Evolutionary trajectory and origin of SARS-CoV-2 variant

Anyou Wang (✉ anyou.wang@alumni.ucr.edu)

University of Memphis <https://orcid.org/0000-0002-4981-3606>

Article

Keywords: SARS-CoV-2, COVID-19, variant, evolutionary trajectory, Fréchet distance, artificial recurrent neural network

Posted Date: October 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1021653/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Evolutionary trajectory and origin of SARS-CoV-2 variant

Anyou Wang ORCID 0000-0002-4981-3606

Feinstone Center for Genomic Research, University of Memphis, Memphis, TN 38152, USA

Contact:

anyou.wang@alumni.ucr.edu

Project website:

<https://combai.org/ai/covidvariant/>

Abstract: Alignment-based quality attributes have identified several variants of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) that causes COVID-19 pandemic, but they face challenges to provide dynamic trajectory and origin of these variants. This study employs an alignment-free approach combining Fréchet distance (Fr) and artificial recurrent neural network to reveal evolutionary trajectory and origin of SARS-CoV-2 variant. Fr generates a distance matrix of 84 genome features and more than one million of genome sequences. Recurrent neural networks use this Fr matrix to quantitatively identify variants and reveal the evolutionary trajectory and origin of SARS-CoV-2. Total 34 SARS-CoV-2 variants have been identified. All these variants dynamically delete their genome during evolution, but their trajectory and deletion degree varies with individual variants, which can be classified into 3 groups, slight mutation group (13 members), middle level deletion (17 members), and high deletion (4 members). The slight deletion group works like wild type and its trajectory waves only slightly and temporarily, which has very low infection capacity. The high deletion group fluctuates with a rough trajectory characterized as a large loss and it also infects humans lightly. The middle deletion group gradually deletes their genome with a certain rhythm trajectory, corresponding to the pandemic peaks. This group causes most of the global COVID-19 cases. At least 3 mink coronavirus variants pose 56 genome features similar to SARS-CoV-2 and they are predicted to be able to infect human, and thus mink is the most likely origin of SARS-CoV-2, and the origin path follows this order: mink, cat, tiger, mouse, hamster, dog, lion, gorilla, leopard, bat, and pangolin. Therefore, this mink-origin SARS-CoV-2 evolves with a gradual deletion rhythm to infect humans

SARS-CoV-2; COVID-19; variant; evolutionary trajectory; Fréchet distance; artificial recurrent neural network

Introduction

Since 2019 when SARS-CoV-2 emerged, its origin and evolutionary trajectory have been intensely investigated¹⁻¹⁴ because understanding its origin and evolutionary trajectory helps not only to combat the pandemic but also to prevent future pandemic.

Current studies on the origin and evolutionary trajectory depend on the traditional alignment-based methods. These studies have identified variants on the basis of mutation characteristics and have provided useful information to understand the static states of virus mutations^{4,15-20}. However, SARS-CoV-2 have accumulated more than 50,000 mutations²¹ and a variant (e.g alpha) usually carries many mutations. It is biased to discriminate variants based on one or two mutation signatures although much effort has been paid to select variant signatures^{22,23}. More importantly, virus variants keep mutating during their dynamical evolution. A certain number of signatures might disappear at a time point but come back again later for a variant. Therefore, it is challenging for these static state approaches to provide a clear picture for SARS-CoV-2 origin and evolution trajectory.

Recently I created a quantitative system using Fréchet distance²⁴ (Fr) to compute distance between a variant and reference genome decomposing into 84 genome features (4 single nucleotides, 16 dinucleotides and 64 codons)²⁵. Combining Fr and artificial neural network has provided a clear picture for SARS-CoV-2 genome evolution²⁵. Here I expanded this system to investigate SARS-CoV-2 variant evolution trajectory and origin.

Results

Variant identification

This study identified variant clusters(referred as variants) by two steps, UMAP²⁶ pre-classification and confirmation by long short-term memory²⁷ (LSTM, an artificial recurrent neural network architecture) (Figure 1A). All genome sequences used here were the same as that of my recent study²⁵, including 1,128,954 genome sequences filtered out from 2,212,864 that contained 12 officially defined variants^{18,19}. Most of sequences were identified as alpha (973995 samples) and delta (116104 samples) deposited on July 4, 2021 in GISAID database. The study also employed the quantitative matrix pre-created for studying genome evolution²⁵, which contains Frs for 1,128,954 genome sequences and 84 genome features, including 4 single nucleotides, 16 dinucleotides and 64 codons (materials and methods). This present study split these 1,128,954 samples into weekly chunks following the time-series order from December, 2019 to July 4, 2021. These weekly chunk data were used to identify variants.

