Phylodynamics of Infectious Diseases (Part II)

Mandev Gill

Department of Statistics and Institute of Bioinformatics

University of Georgia

Example: 2020 COVID-19 Resurgence in Europe

Paris, April 24, 2020

In spring 2020, Europe experienced a wave of SARS-CoV-2 infections

- Governments imposed lockdowns and social restrictions to try to contain the spread
- By mid-April, the European Commission had constructed a roadmap to lift containment measures in a coordinated and cautious way to revive social and economic activities

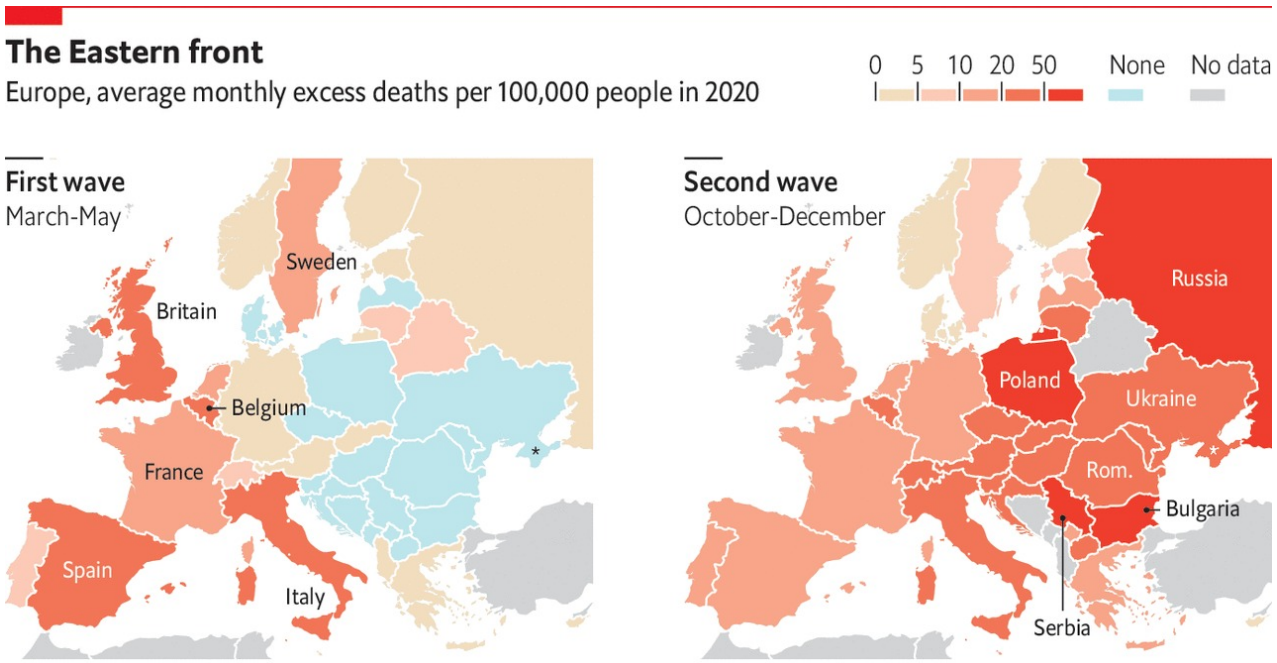
Example: 2020 COVID-19 Resurgence in Europe

Brussels, June 20, 2020

By the summer of 2020, many of the containment measures had been relaxed

On June 15, most EU and Schengen-area countries opened their borders to other countries

Example: 2020 COVID-19 Resurgence in Europe



COVID-19 Second Wave in Europe, 2020

Many countries in Europe experienced a resurgence in cases in the late summer

There was a rapid increase in the number of cases, and it became clear by the end of October that there was a second epidemic wave, causing governments to reimpose containment measures

- Cross-border travel during the summer was largely blamed for the resurgence

Example: 2020 COVID-19 Resurgence in Europe

Lemey et al. (2021) sought to clarify the role of cross-border summer travel in the fall 2020 COVID-19 resurgence in Europe through a phylodynamic analysis

- Even without resumption of cross-border travel, relaxing containment measures during ongoing transmission can allow for the proliferation of locally circulating strains
- What was the relative importance of newly introduced lineages vs. rekindling of persistent lineages?

Example: 2020 COVID-19 Resurgence in Europe

Lemey et al. (2021) considered all SARS-CoV-2 genomes from 10 European countries available on November 3, 2020 on GISAID (total number: 39,812)

- Only retained unique sequences from each country
- Subsampled each country proportionally to the cumulative number of cases by setting arbitrary threshold of 6.5 sequences per 10,000 cases, with minimum of 100 sequences per country
- To maximize spatial and temporal coverage, genomes were binned by epi-week and sampled as evenly as possible
- Only sequences from B.1 lineage were considered

This yielded a data set of 2,909 genomes

Example: 2020 COVID-19 Resurgence in Europe

This preliminary data set had relatively fewer recently sampled genomes, so on January 4, 2021, Lemey et al. (2021) updated the data set with genomes collected between August 1 (June 22 for Portugal, corresponding to most recent sampling date in preliminary data set) and October 31

- The number of genomes to add per country was obtained by raising the threshold to 8.5 sequences per 10,000 cases and increasing the minimum number of sequences to 200
- To bias temporal coverage towards more recent samples, genomes from each country were binned by week and sampled proportionally to an exponential function
- Only unique sequences were retained

This amounted to an additional 1,050 genomes, bringing the total number to 3,959

Example: 2020 COVID-19 Resurgence in Europe

OXFORD JOURNALS Molecular Biology and Evolution

Mol Biol Evol. 2021 May; 38(5): 1777–1791.
Published online 2020 Dec 15. doi: 10.1093/molbev/msaa314

PMCID: PMC7798910
PMID: 33316067

Journal Article

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

Benoit Morel,[1] Pierre Barbera,[1] Lucas Czech,[2] Ben Bettisworth,[1] Lukas Hübner,[1,3] Sarah Lutteropp,[1] Dora Serdari,[1] Evangelia-Georgia Kostaki,[4] Ioannis Mamais,[5] Alexey M Kozlov,[1] Pavlos Pavlidis,[6] Dimitrios Paraskevis,[4] and Alexandros Stamatakis[1,3]

Unfortunately, phylogenetic analysis of SARS-CoV-2 data is difficult!

Because of relatively low degree of resolution offered by the sequence data, Lemey et al. (2021) relied on

- Relatively simple model specifications
- Informing parameters by additional non-genetic sources of information

Evolutionary Model: HKY nucleotide substitution model with gamma-distributed across-site rate variation and strict molecular clock

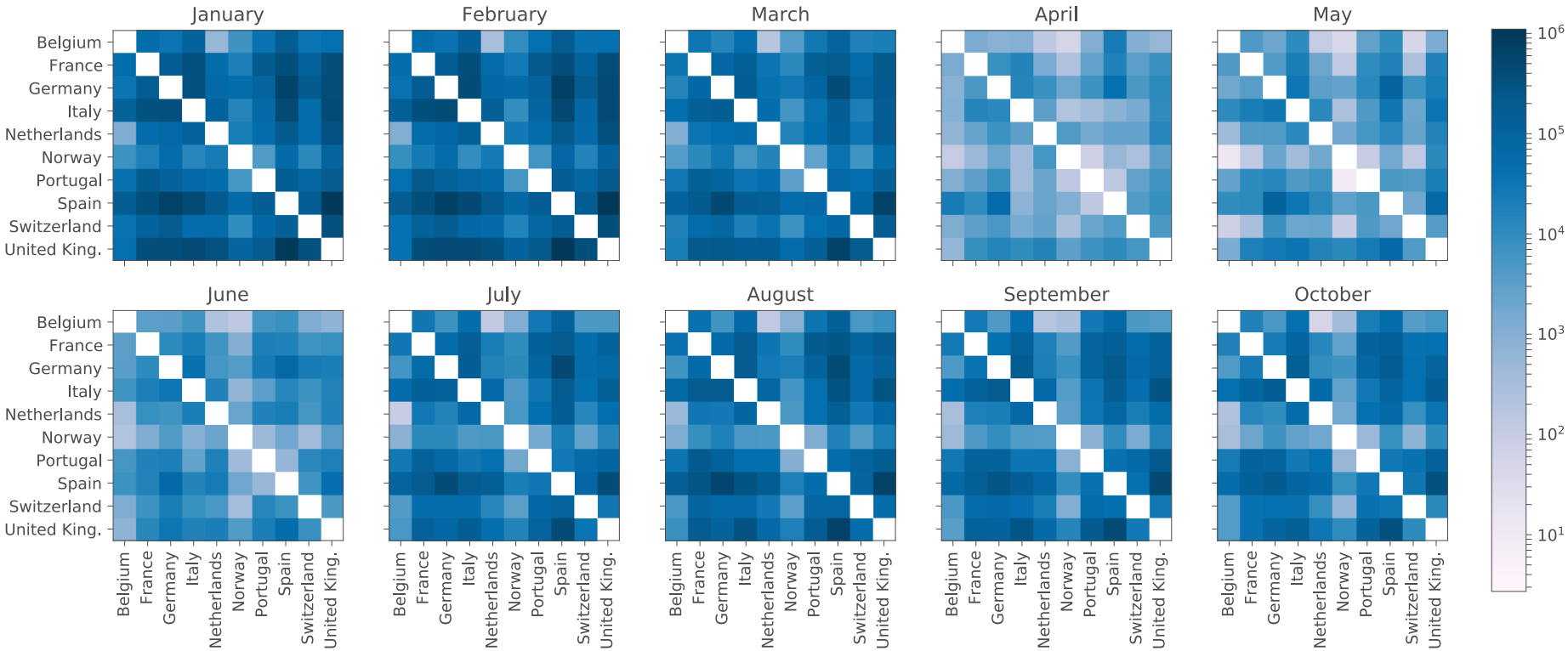
Example: 2020 COVID-19 Resurgence in Europe

