29

Evaluation of amplicon sequencing tools on mock complex *P. falciparum* infections

Timothy M. Farrell[1,2], Angela M. Early[1,2], Rachel F. Daniels[1,2], Sarah K. Volkman[1,2], Dyann F. Wirth[1,2], Bronwyn L. MacInnis[1,2], Daniel E. Neafsey[1,2]

1. Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, 02142 USA
2. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, 02115 USA

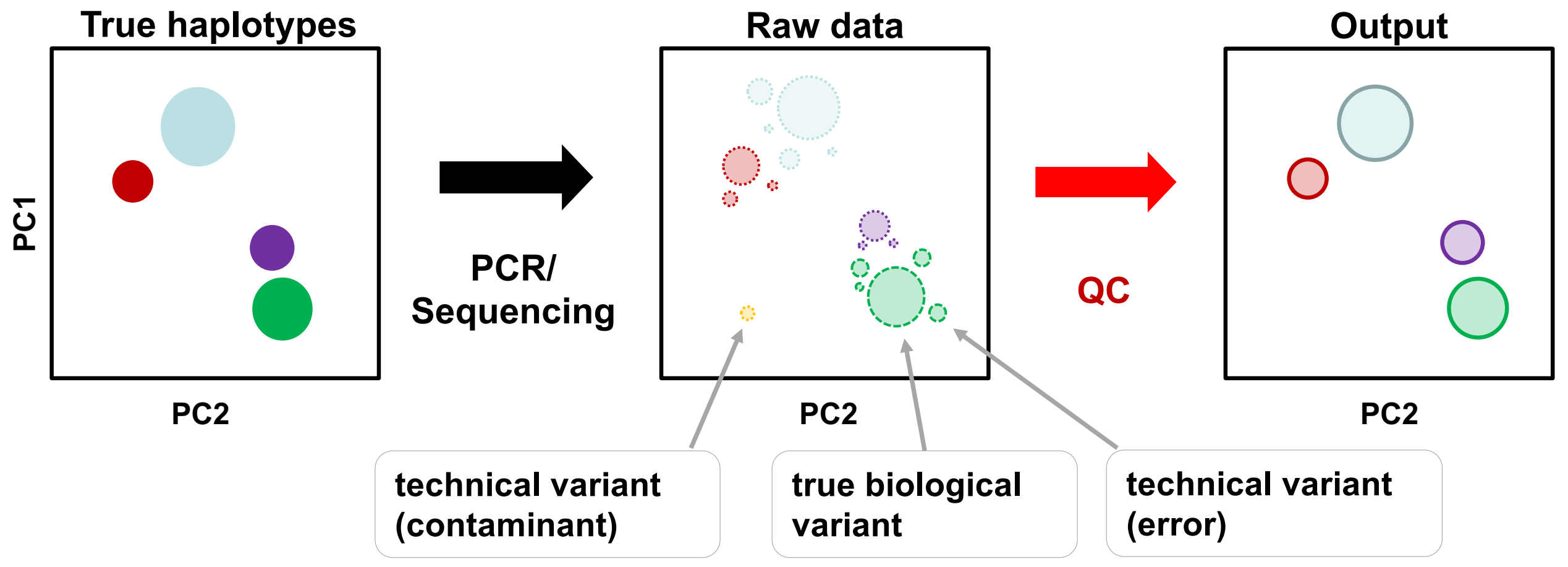
Introduction

Accurate genotyping of polyclonal malaria infections is important for assessing population-level effects of interventions over geography and time.[1]

Amplicon sequencing is a cost-effective, high-throughput genotyping technology, but its data can be subject to technical artifacts, especially when applied to difficult-to-sequence genomes like *P. falciparum*.[2,3]

