

MPG NGS workshop I: SNP calling

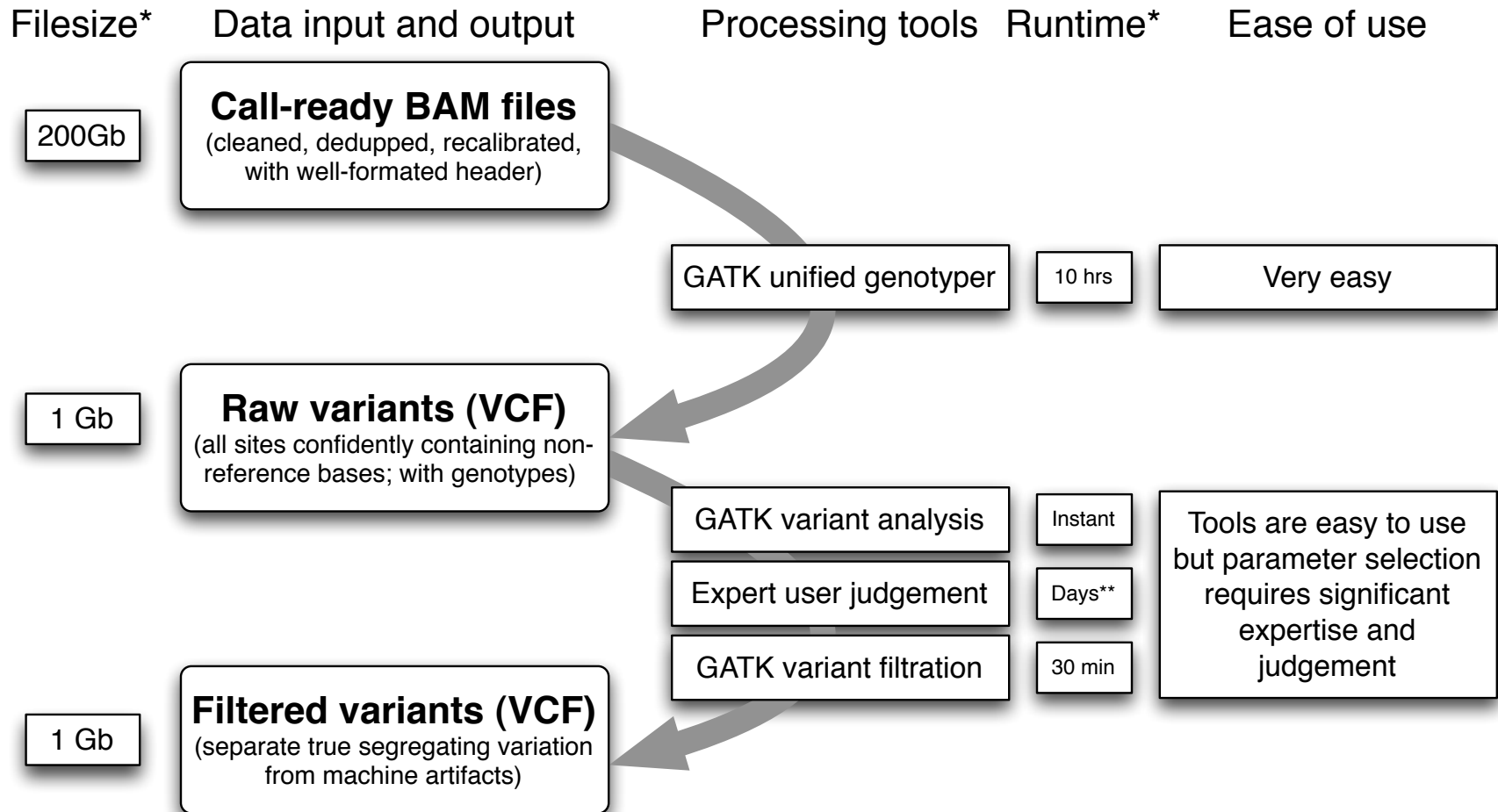
Mark DePristo

Manager, Medical and Population Genetic Analysis
Genome Sequencing and Analysis Group
Medical and Population Genetics Program
Broad Institute of Harvard and MIT

