Day Zero Diagnostics
Interview Talk

Tim Farrell

April 2, 2019

Overview

- NEB Internship Summer 2015 (ONT)
- HMS Internship 2015-2016 (Illumina + ML)
- Broad Malaria Group 2017-2019 (Microbial Genomics + Cloud Computing)
  - Amplicon sequencing pipeline validation
  - Phase IV anti-malarial vaccine project (60K+ longitudinal samples)
  - Malarial genomic analysis pipelines in the cloud

Towards Error Mitigation Applications for Oxford Nanopore Technologies

Advisors: Zhiyi Sun
Laurence Ettwiller

Genome Biology Division

August 4, 2015

3GS Tech Overview

Platform	PacBio RS	ONT MinION
Cost ($ K)	695	MAP (1 + .270/run)
Size (in^3 / lbs)	176,000 / 1,895	12 / <2
Throughput (Gb)	0.5	0.05
Run time (hrs)	3-4	48-72
Read length (bp)	10K	8K
Observed error	~11% (single-pass)	>20%
Quality score	Q40	<Q10

ONT Min

(++) portability

(+) cost

(+) direct interrogation of

(--) high error rates/ lo

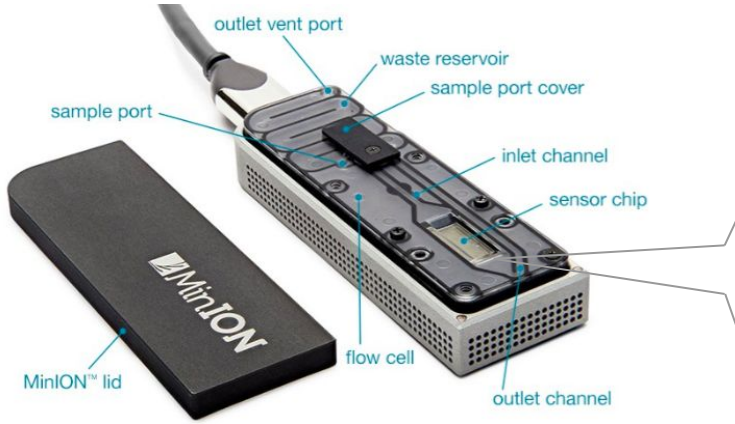
PacBi

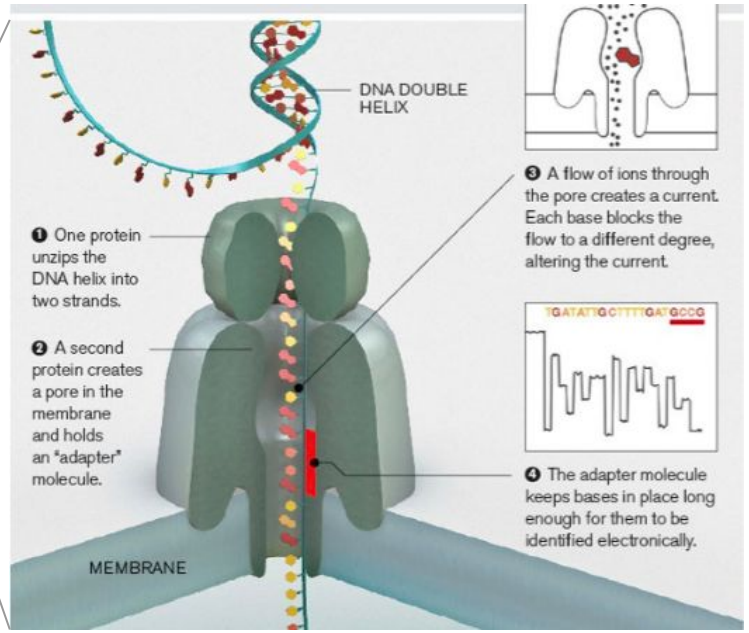
(--) size

ONT-specific Applications

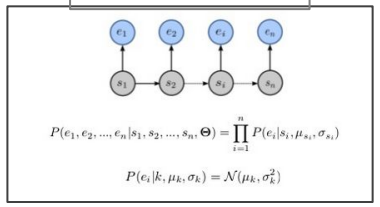
- Clinical microbiology
- Precision medicine
- In field *de novo* assembly
- Epigenomics
- Structural variation analysis

ONT Sequencing Mechanism





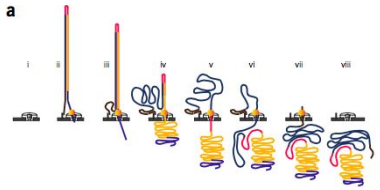
Metrichor

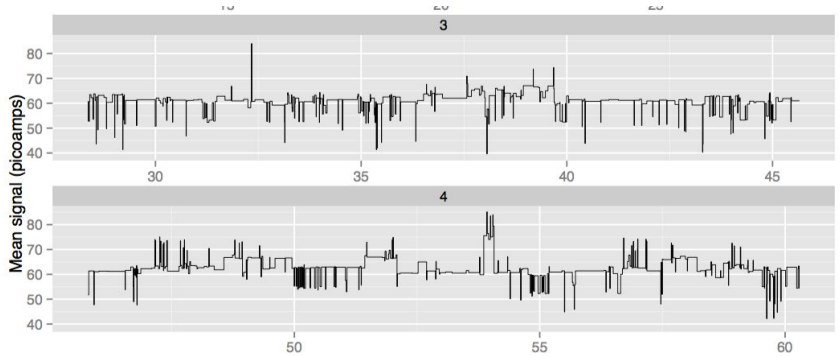


$$P(e_1, e_2, ..., e_n | s_1, s_2, ..., s_n, \Theta) = \prod_{i=1}^{n} P(e_i | s_i, \mu_{s_i}, \sigma_{s_i})$$

$$P(e_i | k, \mu_k, \sigma_k) = \mathcal{N}(\mu_k, \sigma_k^2)$$

…CGATC





Project Concept

Primary:

Use ONT to analyze modified DNA to determine feasibility of pre-sequencing modification of substrates for ‘error mitigation’.

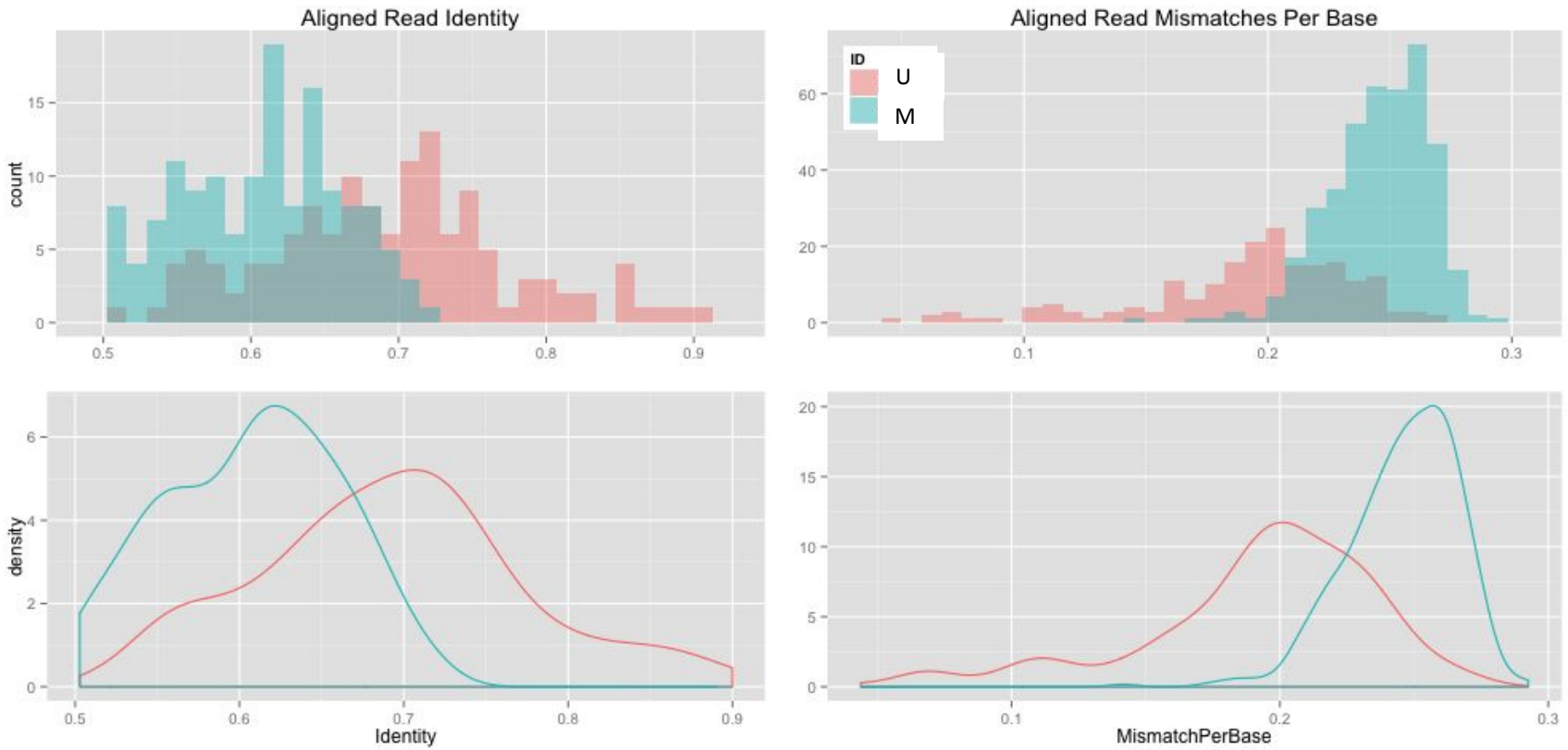
(1) Do base modifications affect ONT read distributions?

(2) Can modifications produce more easily distinguishable signal patterns?

