次世代定序、生物資訊學與基因體醫學
NGS, bioinformatics and genomic medicine
(Genom7009)

Human Reference Genome(s) & NGS quality
FastQC (HW1)

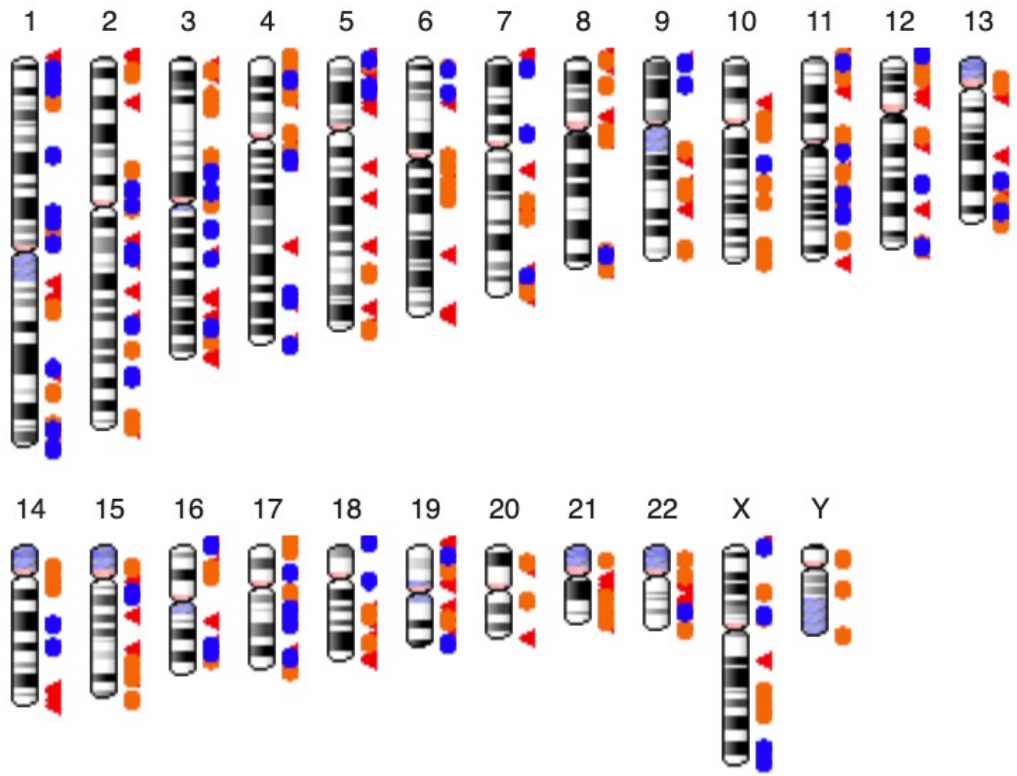
Jacob Shujui Hsu 許書睿

國立臺灣大學醫學院基因體暨蛋白體醫學研究所

Graduate Institute of Medical Genomics and Proteomics

Genome Reference Consortium

- Main Builds (major)
  - GRCh37 (Feb 27, 2009)
  - GRCh38 (May 07,2014)
- Patches (minor)
  - GRCh37.p13 (June 28,2013)
  - GRCh38.p13 (March 01,2019)
  - GRCh38.p14 (**February 03, 2023**)
- Contigs
- GRch39? Telemore-to-Telomere?
- Genome graphs?
