PCV Modeling Update 02/2017

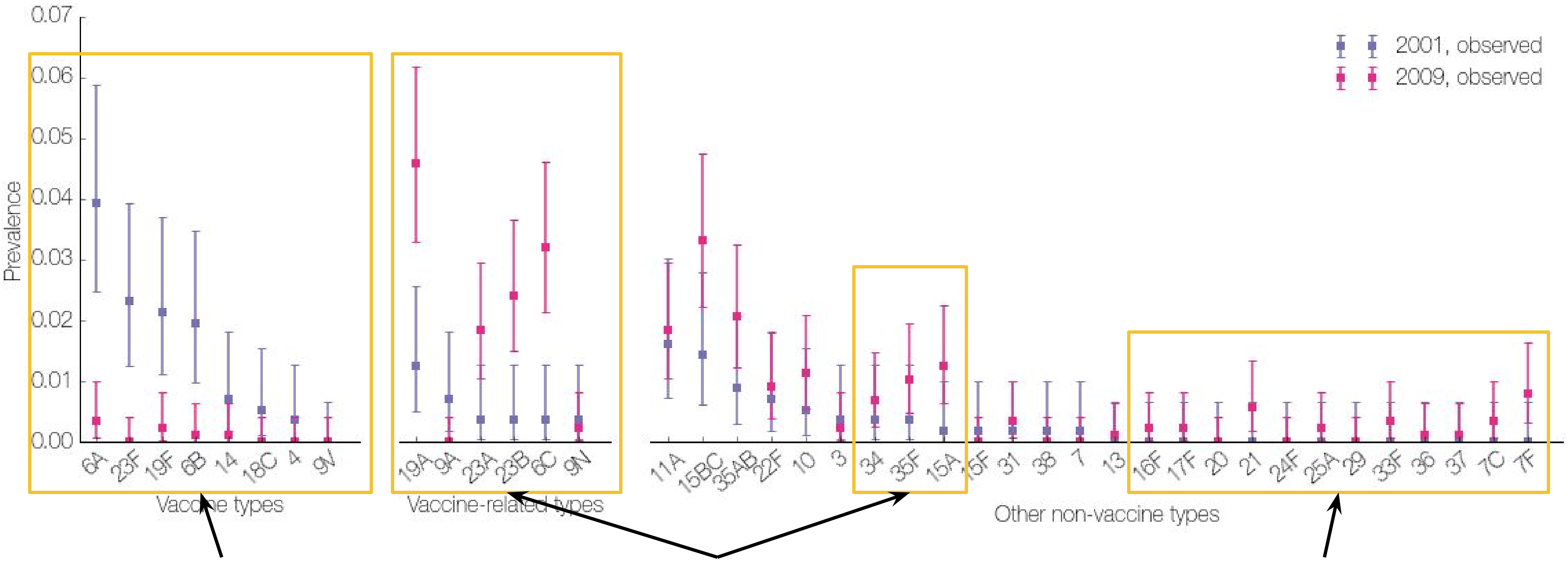
Marc Lipsitch, PI and Francisco Cai, Programmer

Recap

- **Fitness parameters of the model are fit using peri-PCV7 (2001) data.**
- **Model reproduces peri-PCV7 serotype-specific prevalences.**
    - Expected, since number of free parameters = number of observed quantities.
- **Model had trouble reproducing PCV7-era serotype-specific prevalences.**
    - Not surprising, since there are 40 more quantities, but only 1 new parameter, vaccine efficacy.
    - 4 of 6 vaccine-related types (VRTs) were consistently underestimated (19A, 23A, 23B, 6C).
    - In general, model could not accommodate changes to the fitness ordering of serotypes.
    - Serogroup cross-immunity and shortened colonizations led to mild improvements.
    - Details in Phase II Tasks 3 and 4 Report.

Next step

- Fit model using both peri-PCV and PCV-era data.



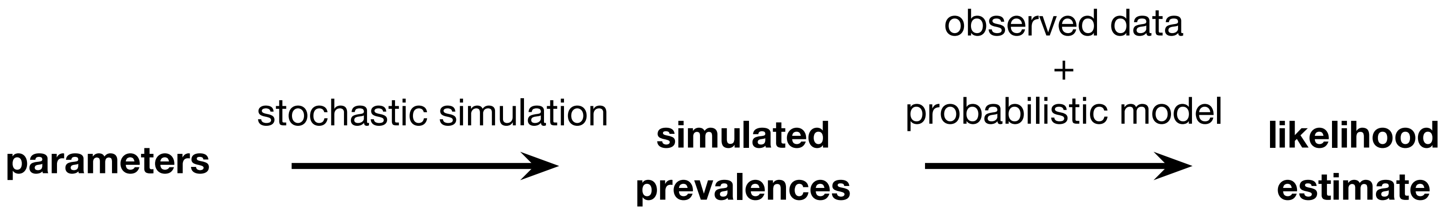
Peri-PCV data is more informative of vaccine types

PCV7-era data is more informative for VRTs and non-vaccine types (NVTs) that expanded after vaccine introduction...

