

Next Generation Sequencing (NGS): Introduction, Data Formats, Processing & Variant Detection: Part I

Dr Stephen J Newhouse

Lead Data Scientist & Senior Bioinformatician @ NHIR BRC-MH SlaM NHS & IoPPN KCL & UCL Farr Institute

stephen.j.newhouse@kcl.ac.uk, [@s_j_newhouse](https://twitter.com/s_j_newhouse)

Genomic Medicine MSc

15 Feb 2016

Overview

1. DNA Sequencing the very basics
2. NGS technologies and terminology
3. Experimental designs
4. Data formats
5. Data processing overview

Special Acknowledgements:

many slides taken from :-

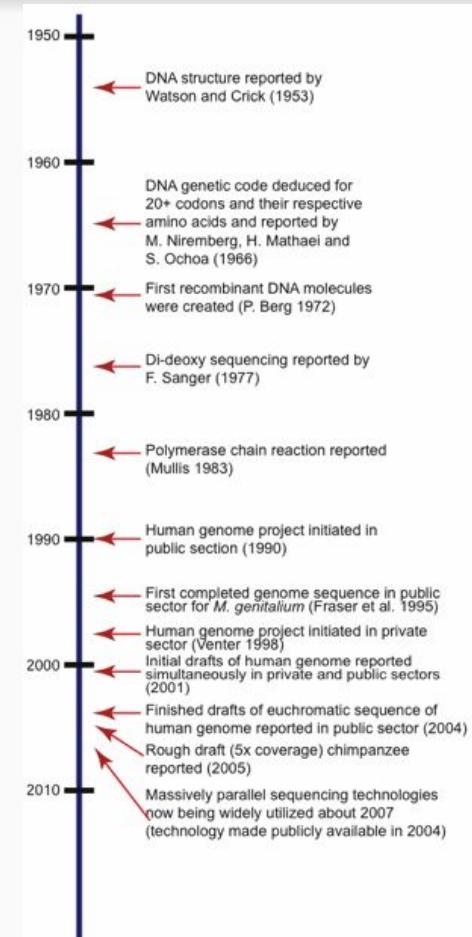
- <https://www.broadinstitute.org/gatk/events/slides/1506>
- https://www.broadinstitute.org/gatk/events/slides/1506/GATKwr8-A-2-Intro_toHTS.pdf

(these are the experts & all credit goes to them!)

1. DNA Sequencing

Some Milestones in Molecular Biology

- First Isolation of DNA : 1867 (Freidrich Meisher)
- Composition of nucleic acids; tetranucleotide theory : 1909 - 1940 (Phoebus Levine)
- Base Pair rules: G=C and A=T : 1950 (Edwin Chargaff)
- DNA Structure: 1953 (Watson & Crick)
- Genetic Code deduced, AA>Condons>Proteins: 1966 (Niremberg,Mathael, Ochoa)
- **Sanger Sequencing : 1977 (Frederick Sanger)**
- **Polymerase Chain Reaction: 1983 (Kary Mullis)**
- **Human Genome Project Started: 1990**
- **Next-Generation Sequencing : 2005**

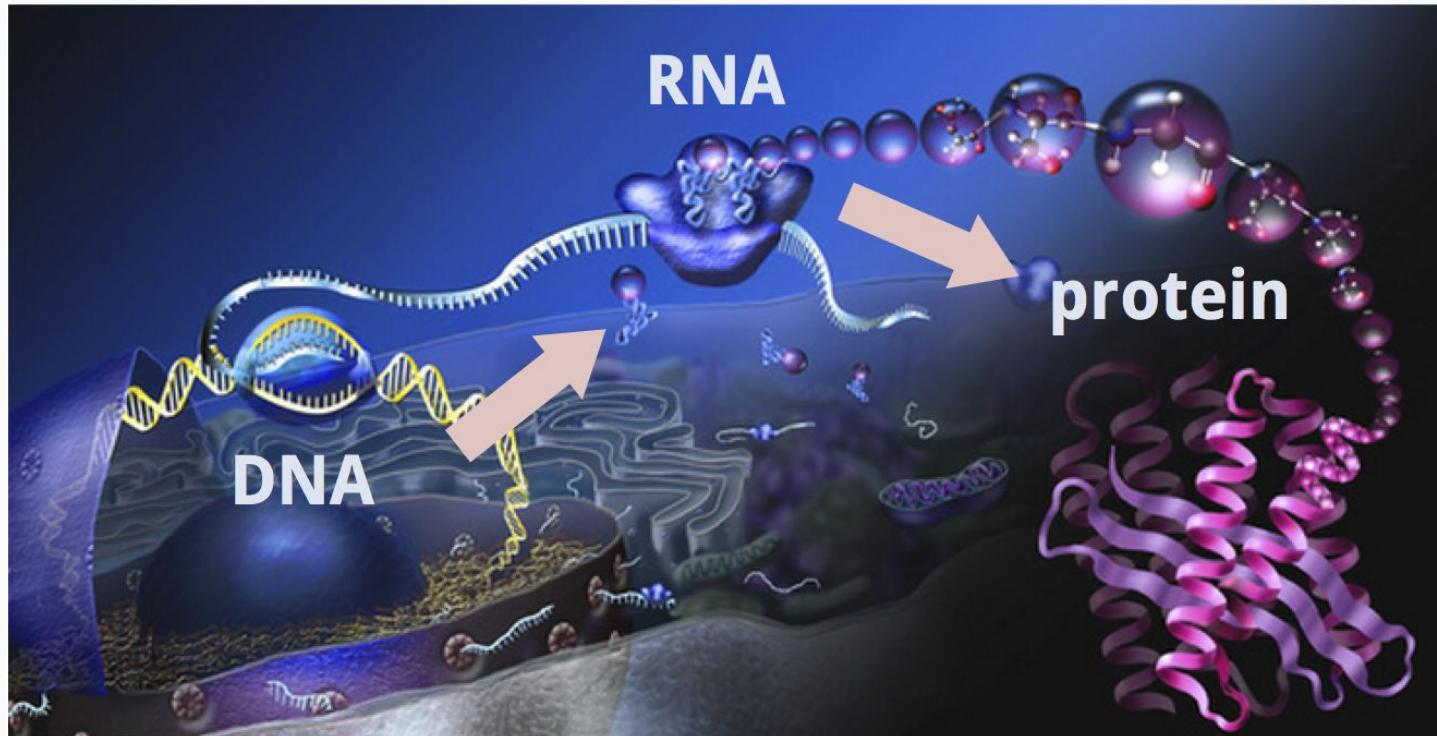


What is DNA Sequencing?

- DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule
- Any method or technology that is used to determine the order of the four (5) bases—adenine, guanine, cytosine, and thymine (A,G,C,T) in a strand of DNA
- DNA can be isolated from cells of animals, plants, bacteria, archaea, or virtually any other source of genetic information.
- The DNA sequencing revolution has impacted nearly every field of biological research!

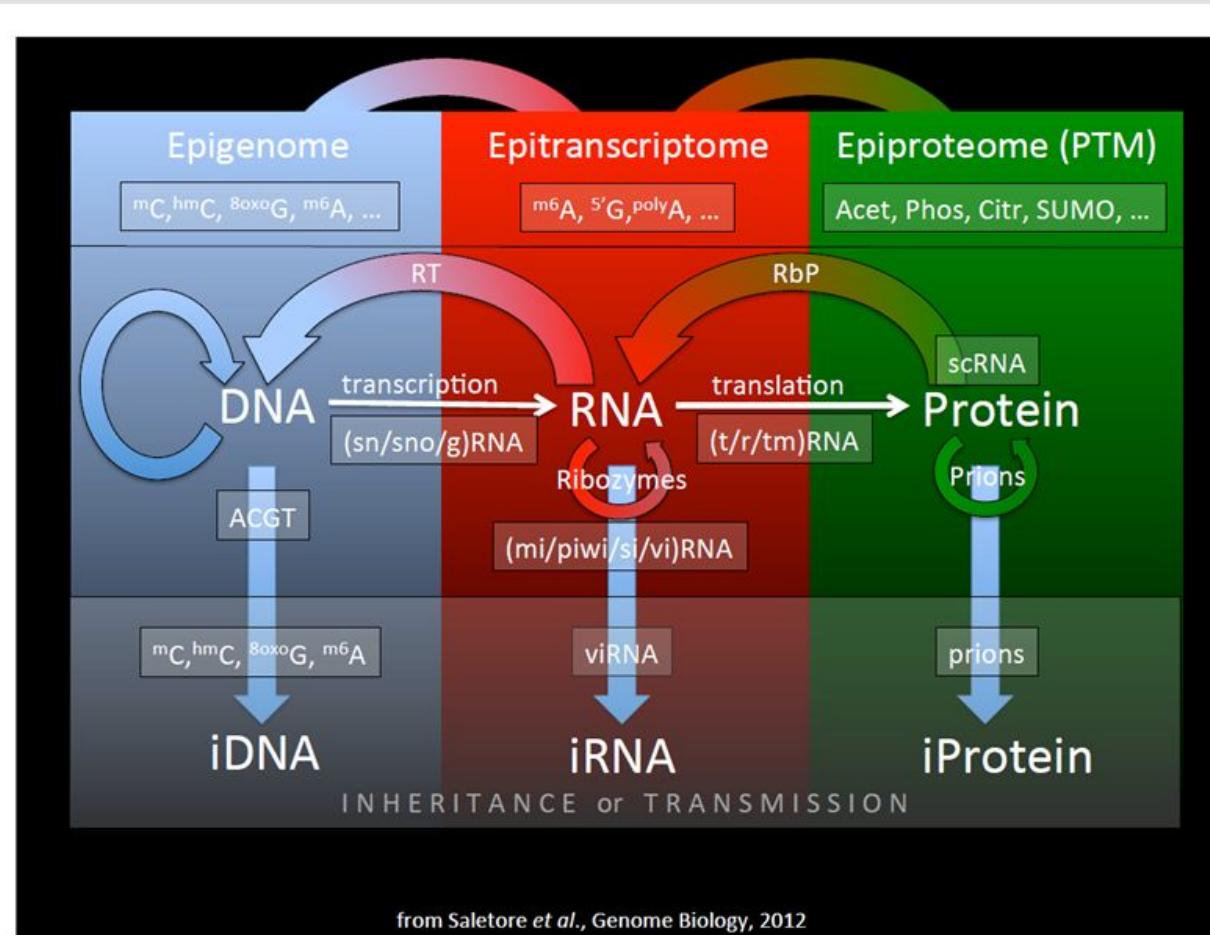
The Central Dogma of Molecular Biology

James Watson version - 1965



So once we have the genomic DNA sequence of a species we have all of the information there is?.....Not Really....

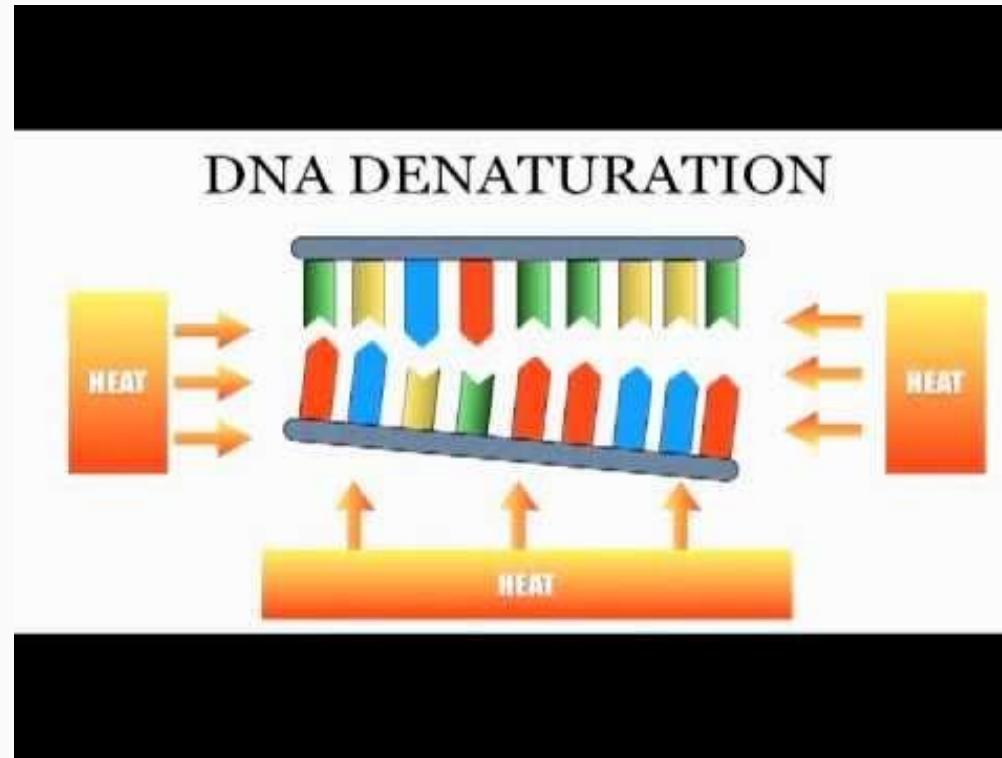
The Central Dogma : 2016



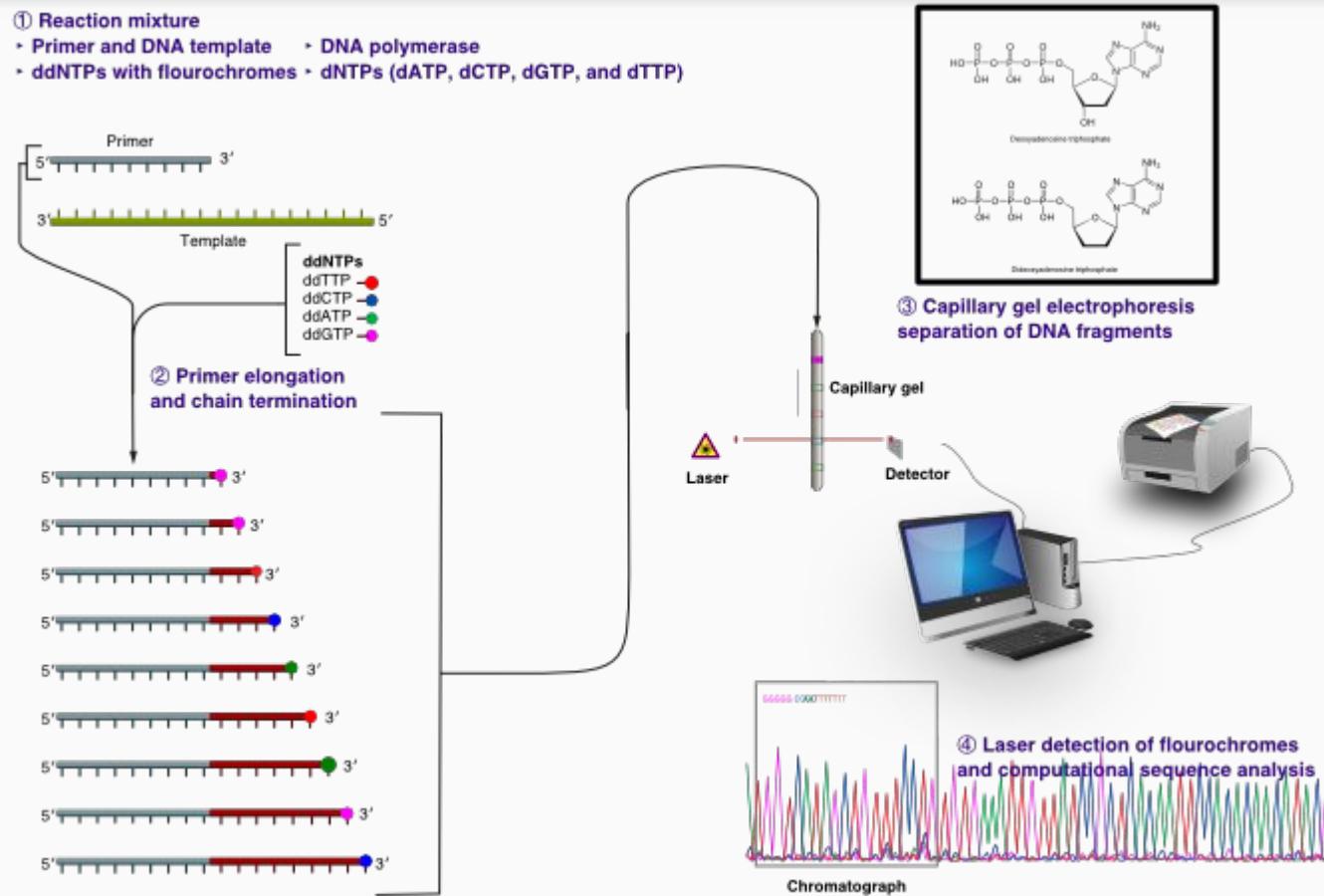
Why Sequence DNA?

- Determine sequence of:
 - Genes, Gene clusters or pathways
 - Whole Chromosomes
 - Whole Exomes
 - Entire Genomes
- Studying the Genome enables:
 - Identify new genes & new drug targets
 - Look for association of genes/mutations with diseases and phenotypes
 - **Personalised/Stratified medicine & Clinical Diagnostics**
 - Study Evolutionary biology - relationship between species
 - Plants & Agriculture: Breeding and Improved strains
 - Microbes, Viruses and Infectious diseases
 - Environmental Genomics & Metagenomics - id species present in water, dirt, debris filtered from the air, or swab samples of organisms
 - Forensics, Paternity testing...

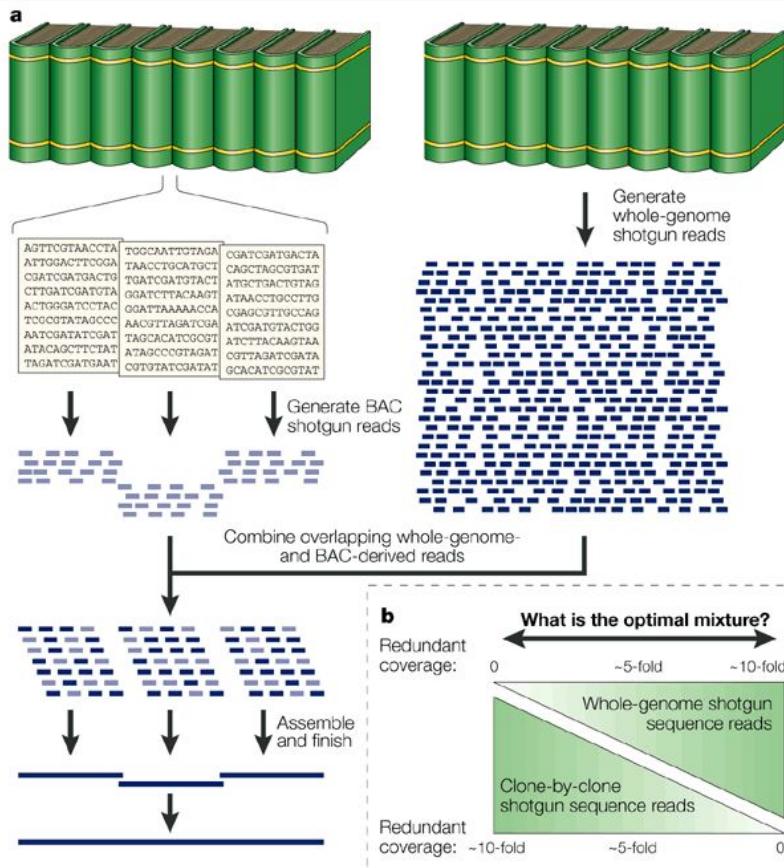
Sanger Sequencing: Chain Termination Method (the '90s)..



