Next Generation Sequencing and bioinformatic tools for malaria epidemiology

Valentina Mangano (1) & Alessandro Renda (2)

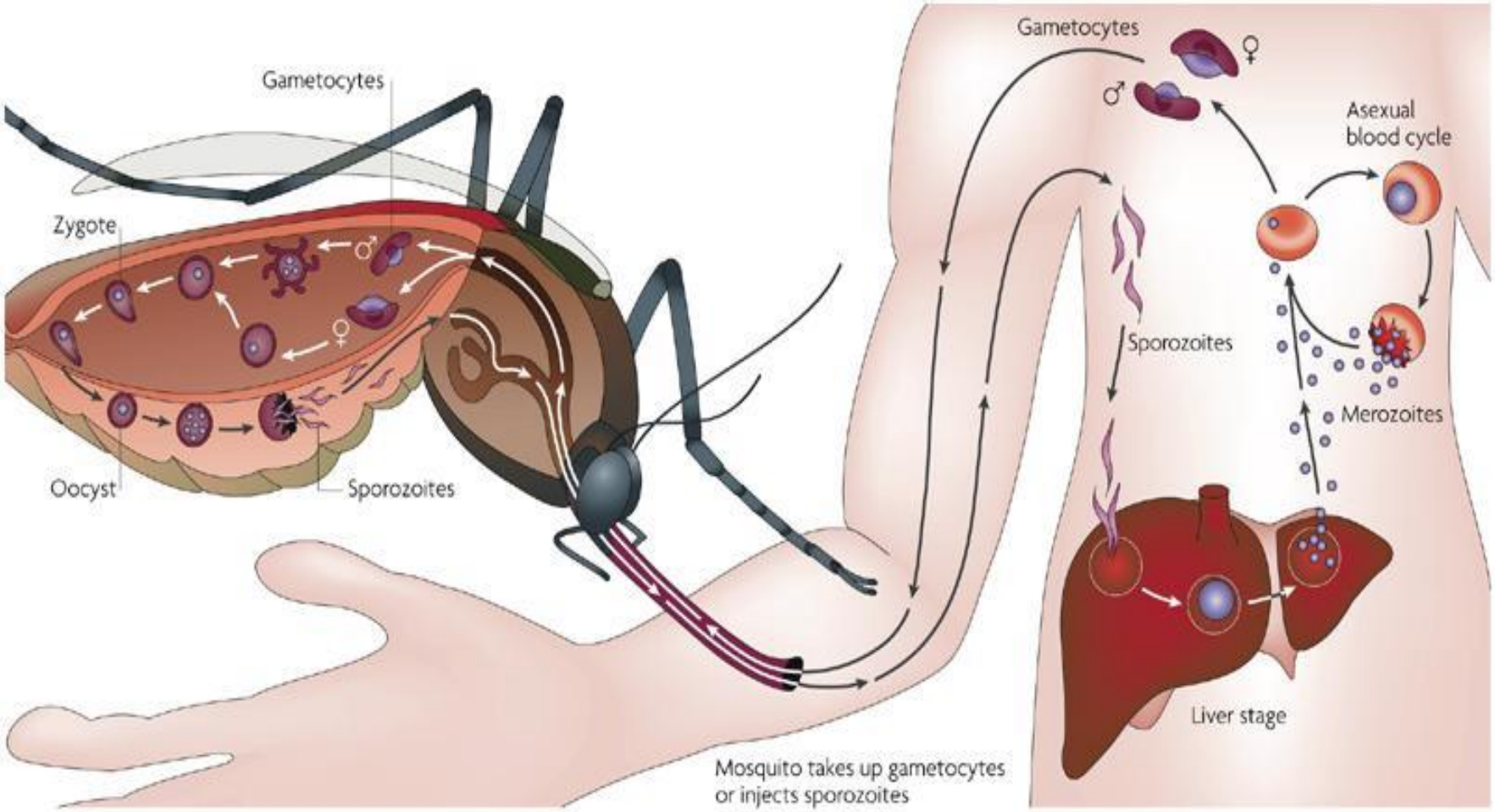
1. Dipartimento di Ricerca Traslazionale e Nuove Tecnologie in Medicina e Chirurgia

2. Dipartimento di Ingegneria dell'Informazione

Convegno Prosit 2022 «Le sfide delle tecnologie digitali per la salute del futuro»
8 luglio 2022, Polo Didattico San Rossore

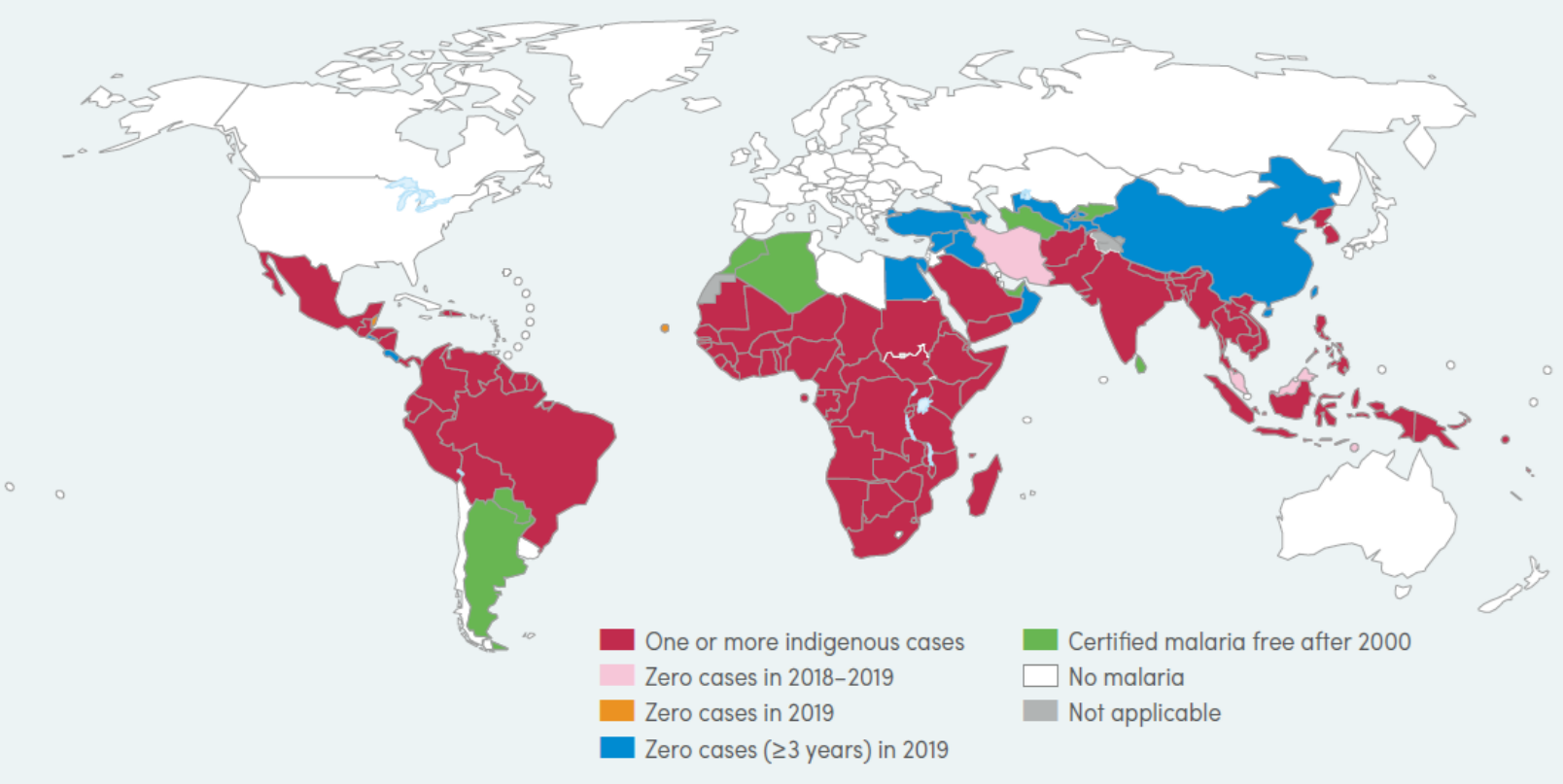
CENTRO INTERDIPARTIMENTALE PROSIT PROMOZIONE DELLA SALUTE E INFORMATION TECHNOLOGY

