# Review
# BIFX-540

S. *Ravi*chandran, Ph.D.
Hood College, Frederick, MD 21701

# Agenda

- Review of materials covered in the previous classes
- DBs (NCBI, Ensembl, UniProt)
- Sanger Sequencing
- Microarray
- SNP or other genomic variations
- Impact Analysis (PolyPhen2, Ensembl Tools)

# Agenda

- GWAS Tutorial
- https://zzz.bwh.harvard.edu/plink/tutorial.shtml
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

# How does sequencing work?

- DNA
- Create fragments
  - Reads
  - Measurement can result in the reads of same length or shorter
    - Or longer when retained with adapters
    - Mostly the instrument knows the adapters (at the 5’ end but not the 3’ end) and skips them

XXXXGGGGTTTTYYYY

Red font indicate adapters

# Alignment

- Collection of short DNA pieces called library
  - Short DNA is called fragment (F ~ 450 bp)
- Each segment can be sequenced
  - One end or both ends (Paired-End Sequencing)
  - HiSeq or MiSeq (instruments that can do PES)
- Famous aligner is called BWA

# Adapters

```
Illumina Universal Adapter AGATCGGAAGAG
Illumina Small RNA 3' Adapter TGGAATTCTCGG
Illumina Small RNA 5' Adapter GATCGTCGGACT
SOLID Small RNA Adapter CGCCTTGGCCGT
```

- You don’t want adapters in your reads.
- Software packages are available to trim them
- Trimmomatic

# Adapters

- The sequencer will typically recognize the XXXX at the beginning and will skip reading and not report the fragment
- We can illustrate the process with arrows that show the direction of sequencing. Suppose that the sequencer can generate reads of lengths

```
5:---->
  AAAAT
  AAAATGGGG
  TTTTAGGGG
      AGGGG
      <---- 3:
```

# Adapters

- “read-through”, where the sequencing reads are longer than the fragments
  - Reads will include small section of the 3’ adapter ends. They can be removed computationally after the sequencing

# Adapters

During the library preparation, uniquely designed DNA adapters of lengths that are ~ 30 bases are appended to the 5', and 3' ends of

each sequence.

Here is an example of the forward and reverse strands of a sequence

**XXXXAAAACCCCYYYY** and **XXXXGGGGTTTTYYYY**

Note that both or neither of these strands may be sequenced for any given fragment.

# Sequencing will be done on both strands

- Size distribution of the reads can vary (typically 25-50 bp)
- **Paired-end Sequencing**
  - Method to sequence both ends of a fragment and to make the pairing information available
  - Illumina PE sequencing
  - The two reads are stored in separate files (fastq)

```
First step "single end reads"
----->
AAAATTTTGGGGCCCC

Then the sequence is flipped
and reverse complemented and
a second measurement is taken
----->
AAAATTTTGGGGCCCC
TTTTAAAACCCCGGGG
          <-------
```

# Sequencing Instruments

- Illumina, MiniSeq, MiSeq, NextSeq, HiSeq
  - Undisputed leader in HTS
  - Up to 300M reads
  - Up to 1500 GB/run (GB: 1 Billion bases)
- IonTorrent, PGM, Proton
  - Up to 400 bp long reads;  Up to 12 GB/run
- PacBio Sequel
  - Specialized in long-read sequencing
  - Up to 12,000 long paired-end reads; up to 4 GB/run

# Sequencing outputs

- Sequencing runs outputs
  - Gzip compressed FASTQ files/sample
  - # of samples = # of FASTQ files
- Raw data
  - Usually submitted to NCBI SRA db
- Sequencing Coverage?
  - # measurements (on average) will be carried out on base of the genome
  - 10x means each base on average sequenced 10 times

# Miscellaneous

- Typical coverages for DNA sequencing
  - Genome assembly ~ 50x & above
  - Variation calling ~ 20x & above
- RNA
  - Since different transcripts are expressed at different levels, no common measure is available
- C = # of sequenced bases/total-genome-size
  - There are other methods available (not covered)

# Miscellaneous

- Note in an expt not all bases will be sequenced. There is a theoretical formula to calculate this measure
  - Lander/Waterman model

# NGS Reads

- SRA (Short Read Archive)
  - Very convoluted, unfriendly database
  - Each Bioproject; PRJNA616147
  - NCBI Biosample SAMN14483190
    - Source of the sample
  - SRA experiment; SRX8023307
  - SRA Run; SRR11445485
    - Data files linked to a given experiment

**COVID-19 virus**

*More than one can be available*

# NGS Reads

- SRA
  - How can I download data?
  - SRA-toolkit
- ENA (SRA equivalent)
- What is a spot?

# FASTQ

- Here is an example of a file with 1 read
- Fist line ID
- Second line read letters
- Third line id
- Fourth line Quality measure

@ID length=5
GTCCA
+ID length=5
CBCFF

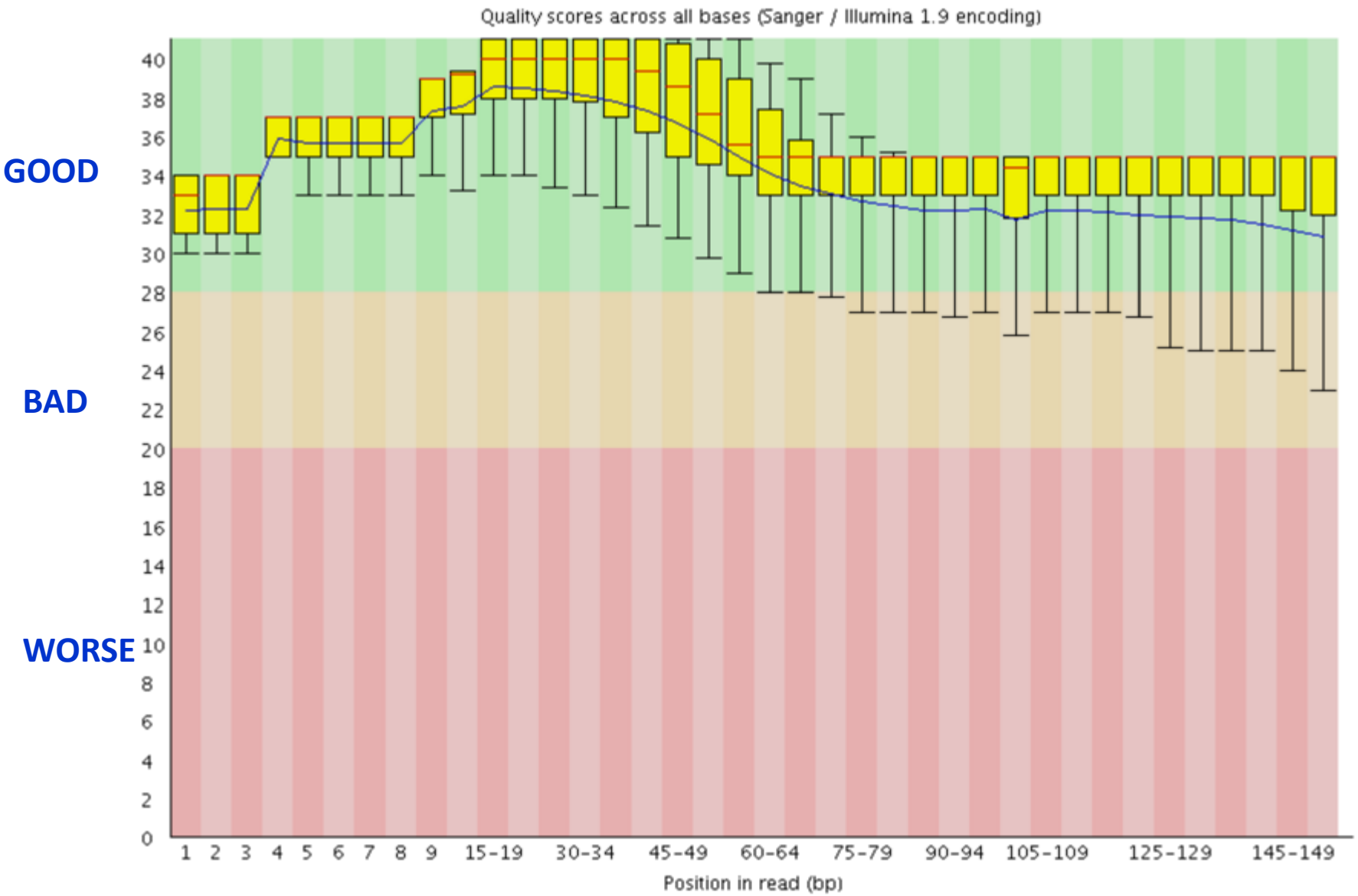
$$P = 10^{\frac{-Q}{10}}$$

Let us calculate Q scores (SANGER) for the first read
ASCII codes for
CBCFF are
67, 66, 67, 70, 70

a) Take each ASCII # subtract 32
b) Use that number from a in the following formula
c) Q scores
d) 67 → 35 → 10^(-3.5) → 0.03% (prob of error)

# FASTQC

- Tool (*analysis/plots are not intuitive*) to analyze the quality of reads
  - Use Galaxy to do a sample QC analysis
- Note that FASTQC only visualizes/analyzes the QC doesn't carry out QC on the reads
- How does this work?
  - FASTQC takes a sample and carry out the analysis to make predictions


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
GOOD
BAD
WORSE
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 125-129 145-149
Position in read (bp)

# After QC what steps?

- Visual evaluation of QC (FASTQC)
- Move to next step if everything is fine. If not remove data that is bad and move to previous step

