Data in Brief 59 (2025) 111343

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

SNP-array profiling data from breast cancer patients and healthy women's blood DNA samples

Rafika Indah Paramita[a,b,c], Sonar Soni Panigoro[d,e,*], Fadilah Fadilah[b,c,d,*], Septelia Inawati Wanandi[d,f,g], Noorwati Sutandyo[h,i]

[a] Doctoral Program in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia
[b] Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia
[c] Bioinformatics Core Facilities-IMERI, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 6, Jakarta, 10430, Indonesia
[d] Master's Programme in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia
[e] Surgical Oncology Division, Department of Surgery, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 6, Jakarta, 10430, Indonesia
[f] Department of Biochemistry and Molecular Biology, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia
[g] Molecular Biology and Proteomics Core Facilities-IMERI, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 6, Jakarta, 10430, Indonesia
[h] Department of Hematology and Medical Oncology, Dharmais National Cancer Center Hospital, Jalan Letjen S. Parman, Jakarta, 11420, Indonesia
[i] Department of Internal Medicine, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia

ARTICLE INFO

Article history:
Received 13 December 2024
Revised 10 January 2025
Accepted 21 January 2025
Available online 27 January 2025

ABSTRACT

Breast cancer is commonly acknowledged as the primary type of cancer on a global scale, exerting a substantial influence on death rates, particularly in developing countries. The aforementioned discovery provides evidence in favor of the concept that genetic factors may contribute to the onset of breast cancer. This paper presents the unprocessed idat data containing single nucleotide polymorphisms (SNPs) acquired from breast cancer patients and a control group comprising

* Corresponding authors.
E-mail addresses: sonar.soni@ui.ac.id (S.S. Panigoro), fadilah.msi@ui.ac.id (F. Fadilah).

https://doi.org/10.1016/j.dib.2025.111343

Dataset link: SNP profiling of blood DNA samples of breast cancer patients and healthy women (Original data)

Keywords:
Breast cancer
Blood DNA samples
Gene Expression Omnibus (GEO)
DNA mutations
SNP-array

of healthy women. The DNA was obtained from stored blood samples that were collected from a total of 48 female patients diagnosed with breast cancer at Cipto Mangunkusumo National Hospital Jakarta and Dharmais National Cancer Center Hospital Jakarta. Additionally, 24 healthy women were included as control subjects. Subsequently, the DNA samples were subjected to hybridization onto Infinium Asian Screening Array (ASA)'s beadchips. The chip was then subjected to fluorescence intensity measurements using an iScan machine manufactured by Illumina. The data output is produced in the form of a .idat file for each sample. Subsequently, further quality control measures and population stratification analysis were conducted using PLINK (v1.9). After the conclusion of the quality control procedure, 72 individuals and a dataset consisting of 424,285 genetic variants were selected for further analysis. The idat raw data files have been added to the Gene Expression Omnibus (GEO) with accession number: GSE245794 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245794).

Specifications table

Subject	Cancer research, Bioinformatics.
Specific subject area	Breast cancer patients and healthy women as a control.
Type of data	Raw data: idat files Table: quality control of SNP-array results Figure: Stratification population analysis of research subjects and individuals from 1000 Genomes Project's Superpopulation Supplementary file: Subjects' metadata.
Data collection	The DNA purification process involved the extraction of samples from the blood, specifically targeting the buffy coat. The isolation of DNA was carried out using the Genomic DNA Mini Kit (GeneAid). The DNA isolate's purity was assessed by employing Nanodrop to determine the absorbance ratio of 260/280. The concentration of double-stranded DNA was determined by employing the Qubit® 3.0 Fluorometer and the Qubit dsDNA BR Assay Kit, both manufactured by Thermo Fisher Scientific. In SNP-array analysis, the DNA samples will be hybridized into beadchips. Each of the DNA isolates were subjected to a normalization technique, resulting in a concentration of 50 ng/uL. This standardized concentration was then utilized in the microarray processing method, specifically employing the Infinium Asian Screening Array (ASA)-24 v1.0 reagent. The measurement of fluorescent intensity on the chip are performed using an iScan machine manufactured by Illumina. The data output is generated as a .idat file for each sample, which is subsequently subjected to analysis using the gtc2vcf software [1]. This software facilitates the conversion of the .idat file format to .vcf format. Following this conversion, quality control analysis is performed using PLINK (v1.9) [2]. The recommended thresholds for quality control in genotyping studies are as follows: a genotyping rate of greater than 98%, an SNP missingness rate of less than 0.02, an individual missingness rate of less than 0.02, a gender check threshold for females of less than 0.2, a minor allele frequency (MAF) greater than 0.01, a Hardy-Weinberg Equilibrium (HWE) threshold of less than 0.001, and a heterozygosity rate with a standard deviation less than 3.

