Institut Pasteur de Dakar

Computational Complexity and Resource needs in Genomic Data Analysis

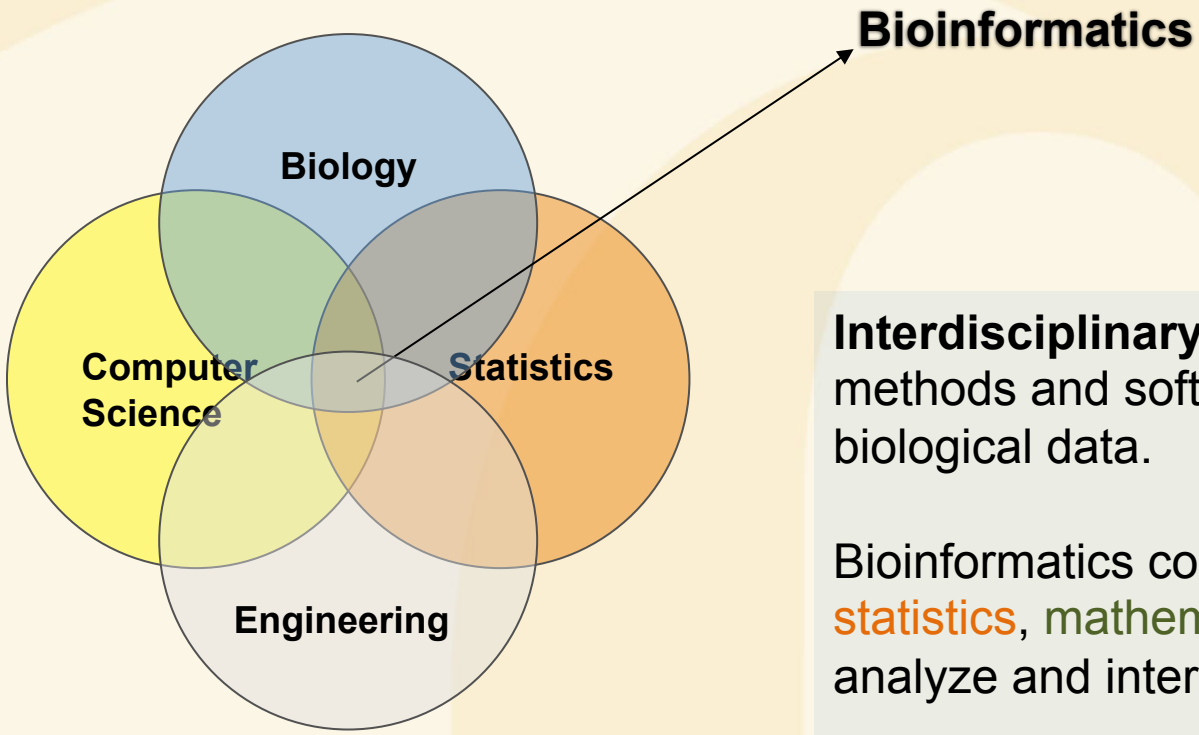
Cheikh LOUCOUBAR, PhD in Statistical Genetic
Group of "Biostatistics, Bioinformatics and Modeling"
Institut Pasteur de Dakar

Outline

- Bioinformatics Context
- Our Research Group
- Our Research Project
  - Objectives
  - Themes
  - Data types

- Studied diseases
- Examples of analysis and computational characteristics
  - Genome-wide linkage on SNPs + microsatellites
  - Genome-wide association studies (single-SNPs GWAS)
  - Detecting multi-way epistasis in family-based association studies (M-TDT)

What's Bioinformatics ?



Interdisciplinary science developing methods and software tools for studying biological data.

Bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

(Wikipedia)

What are the input data ?

A succession of 4 letters: **A** , **T** , **C** et **G**
Coding the genetic information

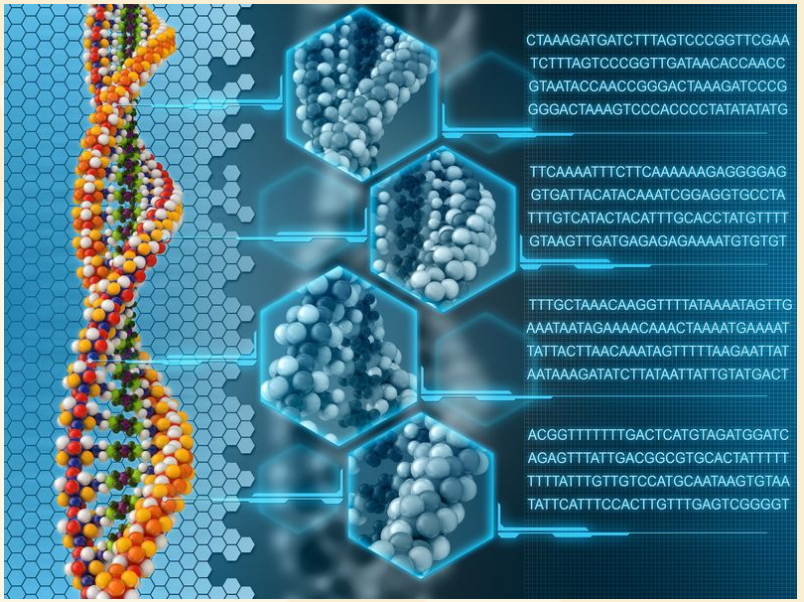
This **Information** is contained in nucleus of cells of an organism

This **information**, as a whole is called:
Genome

The **genome** codes life (such as a source code for a computer program)

The code can be divided into functional units:
Genes

From genes, by a complex process within cells, interacting with its environment, **proteins** are produced: **the expression of the source code**



