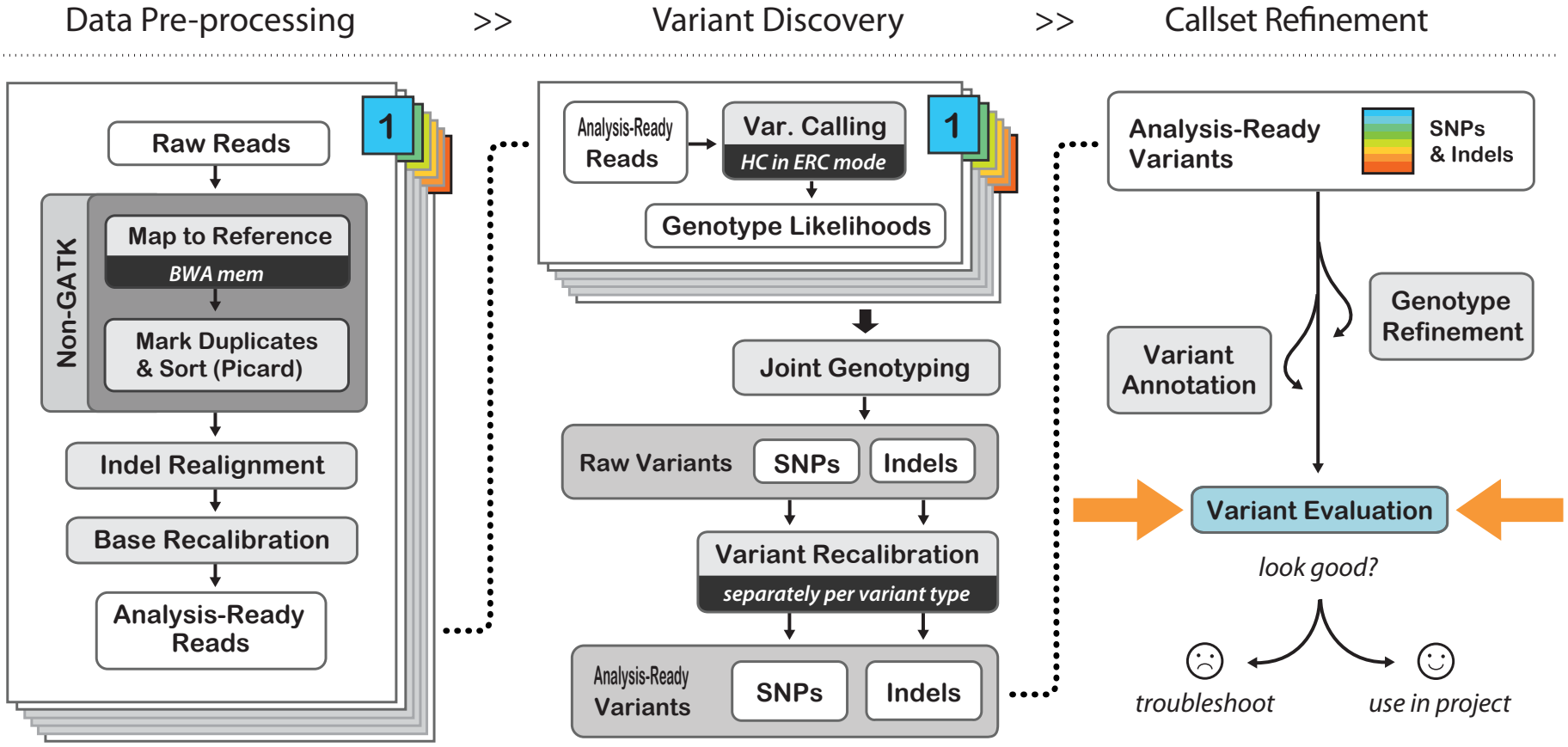
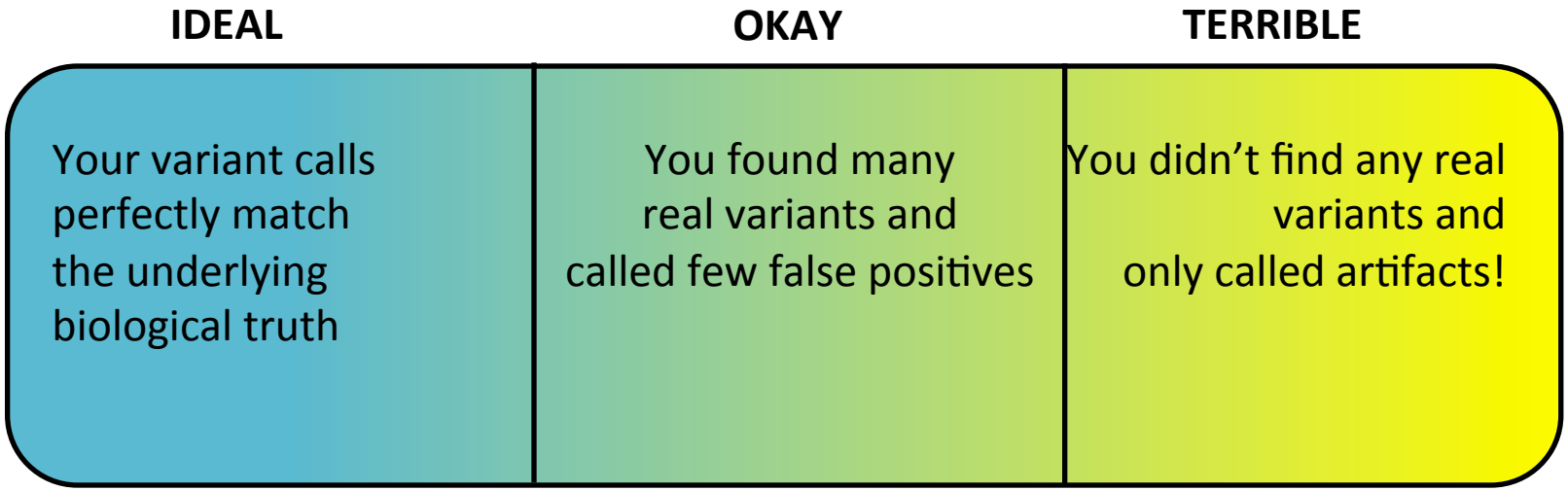
Callset Evaluation

