EBOV phylodynamics using regression-ABC

CBGP seminar

Emma Saulnier

Samuel Alizon and Olivier Gascuel

21/11/2017

CNRS, UM, MIVEGEC, LIRMM

Introduction

Mathematical epidemiology

Phylogenies of viral infections

Phylodynamics

Approximate Bayesian Computation (ABC)

Major human viral outbreaks during the past decade

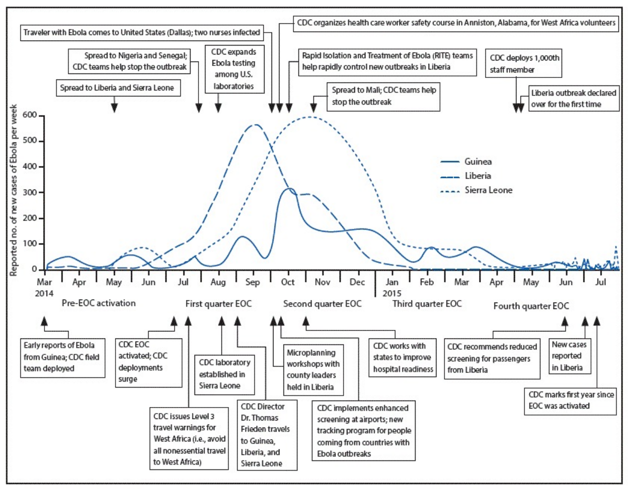
- 214 countries (worldwide)
- 18,449 deaths

- 3 countries (west Africa)
- 28,616 cases
- 11,310 deaths

- 84 countries (Americas, Africa, Asia)
- >2,000 cases of microcephaly in Brazil

according to the WHO (2017)

Public health interventions



2014-2016 Ebola outbreak in west Africa

CDC (2016)

Basic reproduction number

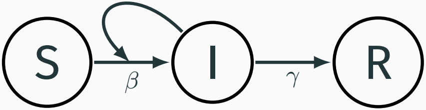
- $\mathcal{R}_0$: expected number of secondary infections caused by an infected individual during its entire infection, in a fully susceptible population of hosts
- Early estimations for the 2014-2016 Ebola outbreak in Sierra Leone : $\mathcal{R}_0 = 2.02\ [1.79 - 2.26]$

- $\mathcal{R}_0 > 1$: the epidemic spreads
- $\mathcal{R}_0 < 1$: the epidemic is under control

WHO Ebola Response Team (2014)

Mathematical epidemiology

Susceptible-Infected-Removed (SIR) epidemiological model:



Ordinary differential equations (ODEs):

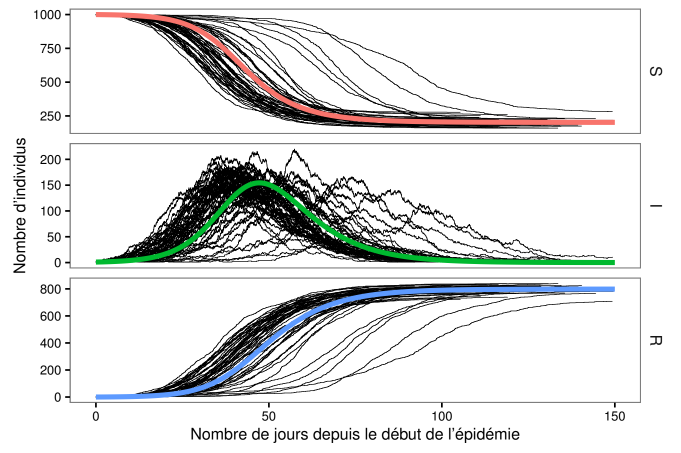
$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta I(t)S(t), \qquad \frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta I(t)S(t) - \gamma I(t), \qquad \frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma I(t)$$

Reproduction number:

$$\mathcal{R}(t) = \frac{\beta S(t)}{\gamma}$$

$$\mathcal{R}_0 = \mathcal{R}(t_0) = \frac{\beta N}{\gamma}, \; (S(t_0) = N)$$

SIR model trajectories

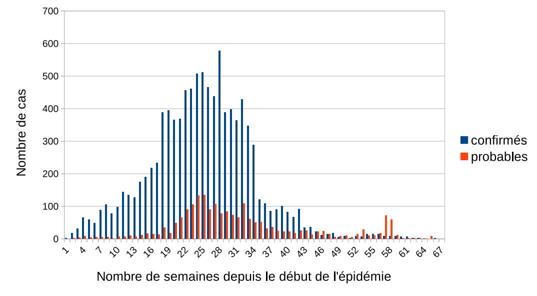


100 simulations with $\mathcal{R}_0 = 2$, $N = 1000$ et $d_I = 7$ days

(expected duration of infection $d_I = \frac{1}{\gamma}$).

Epidemiological or surveillance data

Incidence time series



2014-2016 Ebola outbreak in Sierra Leone.