- Personal Genome Assembly



Region containing alternate loci
Region containing fix patches
Region containing novel patches

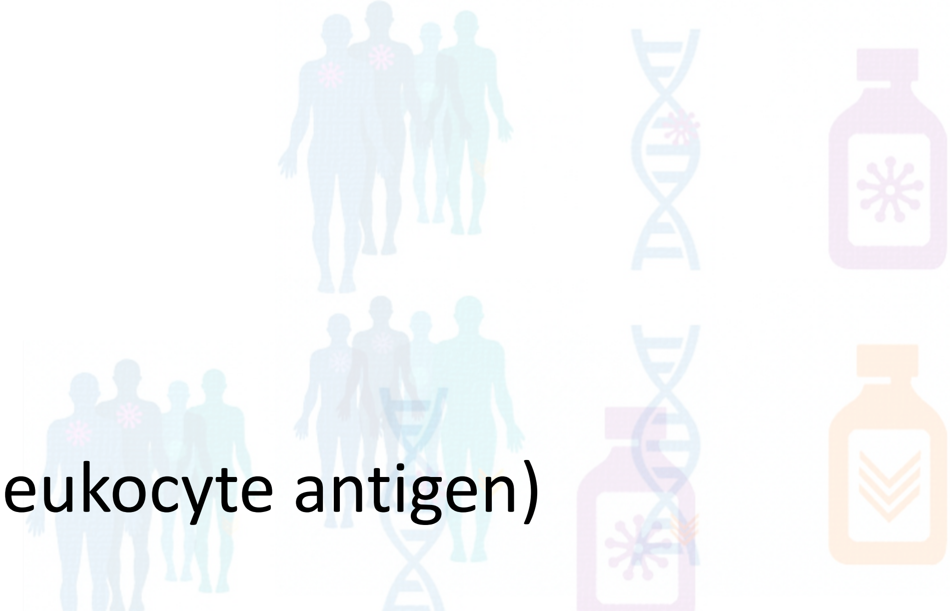
Ideogram of the latest human assembly, GRCh38.p14

https://www.ncbi.nlm.nih.gov/grc/human

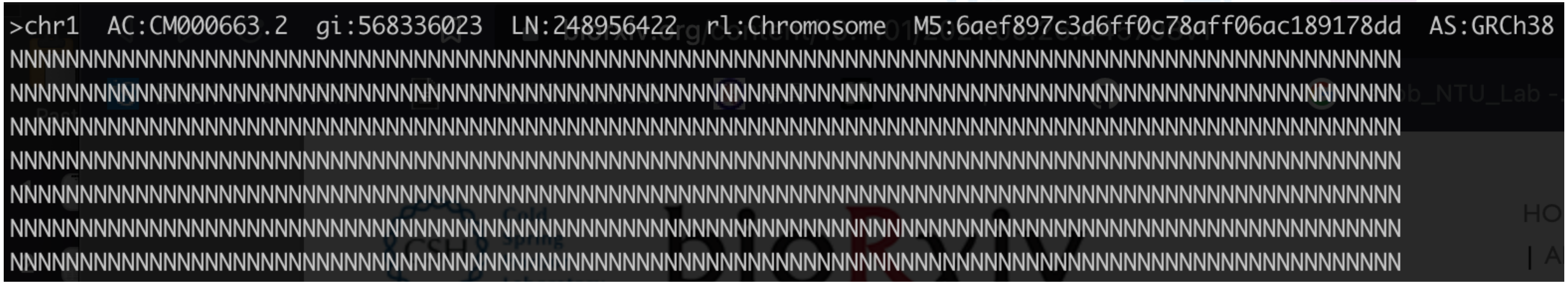
Incomplete human genome (~50% repeats)

- GRCh37 (hs37d5, d: decoy)

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```



- GRCh38 (hs38DH, D: decoy; H: Human leukocyte antigen)



```
>chr1  AC:CM000663.2  gi:568336023  LN:248956422  rl:Chromosome  M5:6aef897c3d6ff0c78aff06ac189178dd  AS:GRCh38
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

https://sites.google.com/ucsc.edu/t2tworkinggroup/publications?authuser=0

Liao et al., Communications Biology (2023)

Human Genome Reference(s)

GATK reference type	B37	HG19	GRCh37(GRC)
(.fasta) file/MD5	human_g1k_v37 human_g1k_v37_decoy	ucsc.hg19	
1b22b98cdeb4a9304cb5d48026a85128	1	chr1	chr1
a0d9851da00400dec1098a9255ac712e	2	chr2	chr2
23dccd106897542ad87d2765d28a19a1	4	chr4	chr4
1fa3474750af0948bdf97d5a0ee52e51	Y	----	----
6743bd63b3ff2b5b8985d8933c53290a	NC_007605	----	----
c68f52674c9fb33aef52dcf399755519	MT	----	chrM
fdfd811849cc2fadebc929bb925902e5	3	----	--
1e86411d73e6f00a10590f976be01623	----	chrY	chrY
