Comparison of variant-calling pipelines for the genomic analysis of Influenza A Virus

A thesis submitted by

Minju Kim

in fulfillment of the requirements for the degree of

Master of Science

in

Department of Biology

New York University

May, 2018

Adviser: Elodie Ghedin

Summary

Influenza virus is a single strand negative-sense RNA virus that has a high error rate during replication. Consequently, an influenza infection is comprised of a diversity of minority genomic variants that is present in the virus population. Accurately and consistently detecting these minority variants in infected hosts is needed to better surveil for the emergence and circulation of new virus strains and study the evolution of the virus. I performed variant calling analyses on influenza next-generation sequencing (NGS) data by testing 6 different methods using multiple pipelines based on different algorithms. The methods based on the Genome Analysis Toolkit (GATK), developed by the Broad Institute, consistently detected the largest number of variants including single-nucleotide variants, insertions and deletions regardless of data types or read aligner used. Throughout variant concordance analyses of the various pipelines, I determined that variant-callers, rather than read aligners, impact the discordance of variant detection the most. In addition, I also found that a number of variants were exclusively detected by a single variant calling pipeline. However, there is no evidence that concordantly detected variants have significantly higher Phred quality scores and depth of coverage than those uniquely detected, which means that those cannot be regarded as false positive calls. Therefore, in order to obtain sufficient genomic variant information and minimize missing variant calls from influenza NGS data, it is recommended to combine high confident call-sets detected by multiple variant calling pipelines.

Introduction

Current advances in Next Generation sequencing (NGS) technology have made it more affordable for researchers to perform genome sequencing as a routine method [1]. Many studies are now being performed to identify genomic variants for clinical purposes and in for biological studies. Genomic variants include Single Nucleotide Polymorphisms (SNP) and Single Nucleotide Variants (SNV), as well as insertions and deletions (indels). These provide important clues about genetic changes with potential phenotypic effects in target organisms or pathogens, which researchers and clinicians can use for clinical applications and further biological studies [1]. For example, studies on Mendelian diseases have been using patient-specific variant information to identify causal genes of disease and there has long been studies on the impact of SNPs on phenotypic variations in human populations, such as lactase persistence and skin pigmentation [2]–[4]. For the success of such genomic studies, however, a verification of the accuracy and consistency of the variants identified should first be performed because false variant calls (false positives) or failure to detect key genetic variants (false negatives) could lead to inappropriate information that is being followed-up on in research and in the clinic [1]. As a result, to test for the accuracy of variant information obtained from genomic data, a number of bioinformatics pipelines that use different statistical models and algorithms have been developed [5]. Studies have compared the performance of a few of these tools to determine how consistent the variant calls are across pipelines [1], [5], [6].

Influenza is a highly contagious respiratory infection caused by influenza viruses (mostly by types A, B, and occasionally by C) [7]. The viruses are characterized by high a mutation rate (2.3×10^{-5}) and genetic diversity based on their error-prone replication mechanism which, combined with the selective pressure imposed by host immune responses, leads to rapid

evolution and emergence of new strains, with seasonal outbreaks and occasional pandemics [7], For more effective surveillance of emerging influenza virus strains and to study genetic changes, accurate and consistent detection of influenza virus minority variants is necessary. Although there has been a lot of comparative research on the pipelines used with human genomics data [1], [6], [8], [9], there has not been a similar study done with influenza virus data, although NGS is now routinely used in molecular epidemiology studies of this important pathogen.

In my research project, I employed commonly used open source read aligners (Bowtie2 and BWA (BWM-MEM) [10], [11] and variant-callers (FreeBayes; Genome Analysis Toolkit; SAMtools) [12]–[14] (**Fig. 2**) benchmarked the comparative analytic methods previously used and applied those to influenza NGS data obtained from infected ferret experiments and from patient data to compare the performance of multiple variant calling pipelines.

Unlike for human exome/genome data, there is no gold-standard reference variant information/database of human influenza A virus (H1N1pdm and H3N2) to help define true/false - positive/negative variants as used in comparative studies of human data [1], [5], [6], [15]. Thus, such comparative methods were not used here. Instead, hard-filtering with relatively strong thresholds (Phred quality score > 20 and depth of coverage of the nucleotide position > 200) was performed on the entire call-sets to minimize the number of false positive calls and increase true positive calls before conducting downstream comparisons. Only the variants, which passed the filter, whose depth of coverages are significantly higher than for true positives identified from human data (43.60x ~ 298.45x [1]) (H1N1pdm: ~1300x; H3N2: ~3315x) were analyzed. In this study, the performance was measured by the number of variants the pipeline verified. With high coverage variant results, systematic comparisons of the pipelines followed and consisted in three steps. First, in order to find the differences in overall performance among multiple pipelines, I

compared the total number of variants found by each pipeline. Then, I calculated the concordance ratios of the variants detected by different pipelines to test how consistently the variants were detected with different pipelines .This analysis was conducted at two different levels: 1. Between variant callers that used the same read aligner, and 2. Between read aligners combined with different variant callers. This would help determine whether the discordant results between the pipelines were due to the variant callers or to the read aligners. Finally, I performed statistical tests, such as one-way ANOVA tests to test whether there were significant differences in the means of Phred quality scores and depth of coverage among the concordantly found variants and uniquely detected variants using one variant caller, and pairwise t-tests to find significant differences in Phred quality scores and depth of coverage between variants concordantly discovered by the pipeline using different read aligners or uniquely detected by either Bowtie2 or BWA pipelines.

Materials and method

1. Sample Collection

1.1. Samples from H1N1pdm ferret infections

For these analyses, I used deep sequence data from an experiment on the effect of pre-immunity on virus transmission and evolution. The experiment was comprised of transmission pairs of ferrets where one set of ferrets was preimmunized with either of 4 seasonal H1N1 strains from different decades of epidemics (A/Denver/1/1957, A/Brisbane/59/07, A/Texas/36/91, or A/PR/34). After 12 weeks, this pre-immunized group was challenged with pandemic H1N1 virus (A/California/07/2009) and co-caged with naïve ferrets. A control group was comprised of ferrets who did not get pre-immunized before the challenge with pandemic H1N1 (**Fig. 1**). We

thus had 4 sets of ferrets based on their pre-immune status and method of infection: Set 1 (Naïve, direct infected); Set 2 (Naïve, contact infected); Set 3 (Pre-immunized, direct infected); Set 4 (naïve, contact infected from pre-immunized ferret) (**Table 1**).

