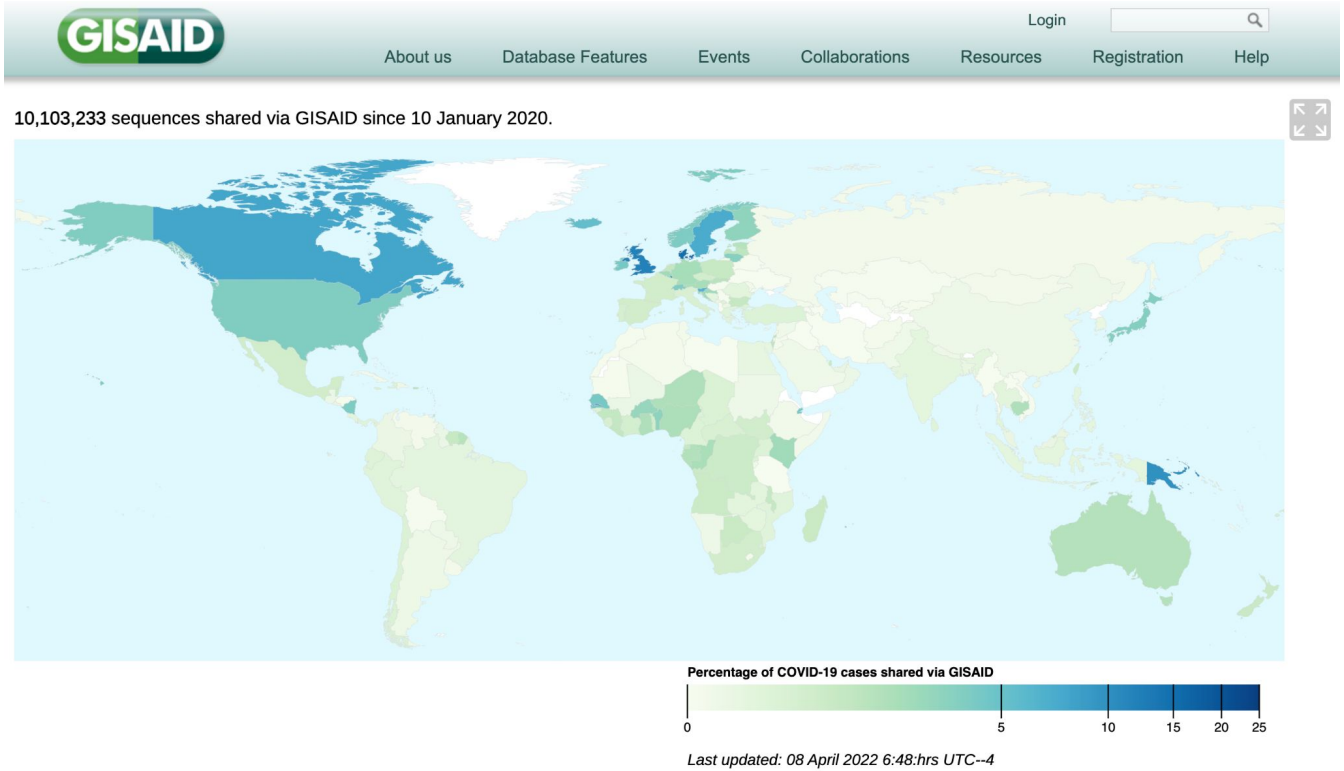
After the data deluge: scaling strategies for massive genomic datasets

Molecular Epidemiology of Infectious Diseases

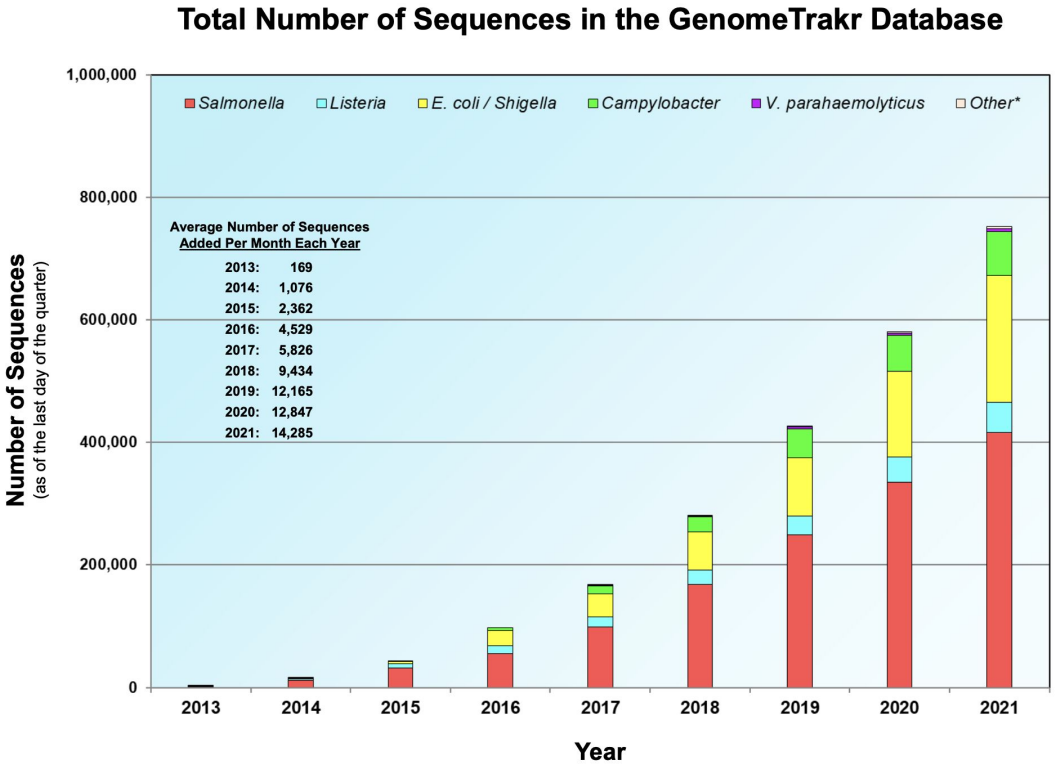
Lecture 11

April 11th, 2022

The data deluge



The data deluge



https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-fast-facts

Massive genomic datasets have pushed the limits of existing methods

The benefits of Bayesian inference have a cost

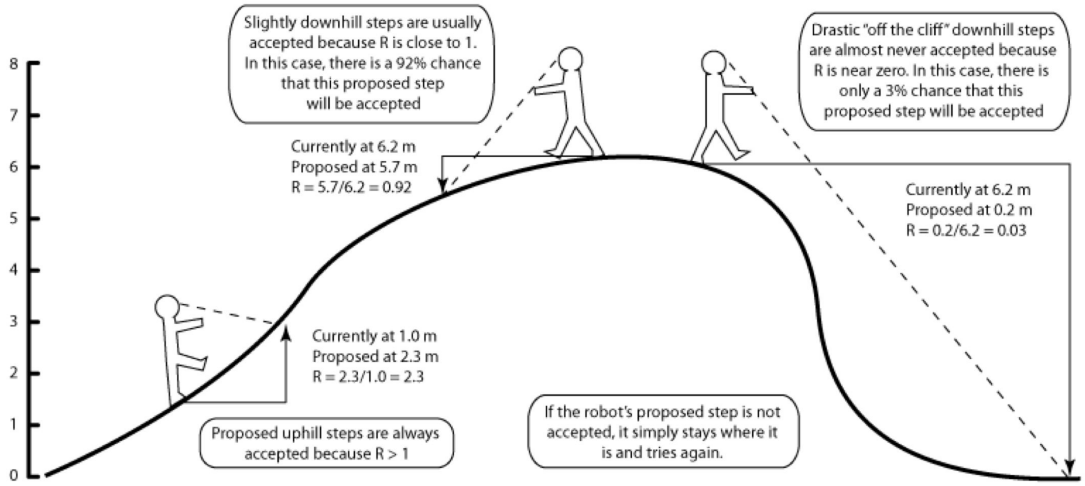
We have primarily focused on Bayesian phylogenetic methods because they allow for flexible, probabilistic modeling of pathogen evolution

While quantifying our uncertainty in epi/evo parameters as well as the pathogen phylogeny.

BUT these Bayesian methods are highly reliant on MCMC sampling.

The costs of Bayesian MCMC

Like a blind robot on a random walk, MCMC is inherently inefficient



Problems with MCMC efficiency

Sampling via a random walk is inherently inefficient as it may take a long time for the chain to converge on the posterior distribution.

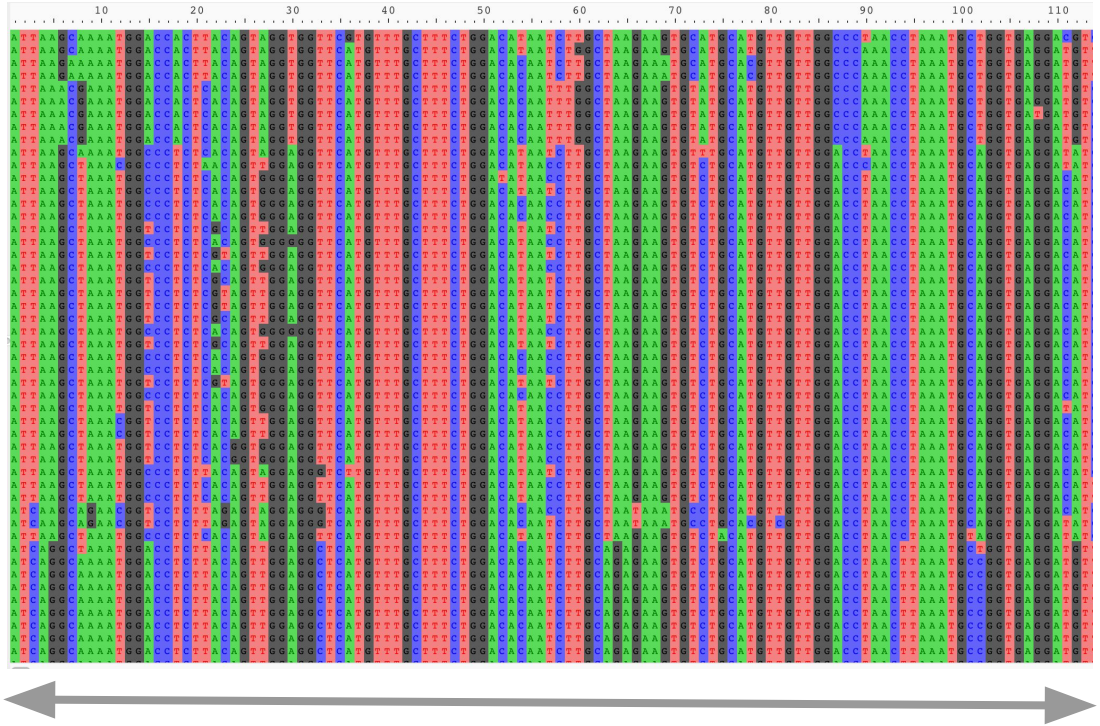
Using a Metropolis-Hastings step means many proposals are rejected.

