UK Health
Security
Agency

Consensus & Variant Calling

Dr. David Bibby,
Genomics and Clinical Virology,
5th March 2025

Overview

Consensus & Variant – what are they?

- How to build a consensus and define variants

What can be done with variant analysis - examples

- Features: Drug resistance, epitopes, species/strain identification
- Phyletics: Linkage, Dual infections, Transmission, Quasispecies reconstruction

Technical pitfalls - examples

- Virus – Laboratory – Bioinformatics

Validation, validation, validation

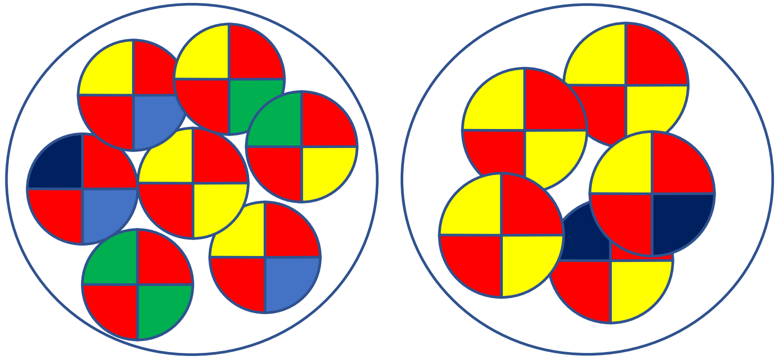
- Reproducibility
- Standardised materials
- EQA schemes
- Clinical validation?

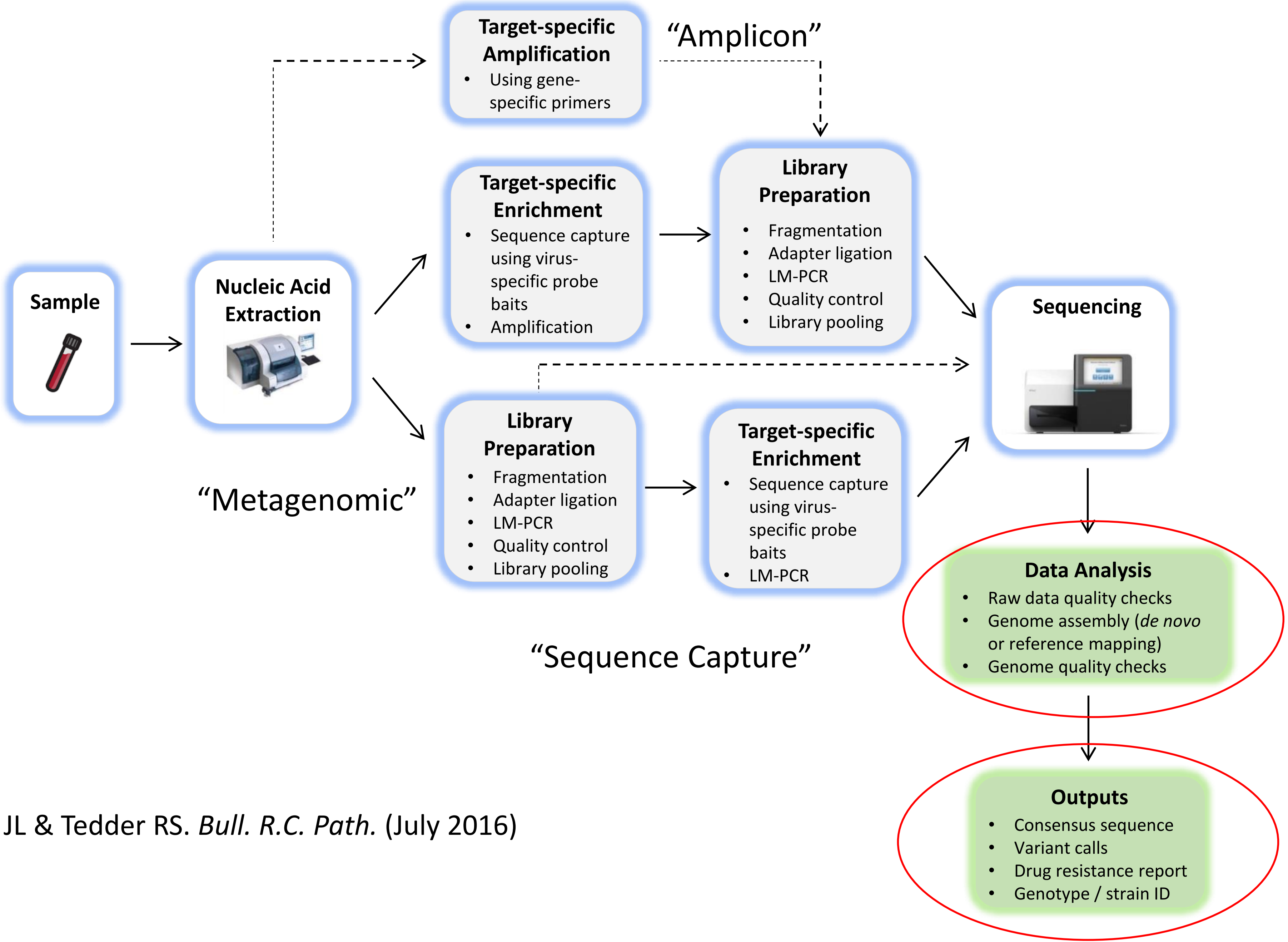
Consensus and Variant

Consensus: "The sequence of the most frequent nucleotides at each position"

Variants: “Differences between a test sequence and a reference”

- Viruses often exist in populations of related sequences, i.e.‘quasi-species’
- A consensus of a viral sequence may often contain mixed bases, incorporating the variants above a set frequency
  - e.g. 15-20% to mimic Sanger detection





Sample

Nucleic Acid Extraction

“Amplicon”

Target-specific Amplification

- Using gene-specific primers

Target-specific Enrichment

- Sequence capture using virus-specific probe baits
- Amplification

Library Preparation

- Fragmentation
- Adapter ligation
- LM-PCR
- Quality control
- Library pooling

“Metagenomic”

Library Preparation

- Fragmentation
- Adapter ligation
- LM-PCR
- Quality control
- Library pooling

Target-specific Enrichment

- Sequence capture using virus-specific probe baits
- LM-PCR

“Sequence Capture”

Sequencing

Data Analysis

- Raw data quality checks
- Genome assembly (*de novo* or reference mapping)
- Genome quality checks

Outputs

- Consensus sequence
- Variant calls
- Drug resistance report
- Genotype / strain ID

Mbisa JL & Tedder RS. *Bull. R.C. Path.* (July 2016)

How to build a consensus

Sequencer output:

- Giant file containing all sequences from all samples (and controls)
- Each read has an adapter sequence added during the sample library prep
- These enable the reads to be ‘binned’ according to sample ID

The bins are FASTQ files

- Paired end – Forward and Reverse (often R1 & R2 files)
- Adapters usually trimmed before further analysis

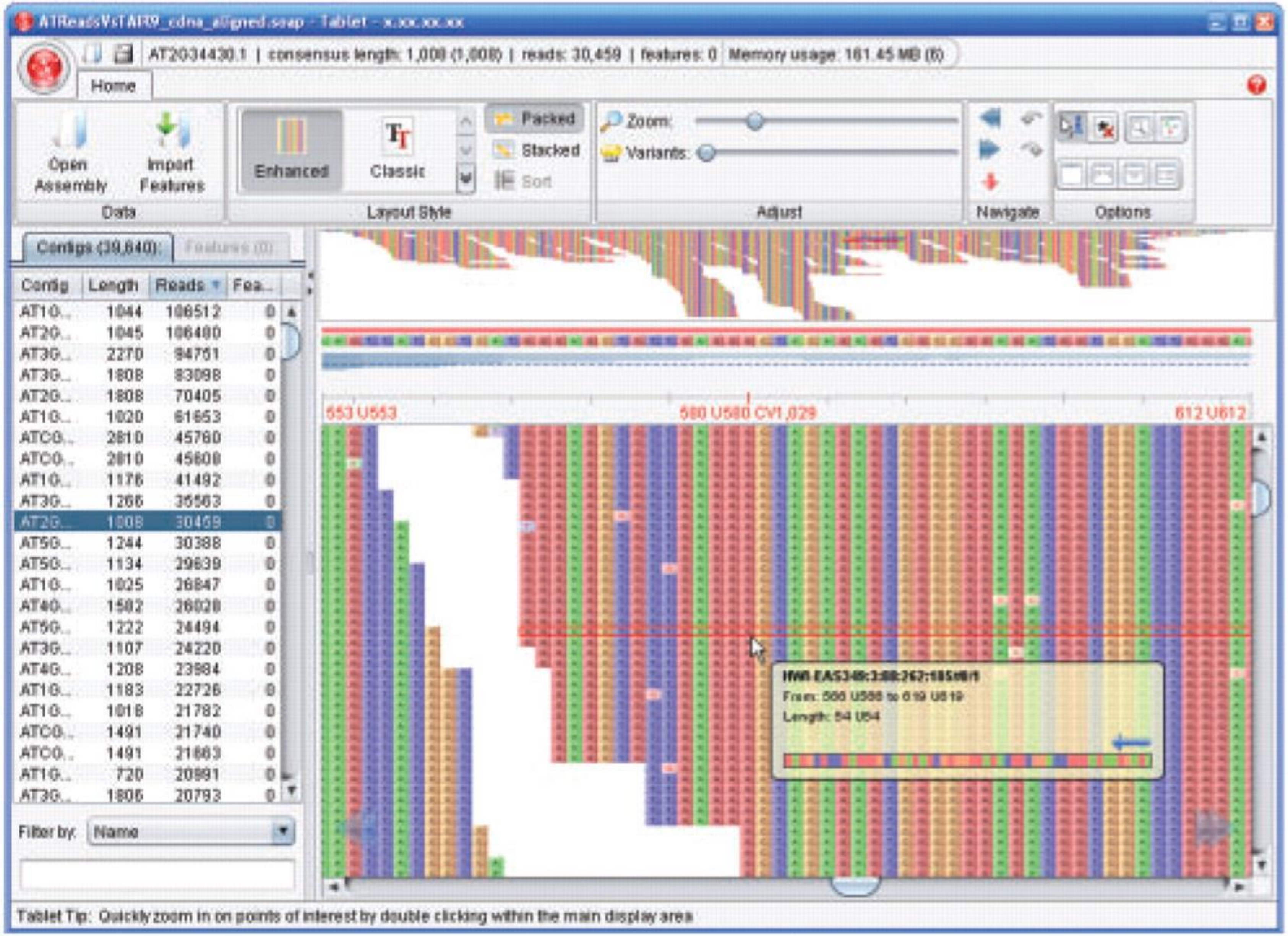
Reads are e.g. Reference Mapped → SAM file

```
bwa mem my_virus_ref.fasta sample1_R1.fastq sample1_R2.fastq > sample1.sam
```

- SAM files are converted to BAM files

```
samtools view -Sbhu sample1.sam | samtools sort > sample1.bam
samtools index sample1.bam
```

- These can be viewed in Tablet / Genious / IGV etc.



AT2G34430.1 | consensus length: 1,008 (1,008) | reads: 30,459 | features: 0 | Memory usage: 161.45 MB (6)

Contig	Length	Reads	Fea...
AT1G...	1044	108512	0
AT2G...	1045	108480	0
AT3G...	2270	94751	0
AT3G...	1808	83098	0
AT2G...	1808	70405	0
AT1G...	1020	61653	0
ATCG...	2810	45760	0
ATCG...	2810	45608	0
AT1G...	1176	41492	0
AT3G...	1266	35563	0
AT2G...	1008	30459	0
AT5G...	1244	30388	0
AT5G...	1134	29639	0
AT1G...	1025	26847	0
AT4G...	1582	26028	0
AT5G...	1222	24494	0
AT3G...	1107	24220	0
AT4G...	1208	23984	0
AT1G...	1183	22726	0
AT1G...	1018	21782	0
ATCG...	1491	21740	0
ATCG...	1491	21663	0
AT1G...	720	20891	0
AT3G...	1806	20793	0

Filter by: Name

HWI-EAS348:3:88:262:185#0/1

From: 566 U566 to 619 U619

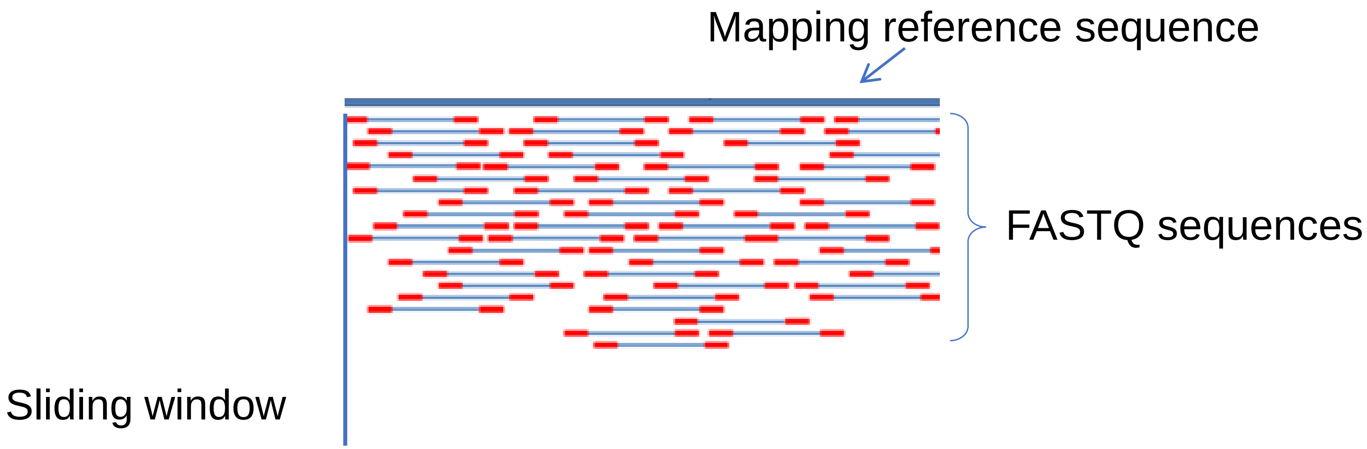
Length: 54 U54

Tablet Tip: Quickly zoom in on points of interest by double clicking within the main display area

How to build a consensus

- Several tools are available to derive consensus sequences from a SAM/BAM file:
    - `mpileup`
    - `V-Phaser`
    - `QuasiBAM`

- Slide along the sequence, interrogating all reads covering each position



Mapping reference sequence

FASTQ sequences

Sliding window

How to build a consensus

Considerations

- Quality of bases within a read

    Phred score exclusion thresholds (usually 30, sometimes 20)

- Quality of the read mapping

    Map Quality exclusion thresholds

        i.e. where the degree of homology to the reference sequence is low

        Are these contaminants or rare sequence motif(s)?

