A possible way to identifying mixed infected samples

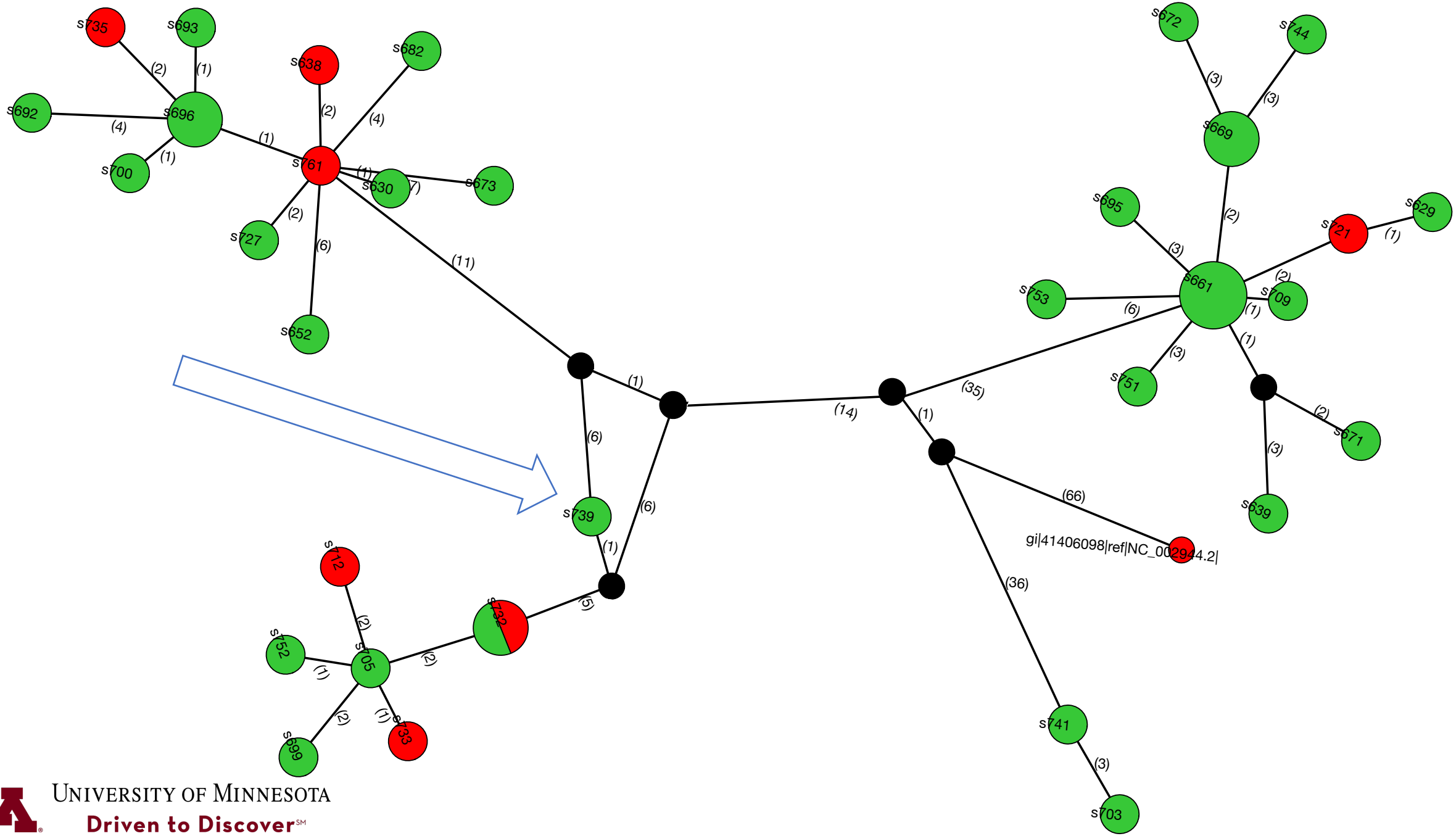
Ph. D candidate Yuanyuan Wang

Computational Biology and Bioinformatics

University of Minnesota, Twin Cities

1. Multiple infections need to be identified

- Why do we care about multiple infections?
  - No recognition of mixture sample will result in wrong phylogenetic trees
  - Multiple infected individuals associated increased transmission intensity in the local environment. (Manske M, Miotto O, et al., 2012). Identifying those individuals maybe crucial to disease control.



