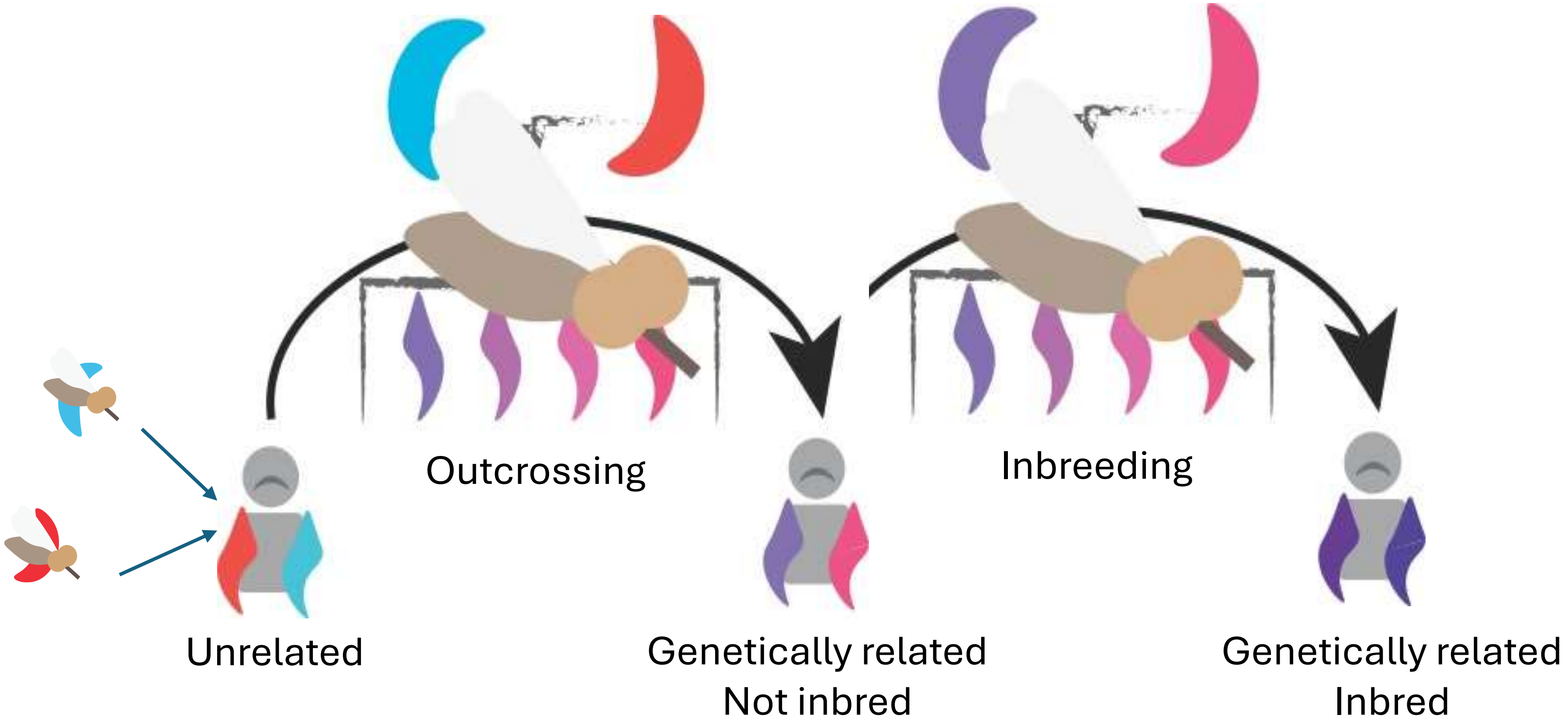
MalKinID: A likelihood-based classification model for identifying malaria parasite genealogical relationships to study transmission



End-to-end genetic surveillance for enabling public health decision-making

Integrate genetics, epidemiology, mapping, and modeling

- Create countrywide genetic epidemiology map and use routine data for ongoing surveillance
- Use models to integrate genetic and epidemiology data

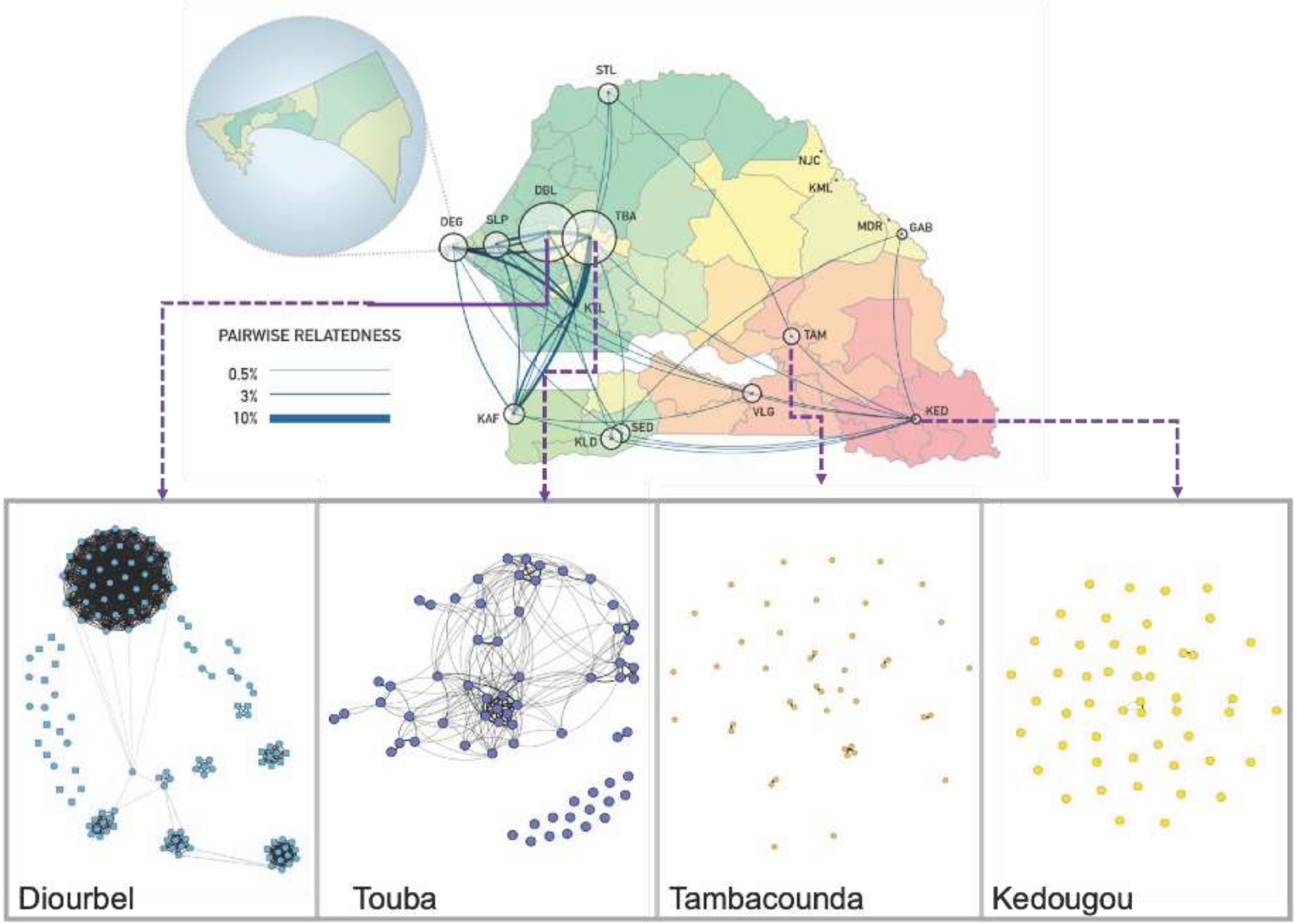
Leverage genetic data to inform decision-making

- Study transmission patterns, identify metrics for intervention placement and assessment
- Assess current and emerging drug resistance risk


Sarah Volkman

Daouda Ndiaye

Senegal reveals structural changes in genetic relatedness networks associated with declining transmission