Samples are generally highly autocorrelated, requiring a lot of wasted computation (thinning) to get pseudo-independent samples from the posterior.

Faster gradient-based MCMC methods (e.g. Hamiltonian Monte Carlo, NUTS) cannot be easily adapted to sampling from tree space.

What do we even mean by big?

Long (many samples)



Wide (many sites/characters)

What is the bigger problem: the length or width of a genomic dataset?

Scaling with the number of sites

The likelihood of the sequence data is generally computed independently at each site.

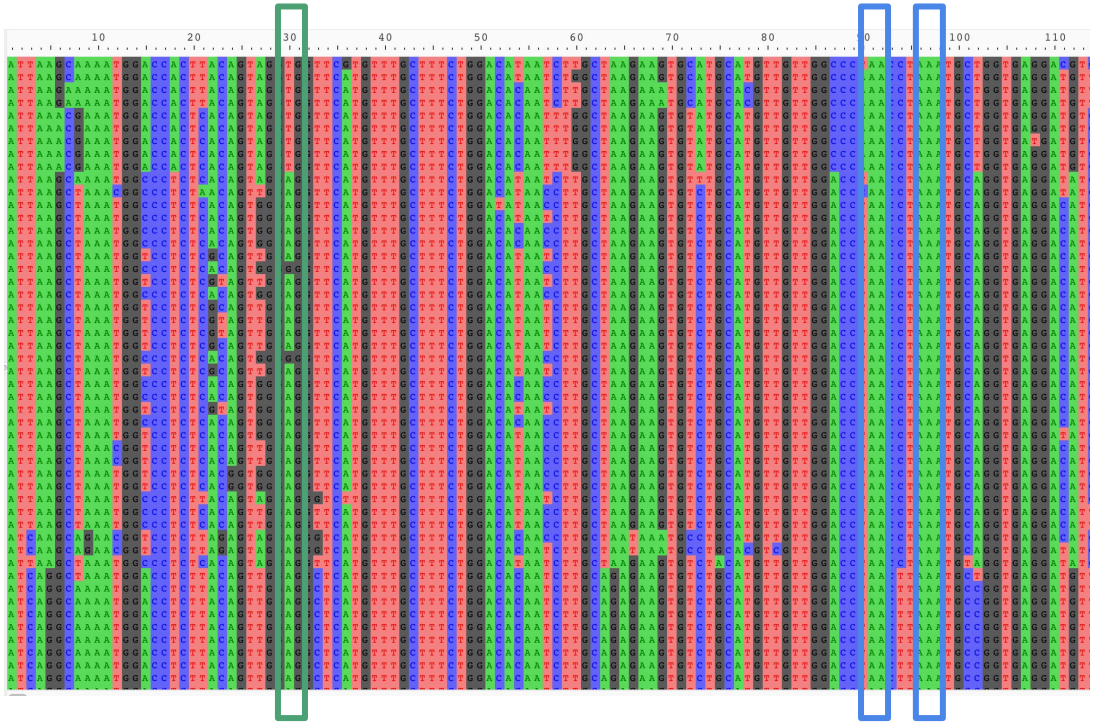
$$L(Seq|Tree) = \prod_{i=1}^{i=N} L(Seq_i|Tree)$$

Computation time therefore generally increases linearly with the number of sites (width) of an alignment.

Or even less than linearly because we actually only need to compute the likelihood once for each **unique site pattern** in an alignment.

Site patterns

Many site patterns will not be unique such that the cost of computing the likelihood will generally increase slightly less than linearly with the number of sites in an alignment.

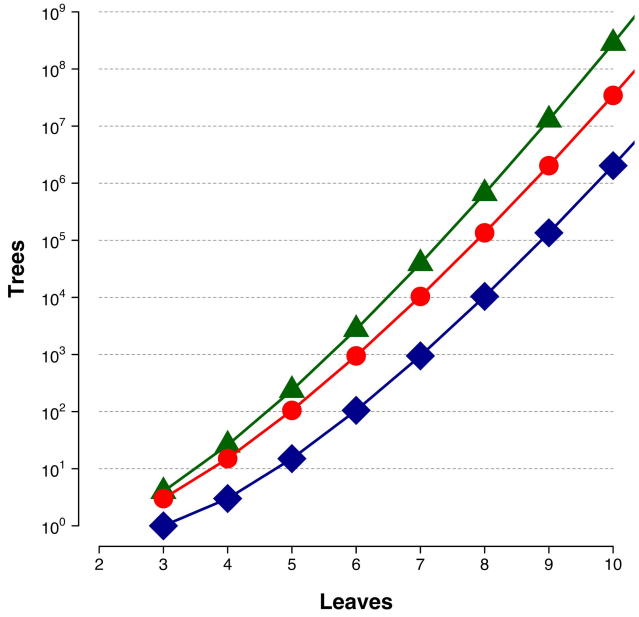


Scaling with the number of samples

On the other hand, tree space grows exponentially with the number of tips or samples

We also need to estimate an additional branch length parameter for each sample we add.

MCMC methods thus scale very poorly with the number of samples

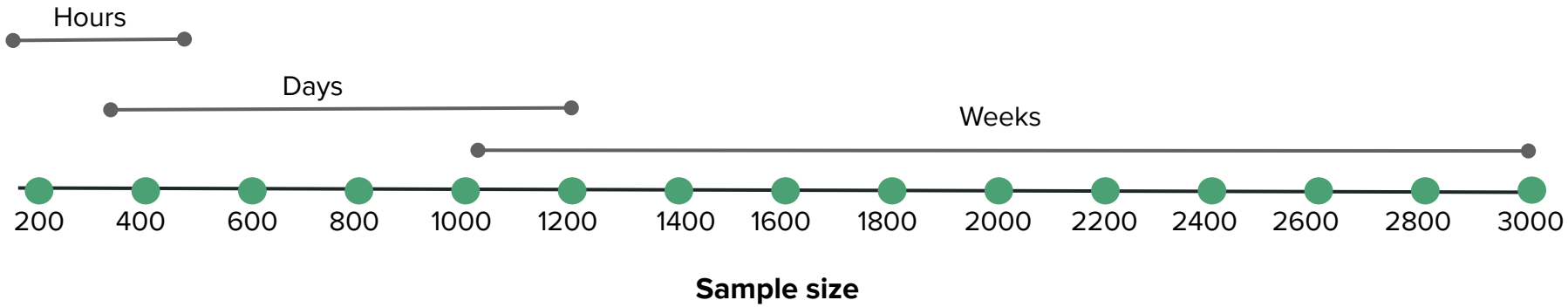


Red shows rooted binary trees.

Practical limits of Bayesian phylogenetics

Generally Bayesian phylogenetic inference using MCMC is limited to at most a few thousand samples.

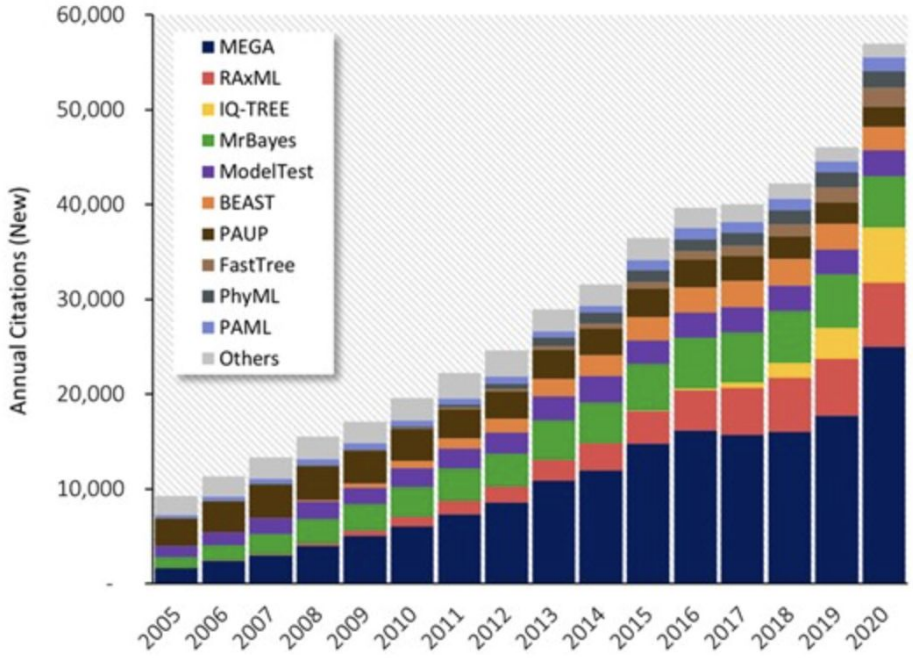
Assuming a typical 10kb alignment, the figure below gives a rough approximation of how run times scale with sample size.



Faster tree reconstruction methods

Nothing wrong with ML

A large number of biologists still use ML or other fast tree reconstruction methods.



