If you downloaded it already yesterday, type:

cd botswanatraining
git pull

If you haven’t

git clone https://github.com/sunandoroy/botswanatraining

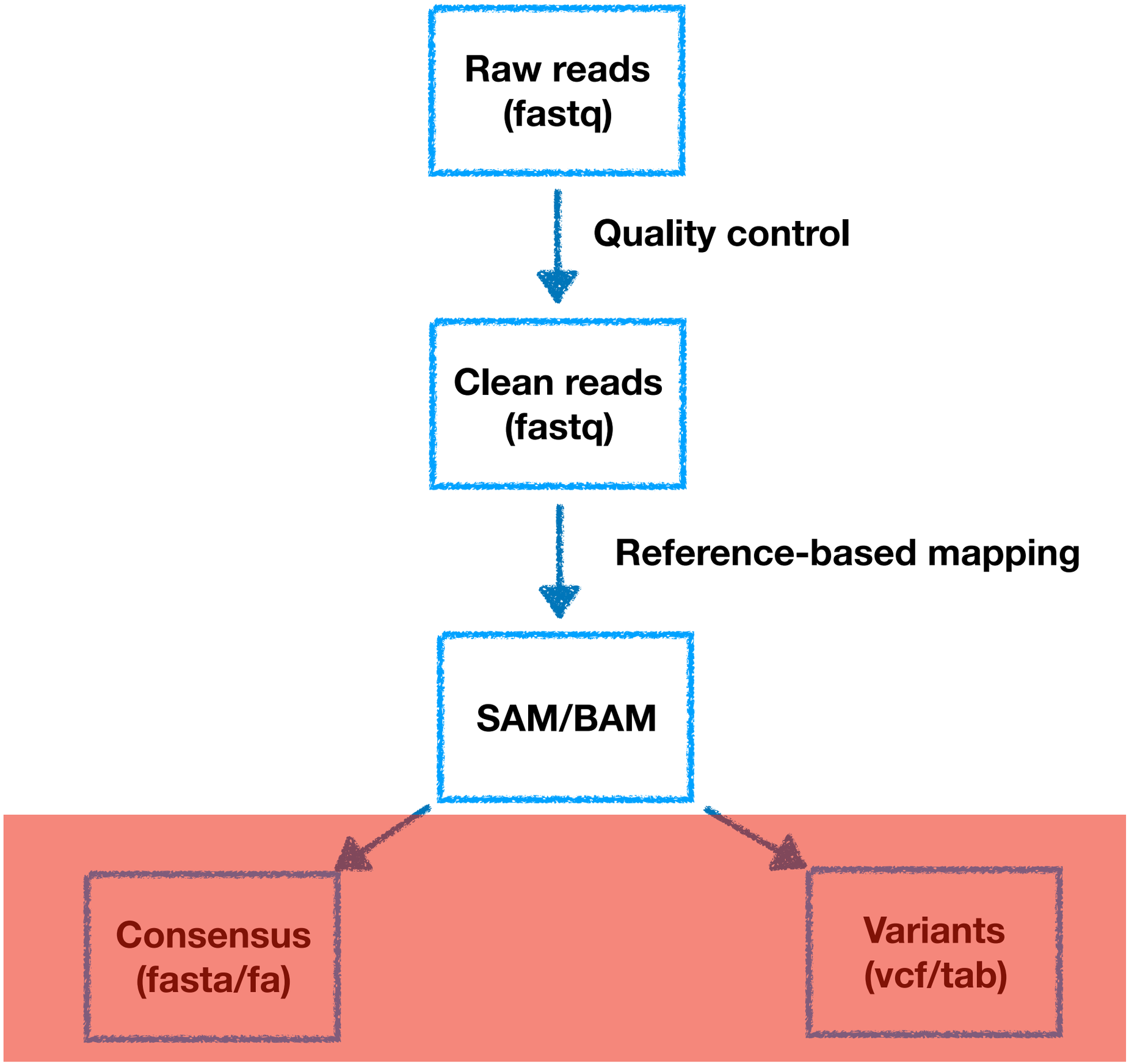
Consensus&variants calling

Dr Cristina Venturini
University College London (UCL)