Secondary:

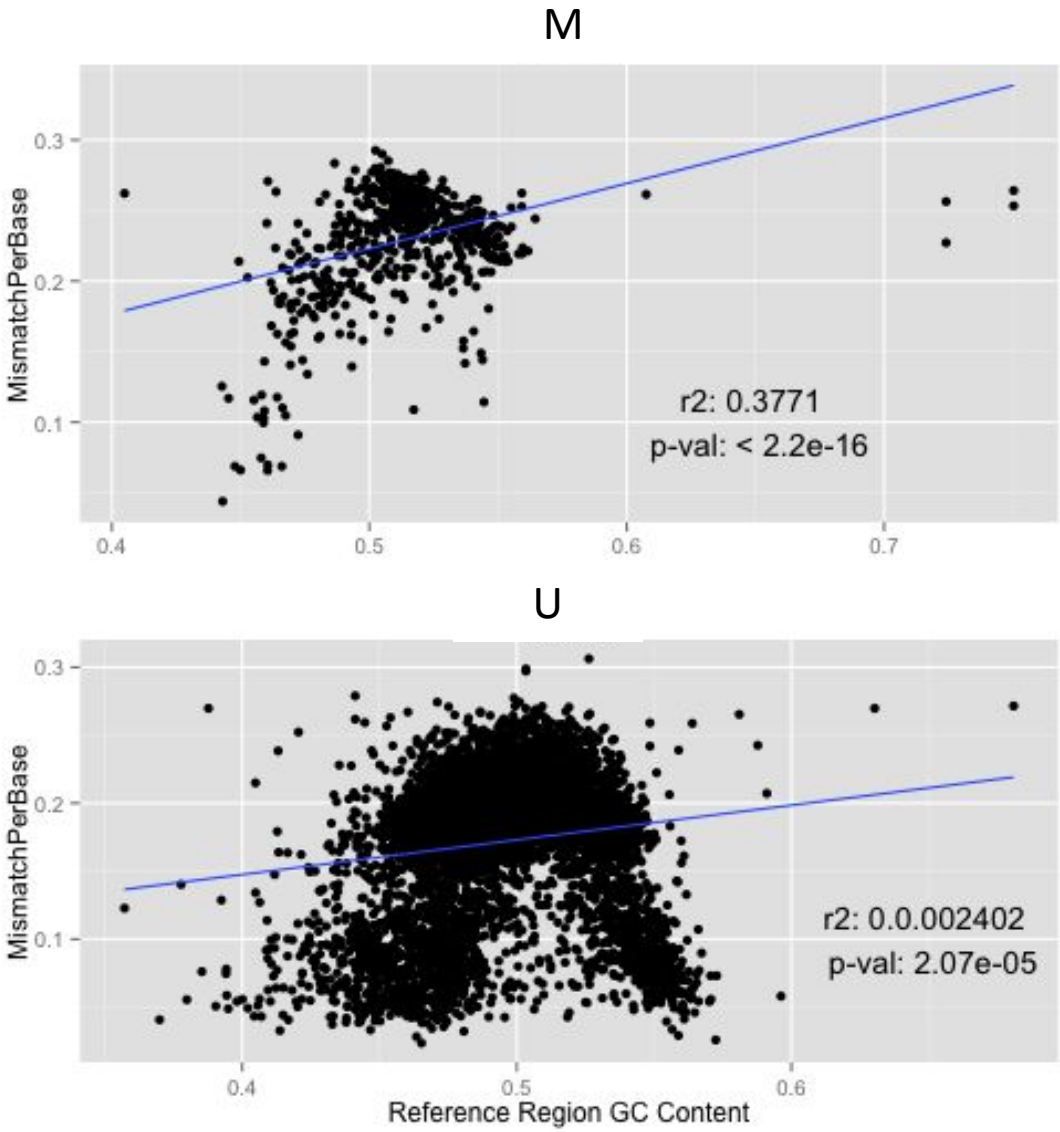
- Assess computational tools available for ONT data
- Build pipeline for future ONT data processing/ analysis

M and U give different distributions



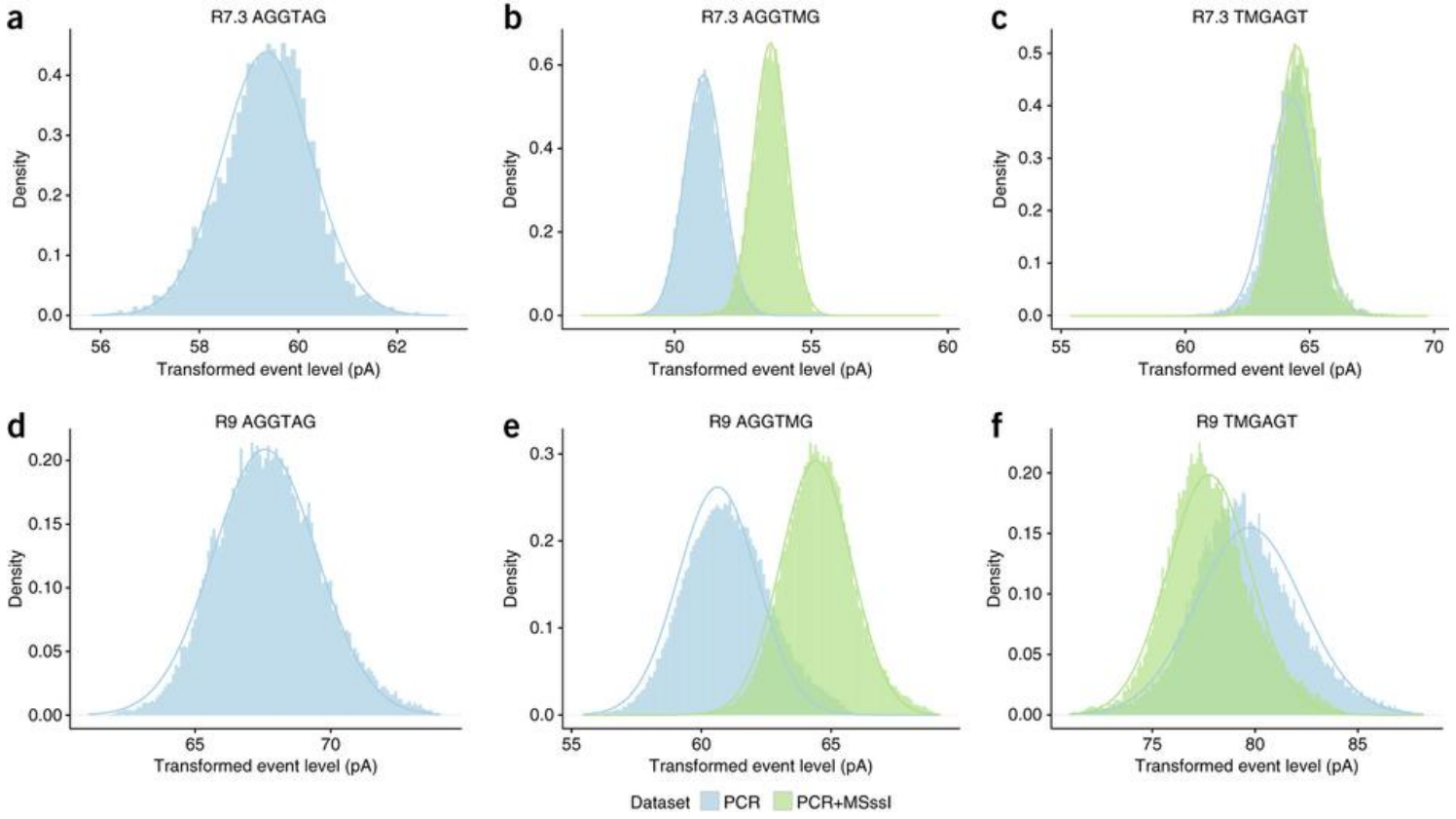
U	M	
55.6%	36.3%	Mean Identity
0.192	0.245	Mean MismatchPerBase

M positively correlated with error



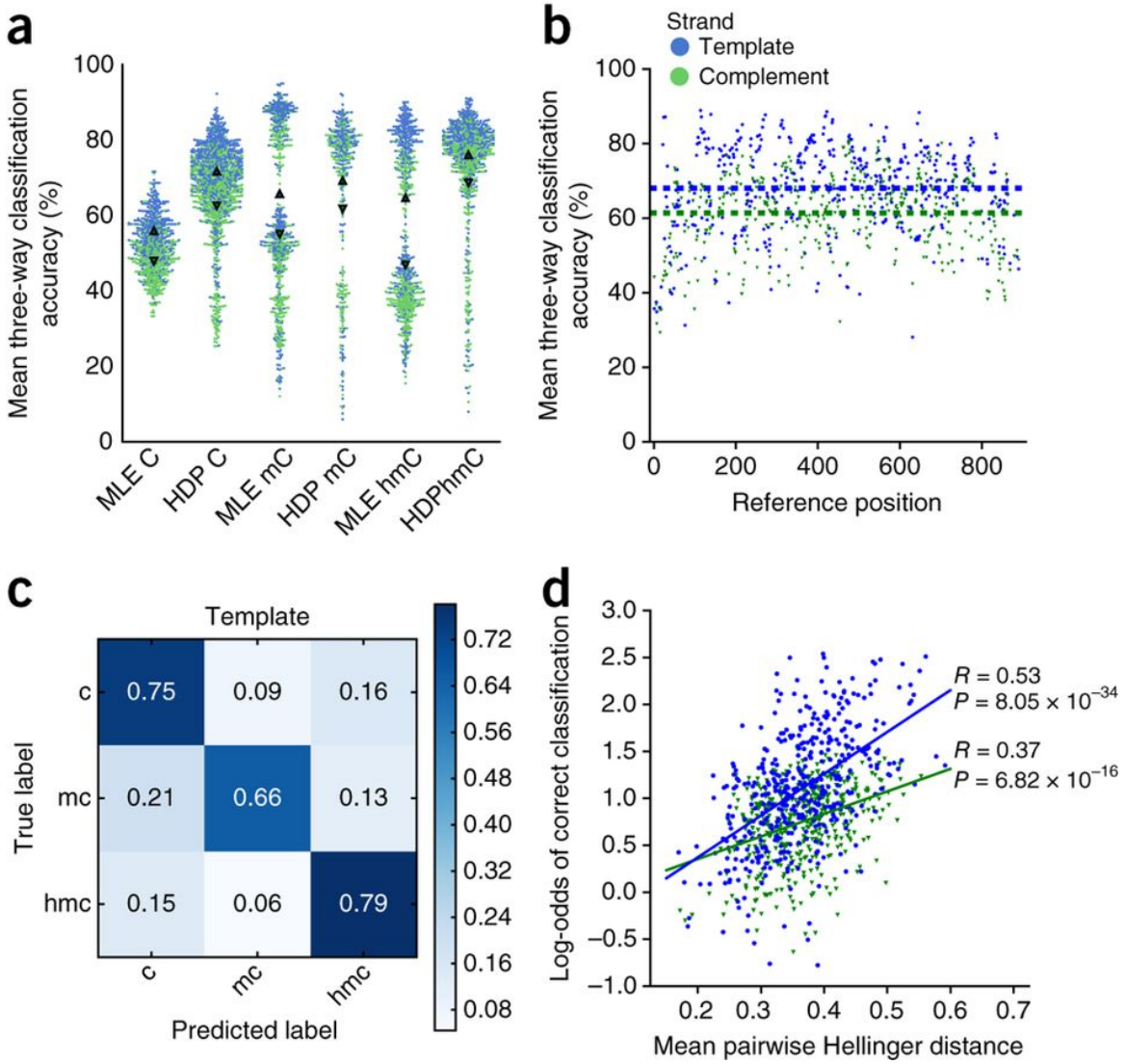
More Recent Developments

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. **Detecting DNA cytosine methylation using nanopore sequencing.** *Nat Methods*, 14: 407–410. doi:10.1038/nmeth.4184.



