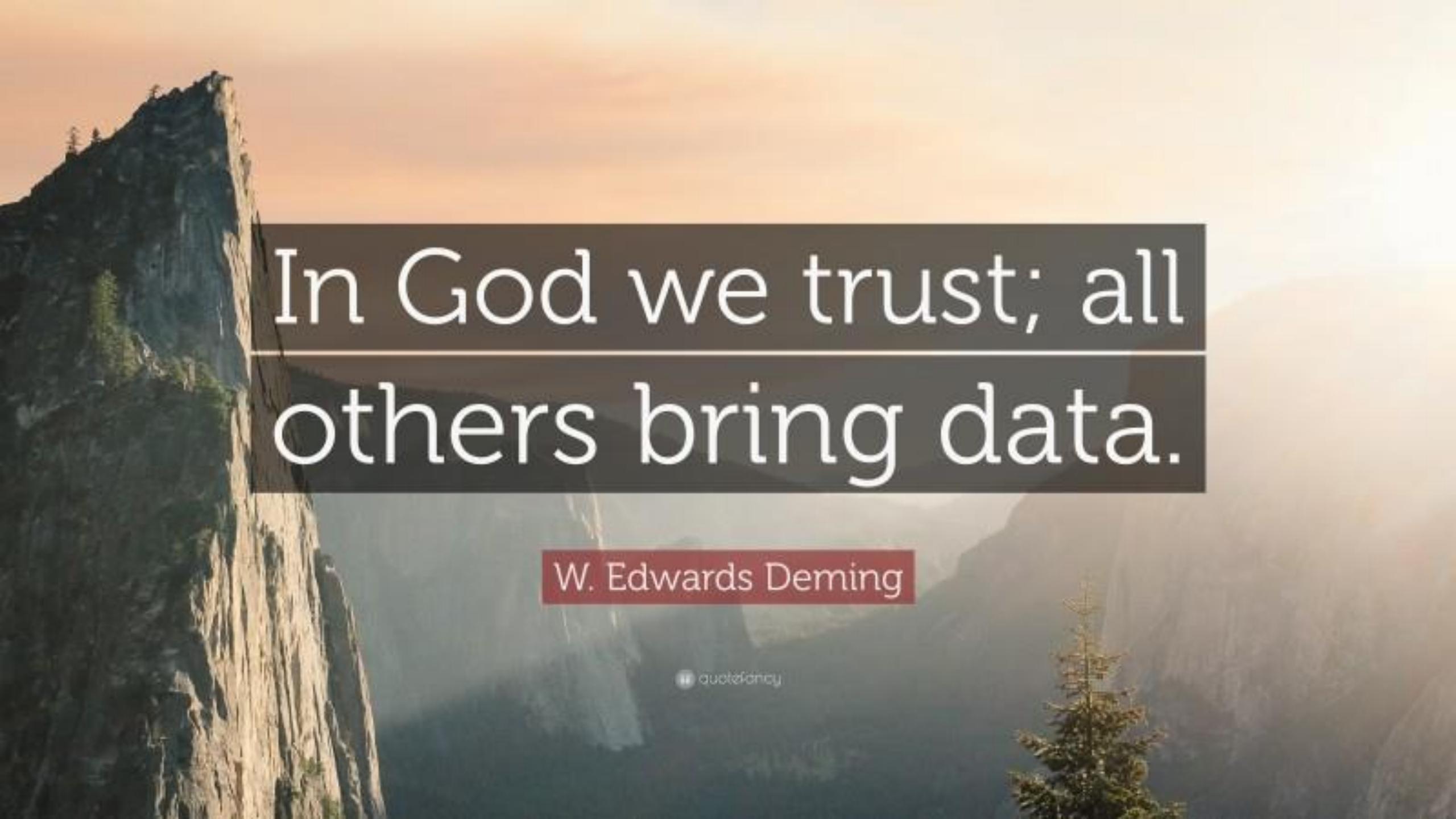
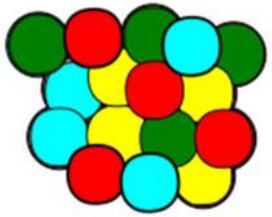

In God we trust; all others bring data.
W. Edwards Deming
quotefancy

BULK RNASEQ WORKSHOP

What we want ?

heterogenous tissue

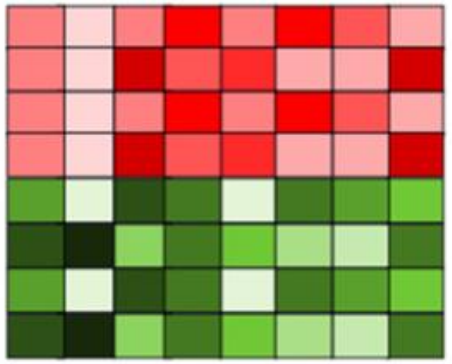


total RNA extraction

↓

...CATCCTAGCTA...

sequencing

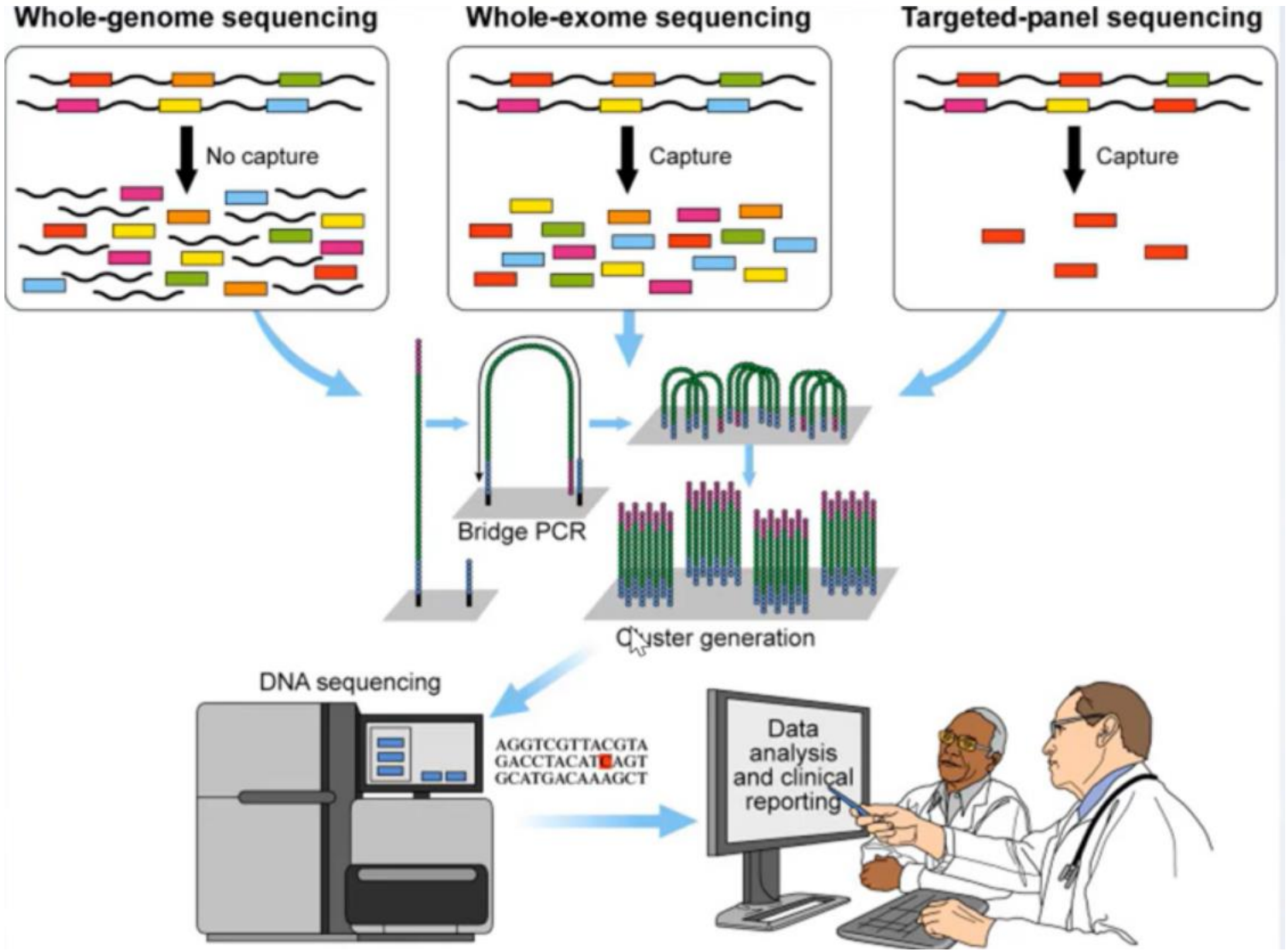


average expression data

What is Bulk RNA sequencing (Bulk RNA-seq)?

- Bulk RNA sequencing is the method of choice for transcriptomic analysis of tissue sections, or biopsies.
- It measures the average expression level of individual genes across hundreds to millions of input cells and is useful to get a global idea of gene expression differences between samples.