- QC tools
  - Trimmomatic, BBDuk, cutadapt

FASTA/FASTQ

**Compute sequence length**

**Concatenate** FASTA alignment by species

**Filter sequences by length**

**FASTQ Quality Trimmer** by sliding window

**FASTQ Trimmer** by column

**Combine FASTA and QUAL** into FASTQ

**Filter FASTQ** reads by quality score and length

**FASTQ Groomer** convert between various FASTQ quality formats

**Manipulate FASTQ** reads on various attributes

**FASTQ Masker** by quality score

**FASTQ splitter** on joined paired end reads

**FASTQ de-interlacer** on paired end reads

**FASTQ interlacer** on paired end reads

**FASTQ joiner** on paired end reads

**Remove sequencing artifacts**

**Rename sequences**

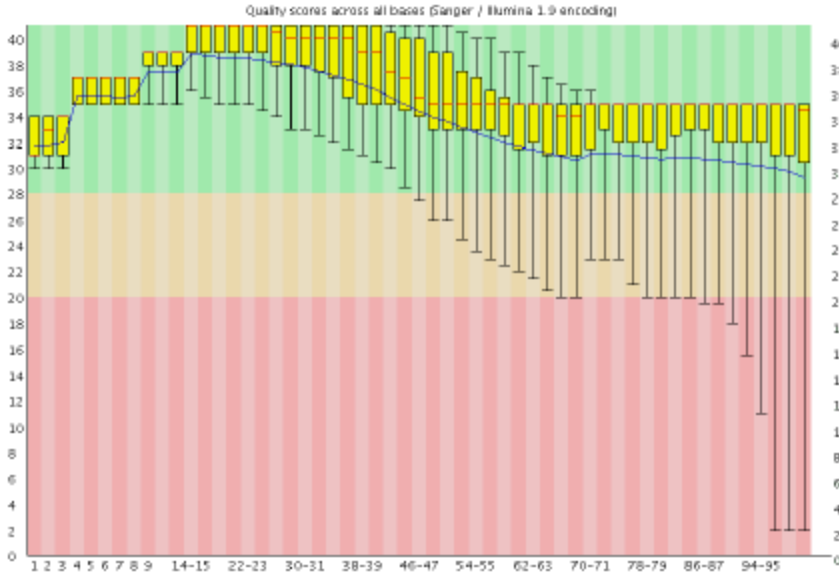
**Filter by quality**

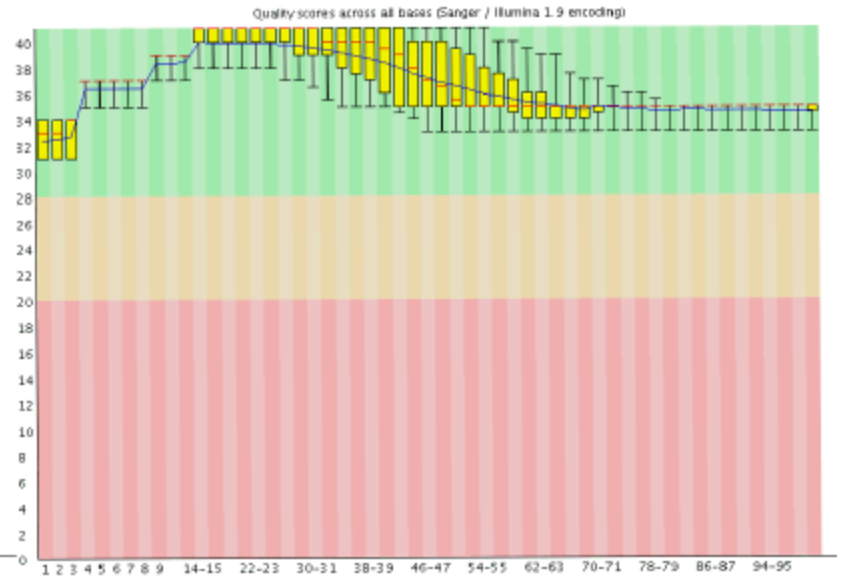
**Clip** adapter sequences

**Reverse-Complement**

**Trim sequences**

# Trimmomatic; trimming the reads



# SAM/BAM

- SAM
  - Aligned file
  - Different formats
- BAM
  - Binary version of SAM
- SAMtools can be used to view BAM file alignments

# Technologies

- PyroSequencing
  - Becoming very popular; accurating sequencing with extremely low error rates

# Technologies

- Illumina
- IonTorrent
  - Do not make use of optical signals; pH based;
  - Higher error rate
- PacBio
  - Mean fragment length of sequence is about 15kb
  - Unlike Illumina, it can generate longer reads
  - More expensive than Illumina
- Nanopore

# Other Sequencing Methods

- Pyrosequencing
  - Alternative approach to Illumina
  - Uses a Bead based approach instead of a “flow cell”
- Sequencing by Ligation
  1. Library preparation
  2. Bead coupling → PCR amplification
  3. Bead deposition on glass slide
  4. Sequencing by ligation
  5. Data Analysis (repeat steps 1 to 5 many times)

# IonTorrent PGM, Proton

Designed to more specialized clinical applications

- Up to 400 bp long reads
- Up to 12 GB per run

# Illumina MiniSeq, MiSeq, NextSeq HiSeq

- Illumina is the current leader in HTS
- Illumina currently offers sequencers that cover the full range of data
- output. More details are available on the Illumina sequencers page.
- Up to 300 million reads (HiSeq 2500)
- Up to 1500 GB per run (GB = 1 billion bases)

# MinIon

A portable, miniaturized device that is not yet quite as robust and reliable compared to other relevant platforms

- More details on the MinION sequencers page.

- Up to 10,000 long reads
- Up to 240 MB per run (MB = 1 million bases)

# PacBio

This company is the leader in long-read sequencing:

- Up to 12,000 bp long paired-end reads
- Up to 4 GB per run
- Details on PacBio page

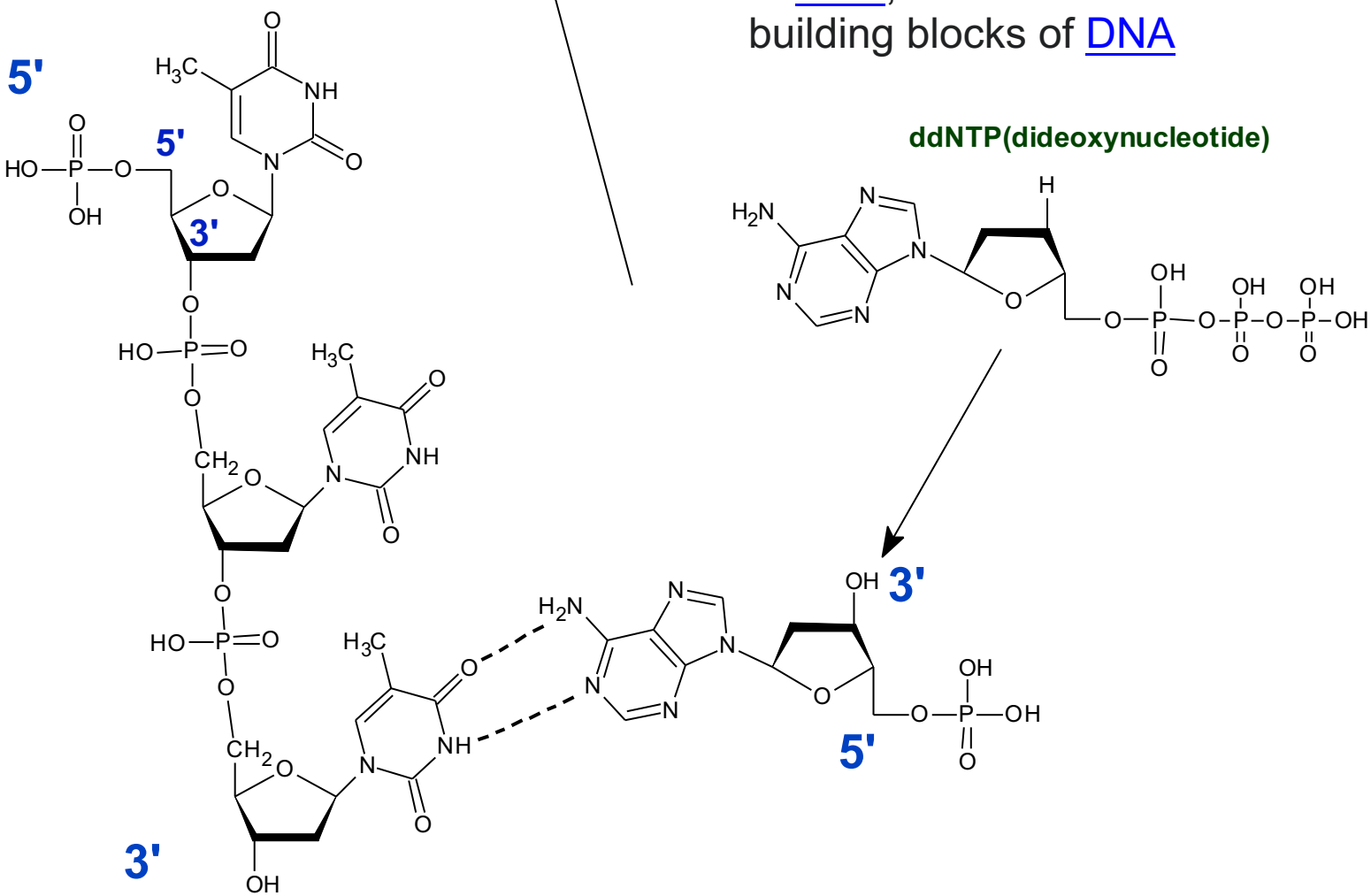
# Sanger Sequencing

# Basic Steps

- Dideoxynucleotide sequencing
- Obtain a template (DNA)
  - Denature to get a single strand DNA
- Primer (20 nts complementary to the strand being sequenced)
- DNA Polymerase
- 4 dNTPs