Nasal washes were collected from each ferret at different time points post challenge. From the nasopharyngeal swab, RNA was isolated and influenza A virus segments were amplified using the Multisegment-RT-PCR (M-RT-PCR) approach and random primed using the Sequence Independent Single Primer Amplification (SISPA) methodology, and the fragments were pair-end sequenced on the MiSeq [16]–[18]. Total 142 sample data were used from the experiment (**Table 1**) and the mean coverage for each sample was 2092 with the standard deviation of 740 (**Table 2**).

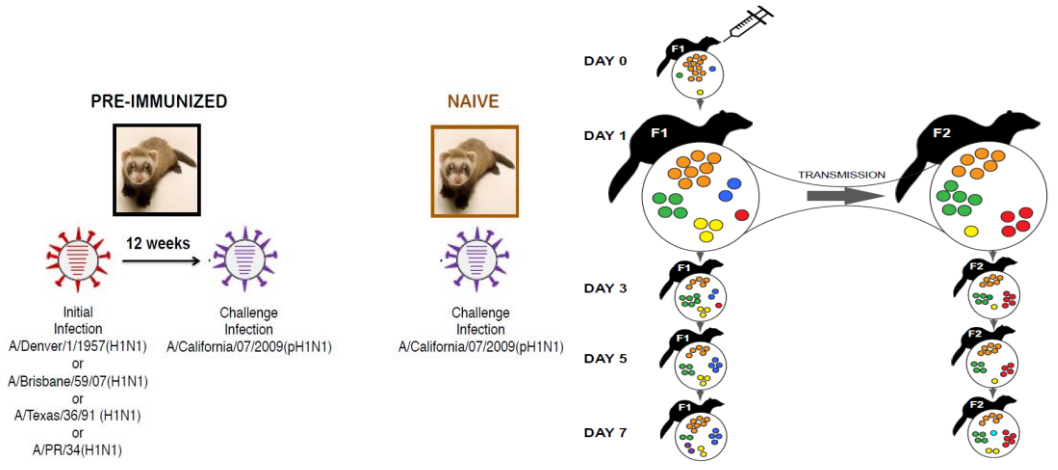

Figure 1. Ferret infection scheme

Table 1. Total number of samples data from 4 groups of ferrets

Naïve Group		Pre-immunized Group	
63		79	
Naïve, Direct Infected (Set 1)	Naïve, Contact Infected (Set 2)	Pre-Immunized, Direct Infected (Set 3)	Naïve, Contact Infected (Set 4)
23	40	47	32
142			

1.2. Samples from H3N2 human infections

Human influenza A H3N2 samples were collected at Weill Cornell Medical College/New York Presbyterian Hospital (WCMC/NYP), located in Manhattan, New York City between September 2014 and January 2015. Subjects were in outpatient clinics, in the ER or were admitted to NYP and diagnosed with acute influenza based on clinical symptoms and confirmed by Film Array PCR for respiratory viruses on nasopharyngeal (NP) swabs, or Bronchoalveolar lavage (BAL). Total RNA was isolated form the NP swabs and influenza A H3N2 virus segments were amplified using the M-RT-PCR approach and random primed using the SISPA methodology and the fragments were sequenced on the MiSeq, 2x250 bp paired-end Illumina platform [17], [18]. The mean coverage for each sample was 5,579 with the standard deviation of 3,355 (**Table 2**).

Table 2. Mean coverage (mean number of reads across all 8 segments) for each sample

	H1N1pdm (142)	H3N2 (107)
Mean	2,092	5,579
SD	740	3,355

2. Alignment and variant calling

2.1. Reference sequence

The FASTA files of reference sequences (A/California/07/2009(H1N1)) and A/Perth/16/2009(H3N2)) were downloaded from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) database (**Table 3**). In this project, across all 8 segments, only the coding sequences from the first ATG to the last stop codon were used from the reference sequences.

Table 3. Accession codes of the reference sequences across 8 segements

Segment	H1N1pdm	H3N2
PB2	CY266350	KJ609203
PB1	FJ966978	KJ609204
PA	CY121685	KJ609205
HA	FJ966974	KJ609206
NP	CY121683	KJ609207
NA	CY266193	KJ609208
MP (M1 + M2)	FJ969527	KJ609209
NS (NS1 + NS2or NEP)	CY121684	KJ609210

