Four applications of statistics to COVID-19 modelling

The Canadian Network for Modelling Infectious Diseases: Progress and Next Steps
BIRS 23w5151

Lloyd T. Elliott, November 16 2023

Statistic support for COVID-19 response

- Epidemiology:
    1. NPI modulation (MAGPIE)
    2. Prevalence estimates from serology (MAGPIE)
    3. True prevalence from case counts (UVic)
- Host genetics:
    4. The HostSeq project: Host genetics in Canada (Genome Canada/ CanCOGeN/CGEn & SickKids)

Acknowledgements

Statistical support team
Sonny Min
Renny Doig
Elika Garg
Olga Vishnyakova

MAGPIE collaboration
Caroline Colijn
Jessica Stockdale
Nicola Mulberry
Liangliang Wang

UVic collaboration
Matthew Parker
Junling Ma
Yangming Li
Laura Cowen
Jiguo Cao

HostSeq collaboration
Elika Garg
Paola Arguello Pascualli
Olga Vishnyakova
Steve Jones
Lisa Strug
Jennifer Brookes
Shelley Bull
France Gangnon
Celia Greenwood
Rayjean Hung
Jerry Lawless
Andrew Patterson
Lei Sun
Study PIs
Sub-Committees
Study participants

NSERC CRSNG

Michael Smith Health Research BC

Genome BritishColumbia
Leading ▸ Investing ▸ Connecting

CAN MOD
Canadian Network for Modelling Infectious Diseases
Réseau canadien de modélisation des maladies infectieuses

GenomeCanada

1. NPI modulation

Conclusion

After health authorities recommend or require changes to *non-pharmaceutical interventions* (e.g. adding or removing a mask mandate, starting or stopping a lockdown), it can take some time for the changes to be reflected in case counts. Sometimes up to 2.5 months.

J Stockdale, R Doig, J Min, N Mulberry, L Wang, L Elliott, C Colijn. *Long time frames to detect the impact of changing COVID-19 measures, [BC] Canada, March to July 2020*. Eurosurveillance 2021

Modelling effect of NPI change

- **If non-pharmaceutical interventions (NPIs) change, how long until we see a statistically significant change in case counts?** (Case counts are "noisy")

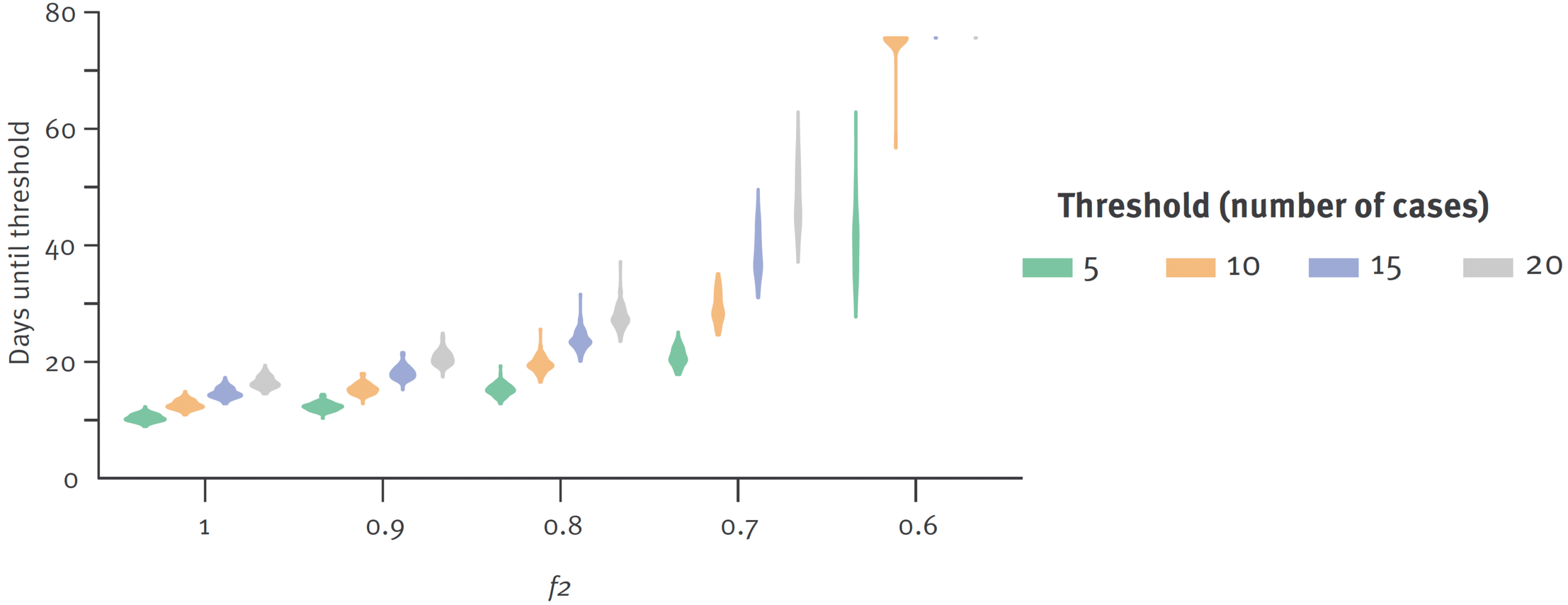
$$f(t) = \begin{cases} f_0 & \text{before distancing is enacted} \\ f_1 & \text{when distancing measures are in place} \\ f_2 & \text{after relaxation of distancing} \end{cases}$$

