Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 5 No. 1 (2023)
https://doi.org/10.35877/454RI.asci1608

The Bioinformatics Application in Detecting Germline and Somatic Variants towards Breast Cancer using Next Generation Sequencing

Rizka Retnomawarti[a], Sonar Soni Panigoro[b], & Rafika Indah Paramita[c,d,e]

[a]*Master's Programme in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Depok, Indonesia*
[b]*Surgical Oncology Division, Department of Surgery, Faculty of Medicine, Universitas Indonesia, Depok, Indonesia*
[c]*Doctoral Program in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Depok, Indonesia*
[d]*Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Depok, Indonesia*
[e]*Bioinformatics Core Facilities – IMERI, Faculty of Medicine, Universitas Indonesia, Depok, Indonesia*

Abstract

Breast cancer is the type of cancer with the most and the highest cases causing mortality in Indonesia, so an effective treatment is required to reduce the incidence and mortality rate due to cancer breasts. Most breast cancer patients are diagnosed at an advanced stage so the treatment used are limited and the risk of death becomes higher. Along with the development of human genome sequencing technology, the genetic examination of breast cancer is considered as an examination that can be used for early prevention and treatment management personally. Based on the target variants detected, the genetic examination of breast cancer can be divided into two, namely the examination of germline variants and somatic variants. Germline variant examination is intended to predict the risk of breast cancer which can be used as an early preventive measure, while somatic variant examination is intended for cancer diagnosis and management therapy. NGS technology is able to detect both types of variants in a number of genes associated with breast cancer in several samples effectively and quickly. However, the data generated from NGS technology is very large and complex, so the role of bioinformatics is required in analyzing and interpreting data. By utilizing bioinformatics pipelines and tools, analysis of germline variants and somatic variants in breast cancer can be carried out accurately so that the results of genetic examinations can be used as a step to treat breast cancer.

Keywords: Germline, NGS, pipeline, somatic.

Received: 3 March 2023 Revised: 29 April 2023 Accepted: 12 May 2023

1. Introduction

Breast cancer is a form of uncontrolled growth accumulation of breast cell, causing the formation of masses or tissues characterized by lumps on the breast. The uncontrolled growth of cells can spread to other body tissues and form new tumors or called metastases (Bai et al., 2021; Sun et al., 2017). The majority of breast cancers are experienced by women and are known to be the type of cancer that has the highest prevalence rate in the world rather than other types of cancer. Early detection and treatment is one of the strategies for breast cancer treatment which can increase survival by 90% (Huang & Davidson, 2006; Lynch et al., 2015; Wang et al., 2021).

Genetic examination of genes known to be associated with breast cancer is one of the early detection methods for cancer treatment. In conducting genetic diagnostic testing, understanding the difference between germline variants and somatic variants is essential to be recognized. The germline variant is inherited from the mother or father during the conception or fertilization process (Ginsburg et al., 2020; Tschiderer et al., 2022). The term of "germ" refers to variants present in egg and sperm cells or referred to as "germ cells." This variant exists in all cells of the body and remains as long as the individual lives. Germline variants can be "dominant" or "recessive". A dominant variant means that one parent has a normal copy of the gene and a mutated copy where there is a 50% chance that the variant is

* Corresponding author.
E-mail address: sonar.soni@ui.ac.id

inherited. Meanwhile, the recessive variants mean that as many as two copies of the gene mutate are required to cause the disease, where there is a 25% chance of variants inherited. Somatic variants are the variants obtained after fertilization or during the fetus growth in the womb. This variant appears only on tumor cells and it is not in all tissues of the body and cannot be inherited. Somatic variants are often referred to as the driver variants because they can initiate cancer growths. Several drugs have been developed targeting such variants to control cancer growth. Therefore, the identification of somatic variants is often associated with treatment management and therapy of a disease (What_is_the_Difference_Between_Potentia, n.d.) .

Next Generation Sequencing (NGS) technology is a technology that can be used in genetic examinations related to breast cancer. Accurate invocation of variants in NGS data is a very important stage in the entire process of data analysis and interpretation. Along with the development of NGS technology in the last 10 years. The development of analytical tools and pipelines is used to detect variants in clinical sample sequence results that is also developing very quickly. Until now, there are several bioinformatics pipelines that can be used to analyze and interpret NGS result data (Lynch et al., 2015). This paper aims to discuss about NGS technology in the genetic examination of breast cancer and the role of bioinformatics in analyzing germline and somatic variants in cancer, especially for breast cancer.