For the phylogenetic tree prior, the study used a simplified coalescent-based model with a covariate:

- Log-transformed effective population size parameterized as a piecewise constant function that changes values at pre-specified time points spaced two weeks apart (Skygrid parameterization)
- Log effective population size is deterministic function of a covariate: the total log-transformed COVID-19 case counts for the 10 European countries of interest for each two week interval
- Constant estimable intercept (α) and regression coefficient (β)
- To test whether a lag time was appropriate for the covariate, the marginal likelihood was estimated after shifting case counts by 0, 1, 2, 3, and 4 weeks. A 2 week lag time resulted in the best model fit.

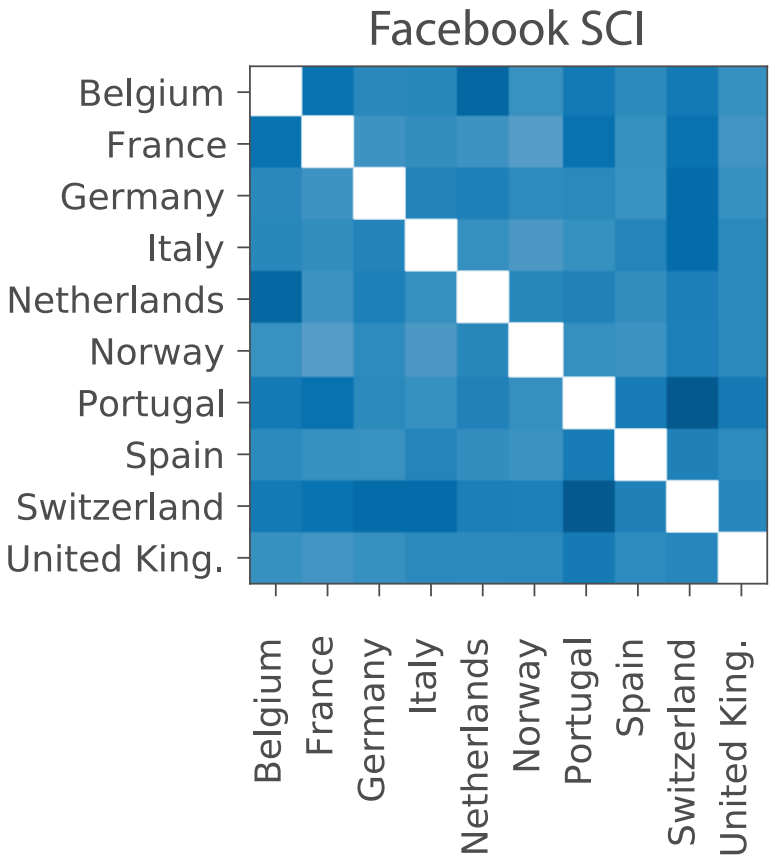
Example: 2020 COVID-19 Resurgence in Europe

The study employed a discrete trait analysis phylogeographic model with three different covariates



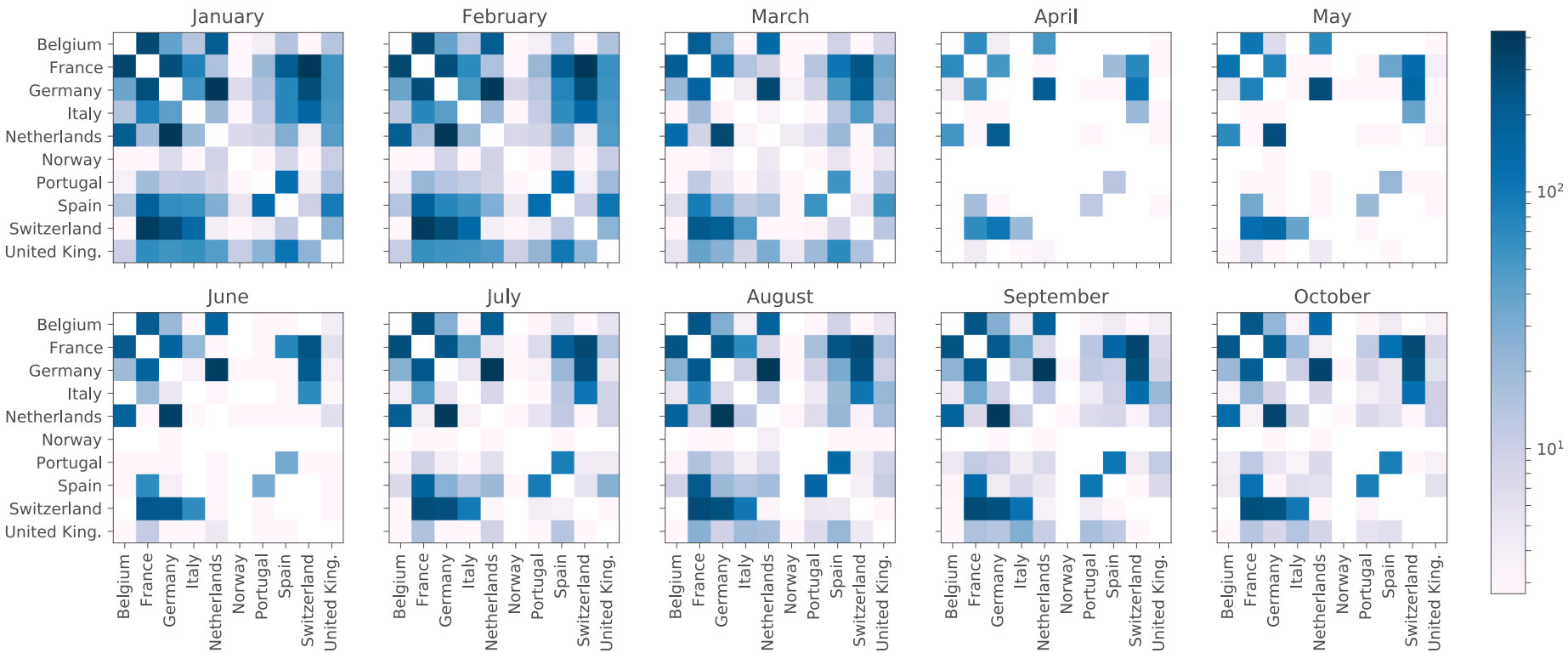
Covariate 1: Monthly Air Traffic Flows (based on origin-destination tickets between the 10 countries between January 2020 and October 2020)

