nature portfolio

Corresponding author(s): Young Seok Ju, Yeon Hee Park, Jeongmin Lee

Last updated by author(s): Aug 11, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All sequences were generated by Illumina sequencer (Novaseq 6000).
Data analysis	Whole-genome sequencing (WGS) was performed using the CancerVision™ assay (Inocras, San Diego, CA, USA). Sequencing reads were aligned to the GRCh38 reference genome using bwa-mem (v0.7.17-r1188). Somatic SNVs and indels were called using Mutect2 (GATK v4.0) and Strelka2 (v2.9.10). Tumor purity, ploidy, and segmented copy number profiles were estimated with Sequenza (v3.0.0). Somatic structural variants were identified using Delly (v0.7.6). Variant annotation was performed using the Ensembl Variant Effect Predictor (VEP, release 112), and VCF files were processed with bcftools (v1.9). Recurrent copy number alterations were detected using GISTIC (v2.0.23), and extrachromosomal DNA (ecDNA) amplicons were identified using AmpliconArchitect (v1.2). For mutational signature analysis, an in-house algorithm was validated using SigProfiler (v3.4) and SignatureAnalyzer (v0.0.8). Driver gene detection was performed using the IntOGen pipeline (v2023). Statistical analyses were conducted in R (v3.4.0) with associated packages, including DESeq2 (v1.26.0) for RNA-seq differential expression analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Whole-genome and transcriptome sequencing data generated in this study are publicly available. FASTQ files: Korean Sequence Read Archive (KRA), accession KAP240849 (https://kbds.re.kr/KRA). CRAM files: European Genome Phenome Archive (EGA), accession EGAD50000001546 (https://ega-archive.org). The human reference genome GRCh38 was used for alignment (Genome Reference Consortium: https://www.ncbi.nlm.nih.gov/grc/human).

Previously published datasets analyzed as summary-level data:
• TransNEO cohort: Sammut, S.-J. et al. Nature 601, 623–629 (2022). DOI: 10.1038/s41586-021-04278-5.
• METABRIC dataset: Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. 7, 11479 (2016). DOI: 10.1038/ncomms11479.

All other data supporting the findings are available within the paper and Supplementary Information.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Although patients were not recruited to the study according to any sex- or gender-based criteria, all participants were female reflecting exclusively higher prevalence of breast cancer in female. Accordingly, exploratory sex-specific analyses could not be performed.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity, or other socially relevant groupings were not considered in study design.
Population characteristics	The study cohort consists of 1,364 breast cancer patients, with 904 patients from a prospective cohort and 460 patients from a retrospective cohort (Supplementary Table 1). The retrospective cases predominantly had early-stage disease and underwent tumor sampling shortly after initial diagnosis, with sequencing performed on surgical specimens obtained during curative-intent surgery. Key clinical characteristics are: Age: The median age of participants was 44.0 years (range: 21.0–85.0), with significant differences between the two cohorts (prospective: 39.0.3 years, retrospective: 51.0 years; p<0.001). Gender: All patients were female. Menopausal status: 69.4% of patients were premenopausal, with a significantly higher proportion in the prospective cohort (78.8%) compared to the retrospective cohort (50.9%). Biopsy site: The majority of samples were obtained from the breast (93.0%), with other sites including lymph nodes, liver, lung, pleura, brain, and bone. Histological subtypes: The most common subtype was invasive ductal carcinoma (86.4%), followed by invasive lobular carcinoma (3.8%) and mucinous carcinoma (1.9%). Hormone receptor/HER2 status: Patients were classified as hormone receptor-positive, HER2-negative (63.4%), HER2-positive (20.9%), and triple-negative breast cancer (15.7%), with no significant differences between the two cohorts (p=0.185). Initial clinical stage: There was a significant difference in clinical stage distribution (p<0.001), with a higher proportion of early-stage patients (stage I-II) in the retrospective cohort, while more advanced stages (stage III-IV) were seen in the prospective cohort. Notably, two patients in the cohort (cases #729 and #1289) were initially diagnosed with ductal carcinoma in situ (DCIS) based on surgical pathology. However, case #729 later developed a metastatic lesion in the lung, which was collected and subjected to whole-genome sequencing. In the other case (#1289), although the tumor retained a non-invasive histologic classification, it exhibited an high somatic mutation burden (carrying 3,213 point mutations) and extensive copy number alterations, suggesting a biologically more advanced disease state. Although genomes of these cases were included in this study, they were excluded from all outcome-related analyses. Curative surgery: The majority of patients (92.9%) underwent curative surgery, with higher rates in the retrospective cohort (98.9%) compared to the prospective cohort (89.8%) (p<0.001). Distant metastasis: Among 1,364 patients, 247 (18.1%) developed distant metastasis, while metastatic status remained unknown for 1,117 (81.9%). A higher proportion of patients in the prospective cohort (20.9%) developed distant metastasis compared to the retrospective cohort (12.6%; p<0.001). This difference is likely due to the higher inclusion of initially late-

