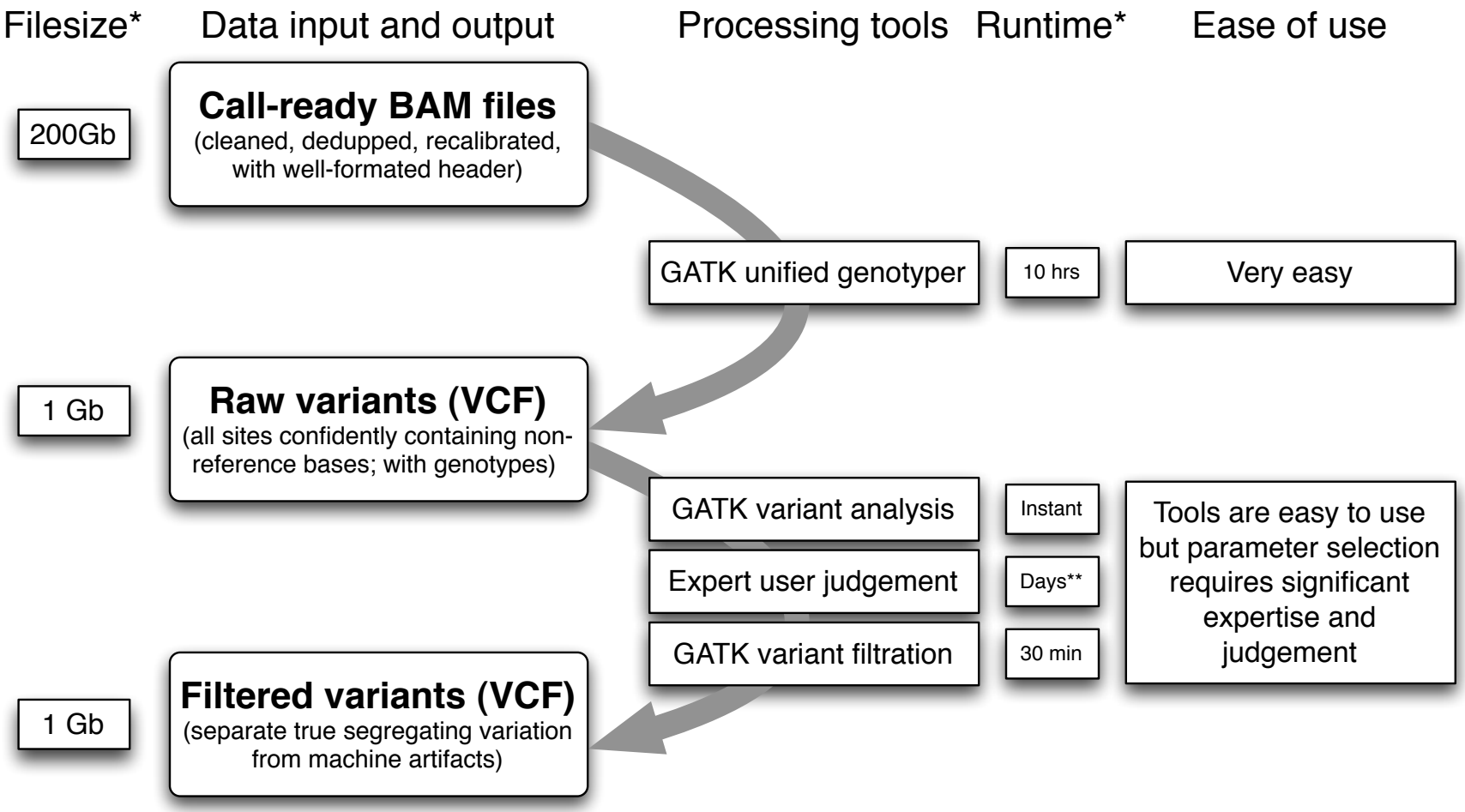
MPG NGS workshop I: Quality assessment of SNP calls

Kiran V Garimella (kiran@broadinstitute.org)
Genome Sequencing and Analysis
Medical and Population Genetics
February 4, 2010