  - The distancing function f gives the amount of distancing at time t (f = 1 is no distancing)
  - For a given function f, we can simulate case counts (SEIR-type with quarantine and distancing; Anderson et al. 2020)
  - We can then compare the simulated case counts, and find out at what time they "differ significantly" (i.e., they differ more than what we'd expect from random fluctuations in case counts)

Definition of "differing significantly"

- Suppose scenario M1 is with more relaxation and scenario M2 is with no relaxation (the scenarios match until the time of relaxation)
- Simulate 100 case count trajectories from scenario M1, and 100 case count trajectories from scenario M2
- Find the first date for which 95% of the time, scenario M1's case count is higher than scenario M2's case count by an "alarm" threshold of "10" or more

M1: Relaxation to f2. M2: No relaxation



- We vary the "alarm" threshold: 5, 10, 15, 20
- In another section of work, we find MLE estimates: f2 = 0.65, f1 = 0.36
- Simulations were based on epidemiological parameter priors drawn from literature

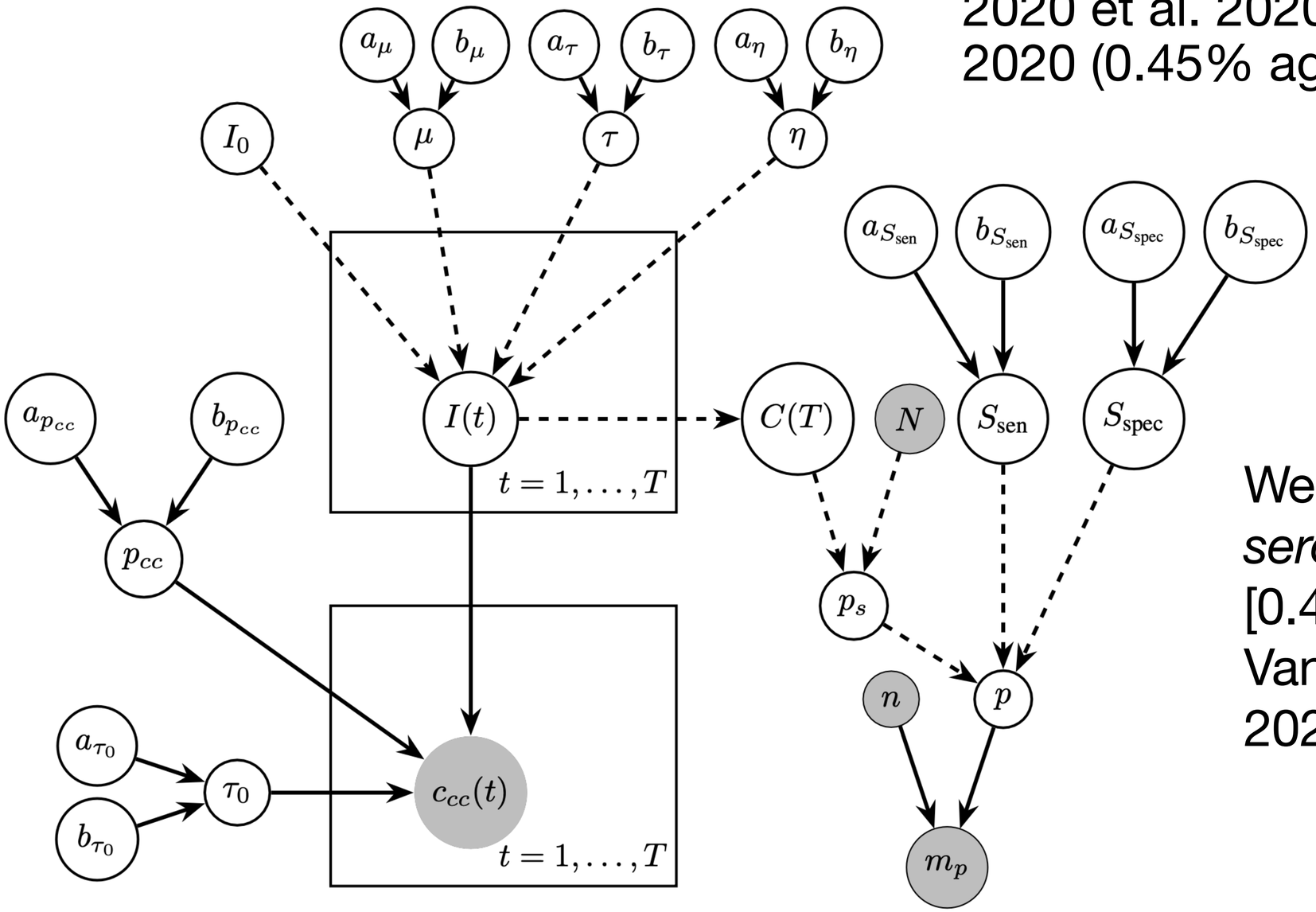
2. Prevalence estimates from serology

Conclusion

