

ENABLING MULTISCALE VARIATION ANALYSIS IN GENOME GRAPHS

SEQBIM 2020

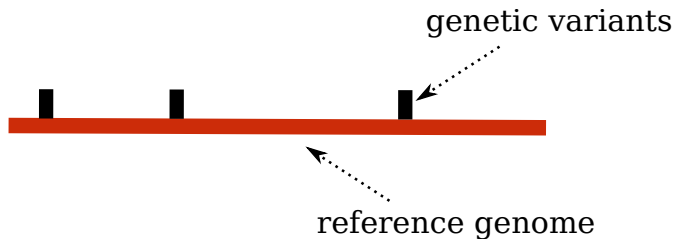
Brice Letcher

23rd November 2020

EMBL-EBI

INTRODUCTION: GENOME GRAPHS

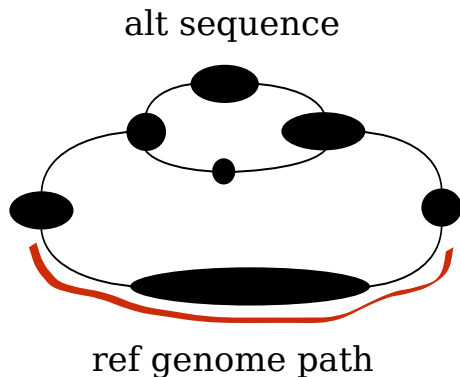
REFERENCE + DEVIATION MODEL OF GENETIC VARIATION



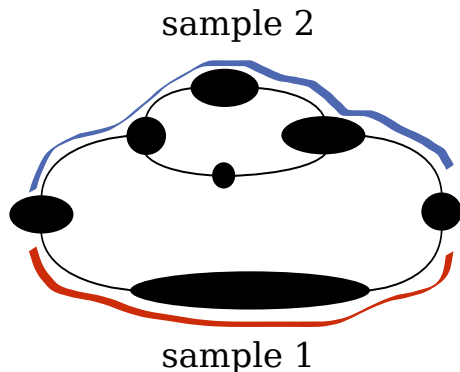
```
##fileformat=VCFv4.2
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample
```

300 Mbp of new sequence found by
assembling Illumina reads not mapping to
GRCh38 in 910 humans of African descent
(Sherman et al., 2019)



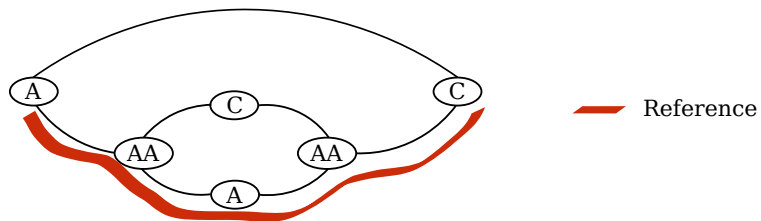
300 Mbp of new sequence found by assembling Illumina reads not mapping to GRCh38 in 910 humans of African descent (Sherman et al., 2019)



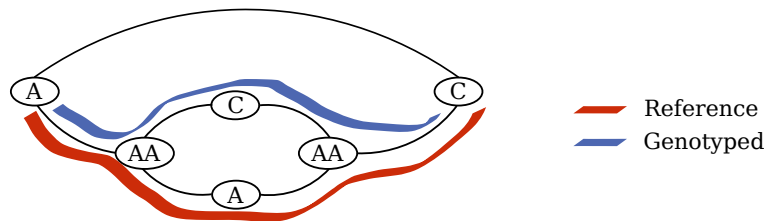
For multiple samples, sequencing reads can be mapped to the graph, genotyped and compared on the same set of variants.

MULTISCALE VARIATION

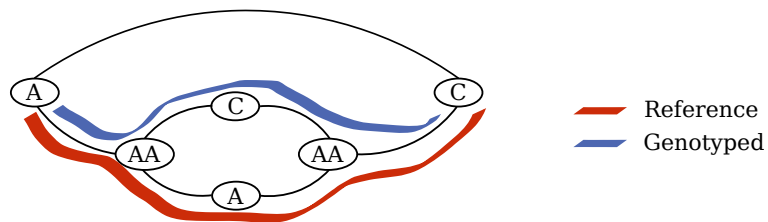
VARIATION AT MULTIPLE LENGTH SCALES



VARIATION AT MULTIPLE LENGTH SCALES

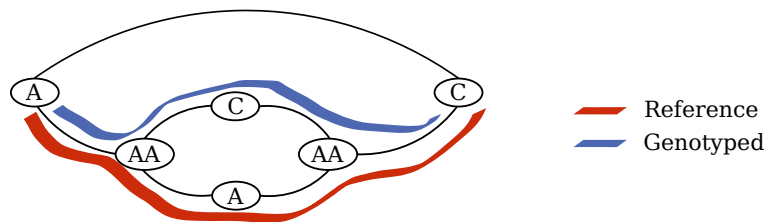


VARIATION AT MULTIPLE LENGTH SCALES



```
##fileformat=VCFv4.2
#CHROM POS REF ALT FORMAT sample
myRef 1 AAAAAA A GT 0
myRef 5 A C GT 1
```

VARIATION AT MULTIPLE LENGTH SCALES



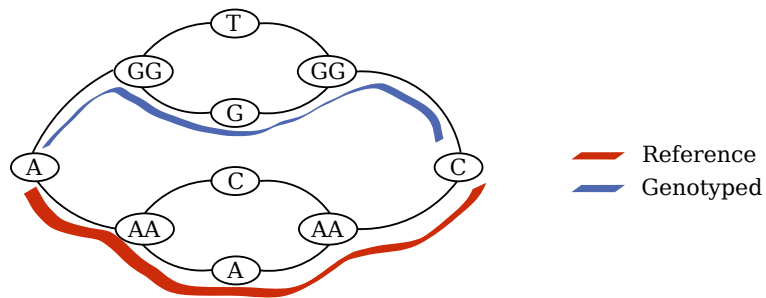
```
##fileformat=VCFv4.2
```

```
#CHROM POS REF ALT FORMAT sample
```

