Influenza Phylodynamics: Stochastic simulation and analysis of seasonal influenza in New Zealand

Allan Wilson Centre
University of Auckland
Summer Internship 2012/2013

Student: Alexandra N Popinga
Supervisor: Alexei Drummond

I. Introduction

Influenza

Human influenza viruses cause contagious respiratory illnesses in the global population, resulting in ~500,000 human deaths annually. High-latitude countries experience influenza outbreaks primarily during winter months, and the asynchronous seasonality between the Southern and Northern Hemispheres results in two distinct flu seasons per year on the global level. This biannual nature of human influenza raises a fundamental question regarding its dynamics: do new seasonal epidemics derive from low-level viral lineage persistence within the local population, or instead stem from other regions? If the emergence of new strains could be attributed to viral migration from outside regions it may provide a means of predicting and minimizing spread (Russell *et al.*, 2008).

Phylodynamics

A complete understanding of global influenza dynamics is vital to the prevention of future outbreaks, and the perspective of phylodynamics plays an increasingly important role in the development of strategies for control and surveillance of influenza viruses. Phylodynamics is a computational study of behavior resulting from an interplay between ecological and evolutionary processes. It is particularly useful in characterizing viral phylogenies and transportation due to the complex interaction between rapid evolution and epidemiological processes of these species. The genetic variation and evolutionary trajectory in viral pathogens can be further utilized to describe epidemic transmissions; these systems can be combined by placing them on a common time scale or spatial frame of reference and can be applied to both local and global investigations (Vijaykrishna *et al.*, 2011).

New Zealand

As a region located primarily between 35 and 48 degrees latitude, New Zealand is geographically interesting for the study of both local and global seasonal dynamics of human influenza. It has been assumed by prevailing models that the local dynamics will reflect the transitory expression of global human influenza dynamics at the peripheries of its distribution. Due to its presumed role as an endpoint (sink) rather than a source of new influenza strains destined for global circulation, data from locations such as Christchurch, NZ, have proved key in the understanding of annual rhythms in seasonal influenza (Rambaut *et al*, 2008, Bedford *et al*, 2010). Furthermore, New Zealand is isolated. This allows for the study of both local persistence as well as high-latitude contribution to global circulation by way of human air traffic, as air traffic remains the primary mode of transportation for the inflow and outflow of people.

II. Project

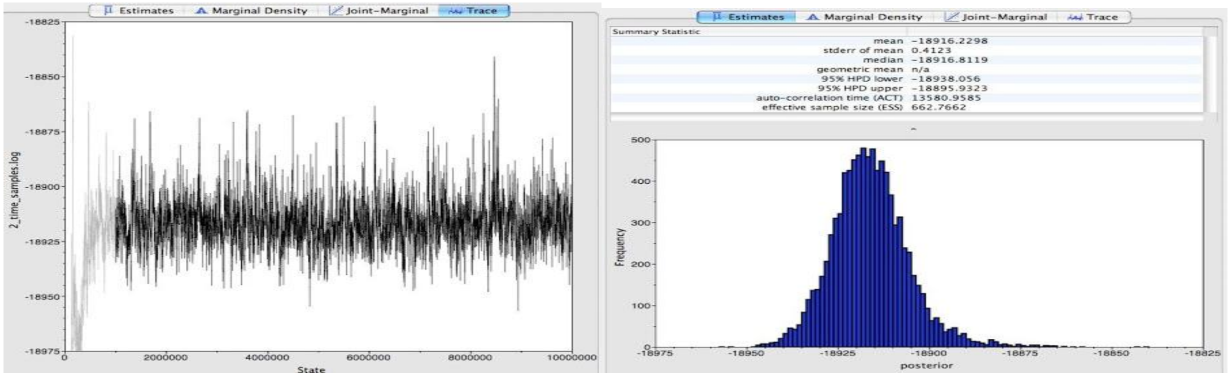
The goal of this AWC summer project was to set the stage for the development of a geographically-aware forward stochastic model of influenza viruses in New Zealand. Such a model would allow researchers to compare results over multiple years using sequence data from public sources and local diagnostic sequencing labs.

