Songklanakarin J. Sci. Technol.
43 (6), 1726-1735, Nov. - Dec. 2021

http://www.sjst.psu.ac.th

Original Article

Establishing an automated graphical genome analysis platform♣

Wison Laochareonsuk[1], Komwit Surachat[2], and Surasak Sangkhathat[3*]

[1] *Department of Biomedical Science, and Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, 90110 Thailand*

[2] *Department of Computing Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, 90110 Thailand*

[3] *Department of Surgery and Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, 90110 Thailand*

Received: 11 January 2021; Revised: 20 May 2021; Accepted: 1 June 2021

Abstract

Precision medicine is a modern health concept which involves using a specific treatment for individuals based on genetic background. This new approach not only concerns personalized prescriptions to minimize adverse events and targeted therapy but also extends to diagnosis, prognosis and prevention of relevant diseases. Emerging high-throughput sequencing technology allows widespread identification of genetic variations. However, the downstream workflow of sequencing raw data is based on command-line interface (CLI) and open-source software, and general users may be confronted with the difficulty of using the basic CLI language and constructing pipelines. Our project aimed to establish a fully automated web-based analysis system for cancer genomic medicine and other human genomic fields to query and visualize genomic data**.** The web-based analysis for cancer biomedicine named iCBC was prepared using a user-friendly graphical interface on a high-performance computing system**.** We also implemented various bioinformatic tools and assembled automated pipelines for genomic sequencing**.** The implemented workflows included quality control, alignment and pre-processing, variant discovery, annotation and prioritization with multiple filtering criteria. Ultimately, the iCBC was developed, which is an open access online analysis platform for human high-throughput genomic data in biomedical cancer research.

Keywords: cancer research, medical bioinformatics, online analysis platform

1. Introduction

Precision medicine has become a modern healthcare paradigm in the new era involving specific medical treatments based on individual lifestyle and genetic background (Goetz & Schork, 2018). This novel approach not only involves personalized treatment to decrease adverse events and targeted therapeutic effects but also influences solving undiagnosed diseases, predicting outcomes and encountering for preventable diseases to improve the individual's quality of life. The concept of genomic medicine was introduced after the fully developed human genome projects and 1000-genome project were published in the 2010s which provide a reference for human genetic mapping and exploring the variations in regions of specific interest (The Genomes Project *et al.*, 2015). Nowadays, emerging high-throughput sequencing technology is enhancing the power of genetic studies. Because of high-capability sequencing, this platform allows clinicians and researchers to extend their identification of disease pathogenesis, application for diagnosis of diseases, selection of suitable treatments and provision of better prognoses. Together with the rapid decline of sequencing costs, next generation sequencing has become a current application for biomedical research that delivers novel medical knowledge worldwide (Nambot *et al.*, 2018; Warr *et al.*, 2015).

♣Peer-reviewed paper selected from The 9th International Conference on Engineering and Technology (ICET-2021)
*Corresponding author
Email address: surasak.sa@psu.ac.th

As precision medicine becomes a greater part of healthcare systems, with the emergence of sequencing data, including whole exome sequencing or targeted sequencing, the resources available for this work face the danger of becoming overwhelmed (He, Ge, & He, 2017). Moreover, downstream analysis process of the data is based on open-source software (i.e., FastQC, Burrows-Wheeler Aligner, Samtools and Genome Analysis Tool Kit), running on command line interface (CLI), which may present challenges to researchers or clinicians due to the difficulty of the basic computer language and workflow assembly. According to immersion of the data, bioinformatic analysis is usually operated on an effective infrastructure system including high performance computing and extensive storage. Moreover, variant prioritization is a crucial step in genomic analysis that requires various criteria to filter out non-related variants and select only possible pathogenic variants (DePristo *et al.*, 2011; Gong *et al.*, 2016).

Therefore, a web-based genomic analysis program for cancer biomedicine named integrative Computational Biology for Cancer (iCBC) was established using a graphical interface on a cloud computing platform. We implemented various bioinformatic tools and assembly analysis pipelines for whole exome sequencing and targeted sequencing. The provided analyses include sequence quality control, reference alignment and pre-processing, variant calling, variant annotation and variant prioritization by multiple filtering criteria for individual sequence analysis. In addition, somatic and comparative germline analyses are also embedded in an integrative sequence analysis, in which the user can select separate conditions for each sample between case-control for germline analysis or matched tumor-normal sample for somatic analysis. Finally, the integrative analysis archives workflows with variant annotation and prioritization, similar to individual sample analysis.