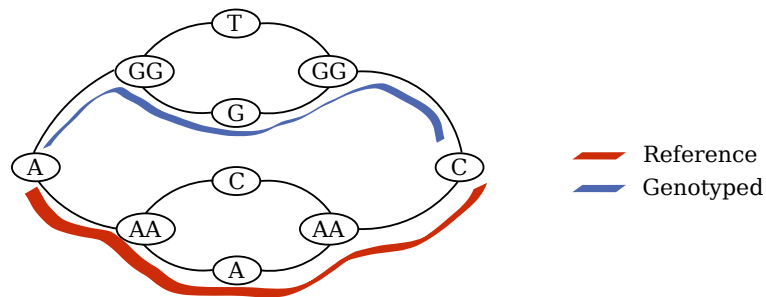
```
myRef 1 AAAAAA A GT 0 <..... Implied sequence: AAAAAA
```

```
myRef 5 A C GT 1 <..... Implied sequence: AAACAA
```

VARIATION ON MULTIPLE SEQUENCE BACKGROUNDS



VARIATION ON MULTIPLE SEQUENCE BACKGROUNDS



```
##fileformat=VCFv4.2
#CHROM POS REF ALT FORMAT sample
myRef 5 A C GT .
??? ? G T GT 0
```

- How to represent variant calls in graphs with multiscale variation?
- **How to genotype graphs with multiscale variation?**

Object Directed, acyclic genome graph G with nodes V and edges E

Object Directed, acyclic genome graph G with nodes V and edges E

Notions

1. $outdegree(v)$: number of edges $(v, w) \in E$ such that $w \in V$
2. topological ordering on V : for all paths $(v_i, \dots, v_j, \dots, v_n) \in G$ and $j \in [i, n)$, $v_j \leq v_{j+1}$

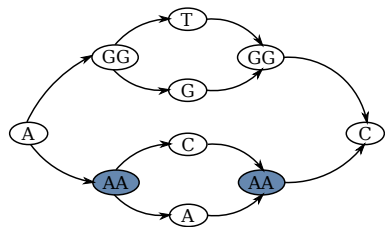
Object Directed, acyclic genome graph G with nodes V and edges E

Notions

1. $outdegree(v)$: number of edges $(v, w) \in E$ such that $w \in V$
2. topological ordering on V : for all paths $(v_i, \dots, v_j, \dots, v_n) \in G$ and $j \in [i, n)$, $v_j \leq v_{j+1}$

A **variant site** is a set of nodes

$s(v_1, v_2) = \{v \in V \mid v_1 \leq v \leq v_2\}$ for which $outdegree(v_1) > 1$ and v_2 is the first node where all paths from v_1 end.



Object Directed, acyclic genome graph G with nodes V and edges E

Notions

1. $outdegree(v)$: number of edges $(v, w) \in E$ such that $w \in V$
2. topological ordering on V : for all paths $(v_i, \dots, v_j, \dots, v_n) \in G$ and $j \in [i, n)$, $v_j \leq v_{j+1}$

A **variant site** is a set of nodes

$s(v_1, v_2) = \{v \in V | v_1 \leq v \leq v_2\}$ for which $outdegree(v_1) > 1$ and v_2 is the first node where all paths from v_1 end.

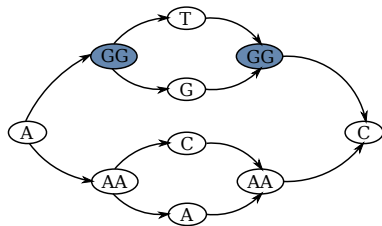
Object Directed, acyclic genome graph G with nodes V and edges E

Notions

1. $outdegree(v)$: number of edges $(v, w) \in E$ such that $w \in V$
2. topological ordering on V : for all paths $(v_i, \dots, v_j, \dots, v_n) \in G$ and $j \in [i, n)$, $v_j \leq v_{j+1}$

A **variant site** is a set of nodes

$s(v_1, v_2) = \{v \in V \mid v_1 \leq v \leq v_2\}$ for which $outdegree(v_1) > 1$ and v_2 is the first node where all paths from v_1 end.



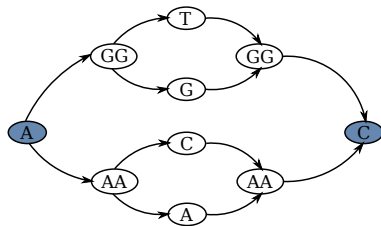
Object Directed, acyclic genome graph G with nodes V and edges E

Notions

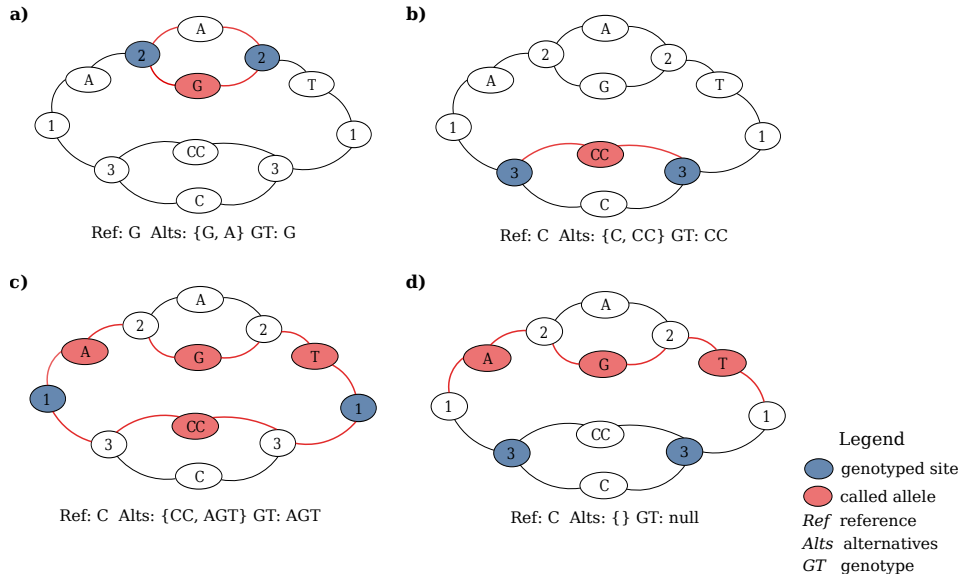
1. $outdegree(v)$: number of edges $(v, w) \in E$ such that $w \in V$
2. topological ordering on V : for all paths $(v_i, \dots, v_j, \dots, v_n) \in G$ and $j \in [i, n)$, $v_j \leq v_{j+1}$

A **variant site** is a set of nodes

$s(v_1, v_2) = \{v \in V \mid v_1 \leq v \leq v_2\}$ for which $outdegree(v_1) > 1$ and v_2 is the first node where all paths from v_1 end.

