Lecture 4

Introduction to bioinformatics (MVE510)

Autumn 2020

Additional reading: A beginner's guide to SNP calling from high-throughput DNA-sequencing data. Altmann A, Weber P, Bader D, Preuss M, Binder EB, and Müller-Myhsok B. Human Genetics, 131(10) 2012.

Repetition

- Errors are common in DNA sequence data. The error patterns depends on the underlying chemistry.
- Illumina has a high error rate for substitutions and a low error rate for indels. GC-rich regions has the highest error rate.
- Pre-processing aims to remove reads and bases that are not of sufficient quality.
- Filtering and trimming are two common ways to do pre-processing.

Todays agenda: genome sequencing

- Genome evolutions – SNPs, indels and structural variation
- Analysis of data from genome sequencing
- Coverage, quality refinement, score recallibration
- SNP calling
  - A naïve approach
  - The GATK unified genotyper
- Introduction to computer exercise 2

Genome evolution

- The genome of an organism is evolved through random mutations
- For humans and many other multicellular organism mutations can be either germline or somatic
- Germline mutations
  - Inherited from the parents
  - Passed to the offspring (present in germ cells)
- Somatic
  - Mutations that are not inherited
  - Not passed to the offspring

Genome evolution – SNPs and indels

- Single nucleotide polymorphisms (SNPs) are substitutions affecting a single nucleotide position
- If a SNP is in a coding region and results in an amino acid change it is called non-synonymous otherwise it is called synonymous
- Indels are insertion or deletions affecting one or a few nucleotides
- Indels can result in frameshifts which can impair a complete gene

Example of SNPs and indels

Single nucleotide polymorphisms (SNPs)

Genome 1 AGTATAGTAGGGGTACAGTGGGTAAG

Genome 2 AGTATAGTAGCGGTACAGAGGGTAAG

Insertions and deltions (indels)

Genome 1 CGATAGGTATTT-ACCCAGACCCCCTGT

Genome 2 CGATAGGTATTTTACCCAGAC---CTGT

Frame shifts

Before indel

M L V V D P P G

AGATGCTGGTGGTCGATCCGCCGGGA

After indel

M L C G R S A G

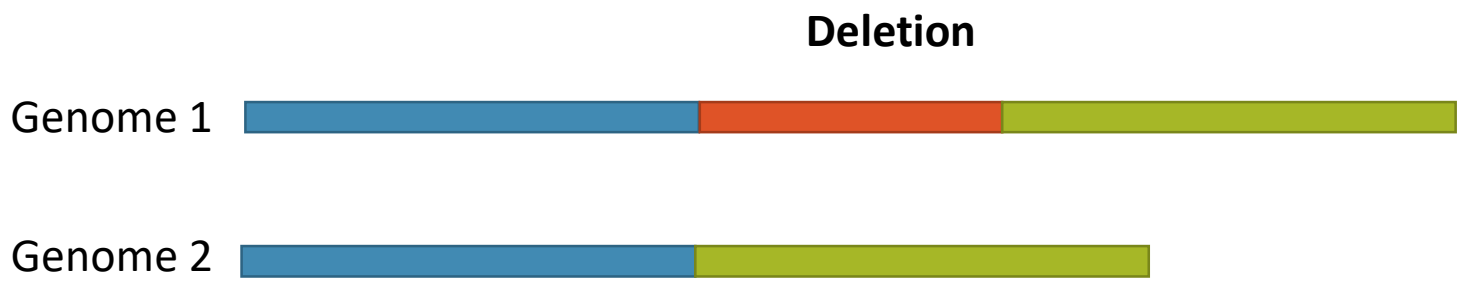
AGATGCTGTGTGGTCGATCCGCCGGG

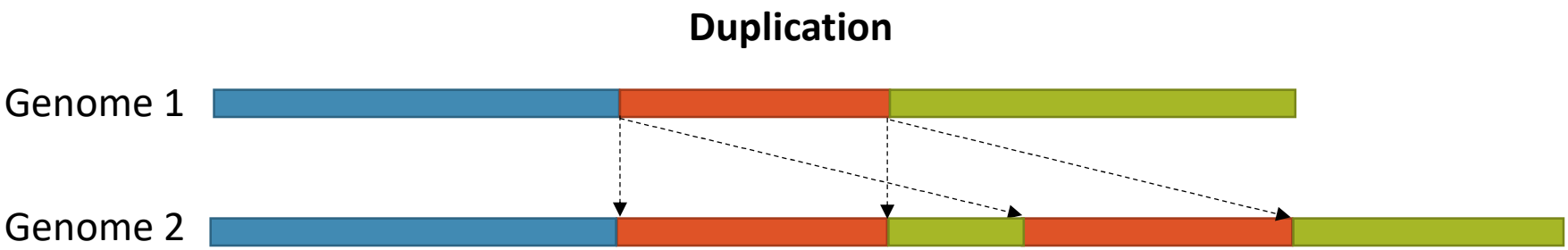
Genome evolution – structural variants

Copy number variation (CNV)

- A chromosomal region that is duplicated or deleted
- Can be caused by many mechanisms including impaired DNA replications mechanisms (e.g. in cancer)
- Used by e.g. bacteria to regulate gene expression – more copies of a genes means higher expression
- Important in evolution: the genes in the new region can evolve into completely novel biochemical functions (neofunctionalization)

Genome evolution



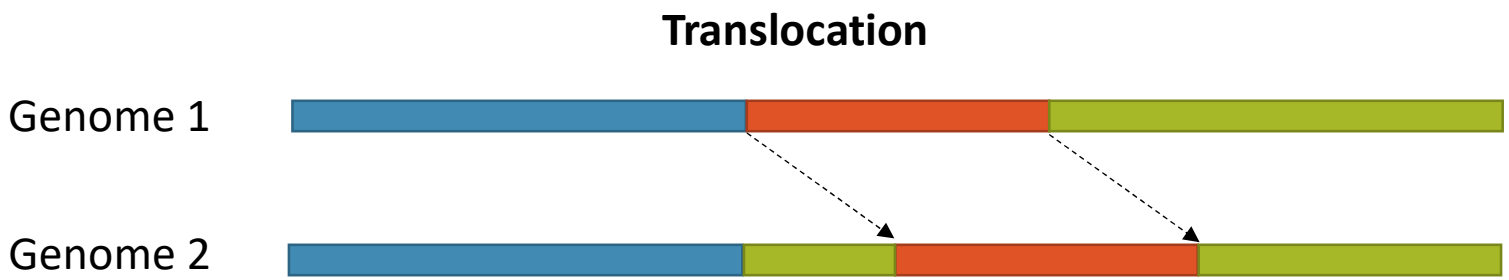


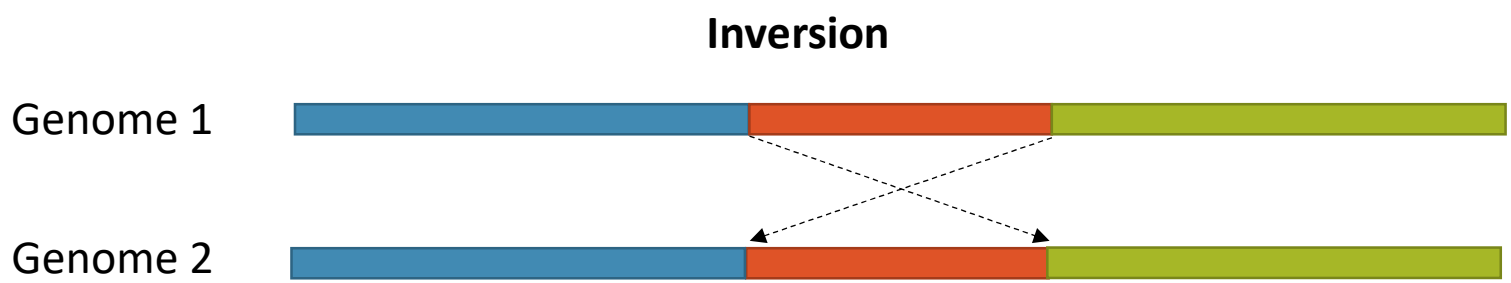
Genome evolution - structural variants

Chromosomal rearrangements

- A chromosomal region that has been removed and inserted at another place in the genome
- Can be caused by e.g. impaired DNA replication mechanisms
- Also important in evolution. Can give rise to fusion genes i.e. two different genes that have been fused together. If functional, the new product can benifitial or harmful

Genome evolution





The frequency of mutations varies

Haploid organisms (e.g. bacteria and many single cell eukaryotes)

- One copy of the chromosome: Mutations are present (100% of the DNA) or absent (0% of the DNA

Diploid organisms (e.g. humans)

- Two copies of the chromosome: Mutations are present (100%, homozygous), present in one chromosomal copy (50%, heterozygous) or absent (0%) of the DNA

The frequency of mutations varies

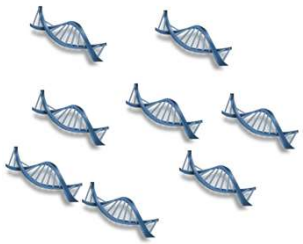
Polyploid organisms (>2 copies)

- Many copies of the chromosome: A wide range of mutation frequencies are possible (Fern has for example 630 chromosome copies!!)