(*continued on next page*)

Data source location	This research is a cross-sectional study conducted using stored blood samples that were collected from 48 female patients with breast cancer from Cipto Mangunkusumo National Hospital Jakarta-Indonesia and Dharmais National Cancer Center Hospital Jakarta-Indonesia. Blood samples of 24 healthy women were also collected as controls that were taken from Faculty of Medicine Universitas Indonesia's staff.
Data accessibility	Repository name: Gene Expression Omnibus (GEO) Data identification number: GSE245794 Direct URL to data: (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245794) Subjects' metadata: https://doi.org/10.6084/m9.figshare.28172903.v2
Related research article	None.

1. Value of the Data

- The raw .idat data provide SNP profile of breast cancer patients that may have potential as diagnostics or therapeutics biomarkers.
- The pre-processing steps show better quality data for further analysis
- The utilisation of the dataset has the potential to expedite cancer genetic research in the pursuit of precision oncology

2. Background

Breast cancer is widely recognized as the predominant form of cancer globally, exerting a significant impact on mortality rates, particularly in developing nations. According to the statistics provided by GLOBOCAN in the year 2020, there were 2.3 million newly reported cases of cancer, resulting in its ranking as the fifth most prevalent cause of mortality. Furthermore, it was observed that Asia exhibited the highest prevalence of cancer compared to other continents. Breast cancer accounts for the highest incidence of new cancer cases (30.8%) and mortality rates (15.3%) among women in Indonesia [3]. The data suggests that 10% of newly diagnosed breast cancer cases can be attributed to individuals who have a family history of the disease and share similar characteristics. This finding supports the hypothesis that hereditary factors may play a role in the development of breast cancer [4]. Hence, a more effective approach to the management and prevention of breast cancer can be achieved by a comprehensive understanding of its genetic etiology.

The current state of knowledge on the molecular characteristics of breast cancer has advanced, resulting in the acquisition of more biomarkers that can aid in the identification and prediction of targeted therapeutic approaches for breast cancer. A number of genomic investigations have been conducted in Indonesia, one of which is a pilot study on breast cancer patients (without healthy women as a control) in a limited population in Jakarta, performed by Haryono et al (2015) utilizing the SNP-Array technique. According to the findings of this study, a total of 11 mutations were identified, which exhibited a significant correlation with the susceptibility to breast cancer within a limited sample of individuals residing in Indonesia. Nevertheless, a limited number of mutations were observed in the genes under investigation. Specifically, these mutations were found in the introns of CTNNA2, SOGA2, SSBP2, and TEX10 genes [5].

Advancements in sequencing and microarray technology, along with the development of computational tools, have facilitated the utilization of genomic sequencing in clinical settings. These improvements have contributed to the therapeutic significance of genomics in the context of cancer treatment [3]. In this study, we report the raw .idat data encompassing single nucleotide polymorphisms (SNPs) obtained from both breast cancer patients and a control group of healthy women. The data collected from this project will contribute to the comprehension of breast cancer etiology, facilitate the prediction of suitable therapeutic interventions, and enable prognostic assessments based on germline gene mutations linked to the disease.

Table 1
Data quality control of SNP-array results.

No	Parameters	Before filtering	After filtering
1	Genotyping rate (>98%)	72 subjects 656890 variants	72 subjects 656890 variants
3	*SNP missingness* (<0.02)	72 subjects 656890 variants	72 subjects 640800 variants
2	*Individual missingness* (<0.02)	72 subjects 640800 variants	72 subjects 640800 variants
4	*Gender check* (Female: <0.2)	72 subjects 640800 variants	72 subjects 640800 variants
5	*Minor Allele Frequency* (MAF; >0.01),	72 subjects 640800 variants	72 subjects 429993 variants
6	*Hardy-Weinberg Equilibrium* (HWE; <0.001)	72 subjects 429993 variants	72 subjects 429723 variants
7	*Heterozygosity rate* (standard deviation < 3)	72 subjects 429723 variants	72 subjects 429723 variants
8	*Variants duplications*	72 subjects 429723 variants	72 subjects 424285 variants

3. Data Description

The outcomes derived from SNP-array were intensity images of each single nucleotide polymorphism (SNP) present in the sample, which are identified by the principle of hybridization between the oligonucleotide probe sequence and the DNA sequence within the sample. The output of the scanning process yields images stored in the .idat file format, comprising two distinct intensity colors: red and green. The file is subsequently transformed into the variant calling format (.vcf) using the gtc2vcf pipeline [1]. Following this, the file is further converted into the .bim, .bed, and .fam formats using PLINK 1.9 [2] for subsequent analysis.