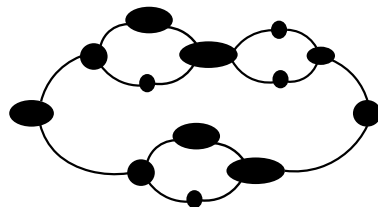
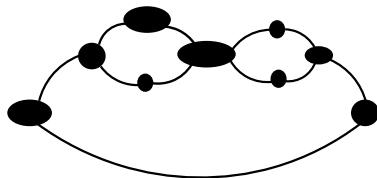


ALGORITHM FOR GENOTYPING MULTISCALE VARIATION



APPLICATIONS OF MULTISCALE GENOTYPING

DATASETS EXHIBITING MULTISCALE VARIATION



Organism

M. tuberculosis

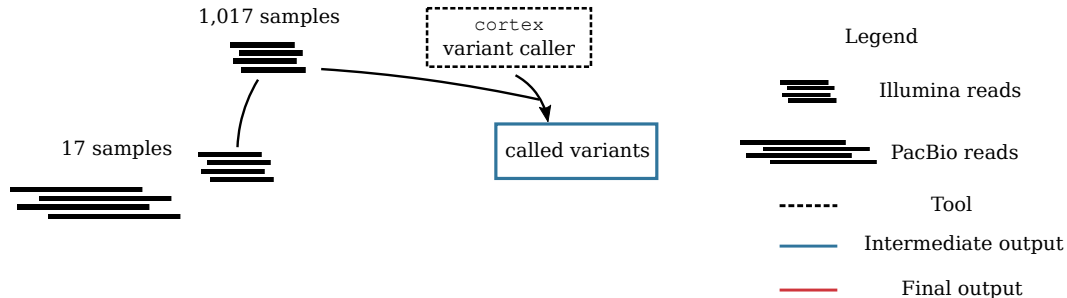
P. falciparum

Context

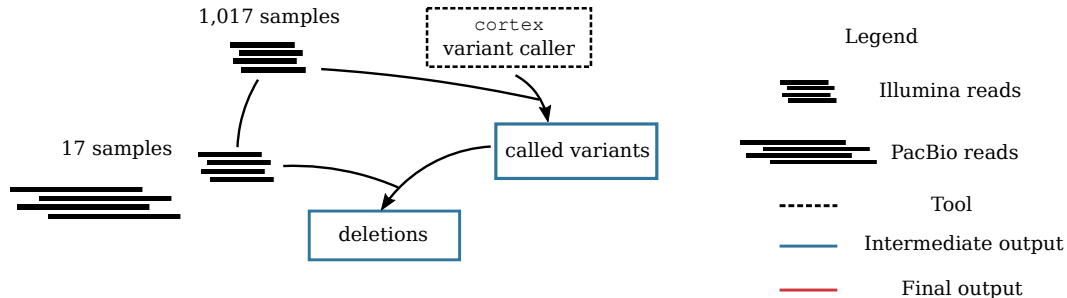
Map genetic determinants
of drug resistance

Map candidate vaccine
gene diversity

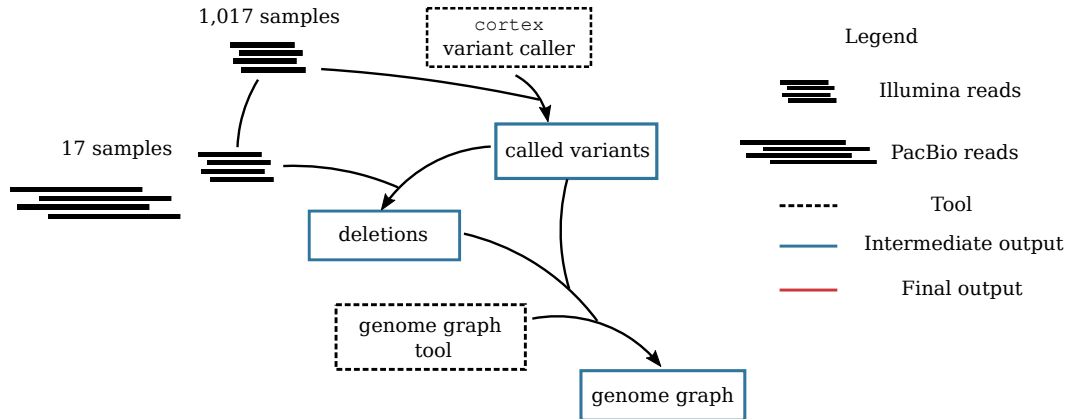
LARGE DELETIONS AND OVERLAPPING VARIATION IN *M. tuberculosis*