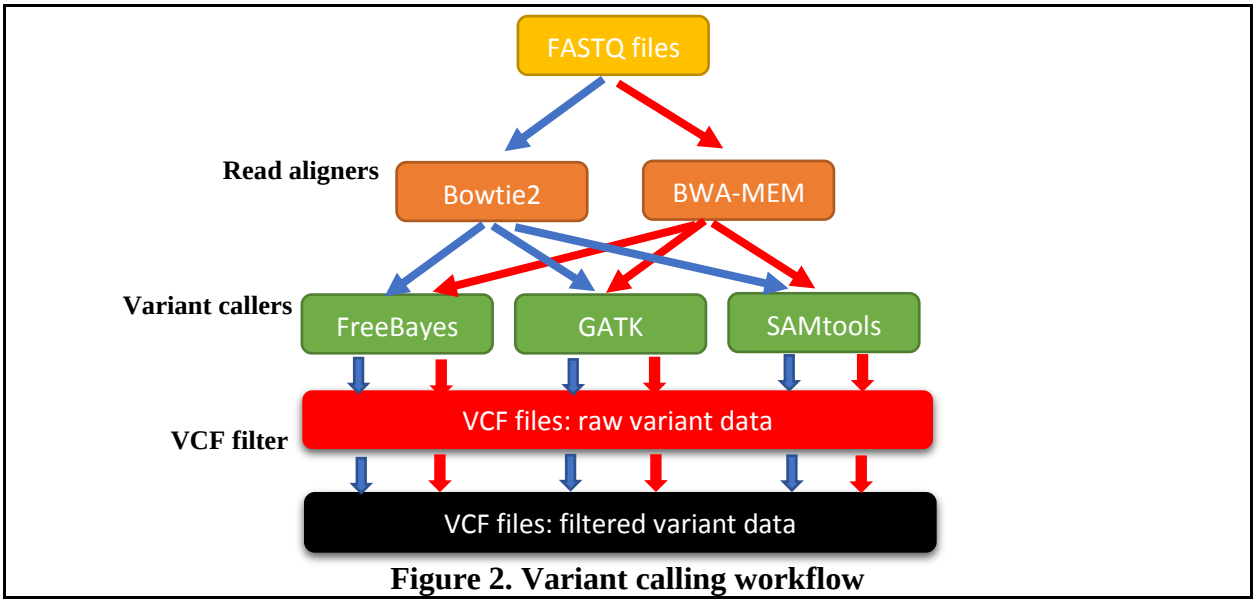

Figure 2. Variant calling workflow

2.2. Alignment

As a result of sequencing, a total of 142 H1N1pdm (**Table 1**) and 107 H3N2 of FASTQ files were obtained. Sequence reads were trimmed using a quality score cutoff of Q20 using Trimmomatric [19] to reduce the noise in detecting SNVs. Pair by pair, the entire FASTQ files were first aligned to the reference sequences (A/Perth/16/2009 H3N2 or A/California/07/2009 H1N1) (FASTA format) using the two different read aligners: Bowtie2 (version 2.3.2); Burrows-Wheeler Aligner (BWA) (version 0.7.15) with default parameters [11], [10]. The reference sequences were indexed by Bowtie2 and BWA and their dictionary files were also prepared

using Picard-tools (version 2.8.2) (http://broadinstitute.github.io/picard/) before performing alignment of the FASTQ files (**Fig. 2**).

2.3. Variant calling

2.3.1. Variant-calling using FreeBayes (version 1.1.0)

The sequence alignment/map (SAM) format files obtained independently via two different aligners were converted to binary alignment map (BAM) format files and sorted by Samtools [14]; the duplicate reads of the sorted BAM files were marked by Picard-tools to avoid over-representation of sequence regions that were preferentially amplified during the PCR reaction. After indexing the processed BAM files with Samtools, FreeBayes [12] was used to identify raw variants from the BAM files. The variant call format (VCF) files obtained were then hard-filtered with the vcflib package [20] with a depth of coverage (DP) > 200 and a Phred quality score (QUAL) > 20).

2.3.2. Variant-calling using Genome Analysis Toolkit (GATK; version 3.8)

After performing alignments with default parameters, SAM format files were obtained, and were then sorted and converted to indexed BAM format files using Picard-tools (version 2.8.2). Picard-tools was used again for marking PCR duplicate reads so that they could be ignored in the downstream GATK analysis. During the sequencing process, identical DNA fragments can be sequenced multiple times. Thus, by marking duplicate reads, the variant caller can exclude uninformative reads from its calculation; otherwise those could provide additional evidence for or against a putative variant [13]. After adding read group tags to the duplicate-marked BAM files using Picard, the BAM files were re-aligned by GATK IndelRealigner to reduce artifacts created during the initial alignment. With the use of the GATK haplotype caller,

the variant call format (VCF) files were created containing raw variants for each sample across the viral genome. Finally, all the vcf files were filtered by the vcffilter function of the vcflib package [20] to reduce the number of false positive calls. Only the qualified variants passing the threshold whose depth of coverages are greater than 200 and Phred quality scores are greater than 20 were included in the final vcf files.

2.3.3. Variant-calling using Samtools (version 1.3.1)

The results from two aligners were converted to the BAM format and sorted using Samtools [14]; Picard-tools marked the duplicate reads for each BAM file. After building the BAM index, the 'mpileup' command in Samtools was executed for the processed BAM files to identify single nucleotide variants (SNV) and indels; the command yielded binary call format (BCF) files containing the raw variant information. The bcf files were then converted to the vcf format, which is human readable using BCFtools [21] and hard-filtered by the vcflib package [20] with the same threshold used for the previous two pipelines.

3. Identification of the concordantly/uniquely detected variants

As a result of variant calling, total 12 datasets (6 H1N1pdm and 6 of H3N2 variant datasets) of filtered vcf files were generated. For more detailed calculation of statistics and comparison, the information on insertions and deletions (indel) was extracted across all vcf files using a function, 'SelectVariants' of GATK software, and saved as an independent file (vcf format). Then, using a function, 'VariantToTable', all the vcf files were regularized to tabular

formatted text files. Concordant/discordant variants between aligners and among variant callers were identified by applying the following linux commands.

(1) Concordance between the two files
awk 'NR==FNR{a[$1 FS $2 FS $4]++;next}a[$1 FS $2 FS $4] ' <file1> <file2> > <output>

(2) Discordance between the two files
awk 'NR==FNR{a[$1 FS $2 FS $4]++;next}!a[$1 FS $2 FS $4] ' <file1> <file2> > <output>

* $1: segment column; $2: nucleotide position column; $4: alternative allele column

The total number of variants and the number of concordantly/discordantly detected variants were counted sample by sample, merged and saved as comma separated values (CSV) format files. With those csv files, barplots showing the total number of variants per pipeline and venn diagrams showing variant concordance between pipelines were drawn by R (version 3.4.2) by utilizing libraries: ggplot2 , VennDiagram, gridExtra [22]–[24].

4. Statistical test

To test whether there were differences in the means of the Phred quality scores and the depth of coverage between the groups of variants concordant across variant calling pipelines and the groups of variants uniquely discovered by individual variant calling pipelines, one-way ANOVA tests were performed and Tukey's HSD (honest significant difference) tests (95% confidence interval) were applied to the results of the ANOVA tests. All statistical tests were carried out using R (version 3.4.2).

Results

1. Alignment results

Table 4. Alignment statistics of the reads (SD=standard deviation; 'properly paired' indicates the paired-reads mapped to the same reference sequence.)

Dataset	Aligner	Mean % of mapped reads (SD%)	# of mapped reads	Mean % of the properly paired reads (SD%)	# of properly paired reads
H1N1pdm	Bowtie2	78.76(11.07)	1426199	65.94(13.14)	1186697
	BWA (BWA-MEM)	92.58(9.60)	1723811	85.50(10.84)	1554669
H3N2	Bowtie2	89.55(11.16)	286197	80.15(10.48)	256225
	BWA (BWA-MEM)	98.59(5.21)	337000	97.10(5.19)	320348

In our comparison of BWA and Bowtie2, two popular alignment tools, BWA (and more specifically, BWA-MEM) performed better to align reads against the reference sequences, regardless of datasets (**Table 4**). BWA yielded a higher percentage of successfully mapped reads and properly paired alignment categories for H1N1pdm and H3N2 data.