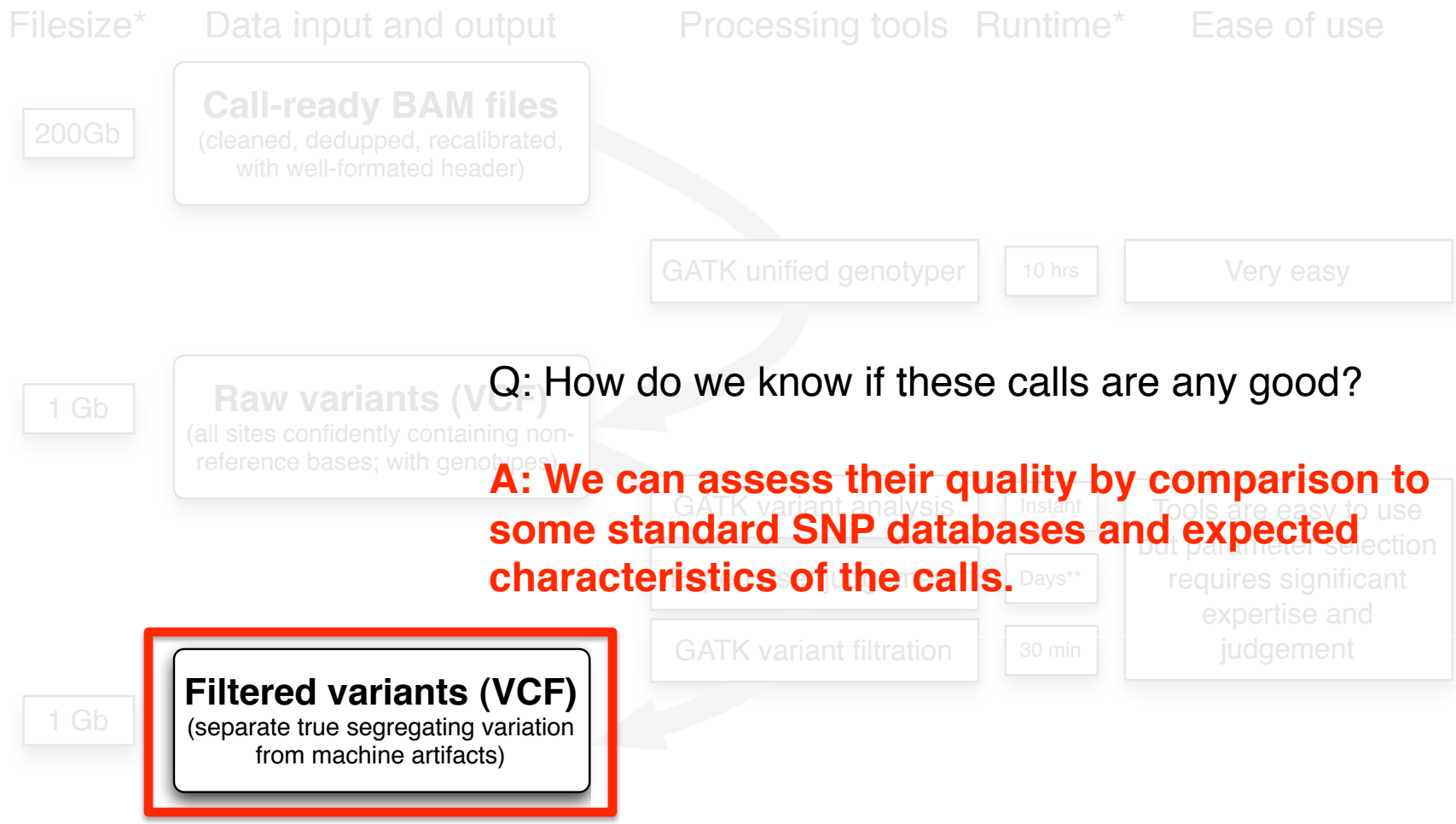
SNP calling workflow



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

** Potentially requires many rounds of experimentation and evaluation

SNP calling workflow



Q: How do we know if these calls are any good?

A: We can assess their quality by comparison to some standard SNP databases and expected characteristics of the calls.

* Runtime and file sizes are for a single sample 30x whole genome BAM
** Potentially requires many rounds of experimentation and evaluation

Common indicators of variant quality

- Metrics
  - Number of variants
  - Transition/transversion ratio
  - Hardy-Weinberg Equilibrium violations
- Presence in variant databases
  - % in dbSNP build 129/130
  - Concordance to Hapmap (II, II+III consensus, III)
  - Concordance to validation data (i.e. array-based genotyping)
- Visualization (IGV*)
  - Examination of alignments, local sequence context, error covariates, etc.