2. Materials and Methods

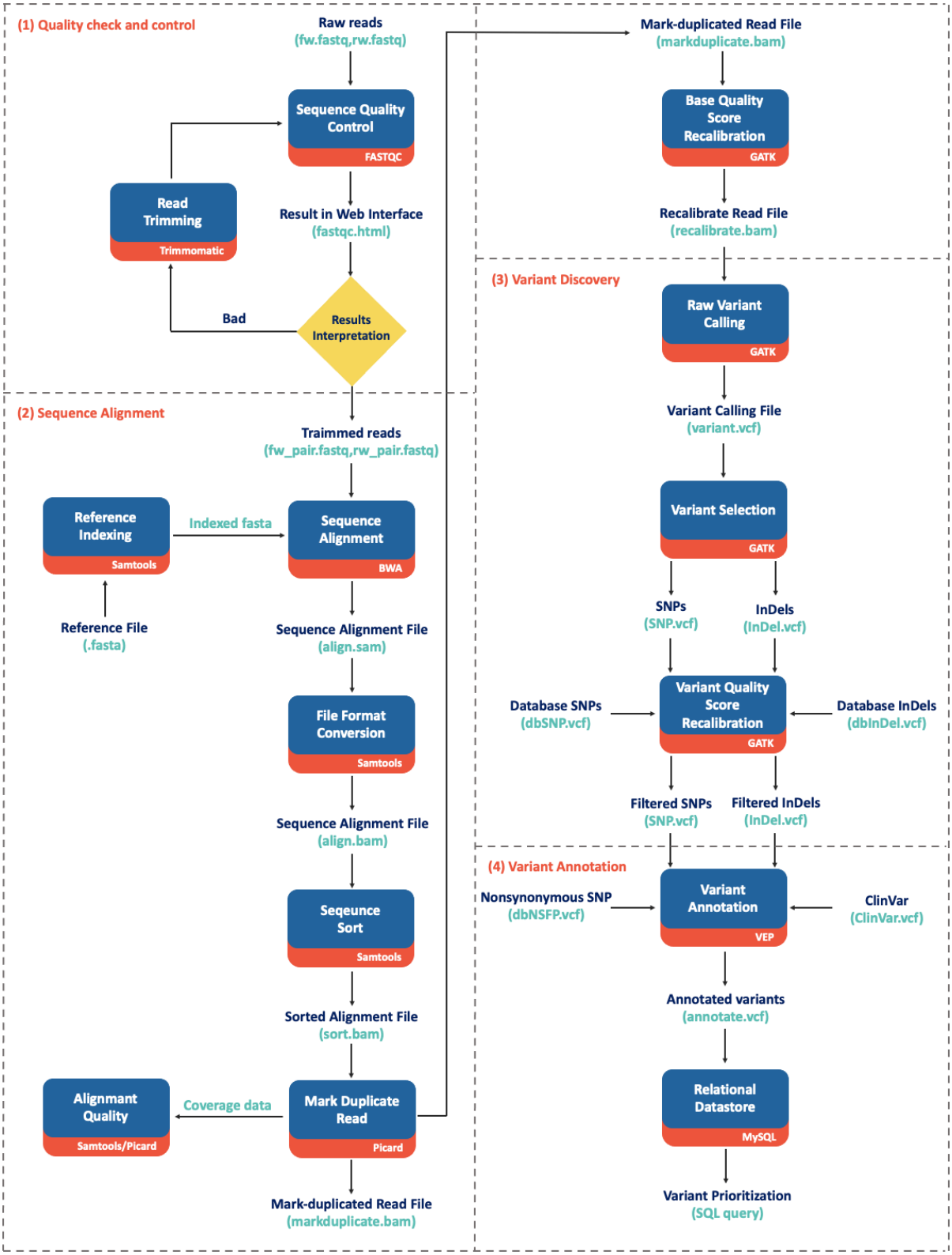
The iCBC was developed as a web-based bioinformatic tool with selectable and automated pipelines for individual and integrative human genomic analysis. The development of the in-house pipeline applied open-source software with the Linux scripting language. Focusing on human genomic studies, short read sequencing has been the most widely used platform for single-end and paired-end reads. The best practice guideline recommends four principle processes for an NGS pipeline including (1) quality control, (2) pre-processing, (3) variant discovery, and (4) functional annotation (Figure 1) (DePristo *et al.*, 2011). The analysis usually starts with a sequence quality check and control followed by trimming out parts with low base quality scores, short sequence lengths and remaining adaptors. Alignment is the next crucial step in which sequences are mapped with a pre-processing reference at the best position. When finished, the data are stored in a format of sequence align map (SAM)(H. Li *et al.*, 2009). Because of the large size of the data, the aligned reads are usually converted to a binary format (binary align map-BAM) and sorted by order of chromosomes and their positions to reduce the time of computation. In general, short read platforms regularly need library construction before sequence reading which may lead to non-random enrichment and cause duplicated sequences. These sequences require re-analysis for duplication and data bias. Systematic base quality score errors in each position are improved by correcting the errors with known variant databases (Tian, Yan, Kalmbach, & Slager, 2016). After pre-processing, the sequence quality is satisfied for the discovery of nucleotides that differ between the read sequences and reference sequences in each position, relying on a diploid assumption. The variants are generally recorded in variant call format (VCF). Alternatively, the variants may be identified in all represented sites (genomic VCF or GVCF) that allows joining multiple analyses together (McKenna *et al.*, 2010). Moreover, somatic variant calling is a special analysis for matched tumor-normal samples which may be interfered with by chromosomal aneuploidy or polyploidy. Some variants may be artifacts and need to be filtered out. Finally, functional annotation is the last important step for variant discovery, which is the clinical implication of the analysis results (Roca, Fernández-Marmiesse, Gouveia, Segovia, & Couce, 2018). Functional alterations resulting from those variants need further prioritization based on the basic characteristics of each disease and variant-confident statistics correlated with the reported database (Ebiki, Okazaki, Kai, Adachi, & Nanba, 2019).

