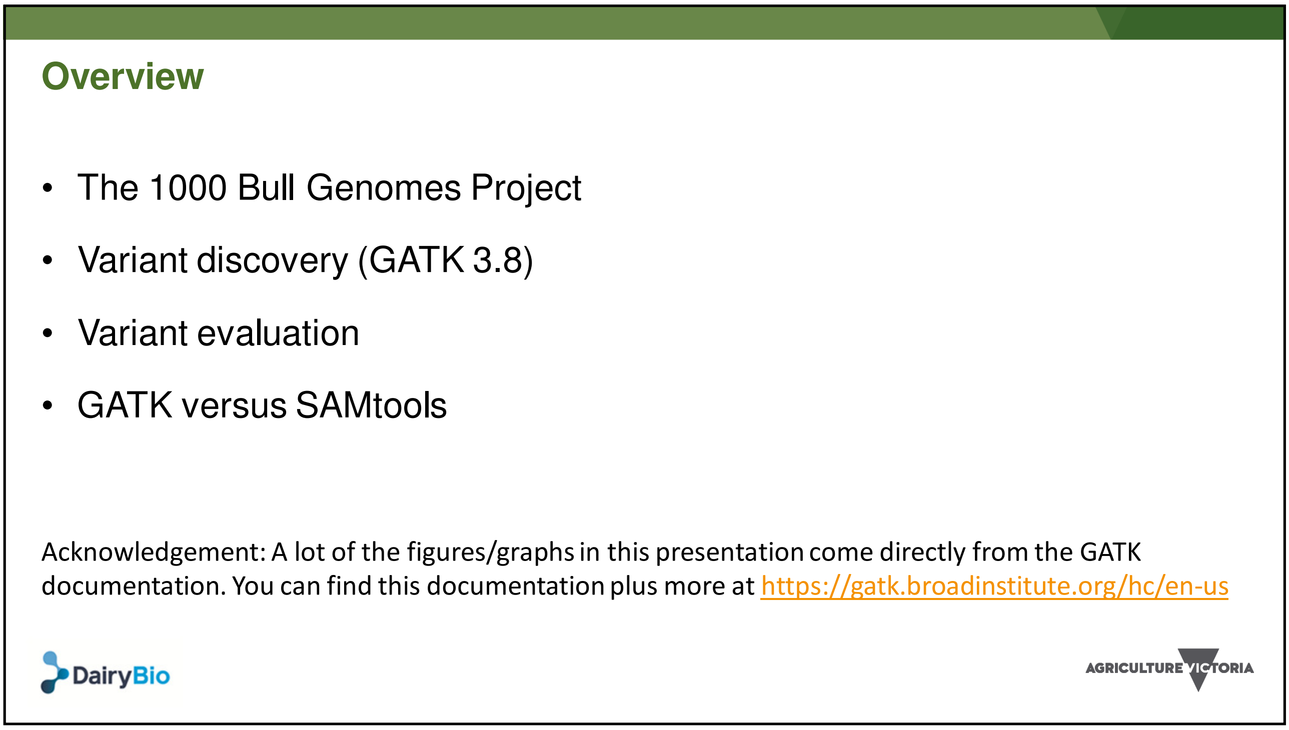
Variant Detection and Genotype Calling Using GATK

Part 2

Dr Amanda Chamberlain and Dr Christy Vander Jagt

1



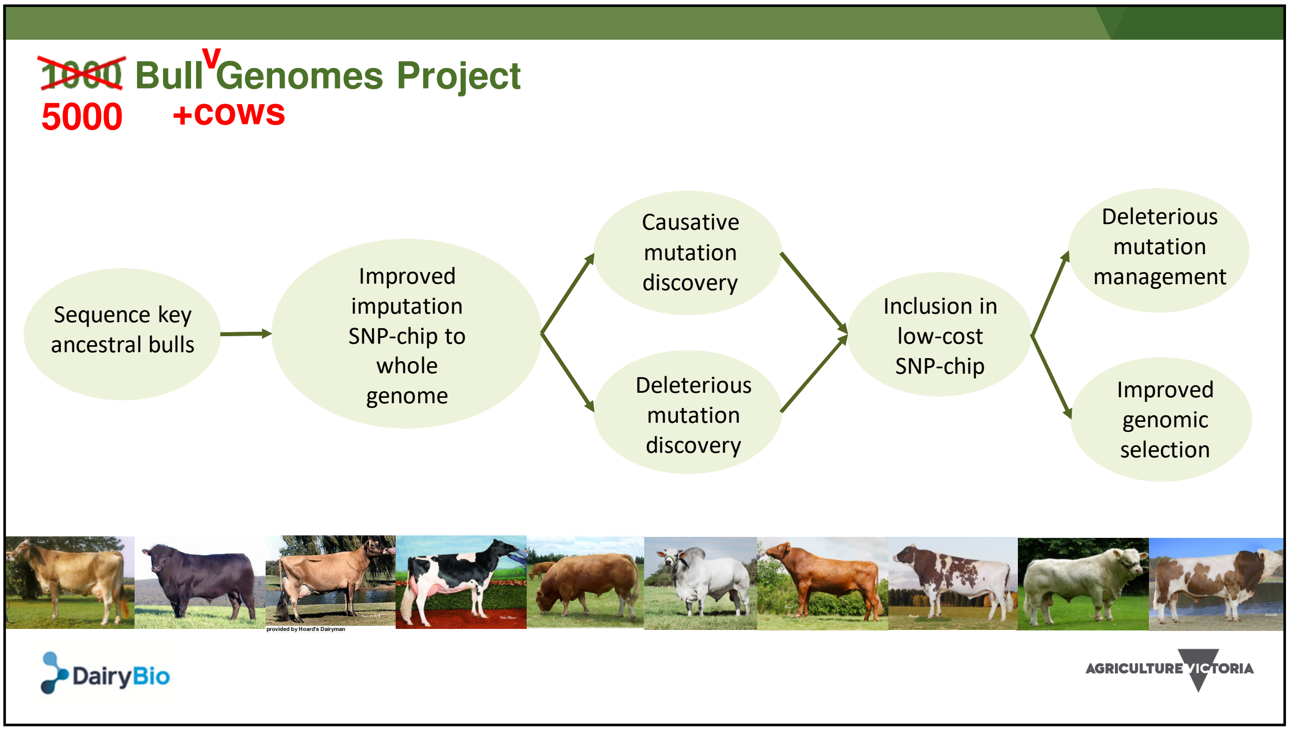
Overview

- The 1000 Bull Genomes Project
- Variant discovery (GATK 3.8)
- Variant evaluation
- GATK versus SAMtools

Acknowledgement: A lot of the figures/graphs in this presentation come directly from the GATK documentation. You can find this documentation plus more at https://gatk.broadinstitute.org/hc/en-us

DairyBio
AGRICULTURE VICTORIA

2

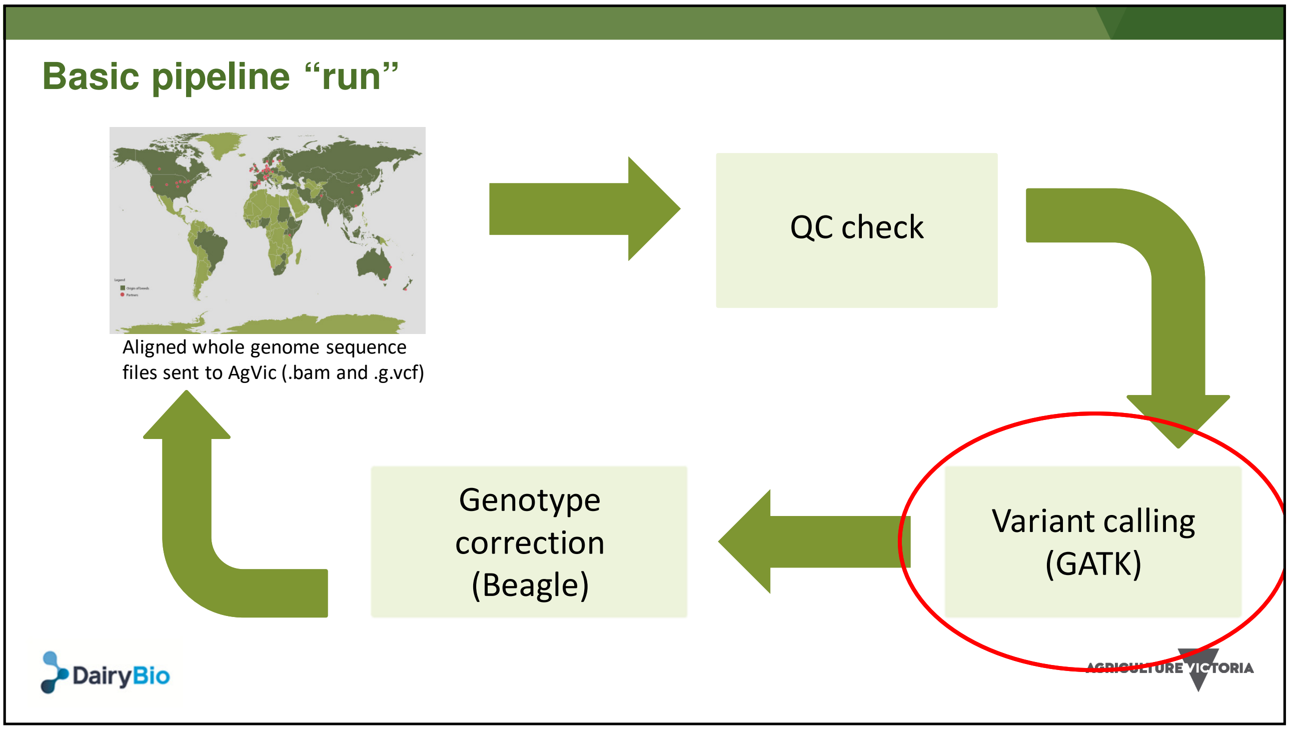


~~1000~~ Bull Genomes Project

5000 +cows

Sequence key ancestral bulls
Improved imputation SNP-chip to whole genome
Causative mutation discovery
Deleterious mutation discovery
Inclusion in low-cost SNP-chip
Deleterious mutation management
Improved genomic selection
provided by Hoard's Dairyman
DairyBio
AGRICULTURE VICTORIA