*** See http://www.broadinstitute.org/igv/ for more details.**

VariantEval: rapid assessment of SNP quality metrics and presence in external databases

```
java -Xmx2048m -jar GenomeAnalysisTK.jar \
     -T VariantEval \
     -R /broad/1KG/reference/human_b36_both.fasta \
     -L 1 \
     -D dbsnp_129_b36.rod \
     -G \
     -A \
     -V \
     -B eval,VCF,NA19240.filtered.vcf \
     -l INFO \
     -o NA19240.filtered.eval
```

the module to run

Indicates that your VCF file has genotypes

Print extended evaluation information

Print a list of “interesting” sites (FPs, FNs, etc).

Running time for ~270K variants in a ~247mb region: **1 minute**

NA19240.filtered.eval: dbSNP section for all variants

```
all,summary,db_coverage       Analysis name      db_coverage
all,summary,db_coverage       Analysis params
all,summary,db_coverage       Analysis class     ...varianteval.VariantDBCoverage
all,summary,db_coverage       Analysis time      Tue Feb 02 17:11:15 EST 2010
all,summary,db_coverage       name                    dbSNP
all,summary,db_coverage       n_db_snps               966213
all,summary,db_coverage       n_eval_sites            269661
all,summary,db_coverage       n_overlapping_sites     222357
all,summary,db_coverage       n_concordant            222118
all,summary,db_coverage       n_novel_sites           47304
all,summary,db_coverage       dbsnp_rate              82.46       # percent eval snps at dbsnp snps
all,summary,db_coverage       concordance_rate        99.89
... (~,1000 more lines)
```

See http://www.broadinstitute.org/gsa/wiki/index.php/VariantEval for more detail.

Selected VariantEval output

```
all,summary,variant_counts                  variants             269661
all,summary,variant_counts                  heterozygotes        178616
all,summary,variant_counts                  homozygotes          91045
all,summary,db_coverage                     n_novel_sites        47304
all,summary,db_coverage                     dbsnp_rate           82.46       # percent eval snps at dbsnp snps
all,summary,db_coverage                     concordance_rate     99.89
all,summary,genotype_concordance            name                 hapmap-chip
all,summary,genotype_concordance            TRUTH_STATE   CALLED_REF   CALLED_VAR_HET   CALLED_VAR_HOM  NO_CALL...
all,summary,genotype_concordance            IS_REF               0      40               0               58580
all,summary,genotype_concordance            IS_VAR_HET           0      35098            35              400
all,summary,genotype_concordance            IS_VAR_HOM           0      42               23557           322
all,summary,genotype_concordance            UNKNOWN              0      143436           67453           0
all,summary,genotype_concordance            VARIANT_SENSITIVITY  98.79%
all,summary,genotype_concordance            VARIANT_CONCORDANCE  99.87%
all,summary,genotype_concordance            OVERALL_CONCORDANCE  99.80%
all,summary,transitions_transversions       ratio                2.15

known,summary,variant_counts                variants             222357
known,summary,transitions_transversions     ratio                2.16

novel,summary,variant_counts                variants             47304
novel,summary,transitions_transversions     ratio                2.12

filtered,summary,variant_counts             variants             88616
filtered,summary,transitions_transversions  ratio                1.32
```

summary for all variants

summary for known variants

summary for novel variants

summary for filtered variants

VariantEval computes relevant quality metrics for interesting SNP subsets (all, known, novel, filtered), making it easy to evaluate their aggregate performance.

See http://www.broadinstitute.org/gsa/wiki/index.php/VariantEval for more detail.

CallsetConcordance: compute the overlap and disjoint for N-callsets

```
java -Xmx2048m -jar GenomeAnalysisTK.jar \
     -T CallsetConcordance \
     -L 1 \
     -R /broad/1KG/reference/human_b36_both.fasta \
     -CT NWayVenn \
     -B callset1,VCF,NA19240.SeqCenterA.vcf \
     -B callset2,VCF,NA19240.SeqCenterB.vcf \
     -CO NA19240.venn.vcf
```

