Using mobile phone data to model influenza epidemics

André Lichtsteiner

Supervisors: Tim Vaughan, David Welch and Alexei Drummond

Summer Research Project SCI 129 (2016/2017)

Introduction

Purpose of the project

The project aims to improve scientific understanding of the spread of influenza in New Zealand, by incorporating data about movement of people into the model.

This is important for informing the New Zealand health sector about how best to respond to and manage future outbreaks of influenza and other viral pathogens.

Project Overview

We created a higher spatial resolution model of the spread of influenza in New Zealand by considering this at the level of District Health Boards (DHBs), of which there are twenty.

The number of genetic sequences available is quite limited, which adds to the challenge of inferring population sizes for, and rates between all of the DHBs. This is particularly impractical for those DHBs for which there are no genetic sequences sampled. It is simply asking too much of the structured coalescent model to accurately infer 380 migration rates and 20 population sizes while simultaneously inferring the phylogeny from 50 genetic sequences from only some of the DHBs. This is the motivation for using an external source to inform the migration model (population sizes and migration rates).

We used aggregated mobile phone location data provided by Qrious (a subsidiary of Spark) to approximate movement of people.

We were able to incorporate the phone data into the model, but we ran into challenges when trying to compare that model to the fully inferred model. The majority of my time was used in investigating ways that we could compare the models, to see if the model informed by the phone data is actually better than other models.

The big advancement towards comparing the models came from a paper by Lartillot and Philippe, which outlines a method they call “Model Switch Integration” (Lartillot & Philippe, 2006).

This method progressively changes the analysis from using one model to using the other, ie. it “switches” between them slowly. The advantage of this approach was that it works in situations (such as ours) where other approaches such as path sampling are unsuccessful. Path sampling for our data and model is infeasible because it requires you to sample from the prior, which in our case is highly dimensional and difficult to sample from, so using this method was a big step forward.

I implemented this method in a new BEAST 2 package so that we could use it with our models, and this is now available for use. *(Page 4).*

Overview of results

Our results show that the mobile phone data is better for informing the model than a migration model having all migration rates and populations being the same. Other possible explanatory migration models, such as distance based models have also been considered and compared. *(Page 5).*

Overview of the Topic

Bayesian Phylogenetics

This project is based around Bayesian inference of phylogenetic trees, which is based on Bayes' theorem, and uses the Markov Chain Monte Carlo (MCMC) technique to effectively obtain a sample from all possible trees which could explain the genetic sequence data.

MCMC is invaluable because it is facilitates sampling from distributions where direct sampling is difficult or infeasible, such as in the case of phylogenetics.

Bayesian inference relies on the Bayes rule, which is:

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

Terms:

- D - sequence data
- θ - model parameter

Components:

- P(θ) - the prior distribution
- P(D|θ) - the likelihood function
- P(θ|D) - the posterior distribution

An explanation of these components can be found on Wikipedia.

The BEAST 2 (Bayesian Evolutionary Analysis by Sampling Trees) software uses MCMC for inferring phylogenies, and this is what we used in this project. It was developed primarily by people in the Centre for Computational Evolution (Bouckaert et al, 2014).

We are interested in movement of flu between DHBs, and this separation of a population (NZ, or the world) into smaller groups (DHBs) is called having a population structure. The structured coalescent model allows us to incorporate population structure into the inference, and Tim Vaughan's MultiTypeTree package adds support for this analysis to BEAST (Vaughan et al, 2014).

All of our phylogenetic analyses are using BEAST 2 and MultiTypeTree for the core operations.

Technical

MultiTypeTree has a migration model element, which itself has migration rate matrix and population sizes. The rate matrix represents the probability of the phylogeny moving (backward in time) between locations. The population sizes are in terms of effective population size, which is distinct from the actual number of individuals, and is related to the population structure. BEAST is able to sample the distribution for these when they are unknown, but this involves significant dimensionality. For an idea of this - having 20 DHBs means sampling for 20 population sizes, along with 380 migration rates. Compared to a model with only 2 areas (eg. North Island/South Island split), adding an additional 18 areas adds significant complexity to the model and makes sampling much more computationally demanding. Some DHBs have not had any sequences sampled, which makes this even more challenging.

