Improving the Specificity of SNP Calls in the 1000 Genomes Project

Michael Melgar
Medical & Populations Genetics
Genome Sequencing & Analysis Group
August 7, 2009

Summer Research
Program in Genomics

1000 Genomes Project seeks to find rarer variants

- Genetic disease studies have used Hapmap to associate variants with ischemic heart disease[1,2], obesity[3,4], and other diseases

- 1000 Genomes takes the next step after Hapmap: catalogs variants with minor allele frequencies down to ~1% by sequencing more individuals

- More and rarer variants provide enhanced genomic resolution to disease association studies

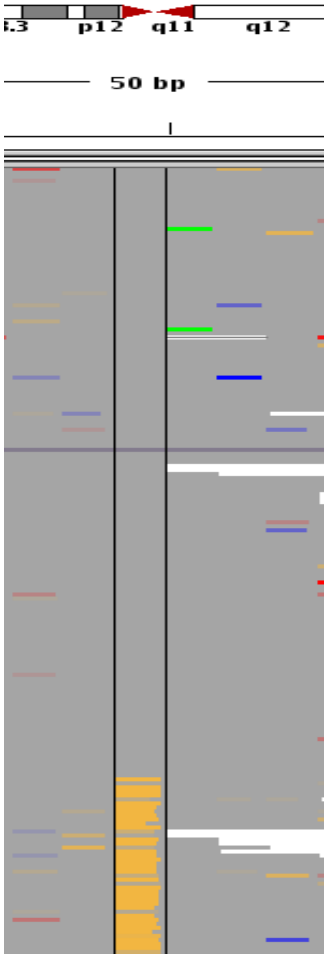
1. Nora et al., Circulation 61, 503–508 (1980)
2. Myocardial Infarction Genetics Consortium, Nature Genetics 41, 334 - 341 (2009)
3. Wardle et al., J. Clin. Nutr. 87, 398–404 (2008)
4. Thorleifsson et al., Nat. Genet. 41, 18–24 (2009)

SNP calling from high-throughput resequencing is still in its infancy

- Sequences from 454, Illumina, SOLiD sequencers contain errors, leading to false SNP detection

- Approaches to modeling error are essential to identify & remove false positive SNP calls