Comparing statistics between your callset and a truth set

You are here in the GATK Best Practices workflow for **germline variant discovery**



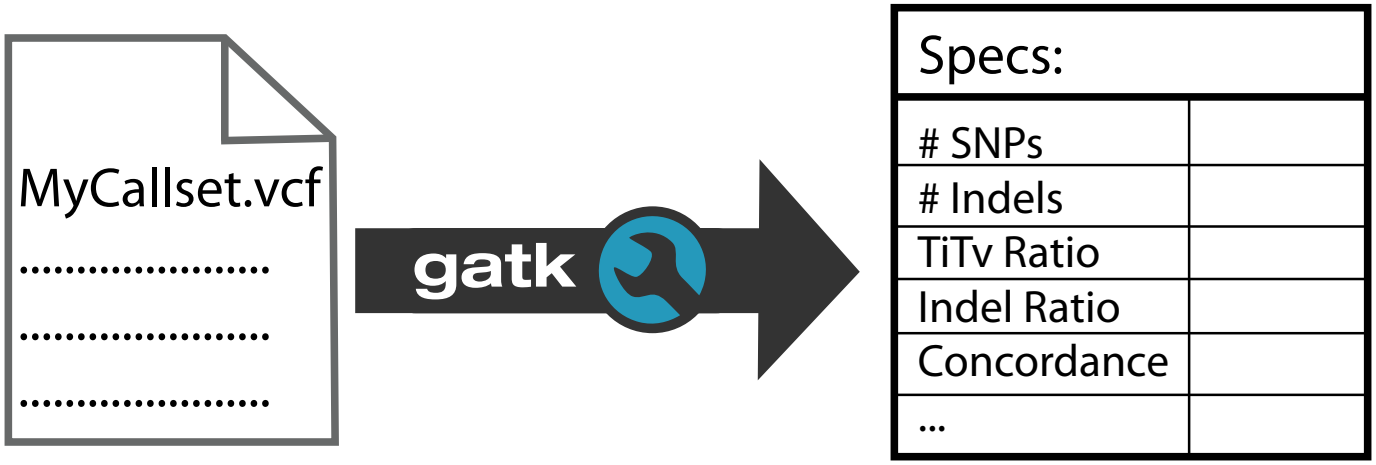
Where are you on this spectrum?