641e4338fa8d52a5b781bd2a2c08d3c3	----	chr3	chr3
d2ed829b8a1628d16cbeee88e88e39eb	----	chrM	----

https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies

Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	mitochondria	Epstein-Barr virus	decoy sequences	Remarks
GRCH		GRCh37	No canonical name	No canonical name	No canonical name	Maintained by Mitomap, distributed for convenience	✗	✗	
UCSC	GRCh37	hg19	chrUn_gl000212	chr1_gl000191_random	chr6_apd_hap1	NC_001807 (from build 36)	✗	✗	Chromosome names start by "chr" PAR regions on chrY are hard masked
Ensembl	GRCh37.p13	Ensembl API release 75 Homo_sapiens.GRCh37.75.dna.primary_assembly.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	Chromosome named "1" to "22", "X", "Y" and "MT"
1000 genomes project phase I & III	GRCh37.p2	hs37 g1k_v37 b37 human_g1k_v37.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	"1" to "22", "X", "Y" and "MT"

https://bioinfo-diag.fr/wp-content/uploads/2019/10/BioinfoDiag_2019_EAY.pdf

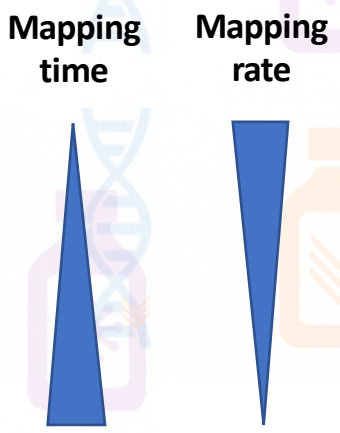
Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	mitochondria	Epstein-Barr virus	decoy sequences	Remarks
1000 genomes project phase II	GRCh37.p4	hs37d5 b37+decoy +herpes hs37d5.fa.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	NC_007605	hs37d5 ss	pseudo-autosomal regions are hard-marked on Y chromosome