2.1 Quality control

The base quality scores have a probability of incorrect calling of sequencer (Peng *et al.*, 2015). These scores are recorded by a logarithmic scale called the Phred score (-10 log P) (Liao, Satten, & Hu, 2017). The statistical adjustment should be performed before variant discovery by observing any errors with known variants and calculating the adjusted score (Tian *et al.*, 2016). In our tool, quality assessment was powered by FastQC which is suitable for various platforms. The implemented tools not only evaluate base quality score but also estimate sequence length distribution, G-C content distribution, duplication levels and remaining adaptors. Low quality sequences and remaining adaptors will be discarded during the quality check.

2.2 Alignment

Alignment is a process which specifically matches the reading sequences with each reference base position in the reference genome sequence. The Burrows–Wheeler transform, or block-sorting compression, is a rearrangement data algorithm that prepares character strings into similar character runs using a suffix sorting array. The BWT provides a faster and memory efficient method to alignment with reference genome. Although there are many alignment programs based on the Burrows–Wheeler transform (BWT) such as Bowtie or the Burrows-Wheeler Alignment Tool (BWA)(Li & Durbin, 2009) and a short oligonucleotide alignment program (SOAP)(R. Li *et al.*, 2009), the BWA has well-balanced computing performance between speed, memory usage, and accuracy. The mapped reads are commonly recorded in the SAM format and converted to a binary format to minimize storage and computing time. Alignment accuracy is assessed by mapping quality scores generated to indicate the confidence of precise mapping.



Fw: forward, Rw: reverse, SAM: sequence alignment map, BAM: binary alignment map, VCF: variant calling format, SNPs: single nucleotide polymorphisms, InDels: insertions/deletions, SQL: structured query language

Figure 1. Analysis schema of human genomic sequence

2.3 Duplicated reads identification

Second-generation sequencing frequently needs a polymerase chain reaction (PCR)-supplemented preparation. The preparation may lead to enrichment bias by selectively amplifying some regions, which may interfere with the accuracy of the variants by increasing confidence (Ebbert *et al.*, 2016). In this step, duplicated sequences will be marked by using the MarkDuplicates function in the Picard tool which is a widely recommended program.

2.4 Base quality score recalibration

Base Quality Score Recalibration (BQSR) is the last step of data pre-processing before variant discovery by machine learning methods. This process detects systematic errors in the sequencer for each base. Mismatch errors usually occur at the end of a read and are generally found using a sequencing-by-synthesis method (Yu, Yorukoglu, Peng, & Berger, 2015). Recalibration is accomplished by analyzing the covariations among known databases and estimating an adjusted quality or empirical score.

2.5 Variant discovery

Variant discovery or variant calling is an important step of genomic analysis which identifies single nucleotide polymorphisms (SNPs) and insertion-deletions (InDels). In general, variants are discovered by comparing allelic alterations between reads using a reference from a process of local re-alignment. There are widely used variant calling programs including SAMtools (H. Li *et al.*, 2009), the Genome Analysis ToolKit (GATK) (McKenna *et al.*, 2010), and Atlas2 (Challis *et al.*, 2012). GATK seems to provide the highest quality variant identification. To reduce overwhelming of the records, only identified variants together with their general information are collected into a variant calling format (VCF) file (Danecek *et al.*, 2011).

