Sequence Data

Armidale Summer Course 2015

Ben Hayes and Hans Daetwyler

Using sequence data in genomic selection and GWAS

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

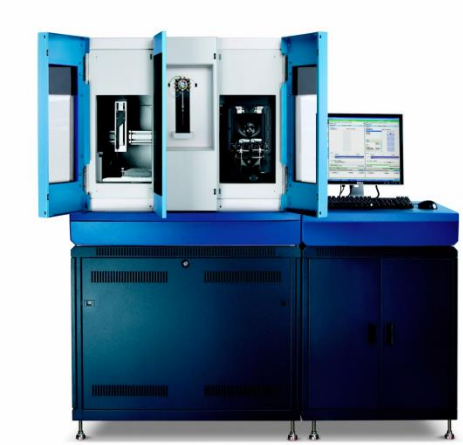
Using sequence data in genomic selection and GWAS

- 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

Technology – NextGenSequence

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

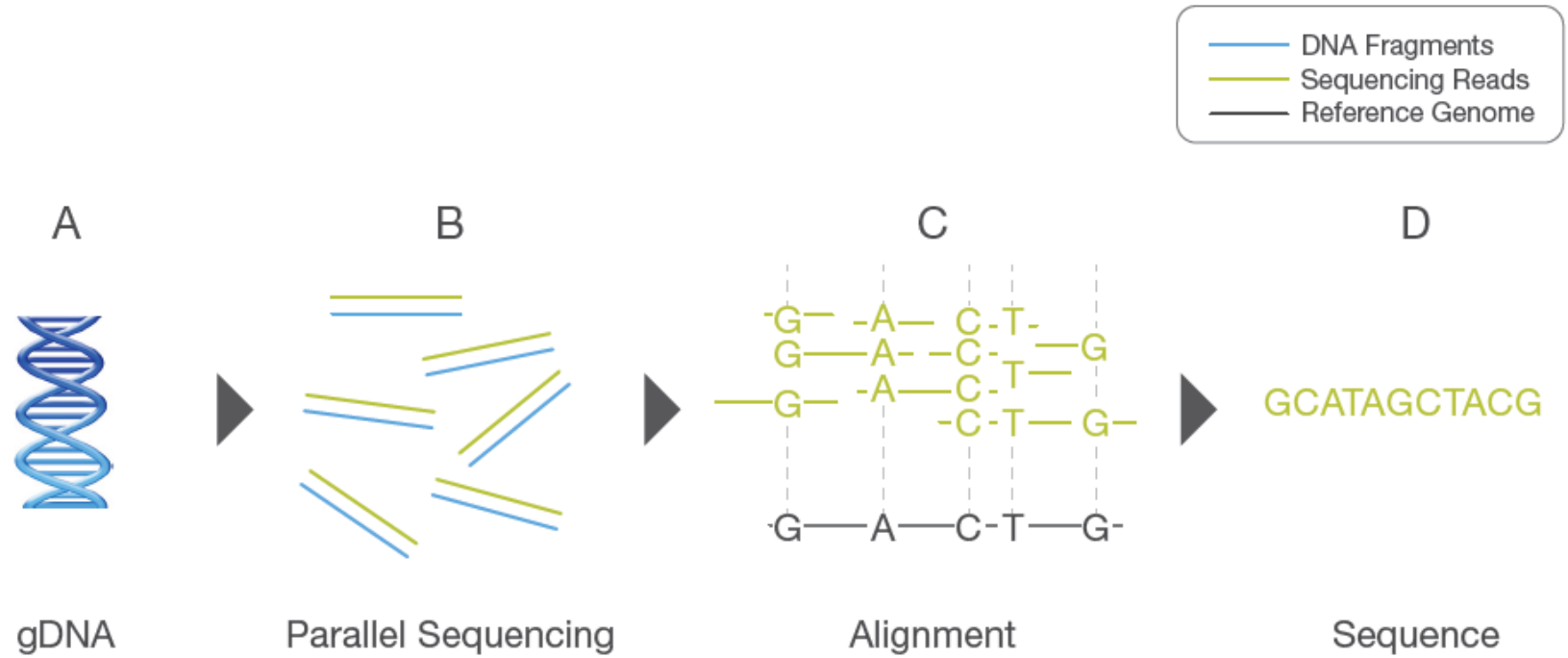
Roche: GS-FLX



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

Whole Genome Sequencing

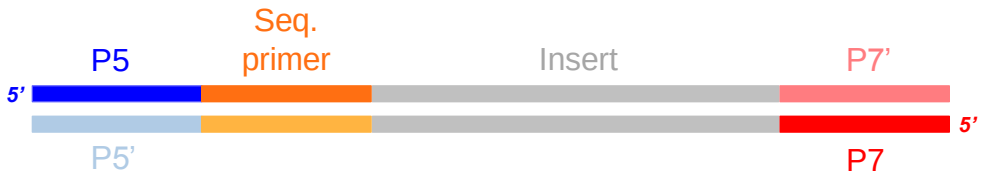
Illumina HiSeq3000

- 1 flow cell 100bp paired end run takes 3.5 days
- Generates 750 Gbases of sequence.
  - HiSeq2000 was 200 Gbases in 10 days

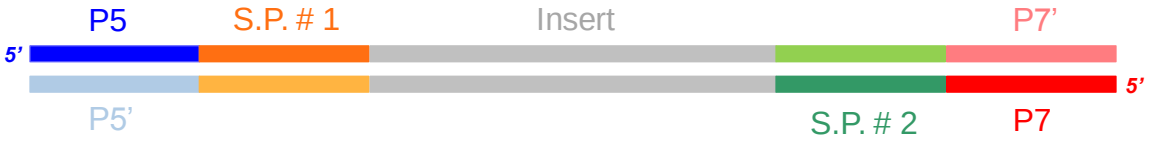
Library Preparation

- DNA sheared into smaller fragments and adapters ligated to ends.

