

STATISTICAL ANALYSIS PIPELINE ADMIXTURE DATA FROM A HUMAN POPULATION SETTLEMENT MODEL

Aurélien LUCIANI
Master 2 Bioinformatique
Université de Bordeaux
11 September 2015



INTRODUCTION

- ▶ Internship at Massey University (New Zealand)
- ▶ Computational Biology Research Group
- ▶ Part of a 2 years project
- ▶ Supervisors:
 - ▶ Murray Cox
 - ▶ Marie Noelle Beurton-Aimar

TABLE OF CONTENTS

Context

- Geological and anthropological context
- Previous papers
- Model

Measures and comparisons

Pipeline

Statistical analysis





**FIRST SETTLEMENT WAVE
AROUND 45,000 YEARS AGO**



SECOND SETTLEMENT WAVE
AROUND 4,500 YEARS AGO

CONTEXT

PREVIOUS PAPERS

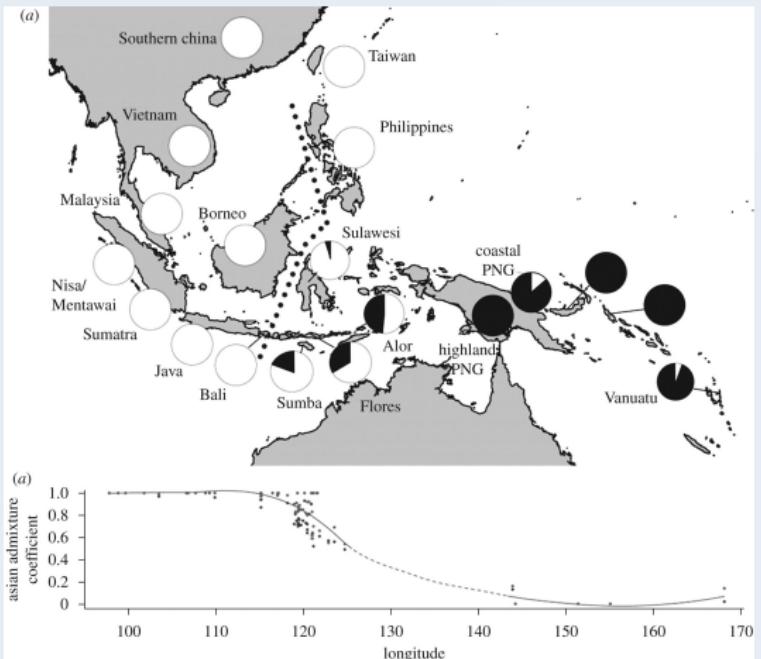


Figure : Cox et al. [2010]

