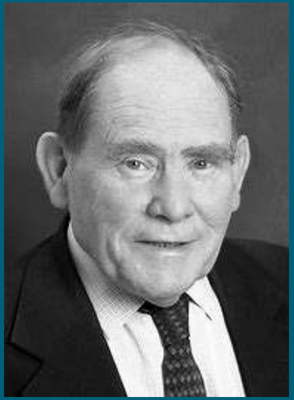
Human genome, genetic variation, genomic technologies

Dr Bilal Alobaidi
Biosciences Institute
Faculty of Medical Sciences

Newcastle University

“Progress in science depends on new techniques, new discoveries, and new ideas, probably in that order.”

- Sydney Brenner, 2002 Nobel Prize Winner



Disease

Major challenges in medical genetics

- Identifying genetic variation
- Interpreting genetic variation

In research: Large, longterm studies to identify genetic variation that increases/decreases risk of disease, and functional studies to confirm pathogenicity & unravel the mechanism

In clinic: Diagnosis in individual patients, who want reliable and useful answers fast and affordable....

1989
Genetics is an art

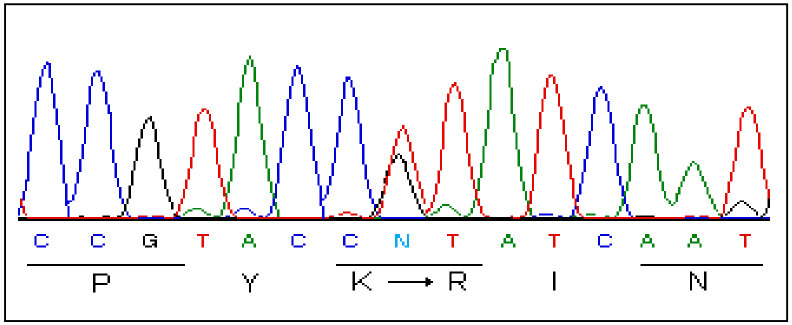
2019
Genetics is an industry

Our genome: Full of variation

- 6 billion nucleotides per genome
- 2 people vary at 4 million positions
- 1 variation (mutation) can cause a rare disease

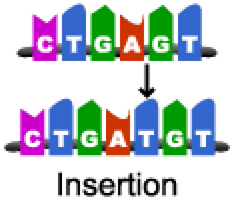
Full of variation indeed!

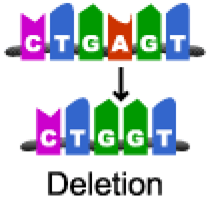
Single nucleotide variation



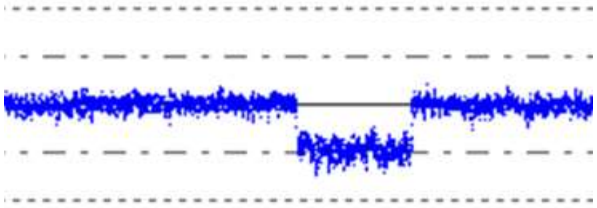
Additional chromosome

Insertion-deletions

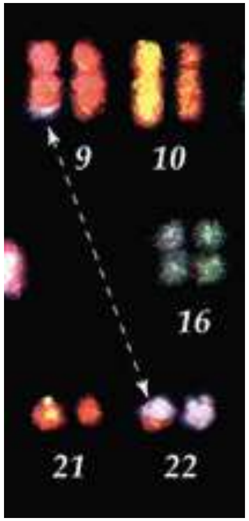




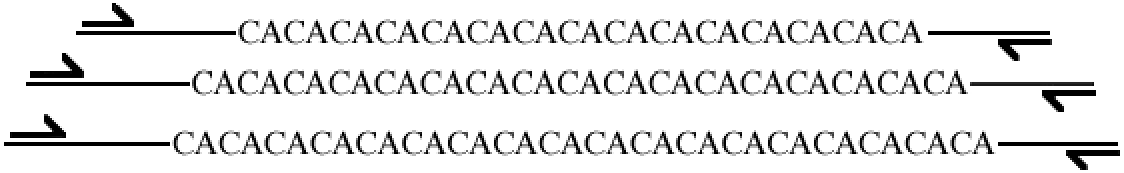
Deletions of DNA sequence



Chromosome rearrangements



Microsatellite repeat variations

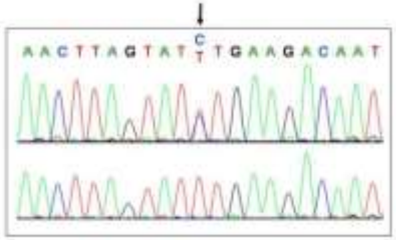

CACACACACACACACACACACACACACACACA
CACACACACACACACACACACACACACACACACACA
CACACACACACACACACACACACACACACACACACACA

Variation per genome

- SNVs (single nucleotide variants):
  - ~ **3-3.5 Million SNVs**
    - **of which vast majority SNPs – single nucleotide polymorphisms**
    - ~ 500 private/rare coding variants
    - ~50-100 *de novo* mutations per genome (0-4 coding)
- Indels (insertions/deletions)
  - **~300,000 indels**
  - Largest uncertainty, still difficult to detect
- CNVs (copy number variants)
  - **100bp-10Mb: ~1000 per genome**
  - >50kb: ~30/genome
  - *De novo*, >100kb: <1 per genome

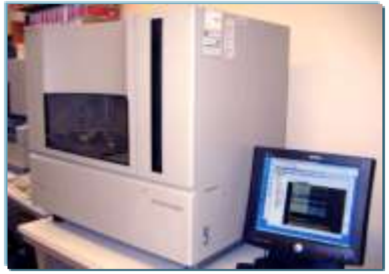
Detection of genomic variation at all resolutions From nucleotides to chromosomes!

1-500 nucleotides



DETAILED LEVEL

Sanger Sequencing



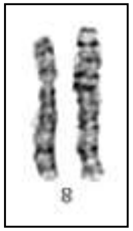
Next Generation Sequencing

BRIDGING THE GAP

TECHNOLOGY DEVELOPMENT

Whole Genome Sequencing

>5 million nucleotides

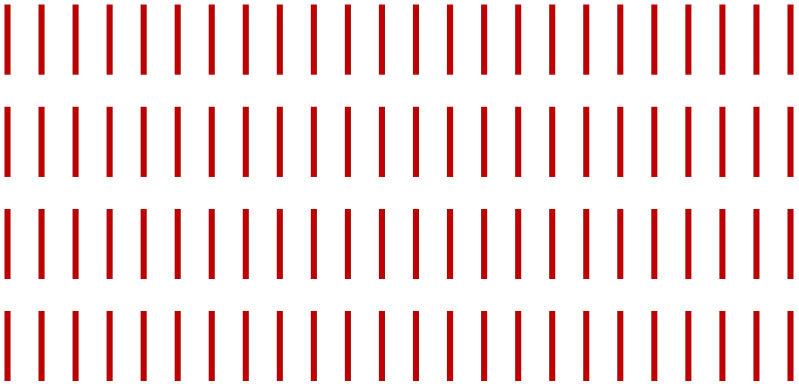


WHOLE GENOME VIEW

Karyotyping

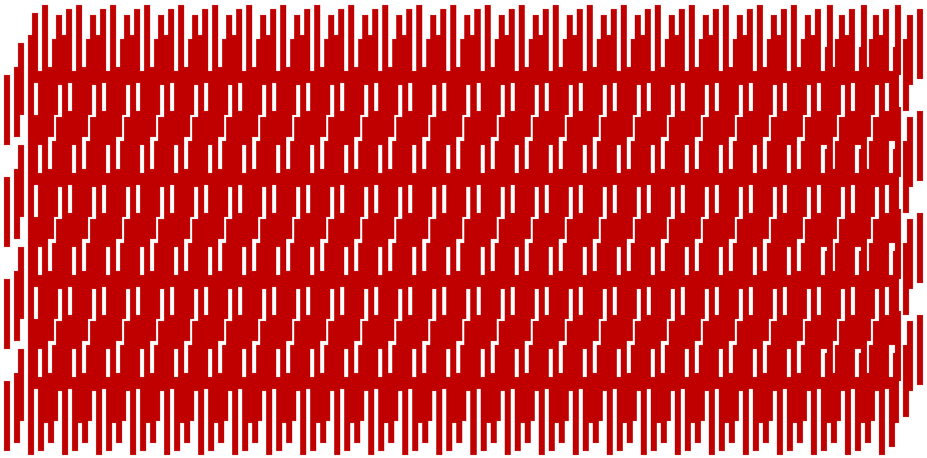
Microarrays

Traditional *vs.* Next generation sequencing



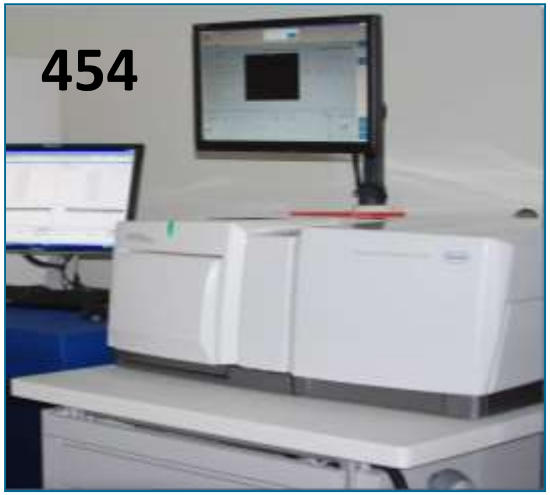
96 DNA fragments sequenced simultaneously