Note that germline mutations are present in <u>all</u> the cells while somatic mutations are present in a subset of the cells. In a sample with many cells, somatic mutations has thus typically a lower frequency.

Genome sequencing

- Aims to characterize DNA in an organism to study the presence of mutations
- The process is reference-based where the DNA is compared against a <u>reference sequence.</u>



DNA from an organism of interest

NGS sequencer

```
ACGATAGGTACGAT
AGTACGGTACGTAC
GTGACAGGTTTTAC
CCCGGATATCCATT
AAATTTAACGAATA
CGATAAAGTAATAA
```

Reads

```
ACGATAGGTACGATGTACGTAGGATACGATAGAGCTAGTGATAGATAAG
      GTACGATGTA
          GTACGTACGA
                      ATAGAGCTAG
                          GCTAGTGATA
    ATGTACGTAC
            CGATACGATA
```

Sequence comparison (alignment)

```
ACGATAGGTACGATGTACGTAGGATACGATAGAGCTAGTGATAGATAAG
      GTACGATGTA
          GTACGTACGA
                      ATAGAGCTAG
                          GCTAGTGATA
    ATGTACGTAC
            CGATACGATA
```

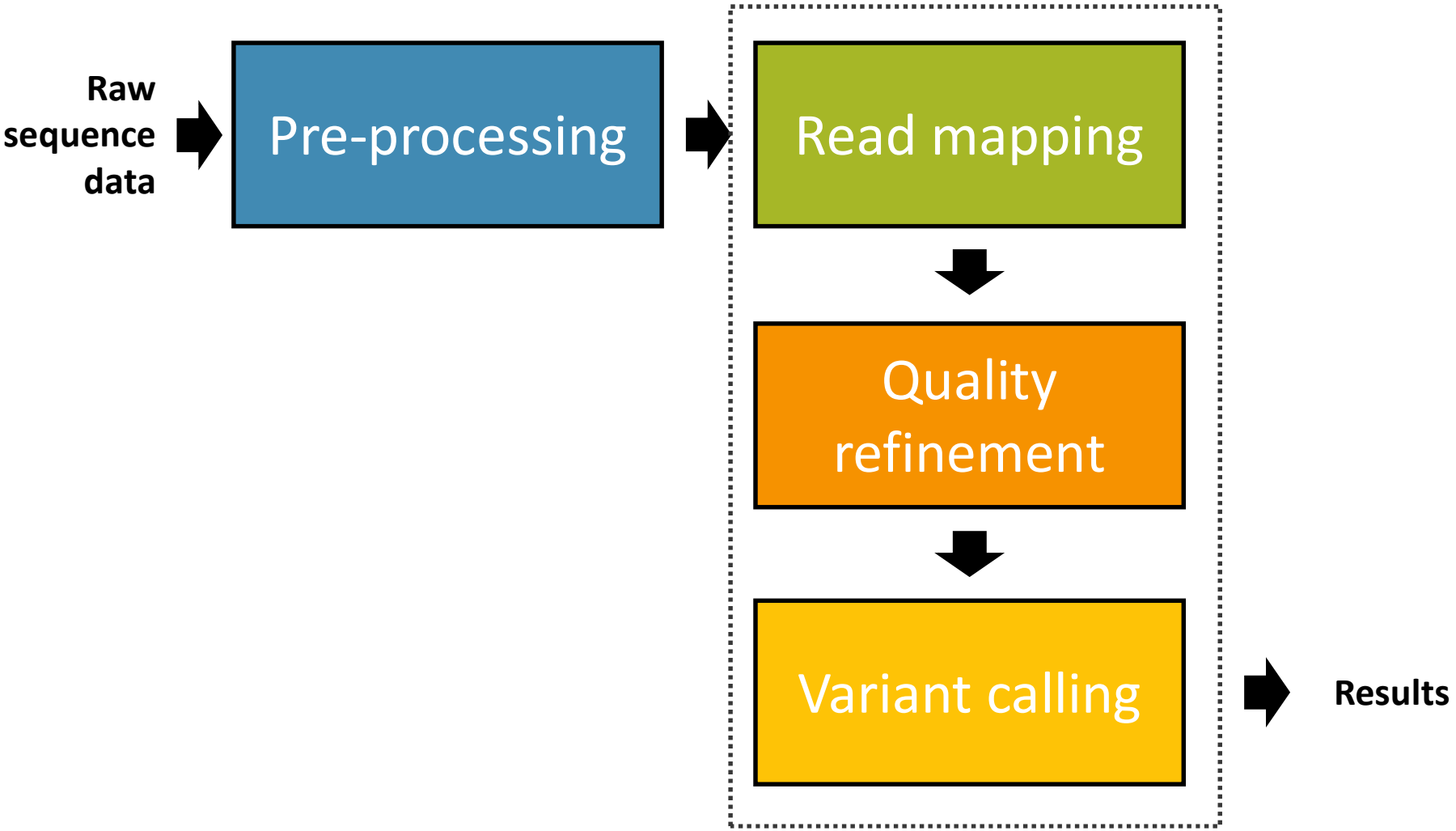
Identification of differences (variant calling)

Genome sequencing: before you start

A few important considerations

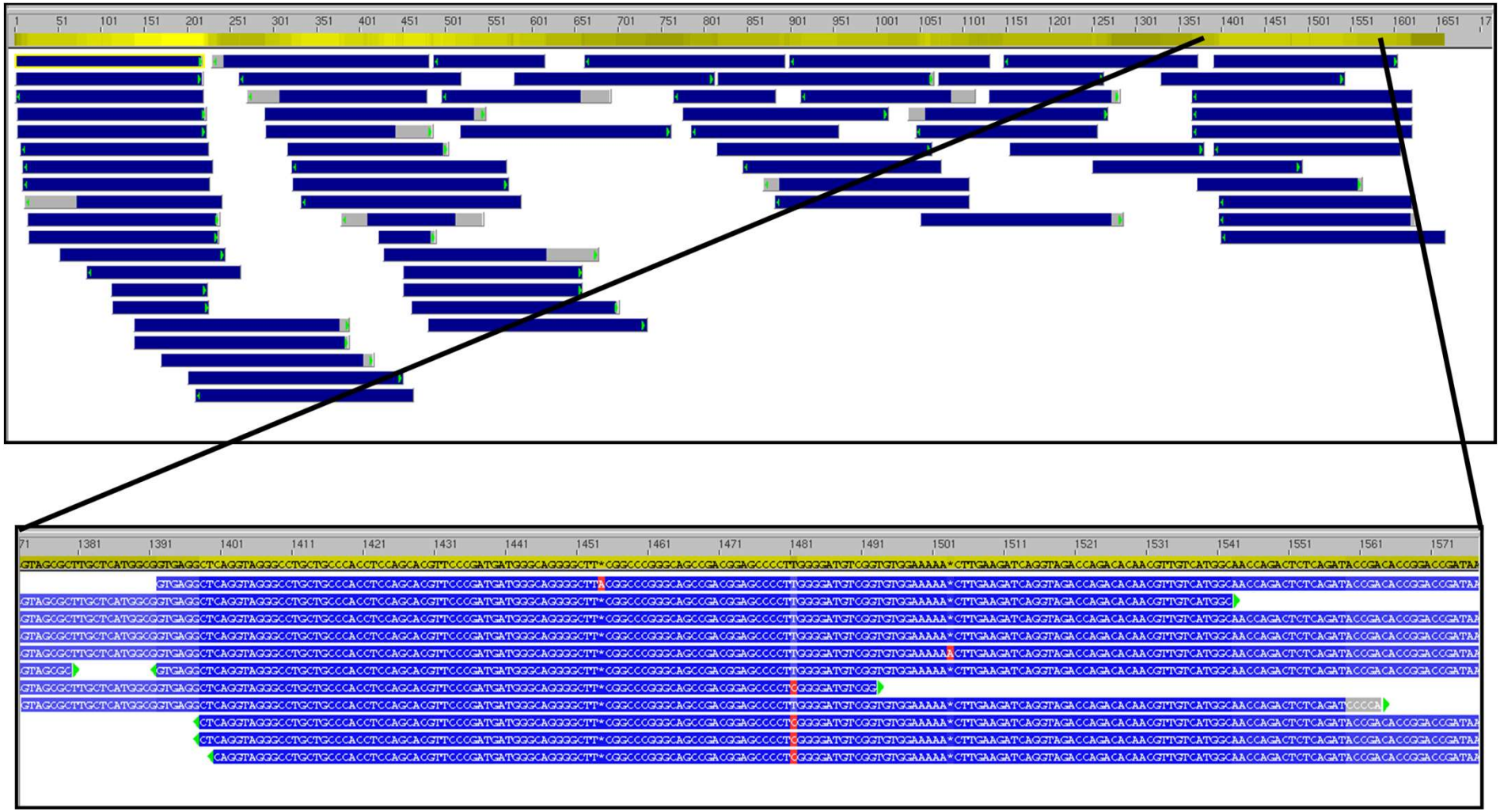
- What should we sequence? Whole genome, whole exome, set ('panel') of genes, a single gene.
- What mutations are we looking for? Germline, somatic or both?
- What sequencing technique should we use? What coverage do we need?
- What is a suitable reference?