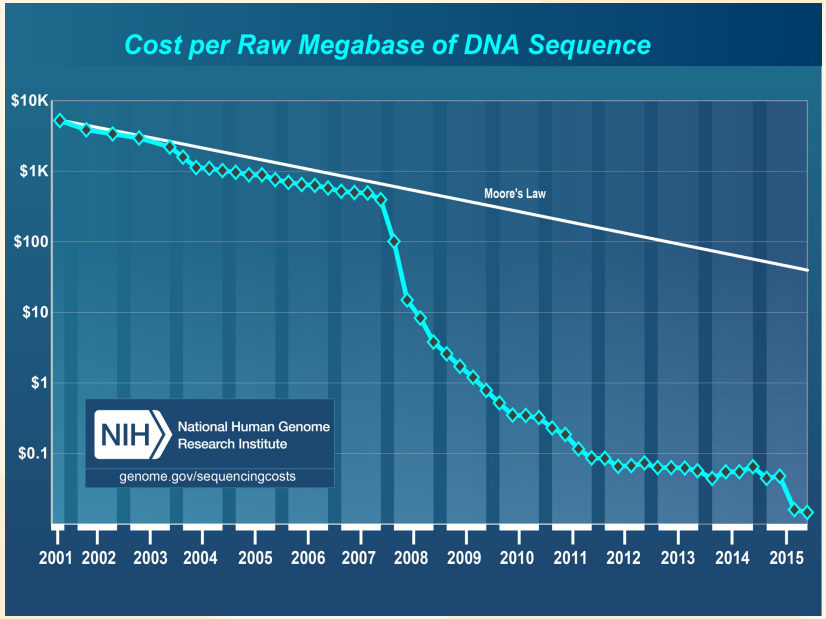
Institut Pasteur de Dakar

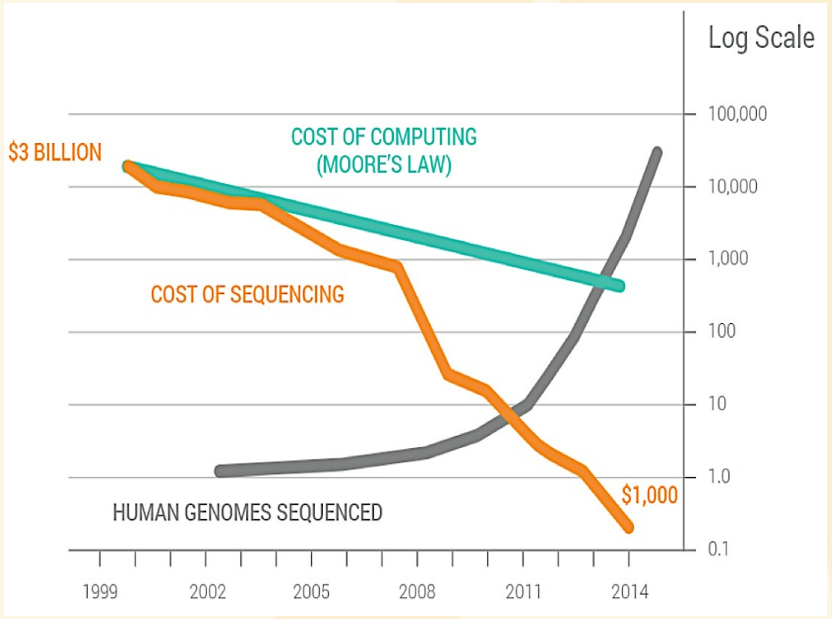
Cost for data production ?

Since the 2000s

Next-Generation Sequencing (**NGS**)

a set of High-throughput sequencing technologies costing less and less expensive





Institut Pasteur de Dakar

Institut Pasteur de Dakar (IPD)

The Institut Pasteur de Dakar (IPD) is a Senegalese Foundation

- recognized **charity**,
- **non-profit**,
- is allowed to **contribute to public health** in Africa and particularly in Senegal

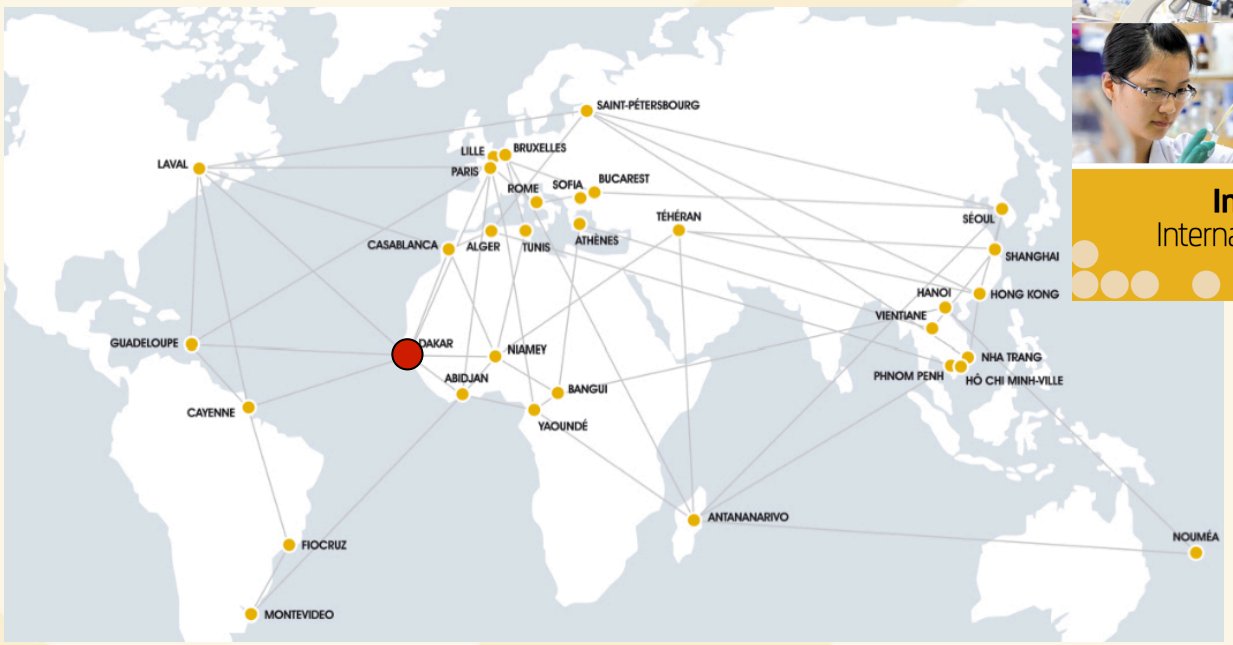
IPD's main activities:

- Research
- Education
- Training
- Medical expertise
- Production of Yellow Fever vaccine

Institut Pasteur de Dakar

IPD belongs to a network of 33 Pasteur Institutes

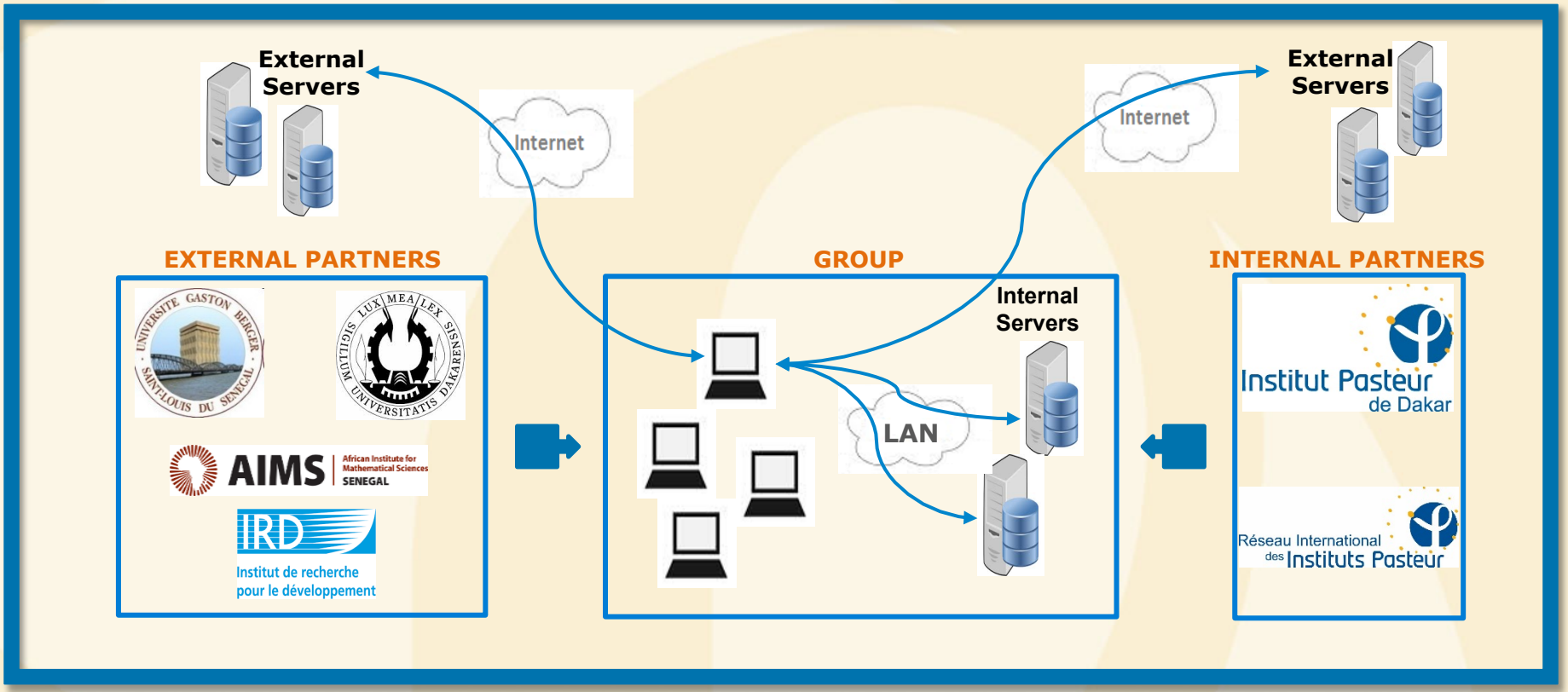
IPIN: Institut Pasteur International Network