Kumar (MBE, 2022)

Fast ML phylogenetic methods

Many popular ML-based methods use a similar reconstruction strategy:

1. A starting tree is obtained using a very fast approximate method like Neighbor Joining or Maximum Parsimony.
2. A hill climbing algorithm is used to search for trees with higher likelihood:
    a. A new tree is proposed by a tree-rearrangement move like NNI or SPR.
    b. Tree is accepted only if it increases the likelihood of the sequence data.
3. The search stops once no further improvements to the likelihood can be found
4. Optional: The entire process starting from step 1 is repeated with a different starting tree.

Fast ML phylogenetic methods

Zhou *et. al.* compared the performance of four of the most popular ML methods.

Table 1. Overview of the Four Fast ML-Based Phylogenetic Programs Evaluated in This Study.

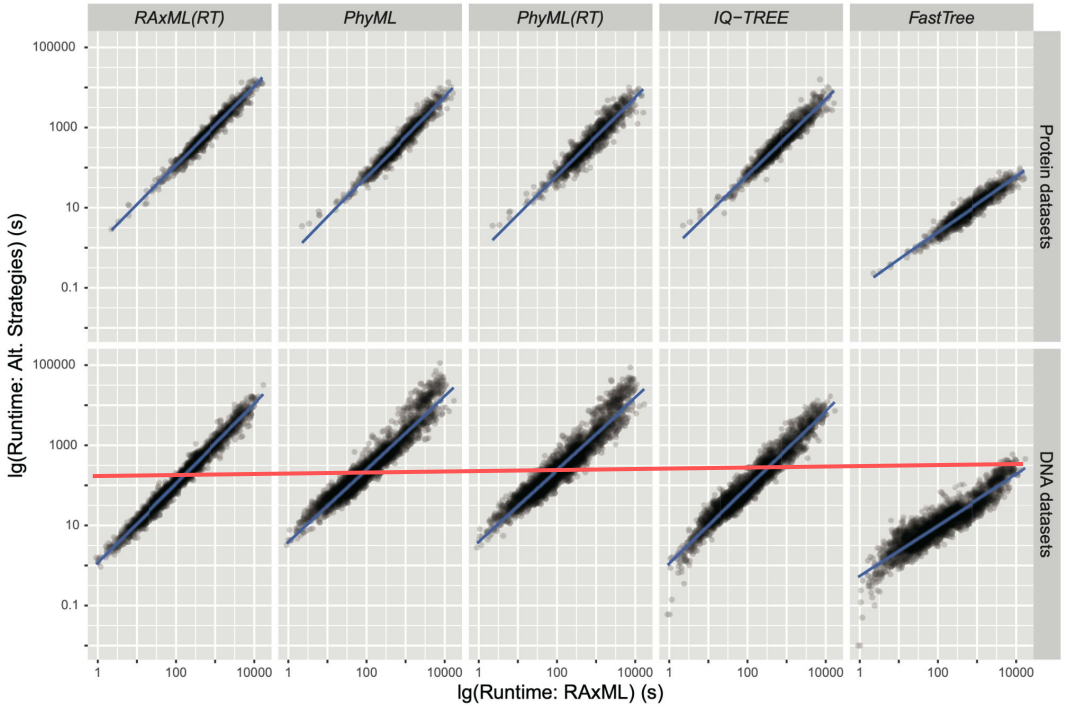
Programs	Optimality Criterion	Starting Tree	Topological Moves	Supported Models		Partitioned Analysis
				AA	DNA	
RAxML v8.2.0 (ExaML v3.0.17)	ML	Parsimony/random/custom	SPR	Common and custom models	JC69, K80, HKY85, GTR	Y
PhyML v20160530	ML	Parsimony/random/custom	Interleaved NNI and SPR	Common and custom models	Common and custom models	Y
IQ-TREE v1.4.2	ML	BIONJ and multiple parsimony/random/custom	NNI and stochastic perturbation	Common and custom models	Common and custom models	Y
FastTree v2.1.9	ML	Heuristic NJ	NNI and SPR (ME) followed by NNI (ML)	JTT, WAG, LG	JC69, GTR	N

NOTE.—ML, maximum likelihood; ME, minimum evolution; NJ, neighbor joining; NNI, nearest neighbor interchange; SPR, subtree pruning and re-grafting.

Zhou et al. (MBE, 2017)

Fast ML phylogenetic methods

FastTree can be 10-100X times faster than similar ML tree approaches.



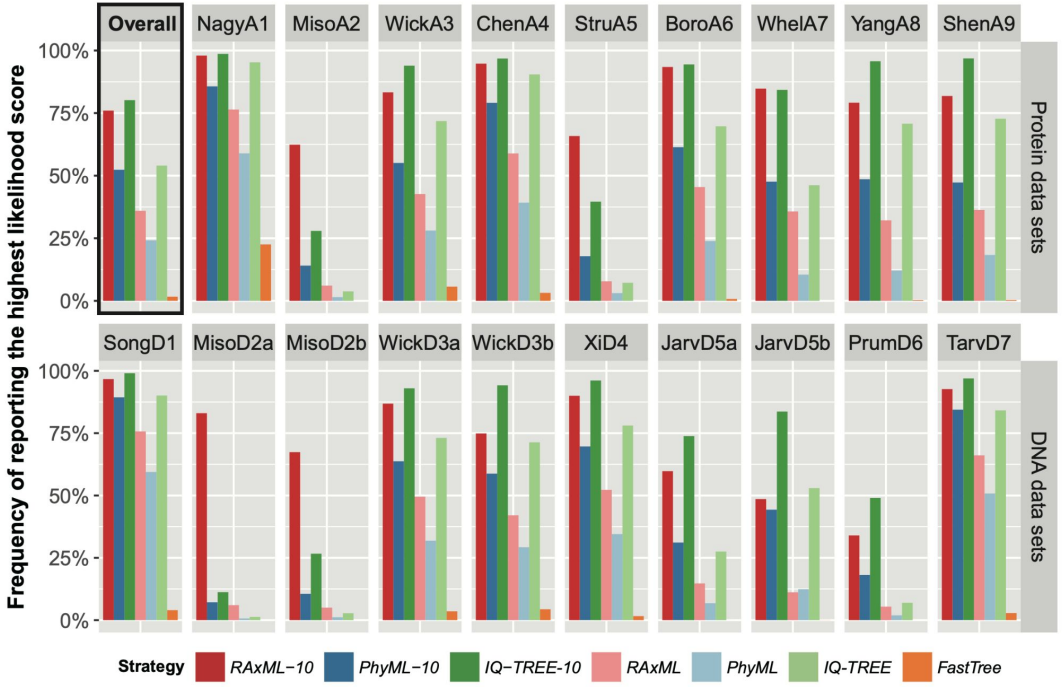
Zhou et al. (MBE, 2017)

Fast ML phylogenetic methods

RAxML, PhyML and IQ-TREE perform about equally well.

Using multiple starting trees (x-10 runs) considerably helps in finding the best tree.

FastTree is considerably less accurate and rarely recovers the highest likelihood tree.



Zhou et al. (MBE, 2017)

Fast ML phylogenetic methods

There is an essential tradeoff between speed and accuracy determined by how exhaustively different ML methods search tree space.

My vote: These days I generally use RAxML or IQ-TREE. IQ-TREE might be gaining an edge since it has some nice features like built-in dating and ultra-fast bootstrap searches.

But I still use FastTree for explorative analyses with very large alignments (>10,000 sequences) when fine-scale accuracy is less important than speed.

Don’t “reinvent the tree”

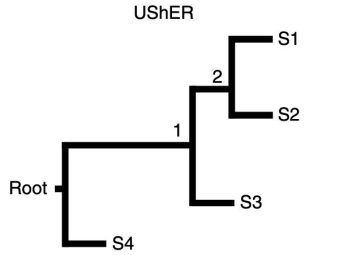
During epidemics, the number of sequences will generally grow rapidly over time, requiring *online* or *real-time* methods that can quickly incorporate new data.

Many fast ML methods like IQ-TREE and RAxML (via EPA-ng) allow for new samples to be placed on existing trees. This is known as ***sample placement***.

This can however still be time consuming. Turakhia *et al.* (2021) found that it takes about 28 minutes to place one SARS-CoV-2 sample on an existing reference tree with 38,342 tips.

Ultra-fast placement algorithms

UShER uses a maximum parsimony approach to search a *mutation-annotated tree* for a placement that requires the fewest additional mutations.



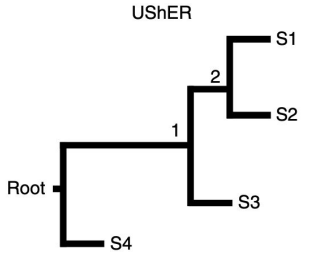
Node	List of mutations
0 (root)	( )
1	(G1U, C4U, G6A)
2	(C7A)
S1	(C8U)
S2	(A9G)
S3	(A12G)
S4	(A13G)
S5	(G1U, C7A)

206 bytes

Turakhia et al. (Nat. Genetics, 2021)

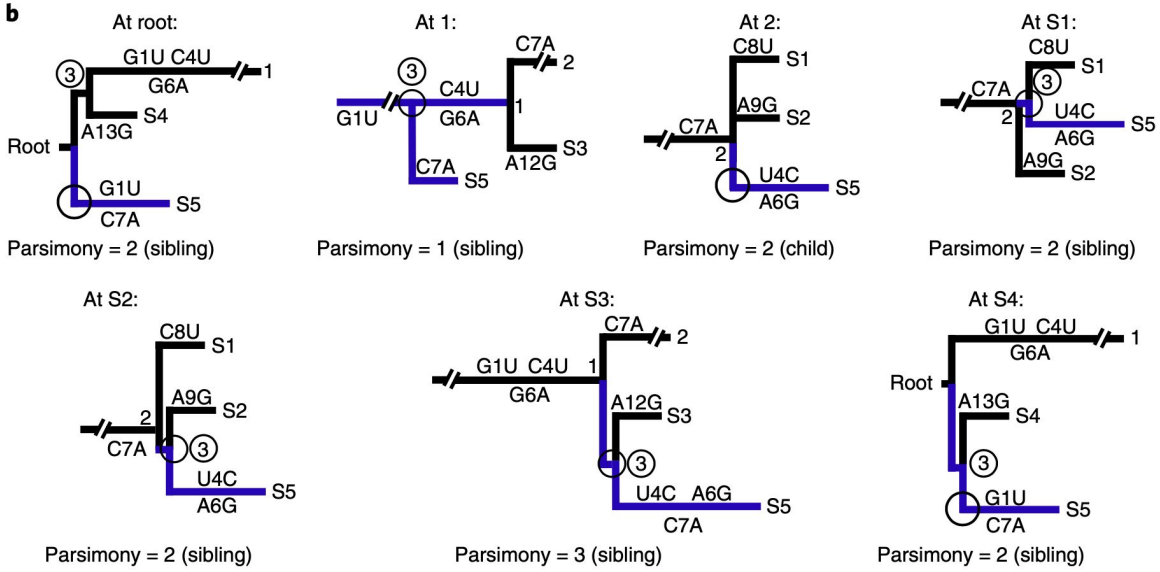
Ultra-fast placement algorithms

UShER uses a maximum parsimony approach to search a *mutation-annotated tree* for a placement that requires the fewest additional mutations.



