Michigan 1000 genome SNP Calls

- whole genome summary from 2010/08/04 freeze -

October 26, 2010

Hyun Min Kang

Materials and Methods

- Samples
  - 602 samples without SOLiD data
    - 174 AFR, 177 ASN, 280 EUR
  - DCC alignment (2010/08/04 freeze)
- Methods
  - BAQ adjustment
  - glfMultiples with ‘uniform’ Prior
  - Filtering of SNPs based on empirical distribution of various statistics (calibrated from chr20)

WG call Summary Statistics

Category (Marginal)	#SNPs	%dbSNP /SNP	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	%HM3 found	%HM3 /SNP
Unfiltered	**27,143,518**	**30.2**	**2.12**	**2.05**	**2.07**	**99.371**	**5.160**
AvgDepth > 10x	274,362	20.2	1.45	1.20	1.24	0.012	0.061
AvgDepth < 0.5x	70,291	30.4	1.02	1.21	1.15	0.005	0.101
#Samples < 10%	11,860	20.4	0.79	0.87	0.85	0.001	0.059
NearInDel < 5bp	338,854	30.3	1.26	1.25	1.26	0.315	1.309
MQbias (Ref>Alt)	87,118	30.2	1.54	1.06	1.19	0.024	0.396
MQbias (Ref<Alt)	224,655	80.5	2.34	1.30	2.07	0.023	0.143
StrandBias ($p<10^{-7}$)	795,990	16.7	1.78	0.50	0.62	0.107	0.189
ReadTailBias (Ref)	150,271	48.6	1.44	1.11	1.26	0.019	0.176
ReadTailBias (Alt)	256,731	13.2	1.43	1.04	1.09	0.014	0.079
All Filtered Out	**1,709,282**	**27.4**	**1.72**	**0.80**	**0.99**	**0.492**	**0.405**
Filtered	**25,434,236**	**30.4**	**2.15**	**2.20**	**2.18**	**98.880**	**5.480**