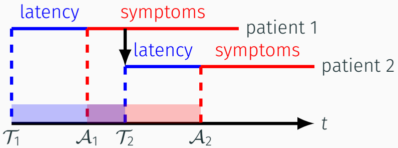
- Actual incidence = $\beta S(t)I(t)\mathrm{d}t$
- Observed incidence $\propto$ actual incidence $\times$ sampling proportion

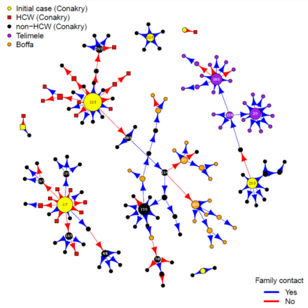
data from the WHO (2016)

Epidemiological or surveillance data

Questionnaires

- Generation time ($\mathcal{T}_2 - \mathcal{T}_1$)
- Serial interval ($\mathcal{A}_2 - \mathcal{A}_1$)
- Transmission networks



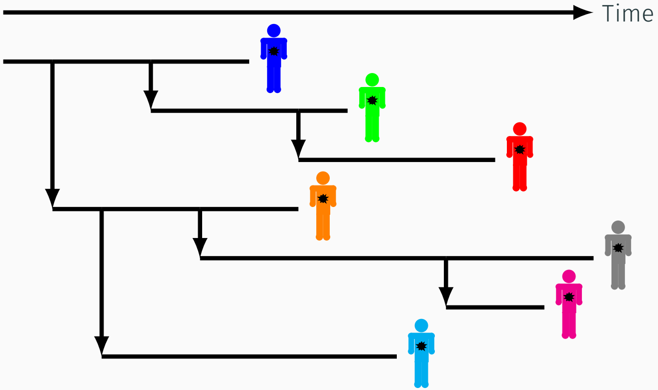


2014-2016 Ebola outbreak in Guinea

Faye *et al.* (2015)

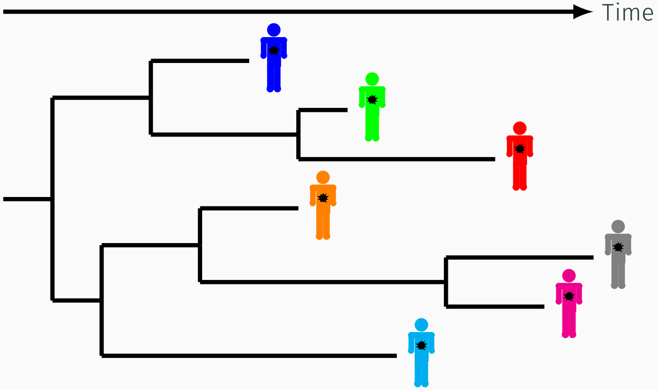
Phylogeny of infections and transmission tree

Full transmission tree



Phylogeny of infections and transmission tree

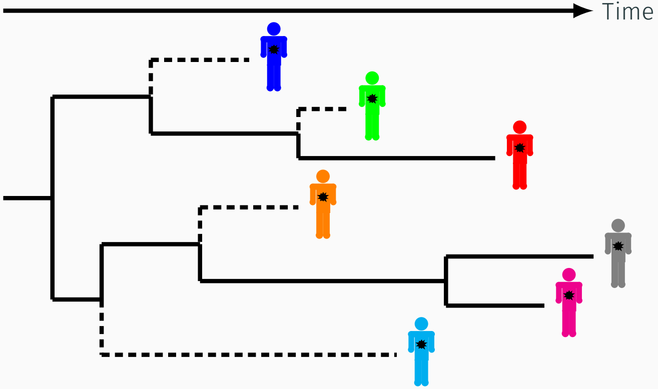
Full phylogeny of infections



- Neutral evolution assumption
- Loss of transmission directionality

Phylogeny of infections and transmission tree

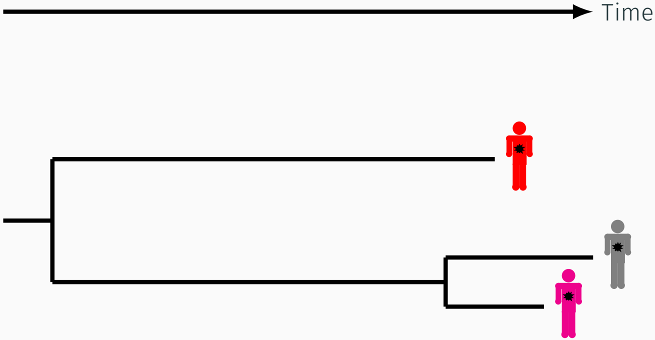
Phylogeny of sampled infections



- Loss of some transmission events (branchings)

