Sequence Data

SISG 2017

Ben Hayes and Hans Daetwyler

Using sequence data in genomic selection

- Generation of sequence data (Illumina)
- Characteristics of sequence data
- Quality control of raw sequence
- Alignment to reference genomes
- Variant Calling

Using sequence data in genomic selection and GWAS

- Motivation

- Genome wide association study
  - Straight to causative mutation
  - Mapping recessives

- Genomic selection (all hypotheses!)
  - No longer have to rely on LD, **causative mutation actually in data set**
    - Higher accuracy of prediction?
    - Not true for genotyping-by-sequencing
  - Better prediction across breeds?
    - Assumes same QTL segregating in both breeds
    - No longer have to rely on SNP-QTL associations holding across breeds
  - Better persistence of accuracy across generations

Using sequence data in genomic selection and GWAS

- Motivation

- Genotype by sequencing
  - Low cost genotypes?
  - Need to understand how these are produced, potential errors/challenges in dealing with this data

Technology – NextGenSequence

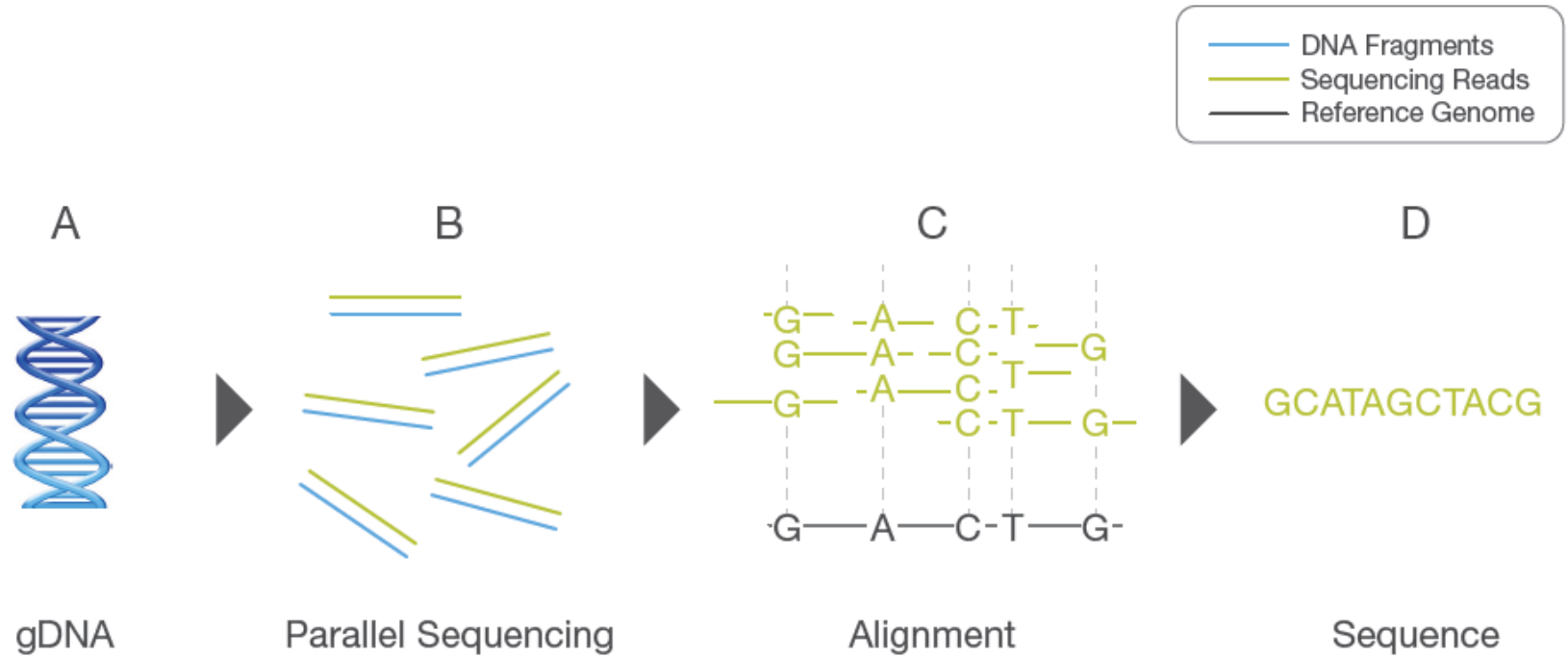
• Over the past few years, the “Next Generation” of sequencing technologies has emerged.

Roche: GS-FLX

Applied Biosystems: SOLiD 4

Illumina: HiSeq3000, MiSeq

Figure 1: Conceptual Overview of Whole-Genome Resequencing



A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

An Introduction to Next-Generation Sequencing Technology

Illumina, accessed August 2013

Sequencing Workflow

- Extract DNA
- Prepare libraries
  - ‘Cutting-up’ DNA into short fragments
- Sequencing by synthesis (HiSeq, MiSeq, NextSeq, …)
- Base calling from image data -> **fastq**
- Alignment to reference genome -> **bam**
  - Or de-novo assembly
- Variant calling -> vcf

‘Raw’ sequence to genotypes

Illumina HiSeq 2000

FASTQ File
- Unfiltered
- Quality metrics for all bases

- *Quality Control*
- *Filters*
- *Align to reference*

SAM/BAM File for each FASTQ

- *Remove PCR duplicates*
- *Locally Realign*
- *Sort*
- *Index*
- *Merge*

QC’d merged SAM/BAM File

- *Variant calling*

Variant Call Format (VCF) File
- SNP and Indels
- Genotypes
- Quality metrics
  - Variants
  - Genotypes

‘Raw’ sequence to genotypes

Illumina HiSeq 2000

FASTQ File
- Unfiltered
- Quality metrics for all bases

- *Quality Control*
- *Filters*
- *Align to reference*

SAM/BAM File for each FASTQ

- *Remove PCR duplicates*
- *Locally Realign*
- *Sort*
- *Index*
- *Merge*

QC’d merged SAM/BAM File

- *Variant calling*

Variant Call Format (VCF) File
- SNP and Indels
- Genotypes
- Quality metrics
  - Variants
  - Genotypes

Raw data - FASTQ Files

```
@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCACTCCAATATTCTTGCCTGGAGAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
@<;DDD;;DFDFDEGFHDFGIEHE<B<<CEGEF=;?CGAG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACXX:5:1101:1788:2094 2:N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTCCTATAATGCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHGJJIICCGHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:2150:2031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAACAGTAGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIHHIIJJJIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:1101:2415:2093 2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAAGTTAAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJJIJJJIHHIJJJJIJJJJJJJJJIJJJJIGJIJJJIJJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:1101:2445:2236 2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAGGGTCTTTTCCAATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD<F<ECBAA3A<AHFAF@H<@91?CFHIEHHAGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############
```

.

Raw data - FASTQ Files (Casava1.8)

```
@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCA[illegible]CAATATTCTTGCCTGGAGAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
@<;DDD;;DFDFDEGFHD[illegible]FHE<B<<CEGEF=;?CGAG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACX[illegible]101:1788:2094 2:N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTC[illegible]AATGCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJ[illegible]HGJJIICCGHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1[illegible]2150:2031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAA[illegible]CAGTAGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIH[illegible]IJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:1101:[illegible]:2093 2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAAGTTA[illegible]AATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJJIJJJIHHIJJJJJIJJ[illegible]IJJJJJIGJIJJJIJJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:1101:2445:[illegible]6 2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAGGGTCTTTTCCAA[illegible]GTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD<F<ECBAA3A<AHFAF[illegible]@91?CFHIEHHAGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############
```

@HISEQ:185:D2E63ACXX:5:1101:1518:2222

@InstrumentName:RunID:FlowCellID:FlowCellLane:TileNumber:x-CoordInTile:y-CoordInTile

Raw data - FASTQ Files (Casava1.8)

@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCACTCCAATATTCTTGCCTGG[illegible]GAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
@<;DDD;;DFDFDEGFHDFGIEHE<B<<CEGEF=[illegible]GAG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACXX:5:1101:1788:20[illegible]:N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTCCTATAATGCAG[illegible]TCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHGJJI[illegible]HGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:21[illegible]031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAA[illegible]AGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIH[illegible]JIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:11[illegible]415:2093 2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAA[illegible]AAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJIJJ[illegible]JJJJJIJJJJJJJJJJIJJJJIGJIJJJIJJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX[illegible]101:2445:2236 2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAC[illegible]CTTTTCCAATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD[illegible]CBAA3A<AHFAF@H<@91?CFHIEHHAGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############

2:N:0:AATCGTGGTACGGTGA

WhichMemberOfPair:FailedFilterYorN:ControlBitNumber:Barcode

Raw data - FASTQ Files (Casava1.8)

```
@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCACTCCAATATTCTTGCCTGGAGAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
@<;DDD;;DFDFDEGFHDFGIEHE<B<<CEGEF=  G<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACXX:5:1101:1788:209  N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTCCTATAATGCAGG  TCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHGJJI  HGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:215  031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAAC  AGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIH  JIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:110  415:2093 2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAA  AAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJIJJJ  JJJJJIJJJJJJJJJJIJJJJIGJIJJJIJJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:  101:2445:2236 2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAG  CTTTTCCAATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD  CBAA3A<AHFAF@H<@91?CFHIEHHAGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############
```

GGAAATGGCAACCCACTCCAATATTCTTGCCTG….