Sanger Sequencing: Chain Termination Method

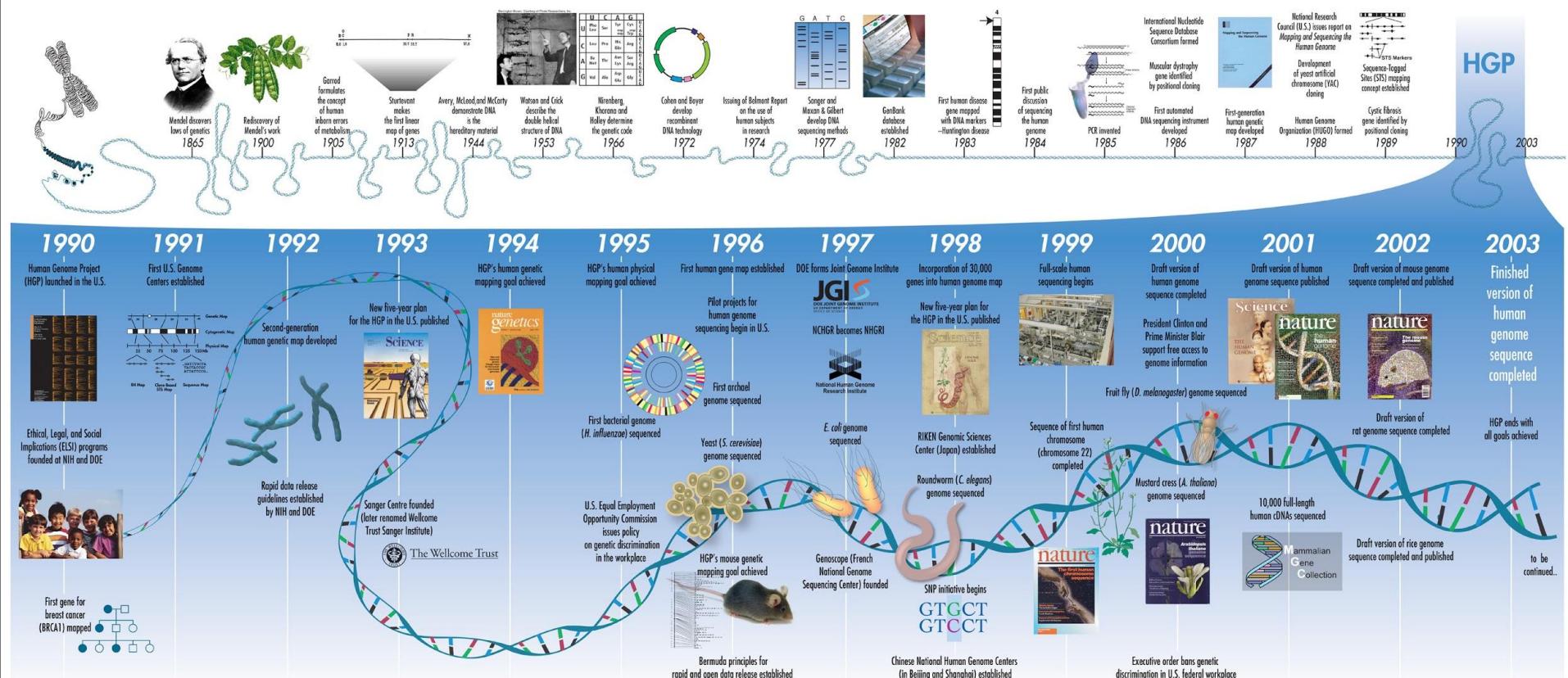


Shotgun-sequencing the human Genome



a | In a hybrid approach, elements of both clone-by-clone and whole-genome shotgun sequencing are amalgamated. A subclone library from a whole genome (represented by an encyclopaedia set) is prepared, and numerous sequence reads (depicted in dark blue) are generated in a genome-wide fashion. Meanwhile, individual mapped BACs are also subjected to shotgun sequencing. The BAC-derived sequence reads (depicted in light blue) can then be used to identify overlapping sequences in the larger collection of whole-genome-derived sequence reads, in essence reducing the complexity of the whole-genome shotgun data set to a series of individual BAC-sized bins. The combined set of sequence reads for each BAC can then be individually assembled and subjected to sequence finishing. **b** | Optimal balance of generating clone-by-clone versus whole-genome shotgun sequence reads in a hybrid sequencing strategy. In the hybrid shotgun-sequencing approach illustrated in **a**, sequence reads derived from individual clones and those generated in a genome-wide fashion are assembled together. Although an overall sequence redundancy of 8–10-fold is typically desired, the optimal mixture of clone-by-clone and whole-genome shotgun sequence reads is at present not well established. Current projects that are using a hybrid approach for sequencing the mouse, rat and zebrafish genomes (see Table 1) should provide valuable insight into this issue.

The Human Genome Project: 13 years in the making!



2. NGS Technology & Terminology

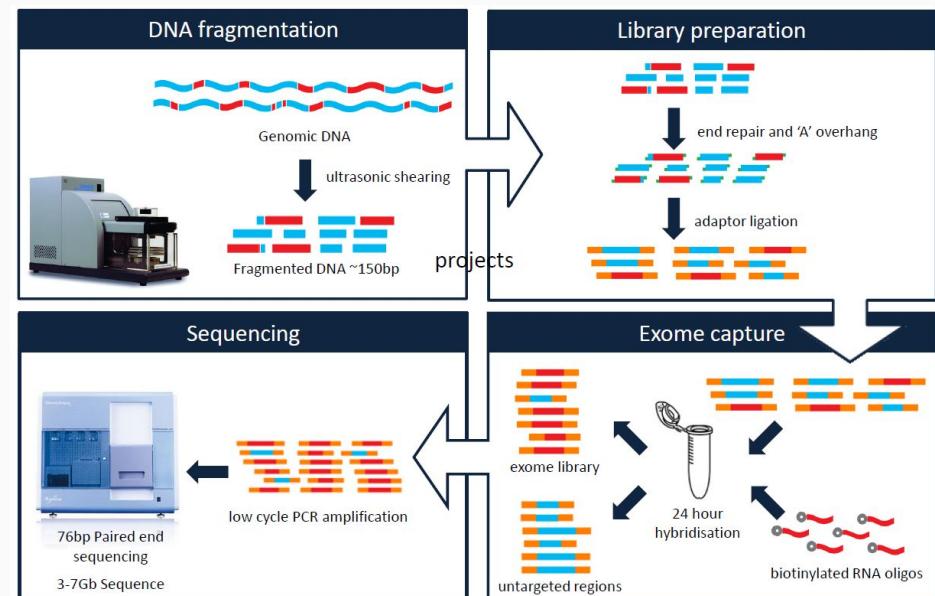
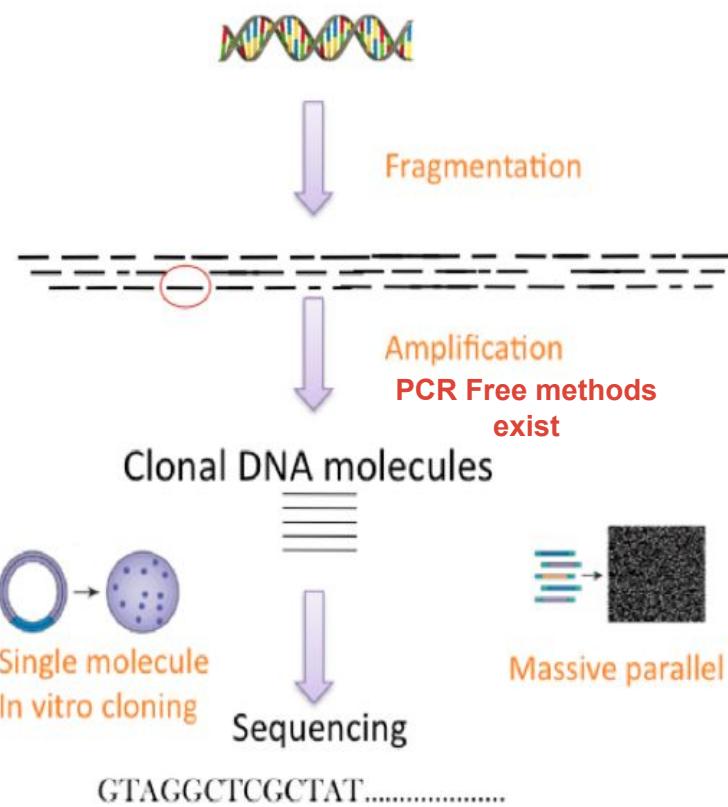
Next Generation Sequencing

- **Shotgun sequencing on steroids!**
- Cost effective, fast in comparison, Clone-free, generates millions of short reads that can be assembled and mapped to a reference genome (Human Genome Project)
- **A Human Genome Project now takes ~ 1 month vs 13 years!**
- NGS has brought high speed, high resolution molecular genetics to science and medicine and the benchtop!
- Will revolutionise medical diagnostics through rapid identification of mutant alleles that cause disease - **High hopes for Personalised Medicine!**

Terminology: Next Generation Sequencing

- **Second/Third Generation Sequencing**
 - Next Generation Sequencing (NGS)
 - High-Throughput DNA Sequencing (HTS)
 - Massively Parallel Sequencing (MPS)
 - **Illumina: sequencing by synthesis (SBS):**
 - <http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html>
- **First Generation Sequencing...big in the 90's**
 - Sanger Sequencing

How does Next Generation Sequencing work?



Thanks to Dr Mike Simpson (michael.simpson@kcl.ac.uk)

How does Next Generation Sequencing work?

- generates millions of short nucleotide sequences (reads)
- Massive amount of data!
- Raw files ~2-30GB x 2 per sample (Illumina)

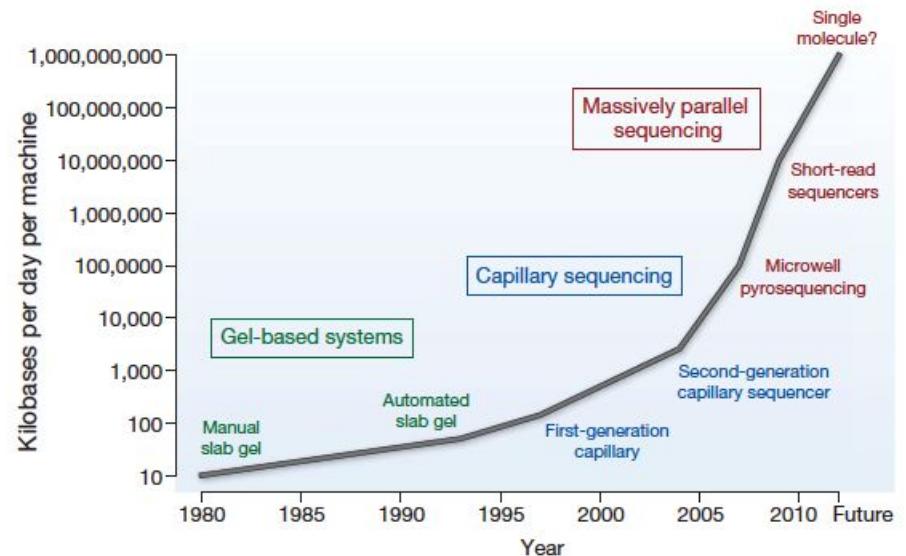
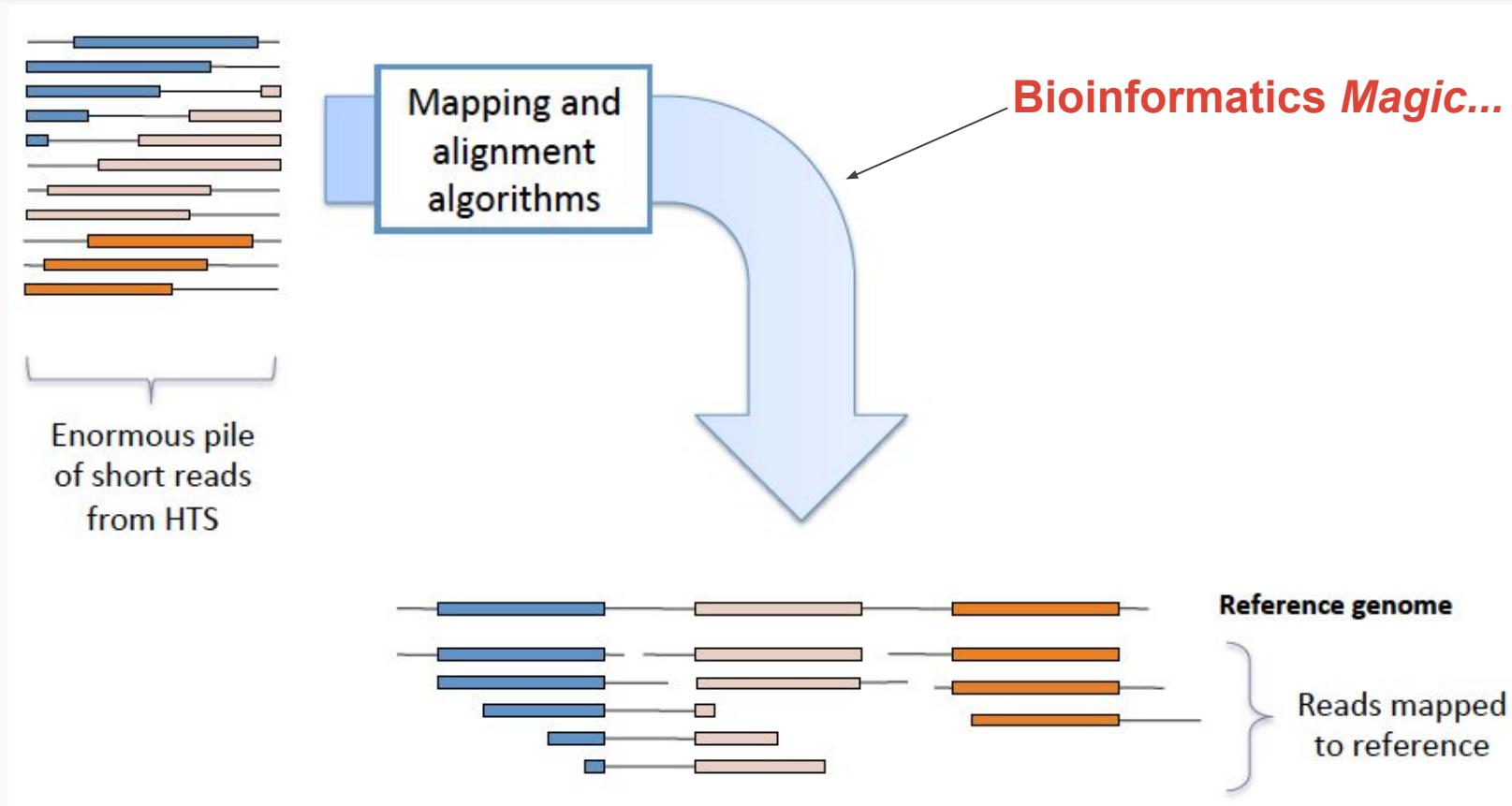


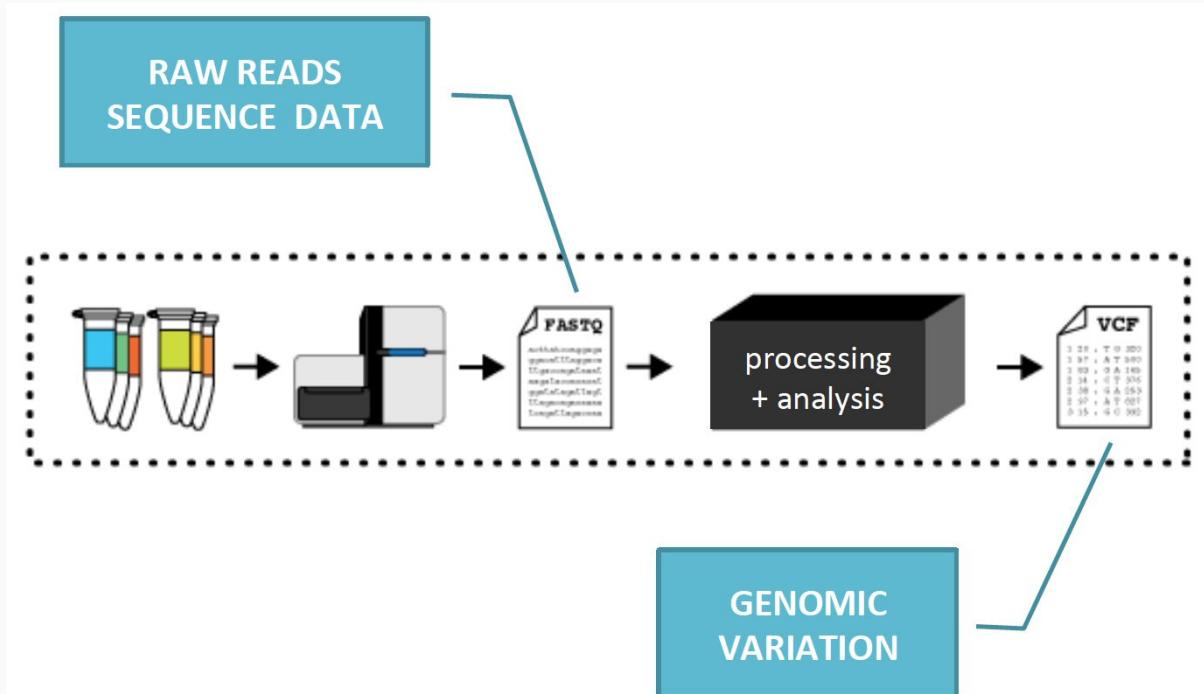
Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

How does Next Generation Sequencing work?