2. Literature Review

2.1. Germline Variants

The *germline* variant is defined as a variant caused by a gene change in reproductive cells (eggs or sperm cells) that then impacts the DNA present in every cell throughout the body . The germline variant can be inherited from parents to their children. So, this variant is also called as the hereditary variant. Germline variants have been widely studied by various researchers regarding their correlation with cancer risk where the results of this study can improve clinicians' understanding of the process cellular cancer, as well as guidance related to cancer screening and treatment. Growing germline variant studies found a correlation between the variant and tumor development and therapeutic response. Germline variants are also known to impact molecular pathways through changes in amino acids, splicing patterns or expression in genes, increasing the occurrence of somatic mutations, and broad genome mutations. These molecular changes cause the tumor to develop and metastasize, as well as changes in the immune environment and therapeutic response (Chatrath *et al.,* 2021).

In breast hereditary cancer, there a several germline variants in certain genes that have been identified as having strong associations with cancer, so they are considered as the main genes needed in a hereditary breast cancer genetic examination. These genes include the BRCA1 and BRCA2 genes *(BRCA1/2), TP53*, *PTEN*, *CDH1*, *STK11*, and *PALB2*, which are included in the high-penetration gene groups. In addition, it can increase the risk of breast cancer by 4four times, as well as the *ATM*, *NF1,* and *CHEK2* which is a moderate penetration gene group and can increase the risk of breast cancer by 2-3 times (Desai *et al.,* 2021).

Detection of germline variants associated with breast cancer is beneficial for the cancer prevention, early detection, and treatment for both patients and their families. The management of patients who have carried out genetic examinations can be adjusted to the results of germline variants obtained, which includes decisions in performing surgical procedures ,such as bilateral mastectomy for patients carrying a high penetration gene variants, avoidance of radiation action for patients with Li-Fraumeni syndrome who have variants of TP53 gene, implementation of cancer prevention measures using the Magnetic Resonance Imaging (MRI) method, and the use of poly inhibitor drug compounds (adenosine diphosphate ribosa) polymerase in BRCA1/2 mutation-carrying patients diagnosed with metastatic breast cancer. Besides the benefits for patients, germline variant testing is also considered essential for the patient's family members. Germline variants detected in an individual have a 50% chance of decreasing to other family members. This variant information can be a guide to implement several strategies for preventing and screening hereditary cancer diseases early to increase the potency of higher cancer recovery (Desai et al., 2021).

2.2. Somatic Variants

Somatic variants, known as genomic variants, is the changes DNA that occur after fertilization in all cells except germ cells (sperm and egg cells). These DNA changes are generally caused by exposure to environmental factors, such as ultraviolet (UV) light, cigarette smoke, radiation, viruses, chemical compounds, age, and alcohol. Unlike the germline

variant, the somatic variant cannot be inherited from the parent to the child, besides the somatic mutation rate is much higher than germline, which is $6.4\times10-10$ -7.8×10^{-10} mutations per base pair per cell division across cell types that can be even higher depending on the level of affected cell growth (National Cancer institute, n.d.; National Health Service, 2016; Van Horebeek *et al.,* 2019).

Although most somatic variants accumulated in cells are harmless or do not have a significant effect on individuals, variants that affect a gene with a specific regulatory function can result in phenotype changes that lead to diseases or abnormalities of the body. The impact of somatic variants on phenotype changes depends on the large differences in variance factors influenced by the number of variants (abundance/ variant allele fraction (VAF)), the type of cells affected by the variant, and the type of variant. The higher the VAF value, the higher the effect of varietal changes on the phenotype so that it can result in the development of a disease, such as cancer. Studies show that there are around 70 – 80% of cases of cancer caused by somatic variants, so further studies are needed to understand somatic variants and their use of the development of diagnostic examination and treatment of cancer caused by these variants (Campbell & Martincorena, 2015).

Along with the development of genetic technology, researches cancer related to the characteristics of somatic variants has developed much Cancer caused by the accumulation of somatic variants in cells produces heterogeneous tumors consisting of different sub-populations/cell clones with different types of mutations. Unlike the germline variant, the somatic variant type is much more complex. For example, copy number aberrations (CNA) that is the most commonly found somatic mutation of cancer where this type of mutation can increase or decrease one or two alleles (or several kilobases) in a region of the genome, the arm of the chromosome, even in the entire chromosome. CNA is known to have an important role in the development of a cancer so CNA analysis is required in the nature of cancer diagnosis, prognosis, and treatment. Besides CNA mutations, the types of mutations found in somatic cancers are big insertion/delesion mutations, SNV, splicing, non-sense, and frameshifts (Mathioudaki *et al.,* 2020; Zaccaria & Raphael, 2020).

