Genomic alterations converge on lipid metabolism and ERBB2 in breast cancer

Divya Khanna[1#], Arnab Ghosh[2#], Priyanka Bhadwal[1$], Leepakshi Dhingra[1$], Ramakanth Chirravuri-Venkata[3$], Subrata Das[2$], Isha Choubey[3], Ankita Singh[1], Geeta Jadaun[1], Hriday Home Chowdhury[1], Suryanshi Tiwari[1], Vibha Mattoo[1], Pooja Dixit[1], Safeer Khan[1], Vinit Gupta[1], Tuneer R. Mallick[2], Surya Narayan Mishra[3], Tulasi Nagabandi[3], Venkata Krishna Vanamamalai[3], Malini Nemalikanti[3], Renu Kumari[1], Diksha Singhal[1], Laxmi Yadav[1],Vaishali Pandey[1], Shailya Verma[1], Rashmi Gudur[4] , Senthil Kumar[9], Susanta Roychoudhury[5], Rohan Chaubal[6], Sonam Dhamija[1], Koushik Mondal[5], Lipi Thukral[1], Aruna Korlimarla[7], Rekha V Kumar[7], Ashutosh Mishra[8], Sandeep Mathur[8], Bhawna Sirohi[10], SVS Deo[8,9], Samir Bhattacharya[5], B S Srinath[7], Juhi Tayal[11], Anurag Mehta[11], IBCGA Consortium, Shilpak Chatterjee[12] , Sanjeev Khosla[13], Divya Tej Sowpati[3], Kumardeep Chaudhary[1], Karthik Bharadwaj Tallapaka[3✉], Sherry Bhalla[1], Nidhan K. Biswas[2✉], Sudeep Gupta[6,14✉], Shantanu Chowdhury[1,15,16✉].

[1]CSIR Institute of Genomics and Integrative Biology, South Campus, Opp. Bus Depot, Sukhdev Vihar, Mathura Road, New Delhi – 110025

[2]Biotechnology Research and Innovation Council-National Institute of Biomedical Genomics (BRIC-NIBMG), P.O. NSS, Kalyani, West Bengal, India-741251

[3]CSIR- Centre for Cellular & Molecular Biology, Habsiguda, Uppal Road, Hyderabad - 500 007, Telangana.

[4]Krishna Vishwa Vidyapeeth, Dhebewadi Road, Malkapur, Karad, Taluka Karad, District Satara, Maharashtra, India Karad 415539

[5]Saroj Gupta Cancer Centre & Research Institute, Mahatma Gandhi Road, Thakurpukur, Kolkata - 700 063.

[6]Tata Memorial Hospital (TMH), is Dr. Ernest Borges Marg, Parel, Mumbai-400012, Maharashtra

[7]Sri Shankara Cancer Hospital and Research Centre, 1st Cross, Shankara Math Premises, Shankarapuram, Basavanagudi, Bengaluru, Karnataka 560 004.

[8]All India Institute of Medical Sciences, Ansari Nagar, New Delhi – 110029

[9]Oncology Lead & Senior Consultant Surgical Oncology, Apollo Hospitals, New Delhi

[10] BALCO Medical centre, Vedanta Medical Research Foundation, Raipur, Sector-36, Atal Nagar, PO: Uparwara, Raipur, Chhattisgarh 493661.

[11]Rajiv Gandhi Cancer Institute and Research Center, D-18, Sector-5, Rohini, Delhi-110085

[12]CSIR-Indian Institute of Chemical Biology, Raja S.C. Mullick Road, Kolkata, India

[13]CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh – 160036

[14]Department of Medical Oncology, Homi Bhabha National Institute, Mumbai, India.

[15]Integrative and Functional Biology Unit, CSIR Institute of Genomics and Integrative Biology, New Delhi 110025, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India.

[16]Current Address: Bharat Charitable Cancer Hospital & Institute (Trust), Mysuru, 570024, India

✉ Corresponding author
shantanuc@igib.in
sudeep.gupta@actrec.gov.in
nkb1@nibmg.ac.in
karthikt@ccmb.res.in
# First Authors: equal contribution, $ Shared second author

Table of Contents

Supplementary Methods

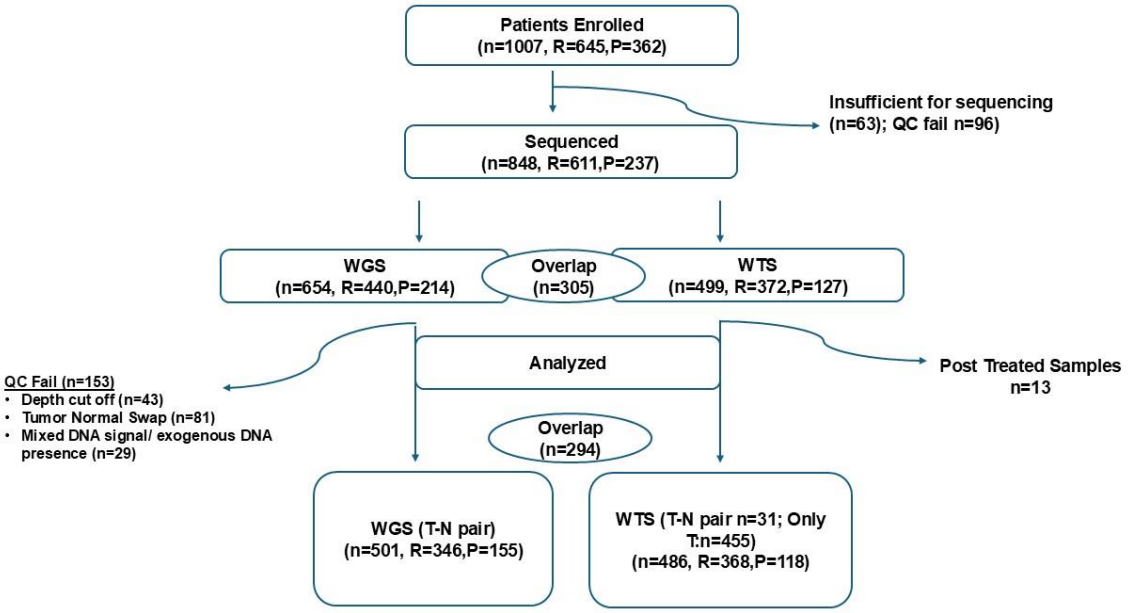
Study cohort and sample processing overview

This ambispective study included both prospectively and retrospectively collected, treatment-naïve breast cancer samples from women aged ≥18 years. Retrospective cases with at least three years of follow-up were obtained from hospital biorepositories after ethical approvals. While the prospective recruitment was conducted across collaborating hospitals in India, with written informed consent obtained from all participants.

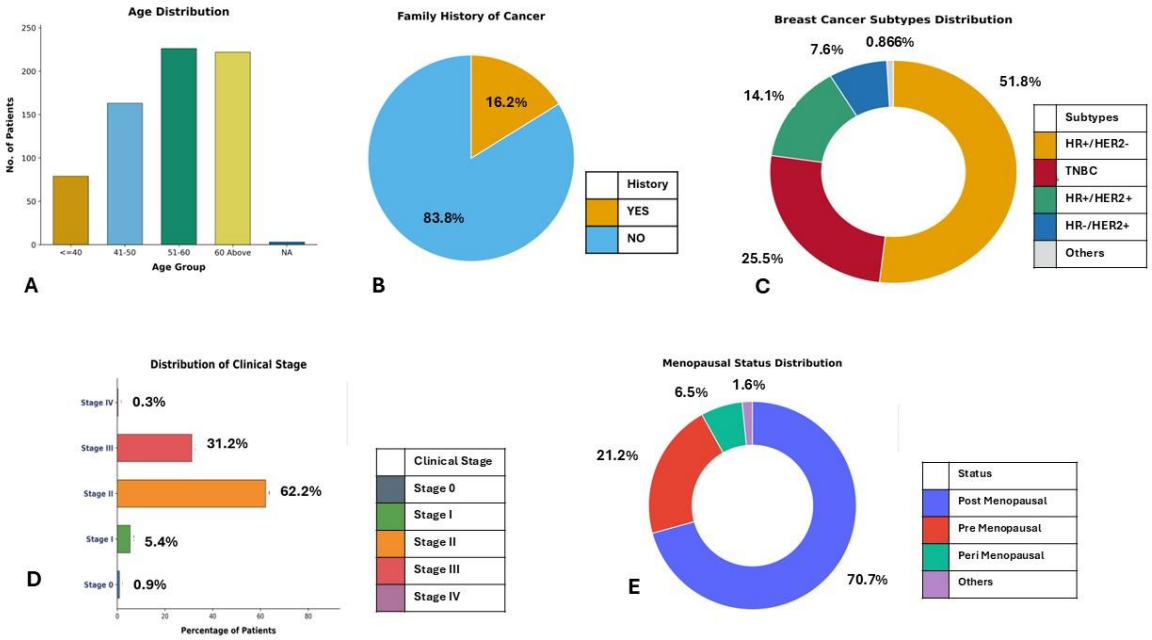
1,007 treatment-naive patients were enrolled in the study including retrospective and prospective cases. Of these, 848 (611 retrospective, 237 prospective) had sufficient tissue for sequencing, after excluding 63 cases with insufficient material and 96 samples with isolation QC failures during DNA/RNA isolation (Supplementary Figure S1). Whole genome sequencing (WGS) was attempted in 654 cases where isolated DNA from paired tumour-normal samples were of adequate quality following WGS library preparation. Whole transcriptome sequencing (WTS) was attempted for 499 patients with adequate RNA quality following library preparation. 305 samples were of adequate quality for both WGS and WTS. Following post-sequencing QC: 153 WGS samples had to be excluded (43 below depth cut-off, 81 tumour/normal swaps, and 29 with mixed or exogenous DNA signal), leaving 501 tumour-normal WGS pairs which were analysed further. In the WTS set, 13 samples were to be found to be collected after treatment had been initiated and therefore excluded to ensure all samples are treatment naïve. Following this we had 486 WTS profiles (455 tumour-only, 31 paired) for analysis (Supplementary Table S12). In total, 294 patients had both WGS and WTS data available. Respective numbers of retrospective and prospective cases in WGS and WTS are given in Supplementary Figure S1.

Clinicopathological characterization of the sequenced cohort (n = 693) was performed across four intrinsic subtypes defined by hormone receptor (HR) and HER2 status: HR+/HER2− (n = 362), HR−/HER2+ (n = 53), HR+/HER2+ (n = 98), and triple-negative breast cancer (TNBC, n = 176), 4 additional cases were processed where IHC was inconclusive and have been classified as "Others."

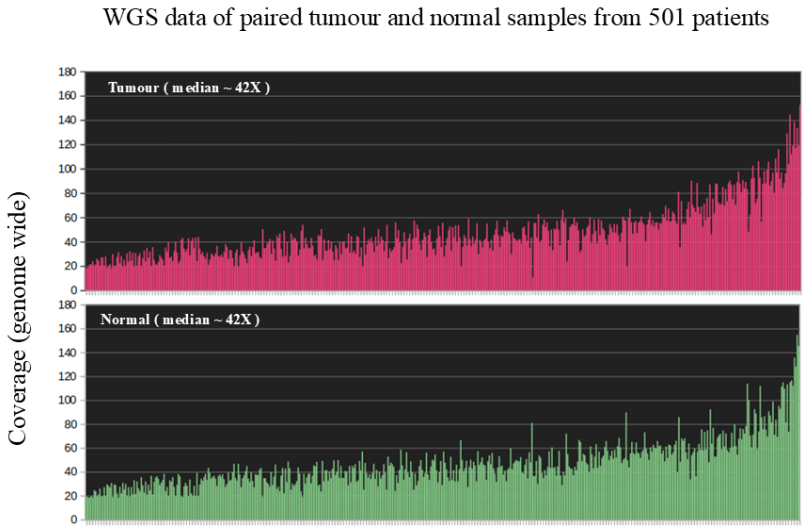
The CONSORT flow (Supplementary Figure S1) and clinical tables (Supplementary Table S1a, Table S24) provide specific characteristics of each patient sample.



Supplementary Figure S1: The CONSORT diagram summarizes patient enrolment, sample processing, sequencing, and downstream analysis, with participants stratified into prospective (P) and retrospective (R) cohorts. The diagram shows sample inclusion, quality control filtering, sequencing modalities (WGS and WTS), and the final datasets used for analysis.



Supplementary Figure S2: Clinicopathological characteristics of the Indian breast cancer cohort. (A) Age distribution of patients at diagnosis, (B) Percentage of patients with a reported family history of cancer, (C) Molecular subtypes of breast cancer, (D) Distribution of cancer stages at presentation, (E) Menopausal status of the cohort.



Supplementary Figure S3: Out of 501 patients 337 (67%) patients have both tumour and normal depth of coverage ≥ 30X. 113 (22%) have both tumour and normal depth of coverage ≥ 50X and 39 (8%) have both tumour and normal depth of coverage ≥ 70X.

Processing and analysis of WGS data

Raw WGS data (40x median depth for paired tumour and normal samples) were demultiplexed with Illumina adapter information through bcl2fastq (version: v2.19.0.316) followed by quality assessment through FastQC (version: 0.11.7) (https://github.com/s-andrews/fastqc). Only high-quality paired-end reads were mapped to the human reference genome (GRCh38) followed by post alignment processing (sorting, duplicate marking, base quality score recalibration, and merging) through BWA-MEM[31], Picard (https://broadinstitute.github.io/picard/) and GATK [1] based accelerated pipeline. Post-alignment QC was done using MapInsights[2] and utilized the CalculateContamination module of GATK (4.0.5.1) to verify tumour-normal parity using the gnomAD[3] resources. Germline joint genotyping was performed from normal samples using gVCF calling with accelerated HaplotypeCaller, followed by joint genotyping with GLnexus (version: v1.4.1), and functional annotation with annovar and ClinVar (version: Release:2025-06-01)[4]. The pathogenic germline variants were identified as per ACMG guidelines for breast cancer. Somatic mutations were identified by jointly analyzing tumour and paired normal samples through Mutect2 (GATK version: 4.0.5.1) using '--normal-lod -10 and --force-active' parameters and gnomAD as germline resources to increase sensitivity. Somatic call sets were further filtered to obtain only high-quality mutations utilizing information such as depth, strand orientation,

variant allele frequency, low complexity regions (homopolymer and/or sequence similarity). Further variant annotations (including pathogenicity prediction by multiple algorithms and InterPro domain annotations) for high-quality somatic mutations were performed through annovar[5], and population-specific germline resources (India-specific in-house control sets including > 1800 individuals). Any polymorphic sites (AF > 0.01 in 1000 Genome[6] populations, population-specific resources) were removed from the somatic mutation call set. We have utilized MutSig2CV[7] for the identification of significantly mutated genes (SMGs), and somatic mutation rate (as mutations per Mb of genome and exome separately), in the overall cohort and within clinical sub-groups. Variants in the SMGs were manually verified through visual inspection using IGV (version: 2.17.4)[8] by two independent individuals, followed by removal of potentially artefactual variants. We utilized a Random Forest framework that integrated pathogenicity scores from 14 existing algorithms to classify missense mutations identified in significantly mutated genes. The tumour and normal WGS data were segmented using dense population-specific germline marker sets through AscatNGS[9], followed by identification of (a) both arm-level and focal copy number amplification and deletions, and (b) significantly amplified and deleted genes through GISTIC (version: v2.0.22)[10]. We have utilized Maftools (version 2.22.0)[11] for statistical assessment of mutual co-occurrence and exclusivity of somatic alteration events in the overall cohort and the clinical subgroups. Selection in driver gene mutations were estimated using the package 'dndscv' [12] using the reference genome and epigenomic covariates data for GRCh38.

Protein Structure-Based Characterization of Somatic Mutations

The genes with mutations were analysed to obtain the list of transcripts with the amino acid changes at respective positions. These files were formatted in maf format and analysed using Maftools to identify the genes with high mutation load. The corresponding data for the top genes were downloaded from Genomic Data Commons (GDC) [13], corresponding to TCGA-BRCA dataset. The mutations of IBCGA were compared with TCGA-BRCA data and unique mutations were selected. These mutations were annotated using ProtVar [14] and features like residue conservation, stability of protein (based on FoldX v5.0 calculations), pathogenicity prediction based on AlphaMissense, CADD phred, ESM-1b score, localization on protein structural domains (based on Pfam annotation), disease associated with the variants (dbSNP, ClinVar, Ensembl, UniProt) were evaluated. These were studied for top four driver genes, TP53, PIK3CA, KMT2C and MAP3K1. With our interest in PIK3CA we further explored and identified mutations (V344G, N345S, E418K, E453Q, Q546E) at the interacting interface with PIK3R1 protein. We further used the crystal structure of the complex (PDB id: 5XGH [15]) and modelled mutated complex, which were further docked using HADDOCK [16] and 100+ structures were generated. These mutated complexes were analysed in comparison to the wildtype structure and the changes in interaction patterns (hydrogen bonding, salt bridge) was observed using inhouse pml script. Interface residues were identified from modelled PIK3CA–PIK3R1/2/3 complexes, with pDockQ used as an overall

model-confidence score[17] . PDB 5XGH corresponds to the p110α/niSH2–p85α heterodimer, so the PIK3R1 interface is directly represented in the experimental structure.

Structural variation

We identified structural variants (SVs) from WGS data of 501 tumours and their matched normal samples using a multi-caller strategy. The aligned BAM files were analysed with three tools—GRIDSS (Genomic Rearrangement Identification Software Suite)[18], Manta[19] and SvABA[20]the results were compiled in VCF format. Variants detected by at least two of the callers were considered high-confidence, resulting in consensus SV calls for 439 samples. GRIDSS (v2.13.2) was run with its default parameters for somatic SV detection. GRIPSS incorporates evidence from discordant read pairs, split reads, and soft-clipping to generate high-confidence SV calls for the identification of translocations and complex rearrangements. Manta (v1.6.0) was used in somatic mode for tumour-normal pairs, detecting a wide range of SV types. It integrates paired-end and split-read signals and is optimized for rapid SV detection. SvABA (v1.1.0) was used to call somatic SVs via a local assembly-based approach. It uses a custom implementation of SGA (String Graph Assembler). Manta and SvABA were used in a chromosome-wise, chunked mode to increase computational efficiency. However, since this strategy may lead to incomplete detection of inter-chromosomal translocations (TRA), the break ends of translocation calls reported by GRIPSS, which was run on the entire genome without chunking. Following SV calling, individual VCF files were merged using Jasmine v1.1.5 to generate a consensus call set. Jasmine clusters SVs across samples and tools based on breakpoint proximity, strand orientation, and SV type. A maximum positional distance threshold of 500 bp was used for merging, and the SVs supported by at least two or three callers were retained to ensure high-confidence consensus calls. The translocations reported by GRIPSS were manually validated for their presence in the merged VCF files. The consensus call sets were annotated with AnnotSV v3.4.4[21]. Then, a BED format file was created focusing on protein-coding regions and guided by a reference GTF annotation file, and coordinates were expanded 2kb upstream using the bedtools slope function. Downstream analysis was performed with in-house custom scripts to process and merge SV calls from three SV callers (GRIDSS, Manta, and SvABA) and to obtain supported variants from Jasmine.