Institut Pasteur de Dakar

Our Research Group

Group of Biostatistics, Bioinformatics and Modeling



Institut Pasteur de Dakar

Our Research Group

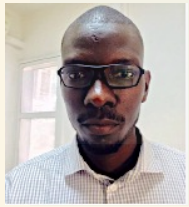
Cheikh LOUCOUBAR

Biostatistician
PhD in Statistical Genetics
(Head of the Group)

Maryam DIARRA

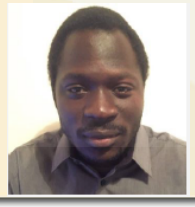
Biostatistician
PhD in Applied
Mathematics

Mamadou DIOP



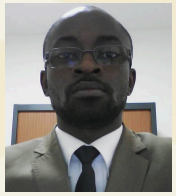
Bioinformatician
Master in Computer
Science

Mor A. LOUM



PhD Student
in Mathematics
Paris-Sud Orsay University

Grace MUPOYI NTUALA



Master Fellow
M2 Mathematics
AIM

Ismaila BA

Master Fellow
M2 Mathematics
AIMS

Raima Appaw CAROL

Master Fellow
M2 Mathematics
AIMS

Institut Pasteur de Dakar

Objectives

We aim at identifying environmental factors, genetic factors and their interactions implicated in arboviral and malaria infections and diseases

- development of statistical methods and tools
- definition of analysis pipeline
- Adapted for genetic association analysis of complex traits

(Multi-markers association methods)

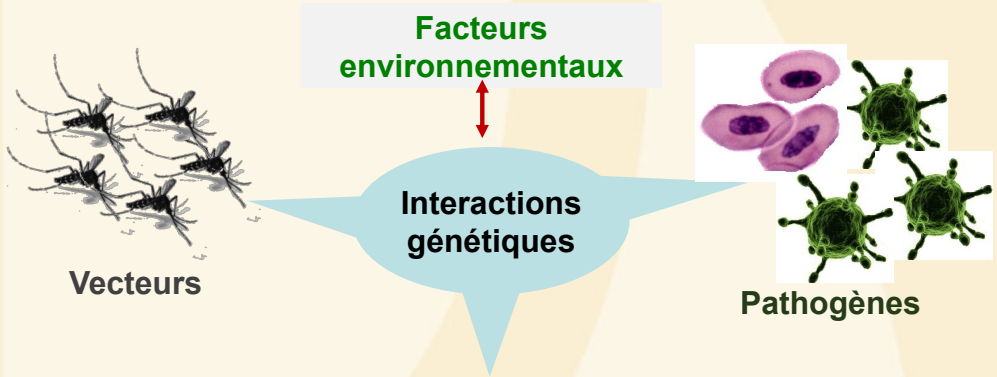
Scientific interests:

- Genes-genes and genes-environment interactions
- Pleiotropic genes within and between diseases
- Disease transmission modeling and Risk assessment

Theme (1)

Study of epidemiological and environmental factors:

- data mining
- Generalized regression models
- Spatial statistics

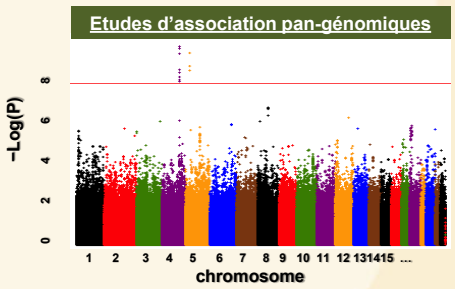


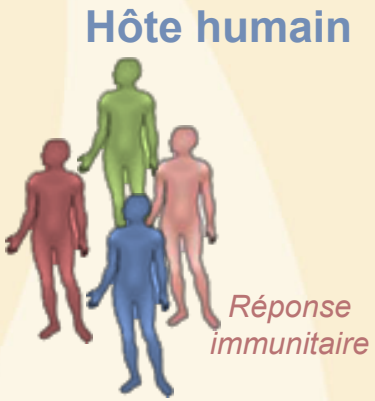
Institut Pasteur de Dakar

Theme (2)

Genetic associations studies

- Sequence data analysis
- GWAS studies (family-based designs)
- Post-GWAS methods including knowledge (Pathways, text mining, gene co-expression networks)





Institut Pasteur de Dakar

Data types - SNPs, Microsatellites

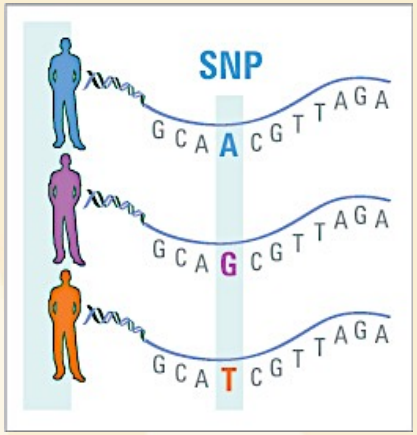
Genetic markers: varying from an individual to another

SNPs (bi-allelic markers)

Total length of Human Genome ~ **3 billion** base pairs
Number of SNPs ~ **1.5 million** (**1 every 2000 bp**)

SNPs are, for the majority, responsible for differences observed between individuals of the same species.

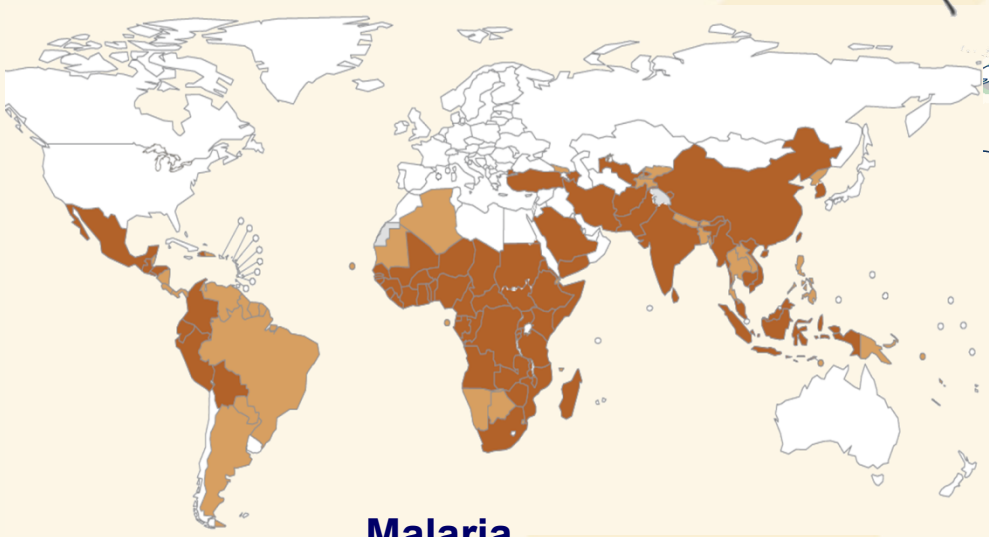
Are implicated in many diseases susceptibility