- Handling of insertions / deletions / variants

Can be very dependent upon choice of mapping software
Its parameters, and/or reference sequence

Variant calling – mpileup

mpileup (samtools) http://www.htslib.org/doc/samtools.html

1. Iterates through each position in a reference (i.e. one row per position)
2. Iterates through each read covering that position and adds a match type symbol…

`. ,`	Match to reference (forward & reverse respectively)
`^ $`	Start and finish of a read respectively
`ACGTN, acgtn`	Mismatch to reference (fwd & rev respectively)
`[+-][0-9]+[ACGTNacgtn]+`	Insertions / deletions

3. …and a Quality symbol (Phred Score)

Variant calling – mpileup

```
1  215906528  G  21  ,,,,,,,,.,,,.,,..,.,,        ;=?./:?>>;=7?>>@A?==:
1  215906529  A  18  ,,,,,,,,.,,,,,..,.           D>AA:@A>9>?;;?>>@=
[...]
1  215906547  C  15  gGg$,GggGG,,....             <;80;><9=86=C>=
1  215906548  G  19  c$,ccC.,.,,,.,....,^.        ;58610=7=>75=7<46;
[...]
1  215906555  G  16  .$aaaaaA.AAAaAAA^A           2@>?8?;<:335?:>
```

apprize.info

Variant calling – VCF

VCF “Variant Call Format”

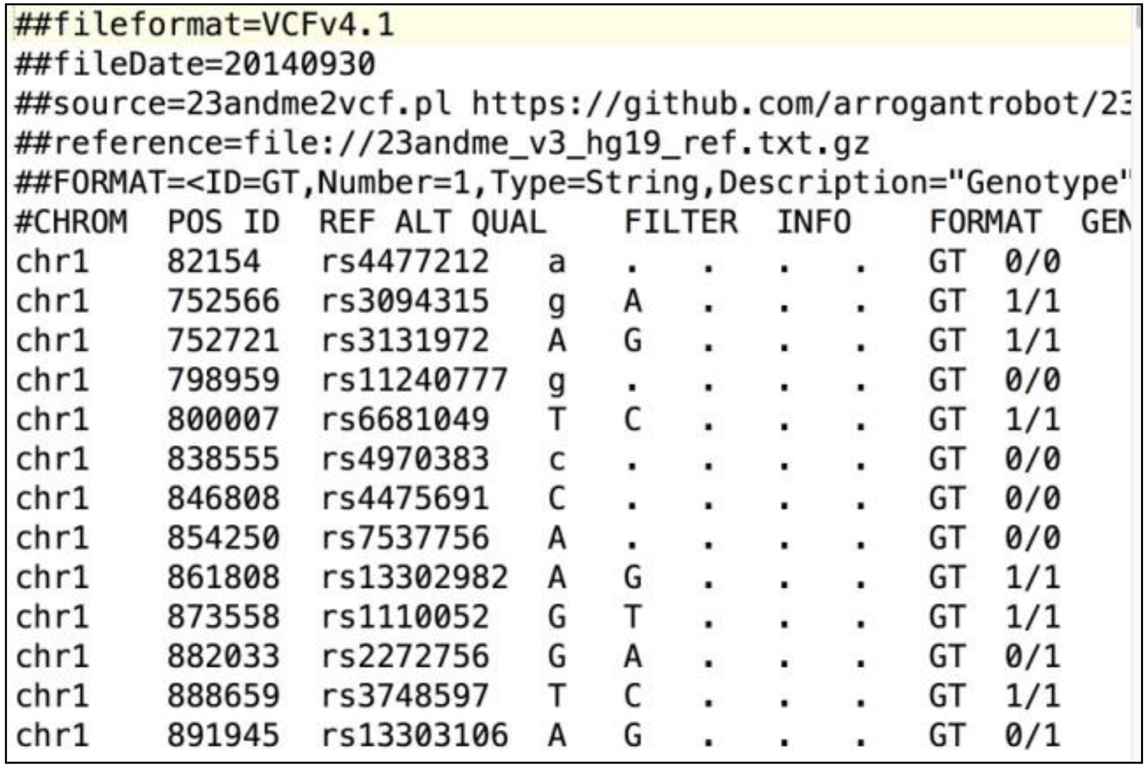
http://vcftools.sourceforge.net

- Developed for human genome annotations by 1,000 Genomes project
- Useful for sparse variation in long, multi-chromosome genomes
- Lists variations from a reference in a tabular format
  - One row per variant
  - (At least) 8 columns:

CHROM POS ID REF ALT QUAL FILTER INFO

CHROM = Chromosome
POS = Position
REF = Reference
ALT = Alternative (variant)

Variant calling – VCF



```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   rs4477212   a   .   .   .   .   GT  0/0
chr1    752566  rs3094315   g   A   .   .   .   GT  1/1
chr1    752721  rs3131972   A   G   .   .   .   GT  1/1
chr1    798959  rs11240777  g   .   .   .   .   GT  0/0
chr1    800007  rs6681049   T   C   .   .   .   GT  1/1
chr1    838555  rs4970383   c   .   .   .   .   GT  0/0
chr1    846808  rs4475691   C   .   .   .   .   GT  0/0
chr1    854250  rs7537756   A   .   .   .   .   GT  0/0
chr1    861808  rs13302982  A   G   .   .   .   GT  1/1
chr1    873558  rs1110052   G   T   .   .   .   GT  1/1
chr1    882033  rs2272756   G   A   .   .   .   GT  0/1
chr1    888659  rs3748597   T   C   .   .   .   GT  1/1
chr1    891945  rs13303106  A   G   .   .   .   GT  0/1
```

genomeintelligence.org

Variant calling – V-phaser

V-Phaser & V-Phaser 2

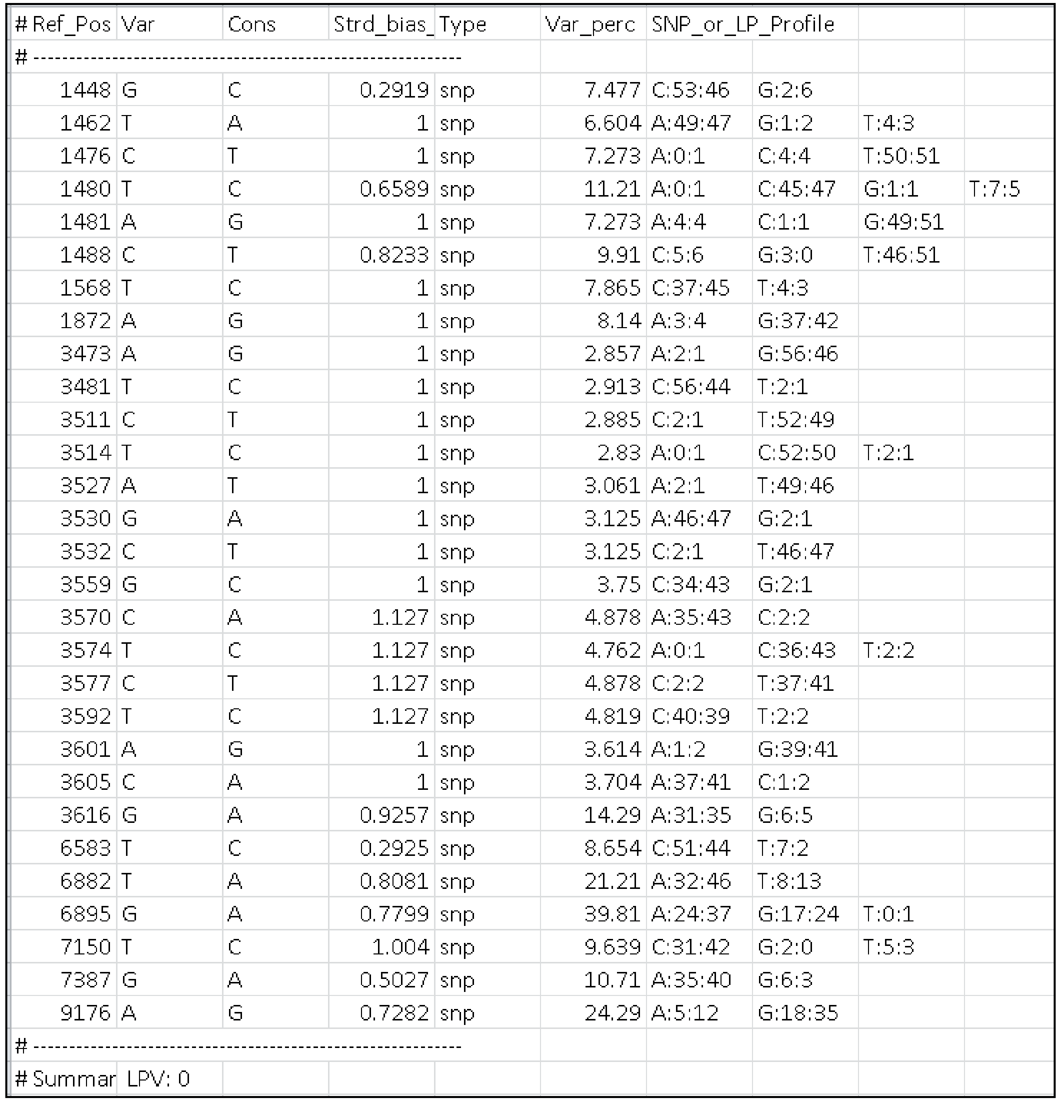
- Developed by the Broad Institute
- Considers read position, strand bias, quality scores, dinucleotide frequency, forward & reverse read, and phasing (linkage)
- Reports variant frequency and absolute read numbers by forward & reverse read
- Similar to VCF, but for viral populations
  - One row per variant
  - Seven columns:

Ref_Pos Var Cons Strd_bias_pval Type Var_perc SNP_or_LP_Profile

Macalalad AR *et al*. PLoS Computational Biology 8(3):e1002417

Yang X, *et al*. BMC Genomics 2013 14:674

Variant calling – V-phaser



# Ref_Pos	Var	Cons	Strd_bias_	Type	Var_perc	SNP_or_LP_Profile			
# --									
1448	G	C	0.2919	snp	7.477	C:53:46	G:2:6		
1462	T	A	1	snp	6.604	A:49:47	G:1:2	T:4:3	
1476	C	T	1	snp	7.273	A:0:1	C:4:4	T:50:51	
1480	T	C	0.6589	snp	11.21	A:0:1	C:45:47	G:1:1	T:7:5
1481	A	G	1	snp	7.273	A:4:4	C:1:1	G:49:51	
1488	C	T	0.8233	snp	9.91	C:5:6	G:3:0	T:46:51	
1568	T	C	1	snp	7.865	C:37:45	T:4:3		
1872	A	G	1	snp	8.14	A:3:4	G:37:42		
3473	A	G	1	snp	2.857	A:2:1	G:56:46		
3481	T	C	1	snp	2.913	C:56:44	T:2:1		
3511	C	T	1	snp	2.885	C:2:1	T:52:49		
3514	T	C	1	snp	2.83	A:0:1	C:52:50	T:2:1	
3527	A	T	1	snp	3.061	A:2:1	T:49:46		
3530	G	A	1	snp	3.125	A:46:47	G:2:1		
3532	C	T	1	snp	3.125	C:2:1	T:46:47		
3559	G	C	1	snp	3.75	C:34:43	G:2:1		
3570	C	A	1.127	snp	4.878	A:35:43	C:2:2		
3574	T	C	1.127	snp	4.762	A:0:1	C:36:43	T:2:2	
3577	C	T	1.127	snp	4.878	C:2:2	T:37:41		
3592	T	C	1.127	snp	4.819	C:40:39	T:2:2		
3601	A	G	1	snp	3.614	A:1:2	G:39:41		
3605	C	A	1	snp	3.704	A:37:41	C:1:2		
3616	G	A	0.9257	snp	14.29	A:31:35	G:6:5		
6583	T	C	0.2925	snp	8.654	C:51:44	T:7:2		
6882	T	A	0.8081	snp	21.21	A:32:46	T:8:13		
6895	G	A	0.7799	snp	39.81	A:24:37	G:17:24	T:0:1	
7150	T	C	1.004	snp	9.639	C:31:42	G:2:0	T:5:3	
7387	G	A	0.5027	snp	10.71	A:35:40	G:6:3		
9176	A	G	0.7282	snp	24.29	A:5:12	G:18:35		
# --									
# Summar	LPV: 0								

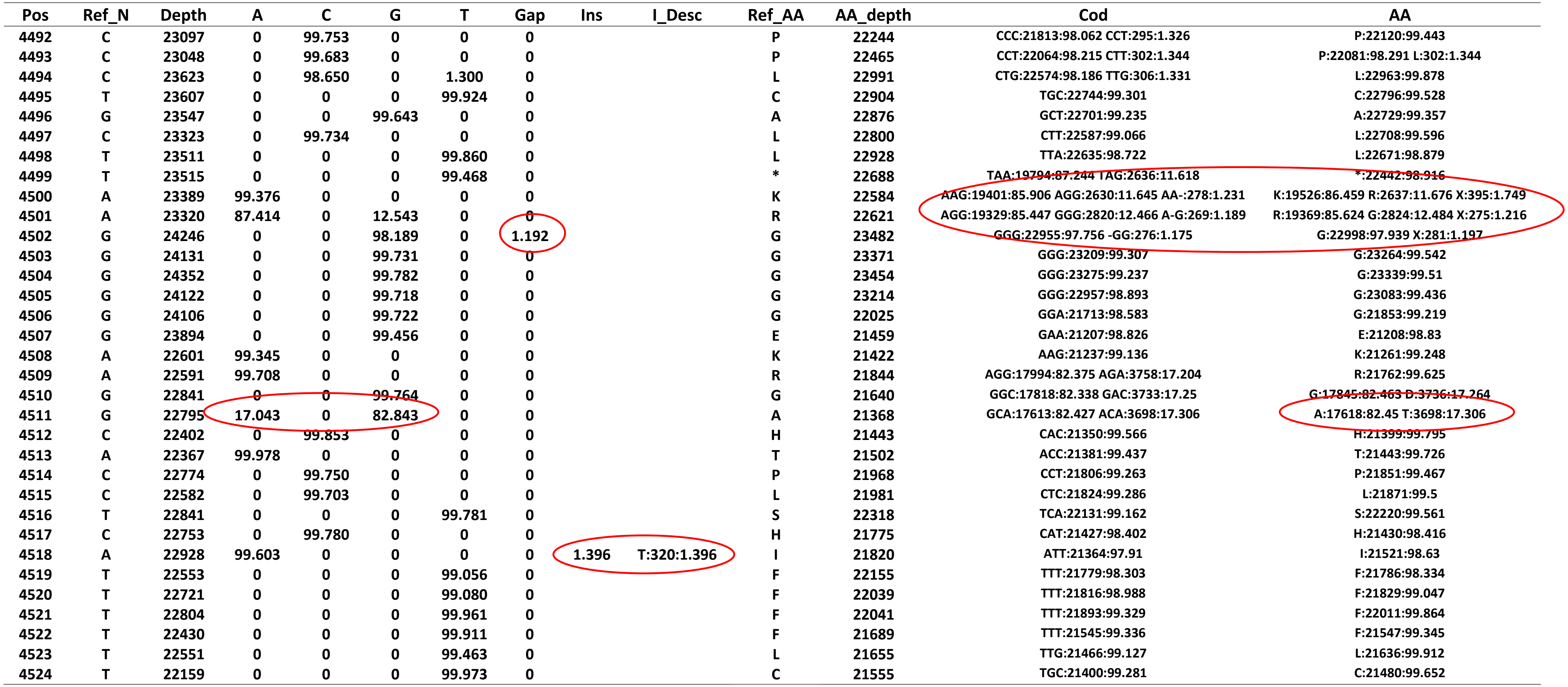
Variant calling – QuasiBAM

- **QuasiBAM** (UKHSA)
- Produces a table of nucleotide & codon frequencies for an entire reference
- One row per nucleotide position, i.e. more like mpileup
- 14 Columns:

1-3	**Position, Reference nucleotide, Depth**
4-9	**A / C / G / T / Gap / Insertion frequencies**
10	**Insertion sequences and their frequencies**
11-12	**Reference Amino Acid, Depth**
13-14	**Codon / Amino Acid frequencies**

- Can be parameterized
  - Strandedness
  - Gap-masking
  - Primer-mediated error filtering