Genomic Medicine & NGS

- **The Ultimate goal is to detect clinically relevant variants from sequencing data**



Types of Variation Detected by NGS

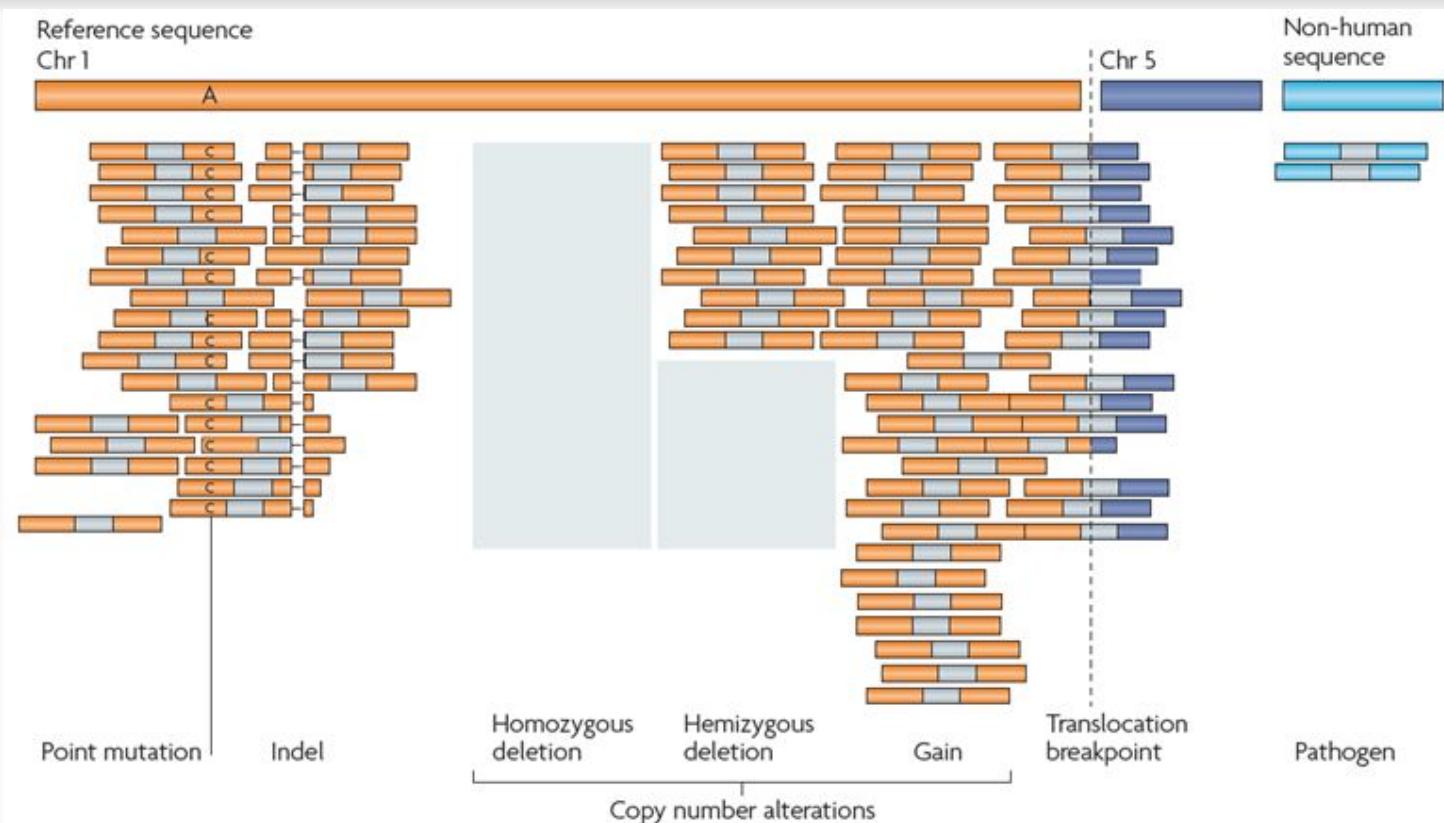
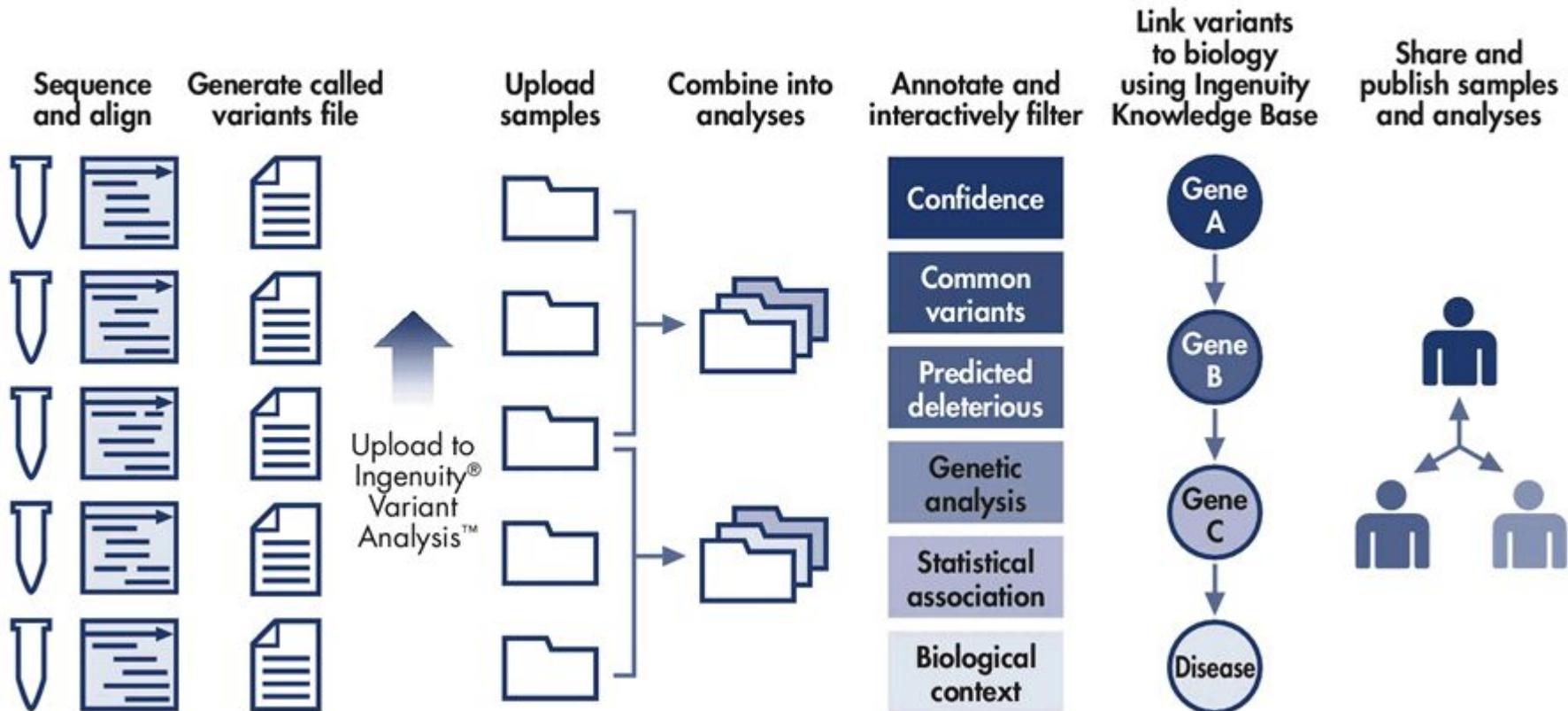


Figure 3 | Types of genome alterations that can be detected by second-generation sequencing. Sequenced

Overview: Interpretation of NGS...

Biological interpretation of human whole genome, exome, and targeted panel samples



NGS Technologies: Illumina & Others....



Genomics
England are using
illumina

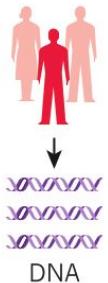
The DNA of A Nation

Vivien Marx

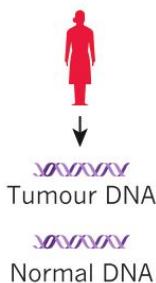
Nature 524, 503–505 (27 August 2015) doi:10.1038/524503a

Published online 26 August 2015

~50,000
people with rare
diseases and
their parents



~25,000
people with
cancer



THE CLINICAL GENOME

Genomics England plans to sequence 100,000 genomes by 2017. The genomic data will be crucial for diagnosing and treating disease, but its interpretation will require automated, specialized software.



RECRUITMENT OF 75,000 PEOPLE

The 100,000 Genomes Project is recruiting people with cancer and rare diseases. The genomes of both normal and tumour cells will be sequenced in people with cancer.

NEXT-GENERATION SEQUENCING

The Californian company Illumina will use UK-based high-throughput sequencing machines to produce whole-genome sequences and identify genetic variants.

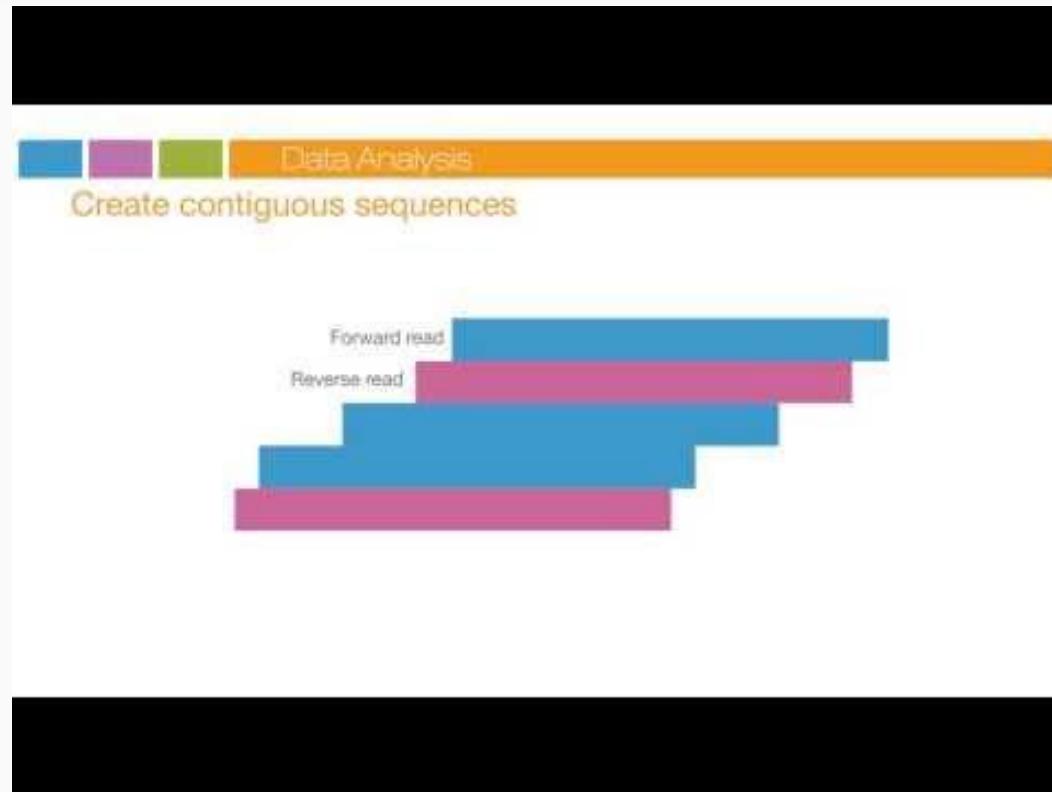
AUTOMATED INTERPRETATION

Four UK and US companies will use specialized software to automatically analyse the genetic variants that may be linked to disease.

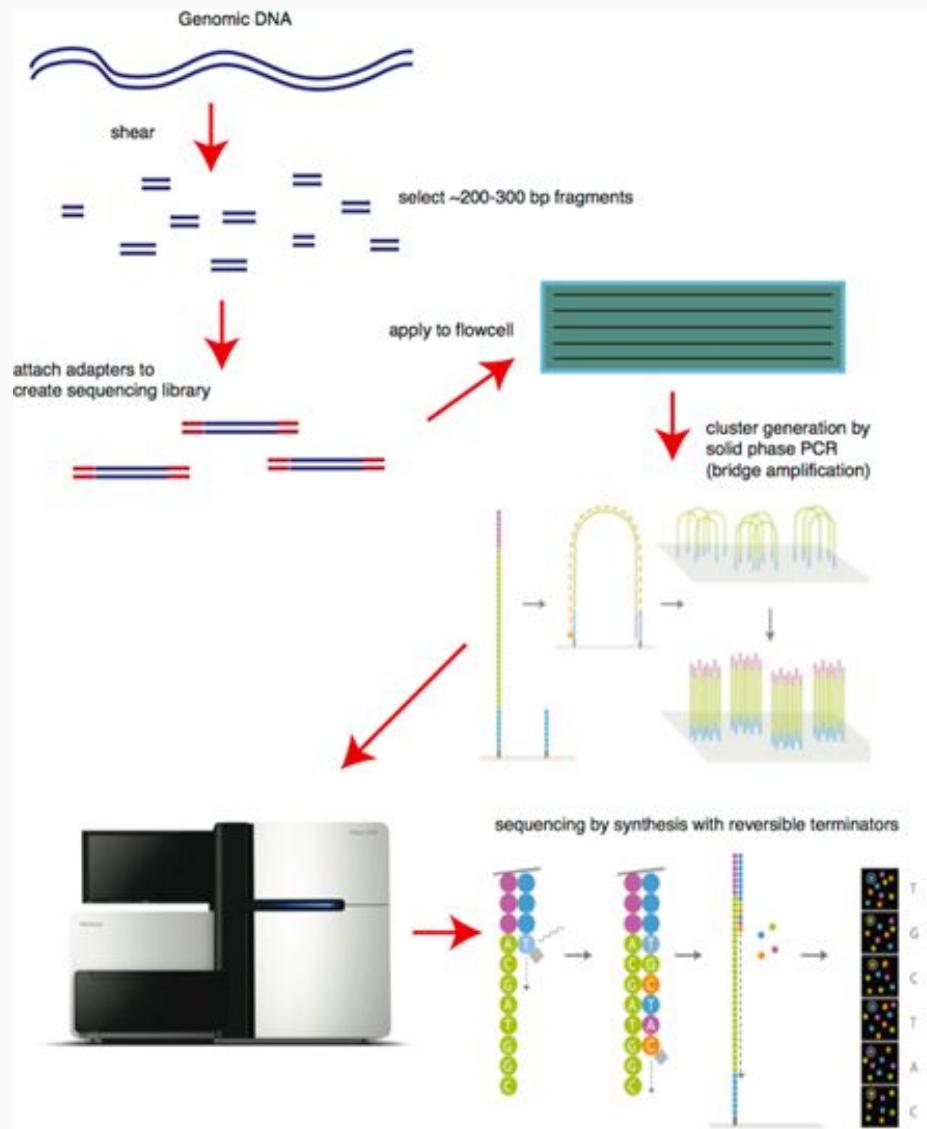
CLINICAL INTERPRETATION

Around 2,000 UK scientists and clinicians will pore over the data to validate or better understand how the variants may cause disease before the information is fed back to patients.

Illumina Sequencing Technology

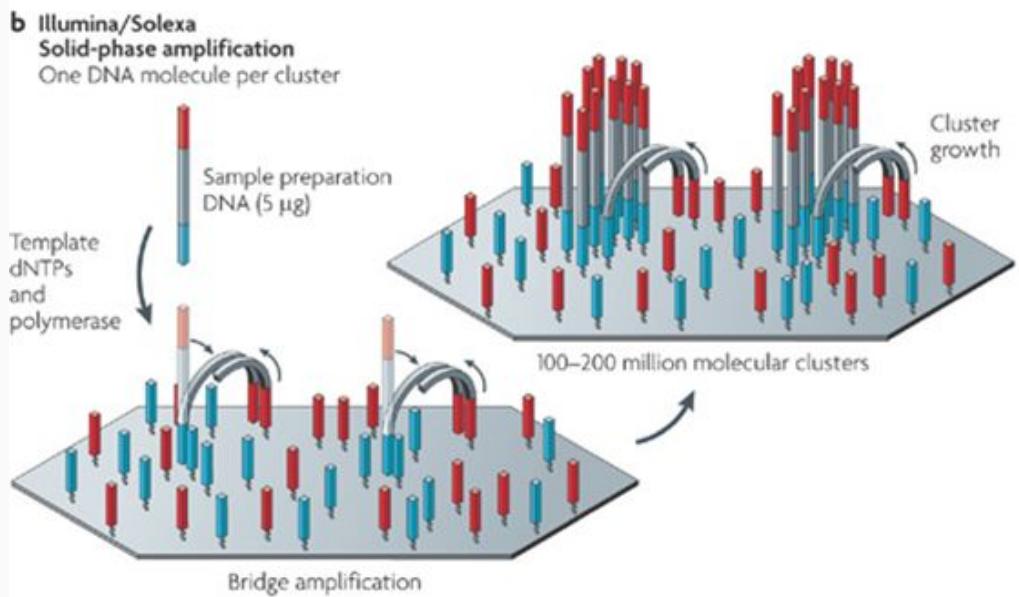


Illumina Sequencing Technology: Overview



- **Fragmentation and tagging of genomic/cDNA fragments** – provides universal primer allowing complex genomes to be amplified with common PCR primers
- **Template immobilization** – DNA separated into single strands and captured onto beads (1 DNA molecule/bead)
- **Clonal Amplification** – Solid Phase Amplification
- **Sequencing and Imaging** – Cyclic reversible termination (CRT) reaction : this is SBS

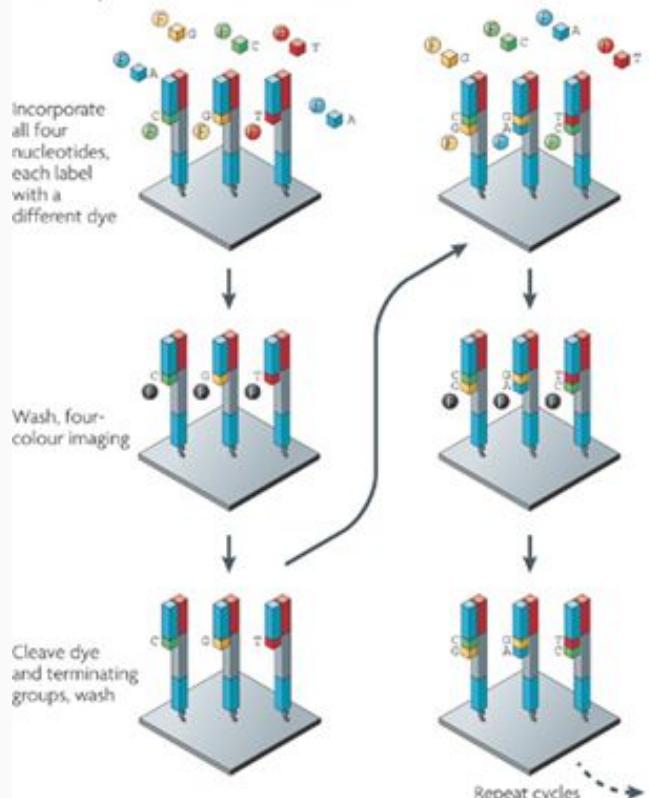
Illumina Sequencing Technology: Clonal Amplification



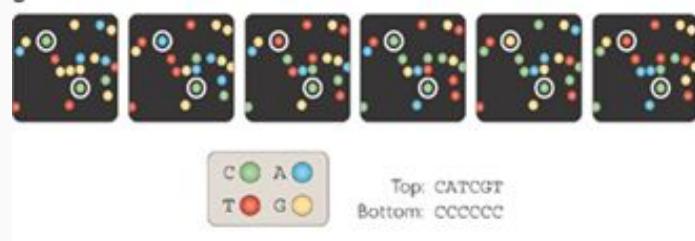
- Clonal Amplification – Solid Phase Amplification
- Priming and extension of single strand, single molecule template
- Bridge amplification of the immobilized template with primers to form clusters (creates 100-200 million spatially separated template clusters)
- Provides free ends to which a universal sequencing primer can be hybridized to initiate NGS reaction – each cluster represents a population of identical templates