CONTEXT

PREVIOUS PAPERS

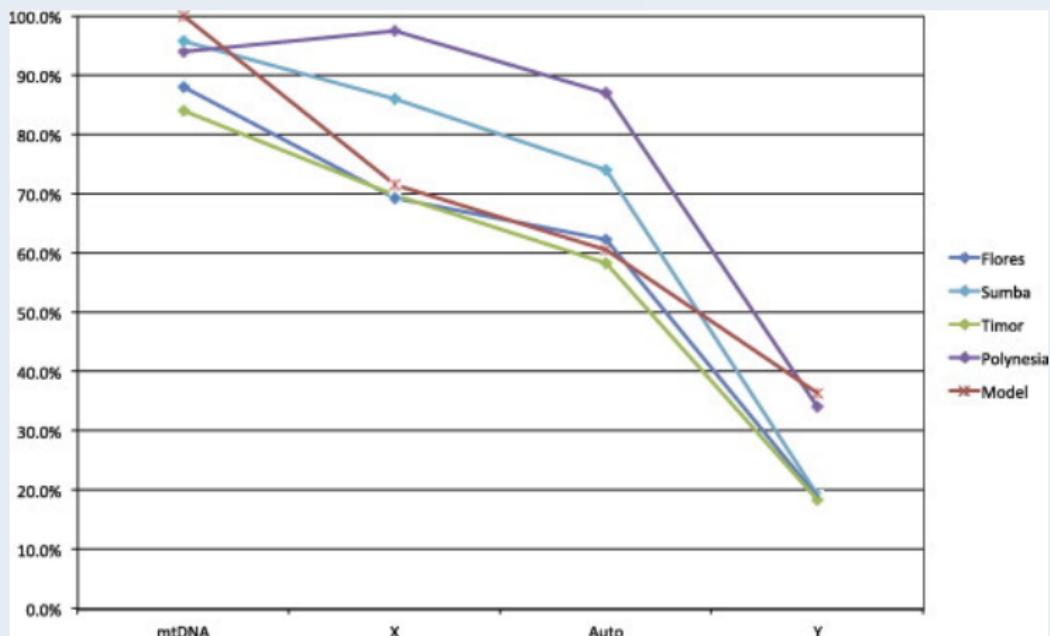
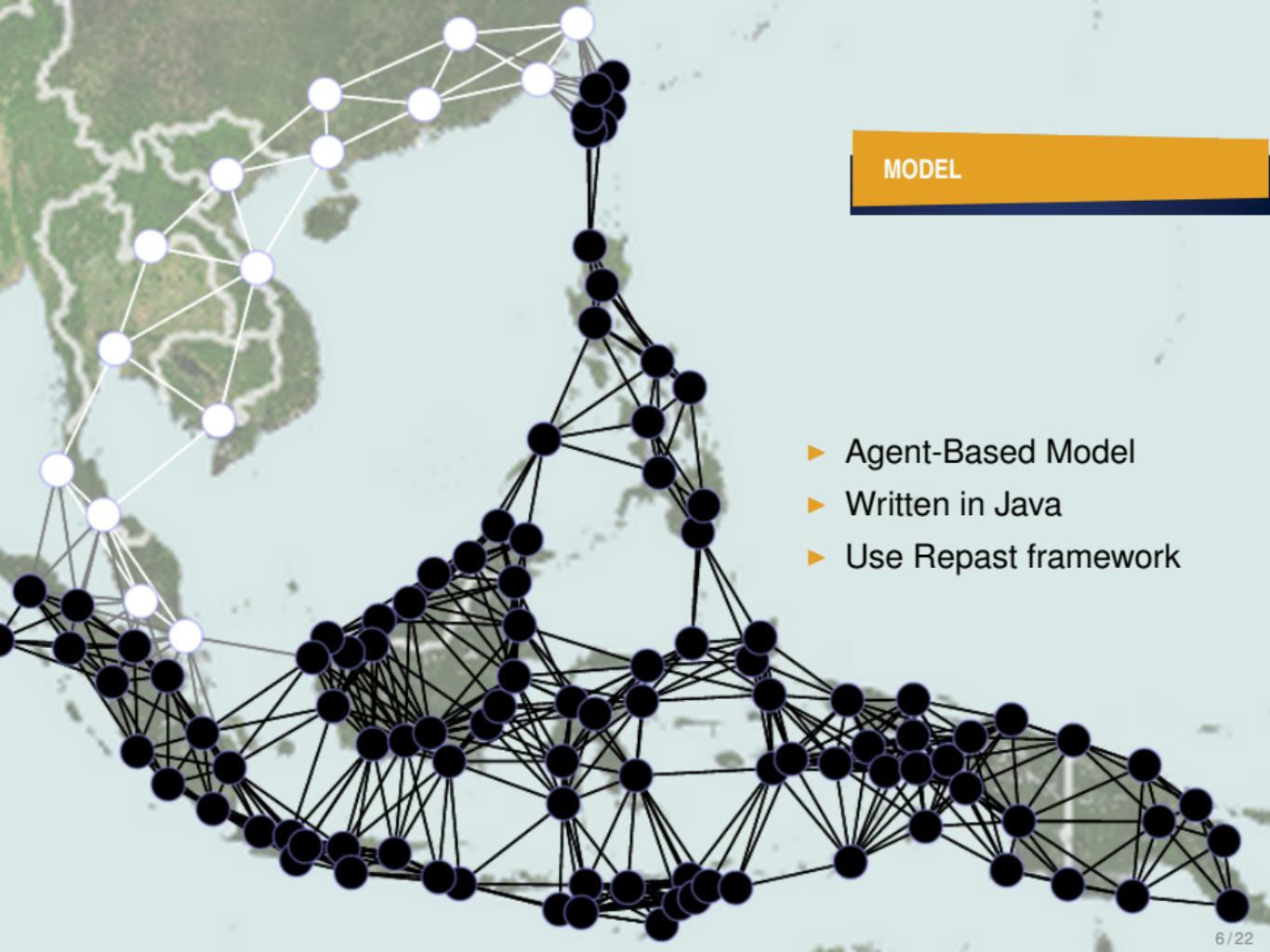


Figure : Lansing et al. [2011]



MODEL

- ▶ Agent-Based Model
- ▶ Written in Java
- ▶ Use Repast framework

TABLE OF CONTENTS

Context

Measures and comparisons

- Observed data

- Parameters

- Comparison functions

 - Mean Square Distance

 - Partial Mantel Correlation

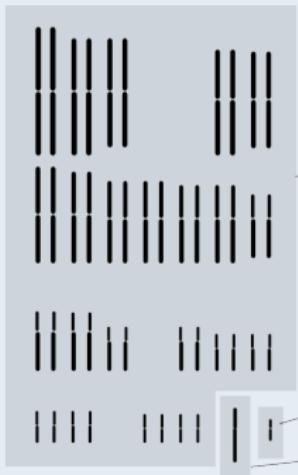
 - Summary statistics

Pipeline

Statistical analysis

MEASURES & COMPARISONS

OBSERVED DATA



Autosome
25 markers

Mitochondria
1 marker

Y Chromosome
1 marker

X Chromosome
25 markers

- ▶ Different parts of the DNA
- ▶ Associated with different origins
- ▶ 52 binary markers in total