Example: 2020 COVID-19 Resurgence in Europe



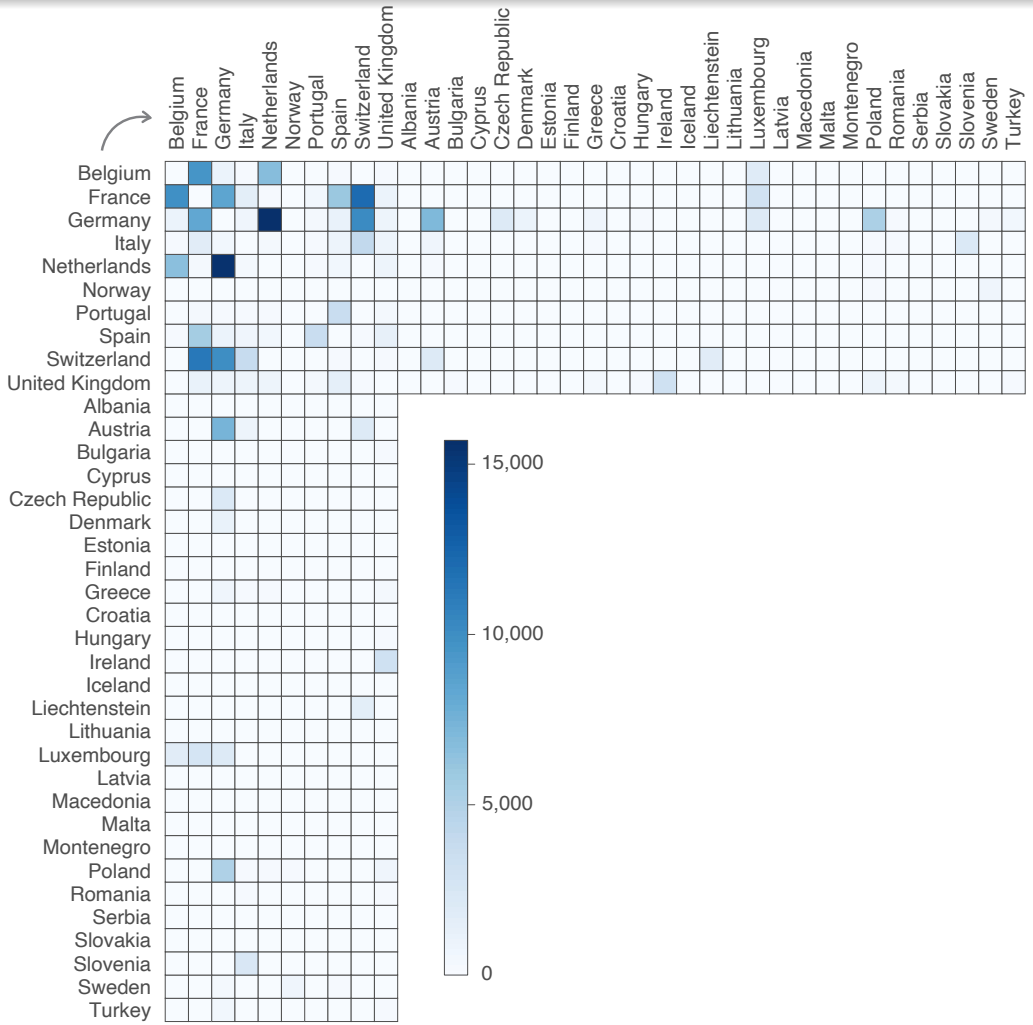
Covariate 2: Facebook social connectedness index (SCI) - anonymized snapshot of active Facebook users and their friendship networks to measure the intensity of social connectedness between countries. Measures relative probability of Facebook friendship link between two users in different countries.

Example: 2020 COVID-19 Resurgence in Europe



Covariate 3: International Google Mobility Data - The Google COVID-19 Aggregated Mobility Research Dataset contains anonymized mobility flows aggregated over users who have turned on location history setting on a range of platforms. We consider aggregated mobility flows between the 10 countries summarized by 2-week or monthly time periods between January and October.

Example: 2020 COVID-19 Resurgence in Europe



Pairwise Google mobility data among 10 countries included in study and other European countries. Mobility to and from each country within 10-country sample covers between 64% and 96% of mobility of these countries to and from all countries within Europe, except for Norway (27%) and UK (49%).

Example: 2020 COVID-19 Resurgence in Europe

Model		**Parameter estimates**
Time-homogenous spatial diffusion	coalescent GLM	α = 2.604 [2.487,2.735], β = 1.711 [1.603,1.898]
	spatial GLM	air travel: E[δ] = 0.018, (β|δ=1) = 0.044 [0.001,0.123] SCI: E[δ] = 0.004, β(|δ=1) = 0.013 [0.003,0.032] mobility: E[δ] > 0.999, β(|δ=1) = 0.358 [0.258,0.456]

Parameter estimates for coalescent model with covariates (coalescent GLM) and discrete trait analysis phylogeographic model with covariates (spatial GLM). δ denotes inclusion probability for phylogeography covariates.

Example: 2020 COVID-19 Resurgence in Europe

To accommodate variability in the mobility measures, it is also possible to apply time-inhomogeneous discrete trait analysis frameworks (Bielejec et al., 2014):

- Monthly covariate data, time-homogeneous effect sizes and inclusion probabilities
- Monthly covariate data, month-specific inclusion probabilities (pooled hierarchically), time-homogeneous effect sizes
- Model transition rates as function of mobility covariate with two-week intervals (include time-homogeneous random effects to account for potential biases in ability of mobility to predict phylogeographical spread). Also, allow overall rate of migration between countries to vary through time via a time-inhomogeneous rate scalar that is parameterized as a log-linear function of the total monthly between-country mobility.

Because of the high computational burden of these time-inhomogeneous models, they are fit to a set of trees inferred under the time-homogeneous model.

Example: 2020 COVID-19 Resurgence in Europe

Time-inhomogeneous spatial diffusion	spatial GLM, constant inclusion probabilities	air travel: $E[\delta] = 0.018$, $\beta(\mid\delta=1) = 0.029$ [0.001,0.105] SCI: $E[\delta] = 0.008$, $\beta \mid \delta=1 = 0.024$ [0.001,0.078] mobility: $E[\delta] > 0.999$, $\beta(\mid\delta=1) = 0.333$ [0.229,0.438]
	spatial GLM, time-variable inclusion probabilities	air travel: $E[\delta_h] = 0.010$, $\beta \mid (\delta_h=1) = 0.047$ [0.002,0.139] SCI: $E[\delta_h] = 0.012$, $\beta \mid \delta_h=1 = 0.018$ [0.000,0.062] mobility: $E[\delta_h] = 0.949$, $\beta(\mid\delta_h=1) = 0.357$ [0.230,0.503]
	spatial GLM time-variable rate scalar GLM	mobility: $\beta = 0.271$ [0.118,0.444] mobility: $\alpha = 0.740$ [0.618,0.856], $\beta = 0.504$ [0.350,0.646]

Here, δ_h denotes the inclusion probability at the hierarchical level

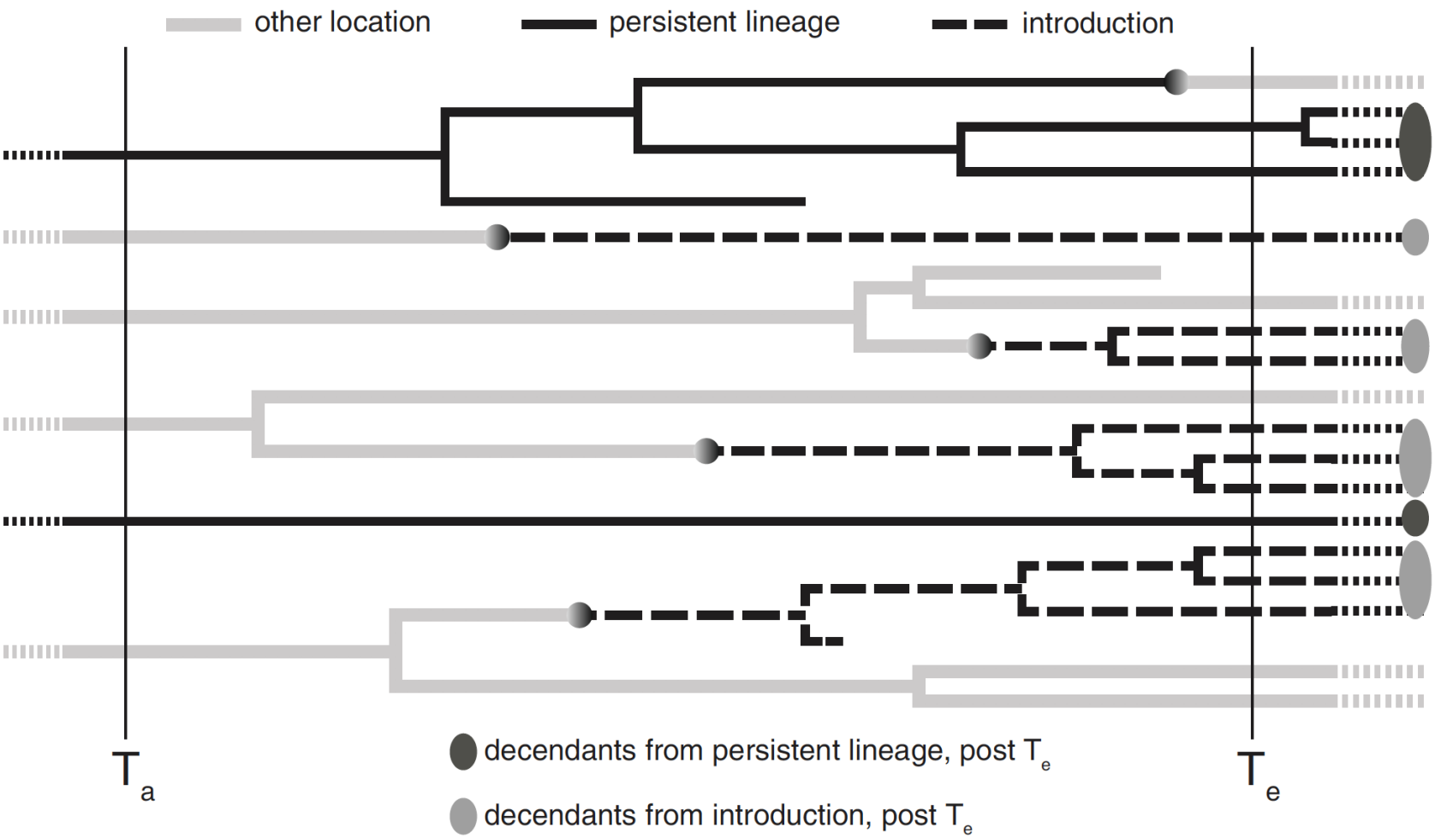
Example: 2020 COVID-19 Resurgence in Europe

