Tutorial

This tutorial will guide you through the basics of how to use the codes used in the manuscript of “Inferring transmission heterogeneity using virus genealogies: estimation and targeted prevention”.

Part I: inferring heterogeneity from virus genealogy. To use the following command, please firstly run the file of “CmpSimu\_Estimation.py”.

1. Simulate an outbreak

To simulate an outbreak, use the command

Re = singleSimu(mu\_InFC, mu\_Rec, cv\_InFC, smpSize, smpRatio, fntNum)

The meaning and default values of these parameters are as follows:

|  |  |  |
| --- | --- | --- |
| parameter | meaning | Default value |
| mu\_InFC | , the average transmissibility rate | 2.5 |
| mu\_Rec | , the rate of being diagnosis | 1 |
| cv\_InFC | , the coefficient of variation of the transmissibility rate. | 1 |
| smpSize | the number of being diagnosed, which determine the time of stopping the simulation. | 100 |
| smpRatio | , the sequencing ratio | 0.9 |
| fntNum | the size of susceptible individuals | 9999 –standing for infinite population size |

To visualize the simulated outbreak, use the command

simutree = Re[3];

Phylo.draw(Phylo.read(StringIO(simutree),"newick"))

1. Estimation of heterogeneity and epidemiological parameters given a virus genealogy

simuG = singleSimu(mu\_InFC, mu\_Rec, cv\_InFC, smpSize, smpRatio, fntNum)

inferRe = singleEst(simuG[0], simuG[1], simuG[2], back\_ratio, = [0.9,0.85,0.8], smpRatio = 0.9)

where simuG[0], simuG[1], simuG[2] are the data about the sampled genealogy from the simulated outbreak; “back\_ratio” means the parameter of “*p*” used in the analysis of sampled virus genealogy; and “smpRatio” stands for the sequencing ratio.

The output of “inferRe” contains three components: the estimated , , and

1. multiple simulations and the visualization of the results

To get the performance of the proposed method under various levels of heterogeneity (as in Figure 2 in the manuscript), use the following command,

mtpSimu = cmpEst()

EstRePlot(mtpSimu)

1. Analysis of real data (Run the file of “AnaRealData.py”)
2. To apply to a real virus genealogy (which has been written in Newick format in a txt file), use the command:

infRe = anaRealData(filename)

the output of inference include the estimated , , as well as the Sackin’s index of the virus genealogy.

1. To get the results of analyzing the three real datasets (as in Figure 6 in the manuscript), use the following command.

idua = anaRealData\_IDUAE(path);

idub = anaRealData\_IDUB(path);

msmB = anaRealData\_MSMB(path);

# plot the results

rePlot(idua,idub,msmB)

where ‘path’ is the folder where the real datasets are storied.

Part II: study the performance of phylogeny-guided prevention based on simulation.

1. Simulation study under the continuous monitoring scenario (Run the file of “SSE\_CMsimu.py”)

To show the relative infection size and the fraction of contact traced of the phylogeny-guided strategy and the random strategy under various level of heterogeneity (as in the Figure 7 in the manuscript), use the command:

comCMPlot(CV, simuK)

where “CV” is an array of the level of heterogeneity; and “simuK” is the times of simulation. The output of this function is similar to Figure 7 in the manuscript.

1. Simulation study under the cross-sectional scenario (Run the file of “SSE\_CSsimu.py”)

To compare the relative effect of the NCE-based strategy and the MRD strategy (as in the Figure 10 in the manuscript), use the command:

comCSPlot (CtRatio, simuK)

where “CtRatio” is an array of the prevention fraction; and “simuK” is the times of simulation.