GATK Best Practices for Variant Discovery

Callset Evaluation

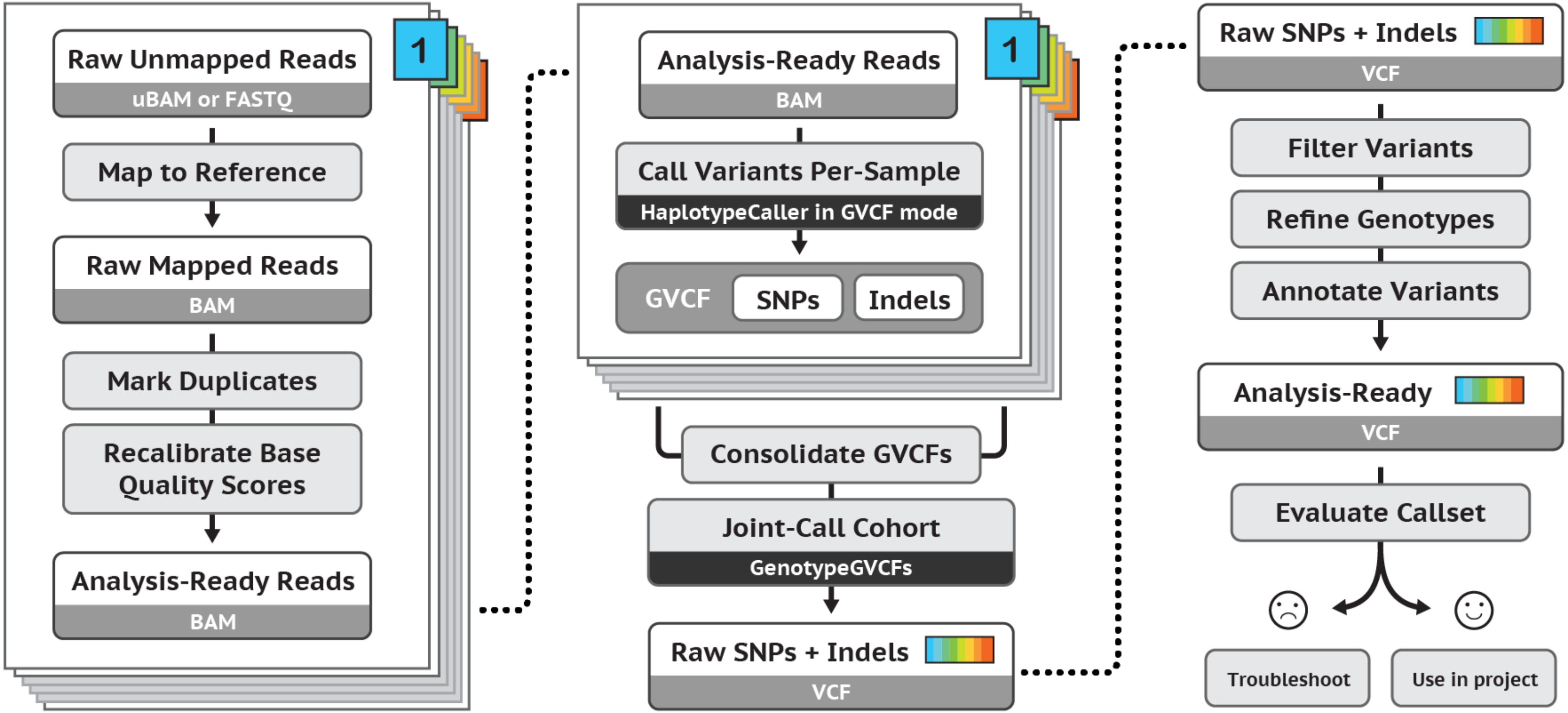
Comparing statistics between your callset and external resources