Multi-type tree simulation study:

Simulations were run in BEAST 2, (Drummond *et al*., 2012), using a new add-on, an MCMC sampler for multi-type trees developed by T. Vaughan, a postdoc in the Computational Evolution Group. Multi-type trees are phylogenetic trees whose tips are labeled with molecular sequence data as well as sampling times and sampled types. These "sampled types" may be used to represent locations where the sequence data were sampled, which is particularly suitable for studies in phylodynamics. The sampled types are visualized by colour; a change of colour along a branch of the tree indicates a *migration event* has occurred, and the consequential formation of these single-child nodes can be quantified and analyzed using the programs TreeStat and Tracer. Both of these can be found within the BEAST package, (Rambaut and Drummond, 2007).

This new sampler developed by Vaughan also possesses the ability to handle serially-sampled data, a feature that is likewise key for analyzing influenza viral lineages sampled across several years.

To validate the sampler a test was conducted simulating a phylogenetic tree with 128 taxa, 2 demes, and 4 sampling times under the structured coalescent, with hardcoded rates of coalescence and migration. A 2kb alignment was simulated down the tree with given □, and the sampler inferred the alignment parameters under the structured-coalescent prior, (Vaughan, pers. comm.). The BEAST log files were analyzed in Tracer post-simulation, and the operators appeared to mix efficiently, (Figures 1-a,b).



```
Summary Statistic
                         mean  -18916.2298
               stderr of mean  0.4123
                       median  -18916.8119
               geometric mean  n/a
                95% HPD lower  -18938.056
                95% HPD upper  -18895.9323
  auto-correlation time (ACT)  13580.9585
   effective sample size (ESS)  662.7662
```

Figures 1-a,b. Visualization of a BEAST log file of simulated molecular data (100 taxa) sampled at 2 epochs in Tracer v1.7.4, (Rambaut). The mixing of our new multi-type sampler greatly improved over the course of testing.

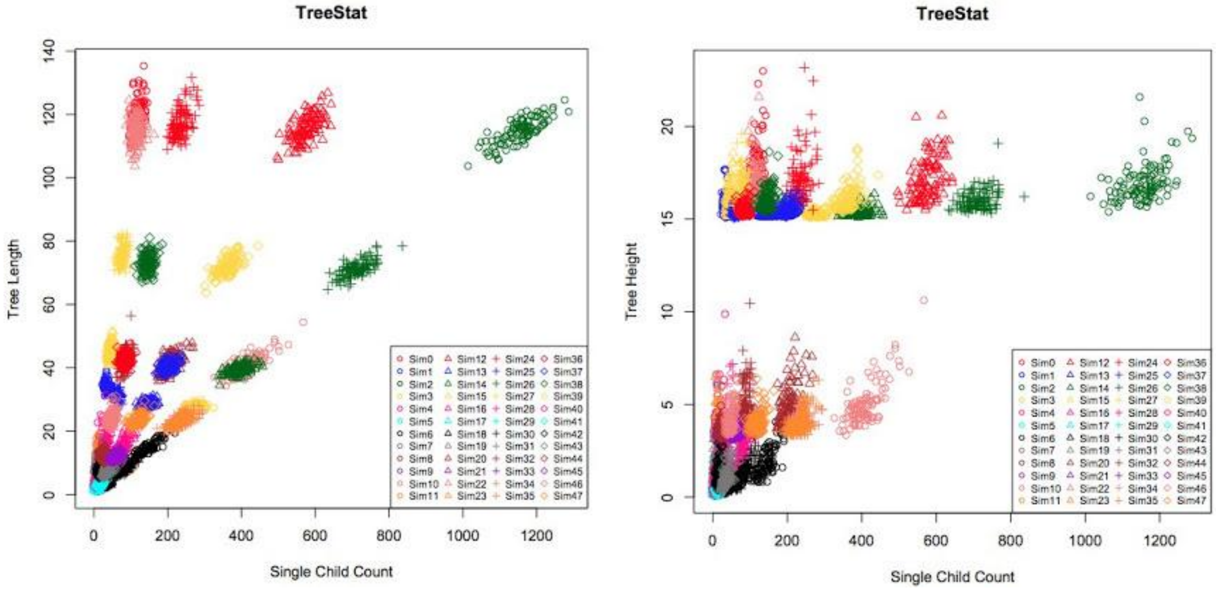
Simulation study of multifarious schemes:

The phylogenetic trees produced in BEAST are reliant upon the manipulation of parameters specified in the input file, including: number of years sampled, mutation rates, number of demes, etc. An amendment to the mutation rate, for example, can affect the entire topology of the tree. Thus, another series of simulations were run in BEAST 2 using Vaughan's multi-type sampler, this time with the aim of testing a variety of sampling schemes by running simulations using varied combinations in a small range of migration and coalescence rates (values estimated *a priori*) across three distinctive numbers of years sampled, (Table 1). Again, molecular sequence data were simulated down the tree and inferences made under the structured coalescent.