More Recent Developments

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. **Mapping DNA methylation with high-throughput nanopore sequencing.** *Nat Methods*, 14: 407–410. doi:10.1038/nmeth.4189.



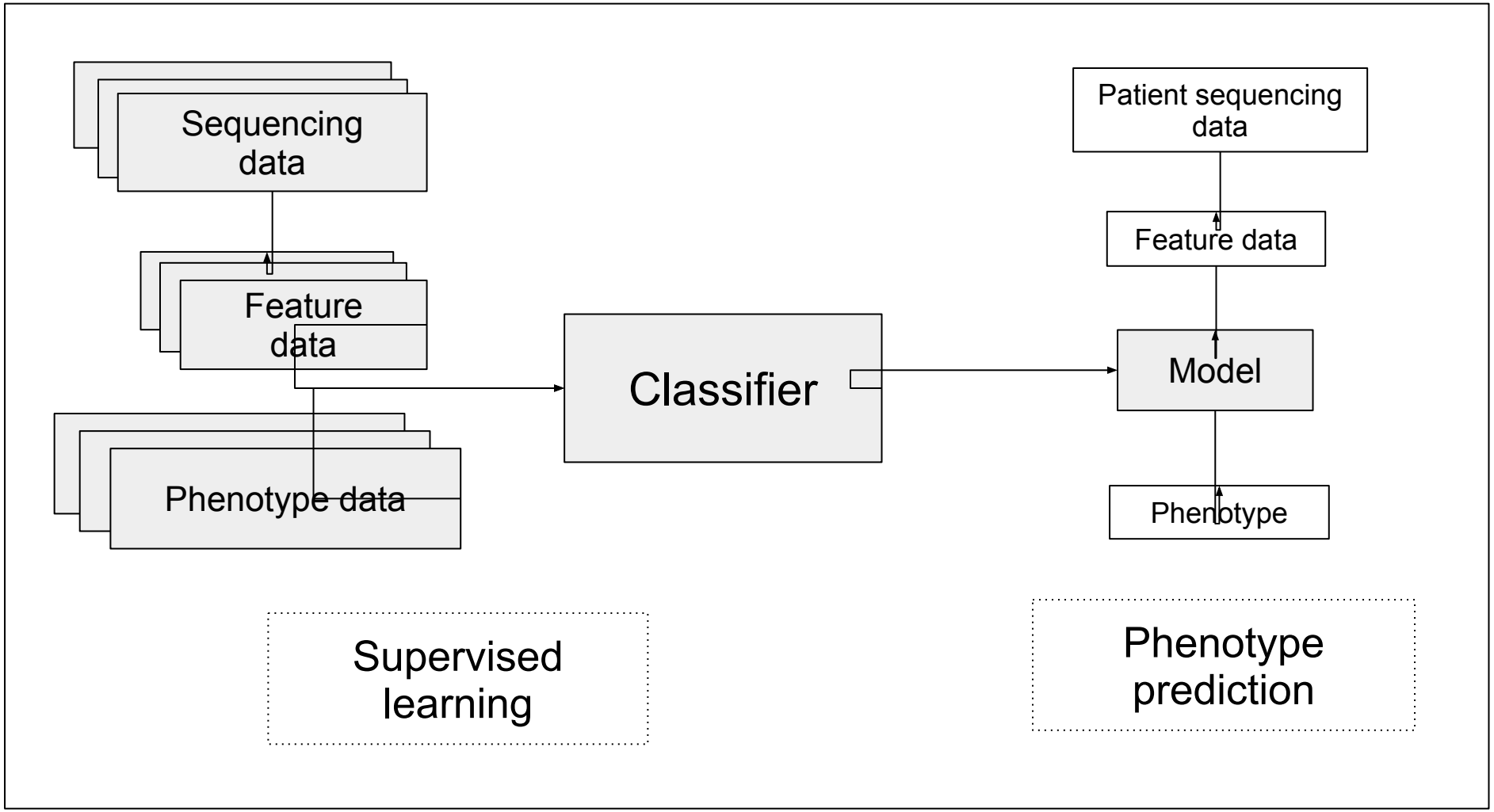
Clinical phenotype prediction from highly-polymorphic structurally-variant genotypes

Tim Farrell
Course Project, BE562
December 11, 2015
tmf@bu.edu

Human genomic variation and clinical sequencing

- 80 million variants identified in human genome (Jun 2015)
  - SNPs
  - CNVs
  - structural (>50bp; inversions, translocations, etc.)
- Discordance b/t sequencing tech and variant callers (VCs)
- Recent study on VC standardization reported 23% of human genome is "difficult" (i.e. not enough consensus among tools to make reasonable prediction)
- Together gives low confidence for "predictive" clinical sequencing

Building robust predictive models for clinical sequencing assays



Rh RBC antigen genes

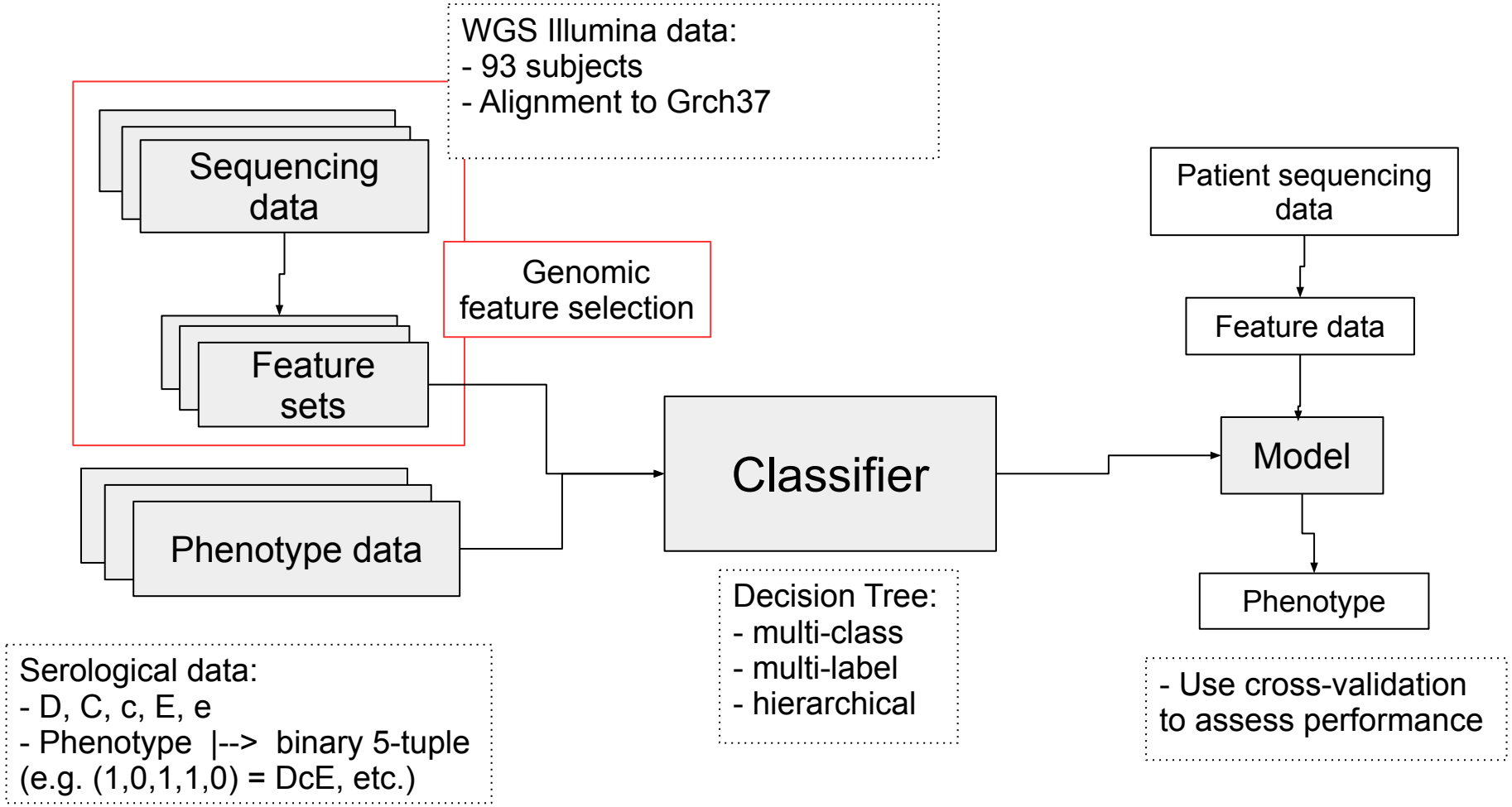
- Rh RBC antigen genomic region exemplifies “difficult”
  - Encodes for highly immunogenic antigens on RBC membranes

- RhCE and RhD
  - Highly similar genes known to undergo complex rearrangements

- 50 known antigens
  - Most significant: D, C, c, E, e
  - Many-to-one relationship haplotypes-to-phenotype (e.g. heterozygosity; but also silent variation, etc)