Illumina Sequencing Technology : Cyclic Reversible Termination

a Illumina/Solexa — Reversible terminators

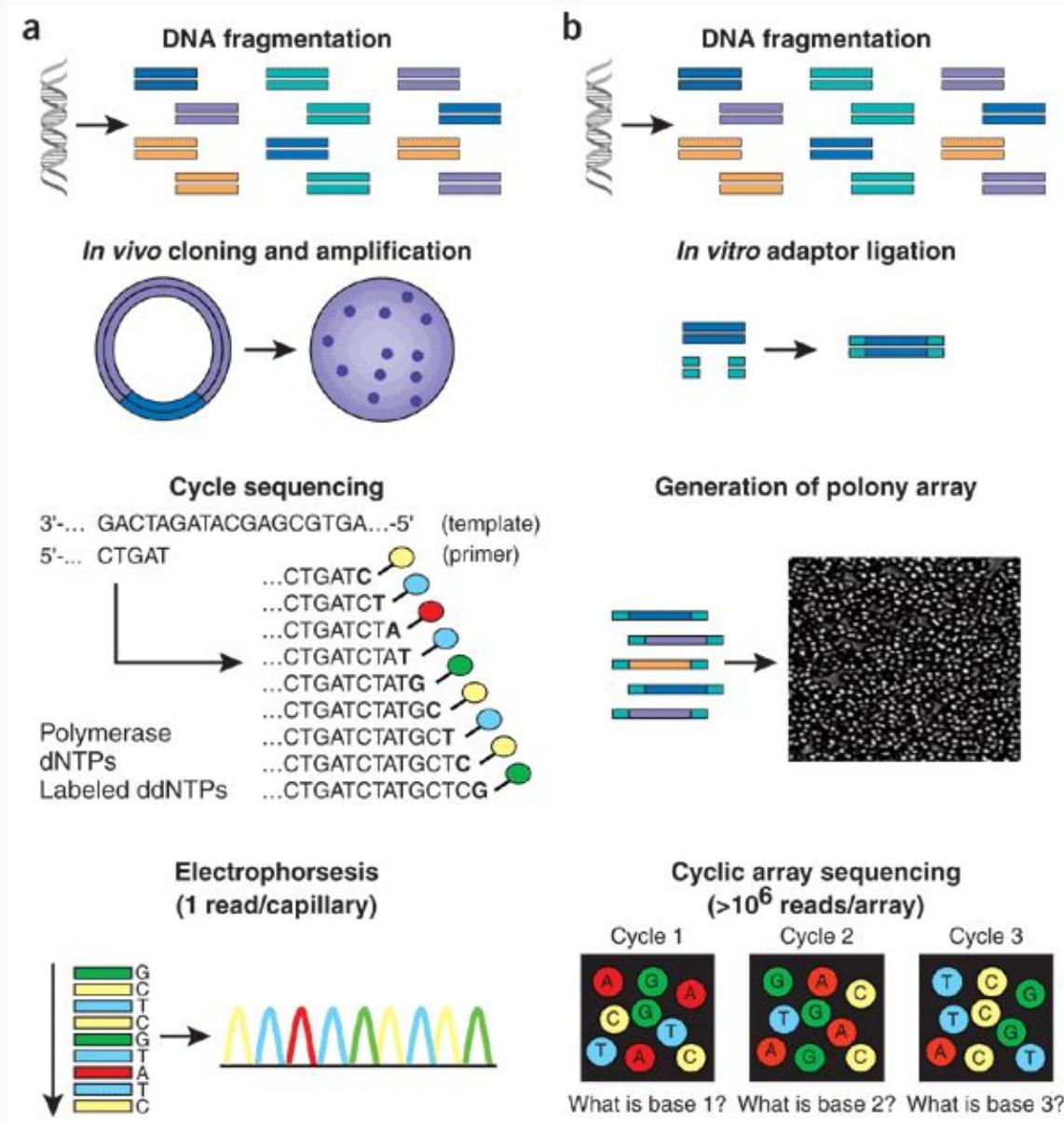


b



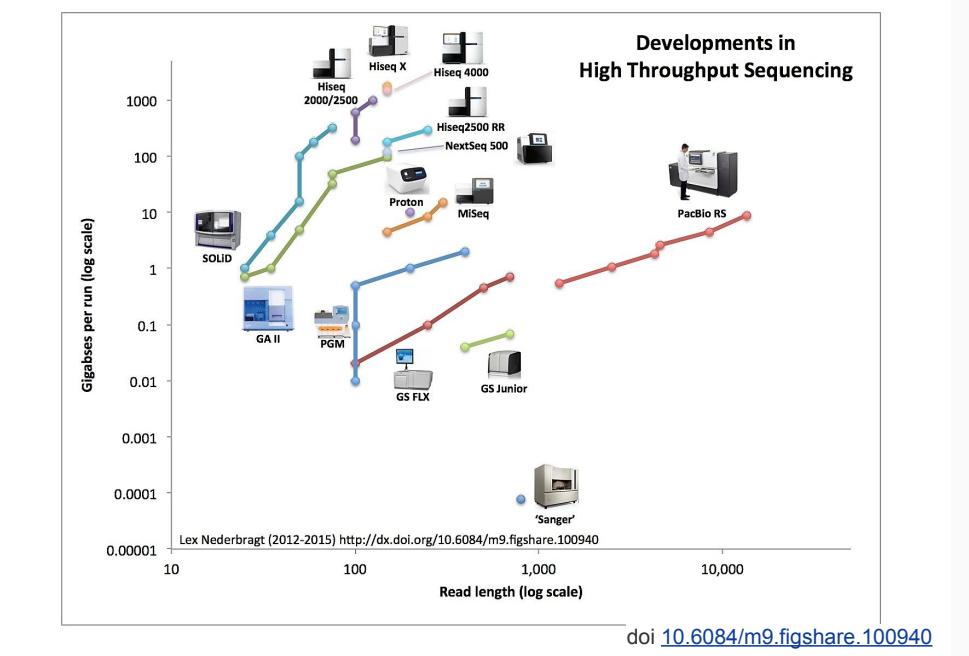
- **Cyclic Reversible Termination – DNA Polymerase bound to primed template adds 1 (of 4) fluorescently modified nucleotide. 3' terminator group prevents additional nucleotide incorporation.**
- Following incorporation, remaining unincorporated nucleotides are washed away. Imaging is performed to determine the identity of the incorporated nucleotide.
- Cleavage step then removes terminating group and the fluorescent dye. Additional wash is performed before starting next incorporation step
- This is repeated ~250 million times (25Gb) with HiSeq2500 (~4 days)

Sanger (*Shotgun*) vs NGS Sequencing: Overview



Terminology: Short/Long Reads

- A **Read**: a sequence of DNA generated (read) by NGS technology
- Short or Long refer to the length of the sequence
- **Short**: 35-150bp (Illumina)
- **Long**: 200bp - 1000's bp (eg Sanger & Pacbio)
- **Read Length**: length of read produced by machine
- *Reads are getting longer as technology improves & with new technologies*

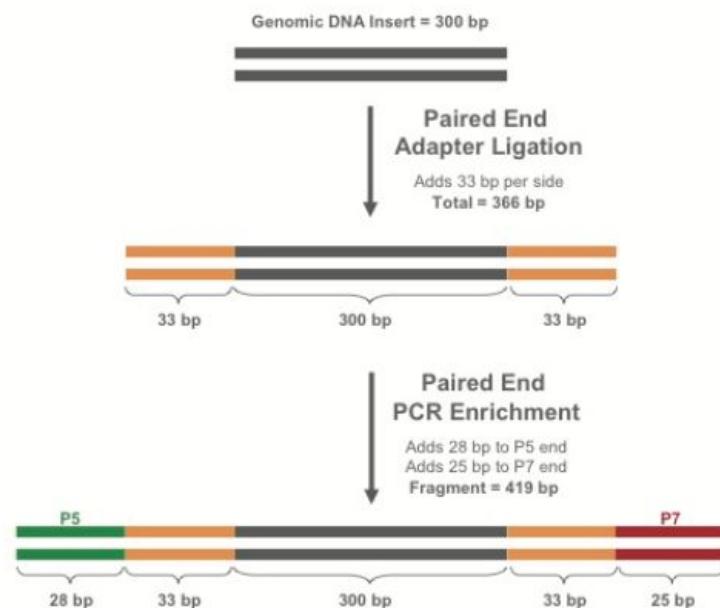


<https://github.com/lexnederbragt/developments-in-next-generation-sequencing>

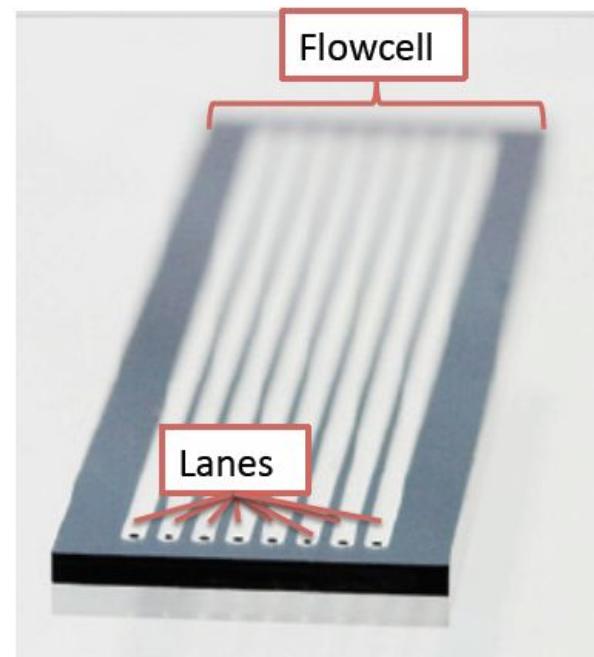
https://figshare.com/articles/developments_in_NGS/100940

<https://fixlexblog.wordpress.com/2015/06/17/developments-in-high-throughput-sequencing-june-2015-edition/>

Terminology: Libraries, Lanes and Flowcells



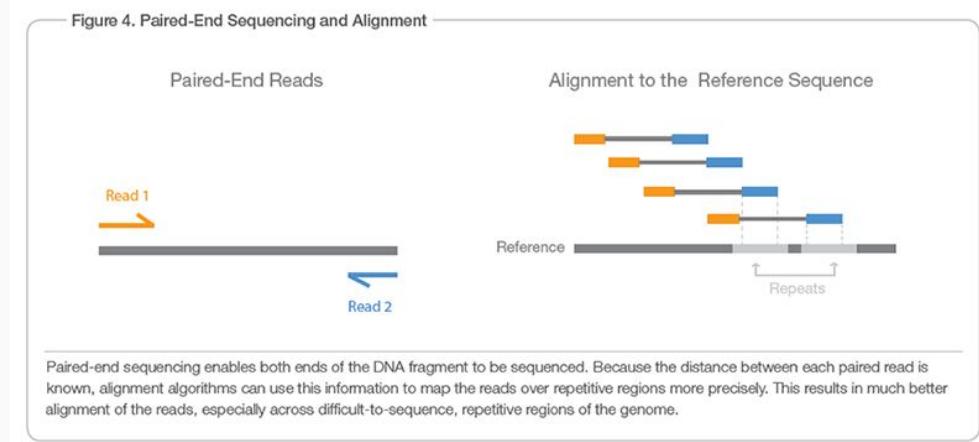
Each reaction produces a unique **library** of DNA fragments for sequencing.



Each NGS machine processes a single **flowcell** containing several independent **lanes** during a single sequencing run

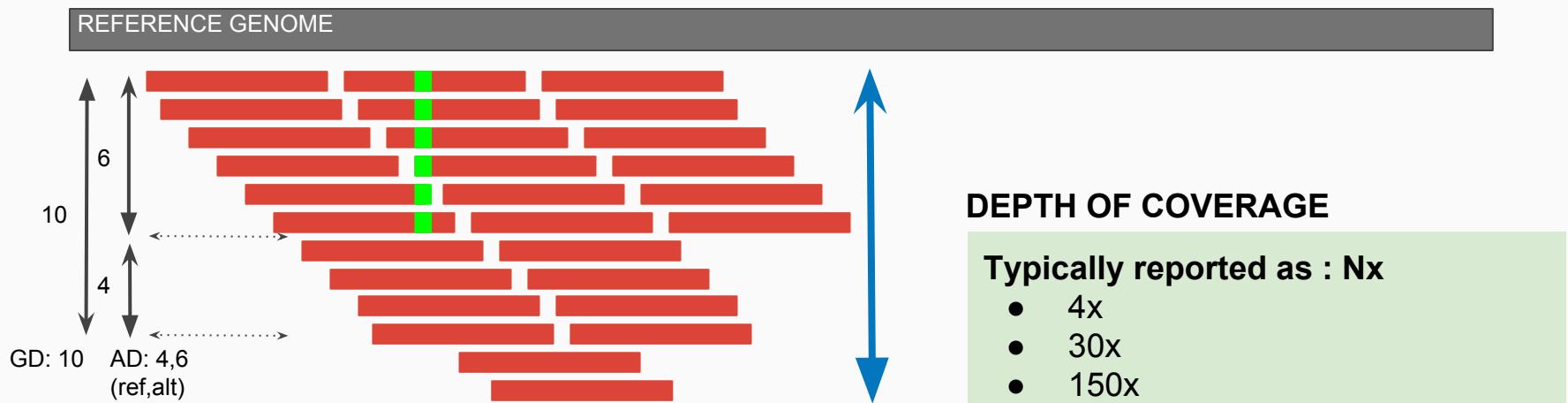
Terminology : Single End (SE) vs Paired End (PE)

- ❑ **Single End:** one read sequenced from one end of each sample DNA insert (Rd1 SP: Read 1 Sequencing Primer)
- ❑ **Paired End:** two reads (one from each end) sequenced from each sample DNA insert (Rd1 and Rd2 sequencing primer)
 - ❑ Akin to F and R reads in Sanger sequencing
 - ❑ More robust Variant Calls
 - ❑ Provides better resolution of INDELS and Copy-Number Variants



Terminology: Depth of Coverage

- **Depth of Coverage:** Number of reads that align to, or "cover," known reference bases.
 - Often reported as an **average** over the whole length of the genome, chromosome, exome...
- **Genotype Depth:** Number of reads at observed **Variant Call** (eg GD:10)
- **Allele Depth:** Number of reads for each observed alleles (eg AD: 4,6)
- Coverage can also refer to the [% region] of genome targeted/sequenced