Variant calling – QuasiBAM



Pos	Ref_N	Depth	A	C	G	T	Gap	Ins	I_Desc	Ref_AA	AA_depth	Cod	AA
4492	C	23097	0	99.753	0	0	0			P	22244	CCC:21813:98.062 CCT:295:1.326	P:22120:99.443
4493	C	23048	0	99.683	0	0	0			P	22465	CCT:22064:98.215 CTT:302:1.344	P:22081:98.291 L:302:1.344
4494	C	23623	0	98.650	0	1.300	0			L	22991	CTG:22574:98.186 TTG:306:1.331	L:22963:99.878
4495	T	23607	0	0	0	99.924	0			C	22904	TGC:22744:99.301	C:22796:99.528
4496	G	23547	0	0	99.643	0	0			A	22876	GCT:22701:99.235	A:22729:99.357
4497	C	23323	0	99.734	0	0	0			L	22800	CTT:22587:99.066	L:22708:99.596
4498	T	23511	0	0	0	99.860	0			L	22928	TTA:22635:98.722	L:22671:98.879
4499	T	23515	0	0	0	99.468	0			*	22688	TAA:19794:87.244 TAG:2636:11.618	*:22442:98.916
4500	A	23389	99.376	0	0	0	0			K	22584	AAG:19401:85.906 AGG:2630:11.645 AA-:278:1.231	K:19526:86.459 R:2637:11.676 X:395:1.749
4501	A	23320	87.414	0	12.543	0	0			R	22621	AGG:19329:85.447 GGG:2820:12.466 A-G:269:1.189	R:19369:85.624 G:2824:12.484 X:275:1.216
4502	G	24246	0	0	98.189	0	1.192			G	23482	GGG:22955:97.756 -GG:276:1.175	G:22998:97.939 X:281:1.197
4503	G	24131	0	0	99.731	0	0			G	23371	GGG:23209:99.307	G:23264:99.542
4504	G	24352	0	0	99.782	0	0			G	23454	GGG:23275:99.237	G:23339:99.51
4505	G	24122	0	0	99.718	0	0			G	23214	GGG:22957:98.893	G:23083:99.436
4506	G	24106	0	0	99.722	0	0			G	22025	GGA:21713:98.583	G:21853:99.219
4507	G	23894	0	0	99.456	0	0			E	21459	GAA:21207:98.826	E:21208:98.83
4508	A	22601	99.345	0	0	0	0			K	21422	AAG:21237:99.136	K:21261:99.248
4509	A	22591	99.708	0	0	0	0			R	21844	AGG:17994:82.375 AGA:3758:17.204	R:21762:99.625
4510	G	22841	0	0	99.764	0	0			G	21640	GGC:17818:82.338 GAC:3733:17.25	G:17845:82.463 D:3736:17.264
4511	G	22795	17.043	0	82.843	0	0			A	21368	GCA:17613:82.427 ACA:3698:17.306	A:17618:82.45 T:3698:17.306
4512	C	22402	0	99.853	0	0	0			H	21443	CAC:21350:99.566	H:21399:99.795
4513	A	22367	99.978	0	0	0	0			T	21502	ACC:21381:99.437	T:21443:99.726
4514	C	22774	0	99.750	0	0	0			P	21968	CCT:21806:99.263	P:21851:99.467
4515	C	22582	0	99.703	0	0	0			L	21981	CTC:21824:99.286	L:21871:99.5
4516	T	22841	0	0	0	99.781	0			S	22318	TCA:22131:99.162	S:22220:99.561
4517	C	22753	0	99.780	0	0	0			H	21775	CAT:21427:98.402	H:21430:98.416
4518	A	22928	99.603	0	0	0	0	1.396	T:320:1.396	I	21820	ATT:21364:97.91	I:21521:98.63
4519	T	22553	0	0	0	99.056	0			F	22155	TTT:21779:98.303	F:21786:98.334
4520	T	22721	0	0	0	99.080	0			F	22039	TTT:21816:98.988	F:21829:99.047
4521	T	22804	0	0	0	99.961	0			F	22041	TTT:21893:99.329	F:22011:99.864
4522	T	22430	0	0	0	99.911	0			F	21689	TTT:21545:99.336	F:21547:99.345
4523	T	22551	0	0	0	99.463	0			L	21655	TTG:21466:99.127	L:21636:99.912
4524	T	22159	0	0	0	99.973	0			C	21555	TGC:21400:99.281	C:21480:99.652

Uses of variant analysis

Features

- Typing
- Resistance
- Epitopes

Quasispecies reconstruction

- Linkage
- Dual infections
- Transmission

Uses of variant analysis - Features

Here, the consensus can be submitted to ‘conventional’ tools for interpretation

Stanford University
HIV DRUG RESISTANCE DATABASE
A curated public database to represent, store and analyze HIV drug resistance data.

geno2pheno[ngs-freq]

IRD Influenza Research Database

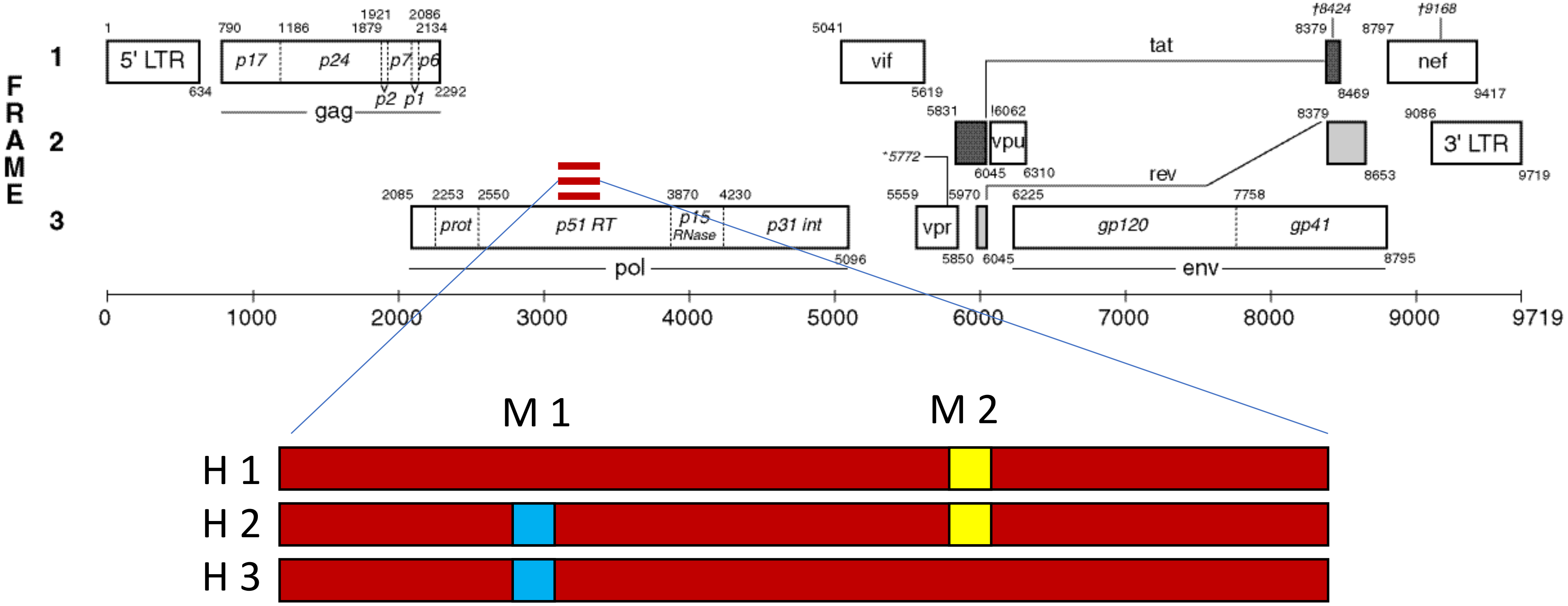
DENGUE, ZIKA & CHIKUNGUNYA
VIRUSES TYPING TOOL

- Use the variant calling tool(s) to produce consensuses at different mixed-base thresholds to interrogate minority variants.
  - *Stanford* (HIV) and *geno2pheno* (HCV) can already take nucleotide frequency files

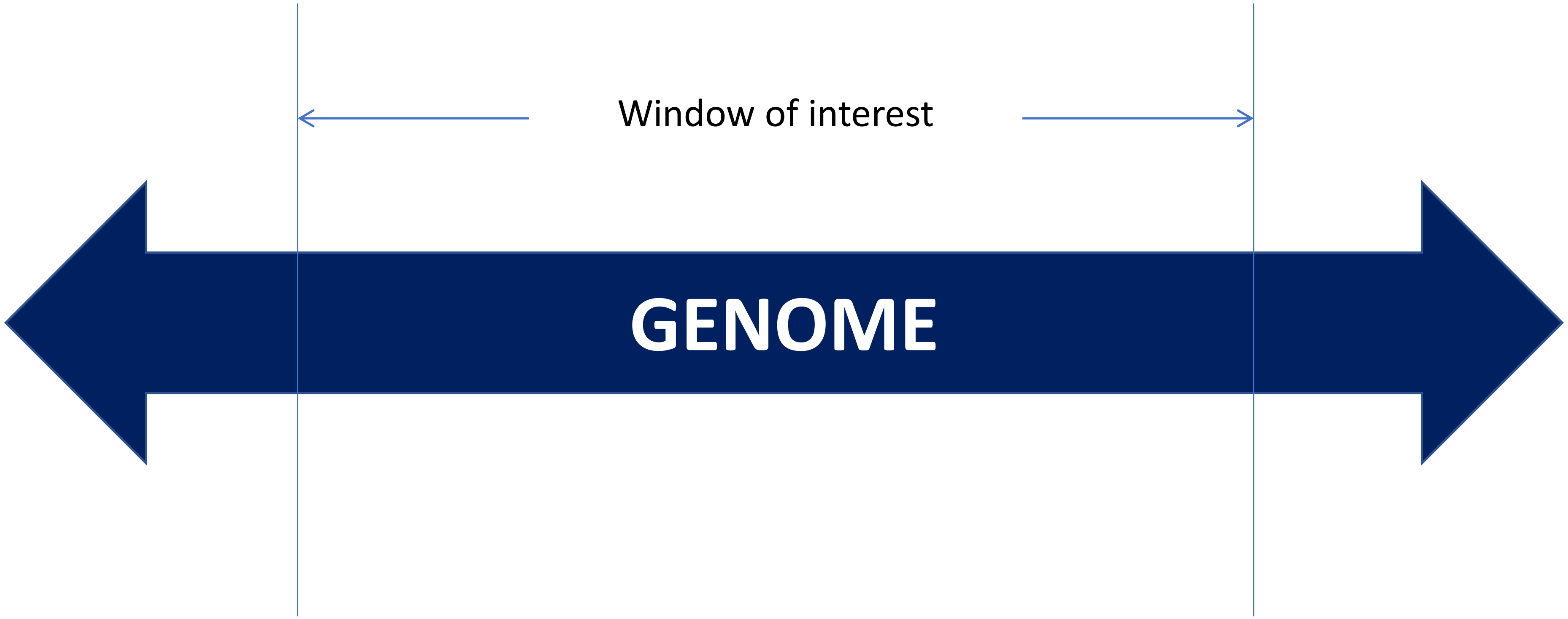
Validation, validation, validation

Uses of variant analysis - Quasispecies

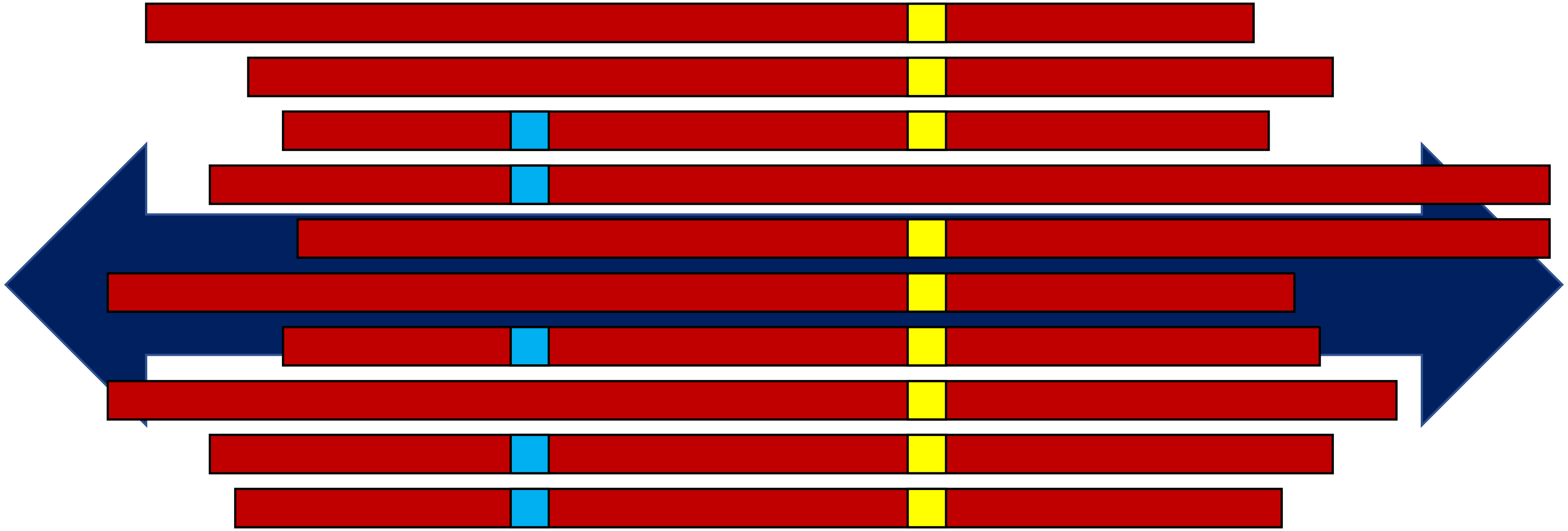
- Each read derives from an individual virus genome molecule
- Linkage of variants on reads enables binning of haplotypes
- Examine all reads that map across a short, specified region of the genome:



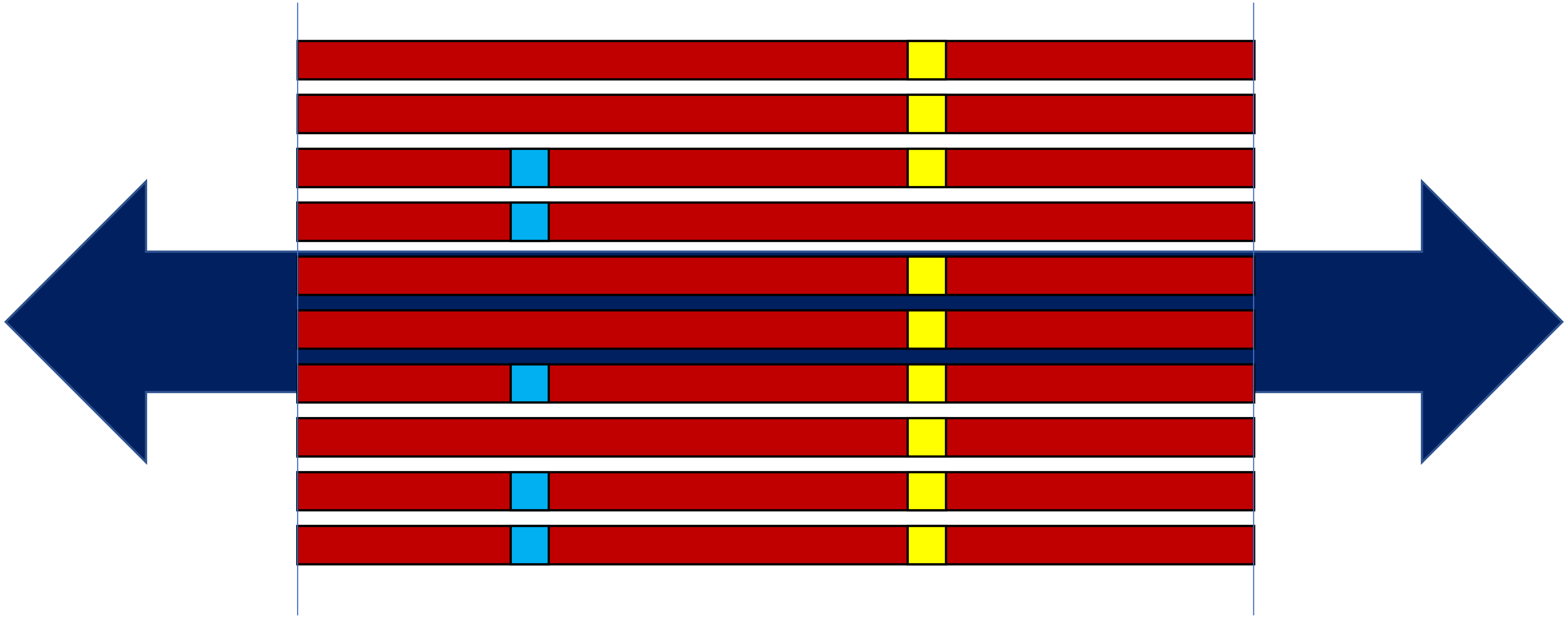
Uses of variant analysis - Quasispecies



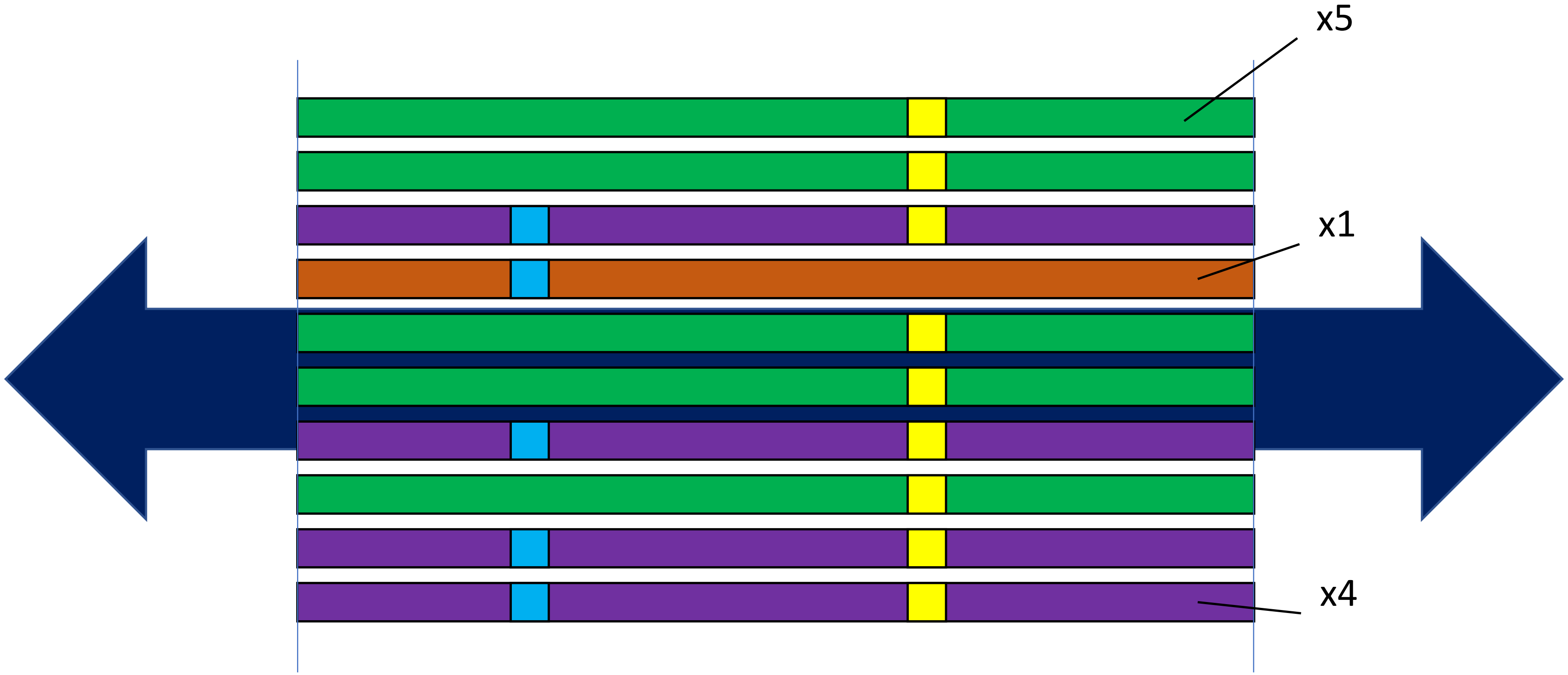
Uses of variant analysis - Quasispecies



Uses of variant analysis - Quasispecies



Uses of variant analysis - Quasispecies



Uses of variant analysis - Quasispecies

- Local data can be expanded to generate longer haplotypes

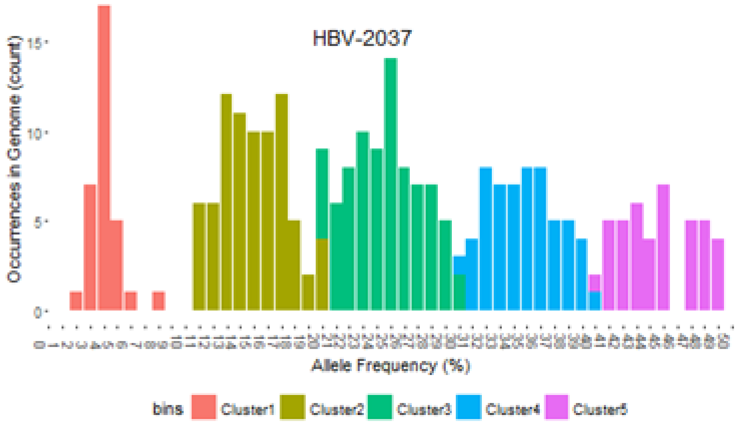
Haploclique	https://github.com/cbg-ethz/haploclique
QuasiRecomb	https://github.com/cbg-ethz/QuasiRecomb
QuRe	https://sourceforge.net/projects/qure
PredictHaplo	http://bmda.cs.unibas.ch/software.html

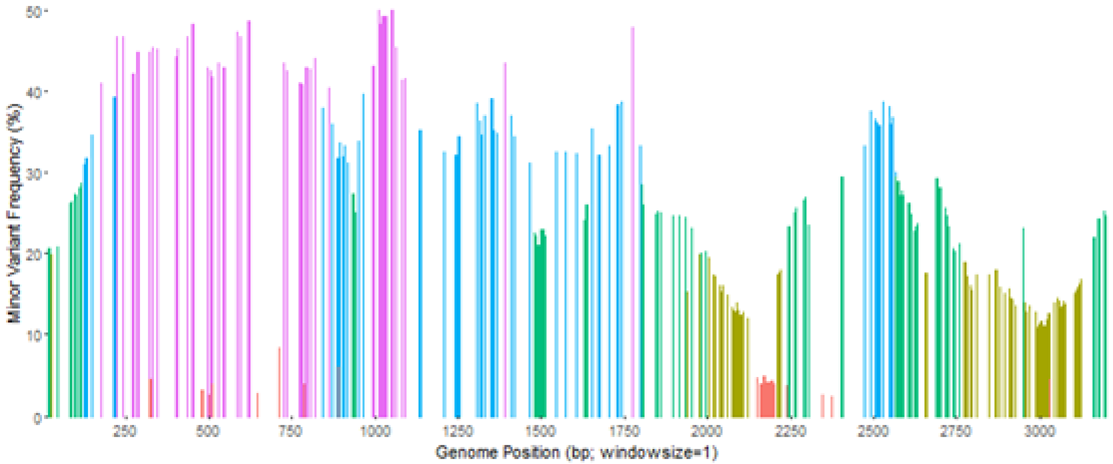
Efficiency of reconstruction is “varied”!

Beerenwinkel N *et al*. Front Microbiol. 2012 3:329
Prosperi MCF *et al*. Sci Rep. 2013 3:2837

Uses of variant analysis - Quasispecies

- Correlate mutation frequencies across the genome

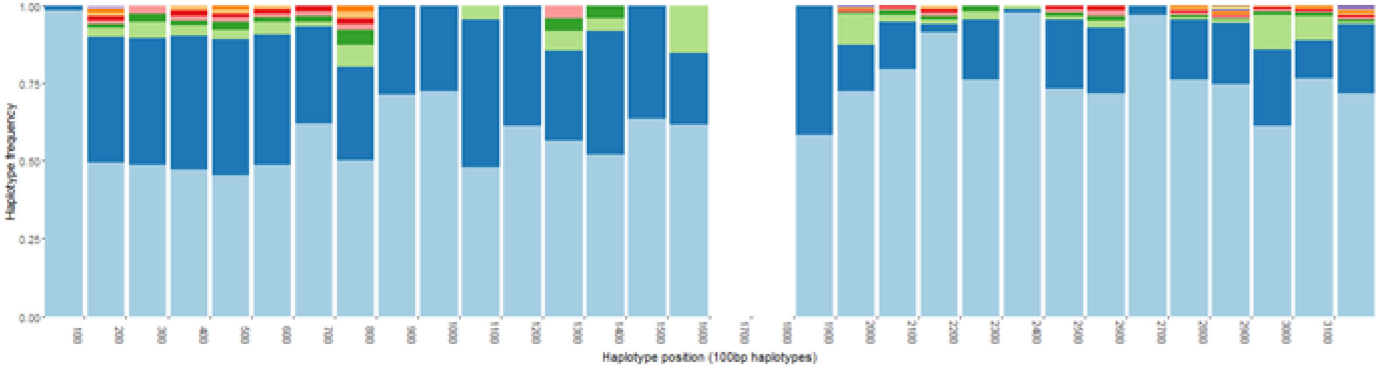




Mathew Beale

Uses of variant analysis - Quasispecies

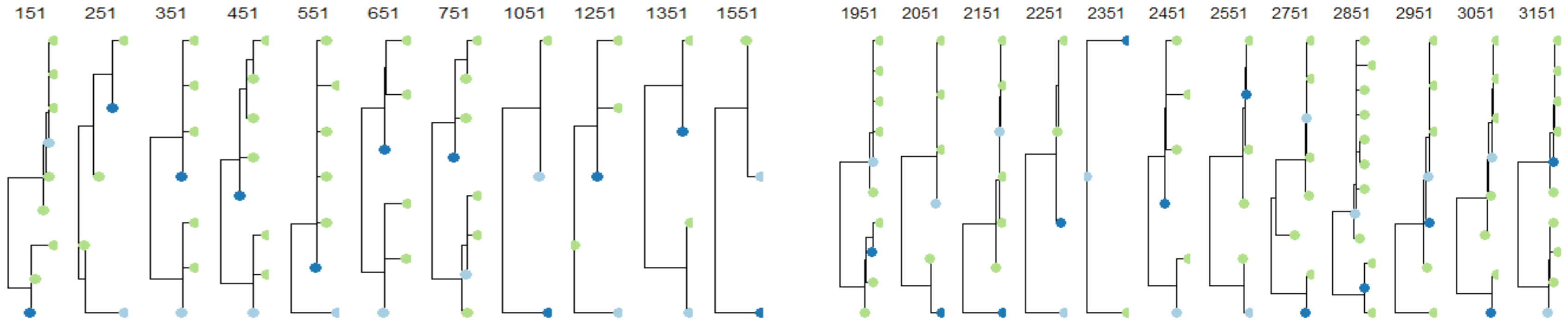
- Correlate mutation frequencies across the genome