MEASURES & COMPARISONS

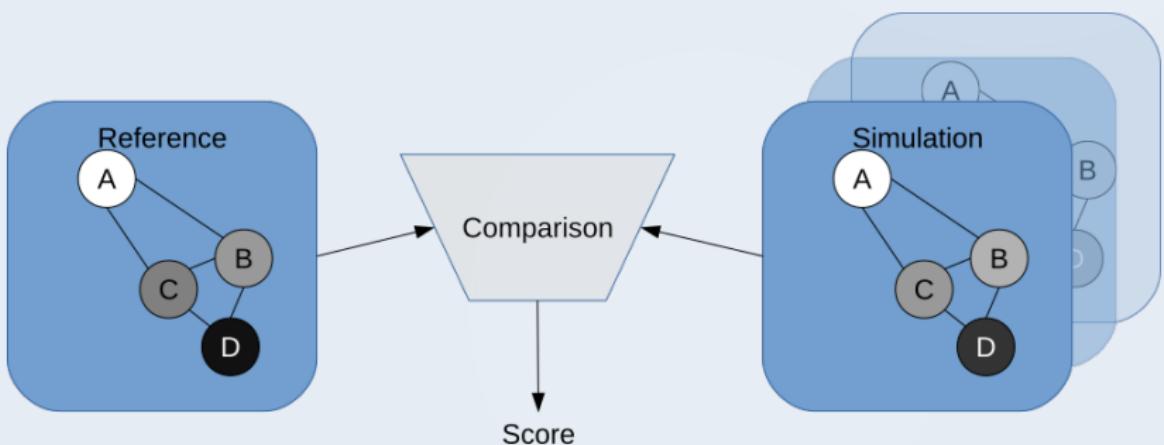
PARAMETERS

Parameter	Estimated	Comment
Migration prob.	$0.0 < x \leq 1.0$	prob. to start migrating for a Melanesian agent
Migration prob. ratio	$1.0 \leq x \leq 4.0$	corresponding ratio for an Asian agent
Fecundity	$2.5 < x < 8.0$	Poisson law mean for a Melanesian agent
Fecundity ratio	$1.0 \leq x \leq 2.0$	corresponding ratio for an Asian agent
Marriage threshold	$0.0 \leq x \leq 0.25$	affects marriages rules
Growth rate	$0.0 < x < 0.001$	limiting rate of pop. growth
Number of agents	$100 \leq x < 400$	pop. size in each deme, initially
Graph	{...}	composition of the graph (nodes and edges)
Starting distribution	{...}	distribution of pop. in the graph

Table : Summary of the changeable model parameters.

MEASURES & COMPARISONS

COMPARISON FUNCTIONS



MEASURES & COMPARISONS

COMPARISON FUNCTIONS

Mean Square Distance

$$MSD = \frac{\sum_{i=1}^n (AdRef_i - AdSim_i)^2}{n}$$

$$0 \leq MSD \leq 1$$

MEASURES & COMPARISONS

COMPARISON FUNCTIONS

Mean Square Distance

$$MSD = \frac{\sum_{i=1}^n (AdRef_i - AdSim_i)^2}{n}$$

$$0 \leq MSD \leq 1$$

Example

Island	A	B	C	D
Reference	1.0	0.5	0.4	0.0
Simulated	1.0	0.4	0.3	0.2
Distance	0.0	0.1	0.1	0.2

$$MSD = 0.1$$

MEASURES & COMPARISONS

COMPARISON FUNCTIONS

Mantel Partial Correlation

$cor = \text{mantel.partial}(M_{\text{Simulated}}, M_{\text{Reference}}, M_{\text{geographical}})$

$$-1 \leq cor \leq 1$$

MEASURES & COMPARISONS

COMPARISON FUNCTIONS

Mantel Partial Correlation

$cor = \text{mantel.partial}(M_{\text{Simulated}}, M_{\text{Reference}}, M_{\text{geographical}})$

$$-1 \leq cor \leq 1$$

Example

$$\text{mantel.partial} \left(\begin{bmatrix} 0.0 & 0.5 & 0.6 & 1.0 \\ 0.5 & 0.0 & 0.1 & 0.5 \\ 0.6 & 0.1 & 0.0 & 0.4 \\ 1.0 & 0.5 & 0.4 & 0.0 \end{bmatrix}, \begin{bmatrix} 0.0 & 0.6 & 0.7 & 0.8 \\ 0.6 & 0.0 & 0.1 & 0.2 \\ 0.7 & 0.1 & 0.0 & 0.1 \\ 0.8 & 0.2 & 0.1 & 0.0 \end{bmatrix}, \begin{bmatrix} 0.0 & 300 & 250 & 400 \\ 300 & 0.0 & 120 & 150 \\ 250 & 120 & 0.0 & 100 \\ 400 & 150 & 100 & 0.0 \end{bmatrix} \right)$$

$$cor = 0.72$$

MEASURES & COMPARISONS

COMPARISON FUNCTIONS

Summary statistics:

- ▶ Mean Square Distance
 - ▶ Autosomal admixture
 - ▶ X-Chromosomal admixture
- ▶ Partial Mantel Correlation
 - ▶ Autosomal admixture
 - ▶ X-Chromosomal admixture

