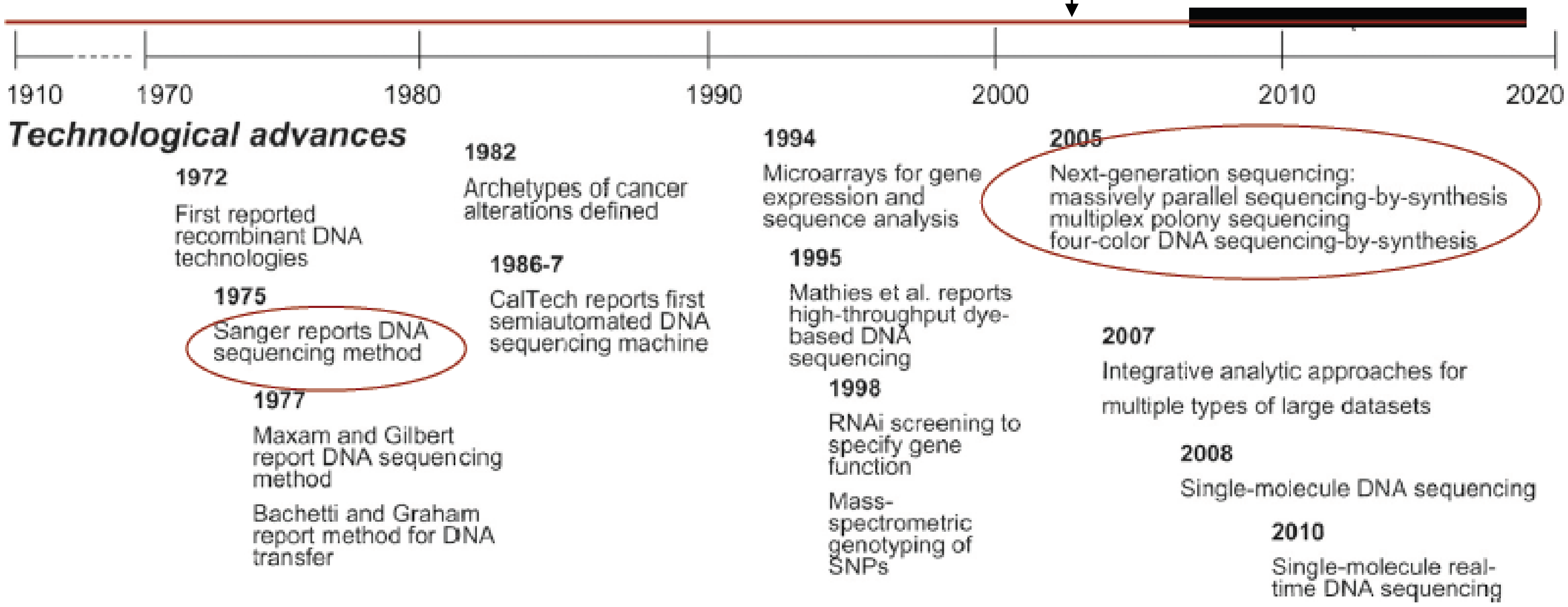
CS342: Bioinformatics
Lecture 2

Assignments

- Read and answer questions about paper
  - "Computational Biology in the 21st Century: Scaling with Compressive Algorithms" by Bonnie Berger, Noah M. Daniels, and Y. William Yu. *Communications of the ACM*, August 2016.
  - Due Wed, Jan 22nd

