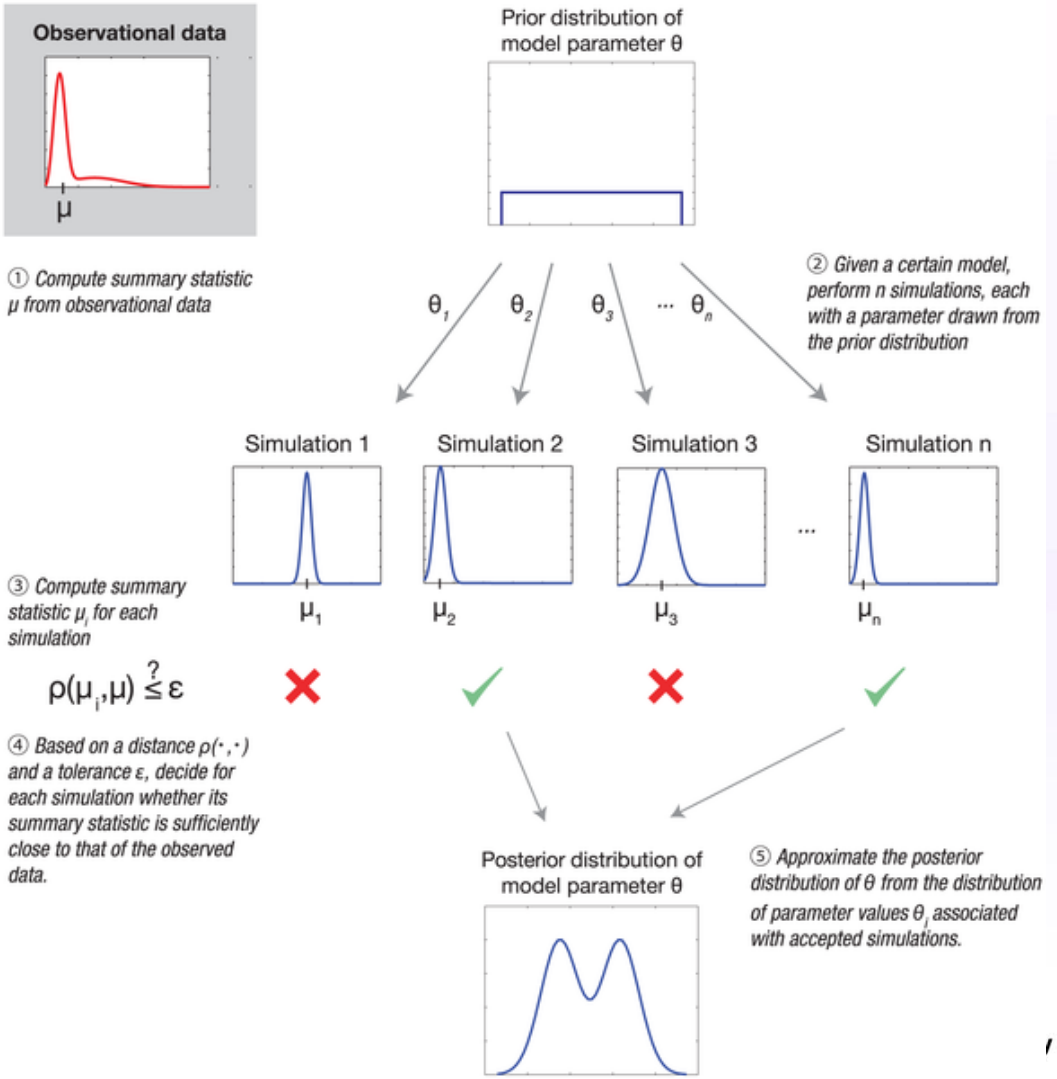
Introduction to ABC with an application to estimating transmission dynamics

Jukka Corander
Department of Mathematics and statistics
University of Helsinki, Finland

What ABC?

- Approximate Bayesian computation (ABC) is a method to do inference for intractable models
- Intractability means here that likelihood calculation is either too expensive or impossible
- Assumes we can still simulate data from our model
- The core idea of ABC was introduced in a seminal paper by Tavaré et al. (1997) to do inference for a coalescence model