TABLE OF CONTENTS

Context

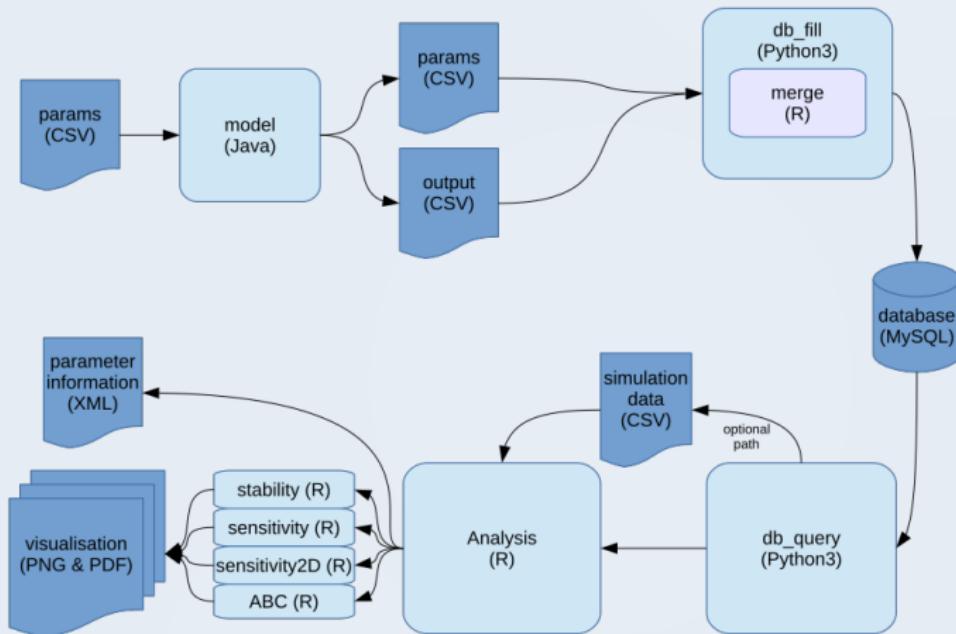
Measures and comparisons

Pipeline
Overview

Statistical analysis

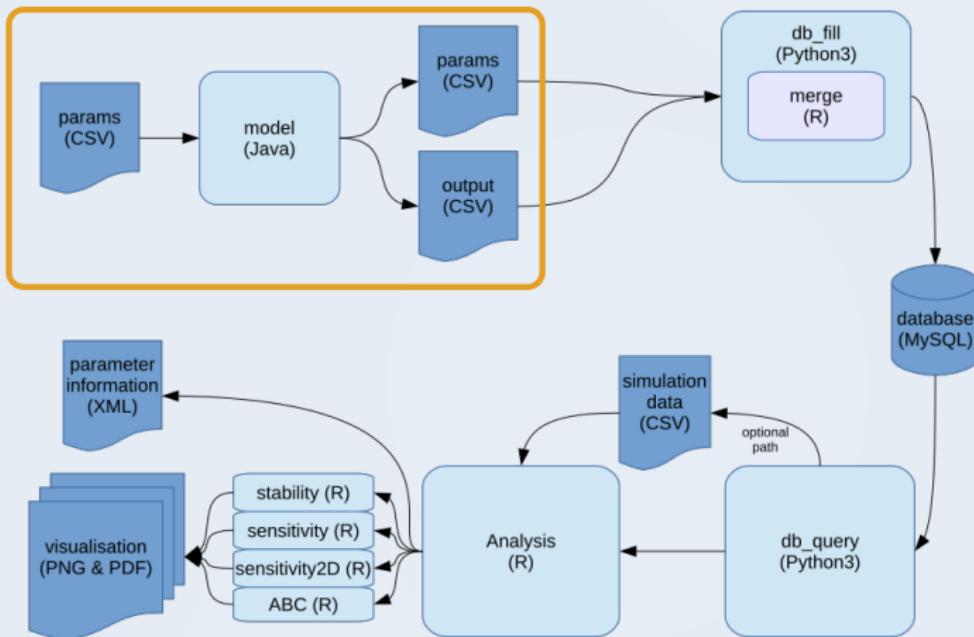
PIPELINE

OVERVIEW



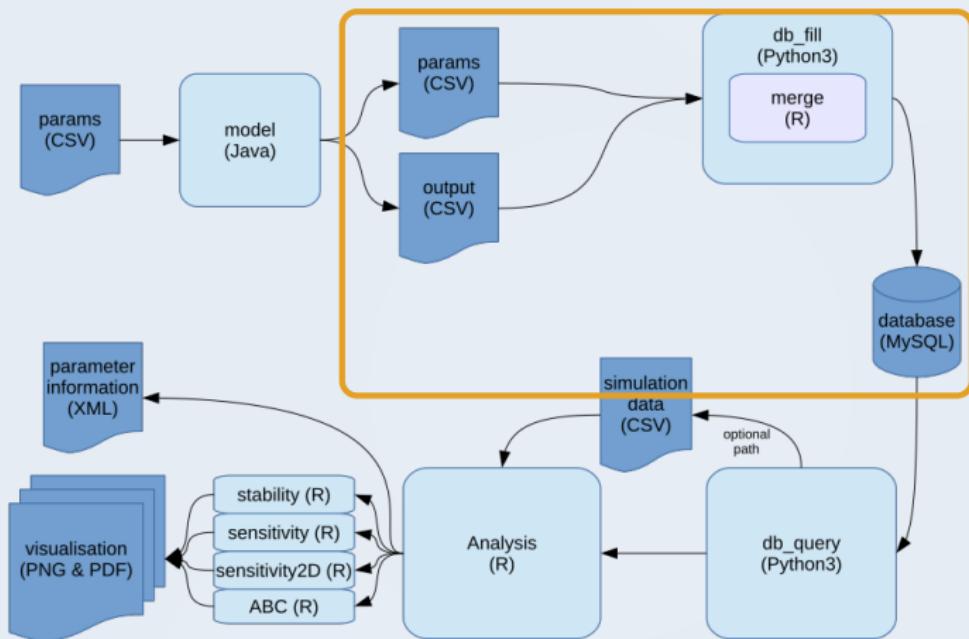
PIPELINE

OVERVIEW



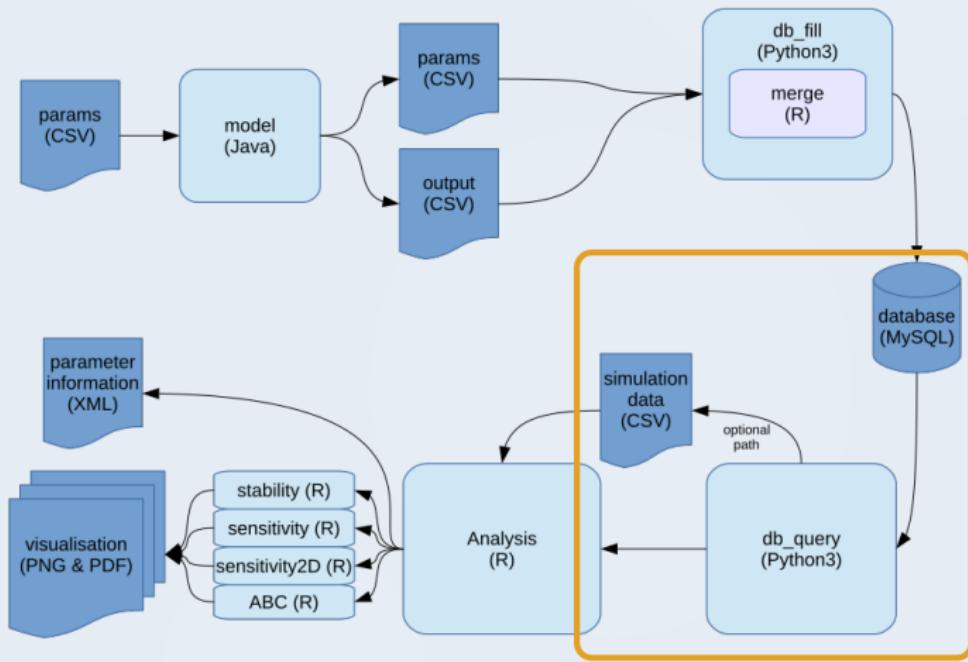
PIPELINE

OVERVIEW



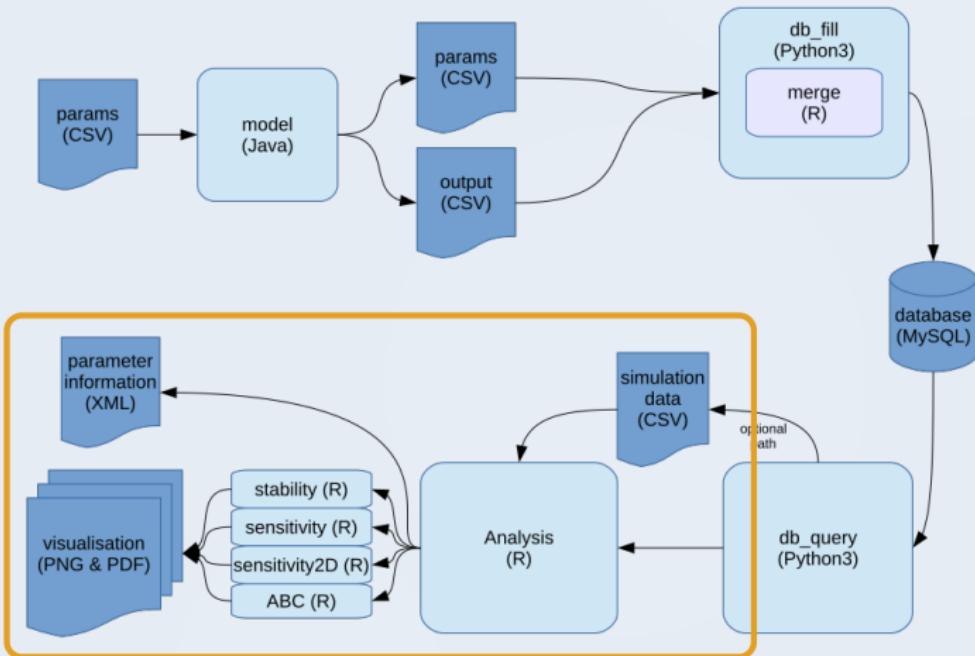