chr20 call Summary Statistics

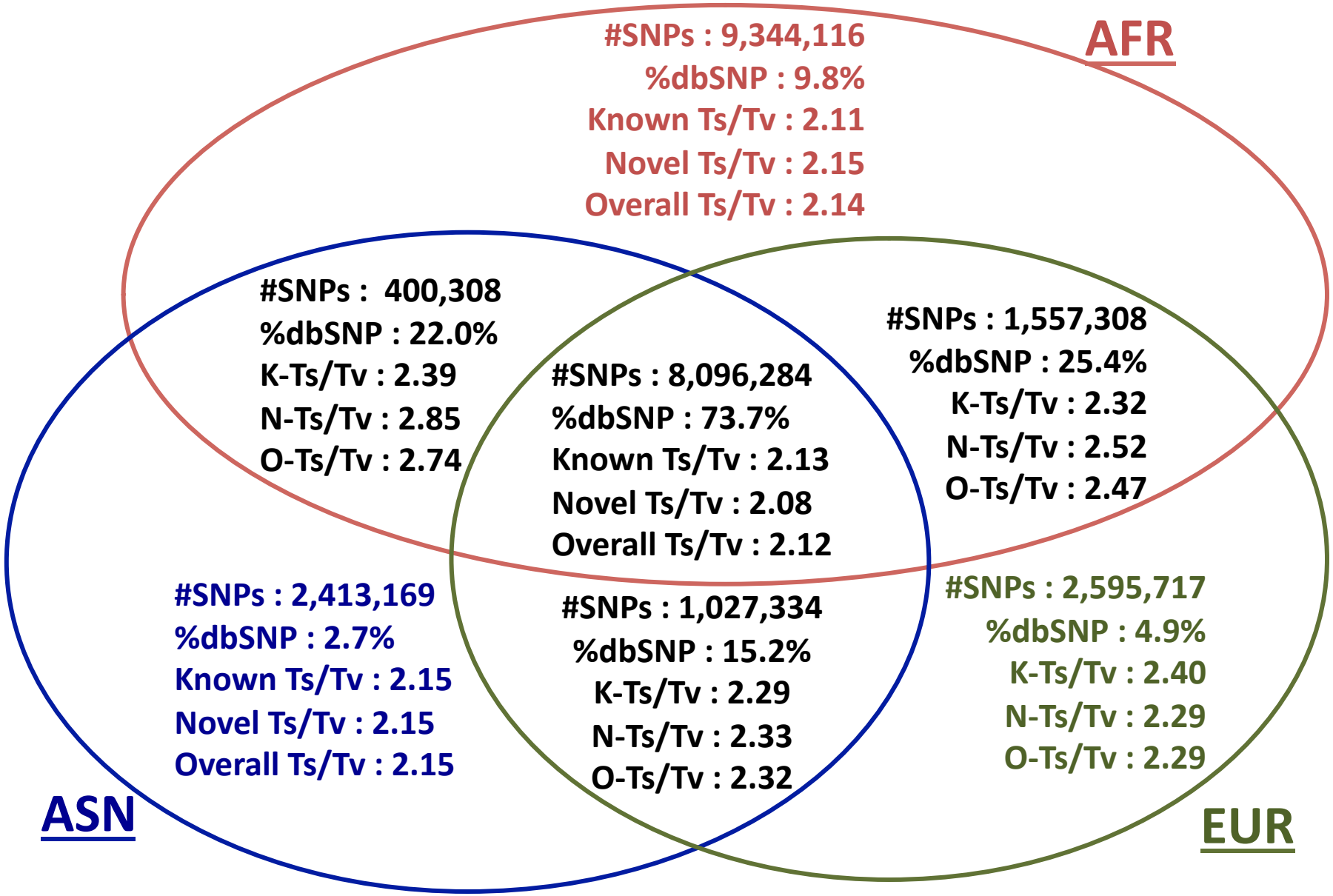
Category (Marginal)	#SNPs	%dbSNP /SNP	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	%HM3 found	%HM3 /SNP
Unfiltered	**624,393**	**31.0**	**2.30**	**2.16**	**2.20**	**99.663**	**5.208**
AvgDepth > 10x	9,005	18.1	1.50	1.21	1.26	0.022	0.089
AvgDepth < 0.5x	1,659	27.2	1.12	1.13	1.12	0.000	0.000
#Samples < 10%	281	21.7	0.69	0.65	0.66	0.000	0.000
NearInDel < 5bp	8,417	29.2	1.48	1.27	1.33	0.353	1.544
MQbias (Ref>Alt)	4,286	24.4	1.49	1.17	1.24	0.022	0.187
MQbias (Ref<Alt)	3,975	82.1	2.67	1.32	2.32	0.014	0.126
StrandBias ($p<10^{-7}$)	23,430	13.7	1.90	0.62	0.72	0.144	0.226
ReadTailBias (Ref)	3,781	40.5	1.43	1.19	1.28	0.016	0.159
ReadTailBias (Alt)	8,219	12.8	1.44	1.02	1.07	0.014	0.061
All Filtered Out	**48,268**	**22.4**	**1.80**	**0.88**	**1.03**	**0.567**	**0.383**
Filtered	**576,125**	**31.7**	**2.33**	**2.38**	**2.36**	**99.096**	**5.612**