Illumina MiSeq Reporter + BSO	hg19	hg19	✗	✗	✗	NC_001807 (from build 36)	✗	✗	hg19 without unplaced/unlocalized contigs nor alternate loci
Ion Torrent	hg19	hg19	✗	✗	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	hg19 without unplaced/unlocalized contigs nor alternate loci
GATK Bundle	GRCh37.p2	b37 + decoy	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	hs37d5 ss	"1" to "22", "X", "Y" and "MT"

https://bioinfo-diag.fr/wp-content/uploads/2019/10/BioinfoDiag_2019_EAY.pdf

Mapping rate & Contigs

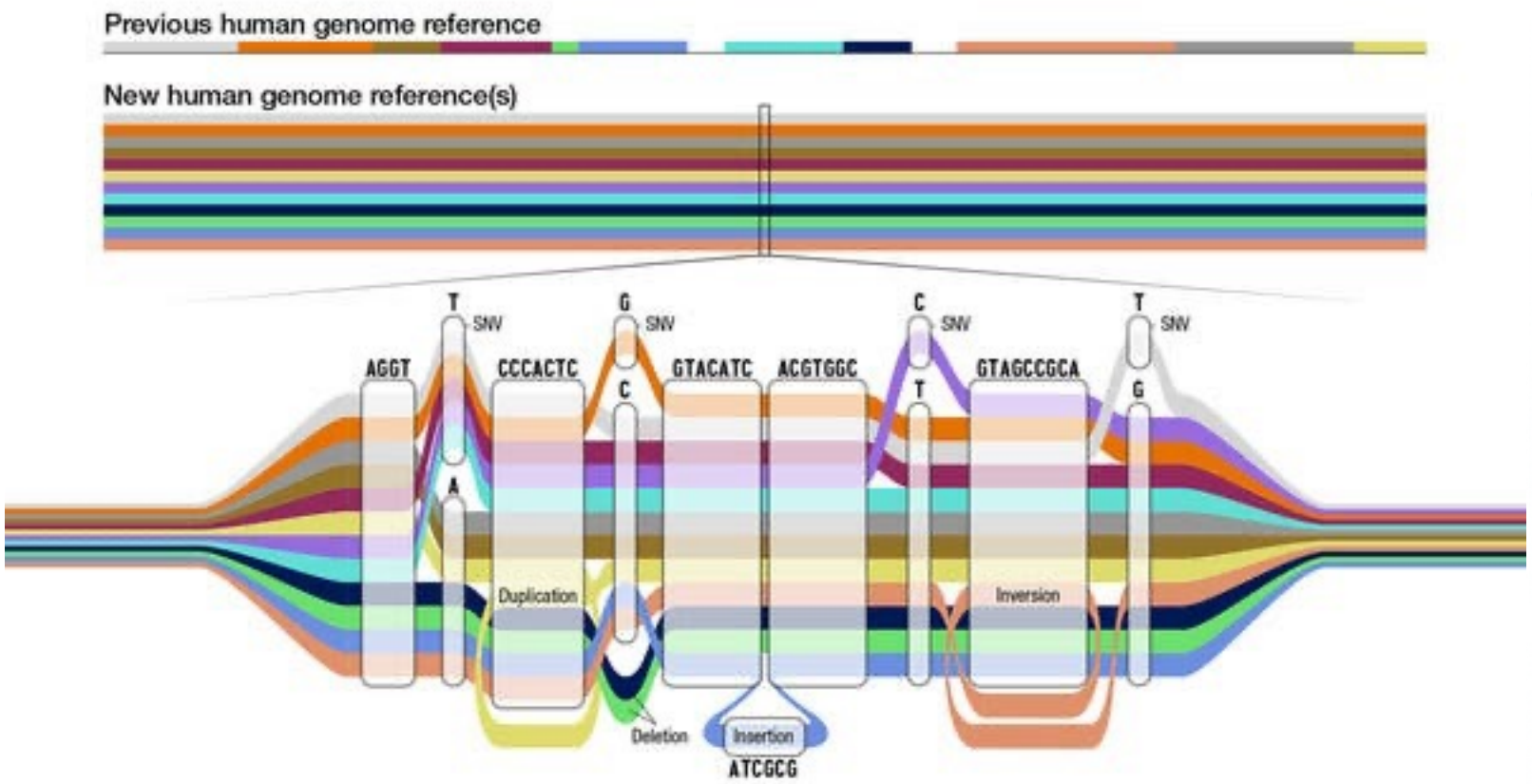
- A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing, Nature Communications (2015)

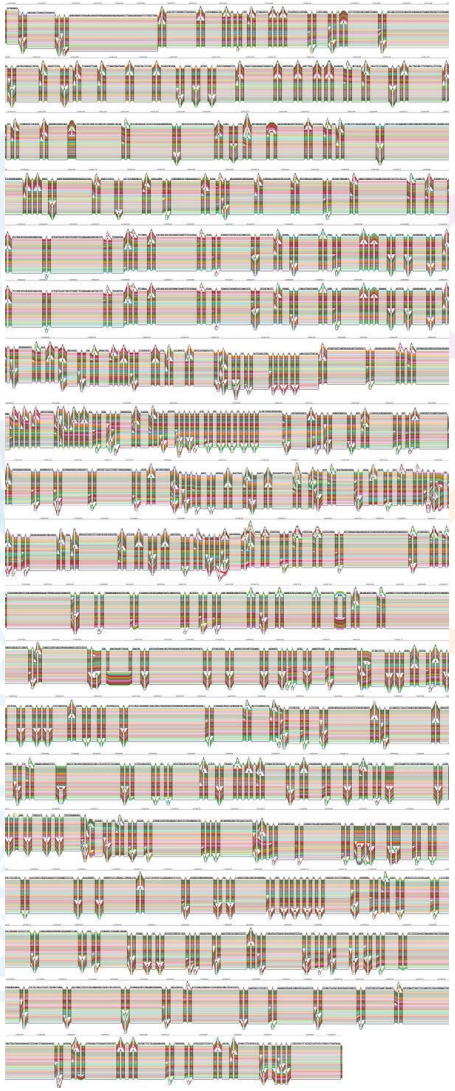
	Reads (control + tumor)			Minutes
	Total	Aligned	Uniquely aligned	CPU time
b37+d	2152793590 (100.0 %)	2112982704 (98.15 %)	2018366589 (93.76 %)	79348.8 (100.0 %)
b37	2152793590 (100.0 %)	2071267988 (96.21 %)	1996504352 (92.74 %)	92517.7 (116.6 %)
hg19r	2152793590 (100.0 %)	2058694644 (95.63 %)	1990780196 (92.47 %)	95894.3 (120.85 %)



Supplementary Table 16. Comparisons of reference genome build versions on alignment rates and alignment times. (All mapping was performed with Novoalign2.) Using a larger reference genome build leads to higher mapping rates and shorter mapping times.