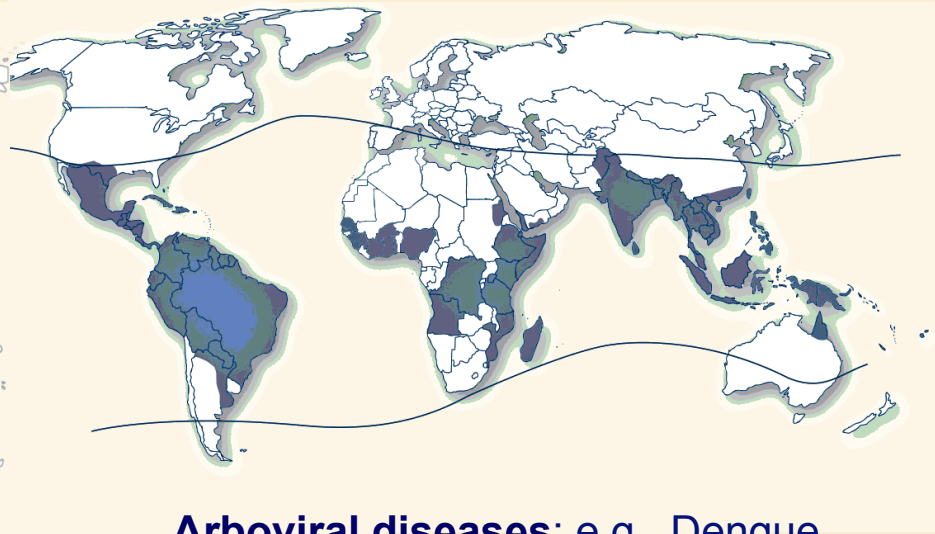


Microsatellites (multi-allelic markers)

Institut Pasteur de Dakar

Diseases under study (*Areas of ongoing transmission)*





Malaria

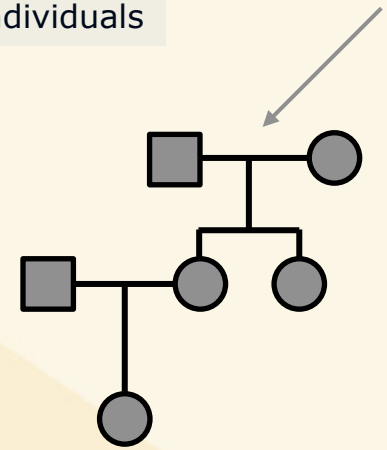
Arboviral diseases: e.g., Dengue

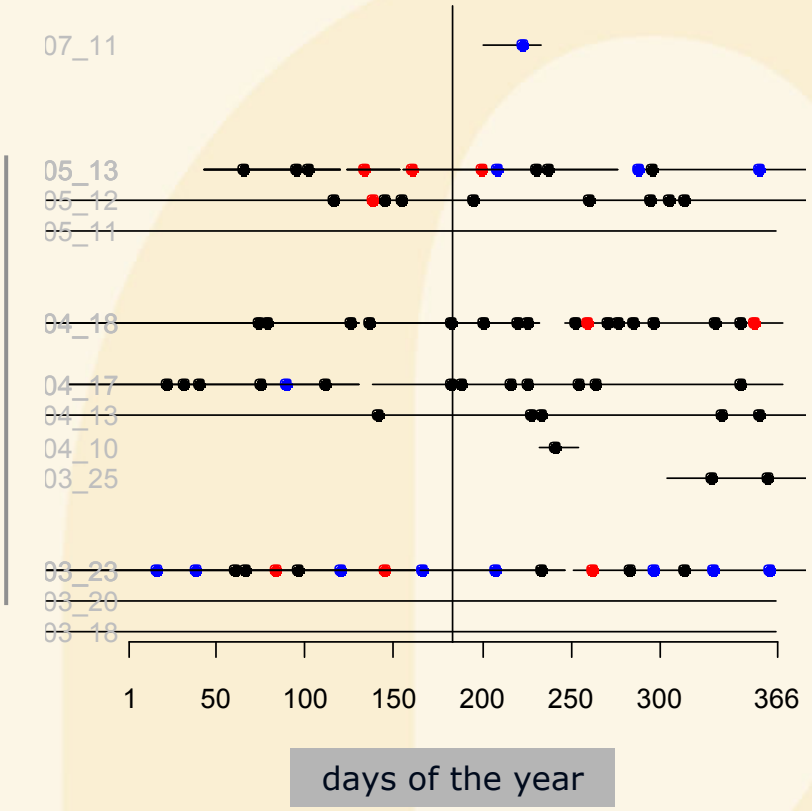
***WHO*, 2013**

Data design

Data design : Longitudinal familial data
(within and between individual correlations)

Related individuals





- Not infected
- Asymptomatic infection
- Symptomatic infections

Institut Pasteur de Dakar

Data collected

	Datasets				
	Phenotypic	**Environment**	**Human genetics**	**Familial**	**Spatial**
Site 1: Dielmo - Ndiop: **Since 1990** **~60 000 clinical episodes on ~1 750 individuals**	▪ Related to malaria ▪ Related to arboviruses: - Zika - West Nile - Yellow Fever - Dengue	▪ Distances to mosquitos breeding sites ▪ On climate ▪ On vectors	▪ Genome-wide SNPs and microsatellites ▪ Gene expression profiles (available within partners or on public databases)	▪ Familial relationships	▪ GPS coordinates of houses
Kedougou: **Since 2009** **~18 000 clinical episodes**	▪ Related to malaria ▪ Related to arboviruses	▪ On climate (rainfall, temperature) ▪ On vectors	NA	NA	▪ GPS coordinates of villages

Institut Pasteur de Dakar

Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

Data: *N* **= 246** individuals
X-vars: **424** microsatellites markers
719,656 SNPs
Y-vars: **4 viruses** (Zika, Yellow Fever, West Nile, Dengue)

Family	Indiv	Father	Mother	Sex	Dengue	M1	M2	M3	…
D1	D1423	X	X	1	1	8 3	9 7	7 7	…
D1	D1424	D99001	D99002	2	0	8 3	9 6	1 1	…
D1	D1425	D99005	D1424	2	0	8 3	6 4	7 1	…
D1	D1426	D1423	D1424	2	1	3 3	7 6	7 1	…
D1	D1427	D1423	D1424	2	1	8 8	9 9	7 1	…
D1	D1430	D1423	D1424	1	0	8 3	9 7	7 1	…
D1	D1433	D1423	D1424	2	1	8 3	9 7	7 1	…
D1	D1437	D1423	D1424	1	0	8 3	9 6	7 1	…
D1	D1441	D99004	D99003	1	0	8 2	9 9	7 1	…
D1	D99001	X	X	1	1				…
D1	D99002	X	X	2	1				…
D1	D99003	D99001	D99002	2	0				…
D1	D99004	X	X	1	1				…
D2	D2201	D99007	D2208	1	0	8 3	9 4	7 5	…
D2	D2202	X	X	2	0				…
…	…	…	…	…	…	…	…	…	…

Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

Software: **MERLIN** (developer: Center of **Statistical Genetics,** Michigan University
Coding language: C++
Interface: command line

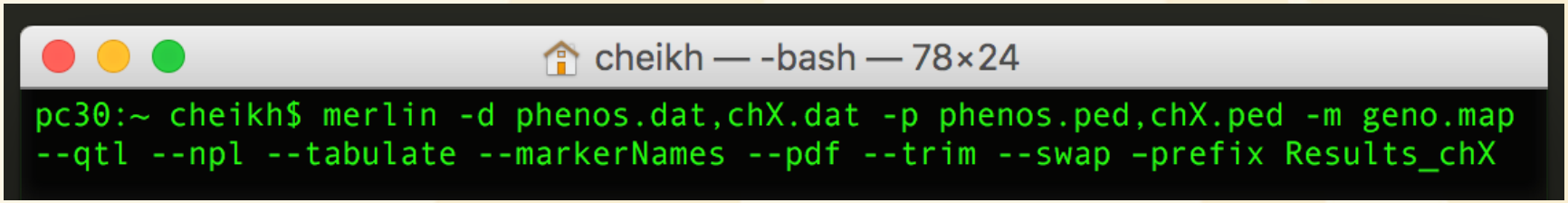
Useful to screen dense genetic markers (variables)
Allows identification of genomic region susceptible to **explain trait variability**