Observed Bases, N for no call

Raw data - FASTQ Files (Casava1.8)

@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCACTCCAATATTCTTGCCTGGAGAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
DD;;DFDFDEGFHDFGIEHE<B<<CEGEF=;?CGAG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
Q:185:D2E63ACXX:5:1101:1788:2094 2:N:0:AATCGTGGTACGGTGA
AAAAGAATCTTCCTATAATGCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DD DFHEHIJJJIJIJGBHHGJJIICCGHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:18 63ACXX:5:1101:2150:2031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACT AGGGAAACAACAGTAGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJC HIIJCIJIHHIIJJJIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63A :1101:2415:2093 2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAA GTTAAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJJIJJJIH IIJJJJJJJJJJIJJJJIGJIJJJIJJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:110 15:2236 2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAGGGTCTTT ATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD<F<ECBAA3A<AH H<@91?CFHIEHHAGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############

+

Next line contains phred quality scores for all bases

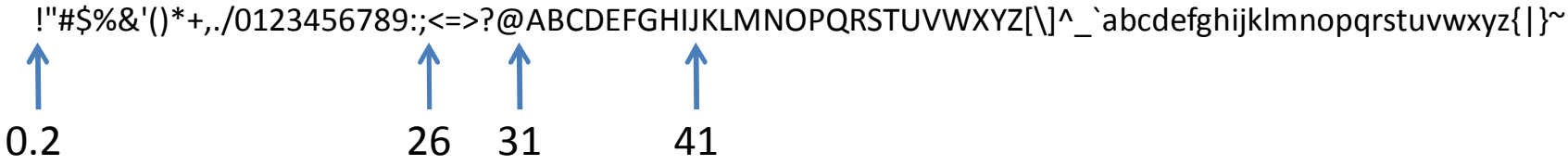
Raw data - FASTQ Files (Casava1.8)

```
@HISEQ:185:D2E63ACXX:5:1101:1518:2222 2:N:0:AATCGTGGTACGGTGA
GGAAATGGCAACCCACTCCAATATTCTTGCCTGGAGAATTCCATGGACAGAGGAGCCTGGCAGACTACAGTCCATGGGGCTGCAAAGAGTCAGACACAACT
+
@<;DDD;;DFDFDEGFHDFGIEHE<B<<CEGEF=;?CGAG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACXX:5:1[illegible]1788:2094 2:N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTCCTATA[illegible]CAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHG[illegible]CGHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:215[illegible]031 2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAACAG[illegible]CATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIHHIIJJJIJJJIE[illegible]=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:1101:2415:2093[illegible]0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAAGTTAAATAATAT[illegible]TGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJJIJJJIHHIJJJJIJJJJJJJJJJIJJJJIGJ[illegible]U=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:1101:2445:2236 2:N:0:A[illegible]GTGGTACGGTGA
AATCTTTCTCAGCATGAGGGTCTTTTCCAATGAGTCAGCT[illegible]GCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD<F<ECBAA3A<AHFAF@H<@91?CFHI[illegible]AGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############
```

@<;DDD;;DFDFDEGFHDFGIEHE<B<<CEGEF=;?CGAG<BBHEGH9BFF;?DF; …

Phred quality scores (Q scores) for all bases in read

FASTQ quality symbols and Phred scores



- Raw Illumina reads generally do not go above 41
- Q scores are calculated from error probabilities are calculated by the following formula:
- Phred = $-10*\log_{10}(\text{error rate})$
  - e.g. Phred of 30 = error rate of 0.001
  - Phred of 20 = error rate of 0.01

Quality control at FASTQ stage

- Trim read ends based on quality
  - Remove base calls < 20 phred
  - Why?
    - Base calls deteriorate at read ends

2:N:0:AATCGTGGTACGGTGA

- Remove reads that fail chastity filter
  - Yes Y → read fails. No N → read passes
  - Chastity filter will fail if more than three bases are N (no Call) within first 25 base calls of read
  - Why?
    - Ns due to low quality signal from colour clusters
      - Ratio of one base to all others unclear
      - Could indicate overloading of flow cell

Quality control at FASTQ stage

After trimming of reads

- Remove reads with <20 mean phred quality score
  - Why?
    - Whole read low quality

- Remove reads that are less than 50% of original read length
  - Why?
    - Short reads are hard to uniquely map
    - Quality questionable

- Remove reads with more than 3 Ns
  - Why?
    - Unreliable base calls

QC and Visualisation of Raw Sequence Fastq files

Å Practical 1a on galaxy

‘Raw’ sequence to genotypes

Illumina HiSeq 2000

FASTQ File
- Unfiltered
- Quality metrics for all bases

- *Quality Control*
- *Filters*
- *Align to reference*

SAM/BAM File for each FASTQ

- *Remove PCR duplicates*
- *Locally Realign*
- *Sort*
- *Index*
- *Merge*

QC’d merged SAM/BAM File

- *Variant calling*

Variant Call Format (VCF) File
- SNP and Indels
- Genotypes
- Quality metrics
  - Variants
  - Genotypes

Alignment of sequence

• If a “reference” genome exists for the organism you are sequencing, reads can be “aligned” to the reference

• This involves finding the place in the reference genome that each read matches to

• Due to high sequence similarity within members of the same species, most reads *should* map to the reference

- Quality of reference genome will influence how much of sequence will map and how good your variant calls are

Tools for generating alignments

• There are MANY software packages available for aligning data from next generation sequencing experiments

•Which software depends on the data you are analysing and the results you hope to achieve

Software Examples: BWA, MAQ, Bowtie, NovoAlign, BFAST, ELAND, MOSAIK, SHRiMP, SOAP, SSAHA and BLAST,…

Å Some aligners will take first match, some other determine best match but are slower

Alignment Steps

- Map each FASTQ file to reference
  - Separately for paired-end (PE) and single-end (SE) reads
  - Generates .sai file
  - Convert to SAM file (sequence alignment format)
  - Convert to BAM file (binary form of sam)
- Merge PE and SE .bam file
  - Sequence alignment map format

QC and Visualisation of a BAM files

Å Practical 1b on galaxy

‘Raw’ sequence to genotypes

Illumina HiSeq 2000

FASTQ File
- Unfiltered
- Quality metrics for all bases

- *Quality Control*
- *Filters*
- *Align to reference*

SAM/BAM File for each FASTQ

- *Remove PCR duplicates*
- *Locally Realign*
- *Sort*
- *Index*
- *Merge*

QC’d merged SAM/BAM File

- *Variant calling*

Variant Call Format (VCF) File
- SNP and Indels
- Genotypes
- Quality metrics
  - Variants
  - Genotypes

Variant Calling

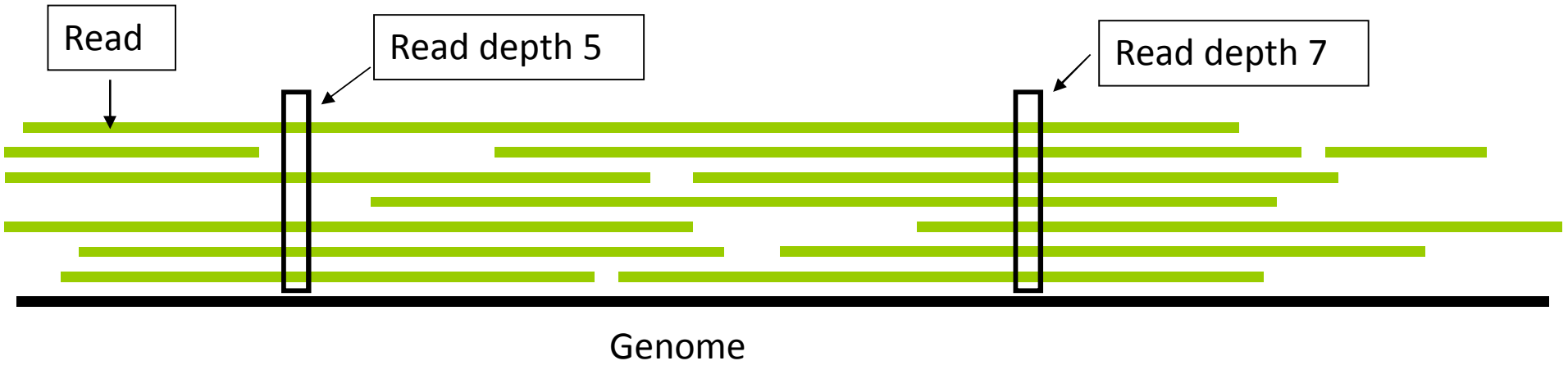
- Using all sequenced individuals we want to:
  - Identify ‘all’ variants; SNP and Indels
  - Genotype all individuals for those variants