Malaria life cycle



Parasite: ***Plasmodium falciparum, P. vivax, P. malariae, P. ovale, P. knowlesi***
Vector: ***Anopheles gambiae, An. funestus, An. arabiensis***

The burden of malaria



WHO, *World Malaria Report 2021*

241 million clinical cases, 627000 deaths in 85 countries in 2020

Control strategies and tools

Prevention

Insecticide Treated Nets

Chemoprevention

(Vaccination)

Diagnosis

Microscopy

Rapid antigen tests

(Molecular assays)

Treatment

Uncomplicated malaria

Severe malaria

(Reservoir)

Vector resistance to insecticides	Parasite genome deletion of antigen encoding loci	Parasite resistance to antimalarial drugs

Biological threats to malaria control

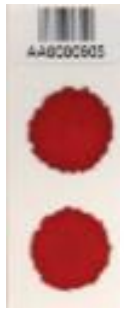
Next Generation Sequencing of malaria parasites

- Epidemiological studies and collection of Dried Blood Spots (Whatmann 903TM) from fingerprick
- DNA isolation from DBS (QIAamp Kit)
- Plasmodium selective Whole Genome Amplification
- Amplicon sequencing (Illumina MiSeq, 150 amplicons of 200bp) and Single Nucleotide Polymorphisms calling (SpotMalaria/GenRe v3.0 pipeline) of:
  - mitocondrial regions for parasite detection and species id
  - regions harbouring drug resistance mutations
  - regions harbouring informative variation for genetic barcode (101 SNPs)
- Whole Genome Sequencing of high quality *P. falciparum* positive samples for SNP and CNV calling

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

wellcome sanger institute

SPOT Malaria



DBS

Illumina MiSeq

Next Generation Sequencing of malaria parasites

- All inhabitants (Mossi, Rimaibe, Fulani ethnicity) of 4 rural villages of the Plateau Central region
- 7016 DBS samples collected during 4 cross sectional surveys

Development and testing of bioinformatic tools

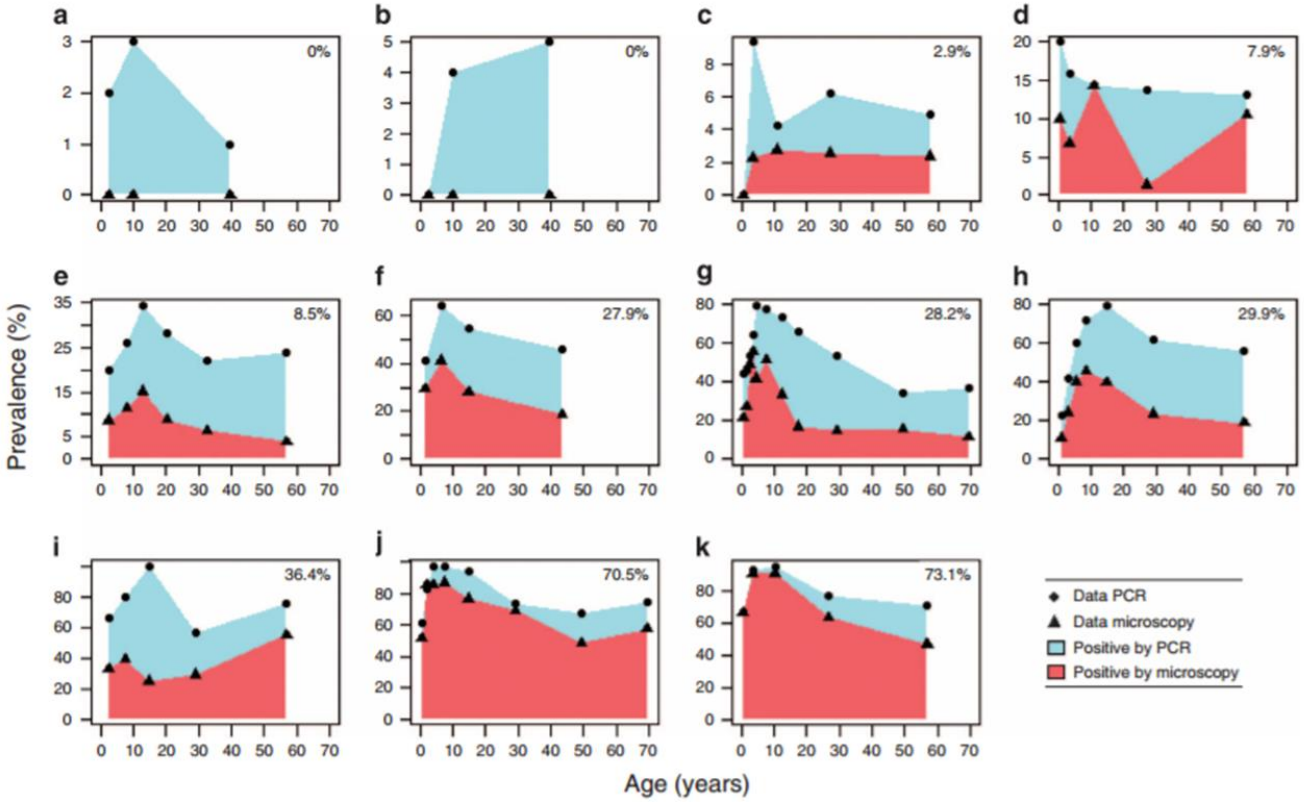
- Children under 5 and pregnant women attending 3 primary care health centres in Western Equatoria State
- 1751 DBS samples collected during 1 cross sectional survey

NGS

NGS analyses provide actionable information

Ultrasensitive detection of infection

- Subjects with parasite density below limit of detection of microscopy/RDT infect mosquitoes and contribute to transmission
- Higher frequency in lower prevalence settings
- Reaching elimination, treatment of low density/asymptomatic infection reservoir is needed to interrupt transmission



Okell et al. Nature Comm 2012

NGS analyses provide actionable information

Resistance to antimalarial drugs

- Resistance to artemisinin emerged in Sout East Asia since 2009 and has been documented in Rwanda in 2020
- Molecular surveillance for early detection of resistance to artemisinin
- Monitor resistance to artemisinin partner/alternative drugs for planning of regimen switches