For explaining the algorithm to identify variants, I described below the detail in identifying variant 0 as an example. A variant is defined here as that it must have at least 50 members in the pre-classification step. To have enough members in a variant, I used first 8 week data to pre-classify variant 0 to variant 4 by UMAP (Figure 1B, materials and methods). These pre-classified members of variant 0 were subjected to further confirmation via using LSTM, in which a LSTM model was built with 4 layers (materials and methods) and members in variant 0 were used as train-set and members of the rest of variants (variant 1-4) as test-set. The mean absolute error(MAE) was calculated for the training set (variant 0 in this case). MAE measures distance between two groups (e.g. actual value and prediction here) and it can infer outliers, which have large MAE. In this study, MAE threshold was set as $\text{mean} + 1.5 \text{ standard deviation}$. MAE distribution for variant 0 was plotted (Figure 1C) and all samples within the threshold were finally classified as variant 0 members and a few samples outside the threshold were

treated as outliers, which were put back to the sample pool waiting for the next cycle UMAP(Figure 1D) .

This model trained by variant 0 was used to search its full members from the rest data pool (from week #9 to July 4, 2021.) week by week by using weekly chunk data.

Similarly, variant 1-4 members were identified. The 84 feature Frs of these 5 variants showed that variant 0 was closed to reference sequence with less Frs changes (Figure 1E), but variant 2-4 had undergone a series of mutations. Variant 4 had already mutated most of its 84 features(Figure 1E) within 8 week's data. It was unbelievably fast. This indicated that SARS-CoV-2 actually rapidly mutated, and also suggested that SARS-CoV-2 actually stayed around the human community for a long time before emerging.

After these variant 0-4 identification, searching for other groups of variants was performed by moving the week window to the next and finally all 34 variants were identified (Figure 1F). From variant 3, all variants identified here contained mix members of existing official variants(Figure 1F). For example, variant 4 already contained members of alpha, gamma, delta and others(Figure 1F). This indicates that this variant 4 already has a mix of multiple features (mutations) of current variants.

Evolutionary trajectories of variants

To understand the evolutionary trajectories of 34 variants, I employed the pre-built LSTM model for investigating the evolutionary trajectories of SARS-CoV-2 genome²⁵ and only changed the input for training and testing with variant data. To diminish noise and be robust, the LSTM model was used to predict the evolutionary trajectory and to forecast the near future, 30 days after July 4, 2021. The

predicted median of all 84 feature Frs was used here to explain the results. Positive Fr represents gain/insertion and negative as deletion/loss.

Globally, the Fr median of all 34 variants went down(Figure 2A), indicating SARS-CoV-2 underwent deletion during evolution. Three waves of deletions were observed in the global trajectory, one from 03/2020 to 05/2020 (the biggest loss with Fr from -50 to -180), another in 08/2020, and final one beginning in 12/2020. The final one went smoothly but it stayed for a very long time and still keeps going down in the near future as forecasted (Figure 2A).

As examples, three individual variants (variant 13, 23, and 24) also showed deletions (Figure 2B-C, complete plots was shown in this project website <https://combai.org/ai/covidvariant/>), but their trajectories were different. Among these three variants, variant 23 posed the biggest deletion in 1/2021 (Fr = -650) and then its deletion recovered a little after 03/2021 but its Fr was always below -550, indicating that this variant 23 holds the biggest deletion during entire evolution. In contrast, variant 13 only suffered the middle level of deletion (the max Fr > -300), and it had three waves of deletion, respectively beginning in 06/2020, 1/2021, and 04/2021(Figure 2B). These deletion waves were corresponding to three waves of global virus outbreaks. Variant 24 was a flexible variant and it mutated frequently, but it kept a stable middle level of deletion(Fr < -300) after 04/2021.

To outlook at the level of mutation over all 34 variants, this study plotted the median of Fr for all 34 variants(Figure 2E) and found that variant 23 did pose the biggest deletion among all variants with Fr near -600. The variant 13 and 24 only underwent a middle level of mutation. Variant 0 and 1 almost behaved like wild type with (Fr = ~0), without any observable deletion.