Observation 1: Parasites in **high transmission regions** are unrelated

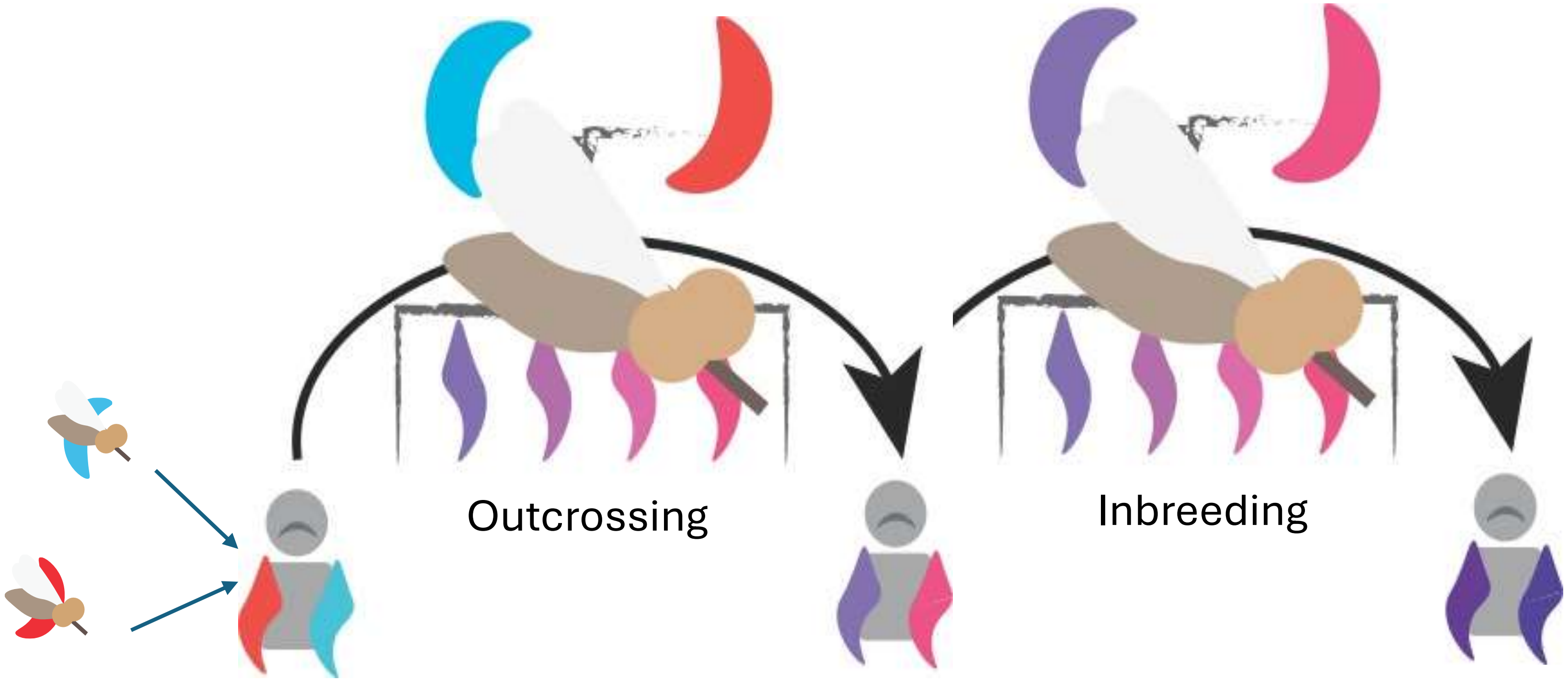
Observation 2: Parasites in **low transmission regions** more likely to be genetically related

Observation 3: **Low transmission regions** with similar incidences (Diourbel, Touba) can **have different relatedness structure**

Schaffner et al 2023

Transmission influences parasite relatedness

Can we use relatedness to reconstruct transmission chains and determine where parasites are coming from?



Outcrossing

Inbreeding

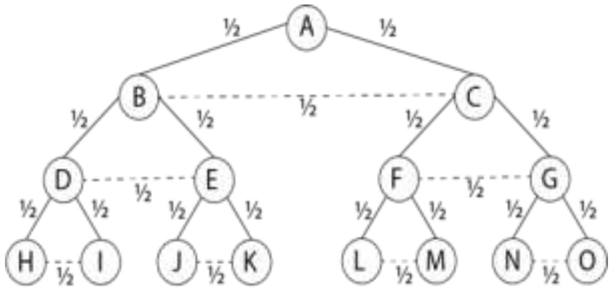
Unrelated

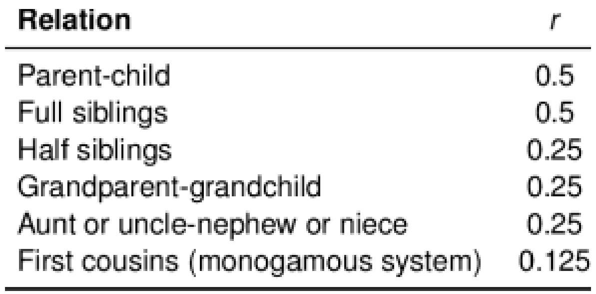
Genetically related
Not inbred

Genetically related
Inbred

MalKinID: Reconstructing transmission history based on patterns of genetic relatedness

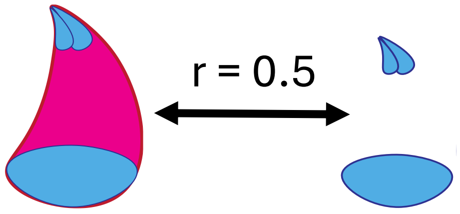
Relatedness reflects genealogy



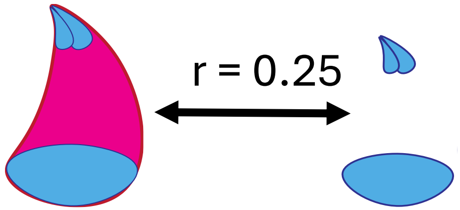


Relation	*r*
Parent-child	0.5
Full siblings	0.5
Half siblings	0.25
Grandparent-grandchild	0.25
Aunt or uncle-nephew or niece	0.25
First cousins (monogamous system)	0.125

Source: adapted from Bergstrom (1996).