Node	List of mutations
0 (root)	( )
1	(G1U, C4U, G6A)
2	(C7A)
S1	(C8U)
S2	(A9G)
S3	(A12G)
S4	(A13G)
S5	(G1U, C7A)

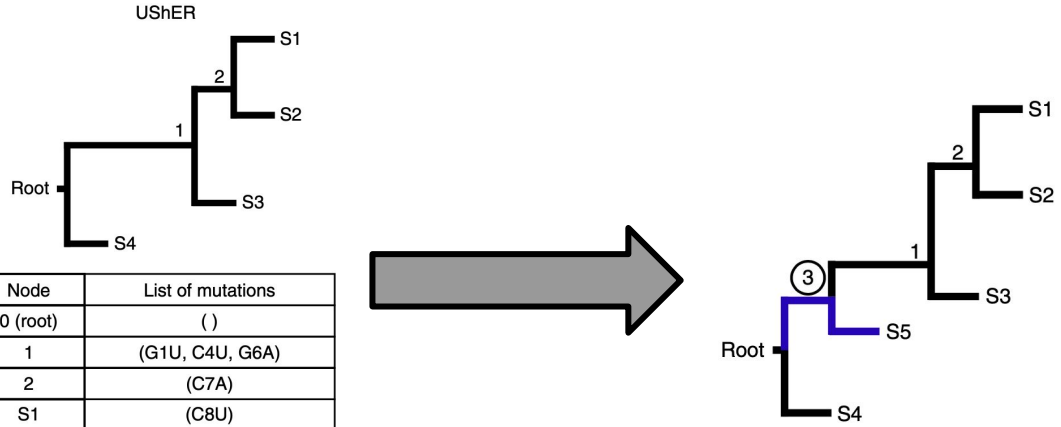
206 bytes



Turakhia et al. (Nat. Genetics, 2021)

Ultra-fast placement algorithms

UShER uses a maximum parsimony approach to search a *mutation-annotated tree* for a placement that requires the fewest additional mutations.



Node	List of mutations
0 (root)	( )
1	(G1U, C4U, G6A)
2	(C7A)
S1	(C8U)
S2	(A9G)
S3	(A12G)
S4	(A13G)
S5	(G1U, C7A)

206 bytes

Node	List of mutations
0 (root)	( )
3	(G1U)
1	(C4U, G6A)
2	(C7A)
S1	(C8U)
S2	(A9G)
S3	(A12G)
S5	(C7A)
S4	(A13G)

Turakhia et al. (Nat. Genetics, 2021)

Ultra-fast placement algorithms

Fast placement algorithms allow for ‘real-time’ phylogenetic inference during large epidemics by adding new samples and only periodically rebuilding the reference tree.

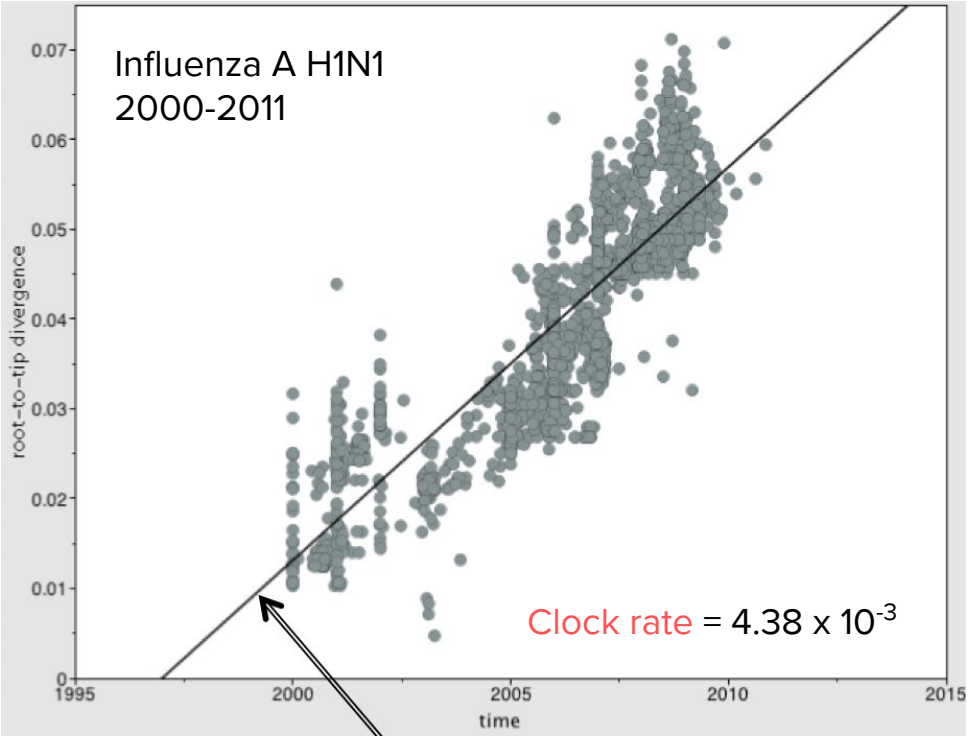
Works well when genetic diversity is limited and sequences are only separated by a few mutations.

Danger zone: Parsimony is known to be less accurate than ML when there’s a reasonably high probability that multiple mutations (including reversions) have occurred along a branch.

Faster dating: methods for time-calibrated phylogenies

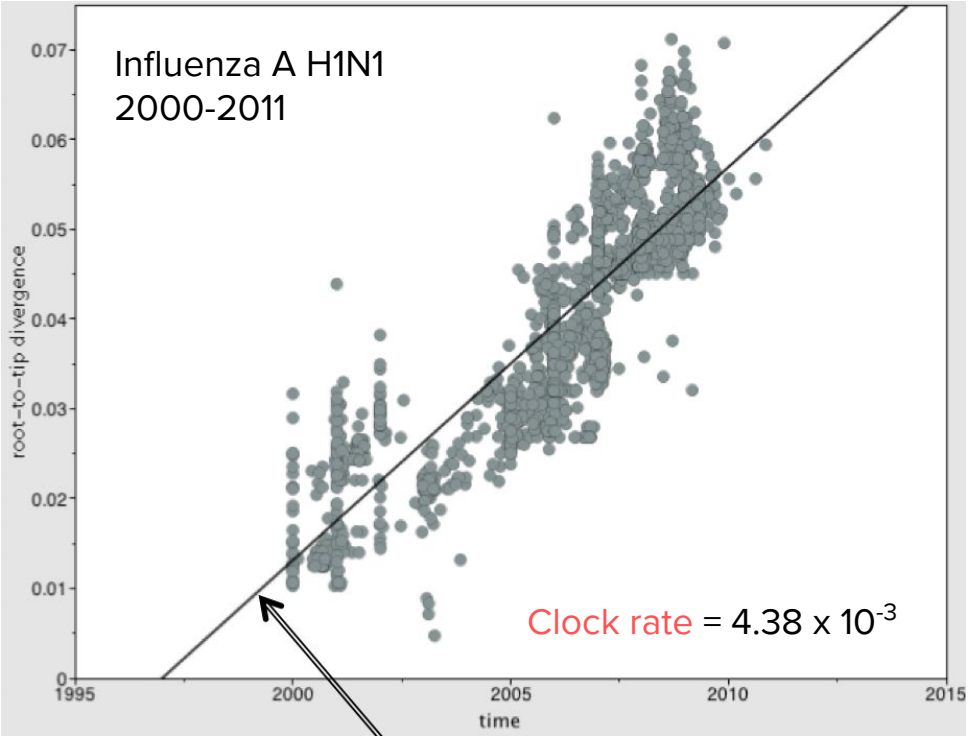
Root-to-tip regression

Recall that under a strict molecular clock model lineages will accumulate mutations at a constant rate such that the number of mutations between a tip and the root should increase linearly with time.



Root-to-tip regression

Thus if we regress root-to-tip divergence (in number of mutations) against sampling times, the slope of the resulting regression line will give us a rough estimate of the molecular clock rate.



Least-squares dating

Least-squares dating methods like LSD (To *et al*., 2016) use root-to-tip regression to estimate clock rates and node heights (i.e. divergence times).