The quality control procedures were conducted utilizing PLINK 1.9, Unix Terminal, and RStudio software. During the data quality control phase, various filtering steps were conducted to ensure the reliability of the data. These steps included assessing the genotyping rate, which was required to be greater than 98%. Additionally, the levels of SNP missingness, individual missingness, and gender check were examined, with thresholds set at less than 0.02 for each. The minor allele frequency (MAF) was also considered, with a minimum threshold of 0.01. Furthermore, the data was evaluated for adherence to Hardy-Weinberg Equilibrium (HWE), with a significance level set at less than 0.001. The heterozygosity rate was assessed, with a standard deviation threshold of less than 3 [6]. Finally, variant duplications were identified and documented in Table 1. Following the completion of the screening process, a total of 72 subjects were retained for examination. These subjects exhibited a high genotyping rate of 0.999432 and possessed a considerable number of variations, specifically 424285, which would be subjected to additional investigation.

In addition to several pre-existing quality control measures, the inclusion of population stratification is a crucial stage in ensuring quality control. Population stratification is conducted in order to examine the dispersion of persons from diverse ethnic backgrounds within certain research. The presence of variations in allele frequencies among subpopulations can result in population stratification, which has the potential to produce erroneous positive connections and obscure genuine associations [7].

The utilization of the multidimensional scaling (MDS) methodology is one of the population stratification approaches employed in this study. The present methodology computes the mean fraction of alleles shared between pairs of people within a given sample, resulting in a quantitative indicator (referred to as a component) that characterizes the genetic variation of each individual. The plotting of individual component scores can serve as a method to ascertain the presence of clusters of individuals exhibiting higher degrees of genetic similarity [7].