BROAD INSTITUTE

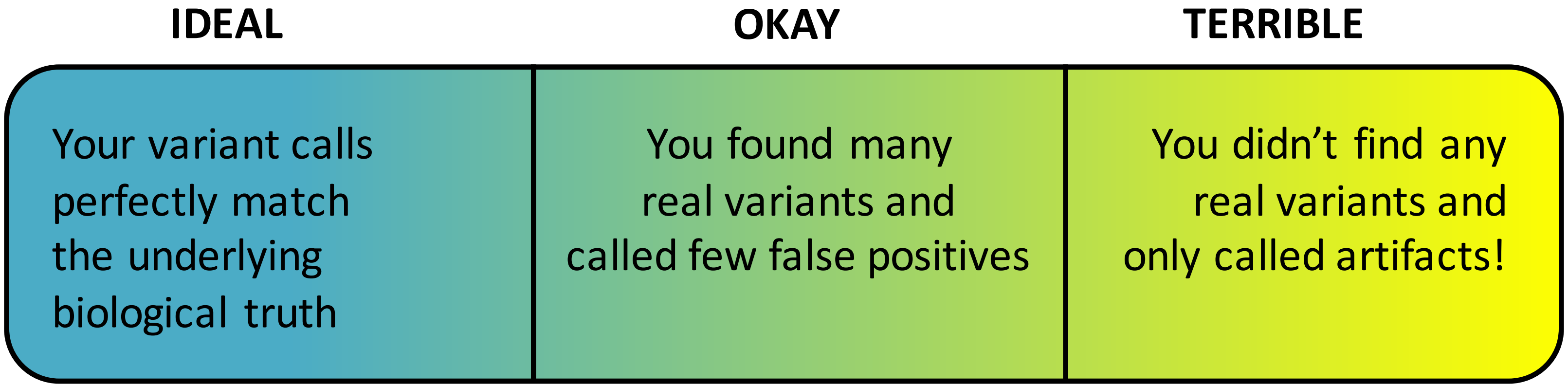
http://software.broadinstitute.org/gatk/

gatk

Best Practices for Germline SNP & INDEL Discovery



What do callset evaluation methods aim to determine?



IDEAL

Your variant calls perfectly match the underlying biological truth

OKAY

You found many real variants and called few false positives

TERRIBLE

You didn’t find any real variants and only called artifacts!

Where are you on this spectrum?

(not veracity of individual variant calls)

How do I figure out how good/bad my callset is?

MyCallset.vcf

....

....

....

Specs:	
# SNPs	
# Indels	
TiTv Ratio	
Indel Ratio	
Concordance	
…	

Compare key statistics from callset and truth set stats

MyCallset Specs	

VS

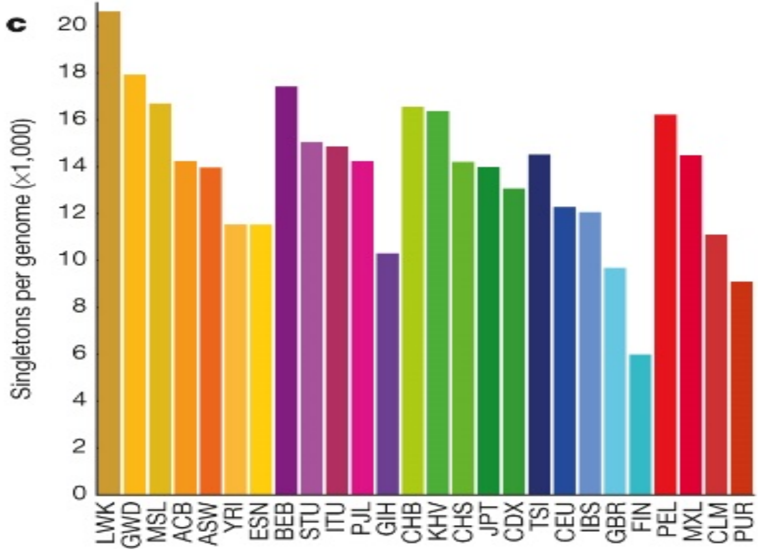
Truth Specs	

Guiding principle: divergence is indicative of error

Key assumption: truth set is representative / comparable

- Important to match dataset properties!
  - Population ethnicity (European, African, etc.)
  - Sequencing / exp. design (WGS vs. WES)
  - Cohort size

Not easy! You might need to use sub-cohorts (of both sets) to match all three.