...particularly for NVTs that were not sampled in 2001.

Model fitting: Before

Goal: Maximize expected likelihood of parameters given 2001 data.



parameters

stochastic simulation

simulated prevalences

observed data + probabilistic model

likelihood estimate

Challenges: The likelihood…

… is a function of many (40) parameters

Parameter space is big.

… has no closed form

→ cannot calculate gradient

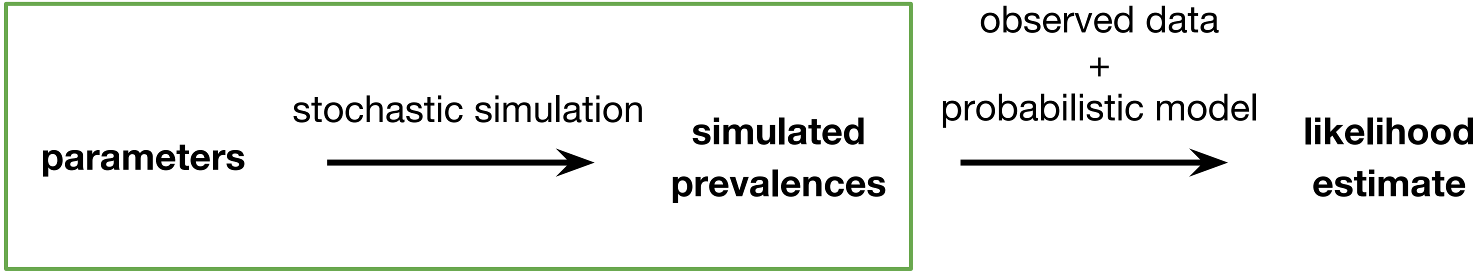
… cannot even be calculated directly

→ we only have a noisy estimate of it by running a stochastic simulation, which takes time (1-2 min.)

Difficult to determine which direction to explore next in parameter space.

Model fitting: Shortcuts exploited

Goal: Maximize expected likelihood of parameters given 2001 data.



parameters

stochastic simulation

simulated prevalences

observed data + probabilistic model

likelihood estimate

Things we had going for us:

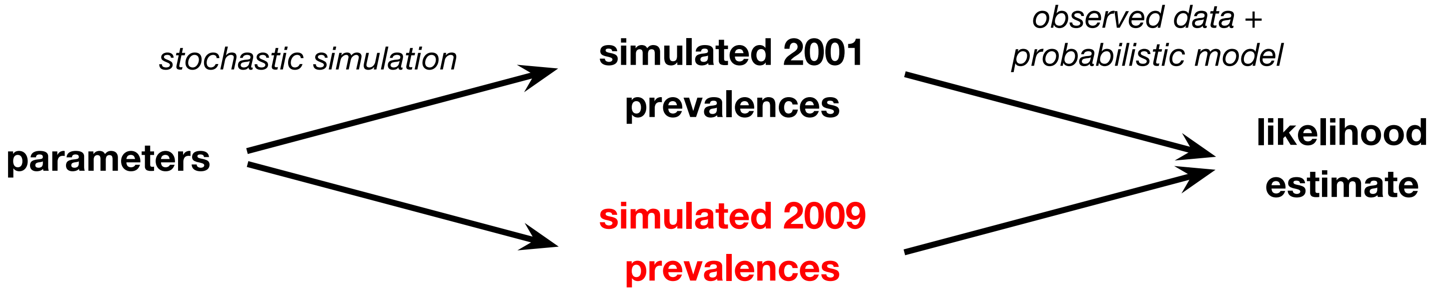
1. If the simulated prevalences matched the observed prevalences, this will maximize the likelihood.
   → Focus on finding parameters that reproduce the observed prevalences (green box).

2. Monotonic relationship between fitness rank and prevalence.
   → Based on our current simulated prevalence, we know how to adjust the fitness rank.

3. Adjusting a serotype’s fitness rank does not affect prevalences of other serotypes excessively.
   → Fitness parameters can be fit in parallel.

