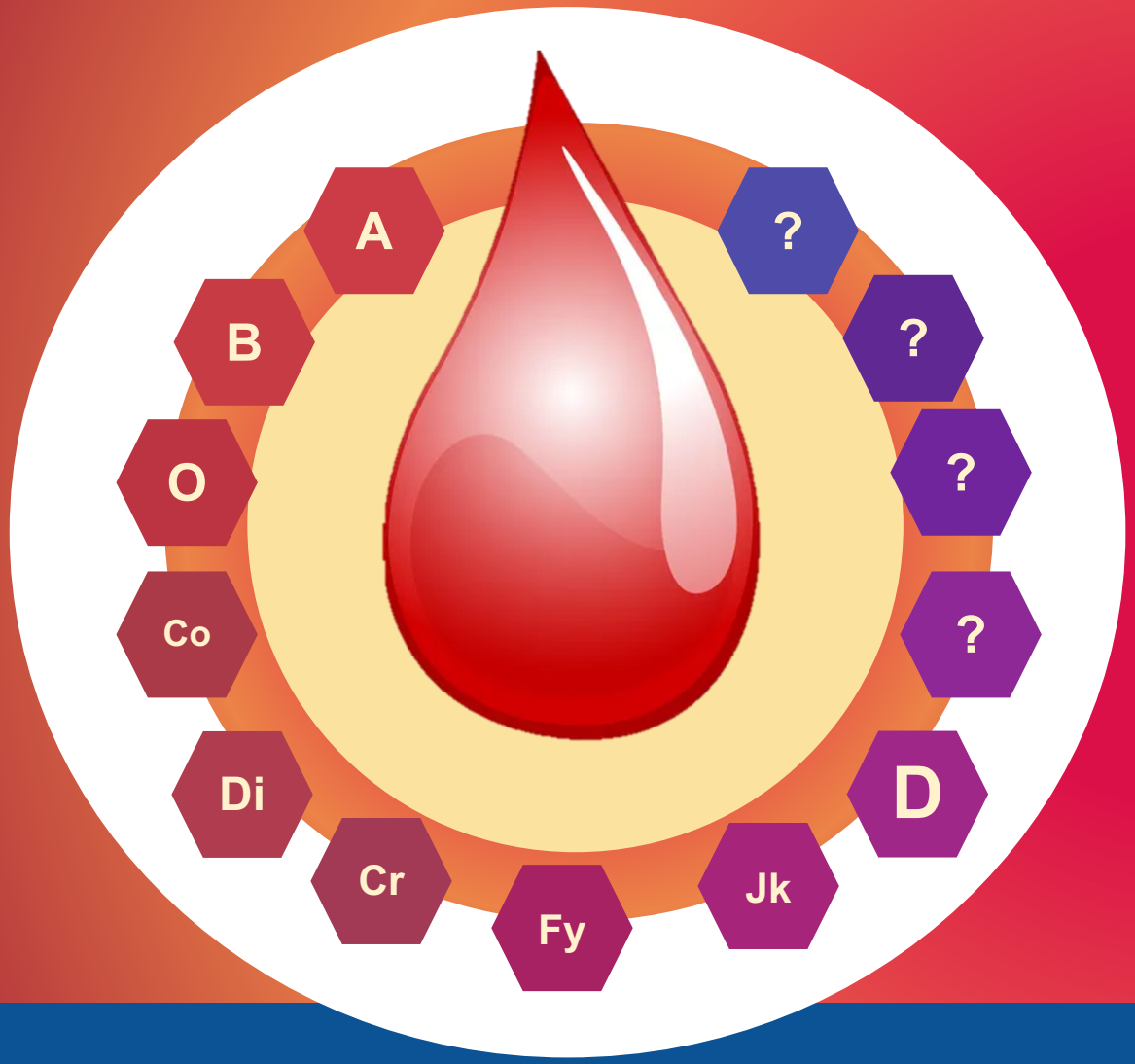


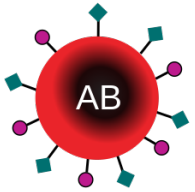
Genomics
The future of blood grouping

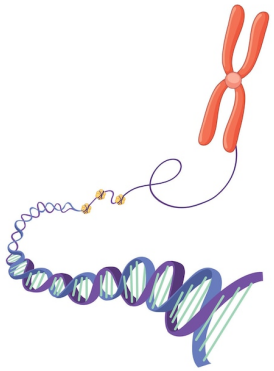
IGIB
Institute of Genomics & Integrative Biology
Genomics Knowledge Partner