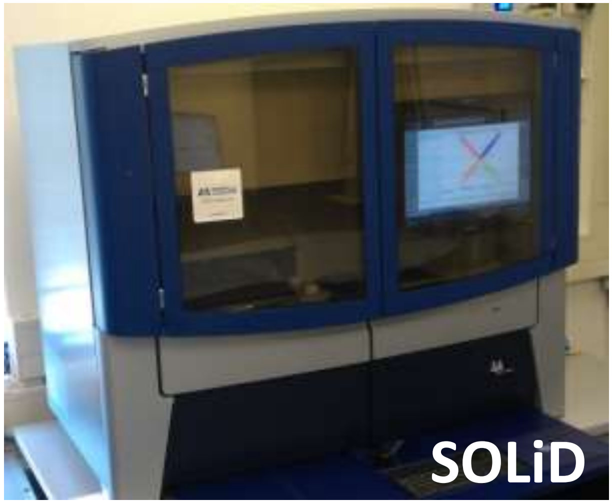
Miniaturization and parallelization



Millions of DNA fragments sequenced simultaneously

Next generation sequencing equipment


454


SOLiD

Ion Torrent

HiSeq

Pacific Biosciences

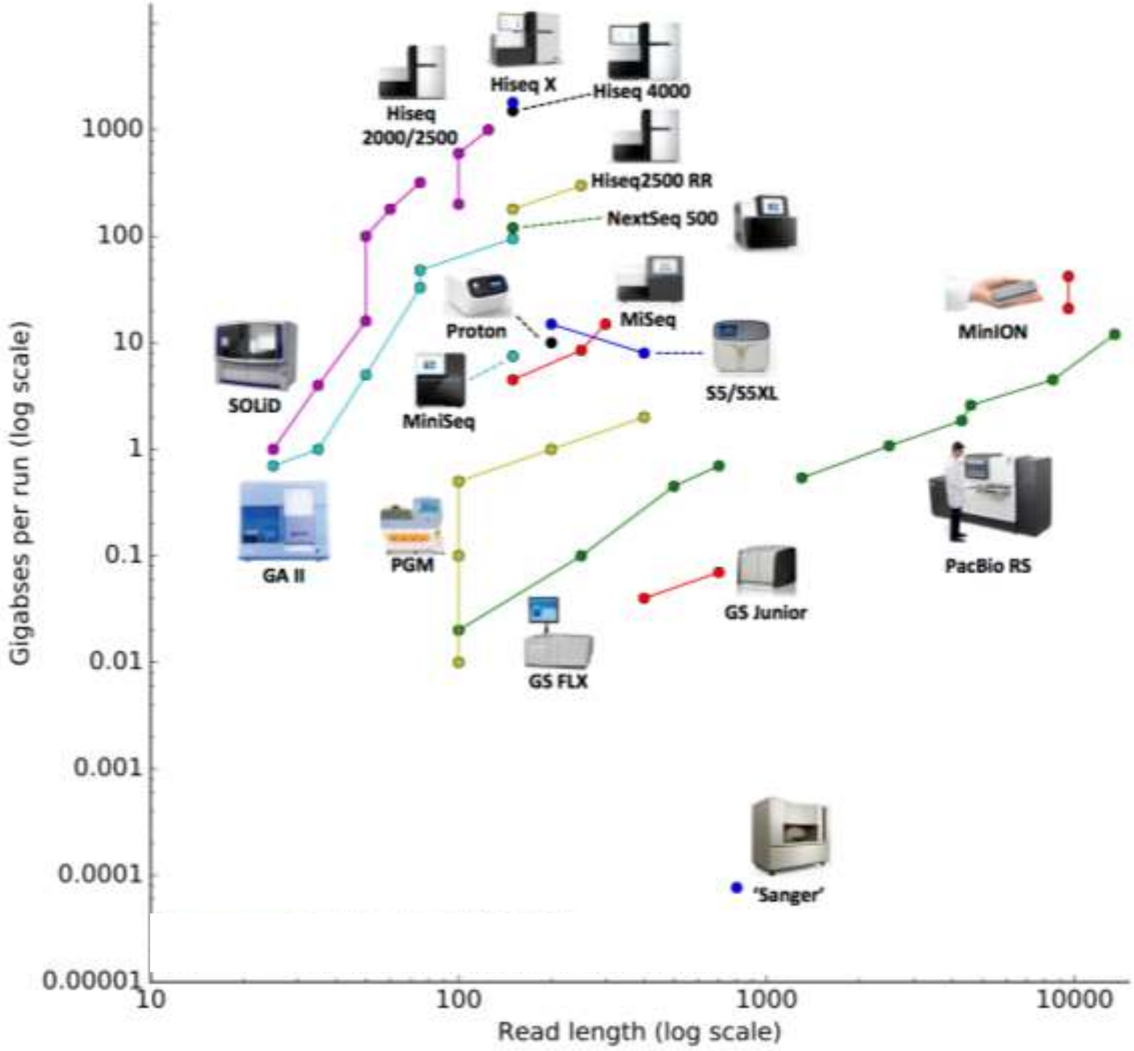
Complete Genomics

Genome technology: Big and small scale

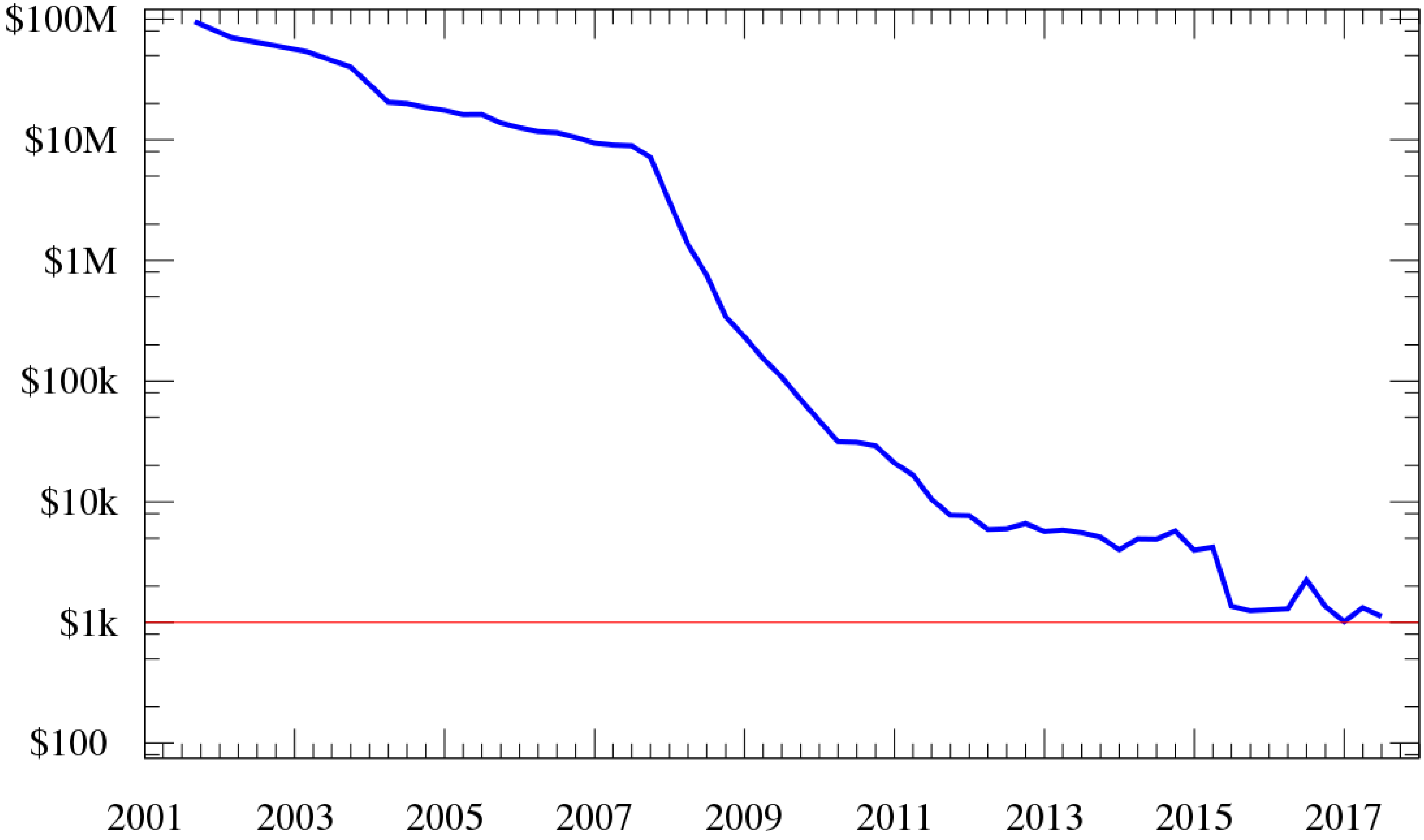
Illumina Novaseq

Oxford Nanopores MinION

Reading the DNA: Throughput and length



Cost of sequencing a human genome (with reasonable quality for variant identification)



$3,000,000,000	2003 Human Genome Project	
$20,000,000	2006 1st individual genome	
$2,000,000	2007 1st NGS Genome	
$200,000	2008 1st 30x genome	
$10,000	2010 1st sub-10K genome	
$1,000	2014 1st $1,000 genome	
$100	2017 1st $100 genome	

Technology choice based on application

- All types of genetic variation or only single nucleotide variants?
- Gene panel, all genes (the exome) or the entire genome?
- Discovery science or diagnostics?
- Germline variant detection or also somatic (and in what tissue)?

And on finances, expertise, bioinformatics capacity, turn-around-time, etc.…

Next generation sequencing workflow

Sequencing

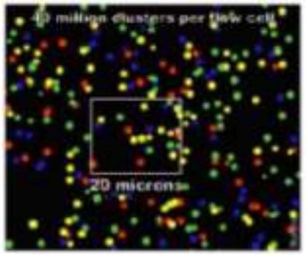
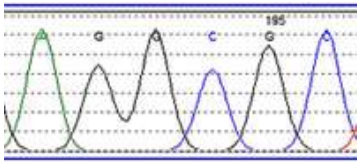


Image processing

Reads generation

Read alignment

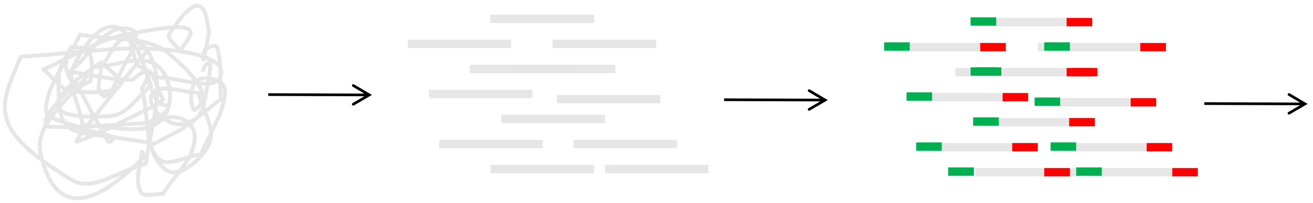


Variant Detection

Variant prioritization

Next generation sequencing basics

Library preparation



Genomic DNA

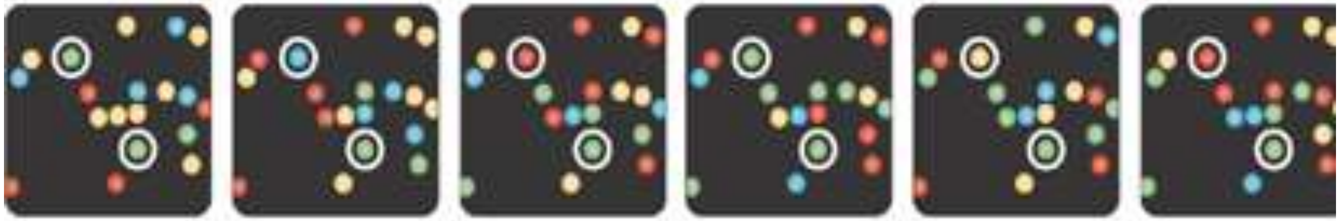
Fragment DNA

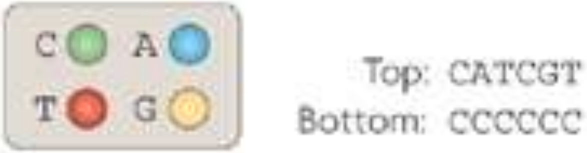
Adaptor ligation

Bridging PCR

Deposit molecules and amplify

Sequencing





Top: CATCGT
Bottom: CCCCCC

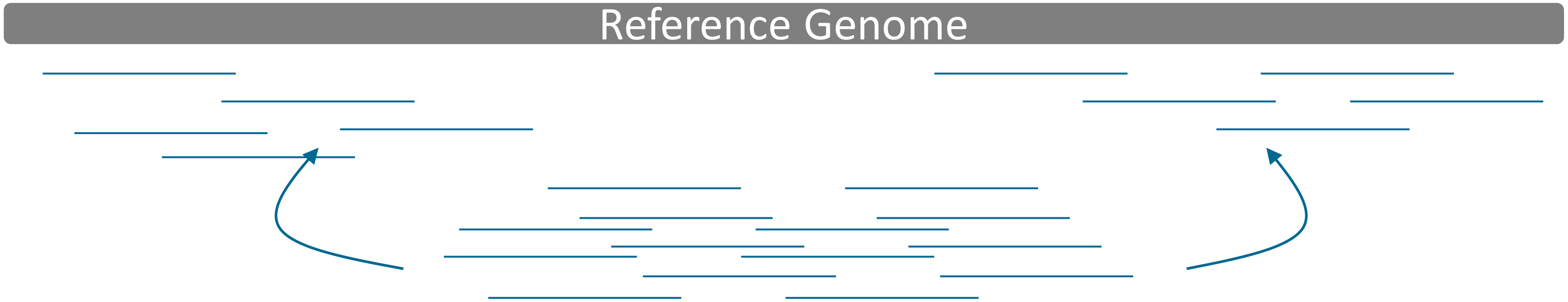
Translate into NGS reads

AAGTGTTGAGGCTTTGTGATGCTTATATTATATTAGCAAACTTAGA

~100 Million Single Reads per Patient!