Model Fitting: Now

Goal: Maximize expected likelihood of parameters given 2001 **and 2009** data.



parameters

stochastic simulation

simulated 2001 prevalences

simulated 2009 prevalences

observed data + probabilistic model

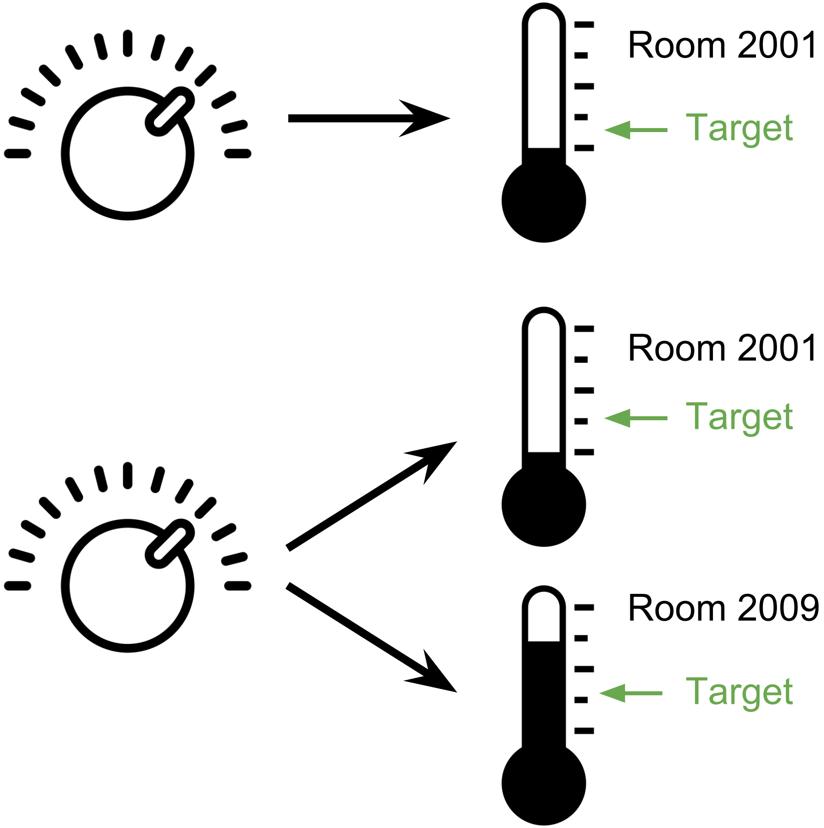
likelihood estimate

Things we used to have going for us (now with complications):

1. If the simulated prevalences matched the observed prevalences, this will maximize the likelihood.
   → Focus on finding parameters that reproduce the observed prevalences.
   The simulated prevalences now have to match the observed data at two time points.
2. Monotonic relationship between fitness rank and prevalence.
   → Based on our current simulated prevalence, we know how to adjust the fitness rank.
   We now adjust one rank and hope it reproduces the observed prevalence at two time points.
3. Adjusting a serotype's fitness rank does not affect prevalences of other serotypes excessively.
   → Fitness parameters can be fit in parallel.
   Parameters to be fit now also include vaccine efficacy.

Analogy

Before:



Raise thermostat a little bit.

Now:

Less clear what to do.

Icons: Ashley Worobec, Sarah Tan. Noun Project. thenounproject.com

Initial plan

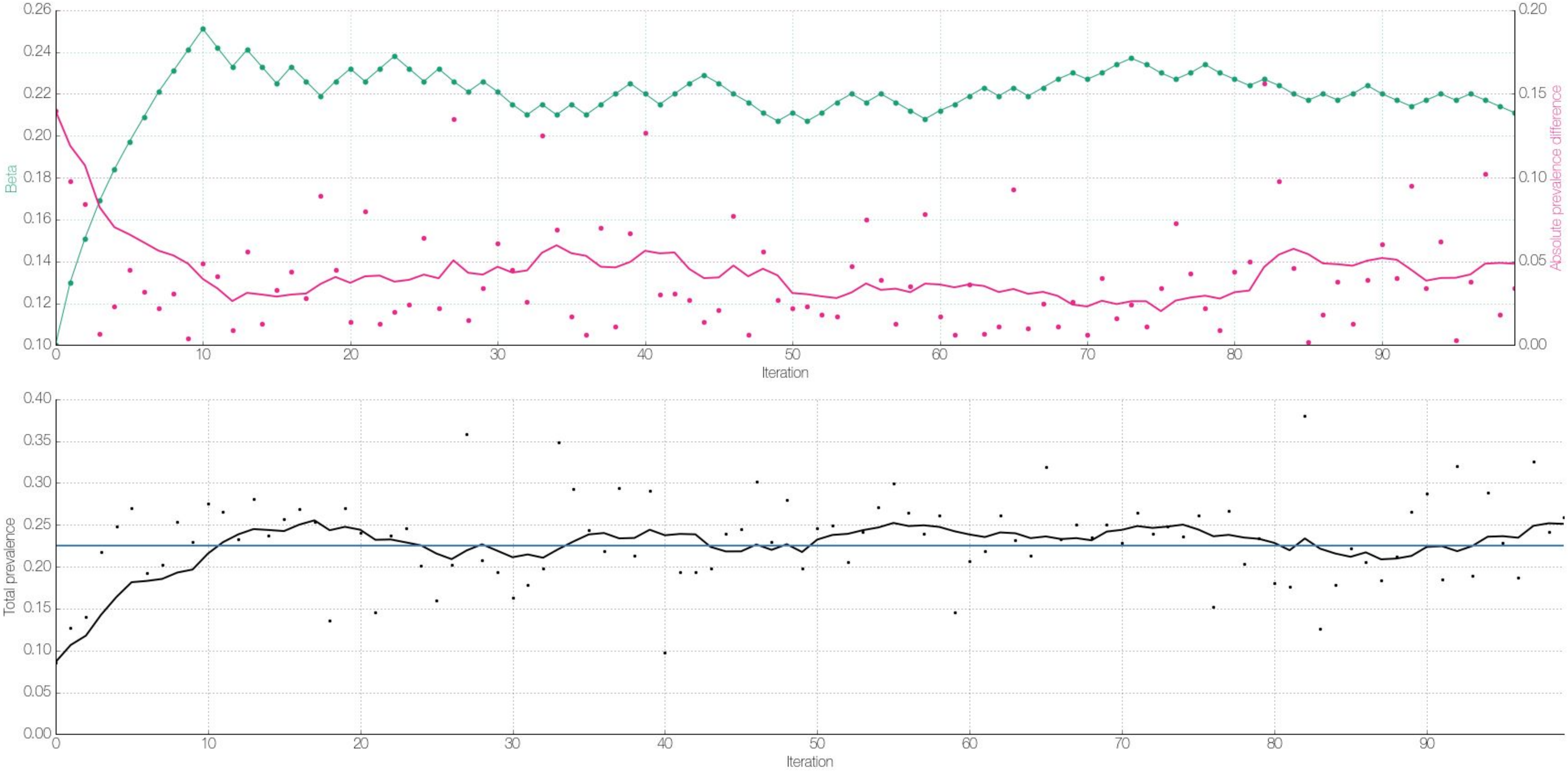
Challenges and how to address them:

1. Difficult to determine which direction to explore next in parameter space (no analytical gradient)
    - **Simultaneous perturbation stochastic approximation**
        - Simultaneously perturb all parameters to estimate all components of the gradient.
            - (*n* parameters requires only 2 simulations / iteration)
        - However, simulations may be too noisy… If so, will try increasing population size or averaging results from multiple simulations.
2. Parameter space is large.
    - **Try fitting fitness ranks in parallel**, i.e. adjust each fitness rank as if it only affected the terms in the log-likelihood involving its serotype.
    - Adjust the fitness ranks and the vaccine efficacy on alternate iterations.

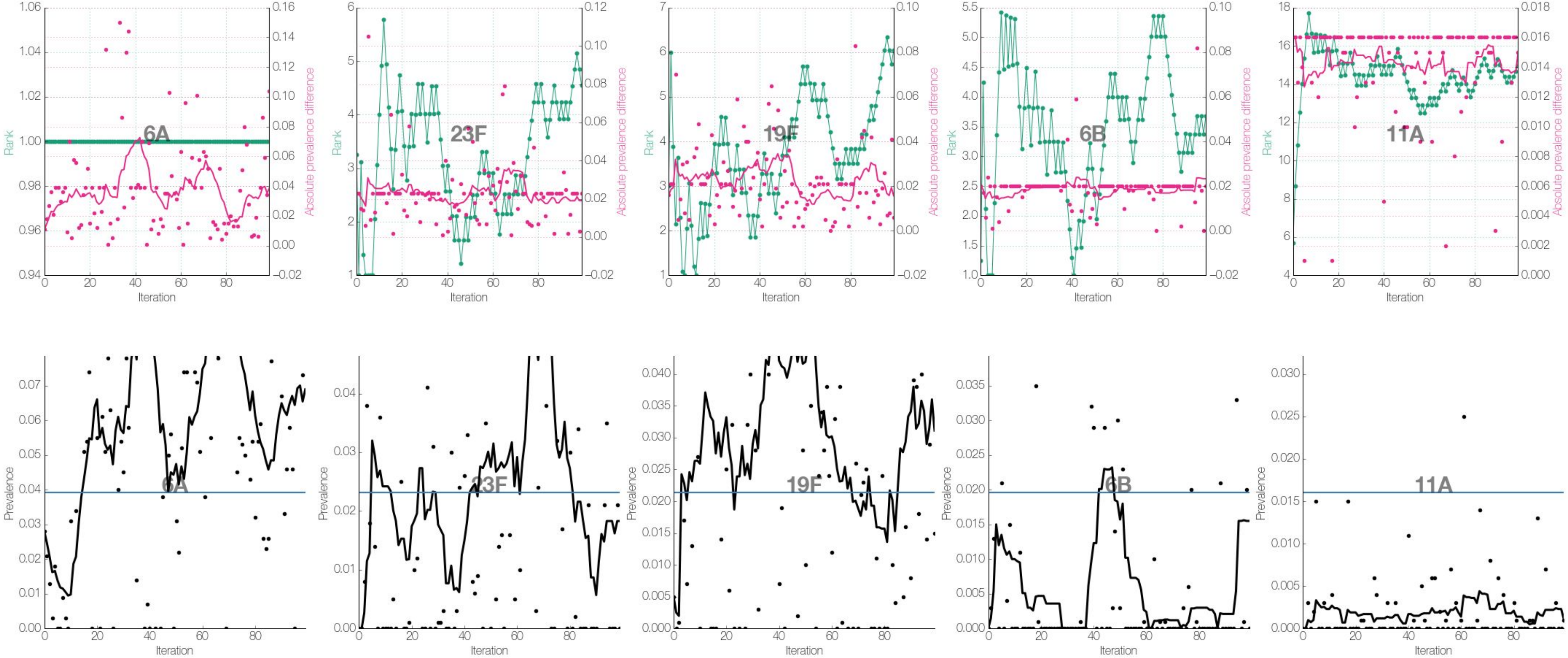
First milestone

- Try to reproduce previous results, i.e. fit only peri-PCV7 data using new algorithm
- As before, use absolute prevalence error, rather than log-likelihood, to quantify how well we are doing.