The results of these simulation tests provide us with some insights concerning realistic priors in future Bayesian analyses of actual data sets, and they illuminate the effects of these altering these rates in relation to each other and on tree size and topology, (Figure 2-a,b).

No. of Years Sampled	Rates of Coalescence	Rates of Migration
1	1	1
4	2	2
16	5	5
	10	10

Table 1. A total of 48 simulations were performed, using all possible combinations of rates and numbers of years sampled as shown in the table above.



Figures 2-a,b. Summary statistics showing *tree length* (minimum number of changes) and *tree height* (length of the longest path from the root to one of its descendants, respectively, vs. the number of single-child nodes in the tree (representing migration events).

Tree simulations using Oceania data:

Haemagglutinin (HA) sequence data for influenza type-A subtype H3N2 were obtained through the public Influenza Virus Database of NCBI, (National Center for Biotechnology Information). The data were filtered to include only well-sampled years (1999-2009) and locations. The locations used in this part of the study are listed below:

New Zealand

- North (mostly Wellington) : 186 seq : label 0
- South (mostly Canterbury) : 271 seq : label 1

Australia

- "Australia" : 33 seq : label 2
- Western Australia : 75 seq : label 3
- New South Wales/Victoria: 28 seq : label 4
- Queensland : 62 seq : label 5
- South Australia: 34 seq : label 6

The nucleotide sequences obtained from these locations were entered in an xml file specifying parameter priors relevant to influenza virus data, as prescribed in work done earlier by Denise Kühnert, a PhD student in the Computational Evolution Group. i.e. The substitution model implemented was the General Time Reversible (GTR) model. As its name implies, GTR is the most general neutral, time-reversible model with finite sites, (Tavaré, 1986). Although statistical analysis could be performed in the future to validate this choice of model, its pre-established affinity in influenza molecular data is satisfactory for the time being.

A short Python script converted the NCBI database date format [`YYYY/MM/DD`] into decimals readable by BEAST 2 for *traitname= "date-forward"*. These real data samples demonstrate the aforementioned utility of an MCMC sampler capable of handling serially sampled data. Additionally, the sampled types (localities) were ariculated by the "colourTraitSet" spec, (Appendix III-b).

The real data simulations were submitted to the New Zealand eScience Infrastructure (NeSI) 'Pan' cluster. With 667 sequences of 1700 bp length, and migration events to account for as well as coalescent events, the jobs required a couple of days to finish. Once completed, the output tree files were viewed in FigTree and DensiTree, (Rambaut, 2012, Bouckaert, 2010), and log files in Tracer.

Discussion:

There is still work to be done. As seen in Figure 3, the tree generated by this initial Oceania xml is rather nonsensical. This is to be expected, considering only the preliminary steps have been taken toward our goal of developing of a geographically-aware forward stochastic model. For example, human traffic data between locations must be incorporated into the model; that is to say, not all migration rates (e.g. an individual's movement from city to city) between each location is going to be equal in the real world.

However, throughout the AWC summer project significant progress has been made toward the stochastic model objective. We have tested the new MCMC sampler for multi-type trees developed by Vaughan using simulated data and will soon employ real data like that discussed above for further validation.

We have also tested several rates of coalescence and migration under the structured coalescent. The coalescent moves “backward” through time and lineages coalesce whenever two or more individuals share the same parent. The “structured” coalescent refers to population sizes that can be spatially subdivided, allowing us to indicate deme locations. Specifically, we employed the Wright-Fisher model, where individuals ‘choose’ their parents independently of another. Of course, these are not the only models that have been proposed for use on influenza data. There is also the Kühnert birth-death model, which marries the Bayesian phylogenetic approach with epidemic modelling of infectious disease, (Kühnert *et al.*, 2011), as well as a variation on the structured coalescent theme described by Volz *et al*., 2012, where the function 1/N becomes birth rates over the square of the population size. The Volz model is an attempt to reconcile these approaches and may prove salutary to forthcoming research.