DNA Sequencing Timeline

April 14, 2003- successful completion of the **Human Genome** Project announced



Technological advances

1972
First reported recombinant DNA technologies

1975
Sanger reports DNA sequencing method

1977
Maxam and Gilbert report DNA sequencing method

Bachetti and Graham report method for DNA transfer

1982
Archetypes of cancer alterations defined

1986-7
CalTech reports first semiautomated DNA sequencing machine

1994
Microarrays for gene expression and sequence analysis

1995
Mathies et al. reports high-throughput dye-based DNA sequencing

1998
RNAi screening to specify gene function

Mass-spectrometric genotyping of SNPs

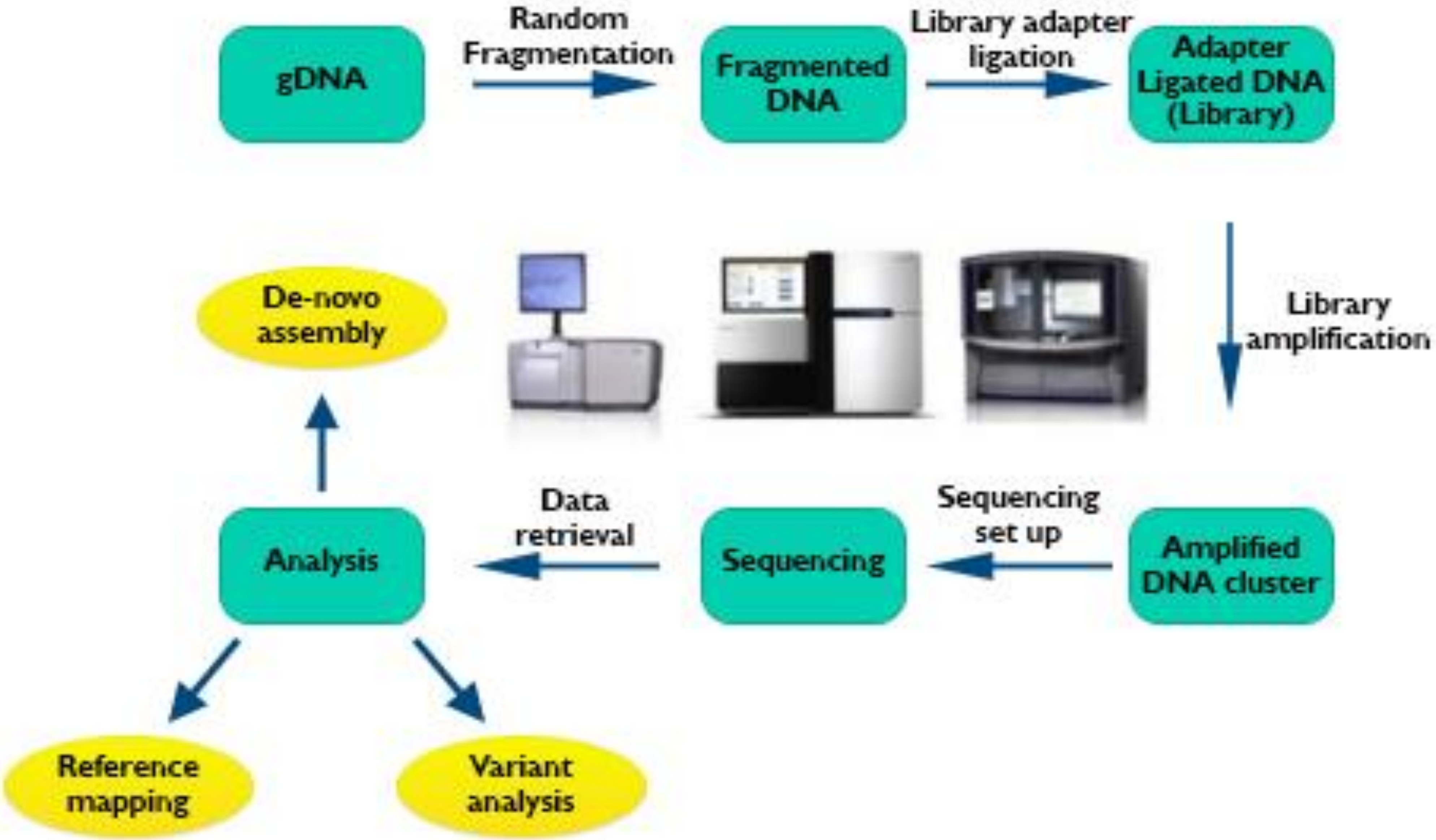
2005
Next-generation sequencing:
massively parallel sequencing-by-synthesis
multiplex polony sequencing
four-color DNA sequencing-by-synthesis

2007
Integrative analytic approaches for multiple types of large datasets

2008
Single-molecule DNA sequencing

2010
Single-molecule real-time DNA sequencing

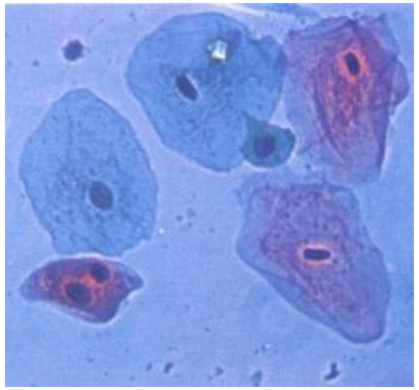
cancerdiscovery.aacrjournals.org


gDNA
Random Fragmentation
Fragmented DNA
Library adapter ligation
Adapter Ligated DNA (Library)
Library amplification
Amplified DNA cluster
Sequencing set up
Sequencing
Data retrieval
Analysis
De-novo assembly
Reference mapping
Variant analysis

DNA Sequencing Technologies

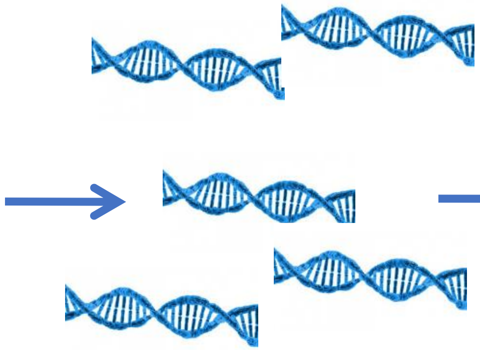
No technology exists that can sequence a complete (human) genome from end to end!!!

Next Generation Bulk Sequencing (10,000 ft view)



DNA extracted from a collection of cells (> 80K cells)

(Human genome is ~3 billion nucleotides long)



DNA sheared into small fragments

Fragments are sequenced

… CATTCAGTAG …

… AGCCATTAG …

… GGTAGTTAG …

… GGTAAACTAG …

… TATAATTAG …

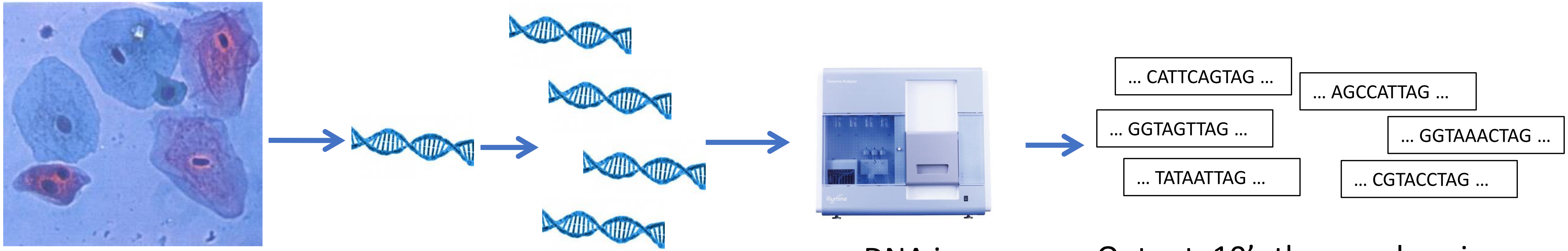
… CGTACCTAG …

Output: 10-100's million noisy ***reads*** (strings)
Reads: 150-1000 nucleotides

DNA Sequencing Technologies

No technology exists that can sequence a complete (human) genome from end to end!!!