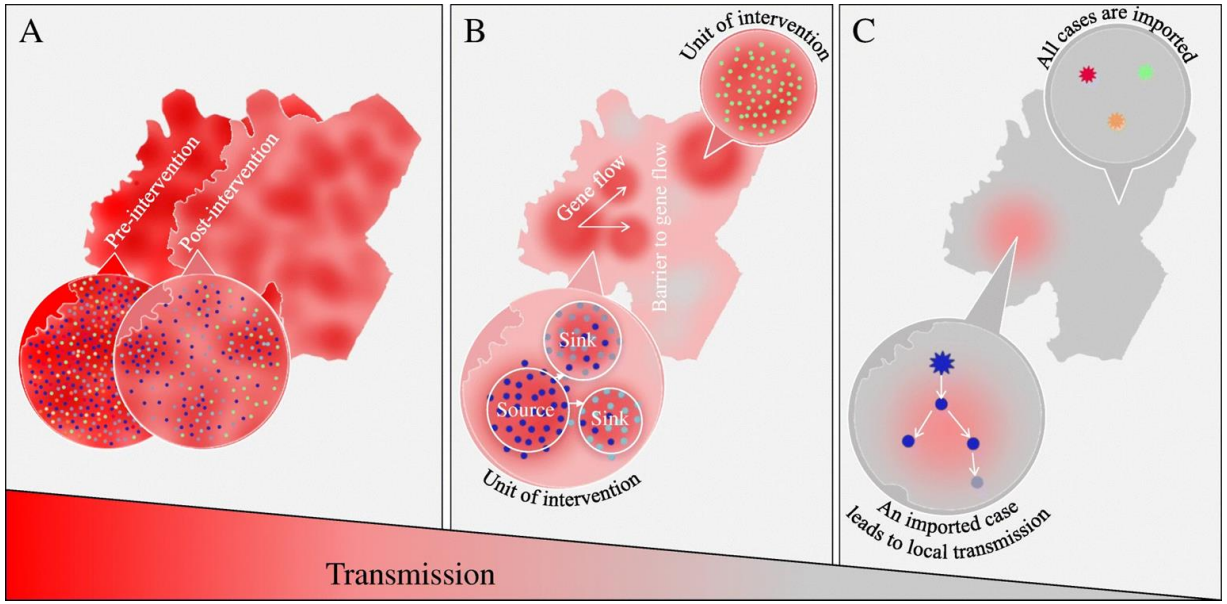
Worldwide Antimalarial Resistance Network

Uwimana et al. Nature Medicine 2020
Stokes et al. Elife 2021

NGS analyses provide actionable information

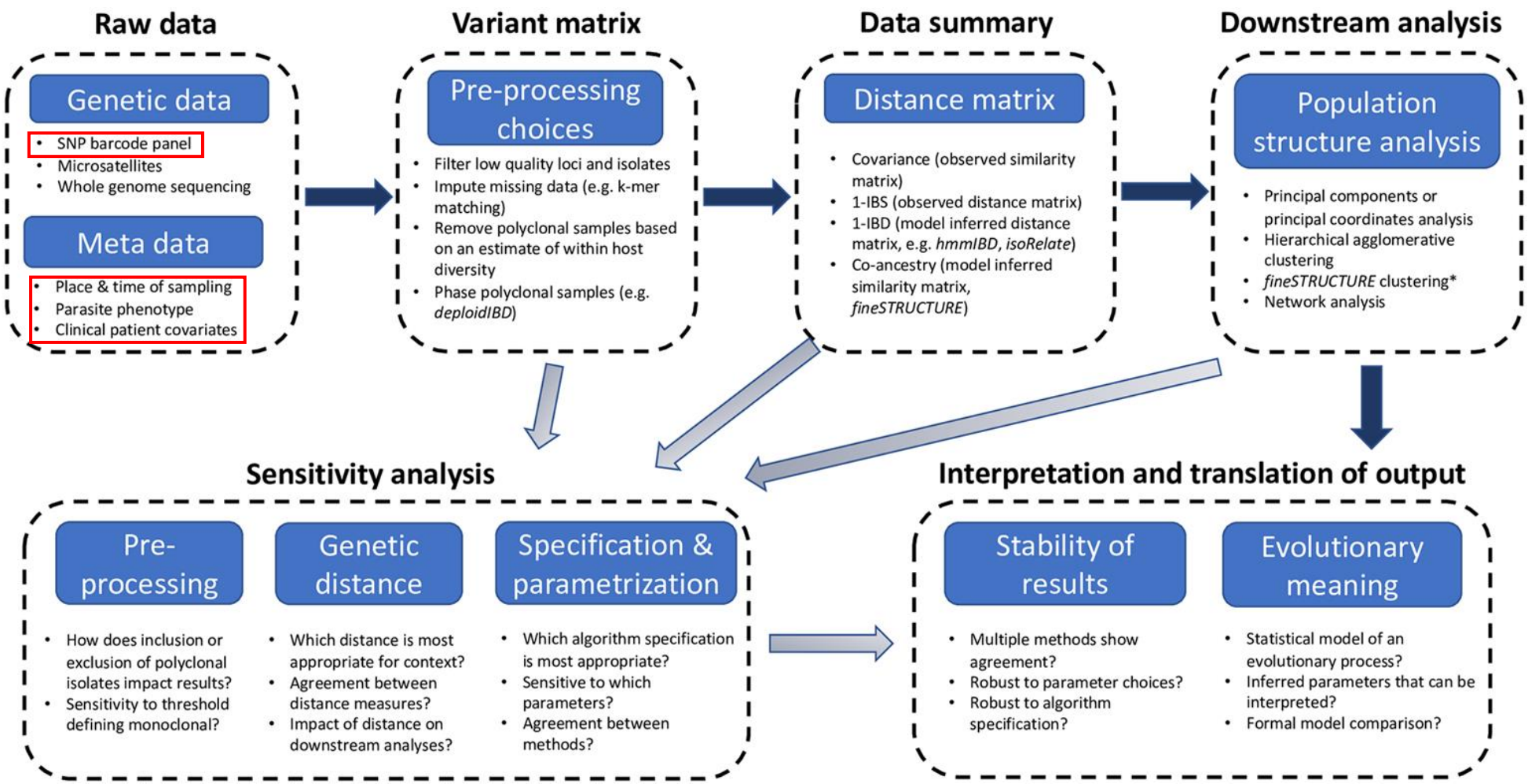
Genetic barcode of *P. falciparum* strains

- Distinction of incident vs chronic infections
- Calculation of Complexity of Infection (COI): number of genetically distinct parasite strains co-infecting a single host; indicator of transmission intensity
- Characterization of parasite relatedness/population structure
- Monitor fine-scale spatiotemporal transmission patterns, control programmes planning and evaluation