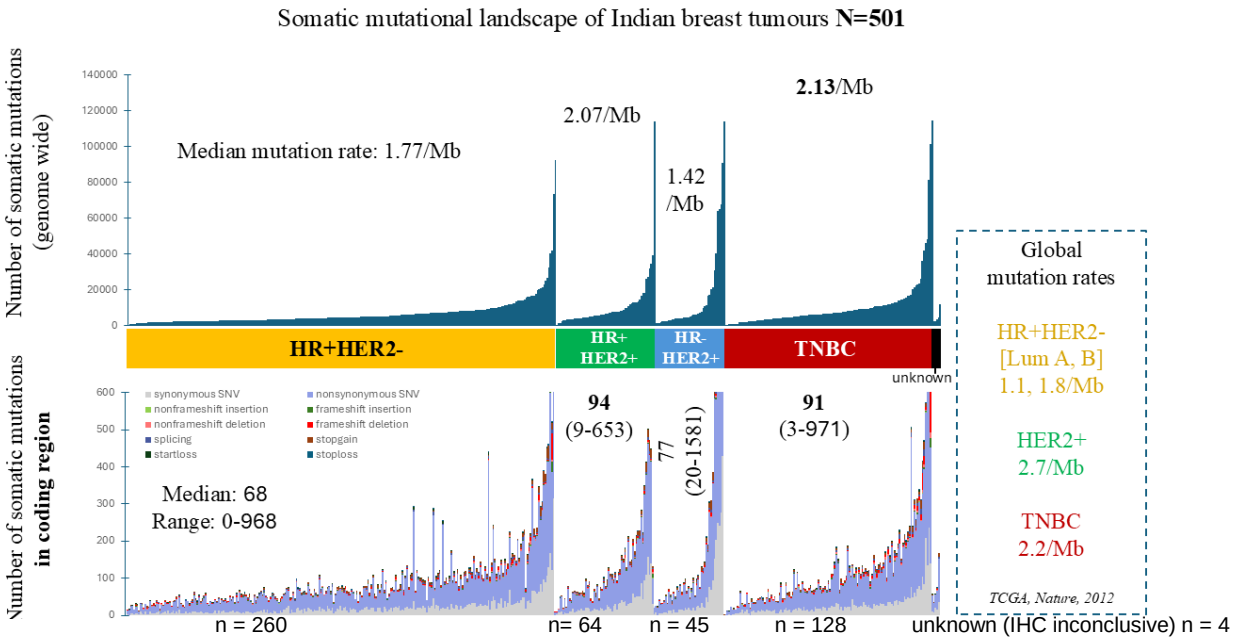
Supplementary Results

Discussion 1:

Detailed somatic mutational landscape of Indian Breast Cancer Genome (IBCGA) cohort

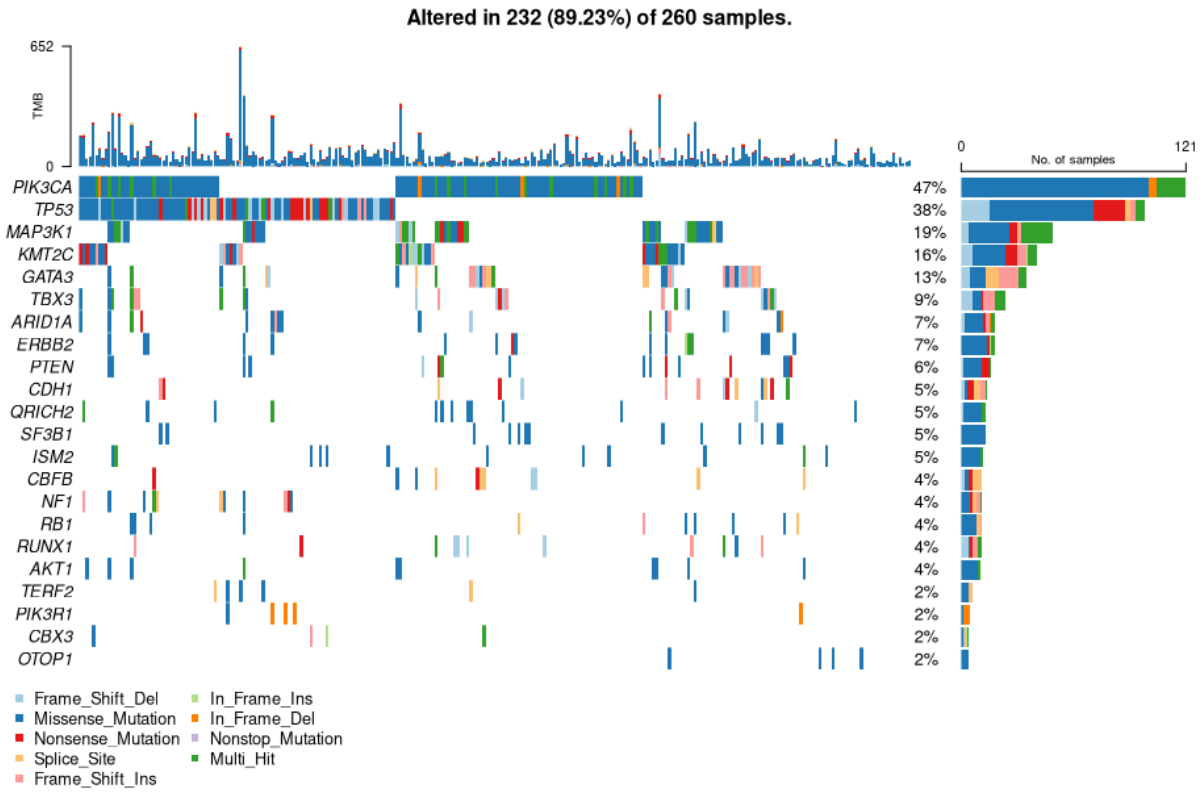
The analysis of 501 paired tumour and normal tissue specimens catalogued a total of 4,593,630 somatic mutations. This comprised 59,234 coding mutations - including 39,141 missense, 2,249 nonsense, 2,640 frameshift/in-frame indels, 13,668 silent, others (nonstop 102, splice 1,097, start loss 337) and 4,534,396 non-coding mutations. The median number of somatic mutations per tumour was 5,209, with a range extending from 158 to 114,814 (Supplementary Table S1b). The variation in mutation burden

across the four major subtypes was statistically significant (p = 1.5e−05, ANOVA). The HER2+ (HR+/-) and TNBC tumours demonstrated a significantly higher number of somatic mutations than HR+/HER2-tumours (p < 0.015) (Supplementary Figure S4).

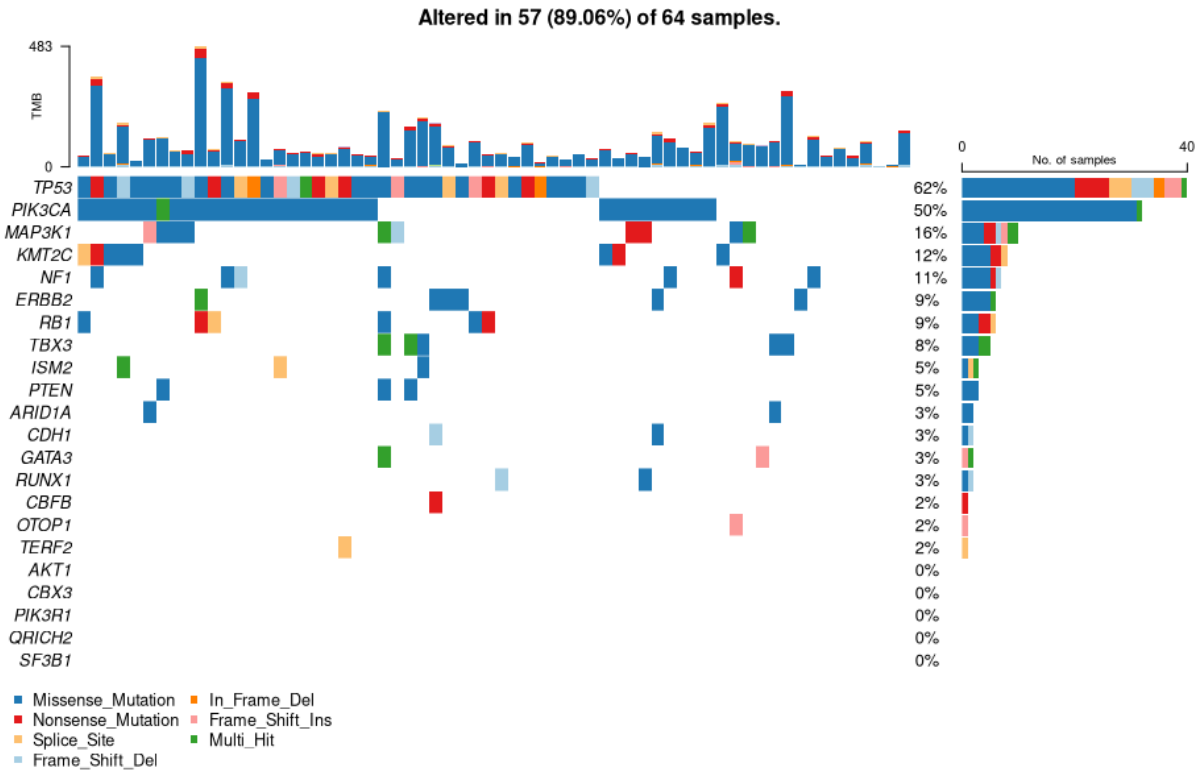


Supplementary Figure S4: The somatic mutation rate (calculated as mutation per Mb of genome) varied significantly (p = 1.5e-05) between subtypes, HR+HER2-, HR+HER2+, HR-HER2+, and TNBC patients harboured 1.42/Mb, 2.07/Mb, 1.78/Mb, and 2.13/Mb genome wide somatic mutation rate with highest obtained for TNBC tumours. The somatic mutation rate of breast tumour from TCGA is shown in right (for five patients IHC was not well defined).

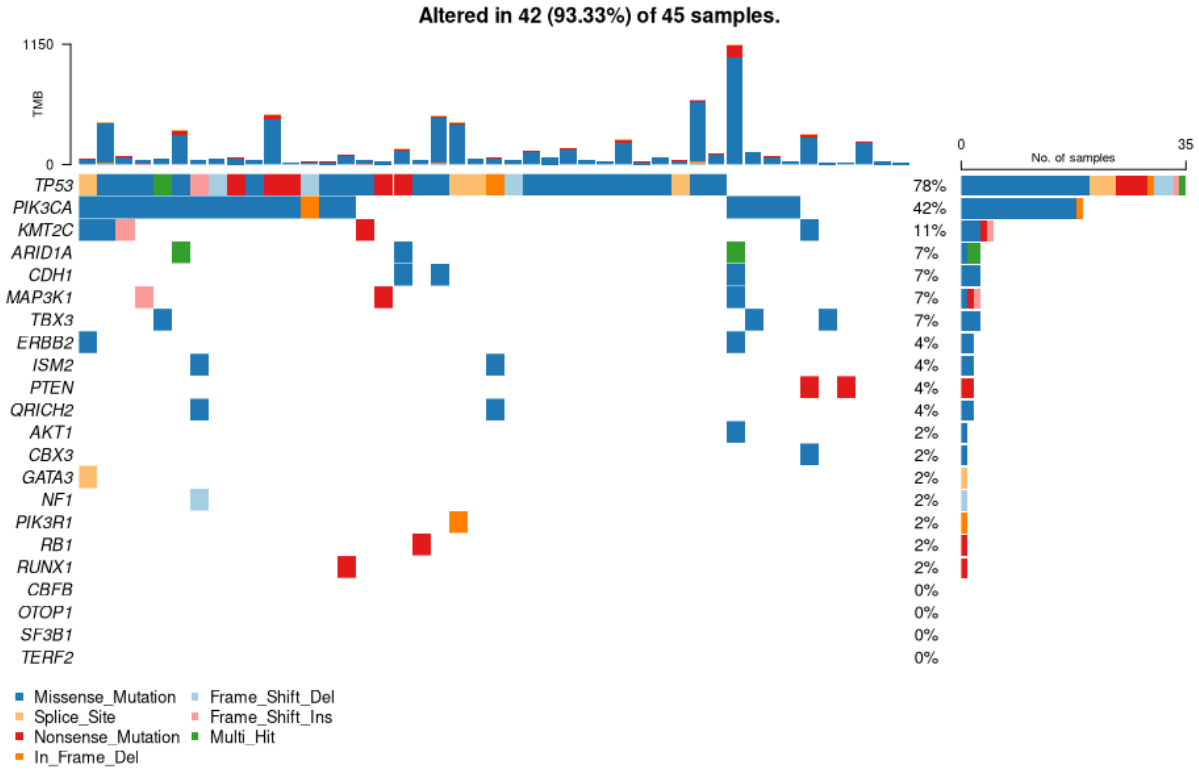
The 22 significantly mutated driver genes—including TP53, PIK3CA, KMT2C, MAP3K1, GATA3, TBX3, PTEN, ARID1A, ERBB2, NF1, RB1, ISM2, CDH1, QRICH2, RUNX1, PABPC1, PIK3R1, SF3B1, AKT1, CBFB, CBX3, and OTOP1—were identified in the cohort, as detailed in (Supplementary Table S2, Supplementary Figure S5, S6, S7, S8).



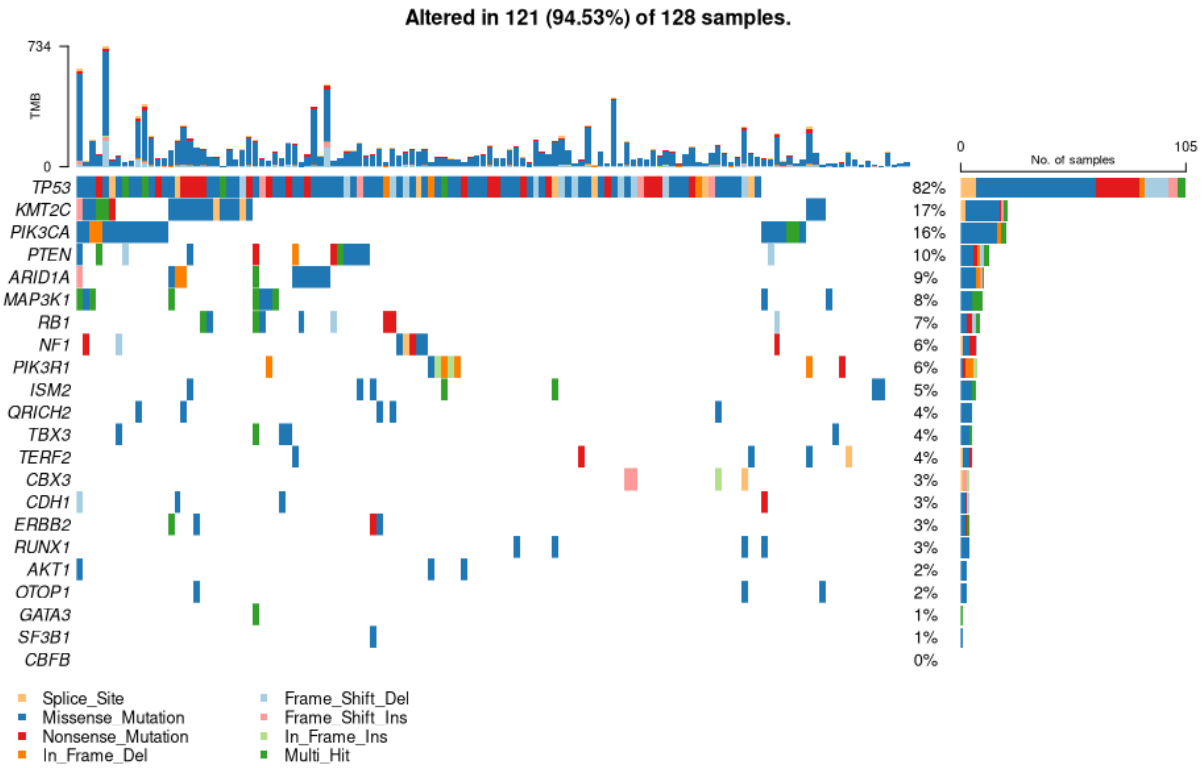
Supplementary Figure S5: Somatic mutational landscape of 260 HR+HER2- patients. A median of 4542 somatic mutations ranging between 158 and 92699. Somatic alteration landscape of 22 breast cancer driver genes in this immunohistochemistry subtype is illustrated.



Supplementary Figure S6: Somatic mutational landscape of 64 HR+HER2+ patients. A median of 6632 somatic mutations ranging between 653 and 114273. Somatic alteration landscape of 22 breast cancer driver genes in this immunohistochemistry subtype is illustrated.