Mathew Beale

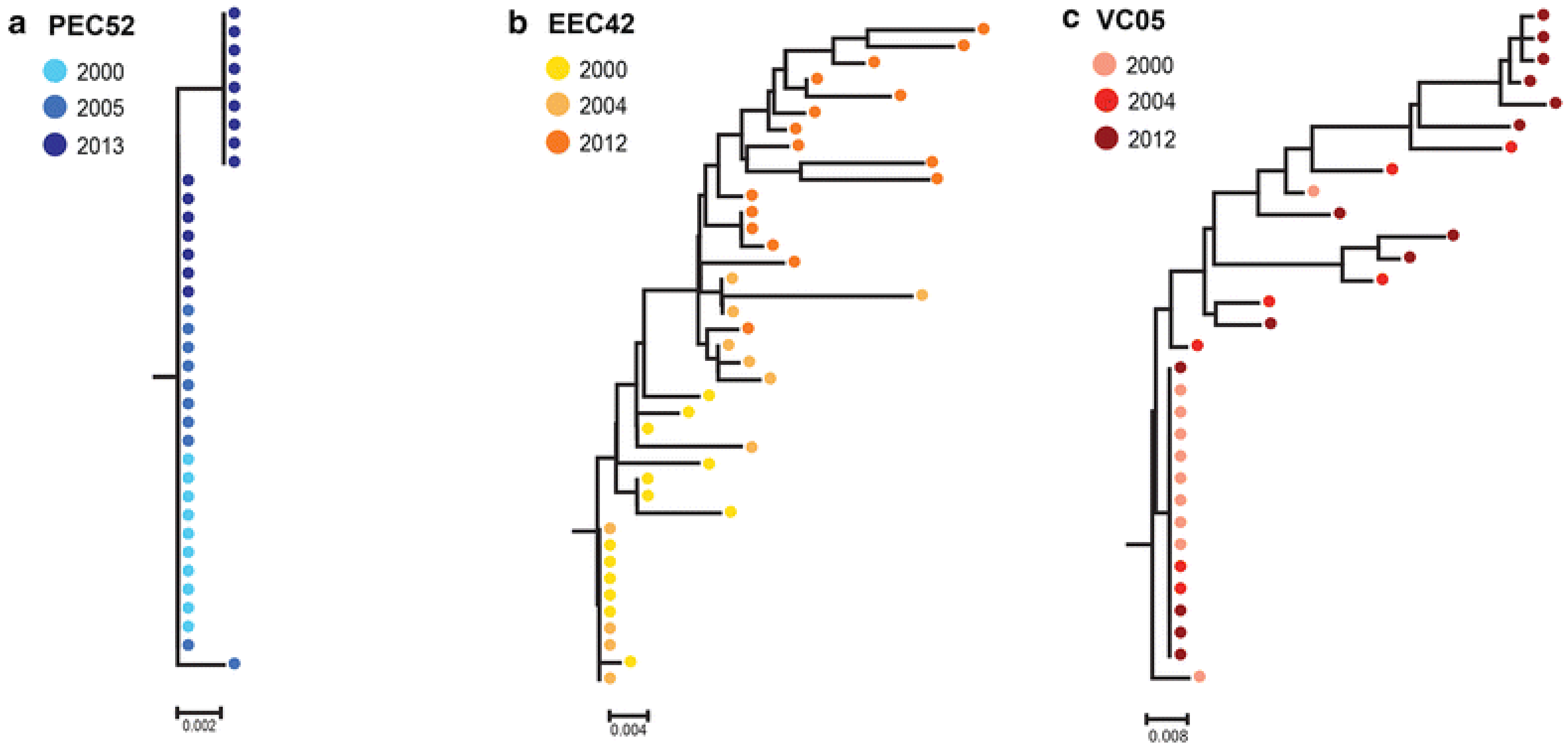
Uses of variant analysis - Quasispecies

- Correlate mutation frequencies across the genome



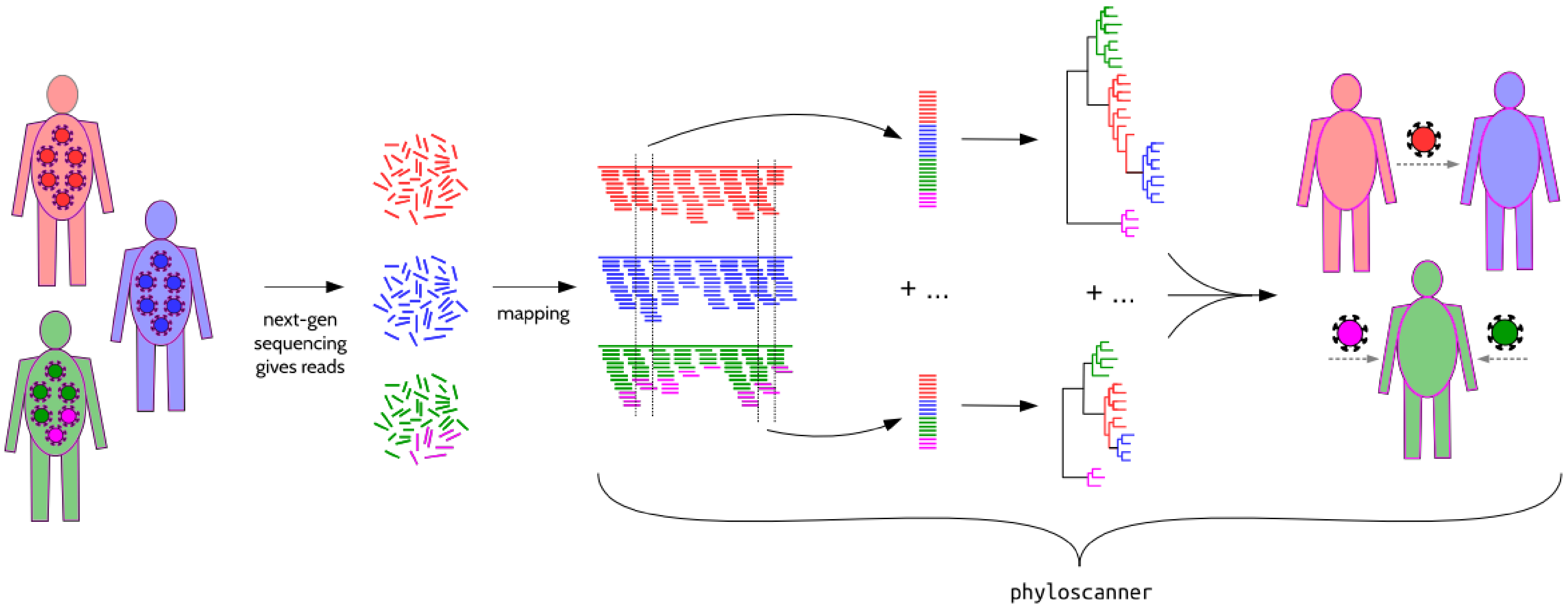
Mathew Beale

Uses of variant analysis - Quasispecies



de Azevedo SSD *et al*. Retrovirology 2017 14:29

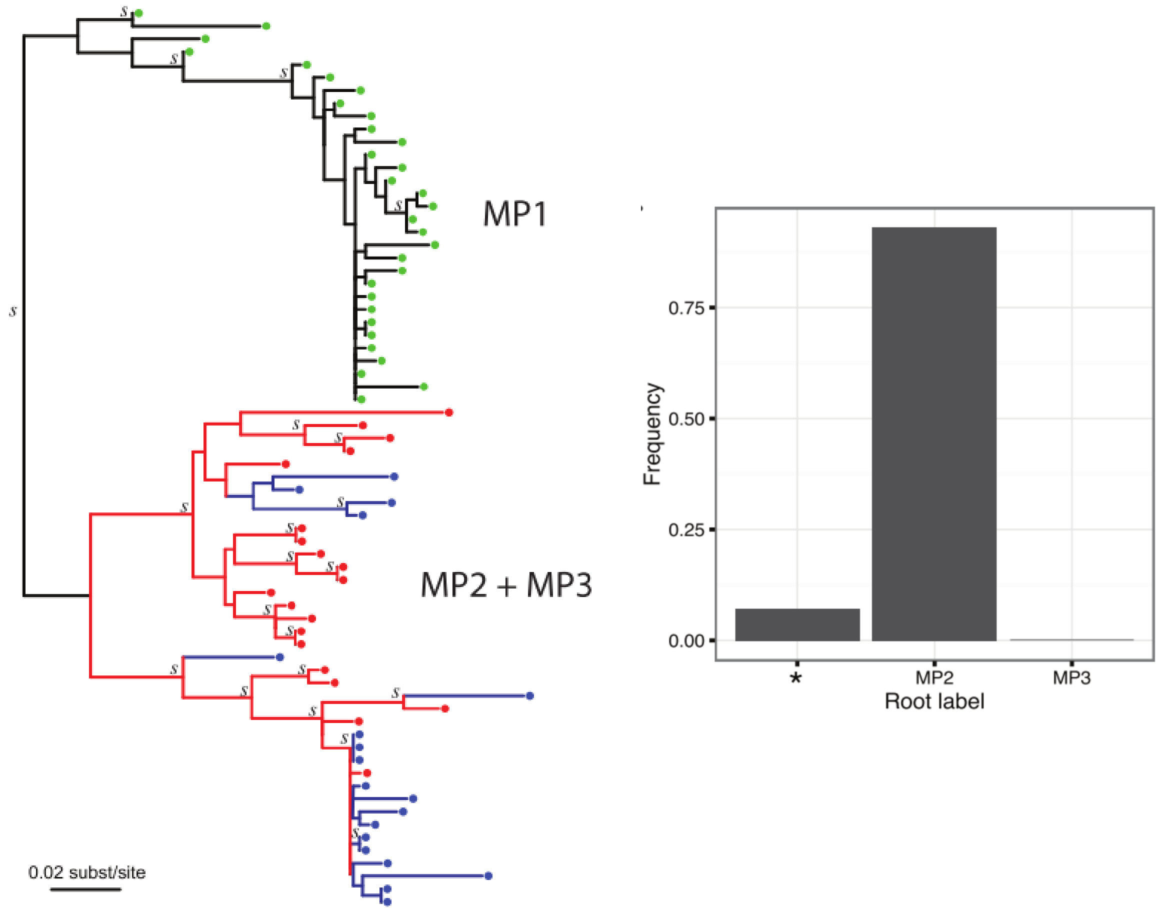
Uses of variant analysis - Quasispecies



Wymant C *et al*. Mol Biol Evol 2017

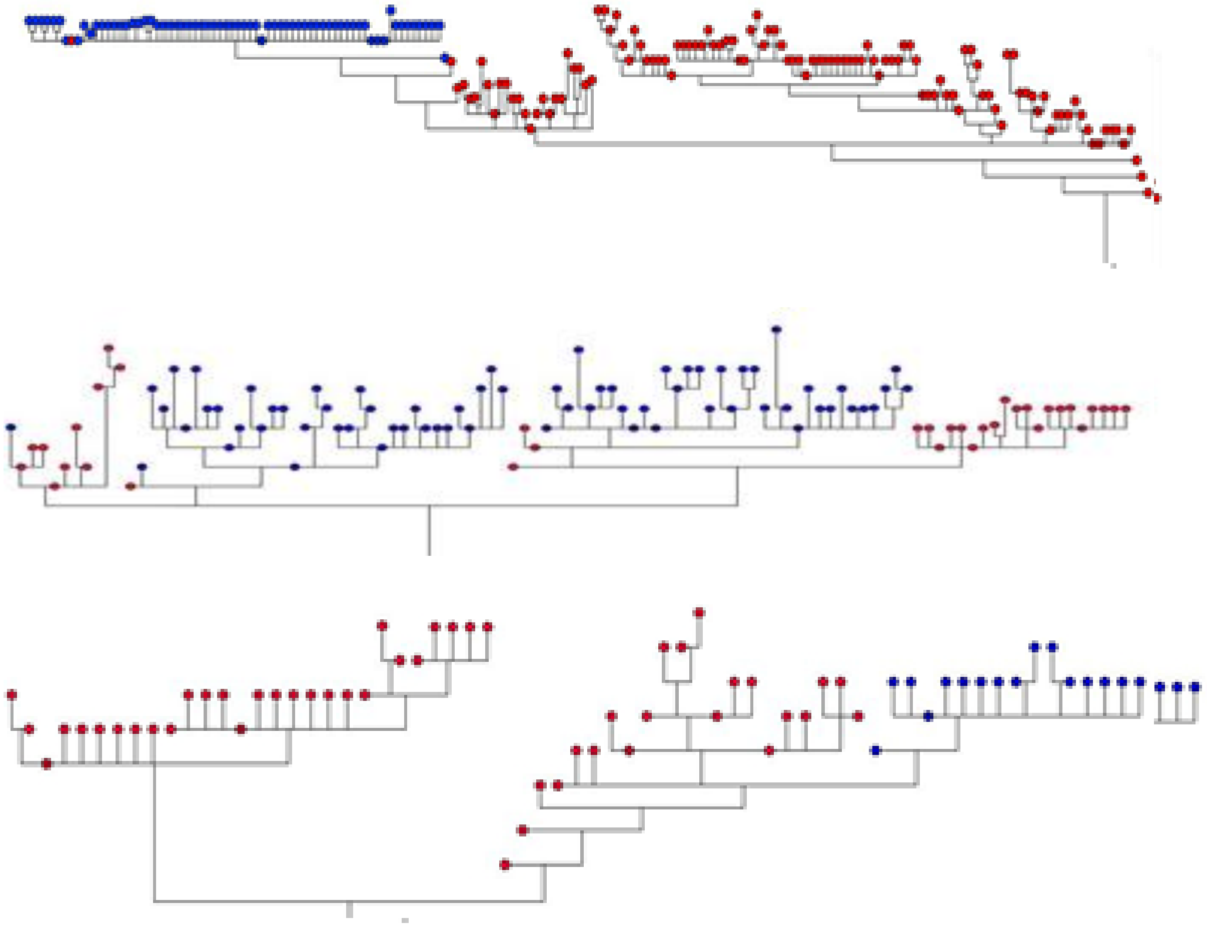
Uses of variant analysis - Quasispecies

- Three patients – MP1, 2 & 3
- **Who infected whom?**
- MP1 is independent from the cluster
- MP2 (red) → MP3 (blue)?
- Or *vice versa*?



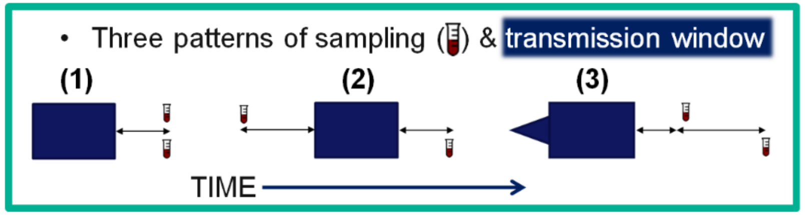
Romero-Severson EO *et al*. Genetics 2017 207(3):1089

Uses of variant analysis - Quasispecies

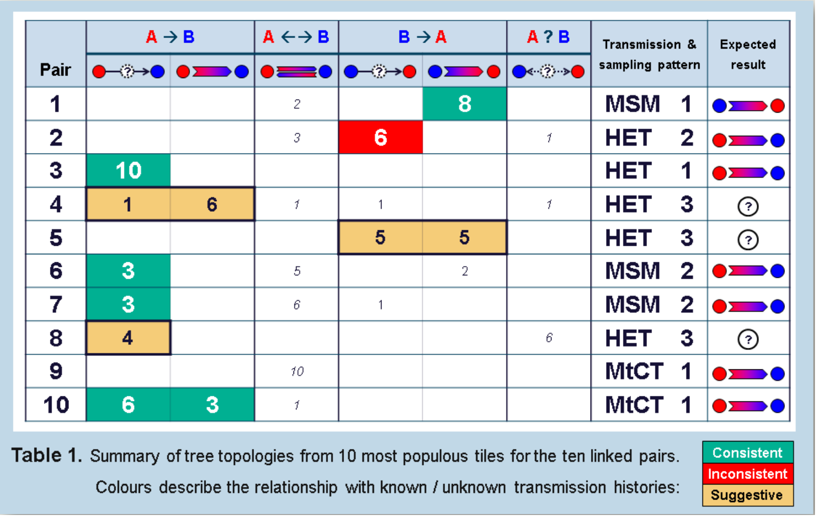


Transmission	
MSM	x3
HET	x5
MtCT	x2

Subtype	
B	x5
C	x2
G, 01, 02	x1 each



- Three patterns of sampling (🧪) & transmission window



Pair	A → B		A ←→ B	B → A		A ? B	Transmission & sampling pattern	Expected result
	●→?→●	●▶●	●═●	●→?→●	●▶●	●←?→●		
1			2		🟩 8		MSM 1	●▶●
2			3	🟥 6		1	HET 2	●▶●
3	🟩 10						HET 1	●▶●
4	🟨 1	🟨 6	1	1		1	HET 3	?
5				🟨 5	🟨 5		HET 3	?
6	🟩 3		5		2		MSM 2	●▶●
7	🟩 3		6	1			MSM 2	●▶●
8	🟨 4					6	HET 3	?
9			10				MtCT 1	●▶●
10	🟩 6	🟩 3	1				MtCT 1	●▶●

Table 1. Summary of tree topologies from 10 most populous tiles for the ten linked pairs. Colours describe the relationship with known / unknown transmission histories:

🟩 Consistent
🟥 Inconsistent
🟨 Suggestive

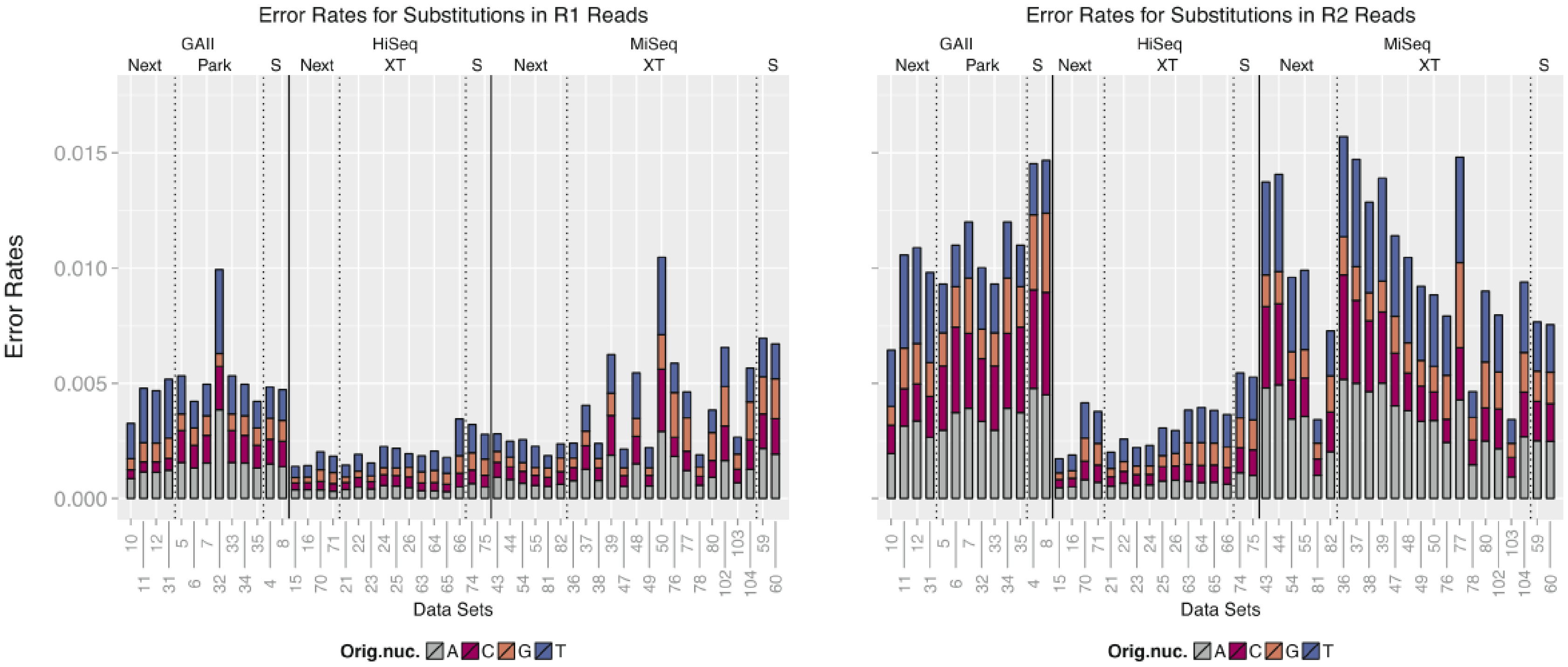
Bibby DF *et al*. HIV Dynamics & Evolution 2017

Technical pitfalls

Frequency of minor variant detection limited by experimental noise
Many sources of error:

1. Sequencing
2. Amplicon-based sequencing
3. Nucleotide content
4. Hexamer priming
5. Product degradation
6. Contamination
7. Bioinformatics