Wesolowski et al. BMC Medicine 2018

But…there is not obvious analysis pipeline



Watson et al. Plos Genetics 2020

Creating the data analysis environment

Jupyter Notebook: web-based interactive development environment

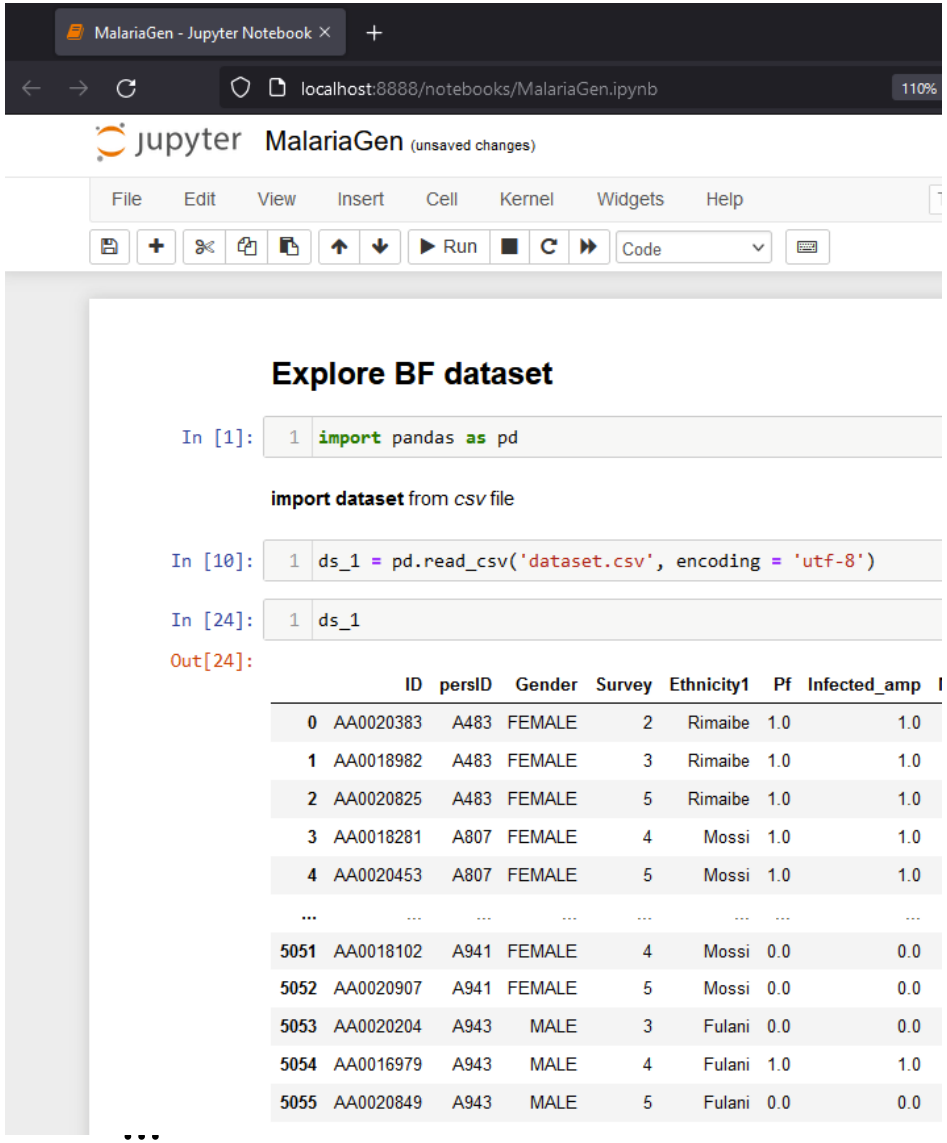
- collection of *text* cells and executable (and updatable) *code* cells, with the respective output
- supporting over 40 programming languages, including Python and R

Python: modern, general-purpose, object-oriented, high-level programming language

- clean, quite simple, expressive: fewer lines of code, fewer bugs
- equipped with large standard library + large collection of add-on packages, including **Pandas**

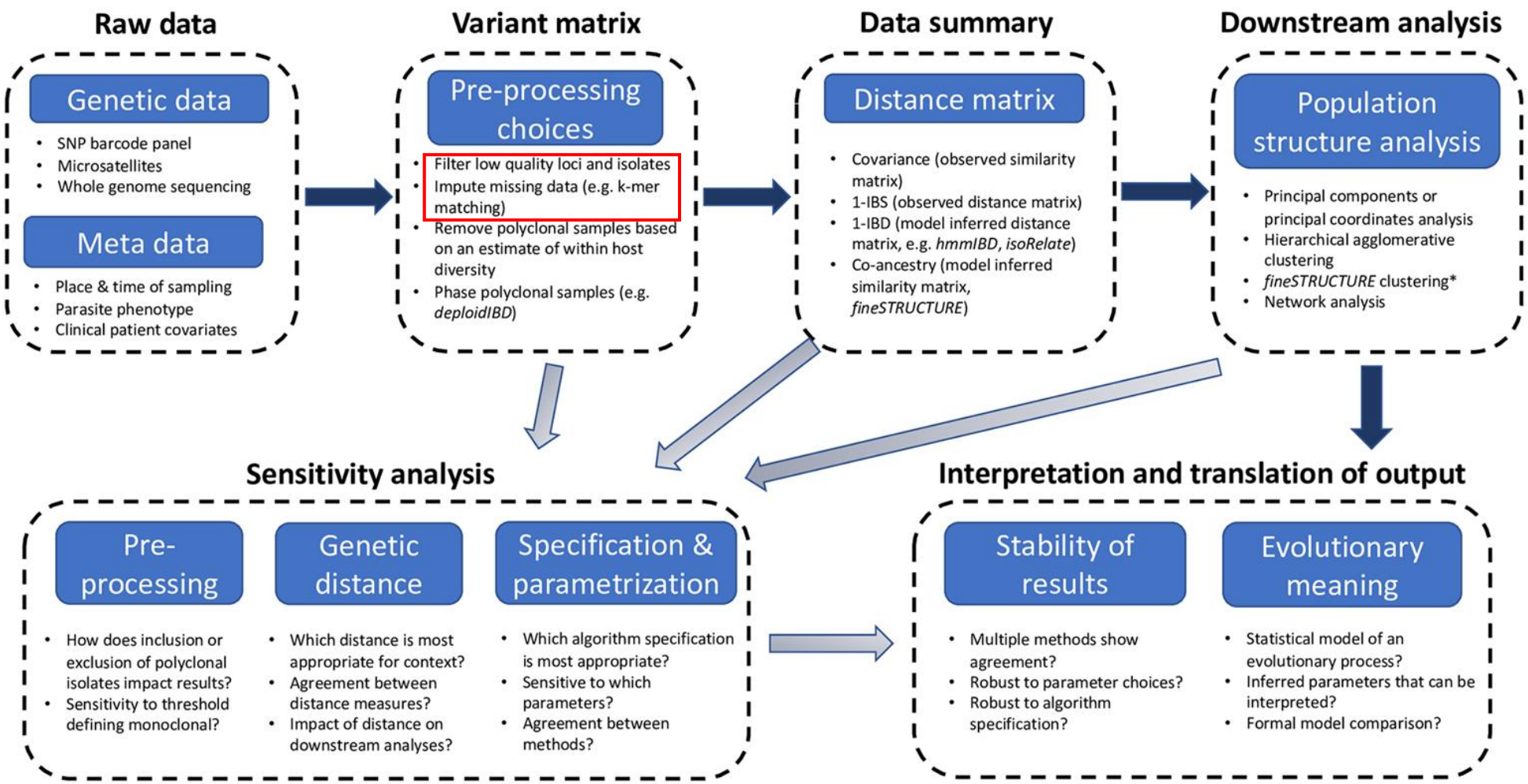
Pandas: Python library designed to make data pre-processing and data analysis fast and easy

- suitable for handling heterogenous data represented in tabular format
- widely adopted in *data science* community



...