2.6 Filtration

The variants are computationally identified based on the variant calling error score including mapping quality (MQ), fisher strand (FS), strand odds ratios (SOR), the mapping quality rank sum test (MQRankSum) and the read position rank sum test (ReadPosRankSum) (Ruffalo, Koyuturk, Ray, & LaFramboise, 2012). In a basic pipeline, hard filtration is frequently preferred to crudely select true variants by setting a threshold for all variants. However, hard filtration may throw out good variants because of the inflexibility of individual thresholds. The problem is solved by creating a contour cut edge using machine learning of data from known highly validated resources (Omni, 1000 Genomes, HapMap), called variant quality score recalibration (VQSR). VQSR generates a Gaussian mixture model to determine any cluster areas of those variants and pre-calculated truth sensitivity for selecting true variants and filtering out artifacts (DePristo *et al.*, 2011; Pirooznia *et al.*, 2014). After applying designated sensitivity, the artifact variants are filtered out.

2.7 Functional annotation

Variant annotation is the final step for the WES analysis pipeline to get more information on the discovered variants. Functional annotation is a process of applying variant information corresponding with diseases or syndromes of interest (McCarthy *et al.*, 2014). This process correlates data between functional impacts based on genetic alterations in various populations and discovered variants. There are many available annotation tools, e.g., the Mutalyzer (Wildeman, van Ophuizen, den Dunnen, & Taschner, 2008), VariantAnnotator (GATK) (McKenna *et al.*, 2010), SnpEff (Cingolani *et al.*, 2012) and variant effect predictor (VEP) (McLaren *et al.*, 2016). For our pipeline, we used VEP to get more complete information and reduce processing time.

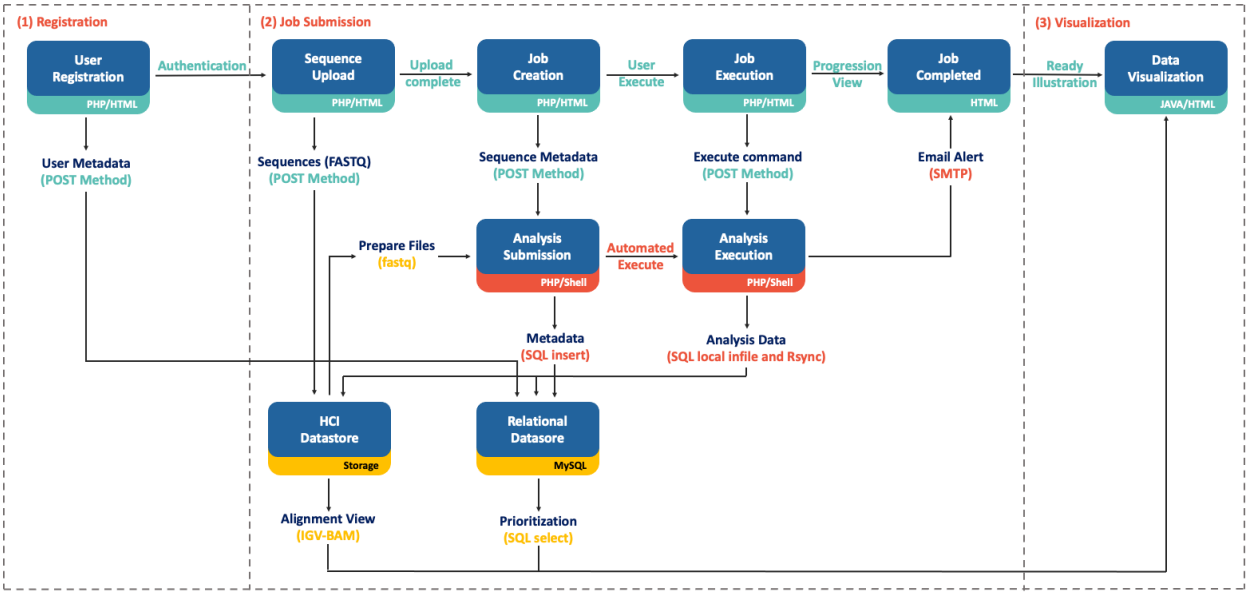
3. Results and Discussion

3.1 Overview

We established the iCBC platform to support user-friendly automated genomic sequence analysis based on an interactive web application. Our platform implemented various bioinformatic tools and constructed analysis workflows for individual sequence and multi-sequence analysis. Registration for the iCBC is free, with no-charge for analysis and basic storage for academic and research purposes. General access and private logins to iCBC are digitally verified for security with an encrypted hypertext protocol certificated by Prince of Songkla University and by generating an encoded key chain. The web-based interface is easy to use for submitting jobs for analysis, setting the parameters of the trimming and filtering processes and visualization of the genomic data using automatic or customized execution. Granted users are allocated storage space of 200 Gigabytes with synchronized duplication. Job creation and sequence submission are private and secure on an implemented open-source cloud storage system. To execute a job, our platform automatically distributes processing resources by using scheduler software on our high-performance computing (HPC) and hyper availability (HA) cluster at the Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University. We distribute 24 core computing units and 64 gigabytes of memory for each job. In general, the system commits to a 2-2.5 hours turnaround time for 9-gigabase whole exome sequences. Additionally, each individual is allowed to download and share archived files on our cloud-based system. The web interface is composed of three layers, registration, job submission and visualization (Figure 2). All computing processes are logged in the representative directory of each analysis which can be tracked back for further information.