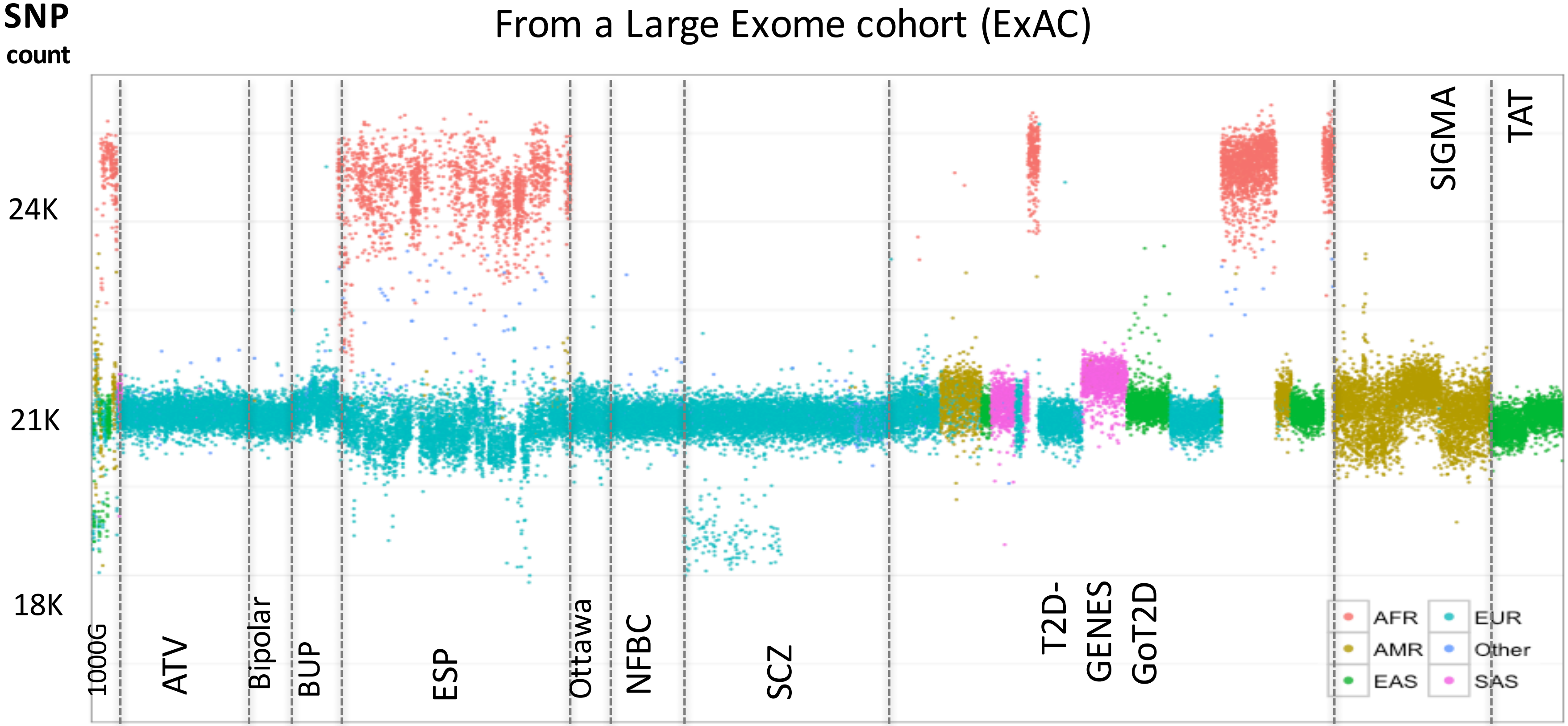
http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html