We can reduce variance of *seroprevalence* estimates using a joint model of both case counts and *serological survey*. We provide an estimate of *seroprevalence* of 0.57% [0.48%, 0.68%] in Vancouver on May 27 2020. This refines Skowronski et al. 2020's estimate of 0.55% [0.15%, 1.37%].

L Wang, J Min, R Doig, L Elliott and C Colijn. *Estimation of SARS-CoV-2 antibody prevalence through integration of serology and incidence data [Vancouver BC Canada, January to May 2020]*. 2022. Canadian Journal of Statistics

Case counts from BCCDC (January to May 2020). Serological survey from Skowronski 2020 et al. 2020: n = 885, mp = 4 on May 27 2020 (0.45% age standardized 0.55%)



We provide an estimate of *seroprevalence* of 0.57% [0.48%, 0.68%] in Vancouver on May 27 2020

4. The HostSeq project: Host genetics in Canada

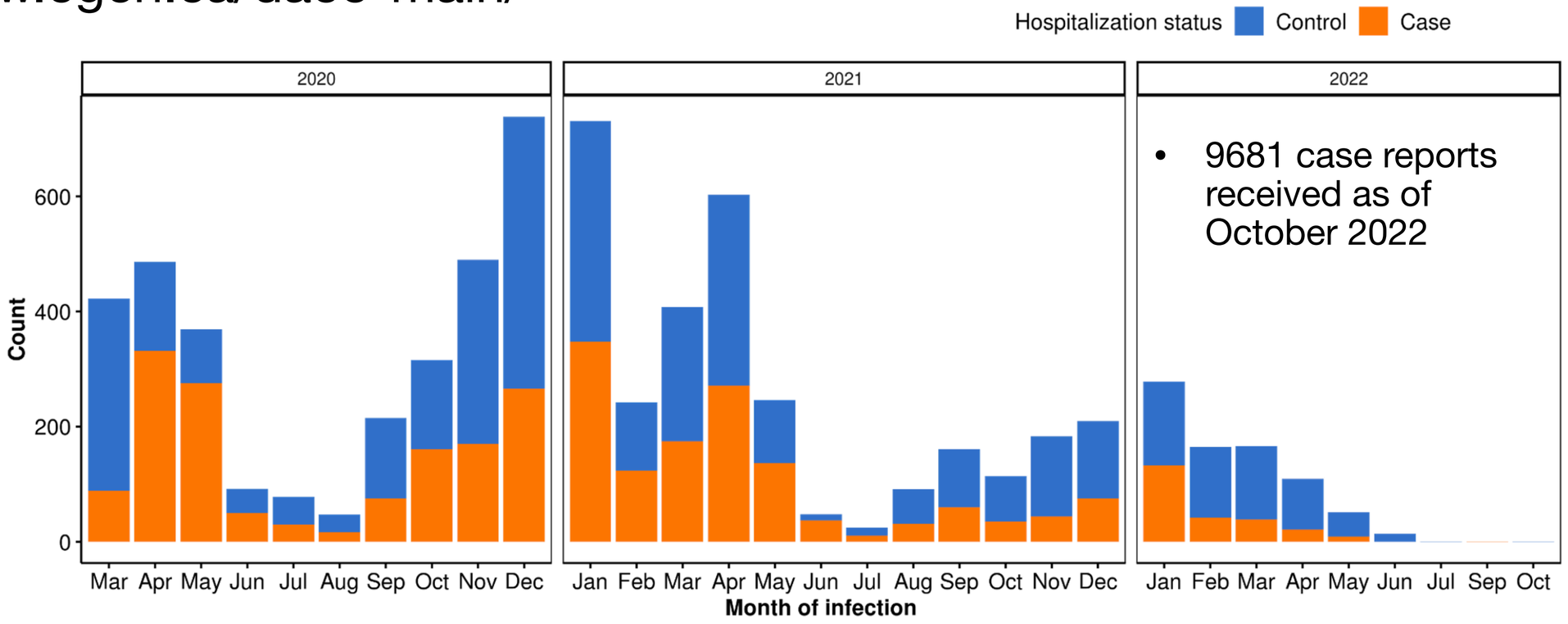
Question

How does human genetic variation modulate COVID-19 susceptibility and severity? The HostSeq project is Canada's contribution to the global effort to answer this question. Our mandate is to DNA sequence the human genome of 10,000 COVID-19 positive Canadians, and make these genomes and case reports available to researchers.