	stage breast cancer patients in the prospective cohort, which inherently carries a greater risk of disease progression and metastasis. Follow-up duration: The follow-up duration was measured from the time of diagnosis. The median follow-up duration from diagnosis was 54.8 months (range: 0.4–215.0 months), with a significantly shorter follow-up period in the prospective cohort (23.3 months) compared to the retrospective cohort (54.5 months; p<0.001).
Recruitment	Participants were recruited from Samsung Medical Center (Seoul, Korea) and Seoul St. Mary's Hospital (Seoul, Korea) in prospective and retrospective manner between 2012 and 2023. In the retrospective cohort, patients were included based on the availability of archived primary tumor and matched normal blood sample, as well as sufficient clinical information for downstream analysis. Case enrollment in the retrospective study was independent of death status at the time of study accrual. All retrospective samples were obtained from patients shortly after initial diagnosis, typically at the time of curative-intent surgery. No specific selection criteria were applied with respect to disease stage, including metastatic status at the time of accrual. Detailed study design and inclusion and exclusion criteria, along with the CONSORT diagram for this study, are available in the Supplementary Methods and Supplementary Fig. 1.
Ethics oversight	The study was approved by the Institutional Review Boards of both institutions (Samsung Medical Center: SMC 2022-05-050, SMC 2013-04-005; Seoul St. Mary's Hospital: KC21TISI0007, KC22TISI0292) and conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines. Written informed consent, including consent to publish de-identified clinical and genomic data, was obtained from all prospective participants. For retrospective cases, the use and publication of de-identified data were approved by the relevant institutional review boards, with a waiver of informed consent where applicable. The study design was informed by, and aligned with, the clinical trials NCT03131089 and NCT06334471.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was performed for this integrated analysis. The study combined whole-genome sequencing datasets from four independent breast cancer cohorts, each of which had predefined enrolment targets determined by their original study objectives and design. All available samples from these cohorts that met our inclusion criteria and had high-quality sequencing data were included in the analysis. The resulting aggregated sample size (n = XXX) provided sufficient statistical power to detect genomic alterations and mutational patterns of interest, as evidenced by the robust effect sizes, and statistically significant associations observed in key analyses.
Data exclusions	Samples with tumor-normal swaps or no detectable somatic mutations were excluded from the analysis.
Replication	This study is based primarily on computational analyses of pre-existing sequencing datasets; therefore, no wet-lab experiments requiring repeated measurements were performed. All analyses were conducted once on the complete datasets. Where applicable, key findings were assessed for reproducibility using independent external datasets (e.g., TransNEO, METABRIC) and showed consistent results.
Randomization	Not applicable since this was not an intervention-based study.
Blinding	Blinding was not applicable to this study because all analyses were performed on pre-existing, de-identified genomic and transcriptomic datasets obtained from independent cohorts. No prospective experimental interventions, clinical outcome assessments, or subjective measurements requiring blinding were conducted. The computational analyses were performed using predefined pipelines, and thus the possibility of observer bias was minimal.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	A subset of participants was enrolled as part of substudy components within the clinical trials NCT03131089 and NCT06334471, which informed specific aspects of the study design.
Study protocol	This study does not have a single overarching clinical trial protocol, as it integrates data from multiple independent cohorts. The referenced trial registrations (NCT03131089 and NCT06334471) apply only to specific substudy cohorts; their full protocols are available via ClinicalTrials.gov (https://clinicaltrials.gov/study/NCT03131089, https://clinicaltrials.gov/study/NCT06334471). No public protocol exists for the remaining institutional cohorts, as these were conducted as independent institutional studies outside the scope of a registered clinical trial.
Data collection	Participants were recruited from Samsung Medical Center (Seoul, Korea) and Seoul St. Mary's Hospital (Seoul, Korea) in prospective and retrospective manner between 2012 and 2023. For retrospective tissue archival, participants were included if they were histologically diagnosed with breast cancer and had breast cancer tissues collected between 2012 and 2023, stored in the Samsung Medical Center BioBank with sufficient amount for sequencing.
Outcomes	The primary outcome—to elucidate the individual genetic traits of breast cancer patients and construct the whole-genome landscape of breast cancer—was pre-defined during the initial study design phase. This was assessed by comprehensive analysis of somatic and germline variants, mutational signatures, copy number alterations, and structural variations derived from whole-genome sequencing (WGS) data. Bioinformatic processing involved mapping reads to GRCh38 and subsequent variant calling, annotation, and classification using established tools (bwa-mem, Mutect2, Strelka2, Sequenza, Delly, GISTIC, AmpliconArchitect, VEP, and bcftools) followed by manual curation. The secondary outcome—to investigate associations between WGS-derived genomic features and clinical outcomes—was pre-defined as statistical analysis correlating genomic metrics (e.g., HRD score, APOBEC mutational signature, focal amplifications, MATH score, TMB) with survival and treatment response. This was assessed using Kaplan–Meier survival analysis, multivariate Cox regression, and correlation tests, with validation in publicly available external datasets (e.g., METABRIC, TransNEO) where applicable.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*