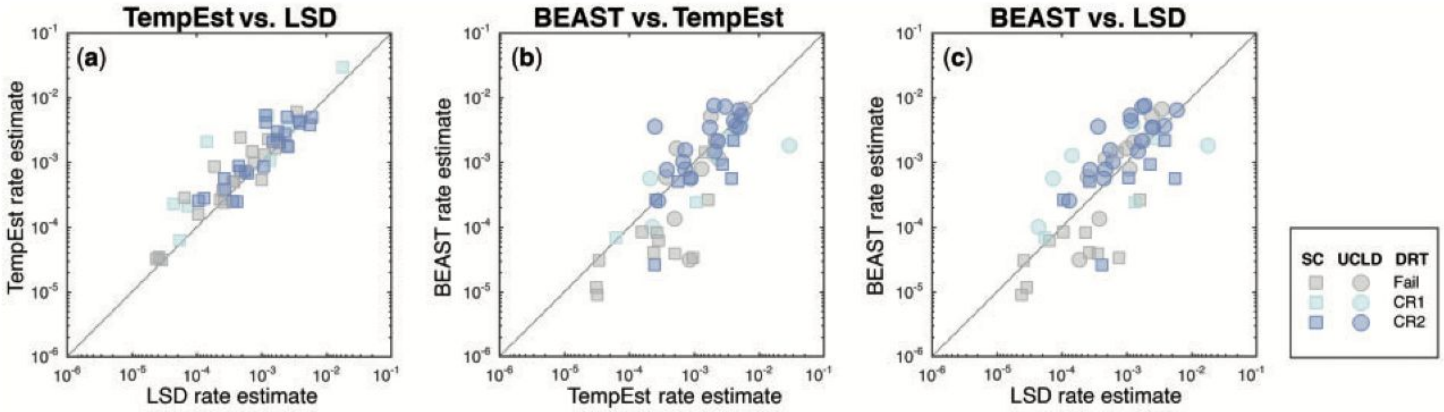
The node heights t_i and clock rate ω can be estimated together using least-squares optimization:

$$\phi(\omega, t_1, \ldots, t_{n-1}) = \sum_{i=2}^{2n-1} \frac{1}{\sigma_i^2} \left(b_i - \omega(t_i - t_{a(i)})\right)^2$$

This is very similar to standard least-squares regression: we try to minimize the deviation between the observed branch lengths b_i and the expected number of mutations $\omega(t_i - t_{a(i)})$ along that branch under the molecular clock assumption.

How do dating methods compare?

Regression-based methods like LSD generally return molecular clock estimates that are very similar to Bayesian dating methods.



Duchene *et al.* (Bioinformatics, 2016)

Other fast dating methods

TempEst: Simple GUI app for root-to-tip regression from the creators of BEAST (Rambaut *et al.*, 2016). See tutorial from week one for an example.

TimeTree: Iteratively optimizes branch lengths and ancestral sequences in a maximum likelihood framework. Available in the *TimeTree* Python package (Sagulenko *et al.*, 2018).

BactDating: Allows for Bayesian inference of divergence times from a fixed tree. Quantifies uncertainty in divergence times and partially accounts for recombination (Didelot *et al.*, 2018).

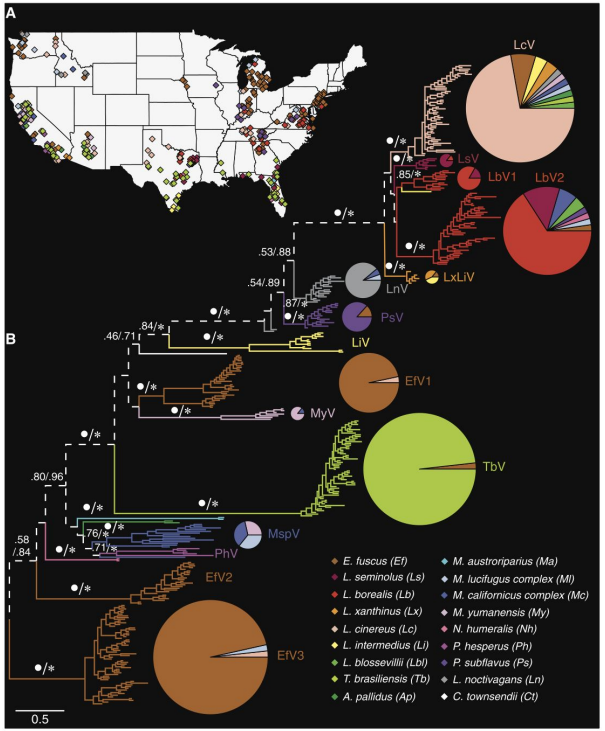
Faster ancestral state reconstruction

Ancestral state reconstruction

Ancestral state reconstruction is widely used to reconstruct changes in the state of lineages along a tree.

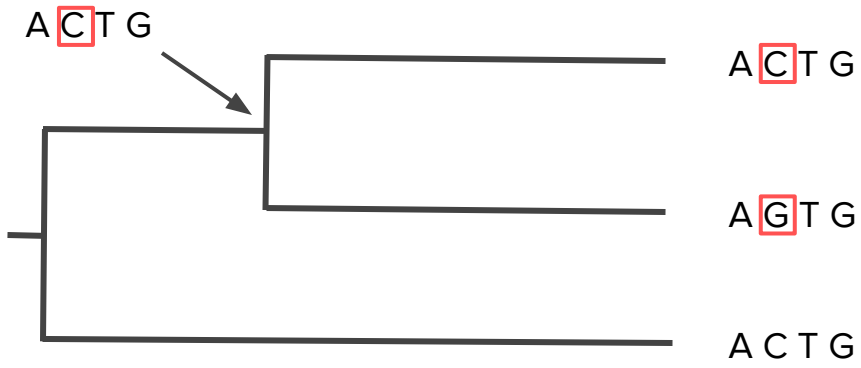
In phylogeography, ancestral states are often locations but a lineage's state could refer to its host population, genotype or even its entire sequence.

Given a fixed tree, maximum parsimony (MP) and maximum likelihood (ML) methods allow for very efficient reconstructions.



Streicker *et al.* (Science, 2010)

Likelihood of sequence data on trees



Maximum likelihood methods work very similar to how we computed the likelihood of sequence data given a tree using Markovian evolutionary models.

Modeling molecular evolution

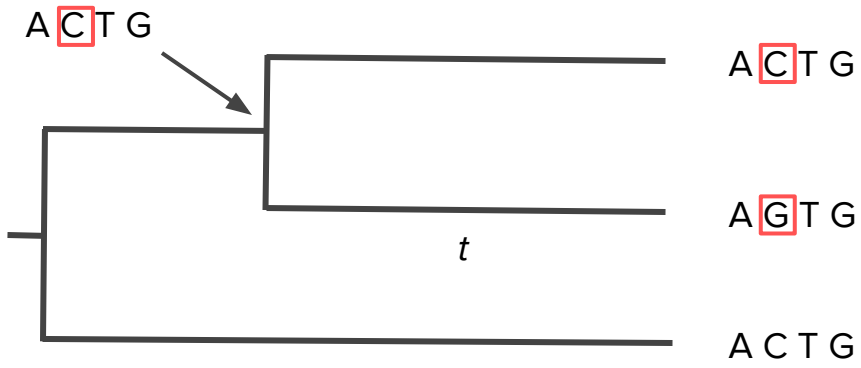
We can compute transition probabilities under a continuous-time Markov model given our substitution matrix ***Q*** and the time elapsed along a branch ***t***.

$$P(t) = e^{Qt}$$

The elements of *P(t)* give us the probability of every possible transition. Importantly, these **transition probabilities take into account every possible substitution path**.

$$P(t) = \begin{bmatrix} P_{T,T} & P_{T,C} & P_{T,A} & P_{T,G} \\ P_{C,T} & P_{C,C} & P_{C,A} & P_{C,G} \\ P_{A,T} & P_{A,C} & P_{A,A} & P_{A,G} \\ P_{G,T} & P_{G,C} & P_{G,A} & P_{G,G} \end{bmatrix}$$

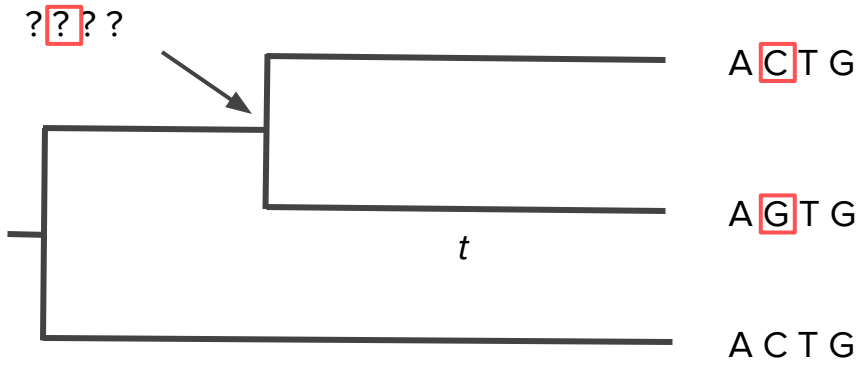
Computing likelihoods at one site



We can compute the likelihood of the sequence data given each possible ancestral state:

$$L(Seq|Tree) = P_{C,C}(t) * P_{C,G}(t)$$

Computing likelihoods at one site



Then find the ancestral state that maximizes the likelihood of the sequence data:

$$\text{ML state} = \underset{X \in \{A,C,T,G\}}{\text{argmax}} \left(P_{X,C}(t) * P_{X,G}(t) \right)$$

ML ancestral state reconstruction

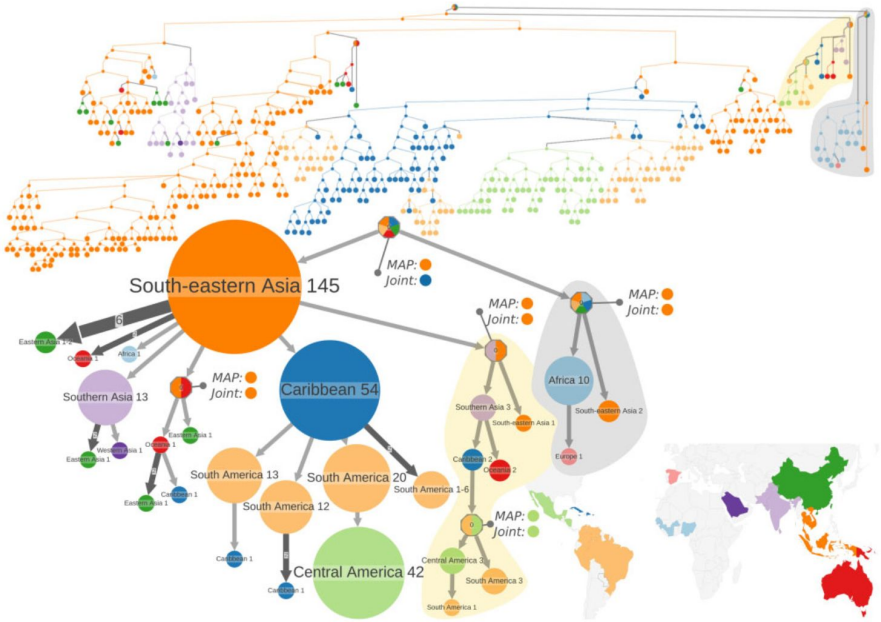
ML inference allows for state reconstruction under standard Markovian models of sequence or character evolution like the General-Time-Reversible (GTR) model that allow for transition rates to differ between states.