Gare du Nord, Paris, August 2020

To assess the impact of summer travel on the second wave in different countries:

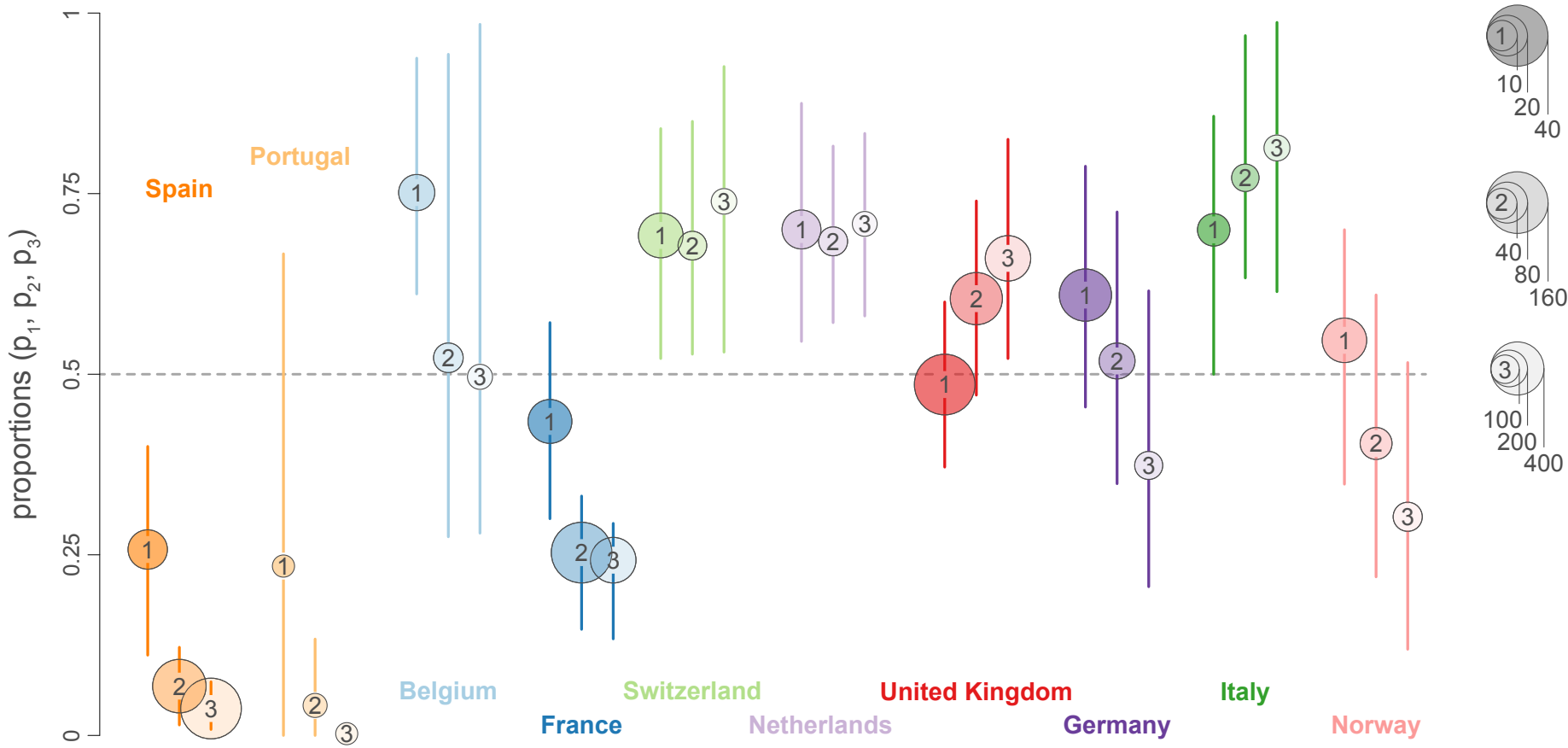
- Focus on a 2-month period between June 15, when borders were opened by most EU and Schengen area countries, and August 15, before which the majority of holiday return travel is expected for most countries
- Identify number of lineages circulating in each country on August 15 and determine whether they resulted from a lineage that persisted since June 15 or from a unique introduction that occurred after June 15

Example: 2020 COVID-19 Resurgence in Europe



12 lineages circulating at evaluation time T_e. Tracking lineages back to ancestral time (T_a), we identify 2 unique persistent lineages and 4 unique introductions. 9 out of 12 lineages at T_e resulted from unique introductions.

Example: 2020 COVID-19 Resurgence in Europe



(1) Ratio of unique introductions over total number of unique persisting lineages and unique introductions between June 15 and August 15 [this is $4/(2+4)$ in example on previous slide], (2) Proportion of descendant lineages on August 15 that resulted from the unique introductions over the total descendants circulating on August 15 [this is $9/12$ in example on previous slide], (3) Proportion of descendant tips (sampled genomes) after August 15 that resulted from unique introductions over total number of descendant tips [not shown on previous slide, but conceptually represented by ovals]. Countries are ordered (left to right) by average incidence between June 15 and August 15.

Example: 2020 COVID-19 Resurgence in Europe

The estimated posterior mean ratio of unique introductions is close to or greater than 0.5 for all countries except Spain and Portugal

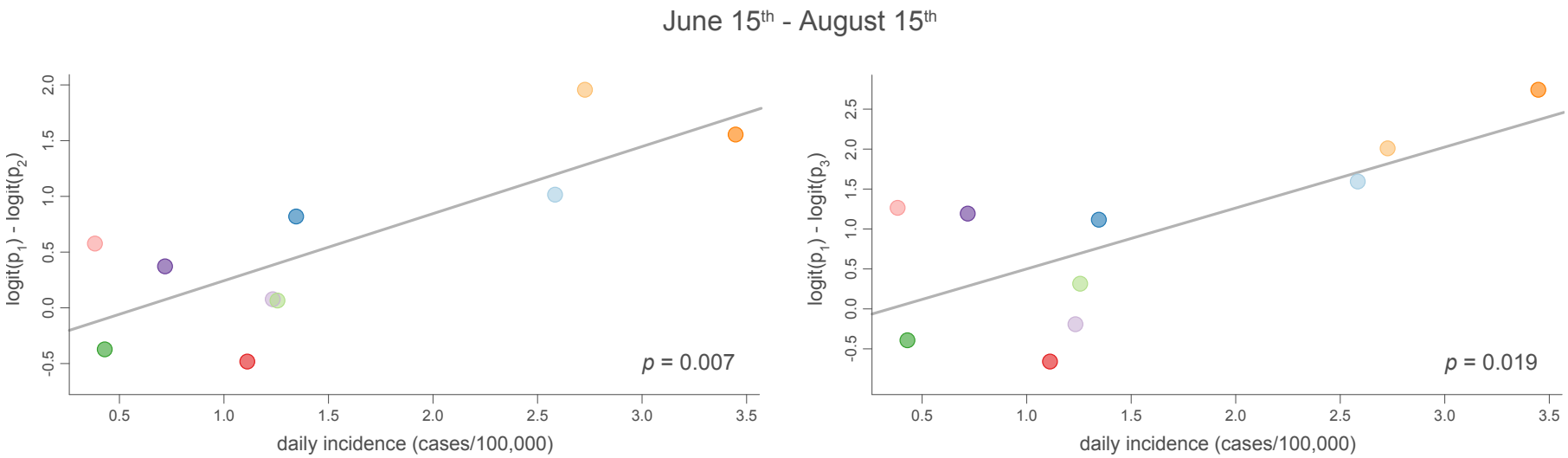
- By August 15, a relatively large number of circulating lineages in most countries was produced by new introductions over summer

Proportions of descendants from these introductions on August 15 and sampling times measure relative success of newly introduced lineages compared to persisting lineages.

- Considerable variation in onward transmission
- In countries that experienced relatively high summer incidence (such as Spain, Portugal, Belgium and France), the introductions lead to comparatively fewer descendants on August 15 or later. In countries that have relatively lower summer incidence, the introductions generally lead to comparatively more descendants on August 15 or later.
- Norway stands out as an outlier. Many lineages that were estimated as persistent may in fact be introductions from Scandinavian countries not represented in genome sample.