In breast cancer, studies show several somatic variant genes that act as gene drivers, including the *TP53, PIK3CA, CDH1, AKT1, GATA3, MAP2K7, MYC, KMT2C, ERBB2, PTEN,* and *MAP2*K4 genes. The *TP53* and *PIK3CA* genes are the most widely studied genes related to breast cancer. The *TP53* gene is a gene that encodes the transcription factor of p53. The *TP53* gene plays a role in the initiation of gene transcription involved in cell cycle processes, cellular senescence, metabolism, apoptosis, DNA repair, and cell stress response. Mutations in the *TP53* gene in breast cancer were found in more than 30% of cases, especially in HER2-positive type breast cancer and Triple Negative Breast Cancer (TNBC) (Bai *et al.,* 2021). Moreover, the *PIK3CA* gene is a gene that encodes the p110 alpha protein, a subunit of the phosphatidylinositol 3-kinase (PI3K) protein that plays an important role in various biological functions, such as cell survival, cell differentiation, and cell proliferation through the PI3K/AKT/mTOR signaling pathway. *PIK3CA* gene mutations are known to play a role in the process of initiation and development of breast cancer. Besides, these are the second most mutations found in breast cancer cases after mutations in the *TP53* gene, especially in ER-positive type breast cancer. Based on the role of the both genes, the *TP53* and the *PIK3CA* genes have the potential to be biomarker genes to predict the development of breast cancer and the selection of cancer treatments which is appropriate for the patient (Zardavas *et al.,* 2014).

2.3. The Workflow of NGS Technology

NGS technology uses a similar principle with capillary electrophoresis sequencing technology, where DNA polymerase enzymes catalyze the process of adding fluorescent of *deoxyribonucleotide triphosphates* (dNTPs) compounds labeled fluorescent on DNA template throughout the DNA synthesis cycle. At each addition of nucleotide bases, the fluorescent signal will be captured by the tool and read as a specific base. This process takes place continuously in each synthesis cycle and is parallel to all millions of DNA fragments. The NGS stages are broadly divided into four major stages, as follows (Ilumina, 2021):

a) Formation of library sequencing (library preparation)

   In the formation of libraries, DNA or cDNA samples are randomly fragmented into small pieces, then a ligation process is carried out with 5' and 3' adapters where these adapters contain of sequencing primers and barcode indexes, which are unique sequence tags that are specific to a particular sample. The barcode index contained in

each sample allows the entire sample to be mixed and sequenced at the same time. After the adapter pasting process is carried out, the amplification, purification, and quality control of library results are carried out.

b) Cluster generation

In most sequencing platforms, the clonal amplification process is an important stage for the sequencing process which functions in forming many copies of libraries both with the PCR emulsion method and the *bridge amplification* method which then generates the similar library cluster as the template in the sequencing process.

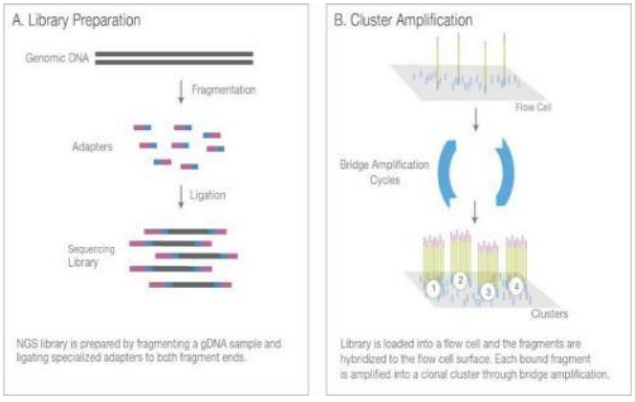
c) Sequencing process

Sequencing technology generally uses the *sequencing by synthesis* (SBS) method, which is a nucleotide detection method by utilizing the DNA polymerase enzyme. Some of the SBS methods include the sequencing method with reversible terminators used by the Illumina platform, pyrosequencing used by the Roche platform and the sequencing of semi-conductor ions by the ion platform Torrent. In the reversible terminator method, each strand of DNA template is attached to a chip that is read by synthesis method using polymerase enzymes and the help of *dideoxytnucleotide triphosphate* (ddNTP) compounds that have characterized by fluorescent dyes. ddNTP is a compound that aims to stop the process of DNA synthesis in a certain base and then emission wavelengths of different colors to be trapped by a sequencing machine camera. Furthermore, the reverse termination enzyme plays an important role in converting ddNTP into dNTP after being bound to the template strand so that the synthesis reaction takes place continuously (Ilumina, 2021; *NGS Workflow and Fundamentals of Sample Preparation - Enzo Life Sciences*, n.d.) .