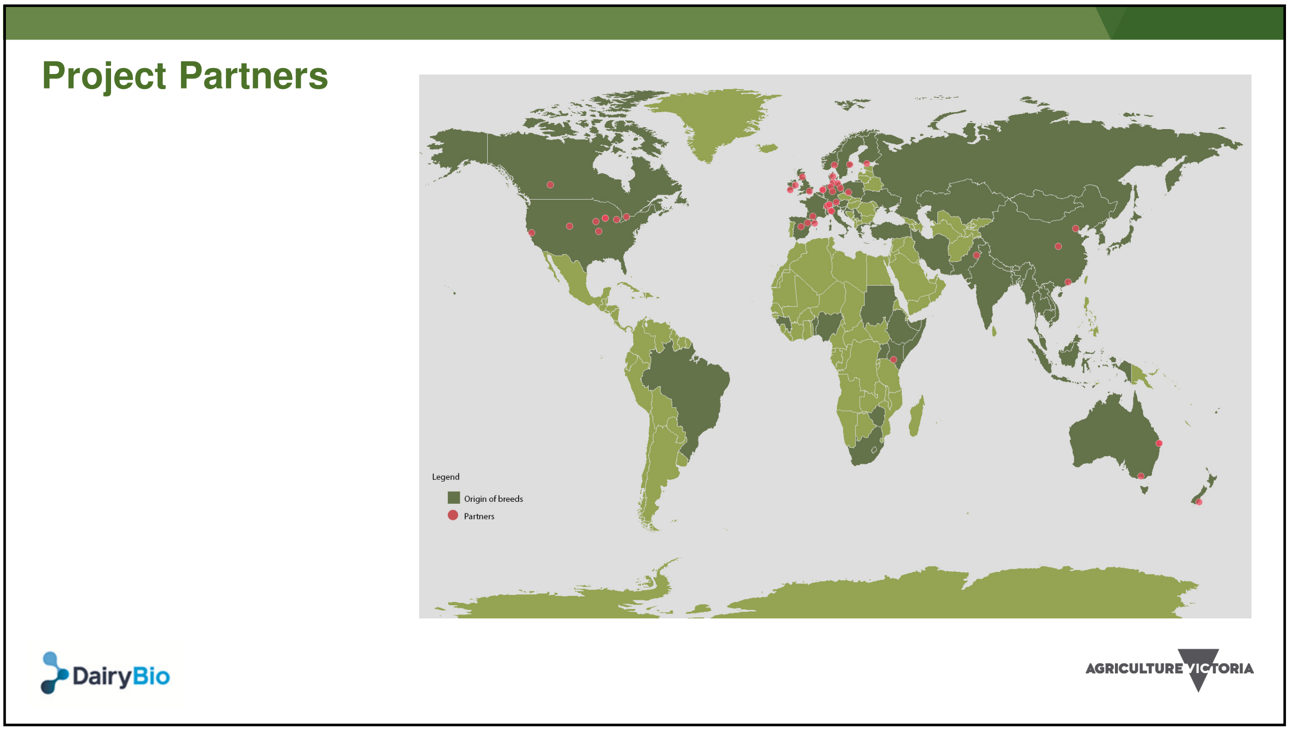
3



Basic pipeline “run”

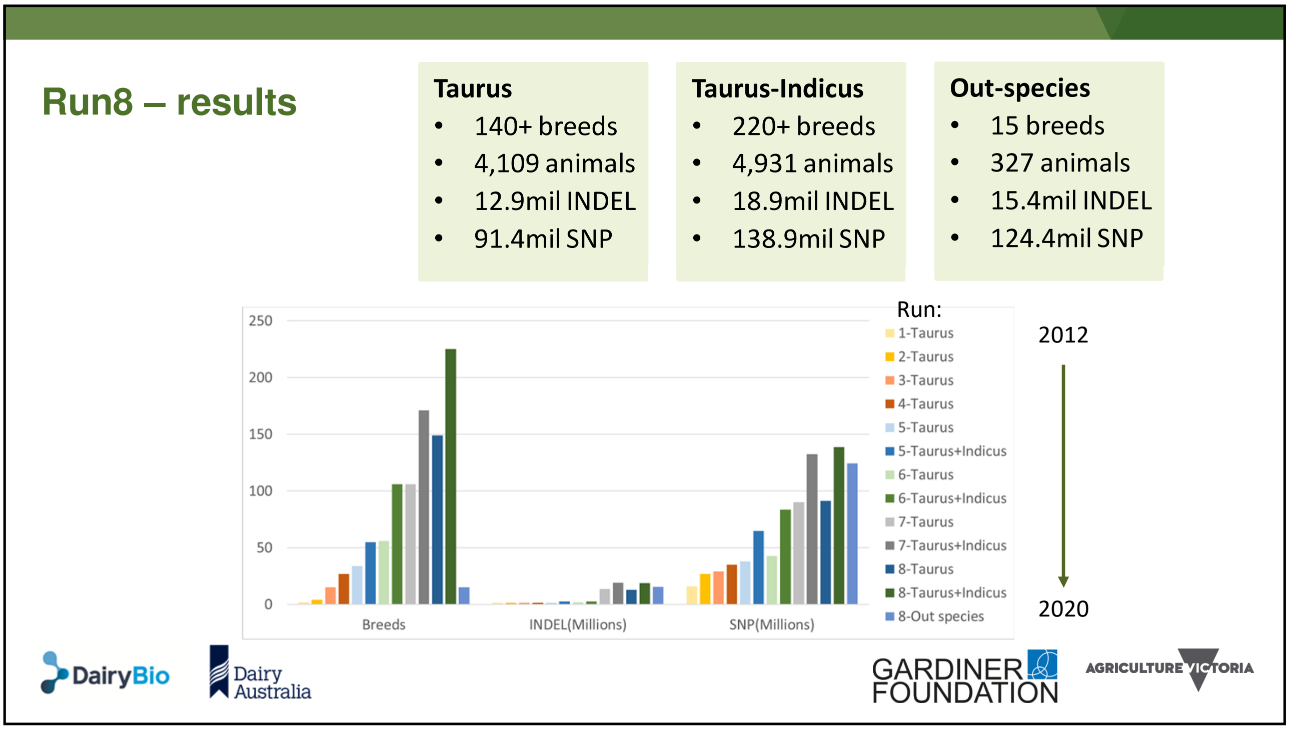
Aligned whole genome sequence files sent to AgVic (.bam and .g.vcf)