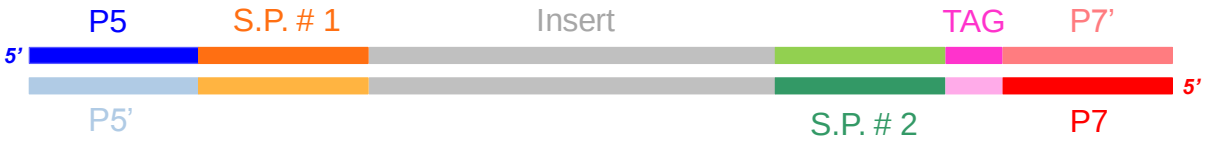
Single Read



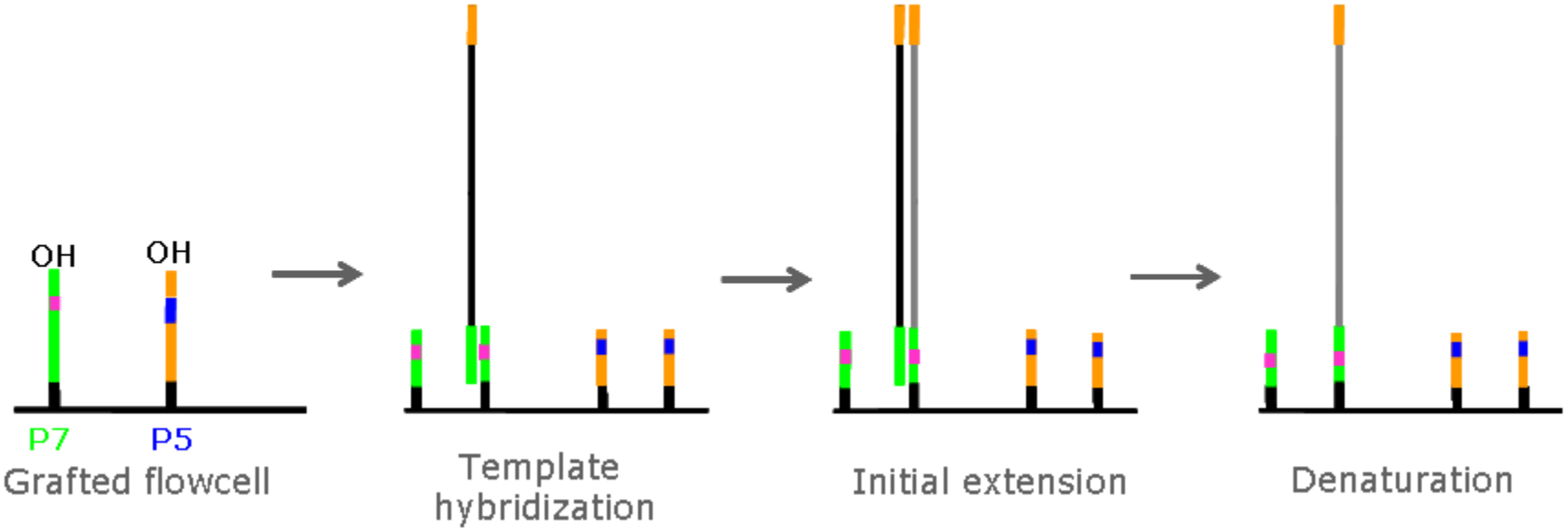
Paired Read



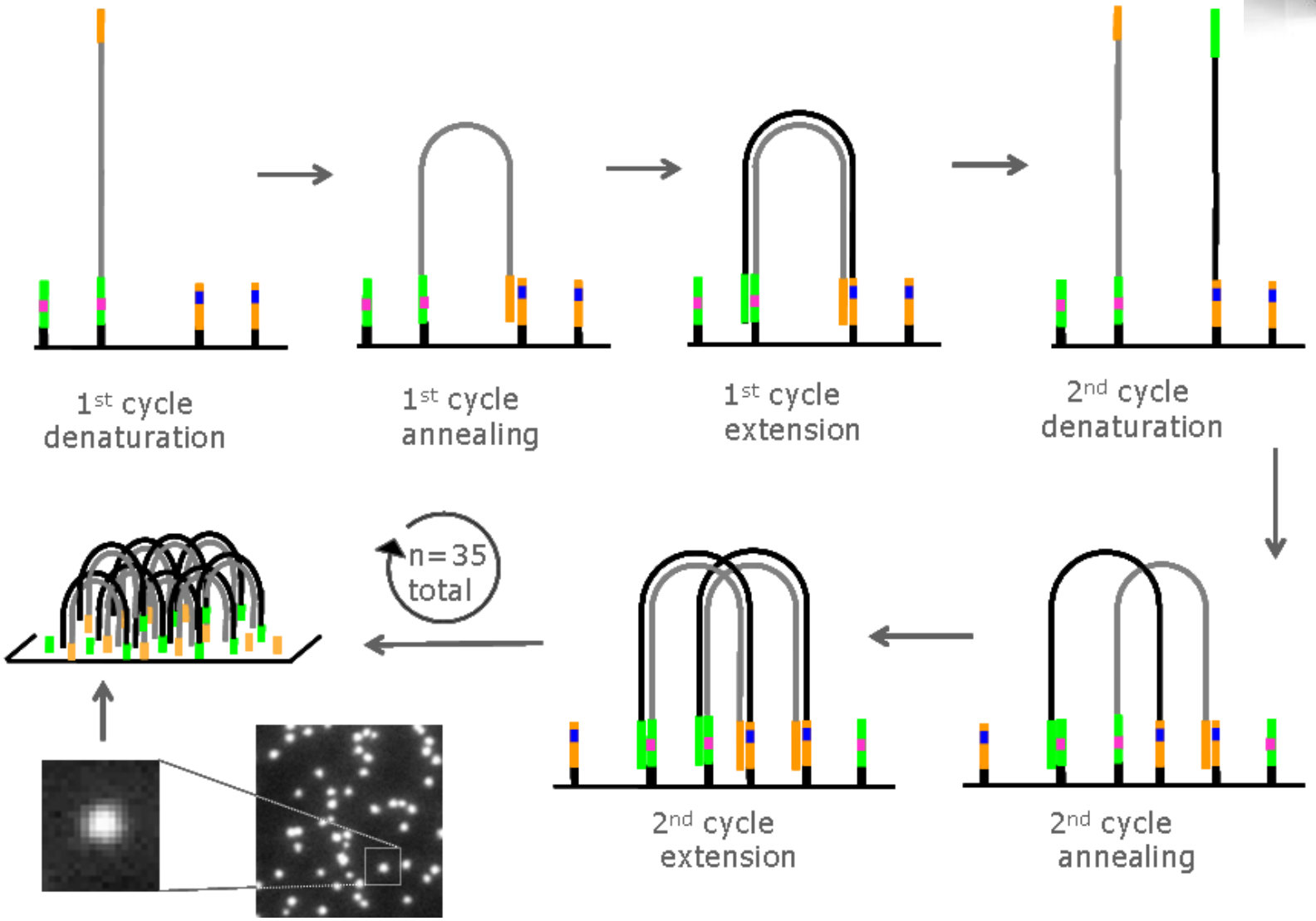
Indexed Paired Read



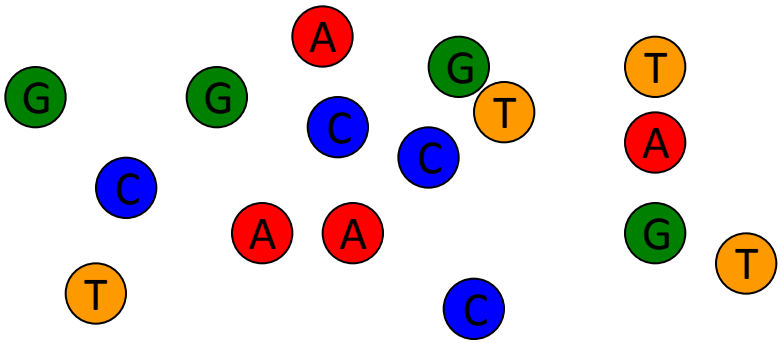
Cluster Generation

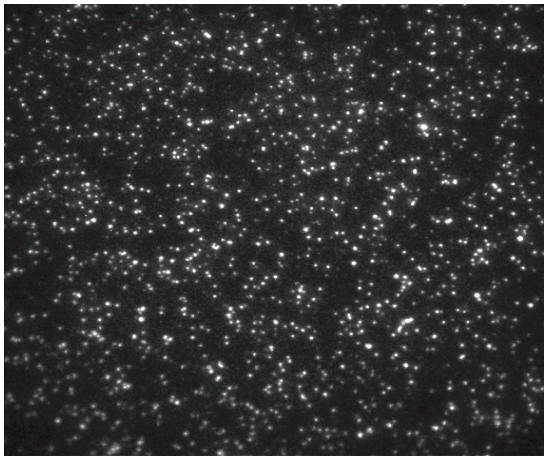


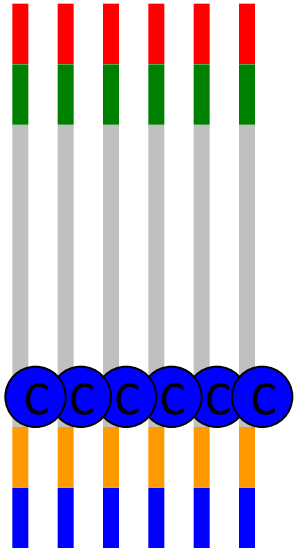
Cluster Generation

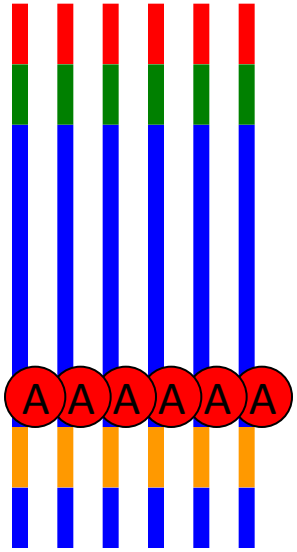


Sequencing by synthesis

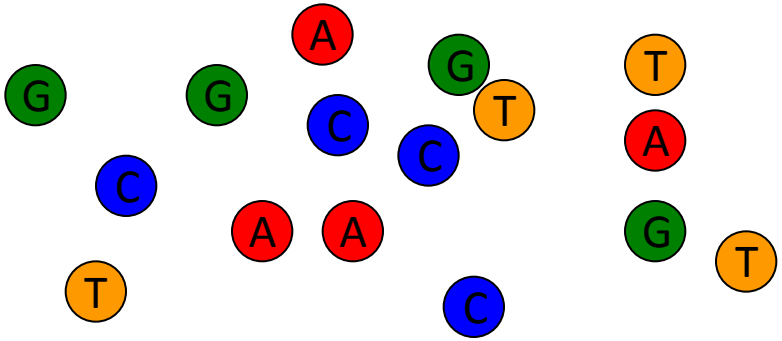


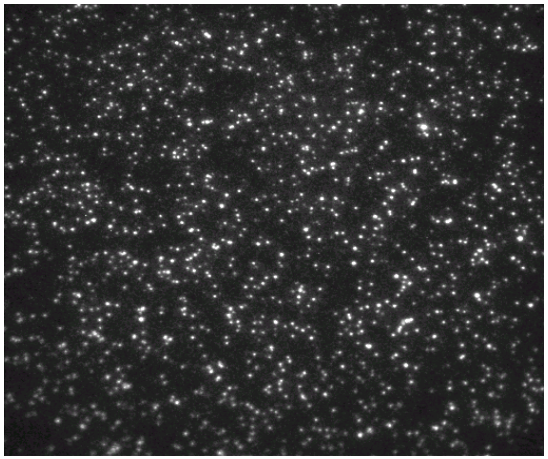


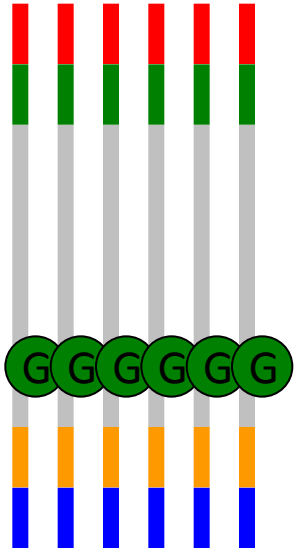




Sequencing by synthesis







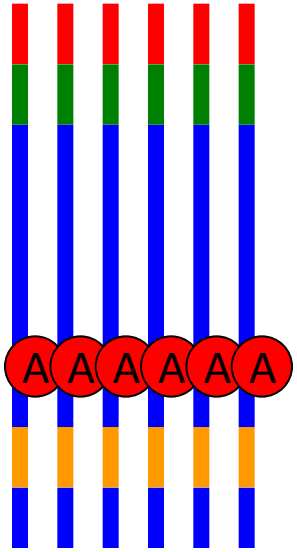
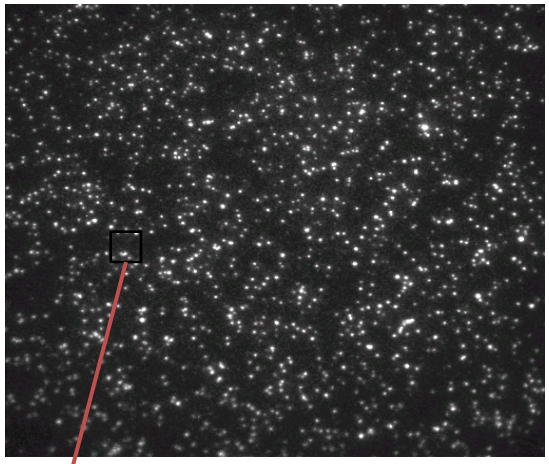
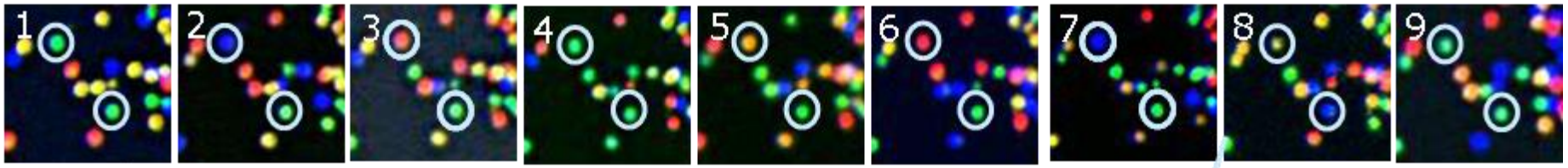


Image Analysis and Base Calling



TGCTACGAT...



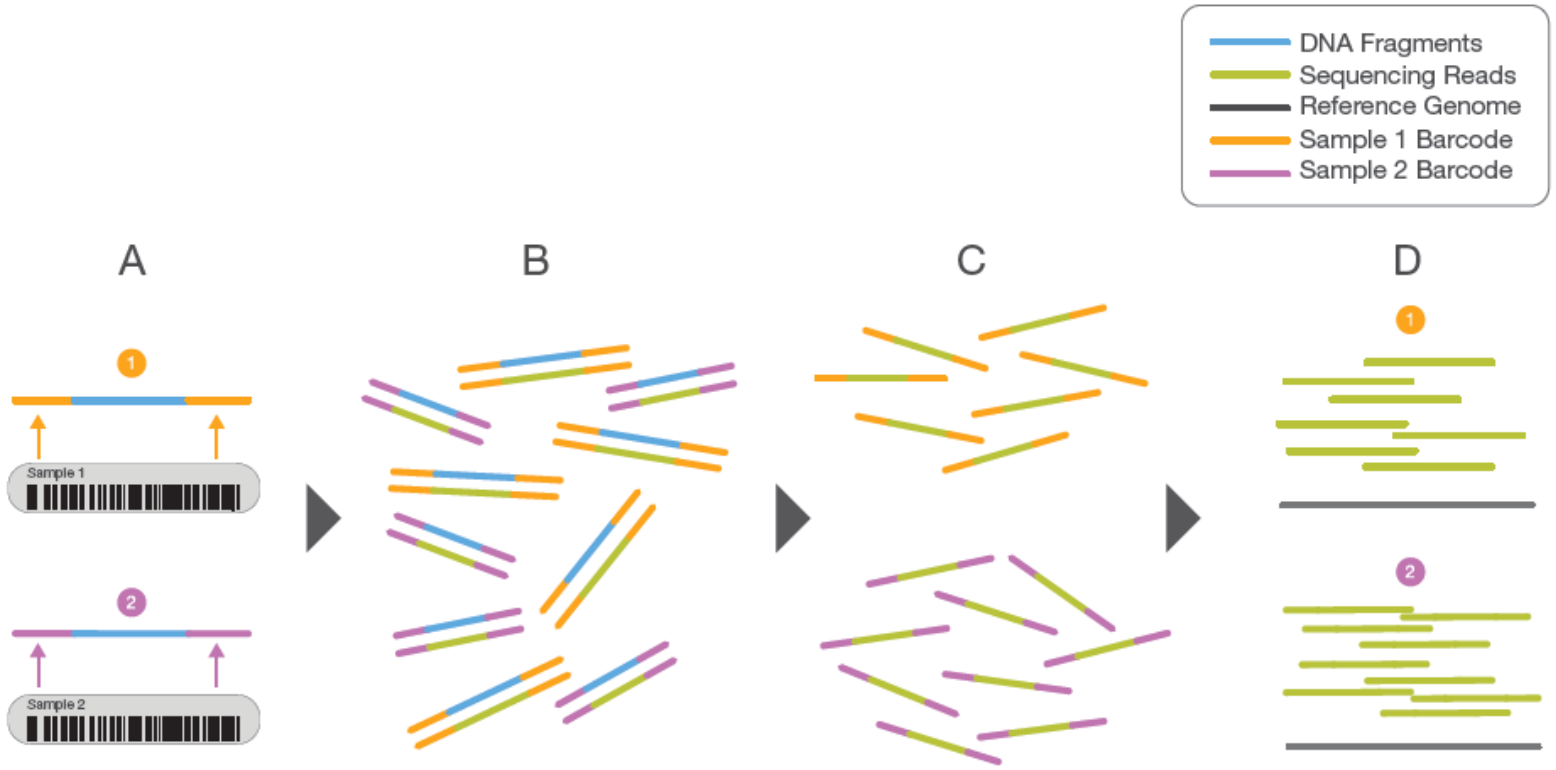
TTTTTTTGT...

Multiplexing or pooling of samples

- Known unique nucleotide tags or “barcodes” can be attached to each read
- Identifies sample
- Allows mixing of many samples on one flowcell
- Can reduce cost substantially
  - Main reason why genotyping by sequencing can be cost effective

Pooling of samples via barcoding

Figure 2: Conceptual Overview of Sample Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

‘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
+
CCCFFFFFHHHHHHJJJJJJJIJJJIHHIJJJJIJJJJJJJJJJIJJJJIGJIJJJIJJ=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;;DFDFDEGFHL[illegible]EHE<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]TAATGCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@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:5:1101:2[illegible]2093  2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAAGTTA[illegible]AATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHHJJJJJJJIJJJIHHIJJJJIJJ[illegible]JJIJJJJIGJIJJJIJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:1101:2445:[illegible]  