PIPELINE

OVERVIEW



PIPELINE

OVERVIEW



PIPELINE

OVERVIEW

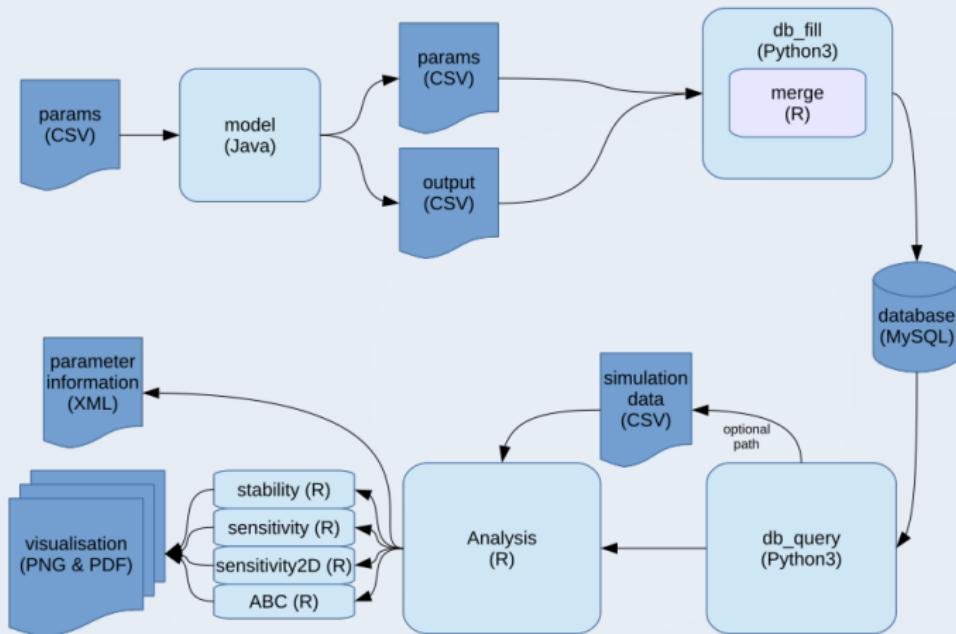


TABLE OF CONTENTS

Context

Measures and comparisons

Pipeline

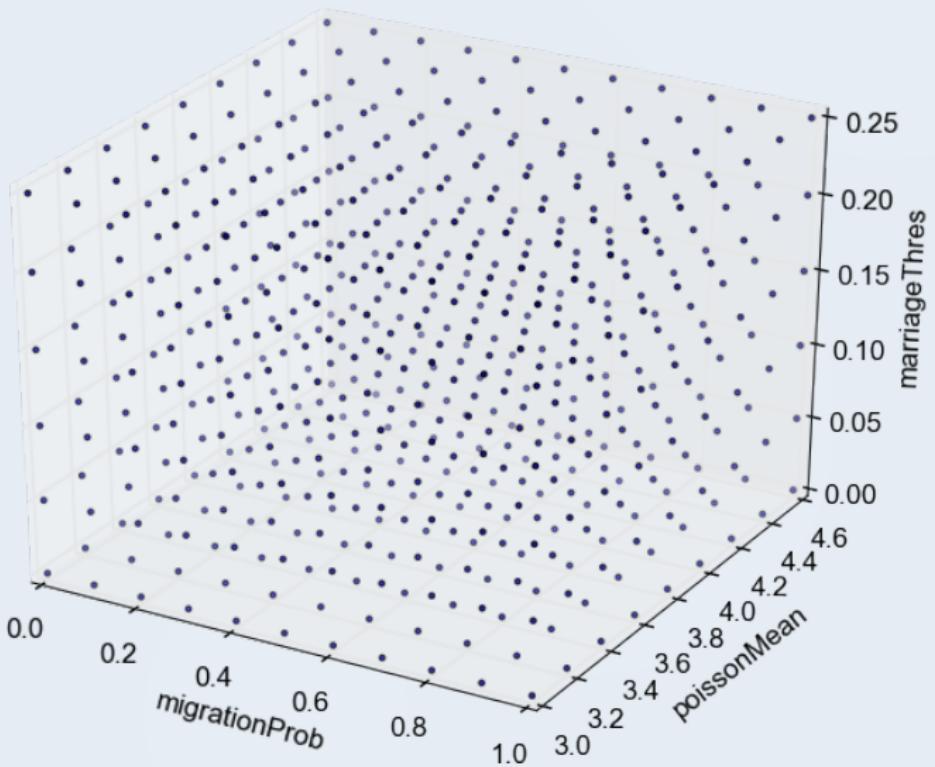
Statistical analysis

Grid search

Approximate Bayesian Computation

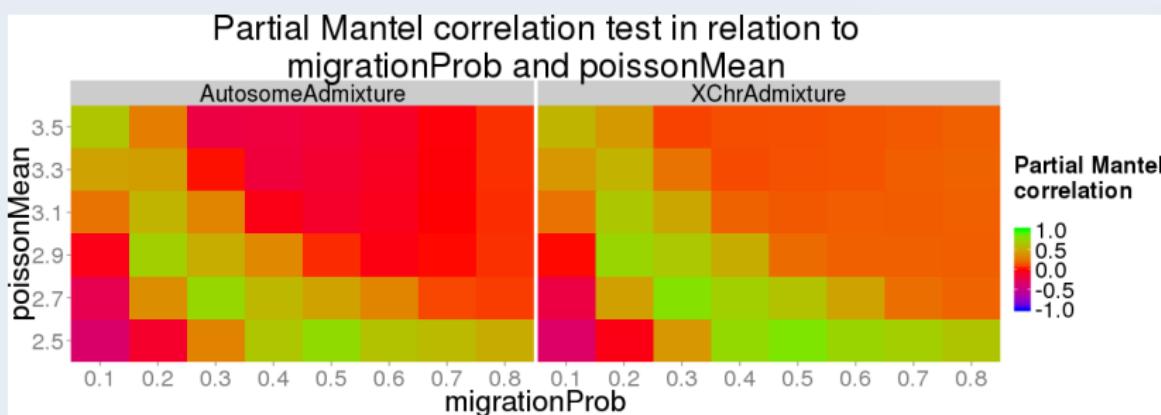
STATISTICAL ANALYSIS

GRID SEARCH



