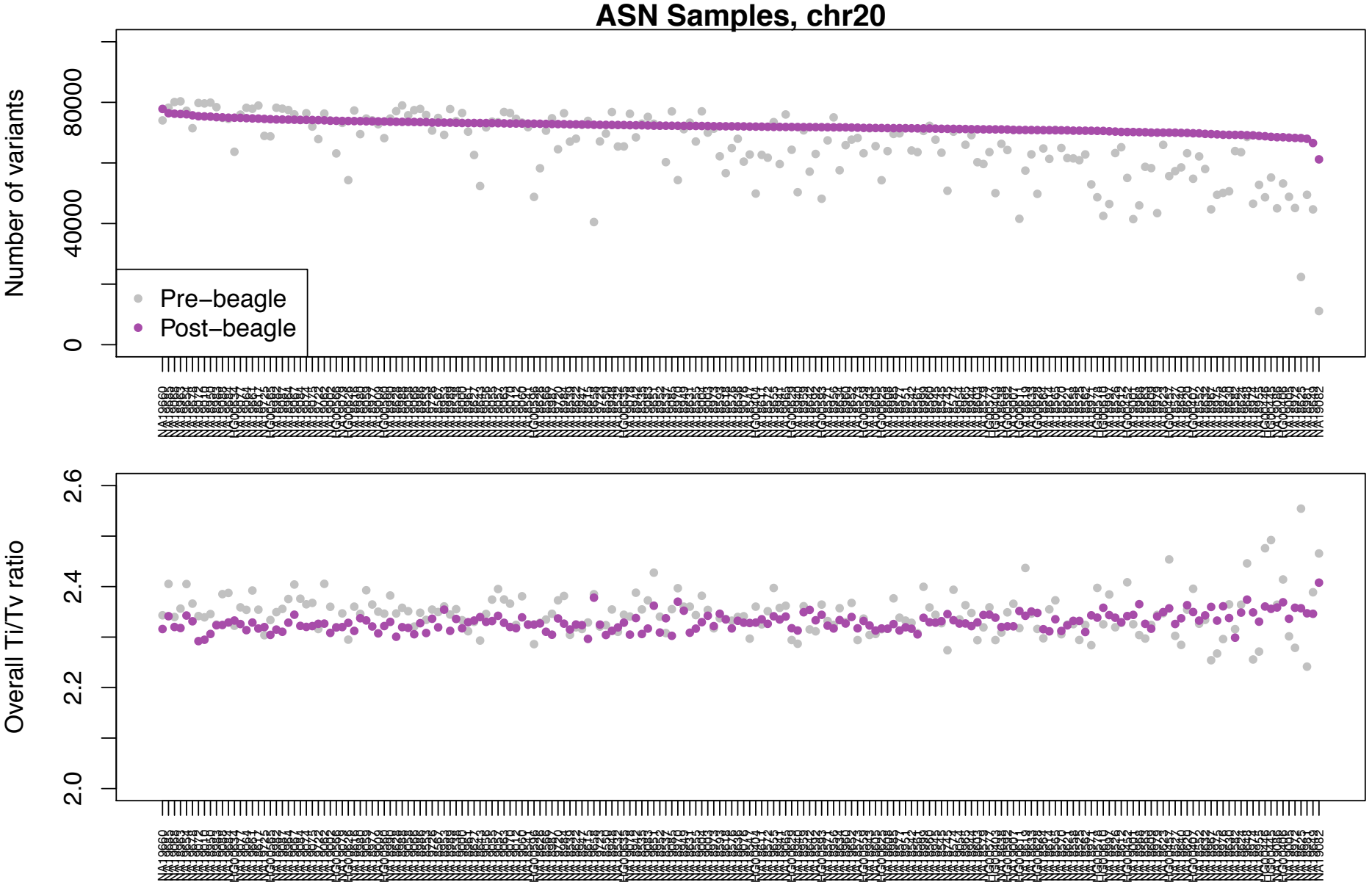
By-sample metrics are consistent



By-sample metrics are consistent



By-sample metrics are consistent



Conclusions

- Whole genome calling is done
- Large number of high quality variant sites
- Call set metrics look quite good
- Genotype refinement improves sample-level consistency tremendously