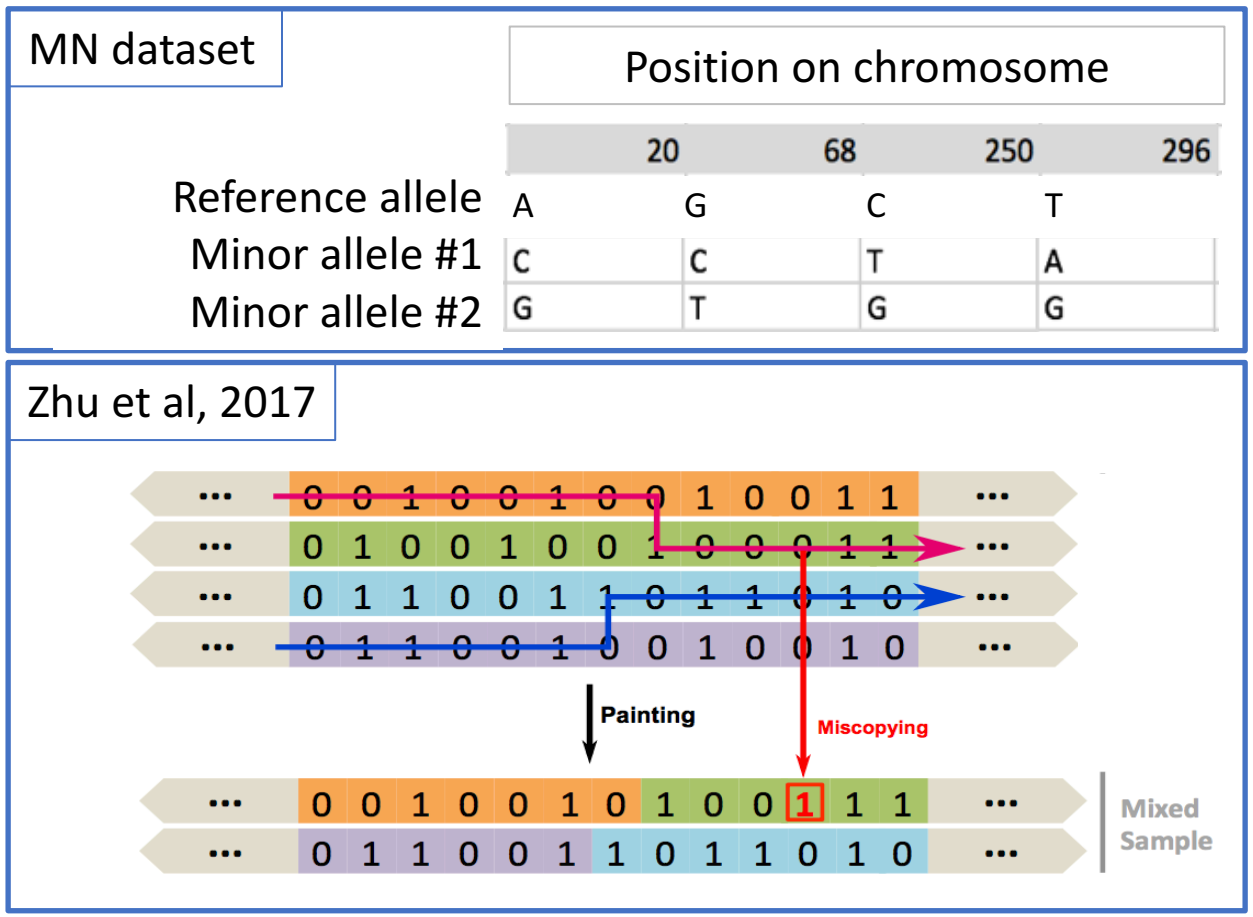
UNIVERSITY OF MINNESOTA
Driven to Discover℠

2. Definitions of multiple infections

- From WGS data, how to tell if the animal has multiple infection?
  1) [Biology]
     Heterozygous loci:
     - when a loci that has multiple alternative alleles. (not suitable for single-colony isolates)
  2) [Transmission]
     Patterns/Haplotypes:
     - when a sequence (sample) shares haplotypes/ evolutionary path with other sequences.