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]GTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHGJJ[illegible]GHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:215[illegible]031  2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAA[illegible]TAGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJI[illegible]JJIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:110[illegible]415:2093  2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAA[illegible]AAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJIJJ[illegible]JJJJJIJJJJJJJJJIJJJJIGJIJJJIJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX[illegible]101:2445:2236  2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGA[illegible]TCTTTTCCAATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD[illegible]ECBAA3A<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=  AG<BBHEGH9BFF;?DF;;8CHEIGIHJHHHH=?7?D7)6;A;>?C@A>ACCDC(:ABC902>
@HISEQ:185:D2E63ACXX:5:1101:1788:209  N:0:AATCGTGGTACGGTGA
AGATGGTAAAGAATCTTCCTATAATGCAG  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  TAGCATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIH  JJIJJJIEGGII=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:110  15:2093  2:N:0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAA  AAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJIJJJ  JJJJJIJJJJJJJJJIJJJJIGJIJJJIJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:  01:2445:2236  2:N:0:AATCGTGGTACGGTGA
AATCTTTCTCAGCATGAG  CTTTTCCAATGAGTCAGCTCTTCGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD  ECBAA3A<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
AGA  AAAGAATCTTCCTATAATGCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@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  AGTTAAATAATATGGATGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHJJJJJJJIJJJIH  IIJJJJJJJJJJIJJJJIGJIJJJIJJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:110  5:2236  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