With reduced number of variables
- More statistical power
- Less resource needs
- Quick run

Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

Example de run
Inputs data file: 720,080 genetic markers (719656+424) ~ 2.1GB

Commande line:



Performance options

- ***--trim*** *: To drop non-informative individual from the analysis*
- ***--swap****: using exchange files to not overload RAM memory*

Institut Pasteur de Dakar

Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

On good office lap top

Hardware Overview:

Model Name:	MacBook Pro
Model Identifier:	MacBookPro12,1
Processor Name:	Intel Core i5
Processor Speed:	2.7 GHz
Number of Processors:	1
Total Number of Cores:	2
L2 Cache (per Core):	256 KB
L3 Cache:	3 MB
Memory:	8 GB
Boot ROM Version:	MBP121.0167.B15
SMC Version (system):	2.28f7
Serial Number (system):	C02PTB5DFVH5
Hardware UUID:	03772EDE-AA15-5D4F-B092-AD471E2F396A

25 hours to analyze chromosome 1 only (~ 4.5% of input data)

Time : ~22 days needed for one complete run

Not reasonable,
Particularly when testing several setting

Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

On collaborators' servers (Paris)

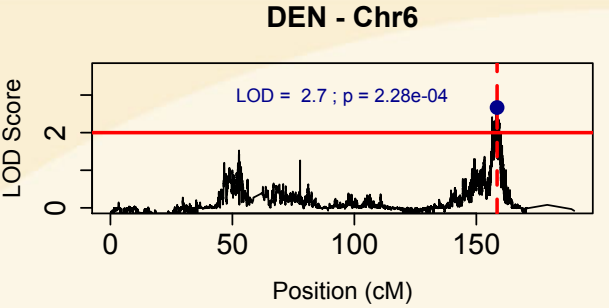
configuration:

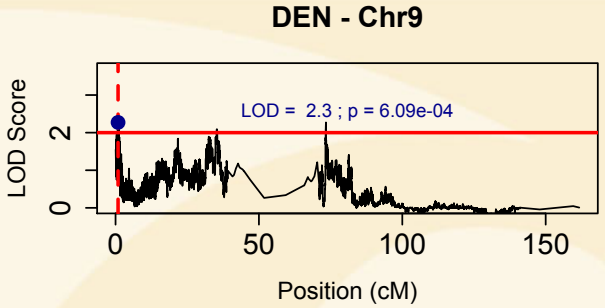
- SSD hard drive
- 400GB de RAM
- 80 core Intel

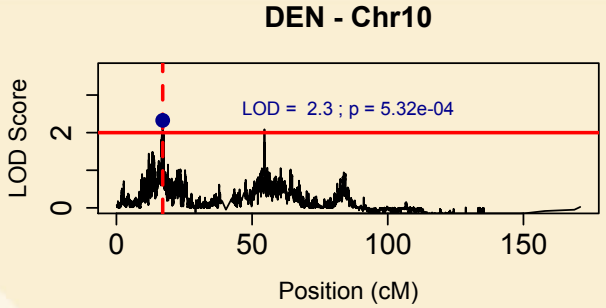
Parallelized on different cores

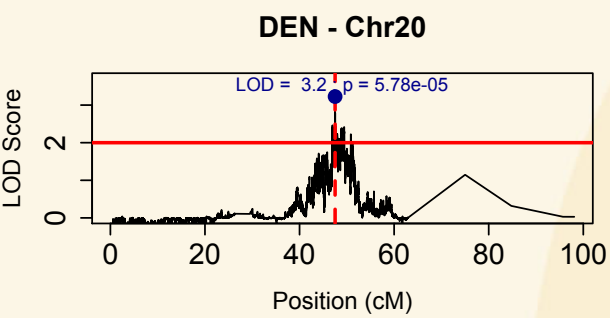
Time: 2 days for one complete run

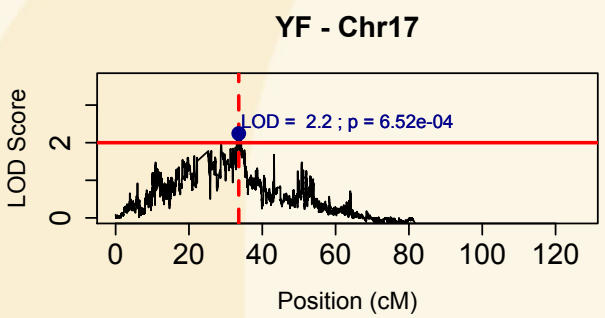
Example of analysis 1:
Genome-wide linkage using SNPs & Microsatellites

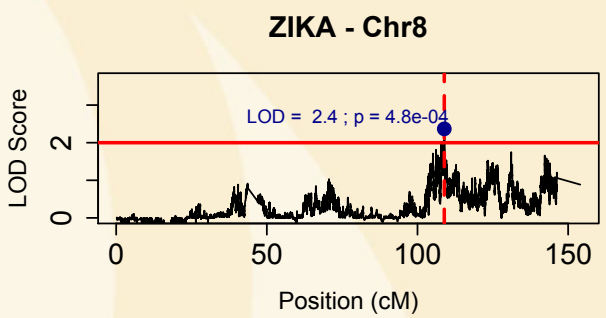


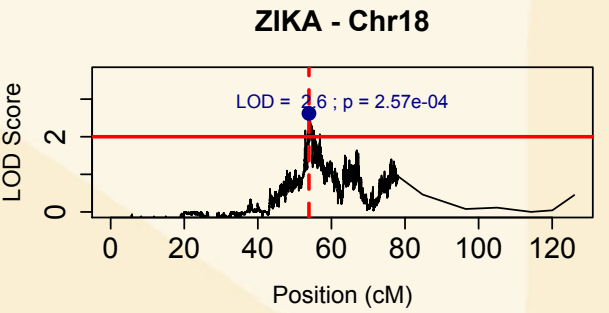


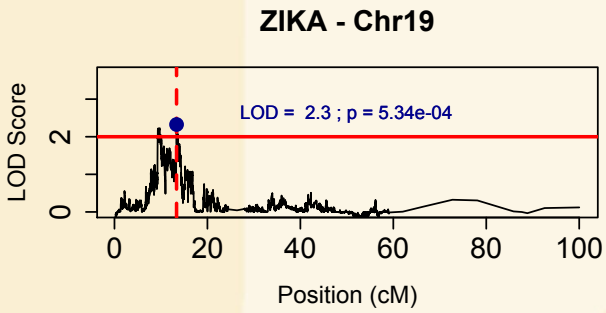


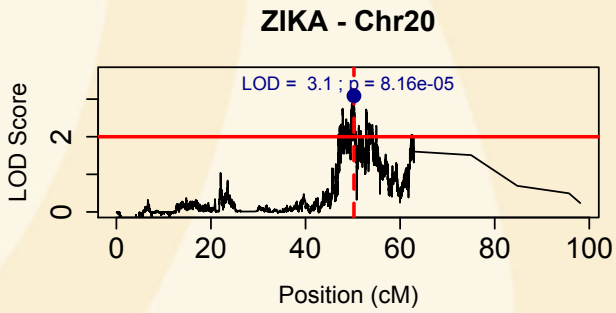








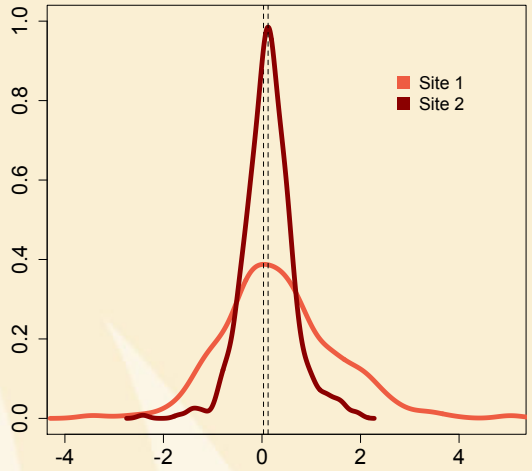




Institut Pasteur de Dakar

Example of analysis 2:
Genome-wide association studies (GWAS)