Phylogeny of infections and transmission tree

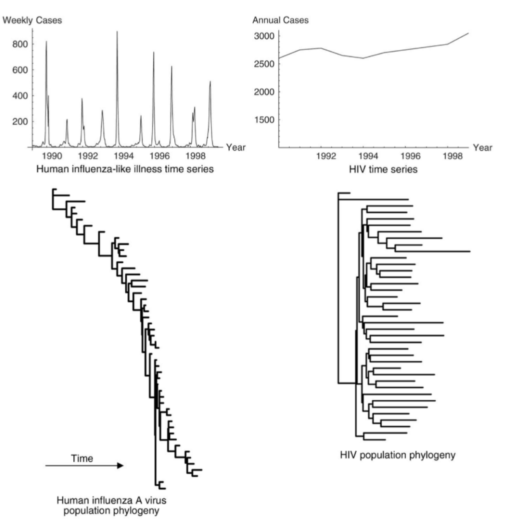
Phylogeny of sampled infections



- Still contains information about the epidemiological dynamics provided:
  - a sufficient number of sequences
  - a good sampling proportion

The rise of phylodynamics

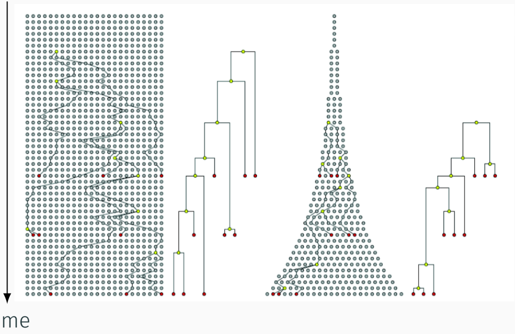
Human influenza A virus HIV



Grenfell *et al.* (2004)

Phylodynamic inference of the $\mathcal{R}_0$

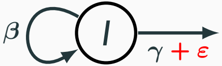
Coalescent models

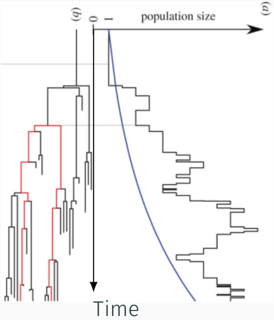


- Reconstruct the phylogeny going **backward in time**
- **Strong assumption** on the demographic history (eg: constant population size or exponential growth)
- Infer population size and growth rate
- Relationship between population growth rate and $\mathcal{R}_0$

review from Kühnert *et al.* (2011)

Birth-death (BD) models





- Reconstruct the phylogeny and the demographic history **simultaneously**, going **forward in time**
- Assume a birth-death process **with sampling**
- Relationship between birth and death rates and the $\mathcal{R}_0$ for simple epidemiological models

Stadler *et al.* (2015)

Limits of the current approaches aiming to infer the $\mathcal{R}_0$

from epidemiological data:

- Under-reported or memory-based data

from phylogenies of infections:

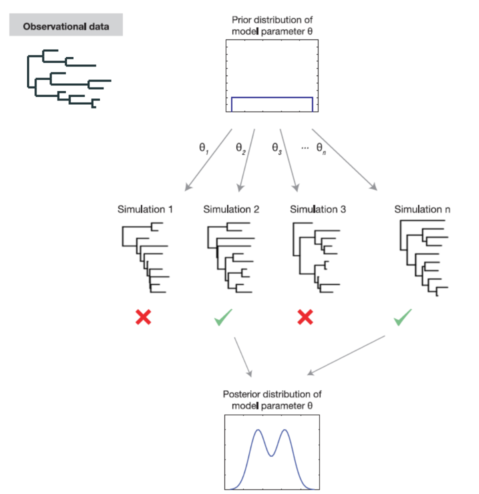
- Rely on simple demographic models that are different from the epidemiological models

more broadly:

- No integration of both types of data (genetic and epidemiological)
- Based on the computation of a likelihood function
  - limited by the model complexity
  - limited by the dataset size

Approximate Bayesian Computation (ABC)

- Approach based on **simulation** from **any kind of model**
- Comparison between observed and simulated data using a **distance** frequently involving **summary statistics**
- Potentially **not limited** by the model complexity nor by the dataset size
- Would allow to infer more than just the $\mathcal{R}_0$



Sünnaker *et al.* (2013)

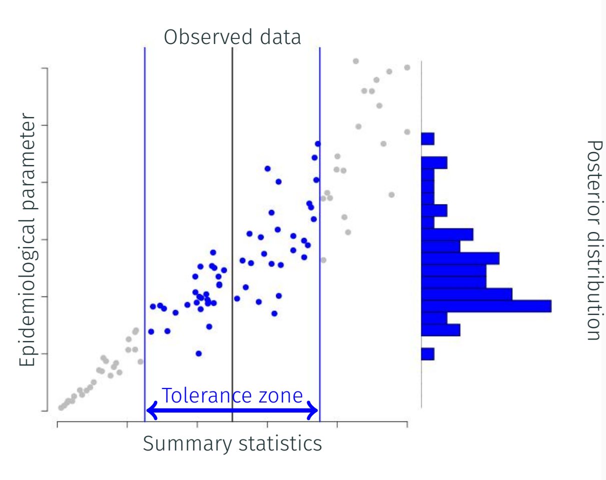
Regression-ABC

Simulations and summary statistics computation

Csilléry *et al.* (2010)
Beaumont *et al.* (2002)

