AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Bioinformatics and Genome Analyses Course

Lecture 2 (Nov 13)

Stephan Pabinger
stephan.pabinger@ait.ac.at

http://pabinger.site44.com

Recap

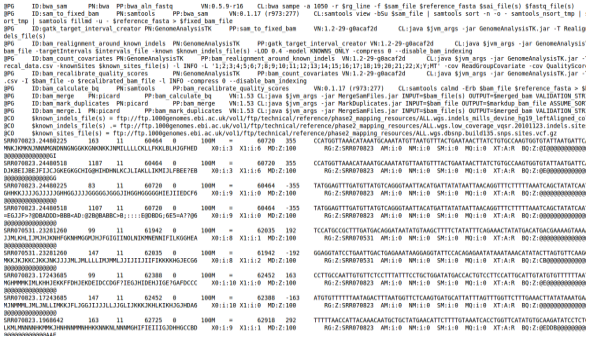
History

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology
1977	DNA sequencing
1985	*In Vitro* Amplification of DNA (PCR)
2001	The Human Genome Project

Different sequencing technologies

SAM/BAM files

Read duplicates



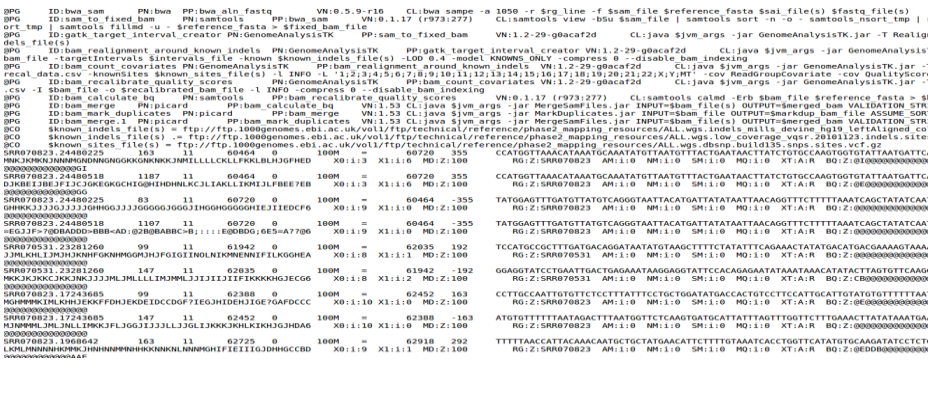
Recap

Python programs

- fasta_extractor.py

```
>Sequence_1 assembly1
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
>Sequence_2
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
```

- sam_examiner.py



Go to https://github.com/spabinger/BioinformaticsAndGenomeAnalyses2018

Coordinate systems

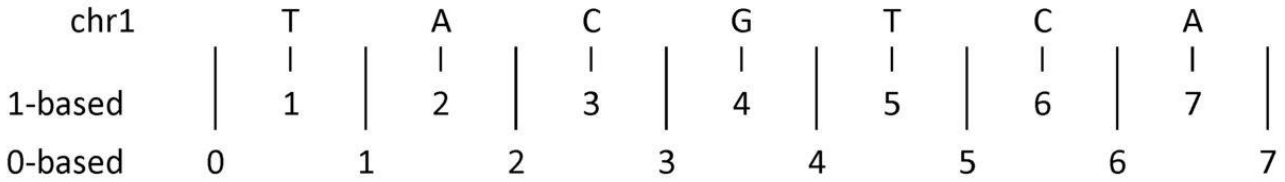
Coordinate system

- 0 based → 0, 1, 2, … 9 | 1 based → 1, 2, 3, … 10

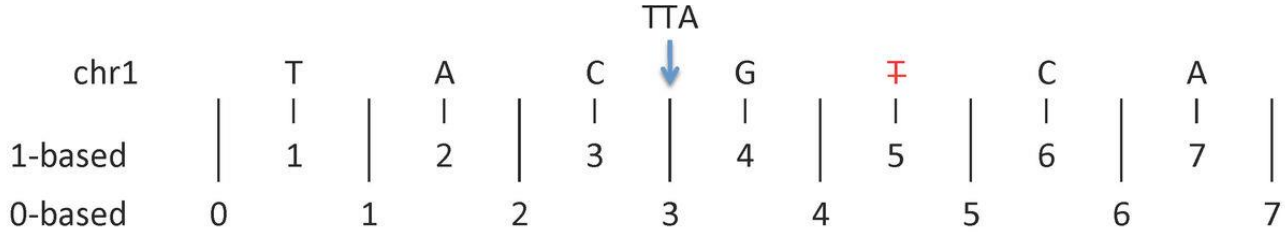
- BED – 0 based
- GFF – 1 based
- Ensembl uses a one-based coordinate system - UCSC use a zero-based coordinate system

	1 based	0 based
Third element	3	2
First ten	1, 10	0, 10
Second ten	11, 20	10, 20
One base long at 10	10,10	9,10
Interval	end – start + 1	end – start
Five elements at 100	100, 104	99, 104

Coordinate system



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A



	1-based	0-based
Indicate a deletion	chr1:5-5 T/-	chr1:4-5 T/-
Indicate an insertion	chr1:3-4 -/TTA	chr1:3-3 -/TTA

https://www.biostars.org/p/84686/ https://www.biostars.org/p/6373/

Conversion tool

convert_zero_one_based

- Python CLI
- convert between zero and one based coordinate systems
- Use: `convert_zero_one_based --help`

https://github.com/griffithlab/convert_zero_one_based

AUSTRIAN INSTITUTE OF TECHNOLOGY

Genome reference

Genome reference

Reference genome is a “consensus” across all chromosomes of DNA pooled from multiple individuals

UCSC and Genome Reference Consortium (GRCh)

hg18, hg19, hg38 ←→ GRCh36, GRCh37, GRCh38

Newest version (hg38) release on Dez 24th 2013

Download (e.g.)

- http://hgdownload-test.cse.ucsc.edu/goldenPath/hg38/bigZips/
- ftp://gsapubftp-anonymous@ftp.broadinstitute.org

	HG38 (UCSC)	GRCh38
Prefix	Chr	-
Mitochondrial	chrM	MT
Order	chrM, chr1, chr2, …chrX, chrY	1,2, …, X, Y, MT

Genome reference

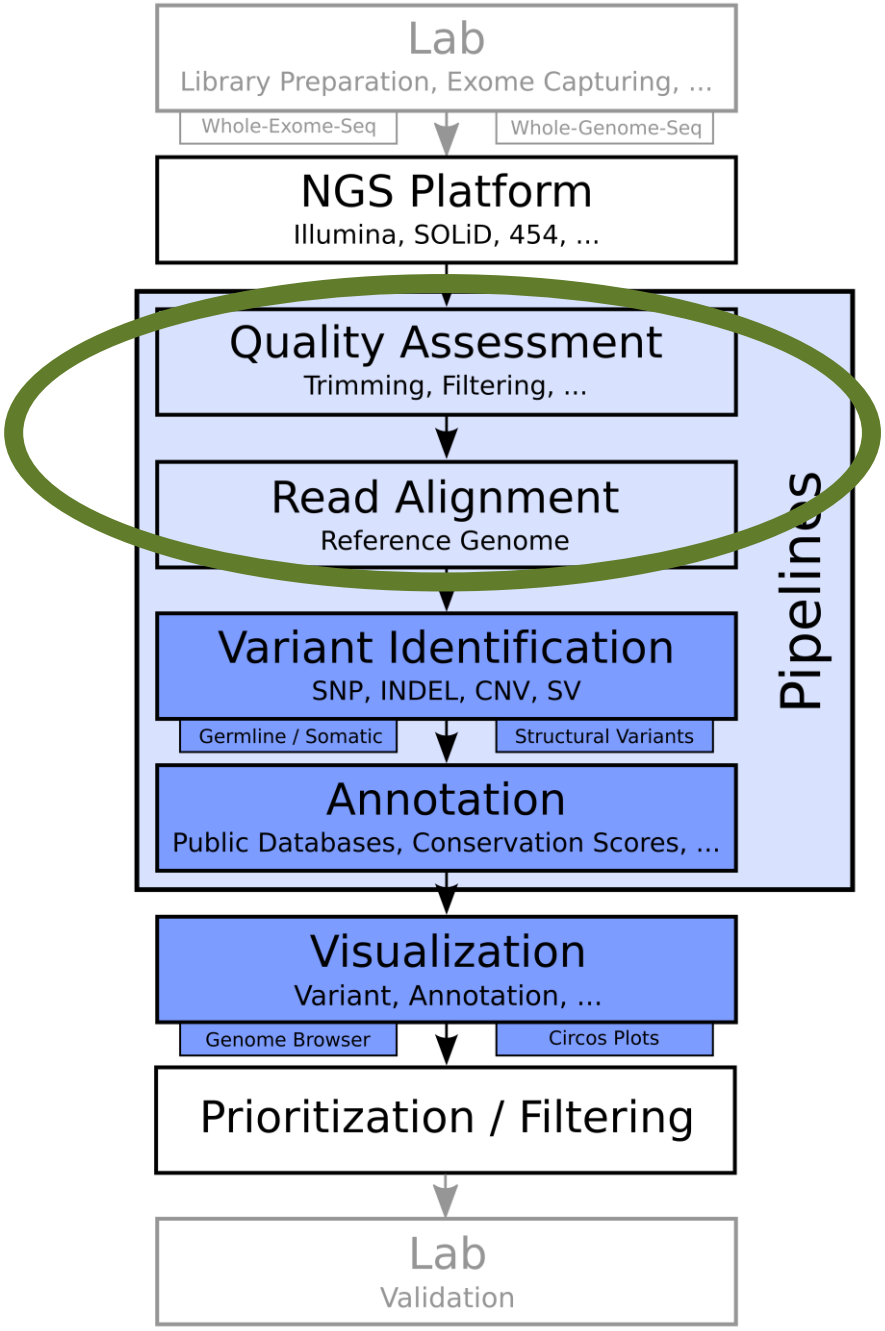
Indexing

- Fai file (created by samtools faidx)
  contig, size, location, bases-per-line and for efficient random
- Dict file (created by Picard CreateSequenceDictionary)
  SAM style header describing the contents of the fasta file
- Different mapping programs

Important

- Choose one reference genome (well sorted, indexed) and stick to it
- Be sure that previous variant calls use same reference - otherwise convert coordinates (lift-over)

http://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq#data-processing-ovw


Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

AUSTRIAN INSTITUTE OF TECHNOLOGY

FASTQ

FASTQ

Storing and defining sequences from next-generation sequencing technologies

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
T
+
!''*((((***+))%%%++)(%%%%).1***-
+*''))**55CCF>>>>>>CCCCCCC65
```

Sequence ID
Sequence
‹separator›
Quality scores

Old format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

New format

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Cleaning up FASTQ files

Before mapping make sure

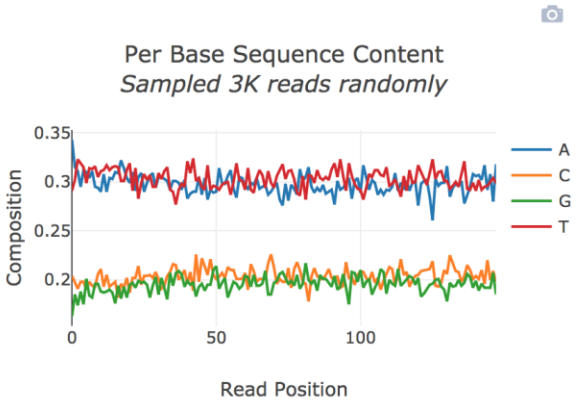
- Non genomic sequences are removed (barcodes, …)
- Adapter sequences are removed
- Clean contaminations (PRINSEQ, DeconSeq)
- Trim Ns
- Trim bad quality reads

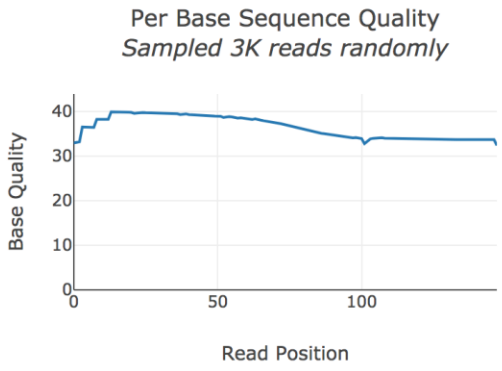
Skip cleaning → less read mapping; if not randomly distributed some areas wont get enough coverage

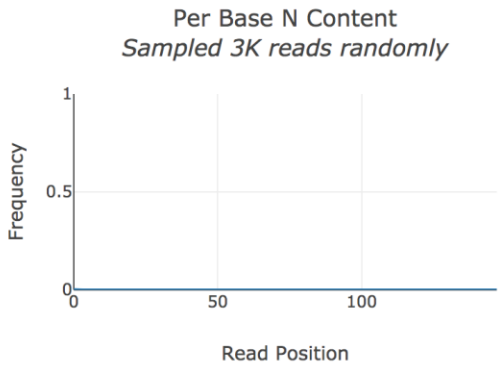
Tools for FASTQ manipulation

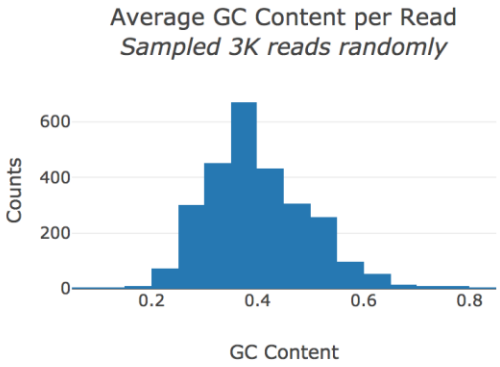
- FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  HTML output
- Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit/
  Lots of tools, charts, trimming, clipping, filtering
- Cutadapt https://code.google.com/p/cutadapt/
  Remove adapter sequences
- DeconSeq http://deconseq.sourceforge.net/
  User friendly interface, coverage plots, metagenomics datasets
- PRINSEQ: http://prinseq.sourceforge.net/
  HTML output, trimming, filtering, contaminations
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic
  Paird-end trimming
- Atropos https://github.com/jdidion/atropos
  Multi-threading, paired-end, bisulfite-seq, …
- ...

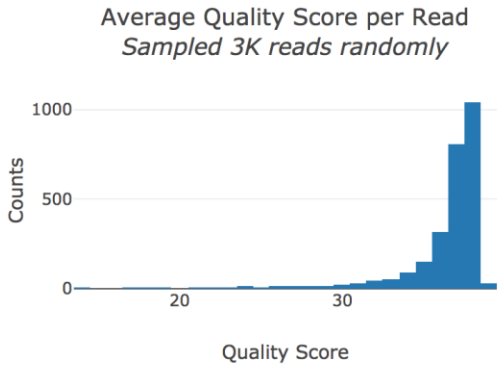
www.fastq.bio

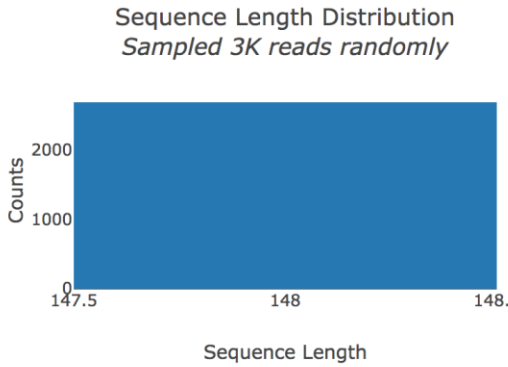








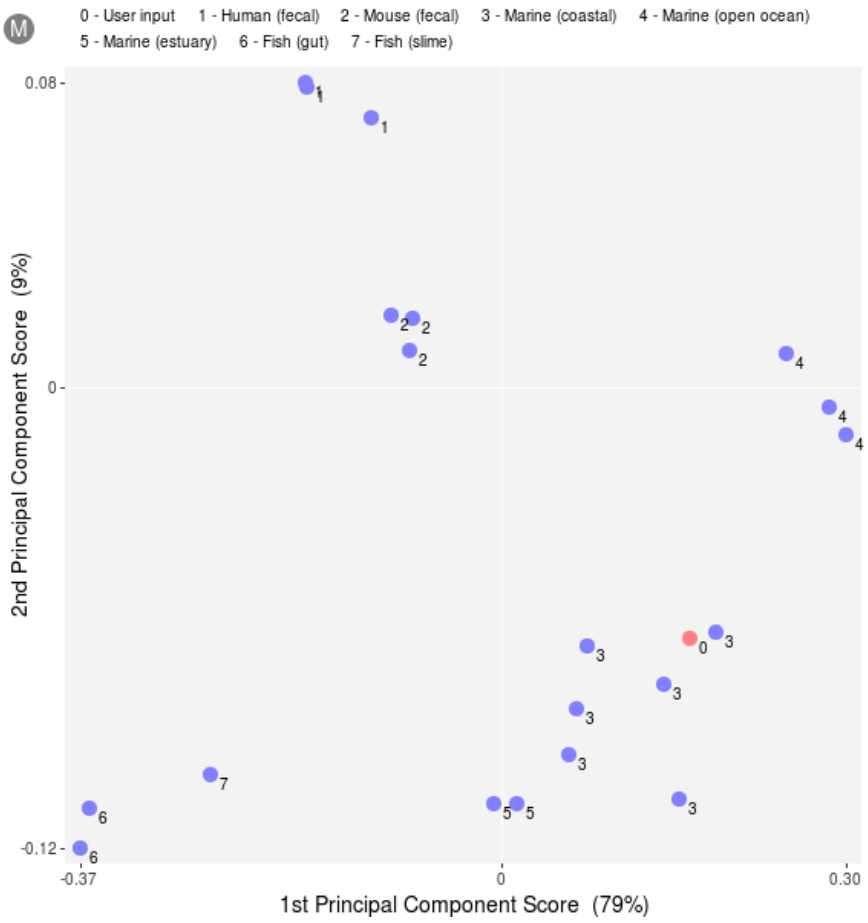




Check for contamination

PRINSEQ