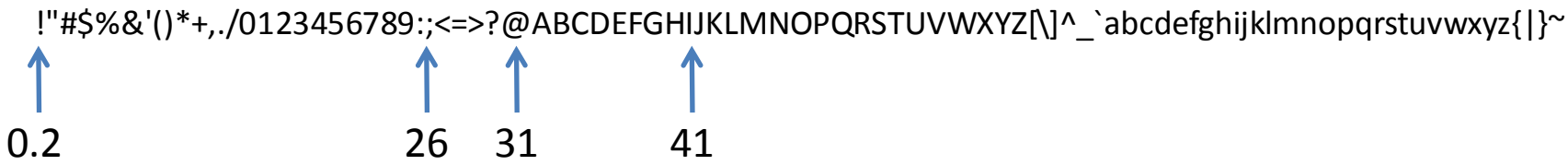
Raw data - FASTQ Files (Casava1.8)

```
@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
AGATGGTAAAGAATCTTCCTAT[illegible]GCAGGAGTCCCGGGTTCAATCCCTGGGTCAGGAAGATCCCCTGGATCTTCCACTCCAGTATTCTTGCCTGGAGAAT
+
?@@DDF=AFFDFHEHIJJJIJIJGBHHG[illegible]CGHGIHFFFHIGGIIIIJ?FGHJHEIID@>@EC>AEFFDDCDC>CE@>@CACACDD@C>?CDD?BD@C
@HISEQ:185:D2E63ACXX:5:1101:215[illegible]31  2:N:0:AATCGTGGTACGGTGA
CAAAAGCACTCTGGAGGGAAACAACAG[illegible]CATAACTGAGGCAGAAGATAGGATAAGTGAGGTGGAAGATAGGATGGTGGAAGTAAATGAAGCAGAGAGGA
+
CCCFFFFAHFHHHJGHJJAHIIJCIJIHHIIJJJIJJJIE[illegible]=FGFHGFFGHCHIFHIGIGEGAEHHEEFBDFEDACCBCCCCDDADCCC@?C??<???
@HISEQ:185:D2E63ACXX:5:1101:2415:2093[illegible]0:AATCGTGGTACGGTGA
AGCAAAACTGACTTGAAGAAAGTTAAATAATAT[illegible]TGCCCACATATATTTAGAGAGGCGGTTTGATCTCATTTTTATTTTGACTTTTATGAAACTTTCAG
+
CCCFFFFFHHHHHHJJJJJJJIJJJIHHIJJJJIJJJJJJJJJJIJJJJIGJ[illegible]IJ=EF?@DCDC@CEEED=ADEE@?CDDDDCDACEDCCDDCCA
@HISEQ:185:D2E63ACXX:5:1101:2445:2236  2:N:0:A[illegible]GTGGTACGGTGA
AATCTTTCTCAGCATGAGGGTCTTTTCCAATGAGTCAGC[illegible]CGCATCAGGTGGCCAAAGTATTAGAGTTTCAGGTTCATTATCAGTCCTTCCAATGAAC
+
?@@DDDDB?DADFD<F<ECBAA3A<AHFAF@H<@91?CFH[illegible]AGBGHID>0?B==E;F9F@7=FC).=CG))7)==?;CD@7@D###############
```

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

Phred quality scores for all bases in read

FASTQ quality symbols and Phred scores



- Raw Illumina reads generally do not go above 41
- 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

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

Possible problems

- Flow cell overloading
  - Too many different library fragments loaded onto flow cell
  - Clusters are too close together
  - Difficult to say which base it is from colours

QC and Visualisation of Raw Sequence Fastq files

- 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,…

- 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
- We now have multiple bam files per individual

Processing bam files

- Merging multiple bam files
  - Not strictly necessary
  - But, hard to keep track of all bams of an individual

- After merging
  - Resort and reindex merged bam

Processing bam files

- Remove PCR duplicates
  - Duplicate reads have all exact same start base pair (and finish before trimming)
  - Keep 1 unique read

  - Why?
    - Library problems: some segments overamplified
      - Did to many PCR cycles
        - Library segment drop out
        - Just did too many…

Processing bam files

- Local realignment around known indels
  - Refines mapping around indels
  - Why?
    - More accurate variant calling

- Sorting bam file
  - Needed for variant calling of multiple individuals

- Indexing bam and reference file
  - Samtools needs this to find reads

Variant Calling

- Using all sequenced individuals we want to:
  - Identify ‘all’ variants; SNP and Indels
  - Genotype all individuals for those variants
- Various programs available:
  - Two popular ones are Samtools and Genome Analyser Tool Kit (GATK)
  - We use Samtools, but are testing GATK

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).

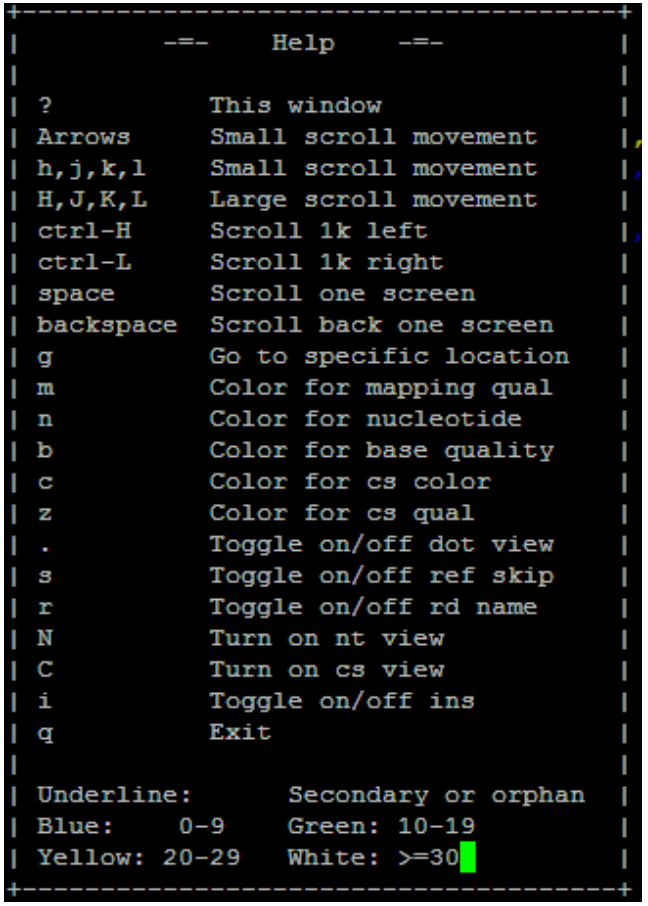
- Able to:
  - Extract reads from specific genomic region
  - Sort, index, merge bams
  - Visualise bams (command line)
  - Call variants
  - etc

Samtools tview

- We can visualise alignements by:
- $samtools tview SomeFile.bam Reference.fa

- Go to a specific place
- $g
- “Chr18”