2. Total number of variants

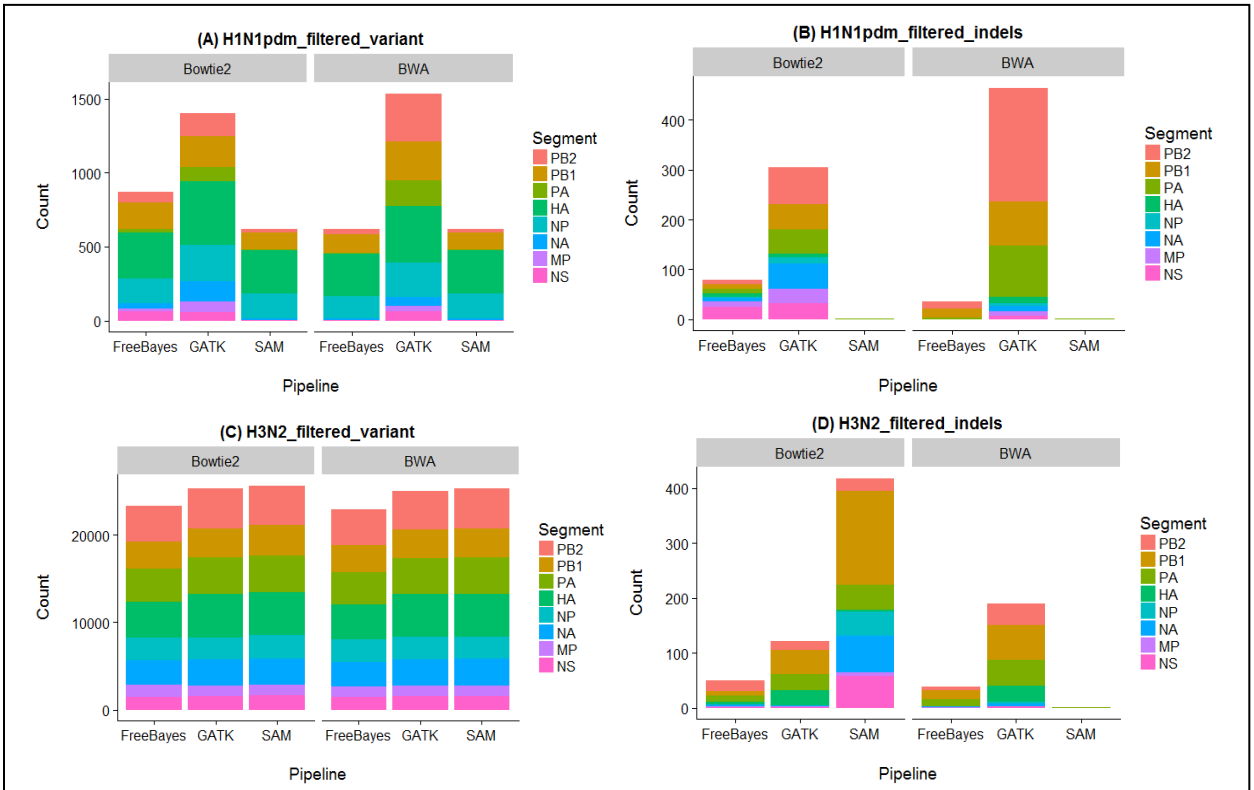

Figure 3. Total number of variants and indels across the viral genomes detected by 6 variant calling pipelines

The total number of variants detected varied according to the pipeline applied. From the H1N1pdm data, regardless of types of aligners, the GATK pipeline identified more variants than any other variant caller. As shown on **Fig. 3A**, the BWA-GATK pipeline succeeded in finding a total of 1,537 variants (mean per sample is 10.4±6.0) while the Bowtie2-GATK pipeline found 1,400 (mean per sample 9.9±4.2) variants. Both FreeBayes and SAMtools discovered the same number of variants – 621 (mean per sample 4.7±1.1)) with BWA as the aligner, which is only 40 ~ 44% of the total number of variants detected with the GATK pipelines (**Fig. 3A**). Although the variant calling results were similar for both SAMtools pipelines irrespective of aligners used,

FreeBayes performed better when it was combined with Bowtie2 than BWA in terms of detecting more variants. Specifically, the Bowtie2-FreeBayes pipeline found 250 more variants than BWA-FreeBayes. However, for the H3N2 dataset, the performance of the pipelines was not consistent in terms of number of variants detected as compared to the results from the H1N1pdm data. For the H3N2 data, the pipelines using SAMtools as their variant caller found more variants than the other pipelines regardless of types of aligners combined (**Fig. 3C**). Each pipeline discovered 25,670 (mean per sample 240.0±23.5) (Bowtie2-SAMtools) and 25,336 (mean per sample 236.8±22.2) (BWA-SAMtools). The capacity of the two pipelines that used GATK as their variant caller was lower, at 98~99% of the total number of variants detected by the SAMtools pipelines. The pipelines using FreeBayes as their variant caller showed a relatively lower performance than the other pipelines (~91% of the results of SAMtools and ~90% of the results of GATK). FreeBayes when combined with Bowtie2 also performed better than when it was used with BWA, as seen for the H1N1pdm data (**Fig. 3A and 3B**).

From the perspective of discovering indels, the two pipelines using GATK were consistently better at detecting both insertions and deletions than the other pipelines, except in one case (**Fig. 3B and 3D**). Specifically, Bowtie2-GATK and BWA-GATK pipelines detected 305 (mean per sample 2.7±2.1) and 465 (mean per sample 4.4±3.9) indels, respectively, in the H1N1pdm data. This accounts for ~21% and ~30% of the total number of variants discovered by each of pipeline. From the H3N2 data, they detected 121 (mean per sample 2.3±1.7) and 189 (mena per sample 3.2±2.5) indels, respectively, which accounts for 0.4 and 0.7% of the total variants detected in that dataset. The SAMtools pipelines detected the fewest number of indels (1, 0.1% of the total variants) in the H1N1pdm data. However, the pipelines showed a highly different performance in detecting indels in the H3N2 data depending on their read aligner.