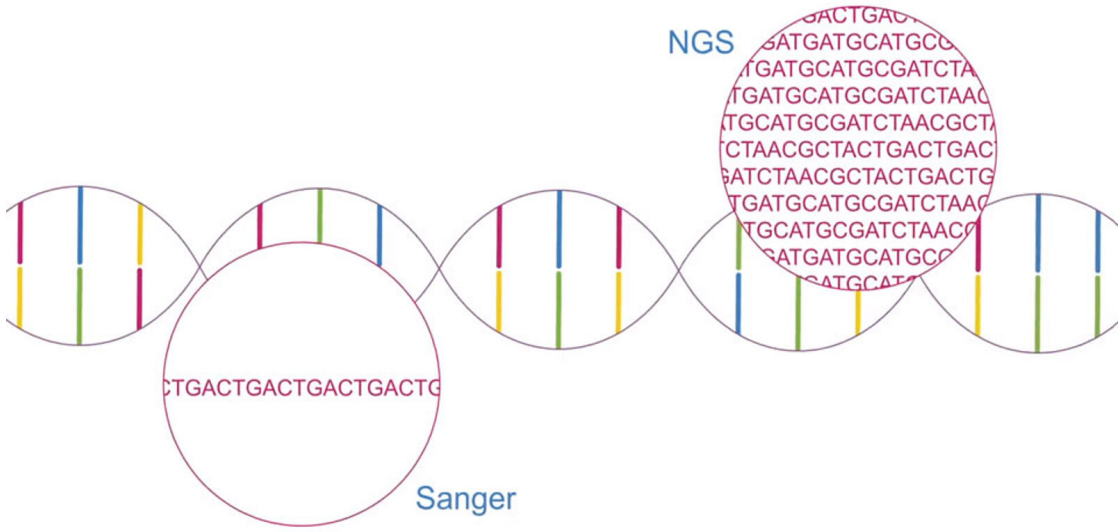
Mercy Rophina
Senior Research Fellow
CSIR – Institute of Genomics and Integrative Biology (IGIB)

So far in our course ...

A quick recap !