the module to run

compare N callsets

callsets to compare

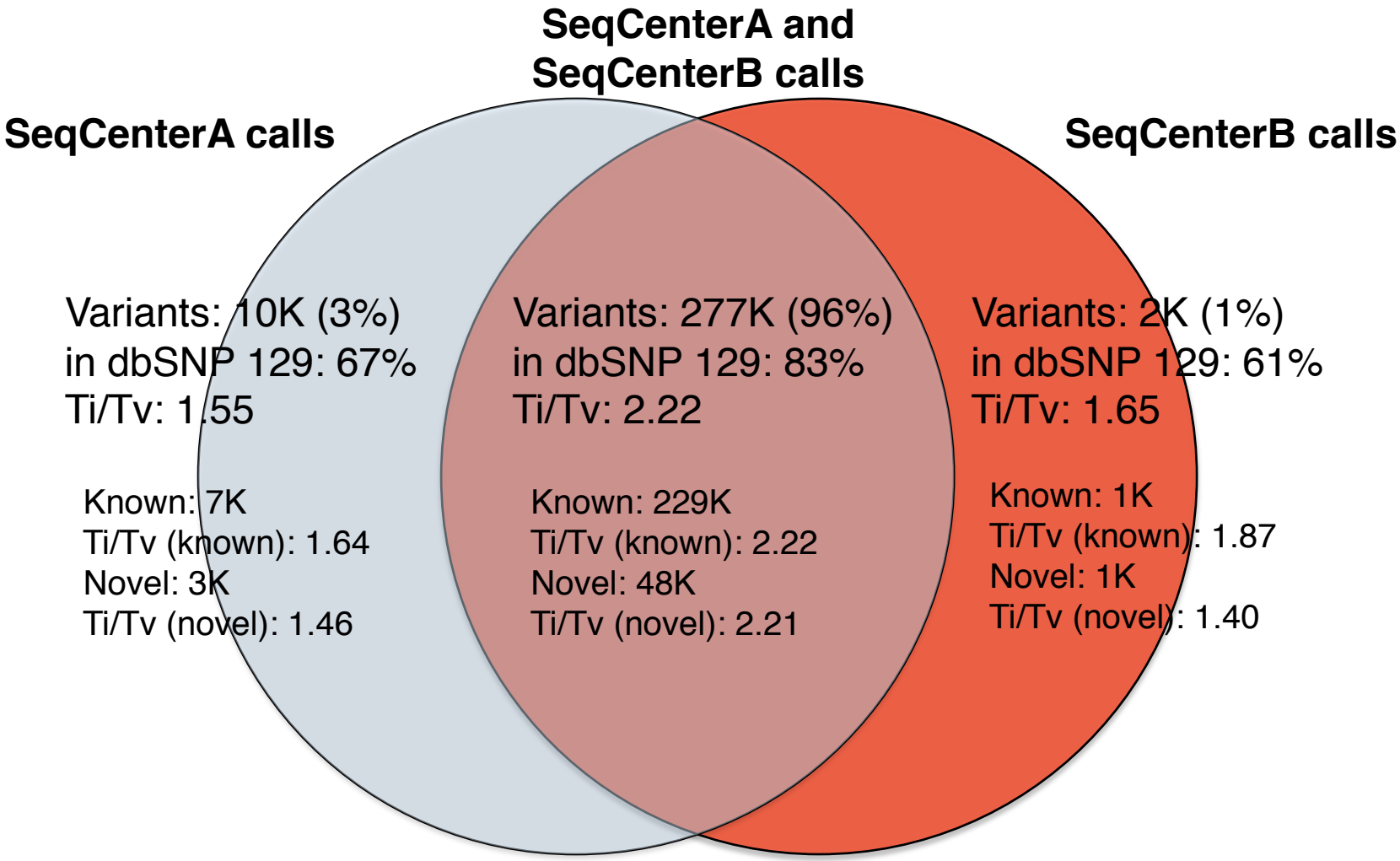
NA19240.venn.vcf: excerpt

#CHROM	POS	…	INFO	FORMAT	NA19240.callset2	NA19240.callset1
1	744045	…	…;NwayVenn=NA19240.callset1-NA19240.callset2	GT:DP:GQ	1/1:-1:99.00	1/1:23:68.04
1	804163	…	…;NwayVenn=NA19240.callset1	GT:DP:GQ	./.	1/0:47:99.00
1	891200	…	…;NwayVenn=NA19240.callset2	GT:DP:GQ	1/0:-1:99.00	./.

CallsetConcordance tells you if a SNP call appears in one or more callsets. VariantEval allows for more sophisticated selection operators to evaluate things occurring in one Venn segment (see docs for details).