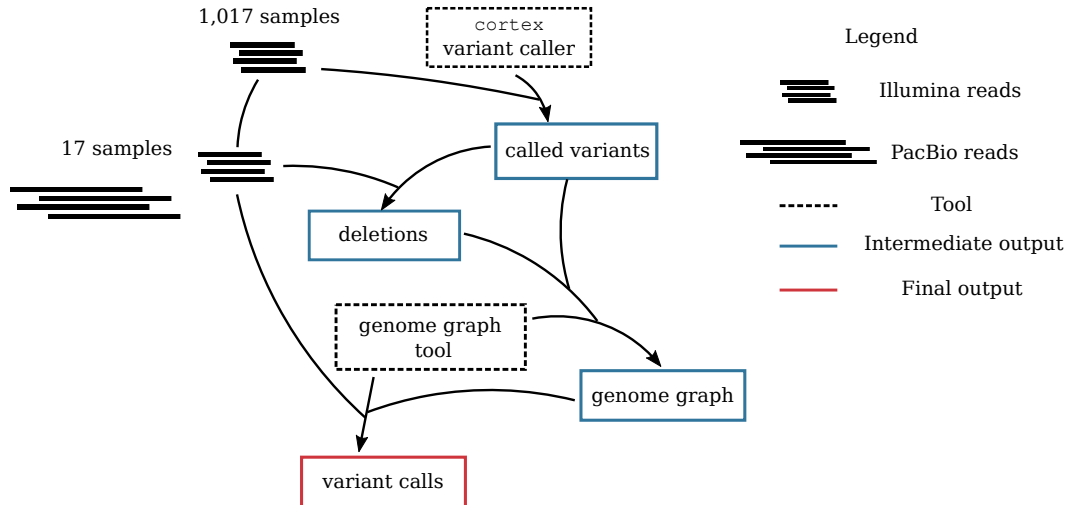
LARGE DELETIONS AND OVERLAPPING VARIATION IN *M. tuberculosis*



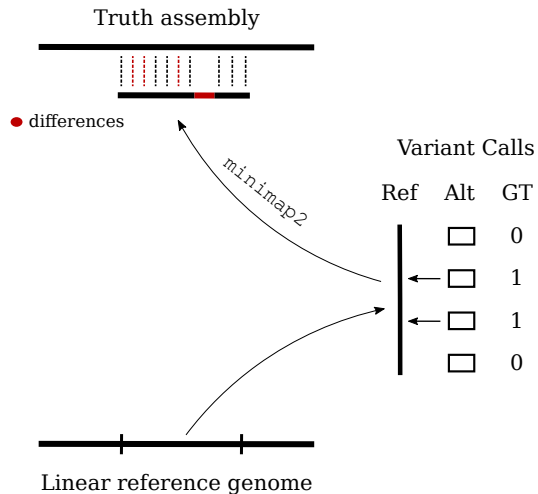
LARGE DELETIONS AND OVERLAPPING VARIATION IN *M. tuberculosis*



LARGE DELETIONS AND OVERLAPPING VARIATION IN *M. tuberculosis*

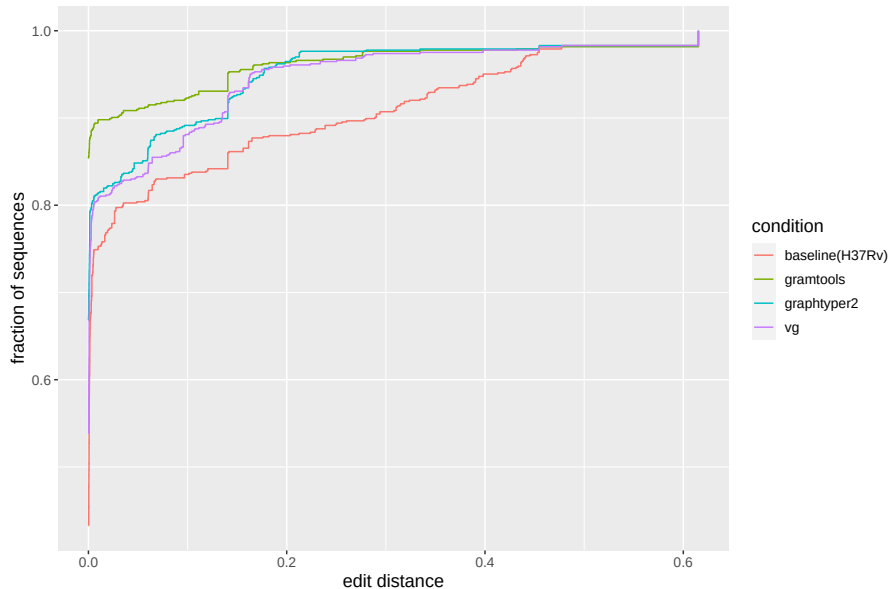


MEASURE OF GENOTYPING PERFORMANCE

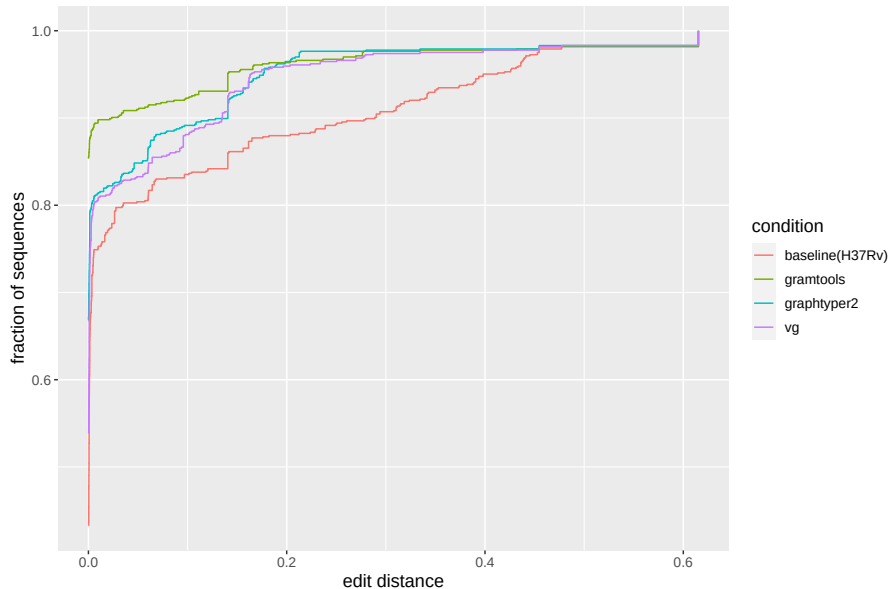


Tools:

gramtools,
vg (Garrison
et al., 2018),
graph typer2
(Eggertsson
et al., 2019)



Tools:
gramtools,
vg (Garrison
et al., 2018),
graphtyper2
(Eggertsson
et al., 2019)



