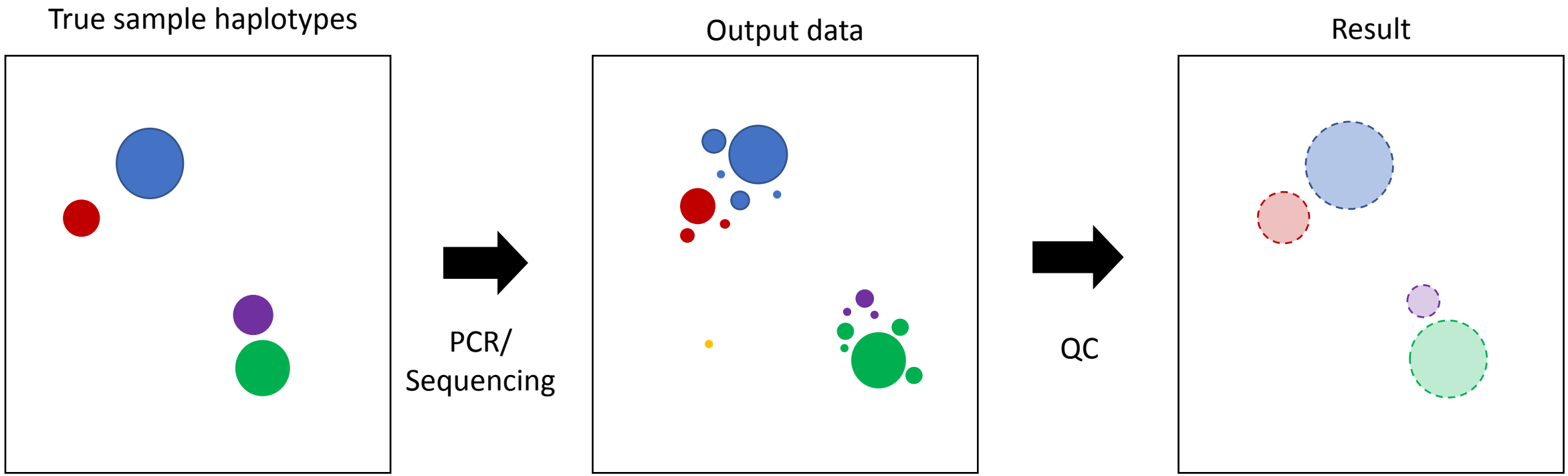
Amplicon sequencing pipeline validation

2017-2018

Amplicon sequencing analysis



Task: How to eliminate technical variation without compromising biological variation?

Amplicon seq error correction tools

- PASEC (Early *et al*, **Malaria J** 2019 ; Neafsey *et al*, **NEJM** 2015)
  - Clusters based on distance and coverage
  - Manually mask difficult-to-sequence regions (homopolymers, etc)

- SeekDeep (Hathaway *et al,* **Bioinformatics** 2017)
  - Iteratively clusters based on weighted-distance, where weight is a function of difference type (mismatch or indel) and base quality
  - Derives power from duplicate PCRs

- DADA2 (Callahan *et al,* **Nature Methods** 2016)
  - Clusters based on error model prediction

$$p_A(i \to j) = \frac{1}{1 - p_{pois}(c_i \lambda_{ij}, 0)} \sum_{c' = c_j}^{\infty} p_{pois}(c_i \lambda_{ij}, c')$$

PASEC pipeline for N samples over M amplicons

{wdl}

sampleN .demultiplexed.bam

Merge paired end reads using FLASH

sample1 .merged.fastq ... sampleN .merged.fastq

Align merged reads using BWA mem

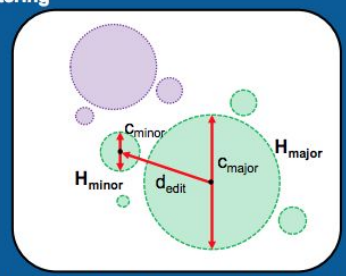
sampleN.aln.bam

Extract haplotypes for each amplicon for each sample (in parallel)

Extraction
- dereplicate reads within amplicon coordinates
- trim to amplicon coordinates (optional)
- mask specified regions (optional)

Filtering
- quality < min mapping quality
- length < min amplicon length
- frequency < min haplotype frequency