Infection trajectory of SARS-CoV-2 variant

With the waving trajectory of SARS-CoV-2 genome evolution as shown above, it is expected that the infection capacity of a variant would change dynamically along the evolutionary trajectory of its genome. To appreciate the infection capacity of individual variants, this study used my pre-developed strategy and LSTM model²⁵ to predict infection cases for each variant(materials and methods). Briefly, this study first trained the pre-built LSTM model with global infection cases as a response and the multivariate (84 feature Frs) matrix as training matrix and predicted the global cases(**Figure 3A**). Interestingly, the prediction cases were higher than actual after 6/2021, indicating vaccination effect. This trained LSTM model with its parameters was kept and then I used the median of 84 feature Frs for each variant (e.g. variant 13) to feed this trained LSTM model to predict infection cases for this variant(e.g. variant 13).

The above three variants, variant 13, 24, and 23, were also used here as examples to demonstrate the predicted infection trajectory(Figure B-D, full plots shown in project website <https://combai.org/ai/covidvariant/>). The number of genome sequences for each variant was also plotted with prediction infection cases to examine if my prediction was corresponding to the sequence number. Of course, a variant infection capacity was dependent on genome composition (feature Fr in this case) instead of its sequences deposited in the database. However, if samples collected for sequencing were random enough, the sequence number could reflect a certain degree of infection cases. As expected, variant 13, 24 and 23 had 101535, 200524 and 3347 sequences respectively and (Figure 1F) and they were predicted to infect a maximum 550k, 600k, and 140k cases a day (Figure B-D). As they evolved in their genome, variants 13 and 24 increased their infection capacities along their evolutionary trajectories.

To understand the evolution across all variants, I need to find a way to compare their similarity. MAE is a metric because MAE actually calculates the distance between samples. A variant has a larger MAE, it has longer distance from wild type and it is less similar to wild type. I treated variant 0 as a wild type and used its Fr matrix as train-set and the Fr matrix of other 33 individual variants as test-set to compute the MAE of samples.

Plotting the median of MAE and prediction cases of each variant showed that variant 1 was closed to wild type as expected and variant 23 was the one with most mutations (Figure 4E). From the Loess regression curve derived from prediction cases, variants(e.g. 13) with middle level of mutation have more capacity to infect humans. Variants carrying low and high mutations had low infection capacity. Variant 24 also had the highest prediction cases, but it already over mutated so its infection capacity might go down soon if it continued to mutate. However, its genome was forecasted as stable in the near future (purple line in figure 2C), its infection capacity would be still high in the near future. The troublesome variant was variant 13, with up-hill infection capacity and a middle level of mutations. It could cause more infection cases in the near future.

SARS-CoV-2 origin path

To understand the origin path of SARS-CoV-2, I also calculated the MAE between human wild type and animal coronavirus samples from GISAID, but human samples were reversed in time-series order, from 2021 to 2019. As done before, the Fr matrix of variant 0 was used as train-set to fit the pre-built LSTM model and animal samples were treated as a test to calculate MAE for samples. Animal samples with lowest MAE were closed to human wild types. Ranking minimum MAE for all animals revealed that mink was very close to human wild type (MAE near 0, Figure 4A), followed in order by cat, tiger, mouse, hamster, dog, lion, gorilla, leopard, bat, and pangolin. This indicated that mink coronavirus had

the ability to infect humans directly and it was the most likely origin of SARS-CoV-2. In contrast, it is unlikely for a coronavirus from bat or pangolin to infect humans.

To understand how mink coronavirus was so close to SARS-CoV-2, I plotted the MAE between human and mink (Figure 4B) and found that several of the mink samples had mutated to be similar to human wild type. Moreover, these mink mutants had 56 consensus features (defined as ones with the same sign of positive and negative Fr, Figure 4C) and they had 25 features(out of 56) different from normal mink viruses(Figure 4D). These mink viruses actually shared 57% (32 out of 56) features with humans (Figure 4E). Except for 0, only 16% (9/56) was different between mink and human. Therefore, mink is predicted as the SARS-CoV-2 origin.

