Supplement to Population divergence time estimation using individual lineage label switching

Peter Beerli[*,1,+], Haleh Ashki[2], Somayeh Mashayekhi[3] and Michal Palczewski[4,+]

[1]Department of Scientific Computing, Florida State University, Tallahassee, Florida, 32306, USA
[2]Current Address: Foundation Medicine Inc, San Diego, California, 92121, USA
[3]Department of Mathematics, Kennesaw State University, Marietta, Georgia, 30060, USA
[4]Maplebear Inc. San Francisco, California, 94105, USA

[*]Corresponding author: beerli@fsu.edu

Abstract

This supplement contains:
– Simulation conditions.
– Probabilities of divergence events after migration or coalescence event happened.
– Alternative example of the *Migrate* software to evaluate different models for the distribution of the Zika virus.

Keywords: Coalescence, Gene tree, Species tree, Bayesian inference, Divergence time

Simulation conditions

The trees were simulated using MS by Hudson (2002) and our own sequence simulator POPSIM. We ran test runs using MS and SEQGEN (Rambaut and Grassly 1997) and that delivered equivalent results. 4 replicates were run for each true divergence parameter of $N_e \times 8.0, 4.0, 2.0, 1.0, 0.5, 0.25, 0.125, ..., 0.001953125$ for each set. The sum of all population sizes for all simulations was set to $\Theta = 0.02$. All dataset were run using the parallel version of MIGRATE that can run n loci in the same time if there are n compute nodes available. With fewer compute nodes MIGRATE uses a load-balancing scheme. All runs were run on the high performance computing cluster at Florida State University and used maximally $n = 501$ nodes, this setup allowed to finished each run in less than 4 hours which is the maximum allowed run-time on the particular queue used. We tuned the run time so that most of the loci would deliver acceptable results: in particular we used the options in MIGRATE: long-inc=1000, long-sample=10000, and replicate=YES:2. The number of cores and the approximate runtime for an individual run is also given. The comment shows deviations from the scheme.

Figure	Loci	Populations	Fixed parameters	Mutation model	Cores	Runtime	Comment
5A	2	2	$4Nm = 0$	F84	4+1	1h	
	10	2	4Nm=0	F84	20+1	1h	
	1000	2	4Nm=0	F84	500+1	4h	
5B	10	2	4Nm=0.25	F84	20+1	1h	
	10	3	4Nm=0.25	F84	20+1	15 min	$long - inc = 100$
	10	2	4Nm=1.00	F84	20+1	1h	
	10	3	4Nm=1.00	F84	20+1	15 min	$long - inc = 100$
5C	2	10	4Nm=0	F84	20+1	1 min	$long - inc = 10^*$

* using long-inc=1000 delivered the same results

The outputs and parameter files for all runs in Figure 5 are available as compressed archives on github.com/pbeerli/divergencesupplement.

[+]These authors contributed equally to this work.

Calculations of the probabilities of events

The probabilities for the events depend on rates λ_i that may be constant through time or not. We use for coalescence and migration events $lambda_1$ and λ_2, respectively, and for the divergence time λ_3.

Constant λ_1 and λ_2

In the standard structured coalescence with two forces, genetic drift and recurrent gene flow between isolated populations we have two different types of rates, one type for coalescences and another type for immigration events. Both types can be considered exponential rates; there is a constant risk over time that one or the other event happens, so we can calculate

$$f_1(t)=\lambda_1 e^{-\lambda_1 t} \qquad f_2(t)=\lambda_2 e^{-\lambda_2 t} \tag{1}$$

We consider first the case where the event with λ_1 happens first $(T_1 < T_2)$. We find $P(T_1 < T_2)$ assuming that

$$P(T_2 > t) = \int_t^\infty -(-\lambda_2)e^{-\lambda_2 u}du = -e^{-\lambda_2 u}\Big|_t^\infty = -e^{-\lambda_2 \infty} + e^{-\lambda_2 t} = e^{-\lambda_2 t} \tag{2}$$

$$P(T1 < T2) = \int_0^\infty f_1(u)P(T_2 > u)du = \int_0^\infty \lambda_1 e^{-\lambda_1 u}e^{-\lambda_2 u}du \tag{3}$$

$$= \lambda_1 \int_0^\infty e^{-\lambda_1 u}e^{-\lambda_2 u}du = \lambda_1 \int_0^\infty e^{(-\lambda_1-\lambda_2)u}du \tag{4}$$

$$= \lambda_1 \int_0^\infty e^{-\lambda_1 u}e^{-\lambda_2 u}du = \lambda_1 \int_0^\infty e^{(-\lambda_1-\lambda_2)u}du \tag{5}$$