Genome sequencing



Genome sequencing: read mapping

- Aims to identify where each sequence read match the reference genome
  - Chromosome and position
  - Strand
- Take differences between the reads and the reference into account
  - Biological differences between the sequenced genome and the reference
  - Sequencing errors in the reads
- This is the topic of the entire next week!



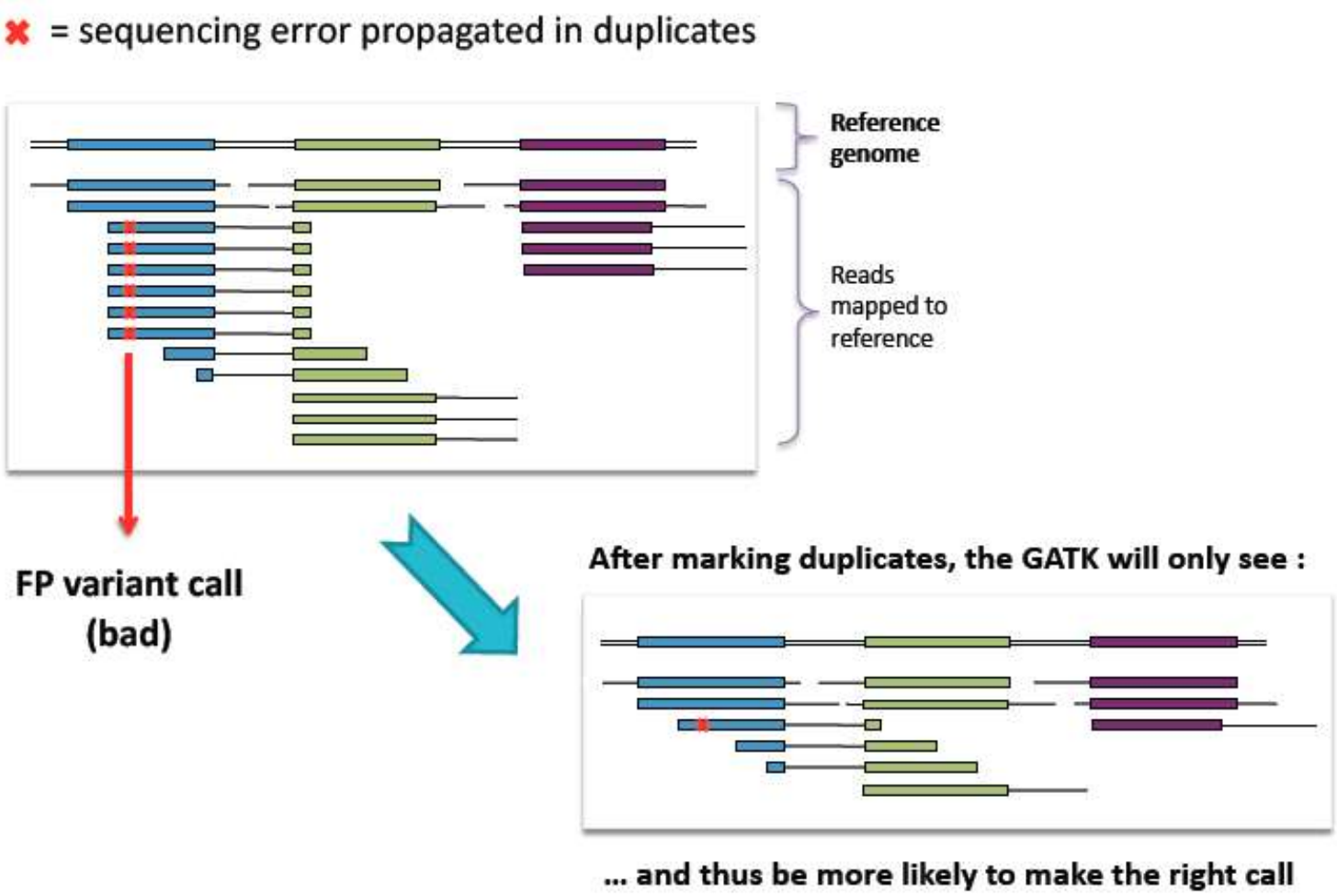
Coverage

- The <u>coverage</u> is the number of times a nucleotide position in the reference is ’covered’ by reads
- The <u>average coverage</u> is the average coverage of all nucleotide positions in the reference
- Higher coverage means that there is more information and thus higher accuracy in identifying mutations
- The coverage depends on the experimental design, the amount of sequencing data generated, quality of the sequencing data etc, but is typically in the range 15x-200x.

Genome sequencing: Quality refinement

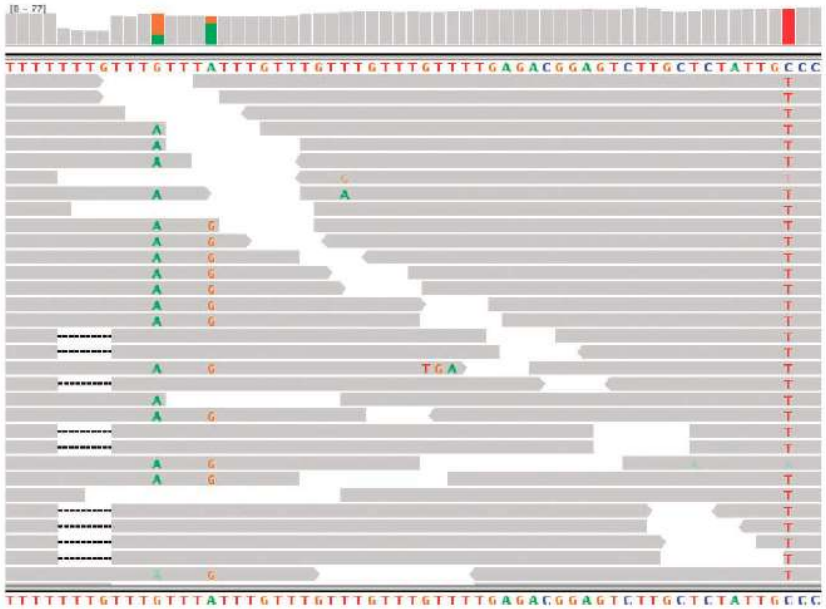
- There are several errors in NGS data that needs to be removed to reduce the number of false positives.
- Errors that are particularly problematic in genome sequencing include
  - Duplicates
  - Incorrectly aligned reads
- The quality refinement step aims to remove errors in the data and errors made in the read mapping.