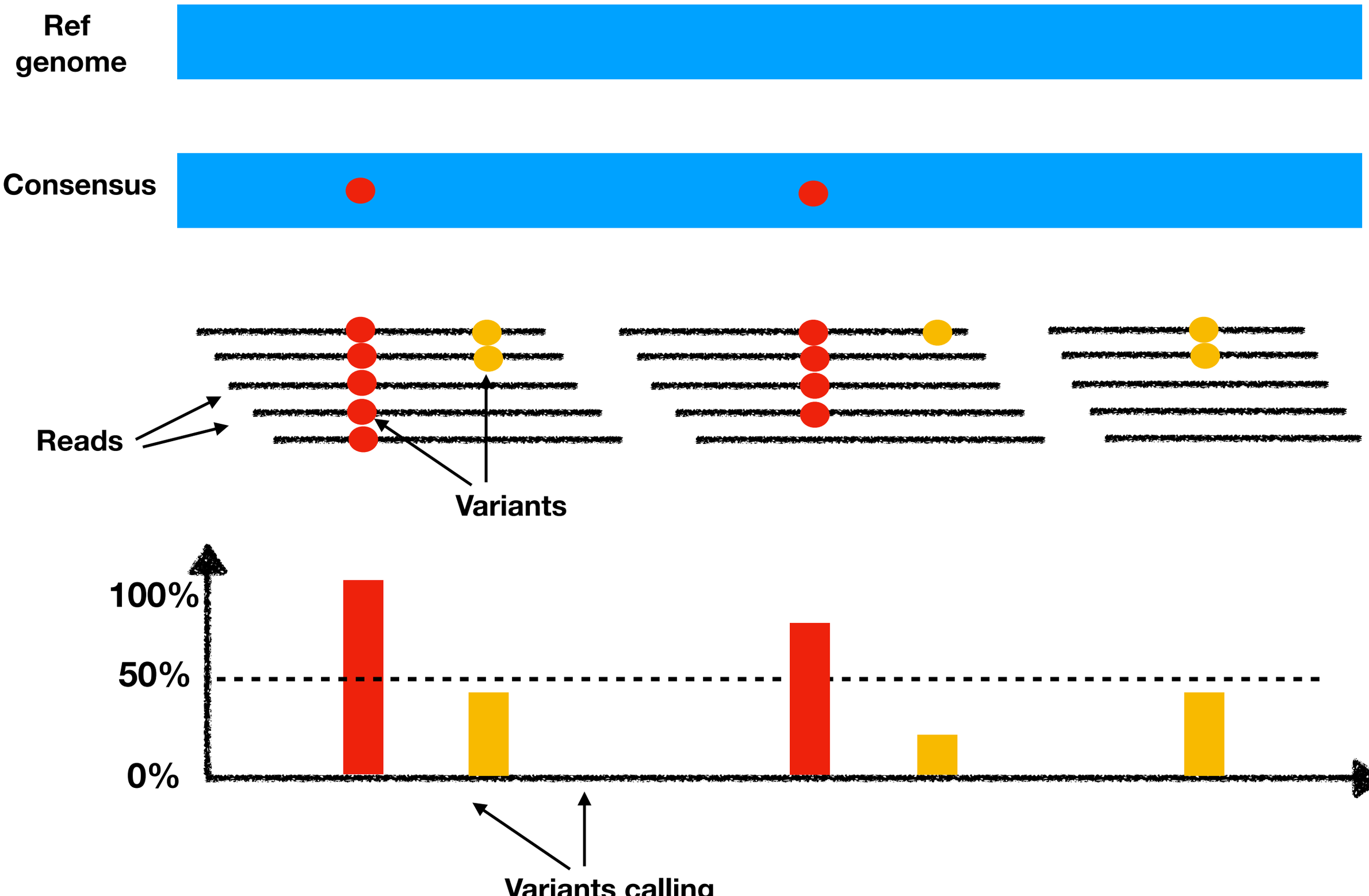
Overview

- Consensus&variants - what are they?
- What can be done with consensus sequences - examples
- What can be done with variants table - examples

Where are we in the analysis?



