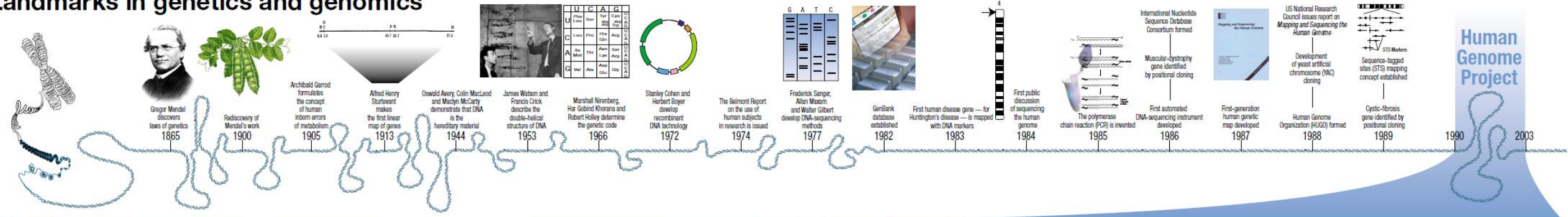
中央研究院．臺灣人體生物資料庫
Academia Sinica · Taiwan Biobank

基因體定序資料品質檢查與前處理

蘇明威 bioperl@as.edu.tw

Support science | Improve our health | Accelerate your discovery

Landmarks in genetics and genomics



1990

The Human Genome Project (HGP) launched in the United States

Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)

First gene for breast cancer (*BRCA1*) mapped

1991

First US genome centres established

1992

Second-generation human genetic map developed

Rapid-data-release guidelines established by the NIH and DOE

1993

New five-year plan for the HGP in the United States published

The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)

1994

The HGP's human genetic mapping goal achieved

1995

The HGP's human physical mapping goal achieved

First bacterial genome (*Haemophilus influenzae*) sequenced

US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996

First human gene map established

Pilot projects for human genome sequencing begin in the United States

First archaeal genome sequenced

Yeast (*Saccharomyces cerevisiae*) genome sequenced

The HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997

The DOE forms the Joint Genome Institute

US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)

Escherichia coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded near Paris

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the United States published

RIKEN Genomic Sciences Center established in Japan

Roundworm (*Caenorhabditis elegans*) genome sequenced

Single-nucleotide polymorphism (SNP) initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers established in Beijing and Shanghai

1999

Full-scale human genome sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced

Mustard cress (*Arabidopsis thaliana*) genome sequenced

Executive order bans genetic discrimination in US federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rice genome sequence completed and published

Draft version of rat genome sequence completed

2003

Finished version of human genome sequence completed

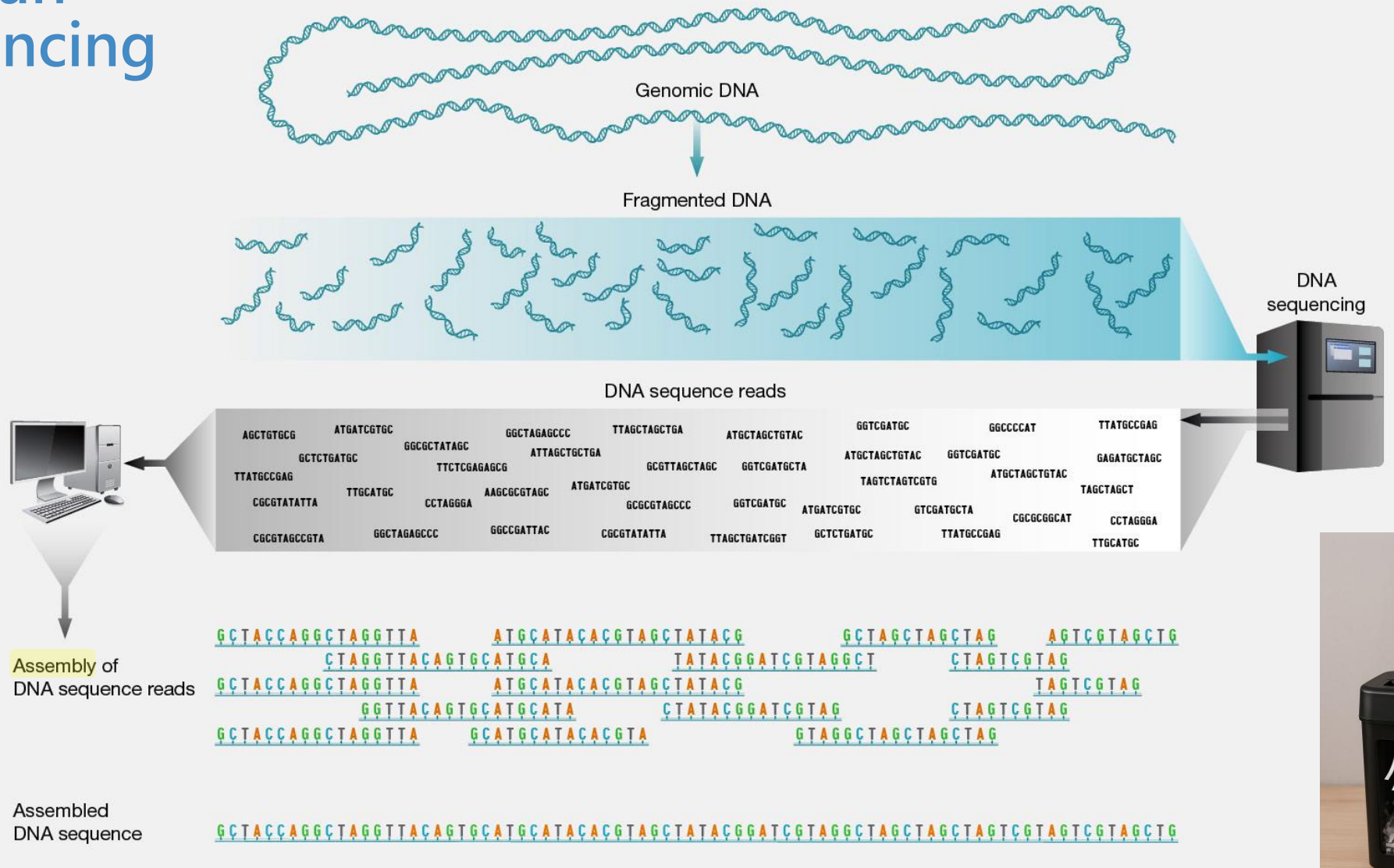
The HGP ends with all goals achieved

To be continued...

PEAS COURTESY I. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; SCIENCE COVERS COURTESY AAAS

https://www.genome.gov/human-genome-project

Shotgun Sequencing



https://www.genome.gov/genetics-glossary

TWB申請使用資料流程及注意事項

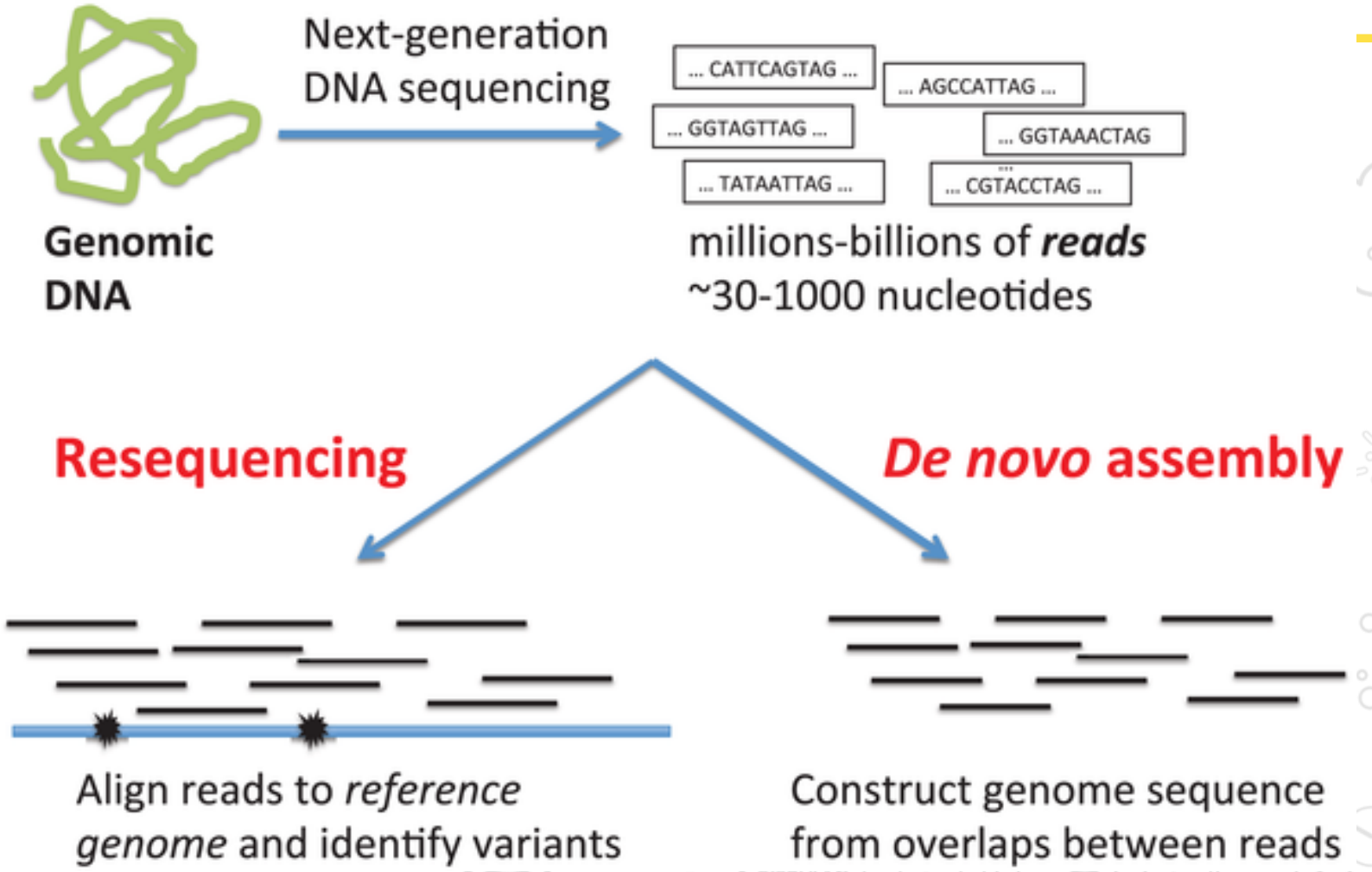
- TWB資料申請流程：線上申請— 檢體與數位資料釋出管理系統

- IRB撰寫注意事項

① 研究對象—TWB收案對象：	30~70歲(X)	年滿20歲本國籍成年人(O)
② 收案人數—「數位資料集」	197,000筆(X)	200,000人(O)
③ 研究材料—資料名稱完整：	環境暴露資料(X)	尿液三聚氰胺含量數位資料集(O)
④ 範例(以TWB申請書為例)		

Assembly and alignment

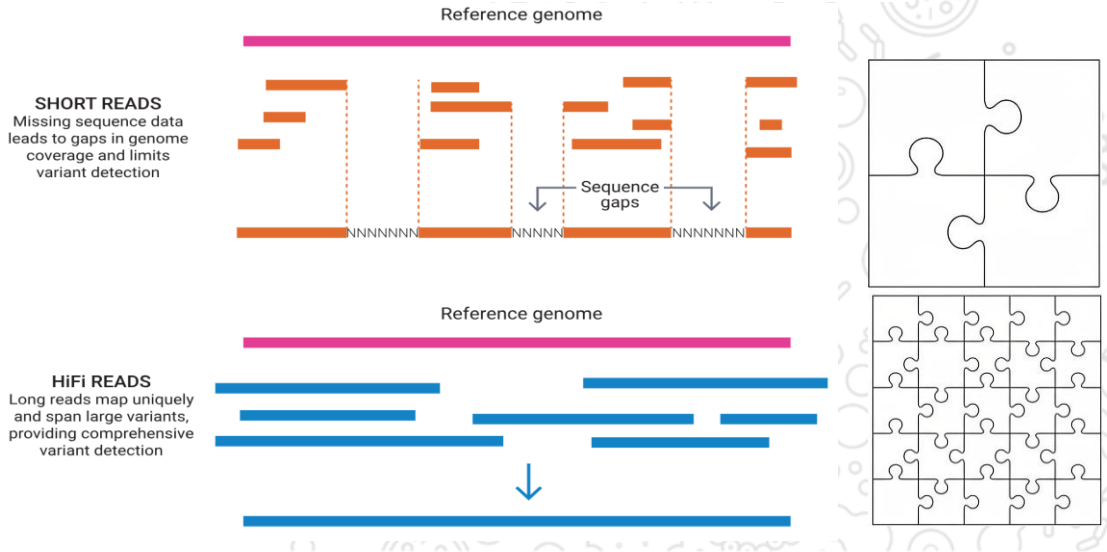
https://thejigstore.co.nz/



Raphael BJ (2012) Chapter 6: Structural Variation and Medical Genomics. PLoS Comput Biol 8(12): e1002821. https://doi.org/10.1371/journal.pcbi.1002821

Short read and long read

Comparison Dimension	PacBio	Nanopore
Principle	Fluorescently labeled dNTPs + ZMW	Nanopore current sensing
Read Length	10–20 kb (HiFi)	Up to Mb levels
Initial Accuracy	~85%	~93.8% (R10 chip)
Corrected Accuracy	>99.9%	~99.996% (consensus sequence, 50X depth)
Throughput	120 Gb/run (HiFi)	1.9 Tb/run (PromethION)
Equipment Cost	High	Low (MinION portable)
Suitable Applications	Clinical research, genome assembly	Field monitoring, real-time analysis



https://www.cd-genomics.com/resource-pacbio-oxford-nanopore-comparison.html

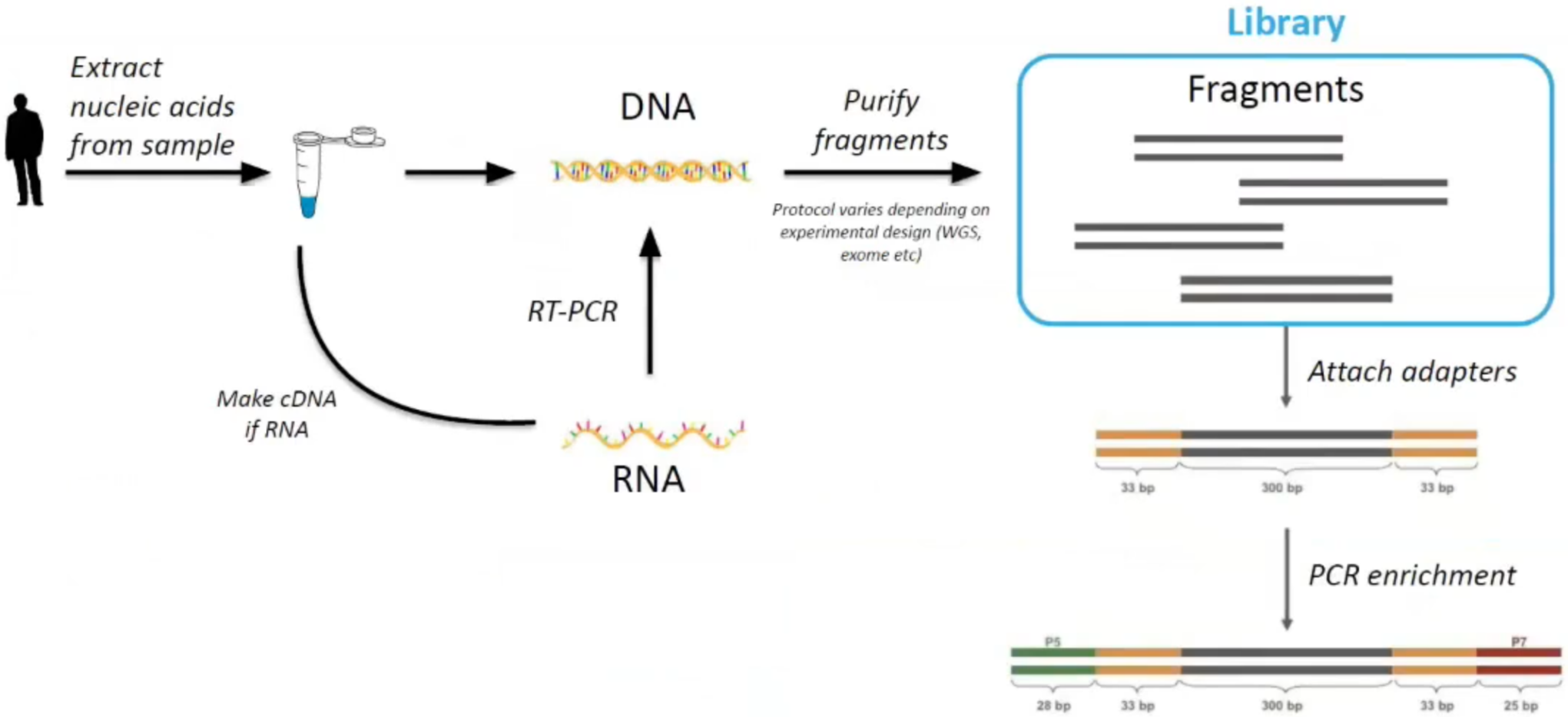
https://www.pacb.com/blog/sequencing-101-comparing-long-read-sequencing-technologie

Technology Selection Based on Research Objectives

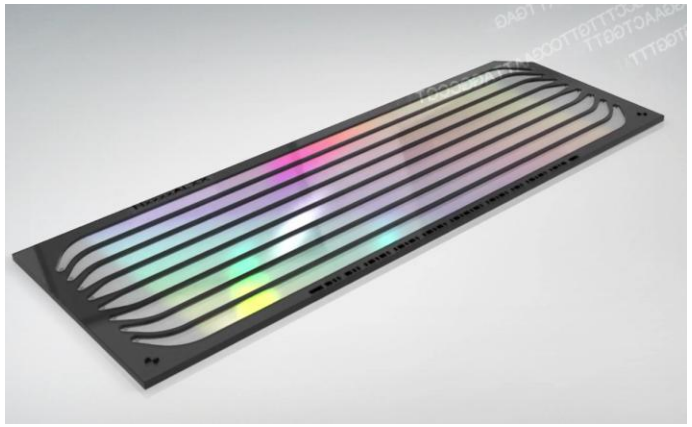
Research Objective	Recommended Technology	Justification
High-Precision Genome Assembly	PacBio	HiFi reads offer post-error correction accuracy >99.9%, making them ideal for assembling complex genomes and detecting structural variations.
Real-Time Monitoring and Rapid Response	Nanopore	Supports real-time data streaming, suitable for infectious disease monitoring and timely clinical diagnostics.
Full-Length Transcriptome Analysis	PacBio	HiFi reads facilitate full-length transcript sequencing, directly capturing RNA isoforms and elucidating gene expression regulatory mechanisms.
Sequencing in Extreme Environments	Nanopore	The portable MinION device is well-suited for genomic sequencing in field, polar, and even space environments.
Epigenetic Research	PacBio	Direct detection of DNA methylation and base modifications provides high-resolution data for studying epigenetic regulatory networks.
Large-Scale Pathogen Genome Research	Nanopore	With a PromethION platform throughput of up to 1.9 Tb per run, it supports high-throughput pathogen genome sequencing and analysis.