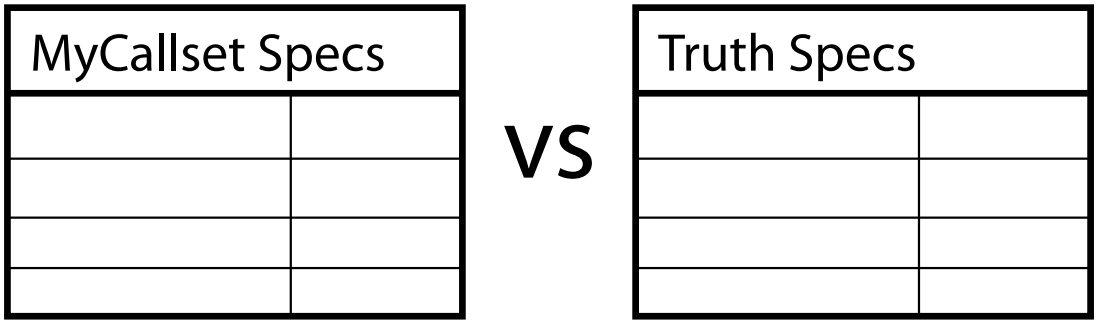
= What callset evaluation methods aim to determine

(not veracity of individual variant calls)

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



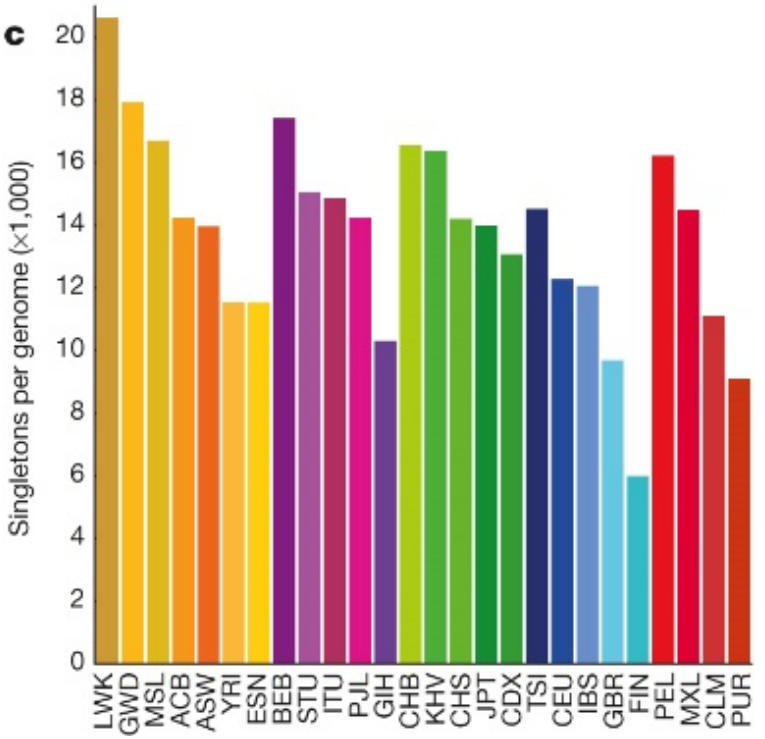
Extract key statistics and compare to truth set stats