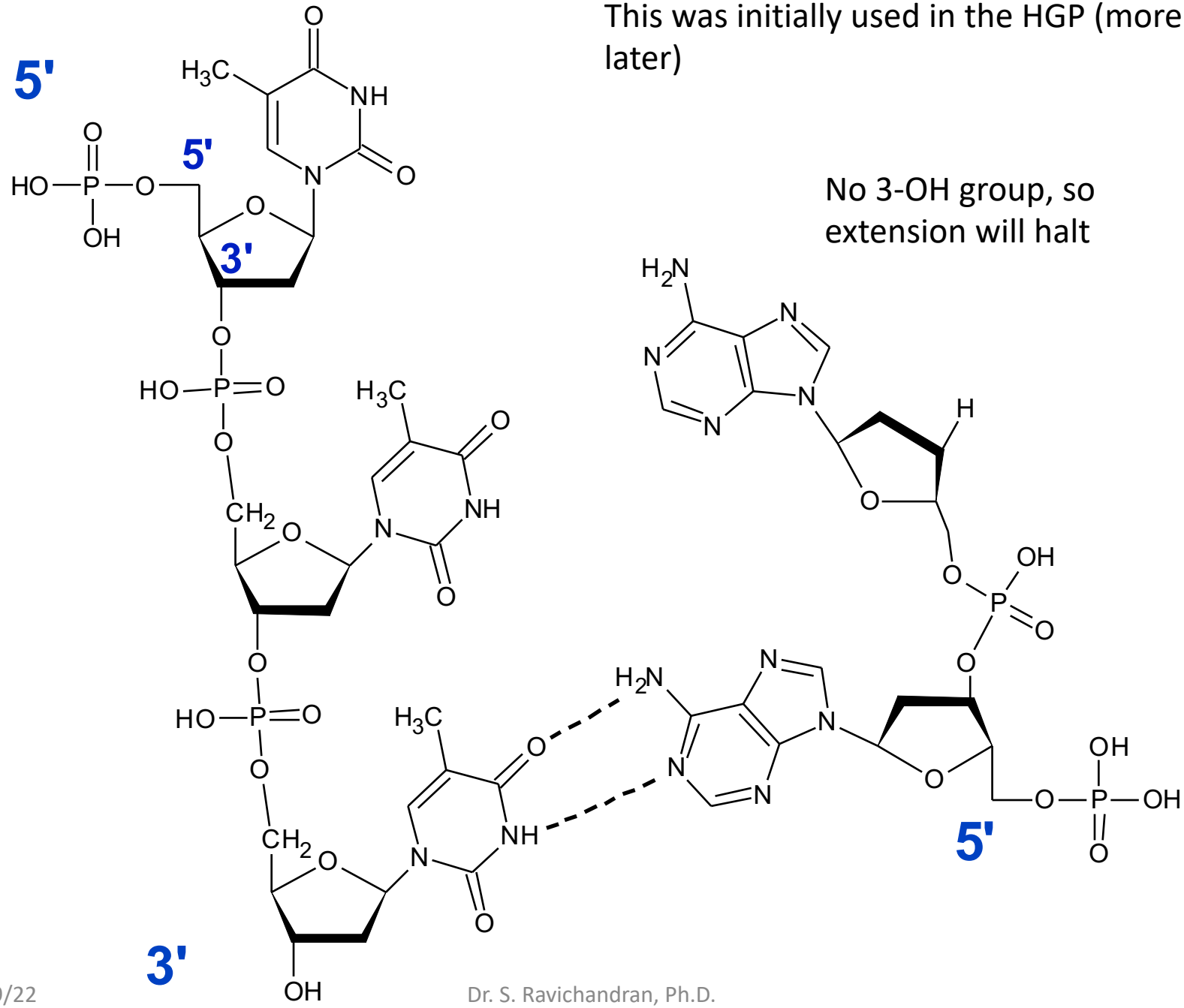
2′-deoxynucleotides (dNTPs) for extension, For example, 2′-deoxythymidine triphosphate
dideoxynucleotide (ddNTP) for inhibition. For example, 2′,3′-dideoxythymidine triphosphate

NTPs are the building blocks of RNA, and dNTPs are the building blocks of DNA



This was initially used in the HGP (more on this later)

No 3-OH group, so extension will halt

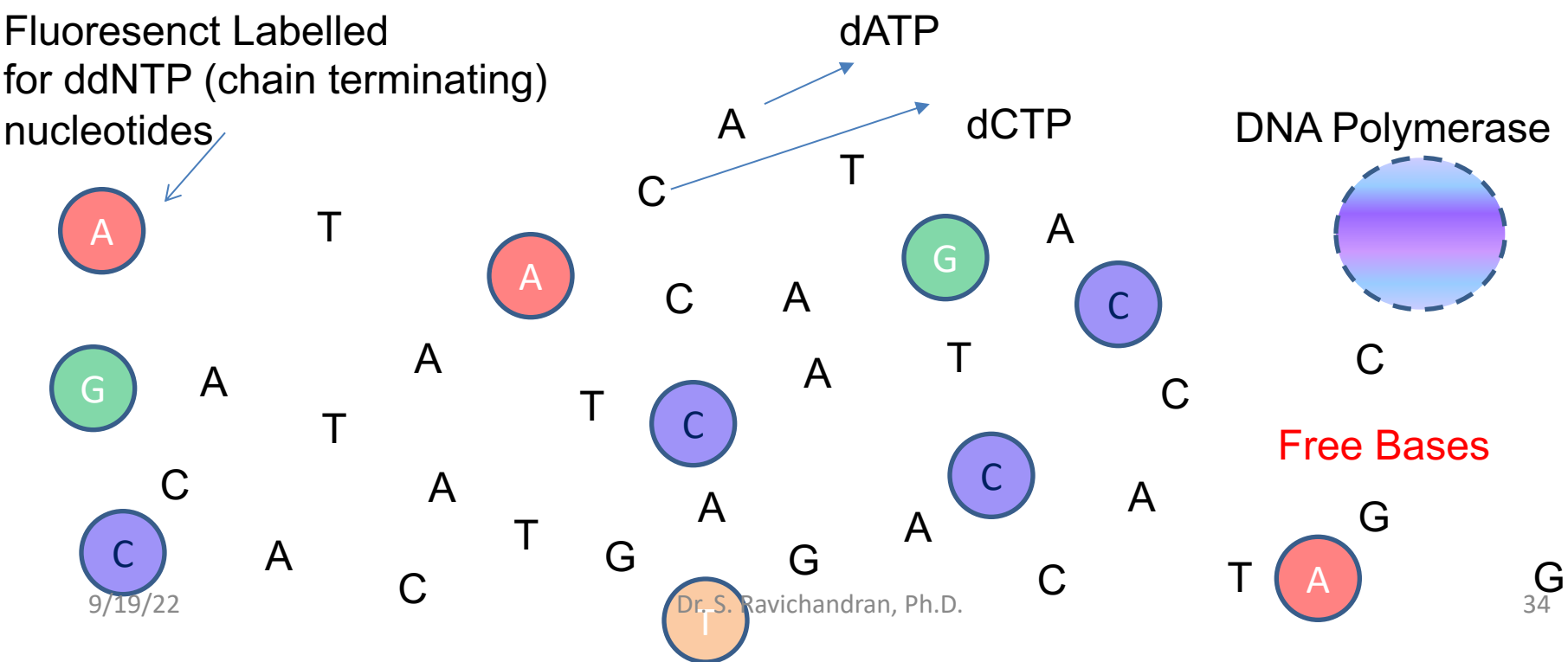


# Sanger Sequencing
# Pre next-gen or First-gen sequencing

- Sequencing Reaction

**DNA chain grow only from 3' end**

```
5'   GTACGG ..3'   Primer
3'   CATGCCATCGGGGCATGG ..5'  Template
```



# ddCTP focused reaction

**3′- CATGCCATCGGGGCATGG -5′ Template**

**5′- GTACGG -3′** Primer

DNA chain grow only from 3' end

**5′- GTACGGTAGC -3′**

**5′- GTACGGTAGCC -3′**

**5′- GTACGGTAGCCC -3′**

**5′- GTACGGTAGCCCCGTAC -3′**

**5′- GTACGGTAGCCCCGTACC -3′**

If you are using **ddCTP** dideoxyribonucleic nucleotide triphosphate

These are chain-terminating nucleotides. Once added they cannot extend the chain

# Sequencing Reaction Products

GTACGGT
GTACGGTA
GTACGGTAG
GTACGGTAGC
GTACGGTAGCC
GTACGGTAGCCC

PRIMER 5′ GTACGG 3′

dNTPs

A T G C

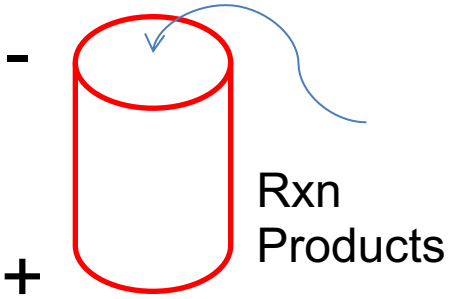
PLUS Free Bases

Handle 500-800 bases long

Sequencing runs out of steam roughly after about 800 bases.