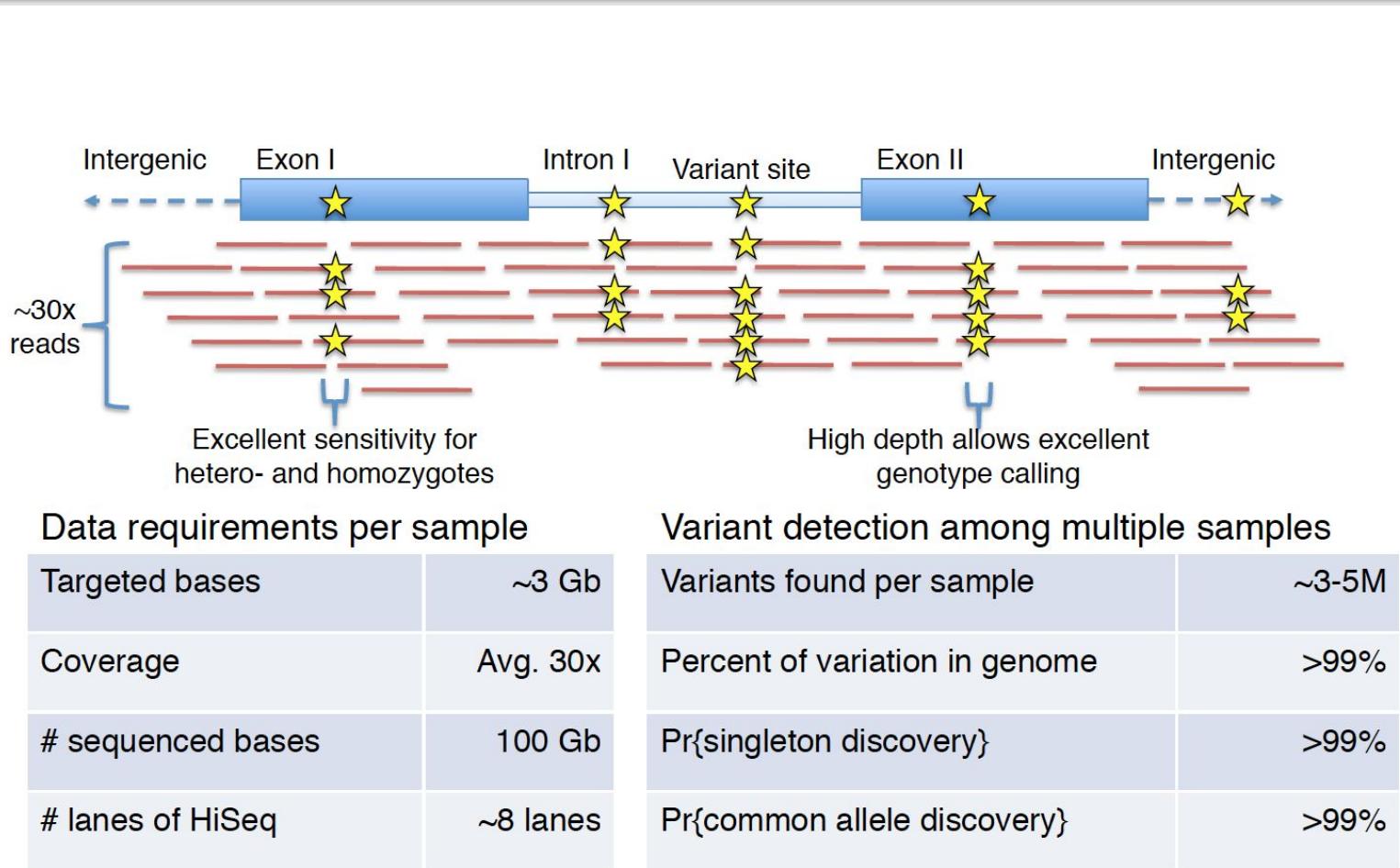


3. Experimental Design

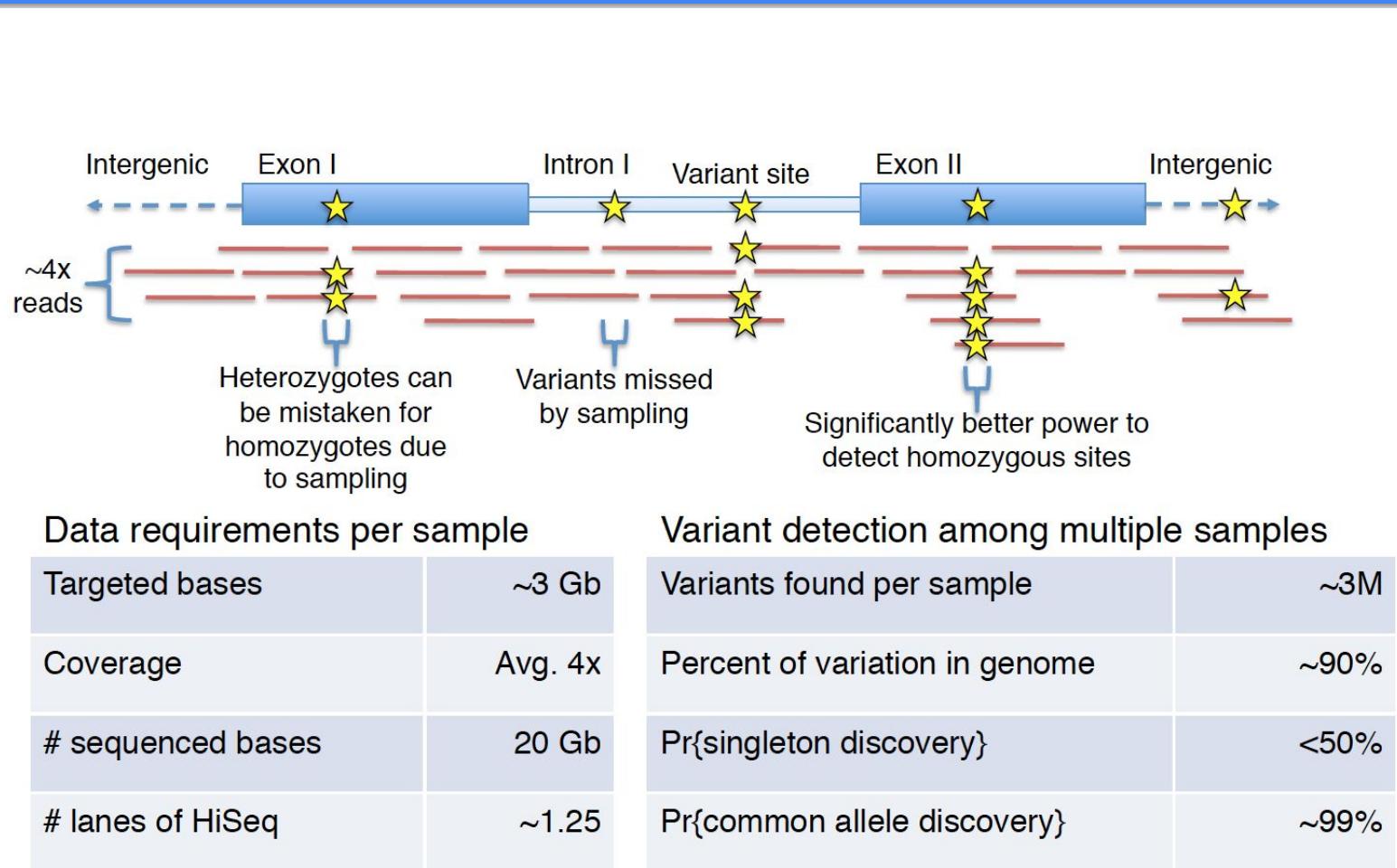
NGS Design Terminology : Some Key Terms

- High depth/coverage (Deep Coverage)
- Low depth/coverage (Shallow Coverage)
- Targeted(Gene Panel) v Exome v Whole Genome
- Single sample v Multiple Sample
- Whole Genome (WGS,WG)
- Whole Exome (WES, WEX, WEx)
- Targeted or Panel (Clinically relevant genes)

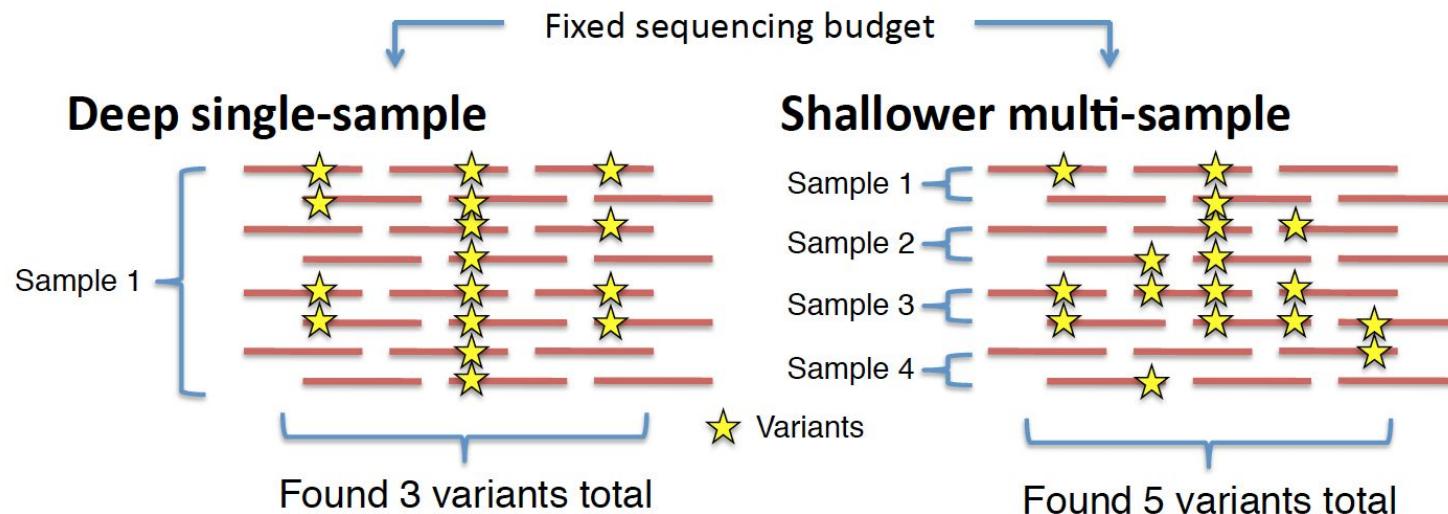
High Depth



Low Depth



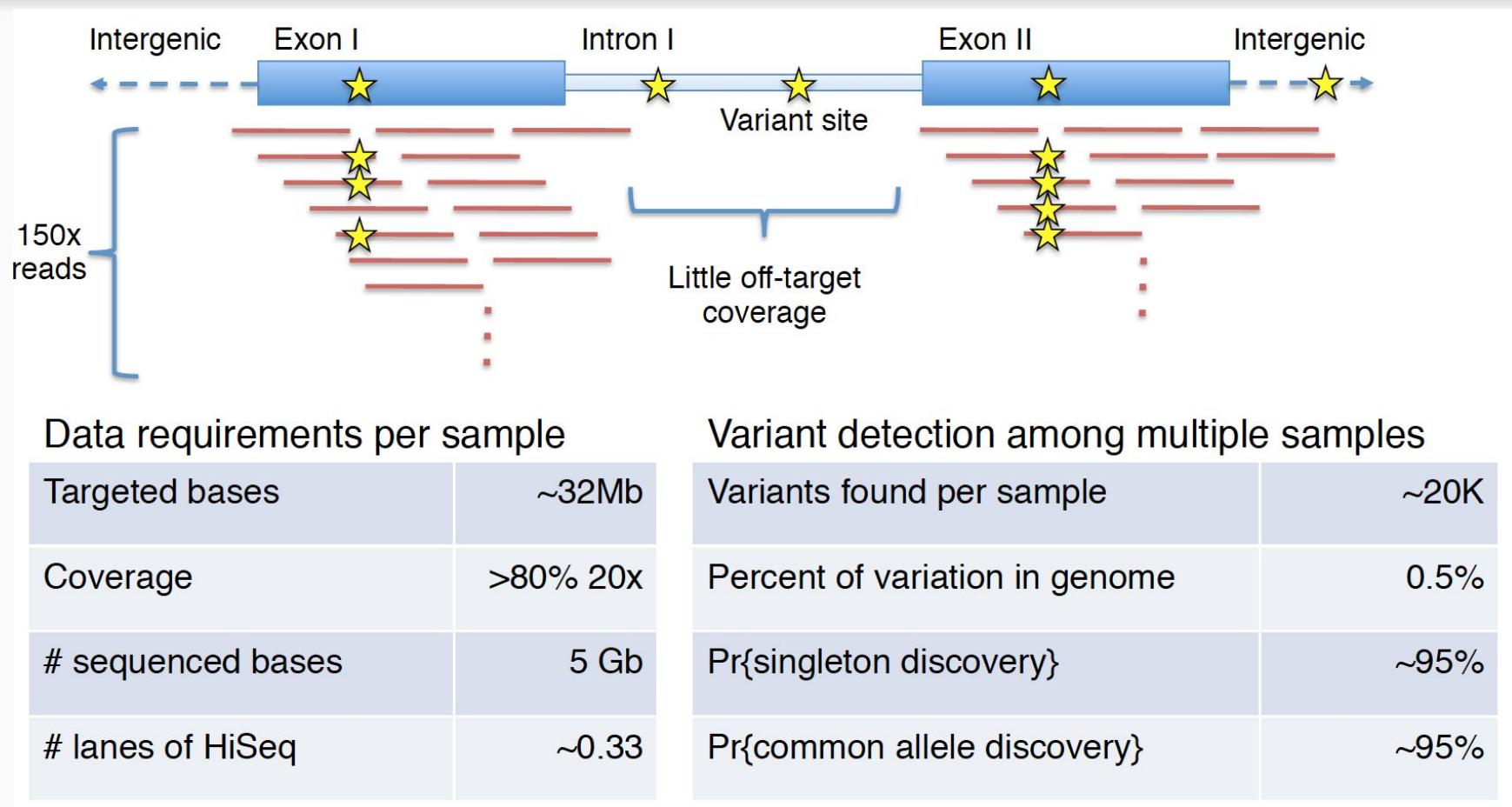
Single Sample v. Multiple Sample



- Higher sensitivity for variants in the sample
- More accurate genotyping per sample
- Cost: no information about other samples

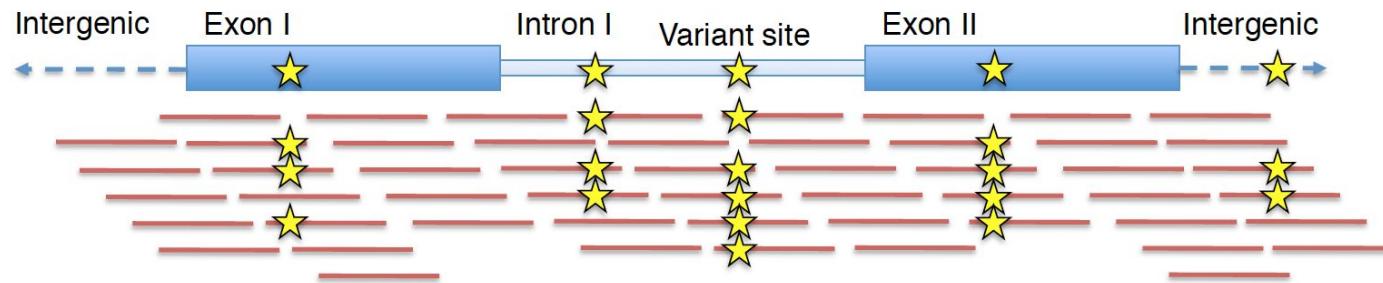
- Sensitivity dependent on frequency of variation
- Worse genotyping
- More total variants discovered

Exome Capture

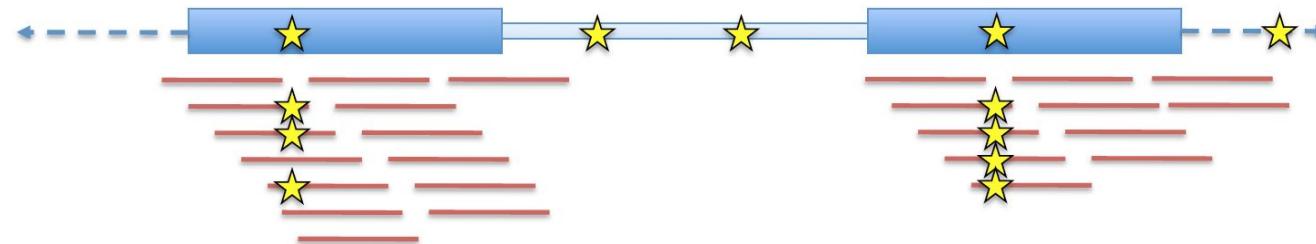


Whole Genome v Exome (Panel/Targeted) WGS v WEx

Whole genome



Exome



Small|targeted experiments, gene panels, RADseq

restriction site-associated DNA sequencing

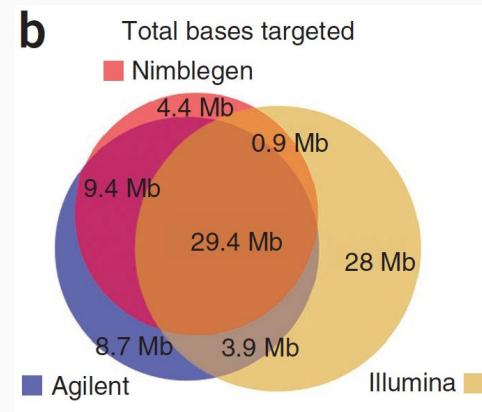
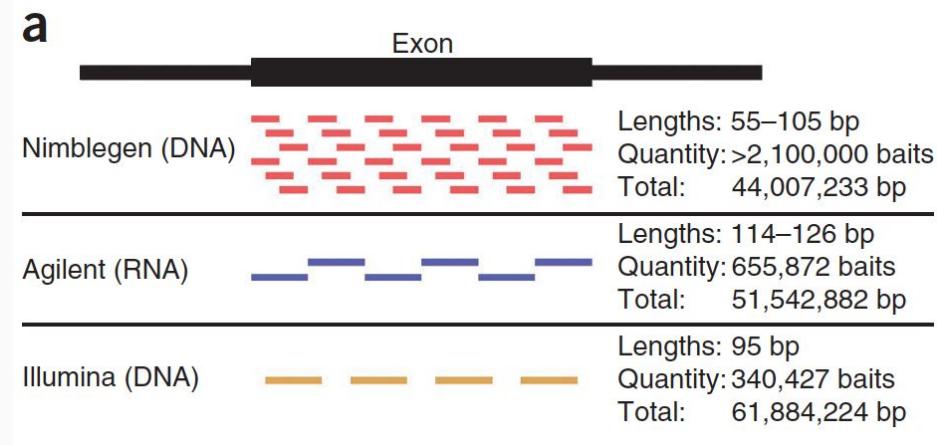
- Similar to exomes for most purposes

Key Differences between Whole Genome and Exome

- **Whole Genome**
 - Entire genome is prepared
 - PCR free methods possible
 - Higher price
 - Exons + Introns
 - Coverage: Entire Genome
 - Almost: some missing
 - some not “accessible”
- **Exome or Targeted Gene Panel**
 - Capture of target regions
 - PCR Amplification often required
 - Lower price
 - Exons only (+/- adjacent)
 - Coverage: targets only
 - *for regions where we already know the sequence!*
- **Genome & Targeted...**
 - *We will always miss something*
 - *Never 100% coverage*
 - *based on our current knowledge: the current reference genome is not perfect or complete*

Targeted and Exome Capture Regions

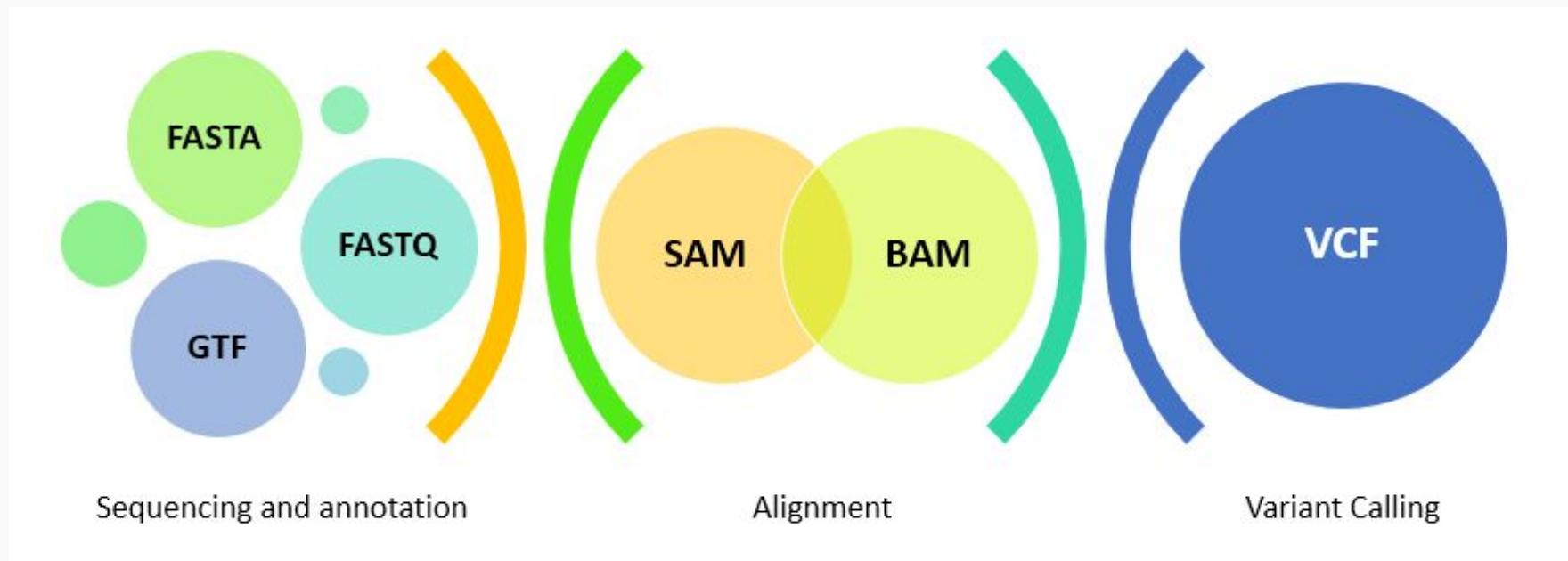
- Involves **BAITS**: complementary sequences designed to “capture” segments of target DNA
- **Resulting covered intervals are specific to capture kit manufacturer**
- **Capture/Enrichment designs include different**
 - biochemical methods
 - bait lengths
 - quality and overlap of baits
 - number of DNA bases targeted



4. Data Formats

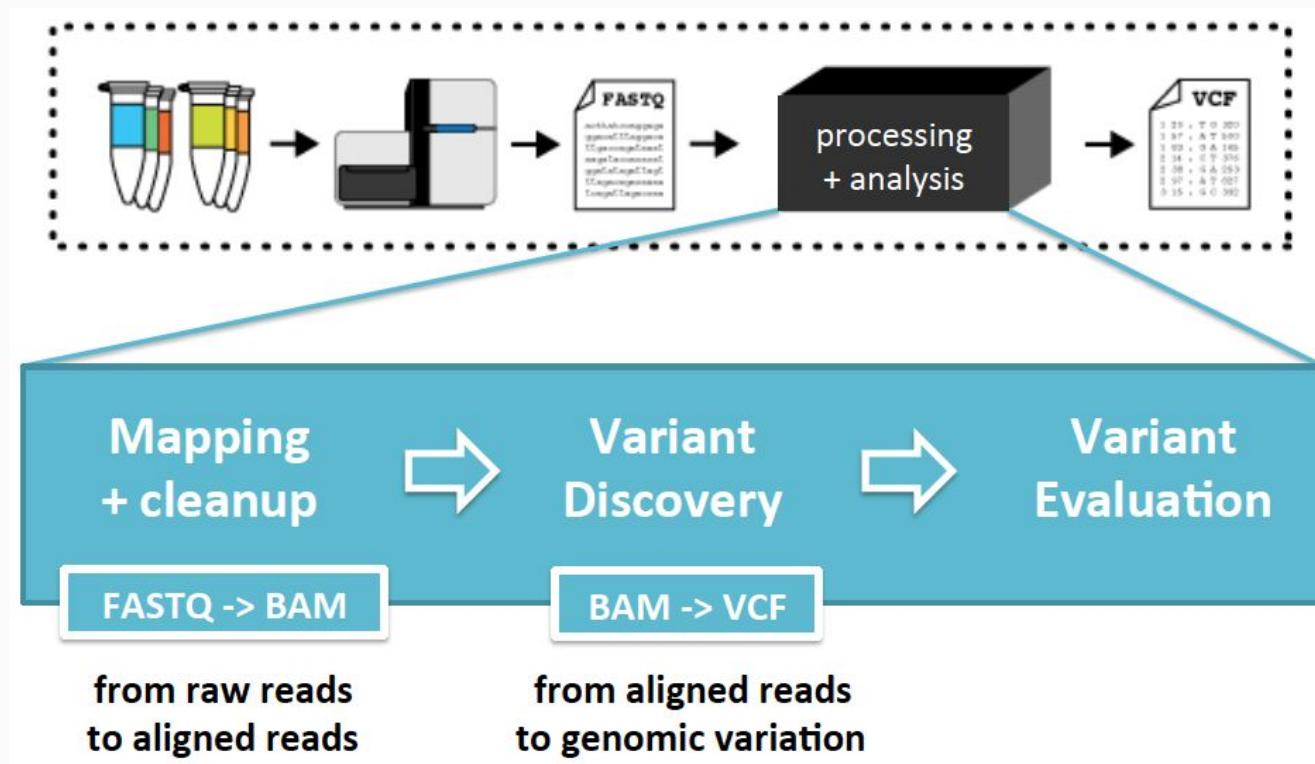
NGS File Formats

- Most common file formats for Next Generation Sequence Analysis
 - <https://docs.cyfronet.pl/display/PLGDoc/Most+common+file+formats+for+Next+Generation+Analysis>



NGS File Formats: from reads to variants

Major steps involve transforming data and storing results in **SPECIFIC FORMATS**



FASTA Format

- Standard for displaying (nucleotide or protein) sequences in a text file.
- An entry for a sequence takes up two lines in the file:
 - The first line begins with a ">" symbol, followed by the sequence description
 - The second line contains the sequence itself.
 - A sequence file in can contain several FASTA sequences.