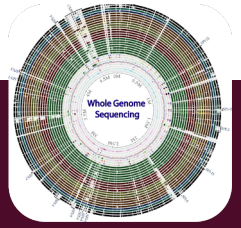




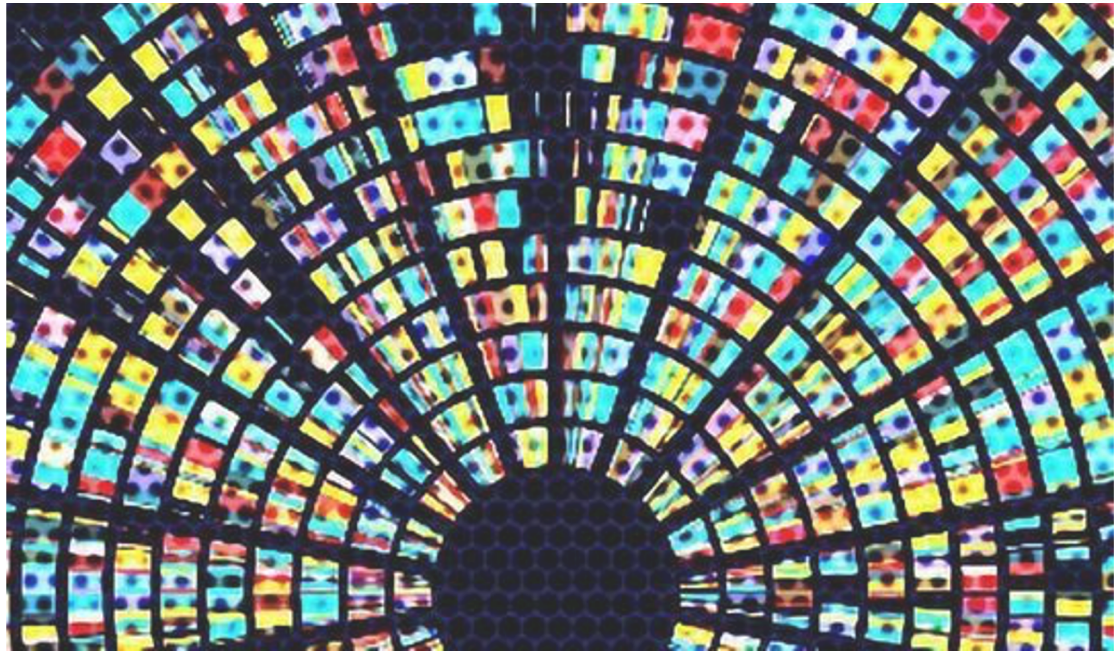


Serological techniques

Molecular techniques

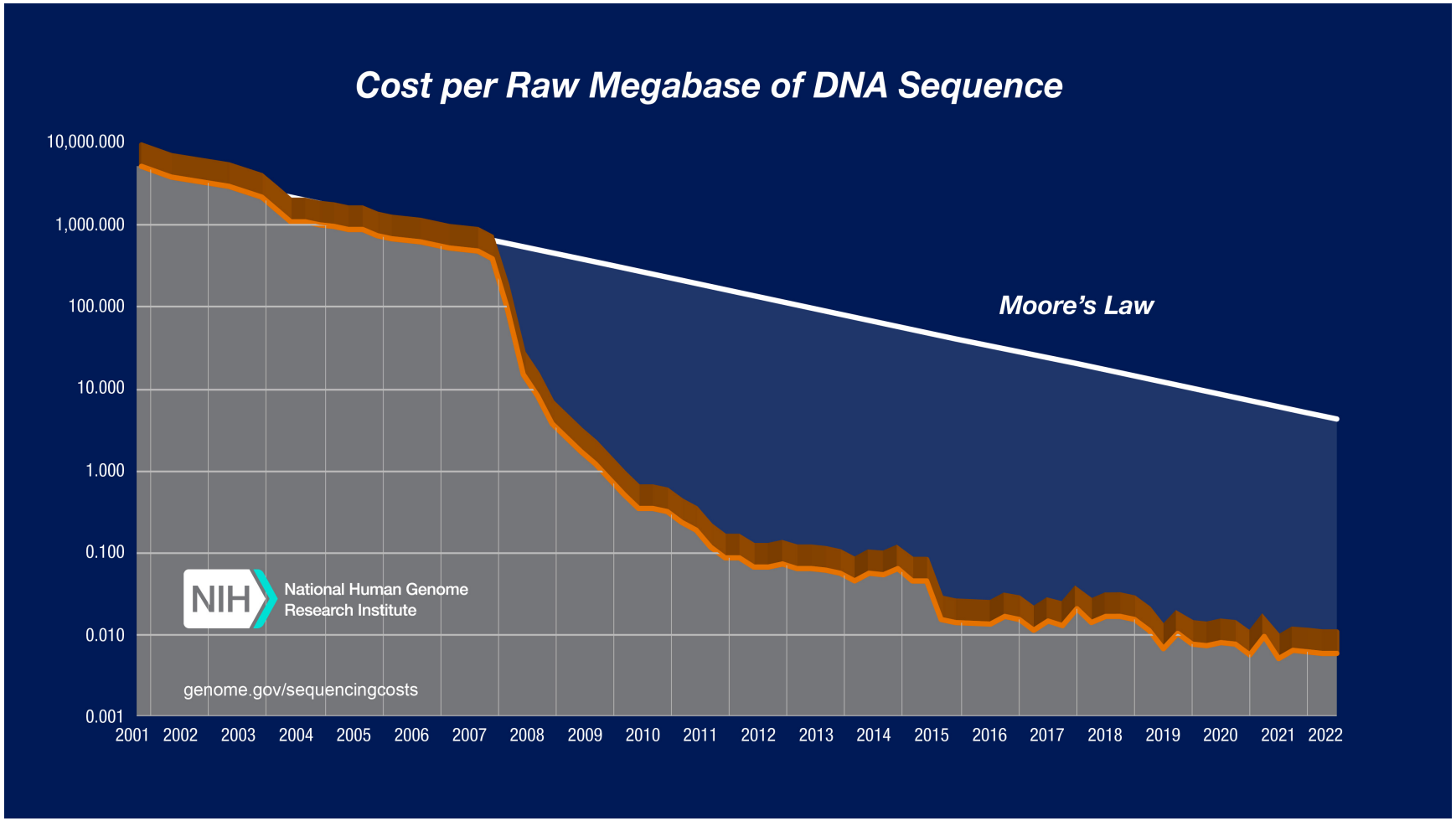


Sequencing (NGS) techniques



Future of Immunohematology

*Advent of **NGS***



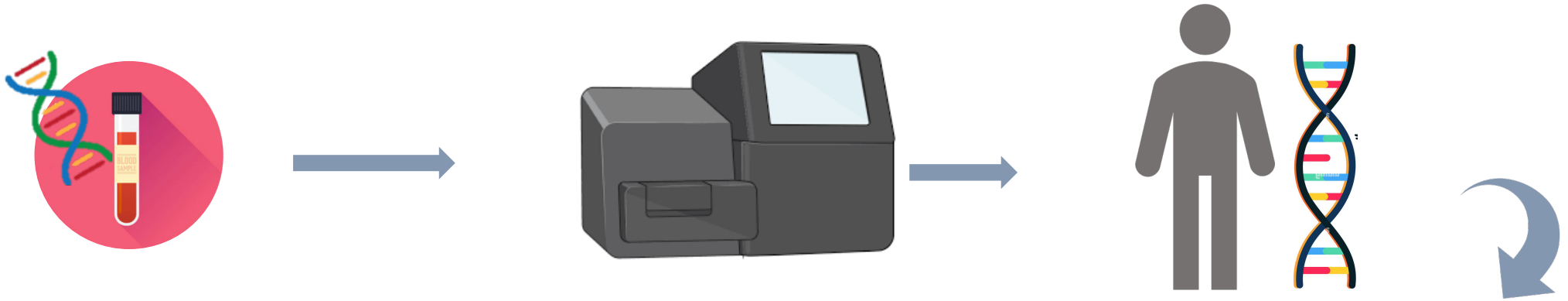
Constant **drop in the cost** of sequencing over the years

More number of sequencing projects performed globally

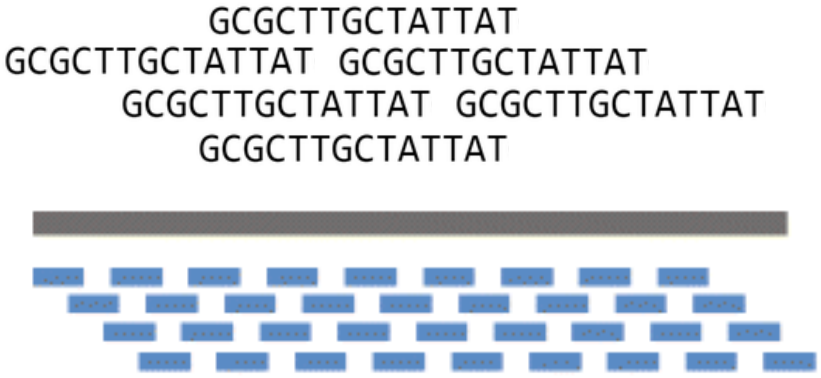
Availability of bulk **sequencing data** in public