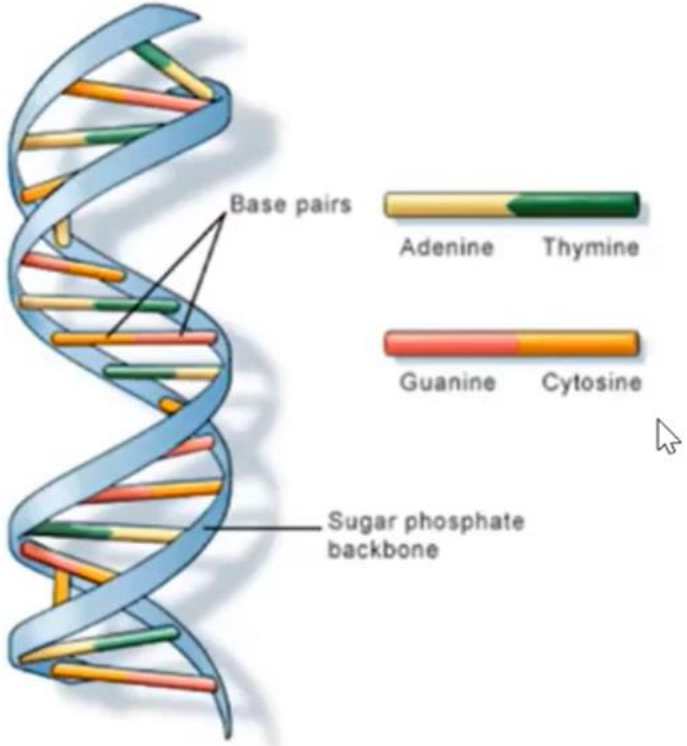
Bulk RNA-seq Pipeline

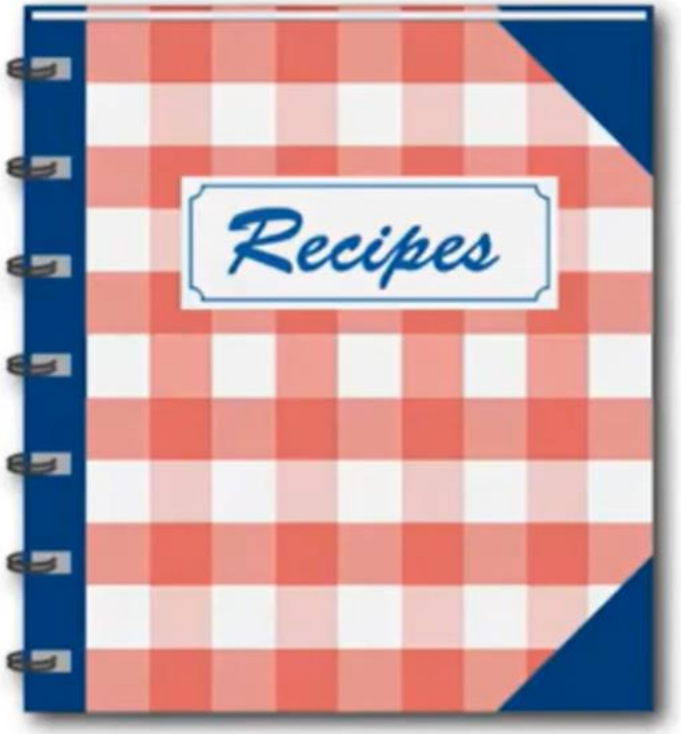


DNA

DNA

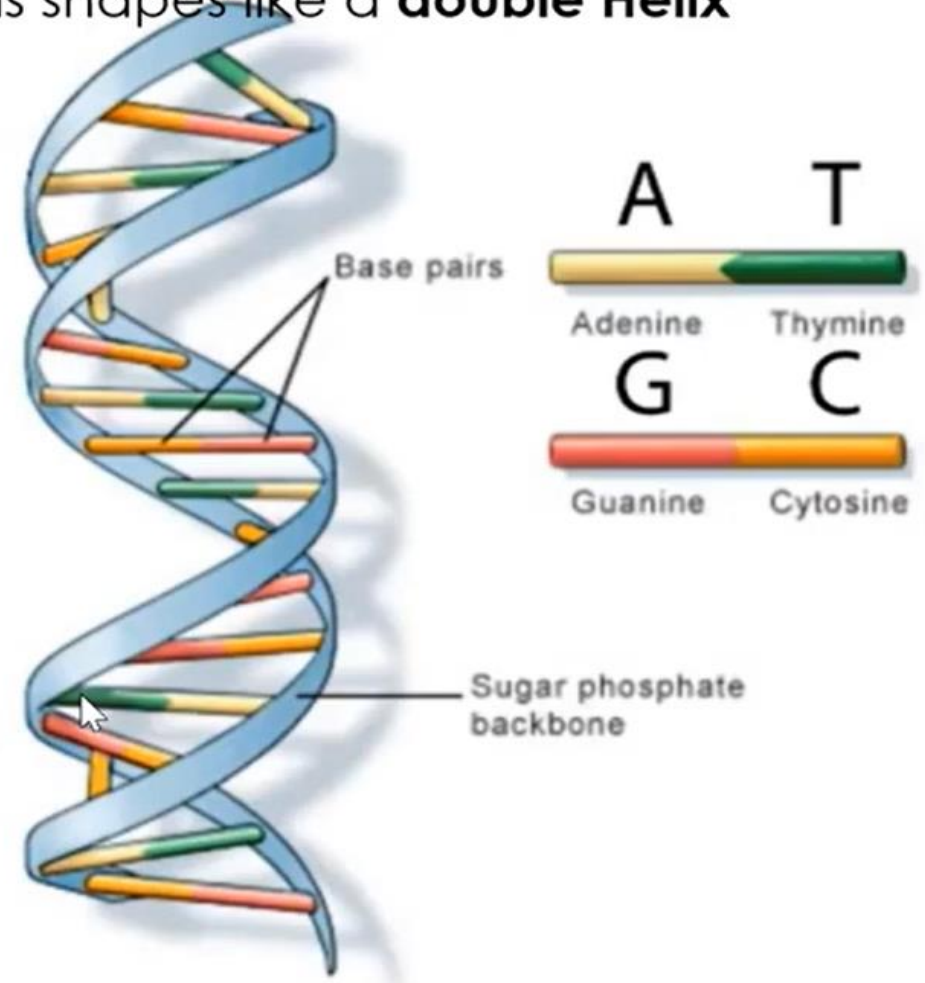
DNA is the kind of **Molecule** that **encodes your Genome** (all your **Genetic Information**). Separate **recipes** for each type of molecule (each little physical piece that make up brain cells, skin cells, heart cells building and maintaining you)

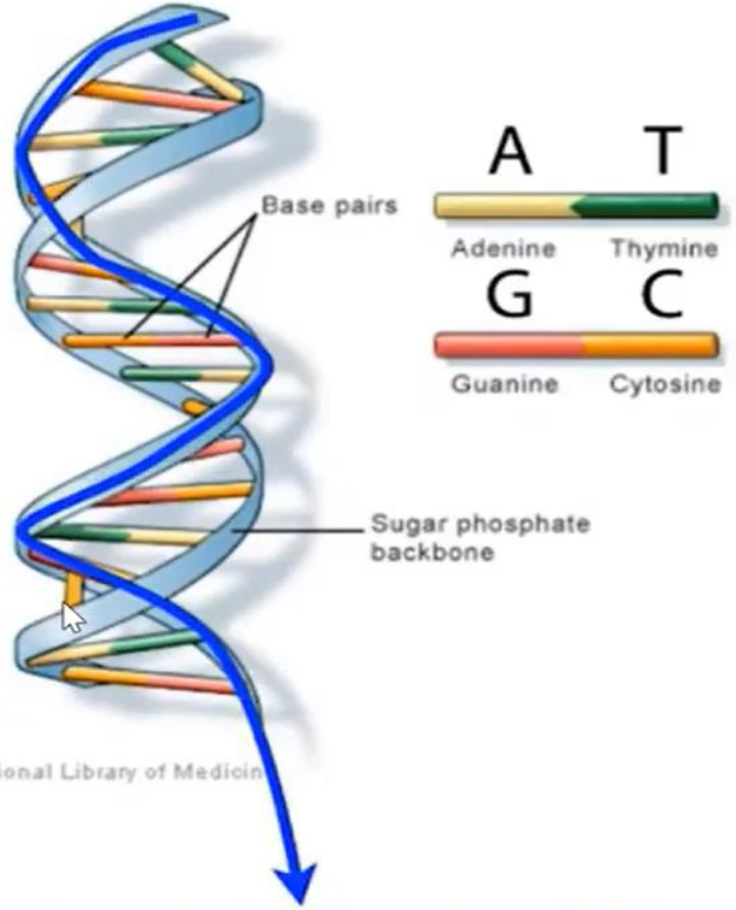




DNA

This recipe is not written in English.
DNA Molecule is shapes like a **double Helix**





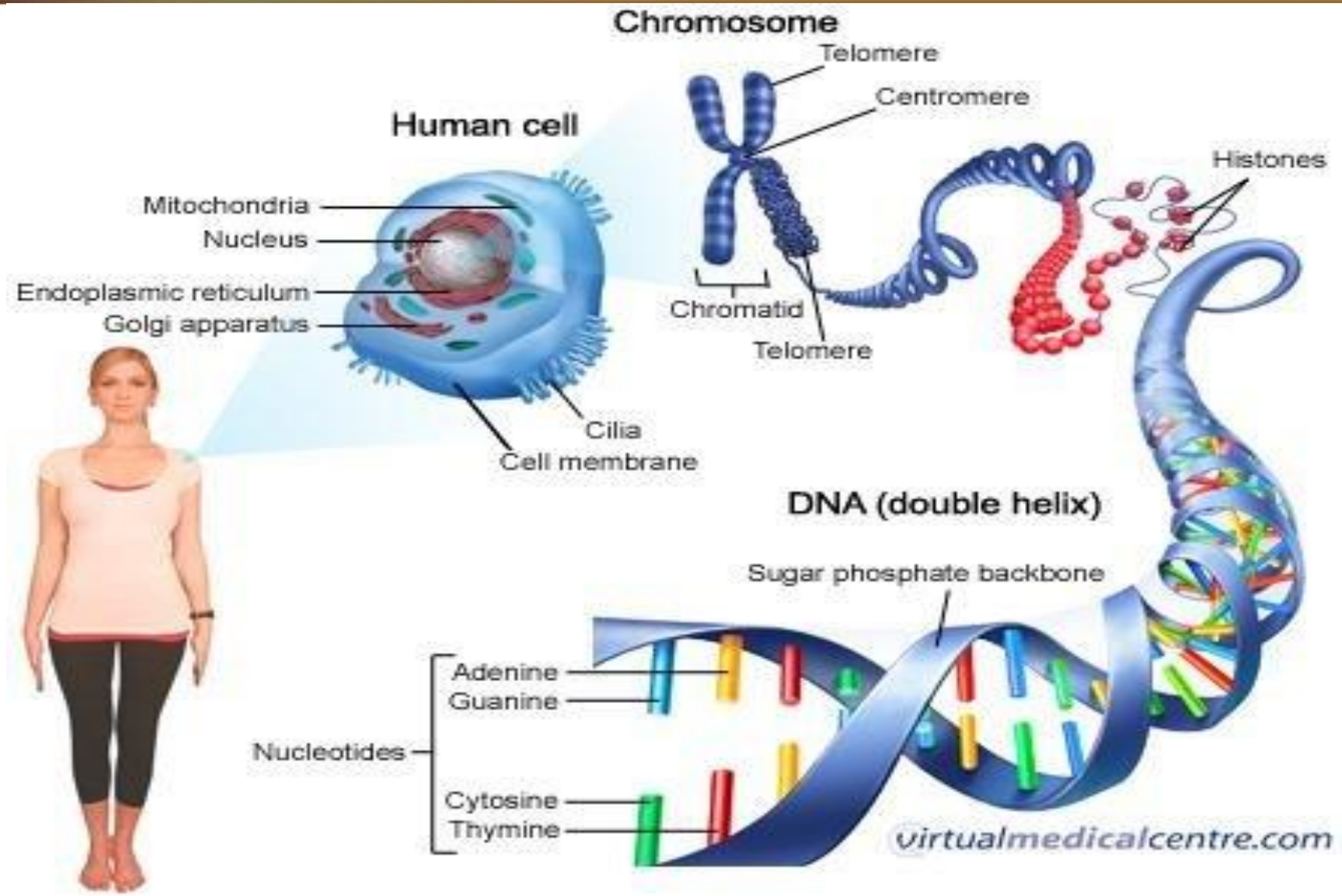
TCACACTGAGCGTGCTG

Miracle in compression

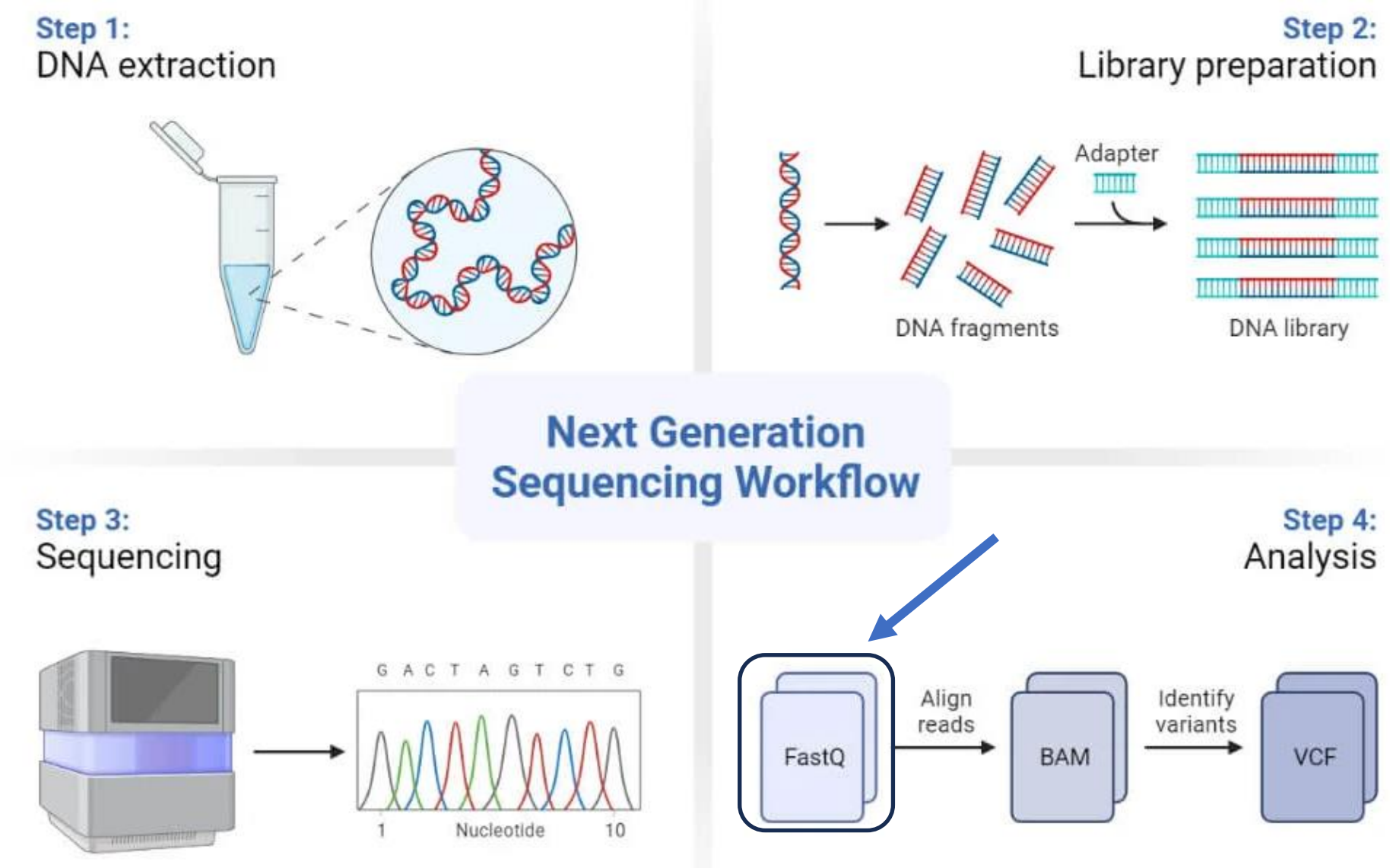
48 volumes of

1000 pages

Every 1mm one letterA4



Pipeline


Step 1:
DNA extraction
Step 2:
Library preparation
Adapter
DNA fragments
DNA library
Next Generation Sequencing Workflow
Step 3:
Sequencing
G A C T A G T C T G
1
Nucleotide
10
Step 4:
Analysis
FastQ
Align reads
BAM
Identify variants
VCF