- Dinucleotide (e.g., TA, GC, …) odds ratios
- Principal component analysis (PCA) to group metagenomes

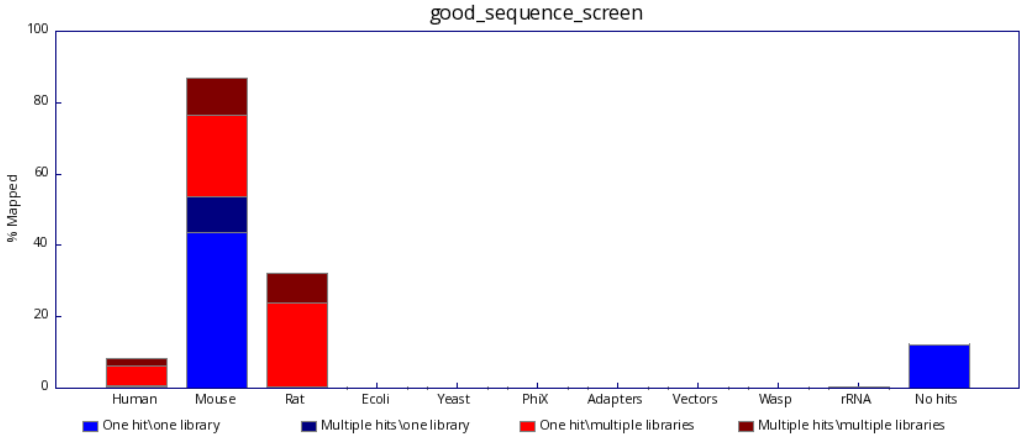


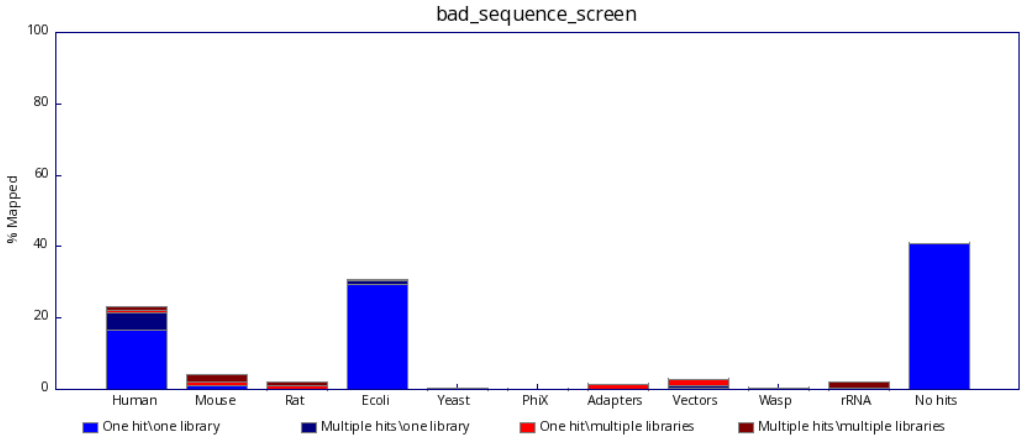
Check for contamination

FastQ Screen

Screen against a set of sequence databases:

- genomes of all of the organisms you work on
- PhiX
- Vectors
- …

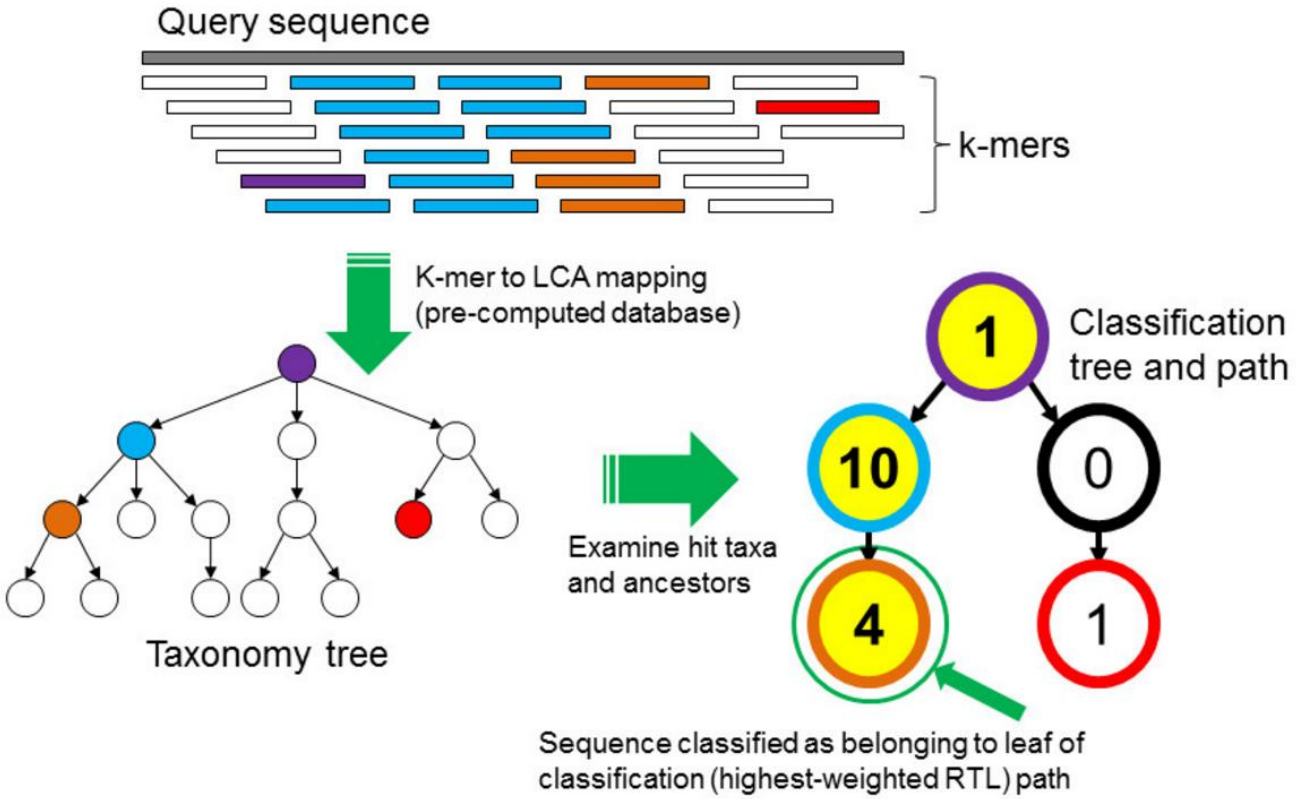




Check for contamination

Kraken (https://ccb.jhu.edu/software/kraken/)

- Assigning taxonomic labels to short DNA sequences
- Detect metagenomics contaminations



Phred quality score

Characterize the quality of DNA sequences

$$q = -10 \log_{10}(p)$$

p = error probability for the base

Phred quality score	Probability	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Widely used tools

SAMtools

- Utilities for manipulating SAM files

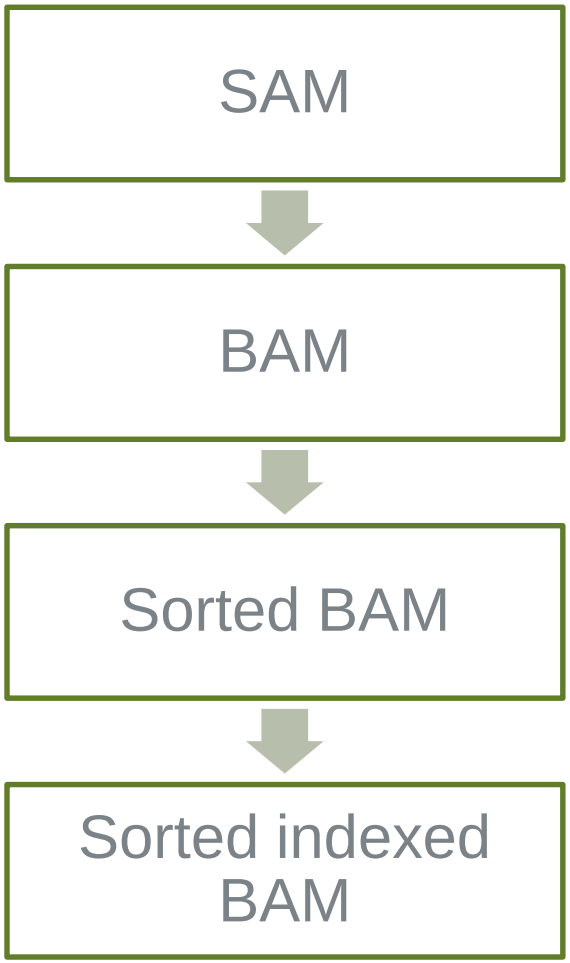
GATK

- Genome Analysis Toolkit - wide variety of tools, variant discovery and genotyping, quality assurance

Picard

- Utilities for manipulating SAM files

SAM / BAM



Convert (view)

sort

index

Quality check of alignment

Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

Metrics

- Coverage / nucleotide distribution
- Reads mapped outside of a target (e.g., Exome sequencing)
- Number of mapped reads (wrong reference genome?)
- Insert size statistics
- Mapping quality - rule of thumb: Anything less than Q20 is not useful data

Tools

- Qualimap 2
  http://qualimap.bioinfo.cipf.es/
- bamstats
  http://bamstats.sourceforge.net/

Read Group Tag (RG)

RG - Meta information

- ID: unique e.g., SRA number (Sequence read archive)
- PL: Sequencing platform
- PU: Platform unit (run name / flowcell-barcode-lane)
- LB: Library name
- PI: Insert size (Predicted mean insert size)
- SM: Sample (Individual)
- CN: Sequencing center

PICARD