Human Pangenome Reference Consortium (HPRC)





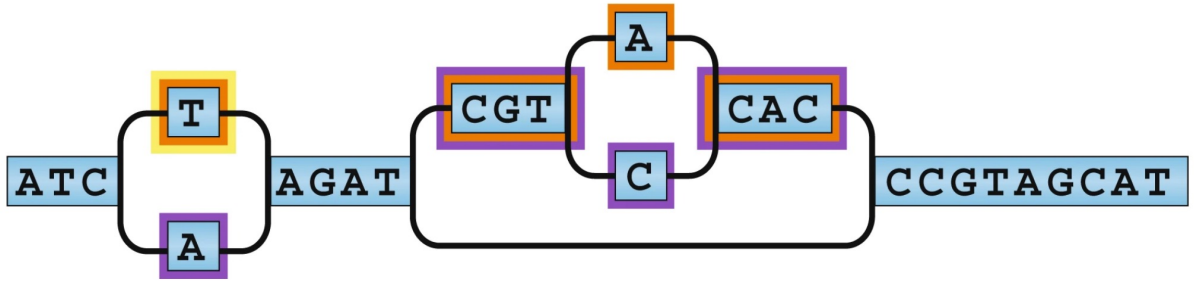
https://www.nytimes.com/2023/05/10/science/pangenome-human-dna-genetics.html

We Need More Than One Human Reference Genome (Video)

https://humanpangenome.org/

Further information

- Graph genome



- Toward a better human genome reference
  - Complete (centromere, telomere, repeats, pseudogenes)
  - Representative (from all populations to each individuals ?)
  - Compatible (genomic coordinates, annotations)

科普文章:

- Human Genome Reference
- Demystifying the versions of GRCh38/hg38 Reference Genomes, how they are used in DRAGEN™ and their impact on accuracy