False positive heterozygotes are often characterized by **allelic imbalance**



Individual ID: **NA12878** Chr1 : 245,047,907

Called genotype: CG

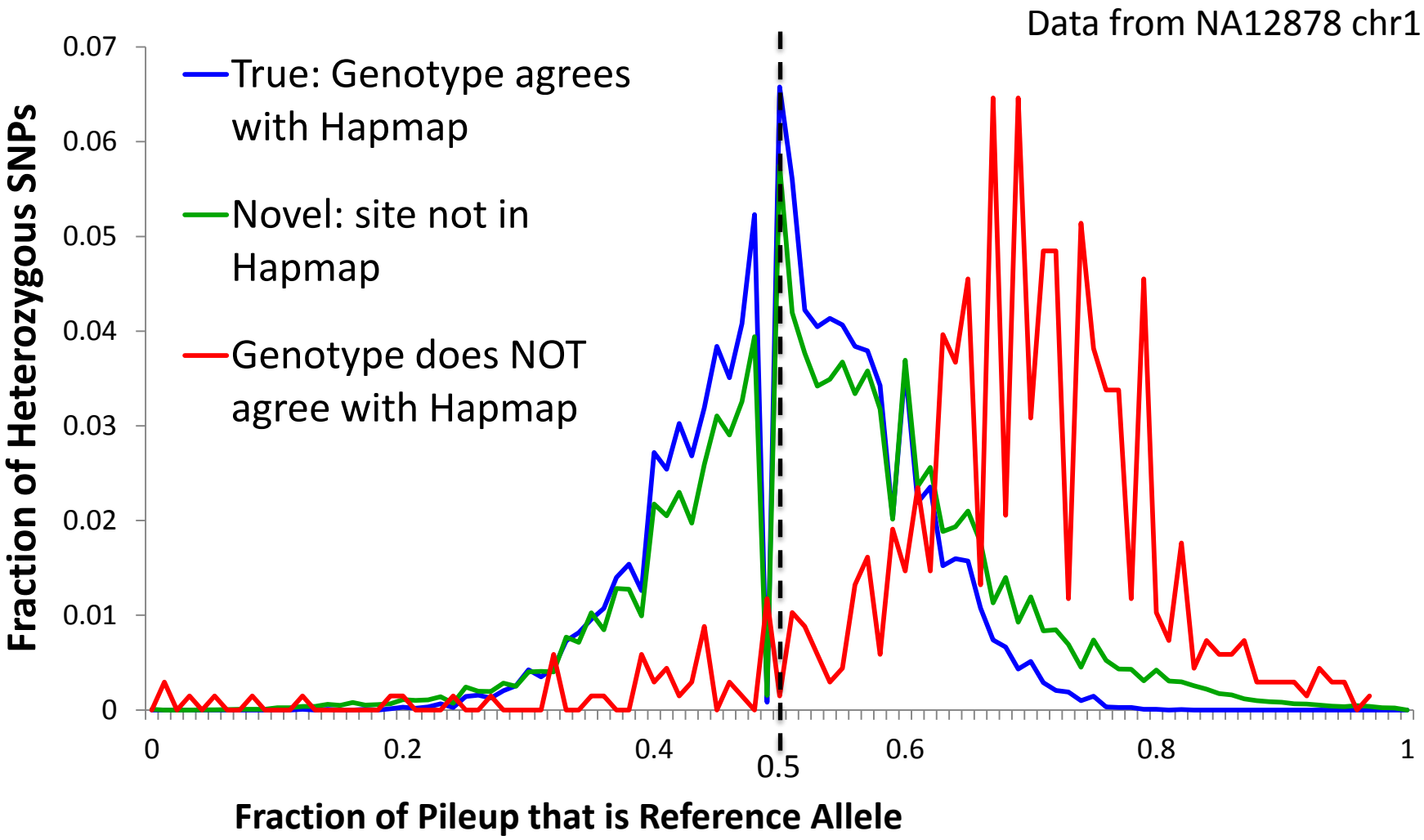
Fraction of alleles in pileup: C:0.76 G:0.21