Discussion

This study employed LSTM and Fr to quantitatively identify 34 SARS-CoV-2 variants. Traditionally, variants have been identified on the basis of quality attributes, especially point mutation, leading to a bottleneck for understanding their dynamic evolution trajectories. In contrast, quantity attributes can be used for downstream modeling and pattern recognition and appreciating their evolutionary trajectories.

Evolutionary trajectory of SARS-CoV-2 have been intensely reviewed and discussed^{5,6}, but they have not been revealed. This study uncovered its trajectory, in which it overall continues to shorten its genomes during evolution. Individual variants have their own trajectories, but those with middle level of deletion have more infection capacity and these variants (e.g. variant 13 and 24) are the real source causing COVID-19 pandemic. The wild type SARS-CoV-2 and highly mutated variants have low infection capacity.

The wild type SARS-CoV-2 has less capacity to infect humans and the early samples already carry many mutations. It is unlikely for these viruses to mutate suddenly, suggesting that SARS-CoV-2 or its similar virus already stay inside the human community for a long time, like years before 2019 emerging.

The origin of SARS-CoV-2 have been a hot topic to debate and it has been widely reported^{6,8,11–13,28–30}. Recent researches have found that mink coronavirus can trans-infect human⁷. Consistently, this study also predicted mink coronavirus as the origin of SARS-CoV-2. Furthermore, this study further uncovered the origin path of SARS-CoV-2, a mystery hanging on for years, which follows this order: mink, cat, tiger, mouse, hamster, dog, lion, gorilla, leopard, bat, and pangolin. Therefore SARS-CoV-2 comes from our neighbors like mink, cat and mouse, instead of bat and pangolin as thought.

Materials and methods

Except the method for identifying variant, all other materials and methods were referred to my paralleled study²⁵

Briefly, this study directly used the matrix (1,128,954*84) containing Fréchet distance of 84 genome features (4 single nucleotides, 16 dinucleotides and 64 codons) for 1,128,954 genome sequences filtered from total 2,212,864 sequences, which was downloaded on July 4, 2021.

All computations were done under Linux with python 3.8, TensorFlow 2.4.0 and Scikit-learn 0.24.0.

Discrete Frechet distance (Fr)

This study computed the coupling Fr^{24} for each feature (e.g. “A”) for a given individual virus genome (e.g. EPI_ISL_601443, alpha variant) against the reference genome(NC_045512).

Given P and Q as a feature (e.g. “A”) trajectory for the reference and a variant respectively.

$P=\{p[1], p[2],...p[n]\}$, ordered positions and contents in reference genome

$Q=\{q[1], q[2],...q[m]\}$, ordered positions and contents in variant genome

A coupling L between P and Q is a sequence

$(P[a_1], Q[b_1]), (P[a_2], Q[b_2]), \dots, (P[a_l], Q[b_l])$

where $a_1 = 1, b_1 = 1, a_l = n, b_l = m$. For all $i = 1, \dots, l$, $a_{i+1} = a_i$ or $a_{i+1} = a_i + 1$, and $b_{i+1} = b_i$ or $b_{i+1} = b_i + 1$.

Fr between Q and P is defined as

$Fr(Q,P) = \min\{ \max \text{distance}(Q[b_i], P[a_i]) \text{ for all possible couplings between P and Q} \}$

Frs for all 84 features of more than one million virus were deposited in the project website

Variant identification

The UMAP²⁶ was used to pre-cluster variants and the variants were confirmed by using a LSTM model.

This LSTM model contained 4 hidden layers, the activation was set to 'relu', and adam was used as the optimizer. MAE was used for the loss. 150 and 20 was respectively set as epoches and batch size.

Validation split was set to 0.1.

Evolutionary trajectory and infection case prediction and forecast

This part used the same LSTM model as my paralleled study. Two independent machine-learning sets were prepared in this study, including test (5%) and training (95%) following the order of time series. Batch size and epochs were set to 64 and 50 for all machine learning. dropout (0.2) was set for two model layers for all models in this study. 4 hidden layers were set to a typical running. Adam was used for model optimizer and mean_squared_error was set for the loss.

The forecasts were set 30 days after July 4, 2021.

Final graphing

Several final summary were drawn by using ggplot2 in R. Otherwise, they were completed by python.

All detailed figures and data was deposited in the project website

<https://combai.org/ai/covidvariant/>

Conflict interests

NO

Funding

No funding resource associated with for this project

Acknowledgment

Thank GISAID (<https://www.gisaid.org>) for providing the full data.