This project investigates how we can eliminate (some of) that complexity by providing the model with migration and population data from elsewhere (ie. mobile phone movement) rather than inferring it.

Project Development

Preparing the phone data

The mobile phone data we used was provided to us by Qrious, a Spark subsidiary, and is highly aggregated.

The data provided is split up into approximately 30 Regional Tourism Organisations (RTOs), but our genetic sequences have locations associated with them based on the DHB they are from. For each RTO, we were given information about how many hours were spent by people whose home is in that RTO, in each of the other RTOs. This was the basis for our movement model.

Tim wrote a script which calculated the intersection between RTOs and DHBs. I used the intersection to convert the data into a format that can be understood by BEAST and MTT (a backward in time rate matrix). For simplicity, we assumed uniform spatial distribution of people across NZ (which is not realistic). We used human population numbers from Statistics New Zealand as the base for the population sizes in our model (assuming that this should roughly correlate with flu population size).

In the structured coalescent model, the rate matrix and population sizes have specific meanings and units, which our transformed data did not match - namely we didn't know the true scale of any of the values, so we treated them as relative values, and inferred a scaling parameter for the population sizes and another for the rate matrix.

The code used in the transformation is available at: https://github.com/andre-lichtsteiner/summer_research_misc

Considering the wider world

Our original DHB model based on the phone data did not consider the rest of the world - that is it did not have any location other than the DHBs. This forces the model to work under the implicit assumption that there must be a common ancestor of all strains of the flu genomes sampled somewhere in New Zealand. This is not realistic, as there is evidence for seasonal influenza entering New Zealand from elsewhere in the world (Bedford et al., 2010).

To account for this, we added a world deme (location) and sampled for the size of the effective population of it, along with the rates of migration, and by extension, the number of migration events over the time period of the tree. We limited this to migrations into New Zealand from the world only, that is no migrations to the world from New Zealand, to simplify the model.

Model Validity

Using the methods outlined in the next section, we tested the various models against each other, estimating the Bayes factor. Note that small changes to the way that a model is set up can have considerable effects on the Bayes factor result, so these numbers are very specific to the exact models and priors used. That is, when comparing models, the "model" is more specific than just any structured coalescent model with a particular migration rate matrix and population size parameter.

I have tried to account for this by, as much as possible, in each analysis keeping both models the same *except* for the migration rates and population sizes.

Assessing the usefulness of the phone data

Once we had a basic model which incorporated the mobile phone data, we needed to quantitatively measure how well that model fits the data. In the Bayesian paradigm, the most common and general measure of relative model fit is the Bayes factor, which reflects how much better one model is than another at explaining the data.

Challenges of doing this

There are a few methods such as the AIC (Akaike Information Criterion) and HME (Harmonic Mean Estimator) which provide a rudimentary way to calculate a Bayes factor. In more recent years, these methods have been shown to be very inaccurate (Lartillot & Philippe, 2006), so with that in mind we wanted to use a more different method.

A relatively recent development in this area is path sampling (as well as the similar stepping stone). These methods rely on progressively running many short analyses, moving along a path from the prior to the unnormalised posterior of the model. This works well in many cases and is known to be accurate.

We attempted to use this method, but it was not successful - sampling from the prior was not feasible due to how many dimensions our data has, as well as how wide some of our non-informative (not highly specific) priors needed to be. These included our priors on the scale parameters controlling the overall size of the rate matrix and population sizes. We used log normal priors for these to reflect that we had no information as to the scale these parameters should be.

Search for a different approach

When it became apparent that we would not be able to assess our model using path sampling, we tried other techniques, including various linear model style approaches. None of these really worked as we had hoped.

We found a paper by Lartillot and Philippe (2006) which describes a method called model-switch integration. In this method, rather than sampling along a path from prior to posterior for each model, the models are progressively switched from one to the other. This seemed promising for us, as it doesn't involve sampling from the prior, which was the root of our difficulty in using path sampling.