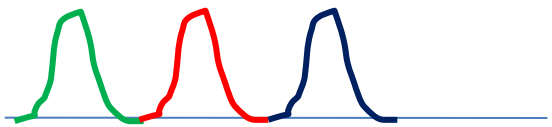
3′ -CATGCCATCGGGGCATGG- 5′

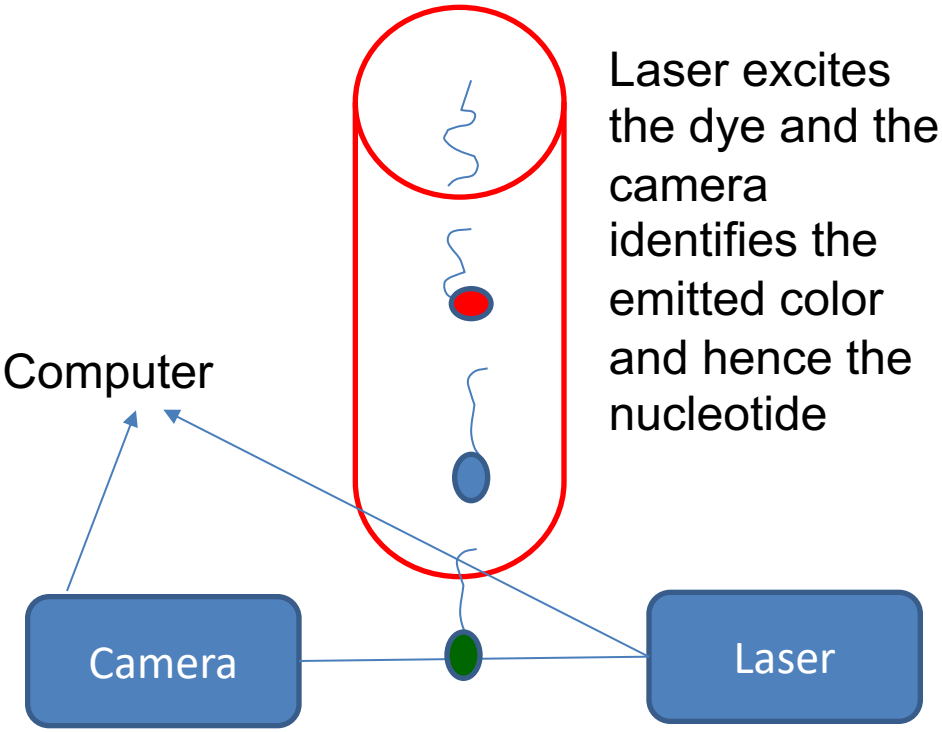
**DNA electrophoresis: Used for separation the fragments**



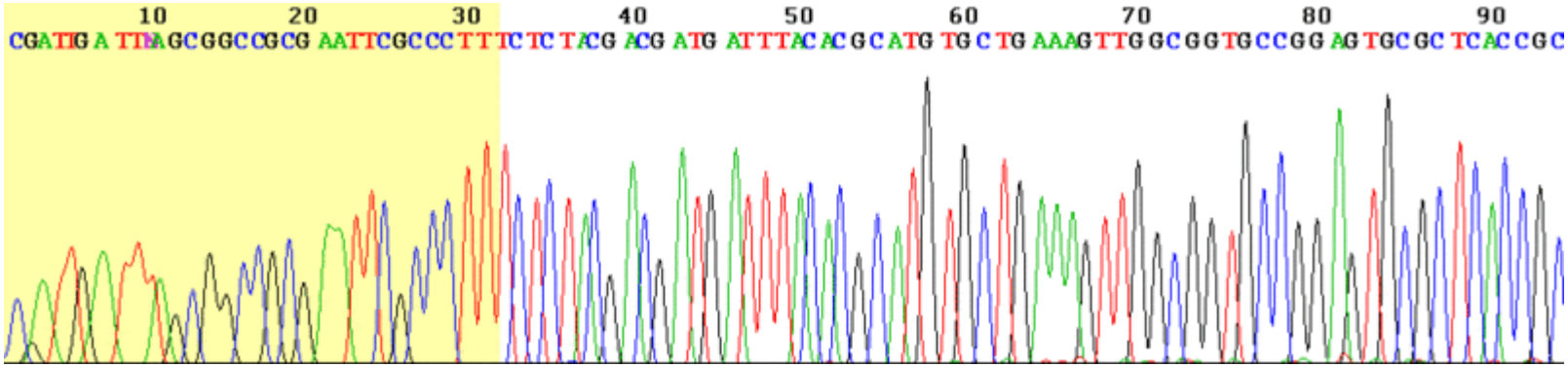
The reaction fragments (the fluorescent tagged fragments) are poured in a gel containing tube and electrophoresis is applied. The smallest fragment moves faster towards +ve end and read by the laser. The next larger fragment will follow it.

THERE ARE NEWER VARIANTS BUT THE CONCEPT IS THE SAME






Laser excites the dye and the camera identifies the emitted color and hence the nucleotide




Author: Loris; http://en.wikipedia.org/wiki/File:Sanger_sequencing_read_display.gif

Chromatogram

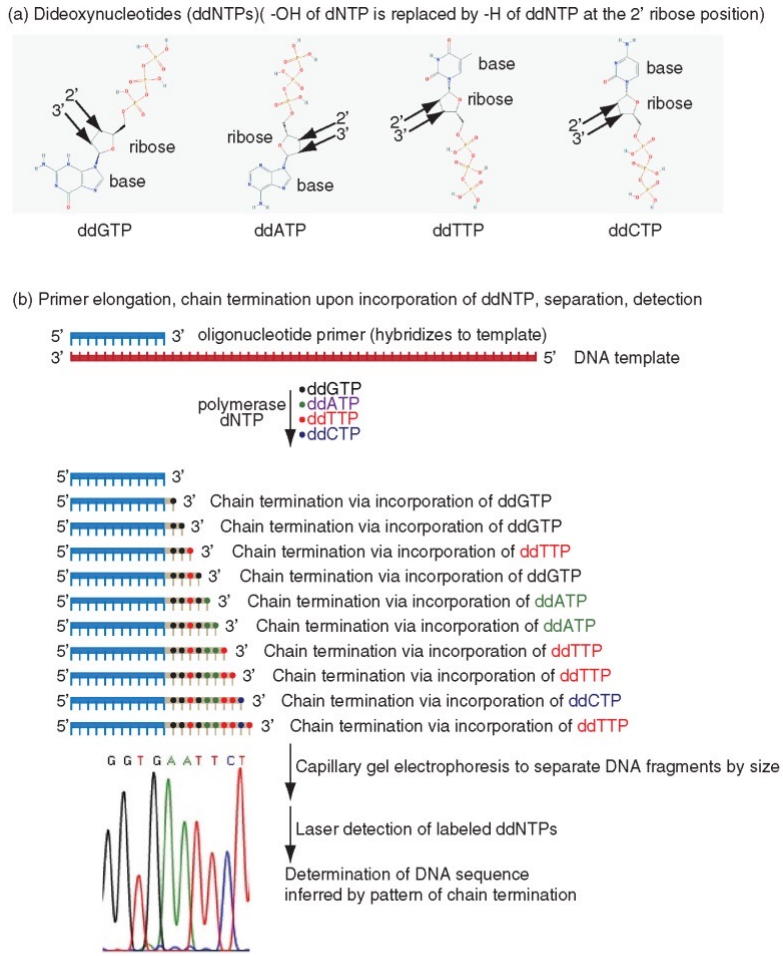



**FIGURE 9.1** DNA sequencing by the Sanger method. (a) Structures of the four modified dideoxynucleotide (ddNTP) bases: 2′-3′-dideoxyguanosine 5′-triphosphate (ddGTP), 2′,3′-dideoxyadenosine-5′-triphosphate (ddATP), 2′,3′-dideoxythymidine 5′-triphosphate (ddTTP), and 2′,3′-dideoxycytidine 5′-triphosphate (ddCTP). The 2′ and 3′ ribose positions have hydrogen atoms in the ddNTPs, while they have a 3′ hydroxyl in DNA. (b) An oligonucleotide primer (in blue) (e.g., a 22-mer or synthetic nucleic acid of length 22 nucleotides) is hybridized to a single-stranded template (red) then extended using a DNA polymerase in the presence of dNTPs and a limited amount of one of the four ddNTPs. Chain termination occurs at one of the sites containing the ddNTP. The resulting synthesized fragments can be separated using a method such as capillary electrophoresis, and the products can be detected to infer the DNA sequence (bottom). The sequence in this example (GGTGAATTCT) corresponds to beta globin (**Fig. 9.2**). Structures are from the NIH PubChem Open Chemistry Database at NCBI (⊕ http://pubchem.ncbi.nlm.nih.gov/; compounds 446577, 65304, 65051, and 119119).

*Bioinformatics and Functional Genomics*, Third Edition, Jonathan Pevsner.

Companion Website: www.wiley.com/go/pevsnerbioinformatics

# Sanger Sequencing

- 1979-2003 Dominant sequencing method
- Can produce high quality reads
  - Error rate < 1% /base
- Still used in large sequencing centers
- Currently, can read 800 or more bases in one reaction

# Microarray

- How do we know whether a gene is turned on or off?
  - By measuring the mRNA
- Example
  - Muscle cell will express muscle related proteins
    - Actin, myosin and not Insulin (Hormone) or Melanin (Pigment)
- One could collect the expression profiles of two different cells
  - If we compare we will know what is expressed and where?