- Importance of read depth

Read depth

- Once aligned we can investigate how well our reads cover the genome

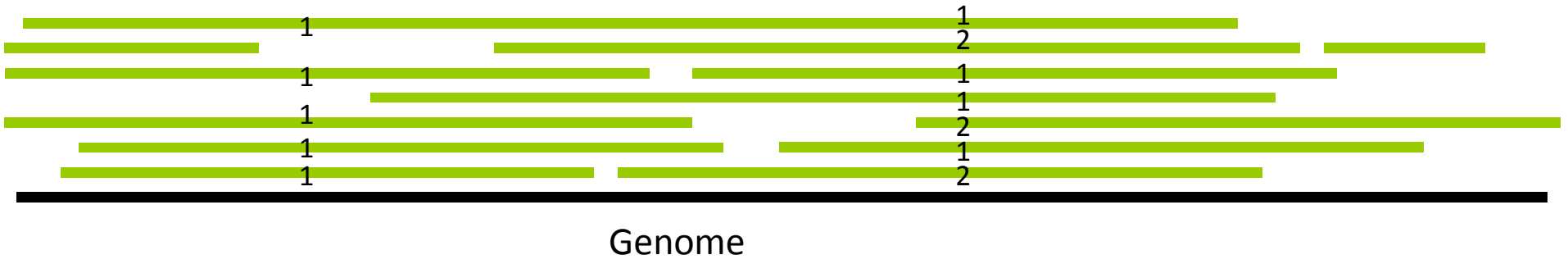
- Read depth or fold coverage



Importance of read depth

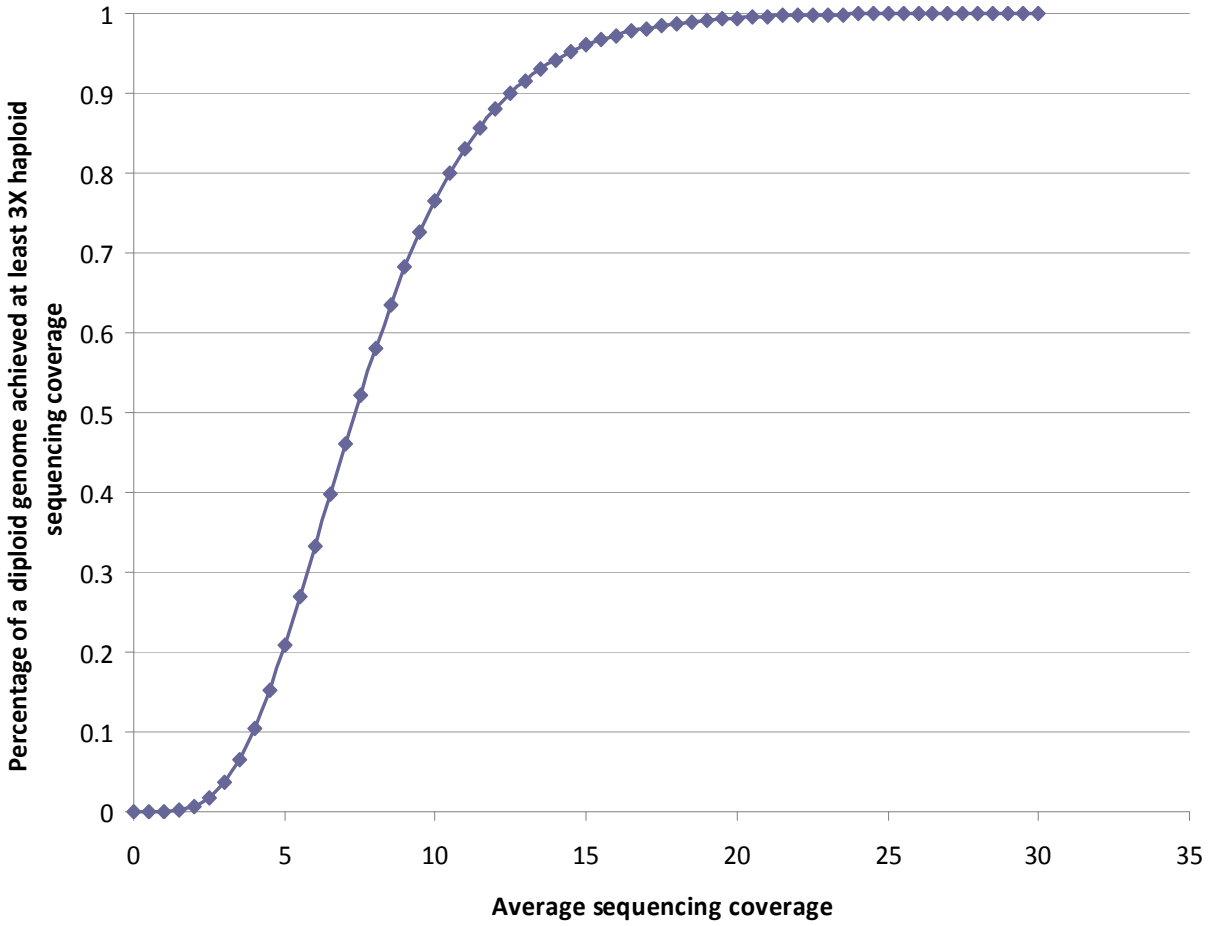
- Consider a diploid heterozygous locus (individual carries 2 different alleles)
  - ï 50/50 chance of observing each allele in every read

- If read depth is low, it is possible to not observe an allele and therefore call a heterozygous locus homozygous → errors
  - ï Read depth 5 → $0.5^5 = 0.03125$



What read depth is sufficient

- Proportion of genome achieving at least 6x diploid coverage
- 12.5x achieves 90% in simulation below (Shen et al. 2010, Suppl. Material)
- In a Japanese bull, 16x achieved 93% coverage (Kawahara-Miki et al. 2011)



Heterozygosity and read depth

- Importance of read depth for SNP discovery or accurate genotype calling
  - SNP discovery
    - Missing some heterozygotes is not critical
      - Hopefully picked up in other individuals
    - Just do more individuals to identify SNP
    - Individual genotype not used directly
  - Genotype calling
    - Missing heterozygotes a problem because incorrect genotype included in downstream analysis
    - Statistical methods can be used to correct incorrect genotype calls

‘Raw’ sequence to genotypes

Illumina HiSeq 2000

FASTQ File
- Unfiltered
- Quality metrics for all bases

- *Quality Control*
- *Filters*
- *Align to reference*

SAM/BAM File for each FASTQ

- *Remove PCR duplicates*
- *Locally Realign*
- *Sort*
- *Index*
- *Merge*

QC’d merged SAM/BAM File

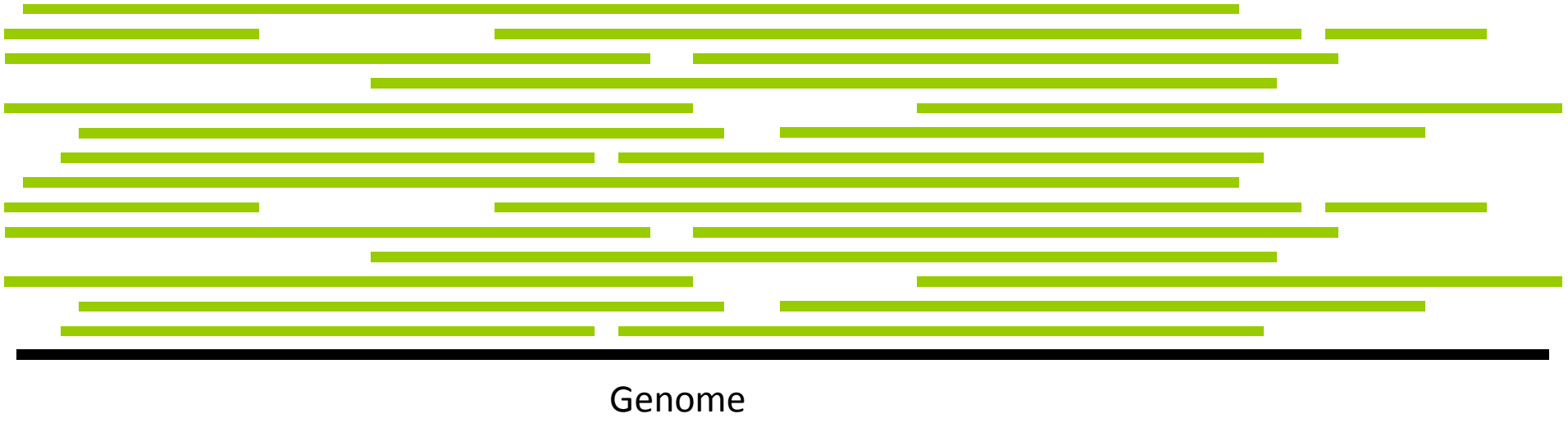
- *Variant calling*

Variant Call Format (VCF) File
- SNP and Indels
- Genotypes
- Quality metrics
  - Variants
  - Genotypes

Identification of variants

Å Program SAMtools

- stacks aligned bam files of multiple individuals
- Calls variants and calculates quality/confidence statistics for calls
- http://samtools.sourceforge.net/mpileup.shtml



Samtools

Samtools provides a command line interface for manipulation of SAM/BAM formatted data.
(http://samtools.sourceforge.net)

- Open source and multi-platform (R package available: Rsamtools).

