idCOV: a pipeline for quick clade identification of SARS-CoV-2 isolates

Xun Zhu^{1,†}, Ti-Cheng Chang^{1,†}, Richard Webby², and Gang Wu^{1,*}

¹ Center for Applied Bioinformatics, St. Jude Children's Research Hospital. 262 Danny Thomas Pl, Memphis, TN 38105

² Department of Infectious Disease, St. Jude Children's Research Hospital. 262 Danny Thomas Pl, Memphis, TN 38105

* To whom correspondence should be addressed.

† The authors wish it to be known that these authors contributed equally.

Abstract

idCOV is a phylogenetic pipeline for quickly identifying the clades of SARS-CoV-2 virus isolates from raw sequencing data based on a selected clade-defining marker list. Using a public dataset, we show that idCOV can make equivalent calls as annotated by Nextstrain.org on all three common clade systems using user uploaded FastQ files directly. Web and equivalent command-line interfaces are available. It can be deployed on any Linux environment, including personal computer, HPC and the cloud. The source code is available at <https://github.com/xz-stjude/idcov>. A documentation for installation can be found at <https://github.com/xz-stjude/idcov/blob/master/README.md>.

Introduction

The on-going Coronavirus disease 2019 (Covid-19) pandemic has resulted in over 734,000 deaths, affecting more than 188 countries and territories (CSSE, 2020; Dong, et al., 2020). The general impact of the disease calls for a quick response from the local health system at every stage of its transmission.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; previously known as 2019-nCoV) is the pathogenic cause of Covid-19 (Lescure, et al., 2020). Quickly identifying the genetic clade of the virus helps develop an understanding of the local transmission of the disease. The recent NYC study (Gonzalez-Reiche, et al., 2020), for example, identified crucial local transmission events, plurality of virus introductions, and the possible sources of these introductions. The classification of the infected virus can also provide more granular information to contact tracing and for epidemiological studies to understand disease severity and host genetic susceptibility to different lineages of SARS-CoV-2.

In order to quickly identify the clade of an isolate of SARS-Cov-2 given its sequencing FastQ files, we have developed a bioinformatics pipeline and a user-friendly web interface. An equivalent command-line interface is also supplemented for batch-processing many samples in a terminal environment. We introduce this system as *idCOV*.

Pipeline

Input

The pipeline (Figure 1) starts with the sequencing data derived from SARS-Cov-2 isolates in the format of pair-end FastQ files. These files are either available on the server or uploaded by the user through the web or the command-line interface. The fragments in the FastQ files are then aligned to a reference genome.

Alignment

Normally a virus reference genome is used for alignment. However, since sequencing library preparations might be generated from clinical material, some of the sequencing reads in the FastQ files may be from the human host. To account for this, we aggregated the Wuhan-Hu-1 genome (Wu, et al., 2020) with the GRCh38 human genome and used the combined genome as the reference for the alignment. The alignment is done using BWA, version 0.7.17 (Li and Durbin, 2009). The “bwa mem” algorithm is used by default. After the alignment, the result BAM files are first sorted by the fragment locations and then in-

dexed, both using the SAMtools software, version 1.10 (Li, et al., 2009). This post-processing is to improve the performance of subsequent steps.

Mutations Calculation

The virus reference genome (Wuhan-Hu-1) is seen as the original genome and any deviation from it is seen as a mutation. We run FreeBayes, version 1.3.2 (Garrison and Marth, 2012) on the BAM files generated above and the reference genome to calculate the mutations. The ploidy is set to one and the fragments mapping to the human genome are removed in the input to FreeBayes. The default values are used for the rest of the FreeBayes parameters. The output from FreeBayes is a list of confident mutations stored in a VCF file.

Partial genome reconstruction

Depending on the depth and distribution of the sequencing, there might be regions in the viral genome to which no fragments were mapped. In the VCF output from FreeBayes, an absence of mutations due to unmapped regions must be distinguished from an actual absence of mutations (as deemed by FreeBayes.) Therefore, before we compare the markers, we reconstruct the partial genome of the isolate by masking the un-mapped regions in the reference genome, and then apply the mutations as called by FreeBayes to the masked genome.

Isolate identification

A marker profile is generated from the reconstructed genome. The marker profile consists of the base pairs at a list of predefined marker locations. These marker locations are dictated by the clade system. Currently, three clade systems are used in the classification step: the old Nextstrain, the new Nextstrain (Hodcroft, et al., 2020), and the GISAID clade system (e.V., 2020).