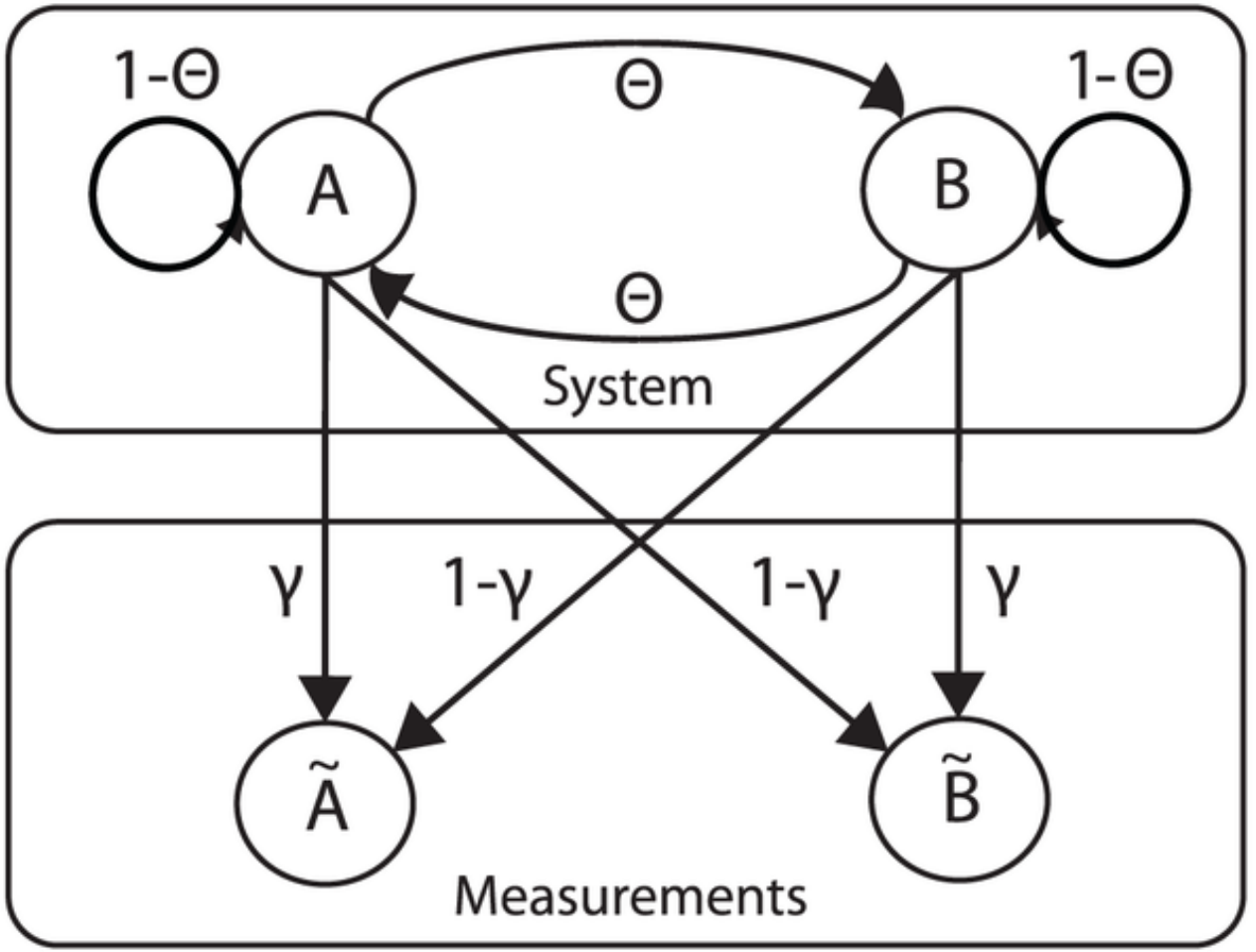
ABC - how does it work?



Sunnåker et al. (PLoS Computation Biology, 2012)

ABC – simple HMM example

Data: AAAABAABBAAAAAABAAAA, summary statistic #switches = 6



Sunnåker et al. (PLoS Computation Biology, 2012)