But…there is not obvious analysis pipeline

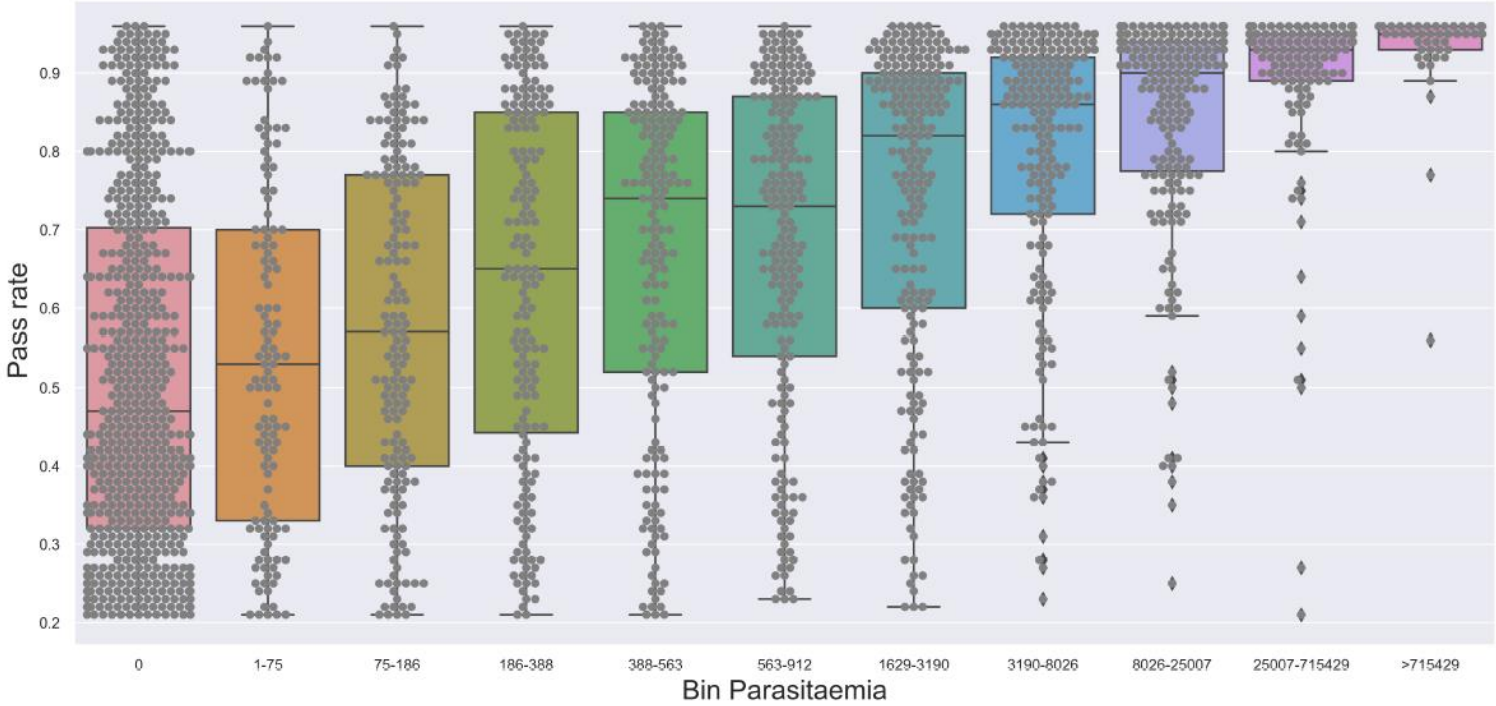


Watson et al. Plos Genetics 2020

Data QC and ascertainment bias

Preliminary analysis

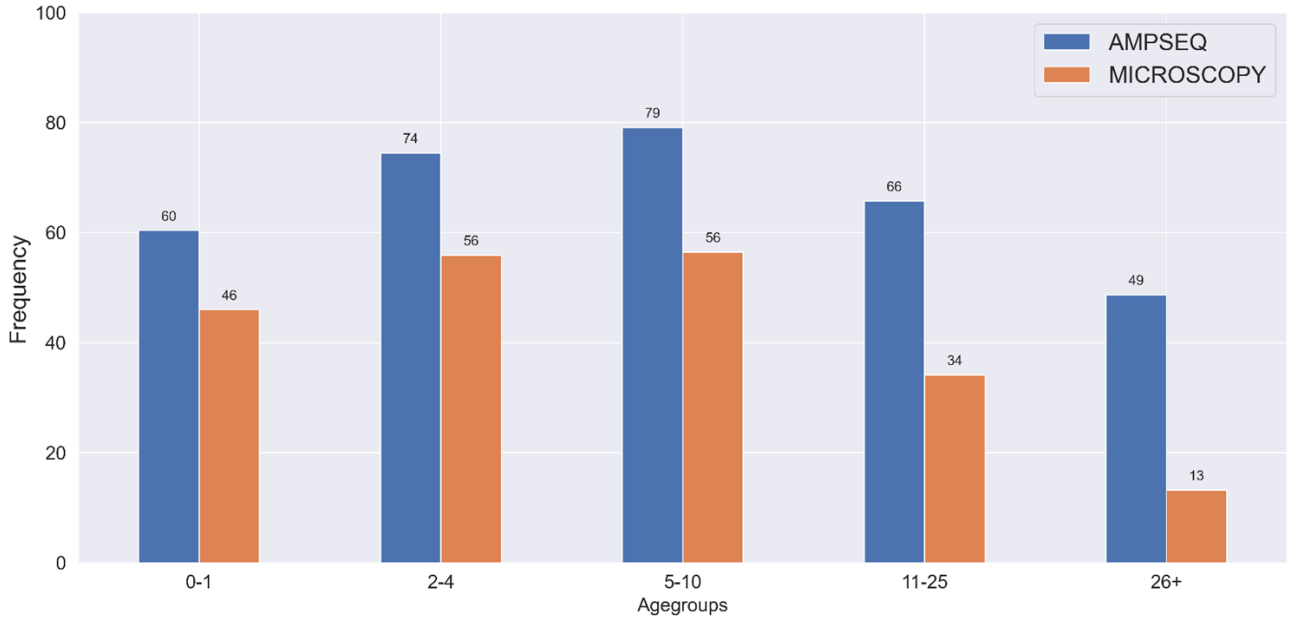
- Effect of parasite density on barcode pass rate
- Barcode pass rate thresholds impose ascertainment bias
- Consider alternative of imputing missing positions



Ultrasensitive detection of infection

Preliminary analysis

- Overall, infection is detected in 65% of the population by NGS (AmpSeq) vs 39% by microscopy
- Same age-group profile
- The frequency of infections undetected by microscopy is larger in older age-groups (lower parasite densities)
- Adults are not target of preventive strategies (e.g. SMC, LLITN) while representing a reservoir for transmission