Finally, given a clade system, the marker profile is compared with the reference clade-defining marker profiles (Supplementary Table 1). If the marker profile of the isolate is the same as that of a designated clade, the isolate is called as the said clade. If the marker profile of the isolate is different from that of any of the clades, the clade with a marker profile most similar (in terms of Manhattan distance) is

reported, along with the Manhattan distance between the two. When multiple clades are tied as having the least Manhattan distance, the clade of the isolate is considered ambiguous between the tied clades. If none of the Manhattan distances is below 4.5, the clade of the isolate is considered undetermined.

For two marker profiles $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$, their Manhattan distance is defined as $D_M(A, B) = \sum_{i=1}^n M_i$, where $M_i = 0$ when $a_i = b_i$, $M_i = 1$ when $a_i \neq b_i$, and $M_i = 0.5$ when either a_i or b_i is undetermined due to lack of mapped reads at the marker region.

Software

The web user interface

After idCOV is deployed on a Linux environment, it can be accessed using any modern browser. A user can either upload the gzipped FastQ files stored on the client machine or specify a list of gzipped FastQ files already on the server file pool. Each pair of gzipped FastQ files needs to be named “*R1.fastq.gz” and “*R2.fastq.gz” so that their parity can be determined. A project can be created with the uploaded or specified files. The pipeline can be started by running a project. The sessions are tracked on the server, so the progress of any running project will not be lost when the client browser is closed. (Supplementary Figure 1)

The command line interface

A command line interface will also be provided on the server-side after idCOV is deployed. Assuming terminal access (either direct or through SSH), a user can use the command “idcov” in the command line to run the pipeline directly on any gzipped FastQ files on the server.

Data security

When deployed on a private computer or a cloud VM, all uploaded data is stored in a folder specified by a Docker volume binding parameter. The administrator of the computer can set the appropriate access permissions for this folder to ensure the data security.

Implementation

The workflow management is done by using the Nextflow system (Di Tommaso, et al., 2017). The individual steps are broken down into Nextflow processes, and the deployment on multiple platforms is done by preparing a configuration for HPC and another for cloud environment. The latter can also be used on personal computers. A Datahike database (Kühne, et al., 2020) is used for storing user accounts, sessions, and job tracking. The server is written in Clojure using the Ring library. The web frontend is written in ClojureScript using the Fulcro library. The command-line interface is written in Python.

Testing

The Clojure/ClojureScript code is unit-tested using the clojure.test library. The web UI is tested in Google Chrome Version 84.0.4147.105 (Official Build) (64-bit) and Mozilla Firefox Version 79.0.

We tested the entire workflow on 18 representative samples from a SARS-Cov-2 genomic sequencing dataset reported by the Peter Doherty Institute for Infection and Immunity and the Victorian Department of Health and Human Services in Australia (Seemann, et al., 2020). These samples included viral sequences representative of L, V, S, G, GR, and GH clades (GISAID clade nomenclature). idCOV correctly identified the clade of all viruses, regardless of the clade-defining marker profile used (i.e., Nextstrain old, new, and GISAID) (Figure 3). The full report is attached in the Supplementary Materials.

Conclusion

Rapid classification of SARS-Cov-2 viruses into their respective genetic clades helps characterize transmission patterns and assess local and global epidemiologic trends. idCOV provides an easy portal to a complete bioinformatics pipeline that identifies a virus' clade using the three most widely used clade systems. The system is internally tested and has been shown to provide valuable insights to clinicians and biologists in a timely manner.

Funding

This research was supported in part by American Lebanese Syrian Associated Charities (ALSAC) and the National Cancer Institute grant P30 CA021765. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

None declared.

Figures

Figure 1: The bioinformatics pipeline of idCOV.

Figure 2: idCOV user experience.

Figure 3: Screenshots of an example report of idCOV. The dataset consists of a random selection of isolates from Seemann, et al., 2020 covering all major clades. Table A) shows the final calls with color scheme consistent with Nextstrain.org. Table B) shows the Manhattan distances between every sample and every clade. A smaller Manhattan distance indicates a higher degree of similarity. Table C) shows the marker profiles and Table D) shows the marker coverages. When a marker is covered by fewer than 10 reads, the marker is deemed as undetermined and denoted as a question mark (?).

References

- CSSE. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). In. ArcGIS Online (arcgis.com); 2020.
- Di Tommaso, P., *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* 2017;35(4):316-319.
- Dong, E., Du, H. and Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 2020;20(5):533-534.
- e.V., F.v.G. GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. In. Gisaids.org: GISAID; 2020.

Garrison, E. and Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* 2012.

Gonzalez-Reiche, A.S., *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 2020.

Hodcroft, E.B., *et al.* Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org. In. Nextstrain.org: Nextstrain; 2020.