Commonly used truth sets

- **dbSNP**
  All previously reported variation (lots of junk!)
- **ExAC and GnomAD**
  Extensive catalog of human variation built by aggregating results from many studies
- **HapMap**
  Highly validated common human variants
- **OMNI**
  Common variation validated by array
- **NIST's Genomes in a Bottle, or Illumina's Platinum Genomes**
  high confidence callsets from a handful of common benchmarking samples

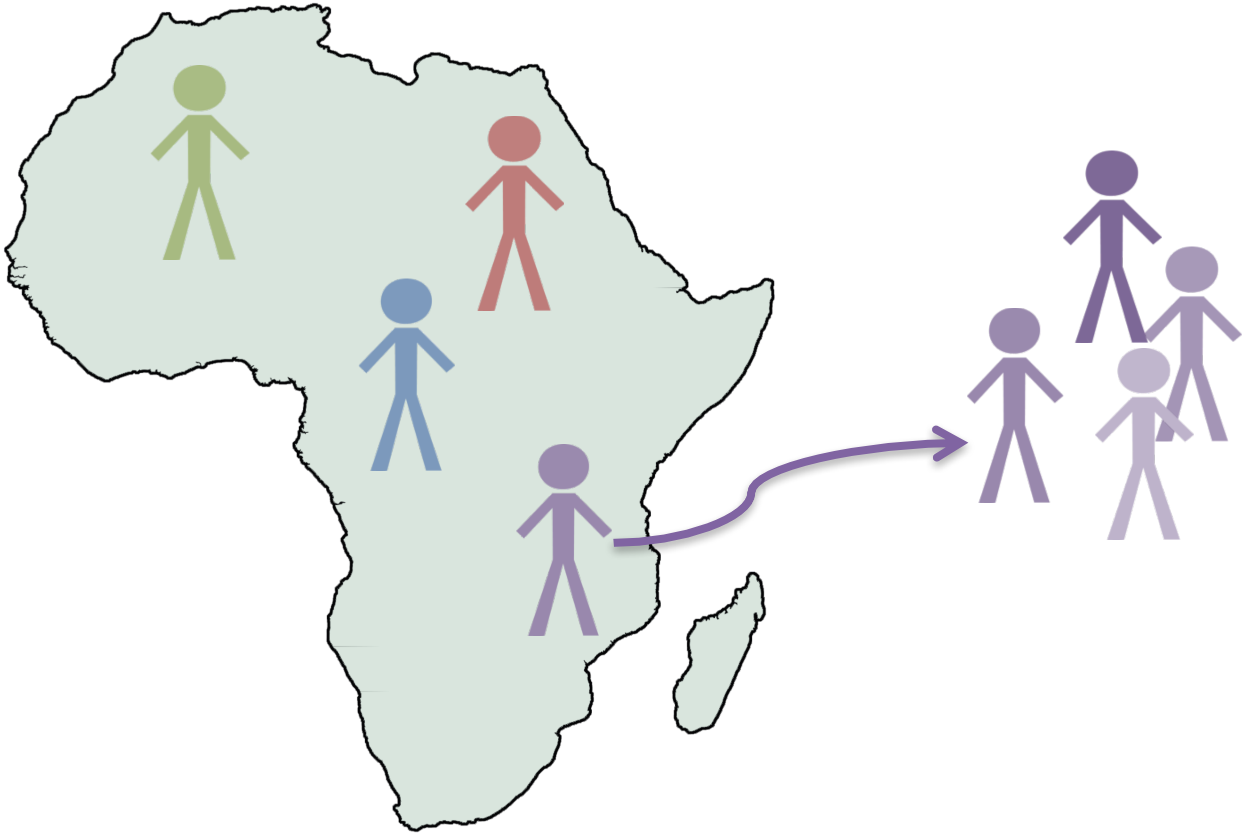
Ethnicity affects many variant call metrics