Example: 2020 COVID-19 Resurgence in Europe

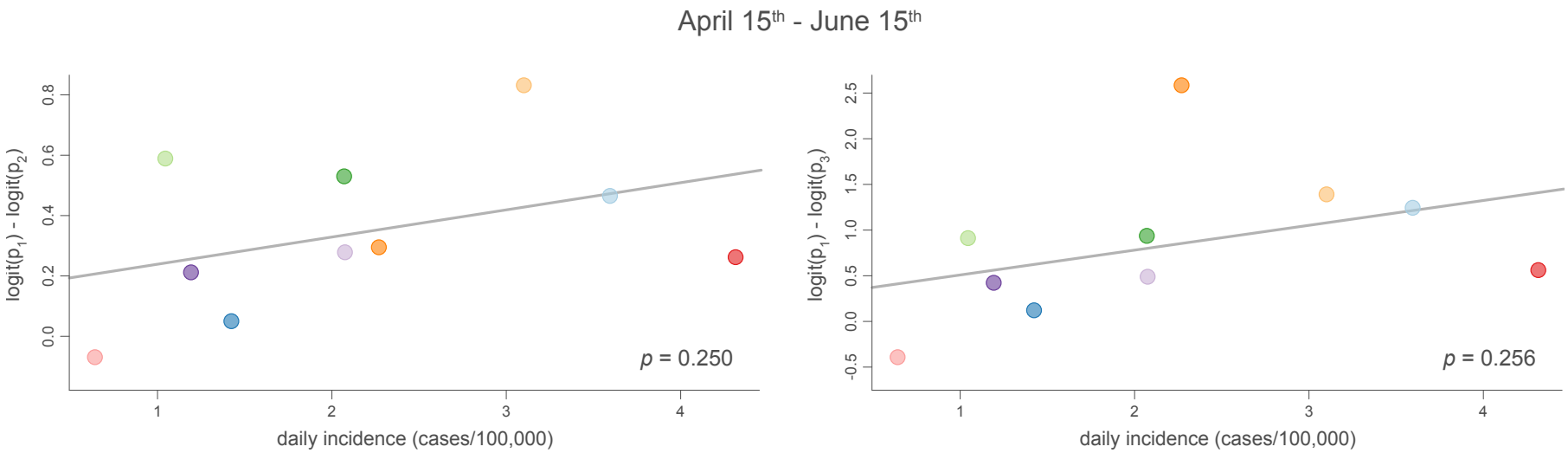
Does incidence shape the outcome of introductions?



Statistically significant overall associations between incidence and success of introductions as characterized by difference between logit-scaled proportion of unique introductions (p_1) and logit-scaled proportion of their descendants on August 15 (p_2)/descendant tips after August 15 (p_3).

Example: 2020 COVID-19 Resurgence in Europe

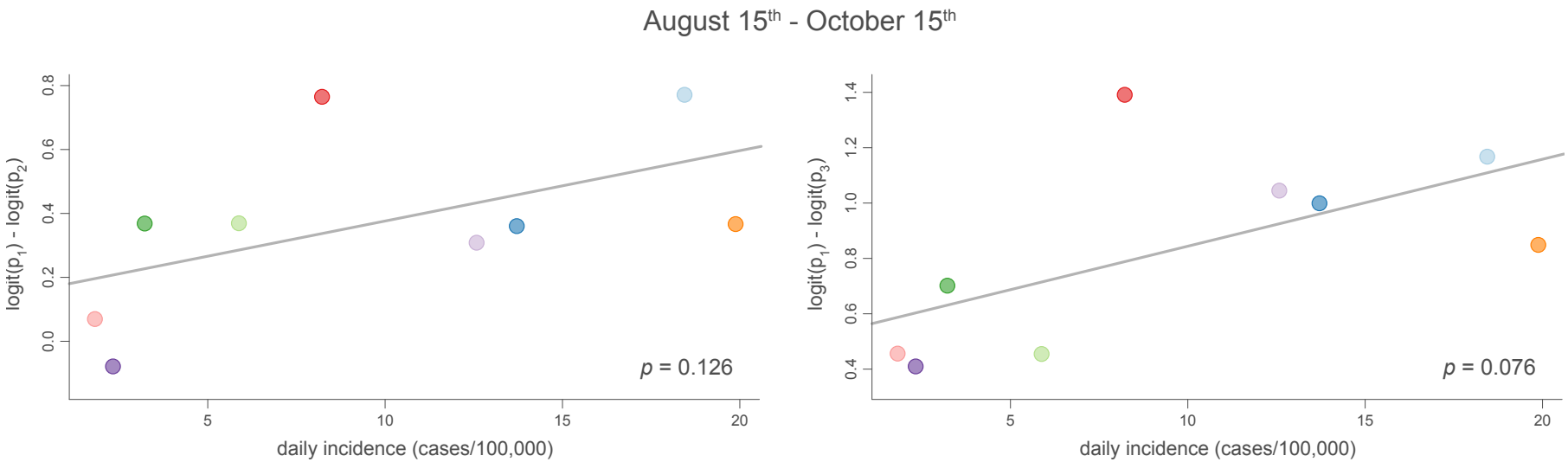
Assessing the relationships for the period April 15 - June 15



Relationship is qualitatively similar, but more variable and not statistically significant

Example: 2020 COVID-19 Resurgence in Europe

Assessing the relationships for the period August 15 - October 15



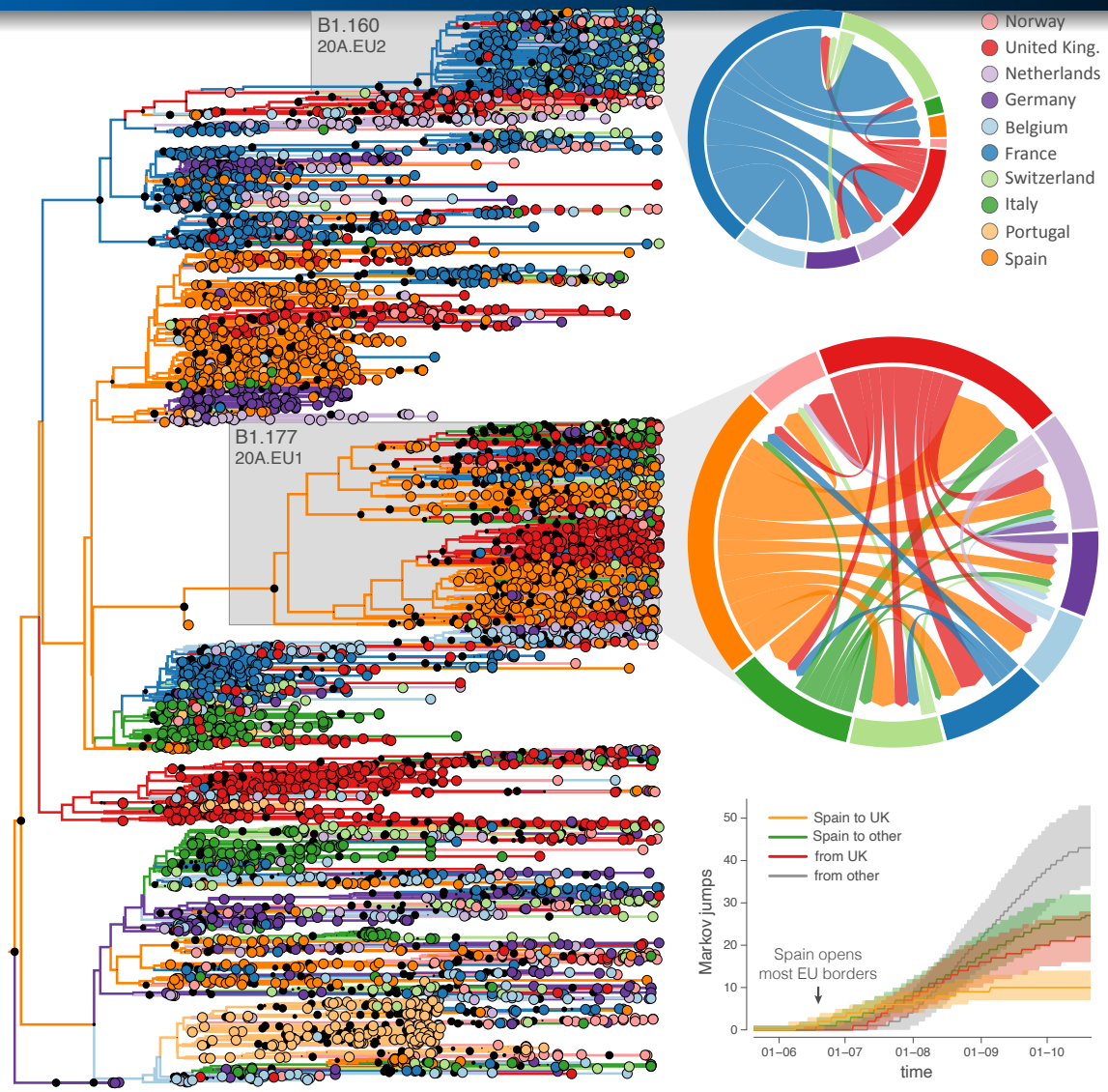
Relationship is qualitatively similar, but more variable and not statistically significant

Example: 2020 COVID-19 Resurgence in Europe

Lemey et al. (2021) conclude in their study that new introductions were able to thrive in low incidence settings: the comparatively higher proportion of introductions as well as lower and more stable incidence between June 15 and August 15 provided favorable conditions for a process of genetic drift by which introductions were able to fuel transmission.