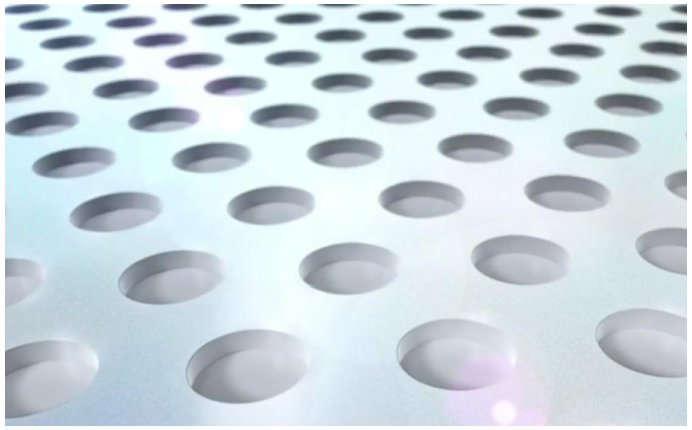
https://www.cd-genomics.com/resource-pacbio-oxford-nanopore-comparison.html

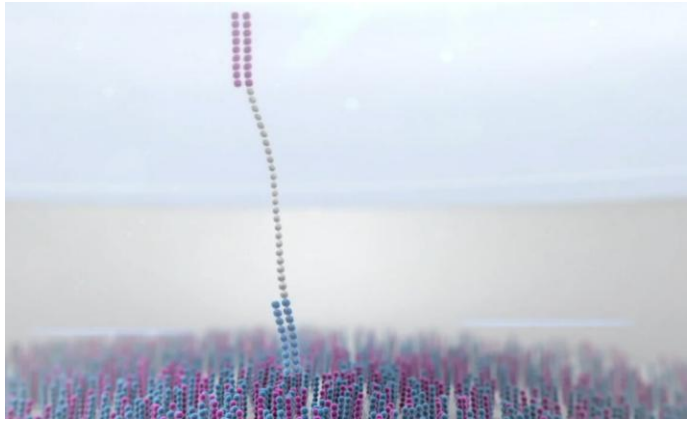
Library preparation

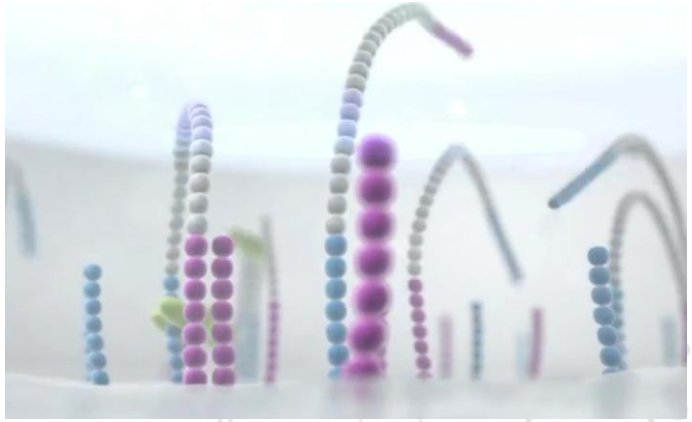


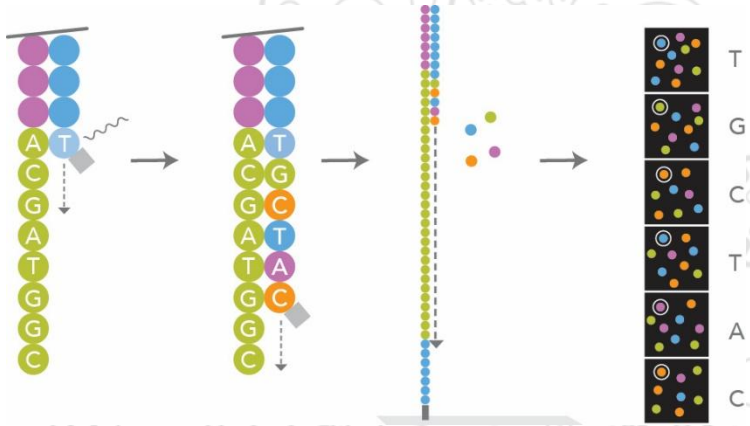
Illumina flow cell









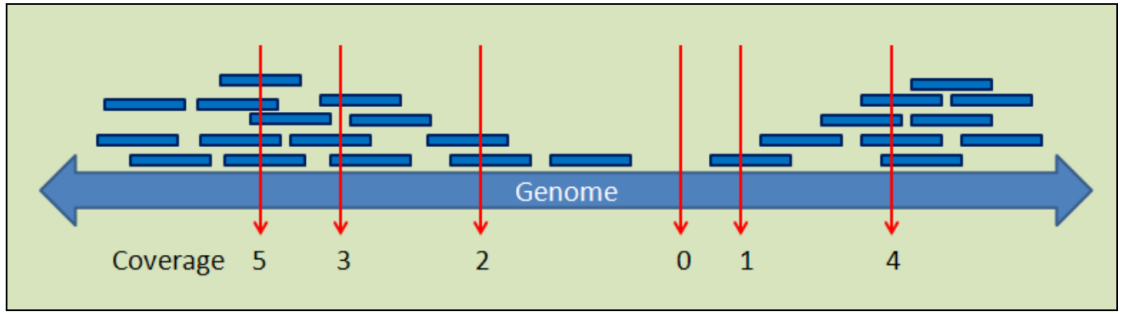


Sequencing by Synthesis

https://www.youtube.com/watch?v=pfZp5Vgsbw0&t=14s

Depth and coverage

- Coverage – What % of the genome sequence is represented in the sequencing data
- Depth of coverage – How many times is every base in the genome represented (on average)



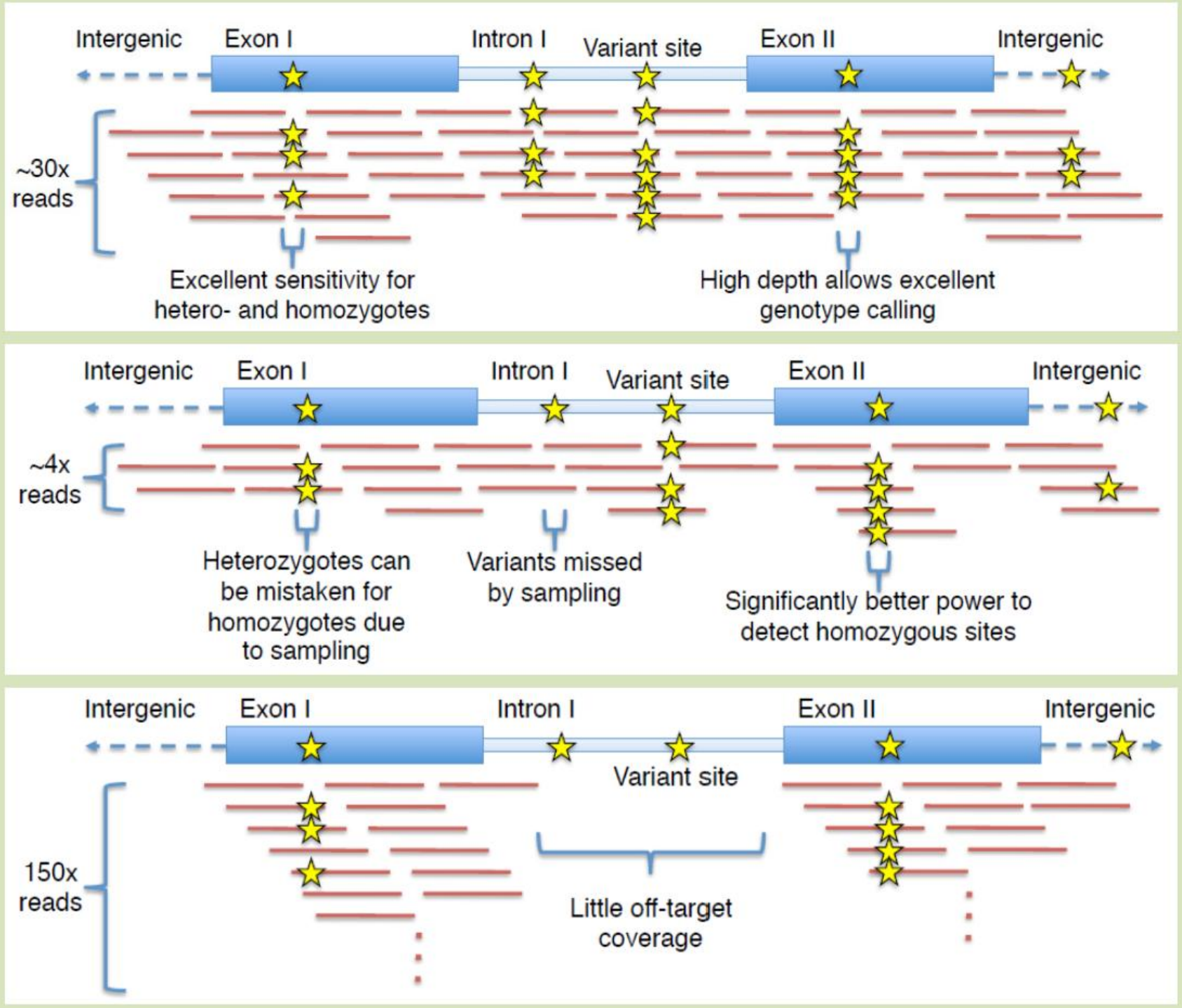
For WGS

- Haploid genome size => 3.2 Giga base pairs (3.2 billion)
- 50x coverage => ~160 Gbp

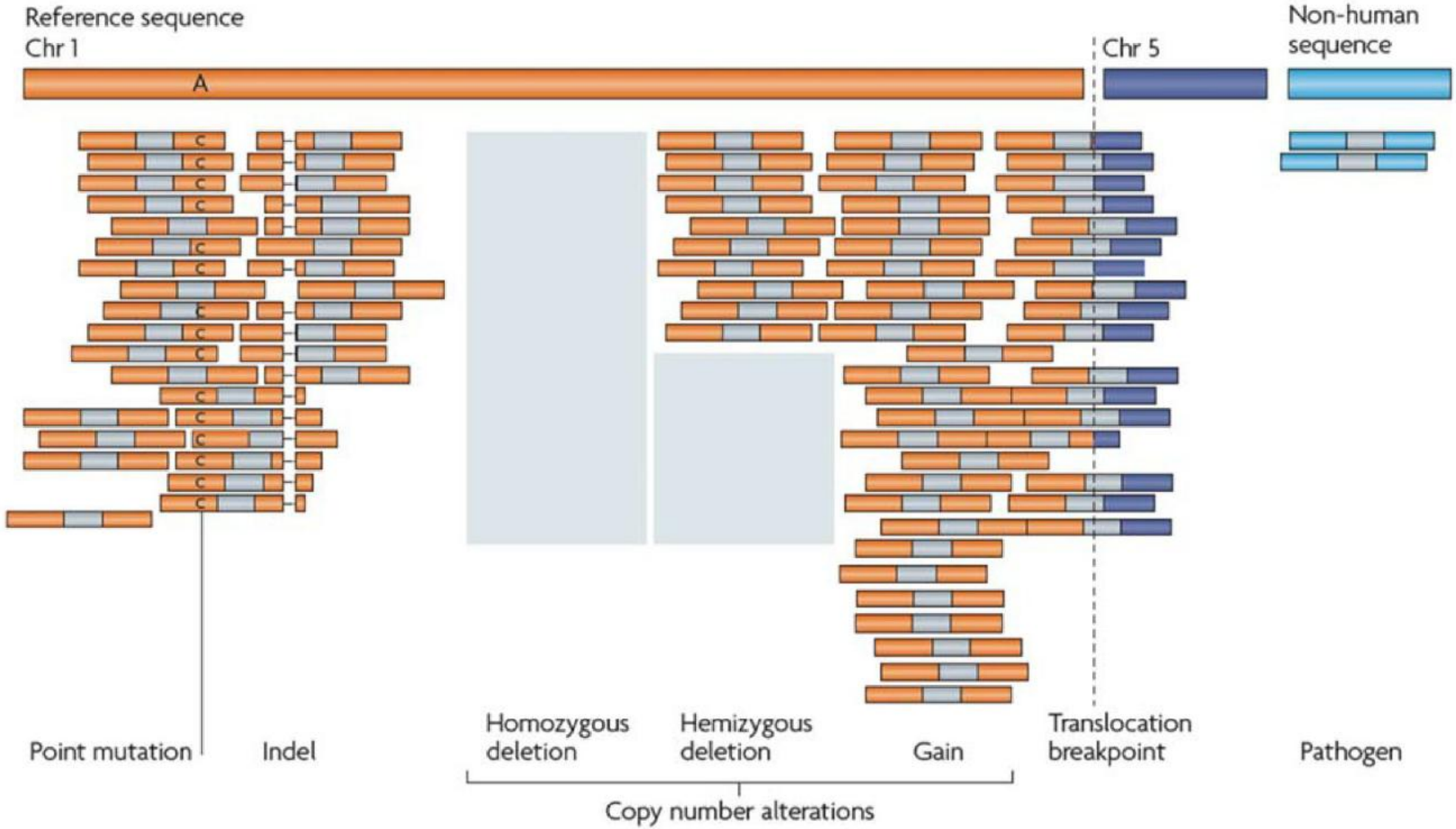
For Exome Sequencing

- Exome size => 33 Mega base pairs (33 million bases)
- 50x coverage => ~1.65 Gbp
- About 100 times smaller than WGS
- Depending on your method of capture, this number can vary

Design



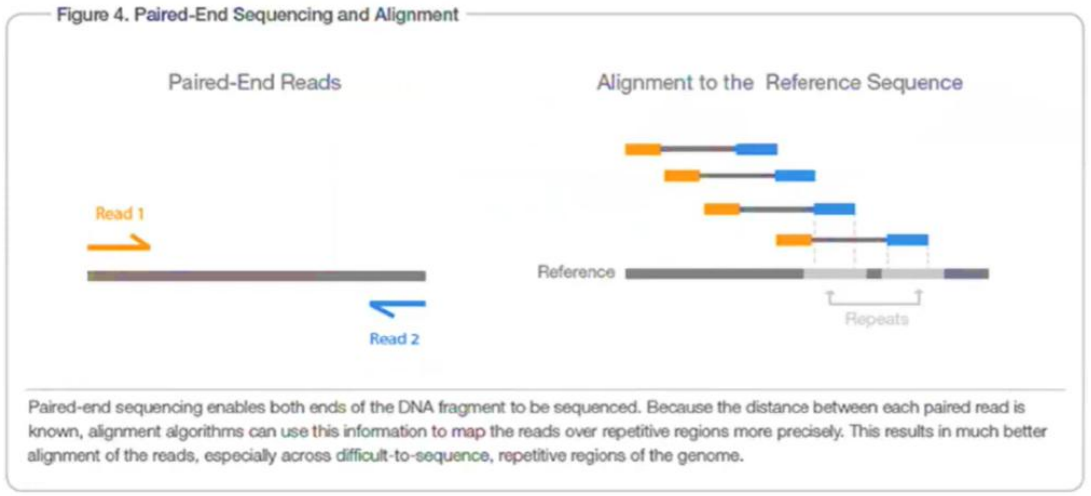
Re-sequencing: variation detection



Nat Rev Genet. 2010 Oct; 11(10): 685-96.

Mapping reads to reference is simple in principle…

Insert 的作用與意義



https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

1. 幫助比對與定位
- 雙端定序可以提供「兩個位置」的比對資訊
- 如果 Read 1 和 Read 2 都能準確比對到參考基因組，且距離合理 → 增加比對可信度

2. 偵測結構變異（Structural Variants）
- 如果 insert size 明顯偏大或偏小 → 可能是有 **插入、缺失、倒位、融合等結構變異**
- 例如：正常 insert 是 300 bp，但某對 reads 間距變成 1000 bp → 可能是插入了外來序列

3. 過濾錯誤比對
- 如果 Read 1 和 Read 2 比對到不同染色體，或方向不對 → 可能是 chimeric read 或污染
- Insert size 可以用來判斷是否合理

4. 估算文庫品質
- Insert size 的分布可以反映文庫製備是否均勻
- 太短或太長都可能影響定序效率與分析準確度

Human reference genome

Next assembly update
The GRC remains committed to its mission to improve the human reference genome assembly, correcting errors and adding sequence to ensure it provides the best representation of the human genome to meet basic and clinical research needs. We will continue to make these updates publicly available at regular intervals in the form of patch releases, but have decided to indefinitely postpone our next coordinate-changing update (GRCh39) while we evaluate new models and sequence content from ongoing efforts to better represent the genetic diversity of the human pangenome, including those of the Telemore-to-Telomere Consortium and the Human Pangenome Reference Consortium.

Release name	Date of release	Equivalent UCSC version
GRCh39	Indefinitely postponed[29]	-
T2T-CHM13	January 2022	hs1
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Donors were recruited by advertisement in The Buffalo News, on Sunday, March 23, 1997. As a result of how the DNA samples were processed, about 80 percent of the reference genome came from eight people and one male, designated RP11, accounts for 66 percent of the total. The ABO blood group system differs among humans, but the human reference genome contains only an O allele, although the others are annotated.

https://en.wikipedia.org/wiki/Reference_genome
https://www.ncbi.nlm.nih.gov/grc/human

GRCh38/hg38

Component	Description	Number of Contigs
Main Reference	chr1-22, chrX, chrY, chrM	25
Unlocalised Sequences	chr*_random	42
Unplaced Sequences	chrUn*	127
Alternative Loci	chr*_alt	261
Epstein-Barr Virus	chrEBV	1
Decoy Sequences	chrUn*_decoy	2385
HLA Sequences	HLA-*	525

- Unlocalized sequences: known to belong on a specific chromosome but with unknown order or orientation.
- Unplaced sequences: chromosome of origin unknown.
- Alternate contig sequences can be novel to highly diverged or nearly identical to corresponding primary assembly sequence.

Note.

- In b37 where chromosomes are called "1" to "22", "X", "Y" and "MT")
- human genome" refers to the approximately 3 billion ($3*10^9$) base pairs in a single copy (haploid) of the human genome.