Single Cell Sequencing (10,000 ft view)



… CATTCAGTAG …

… AGCCATTAG …

… GGTAGTTAG …

… GGTAAACTAG …

… TATAATTAG …

… CGTACCTAG …

DNA extracted from one cell

(Human genome is ~3 billion nucleotides long)

DNA is amplified

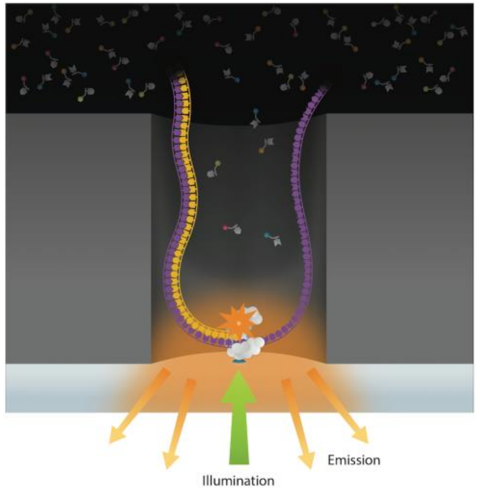
DNA is sequenced

Output: 10’s thousands noisy ***reads*** (strings) per cell

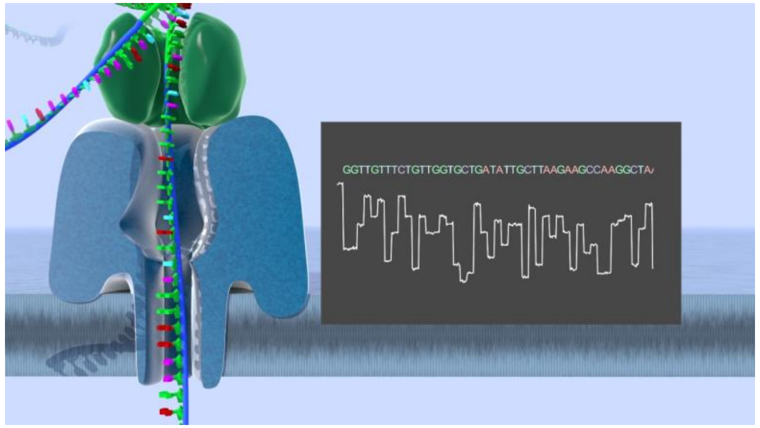
DNA Sequencing Technologies

No technology exists that can sequence a complete (human) genome from end to end!!!

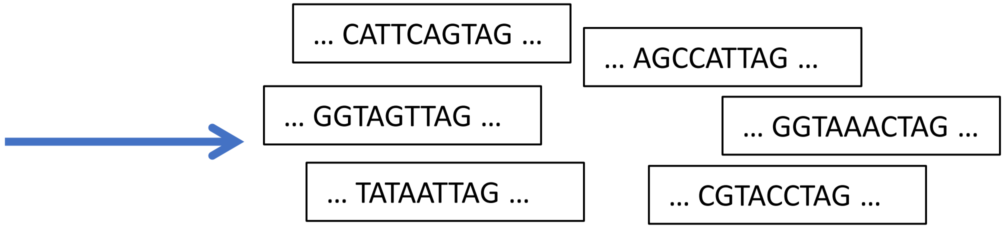
Long Read Sequencing* (10,000 ft view)



DNA Passes through polymerase
PacBio SMRT Seq



DNA Passes through a Nanopore
Oxford Nanopore



... CATTCAGTAG ...

... AGCCATTAG ...

... GGTAGTTAG ...