Different with the reversible terminator method, the *pyrosequencing* method uses non-electrophoretic bioluminescence compounds that make measurements based on the release of inorganic pyrophosphates converted into light through enzymatic cascade reaction. Meanwhile, the semi-conductor ion sequencing method is a DNA sequencing method which is based on the detection of hydrogen ions released into the DNA mold. In ion-semiconductor sequencing *chips*, hydrogen ions are detected after a clear correlation is formed between chemical and digital processes where nucleotide bases are read as unprocessed scanning, camera, and light (Behjati & Tarpey, 2013; *Ion Semiconductor Sequencing - an overview | ScienceDirect Topics*, n.d.) .

d) Data analysis

Sequencing data analysis is divided into three important stages, primary, secondary, and tertiary analysis. Primary analysis involves the drawing process, the calculation of signal intensity, and the process of converting signal data into sequencing data. Sequencing data in the form of FASTQ files is then processed more directly based on the quality of the data produced. Secondary analysis involves the process of aligning sequence readings with genomic references or *de novo assembly* processes and *variant calling* processes based on the bioinformatics pipeline used. Tertiary analysis involves the process of finding a correlation between the variance data of sequencing results and the patient's phenotype. Tertiary analysis starts from annotations and interpretation of variants, which then the results of the interpretation can be further used based on the application of the NGS technology studied (McFadyen R, 2020).



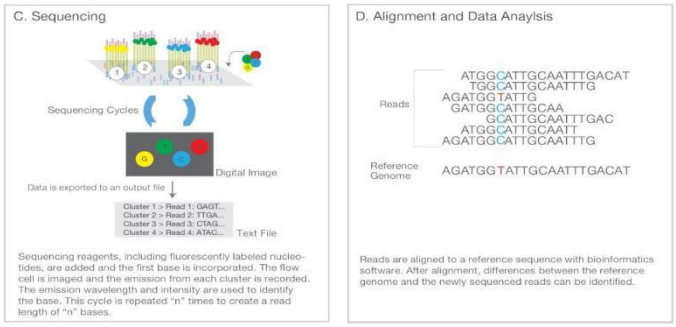


Figure 1. Next *Generation Sequencing* (NGS) stages grouped into four major stages, namely a) Formation of sequencing libraries, b) Formation of the clist, c) Sequencing process, and d) Data analysis (Ilumina, 2021)

2.4. The Role of Bioinformatics in Breast Cancer Detection

The sequencing process using NGS technology produces very large data outputs (gigabytes to terabytes) as well as millions of sequences reads so that computational methods are required in processing and analyzing the data. Bioinformatics is a field of science that uses computational, mathematical, and statistical approaches to collect, organize, and analyze, large and complex genomic data. The set and stages of bioinformatics algorithms used to process raw data sequencing to generate annotations and variant interpretations that are referred to as pipelines. A bioinformatics pipeline is designed, developed, validated by various communities or organizations to be used by open researchers or academics (open source) or requires a special license for its use. In the selection of a bioinformatics pipeline, it is necessary to understand the function of each algorithm used to be adjusted to the characteristics of the NGS experiment so to get the expected results based on specific purpose of study or examination. Some experimental characteristics to be concerned in pipeline selection are (McFadyen R, 2020; Ozcelik *et al.*, 2012; Roy *et al.*, 2018; Walsh *et al.*, 2010; Welch *et al.*, 2011) genetic study/examination (identification of somatic variants/germline).

In the genetic examination of breast cancer, it commonly uses the targeted sequencing type because there have been quite a lot of studies that prove the role of several important genes in the growth of cancer, so that in the diagnosis of breast cancer can focus on that gene collection. For example, in breast hereditary cancer examination, commonly identified genes are *BRCA1*, *BRCA2, PALB2, CHECK2, ATM*, *MLH1*, *MSH2*, *CDH1*, *PTEN*, *SK11*, and *TP53*. Meanwhile, in the examination somatic breast cancer, the genes identified are *TP53*, *PIK3CA*, *CDH1*, *AKT1*, *GATA3*, MAP2K7, *MYC*, *KMT2C*, *ERBB2*, *PTEN*, *BRCA1*, *BRCA2*, and *MAP2K4*. The type of sample used for diagnosis is DNA obtained from blood/saliva samples (for germline variant detection) or tumor samples (for somatic variant detection). The type and target of variants to be detected in a breast cancer diagnosis examination determine what bioinformatics analysis method will be used due to the complexity of different variants between germline and somatic variants so that the selection of the right pipeline is required so that the interpretation results are in accordance with the purpose of diagnosis (Lynch *et al.*, 2015) .