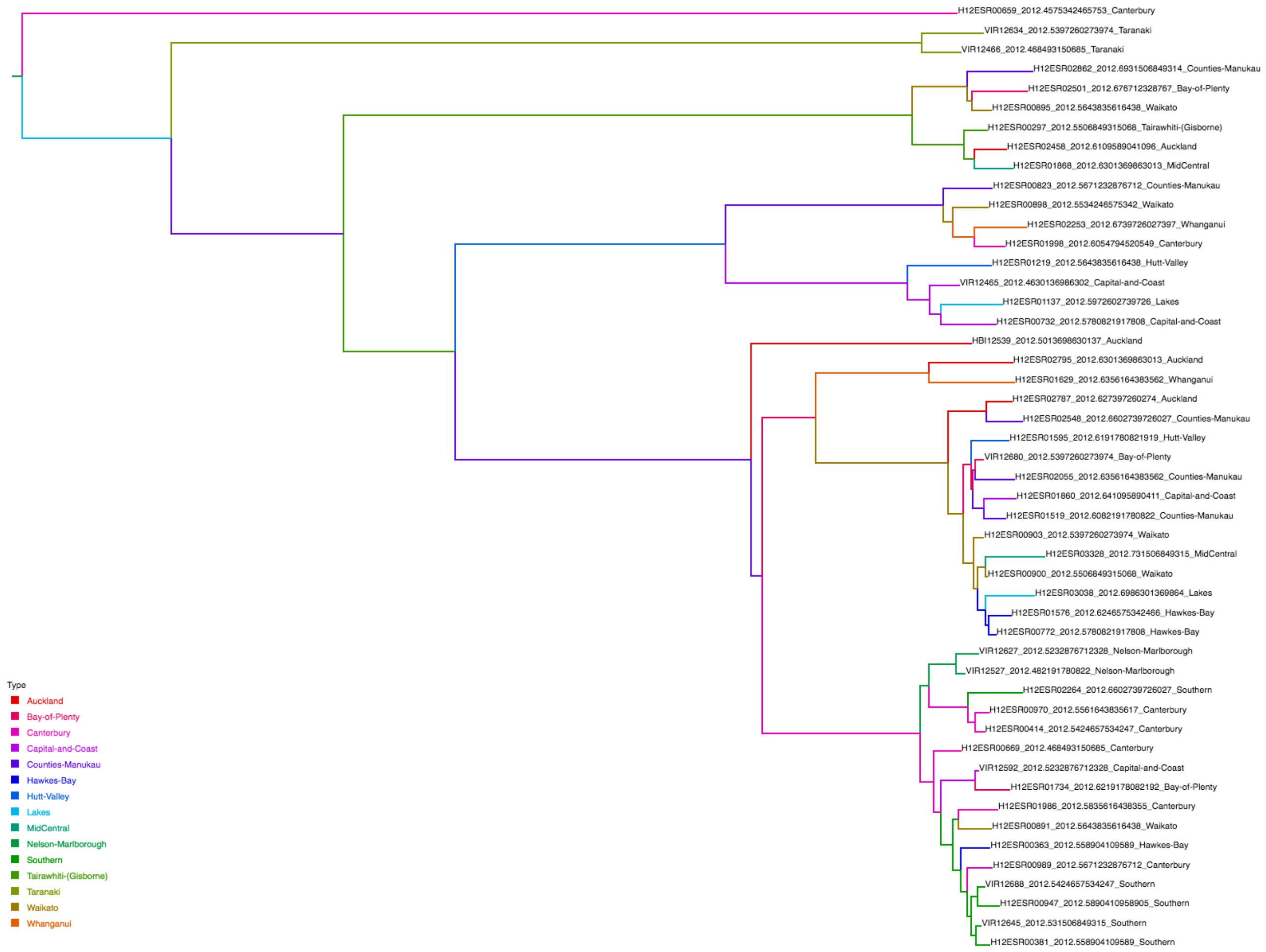
I implemented their technique in a BEAST 2 package in order to use it to assess our model. This is open source and is available for other researchers to install and use. This should be useful for other projects where path sampling is not feasible. More information can be found at: https://andre-lichtsteiner.github.io/ModelComparison/

We did make one main change to the technique in our implementation - we set it up such the models are moved between constantly - incrementing slightly after every step in the MCMC chain. In the paper, this increment was done at set intervals rather than at every step. We expect that this will increase the accuracy of the results.