Data: *N* = **481** individuals (34 founders, 447 offspring)
X-vars: **719,656** SNPs
Y-vars: **1 quantitative trait** (Malaria risk score)



village	family	indiv	father	mother	sex	Risk_Score	rs13137	rs10128	rs13541	rs14073	rs13399
1	8	**D0101**	0	0	1	3,872	-	-	-	-	-
1	8	**D0102**	0	D0307	2	2,014	A A	A A	G G	A A	A G
1	8	**D0103**	0	D0102	1	5,094	A A	A C	G G	A A	G G
1	8	**D0104**	0	D2005	2	2,566	A A	C C	G G	G A	A G
1	8	**D0105**	D0103	D0104	1	0,372	A A	C C	G G	G A	G G
1	8	**D0107**	D0101	0	1	0,560	A A	C C	G G	A A	G G
1	8	**D0108**	D0101	0	1	-0,298	-	-	-	-	-
1	8	**D0109**	D0101	0	2	0,998	A A	C C	G G	G A	G G
1	8	**D0110**	D1812	D0109	1	0,232	A A	A C	G G	A A	G G
1	8	**D0111**	D0313	D0314	2	0,433	-	-	-	-	-
1	8	**D0112**	D0103	D0104	1	0,532	A A	A C	G G	G A	A G
1	8	**D0113**	D1812	D0109	2	0,760	A A	A C	G G	G A	G G
1	8	**D0114**	D0103	D0104	1	-0,169	A A	A C	G G	G A	A G

Institut Pasteur de Dakar

Example of analysis 2:
Genome-wide association studies (GWAS)

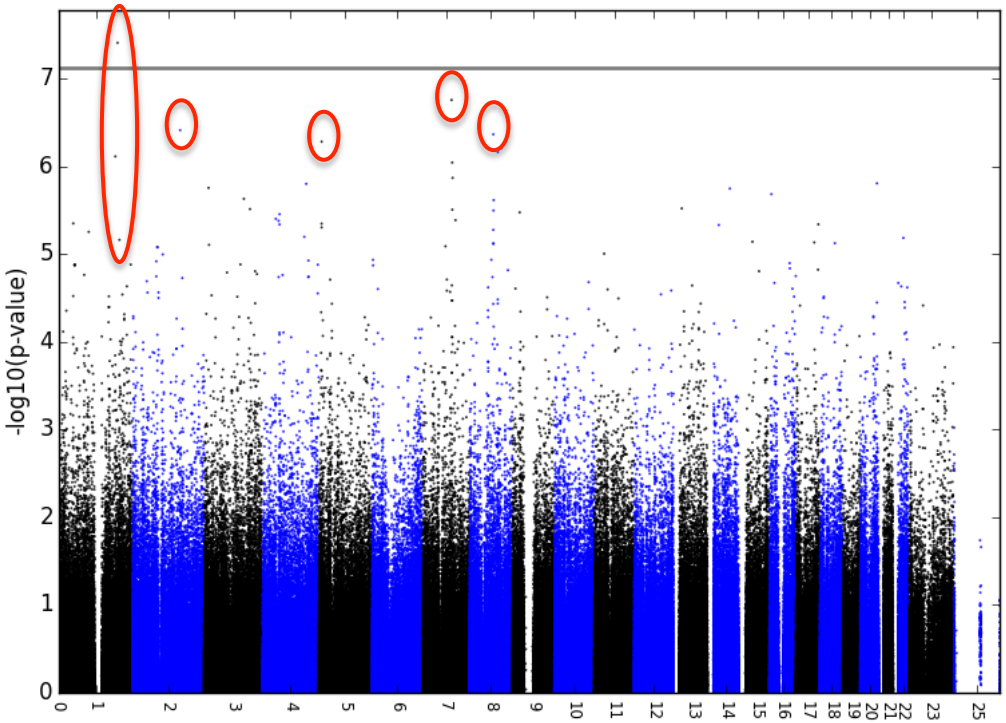
On a good office lap top

Hardware Overview:

Model Name:	MacBook Pro
Model Identifier:	MacBookPro12,1
Processor Name:	Intel Core i5
Processor Speed:	2.7 GHz
Number of Processors:	1
Total Number of Cores:	2
L2 Cache (per Core):	256 KB
L3 Cache:	3 MB
Memory:	8 GB
Boot ROM Version:	MBP121.0167.B15
SMC Version (system):	2.28f7
Serial Number (system):	C02PTB5DFVH5
Hardware UUID:	03772EDE-AA15-5D4F-B092-AD471E2F396A

Reasonable run time
~ 40 minutes for complete run
(including Quality Control)

Results: Manhattan plot



Institut Pasteur de Dakar

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

Design

- N independent "parents – affected offspring" trios
- K multi-allelic markers to test simultaneously: $L^1, L^2, \ldots, L^K$
  having respectively: $l_1, l_2, \ldots, l_K$ alleles

  e.g., marker locus L^i has l_i alleles denoted $a_1^i, a_2^i, \ldots, a_{l_i}^i$

Remarks

- Maximum number of observed combinations of alleles, sampling one allele from each of the K loci, (K-tuples of alleles) is:

$$\prod_{i=1}^{K} l_i = l_1 \times l_2 \times \ldots \times l_K$$

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

On the data:

$$\log L\left(\alpha^1,\alpha^2,...,\alpha^K\right)=\sum_{S\neq S'} n_{SS'}\times\log\left(p_{SS'}\right)+\sum_{S\neq S'}(m_{SS'}-n_{SS'})\times\log\left(1-p_{SS'}\right)$$

Where $\alpha^i=\begin{pmatrix}\alpha_a^i\\ \vdots\\ \alpha_v^i\end{pmatrix}$ transmission intensities for each allele "*j* " at each locus *i* among heterozygous parents bearing the "*j* " allele.

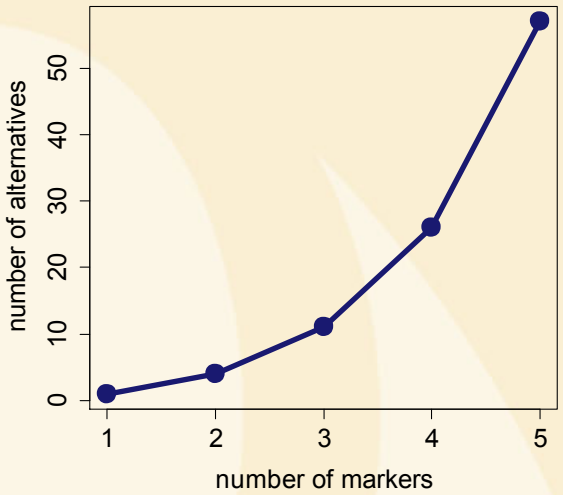
Under the null hypothesis of 50/50 risk: **($p_{SS'} = 1 - p_{SS'} = 0.5$)**

$$\log L_0=-\log(2)\times\sum_{S\neq S'} m_{SS'}$$

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

Considering *K* loci simultaneously leads to *m* different models to test:

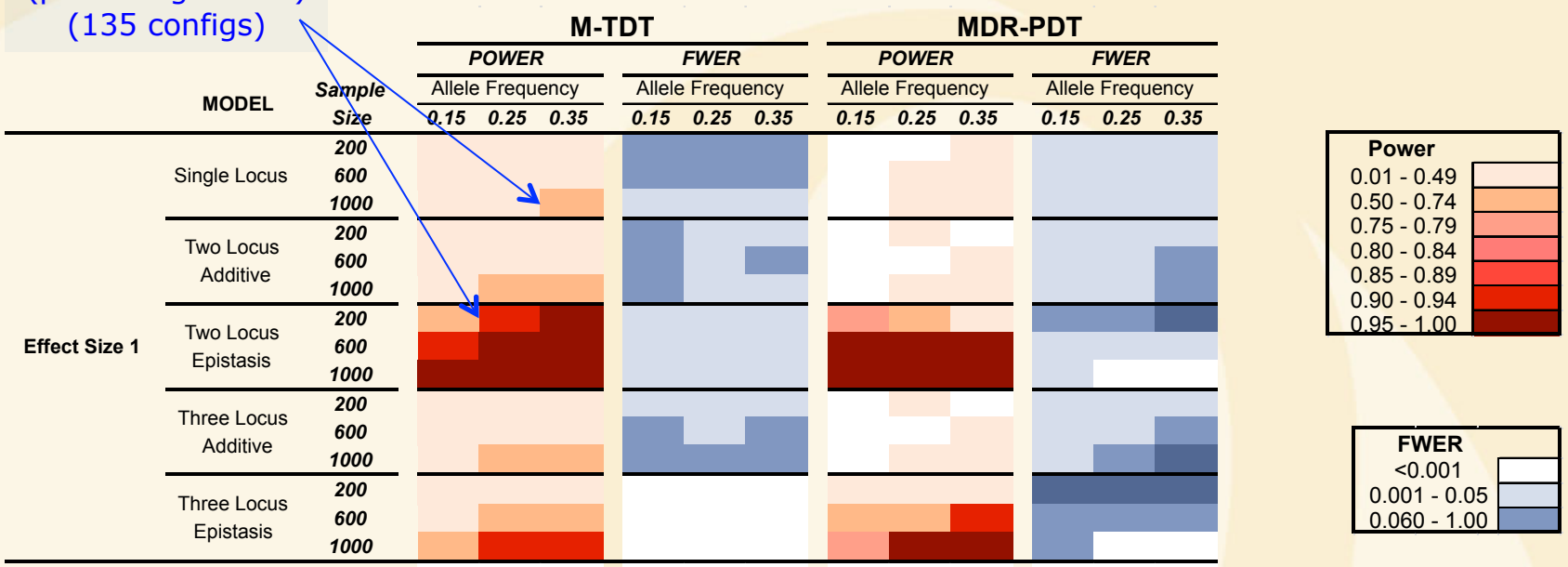
$$m = K + 2 \sum_{q=2,\ldots,K} \binom{K}{q}$$
$$= 2^{K+1} - K - 2$$



➔ P-values are corrected to control the FDR (*e.g., Benjamini and Hochberg's, 1995*)

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

1000 simulates dataset per cell (per configuration) (135 configs)



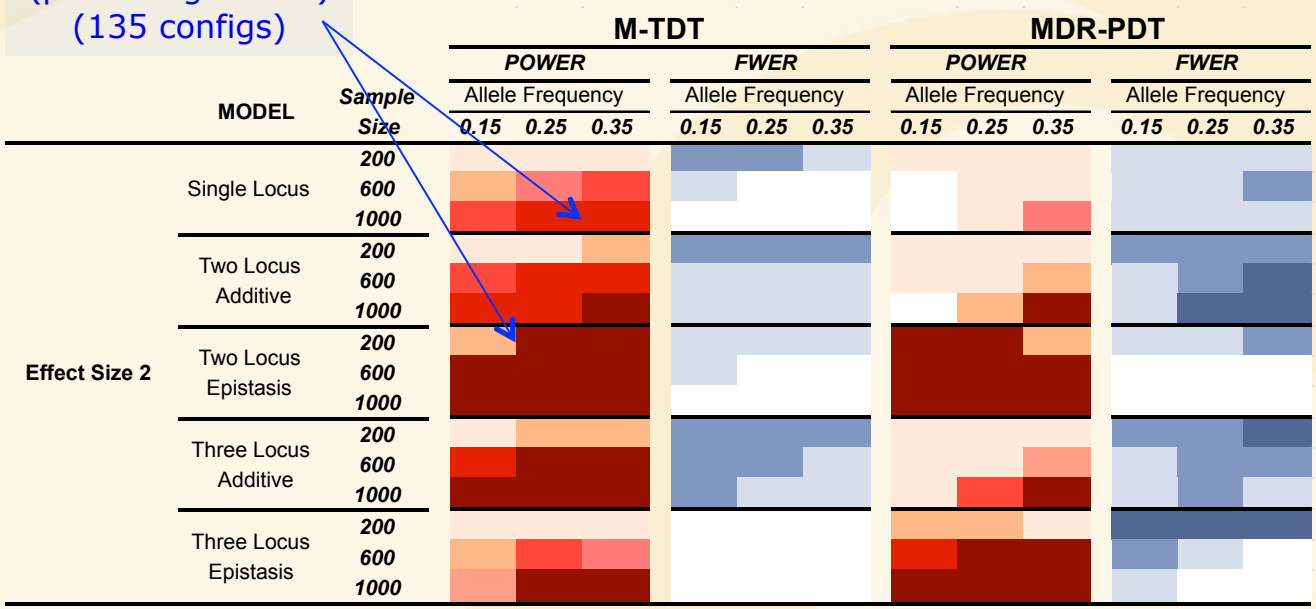
M-TDT (*Loucoubar C, et al., under review, 2016*)

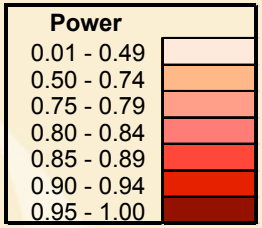
MDR-PDT (*Martin ER, Ritchie MD, et al., 2006*)

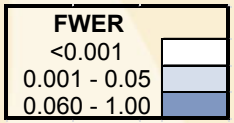
Institut Pasteur de Dakar

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

1000 simulates dataset per cell (per configuration) (135 configs)







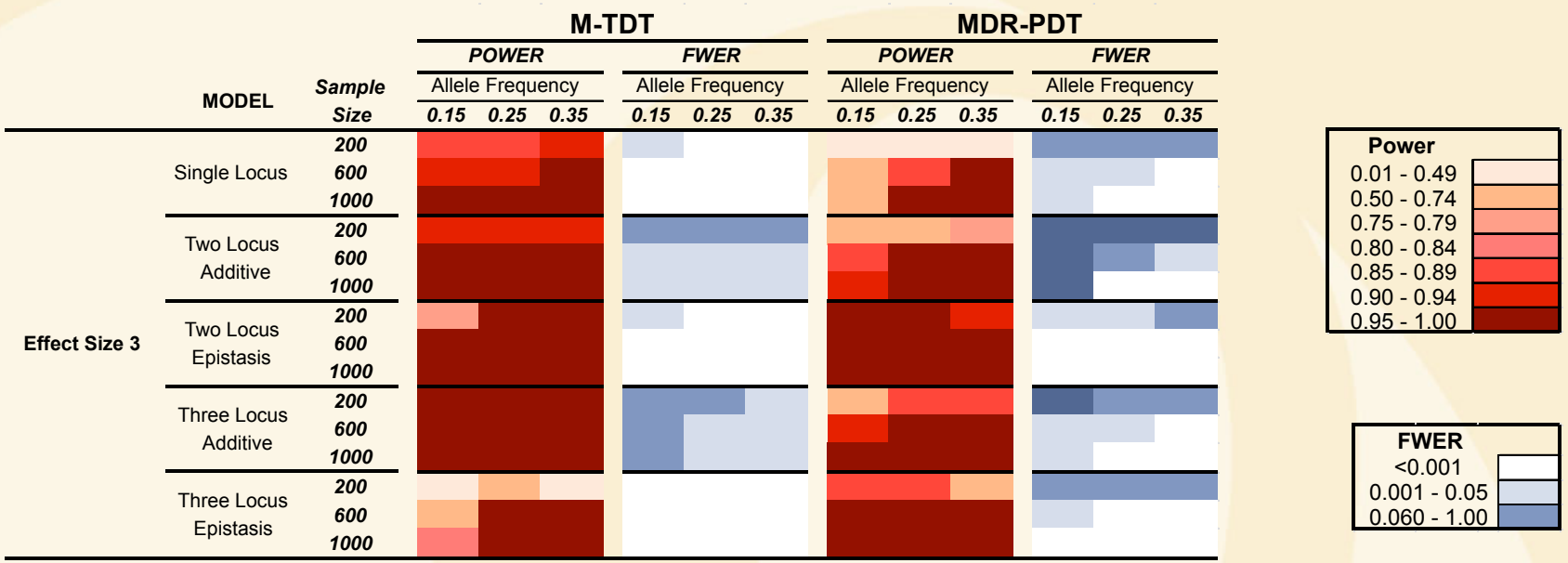
For **single-locus** scenarios: M-TDT more Power ++, lower FWER ++