Population-subsetted calls

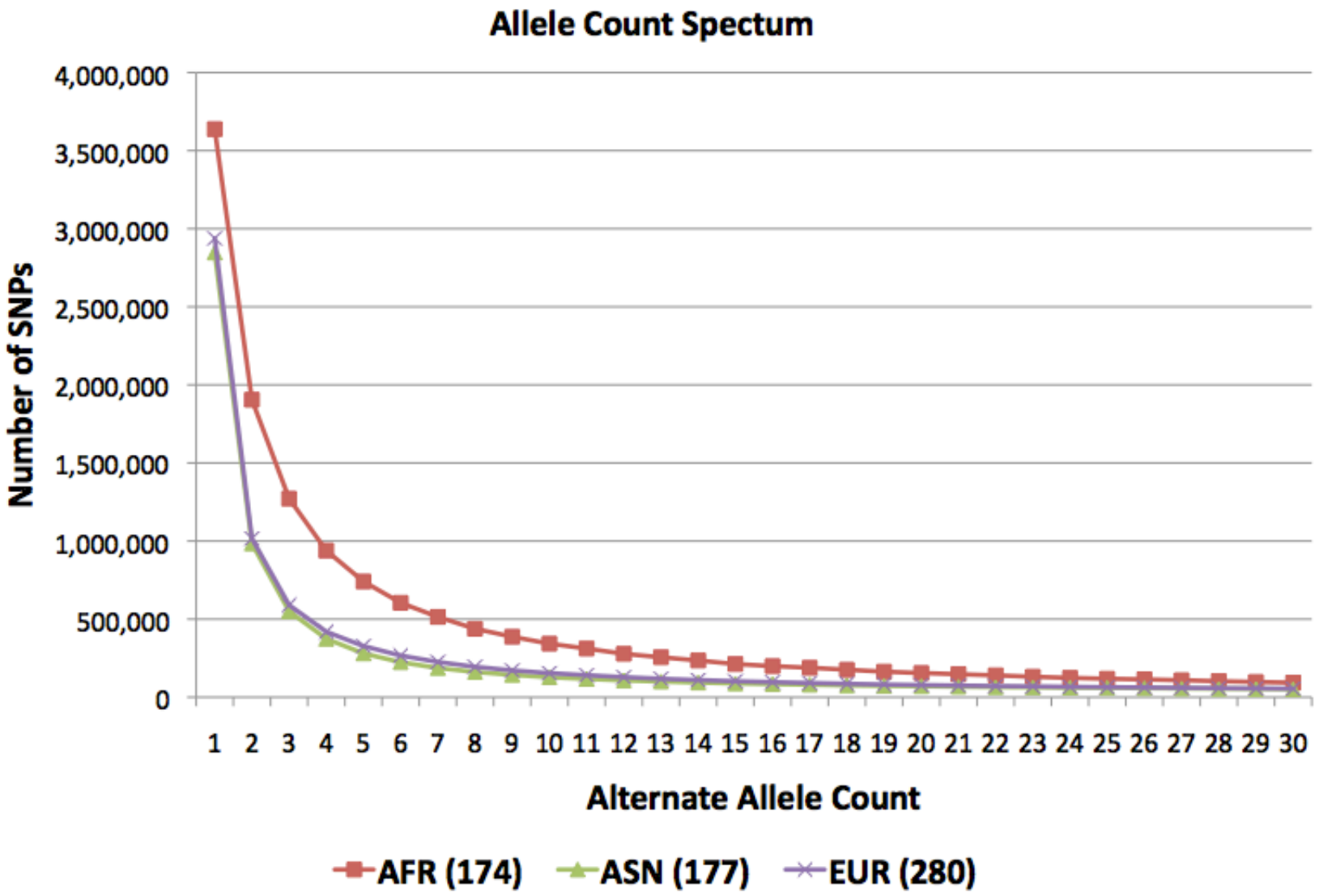
Category (Marginal)	#SNPs	%dbSNP /SNP	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	%HM3 found	%HM3 /SNP
Filtered/ALL	25,434,236	30.4	2.15	2.20	2.18	98.880	5.480
Filtered/AFR	19,398,016	38.0	2.14	2.18	2.17	98.929	6.999
Filtered/ASN	11,937,095	52.6	2.14	2.18	2.16	97.778	10.397
Filtered/EUR	13,276,643	50.1	2.15	2.26	2.20	98.205	9.602

* *%HM3 are evaluated from overlapping individuals between 1KG (629 – w/ SOLiD) and HM3 release 3 (477 ALL, 180 EUR, 156 AFR, 156 ASN), considering only polymorphic sites within the subset*
* *%dbSNPs evaluation and Ts, Tv classification are based on dbsnp_129_b37.rod from GATK resource files, taking only subset with "class = single" AND "end – start = 1"*