02/04/10

Three slide background on SNP calling in the GATK

SNP calling workflow



* Runtime and file sizes are for a single sample 30x whole genome BAM

** Potentially requires many rounds of experimentation and evaluation

GATK single sample genotype likelihoods

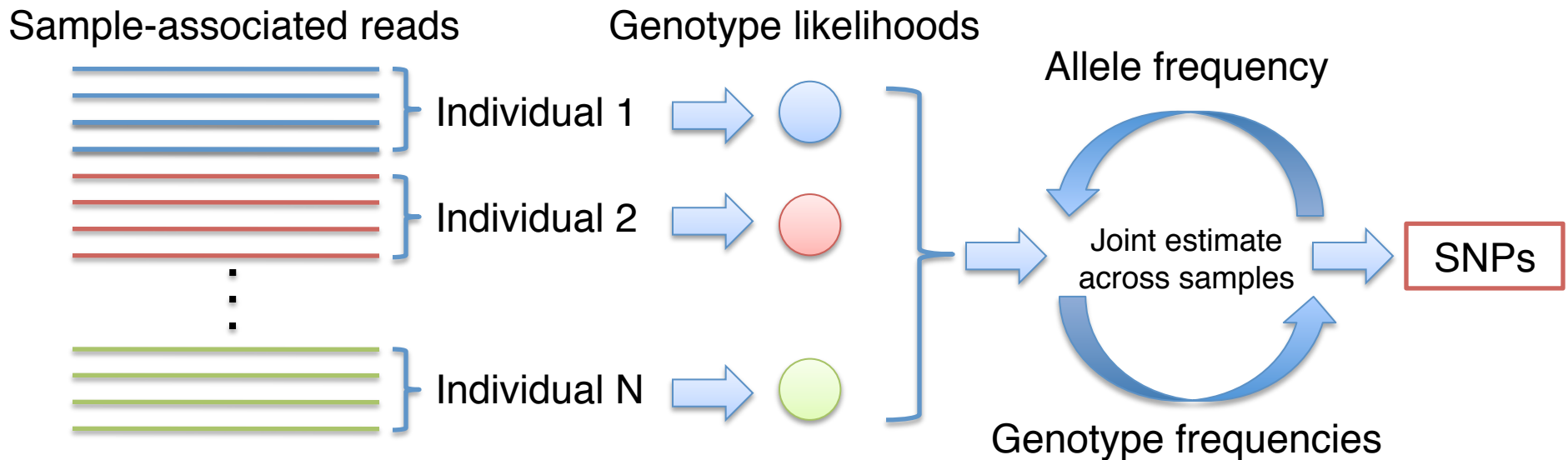
Bayesian model

Likelihood for the genotype Prior for the genotype Likelihood of the data given the genotype Independent base model

$$L(G | D) = P(G)P(D | G) = \prod_{b \in \{good_bases\}} P(b | G)$$

- Priors applied during multi-sample calculation; $P(G) = 1$
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- $P(b | G)$ uses platform-specific confusion matrices
- $L(G/D)$ computed for all 10 genotypes

We apply a generalization of the single sample SNP caller to Pilot 1



- This approach allows us to combine weak single sample calls to discover variation among samples with high confidence

Making raw variant calls with the GATK unified genotyper

Running the Unified Genotyper

```
java -Xmx2048m -jar GenomeAnalysisTK.jar
-R /broad/1KG/reference/human_b36_both.fasta
-T UnifiedGenotyper
-D dbsnp_129_b36.rod
-varout NA19240.raw.vcf
-confidence 50
--heterozygosity 1.000000e-03
-I NA19240.SLX.bam
```

Minimum phred-scaled confidence required to emit a SNP

1 het per 1000 reference bases on average for a Yoruban

BAM file containing NA19240 SLX reads



Long string of variant annotations (more info in a few slides)

Raw VCF calls (NA19240.raw.vcf)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19240
1	36496	.	T	A	53.13	0	<ATTRIBUTES>	GT:DP:GQ	1/0:6:84.70
1	45162	rs10399749	C	T	331.37	0	<ATTRIBUTES>	GT:DP:GQ	0/1:27:99.00
1	48677	.	G	A	399.86	0	<ATTRIBUTES>	GT:DP:GQ	1/0:25:99.00