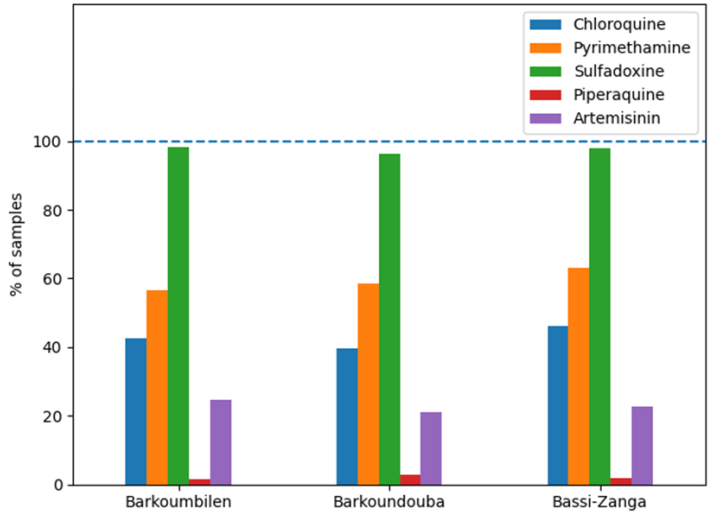
Resistance to antimalarial drugs

Frequency of markers of resistance to antimalarials, alone and in combination

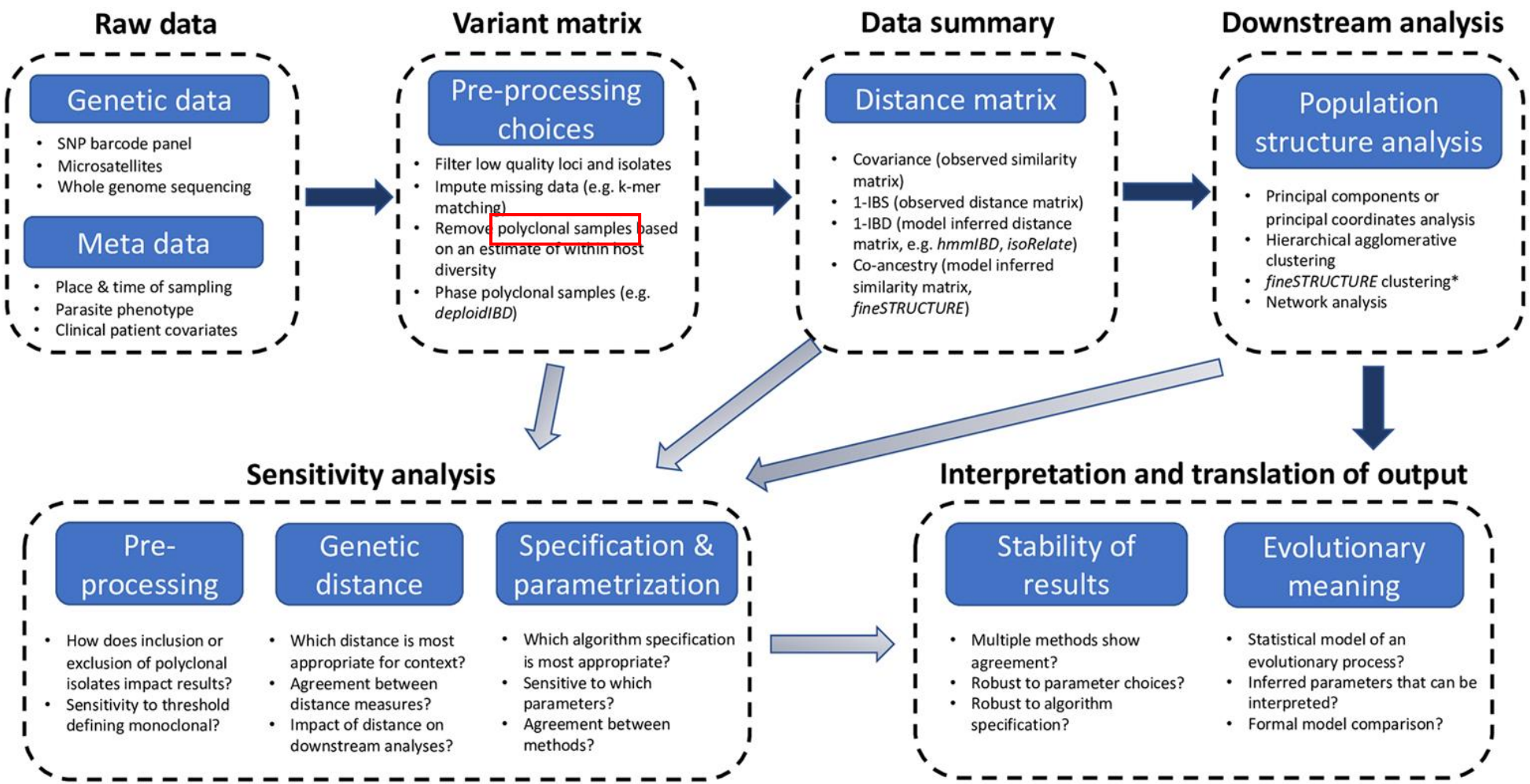
	Chloroquine	Pyrimethamine	Sulfadoxine	Piperaquine	Artemisinin
Chloroquine	43	27,1	42,4	1	12,4
Pyrimethamine		58,9	57,8	1,3	16,1
Sulfadoxine			97,9	1,8	23
Piperaquine				1,8	1,3
Artemisinin					23,3

Preliminary analysis

- Mutations causing resistance to both Pyrimethamine and Sulfadoxine, used in combination for Intermittent Preventive Treatment of pregnant women, were observed in 57.8% of parasites
- Variants at the *Kelch13* locus, involved in resistance to artemisinin, the first line drug used for treatment of clinical cases, were detected in 23.3% of parasites