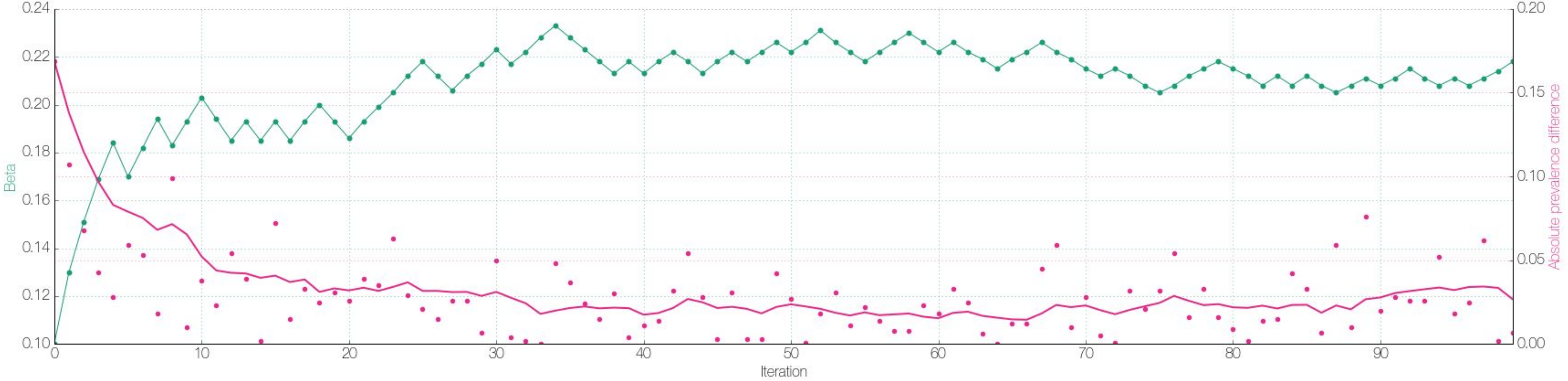
First test

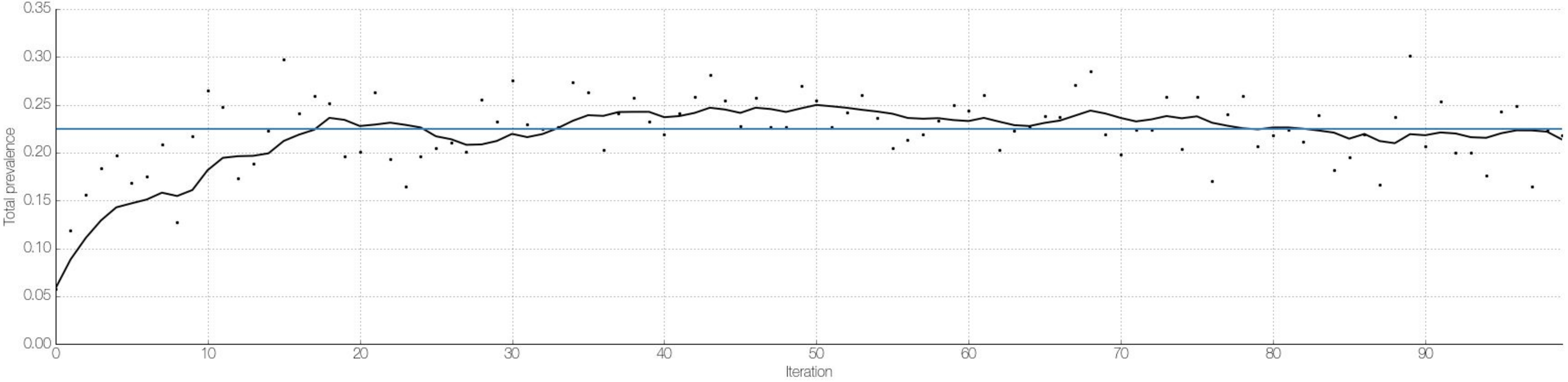

0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
Beta
0.20
0.15
0.10
0.05
0.00
Absolute prevalence difference
0
10
20
30
40
50
60
70
80
90
Iteration
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Total prevalence
Iteration

First test

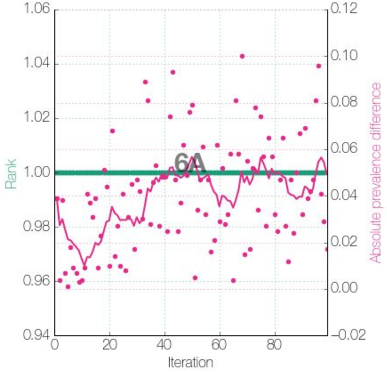

6A
23F
19F
6B
11A
Rank
Absolute prevalence difference
Iteration
0
20
40
60
80
1.06
1.04
1.02
1.00
0.98
0.96
0.94
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
6
5
4
3
2
1
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
7
6
5
4
3
2
1
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
18
16
14
12
10
8
6
4
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
6A
23F
19F
6B
11A
Prevalence
Iteration
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.04
0.03
0.02
0.01
0.00
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.030
0.025
0.020
0.015
0.010
0.005
0.000