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information⁷

SNP calling artifacts

- SNP calls are generally infested with false positives
 - From systematic machine artifacts, mismapped reads, aligned indels/CNV
 - Raw SNP calls might have between 5-20% FPs among novel calls
- Separating true variation from artifacts depends very much on the particulars of one's data and project goals
 - Whole genome deep data, WG low-pass, hybrid capture, pooled PCR are have significantly different error modes

Filtering artifacts out of your SNP calls

- The GATK uses a three pass approach
 - First emit all sites potentially containing a true variant
 - Aggregate SNP covariates in the raw VCF to determine the relationship between each covariate and error
[warning: requires user expertise]
 - Finally, apply these filters to the raw VCF using the GATK VariantFiltration tool
- We are currently working on a robust, easy-to-use automated tool

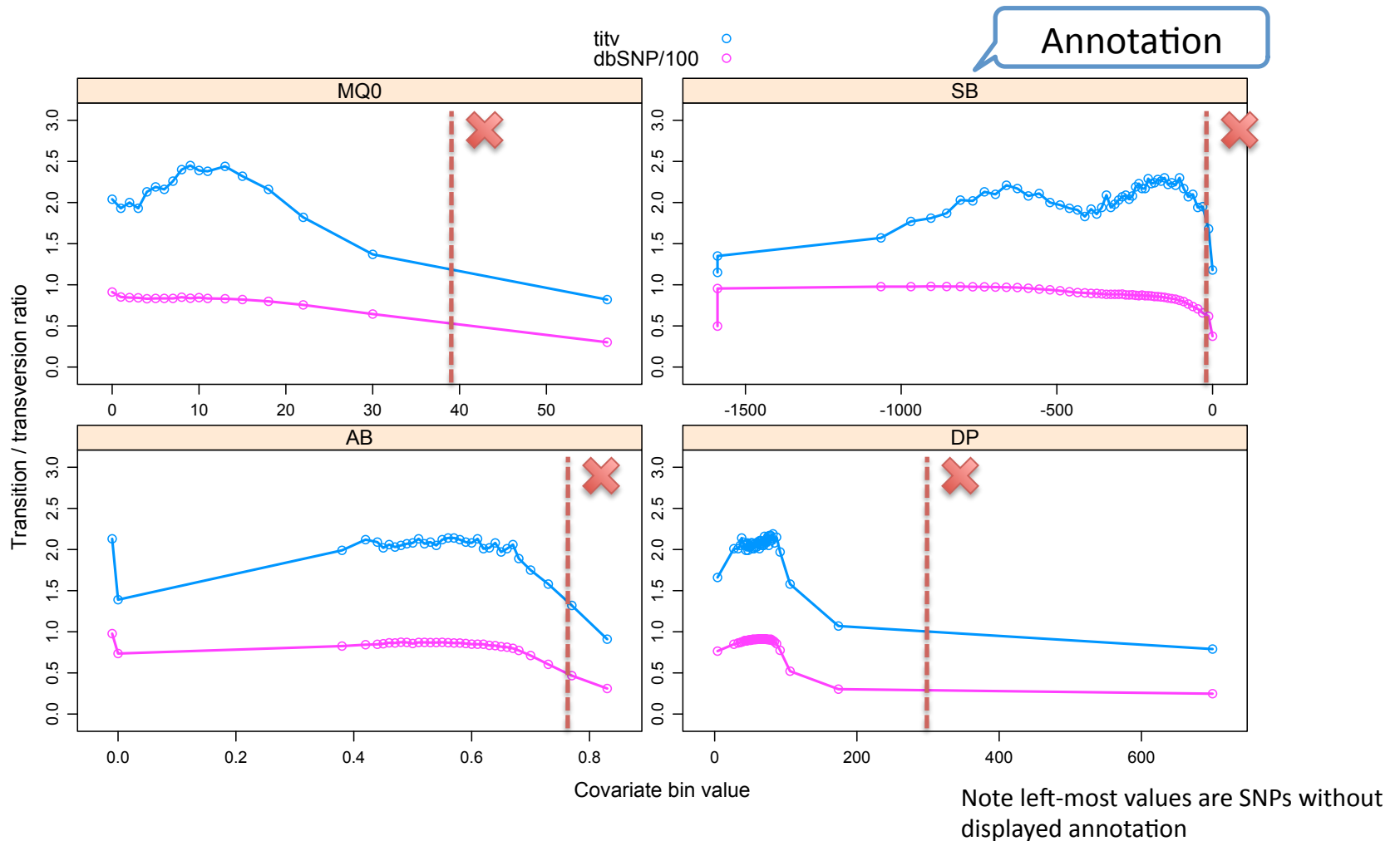
Variant annotations and filters

VCF record for an A/G SNP at 22:49582364

22	49582364	.	A	G	198.96	0
AB=0.67; AC=3; AF=0.50; AN=6; DP=87; DeIs=0.00; HRun=1; MQ=71.31; MQ0=22; QD=2.29; SB=-31.76 GT:DP:GQ						
INFO field	AC	No. chromosomes carrying alt allele		AB	Allele balance of ref/alt in hets	
	AN	Total no. of chromosomes		Hrun	Length of longest contiguous homopolymer	
	AF	Allele frequency		MQ	RMS MAPQ of all reads	
	DP	Depth of coverage		MQ0	No. of MAPQ 0 reads at locus	
	QD	QUAL score over depth		SB	Estimated SB score	
		0/1:12:99.00	0/1:11:89.43	0/1:28:37.78		

Heterozygous genotype A/G in all three individuals