S Yoo, E Garg et al. *HostSeq: a Canadian whole genome sequencing and clinical data resource [Canada, March 2020 to October 2022]*. BMC Genomic Data 2023

HostSeq version 9

- 10,059 samples at 15 study sites across Canada released March 2023
- WGS and joint call with ~153M variants available through DACO https://www.cgen.ca/daco-main/

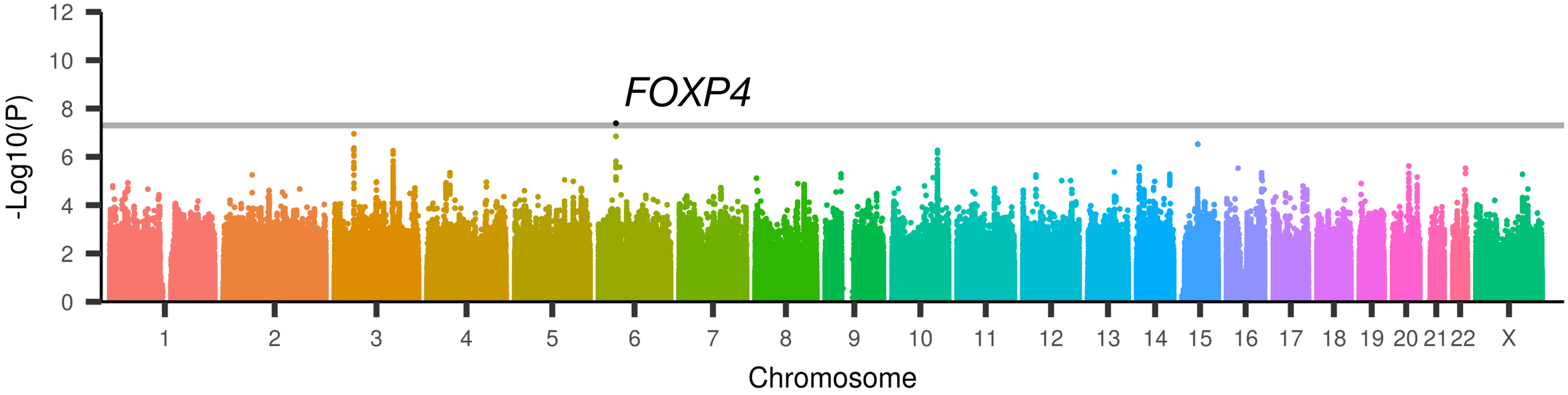


Analyzing HostSeq

- After QC (subset to complete cases), we retain 8,474 samples
- Imputed ancestries: 455 African (5.4%), 537 Admixed American (6.3%), 519 South Asian (6.1%), 654 East Asian (7.7%), 6107 European (72.1%), and 202 uncategorized (2.4%)
- We perform a genome-wide association study (GWAS) on the contrast B1: hospitalized cases *vs* non-hospitalized cases (HGI's B1)