3.2 Submission of genome sequence

The job submission allows the user to create an analysis workflow including sequences in the FASTQ or compressed FASTQ format with metadata information. Genomic data is stored in a cloud space providing up to 100 gigabytes before submission. iCBC also supports both single-



HCI: hyperconverged infrastructure, SMTP: simple mail transfer protocol, HTML: hypertext markup language, SQL: structured query language, IGV: integrative genomics viewer, BAM: binary alignment map

Figure 2. Three layers of web-based genomic analysis platform

end and pair-end read formats. To maximize space allocation and speed up the upload process, compressed formats can be submitted for analysis. The metadata information includes the genome reference of the target construction, an illumina sequence adapter, target capture regions and selection of automated analysis.

3.3 Analysis progression

iCBC creates automated and selectable pipelines which execute a job until finished and allows users to customize each step on the pipeline. When a job is created and sequences are submitted, the user can start an analysis by clicking on a suggested button or select a step from the overall analysis box (Figure 3). When started, the progression bar is automatically switched from suggestion to running. In cases of error, the status bar will be marked as false, and the suggestion bar will shift to the previous task. At the end of each step, the bar will move to the next task and the user is allowed to check the results or download relevant files. After the analysis is completely executed, annotated variants are dumped into a MySQL database specified by the job and username. All of the representative files including raw sequences, trimmed sequences, and aligned and annotated files are stored in our provided space and allowed to be downloaded for each until the job is manually deleted by the user.

3.4 Sequence quality control

Raw sequence quality is proven by the FastQC program. FastQC provides fast and accurate quality screening including distribution of base quality score, base and GC content deviation, sequence length distribution, degree of duplication and remaining adapter or sequence index (Figure 3). After the raw data are evaluated, the sequence continues to quality control by the Trimmomatic program that allows the user to trim head and tail bases, screen and disregard low quality sequences, cut remaining adapters, control fragment length and assembly paired-end matched sequences. For the quality control section, users are authorized to assign protocols simply by selecting parameters suitable for their data. Finally, the processed sequences will be automatically qualified and displayed to users again.

3.5 Alignment and statistical matrix

Qualified sequences are aligned with selected genome references using BWA-MEM together with converting from SAM to BAM and sorting by reference position using Samtools view and sort respectively. Before pre-processing alignment, the sequences are split by genomic contigs and their position into 24 chunks of data (Table 1) to maximize the computational speed of the downstream processes. Then the sequences are sent in parallel to the Picard program for marking the duplicated reads of each chunk. The alignment statistics are measured by Samtools stats and Picard QualityScoreDistribution is derived from the completely merged BAM. The summary statistics include total number of reads, number of mapped and unmapped reads and the fractional proportion of paired reads. Depth of coverage is one of the most important factors for gaining variant discovery performance of genomic sequencing verified by Samtools depth. An overview of the coverage is demonstrated by bar graphs of reading depth (Figure 5). Finally, sequence quality is adjusted for systemic errors by correcting with known highly validated databases using GATK base quality score recalibration.