Consensus&variants - what are they?

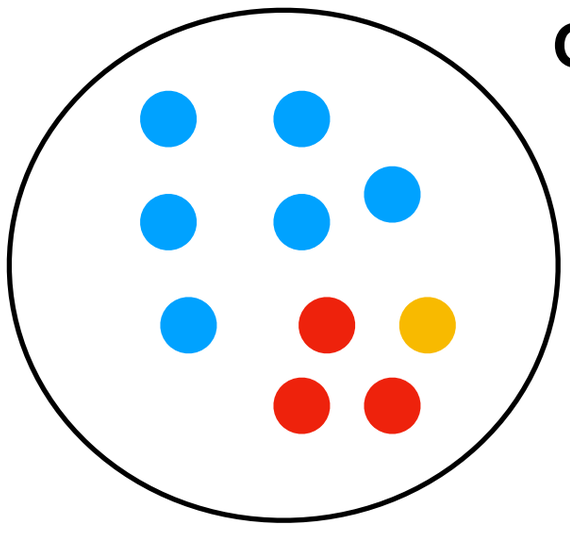


Consensus&variants - what are they?

- Consensus: the sequence of the most frequent nucleotides at each positions

- Variants: difference between a sample sequence and a reference —> variants can be frequent at different frequency

Viral populations/ Quasispecies

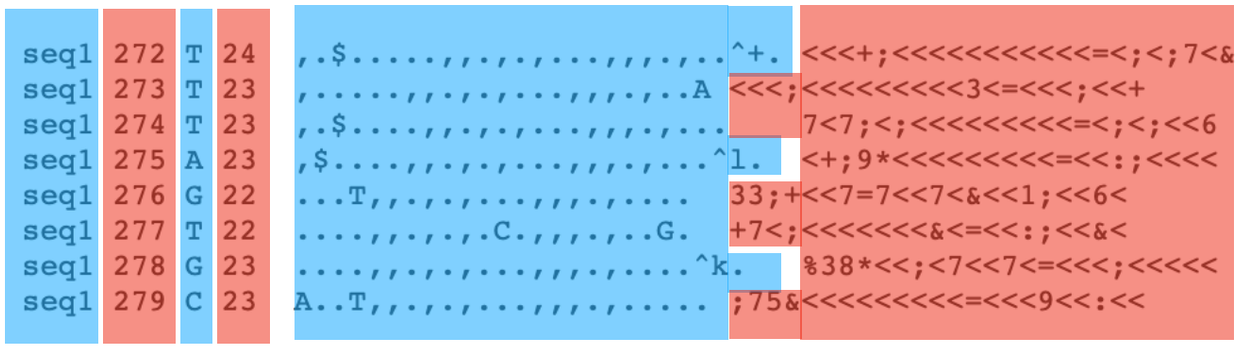


How to build a consensus?

- **From a BAM file: mpileup (samtools)**
    - + bcftools (tutorial)
    - QUASR
- **Considerations:**
    - Quality of bases within a read (phred score)
    - Quality of the read mapping
    - Insertions/deletions/variants
- **Can be very dependent upon choice of mapping software (+ parameters) and reference sequence**

Pileup file - example

Base-pair informations at each chromosomal position.



- Chromosome
- 1-based coordinate
- Reference base
- Number of reads covering the site
- Read bases
- Base qualities

- Read bases:
  - ./, Match ref base (f/r strand)
  - ACGTN/acgtn mismatch ref base (f/r)
  - ^/$ start/finish of a read

What can be done with consensus sequence?

- Drug resistance mutations at consensus level (DRM)
- How similar are two sequences? (between/within patients)
- Phylogeny analysis

Variants calling

- **From a BAM file: mpileup (samtools)**
  - + Varscan (tutorial)
  - FreeBayes
- **VCF file**

VCF file example

```
##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">
```

The header provides metadata describing the body of the file
Always start with # or ##