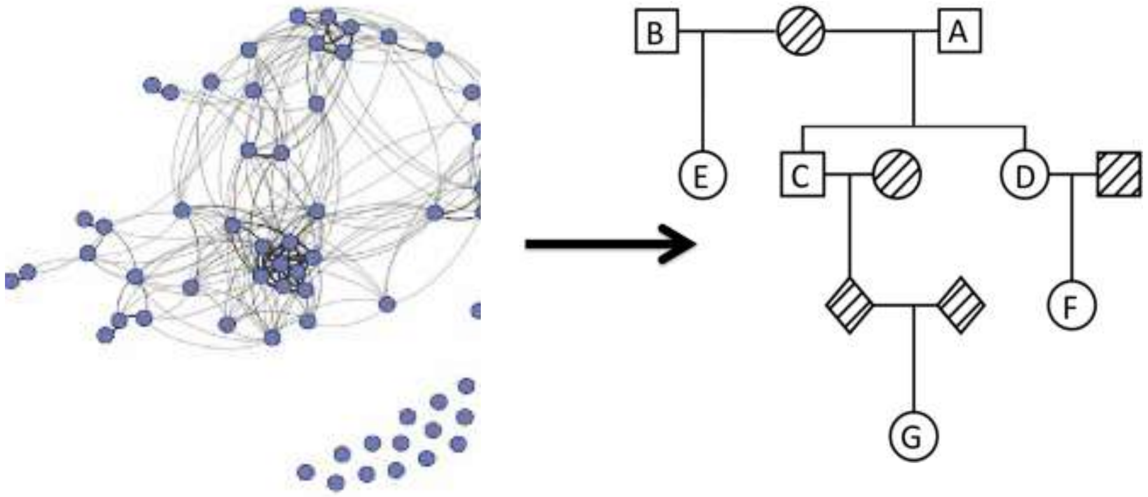


Likely a Parent-Child
Or a full-sibling



Likely a half-sibling
Or avuncular relationship

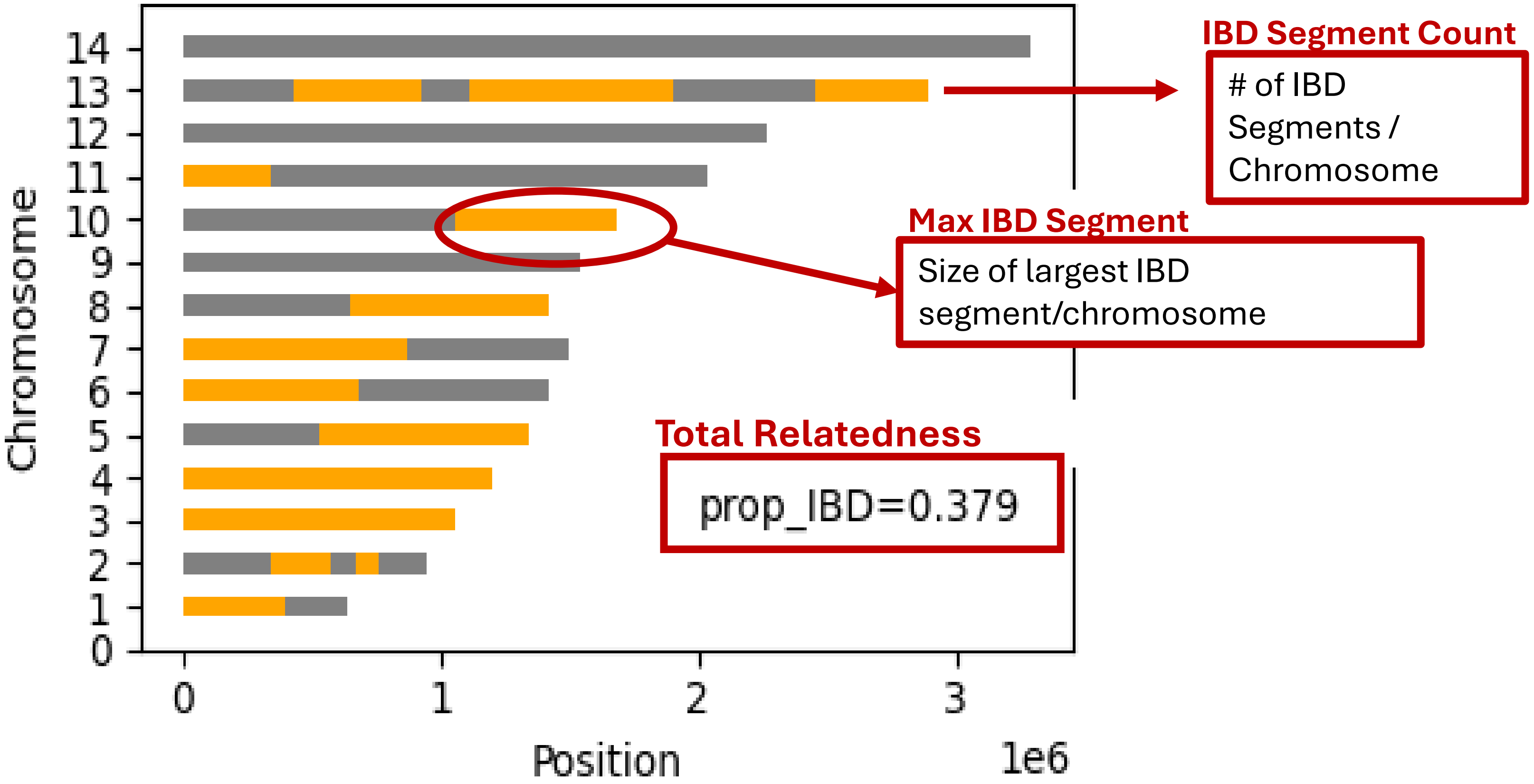
Pedigrees can allow us to identify **parent-offspring pairs**



Parent-offspring pairs could be most informative for reconstructing transmission history (because parents transmit offspring)

What can we use to determine genealogical history?

$$L(G\ |Data\ ) = \underbrace{P_G(r_{total} = r)}_{\text{Total Relatedness}} * \prod_{i=1}^{14}[\underbrace{P_G(IBD_{max,chrom=i} = ibd_{max,i})}_{\text{Max IBD Segment}}\underbrace{P_G(N_{segments,i} = n_{segments,i})}_{\text{IBD Segment Count}}]$$



IBD Segment Count

# of IBD Segments / Chromosome

Max IBD Segment

Size of largest IBD segment/chromosome

Total Relatedness

prop_IBD=0.379

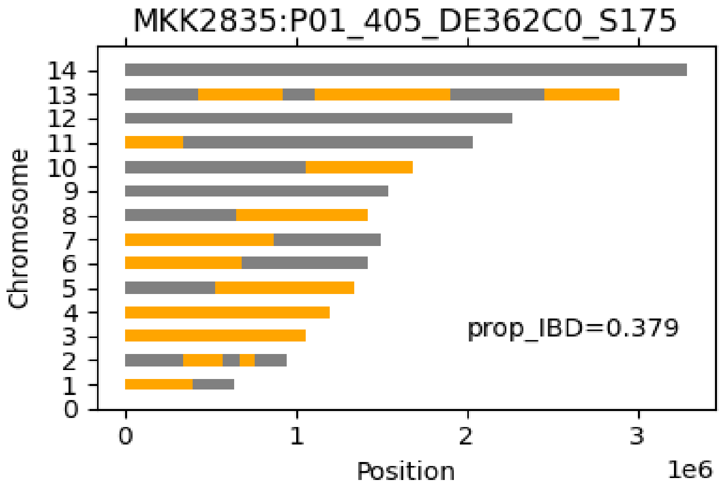
Meiosis model Recalibration

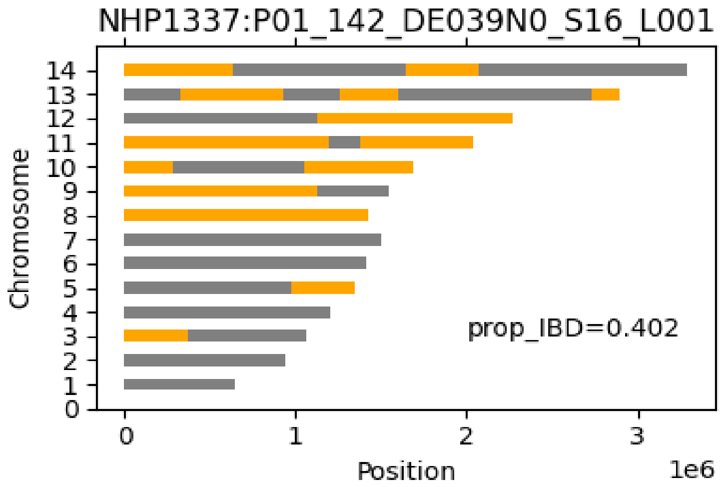
NHP4026 x NF54HT-GFP–luc, MKK2835 × NHP1337, Mal31 X Kh004

Crosses generated by Ferdig, Anderson, and Cheeseman groups

Three crosses:

1. NHP4026 and NF54HT-GFP–luc
    - Lab adapted X SE Asian
    - 174 PC, 3845 Sibs
2. MKK2835 × NHP1337
    - SE Asian strain X SE Asian
    - 66 PC, 564 Sibs
3. Mal31 X Kh004
    - Africa X SE Asia
    - 175 PC, 4073 Sibs