MN dataset

	Position on chromosome			
	20	68	250	296
Reference allele	A	G	C	T
Minor allele #1	C	C	T	A
Minor allele #2	G	T	G	G



3. A quick way to screening for mixture samples using allele frequency plots

RESEARCH ARTICLE

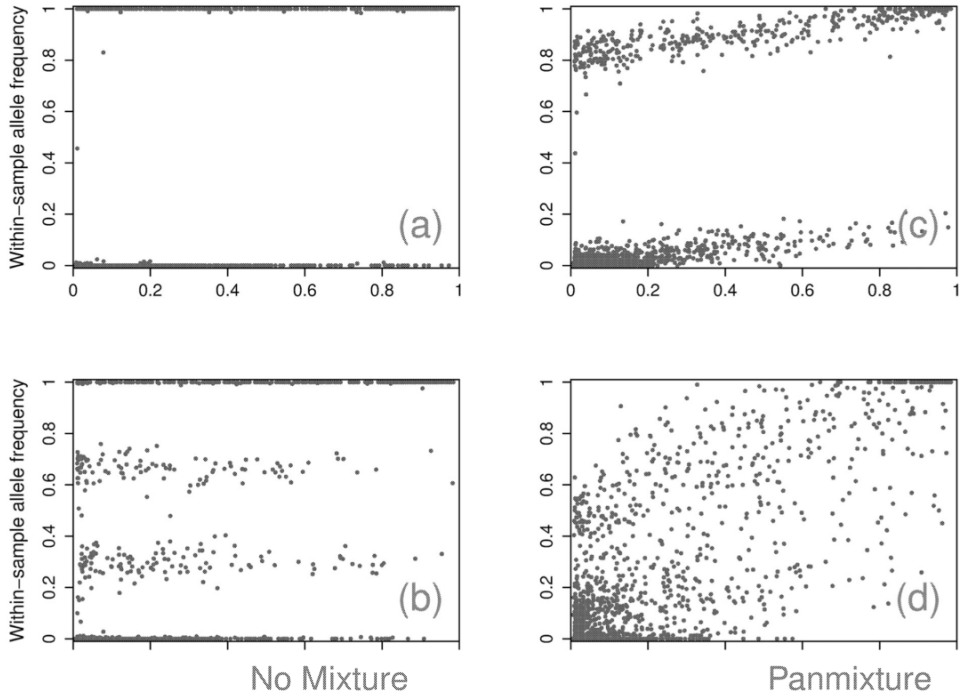
Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data

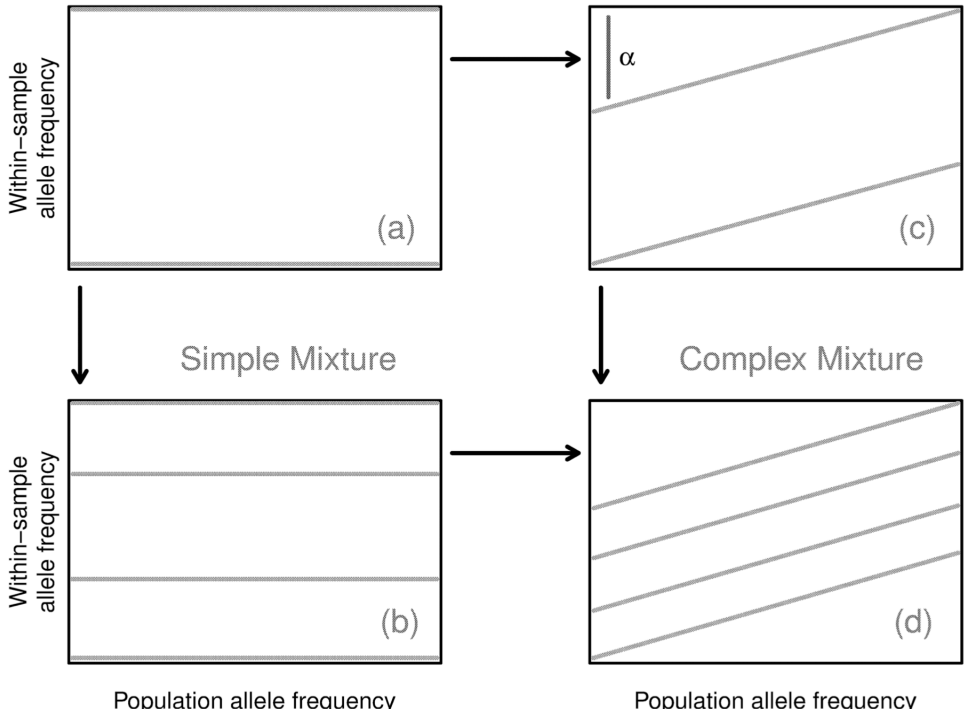
John D. O'Brien[1]*, Zamin Iqbal[2], Jason Wendler[3], Lucas Amenga-Etego[2,4]