- Clinical relevance:
  - Blood transfusion
  - Hemolytic disease of the newborn

Rh antigen prediction pipeline



Feature selection: crude

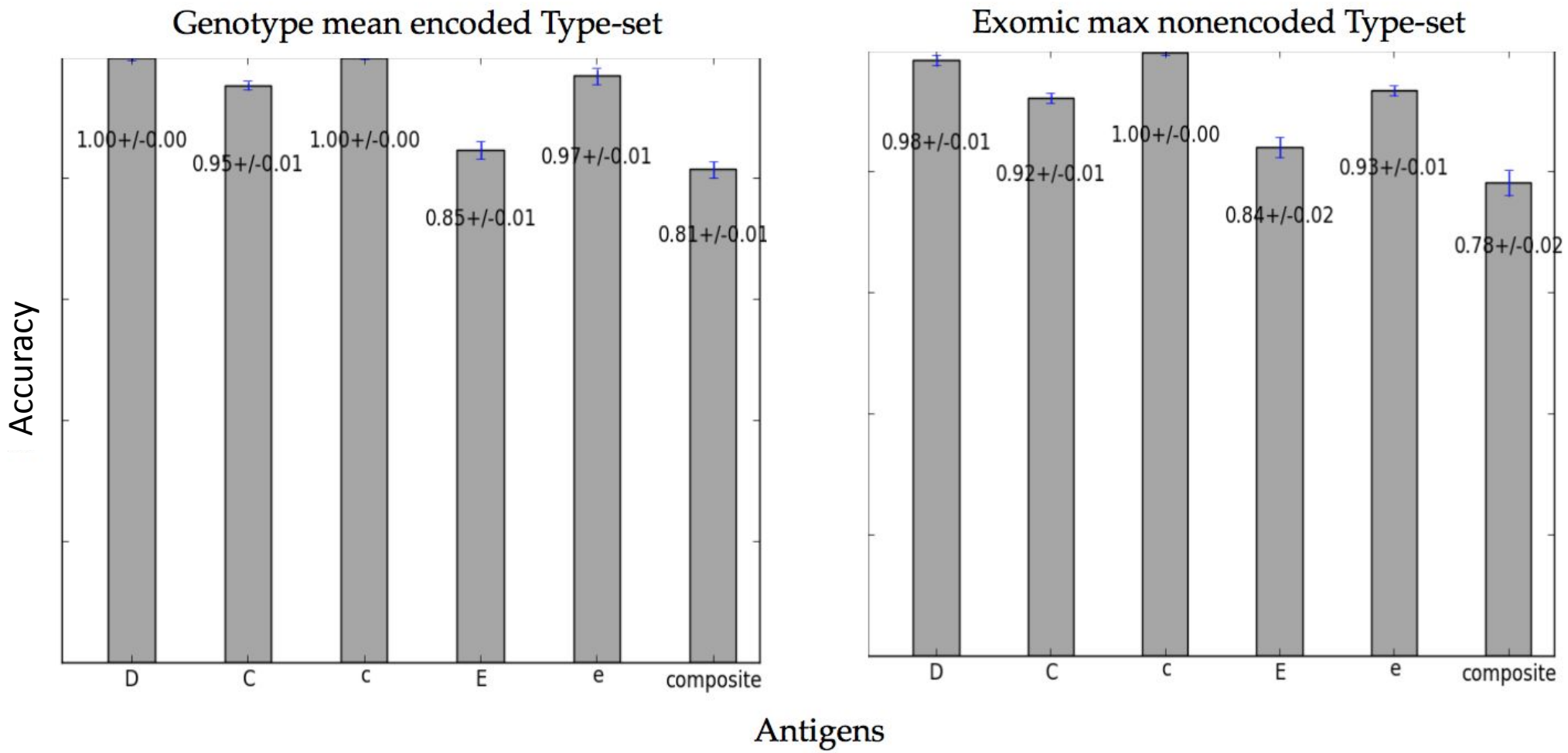
Build PFM for each sample for each gene's exon, then...

- Select
  - Whole exome
  - Variant positions associated with differential phenotypes:
    - dbRBC, ClinVar, dbSNP, dbVar, etc.
    - Call 'genotype'

- Measure:
  - Categorical: call base with highest frequency
  - Position frequency/ max coverage

- Encode:
  - Encoding | Nonencoding
  - e.g. [(1, 4), (2, 3)] |--> [(1, 0, 0, 1), (0, 1, 1, 0)]

Two Best Performing Feature Typesets



Feature selection: fully-featured

- Apply well-established bioinformatics tools to better characterize and differentiate genomic architectures
  - MEME/ DREME:
    - call motifs within exons to eliminate commonalities across genotypes
    - look for motifs in introns that may add specificity
  - HaplotypeCaller: calls SNPs and SV

References

[1] Jameson JL and Longo DL. 2015. Precision medicine – personalized, promising and problematic. *N Engl J Med.* 372(23): 2229-2234.

[2] Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods.* 9(2): 133-139.

[3] Silvestri GA, Vachani A, Whitney D, Elashoff M, Smith KP, Ferguson JS, Parsons E, Mitra N, Brody J, Lenburg ME, and Spira A. 2015. A bronchial genomic classifer for the diagnostic evaluation of lung cancer. *N Engl J Med* 373;3.

[4] Qiu P, Cai X, Ding W, Zhang Q, Norris ED and Greene JR. 2009. HCV genotyping using statistical classifcation approach. *J of Biomed Sci*, 16:62. doi:10.1186/1423-0127-16-62.

[5] Abel HJ , Duncavage EJ. 2014. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics* 206 (2014) 432e440.

[6] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W and Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 30(2): 246- 251.

[7] Seringhaus M and Gerstein M. 2008. Genomics confounds gene classification. *American Scientist*, 96(6) p.466-473.

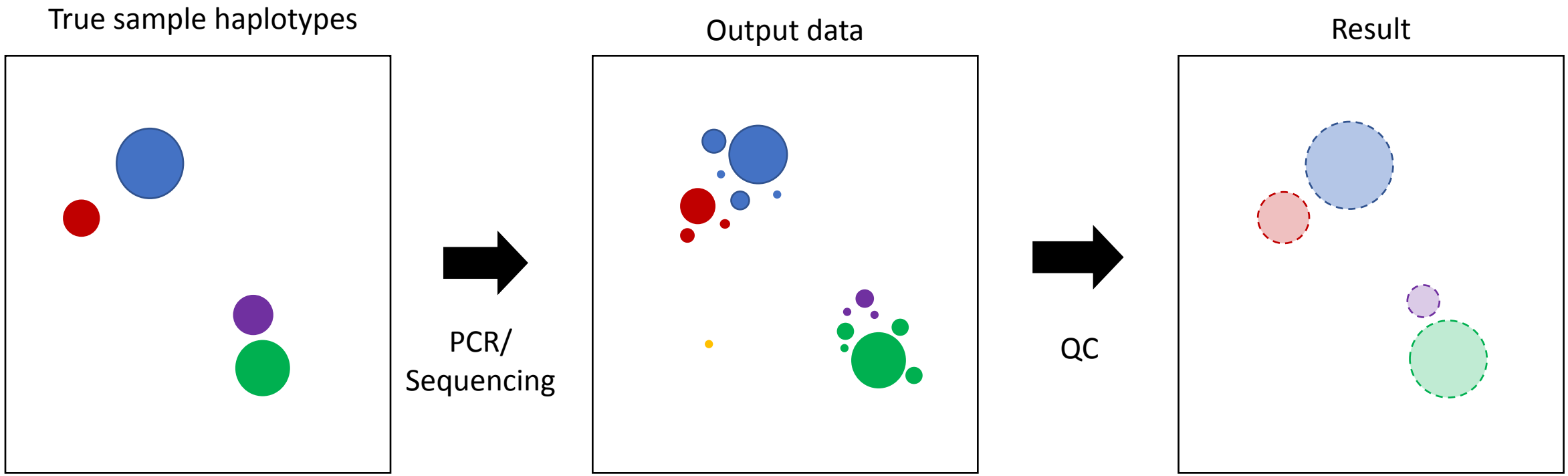
Bill Lane, BWH Pathology

Peter Tonellato, DBMI HMS

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

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

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

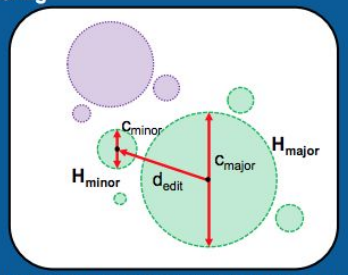
sampleN.aln.bam

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

known_ haplotypes.db

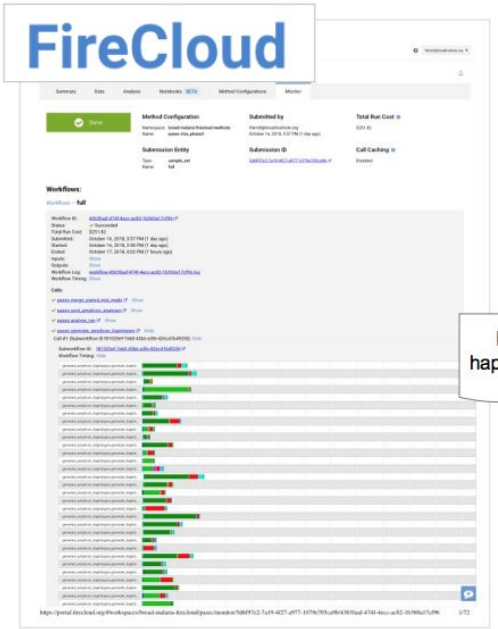
Index amplicon haplotype sets and classify against database