```
java -jar picard/AddOrReplaceReadGroups.jar I=exampleWO_RG.bam O=
exampleW_RG.bam SORT_ORDER=coordinate RGID=superID RGLB=superLib
RGPL=illumina RGSM=superSample
```

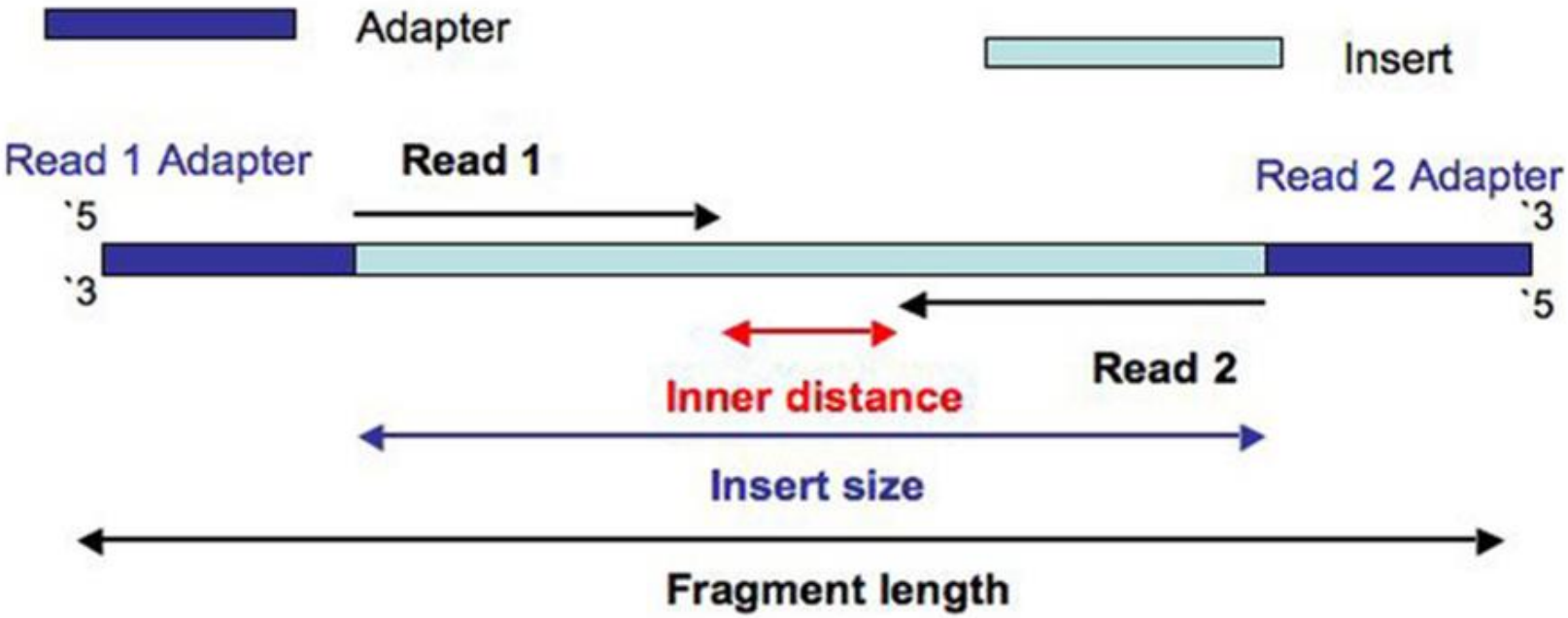
Read Group Tag – why important?

- It refers to a set of reads that were generated from a single run of a sequencing instrument.
- When multiplexing is involved, then each subset of reads originating from a separate library run on that lane will constitute a separate read group.
- To see the read group information for a BAM file, use the following command:

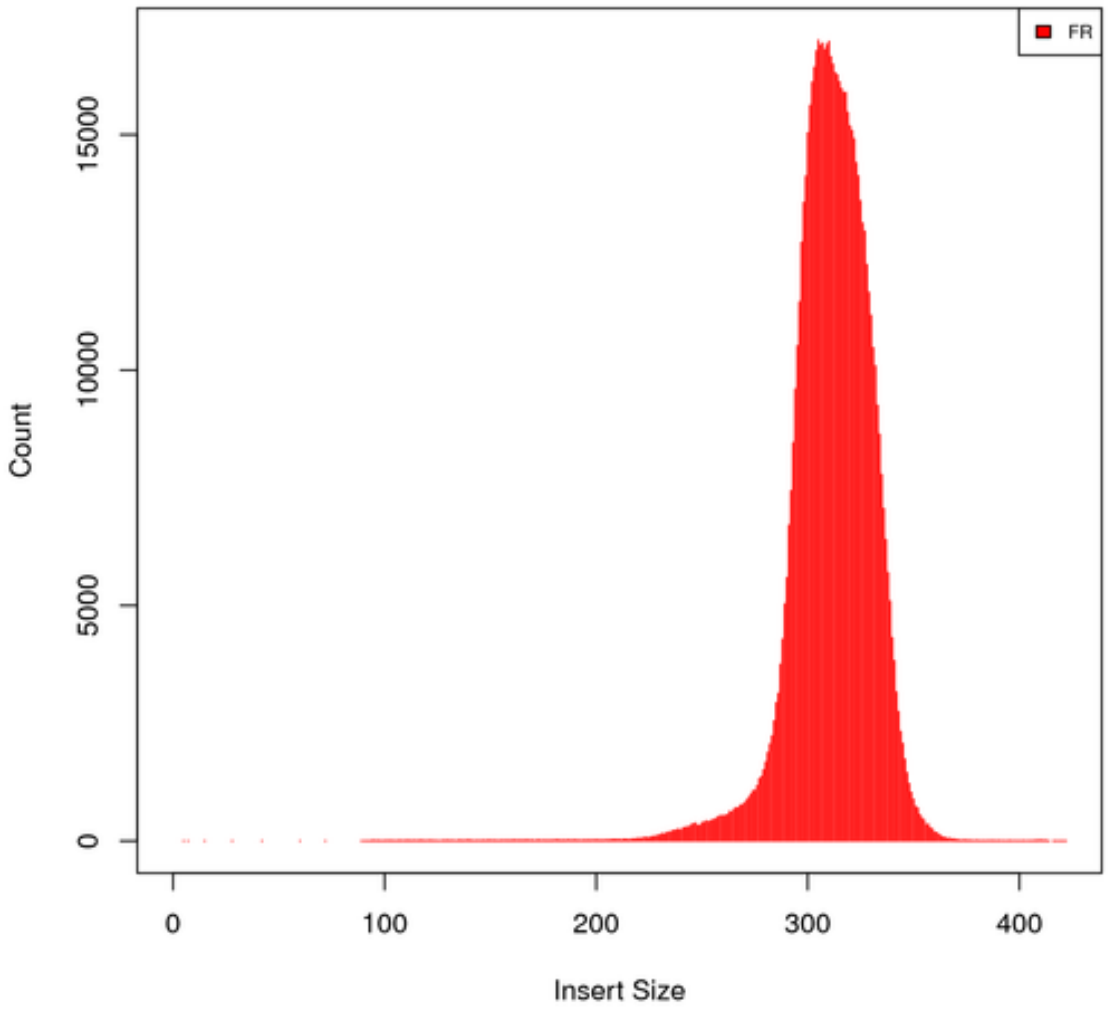
```
samtools view -H sample.bam | grep '@RG’
```

- Check that your FASTQ files have appropriate RQ when performing demultiplexing
- Important to have a correct RG tag → required by bioinformatics analysis tools (GATK, …)

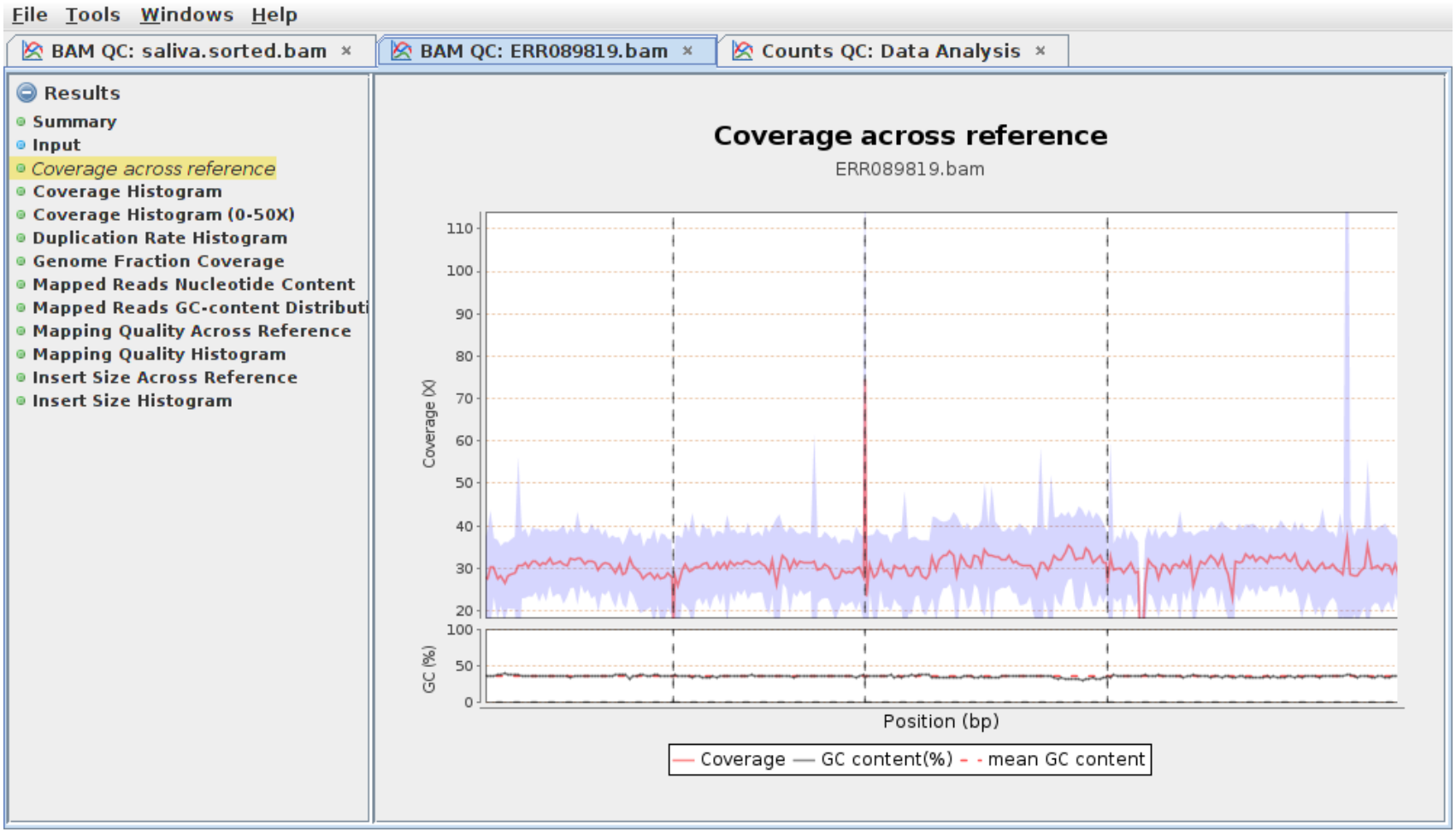
Insert size


Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length

Insert size histogram



Qualimap 2


File Tools Windows Help
BAM QC: saliva.sorted.bam
BAM QC: ERR089819.bam
Counts QC: Data Analysis
Results
Summary
Input
Coverage across reference
Coverage Histogram
Coverage Histogram (0-50X)
Duplication Rate Histogram
Genome Fraction Coverage
Mapped Reads Nucleotide Content
Mapped Reads GC-content Distributi
Mapping Quality Across Reference
Mapping Quality Histogram
Insert Size Across Reference
Insert Size Histogram
Coverage across reference
ERR089819.bam
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

Read duplicates

Origin

- PCR amplification step in library preparation
  1. Get DNA pieces (shatter / enrich DNA)
  2. Ligate adapters to both ends of the fragments
  3. PCR amplify the fragments with adapters
  4. Put fragments on beads or across flowcells
  5. Amplify fragments
  6. Sequence

Identification

- Have the same starting position

Read duplicates

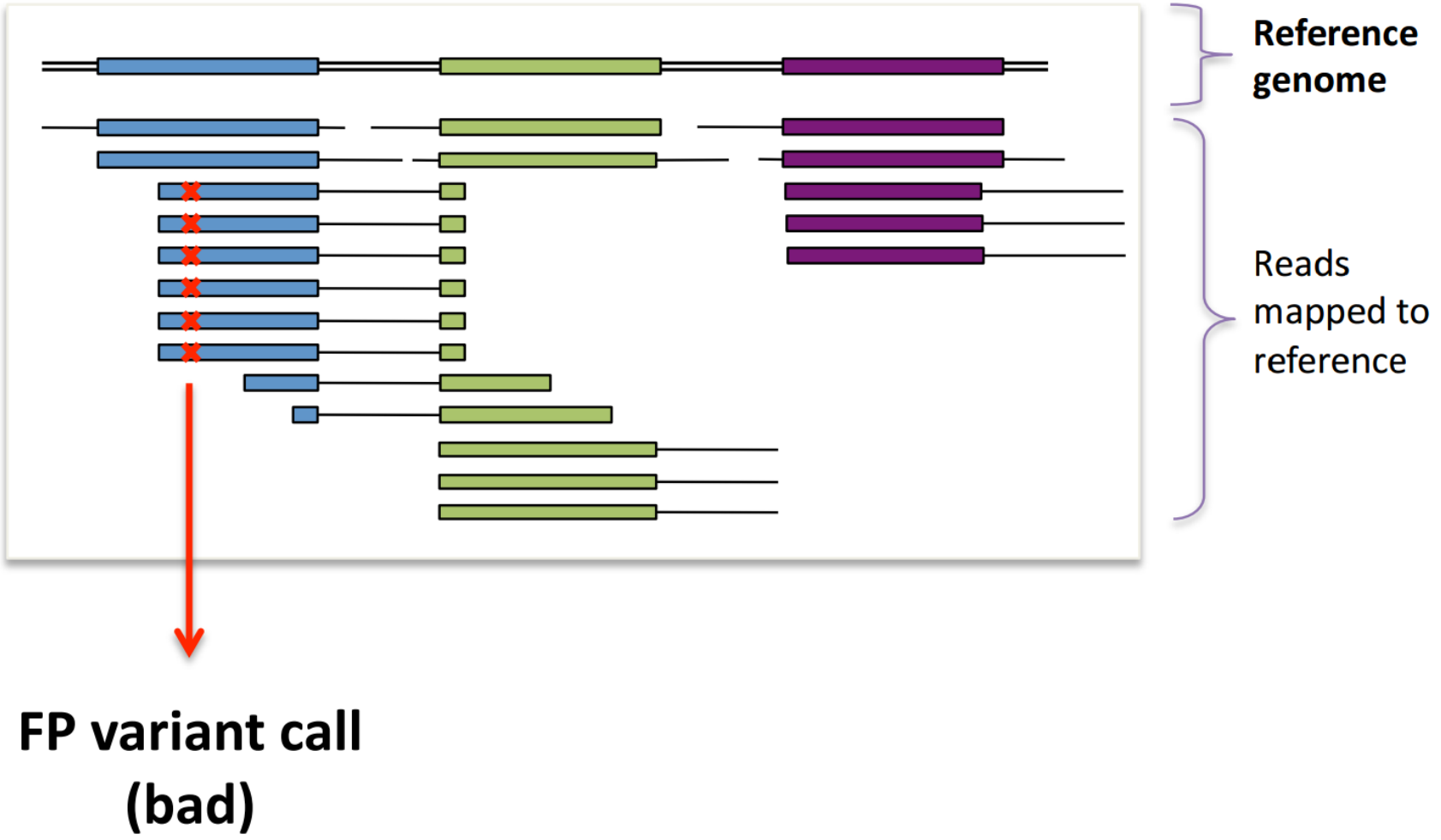
Problems/causes

- More steps during PCR amplification with little input material → more duplicates
  - This may result in duplicate DNA fragments in the final library