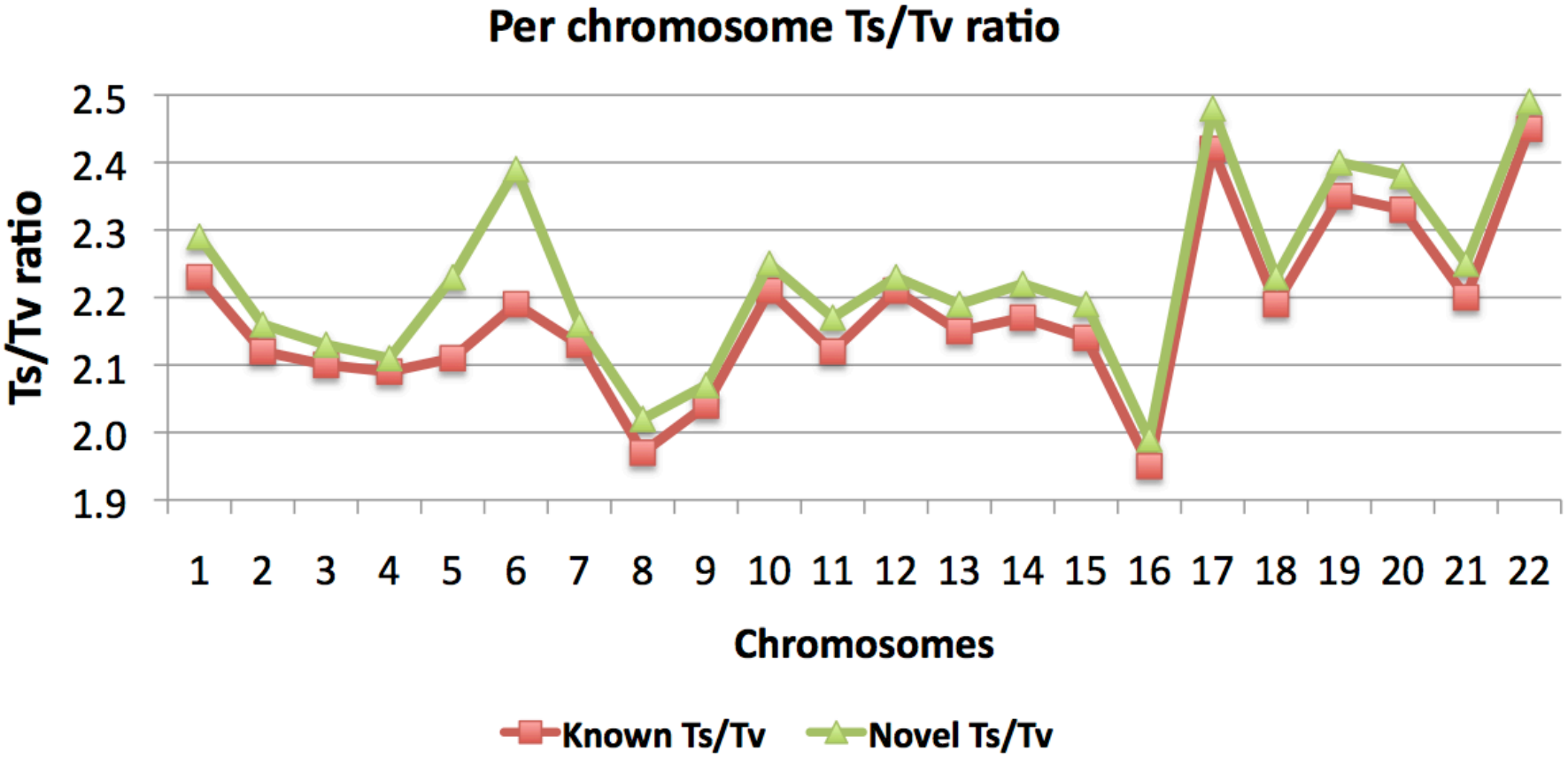
Population-wide Venn Diagram



Allele Count Spectrum

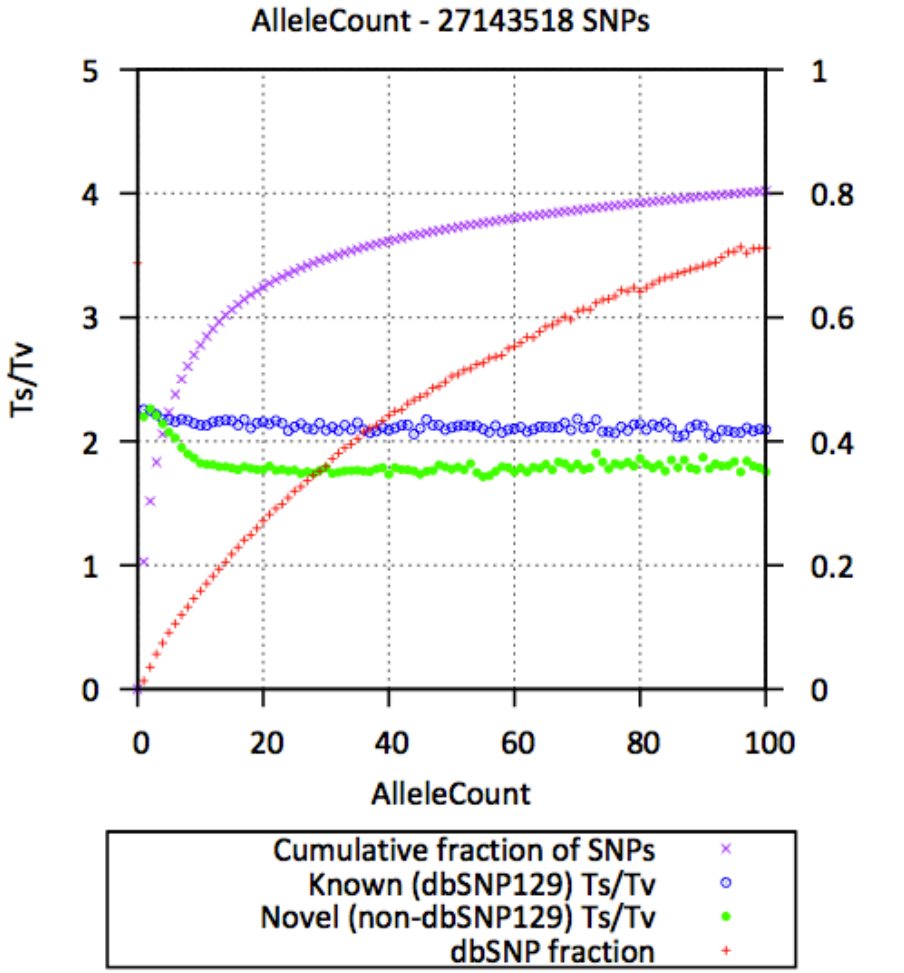


Ts/Tv largely vary across chromosomes