From a Large Exome cohort (ExAC)



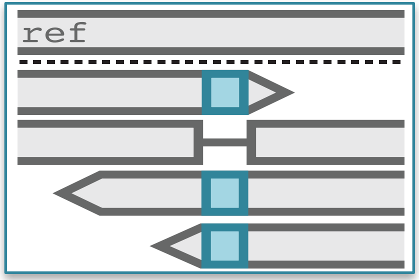
Monkol Lek, 2014

Older populations tend to display more heterogeneity

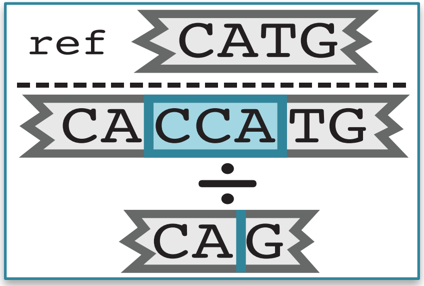


Recommended metrics for callset evaluation

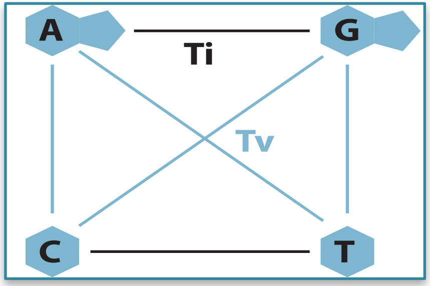
Number of Indels & SNPs



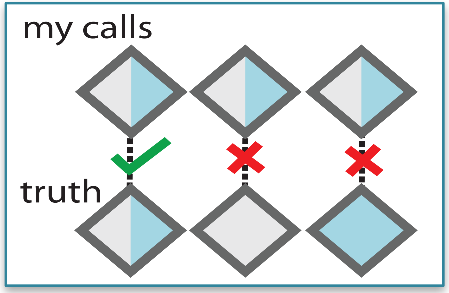
Indel Ratio



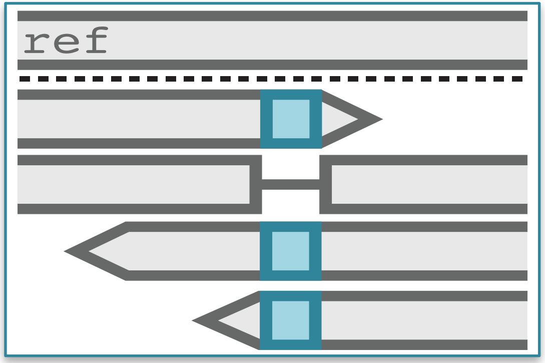
TiTv Ratio



Genotype Concordance



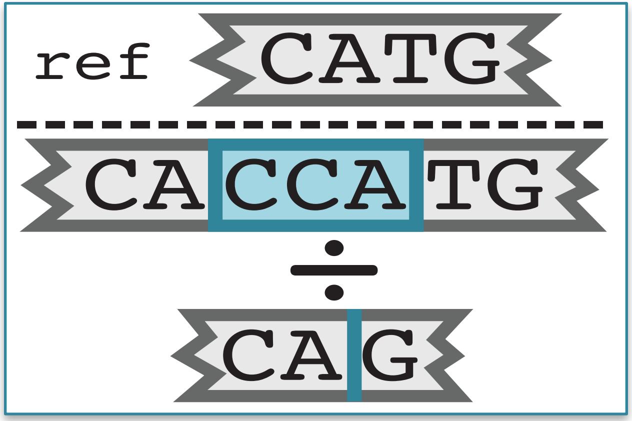
Number of Indels & SNPs



Sequencing Type	# of Variants (in 1 sample)
WGS	~4.4 M
WES	~21 k

- Variants = Indels + SNPs
- Useful for order-of-magnitude sanity check
- Vary by size and diversity of cohort