Figure 3. Overview of sequence and alignment quality

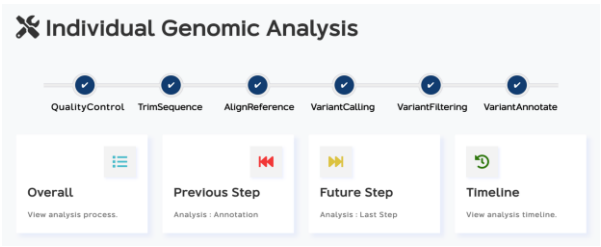


Figure 4. Overview of analysis timeline

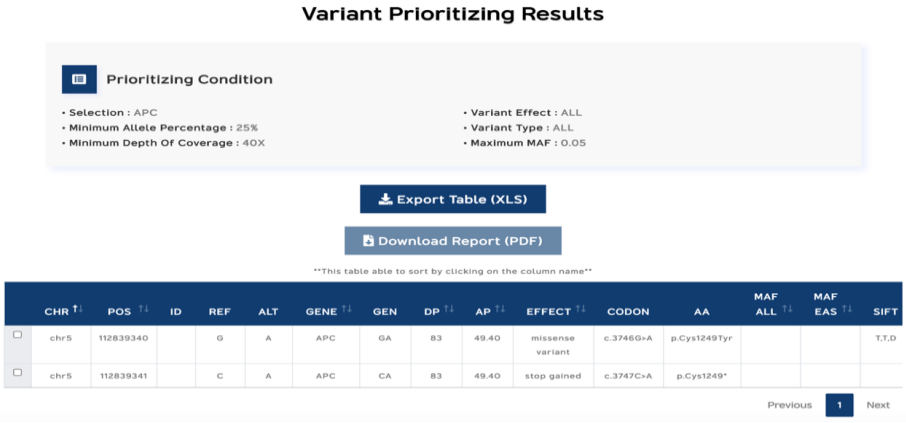


Figure 5. Genomic variant prioritization results

Table 1. Reference genomic contig separation for pre-processing and variant discovery process

Chunk	Chromosome	Start position	End position	Total base pairs
chunk1	1	1	140,000,000	140,000,001
chunk2	1	140,000,001	248,956,422	139,956,422
	2	1	31,000,000	
chunk3	2	31,000,001	171,000,000	140,000,000
chunk4	2	171,000,001	242,193,529	139,193,530
	3	1	68,000,000	
chunk5	3	68,000,001	198,295,559	130,295,559
chunk6	4	1	140,000,000	140,000,001
chunk7	4	140,000,001	190,214,555	140,214,556
	5	1	90,000,000	
chunk8	5	90,000,001	181,538,259	
	6	1	50,000,000	141,538,260
chunk9	6	50,000,001	170,805,979	
	7	1	21,000,000	141,805,980
chunk10	7	21,000,001	159,345,973	
	8	1	3,500,000	141,845,974
chunk11	8	3,500,001	145,138,636	141,638,636
chunk12	9	1	138,394,717	138,394,718
chunk13	10	1	133,797,422	133,797,423
chunk14	11	1	135,086,622	135,086,623
chunk15	12	1	133,275,309	133,275,310
chunk16	13	1	114,364,328	141,364,330
	14	1	27,000,000	
chunk17	14	27,000,001	107,043,718	140,043,719
	15	1	60,000,000	
chunk18	15	60,000,001	101,991,189	132,329,535
	16	1	90,338,345	
chunk19	17	1	83,257,441	140,257,443
	18	1	57,000,000	
chunk20	18	57,000,001	80,373,285	131,990,903
	19	1	58,617,616	
	20	1	50,000,000	
chunk21	20	50,000,001	64,444,167	111,972,620
	21	1	46,709,983	
	22	1	50,818,468	
chunk22	X	1	110,000,000	110,000,001
chunk23	X	110,000,001	156,040,895	103,268,311
	Y	1	57,227,415	
chunk24	M	1	16,569	16,570

3.6 Variant discovery

Pre-processed sequences are re-aligned with a simultaneous reference and the different nucleotides in each position discovered with the double precision pair Hidden Markov Model in GATK. In addition, we separately extract variants from the previous step and merge them together for further analysis. To filter out artifact variants, GATK-VQSR is used to create a cluster model of variants and to identify flexible cut points for each. However, variant filtration is only supported for selected pipelines. Identified variants are functionally annotated with database SNP version 151, non-synonymous SNP database version 4.03c, genome aggregation database version r2.1.1, the last updated ClinVar database (2021-01-04) and a protein predicting score. For each possible variant, summary information is recorded in an individual table of each private database including chromosomes, position, SNP identification number, reference nucleotide, alternative nucleotide, genotype, depth of coverage, allele depth, gene, biotype, variant effect, functional class, amino acid change, minor allele frequency, predicting summary and clinical significance. In addition, the demonstrated variant statistics include mapping quality, quality by depth, fisher strand, mapping quality rank sum test and read position rank sum test.