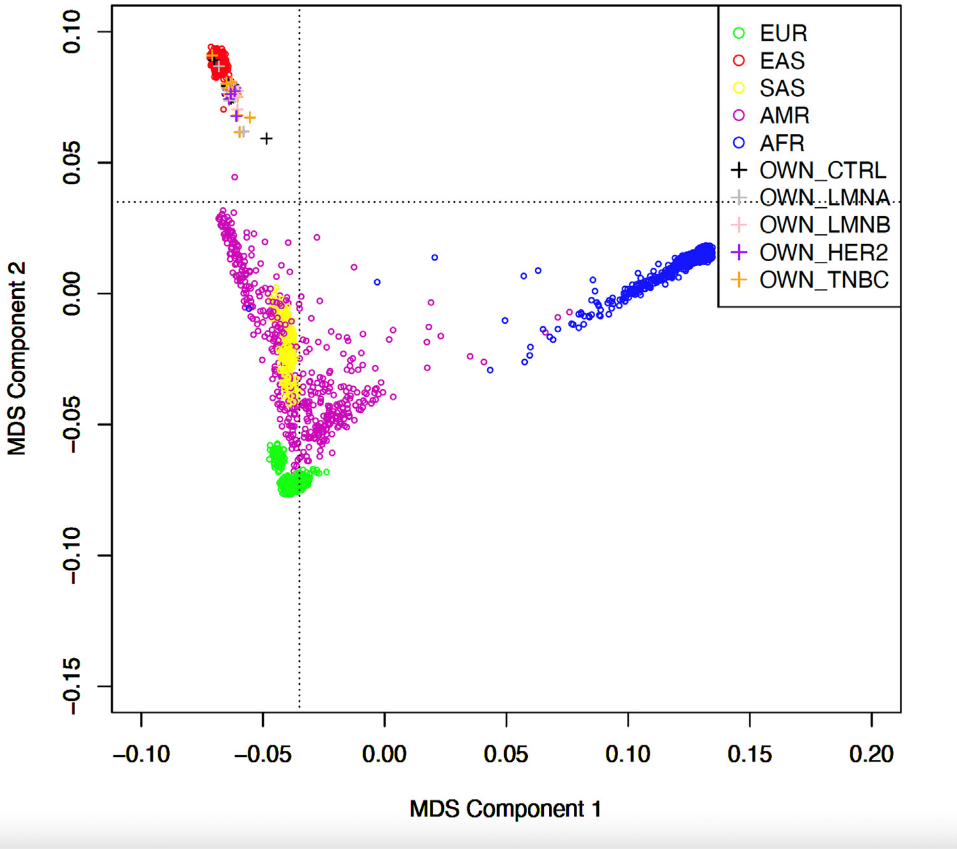

Fig. 1. Stratification population analysis of research subjects and individuals from 1000 Genomes Project's Superpopulation.
EUR: European subjects from 1000 Genomes Project; EAS: East Asian subjects from 1000 Genomes Project; SAS: South Asian subjects from 1000 Genomes Project; AMR: American subjects from 1000 Genomes Project; AFR: African subjects from 1000 Genomes Project; OWN_CTRL: Our healthy control subjects; OWN_LMNA: Our Luminal A Breast Cancer Patients; OWN_LMNB: Our Luminal B Breast Cancer Patients; OWN_HER2: Our HER2+ Breast Cancer Patients; OWN_TNBC: Our TNBC Breast Cancer Patients.

A score plot (Fig. 1) was generated to compare our subjects with population of known ethnic structure. This population consisted of 3202 individuals from five superpopulations: East Asia, South Asia, Africa, Europe, and America that were part of the 1000 Genomes Project [8]. None of our subjects were included to the 1000 Genomes Project. Before proceeding with further study, it is necessary to eliminate individuals who are considered outliers based on MDS analysis. The findings from the analysis of population stratification indicate that healthy female participants serving as controls, as well as breast cancer patients from Indonesia, form a distinct subgroup that exhibits proximity to the East Asian superpopulation. This superpopulation includes individuals from various East Asian regions, such as Chinese Dai in Xishuangbanna, China; Han Chinese in Beijing, China; Southern Han Chinese, China; Japanese in Tokyo, Japan; and Kinh in Ho Chi Minh City, Vietnam. Furthermore, no instances of population outliers were observed within this subgroup.

4. Experimental Design, Materials and Methods

4.1. Sample collection

This research is a cross-sectional study conducted using stored blood samples that were collected from 48 female patients with breast cancer from Cipto Mangunkusumo National Hospital Jakarta and Dharmais National Cancer Center Hospital Jakarta. Blood samples of 24 healthy women were also collected as controls.

4.2. DNA purification

The DNA purification process involved the extraction of samples from the blood, specifically targeting the buffy coat. The buffy coat is a component of whole blood that has a high concentration of leukocytes. This process has the potential to yield a significantly higher quantity of DNA, ranging from five to ten times the initial amount. Whole blood can be separated into three distinct fractions, namely plasma, buffy coat, and erythrocytes, by subjecting it to centrifugation at a speed of 2500 times the force of gravity (2500x g) for a duration of 10 minutes at room temperature, which typically ranges from 15 to 25 degrees Celsius. During this process, the plasma, which appears transparent, will be found in the top layer, followed by the buffy coat in the middle layer, and the erythrocytes in the bottom layer. The utilization of the buffy coat is advisable in cases where more DNA outcomes are required. The isolation of DNA was carried out using the Genomic DNA Mini Kit (GeneAid). The DNA isolate's purity was assessed by employing Nanodrop to determine the absorbance ratio of 260/280. The concentration of double-stranded DNA was determined by employing the Qubit® 3.0 Fluorometer and the Qubit dsDNA BR Assay Kit, both manufactured by Thermo Fisher Scientific [9].

4.3. Beadchips preparations

In SNP-array analysis, the DNA samples will be hybridized into beadchips. Each of the DNA isolates were subjected to a normalization technique, resulting in a concentration of 50 ng/uL. This standardized concentration was then utilized in the microarray processing method, specifically employing the Infinium Asian Screening Array (ASA)-24 v1.0 reagent. The microarray process typically spans three days. On the initial day, DNA amplification is conducted, followed by an incubation period lasting approximately 20 hours. The subsequent day involves enzymatic fragmentation, alcohol precipitation, DNA resuspension, DNA hybridization onto a chip in the capillary flow-through chamber, and an incubation period of roughly 16 hours. Finally, on the third day, enzymatic extension, fluorescent staining, and measurement of fluorescent intensity on the chip are performed using an iScan machine manufactured by Illumina [10].