- Higher rates (~30%) arise when too little starting material is used → more amplification of the library is needed

- May result in false SNP calls (statistical model gets mixed up)

Read duplicates

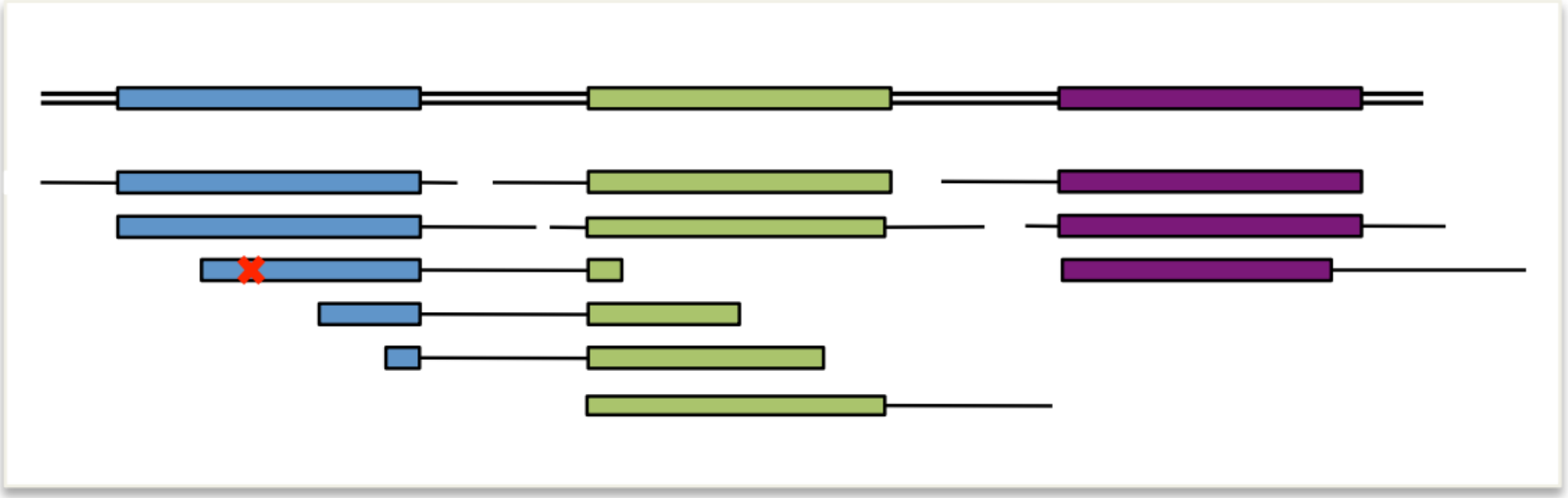


http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

Read duplicates - removal

- → Identify reads that map to the same location
- → Remove all but one

After marking duplicates



Read duplicates - removal

Attention!

Do not remove for

- Haloplex enrichment (nonrandom fragmentation method)
- PCR based enrıchment

You would remove results

Base Quality Score Recalibration

- Various sources of systematic error
  → over / under estimated base quality scores

- Quality score assigned to single base in isolation
  (assigned by the sequencing machines)

- Variant calling algorithms rely heavily on base quality scores

Solution → Correct base quality scores

Base Quality Score Recalibration

Apply machine learning to model these errors empirically and adjust the quality scores accordingly

- First the program builds a model of covariation based on the data
  - Reported quality score
  - Position in the read (cycle)
  - Preceding and current base – sequence context (homopolymer, ...)
- and a set of known variants (1000g, dbSNP, large private cohort)
  - discount most of the real genetic variation

- First pass: calculate new QS based on the model
  Second pass: adjust the base quality scores

→ Visual inspection with before/after plots

Good explanation: http://zenfractal.com/2014/01/25/bqsr/

Base Quality Score Recalibration

WES

- For WES restrict to capture targets
  off-target sites are likely to have higher error rates

Organism with no known variants?

- Call variants -> apply stringent filter -> use these for recalibration
- Repeat previous steps

https://www.broadinstitute.org/gatk/events/slides/1506/GATKwr8-B-3-Base_recalibration.pdf

Informative NGS QC page

Working with SAM/BAM files

SAMTOOLS

http://samtools.sourceforge.net

View