4



Project Partners

5



Run8 – results

Taurus

- 140+ breeds
- 4,109 animals
- 12.9mil INDEL
- 91.4mil SNP

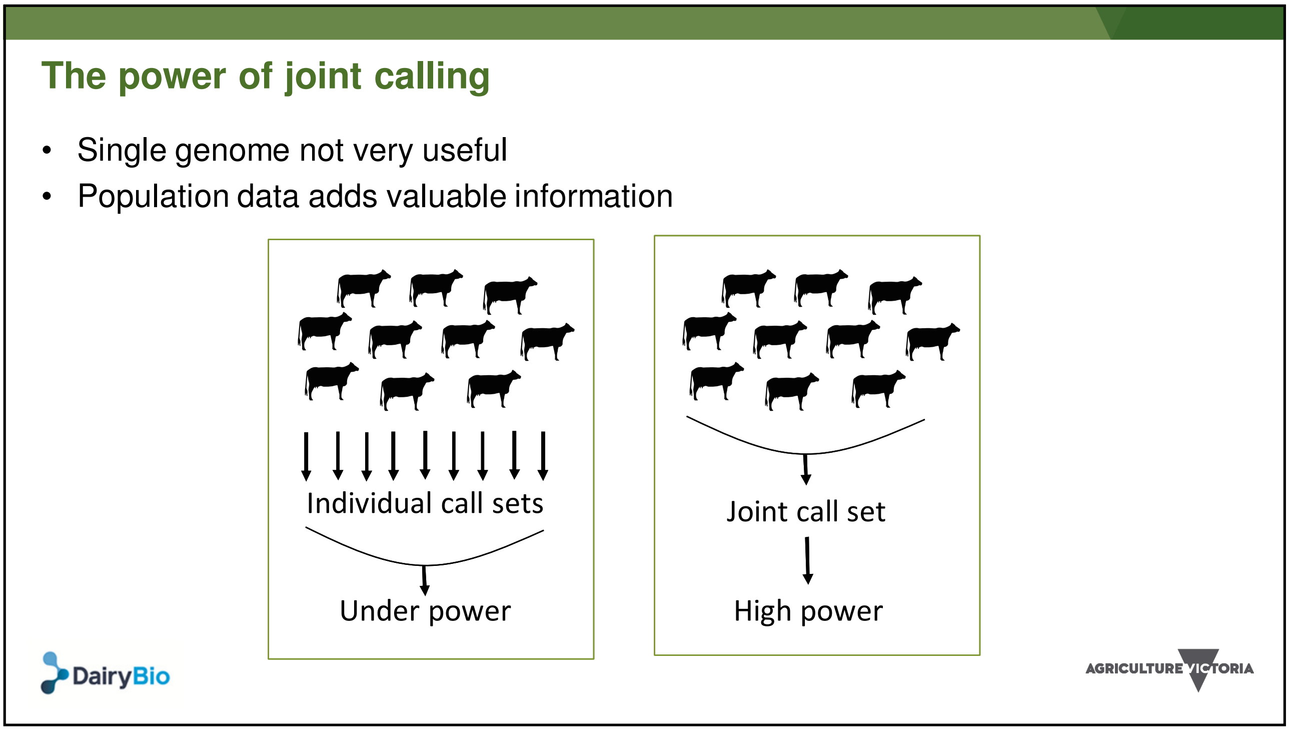
Taurus-Indicus

- 220+ breeds
- 4,931 animals
- 18.9mil INDEL
- 138.9mil SNP

Out-species

- 15 breeds
- 327 animals
- 15.4mil INDEL
- 124.4mil SNP

6



The power of joint calling

- Single genome not very useful
- Population data adds valuable information

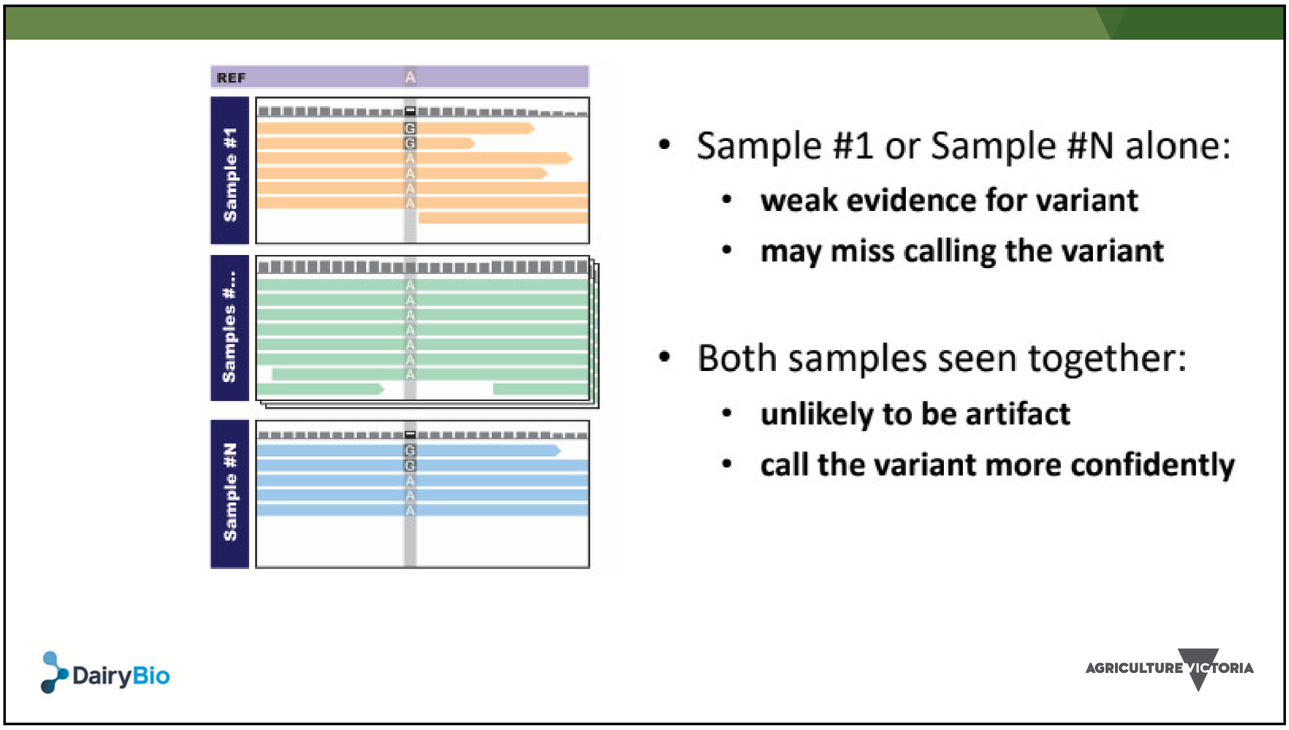
Individual call sets

Under power

Joint call set

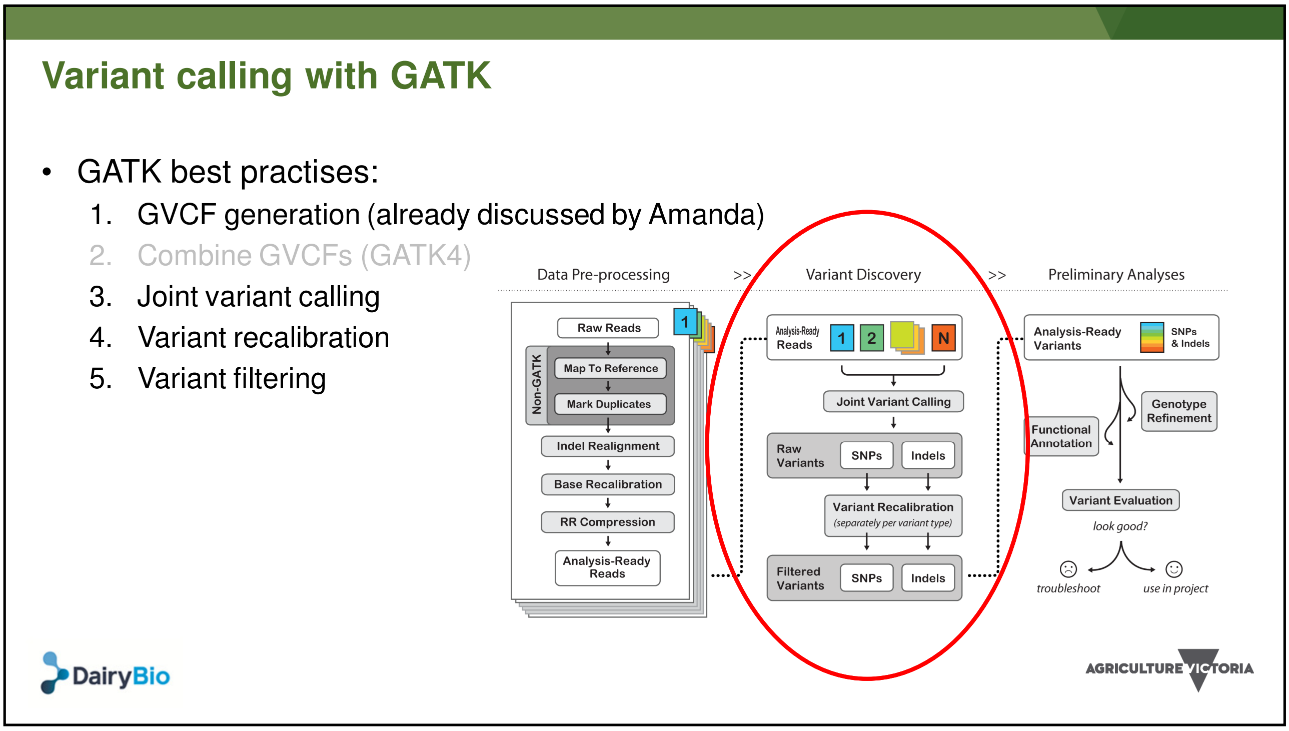
High power

7



- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**
- Both samples seen together:
  - **unlikely to be artifact**
  - **call the variant more confidently**

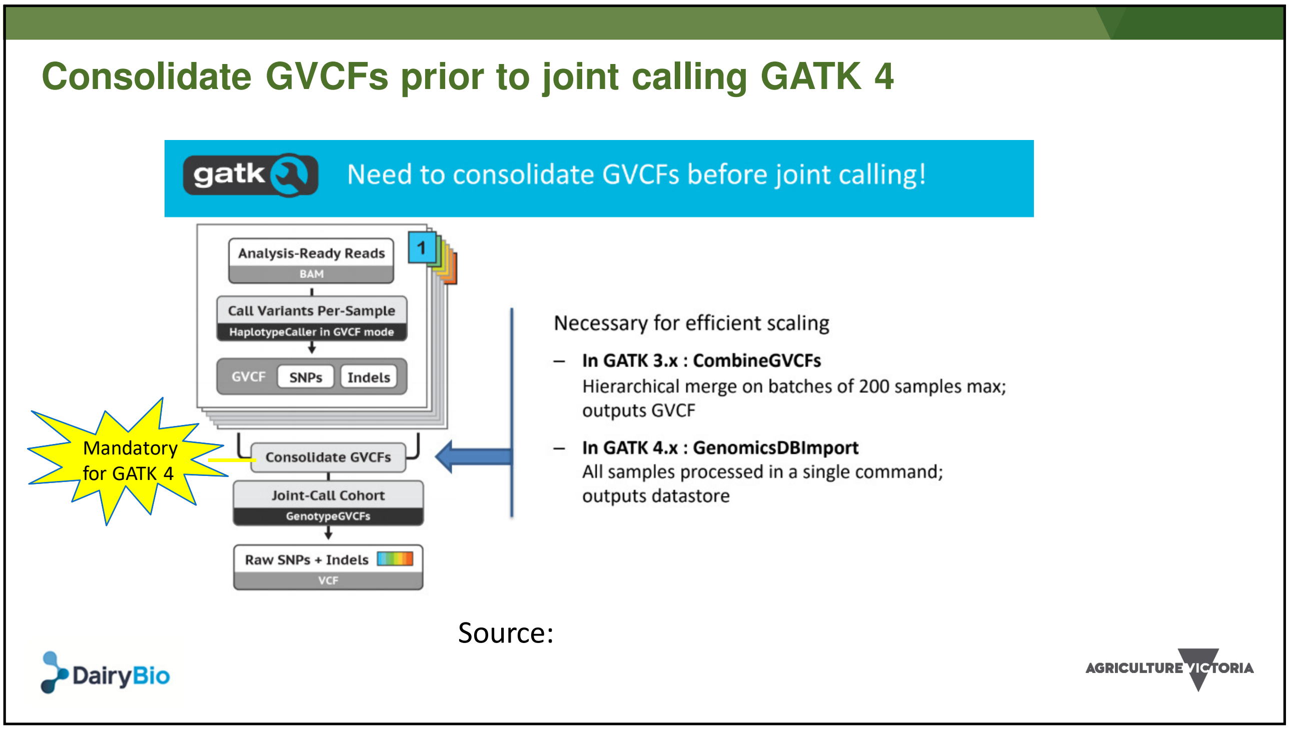
8



Variant calling with GATK

- GATK best practises:
  1. GVCF generation (already discussed by Amanda)
  2. Combine GVCFs (GATK4)
  3. Joint variant calling
  4. Variant recalibration
  5. Variant filtering

9



Consolidate GVCFs prior to joint calling GATK 4

Necessary for efficient scaling

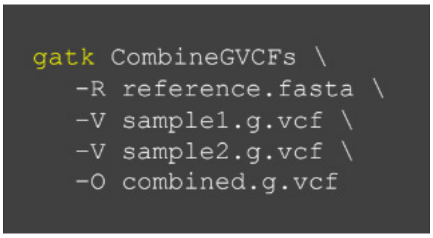
- **In GATK 3.x : CombineGVCFs**
  Hierarchical merge on batches of 200 samples max; outputs GVCF
- **In GATK 4.x : GenomicsDBImport**
  All samples processed in a single command; outputs datastore

Source:

10

Consolidate GVCFs prior to joint calling – GATK 3.8

- Not 100% necessary
- Recommended to run CombineGVCFs tool on batches of 200 samples
- Command:



```
gatk CombineGVCFs \
   -R reference.fasta \
   -V sample1.g.vcf \
   -V sample2.g.vcf \
   -O combined.g.vcf
```

- More information: https://gatk.broadinstitute.org/hc/en-us/articles/360037053272-CombineGVCFs