Fraction of alleles expected for a true heterozygote: C:0.50 G:0.50

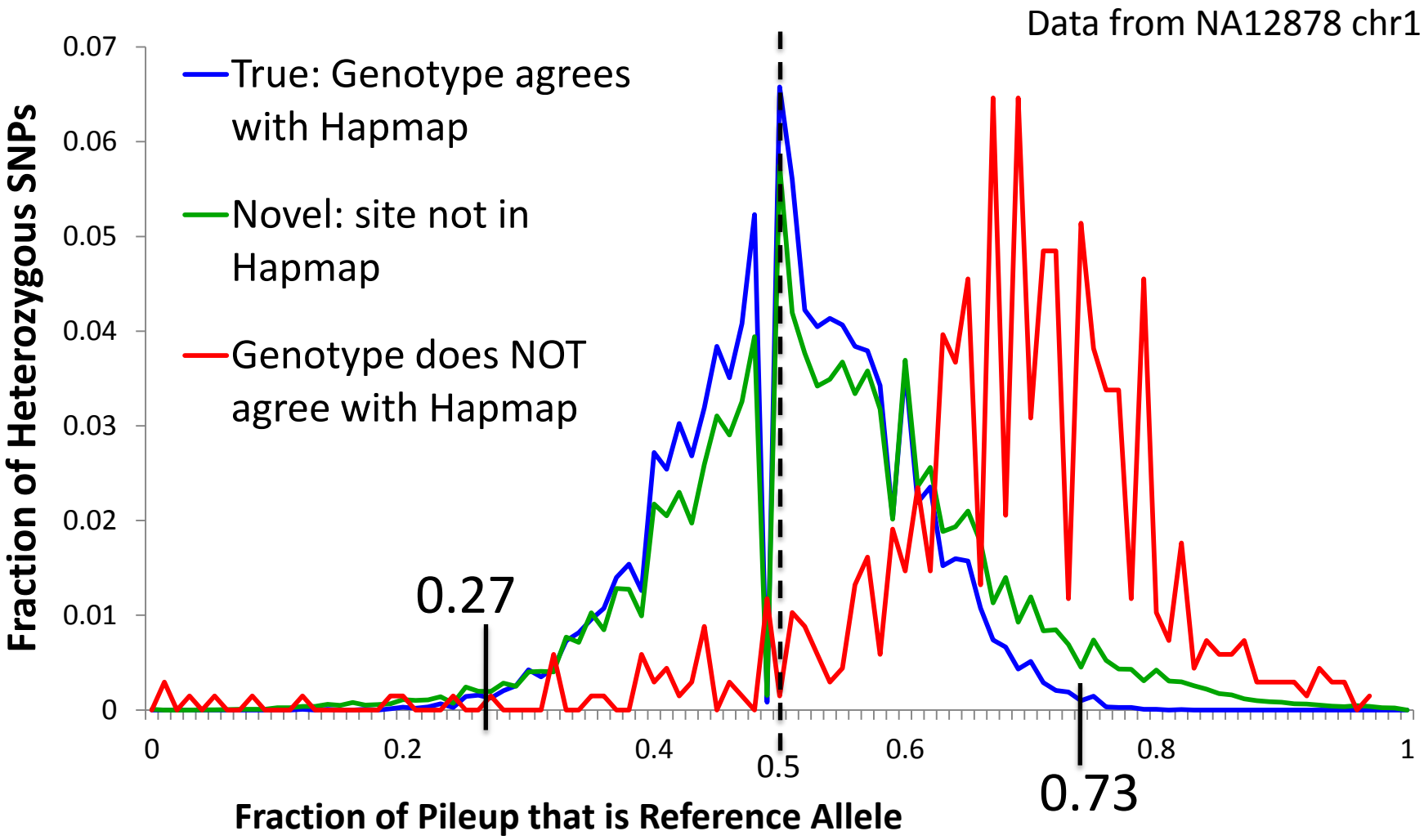
Deviations from this expectation are suggestive of false positives…

Image generated using Integrated Genome Viewer (Jim Robinson)