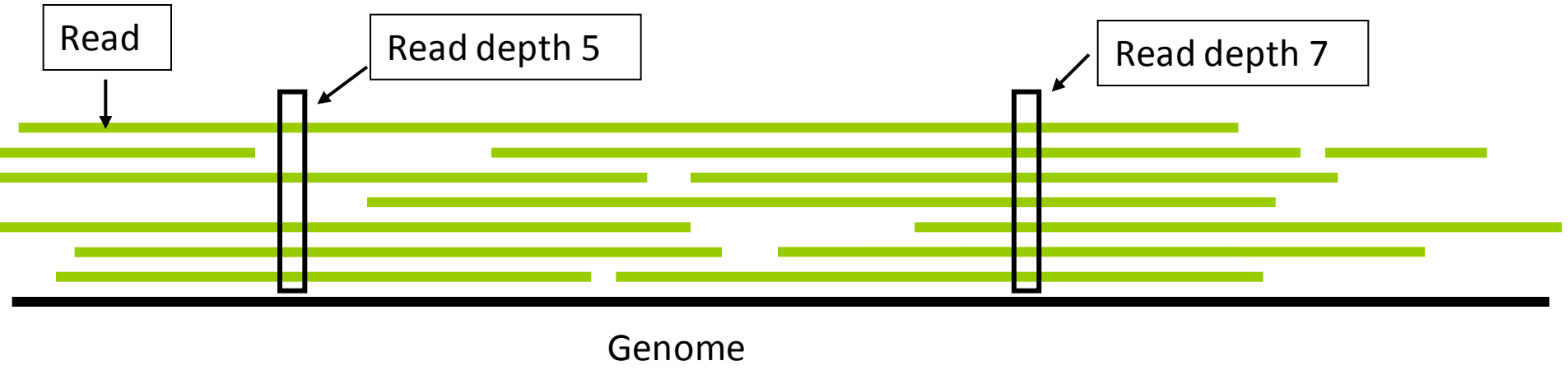
- Get help
- $Shift+/



Read depth

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

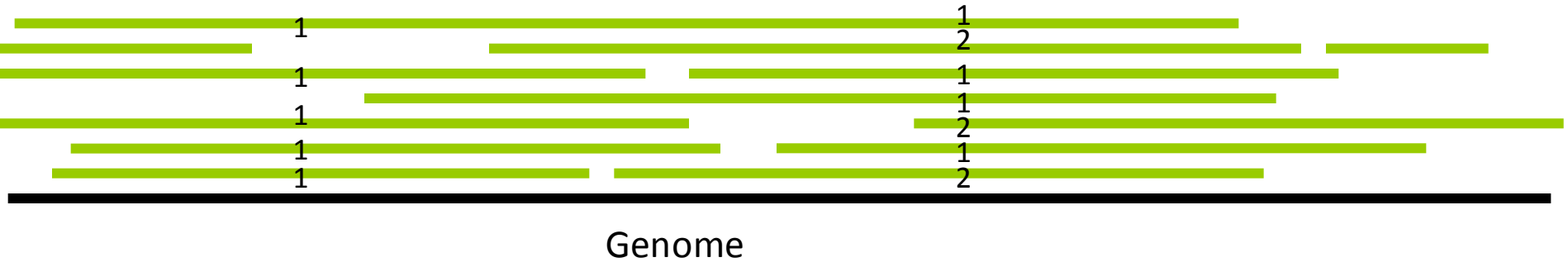
- Read depth or fold coverage



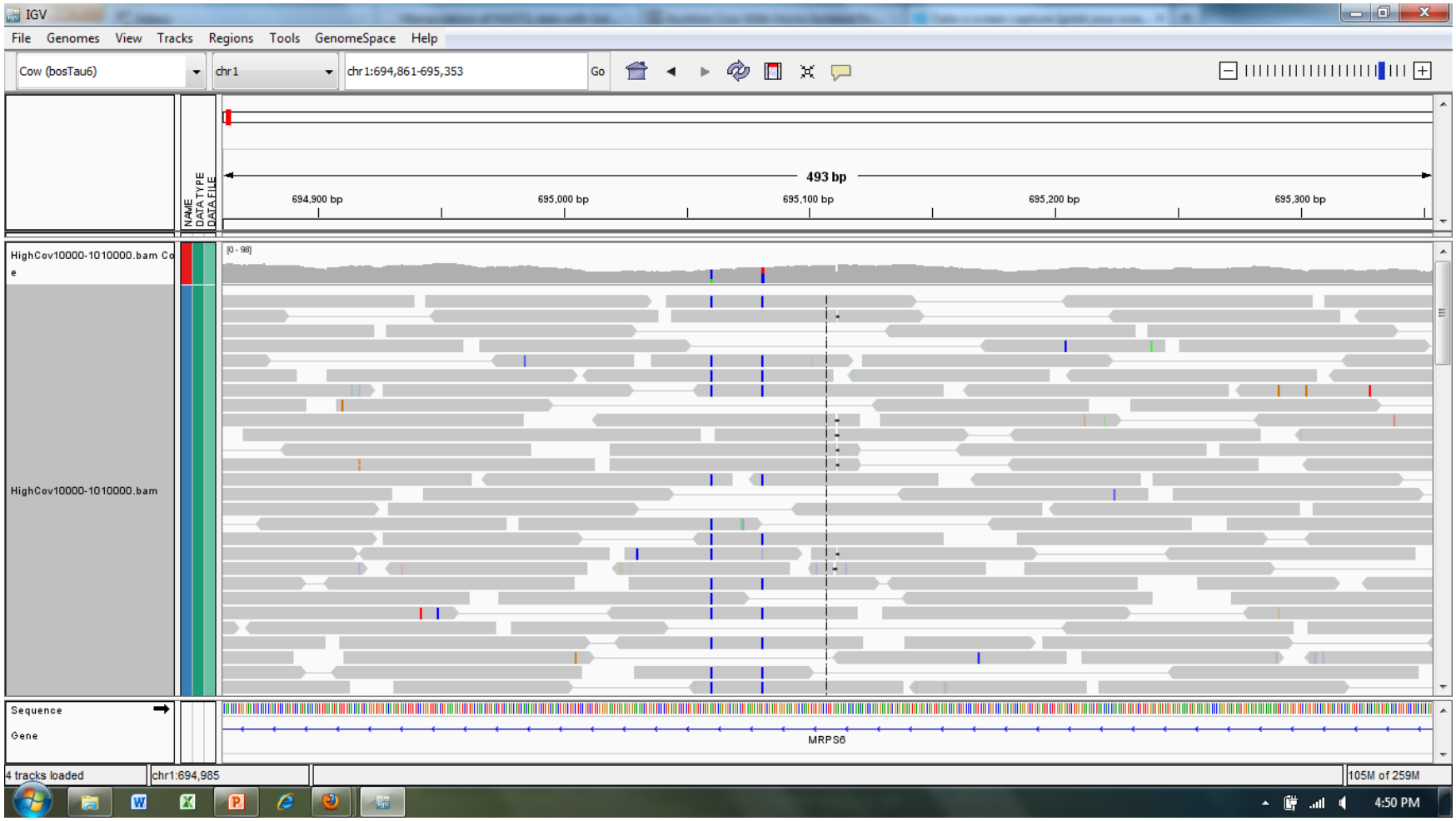
Importance of read depth

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

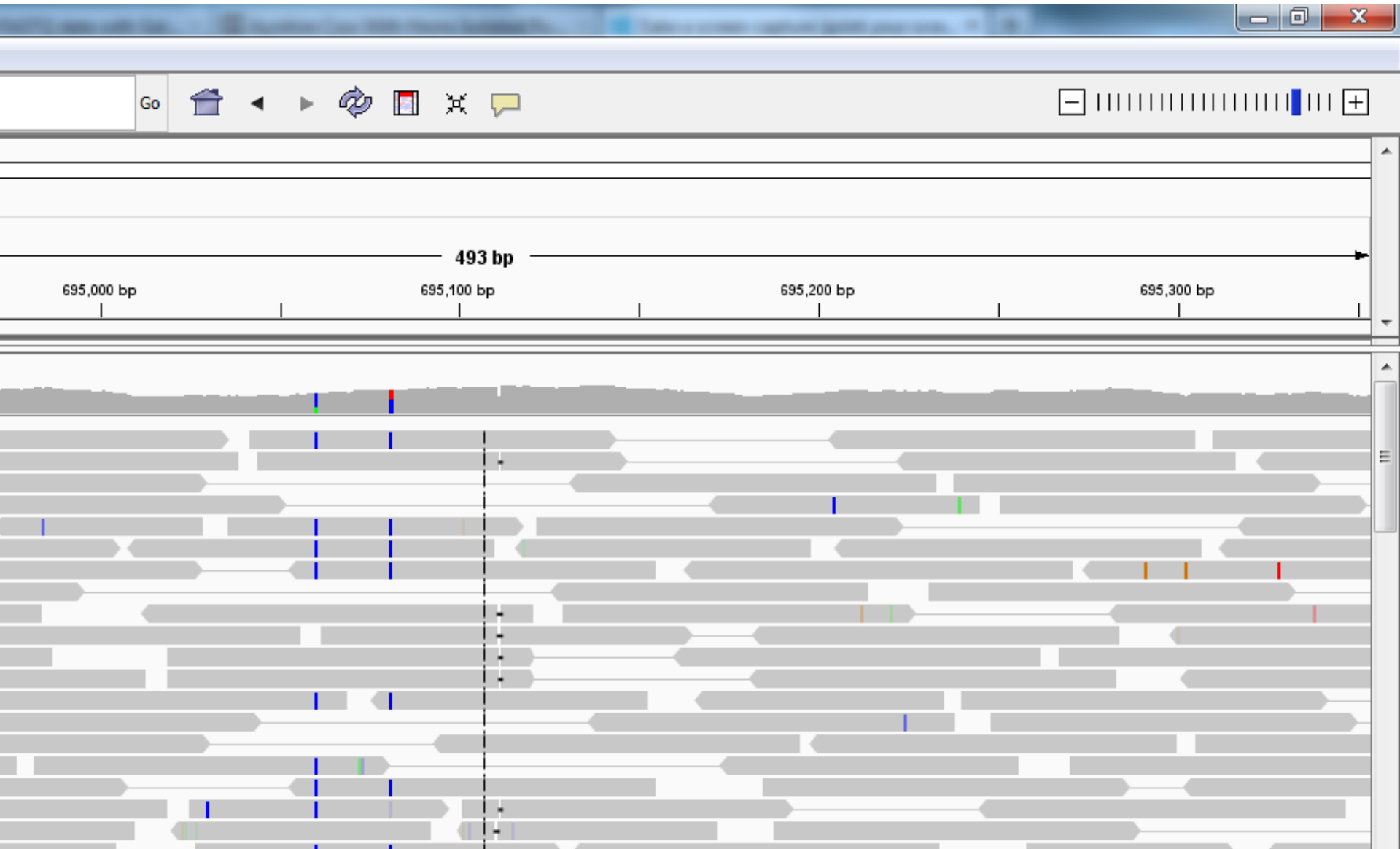
- If read depth is low, it is possible to not observe an allele and therefore call a heterozygous locus homozygous → errors
  - Read depth 5 → $0.5^5 = 0.03125$