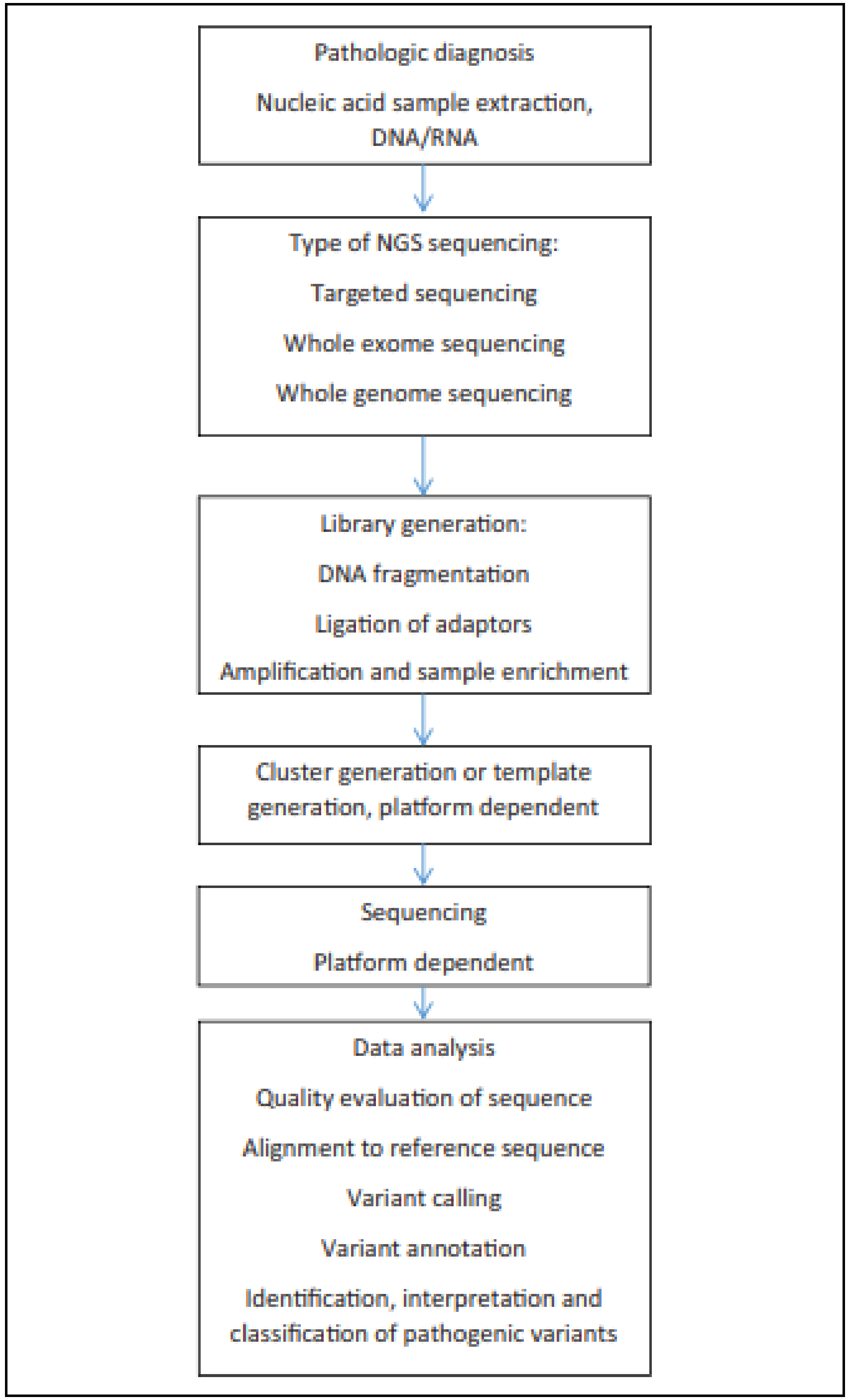
... GGTAAACTAG ...

... TATAATTAG ...

... CGTACCTAG ...

Reads: 10,000's nucleotides

*Emerging technologies
(sometimes called 3rd Generation)



Pathologic diagnosis

Nucleic acid sample extraction, DNA/RNA

Type of NGS sequencing:

Targeted sequencing

Whole exome sequencing

Whole genome sequencing

Library generation:

DNA fragmentation

Ligation of adaptors

Amplification and sample enrichment

Cluster generation or template generation, platform dependent

Sequencing

Platform dependent

Data analysis

Quality evaluation of sequence

Alignment to reference sequence

Variant calling

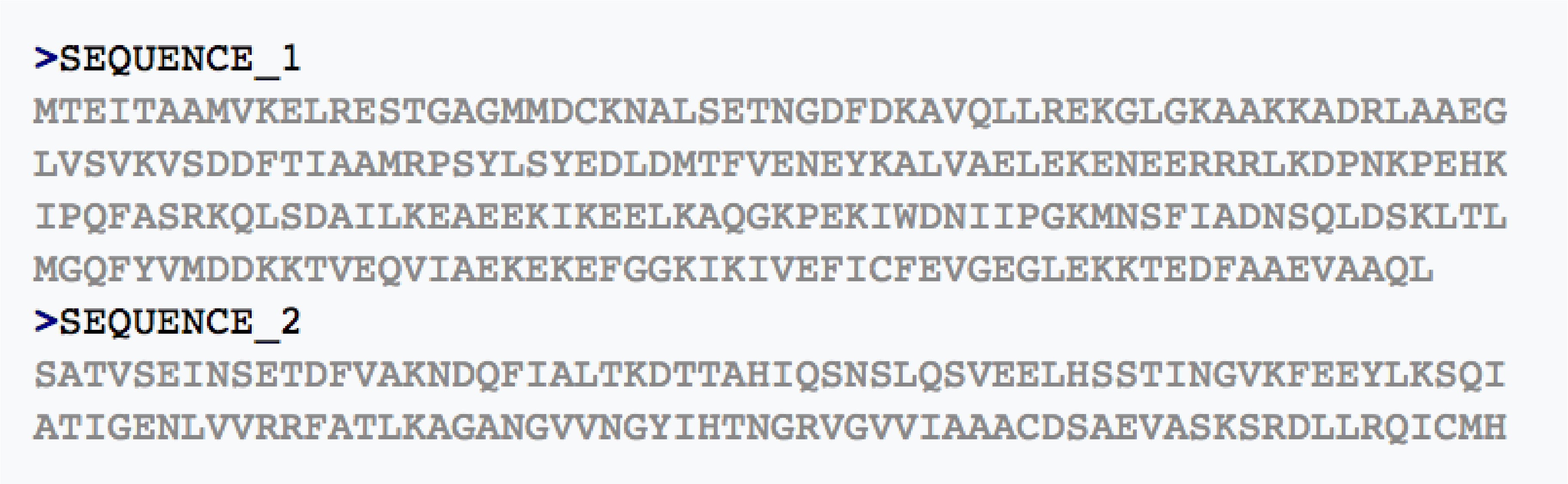
Variant annotation

Identification, interpretation and classification of pathogenic variants

Alekseyev et al., A Next-Generation Sequencing Primer— How Does It Work and What Can It Do?, Academic Pathology Volume 5: 1–11, 2018.

File Format for Sequencing Data?