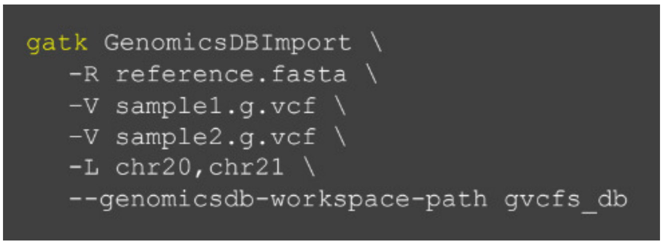
DairyBio

AGRICULTURE VICTORIA

11

Consolidate GVCFs prior to joint calling – GATK 4

- Compulsory for GATK 4
- Can use CombineGVCFs tool but best practises recommends GenomicsDBImport
- Command:



```
gatk GenomicsDBImport \
   -R reference.fasta \
   -V sample1.g.vcf \
   -V sample2.g.vcf \
   -L chr20,chr21 \
   --genomicsdb-workspace-path gvcfs_db
```

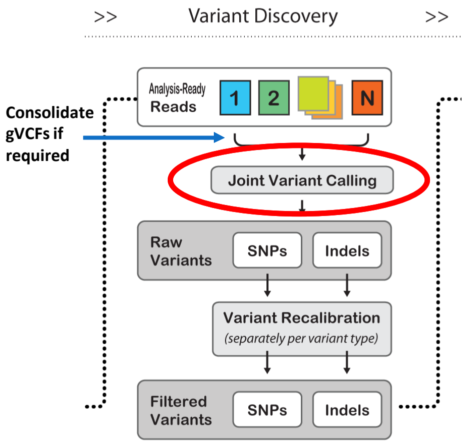
- More information: https://gatk.broadinstitute.org/hc/en-us/articles/360036883491-GenomicsDBImport

DairyBio

AGRICULTURE VICTORIA

12

Joint Variant Calling – GenotypeGVCFs



GenotypeGVCFs can take multiple GVCF files in GATK 3.8 (multiple –V variants), but only a single file in GATK 4

GATK 3.8 command:

```
java –jar GenomeAnalysisTK.jar
    –T GenotypeGVCFs \
    –R human.fasta \
    –V sample1.g.vcf \
    –V sample2.g.vcf \
    –V sampleN.g.vcf \
    –o output.vcf
```

GATK 4 commands:

```
gatk GenotypeGVCFs \
   -R reference.fasta \
   -V variants.g.vcf \
   -O final_variants.vcf
```

```
gatk GenotypeGVCFs \
   -R reference.fasta \
   -V gendb://gvcfs_db \
   -O final_variants.vcf
```

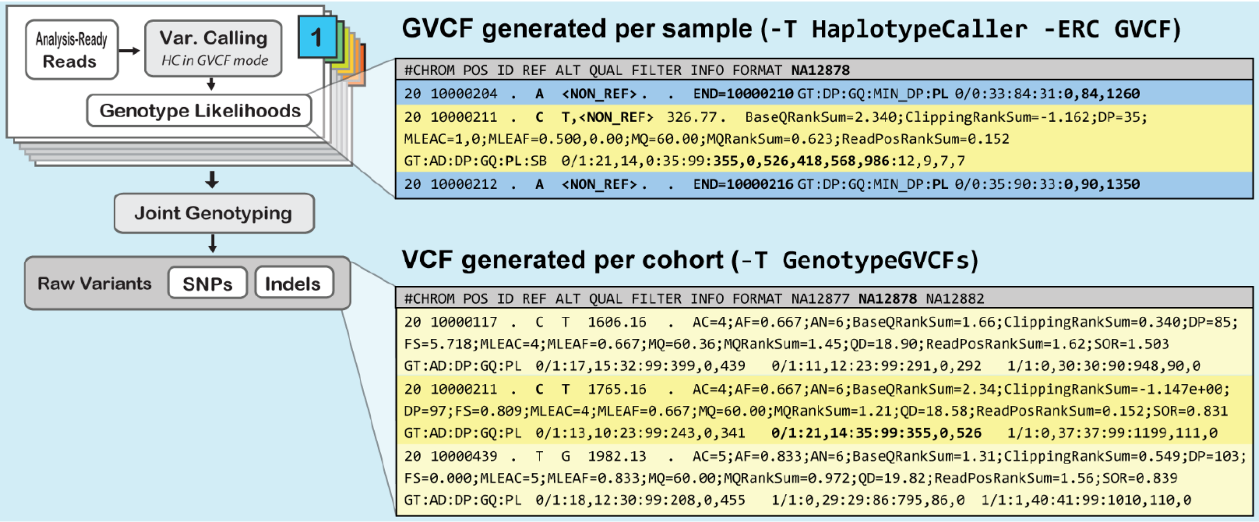
DairyBio

AGRICULTURE VICTORIA

More information: https://gatk.broadinstitute.org/hc/en-us/articles/360037594731-GenotypeGVCFs

13

Joint Variant Calls



GVCF generated per sample (-T HaplotypeCaller -ERC GVCF)

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 .  A  <NON_REF> .  .  END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 .  C  T,<NON_REF> 326.77 .  BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB  0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 .  A  <NON_REF> .  .  END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

VCF generated per cohort (-T GenotypeGVCFs)

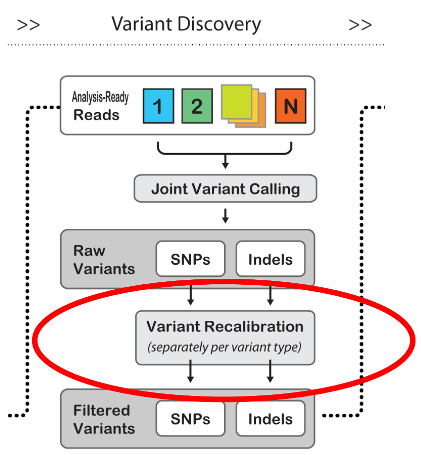
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 .  C  T  1606.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL  0/1:17,15:32:99:399,0,439   0/1:11,12:23:99:291,0,292   1/1:0,30:30:90:948,90,0
20 10000211 .  C  T  1765.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL  0/1:13,10:23:99:243,0,341   0/1:21,14:35:99:355,0,526   1/1:0,37:37:99:1199,111,0
20 10000439 .  T  G  1982.13  .  AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL  0/1:18,12:30:99:208,0,455   1/1:0,29:29:86:795,86,0  1/1:1,40:41:99:1010,110,0
```

DairyBio

AGRICULTURE VICTORIA

14

Variant call set filtering



- Variant calling algorithms are very permissive by design
- How to filter?
  - Hard filtering
    - Multiple threshold values
    - Binary choice: pass or fail
  - Variant “recalibration”
    - Machine learning
    - Annotation profile of “good” vs “bad” variants
    - Multiple annotations
- Trade-off between sensitivity and specificity

More information: https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering

15

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