Guiding principle: divergence is indicative of error

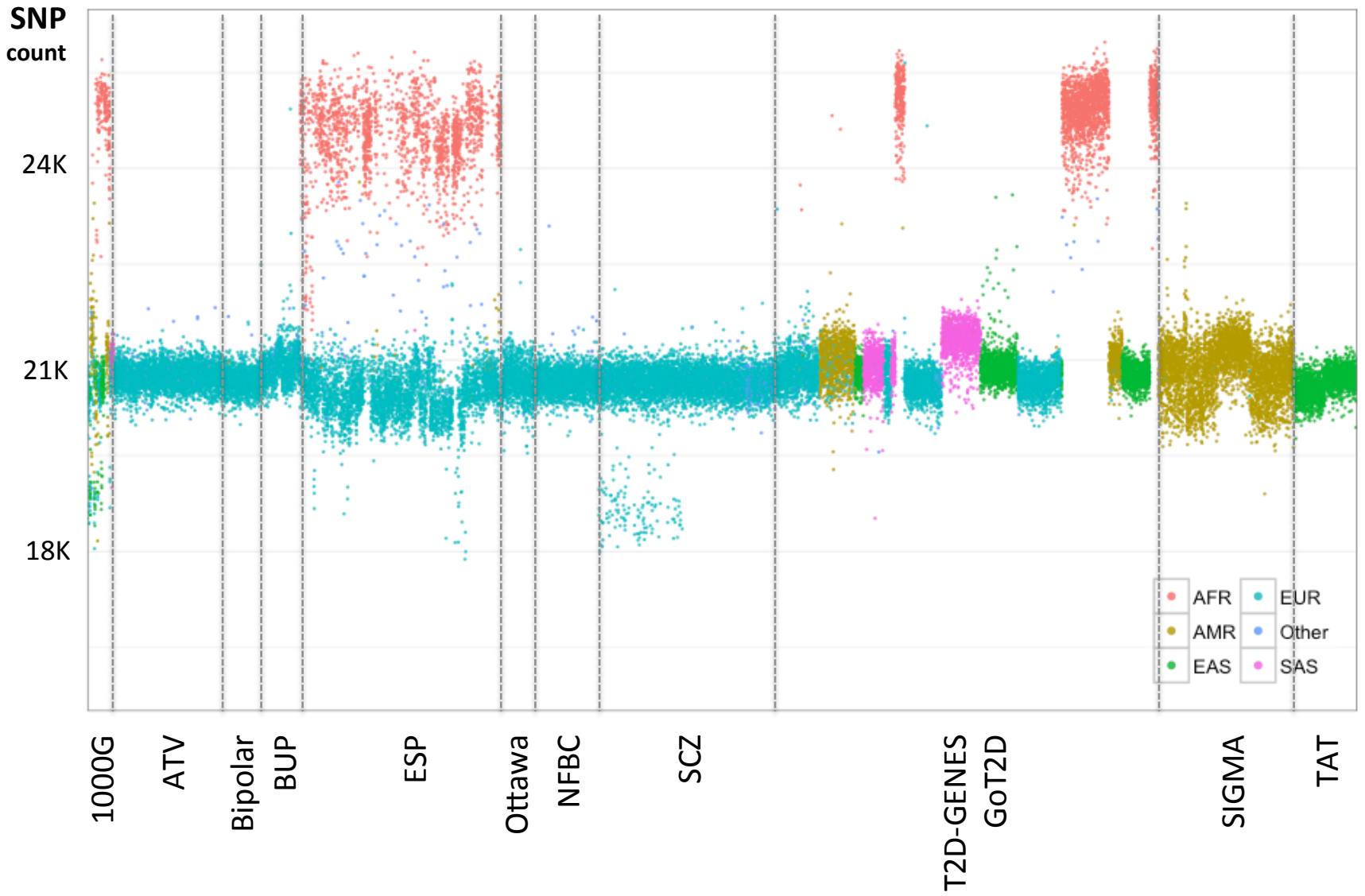
Key assumption: truth set is representative / comparable

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



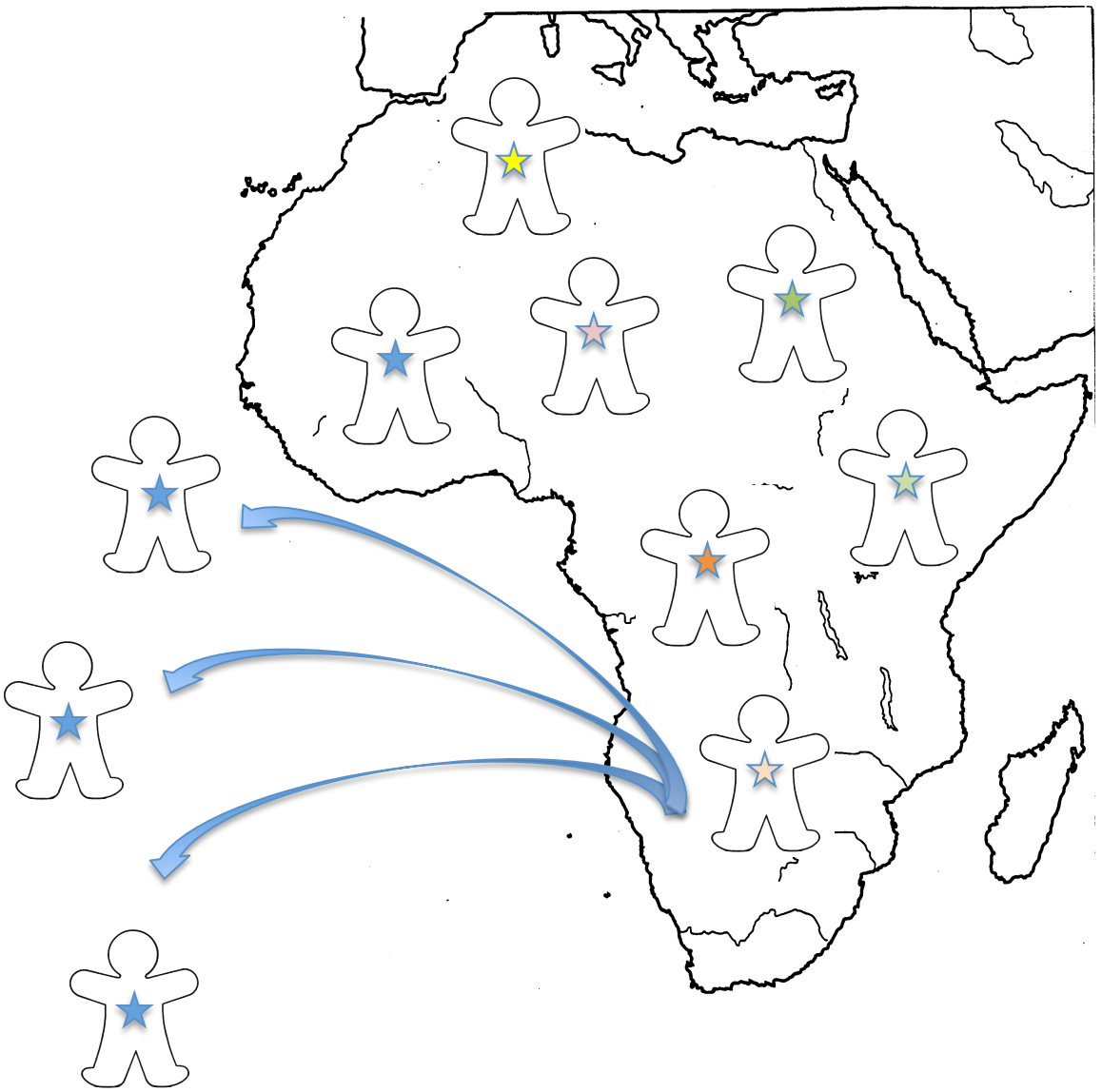
http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html