ABC – simple HMM example

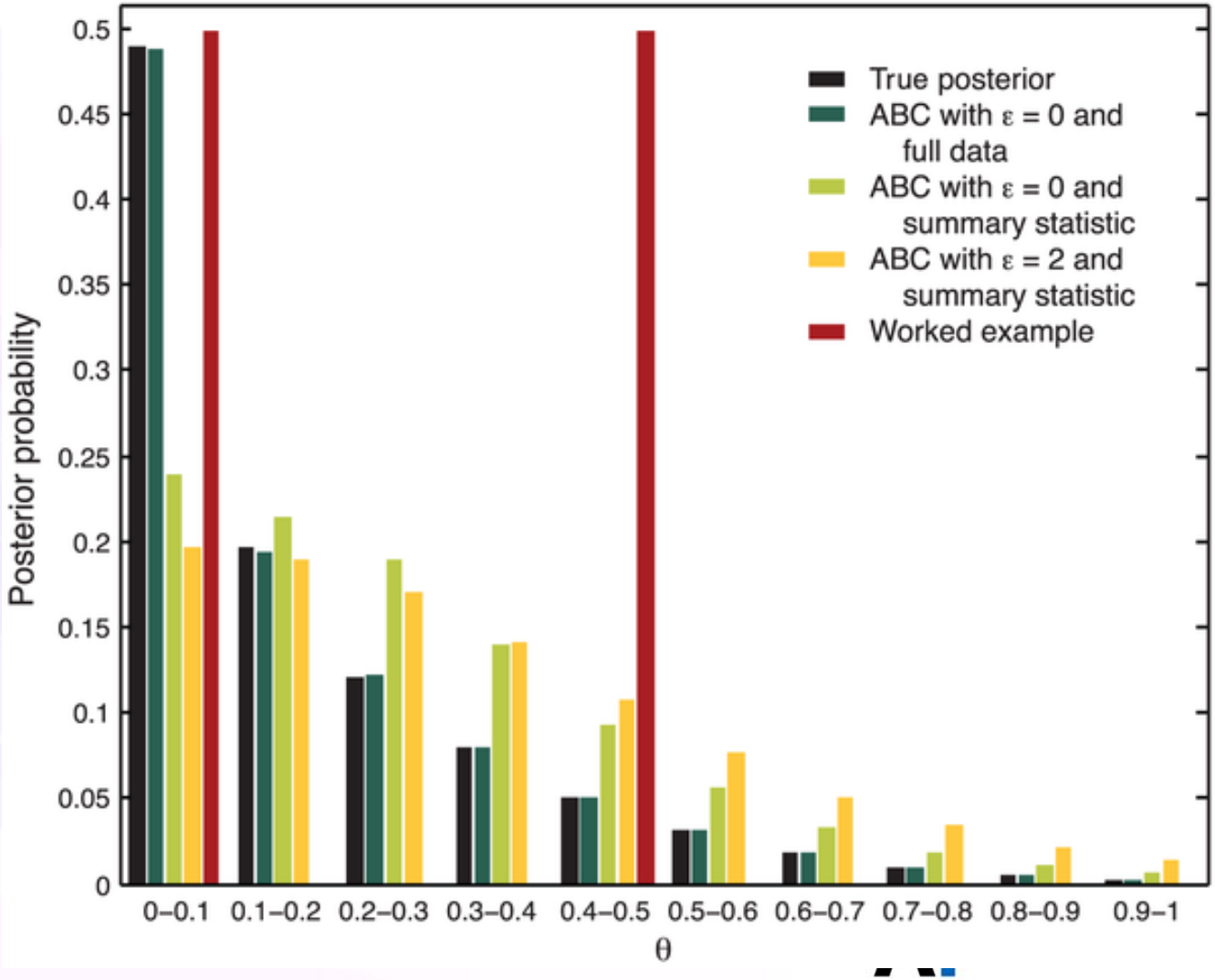
Assume prior $P(\theta) \sim U(0,1)$ and simulate data given random draws θ_i, $i=1,\ldots,n$

I	θ_i	Simulated Datasets (Step 2)	Summary Statistic ω_{S_i} (Step 3)	Distance $\rho(\omega_{S_i}, \omega_{S_0})$ (Step 4)	Outcome (Step 4)
1	0.08	AABAAAABAABAAABAAAAA	8	2	accepted
2	0.68	AABBABABAAABBABABBAB	13	7	rejected
3	0.87	BBBABBABBBBABABBBBBA	9	3	rejected
4	0.43	AABAAAAABBABBBBBBBBA	6	0	accepted
5	0.53	ABBBBBAABBABBABAABBB	9	3	rejected

doi:10.1371/journal.pcbi.1002803.t001

Sunnåker et al. (PLoS Computation Biology, 2012)

A look at inferences

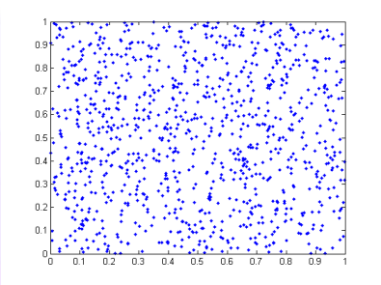


Sunnåker et al. (PLoS Computation Biology, 2012)