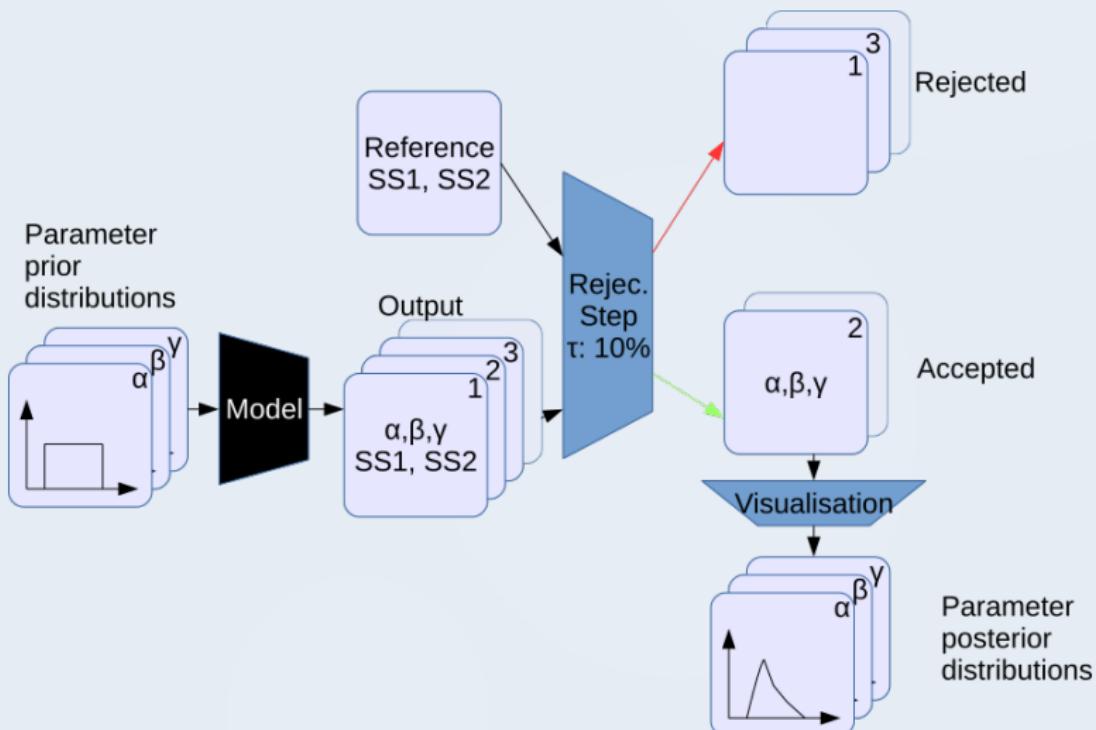
STATISTICAL ANALYSIS

GRID SEARCH



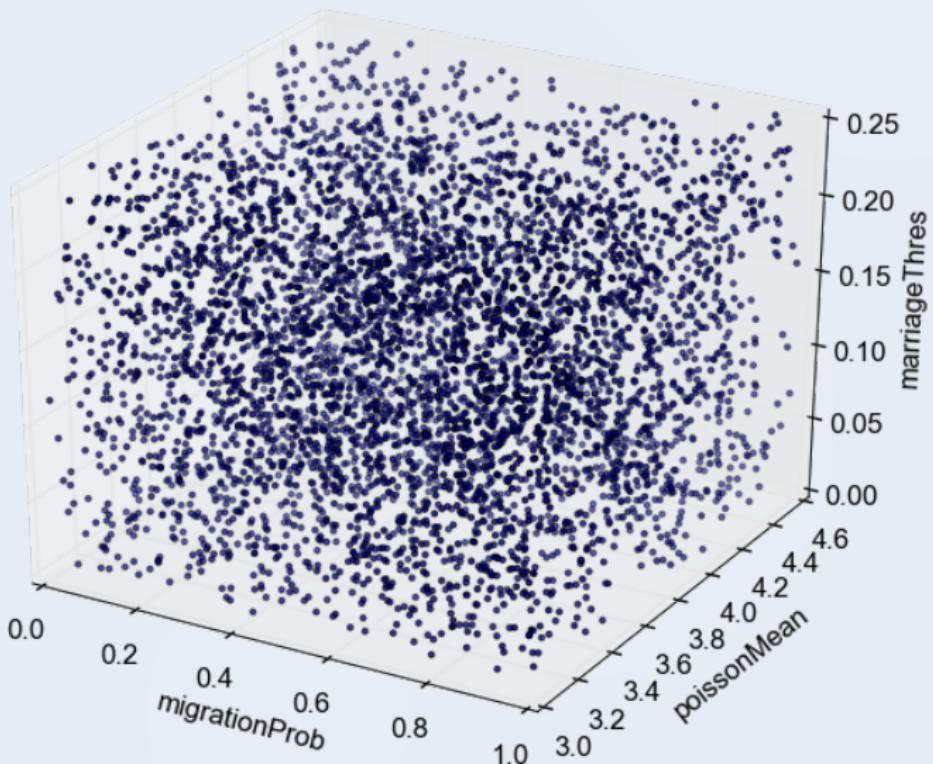
STATISTICAL ANALYSIS

APPROXIMATE BAYESIAN COMPUTATION



STATISTICAL ANALYSIS

APPROXIMATE BAYESIAN COMPUTATION



STATISTICAL ANALYSIS

APPROXIMATE BAYESIAN COMPUTATION

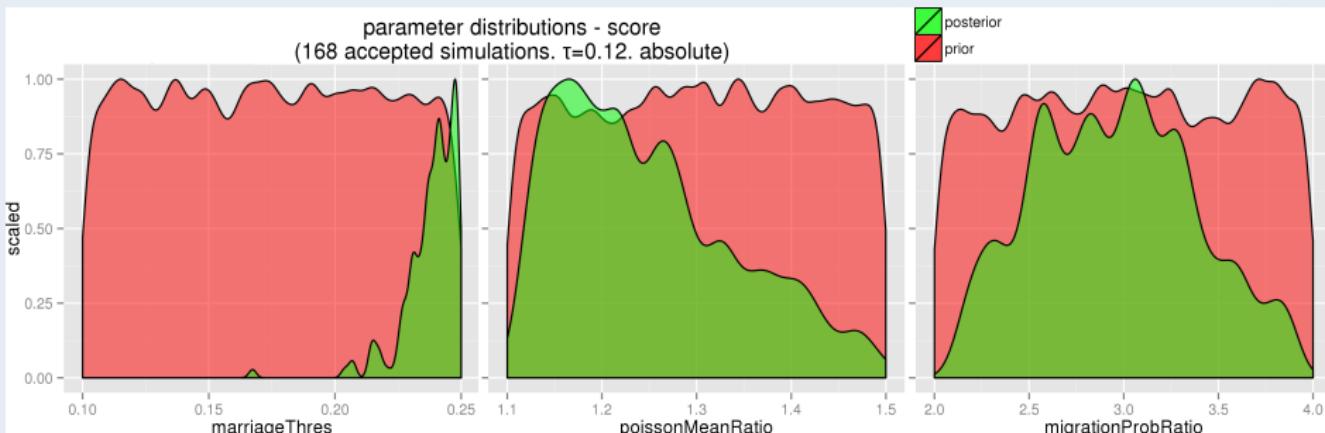


Figure : Example output from an ABC analysis