True SNPs tend to have reference allele fraction 0.5



True SNPs tend to have reference allele fraction 0.5

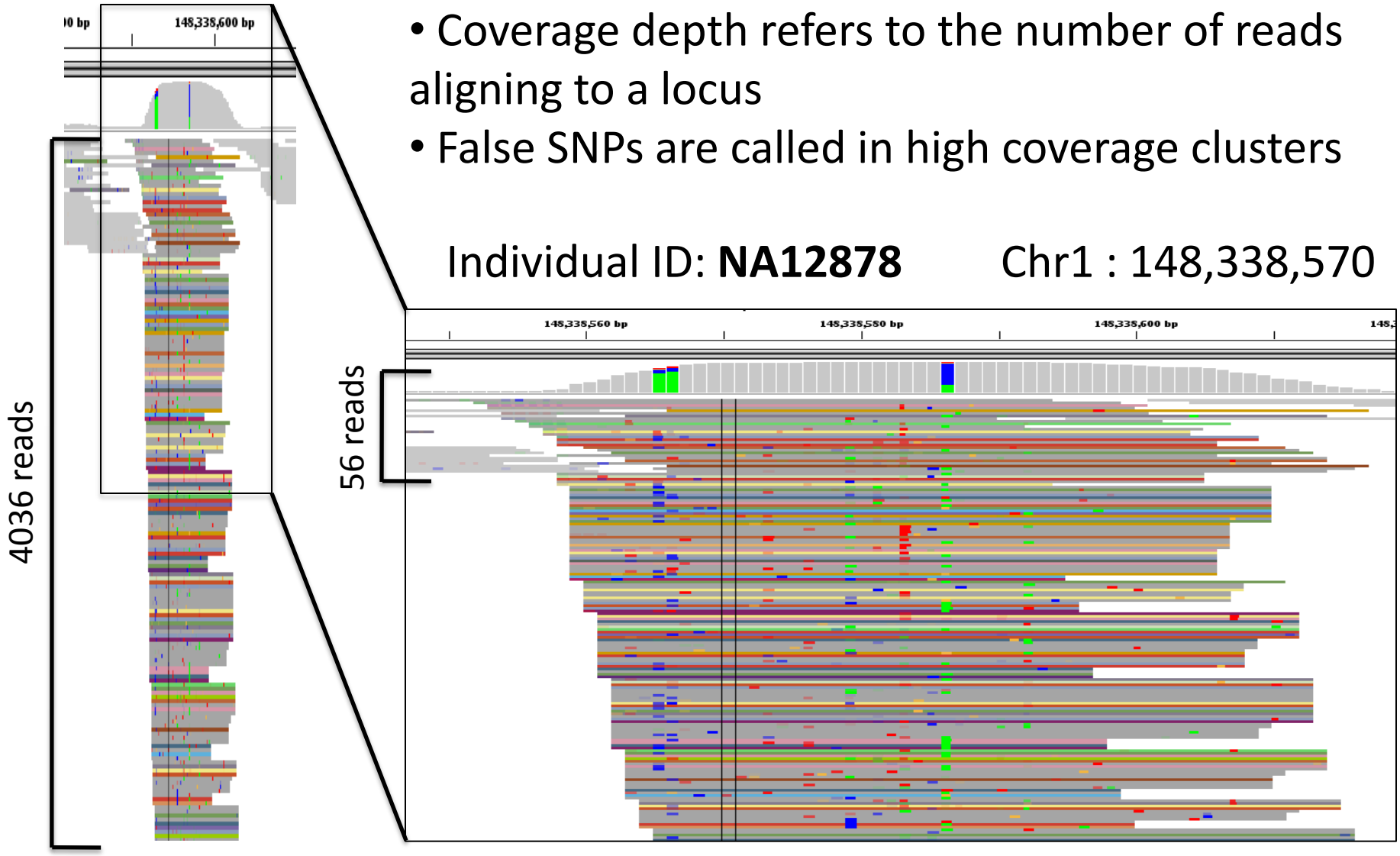


Allelic imbalance filter removes false positives

	All SNPs called in NA12878 chr1	SNPs that passed filter	SNPs removed by filter
# of SNPs	~262k	~260k	**2,809**
dbSNP %[1]	90 %	91 %	**33 %[2]**
Transition[3]/Transversion[4] Ratio[5]	2.1	2.1	**1.0**

1. SNP Database (dbSNP) is a catalog of known common variants in the genome
2. A low dbSNP % indicates a high number of false positives
3. Substitutions from one pyrimidine to another (C ↔ T) or from one purine to another (A ↔ G) are called transitions
4. Substitutions from a pyrimidine to a purine or from a purine to a pyrimidine are called transversions
5. The expected ratio of transitions to transversions in the whole genome is near 2.0