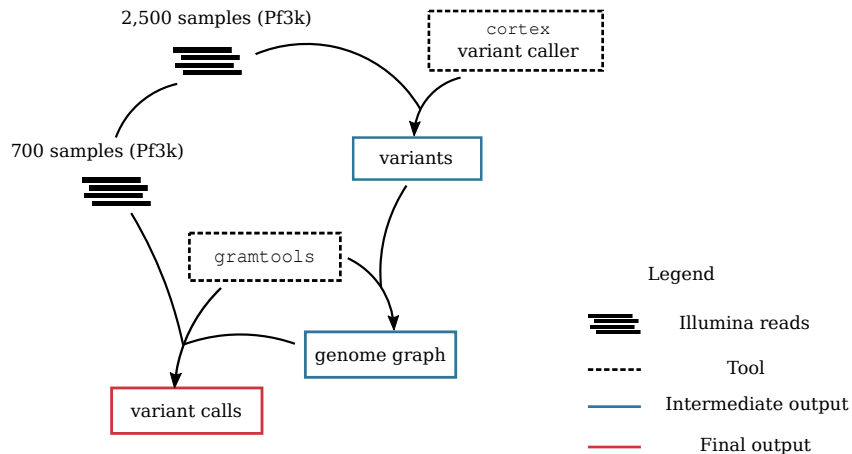
Average
distance:
gramtools
(1.37%),
graphtyper2
(2.09%),
vg(2.41%),
baseline_ref
(4.78%)



Genotype *P. falciparum* surface antigen DBLMSP2

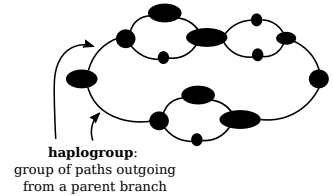
Has two genetically diverged forms, likely due to immune-system driven balancing selection (Amambua-Ngwa et al., 2012)

DIMORPHIC SURFACE ANTIGEN IN *P. falciparum*

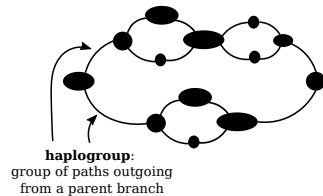
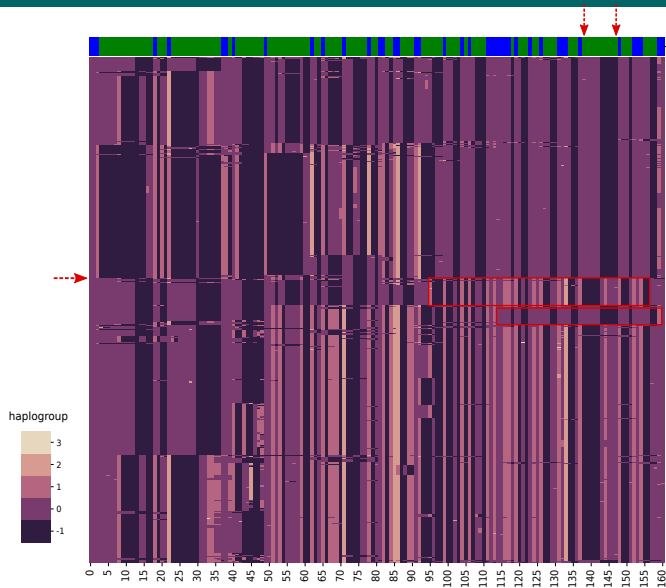


Pf3k: Pearson et al., 2019

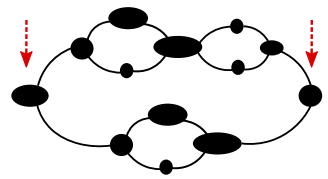
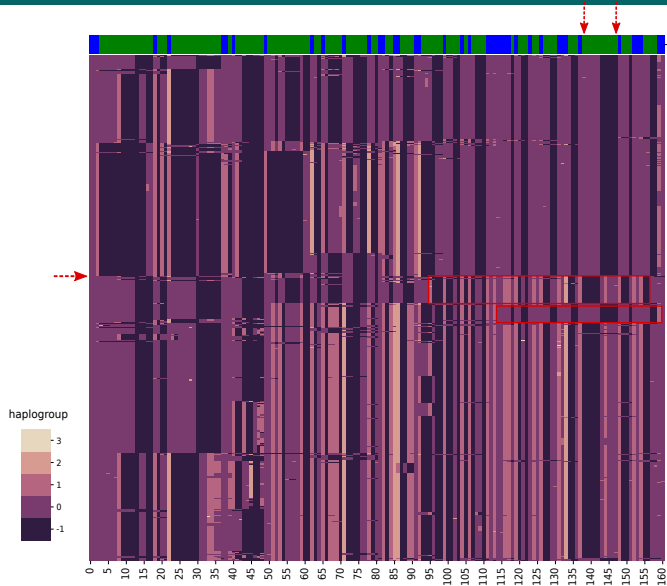
GENOTYPING THE TWO FORMS OF DBLMSP2



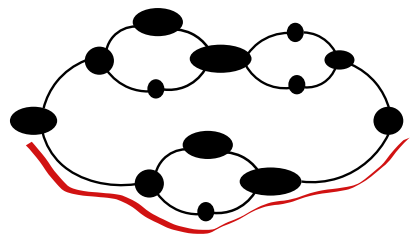
GENOTYPING THE TWO FORMS OF DBLMSP2



GENOTYPING THE TWO FORMS OF DBLMSP2

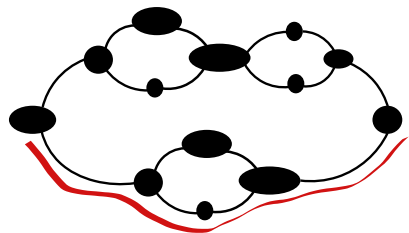


A NEW FORMAT FOR VARIANT CALLS IN GENOME GRAPHS



linear reference

how do we record
relationships between
variant sites?



linear reference

how do we record
relationships between
variant sites?

JSON-encoded extension
to VCF for recording
variant calls in genome
graphs.