Selecting filtering thresholds



Selected filters are: $AB > 0.75$ || $DP > 300$ || $MQ0 > 40$ || $SB > -0.10$ || 3 snps within 10bp

See <http://www.broadinstitute.org/gsa/wiki/index.php/VariantFiltrationWalker> for more information

Running Variant Filtration

```
java -Xmx2048m -jar GenomeAnalysisTK.jar
-R /broad/1KG/reference/human_b36_both.fasta
-T VariantFiltration
-B variant,VCF,NA19240.raw.vcf
-D dbsnp_129_b36.rod
--clusterWindowSize 10
--filterExpression "AB > 0.75 || DP > 300 || MQ0 > 40 || SB > -0.10"
-l INFO
-o NA19240.filtered.vcf
```

Filters out any group of 3 SNPs
within 10 bp of each other

Expression describing SNPs that
should be filtered out



Filtered VCF calls (NA19240.filtered.vcf)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19240
1	36496	.	T	A	53.13	<u>GATK_FILTER</u>	<ATTRIBUTES>	GT:DP:GQ	1/0:6:84.70
1	45162	rs10399749	C	T	331.37	0	<ATTRIBUTES>	GT:DP:GQ	0/1:27:99.00
1	48677	.	G	A	399.86	0	<ATTRIBUTES>	GT:DP:GQ	1/0:25:99.00

SNPs with poor characteristics have their FILTER field filled in

Raw and filtered autosomal calls for YRI daughter and trio

Call set	Callable bases ¹	# variants	dbSNP%	Ti/Tv (Est. FP rate ²)		Hapmap 3 Sensitivity ³	Hapmap 3 Concordance ³
				Known	Novel		
Single individual calls from the GATK							
Raw NA19240	2.70B (89%)	4.52M	77.83	2.07 (1.9%)	1.81 (18.1%)	99.41	99.85
Filtered NA19240		4.26M	80.42	2.10 (~0.0%)	2.01 (5.6%)	99.14	99.85
Daughter + parents multi-sample calls from the GATK							
Raw YRI trio together	2.5B (81%)	6.24M	71.65	2.07 (1.9%)	1.80 (18.8%)	99.62	99.85
Filtered YRI trio together		5.60M	74.86	2.11 (~0.0%)	2.02 (5.0%)	99.29	99.85