Quality refinement



Figures taken from GATK presentations, https://software.broadinstitute.org/gatk/documentation/presentations.php

Quality refinement



HiSeq data, raw BWA alignments

HiSeq data, after MSA

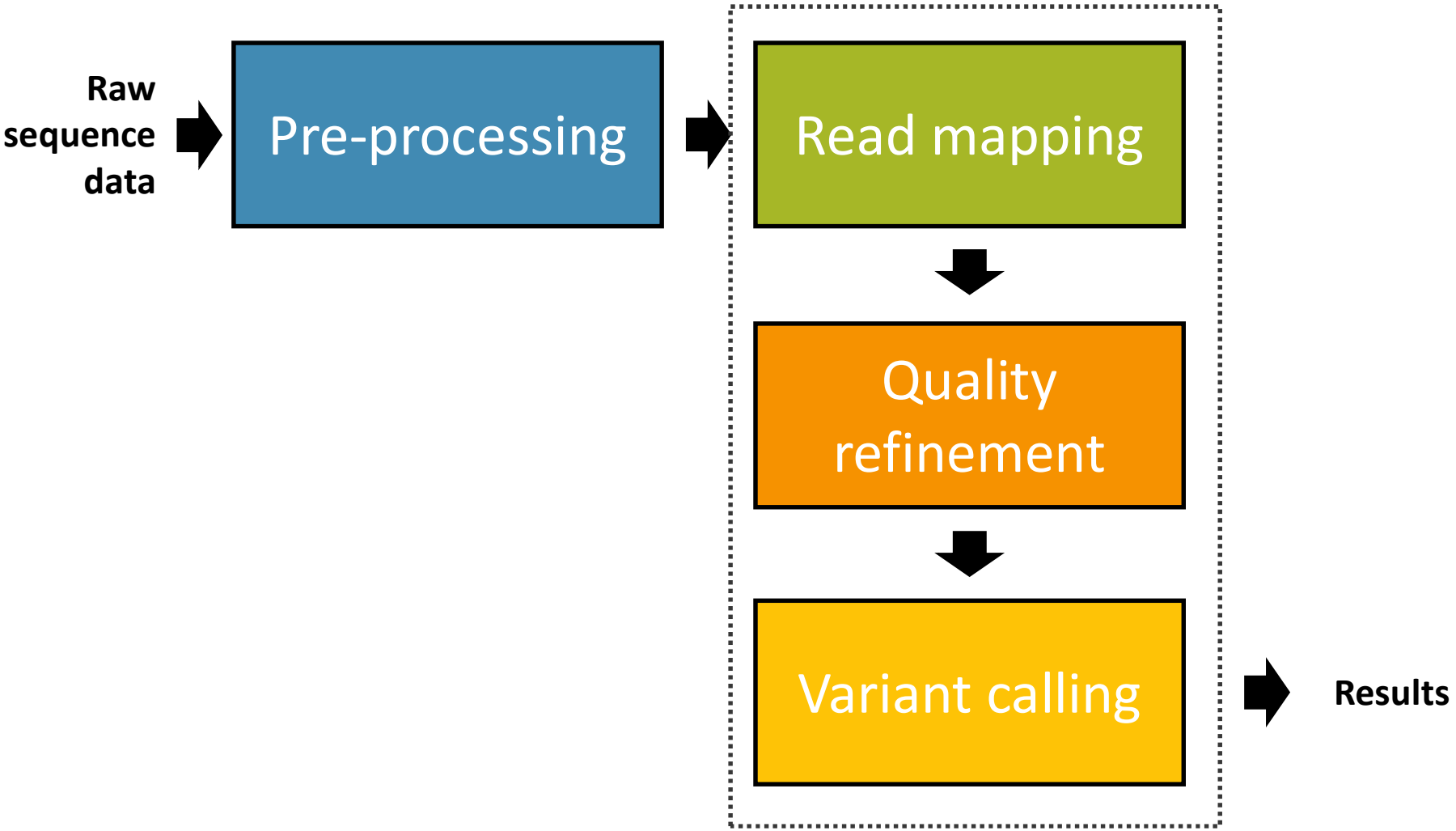
Sensitive read matching can remove ‘artificial’ mutations.

DePristo, M., Banks, E., Poplin, R. et. al, A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Gen.

Quality score recalibration

- Correct quality scores are critical for downstream analysis. Systematic biases may contribute to false results when calling variants.
- The PHRED quality score is therefore recalibrated by dividing the reads into groups based on its sample, sequencing run, sequencing lane, dinucleotide context, etc.
- The recalibrated quality score is then calculated based on the actual mismatch frequency. Common SNPs are removed in this analysis.

Genome sequencing



Variant calling of SNPs

- Variant calling aims to identify SNPs in the sequenced genome compared to the reference
- This is done by analyzing each nucleotide position in the reference and compare it to the data (the reads)
- The aim is to distinguish between true mutations and sequencing errors
- A good caller should have a sensitivity (find all true mutations) and a high specificity (ignore all false positives)

Variant calling of SNPs

Challenges in variant calling

- Sequencing error rate is high, especially in certain genomic regions
- The coverage is varying over the genome and may be low in some regions
- A large number of positions needs to be analyzed
  - Escherishia coli genome: 4.6×10^6 positions
  - Human exome (all genes): 50×10^6 positions
  - Entire human genome: 3.2×10^9

Are the changes a result of a true mutation or only the results of many reads with sequencing errors?

(This is a major problem which we will come back to later!)

A naive variant caller for SNPs

We will use a binomial test to detect <u>germline</u> SNPs in a genome. Assume that we are interested in analyzing position i and that the reference has an 'A' at this position.

```
                          position i
ACGATAGGTACGATGTACGTAGGATACGATAGAGCTAGTGATAGATA
        GTACGATGTACGTAG
            GATGTACGTGGATA
              TGTACGTGGATACG
                TACGTGGATACGAT
                 ACGTAGGATACGATA
```

Assume that the coverage at position i is N_i. Define

$$X_{i,j} = \begin{cases} 1 & \text{if read } j \text{ at position } i \text{ an not an A,} \\ 0 & \text{if read } j \text{ at position } i \text{ is an A.} \end{cases}$$

A naive variant caller for SNPs

Let

$$Y_i = \sum_{j=1}^{N_i} X_{i,j} \, .$$

Y_i is the total number of reads that does <u>not match</u> the reference at position i.

If the reads and their errors are independent, it follows that

$$Y_i \sim \mathrm{Bin}(N_i, p_i)$$

where p_i is the probability of observing another base than the reference (i.e. not an "A") at position i.

A naive variant caller for SNPs

We can assess if there is a mutation at position i by using a statistical tests:

$$H_0: p_i = p_{\text{error}}$$
$$H_1: p_i > p_{\text{error}}$$

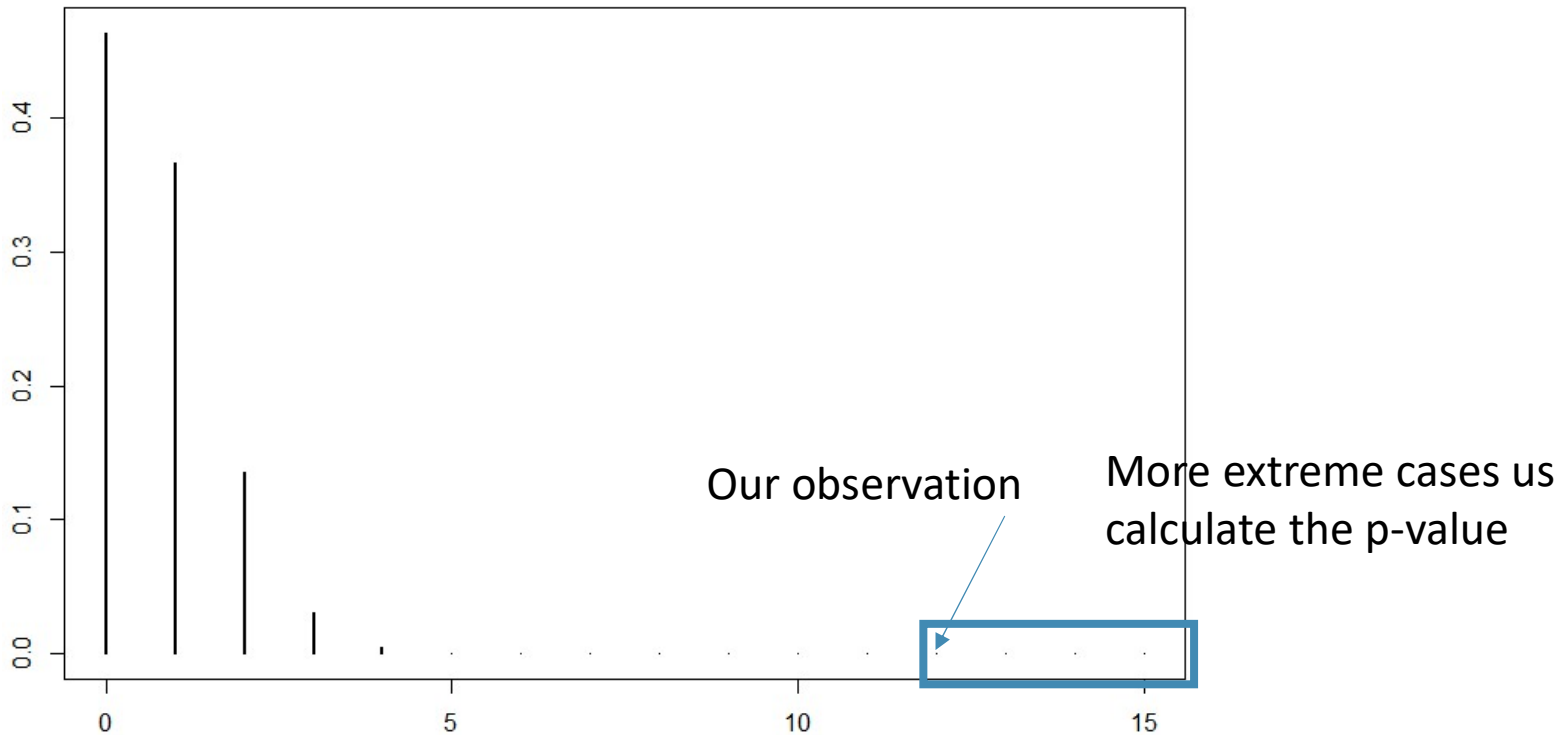
Here, p_{error} is the probability of a sequencing error, which is the lowest value p_{i} can take (if there is no mutation we only observe sequencing errors).