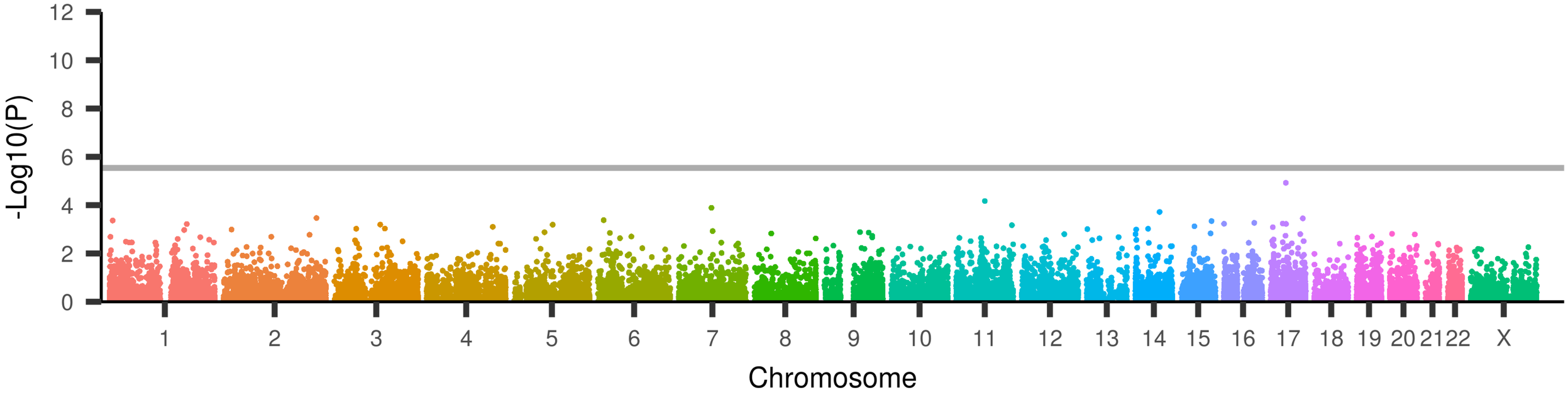
HostSeq GWAS results

- HostSeq univariate GWAS with LMM (regenie; N = 8,474; 4,708,250 variants with MAF<0.05 on GNOMAD easy to sequence regions)



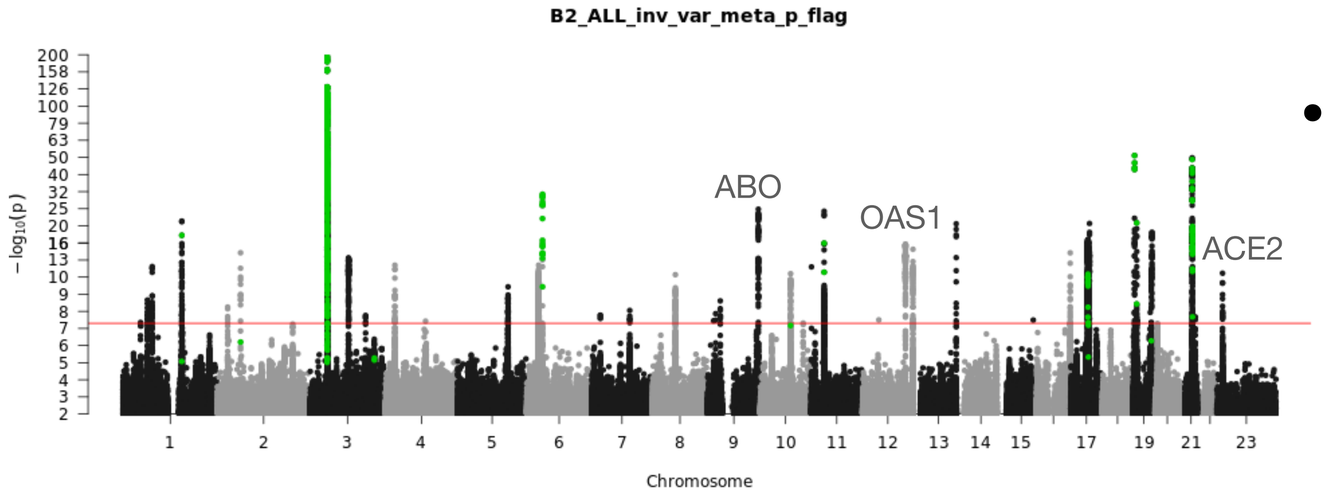
HostSeq GWAS results

- HostSeq rare variant *SNP-Set Kernel Association Optimal Unified Test* GWAS for Hospitalization (N = 8,474; variants on GNOMAD easy to sequence regions, 17351 genes)



HGI GWAS on B2 (hospitalized *vs* population)

- Host Genetics Initiative (HGI): Global consortium to meta-analyze COVID-19 host genetics
- Release 2: No hits. N = 1,332 (May 15 2020)
- Release 7: 91 hits. N = 289,919 cases (April 2022; 119 studies)



- Contrast: Hospitalized covid vs. population

We replicate 2/3 top B1 hits from HGI7

RSID	Nearest-Gene	Locus	REF	EFF	Data	EAF	BETA	SE	P
rs17763742	*SLC6A20*	chr3:45805277	A	G	HGI7	0.16	0.38	0.0320	2.40E-32
rs17763742	*SLC6A20*	chr3:45805277	A	G	HostSeq	0.10	0.33	0.0700	2.50E-06
rs2496646	*FOXP4-AS1*	chr6:41515629	T	C	HGI7	0.85	-0.29	0.0430	2.20E-11
rs2496646	*FOXP4-AS1*	chr6:41515629	T	C	HostSeq	0.91	-0.29	0.0760	1.80E-04
rs2834164	*IFNAR2*	chr21:33249643	A	C	HGI7	0.43	-0.10	0.0180	1.70E-08
rs2834164	*IFNAR2*	chr21:33249643	A	C	HostSeq	0.48	-0.09	0.0410	2.90E-02

- A polygenic risk score (PRS) with these genetic variants explains 1.01% of the variance in severe COVID-19 in our sample

E Garg et al. *Canadian COVID-19 host genetics cohort replicates known severity associations [Canada, March 2020 to October 2022]*. Under review

Thank you!