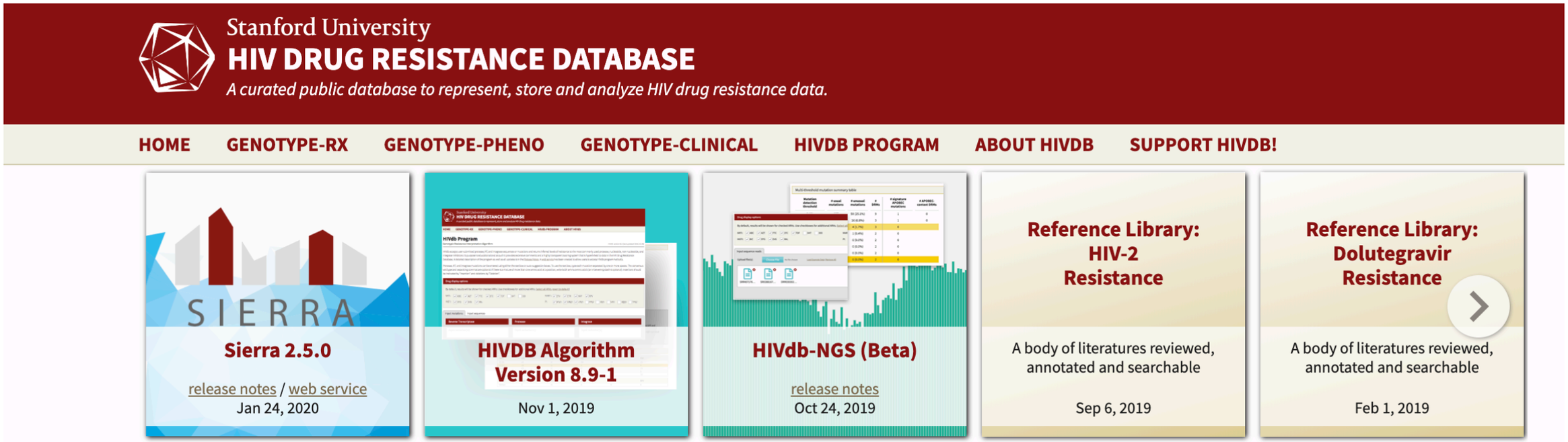
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	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
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

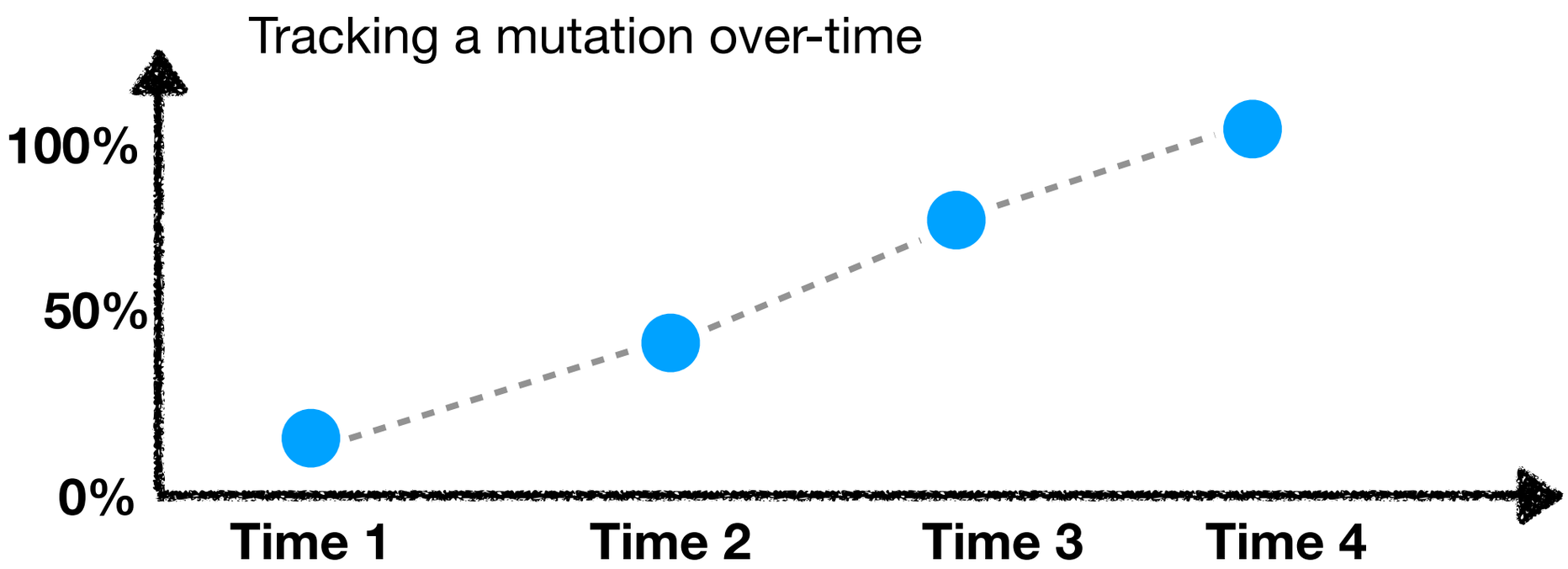
1	CHROM	The name of the sequence (typically a chromosome)
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a “."
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation (i.e. frequency, allele count, CIGAR)
9	FORMAT	An (optional) extensible list of fields for describing the samples (i.e. read depth, genotype)
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