Regression-ABC

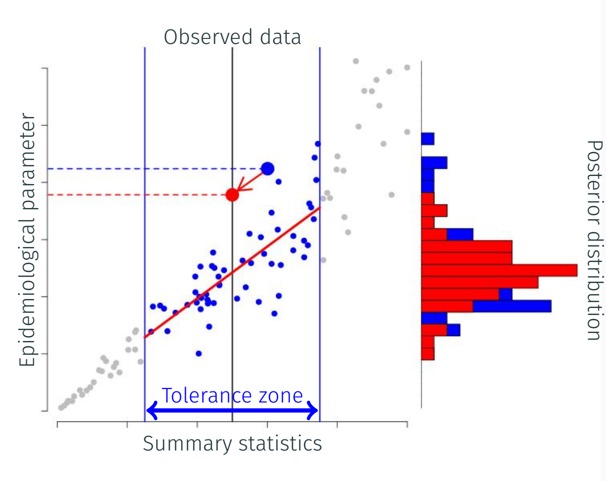
Rejection algorithm



Csilléry *et al.* (2010)
Beaumont *et al.* (2002)

Regression-ABC

Adjustment using regression



Csilléry *et al.* (2010)
Beaumont *et al.* (2002)

Goals

- Develop an ABC approach for phylodynamics
- Validate this approach by comparison with current approaches
- Apply this approach to:
    - a large dataset (phylogeny + epidemiological data)
    - a complex epidemiological model

Inferring epidemiological parameters from phylogenies using regression-ABC

Simulating phylogenies of infections from epidemiological models

Comparing simulated phylogenies to the observed phylogeny

Comparison study

Simulating phylogenies of infections from epidemiological models

The direct approach

- Simulation of the phylogeny of sampled infections and the epidemiological trajectory **simultaneously**, going **forward in time**
- Requires to model the sampling process (ε)
- Implemented in MASTER [Vaughan *et al.* (2013)]

The two-step approach

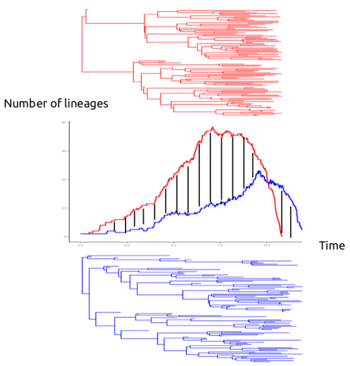
- Simulation of the phylogeny of infections **after** the epidemiological trajectory, going **backward in time**
- Uses the sampling dates
- Implemented in Rcolgem [Volz (2012)]

Comparing simulated phylogenies to the observed phylogeny

Functional distance

$$d_f(\Phi_{\mathrm{obs}}, \Phi_{\mathrm{sim}})$$

- Difficult to design
- ABC-MCMC [Marjoram *et al.* (2003)]
- Kernel distance [Poon *et al.* (2013)]
- Distance between two LTT plots [Saulnier *et al.* (2017)]

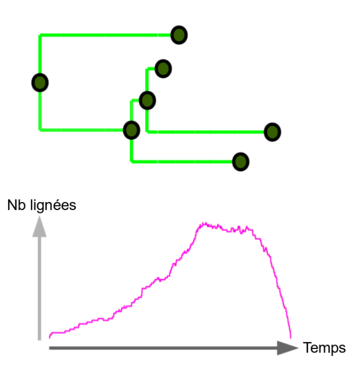


Comparing simulated phylogenies to the observed phylogeny

Summary statistics

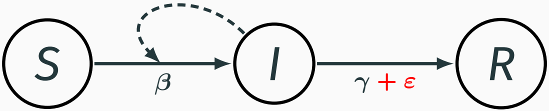
$$d(s(\Phi_{\text{obs}}), s(\Phi_{\text{sim}}))$$

- Easy to design
- Regression-ABC [Blum *et al.* (2010)]
- 83 statistics :
  - Branch lengths (26)
  - Topology (8)
  - LTT plot (9)
  - X-axis coordinates of the LTT plot (20)
  - Y-axis coordinates of the LTT plot (20)



Comparison study

SIR model with sampling:



Parameters:

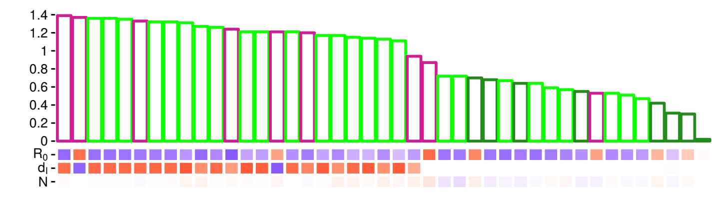
- $\mathcal{R}_0 = \beta N/(\gamma + \varepsilon)$
- $d_I = 1/(\gamma + \varepsilon)$
- $N = S + I + R$

Comparison study

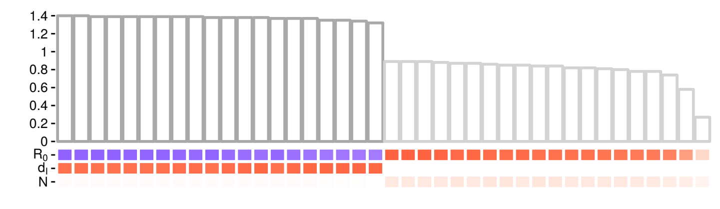
Methods

- ABC-D [Saulnier *et al.* (2017)]: Rejection algorithm with distance between two LTT plots
- ABC: Rejection algorithm with the 83 summary statistics
- ABC-FFNN [Blum *et al.* (2010)]: Rejection algorithm with the 83 summary statistics + adjustment using FFNN regression (non-linear + variable selection)
- ABC-LASSO [Saulnier *et al.* (2017)]: Rejection algorithm with the 83 summary statistics + adjustment using LASSO regression (linear + optimized variable selection)
- BDSIR [Kühnert *et al.* (2014)]: Approach based on the likelihood of the BDSIR model and using MCMC (BEAST)
- Kernel-ABC [Poon (2015)]: ABC-MCMC method using the kernel distance (simulations using Rcolgem)

Epidemiological information captured by the summary statistics



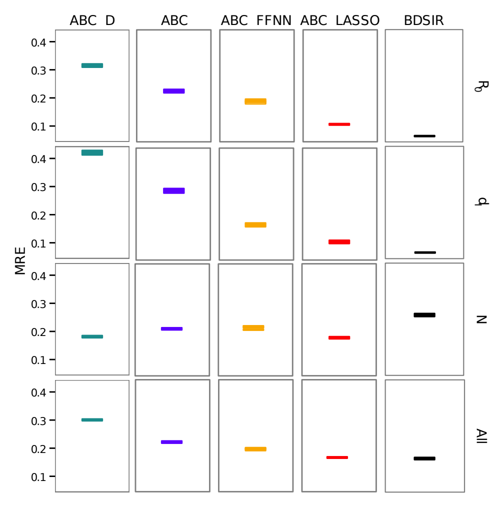
- LTT plot
- Branch lengths
- Topology



- X-axis coordinates of the LTT plot
- Y-axis coordinates of the LTT plot

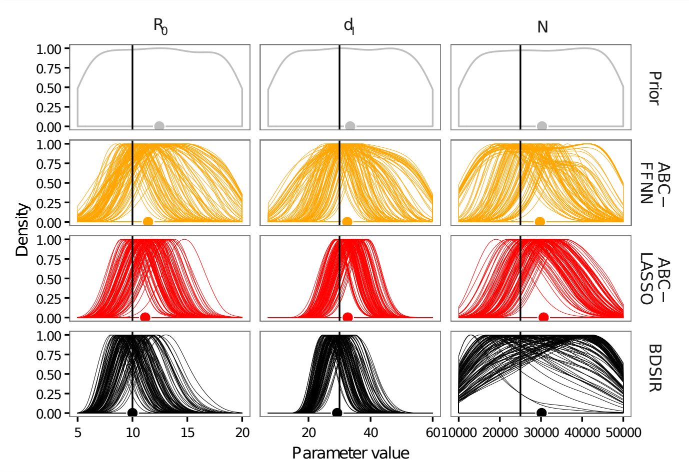
Saulnier *et al.* (2017)

Similar accuracies for ABC-LASSO and BDSIR methods on large phylogenies



Saulnier *et al.* (2017)

The BDSIR method hardly converges towards a posterior distribution for N



Saulnier *et al.* (2017)

Conclusions on this first section

- 83 summary statistics contains information about the epidemiological parameters
  - LTT plot > branch lengths » topology
- Similar accuracies for ABC-LASSO and BDSIR methods on large phylogenies
- Adjustment using regression reduces the inference error
- ABC-LASSO is more robust than ABC-FFNN
- Bad convergence for N using BDSIR

Inferring epidemiological parameters using regression-ABC and combining phylogeny and incidence data