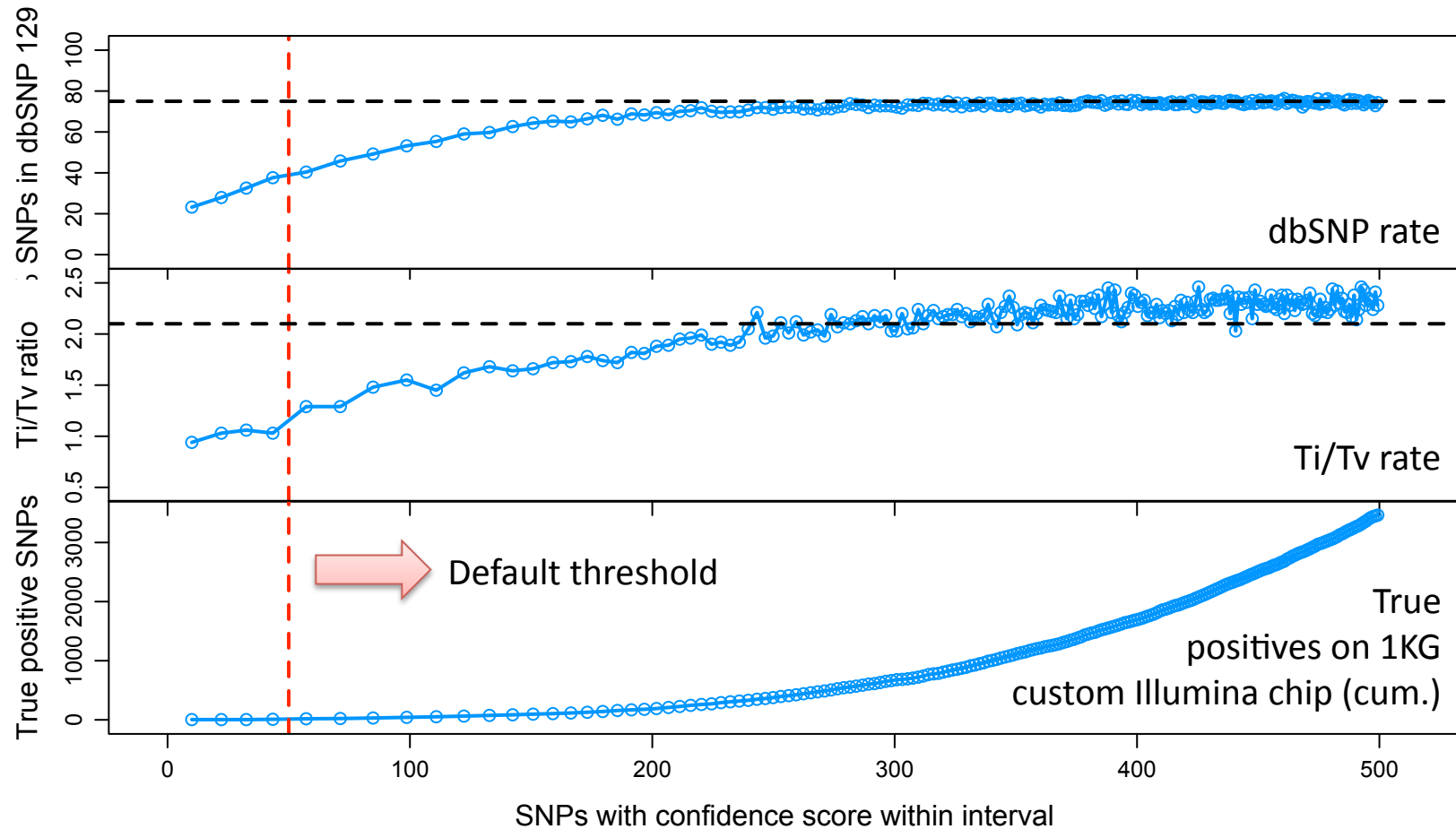
1. % of all 3.1B bases of the B36 human genome called with at least Q50 confidence
2. Calculated as $1 - (\text{titv_Observed} - 0.5) / (\text{titv_Expected} - 0.5)$ with titv_Expected of 2.1
3. NA19240 sensitivity and concordance results

Example scripts

- 1000 Genomes SLX YRI BAM files:
 - Locally available at:
/humgen/gsa-hpprojects/1kg/1kg_pilot2/
useTheseBamsForAnalyses/<sample>.SLX.bam
 - Available for download at 1000genomes.org
- Scripts and VCF files:
 - /humgen/gsa-scr1/pub/tutorials/MPG_workshop

Appendix

Choosing a minimum confidence score for a SNP



- Each point on plot includes ~3000 SNPs from NA19240
- The density of points across the confidence interval indicates the number of SNPs
- ~0.5% of SNPs have $Q < 100$, and only 2% are less than $Q < 200$
- The default Q_{50} threshold results in an highly sensitive call set