SAM File

chr11:5246500-5248500 (reverse strand):

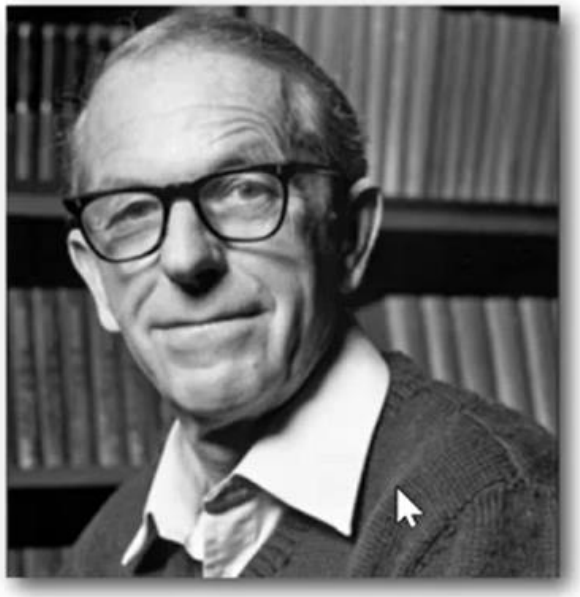
ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT

chr11:5246500-5248500 (reverse strand):

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCT TTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGAT TACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTCTTTTATTTTTAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
TTATGGGTTAAAGTTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
AATGCTTTCTTCTTTTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA
TTATGGGTTAAAGTCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC
AATGCTTTCTTCTTTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTT

Homo sapiens hemoglobin, beta (HBB)

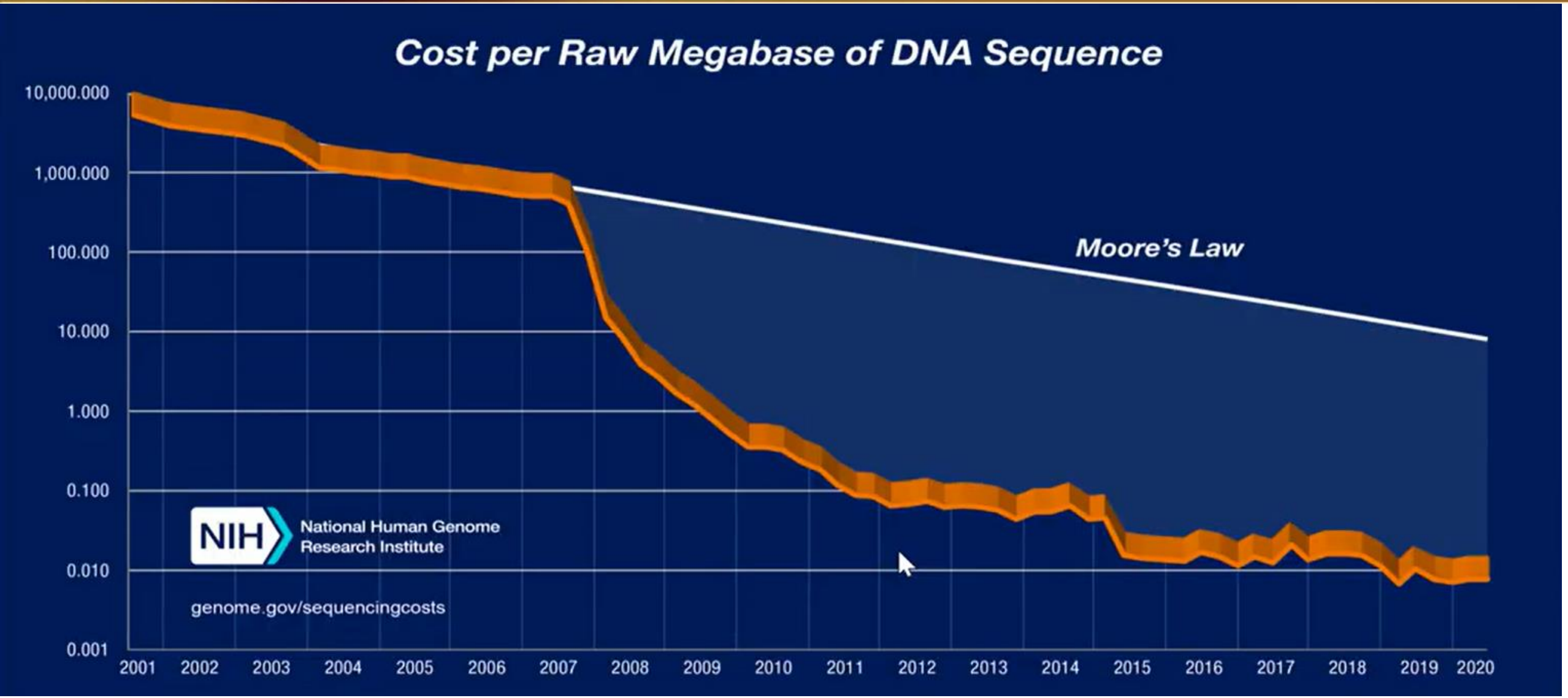
First generation DNA sequencing

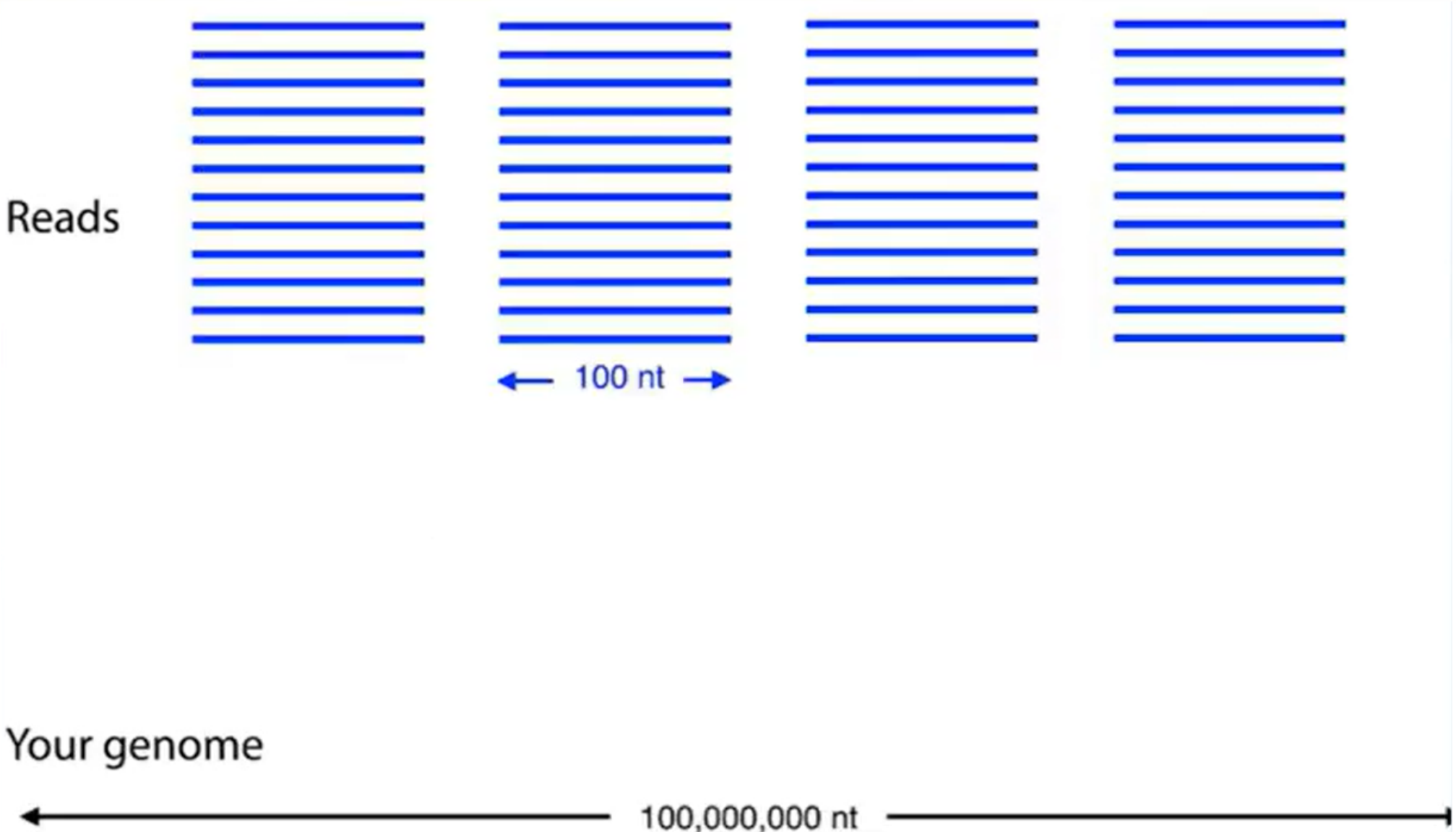


Fred Sanger

"Chain termination" sequencing

First generation DNA sequencing


Cost per Raw Megabase of DNA Sequence
10,000.000
1,000.000
100.000
10.000
1.000
0.100
0.010
0.001
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts


Reads
100 nt
Your genome
100,000,000 nt

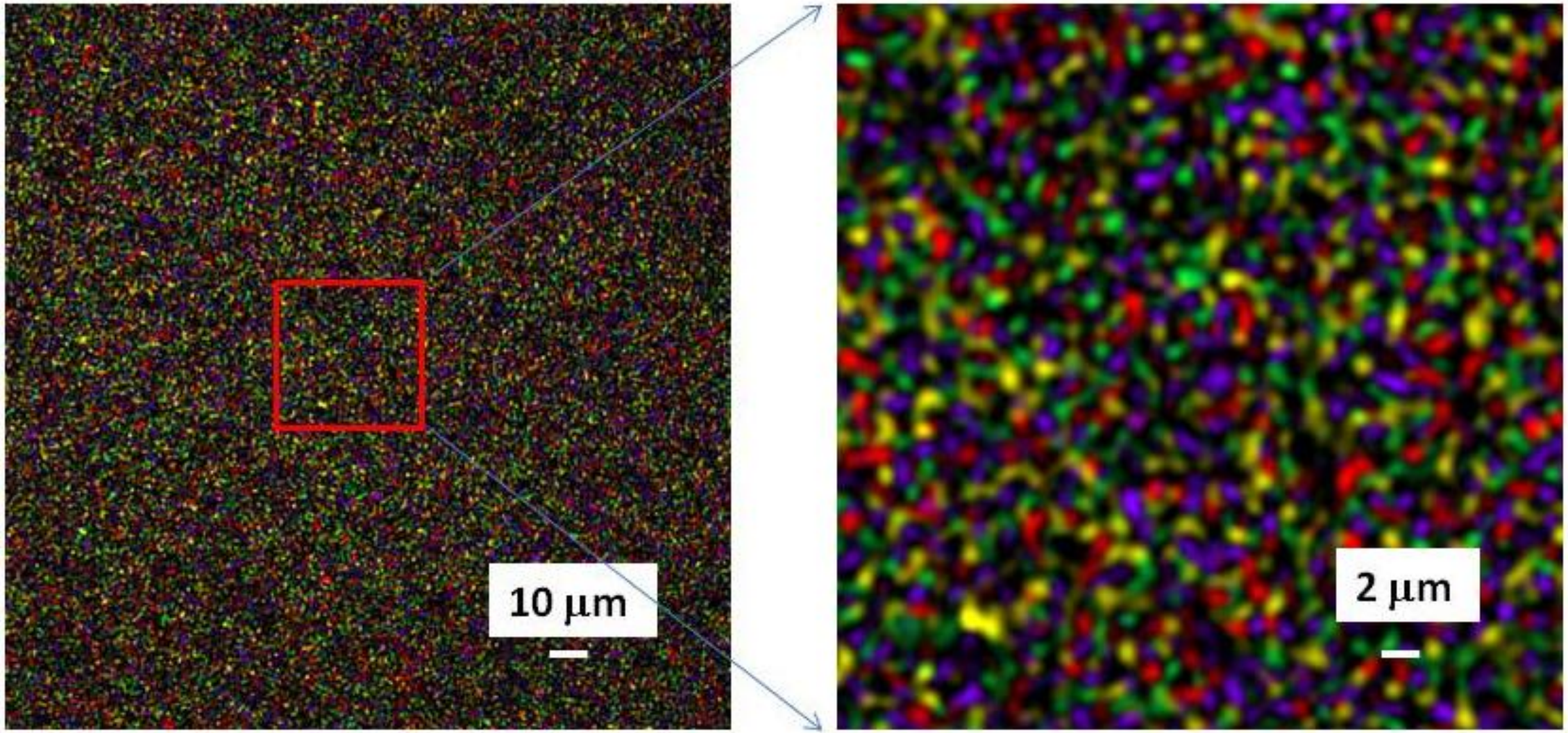
Input DNA

CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

Cut into snippets

CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

Deposit on slide


10 µm
2 µm

$$Q = -10 \cdot \log_{10} p$$

Base quality

Probability that base call is incorrect

$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

$Q = 30 \rightarrow$ 1 in 1,000

The score values

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%

Call: orange (C)