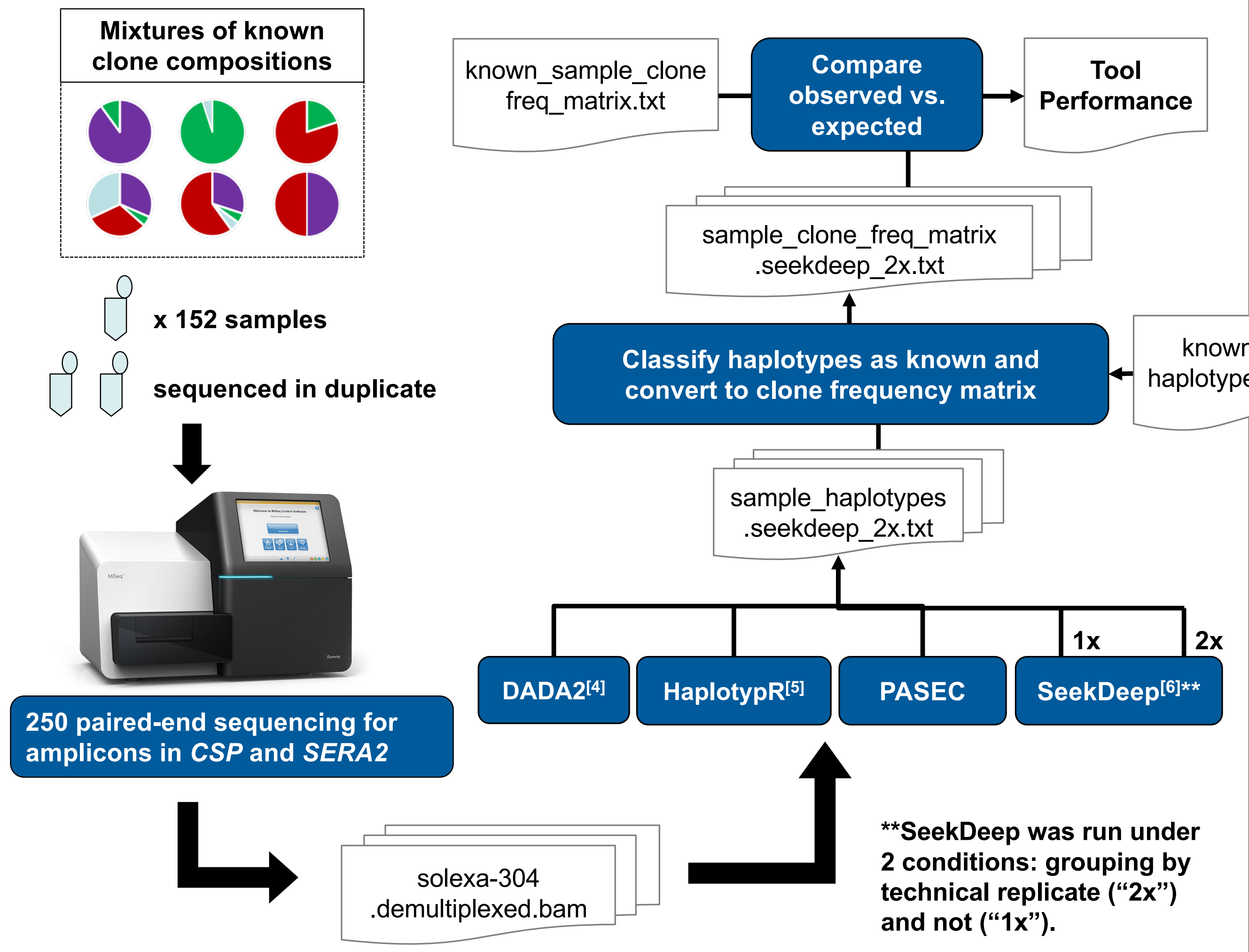
Here, we benchmarked a set of state-of-the-art amplicon sequencing analysis tools on a controlled dataset with known clone compositions at low parasite density (1-200 genomes/ul) to quantify their performance at resolving **exact** haplotypes and their frequencies of two amplicons in *P. falciparum* genes *CSP* and *SERA2*.

How to eliminate technical variation without compromising biological variation?



Graphic 1. *PCA plots of a single sample's haplotype sequences, where each point is sized by coverage, at each step in an amplicon sequencing workflow.*