3. True prevalence from case counts

Question

- **How many COVID-19 infections are there at a given time?**
- Due to testing and reporting protocols, and asymptomatic cases, case counts published by centres for disease control may be underestimates
- Seroprevalence studies show that case counts underestimate (Skowronski et al. 2020)
- What is the true prevalence of COVID-19?

M Parker, Y Li, L Elliott, J Ma and L Cowen. *Under-reporting of COVID-19 in the Northern Health Authority region of British Columbia [March to October 2020]*. Canadian Journal of Statistics 2021

True prevalence

Methods

- Adapt the open population model from ecology (growth, importation, recovery, death)
- Usual open population model:

$$\mathscr{L} = \prod_{i=1}^{U} \left[\sum_{N_{i1}=n_{i1}}^{K} \cdots \sum_{N_{iM}=n_{iM}}^{K} \left\{ \left(\prod_{t=1}^{M} \text{Binom}(n_{it}; N_{it}, p) \right) \text{Pois}(N_{i1}; \lambda) \prod_{t=2}^{M} P_{N_{it-1}, N_{it}} \right\} \right]$$

$$P_{a,b} = \sum_{c=0}^{m=\min\{a,b\}} \text{Binom}(c; a, \omega) \text{Pois}(b - c; \gamma)$$

True prevalence

Methods

- Open population model for epidemiology:

$$\mathscr{L} = \sum_{N_1=n_1}^{K} \cdots \sum_{N_T=n_T}^{K} \left\{ \text{Pois}(N_1;\lambda) \cdot \left(\prod_{t=1}^{T} \text{Binom}(n_t; N_t - a_{t-1} + r_{t-1} + D_{t-1}, p) \right) \cdot \left(\prod_{t=2}^{T} P_{N_{t-1},N_t} \right) \cdot \left(\prod_{t=1}^{T-1} \text{Mult}(a_t - D_t - r_t, D_t, r_t; a_t, p_a, p_d, p_r) \cdot \sum_{R_t=r_t}^{N_t-D_t} \text{Mult}(A_t, D_t, R_t; N_t, p_a, p_d, p_r) \right) \right\}$$

$$P_{a,b} = \sum_{c=0}^{m=\min\{a,b\}} \text{Pois}(c;\omega a) \cdot \text{Pois}(b-c;\gamma)$$

3. True prevalence definitions

Statistics	
M	number of sampling sites
T	number of sampling occasions
a_{it}	detected cases still active at site i, time t
	$i \in \{1, 2, \ldots, M\}$, $t \in \{1, 2, \ldots, T\}$
n_{it}	new detected cases at site i, time t
d_{it}	new detected deaths at site i, time t
r_{it}	new detected recoveries at site i, time t
H_i	total population size at site i
Latent States	
N_{it}	total active cases at site i, time t
A_{it}	cases which remain active in site i from time t to $t+1$
D_{it}	cases which die in site i from time t to $t+1$
R_{it}	cases which recover in site i from time t to $t+1$
S_{it}	new cases from domestic spread in site i from time t to $t+1$
G_{it}	new cases from importation in site i from time t to $t+1$

Parameters	
λ	expected initial active cases per site
p_a	probability of a case remaining active
p_d	probability of a case dying
p_r	probability of a case recovering
ω_1	expected new domestic spread from unobserved cases
ω_2	expected new domestic spread from observed cases
γ	expected new imported cases per site
p	probability of detecting an active case
Derived Variables	
δ_i	fraction of population susceptible at site i
Ω_{it-1}	expected new domestic spread in site i time $t-1$ from all sources
α	inflation factor for proportion of observed deaths

3. True prevalence model

(1) Initial Active Cases: $N_{i1} \sim \text{Poisson}(\lambda)$

(2) Latent State Process: $\{A_{it}, D_{it}, R_{it}\} \sim \text{Multinomial}(N_{it}; p_a, p_d, p_r)$

(3) Detected Active Cases: $a_{it} = n_{it} + a_{it-1} - r_{it-1} - d_{it-1}$, for $t > 0$

(4) Domestic Spread: $S_{it} \sim \text{Poisson}(\Omega_{it-1})$, for $t > 1$

(5) Ω_{it-1} : $\omega_1(N_{it-1} - a_{it-1}) \cdot \delta_i + \omega_2 a_{it-1}$ (mean domestic spread)

(6) δ_i : $(H_i - N_{it})/H_i$ (fraction of susceptible population, where H_i is the total population size)

(7) Imported Cases: $G_{it} \sim \text{Poisson}(\gamma)$, for $t > 1$

(8) Active Cases Updates: $N_{it} = A_{it-1} + S_{it} + G_{it}$, for $t > 1$

(9) Observation Process I: $n_{it} \sim \text{Binomial}(N_{it} - a_{it-1}, p)$

(10) Observation Process II: $\{a_{it} - d_{it} - r_{it}, d_{it}, r_{it}\} \sim \text{Multinomial}(a_{it}; p_a^*, \alpha p_d, p_r)$