Estimate p, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

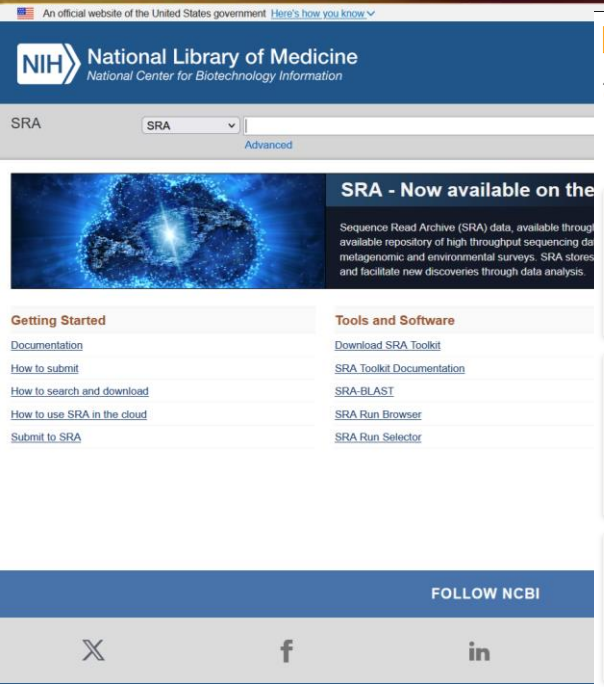
$Q = -10 \log_{10} 1/3 = 4.77$

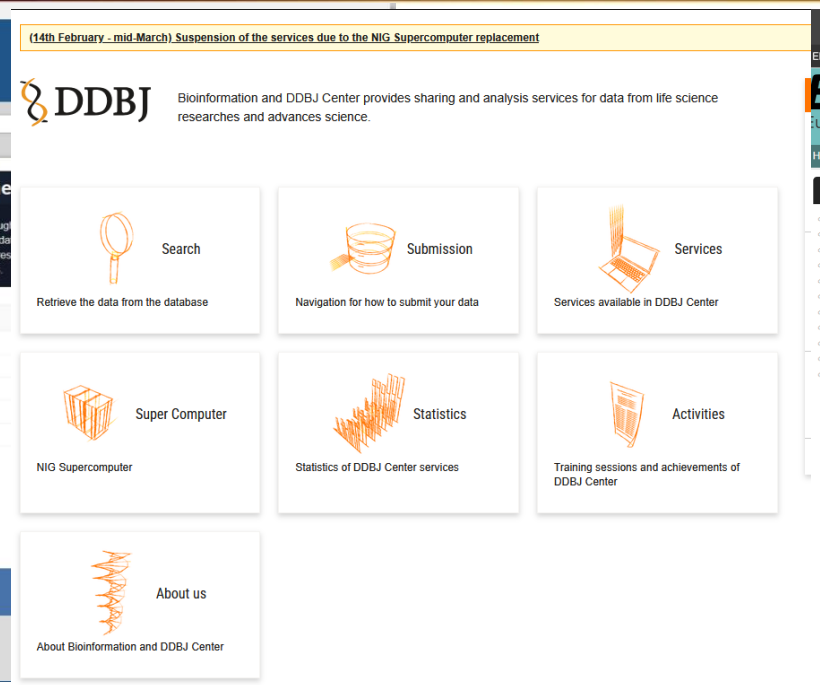
A read in FASTQ format

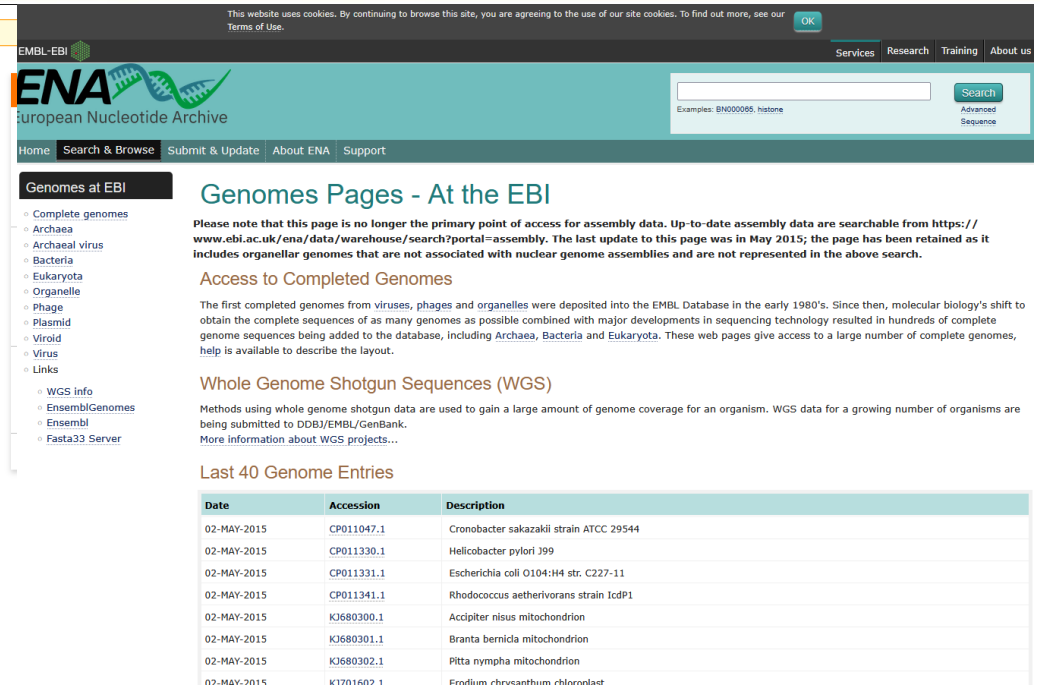
Name	@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1
Sequence	ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
(ignore)	+
Base qualities	?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G

Space	32	0	48	@	64	P	80
!	33	1	49	A	65	Q	81
"	34	2	50	B	66	R	82
#	35	3	51	C	67	S	83
$	36	4	52	D	68	T	84
%	37	5	53	E	69	U	85
&	38	6	54	F	70	V	86
'	39	7	55	G	71	W	87
(	40	8	56	H	72	X	88
)	41	9	57	I	73	Y	89
*	42	:	58	J	74	Z	90
+	43	;	59	K	75	[	91
,	44	>	60	L	76	\	92
-	45	=	61	M	77	]	93
.	46	<	62	N	78	^	94
/	47	?	63	O	79	_	95

Data gathering







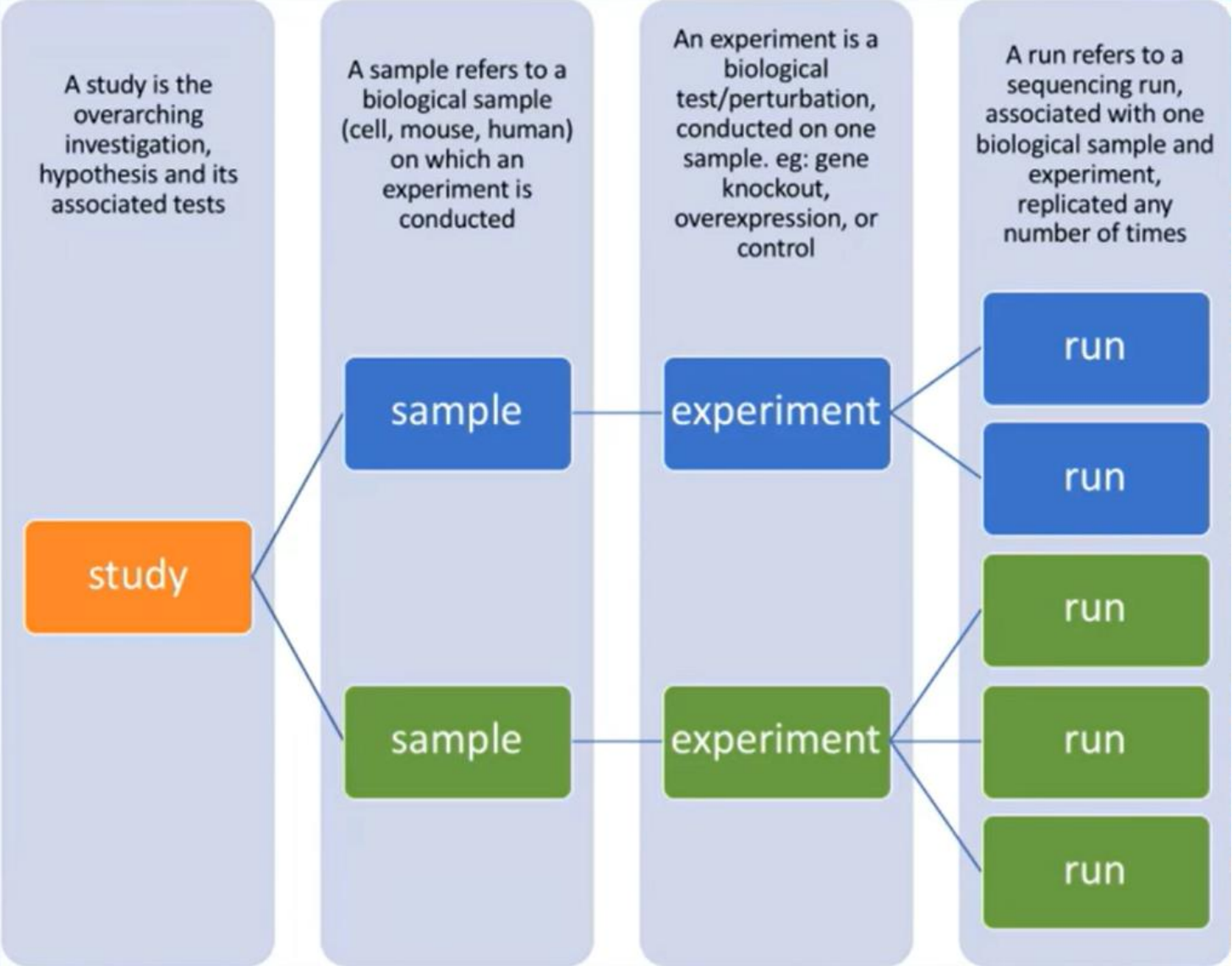
SRP : Study	DRP	ERP
SRX : Experiment	DRX	ERX
SRS : Sample	DRS	ERS
SRR: Run	DRR	ERR

SRP (Project/Study)

SRS (Sample)

SRX (Experiment)

SRR (Run)



SRR : SRR6468671.SRA (Too Dense) 326 M

SRAtoolkit : Fastq-dump

```
fasterq-dump --split-3 SRR11192680.SRA
```

FastQ : SRR6468671_1.fastq 738 M
SRR6468671_2.fastq 738 M
SRR6468671.fastq 1.5 K (Not pared)

Paired End

```
1.6G 10:16 4    جوالى SRR6468540.sra
738M 18:00 29   جوالى SRR6468671_1.fastq
734M 18:00 29   جوالى SRR6468671_2.fastq
1.5K 18:00 29   جوالى SRR6468671.fastq
326M 15:01 26     مف SRR6468671.sra
```

Fastq File head !