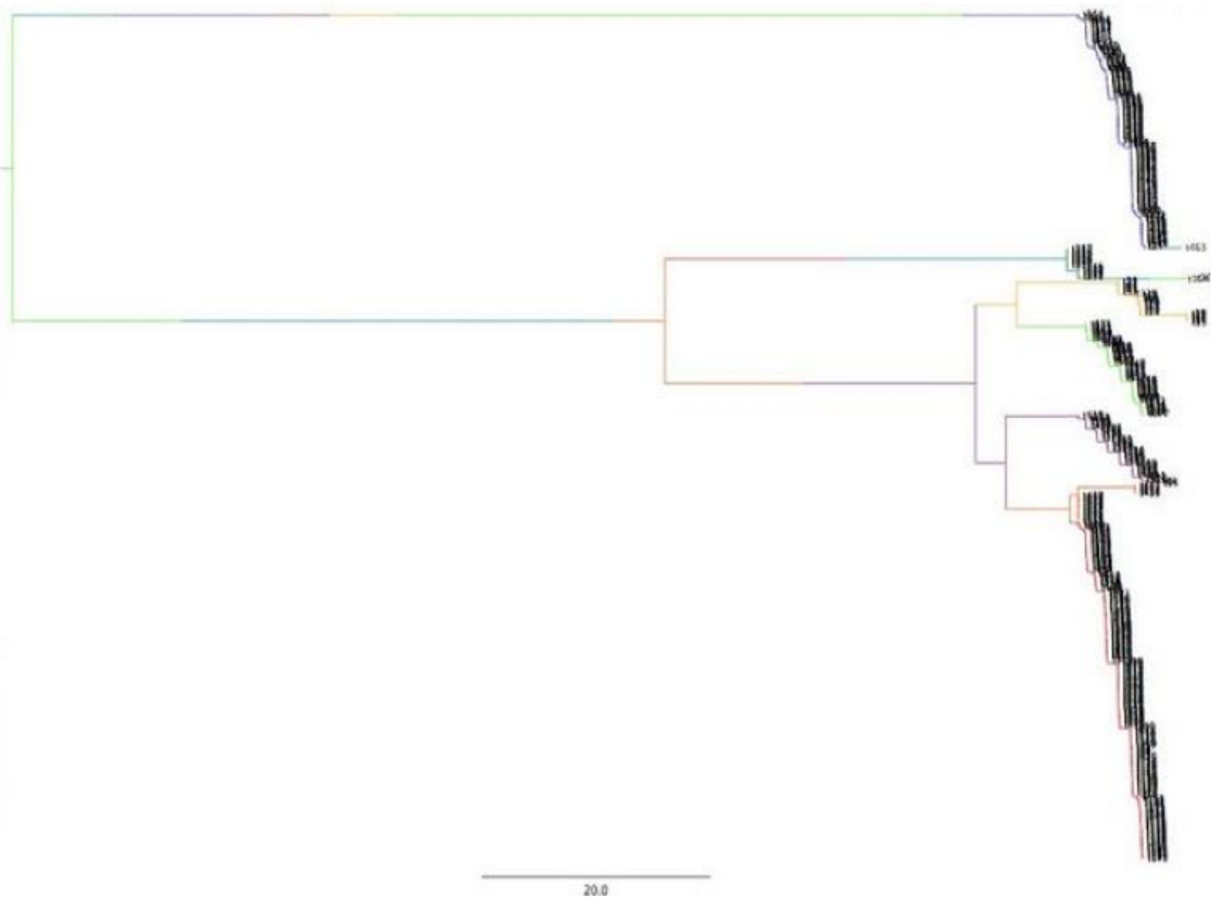


Figure 3. Preliminary Oceania influenza A subtype H3N2 tree for years 1999-2009, Australia and NZ.