Ethnicity affects many variant call metrics



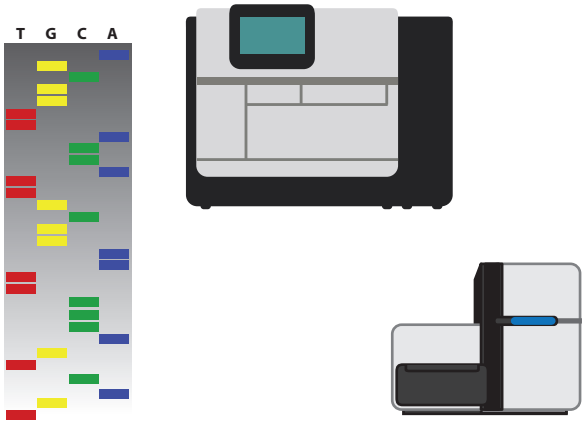
Monkol Lek, 2014

Older populations tend to display more heterogeneity



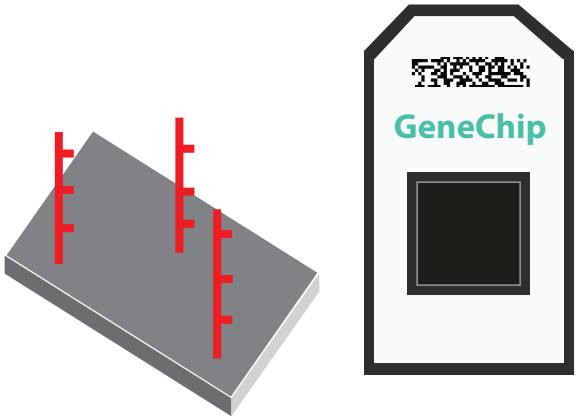
If possible, use truth sets generated with orthogonal methods

Sequencing



- Sanger sequencing
- Other HTS technologies

Probe/Array-based



- GeneChip
- Microarrays

Commonly used truth sets

- **dbSNP**

  All previously reported variation (lots of junk!)
- **Sample-matched genotyping chip**

  Awesome! But adds cost & limited to known variants
- **HapMap**

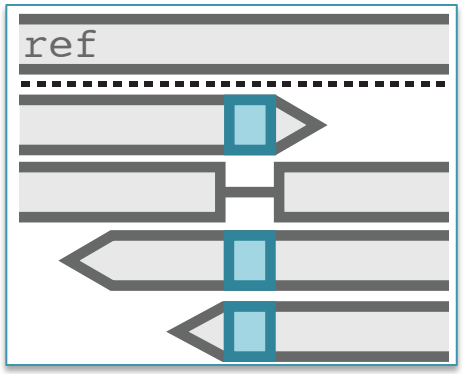
  Highly validated common human variants
- **OMNI**

  Common variation validated by array
- **NIST Genomes in a Bottle (single sample evaluation)**