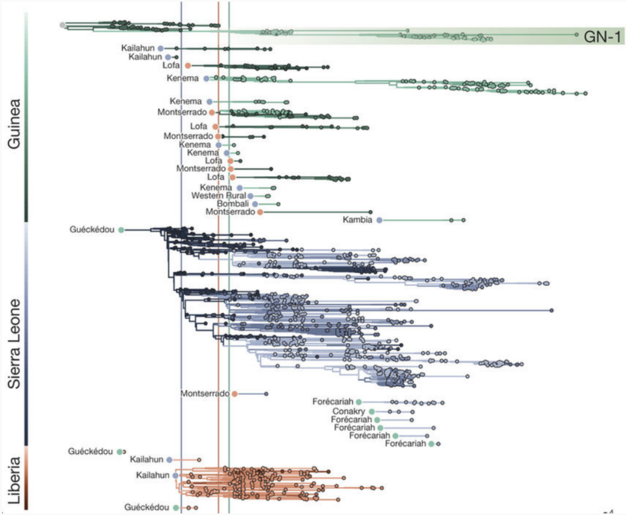
2014-2016 Ebola outbreak in Sierra Leone

SEIDR model

ABC-regression inferences using the phylogeny and/or the incidence data

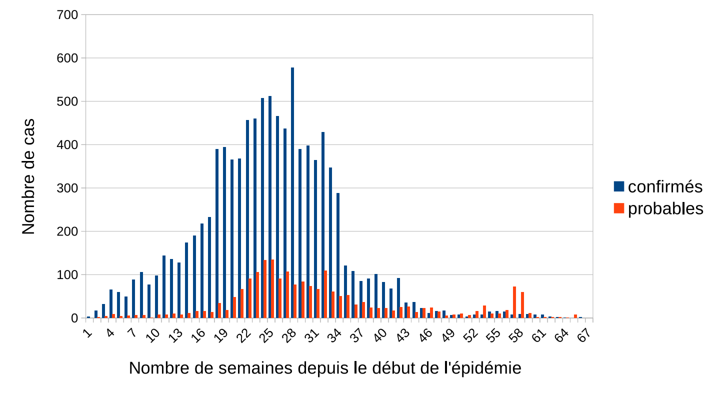
Sensitivity of our phylodynamic approach to phylogenetic uncertainty

Phylogeny of the 2014-2016 Ebola outbreak in Sierra Leone



Dudas *et al.* (2017)

Incidence data



2014-2016 Ebola outbreak in Sierra Leone

data from the WHO (2016)

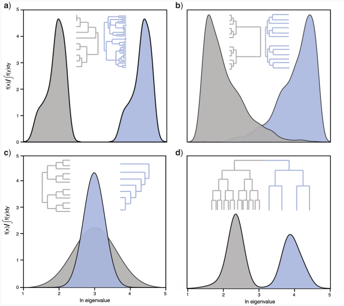
New summary statistics

computed on incidence data:

- Date of the maximal incidence value
- Slope of the exponential growth phase
- Slope of the exponential decrease phase
- Slope ratio
- Auto-correlation coefficients

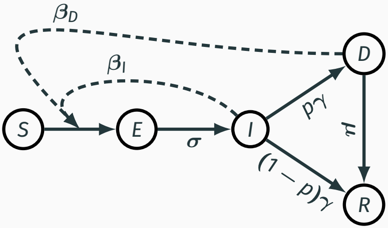
computed on phylogenies:

- Statistics of the Laplacian spectrum



Lewitus *et al.* (2015)

SEIDR model



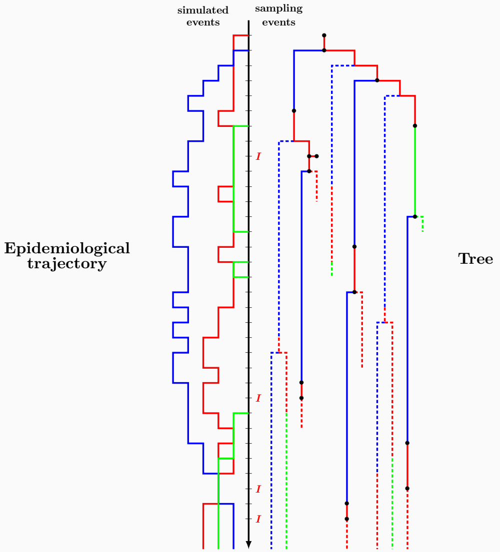
Fixed parameters (according to [WHO Ebola Response Team (2015)]):

- Expected duration of the latency phase: $1/\sigma = 11.8$ days
- Expected duration of the symptomatic phase: $1/\gamma = 6.2$ days
- Lethality rate: $p = 0.765$

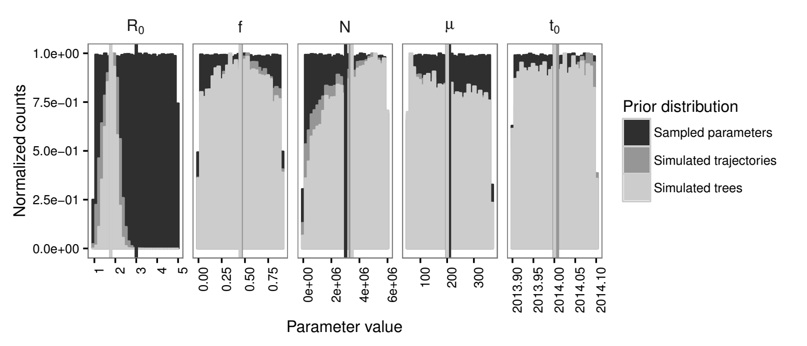
Variable parameters:

- Global basic reproduction number: $\mathcal{R}_0 = \mathcal{R}_{0,I} + \mathcal{R}_{0,D}$
- Fraction of the $\mathcal{R}_0$ associated to dead bodies: $f = \mathcal{R}_{0,D}/\mathcal{R}_0$
- Expected duration of *post-mortem* transmissibility: $1/\mu$
- Total population size: N
- Date of origin of the epidemic: t_0

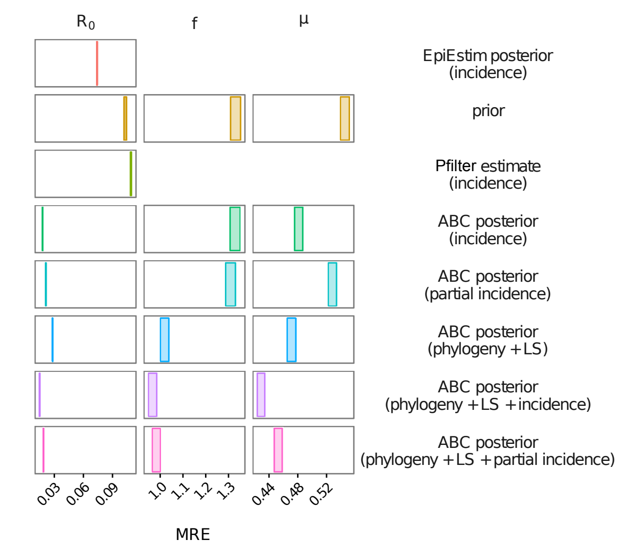
Simulation of large phylogenies of infections from the SEIDR model



Modifications of the prior distributions after simulations



More accurate estimations using ABC-regression with both types of data

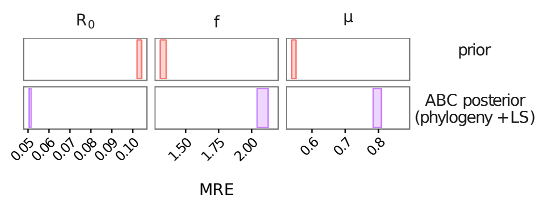


Important sensitivity to phylogenetic uncertainty due to the low substitution rate

Procedure:

1. Sequence simulation for simulated phylogenies of infections
2. Phylogenetic inference using the simulated sequences
3. Time-scaling of the phylogenetic trees
4. Inference using regression-ABC

Ebola virus substitution rate: 0.0012 subst.site^{-1}.year^{-1}

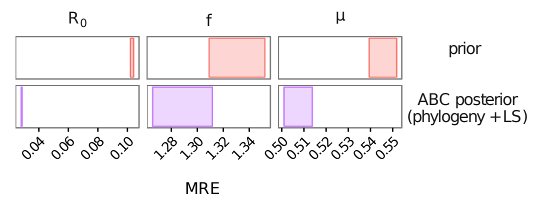


Important sensitivity to phylogenetic uncertainty due to the low substitution rate

Procedure:

1. Sequence simulation for simulated phylogenies of infections
2. Phylogenetic inference using the simulated sequences
3. Time-scaling of the phylogenetic trees
4. Inference using regression-ABC

Ebola virus substitution rate $\times 10$



$\mathcal{R}_0$ inferences using the Sierra Leone dataset

Incidence:

$$\mathcal{R}_0 = 1.44\ [1.39 - 1.58]$$

Phylogeny + incidence:

$$\mathcal{R}_0 = 1.65\ [1.58 - 1.81]$$

Phylogeny:

$$\mathcal{R}_0 = 1.68\ [1.60 - 1.80]$$

Conclusions on this second section

- The new simulation approach induces a modification of the prior distributions
- The parameter inference is improved by the use of both types of data
- Regression-ABC inferences are impacted by the phylogenetic uncertainty
- This is especially the case for f and μ
- A higher substitution rate improves the phylogenetic inference and therefore the parameter inference using regression-ABC
- We re-estimated the $\mathcal{R}_0$ of the 2014-2016 Ebola outbreak in Sierra Leone using the phylogeny and incidence data

Conclusions and perspectives

Conclusions

- We developped a regression-ABC approach for phylodynamics
- We validated it by comparing it to several existing approaches
- We applied it to the dataset of the 2014-2016 Ebola outbreak in Sierra Leone