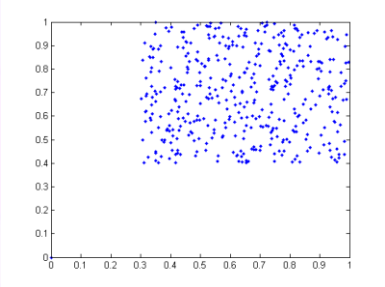
ABC-MCMC

1. Sample candidate θ^* from proposal $q(.,\theta)$ where θ is the current value of the parameter
2. Sample new data set using θ^* and calculate new summary statistic S^* (was S for θ)
3. If $\rho(S^*,S)<\varepsilon$, go to #4, else discard θ^* and go to #1
4. Accept θ^* with probability $[\pi(\theta^*)/\pi(\theta)]\cdot[q(\theta,\theta^*)/q(\theta^*,\theta)]$
5. Return to #1

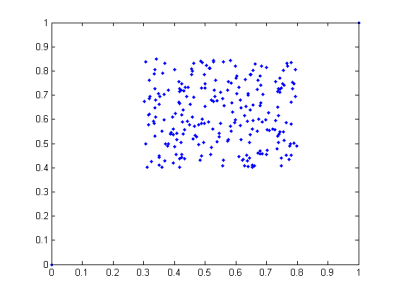
ABC-SMC



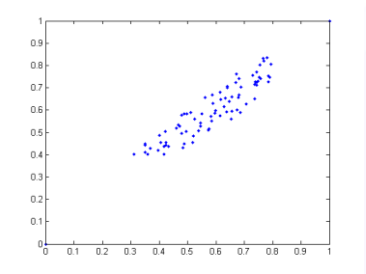
$\pi(\theta)$



$P(\theta|\rho(S(\theta),S)<\varepsilon_1)$



$P(\theta|\rho(S(\theta),S)<\varepsilon_2)$



$P(\theta|\rho(S(\theta),S)<\varepsilon_3)$

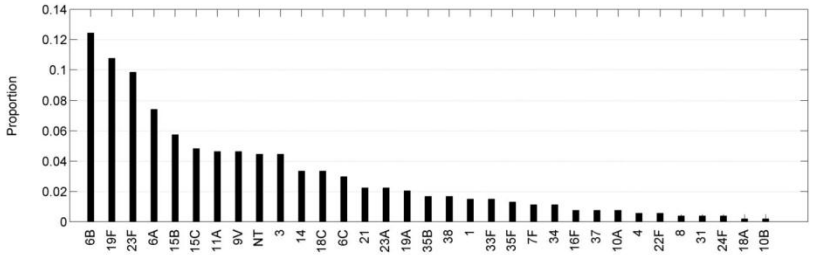
Aalto University

UNIVERSITY OF HELSINKI

ABC recap

- In reality more complex sampling algorithms: ABC-MCMC, particle filtering, etc
- Necessitates quality controls, predictive checks,…
- Formal ABC-based model comparison is an issue (Robert et al. PNAS, 2011), but latest results give more promising insight (Marin et al. JRSS B 2014, http://arxiv.org/abs/1110.4700)
- Very intensive research area at the moment!

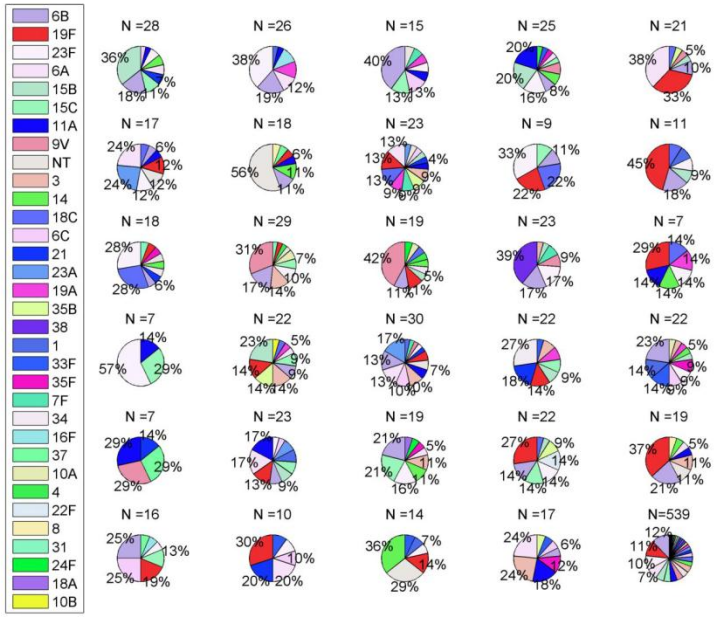
Pneumococcus strain incidences in Oslo DCCs (data sampled once in 2006)



Globally

Diversity of strains in the data:

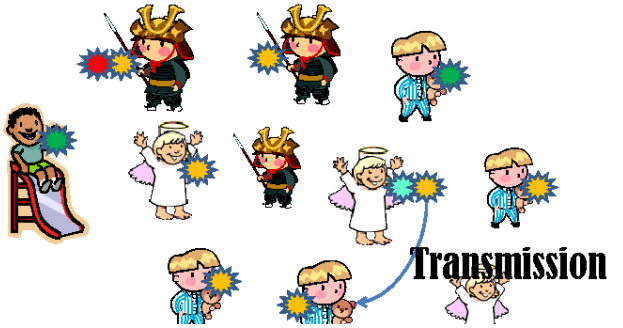
Locally:

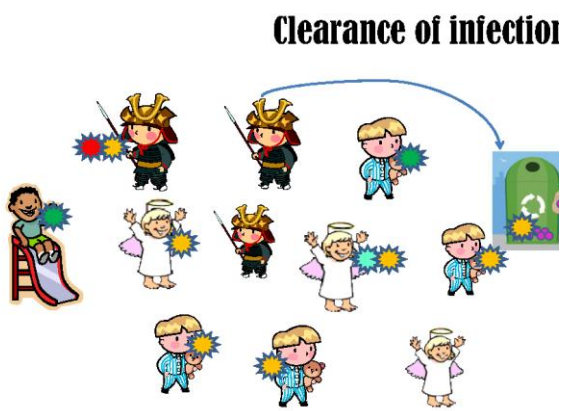


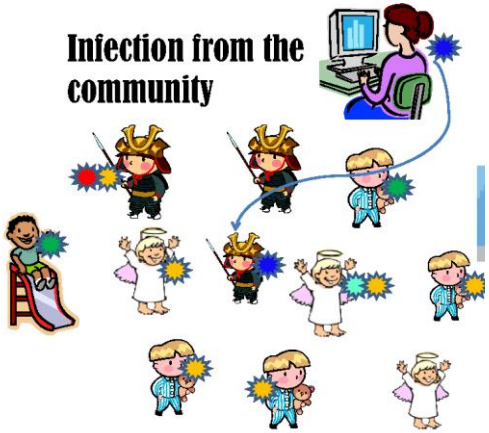
Numminen et al. (Biometrics 2013)

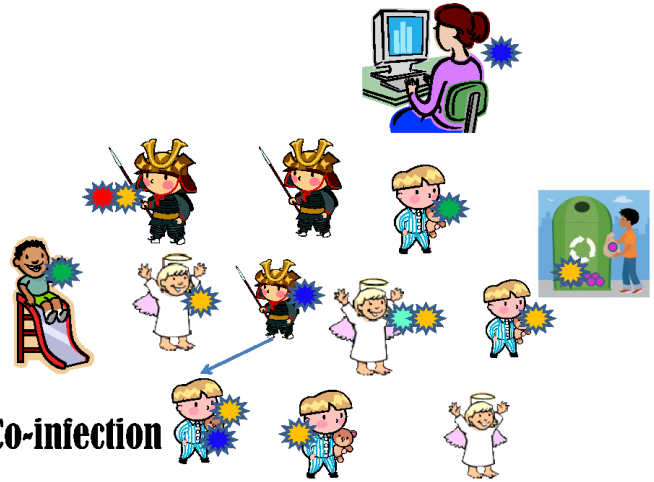
Stochastic microepidemics in a metapopulation

What is happening in a DCC?









Aalto University

UNIVERSITY OF HELSINKI

Numminen et al. (Biometrics 2013)

Stochastic SIS-model for DCC transmissions

- Ingredients for a stochastic soup within a DCC:
- $I_{ij}(t)$ indicator for kid *i* carrying strain *j* at time *t*
- β rate parameter for transmission from someone in DCC
- Λ rate parameter for transmission from outside DCC
- θ competition parameter scaling the probability of co-infection
- γ clearance rate parameter, since we have data from single time point only, all other parameters are estimated relative to a fixed clearance rate

Aalto University

UNIVERSITY OF HELSINKI

Stochastic SIS-model for DCC transmissions

- **Continuous-time Markov chain with transition probabilities:**

$$
\begin{aligned}
Pr\left(I_{is}(t+\delta t)=1 | I_{is}(t)=0\right) &= \beta E_s(I(t)) + \Lambda P^s + o(\delta t), \\
\text{if} \quad \sum_{j=1}^{N^s} I_{ij}(t) &= 0 \\
Pr\left(I_{is}(t+\delta t)=1 | I_{is}(t)=0\right) &= \theta\left(\beta E_s(I(t)) + \Lambda P^s\right) + o(\delta t)\right), \\
\text{if} \quad \sum_{j=1}^{N^s} I_{ij}(t) > 0 & \quad \text{and} \quad I_{is}=0. \\
Pr\left(I_{is}(t+\delta t)=0 | I_{is}(t)=1\right) &= \gamma + o(\delta t) \qquad (2)
\end{aligned}
$$