# Microarray

- Molecular basis for phenotypic differences
- Questions?
  - Why we are all different?
  - Why cancer cells live longer?
- We know the differences are due to genomic differences
- Can we identify variations and relate them to phenotypic differences?

# Microarray

- SNP
- Genotypes
  - AA
  - AC
  - CC

**C**

**GGATCGAGTNTTAAGCCTA**

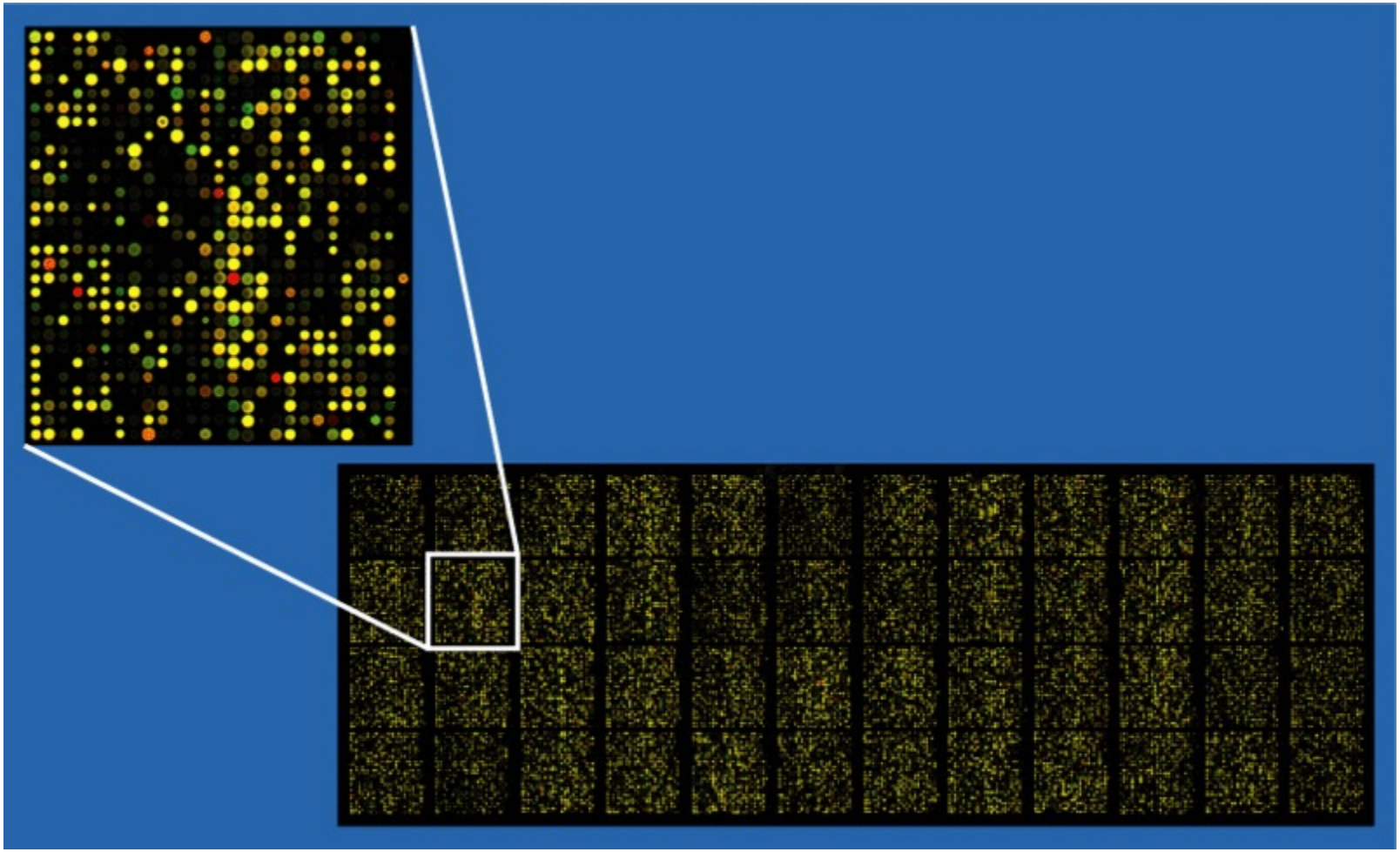
**A**

# Microarray

- Method to count molecules
- Steps
  - Denaturation
  - Hybridization
  - Convert intensity to number
- Important
  - Works only when you have many many copies of molecules

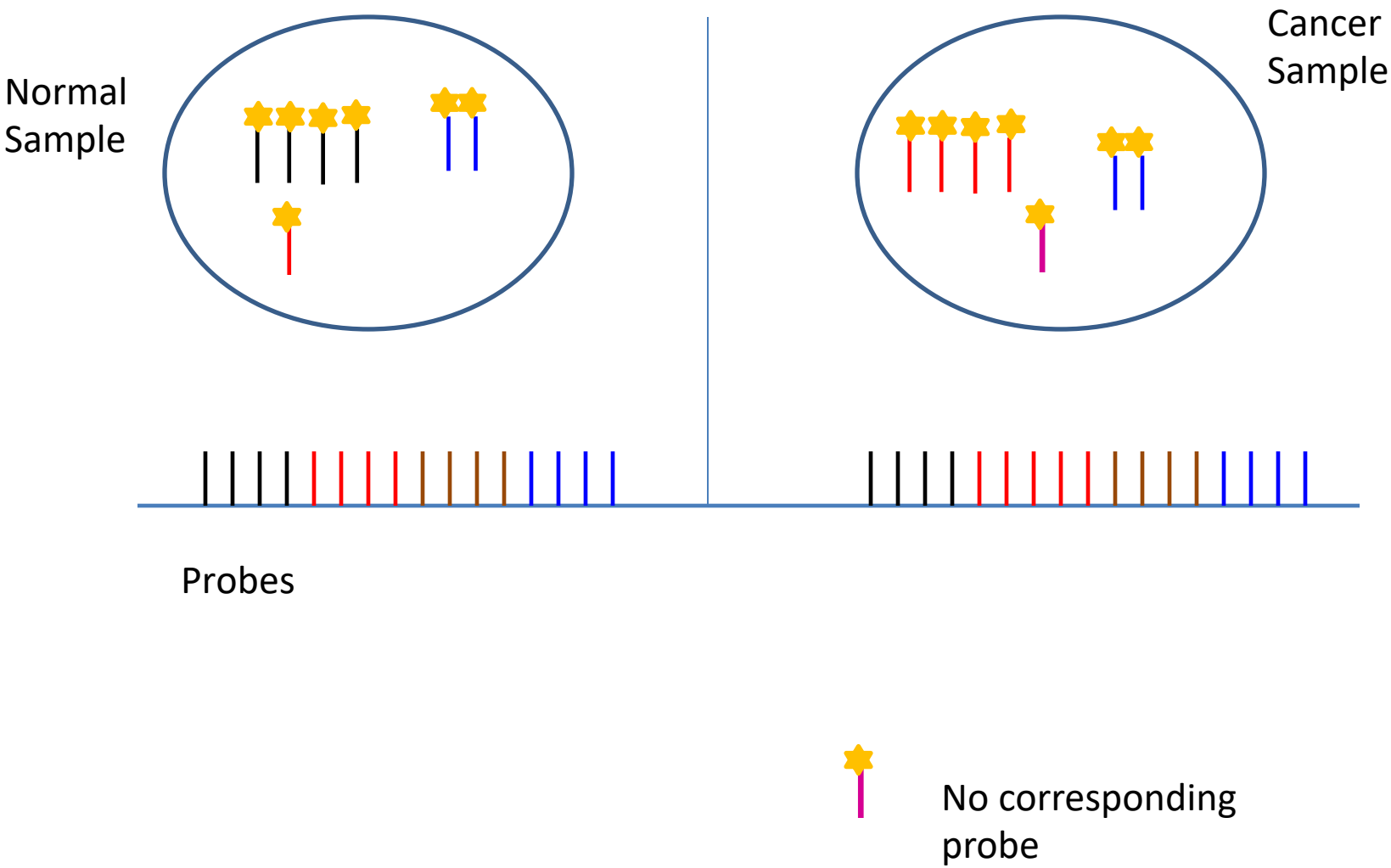
# Microarray

- Different microarray methods
  - Density
  - one or two labels
- Platforms
  - Affymetrix (high density; one color)
  - Agilent (circles on grid, 1 or 2 color)
  - Illumina (high density, 1 or 2 color)

