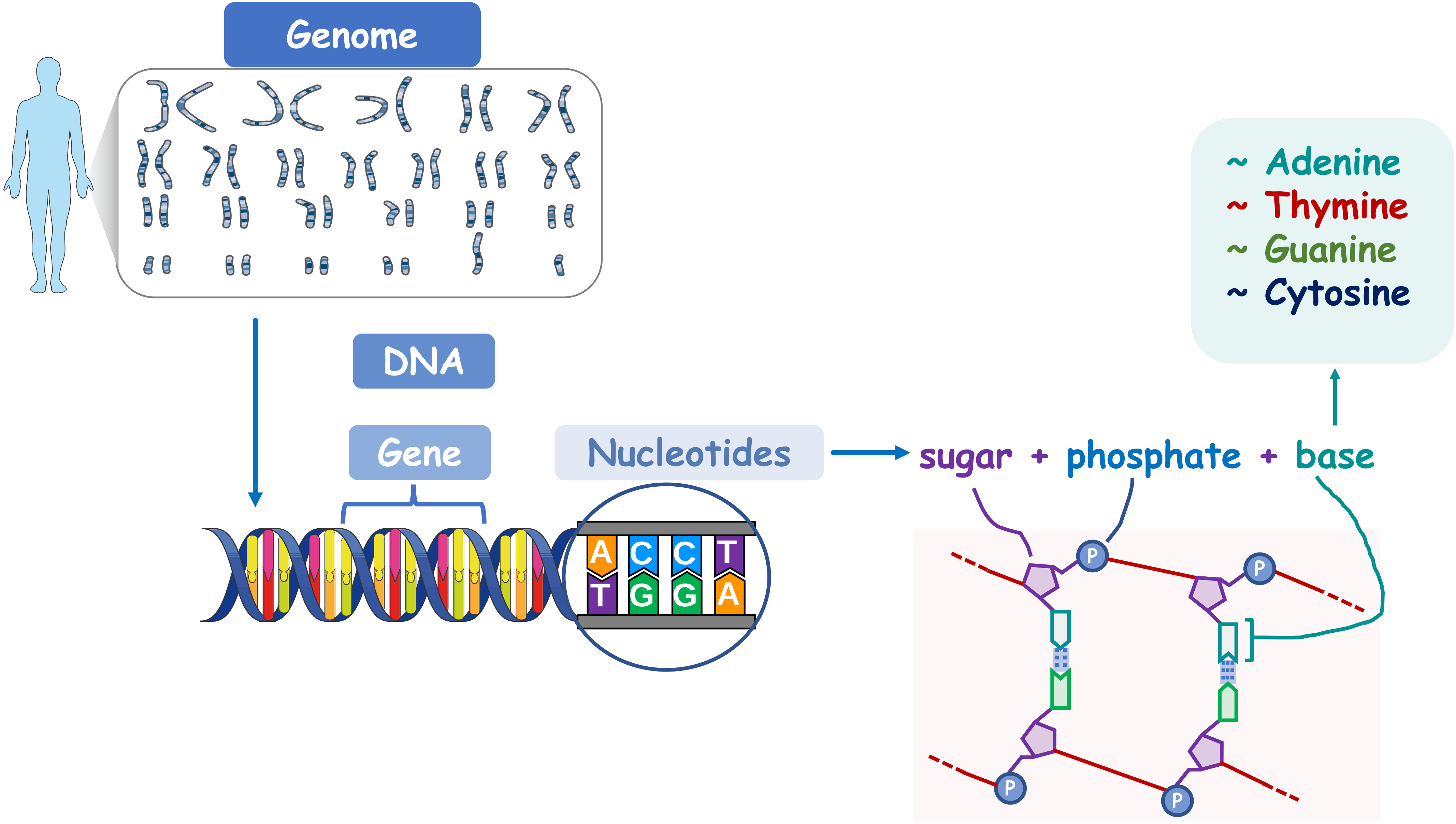
组学与大数据分析：
群体遗传变异数据的测序与质控

盛 欣

良渚实验室

Email: shengxin@zju.edu.cn

Looks for variation in genome



Genome

DNA

Gene

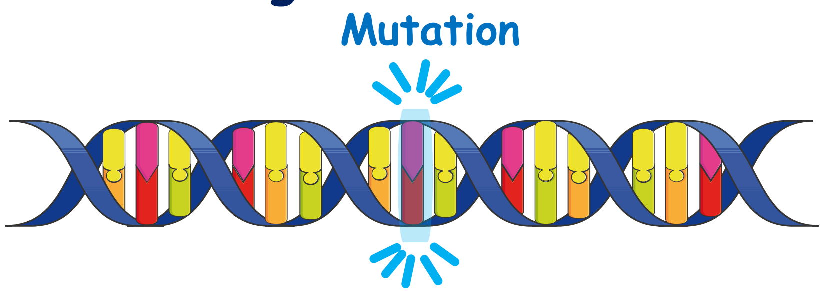
Nucleotides

sugar + phosphate + base

~ Adenine
~ Thymine
~ Guanine
~ Cytosine

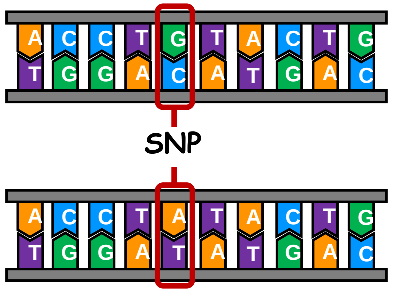
Human Genetic Diversity

Slight differences in DNA sequences

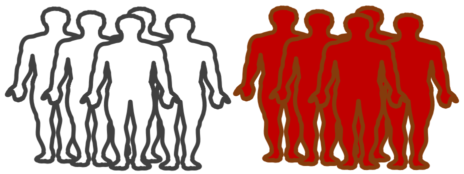


Genetic Variants

e.g. SNP- Single Nucleotide Polymorphism



Variation in traits & susceptibility to diseases



Variants

Naming

e.g. 185delAG variant in *BRCA1* gene

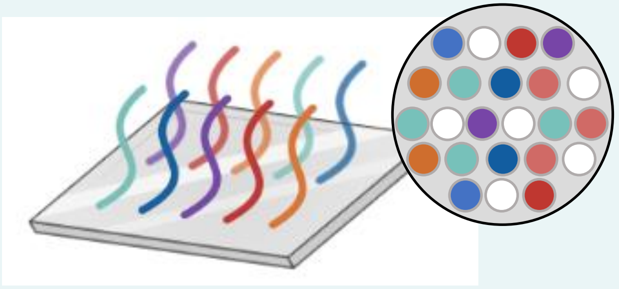
e.g. factor V Leiden variant in *F5* gene

e.g. Δ F508 variant in *CFTR* gene

Detection

Genotyping

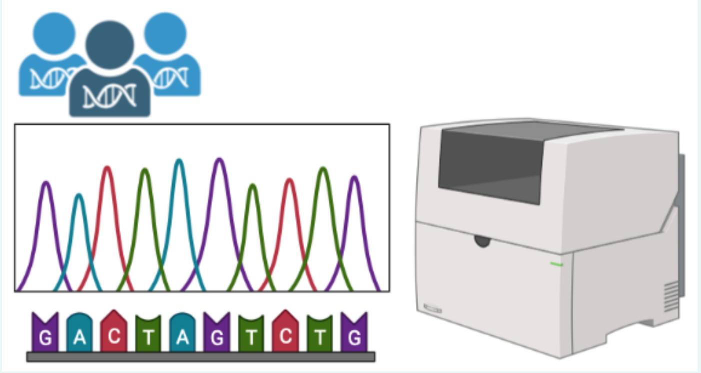
probes to detect
variants of interest



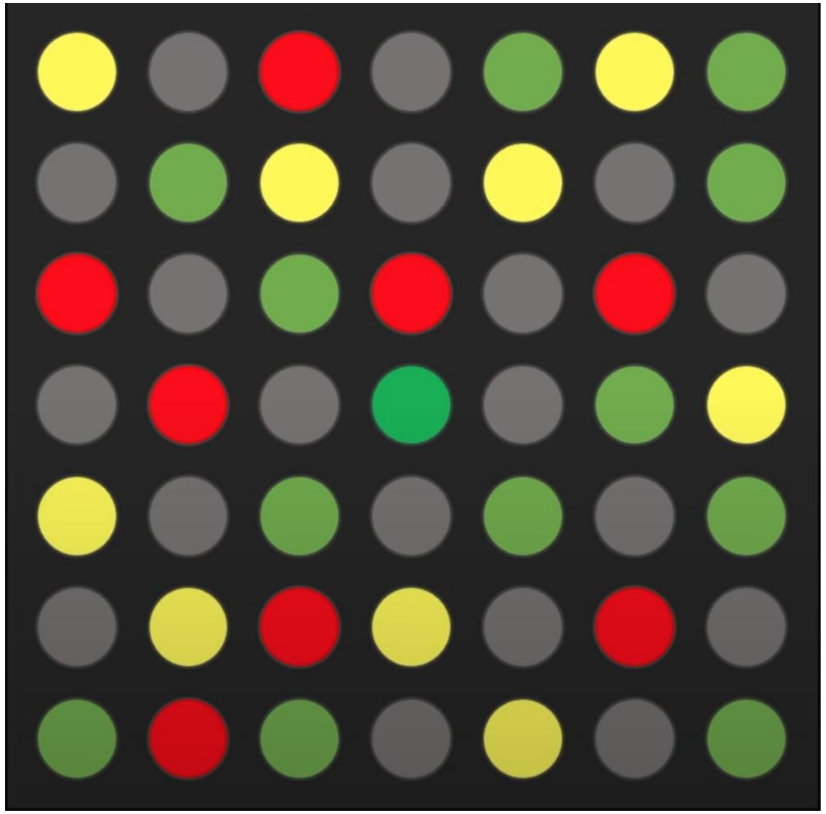
DNA microarrays

DNA Sequencing

exact sequence of continuous DNA



DNA Microarray



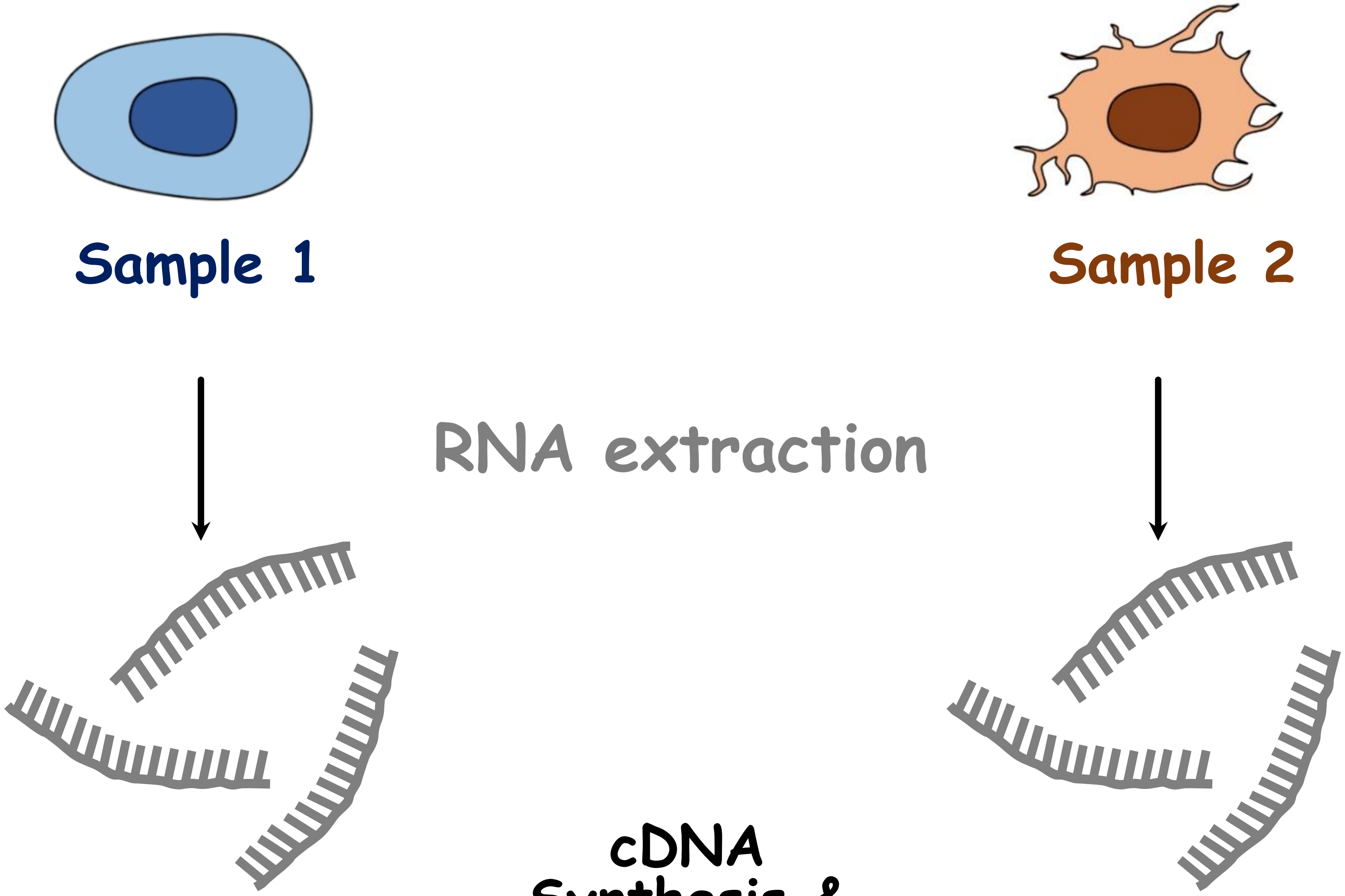
~ Gene Expression

~ Genotyping

*Note:

Different types of
DNA microarrays!

Microarray



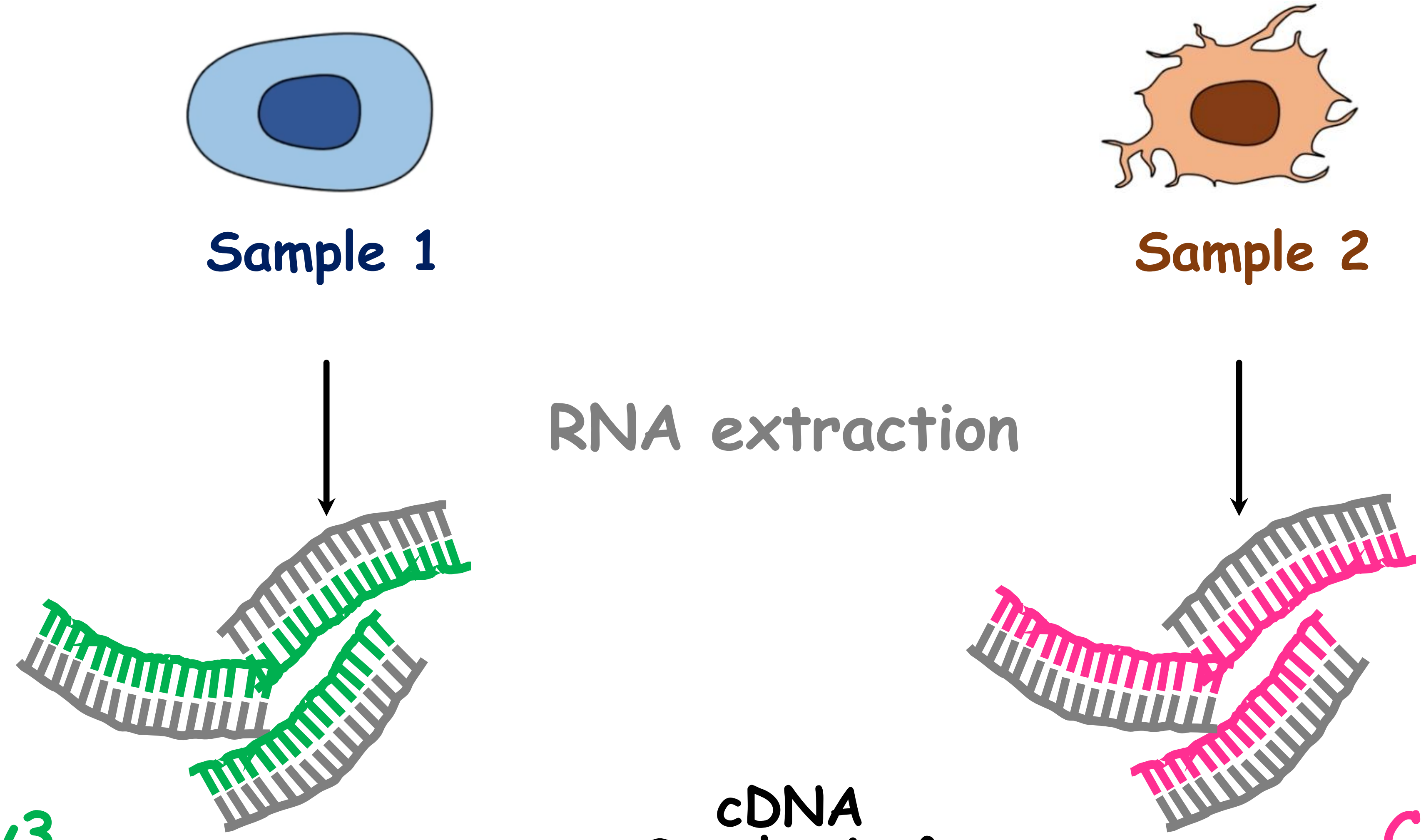
Sample 1

Sample 2

RNA extraction

cDNA
Synthesis &
Fluorescence
Labelling

Microarray



Sample 1

Sample 2

RNA extraction

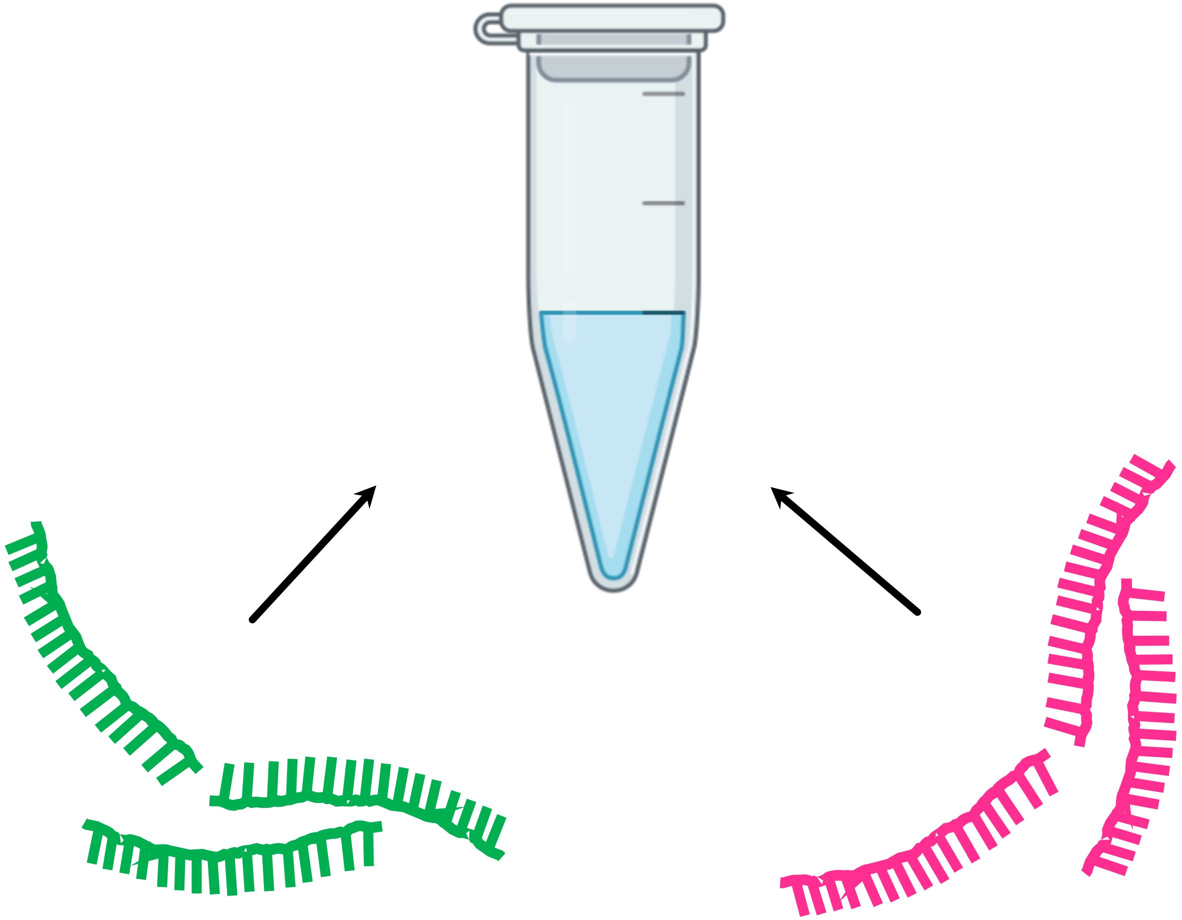
Cy3

Cy5

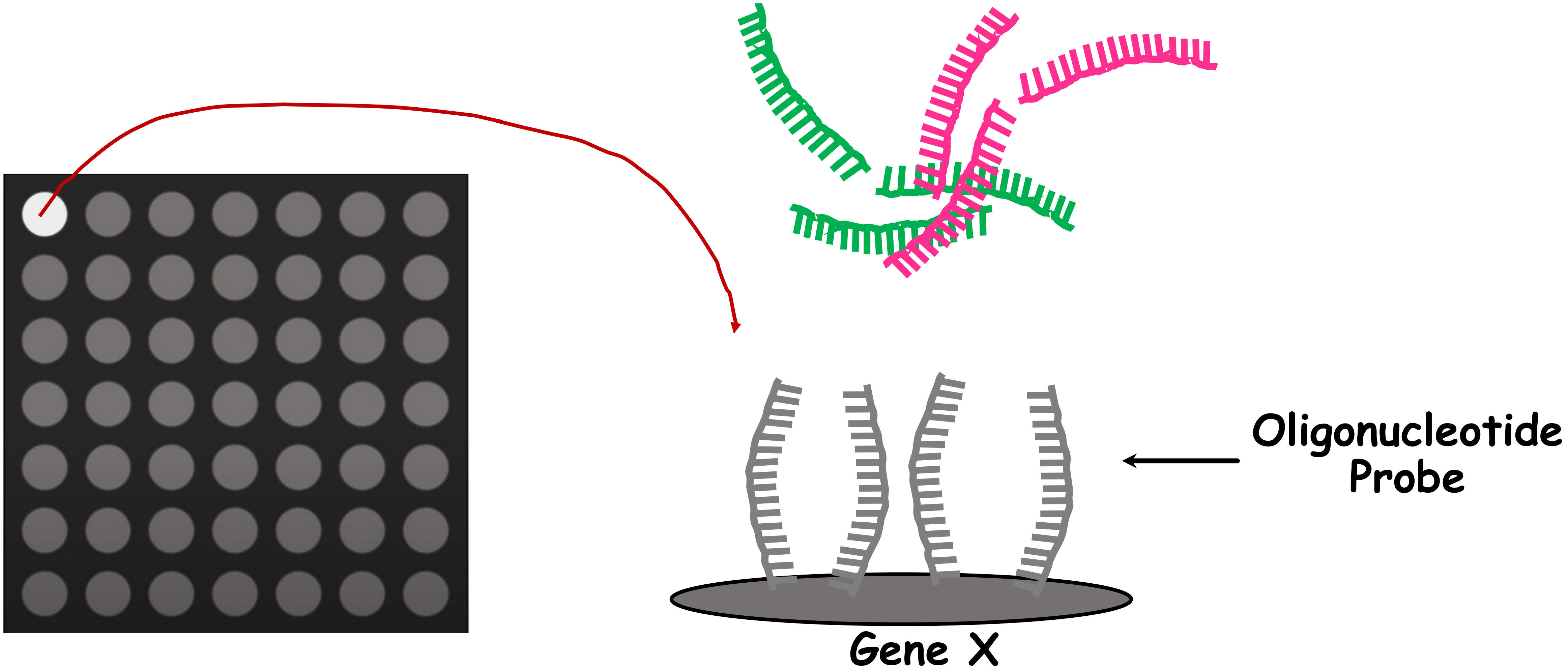
cDNA
Synthesis &
Fluorescence
Labelling

Microarray

Sample Preparation



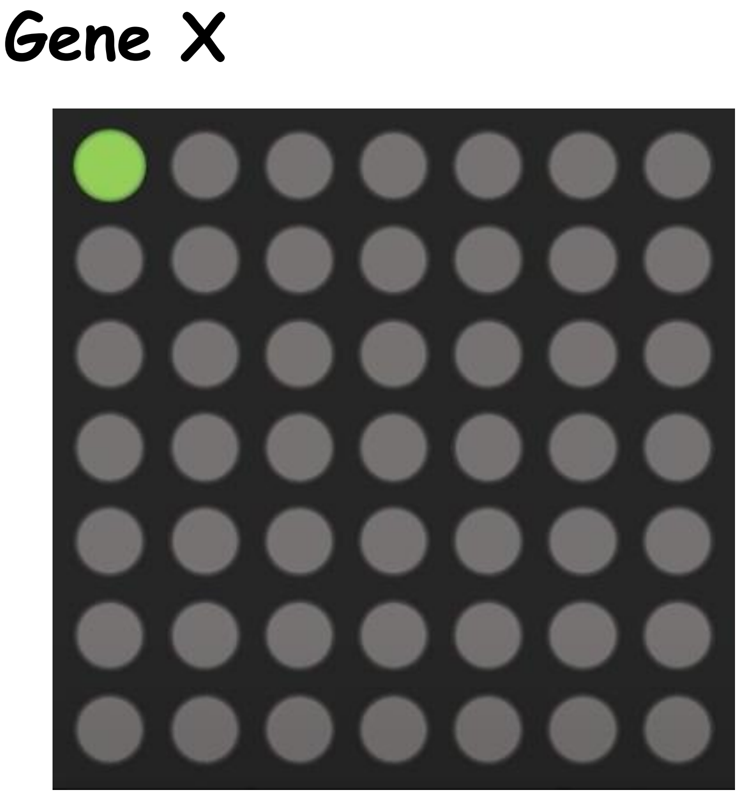
Microarray

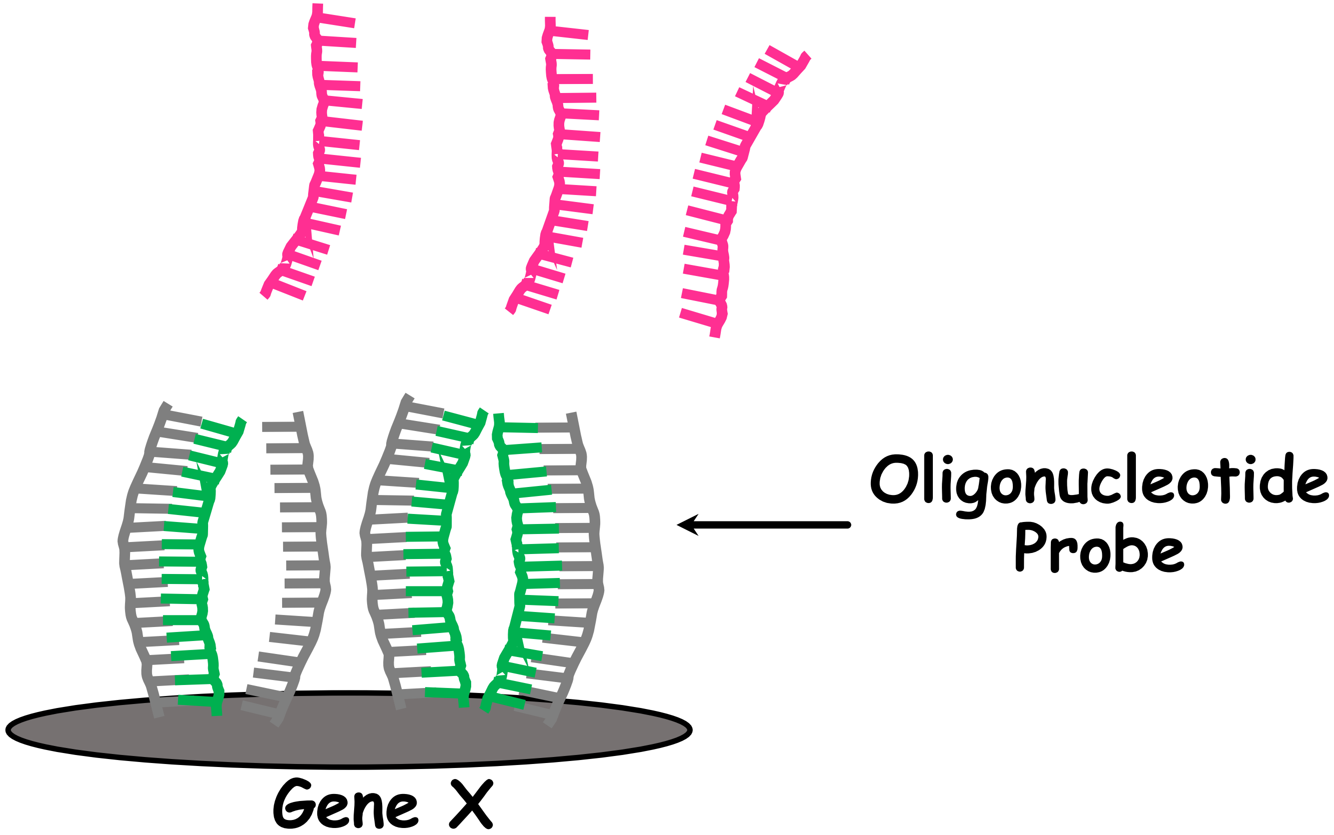


Oligonucleotide Probe

Gene X

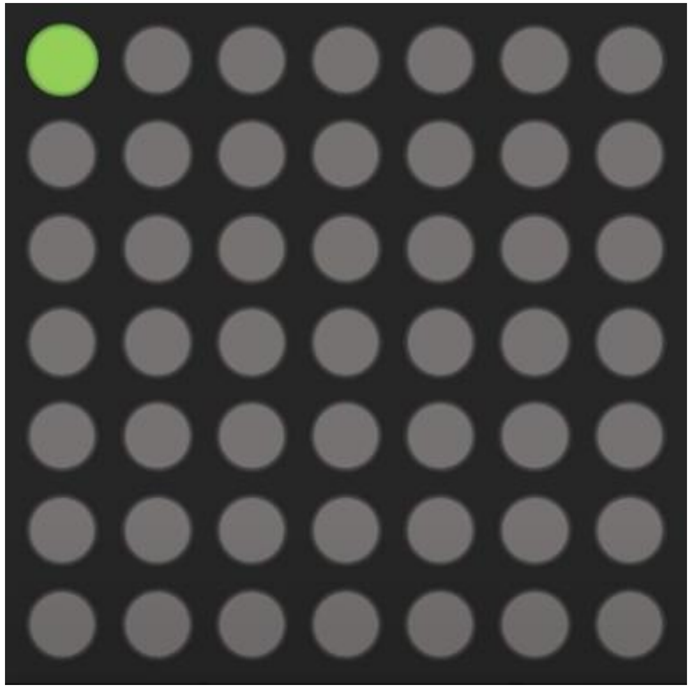
Microarray





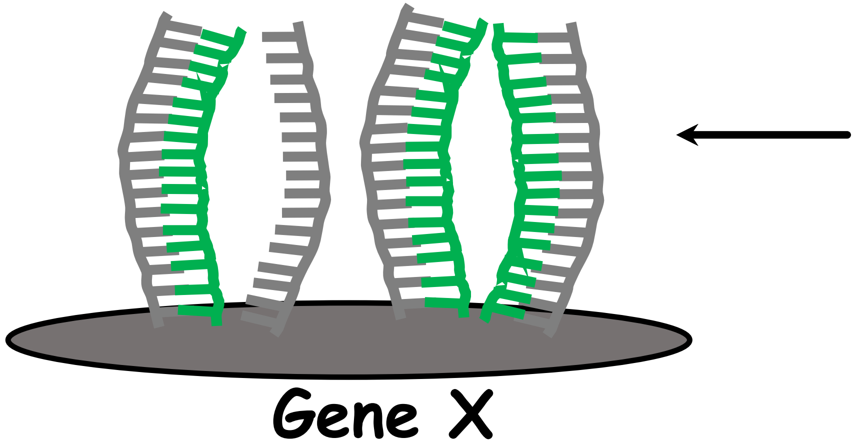
Microarray

Gene X

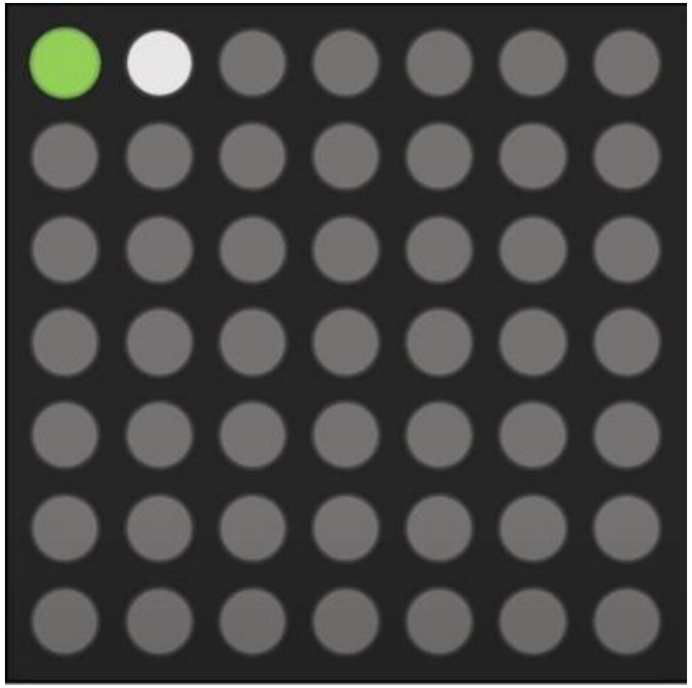


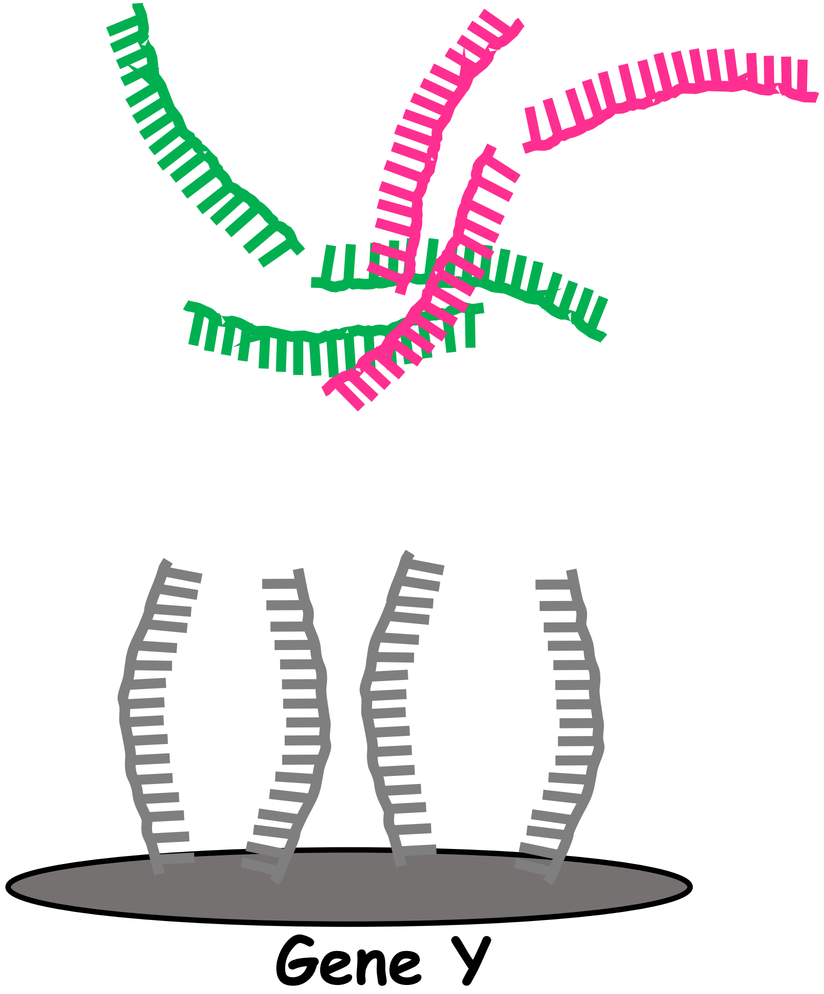
Conclusion:

Gene X is expressed in Sample 1
Gene X is not expressed in Sample 2

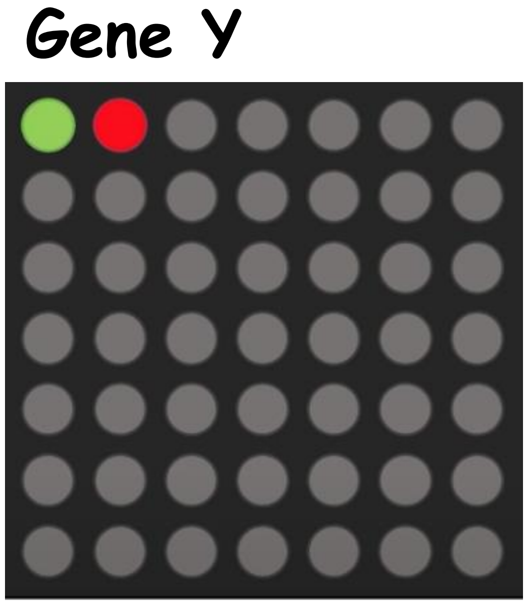


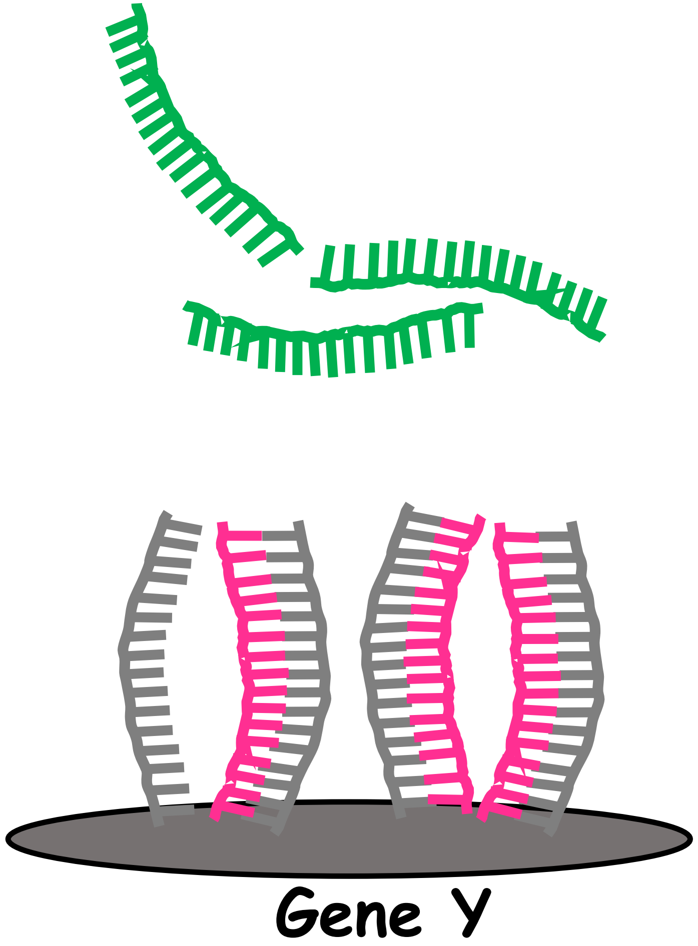
Microarray



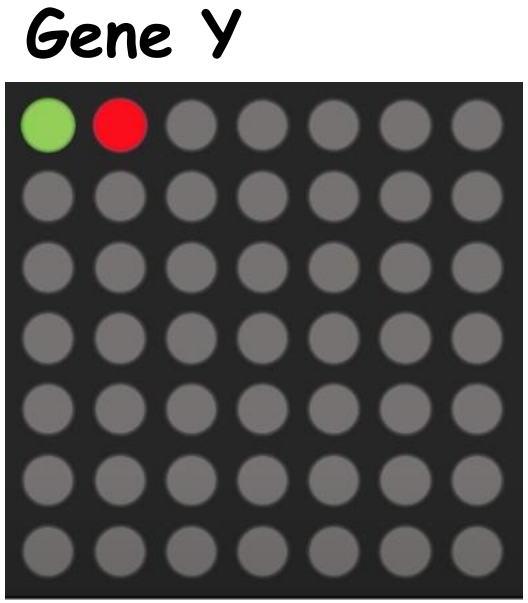


Microarray



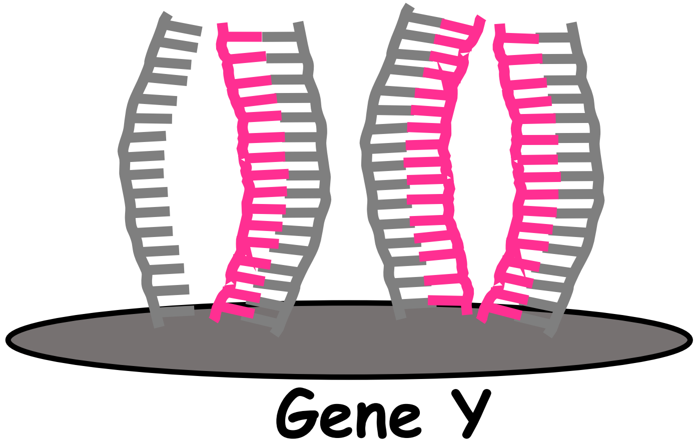


Microarray

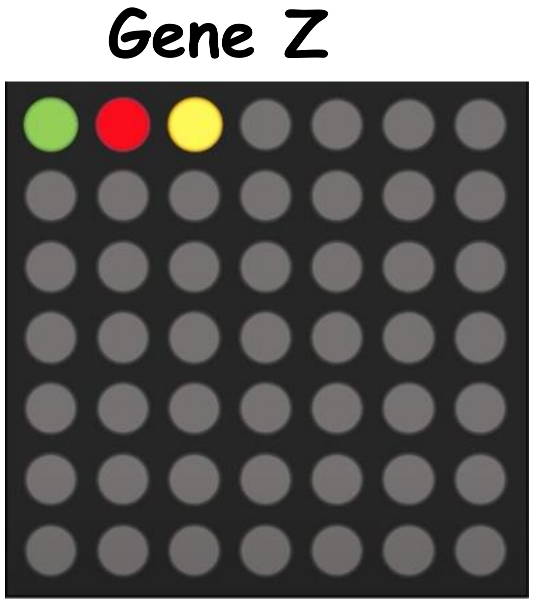


Conclusion:

Gene Y is not expressed in Sample 1
Gene Y is expressed in Sample 2

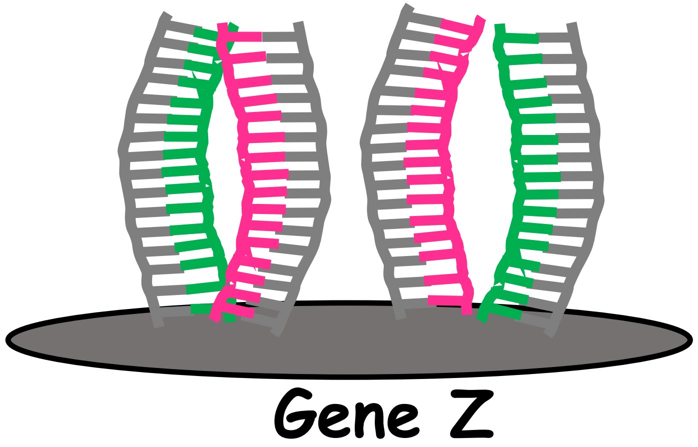


Microarray

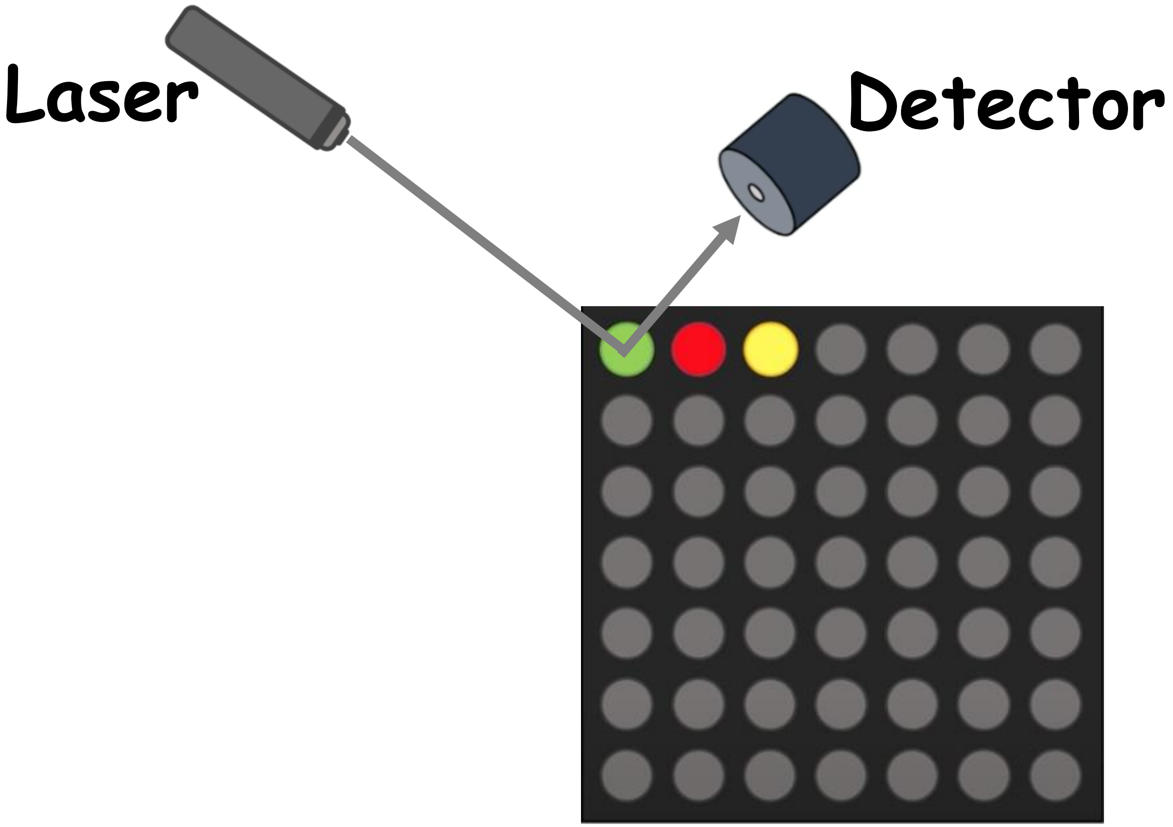


Conclusion:

Gene Z is expressed in Sample 1 and in Sample 2

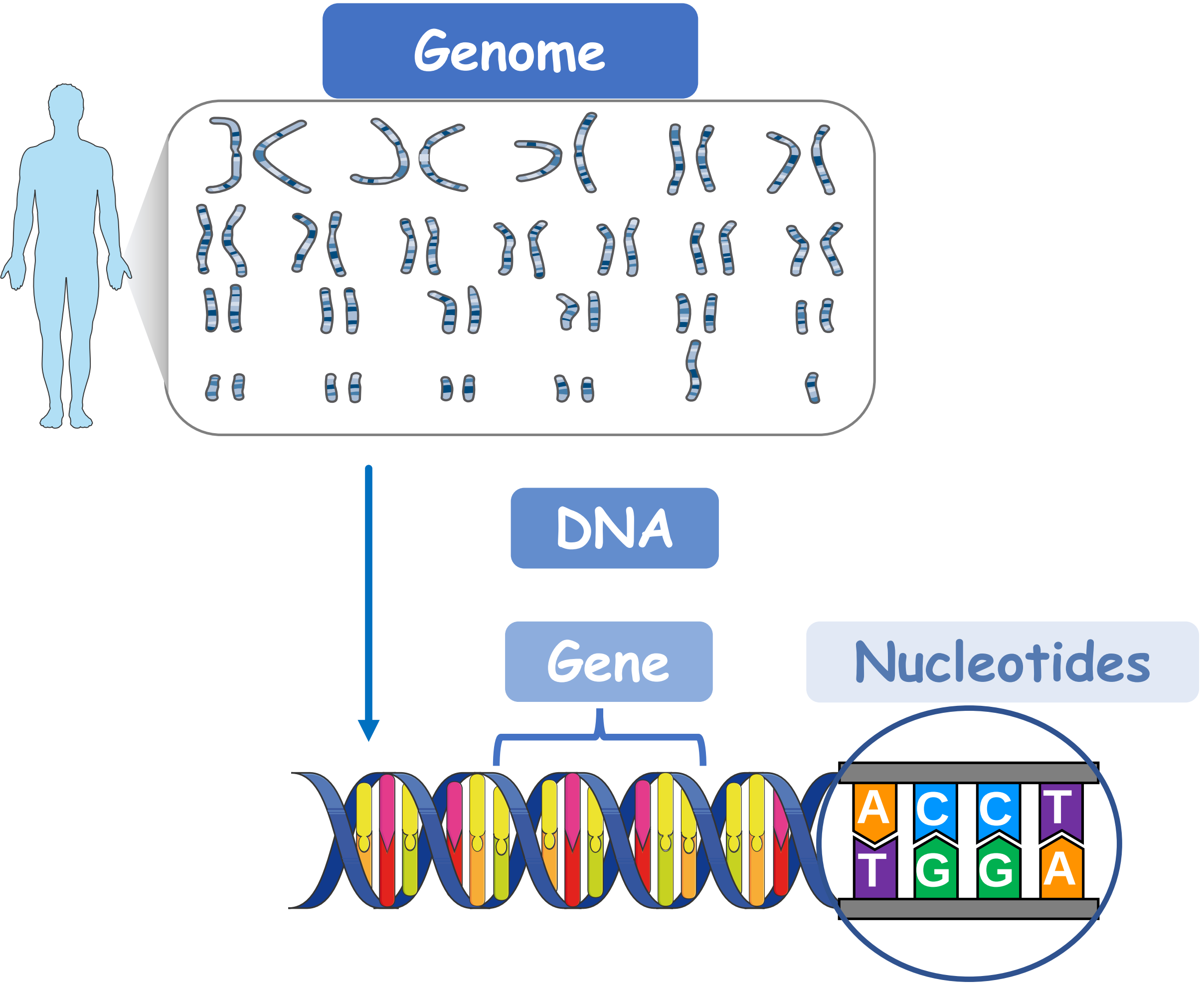


Microarray

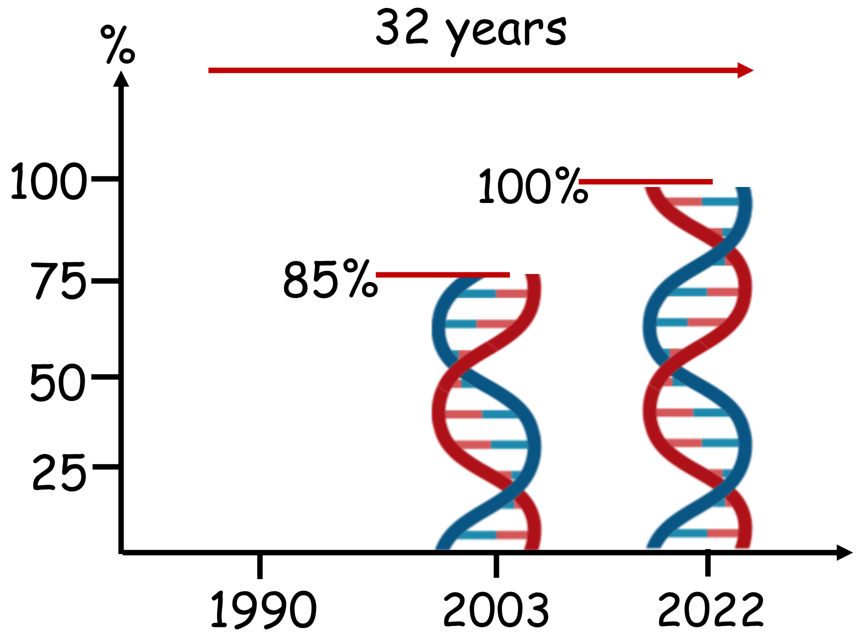


DNA Sequencing

The Human Genome Project

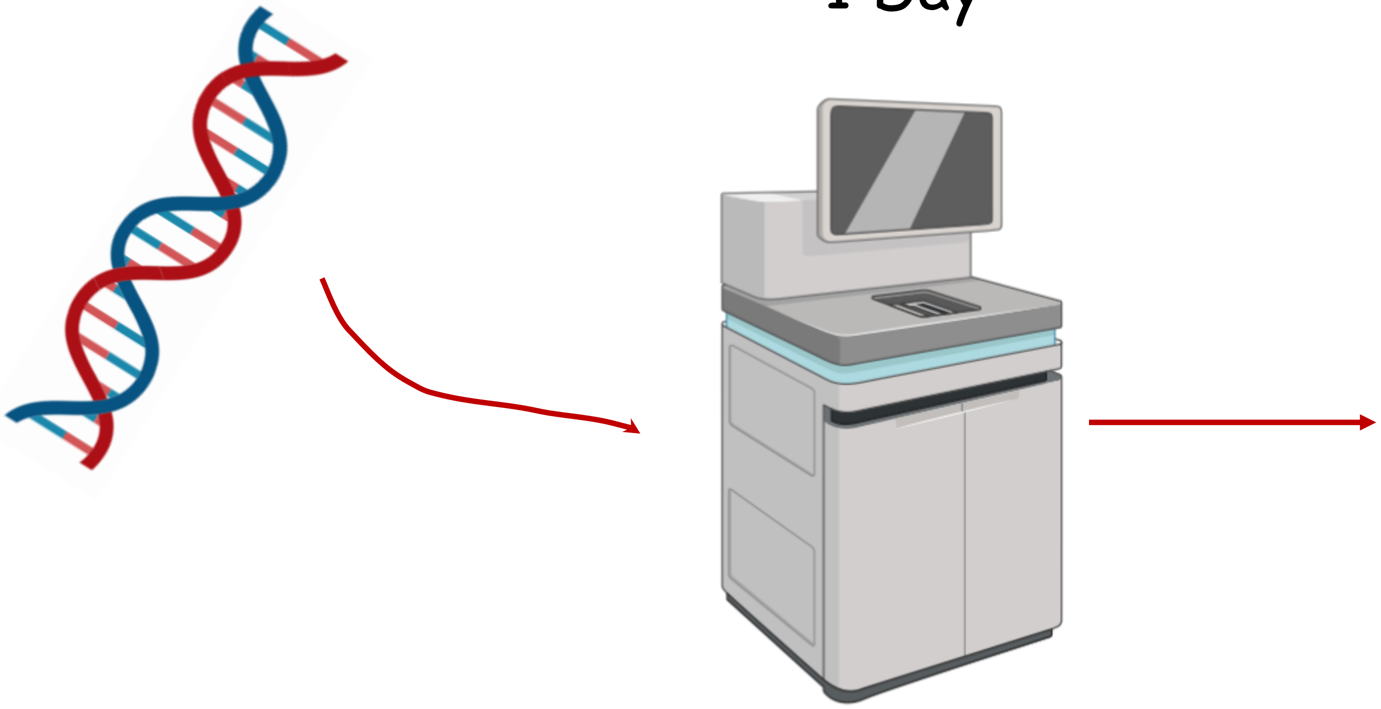


3.2 Billion Bases



Next Generation Sequencing (NGS)

1 Day



Entire Genome

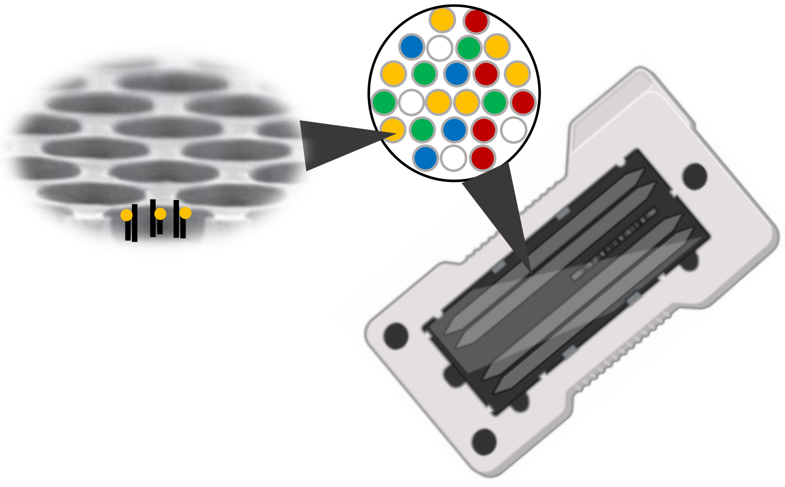
```
GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTGCTCA
     AAGAGGCCATATTAGCTAGG
                     TAGGCTGAATTTTTGCTCA
            CATATTAGCTAGGCTGAATT
GGTGAAAGAGGCCATATTAG
                TTAGCTAGGCTGAATTTTTG
               ATTAGCTAGGCTGAATT
```

Next Generation Sequencing (NGS)

Number of DNA Strands Sequenced

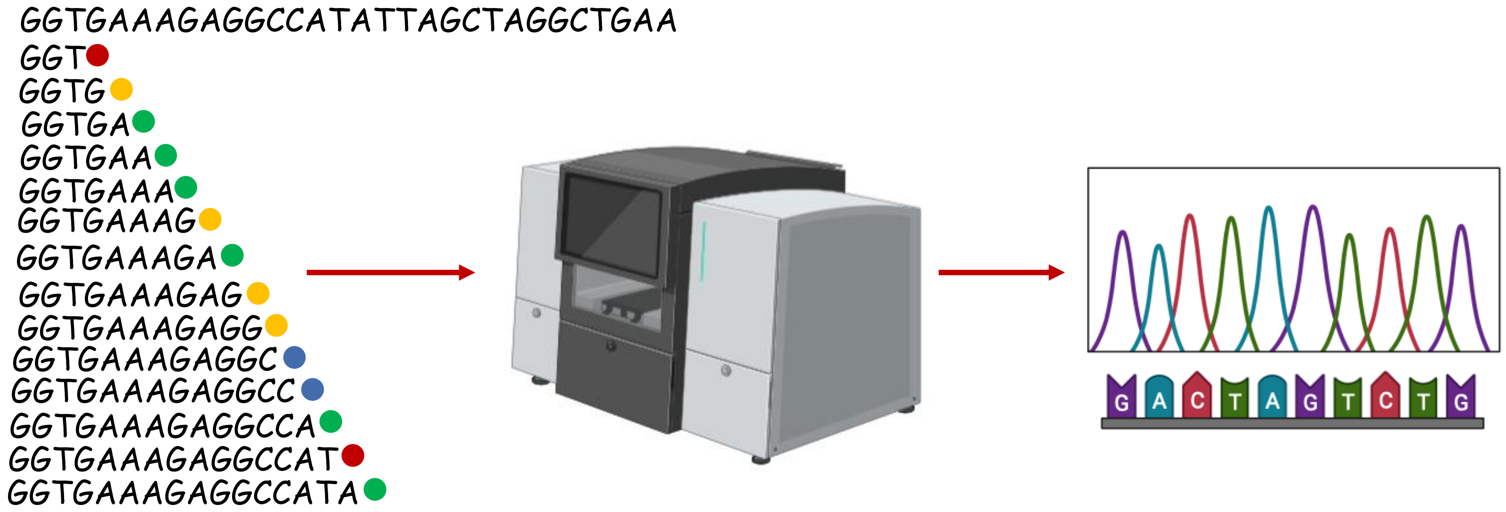
NGS

Billions of Strands



Sanger Sequencing

One Strand



GGTGAAAGAGGCCATATTAGCTAGGCTGAA

GGT

GGTG

GGTGA

GGTGAA

GGTGAAA

GGTGAAAG

GGTGAAAGA

GGTGAAAGAG

GGTGAAAGAGG

GGTGAAAGAGGC

GGTGAAAGAGGCC

GGTGAAAGAGGCCA

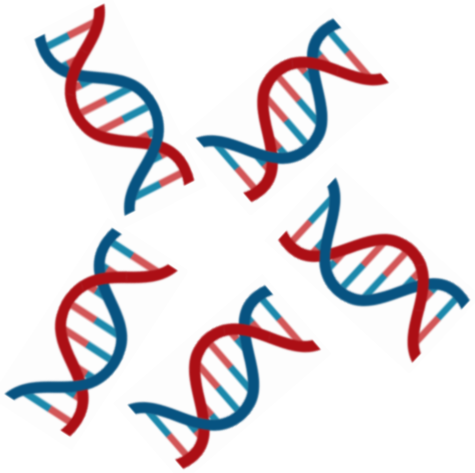
GGTGAAAGAGGCCAT

GGTGAAAGAGGCCATA

Next Generation Sequencing (NGS)

Human Genome Project → Human Reference DNA

………………GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTGCTCA………………



Billions of Strands

→

Sequencer

→

GGTGAAAGAGGCCATATTAG

AAGAGGCCATATTAGCTAGG

TTAGCTAGGCTGAATTTTTG

ATATTAGCTAGGCTGAATTT

CATATTAGCTAGGCTGAATT

TAGGCTGAATTTTTGCTCA

Sequences

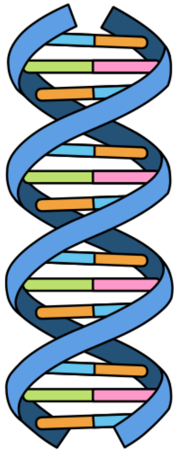
Next Generation Sequencing (NGS)

Human Genome Project → Human Reference DNA

```
..................GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTTGCTCA..................
                       AAGAGGCCATATTAGCTAGG
                                       TAGGCTGAATTTTTTGCTCA
                              CATATTAGCTAGGCTGAATT
                  GGTGAAAGAGGCCATATTAG
                                  TTAGCTAGGCTGAATTTTTTG
                               ATATTAGCTAGGCTGAATTT
```

Next Generation Sequencing (NGS)

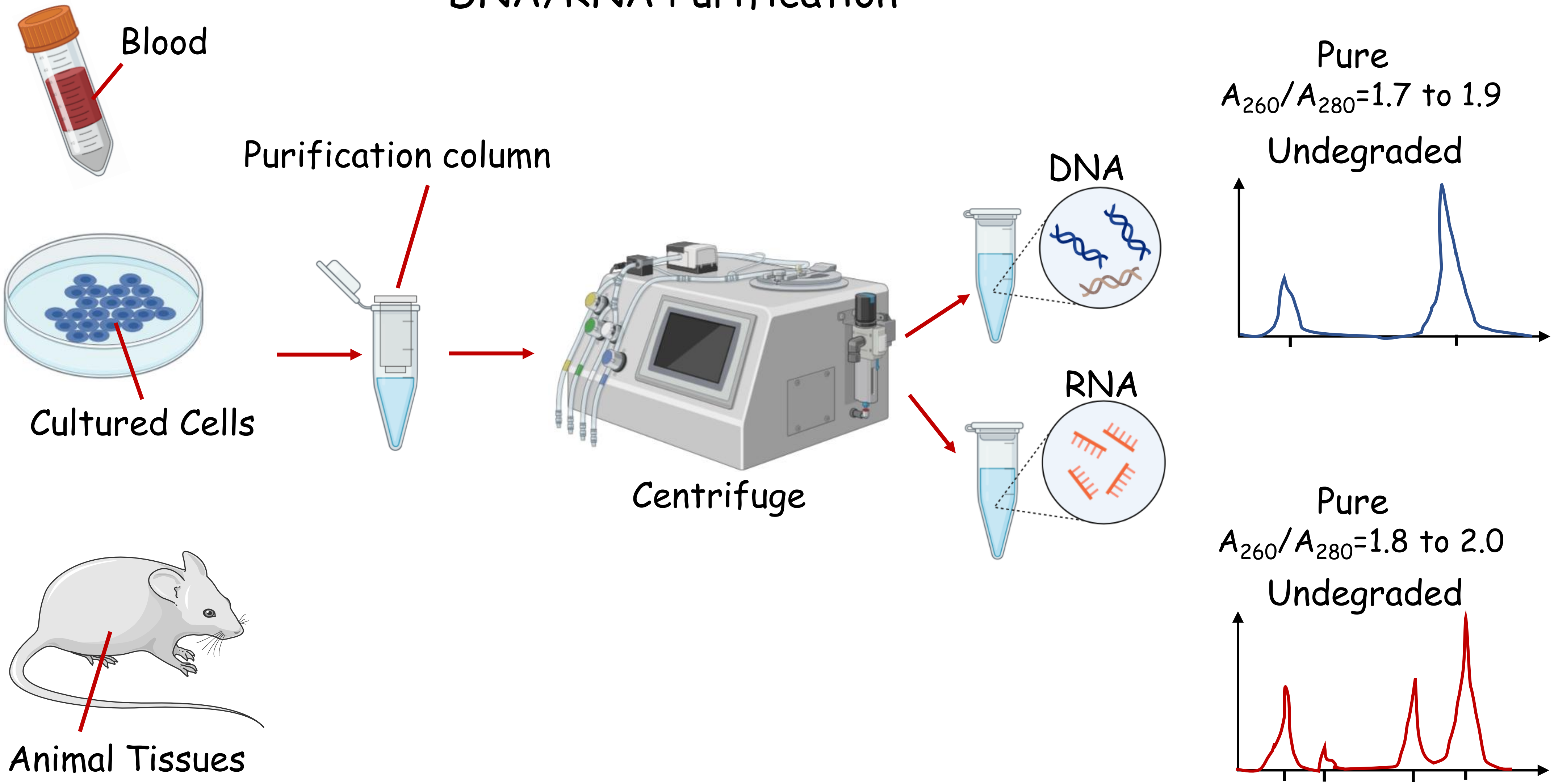
DNA



RNA

Next Generation Sequencing (NGS)