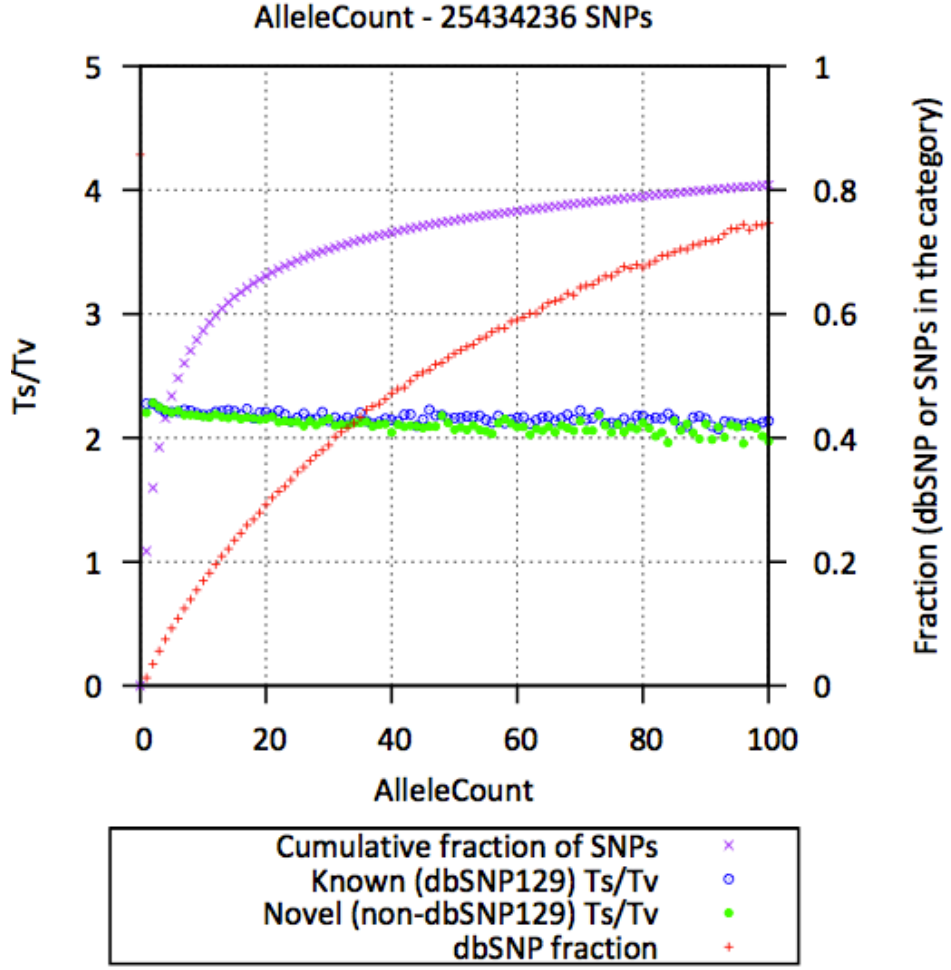


Ts/Tv at low AF (1-10%) SNPs improves after filtering

Unfiltered

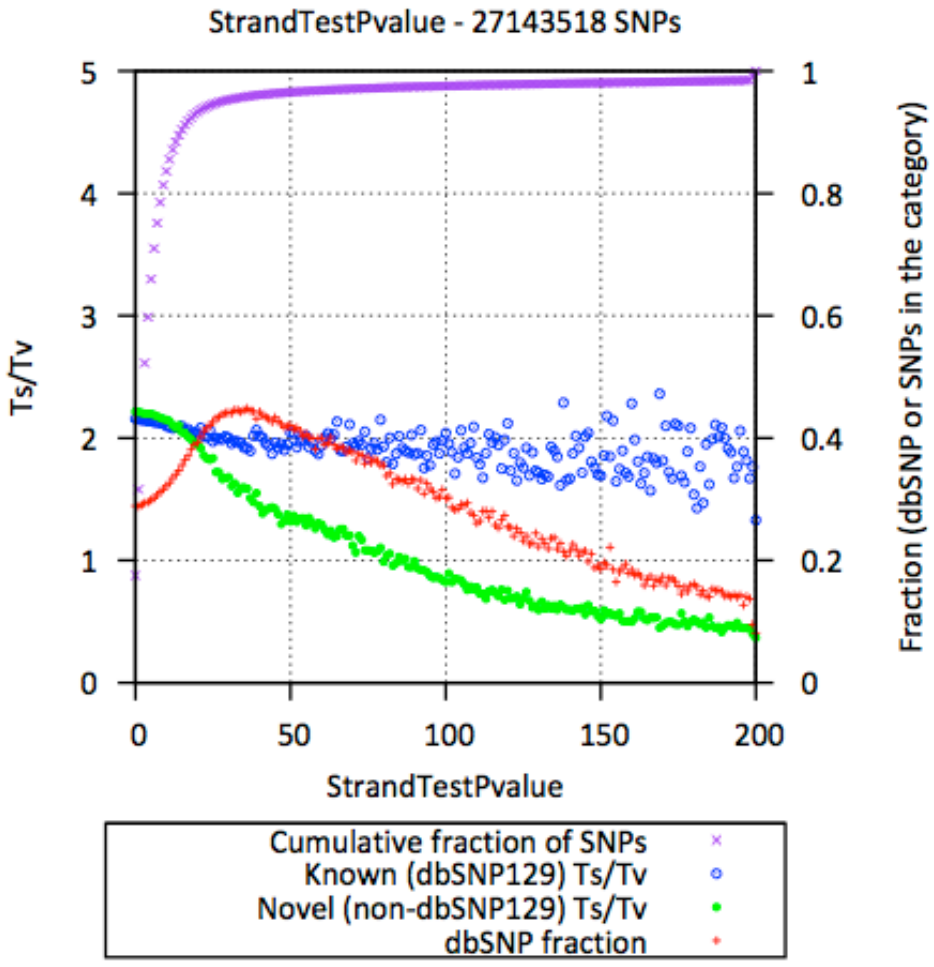


Filtered

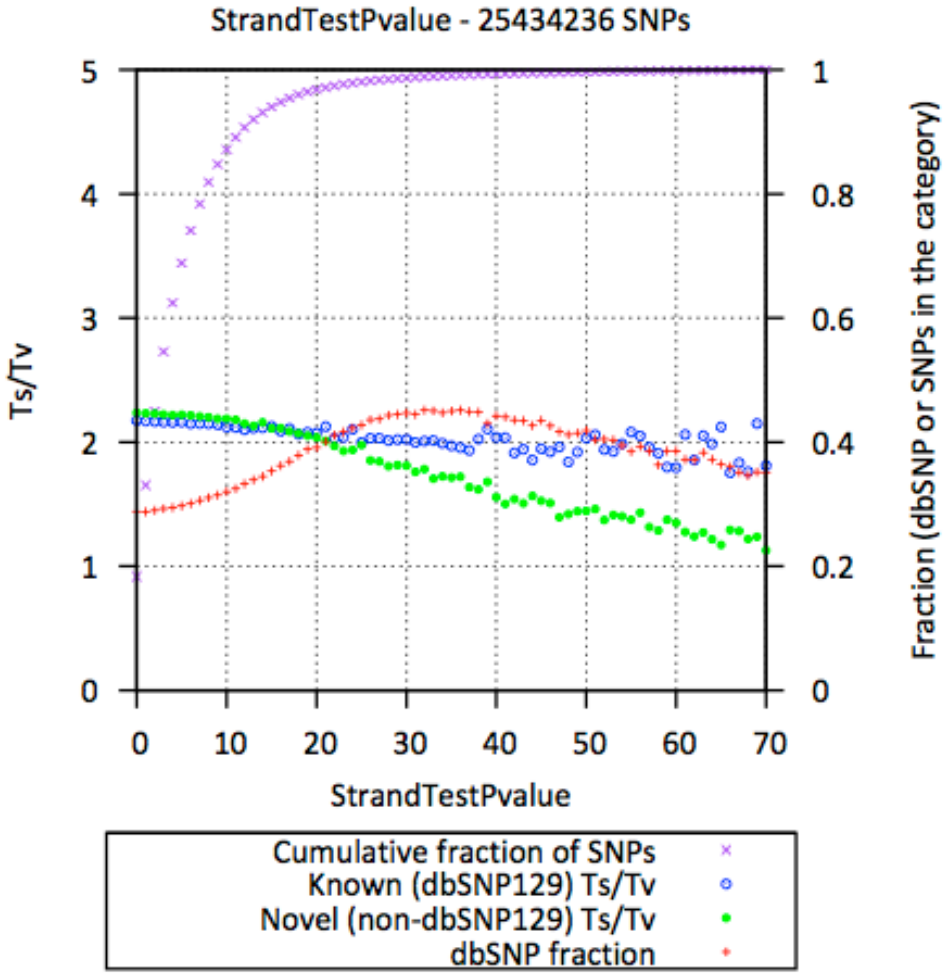