Modified IBD HMM probabilities

$$P(Concordance|IBD) = (1-\varepsilon)^2 + (\varepsilon)^2$$

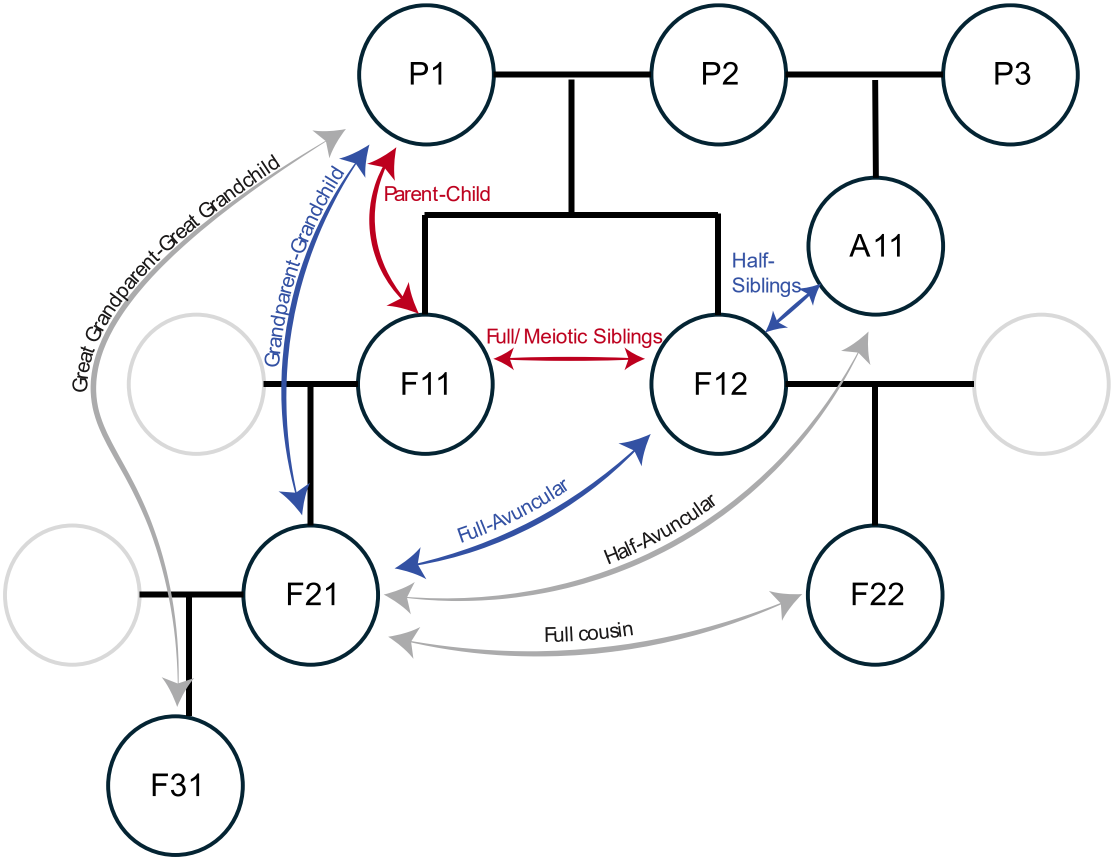
$$P(Concordance|non-IBD) = 2\varepsilon(1-\varepsilon)$$

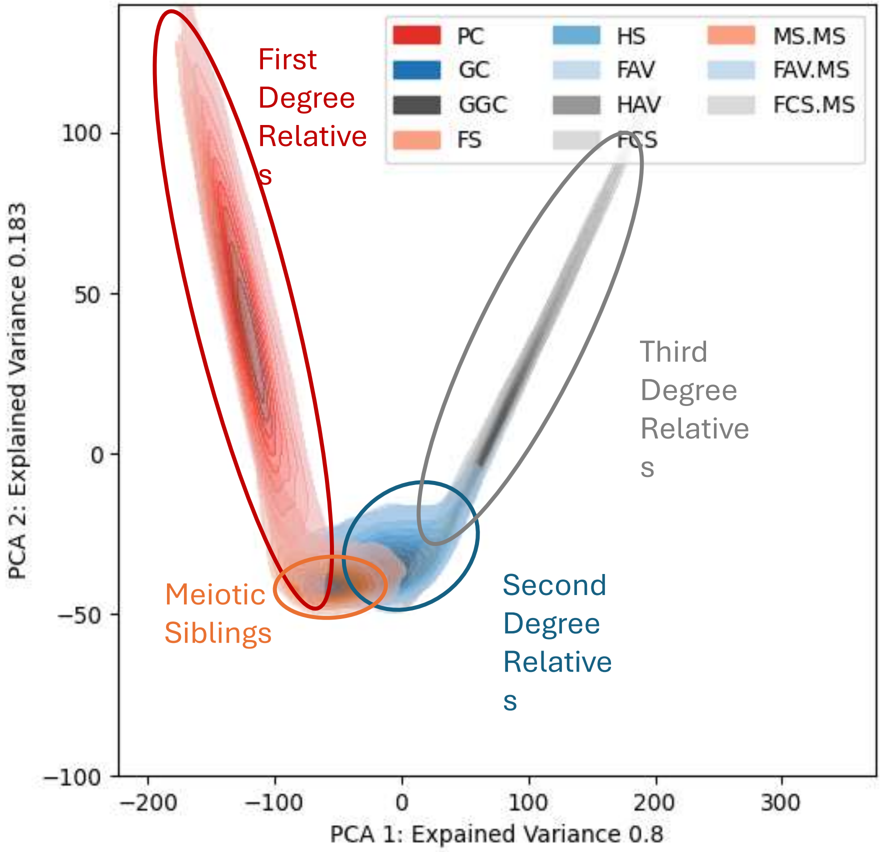
$$P(Discordance|IBD) = 2\varepsilon(1-\varepsilon)$$

$$P(Discordance|non-IBD) = 1 - 2\varepsilon(1-\varepsilon)$$

WGS reduced to set of variants that differ between the two parental strains

MalKinID: Accurately quantifies relationships during outcrossing

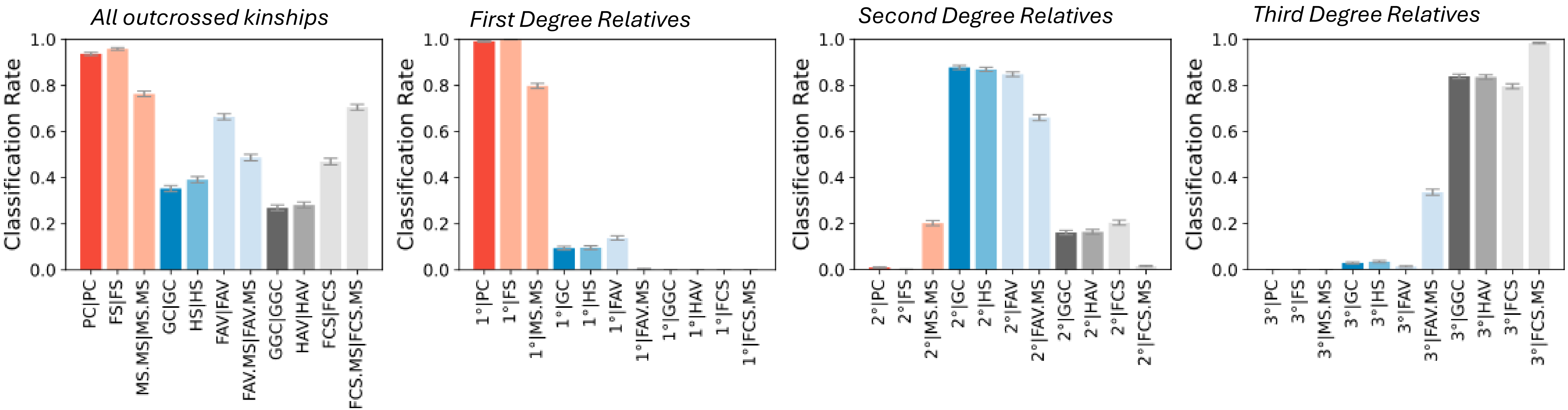




First, **second**, and **third** degree relatives occupy distinct niches in mathematical parameter space

