Next-Generation Sequencing on COVID-19 Pandemic

Jiahuan He*
McGill University, Montréal, Québec, Canada.

* Corresponding author. Tel.: (1)4388661160; email: jiahuan.he@mail.mcgill.ca
Manuscript submitted October 11, 2021; accepted January 8, 2022.
doi: 10.17706/ijbbb.2022.12.2.30-38

Abstract: Next-generation sequencing (NGS), with Illumina sequencing being the most well-known and most widely used NGS platform, is a high-throughput DNA (or RNA) sequencing technology that allows massive parallel sequencing. There are numerous applications of NGS, with some of them being related to the pandemics or epidemics caused by viral infection, and an example of that is the coronavirus disease 2019 (COVID-19), which is a pandemic that began in December 2019 caused by a novel coronavirus—the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This review then focuses on the discussion of how NGS technology has been applied to the study of the COVID-19 pandemic. Specifically, the review describes the applications of NGS in whole-genome sequencing of SARS-CoV-2, identification of novel COVID-19 viral mutations, tracking the variant viral lineages, and providing insights into the origin and transmission pattern of the current pandemic.

Key words: Next-generation sequencing, Illumina sequencing, COVID-19, novel virus identification, viral variant, transmission pattern.

1. SARS-CoV-2 Molecular Biology & Infection

1.1. Molecular Biology of SARS-CoV-2

Coronavirus disease 2019 (COVID-19) emerging in December 2019 is caused by a novel coronavirus—the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The disease initially appeared in Wuhan city, Hubei province in China, and then outbroke into a global pandemic [1]. People contracting SARS-CoV-2 can present fever, upper or lower respiratory tract symptoms, pneumonia and etc., and under severe conditions, death can be resulted [2]. The genetic material of SARS-CoV-2 is a single-stranded positive-sense genomic RNA stabilized by nucleocapsid proteins (N proteins) and encapsulated by a viral envelop [3]. The envelop has three structural proteins on it: membrane protein (M protein), spike protein (S protein) and envelope protein (E protein) [3]. The way that SARS-CoV-2 enters the host cells is via the transmembrane S protein, which is composed of two subunits, S1 and S2. The S1 subunit contains the receptor-binding domain (RBD), whereas S2 subunit has the fusion peptide (FP) [4]. The fusion is triggered when RBD on S1 subunit binds to the angiotensin-converting enzyme 2 receptor (ACE2 receptor), which results in cleavage at S1/S2 cleavage site [5]. A subsequent cleavage at the S2' site activates the fusion machinery, which fuses the viral membrane with the host cell membrane to let the virus enter the host cell [5]. Following the entry, the viral genomic RNA undergoes translation and produces pp1a and pp1ab which are further processed into viral replication and transcription-polymerase [6]. During the transcription of viral genomic RNA, a set of subgenomic mRNA (sg mRNA) is generated to be translated later into different viral structural proteins [7]. The replicated viral RNA bound by N proteins enters the ER-Golgi intermediate cavity, which is then

assembled with other structural proteins (S, M, E proteins) into an entire virus, and eventually, the virus exits the host cell by exocytosis [7]. After releasing hundreds of viruses, the host cell usually dies owing to running out of resources or being killed by the immune system, and this is how SARS-CoV-2 causes cell/tissue damage.

1.2. Infection Mechanism & Immune Responses

When SARS-CoV-2 viruses enter the lungs and encounter the alveolar macrophages or dendritic cells, they will be engulfed by those cells via the interaction between pathogen-associated molecular patterns (PAMPs) and pattern recognition receptors (PRR) [8]. Inside the dendritic cells, the viral RNA will trigger type I interferon production, which in turn activates natural killer cells. In addition, the viral particles will be combined with dendritic cells' type I and type II major histocompatibility (MHCI & MCHII) and will be presented on dendritic cells' membrane [9]. The dendritic cells then mature and migrate to lymph nodes, where they present the antigen to naïve CD4+ T cells [10]. The naïve CD4+ T cells' T cell receptor (TCR) binds to the antigen-MHCII complex and differentiates into T follicular helper cells (Tfh cells) or type I helper T cells (Th1 cells) with the co-stimulatory signal and cytokines released by dendritic cells [11]. Tfh promotes B cells to proliferate and differentiate into memory B cells or plasma cells to produce antibodies, and Th1 cells co-activate CD8+ T cells to cytotoxic T cells when the TCR of CD8+ T cells bind with MHCI-antigen complex on dendritic cells [12].