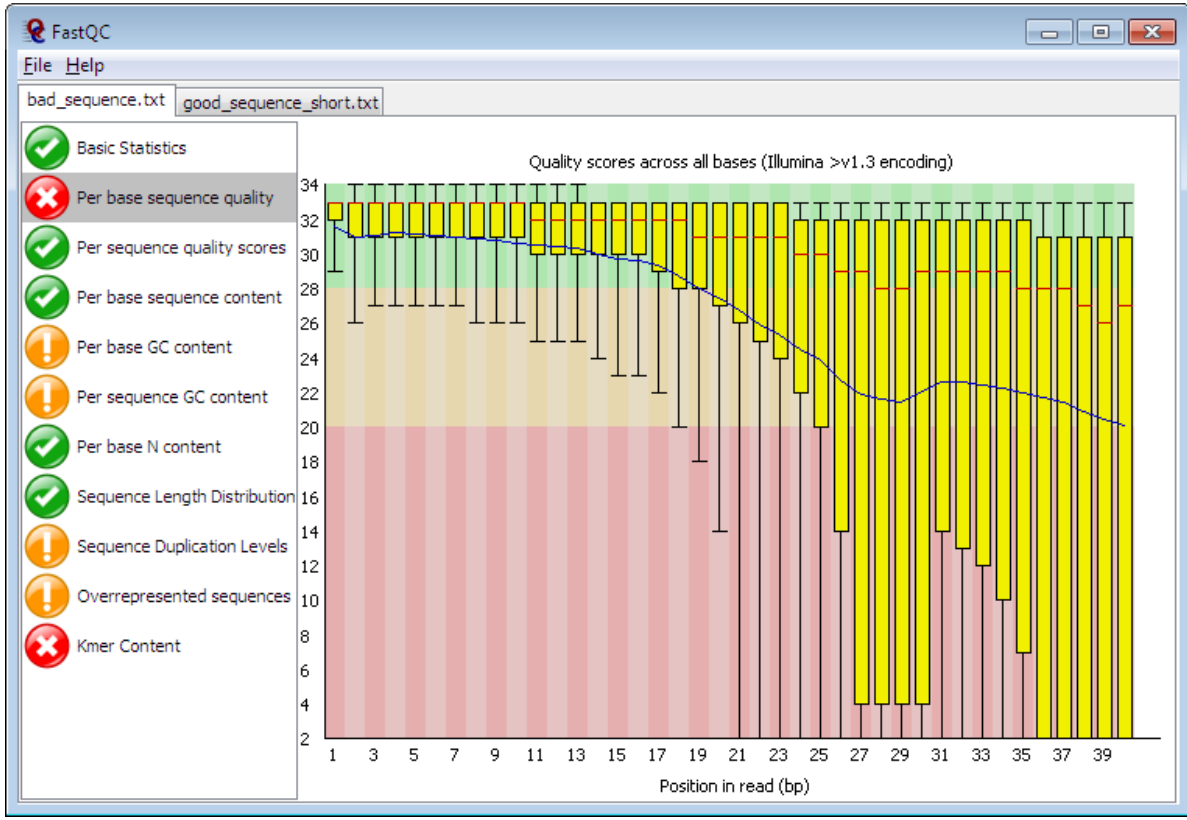
```
SRR6468540.sra  SRR6468671_1.fastq  SRR6468671_2.fastq  SRR6468671.fastq  SRR6468671.sra
(base)                my-pc                :~/SRR$ head SRR6468671_1.fastq
@SRR6468671.1 1 length=101
TAGATCAATTTCATTTATTTTGGACTATGTTGTAATTTTTATTCTTTGCAATGTTTTGAGATTCACTTTAGGAGATCGGAAGAGCACACGTCTGAACTCCA
+SRR6468671.1 1 length=101
CCCDFFDFHHDGDHIIBGGGIGG?FEGIIFH<FHEHGGIIGGGEGHI@GGIIGHGGBDHG8DGG<FHEGIIIHGEEHIGEHFFFFCC@A???CD>AACDCC
@SRR6468671.2 2 length=101
GTACCGCCAATAAGCGTTTGAGAGATGGAGTGACACAGTAGGATAAGCTAACCGTACTGTTGGTTATGTACGGAGATCGGAAGAGCACACGTCTGAACTCC
+SRR6468671.2 2 length=101
@?@FFFDDFHHDHJIJFFGGHEGHHBGHGH0BDHIEIJBFEDEHIIJ>FGGGIH;=CAAEEHF?;@CC>>CDDDD@>CB=8@BABCCBDDDDDDB>ACCCD
@SRR6468671.3 3 length=101
CCGTCCCTTGGGTGCCGCCTTTTTTGTTTTTCATCAGATAAACAGGGTGGTACCGCGATGAGTCCCGTCGTCCCTTGCAATTAGATCGGAAGAGCACACGT
(base)                my-pc                :~/SRR$ wc SRR6468671_1.fastq
 11191060  22382120 773622054 SRR6468671_1.fastq
(base)                my-pc                :~/SRR$ wc -l SRR6468671_1.fastq
11191060 SRR6468671_1.fastq
```

~ 12000000 line
~ 3000000 Read

FASTQC

Aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/




```
$ fastqc -o ~/results/ *.fq
```

Command

Output

Path

Filename

- You may have FastQC or Need to Download .
  - Then :

```
$ sudo apt-get  update        ( Prepare to download )
$ sudo apt-get  install fastqc
$ fastqc –f  fastq SRR6468671_1.fastq   SRR6468671_2.fastq
```

LET'S CODE

FastQC Report

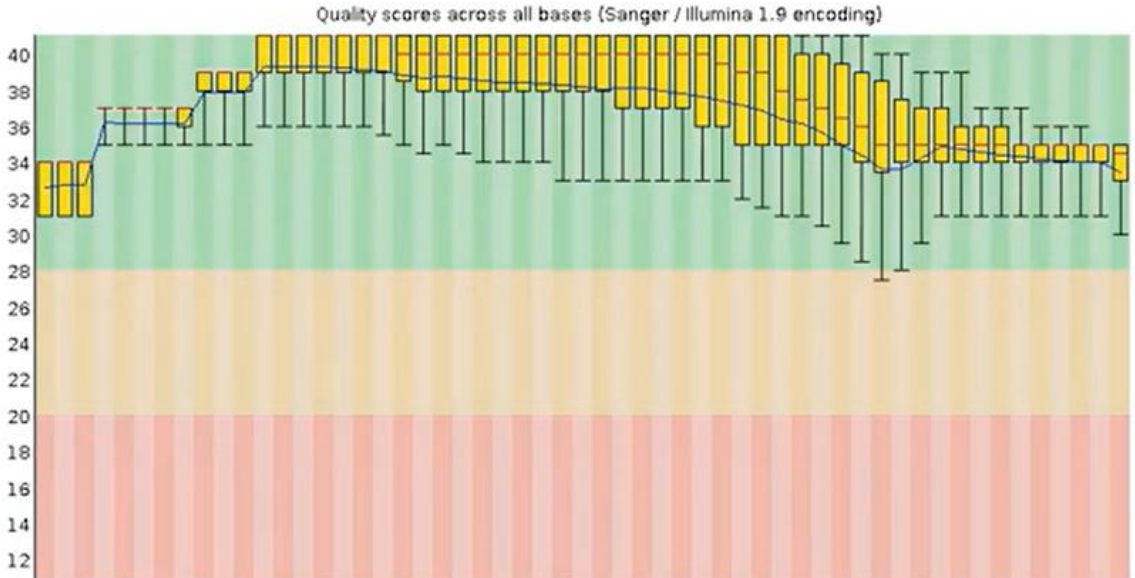
Summary

- Basic Statistics
- Per base 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

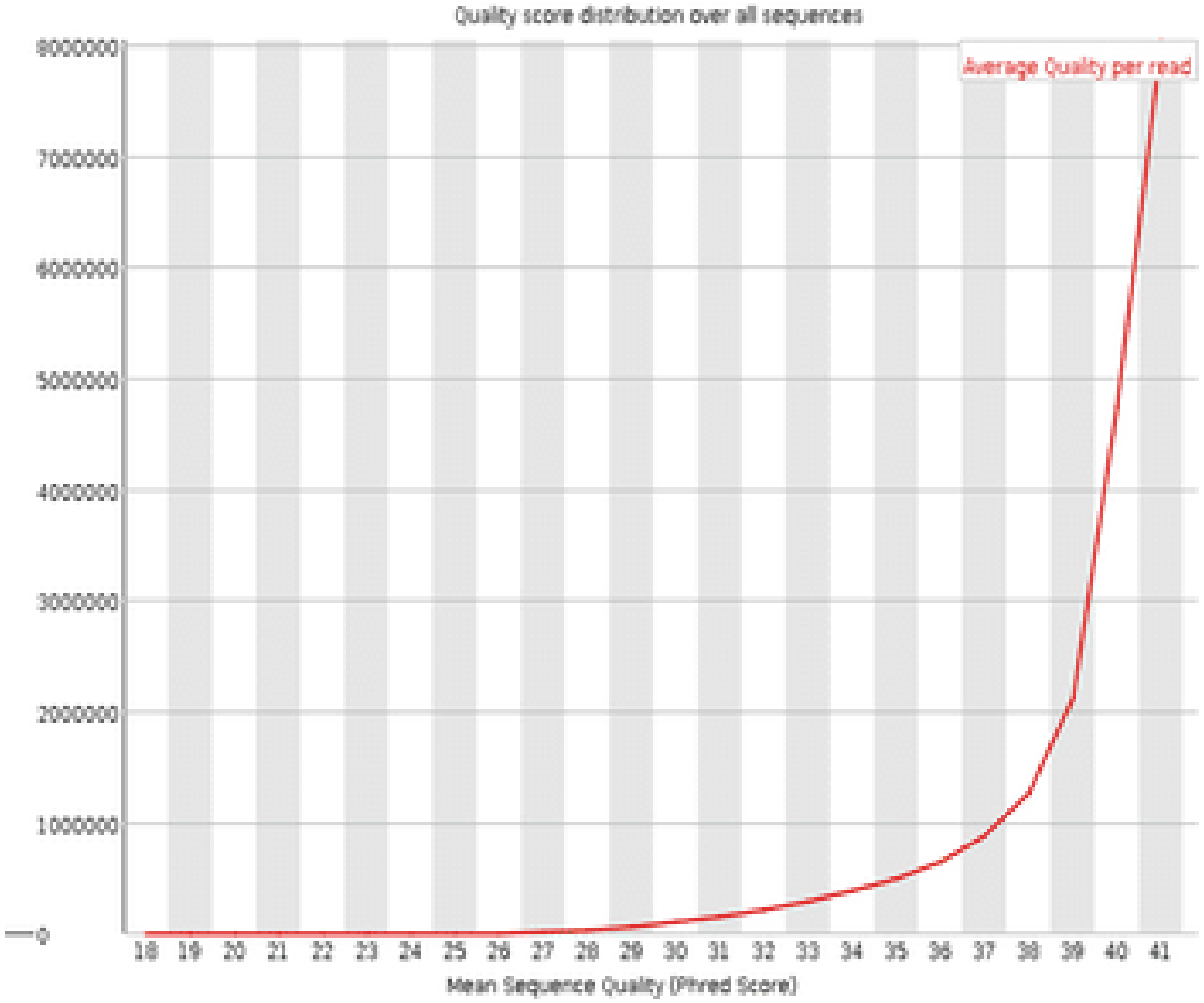
Basic Statistics

Measure	Value
Filename	SRR6468671_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2797765
Sequences flagged as poor quality	0
Sequence length	78-101
%GC	43