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

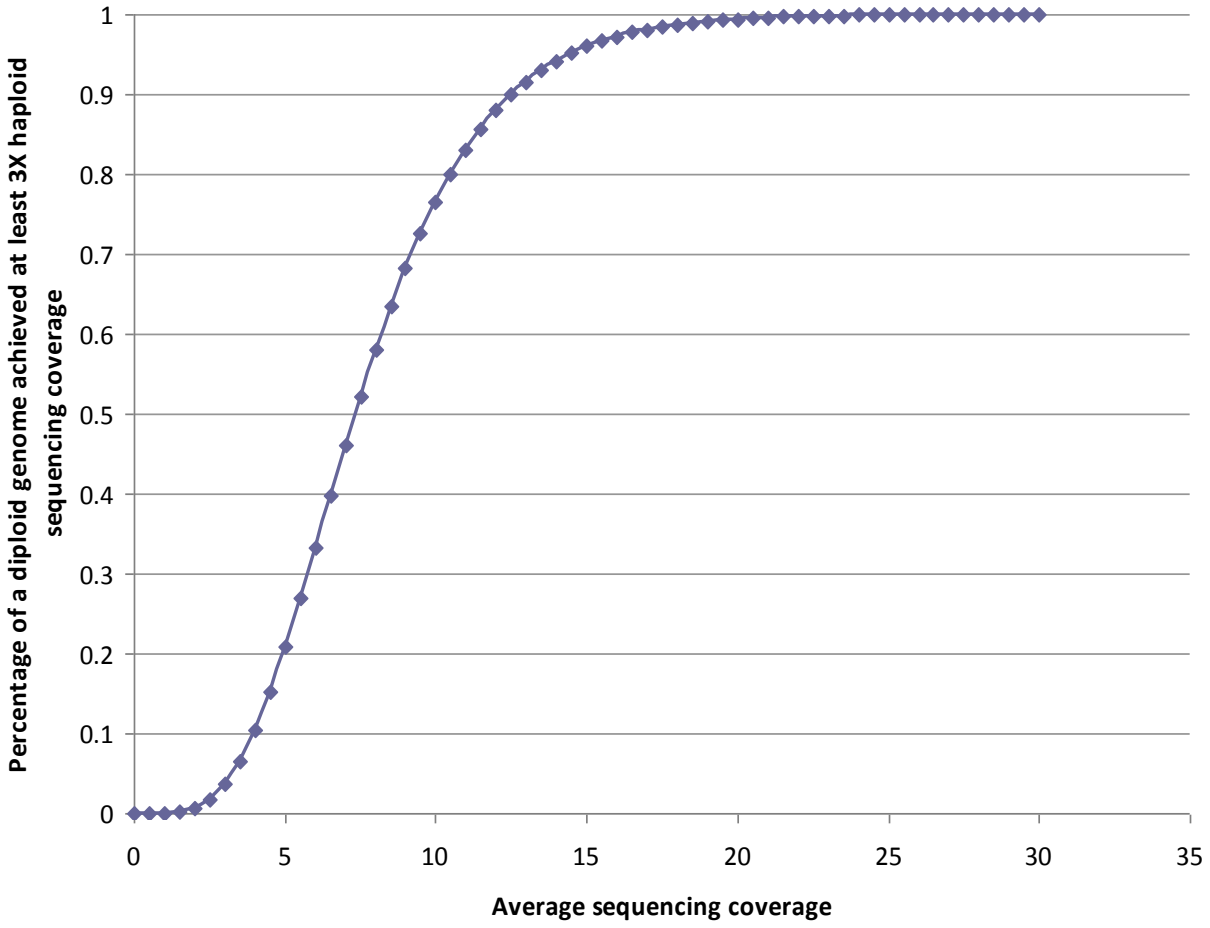


Viewing aligned reads with Integrated Genome Viewer (IGV, Broad Institute)



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)



QC and Visualisation of a BAM files

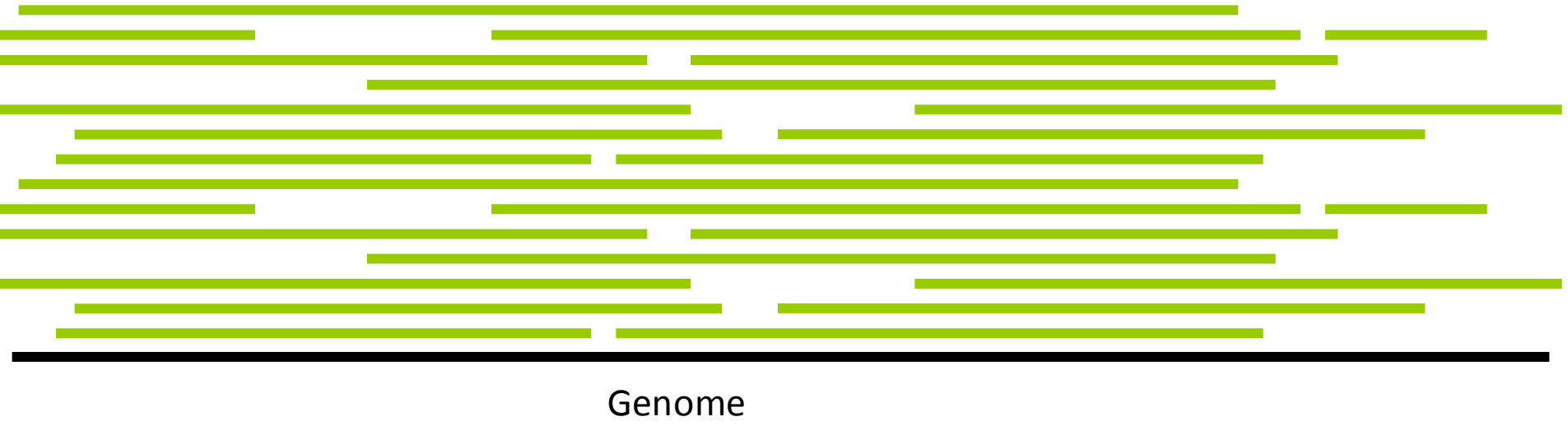
- Practical 1b on galaxy

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

Identification of variants

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



Variants in sequence

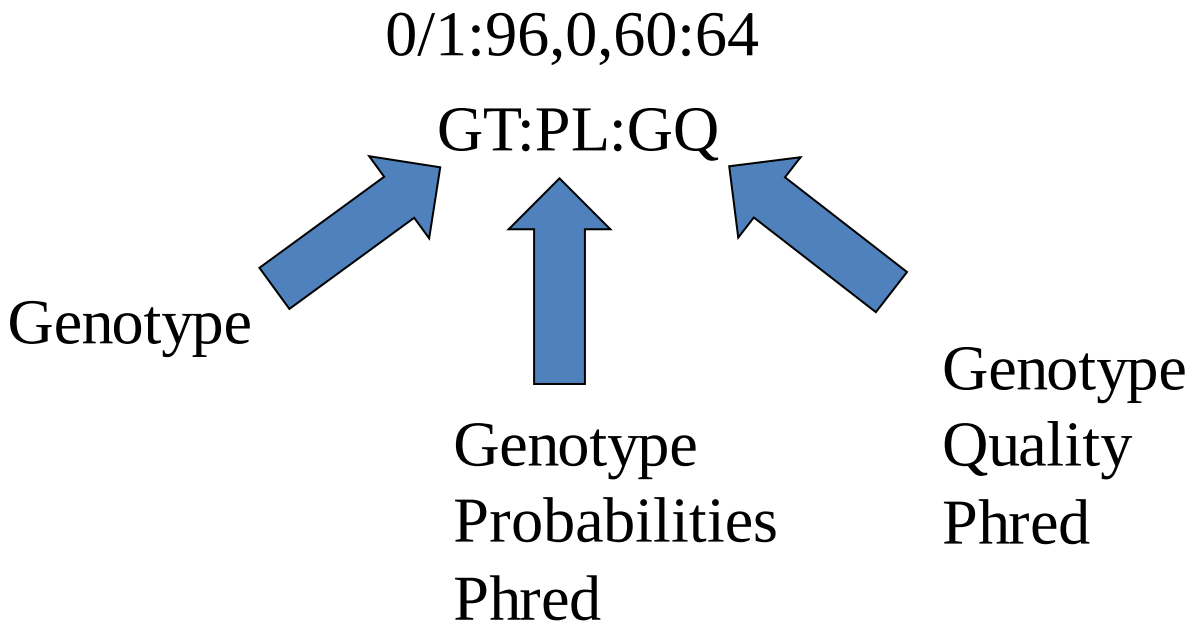
- SNP
- INDEL
  - INsertions and DELetions of DNA sections
- Copy number variants (CNV)
  - Repeated sections of DNA of various lengths
- Most studies to date have concentrated on SNP