With natural killer cells, cytotoxic T cells and the neutralizing antibodies produced by plasma cells, SARS-CoV-2 infection should be under the control by immune system in theory. However, SARS-CoV-2 virus is capable of evading the innate immune system, resulting in failure to prime the adaptive immunity and hence the increasing the severity of infection [13].

2. Next-Generation Sequencing—Background, Principle and Method

2.1. Background of NGS

DNA sequencing is the decryption of the nucleotide sequence in DNA, and the current DNA sequencing method used by most biologists is the next generation sequencing (NGS) that is derived from the first-generation sequencing technologies developed in 1970s, which include Sanger sequencing and Maxam-Gilbert sequencing [14]. Maxam-Gilbert sequencing involves chemical modification on DNA and subsequent cleavage of the DNA backbone on the sites of modification, and the one end of each DNA fragment generated by cleavage is labeled with radioactive substance. Sanger sequencing uses terminating nucleotides (dideoxynucleotide triphosphates (ddNTPs)) that have a terminator on their 3' hydroxyl group which prevents DNA polymerase to extend the chain, and these terminating nucleotides are also fluorescently labeled for detection [14]. These two first generation sequencing methods were utilized until the emergence of NGS, also known as "second generation sequencing", which can be further divided into two categories: sequencing by hybridization (SBH) and sequencing by synthesis (SBS) [14]. SBH utilizes hybridization between oligonucleotides and the sample DNA that needs to be sequenced, whereby the overlapping regions between those oligonucleotides that have successfully hybridized to the sample DNA are assembled together to form a larger continuous sequence [15]. SBS, with Illumina Sequencing by Synthesis being the most known version, is an improved version of Sanger sequencing, and it is also the most widely used sequencing method today [14].

2.2. Workflow of Illumina NGS

The workflow of Illumina sequencing involves four main steps—library preparation, cluster generation, sequencing and data analysis [16]. In the library preparation stage, the DNA or cDNA sample (used for RNA sequencing) are randomly fragmented, and then both ends of the fragments are ligated with adapter sequences that consist of a capture sequence and a primer sequence, where capture sequence allows the

fragments to be captured by the oligo sequences in the flow cell and the primer sequence enables DNA polymerase to bind and to elongate the fragment. As for the cluster generation, the library of fragments are denatured from double-stranded to single-stranded and are then loaded into the flow cell, which is a glass slide with several lanes, each of which is randomly coated with oligos that are complementary to the capture sequence of the library fragments. Then DNA polymerases are added into the flow cell to elongate the library fragments that are bound to the oligos. After the polymerization, the original template strand (the bound library fragment) is washed away, whereas the newly synthesized strand is covalently attached to the flow cell surface. The newly synthesized strand now becomes a template strand, bending over with its capture sequence on its top hybridizing to a complementary oligo on the flow cell surface, forming a "single-stranded bridge", and the DNA polymerase then elongates from the oligo and forms a new strand, generating a "double-stranded bridge". The "double-stranded bridge" is denatured, generating two copies of single strand that are covalently attached to the flow cell surface. The transition from "single-stranded bridge" to "double-stranded bridge" and again to "single-stranded bridge" is repeated until a cluster (about 1000 copies) of single-stranded DNA forms which include both forward and reverse strands. The reverse strands are cleaved away, leaving only the forward strands. Then, the free 3' ends are blocked to prevent unwanted DNA priming in the following sequencing stage, and primer sequences are hybridized onto the single strands for DNA polymerase to attach [16].

There are four different nucleotides, adenine (A), guanine (G), thymine (T) and cytosine (C), and they are labeled with different fluorescent dye for their differentiation, but all of them have a terminator molecule attached on their 3' carbon position. During sequencing, one fluorescently labeled terminator nucleotide is added at a time, and a picture of the clusters is taken. After that, the terminator molecule and the fluorescent dye are removed from the nucleotide, and next fluorescently labeled terminator nucleotide is added. Thus, by synthesizing the DNA strand, the sequence of that strand can be known based on the fluorescent dye color shown in the picture of each round. The last step is data analysis, where the newly identified sequences are assembled together are aligned with a reference genome, and by doing so mutations, such as point mutation, insertion and deletion, can possibly be found [16].