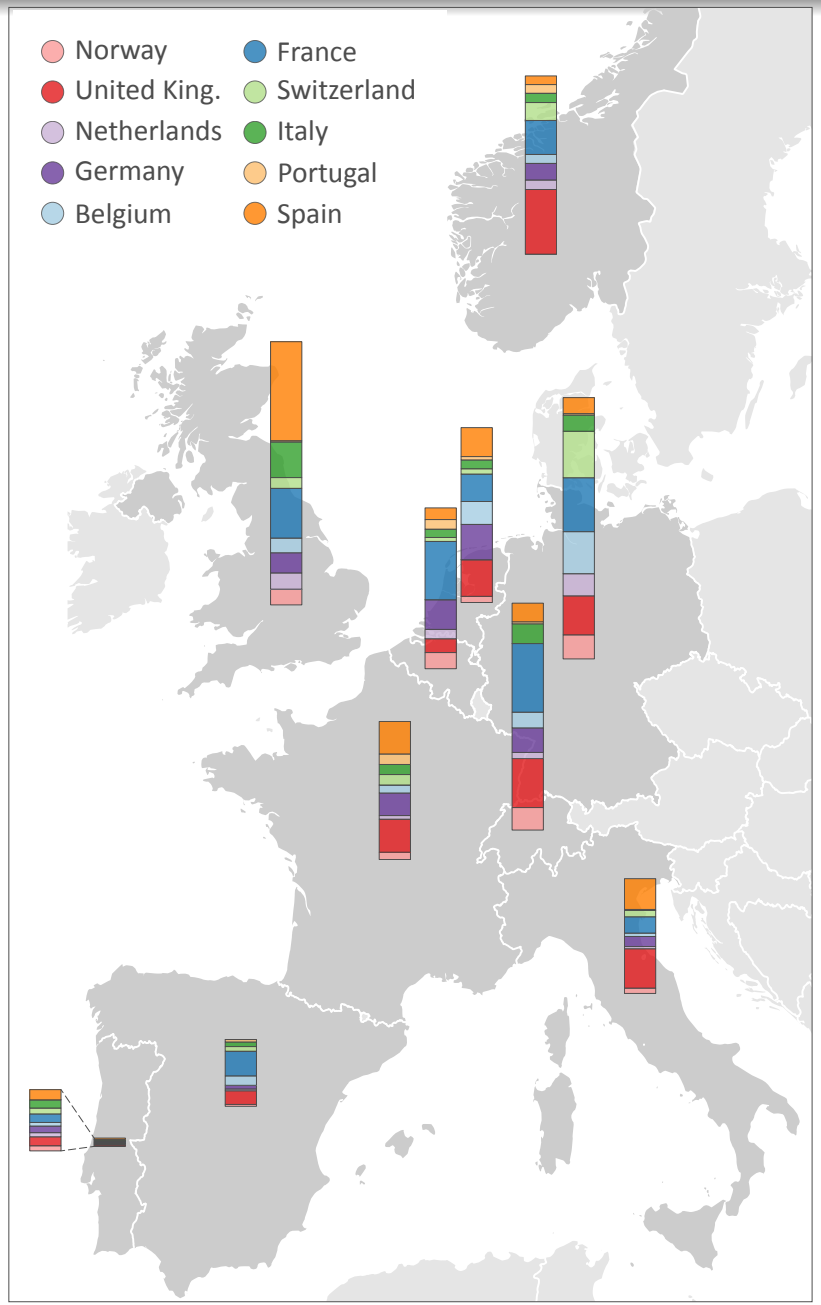
Introductions in the UK particularly benefitted from conditions for successful onward transmission, with a considerable fraction of introductions originating from Spain, reflecting the spread of B.1.177/20E (EU1), which quickly became the most dominant strain in the UK. Although Spain was inferred as the origin of this strain, the UK contributed substantially to its spread.

Example: 2020 COVID-19 Resurgence in Europe



Maximum clade credibility tree annotated with ancestral location realizations. Circular migration flow plots show migration flow out of a particular location, starting close to the outer ring, ending with an arrowhead more distant from the destination location. Bottom right shows posterior mean estimates with 95% HPD intervals over time for four types of transitions for B1.177/20E (EU1): from Spain to the UK, from Spain to other countries, from the UK, and from other countries.

Example: 2020 COVID-19 Resurgence in Europe

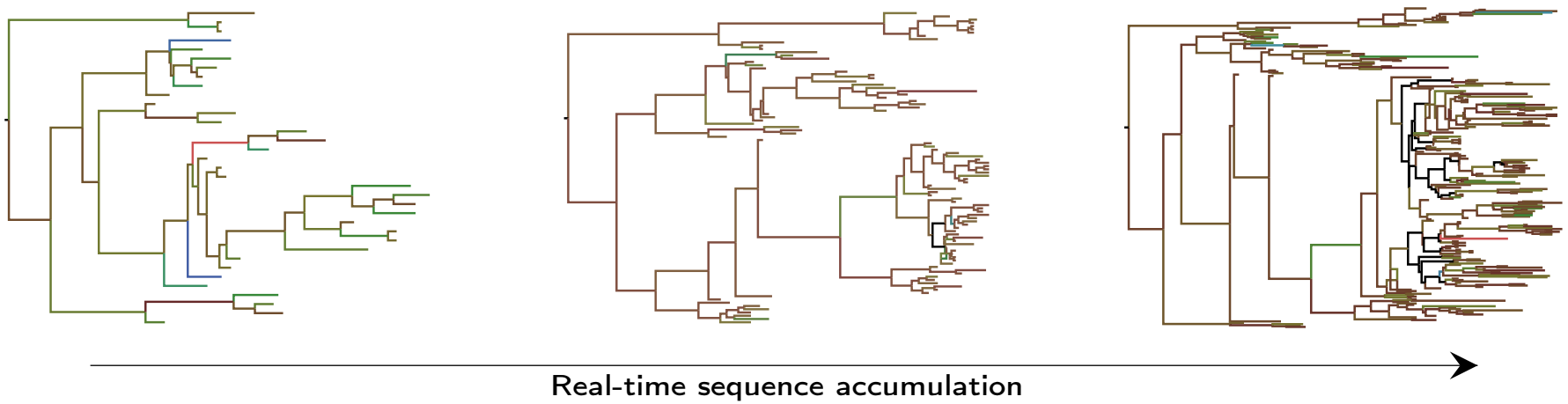


Estimated geographical origin of viral influx between June 15, 2020 and August 15, 2020 in each country.

Each bar plot summarizes the posterior transition estimates into a specific country.

The bar plot for Portugal is very small (representing a low number of introductions into Portugal), so a magnified view is provided.

Need for "Real-Time" Phylodynamic Inference

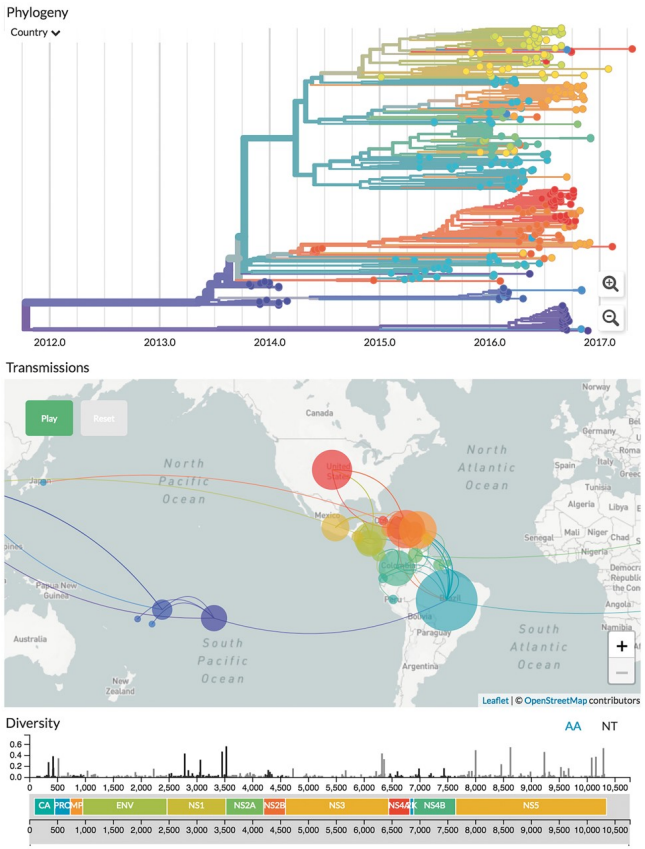


Advances in sequencing technology have enabled genomic surveillance during outbreaks and epidemics in close to real-time

Phylodynamic inference can provide valuable information about pathogen evolution and spread

If the "real-time" sequencing could be matched by "real-time" phylodynamic inference that can rapidly extract updated epidemiological insights as new data become available, it would be very helpful for formulating timely control and intervention strategies

Need for "Real-Time" Phylodynamic Inference



Nextstrain display

Researchers have developed phylogenetic inference software, such as TreeTime, to help meet this challenge

There has also been development of valuable platforms, such as Nextstrain, that facilitate real-time epidemic tracking through bioinformatics pipelines for phylodynamic analysis and visualization tools