Aalto University

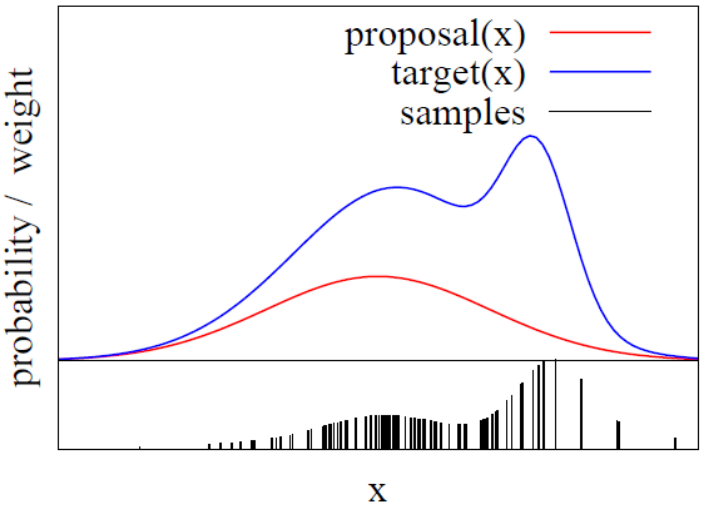
UNIVERSITY OF HELSINKI

How to do the ABC inference?

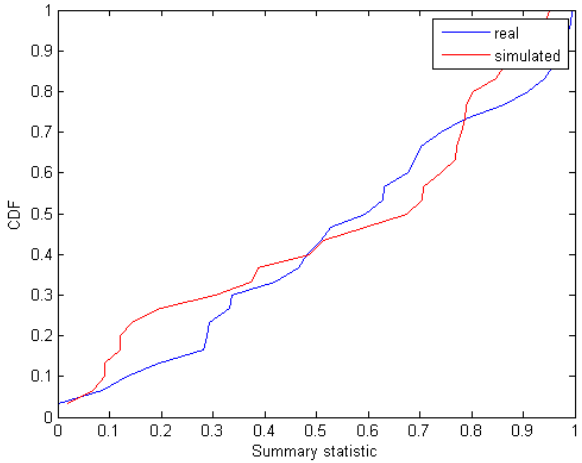
Summaries & discrepancies used in sequential importance sampling

(1) Shannon index of diversity (Peet, 1974) of the distribution of observed strains.
(2) Number of different strains.
(3) Prevalence of carriage among the observed individuals.
(4) Prevalence of multiple infections among the observed individuals.