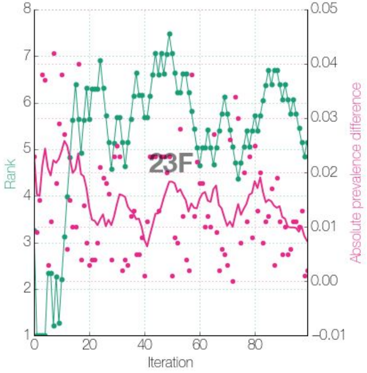
2.5x population size

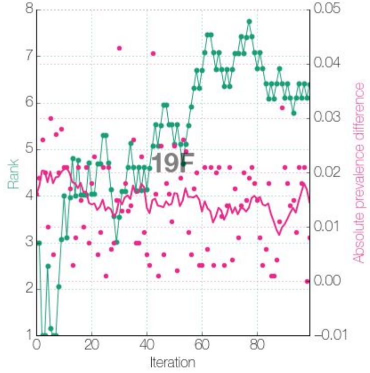


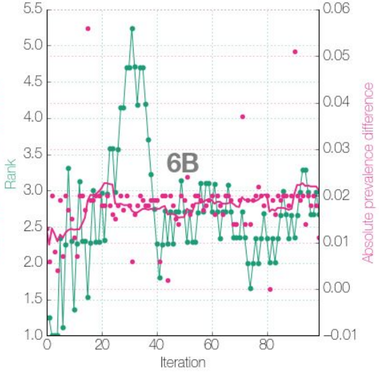


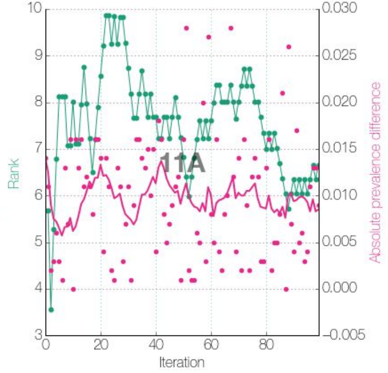
2.5x population size

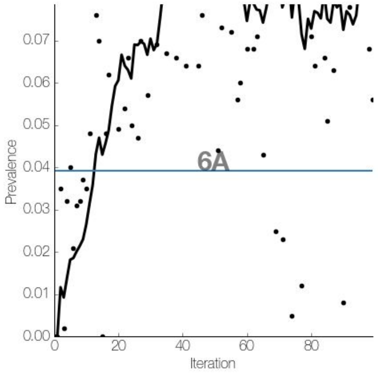


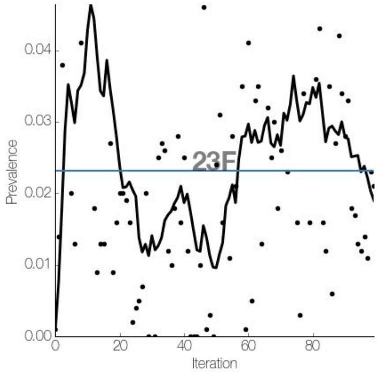


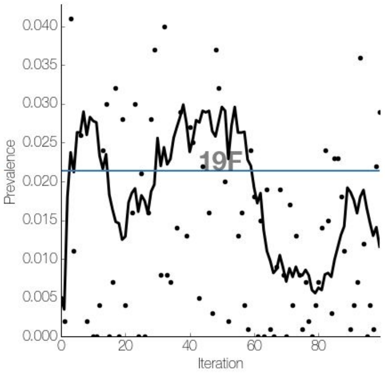


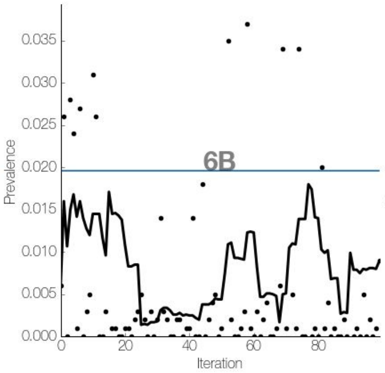


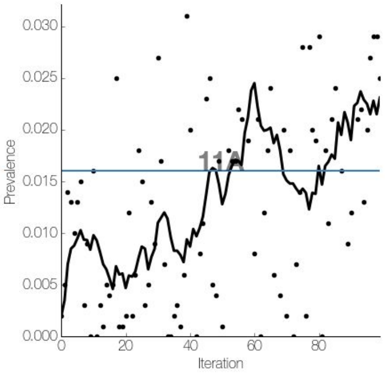