```
22 49582364              .         A         G         198.96   .
    AC=3;
    AF=0.50;
    AN=6;
    DP=87;
    MLEAC=3;
    MLEAF=0.50;
    MQ=71.31;
    MQ0=22;
    QD=2.29;
    SB=-31.76
    GT:DP:GQ   0/1:12:99     0/1:11:89     0/1:28:37
```

INFO field

AC	No. chromosomes carrying alt allele	MLEAF	Max likelihood AF
AN	Total no. of chromosomes	MQ	RMS MAPQ of all reads
AF	Allele frequency	MQ0	No. of MAPQ 0 reads at locus
DP	Depth of coverage	QD	QUAL score over depth
MLEAC	Max likelihood AC		

DairyBio

AGRICULTURE VICTORIA

16

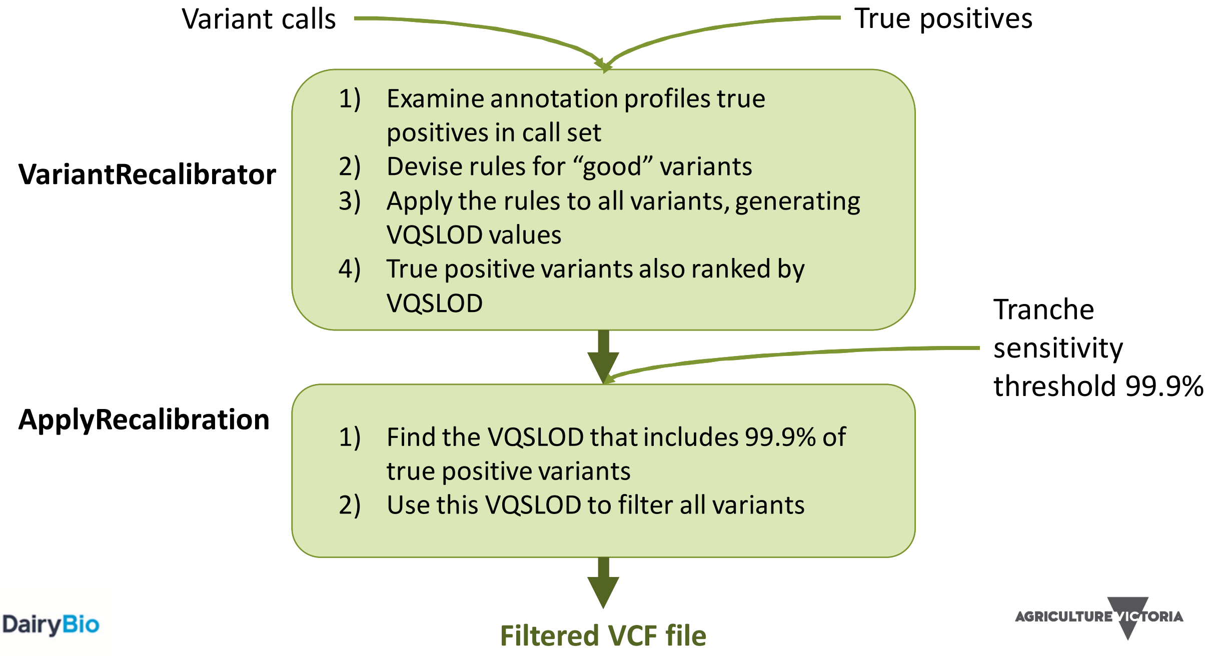
Best practises: Variant Quality Score Recalibration (VQSR)

- "Sophisticated filtering technique applied on the variant callset that uses machine learning to model the technical profile of variants in a training set and uses that to filter out probable artifacts from the callset."

- Two step process:
  1. Variant recalibration (VariantRecalibrator)
  2. Applying the recalibration (ApplyRecalibration)

More information: https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR-

17



Variant calls

True positives

VariantRecalibrator

1) Examine annotation profiles true positives in call set
2) Devise rules for "good" variants
3) Apply the rules to all variants, generating VQSLOD values
4) True positive variants also ranked by VQSLOD

Tranche sensitivity threshold 99.9%

ApplyRecalibration

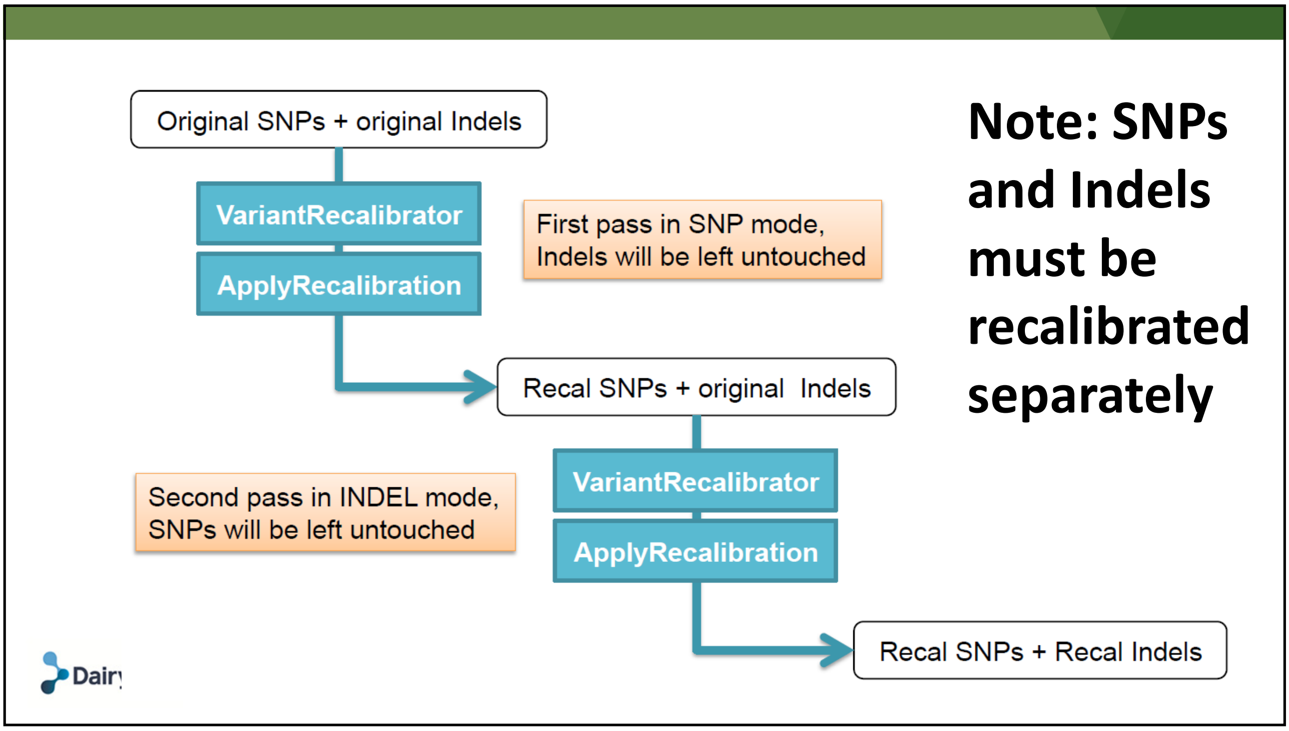
1) Find the VQSLOD that includes 99.9% of true positive variants
2) Use this VQSLOD to filter all variants

Filtered VCF file

DairyBio

AGRICULTURE VICTORIA

18



Original SNPs + original Indels

VariantRecalibrator

ApplyRecalibration

First pass in SNP mode, Indels will be left untouched

Recal SNPs + original Indels

VariantRecalibrator

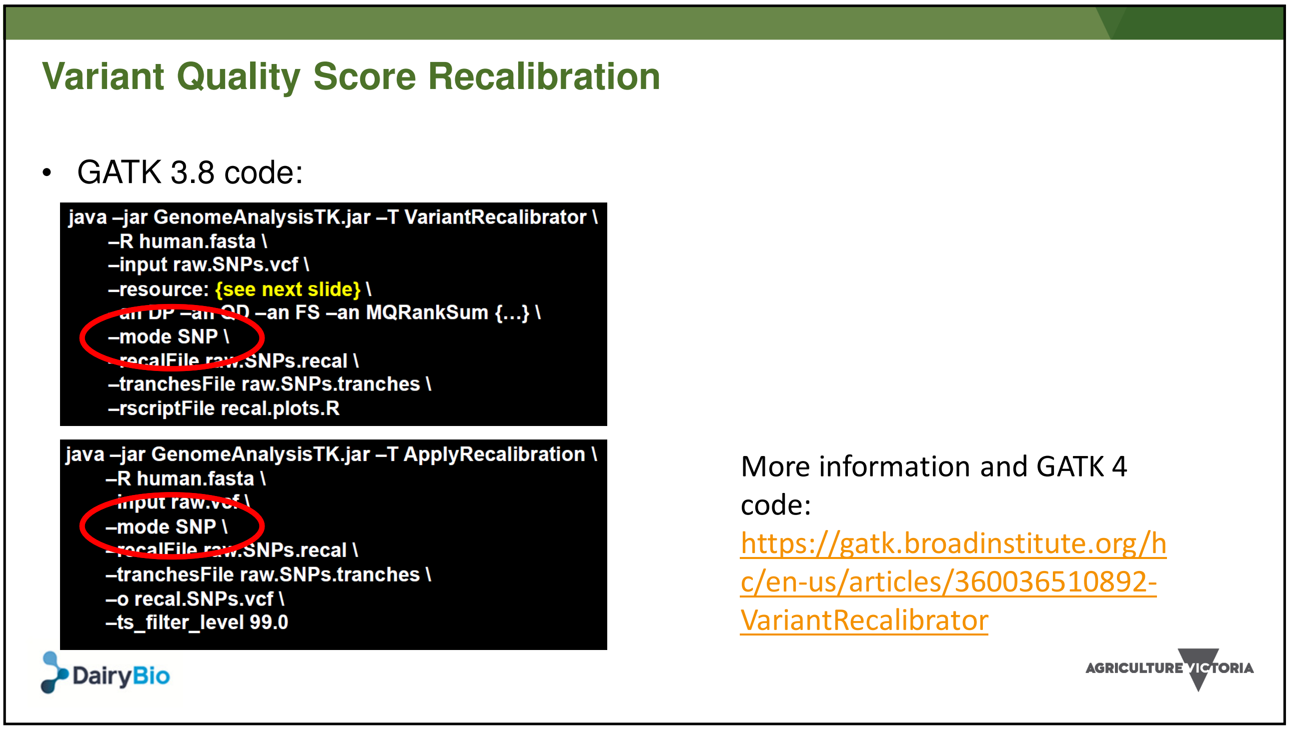
ApplyRecalibration

Second pass in INDEL mode, SNPs will be left untouched

Recal SNPs + Recal Indels

Note: SNPs and Indels must be recalibrated separately

19



Variant Quality Score Recalibration

- GATK 3.8 code:

```
java –jar GenomeAnalysisTK.jar –T VariantRecalibrator \
      –R human.fasta \
      –input raw.SNPs.vcf \
      –resource: {see next slide} \
      –an DP –an QD –an FS –an MQRankSum {…} \
      –mode SNP \
      –recalFile raw.SNPs.recal \
      –tranchesFile raw.SNPs.tranches \
      –rscriptFile recal.plots.R
```

```
java –jar GenomeAnalysisTK.jar –T ApplyRecalibration \
      –R human.fasta \
      –input raw.vcf \
      –mode SNP \
      –recalFile raw.SNPs.recal \
      –tranchesFile raw.SNPs.tranches \
      –o recal.SNPs.vcf \
      –ts_filter_level 99.0
```

More information and GATK 4 code: https://gatk.broadinstitute.org/hc/en-us/articles/360036510892-VariantRecalibrator

20

Resource datasets

- Three types of resources:
  1) Truth
     - Validated to a high degree of confidence
     - Representative of “true” sites (truth=true)
     - Used to train recalibration model (training=true)
     - Used to determine where to set cutoff in VQSLOD sensitivity
  2) Training
     - Validated to some degree of confidence
     - May contain false positives (truth=false)
     - Used to train recalibration model (training=true)
  3) Known
     - Not validated to a high degree of confidence (truth=false)
     - Not used to train recalibration model (training=false)
     - Only for reporting purposes, not used in any calculations
  - Prior
     - Phred-scaled estimate of data accuracy