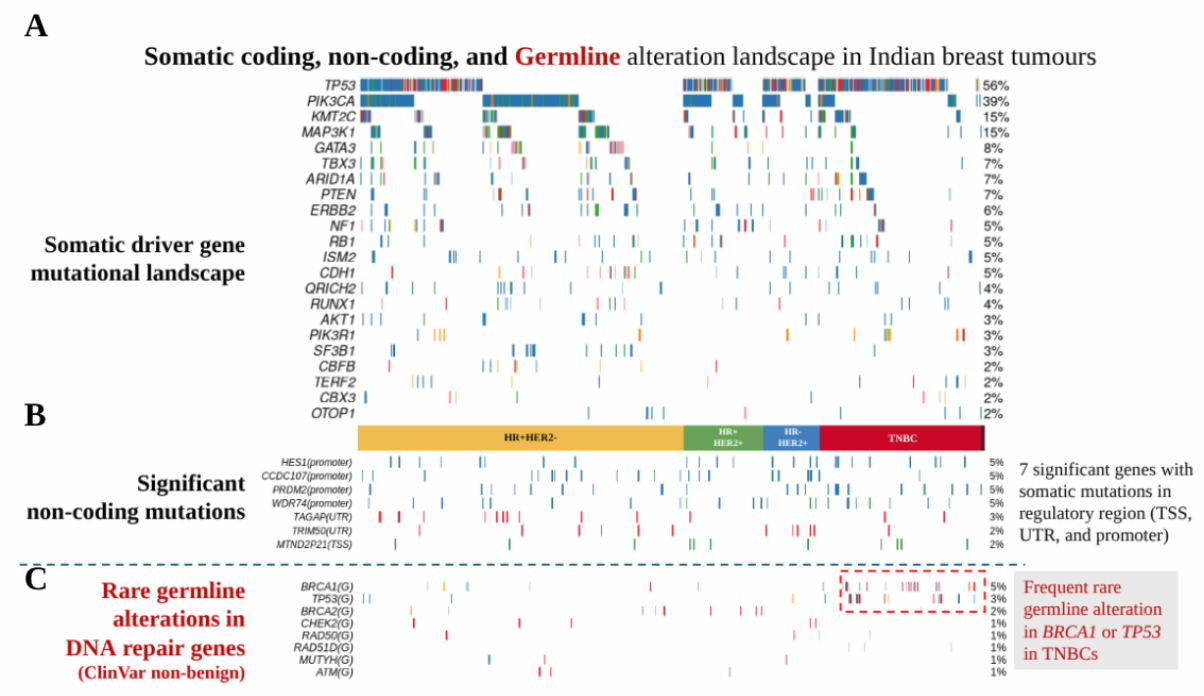


Supplementary Figure S7: Somatic mutational landscape of 45 HR-HER2+ patients. A median of 5690 somatic mutations ranging between 1241 and 114214. Somatic alteration landscape of 22 breast cancer driver genes in this immunohistochemistry subtype is illustrated.



Supplementary Figure S8: Somatic mutational landscape of 128 TNBC patients. A median of 6812 somatic mutations ranging between 256 and 114814. Somatic alteration landscape of 22 breast cancer driver genes in this immunohistochemistry subtype is illustrated.

The screening of rare pathogenic or likely pathogenic germline variants in DNA repair genes across the cohort (N=501) identified *BRCA1*as the frequently altered gene (25/501, 5%), with enrichment in TNBC (18/128,14.1%) relative to other subtypes. This was followed by TP53 (3%) and BRCA2(2%). Lower frequency alterations are also observed in *CHEK2*, *RAD50*, *RAD51D*, *MUTYH*, and *ATM* (Supplementary Table S3; Supplementary Figure S9).



Supplementary Figure S9: (A-B) Somatic ((A) coding significantly mutated driver genes, (B) significantly altered noncoding mutations), and (C) germline (rare germline pathogenic or likely pathogenic alterations in previously known breast cancer associated genes altered over 1% patients).

Several known hotspot mutations were detected at notable frequencies, including H1047R/L/Y (17.17%), E545K/A (7.98%), and E542K (4.59%) in PIK3CA, and R175H/G (2.59%) and R273H/C/L/P (4.39%) in TP53. No *ESR1* hotspots were identified in these primary tumours, which are often associated with poor outcomes. The PIK3CA E545K mutation was predominantly found in tumours lacking a *TP53* mutation. We also report that 63.26% of *GATA3* mutations were truncating in nature. We compared somatic SNVs found to be present within the 22 significantly mutated genes of the IBCGA cohort against 35 global breast cancer cohorts totalling 17,708 samples (Supplementary Table S4a, S4b). We observed that frequently mutated genes, such as TP53 and PIK3CA, exhibited a high proportion of known mutations (81% and 82% respectively) in comparison to global cohorts. In contrast, other genes, including *RB1*, *PIK3R1*, *NF1*, and *KMT2C*, consisted entirely of unique mutations. Further analysis identified significant ($q < 0.2$) enrichment of somatic mutations in the regulatory regions of nine genes, including the promoters of *CCDC107* and *HES1*, which were present in 2-5% of patients.

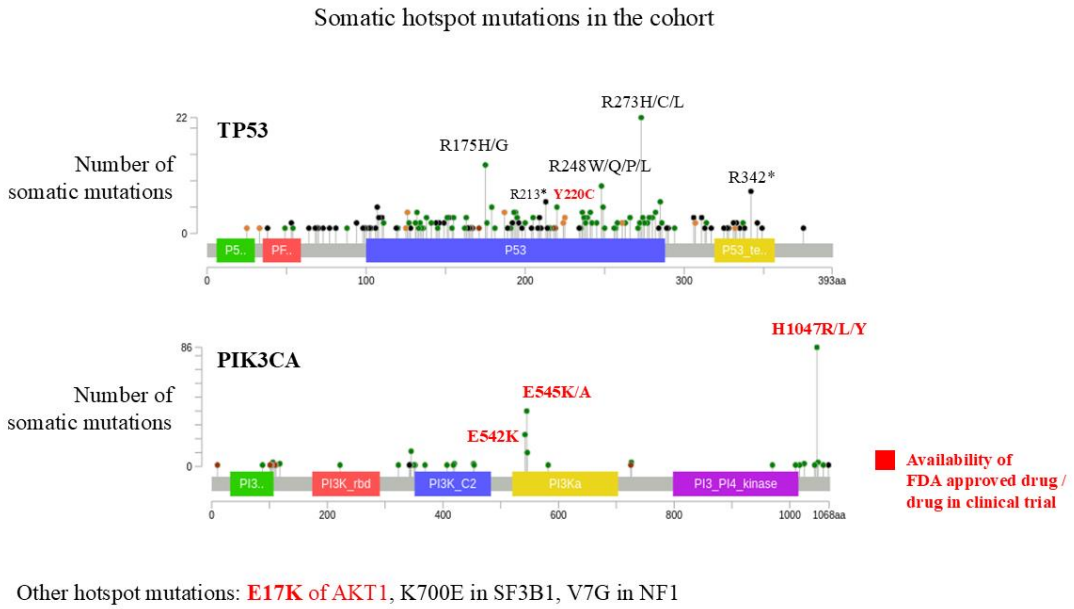
Germline alteration and somatic non-coding mutational landscape

We detected rare germline mutations (excluding "benign" reported in ClinVar database) in breast cancer associated DNA repair genes - BRCA1 (in 5% patients), TP53 (3%), BRCA2 (2%) to be present in ≥ 2% patients and *CHEK2*, *RAD50*, *RAD51D*, *MUTYH*, and *ATM* to be present between 1-2% Indian breast cancer patients. While rare germline mutations in *BRCA2* and *TP53*, were significantly ($p < 0.05$, Fisher's exact test) enriched in TNBC tumours, rare germline *BRCA2* mutations were predominantly detected in HER2+ tumours (irrespective of HR status). We found only 2.5% of patients with reported family history of cancer harboured pathogenic rare germline alterations in *BRCA1*, *BRCA2*, or *TP53* genes and 0.8% of the patients had rare germline alteration in *CHECK2*, *RAD51D*, *MUTYH* or *ATM* genes (Supplementary Figure S9)

We found significant ($q < 0.2$) enrichment of somatic mutations in regulatory regions of 9 genes - in promoters of *CCDC107*, *EIF2S3L*, *PRDM2*, *HES1*, in UTRs of *TAGAP*, *TRIM50*, both promoter and UTR (3' or 5' untranslated region) of *WDR74*, and TSS (transcription start site) of *RMRP*, *MTND2P21*. These mutations were present in 2-5% of the patients in this cohort. 85.71% promoter mutations in *EIF2S3L* were detected in the HR+/HER2- tumours.

Assessment of pathogenicity of somatic driver gene mutations and mutational hotspots

All missense mutations in *PIK3CA, TP53, RB1, NF1, AKT1, SF3B1, CBFB, PTEN, PIK3R1*, and 96.5% of *ARID1A*, *ERBB2* and 85.7% of *CDH1* gene were predicted to be pathogenic by the machine learning based framework (Supplementary Table S5) on the other hand less proportion of missense mutations in *RB1* (62.5%), *GATA3* (47.3%), *MAP3K1* (29.2%), *KMT2C* (23.0%), and *TBX3* (20.5%) were classified to be deleterious. We found all PIK3CA missense mutations including - D1017H, G106R, G106V, V344G, N1044Y are found to be pathogenic, along with known oncogenic mutations like H1047R, E542K, E545K, N345S etc. Likewise, all TP53 missense mutations that include known oncogenic mutations like R175H, R273H etc. were predicted to be pathogenic (Supplementary Figure S10).



Other hotspot mutations: **E17K** of AKT1, K700E in SF3B1, V7G in NF1

Supplementary Figure S10: Prevalence of somatic hotspot mutations in Indian breast cancer cohort. FDA approved drugs are available for multiple of these hotspot mutations in PIK3CA, and Y220C of TP53, and E17K of AKT1 genes

Somatic hotspot mutations *H1047R/L/Y, E545K/A*, and *E542K* in *PIK3CA*, and *R175H/G, R273H/C/L/P, R342**/frameshift, *R213*, Y220C, R248W/Q/P/L, p.R282W/G in TP53*, *E17K* in *AKT1* genes were detected in the Indian breast cancer cohort (Supplementary Table S6). Unlike *H1047R, E545K* hotspot mutation of *PIK3CA* was predominantly detected in tumours in absence of mutations in TP53. Overall, the *E545K* hotspot of *PIK3CA* was less frequent among both *HER2+* and TNBC tumours. We report frequent S133R mutations in MAP3K1 which was exclusively present in HER2-negative tumours irrespective of HR positivity status. Frequent N345K mutation in PIK3CA, K700E in SF3B1, and V7G in NF1 genes were exclusive to HR+HER2- tumours. The NF1 hotspot mutation was found predominantly in tumours with both *TP53*, and *PIK3CA* mutated. Majority of *GATA3* mutations detected in this cohort (predominantly found in HR+HER2- tumours) were truncating (63.26%) in nature (frameshift insertion/deletion, splice-site, and start codon). We have noted frequent multiple *MAP3K1* mutations in the same tumours.

Functional role of candidate driver mutations in IBCGA

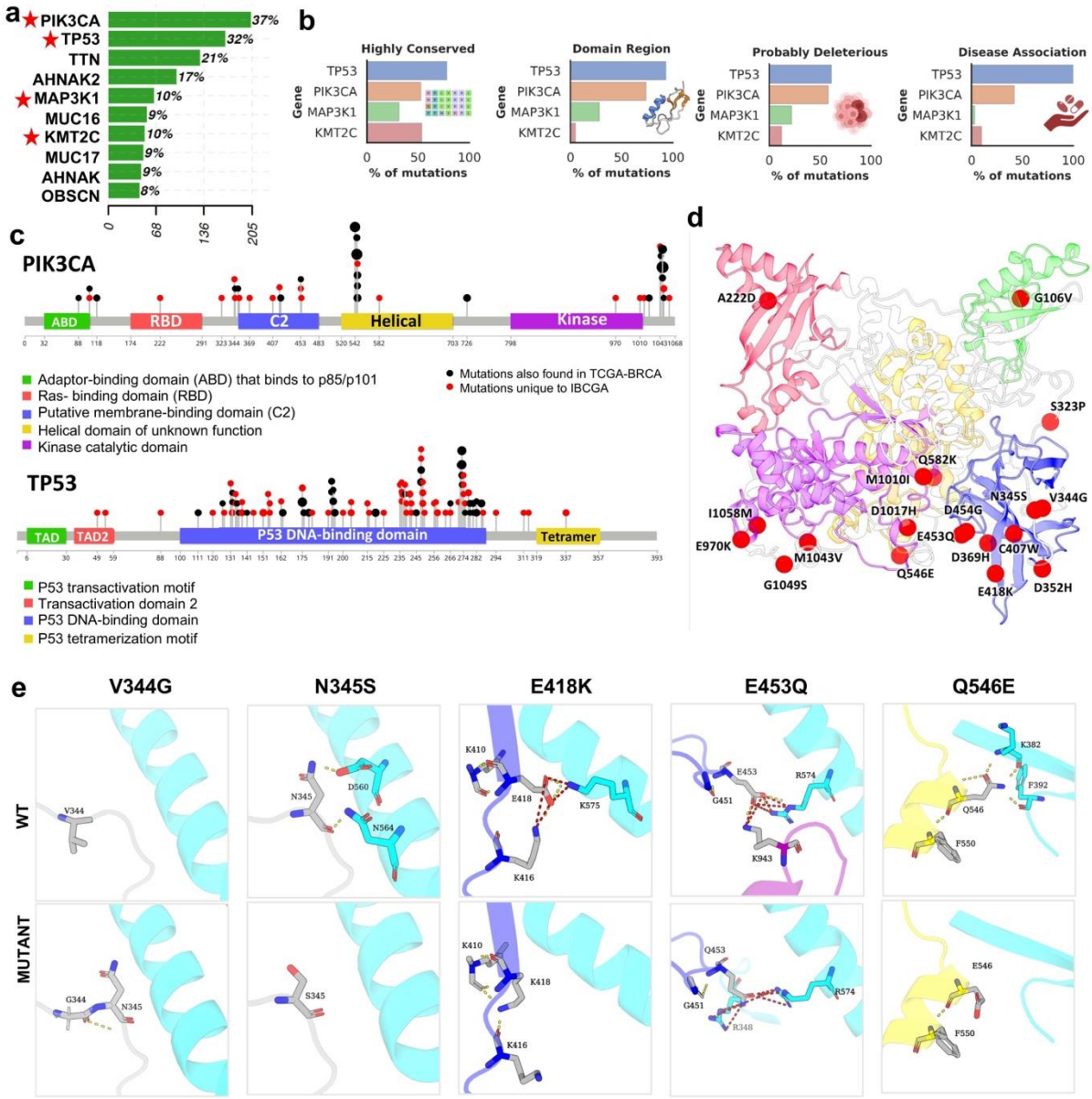
We identified recurrent and population-enriched somatic alterations, including four candidate driver genes, ISM2, QRICH2, CBX3 and OTOP1. Although biological functions are heterogeneous their presentation at the population-scale presents opportunity for investigating additional layers of tumour biology. However, functional studies will be required to establish their contributions to breast tumour development. ISM2 mutations may plausibly perturb the PI3K–lipid interface, by analogy with its

paralogue ISM1 — which suppresses hepatic lipogenesis via receptor-independent activation of PI3K–AKT signalling[22]; however, ISM2 itself has been characterised only as a prognostic biomarker linked to immune infiltration in colorectal cancer, with its underlying mechanisms not yet defined[23]. CBX3/HP1γ overexpression silences tumour suppressors and shapes immune exclusion across cancers [24,25]. QRICH2 is known for its role in ubiquitin -mediated degradation of specific substrate proteins[26], so by extension we speculate that an analogous proteostatic mechanism may operate on oncoproteins in tumour cells, potentially sustaining elevated ERBB2 and lipogenic signalling. And OTOP1 was shown to connect intracellular proton homeostasis[27,28] and has been shown to protect against adipose tissue wasting during cancer cachexia[29].

Protein structure-mapped mutations reveal *PIK3CA* to harbour unique mutations at previously unknown functional surfaces

For structure-based understanding of mutations, genomic data was integrated with high-confidence protein models. We focussed on missense mutations from 501 patients with annotations of mutation positions within the corresponding transcript: This revealed mutations in 13,568 genes at 38,125 unique positions. For further analysis, we focused on mutations in the top four genes (*PIK3CA, TP53, MAP3K1,* and *KMT2C*), and compared mutations reported in the TCGA-BRCA dataset to identify mutations that are unique to IBCGA (that is, not present in TCGA-BRCA). Among these mutations (n=151), 58.27% occurred at highly conserved residues, with 52.98% within functional domains, 39.07% predicted to be deleterious, and 46.35% had known disease associations (Supplementary Figure S11). *TP53* and *PIK3CA* exhibited the highest proportion of these features. *PIK3CA*, revealed 19 unique mutations with similar frequency, amongst which 52.63% are at highly conserved residues. These mutations are also located on the protein functional domains: like p85-binding domain (G106V), ras-binding domain (A222D), Phosphoinositide 3-kinase C2 (D352H, D369H, C407W, E418K, E453Q), Phosphoinositide 3-kinase family, accessory domain (Q546E, Q582K) and Phosphatidylinositol 3- and 4-kinase domain (M1010I, E970K) that forms the catalytic domain of the protein. Six variants (G1049S, M1043V, C407W, V344G, A222D and G106V) were additionally predicted to be destabilising, which may affect folding or function. Around 40% of these mutations (G106V, V344G, E418K, E453Q, Q546E, E970K, M1043V, and G1049S) are also related to several cancer and non-cancerous diseases like CLOVES syndrome, Cowden syndrome, PIK3CA-related disorder, Neoplasm, etc. Apart from these 60% of them are predicted to be deleterious and pathogenic in nature (G106V, A222D, V344G, N345S, D352H, D369H, E418K, Q582K, D1017H, G1049S). These mutating residues are present on the different protein-protein interaction interfaces. Ten mutants (D352, C407, E453, V344, N345, D369, E418, D454, Q546, D1017) are at the interface of three PIK3CA interacting partners (PIK3R1, PIK3R2, PIK3R3), four of them (Q582, E970, M1010, G1049) at the interface of two partners (PIK3R1, PIK3R2) and one mutant, I1058M is at interaction interface of PIK3R3 protein in modelled complexes for which pDockQ scores supported a plausible interaction; pDockQ reflects overall model confidence

and does not by itself validate individual interface residues. We further used the crystal structure of the complex (PDB id: 5XGH)[15] and modelled mutated complex, which were further docked using HADDOCK[16] and 100+ structures were generated. These mutated complexes were analysed in comparison to the wildtype structure and the changes in interaction patterns. Molecular docking of PIK3CA and PIK3R1 showed the effect of mutations at the interface. Docking of PIK3CA with PIK3R1 (Methods) was used to examine the predicted effect of these mutations at the interface. Complexes showed that critical salt bridges, hydrogen bonds were lost in the case of some mutants (N345S, E418K, E453Q, Q546E), as compared to the wild type. Notably, the wild-type N345 contacts recovered in our models (D560, N564 of PIK3R1) correspond to those observed crystallographically[30], supporting the modelling approach. These predictions indicate that V344G, N345S, E418K and Q546E perturb interface contacts, with V344G and E418K showing the largest predicted loss. Release of p85-mediated inhibition is an established activating mechanism for the canonical PIK3CA helical- and C2-domain hotspots [30], and the position of these variants is consistent with a comparable mode of action; we note that N345K at one of these residues is itself a recurrent hotspot, so novelty here is at the level of the substitution rather than the residue. As these are computational predictions, experimental validation of the effect on PI3K–p85 regulation will be required. Lastly, analysis of recurrent mutations revealed high-frequency hits in TP53, MAP3K1, and KMT2C. TP53 variants clustered in the DNA-binding domain (K132R, A138V, P151S, R175G, Y205H, M237I, S241F, R249S/W, E258K, C277F) and tetramerization motif (R337L), both highly conserved and disease-associated. In MAP3K1, S133R and F14V occurred in 13 and 9 samples, respectively; all MAP3K1 variants (F14V, K32T, V128G, K146Q, D132A, S133R, P257L) localised to unstructured loop regions; interpretation is limited, as MAP3K1 is predominantly inactivated by truncating events. KMT2C mutations (F48V, V125I, R886H, P1606L, Q3836K) were also confined to unstructured regions, with R886H (8 samples) predicted to be highly destabilising and pathogenic. Together, these observations indicate that population-enriched missense variants in this cohort are preferentially positioned at conserved, functionally constrained surfaces — most notably the PIK3CA–regulatory subunit interface — consistent with the evidence for positive selection at these loci presented in the main text (Figure 1C).



