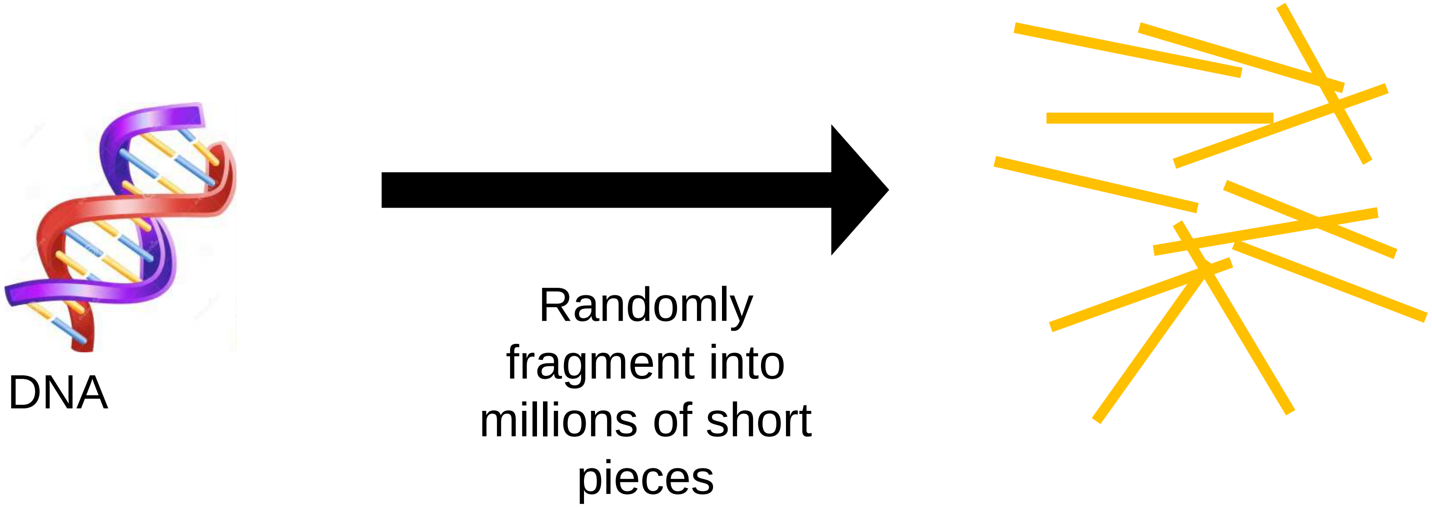
Overview of variant analysis

MSc in Genomic Medicine

Lucy Crooks

25/1/2016

Principle of next generation sequencing

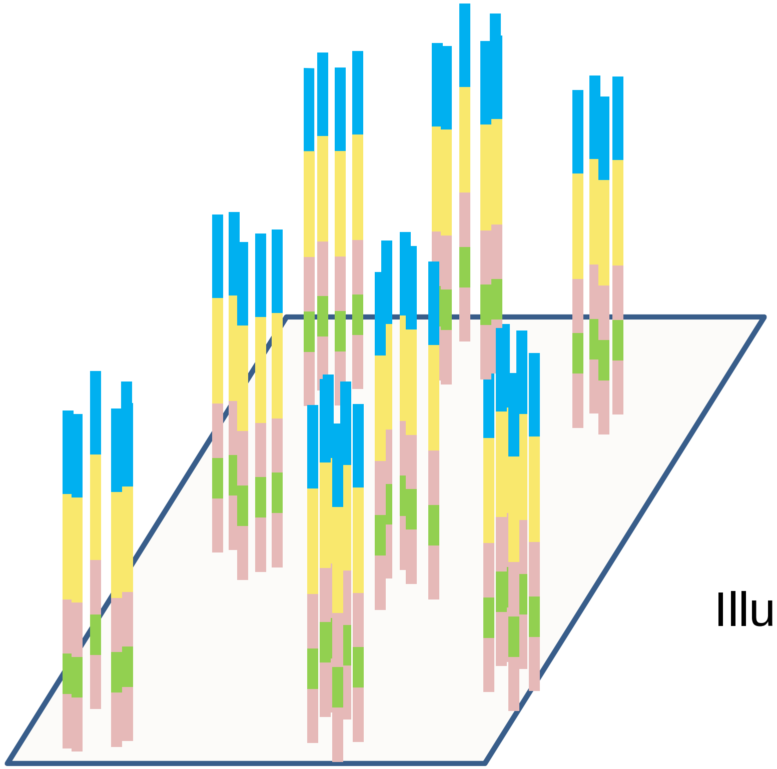


DNA

Randomly fragment into millions of short pieces

Prepared fragmented DNA = library

- Attach library to substrate so fragments can be distinguished
- Sequence massively in parallel



Parallelisation greatly reduces cost

Illumina clusters on a flow cell

- Output from the sequencer is reads

- A read is the set of the DNA bases in order from a fragment

- Each read is small ~100 bases and can have a mistake

- Method works because we combine information from many reads all starting and ending in different places

How do we get from reads to identifying the genetic change that has caused a patient's disease?