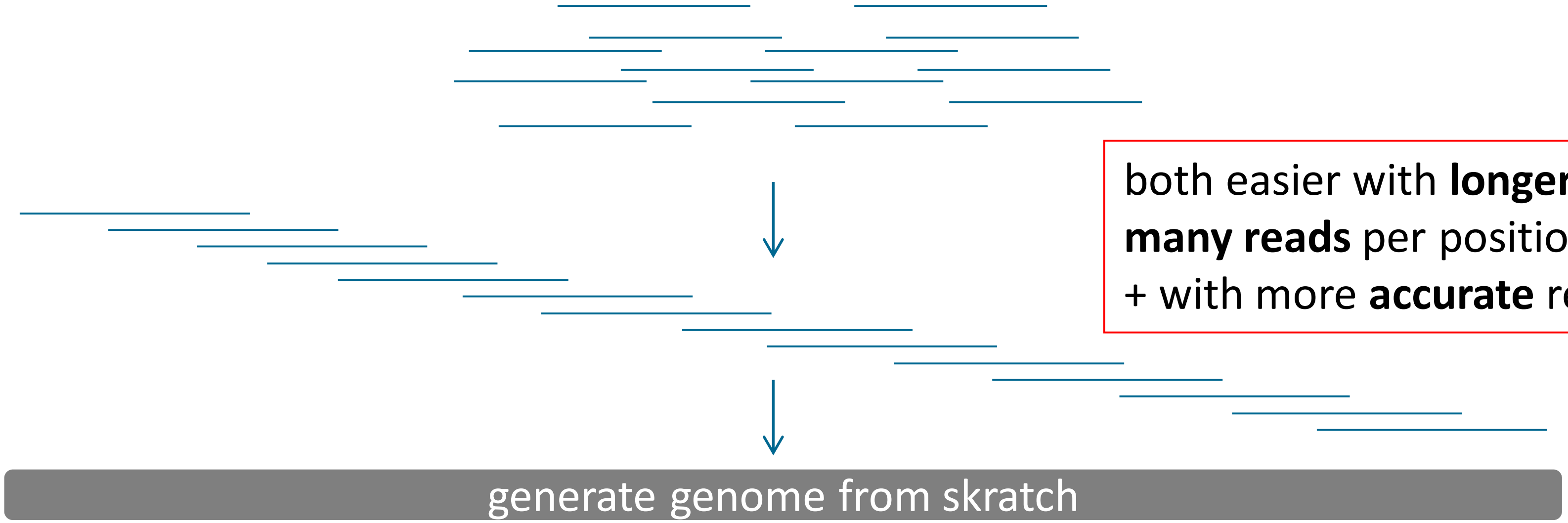
Mapping sequencing reads

- Mapping the reads to a **reference genome**



Reference Genome

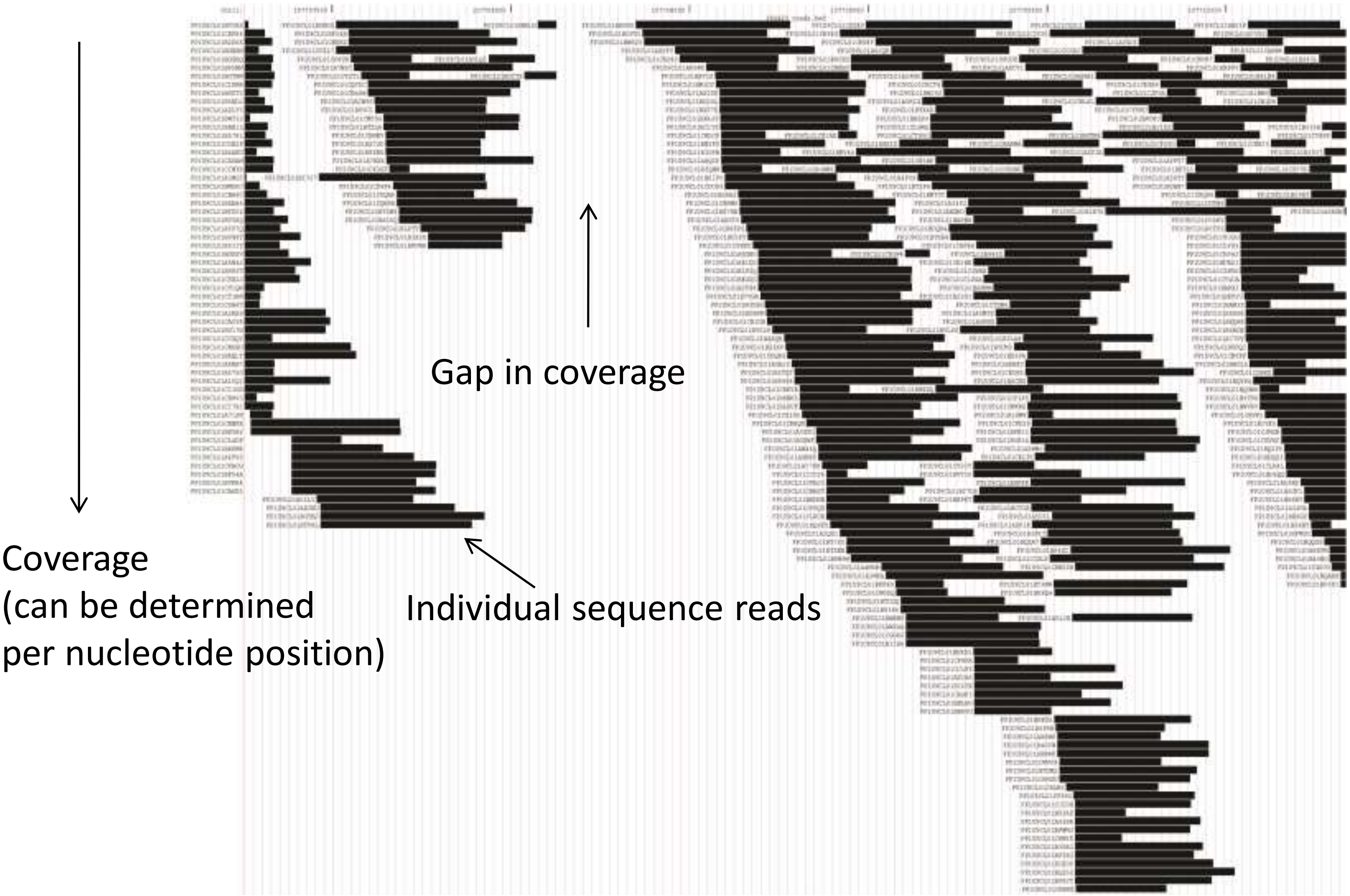
- Mapping the reads in a ***de novo* assembly**



generate genome from skratch

both easier with **longer** reads
many reads per position
+ with more **accurate** reads

Aligning sequencing reads to the reference genome

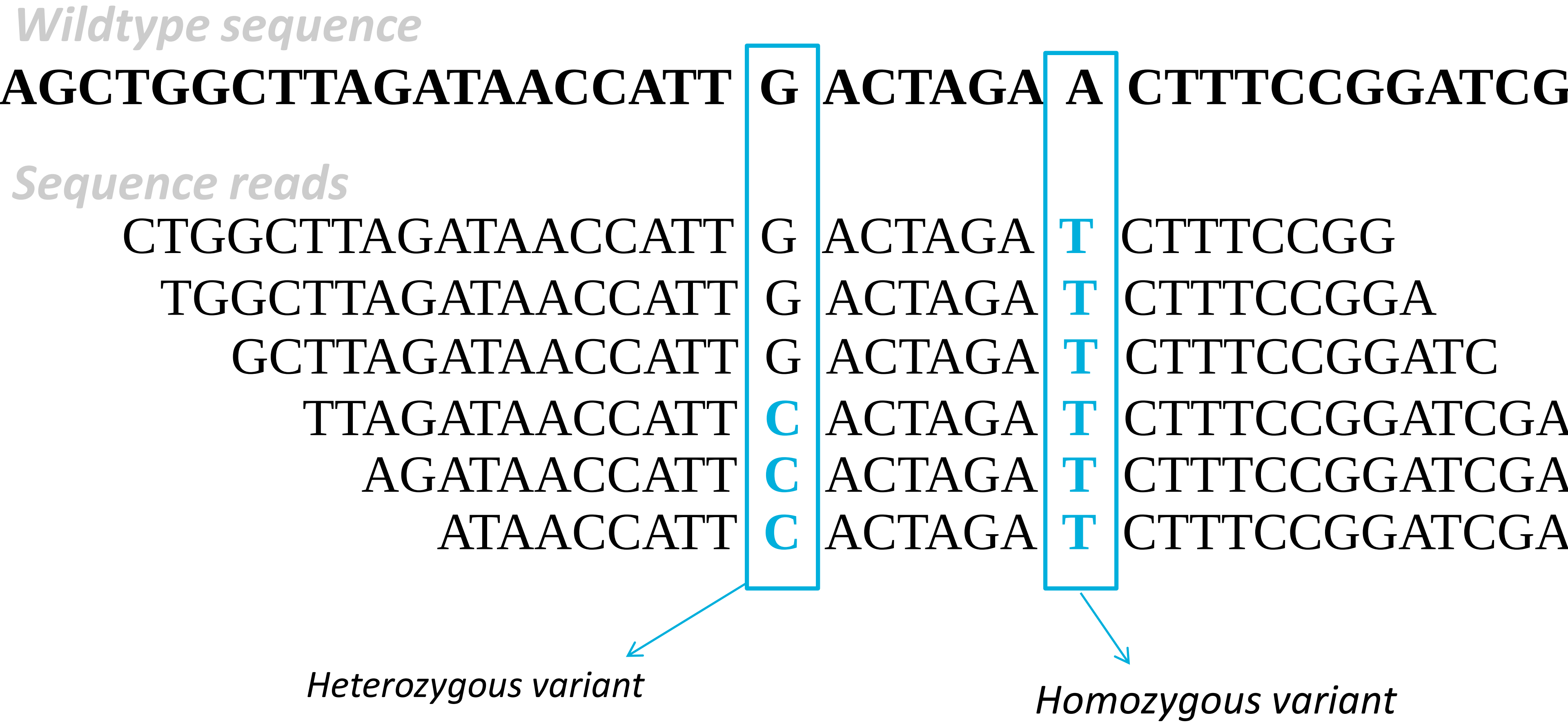


Gap in coverage

Coverage
(can be determined
per nucleotide position)

Individual sequence reads

Variant detection - Theory



Wildtype sequence

AGCTGGCTTAGATAACCATT G ACTAGA A CTTTCCGGATCG

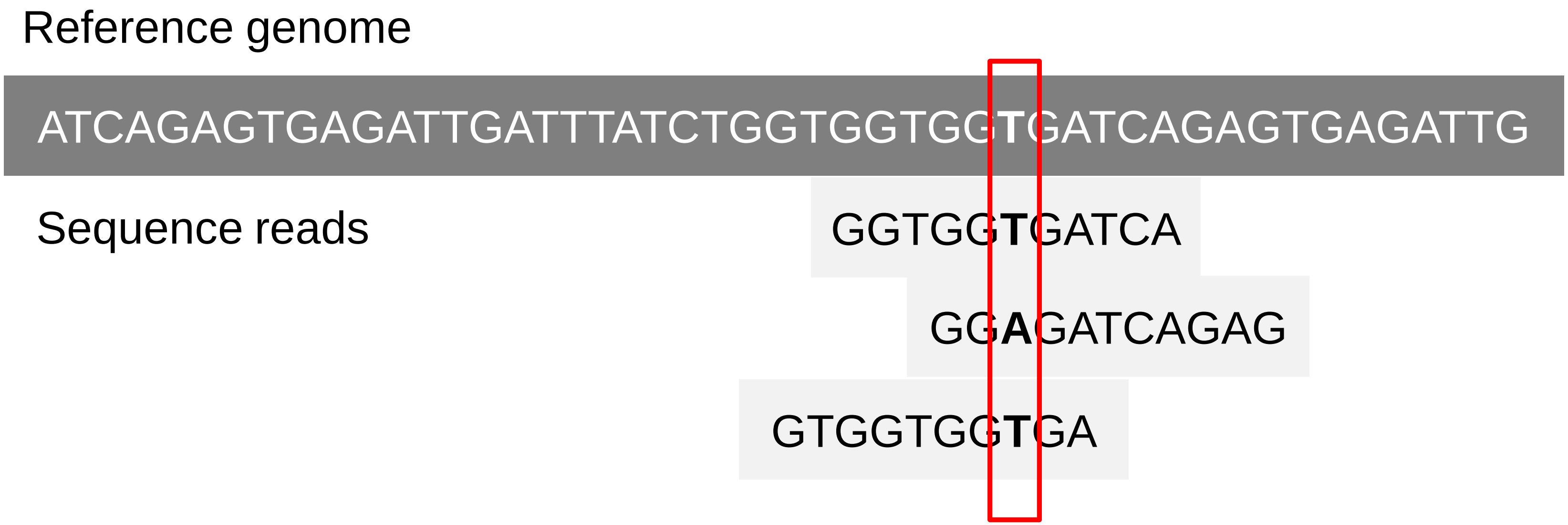
Sequence reads

CTGGCTTAGATAACCATT G ACTAGA **T** CTTTCCGG
TGGCTTAGATAACCATT G ACTAGA **T** CTTTCCGGA
GCTTAGATAACCATT G ACTAGA **T** CTTTCCGGATC
TTAGATAACCATT **C** ACTAGA **T** CTTTCCGGATCGA
AGATAACCATT **C** ACTAGA **T** CTTTCCGGATCGA
ATAACCATT **C** ACTAGA **T** CTTTCCGGATCGA