But…there is not obvious analysis pipeline



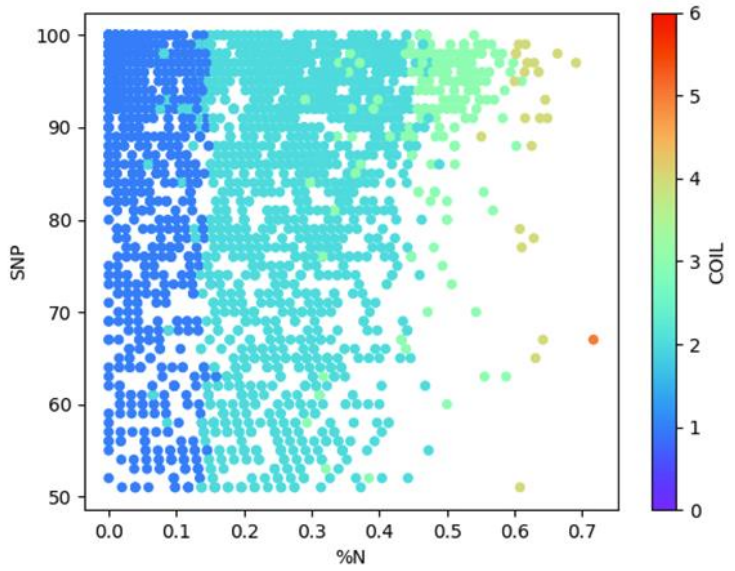
Watson et al. Plos Genetics 2020

COIL vs Real McCOIL

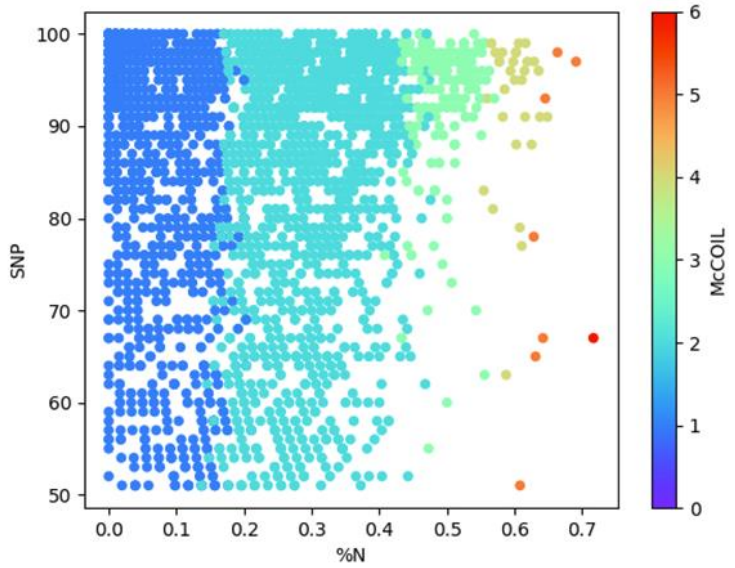
Preliminary analysis

- Comparison of methods for estimating COI
  - **COIL**: COI using Likelihood
  - **The Real McCOIL**: Turning HEterozygous SNP data into Robust Estimates of ALelle frequency, via Markov chain Monte Carlo, and Complexity Of Infection using Likelihood
- Obtain similar results on our data, but The Real McCOIL better captures the relationship beetween COI and the number of heterozygous calls

COIL



The Real McCOIL

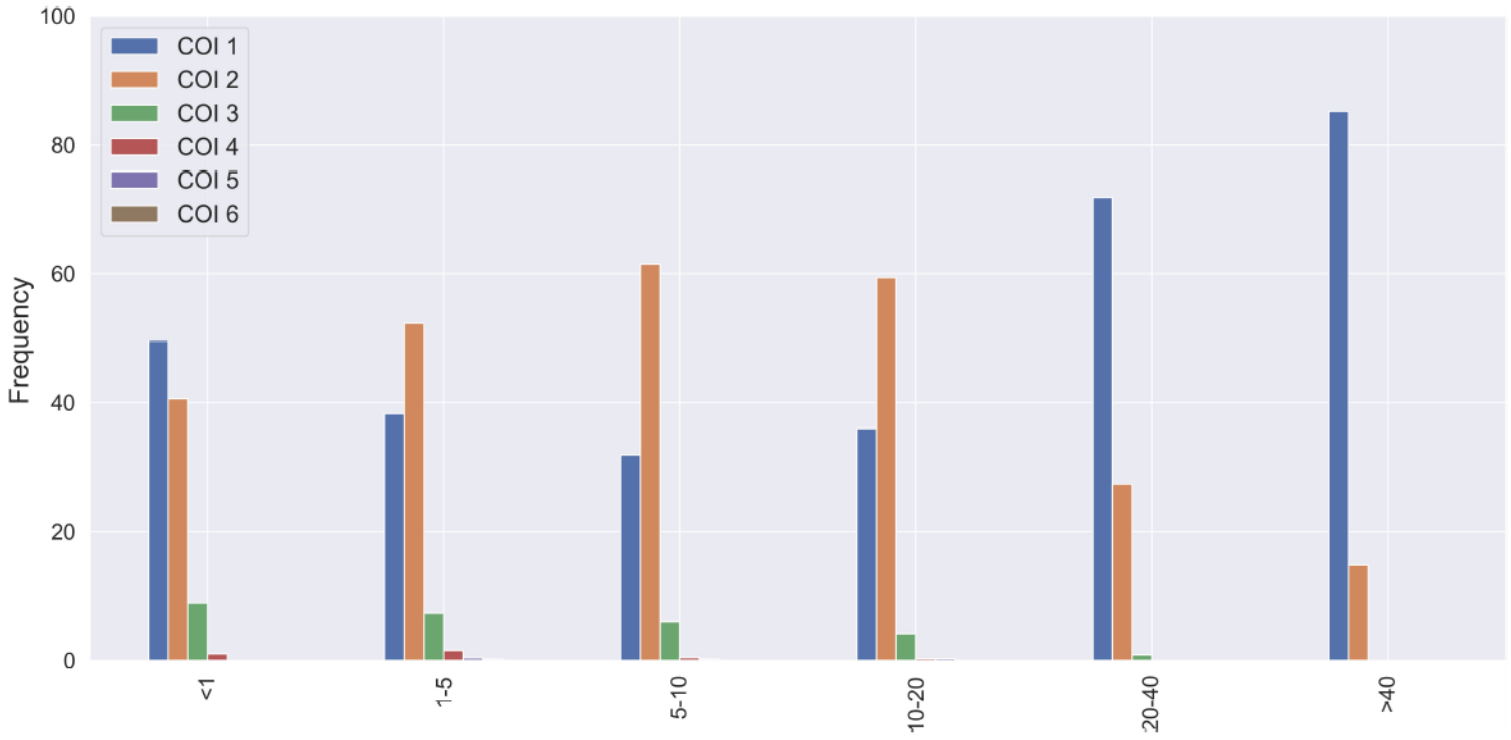


Complexity of Infection (COI)