```
samtools view -h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view -f 0x02 <file.bam> > <only_proper_paired.sam>
```

SAMTOOLS

http://samtools.sourceforge.net

View

```
samtools view –h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view –f 0x02 <file.bam> > <only_proper_paired.sam>
```

SAMTOOLS

http://samtools.sourceforge.net

View

```
samtools view –h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view –f 0x02 <file.bam> > <only_proper_paired.sam>
```

Sort

```
samtools sort -o <sorted.bam> <aln.bam>
```

Index

```
samtools index <sorted.bam>
(only for BAM)
```

SAMTOOLS

Simple stats (reads mapped, reads paired, ...)

```
samtools flagstat <file.bam>
```

Stats for each chr/contig - reads mapped and unmapped

```
samtools idxstats <sorted.bam> (and indexed)
```

Convert BAM to SAM

```
samtools view -S -h -b <aln.bam> > <aln.sam>
```

Converting BAM to a sorted BAM file (without intermediate file)

```
samtools view -bSh <file.sam> | samtools sort -o
file_sorted.bam -
```

http://davetang.org/wiki/tiki-index.php?page=SAMTools

Other tools

SAMBAMBA - „Multithreaded" SAMtools - https://github.com/lomereiter/sambamba

- view
- sort
- index
- merge
- flagstat
- markdup

elprep - https://github.com/exascience/elprep

- High-performance tool for preparing .sam / .bam / .cram files
- In-memory and multi-threaded application
- Requires lots of memory (WGS ~256GB)
- Replacement for SAMTOOLS & Picard

SAMBAMBA

„Multithreaded” SAMtools

- view
- sort
- index
- merge
- flagstat
- markdup

- Binaries available

https://github.com/lomereiter/sambamba

elprep

- High-performance tool for preparing .sam / .bam / .cram files
- In-memory and multi-threaded application
- Requires lots of memory (WGS ~256GB)
- Replacement for SAMTOOLS & Picard

https://github.com/exascience/elprep

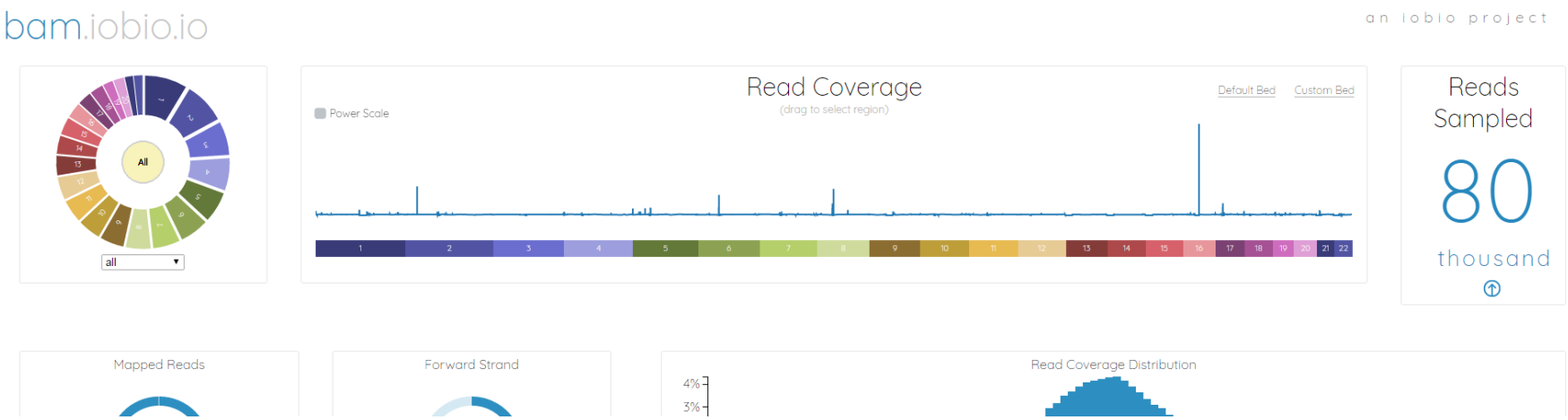
PICARD & iobio

PICARD

- JAVA based tool (http://picard.sourceforge.net/)
- BuildBamIndex, FastqToSam, MergeSamFiles, …

bam.iobio.io

- Web-based (http://bam.iobio.io)
- Coverage overview
- Mapping overview



Multiple mapping

Multiple mapping - MAPQ

Bowtie 1

- 255 for uniquely mapped reads (default)
- 0 for multiply mapped reads

Bowtie 2

- MAPQ filter of >=40 to get reads which had only 1 convincing alignment
- Lower filter to allow multi-mapped reads; secondary alignment with varying degrees of difference to the primary

BWA

- Phred scores as MAPQ values
- 0 if read maps to multiple positions

Genetic variations

Genetic variation

A regular diploid human cell contains 46 chromosomes (23 pairs)

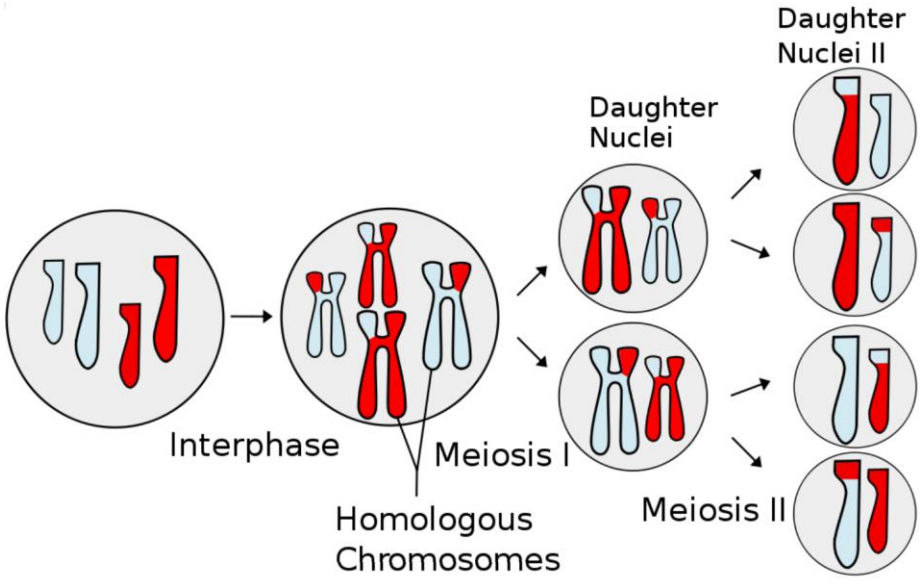
- 22 pairs + sex chromosomes XX(female) XY(male)
- One set of chromosomes inherited from each parent