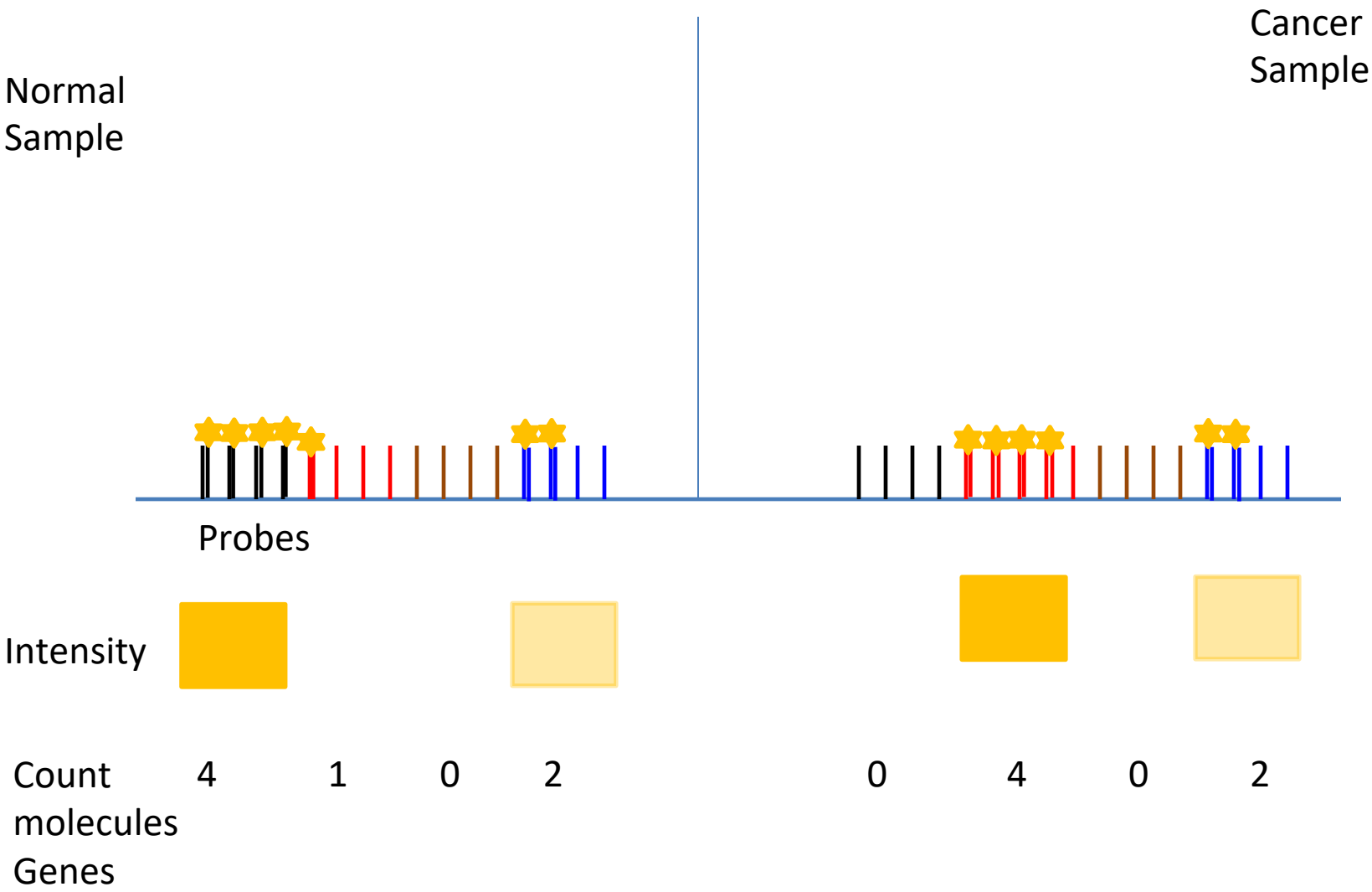


Example of an approximately 37,500 probe spotted oligo microarray with enlarged inset to show detail

https://commons.wikimedia.org/wiki/File:Microarray2.gif

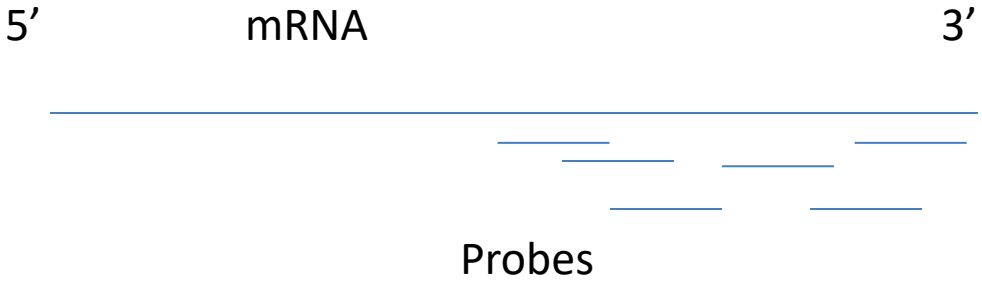



- Pour and Wash
- Measure the intensity
- Compare the normal vs cancer samples


Normal Sample
Cancer Sample
Probes
Intensity
Count molecules Genes
4
1
0
2
0
4
0
2

# Microarray Applications

- Gene expression
  - Why 3'
  - Target the slow decaying end
- SNP
  - Person will have **AA, AG, GG**




**G**

**GGATCGAGTNTTAAGCCTA**

**A**

**Probe for allele-1**

**CCTAGCTCACAATTCGGAT**

**Probe for allele-2**

**CCTAGCTCATAATTCGGAT**

# Microarray Applications

- CHIP Microarray
  - Transcription factor binding sites
    - Proteins bind to DNA
    - Identifying the sites

# Compare Sanger with NGS Sequencing

# NGS

- WGS
  - A name applied to New & Powerful sequencing technologies developed in the last 10 years
  - Popular platforms are PacBio, Illumina, IonTorrent
- Early days of WGS
  - 35–50 base pairs,
- As of Recent days
  - hundreds of base pairs.
    - Note PacBio can handle thousands of base pairs.
    - This can be extraordinarily important in resolving duplicated regions and in genome assembl

# NGS

- WGS
  - sequencing reads
    - Millions to Billions
    - Illumina can produce roughly 1TB or more of data per run (HiSeq technology)
  - The time required for a run
    - Hours to days.
  - The cost HGP
    - US$ 1–3 billion over a 15 year period,
  - Today can be whole-genome sequence < $2K
  - Each technology introduces different, characteristic types of errors that influence the variants that are called at the end of the data analysis pipeline.

# Table 9.1 from the Jonathan Pevsner's Book

**TABLE 9.1 Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, http://en.wikipedia.org/wiki/DNA_sequencer, and literature cited for each technology.**

| Technology | Read length (bp) | Reads per run | Time per run | Cost per megabase (US$) | Accuracy (%) |
|---|---|---|---|---|---|
| Roche 454 | 700 | 1 million | 1 day | 10 | 99.90 |
| Illumina | 50–250 | <3 billion | 1–10 days | ~0.10 | 98 |
| SOLiD | 50 | ~1.4 billion | 7–14 days | 0.13 | 99.90 |
| Ion Torrent | 200 | <5 million | 2 hours | 1 | 98 |
| Pacific Biosciences | 2900 | <75,000 | <2 hours | 2 | 99 |
| Sanger | 400–900 | N/A | <3 hours | 2400 | 99.90 |

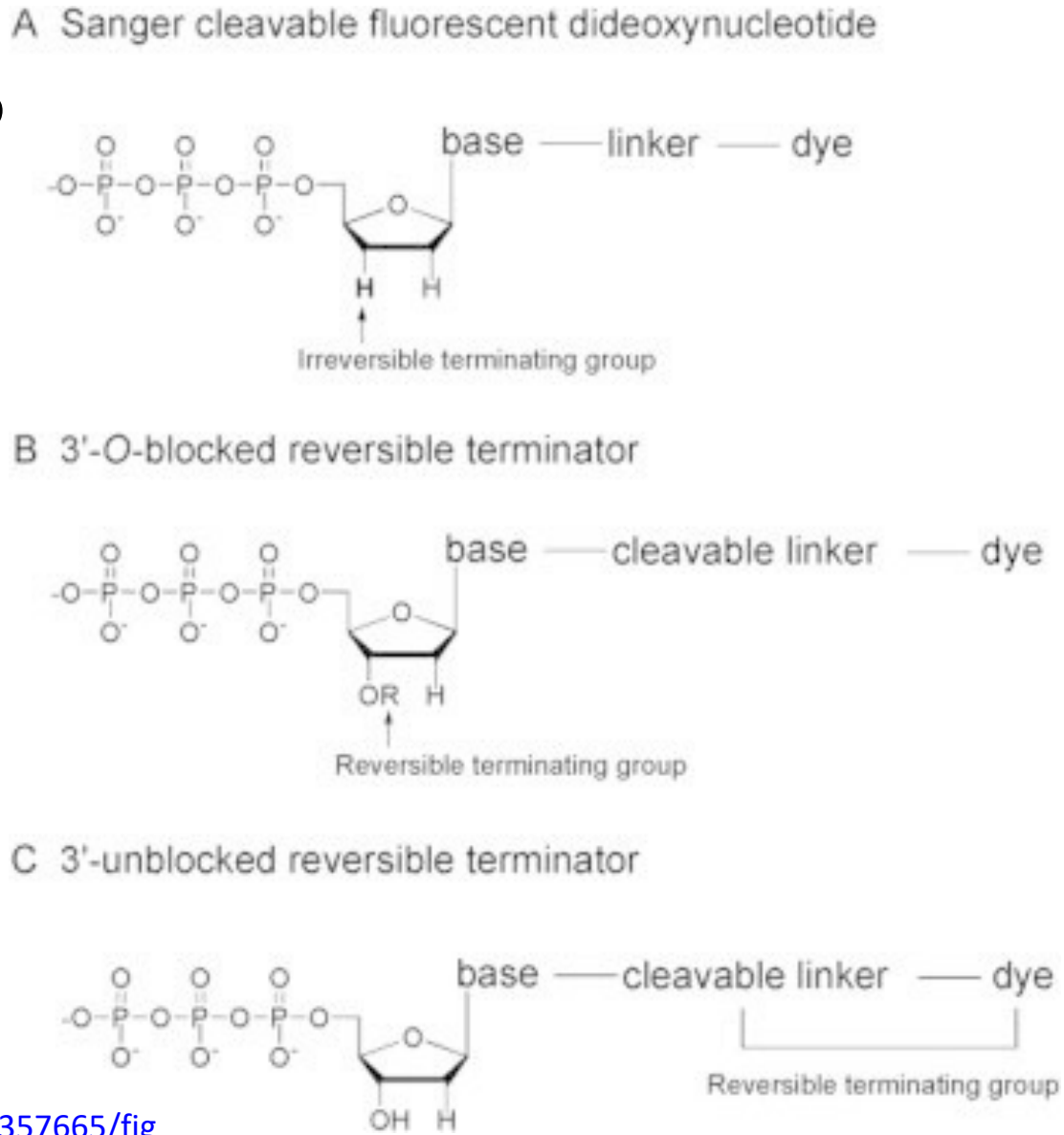
https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8

# Watch the Video

# Next Gen Seq. technology: Applications

- 1000 Genomes Project
- HEP
- How do go from sequence reads to applications?
  - Step1 answers the question for each read,
    - where does this read came from?