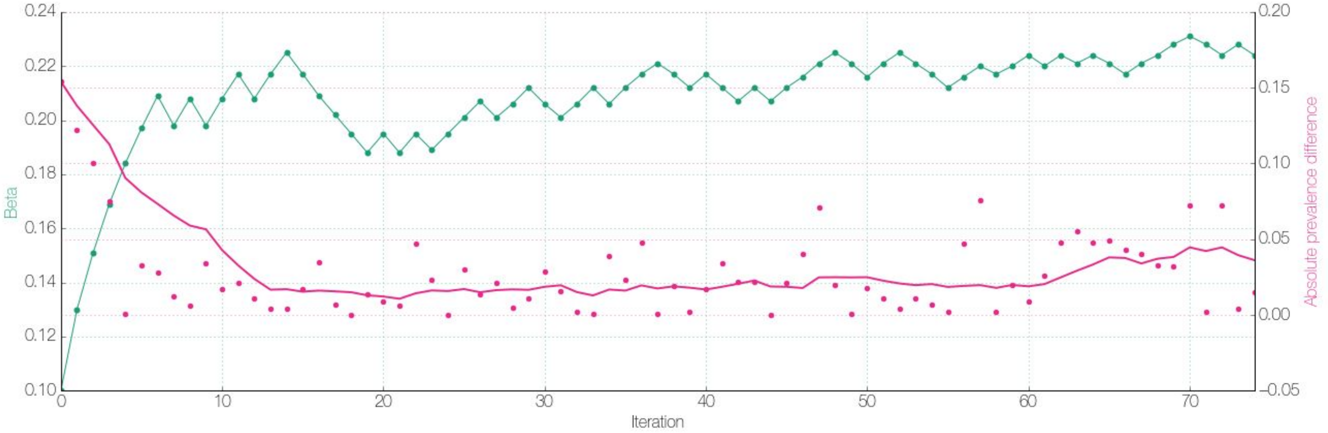


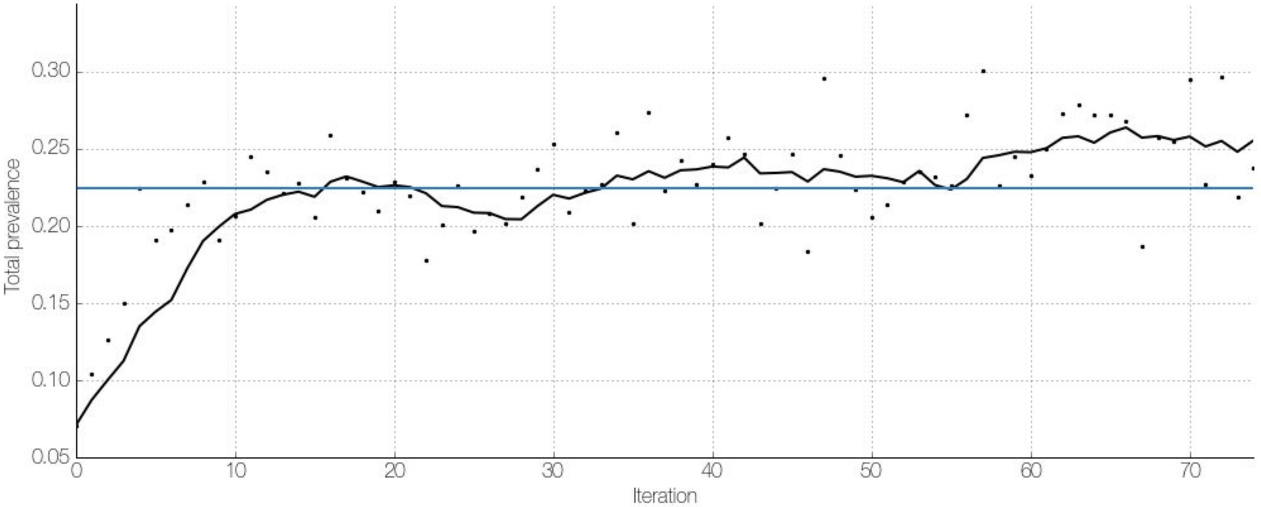




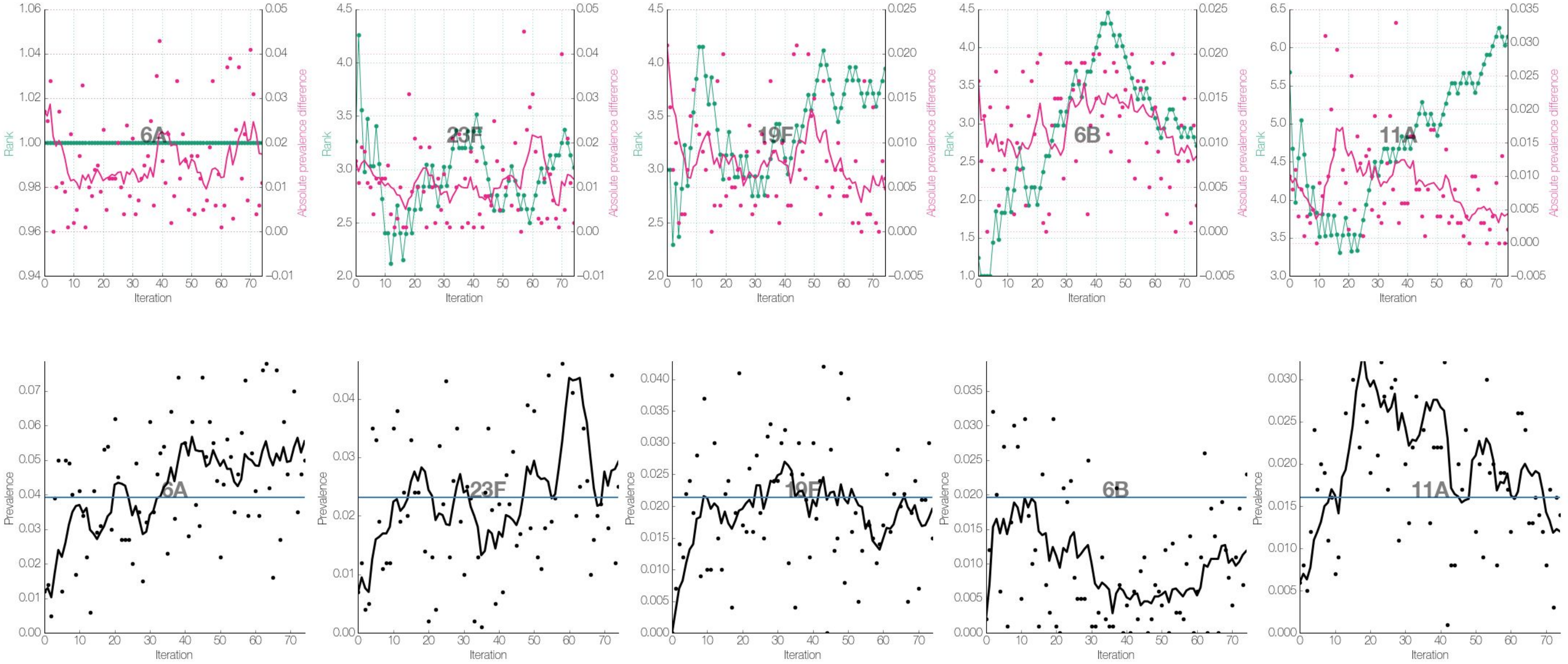


Averaging 5 simulations





Averaging 5 simulations



Next steps

- Immediate challenge seems to be stochastic noise affecting the fitting of individual serotype parameters.
- How do we reduce noise, without increasing computational time too much?
- Try simple ideas:
    - Use a combination of larger population sizes and averaging more simulations.
    - Instead of averaging, consider the *distribution* of results in each simulation set.
    - Perturb one parameter at a time for gradient estimation (runtime would increase dramatically, however).
- Look for a more principled approach to statistical inference in individual-based models
    - Currently exploring this as a possible project, with Professor Pierre Jacob at Harvard (Statistics)