Future of Immunohematology

*Advent of **NGS***

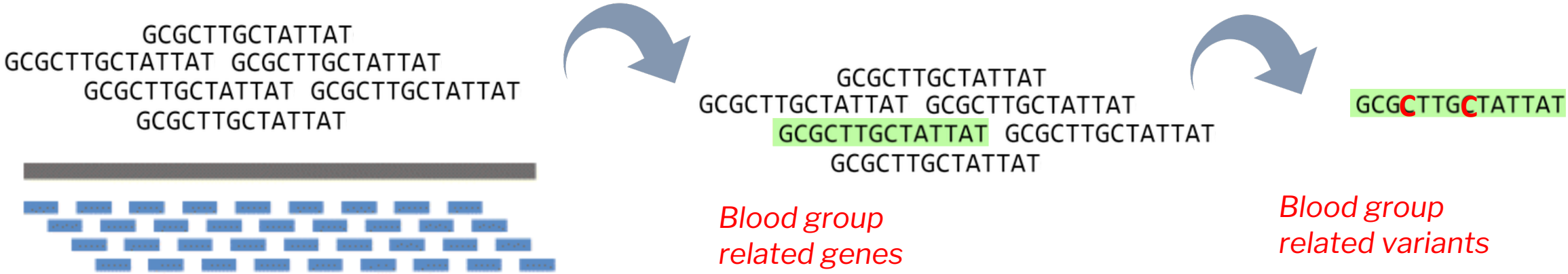


Blood group genotyping from sequencing data ?



Blood typing from NGS data

The basic concept



2 basic pre-requisites

How does a sequencing data file look like ?

How are we going to search for variations and correlate the blood phenotype ?

Blood typing from NGS data

The reference data

51

Unique genes

Single Nucleotide Variations (SNVs)

Deletions

Insertion

Variable expression of antigens

Variation in Blood Phenotype

Need to look for these variations in the sequencing data

Blood typing from NGS data

The reference data

51

Unique genes

SLC14A1 gene

Chromosome : ***18***

Genomic locus : ***45724181 - 45752520***

(NCBI, Ensembl, LRG)

Blood group associated variants : ***149***

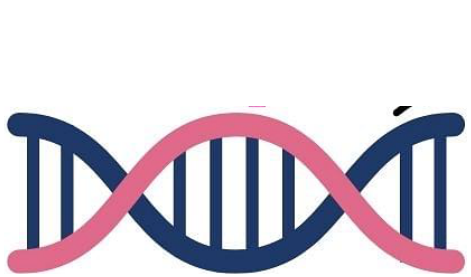
Variant details

Jk(b+) phenotype

Protein position	***cDNA position***	***Genomic position***
NP_056949.4: p.Asp280Asn	NM_015865.7: c.838G>A	NC_000018.10: g.45739554G>A

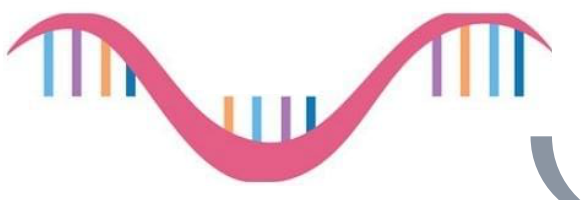
Blood typing from NGS data

The reference data

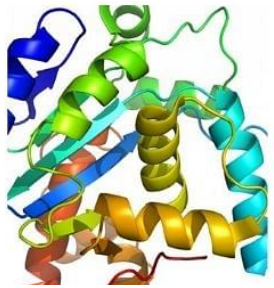


DNA

RNA



cDNA



Protein

Genomic position

NC_000018.10:

g.45739554G>A

mutalyzer

cDNA position

NM_015865.7:

c.838G>A

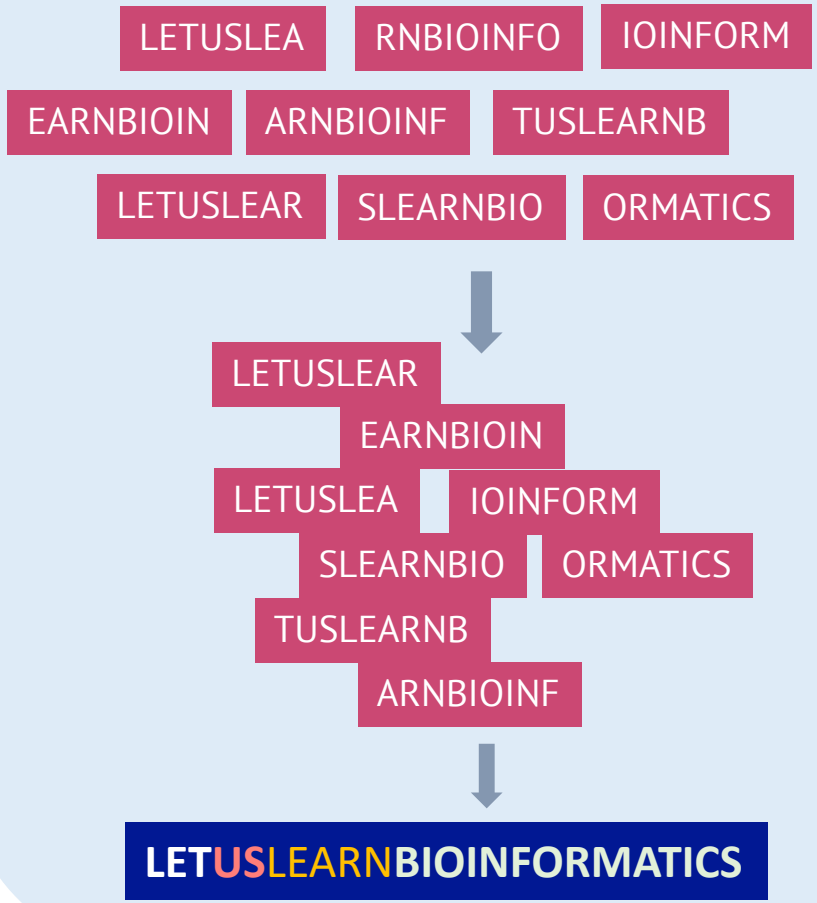
Protein position

NP_056949.4:

p.Asp280Asn

NGS data files – *A quick visit*

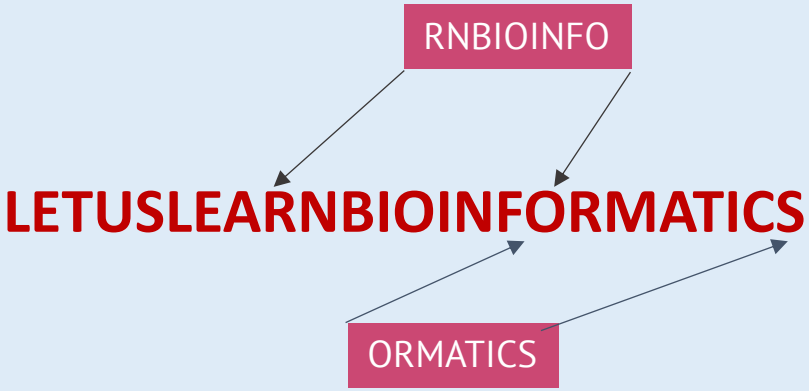
Read assembly



(Fastq files)

(Fasta files)

Reference alignment



(SAM files)

Comparison

(VCF files)

NGS data files – *A quick visit*

Fasta File

- Text based format representing either nucleotide or peptide sequence
- Nucleotides and amino acids represented by **single – letter codes**

Nucleotide Fasta

```
>X81322.1 E.coli hpcC gene
GAAGTAGAAGGCGTGGGCCGCCTGGTGAACCGAATTGTTGAGTGAGGAAACAGCGAAATGAAAAAAGTAA
ATCATTGGATCAACGGCAAAAATGTTGCAGGTAACGACTACTTCCTGACCACCAATCCGGCAACGGGTGA
AGTGCTGGCGGATGTGGCCTCTGGCGGTGAAGCGGAGATCAATCAGGCGGTAGCGACAGCGAAAGAGGCG
TTCCCGAAATGGGCCAATCTGCCGATGAAAGAGCGTGCGCGCCTGATGCGCCGTCTGGGCGATCTGATCG
```

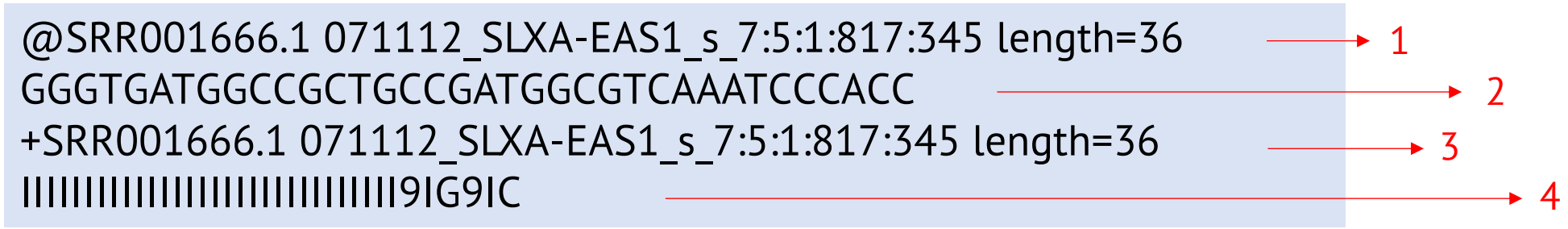
Protein Fasta

```
>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ
QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ
LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK
```

NGS data files – *A quick visit*

Fastq File

- Text based format storing both **biological sequence** along with **quality scores**
- Each sequence requires at least **4 lines**