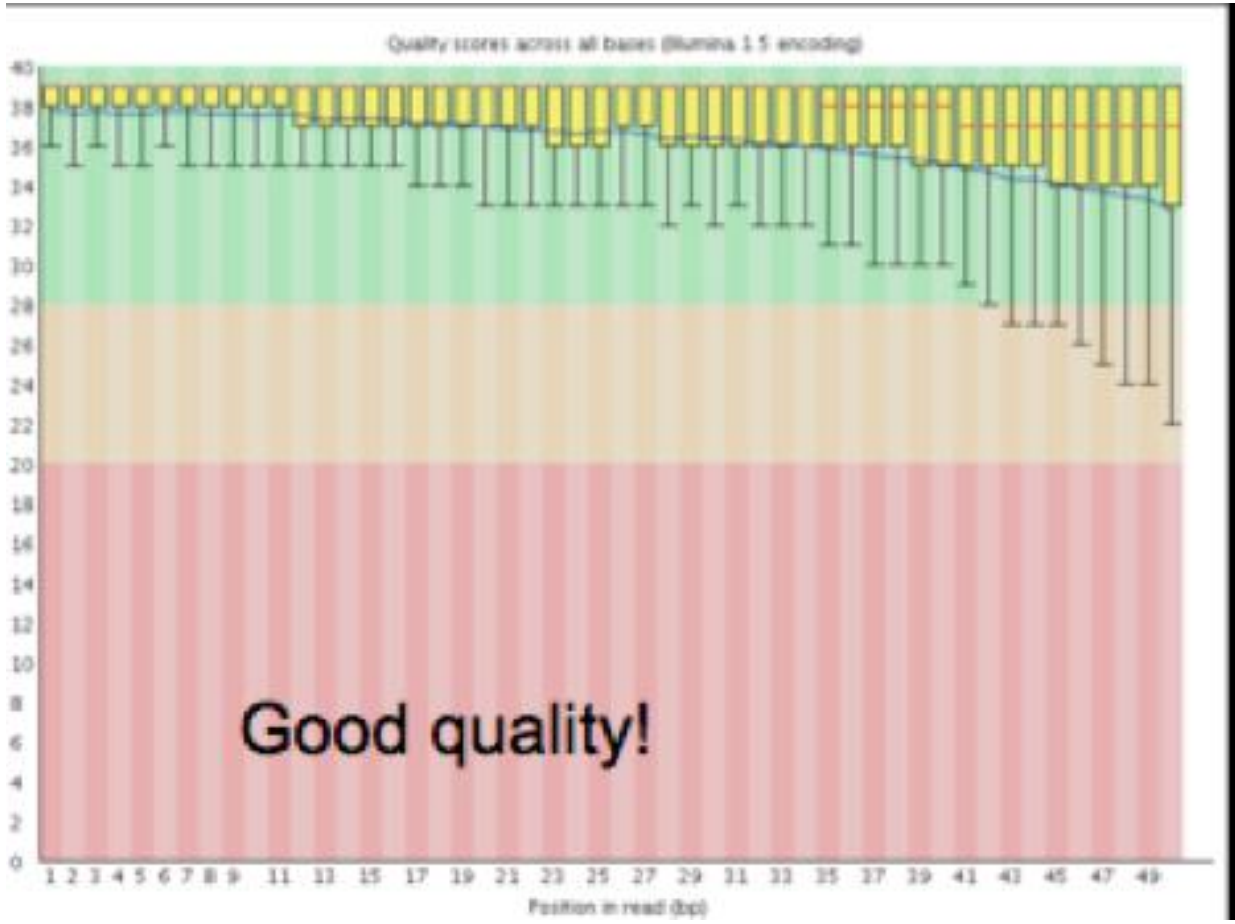
Per base sequence quality

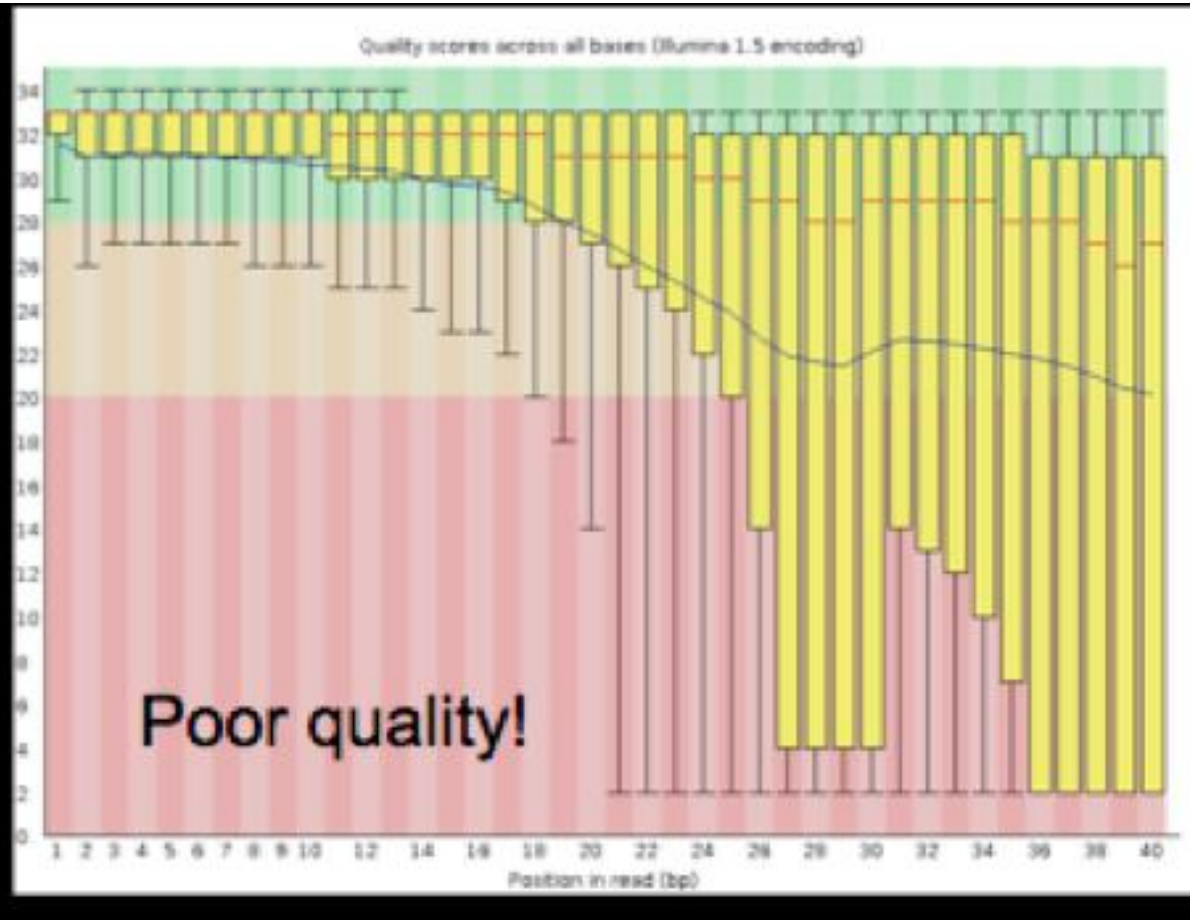


X : Position
Y : Quality



More than 8000000 base pairs Quality are 40


Quality scores across all bases (Illumina 1.5 encoding)
Good quality!
Position in read (bp)


Quality scores across all bases (Illumina 1.5 encoding)
Poor quality!
Position in read (bp)

Trimming

Trimmomatic: A flexible read trimming tool for Illumina NGS data .

Download :

http://www.usadellab.org/cms/index.php?page=trimmomatic

Install :

$ sudo apt-get install trimmomatic

$ java -jar trimmomatic-0.35.jar SE/PE 2 input files and 4 outputs

LEADING:25 (Trim from left side , quality < 25)

TRAILING: 25 (Trim from right side , quality < 25)

SLIDINGWINDOW:4: 25 (Slidin window size 4 pair, trim next by meanquality < 25)

MINLEN:36 (Remove if Remaining real < 36)

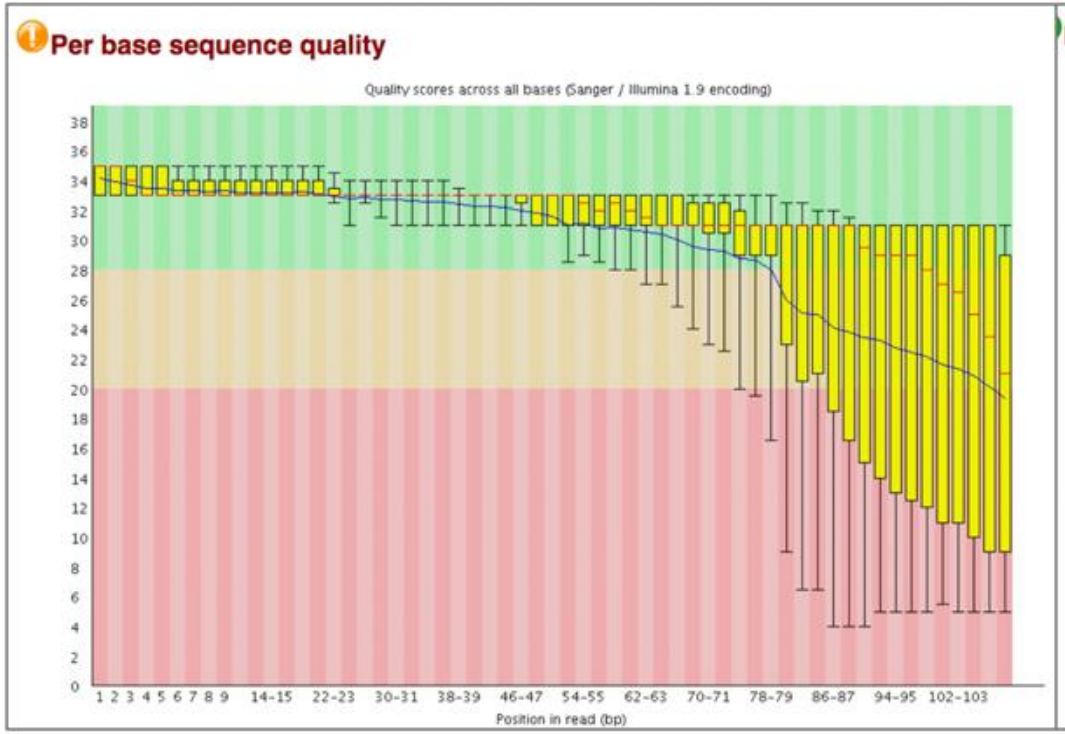
Required :

jre-8v261-linux-x64.tar.gz --------→ tar xvzf jre-8v261-linux-x64.tar.gz

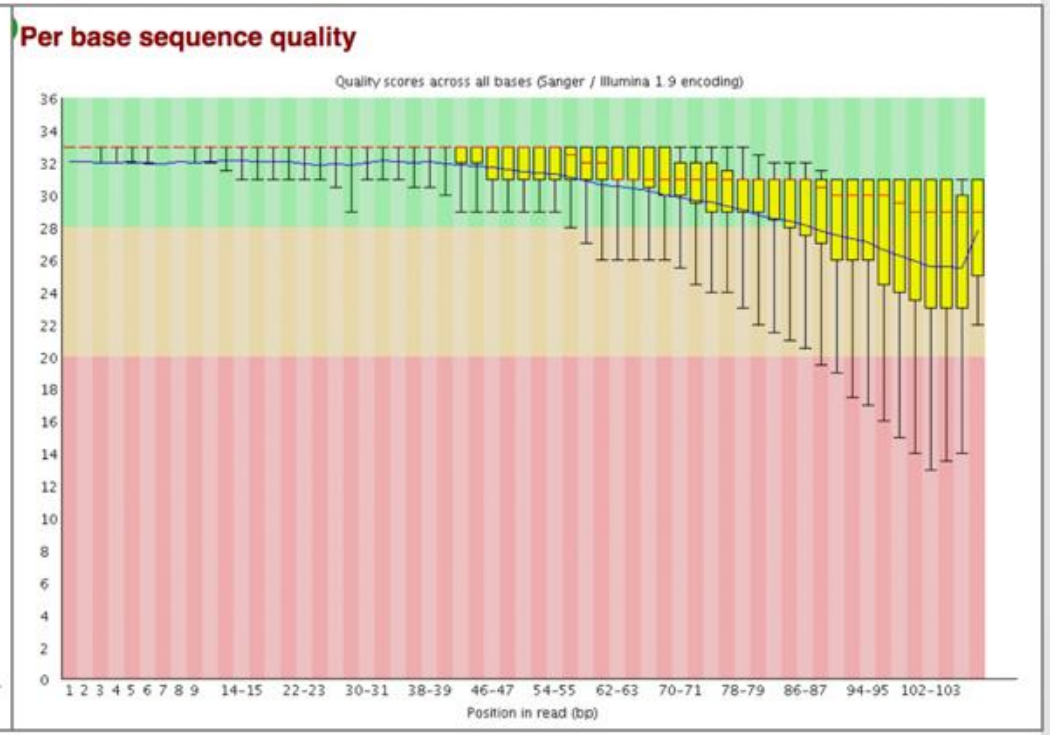
Input : SRR6468671_1.fastq SRR6468671_2.fastq

Output : SRR6468671_1_trimmed.fastq SRR6468671_1_untrimmed.fastq
SRR6468671_2_trimmed.fastq SRR6468671_2_untrimmed.fastq

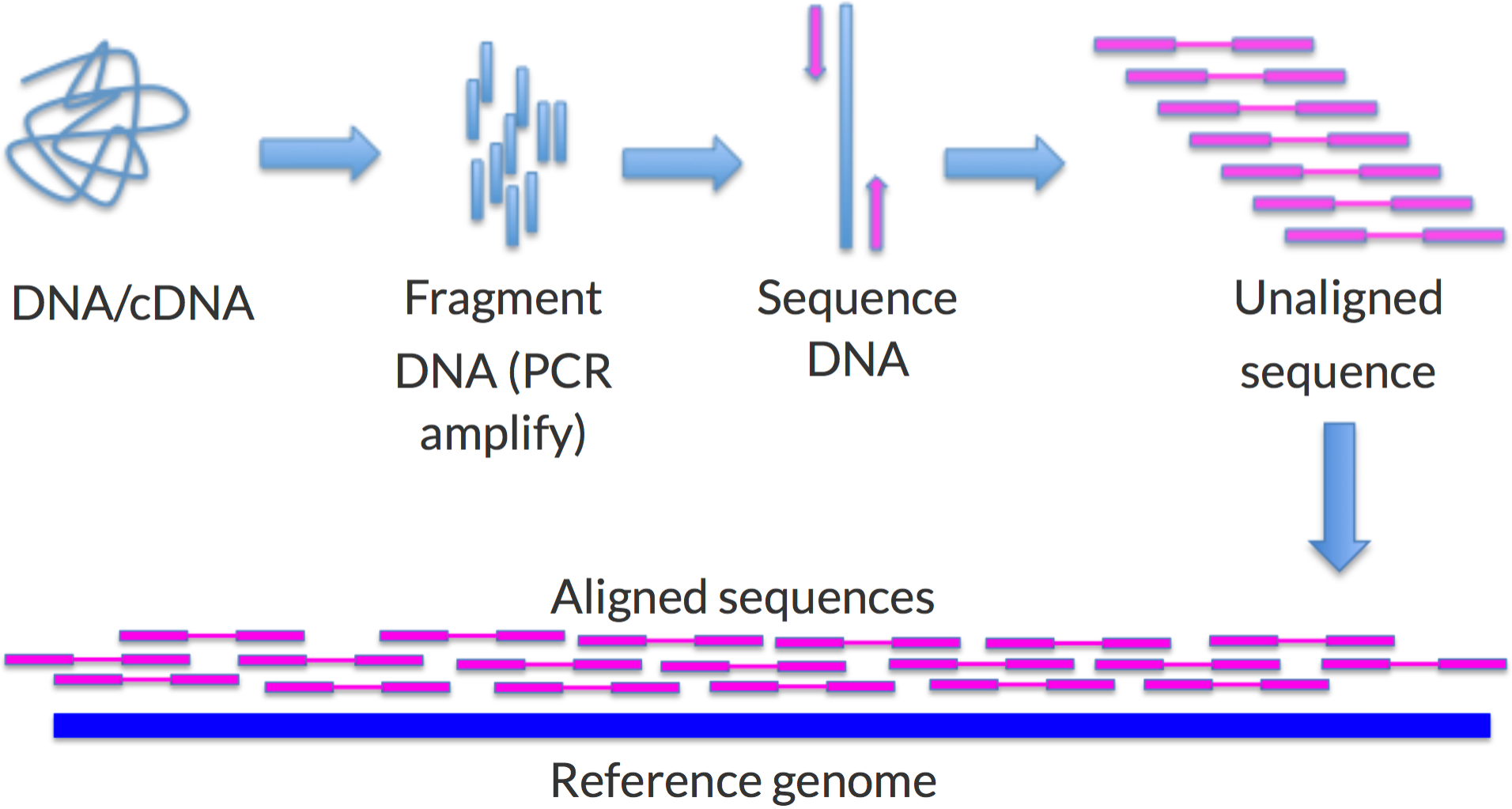
Before trimming



After trimming



Alligning


DNA/cDNA
Fragment DNA (PCR amplify)
Sequence DNA
Unaligned sequence
Aligned sequences
Reference genome