True prevalence

Results

- We found under-reporting rates up to 85% in Northern Health Authority of BC. Covariates considered:

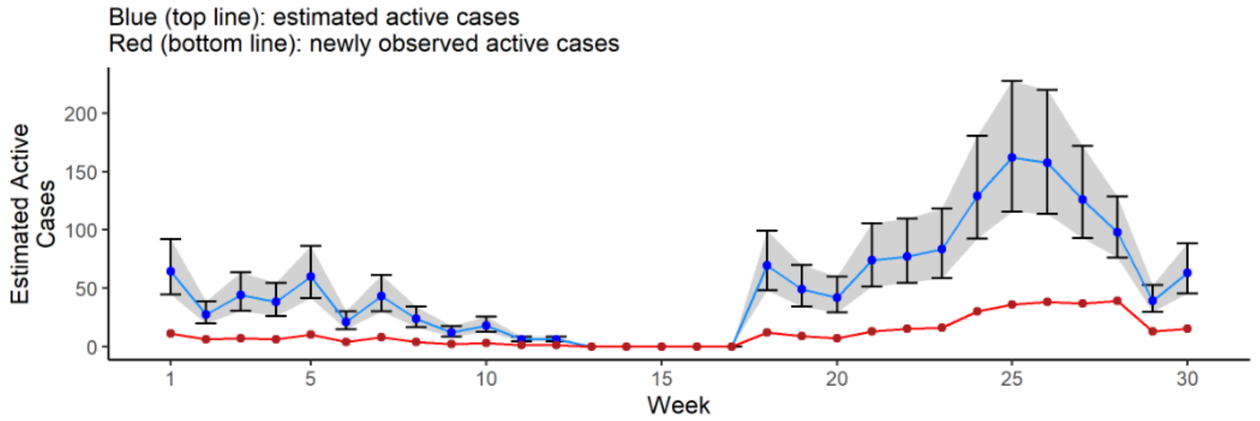


FIGURE 6: Estimated active cases $\hat{N}_t$ per week from the best fitted model (*mob*, *vol*), as chosen by AIC. Bottom line (red): newly observed active cases. Top line (blue): estimated active cases with 95% confidence intervals. $\hat{N}_t$ are calculated from the estimated probability of detection $\hat{p}_t$ and newly observed active cases by $\hat{N}_t = n_t/\hat{p}_t$.