2.4.1. Stages of Bioinformatics Analysis (Pipeline)

The stages of NGS sequencing data analysis mostly are divided into three main stages, namely primary analysis, secondary analysis, and tertiary analysis (McFadyen R, 2020). Primary analysis is an analysis process that starts from the conversion of signal data generated by the instrument into sequencing data containing nucleotide base calls. This primary analysis is generally performed on instrument software that converts binary base call (BCL) file data into FASTQ files. After obtaining the FASTQ file, the pre-processing reads process is carried out as a quality control process for the sequencing results obtained, several stages of pre-processing include the filtering process, demultiplexing, and trimming (McFadyen R, 2020). Filtering is a process of filtering the readings that do not fulfill the quality of alkaline length and base call based on Phred values. The readings that do not meet quality requirements need to be eliminated to minimize false positives and mapping results with bad genome references. Demultiplexing is the process of separating sequence readings into separate files based on the sample barcode index. Trimming is the process of base removal that has a low quality (<Q30) or removal of adapter sequences that are read on the results of

the sequence. The trimming process can be done using bioinformatics tools, such as Trimmomatic (McFadyen R, 2020).

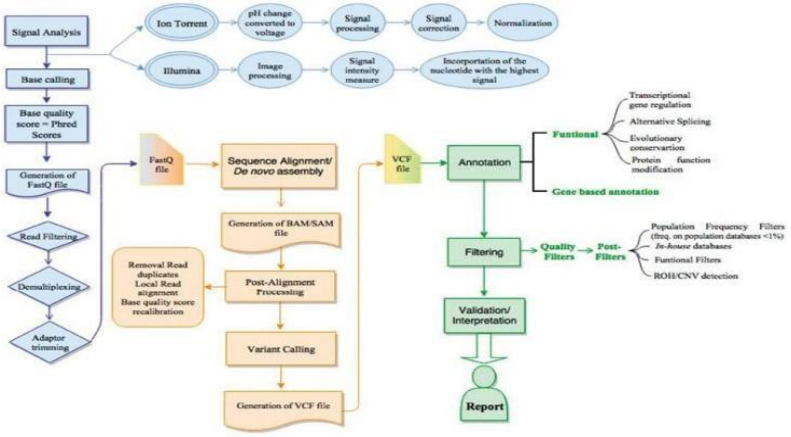


Figure 2. Illustration of bioinformatics pipelines in NGS sequencing data analysis *(30)*

After obtaining the results of sequence readings that have high quality from the primary analysis, secondary analysis is carried out. The process here is the alignment of reads with genome references (alignment) and variant calling (McFadyen R, 2020). The alignment process aims to find the location of the read genome based on the references used and determine how many reads are successfully lined up in that position. A commonly used reference genome is GRCh37/hg19 or GRCh38/hg38 which is the result of a human genome assembly obtained based on the publication of the *Human Genome Project.* The latest reference genome published in December 2013 is GRCh38/hg38 which is a corrected and updated version of the previous version, hg19. The longer the reads produced, the alignment process will be easier, and if the type of reads used is *paired end*, then the position alignment can be validated using a second base reading, so it reduces *false positives* or alignment error (McFadyen R, 2020).

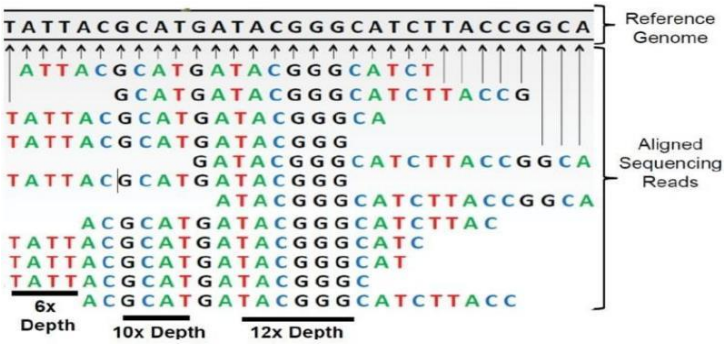


Figure 5. Visualization of the reads alignment process with genome references (McFadyen R, 2020)

Before entering the *variant calling* process, the next process is post-alignment processing. This process aims to control quality so that only the reads with the best quality are generated on a BAM file or a file containing alignment results in binary form. This process is divided into three stages, namely duplicate removal, removal of reads process derived from the same DNA molecule or duplication of PCR which can result in false positives on interpretation, local read realignment, which is a process to improve the accuracy of detection of insertion/deletion variants and reduce *mismatchingi* from aligned reads, base quality score recalibration, which is an alignment metric used to estimate the true value of base call quality (McFadyen R, 2020).