Variant Call Format (VCF) file

- ##fileformat=VCFv4.1
- ##samtoolsVersion=0.1.18 (r982:295)
- ##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
- ##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
- ##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
- ##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
- ##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
- ##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
- ##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
- ##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
- ##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
- ##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
- ##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
- ##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
- ##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
- ##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than in group2.">
- ##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
- ##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
- ##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
- ##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
- ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
- ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
- ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
- ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
- ##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
- ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
- #CHROM POS ID REF ALT QUAL FILTER INFO
  FORMAT Individual1 Individual2
- 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
- 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
- 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
- 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
- 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
  - Quality scores low

Reasons against filters:

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

Variant filters

- Number of reads per allele
  - Observe alleles on forward /reverse strand reads
  - Sequencing occurs in both directions on DNA fragment
  - Should observe allele in both directions
  - Why?
    - Not observing reads could point to systematic errors

Variant filters 2

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 3

- Mapping quality
  - Reads not well mapped
  - Why?
    - Repetitive regions of genome
    - Assembly errors
    - Errors in read

- Quality (overall)
  - Overall score taking into account some of measures above
  - Catch all, but strong quality filters will remove low MAF SNP

Variant filters 4

- Multiple variants within 3bp window
  - Remove:
    - SNP close to SNP
    - SNP close to indels
    - Indels close to Indels
  - Why?
    - Alignment errors and indels can cause shifts → call 2 SNP close together instead of 1
    - Local realignment around indels may help

Variant filters 5

- 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?

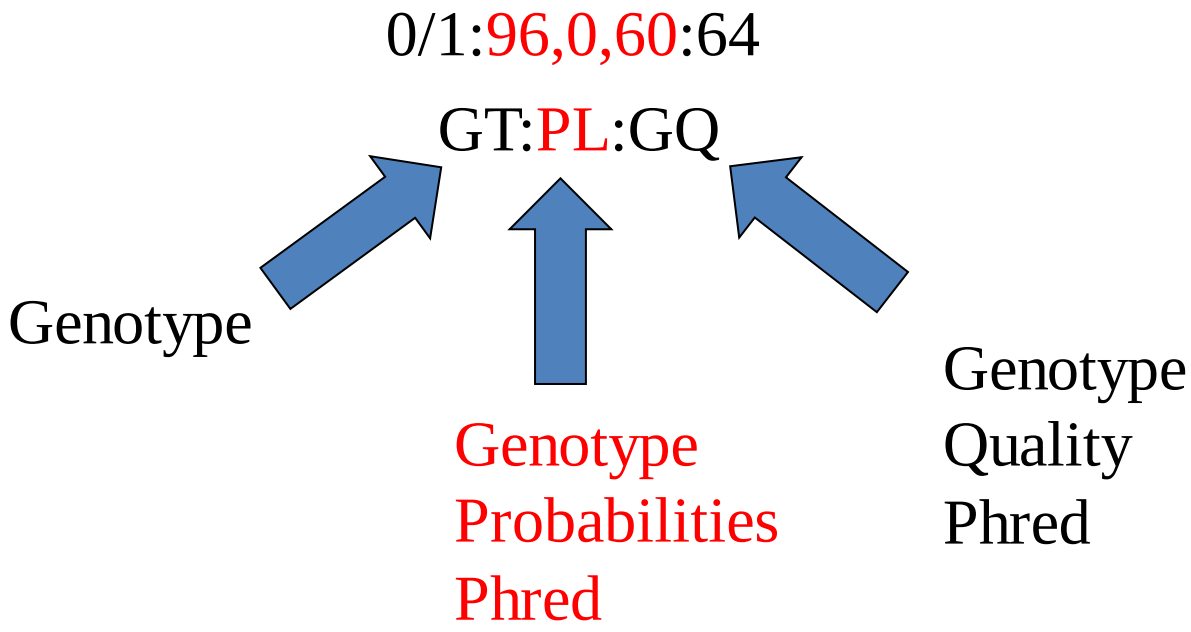
- Require minor allele is in at least 2 animal in sample
  - BUT will lead to a threshold on minor allele frequency
  - E.g. 50 seq. animals → 2/50=0.04
    - Thus, MAF cut off is 0.04
- Heterozygosity (Hardy-Weinberg)
- etc

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	1	1	no
0	5	110	1	0	yes
0	50	60	0	0	no
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

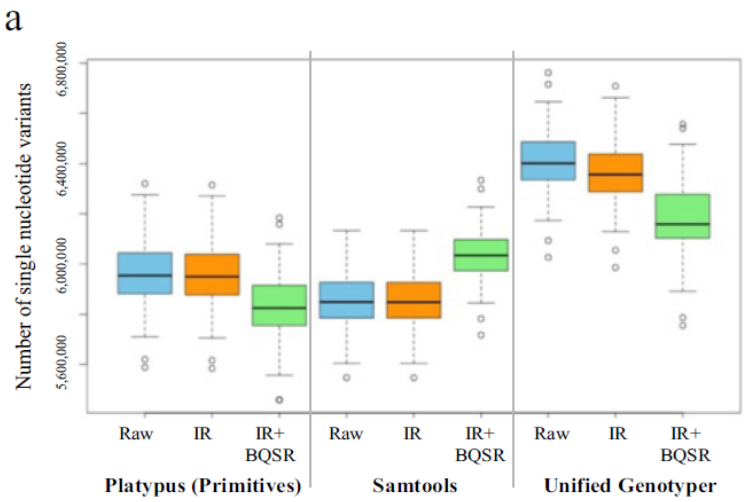
Prob(0)*0 + Prob(1)*1 + Prob(2)*2 = 1*0 + 0.0036*1 + 0*2 = 0.0036

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

QC and Filtering VCF Files

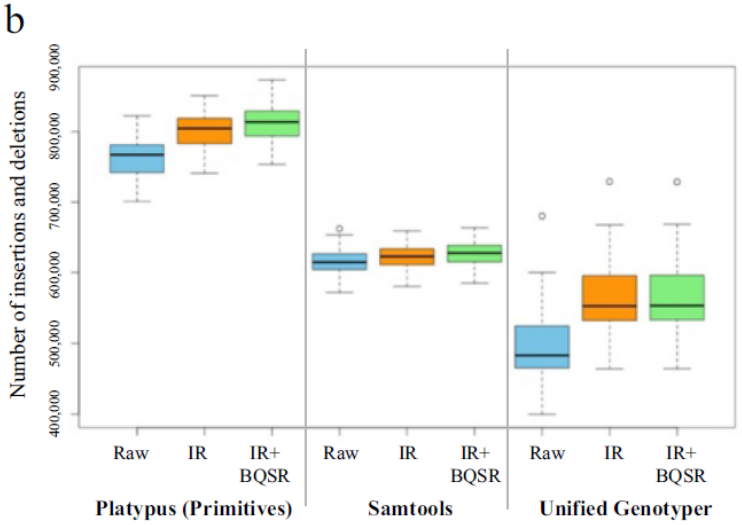
- Practical 1c

Comparisons of variant callers



SNP

Raw = no indel realignment or base quality score recalibration
IR = InDel realignment
IR + BQSR = InDel realignment followed by base quality score recalibration



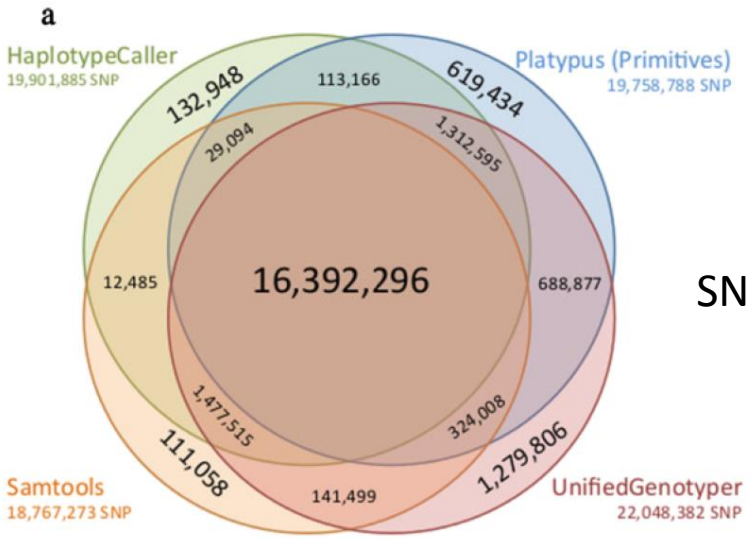
indels

Raw = no indel realignment or base quality score recalibration
IR = InDel realignment
IR + BQSR = InDel realignment followed by base quality score recalibration