Under the null hypothesis, a p-value for the test can be calculated by

$$\text{p}-\text{value} = \text{Prob}(Y_i \geq y_i) = \sum_{j=y_i}^{N_i} \text{Prob}(Y_i = j)$$

Example

Assume that the sequencing error for a specific position is $p_{\text{error}} = 0.05$ (5%) and the coverage is 15. If 3 reads are matching the reference while reads 12 have a mutation (say a ”C” instead of an ”A”).



$$\text{pvalue} = 9.6 \times 10^{-14}$$

Why is this caller naïve?

- The assumptions about the errors are not true
  - The error rate differs considerably between positions
  - The errors are not independent – they depend on the context
- We do not use any information about the quality of the sequenced nucleotides
- The caller does not distinguish between genotypes which will make it inefficient in diploid genomes

The GATK unified genotyper

GATK (Genome analysis toolkit) contain the unified genotyper, which is a more advanced mutation caller.

The unified genotyper calculates

$$\mathrm{Prob}(G_i|D_i)$$

for each position i. Here G_i set to any of the possible genotypes (AA, AC, AG, AT, CC, CG, CT, GG, GT and TT for a diploid organism). D_i denotes the data available at the position i.

The GATK unified genotyper

Model

Using Bayes theorem, this can be rewritten as

$$\mathrm{Prob}(G_i|D_i) = \frac{\boxed{\mathrm{Prob}(D_i|G_i)} \times \boxed{\mathrm{Prob}(G_i)}}{\mathrm{Prob}(D_i)}$$

Prior knowledge

where

$$\mathrm{Prob}(D_i) = \sum_i \mathrm{Prob}(D|G_i) \times \mathrm{Prob}(G_i)\,.$$

The sum is calculated over all 10 genotypes.

The GATK unified genotyper

The unified genotyper assumes that

$$\mathrm{Prob}(D_i|\, G) = \prod_{\substack{r \in \{\text{good read} \\ \text{bases at } i\}}} \frac{\mathrm{Prob}(r|H_1)}{2} + \frac{\mathrm{Prob}(r|H_2)}{2}$$

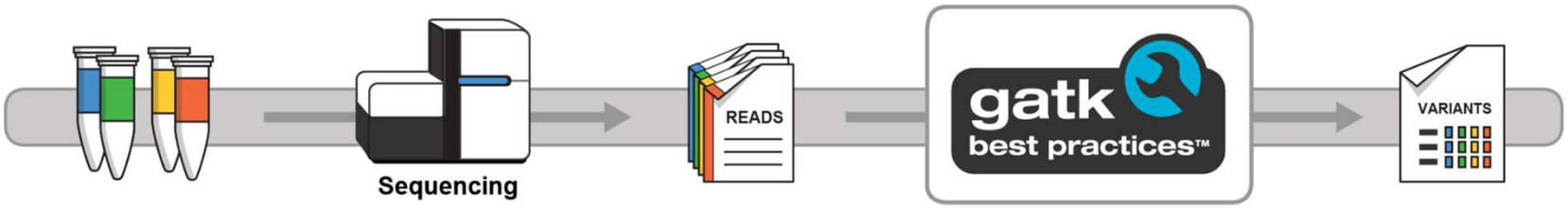
where $G = H_1 H_2$ and, with ε_i=error probability of position i,

$$\mathrm{Prob}(r|H) = \begin{cases} 1 - \varepsilon_i & \text{if r} = H \\ \varepsilon_i & \text{if } r \neq H \end{cases}$$

ε_i is derived from a site-specific error model.

The GATK unified genotyper

- The unified genotyper has several advantages compared to the naive caller.
  - Only reads of sufficient quality are included
  - A more sophisticated error model is used. The quality values are taken into account,.
  - The probability of each possible genotype is estimated
- GATK also offers information on 'best practices' in mutation calling