Index sample haplotypes lists and pool results from all amplicons

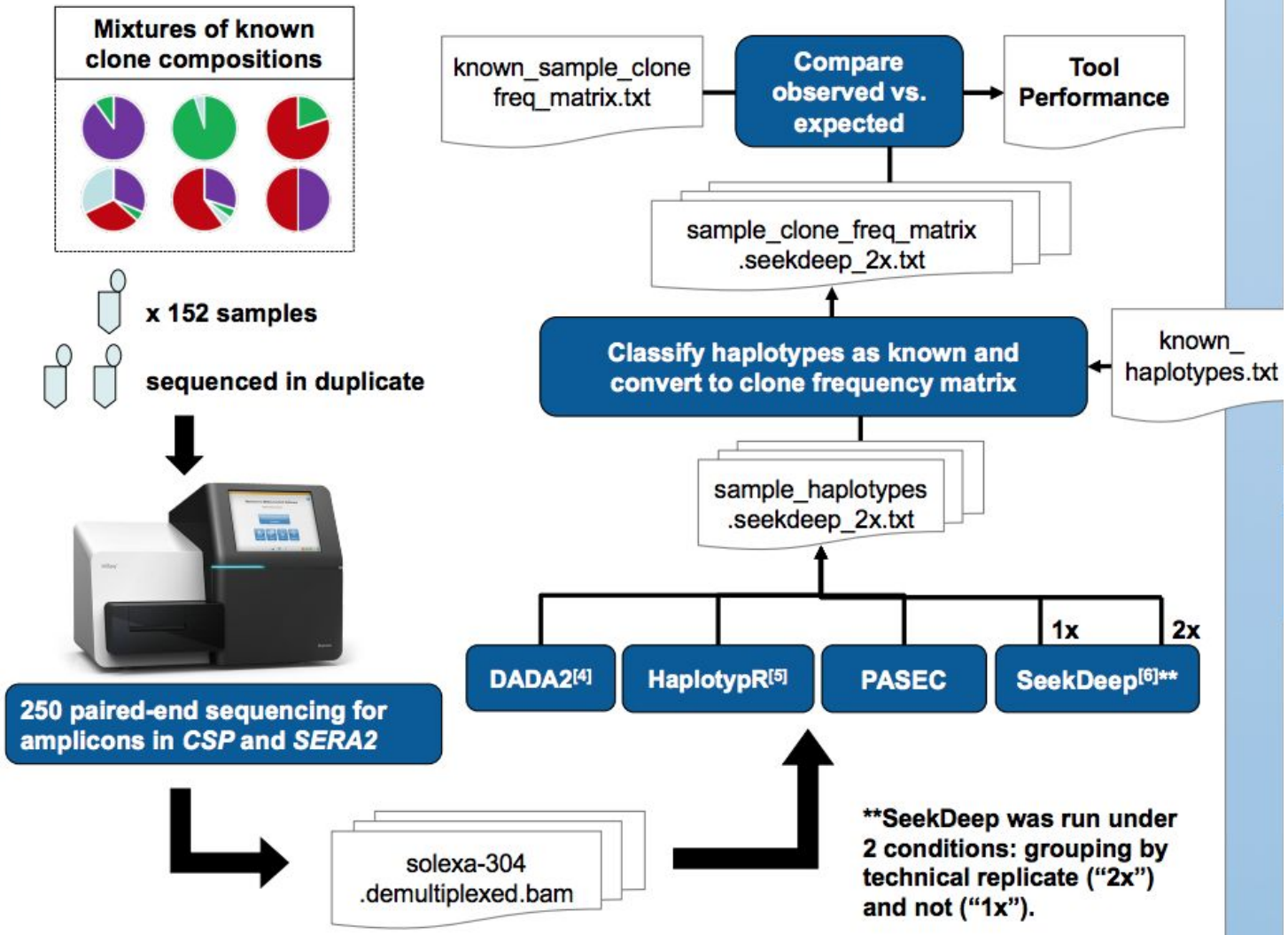
sample_haplotypes .indexed.txt

haplotype_set.index.txt

FireCloud



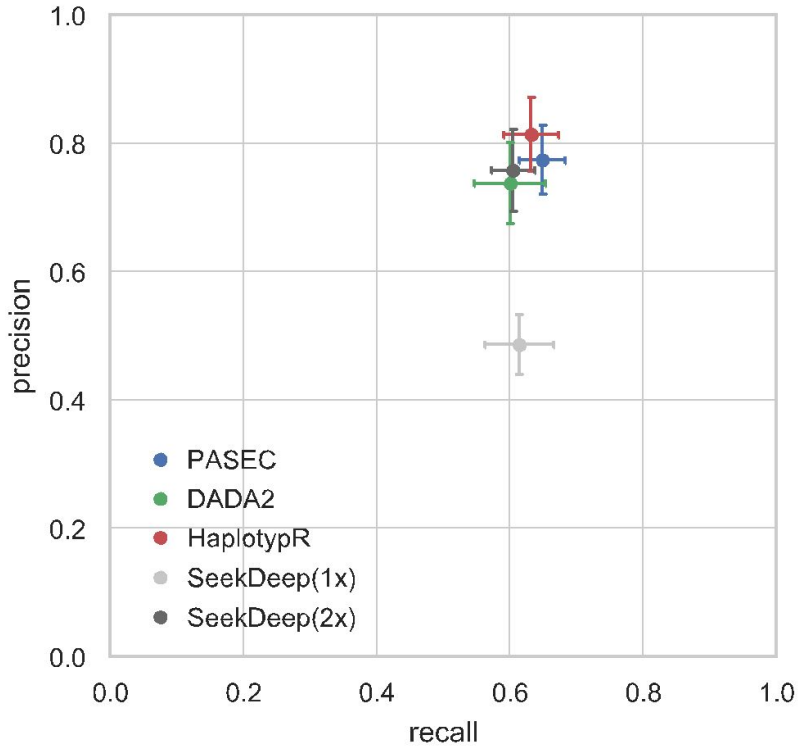
Methods

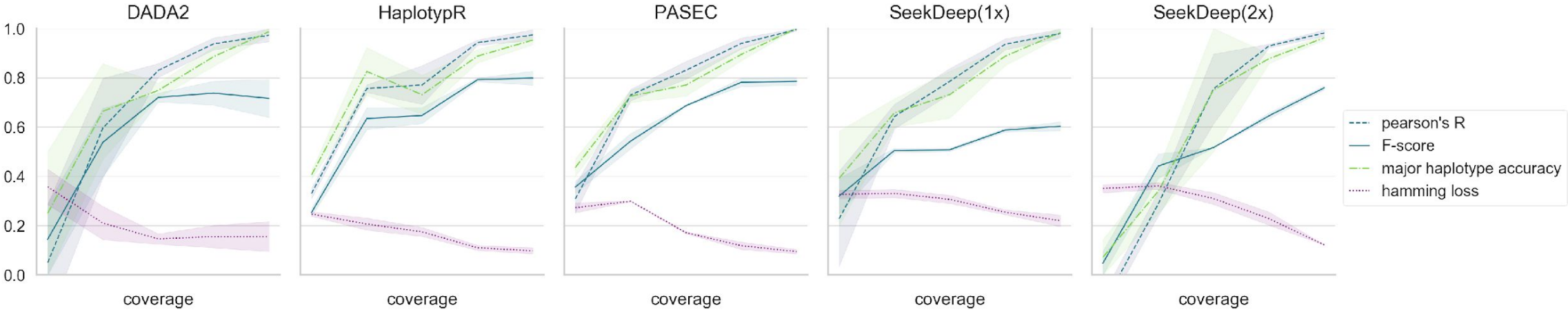


**SeekDeep was run under 2 conditions: grouping by technical replicate ("2x") and not ("1x").

Tool performance comparison

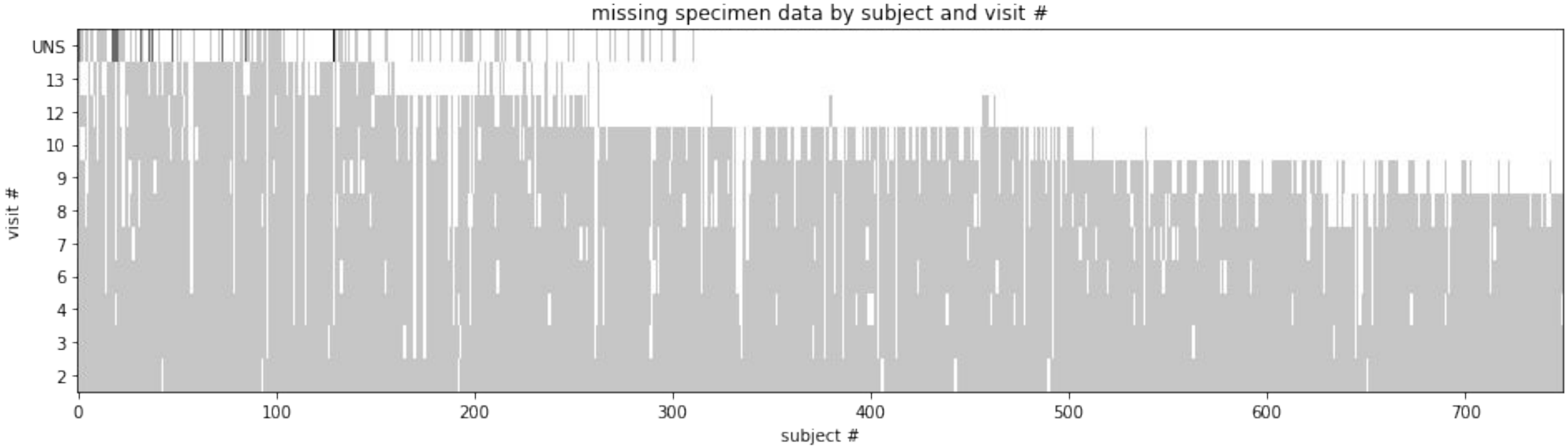
https://www.biorxiv.org/content/early/2018/10/25/453472.