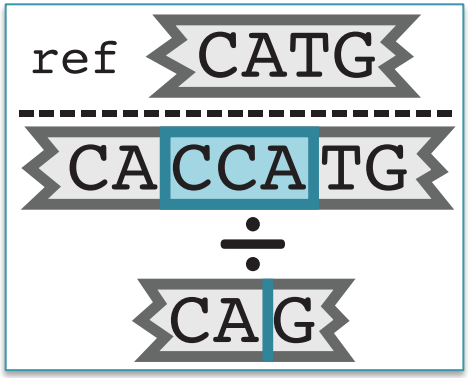
  Consensus callsets from common benchmarking samples

Recommended metrics for callset evaluation

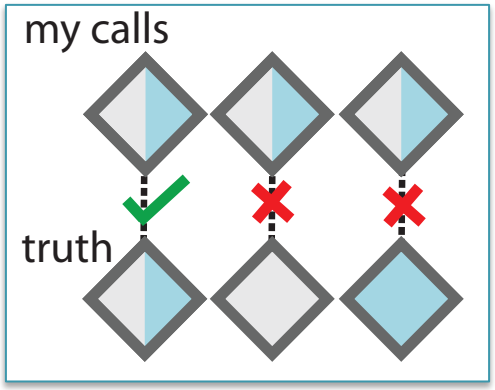
Number of Indels & SNPs



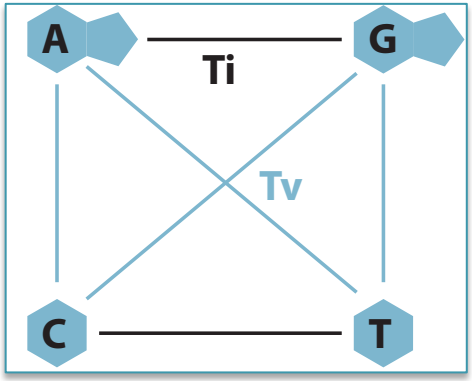
Indel Ratio



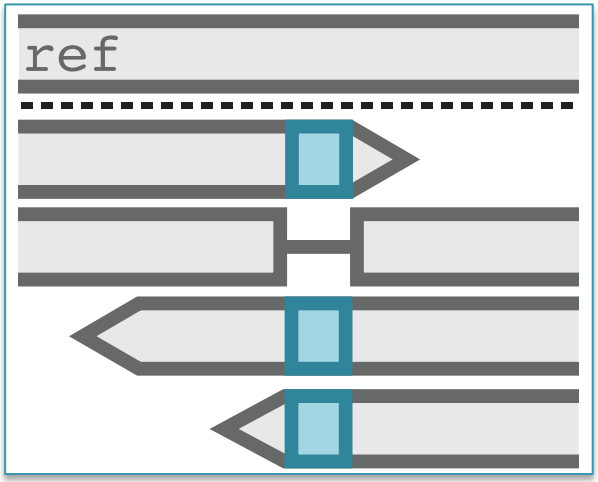
Genotype Concordance



TiTv Ratio



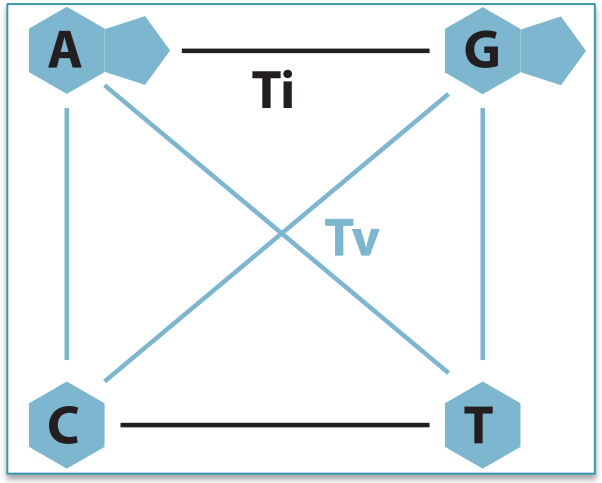
Number of Indels & SNPs



Sequencing Type	# of Variants
WGS	~4.4 M
WES	~41 k

- Variants = Indels + SNPs
- Useful for order-of-magnitude sanity check

TiTv Ratio

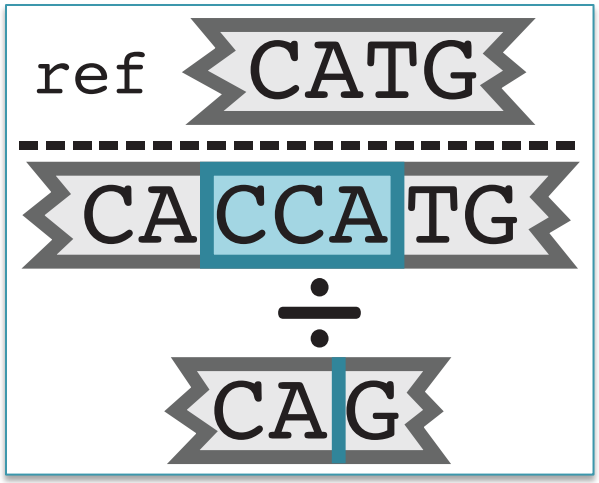


Sequencing Type	TiTv Ratio
WGS	2.0-2.1
WES	3.0-3.3

- If random: expect ratio of 0.5

  Twice as many possible transversions vs transitions!
- Low TiTv ratio indicates high rate of false positives

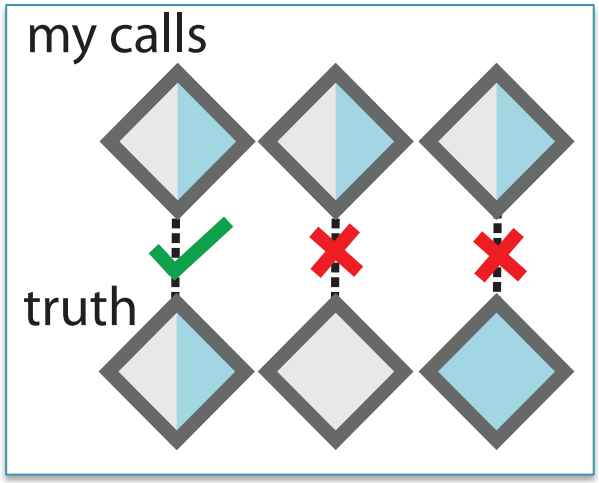