Heterozygous variant

Homozygous variant

The importance of coverage



Reference genome

ATCAGAGTGAGATTGATTTATCTGGTGGTGG**T**GATCAGAGTGAGATTG

Sequence reads

GGTGG**T**GATCA

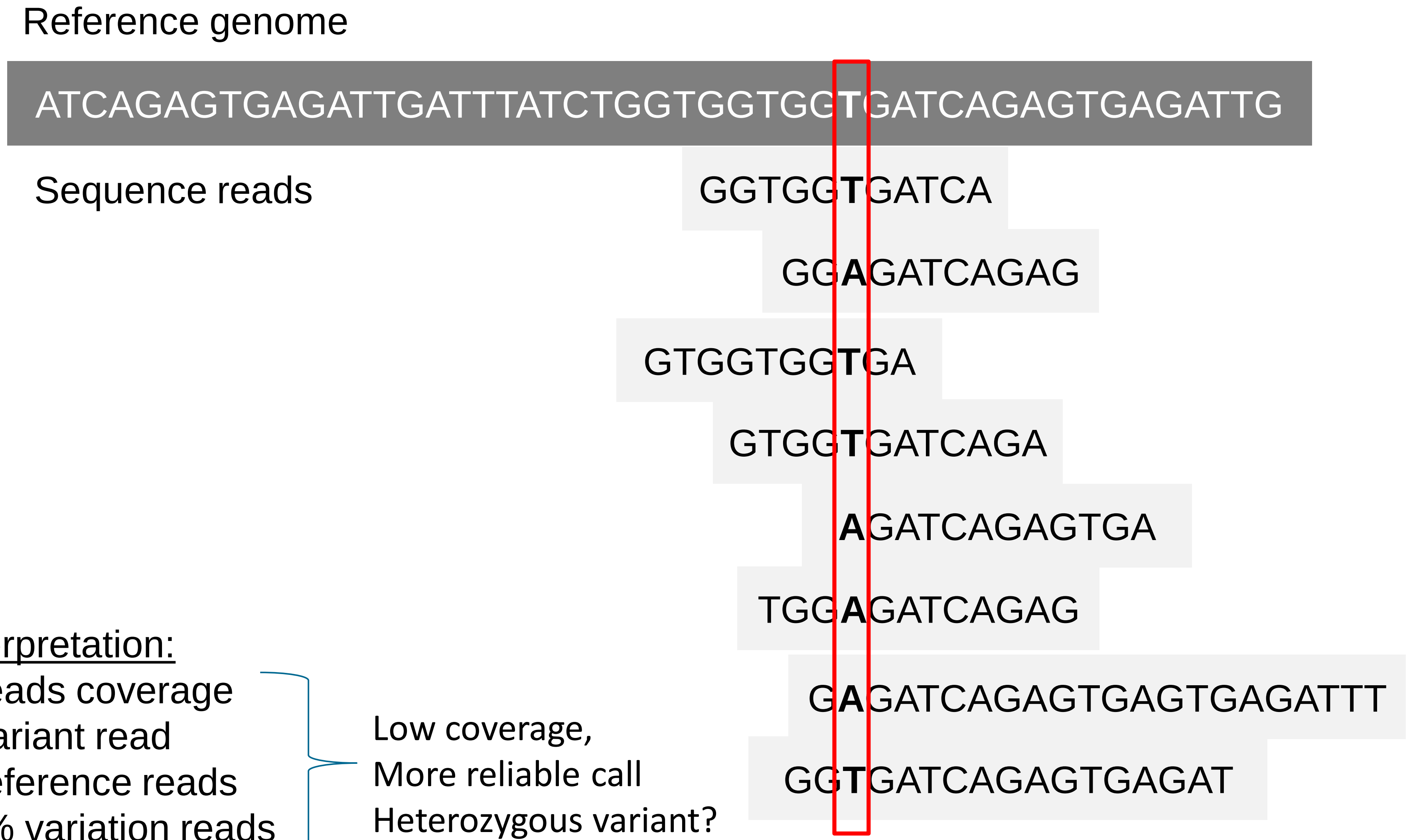
GG**A**GATCAGAG

GTGGTGG**T**GA

<u>Interpretation:</u>
3 reads coverage of nucleotide position (all unique)
1 variant read
2 reference reads
33% variation reads

Very low coverage, no reliable call

The importance of coverage



Reference genome

ATCAGAGTGAGATTGATTTATCTGGTGGTGG**T**GATCAGAGTGAGATTG

Sequence reads

GGTGG**T**GATCA

GG**A**GATCAGAG

GTGGTGG**T**GA

GTGG**T**GATCAGA

AGATCAGAGTGA

TGG**A**GATCAGAG

G**A**GATCAGAGTGAGTGAGATTT

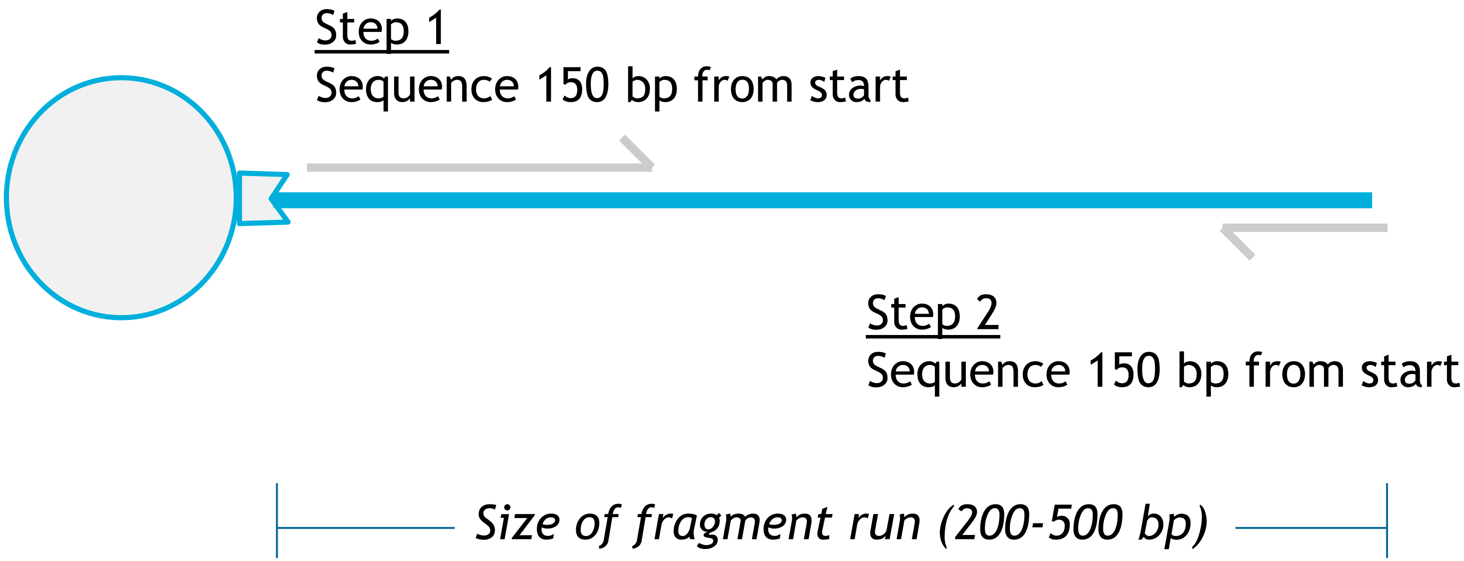
GG**T**GATCAGAGTGAGAT

<u>Interpretation:</u>

8 reads coverage
4 variant read
4 reference reads
50% variation reads

Low coverage,
More reliable call
Heterozygous variant?

Paired-End sequencing



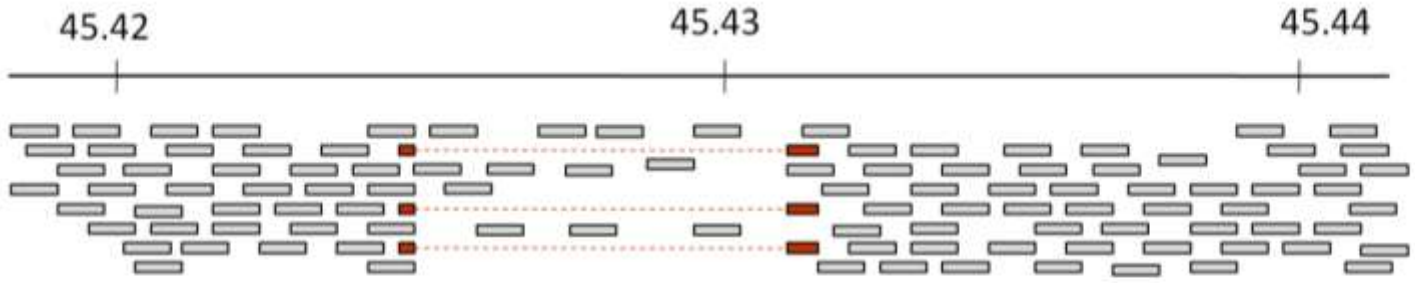
Step 1
Sequence 150 bp from start

Step 2
Sequence 150 bp from start

Size of fragment run (200-500 bp)

- More data from 1 sequence run (but also takes longer)
- Higher confidence mapping due to relation between sequences
- Useful for studying structural genomic variation

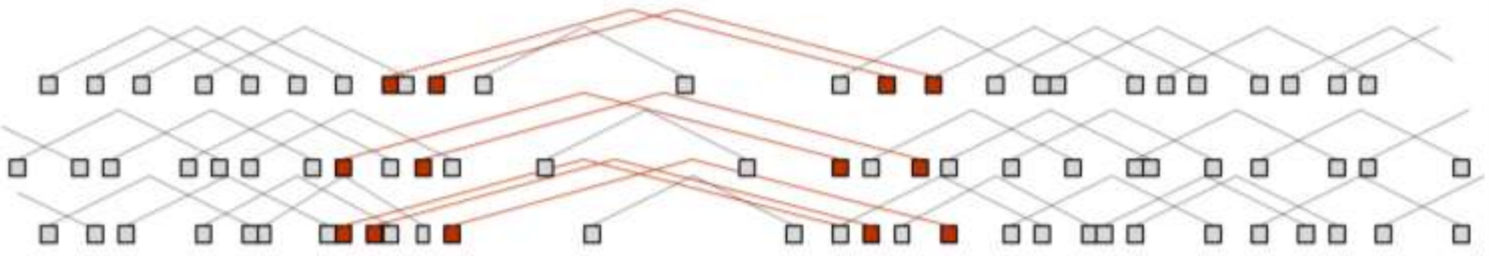
Detecting structural genomic variation by NGS