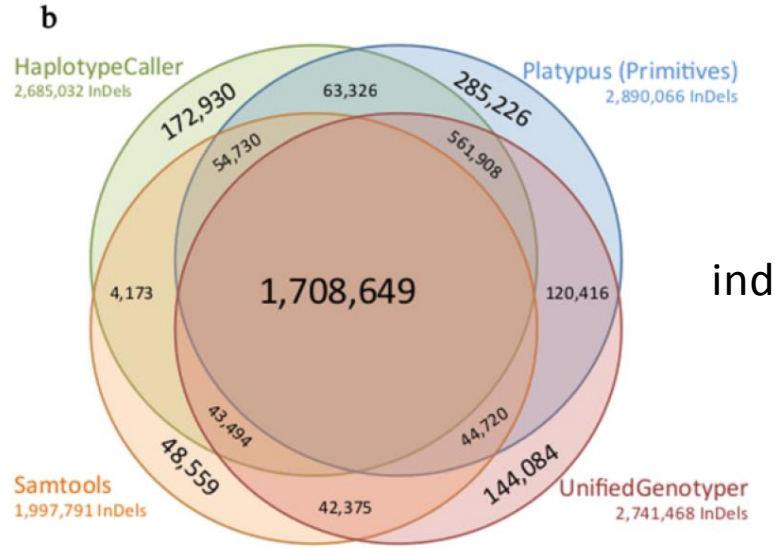
- GATK UG finds more SNP but fewer indels
- Benefit from indel realignment variable

Baes et al., 2014, BMC Genomics

Variants called



SNP

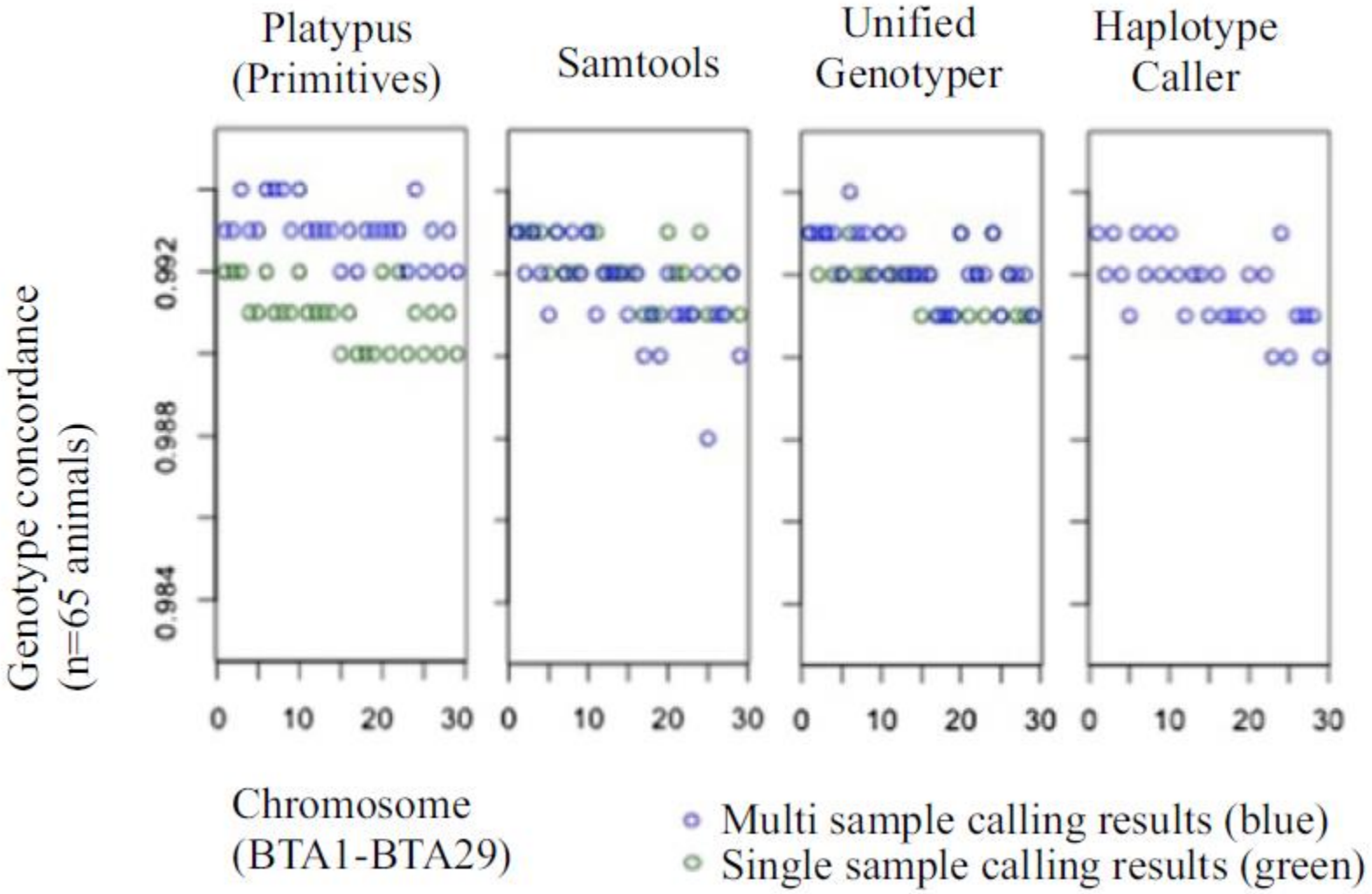


indels

- Most variants identified with all callers

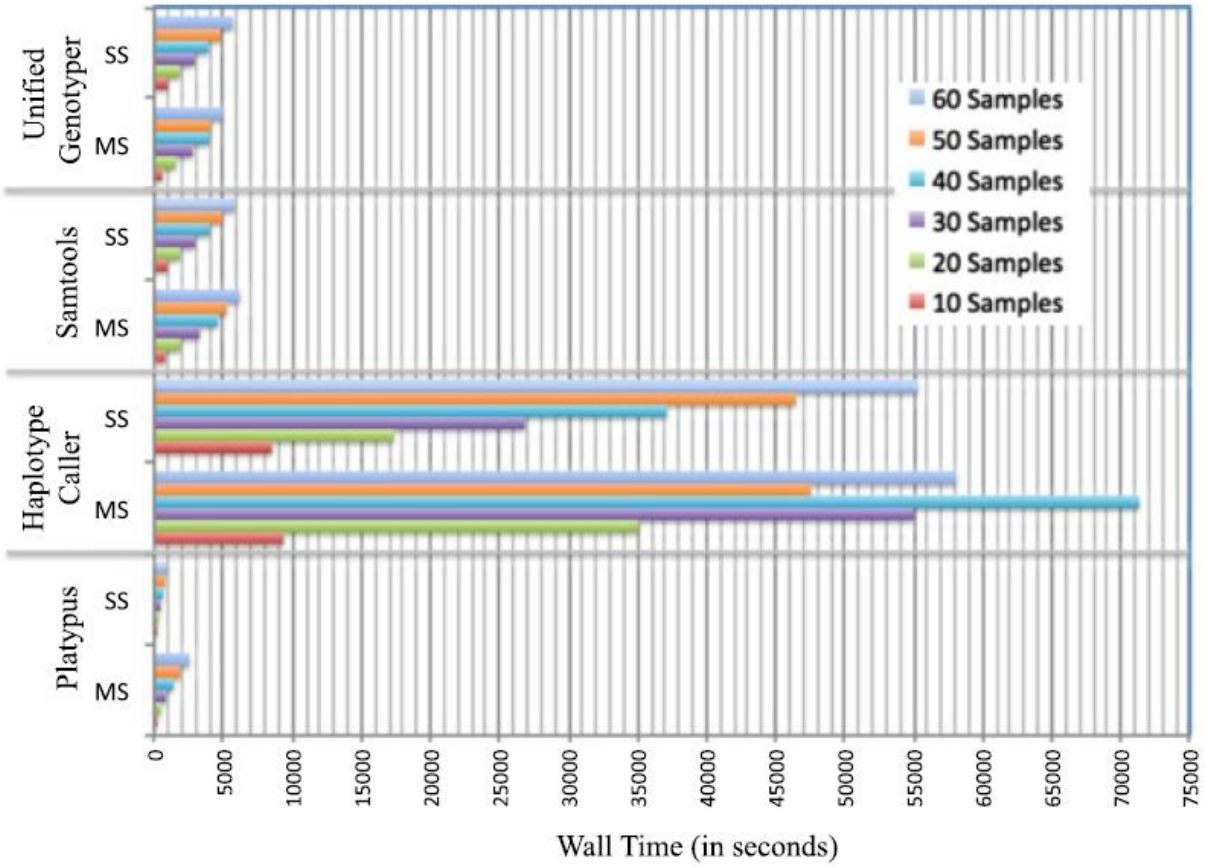
Baes et al., 2014, BMC Genomics

Concordance with bovine 800k SNP chip



Baes et al., 2014, BMC Genomics

Running time comparison



SS = single sample variant identification
MS = multi sample variant identification

Baes et al., 2014, BMC Genomics