Germline

- Meiosis → four genetically unique **haploid** gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell

Somatic mutation

- Not inherited from parent
- Acquired from spontaneous mutations during DNA replication



Types of genetic variations

SNV / SNP

- A single nucleotide — A, T, C or G — in the genome differs between members of a population or chromosome pairs
- Originally defined as occurring at least in one individual of the population (these definitions may shift in time)
- SNV (single nucleotide variant) if observed very rarely
- SNP, SNV → may fall within
  - coding sequences of genes
  - non-coding regions of genes
  - intergenic regions

Types of genetic variations

INDEL

- Insertion / deletion of bases
- Coding regions of the genome - produce a frameshift mutation (unless multiple of 3)
- There are approximately 190-280 frameshifting INDELs in each person.
  A map of human genome variation from population-scale sequencing". Nature 467 (7319)

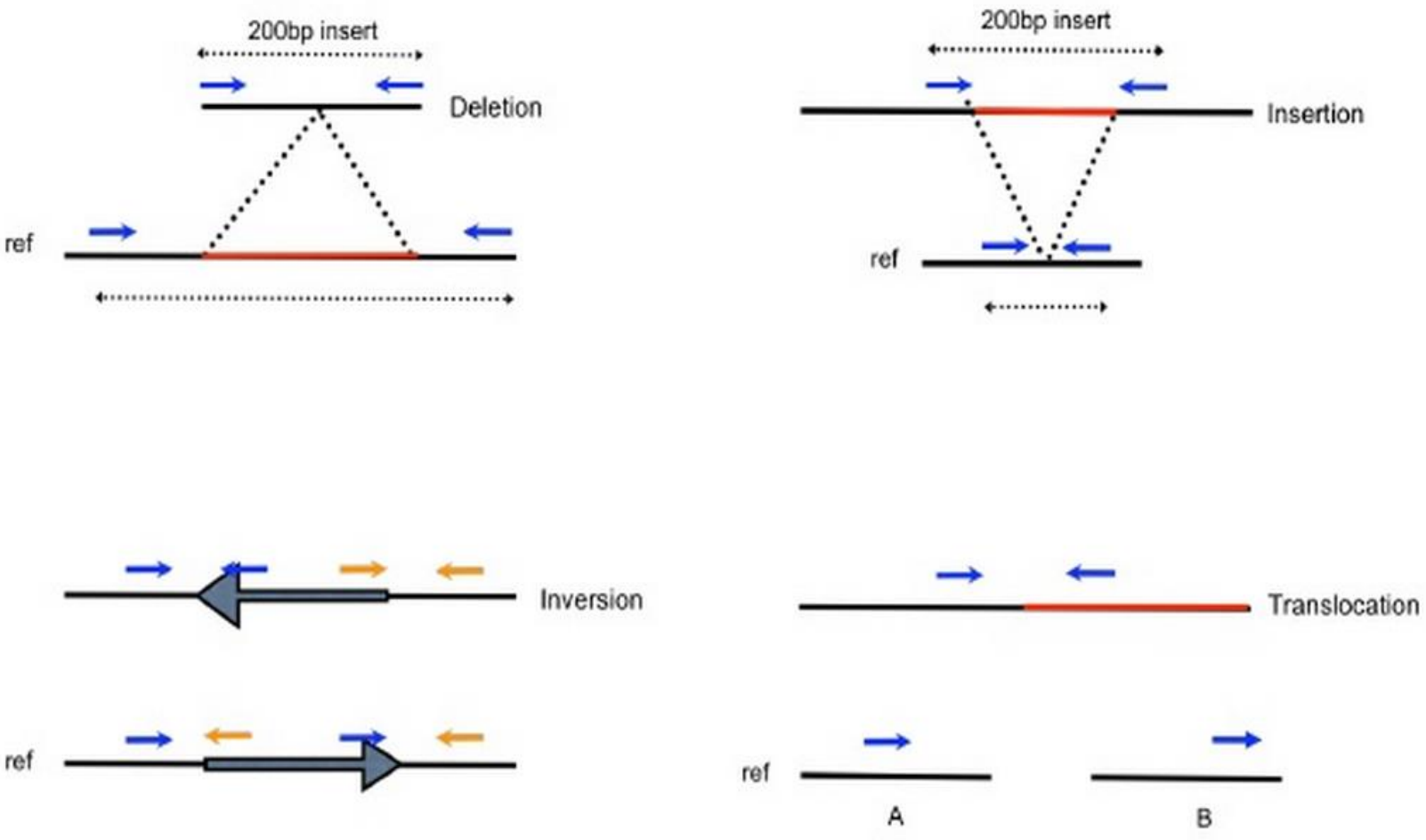
Types of genetic variations

Structural variations (SV)

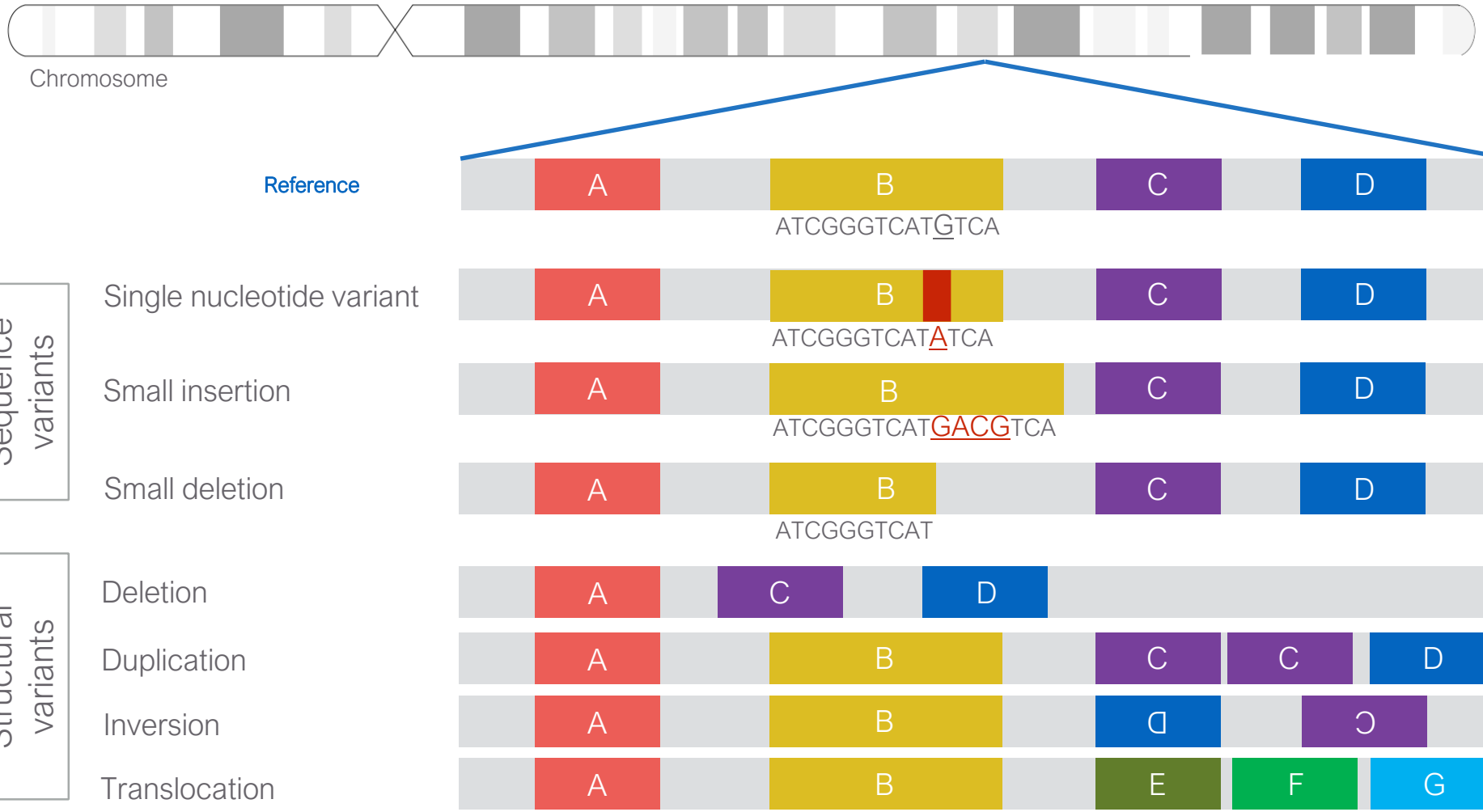
- Variation in structure of an organism's chromosome

- Insertions
  Deletions
  CNV
  Inversions
  Translocations

Types of structural genetic variations



Summary DNA Variants



Credit: Elizabeth Ruzzo, PhD, CHGV

Human genome

- Typical genome differs from the reference human genome at **4.1 million to 5.0 million sites**.

- **~99.9%** of variants consist of **SNVs and INDELs**

- Structural variants affect more bases
  - 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~1100 insertions)
  - Affecting ~20 million bases of sequence

What kind of data do you have?

Genotyping

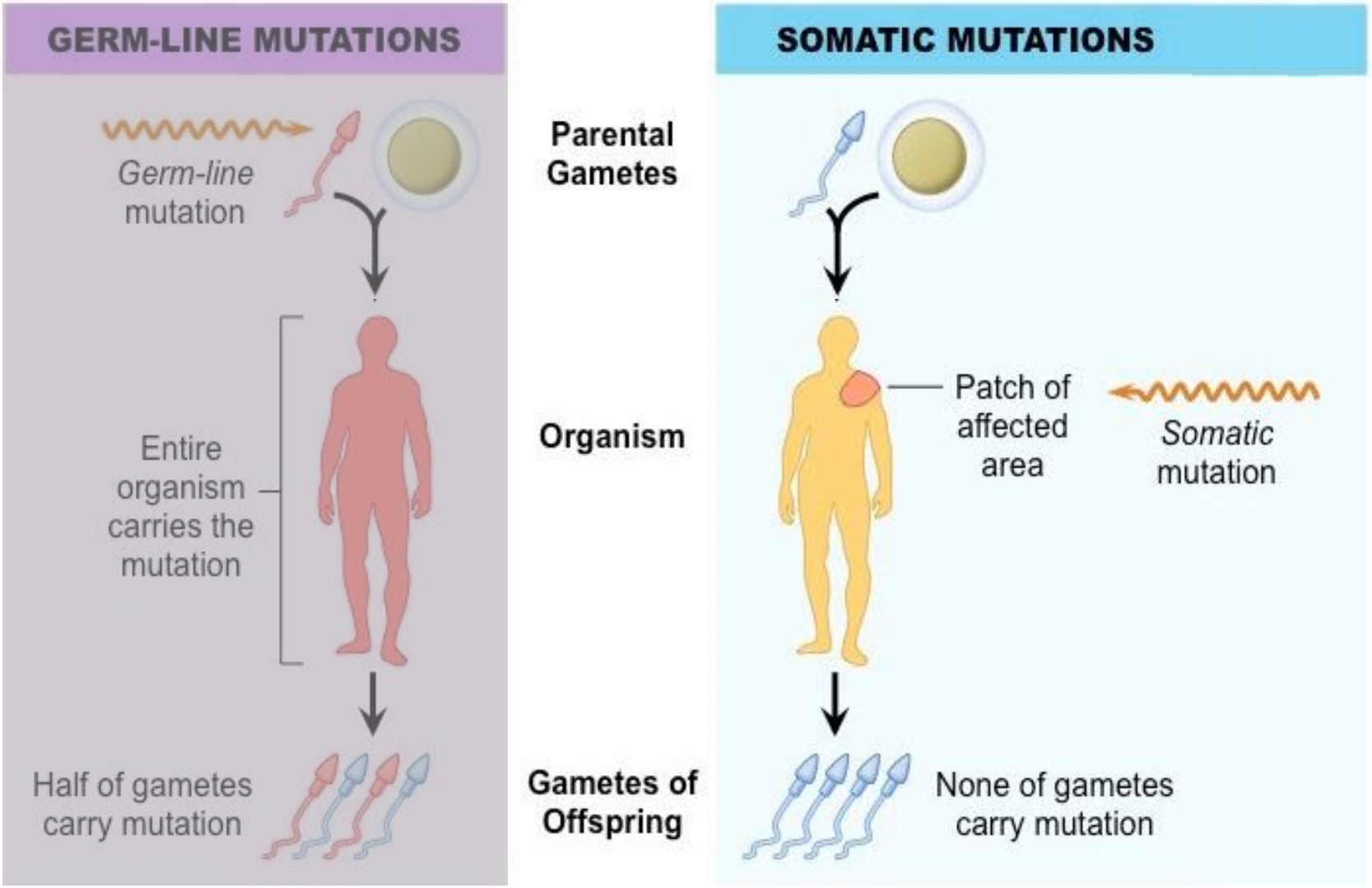
- Single samples
- Family → finding causing mutation of rare disease
- Time series

Somatic mutations

- Primary tumor tissue
- Blood samples

→ Different callers / strategies for respective samples

Somatic mutations



Somatic mutations

- Change in genetic structure **not** inherited from a parent
- Constantly happening in living organisms
- Can be used for detection of disease

- Detected by comparison to „normal“ cells

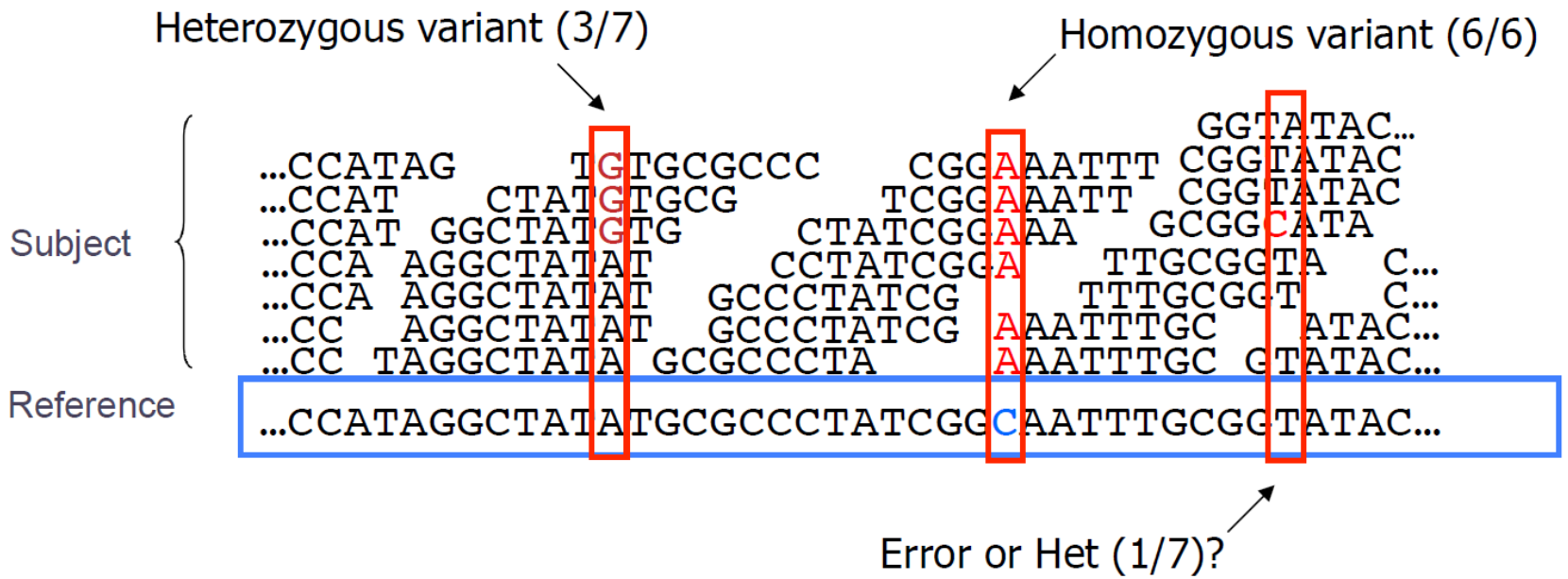
Cancer

- Used to characterize tumor cells

AUSTRIAN INSTITUTE OF TECHNOLOGY

Variant Calling

Genotyping theory



- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!
- Sequencing instruments make mistakes
  – Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

What are variant calls?

Find differences to a reference (hg19, GRCh38)

Naive variant calling

- Check all the reads that cover base chr11:1234567
- Add up the bases at chr11:1234567