DNA/RNA Purification



Blood

Cultured Cells

Animal Tissues

Purification column

Centrifuge

DNA

RNA

Pure

A_{260}/A_{280}=1.7 to 1.9

Undegraded

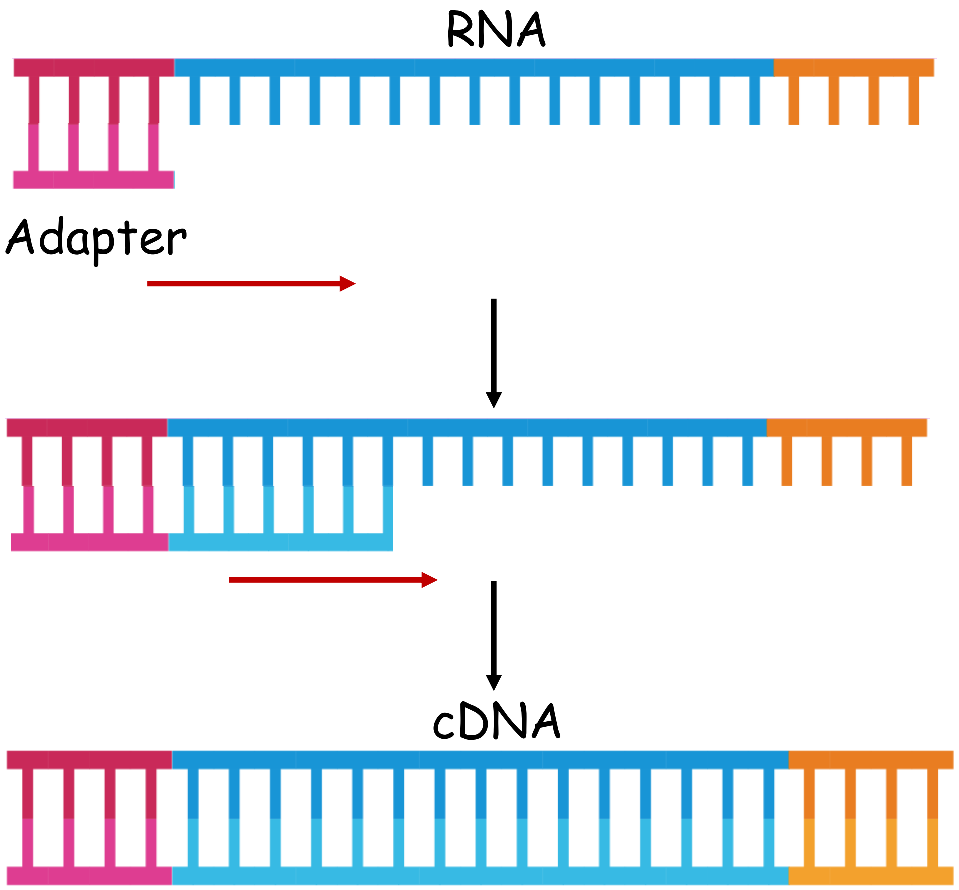
Pure

A_{260}/A_{280}=1.8 to 2.0

Undegraded

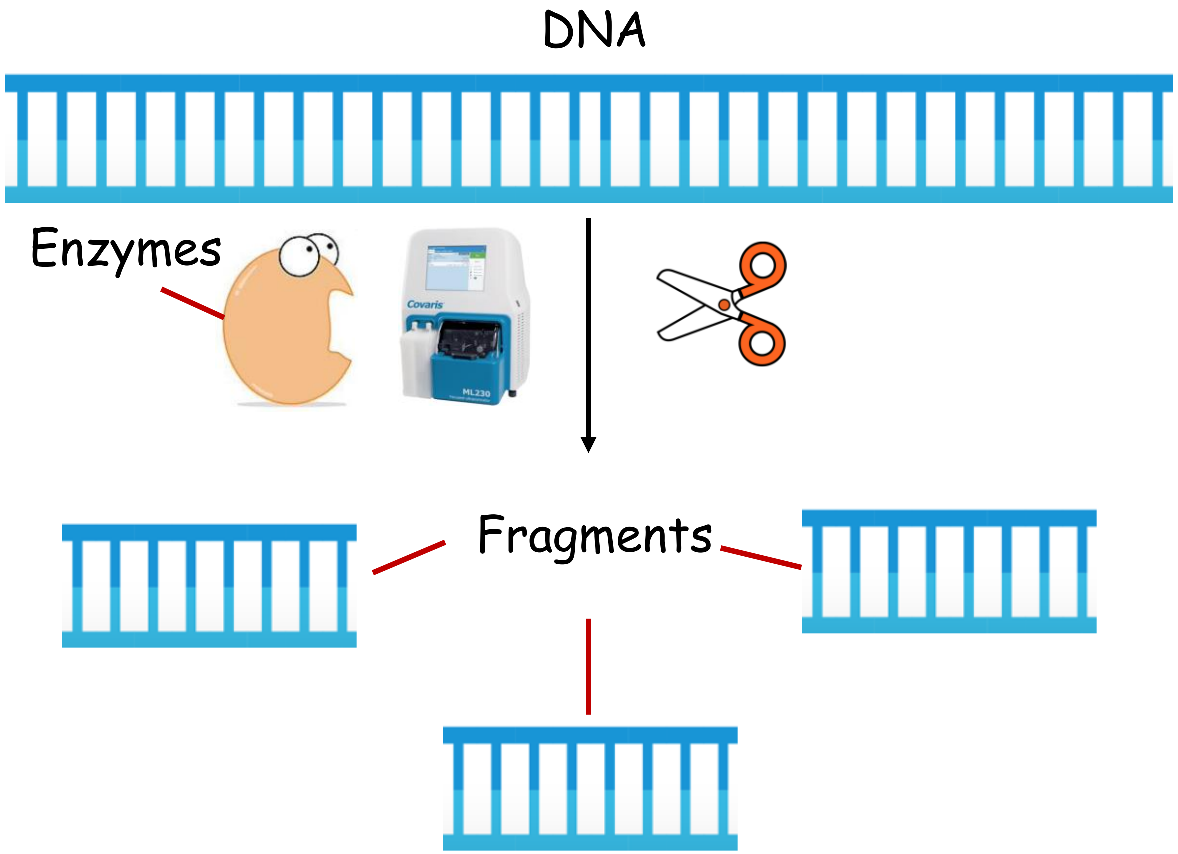
Next Generation Sequencing (NGS)

RNA is Reverse Transcribed



Next Generation Sequencing (NGS)

Library Preparation



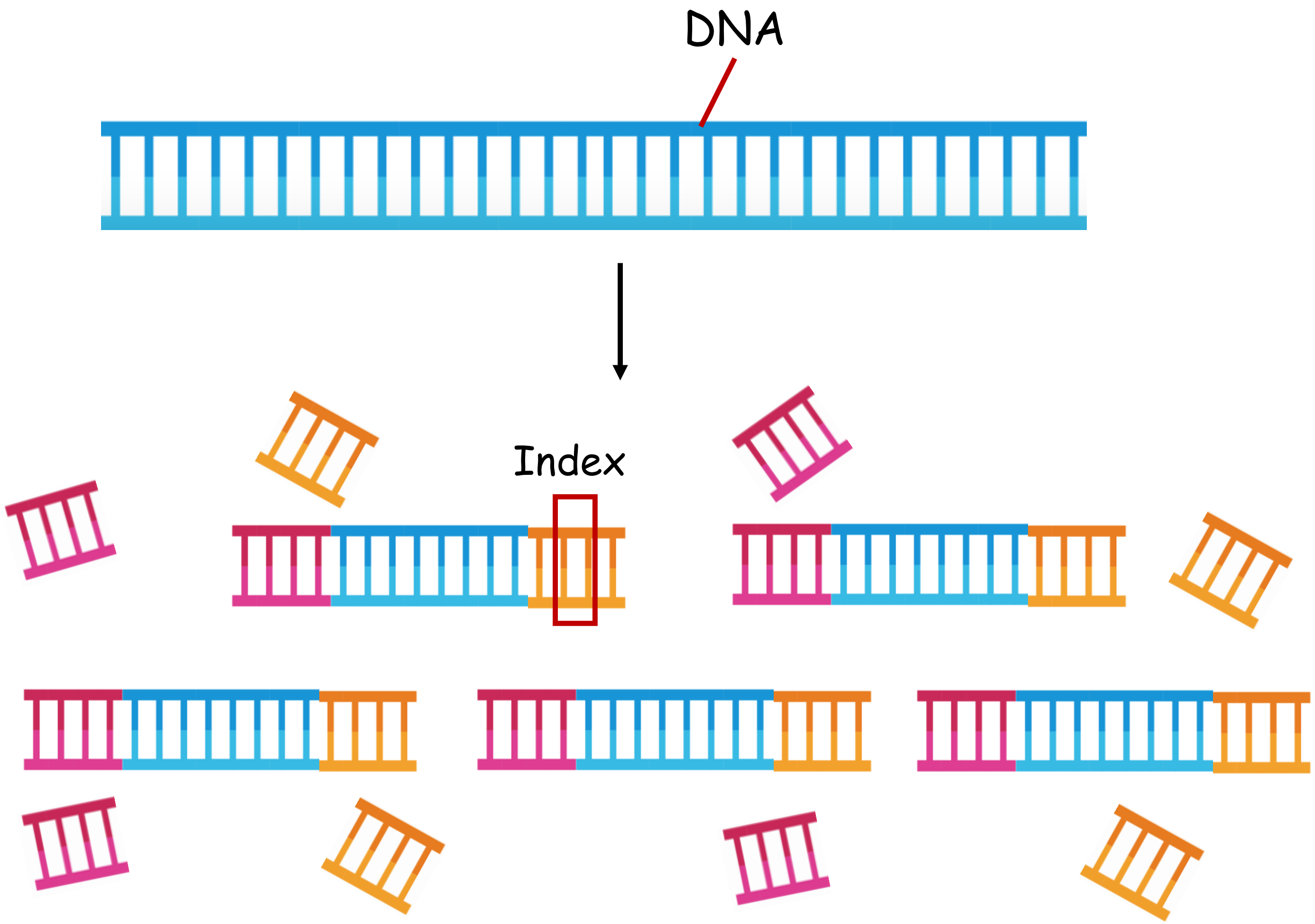
DNA

Enzymes

Fragments

Next Generation Sequencing (NGS)

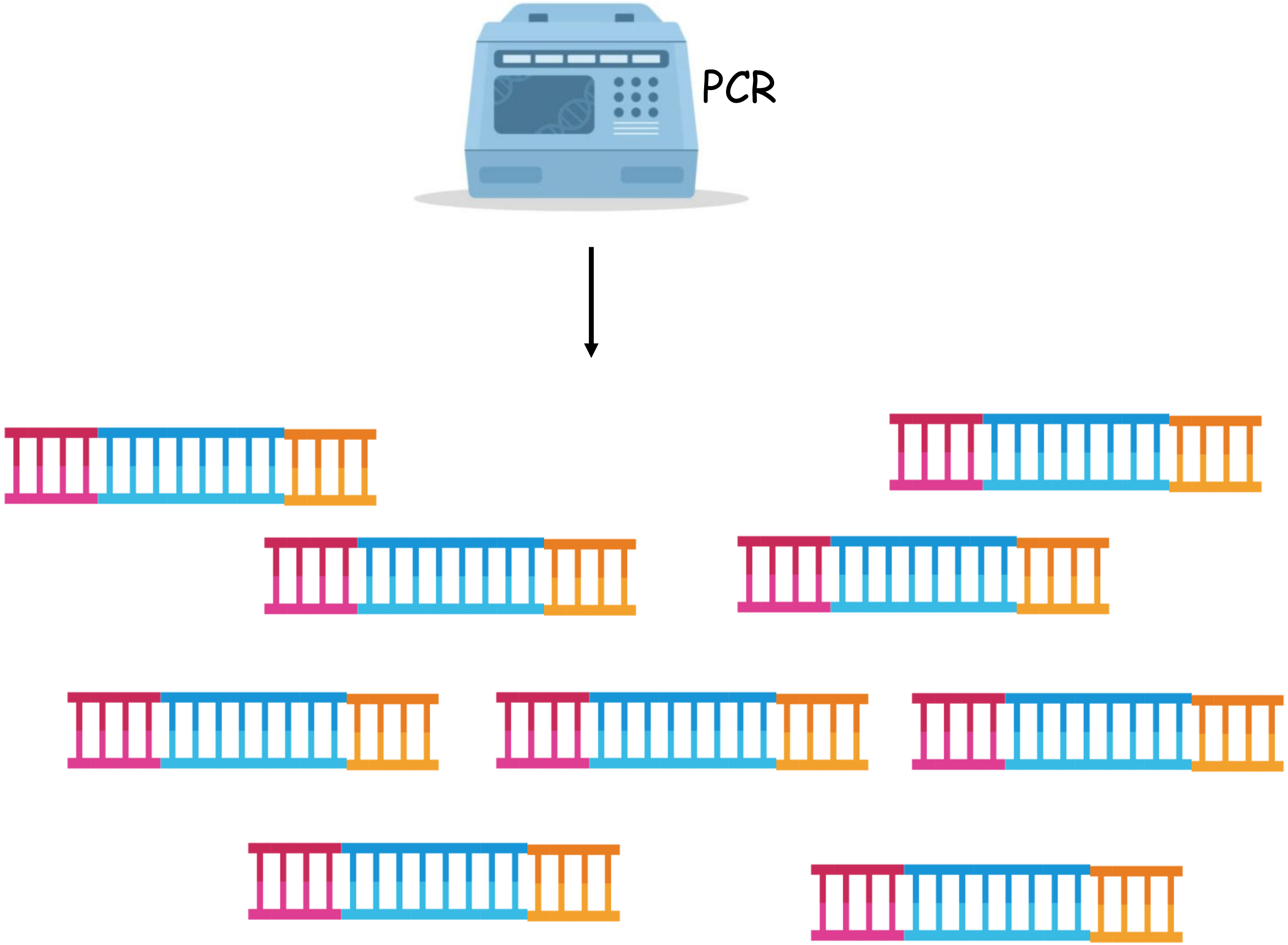
Library Preparation



Information needed for sequencing

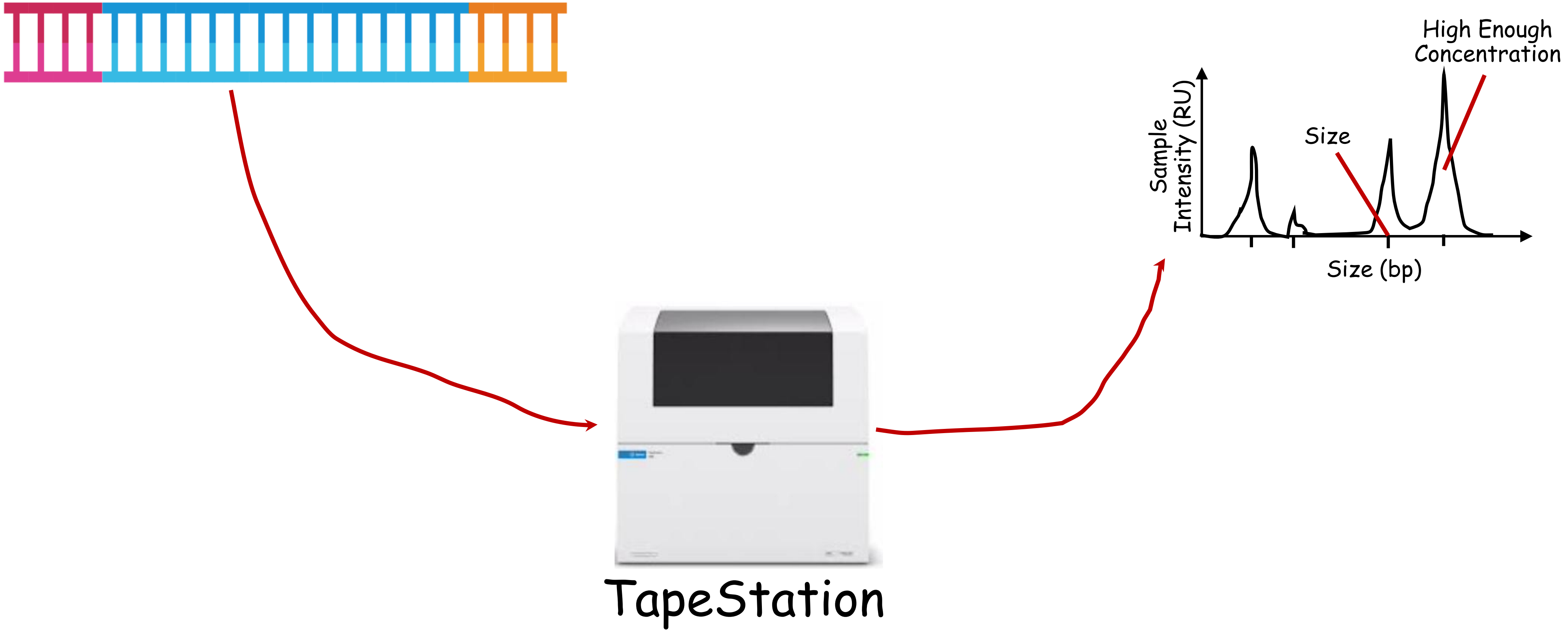
Next Generation Sequencing (NGS)

Library Preparation



Next Generation Sequencing (NGS)

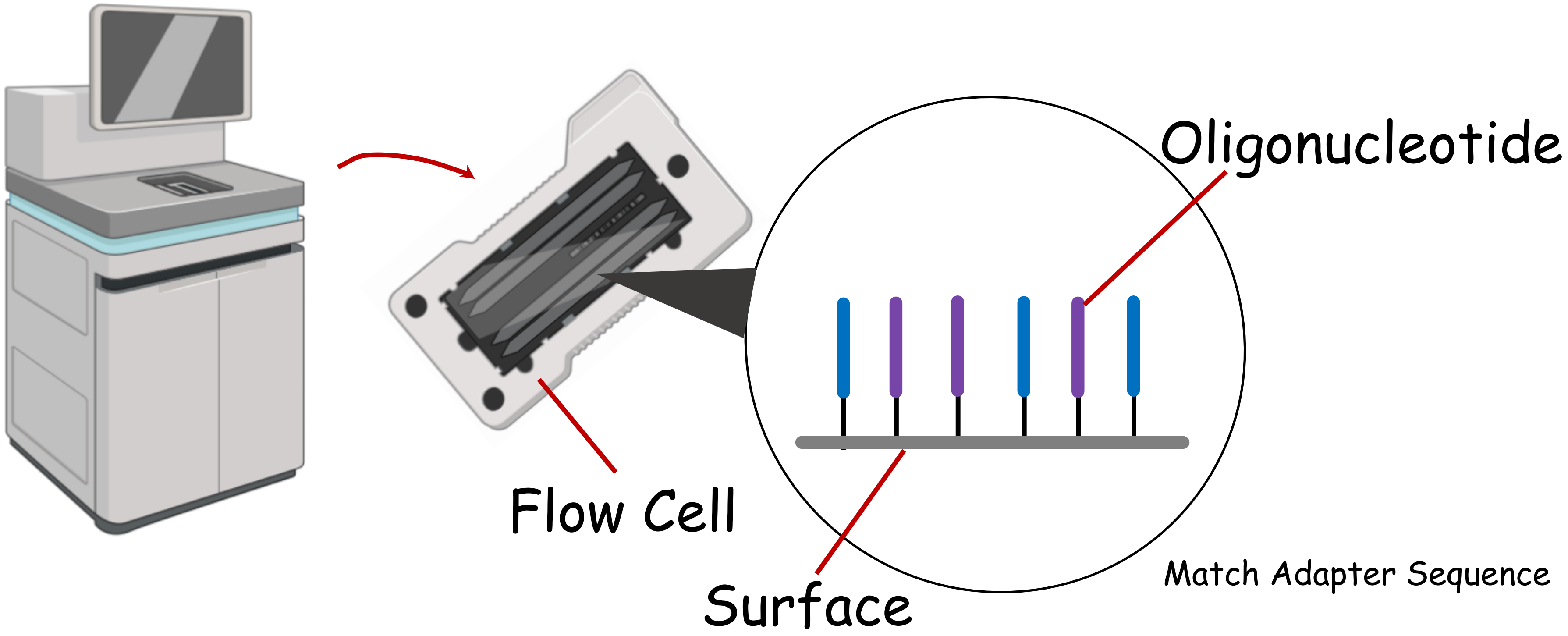
Library Preparation



Next Generation Sequencing (NGS)

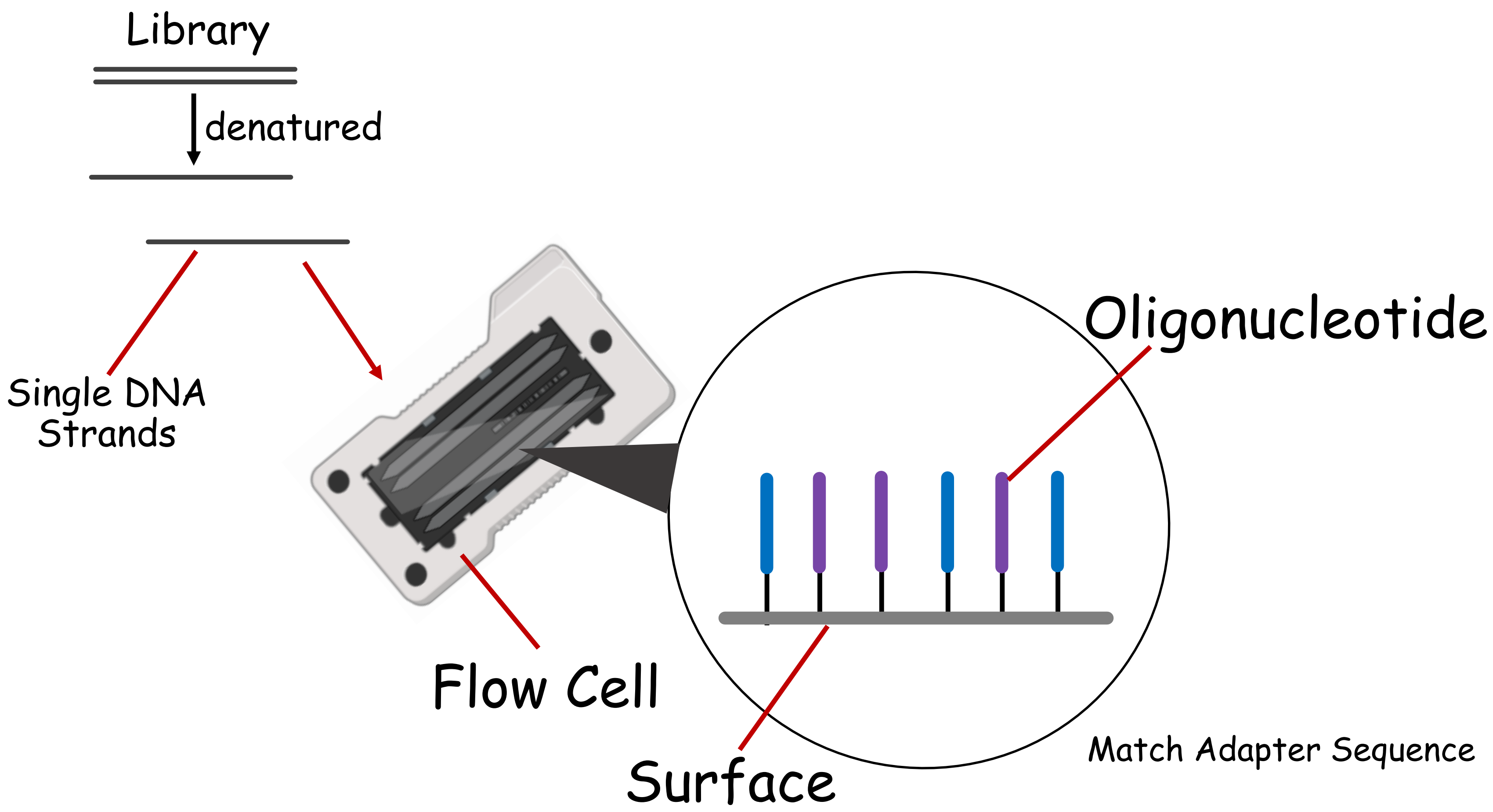
Illumina

Sequencing by Synthesis (SBS)



Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



Library

denatured

Single DNA Strands

Flow Cell

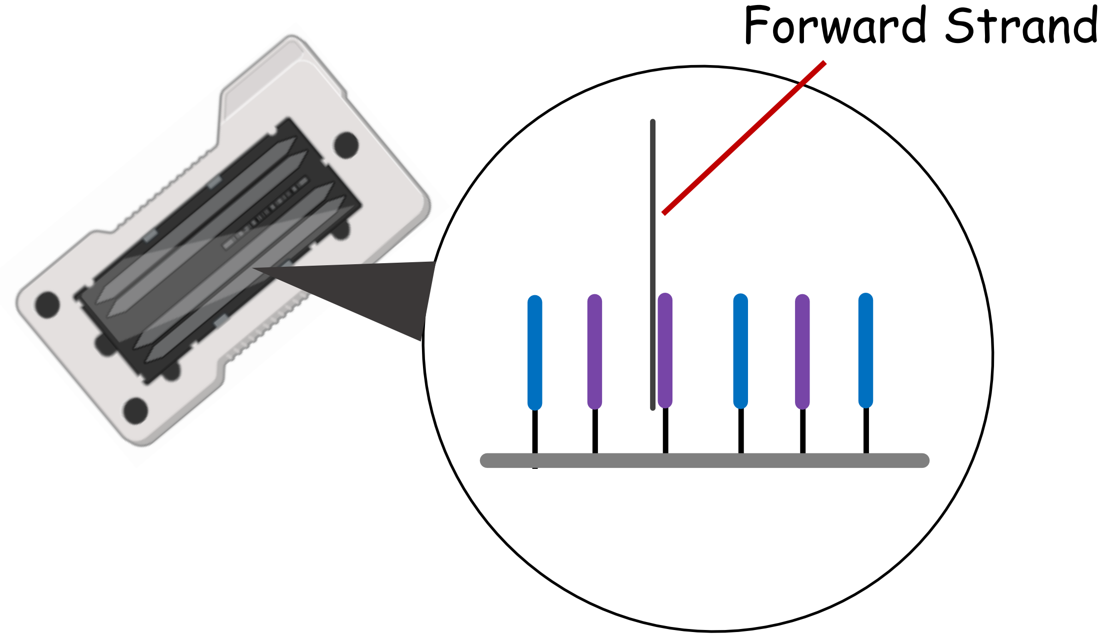
Surface

Oligonucleotide

Match Adapter Sequence

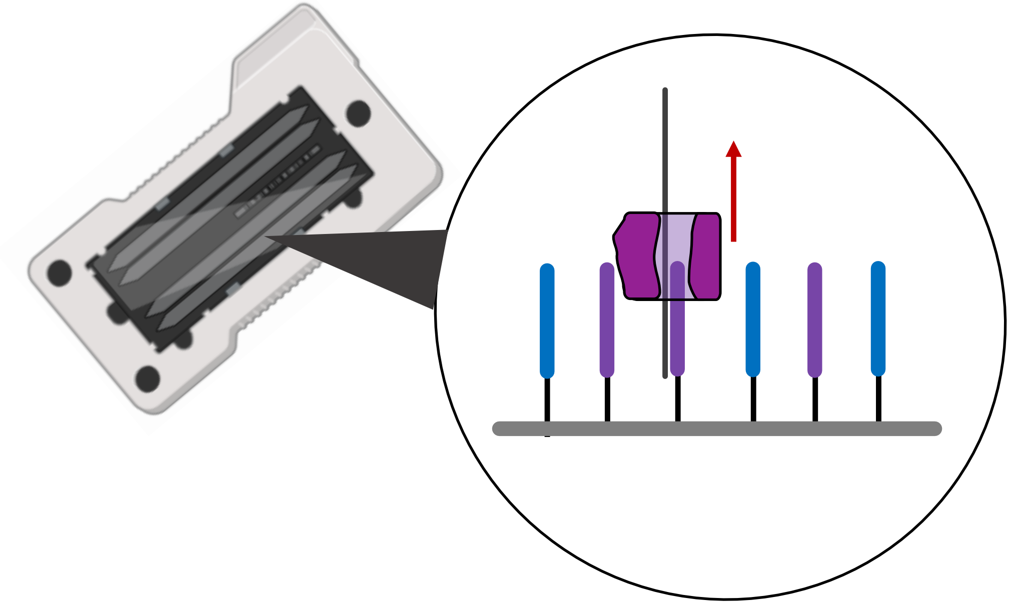
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



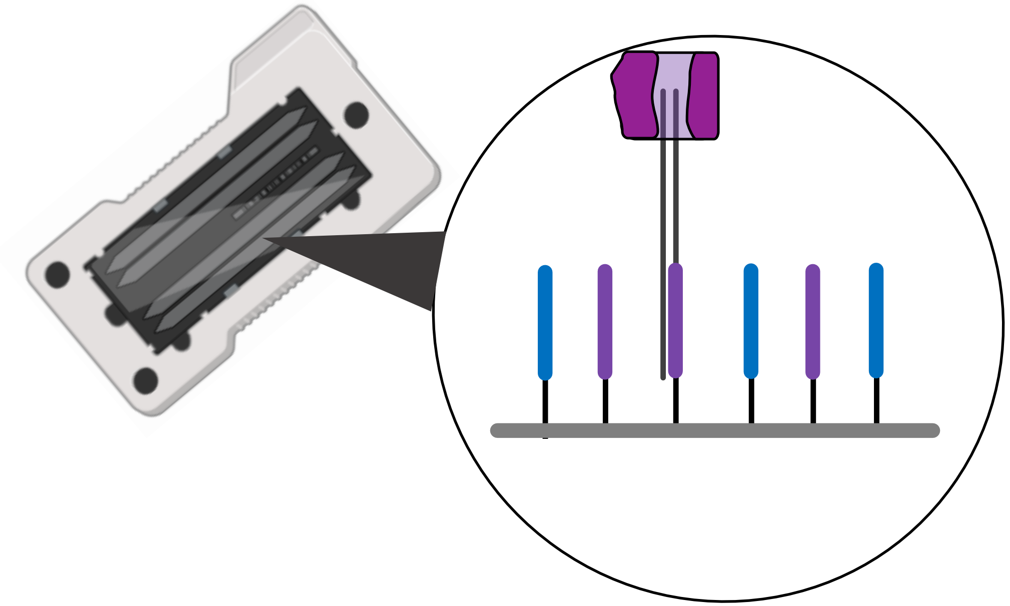
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



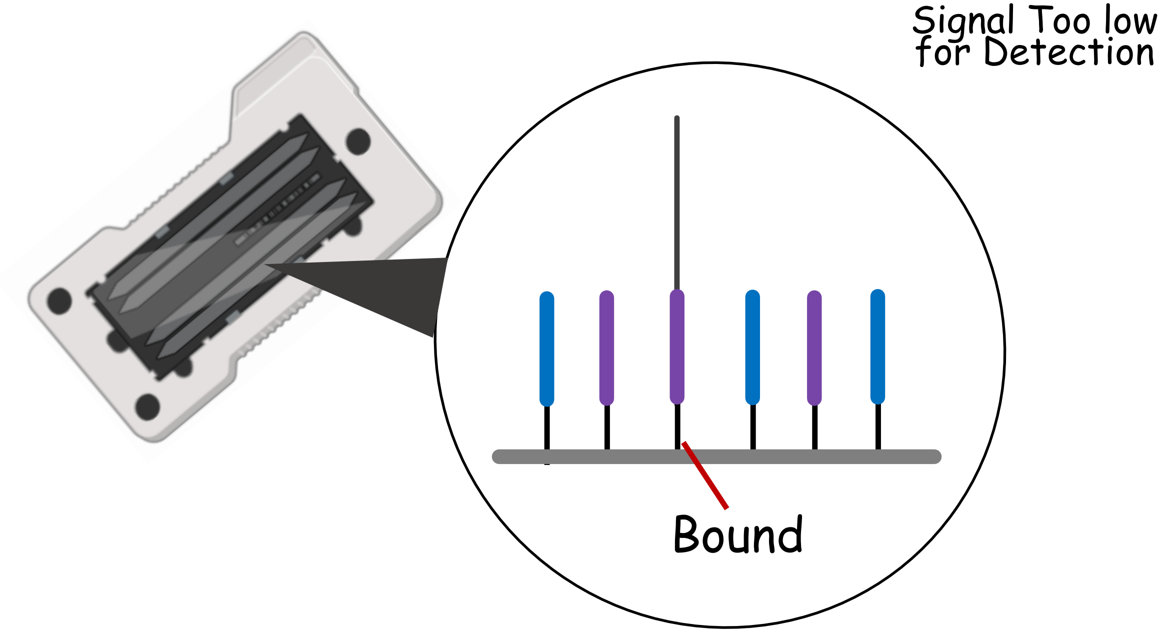
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