Variant calling is a process that aims to detect variants based on nucleotide differences that exist in sequence results with reference genomes. The input data entered is a BAM file that contains the entire reads information and its quality values. There are more than 15 algorithms that can be used for variant switching processes, which are grouped into

three categories, namely the allele calculation algorithm, probability method (Bayesian model), as well as the Heuristic approach based on the quality, frequency, and significance of variants. Some bioinformatics tools that are commonly used for the variant calling process are Strelka, FreeBayes, SpeedSeq, Samtools, Varscan2, GATK, DRAGEN, and DeepVariant with output in the form of VCF files. In these tools, researchers can determine the characteristics of the variants they want to identify. It can determine whether only the SNP variant and the small InDel variant (<50 bp) are to be identified or if identification is required in the large deletion variant large/structural variants, such as CNV. This is important to determine because there are differences in algorithms and parameters in tools for detecting variant types. The determination of the variant to be identified can also be based on the limitation of the panel/library reagent used or the purpose of the diagnostic study/examination performed (McFadyen R, 2020).

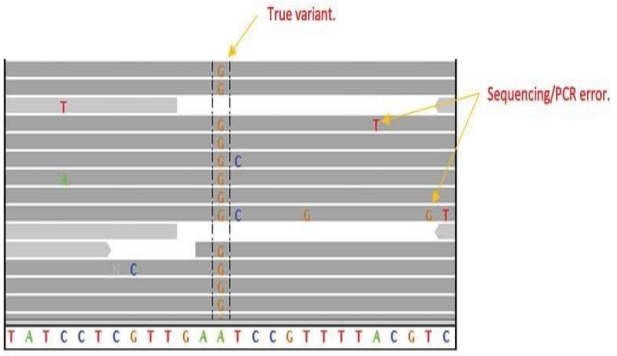


Figure 6. Variant calling process based on BAM files visualization between reads sequencing results and reference genomes [(30)]

Tertiary analysis is the final process of NGS data analysis which involves the process of annotating variants and interpreting variants. Variant annotation is a process of predicting the biological and functional effects of a genetic variant on human clinical status based on a database of variants, such as dbSNP, dbVar, OMIM, ClinVar, COSMIC, PharmGKB, and 1000Genome. Variant interpretation is a process of classifying variants based on the guidance of *The American College* of *Medical Genetics and Genomics* (ACMG), *the Association for Molecular Pathology* (AMP), and *The College of American Pathologists* (CAP). This guide was established by an organization of geneticists and clinicians around the world and became a basic guideline for the interpretation of a particular variant based on the principle of *Mendelian disorder*. There are five variants classifications, namely 'pathogenic', 'likely pathogenic', '*uncertain significance*', 'likely benign', and '*benign'*. The results of this interpretation are then provided to the doctor to be conveyed to the patient and adjusted to the clinical condition of the patient concerned (McFadyen R, 2020).

2.4.2. Analysis and Interpretation of Germline and Somatic Variants

The main difference in germline and somatic variants is in how the both types of variants are distinguished by reference. In the discussion related to the differences in germline and somatic variant analysis, this paper discussed more deeply related to the analysis pipeline provided by GATK because this pipeline is the most widely used analytical tool and provides a complete statistical method. Besides, this has high accuracy to identify variants based on differences in readings of sequencing results with reference genomes (Zhao *et al.*, 2020).

The germline variant is an easy variant to analyze rather than somatic variant because this variant has very clear differences in reference. The summoning of germline variants generally assumes the presence of a fixed ploidy and so the summoning process only includes the genotype site. HaplotypeCaller is a tool used in GATK pipelines for calling germline variants. It uses an invocation variant setting based on ploidy differences with diploids on one or more samples without necessarily relying on the allele balance in the genotype. Unlike the germline variant, the somatic variant has a contrast between the two samples on the reference. Detection of somatic variants in GATK pipelines using a tool called Mutect2. It will distinguish the presence or absence of evidence on a variant between two samples, i.e. a tumor sample and a normal sample of the same individual, and then, the variance will be called a somatic variant if it has differences from the control and reference samples. Both tools of HaplotypeCaller and Mutect2 can call SNV and indel variants simultaneously. However, the Mutect2 GATK tools have the ability to call the copy

number variant and structural variant using frequency information variant alleles in tumors for validation of variant invocation results. The process of invoking variants has the same principle as invoking SNV and indel variants, namely by comparing variants with references to genome regions with high copy number variations and repeated structural variation (Shlee, 2015).