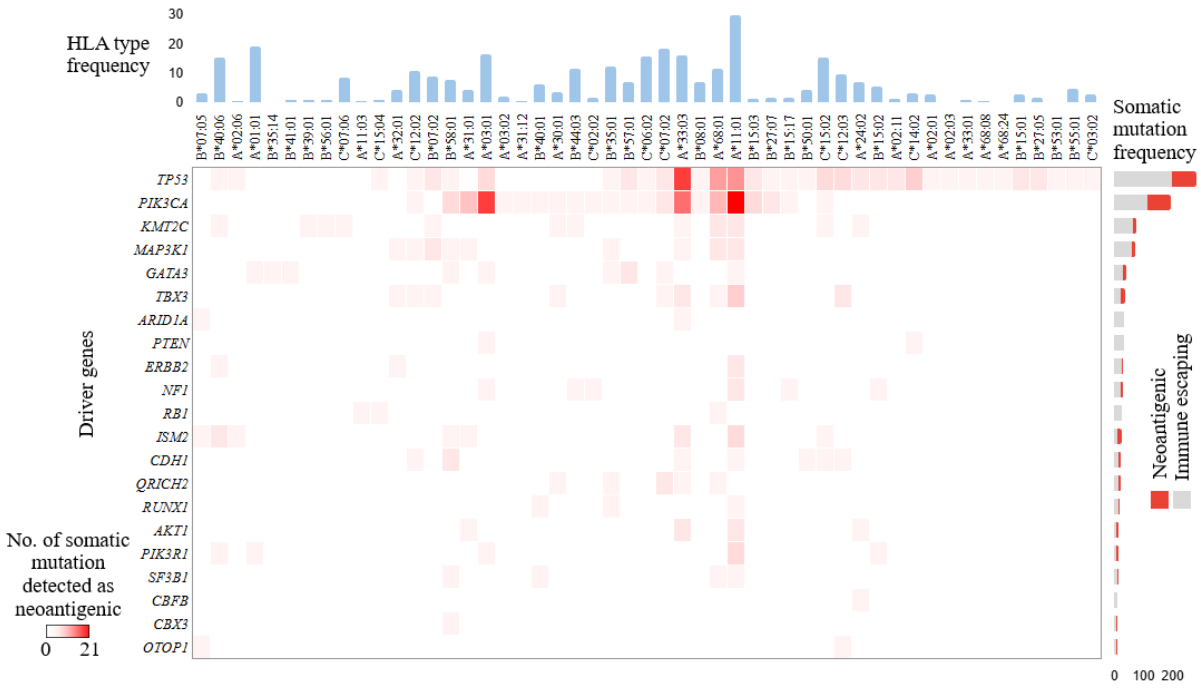
Supplementary Figure S11 (a) The horizontal bar graph shows the top 10 genes that harbour major missense mutations, amongst which *PIK3CA*, *TP53*, *MAP3K1* and *KMT2C* are established breast cancer driver genes. (b) The horizontal bar plots show the percentage of unique IBCGA mutations (that is, not present in TCGA-BRCA), localized at the highly conserved residue position, present in the functional domain region, predicted to be deleterious (CADD Phred-like score >25.0), and having disease associations in case of the top four genes. (c) The lollipop plot depicts the mutation on the PIK3CA and TP53 protein. The size of the circle represents the comparative frequency of the mutations. The mutations that present in IBCGA, and not reported in TCGA-BRCA, are shown in red. (d) Mutations from IBCGA, but not present in TCGA-BRCA, are highlighted and labelled on the PIK3CA protein in red. (e) Interaction of PIK3CA wildtype and mutants with the PIK3R1(cyan) protein. The residue interaction at the interface shows the loss of interaction in mutants (V344G, N345S, E418K, Q546E)

as compared to wild type. The dashes represent the hydrogen bond (yellow) and salt bridge interactions (red).

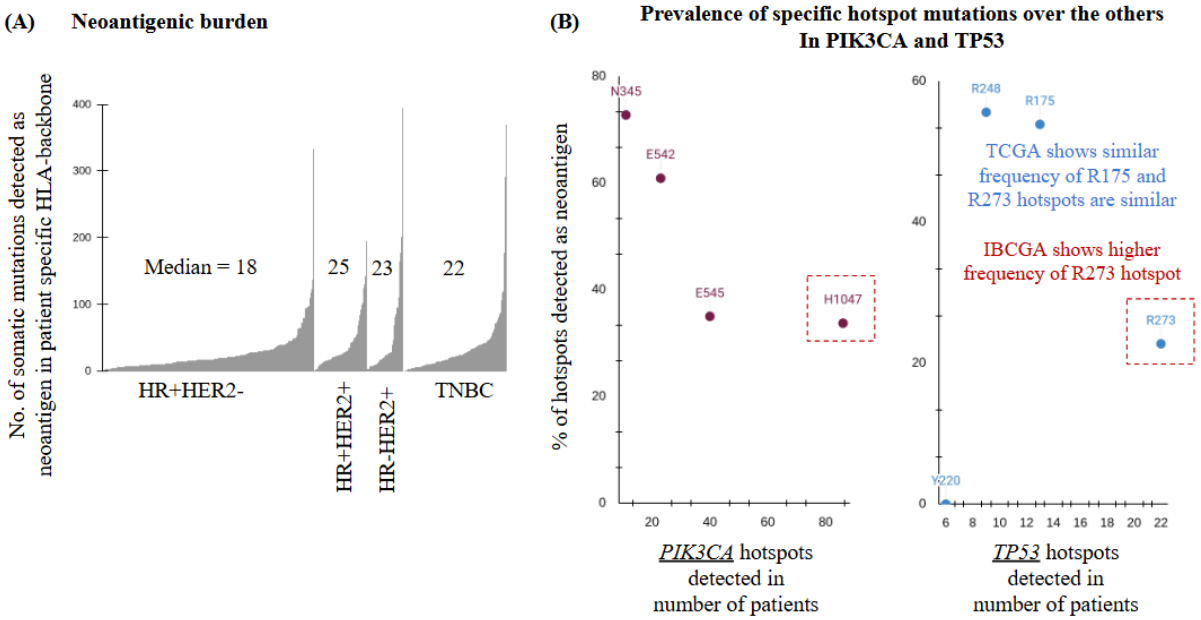
Prediction of neoantigens from somatic mutation data

Germline HLA class I alleles were inferred from normal-tissue WGS data using OptiType[31]. Quality-controlled, non-synonymous somatic variants were annotated using Ensembl Variant Effect Predictor[32] (VEP), incorporating all available GENCODE mRNA isoforms. For each variant, corresponding wild-type and mutant peptide sequences were generated and processed into 8–15 amino-acid peptides. Peptide–HLA class I binding affinities were predicted against patient-specific HLA alleles using the pVAC-Seq pipeline[33], with default parameters and filtering criteria. Peptides meeting the pVAC-Seq criteria were considered candidate neoantigens. The resulting neoantigen predictions were used to assess the presentation potential of recurrent somatic driver and hotspot mutations across the IBCGA cohort. (Supplementary Figure S12, S13).

.



Supplementary Figure S12: Number of somatic driver gene mutations detected as neo-antigenic on specific HLA germline backbone.



Supplementary Figure S13: Landscape of neoantigenic somatic mutations in breast tumours from Indian patients. (A) Distributions of neoantigenic somatic mutations (on patient specific MHC-I backbone), among breast cancer IHC subtypes. (B) Prevalence of specific hotspot mutations in PIK3CA and TP53 are shown to be correlated with immune escape capability. Immune escape capability is defined such that it has an inverse correlation of the proportion of mutations detected as neoantigen in the number of patients.

Quantitative analysis of copy number alterations across breast tumour subtypes

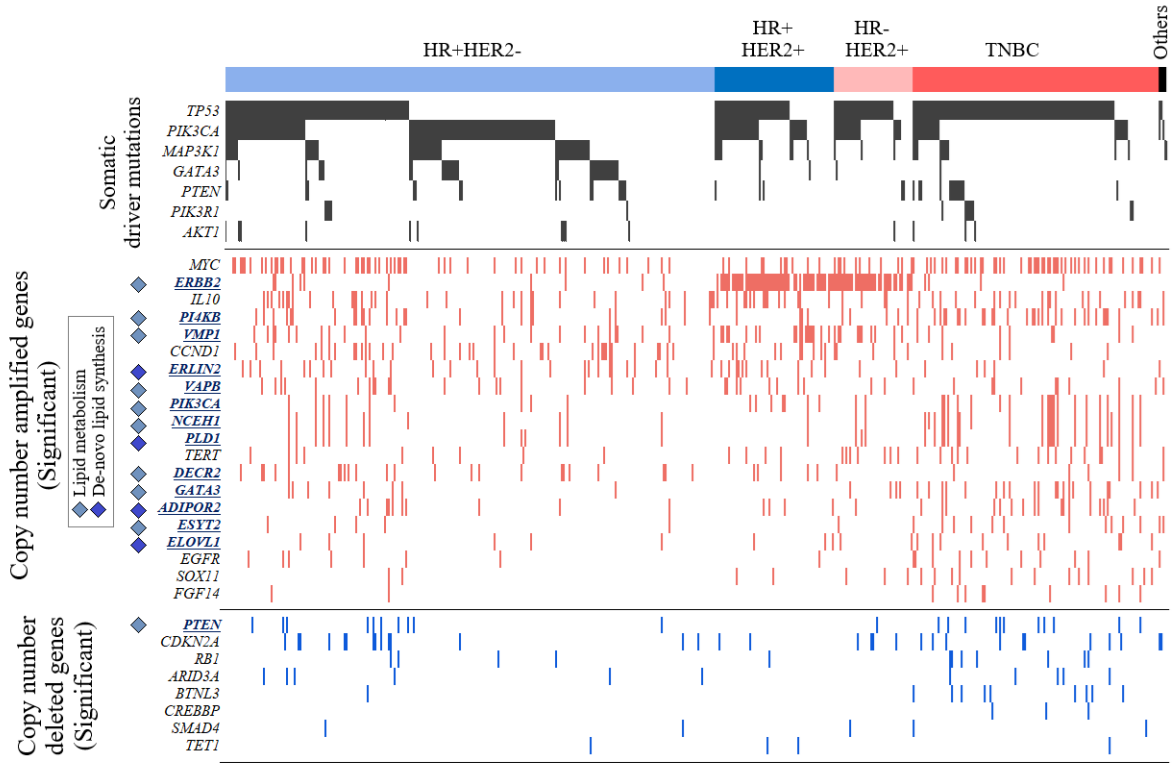
Genomic analysis identified six chromosome arms (1q, 7q, 8q, 16p, 20p, and 20q) that were significantly amplified and twenty-one arms that were significantly deleted ($q < 0.05$) across the 490 tumours (11 cases CNV could not be resolved unambiguously; Supplementary Table S7). Significant focal amplifications were quantified for key oncogenes, including *MYC* (25%), *ERBB2* (20%), *ECM1* (15.89%), *PI4KB* (14%), *CCND1* (12%), and *PIK3CA* (8%). Deletions were quantified for tumour suppressors such as *CDKN2A* (6%) and *PTEN* (4%). Subtype-specific patterns were evident, with *ERBB2* amplification occurring in 78.89% of HER2+ tumours. Although somatic mutations in *PIK3CA* were significantly less in TNBCs, the incidence of the copy number amplification of the gene was prevalent. The overall copy number alteration burden was significantly higher in TNBC tumours, with a median proportion of the genome altered of 42.51%, compared to 26.62% in other subgroups ($p = 1.582e{-}05$, t-test). This was driven by both significantly higher amplification ($p = 0.0003$) and deletion ($p = 4.757e{-}06$) events in TNBCs.

Comparison of somatic mutational and copy number alterations of IBCGA cohort with TCGA breast DCIS

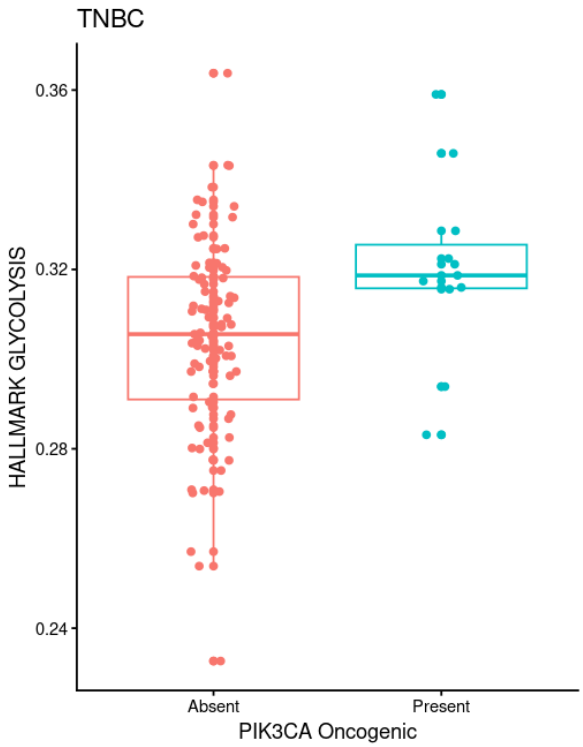
Somatic mutations in *TP53*, *TBX3*, and *RB1*, in both HR+HER2- and HER2+ (irrespective of HR status) are found to be altered in a significantly higher proportion of patients in the Indian cohort as compared to the similar tumours in TCGA cohort. Mutations in *GATA3* in HER2+ tumours, and *CDH1* in both HR+HER2- and HER2+ tumours were found to be significantly less in this cohort in comparison with the TCGA cohort. In contrast to the TCGA cohort, somatic mutations in *SF3B1* were not found among Indian HER2+ tumours. Frequency of somatic mutations of *TP53* (82%), *PIK3CA* (16%), *PTEN* (10%), *PIK3R1* (6%), *NF1* (6%), and *ERBB2* (3%) was found to be very similar ($p > 0.05$, Fisher's exact test) with the TNBC tumours of TCGA cohort [*TP53*: 83%, *PIK3CA*: 12%, *PTEN*: 7%, *PIK3R1*: 4%, *NF1*: 4%, and *ERBB2*: 3%]. In contrary, we found significantly ($p < 0.05$, Fisher's exact test) higher mutational frequency of *KMT2C* (17%), *ARID1A* (9%), and *MAP3K1* (8%) in TNBC in Indian cohort compared to the TCGA [*KMT2C*: 6%, *ARID1A*: 2%, and *MAP3K1*: 1%]. *AKT1* was found to be mutated in 2% of Indian TNBC patients, while it is not mutated in TNBC patients in TCGA.

Focal amplification of *MYC*, *PIK3CA*, *PI4KB*, *TERC*, *GATA3*, *CCND2*, *EHF*, *ELOVL1*, etc., were common among Indian TNBC tumours and TCGA cohort, in contrary, amplification of *AKR1C3* (in 20% patients of Indian TNBC cohort), *NCEH1* (19%), *PLD1* (17%), *IL10* (16%), *ADIPOR2* (13%), *TERT* (13%), deletion of *BTNL3* (7%) were exclusive to the Indian cohort compared to TCGA. We also noted a lower fraction of Indian TNBC tumours harbouring focal deletion of *PTEN* (19%), in comparison with the TCGA cohort. Amplifications of *ERBB2*, *MYC*, *CCND1*, *ERLIN2*, *PI4KB* etc. were found in HER2+ (irrespective of HR status) tumours from both Indian and the TCGA cohort, while amplifications of *VMP1*, *IL10*, *FGFR1*, *TERT*, *GATA3*, *AKR1C3*, *MDM2*, *ELOVL1*, *TERC*, and

IGF1R were almost exclusively found in the subtype of Indian patients. Interestingly, the frequency of *MYC* amplification was significantly ($p < 0.05$, Fisher's exact test) higher among Indian HR+HER2- tumours (23%) as compared to TCGA (9%). Alongside, amplification of *VMP1*, *FGFR1*, *DECR2*, *BIRC5*, *EHF*, *NCEH1*, *MDM2*, *AKR1C3*, *ADIPOR2*, *LSS*, *TERT* etc. were almost exclusive to the subtype among Indian patients compared to TCGA (Supplementary Table S9, S10).



Supplementary Figure S14: Significantly amplified or deleted genes in breast cancer cohort shown alongside frequently mutated genes.



Supplementary Figure S15: ssGSEA scores across hallmark pathways in TNBC samples with PIK3CA oncogenic mutations.