Pan-genomics in the human genome era Nature Reviews Genetics volume 21, pages243–254 (2020)

For reference genome(s)

- About 2-4% of WGS reads are not in reference genome, and personal SV may affect allele specific expression
- Current reference genome missed 8% of the real genomic regions. The next phase of HPRC is to have 350 sample which are expected to cover 99.94% variant with MAF > 1% in All of US
- HG002 diploid T2T is about to finished

GRCh38.p13
(2019 March)

T2T
(2022 March)

GRCh38.p14
(2023 February)

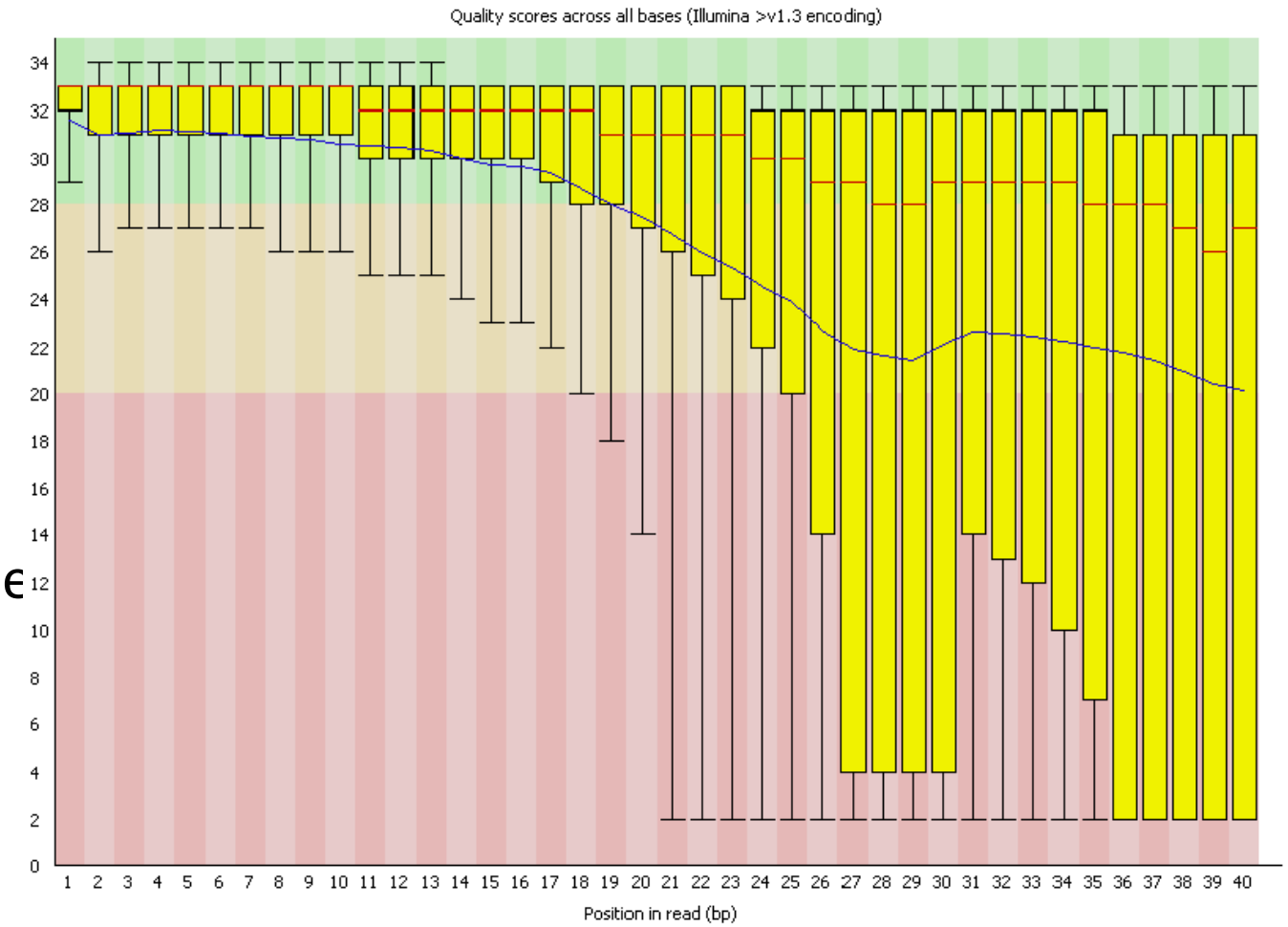