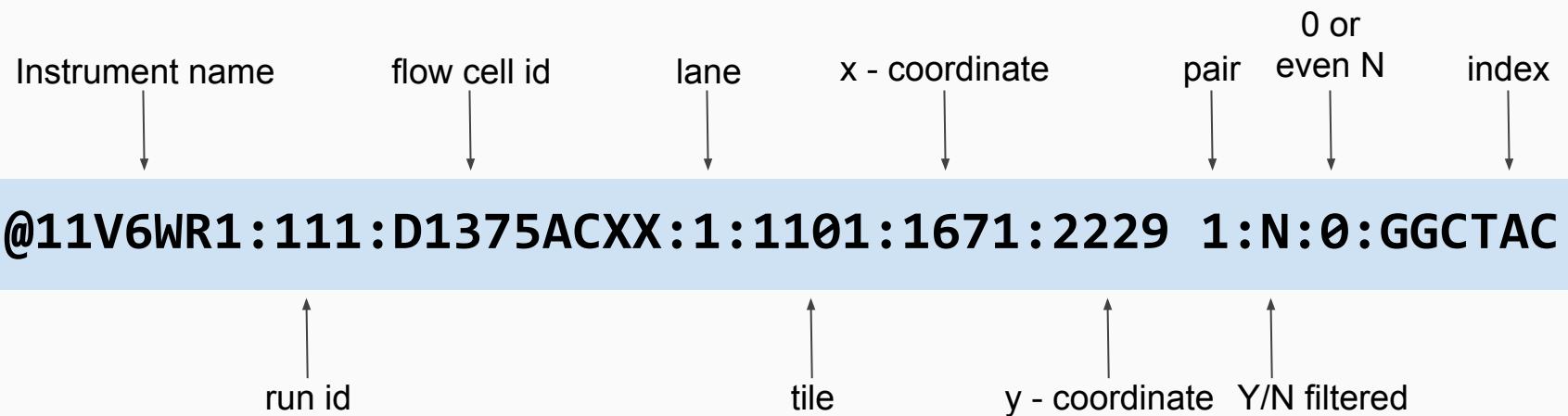
```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
GTACGACGGAGTGTATAAGATGGAAATCGGATACCAAGATGAAATTGTGGATCAGGTGCAAAAGTCGGC
AGATATCGTTGAAGTCATAGGTGATTATGTTCAATTAAAGAACAGCAAGGCCGAAACTACTTGACTCTGT
CCTTTCATGGAGAAAGCACACCTCGTTCCGTATGCCCGACAAACAGATTTTCATTGCTTGGCT
GCGGAGCGGGCGGCAATGTTCTTTTAAGGCAGATGGAAGGCTATTCTTGCCAGTCGGTTTC
>2 dna:chromosome chromosome:GRCh37:2:1:243199373:1
TCTTCTGGAGAACAAAAATGGCTGAGGCACATGAGCTCCTGAAGAAATTTACCATCATTGTTAATAA
ATACAAAAGAAGGTCAAGAGGCCTGGATTATCTGCTTCTAGGGGCTTACGAAAGAGCTGATTAATGA
ATTCAGATTGGCTATGCTTGTGATTCTGGACTTATCACGAAATTCTGTAAAGAGGGGATTAGT
GAGGCGCAAATGGAAAAGCAGGGCTCCTGATCAGACGCGAAGACGGAAGCGGATATTCGACCGCTCA
```

FASTQ Format

- Raw Reads
- DNA sequence and PHRED Quality Score in a single text file
- 4 lines per sequence:
 - Line 1 - begins with a '@' character and is followed by a sequence identifier and an optional description (like a sequence description),
 - Line 2 - is the raw sequence letters,
 - Line 3 - begins with a '+' character,
 - Line 4 - encodes the quality values for the sequence in Line 2. **Base Qualities (ASCII 33 + Phred scaled Q)**

```
@1V6WR1:111:D1375ACXX:1:1101:1671:2229 1:N:0:GGCTAC ← Identifier
TTTATAAAAAATTTCACATGTGACTAACATTACTCACATTGTTGAGATGGGGATAAAAACAGAAGGAACATAACCTT ← sequence
+
=?@DAB22A?FDBBGHF@E@EAAIG9CEFH9CHHIEFI3CFCHBFDFHDFCGB6??8*?ABG)8CC(7=2=)7=?>?E ← base qualities
```

Illumina Sequence Identifiers



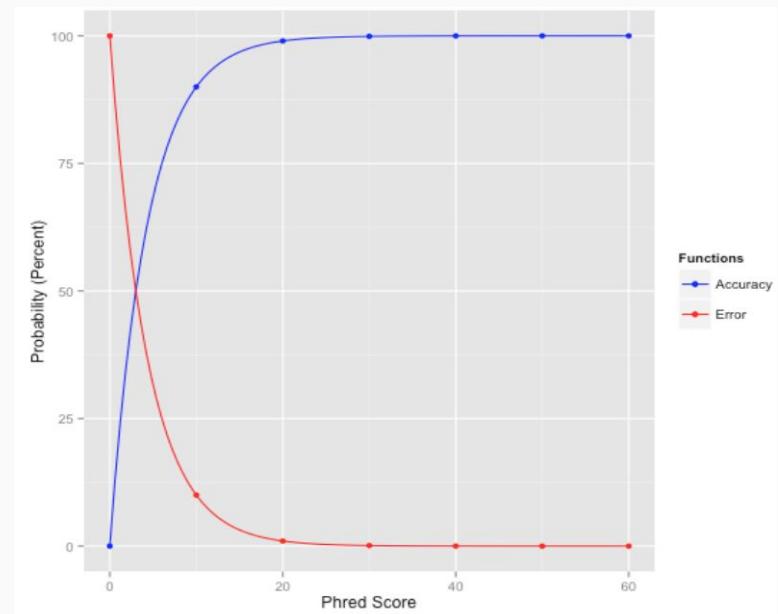
More details at: https://en.wikipedia.org/wiki/FASTQ_format

Phred Scores

- Phred value or Q score : Probability that the corresponding base call is incorrect
- $Q = -10 * \log_{10}(e)$
 - e: probability of a base being wrong
- Examples
 - Q10 = 90% Confident = 10% error rate
 - Q30 = 99.9% Confident = 0.1% error rate

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

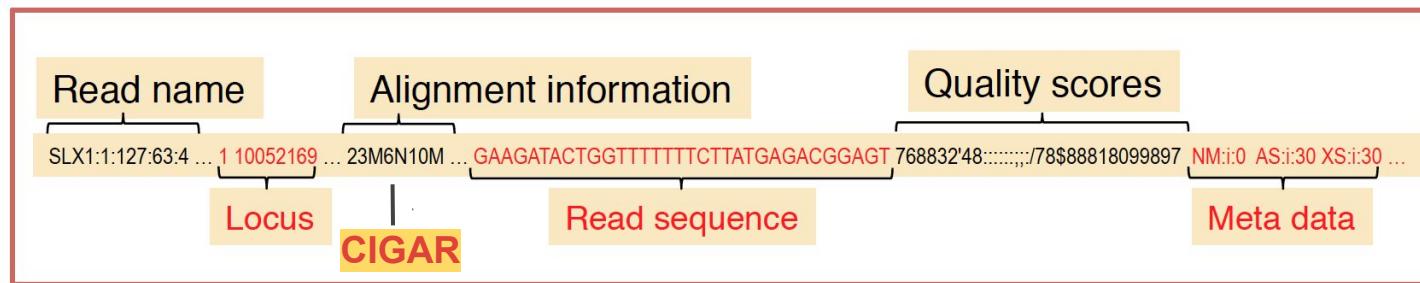


SAM & BAM Format

Official Specification : <https://samtools.github.io/hts-specs/SAMv1.pdf>

- Text format for storing information about aligned reads in a series of tab delimited ASCII columns

Sequence Alignment Map / Binary Alignment Map (compressed)



BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

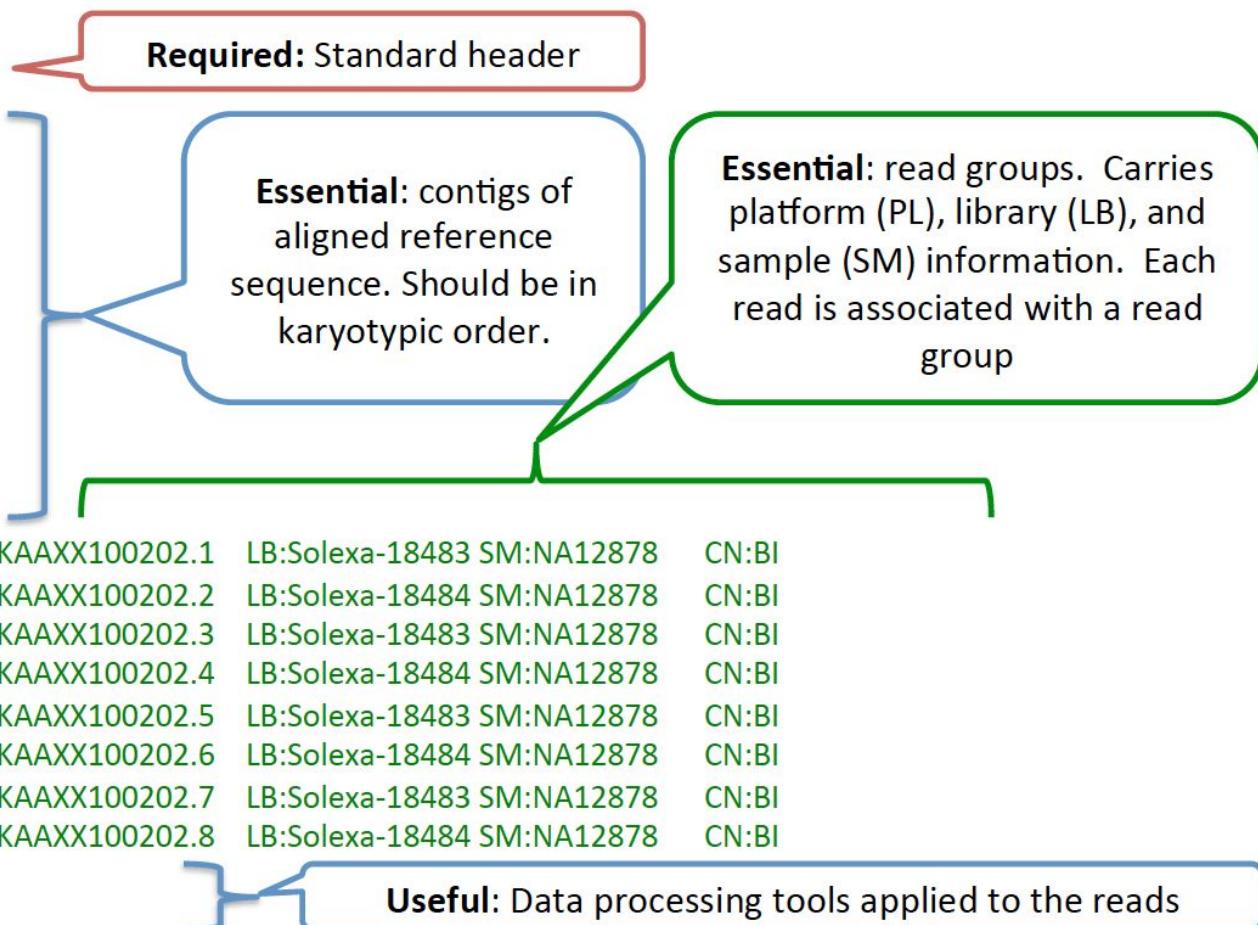
-> technology-independent

Data processing and analysis

A BAM file can contain data from a single or from several samples

BAM Headers: An essential part of a BAM file

```
@HD VN:1.0 GO:none SO:coordinate  
@SQ SN:chrM LN:16571  
@SQ SN:chr1 LN:247249719  
@SQ SN:chr2 LN:242951149  
[cut for clarity]  
@SQ SN:chr9 LN:140273252  
@SQ SN:chr10 LN:135374737  
@SQ SN:chr11 LN:134452384  
[cut for clarity]  
@SQ SN:chr22 LN:49691432  
@SQ SN:chrX LN:154913754  
@SQ SN:chrY LN:57772954  
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI  
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI  
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI  
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI  
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI  
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI  
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI  
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI  
@PG ID:BWA VN:0.5.7 CL:tk  
@PG ID:GATK PrintReads VN:1.0.2864
```



20FUKAAXX100202:1:1:12730:189900 163 chrM 1 60 101M = 282 381
GATCACAGGTCTATCACCTATTAAACCACTCACGGGAGCTTCCATGCATTGGTA...[more bases]
?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCAB...[more quals]
RG:Z:20FUK.1 NM:i:1 AM:i:37 MD:Z:72G28 MQ:i:60 PG:Z:BWA UQ:i:33

The CIGAR String: describes how a read aligns to the reference

The sequence being aligned to a reference may have additional bases that are not in the reference or may be missing bases that are in the reference. The CIGAR string is a sequence of base lengths and the associated operation. They are used to indicate things like which bases align (either a match/mismatch) with the reference, are deleted from the reference, and are insertions that are not in the reference

Operator Description

- **D** Deletion; the nucleotide is present in the read but not in the reference
- **H** Hard Clipping; the clipped nucleotides are not present in the read.
- **I** Insertion; the nucleotide is present in the reference but not in the read
- **M** Match; can be either an alignment match or mismatch. The nucleotide is present in the reference.
- **N** Skipped region; a region of nucleotides is not present in the reference
- **P** Padding; padded area in the read and not in the reference
- **S** Soft Clipping; the clipped nucleotides are present in the read
- **X** Read Mismatch; the nucleotide is present in the reference
- **=** Read Match; the nucleotide is present in the reference

The CIGAR String: An Example

http://genome.sph.umich.edu/wiki/SAM#What_is_a_CIGAR.3F

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:	ACTAGAATGGCT																		

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Alignment:					<u>A</u>	<u>C</u>	<u>T</u>	<u>A</u>	<u>G</u>	<u>A</u>	<u>A</u>		<u>T</u>	<u>G</u>	<u>G</u>	<u>C</u>	<u>T</u>		

CIGAR: 3M 1I 3M 1D 5M

POS: 5

CIGAR: 3M1I3M1D5M

The **POS** indicates that the read **aligns starting at position 5** on the reference. The CIGAR says that the **first 3 bases in the read sequence align with the reference**. The next base in the read does not exist in the reference. Then 3 bases align with the reference. The next reference base does not exist in the read sequence, then 5 more bases align with the reference. **Note that at position 14, the base in the read is different than the reference, but it still counts as an M since it aligns to that position.**

VCF Format: Stores variant information

<https://vcftools.github.io/specs.html>

Example

```

##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1>Type=String>Description="Ancestral Allele">
##INFO=<ID=H2,Number=0>Type=Flag>Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1>Type=String>Description="Genotype">
##FORMAT=<ID=GQ,Number=1>Type=Integer>Description="Genotype Quality (phred score)">
##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="Read Depth">
##ALT=<ID=DEL>Description="Deletion">
##INFO=<ID=SVTYPE,Number=1>Type=String>Description="Type of structural variant">
##INFO=<ID=END,Number=1>Type=Integer>Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT .
1 2 rs1 C T,CT .
1 5 . A G .
1 100 . T <DEL> .
. . PASS .
. H2;AA=T .
. PASS .
. PASS .
. SVTYPE=DEL;END=300 GT:DP 1/2:13 0/0:29
. GT:GQ 0|1:100 2/2:70
. GT:GQ 1|0:77 1/1:95
. GT:GQ:DP 1/1:12:3 0/0:20