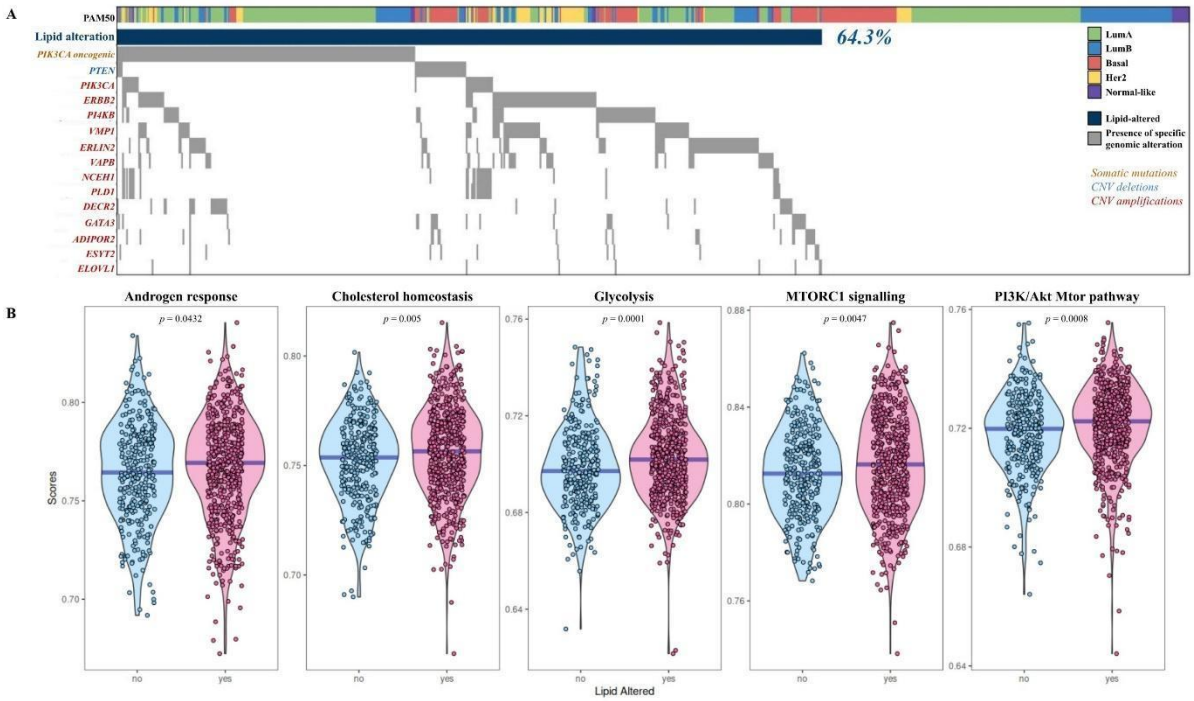
Genomic correlates of distant metastasis

In the 36 patients (out of 501, 7%) who developed distant metastasis, primary tumours showed a significantly higher frequency of copy number amplifications in *MYC* (47% vs. 23% in non-metastatic) and *ERBB2* (42% vs. 18%). Somatic mutations in *CDH1* were also significantly higher in tumours that developed metastasis (14%) compared to those that did not (4%). Furthermore, the TP53 hotspot mutation R175H was significantly enriched among metastatic tumours (8%) compared to non-metastatic cases (2%) ($p < 0.05$, Fisher's exact test) (Supplementary Table S11).

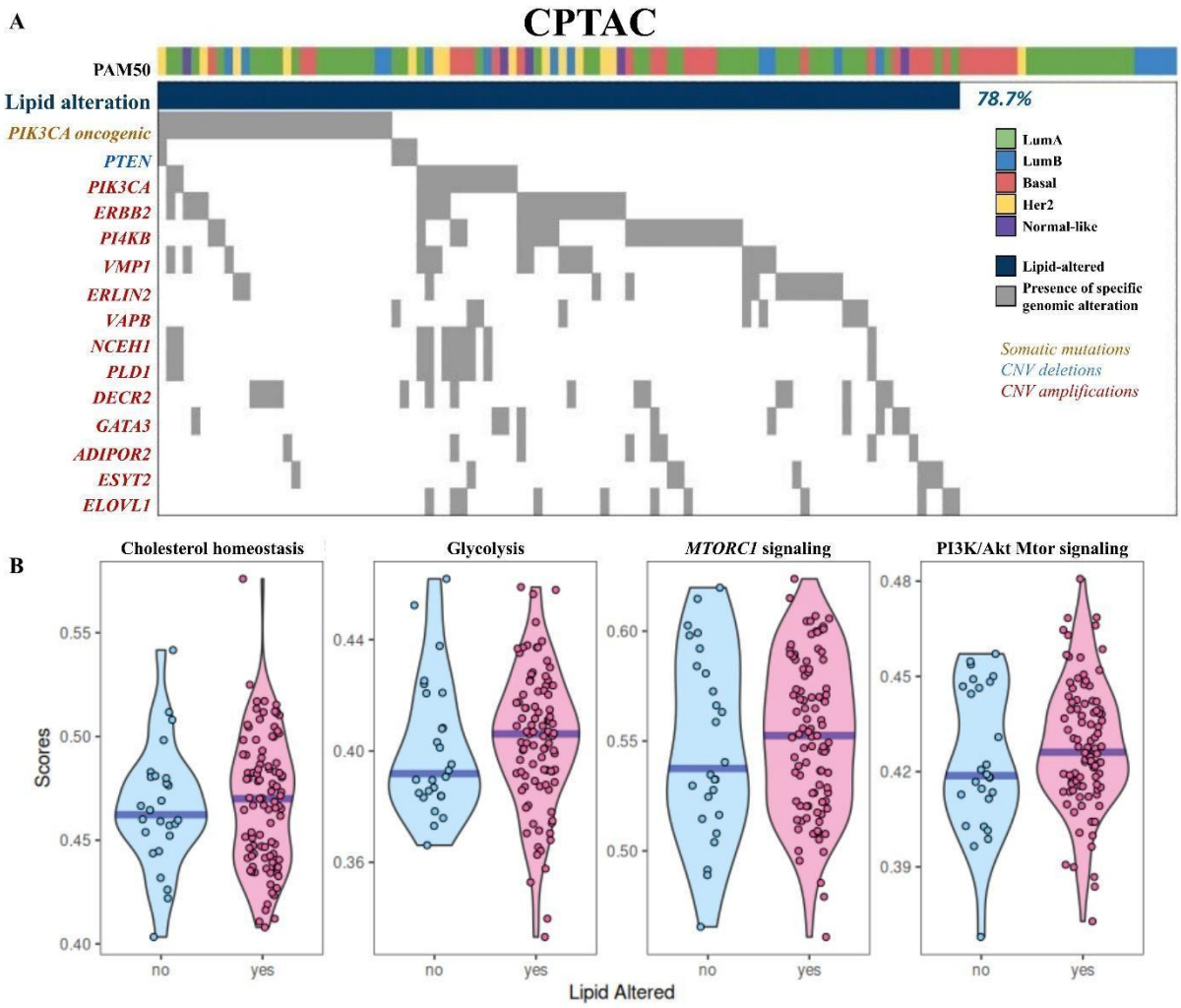
Cross-Cohort Validation of the Lipid-Alteration Signature

We further validated the lipid-alteration genomic signature in the TCGA-BRCA cohort, where 64.3% (551 of 857) of the TCGA patients were identified to harbour any one of the lipid-alteration features (Supplementary Figure S16A). We performed ssGSEA expression analysis of these two groups (lipid altered vs unaltered in the genomic level) of TCGA samples for the hallmark pathways. We identified lipid-related hallmarks to be significantly higher ($p < 0.05$, one-tailed wilcox) in TCGA tumours harboring lipid-signatures (Supplementary Figure S16B): 1) Androgen response, 2) Cholesterol homeostasis, 3) PI3K-Akt-Mtor, 4) Mtorc1 signalling, 5) glycolysis. In addition, EMT, hypoxia, TGF-beta, Notch, etc. were also higher in these patients (not shown). Therefore, this provides validation of our lipid alteration signature in an external cohort and further strengthens our study.

Next, we tested the lipid alteration features in the CPTAC dataset of 122 breast cancers, where 78.7% of the patients harboured the lipid-related genomic alterations (Supplementary Figure S17A). We further performed ssGSEA for these samples: four out of five lipid-related hallmark pathways (cholesterol homeostasis, glycolysis, Mtorc1 signalling, and PI3K/Akt Mtor signalling) which were identified to be highly expressed in lipid-altered patients both in IBCGA and TCGA. In the CPTAC cohort, patients harbouring lipid-related alterations also had higher expression of these pathways, although directionally consistent these were not statistically significant, likely due to the low number of patients data available from CPTAC (Supplementary Figure S17B).



Supplementary Figure S16: Validation of the Lipid-Alteration Genomic Signature Using the TCGA-BRCA Cohort. **A.** Lipid-alteration features **B.** Lipid related hallmark pathways



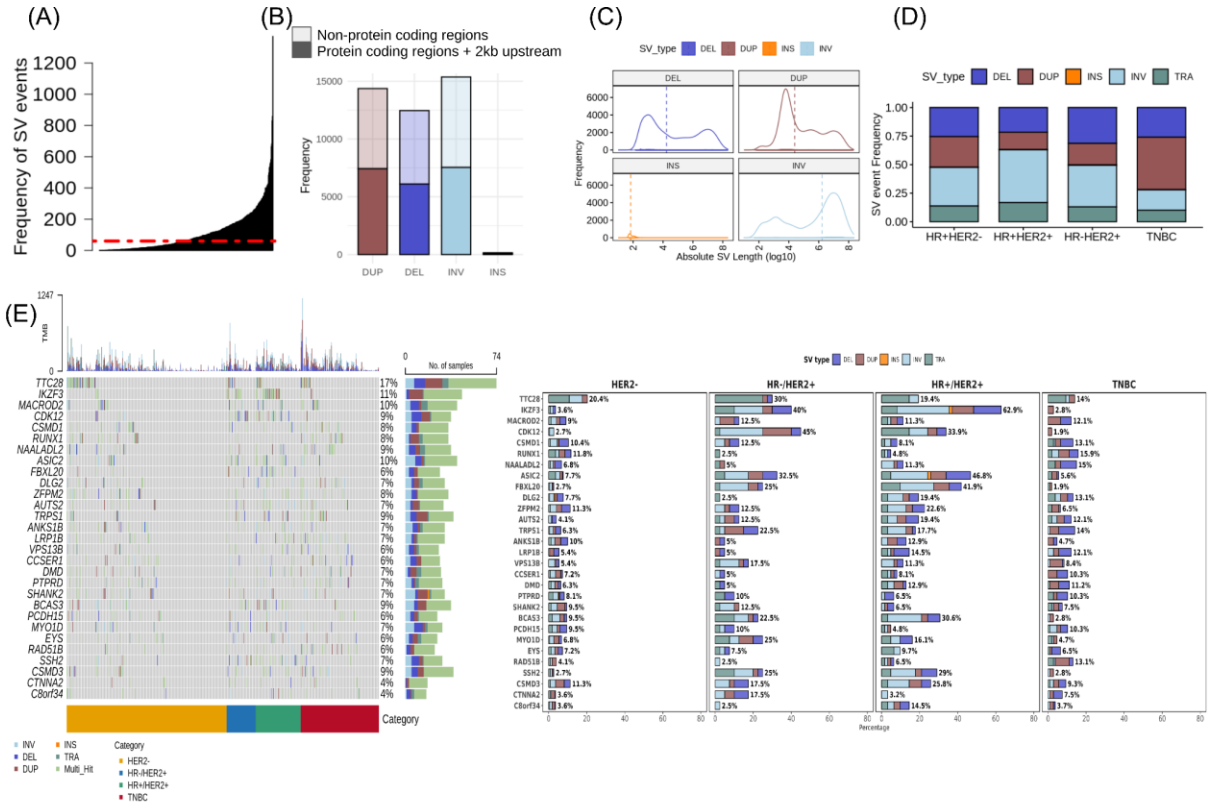
Supplementary Figure S17: External Validation of the Lipid-Alteration Genomic Signature in CPTAC Cohort A. Lipid-related genomic alterations B. Lipid related hallmark pathways

Discussion II

Somatic structural variation landscape in breast cancer cohort

We identified 90,553 somatic structural variations (SVs) encompassing 439 patients in our cohort, of which 59,078 (65.2%) represented local intrachromosomal breaks (<5 Mb), although intrachromosomal (20.2%; 18,331) and interchromosomal (14.5%, 13,144) genomic reshuffling events (>5 Mb) were also noted. Although, HR+HER2+ tumours had the highest SV burden (median: 214 SVs per patient), followed by TNBC (157 SVs), HR-HER2+ (131 SVs), and HR+HER2- (69 SVs) tumours, we noted subtype-specific variations in the type of SVs as, TNBC tumours had the highest burden of duplication events (median: 37 per patient), whereas, HR+HER2+ had highest burden of deletion (50 SVs), inversions (94 SVs), and translocations (32 SVs). We identified pathogenic deletion and duplication events (ACMG classification), which showed high tumour allele fractions (TAF $\geq$ 10%). HER2+

tumours showed a higher burden of pathogenic events (median: 74 per patient), followed by TNBC (38 pathogenic SVs) and HR+HER2- (29 pathogenic SVs), indicating possible clinical actionability. We integrated the large-scale copy-number patterns (G-scores) along with the SV calls to construct a genomic interaction map between 5Mb genomic segments. We noted enrichment of super-enhancers in breast cancer and normal breast epithelium, in the regions flanking 1Mb of ERBB2 (median 76 kb upstream and 143 kb downstream), which might disrupt enhancer-promoter interactions, TAD boundaries, and chromatin organisation. Recurrent SVs included genes, like, *ASIC2* (22% of HER2+ patients), *MACROD2* (9.5% of patients), *RAD51B* (12% of TNBC patients), as well as, several chromatin remodelers (total 466 SVs across 159 patients), like, *NF1* (23% of patients), *ARID1B* (4% of patients), *ESR1* (3.8% of patients), *RB1*, *BRD4*, etc (Supplementary Figure S18A-E).

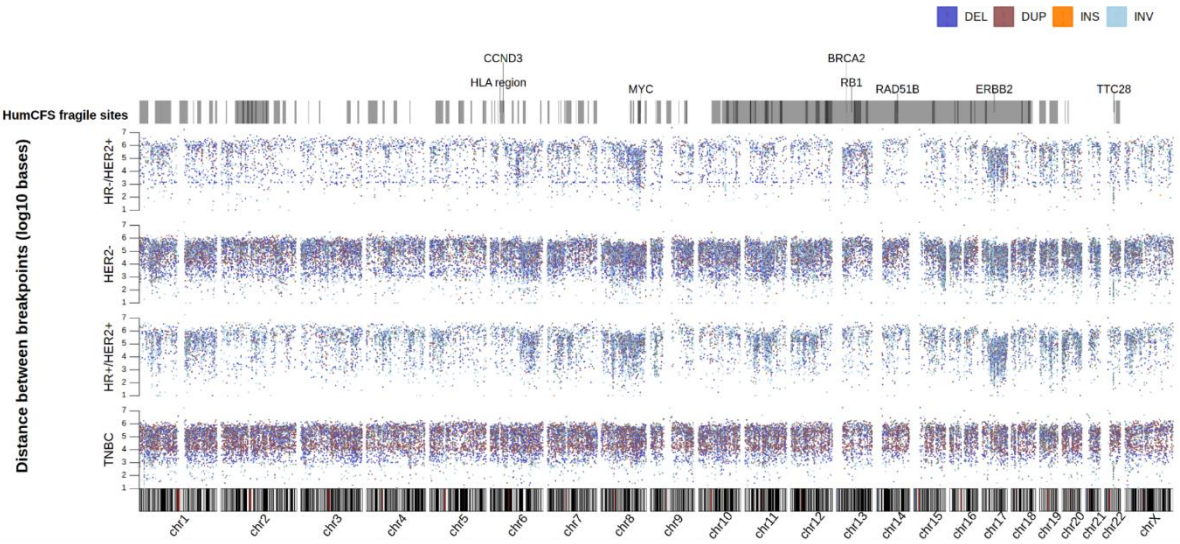


Supplementary Figure S18: Landscape of structural variants (SVs) across breast cancer subtypes.(A) Distribution of SV event frequency across all samples, highlighting their overall occurrence.(B) Illustrates the distribution of SV overlap with 2kb upstream of protein coding genes and non-coding regions with respect to SV types, the distribution for [Duplication (DUP), Deletion (DEL), Inversion (INV), Insertion (INS)],the distribution Other regions: DEL (6355), DUP (6939), INS (92), INV (7829), 2kb upstream of protein coding region: DEL (6108), DUP (7428), INS (64), INV (7539).(C) Length distributions for major SV classes (DEL, DUP, INS, INV) Log-scaled density plots depict the distribution of absolute SV lengths.(D) Shows the proportion of each SV type across cancer subtypes (HER2-, HR+/HER2-, HR+/HER2+, TNBC). This illustrates how the SV frequencies vary for each type like HER2+ and TNBC show enrichment for specific SV types. (E) Left:Oncoplot depicting SV

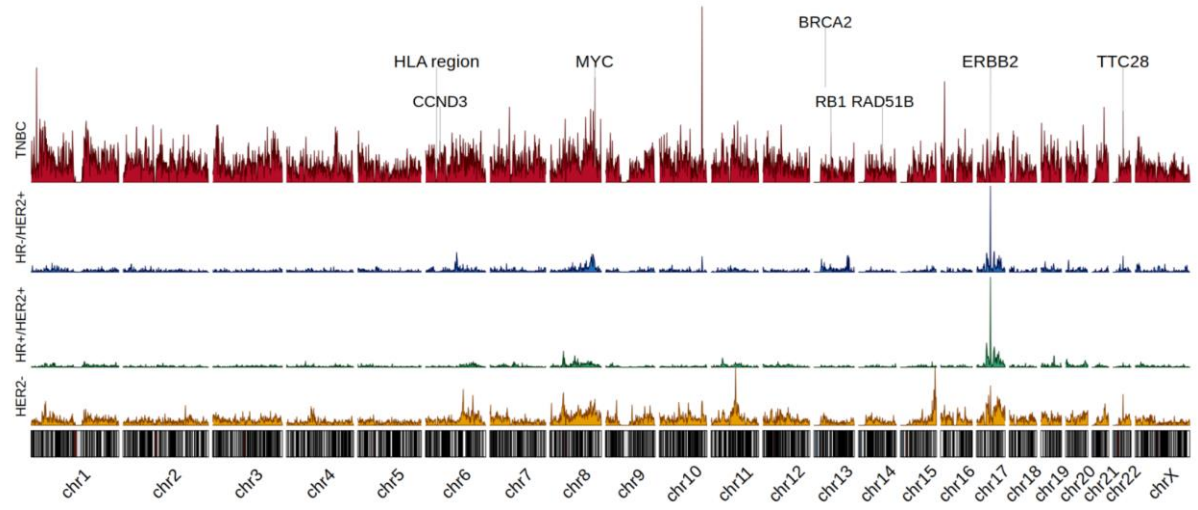
events for 30 recurrently altered genes, stratified by breast cancer subtype and SV type. Top bars represent tumour mutational burden (TMB). Right: Barplots of subtype-specific frequency of SVs in each gene, coloured by type.

A large fraction of SV breakpoints (50-53%) were enriched in protein-coding genes and their 2-kb upstream regions. SV breakpoints showed a preferential overlap with known fragile sites annotated in the HumCFS database (Z-score = 19.196, $p < 0.001$)[34] (Supplementary Figure S19) Genes proximal to fragile sites and frequently disrupted include *TTC28*, *IKZF3*, *MACROD2*, *ASIC2*, and *CDK12*, among others. Comparison with PCAWG confirmed recurrent alterations in these and other genes, reflecting both shared and cohort-specific SV patterns (Supplementary Figure S20).