MalKinID: Accurately identifies PC relationships following outcrossed conditions

Simulated Data



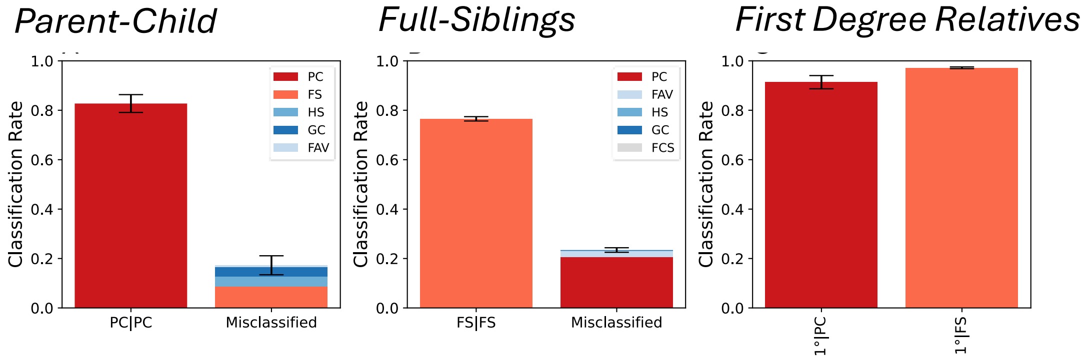
All outcrossed kinships

First Degree Relatives

Second Degree Relatives

Third Degree Relatives

Empirical Lab Cross Data



Parent-Child

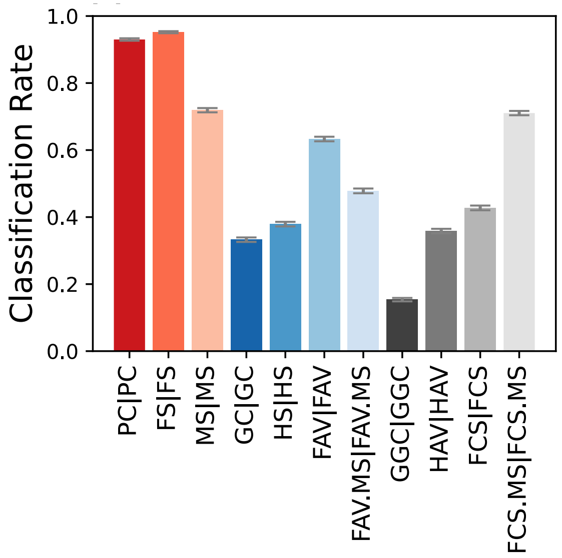
Full-Siblings

First Degree Relatives

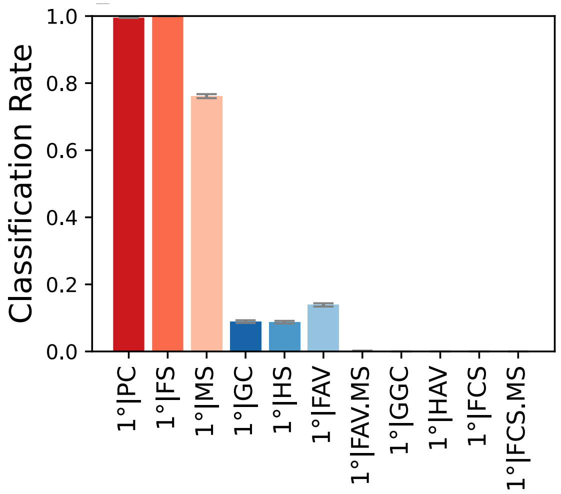
For outcrossed relatives, *MalKinID* can distinguish :

1) between first, second, and third degree relatives
2) Within first degree relatives

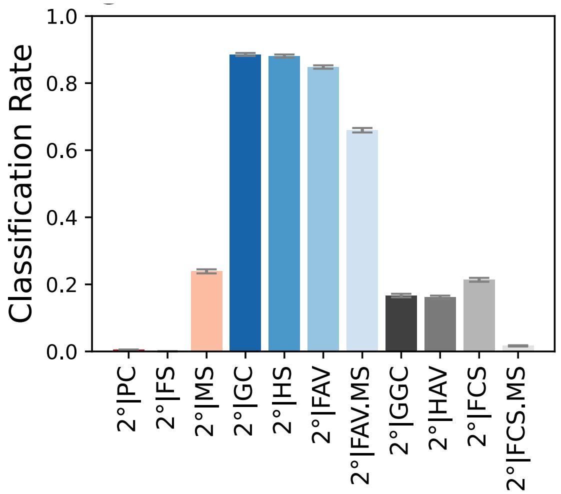
Likelihood approach performs equivalently to machine-learning based neural network