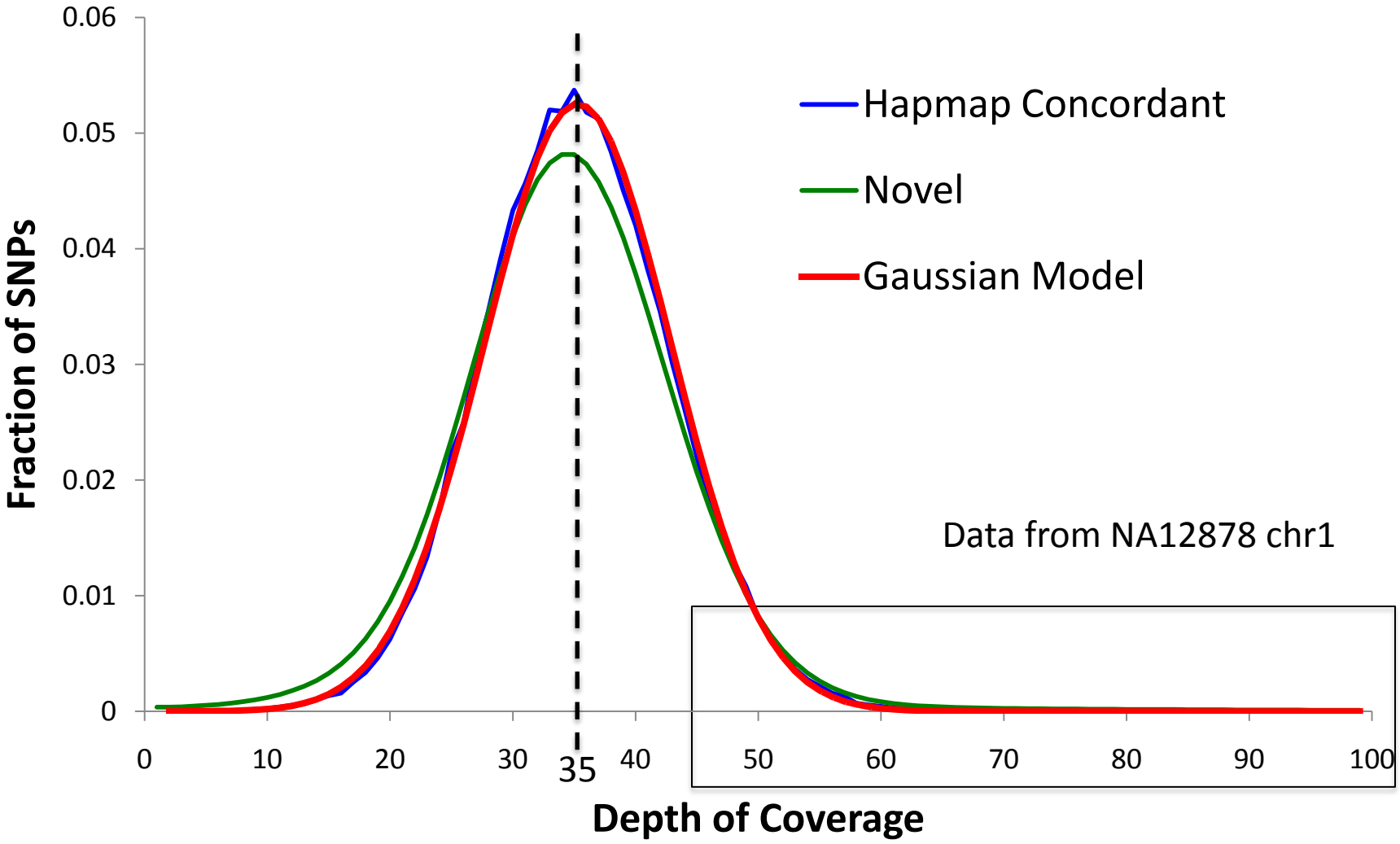
Excessive **depth of coverage** indicates misalignment