Kühne, K., Weilbach, C. and Prokopov, N. Datahike is a durable Datalog database powered by an efficient Datalog query engine. In. Github.com; 2020.

Lescure, F.-X., *et al.* Clinical and virological data of the first cases of COVID-19 in Europe: a case series. *The Lancet Infectious Diseases* 2020.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* 2009;25(14):1754-1760.

Li, H., *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Seemann, T., *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv* 2020.

Wu, F., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 2020;579(7798):265-269.

Calls

[Download CSV](#)

Sample ID	GISAID	New Nextstrain	Old Nextstrain
TAS111	GH_1	20C	A2
TAS221	S_2	19B_2	B1
TAS255	S_1	19B_1	B/B4
TAS78	V_1	19A_2	A1a
TAS87	S_1	19B_1	B/B4
VIC07	S_1	19B_1	B/B4
VIC146	V_1	19A_2	A1a
VIC147	GH_1	20C	A2
VIC149	G	20A_1	A2
VIC1490	L/O_1/O_2/V_2	19A_1	A3
VIC1783	GH_2	20A_2	A2
VIC2284	GR	20B	A2a
VIC2646	GR	20B	A2a
VIC4055	GR	20B	A2a
VIC527	G	20A_1	A2
VIC61	V_1	19A_2	A1a
VIC803	G	20A_1	A2
VIC951	V_1	19A_2	A1a

Call Manhattan Scores

[Download CSV](#)

Sample ID	GISAID												New Nextstrain								Old Nextstrain							
	G	GH_1	GH_2	GR	L	O_1	O_2	S_1	S_2	V_1	V_2	19A_1	19A_2	19B_1	19B_2	20A_1	20A_2	20B	20C	A1a	A2	A2a	A3	A6	B	B1	B4	
TAS111	2	0	1	5	6	6	8	8	11	9	8	6	7	8	11	2	1	5	0	8	3	5	7	6	7	11	7	
TAS221	8	10	9	11	4	4	6	2	1	7	6	4	5	2	1	8	9	11	10	6	7	11	5	4	3	1	3	
TAS255	7	9	8	10	3	3	5	1	4	4	5	3	4	1	4	7	8	10	9	3	6	10	4	3	2	4	2	
TAS78	7	9	8	10	3	3	1	5	8	0	1	3	2	5	8	7	8	10	9	1	6	10	2	3	4	8	4	
TAS87	7	9	8	10	3	3	5	1	4	4	5	3	4	1	4	7	8	10	9	3	6	10	4	3	2	4	2	
VIC07	6	8	7	9	2	2	4	0	3	5	4	2	3	0	3	6	7	9	8	4	5	9	3	2	1	3	1	
VIC146	7	9	8	10	3	3	1	5	8	0	1	3	2	5	8	7	8	10	9	1	6	10	2	3	4	8	4	
VIC147	2	0	1	5	6	6	8	8	11	9	8	6	7	8	11	2	1	5	0	8	3	5	7	6	7	11	7	
VIC1490	5	7	6	8	1	1	1	3	6	2	1	1	2	3	6	5	6	8	7	3	4	8	0	1	2	6	2	
VIC149	0	2	1	3	4	4	6	6	9	7	6	4	5	6	9	0	1	3	2	6	1	3	5	4	5	9	5	
VIC1783	2	2	1	5	6	6	6	8	11	7	6	6	7	8	11	2	1	5	2	8	3	5	5	6	7	11	7	
VIC2284	3	5	4	0	7	7	9	9	12	10	9	7	8	9	12	3	4	0	5	9	4	0	8	7	8	12	8	
VIC2646	3	5	4	0	7	7	9	9	12	10	9	7	8	9	12	3	4	0	5	9	4	0	8	7	8	12	8	
VIC4055	3.5	5.5	4.5	0.5	7.5	7.5	9.5	8.5	11.5	10.5	9.5	7.5	8.5	8.5	11.5	3.5	4.5	0.5	5.5	9.5	4.5	0.5	8.5	7.5	7.5	11.5	7.5	
VIC527	0	2	1	3	4	4	6	6	9	7	6	4	5	6	9	0	1	3	2	6	1	3	5	4	5	9	5	
VIC61	7	9	8	10	3	3	1	5	8	0	1	3	2	5	8	7	8	10	9	1	6	10	2	3	4	8	4	
VIC803	0	2	1	3	4	4	6	6	9	7	6	4	5	6	9	0	1	3	2	6	1	3	5	4	5	9	5	
VIC951	7	9	8	10	3	3	1	5	8	0	1	3	2	5	8	7	8	10	9	1	6	10	2	3	4	8	4	