Variant Analysis

Steps in variant analysis

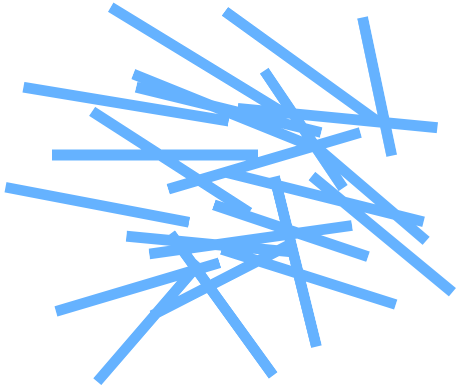
Alignment

Variant calling

Quality filtering

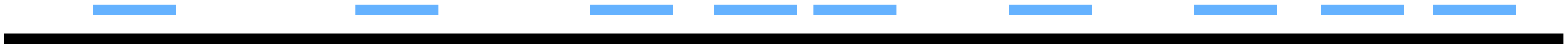
Identify key variant

1. Alignment



Millions of jumbled-up reads

Find where on the genome they came from



Reads are 100 bases long

Genome is 3,000,000,000 bases long!

- Aligning to a reference is much easier than *de novo* genome assembly

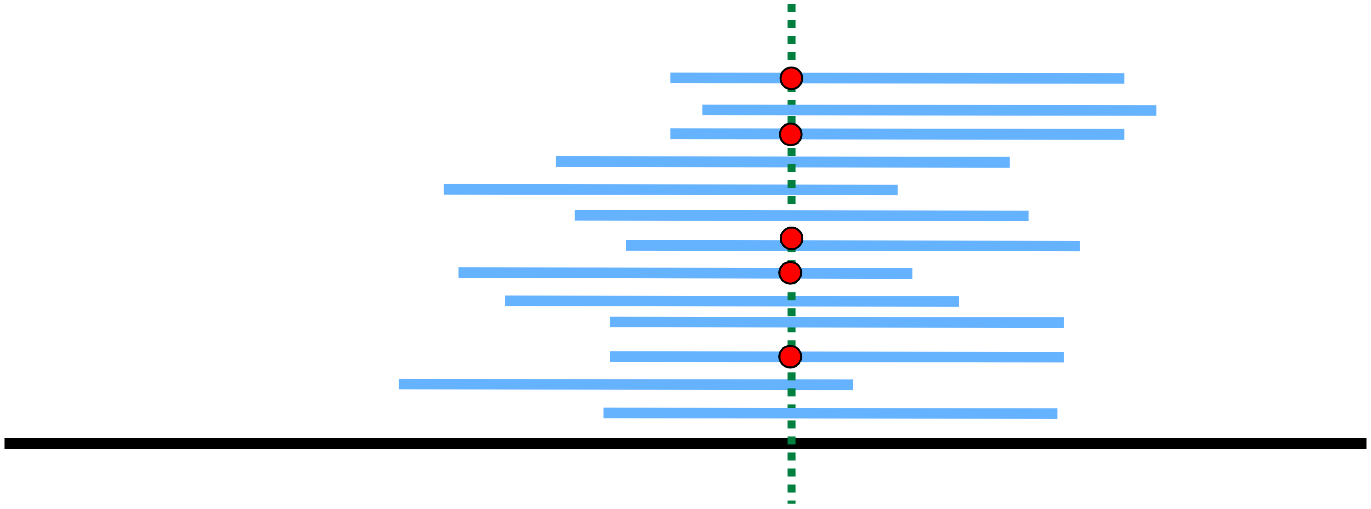
- The original human genome reference was completed in 2003, taking 13 years

- It is a mixture from several individuals

- It is not a consensus

2. Variant calling

Look for differences from the reference at each position



Depth/coverage = how many reads map are over the position

Types of Variant

- SNV – single nucleotide variation
- InDel – small insertion or deletion
- SV – structural variant (longer insertion, deletion or rearrangement, also called CNV – copy number variant)

Also want to know if the individual is homozygous or heterozygous for a variant

3. Quality filtering

- Many quality scores are generated that can be used to filter variants

- Unclear which are most useful and how they relate to each other

- Best is to have ‘truth set’ to test filtering strategy

- Have to chose where to balance missing true variants (sensitivity) against calling a variant by mistake (specificity)

4. Identify key variant

~ 4 million variants per person

- Restrict by gene

Diagnostic genetics
Small set of genes connected to specific disease

Research
Exploratory gene list based on biological pathways or gene expression data

- Test for association in large population of cases and controls

4. Identify key variant

- Has variant been published for a same or similar condition?

- Is it reported above low frequency in healthy populations?

- Where does it occur in relation to parts of genes?
  - Should it affect protein sequence or transcription?

- What are the predicted functional consequences?

4. Identify key variant

Aspects of this process are referred to as

- Variant annotation
- Variant interpretation
- Variant prioritisation

Caveats of NGS data

- Some regions of genome do not sequence well

  GC rich regions are problematic for PCR

  If there is no coverage you cannot see variants

- Short reads are not effective in repetitive regions and when there are gene copies

- Need bioinformatics skills!

- High requirements for computer processing and storage

  Data from 1 HiSeq run equivalent to 48 HD movies