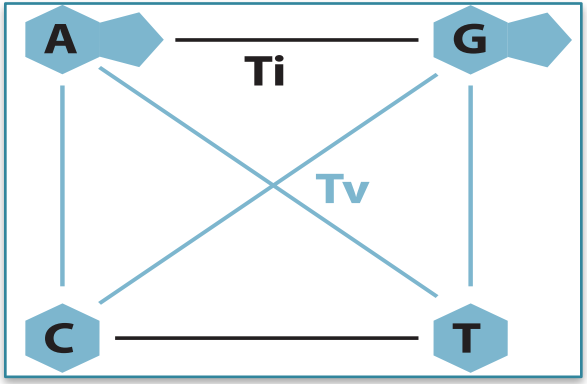
Indel Ratio



Variant prevalence	Indel Ratio
Common	~1
Rare	0.2-0.5

- Ratio of **in**sertions to **del**etions
- Varies by allele frequency: common (”known”) vs. rare (“novel”)

TiTv Ratio (Transitions/Transversions)

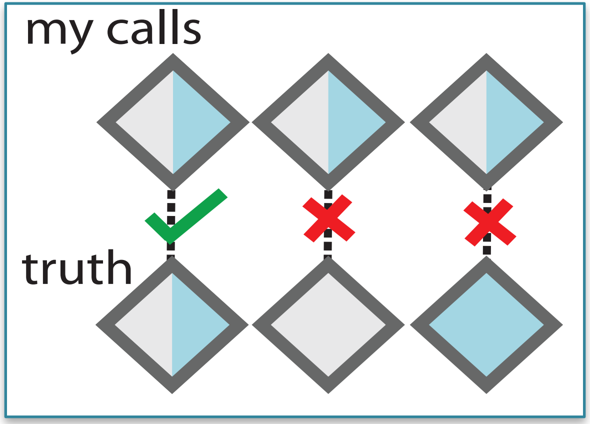


Sequencing Type	TiTv Ratio
WGS	2.0-2.1
WES	3.0-3.3

In Humans…

- Used for SNPs only
- If variation were random: expect ratio of 0.5 as there are twice as many possible transversions vs transitions!
- Low TiTv ratio indicates high rate of false positives

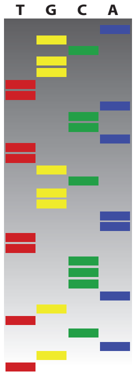
Genotype Concordance

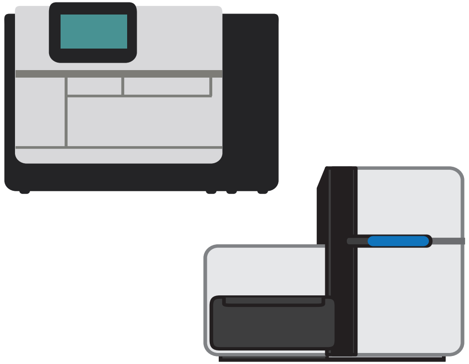


- Most appropriate truth set is genotyping chip for same sample
- % Genotype calls in callset that match calls in truth set
- Unmatched variants considered false positives

For evaluating concordance, it is best to use truth sets generated with orthogonal methods