$$= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)u}du = \lambda_1 \int_0^\infty \frac{-(\lambda_1+\lambda_2)}{-(\lambda_1+\lambda_2)}e^{-(\lambda_1+\lambda_2)u}du \tag{6}$$

$$= \frac{\lambda_1}{-(\lambda_1+\lambda_2)} \int_0^\infty -(\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)u}du \tag{7}$$

$$= \frac{\lambda_1}{-(\lambda_1+\lambda_2)} \left[e^{-(\lambda_1+\lambda_2)u}\Big|_0^\infty\right] = \frac{\lambda_1}{-(\lambda_1+\lambda_2)} \left[e^{-(\lambda_1+\lambda_2)\infty} - e^{-(\lambda_1+\lambda_2)0}\right] \tag{8}$$

$$= \frac{\lambda_1}{-(\lambda_1+\lambda_2)}(0-1) = \frac{\lambda_1}{\lambda_1+\lambda_2} \tag{9}$$

Non-constant and constant rates

In this section, we consider three different cases. Suppose we have three events

- The divergence event $\rightarrow f_{T_1}(t) = \lambda_1'(t)e^{-\lambda_1(t)}$, Time related to it call T_1
- The coalescent event $\rightarrow f_{T_2}(t) = \lambda_2 t e^{-\lambda_2 t}$, Time related to it call T_2
- The Migration event $\rightarrow f_{T_3}(t) = \lambda_3 t e^{-\lambda_3 t}$, Time related to it call T_3

Divergence has a rate that changes with the time, the risk of switching increases the longer we wait and is non-constant, This leads to complication in finding a solution to the integral.

Divergence happen first

We need to find $P(T_1 < T_2 \& T_1 < T_3)$. We know

$$P(T_1 < T_2 \& T_1 < T_3) = \int_0^\infty P(T_2 > t)P(T_3 > t)f_{T_1}(t)dt. \tag{10}$$

Fist we find $P(T_2 > t)$ and $P(T_3 > t)$ as

$$P(T_2 > t) = \int_t^\infty \lambda_2 e^{-\lambda_2 u}du = -(e^{-\lambda_2 u})|_t^\infty = e^{-\lambda_2 t}. \tag{11}$$

$$P(T_3 > t) = \int_t^\infty \lambda_3 e^{-\lambda_3 u}du = -(e^{-\lambda_3 u})|_t^\infty = e^{-\lambda_3 t}. \tag{12}$$

Using Eqs. (10)-(12) we have

$$P(T_1 < T_2 \& T_1 < T_3) = \int_0^\infty e^{-\lambda_2 t} * e^{-\lambda_3 t} * \lambda_1'(t)e^{-\lambda_1(t)}dt. \tag{13}$$

Coalescent happen first

We need to find $P(T_2 < T_1 \& T_2 < T_3)$. We know

$$P(T_2 < T_1 \& T_2 < T_3) = \int_0^\infty P(T_1 > t) P(T_3 > t) f_{T_2}(t) dt. \tag{14}$$

Fist we find $P(T_1 > t)$ and $P(T_3 > t)$ as

$$P(T_1 > t) = \int_t^\infty \lambda_1'(u) e^{-\lambda_1(u)} du = -(e^{-\lambda_1(u)})|_t^\infty = e^{-\lambda_1(t)}. \tag{15}$$

$$P(T_3 > t) = \int_t^\infty \lambda_3 e^{-\lambda_3 u} du = -(e^{-\lambda_3 u})|_t^\infty = e^{-\lambda_3 t}. \tag{16}$$

Using Eqs. (14)-(16) we have

$$P(T_2 < T_1 \& T_2 < T_3) = \int_0^\infty e^{-\lambda_1(t)} * e^{-\lambda_3 t} * \lambda_2 e^{-\lambda_2 t} dt. \tag{17}$$

Migration happen first

We need to find $P(T_3 < T_1 \& T_3 < T_2)$. We know

$$P(T_3 < T_1 \& T_3 < T_2) = \int_0^\infty P(T_1 > t) P(T_2 > t) f_{T_3}(t) dt. \tag{18}$$

Fist we find $P(T_1 > t)$ and $P(T_2 > t)$ as

$$P(T_1 > t) = \int_t^\infty \lambda_1'(u) e^{-\lambda_1(u)} du = -(e^{-\lambda_1(u)})|_t^\infty = e^{-\lambda_1(t)}. \tag{19}$$

$$P(T_2 > t) = \int_t^\infty \lambda_2 e^{-\lambda_2 u} du = -(e^{-\lambda_2 u})|_t^\infty = e^{-\lambda_2 t}. \tag{20}$$