HPRC
(2023 May)

GRCh38+HPRC
(2024 ??)

~~GRCh37.p13~~
~~**(2013 June)**~~

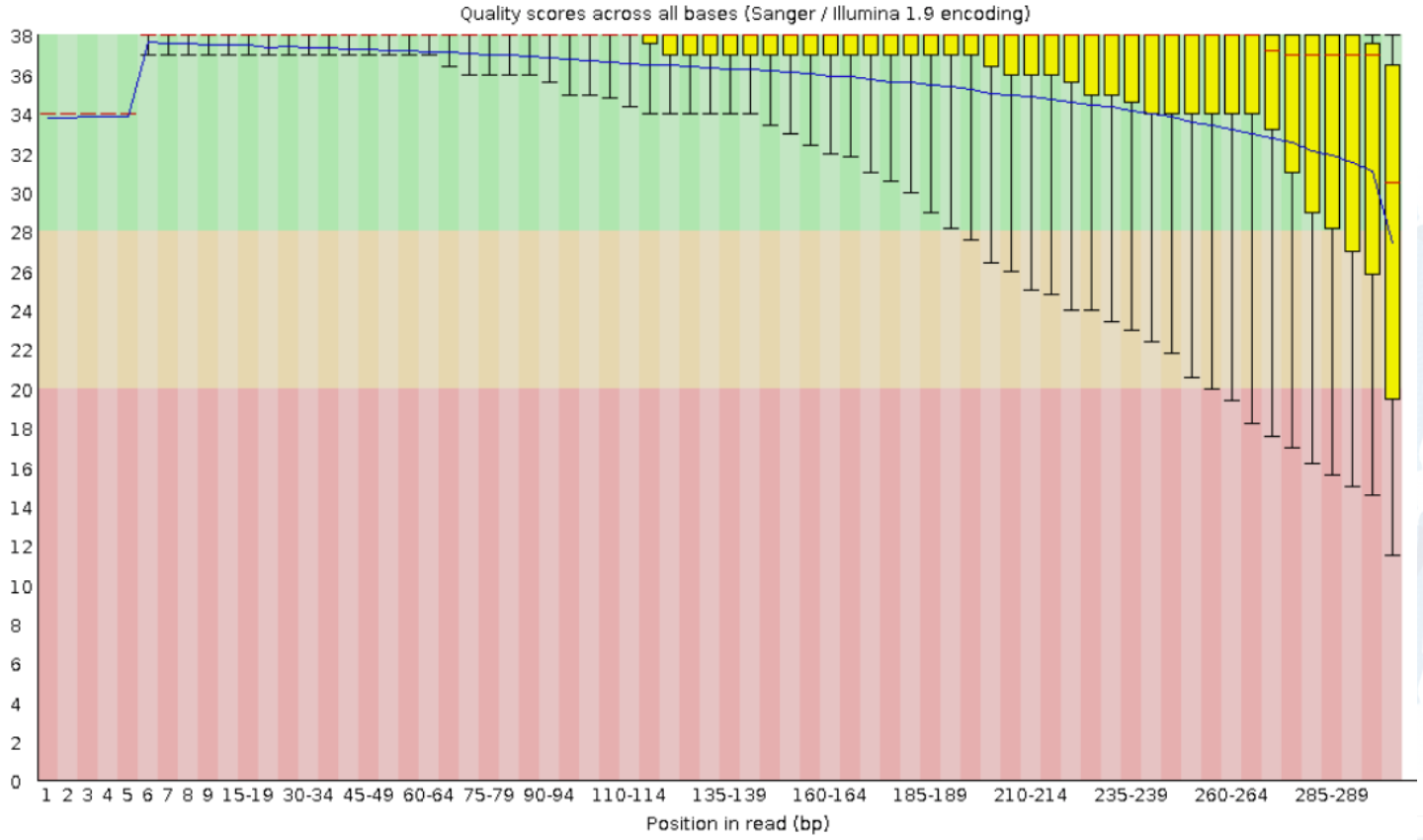
HW1: FastQC (quality score)

- Pair-end WES
- Three dataset
  - A/B/C
- The Median value of quality score
  - x-axis: position in read (bp)
  - y-axis: Phred score
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality



FastQC report

- Pass
- Warning
- Failure

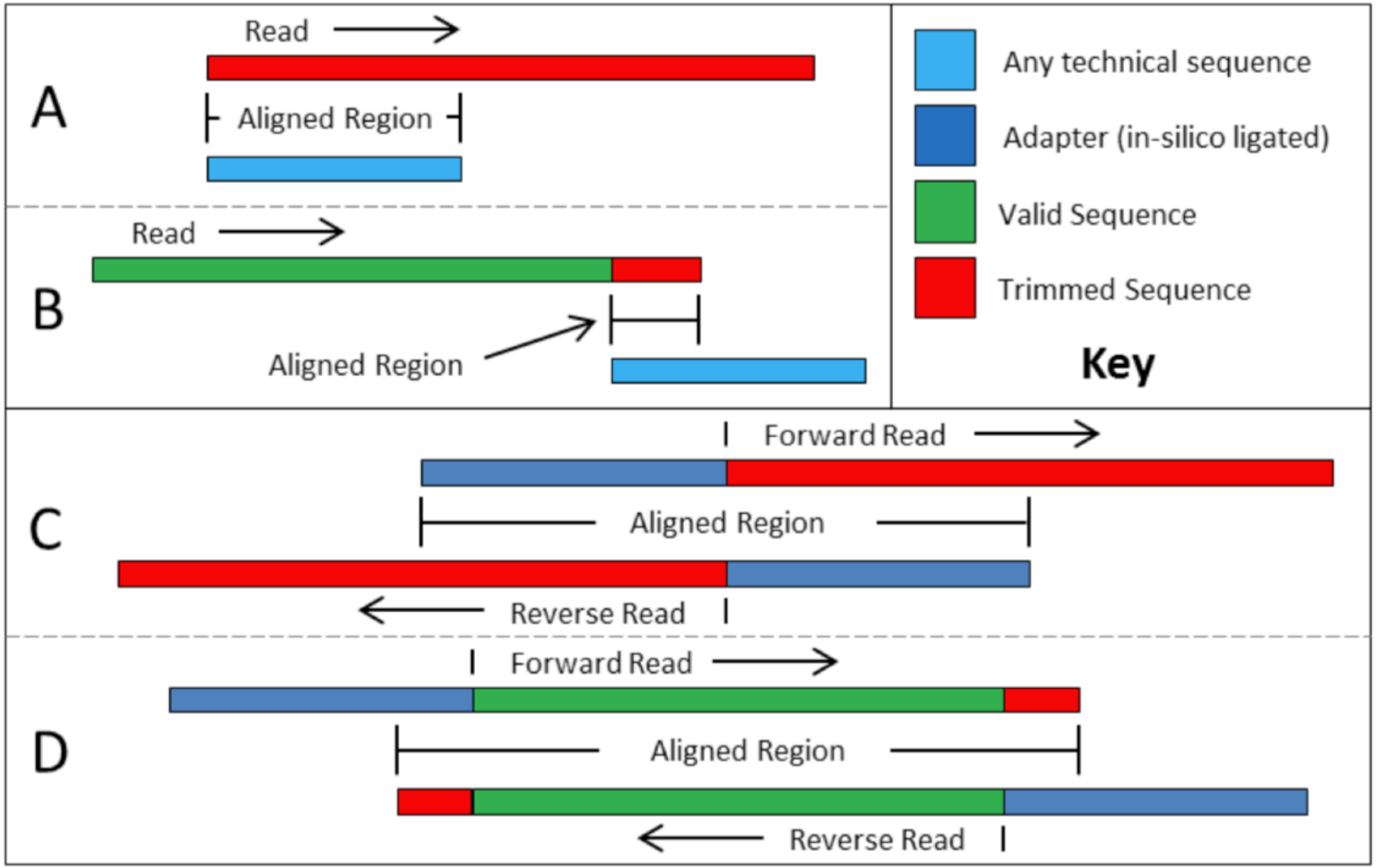


- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

From FastQC
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Trim or not to trim adaptor? quality?