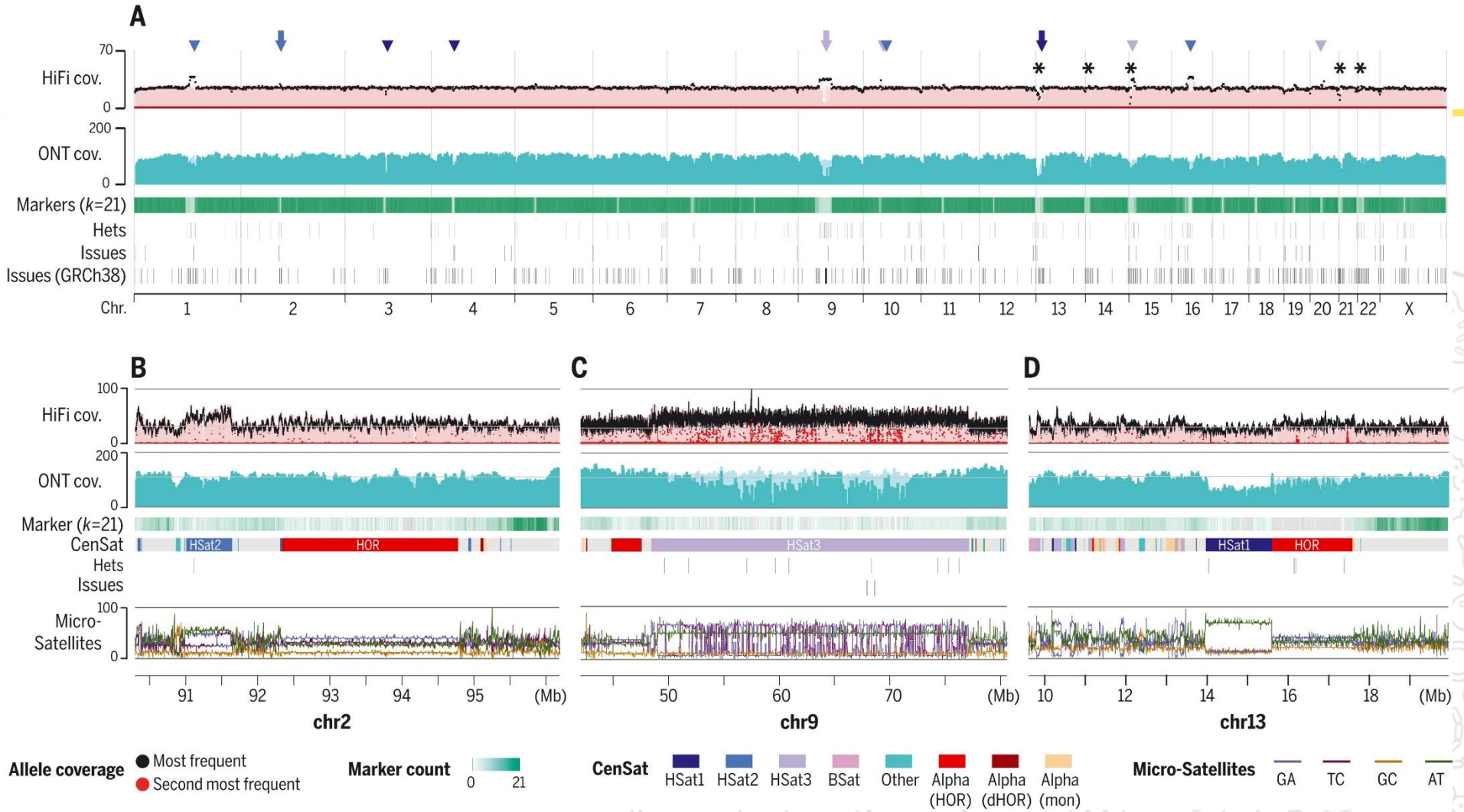
https://www.illumina.com/science/genomics-research/articles/dragen-demystifying-reference-genomes.html

Simons Genome Diversity Project

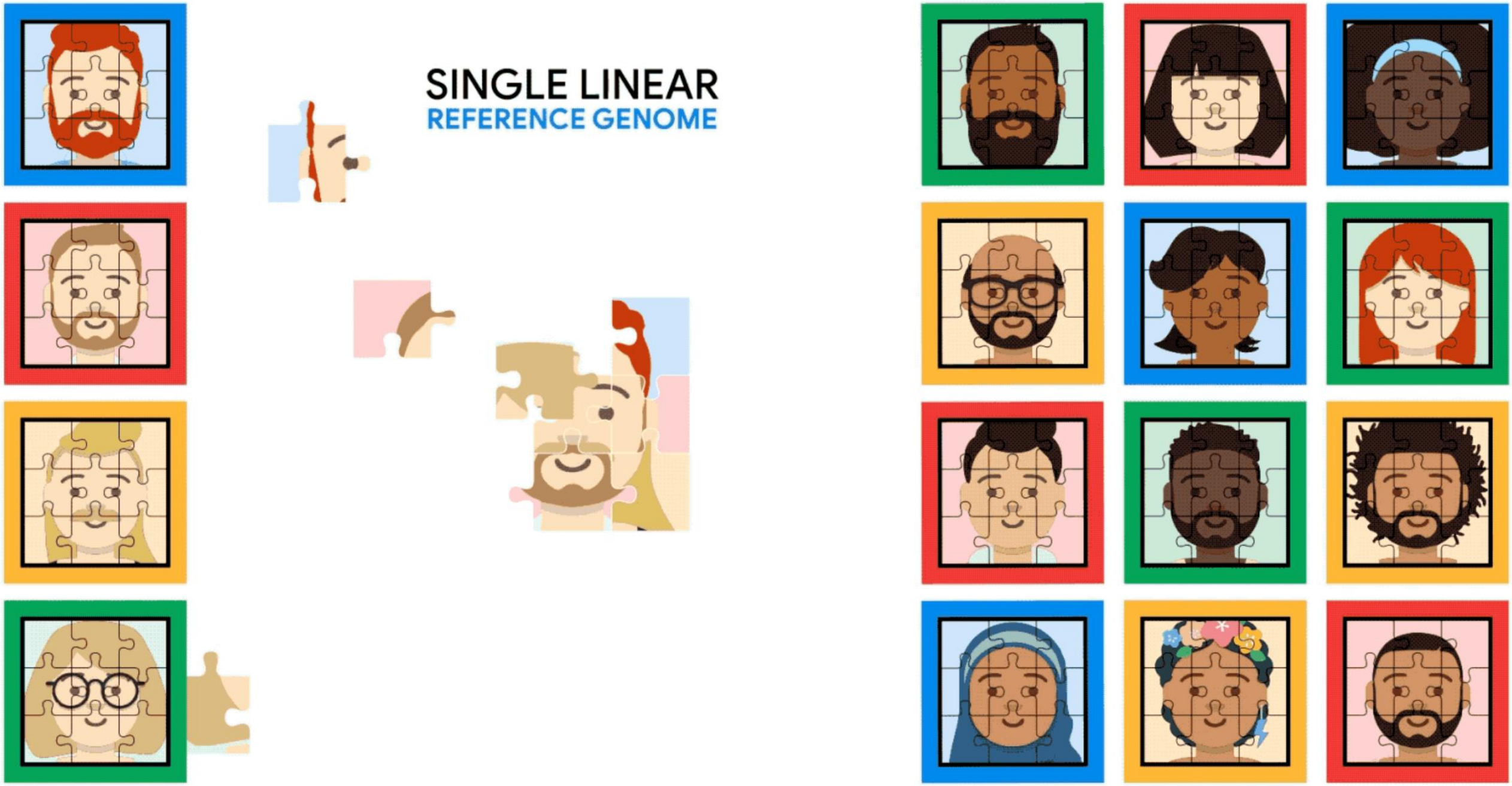
260 genomes from 127 populations at least 30x coverage using Illumina technology.

https://www.simonsfoundation.org/simons-genome-diversity-project/

Telomere-to-Telomere (T2T) consortium

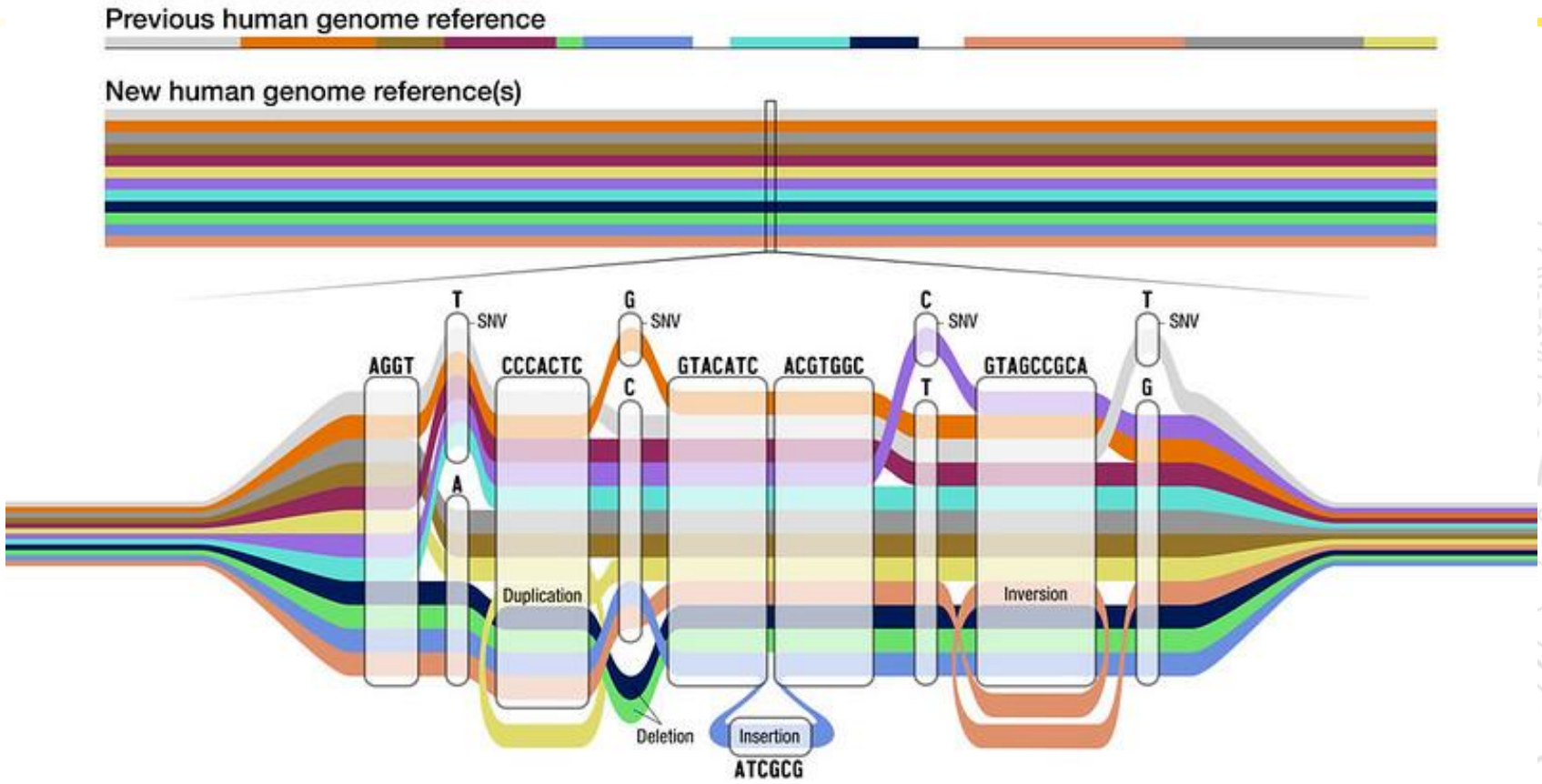


https://www.science.org/doi/10.1126/science.abj6987



https://research.google/blog/building-better-pangenomes-to-improve-the-equity-of-genomics/

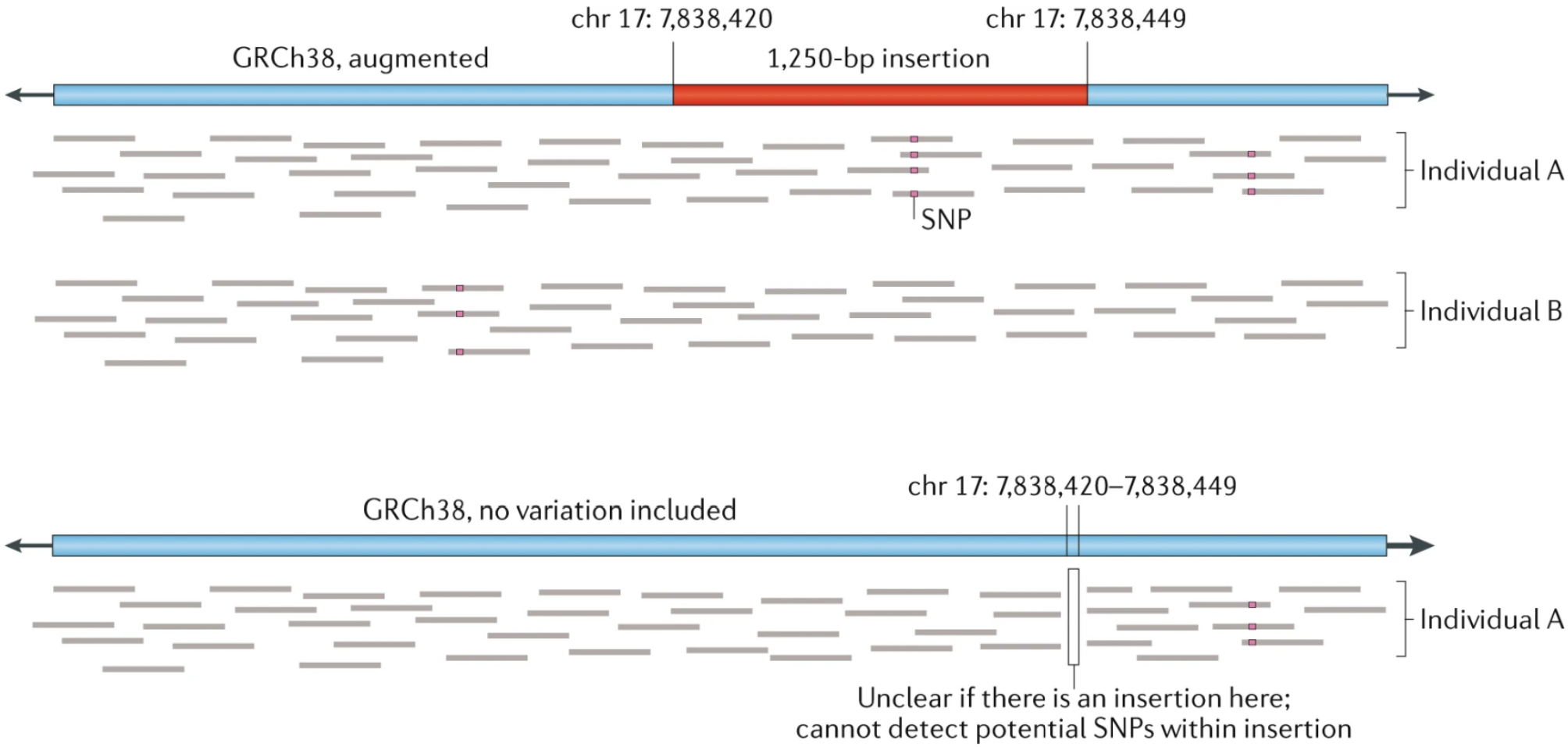
human “pangenome” reference



亦可將這些額外的序列直接作為 Alternate 序列記錄下來。GRCh38 中大量的 ALT 序列即是。但如它們原本應該是在某條染色體上，特定位置的資訊就沒被記下。

https://www.nih.gov/news-events/news-releases/scientists-release-new-human-pangenome-reference

Variant discovery from a pan-genome reference.

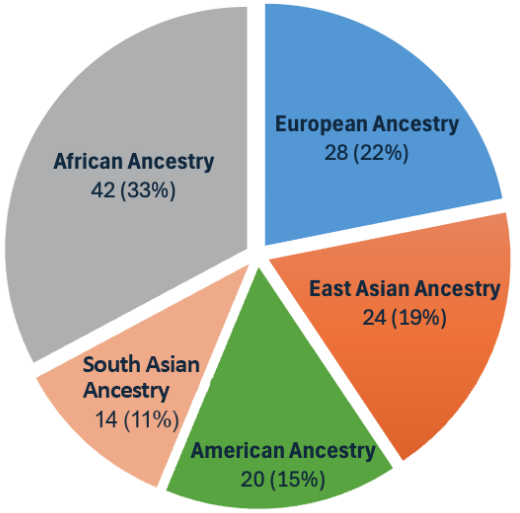


https://www.nature.com/articles/s41576-020-0210-7

Evolution of the DRAGEN

DRAGEN	3.4-3.6	3.7-3.8	3.9	3.10	4.0	4.2	4.3
Native alt contigs handling	alt_aware		alt_masked	alt_masked_v2		alt_masked_v3	alt_masked_v4
Multigenome mapper		First generation					Second generation
Pangenome reference *		16 European ancestry samples		16 European ancestry samples + MHC		32 Global ancestry samples	128 Global ancestry samples
Machine learning			v2	v3.1	v4	v7	v12

* historically referred to as "multigenome (graph) reference".



The pangenome reference in DRAGEN v4.3 is made up of 128 population samples from 26 ancestries around the world.

https://sapac.illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html

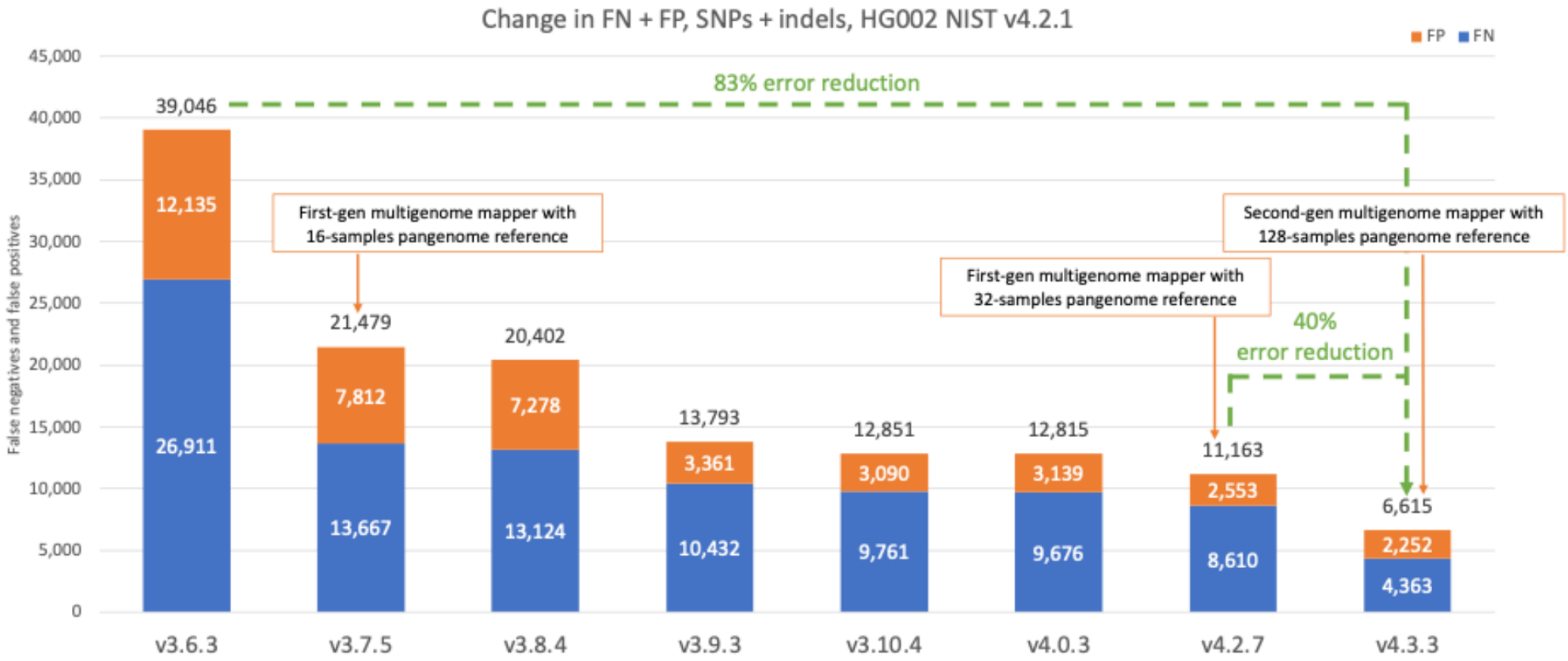


Figure 4: HG002 NIST v4.2.1 SNPs and indels false positives and false negatives error count on successive DRAGEN versions.

https://sapac.illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html

Quality control is essential to catch problems early

Library Prep

- Normal amounts of raw data (in Gb), but poor target coverage
- High proportion of chimerism
- Bad insert size distribution (too big/small, weird peaks)
- Shearing-based oxidation (poor *OxoQ* values)
- Library size too small