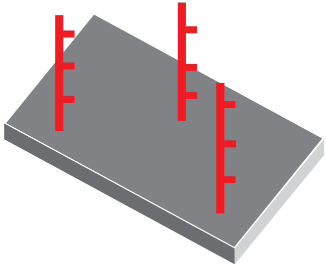
SEQUENCING





- Sanger sequencing
- Other HTS technologies

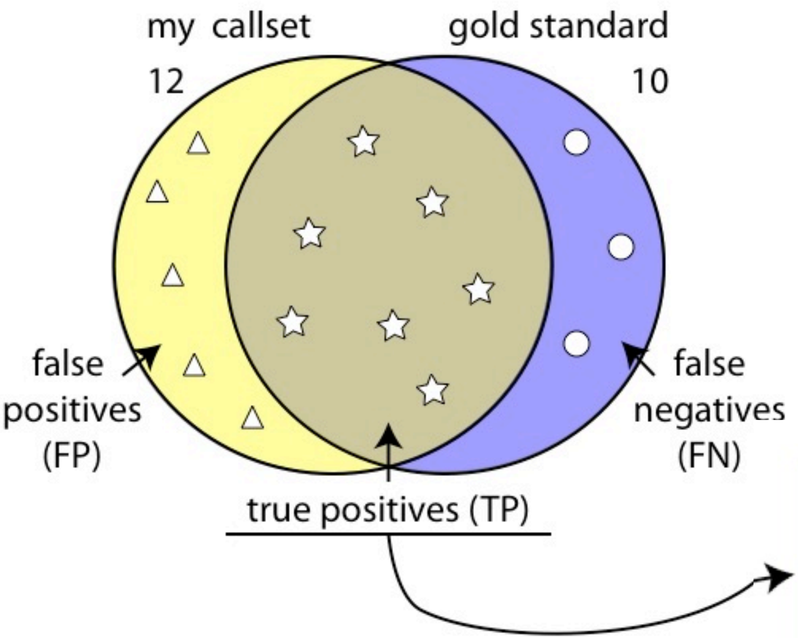
PROBE/ARRAY-BASED



- GeneChip
- Microarrays

Cheat sheet of concordance metrics

Variant Concordance

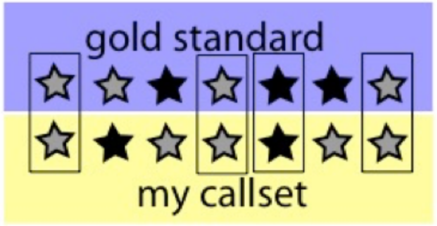


SENSITIVITY

$$\frac{TP}{TP + FN} = \frac{7}{7 + 3} = 70\%$$

FALSE DISCOVERY RATE

$$\frac{FP}{FP + TP} = \frac{5}{5 + 7} = 42\%$$



GT CONCORDANCE

$$\frac{\sum \text{matches}}{TP} = \frac{4}{7} = 57\%$$

So how do I get these metrics?

	Variant Level Evaluation	Genotype Level Evaluation

GATK	VariantEval
	`java -jar GenomeAnalysisTK.jar \`
	`-T VariantEval \`
	`-R reference.b37.fasta \`
	`-eval callset.vcf \`
	`--comp truthset.vcf \`
	`-o results.eval.grp`
		GenotypeConcordance
		`java -jar GenomeAnalysisTK.jar \`
		`-T GenotypeConcordance \`
		`-R reference.b37.fasta \`
		`--comp truthset.vcf \`
		`--eval callset.vcf \`
		`-o results.grp`

Picard	CollectVariantCallingMetrics (CVCM)
	`java -jar picard.jar \`
	`CollectVariantCallingMetrics`
	`INPUT=callset.vcf \`
	`DBSNP=truthset.vcf \`
	`OUTPUT=results`
		GenotypeConcordance
		`java -jar picard.jar \`
		`GenotypeConcordance \`
		`CALL_VCF=callset.vcf \`
		`TRUTH_VCF=truthset.vcf \`
		`CALL_SAMPLE=sampleName \`
		`TRUTH_SAMPLE=sampleName \`
		`OUTPUT=results`

- Genotype level evaluation tools equivalent—these tools will be merged in GATK4
- Variant level evaluation tools are different

Which variant-level evaluator should I use?