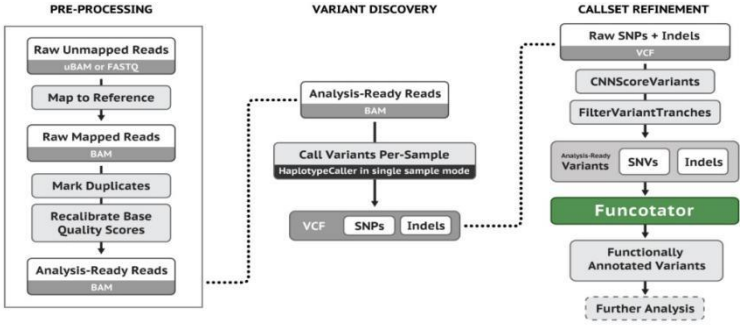


Figure 7 Illustration of germline variant analysis pipeline on GATK's HaplotypeCaller tools (Institute, n.d.)

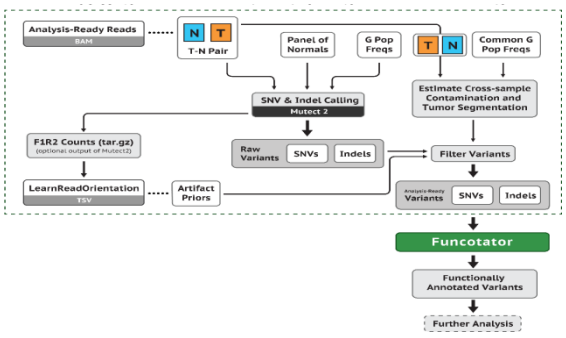


Figure 8 Illustration of somatic variance analysis pipeline on GATK's Mutect2 tools (Broad Institute, 2020)

Besides differences in analysis, germline and somatic variants are also distinguished by categories and evidence supporting interpretation. The classification of the interpretation of germline variants and somatic variants based on their pathogenicity is categorized into five, namely pathogenic, likely pathogenic, Variant of Uncertain Significance (VUS), likely benign, and benign based on ACMG guidelines. A variant is stated *pathogenic* if it is proven to interfere with gene function and can cause a disease. Meanwhile, a variant is stated a *benign* if the variant has high frequency and has no (neutral) influence on the increased risk of a disease. A variant is stated a *Variant of Uncertain Significance* (VUS) if the variant is not yet known for its biological significance because there have not been many studies studying the variant so its impact on the risk of disease is also uncertain. A variant is stated as *likely pathogenic* or *likely benign* if there is evidence with more than 90% confidence that the variant is the cause of a disease or does not cause a disease (Li *et al.*, 2017).

3. Conclusion

Breast cancer is the most studied type of cancer related to genetic factors and the mechanism of cancer growth. There are several genes that have been identified as having a high association with breast cancer where variations in those genes can cause a person has a higher risk of disease than the population. Then, the gene collection is used as a genetic test target to identify variants in patient samples. Based on the cell type, genetic variants can be classified into two types, namely germline and somatic variants. Next Generation Sequencing (NGS) technology is a very useful method in genetic testing to detect both variants in a number of gene targets for several samples at once. Sequencing data

generated by NGS technology can be analyzed using bioinformatics pipelines and tools. Bioinformatics has an important role in accurately identifying and distinguishing germline and somatic genetic variants. By having accurate variant interpretation results, breast cancer genetic examination can be used as a good step for cancer management in terms of prevention, diagnosis, and treatment of diseases.

References

Bai, H., Yu, J., Jia, S., Liu, X., Liang, X., & Li, H. (2021). Prognostic value of the tp53 mutation location in metastatic breast cancer as detected by next-generation sequencing. *Cancer Management and Research*, *13*, 3303–3316. https://doi.org/10.2147/CMAR.S298729

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Broad Institute. (2020). Somatic short variant discovery (SNVs + Indels). *Broad Institute*.

Campbell, P. J., & Martincorena, I. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1488.

Chatrath, A., Ratan, A., & Dutta, A. (2021). Germline Variants That Affect Tumor Progression. *Trends in Genetics*, *37*(5), 433–443. https://doi.org/10.1016/j.tig.2020.10.005

Desai, N. V., Yadav, S., Batalini, F., Couch, F. J., & Tung, N. M. (2021). Germline genetic testing in breast cancer: Rationale for the testing of all women diagnosed by the age of 60 years and for risk-based testing of those older than 60 years. *Cancer*, *127*(6), 828–833. https://doi.org/10.1002/cncr.33305

Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Yataco, J. A. D., Gyawali, B., McCormack, V., de Anderson, M. M. L., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M., Unger-Saldaña, K., … Anderson, B. O. (2020). Breast Cancer Early Detection: A Phased Approach to Implementation. *Cancer*, *126*(S10). https://doi.org/10.1002/cncr.32887

Huang, Y., & Davidson, N. E. (2006). Breast cancer. *Principles of Molecular Medicine*, 728–735. https://doi.org/10.1007/978-1-59259-963-9_74

Ilumina. (2021). Next-Generation Sequencing (NGS) | Explore the technology. *Innovative Technologies*.