While only 1 indel was reported via the BWA-SAMtools pipeline, the Bowtie2-SAMtools pipeline succeeded in finding a total of 418 (mean per sample 4.5±3.0, 1.6%) indels from the data, which is substantially more than for any of the other 5 combinations of pipelines. The results obtained through the FreeBayes pipelines showed a different trend depending on the dataset. In the H3N2 data, each of the FreeBayes pipelines succeeded in detecting a comparable number of indels: 49 (mean per sample 1.3±0.6, 0.2%) with Bowtie2-FreeBayes, and 39 (mean per sample 3.1±0.8, 0.1%) with BWA-FreeBayes, which is not significantly different. However, the Bowtie2-FreeBayes pipeline detected 78 (mean per sample 1.5 ± 0.6, 9%) indels which is more than double the number of indels detected by the BWA-FreeBayes pipeline. BWA-FreeBayes detected only 35 (mean per sample 1.4 ± 0.7, 5.6%) in the H1N1pdm data (**Fig. 3B** and **3D**).

3. Concordance between pipelines

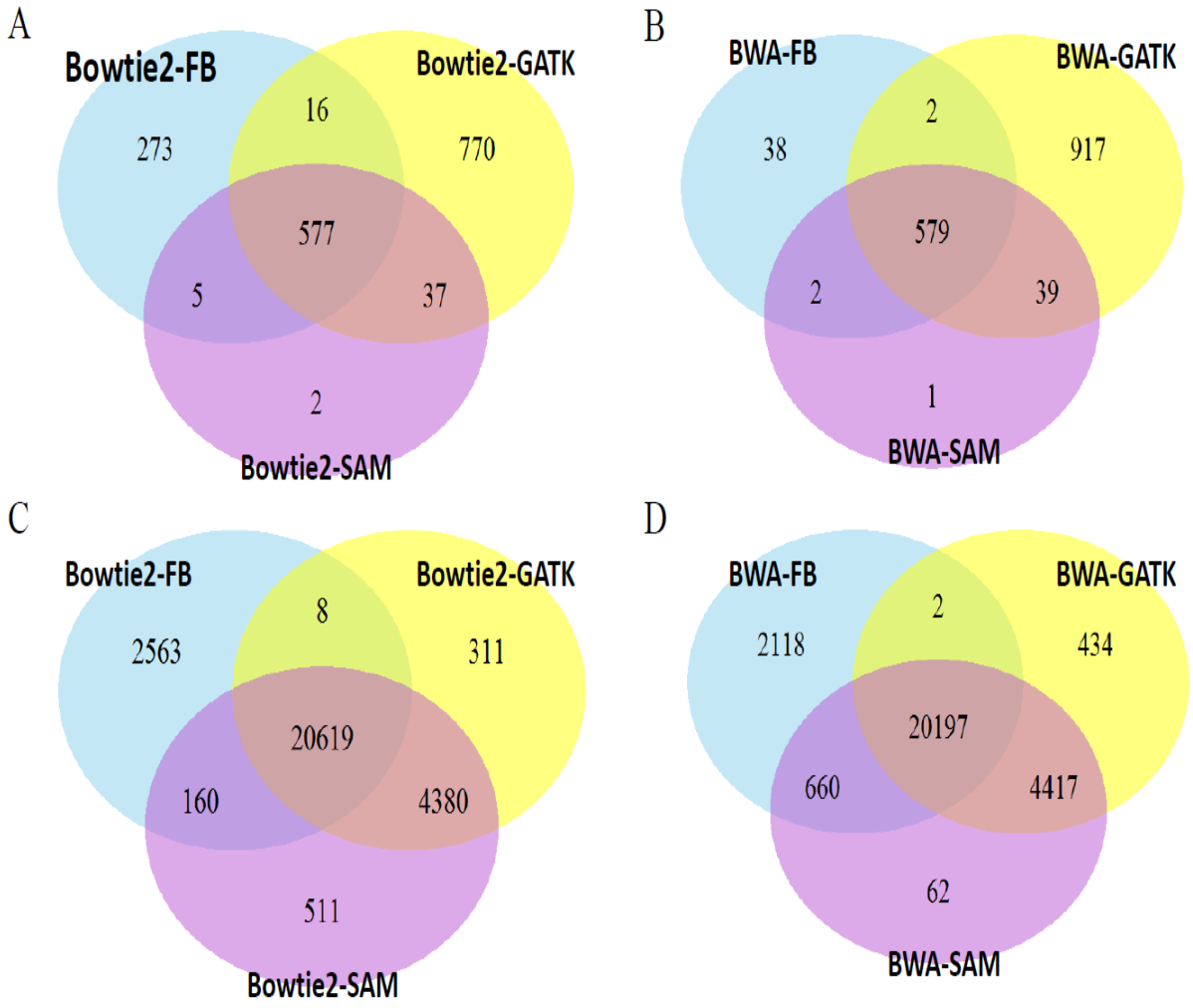

Figure 4. Total variant concordance among the variant callers using identical read aligners for A and B H1N1pdm; for C and D H3N2

Table 5. Variant concordance percentages among variant callers and total number of variants of the union of the three pipelines using the same read aligner

Subtype	Aligner	FB-only (%)	GATK-only (%)	SAM-only (%)	FB∩ GATK (%)	FB∩ SAM (%)	GATK∩ SAM (%)	All (%)	Total variant (count)
H1N1 pdm	Bowtie2	16.25	45.83	0.12	35.30	34.64	36.55	34.35	1680
	BWA	2.41	58.11	0.06	36.82	36.82	39.16	36.69	1578
H3N2	Bowtie2	8.98	1.09	1.79	72.24	72.78	87.56	72.22	28552
	BWA	7.59	1.56	0.22	72.42	74.78	88.25	72.42	27890

I assessed the concordance among the variant calling pipelines to test how consistently the variants were detected by multiple pipelines and how frequently the variants were uniquely discovered by a single specific pipeline. The comparisons were performed at two different levels: 1. Concordance among variant callers with the use of the same read aligner; 2. Concordance between different pipelines comprised of different aligners but with the same variant callers. From the first level comparison, variant concordance ratio between variant callers was low in the H1N1pdm data compared to the H3N2 data. Among the groups using Bowtie2, only 577 variants were consistently detected by the 3 pipelines (**Fig. 4A** and **Table 5**). This accounts for ~34% of the union (Bowtie2-FreeBayes ∪ Bowtie2-GATK ∪ Bowtie2-SAMtools) (1,680) of the variants. Among the BWA methods, 579 of 1,578 variants (BWA-FreeBayes ∪ BWA-GATK ∪ BWA-SAMtools) were detected, which accounts for ~36% of the union (**Fig. 4B** and **Table 5**). The percentage of variants detected by the two pipelines were between 34~39% of the union in both groups (Bowtie2 and BWA) (**Table 5**). Such wide variant discordance was caused by the number of variants uniquely detected by GATK, regardless of the aligner. From the Bowtie2 group, the GATK pipeline uniquely reported 770 variants, which is ~46% of the total number of variants found by Bowtie2 pipelines (**Fig. 4A** and **Table 5**). In the BWA group, 917 variants were uniquely detected by the GATK pipeline, which accounts for 58% of the union(**Fig. 4B** and **Table 5**). If the sets of variants uniquely detected by GATK in both groups were excluded, each variant concordance ratio goes up to 66% (Bowtie2-FreeBayes ∩ Bowtie2-SAMtools) and 94% (BWA-FreeBayes ∩ BWA-SAMtools).