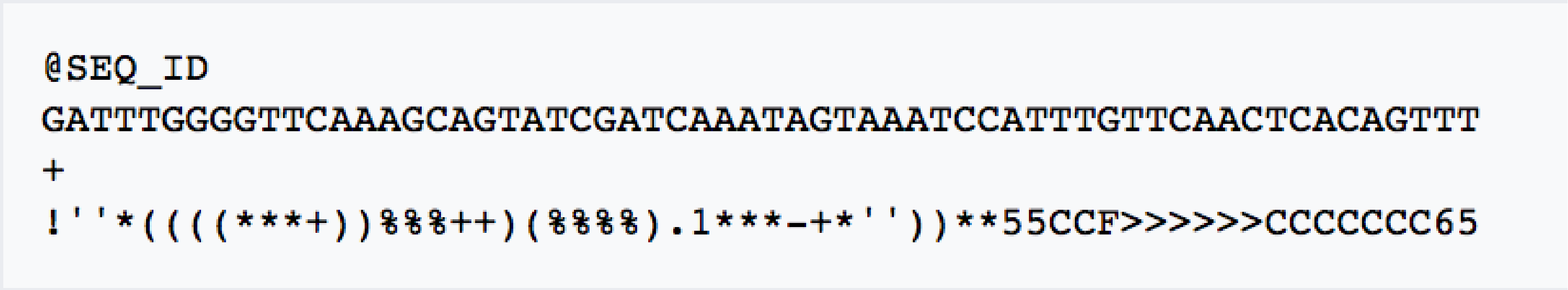
FASTA Format



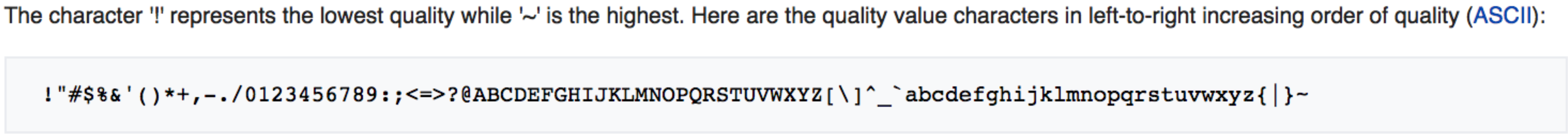
```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

FASTQ

FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores.



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality (ASCII):

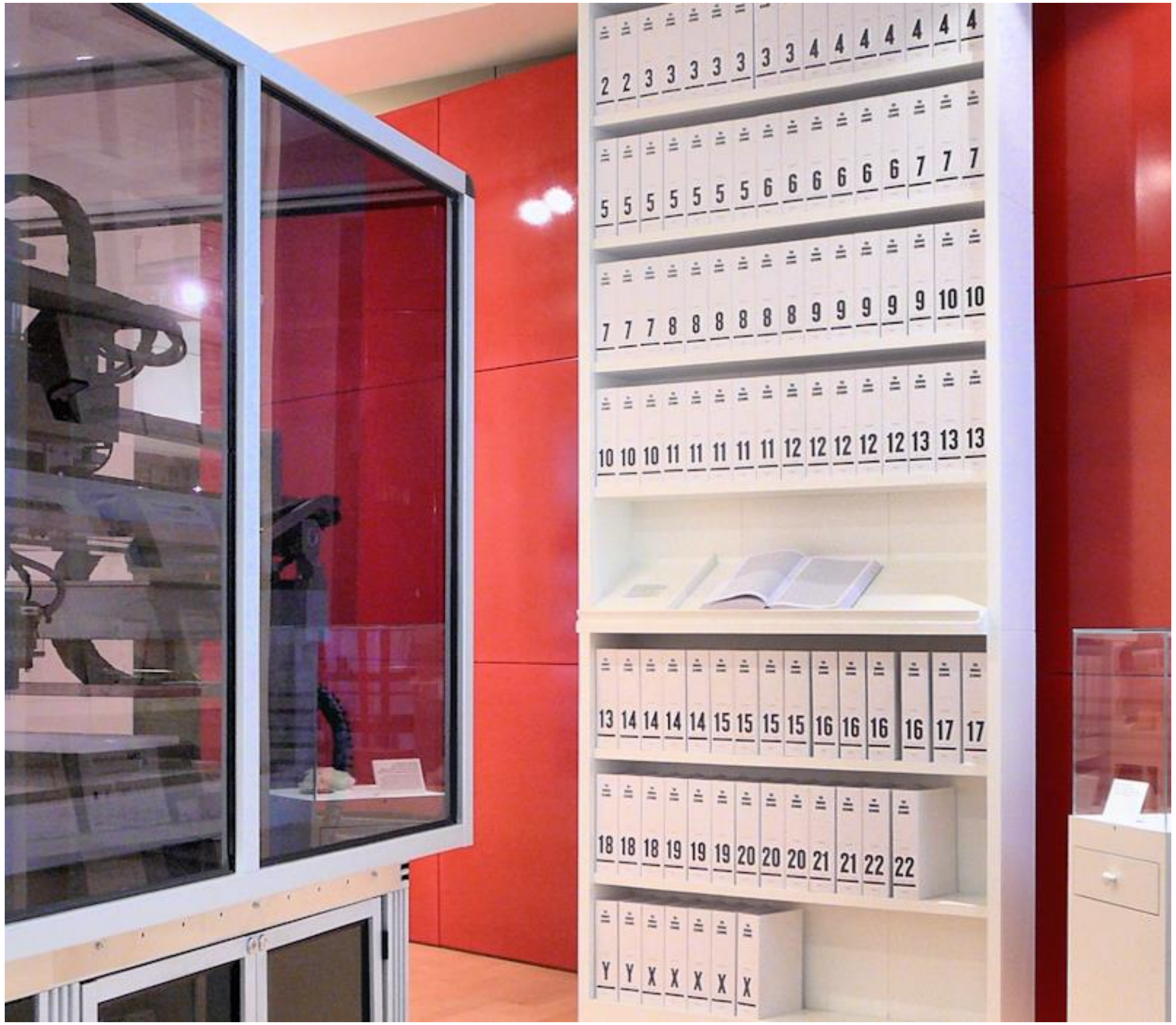
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Reference Genomes

For many species we have compiled what’s called a “reference genome” that indicates what we expect a “typical” genome to look like.