Institute, B. (n.d.). Germline short variant discovery (SNPs + Indels). *Germline Short Variant Discovery (SNPs + Indels)*.

Ion Semiconductor Sequencing - an overview | ScienceDirect Topics. (n.d.).

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *Journal of Molecular Diagnostics*, *19*(1), 4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002

Lynch, J. A., Venne, V., & Berse, B. (2015). Genetic tests to identify risk for breast cancer. *Seminars in Oncology Nursing*, *31*(2), 100–107. https://doi.org/10.1016/j.soncn.2015.02.007

Mathioudaki, A., Ljungström, V., Melin, M., Arendt, M. L., Nordin, J., Karlsson, Å., Murén, E., Saksena, P., Meadows, J. R. S., Marinescu, V. D., Sjöblom, T., & Lindblad-Toh, K. (2020). Targeted sequencing reveals the somatic mutation landscape in a Swedish breast cancer cohort. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-74580-1

McFadyen R. (2020). *NGS Data analysis workflow*.

National Cancer institute. (n.d.). *Definition of locus - NCI Dictionary of Genetics Terms - NCI*.

National Health Service. (2016). *Genomics Education Programme*. *21*(6), 747–754.

NGS Workflow and Fundamentals of Sample Preparation - Enzo Life Sciences. (n.d.).

Ozcelik, H., Shi, X., Chang, M. C., Tram, E., Vlasschaert, M., Di Nicola, N., Kiselova, A., Yee, D., Goldman, A., Dowar, M., Sukhu, B., Kandel, R., & Siminovitch, K. (2012). Long-range PCR and next-generation sequencing of BRCA1 and BRCA2 in breast cancer. *Journal of Molecular Diagnostics*, *14*(5), 467–475. https://doi.org/10.1016/j.jmoldx.2012.03.006

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V., Wang, C., & Carter, A. B. (2018). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *Journal of Molecular Diagnostics*, *20*(1), 4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003

Shlee. (2015). *Somatic calling is NOT simply a difference between two callsets*.

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. In *International Journal of Biological Sciences* (Vol. 13, Issue 11). https://doi.org/10.7150/ijbs.21635

Tschiderer, L., Seekircher, L., Kunutsor, S. K., Peters, S. A. E., O'keeffe, L. M., & Willeit, P. (2022). Breastfeeding Is Associated With a Reduced Maternal Cardiovascular Risk: Systematic Review and Meta-Analysis Involving Data From 8 Studies and 1 192 700 Parous Women. In *Journal of the American Heart Association* (Vol. 11, Issue 2). https://doi.org/10.1161/JAHA.121.022746

Van Horebeek, L., Dubois, B., & Goris, A. (2019). Somatic Variants: New Kids on the Block in Human Immunogenetics. *Trends in Genetics*, *35*(12), 935–947. https://doi.org/10.1016/j.tig.2019.09.005

Walsh, T., Lee, M. K., Casadei, S., Thornton, A. M., Stray, S. M., Pennil, C., Nord, A. S., Mandell, J. B., Swisher, E. M., & King, M. C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(28), 12629–12633. https://doi.org/10.1073/pnas.1007983107

Wang, L., Zhang, S., & Wang, X. (2021). The Metabolic Mechanisms of Breast Cancer Metastasis. In *Frontiers in Oncology* (Vol. 10). https://doi.org/10.3389/fonc.2020.602416

Welch, J. S., Larson, D. E., Wallis, J., Chen, K., Payton, J. E., Fulton, R. S., Veizer, J., Schmidt, H., Vickery, T. L., Watson, M. A., Link, D. C., Graubert, T. A., Mardis, E. R., Ley, T. J., & Wilson, R. K. (2011). Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene. *Jama*, *305*(15), 1577–1584.

What_is_the_Difference_Between_Potentia. (n.d.). Verywell Health.

Yaoting Gui1, 12, Guangwu Guo2, 12, Yi Huang1, 12, Xueda Hu2, 12, Aifa Tang1, 3, 12, Shengjie Gao2, Renhua Wu2, Chao Chen2, Xianxin Li1, Liang Zhou1, Minghui He2, Zesong Li1, 3, Xiaojuan Sun3, Wenlong Jia2, Jinnong Chen2, Shangming Yang2, Fangjian Zhou4, C. L. (2017). *Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder*. *21*(2), 129–139.

Zaccaria, S., & Raphael, B. J. (2020). Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17967-y

Zardavas, D., Phillips, W. A., & Loi, S. (2014). PIK3CA mutations in breast cancer: Reconciling findings from preclinical and clinical data. *Breast Cancer Research*, *16*(1). https://doi.org/10.1186/bcr3605

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., & Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-77218-4