Limits of our regression-ABC phylodynamic approaches

- Ability to rapidly simulate a large dataset
- Identifiability of the parameters from the data and through the summary statistics
- Rejection algorithm based on the Euclidian distance computed on large vectors of unweighted statistics
- Non-linear regression method with optimized variable selection
- Sensitivity to phylogenetic uncertainty
- No model comparison

Perspectives

short term

- Test other regression models
    - random forests, deep learning
- Applications to other datasets (flu virus, HIV)
    - more data
    - more complex models (seasonality, spacial spread, host and contact heterogeneity)
    - new statistics on labellized trees

Perspectives

long term

- Develop a model comparison approach
- Use sequences instead of phylogenies
  - simulate sequence evolution during an epidemic
  - develop new summary statistics on sequences
  - results in removing the problem of phylogenetic uncertainty
  - enable to test other assumptions about the sequence evolution

Thanks

Fundings

PEPS (CNRS, UM), Sidaction

MIVEGEC

- Samuel Alizon
- Members of the ETE team

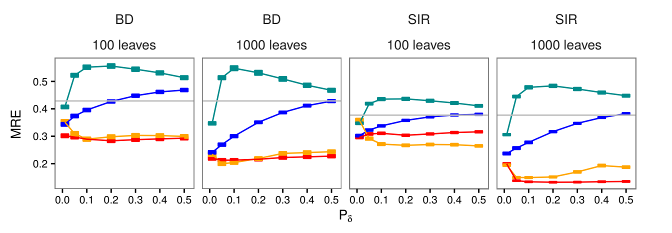
LIRMM

- Olivier Gascuel
- Members of the MAB team

Thank you for your attention

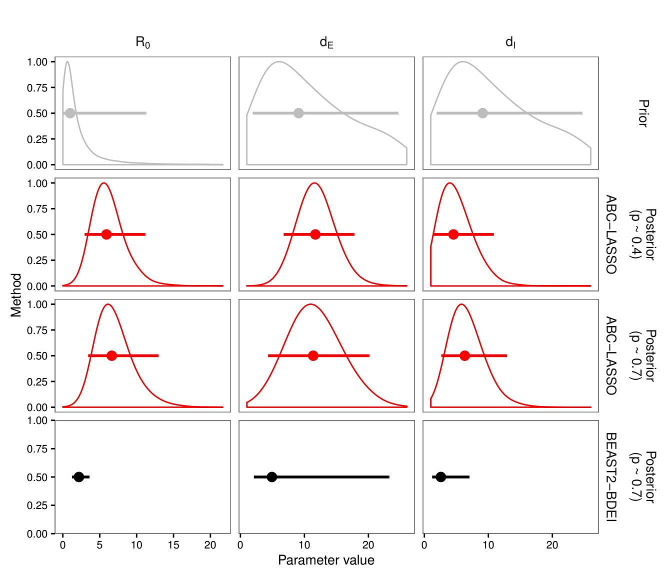
Questions ?

Erreur d'inférence de plusieurs méthodes ABC en fonction de la tolérance

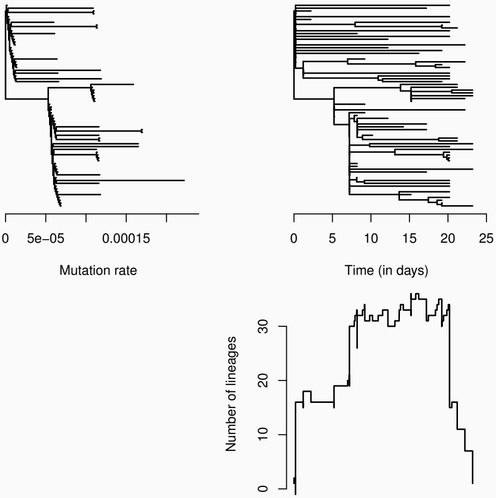


- ABC-D
- ABC
- ABC-FFNN
- ABC-LASSO

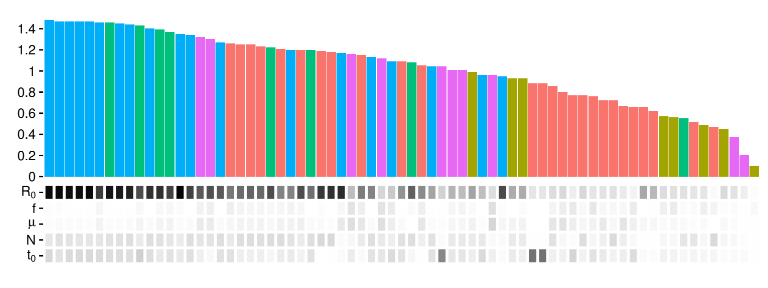
Inférences à partir de la phylogénie du début de l’épidémie d’ Ebola en Sierra Léone en 2014



Phylogénie du début de l'épidémie d'Ebola en Sierra Léone en 2014



Nouvelles statistiques de résumé



Inc > LTT > LS > BL > Topo

Algorithme itératif de filtres à particules