2.3. Advantages & Limitations of Illumina NGS

One of the greatest breakthroughs by Illumina SBS is the parallel sequencing of millions of different DNA-strand clusters on the flow cell surface, and this give rise to much higher throughput compared to that of Sanger sequencing—AmpliSeq of Illumina can produce around 300 kb with targeted resequencing (250 bp amplicon length X 1536 amplicons) from 10 ng DNA, whereas Sanger sequencing can only produce around 1 kb from the same amount of DNA samples [17]. Despite the high throughput, Illumina SBS still has limitations, and the major one is the lack of synchrony for synthesis of DNA strands within a cluster. During the synthesis, the synthesis of some strands in a cluster may jump ahead and some may lag behind, and as synthesis goes on, more and more strands become asynchronized, and the asynchronized signals will aggravate to an extent that the true signal is overwhelmed. Hence, there is a length limit to Illumina SBS, which is from 200 to 400 bases for most sequencing platforms of Illumina [18].

2.4. Nanopore NGS

Besides Illumina SBS, nanopore sequencing is another NGS method (but instead of being "second generation", it is "third generation sequencing") applied in a study by Zhu, *et al.* published in January 2020, which is one of the earliest researches for SARS-CoV-2 genome sequencing. Nanopore sequencing works by letting a single-stranded DNA that needs to be sequenced to pass through a nanopore—every time a nucleotide passes the pore, there will be a change in ion currents across the pore, with different nucleotides generating distinctive current changes, and by detecting these current changes, one can know the identity of

the nucleotide passing through the pore [19]. The greatest advantage of nanopore sequencing is its portability, and the largest limitation is its high error rate. Nevertheless, when nanopore sequencing is applied in combination with other sequencing method, such as Illumina SBS, in which the sequencing volume is high enough, that limitation can be circumvented [14].

3. NGS Assembly

NGS provides short read sequences rather than the whole genome sequence, so in order to obtain sequence for the whole genome, NGS assembly is required, which is the reconstruction of sequence up to chromosome length [20]. The general process of NGS assembly is connecting contiguous short reads into longer reads (contigs), and then the contigs are joined together into even longer reads (scaffolds) [21]. There are four steps for NGS assembly: preprocessing filtering, graph construction, graph simplification, and post-processing filtering [21]. In preprocessing filtering stage, the erroneous reads are corrected or eliminated, thereby minimizing the misassembled contigs in following steps [21]. After the preprocessing filtering, the short reads are stored in an abstract data structure, graph, which in this case is used to show relationship/similarity (edges) between different reads (nodes) [21]. The next stage, graph simplification, is simplifying the graph by reducing the number of edges and nodes and by eliminating erroneous reads. The final stage, post-processing filtering, assembles short reads to contigs, filters out misassembled ones, and join contigs into scaffolds [21].

There are four approaches for preprocessing filtering: K-spectrum approach, Suffix Tree/Array approach, Multiple Sequence Alignment approach (MSA), and Hybrid approach [21]. Graph construction also has four main methods: Overlap/Layout/Consensus (OLC) method, de Bruijn Graph (DBG) method, Greedy method, and Hybrid method [20]. Because Zhu et al.'s study in 2020 (one of the earliest researches on whole genome sequencing of SARS-CoV-2 as mentioned previously) utilized K-spectrum preprocessing filtering and DBG method on CLC Bio, the following discussion will mainly focus on these two [22]. K-spectrum preprocessing filtering approach works by extracting all k-mers (k stands for the length of a read) from the reads and then assigning each k-mer with a weight value (determined by the k-mer's frequency and the quality scores of the bases in that k-mer) [21]. The k-mers are then sorted based on their weight values, and all the k-mers whose weight value is below the threshold will be sent to correct their errors [21]. As for DBG graph construction method, it is based on K-mer graphs. The nodes in the graph represent the k-mers, with the forward neighbor of each node representing the first k-1 bases of the current node and the backward neighbor representing the last k-1 bases [22]. The edges in the graph denote the shared k-1 bases between two adjacent nodes [22]. In the graph simplification process, the K-mer graph is simplified by combining linearly connected n nodes with k bases into one node with k+n bases, but if there is an error or single nucleotide polymorphism (SNP), (a) bubble(s) (one line of nodes diverges into multiple paths and then converges back to one line again) will form in the graph [22]. The frequency of the multiple paths in the bubble scan be used to distinguish SNPs from errors—paths at equal frequency means SNPs, whereas if one path is at low frequency and the others have high frequency, the low-frequency path is highly likely a sequencing error [22]. If there are repeats in the reads, several paths in the graph will converge to one path (the repeat) and then diverges back to multiple paths [22]. To make repeat sequences less complicated, CLC bio increases k's value with larger number of input reads—longer k-mers can more likely incorporate uniqueness of sequence, thereby reducing repeats in the graph [22].