Results

Inference: all rates and population sizes equal, 15 DHBs only *(tree 1)*

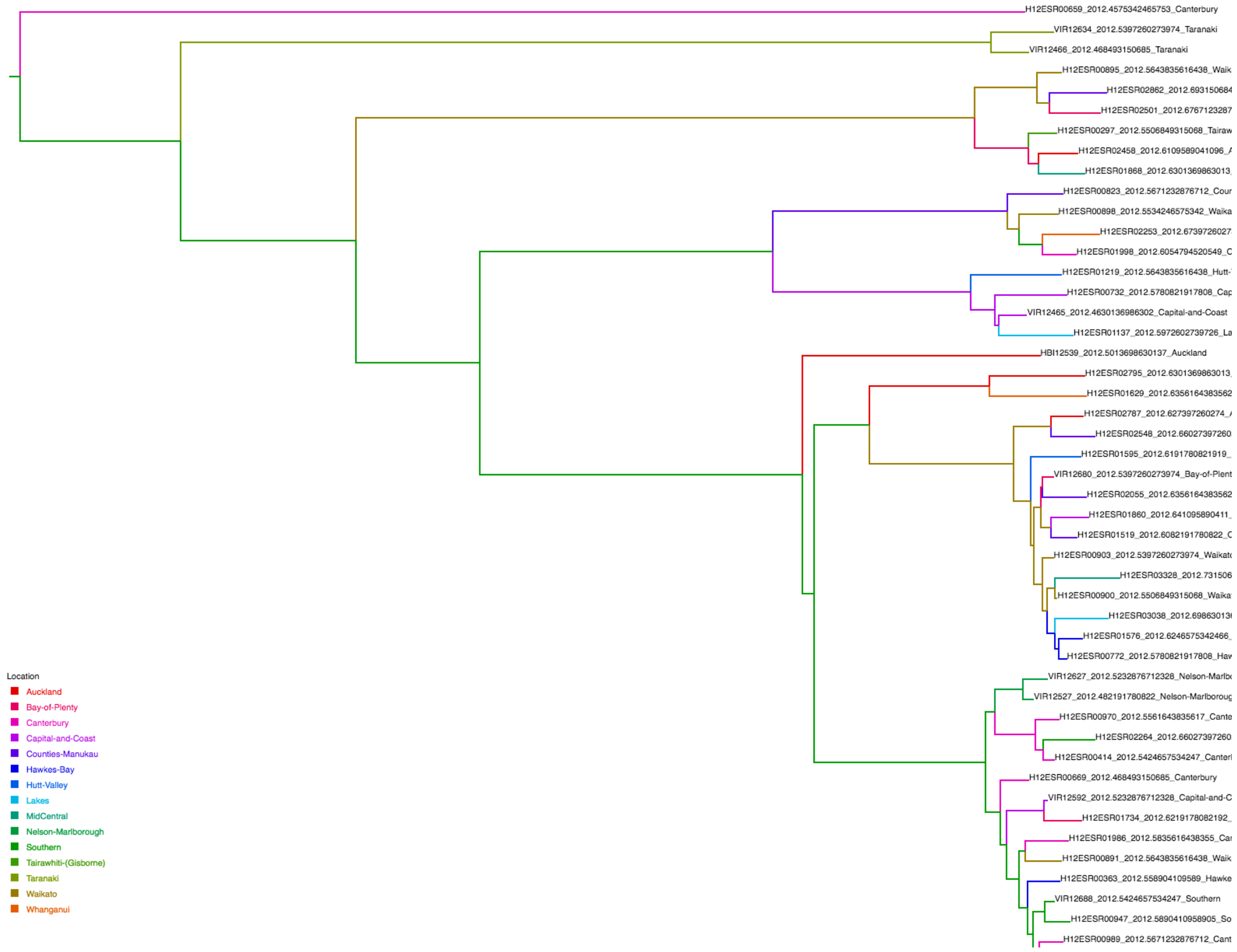
This only includes the 15 DHBs which had genetic sequence samples in 2012, 50 sequences in total. (Note that this was a relatively short chain, as the results from this are not central to this project)