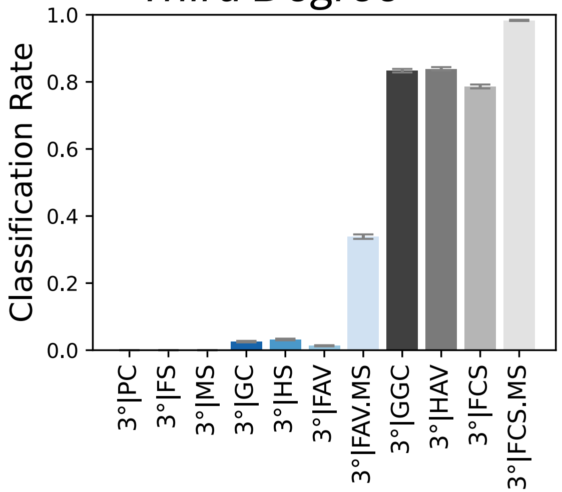
All Genealogies



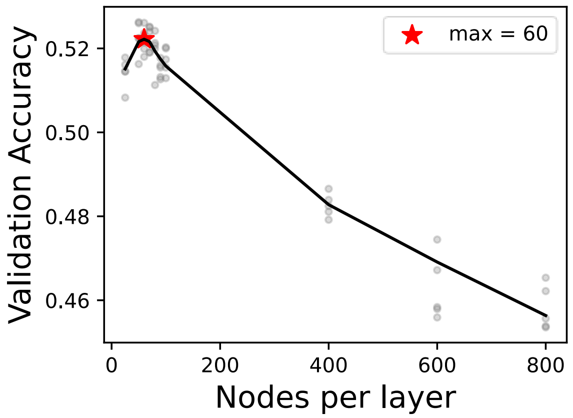
First Degree



Second Degree



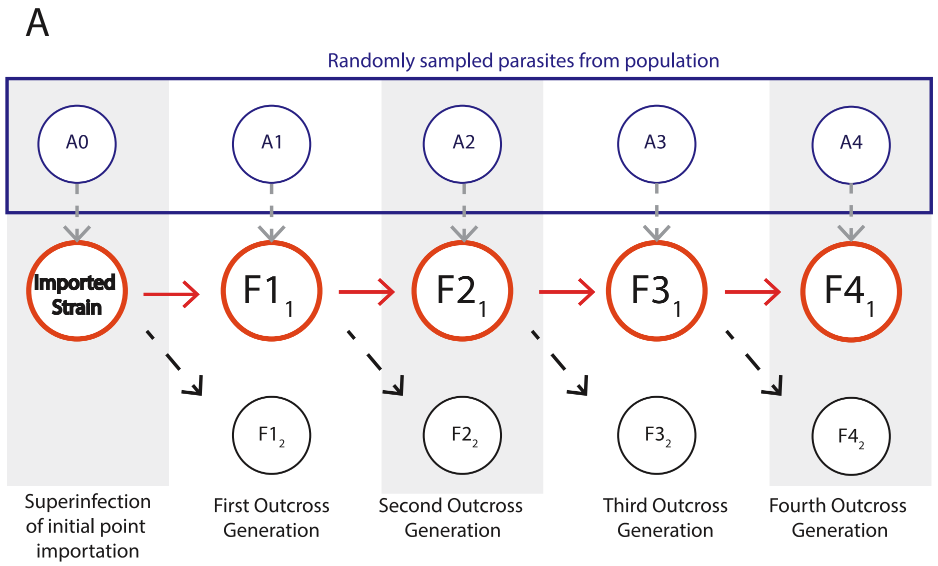
Third Degree

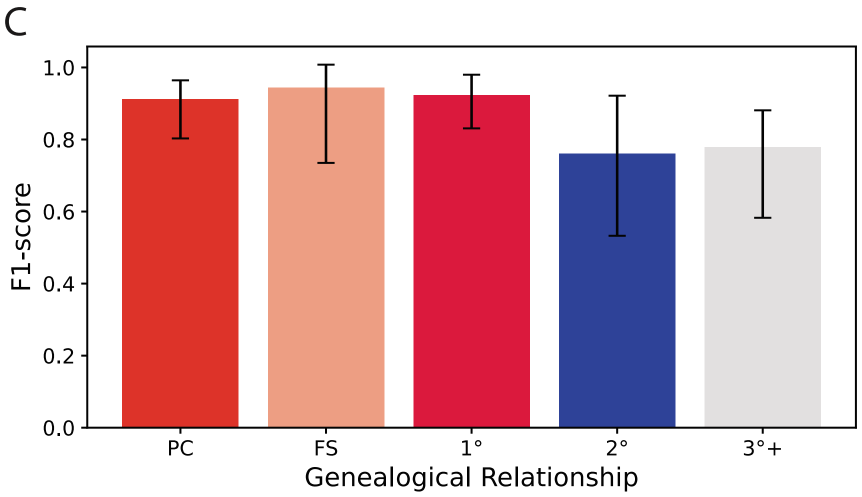


Neural Network Parameterization

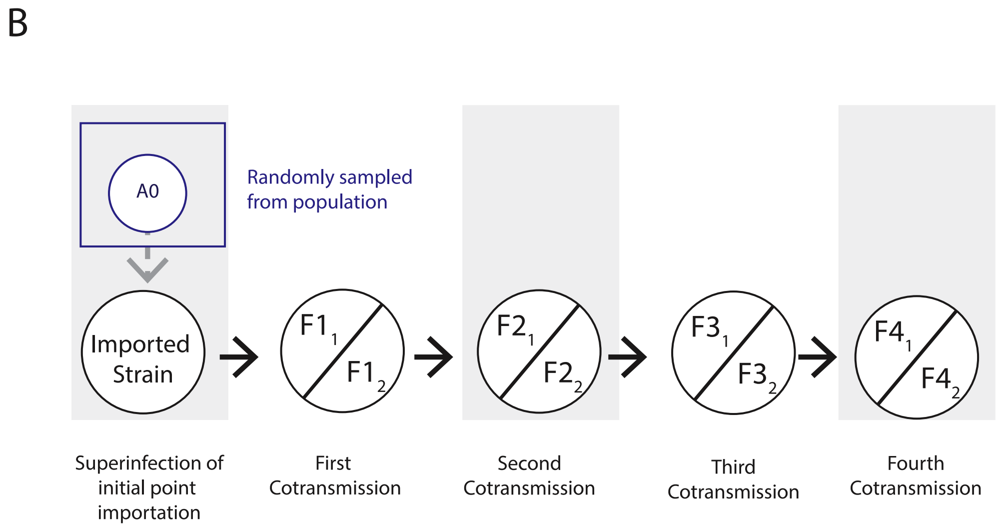
MalKinID: Accurately identifies PC relationships following outcrossed conditions

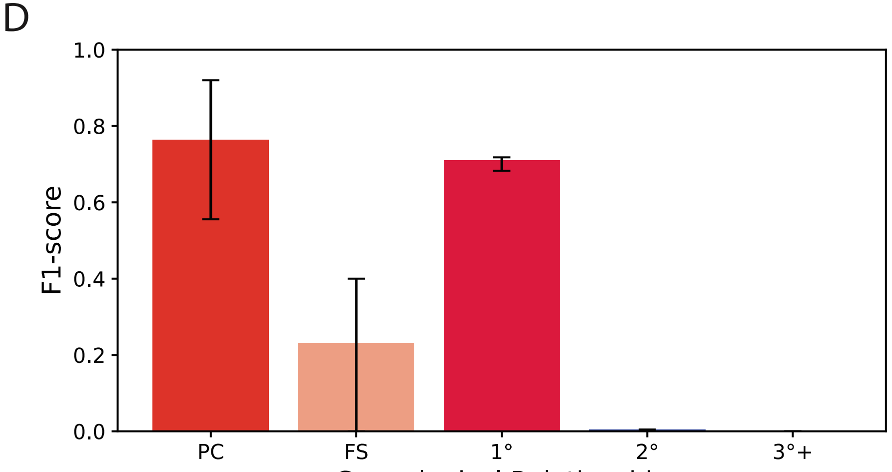
Outcross Scenario





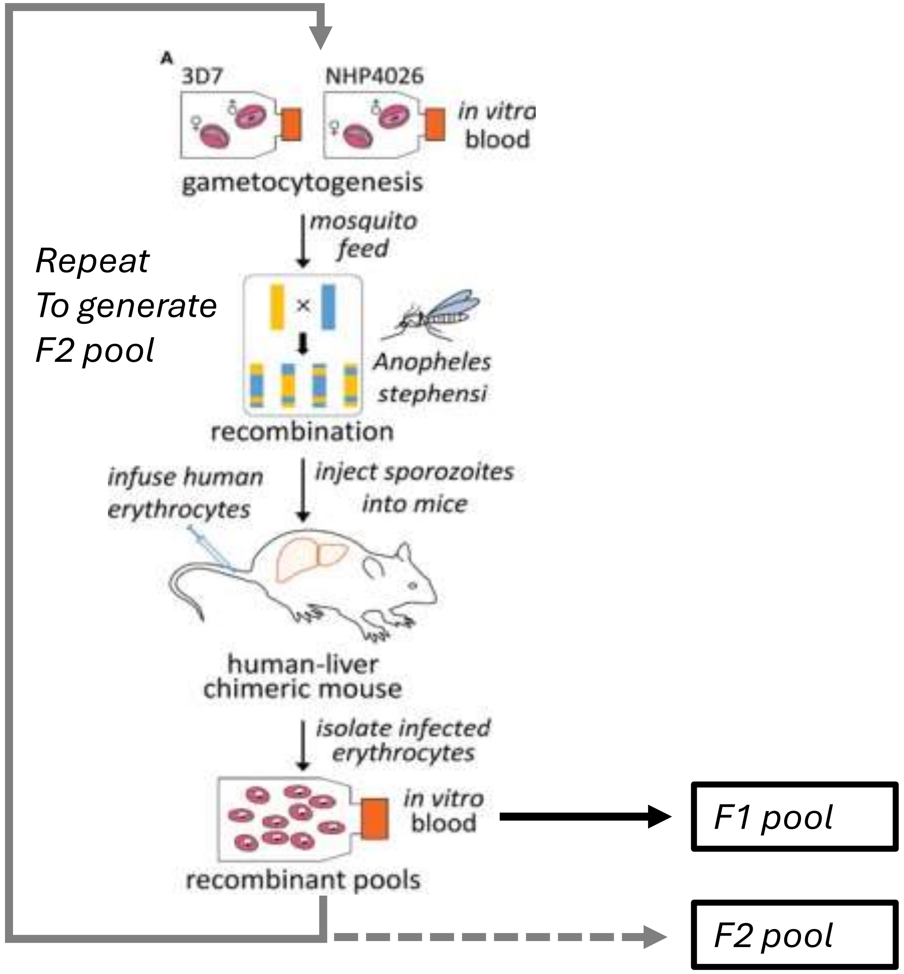
Pairwise relatedness cannot be used to reconstruct inbred lineages





Disentangling inbred genealogies requires multiple comparisons using a known tree

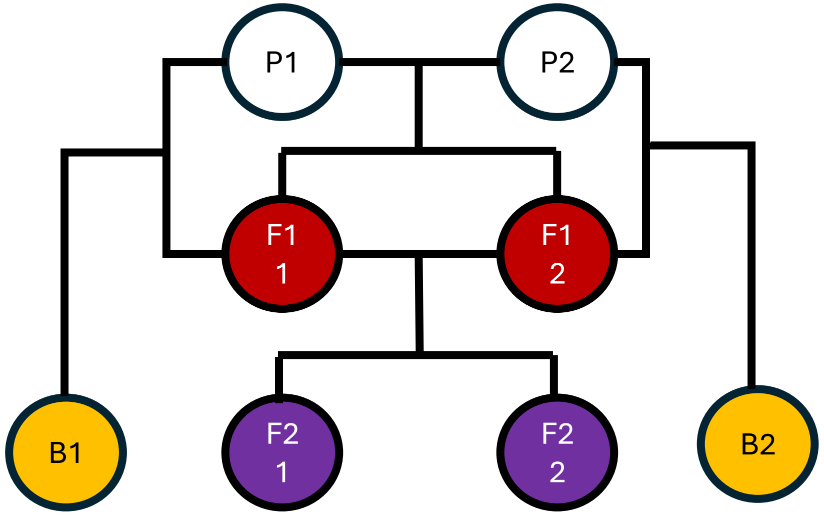
Lab Cross Design

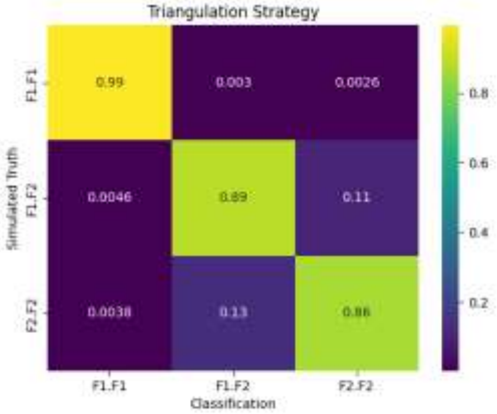


Kumar et al 2022

Can classify parasites based on **trio comparisons** to **known parasites** in the genealogy tree