References:

Bedford, T., S. Cobey, P. Beerli, and M.Pascual. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLoS Pathog.* 2010 May; 6(5): e1000918. Published online 2010 May 27. [DOI:10.1371/journal.ppat.1000918]

Bouckaert, R. R. 2010. "DensiTree: Making Sense of Sets of Phylogenetic Trees." *Bioinformatics* 26 (10) (March 12): 1372–1373. [DOI:10.1093/bioinformatics/btq110].

Drummond AJ, Suchard MA, Xie D, and Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 2012 August; 29(8): 1969–1973. Published online 2012 February 25. [DOI:10.1093/molbev/mss075]

Kühnert D, Wu CH, Drummond AJ. Phylogenetic and epidemic modeling of rapidly evolving infectious disease. *Infect Genet Evol*. 2011 Dec;11(8):1825-41. Epub 2011 Aug 31. [DOI: 10.1016/j.meegid.2011.08.005].

Rambaut, A., and A.J. Drummond. "Tracer v1. 4." (2007).

Rambaut, A., and A.J. Drummond. "TreeStat" version 1.1." (2007).

Rambaut, A., O.G. Pybus, M.I. Nelson, C. Viboud, J.K. Taubenberger, and E.C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008 May 29; 453(7195): 615–619. Published online 2008 April 16. [DOI:10.1038/nature06945]

Russell, T.C. Jones, I.G. Barr, N.J. Cox, R.J. Garten, V. Gregory, I.D. Gust, A.W. Hampson, A.J. Hay, A.C. Hurt, Jan C. de Jong, A. Kelso, A.I. Klimov, T. Kageyama, N. Komadina, A.S. Lapedes, Y.P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A.D.M.E. Osterhaus, G.F. Rimmelzwaan, M.W. Shaw, E. Skepner, K. Stohr, M. Tashiro, M., R.A.M. Fouchier, and D.J. Smith. The global circulation of seasonal influenza A (H3N2) viruses. *Science* 18 April 2008: **320** (5874), 340-346. [DOI:10.1126/science.1154137]

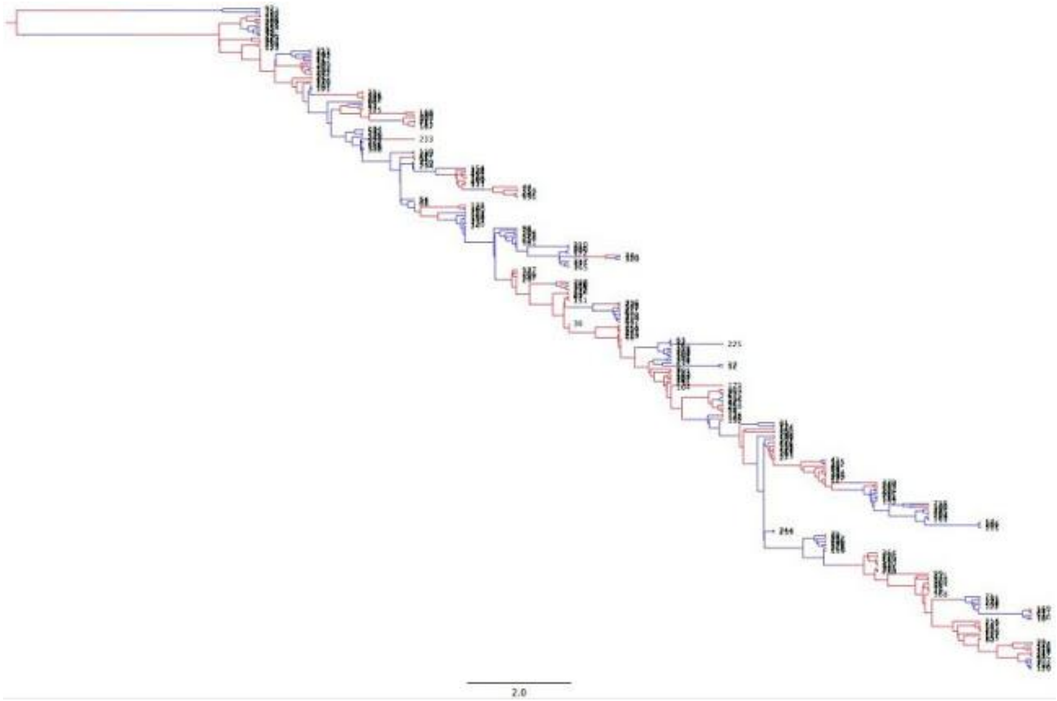
Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences* (American Mathematical Society) 1986. **17**: 57–86.