Need for "Real-Time" Bayesian Phylodynamic Inference

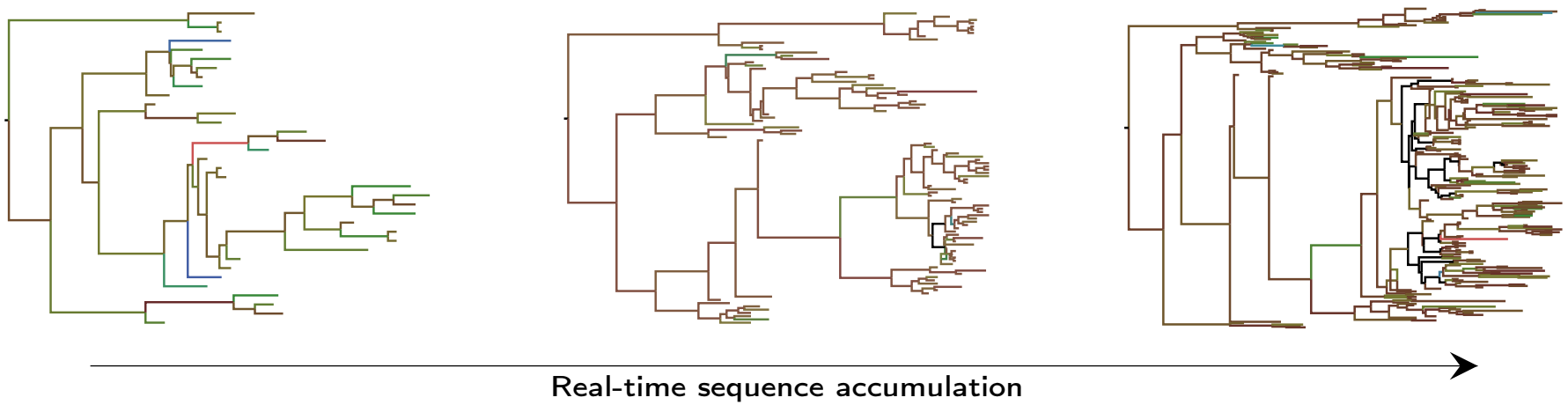
Current methods for rapid phylodynamic inference are based on approximate maximum likelihood methods and do not account for phylogenetic uncertainty

Bayesian phylogenetic inference frameworks naturally account for all sources of model uncertainty

- Especially valuable for short outbreak timescales for which the phylogeny may not be well-resolved

However, Bayesian phylodynamic inference relies on Markov chain Monte Carlo (MCMC) methods that can have very long run times

Need for "Real-Time" Bayesian Phylodynamic Inference

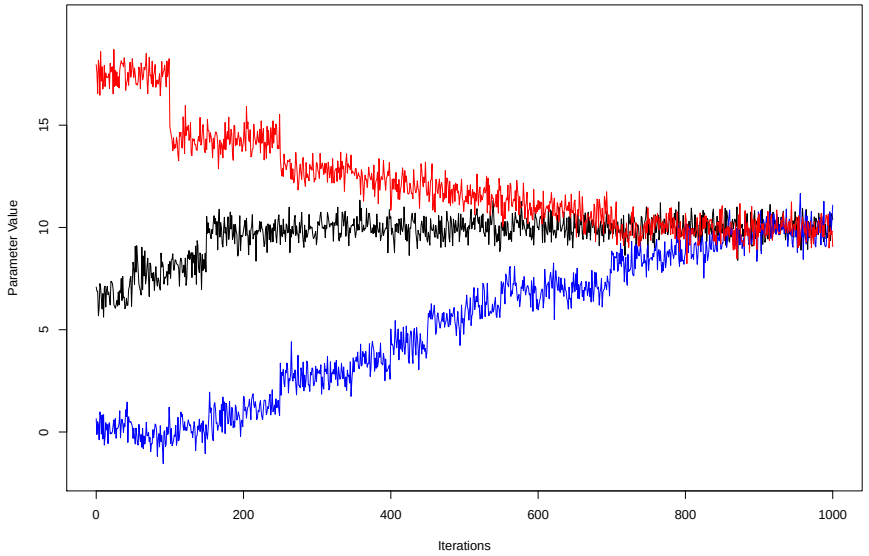


How can we move closer to "real-time" Bayesian phylodynamics?

Run times can be very long, and this situation is exacerbated by the fact that when new data become available, we have to interrupt the current analysis, throw away our results, and restart the analysis from scratch on the enlarged data set

We propose: more efficient "online" approach that can accommodate a continuous stream of new data, using previous inferences to help compute updated inferences

Markov chain Monte Carlo Convergence



Burn-in: transient phase at beginning of chain, prior to convergence to stationary distribution, during which simulated values are influenced by starting values of chain. Discarded to avoid bias.

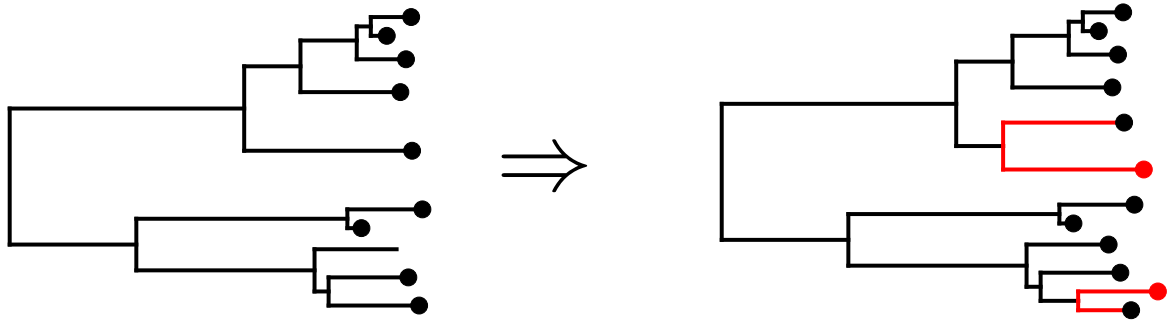
- Length can depend on how far starting values are from true distribution
- Burn-in for phylogenetic models on realistic data can be very long and represent large proportion of iterations required to generate sufficient posterior sample

Goal: Minimize burn-in when updating inferences with new data

Strategy: Use previous inferences along with new sequences to construct good starting values for analysis of expanded data set

Online Bayesian Phylogenetic Inference

phylogeny with sequences at tips:



Online Bayesian phylogenetic inference in BEAST X:

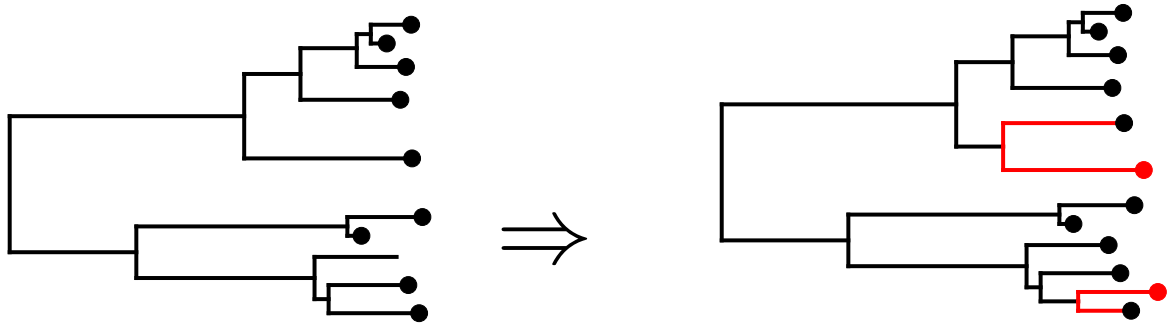
- After chain has achieved stationarity, and after arrival of new sequences, interrupt analysis
- Extract parameter values from last iteration in interrupted analysis
- **Sequence Insertion:** Insert each of the new sequences into the phylogeny extracted from the interrupted analysis

  For each new sequence, compute genetic distance from sequences already in tree

  Minimum genetic distance determines height at which to insert a common ancestor node for the new sequence and the most genetically similar sequence

Online Phylogenetic Inference

phylogeny with sequences at tips:

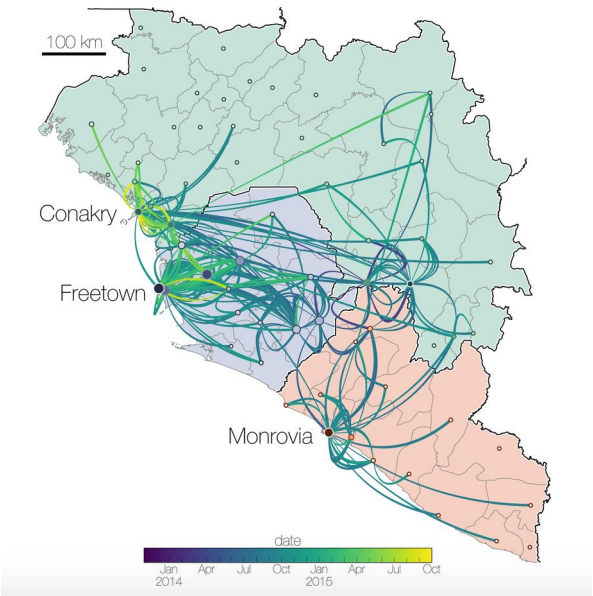


branch-specific parameters: $(\phi_1, \ldots, \phi_{18})$ $(\phi_1, \ldots, \phi_{18}, \phi_{19}, \ldots, \phi_{22})$

other parameters: $(\boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\Pi})$ $(\boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\Pi})$