Technical pitfalls – Sequencing



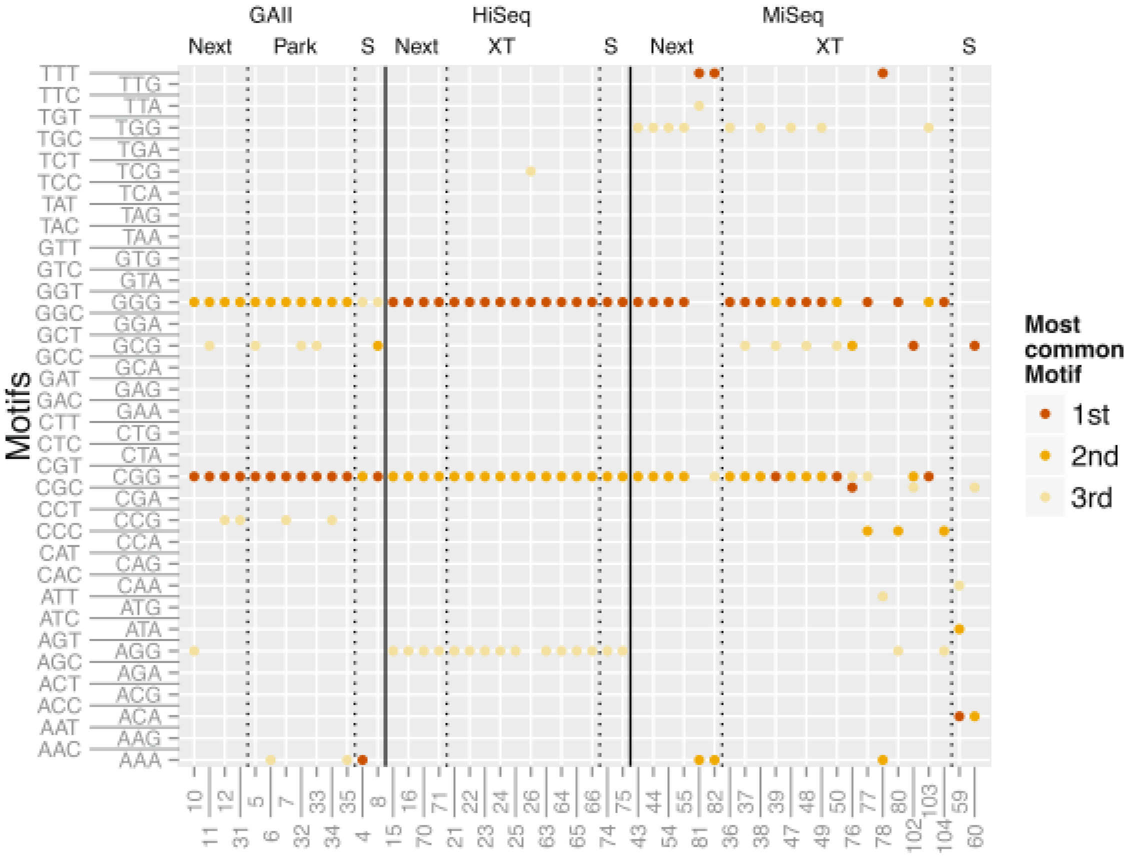
Schirmer M *et al*. BMC Bioinformatics 17(1):125

Technical pitfalls – Sequencing

<u>Substitutions</u>

<u>xGG</u> motif

GGG = **CGG** > AGG > TGG



Schirmer M *et al*. BMC Bioinformatics 17(1):125

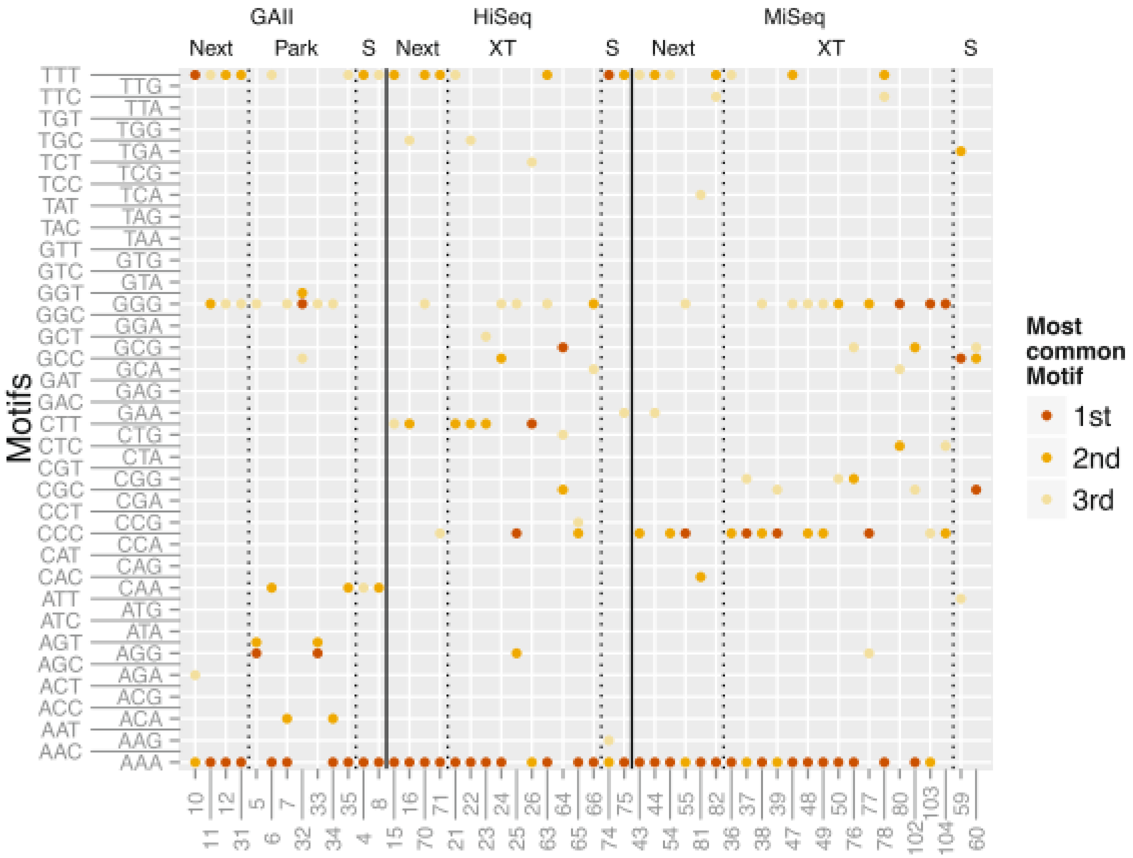
Technical pitfalls – Sequencing

Indels

Homopolymeric tracts

AAA > CCC = GGG = TTT

High Quality Scores



Schirmer M *et al*. BMC Bioinformatics 17(1):125

Technical pitfalls – Amplicons

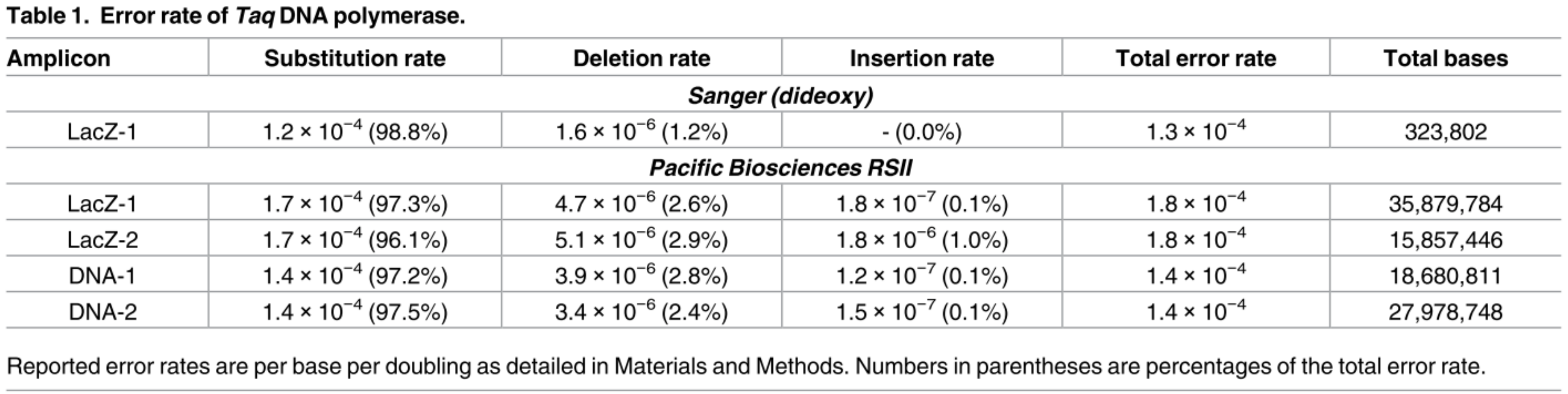


Table 1. Error rate of *Taq* DNA polymerase.

Amplicon	Substitution rate	Deletion rate	Insertion rate	Total error rate	Total bases
			Sanger (dideoxy)		
LacZ-1	1.2×10^{-4} (98.8%)	1.6×10^{-6} (1.2%)	- (0.0%)	1.3×10^{-4}	323,802
			Pacific Biosciences RSII		
LacZ-1	1.7×10^{-4} (97.3%)	4.7×10^{-6} (2.6%)	1.8×10^{-7} (0.1%)	1.8×10^{-4}	35,879,784
LacZ-2	1.7×10^{-4} (96.1%)	5.1×10^{-6} (2.9%)	1.8×10^{-6} (1.0%)	1.8×10^{-4}	15,857,446
DNA-1	1.4×10^{-4} (97.2%)	3.9×10^{-6} (2.8%)	1.2×10^{-7} (0.1%)	1.4×10^{-4}	18,680,811
DNA-2	1.4×10^{-4} (97.5%)	3.4×10^{-6} (2.4%)	1.5×10^{-7} (0.1%)	1.4×10^{-4}	27,978,748

Reported error rates are per base per doubling as detailed in Materials and Methods. Numbers in parentheses are percentages of the total error rate.

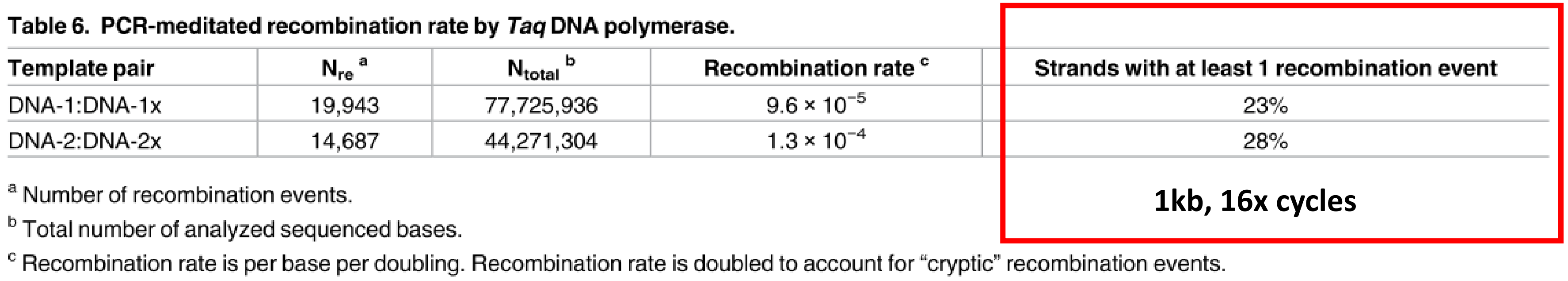


Table 6. PCR-meditated recombination rate by *Taq* DNA polymerase.

Template pair	N_{re} [a]	N_{total} [b]	Recombination rate [c]	Strands with at least 1 recombination event
DNA-1:DNA-1x	19,943	77,725,936	9.6×10^{-5}	23%
DNA-2:DNA-2x	14,687	44,271,304	1.3×10^{-4}	28%

[a] Number of recombination events.

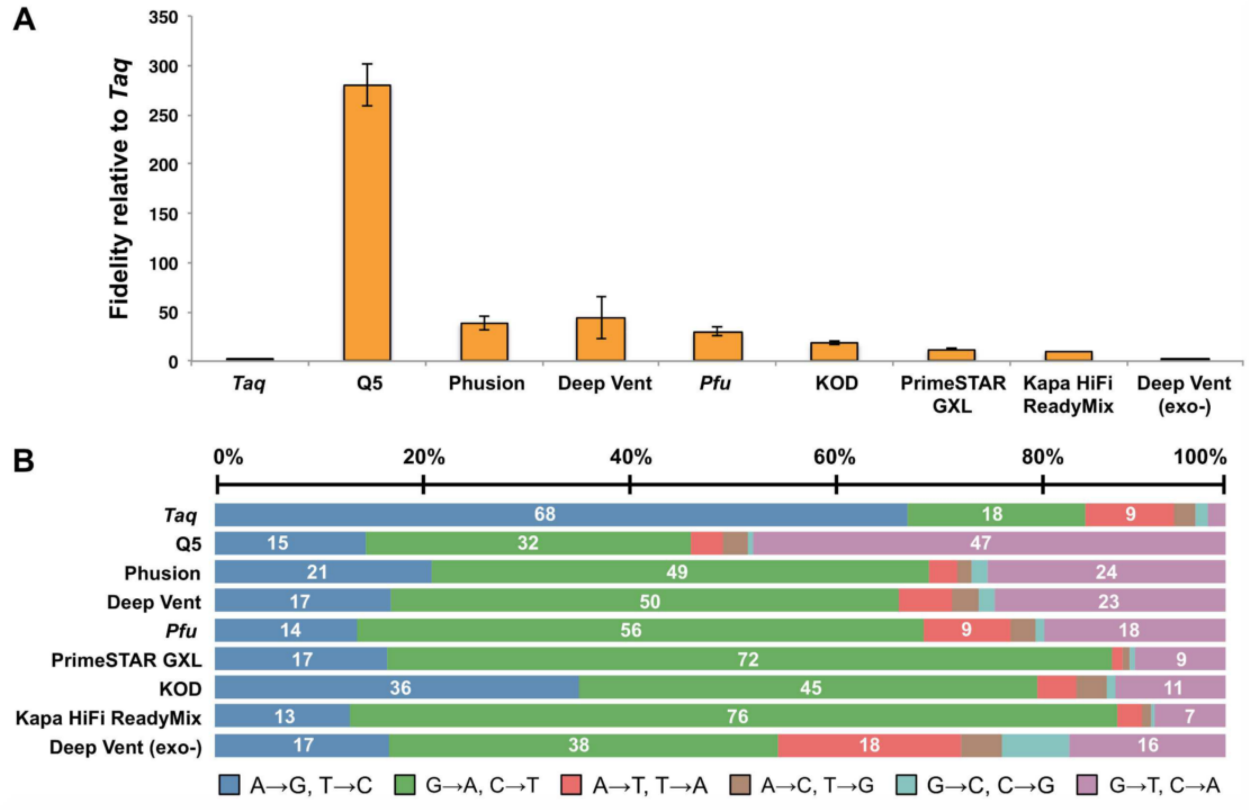
[b] Total number of analyzed sequenced bases.

[c] Recombination rate is per base per doubling. Recombination rate is doubled to account for "cryptic" recombination events.