Vijaykrishna, D., G.J.D. Smith, O.G. Pybus, H. Zhu, S. Bhatt, L.L.M. Poon, S.Riley, J. Bahl, S.K. Ma, C.L. Cheung, R.A.P.M. Perera, H. Chen, K.F. Shortridge, R.J. Webby, R.G. Webster, Y. Guan, and J. S. Malik Peiris. Long-term evolution and transmission dynamics of swine influenza A virus. *Nature* 473, 519–522. 26 May 2011. Published online 25 May 2011. [DOI:10.1038/nature10004]

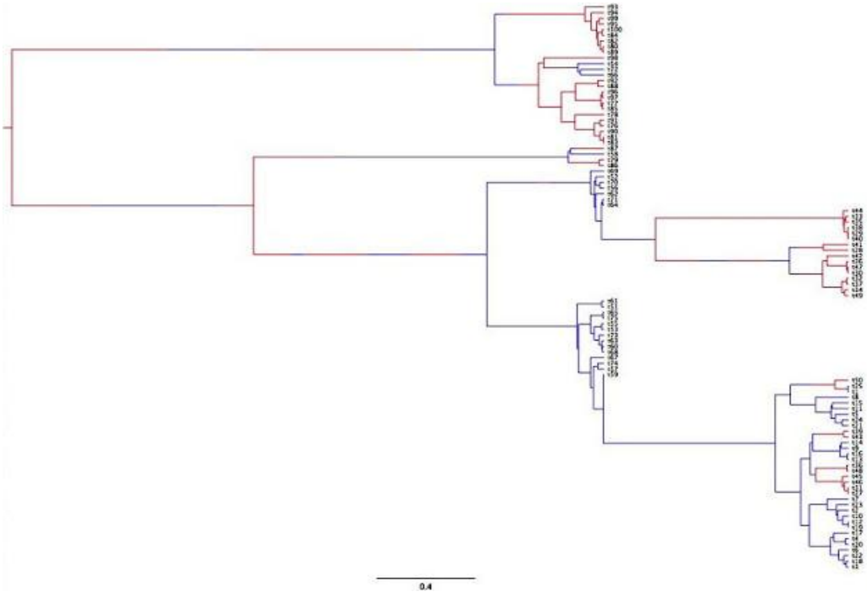
Volz, E.M. Complex Population Dynamics and the Coalescent Under Neutrality. *Genetics.* January 1, 2012 vol. 190 no. 1 187-201. Published online before print, October 31, 2011. [DOI:10.1534/genetics.111.134627].

Appendices:

Appendix I. Multi-type tree with 16 sampling times and 2 sampled types (indicated by red and blue on the branches of the tree. This tree shape (topology + branch lengths) is consistent with influenza phylogeny.



Appendix II. Multi-type tree with 2 sampling times, 2 sampled types, and 100 taxa of simulated sequences. A relatively high rate of migration will result in multiple colour changes (single child nodes) per branch.



Appendix III.

(a)

```
 <!-- Parameter priors -->
   <input spec='CompoundDistribution' id='parameterPriors'>
     <distribution spec='ExcludablePrior' x="@rateMatrix"
     xInclude="0 1 1 1 1 1 1
         1 0 1 1 1 1 1
         1 1 0 1 1 1 1
         1 1 1 0 1 1 1
         1 1 1 1 0 1 1
         1 1 1 1 1 0 1
         1 1 1 1 1 1 0">
<distr spec='LogNormalDistributionModel' M="-0.5" S="2.0"/>
     </distribution>
     <distribution spec='beast.math.distributions.Prior' x="@popSizes">
              <distr spec="LogNormalDistributionModel"  M="-2.0" S="2.0"/>
       </distribution>
```

(b)

```
   <colourTraitSet spec='TraitSet' id='colourTraitSet'
                traitname="discrete"
                value=
                            "t1=1,
                             t2=0,
                             t3=0,
                             … t667=1”>
     <taxa spec='TaxonSet' alignment='@H3N2_Oceania'/>
   </traitSet>
```