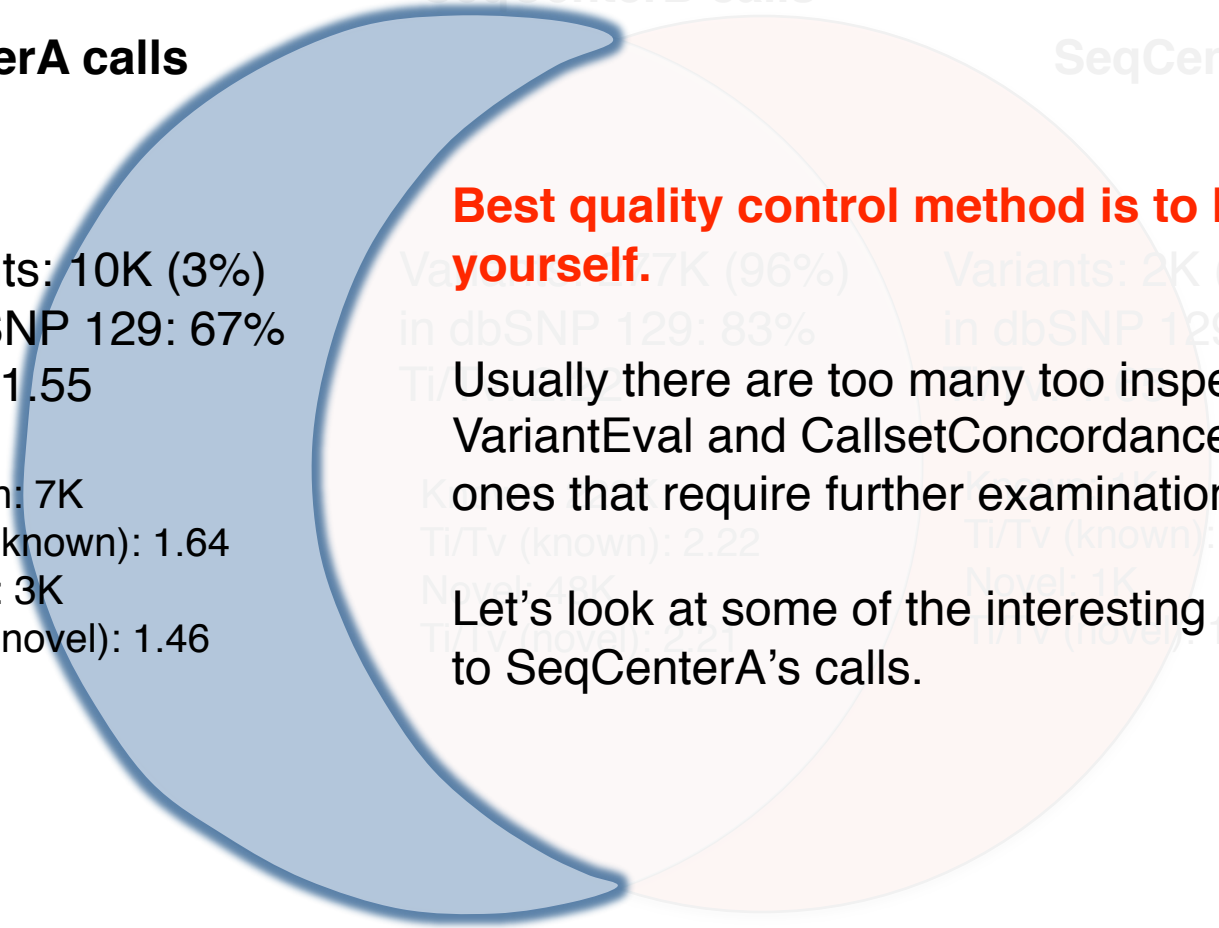
See http://www.broadinstitute.org/gsa/wiki/index.php/Callsets_Concordance_Tool for more detail.

CallsetConcordance & VariantEval help compare and contrast callsets



See http://www.broadinstitute.org/gsa/wiki/index.php/Callsets_Concordance_Tool for more detail.

CallsetConcordance & VariantEval help compare and contrast callsets



Best quality control method is to look at the data yourself.