H12ESR00659_2012.4575342465753_Canterbury
VIR12634_2012.5397260273974_Taranaki
VIR12466_2012.468493150685_Taranaki
H12ESR02862_2012.6931506849314_Counties-Manukau
H12ESR02501_2012.676712328767_Bay-of-Plenty
H12ESR00895_2012.5643835616438_Waikato
H12ESR00297_2012.5506849315068_Tairawhiti-(Gisborne)
H12ESR02458_2012.6109589041096_Auckland
H12ESR01868_2012.6301369863013_MidCentral
H12ESR00823_2012.5671232876712_Counties-Manukau
H12ESR00898_2012.5534246575342_Waikato
H12ESR02253_2012.6739726027397_Whanganui
H12ESR01998_2012.6054794520549_Canterbury
H12ESR01219_2012.5643835616438_Hutt-Valley
VIR12465_2012.4630136986302_Capital-and-Coast
H12ESR01137_2012.5972602739726_Lakes
H12ESR00732_2012.5780821917808_Capital-and-Coast
HBI12539_2012.5013698630137_Auckland
H12ESR02795_2012.6301369863013_Auckland
H12ESR01629_2012.6356164383562_Whanganui
H12ESR02787_2012.627397260274_Auckland
H12ESR02548_2012.6602739726027_Counties-Manukau
H12ESR01595_2012.6191780821919_Hutt-Valley
VIR12680_2012.5397260273974_Bay-of-Plenty
H12ESR02055_2012.6356164383562_Counties-Manukau
H12ESR01860_2012.641095890411_Capital-and-Coast
H12ESR01519_2012.6082191780822_Counties-Manukau
H12ESR00903_2012.5397260273974_Waikato
H12ESR03328_2012.731506849315_MidCentral
H12ESR00900_2012.5506849315068_Waikato
H12ESR03038_2012.6986301369864_Lakes
H12ESR01576_2012.6246575342466_Hawkes-Bay
H12ESR00772_2012.5780821917808_Hawkes-Bay
VIR12627_2012.5232876712328_Nelson-Marlborough
VIR12527_2012.482191780822_Nelson-Marlborough
H12ESR02264_2012.6602739726027_Southern
H12ESR00970_2012.5561643835617_Canterbury
H12ESR00414_2012.5424657534247_Canterbury
H12ESR00669_2012.468493150685_Canterbury
VIR12592_2012.5232876712328_Capital-and-Coast
H12ESR01734_2012.6219178082192_Bay-of-Plenty
H12ESR01986_2012.5835616438355_Canterbury
H12ESR00891_2012.5643835616438_Waikato
H12ESR00363_2012.558904109589_Hawkes-Bay
H12ESR00989_2012.5671232876712_Canterbury
VIR12688_2012.5424657534247_Southern
H12ESR00947_2012.5890410958905_Southern
VIR12645_2012.531506849315_Southern
H12ESR00381_2012.558904109589_Southern
Type
Auckland
Bay-of-Plenty
Canterbury
Capital-and-Coast
Counties-Manukau
Hawkes-Bay
Hutt-Valley
Lakes
MidCentral
Nelson-Marlborough
Southern
Tairawhiti-(Gisborne)
Taranaki
Waikato
Whanganui

This inferred phylogeny (maximum clade credibility) has locations being all over the place throughout the tree.
(Here we also ignore any effect from migration between the DHBs for which there were no genetic samples, as these were completely excluded from the analysis.)

Inference: using the phone data, 15 DHBs only *(tree 2)*

Similar setup to the previous, but using the transformed mobile phone data as the migration rates and human populations in DHBs as the population size.