Exomes

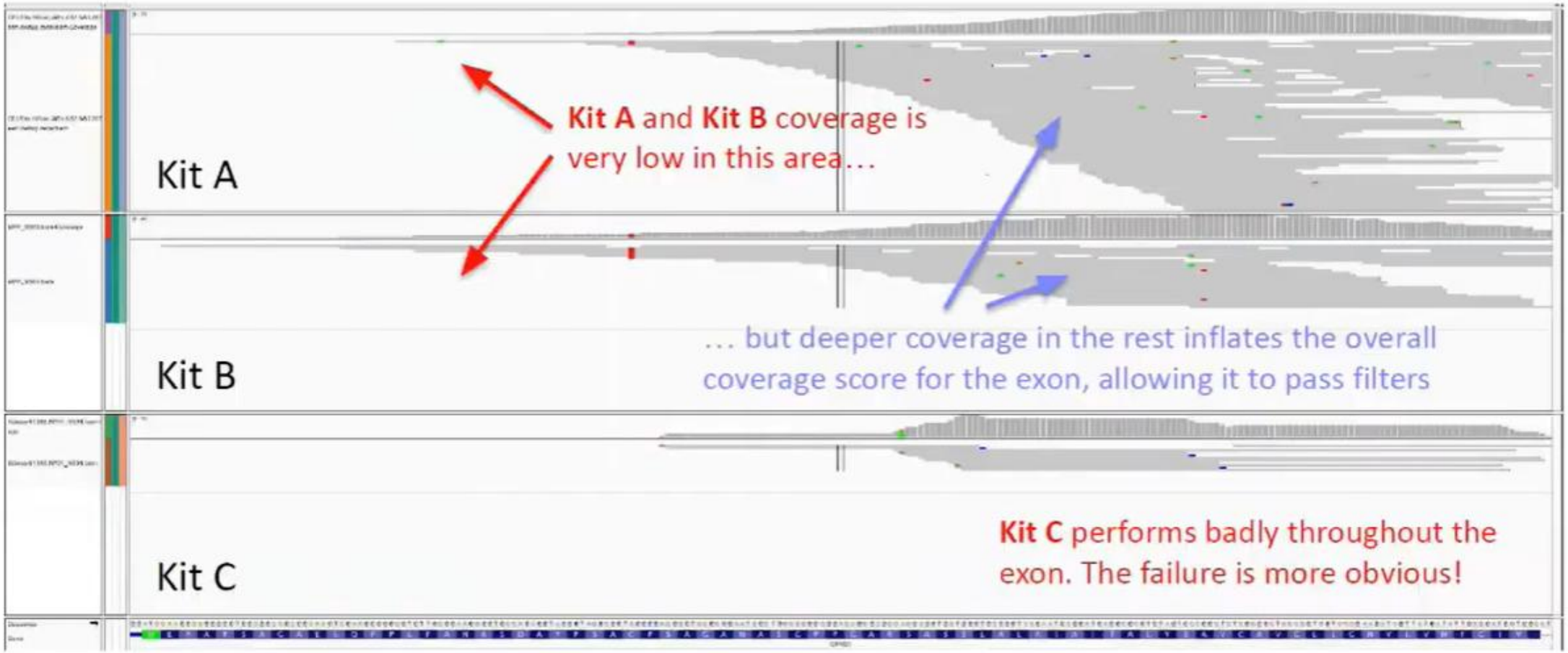
- Severe unevenness in distribution of coverage (“*Fold80*” penalty values)
- Strand biased oxidation artifacts (e.g poor *OxoQ* values for *C* reference bases)

Whole Genomes

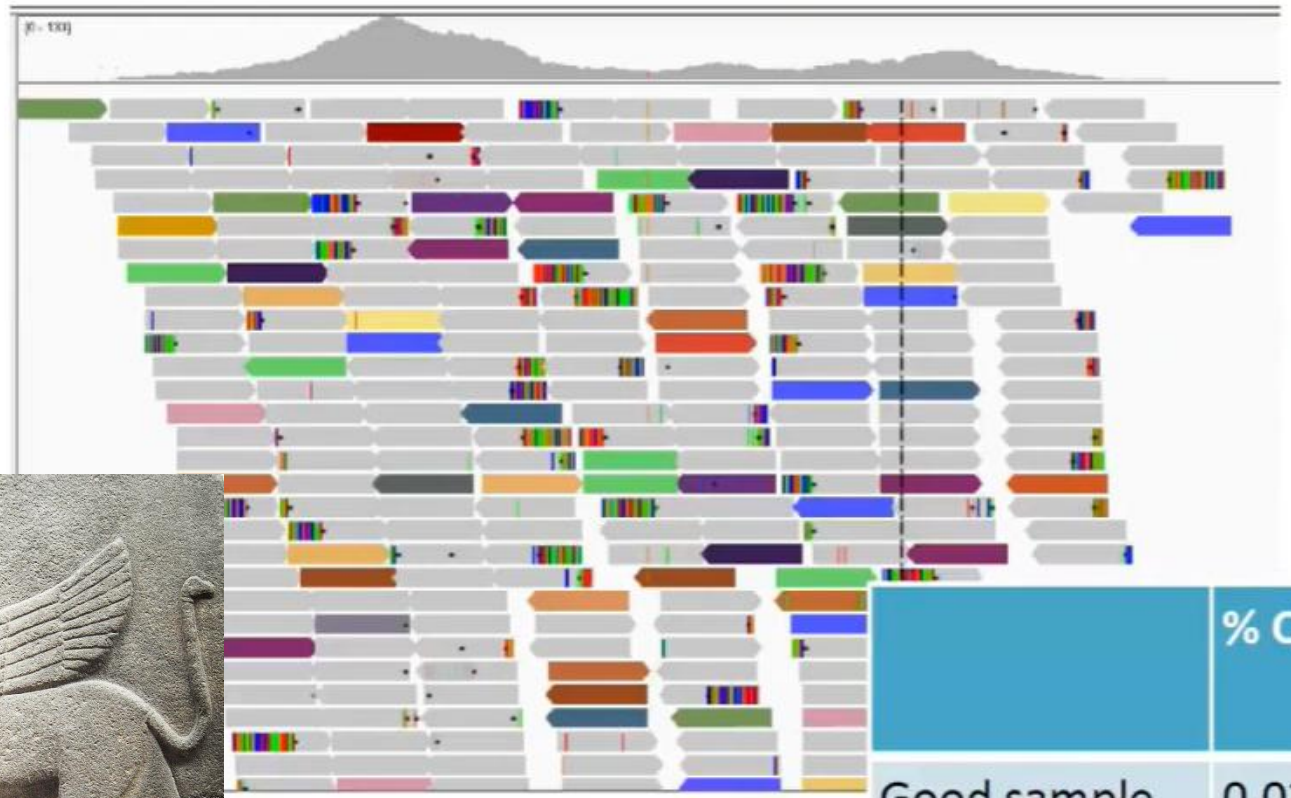
- High proportion of unmapped reads
- High percentage of adapter sequence

Fold80 = 平均深度 / 80%區域覆蓋的深度，Fold80小，覆蓋均一性好，大部分區域的覆蓋深度比較接近。
OxoQ = −10·log10(氧化錯誤率)

Uneven coverage in a PCR-Free whole genome



High percentage of chimerism



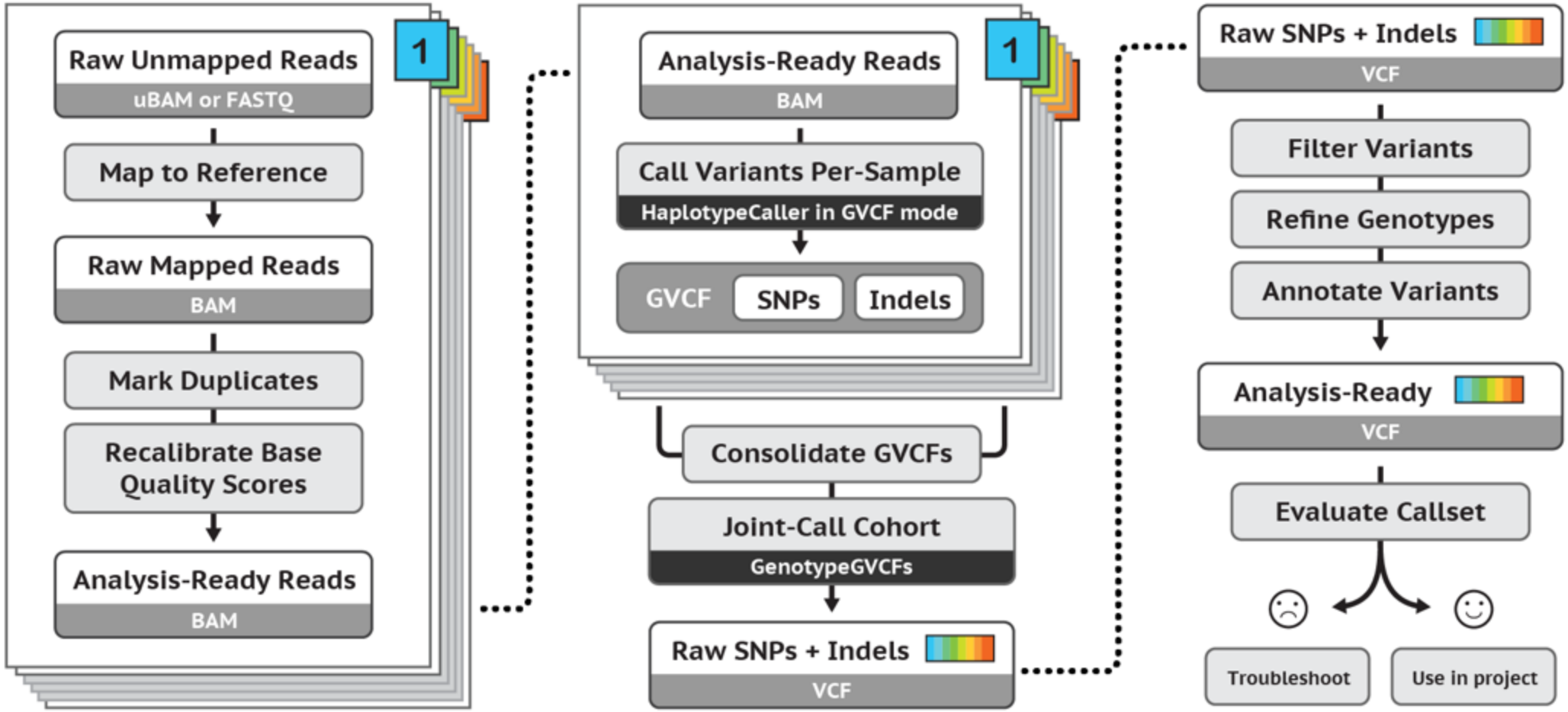
- Reaches coverage goals but data integrity may be an issue as number of chimeric reads is so high.
- This could lead to false positive indel and structural variant calls

	% Chimeras	% Selected bases	% Target bases 20X
Good sample	0.0214	86.296	88.915
Bad sample	30.272	65.438	78.478

Chimeric read 指的是一條 read，同時對參考基因體的兩個不同位置或不同染色體有高相似度比對

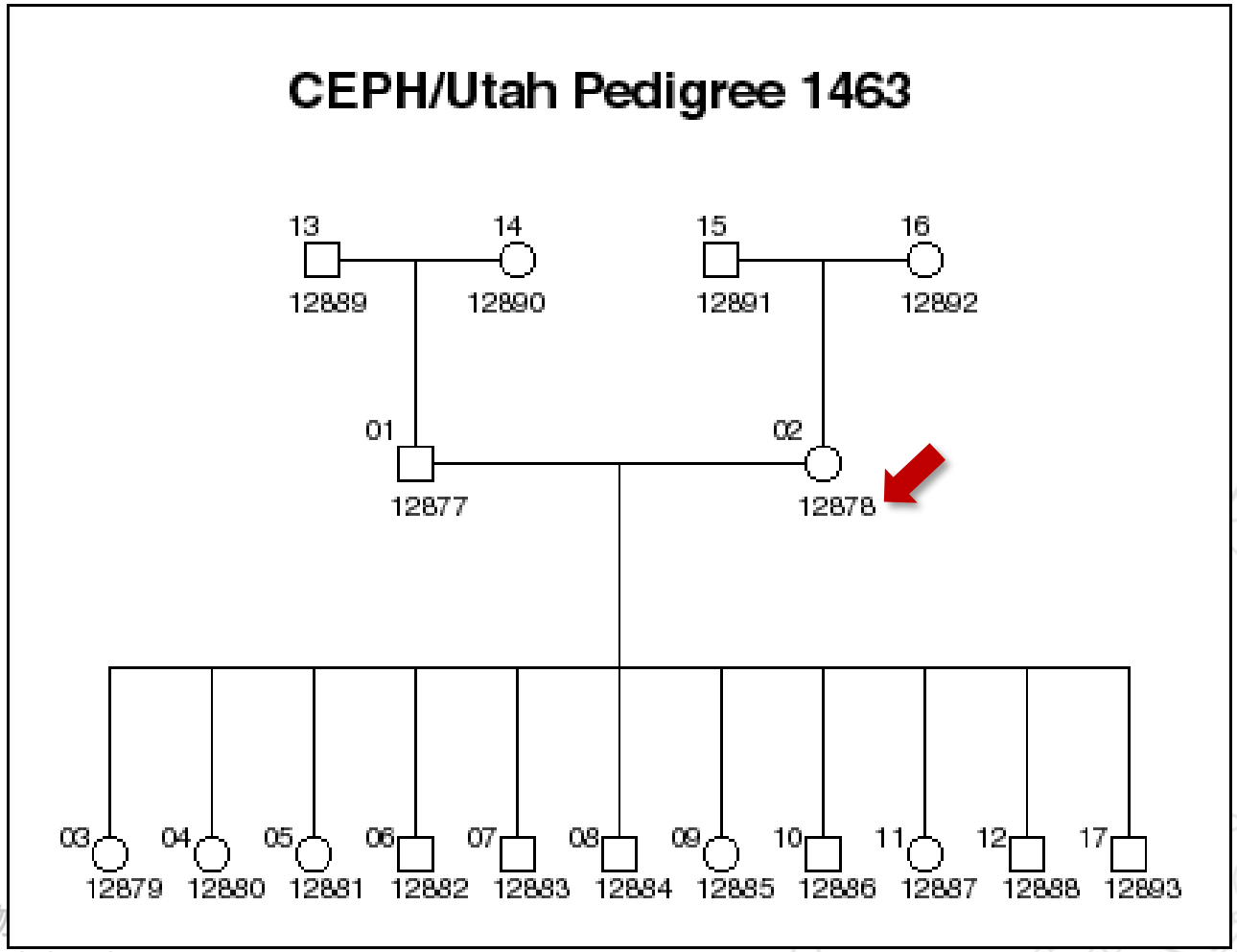
https://zh.wikipedia.org/zh-tw/%E5%96%80%E9%82%81%E6%8B%89

The industry-standard GATK Best Practices



https://gatk.broadinstitute.org/hc/en-us

The most famous family in genomics.



NA12878 was sequenced at a depth of 300x

https://github.com/genome-in-a-bottle

https://catalog.coriell.org/0/Sections/Collections/NIGMS/CEPHFamiliesDetail.aspx?PgId=441&fam=1463&

FastQ (Reads) Format


```
@K00236:146:HLJ5WBBXX:1:1101:0:23614 1:N:0:NTTACT+TAAGAC
NATGGTCTCCATCTCCTGACCTCGTGATCCACCCGCCTCGGCCTCCCAAAATGCTGGGATTACAGGCGTGAGCCACCGTGCCCGGCCTGTTTTTTTTTTCTTTT
TAACCCTTGCCTTGCCTGTTACCTTTCTTCGAGTAAATACATACATA
+
#AAA<AJFAAJFF-FFFAJA<7JFJJJJJJFJJJJJFFFJJJFFJJJJF<-<JJAAJ<FJ7FFJJFFFF-A<JJFJF7-AFJJ<AJJ7<77<F--7-F7FF7-AFJJAJ7<A-AAFA-FF-
JAFJJ7A77FFA-7A7<7FF-AFJAF--77
```

- 1st line is the sequence header which starts with an '@' (not a '>'!).
- 2nd line is the sequence.
- 3rd line starts with '+' and can have the same sequence identifier appended (but usually doesn't anymore).
- 4th line are the quality scores

K00236	the unique instrument name
146	the run id
HLJ5WBBXX	the flowcell id
1	flowcell lane
1101	tile number within the flowcell lane
0	'x'-coordinate of the cluster within the tile
23614	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on
NTTACT+TAAGAC	index sequence

Phred Quality Score Encoding

Formula: score + offset => look for American Standard Code for Information Interchange (ascii) symbol
Two variants: offset=64(Illumina 1.0-before 1.8); offset=33(Sanger, Illumina 1.8+).
A quality score is typically: [0, 40]

```
(33):  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
(64):  @ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
```

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

Why is ASCII code used to record quality scores??

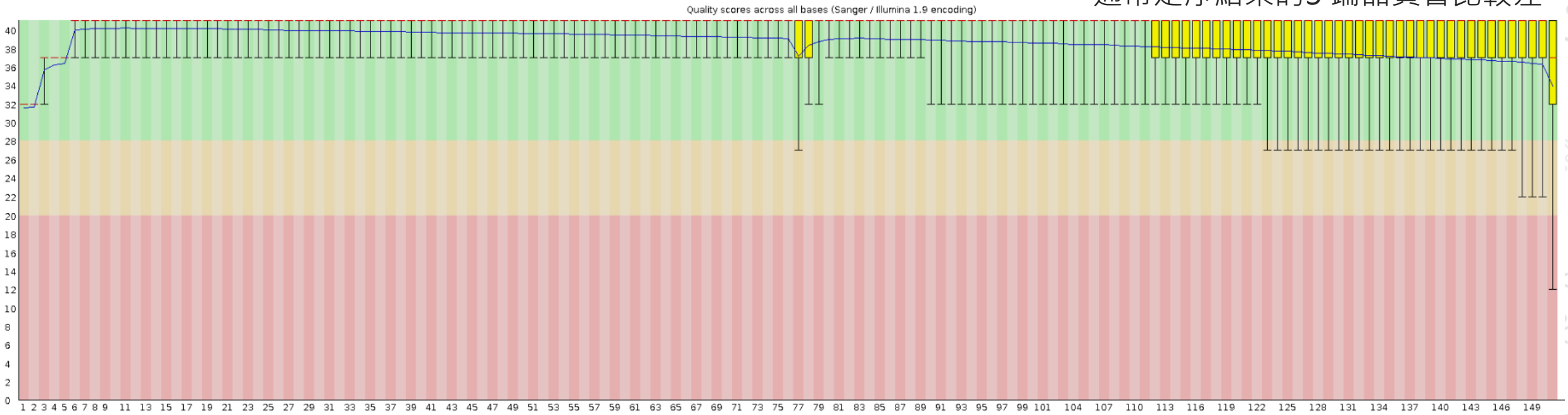
https://learn.gencore.bio.nyu.edu

Basic Statistics

Measure	Value
Filename	NGS1_20170101A_S1_S99_L999_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	464861812
Sequences flagged as poor quality	0
Sequence length	151
%GC	41