```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

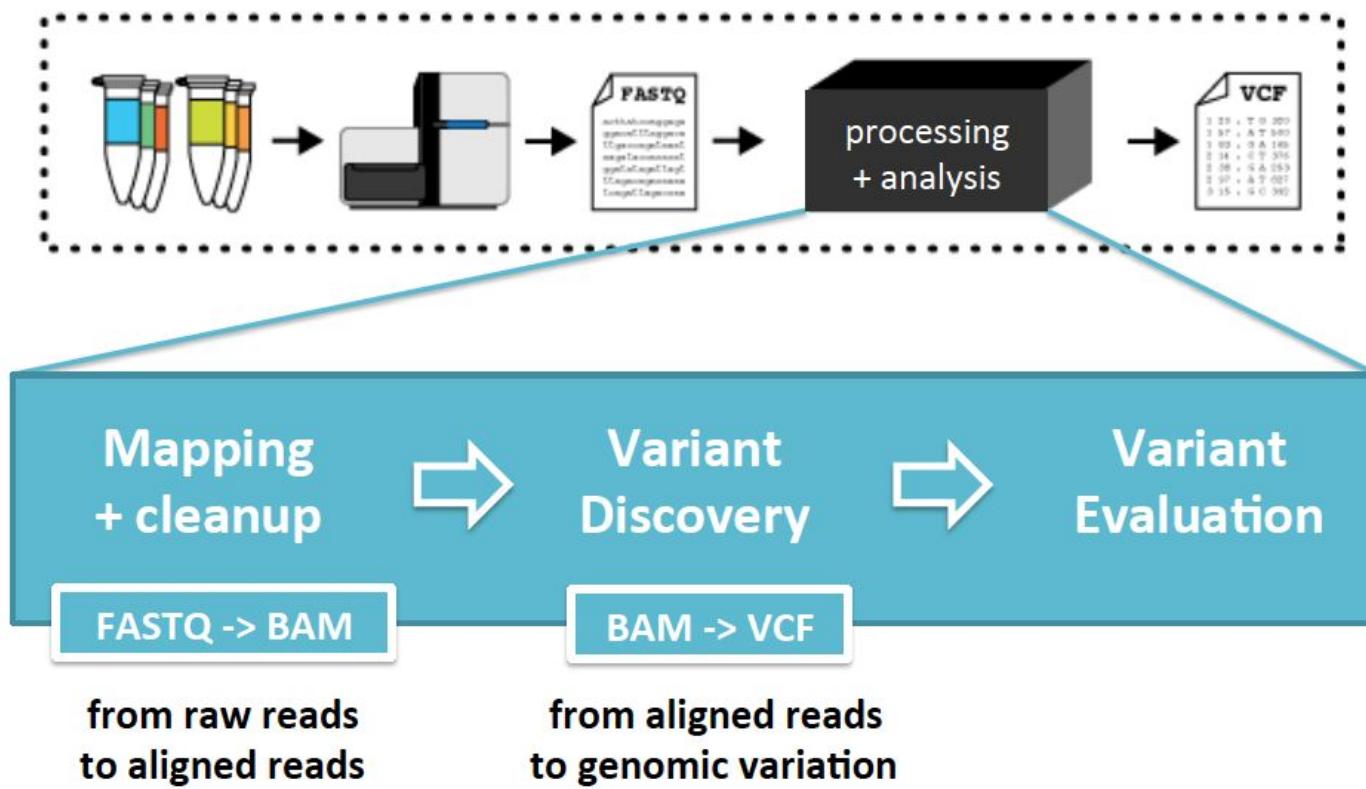
Large SV

Insertion

Other event

5. NGS Data Processing

The Basic Pipeline



Bioinformatic Challenges of NGS

- NGS pushes bioinformatics needs way up!
- Large amount of CPU power
- Informatics groups must manage and maintain large compute clusters and storage
- Challenges in parallelizing software/algorithms for NGS
- VERY large text files (> 10 million lines)
- Standard tools face problems with memory usage, execution times and impossible to browse for problems

Bioinformatic Challenges of NGS

- Data management issues
- Raw data is large (5GB - 30GB per patient)
- How long should they be kept?
- Processed data are manageable for most small studies
- More of an issue for NGS facilities and Bioinformatics services running multiple projects
- WGS 900 samples > 200TB processed data!
- 100K Genomes >>> PETABYTES or More!

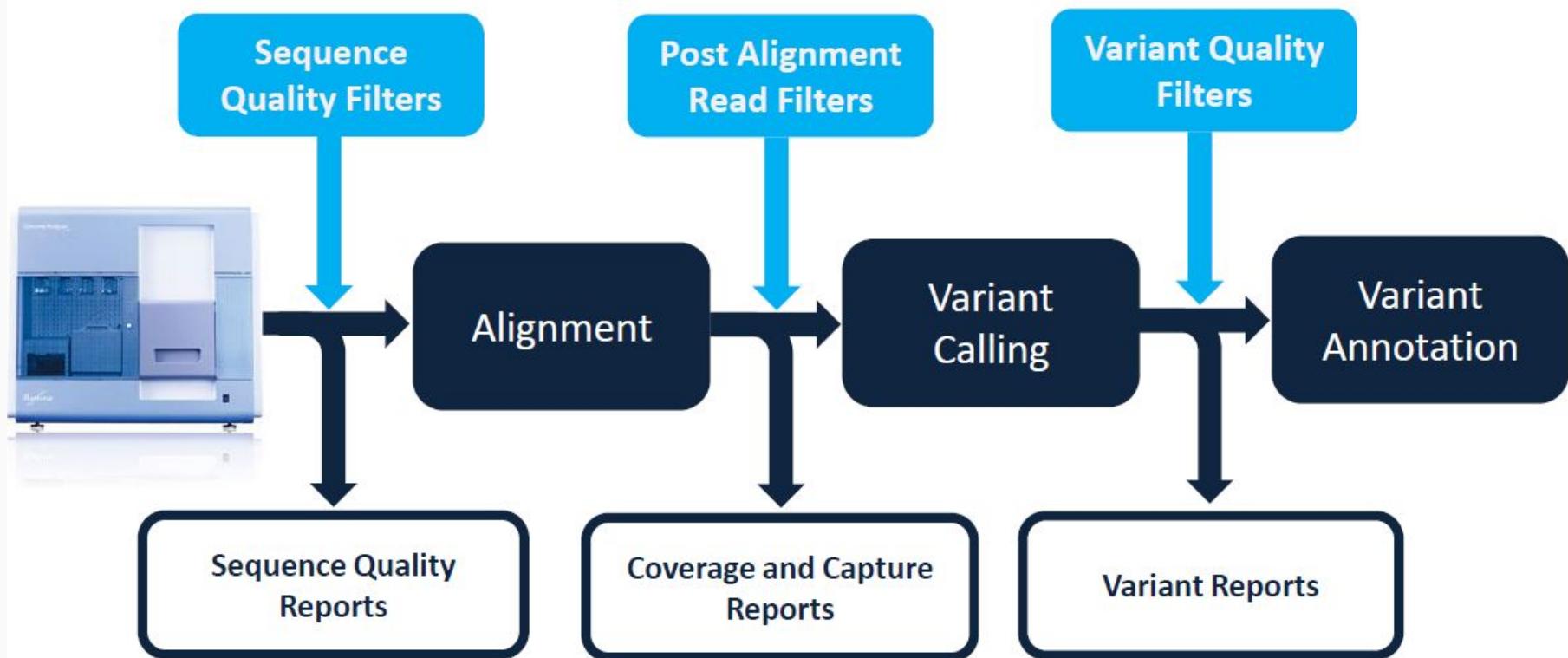
Bioinformatic Challenges of NGS

- In NGS we have to process LARGE amounts of data - non trivial compute time
- Big NGS projects require supercomputing infrastructures and specialist software
- Can't run applications on a standard desktop
- Not everyone can study everything and small facilities,research labs must carefully choose/run projects inline with their compute capabilities & expertise

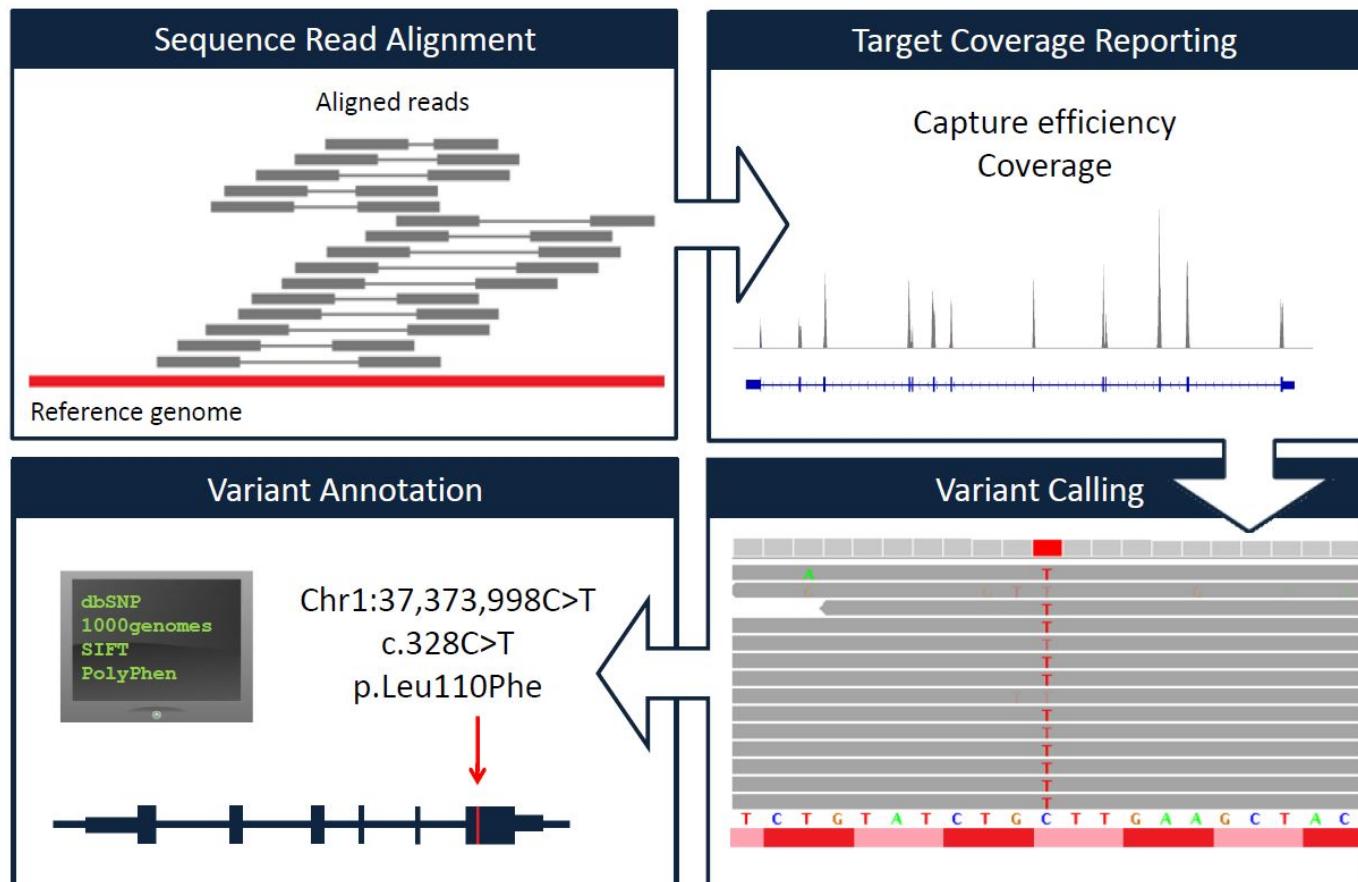
Computational infrastructure for NGS

- **General Requirements**
 - Large Compute cluster
 - Multiple nodes (servers) with multiple cores (500 cores, 50-128 GB RAM)
 - High performance storage (TB,PB levels)
 - Fast network
 - Skilled Bioinformaticians (sysadmin, developers)
- **Bare minimum for a small project**
 - EG. Families Exome Seq, single sample WGS
 - 1-2 Workstations
 - 8-12 cores, 48-64GB RAM, at least 4TB
 - Skilled Bioinformaticians (sysadmin, developers)

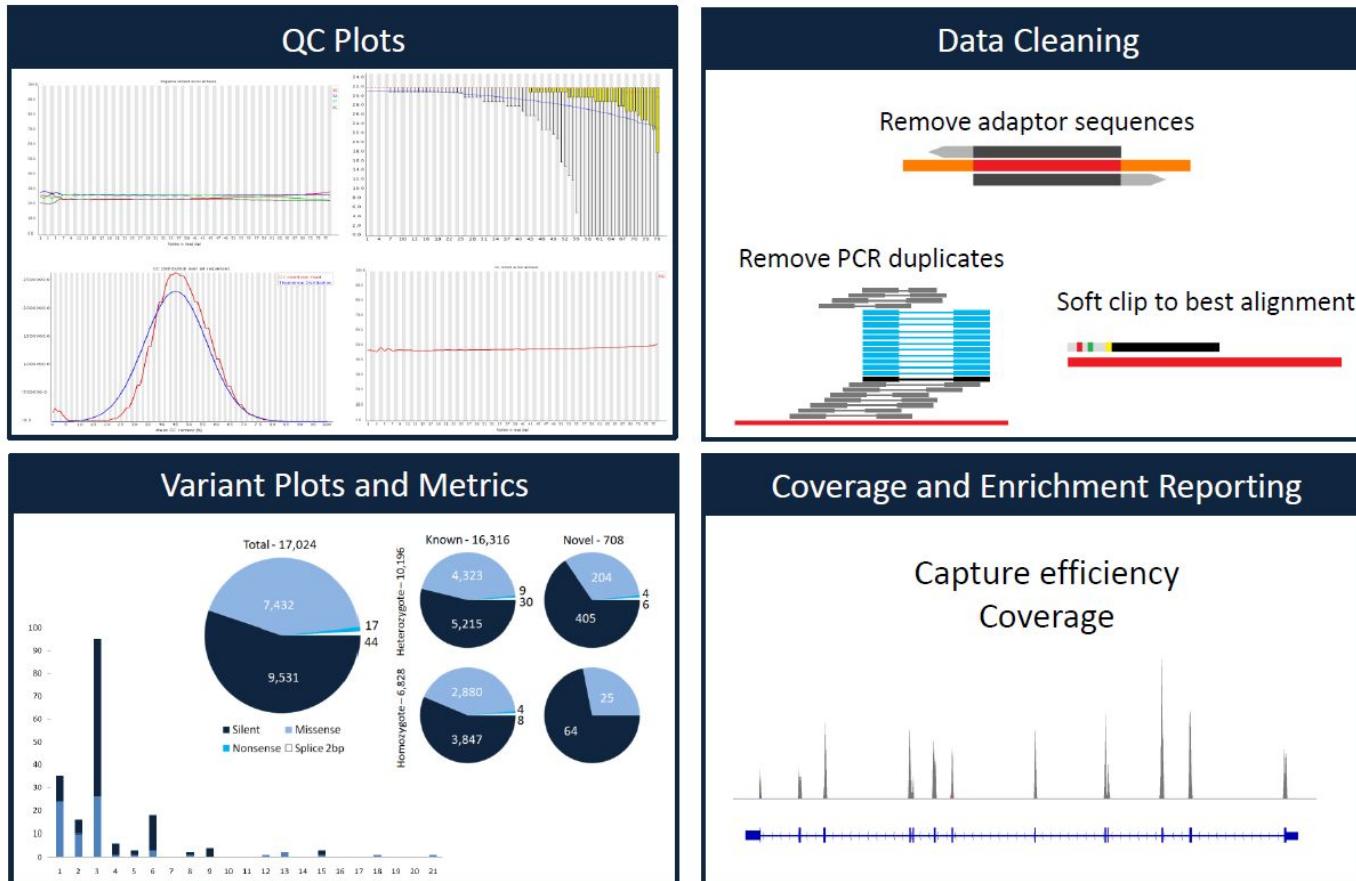
NGS Data Analysis: Basic Bioinformatic Workflow I



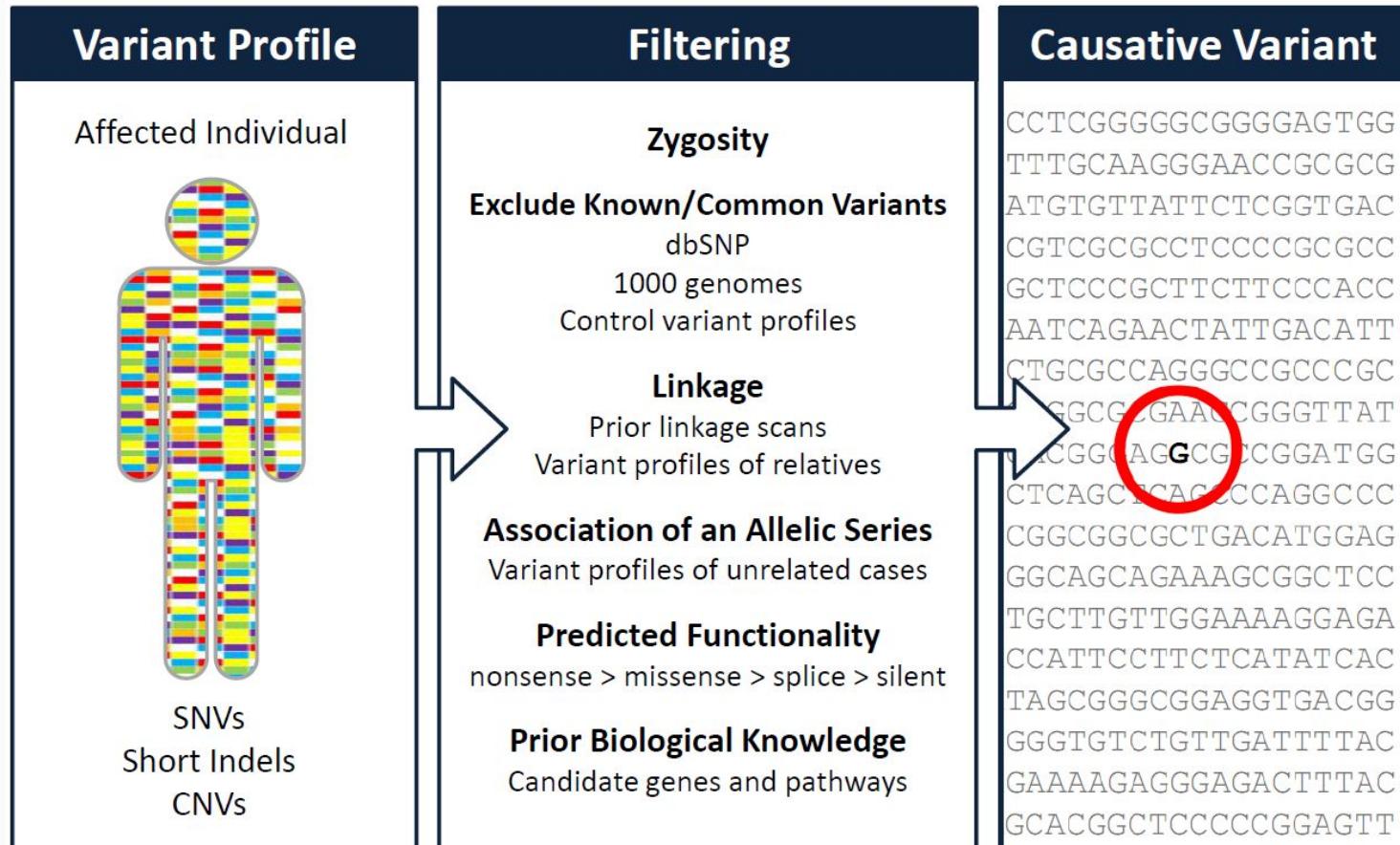
NGS Data Analysis: Basic Bioinformatic Workflow II