```
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.b37.sites.vcf
-resource:omni,known=false,training=true,truth=false,prior=12.0
omni2.5.b37.sites.vcf
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G.b37.sites.vcf
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_137.b37.vcf
```

21

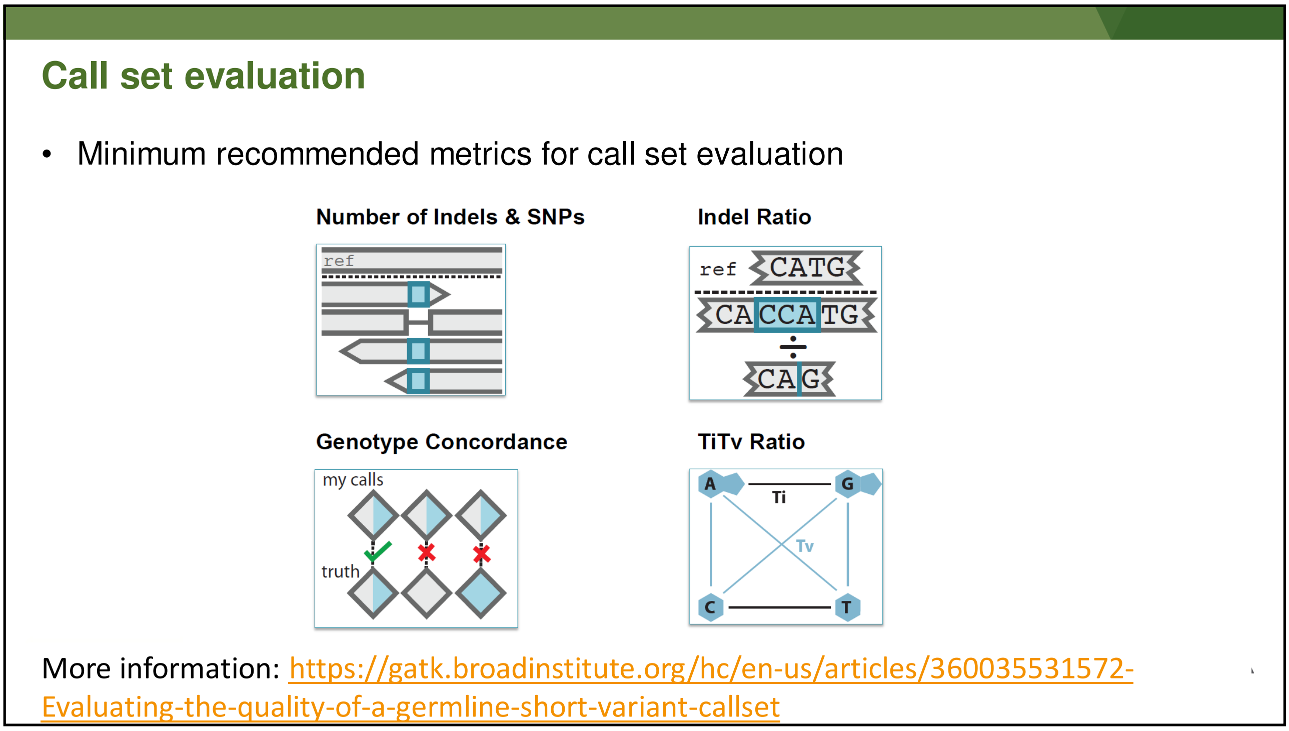
Where to find resources

- Human genome training, truth and known resource datasets: https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle
- GATK do not provide resources for non-human organisms
  - you need to have at least truth and training resource datasets (with assigned prior likelihoods)
  - GATK forum topic called “Non-Human”: https://gatk.broadinstitute.org/hc/en-us/community/topics/360001496611-Non-Human

DairyBio

AGRICULTURE VICTORIA

22



Call set evaluation

- Minimum recommended metrics for call set evaluation

Number of Indels & SNPs

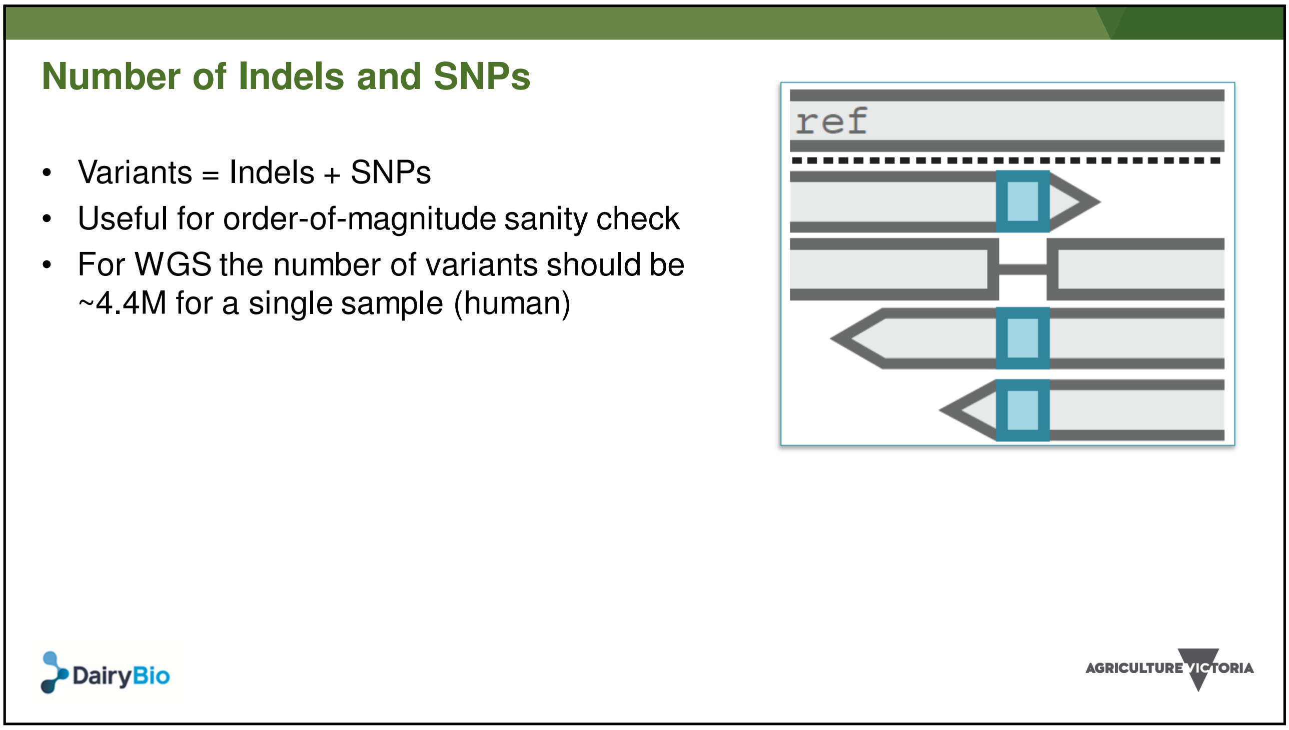
Indel Ratio

Genotype Concordance

TiTv Ratio

More information: https://gatk.broadinstitute.org/hc/en-us/articles/360035531572-Evaluating-the-quality-of-a-germline-short-variant-callset

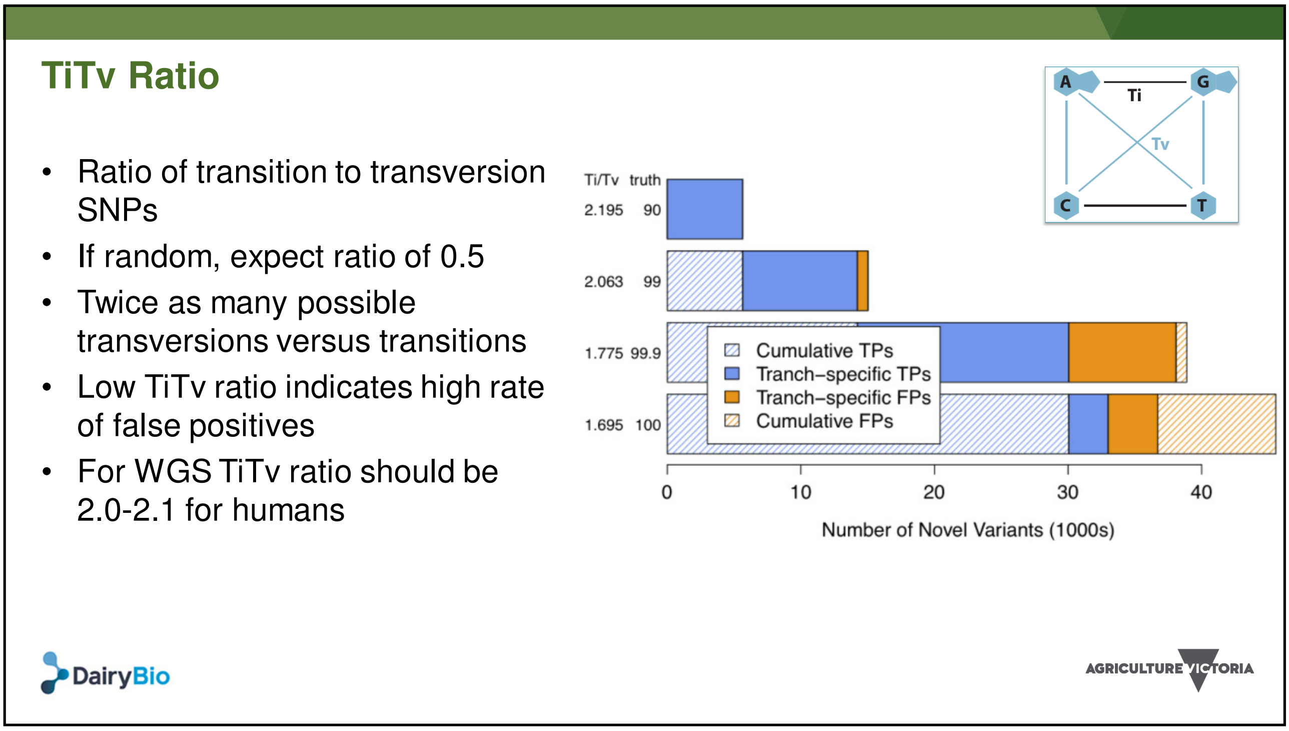
23



Number of Indels and SNPs

- Variants = Indels + SNPs
- Useful for order-of-magnitude sanity check
- For WGS the number of variants should be ~4.4M for a single sample (human)

24

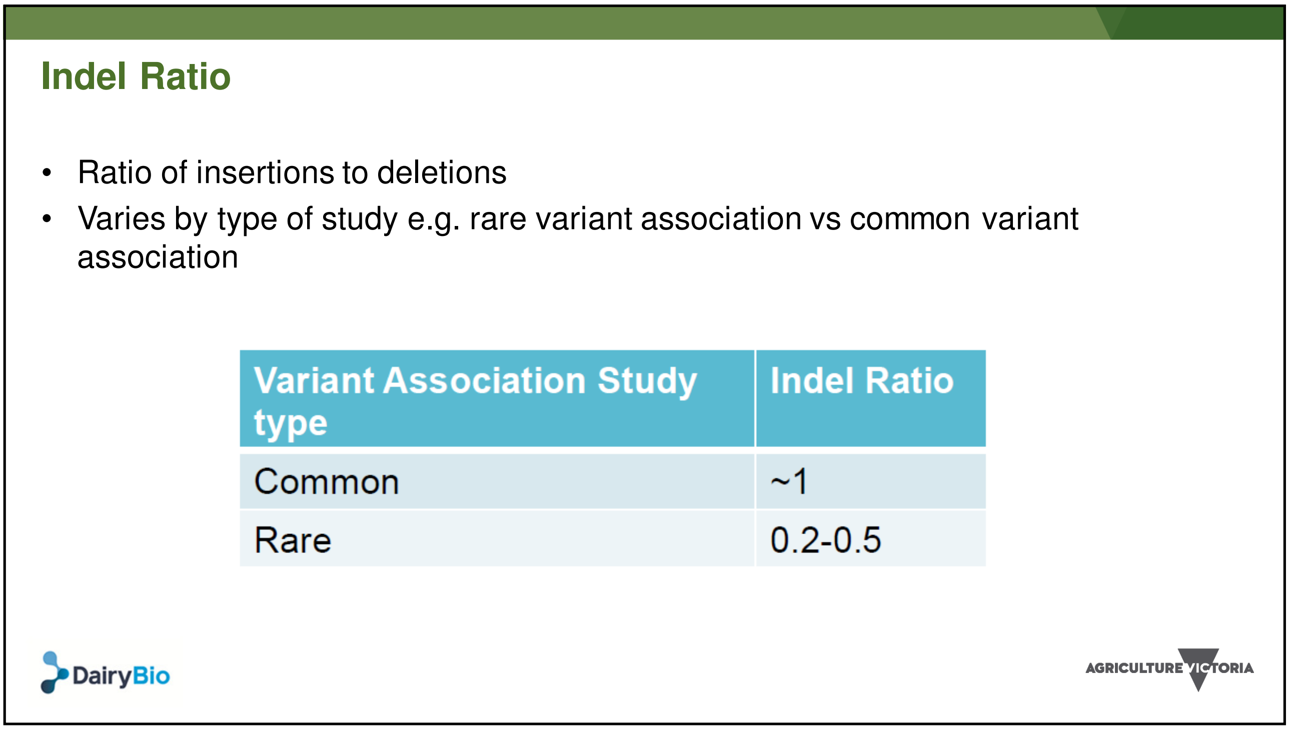


TiTv Ratio

- Ratio of transition to transversion SNPs
- If random, expect ratio of 0.5
- Twice as many possible transversions versus transitions