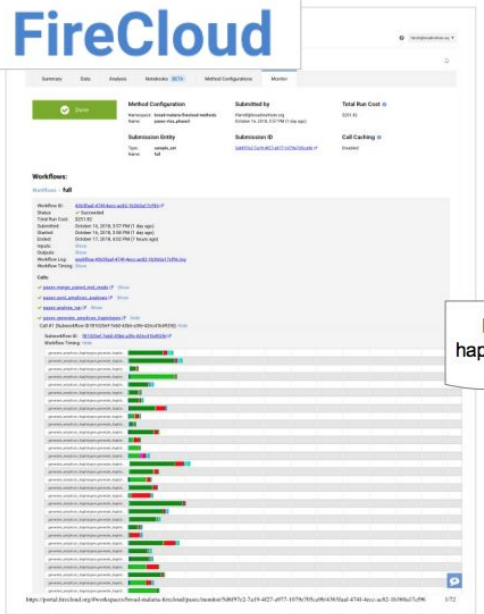
Clustering



From highest to lowest coverage haplotype, cluster H_{minor} to H_{major} if:
- d_{edit} < max cluster edit distance
- (c_{minor} / c_{major}) < max cluster coverage ratio

sample_haplotypes .amplicon1.raw.txt ... sample_haplotypes .ampliconM.raw.txt

FireCloud



known_ haplotypes.db

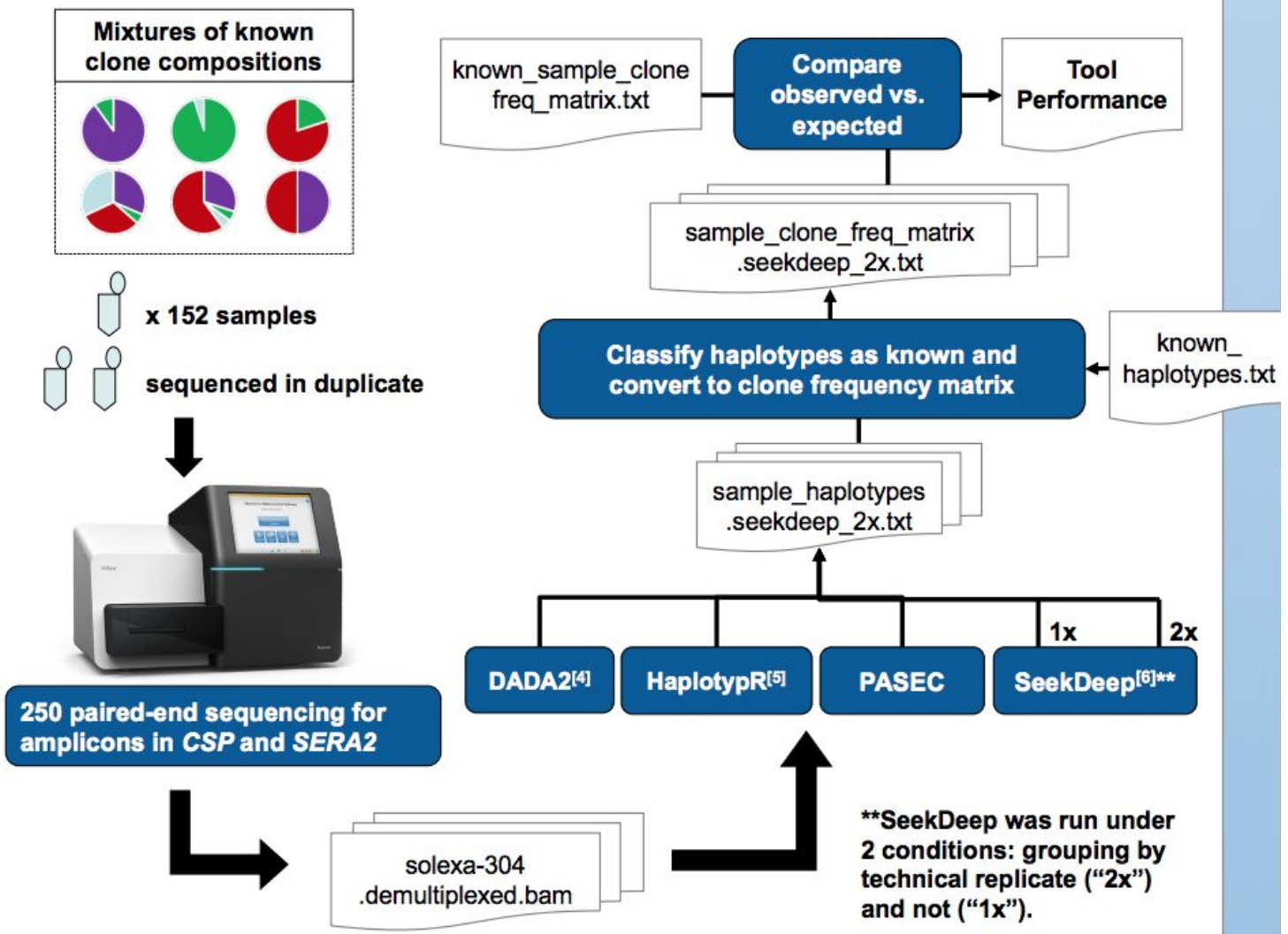
Index amplicon haplotype sets and classify against database