Current human reference is called GRCh38

https://www.ncbi.nlm.nih.gov/genome/guide/human/

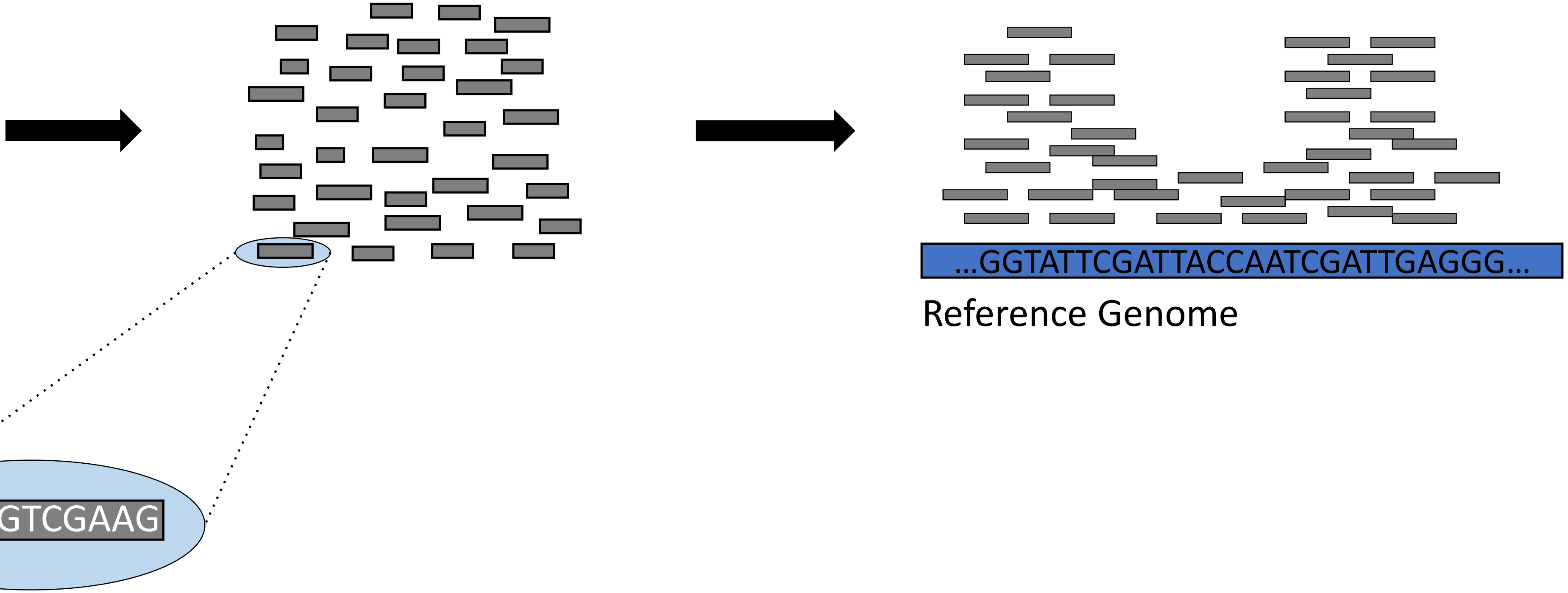


First printout of the human reference genome

Aligning Reads to a Reference

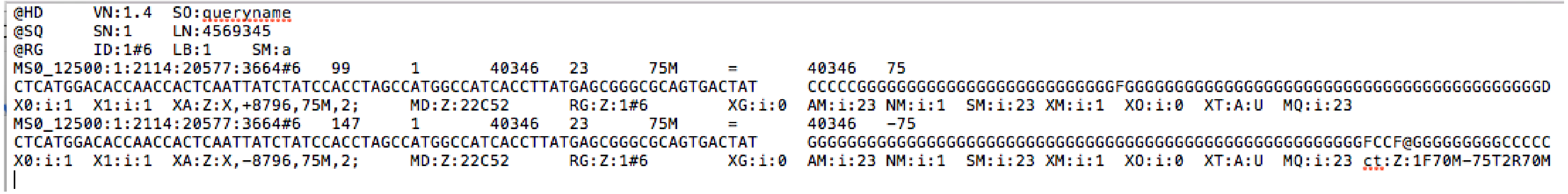
DNA sequenced reads

Reads are aligned to where they “match the best” to reference genome

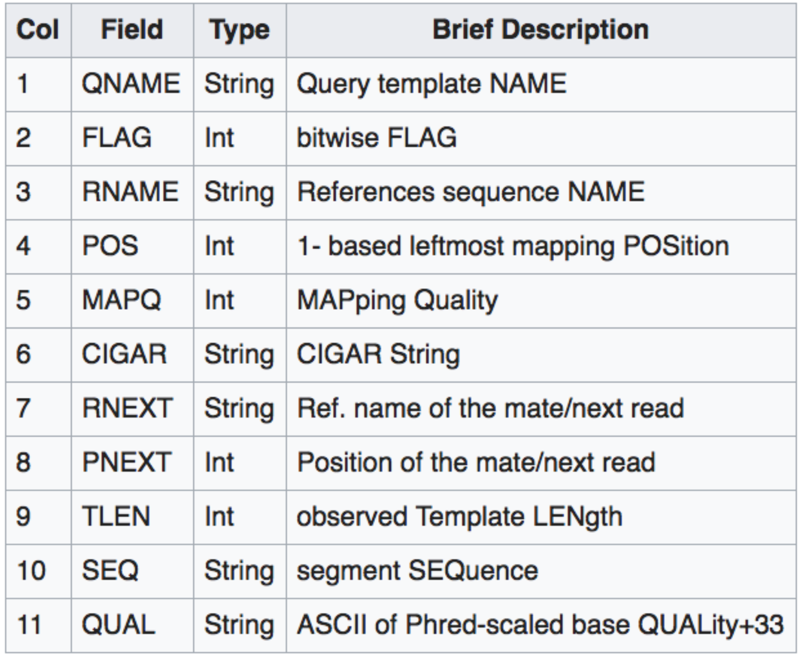


SAM Files

Sequence Alignment Map (SAM) is a text-based format for storing biological sequences aligned to a reference sequence