$$d_k = \int |F^k(x) - \hat{F}^k(x)| dx.$$



IS fig by Cyrill Stachniss



Aalto University

UNIVERSITY OF HELSINKI

Some results

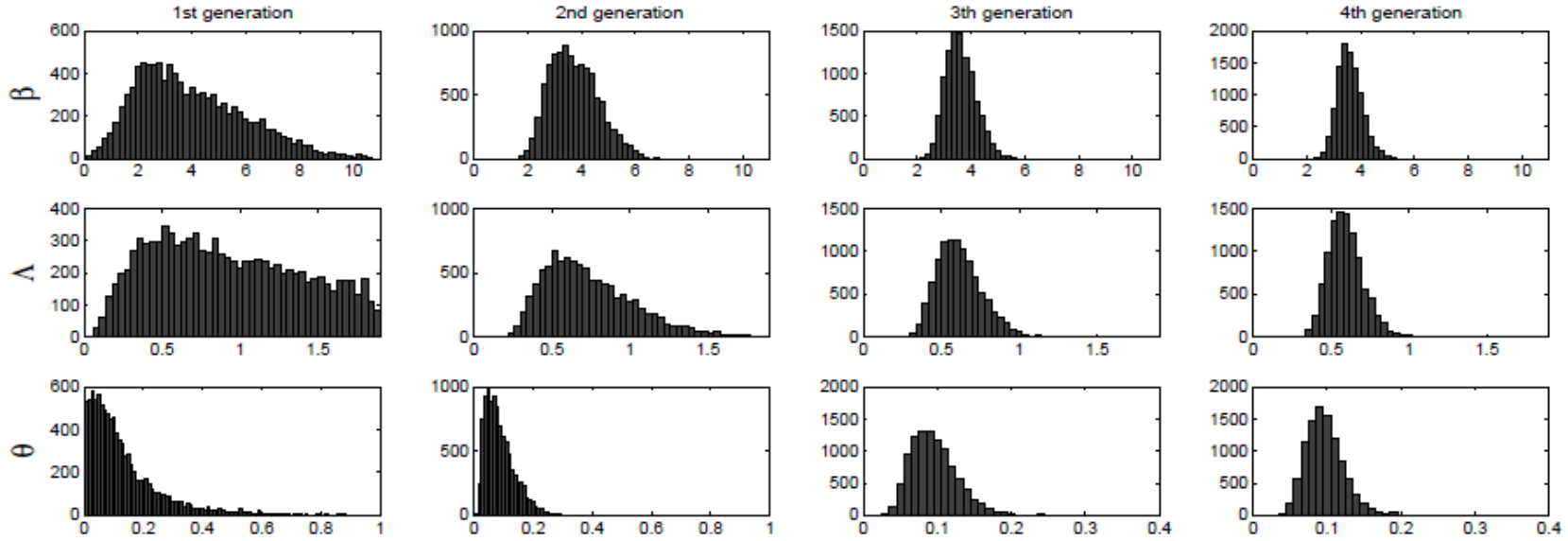


Table 1

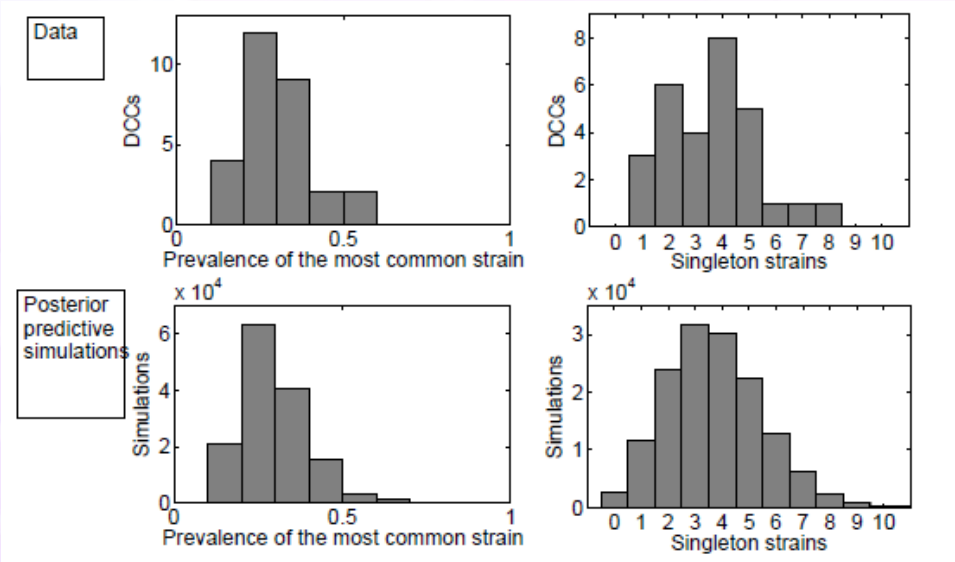
Summaries of the posterior distribution of the estimated parameters, with two different simulation times for the transmission model

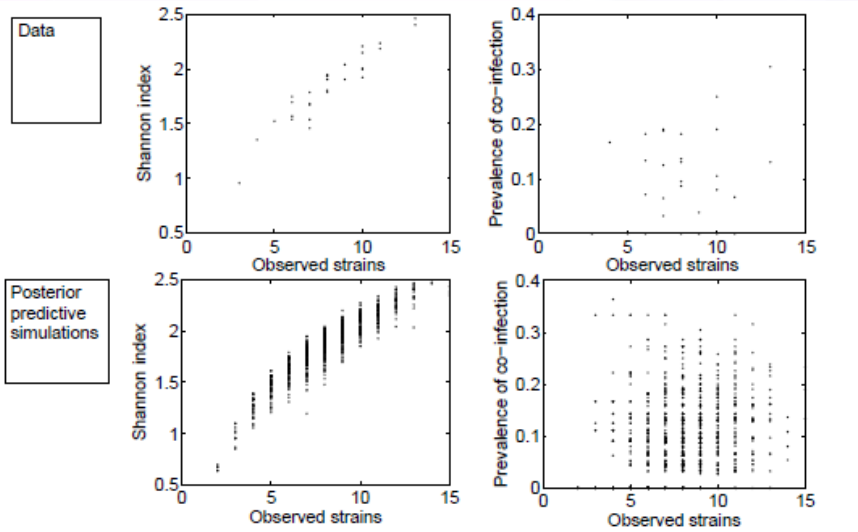
	Mean $T = 10$	Mean $T = 20$	95% CI $T = 10$	95% CI $T = 20$
β	3.589	3.594	(2.8157, 4.5785)	(2.8113, 4.5621)
Λ	0.593	0.584	(0.4017, 0.8359)	(0.3875, 0.8407)
θ	0.097	0.097	(0.0605, 0.1422)	(0.0604, 0.1427)