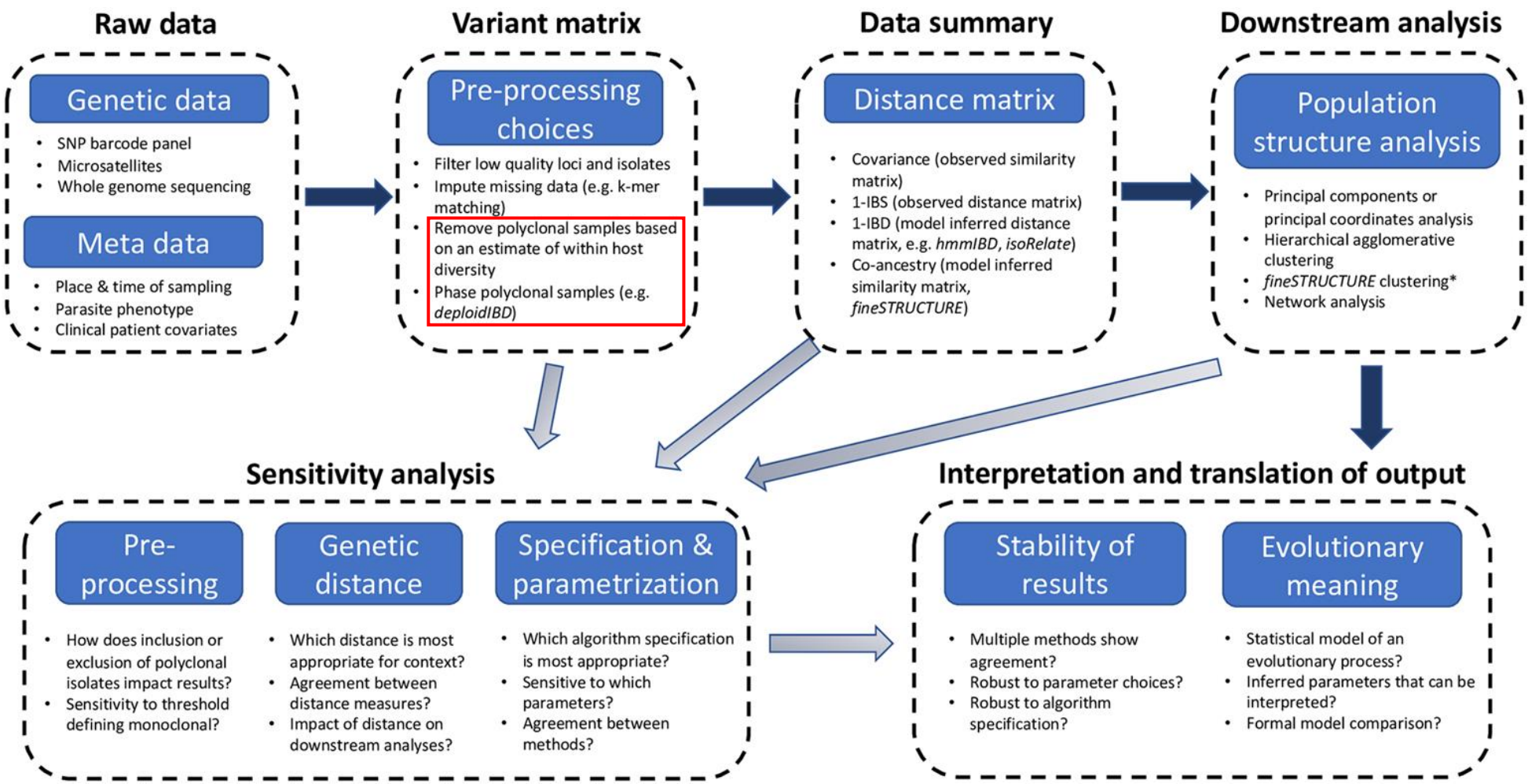
Preliminary analysis

COI decreases with increasing age, reflecting acquired immunity to *P. falciparum* circulating strains

Is clinical immunity acquired faster or stronger to certain strains than others?



But…there is not obvious analysis pipeline



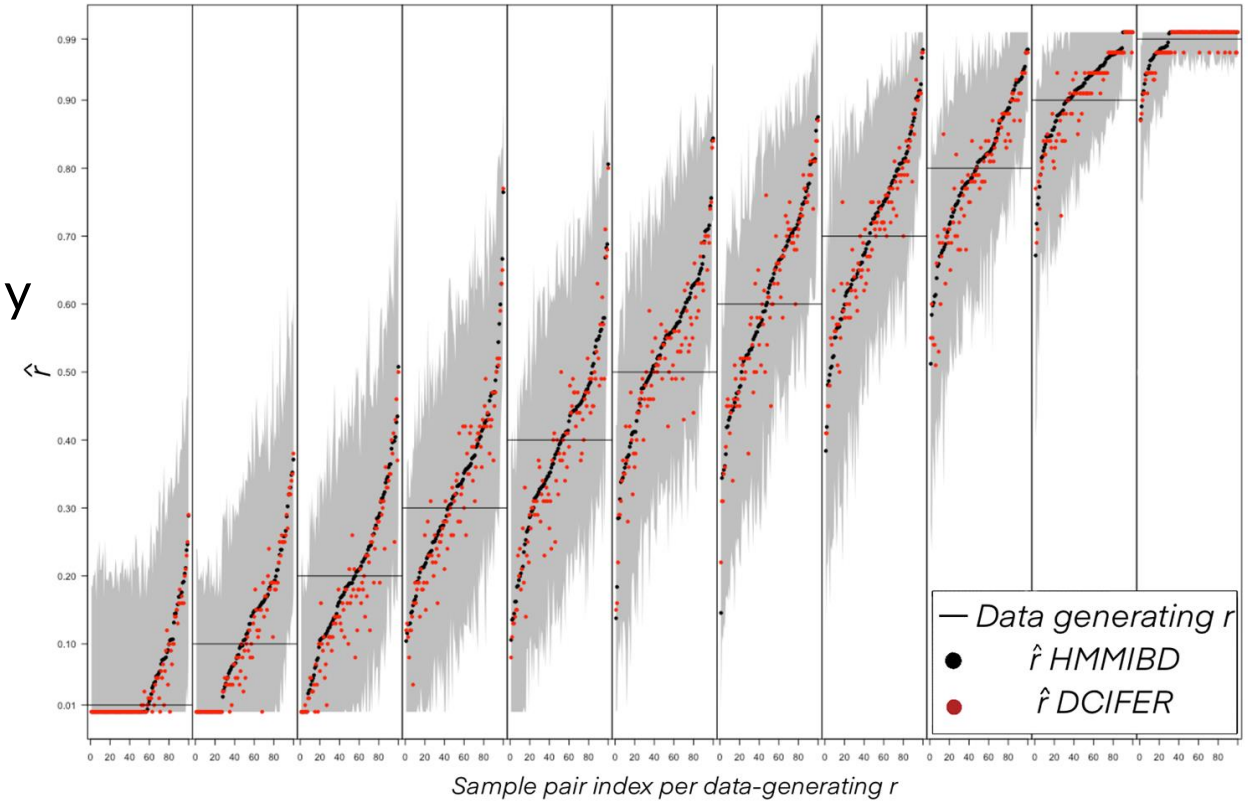
Watson et al. Plos Genetics 2020

Dealing with polyclonal infections

Thanks
Inna Gerlovina - UC, San Francisco
Aimée Taylor - Institut Pasteur

Identity By Descent (IBD): proportion of identical DNA sites inherited without recombination from a common ancestor

- **hmmIBD** (2017) – monoclonal infections only (*ascertainment bias*)
- deploidIBD (2019) – phased polyclonal infections (*computationally intensive*)
- **Dcifer** (2022) – unphased polyclonal infections!



Preliminary analysis

- evaluate the effectiveness of two methods (Dcifer vs hmmIBD) on synth data (known fixed *relatedness*)
- methods yield coherent results, with distributions of estimates centered on the true value
- Dcifer will be used to estimate IBD and analyse population structure on the whole BF dataset

Acknowledgments

- Bienvenu Sirima, Issa Nebie, Youssouf Kabore
- David Modiano, Federica Verra
- Victoria Alawia Alberto, Gabriel Joseph Morbe Tangun
- Chiara Scanagatta, Giampietro Pellizzer, Diego Longoni
- Dominic Kwiatkowski, Kirk Rockett, Kimberly Johnson, Sonia Goncalves
- Inna Gerlovina
- Aimee Taylor
- Alessio Bechini, Andrea Arcangeli, Federica Bassi, Giulia Bonizzi, Mattia Celia Magno, Stefania Bertoncini