H12ESR00659_2012.4575342465753_Canterbury
VIR12634_2012.5397260273974_Taranaki
VIR12466_2012.468493150685_Taranaki
H12ESR00895_2012.5643835616438_Waik
H12ESR02862_2012.693150684
H12ESR02501_2012.6767123287
H12ESR00297_2012.5506849315068_Tairaw
H12ESR02458_2012.6109589041096_A
H12ESR01868_2012.6301369863013_
H12ESR00823_2012.5671232876712_Cour
H12ESR00898_2012.5534246575342_Waika
H12ESR02253_2012.6739726027
H12ESR01998_2012.6054794520549_C
H12ESR01219_2012.5643835616438_Hutt-
H12ESR00732_2012.5780821917808_Cap
VIR12465_2012.4630136986302_Capital-and-Coast
H12ESR01137_2012.5972602739726_La
HBI12539_2012.5013698630137_Auckland
H12ESR02795_2012.6301369863013_
H12ESR01629_2012.6356164383562
H12ESR02787_2012.627397260274_A
H12ESR02548_2012.660273972603
H12ESR01595_2012.6191780821919_
VIR12680_2012.5397260273974_Bay-of-Plent
H12ESR02055_2012.6356164383562
H12ESR01860_2012.641095890411_
H12ESR01519_2012.6082191780822_C
H12ESR00903_2012.5397260273974_Waikato
H12ESR03328_2012.731506
H12ESR00900_2012.5506849315068_Waika
H12ESR03038_2012.69863013
H12ESR01576_2012.6246575342466_
H12ESR00772_2012.5780821917808_Hav
VIR12627_2012.5232876712328_Nelson-Marlb
VIR12527_2012.482191780822_Nelson-Marlboroug
H12ESR00970_2012.5561643835617_Cante
H12ESR02264_2012.660273972603
H12ESR00414_2012.5424657534247_Canterl
H12ESR00669_2012.468493150685_Canterbury
VIR12592_2012.5232876712328_Capital-and-C
H12ESR01734_2012.6219178082192_
H12ESR01986_2012.5835616438355_Car
H12ESR00891_2012.5643835616438_Waik
H12ESR00363_2012.558904109589_Hawke
VIR12688_2012.5424657534247_Southern
H12ESR00947_2012.5890410958905_So
H12ESR00989_2012.5671232876712_Cant
Location
Auckland
Bay-of-Plenty
Canterbury
Capital-and-Coast
Counties-Manukau
Hawkes-Bay
Hutt-Valley
Lakes
MidCentral
Nelson-Marlborough
Southern
Tairawhiti-(Gisborne)
Taranaki
Waikato
Whanganui

This inferred phylogeny (MCC) shows the most recent common ancestor (MRCA) as being in the Southern DHB, (the long green line segments). This tree differs in the distribution of locations from the previous tree (as we would expect). However, we do not expect that this is a reliable phylogeny, as this model forces the MRCA to be somewhere in New Zealand.
(As in the previous tree. we ignore any effect from migration between the DHBs for which there were no genetic samples, as these were completely excluded from the analysis.

Inference of <u>all</u> DHBs (plus "rest of the world")