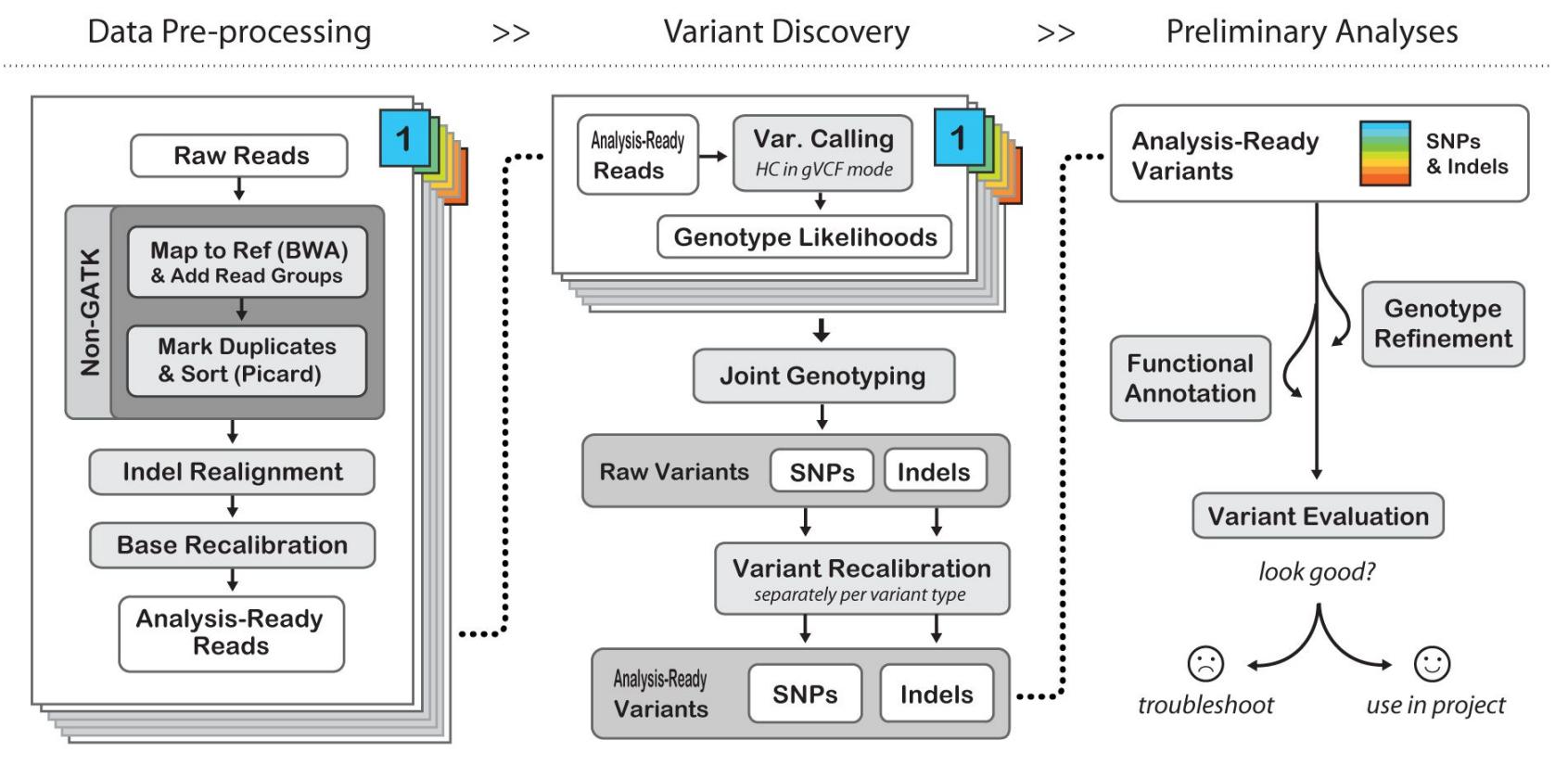
NGS Data Analysis: Basic Bioinformatic Workflow III



NGS Data Analysis: Basic Bioinformatic Workflow IV

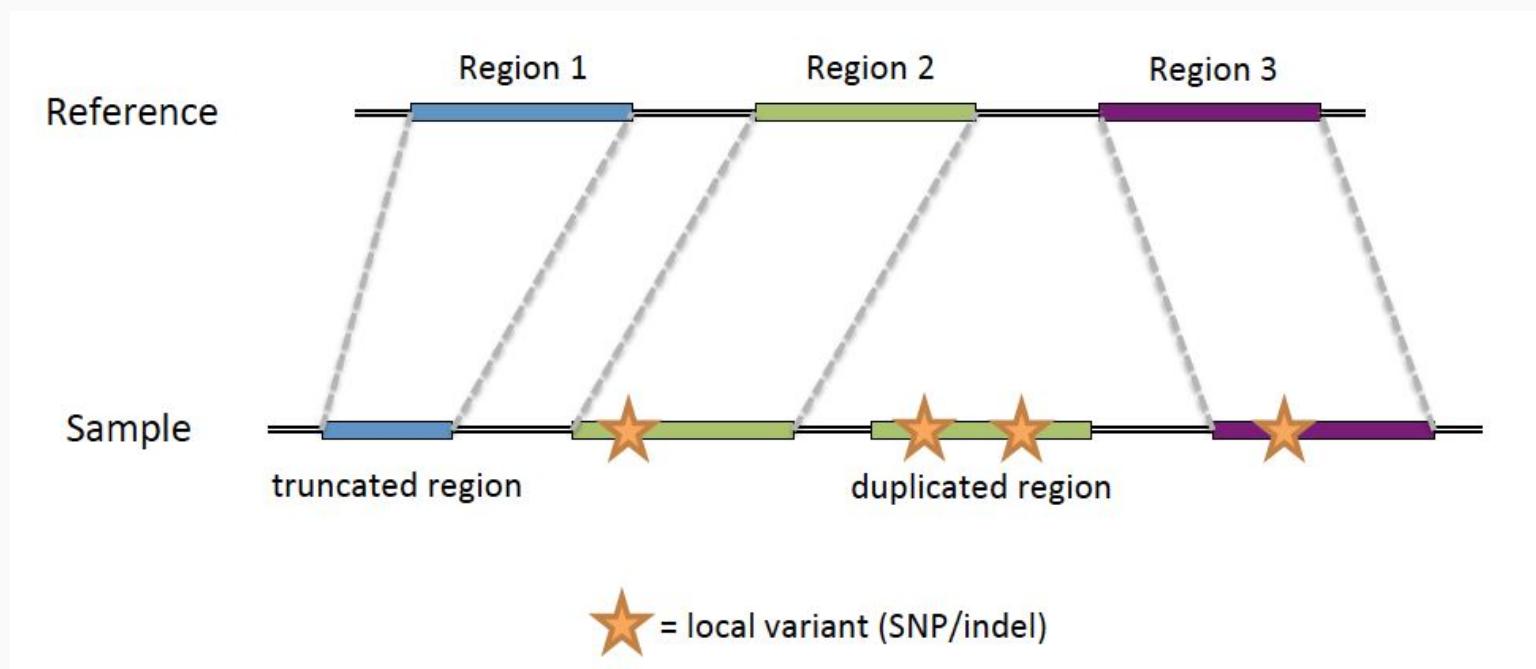


GATK “Genome Analysis Toolkit” : Community Driven Best Practices



Alignment: the most important step in the pipeline

- Ideally we'd align the sample genome to the reference genome....



...But we don't have the whole sample...just a bunch of short reads...

We have a pile of reads that need to be mapped individually

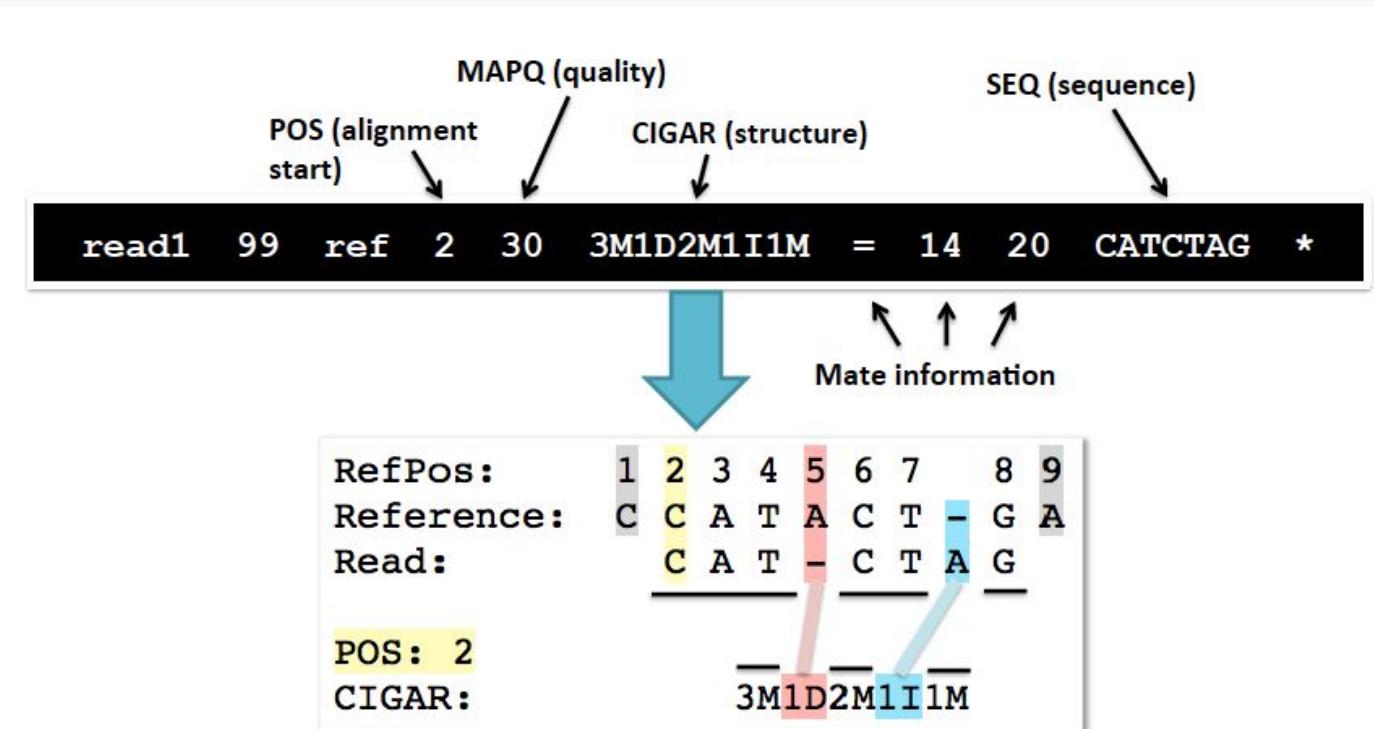
Mapping is complicated by mismatches (true mutations or sequencing errors), indels, duplicated regions etc.



Many alignment tools available...

- All concerned with finding where short reads map to the reference genome
- Details outside the scope of this course...
- **Popular aligners**
 - bwa <http://bio-bwa.sourceforge.net/>
 - bowtie <http://bowtie-bio.sourceforge.net/index.shtml>
 - novoalign <http://www.novocraft.com/products/novoalign/>
 - stampy <http://www.well.ox.ac.uk/project-stampy>
 - **isaac** (illumina) <https://github.com/Illumina/isaac2>
 - many, many more.....

SAM revisited: Mapping produces SAM summarizing position, quality and structure for given sequence alignment..

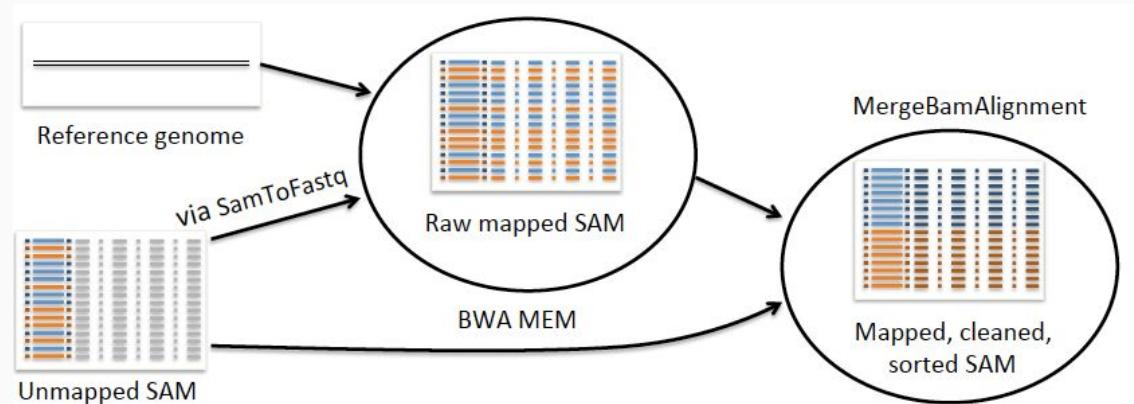


See also:

- SAM format spec: <http://samtools.github.io/hts-specs/SAMv1.pdf>
- Explain SAM flags: <http://broadinstitute.github.io/picard/explain-flags.html>

Post-processing of aligned files

- **Marking of PCR duplicates**
 - PCR amplification errors can cause some sequences to be over-represented
 - Chances of any two sequences aligning to the same position are unlikely
 - Caveat: obviously this depends on amount of the genome you are capturing
 - Such reads are marked but not usually removed from the data
 - Most downstream methods will ignore such reads
 - Typically, [picard](#) is used
- **Sorting**
 - Reads can be sorted according to genomic position
 - [samtools](#)
- **Indexing**
 - Allow efficient access
 - [samtools](#)
- **Mapping artefacts**
 - picard : [CleanSam](#)
 - picard: [FixMateInformation](#)



Genomics England “Pipeline”

as of Feb 2016...

- NGS Provider: Illumina Ltd
- NGS Technology: HiSeq x10
 - Paired End
 - 150bp
- Single Sample Whole Genome
- Depth: min average depth 30x (Tumour 75x)
- Pipeline: **ISSAC2** (Aligner), **Starling** (SNPs), **Canvas** (CNV), **Manta** (SV)
- Alignment: **BAM**
- Variant Calls: **VCF** or gVCF
 - SNP, CNV, Small
- Fast Track Whole Genome Sequencing Services: <https://goo.gl/uMgvgR>

<https://github.com/Illumina/isaac2>

<https://github.com/Illumina/manta>

<https://github.com/Illumina/canvas>

Further Reading

- See all slide links & Handbook



Next Generation Sequencing (NGS): Introduction, Data Formats, Processing & Variant Detection: Part II

Dr Stephen J Newhouse

Lead Data Scientist & Senior Bioinformatician @ NHIR BRC-MH SlaM NHS & IoPPN KCL & UCL Farr Institute

stephen.j.newhouse@kcl.ac.uk, [@s_j_newhouse](https://twitter.com/s_j_newhouse)

Genomic Medicine MSc

16 Feb 2016

Overview

1. Variant Calling
2. Types of Variants
3. Tools for Variant Calling
4. Things to Think About (Depth, Quality, Errors?)

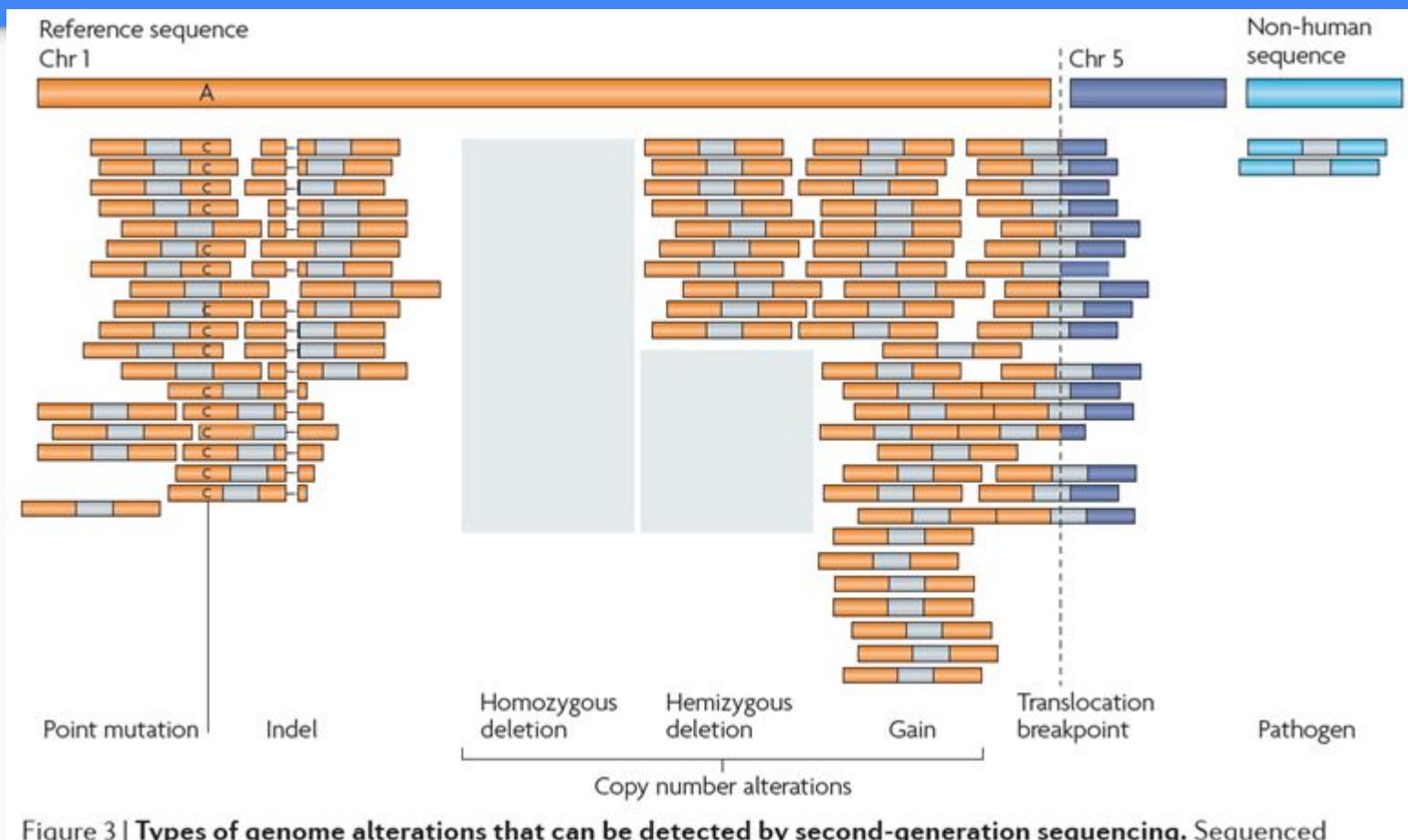


Figure 3 | Types of genome alterations that can be detected by second-generation sequencing. Sequenced

ARTICLE

Received 15 Sep 2014 | Accepted 13 Jan 2015 | Published 25 Feb 2015

DOI: 10.1038/ncomms7275

OPEN

An analytical framework for optimizing variant discovery from personal genomes

Gareth Highnam¹, Jason J. Wang¹, Dean Kusler¹, Justin Zook², Vinaya Vijayan³, Nir Leibovich¹
& David Mittelman^{1,3}

The standardization and performance testing of analysis tools is a prerequisite to widespread adoption of genome-wide sequencing, particularly in the clinic. However, performance testing is currently complicated by the paucity of standards and comparison metrics, as well as by the heterogeneity in sequencing platforms, applications and protocols. Here we present the genome comparison and analytic testing (GCAT) platform to facilitate development of performance metrics and comparisons of analysis tools across these metrics. Performance is reported through interactive visualizations of benchmark and performance testing data, with support for data slicing and filtering. The platform is freely accessible at <http://www.bioplanet.com/gcat>.