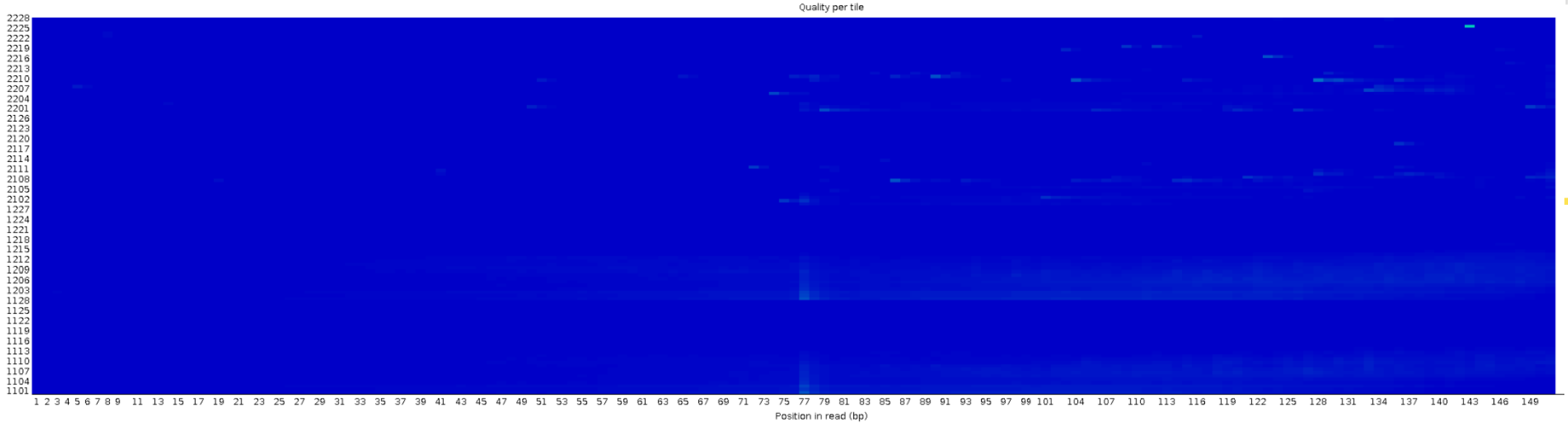
Per base sequence quality

通常定序結果的3'端品質會比較差

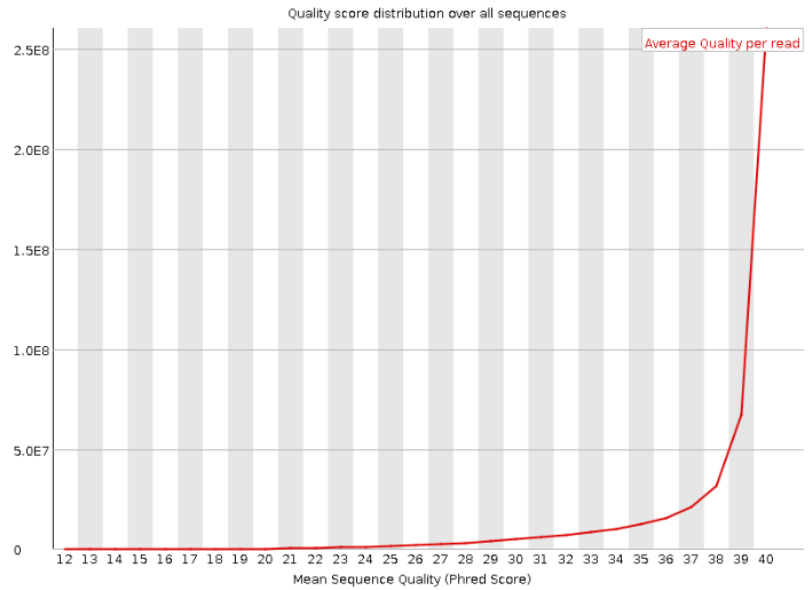


平均 Q >= 30

Per tile sequence quality

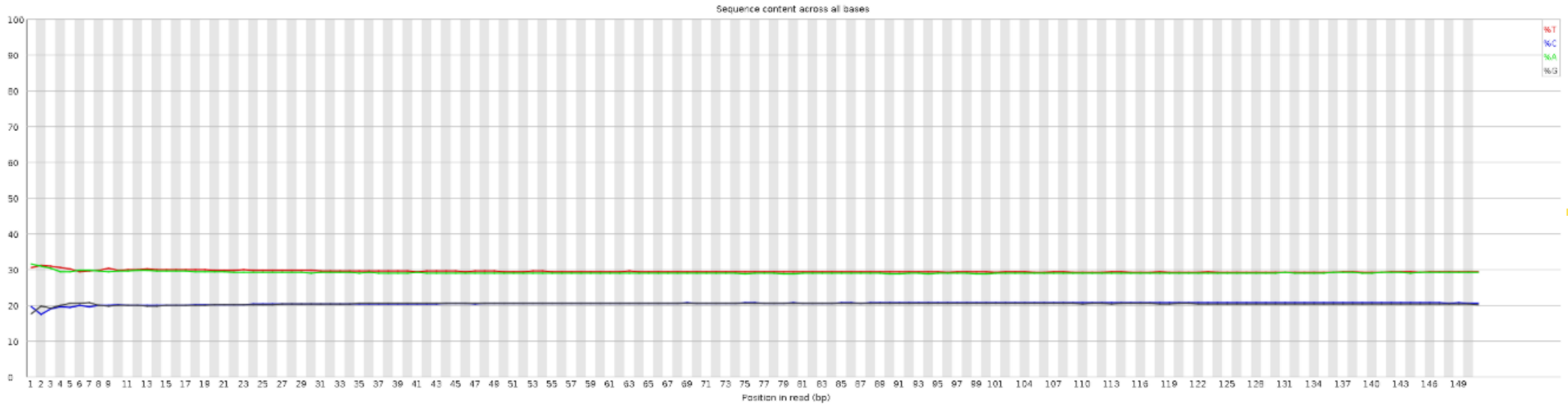


Per sequence quality scores

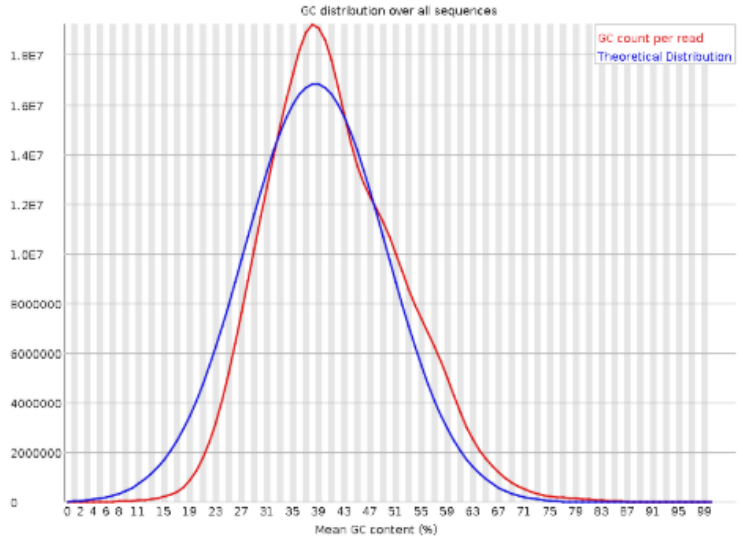


- Per tile sequence quality 顯示 fail 或者 warning，表明某個 lane 或某個 run 中出現出現了部分故障，從而影響一些特定的區域。
- Per sequence quality scores 可用來看是否有特別低品質的定序序列。經過 quality trimming 後可能會偏差，或是看不出 error batch?

Per base sequence content



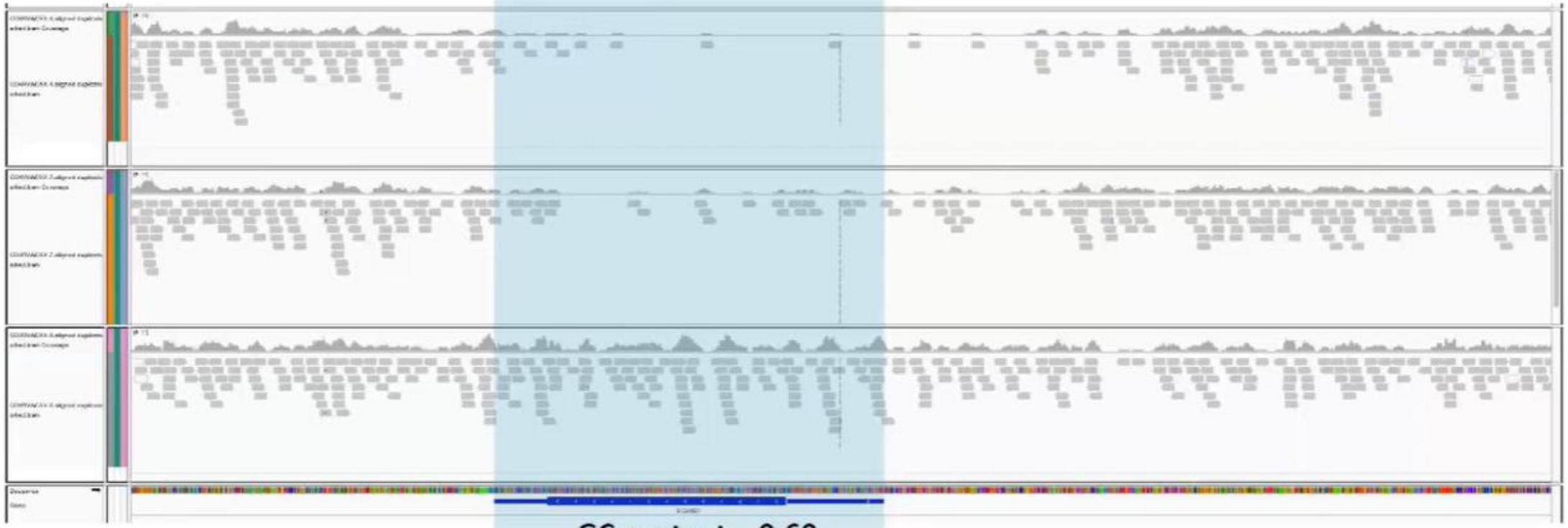
Per sequence GC content



- Per base sequence content 每一位置的ATCG含量。前處理經過隨機片段化成短序列，故位置的含量應是平均的。已有學者發現RNA-Seq接近5‘端的鹼基含量特別容易出現明顯差異。
- Per sequence GC content GC 含量的分布，藍色為整體 GC 含量平均值所繪製的理論分布，紅色為實際計算個別序列 GC 含量。若紅色分布線偏離藍色分布線且出現一個或多個突起的峰值，則可推論樣本當中有序列過度表現或含有不屬於定序物種的序列存在。(人類基因體約 40%)

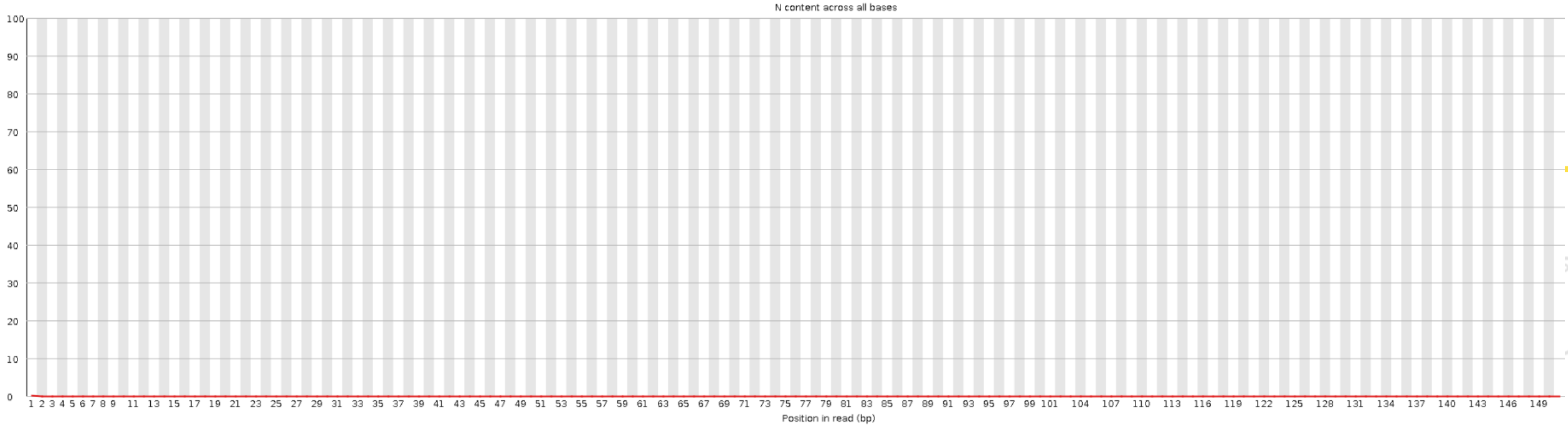
High GC regions affect coverage

This GC-rich region is poorly covered in the first two samples -> **unusable**



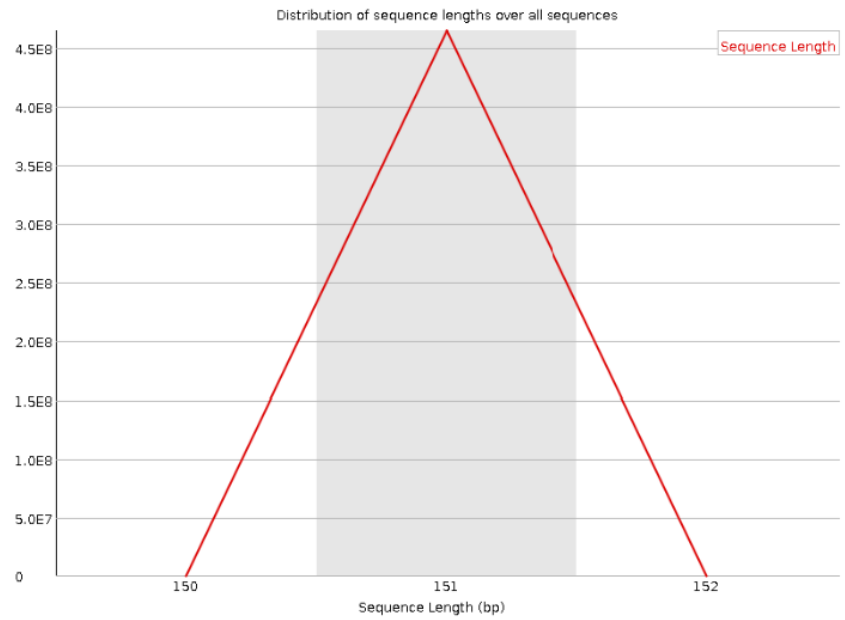
GC content = 0.69

Per base N content

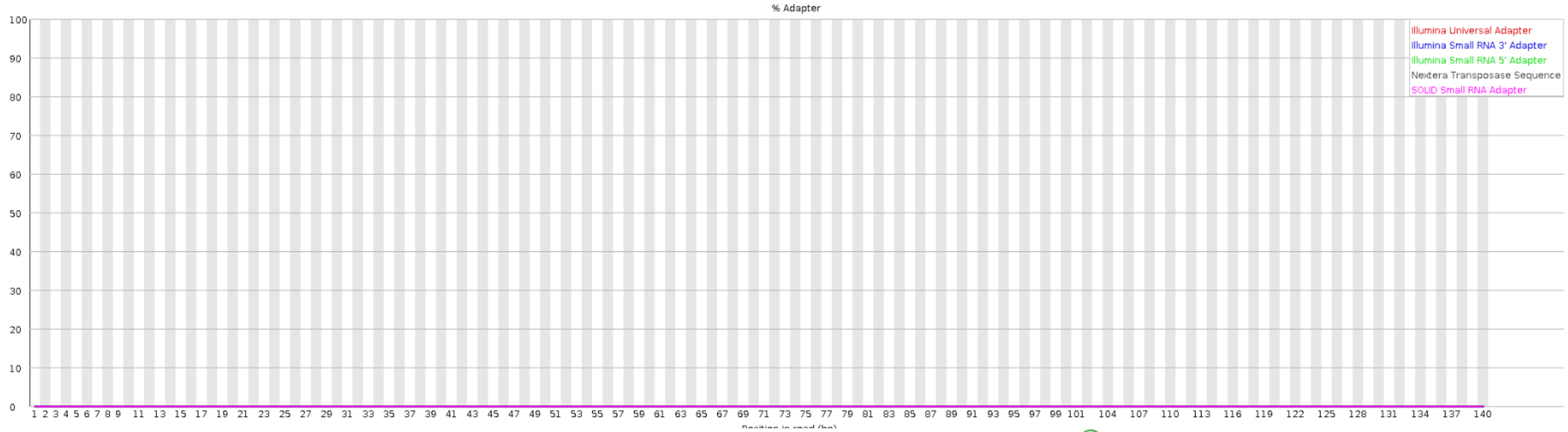


- Per base N content 各位置為N的百分比，用於評估在哪個位置以後可能不適合後續分析，需要進行序列裁切(trimming)。

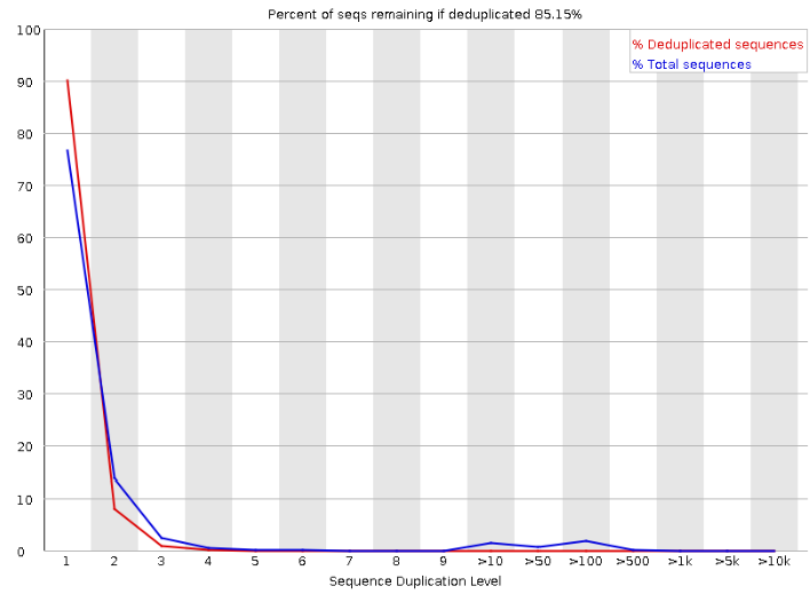
Sequence Length Distribution



Adapter Content



Sequence Duplication Levels



- duplicate sequences 重複序列出現的比例。藍色線為所有序列中重複序列重複次數的分布，紅色線為刪去重複序列後的序列數量作為分母除重複序列筆數的分布。提醒定序資料是否有過度表現的序列影響整體資料的變異度(diversity)。
- overrepresented sequences 重複出現且佔整體定序資料0.1%的序列，被列為過度表現序列。
- Adapter Content 定序資料內的adapter有哪些。

Overrepresented sequences

No overrepresented sequences

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence

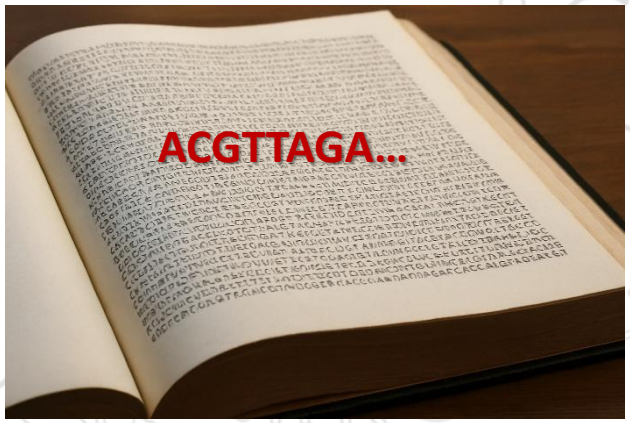
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

https://www.majordifferences.com/2016/05/difference-between-pairwise-and-multiple-sequence-alignment.html

BWA MEM

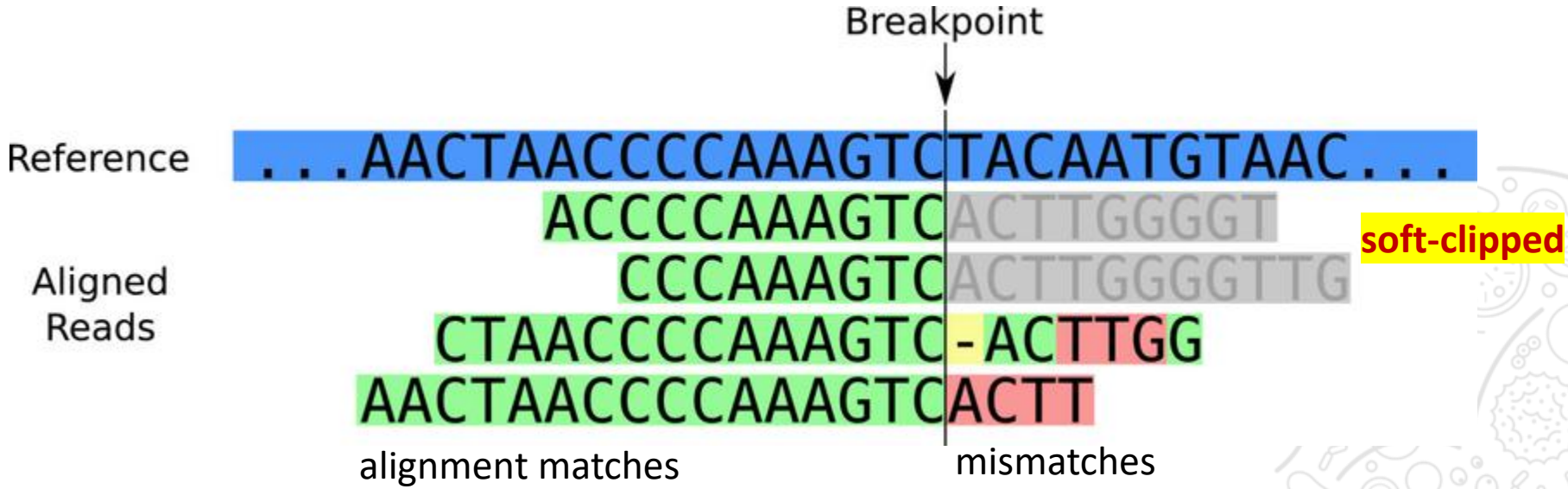
- BWA = Burrows Wheeler Aligner (uses BW transform to compress data)
- MEM = Maximal Exact Match (how alignment “seeds” are chosen)
- **Performs local alignment** (rather than end-over-end)
- Can **clip ends of reads**, if they do not match
- Can **split a read** into pieces, mapping each separately
- **Performs gapped alignment**
- **Utilizes PE reads** to improve mapping
- **Reports only one alignment** for each read
- If ambiguous, one of the equivalent best locations is chosen at random
- Ambiguously mapped reads are reported with low **Mapping Quality**




```
$ bwa mem -t 4 -R '@RG\tID:foo_lane\tPL:illumina\tLB:library\tSM:sample_name'
/path/to/human.fasta read_1.fq.gz read_2.fq.gz | samtools view -S -b - > align.bam
```


Li H. et al. (2009). *Bioinformatics*, 25:1754-60.