Single-end reads

Detect CNVs by looking at **coverage of sequence reads** and at **split reads**

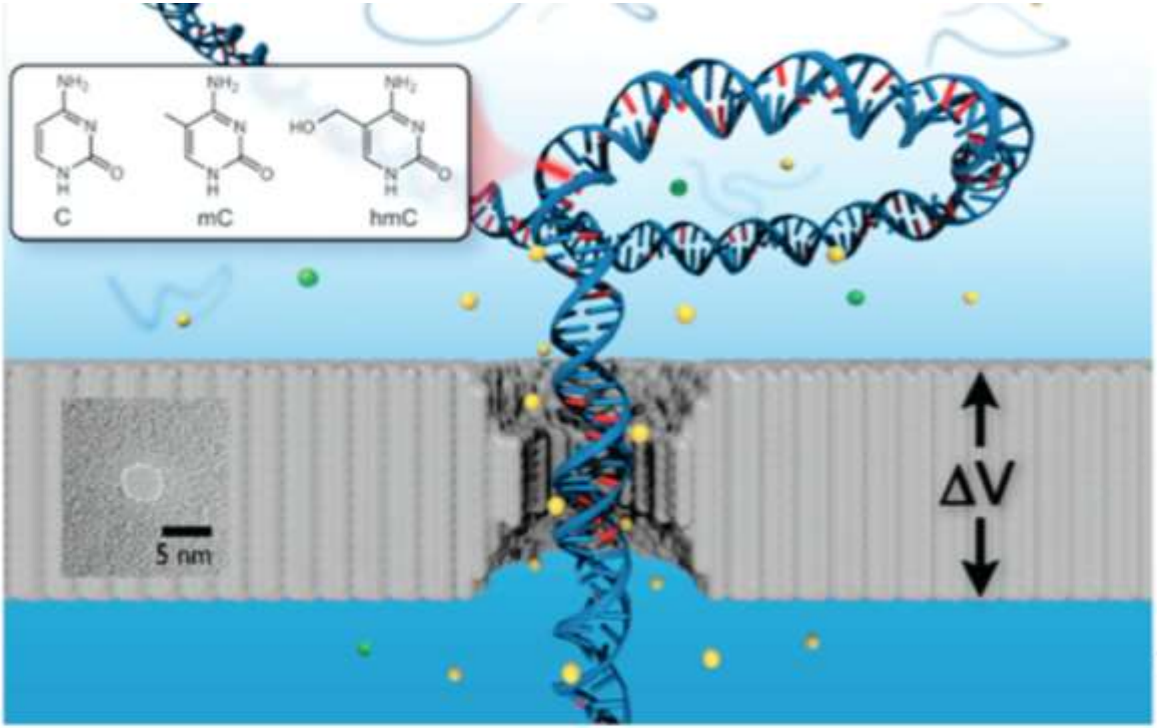
Paired-end reads (some mate-pair)

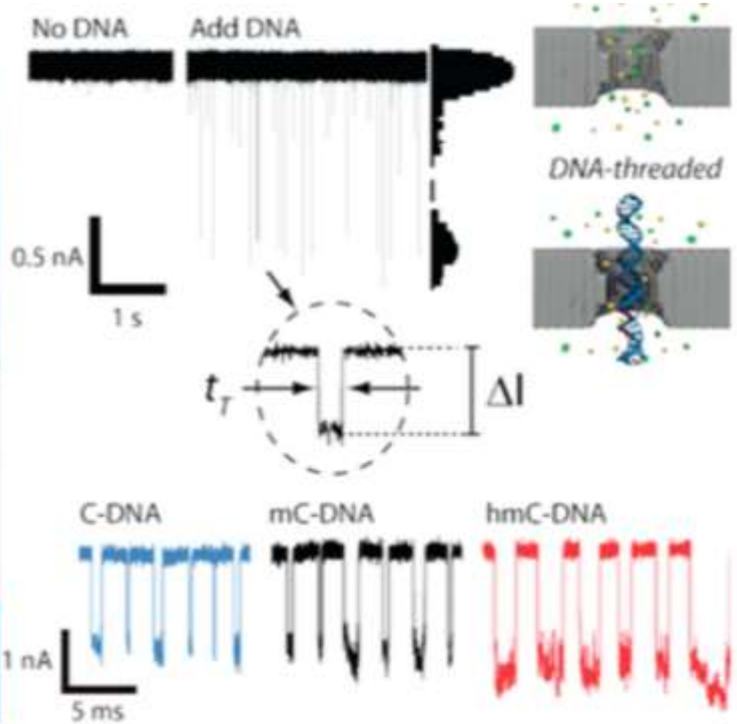


Detects also **balanced rearrangements**!

Vissers, de Vries & Veltman, JMG 2009

Single molecule sequencing without amplification; the future?





Feng et al. Genomics Proteomics Bioinformatics 2015;
Wang et al. Frontiers in Genetics 2015

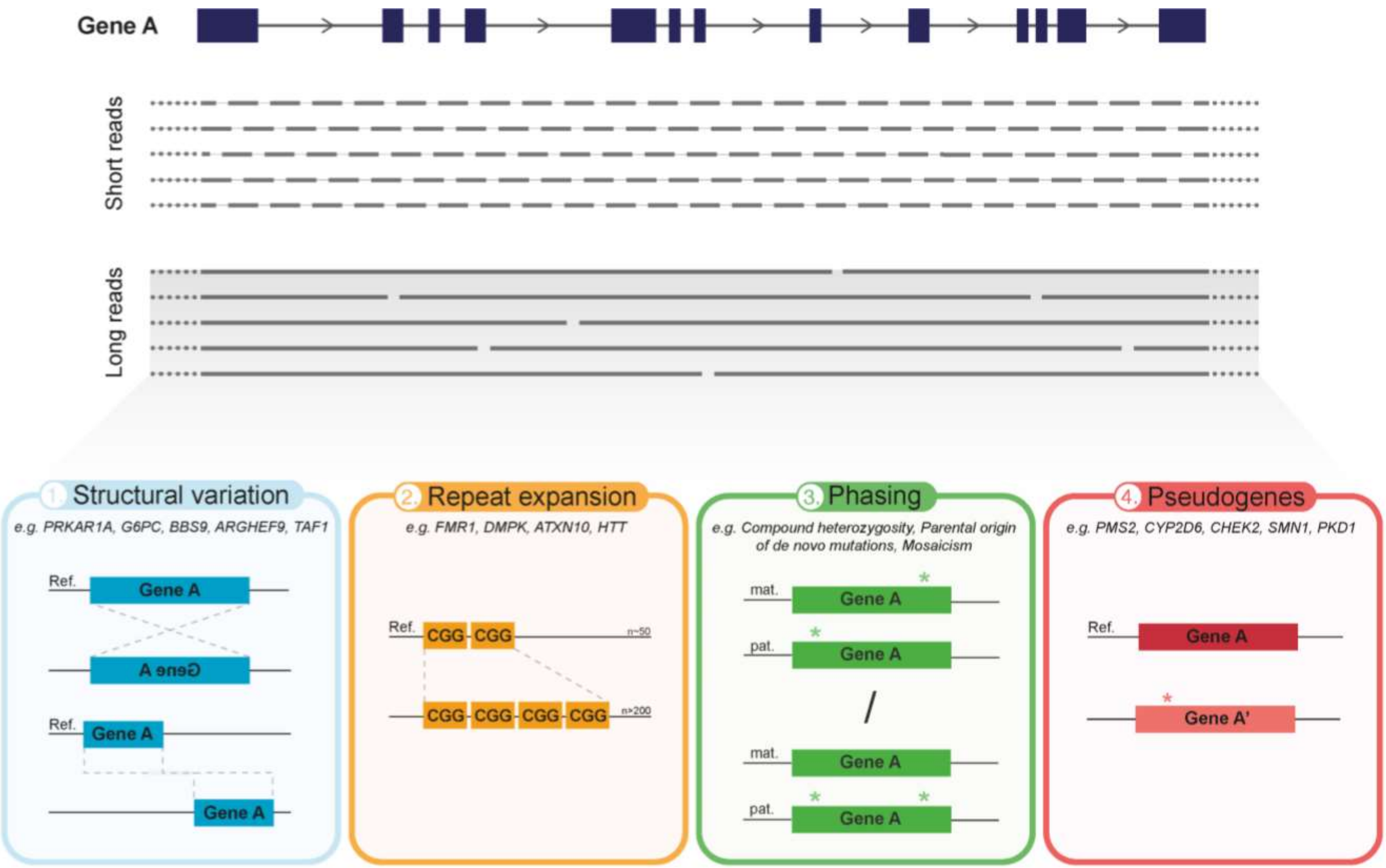
The promise of single molecule sequencing

- Amplification of DNA prior to sequencing introduces artefacts, DNA needs to be chopped in small fragments, it takes time and is expensive

- Sequencing of one molecule (chromosome) at the time is potentially ideal, especially for analyzing complex genomic regions (e.g. HLA)

- Major advantage: Long sequencing reads
- Major Challenge: Raw sequencing accuracy
- Major companies: Pacific Biosciences & Oxford Nanopores

Advantages of long read sequencing



Mantere, Kersten & Hoischen. Frontiers in Genetics 2019.

Typical questions for which NGS can be used NOW!

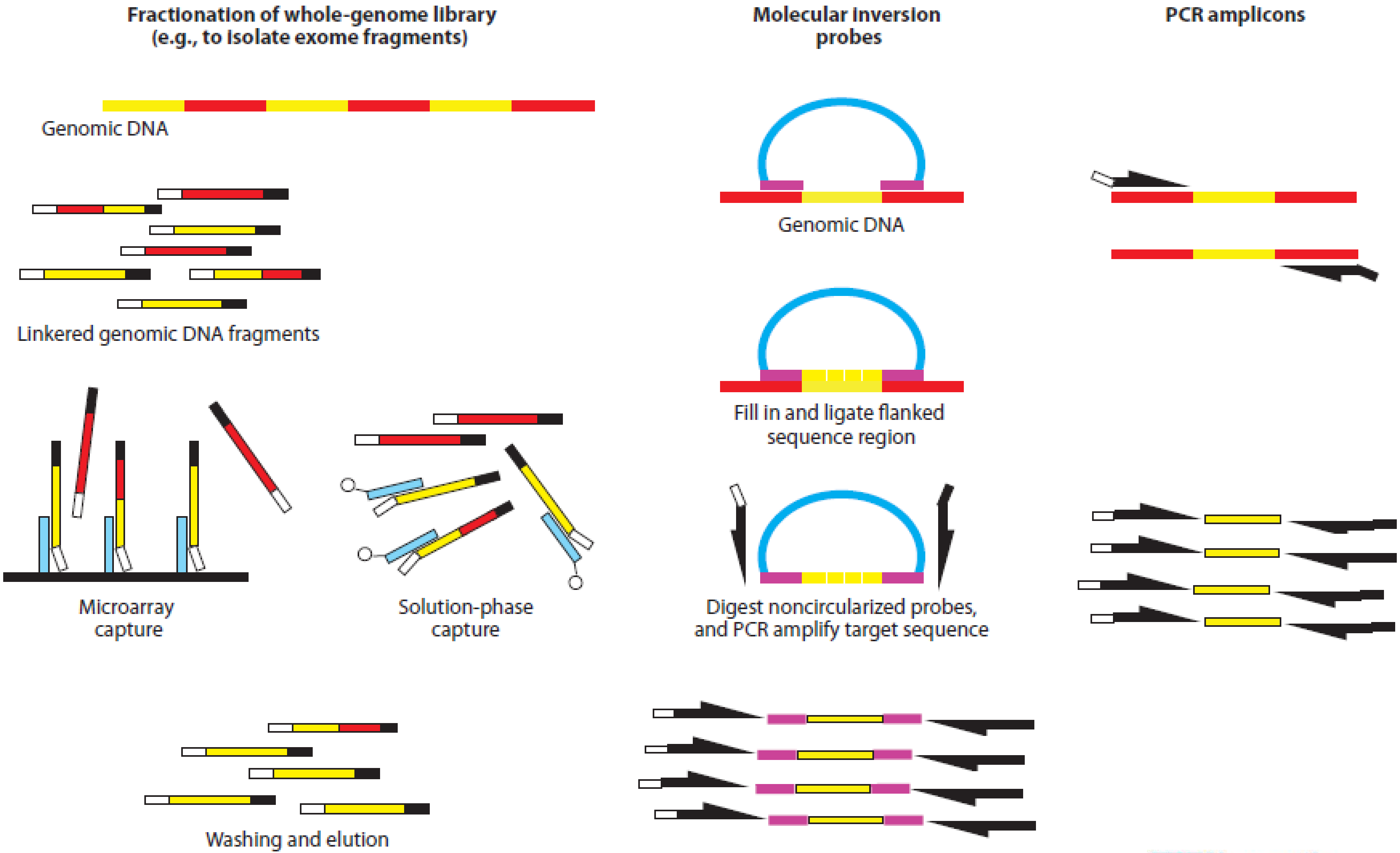
- Can we sequence all known disease genes of genetically heterogeneous diseases in parallel (*e.g.* hereditary breast cancer, ataxia, hereditary blindness)?