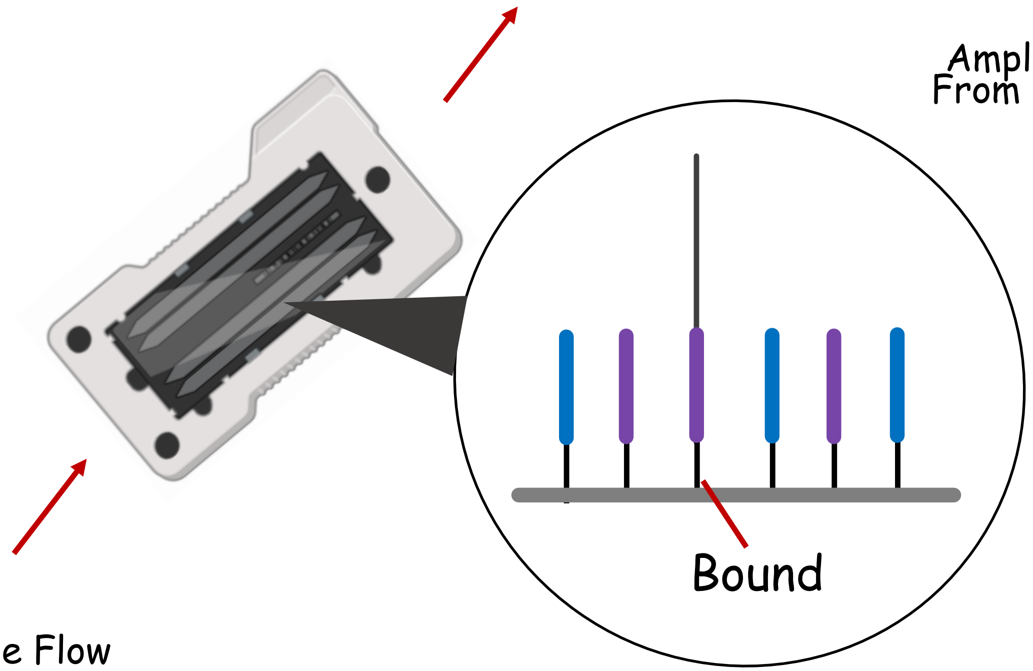


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



Amplified to
From Clusters

Annealing

Extension

Melting

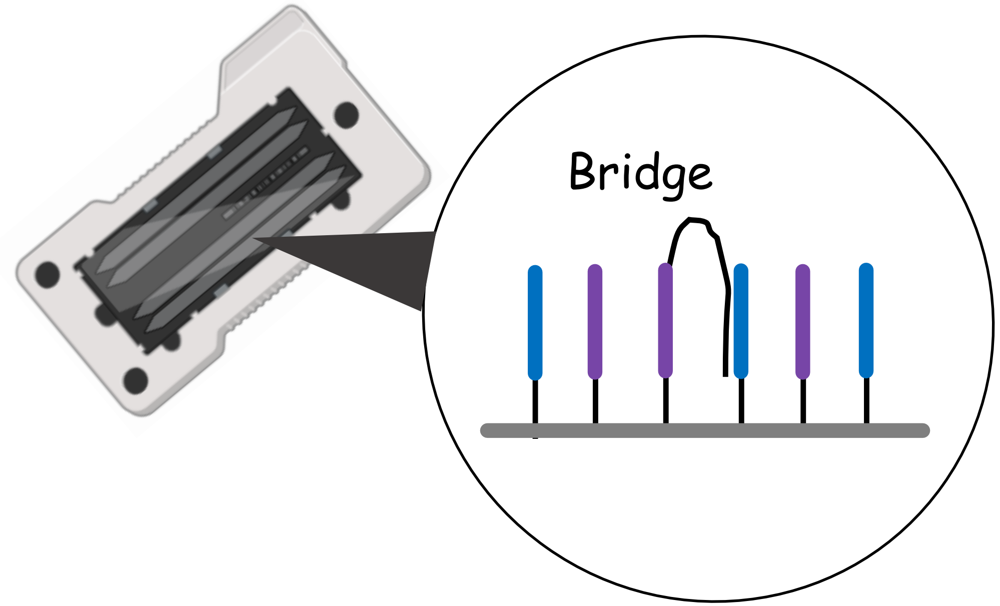
Changing the Flow
Cell Solution

Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



Annealing

Extension

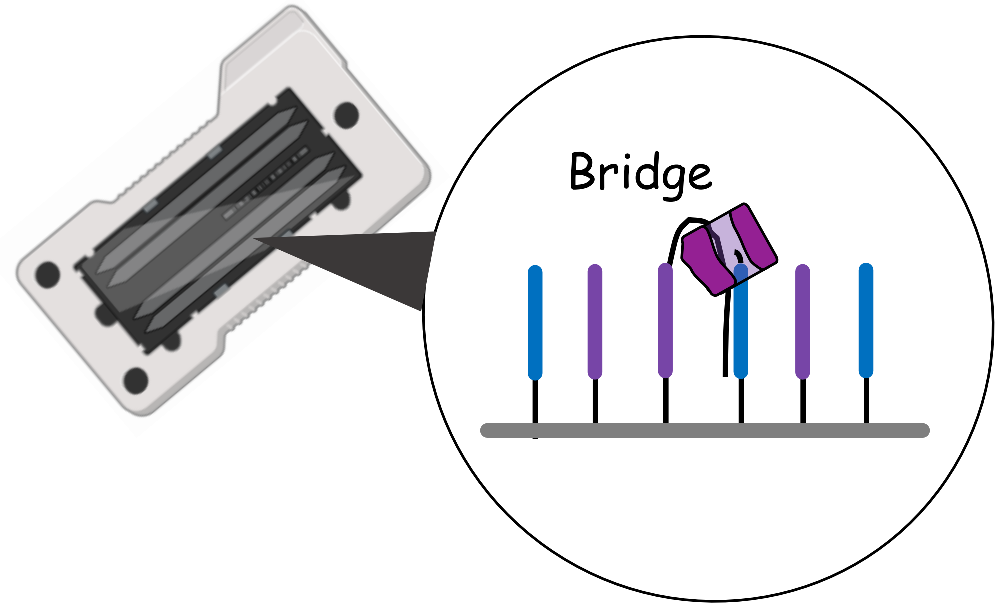
Melting

Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



Annealing

Extension

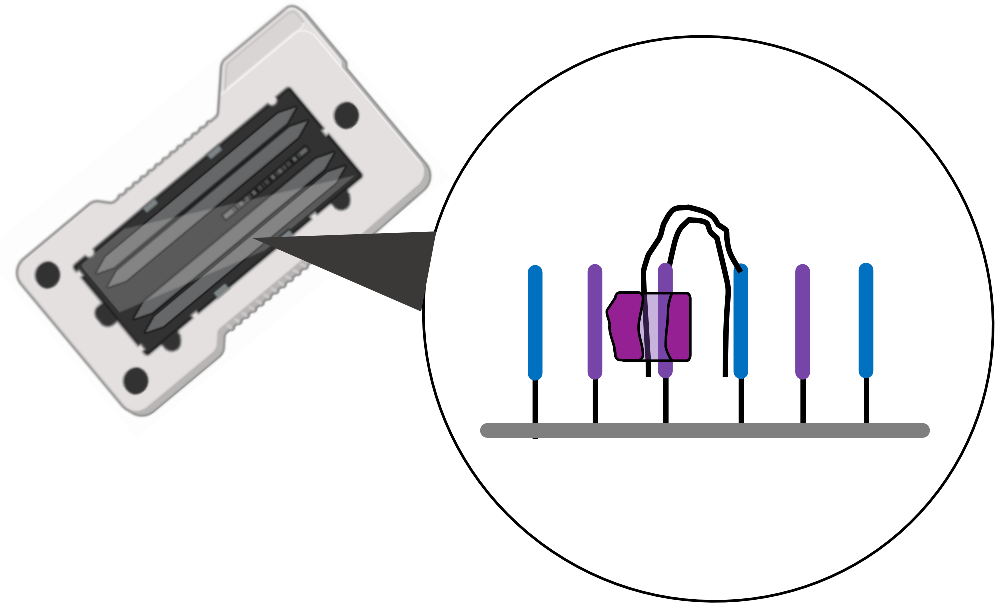
Melting

Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



Annealing

Extension

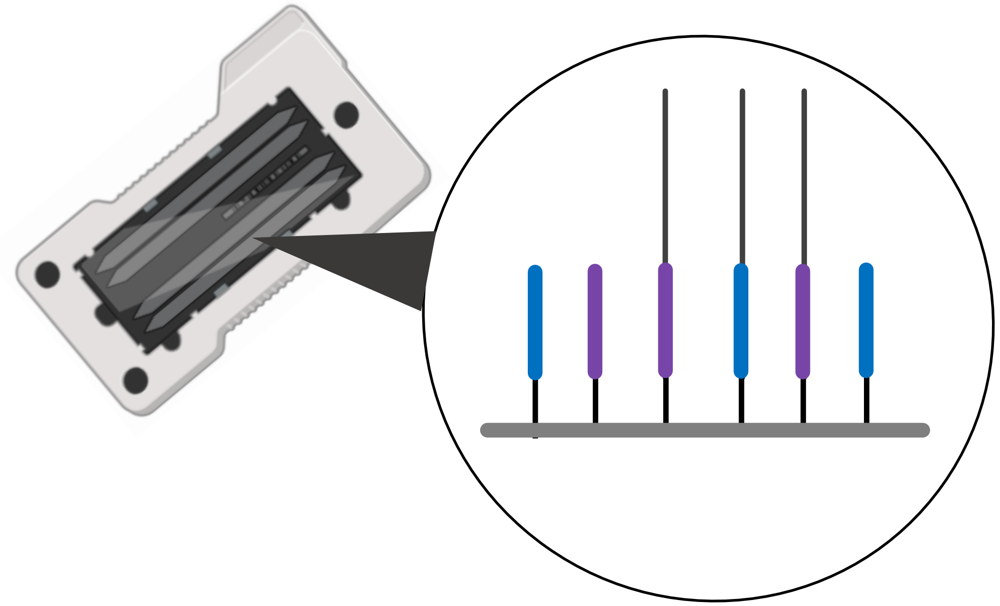
Melting

Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



Annealing

Extension

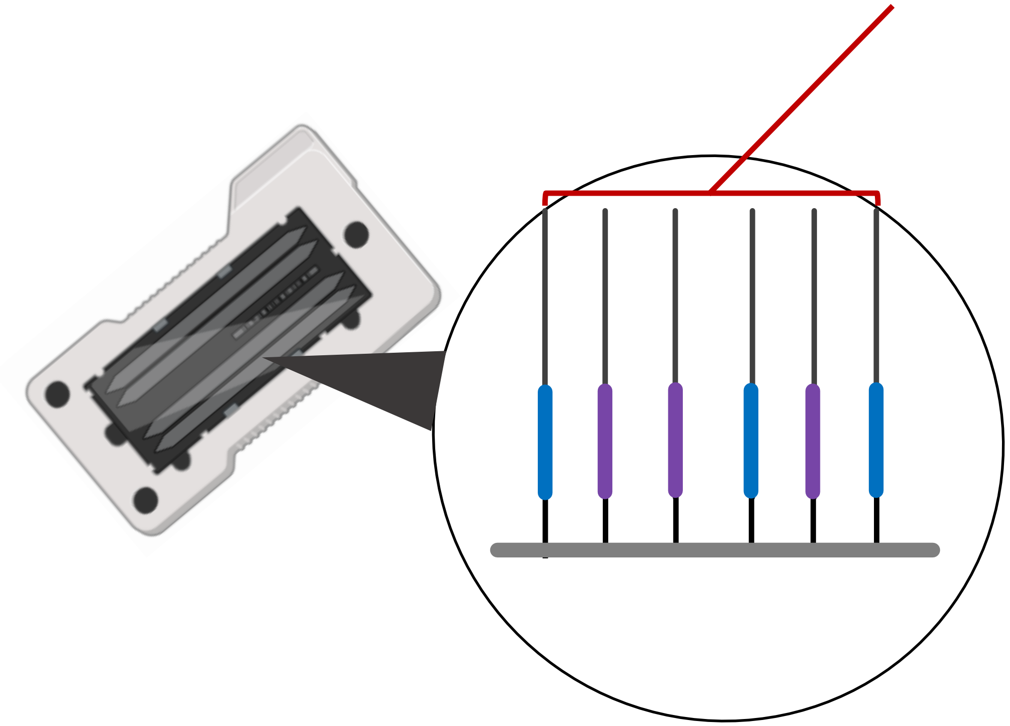
Melting

Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Clonal Amplification

PCR at a Single Temperature



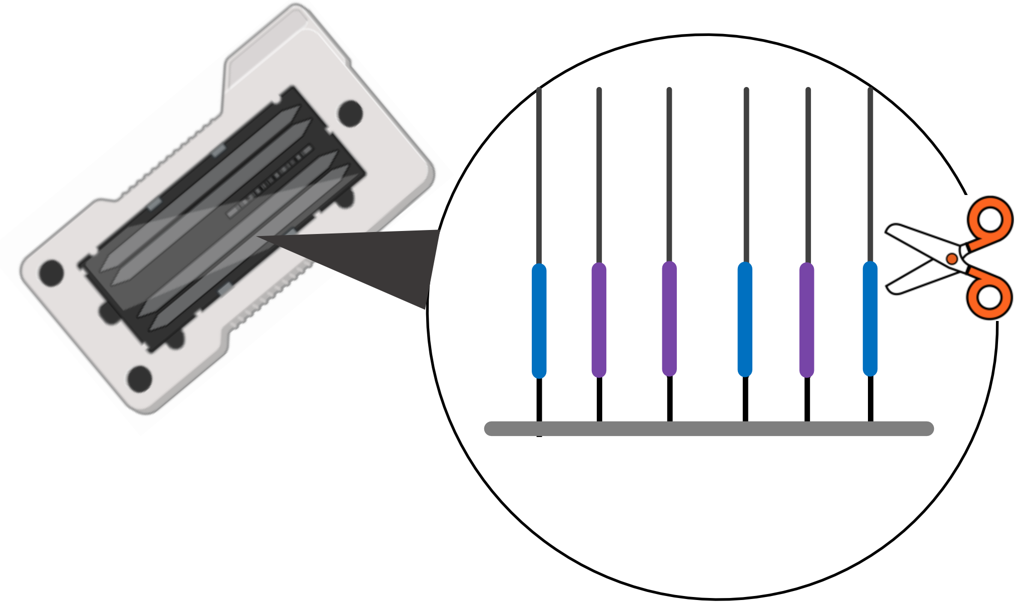
Annealing

Extension

Melting

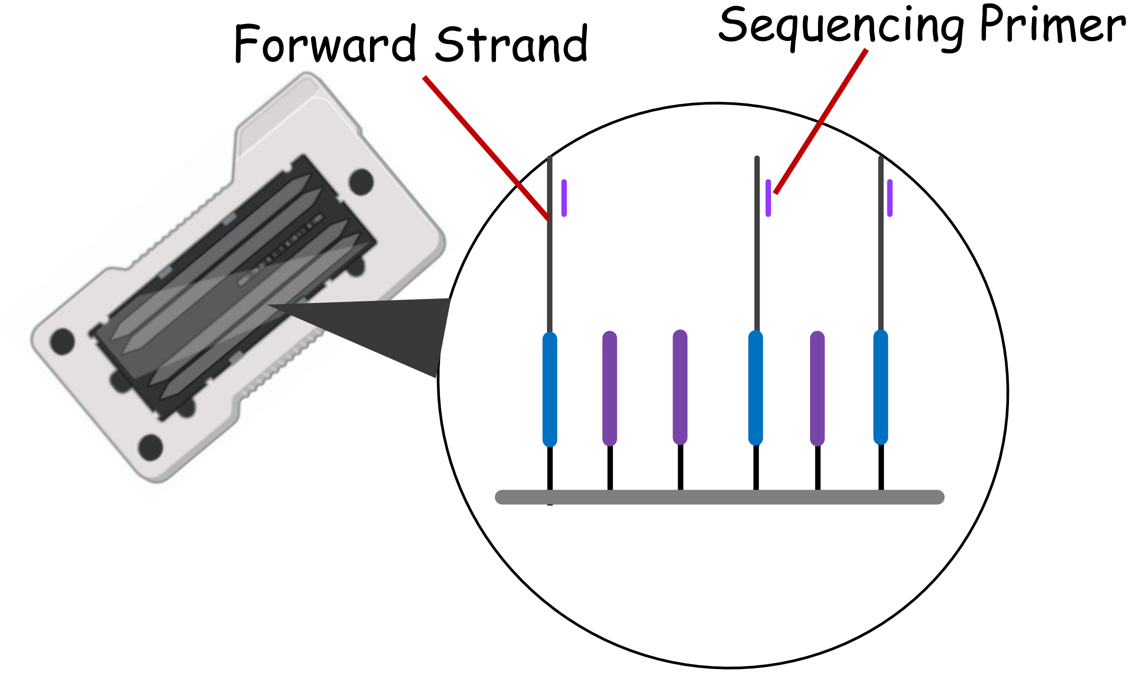
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

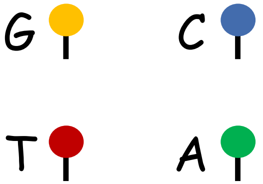


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



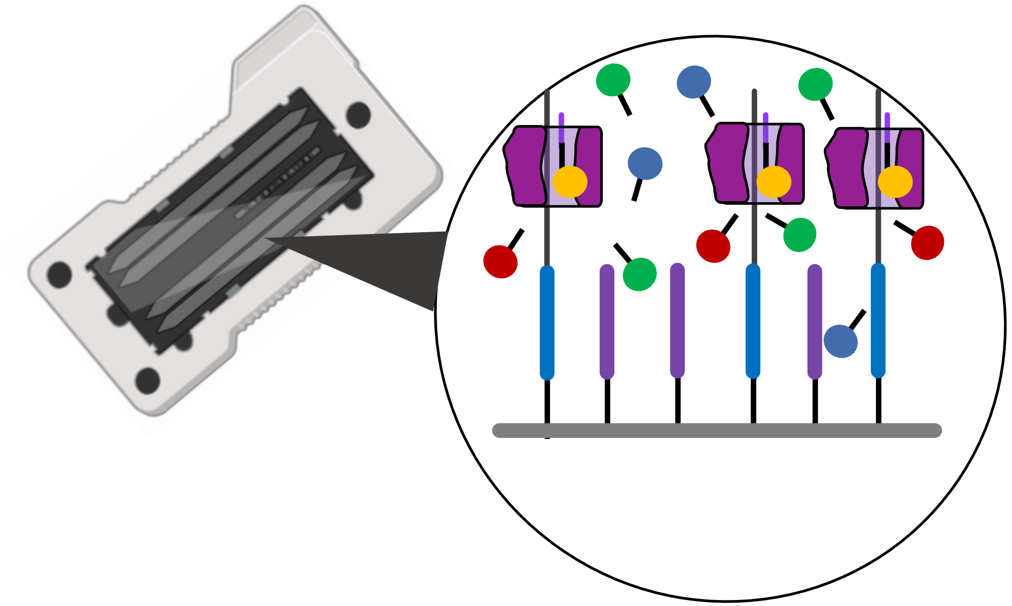
Fluorescent Nucleotides



Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

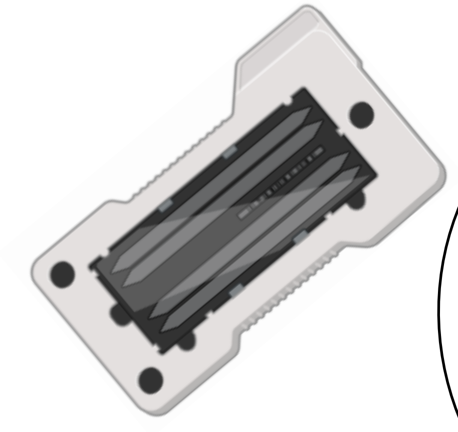
- Fluorescent Tag

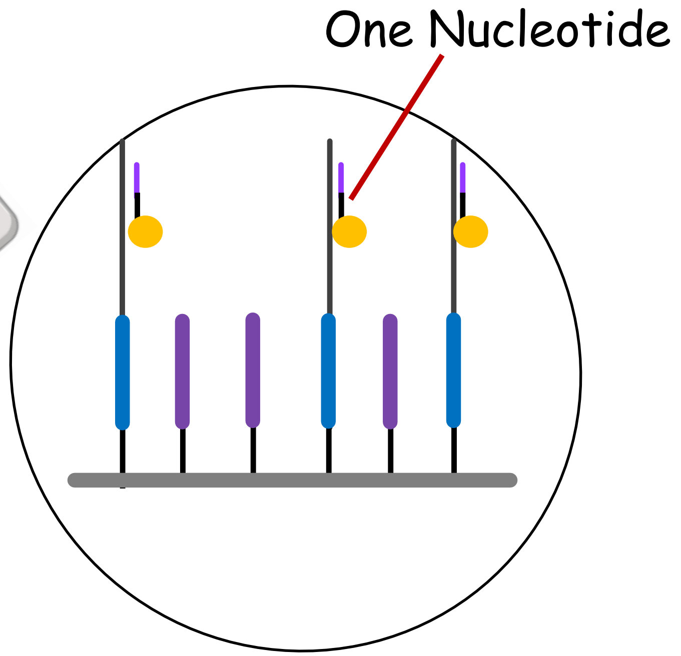


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

- Fluorescent Tag
- Terminator

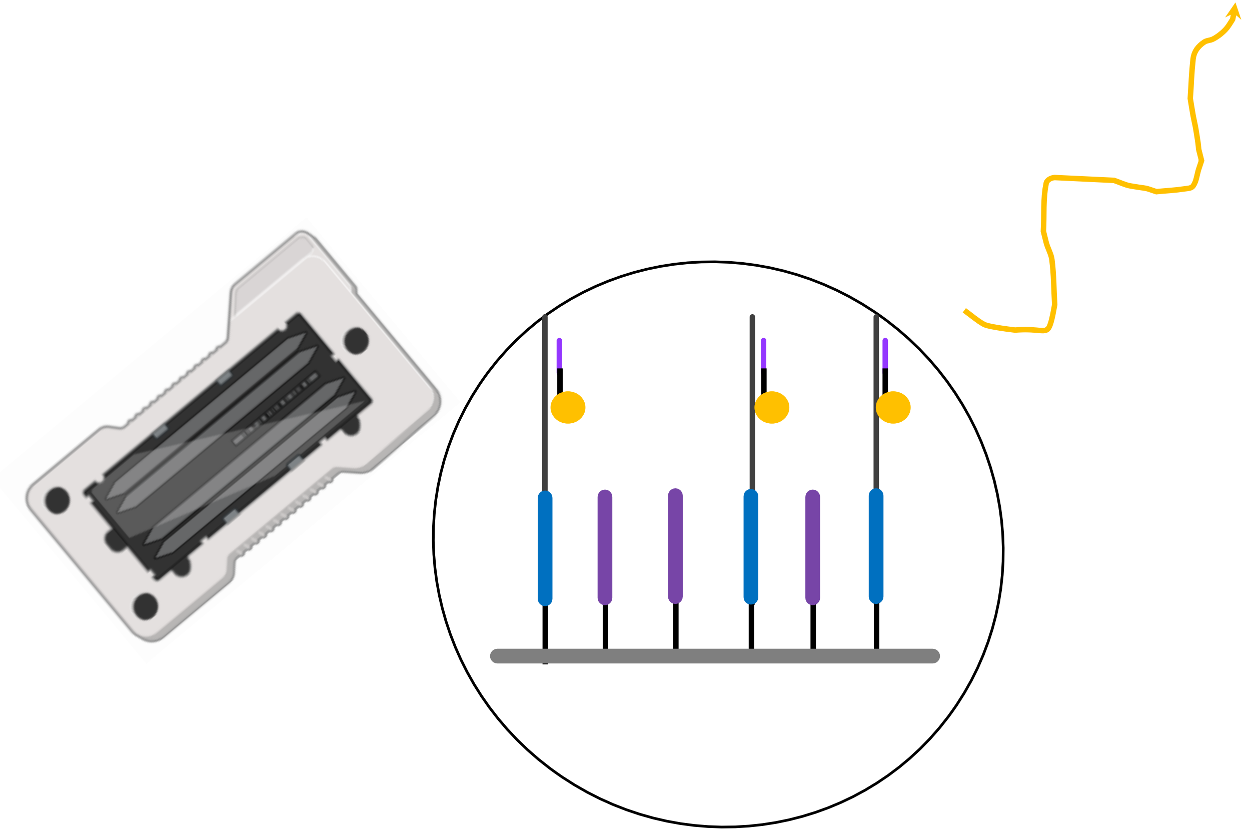




Next Generation Sequencing (NGS)

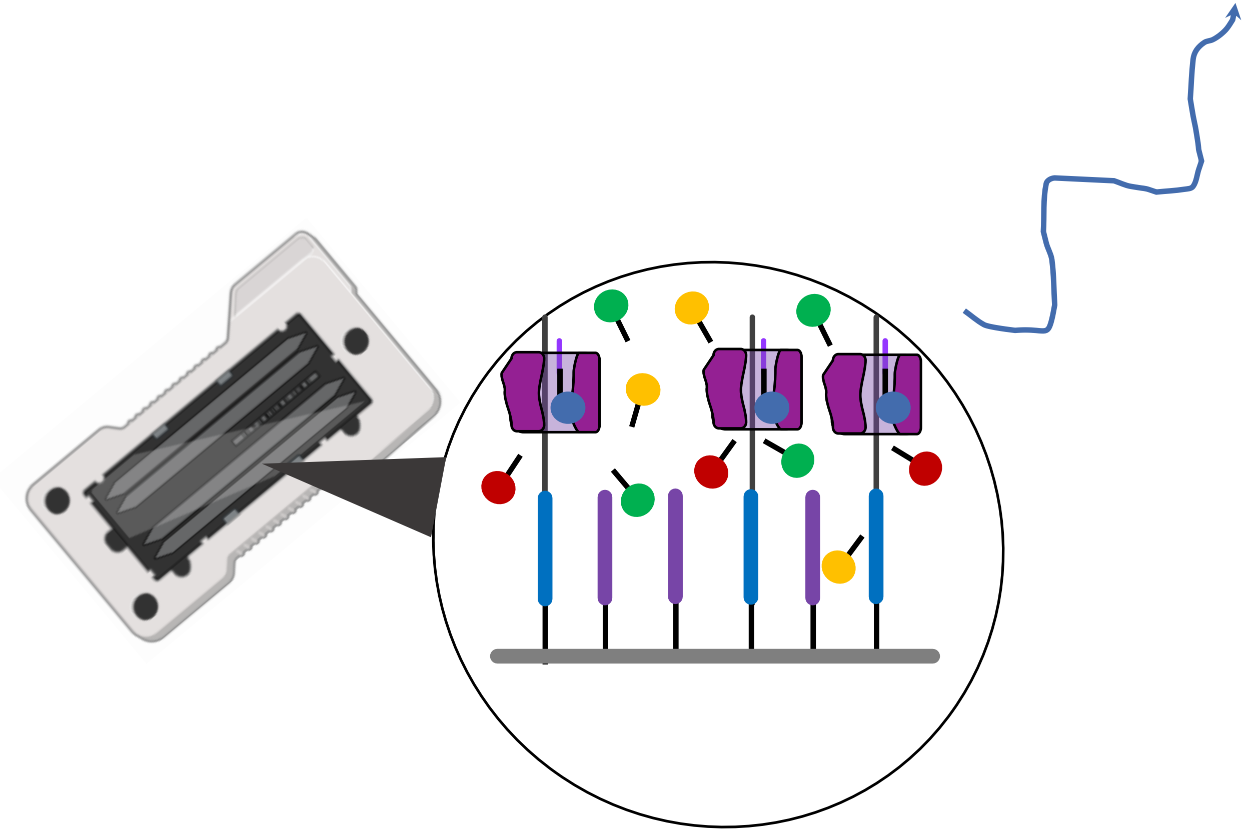
Sequencing by Synthesis (SBS)

- Fluorescent Tag
- Terminator



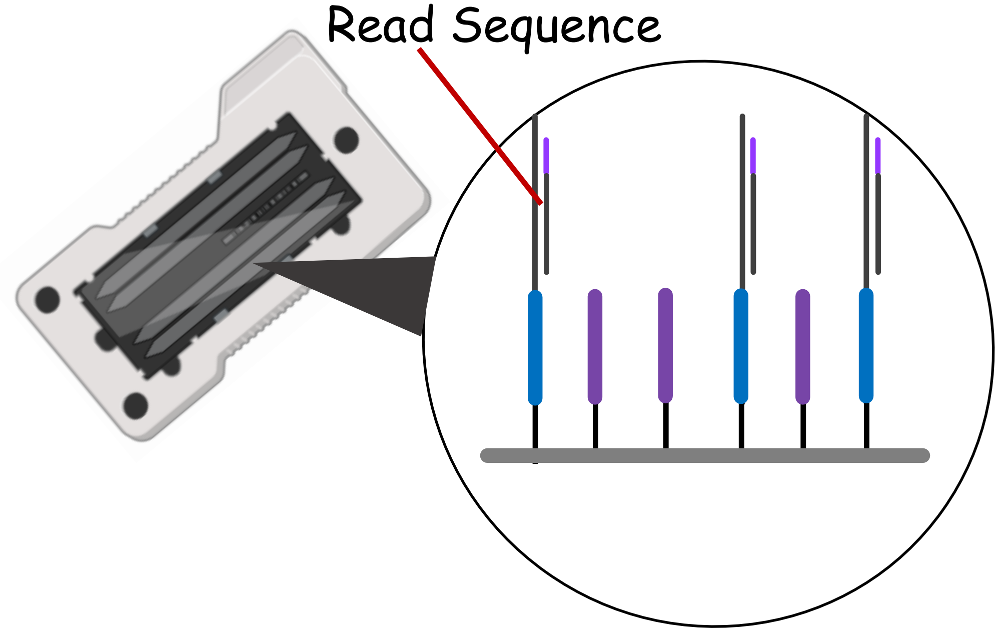
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



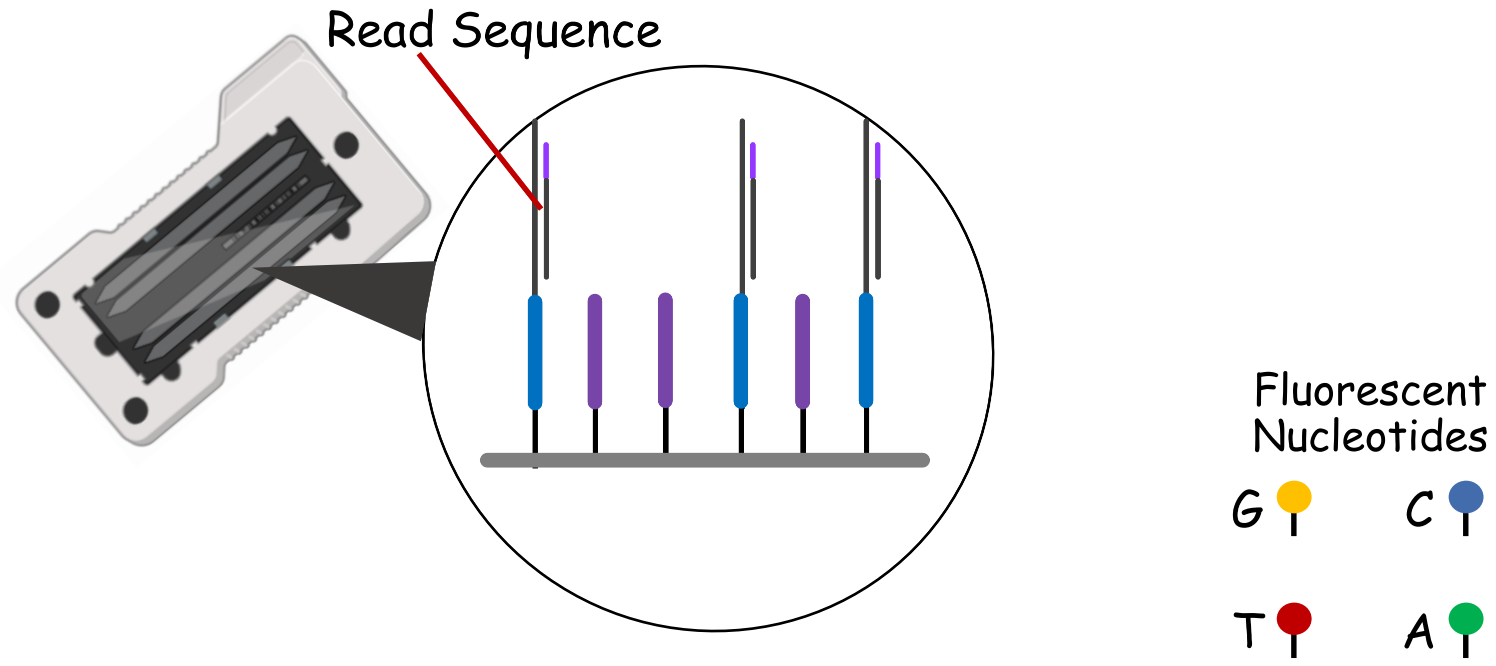
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



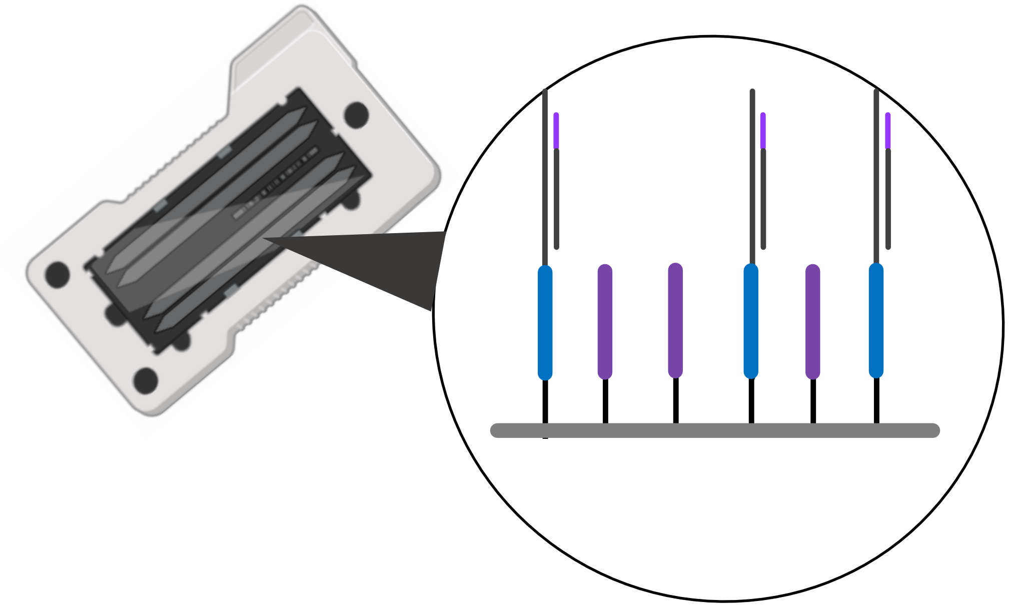
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