In the H3N2 data, the variants were more consistently discovered by multiple variant-callers in both aligner groups. The concordance ratio was approximately twice as high as those seen in the H1N1pdm data. Specifically, more than 72% of the variants in both groups were

commonly detected by the three variant calling pipelines. Also, the variant concordance ratio between GATK and SAMtools was the highest of all in both groups (Bowtie2 group: 87%; BWA group: 88%) (**Table 5**). Unlike the results from the H1N1pdm variant data, the variants uniquely detected by FreeBayes outnumbered those uniquely detected by GATK and SAMtools in both groups. In the Bowtie2 group, FreeBayes uniquely detected 2,563 variants, ~9% of the union variants, which is ~5 times as high as the number (511) of variants detected only by SAMtools and ~8 times as high as the number (311) of variants uniquely detected by GATK (**Fig. 4C** and **Table 5**). In the BWA group, 2,118 variants were uniquely discovered by FreeBayes, which is 7.5% of the union, ~5 times as high as the number (434) of variants detected only by GATK, and ~34 times as much as the number (62) of variants uniquely discovered by SAMtools (**Fig. 4D** and **Table 5**).

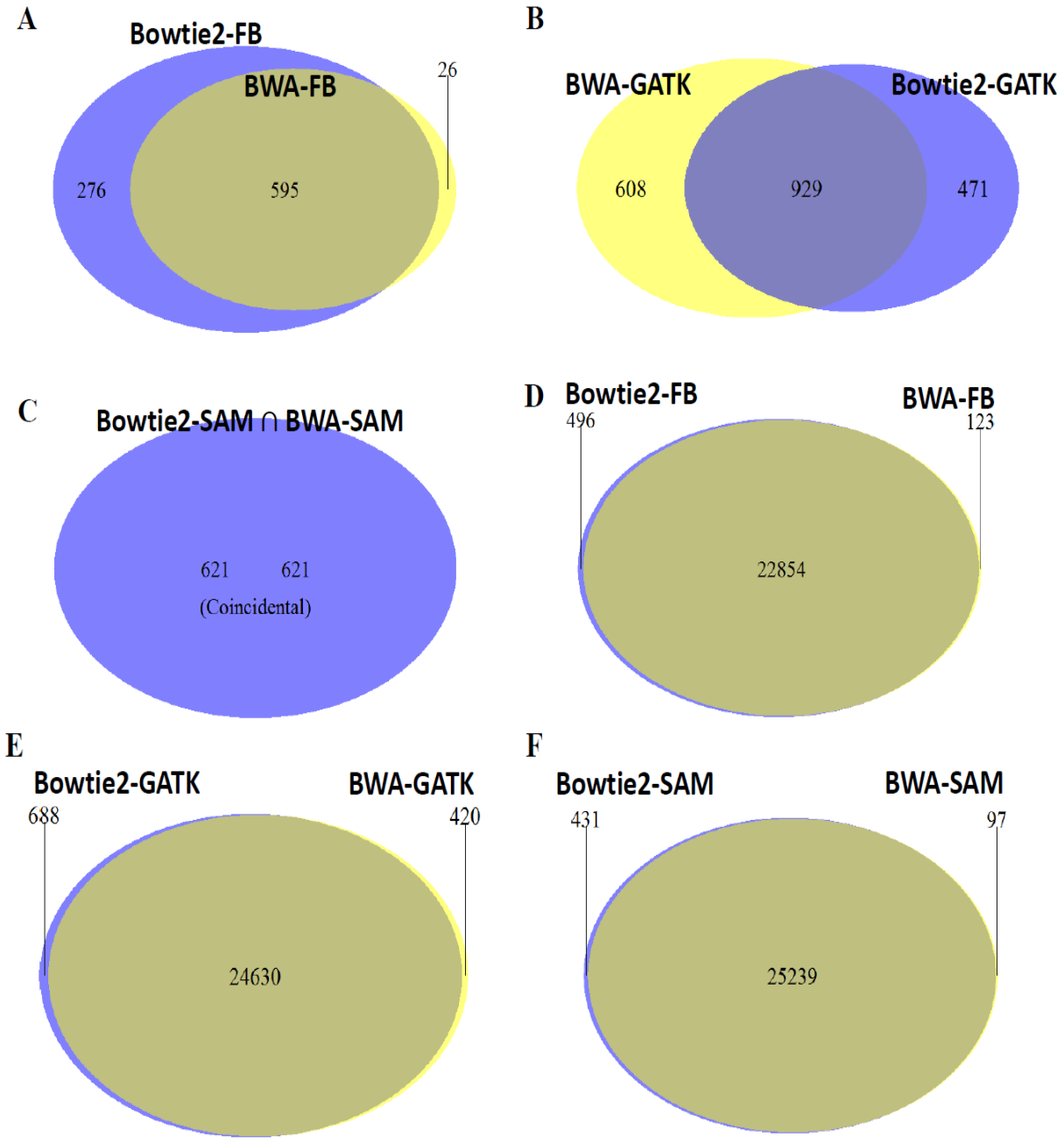

Figure 5. Variant concordance between the pipelines using identical variant caller (A ~ C: H1N1pdm and D ~ F: H3N2)

Table 6. Variant concordance percentages between the pipelines using identical variant-caller and the total number of variants of the union (Bowtie2∪BWA)

		Bowtie2-only (%)	BWA-only (%)	Bowtie2∩BWA (%)	Bowtie2∪BWA (Total count)
H1N1pdm	FreeBayes	30.77	2.90	66.33	897
	GATK	23.46	30.28	46.26	2008
	SAMtools	0	0	100	621
H3N2	FreeBayes	2.11	0.52	97.36	23473
	GATK	2.67	1.63	95.70	25738
	SAMtools	1.67	0.38	97.95	25767