- Can we sequence entire candidate disease gene loci (*e.g.* from linkage studies/homozygosity mapping)?

- Can we sequence the whole exome (all **ex**ons of a gen**ome**) to decipher unknown syndromes/diseases?

→ **Enrichment prior to sequencing required!**

Enriching your DNA to be sequenced


Fractionation of whole-genome library
(e.g., to isolate exome fragments)
Molecular inversion probes
PCR amplicons
Genomic DNA
Genomic DNA
Linkered genomic DNA fragments
Fill in and ligate flanked sequence region
Microarray capture
Solution-phase capture
Digest noncircularized probes, and PCR amplify target sequence
Washing and elution

Targeted next generation sequencing examples

- **CFTR diagnostics**
  Enrichment: Amplicon, Molecular inversion probes
  Sequencing: Ion Torrent, Illumina Miseq & Nextseq

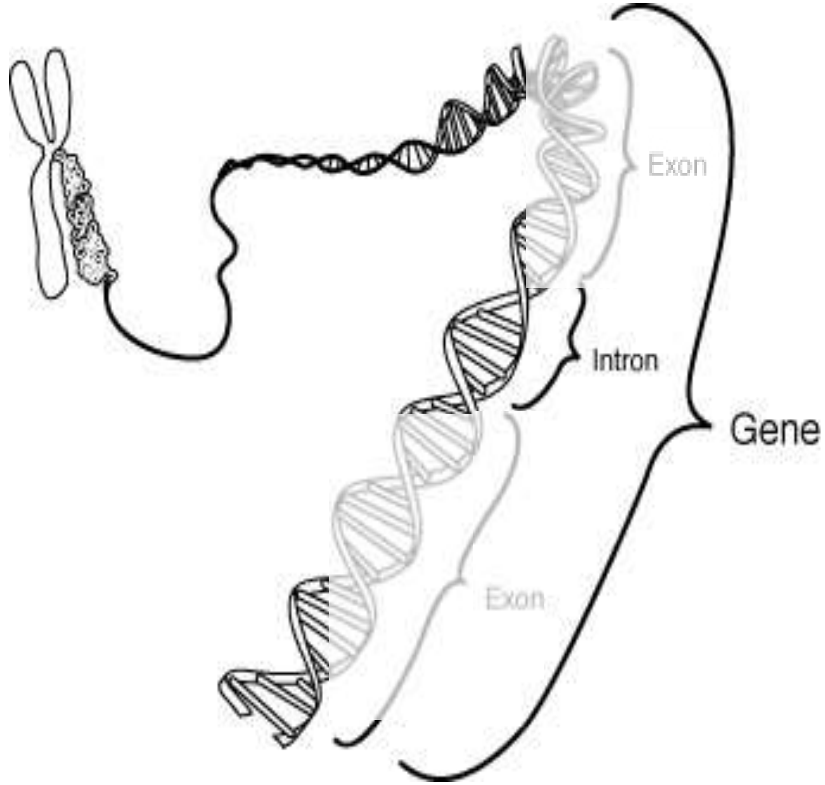
- **Sequencing of 100 candidate male infertility genes**
  Enrichment: Amplicon, Molecular inversion probe, NimbleGen / Agilent in solution enrichment, Fluidigm, Raindance
  Sequencing: Ion Torrent, Illumina Miseq & Nextseq

- **Exome sequencing (diagostics/research)**
  Enrichment options: NimbleGen/Agilent/Twist biosciences
  Sequencing: Illumina Hiseq, Novaseq

Exome sequencing; Practicing for genomes

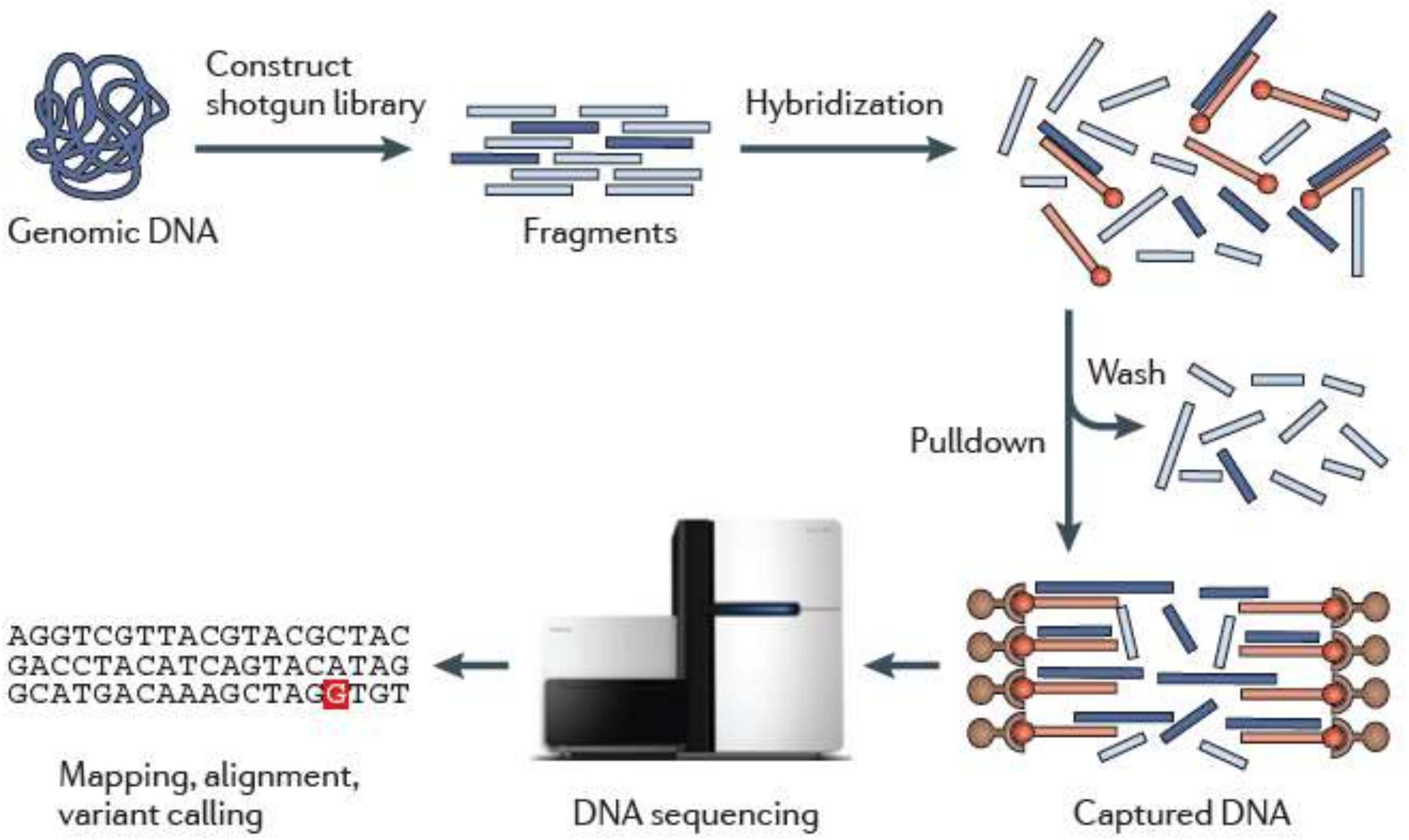
'Exome' (all exons of a genome)

~1% of the human genome



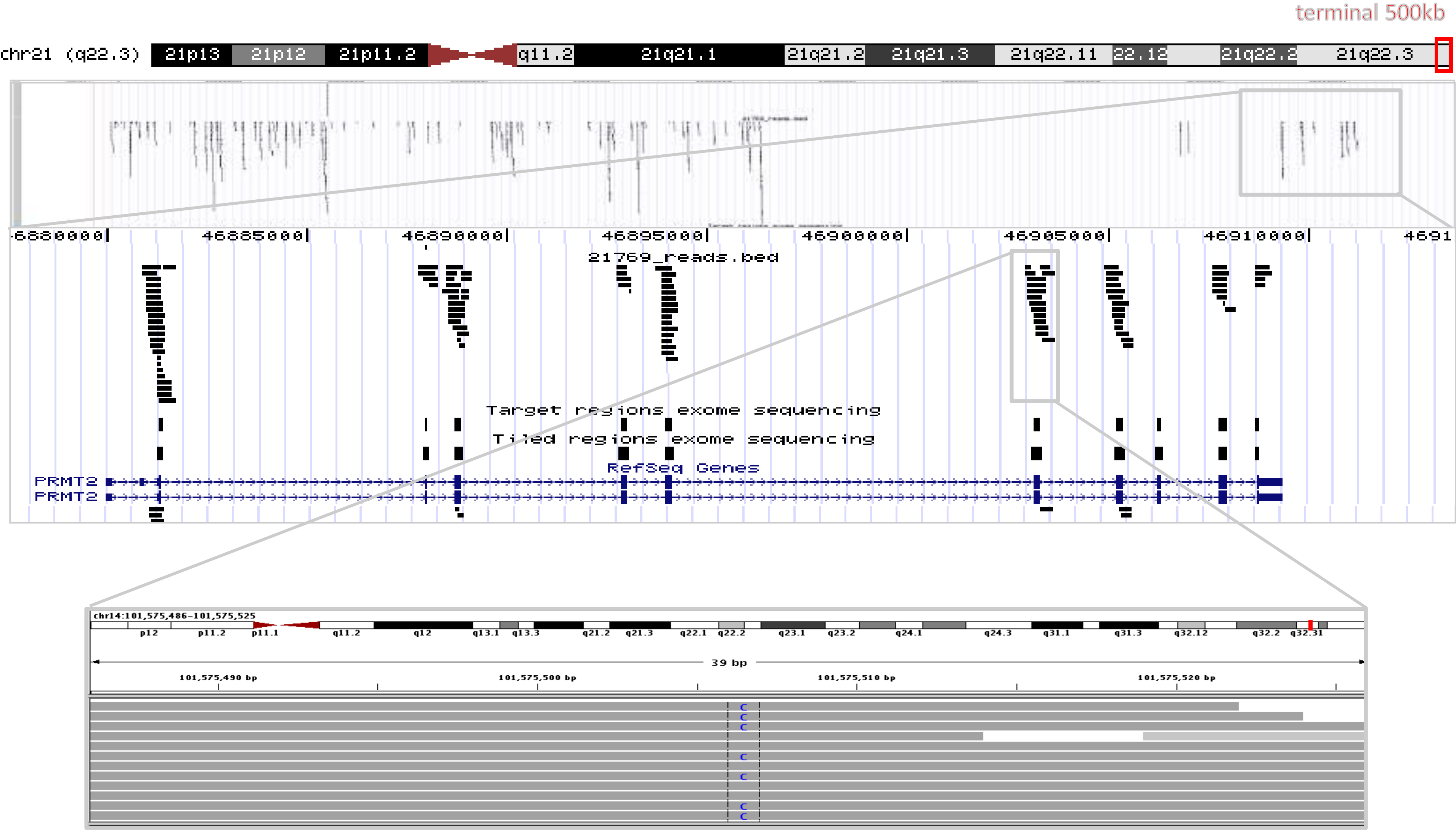
'All' coding sequences of a human genome (>200,000 exons), sequenced and analyzed in **one** experiment