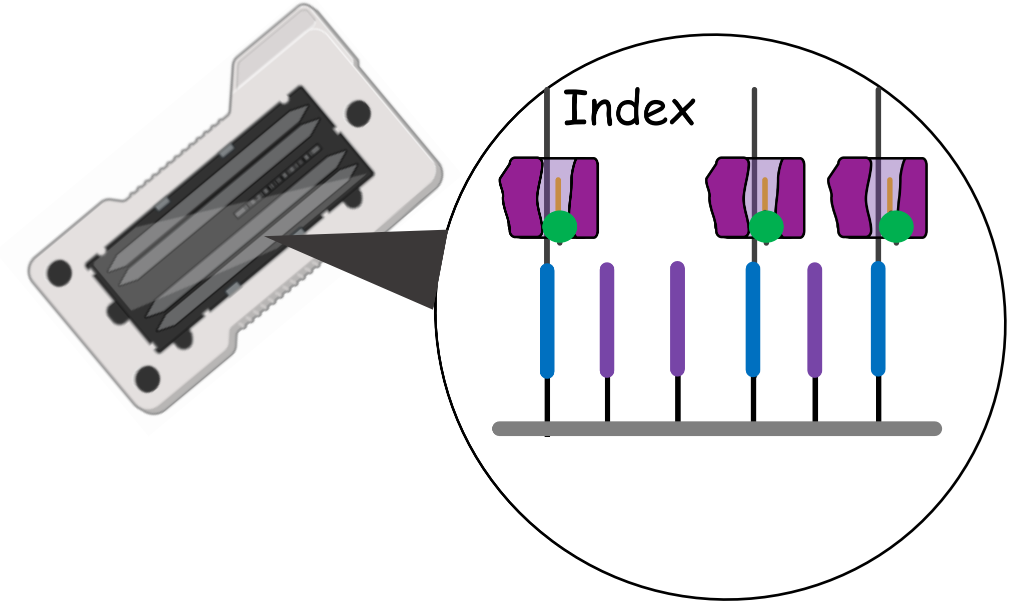
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



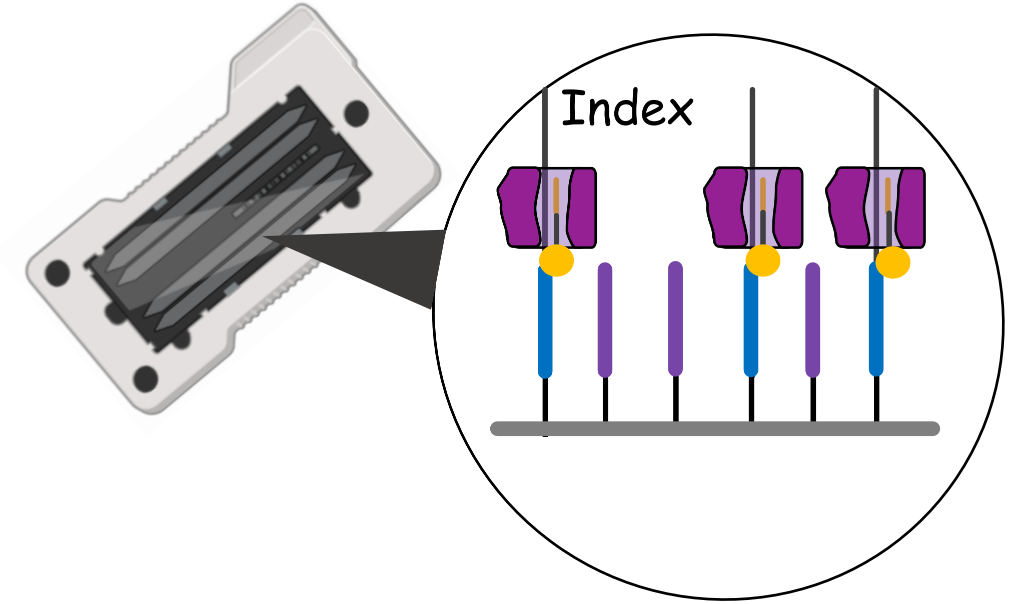
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



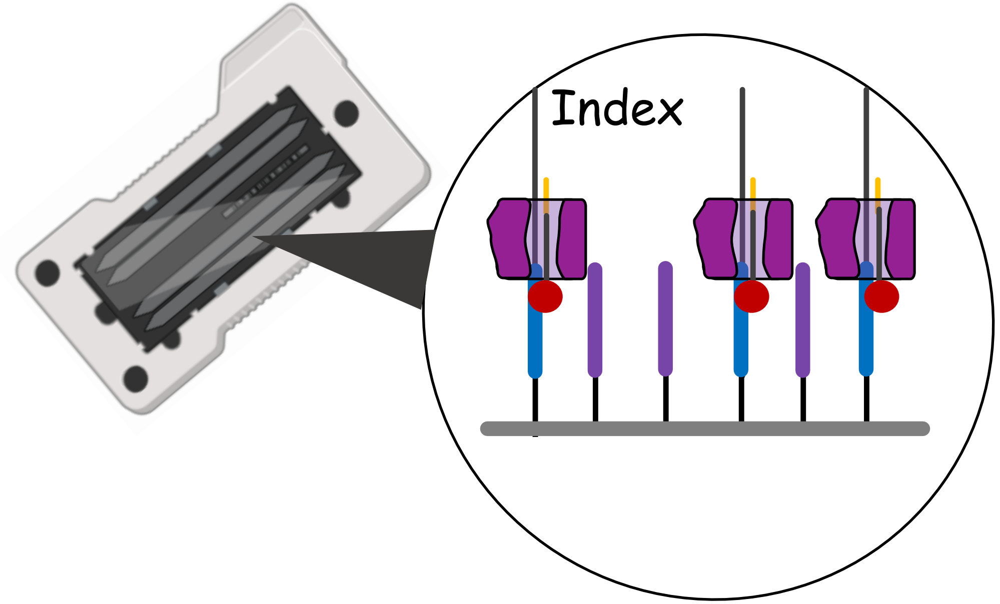
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

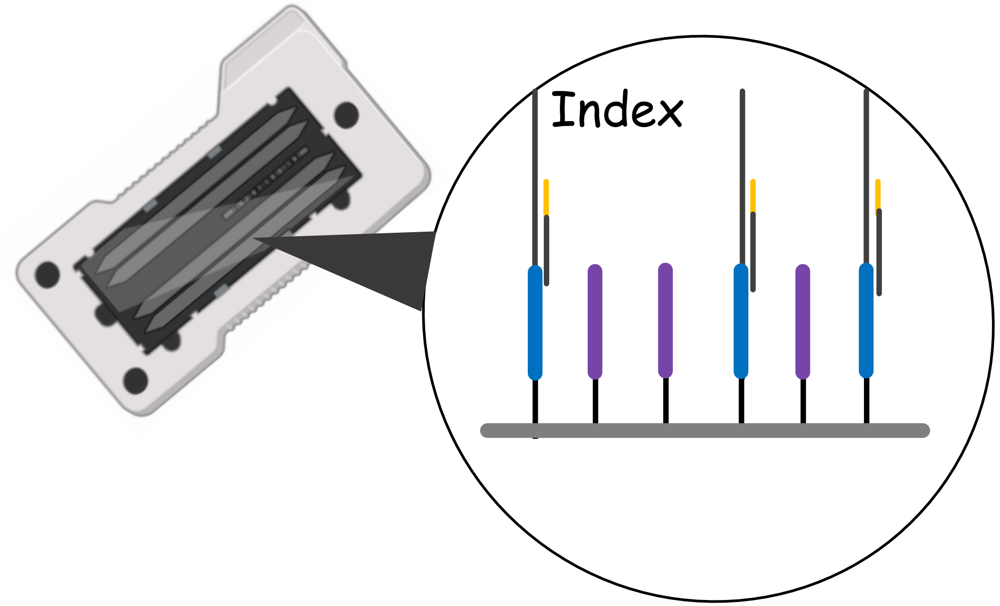


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Single Read

Sequencing Ends



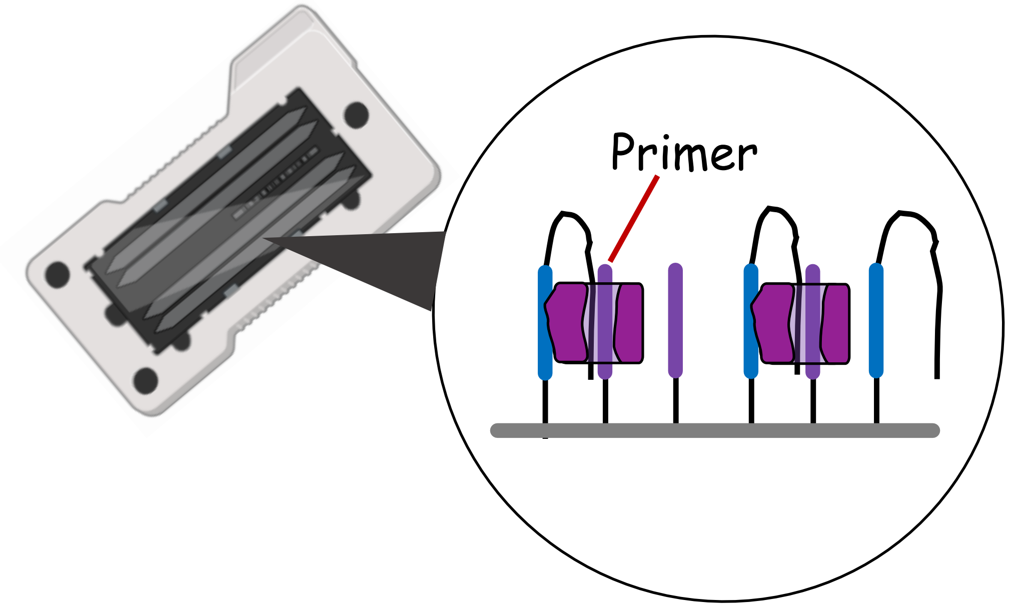
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index

Reverse Strand

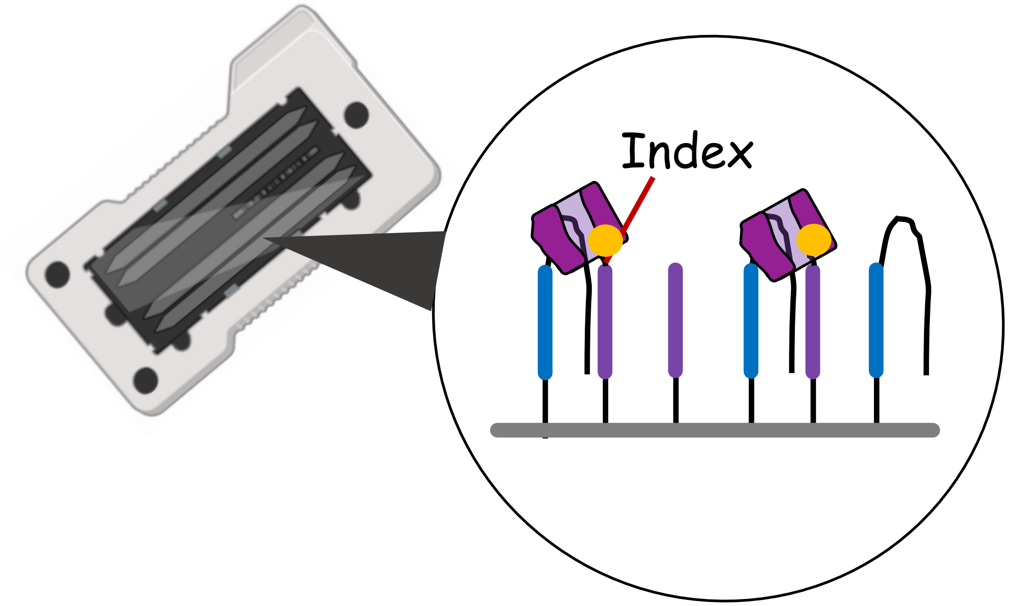


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index
Reverse Strand



Unique Dual Indexes
384 samples/flowcell

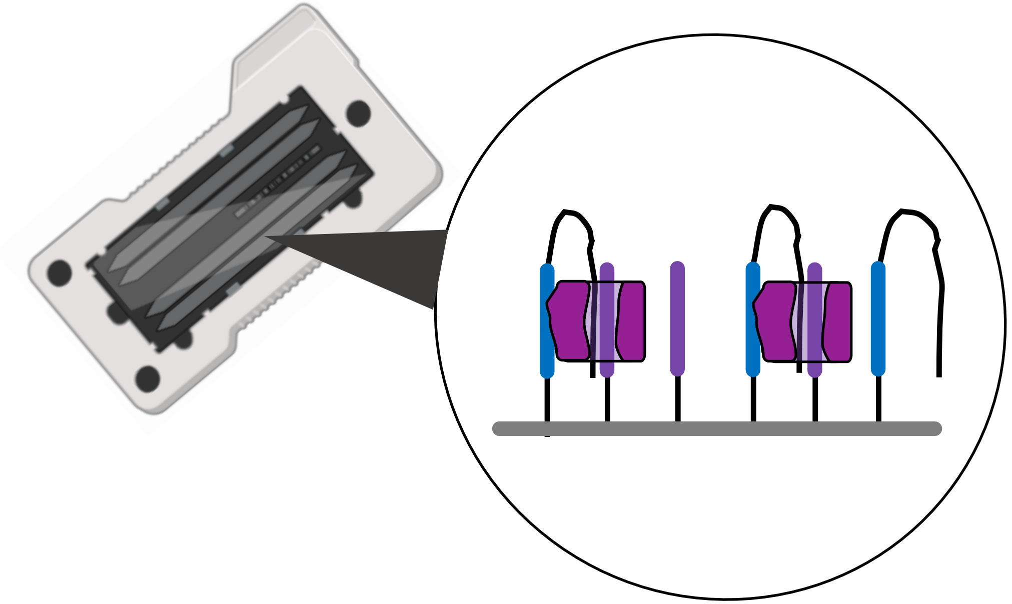
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index

Reverse Strand



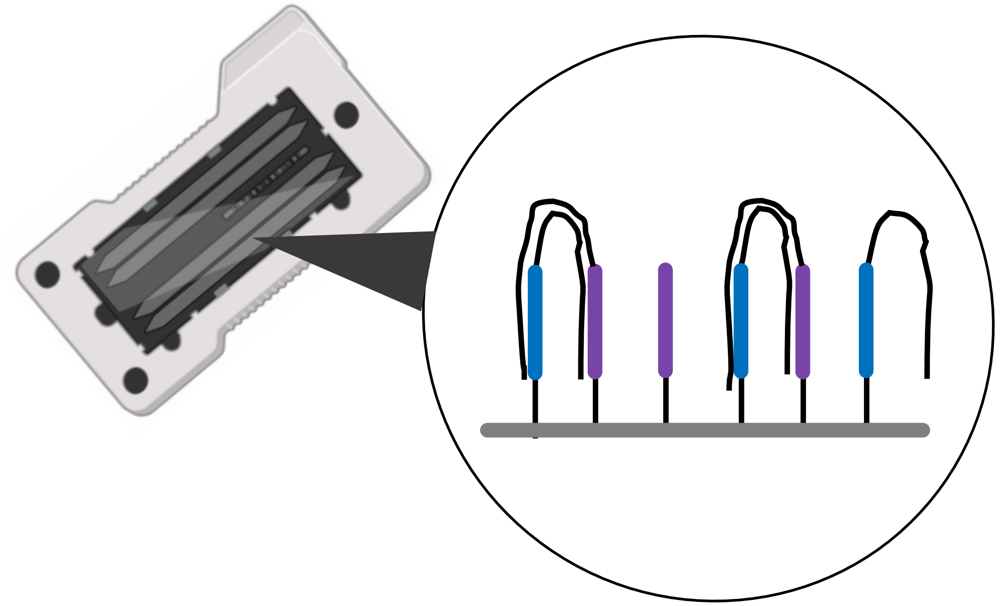
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index

Reverse Strand



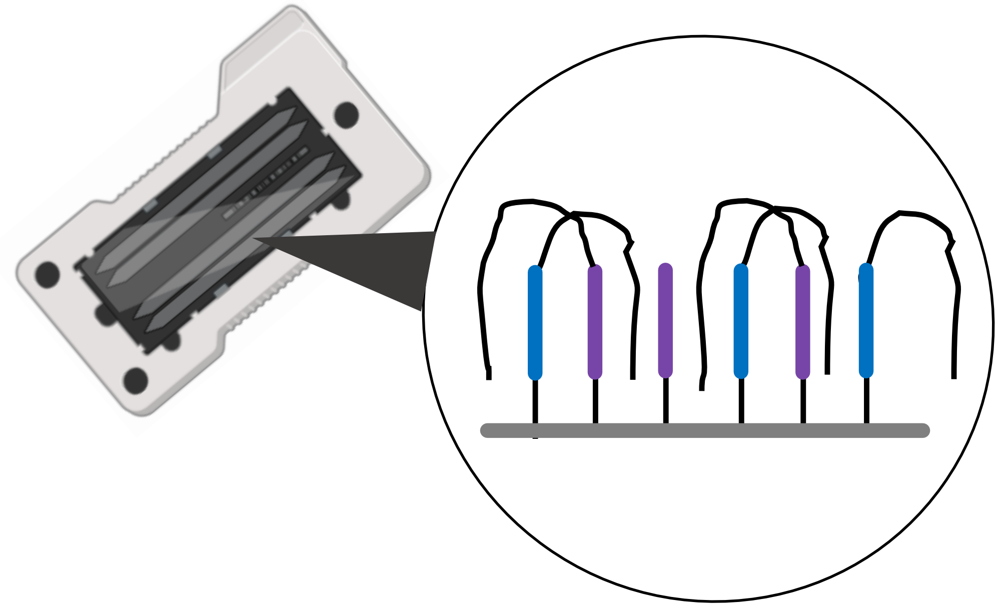
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index

Reverse Strand



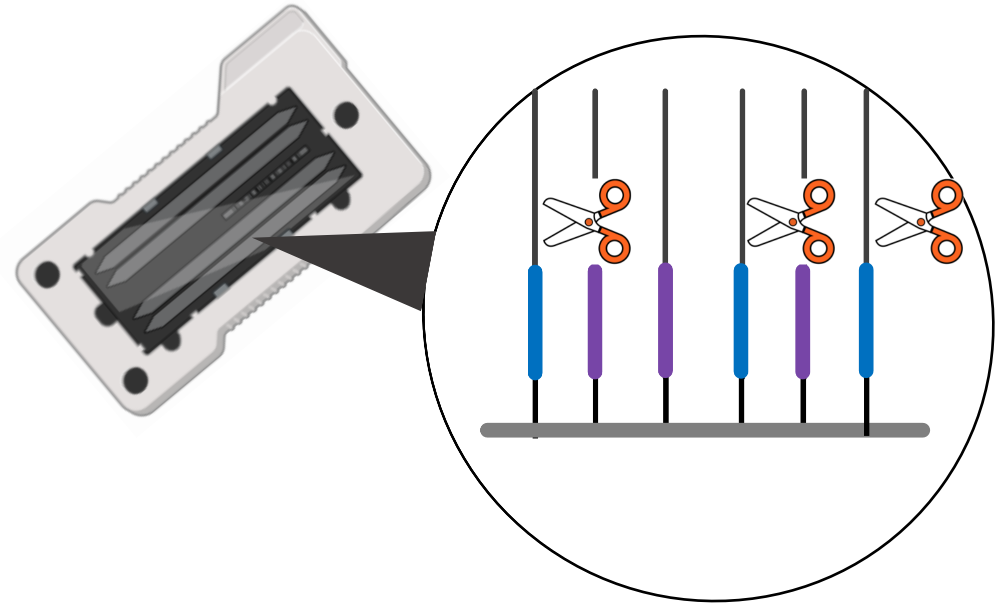
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

Second Index

Reverse Strand



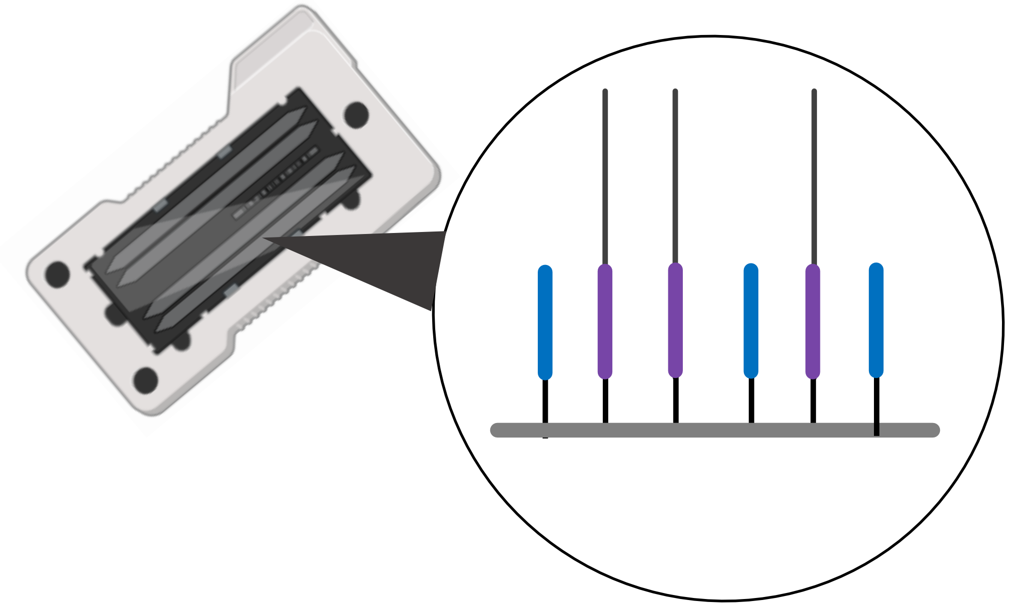
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

Paired Read

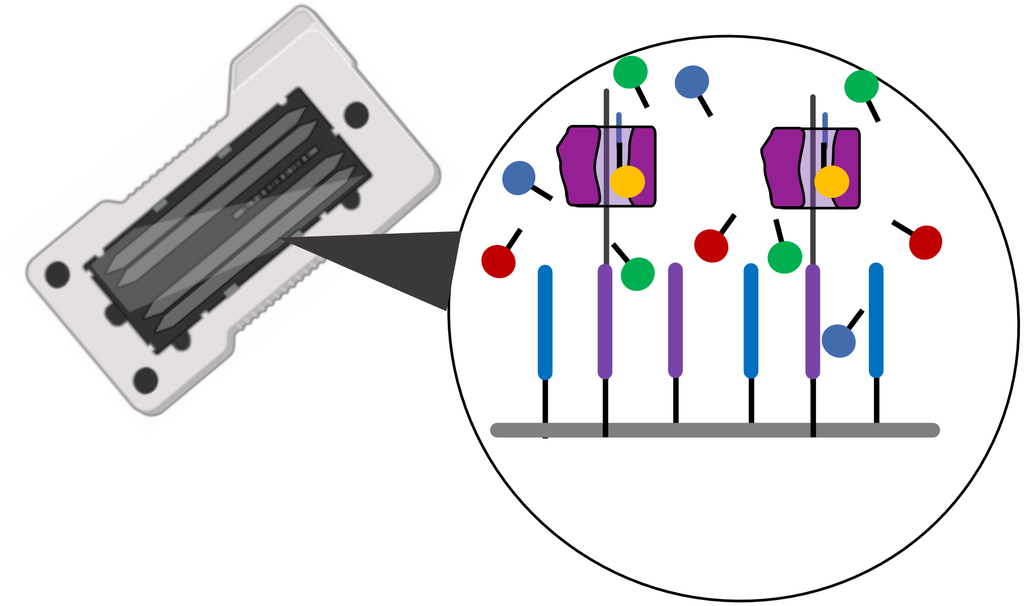
Second Index

Reverse Strand



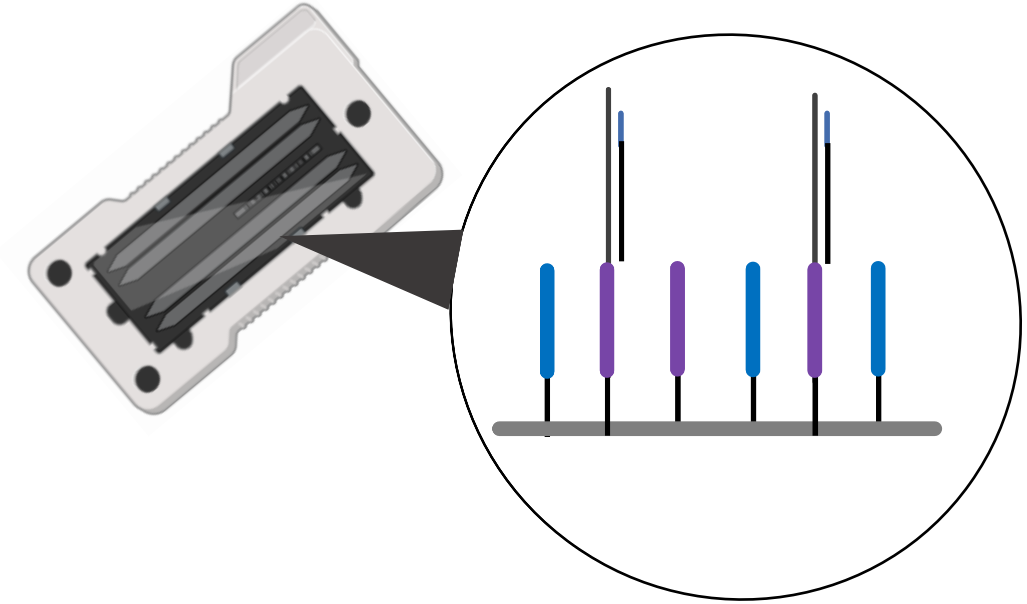
Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)

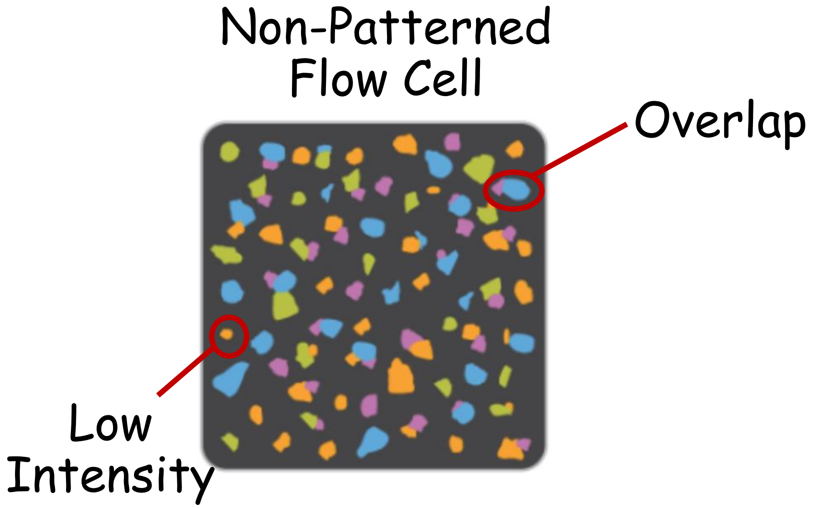


Next Generation Sequencing (NGS)

Sequencing by Synthesis (SBS)



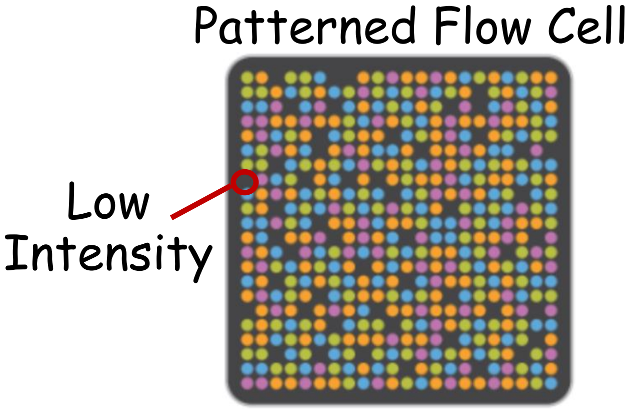
Filtering and Mapping



Non-Patterned Flow Cell

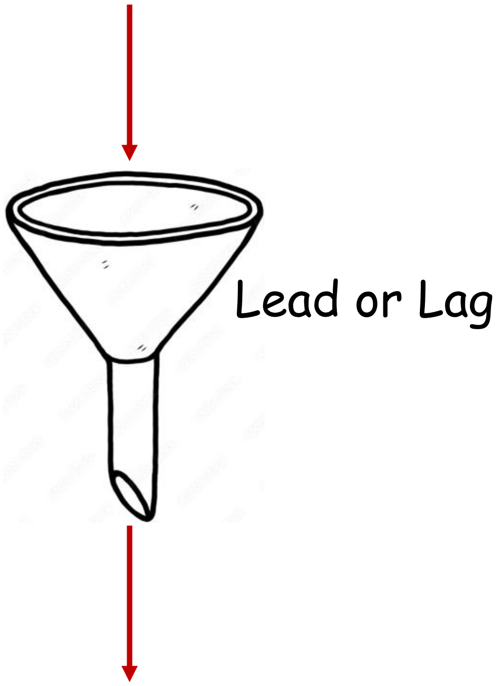
Overlap

Low Intensity



Patterned Flow Cell

Low Intensity



Lead or Lag

More Than One Library

Filtering and Mapping

TAGGCTGAATTTTTTGCTCA

AAGAGGCCATATTAGCTAGG

GGTGAAAGAGGCCATATTAG

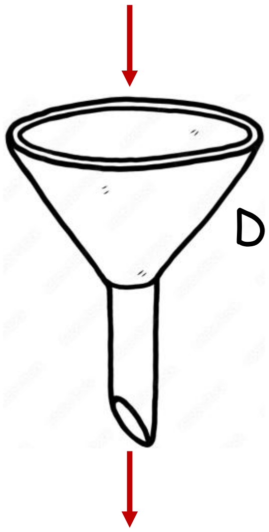
ATATTAGCTAGGCTGAATTT

CATATTAGCTAGGCTGAATT

TAGGCTGAATTTTTTGCTCA

TTAGCTAGGCTGAATTTTTTG

CATTAGCTAGGCTGAATT



Demultiplexed

Sample 1

ATATTAGCTAGGCTGAATTT
AAGAGGCCATATTAGCTAGG

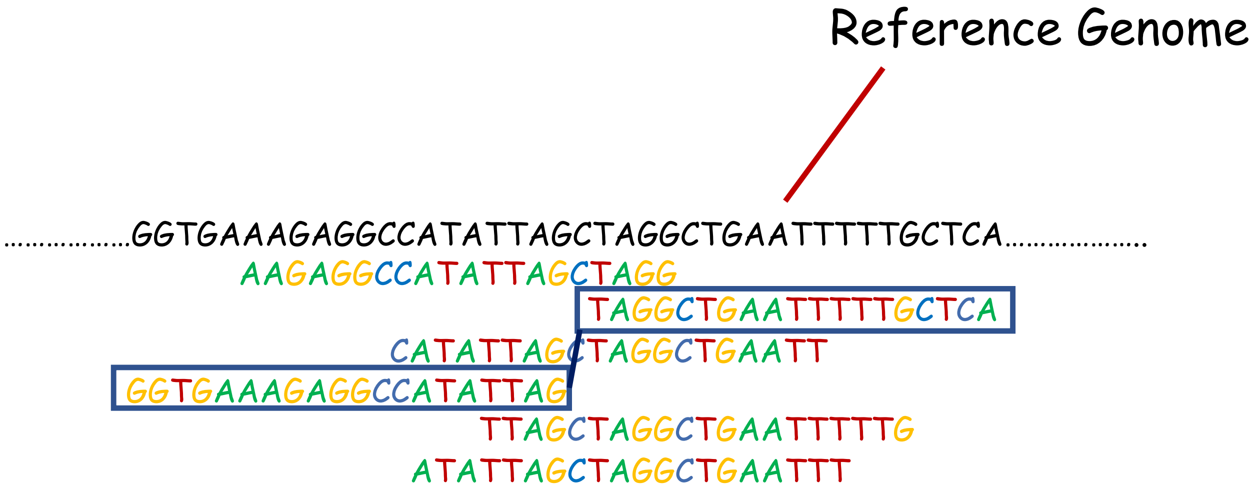
Sample 2

TAGGCTGAATTTTTTGCTCA
CATATTAGCTAGGCTGAATT

Sample 3

TTAGCTAGGCTGAATTTTTTG
CATATTAGCTAGGCTGAATT

Filtering and Mapping



Reference Genome

………………GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTTGCTCA………………..