4. Application of NGS in COVID-19

4.1. Identification of Novel Virus

As what's mentioned previously, Zhu, N., *et al.* conducted the first most prominent COVID-19 study, in which

they performed de novo sequencing of 2019-nCoV genome by Illumina and nanopore sequencing. In the study, three bronchoalveolar-lavage samples from patients with pneumonia of unknown cause were obtained from Wuhan Jinyintan Hospital. Zhu, N., *et al.* extracted RNA from the bronchoalveolar-lavage fluid samples from three patients and used a combination of Illumina sequencing and nanopore sequencing to figure out a genome from those RNA molecules. More than 20,000 viral reads were obtained, and they were assembled into contig map, which is a set of overlapping DNA segments used to construct a genome. With the majority of the sequence already deciphered by Illumina sequencing and nanopore sequencing, 5'- and 3'-RACE (rapid amplification of cDNA ends) was performed to fill in the gaps between contigs to obtain a full sequence of RNA. The complete genome sequences of the novel coronavirus obtained from the three patients showed 86.9% resemblance to the genome of a bat SARS-like coronavirus (bat-SL-CoVZC45, MG772933.1), and the three genomes all fall into sarbecovirus subgenus, which typically had the following genome organization: a 5' untranslated region (UTR), replicase complex (ORF 1ab), S gene, E gene, M gene, N gene, 3' UTR, and other unidentified sequences for non-structural proteins [1]. Despite the high similarity (86.9%) to the two bat coronavirus strains, because sequence of the conserved replicase complex domains (ORF 1ab) of the newly identified coronavirus is less than 90% identical to that of other members of beta-coronavirus, this virus is considered a novel beta-coronavirus. Thus, by using NGS with other supplementary methods (e.g. 5'- and 3'-RACE), the complete genome of a virus can be obtained, and it can then be compared to the genomes of other previously identified viruses to see whether the new virus is novel.

4.2. Detection of Viral Variants

Not only can NGS be used for determining whether the virus is novel, but also can it be applied to identify and track new mutations of the virus. In principle, NGS is used to sequence new viral samples, and usually there is a reference genome for the new sequences to compare with. Different viruses, even the viruses of the same virus strain, can have differences in their sequences due to sequence variation caused by error-prone replication with or without proof-reading system as well as spontaneous editing and damage to nucleotides [23]. Some of the sequence variations can persist because they are advantageous for the virus to infect the hosts and replicate, and the viruses with those sequence variations will become dominant lineage of the originally identified virus strain. For the viruses that acquired undesired sequence variations, they will be outcompeted by the dominant lineages and will be wiped out eventually. D614G variant is an example of SARS-CoV-2 mutation, with Spike D614G amino acid change (change from D614 TO G614) caused by an A-G nucleotide mutation at position 23,402 in the first-identified Wuhan strain (the reference sequence) [24]. In the study conducted by Korber, B., et al., the way they identified SARS-CoV-2 mutations was to look at the SARS-CoV-2 sequences from Global Initiative for Sharing All Influenza Data (GISAID) database, to which hundreds of new SARS-CoV-2 sequences are added daily. The criterium that they set for a SARS-CoV-2 Spike mutation being worth to notice was that the sequence should differ more than 0.3% from the Wuhan reference sequence MN908947v3 [24], [25]. By examining the new sequences added to the GISAID that meet that criterium, Spike D614G variant was identified. The D614G variant was found to account for only 10% of the 997 global sequences before March 1, 2020, but it gradually became more prevalent and eventually dominant, representing 78% of the 12,194 global sequences between April 1 and May 18, 2020. The increase of D614G variant prevalence might be relevant to its higher viral loads in COVID-19 patients, which indicate its higher infectivity [26]. Another more recent example for SARS-CoV-2 variant is B.1.617.2 lineage (delta variant), which was firstly identified in India in December 2020, but then spread widely across the globe and started to become the dominant variant. According to the study done by Liu, C., et al., B.1.617.2 proportion in the global sequences surged around June 4, 2021 and superseded the previously more dominant B.1.617.1 variant [27]. B.1.617.2 variant has mutation L452R and T478K in the receptor-binding domain (RBD) region, T19R, G142D, R158G, A222V substitutions and a double deletion (156-157) in the N-terminal domain (NTD),