- Able to:
  - Extract reads from specific genomic region
  - Sort, index, merge bams
  - Visualise bams (command line)
  - Call variants
  - etc

Variants in sequence

- SNP
- INDEL
  - ï INsertions and DELetions of DNA sections
- Copy number variants (CNV)
  - ï Repeated sections of DNA of various lengths
- Most studies to date have concentrated on SNP

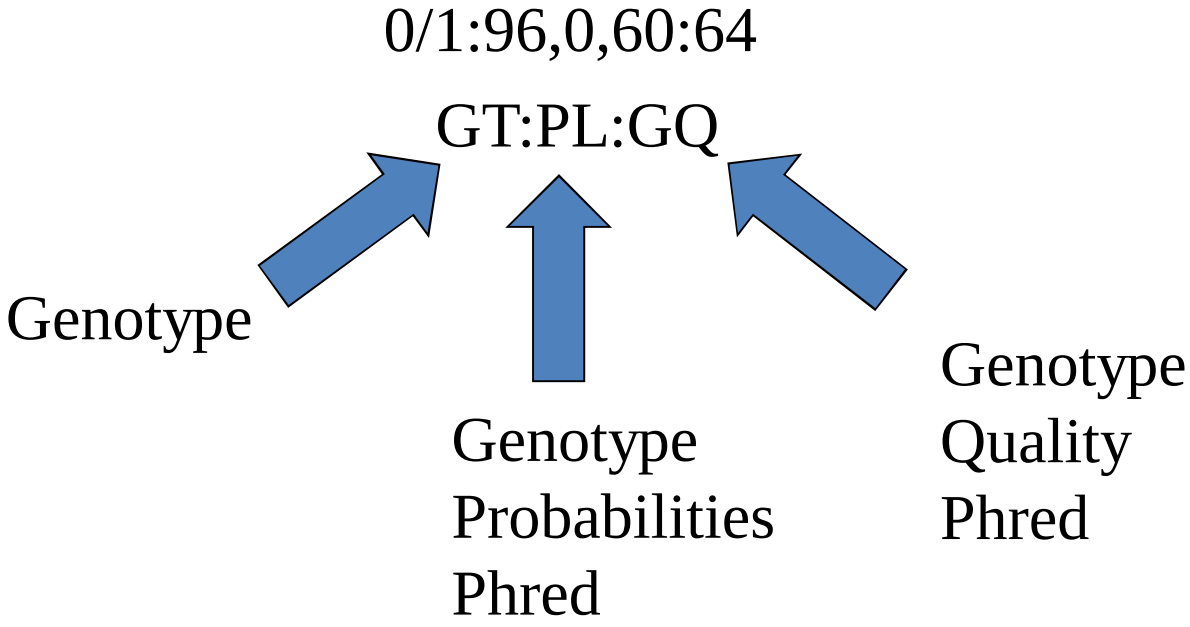
Variant Call Format (VCF) file

Å ##fileformat=VCFv4.1
Å ##samtoolsVersion=0.1.18 (r982:295)
Å ##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
Å ##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
Å ##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
Å ##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
Å ##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
Å ##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
Å ##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
Å ##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
Å ##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
Å ##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
Å ##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
Å ##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
Å ##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
Å ##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than in group2.">
Å ##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
Å ##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
Å ##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
Å ##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
Å ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
Å ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
Å ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
Å ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
Å ##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
Å ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">

Å	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
		FORMAT	Individual1	Individual2				
Å	Chr29	39484430	.	C	A	277		
		DP=30;VDB=0.0178;AF1=0.4455;AC1=5;DP4=11,4,7,8;MQ=52;FQ=279;PV4=0.26,0.43,0.0066,0.11			GT:PL:GQ	0/0:0,9,113:8	0/1:96,0,60:64	
Å	Chr29	39484455	.	TGGG	TGG	18.6	INDEL;DP=23;VDB=0.0316;AF1=0.2602;G3=0.75,1.412e-	
	06,0.25;HWE=0.0458;AC1=2;DP4=12,3,4,2;MQ=54;FQ=19.8;PV4=0.6,1,6.4e-05,1				GT:PL:GQ	0/0:0,9,90:11	0/0:0,9,93:11	
Å	Chr29	39484540	.	A	G	999	.	
		DP=44;VDB=0.0356;AF1=0.588;AC1=6;DP4=7,8,14,14;MQ=46;FQ=999;PV4=1,1,0.079,1			GT:PL:GQ	0/0:0,15,157:11	0/1:101,0,81:83	
Å	Chr29	39484790	.	T	A	408	.	
		DP=33;VDB=0.0381;AF1=0.6663;AC1=7;DP4=6,2,14,11;MQ=50;FQ=413;PV4=0.43,0.21,0.0055,0.31			GT:PL:GQ	0/0:0,9,85:5	0/1:0,0,0:3	
Å	Chr29	39484791	.	A	C	999	.	
		DP=33;VDB=0.0381;AF1=0.6663;AC1=7;DP4=6,2,13,11;MQ=50;FQ=999;PV4=0.42,1,0.0069,0.33			GT:PL:GQ	0/0:0,9,88:5	0/1:0,0,0:3	

VCF file (genotype probabilities)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Individual1	Individual2
Chr29	39484430	.	C	A	277	.		LocusStats GT:PL:GQ	0/0:0,9,113:8	0/1:96,0,60:64
Chr29	39484455	.	TGGG	TGG	18.6	.	INDEL	LocusStats GT:PL:GQ	0/0:0,9,90:11	0/0:0,9,93:11
Chr29	39484540	.	A	G	999	.		LocusStats GT:PL:GQ	0/0:0,15,157:11	0/1:101,0,81:83
Chr29	39484790	.	T	A	408	.		LocusStats GT:PL:GQ	0/0:0,9,85:5	0/1:0,0,0:3
Chr29	39484791	.	A	C	999	.		LocusStats GT:PL:GQ	0/0:0,9,88:5	0/1:0,0,0:3



VCF file (FORMAT - locus quality stats)

In field FORMAT

DP=30; Read depth

VDB=0.0178; Variant distance bias

AF1=0.4455; Maximum likelihood estimate of 1st alternative allele frequency

AC1=5; ML estimate of 1 alternative allele count

DP4=11,4,7,8; Number reads on: Ref-Forward, Ref-Reverse, Alt-Forward, Alt-Reverse

MQ=52; Mapping quality

FQ=279; Phred probability of all samples being the same

PV4=0.26,0.43,0.0066,0.11

P-values for strand bias, baseQ bias, mapQ bias and tail distance bias

Filtering of variants

Reasons for filters:

- Number of artefacts of the sequencing process that lead to falsely identified variants

- Little evidence for a variant
  - ï Quality scores low

Reasons against filters:

- Real variants may be lost
  - ï Low frequency SNP often have lower quality scores

Variant filters

- Read depth
- Minimum read depth
  - Require >5 reads
  - Why?
    - Individual genotype calls will be low quality
- Maximum read depth
  - Short reads of repetitive regions may be mapped to same locations causing massive read depth
  - Why?
    - Reference assembly problems
    - If regions are repeats of 100+ bases, difficult to map uniquely

Variant filters

- Opposing homozygotes
  - Check Parent-Offspring pairs
    - Mendelian Rules
    - If father is homozygous then offspring cannot be homozygous for opposite allele at same locus
  - Why?
    - Inconsistencies due to poor mapping of reads
    - Likely in repetitive genome areas with assembly issues

Additional filters?

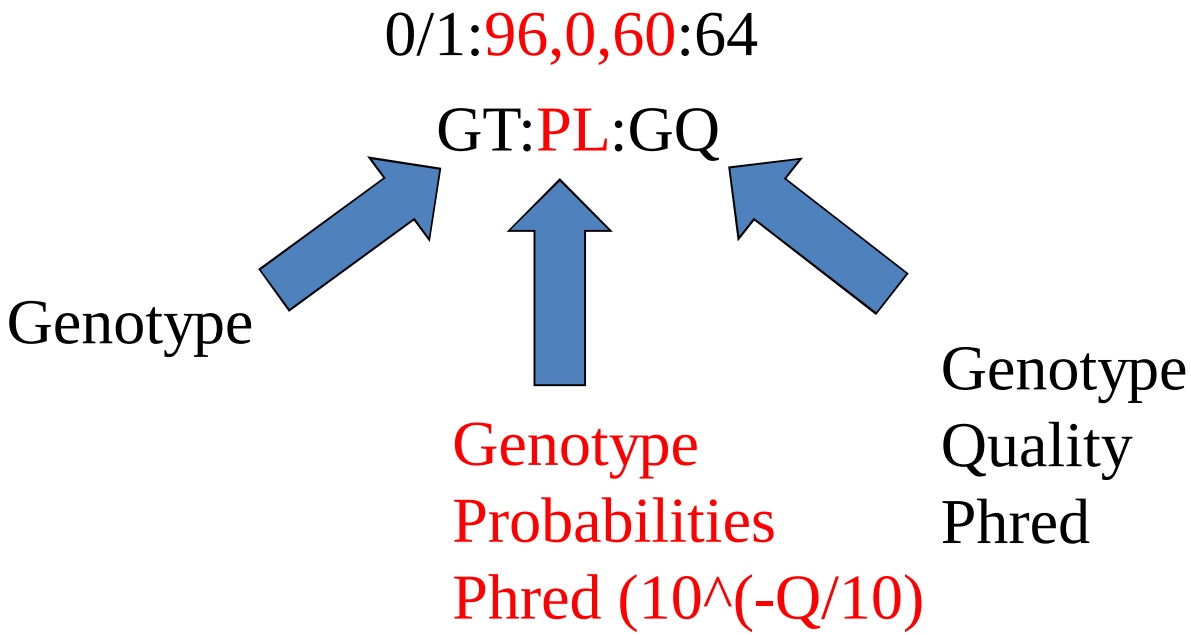
- Require minor allele is in at least 2 animal in sample
  - ï BUT will lead to a threshold on minor allele frequency
  - ï E.g. 50 seq. animals → 2/50=0.04
    - Å Thus, MAF cut off is 0.04
- Heterozygosity (Hardy-Weinberg)
- etc