For **multi-locus** scenarios: Additive: M-TDT more Power ++, lower FWER ++

Epistasis: MDR-PDT more Power ++, higher FWER --

Institut Pasteur de Dakar

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests



Institut Pasteur de Dakar

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

Mean **Power** and **FWER** across all scenarios

	Power	FWER
M-TDT	0.88	0.03
MDR-PDT	0.57	0.07

Not reasonable to run on a classical computer
135,000 datasets to analyse
11 models per dataset

Run on IP Paris cluster on ~4 days

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

	Malaria resistance
Number of individuals in the analyzed sample :	276
Number of parents :	135
Number of offspring :	147
Number offspring also having the parent status :	6
Number of independent families :	11
Number of nuclear families :	73
Mean number of offspring per nuclear family :	2
Number of nuclear families with 1 *offspring :*	28
2 *offspring :*	26
3 *offspring :*	10
4 *offspring :*	8
5 *offspring :*	1

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

									M-TDT		FBAT	
Gene symbol	**Chr**	**rs Number**	**Position (GRCh38)**	**M/m**	**MAF**	***N***	***T***	***NT***	***nominal* P-value**	***empirical* P-value**	***nominal* P-value**	***empirical* P-value**
Pre-selected SNPs												
CR1	1	rs17047661	207609544	A/G	0.21	108	57.8	40.1	0.07	0.139	0.15	0.14
HBB	11	rs334	5227002	T/G	0.06	50	39.3	19.1	0.008	0.047	0.05	0.04
Screened SNP												
UBTF	17	rs9910055	44205669	C/T	0.4	15	4.3	11.5	0.07	0.067	0.12	0.12
UBTF	17	rs2071167	44210151	A/G	0.36	5	1.6	4.7	0.22	0.259	0.32	0.31
UBTF	17	rs9901595	44228331	A/G	0.17	3	2.6	0.3	0.15	0.484	0.3	0.50
UBTF	17	rs9906669	44228338	A/G	0.2	3	0.4	3.1	0.12	0.535	0.3	0.50
SLC4A1	17	rs8066822	44247988	C/T	0.33	7	5.6	2	0.18	0.291	0.28	0.26
SLC4A1	17	rs1465204	44250109	C/T	0.2	7	2.1	5.2	0.25	0.347	0.35	0.35
SLC4A1	17	rs2072081	44250125	A/C	0.35	7	5.6	2	0.18	0.291	0.28	0.26
SLC4A1	17	rs45497993	44250506	T/C	0.33	152	52.2	89.2	0.002	0.013	0.016	0.02
SLC4A1	17	rs2857078	44252803	G/T	0.39	135	60.1	66.3	0.58	0.644	0.67	0.70
SLC4A1	17	rs45530735	44256644	A/G	0.25	106	42.6	55.2	0.2	0.302	0.32	0.34
SLC4A1	17	rs2074108	44258781	A/G	0.17	7	5.3	2.1	0.24	0.394	0.35	0.38
SLC4A1	17	rs2857082	44259723	A/G	0.3	15	10.2	5.6	0.24	0.262	0.31	0.30
SLC4A1	17	rs2074107	44260608	C/T	0.25	3	1.7	1.6	0.98	1	0.98	1
SLC4A1	17	rs5036	44261577	A/G	0.08	5	4.8	0.3	0.03	0.067	0.1	0.06
SLC4A1	17	rs2074106	44261935	A/C	0.16	84	28.8	60.8	6.3×10^{-4}	0.007	0.01	0.01
SLC4A1	17	rs9916116	44268611	A/G	0.46	11	8.7	3.5	0.13	0.164	0.21	0.20
RUNDC3A	17	rs7222501	44277289	C/T	0.45	7	4.3	2.8	0.58	0.612	0.62	0.62
RUNDC3A	17	rs6503366	44280405	A/G	0.26	10	3.6	6.6	0.33	0.455	0.41	0.46
RUNDC3A	17	rs2879165	44283992	A/G	0.33	10	6.6	3.6	0.33	0.439	0.41	0.44
RUNDC3A	17	rs10852960	44302585	A/G	0.18	10	4	5.3	0.66	0.702	0.69	0.72
RUNDC3A	17	rs708386	44315254	A/G	0.22	10	4	5.3	0.66	0.719	0.69	0.72
RUNDC3A	17	rs2074104	44316126	A/G	0.39	11	6.5	3.8	0.39	0.432	0.44	0.46

Example of analysis 3:
Detecting multi-way epistasis in family-based association tests

Association with malaria resistant phenotype

								subset comprised of independent trios			
Genes	**Models**	***N***	**M-TDT statistic**	**DF**	**M-TDT nominal P**	**M-TDT empirical P**	**FDR-adjusted empirical P**	***N***	**M-TDT nominal P**	**M-TDT empirical P**	**FDR-adjusted empirical P**
HBB (**rs334**)	Single	**49**	7.7	1	0.006	0.03	**0.03**	21	0.05	0.14	0.14
SLC4A1 (**rs2074106**)	Single	**84**	11.7	1	6.2×10^{-4}	0.007	**0.01**	45	0.01	0.04	0.08
CR1 (**rs17047661**)	Single	**95**	4.6	1	0.03	0.07	0.07	49	0.03	0.06	0.08
HBB, SLC4A1	Additive	**115**	17.2	2	1.9×10^{-4}	0.005	**0.01**	60	0.01	0.06	0.08
HBB, CR1	Additive	**124**	13.3	2	0.001	0.01	**0.02**	62	0.01	0.05	0.08
SLC4A1, CR1	Additive	**148**	14.1	2	8.7×10^{-4}	0.009	**0.01**	75	0.01	0.04	0.08
HBB, SLC4A1	Epistasis	**115**	15	3	0.002	0.01	**0.02**	60	0.04	0.11	0.12
HBB, CR1	Epistasis	**124**	19.3	3	2.4×10^{-4}	0.002	**0.008**	62	0.004	0.02	0.06
SLC4A1, CR1	Epistasis	**148**	27.7	3	4.2×10^{-6}	8×10^{-5}	**4.4×10^{-4}**	75	1.8×10^{-5}	2.8×10^{-4}	**0.003**
HBB, SLC4A1, CR1	Additive	**173**	20.6	3	1.3×10^{-4}	0.004	**0.01**	93	0.01	0.04	0.08
HBB, SLC4A1, CR1	Epistasis	**173**	39.4	7	1.6×10^{-6}	3×10^{-5}	**3.3×10^{-4}**	93	1.3×10^{-4}	0.001	**0.01**

N: number of informative transmissions

Acknowledgments

THANK YOU