- Low TiTv ratio indicates high rate of false positives
- For WGS TiTv ratio should be 2.0-2.1 for humans

A
G
Ti
Tv
C
T
Ti/Tv truth
2.195 90
2.063 99
1.775 99.9
1.695 100
▨ Cumulative TPs
🟦 Tranch–specific TPs
🟧 Tranch–specific FPs
▨ Cumulative FPs
0
10
20
30
40
Number of Novel Variants (1000s)

25

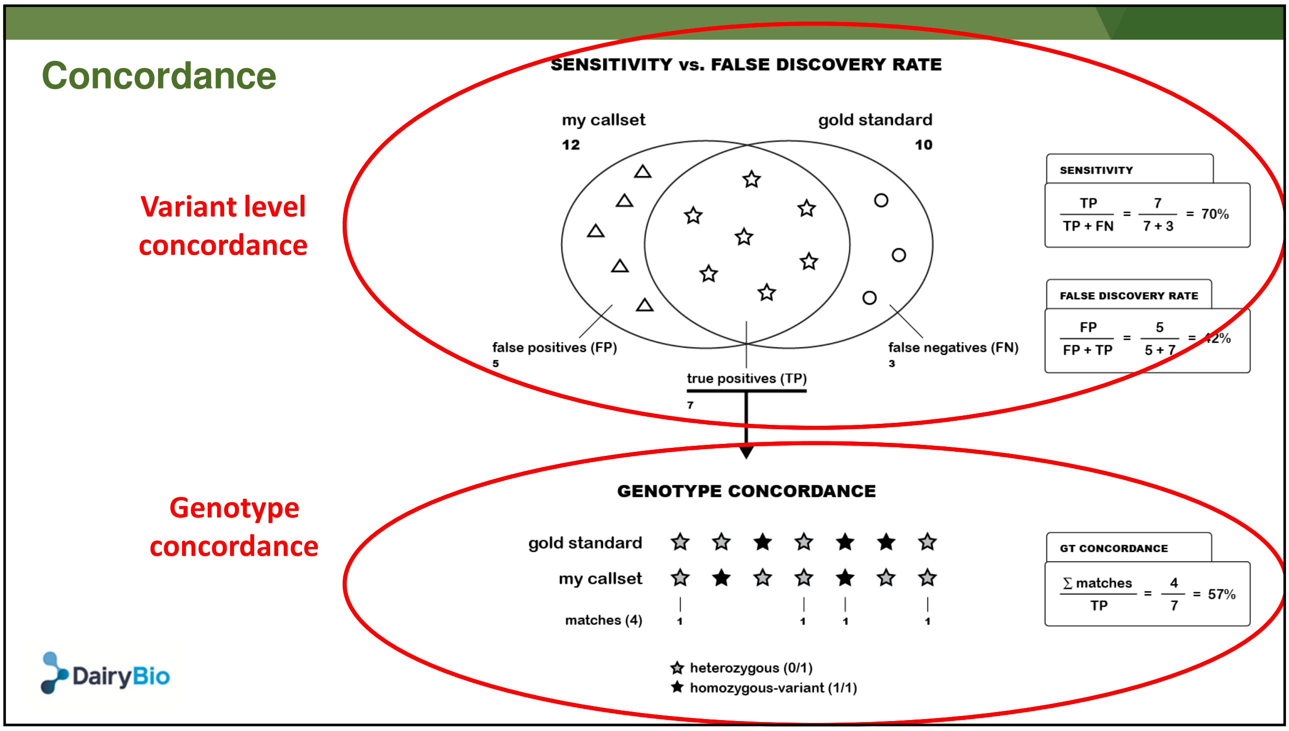


Indel Ratio

- Ratio of insertions to deletions
- Varies by type of study e.g. rare variant association vs common variant association

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

26



Concordance

Variant level concordance

SENSITIVITY vs. FALSE DISCOVERY RATE

my callset 12

gold standard 10

false positives (FP) 5

false negatives (FN) 3

true positives (TP) 7

SENSITIVITY

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

FALSE DISCOVERY RATE

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

Genotype concordance

GENOTYPE CONCORDANCE

gold standard	☆	☆	★	☆	★	★	☆
my callset	☆	★	☆	☆	★	☆	☆
matches (4)	1			1	1		1

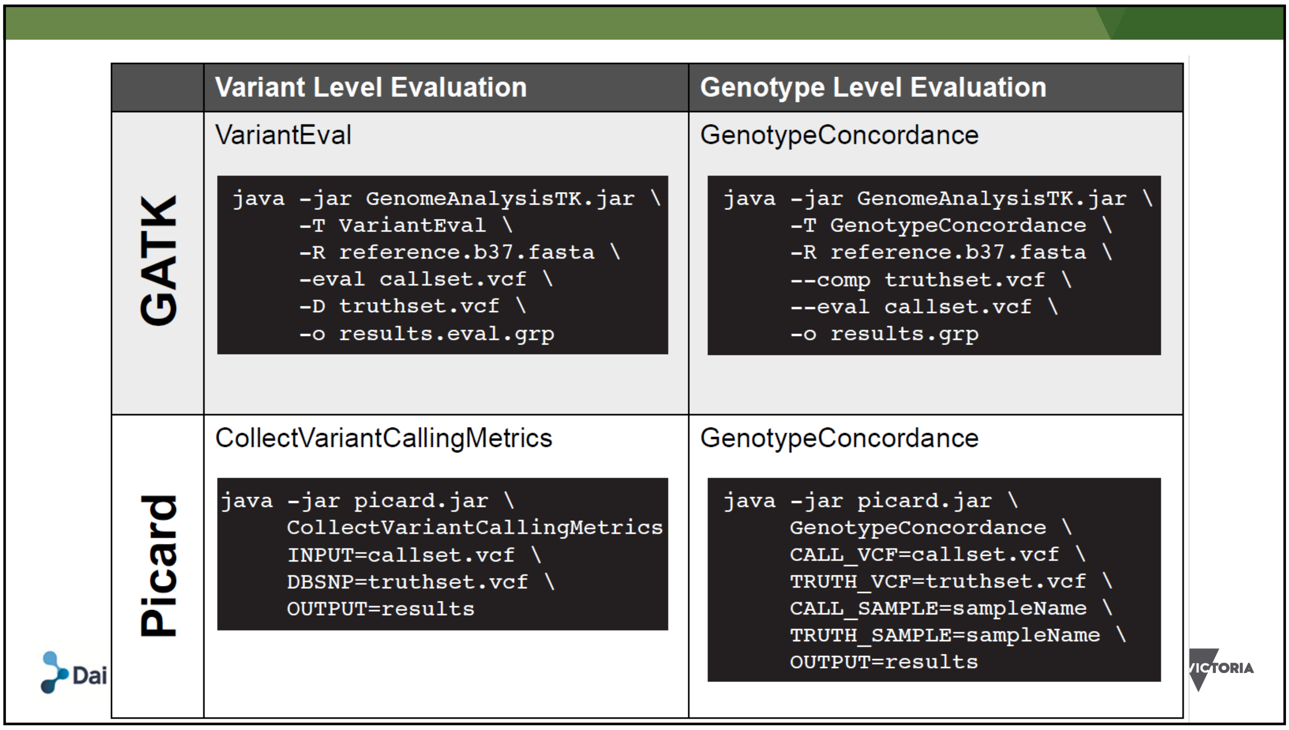
☆ heterozygous (0/1)
★ homozygous-variant (1/1)

GT CONCORDANCE

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

DairyBio

27



	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`

Dai

ICTORIA

28

Call set evaluations used for the 1000 Bull Genomes Project

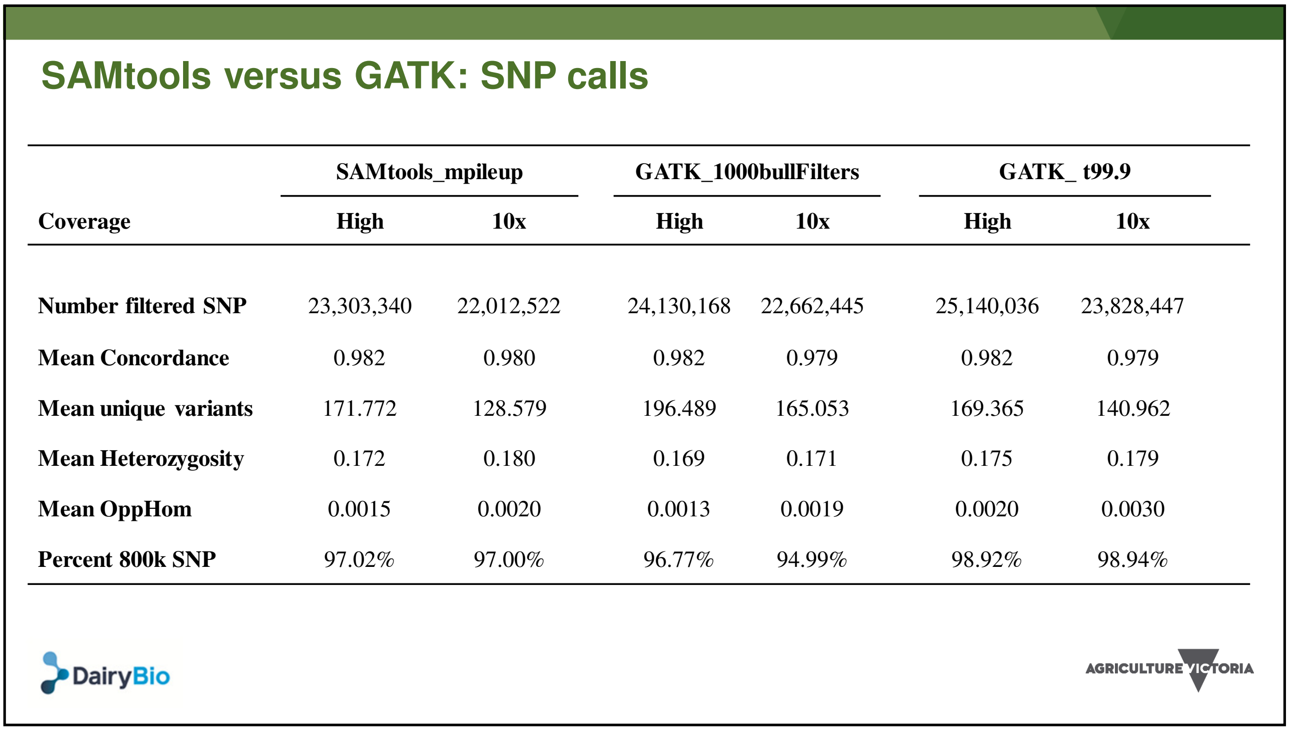
- Concordance to HD chip (~1000 samples)
- Opposing homozygotes
  - Requires pedigree information
- Number of unique variants
- Level of heterozygosity

29