Genotype correction/imputation with phasing/imputation software

- We now have a VCF file with filtered variants
- Could just use genotypes given
- BUT, some individuals will have no reads at certain positions, others have poor evidence to call heterozygotes…
- Can correct with imputation/phasing using genotype probabilities

VCF file (genotype probabilities)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Individual1	Individual2
Chr29	39484430	.	C	A	277	.		LocusStats GT:PL:GQ	0/0:0,9,113:8	0/1:96,0,60:64
Chr29	39484455	.	TGGG	TGG	18.6	.	INDEL	LocusStats GT:PL:GQ	0/0:0,9,90:11	0/0:0,9,93:11
Chr29	39484540	.	A	G	999	.		LocusStats GT:PL:GQ	0/0:0,15,157:11	0/1:101,0,81:83
Chr29	39484790	.	T	A	408	.		LocusStats GT:PL:GQ	0/0:0,9,85:5	0/1:0,0,0:3
Chr29	39484791	.	A	C	999	.		LocusStats GT:PL:GQ	0/0:0,9,88:5	0/1:0,0,0:3



Phred quality scores (Q)

- Related to base-calling error probabilities.
  Expressed in a range from 0 to 999 in our data.
- Probabilities are calculated by the following formula:
- e.g. Phred of 30 = error rate of 0.001
- Phred of 20 = error rate of 0.01

$$P = 10^{\frac{-Q}{10}}$$

- Result is probability of each genotype at each variant eg. AA=0.95 AT=0.05 TT=0.00
- Use these in BEAGLE, EMMAX!

Some GT probabilities (from samtools)

Genotype Prob. 0	Genotype Prob. 1	Genotype Prob. 2	GT called	GT most probable	Correction?
96	0	60	0/1	0/1	no
0	5	110	0/1	0/0	yes
0	50	60	0/0	0/0	no
0	0	0	0/1	?	Impute!

- Lowest phred is best.
- Phasing software considers GT probabilities and haplotypes in population

Example output: beagle dose file

	Starlite	Shotime	Goldsmith	Gravita	Orana	Beau	OVGM	Goldwyn	Starbuck
Chr1:62598	2	2	2	2	2	2	2	2	2
Chr1:62612	0.0036	0.0005	1	0	0.0001	0.983	0.0001	0.0001	0
Chr1:62635	0.45	1.0013	0.2088	0.05	0.997	1	1	1	0.9998
Chr1:63919	1.99	2	1.9829	2	1.9914	1.9892	1	1.9973	2

Prob(0)*0 + Prob(1)*1 + Prob(2)*2 = 1*0 + 0.0036*1 + 0*2 = 0.0036

- Use these instead of integer genotypes in analyses
  - Captures uncertainty in variant calls
- Haplotype (.phased in Beagle) file gives most probable genotype

Conclusion

- Before using sequence data in genomic prediction/GWAS, check quality!
- Phred scores, proportion of reads mapping
- Low sequence depth -> reduced probability of calling heterozygotes
- Convert VCF -> dose file for genomic prediction/GWAS, takes uncertainty in genotype calls into account

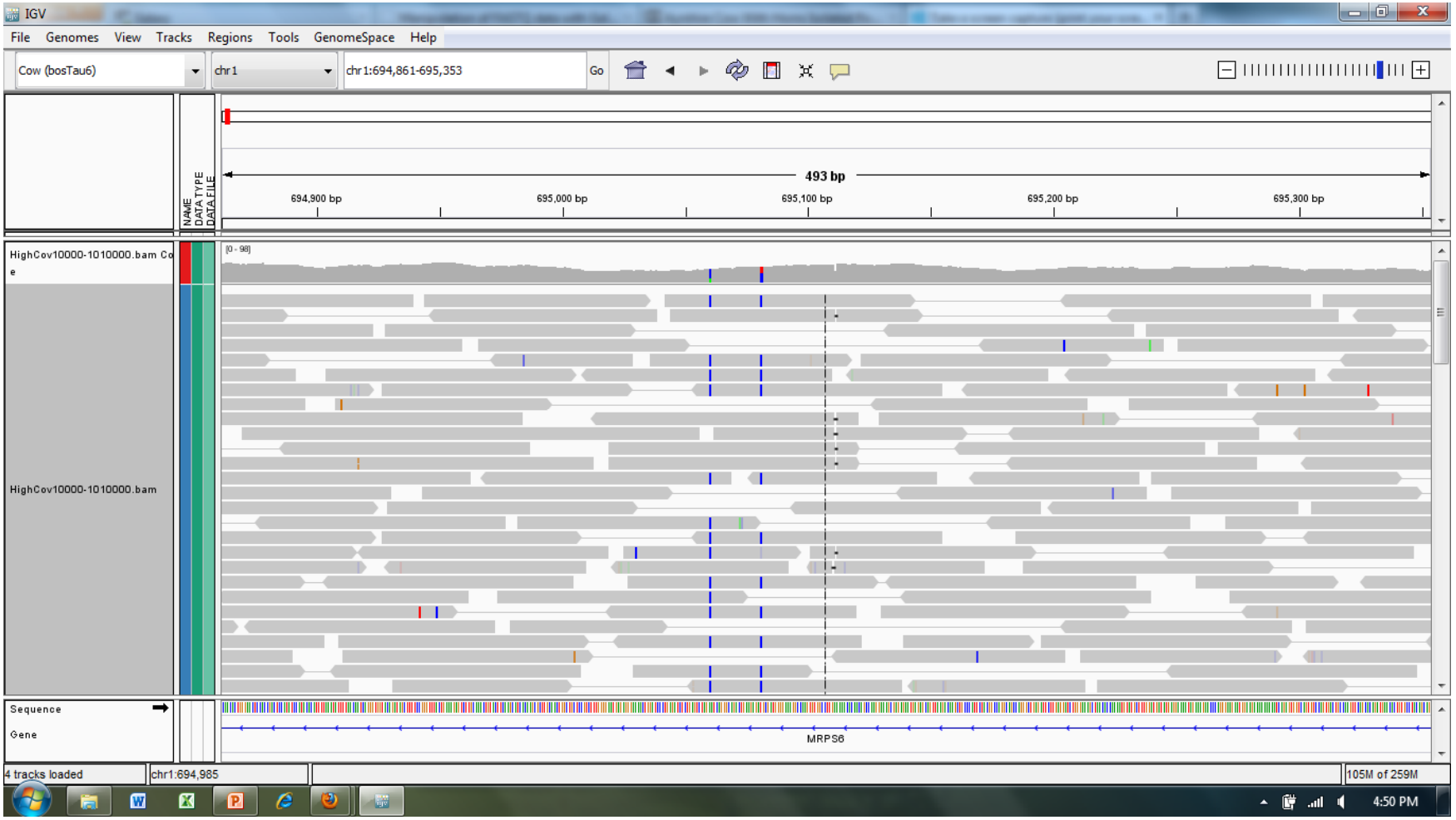
Useful papers

- Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 2011, 21:952

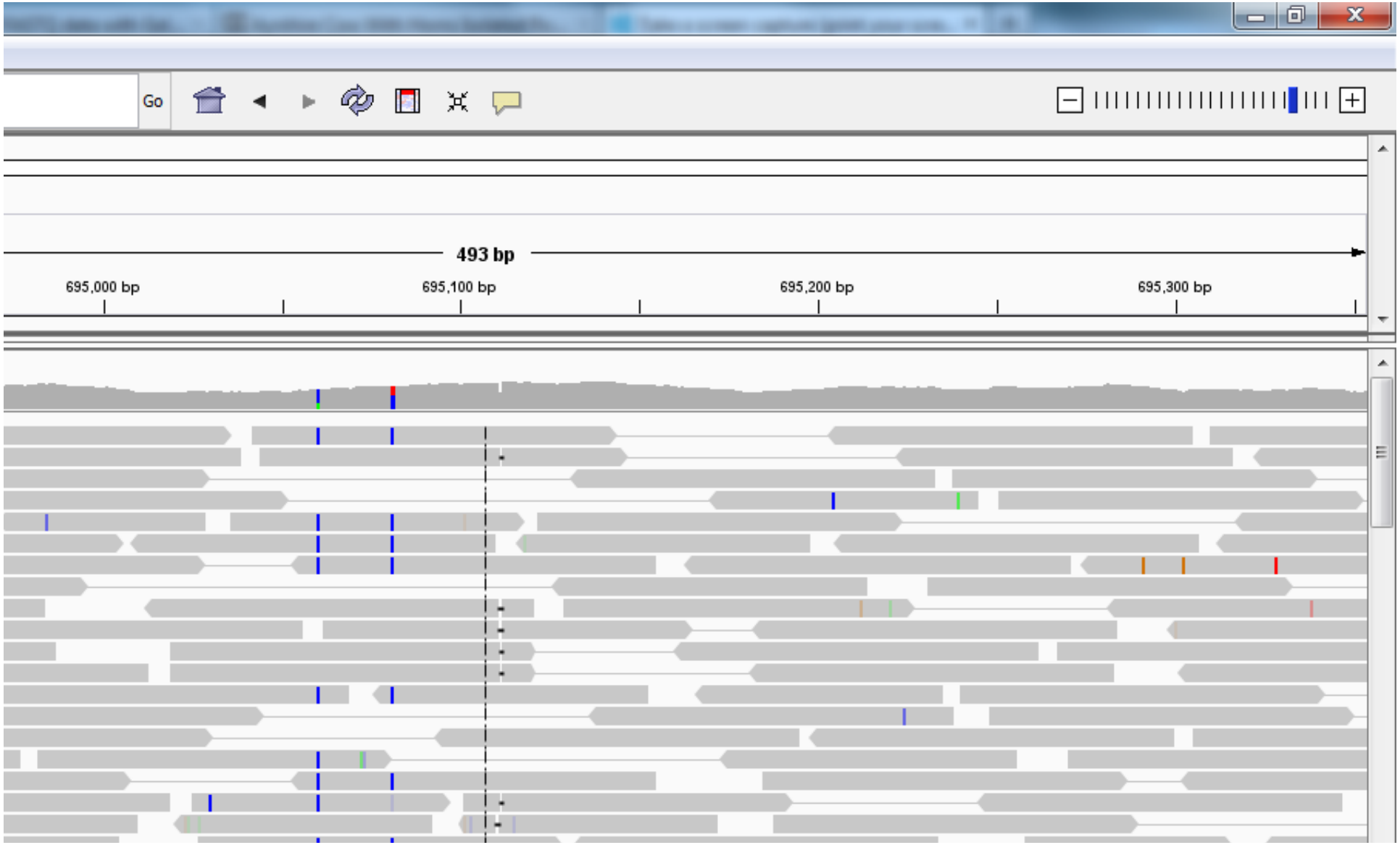
- Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 2014, 112:39

- Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016, 48:965 (STITCH)

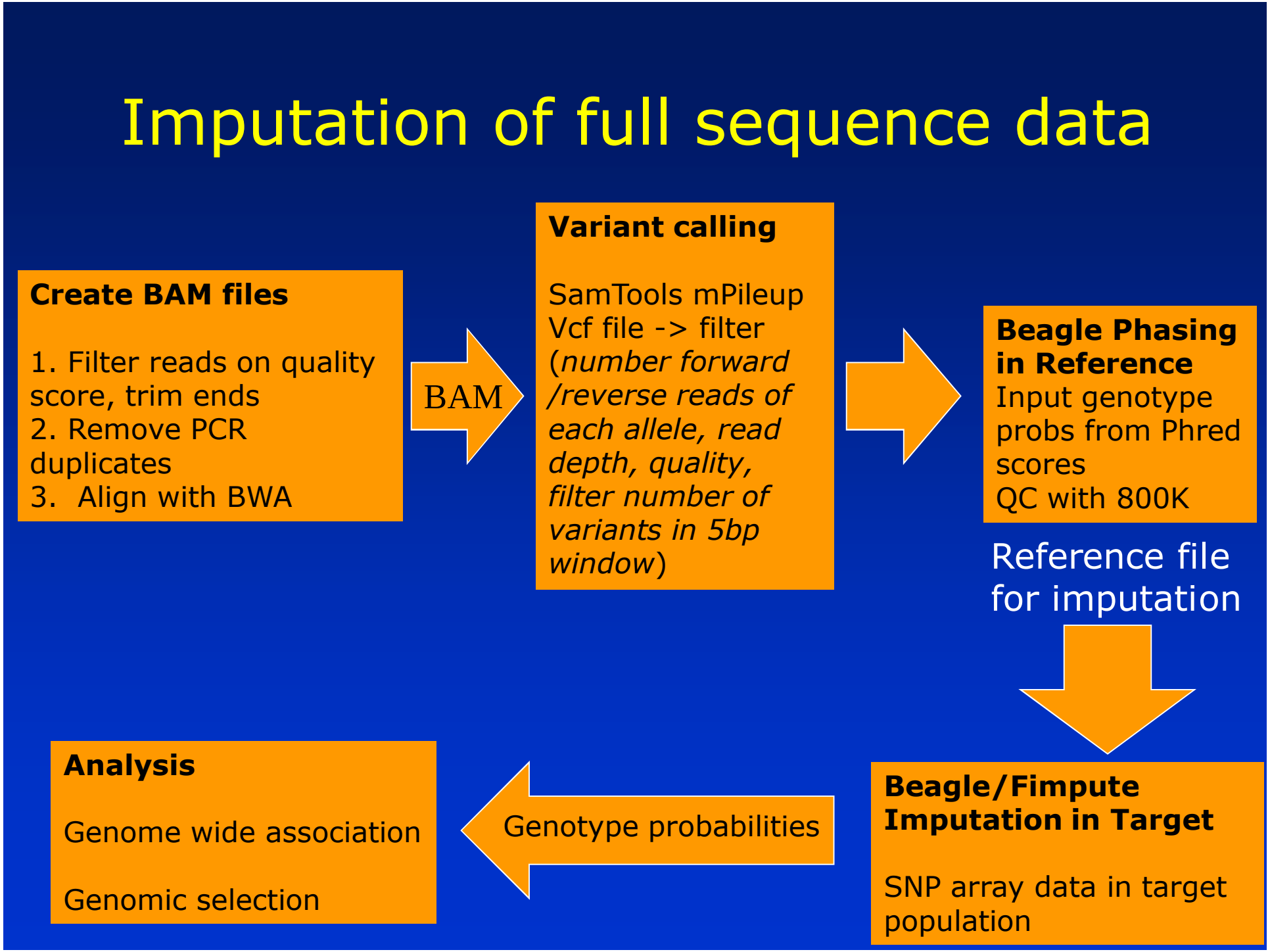
Viewing aligned reads with Integrated Genome Viewer (IGV, Broad Institute)



Viewing aligned reads with Integrated Genome Viewer (IGV, Broad Institute)



Imputation of full sequence data



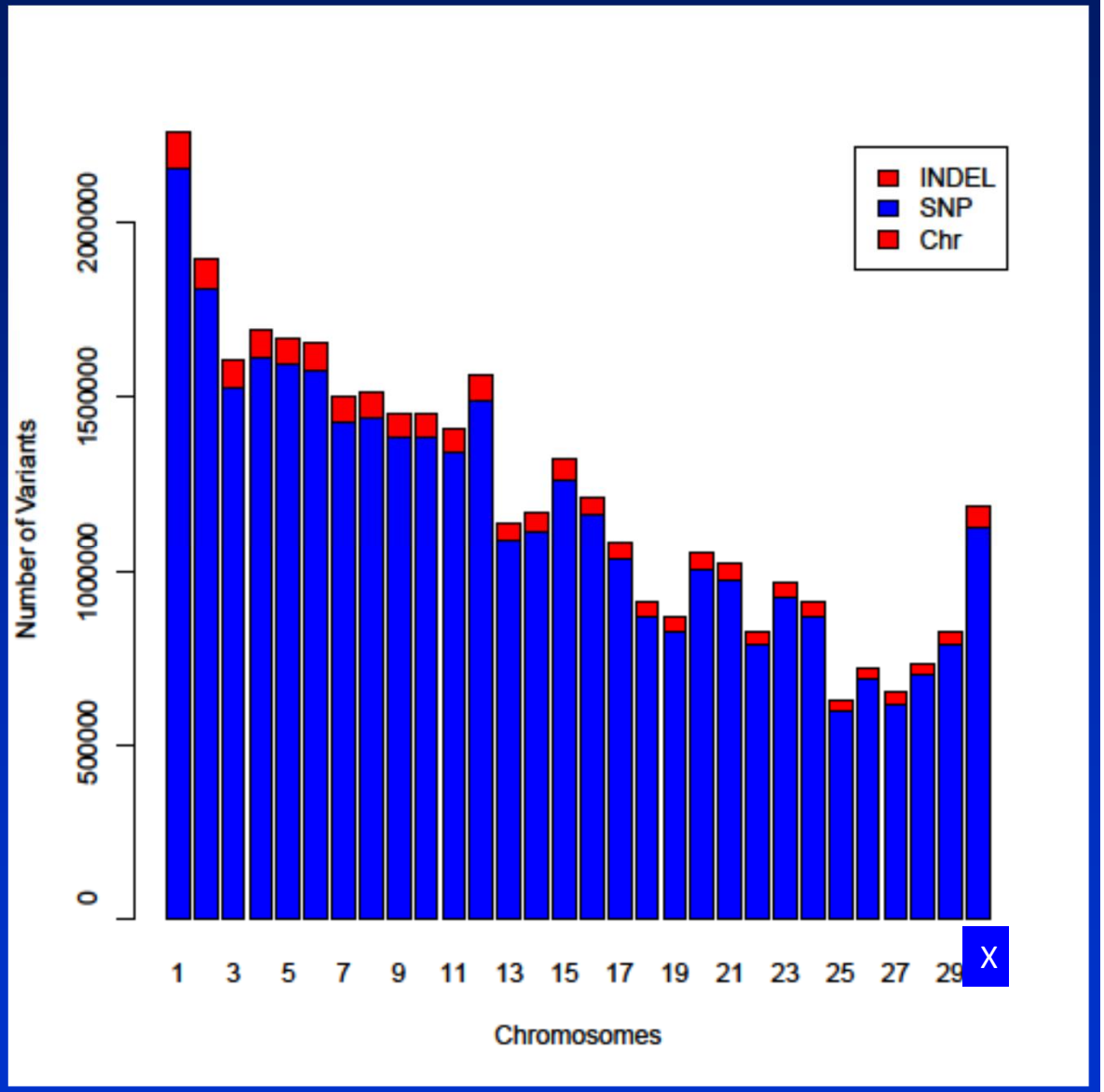
Run4.0 1000 bull genomes

- 1147 animals sequenced
- 27 breeds
- 20 Partners
- Average 11X

Breed/Cross	Number
Holstein (Black and White)	288
Simmental (Dual and Beef)	216
Angus (Black and Red)	138
Jersey	61
Brown Swiss	59
Gelbvieh	34
Charolais	33
Hereford	31
Limousin	31
Guelph Composite	30
Beef Booster	29
Alberta Composite	28
Montbeliarde	28
AyrshireFinnish	25
Normande	24
Holstein (Red and White)	23
Swedish Red	16
Danish Red	15
Other Crosses	11
Belgian Blue	10
Piedmontese	5
Eringer	2
Galloway	2
Unknown	2
Scottish Highland	2
Pezzata Rossa Italiana	1
Romagnola	1
Salers	1
Tyrolean Grey	1
Total	**1147**