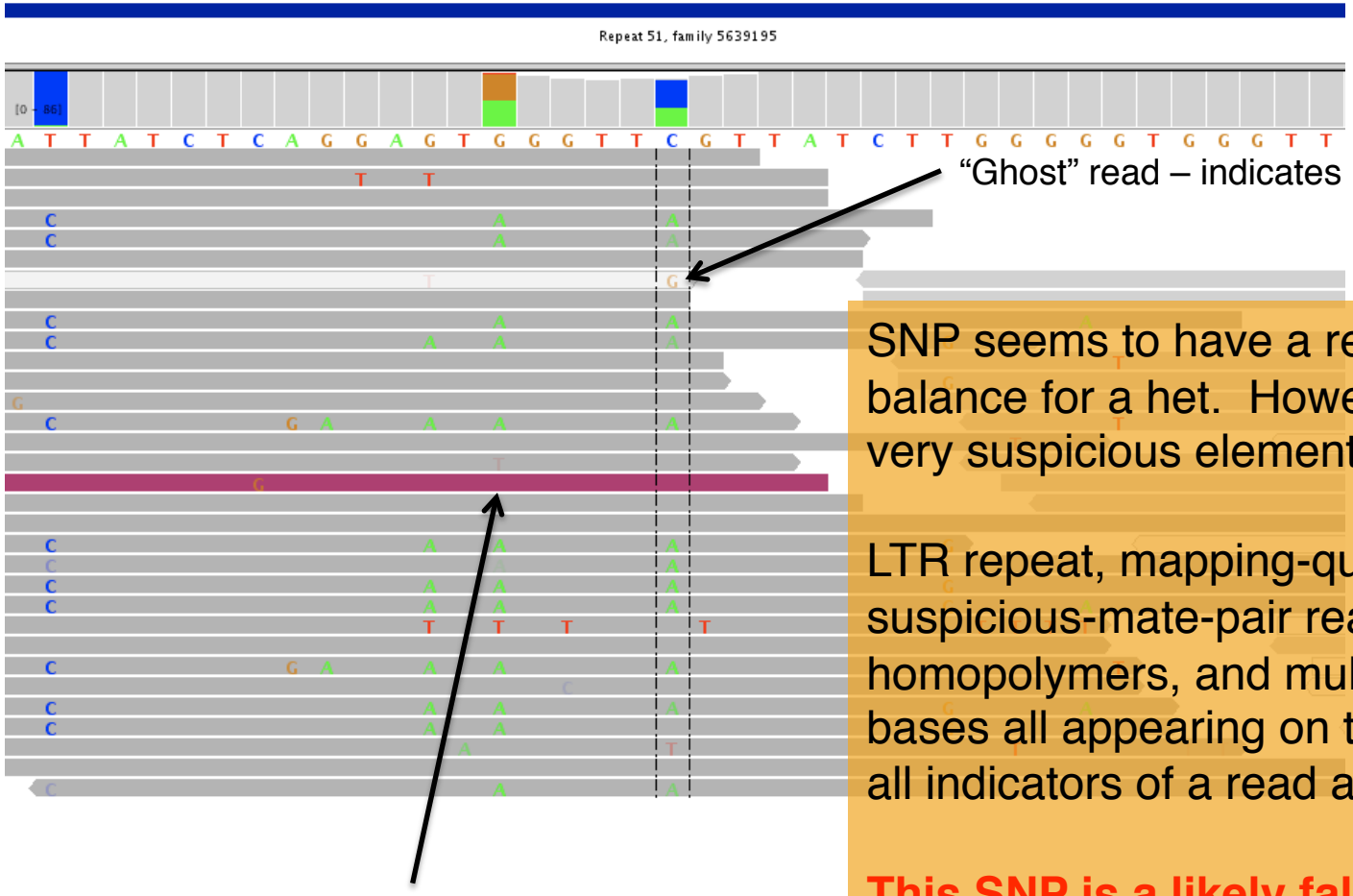
Usually there are too many too inspect, but VariantEval and CallsetConcordance help find the ones that require further examination.

Let's look at some of the interesting that are exclusive to SeqCenterA's calls.

See http://www.broadinstitute.org/gsa/wiki/index.php/Callsets_Concordance_Tool for more detail.

A likely false-positive SNP, suspicious for its proximity to other SNPs

NA19240, chr1:5,639,327-5,639,365



SNP seems to have a reasonable allele balance for a het. However, there are some very suspicious elements in the vicinity.

LTR repeat, mapping-quality-zero read, suspicious-mate-pair read, short homopolymers, and multiple non-reference bases all appearing on the same reads are all indicators of a read alignment error.