Prototype specification:

[https://github.com/
iqbal-lab-org/
jVCF-spec](https://github.com/iqbal-lab-org/jVCF-spec)

ACKNOWLEDGEMENTS

SeqBIM

EMBL-EBI 

Martin Hunt
Zamin Iqbal

<https://www.ebi.ac.uk/research/iqbal/members>



bricoletc



@bricoletc

bletcher@ebi.ac.uk

10/2018



2020

2021

2022

Mid-2022



REFERENCES



Amambua-Ngwa, Alfred et al. (Nov. 2012). “Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites”. en. In: *PLOS Genetics* 8.11, e1002992. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002992.



Eggertsson, Hannes P. et al. (Nov. 2019). “GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs”. en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 5402. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13341-9.



Garrison, Erik et al. (Aug. 2018). “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. en. In: *Nature Biotechnology* 36.9, pp. 875–879. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4227.



Li, Heng, Xiaowen Feng, and Chong Chu (Dec. 2020). “The design and construction of reference pangenome graphs with minigraph”. en. In: *Genome Biology* 21.1, p. 265. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02168-z.



Otto, Thomas D. et al. (May 2018). “Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres”. In: *Wellcome Open Research* 3. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.14571.1.



Pearson, Richard D. et al. (Nov. 2019). “An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples”. en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 824730. DOI: 10.1101/824730.



Sherman, Rachel M. et al. (Jan. 2019). “Assembly of a pan-genome from deep sequencing of 910 humans of African descent”. en. In: *Nature Genetics* 51.1, pp. 30–35. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0273-y.

jVCF format

TB benchmark

DBLMSP2 genotyping

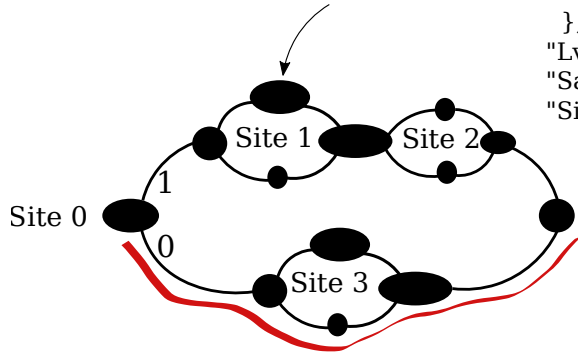
Graph construction

gramtools mapping and genotyping

GRAPHICAL ILLUSTRATION OF JVCF

```
{ "ALS":  
  [ "CTGATGTTAAT" ],  
  "GT": [ [ 0 ] ],  
  "HAPG": [ [ 0 ] ],  
  "POS": 11826,  
  "SEG": "Pf3D7_10_v3" }
```

```
{ "Child_Map":  
  {  
    "0":  
      { "1": [1, 2], "0": [3] }  
  },  
  "Lvl1_Sites": [0],  
  "Samples": [...],  
  "Sites": [...] } }
```



linear reference

RGFA vs JVCF

rGFA describes genome graphs
with stable alt. references



```
S s1 CTGAA SN:Z:chr1 SO:i:0 SR:i:0
S s2 ACG SN:Z:chr1 SO:i:5 SR:i:0
S s3 TGGC SN:Z:chr1 SO:i:8 SR:i:0
S s4 TGTGA SN:Z:chr1 SO:i:12 SR:i:0
S s5 TTTC SN:Z:foo SO:i:8 SR:i:1
S s6 CTGA SN:Z:foo SO:i:12 SR:i:1
S s7 GTTAC SN:Z:bar SO:i:5 SR:i:2
L s1 + s2 + 0M SR:i:0
L s2 + s3 + 0M SR:i:0
L s3 + s4 + 0M SR:i:0
L s2 + s5 + 0M SR:i:1
L s5 + s6 + 0M SR:i:1
L s6 + s4 + 0M SR:i:1
L s1 + s7 - 0M SR:i:2
L s7 - s6 + 0M SR:i:2
```

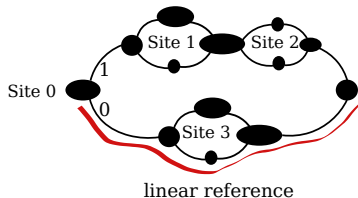
VCF describes variant calls
on linear references

Ref	Alt	GT
-----	-----	----

	☐	0
←	☐	1
←	☐	1
	☐	0

jVCF describes
variant calls on genome graphs

```
{ "Child_Map":
{
"0":
{"foo": [1, 2], "chr1": [3]}
},
"Lvl1_Sites: [0],
"Samples": [...],
"Sites": [...] }
```

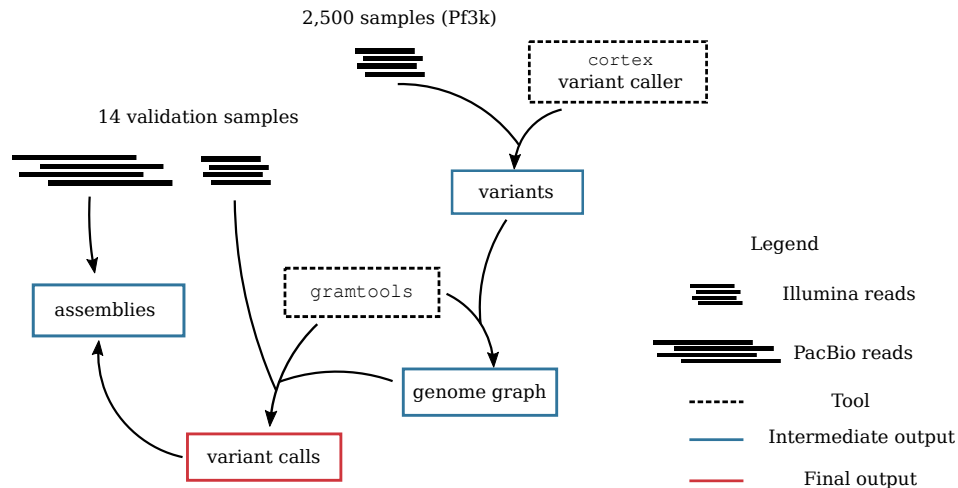


COMPUTATIONAL PERFORMANCE ON *M. tuberculosis* DATASET

	Index			Genotype	
	Disk(Mb)	Max RAM(Mb)	Speed(sec)	Max RAM(Mb)	Speed(reads/sec)
vg	29	697	171	605(158)	3,869
gramtools	153	411	29	446	28,382
graphtyper2	-	-	-	869(88)	8,024