Uses of variant analysis

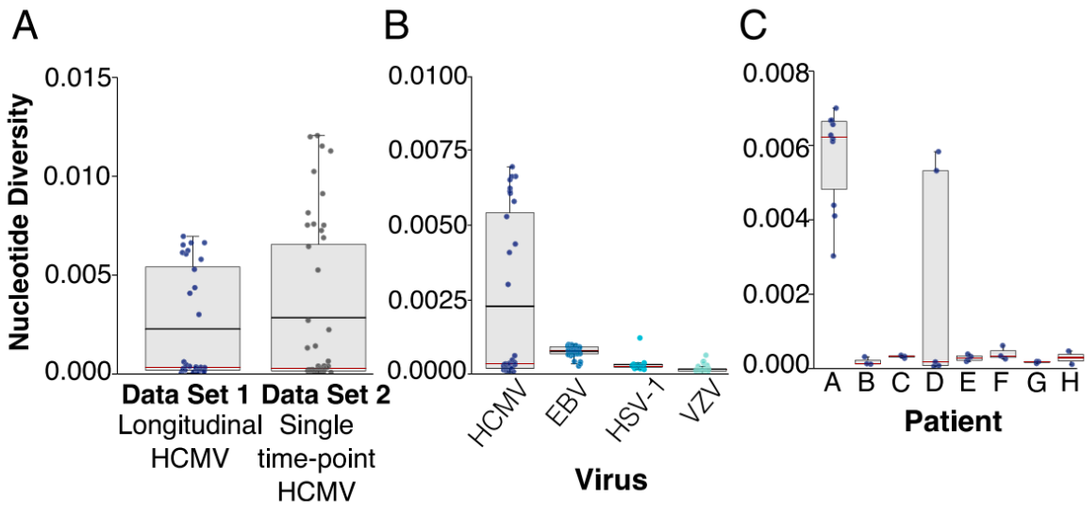
- Drug resistance mutations - tracking overtime
- Typing (genotypes)
- Quasispecies reconstructions:
    - Mixed infections
    - Recombination
    - Transmission