This SNP is a likely false-positive.

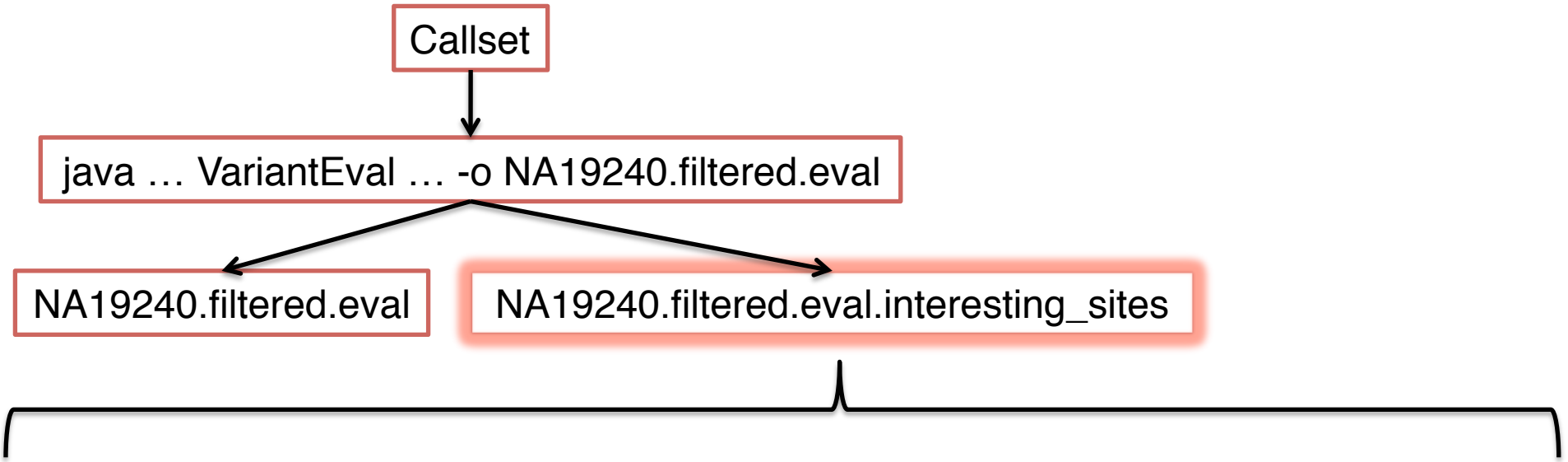
See http://www.broadinstitute.org/igv/ for more details.

Conclusion

- VariantEval computes quality metrics for subsets of SNP calls (all, known, novel, filtered).
  - See http://broadinstitute.org/gsa/wiki/index.php/VariantEval for more detail
- CallsetConcordance is useful for quickly determining shared and private mutations between callsets.
  - See http://www.broadinstitute.org/gsa/wiki/index.php/Callsets_Concordance_Tool for more detail
- IGV is phenomenally helpful for looking at putative variant calls and seeing many potential quality covariates all at once.
  - See http://www.broadinstitute.org/igv/ and Jim Robinson's talk
- Additional help with GATK tools:
  - Wiki: http://www.broadinstitute.org/gsa/wiki/index.php/Main_Page
  - Community Forum: http://getsatisfaction.com/gsa

APPENDIX

VariantEval can provide a list of interesting sites that may warrant further investigation




```
all,flagged,db_coverage                    Discordant […  rs28587348 … A/C …] […  rs28587348      A       G]
all,known,db_coverage                      Discordant […  rs28587348 … A/C …] […  rs28587348      A       G]
novel,flagged,neighbor_distances           neighbor-distance 5 1:1638645 1:1638640
novel,flagged,cluster_counter_analysis     snp_within_cluster 5 1:1638645 1:1638640 0 0 5 10 20 50 100
all,flagged,genotype_concordance           False negative: ref=T chip=[1:1682181 CT] call=null
all,flagged,genotype_concordance           False negative: ref=G chip=[1:1685434 GT] call=null
all,flagged,genotype_concordance           False negative: ref=T chip=[1:1695996 CT] call=null
```

VariantEval's "interesting_sites" list contains dbSNP discordancies, suspicious novel calls, false-negative sites, and other flagged items to make follow-up investigation easier.

See http://www.broadinstitute.org/gsa/wiki/index.php/VariantEval for more detail.