Soft-clipped reads vs. improperly aligned reads at a structural variant breakpoint.



All four reads are sampled from the variant sequence and span the breakpoint, but only the top two are properly soft-clipped. The bottom two reads are aligned with mismatches and skips because the portion that should be soft-clipped is only a few bases long and the global alignment is used instead. Inaccurate soft-clipping can lead to false negatives for split-read based SV prediction algorithms and also lowers the accuracy of SHEAR's heterogeneity estimation algorithm.

https://pmc.ncbi.nlm.nih.gov/articles/PMC4007568/

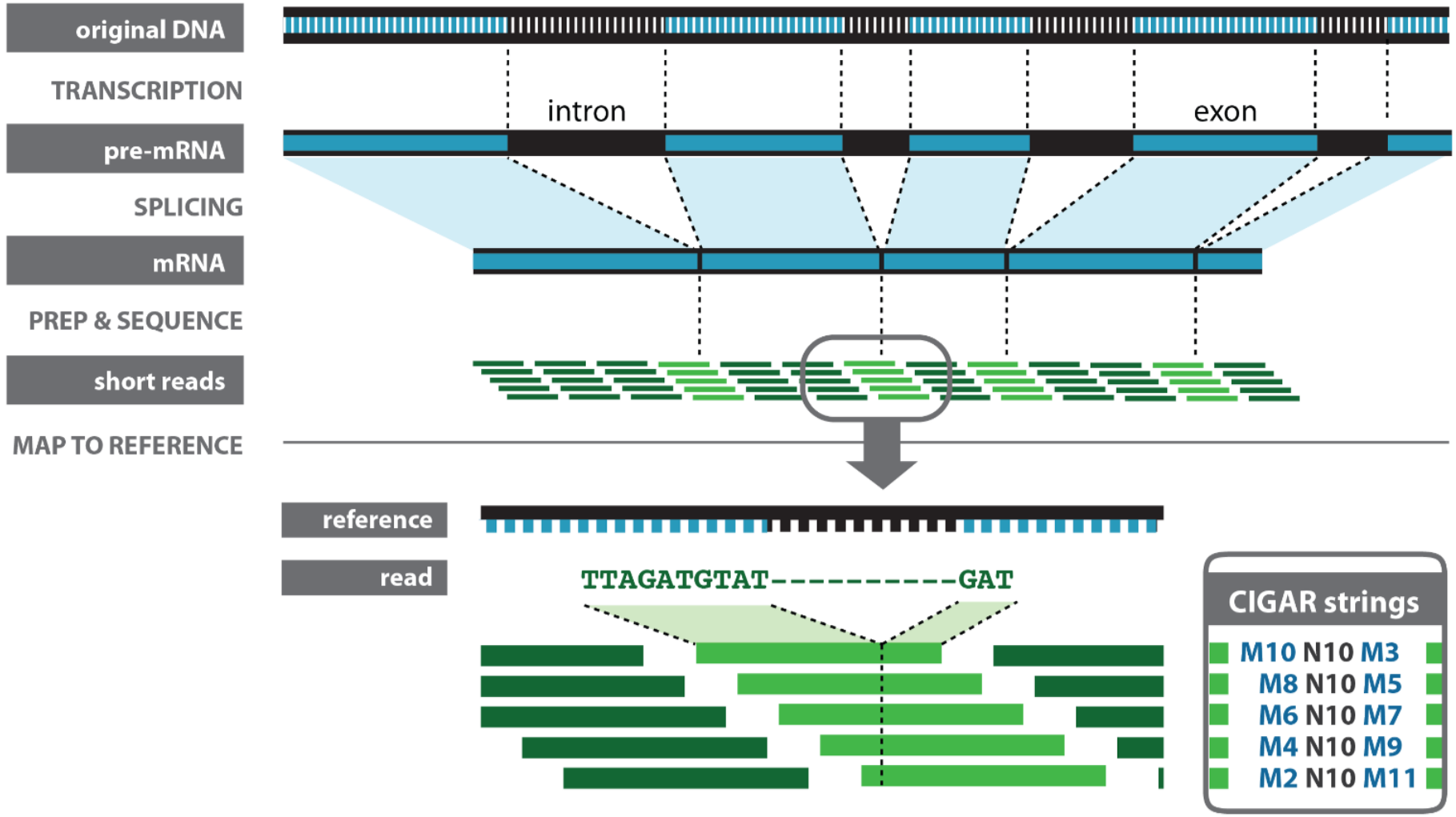
When trimming may not be necessary

- **Applications like RNA-seq:** For some applications, such as differential gene expression analysis using tools like Sailfish, a **quasi-mapping approach is robust enough** to handle low-quality and adapter-contaminated reads without the need for trimming.
- **Short, high-quality reads:** If the **quality** of the reads is already **very high**, trimming may not provide any benefit and could potentially remove useful data.
- **Ancient DNA sequencing:** In this case, degradation is a key feature of the DNA. Trimming would **remove the very information** you are trying to analyze.

When trimming might be harmful

- **De novo assembly:** Some studies have shown that **excessive trimming** can be harmful to de novo assembly and transcriptome characterization, potentially removing reads that could be used for assembly.
- **Variable read length:** If you trim a variable amount from each read based on its quality, you can **create regions of low coverage** in your data, which can negatively impact downstream analysis.

RNAseq reads mapped across splice junc)ons need special handlin~



Mapped seq. files in sam format

```
@SQ     SN:chr2L        LN:23011544
@SQ     SN:chr2LHet     LN:368872
@SQ     SN:chr2R        LN:21146708
@SQ     SN:chr2RHet     LN:3288761
@SQ     SN:chr3L        LN:24543557
@SQ     SN:chr3LHet     LN:2555491
@SQ     SN:chr3R        LN:27905053
@SQ     SN:chr3RHet     LN:2517507
@SQ     SN:chr4         LN:1351857
@SQ     SN:chrM         LN:19517
@SQ     SN:chrX         LN:22422827
@SQ     SN:chrXHet      LN:204112
@SQ     SN:chrYHet      LN:347038
@RG     ID:SRR1663609   SM:ZW177        LB:ZW155        PL:ILLUMINA
@PG     ID:bwa  PN:bwa  VN:0.7.8-r455   CL:bwa mem -M -t 4 -R @RG\tID:SRR1663609\tSM:ZW177\tLB:ZW155\tPL:ILLUMINA
/local_data/Drosophila_melanogaster_dm3/BWAIndex/genome.fa SRR1663609_1.fastq.gz SRR1663609_2
.fastq.gz
SRR1663609.1  97   chrX      2051224  60   6M54S     chrYHet     4586 0   GGATCGTGAT… gggfgg[gfg… NM:i:0  MD:Z:46     AS:i:46  XS:i:0   RG:Z:SRR1663
SRR1663609.1 145   chrYHet      4586   0   100M      chrX     2051224 0   ACTTCTCTTC… BBBBBbdd]c… NM:i:0  MD:Z:100    AS:i:100 XS:i:99  RG:Z:SRR1663
SRR1663609.2  65   chr3RHet  2308288   0   100M      chrYHet     4712 0   AGAAGAGAAG… Y_b`_ccTccB… NM:i:0  MD:Z:100    AS:i:100 XS:i:100 RG:Z:SRR1663
SRR1663609.2 129   chrYHet      4712  60  38M62S     chr3RHet 2308288 0   CTTCTCTTCT… eeeae`edee… NM:i:1  MD:Z:17T20  AS:i:33  XS:i:21  RG:Z:SRR1663
SRR1663609.3  65   chr3RHet  2308278   0   100M      chrYHet     4649 0   AGAAGAGAAG… fffffffffff… NM:i:0  MD:Z:100    AS:i:100 XS:i:100 RG:Z:SRR1663
SRR1663609.3 129   chrYHet      4649   0  41M59S     chr3RHet 2308278 0   TCTCTTCTCT… fffffffff…   NM:i:0  MD:Z:41     AS:i:41  XS:i:41  RG:Z:SRR1663
SA:Z:chrX,5036484,-,16S41M43S,0,2;
SRR1663609.3 401   chrX      5036484   0  16H41M43H chr3RHet 2308278 0   AAAAGAAGAA… BBBBBBBBBB… NM:i:2  MD:Z:7A4G28 AS:i:31  XS:i:28  RG:Z:SRR1663
SA:Z:chrYHet,4649,+,41M59S,0,0;
SRR1663609.4  99   chr3RHet   854491   0   100M          =    854876 485   AGAAGAAGAA… BBBBBBBBBB… NM:i:0  MD:Z:100    AS:i:100 XS:i:100 RG:Z:SRR1663
SRR1663609.4 147   chr3RHet   854876   0   100M          =    854491 -485  GAGAAGAGAA… fffffffffff… NM:i:0  MD:Z:100    AS:i:100 XS:i:100 RG:Z:SRR1663
```

Reference genome information

Read group information

Header

Program infromation

read name

chr

flag

position on chr

mapping quality

CIGAR string

chr of mate

mate position on chr

frag length

Read sequence

Read qualities

(shortened for clarity)

edit dist

match str

best aln score

next aln score

Read group

TAGS

flag

145 = 00010010001

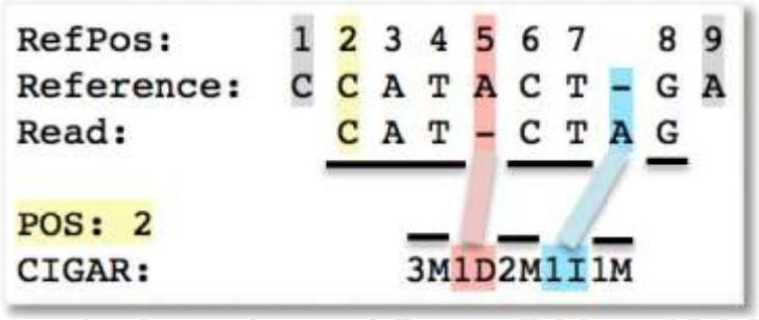
- Read is a part of a fragment in sequencing (000000000001) – they all do, because our data is all PE reads
- Read aligns to reference as reverse complement (00000010000)
- Read is the second end of the fragment (00010000000)

Binary	Hex Bit	Description
000000000001	0x1	template having multiple segments in sequencing
000000000010	0x2	each segment properly aligned according to the aligner
000000000100	0x4	segment unmapped
000000001000	0x8	next segment in the template unmapped
000000010000	0x10	SEQ being reverse complemented
000000100000	0x20	SEQ of the next segment in the template being reversed
000001000000	0x40	the first segment in the template
000010000000	0x80	the last segment in the template
000100000000	0x100	secondary alignment
001000000000	0x200	not passing quality controls
010000000000	0x400	PCR or optical duplicate
100000000000	0x800	supplementary alignment

CIGAR string

CIGAR: Compact Idiosyncratic Gapped Alignment Report – shows how many indels, how many bases soft- or hard-clipped

- **100M** whole read aligned (no clips), no indels
- **16H41M43H** 16 bp clipped from the beginning of the read, 43 bp clipped from the end, 41 remaining bases aligned with no indels
- **52S48M** 52 bp soft-clipped from the beginning (i.e., these bases are still shown, but do not take part in alignment), the other 48 aligned without indel
- **3M1D2M1I1M** 3 bases aligned followed by 1 base deleted, 2 next ones aligned, 1 base inserted and the last one aligned



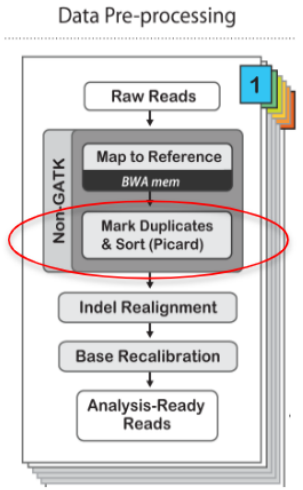
TAGs

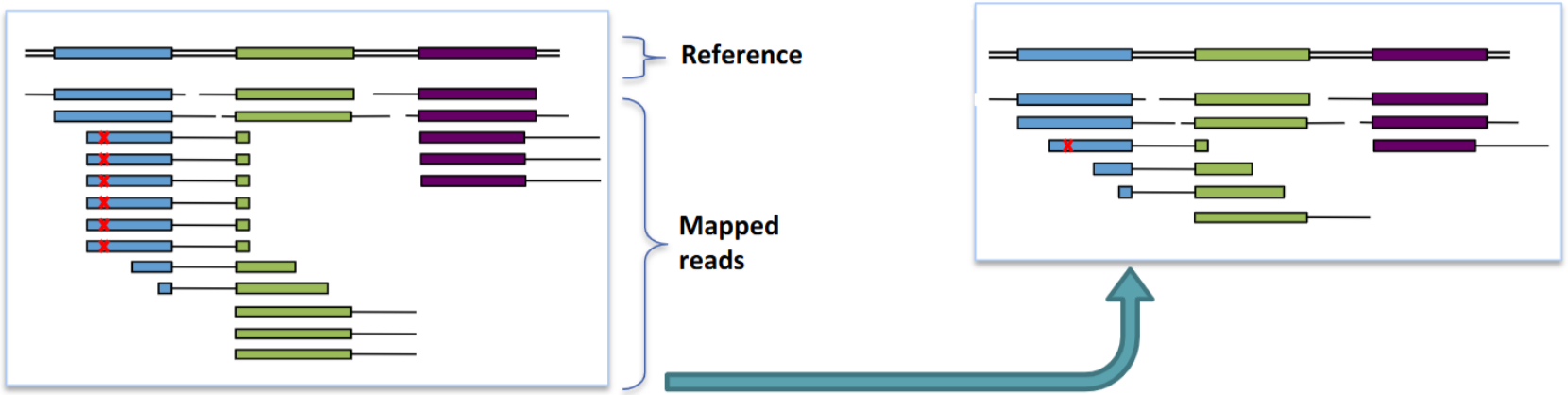
TAG	Example	What it means
NM	NM:i:1	Number of mismatches (integer value)
MD	MD:Z:17T20	17 matches, then some other base in place of T, then 20 more matches (counting from beginning of read)
AS	AS:i:100	Alignment score 100 (integer)
XS	XS:i:21	Second-best alignment score 21 (integer)
RG	RG:Z:SRR166309	Read Group ID
SA	SA:Z:chrYHet,4649,+,41M59S,0,0	Location and tags of second-best hit

Mark duplicates

- Duplicates = **non-independent measurements** of a sequence
  - Sampled from same template of DNA
  - Violates assumptions of variant calling
- Errors in sample/library prep will get propagated to *all* the duplicates

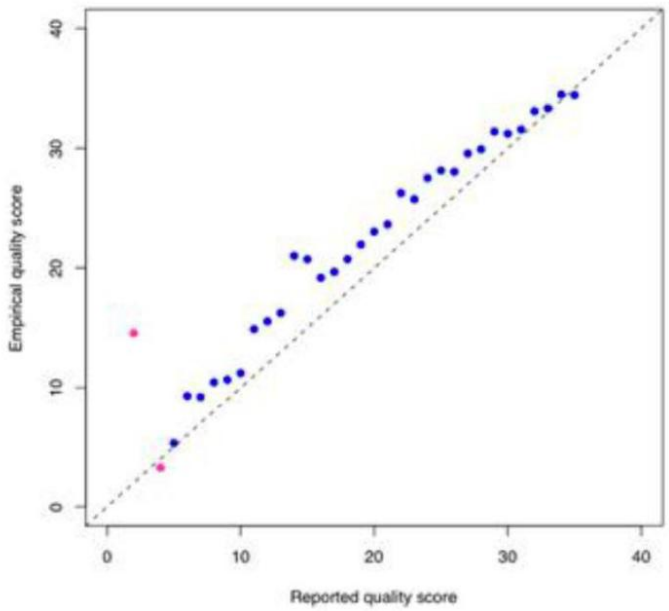
-> Just pick the "best" copy – mitigates the effects of errors

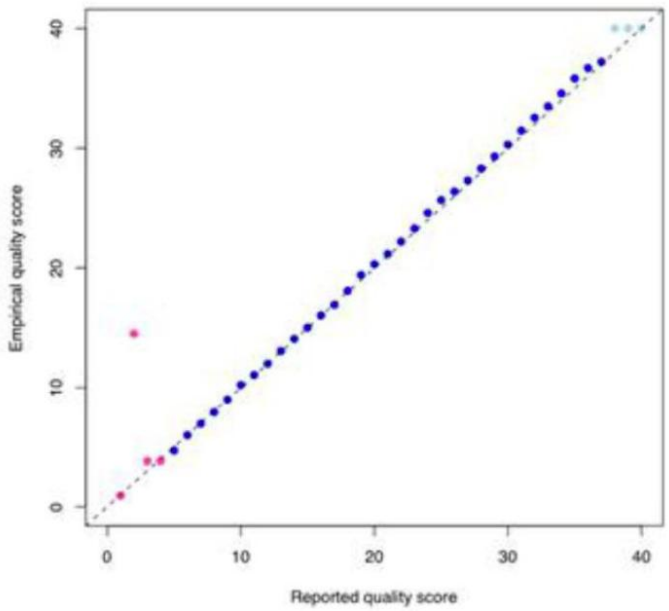




✖ = sequencing error propagated in duplicates

Recalibrate Base Quality Score





分析錯誤模式：工具會掃描整個 BAM 檔（定序結果），找出哪些品質分數、鹼基組合、定序位置（cycle）容易出錯。例如：發現「AA 後面接的鹼基」錯誤率偏高 → 那就調低這類情況的品質分數。