C

Marker profiles

[Download CSV](#)

sample_id	241	1059	3037	8782	11083	14408	14805	17747	17858	18060	23403	25563	26144	28144	28881	28882	28883
TAS111	T	T	T	C	G	T	C	C	A	C	G	T	G	T	G	G	G
TAS221	C	C	C	T	G	C	C	C	G	T	A	G	G	C	G	G	G
TAS255	C	C	C	T	G	C	T	C	A	C	A	G	G	C	G	G	G
TAS78	C	C	C	C	T	C	T	C	A	C	A	G	T	T	G	G	G
TAS87	C	C	C	T	G	C	T	C	A	C	A	G	G	C	G	G	G
VIC07	C	C	C	T	G	C	C	C	A	C	A	G	G	C	G	G	G
VIC146	C	C	C	C	T	C	T	C	A	C	A	G	T	T	G	G	G
VIC147	T	T	T	C	G	T	C	C	A	C	G	T	G	T	G	G	G
VIC149	T	C	T	C	G	T	C	C	A	C	G	G	G	T	G	G	G
VIC1490	C	C	C	C	T	C	C	C	A	C	A	G	G	T	G	G	G
VIC1783	T	C	T	C	T	T	C	C	A	C	G	T	G	T	G	G	G
VIC2284	T	C	T	C	G	T	C	C	A	C	G	G	G	T	A	A	C
VIC2646	T	C	T	C	G	T	C	C	A	C	G	G	G	T	A	A	C
VIC4055	T	C	T	?	G	T	C	C	A	C	G	G	G	T	A	A	C
VIC527	T	C	T	C	G	T	C	C	A	C	G	G	G	T	G	G	G
VIC61	C	C	C	C	T	C	T	C	A	C	A	G	T	T	G	G	G
VIC803	T	C	T	C	G	T	C	C	A	C	G	G	G	T	G	G	G
VIC951	C	C	C	C	T	C	T	C	A	C	A	G	T	T	G	G	G

D

Coverages

[Download CSV](#)

sample_id	241	1059	3037	8782	11083	14408	14805	17747	17858	18060	23403	25563	26144	28144	28881	28882	28883
TAS111	1113	3541	2281	1413	2548	13151	2892	2179	208	4541	6081	7736	7157	4404	48	49	48
TAS221	343	4061	2942	20	333	4051	1175	1332	110	4957	3842	9838	2268	3382	332	338	338
TAS255	2010	3293	5381	1861	5409	5565	6360	1542	1084	3225	5649	5850	7573	4049	960	976	972
TAS78	5139	2270	4553	8301	3509	5831	5181	2320	2534	2714	4589	3351	9394	3815	325	330	325
TAS87	5122	2840	5242	9564	4017	6508	5814	2538	2993	2832	5984	3403	9493	4324	761	759	750
VIC07	81	2351	3756	889	5263	8551	3419	1375	868	3143	7999	7547	11860	5803	293	293	290
VIC146	1134	3440	3933	134	4515	7436	5500	1575	130	2283	4195	7138	15864	6117	654	656	643
VIC147	2825	2686	4042	454	5476	7477	5322	1270	169	2531	3464	5896	12948	5227	456	456	453
VIC149	1014	1328	1587	92	1889	3144	2172	552	38	1144	1413	2826	6102	2351	228	230	231
VIC1490	3794	2768	2954	2796	1772	2498	2084	1396	1815	1328	4722	2312	211	1960	1657	1656	1652
VIC1783	479	1276	2730	644	2480	1854	2571	438	359	1549	3731	1806	3546	2655	410	417	408
VIC2284	1638	3757	4090	2696	6238	6073	8218	2522	2460	2994	7585	7685	8515	5481	2220	2218	2240
VIC2646	8395	5293	7220	2369	6930	7852	9449	4811	5469	4572	7063	8974	14717	6028	2266	2261	2295
VIC4055	11	4418	5709	6	144	7465	18	1058	22	3545	1292	8002	9392	683	13	13	13
VIC527	4461	2633	5538	4410	4662	6503	6088	1043	580	2709	5345	3042	8384	3781	979	992	983
VIC61	13135	9784	5006	144	10806	23760	6922	2384	604	7179	6795	8752	2558	6003	48	47	47
VIC803	3474	2584	3140	46	4864	6415	5543	1191	641	2927	6388	3440	4426	4692	198	199	194
VIC951	5095	4003	6828	753	6385	7595	8279	2390	1992	3636	8309	4765	10677	6069	1162	1165	1152