3.7 Variant prioritization and visualization

To collect the enormous data of genomic variants, MySQL, an open-source relational structure database, is implemented in our analysis system to store variant information and query each item with multiple selecting algorithms. Analyzed data are provided to users, both overview variant call results and specific prioritization. The overview of genomic variants is simply displayed in colorful graphs sorted by essential circumstances. Moreover, we provide basic query options including the gene of interest, depth of coverage, allele depth, type of variant, impact of variant effect and minor allele frequency in the specific population. iCBC also supports advanced query preference related to suspicious diseases or specific oncologic pathways in which we apply array and nested algorithms to match the variant with various preferred genes, mode of inheritance and others in basic options. In the visualization section, selected

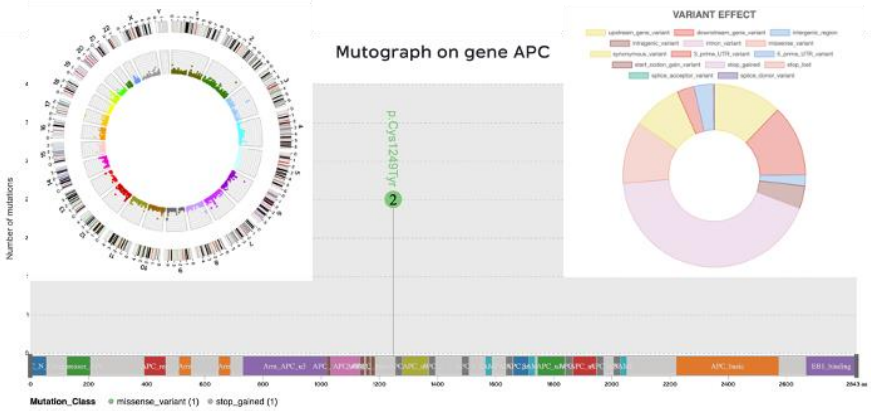


Figure 6. Visualization and summarization of genomic variant data

variants are illustrated for their general information in a single-page table (Figure 5) and an overview of the summarized graphs (Figure 6). Individual reports composed of clinically significant and variant region views appear on a background screen when a user clicks on a unique row of the table.

3.8 Discussion

Bioinformatics is a crucial computational process for dealing with the massive amount of information in genomic variations (Jin *et al.*, 2018). Nowadays, genetic information gives great support to precision medicine which is a modern medical concept (Roden & Tyndale, 2013). The rapid decline of sequencing costs extendedly translates them to clinical application and contributes to the discovery of disease pathogenesis, diagnostic investigations, selecting suitable treatments and prognosis (Qin, 2019). In general, most analysis tools are based on open-source CLI software and need precise ordering of the analysis pipeline which may be difficult and complicated for biomedical researchers (Bodi, 2011). Although there are commercial bioinformatic analysis tools with which users are more familiar, many of these tools require a genuine license and/or annual subscription. Considering the enormous amount of genomic data, the analysis must be executed and processed on a high performance infrastructure together with expandable storage (Papageorgiou *et al.*, 2018). To select only specific genomic information, the researchers frequently prioritize possible variants related to the disease or scope of the study. Variant prioritization, the last important bioinformatic step, is a term of big data analysis related to basic principle of biomedical and computational science (Roca *et al.*, 2018).

The establishment of our web-based analytical tool call iCBC aimed to provide an open access user-friendly graphical interface for computational analysis in biomedical research and related fields. Our tool is composed of two fundamental processes, sequence analysis and data visualization. To assess the genomic sequences, iCBC implements various last updated open-source tools which was suggested in the best practice pipeline. Contributing to the first analytic section, we also included quality check and control, reference alignment and data pre-processing, variant discovery and functional annotation with individual targeted reference. For data visualization, genomic variations are basically displayed in colorful presentations and advanced prioritization with multiple filtering criteria. The criteria are generally applied by selected preferable variants including the gene of interest, depth of coverage, allele depth, type of variants, impact of their effect and minor allele frequency. Moreover, iCBC has comparative genomic analysis features composed of case-control germline analysis and matched tumor-normal somatic analysis. Compared to other commercial tools, we contribute an open-source analysis tool without subscription, with a fully automated pipeline including sequence analysis together with data prioritization, computing infrastructure embedded in high availability and high-performance clusters and secure data storage behind a local firewall and hypertext-encrypted transfer protocol.

4. Conclusions

In conclusion, the iCBC is user-friendly graphical software optimized for human genome sequence analysis and prioritization of discovered variants. iCBC is embedded with various last-update bioinformatic tools and human genome references. Users can easily analyze genomic data by using the automated pipeline and also use customized analysis. The analysis was scheduled for execution in back-end process running on HPC and HA clusters with flexible and scalable data storage secured with a local firewall. Ultimately, the useful information of preferable variants is intensively demonstrated with a paging table and displaying additional information.

Acknowledgements

Financial support from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0185/2561) is acknowledged.