```
@HD     VN:1.4  SO:queryname
@SQ     SN:1    LN:4569345
@RG     ID:1#6  LB:1    SM:a
MS0_12500:1:2114:20577:3664#6   99      1       40346   23      75M     =       40346   75
CTCATGGACACCAACCACTCAATTATCTATCCACCTAGCCATGGCCATCACCTTATGAGCGGGCGCAGTGACTAT     CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGD
X0:i:1  X1:i:1  XA:Z:X,+8796,75M,2;     MD:Z:22C52      RG:Z:1#6        XG:i:0  AM:i:23 NM:i:1  SM:i:23 XM:i:1  XO:i:0  XT:A:U  MQ:i:23
MS0_12500:1:2114:20577:3664#6   147     1       40346   23      75M     =       40346   -75
CTCATGGACACCAACCACTCAATTATCTATCCACCTAGCCATGGCCATCACCTTATGAGCGGGCGCAGTGACTAT     GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFCCF@GGGGGGGGGCCCCC
X0:i:1  X1:i:1  XA:Z:X,-8796,75M,2;     MD:Z:22C52      RG:Z:1#6        XG:i:0  AM:i:23 NM:i:1  SM:i:23 XM:i:1  XO:i:0  XT:A:U  MQ:i:23 ct:Z:1F70M-75T2R70M
```



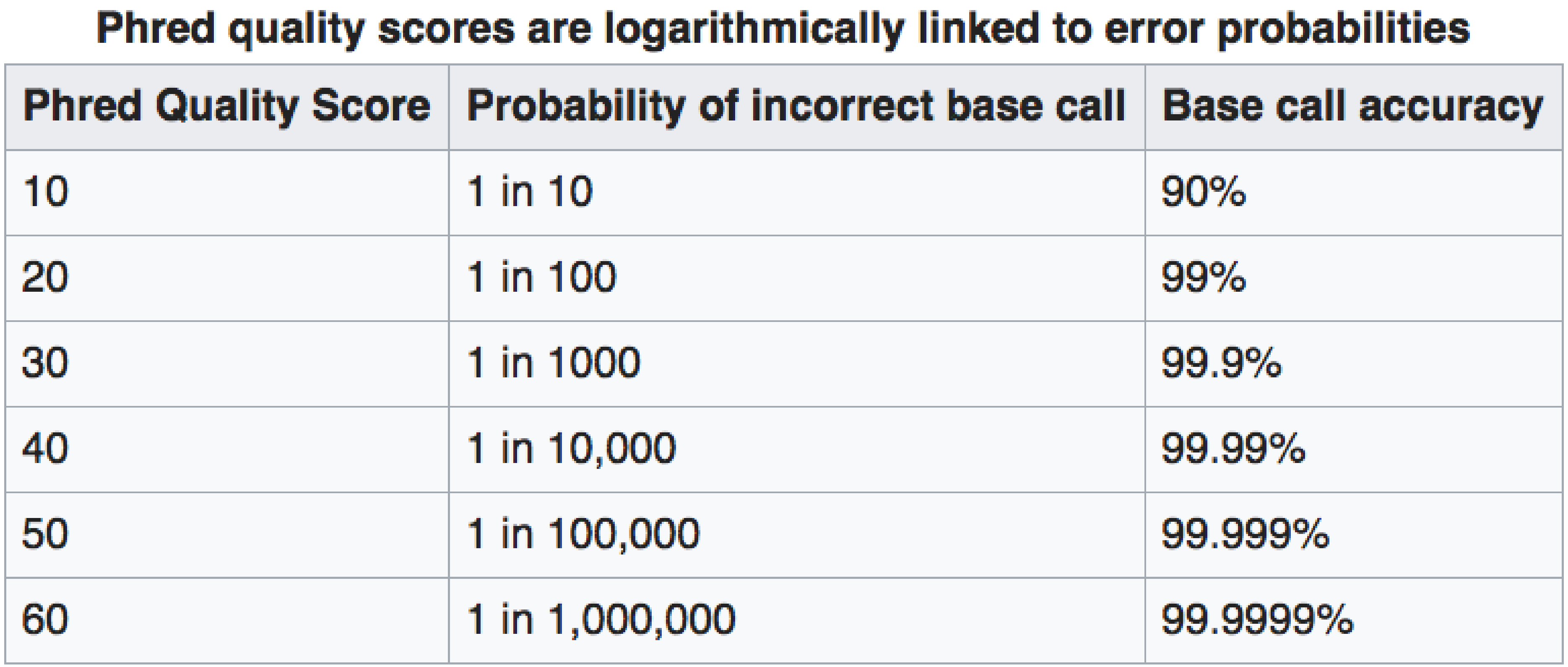
Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Filename.sam

Phred Score

$$Q = -10 \log_{10} P$$

P is base calling error probability.

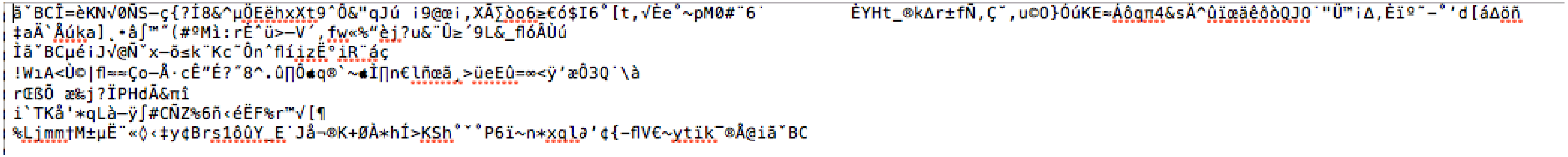


Phred quality scores are logarithmically linked to error probabilities

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%

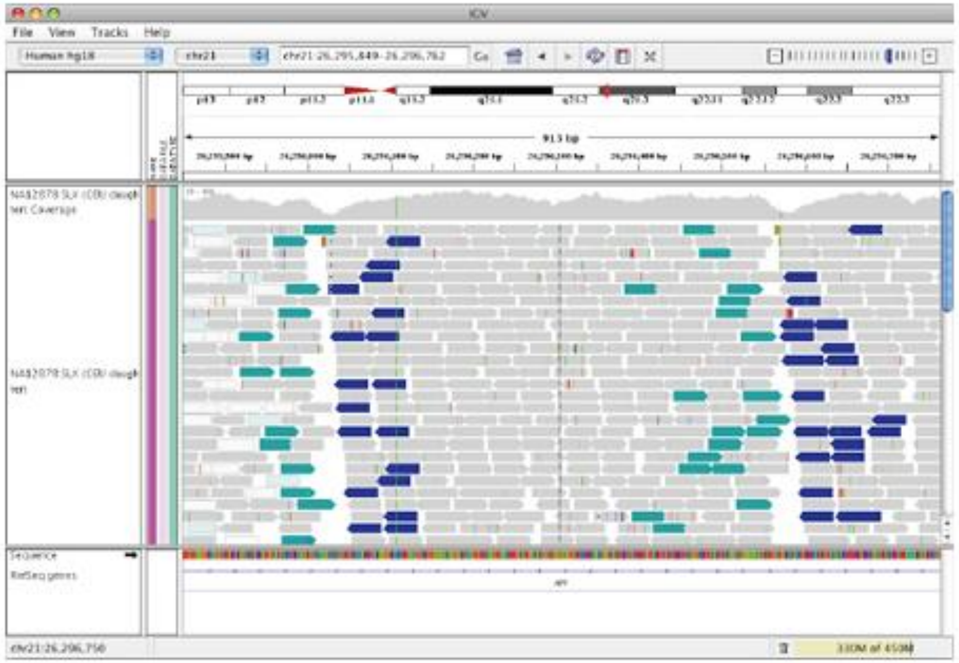
BAM Files

Binary Alignment Map (BAM) is a compressed binary version of the SAM file.