THANK YOU

Context

Geological and anthropological context

Previous papers

Model

Measures and comparisons

Observed data

Parameters

Comparison functions

Pipeline

Overview

Statistical analysis

Grid search

Approximate Bayesian Computation

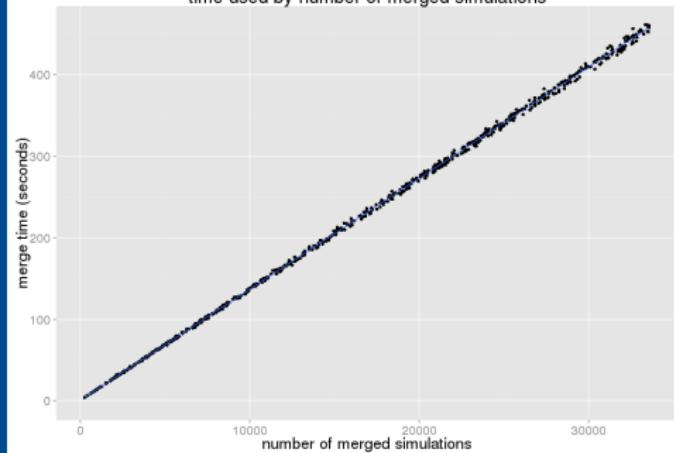
Questions?

map backgrounds from “HERE Satellite”