While ML inference does not allow for full quantification of uncertainty about all parameters, it allows for ancestral state probabilities to be computed using the relative likelihood of each state.

Methods like PastML (Ishikawa *et al*., 2019) allow for efficient inference of both transition rates and ancestral states.

Ancestral reconstructions with PastML

PastML allows for ancestral histories to be compressed into simpler graphs like in this example showing the global dissemination of dengue virus serotype 2.



Ishikawa *et al.* (MBE, 2019)

MP ancestral state reconstruction

Maximum parsimony reconstructions assume a “minimal evolution” model that minimizes the number of state changes required to explain the observed states at the tips of the tree.

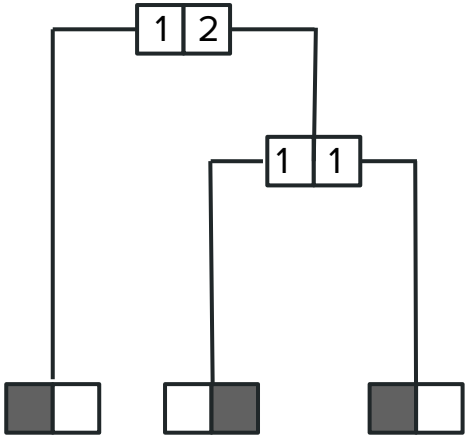
MP works well when transition rates are low such that the probability of multiple state changes along a lineage is low. In this case, the most likely ancestral state mapping will also be the most parsimonious.

But with more rapid character evolution, the most parsimonious history may not be the most likely as it may be unlikely that only a single state change has occurred.

MP algorithms

MP allows for ultra-fast reconstructions using clever dynamic programming methods. Generally these involve just two steps:

A ***post-order*** (tip-to-root) traversal: The number of changes required to explain the tip data subtending each internal node are computed

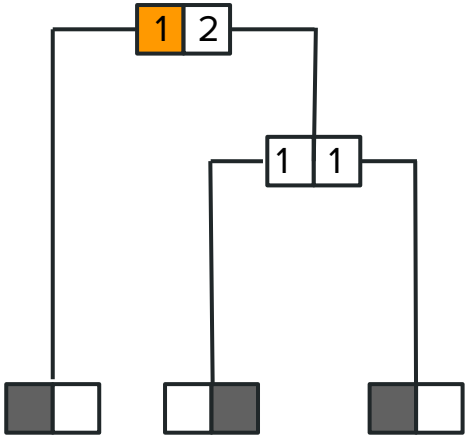


MP algorithms

MP allows for ultra-fast reconstructions using clever dynamic programming methods. Generally these involve just two steps:

A ***post-order*** (tip-to-root) traversal: The number of changes required to explain the tip data subtending each internal node are computed

A ***pre-order*** (root-to-tip) traversal: An ancestral node is chosen at each internal node that minimizes the number of changes required above and below the node.

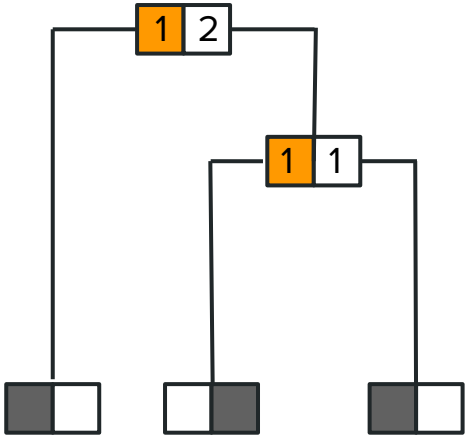


MP algorithms

MP allows for ultra-fast reconstructions using clever dynamic programming methods. Generally these involve just two steps:

A ***post-order*** (tip-to-root) traversal: The number of changes required to explain the tip data subtending each internal node are computed

A ***pre-order*** (root-to-tip) traversal: An ancestral node is chosen at each internal node that minimizes the number of changes required above and below the node.



MP ancestral state reconstruction

The main advantage of MP reconstructions is speed. There are no parameters to estimate and reconstructions generally take less than a second even for trees with thousands of tips.

Thus ultra-fast placement methods like UShER also use MP to place additional samples on existing trees.

Popular phylogenetic software packages like MEGA and Mesquite implement MP reconstructions.

Learning to let go: sub- and re-sampling strategies

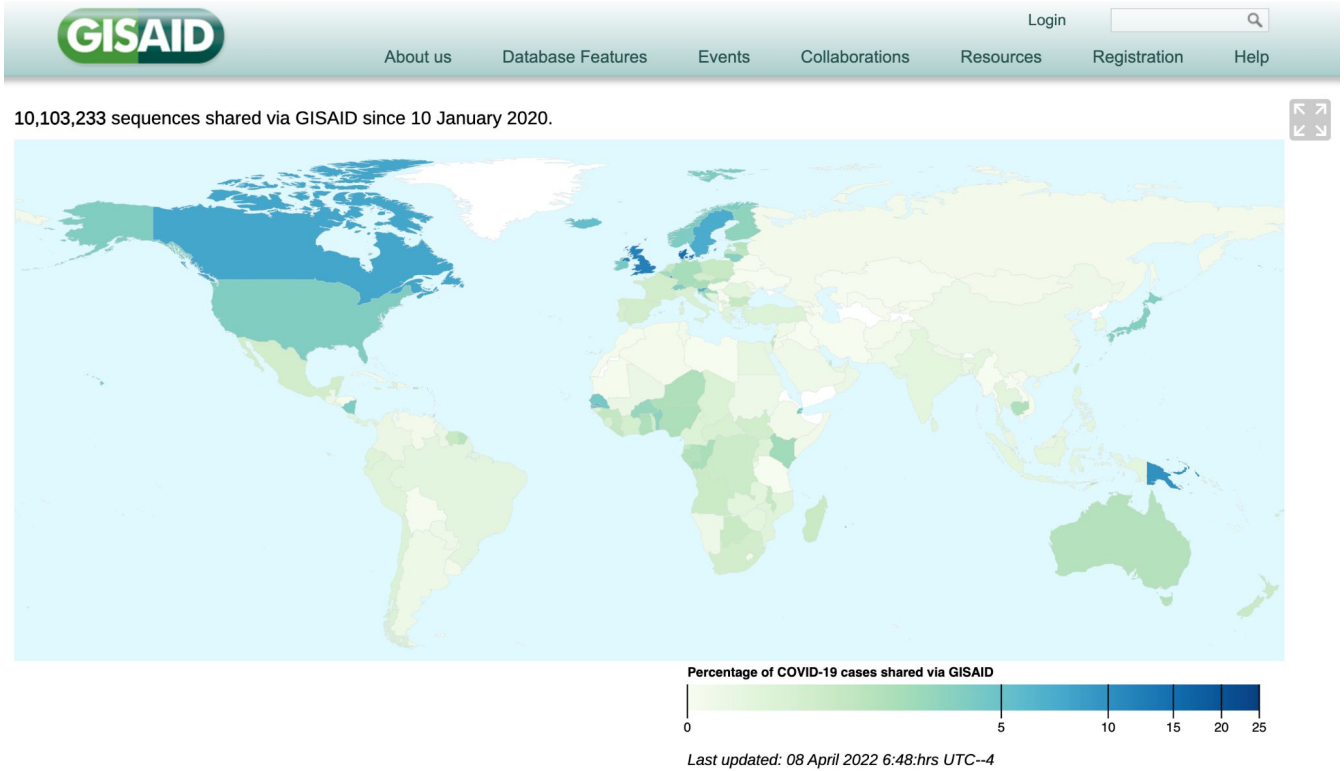
Subsampling

It’s important to remember that it is often possible to answer the questions of interest to us with much smaller datasets than the total available data.

Subsampling strategies allow us to down-sample data sets to more manageable sizes while correcting for biased sampling in the original dataset.

Resampling strategies: allow us to explore how sampling choices may be influencing our inferences.