Trimmomatic : trim adaptor or by quality



DNA TECH

DNA Technologies & Expression Analysis Core Laboratory

UCDAVIS GENOME CENTER

Home | Getting Started | Services | Sample Submission & Forms | Prices | FAQ | Shared Equipment | News & Workshops | Accounts & Calendars | Contact

When should I trim my Illumina reads and how should I do it?

Should I trim adapters from my Illumina reads?

This depends on the objective of your experiments.

In case you are sequencing for **counting applications** like differential gene expression (**DGE**) RNA-seq analysis, **ChIP-seq**, **ATAC-seq**, read trimming is generally not required anymore when using modern aligners. For such studies local aligners or pseudo-aligners should be used. Modern "local aligners" like STAR, BWA-MEM, HISAT2, will "soft-clip" non-matching sequences. Pseudo-aligners like Kallisto or Salmon will also not have any problem with reads containing adapter sequences.

However, if the data are used for **variant analyses, genome annotation or genome or transcriptome assembly** purposes, we recommend read trimming, including both, adapter and quality trimming.

search here … Go

Recent Posts

- Holiday Schedule
- Core operations resume in Phase 2 of the COVID response
- DNA Technologies Core has to ramp down lab work
- Adjusting DNA Tech Core operation to the COVID-19 guidelines
- Join us for the PacBio Day Symposium — February 26th

illumina

Search Illumina.com

Products Learn Company Support Recommended Links

What sequences do I use for adapter trimming?

03/31/21

When performing sequencing on an Illumina instrument, sequences corresponding to the library adapters can be present in the FASTQ files at the 3' end of the reads if the read length is longer than the insert size. To remove these sequences and prevent issues with downstream alignment, adapter trimming is an option in Illumina FASTQ generation pipelines. Sample sheets generated with Illumina Experiment Manager contain the necessary sequences in the **Settings** section for Illumina kits. Illumina kits in BaseSpace™ Sequence Hub Prep, BaseSpace Sequence Hub Instrument Run Setup, and Local Run Manager have adapter information built into the software. However, some third-party tools require the adapter sequence for trimming be specified separately. The recommended sequences to use for each Illumina kit are as follows.

https://sapac.support.illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html?langsel=/au/

Trim or Not to trim, this is the question

Table 2. Correlation of trimmed and untrimmed RNA-seq data with the TaqMan RT-PCR data

	100 bp PE		50 bp SE	
Method	**UHRR**	**HBRR**	**UHRR**	**HBRR**
No trimming + Subread	0.851	0.870	0.848	0.870
Trimmomatic–adapters and SW + Subread	0.850	0.870	0.848	0.869
Trimmomatic–adapters and MI + Subread	0.850	0.871	0.849	0.869
TrimGalore + Subread	0.850	0.870	0.849	0.869

Shown are the coefficients of Pearson correlation between log2 expression values of 949 genes measured by the TaqMan RT-PCR technique and their RNA-seq expression levels generated from using each method (log2-RPKM). '100 bp PE' in the table denotes the 100 bp paired-end SEQC dataset. First reads (R1 reads) in this dataset were extracted and truncated to 50 bp long to generate the 50bp single-end dataset used here ('50 bp SE').

SEQC project:
Universal Human Reference RNA (**UHRR**) and Human Brain Reference RNA (**HBRR**)

SW mode: a sliding window approach is used to remove read bases that have a low sequencing quality.

MI mode : a maximum information quality filtering approach is applied for removing low quality bases.

Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level
NAR Genomics and Bioinformatics, Volume 2, Issue 3, September 2020