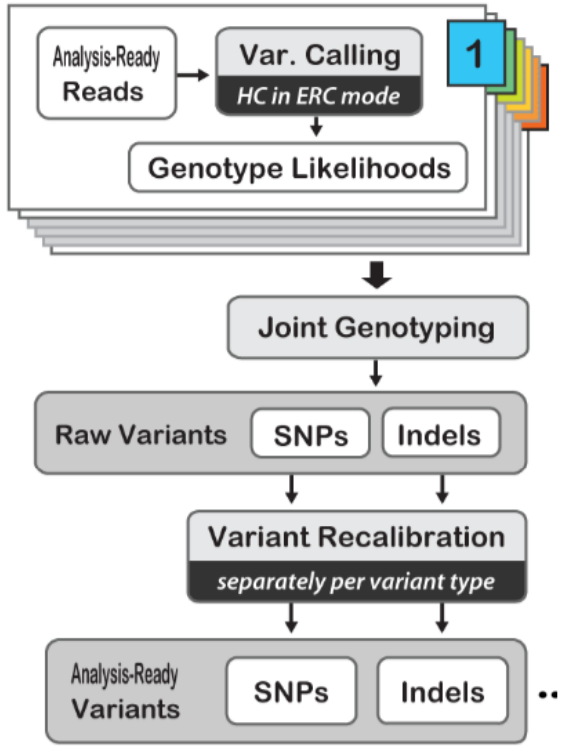
建立錯誤模型：利用已知的變異資料（例如 dbSNP）來排除真正的生物變異，只針對「機器錯誤」來建模。

重新調整品質分數：根據模型，對每個鹼基的品質分數進行加減調整，產生新的 BAM 檔

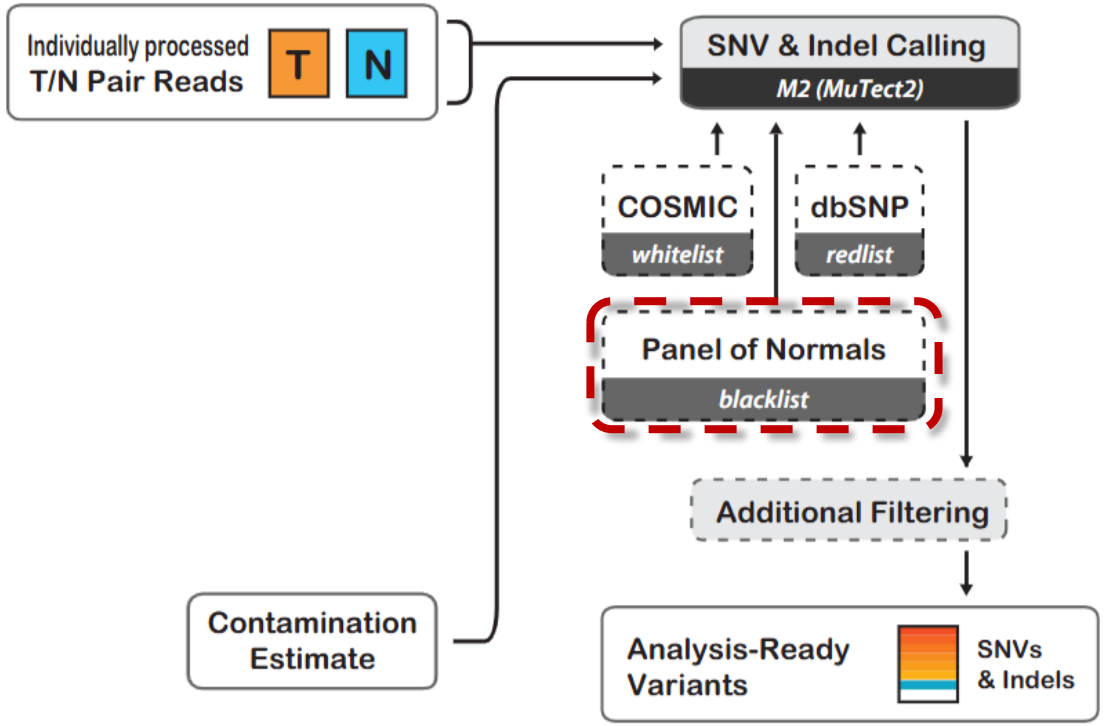
```
java -jar /path/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R /path/to/human.fasta \
-I sorted.markdup.realign.bam \
--knownSites /path/1000G_phase1.indels.b37.vcf \
--knownSites /path/Mills_and_1000G_gold_standard.indels.b37.vcf \
--knownSites /path/to/gatk/bundle/dbsnp_138.b37.vcf \
-o recal_data.table

java -jar /path/GenomeAnalysisTK.jar \
-T PrintReads \
-R /path/to/human.fasta \
-I sample_name.sorted.markdup.realign.bam \
--BQSR recal_data.table \
-o sorted.markdup.realign.BQSR.bam
```

Germline vs. SomaKc Variant Discovery

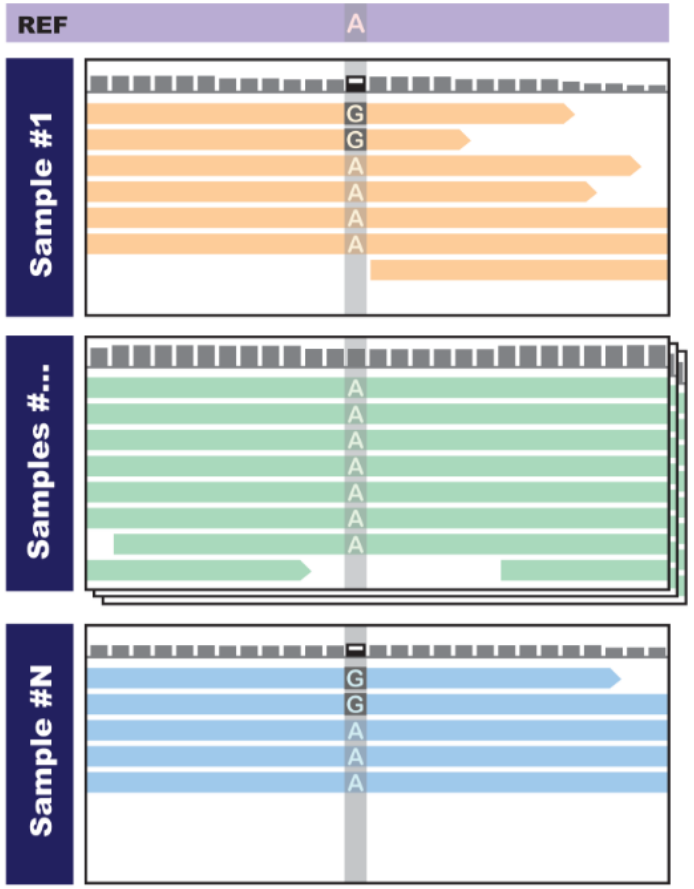
Germline SNPs and Indels



Somatic SNPs and Indels

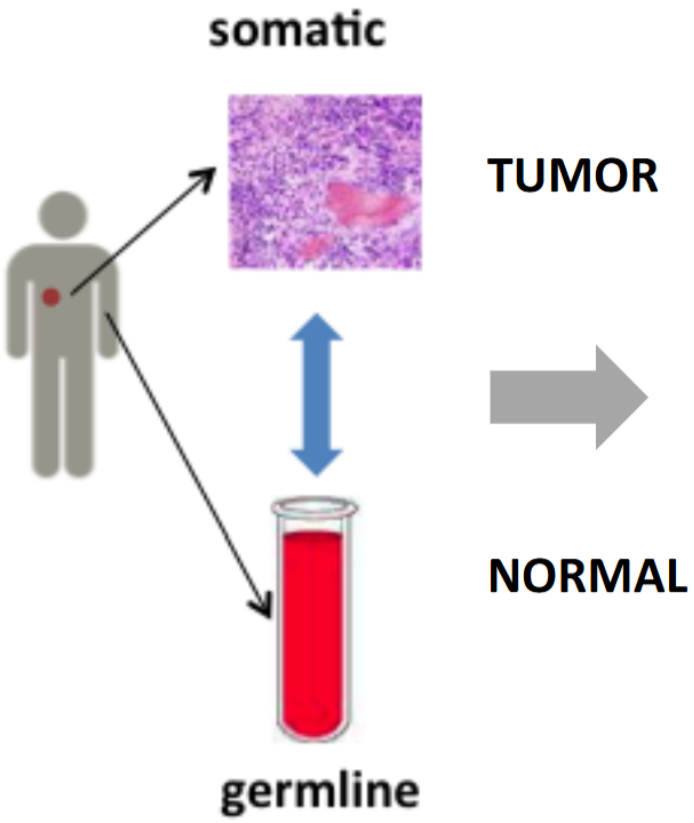


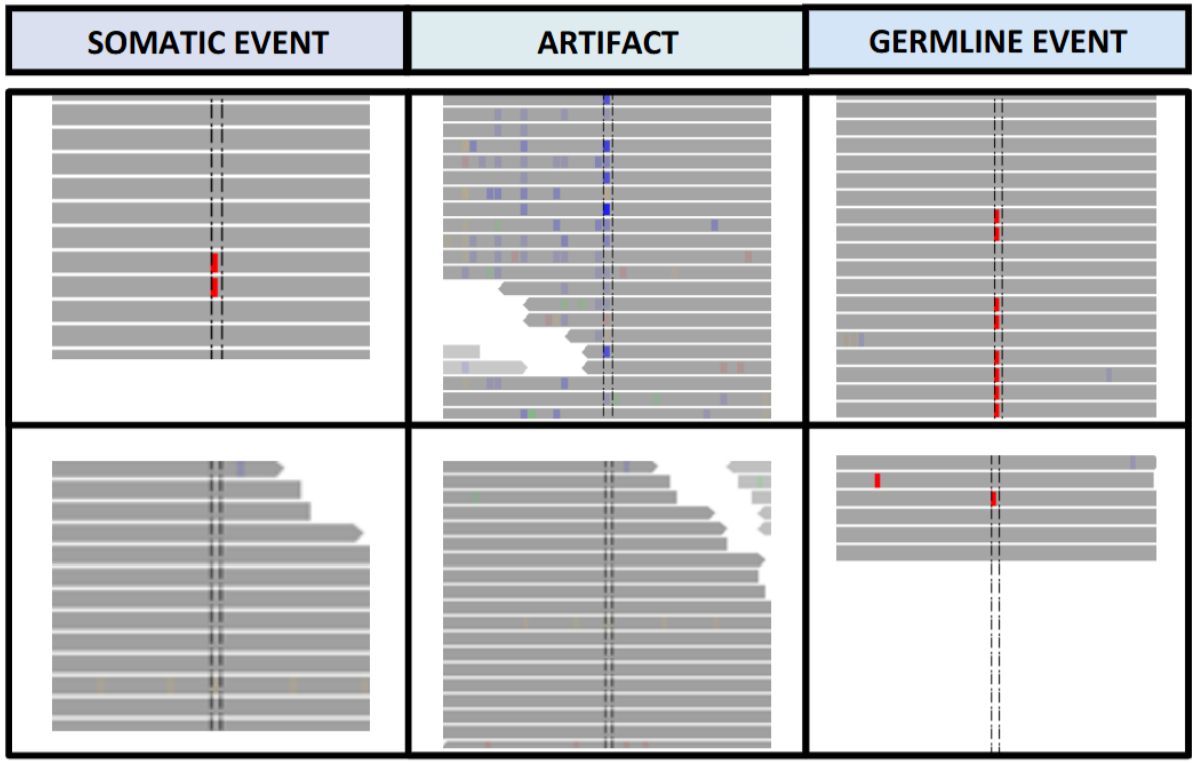
Joint analysis empowers discovery at difficult sites



- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**
- Both samples seen together:
  - **unlikely to be artifact**
  - **call the variant more confidently**

Matched Tumor/Normal samples from the same individual





Filter variants by hard-filtering

```
gatk VariantFiltration \
   -V snps.vcf.gz \
   -filter "QD < 2.0" --filter-name "QD2" \
   -filter "QUAL < 30.0" --filter-name "QUAL30" \
   -filter "SOR > 3.0" --filter-name "SOR3" \
   -filter "FS > 60.0" --filter-name "FS60" \
   -filter "MQ < 40.0" --filter-name "MQ40" \
   -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \
   -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \
   -O snps_filtered.vcf.gz
```

參數名稱	意義	建議過濾門檻
QD (QualByDepth)	將變異的信心分數（QUAL）除以有效覆蓋深度，避免深度過高造成品質膨脹	QD < 2.0
FS (FisherStrand)	使用 Fisher's Exact Test 評估是否有定序方向偏差（strand bias）	FS > 60.0
SOR (StrandOddsRatio)	另一種 strand bias 評估方式，對高覆蓋資料更穩定	SOR > 3.0
MQ (RMSMappingQuality)	所有 Reads 比對品質的均方根值，反映整體比對可信度	MQ < 40.0
MQRankSum	比較支持參考與變異等位基因的 Reads 比對品質差異	MQRankSum < -12.5
ReadPosRankSum	比較參考與變異等位基因在 Reads 中的位置差異，偏向讀段末端可能是錯誤	ReadPosRankSum < -8.0

https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

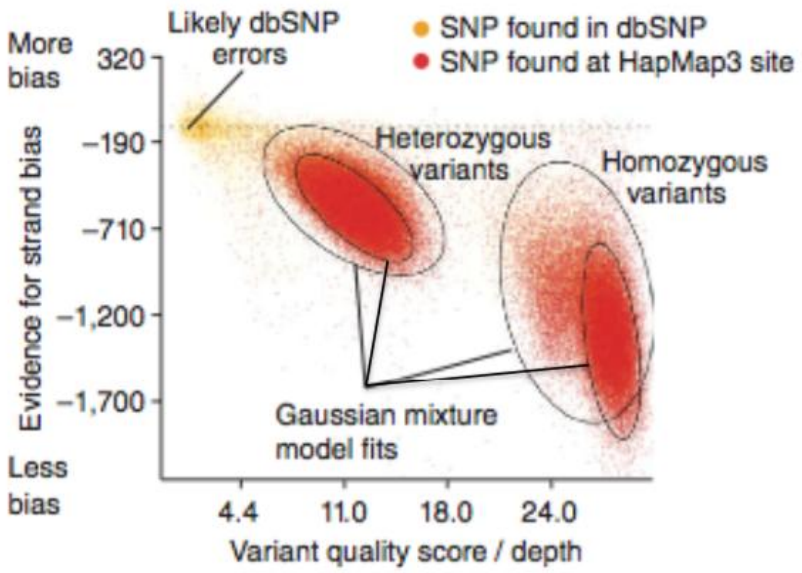
Filter variants by Variant Quality Score Recalibration (VQSR)

項目	**VQSR**（機器學習式）	**Hard Filtering**（門檻式）
原理	使用機器學習建立變異品質模型	根據固定參數門檻進行過濾
依據	多個註解參數的統計分布（如 QD、FS、MQ 等）	每個註解參數設定一個過濾值
彈性	高：根據資料特性自動調整	低：門檻固定，不考慮資料分布
輸出分數	VQSLOD（變異品質對數比）	無分數，直接標記 PASS 或 FILTER
資料需求	建議使用在大樣本（>30 samples）資料集	適用於小樣本或單一樣本分析
訓練資料	需要已知高品質變異集（如 HapMap、Omni）	不需要訓練資料
優點	更準確、可調整、適合大規模分析	簡單快速、容易理解與實作
缺點	計算成本高、參數設定複雜	容易過濾掉真實變異或保留假陽性

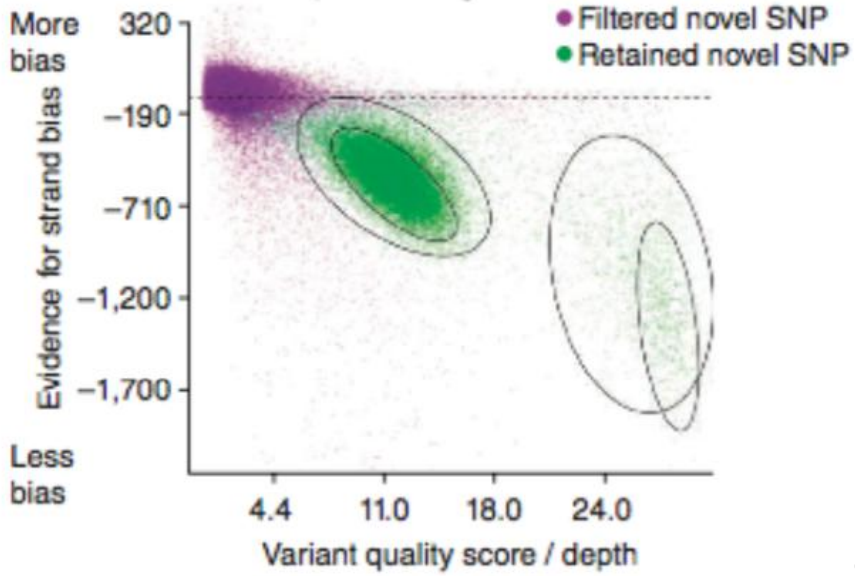
Basic idea

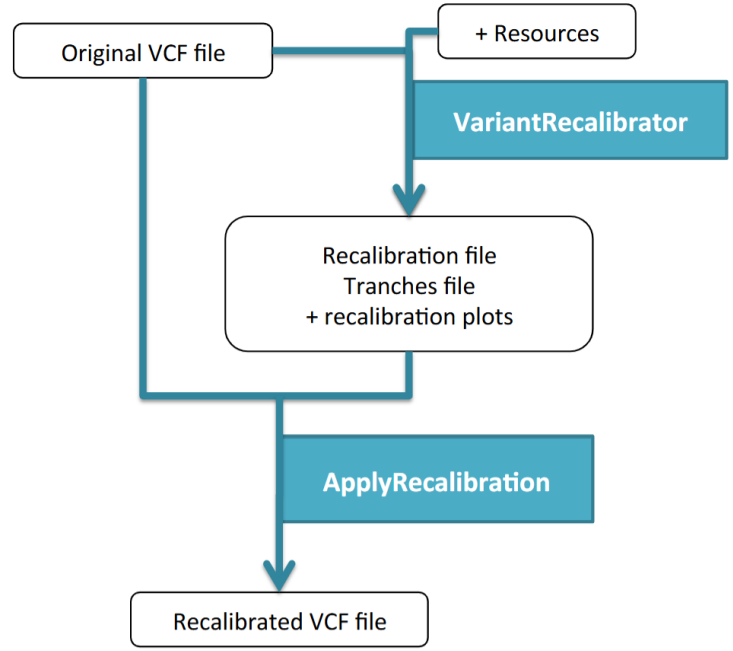
Basic idea: training on high-confidence known sites to determine the probability that other sites are true