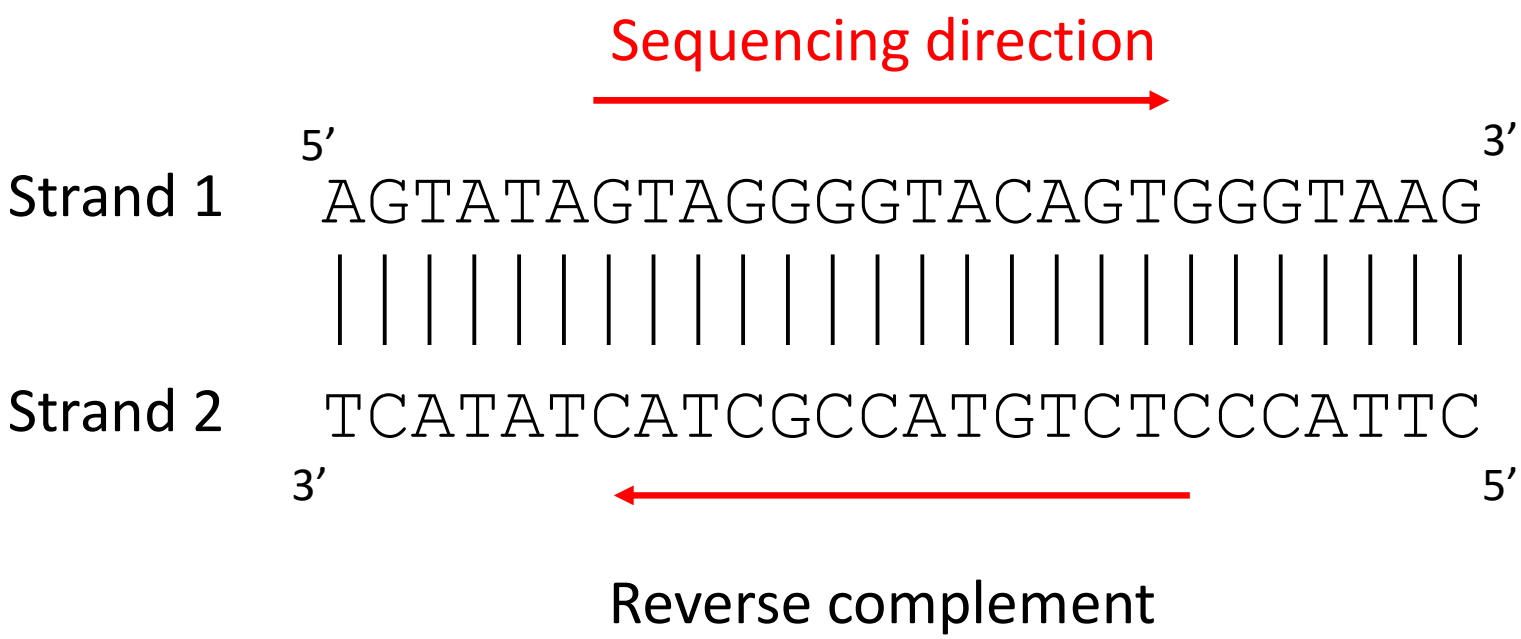


Available at https://software.broadinstitute.org/gatk/

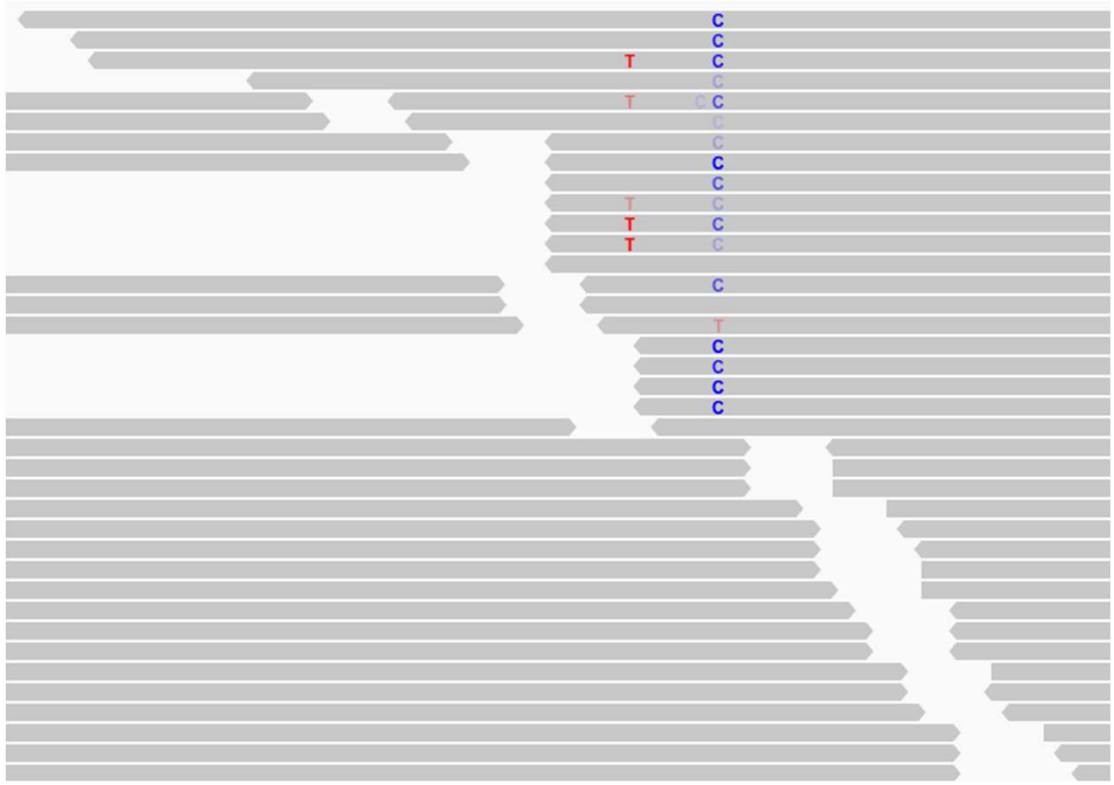
Post-processing – filtering of variants

All the mutations are in the end of the reads.

Different strand have different error probabilities



Post-processing – filtering of variants



All the mutations on one single strand.

Post-processing – filtering of variants

- Many genome sequencing experiments results in a very long list of variants that may need to be filtered before it can be interpreted.
- Critera for filtering typically includes
  - Strand bias, i.e. variant bases only in one read direction
  - Clustered position, e.g. variant bases always at the end of reads
  - Poor mapping, i.e. variant bases only in reads with lower mapper quality

Post-processing – filtering of variants

- It is also possible to filter variants based on their <u>biological function and impact</u>.
- Such criteria include
  - Synonymous/non-synonymous
  - Commonness among healthy individuals in the populations
  - Its predicted biochemical impact: is it located in a protein domain that is likely to have a impact on the protein function
  - Previous knowledge, e.g. its association to other diseases

Databases with information about mutations

dbSNP, https://www.ncbi.nlm.nih.gov/snp

- A collection of human single nucleotide polymorphisms
- Contains information about their frequency in the human population

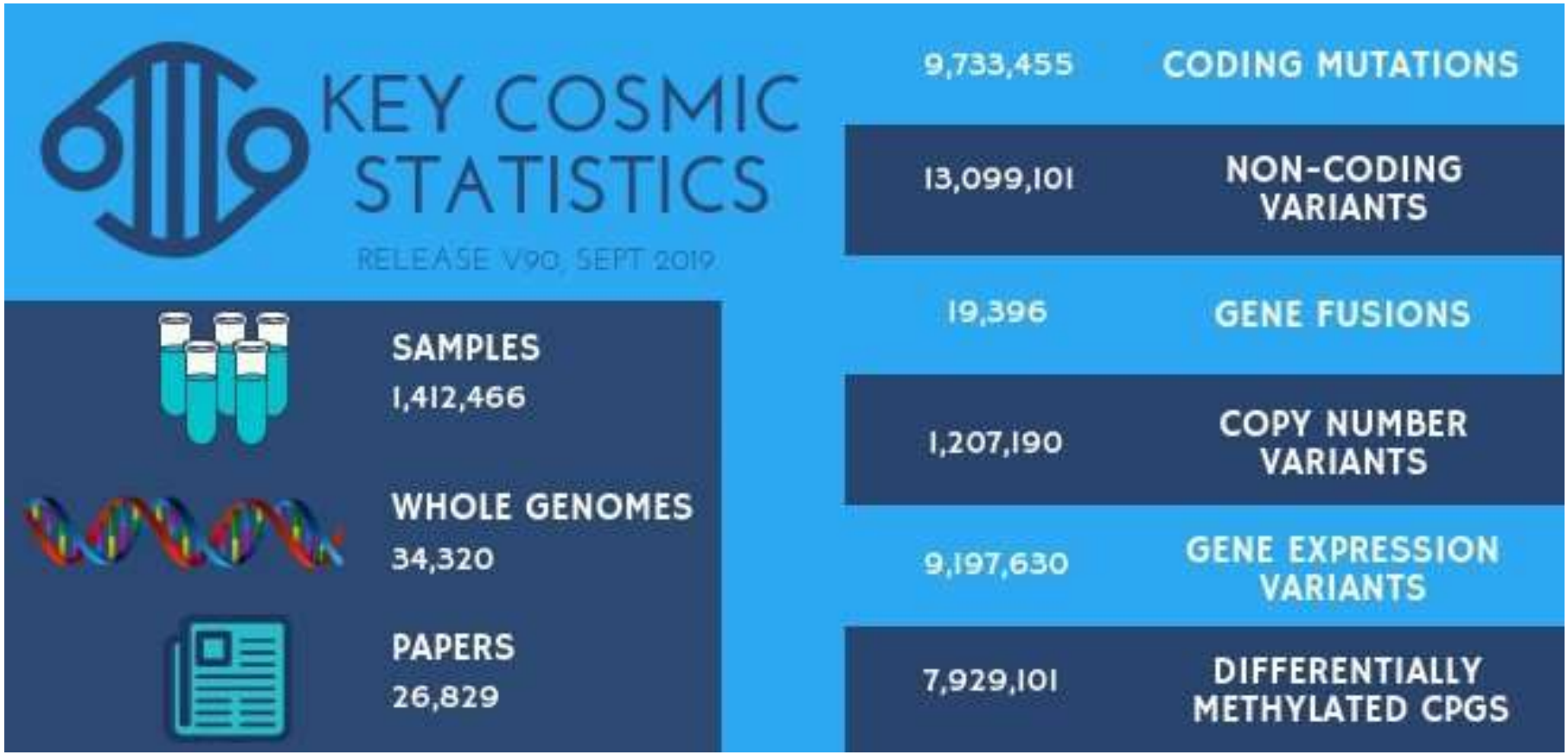
OMIM, https://www.ncbi.nlm.nih.gov/omim

- Database with gene variations associated with human diseases

COSMIC, https://cancer.sanger.ac.uk/cosmic

- A large catalogue of somatic variation related to cancer.