# Sanver vs NGS




https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4357665/figure/f0005/

# Reads matched to Reference

Reads
(NTs are not shown)

Mapping to
Reference Genome

Reference

CTCTGGTCGAGGGTCTAAATCGCCCCTGTGTACCCGGGTAGAGAGAGACCAG

# NGS Application: Variant Detection

READS

READS

READS

Align to the Reference

```
GTCGCAGTANCTGTCT
|||||||||||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||||||||||||
GGATCT-CGATATACC
```

Pile-up:
The collection of NTs from each read at a particular location

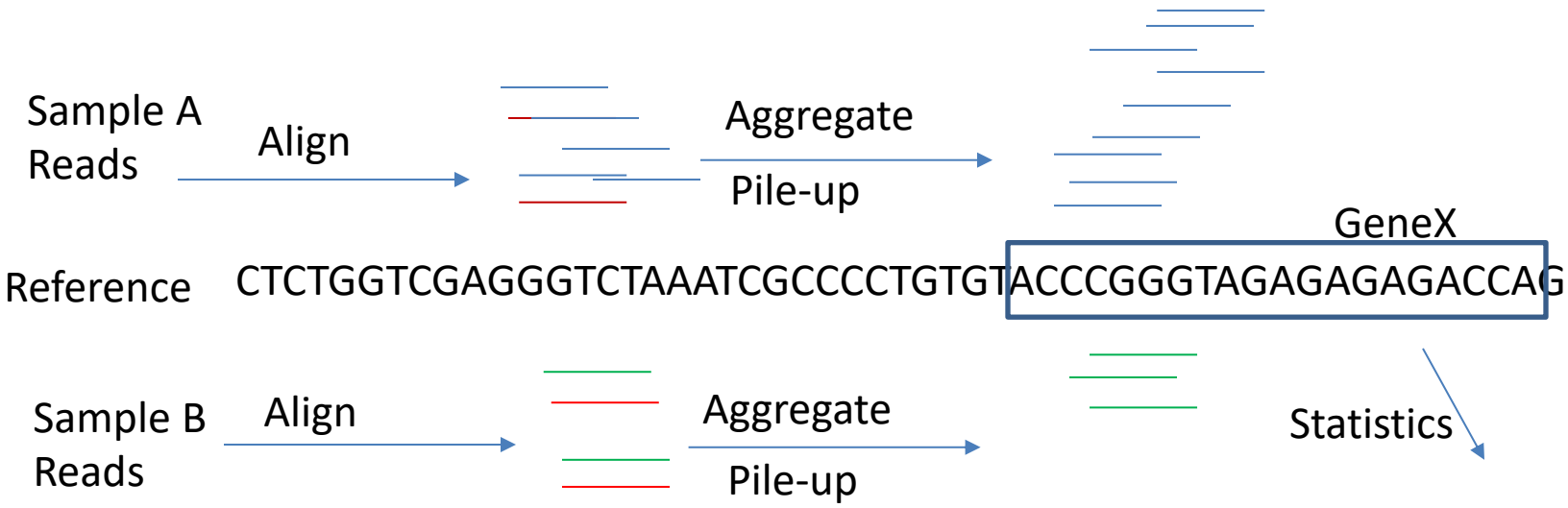
Aggregate/Pile-up

```
TCTCTCCCANNAGAGC
|||||||||||||||
TCTCTCCCAGGAGAGC
```

Statistics

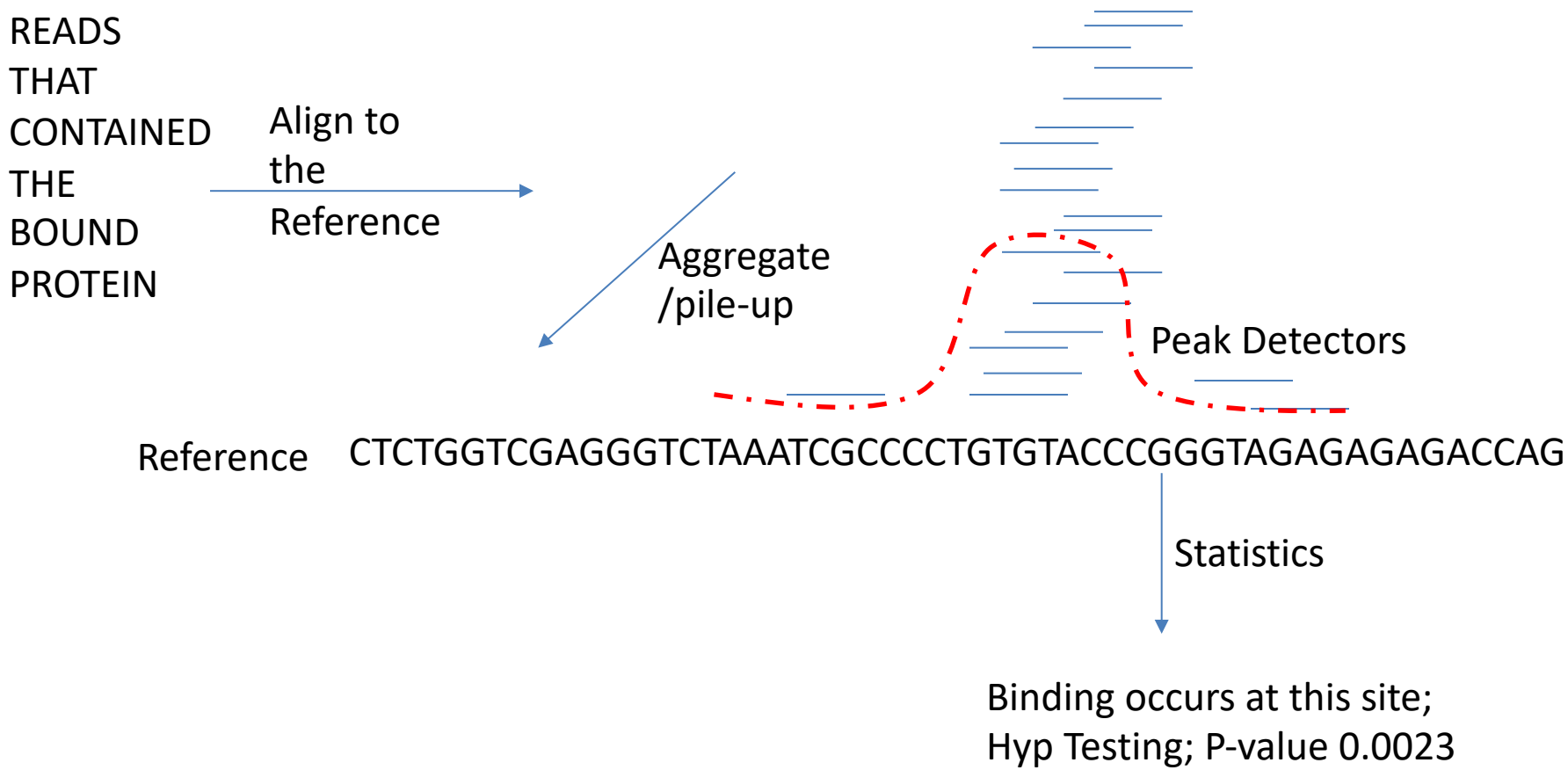
# NGS Application: RNA-seq differential expression (*RNA not DNA*)

Take RNA→ cDNA and sequence them




GeneX is expressed more in Sample A than B

# NGS Application: ChIP-seq



# Analysis of NGS of Genomic DNA

# Commonly used workflow is the Genome Analysis Toolkit (GATK)

| Stage | Examples/explanation | File formats |
|---|---|---|
| Laboratory work | Experimental design<br>Library preparation<br>Enrichment (capture) | |
| Next-generation sequencing | Platforms include Illumina, SOLiD, Pacific Biosciences, other | Output: FASTQ-Sanger, FASTQ-Illumina |
| Analysis pipeline: Quality assessment | Trimming, filtering<br>Software: FastQC | FASTQ |
| Analysis pipeline: Alignment to reference genome | Software: BWA, Bowtie2 | Reference: FASTA<br>Output: SAM/BAM |
| Analysis pipeline: Variant identification | Single nucleotide variants (SNVs), structural variants (e.g. indels)<br>Software: GATK, SAMTools<br>Realignment, recalibration | Variant Call Format ( VCF/BCF) |
| Analysis pipeline: Annotation | Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores | |
| Visualization | Variant visualization; read depth; comparison to other samples<br>Software: IGV, BEDTools, BigBED | |
| Prioritization | Discovery of relevant variants<br>Software: PolyPhen-2, VEP, VAAST | VCF |
| Storage | Deposit data in ENA, SRA, dbGaP | BAM, VCF |