Index sample haplotypes lists and pool results from all amplicons

sample_haplotypes .indexed.txt

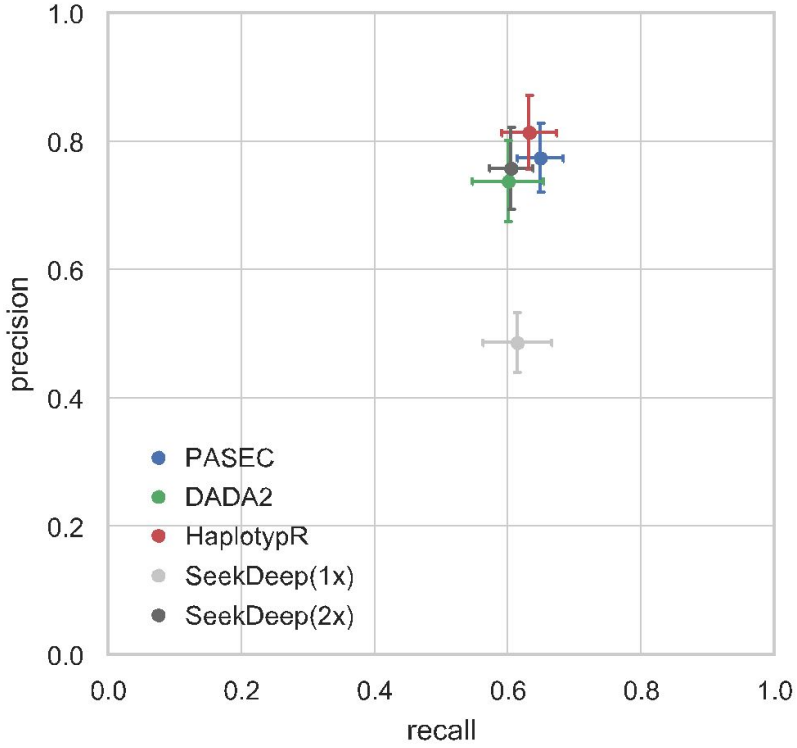
haplotype_set.index.txt

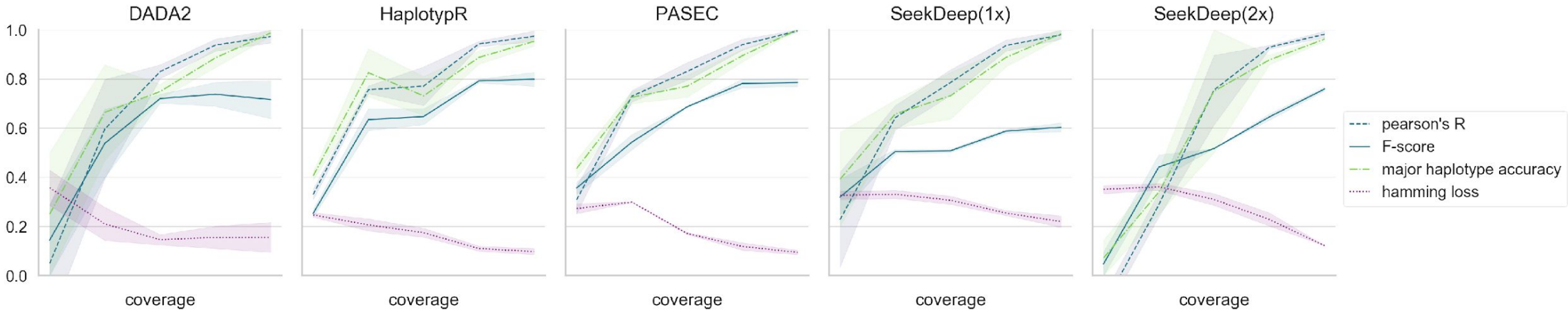
Methods


Mixtures of known clone compositions
x 152 samples
sequenced in duplicate
250 paired-end sequencing for amplicons in *CSP* and *SERA2*
solexa-304 .demultiplexed.bam
known_sample_clone freq_matrix.txt
Compare observed vs. expected