1 *Header/Sequence identifier*

2 *Sequence*

3 *Header/Sequence identifier*

4 *Quality scores*

NGS data files – *A quick visit*

Sequence Alignment Map File

- Stores information of **biological sequences aligned to reference sequence**

Header section

- Starts with an @ symbol

Alignment section

- **11** mandatory fields + numerous optional fields

NGS data files – *A quick visit*

Sequence Alignment Map File

Header

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Body/ Alignment

NGS data files – *A quick visit*

Columns in SAM alignment section

1 *Read name*

2 *SAM flag*

3 *Chromosome/contig name*

4 *Mapped position in the reference*

5 *Mapping quality*

6 *CIGAR String*

7 *Name of the mate*

8 *Position of the mate*

9 *Template length*

10 *Read sequence*

11 *Read quality*

12 *Additional information*

NGS data files – *A quick visit*

Variant Call File – VCF

- Standard format for storing genetic variation information

Header

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial ##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

Body

NGS data files – *A quick visit*

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5; DB;H2	GT:GQ:DP:HQ 0|0:48:1:51,51

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial ##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

Blood variants and phenotypes

The correlation

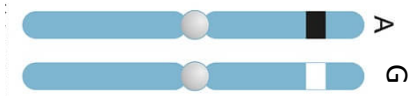
Reference list of known/approved blood group related genetic variants

(Checking for the presence/absence of variants)

VCF file generated as a sequencing output

CHR	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
18	45739554	.	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ **0/1**:48:1:51,51

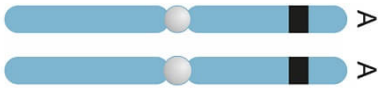
Heterozygous



Single dose

CHR	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
18	45739554	.	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ **1/1**:48:1:51,51

Homozygous



Double dose

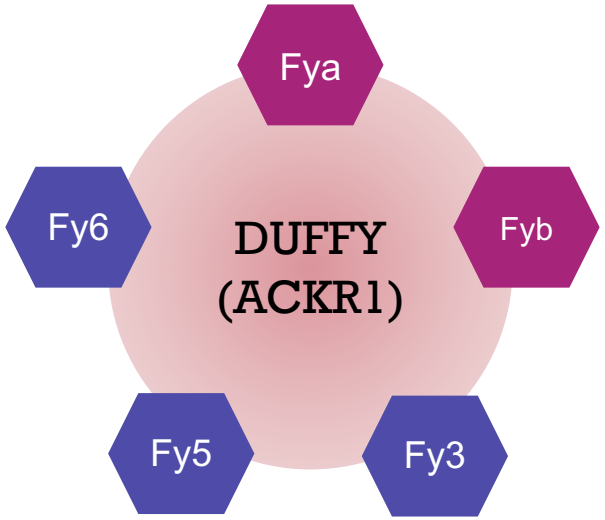
Blood phenotyping methodology

Using genetic variant information

Relating genetic variations with blood phenotype

- Reference Allele – **FY*A (NM_002036.4:c.125G) – Fy(a+) phenotype**
- **FY*B (NM_002036.4:c.125G>A) – Fy(b+) phenotype**

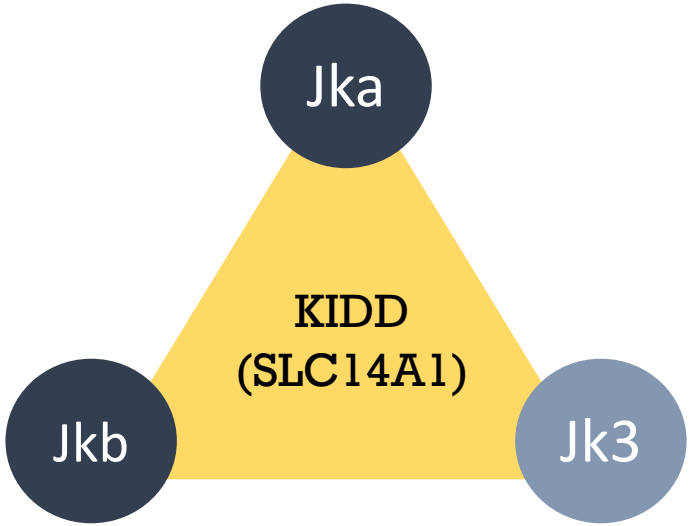
AA change	# Hom	# Het	# Not Present
c.125G>A p.Gly42Asp	1	10	61



- **1** sample → NM_002036.4:c.125A/A → Fy(a-b+) phenotype
- **10** sample → NM_002036.4:c.125G/A → Fy(a+b+) phenotype
- **61** sample → NM_002036.4:c.125G/G → Fy(a+b-) phenotype