Global bias in SAR-CoV-2 sampling



Reconstructing CoV imports and exports

RESEARCH ARTICLE

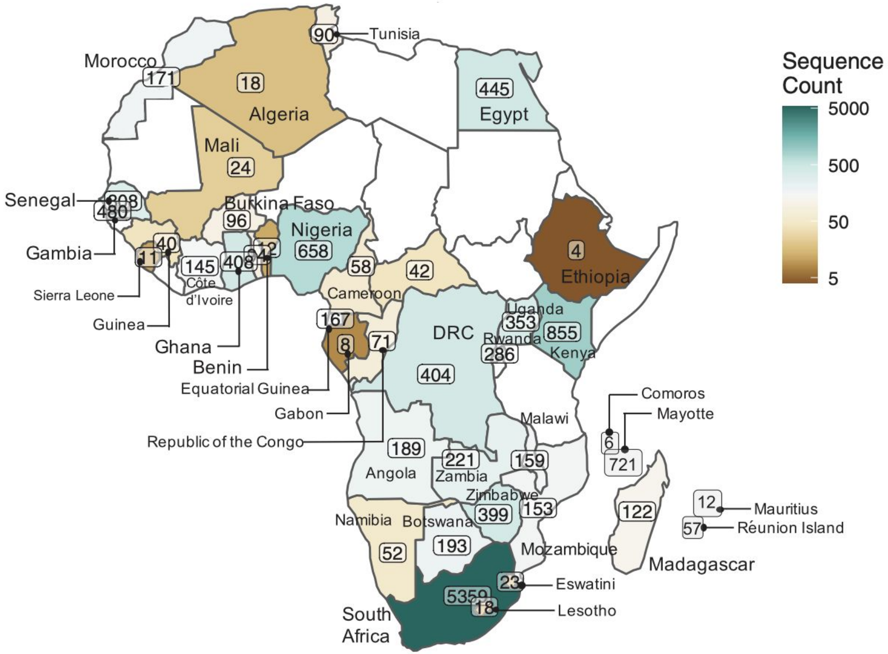
CORONAVIRUS

A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa

Eduan Wilkinson[1,2]†, Marta Giovanetti[3,4]†, Houriiyah Tegally[1]†, James E. San[1]†, Richard Lessells[1], Diego Cuadros[5], Darren P. Martin[6,7], David A. Rasmussen[8,9], Abdel-Rahman N. Zekri[10], Abdoul K. Sangare[11], Abdoul-Salam Ouedraogo[12], Abdul K. Sesay[13], Abechi Priscilla[14], Adedotun-Sulaiman Kemi[14], Adewunmi M. Olubusuyi[15], Adeyemi O. O. Oluwapelumi[16], Adnène Hammami[17], Adrienne A. Amuri[18,19], Ahmad Sayed[20], Ahmed E. O. Ouma[21], Aida Elargoubi[22,23], Nnennaya A. Ajayi[24], Ajogbasile F. Victoria[14], Akano Kazeem[14], Akpede George[25], Alexander J. Trotter[26], Ali A. Yahaya[27], Alpha K. Keita[28,29], Amadou Diallo[30], Amadou Kone[31], Amal Souissi[32], Amel Chtourou[17], Ana V. Gutierrez[26], Andrew J. Page[26], Anika Vinze[33], Arash Iranzadeh[6,7], Arnold Lambisia[34], Arshad Ismail[35], Audu Rosemary[36], Augustina Sylverken[37], Ayoade Femi[14], Azeddine Ibrahimi[38], Baba Marycelin[39], Bamidele S. Oderinde[39], Bankole Bolajoko[14], Beatrice Dhaala[40], Belinda L. Herring[27], Berthe-Marie Njanpop-Lafourcade[27], Bronwyn Kleinhans[41], Bronwyn McInnis[10], Bryan Tegomoh[42], Cara Brook[43,44], Catherine B. Pratt[45], Cathrine Scheepers[35,46], Chantal G. Akoua-Koffi[47], Charles N. Agoti[34,48], Christophe Peyrefitte[30], Claudia Daubenberger[49], Collins M. Morang'a[50], D. James Nokes[34,51], Daniel G. Amoako[35], Daniel L. Bugembe[40], Danny Park[33], David Baker[26], Deelan Doolabh[7], Deogratius Ssemwanga[40,52], Derek Tshiabuila[1], Diarra Bassirou[30], Dominic S. Y. Amuzu[50], Dominique Goedhals[53], Donwilliams O. Omuoyo[34], Dorcas Maruapula[54], Ebenezer Foster-Nyarko[26], Eddy K. Lusamaki[18,19], Edgar Simulundu[55], Edidah M. Ong'era[34], Edith N. Ngabana[18,19], Edwin Shumba[56], Elmostafa El Fahime[57], Emmanuel Lokilo[18], Enatha Mukantwari[58], Eromon Philomena[14], Essia Belarbi[59], Etienne Simon-Loriere[60], Etilé A. Anoh[47], Fabian Leendertz[59], Faida Ajili[61], Fakayode O. Enoch[62], Fares Wasfi[63], Fatma Abdelmoula[32,64], Fausta S. Mosha[27], Faustinos T. Takawira[65], Fawzi Derrar[66], Feriel Bouzid[32], Folarin Onikepe[14], Fowotade Adeola[67], Francisca M. Muyembe[18,19], Frank Tanser[68,69,70], Fred A. Dratibi[27], Gabriel K. Mbunsu[19], Gaetan Thilliez[26], Gemma L. Kay[26], George Githinji[34,71], Gert van Zyl[41,72], Gordon A. Awandare[50], Grit Schubert[59], Gugu P. Maphalala[73], Hafaliana C. Ranaivoson[44], Hajar Lemriss[74], Happi Anise[14], Haruka Abe[75], Hela H. Karray[17], Hellen Nansumba[76], Hesham A. Elgahzaly[77], Hlanai Gumbo[65], Ibtihel Smeti[32], Ikhlas B. Ayed[32], Ikponmwosa Odia[25], Ilhem Boutiba Ben Boubaker[78,79], Imed Gaaloul[22], Inbal Gazy[80], Innocent Mudau[7], Isaac Ssewanyana[76], Iyaloo Konstantinus[81], Jean B. Lekana-Douk[82], Jean-Claude C. Makangara[18,19], Jean-Jacques M. Tamfum[18,19], Jean-Michel Heraud[30,44], Jeffrey G. Shaffer[83], Jennifer Giandhari[1], Jingjing Li[84], Jiro Yasuda[75], Joana Q. Mends[85], Jocelyn Kiconco[52], John M. Morobe[34], John O. Gyapong[85], Johnson C. Okolie[14], John T. Kayiwa[40], Johnathan A. Edwards[68,86], Jones Gyamfi[85], Jouali Farah[87], Joweria Nakaseegu[52], Joyce M. Ngoi[50], Joyce Namulondo[52], Julia C. Andeko[82], Julius J. Lutwama[40], Justin O'Grady[26], Katherine Siddle[33], Kayode T. Adeyemi[14], Kefentse A. Tumedi[88], Khadija M. Said[34], Kim Hae-Young[89], Kwabena O. Duedu[85], Lahcen Belyamani[38], Lamia Fki-Berrajah[17], Lavanya Singh[1], Leonardo de O. Martins[26], Lynn Tyers[7], Magalutcheemee Ramuth[91], Maha Mastouri[22,23], Mahjoub Aouni[22], Mahmoud el Hefnawi[92], Maitshwarelo I. Matsheka[88], Malebogo Kebabonye[93], Mamadou Diop[30], Manel Turki[32], Marietou Paye[33], Martin M. Nyaga[94], Mathabo Mareka[95], Matoke-Muhia Damaris[96], Maureen W. Mburu[34], Maximillian Mpina[49,97,98], Mba Nwando[99], Michael Owusu[100], Michael R. Wiley[45], Mirabeau T. Youtchou[101], Mitoha O. Ayekaba[97], Mohamed Abouelhoda[102,103], Mohamed G. Seadawy[104], Mohamed K. Khalifa[20], Mooko Sekhele[95], Mouna Ouadghiri[38], Moussa M. Diagne[30], Mulenga Mwenda[105], Mushal Allam[35], My V. T. Phan[40], Nabil Abid[79,106], Nadia Touil[107], Nadine Rujeni[108,109], Najla Kharrat[32], Nalia Ismael[110], Ndongo Dia[30], Nedio Mabunda[110], Nei-yuan Hsiao[7,111], Nelson B. Silochi[97], Ngoy Nsenga[27], Nicksy Gumede[27], Nicola Mulder[112], Nnaemeka Ndodo[99], Norosoa H Razanajatovo[44], Nosamiefan Iguosadolo[14], Oguzie Judith[14], Ojide C. Kingsley[113], Okogbenin Sylvanus[25], Okokhere Peter[25], Oladiji Femi[114], Olawoye Idowu[14], Olumade Testimony[14], Omoruyi E. Chukwuma[67], Onwe E. Ogah[115], Chika K. Onwuamah[36,138], Oshomah Cyril[25], Ousmane Faye[30], Oyewale Tomori[14], Pascale Ondoa[56], Patrice Combe[116], Patrick Semanda[76], Paul E. Oluniyi[14], Paulo Arnaldo[110], Peter K. Quashie[50], Philippe Dussart[44], Phillip A. Bester[53], Placide K. Mbala[18,19], Reuben Ayivor-Djanie[85], Richard Njouom[117], Richard O. Phillips[118], Richmond Gorman[118], Robert A. Kingsley[26], Rosina A. A. Carr[85], Saâd El Kabbaj[119], Saba Gargouri[17], Saber Masmoudi[32], Safietou Sankhe[30], Salako B. Lawal[36], Samar Kassim[77], Sameh Trabelsi[120], Samar Metha[33], Sami Kammoun[121], Sanaâ Lemriss[122], Sara H. A. Agwa[77], Sébastien Calvignac-Spencer[59], Stephen F. Schaffner[33], Seydou Doumbia[31], Sheila M. Mandanda[18,19], Sherihane Aryeetey[123], Shymaa S. Ahmed[123], Siham Elhamoumi[33], Soafy Andriamandimby[44], Sobajo Tope[14], Sonia Lekana-Douki[82], Sophie Prosolek[26], Soumeya Ouangraoua[124,125], Steve A. Mundeke[18,19], Steven Rudder[26], Sumir Panji[112], Sureshnee Pillay[1], Susan Engelbrecht[41,72], Susan Nabadda[76], Sylvie Behillil[126], Sylvie L. Budiaki[95], Sylvie van der Werf[126], Tapfumanei Mashe[65], Tarik Aanniz[38], Thabo Mohale[35], Thanh Le-Viet[26], Tobias Schindler[49,97], Ugochukwu J. Anyaneji[1], Ugwu Chinedu[14], Upasana Ramphal[1,69,127], Uwanibe Jessica[14], Uwem George[14], Vagner Fonseca[1,4,128], Vincent Enouf[126], Vivianne Gorova[129,130], Wael H. Roshdy[123], William K. Ampofo[50], Wolfgang Preiser[41,72], Wonderful T. Choga[54,131], Yaw Bediako[50], Yeshnee Naidoo[1], Yvan Butera[108,132,133], Zaydah R. de Laurent[34], Amadou A. Sall[30], Ahmed Rebai[32], Anne von Gottberg[35,139], Bourema Kouriba[12], Carolyn Williamson[7,69,111], Daniel J. Bridges[105], Ihekweazu Chikwe[99], Jinal N. Bhiman[35,139], Madisa Mine[134], Matthew Cotten[40,135], Sikhulile Moyo[54,136], Simani Gaseitsiwe[54,136], Ngonda Saasa[55], Pardis C. Sabeti[33], Pontiano Kaleebu[40], Yenew K. Tebeje[21], Sofonias K. Tessema[21], Christian Happi[14], John Nkengasong[21], Tulio de Oliveira[1,2,69,137]*

The progression of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic in Africa has so far been heterogeneous, and the full impact is not yet well understood. In this study, we describe the genomic epidemiology using a dataset of 8746 genomes from 33 African countries and two overseas territories. We show that the epidemics in most countries were initiated by importations predominantly from Europe, which diminished after the early introduction of international travel restrictions. As the pandemic progressed, ongoing transmission in many countries and increasing mobility led to the emergence and spread within the continent of many variants of concern and interest, such as B.1.351, B.1.525, A.23.1, and C.1.1. Although distorted by low sampling numbers and blind spots, the findings highlight that Africa must not be left behind in the global pandemic response, otherwise it could become a source for new variants.