1 Mathematics Department, Bowdoin College, Brunswick, Maine, United States of America, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, United Kingdom, **3** Pacific Northwest National Laboratory, Richland, Washington, United States of America, **4** Navrongo Health Research Centre, Navrongo, Upper East Region, Ghana

* jobrien@bowdoin.edu

PLOS COMPUTATIONAL BIOLOGY





3. A quick way to screening for mixture samples using allele frequency plots

RESEARCH ARTICLE

Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data

John D. O'Brien[1]*, Zamin Iqbal[2], Jason Wendler[3], Lucas Amenga-Etego[2,4]

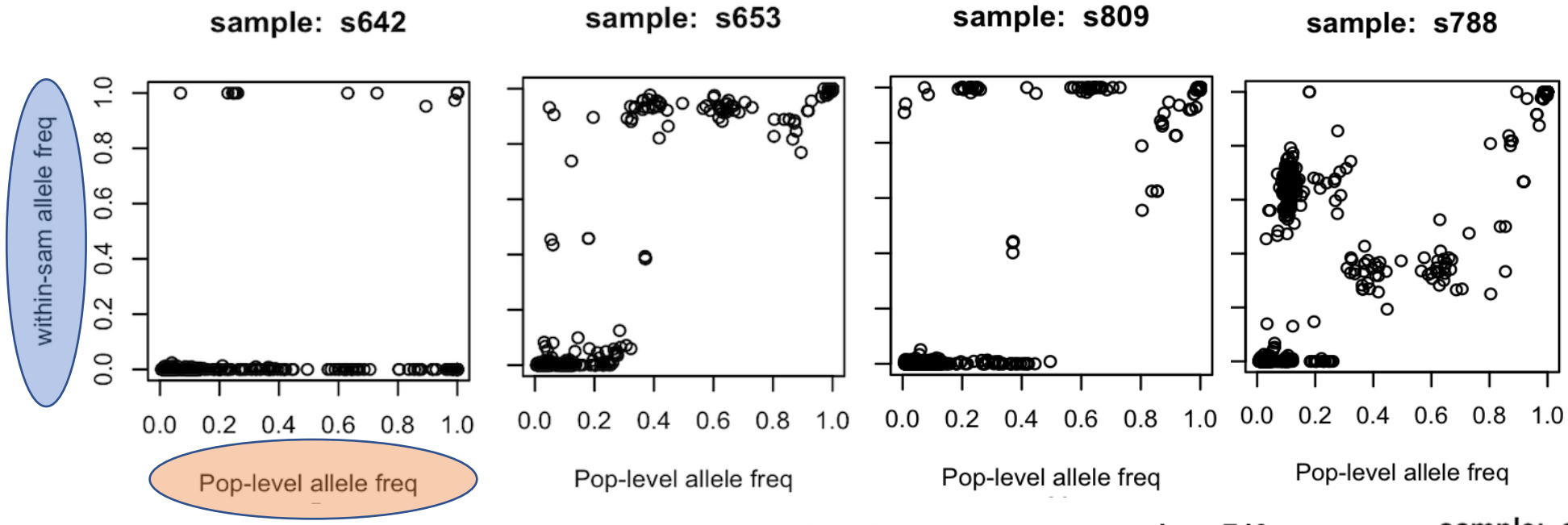
1 Mathematics Department, Bowdoin College, Brunswick, Maine, United States of America, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, United Kingdom, **3** Pacific Northwest National Laboratory, Richland, Washington, United States of America, **4** Navrongo Health Research Centre, Navrongo, Upper East Region, Ghana

$$\mathbb{P}(\mathcal{Q},\mathcal{P},\mathcal{W},\alpha,v,K|\mathcal{D}_i) \propto \mathbb{P}(\mathcal{D}_i|\mathcal{Q},\mathcal{P},v,K)\cdot\mathbb{P}(\mathcal{Q}|\mathcal{P},v,K,\mathcal{W},\alpha)\cdot \mathbb{P}(\mathcal{P})\cdot\mathbb{P}(v)\cdot\mathbb{P}(\mathcal{W}|K)\cdot\mathbb{P}(K)\cdot\mathbb{P}(\alpha).$$