Using Eqs. (18)-(20) we have

$$P(T_3 < T_1 \& T_3 < T_2) = \int_0^\infty e^{-\lambda_1(t)} * e^{-\lambda_2 t} * \lambda_3 e^{-\lambda_3 t} dt. \tag{21}$$

Distribution of the time to the first event

If we have three events, Divergence T_1, Coalescent T_2 and Migration T_3, the distribution of time to the first event $T = min(T_1, T_2, T_3)$ is as

$$P(T > t) = P\{min(T_1, T_2, T_3) > t\} = P\{T_1 > t\} P\{T_2 > t\} P\{T_3 > t\} = \\ e^{-\lambda_1(t)} \times e^{-\lambda_2 t} \times e^{-\lambda_3 t} = e^{-(\lambda_1(t) + \lambda_2 t + \lambda_3 t)}, \tag{22}$$

so $T = min(T_1, T_2, T_3)$ has an exponential distribution.

Alternate example to showcase the software

In the article we use two examples using the relationship among humans that are reconstituting the work of others, Originally we used this virus data set as an example.

Example using samples of complete genomes of the Zika virus

The Zika virus (ZIKV) is a pathogen distributed by mosquitoes. It was originally described in Africa. Subsequently, ZIKV was then brought to all continents via infected hosts. Gatherer and Kohl (2016) discuss the distribution routes of ZIKV based on the dates of incidences. The expansion followed this pattern: Origin in Africa, outbreaks in Asia, and then outbreaks in the Americas.

Complete ZIKV genomes from locations in Nigeria (n=5), China (n=12), Brazil (n=13), Mexico (n=2), Guatemala (n=2), Panama (n=4) and Puerto Rico (n=2) were obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) (Pickett *et al.* 2012) through the web site at `http://www.viprbrc.org/`. The sequences were aligned with MUSCLE (Edgar 2004, aligned dataset in electronic supplement). We pooled the locations Mexico, Guatemala, Panama, and Puerto Rico for the analysis.

We then explored four different main population models (Fig. 1). Model group *a* specified the expansion from Africa to Asia to Brazil to Central America. Model group *b* specifies the expansion from Africa to Asia and, independently, from Africa to Brazil to Central America. Model group *c* is a hybrid of models *a* and *b* where one population is the admixture of two populations. We considered the Brazilian lineages a potentially admixed population from African and Asian lineages. Occurrences of ZIKV in Central Americas and Brazil were reported contemporaneously, suggesting that there may be not enough mutations to separate Brazilian and Central American lineages; we combined the samples from Brazil and Central America (model group *d*), otherwise *d* is equivalent to *c*. The variants a', b', c', and d' include recurrent immigration from the source populations.

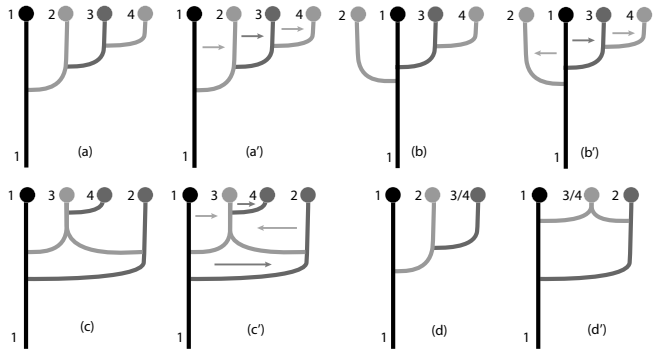

Figure 1 Eight population models used to analyze the ZIKV dataset. Tip labels are 1=Africa, 2=Asia, 3=Brazil, 4=Central America, and 3/4 = Brazil and Central America combined.The arrows mark migration directions.

Zika virus dispersal

Table 1 shows the model probabilities and the log marginal likelihoods for different models (see Fig. 1) for the Zika virus (ZIKV). The model in which the expansion followed a route to the east from Africa is the most likely model. The best model is a simple colonization model without migration. Figure 2 shows the population tree of the best model with mutation-scaled population sizes and divergence times. Yokoyama and Starmer (2017) used an estimate of the mutation rate for various lineages of the Asian ZIKV lineages, suggesting that the mutation rate per year has accelerated and is 0.004/year, but can be as low as 0.0005/year. Faria *et al.* (2016) estimate a phylogenetically derived mutation rate of 0.00098 to 0.00106 per year. *Migrate* estimated the mutation-scaled divergence time $\hat{\tau}_{1\to 2} = 0.05$ and $\hat{\tau}_{2\to 3/4} = 0.0025$ assuming that the mutation rate is per generation. We did not find any clear characterization of generation time for ZIKV in the literature. We equate generation time here as successful transmissions among hosts per year and not the number of replications of an individual ZIKV within a host. Early records from Africa date to 1947 and early records from Asia date to 1951. Thus, gene flow of ZIKV from Africa to Asia was most likely around 1950. The ZIKV outbreak in Brazil started in 2015 (Faria *et al.* 2016). Ignoring the precise sampling dates and assuming the divergences were 67 and 3 years before 2018 when the dataset was assembled then we calculate about 5 generations per year ($67/(0.05/0.004) = 5.2$ and $3/(0.0025/0.004{=}4.8)$ using the high mutation rate. The lower mutation rate (~ 0.001) would lead to 1.34 and 1.2 generations per year, respectively.