as well as D950N substitution in S2 [27]. These mutations render the variant capable to escape from the neutralizing antibodies, with 2.7-fold reduction in convalescent plasma neutralizing capability, 2.5-fold reduction in Pfizer-BioNTech vaccine neutralizing capability and 4.3-fold reduction in Oxford-AstraZeneca vaccine neutralizing capability [27].

4.3. Genomic Surveillance

In addition to identifying the novel virus and viral mutants, NGS can also provide insights into the origin and transmission of epidemics within a certain region or pandemics across the globe. For instance, in the study conducted by Fang, B., *et al.*, 16 sequences of Huanan seafood market were collected, 4 of which were newly sequenced by metagenomic NGS and 12 of which were obtained from GISAID database [28]. The 16 sequences were categorized into 10 haplotypes — a viral haplotype is a collection of variation in the viral genome that are passed to newly synthesized viral genetic materials after replication [29]. Among the 10 haplotypes, H3 haplotype was found in 6 sequences out of the 16 and the other haplotypes were all found to be directly derived from H3 haplotype, which means that there was a circulating infection within the Huanan seafood market in a short-term period. Another example is the identification of the source for 53 cases in Guangdong province in China. The genome sequences were generated from the 53 patients by using metagenomic NGS, and it was found that those sequences are interspersed with virus clades from other provinces in China or even other countries, which means most cases in Guangdong derived from travelling instead of emerging from local communities [30]. One more example can be seen in the report by Nemudryi, A., *et al.*, in which they sequenced 55 whole genomes of SARS-CoV-2 isolated from patients in Bozeman, and they found that there were 14 independent viral lineages, which means there had been multiple introductions of the virus into Bozeman from different sources [31]. As mentioned previously, one of the greatest functions for NGS is to detect new viral mutations and to keep track of the new viral variants in a population. Going back to the study done by Korber, B., *et al.*, by looking at the sequences generated via NGS, they found that the D614G variation was accompanied by three other mutations in most cases, which are C-to-T mutations at position 241, 14408 and 3037 relative to the Wuhan reference sequence [24]. By looking at the proportion of the sequence that possesses these four mutations (the D614G mutation and the three C-to-T mutations) in the regional sequences from GISAID database, the transition from D614 to G614 in a specific region can be monitored — the prevalence of D614G variant increased asynchronously in different regions globally, with Europe being the first region, then North America and Oceania and finally followed by Asia [24].

5. Conclusion

Next-generation sequencing (NGS), as a high-throughput DNA sequencing method, has been applied to many fields, including novel virus identification, mutated viral lineage identification and genomic surveillance. COVID-19 pandemic is a great example of NGS application in pandemics caused by viral infection, and it has indeed been utilized to the aforementioned domains, in that the SARS-CoV-2 sequence generated by NGS allowed people to determine SARS-CoV-2 as a novel virus, to identify mutated viral lineage, to keep track of different viral linages, their origin and their transmission pattern in the entire globe or in certain regions. COVID-19 pandemic reveals the indispensable role played by NGS in tracking the progression of the pandemic, but the application of NGS is not confined only in the viral pandemics—its application can be seen in pandemics caused by other types of antigens as well as in other biological fields including gene editing, cancer and any domains that entail genomic sequencing.