AAGAGGCCATATTAGCTAGG

TAGGCTGAATTTTTTGCTCA

CATATTAGCTAGGCTGAATT

GGTGAAAGAGGCCATATTAG

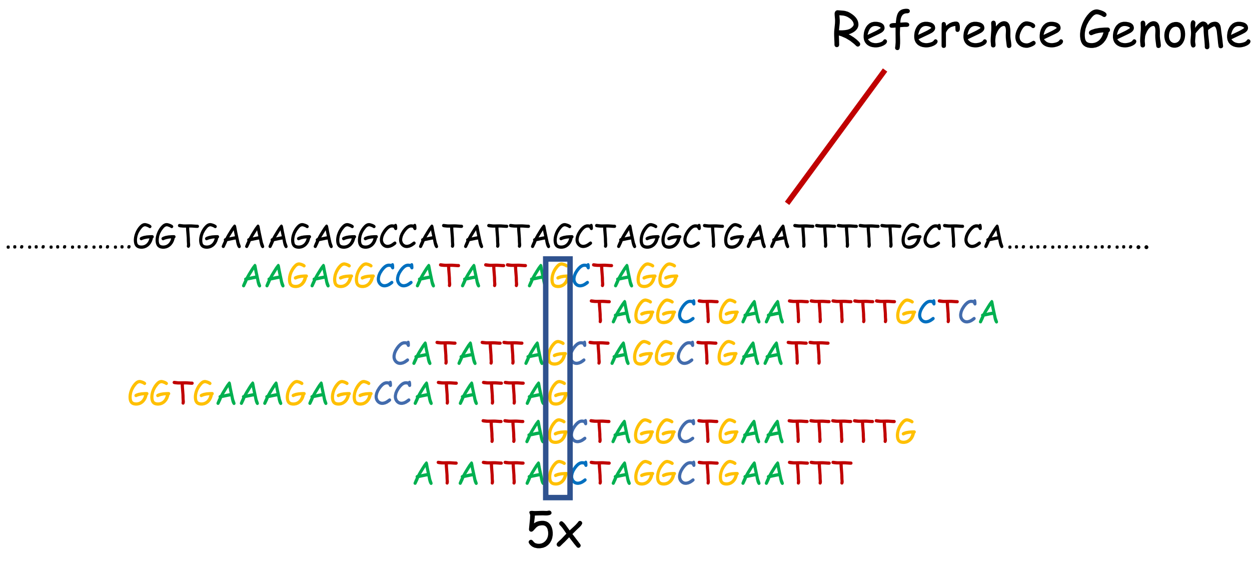
TTAGCTAGGCTGAATTTTTTG

ATATTAGCTAGGCTGAATTT

Paired End Sequencing → Longer Stretches

Paired End Sequencing → Greater Confidence

Filtering and Mapping



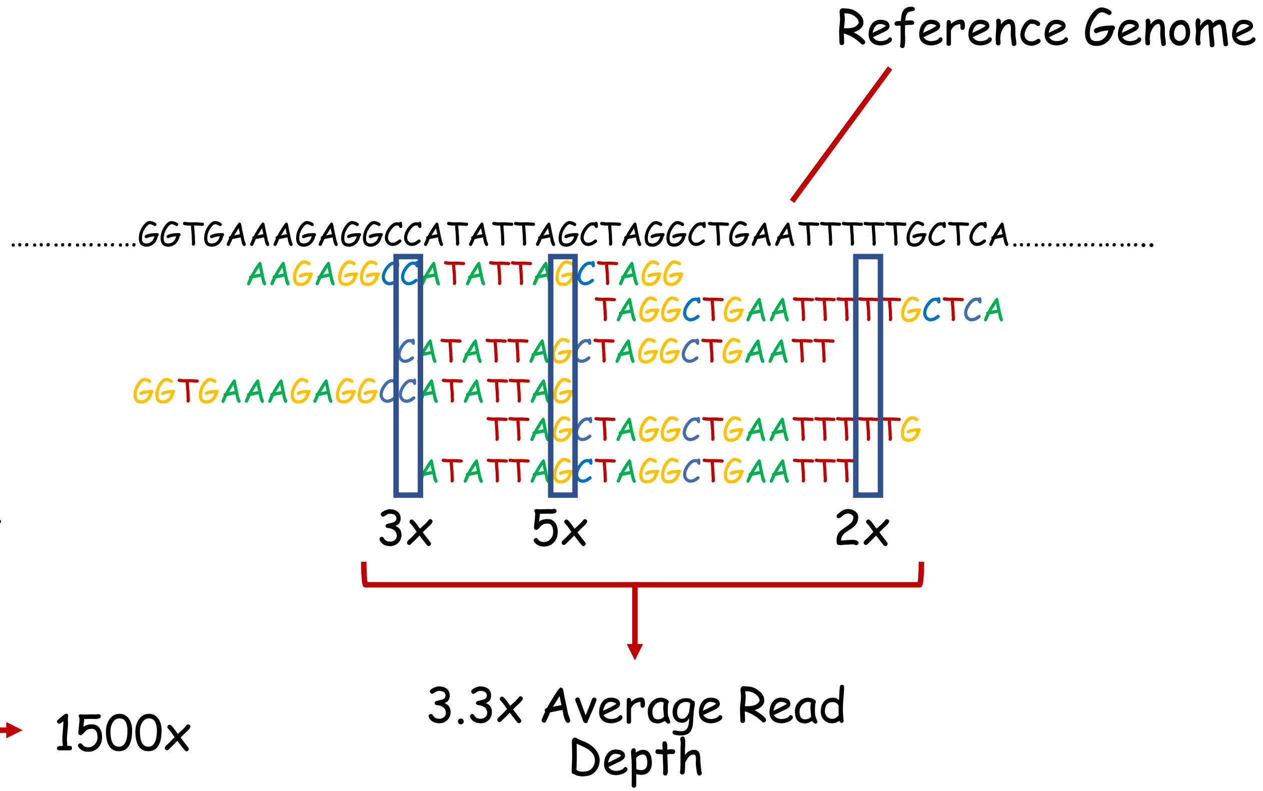
Reference Genome

```
................GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTGCTCA..................
                     AAGAGGCCATATTAGCTAGG
                                     TAGGCTGAATTTTTGCTCA
                            CATATTAGCTAGGCTGAATT
                GGTGAAAGAGGCCATATTAG
                                TTAGCTAGGCTGAATTTTTG
                             ATATTAGCTAGGCTGAATTT
```

5x

Read Depth

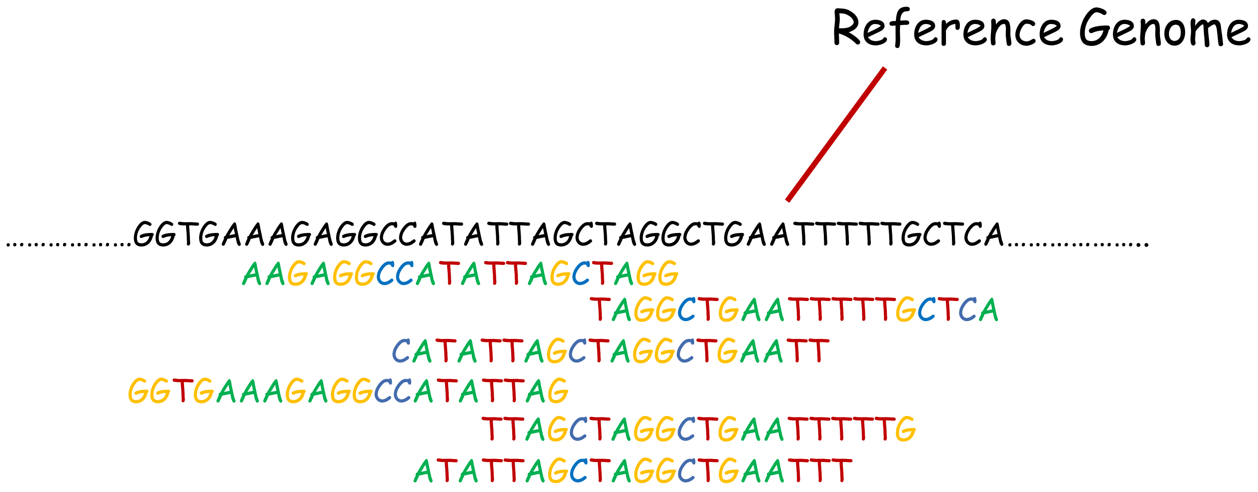
Filtering and Mapping



WGS → 30x

Cancer Rare Mutations → 1500x

Filtering and Mapping



Reference Genome

………………GGTGAAAGAGGCCATATTAGCTAGGCTGAATTTTTGCTCA……………….

AAGAGGCCATATTAGCTAGG

TAGGCTGAATTTTTTGCTCA

CATATTAGCTAGGCTGAATT

GGTGAAAGAGGCCATATTAG

TTAGCTAGGCTGAATTTTTG

ATATTAGCTAGGCTGAATTT

Coverage → Average Read Depth of a Specific Region on DNA

How is NGS Used?

Diagnosis

Cancer Rare Disease

Treatment

Guidance for Cancers/Disease

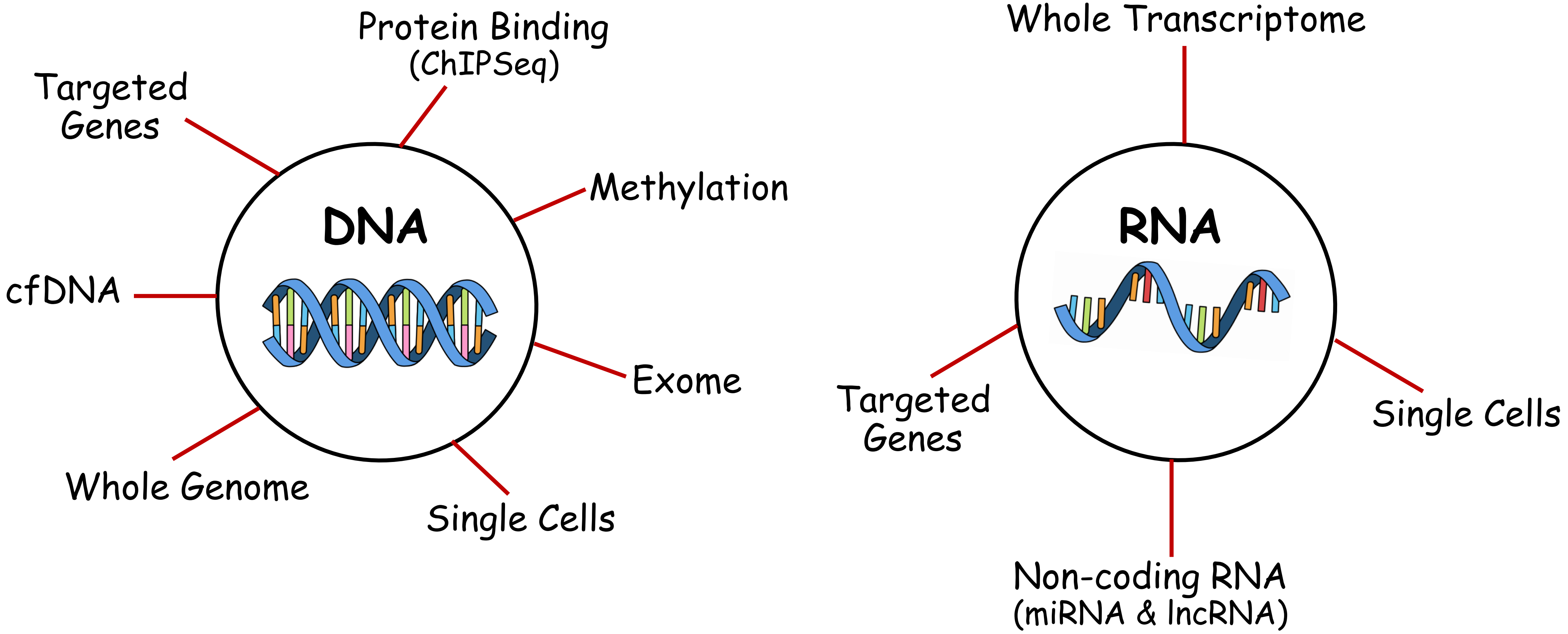
Research

Ecology

Botany

Medical Science

How is NGS Used?



DNA

Targeted Genes

Protein Binding (ChIPSeq)

Methylation

cfDNA

Exome

Whole Genome

Single Cells

RNA

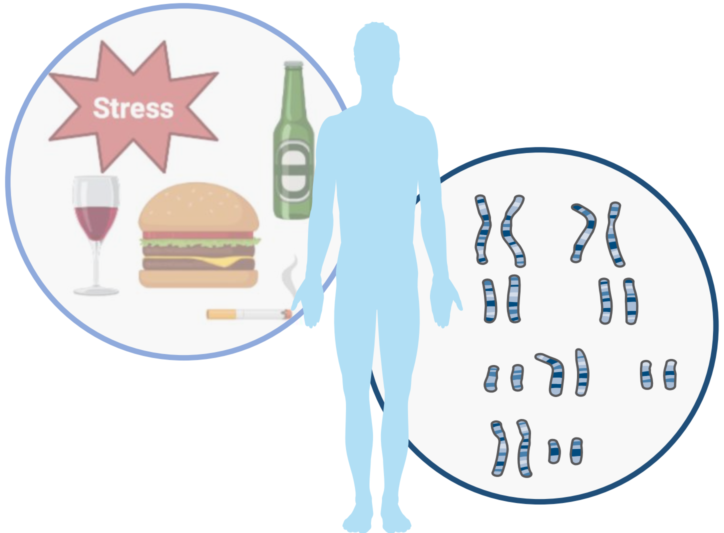
Whole Transcriptome

Targeted Genes

Single Cells

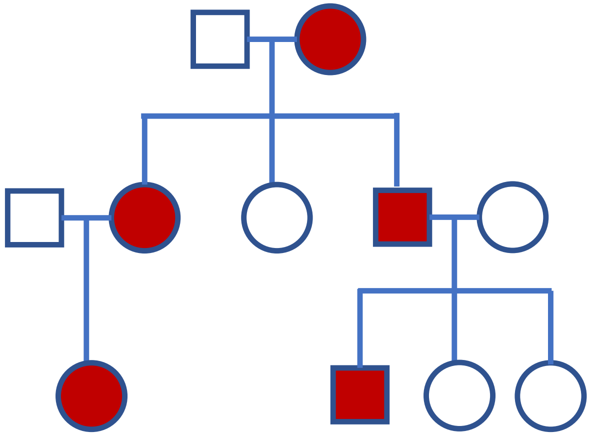
Non-coding RNA (miRNA & lncRNA)

Monogenic Diseases

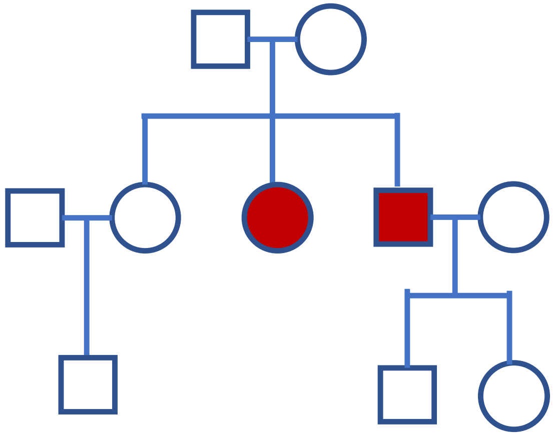


- ~ SINGLE variants in SINGLE genes
- ~ RARE
- ~ RUN in FAMILIES
- ~ MUTIPLE AFFECTED MEMBERS
- ~ EARLIER AGE of DISEASE

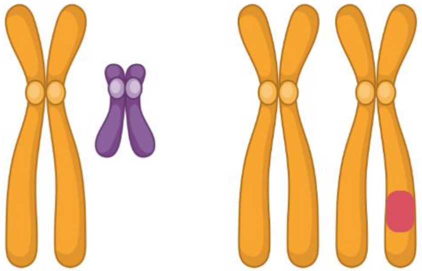
AUTOSOMAL DOMINANT



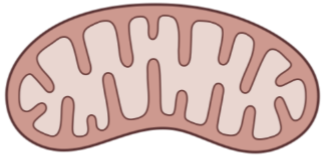
AUTOSOMAL RECESSIVE



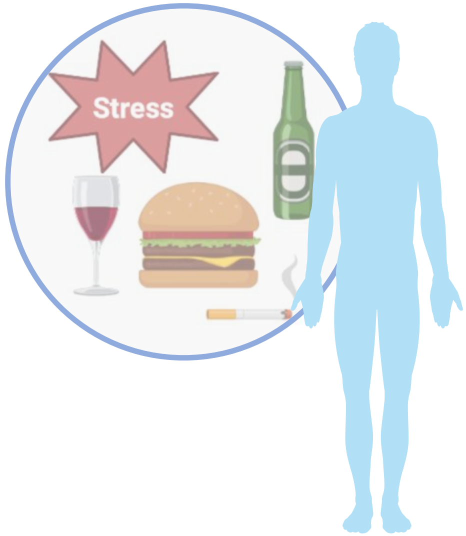
X-LINKED



MITOCHONDRIAL

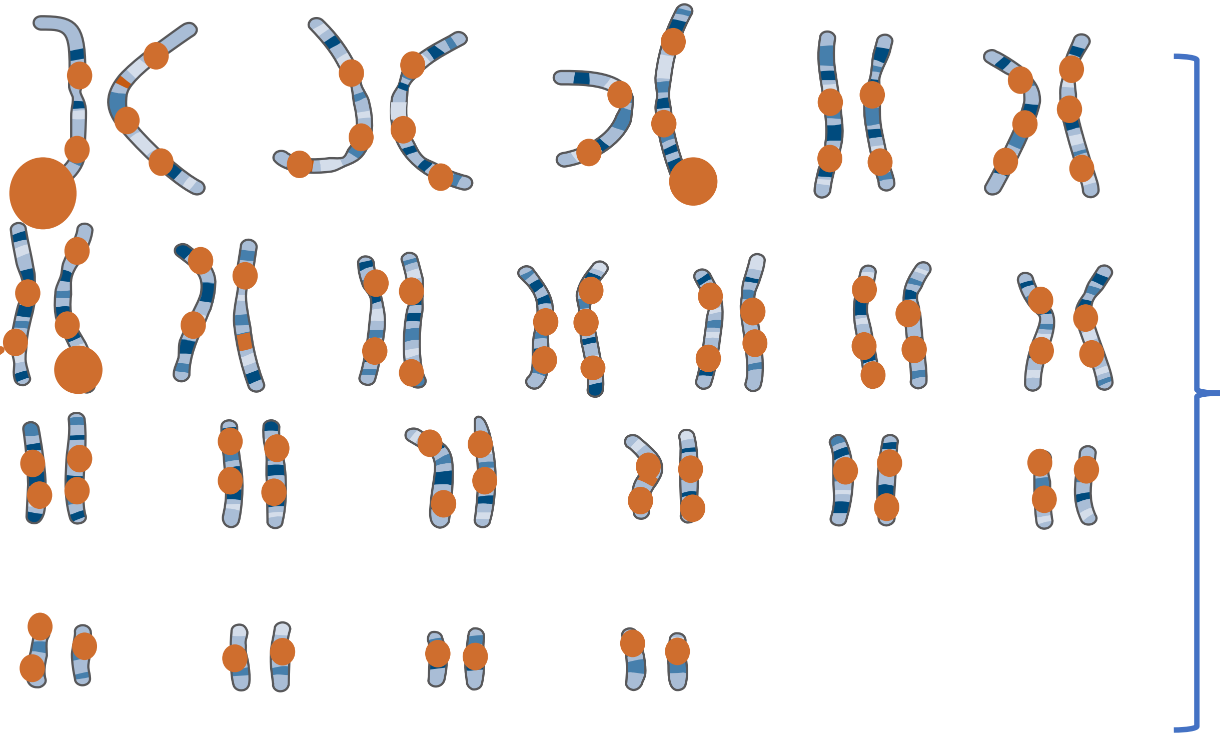


Complex conditions



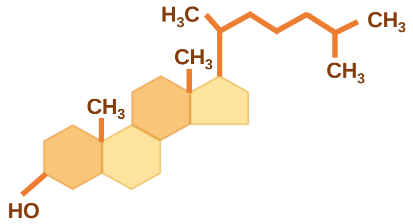
~ MOST COMMON CONDITIONS

~ MANY genes with MANY variants → POLYGENIC



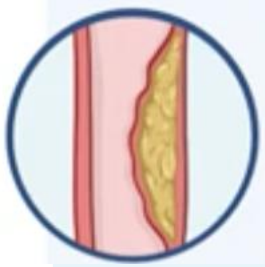
Individual Very SMALL Effect

CUMULATIVE impact on Risk



e.g. High Cholesterol levels

*FAMILIAL hypercholesterolemia



↑↑↑ LDL

&

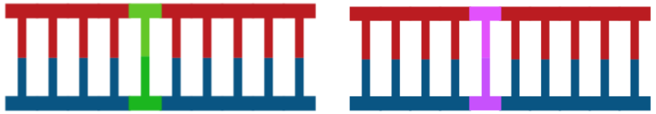
Heart Disease

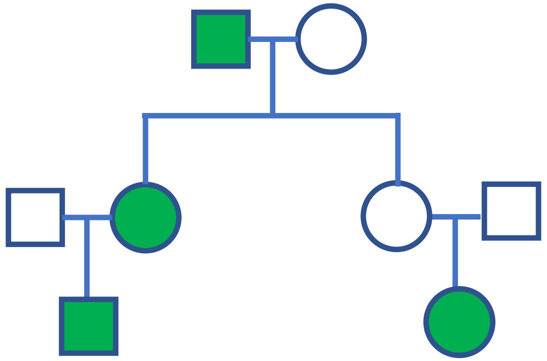
~ MONOGENIC CONDITION

Single variant in one of several genes

~ AUTOSOMAL DOMINANT

↑Risk



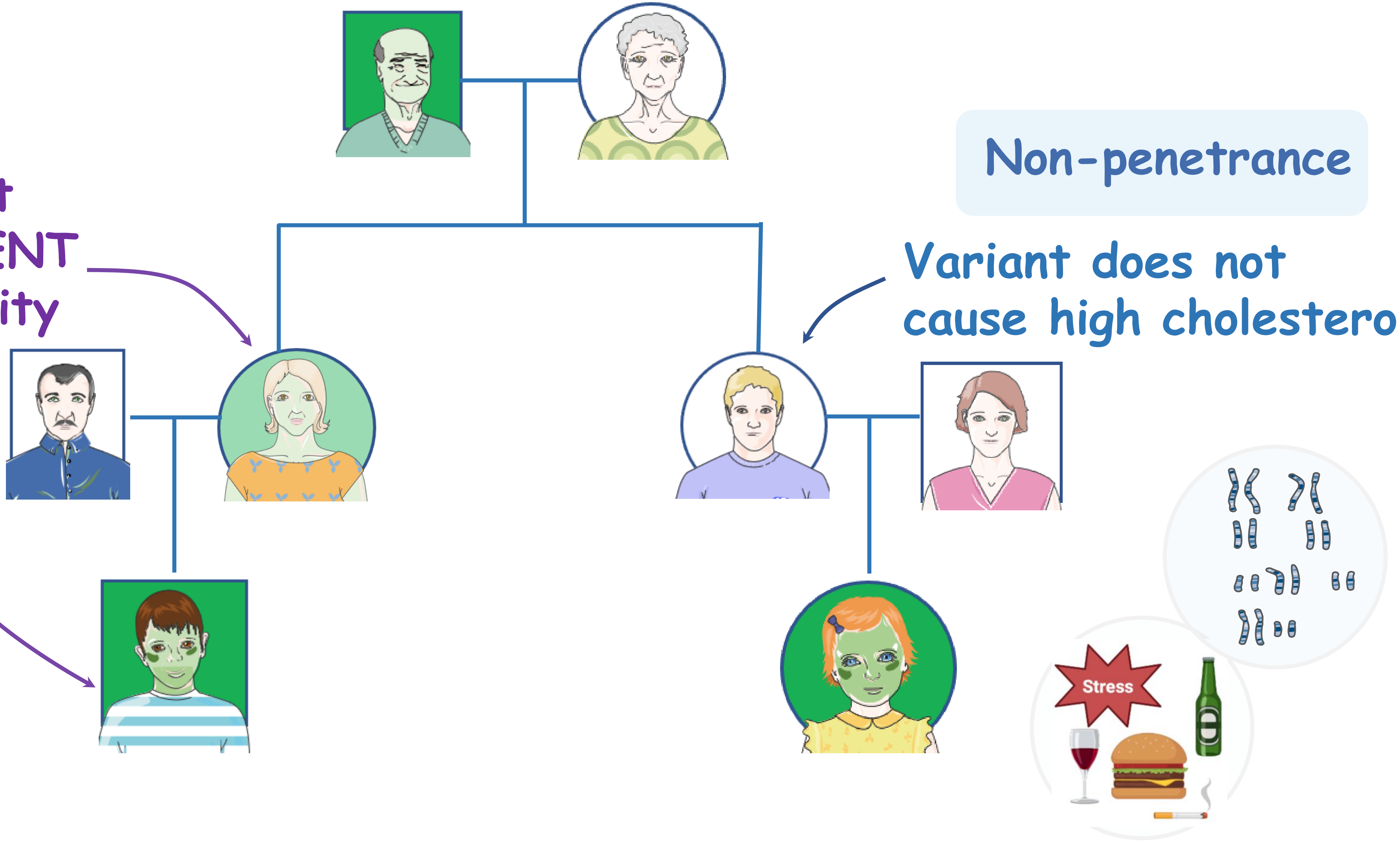


DIFFERENT families

DIFFERENT variants

VARIABLE Expressivity

SAME variant causes DIFFERENT levels of Severity



Non-penetrance

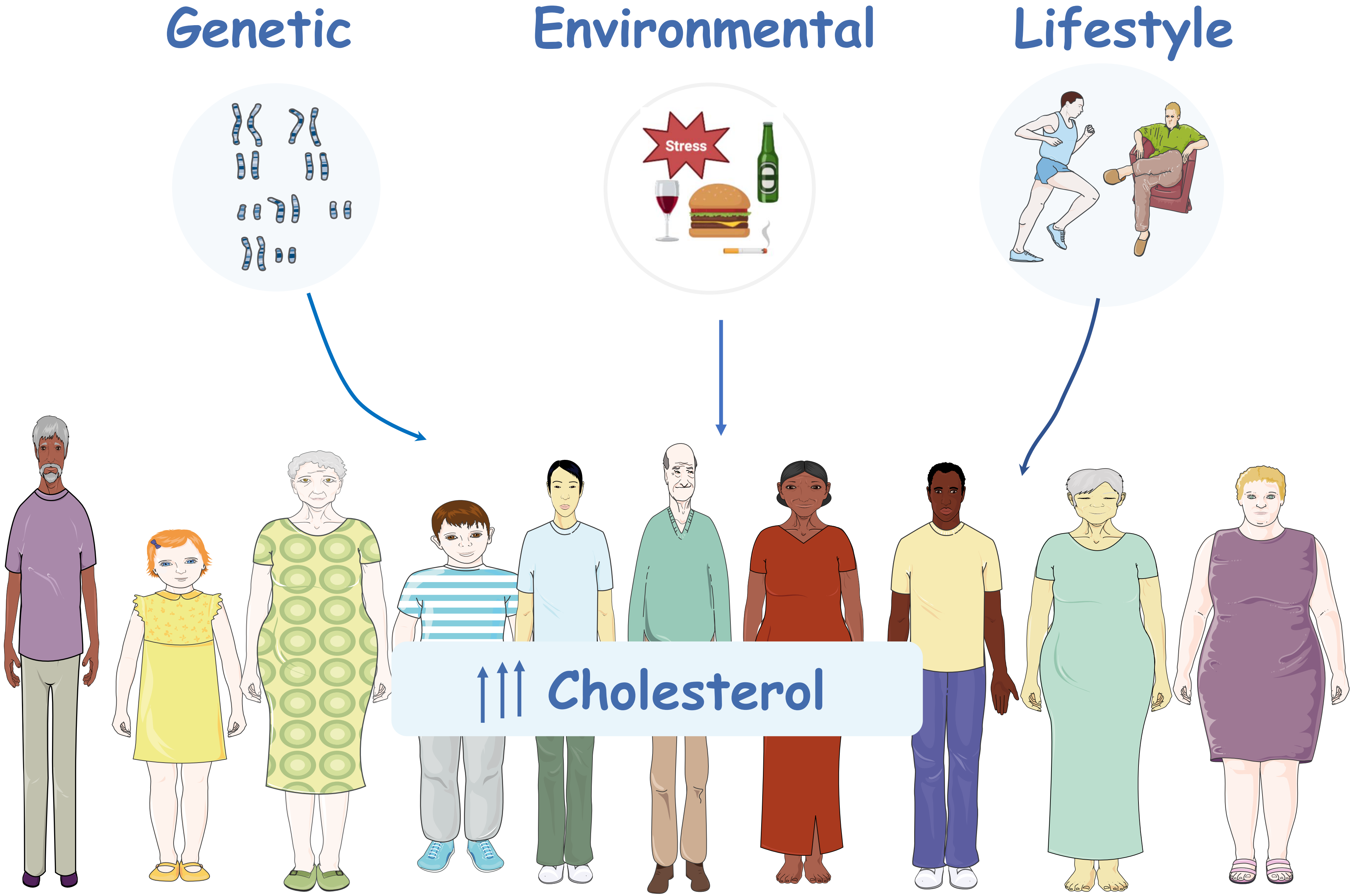
Variant does not cause high cholesterol

Monogenic Diseases

- ✓ Variants in a single gene are enough to cause disease

- ✓ Same variant passed down through a family but can be expressed differently in different members

- ✓ Same disease can be caused by different variants in different genes in different families



Genetic

Environmental

Lifestyle

↑↑↑ Cholesterol

Complex Conditions

✓ Cluster in families

✗ Don't follow an specific inheritance pattern

✓ Have a genetic component → Polygenic (100s to 1000s of variants in many genes)

Genome-wide Association Studies (GWAS)