Exome sequencing workflow

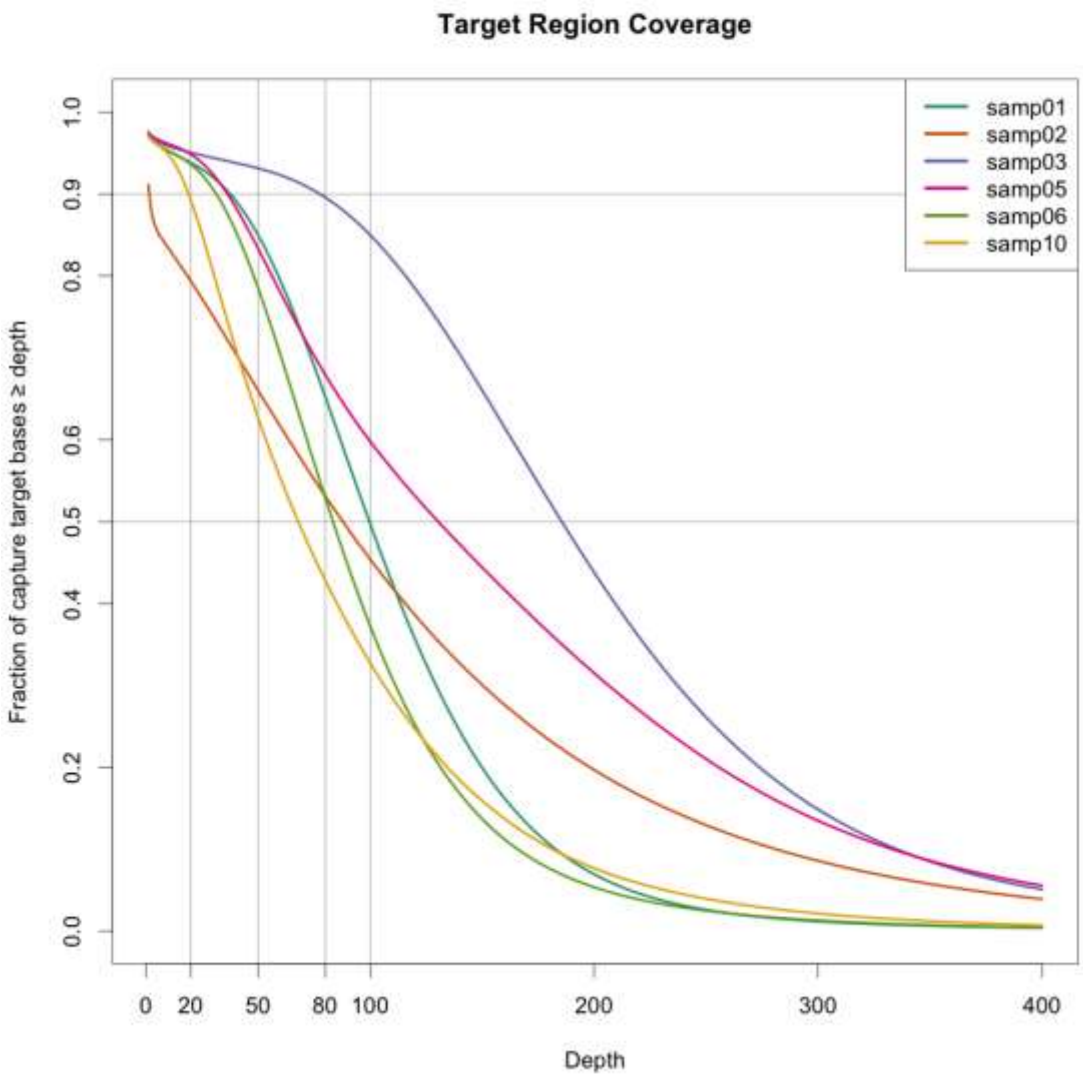


Bamshad et al. Nat Rev Genetics 2011

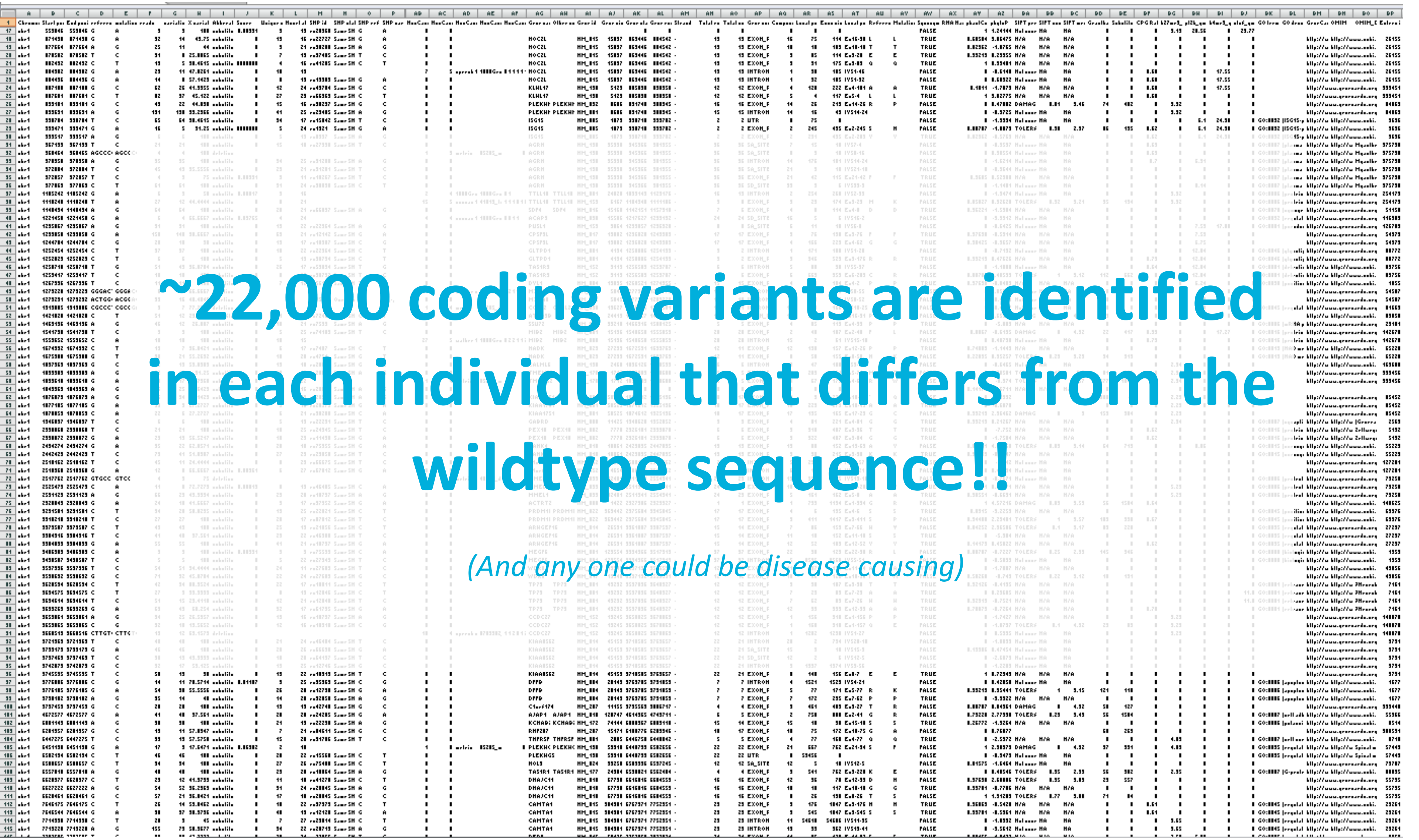
Mapping and annotation of exome sequencing reads



Enrichment is imperfect, varies per sample



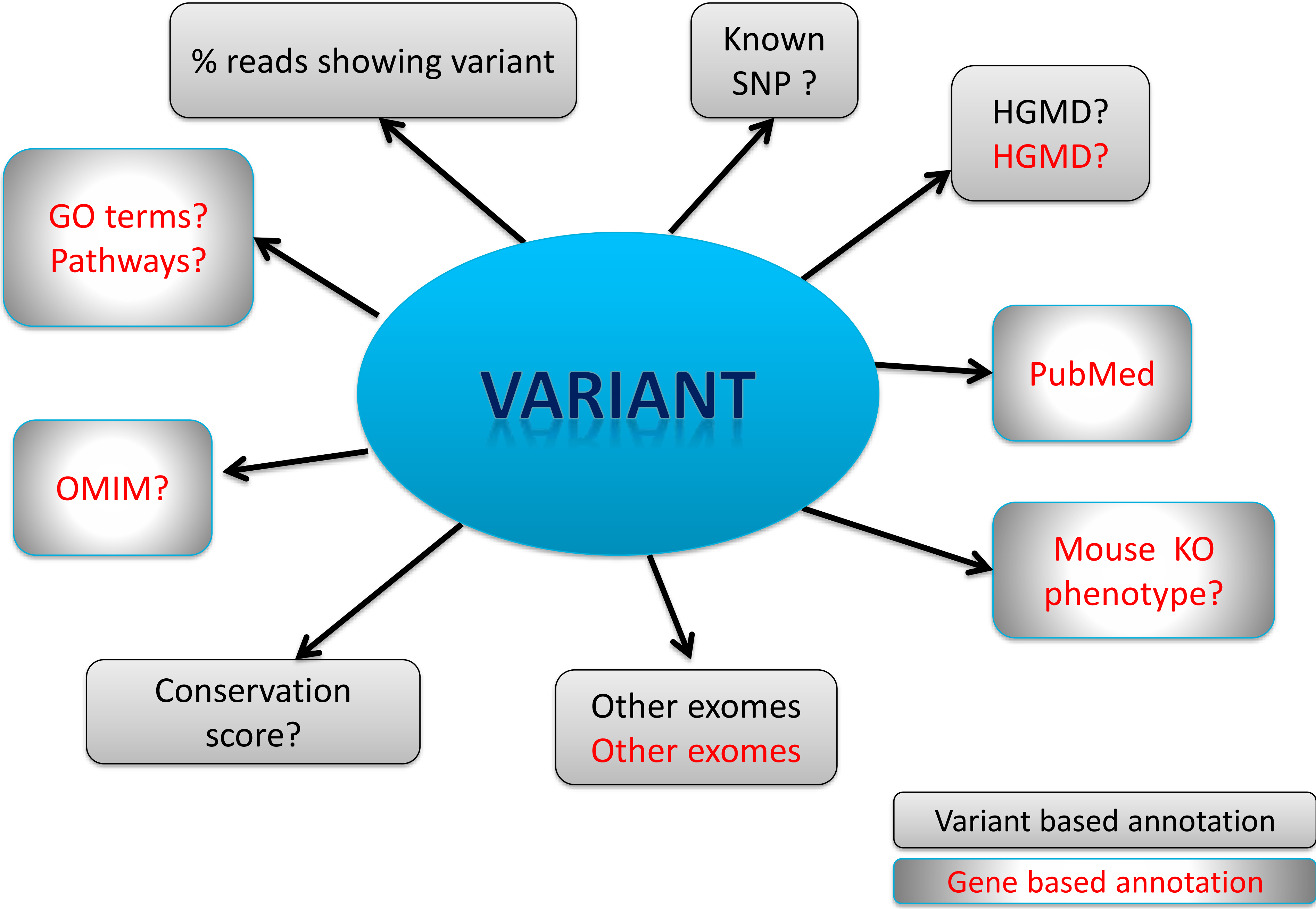
Variant calling and variant annotation



~22,000 coding variants are identified in each individual that differs from the wildtype sequence!!

(And any one could be disease causing)

What do we know about this position in the genome?



Developments in next generation sequencing technology

- Quality

Short read sequencing reliable for most applications
Average sequencing coverage reliable for detection of point mutations
Long read sequencing better for structural variation, repeat expansions, but high error rate for single nucleotide variation detection

- Throughput/Speed

Short read technology allows exome and genome sequencing in days
Thousands genomes can be sequenced on individual systems annually
Long read genomes still take more time and have less throughput

- Costs

Short read technology: 100x coverage exome < €300, 30x coverage genome < €1.200. Long-read genome 30x coverage ~ €10.000
Prices genomes below €1.000 in 1-2 year, below €200 in 5 years?

Genome sequencing: All variation in one experiment!

DNA from
blood/saliva

Genome
with ‘all’ variation

Important:
- Accuracy
- Completeness
(all genome, all variations)
- Speed
- Price

Why perform whole genome sequencing?

If you consider genetics may play a role in your patient:
Why not read the entire book?
Why settle for studying what we now know?
We still live in the dark ages of genetics!