Variant concordance was also assessed between the pipelines using identical variant callers. Overall, the concordance ratio, regardless of variant callers, was higher than in the comparison among variant callers. The pipelines using SAMtools, regardless of type of aligners used detected identical variants (621) (**Fig. 5C**). None of the variants were uniquely detected by either of the SAMtools pipelines (**Table 6**). In the H1N1pdm data, there were 595 variants concordantly detected by both FreeBayes pipelines, which accounts for ~66% of the union (897) (**Fig. 5A** and **Table 6**). As demonstrated in the Venn diagrams in **Fig. 4A** and **Table 5**, FreeBayes detected far more variants in the H1N1pdm data when it is combined with the aligner Bowtie2. In this comparative analysis, the Bowtie2-FreeBayes pipeline succeeded in finding 276 unique variants, which is 30% of the total variants detected in the H1N1 data by the two FreeBayes pipelines. However, only 26 (~3% of the union) were uniquely detected by the BWA-FreeBayes pipeline (**Fig. 5A** and **Table 6**). Both GATK pipelines detected a total of 2008 variants in the H1N1pdm data, which is higher than the union of FreeBayes pipelines and SAMtools pipelines (**Table 6**). In addition, 929 variants were consistently reported by both pipelines (**Fig. 5B**), which also outnumbers the total number of concordant variants (595 - between FreeBayes pipelines; 621 - between SAMtools pipelines) (**Fig. 5A** and **Fig. 5C**). However, the concordance ratio between GATK pipelines is only ~46% because each pipeline detected a relatively large number of unique variants (**Table 6**). Particularly, Bowtie2-GATK

discovered 471 unique variants, which is ~23% of the union. BWA-GATK outperformed Bowtie2-GATK by uniquely detecting 608 variants accounting for ~30% of the union (**Fig. 5B** and **Table 6**). For the H3N2 data, the concordance ratio between the pipelines was high considering the numbers from the H1N1pdm variant concordance statistics. Specifically, a total of 22,854 variants were consistently detected by the two FreeBayes pipelines accounting for 97% of the union (**Fig. 5D** and **Table 6**); GATK pipelines succeeded in finding 24,630 variants simultaneously, which is ~95% of the union (**Fig. 5E** and **Table 6**); and 25,239 variants were detected by all of the pipelines using SAMtools accounting for ~97% of the union. Although the frequencies were relatively low (0.3 ~2.6%), hundreds of variants were uniquely detected by each single pipeline. As shown in **Fig. 5D** Bowtie2-FreeBayes reported ~500 unique variants while BWA-FreeBayes called 123 unique variants. In comparison, Bowtie2-GATK reported 688 and BWA-GATK 420 unique variants (**Fig. 5E**). Although there were no unique variants detected by the SAMtool pipelines in the H1N1pdm data, Bowtie2-SAMtools detected 431 and BWA-SAMtools unique 97 variants in the H3N2 data (**Fig. 5F**). Interestingly, when the same variant caller was selected, the pipelines using Bowtie2 as their aligner outperformed BWA pipelines in terms of detecting more unique variants in the H3N2 data.

Discussion

In this study, I compared multiple variant calling pipelines using NGS data of influenza A viruses (H1N1pdm and H3N2). Using the genomic variant information obtained through 6 different pipelines, I did a comparative analysis on the performance of the pipelines. I first compared the two read aligners selected for this study (BWA and Bowtie2) because they are 2 of the most popular open source software available. As the alignment statistics indicate (**Table 4**), BWA clearly outperformed Bowtie2 in alignment results regardless of datasets. Such a

difference can be explained by the algorithmic steps used. BWA-MEM is equipped with more sophisticated steps which possibly increases the success of the alignment. More specifically, to reduce the number of mismapping events caused by occasional mismatches between the reference sequence and the read of a sample, BWA-MEM uses a process of re-seeding while Bowtie2 performs seeding at the initial stage of the alignment [11], [10]. Seeding is a process of finding matched regions between a query sequence (a reference sequence) and a hit sequence (reads of a sample). By repeating the process, BWA-MEM increases the number of matches. Also, in pair-end mode, BWA-MEM, unlike Bowtie2, takes a local alignment step in case one of the two ends of the paired-end read should be recovered [10], which may increase the success of pair-end alignment of the reads against reference sequences. Li et al. [11] evaluated the performance of BWA-MEM by comparing it to other aligners and reported that it is more performant in terms of accuracy than Bowtie2 for aligning longer reads (> 70bp) due to its advanced seeding algorithm [11]. Accepting the robustness of its read-mapping algorithm, the GATK development team, the Genome Sequencing and Analysis (GSA) group at the Broad Institute has recommended using BWA as a read aligner for the GATK variant calling pipeline [13]. However, BWA does not always guarantee better performance of pipelines in variant detection. For example, even though the BWA-GATK pipeline outperformed any other pipelines that used Bowtie2 in detecting indels, FreeBayes combined with Bowtie2 provided better results than with BWA, regardless of types of data. This is because of its variant detection mechanism which does not overly rely on precise alignment results. FreeBayes is a haplotyped-based variant detector which discovers variants based on the literal sequences of reads aligned to a particular target, not their precise alignment which is used by SAMtools and GATK [12]. Thus, it is

probable that the BWA alignment results, although highly precise, does not much help variant detection for the BWA-FreeBayes pipeline.

In the comparison of the total number of variants among the 6 pipelines tested, the pipelines using GATK succeeded in detecting relatively a higher number of variants regardless of type of aligners combined. Also, the pipelines generally outperformed the others in detecting indels. This can be explained by the algorithm function 'HaplotypeCaller' which was used in the two GATK pipelines at variant calling stage. GATK-HaplotypeCaller detects SNVs and indels at the same time based on 'local *de novo* assembly of haplotypes', which means that as the program meets a certain region showing genetic variation, it gets rid of the previous mapping information and the reads within the region are completely reassembled. This algorithm allows the GATK-HaplotypeCaller to detect variants, particularly indels, more accurately in some regions where different types of variants are present at high frequency [13]. Thus, its unique algorithm may affect the high performance of GATK pipelines in detecting indels.