Strand Bias Filter :
a very effective filter in improving novel Ts/Tv

Unfiltered

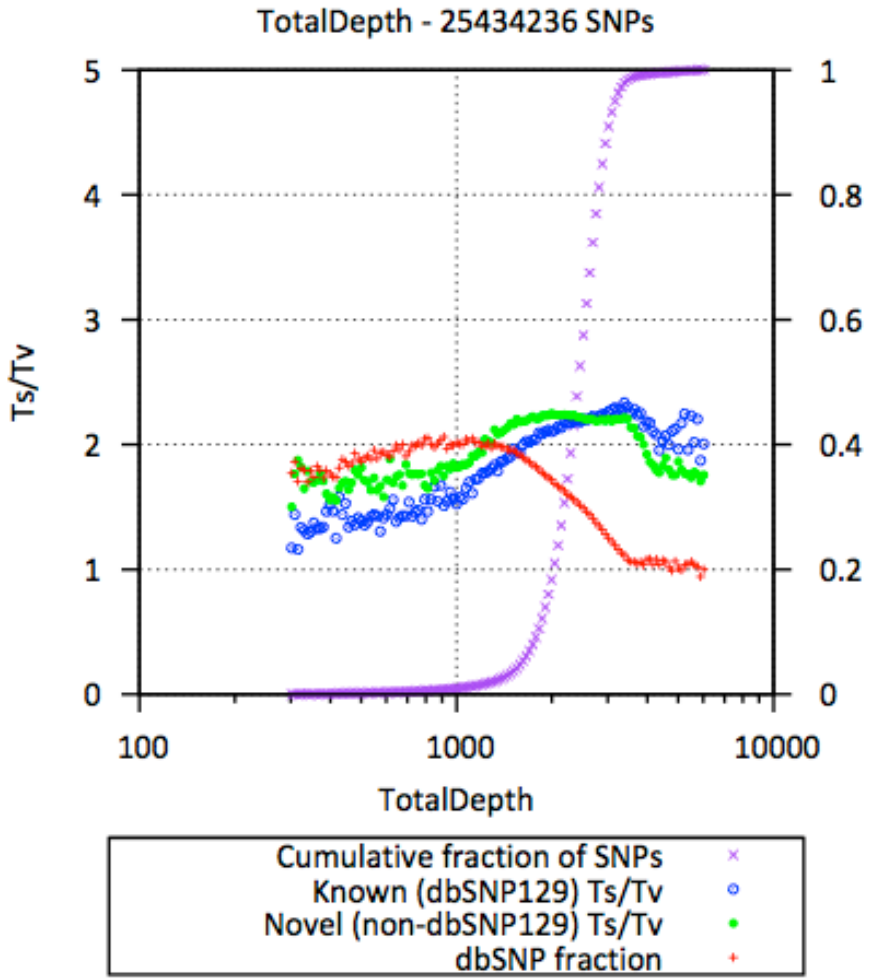


Filtered

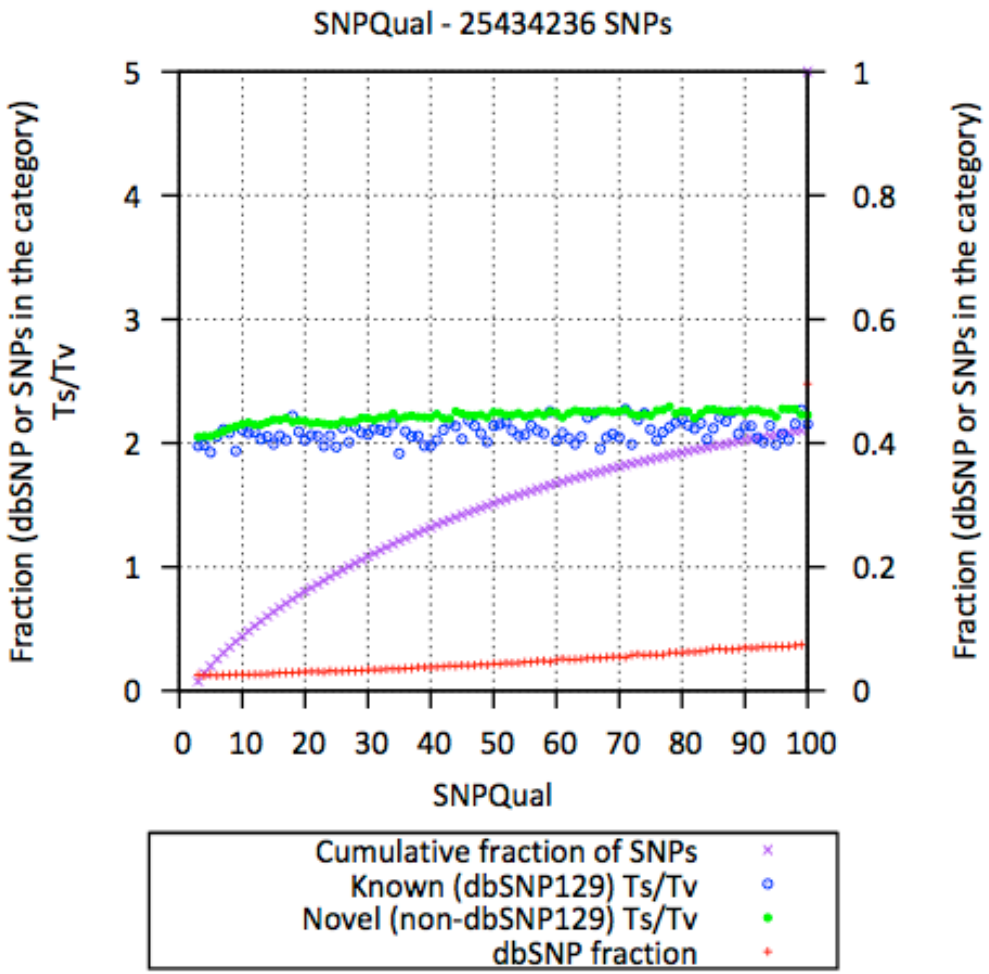


Impact of lenient filters for depth and QUAL

Filtered / Total Depth



Filtered / QUAL



Current Status

- Whole genome SNP calls finished
  - Available at ftp://share.sph.umich.edu/1000genomes/fullProject/2010.08.04.WG/
- LD-aware genotype refinement
  - 25% are currently finished across 3 populations
  - More are running

Acknowledgements

- Goncalo Abecasis
- Carlo Sidore
- Tom Blackwell
- Heng Li