Blood phenotyping methodology

Using genetic variant information



Relating genetic variations with blood phenotype

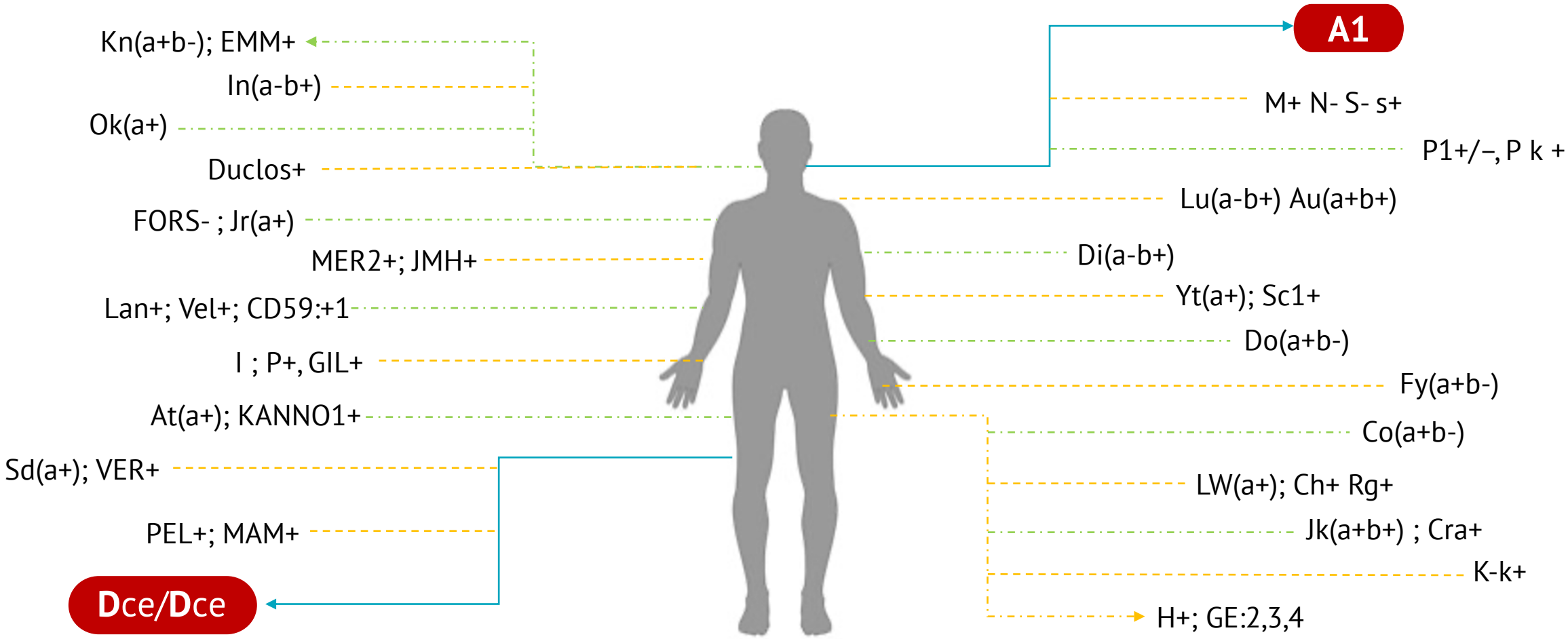
- Reference Allele – **JK*A (NM_015865.7:c.838G)** – **JK(a+) phenotype**
- **JK*B (NM_015865.7:c.838G>A)** – **Jk(b+) phenotype**

AA change	# Hom	# Het	# Not Present
c.838G>A p.Asp280Asn	2	17	35

- **2** sample → **NM_015865.7:c.838A/A** → JK(a-b+) phenotype
- **17** sample → **NM_015865.7:c.838G/A** → JK(a+b+) phenotype
- **35** sample → **NM_015865.7:c.838G/G** → JK(a+b-) phenotype

Complete blood type of a person

Blood Group profile in use : A+ve/A1+ve

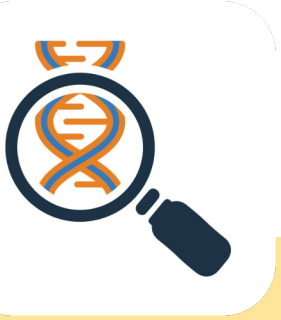


Extremely helpful in avoiding transfusion complications caused by minor blood group alleles

Blood profiling – Past, Present and Future

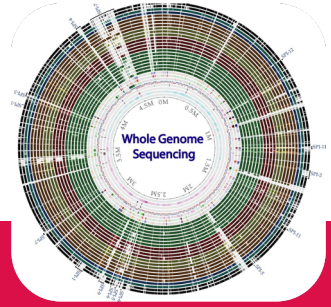
Serological testing

- Based on the principle of **agglutination**
- Reactivity between blood RBCs and antisera



Molecular testing

- **Obtaining DNA sample of the person for sequencing**
- **PCR methods as central tools**
- **Limited to few blood groups**



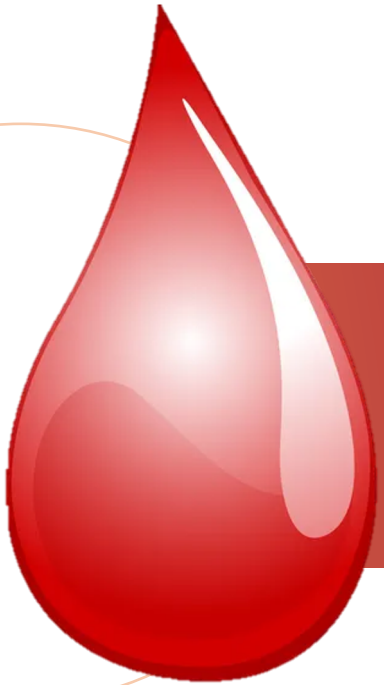
Sequencing based testing

- **Complete variation search of a person's genome/exome**
- **Extensive blood type of a person can be identified**
- **Novel variant discoveries**