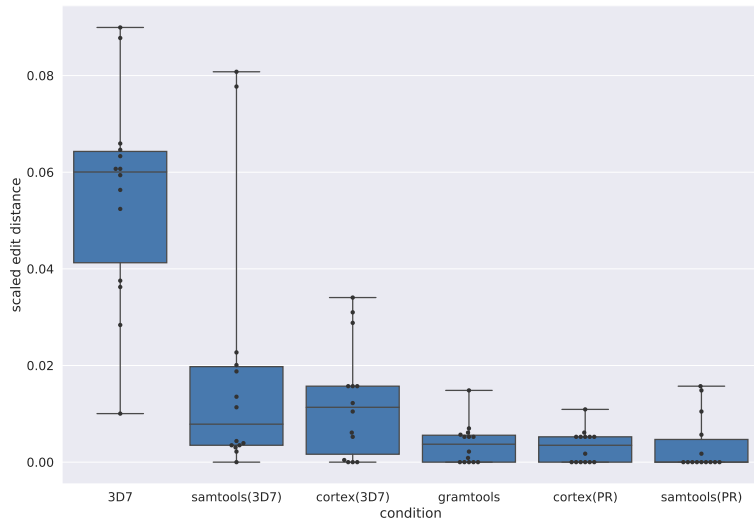
Table 1: *Index:* vg genome graph built from VCF and pruned to reduce graph complexity (else temporary disk use exceeded 500 Gb before completion. *graphtyper2* runs from a VCF and has no separate indexing operation. *Genotype:* Speed shows the average number of reads mapped across the 17 samples (10.7 million) divided by the average CPU time. vg and graphtyper2 have separate read mapping and genotyping steps: for speed, CPU time is summed, and for RAM, mapping is shown followed by genotyping in brackets. *graphtyper2* requires an input file of reads mapped to a linear reference genome; mapping RAM and speed is shown for `bwa mem` with default parameters. *metrics:* Mb: Megabytes; sec: total CPU seconds (accounts for multi-threading. 10 threads used for genotyping in each tool).

SETUP FOR VALIDATING DBLMSP2 GENOTYPING

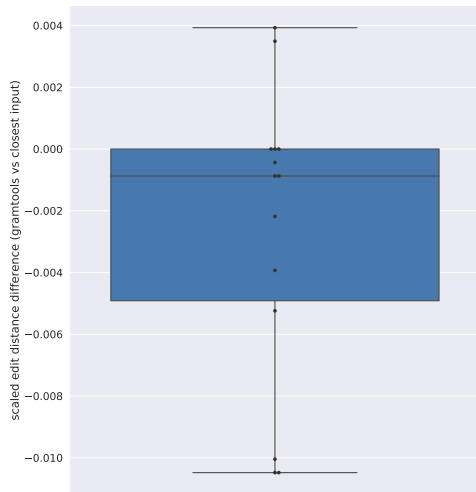


Assemblies: Otto et al., 2018

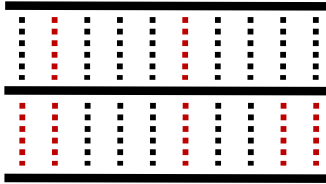
gramtools GENOTYPING COMPARED TO REFERENCE-BASED VARIANT CALLING