Tool Performance
sample_clone_freq_matrix .seekdeep_2x.txt
Classify haplotypes as known and convert to clone frequency matrix
known_ haplotypes.txt
sample_haplotypes .seekdeep_2x.txt
DADA2[4]
HaplotypR[5]
PASEC
1x
2x
SeekDeep[6]**
**SeekDeep was run under 2 conditions: grouping by technical replicate ("2x") and not ("1x").

Tool performance comparison

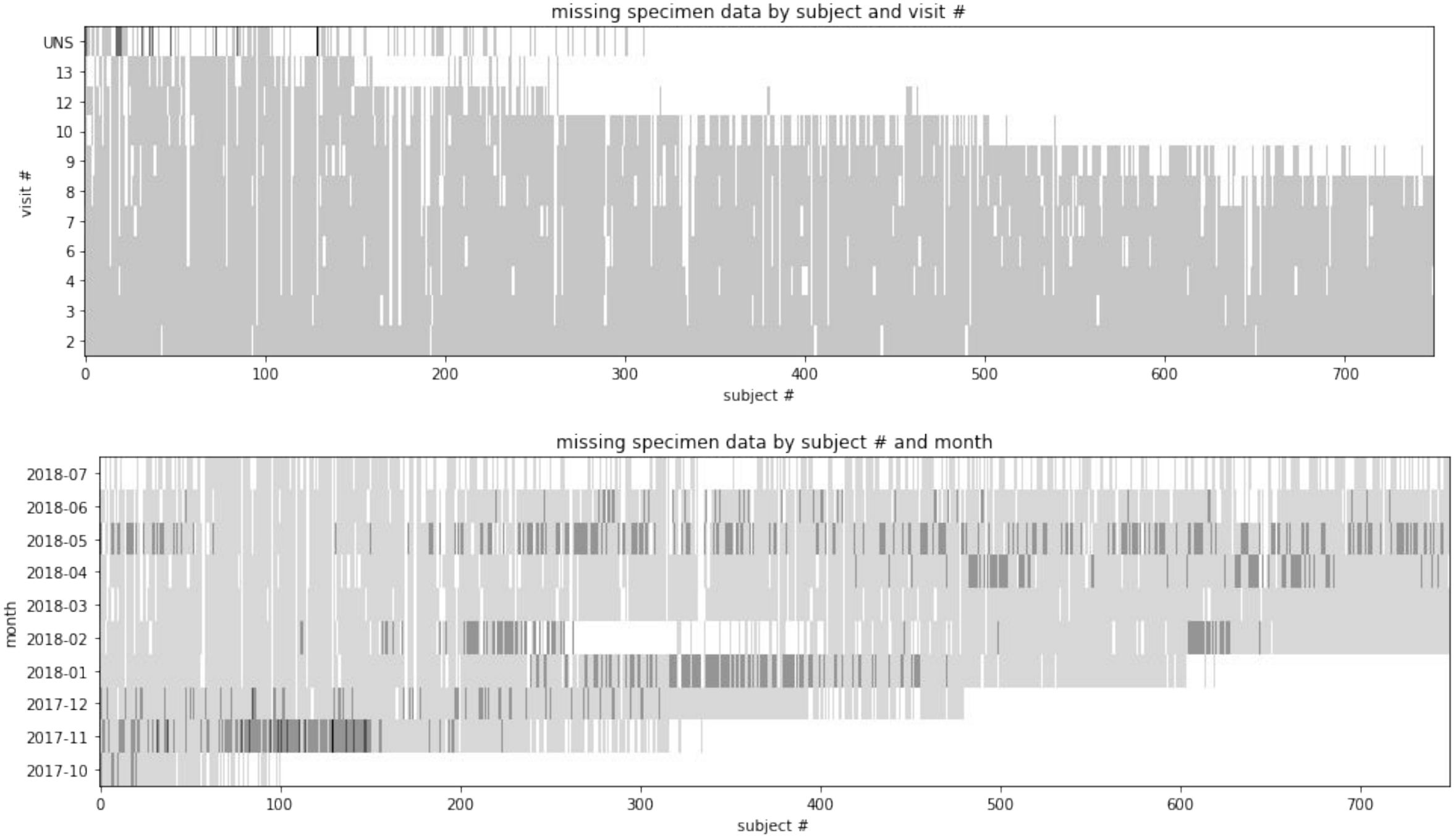
https://www.biorxiv.org/content/early/2018/10/25/453472.