Hisat2

HISAT2 is a fast and sensitive alignment program

for mapping next-generation sequencing reads

(both DNA and RNA) to a population of human

genomes as well as to a single reference genome.

http://daehwankimlab.github.io/hisat2/

Image Description Ge... | Help with Questions | Free Text to Speech O... | Free AI Assistant | Hamster Kombat Daily... | دانلود سریال Arrow | Describe Images With ... | Lumina: Research at th... | BigBlu

HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome. Based on an extension of BWT for graphs (Sirén et al. 2014), we designed and implemented a graph FM index (GFM), an original approach and its first implementation. In addition to using one global GFM index that represents a population of human genomes, **HISAT2** uses a large set of small GFM indexes that collectively cover the whole genome. These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

The HISAT-3N paper published at *Genome Research*. 7/1/2021

HISAT-3N beta release 12/14/2020

HISAT-3N is a software system for analyzing nucleotide conversion sequencing reads. See the HISAT-3N for more details.

Index files are moved to the AWS Public Dataset Program. 9/3/2020

We have moved HISAT2 index files to the AWS Public Dataset Program. See the link for more details.

HISAT 2.2.1 release 7/24/2020

This patch version includes the following changes.

- Python3 support
- Remove the HISAT-genotype related scripts. HISAT-genotype moved to http://daehwankimlab.github.io/hisat-genotype/
- Fixed bugs related to `--read-lengths` option

HISAT 2.2.0 release 2/6/2020

Search

- Main
- About
- Manual
- HISAT-3N
- Download
- HowTo
- Links

Funding

This work was supported in part by the National Human Genome Research Institute under grants R01-HG006102 and R01-HG006677, and NIH grants R01-LM06845 and R01-GM083873 and NSF grant CCF-0347992 to Steven L. Salzberg and by the Cancer Prevention Research Institute of Texas under grant RR170068 and NIH grant R01-GM135341 to Daehwan Kim

Getting Help

Please use hisat2.genomics@gmail.com for private communications only. Please do not email technical questions to HISAT2 contributors directly.

Binaries

Version: HISAT2 2.2.1

Release Date: 7/24/2020

Source	https://cloud.biohpc.swmed.edu/index.php/s/fE9QCsX3NH4QwBi/download
OSX_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/zMgEtnF6LjnjFrr/download
Linux_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/oTtGWbWjaxsQ2Ho/download

Version: HISAT2 2.2.0

Release Date: 2/6/2020

Source	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-220-source/download
OSX_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-220-OSX_x86_64/download
Linux_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-220-Linux_x86_64/download

Version: HISAT2 2.1.0

Release Date: 6/8/2017

Source	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-210-source/download
OSX_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-210-OSX_x86_64/download
Linux_x86_64	https://cloud.biohpc.swmed.edu/index.php/s/hisat2-210-Linux_x86_64/download
Windows	http://www.di.fc.ul.pt/~afalcao/hisat2_windows.html

Index

HISAT2 indexes are hosted on AWS (Amazon Web Services), thanks to the AWS Public Datasets program. Click this link for more details.

H. sapiens

- GRCh38

genome	https://genome-idx.s3.amazonaws.com/hisat/grch38_genome.tar.gz
genome_snp	https://genome-idx.s3.amazonaws.com/hisat/grch38_snp.tar.gz
genome_tran	https://genome-idx.s3.amazonaws.com/hisat/grch38_tran.tar.gz
genome_snp_tran	https://genome-idx.s3.amazonaws.com/hisat/grch38_snptran.tar.gz
genome_rep(above 2.2.0)	https://genome-idx.s3.amazonaws.com/hisat/grch38_rep.tar.gz
genome_snp_rep(above 2.2.0)	https://genome-idx.s3.amazonaws.com/hisat/grch38_snprep.tar.gz

- UCSC hg38

genome	https://genome-idx.s3.amazonaws.com/hisat/hg38_genome.tar.gz
genome_tran	https://genome-idx.s3.amazonaws.com/hisat/hg38_tran.tar.gz

- GRCh37

genome	https://genome-idx.s3.amazonaws.com/hisat/grch37_genome.tar.gz
genome_snp	https://genome-idx.s3.amazonaws.com/hisat/grch37_snp.tar.gz
genome_tran	https://genome-idx.s3.amazonaws.com/hisat/grch37_tran.tar.gz
genome_snp_tran	https://genome-idx.s3.amazonaws.com/hisat/grch37_snptran.tar.gz

If the Index file is not available , we must build one , Then Download The reference Genom as a Fasta format and use given command :

$ hisat-build reference.fasta

Here we have it !
The index files are given for ChrX as 8 files . (chrX_tran_1.ht2 , ..)

$ hisat2 -q -x < Index.ht2 folder > { -1 m1 -2 m2 } --add-chrname S name.sam

To investigate samfile we must use samview but "head" command could be benefitial !

$ head sam ERR188044.9

```
ERR188044.9     83      chrX    12976952        60      76M     =       12976868        -160    AACTACTGACAACGAAGGCCGCGCCT
GCCTTTCCCATCTGTCTATCTATCTGGCTGGCAGGGAAGGAAAGAACTTG      >>CDCDDC@CC?DCEADDHIIJJIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJHHH
HHFFFFFCCC      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:76 YS:i:0  YT:Z:CP NH:i:1
ERR188044.9     163     chrX    12976868        60      76M     =       12976952        160     TACTTCTTTTAGCTGTTTAACTTTGT
AAGATGCAAAGAGGTTGGATCAAGTTTAAATGACTGTGCTGCCCCTTTCA      CCCFFFFFHHHHHJJJJJJJJJJJIHFHGHIJJJIJJJJJGHIJJJJJJJJJJJJJJJJJJJIJJJ
JJJJJIJIJI      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:76 YS:i:0  YT:Z:CP NH:i:1
ERR188044.35    83      chrX    65517987        60      1S75M   =       65512853        -3288   TCTGAGCTAACCTGCCATGCAGAGCC
CTGCAAAGCTCCTCTTTTCTGGGCCAGAAGCCTAGCATTCTCGGCAGTGG      @DDDCC@C>;DFFFHHHHHJJIIIIIJIFJIJJJIJJJJJJJJJJJJJJJIIJJJJJJIJJJIHGH
GHFFFFFCCC      AS:i:-1 XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:0  YT:Z:CP NH:i:1
ERR188044.35    163     chrX    65512853        60      49M1921N27M     =       65517987        3288    GCTGTTGCTGCAGTTGCC
ACTGCCGACACCACAGCTCAGGCCCAAGGTGTCAGTCTTGCATGTTGAGGGTTCCAGC      CCCFFFFFDHHGHHHJJJIJJJJJGGIJJJIJJIJJJJJJJJJJJIJ8@CBEHFHIJJ
JJHHHHHGHFFFEEDEEE      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:76 YS:i:-1 YT:Z:CP XS:A:-  NH:i:1
ERR188044.92    99      chrX    12976800        60      76M     =       12976947        222     GGCGAATCGTAATGAGGCGTGCGCCG
CCAATATGCACTGTACATTCCACAAGCATTGCCTTCTTATTTTACTTCTT      CCCFFFFFHHHHHJJJJIJEHGGGHIIJJJJIJJJJIJJHGHHHHHHGFFFFFEEEEEEDDDDDDD
EEEEDDCCDA      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:76 YS:i:0  YT:Z:CP NH:i:1
ERR188044.92    147     chrX    12976947        60      75M     =       12976800        -222    CAAAGAACTACTGACAACGAAGGCCG
CGCCTGCCTTTCCCATCTGTCTATCTATCTGGCTGGCAGGGAAGGAAAG       @DCB;;->;5@>@=CC@ADD>B8HFGHEJJJJJJJJIJJJJIJIJJJJJJJIIIJJJJJJJJHHHH
HFFFFFCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:0  YT:Z:CP NH:i:1
ERR188044.120   83      chrX    103685235       60      26M1370N50M     =       103677044       -6820   GCAAACTTAGCTTTTCTG
ATGGTGACCTGAAACGAGAATCCAGATCTTCCCAGCAGCCGACGTGTGATCACTGCCG      <C@;3ECEC?7CFFHA=C>C@).>@C8;;GCCAF@D9?4F?GB?8DBFGGA@C6CGFA
CAHE@FFHDBADFDD@C@      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:76 YS:i:0  YT:Z:CP XS:A:-  NH:i:1
```

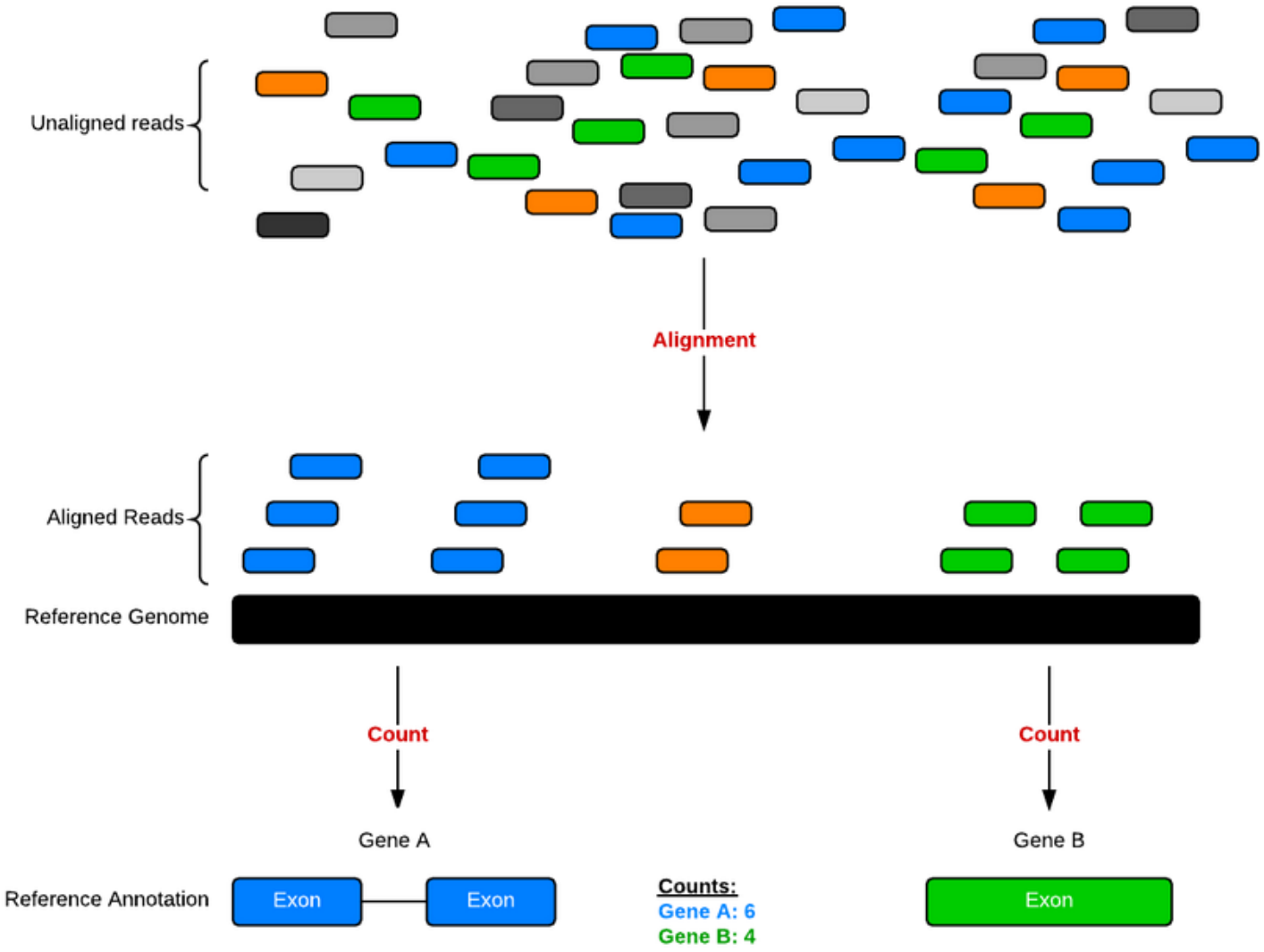
Sam tools

- To identify read count first we must have BAM file .
- Then , converting ***sam*** file to ***bam*** file by using samtools !