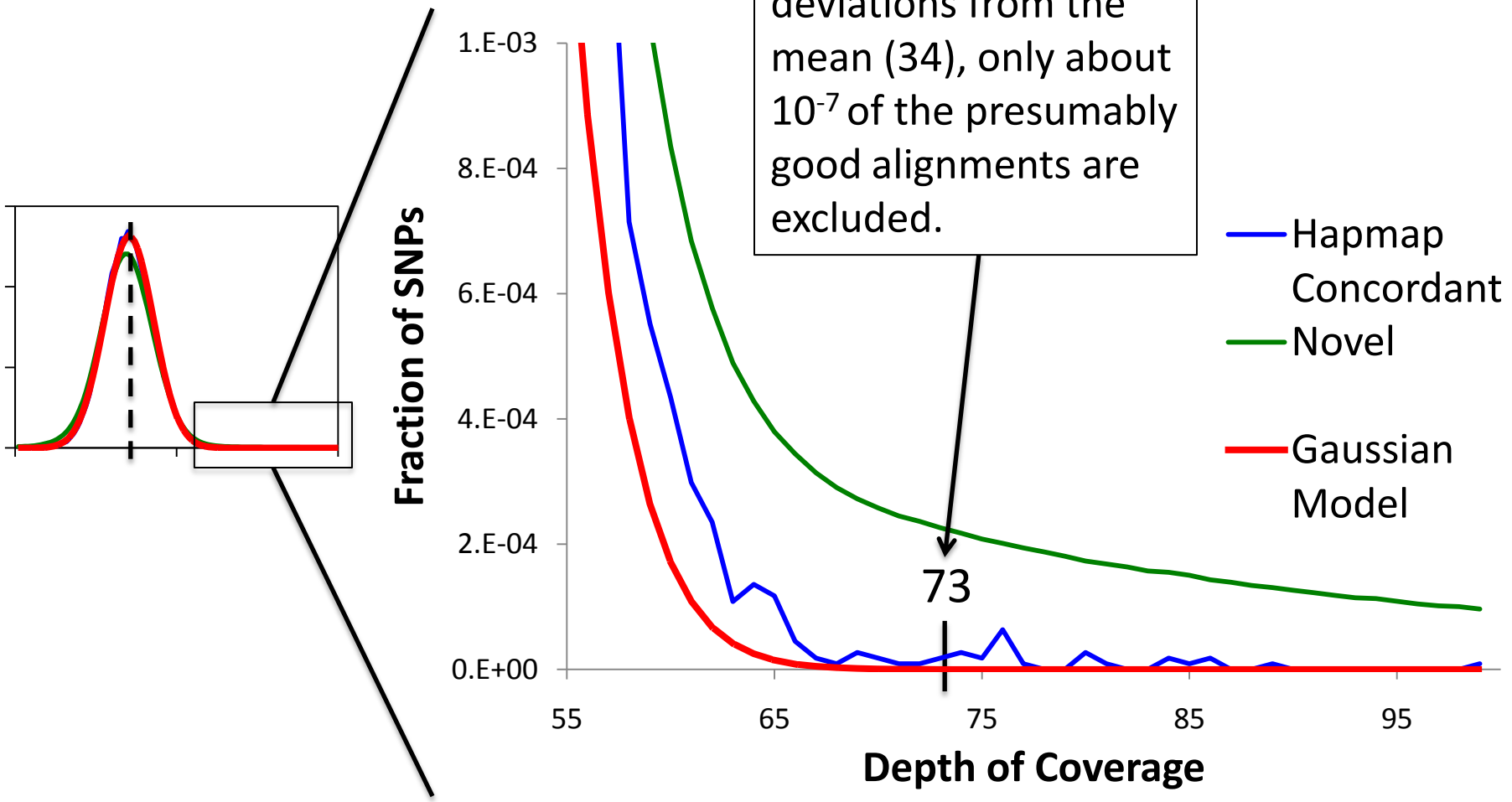
- Coverage depth refers to the number of reads aligning to a locus
- False SNPs are called in high coverage clusters

Individual ID: **NA12878** Chr1 : 148,338,570

True SNPs do not exhibit extremely high coverage depth



True SNPs do not exhibit extremely high coverage depth



Depth of coverage filter removes false positives

	All SNPs called in NA12878 chr1	SNPs that passed filter	SNPs removed by filter
# of SNPs	~262k	~254k	**8,707**
dbSNP %	90 %	91 %	**48 %**[1]
Transition/Transversion Ratio	2.1	2.1	**1.3**[2]

1. Extremely low dbSNP % indicates that the majority of the SNPs removed by the filter are false positives
2. Transition / transversion ratio deviates far from the expected, further confirming the low quality of the SNPs being removed

In Conclusion…

- The **allelic imbalance** filter removes many false positive SNPs at the expense of 1% of true positives
- The **depth of coverage** filter removes false positives with excessive coverage, while removing very few true positives
- Without these obvious false positives, our final validation set should yield more true SNPs for use in disease association studies

Acknowledgements

- Kiran Garimella, Mark DePristo, & the rest of the Genome Sequencing & Analysis Group
- David Altshuler, Stacey Gabriel & the team working on the 1000 Genomes Project
- SRPG: Lucia Vielma, Eboney Smith, Bruce Birren
- Leslie Roldan for communications training
- Jim Robinson for the Integrated Genome Viewer

Appendix Material & Retired Slides