Indel Ratio



Variant Association Study type	Indel Ratio
Common	~1
Rare	0.2-0.5

- Ratio of **in**sertions to **del**etions
- Varies by type of study
  e.g. rare variant association *vs* common variant association

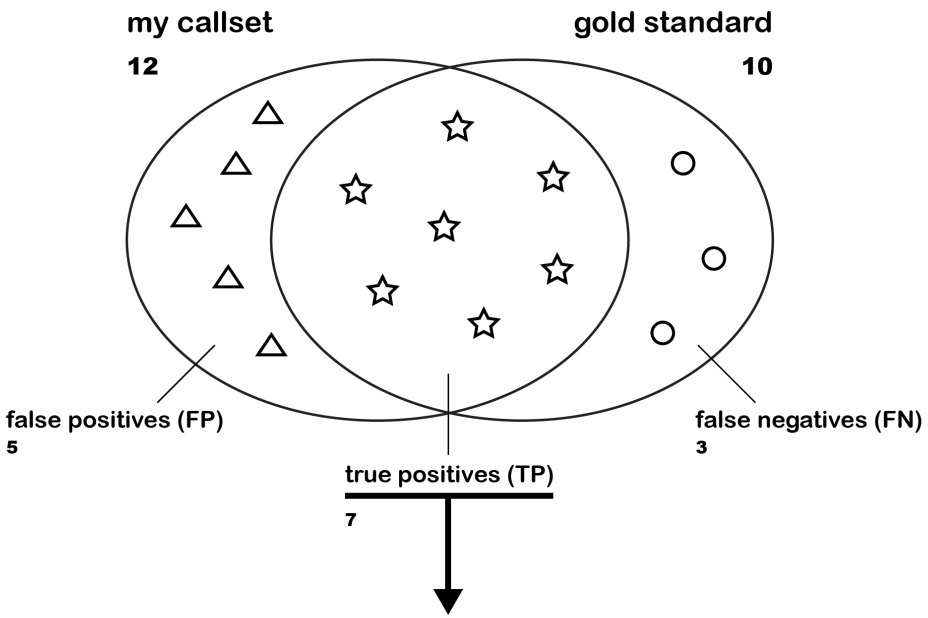
Genotype Concordance



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

Cheat sheet of concordance metrics

SENSITIVITY vs. FALSE DISCOVERY RATE



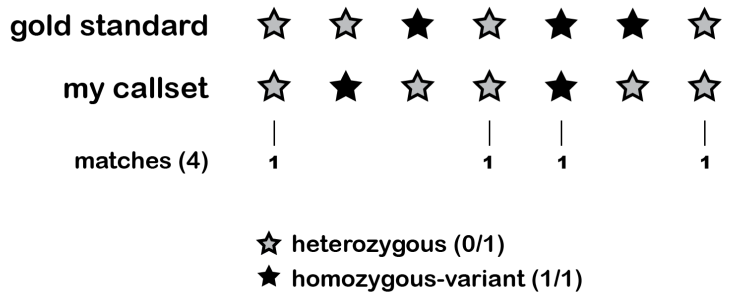
SENSITIVITY

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

FALSE DISCOVERY RATE

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

GENOTYPE CONCORDANCE



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 \`
	`-D 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
	`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`

Which variant-level evaluator should I use?