Drug resistance mutations

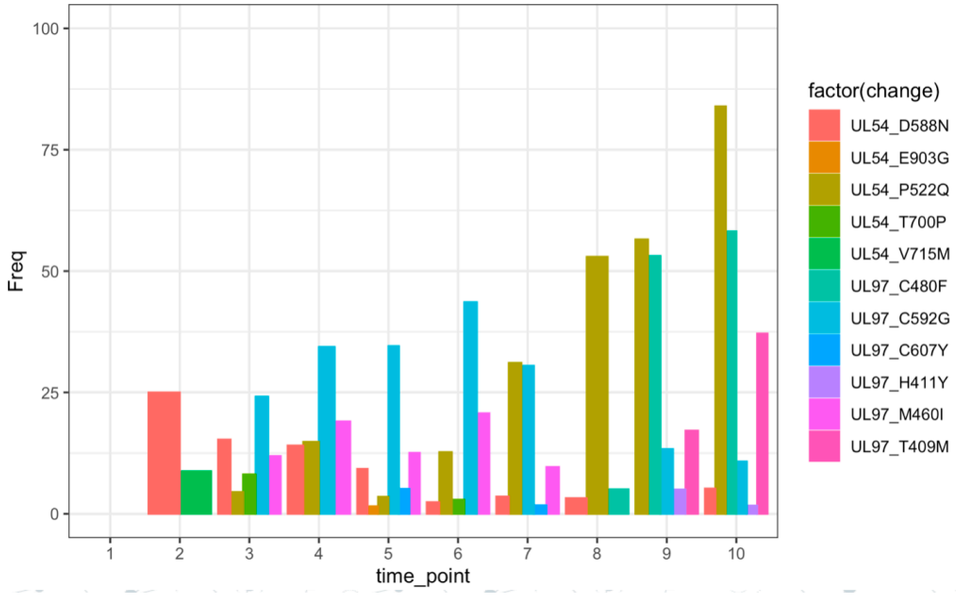


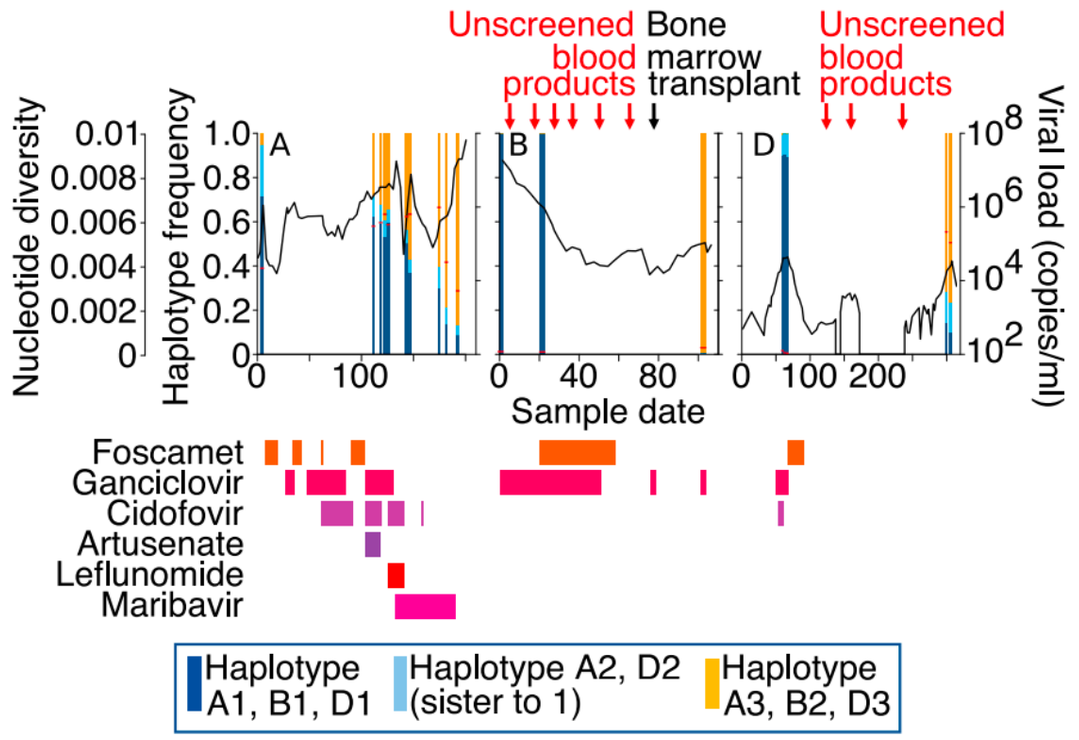


Examples in HCMV