1000 bull genomes Run 4.0

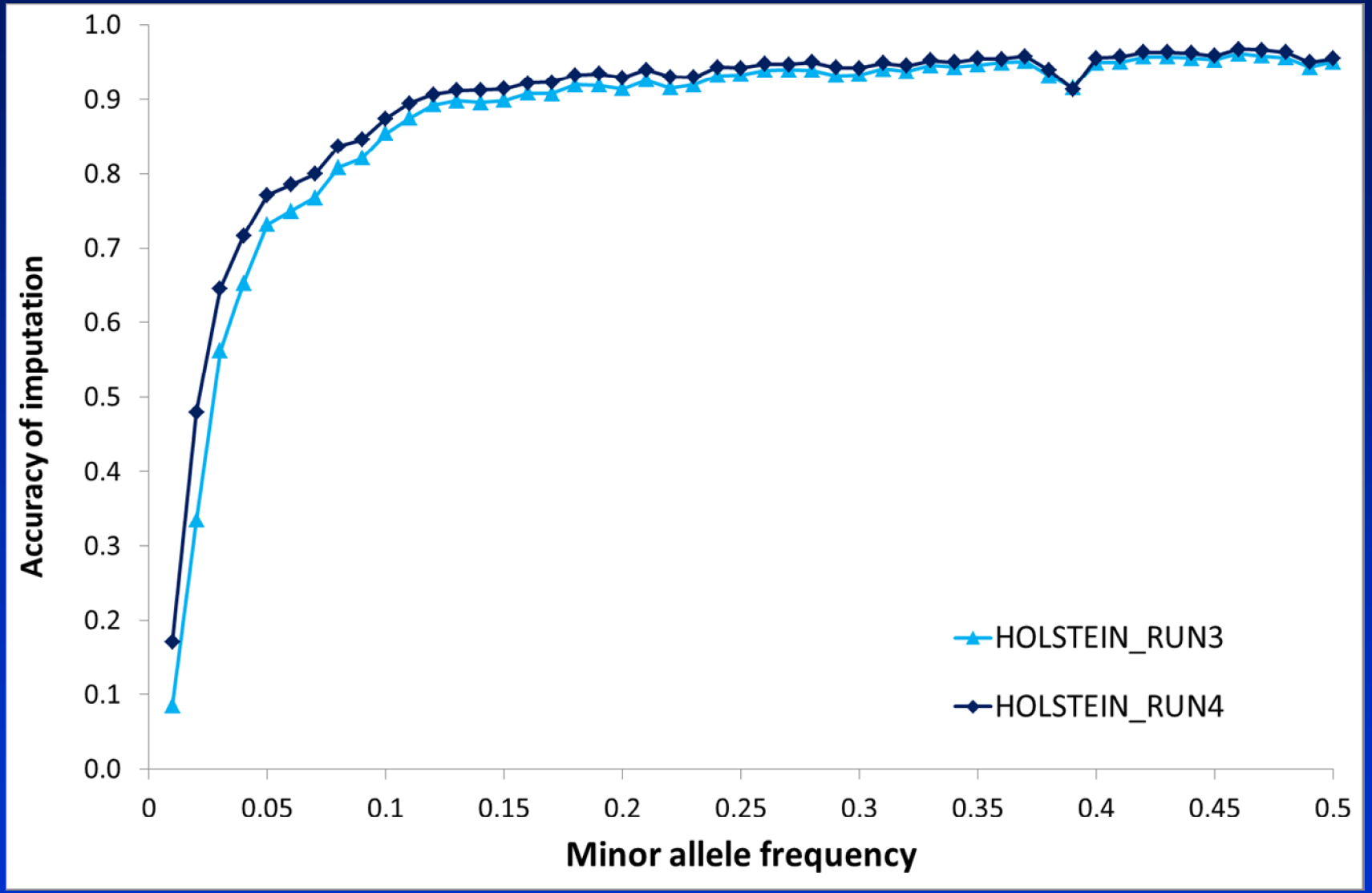
- 36.9 million filtered variants
- 35.2 million SNP
- 1.7 million INDEL



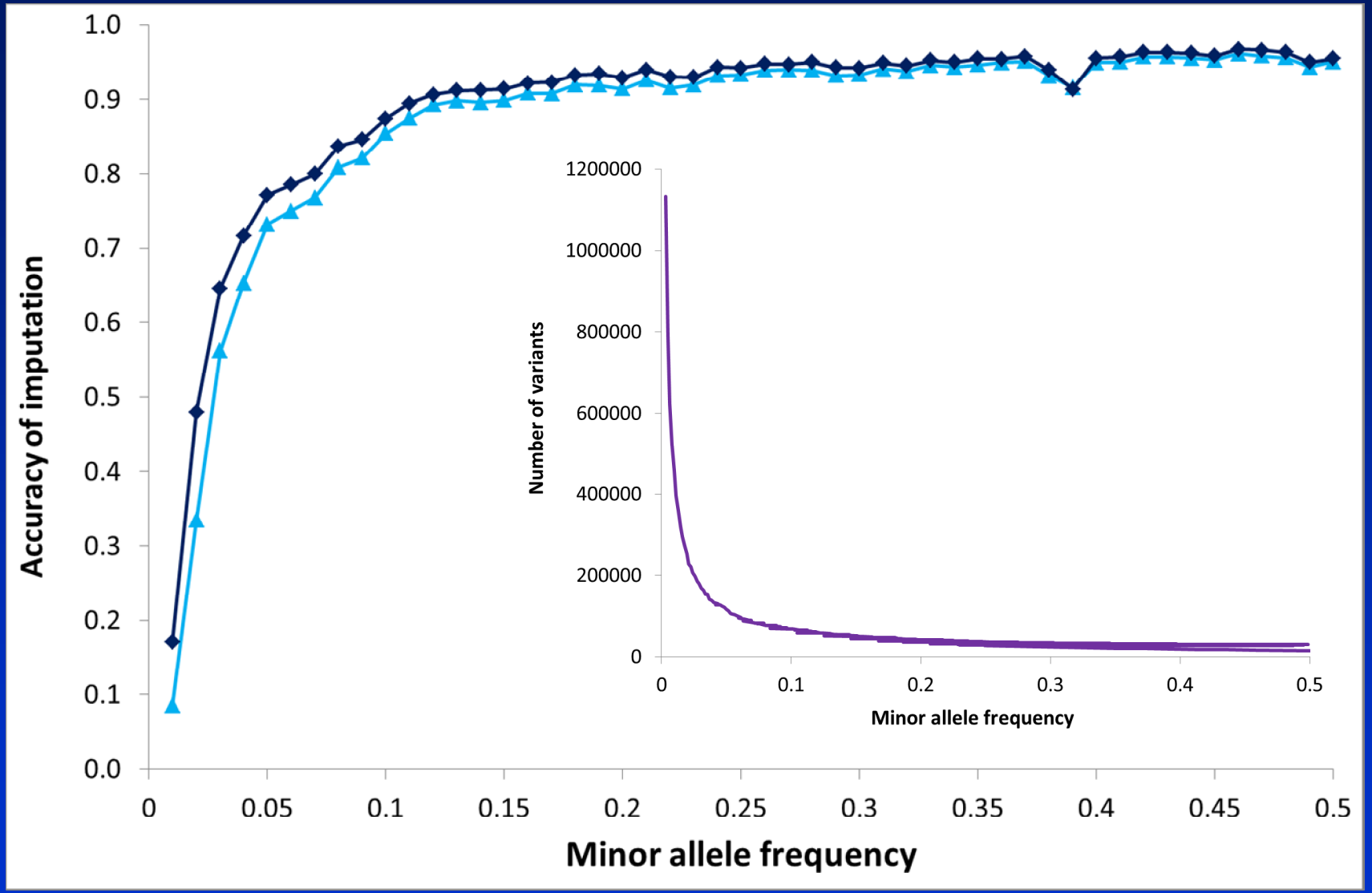
Imputation of full sequence data

- Accuracy?
  - Chromosome 14
  - Remove 50 Holsteins, 20 Jerseys from data set
  - Reduce genotypes to 800K for these animals
  - Impute full sequence using rest of animals as reference