- Impute plausible values for new parameters that are introduced to accommodate new tree branches (e.g., relaxed clock model rates)
- Parameter values for models that do not assume increased dimensionality are carried over (e.g., substitution model parameters)
- Resume by commencing analysis of expanded data set using the "good" starting parameter values that we have constructed

Example: Ebola Virus Epidemic



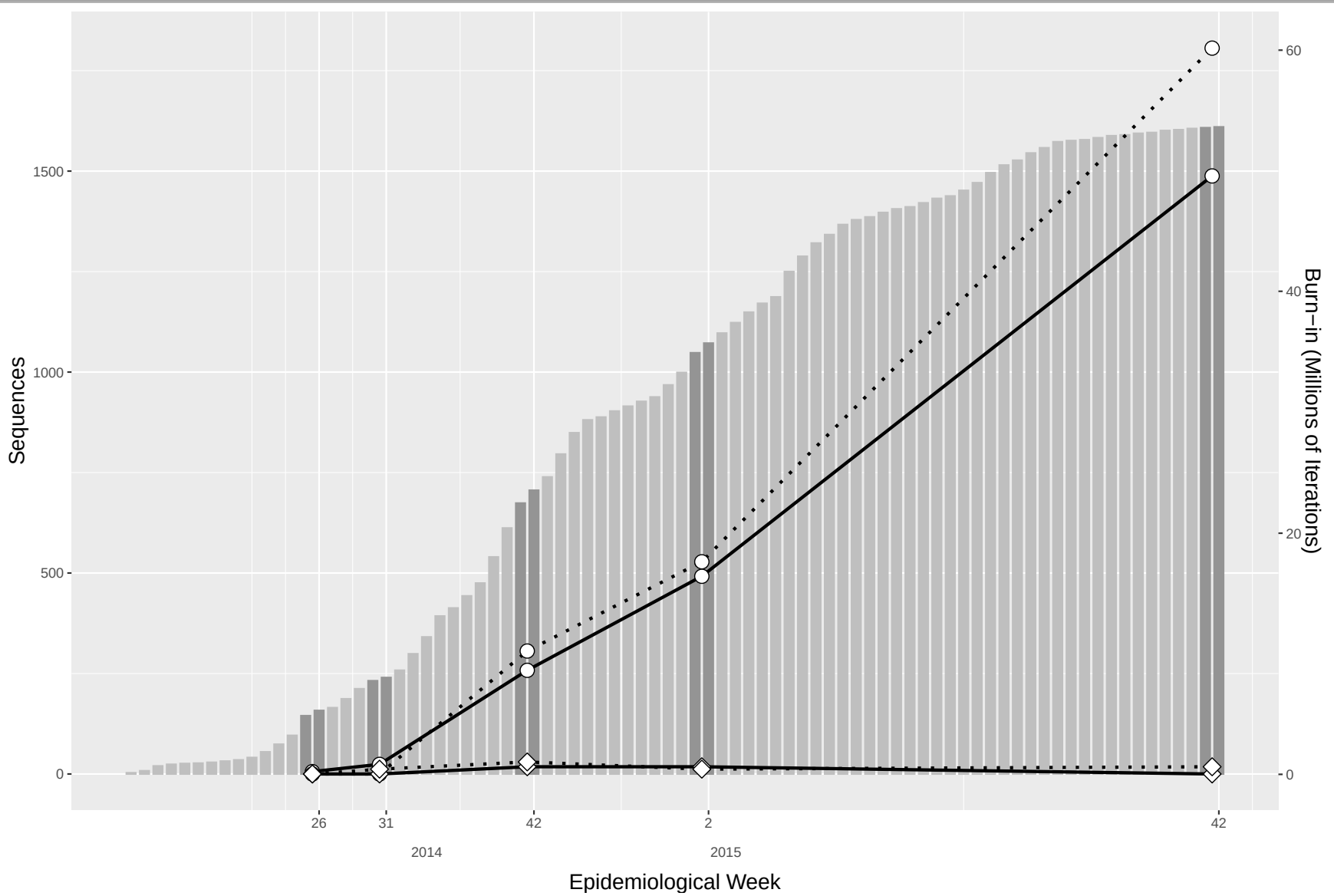
(Dudas et al., 2017)

Evaluate performance on 1610 whole genome sequences from West African Ebola virus epidemic of 2013-2016 (Dudas et al., 2017)

We compute updated inferences at different points of epidemic using two approaches:

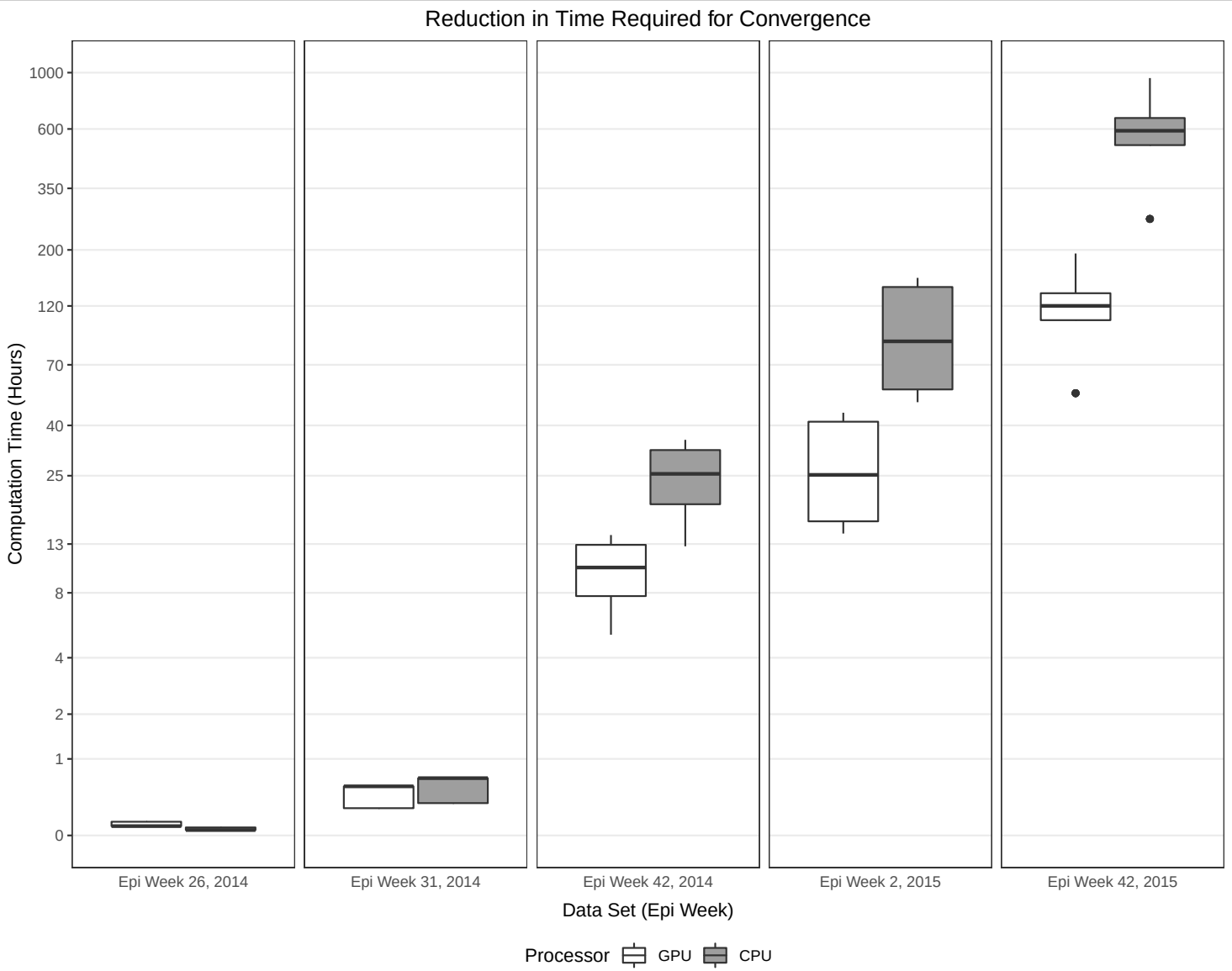
- **Standard Analysis:** analyze complete data set from scratch
- **Online Analysis:** use results from analysis of all data available by previous week, along with newly arrived data from current week, to construct starting values

Results: Burn-in for Updated Inferences



Average burn-in required by standard analyses, represented by circles, and online analyses, represented by diamonds. Solid lines correspond to burn-in estimates based on visual analyses of trace plots while dotted lines correspond to burn-in estimates based on maximizing ESS values.

Savings in Computation Time



Gray: 14-core 2.20 GHz Intel Xeon Gold 5120 CPU

White: Tesla P100 graphics card intended for scientific computing