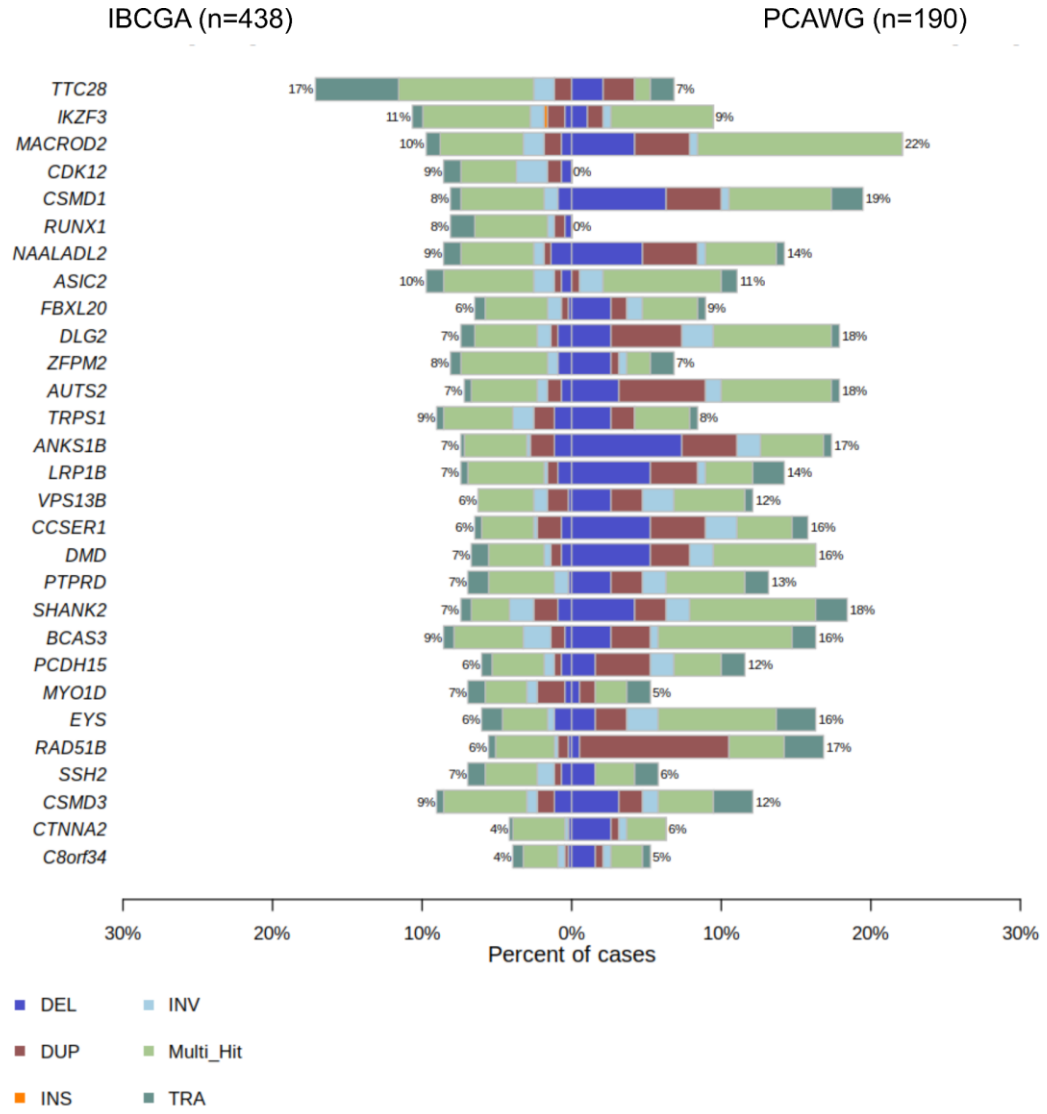
(A)



(B)



Supplementary Figure S19: (A) Genome-wide distribution of structural variant breakpoints across subtypes. (B) Density tracks stratified by SV type highlight subtype-specific enrichment of breakpoints across chromosomes and subtypes. Top vertical track represents annotated common fragile sites.



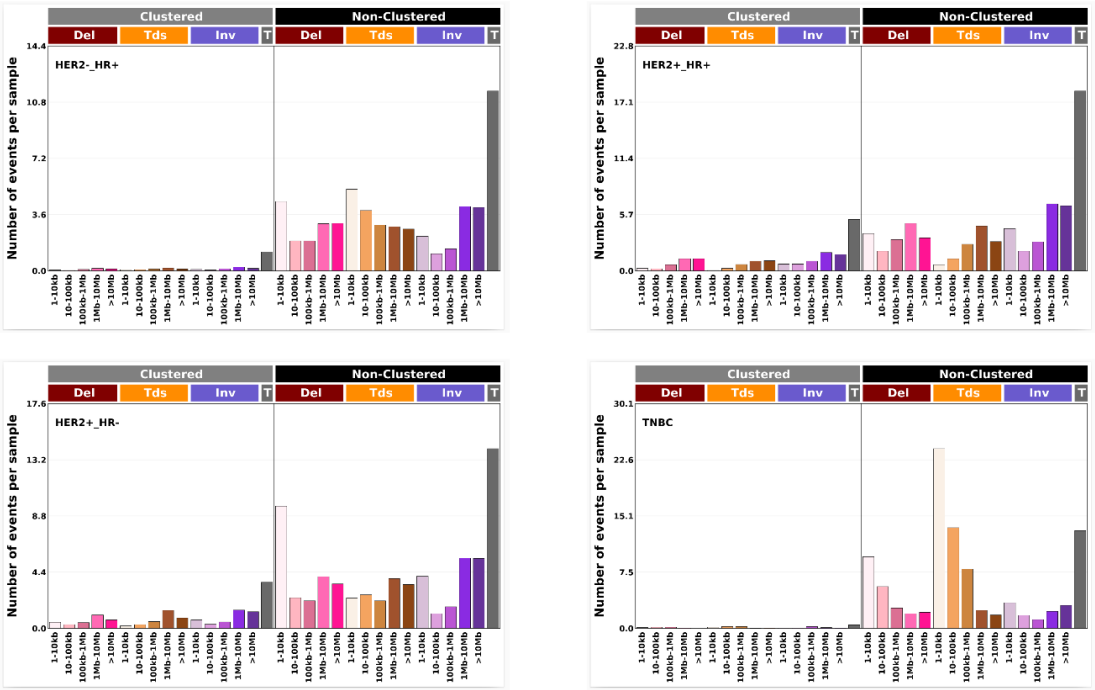
Supplementary Figure S20: This co-barplot compares the distribution and recurrence of structural variants (SVs) in breast cancer samples from the IBCGA and PCAWG cohorts for the 30 recurrently altered genes by SVs, highlighting shared and cohort-specific patterns.

We further observed tumour subtype-specific SV patterns: more duplications in TNBC and inversions in HR+/HER2+ subtypes, with deletions common across all subtypes (Supplementary Figure S21). COSMIC mutational signatures analysis also showed differential distribution of non-clustered SVs across subtypes: TNBC had more small deletions/duplications, HR+/HER2+ had large inversions/translocations. The signature SV3, known to be associated with homologous recombination

deficiency, was enriched in TNBC samples, while the SV2 signature of unknown etiology was prevalent in HER2+ samples[5]. We found that the top 25% of samples with high cosine similarity to the SV3 signature showed significantly better survival probability within the TNBC cohort (log-rank $p < 0.05$). (Supplementary Figure S22). Significant differences were noted in the structural variation landscape of HRD-high versus HRD-low patients ($p < 0.001$, Wilcoxon Rank Sum Test).

Supplementary Figure S21: The karyoplots represents distribution of structural variant breakpoints along each chromosome for IHC subtypes - (A) HR-HER2+, (B) HR+HER2+, (C) HR+HER2-, and (D) TNBCs.

(A)

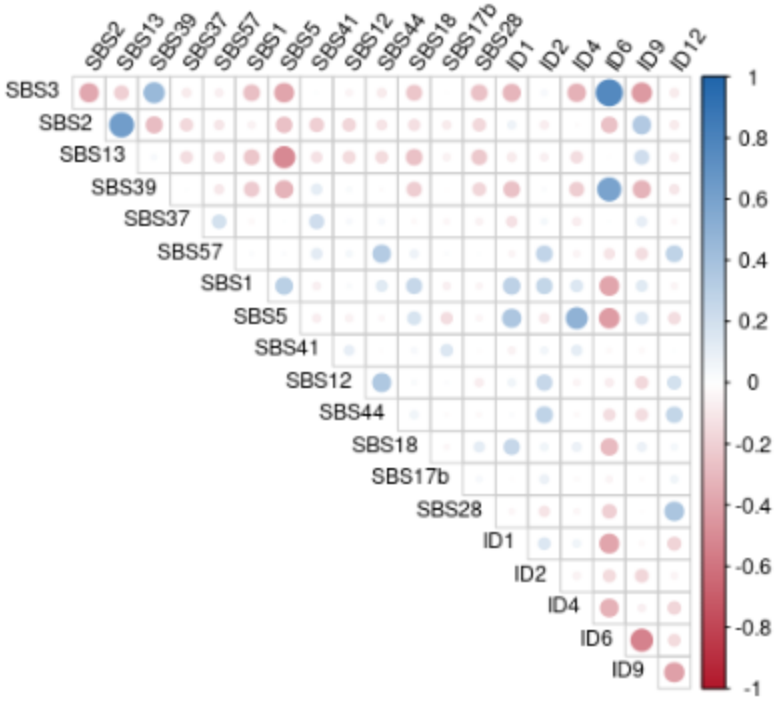


Supplementary Figure S22: (A) Bar plots represent the SV events per subtype categorised as per COSMIC SV32 classification schema.

Signatures of genomic instability in Indian breast tumours

Of the 16 somatic mutational signatures identified, SBS3, which is linked to DNA double strand break repair (DSBR) failure / homologous recombination deficiency (HRD), had a median contribution of 14.78% across the cohort (Supplementary Table S8, Supplementary Figure S23)

The contribution of SBS3 was significantly higher in TNBCs (median contribution of 28.33%) compared to hormone-positive tumours (11.74%) ($p < 5.043e{-}14$, *t*-test), and its prevalence was significantly associated with *TP53* mutational status ($p = 1.494e{-}10$, *t*-test).

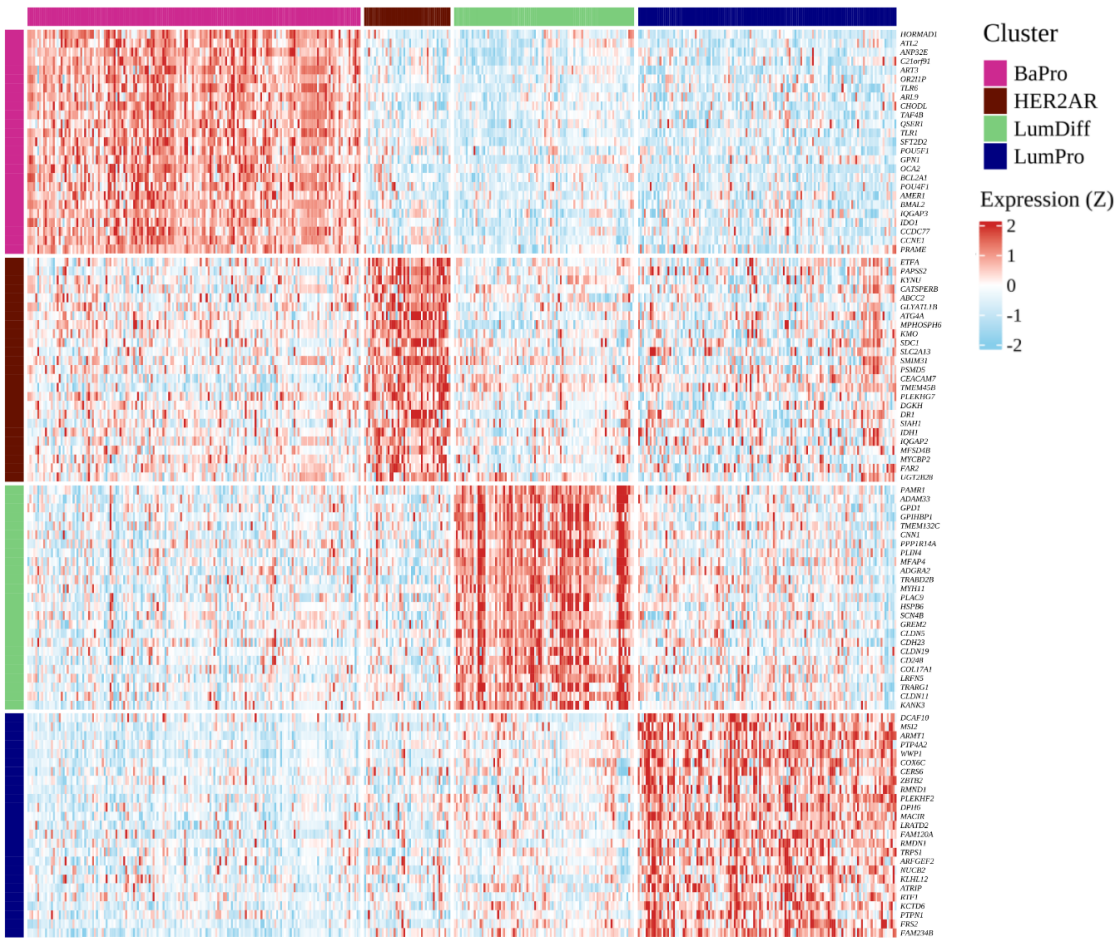


Supplementary Figure S23: Inter correlations between somatic mutational single base substitutions (SBS) and insertion deletion (ID) signatures. SBS2 and SBS13 signatures related to APOBEC hyperactivity were positively correlated. The homologous recombination defect (HRD) related signatures SBS3 and ID6 showed highest positive correlation.

Discussion III

Development of a specificity-based nearest-centroid classifier

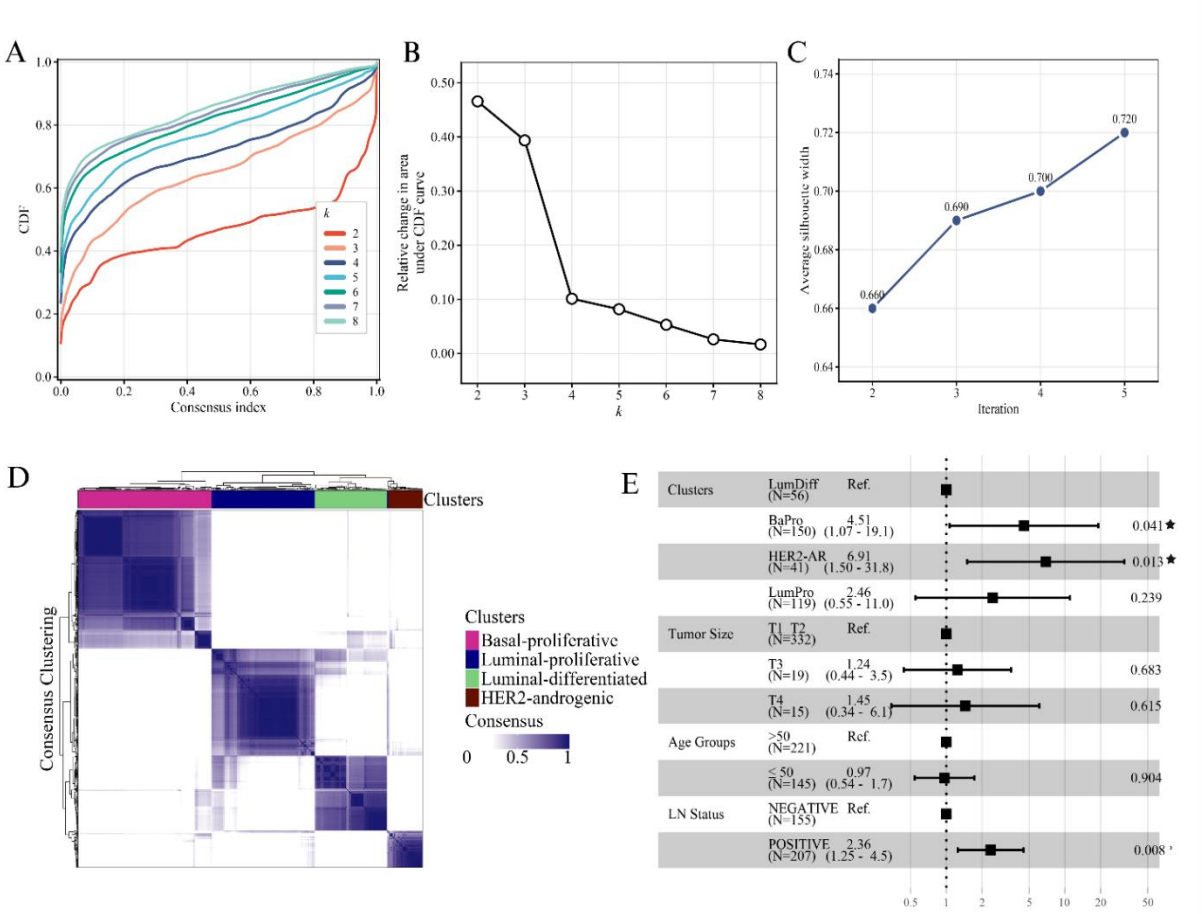
To generate a compact, biologically interpretable classifier for the four IBCGA molecular clusters (BaPro, HER2-AR, LumDiff, LumPro) (Supplementary Table S17), we calculated a specificity score for each gene as the difference between its mean z-scored expression in the target cluster and the maximum mean expression observed across the remaining three clusters. Genes were retained only if their mean expression within the target cluster exceeded a minimum threshold ($z > 0.5$), ensuring that selected markers were both cluster-specific and appreciably expressed rather than simply "least expressed elsewhere." Ranking genes by specificity within each cluster and selecting the top 25 per group yielded a 100-gene signature (25 genes × 4 clusters) restricted to protein-coding genes (Supplementary Figure S24).



Supplementary Figure S24. Specificity-based 100-gene signature demonstrates non-overlapping, cluster-specific expression across the four transcriptional states.

Heatmap of z-scored expression for the top 25 most cluster-specific genes per transcriptional state (100 genes total), selected by maximal differential expression of each cluster relative to the other three (specificity score = target cluster mean − maximum mean of remaining clusters, filtered for target mean expression > 0.5). Rows represent genes, grouped and colour-annotated by their cluster of specificity (left annotation bar); columns represent individual tumour samples, grouped by assigned cluster (top annotation bar). Colour scale indicates row-wise z-score expression from low (blue) to high (red). The block-diagonal pattern demonstrates that each 25-gene module is selectively upregulated within its corresponding cluster and largely absent in the other three, confirming that the four transcriptional states are defined by distinct, mutually exclusive gene expression programmes. This signature was used to construct nearest-centroid classifiers for cross-cohort validation (Supplementary Figure S2 6).