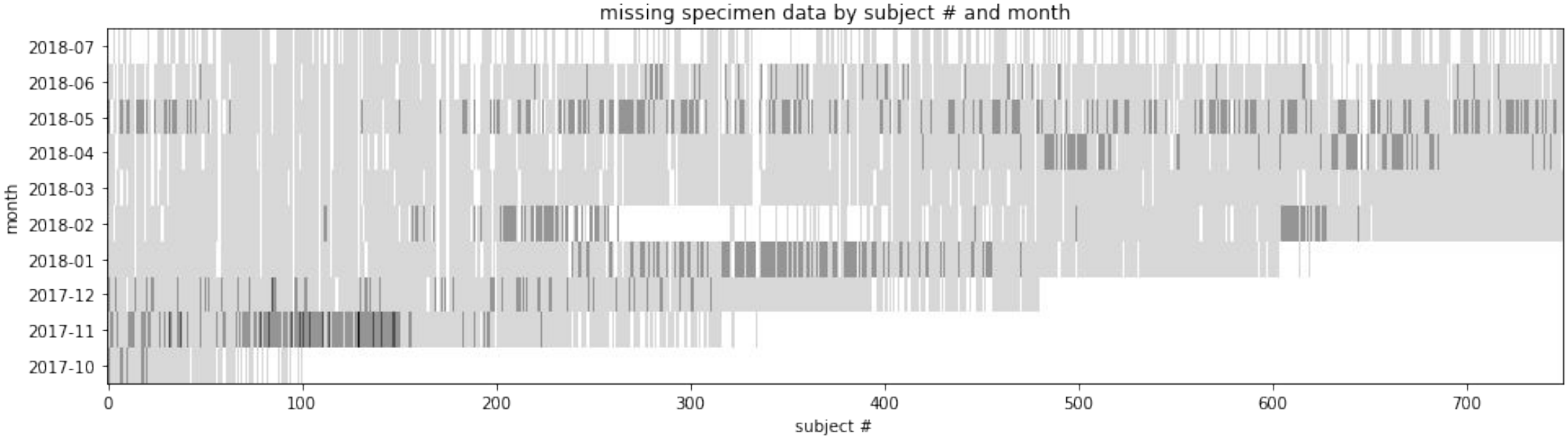


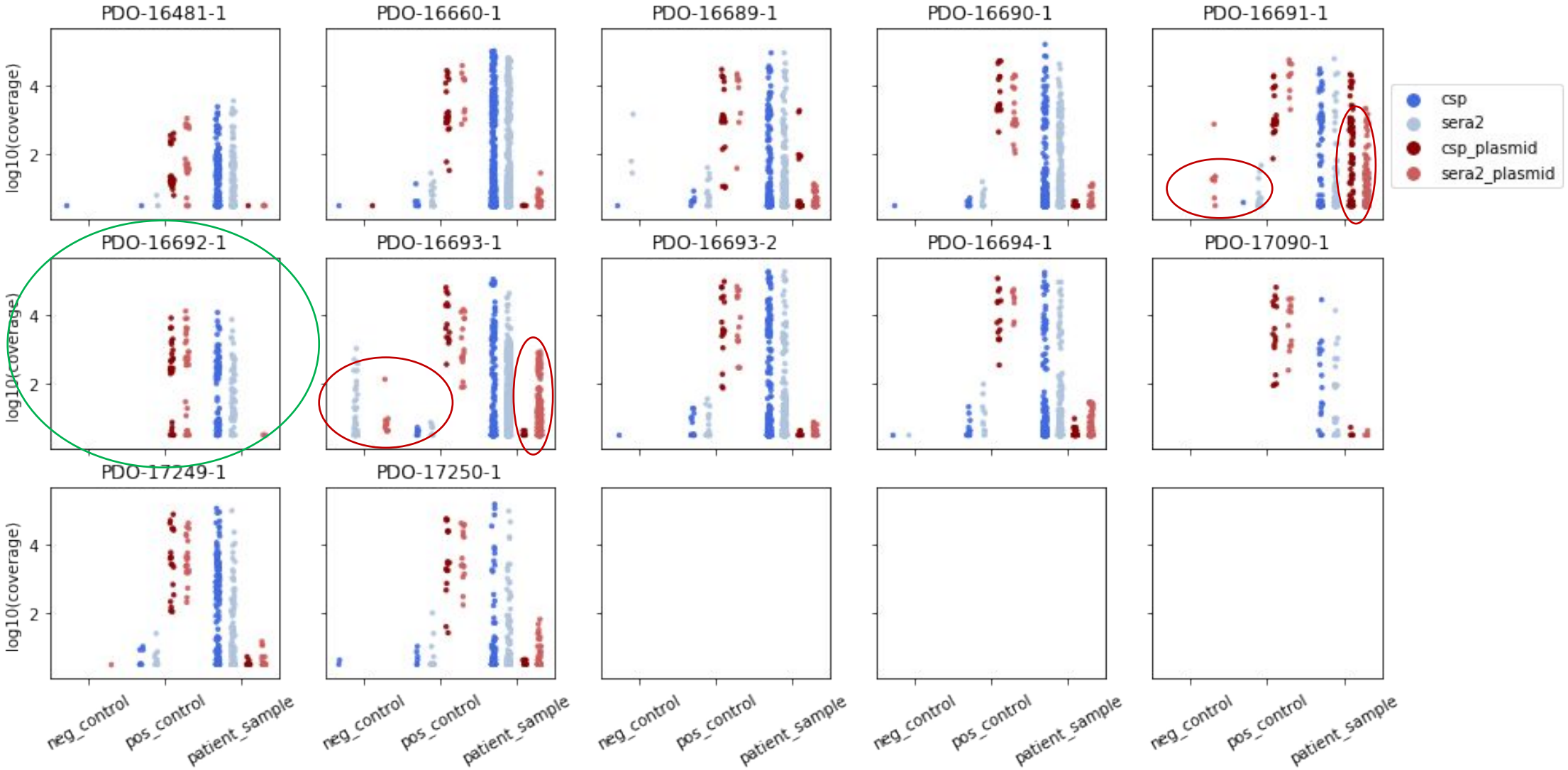


Phase IV anti-malarial vaccine clinical trial project

2018-2019

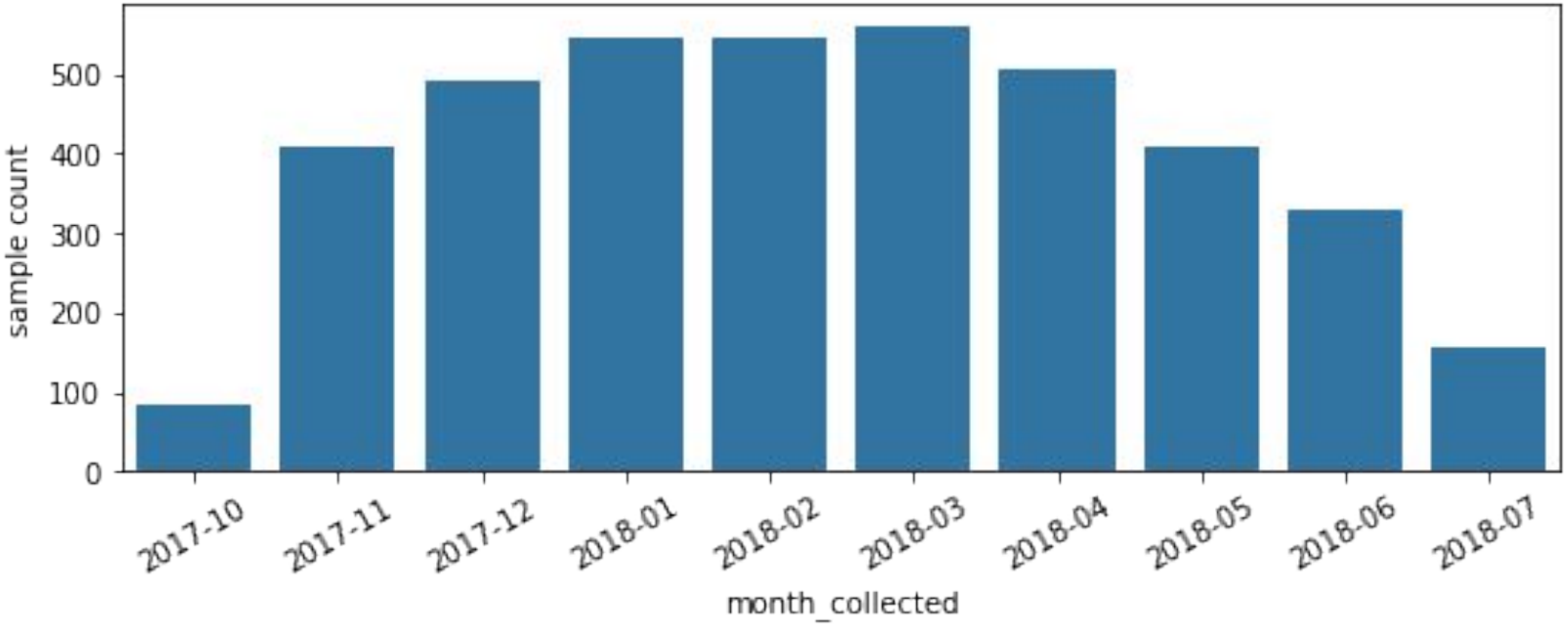

missing specimen data by subject and visit #
visit #
UNS
13
12
10
9
8
7
6
4
3
2
0
100
200
300
400
500
600
700
subject #

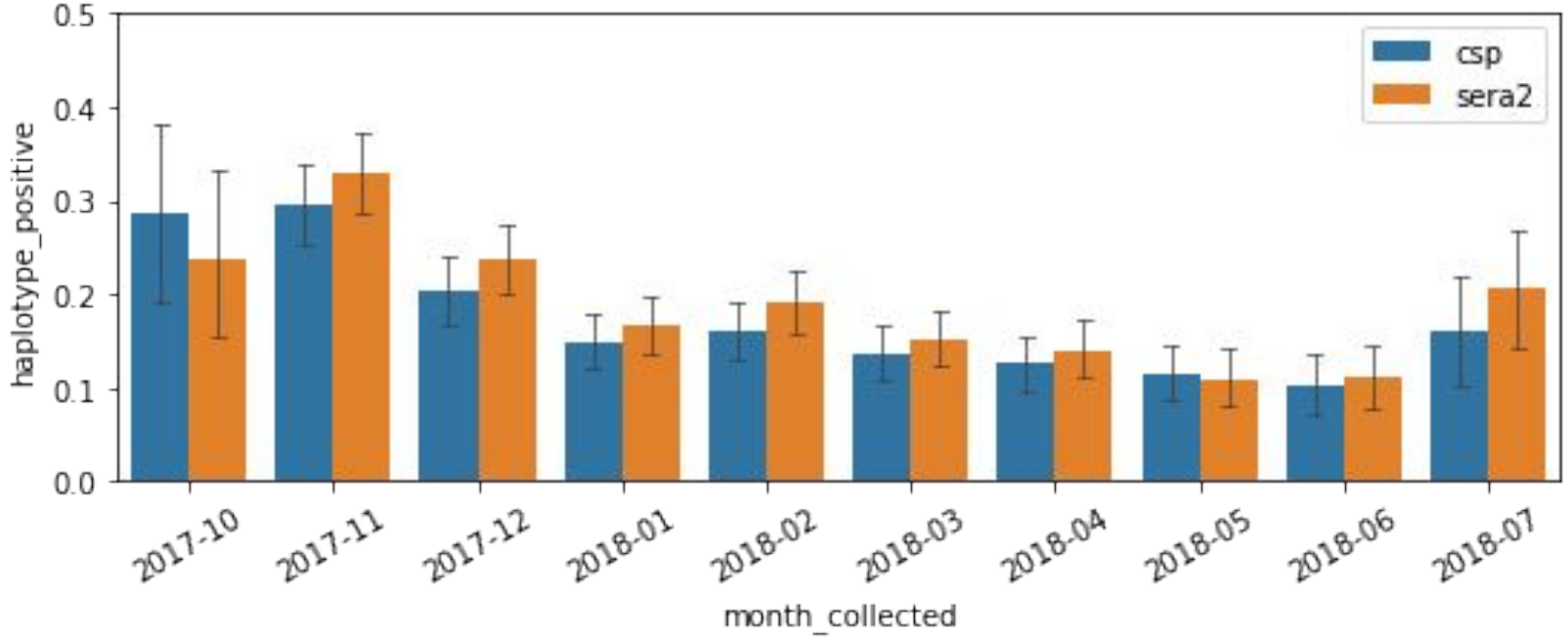

missing specimen data by subject # and month
month
2018-07
2018-06
2018-05
2018-04
2018-03
2018-02
2018-01
2017-12
2017-11
2017-10
0
100
200
300
400
500
600
700
subject #