Methods

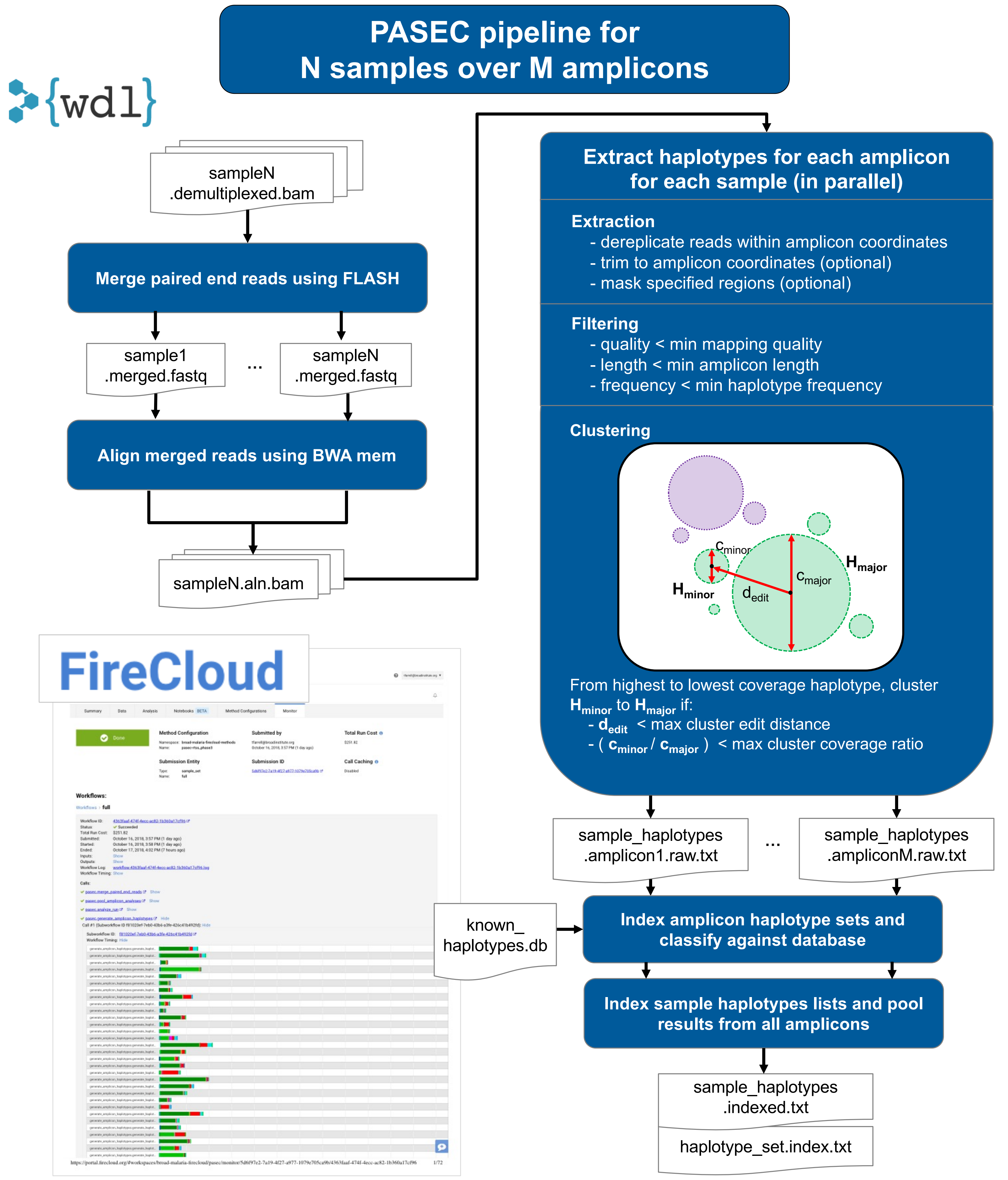


**SeekDeep was run under 2 conditions: grouping by technical replicate ("2x") and not ("1x").

PASEC

PASEC (Parallel Amplicon Sequencing Error Correction) is a highly-scalable amplicon sequencing analysis pipeline with an intuitive error correction clustering algorithm and deployed for use on the FireCloud platform (https://software.broadinstitute.org/firecloud/).

PASEC can be executed through the FireCloud web portal at: https://portal.firecloud.org/#workspaces/broad-malaria-firecloud/pasec.


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
FireCloud
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
c_{minor}
H_{major}
c_{major}
H_{minor}
d_{edit}
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
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

References

1. Koepfli C, Mueller I. Malaria epidemiology at the clone level. *Trends in Parasitology*. 2017;33:974–85.
2. Schirmer M, *et al*. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*. 2016;17:125.
3. Hamilton WL, *et al*. Extreme mutation bias and high AT content in P falciparum. *Nucleic Acid Res*. 2017;45(4):1889–1901.
4. Callahan BJ*, et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13:581–3. doi:10.1038/nmeth.3869.
5. Lerch A, *et al*. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. *BMC Genomics*. 2017;18:864. doi:10.1186/s12864-017-4260-y.
6. Hathaway NJ*, et al.* SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res*. 2018;46:e21–e21. doi:10.1093/nar/gkx1201

Key Results

Tools perform similarly well overall

Despite differences in their algorithms, all tool haplotype resolution performances fell in a comparable range, although tool precision varied much more than sensitivity (Fig. 1A and Fig. 1B). The clear outlier was SeekDeep(1x), which was expected given that SeekDeep was designed to operate on technical replicates. Performance for all tools improved in both metrics by only considering samples with coverage > 100 (Fig. 1B).