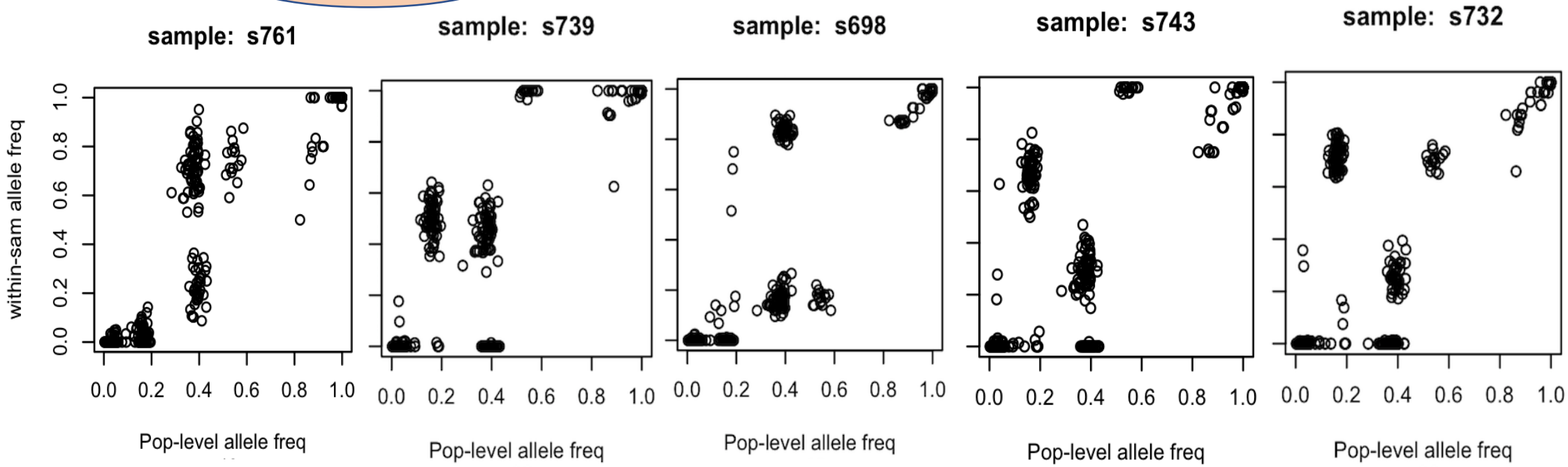
Table 1. Parameters and definitions for the model and its description.

Parameter	Definition
N	Number of samples
M	Number of SNPs
K	Number of strains
$i = 1, \cdots, N$	Index for samples
$j = 1, \cdots, M$	Index for SNPs
$r = 1, \cdots 2^K$	Index for bands / strain mixtures
p_j	(Non-reference) allele frequency for SNP j
$\mathcal{P} = [p_j]$	The PLAF for all SNPs
$\mathcal{Q} = [q_{ij}]$	Within-sample allele frequency for SNP j in sample i
α	Degree of panmixia within a sample, panmixia coefficient
$\mathcal{S} = [s_1, \cdots, s_K]$	Strains in a sample
$\mathcal{W} = [w_1, \cdots, w_K]$	Strain proportions in a sample
λ_r	Band proportions within sample
v	Variation parameter for Beta-binomial
WSAF	Within-sample allele frequency
PLAF	Population-level allele frequency

doi:10.1371/journal.pcbi


sample: s642
sample: s653
sample: s809
sample: s788
within-sam allele freq
Pop-level allele freq
0.0
0.2
0.4
0.6
0.8
1.0

S farm


sample: s761
sample: s739
sample: s698
sample: s743
sample: s732
within-sam allele freq
Pop-level allele freq
0.0
0.2
0.4
0.6
0.8
1.0

Q farm