```
ã˘BCÎ=èKN√0ÑS–ç{?İ8&^µÖEëhxXt9˚Ö&"qJú i9@œi,XÃ∑òo6≥€ó$I6˚[t,√Ėe˚~pM0#¨6˙        ĖYHt_®kΔr±fÑ,Ç˘,u©O}ÓúKE≈Áôqπ4&sÄ^ûïœäêôòQJQ˙"Ü™¡Δ,Ėïº¯–˚'d[áΔöñ
‡aÄ`Åúka]¸•â∫™˝(#ºMì:rÊˆü>–V´,fw«%“èj?u&¨Û≥´9L&_ﬂóÂÙú
Ïã˘BCµéiJ√@Ñ˘x–õ≤k¨Kc˜Ôn^ﬂíizË˚iR¨áç
!WıA<Ù©|ﬂ≈≈Ço–Å·cÊ"É?˝8^.û∏Ô◆q®`~◆Ì∏n€lñœã¸>üeEû=∞<ÿ'æÔ3Q˙\à
rŒßÕ æ‰j?ÏPHdÃ&πî
i`TKå'∗qLà–ÿ∫#CÑZ%6ñ‹éËF%r™√[¶
%Ljmm†M±µË¨«◊‹‡y¢Brs1ôÛY_E˙Jå¬®K+ØÀ∗hÎ>KSh˚˘˚P6ï~n∗xqlə'¢{–ﬂV€~ytïk¯®Å@iã˘BC
```

Filename.bam


IGV
File
View
Tracks
Help
Human hg18
chr21
chr21:26,295,849-26,296,762
Go
913 bp
NA12878 SLX (CEU daughter) Coverage
NA12878 SLX (CEU daughter)
Sequence
RefSeq genes
chr21:26,296,750

SAMtools

DNA Sequence Data

How big do you think a file is that stores DNA sequence data for one human (BAM file)?

DNA Sequence Data

DNA Sequence File: 130 GB

For patients with cancer,

we have two files:

Tumor DNA

Normal DNA

1 patient: 260 GB

My laptop:

16 GB RAM

500 GB hard drive

Memory:

16.25% of 1 patient

Disk:

Less than 2 patients