Aalto University

UNIVERSITY OF HELSINKI

Model validation with ‘unused’ summaries

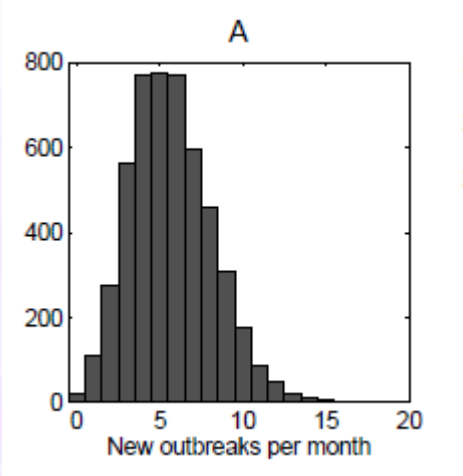


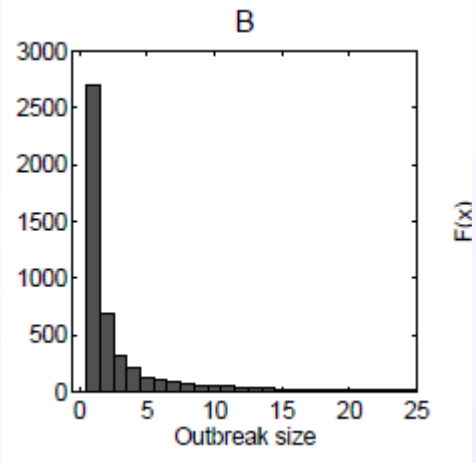


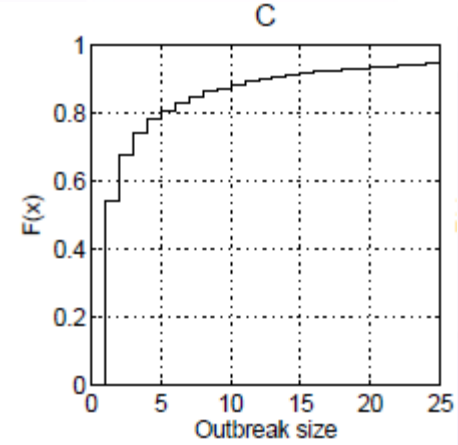
Aalto University

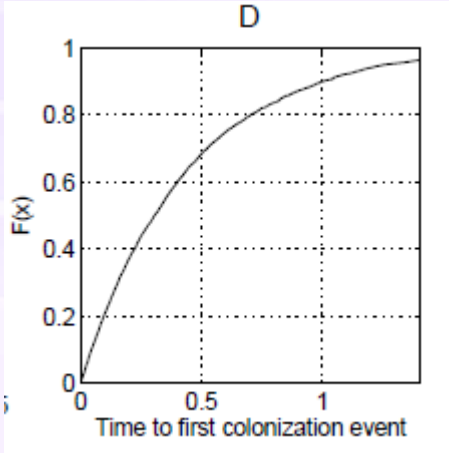
UNIVERSITY OF HELSINKI

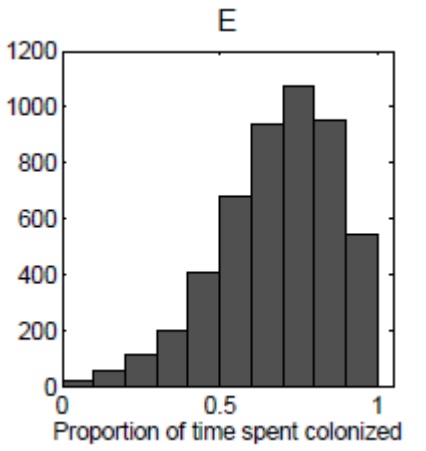
Posterior predictive simulations











Aalto University

UNIVERSITY OF HELSINKI

What else with ABC?

- ABC is particularly attractive for dynamic models with tricky/intractable/expensive likelihood functions
- ABC has grown particularly popular for complex spatio-temporal models in population genetics
- We are currently developing several generic machine learning inspired approaches to solve the key problems in ABC inference: choice of summary statistics, choice of metric to compare synthetic and real summaries, convergence to high likelihood/posterior regions

COIN

With great power comes great responsibility! -Uncle Ben

Hence the ABC sword should never be wielded casually

Aalto University

UNIVERSITY OF HELSINKI