(in this case P1 and P2)

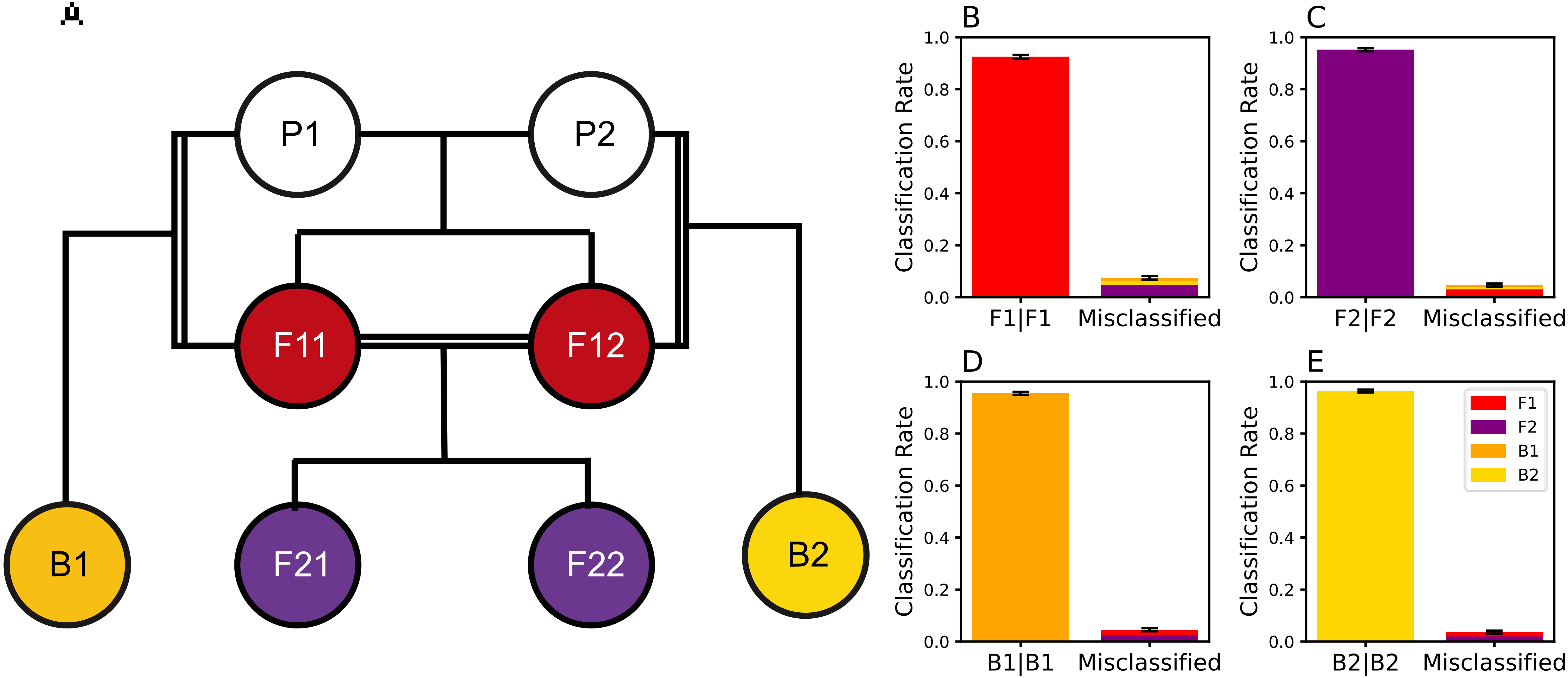




$$L(G_{s1,s2}) = L(G|r_{total,s1,s2}, IBD_{count,s1,s2}, IBD_{max,s1,s2})^{*} \prod_{i=1}^{2}\prod_{j=1}^{2} L\left(G\middle|r_{total,P_i,j}, IBD_{count,P_i,s_j}, IBD_{max,P_i,s_j}\right)$$

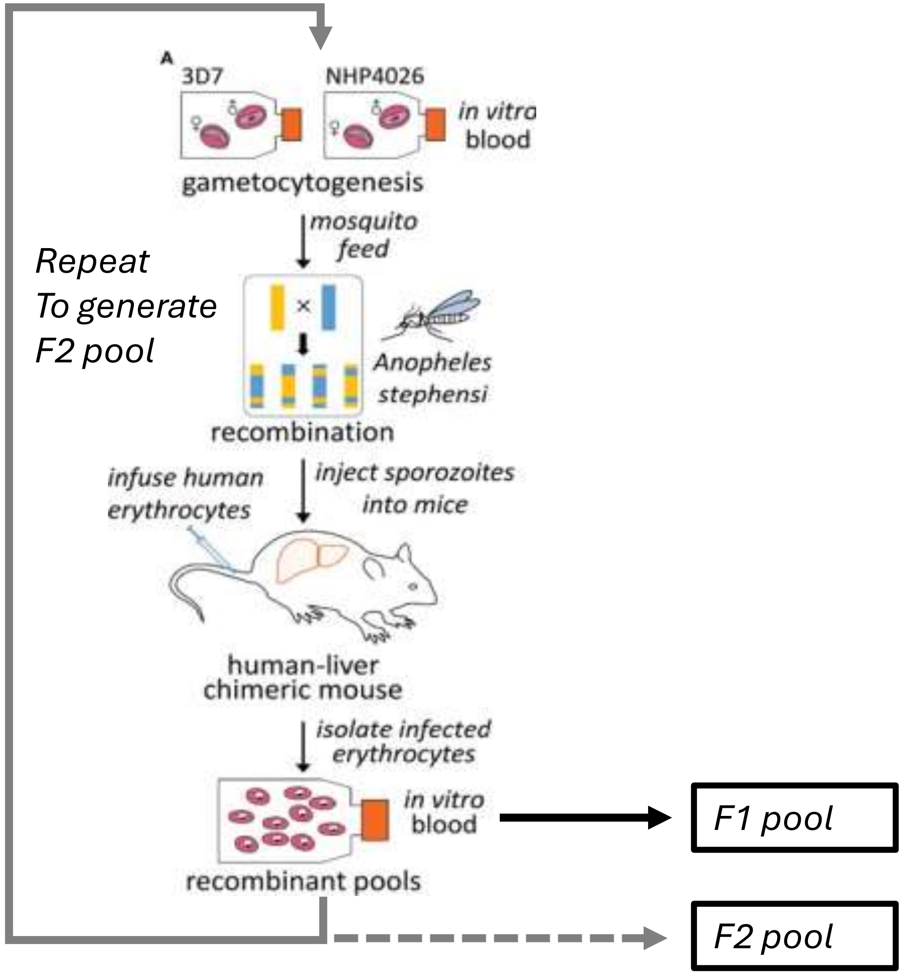
MalKinID can identify inbred genealogical classifications in lab-cross data