Databases with information about mutations

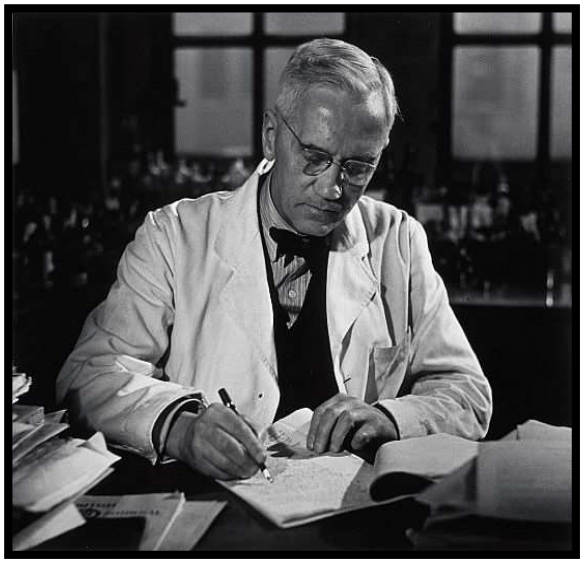


Statistics for the latest version of COSMIC released September 5 2019.

Calling of other forms of mutations

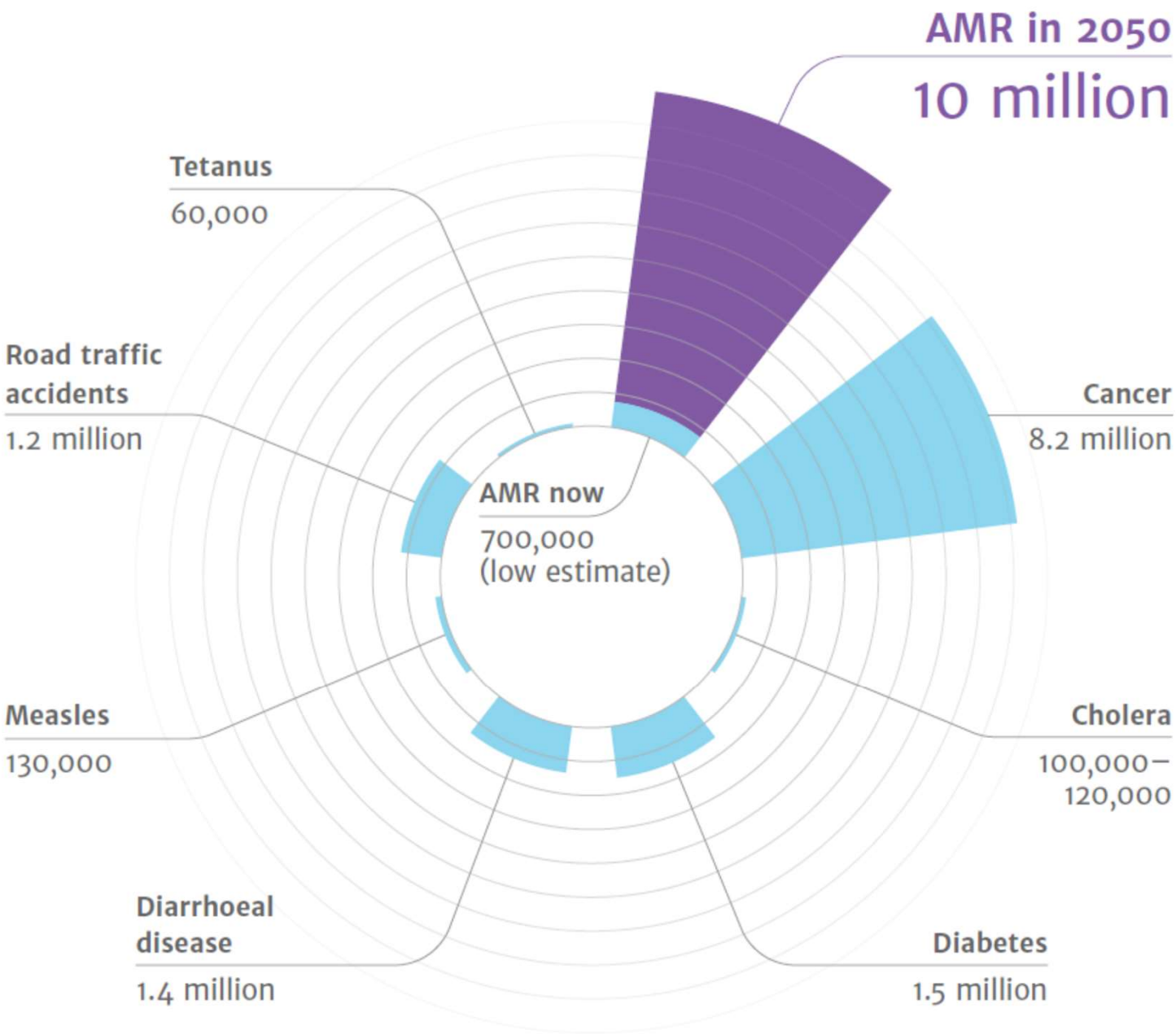
- Calling of other forms of mutations, e.g. indels and other structural variants, is important but not covered by this course.
- For information you can read Wang chapter 9.2 and 9.3