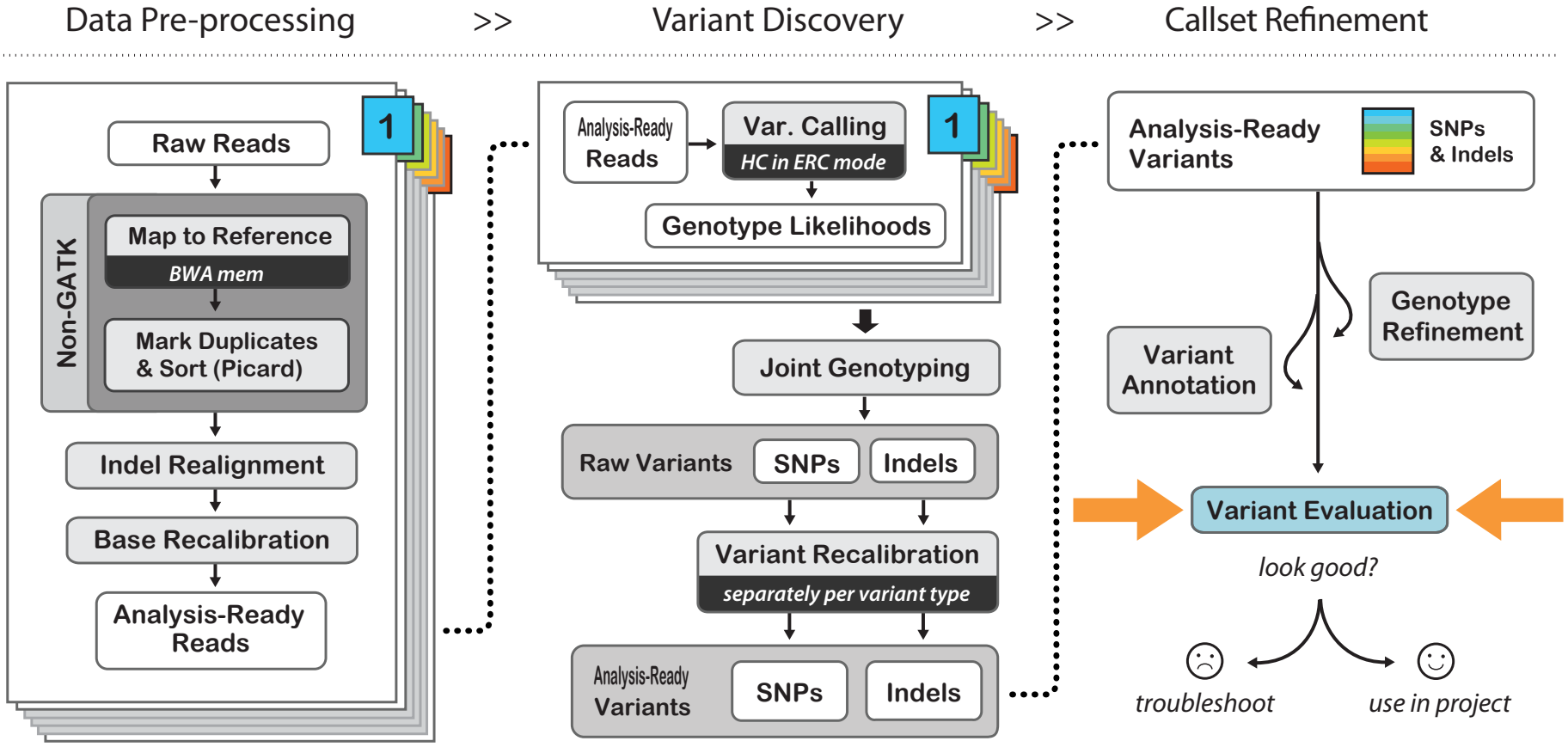
GATK
VariantEval

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

Picard CollectVariantCallingMetrics

- Best performance & speed on very large callsets
- Few options beyond the metrics discussed here

You are here in the GATK Best Practices workflow for **germline variant discovery**



Further reading

http://www.broadinstitute.org/gatk/guide/article?id=6308

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