- $ samtools view –F 4 –b filename.sam output.bam

```
173M 17:46 15    آقوئست  ERR188044.bam
641M 11:19 10    آقوئست  ERR188044.sam
646M 11:22 10    آقوئست  ERR188104.sam
```

Gene Counting



Requirements :

*.sam
*.bam
*.gtf

GTF

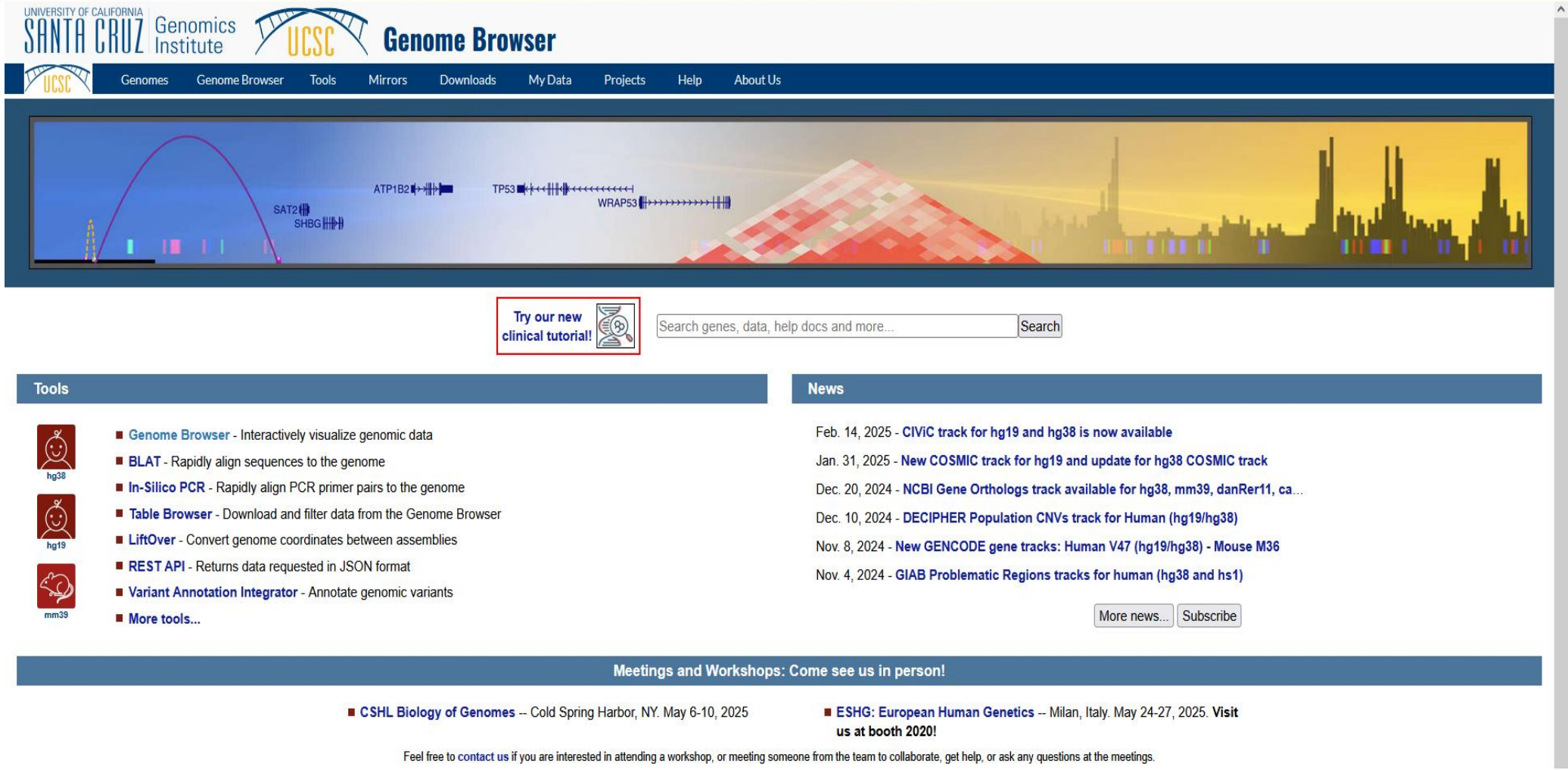
The Gene Transfer Format (GTF) is a widely used format forstoring gene annotations.

You can obtain GTF files easily from the

UCSC table browser and Ensembl.

Gtf file

ucsc genome browser



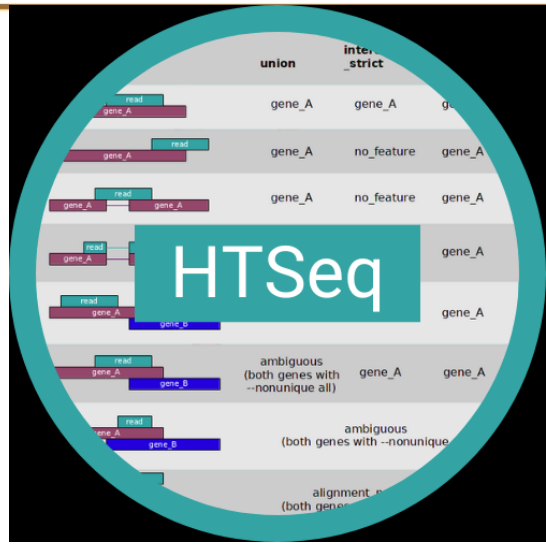
To obtain “gtf” file, identifying the situation of each gene on each chromosome .

For a specific chromosome .

chrUn_KI270751v1.fa.gz	2014-01-23	16:40	48K
chrUn_KI270752v1.fa.gz	2014-01-23	16:40	9.2K
chrUn_KI270753v1.fa.gz	2014-01-23	16:40	20K
chrUn_KI270754v1.fa.gz	2014-01-23	16:40	8.7K
chrUn_KI270755v1.fa.gz	2014-01-23	16:40	11K
chrUn_KI270756v1.fa.gz	2014-01-23	16:40	16K
chrUn_KI270757v1.fa.gz	2014-01-23	16:40	14K
chrX.fa.gz	2014-01-23	16:40	47M
chrX_KI270880v1_alt.fa.gz	2014-01-23	16:40	80K
chrX_KI270881v1_alt.fa.gz	2014-01-23	16:40	46K
chrX_KI270913v1_alt.fa.gz	2014-01-23	16:40	77K
chrY.fa.gz	2014-01-23	16:40	8.0M
chrY_KI270740v1_random.fa.gz	2014-01-23	16:40	5.4K
md5sum.txt	2014-08-21	14:18	26K

Htseq-count

A tool developed with HTSeq that preprocesses RNA-Seq data for differential expression analysis by counting the overlap of reads with genes.



Install :

$ *sudo apt-get install build-essential python2.7-dev python-numpy python-matplotlib python-pysam python-htseq*

Usage :

$ htseq- count [options] <alignment_files(.sam)> <gff_file> > output.count*

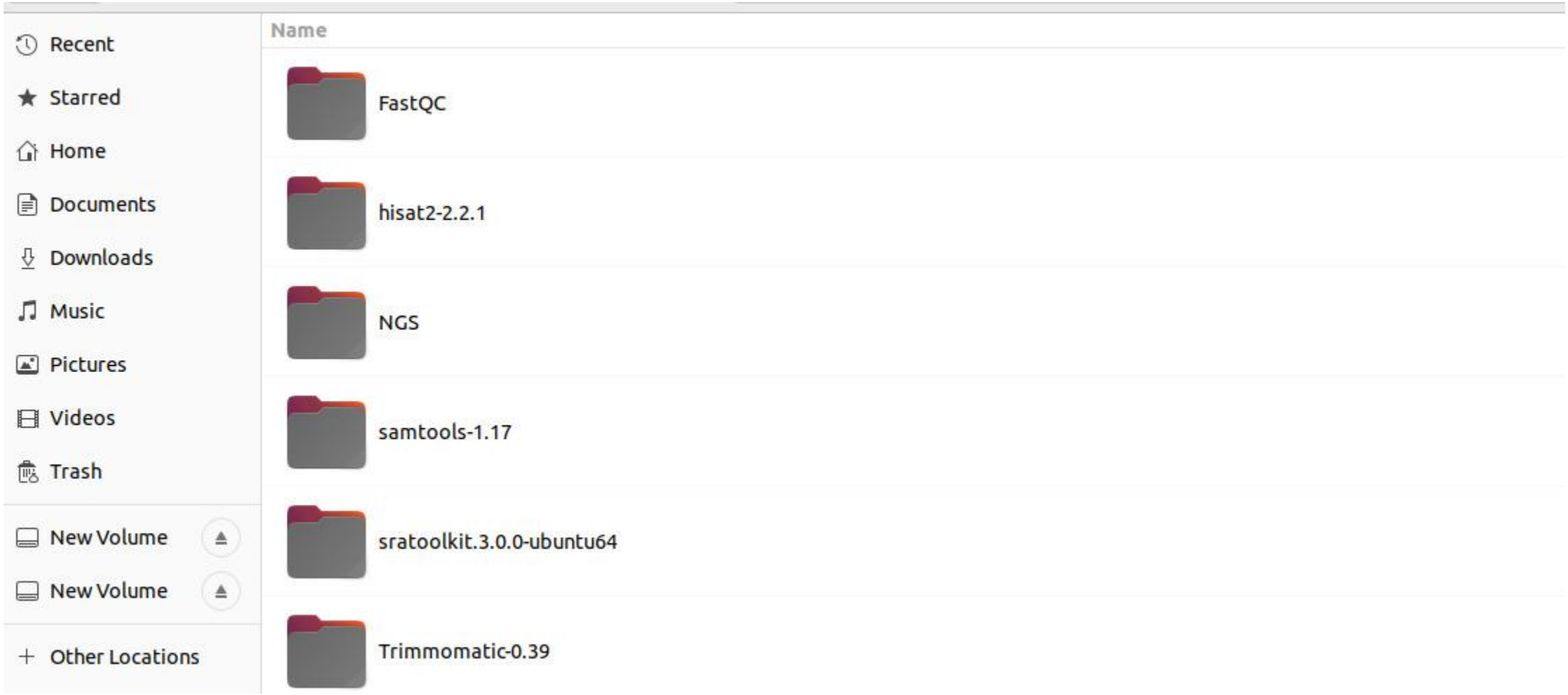
- Count to csv

For each sample , All The pipeline must be done separately .

Samples

Genes

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666


Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
New Volume
New Volume
Other Locations
Name
FastQC
hisat2-2.2.1
NGS
samtools-1.17
sratoolkit.3.0.0-ubuntu64
Trimmomatic-0.39

- fastqc -f fastq ../NGS/samples/ERR188044_chrX_1.fastq ERR188044_chrX_2.fastq

- java -jar trimmomatic-0.39.jar PE ../NGS/samples/ERR188044_chrX_1.fastq ../NGS/samples/ERR188044_chrX_2.fastq ER1trim.fastq ER1untrim.fastq ER2trim.fastq ER2untrim.fastq
  LEADING:25 TRAILING:25
  SLIDINGWINDOW:4:20

hisat2 -q -x ../bulk/NGS/indexes/chrX_tran -1 ../bulk/NGS/samples/ER1trim.fastq -2 ../bulk/NGS/samples/ER2trim.fastq
--add-chrname -S mapped.sam

- htseq-count mapped.sam ../genes/chrX.gtf > ERR18.count

```
sudo -i
wget http://archive.ubuntu.com/ubuntu/pool/main/o/openssl/libssl1.1_1.1.1f-
  1ubuntu2_amd64.deb
sudo dpkg -i libssl1.1_1.1.1f-1ubuntu2_amd64.deb
```

sudo -i

root@s:~# apt --fix-broken install

```
sudo apt-get install build-essential  python3-numpy python3-matplotlib python3-
  pysam python3-htseq
```

Thanks