Conflict of Interest

The author declares no conflict of interest.

Author Contributions

Jiahuan He conducted online research and wrote this review paper.

Acknowledgment

The first author would like to thank Dr. Kai Fu from Sandford University for providing essential background knowledge and feedback for this review paper, and she would also like to thank Dr. Erik S. Yan and lan Deng for their support on paper submission.

References

[1] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., *et al.* (2020). A novel coronavirus from patients with pneumonia in China, 2019. *The New England Journal of Medicine, 382(8)*, 727-733.

[2] Chan, J., Yuan, S., Kok, K. H., To, K. K., Chu, H., Yang, K., *et al.* (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *The Lancet, 395(10223),* 514-523.

[3] Brian, D. A., & Baric, R. S. (2005). Coronavirus genome structure and replication. *Current Topics Microbiology and Immunology, 287*, 1-30.

[4] Li, R., Pei, S., Chen, B., Song, Y., *et al.* (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science, 368(6490)*, 489-493.

[5] Walls, A. C., Tortorici, M. A., Snijder, J., Xiong, X., Bosch, B. J., Rey, F. A., *et al.* (2017). Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proceedings of the National Academy of Sciences of the United States of America: Vol. 114, No. 42.* (pp. 11157-11162).

[6] V'Kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2021). Coronavirus biology and replication: Implications for SARS-CoV-2. *Nature Review of Microbiology, 19(3)*, 155-170.

[7] Li, S., Li, S., Disoma, C., Zheng, R., Zhou, M., Razzaq, A., *et al.* (2020). SARS-CoV-2: Mechanism of infection and emerging technologies for future prospects. *Reviews in Medical Virology, 31(2)*, e2168.

[8] Prompetchara, E., Ketloy, C., & Palaga, T. (2020). Immune response in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. *Asian Pacific Journal of Allergy and Immunology, 38(1)*, 1-9.

[9] Wieczorek, M., Abualrous, E. T., Sticht, J., Alvaro-Benito, M., Stolzenberg, S., Noe, F., *et al.* (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Frontiers in Immunology, 8*, 292.

[10] Cella, M., Salio, M., Sakakibara, Y., Langen, H., Julkunen, I., & Lanzavecchia, A. (1999). Maturation, activation, and protection dendritic cells induced by double-stranded RNA. *Journal of Experimental Medicine, 189*, 821-829.

[11] Luckheeram, R. V., Zhou, R., Verma, A. D., & Xia, B. (2012). CD4+ T cells: Differentiation and functions. *Clinical and Developmental Immunology, 2012*, 925135.

[12] Bousso, P. (2008). T-cell activation by dendritic cells in the lymph node: Lessons from the movies. *Natural Reviews Immunology, 8*, 675-684.

[13] Sette, A., & Crotty, S. (2021). Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell*, *184(4)*, 661-880.

[14] Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology, 122(1)*, e59.

[15] Li, C. M., Dong, H., Zhou, Q., & Goh, K. H. (2008). Chapter 11–Biochips-fundamentals and applications. In X. J. Zhang, H. X. Ju, & J. Wang (Eds.), *Eelectrochemical Sensors, Biosensors and their Biomedical Applications* (pp. 307-383). San Diego: Academic Press.

[16] Illumina. (2015). An introduction to next-generation sequencing technology. Chapter I: Welcome to next-

generation sequencing. Retrieved from chrome-extension://ibllepbpahcoppkjjllbabhnigcbffpi/https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

[17] Illumina. (2021). Differences between NGS and Sanger sequencing. Retrieved from https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html

[18] Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports, 9*, 2856.

[19] Kono, N., & Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. *Development, Growth & Differentiation, 61(5)*, 316-326.

[20] Miller, J. R., Kore, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95(6)*, 315-327.

[21] EI-Metwally, S., Hamza, T., Zakaria, M. & Helmy, M. (2013). Next-generation sequence assembly: Four stages of data processing and computational challenges. *PLOS Computational Biology, 9(12)*, e1003345.