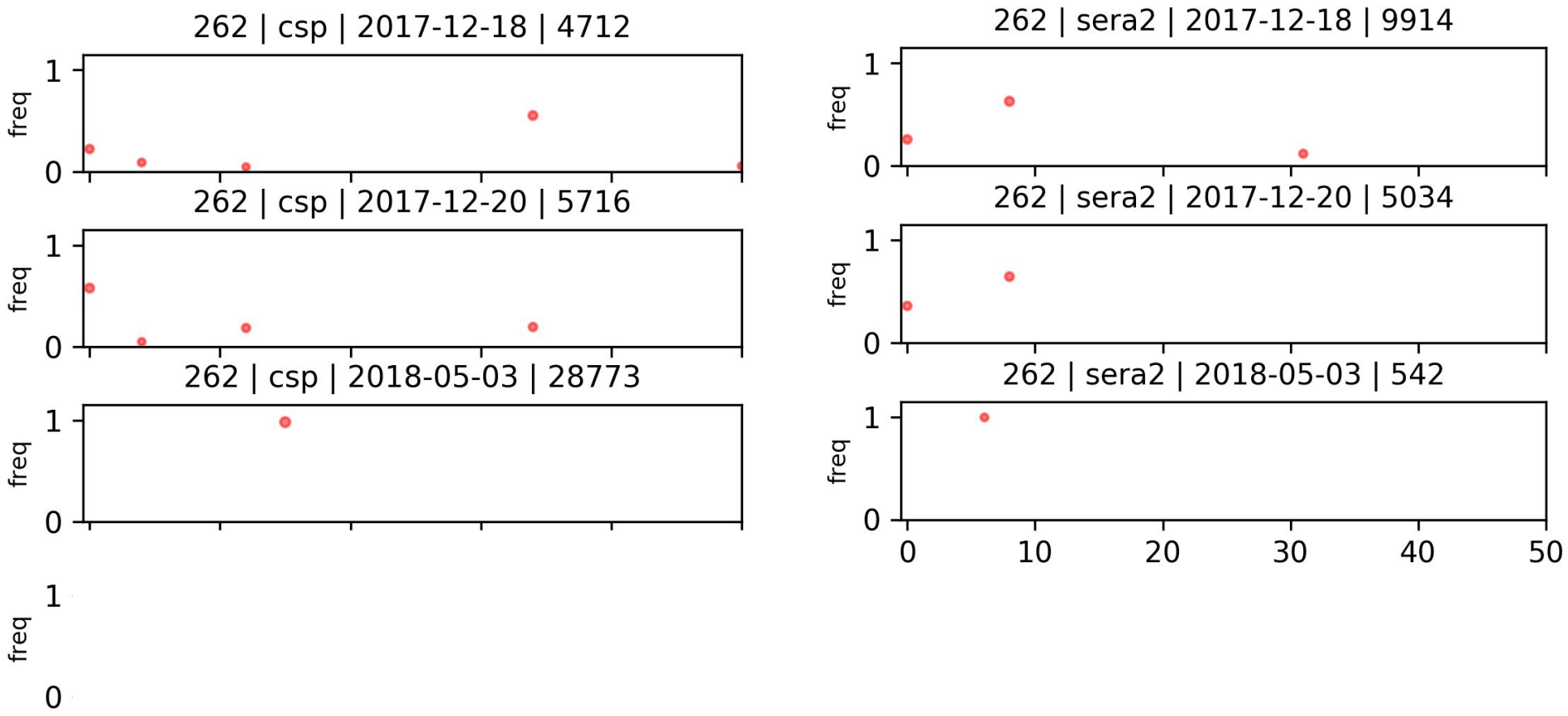
Amplicon (blue) and positive control plasmid (red) log10(coverage) by PDO (each box, with PDO-version as title) and sample type (x-axis). Red circles indicate contamination (amplicons in controls or plasmids in negative controls or patient samples). The green circle indicates one of the cleanest (but 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?



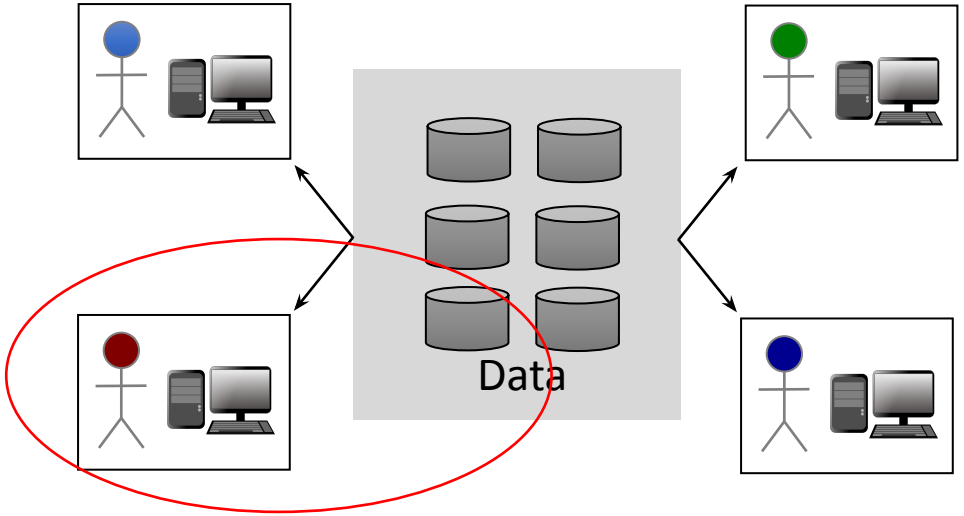
Malarial Genomics on FireCloud

2018-2019

Anthony Philippakis
CDO, Broad Institute

Inverting the Model of Genomic Science

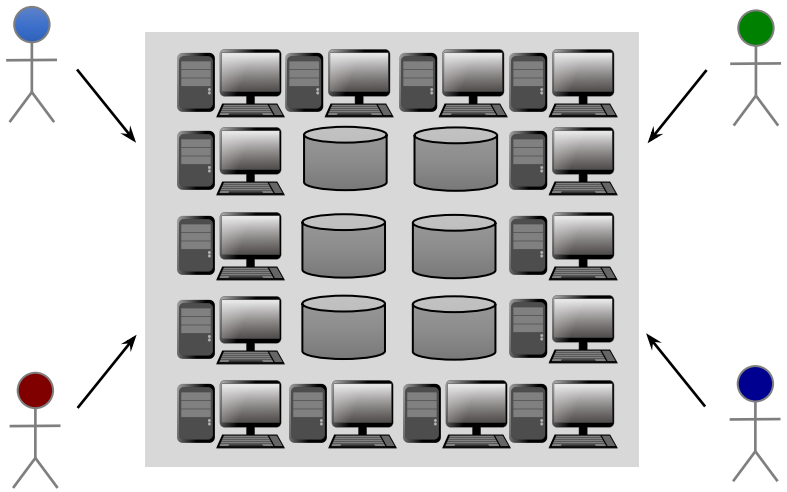
Traditional Approach:
Bring data/ tools to researchers



Problems

- Data Sharing = Data Copying
  - High cost and inconsistency
- Infrastructure Needed
- Siloed Compute
  - Hard to replicate/ reuse

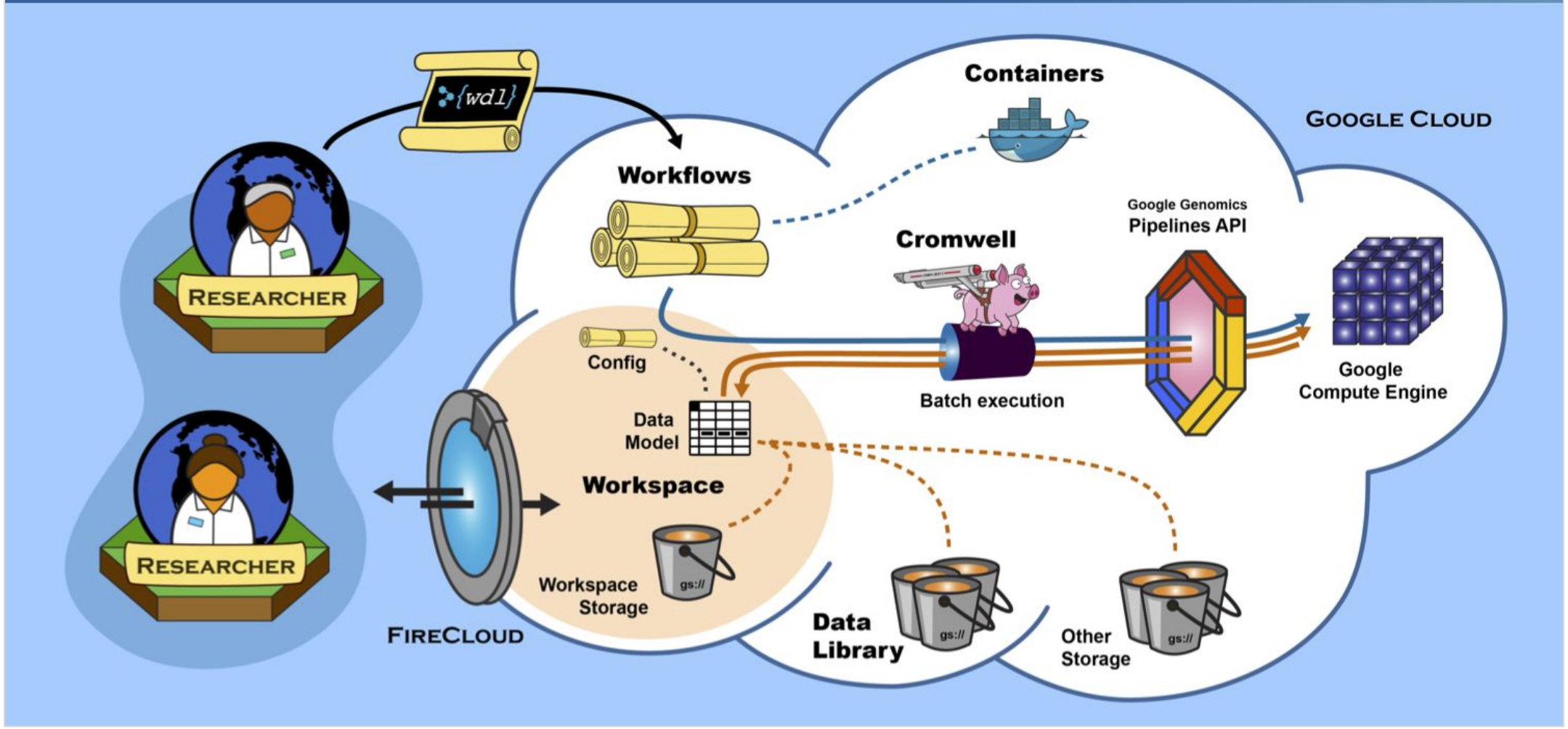
Cloud Approach:
Bring researchers to data/ tools



Advantages

- Cost and Consistency
- Increased Accessibility
  - And more control
- Shared & Elastic Compute