Other variant callers

- SAMtools, FreeBayes, Platypus, VarScan and more…
- Comparison between GATK and SAMtools
  - http://www.wcgalp.org/system/files/proceedings/2018/which-best-variant-caller-large-whole-genome-sequencing-datasets.pdf

30



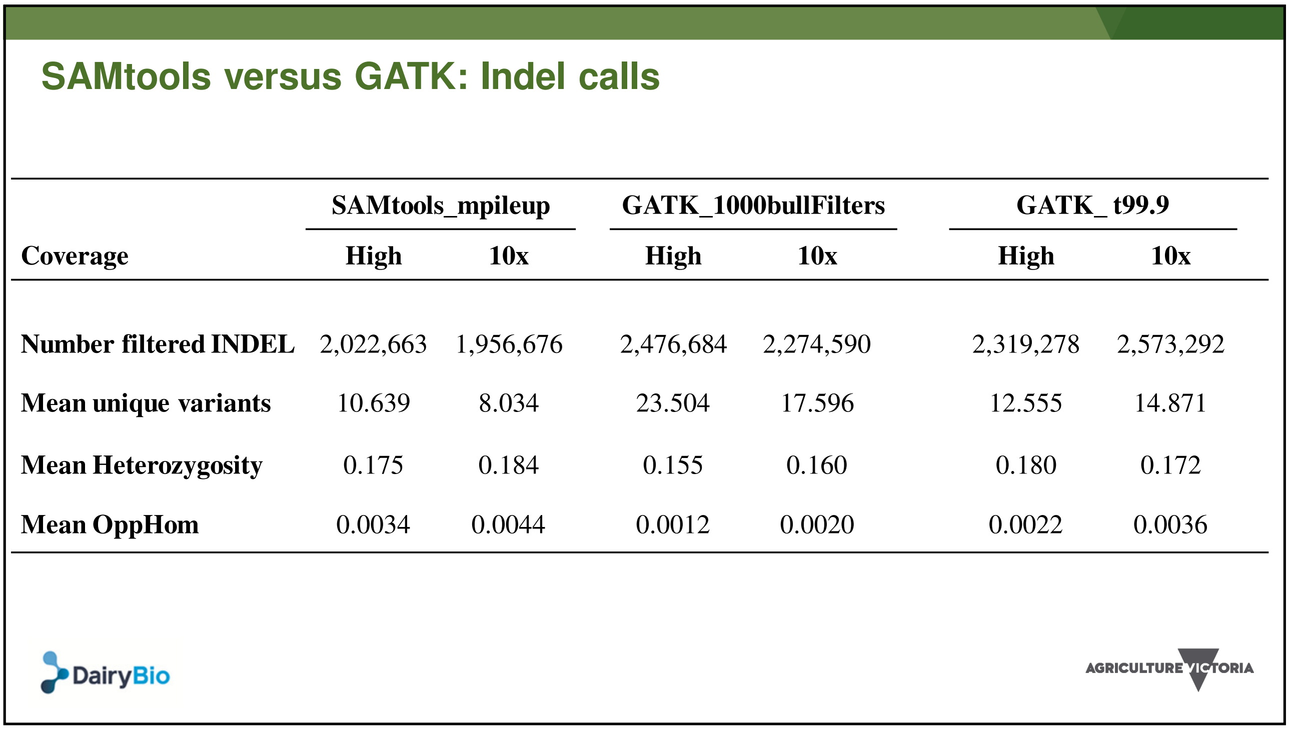
SAMtools versus GATK: SNP calls

	SAMtools_mpileup		GATK_1000bullFilters		GATK_ t99.9	
Coverage	**High**	**10x**	**High**	**10x**	**High**	**10x**
Number filtered SNP	23,303,340	22,012,522	24,130,168	22,662,445	25,140,036	23,828,447
Mean Concordance	0.982	0.980	0.982	0.979	0.982	0.979
Mean unique variants	171.772	128.579	196.489	165.053	169.365	140.962
Mean Heterozygosity	0.172	0.180	0.169	0.171	0.175	0.179
Mean OppHom	0.0015	0.0020	0.0013	0.0019	0.0020	0.0030
Percent 800k SNP	97.02%	97.00%	96.77%	94.99%	98.92%	98.94%

DairyBio

AGRICULTURE VICTORIA

31



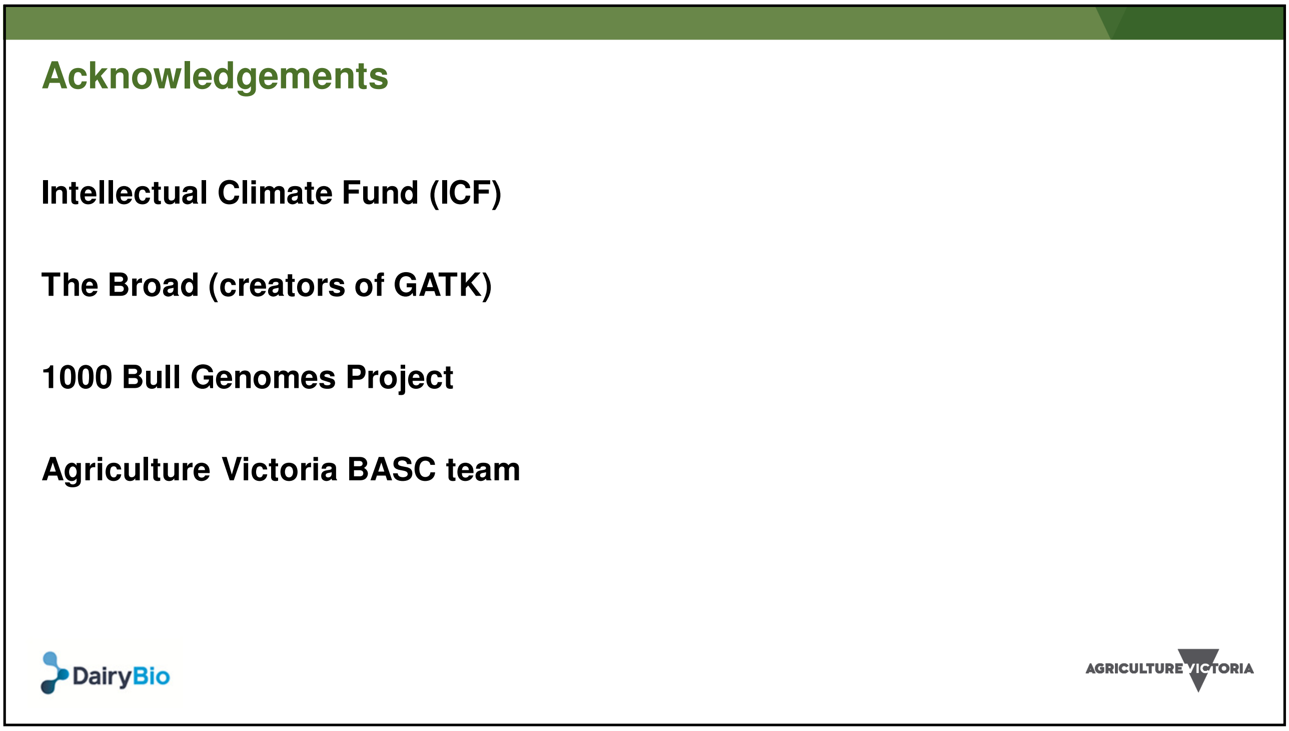
SAMtools versus GATK: Indel calls

	SAMtools_mpileup		GATK_1000bullFilters		GATK_ t99.9	
Coverage	**High**	**10x**	**High**	**10x**	**High**	**10x**
Number filtered INDEL	2,022,663	1,956,676	2,476,684	2,274,590	2,319,278	2,573,292
Mean unique variants	10.639	8.034	23.504	17.596	12.555	14.871
Mean Heterozygosity	0.175	0.184	0.155	0.160	0.180	0.172
Mean OppHom	0.0034	0.0044	0.0012	0.0020	0.0022	0.0036

DairyBio

AGRICULTURE VICTORIA

32



Acknowledgements

Intellectual Climate Fund (ICF)

The Broad (creators of GATK)

1000 Bull Genomes Project

Agriculture Victoria BASC team

DairyBio

AGRICULTURE VICTORIA

33

Thank you

AGRICULTURE VICTORIA

34