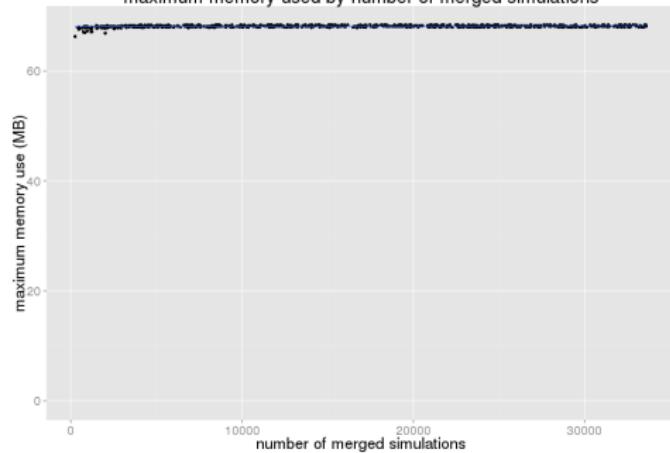
Computational Biology Research Group:
massey.genomicus.com

BENCHMARK MERGE

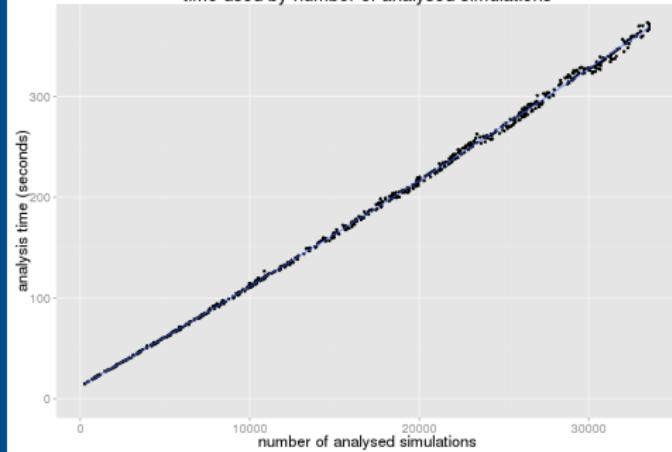
time used by number of merged simulations



maximum memory used by number of merged simulations



time used by number of analysed simulations



maximum memory used by number of analysed simulations