Key advantages of genome sequencing:

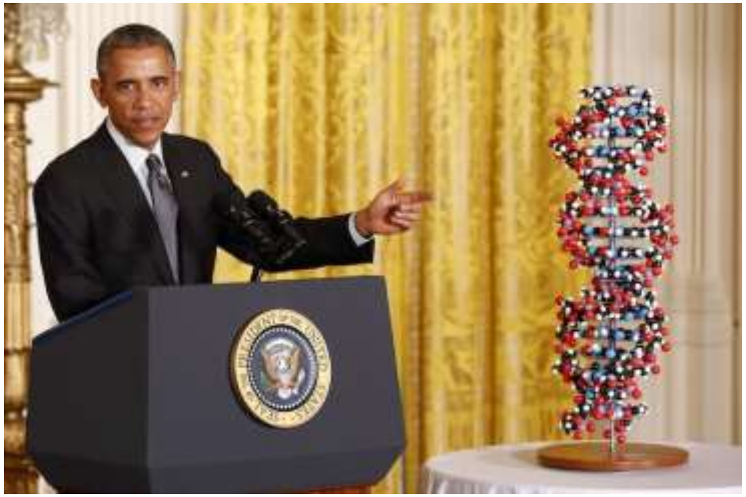
Completeness	All variation
Simplicity	One test

Price is dropping, quality will continue to improve,
no enrichment, better for structural variation

Genome sequencing centers established around the world

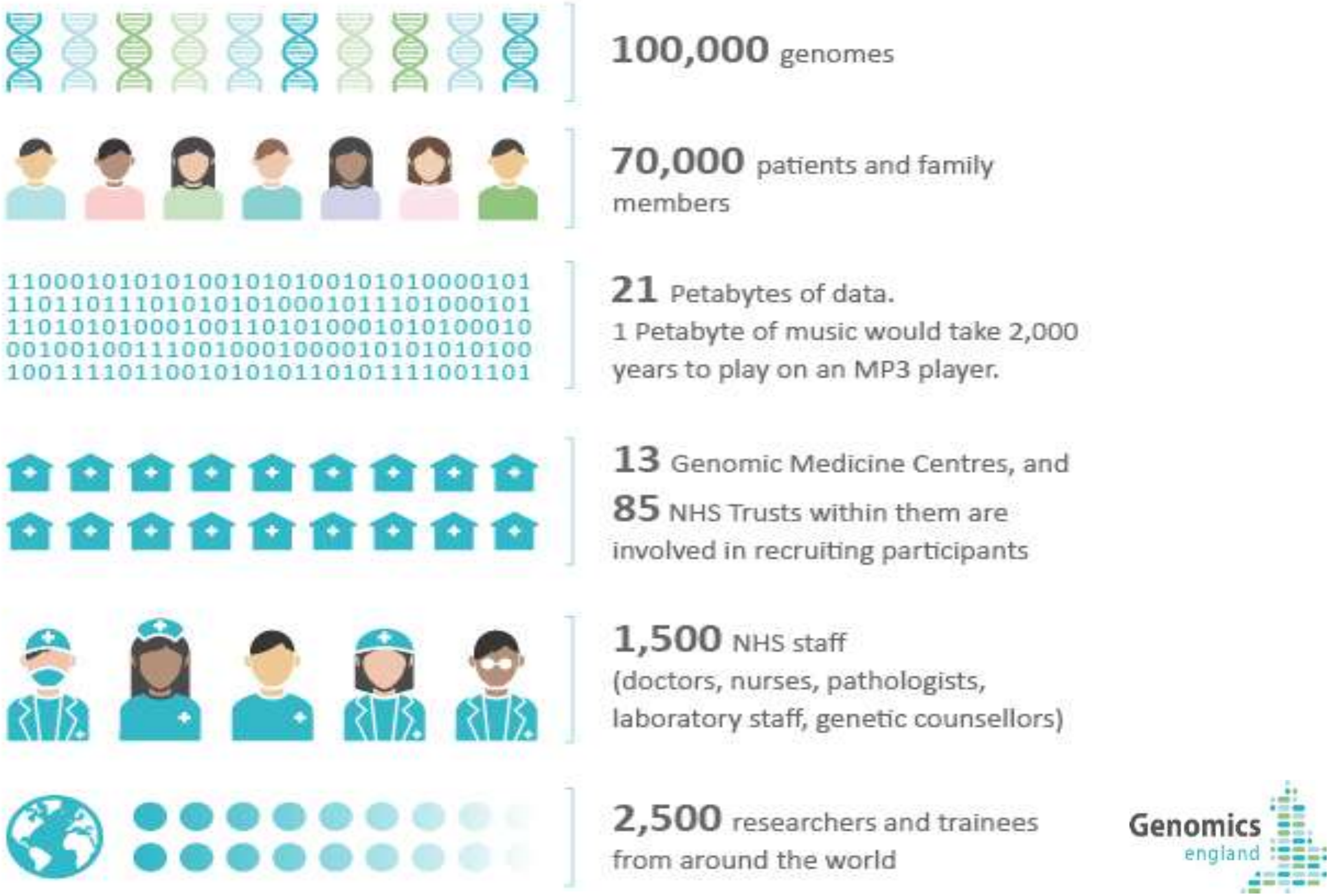
Transformative Genomics: England Begins Daunting Task of Sequencing 100,000 Genomes by 2017

Genomics england



> “Tonight, I’m launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier.”
>
> — President Barack Obama, State of the Union Address, January 20, 2015

The UK 100,000 Genomes Project (and beyond)



100,000 genomes

70,000 patients and family members

21 Petabytes of data.
1 Petabyte of music would take 2,000 years to play on an MP3 player.

13 Genomic Medicine Centres, and
85 NHS Trusts within them are involved in recruiting participants

1,500 NHS staff
(doctors, nurses, pathologists, laboratory staff, genetic counsellors)

2,500 researchers and trainees from around the world

Genomics england

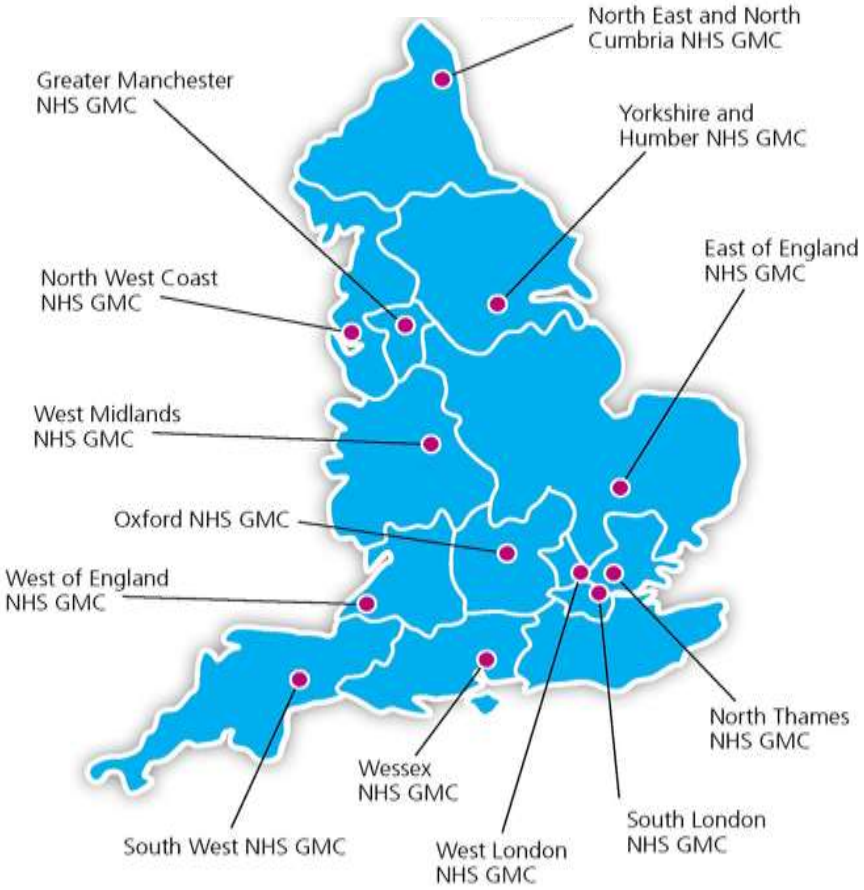
Slides kindly provided by Dr. Richard Scott, clinical lead Genomics England

Newcastle University

Illumina Partnership July 2014

Sequencing Centre January 2016

NHS GMCs December 2014

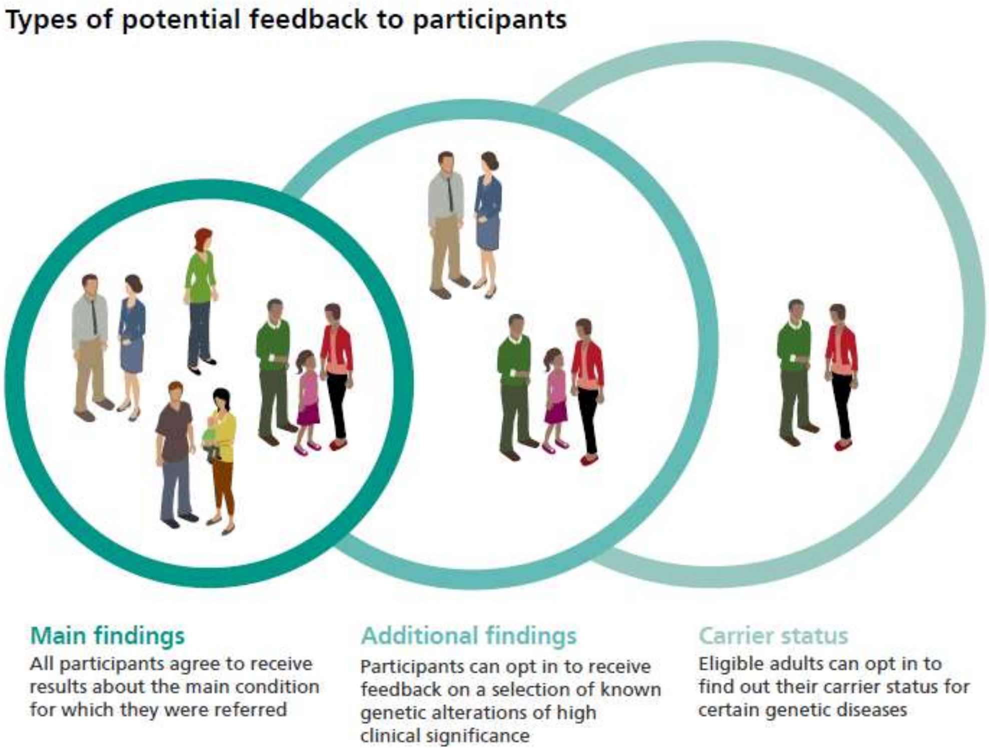


Data Centre November 2014

Biorepository

What are we telling participants?

- Information about a patient’s main condition
- Information about additional ‘serious and actionable’ conditions (optional)
- Carrier status for non affected parents of children with rare disease (optional)



Types of potential feedback to participants

Main findings
All participants agree to receive results about the main condition for which they were referred

Additional findings
Participants can opt in to receive feedback on a selection of known genetic alterations of high clinical significance

Carrier status
Eligible adults can opt in to find out their carrier status for certain genetic diseases

Image courtesy of Health Education England

A new national diagnostic service

- Increasing move towards exome and genome sequencing; further use of patient/parent trios
- Genomics England will provide data infrastructure for a new Genomics Laboratory Service for the NHS in England
- Central test request system with shared ‘genomic test directory’
- Central WGS pipeline (lab and bioinformatics). Reporting by hubs
- Central shared genomic knowledgebase for NHS laboratories
- 7 genomics laboratory hubs for performing other/additional genetic tests, interpretation and reporting

www.genomicsengland.co.uk

NHS England

Genomics england

Newcastle Fertility Centre @ Life

Newcastle University

Institute of Genetic Medicine

NHS

Northern Genetics Service

W

wellcome

IMIGC

INTERNATIONAL MALE INFERTILITY GENOMICS CONSORTIUM