1kb, 16x cycles

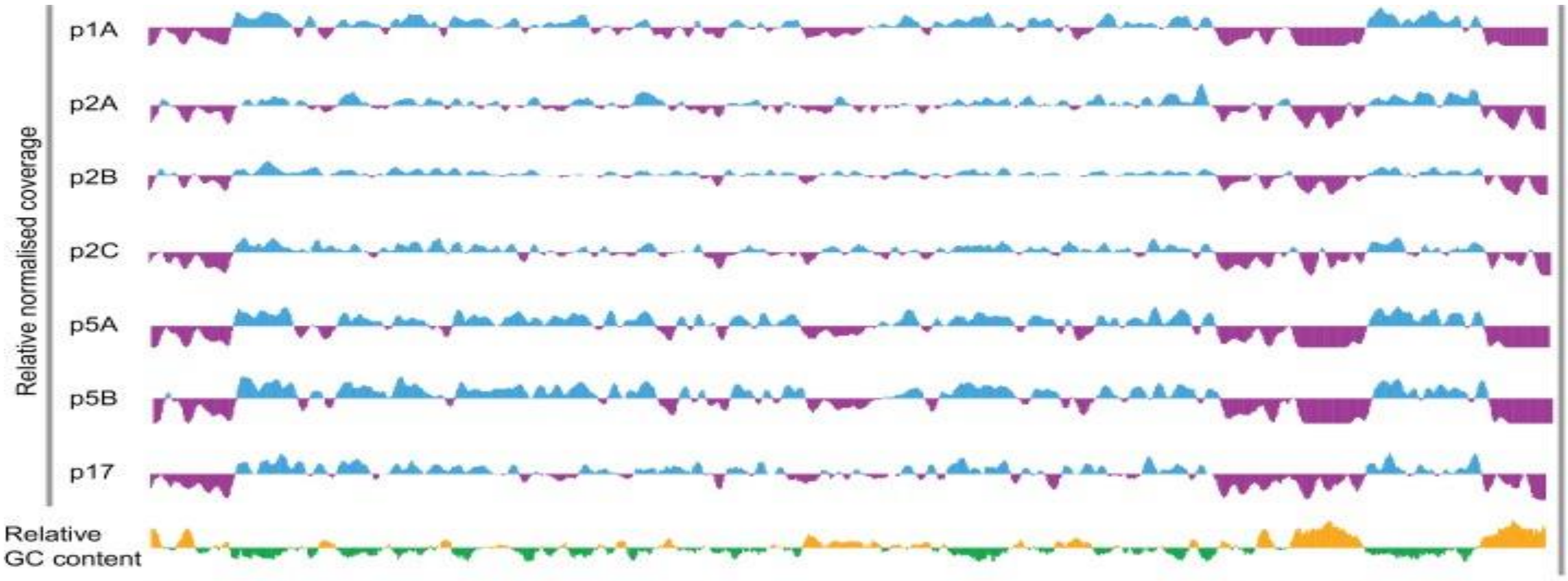
Potapov V *et al*. PLoS One 2017 12(1): e0169774

Technical pitfalls – Amplicons



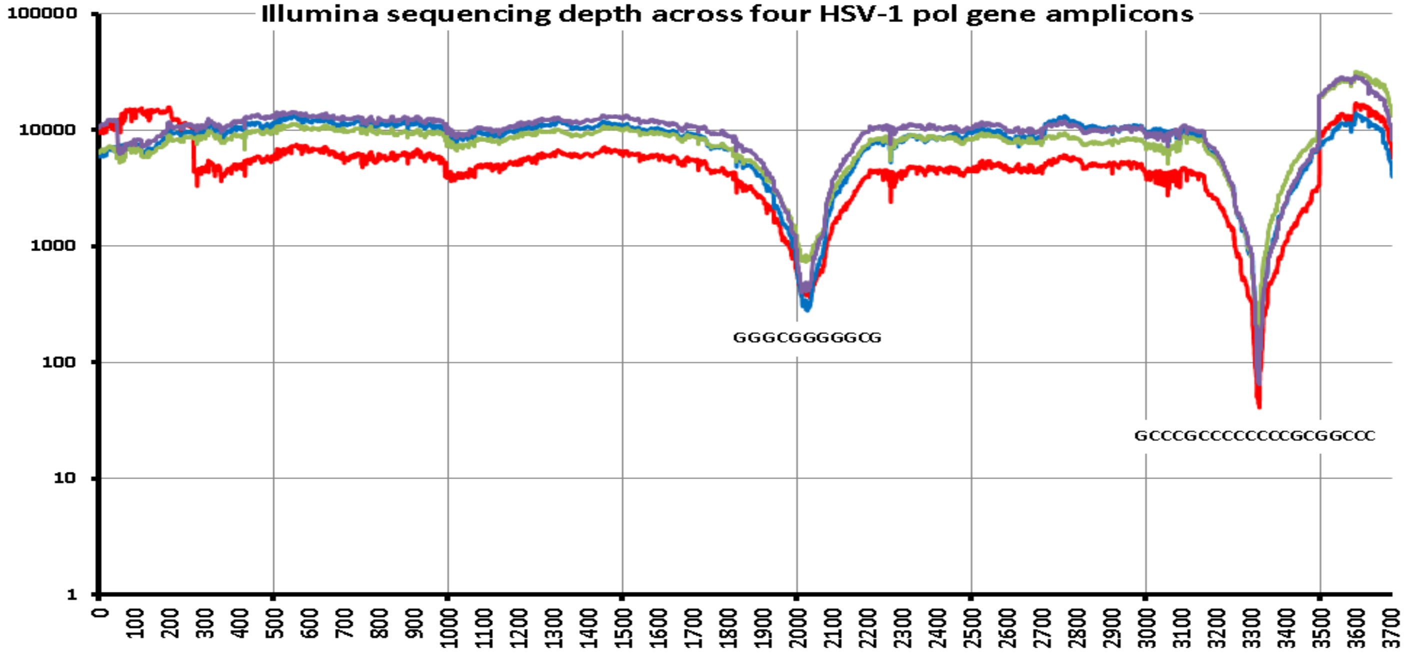
Potapov V *et al*. PLoS One 2017 12(1): e0169774

Technical pitfalls – Nucleotide content



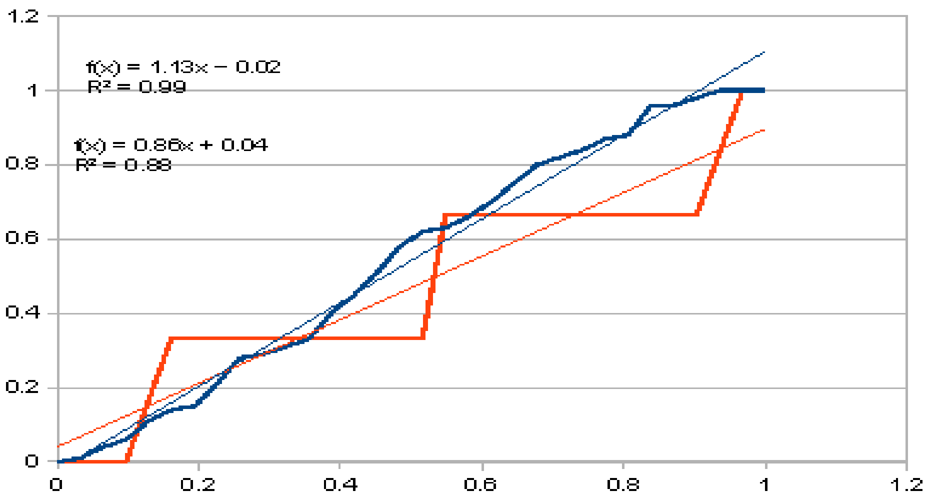
Karamitros T *et al*. PLoS One 2016 11(6):e0157600

Technical pitfalls – Nucleotide content

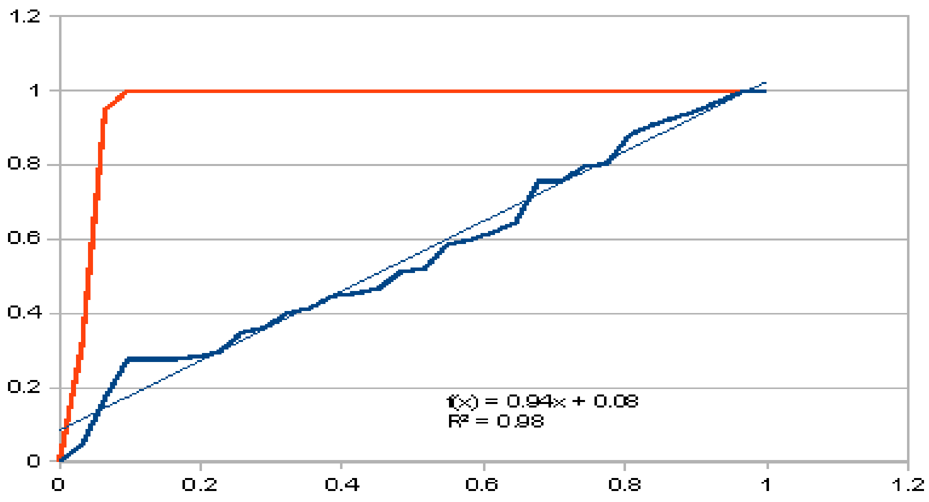


UKHSA (unpublished)

Technical pitfalls – Hexamer priming



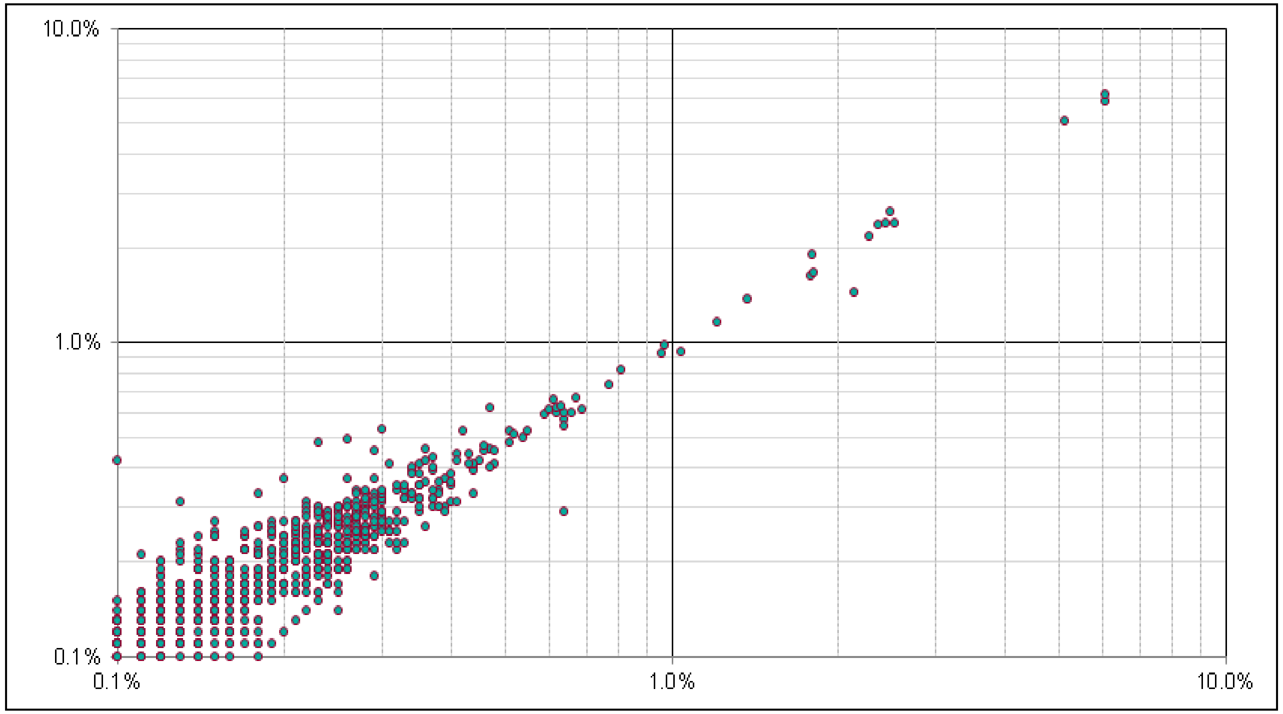
The position of the minority variant nucleotide (red line) is evenly distributed along the read lengths (as is the majority variant in blue)



100% of the minority variants are within 7% (10nt) of a read terminus – artefact from insert-priming

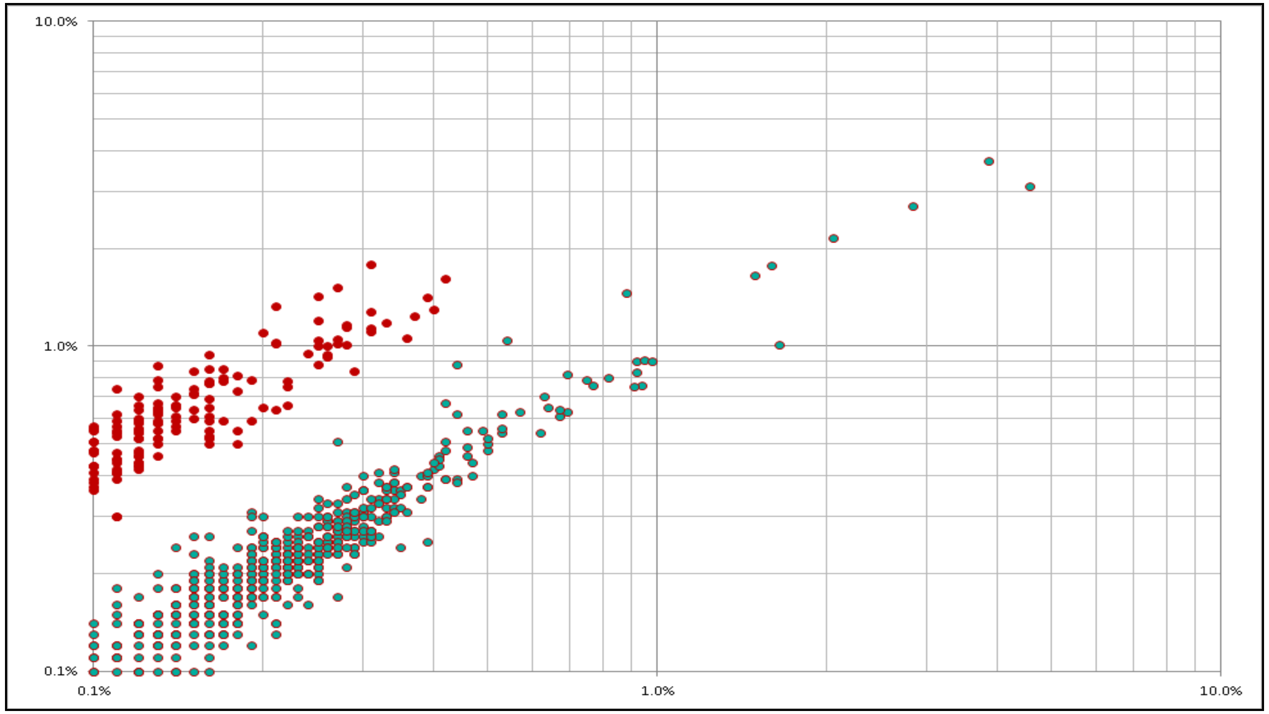
UKHSA (unpublished)

Technical pitfalls – Product degradation



UKHSA (unpublished)

Technical pitfalls – Product degradation



UKHSA (unpublished)

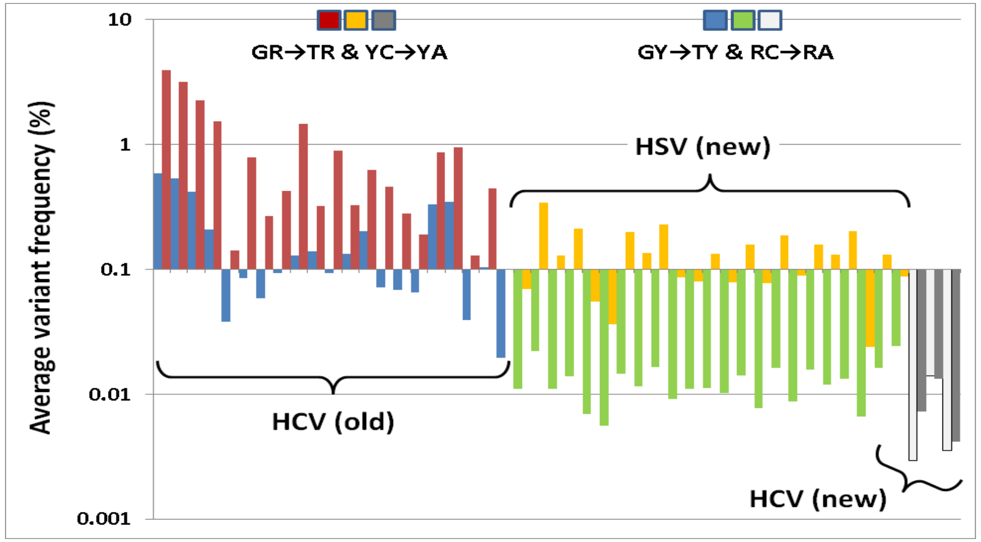
Technical pitfalls – Product degradation