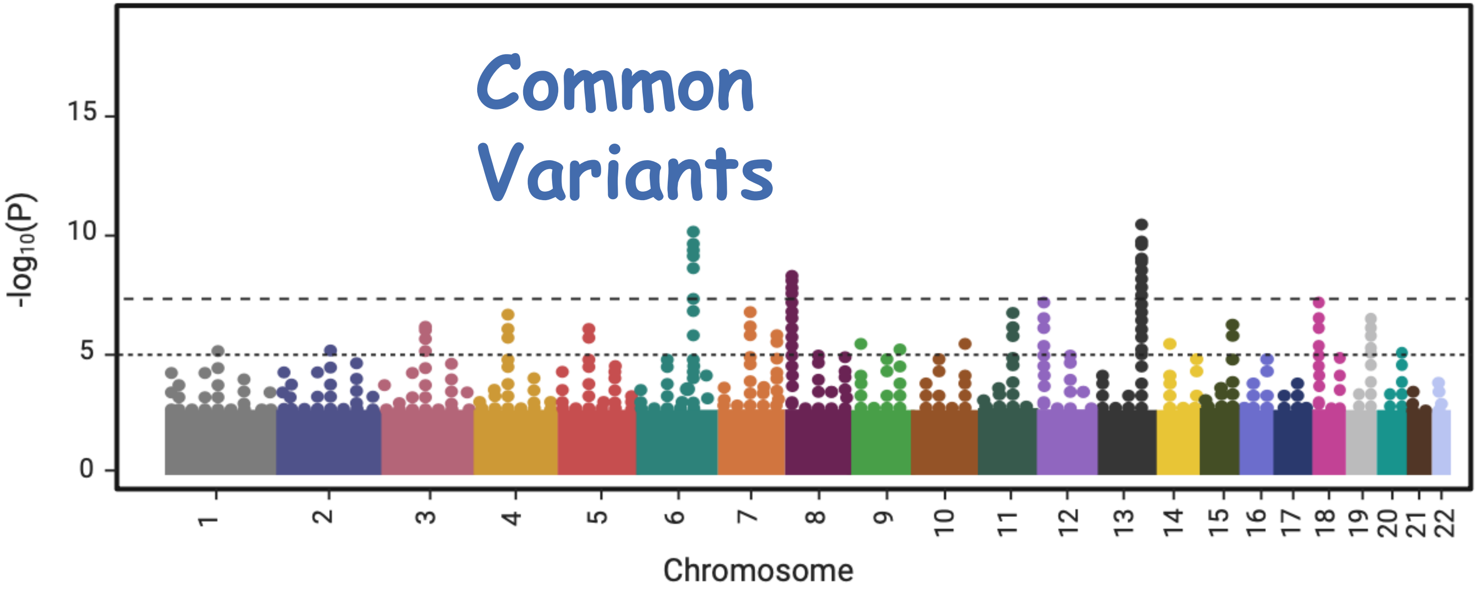
* Determines relative contributions of variants
*

e.g. high cholesterol

e.g. normal cholesterol

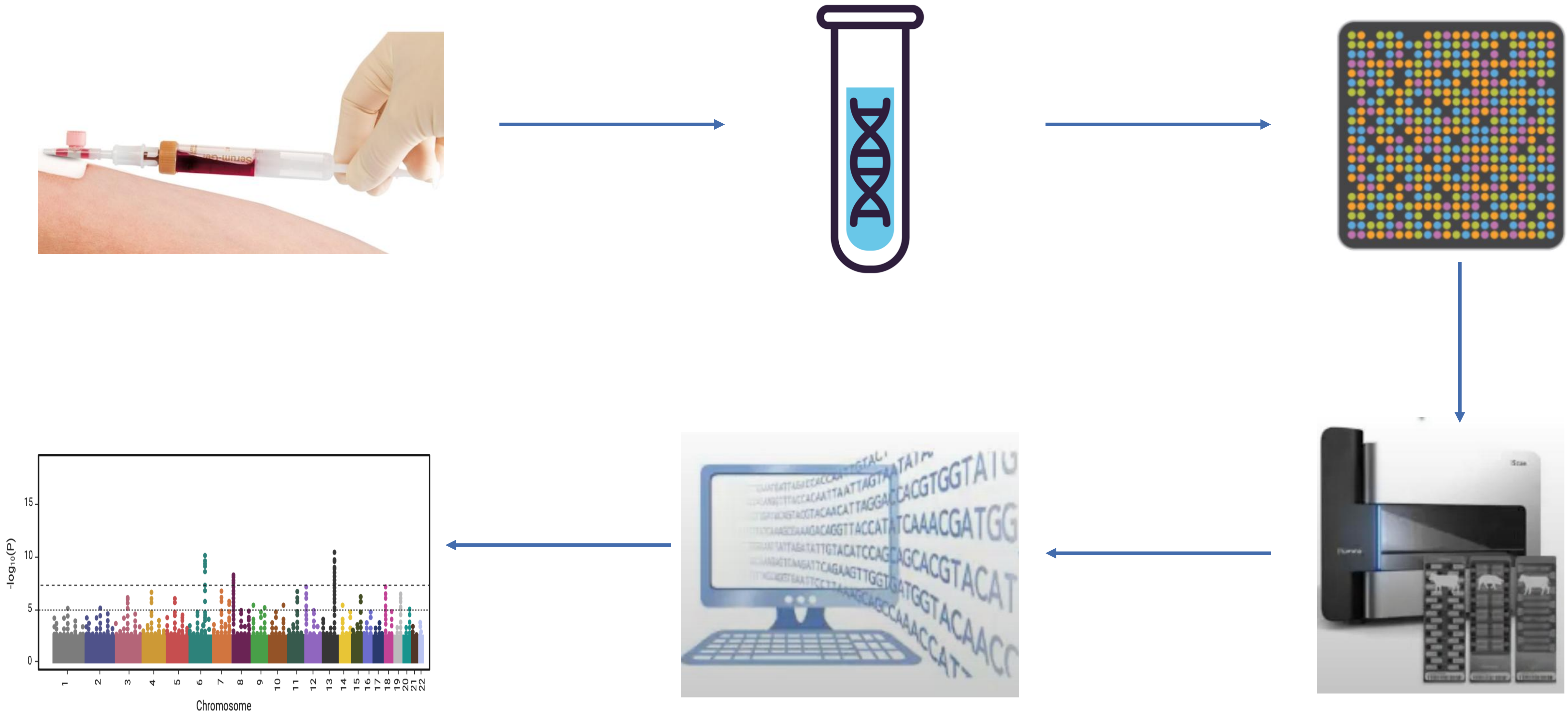
Polygenic Score
(or relative risk) → ✓

~ Encourage healthy choices
~ Early screening
~ Close monitoring



Common
Variants

GWAS Data Generation



Plink format 1

- FAM file – one row per individual

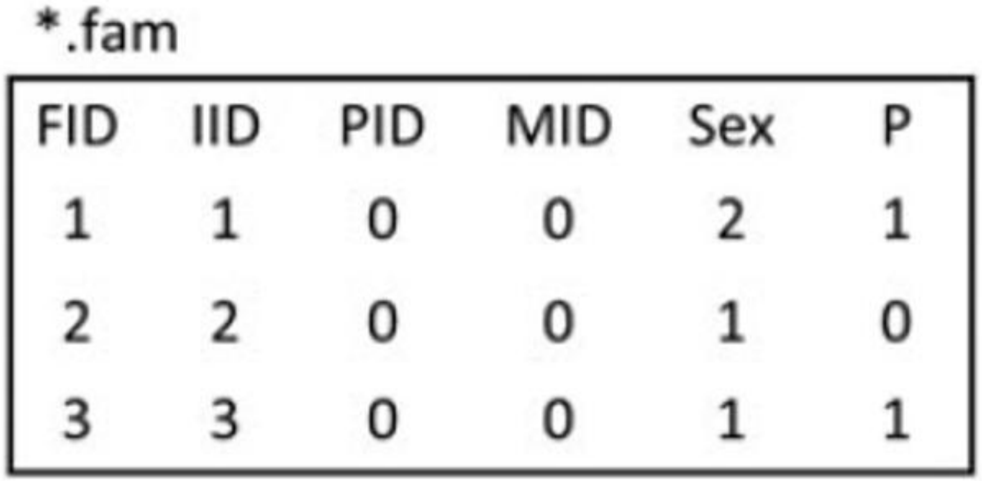
- BIM file – one row per SNP

- BED file – one row per individual – genotype calls for each individual for all SNPs – binary format

- FAM and BIM fie are human readable while BED file is not

```
‘plink –file raw-GWA-data –make-bed –out raw-GWA-data’
```

FAM file – one row per individual

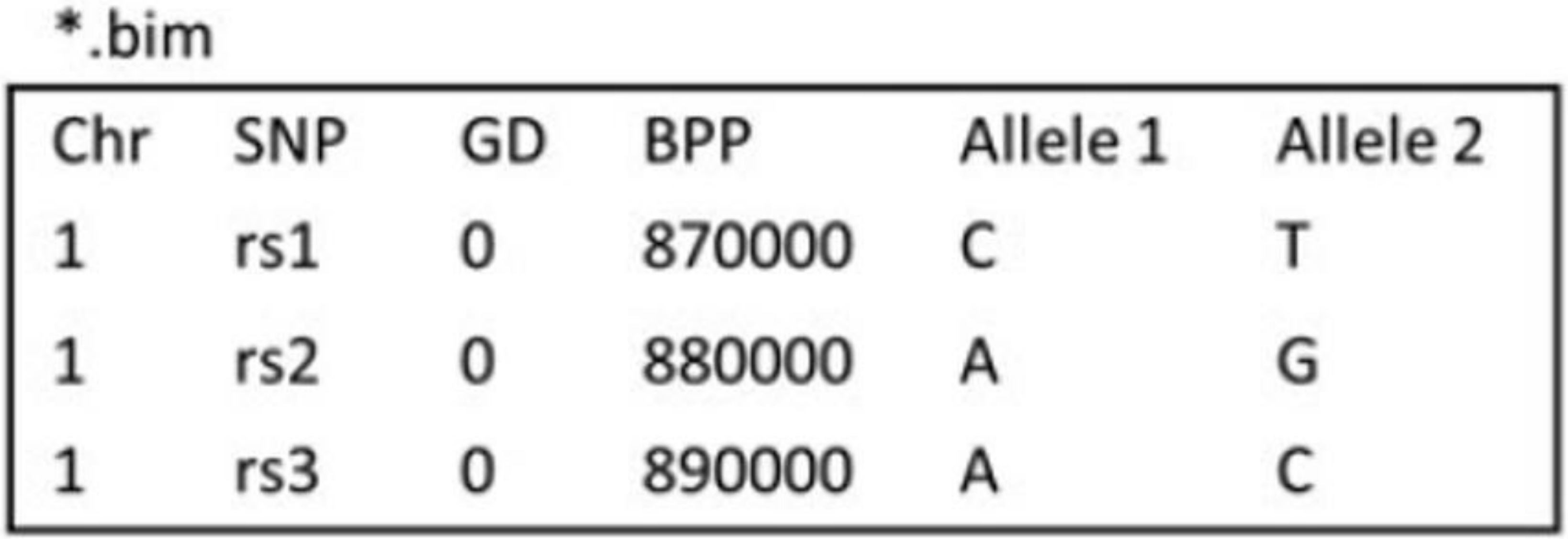


*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

1. FID: Family iD
2. IID: Within-family ID (cannot be '0')
3. PID: Within-family ID of father ('0' if father isn't in dataset)
4. MID: Within-family ID of mother ('0' if mother isn't in dataset)
5. Sex: '1' = male, '2' = female, '0' = unknown)
6. P: Phenotype, '1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control

BIM file – one row per SNP



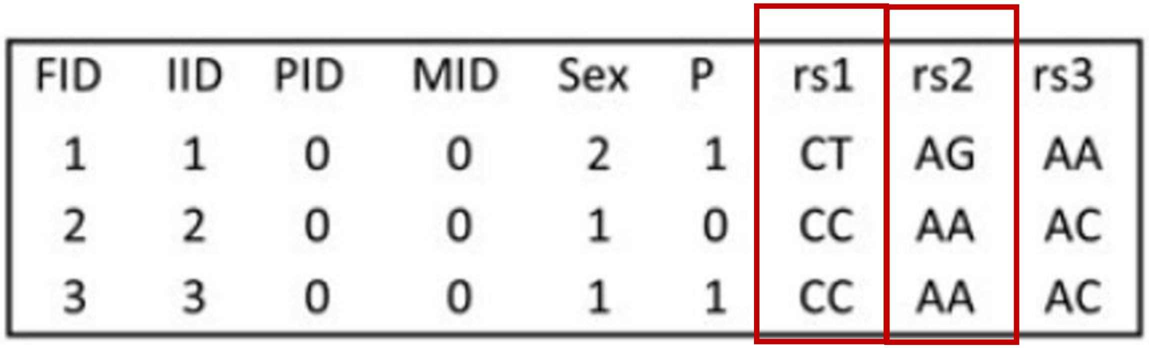
*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

1. Chr: Chromosome code (either an integer, or 'X'/'Y'/'XY'/'MT'; '0' indicates unknown) or name
2. SNP: Variant identifier
3. GD: Position in morgans or centimorgans (safe to user dummy value of '0')
4. BP: Base-pair coordinate (1-based; limited to 2^{31}-2)
5. Allele 1: corresponding to clear bits in .bed; usually minor)
6. Allele 2: corresponding to set bits in .bed; usually major)

Plink format 2

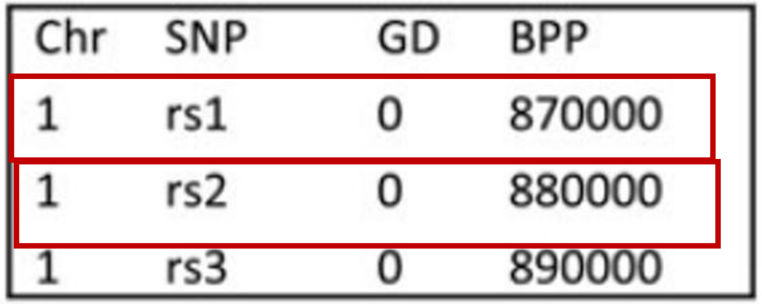
*.PED file
(one row per individual)



FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

1. FID: Family iD
2. IID: Within-family ID (cannot be ‘0’)
3. PID: Within-family ID of father (‘0’ if father isn’t in dataset)
4. MID: Within-family ID of mother (‘0’ if mother isn’t in dataset)
5. Sex: ‘1’ = male, ‘2’ = female, ‘0’ = unknown)
6. P: Phenotype, ’1’ = control, ’2’ = case, ‘-9’/‘0’/non-numeric = missing data if case/control

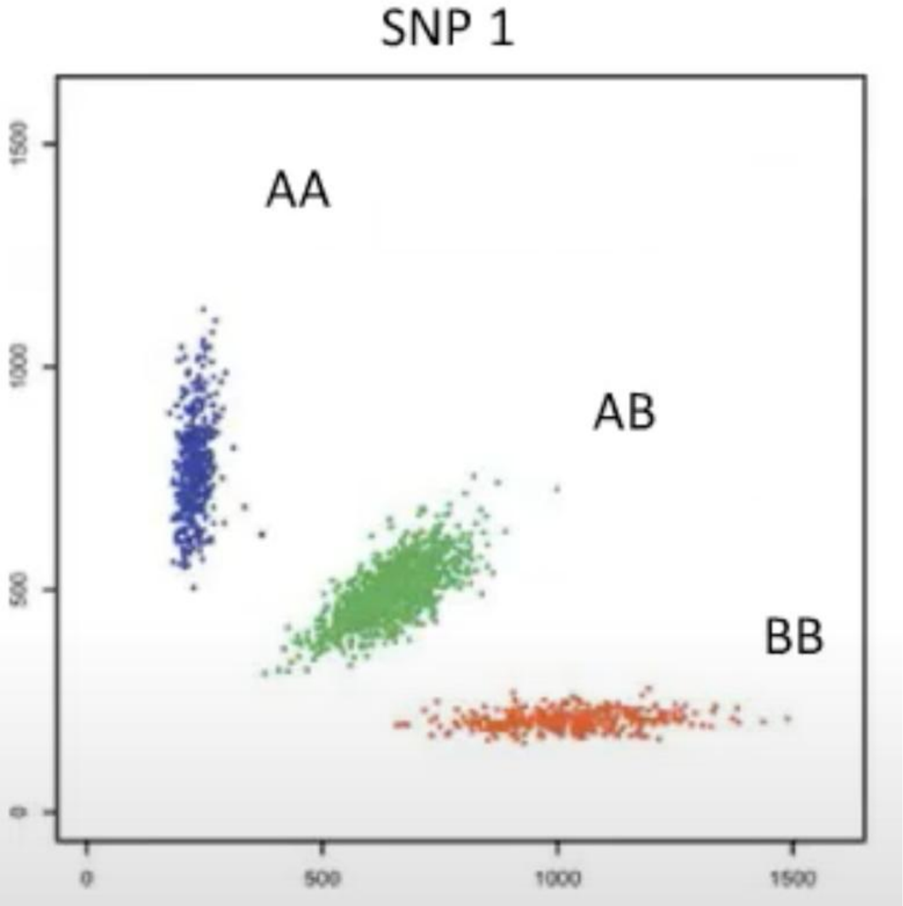
*.MAP file
(one row per SNP)



Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

1. Chr: Chromosome code (either an integer, or ‘X’/’Y’/’XY’/’MT’; ‘0’ indicates unknown) or name
2. SNP: Variant identifier
3. GD: Position in morgans or centimorgans (safe to user dummy value of ‘0’)
4. BP: Base-pair coordinate (1-based; limited to 2^{31}-2)

Why Do We Need Quality Control?



In an ideal world…

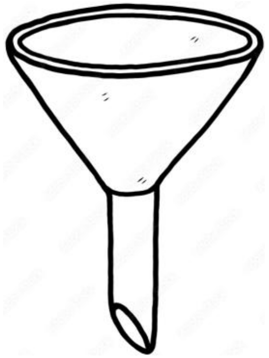
Our sampling practices would be perfect

Our experiments would run perfectly

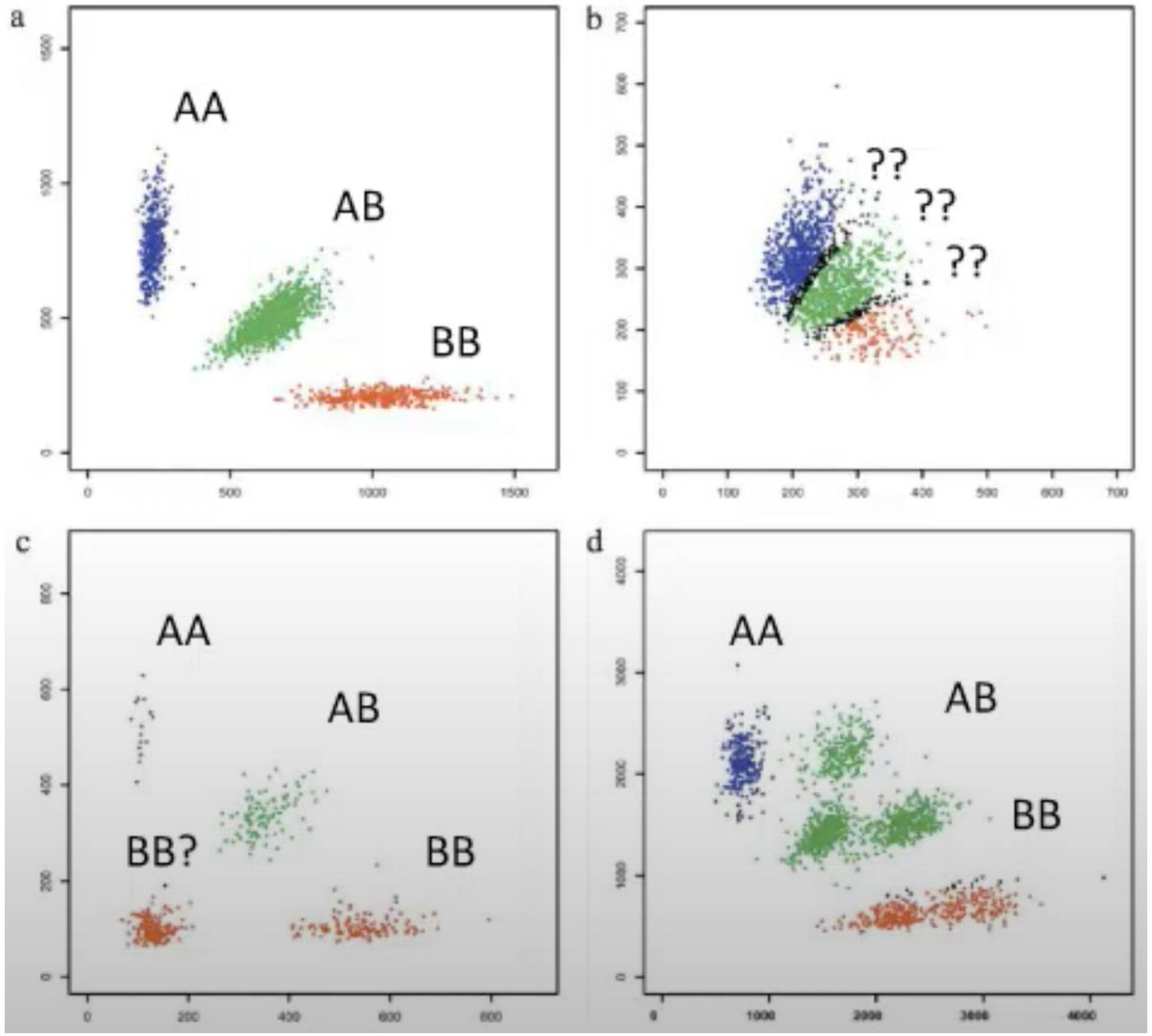
And all our SNP genotypes would look like this

Why Do We Need Quality Control?

- Large-scale experiments generate true results with a certain error rate
- Errors might originate at various steps in the processes:
  - ✓ Sample selection related issues
    - ✓ Cryptic relatedness
    - ✓ Population structure
  - ✓ Sample handling related issues
    - ✓ Labeling/Plating Error
  - ✓ Genotyping array related issues
    - ✓ Genotyping error
  - ✓ Batch effect related issues
    - ✓ Difference in results due to difference in sample processing

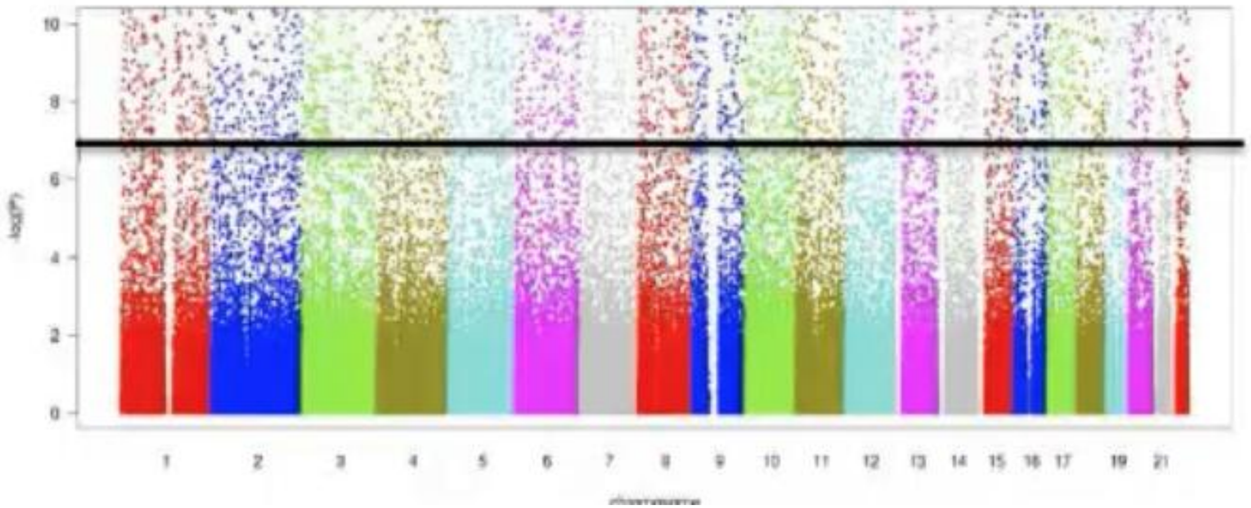


Why Do We Need Quality Control?

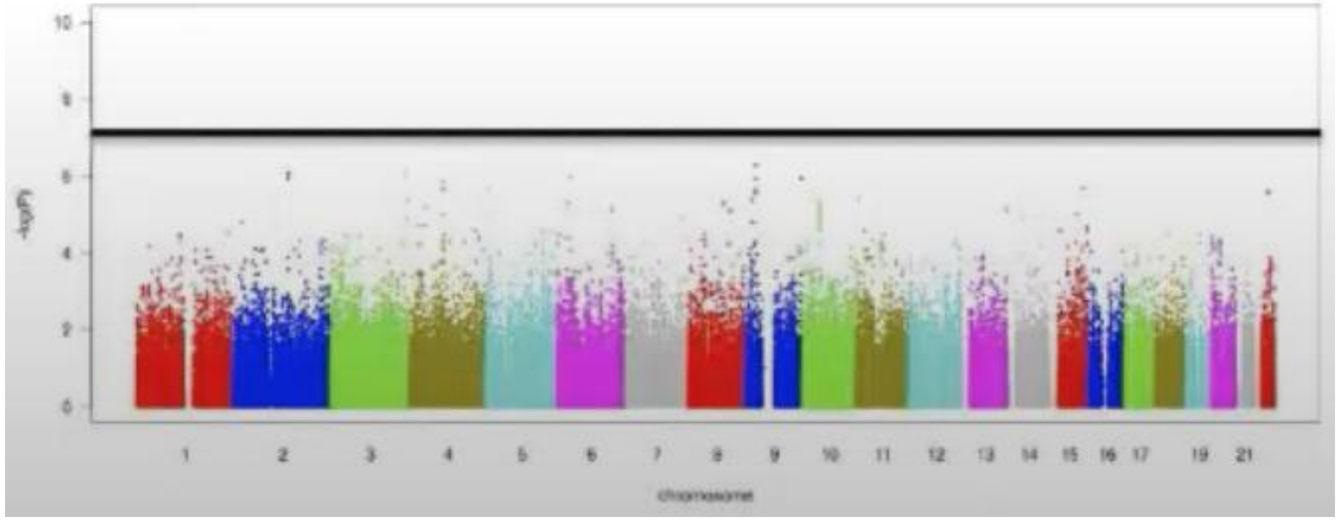


We don't live in an ideal world…

Example: German MI family study Affymetrix 500K Array Set SNPs on chips: 493,840



SNPs passing QC: 270,701



QC Roadmap

Sample QC

Discordant sex information

High Missingness

Excess or deficiency of heterozygosity

Duplicate or related

Divergent ancestry

Batch Effects

SNP QC

Low minor allele frequency

Missingness

Differential missingness

Hardy-Weinberg outliers

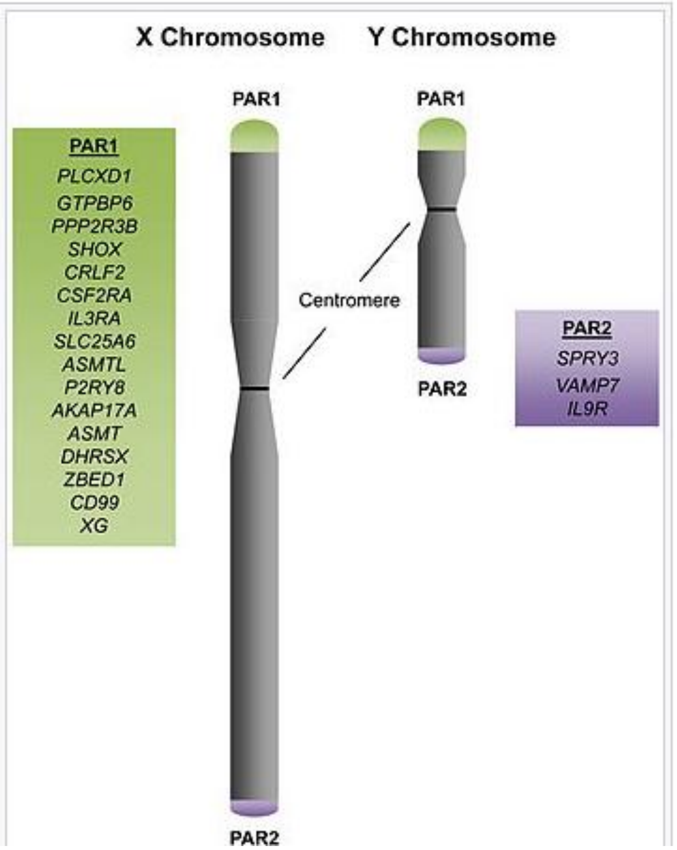
Every marker removed from a study is potentially an overlooked disease association and thus the impact of removing one marker is potentially greater than the removal of one individual.

Implementing QC on a ‘per-individual’ basis prior to conducting QC on a ‘per-marker’ basis to maximize the number of markers remaining in the study.

Gender Check (Genotype Data)

It is useful to begin by using genotype data from the X-chromosome to check for discordance with ascertained sex and thus highlight plating errors.

These are investigated to ensure that another DNA sample has not been genotyped by mistake.

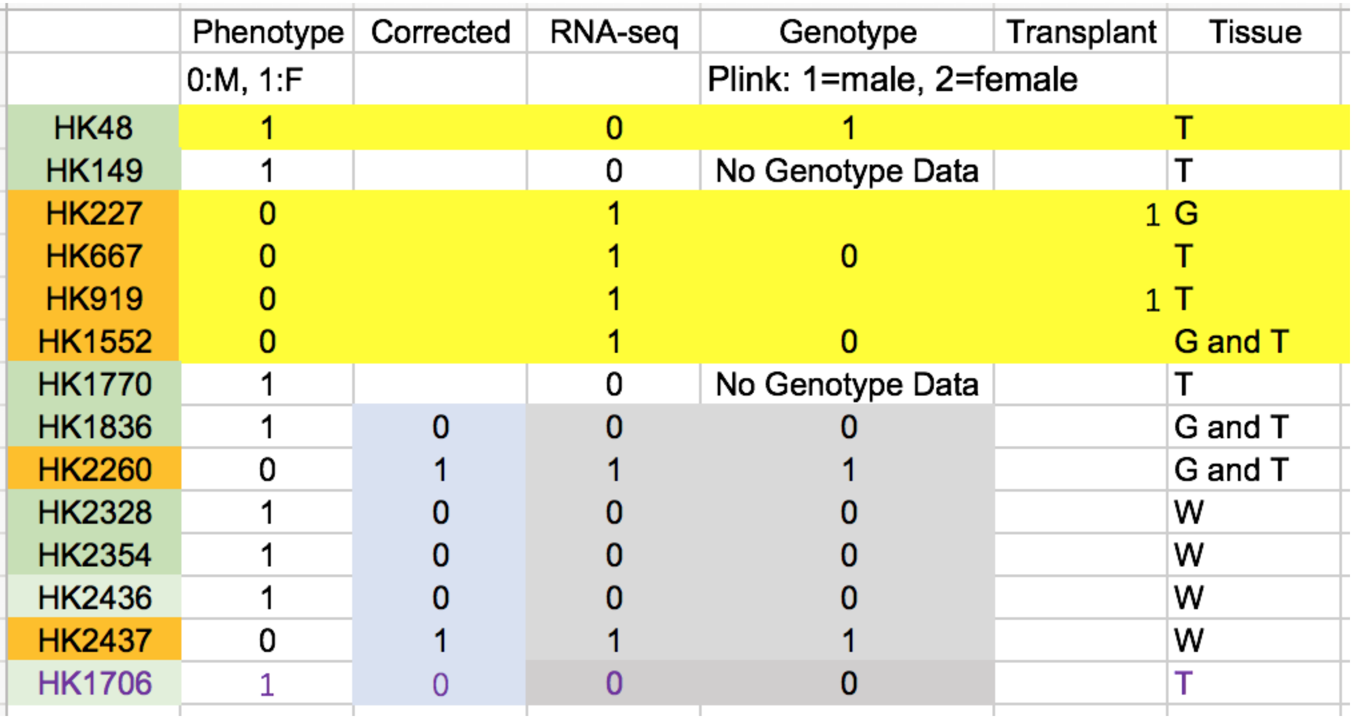


Pseudo-autosomal regions, PAR1, PAR2, are homologous sequences of nucleotides on the X and Y chromosomes.

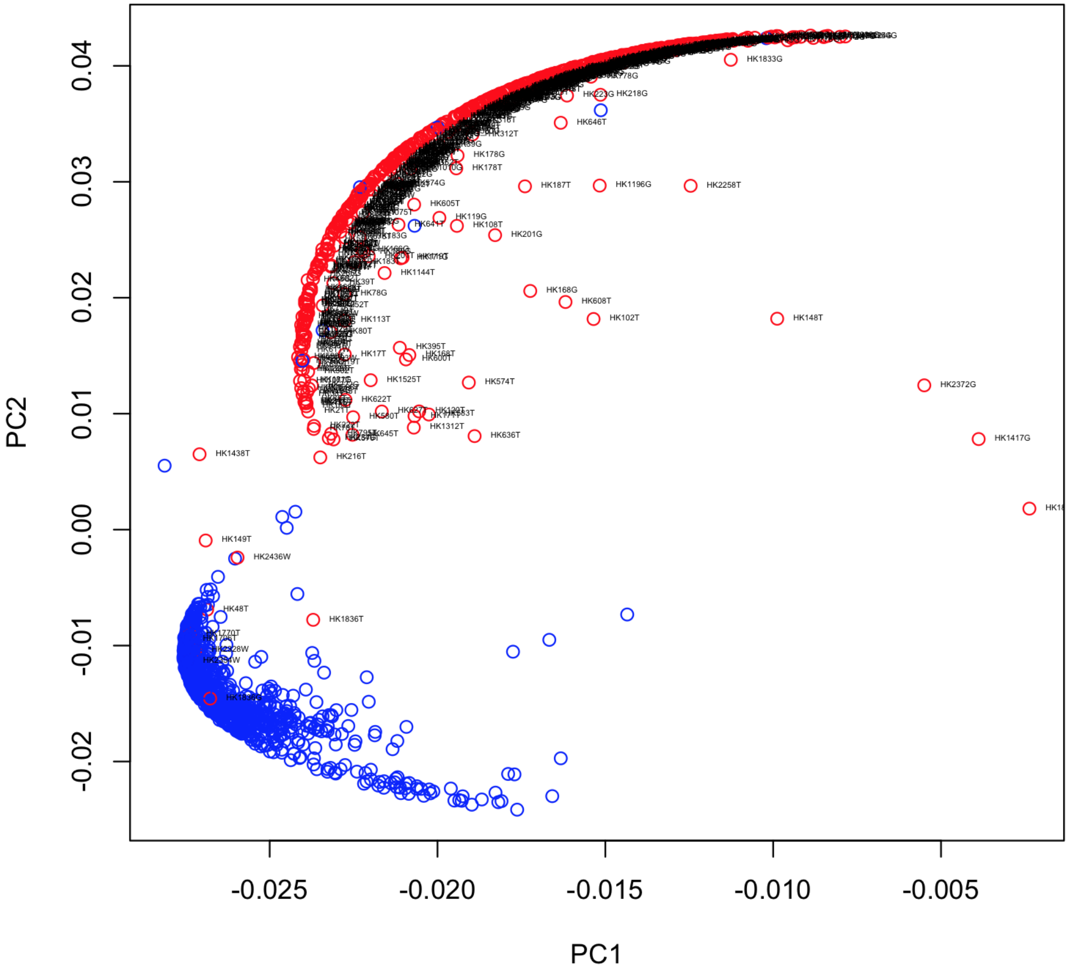
Males only have one copy of the X-chromosome they cannot be heterozygous for any marker not in the pseudo-autosomal region of the Y chromosome.