(1) Train model using HapMap



(2) Apply model to callset

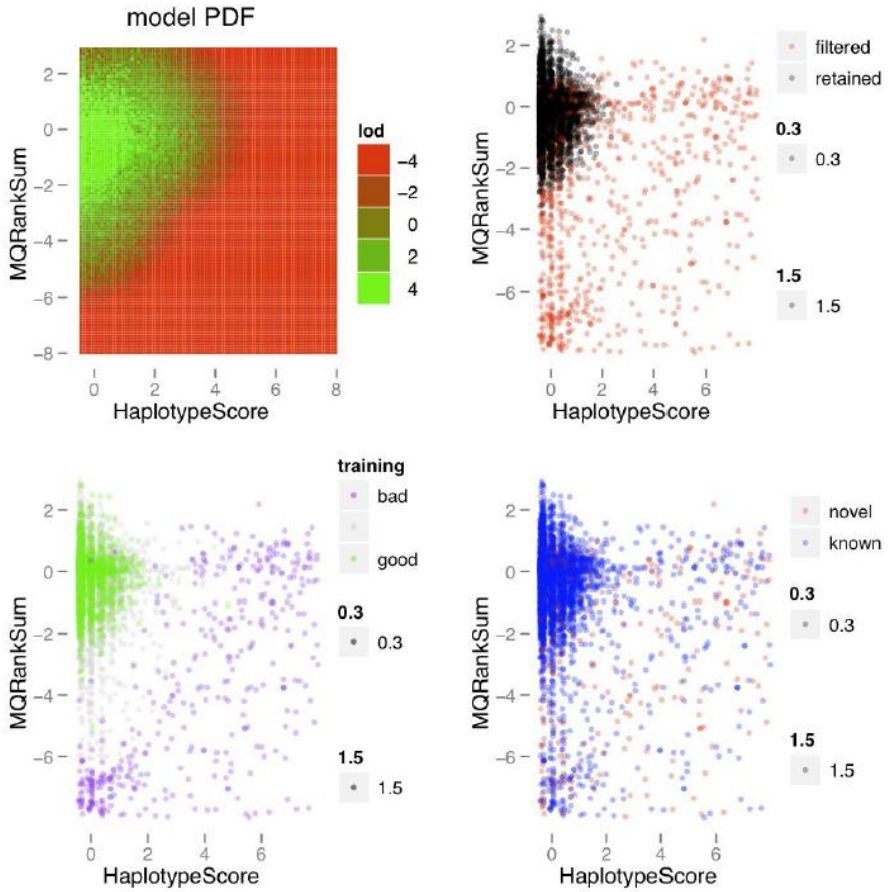


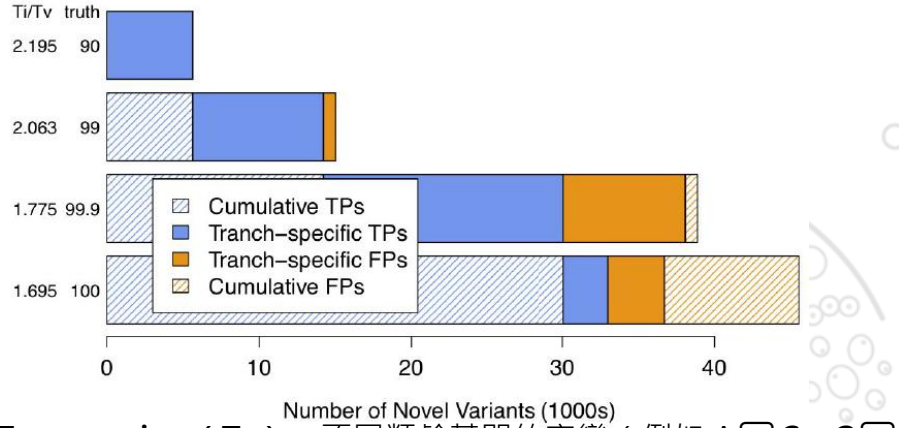


what good variants look like

Modified from DePristo et al. Nature Genetics. 2011

How?





- **Transversion（Tv）**：不同類鹼基間的突變（例如 A↔C、G↔T 等）
- **Transition（Ti）**：同類鹼基間的突變（例如 A↔G 或 C↔T）
- WGS 2.0 ~ 2.2

- VQSR 要做的事是：**根據變異的註解特徵（如 QD、FS、MQ 等），判斷它是「真實」還是「錯誤」**。
- Gaussian Mixture Model 是一種統計模型，它假設資料是由多個高斯分布（鐘型曲線）混合而成。
- GATK 使用已知的高品質變異（如 HapMap、Omni）來訓練 GMM，建立「真實變異」的分布模型。同時也建立一個「錯誤變異」的模型，通常是從資料中推估。
- VQSLOD = log10(真實機率 / 錯誤機率)

https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR

SNP Recalibrator

```
java -jar /path/GenomeAnalysisTK.jar \
 -T VariantRecalibrator \
 -R reference.fasta \
 -input samples.vcf \
 -resource:hapmap,known=false,training=true,truth=true,prior=15.0 /path/hapmap_3.3.hg19.sites.vcf \
 -resource:omini,known=false,training=true,truth=false,prior=12.0 /path/1000G_omni2.5.hg19.sites.vcf \
 -resource:1000G,known=false,training=true,truth=false,prior=10.0 1000G_phase1.snps.high_confidence.hg19.sites.vcf \
 -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 /path/dbsnp_138.hg19.vcf \
 -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an DP -an InbreedingCoeff \
 -tranche 100.0 -tranche 99.5 -tranche 99.0 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
 -mode SNP \
 -recalFile samples.recal \
 -tranchesFile samples.tranches \
 -rscriptFile samples.R
```

Omni，這個資料來源自Illumina的Omni基因型晶片，大概2.5百萬個位點。也可以考慮將其設定為 truth=truth，true=false是較為保守的做法。

```
java -jar /path/GenomeAnalysisTK.jar -T ApplyRecalibration \
 -R human_g1k_v37.fasta \
 -input samples.vcf  \
 --ts_filter_level 99.5 \
 -tranchesFile samples.tranches \
 -recalFile samples.recal \
 -mode SNP \
 -o samples.snps.VQSR.vcf
```

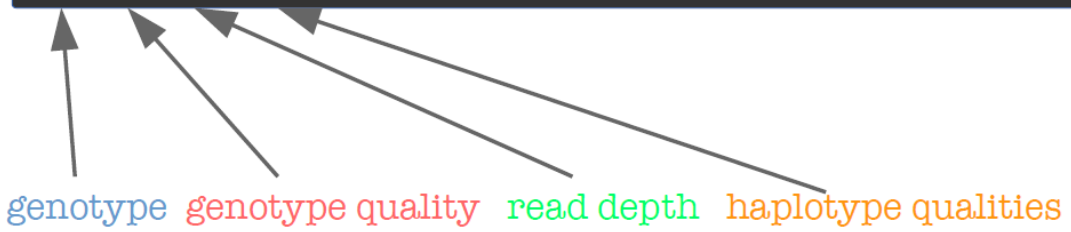
Prior – Phred-scaled estimate of data accuracy
Training –該資料是否作為模型訓練的資料集，用於訓練VQSR模型
Truth -該資料是否作為**驗證模型訓練的真集數據**，這個數據同時還是VQSR訓練bad model時自動進行參數選擇的重要資料.
Known –該資料是否作為已知變異資料集，用於對變異資料的標注

Indel Recalibrator

```
java -jar /path/to/GenomeAnalysisTK.jar \
 -T VariantRecalibrator \
 --maxGaussians 4 \
 -R human_g1k_v37.fasta \
 -input samples.snps.VQSR.vcf \
 -resource:1000G,known=false,training=true,truth=true,prior=10.0 /path/ 1000G_phase1.indels.hg19.sites.vcf \
 -resource:mills,known=false,training=true,truth=true,prior=12.0 /Mills_and_1000G_gold_standard.indels.hg19.sites.vcf \
 -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 /path/dbsnp_138.hg19.vcf \
 -an QD -an DP -an FS -an SOR -an ReadPosRankSum -an MQRankSum -an MQ -an InbreedingCoeff \
 -mode INDEL \
 -recalFile samples.snps.indels.recal \
 -tranchesFile samples.snps.indels.tranches \
 -rscriptFile samples.snps.indels.plots.R

java -jar /path/to/GenomeAnalysisTK.jar \
 -T ApplyRecalibration \
 -R human_g1k_v37.fasta \
 -input samples.snps.VQSR.vcf \
 --ts_filter_level 99.0 \
 -tranchesFile samples.snps.indels.tranches \
 -recalFile samples.snps.indels.recal \
 -mode INDEL \
 -o samples.snps.indels.VQSR.vcf
```

VCF


```
#CHROM  POS     ID        REF    ALT     QUAL    FILTER   INFO
20      14370   rs6054257 G      A       29      PASS     NS=3;DP=14;AF=0.5;DB;H2

FORMAT          NA00001           NA00002           NA00003
GT:GQ:DP:HQ     0|0:48:1:51,51    1|0:48:8:51,51    1/1:43:5:.,.
```



- **CHROM**: chromosome
- **POS**: position
- **ID**: identifier
- **REF**: reference base(s)
- **ALT**: non-reference allele(s)
- **QUAL**: quality score of the calls (phed scale)
- **FILTER**: “PASS” or a filtering tag
- **INFO**: additional information
- **FORMAT**: describes the information given by sample

0/0: homozygous reference.

0/1: heterozygous, one from reference, one from alternate.

1/1: homozygous alternate.

VQSR output VCF (vs. Hard Filter)

- Before VQSR (input vcf):

#CHROM	POS	FILTER	INFO
1	10146	.	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	.	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	.	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

- After VQSR (output vcf):

#CHROM	POS	FILTER	INFO
1	10146	VQSRTrancheINDEL99.30to99.50	AC=1...;NEGATIVE_TRAIN_SITE;VQSLOD=-1.328;culprit=SOR
1	10403	PASS	AC=1;...;QD=0.60; VQSLOD=0.794;culprit=QD
1	234313	VQSRTrancheSNP99.90to100.00	AC=1;...;POSITIVE_TRAIN_SITE;VQSLOD=-5.356;culprit=MQ

- Hard filtering vcf:

#CHROM	POS	FILTER	INFO
1	10146	PASS	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	INDEL_Filter	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	SNP_Filter	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

VQSR problems

1. 樣本數不足 → 模型不穩定。WES 的 variants 數量太少，也不利建立有效的model

- VQSR 需要大量變異資料來建立穩定的高斯混合模型。
- 若樣本太少（例如 <30 個 exome 或 <1 個 WGS），模型可能無法有效分群，導致過濾失準。

2. 註解分布不符合高斯假設

- VQSR 假設所有註解（如 MQ、QD、FS）呈現高斯分布，但某些註解（如 MQ）實際上偏斜或離散，可能使模型失真。
- 解法：可嘗試排除某些註解（例如不使用 MQ）來穩定模型。

3. 訓練資料設定錯誤

- 若使用的 truth/training resource（如 Omni、HapMap）標記錯誤（例如 truth=false 應為 truth=true），會導致模型學習錯誤的變異特徵。
- 這會讓 VQSLOD 分數失準，過濾結果偏離預期。

4. Tranche 分層失效或難以解讀

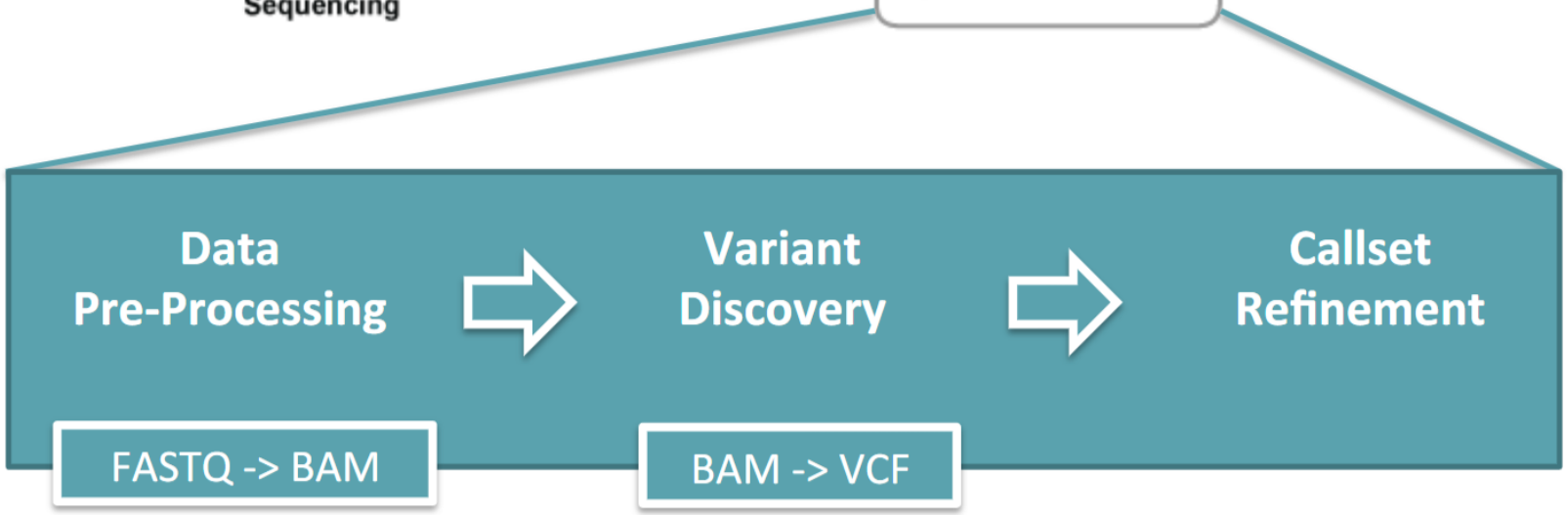
- 有時 tranche plot 中 Ti/Tv 比值不隨 sensitivity 單調下降，代表模型無法清楚區分真實與錯誤變異。
- 這可能是資料品質不佳、模型過度擬合或訓練集不適合。

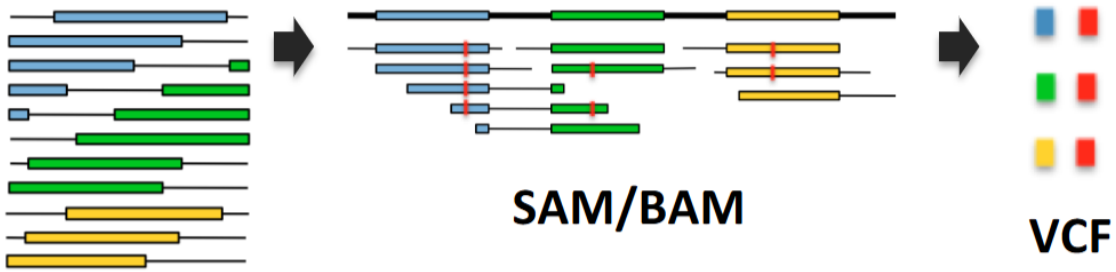
5. Indel 模型更容易失敗

- Indel 的註解分布比 SNP 更複雜，模型更容易不穩。
- 建議 SNP 與 Indel 分開訓練與過濾。

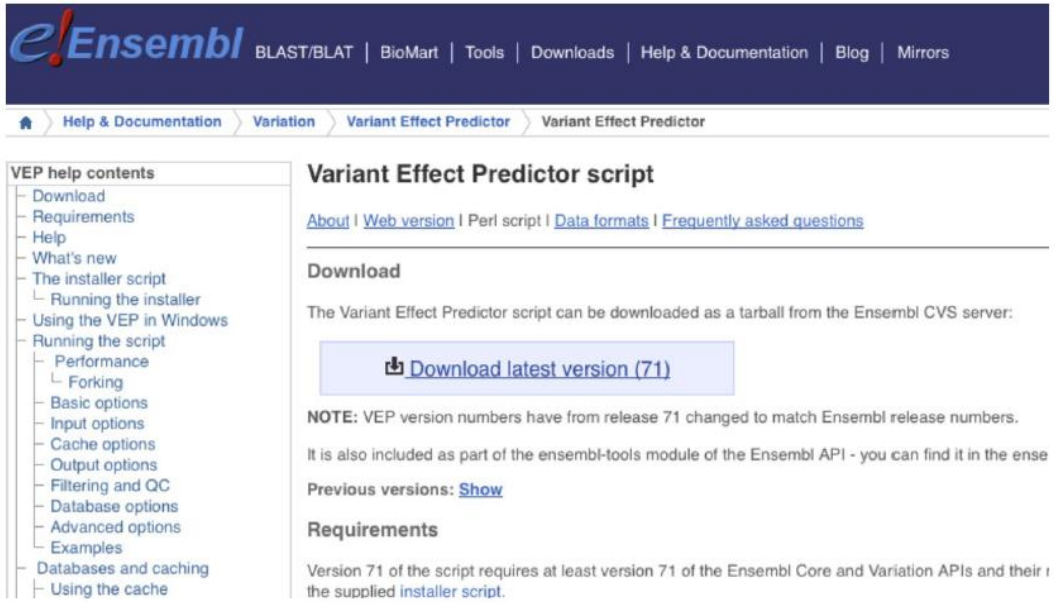

Sequencing
FASTQ
1
2
3
GATK Best Practices
VCF

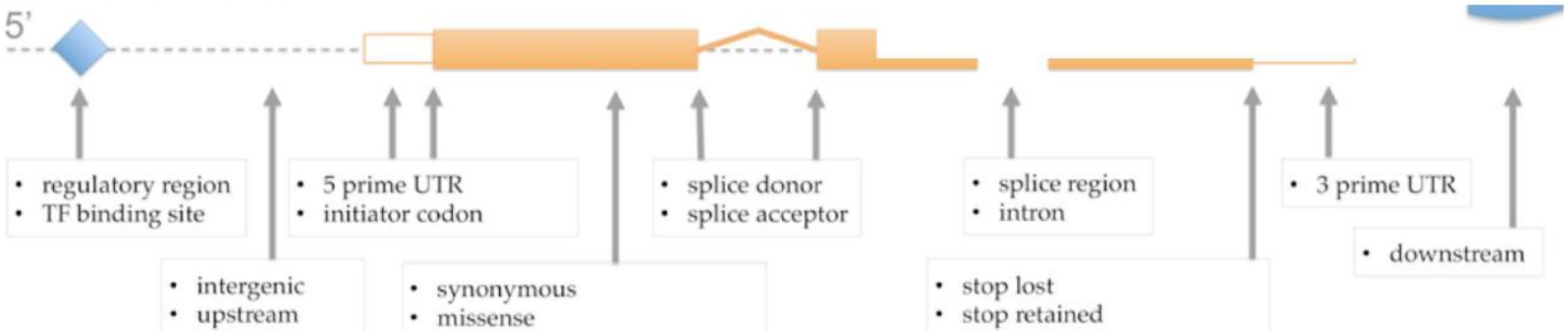
gatk


Data
Pre-Processing
Variant
Discovery
Callset
Refinement
FASTQ -> BAM
BAM -> VCF


FASTQ
SAM/BAM
VCF

Annotation


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Help & Documentation
Variation
Variant Effect Predictor
Variant Effect Predictor
VEP help contents
Download
Requirements
Help
What's new
The installer script
Running the installer
Using the VEP in Windows
Running the script
Performance
Forking
Basic options
Input options
Cache options
Output options
Filtering and QC
Database options
Advanced options
Examples
Databases and caching
Using the cache
Variant Effect Predictor script
About | Web version | Perl script | Data formats | Frequently asked questions
Download
The Variant Effect Predictor script can be downloaded as a tarball from the Ensembl CVS server:
Download latest version (71)
NOTE: VEP version numbers have from release 71 changed to match Ensembl release numbers.
It is also included as part of the ensembl-tools module of the Ensembl API - you can find it in the ense
Previous versions: Show
Requirements
Version 71 of the script requires at least version 71 of the Ensembl Core and Variation APIs and their
the supplied installer script.


5'
regulatory region
TF binding site
5 prime UTR
initiator codon
splice donor
splice acceptor
splice region
intron
3 prime UTR
downstream
intergenic
upstream
synonymous
missense
stop lost
stop retained