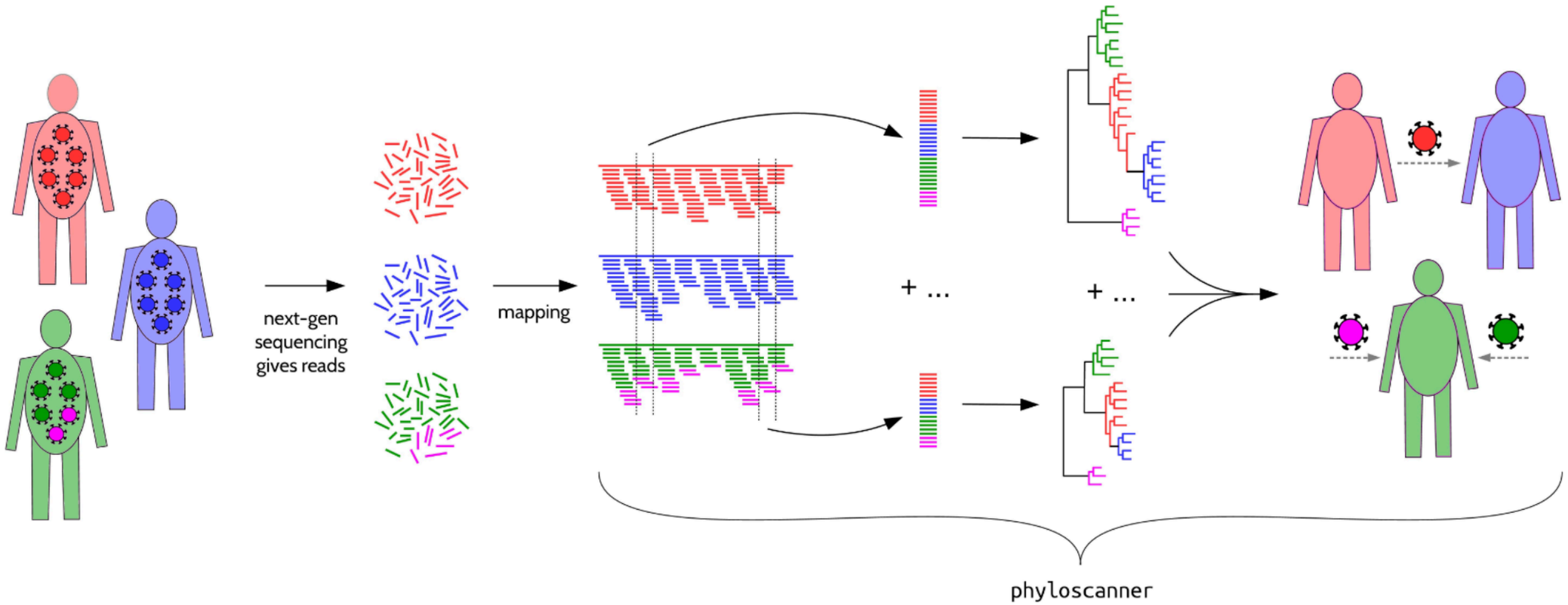
Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination (Cudini J, et al, 2019)