Figure legends

Figure 1. SARS-CoV2 variant cluster identification. A, workflow of this study B, UMAP pre-classified 5 variants, including 0, 1, 2, 3 and 4. C, Loss-MAE distribution of a LSTM model for discriminating variant 0 and 1. The MAE mean+1.5 standard deviation was set as the threshold. D, Outlier detection on the basis of threshold. E, Heatmap of 5 variants. F, compositions of total 34 variants.

Figure 2. Evolutionary trajectory of variant genome. A, Evolutionary trajectory of all 34 variants. B, Evolutionary trajectory of variant 13. C, Evolutionary trajectory of variant 24. D, Evolutionary trajectory of variant 23. E, Ordering individual variants based on their Fr median.

Figure 3. Infection case prediction. A, Actual global infection cases and their predictions. B-D, Predicted infection cases and the number of sequences deposited in GISAID database for variant 13, 24, and 23 respectively. E, Max infection cases and MAE median of 34 variants. The MAE was calculated by using variant genome vs wilt type (variant 0). The blue smooth curve was drawn by loess regression of infection cases.

Figure 4. SARS-CoV-2 Origin path. A, species minimum MAE ranking. B, human and mink variant MAE distribution. C, top 56 consensus features of top 3 mink samples closed to humans. D, comparison of 56 features between top 3 mink samples and total mink samples. E, Comparison of 56 features between top 3 mink samples and top 5 human samples.

Reference:

1. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, (2021).
2. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).

3. Choe, H. & Farzan, M. How SARS-CoV-2 first adapted in humans. *Science* **372**, 466–467 (2021).
4. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).
5. Telenti, A. *et al.* After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596**, 495–504 (2021).
6. Singh, D. & Yi, S. V. On the origin and evolution of SARS-CoV-2. *Exp. Mol. Med.* **53**, 537–547 (2021).
7. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172–177 (2021).
8. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
9. Maxmen, A. & Mallapaty, S. The COVID lab-leak hypothesis: what scientists do and don't know. *Nature* (2021) doi:10.1038/d41586-021-01529-3.
10. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
11. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289 (2020).
12. Lytras, S., Xia, W., Hughes, J., Jiang, X. & Robertson, D. L. The animal origin of SARS-CoV-2. *Science* **373**, 968–970 (2021).
13. Cohen, J. The hunt for SARS-CoV-2's ancestors heats up. *Science* (2021) doi:10.1126/science.acx8985.
14. Pekar, J., Worobey, M., Moshiri, N., Scheffler, K. & Wertheim, J. O. Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412–417 (2021).
15. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
16. Callaway, E. Delta coronavirus variant: scientists brace for impact. *Nature* (2021) doi:10.1038/d41586-021-01696-3.

17. Scudellari, M. How the coronavirus infects cells — and why Delta is so dangerous. *Nature* **595**, 640–644 (2021).

18. Embrace the WHO's new naming system for coronavirus variants. *Nature* **594**, 149–149 (2021).
19. Konings, F. *et al.* SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat. Microbiol.* 1–3 (2021) doi:10.1038/s41564-021-00932-w.
20. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
21. Oude Munnink, B. B. *et al.* The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* **27**, 1518–1524 (2021).
22. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476.e6 (2021).
23. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122–127 (2021).
24. Eiter, T. & Mannila, H. Computing Discrete Fréchet Distance, https://openproceedings.org/2019/conf/edbt/EDBT19_paper_219.pdf. 8.
25. Wang, A. Evolutionary trajectory of SARS-CoV-2 genome. <https://doi.org/10.21203/rs.3.rs-1009010/v1>. (2021).
26. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.4 documentation. <https://umap-learn.readthedocs.io/en/latest/>.
27. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
28. Mallapaty, S. Animal source of the coronavirus continues to elude scientists. *Nature* (2020) doi:10.1038/d41586-020-01449-8.
29. Mallapaty, S. Did the coronavirus jump from animals to people twice? *Nature* **597**, 458–459 (2021).
30. Mink may be host of coronavirus' origin: China's 'Bat Woman' - Global Times. <https://www.globaltimes.cn/page/202101/1212942.shtml>.