gramtools FINDS RECOMBINANTS BETWEEN INPUT HAPLOTYPES IN THE GRAPH

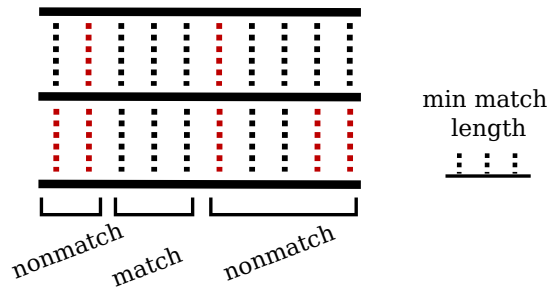


RECURSIVE COLLAPSE AND CLUSTER (RCC) FOR BUILDING GENOME GRAPHS



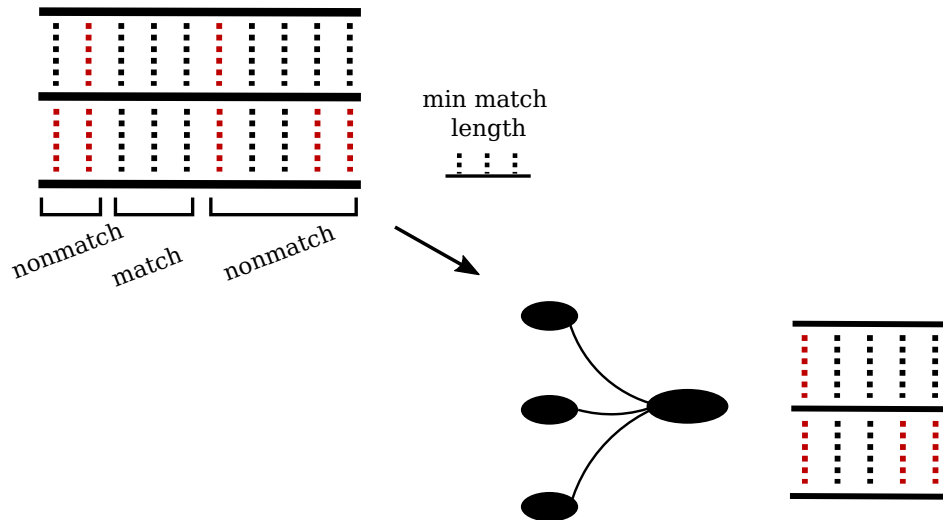
Tool: `make_prg`

RECURSIVE COLLAPSE AND CLUSTER (RCC) FOR BUILDING GENOME GRAPHS



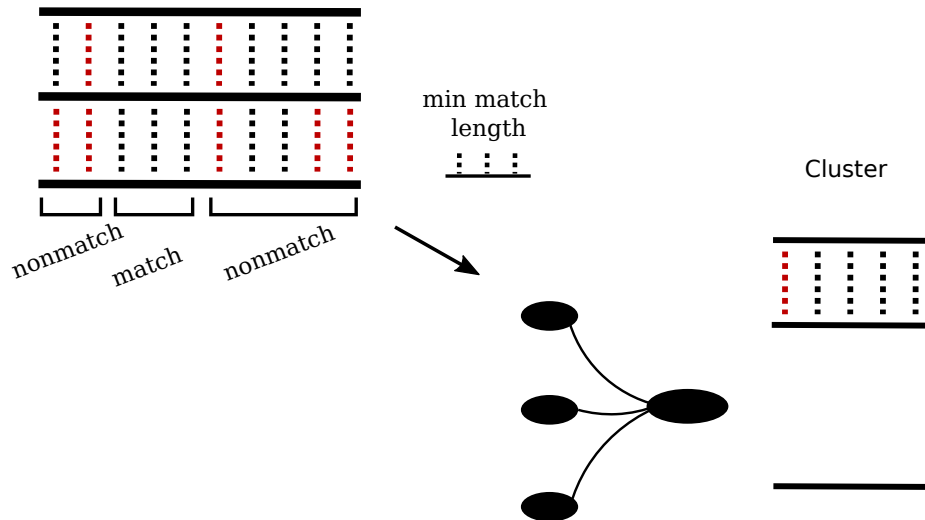
Tool: make_prg

RECURSIVE COLLAPSE AND CLUSTER (RCC) FOR BUILDING GENOME GRAPHS



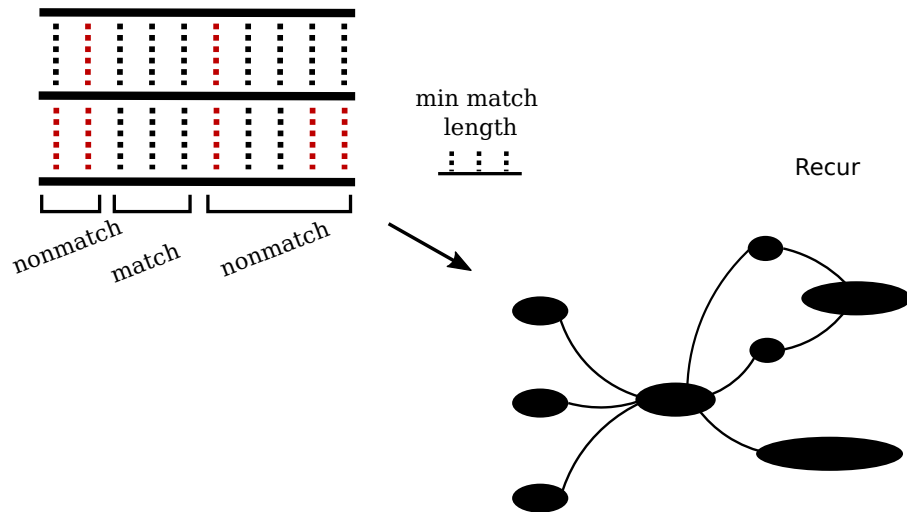
Tool: make_prg

RECURSIVE COLLAPSE AND CLUSTER (RCC) FOR BUILDING GENOME GRAPHS



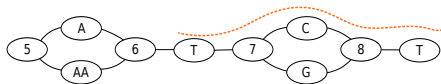
Tool: make_prg

RECURSIVE COLLAPSE AND CLUSTER (RCC) FOR BUILDING GENOME GRAPHS



Tool: make_prg

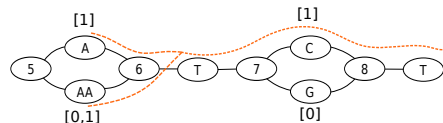
VBWT-BASED READ MAPPING IN gramtools



mapping

BWT	SA	suffix	T	CT	TCT
T	13	0			
5	1	A A 6 A 6 T 7 C 8 G 8 T 0			
A	2	A 6 A 6 T 7 C 8 G 8 T 0			
6	4	A 6 T 7 C 8 G 8 T 0			
7	8	C 8 G 8 T 0			
8	10	G 8 T 0			
8	12	T 0			
6	6	T 7 C 8 G 8 T 0			
0	0	5 A A 6 A 6 T 7 C 8 G 8 T 0			
A	3	6 A 6 T 7 C 8 G 8 T 0			
A	5	6 T 7 C 8 G 8 T 0			
T	7	7 C 8 G 8 T 0			
C	9	8 G 8 T 0			
G	11	8 T 0			

Diagram illustrating the mapping process. A vertical line represents the BWT. Green boxes mark positions in the T, CT, and TCT columns. Red boxes mark positions in the CT and TCT columns. Arrows indicate the mapping from the BWT to the CT and TCT columns.



[{{0,1}: 1}, {{1}: 1}]
Site 5 Site 7

GENOTYPING MODEL IN gramtools

For allele a in candidate set A ,

$$\mathcal{L}(a) = p(c(a))p(i(a))g(a)$$

$P(A)$ powerset of A , c_X
coverage count of set X

$$c(a) = \sum_{X \in P(A): X \ni a} c_X$$

Poisson if $\sigma \leq \lambda$
else Negative Binomial

$$p(c(a) = k | \lambda) = \exp(-\lambda) \frac{\lambda^k}{k!}$$

$$p(c(a) = k | p, r) = \binom{k+r-1}{k} (1-p)^r p^k$$

$$r = \frac{\lambda^2}{\sigma - \lambda} \quad p = \frac{\lambda}{\lambda + r}$$

$$g(a) = p(c(a) = 0)^g (1 - p(c(a) = 0))^{l_a - g}$$

Divide exponents by l_a

$$i(a) = \sum_{X \in P(A): a \notin X} c_X$$

Sequencing error rate ϵ

$$p(i(a) = k | \epsilon) = \epsilon^k$$

Call max. likelihood
allele a' , with confidence

$$\min_{a \in A: a \neq a'} \frac{\mathcal{L}(a')}{\mathcal{L}(a)}$$