- e.g. 15 A's, 4 G's
- Is this an A/G heterozygous site or four sequencing errors?

Actual variant callers

- Estimate likelihood of a variant site vs a sequencing error
- Sequencing error rate
- Quality scores

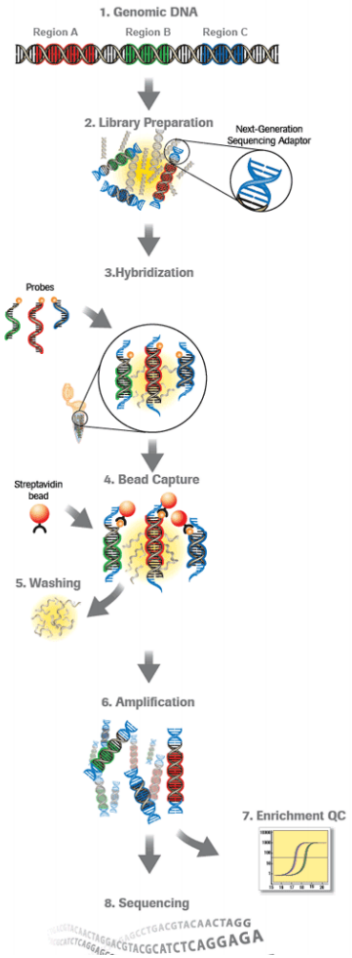
(Other) reasons for a mismatch

- True SNP

OR

- Error generated in library preparation
- Base calling error
  - May be reduced by improved base calling methods, but cannot be eliminated
- Misalignment (mapping error)
  - Local realignment to improve mapping
- Error in reference genome sequence

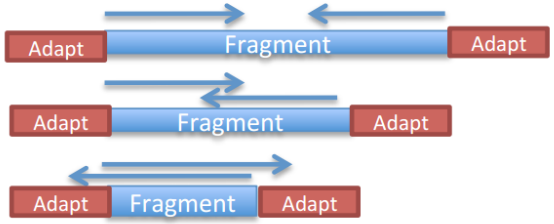
Difficulties



Sonication

Library prep (sequencing adaptors on)

Problem: **error in sonication** >> adaptor seq in reads >> **unmapped reads**



Hybridisation to probes → Possible biases in sequences that hybridise >> **coverage bias**

Bead capture → Possible biases in sequences that elute >> **coverage bias**

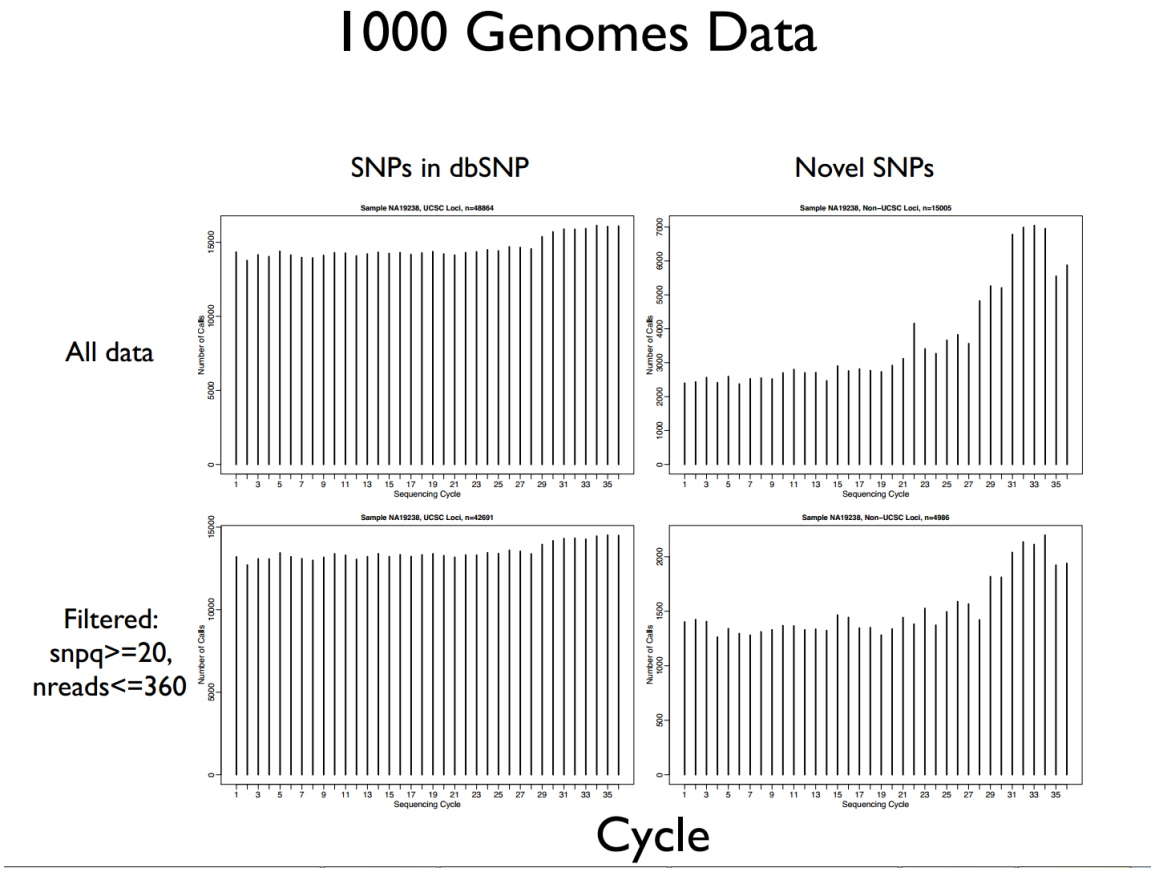
Amplification → Possible biases in sequences that amplify **>> sequence PCR duplicates**

Sequencing → Possible biases in sequences that bridge PCR >> **coverage bias**

https://wiki.uio.no/projects/clsi/images/8/80/VariantCallingCourse_sept2013_Part1.pdf

Difficulties

- Base calling gets more difficult the longer the read gets



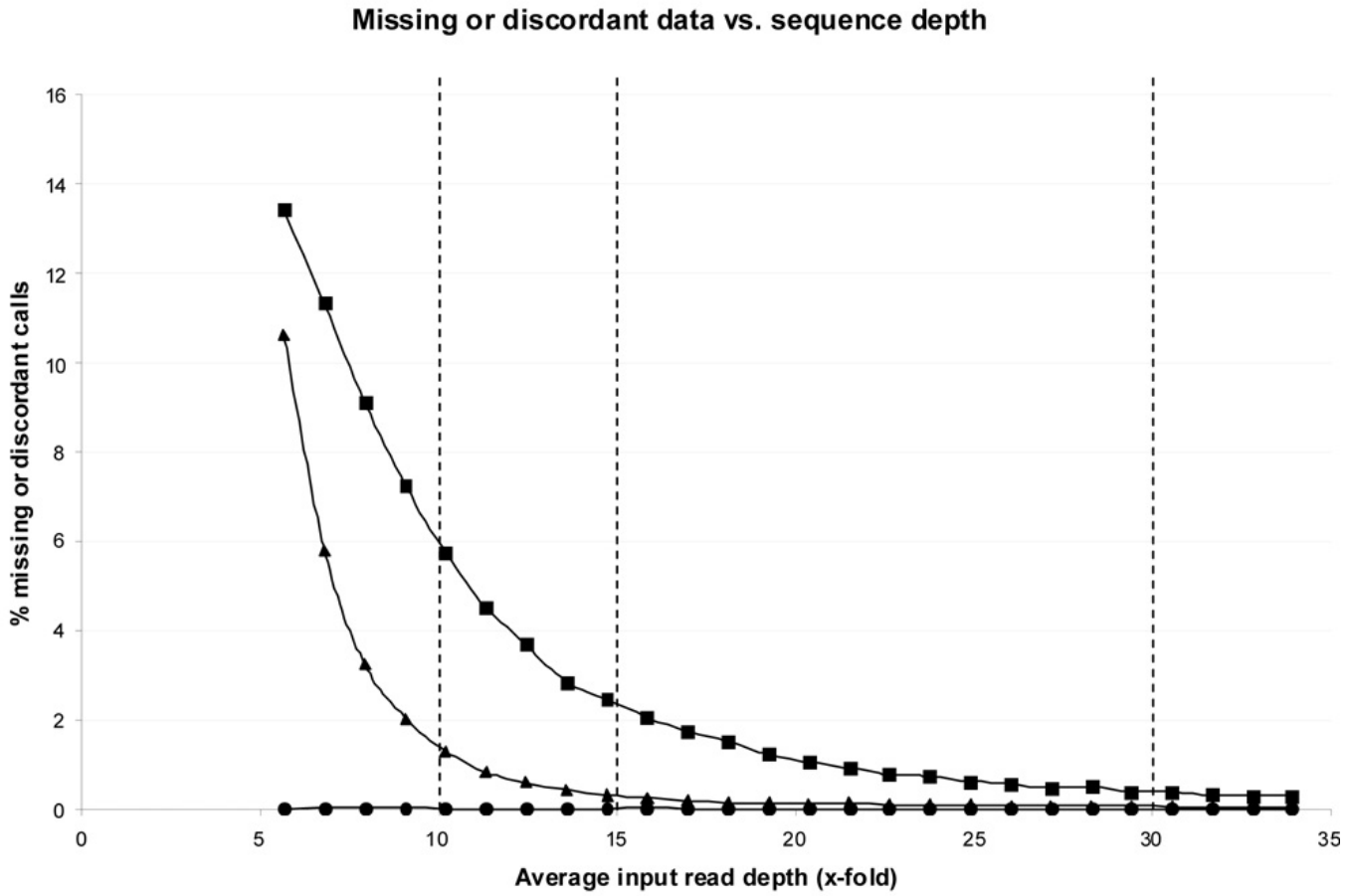
http://www.biostat.jhsph.edu/~khansen/LecIntro1.pdf

More difficulties

- High depth - low quality regions → likely due to copy number or other larger structural events
- Repetitive regions → artefactual variants
- Regions of low complexity (~ 2% of genome)
  polypurine (AG), AT-rich regions, simple tandem repeats

Linderman et al. **Analytical validation of whole exome and whole genome sequencing for clinical applications.** *BMC Medical Genomics* 2014, **7**:20

Heng Li. **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** http://arxiv.org/pdf/1404.0929v1.pdf



Squares Not covered by sequence

Triangles Heterozygote undercalls in sequence data relative to genotype data

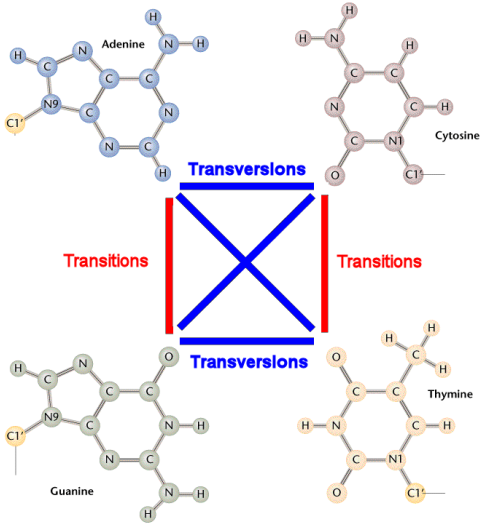
Circles Discordant SNV calls compared to genotypes

Undercall - (false negatives; true genotype is polymorphic yet the call is homozygous reference or 'missing')
Bentley, et al. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry, Nature

Ti/Tv ratio

Transition (Ti)

- purine <-> purine (A <-> G)
- pyrimidine <-> pyrimidine (C <-> T)



Transversion (Tv) purine <-> pyrimidine

A <->C, A <->T, G <->C, G <->T

Transition is more frequent than transversion

- Ti/Tv ~ 2.0 - 2.1 for genome wide
- Ti/Tv ~ 3.0 - 3.3 for exonic variations
- Ti/Tv = 2/4 = 0.5 = 2/4 =0.5 for random, uniform sequencing - sequencing error

Practicals – Day 2