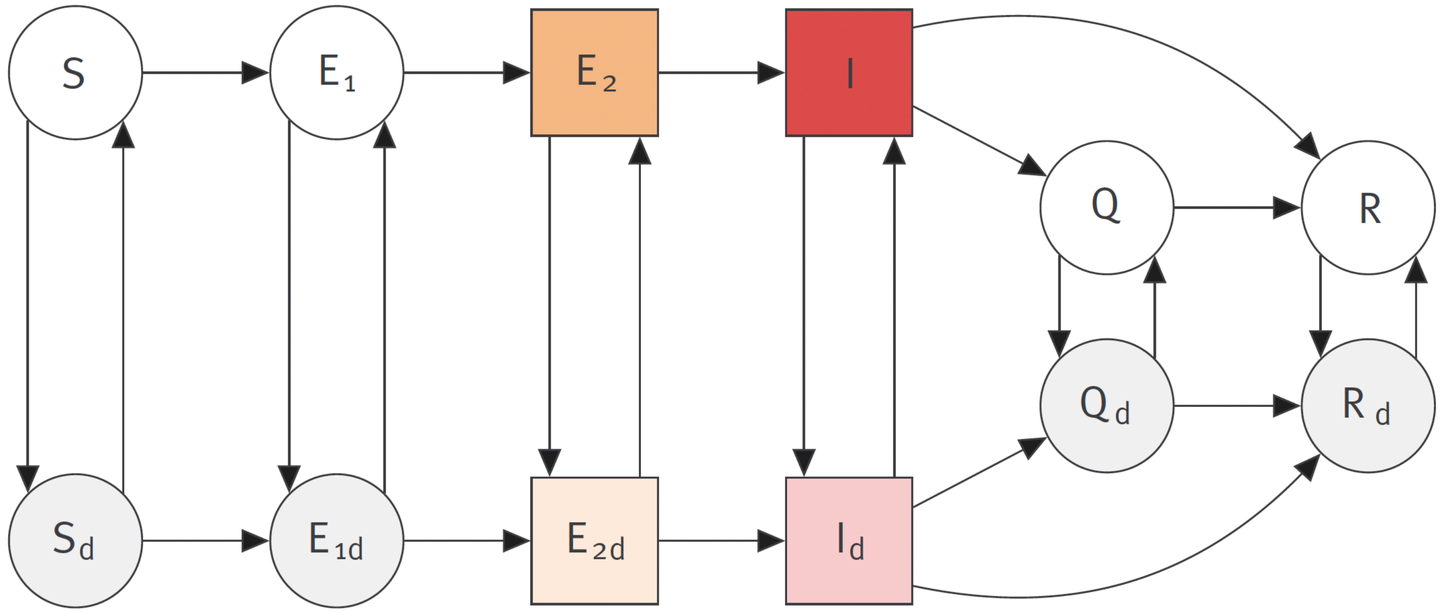
True prevalence

- We have extended this model to a multi-site model to a) all Health Authorities of BC, b) all provinces and territories of Canada. Both considering covariates: vaccine rate, and testing volume

M Parker, J Cao, L Cowen, L Elliott and J Ma. *Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada [Canada, April 2020 to January 2022]*. medrxiv preprint 10.1101/2022.07.11.22277508v1

1. SEIR-type with quarantine and distancing

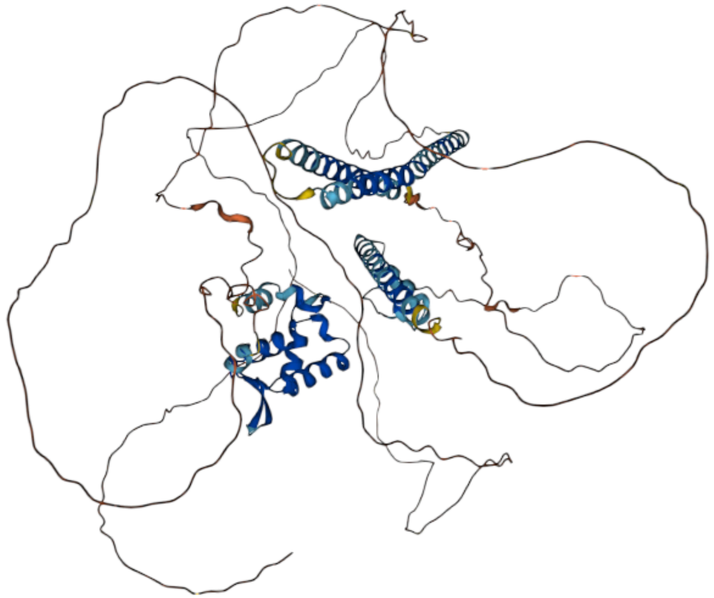
- We define M using an extension of the SEIR-type model (Susceptible, Exposed, Infected, Removed) developed by MAGPIE (Anderson et al. 2020). This is a compartment model, simulated through an ODE.



- Q - quarantined
- E1 - pre-symptomatic & not infectious
- E2 - pre-symptomatic & infectious
- d - physically distancing

4. Gene card: *FOXP4*

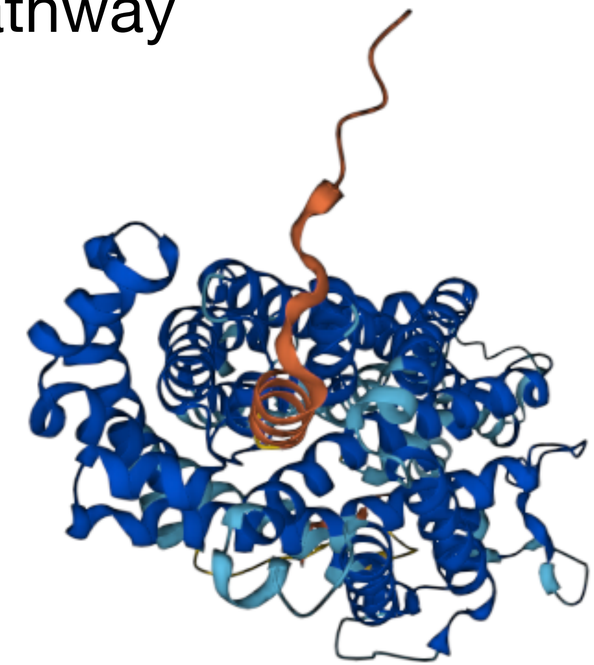
- Forkhead Box P4. A protein coding gene (coding Forkhead Box Protein P4).
- Located on 6p21.1, 56k bases long
- Gene transcription regulator, with some lung-specific regulation. May be involved in repressing some lung-specific expression. "... involved in the upkeep of healthy lung tissue ..."



genecards.org, AlphaFold and HGI

4. Gene card: *SLC6A20*

- Solute Carrier Family 6 Member 20. A protein coding gene (coding protein of same name)
- Located on 3p21.31, 41k bases long
- Transports small molecules across the cell membrane (prolines). Identified as a viral entry pathway



genecards.org, AlphaFold and HGI

4. HostSeq conditions