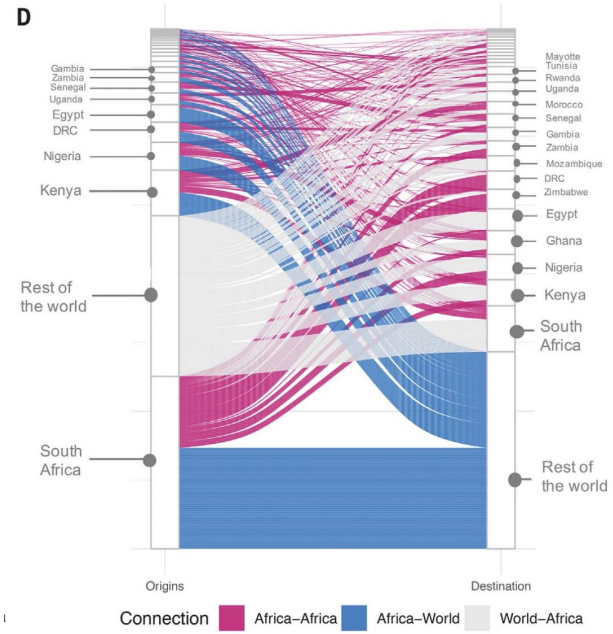
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019 in Wuhan, China (*1*, *2*). Since then, the virus has spread to all corners of the world, causing almost 150 million cases of COVID-19 and more than 3 million deaths by the end of April 2021. Throughout the pandemic, it has been noted that Africa accounts for a relatively low proportion of reported cases and deaths—by the end of April 2021, there had been ~4.5 million cases and ~120,000 deaths on the continent, corresponding to less than 4% of the global burden. However, emerging data from seroprevalence surveys and autopsy studies in some African countries suggest that the true number of infections and deaths may be severalfold higher than reported (*3*, *4*). In addition, a recent analysis has shown that in

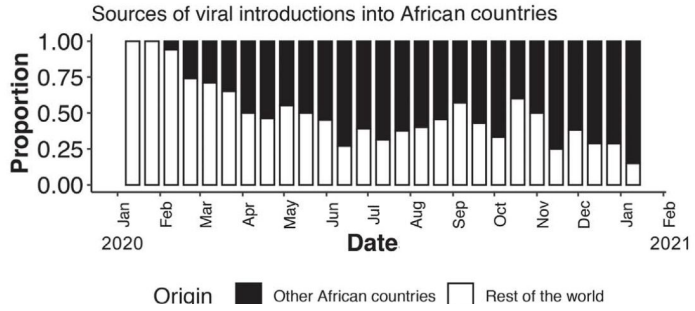


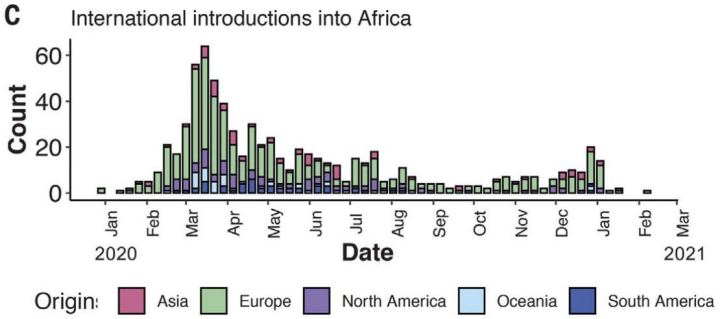
Wilkinson *et al.* (Science, 2021)

Reconstructing CoV imports and exports

A large number of imports were identified early in the epidemic, mostly form Europe.







Wilkinson *et al.* (Science, 2021)

Reconstructing CoV imports and exports

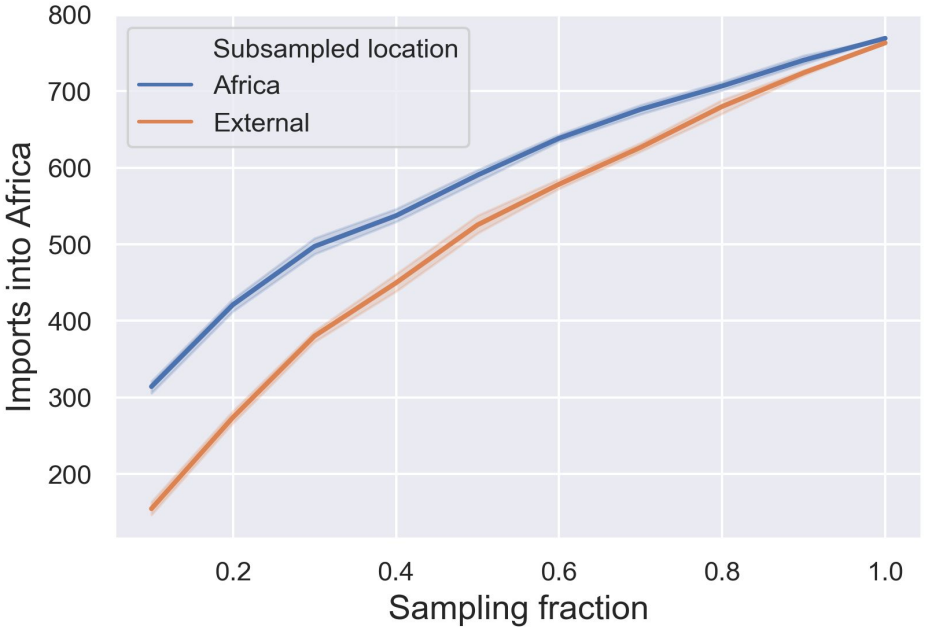
However, the number of imports/exports identified will almost certainly depend on how pathogen genomes were sampled in and outside of Africa.

We therefore performed a rarefaction analysis where we systematically varied the fraction of sampled genomes from either Africa or the rest of the world.

Basic strategy:

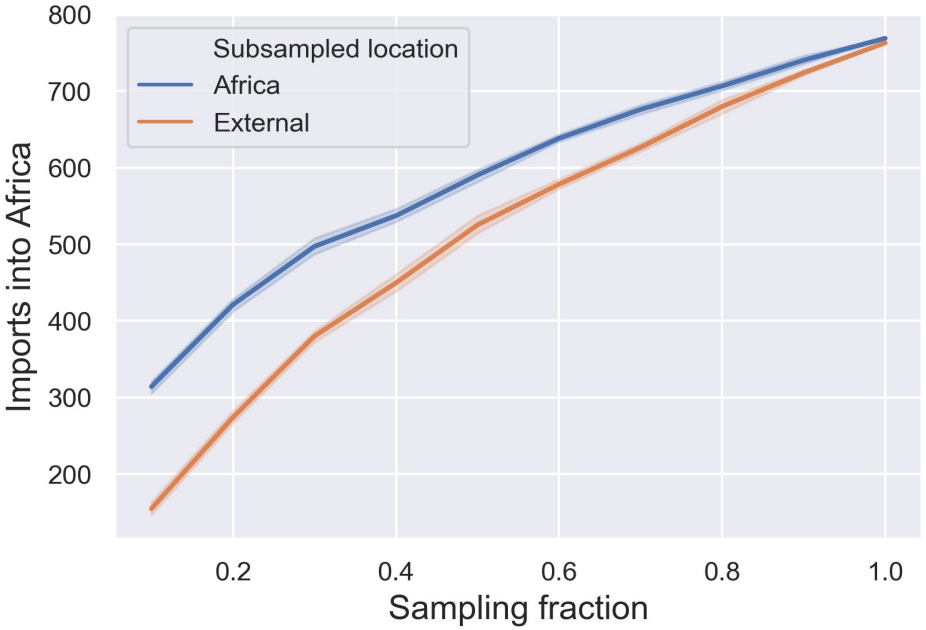
1. Build ML tree for full data set
2. Subsample a fraction *x* of genomes for a given location
3. Reconstruct ancestral locations using MP based on new subsampled tree
4. Count number of imports/exports identified at sampling fraction *x*

Reconstructing CoV imports and exports



Wilkinson *et al.* (Science, 2021)

Reconstructing CoV imports and exports



Vastly more introductions would have been identified with increased sampling in Africa or globally, suggesting that the intros identified are really just the "*ears of the hippo*".

Wilkinson *et al.* (Science, 2021)

Final thoughts

After the data deluge

EVARISTO SA/AFP VIA GETTY

Soldiers disinfecting parts of Brasilia's underground rail network as the coronavirus spread throughout Brazil in late March 2020.

Want to track pandemic variants faster? Fix the bioinformatics bottleneck

Emma B. Hodcroft, Nicola De Maio, Rob Lanfear, Duncan R. MacCannell, Bui Quang Minh, Heiko A. Schmidt, Alexandros Stamatakis, Nick Goldman & Christophe Dessimoz

https://www.nature.com/articles/d41586-021-00525-x

Final thougts

Maximum likelihood and parsimony methods allow us to efficiently explore much larger datasets than currently possible with Bayesian methods.

More efficient tree building, dating, ancestral state reconstruction and subsampling strategies can be combined into efficient workflows for working with massive datasets.

Regardless of how big your dataset is, the methods we discussed allow for easier and more efficient data exploration strategies that can be run upstream of more computationally intensive analyses.