Expects: Male samples to have a homozygosity rate around 1
Females to have a homozygosity rate less than 0.2

Gender Check (RNA-seq data)

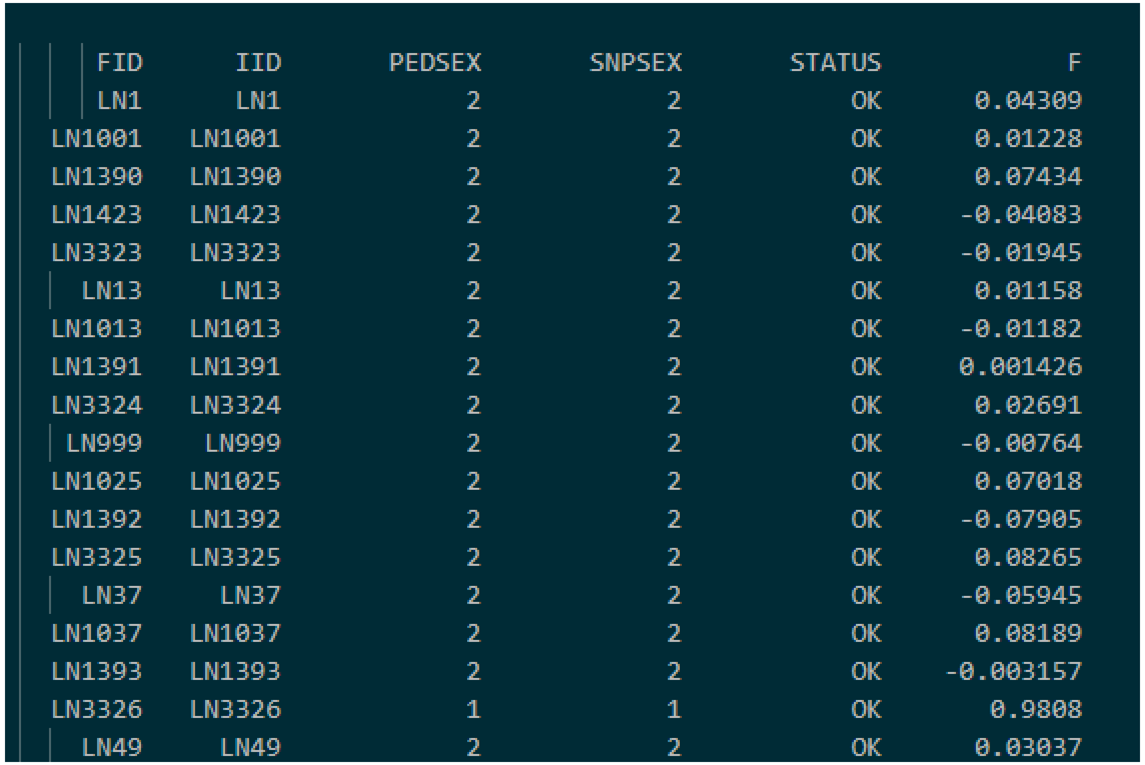


	Phenotype	Corrected	RNA-seq	Genotype	Transplant	Tissue
	0:M, 1:F			Plink: 1=male, 2=female		
HK48	1		0	1		T
HK149	1		0	No Genotype Data		T
HK227	0		1		1	G
HK667	0		1	0		T
HK919	0		1		1	T
HK1552	0		1	0		G and T
HK1770	1		0	No Genotype Data		T
HK1836	1	0	0	0		G and T
HK2260	0	1	1	1		G and T
HK2328	1	0	0	0		W
HK2354	1	0	0	0		W
HK2436	1	0	0	0		W
HK2437	0	1	1	1		W
HK1706	1	0	0	0		T

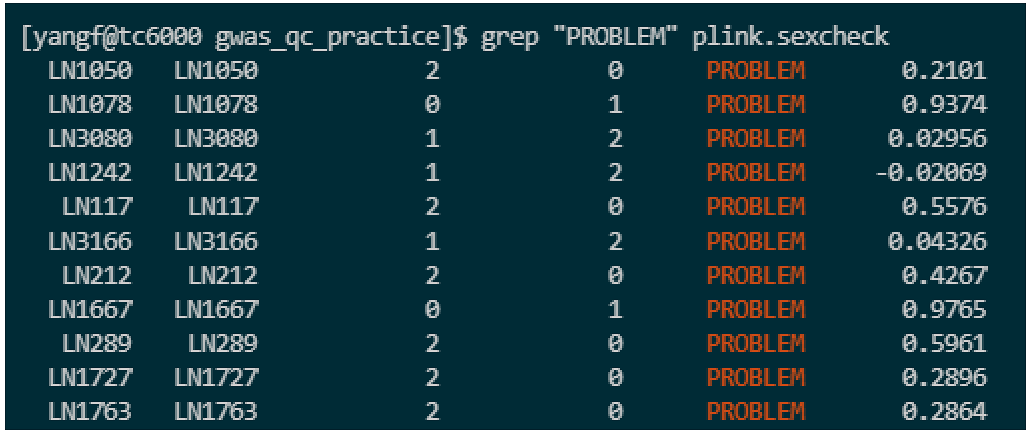


Individuals with Discordant Gender Information

‘plink –bfile raw-GWA-data –check-sex –out raw-GWA-data’
‘grep PROBLEM raw-GWA-data.sexcheck > raw-GWA-data.sexprobs’



```
     FID       IID     PEDSEX     SNPSEX     STATUS        F
     LN1       LN1          2          2         OK  0.04309
  LN1001    LN1001          2          2         OK  0.01228
  LN1390    LN1390          2          2         OK  0.07434
  LN1423    LN1423          2          2         OK -0.04083
  LN3323    LN3323          2          2         OK -0.01945
    LN13      LN13          2          2         OK  0.01158
  LN1013    LN1013          2          2         OK -0.01182
  LN1391    LN1391          2          2         OK 0.001426
  LN3324    LN3324          2          2         OK  0.02691
   LN999     LN999          2          2         OK -0.00764
  LN1025    LN1025          2          2         OK  0.07018
  LN1392    LN1392          2          2         OK -0.07905
  LN3325    LN3325          2          2         OK  0.08265
    LN37      LN37          2          2         OK -0.05945
  LN1037    LN1037          2          2         OK  0.08189
  LN1393    LN1393          2          2         OK -0.003157
  LN3326    LN3326          1          1         OK   0.9808
    LN49      LN49          2          2         OK  0.03037
```



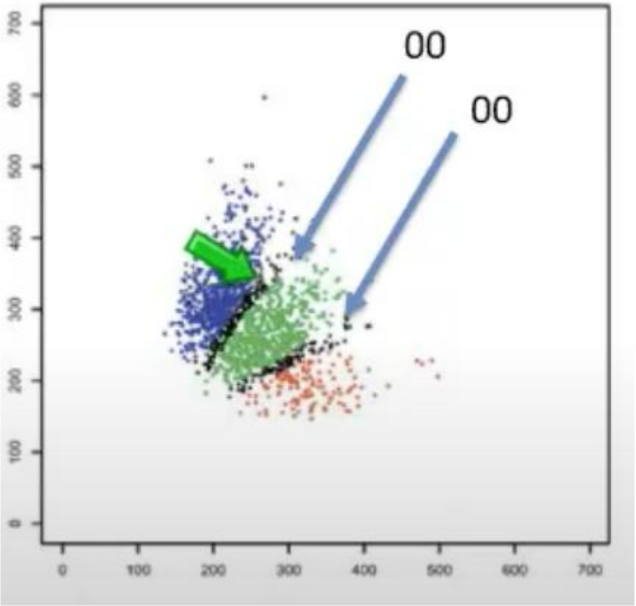
```
[yangf@tc6000 gwas_qc_practice]$ grep "PROBLEM" plink.sexcheck
  LN1050    LN1050          2          0    PROBLEM   0.2101
  LN1078    LN1078          0          1    PROBLEM   0.9374
  LN3080    LN3080          1          2    PROBLEM  0.02956
  LN1242    LN1242          1          2    PROBLEM -0.02069
   LN117     LN117          2          0    PROBLEM   0.5576
  LN3166    LN3166          1          2    PROBLEM  0.04326
   LN212     LN212          2          0    PROBLEM   0.4267
  LN1667    LN1667          0          1    PROBLEM   0.9765
   LN289     LN289          2          0    PROBLEM   0.5961
  LN1727    LN1727          2          0    PROBLEM   0.2896
  LN1763    LN1763          2          0    PROBLEM   0.2864
```

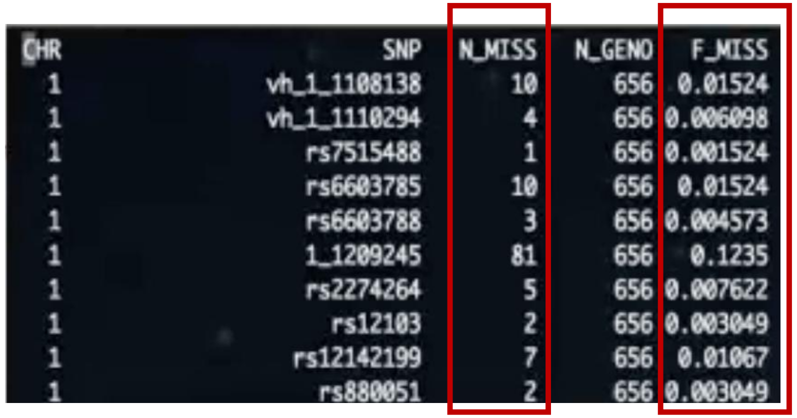
Sample Quality: Failure Rate

Typically, individuals with more than 3-7% missing genotypes should be removed. (Carefully scrutinizing the distribution of missing genotype rates across the entire sample set is the best way to ascertain the most appropriate threshold)

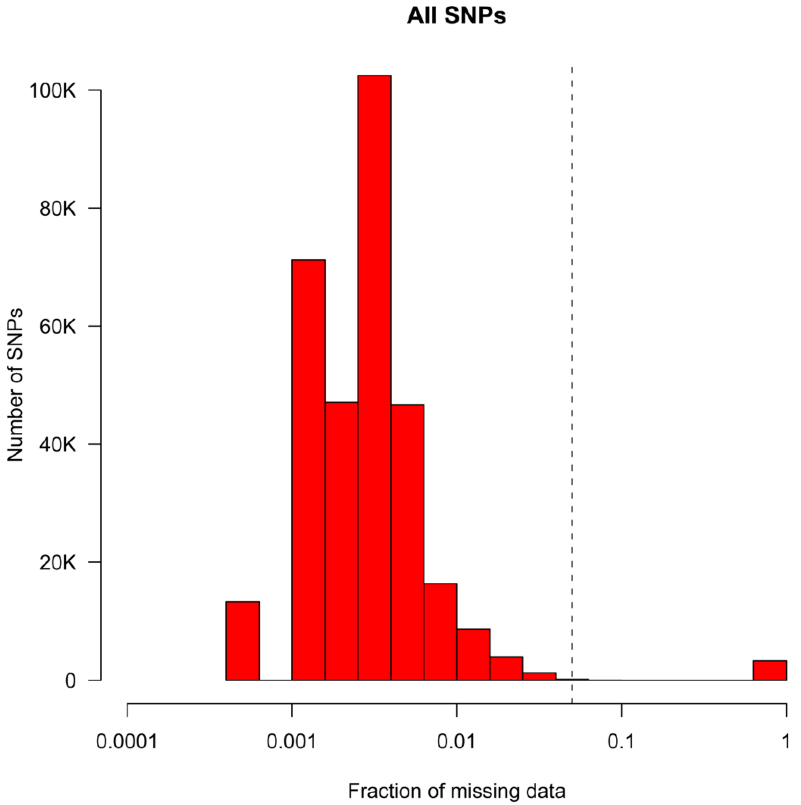
'plink –bfile raw-GWA-data –missing –out raw-GWA-data'

N_MISS: the number of missing SNPs
F_MISS: the proportion of missing SNPs per individual





```
CHR              SNP   N_MISS   N_GENO    F_MISS
  1    vh_1_1108138       10      656   0.01524
  1    vh_1_1110294        4      656  0.006098
  1       rs7515488        1      656  0.001524
  1       rs6603785       10      656   0.01524
  1       rs6603788        3      656  0.004573
  1       1_1209245       81      656    0.1235
  1       rs2274264        5      656  0.007622
  1         rs12103        2      656  0.003049
  1      rs12142199        7      656   0.01067
  1        rs880051        2      656  0.003049
```



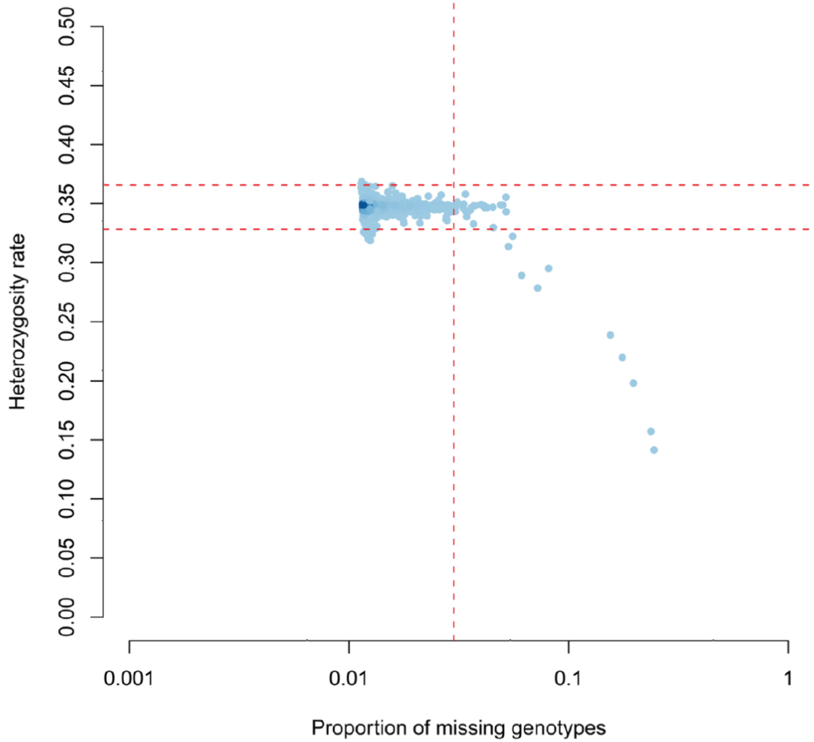
Sample Quality: Heterozygosity Rate

Sample contamination or inbreeding: All individuals should be inspected to identify individuals with an excessive or reduced proportion of heterozygote genotypes.

Mean heterozygosity: *(N-O)/N*
Where *N* is the number of non-missing genotypes and *O* is the observed number of homozygous genotypes for a given individual

'plink –bfile raw-GWA-data –het --out raw-GWA-data'

raw-GWA-data.het
[O(Hom)]: the number of homozygous genotypes
[N(NM)]: the number of non-missing genotypes per individual

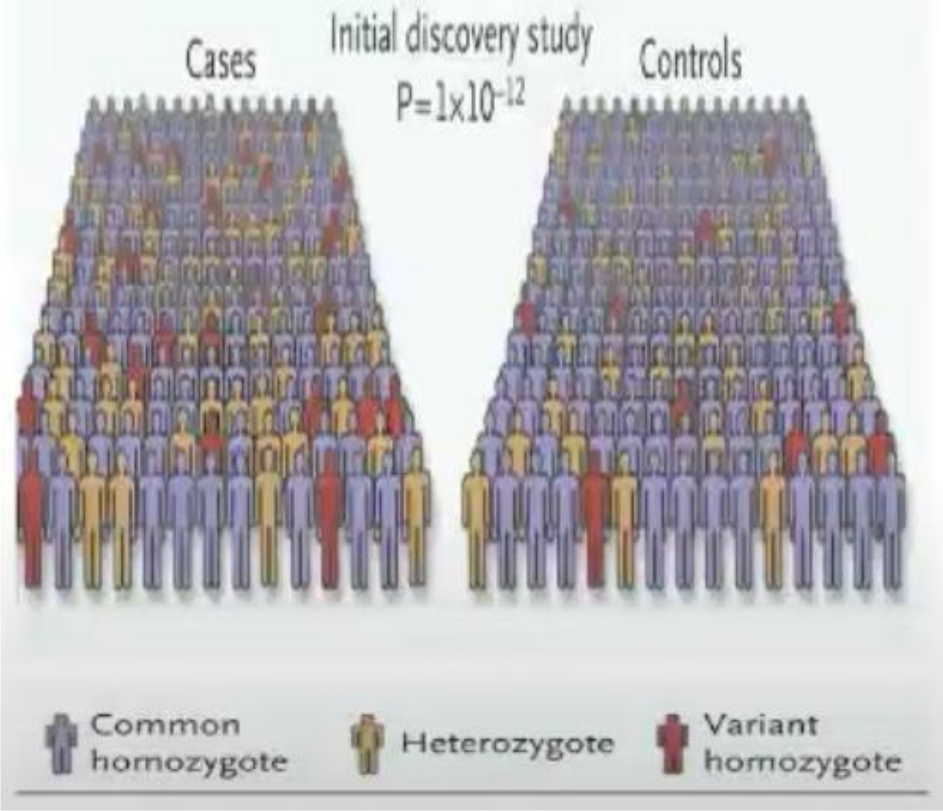


Exclude all individuals with a genotype failure rate ≥ 0.03 and/or heterozygosity rate ± 3 standard deviations from the mean.

Basic Feature (Population): All Sample Are Unrelated

The maximum relatedness between any pair of individuals is less than a second degree relative

If duplicates, first- or second- degree relatives are present, a bias may be introduced to the study because the genotypes within families will be over-represented.



IBD: identity by descent

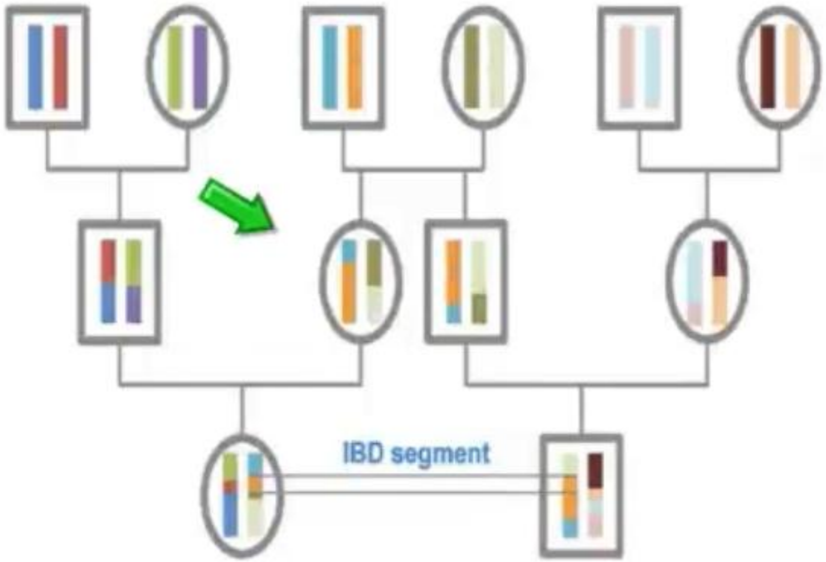
To identify duplicate and related individuals,

IBS (identity by state) is calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosomes)

IBD (identity by descent) can be estimated using genome-wide IBS data (using PLINK)

Duplicates or monozygotic twins: IBS=1; IBD=1
Related individuals: the degree of additional sharing proportional to the degree of relatedness
First-degree relatives: IBD=0.5
Second-degree relatives: IBD=0.25
Third-degree relatives: IBD=0.125

IBD is calculated and denoted in PLINK as Pi-hat



Identification of duplicated or related individuals

1. To reduce the computational complexity, the number of SNPs used to create the IBS matrix can be reduced by pruning the dataset so that no pair of SNPs has an r^2 greater than a given threshold (typically 0.2)

```
‘plink --file raw-GWA-data –exclude high-LD-regions.txt --range –indep-pairwise 50 5 0.2
--out raw-GWA-data’
```

2. To generate pair-wise IBS for all pairs of individuals

```
‘plink --bfile raw-GWA-data --extract raw-GWAS-data.prune.in --genome –out raw-GWA-
data’
```

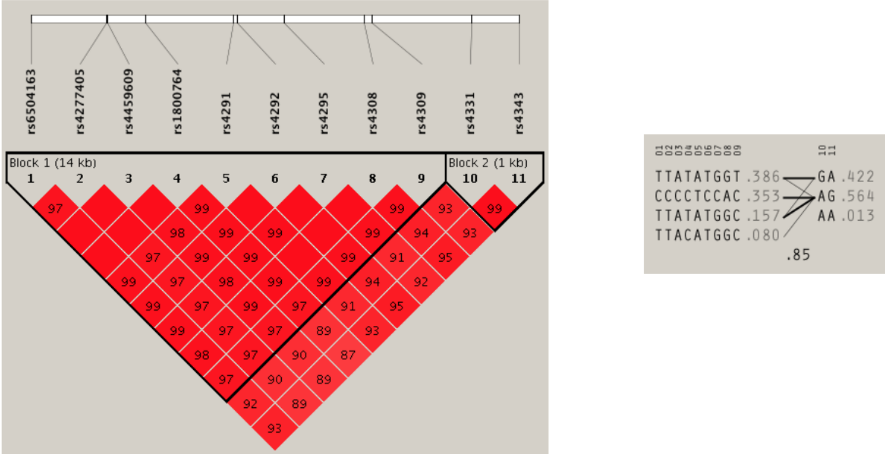
Confounding: population stratification

Confounders: underlying differences between the case and control subgroups other than those directly under study (typically, disease status)

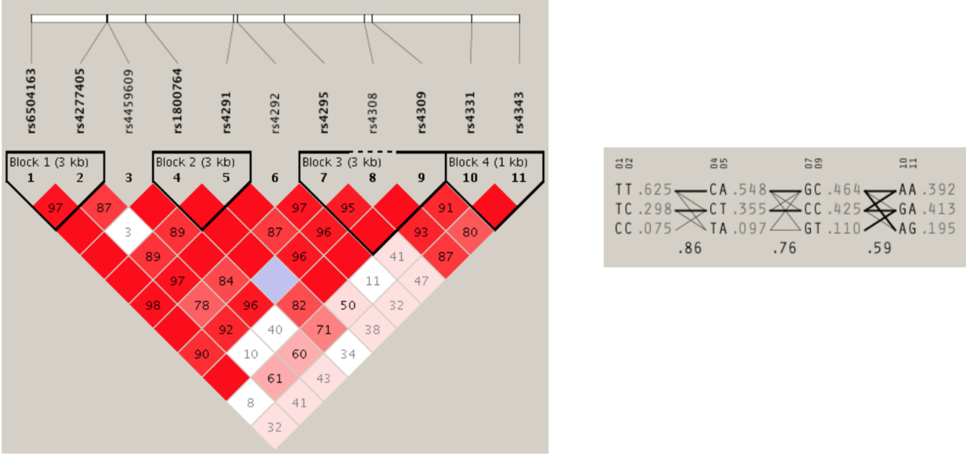
The main source of confounding in genetic studies is **population stratification**.

Population stratification: different population origins rather than any effect on disease risk; hidden fine-scale genetic substructure within a single population cannot be ruled out

EUR Population (1000 Genomes Project)



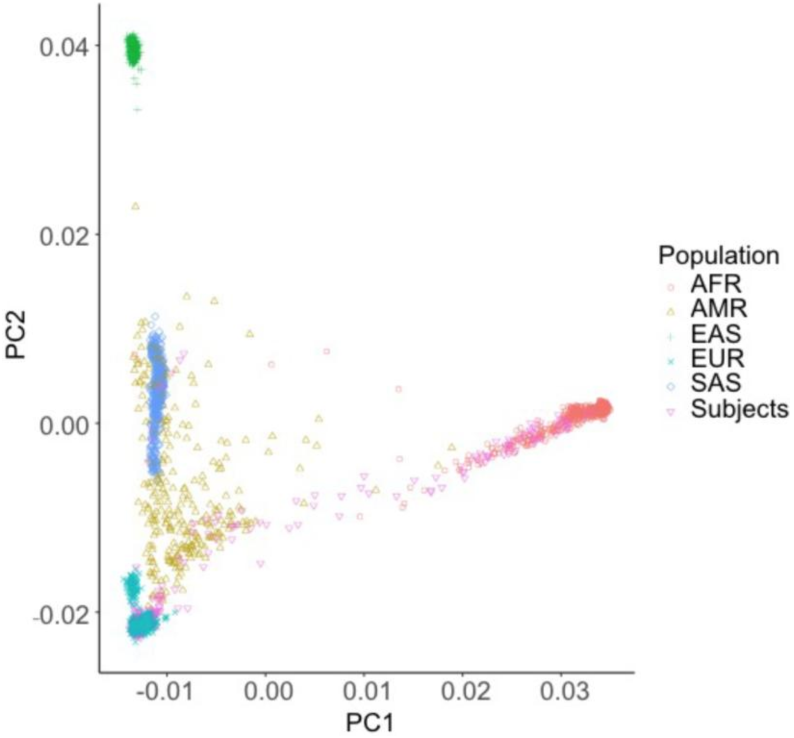
AFR Population (1000 Genomes Project)



Identification of individuals of divergent anscestry

Conduct a principal components analysis on the merged data

```
'perl smartpca.pl –I raw-GWA-data.hapmap3r2.pruned.bed –a raw-GWA-data.hapmap3r2.pruned.pedsnp-b raw-GWA-data.hapmap3r2.pruned.pedind –o raw-GWA-data.hapmap3r2.pruned.pca –p raw-GWA-data.hapmap3r2.pruned.plot –e raw-GWA-data.hapmap3r2.pruned.eval –l raw-GWA-data.hapmap3r2.pruned.log –k 2 –t 2 –w pca-populations.txt'
```



The 1000 genomes project (http://www.1000genomes.org)

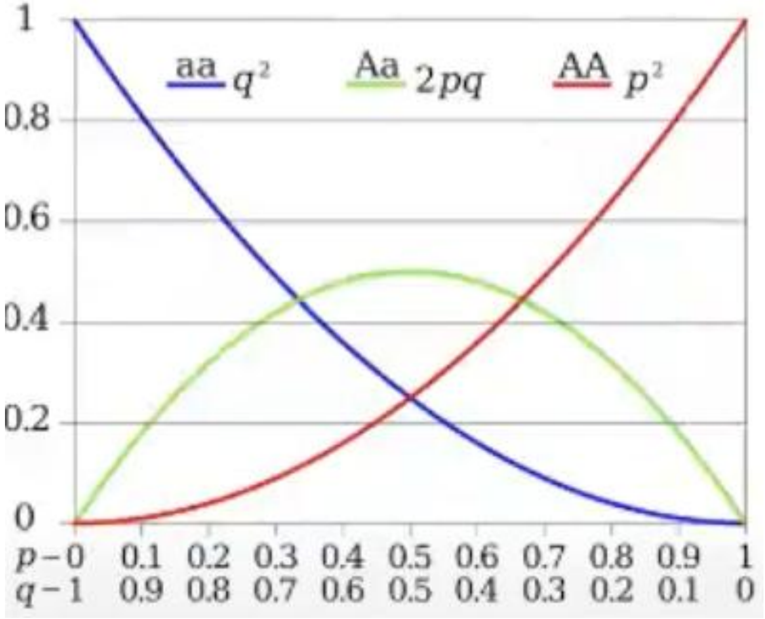
Pre-marker QC

a) SNPs with an excessive missing genotype (e.g. markers with a call rate less than 95% or 99% are removed)

b) SNPs demonstrating a significant deviation from Hardy-Weinberg equilibrium (HWE) (This can be indicative of a genotyping or genotype calling error, e.g. P-value thresholds between 0.001 and 5.7e-07)

c) SNPs with significantly different missing genotype rates between cases and controls

d) Markers with a very low minor allele frequency (e.g. minor allele frequency (MAF) < 5% or 1-2% but studies with small sample size may need to set this threshold higher)



Remove all individuals/SNPs failing QC

To concatenate all the files listing individuals failing the previous QC steps into single file

```
‘cat fail-* |sort –k1 | uniq > fail-qc-inds.txt’
```

To remove low quality samples

```
‘plink –bfile raw-GWA-data –remove fail-qc-inds.txt -–make-bed –-out clean-inds-GWA-data’
```

To calculate the missing genotype rate for each marker type

```
‘plink –-bfile clean-inds-GWA-data --missing –-out clean-inds-GWA-data’
```

To test all markers for differences in call rate between cases and controls

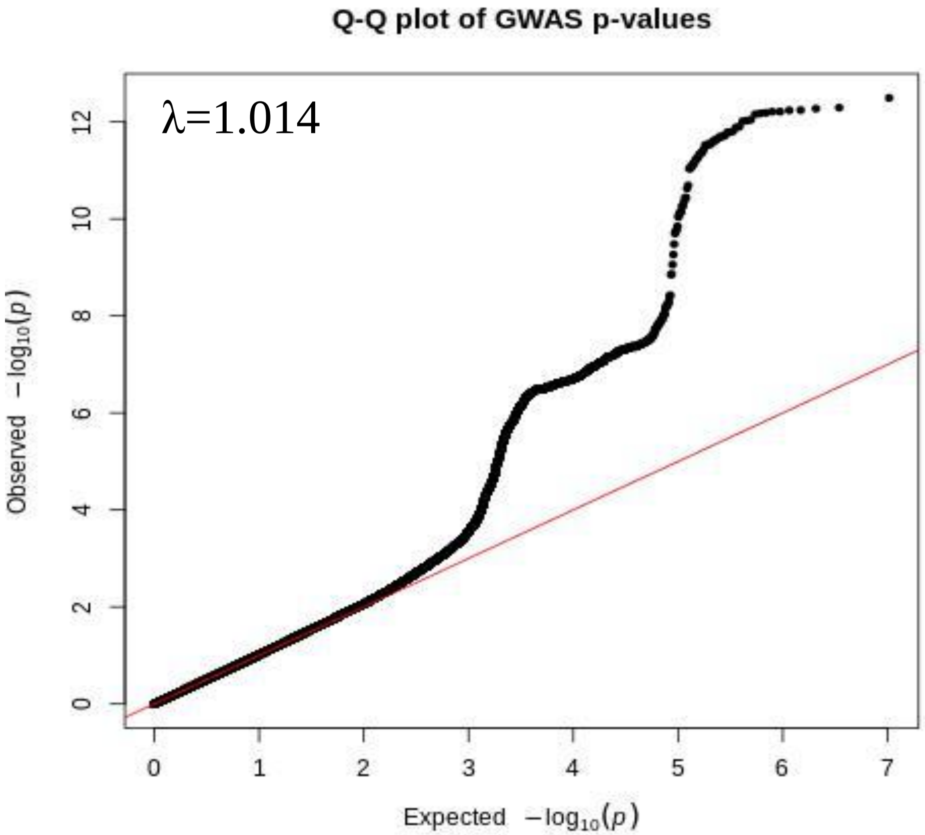
```
‘plink –-bfile clean-inds-GWA-data –-test-missing –out clean-inds-GWA-data’
```

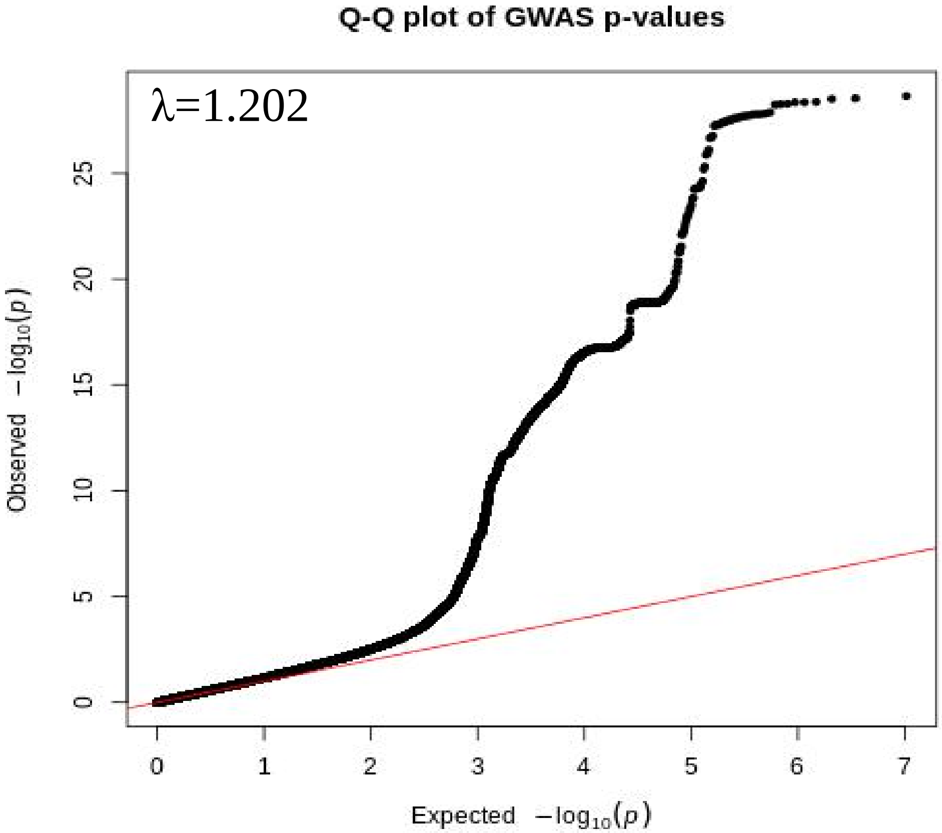
To remove poor SNPs from further analysis and create a clean GWA data file type

```
‘plink –-bfile clean-inds-GWA-data –-exclude fail-diffmiss-qc.txt –-maf 0.01 –geno 0.05 –hwe
0.0001 –make-bed –out clean-GWA-data’
```

GWAS

- More than 99% of the SNPs follow the null distribution of no association.





Linear regression: Y=aX+bU+c,
X: SNP; Y: Trait; c: Intercept; U: confounders

Software

- PLINK software for genotype Quality Control
- SMARTPCA.pl software for running principal components analysis
- Statistical software for data analysis and graphing, such as:
  R