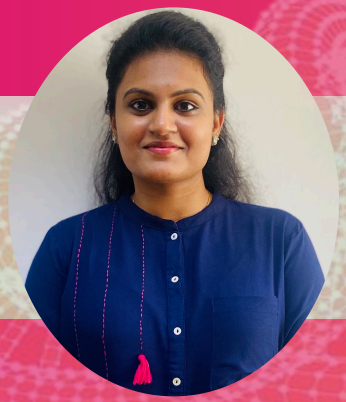
Upcoming courses …

- Almost the end of our lecture series
- From genomes to populations – Population specific blood group registries

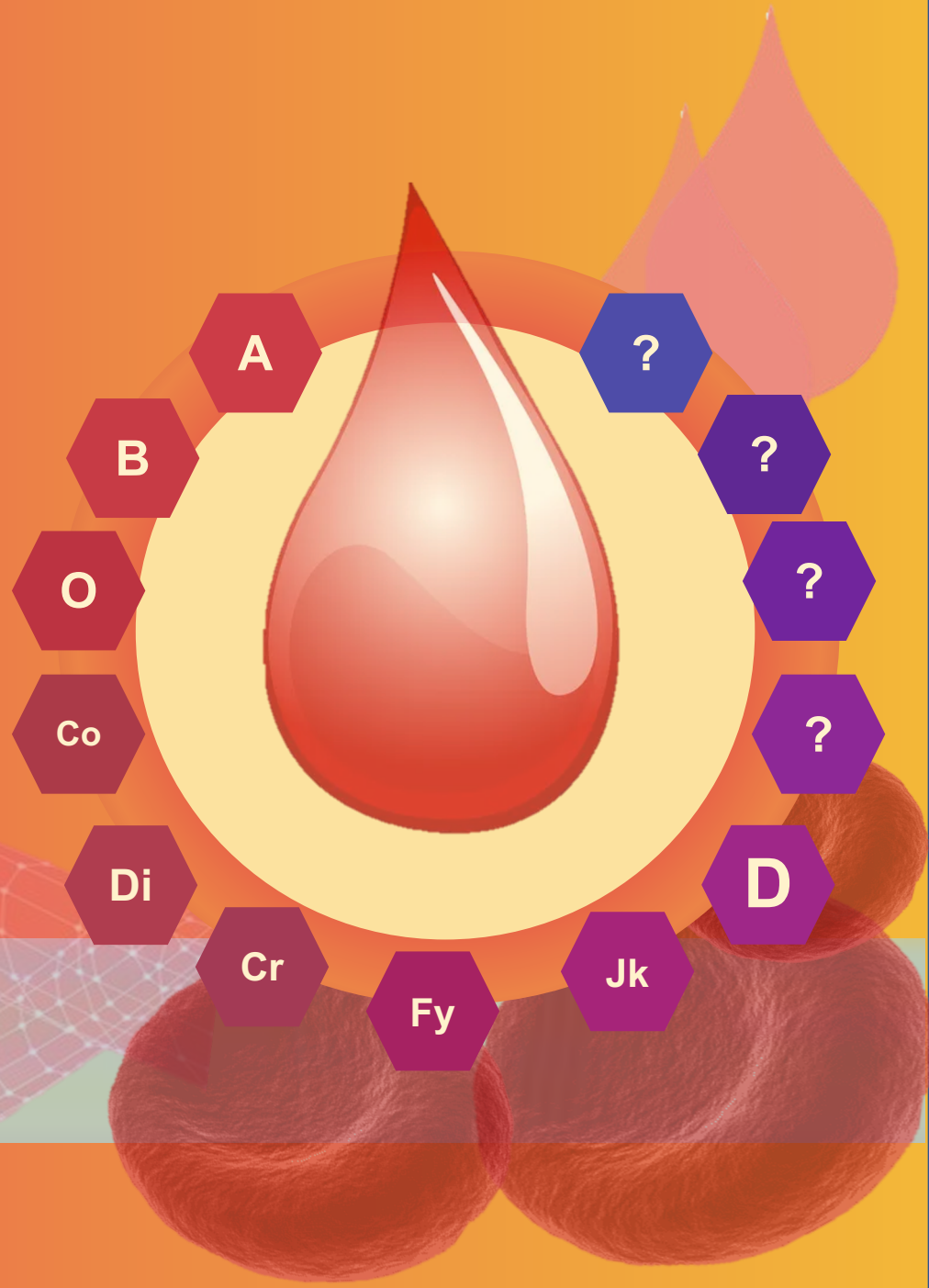
Thank You



Genomics
The future of blood grouping



Mercy Rophina
CSIR-IGIB



Genomics
The future of blood grouping

Mercy Rophina
CSIR-IGIB