References

Bodi, K. (2011). Tools for next generation sequencing data analysis. *Journal of Biomolecular Techniques: JBT, 22*(Supplement), S18-S18. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3186658/

Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., & Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics, 13*, 8-8. doi:10.1186/1471-2105-13-8

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92. doi:10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., & Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England), 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet., 43*(5), 491-498. doi:10.1038/ng.806

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics, 17* (Supplement 7), 239-239. doi:10.1186/s12859-016-1097-3

Ebiki, M., Okazaki, T., Kai, M., Adachi, K., & Nanba, E. (2019). Comparison of causative variant prioritization tools using next-generation sequencing data in japanese patients with mendelian disorders. *Yonago Acta Medica, 62*(3), 244-252. doi:10.33160/yam.2019.09.001

Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. *Fertility and Sterility, 109*(6), 952-963. doi:10.1016/j.fertnstert.2018.05.006

Gong, Y.-N., Chen, G.-W., Yang, S.-L., Lee, C.-J., Shih, S.-R., & Tsao, K.-C. (2016). A Next-Generation sequencing data analysis pipeline for detecting unknown pathogens from mixed clinical samples and revealing their genetic diversity. *PLoS One, 11*(3), e0151495-e0151495. doi:10.1371/journal.pone.0151495

He, K. Y., Ge, D., & He, M. M. (2017). Big data analytics for genomic medicine. *International Journal of Molecular Sciences, 18*(2), 412. doi:10.3390/ijms18020412

Jin, Y., Zhang, L., Ning, B., Hong, H., Xiao, W., Tong, W., Guo, Y. (2018). Application of genome analysis strategies in the clinical testing for pediatric diseases. *Pediatric Investigation, 2*(2), 72-81. doi:10.1002/ped4.12044

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Genome Project Data Processing, S. (2009). The SEQUENCE ALIGNMENT/MAP FORMAT AND SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., & Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England), 25*(15), 1966-1967. doi:10.1093/bioinformatics/btp336

Liao, P., Satten, G. A., & Hu, Y. J. (2017). PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet Epidemiol, 41*(5), 375-387. doi:10.1002/gepi.22048

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., . . . The, W. G. S. C. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine, 6*(3), 26. doi:10.1186/gm543

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., & DePristo, M. A. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., & Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology, 17*(1), 122-122. doi:10.1186/s13059-016-0974-4

Nambot, S., Thevenon, J., Kuentz, P., Duffourd, Y., Tisserant, E., Bruel, A.-L., & Orphanomix Physicians, G. (2018). Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genetics in Medicine, 20*(6), 645-654. doi:10.1038/gim.2017.162

Papageorgiou, L., Eleni, P., Raftopoulou, S., Mantaiou, M., Megalooikonomou, V., & Vlachakis, D. (2018). Genomic big data hitting the storage bottleneck. *EMBnet.journal, 24*, e910. Retrieved from https://pubmed.ncbi.nlm.nih.gov/29782620

Peng, X., Wang, J., Zhang, Z., Xiao, Q., Li, M., & Pan, Y. (2015). Re-alignment of the unmapped reads with base quality score. *BMC Bioinformatics, 16 (Supplement 5)*, S8. doi:10.1186/1471-2105-16-S5-S8

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics, 8*(1), 14. doi:10.1186/1479-7364-8-14

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology and Medicine, 16*(1), 4-10. doi:10.20892/j.issn.2095-3941.2018.0055

Roca, I., Fernández-Marmiesse, A., Gouveia, S., Segovia, M., & Couce, M. L. (2018). Prioritization of variants detected by next generation sequencing according to the mutation tolerance and mutational architecture of the corresponding genes. *International Journal of Molecular Sciences, 19*(6), 1584. doi:10.3390/ijms19061584

Roden, D. M., & Tyndale, R. F. (2013). Genomic medicine, precision medicine, personalized medicine: what's in a name? *Clinical Pharmacology and Therapeutics,*

94(2), 169-172. doi:10.1038/clpt.2013.101

Ruffalo, M., Koyuturk, M., Ray, S., & LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics (Oxford, England), 28*(18), i349-i355. doi:10.1093/bioinformatics/bts408

The Genomes Project, C., Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*, 68. doi:10.1038/nature15393 Retrieved from https://www.nature.com/articles/nature15393#supplementary-information

Tian, S., Yan, H., Kalmbach, M., & Slager, S. L. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bio informatics, 17*(1), 403-403. doi:10.1186/s12859-016-1279-z

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: current and future perspectives. *G3 (Bethesda, Md.), 5*(8), 1543-1550. doi:10.1534/g3.115.018564

Wildeman, M., van Ophuizen, E., den Dunnen, J. T., & Taschner, P. E. M. (2008). Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human Mutation, 29*(1), 6-13. doi:10.1002/humu.20654

Yu, Y. W., Yorukoglu, D., Peng, J., & Berger, B. (2015). Quality score compression improves genotyping accuracy. *Nature Biotechnology, 33*(3), 240-243. doi:10.1038/nbt.3170