Cluster centroids were built by averaging the z-scored expression of the 100 signature genes within each subtype, and each sample was subsequently reassigned to a cluster by Pearson correlation to the four centroids (nearest-centroid classification), providing an internal consistency check on the original cluster assignments. (Supplementary Figure S26).



<u>Supplementary Figure S25</u>: All Breast Tumours: Consensus clustering stability and silhouette assessment.

(A) Consensus cumulative distribution function (CDF) curves for cluster numbers k=2 to k=8, showing increasing clustering stability with k =4. (B) Relative change in the area under the CDF curve (delta area plot), highlighting diminishing gains in stability beyond k=4. (C) Average silhouette widths increased over five iterative re-clustering rounds following removal of negative-silhouette samples, indicating enhanced cluster cohesion and separation. (D) Consensus clustering heatmap of 438 breast tumours showing four robust transcriptional clusters (Basal-proliferative, Luminal-differentiated, Luminal Proliferative, HER2 Androgenic). (E) Multivariable Cox proportional hazards regression confirms the independent prognostic value of cluster assignment after adjustment for key clinical covariates, including age, lymph node status, and tumour stage.

Signature specificity heatmap

Visualization of the 100-gene signature (heatmap) shows clear block-diagonal structure: each 25-gene module is markedly and preferentially upregulated in its corresponding cluster (BaPro markers including *HORMAD1*, *POU5F1*, *PRAME*, *CCNE1*; HER2-AR markers including *ETFA*, *KYNU*, *CEACAM7*, *IDH1*; LumDiff markers including *ADAM33*, *MYH11*, *CLDN5*, *COL17A1*; and LumPro markers including *DCAF10*, *COX6C*, *PTPN1*, *FRS2*), with comparatively low/neutral expression of each module in the other three clusters. This pattern confirms that the specificity-filtering approach isolated non-overlapping, mutually exclusive transcriptional programs for each subtype, and supports high nearest-centroid concordance with the original cluster labels - particularly for BaPro, HER2-AR, and LumDiff, whose blocks show the sharpest contrast; the LumPro block shows comparably strong but slightly less uniform enrichment, consistent with modest within-cluster heterogeneity.

Survival outcomes by molecular subtype

TCGA cohort. Disease-free survival (DFS) did not differ significantly among the four subtypes (log-rank p = 0.88; Supplementary Figure S26, Panel A). In contrast, overall survival (OS) differed significantly by subtype (log-rank p = 0.017; Supplementary Figure S26, Panel B), with HER2-AR showing poorer late OS and LumDiff showing a marked late decline.

METABRIC cohort. Both DFS (log-rank p < 0.001;, Supplementary Figure S26, Panel C) and OS (log-rank p < 0.001; Panel D) differed strongly by subtype in this larger series. HER2-AR consistently showed the worst DFS and OS trajectories, declining earliest and most steeply, while LumDiff showed the most favourable DFS/OS through early-to-mid follow-up before converging with the other subtypes at later time points.

Multivariable Cox regression

After adjustment for age, tumour (T) stage, and lymph node status, molecular subtype retained independent prognostic value in several models:

- *TCGA – DFS (*Supplementary Figure S26, *Panel E):* No subtype reached significance relative to LumDiff (reference), though BaPro trended toward worse DFS (HR 1.66, 95% CI 0.93–3.0, p = 0.085). T-stage (T2–T4) and N4+ nodal status were independently significant (T4 HR 4.62, p = 0.006; N4+ HR 2.45, p < 0.001).

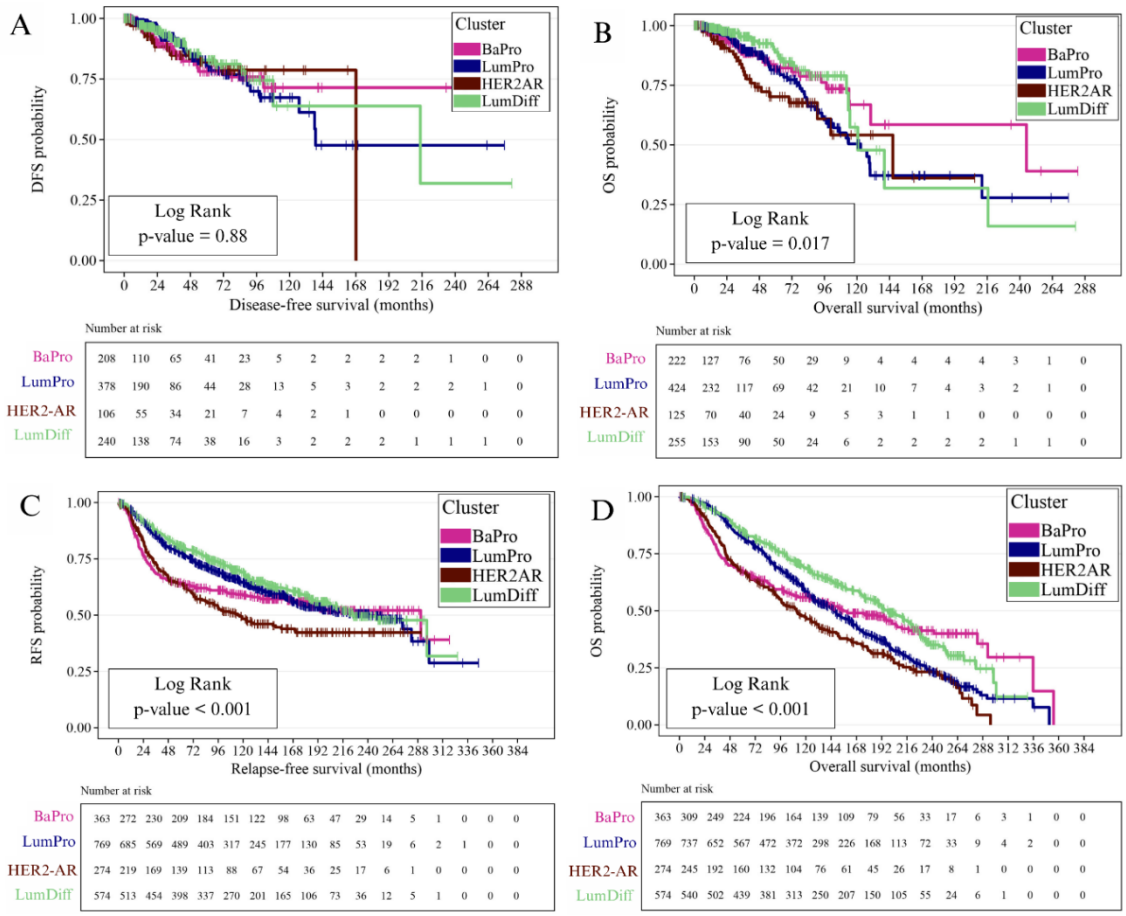
- *TCGA – OS (Supplementary Figure S26, Panel F):* LumPro was independently associated with worse OS versus LumDiff (HR 1.77, 95% CI 1.03–3.0, p = 0.039); BaPro and HER2-AR showed similar-magnitude but non-significant trends (HR ~1.8, p = 0.055 and 0.083,

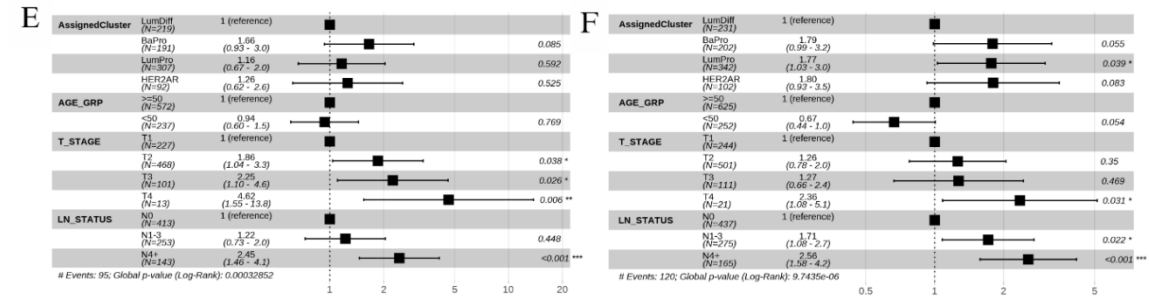
709 respectively). T4 stage (HR 2.36, p = 0.031) and N4+ status (HR 2.56, p < 0.001) were
710 independent adverse factors.

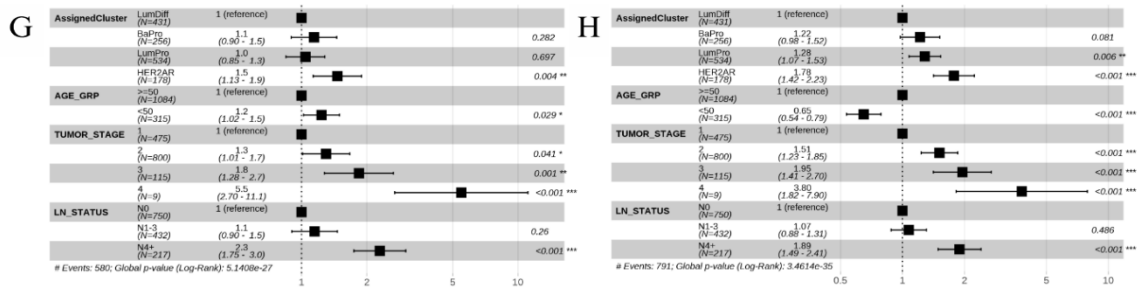
711 • *METABRIC – DFS (*Supplementary Figure S26, *Panel G):* HER2-AR was independently
712 associated with worse DFS versus LumDiff (HR 1.5, 95% CI 1.13–1.9, p = 0.004). Younger
713 age (<50), tumour stage 2–4, and N4+ status were also independently significant.

714 • *METABRIC – OS (*Supplementary Figure S26, *Panel H):* Both LumPro (HR 1.28, p = 0.006)
715 and HER2-AR (HR 1.78, p < 0.001) were independently associated with worse OS relative to
716 LumDiff, alongside tumour stage (stage 4 HR 3.80, p < 0.001) and N4+ nodal status (HR 1.89,
717 p < 0.001).

718 Taken together, the specificity-based centroid classifier produced a compact, non-overlapping 100-gene
719 signature that faithfully recapitulates the four molecular subtypes, and multivariate analysis across two
720 independent cohorts (TCGA and METABRIC) confirms that these subtypes - particularly HER2-AR
721 and LumPro - carry independent prognostic information for OS and DFS beyond standard
722 clinicopathologic variables (age, T-stage, nodal status).


A
DFS probability
Cluster
BaPro
LumPro
HER2AR
LumDiff
Log Rank
p-value = 0.88
Disease-free survival (months)
Number at risk
BaPro 208 110 65 41 23 5 2 2 2 2 1 0 0
LumPro 378 190 86 44 28 13 5 3 2 2 2 1 0
HER2-AR 106 55 34 21 7 4 2 1 0 0 0 0 0
LumDiff 240 138 74 38 16 3 2 2 2 1 1 1 0
B
OS probability
Log Rank
p-value = 0.017
Overall survival (months)
Number at risk
BaPro 222 127 76 50 29 9 4 4 4 4 3 1 0
LumPro 424 232 117 69 42 21 10 7 4 3 2 1 0
HER2-AR 125 70 40 24 9 5 3 1 1 0 0 0 0
LumDiff 255 153 90 50 24 6 2 2 2 2 1 1 0
C
RFS probability
Log Rank
p-value < 0.001
Relapse-free survival (months)
Number at risk
BaPro 363 272 230 209 184 151 122 98 63 47 29 14 5 1 0 0 0
LumPro 769 685 569 489 403 317 245 177 130 85 53 19 6 2 1 0 0
HER2-AR 274 219 169 139 113 88 67 54 36 25 17 6 1 0 0 0 0
LumDiff 574 513 454 398 337 270 201 165 106 73 36 12 5 1 0 0 0
D
OS probability
Log Rank
p-value < 0.001
Overall survival (months)
Number at risk
BaPro 363 309 249 224 196 164 139 109 79 56 33 17 6 3 1 0 0
LumPro 769 737 652 567 472 372 298 226 168 113 72 33 9 4 2 0 0
HER2-AR 274 245 192 160 132 104 76 61 45 26 17 8 1 0 0 0 0
LumDiff 574 540 502 439 381 313 250 207 150 105 55 24 6 1 0 0 0


E
AssignedCluster
LumDiff (N=219) 1 (reference)
BaPro (N=191) 1.66 (0.93 - 3.0) 0.085
LumPro (N=307) 1.16 (0.67 - 2.0) 0.592
HER2AR (N=92) 1.26 (0.62 - 2.6) 0.525
AGE_GRP
>=50 (N=572) 1 (reference)
<50 (N=237) 0.94 (0.60 - 1.5) 0.769
T_STAGE
T1 (N=227) 1 (reference)
T2 (N=468) 1.86 (1.04 - 3.3) 0.038 *
T3 (N=101) 2.25 (1.10 - 4.6) 0.026 *
T4 (N=13) 4.62 (1.55 - 13.8) 0.006 **
LN_STATUS
N0 (N=413) 1 (reference)
N1-3 (N=253) 1.22 (0.73 - 2.0) 0.448
N4+ (N=143) 2.45 (1.46 - 4.1) <0.001 ***
Events: 95; Global p-value (Log-Rank): 0.00032852
F
AssignedCluster
LumDiff (N=231) 1 (reference)
BaPro (N=202) 1.79 (0.99 - 3.2) 0.055
LumPro (N=342) 1.77 (1.03 - 3.0) 0.039 *
HER2AR (N=102) 1.80 (0.93 - 3.5) 0.083
AGE_GRP
>=50 (N=625) 1 (reference)
<50 (N=252) 0.67 (0.44 - 1.0) 0.054
T_STAGE
T1 (N=244) 1 (reference)
T2 (N=501) 1.26 (0.78 - 2.0) 0.35
T3 (N=111) 1.27 (0.66 - 2.4) 0.469
T4 (N=21) 2.36 (1.08 - 5.1) 0.031 *
LN_STATUS
N0 (N=437) 1 (reference)
N1-3 (N=275) 1.71 (1.08 - 2.7) 0.022 *
N4+ (N=165) 2.56 (1.58 - 4.2) <0.001 ***
Events: 120; Global p-value (Log-Rank): 9.7435e-06

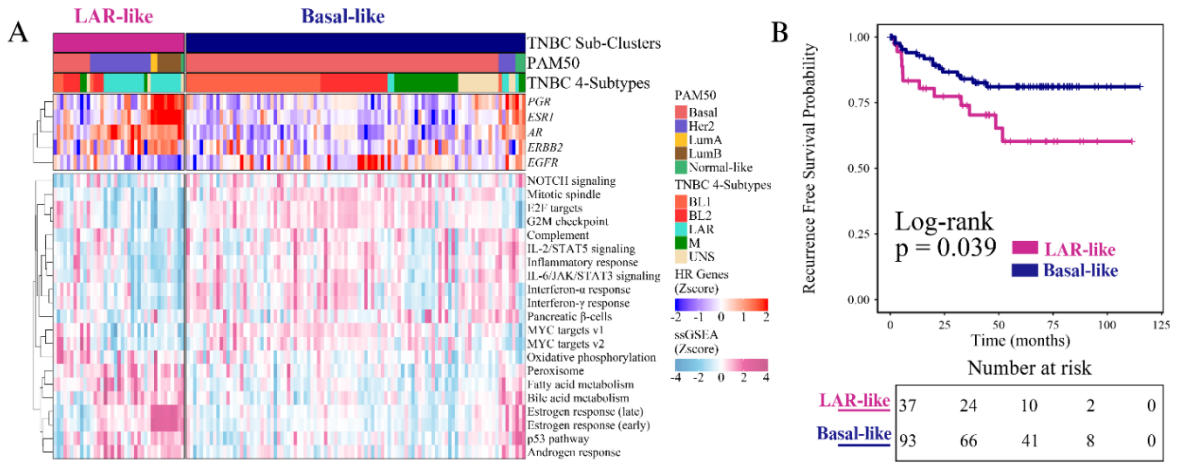

G
AssignedCluster
LumDiff (N=431) 1 (reference)
BaPro (N=256) 1.1 (0.90 - 1.5) 0.282
LumPro (N=534) 1.0 (0.85 - 1.3) 0.697
HER2AR (N=178) 1.5 (1.13 - 1.9) 0.004 **
AGE_GRP
>=50 (N=1084) 1 (reference)
<50 (N=315) 1.2 (1.02 - 1.5) 0.029 *
TUMOR_STAGE
1 (N=475) 1 (reference)
2 (N=800) 1.3 (1.01 - 1.7) 0.041 *
3 (N=115) 1.8 (1.28 - 2.7) 0.001 **
4 (N=9) 5.5 (2.70 - 11.1) <0.001 ***
LN_STATUS
N0 (N=750) 1 (reference)
N1-3 (N=432) 1.1 (0.90 - 1.5) 0.26
N4+ (N=217) 2.3 (1.75 - 3.0) <0.001 ***
Events: 580; Global p-value (Log-Rank): 5.1408e-27
H
AssignedCluster
LumDiff (N=431) 1 (reference)
BaPro (N=256) 1.22 (0.98 - 1.52) 0.081
LumPro (N=534) 1.28 (1.07 - 1.53) 0.006 **
HER2AR (N=178) 1.79 (1.42 - 2.23) <0.001 ***
AGE_GRP
>=50 (N=1084) 1 (reference)
<50 (N=315) 0.65 (0.54 - 0.79) <0.001 ***
TUMOR_STAGE
1 (N=475) 1 (reference)
2 (N=800) 1.51 (1.23 - 1.85) <0.001 ***
3 (N=115) 1.98 (1.41 - 2.70) <0.001 ***
4 (N=9) 3.80 (1.82 - 7.90) <0.001 ***
LN_STATUS
N0 (N=750) 1 (reference)
N1-3 (N=432) 1.07 (0.88 - 1.31) 0.486
N4+ (N=217) 1.89 (1.49 - 2.41) <0.001 ***
Events: 791; Global p-value (Log-Rank): 3.4614e-35

Supplementary Figure S26. Centroid-based classifier validation and survival stratification of transcriptional clusters across independent cohorts.
(A-D) Kaplan-Meier survival curves for the four transcriptional clusters (BaPro, HER2-AR, LumDiff, LumPro) in the TCGA (A-B) and METABRIC (C-D) external cohorts. (A) Disease-free survival (DFS) and (B) overall survival (OS) in TCGA (log-rank p = 0.88 and p = 0.017, respectively). (C) Disease-free survival (DFS) and (D) overall survival (OS) in METABRIC (log-rank p < 0.001 for both). Numbers at risk at each time interval are shown below each plot. (E-H) Forest plots of multivariable Cox proportional hazards regression evaluating assigned cluster (relative to LumDiff, reference), age group, tumour stage, and lymph node status as predictors of (E) DFS and (F) OS in TCGA, and (G) DFS and (H) OS in METABRIC. Boxes represent hazard ratios (HR) with 95% confidence intervals (horizontal lines); box size is proportional to the number of events. HR, 95% CI, and p-values are shown for each covariate level; global log-rank p-values and total event counts are indicated below each panel.