Table 1 Log marginal likelihoods and model probabilities of biogeographic models: (a) eastward, (b) westward, (c) admixture, and (d) three-population models (Fig. 1) and Zika viruses. The $\rightarrow$ mark colonizations, the $\Rightarrow$ mark colonizations with recurrent immigration. Numbers are population labels: 1=Africa, 2= Asia, 3=Brazil, 4=Central America.

	Model	Zika			
		ln(mL)	LBF*	Prob.	Rank
a	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4$	-25762.13	-95.61	0.0	3
a′	$1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4$	-26078.46	-411.94	0.0	6
b	$2 \leftarrow 1 \rightarrow 3 \rightarrow 4\,1$	-25824.41	-157.89	0.0	5
b′	$2 \Leftarrow 1 \Rightarrow 3 \Rightarrow 4$	-26121.92	-455.40	0.0	8
c	$1 \rightarrow 3 \leftarrow 2 \leftarrow 1,3 \rightarrow 4$	-25786.83	-120.31	0.0	4
c′	$1 \Rightarrow 3 \Leftarrow 2 \Leftarrow 1,3 \Rightarrow 4$	-26099.01	-432.49	0.0	7
d	$1 \rightarrow 2 \rightarrow 3/4$	-25666.52	0.0	1.0	1
d′	$1 \rightarrow 3/4 \leftarrow 2$	-25703.50	-36.98	0.0	2

* LBF = ln Bayes factor against the best model *d*

Discussion

The direction of the expansion of ZIKV as estimated by *Migrate* is simply based on genetic data and coalescence-based population genetic models. The used data are not very informative because only a few sequences from Africa are present. This does not allow to pinpoint the expansion from Africa to Asia with good precision. It is also likely that the expansion from Brazil to Mexico and other countries is not very informative because only few sequences from a large area (Mexico, Guatemala, Panama, and Puerto Rico) were used. Gatherer and Kohl (2016) summarized the literature on the spread of the Zika virus and shows an expansion from Africa to Asia to Pacific islands and then to South America. They used incidences of confirmed Zika virus infections and already published phylogenetic trees to report a map of its

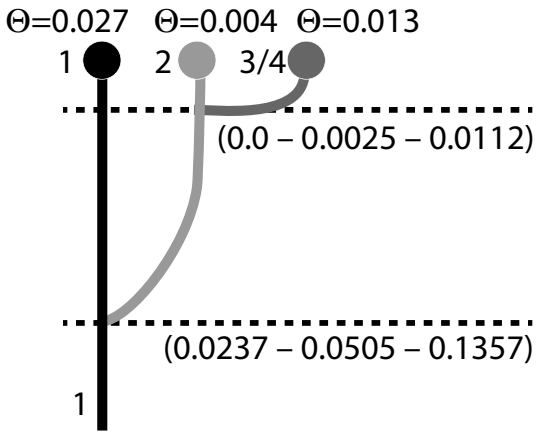

Figure 2 Mode and 50% credibility intervals of the splitting times and population sizes of the best model for Zika virus (model *d*; Fig. 1)

spread. It is comforting that our population genetics approach recovers the same paths as the more detailed historical records of infections; genetic data will be particularly useful for pathogens for which we may not have detailed incidence records.

Literature cited

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32:1792–1797.

Faria NR, do Socorro da Silva Azevedo R, Kraemer MUG, Souza R, Cunha MS, Hill SC, Thézé J, Bonsall MB, Bowden TA, Rissanen I *et al*. 2016. Zika virus in the Americas: Early epidemiological and genetic findings. Science. 352:345–349.

Gatherer D, Kohl A. 2016. Zika virus: a previously slow pandemic spreads rapidly through the Americas. Journal of General Virology. 97:269–273.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z *et al*. 2012. Vipr: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research. 40:D593–D598.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computional Applications in Bioscience. 13:235–238.

Yokoyama S, Starmer WT. 2017. Possible roles of new mutations shared by Asian and American Zika viruses. Molecular Biology and Evolution. 34:525.