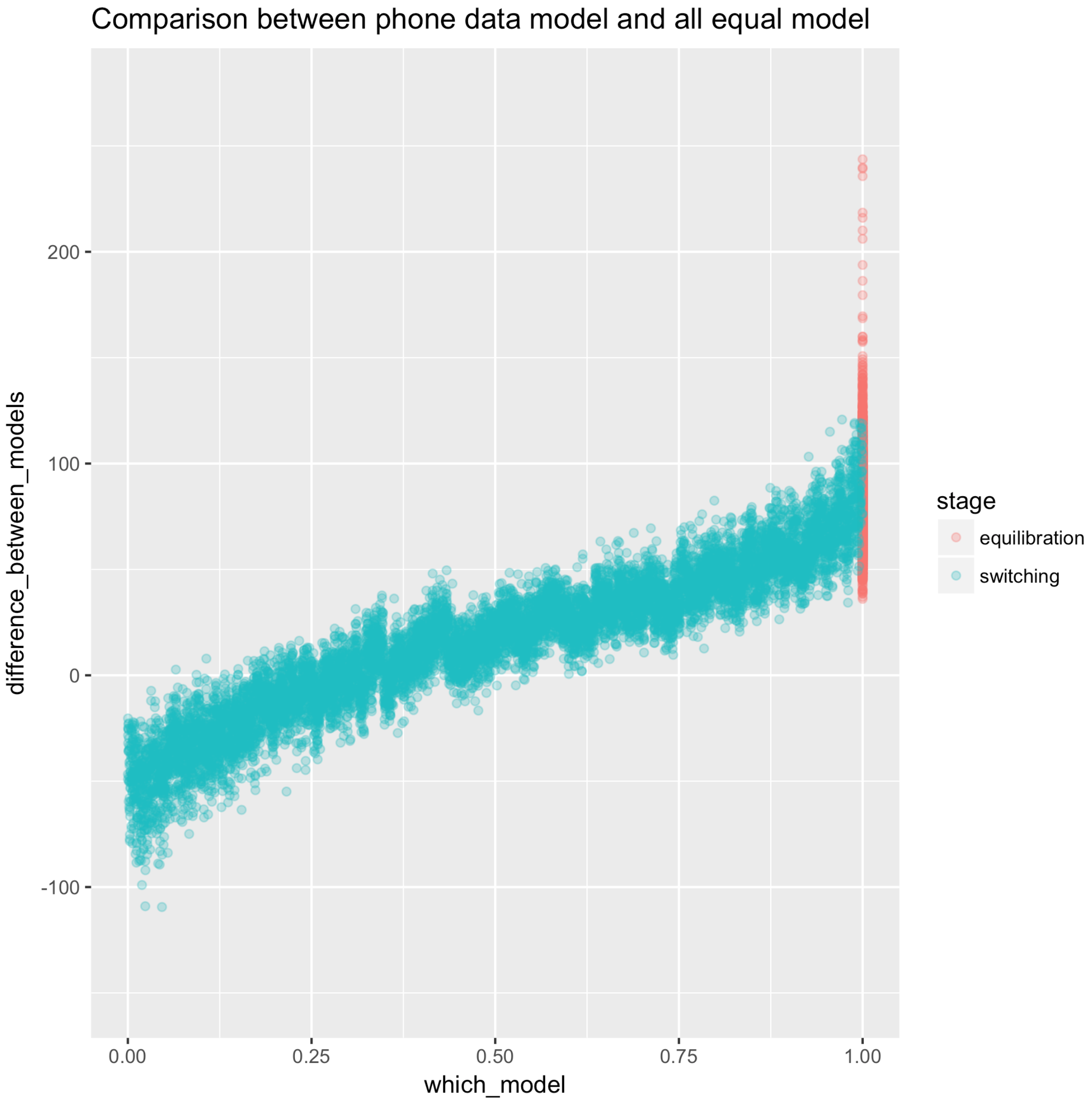
I was unable to produce results where migration rates and population sizes for all DHBs were inferred. I believe this is due to there being insufficient information in the sequence data, but that may not be correct. More sequence data, ideally with multiple sequences for all DHBs would help this significantly, but may still be challenging to sample reliably as there are so many demes.

Comparison: mobile phone data model vs all rates, pop sizes equal

Using the package I created for BEAST 2, I was able to estimate a Bayes factor for the mobile phone data model vs the all equal model.

Using that package I calculated a log Bayes factor of 16.36 (2dp), in favour of the mobile phone data model. Note that "model" here is very specific to the exact specifications of the model including its priors and any restrictions on values that the parameters may take, so this Bayes factor applies specifically to the two models as specified in the BEAST XML file. With that in mind, this Bayes factor indicates very strong support for the mobile phone data model (with human population sizes) over the model where all rates and all population sizes are equal.

See the below graph (from the above chain) for the difference between the two model posteriors along the chain. Note that model 0 is the all equal model and model 1 is the phone data model.



References

Bedford, T., Cobey, S., Beerli, P., Pascual, M. (2010). Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathogens 6*(5) doi: 10.1371/journal.ppat.1000918

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T. G., Wu, CH,, et al. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology, 10*(4). doi: 10.1371/journal.pcbi.1003537

Lartillot, N., Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology. 55*(2), 195-207. doi: 10.1080/10635150500433722

Vaughan, T. G., Kühnert, D, Popinga, A, Welch, D, Drummond, A. J. (2014). Efficient Bayesian inference under the structured coalescent. *Bioinformatics, 30* (16), 2272-2279. doi: 10.1093/bioinformatics/btu201