[22] CLC Bio. (2020). De novo assembly—How it works. Retrieved from http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/852/index.php?manual=How_it_works.html#sec:denovohowdoesitworkj

[23] Sanjuan, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cell Molecular Life Science. 73(23)*, 4433-4448.

[24] Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., *et al.* (2020). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell, 182(4)*, 812-827.

[25] Lam, L. T., Hieu, N. T., Trang, N. H., Thuong, H. T., Linh, T. H., Tien, L. T., *et al.* (2020). Whole-genome sequencing and de novo assembly of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Vietnam. *Vietnam Journal of Biotechnology, 18(2)*. Retrieved from http://dx.doi.org/10.15625/1811-4989/18/2/15082

[26] Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., *et al.* (2021). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature, 592(7852)*, 116-121.

[27] Liu, C., Ginn, H. M., Dejnirattisai, W., Supasa, P., Wang, B., Tuekprakhon, A., et al. (2021). Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell, 184(16)*, 4220-4236.

[28] Fang, B., Liu, L., Yu, X., Li, X., Ye, G., Xu, J., *et al.* (2020). Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). *BioRxiv*. Retrieved from https://doi.org/10.1101/2020.03.04.976662

[29] Shen, L., Bard, J. D., Biegel, J. A., Judkins, A. R., & Gai, X. (2020). Comprehensive genome analysis of 6,000 USA SARS-CoV-2 isolates reveals haplotype signatures and localized transmission patterns by state and by country. *Frontiers in Microbiology.* Retrieved from https://doi.org/10.3389/fmicb.2020.573430.

[30] Chen, X., Kang, Y., Luo, J., Pang, K., Xu, X., Wu, J., *et al.* (2021). Next-generation sequencing reveals the progression of COVID-19. *Frontiers in Microbiology*. Retrieved from https://doi.org/10.3389/fcimb.2021.632490.

[31] Nemudryi, A., Nemudraia, A., Wiegand, T., Nichols, J., Snyder, D. T., Hedges, J. F., *et al.* (2021). SARS-CoV-2 genomic surveillance identifies naturally occurring truncation of ORF7a that limits immune suppression. *Cell Reports, 35(9)*, 109197.

Jiahuan He was born in Nanchang, Jiangxi, China on August 24, 2000. She went to Nanchang Normal School Affiliated Experimental Elementary School for her primary education and graduated from it in 2012. Then, she went to Nanchang No. 2 middle school and Nanchang No.2 High school for her secondary education and graduated from her high school in 2018. To pursue her undergraduate study, she entered McGill University in Montréal, Québec, Canada in September 2018 and expected to graduate in April 2022.

During the first three years of her undergraduate study, she worked as a research assistant (RA) in research labs, as a teaching assistant (TA) in McGill University, and as a volunteer in hospitals and charity institute. From July 2019 to August 2019, she volunteered in Red Cross Society of China by paying return visit to 100 people who donated blood and conducted a survey on their willingness to donate hematopoietic stem cells. From November 2019 to December 2019, she also volunteered in Montréal General Hospital for helping patients finding their way in the hospital. From September 2020 to December 2020 and from January 2021 to April 2021, she was a TA for a General Chemistry course and a Mammalian Physiology course respectively. From January 2020 to August 2020, she was a RA in Professor Joseph Alan Dent's lab in McGill University and worked on a project in which she collected electropharyngeogram (EPG) data from *Caenorhabditis elegans* and helped developing a neural network for analyzing EPG. From May 2021 to present, she has been a RA in Professor Guojun Chen's lab in Rosalind and Morris Goodman Cancer Research Center in Montréal, and her project is about identifying optimal parameters for a cancer immunotherapy. From September 2021 to present, she also has been a RA in Professor Caroline Wagner's lab in McGill and works on compartmental modeling of SARS-CoV-2 dynamics in the following five years. Her current research interests are mainly on the fields of immunology, infection and cancer.

Ms. He is currently a senior undergraduate student at McGill University, and so far she has received Science Undergraduate Research Awards (SURA) in summer 2020, Faculty of Science Scholarship in 2019 and 2021, Major Hiram Mills Scholarship in Biological Sciences in 2021, and was on Dean's Honor List in 2019 and 2021.