Surprisingly, compared to the other 5 pipelines, Bowtie2-SAMtools detected a significantly larger number of indels in the H3N2 data (**Fig. 3D**). This result is unusual considering the vulnerability of SAMtools in detecting indels, as previously demonstrated. In a study targeting human NGS data, it displayed the lowest performance in detecting insertions and deletions of all pipelines [8] and all of the pipelines using SAMtools detect a lower number of indels than the other pipelines in the this study too except the case reported in **Fig. 3D**. Most were found to be uniquely identified by Bowtie2-SAMtools and the mean Phred quality score (136.58) and the depth of coverage (227.70) were significantly lower than those of the other SNVs. However, those indels cannot be regarded as false positive calls because they already passed a filter (Phred quality score > 20 and depth of coverage > 200) at the last stage of the

variant calling pipeline. If those were caused either by sequencing error or by poor alignment results, those indels should have been detected by the other pipelines. However, aforementioned those were mostly verified only by Bowtie2-SAMtools pipeline. Thus, it might be a SAMtools-specific genotyping errors. Further computational analysis should be followed to determine whether those indels are true positives or not.

From the variant concordance analysis of the pipelines, I found that not all of the variants are consistently detected by different pipelines and such discordance is influenced more by the variant callers rather than the read aligners. The variant concordance ratio was higher between the pipelines using identical variant callers than among the variant callers using the same read aligner (**Table 5** and **Table 6**). Moreover, in order to test whether there are differences in the means of the quality scores and the depth of coverage between concordantly detected variants and uniquely detected variants, one-way ANOVA tests were conducted. However, the results showed that their Phred quality scores and the depth of coverages are not significantly different, which indicates that sequence read quality is not a factor in variant concordance or discordance. Through the variant concordance analyses, it became evident that a large number of variants were exclusively detected by a single pipeline. Although the concordance ratio was higher in the H3N2 data, there still were hundreds of unique variants detected by a single pipeline. The sequence qualities were high enough to pass the hard filter. Thus, in the analysis of influenza virus variants, to obtain sufficient genomic variant information and minimize missing variant information, it is recommended to combine high confident call-sets detected by multiple variant calling pipelines.

Nevertheless, there is a limitation in this study. No validation test has been performed on the variants discovered. Even though all the variants identified in this study had high Phred

quality scores and depth of coverages, we could not determine whether any of these were erroneous calls. There could be a systematic error occurring in any step of the sample preparation, sequencing, and even genotyping by the pipelines which possibly impacted the false discovery of the variants or failure of variant detection. Previous studies targeting human NGS data tried minimizing such erroneous calls by using gold-standard sequence information. However, there is no equivalent data for influenza. Thus, to reduce false positive (and negative) calls, further computational method should be developed that are optimized on influenza data.

Reference

[1] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, "Systematic comparison of variant calling pipelines using gold standard personal exome variants," *Sci. Rep.*, vol. 5, no. December, pp. 1–8, 2015.

[2] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nat. Rev. Genet.*, vol. 12, no. 11, pp. 745–755, 2011.

[3] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nat. Rev. Genet.*, vol. 11, no. 6, pp. 415–425, 2010.

[4] C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow, "Lactose digestion and the evolutionary genetics of lactase persistence," *Hum. Genet.*, vol. 124, no. 6, pp. 579–591, 2009.

[5] X. Liu, S. Han, Z. Wang, J. Gelernter, and B. Z. Yang, "Variant Callers for Next-Generation Sequencing Data: A Comparison Study," *PLoS One*, vol. 8, no. 9, pp. 1–11, 2013.

[6] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon, "Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing," *Genome Med.*, vol. 5, no. 3, 2013.

[7] S. Van den Hoecke, J. Verhelst, M. Vuylsteke, and X. Saelens, "Analysis of the genetic

diversity of influenza A viruses using next-generation DNA sequencing," *BMC Genomics*, vol. 16, no. 1, pp. 1–23, 2015.

[8] S. Sandmann, A. O. De Graaf, M. Karimi, B. A. Van Der Reijden, E. Hellström-Lindberg, J. H. Jansen, and M. Dugas, "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data," *Sci. Rep.*, vol. 7, pp. 1–12, 2017.

[9] A. Cornish and C. Guda, "A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference," *Biomed Res. Int.*, vol. 2015, 2015.

[10] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[11] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," vol. 0, no. 0, pp. 1–3, 2013.

[12] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," pp. 1–9, 2012.

[13] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*, no. SUPL.43. 2013.

[14] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[15] A. Cornish and C. Guda, "A Comparison of Variant Calling Pipelines Using Genome in a

Bottle as a Reference," *Biomed Res. Int.*, vol. 2015, 2015.

[16] L. L. M. Poon, T. Song, R. Rosenfeld, X. Lin, M. B. Rogers, B. Zhou, R. Sebra, R. A. Halpin, Y. Guan, A. Twaddle, J. V. DePasse, T. B. Stockwell, D. E. Wentworth, E. C. Holmes, B. Greenbaum, J. S. M. Peiris, B. J. Cowling, and E. Ghedin, "Quantifying influenza virus diversity and transmission in humans," *Nat. Genet.*, vol. 48, no. 2, pp. 195–200, 2016.

[17] A. Djikeng, R. Halpin, R. Kuzmickas, J. DePasse, J. Feldblyum, N. Sengamalay, C. Afonso, X. Zhang, N. G. Anderson, E. Ghedin, and D. J. Spiro, "Viral genome sequencing by random priming methods," *BMC Genomics*, vol. 9, pp. 1–9, 2008.

[18] B. Zhou, M. E. Donnelly, D. T. Scholes, K. St. George, M. Hatta, Y. Kawaoka, and D. E. Wentworth, "Single-Reaction Genomic Amplification Accelerates Sequencing and Vaccine Production for Classical and Swine Origin Human Influenza A Viruses," *J. Virol.*, vol. 83, no. 19, pp. 10309–10313, 2009.

[19] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[20] E. Garrison, "A simple C++ library for parsing and manipulating VCF files.," *https://github.com/vcflib/vcflib*.

[21] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin, "BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data," *Bioinformatics*, vol. 32, no. 11, pp. 1749–1751, 2016.

[22] H. Wickham, "ggplot2:Glegant Graphics for Data Analysis," *Springer-Verlag New York*,

2009.

[23] H. Chen, “Generate High-Resolution Venn and Euler Plots,” 2018.

[24] B. Auguie and A. Antonove, “Miscellaneous Functions for Grid Graphics,” 2017.