Batch execution of workflows in FireCloud



FireCloud

https://portal.firecloud.org

Applicability to Malarial Genomics

Science

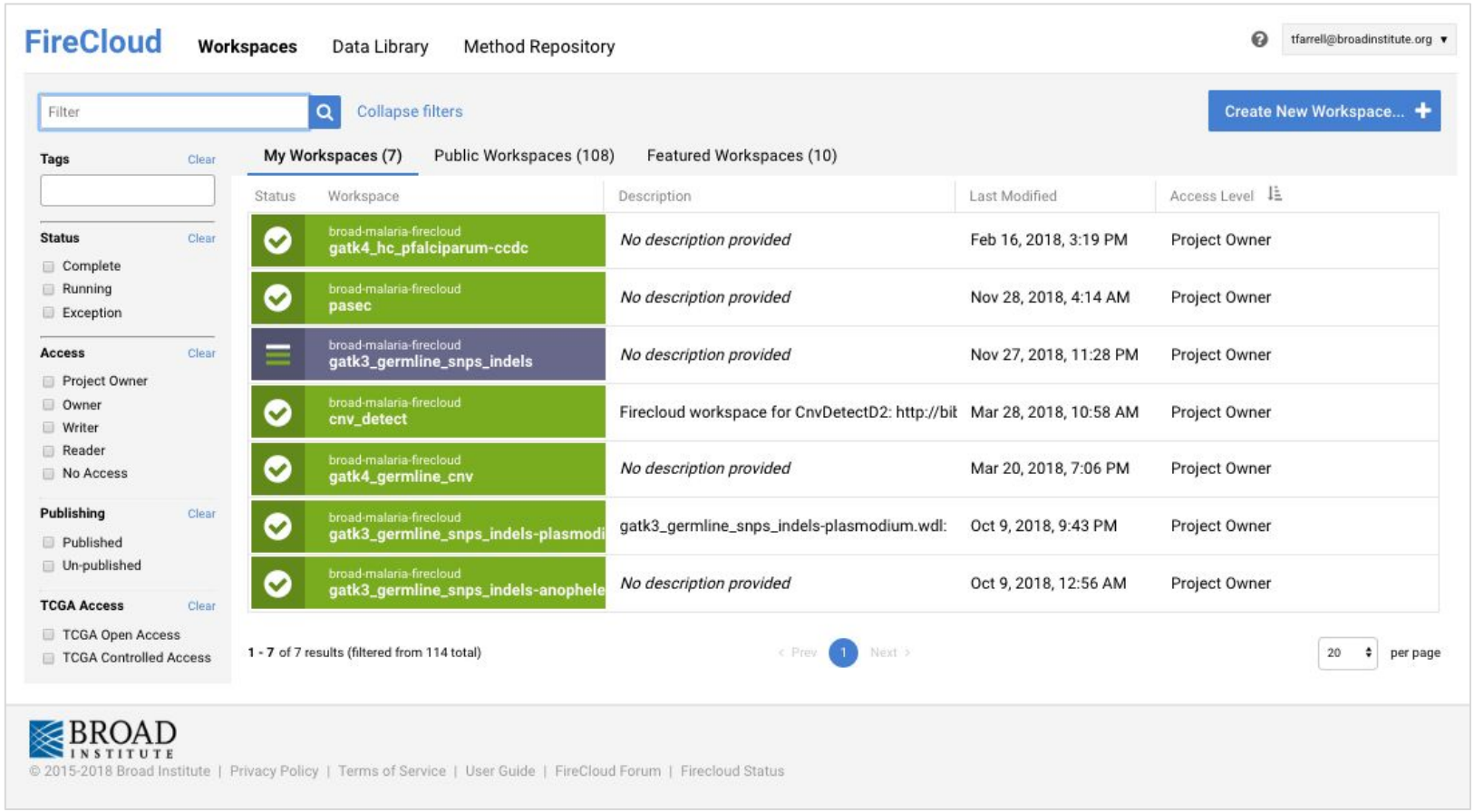
- Data, compute infrastructure and pipelines used both in-house and made available to collaborators
  - Standardized, reusable, transparent workflows
  - Bioinformatics expertise not necessarily required
  - Centralized data store for aggregating datasets

Systems (useful science ⇒ automated for routine use)

- Statistics for use by policy makers and clinical decision-makers
- Surveillance
- Prediction/ classification

broad-malaria-firecloud

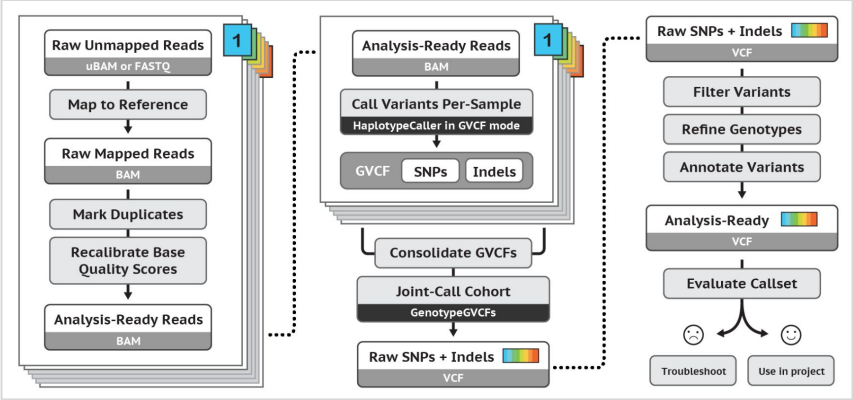
Workspaces for variant-calling, CNV-calling and amplicon sequencing analysis

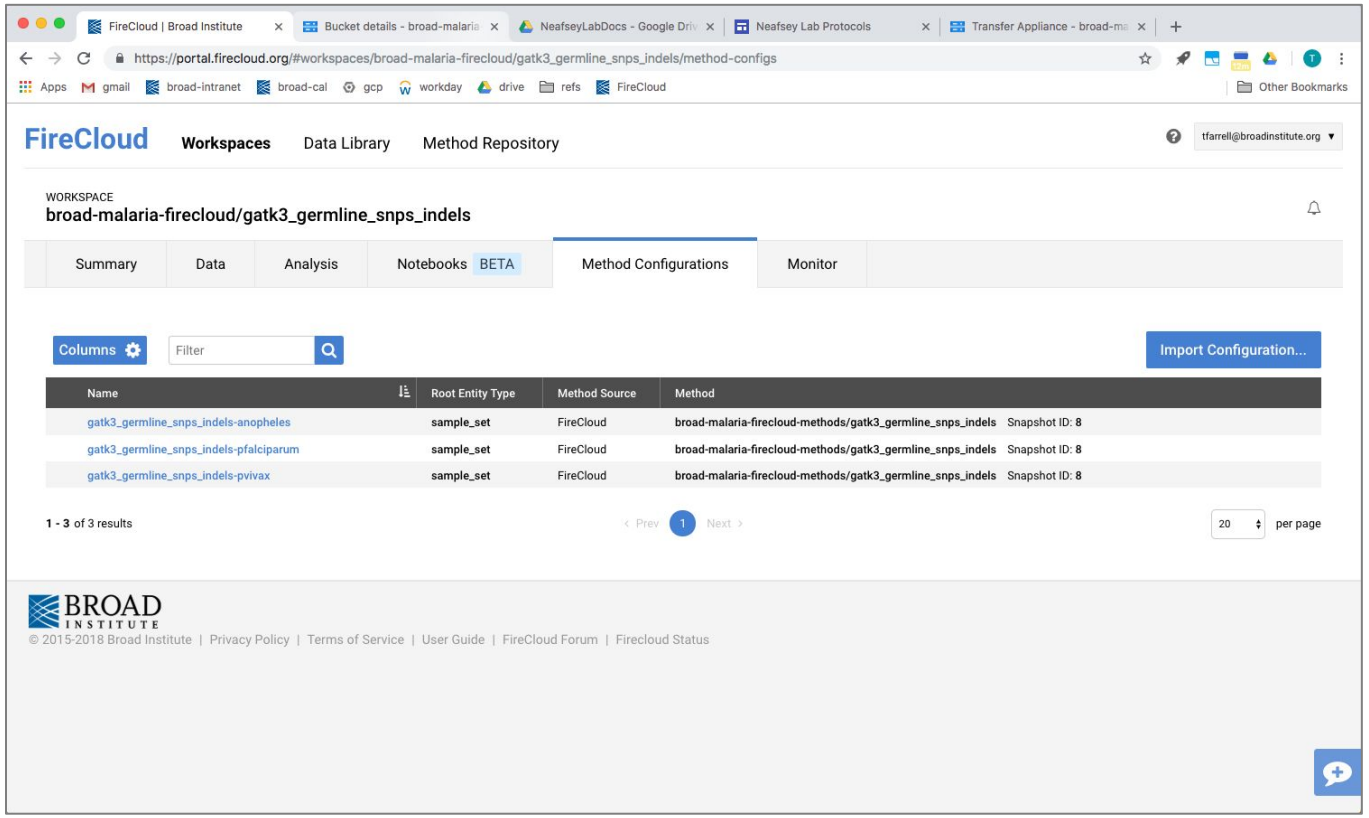


broad-malaria-firecloud/gatk3_germline_snps_indels

- Configurations for parasite and vector variant-calling

- Parallelizes over samples as well as over genomic intervals





broad-malaria-firecloud/gatk3_germline_snps_indels

- 96 anopheles bams, avg. 10 GB
- $400 for compute ($4/ bam)
- 30 hours to compute (10 hrs of variant QC, since been parallelized)

WORKSPACE
broad-malaria-firecloud/gatk3_germline_snps_indels-anopheles

Summary | Data | Analysis | Notebooks BETA | Method Configurations | Monitor

Done

Method Configuration
Namespace: broad-malaria-firecloud-methods
Name: gatk3_germline_snps_indels-anopheles-pignatelli-1_SMv30rH65aU

Submission Entity
Type: sample_set
Name: full

Submitted by
tfarrell@broadinstitute.org
September 21, 2018, 2:09 PM (2 months ago)

Submission ID
e0eb8400-92e7-4475-8c69-438f06ee823e

Total Run Cost
$385.33

Call Caching
Disabled

Workflows:

Workflows > **full**

Workflow ID: 412c262c-d784-481f-bac5-7f067f67b03b
Status: Succeeded
Total Run Cost: $385.33
Submitted: September 21, 2018, 2:11 PM (2 months ago)
Started: September 21, 2018, 2:36 PM (2 months ago)
Ended: September 22, 2018, 8:03 PM (2 months ago)
Inputs: Show
Outputs: Show
Workflow Log: workflow.412c262c-d784-481f-bac5-7f067f67b03b.log
Workflow Timing: Show

Calls:

- GATK3_Germline_Variants.HaplotypeCallerGvcf_GATK3 Show
- GATK3_Germline_Variants.GenotypeGVCFs Show
- GATK3_Germline_Variants.HardFiltration Show
- GATK3_Germline_Variants.CombineGVCFs Show
- GATK3_Germline_Variants.MarkDuplicates Show
- GATK3_Germline_Variants.ReorderBam Show
- GATK3_Germline_Variants.GatherVCFs Show
- GATK3_Germline_Variants.SnpEff Show

Thanks

Dan Neafsey
Bronwyn MacInnis
Dyann Wirth
Seth Redmond
Angela Early
Jacob Tennessen
Thais de Oliveira
Broad GP
Broad DSP

BROAD INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

BILL & MELINDA GATES foundation