Based on simulated data



MalKinID reveals evidence of non-random mating in lab-generated crosses

Lab Cross Design



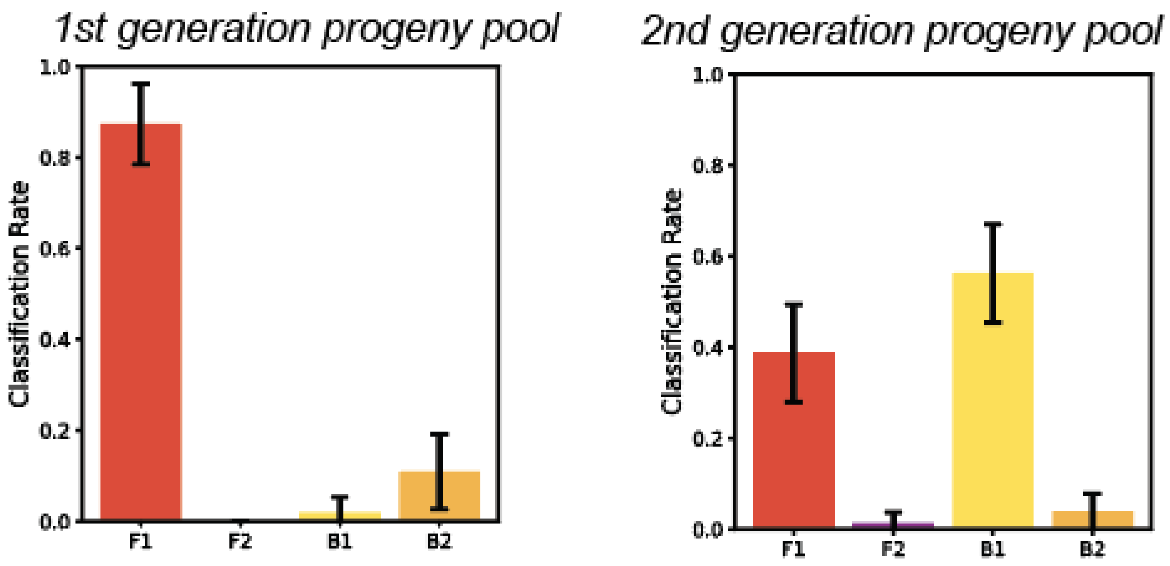
Repeat To generate F2 pool

F1 pool

F2 pool

Kumar et al 2022

Empirical Two Generation Cross (Nf54 X NHP4026)



MalKinID showed that most parasites in the second generation progeny pool where actually F1s or backcrosses of an F1 with the parental NF54

Genetic relatedness can be used to reconstruct transmission lineages and trace parasite movement IF:

Outcrossing is predominant:

- Pairwise relatedness comparisons can distinguish first, second, and third degree relatives

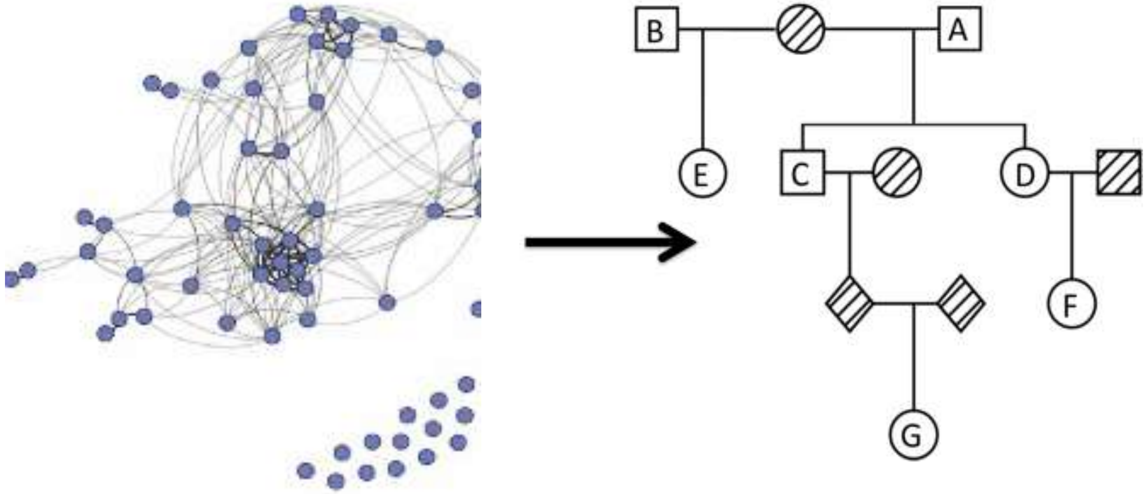
Or:

A hypothesized transmission/pedigree tree is provided

- Use MalKinID to evaluate the likelihood of a hypothesized genealogy/transmission tree

- Pair MalKinID with tree generating algorithm
    - L(Tree | Relatedness Data)

Parent-offspring pairs could be most informative for reconstructing transmission history (because parents transmit offspring)



Acknowledgements

CIGASS
CENTRE INTERNATIONAL DE RECHERCHE, DE FORMATION EN GÉNOMIQUE APPLIQUÉE ET DE SURVEILLANCE SANITAIRE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

BROAD INSTITUTE

Sarah Volkman
Daouda Ndiaye
Dyann Wirth
Aida Badiane
Kate Brenneman
Selina Bopp
Anna Burkhard
Josh Cooke
Awa Deme
Ibrahima Diallo
Mamadou Alpha Diallo
Baba Dieye
Madeline Farringer
Mamane Garba
Kiran Garimella
Amy Gaye
Dan Hartl

Paul Hinkson
Anna Hu
Malhar Khushu
Amanda Lukens
Morgan Martin
Bronwyn MacInnis
Mamy Yaye Ndiaye
Mai Ndiaye
Tolla Ndiaye
El Hadji Amadou Niang
Bassirou Ngom
Connor Payne
Alexandra Probst
Steve Schaffner
Kairon Shao
Julie Thwing
Issiaka Soulama

Djiby Sow
Rob Summers
Abdoulaye Tine
Mamadou Samba Yade
Jonn Smith
Mouhammad Sy
Imran Ullah
Wes Wong
Daba Zoumarou
Patients and their families

PNLP

CDC

CDC USAID FROM THE AMERICAN PEOPLE
President's Malaria Initiative

BILL & MELINDA GATES *foundation*

World Health Organization

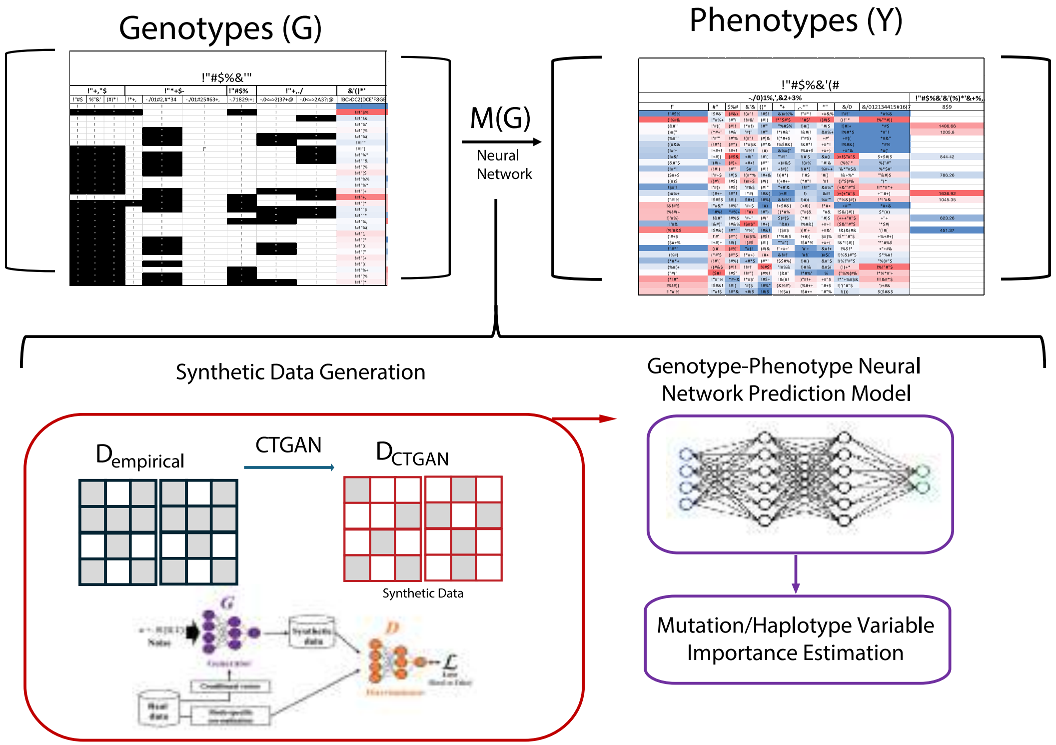
map
malaria atlas project

IDM

SAMS: An ML-based Synthetically Augmented Genotype-Phenotype Mapping Strategy

How can we predict drug resistance risk from parasite genetics?

SAMS



Genotypes (G)

Phenotypes (Y)

M(G)

Neural Network

Synthetic Data Generation

Genotype-Phenotype Neural Network Prediction Model

$D_{empirical}$

CTGAN

D_{CTGAN}

Synthetic Data

Mutation/Haplotype Variable Importance Estimation

Synthetically Generated P Falciparum Data