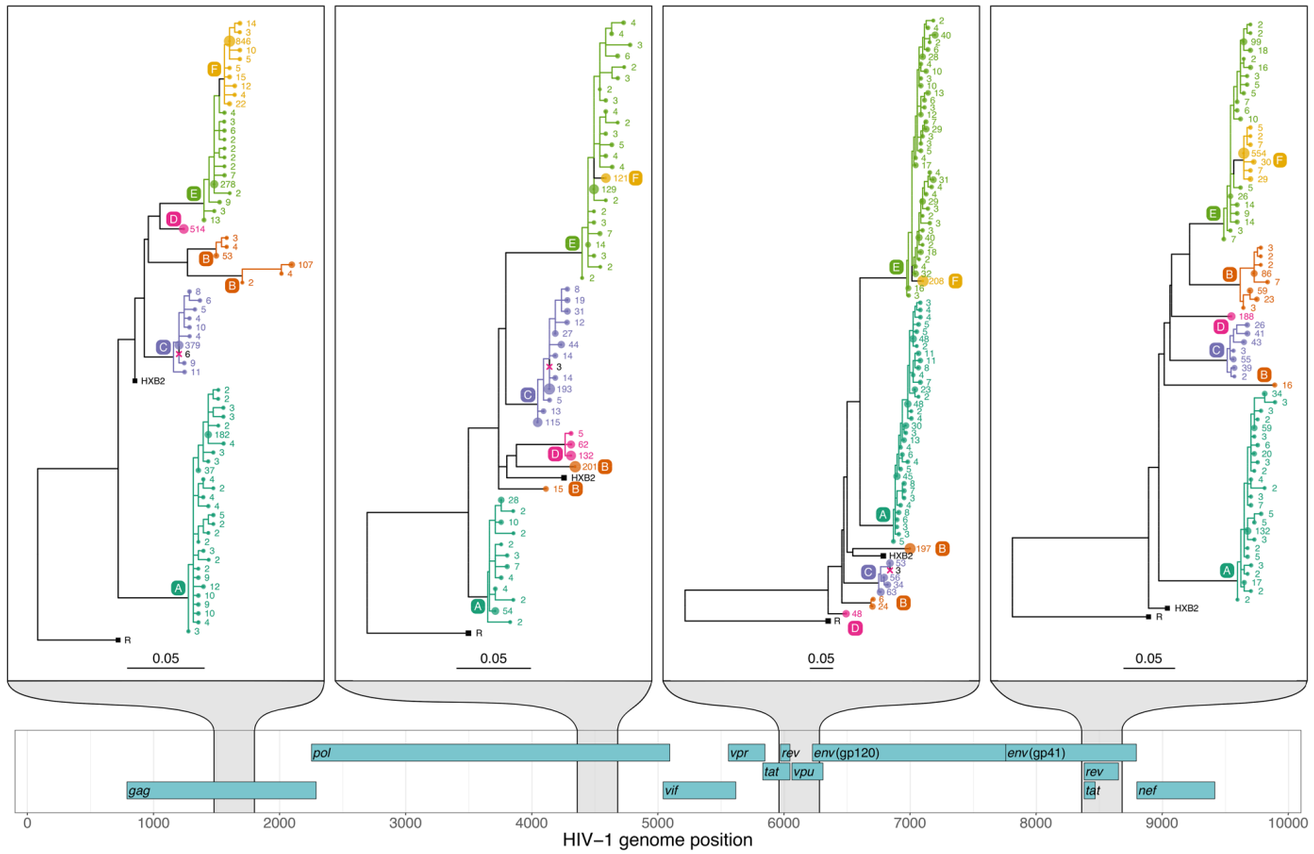
New project where we are tracking drug resistance mutations overtime and we found an association between presence of low-level DRM and poor outcome.

Transmission



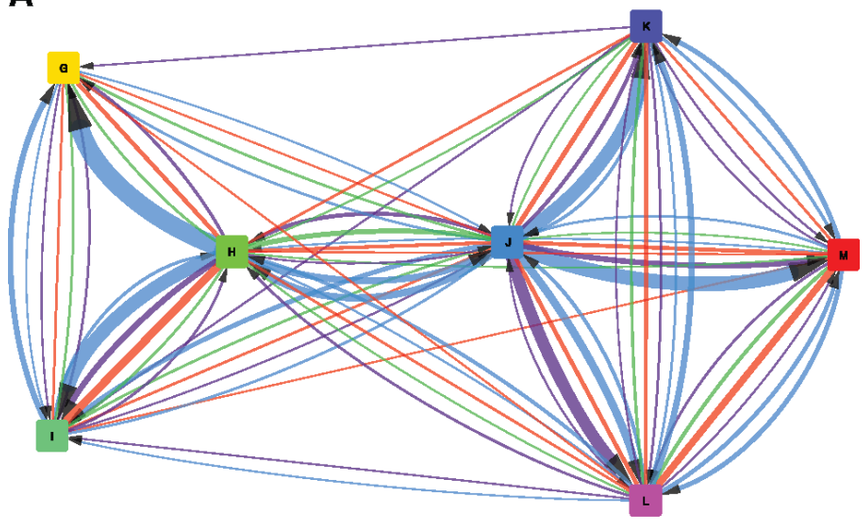
- Phyloscanner: tool to investigate diversity genetic diversity and relationships between and within hosts https://github.com/BDI-pathogens/phyloscanner

Example of Phyloscanner analysis of four illustrative windows of the HiV genome



Relationship between seven patients infected with HIV.

A



B