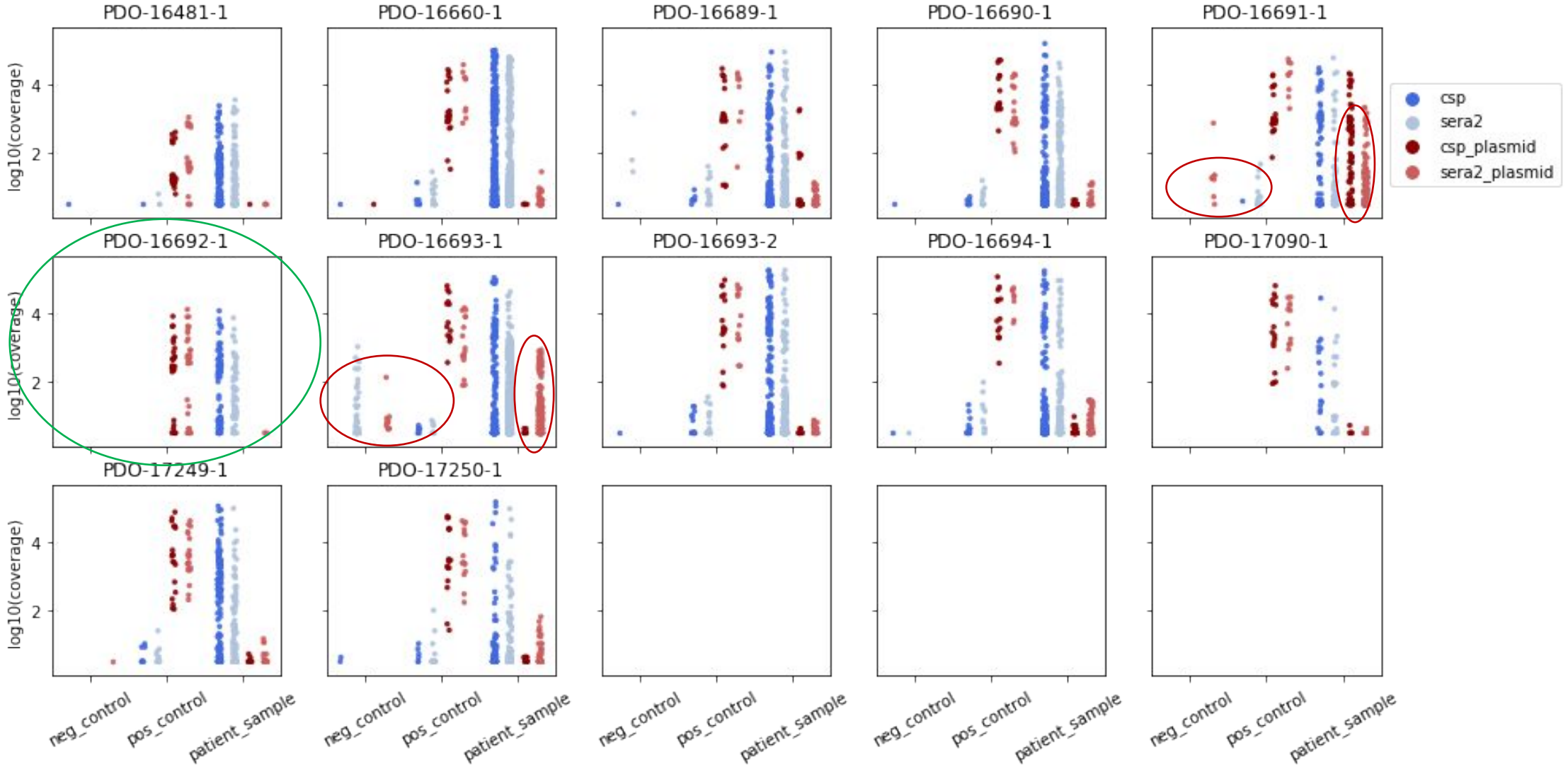




Phase IV anti-malarial vaccine clinical trial project

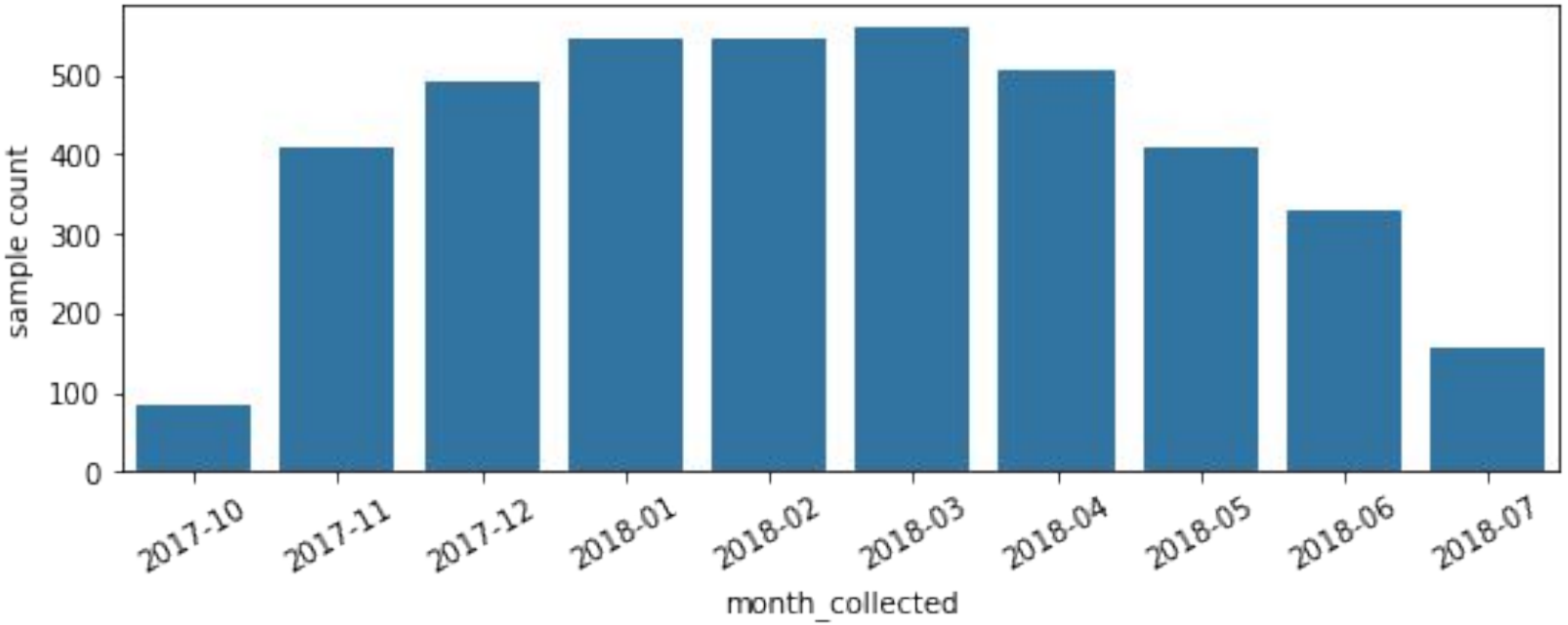
2018-2019

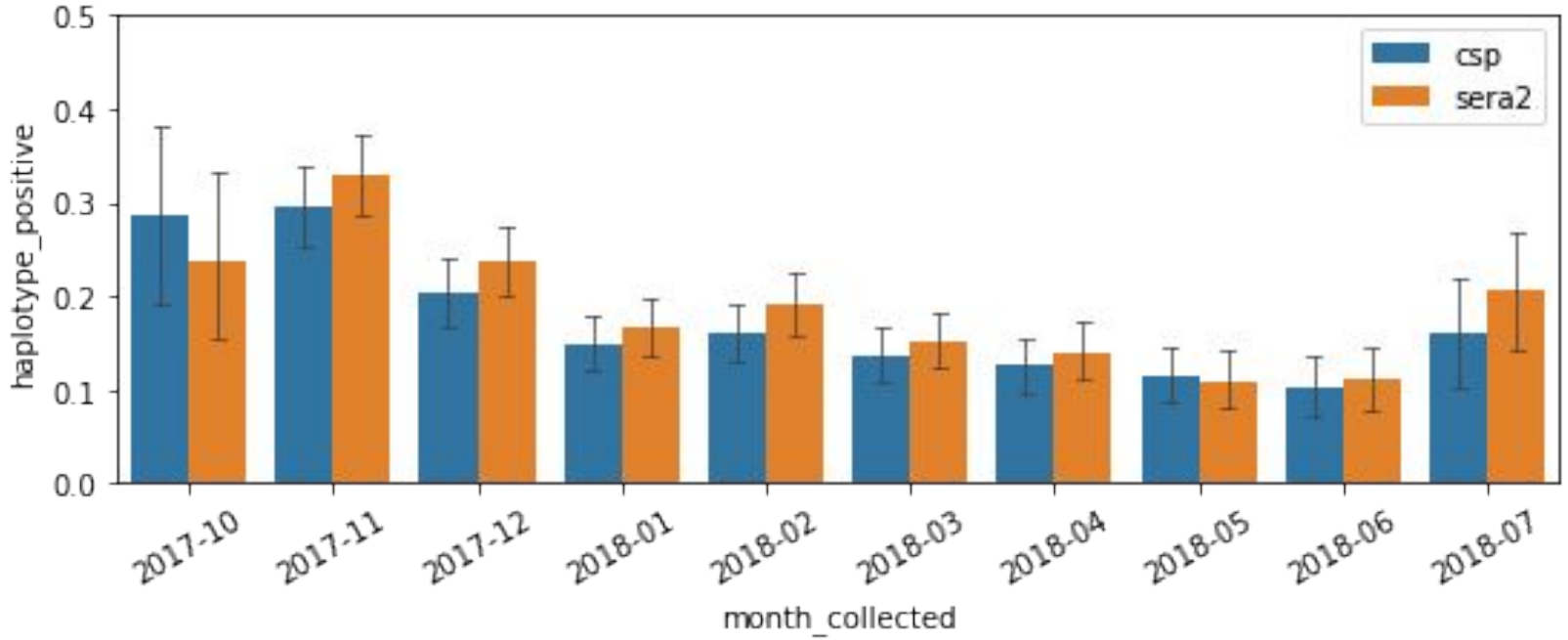

missing specimen data by subject and visit #
visit #
UNS
13
12
10
9
8
7
6
4
3
2
0
100
200
300
400
500
600
700
subject #
missing specimen data by subject # and month
month
2018-07
2018-06
2018-05
2018-04
2018-03
2018-02
2018-01
2017-12
2017-11
2017-10
0
100
200
300
400
500
600
700
subject #



Amplicon (blue) and positive control plasmid (red) log10(coverage) by PDO (each box, with PDO-version as title) and sample type (x-axis). Red circles indicate contamination (amplicons in controls or plasmids in negative controls or patient samples). The green circle indicates one of the cleanest (but lowest coverage) runs. One take away, when we see cross-contamination of the plasmids, there is usually also amplicon cross-contamination. Also of note, PDO-16693-2 (reworked PDO-16693) is a good quality PDO.

Seasonality?





How to define new infections?