Fig.1

F 34 variants and their compositions

Variant	member_number	Alpha	Gamma	Delta	unknown	Kappa	Theta	Beta	Epsilon	Zeta	Eta	Lambda	Iota
0	871	0	0	0	1	0	0	0	0	0	0	0	0
1	272	0	0	0	1	0	0	0	0	0	0	0	0
2	435	0	0	0	1	0	0	0	0	0	0	0	0
3	1564	0.00064	0	0	0.99808	0	0	0.000639	0	0	0	0	0.00064
4	38916	0.00046	0.00069	0.000103	0.99733	7.7E-05	0	5.14E-05	0.000257	0.000694	5.1393E-05	0	0.00028
5	23623	0.00093	0	8.47E-05	0.95602	0.00089	0	0.000593	0.000212	0.000339	0	0	0.04093
6	822	0	0	0	1	0	0	0	0	0	0	0	0
7	2477	0	0	0	0.9996	0	0	0	0.000404	0	0	0	0
8	3920	0	0	0	1	0	0	0	0	0	0	0	0
9	11767	0.00059	0	0	0.99813	0	0	0	8.5E-05	0	0	0	0.00119
10	148492	0.31758	6.7E-06	6.06E-05	0.66179	3.4E-05	0	0.000431	0.002916	0.000929	0.00086873	5.387E-05	0.01533
11	7867	0.00127	0	0.000127	0.98551	0	0	0.01144	0	0.001271	0	0	0.00038
12	799	0	0	0	1	0	0	0	0	0	0	0	0
13	101535	0.14942	0.00853	0.010381	0.64263	0.00718	2.95E-05	0.016684	0.017954	0.012498	0.00521987	0.0009947	0.12849
14	3415	0.24422	0.01347	0.000293	0.60791	0	0	0.02284	0.000586	0	0.01083455	0	0.09985
15	507	0	0	0	0.97633	0	0	0	0.021696	0.001972	0	0	0
16	3657	0.01531	0.0175	0.019962	0.44736	0.00547	0	0.003555	0.329505	0.003828	0	0.0013672	0.15614
17	495215	0.64872	0.00962	0.051493	0.26108	0.00124	3.43E-05	0.004174	0.007506	0.001737	0.00225559	0.0004422	0.0117
18	22664	0.01275	0.00106	0.013766	0.86304	0.00419	0	0.033886	0.048888	0.02087	4.4123E-05	0.0006177	0.00088
19	5002	0.03778	0.0112	0.008597	0.76389	0.0004	0	0.003798	0.097161	0.002199	0.0019992	0	0.07297
20	216	0	0	0	1	0	0	0	0	0	0	0	0
21	38730	0.54312	0.0181	0.0189	0.32342	0.00124	7.75E-05	0.012187	0.052388	0.002505	0.00560289	0.0017816	0.02068
22	5564	0.44123	0.14216	0.016894	0.37203	0	0	0.00647	0.010963	0.003954	0.00017973	0.0008986	0.00521
23	3347	0.82133	0.00418	0.067224	0.09352	0	0	0.001494	0.007768	0.000598	0.00029878	0	0.00359
24	200524	0.74046	0.06382	0.057838	0.11266	0.00046	1.99E-05	0.001696	0.018437	0.000618	0.00079292	0.0007131	0.00248
25	3681	0.56072	0.13366	0.006248	0.17903	0	0.000272	0.001358	0.07335	0.004618	0.01657158	0.0038033	0.02037
26	396	0	0	0	1	0	0	0	0	0	0	0	0
27	314	0.00637	0	0	0.98408	0	0	0	0.003185	0.003185	0	0.0031847	0
28	113	0	0	0	1	0	0	0	0	0	0	0	0
29	604	0	0	0.001656	0.63576	0.00166	0	0.004967	0.35596	0	0	0	0
30	894	0.01678	0.01454	0.011186	0.9094	0.00112	0.001119	0.016779	0.010067	0	0	0.0167785	0.00224
31	251	0	0	0	0.94024	0	0	0	0.059761	0	0	0	0
32	379	0.01583	0.05805	0.055409	0.76517	0	0	0.01847	0.044855	0	0	0.0395778	0.00264
33	116	0.27586	0.02586	0.017241	0.63793	0	0	0.025862	0.017241	0	0	0	0

Fig.2

Fig.3

Fig.4

A Origin path

B Variant MAE

C Top3 mink consensus

D Top3 vs total mink

E Top3mink vs top5 human