Introduction to computer exercise 2



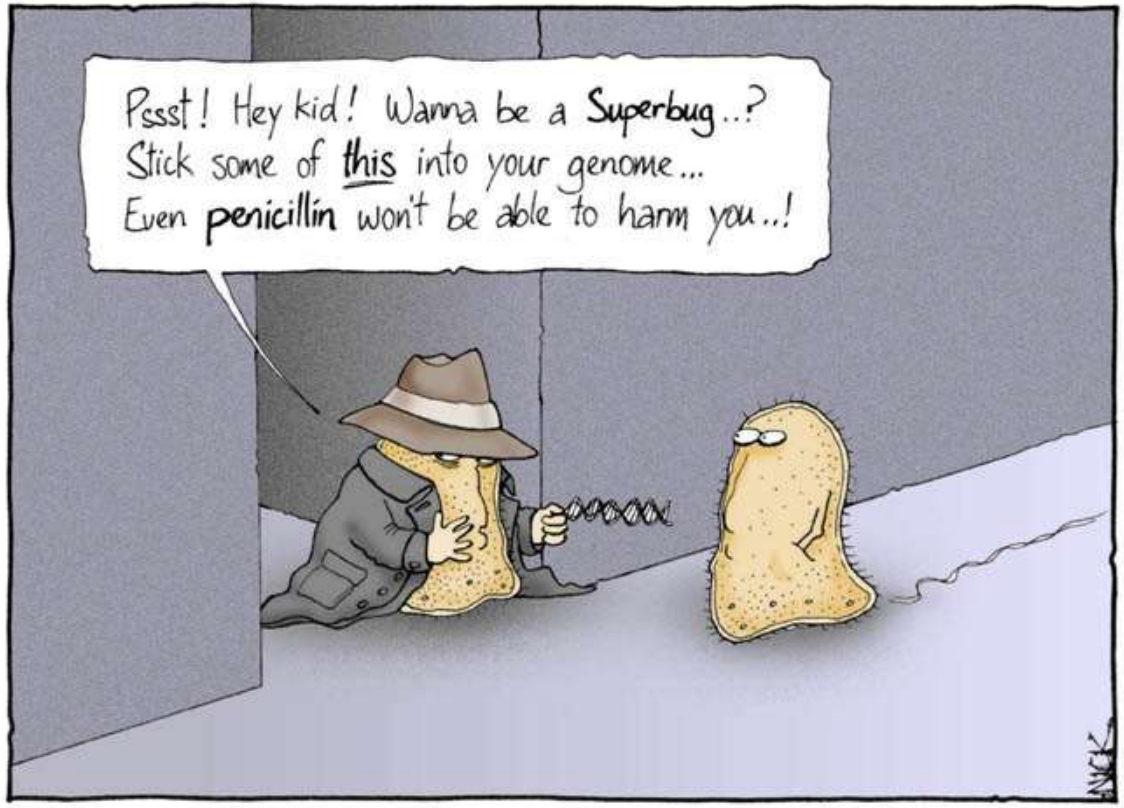
Alexander Fleming

Penicillin-producing fungi

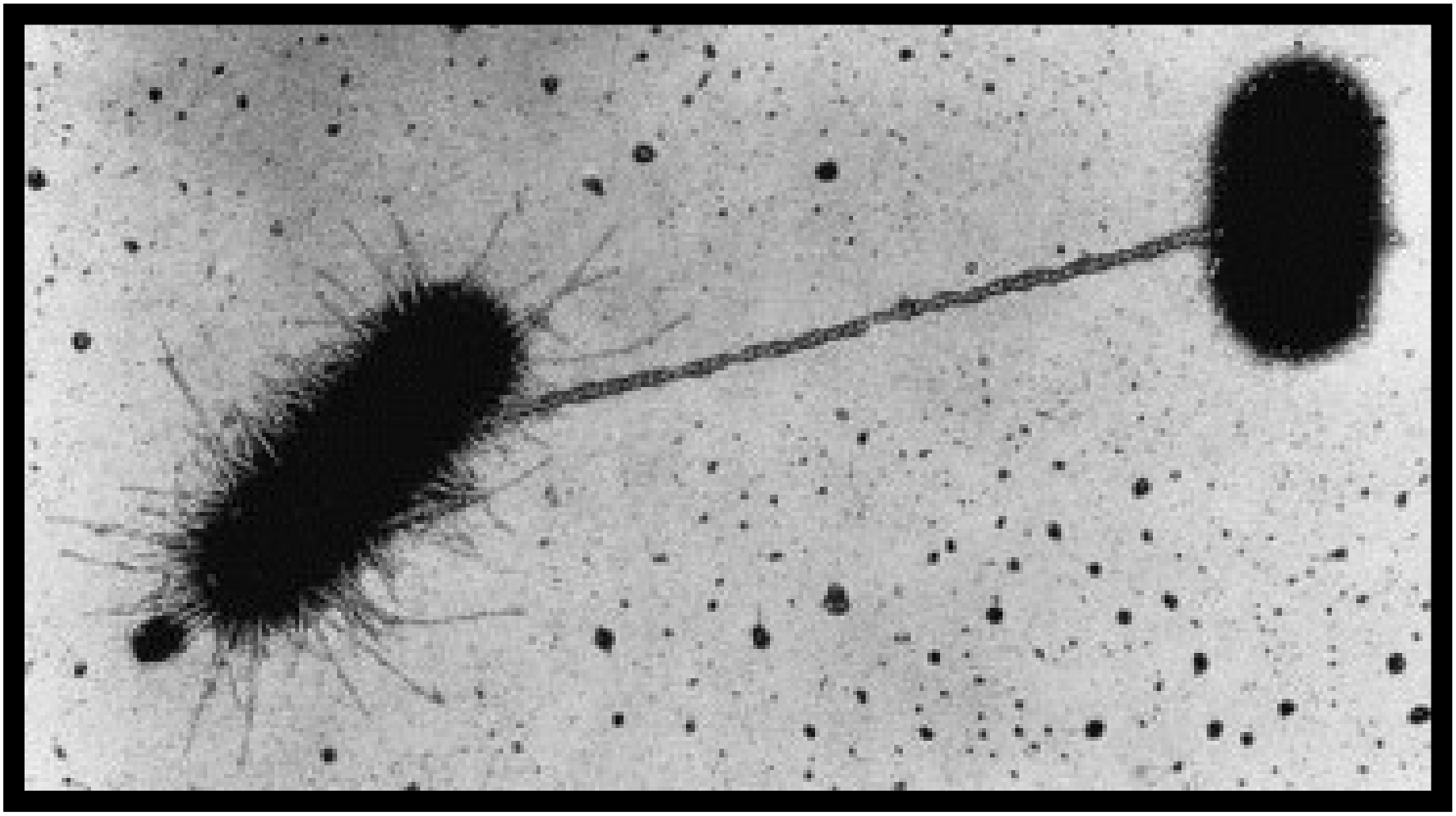

AMR in 2050
10 million
Tetanus
60,000
Road traffic accidents
1.2 million
Cancer
8.2 million
AMR now
700,000
(low estimate)
Measles
130,000
Cholera
100,000–120,000
Diarrhoeal disease
1.4 million
Diabetes
1.5 million

Antibiotic resistance is caused by

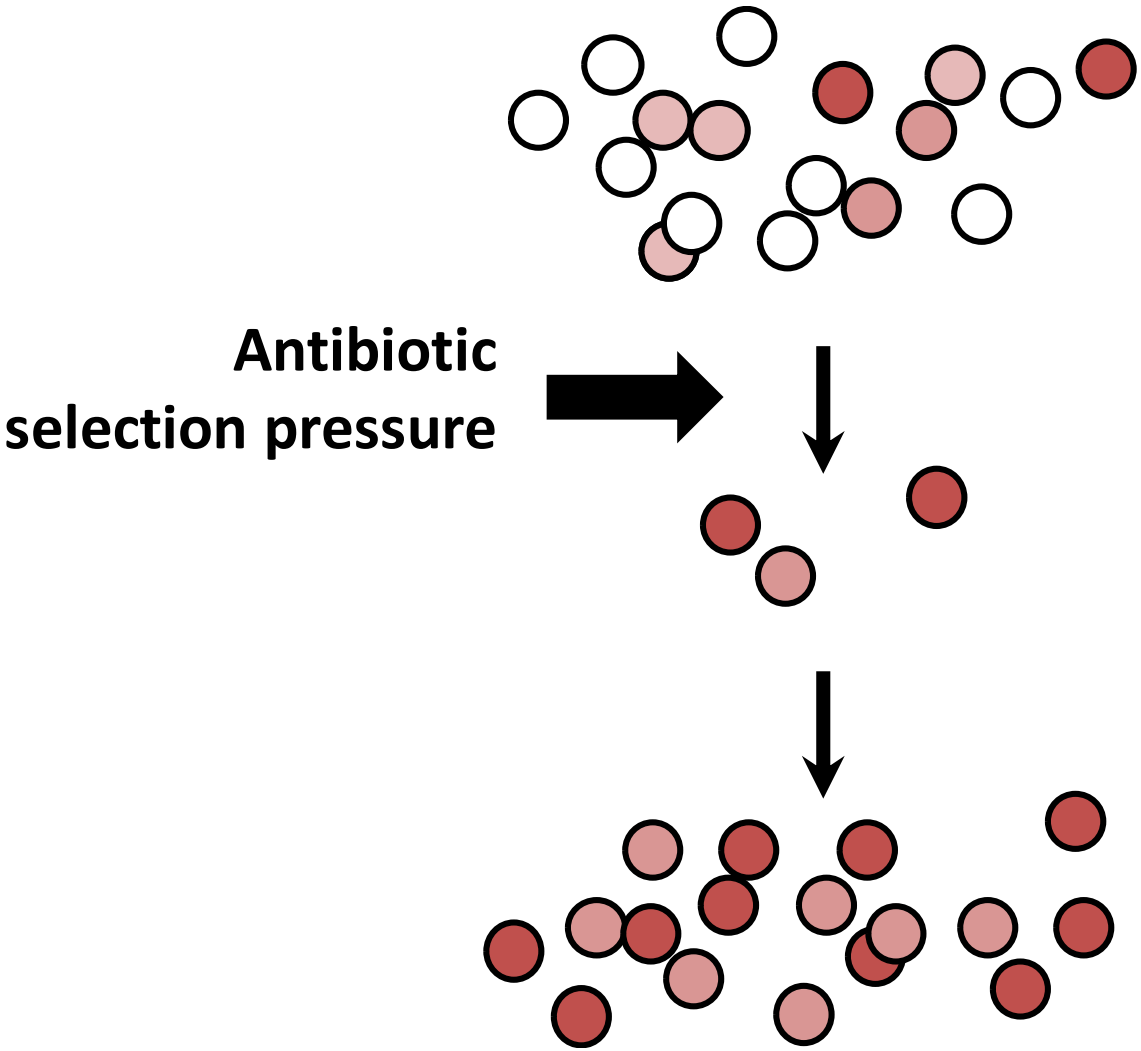
1. Mutations in pre-existing DNA
2. Acquisition of resistance genes



It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.



Selection of antibiotic resistant bacteria



Introduction to computer exercise 2

In this exercise, you are given three bacterial genomes of *Escherichia coli*.

Main questions: Are any of the genomes from a resistant strain?

Introduction to computer exercise 2

Objectives of computer exercise 2

- Provide a first view of running bioinformatics command in Linux environment.
- Analysis of genome sequencing data in R
  - Calculation of summary statistics
  - Implementation of a naïve caller based on the binomial test
- Biological interpretation of the identified mutations

Summary of todays lecture

- Organisms evolve through changes in their genomes. These changes can be single nucleotide polymorphisms (SNPs), indels, and larger structural variants
- Genome sequencings aim to identify mutations in relation to a reference
- Analysis of genome sequence data are dependent on three main steps: read mapping, quality refinement and SNP calling
- SNP calling is done by statistical modelling of the read data and its errors

Summary of todays lecture

- SNP calling results in a huge data reduction. From billions of observations to, in some case, a handful of significant positions. Careful filtering to remove errors in the different steps is therefore vital.