4.4. Data quality control of SNP-array results using bioinformatics tools

The data output is generated as a .idat file for each sample, which is subsequently subjected to analysis using the gtc2vcf software [1]. This software facilitates the conversion of the .idat file format to .vcf format. Following this conversion, quality control analysis is performed using PLINK (v1.9) [2]. The recommended thresholds for quality control in genotyping studies are as follows: a genotyping rate of greater than 98%, an SNP missingness rate of less than 0.02, an individual missingness rate of less than 0.02, a gender check threshold for females of less than 0.2, a minor allele frequency (MAF) greater than 0.01, a Hardy-Weinberg Equilibrium (HWE) threshold of less than 0.001, and a heterozygosity rate with a standard deviation less than 3 [6].

The subsequent phase of quality control entailed the use of population stratification analysis. The analysis of population stratification was conducted with datasets obtained from the 1000 Genomes Project. The dataset comprises a total of 3202 individuals, representing five distinct superpopulations: East Asia, South Asia, Europe, America, and Africa. The reference genome utilized in this study is GRCh38.p13. The utilization of the multidimensional scaling (MDS) methodology is one of the population stratification approaches employed in this study. The present methodology computes the mean ratio of alleles shared between pairs of people in a given sample, so generating a quantitative measure (component) of genetic diversity for each individual. The plotting of individual component scores can be utilized as a method to ascertain the presence of clusters of individuals exhibiting higher degrees of genetic similarity [7].

Limitations

Not applicable

Ethics Statement

This research was approved by the Faculty of Medicine Universitas Indonesia Ethical Committee (approval number: 0450/UN2.F1/ETIK/2018) for blood sample collection and Faculty of Medicine Universitas Indonesia Ethical Committee (approval number: KET-1140/UN2.F1/ETIK/PPM.00.02/2023) for SNP-array profiling.

Credit Author Statement

Rafika Indah Paramita: Conceptualization, methodology, formal analysis, writing—original draft preparation. **Sonar Soni Panigoro**: Conceptualization, methodology, sample collections, supervision. **Fadilah Fadilah:** Conceptualization, methodology, supervision, writing-review and editing. **Septelia Inawati Wanandi:** Conceptualization, methodology, supervision. **Noorwati Sutandyo:** Sample collections, supervision. All authors read and approved the final manuscript.

Data Availability

SNP profiling of blood DNA samples of breast cancer patients and healthy women (Original data) (GEO).

Acknowledgements

This work was supported by Doctoral Grant year 2024 from The minister of education, culture, research, and technology – Republic of Indonesia (Grant number: 051/E5/PG.02.00.PL/2024).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Genovese G. gtc2vcf pipeline 2019.

[2] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, P.C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (3) (2007) 559–575, doi:10.1086/519795.

[3] G.C. Observatory, Breast Fact Sheet 419 (2020).

[4] O.A. Zayas-Villanueva, L.D. Campos-Acevedo, J.D.J. Lugo-Trampe, D. Hernández-Barajas, J.F. González-Guerrero, L.E. Martínez-de-Villarreal, Analysis of the pathogenic variants of BRCA1 and BRCA2 using next-generation sequencing in women with familial breast cancer: a case–control study, BMC Cancer 19 (1) (2019) 722, doi:10.1186/s12885-019-5950-4.

[5] S.J. Haryono, I.G.B. Datasena, W.B. Santosa, R. Mulyarahardja, K. Sari, A pilot genome-wide association study of breast cancer susceptibility loci in Indonesia, Asian Pacific J. Cancer Prev. 16 (6) (2015) 2231–2235, doi:10.7314/APJCP.2015.16.6.2231.

[6] J. Liu, S. Li, X. Li, W. Li, Y. Yang, X.J. Luo, Genome-wide association study followed by trans-ancestry meta-analysis identify 17 new risk loci for schizophrenia, BMC Med. 19 (1) (2021) 1–15, doi:10.1186/s12916-021-02039-9.

[7] A.T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, E.M. Derks, A tutorial on conducting genome-wide association studies: quality control and statistical analysis, Int. J. Methods Psychiatr. Res. 27 (2) (2018) 1–10, doi:10.1002/mpr.1608.

[8] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, M.C. Zody, High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, Cell 185 (18) (2022) 3426–3440 e19, doi:10.1016/j.cell.2022.08.004.

[9] S.S. Panigoro, K.M. Siswiandari, R.I. Paramita, F. Fadilah, L. Erlina, Targeted genome sequencing data of young women breast cancer patients in Cipto Mangunkusumo national hospital, Jakarta, Data Brief 32 (2020), doi:10.1016/j.dib.2020.106138.

[10] Illumina, Infinium HTS Assay Ref. Guide (2019).