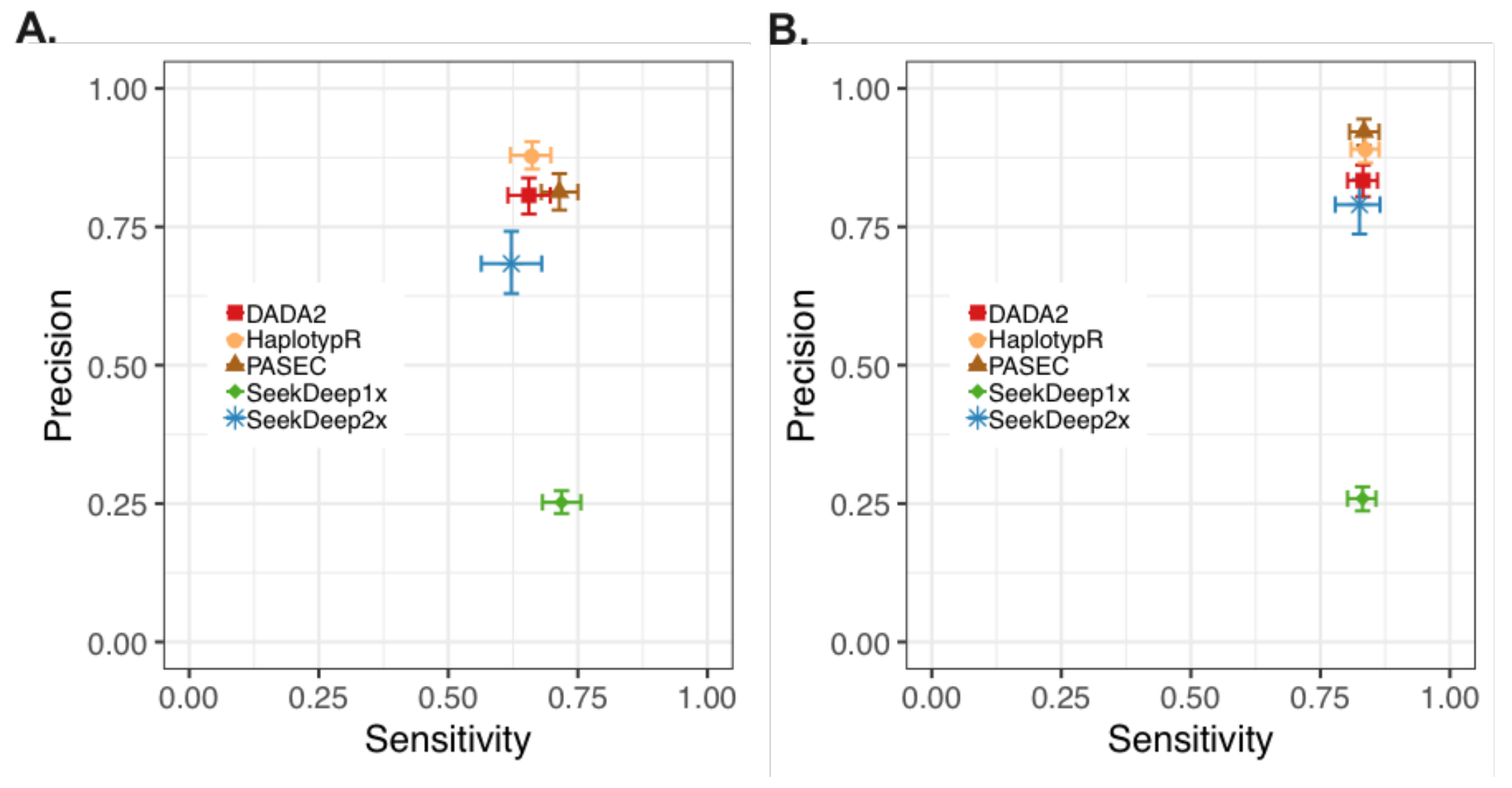


Figure 1. *Precision and recall plots. Error bars represent 95% confidence intervals bootstrapped over 1000 iterations.*

Performance improves with increased coverage

Haplotype set (top) and haplotype frequency (bottom) resolution performance improved with increasing coverage for all tools. F-score (mean of precision and recall) is commonly used as a metric for overall performance of information retrieval algorithms. Haplotype frequency correlation is the Pearson correlation coefficient between observed and expected haplotype frequencies.