- Much investigation revealed context-specific conversion of dinucleotides

  YC → YA
  GR → TR

  previously only seen in sonicated fragments – Costello *et al*. NAR 2013

- The frequency of converted bases is proportional to the time spent at 4ºC



UKHSA (unpublished)

Technical pitfalls – Contamination

“Sequences not belonging to that sample present in the FASTQ set”

- Index-switching
  - Occurs during library prep / sequencing

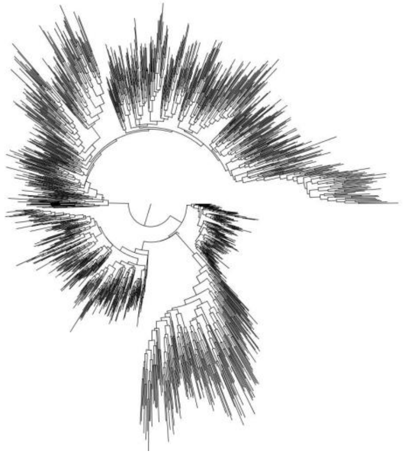
- Laboratory contamination
  - Similar to PCR
  - Spatial and temporal separation of work areas
  - Rotation of adapters
  - Robotics
  - Rotation of control positions
  - Alternation of template types

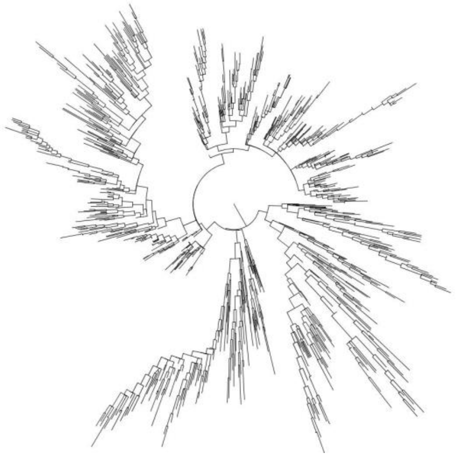
Technical pitfalls – Bioinformatics

- Synthetic FASTQ datasets
  - HIV quasi-species generation

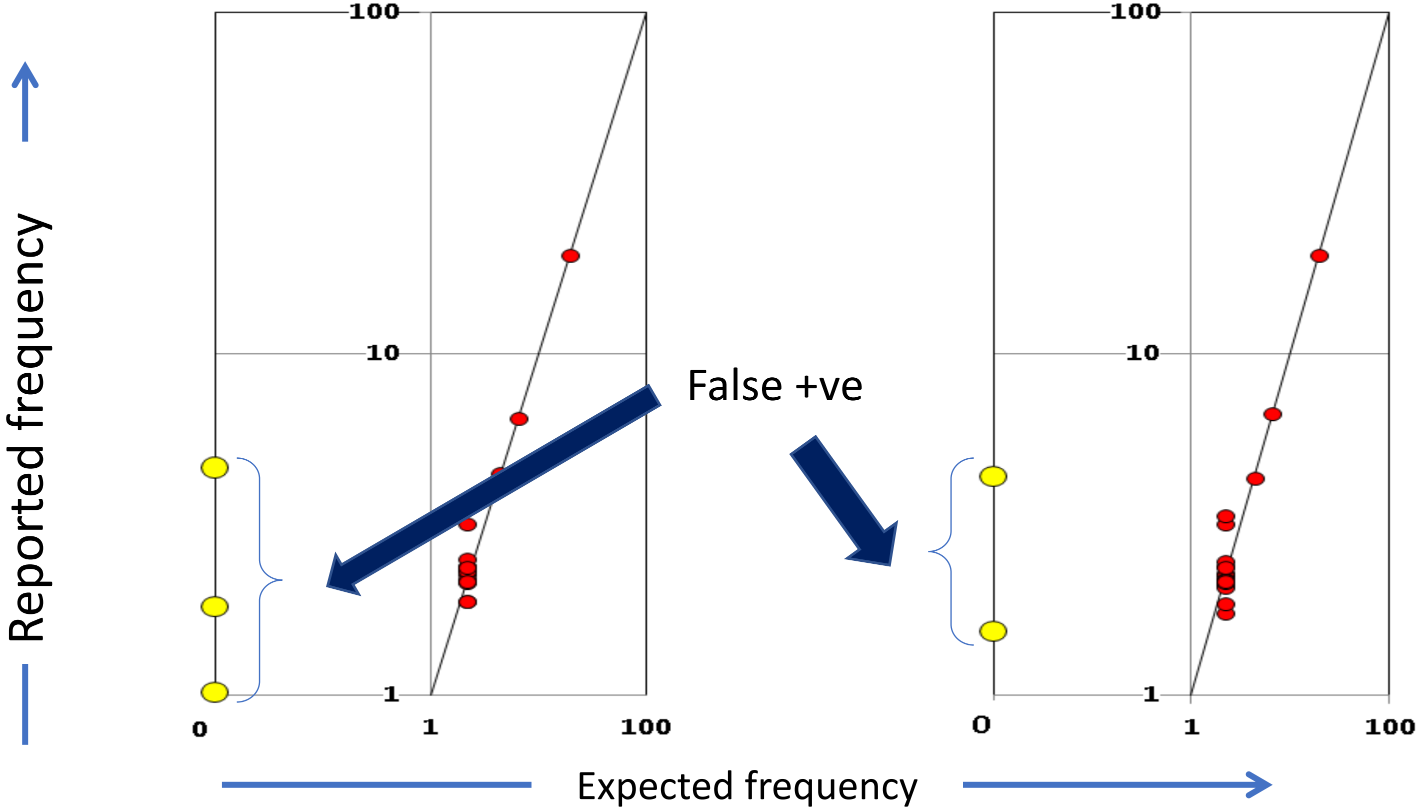
    Adapted from Pandit A & de Boer R, Retrovirology 2014 11(1):56
  - FASTQ generation using empirical error profiles / quality scores

    ART - Huang W, *et al*. Bioinformatics 2012 28(4):593-4
- Two pre-production pipelines tested





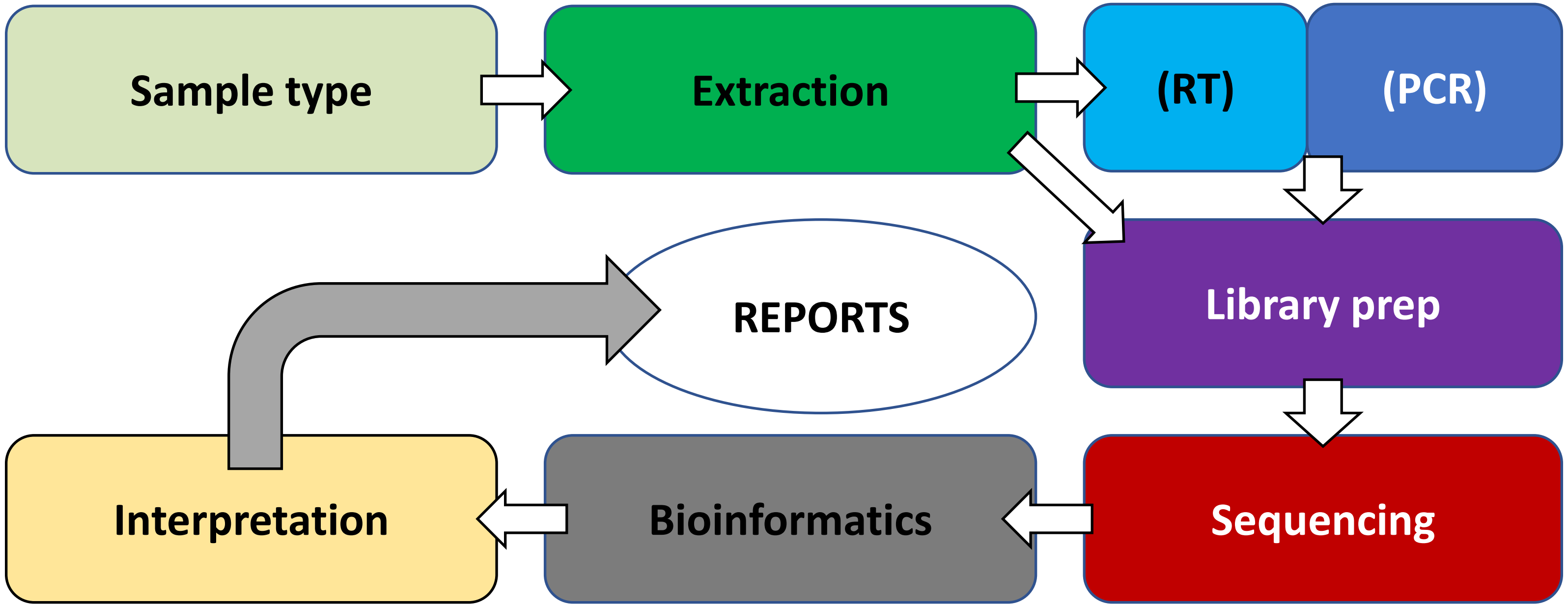
Technical pitfalls – Bioinformatics



False +ve

UKHSA (unpublished)

Validation


Sample type
Extraction
(RT)
(PCR)
Library prep
REPORTS
Sequencing
Bioinformatics
Interpretation

Validation – How low can you go?

- For HIV & HCV, Sanger at 15-20% == resistant
  - What does 10% mean?
  - 5%?
- Reproducibility, repeatability, accuracy and precision is critical
  - Validating against Sanger is relatively straightforward, both clinically and technically
  - Validating lower frequencies is quite the opposite
- Clinical utility of lower frequency variants unproven

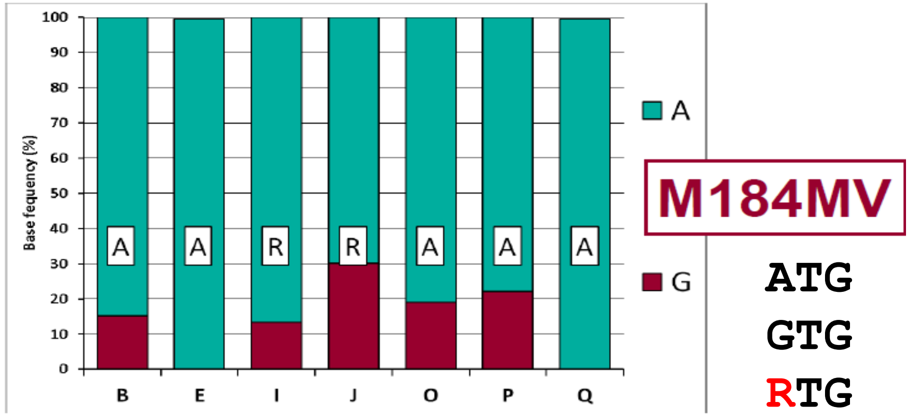
Cut-offs vary considerably between assays

Validation – Copy number & variant frequency

How reliable is a variant frequency call?

- When the depth of coverage (i.e. reads covering that position) is high/low?
- What levels constitute ‘high’ and ‘low’ depths for PCR, sequence capture and metagenomic approaches?

Sample: A Domain: PR/RT Position: 847 Consensus: R



M184MV

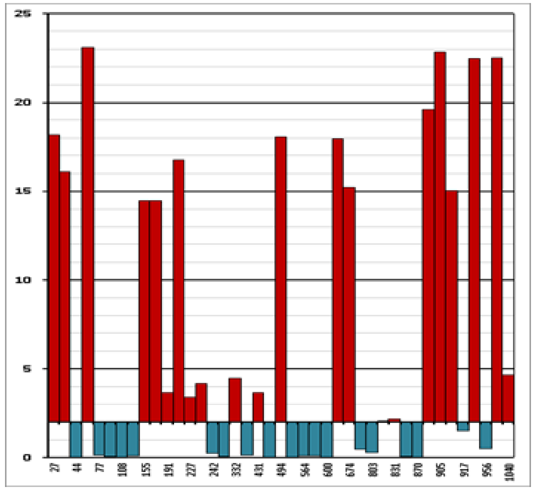
ATG
GTG
RTG

UKHSA (unpublished)

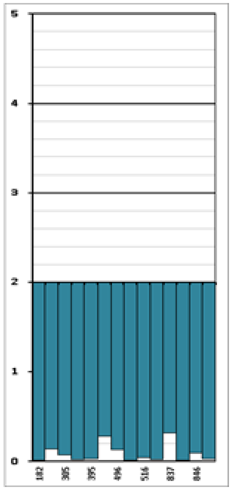
Validation – Copy number & variant frequency

How reliable is a variant frequency call?

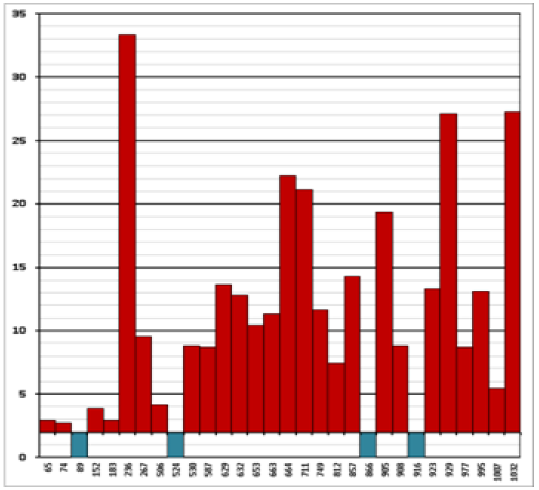
- When the depth of coverage (i.e. reads covering that position) is high/low?
- What levels constitute ‘high’ and ‘low’ depths for PCR, sequence capture and metagenomic approaches?



Sample 1



Sample 2



Sample 3

UKHSA (unpublished)

Validation – Copy number & variant frequency

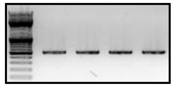
How reliable is a variant frequency call?

- PCR produces large amounts of material
- How representative of the starting population is the amplicon mix?
  - Depth is not correlated with reliability!
- If the starting virus copy number is 100,000 copies per library, 10% = 10,000
  - But at 1,000 copies, how reliable is 5% (50 viruses)?
  - Reverse transcription (to generate cDNA) is notoriously inefficient and error-prone – how many viruses are represented?
  - There are multiple PCR cycles in the library prep too...
- Often, the amount of starting material / viral load is unknown

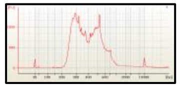
Validation – Standardised materials

Sample – haplotypes mixed at precise frequencies (e.g. 1, 2, 5, 10 & 20%)

DNA / RNA extract



PCR product



Library – storage issues

FASTQ datasets - artificial quasispecies and synthetic FASTQs

```
5'ATGACGTGGGGA3'
3'TACTGCACCCCT5'
```

Consensus sequences – to test interpretation mechanisms

Summary

Consensus & variant calling

- Many tools for mapping (BWA, Bowtie, smalt, Tanoti)
- Several tools for variant calling (QuasiBAM, V-Phaser)

Choice of reference & user-defined parameters is CRITICAL

Using the consensus

- Submit to usual tools
- Different mixed-base thresholds to incorporate minority variants

BEWARE – Interpretations may not be validated on NGS-derived data...

- Phyletic analysis is coming

Summary

Technical pitfalls

- Experimental approach influences the result in unpredictable ways
  - Low-frequency variation can arise through diverse processes
- Reproducibility experiments are essential
  - Across a range of conditions and samples

Validation

ESSENTIAL – especially when detecting ‘new’ data, e.g. <20% minority

- All components of the assay need independent investigation
- Look at all the data
- ESPECIALLY THE BIOINFORMATICS!