FASTQ-Sanger/
FASTQ-Illumina

SAM/BAM

VCF/BCF

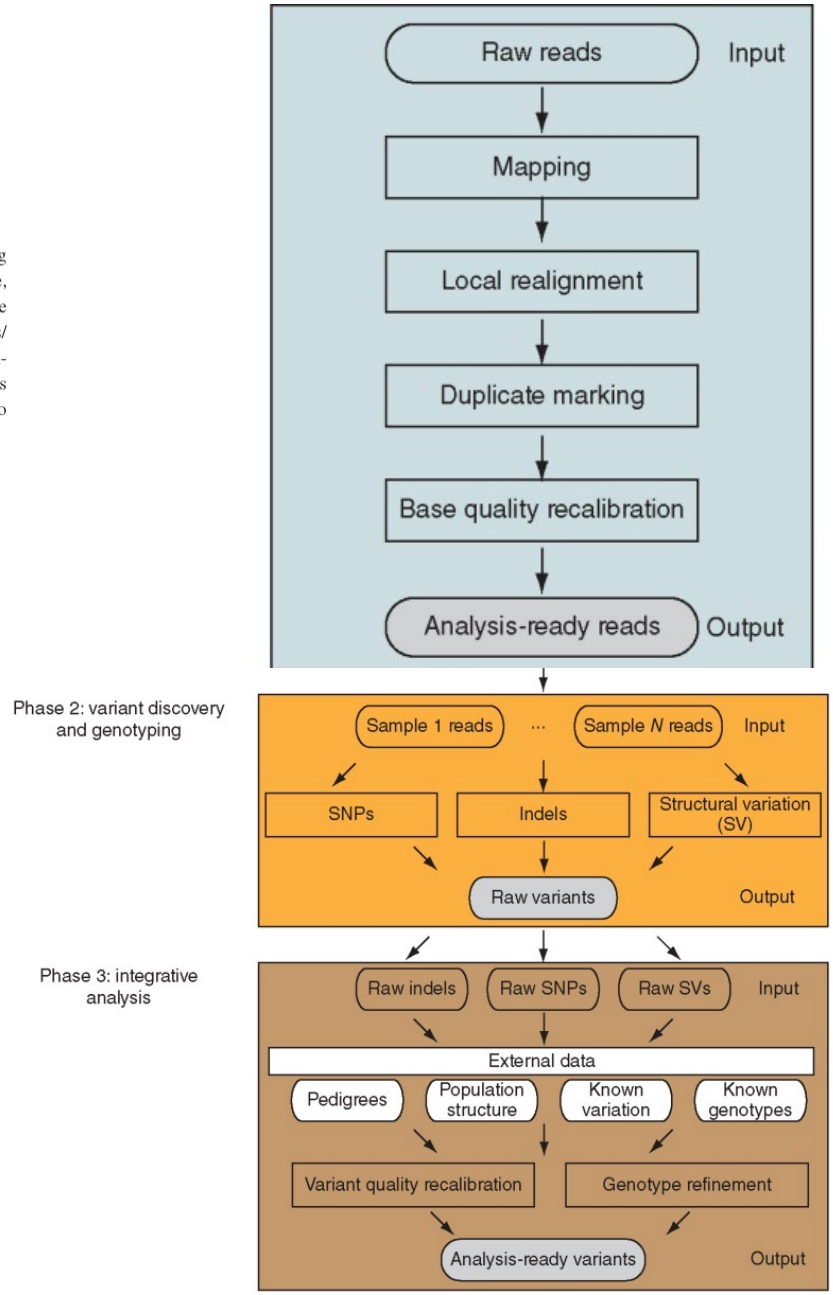
**FIGURE 9.6** Workflow for next-generation sequence experiments: from experimental design to data analysis. We describe software tools and data formats in this chapter.

# GATK: Genome Analysis Toolkit state-of-the-art workflow

**FIGURE 9.7** Workflow for variant discovery and genotyping from next-generation DNA sequencing using GATK. In the first phase, raw reads (in the FASTQ format) are mapped to a reference genome, realigned, duplicate reads are removed, and base quality scores are recalibrated. In the second phase variants are identified in the three categories of single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants (SVs). In the third phase, quality scores of variants are recalibrated and genotypes are refined in the context external data sources that inform the analyses. The steps introduced by GATK greatly reduce both false negative and false positive errors. Adapted from DePristo *et al.* (2011), with permission from Macmillan Publishers.

*Bioinformatics and Functional Genomics*, Third Edition, Jonathan Pevsner.

Companion Website: www.wiley.com/go/pevsnerbioinformatics



www.ncbi.nlm.nih.gov/sra?term=SRR390728

Full

Send to:

**SRX079566: RNA-Seq (polyA+) analysis of DLBCL cell line HS0798**
2 ILLUMINA (Illumina Genome Analyzer IIx) runs: 16.9M spots, 1.2G bases, 1.1Gb downloads

**Design:** PolyA+ RNA was purified using the MACS mRNA isolation kit (Miltenyi Biotec, Bergisch Gladbach, Germany), from 5-10ug of DNaseI-treated total RNA as per the manufacturer's instructions. Double-stranded cDNA was synthesized from the purified polyA+RNA using the Superscript Double-Stranded cDNA Synthesis kit (Invitrogen, Carlsbad, CA, USA) and random hexamer primers (Invitrogen) at a concentration of 5uM. The cDNA was fragmented by sonication and a paired-end sequencing library prepared following the Illumina paired-end library preparation protocol (Illumina, Hayward, CA, USA).

**Submitted by:** BC Cancer Agency Michael Smith Genome Sciences Centre (BCCAGSC)

**Study:** CGCI: Non-Hodgkin Lymphoma (NHL)
PRJNA172563 • SRP020237 • All experiments • All runs
show Abstract

**Sample:** human B cell lymphoma cell line DB
SAMN00630374 • SRS212581 • All experiments • All runs
*Organism:* Homo sapiens

**Library:**
*Name:* HS0798
*Instrument:* Illumina Genome Analyzer IIx
*Strategy:* RNA-Seq
*Source:* TRANSCRIPTOMIC
*Selection:* cDNA
*Layout:* PAIRED

**Spot descriptor:**
1 forward 37 reverse

**Pipeline:** show...

**Runs:** 2 runs, 16.9M spots, 1.2G bases, 1.1Gb

| Run | # of Spots | # of Bases | Size | Published |
|---|---|---|---|---|
| SRR292241 | 9,721,384 | 699.9M | 956.3Mb | 2011-06-24 |
| SRR390728 | 7,178,576 | 516.9M | 184.6Mb | 2011-12-21 |

ID: 90134

Runs: SRR*
Study: SRP*
Study Acc.: SRX

**Related information**
BioProject
BioSample
dbGaP
Taxonomy

**Search details**
SRR390728[All Fields]
Search
See more...

**Recent activity**
Turn Off Clear
SRR390728 (1) SRA
SAMN00630374 (1) SRA
(SRR390728) NOT cluster_dbgap[PROP] (1) SRA
Submission Quick Start Guide - SRA Handbook
SRS212581 (1) SRA
See more...

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)
Write to the Help Desk

sratoolkit.2.6.2-win64.zip | WebDocument_9-02_....gz Canceled | WebDocument_9-01_....gz Canceled | SIFT4G_Annotator_v2.4.zip Canceled | seqcost2015_4.xlsx | Show all downloads...

# FastQ from NCBI SRA using toolkit (Windows)

```
C:\Users\Ravi\Downloads\sratoolkit.2.6.2-win64\Data>.\..\bin\fastq-dump.exe -X 3 -Z SRR390728
Read 3 spots for SRR390728
Written 3 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
@SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACTTATTTATTATTATTATTTTGAGACAGAGCATTGGTC
+SRR390728.3 3 length=72
-;;;8;;;;;;;,*;;';-4,44;,:&,1,4'./&19;;;;;;669;;99;;;;;-;3;2;0;+;7442&2/

C:\Users\Ravi\Downloads\sratoolkit.2.6.2-win64\Data>
```

**-X multiple_runs ; in this case we are asking for 3**

**-Z Spots: In this case, we want to see the first 3 spots**

**Spot is a definite region in the flow cell (illumine)**

# Extracting Fasta file

Fasta format
with
50 columns

```
C:\Users\Ravi\Downloads\sratoolkit.2.6.2-win64\Data>.\..\bin\fastq-dump.exe -X 3 -Z SRR390728 -fasta 50
Read 3 spots for SRR390728
Written 3 spots for SRR390728
>SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACT
TTGACAAGCTGAGAGAAGNTNC
>SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCAC
TCCCTGGGTGGGGGGACTGGGT
>SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACTTATTTATTATTATT
ATTTTGAGACAGAGCATTGGTC

C:\Users\Ravi\Downloads\sratoolkit.2.6.2-win64\Data>
```

**-X multiple_runs ; in this case we are asking for 3**
**-Z Spots: In this case, we want to see the first 3 spots**
**Spot is a definite region in the flow cell (illumine)**
**-fasta 50 : 50 characters/line**

You can look for sequences and take it directly to Galaxy for FastQC or further analysis

# Other sources for Archive (ENA)