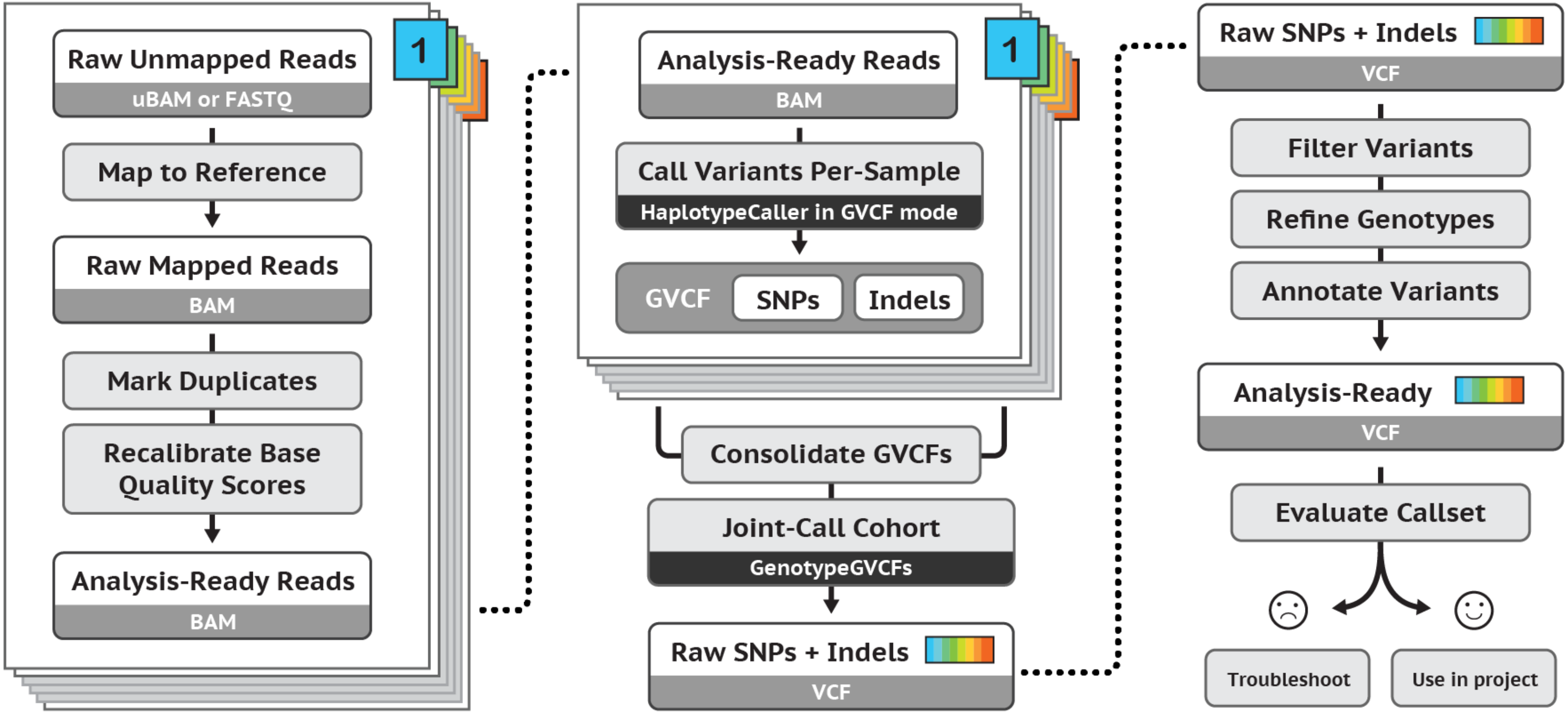
GATK VariantEval

- More detailed analysis
- More options for stratification
- Ability to compare to multiple truth sets

Picard CVCM

- Best performance on very large callsets
- Ability to interpret no-call as confident reference in a “confidence region”
- Few options beyond the metrics discussed here

Best Practices for Germline SNP & INDEL Discovery


Raw Unmapped Reads
uBAM or FASTQ
1
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
Analysis-Ready Reads
BAM
1
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project