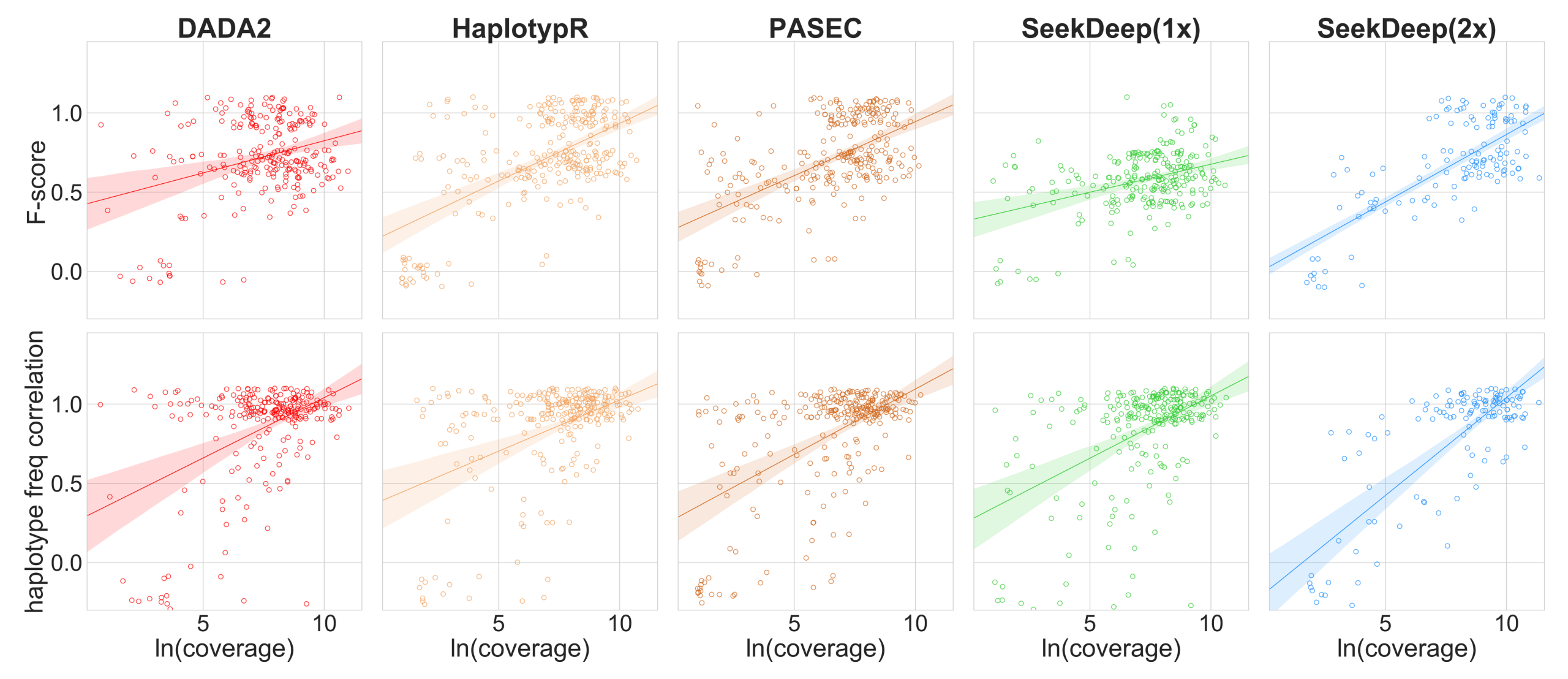


Figure 2. *Correlations between performance metrics and coverage for each tool across all samples.*

Conclusion

Amplicon sequencing is a cost-effective, high-throughput genotyping technology that requires post-sequencing data processing to reduce technical error. We demonstrate that current amplicon sequencing data analysis tools performs similarly well when applied to a dataset of known input. For more details on the methods and analysis, see our preprint here: https://www.biorxiv.org/content/early/2018/10/25/453472.

We also introduce **PASEC**, a cloud-based amplicon sequencing analysis pipeline that is well-validated, highly-scalable and accessible through an easy-to-use web portal.

Together, this work demonstrates progress towards more standardized and validated amplicon sequencing analysis tools for malaria genomic epidemiology.

Contact

Tim Farrell
Infectious Disease and Microbiome Program
tfarrell@broadinstitute.org
https://tmfarrell.github.io