Imputation of full sequence data



Imputation of full sequence data

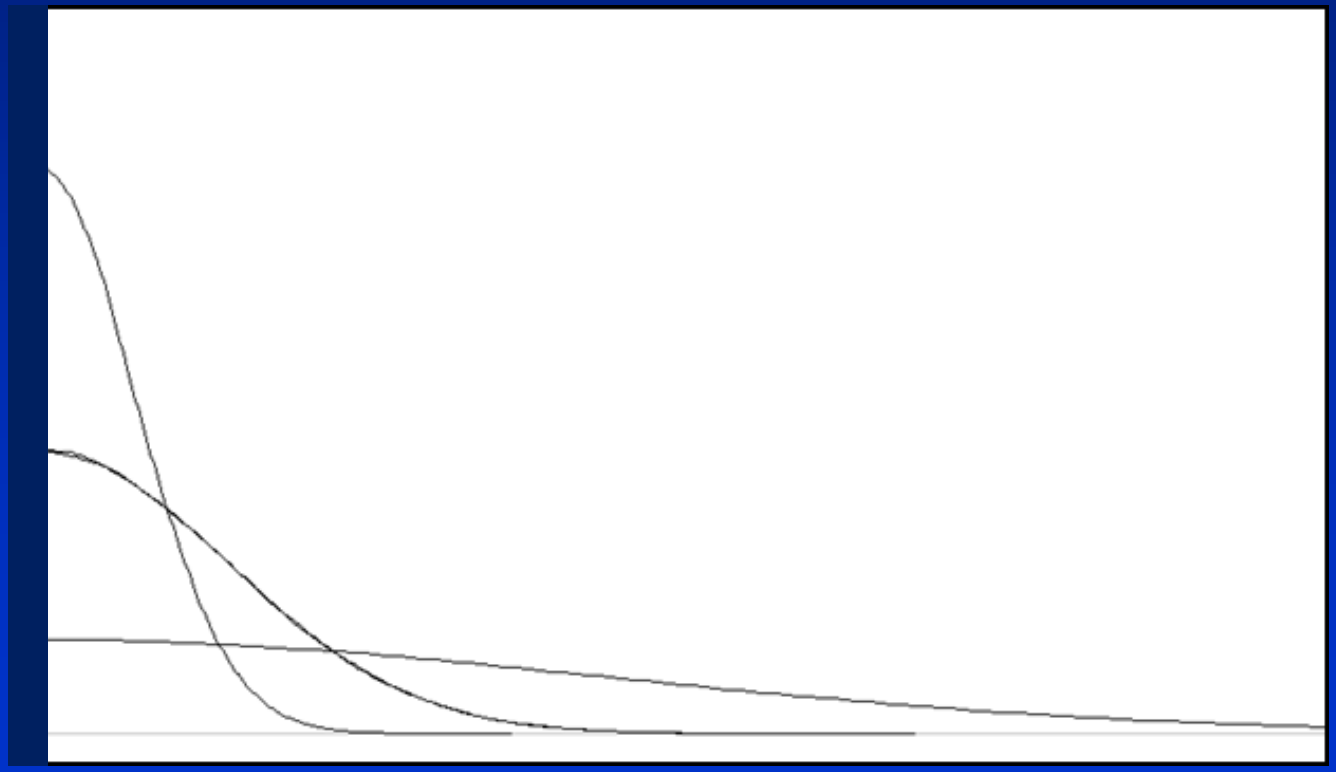


Does sequence improve accuracy?

Holstein validation

Genomic Prediction with Sequence

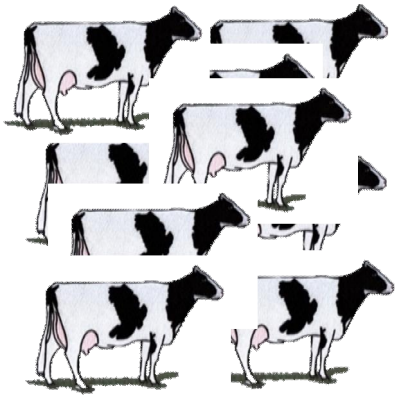
- BayesR -> variants belong to one of 4 distributions, with zero, very small, small, medium variance
- Posterior proportion of variants in each distribution

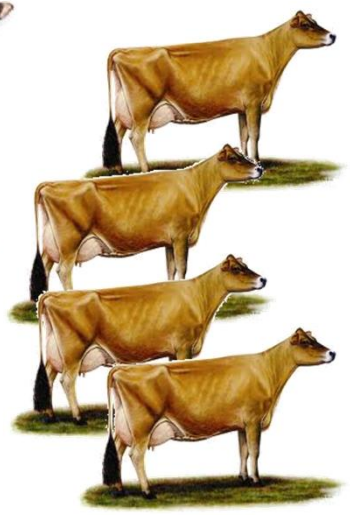


- 1 million Run3 variants in genes, +/- 2kb from genes

Does sequence improve accuracy?

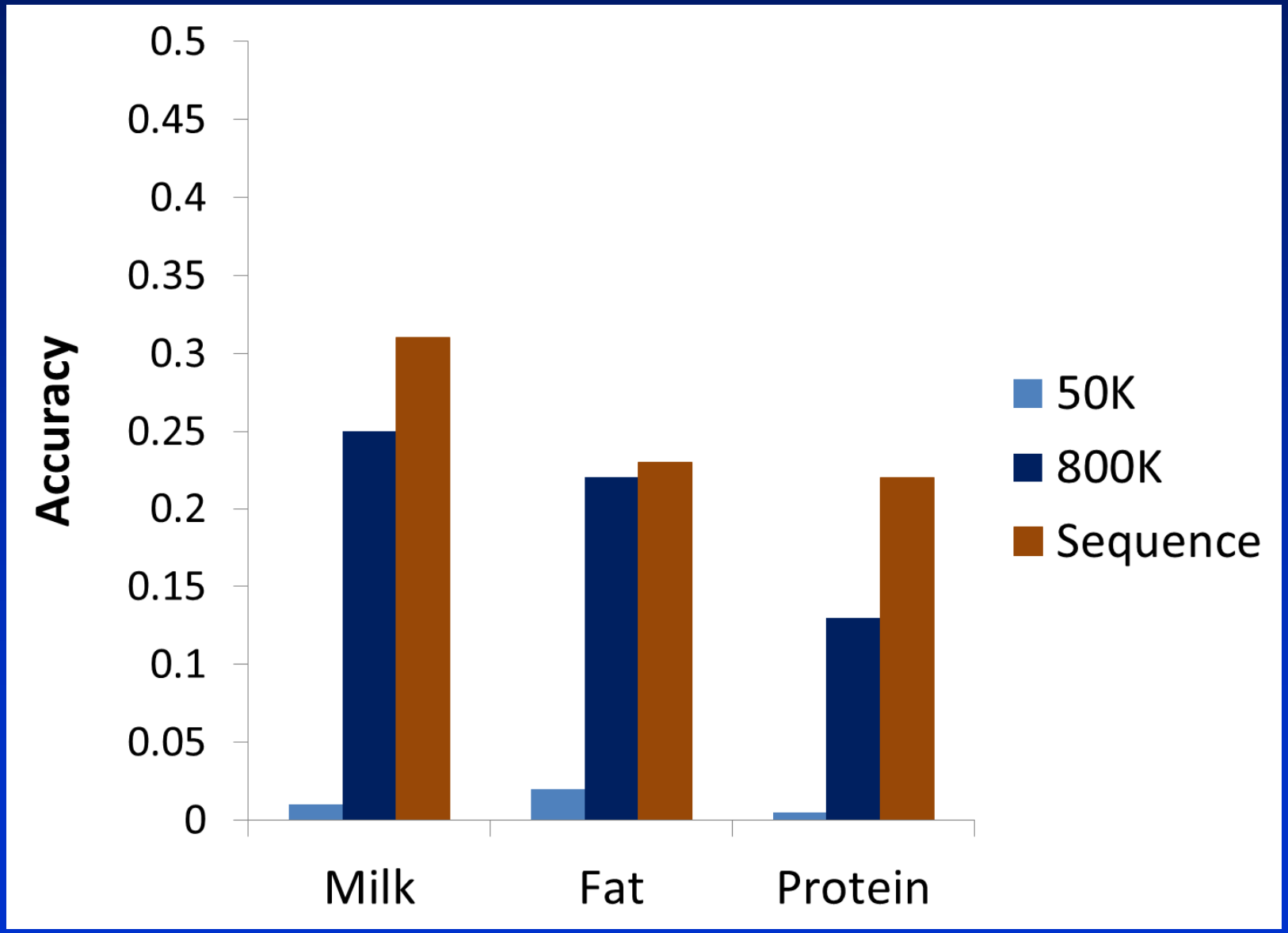
	Protein	SCC	Fertility
50K	0.53	0.48	0.39
777K	0.54	0.50	0.41
Sequence	0.56	0.52	0.44





Aussie Reds

Does sequence improve accuracy?



Genomic Prediction Software

Software	A matrix	G matrix	Weights	Genotype probabilities	Reference
GCTA	No	Yes	No	No	Yang J Am J Hum Genet. 2011 7;88:76-82.
Emmax	No	Yes	No	Yes	Kang HM Nat Genet. 2010;42:348–354
ASREML	Yes	Yes	Yes	Yes	Gilmour et al. 2002