Transcriptional Sub-clustering Reveals Two Distinct TNBC Subtypes with Different Biological Programs

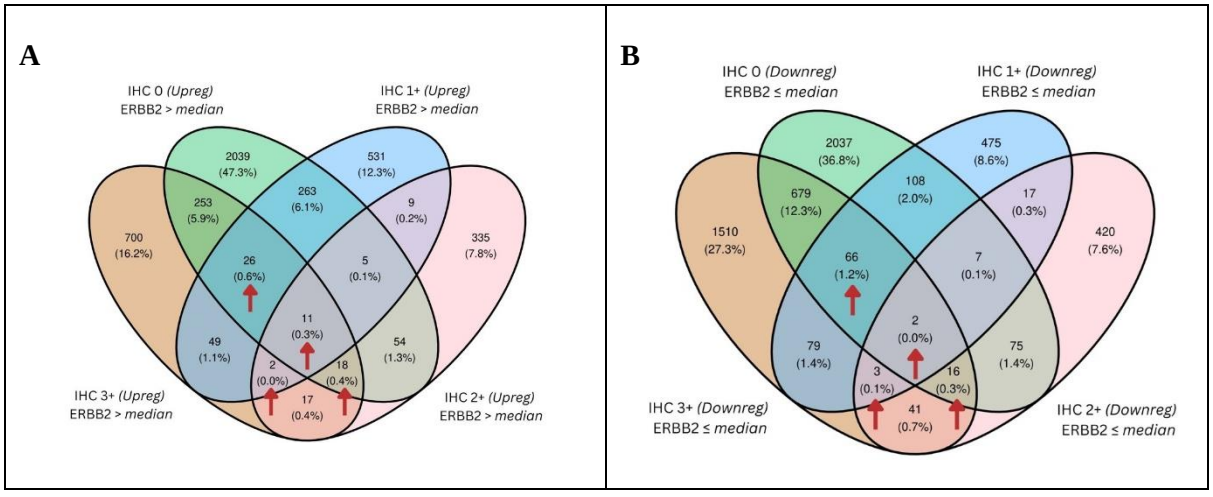
To identify the molecular features driving the divergent clinical outcomes, the tumours were stratified into distinct clusters and then examined the transcriptomic programmes that characterised each group. (Supplementary Tables S17, S18, S19, S20). Consensus clustering of 137 clinically defined TNBC tumours based on the top 3,500 most variable genes identified two major transcriptional subgroups: IBCGA derived clusters- LAR-like TNBC (N=39) and Basal-like TNBC (N=98), which showed strong concordance with PAM50 and TNBC4-subtype[35] classifications. The LAR-like subgroup exhibited elevated expression of hormone receptor genes (AR, PGR, ESR1, ERBB2, EGFR) and significant enrichment of metabolic pathways, including oxidative phosphorylation, fatty acid metabolism, adipogenesis, peroxisome activity, and hormone response signatures (estrogen and androgen response), while showing downregulation of interferon gamma and interferon alpha pathways that may contribute to worse survival outcomes. In contrast, the Basal-like subgroup was characterised by enrichment of proliferative and inflammatory programs such as mitotic spindle assembly, MYC targets, G2M checkpoint, E2F targets, interferon responses (gamma and alpha), inflammatory response, IL2-STAT5 signalling, IL6-JAK-STAT3 signalling, and TNFα signalling via NF-κB ($q < 0.05$) (Supplementary Figure S27A, Supplementary Table S23). The derivation of the HER2-associated 144-gene signature from concordant *ERBB2*-high versus *ERBB2*-low differential-expression patterns across HER2-IHC categories is shown in (Supplementary Figure S28; Supplementary Table S21). Kaplan-Meier survival analysis revealed that patients with LAR-like TNBC experienced significantly worse recurrence-free survival compared to those with Basal-like TNBC (log-rank p = 0.039), with this divergence becoming more pronounced over time. LAR-like TNBC classification remained independently associated with

increased recurrence risk after adjustment for age group and lymph node status, supporting the independent prognostic relevance of these transcriptional subtypes beyond traditional clinicopathological variables (Supplementary Figure S27B).



<u>Supplementary Figure S27</u>: Transcriptional sub-clustering and prognostic relevance of TNBC tumours

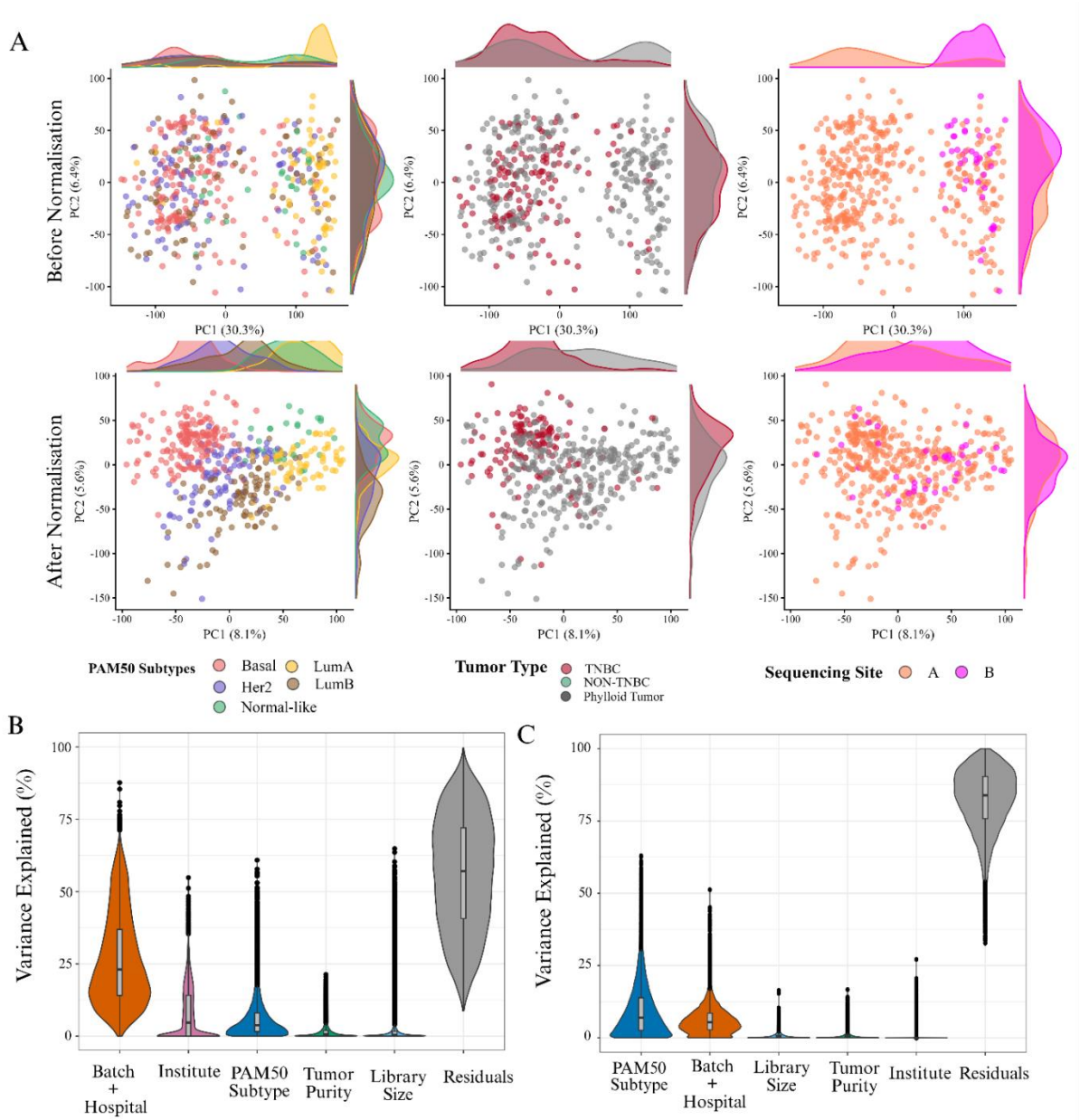
(A) Consensus clustering of 137 clinically defined TNBC tumours based on the 3,500 most variable genes identifies two transcriptional subgroups: IBCGA derived LAR-like and Basal-like TNBC subtypes. Top annotation indicates concordance with PAM50 and TNBC type-4 classifications. Lower panel shows Hallmark pathways significantly enriched by ssGSEA (FDR-adjusted p value $p < 0.05$), highlighting distinct biological programmes across subgroups. (B) Kaplan–Meier analysis of recurrence-free survival (RFS) reveals differential outcomes between LAR-like and Basal-like tumours (log-rank $p = 0.039$).



Supplementary Figure S28: HER2-associated 144-gene signature derived from differentially expressed genes in HER-IHC categories. Each of the tumour categories HER2- IHC 0, IHC 1+, IHC 2+ or IHC 3+ was independently divided into *ERBB2*-high or *ERBB2*-low using the within-category median *ERBB2* expression. Following this differential gene expression (DEG) analysis was performed for each IHC category between *ERBB2*-high vs *ERBB2*-low tumours. DEGs ($p < 0.05$) for each IHC category are shown as four-set Venn diagrams, upregulated genes, log2 fold change > 0 in **(A)** and downregulated genes, log2 fold change < 0 in **(B).** DEGs between IHC 3+ and IHC 0 were retained when concordant with at least two of the three (IHC 0, IHC 1+ or IHC 2+) *ERBB2*-high/low DEGs (marked with arrows in the figure). This gave a 144-gene DEG set which was designated as the HER2-associated signature (57 upregulated, 87 downregulated; Supplementary Table S21, S22).

Evaluation of Batch Effect and Normalization

Post-normalization analyses demonstrated a substantial reduction in variance associated with technical factors, including batch effects, library size, and tumour purity, while enhancing variance attributable to biologically relevant signals such as PAM50 subtypes. PCA performed before normalization revealed pronounced batch effects, with samples clustering predominantly by sequencing site, indicative of strong technical bias. Following normalization, this batch-driven separation was resolved, and samples no longer clustered by sequencing origin. Importantly, normalization also improved biological resolution: samples of the same PAM50 subtype clustered more closely together, and distinct separation between subtypes became more apparent in PCA space. These results confirm that the normalization strategy effectively mitigated technical artifacts while preserving and amplifying meaningful biological variation. (Supplementary Figure S29)



<u>Supplementary Figure S29</u>: Impact of normalization on batch effects and biological signal

(A) PCA of gene expression data before (top) and after (bottom) normalization. Initially, samples cluster by sequencing site with weak separation between biological groups, indicating strong batch effects. After normalization, this technical clustering is reduced, and the samples are grouped more clearly by PAM50 subtype and tumour category. (B) Variance partitioning before normalization shows that technical factors, particularly sequencing sites, explain a large proportion of gene expression variability, exceeding biological contributions. (C) After normalization, the contribution of batch effect is markedly reduced, and PAM50 subtype, accounted for a greater proportion of variance.

PAM50 intrinsic subtype composition highlights distinct patterns in IBCGA

We evaluated the distribution of PAM50 intrinsic molecular subtypes across the IBCGA cohort and compared it with three major external breast cancer datasets (TCGA, METABRIC, and SCAN-B) (Supplementary Table S13). Subtype assignments were generated using the AIMS classifier. A global comparison of subtype frequencies across cohorts demonstrated a highly significant difference in overall distribution (global $p < 0.01$).

Within the IBCGA cohort, Basal tumours constituted the largest subtype, accounting for 38.4% of cases. This proportion was notably higher than that observed in TCGA (17.5%), METABRIC (16.7%), and SCAN-B (9.3%). In contrast, Luminal A tumours represented only 14.3% of IBCGA cases, compared with 36.1% in TCGA, 27.1% in METABRIC, and 48.3% in SCAN-B. Luminal B tumours comprised 19.5% of the IBCGA cohort, a frequency comparable to TCGA (18%) but lower than METABRIC (27.8%) and higher than SCAN-B (11.7%). HER2-enriched tumours accounted for 20.8% of IBCGA cases, like TCGA (20.4%) and exceeding the proportions observed in METABRIC (14.9%) and SCAN-B (11.6%). Normal-like tumours were the least frequent subtype in IBCGA (7%), compared with 8% in TCGA, 13.5% in METABRIC, and 19.1% in SCAN-B (Supplementary Figure S30A, Supplementary Tables S13, S14, S15, S16).

Overall, the intrinsic subtype landscape of the Indian IBCGA cohort is characterised by a marked predominance of Basal-like tumours and a reduced representation of Luminal A disease relative to predominantly Western cohorts, underscoring important population-level differences in tumour biology.

TNBC molecular subtype distribution across cohorts relative to IBCGA

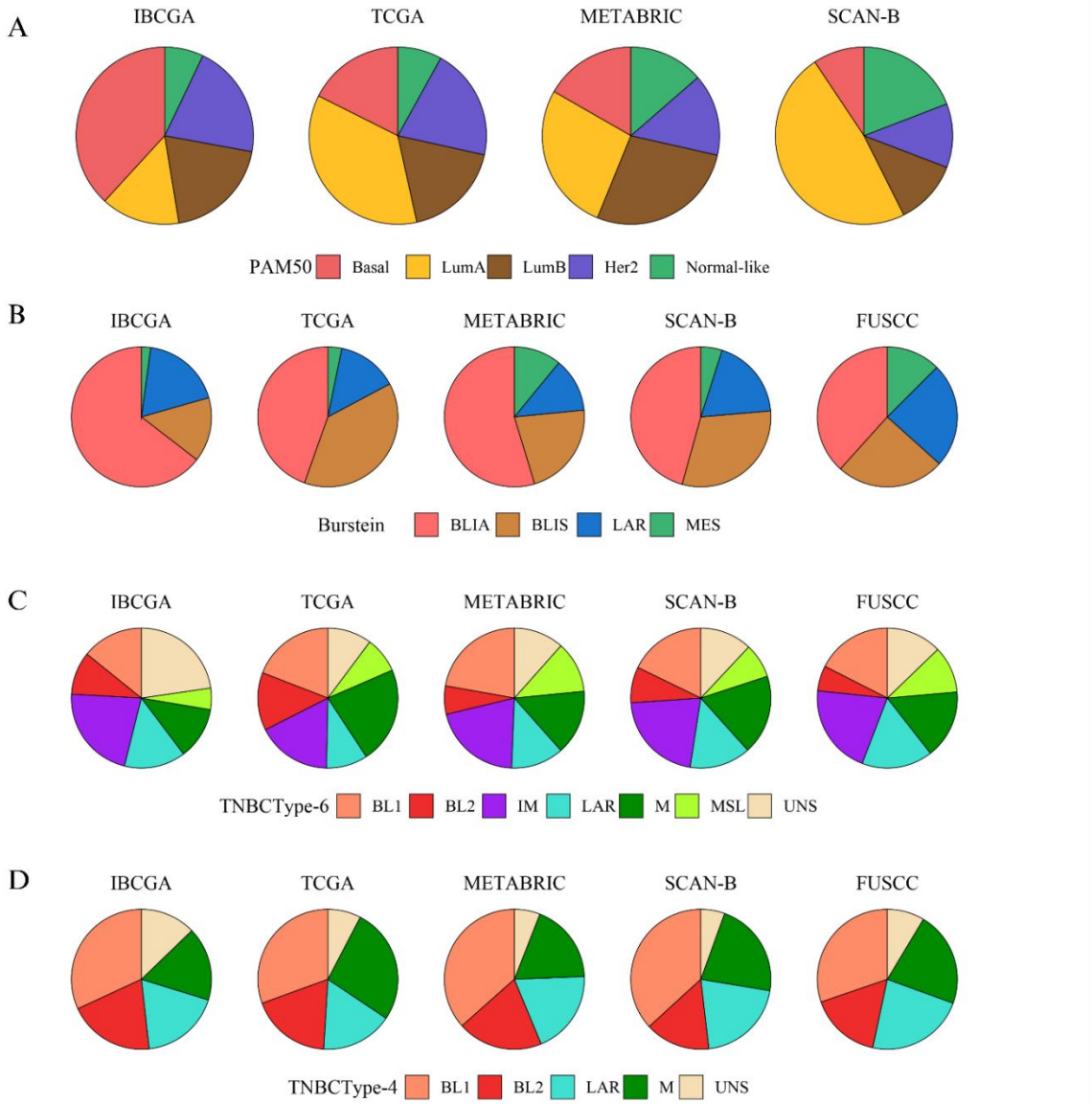
Given the high prevalence of triple-negative breast cancer (TNBC) within the IBCGA cohort, we next examined TNBC-specific transcriptional heterogeneity using three established molecular classification frameworks: TNBC Type-4[35], TNBC Type-6[36], and the Burstein subtype scheme[37]. Subtype distributions were compared across IBCGA, TCGA, METABRIC, SCAN-B, and FUSCC cohorts. Global comparisons were performed for each classification model.

Under the TNBC Type-6 model, subtype composition differed significantly across cohorts (global $p < 0.01$). In the IBCGA cohort, IM (21.9%) and UNS (23.1%) were among the most frequent subtypes, followed by BL1 (14.1%) and LAR (14.1%). Compared with other cohorts, IBCGA demonstrated a higher proportion of UNS tumours relative to TCGA (10.1%), METABRIC (11.5%), SCAN-B (11.9%), and FUSCC (12.7%). The mesenchymal (M) subtype was less frequent in IBCGA (12%) than in TCGA (22.7%) and SCAN-B (18.4%), while MSL tumours were comparatively lower in IBCGA (4.9%) than in most external datasets (Supplementary Figure S30C, Supplementary Table S14) In contrast, the TNBC Type-4 model[35] showed no significant difference in overall subtype distribution across cohorts (global $p = 0.135$). Within IBCGA, BL1 represented the largest group (32.1%), followed by BL2

(19.8%), LAR (18.4%), and M (17%). These proportions were broadly comparable to those observed in TCGA, METABRIC, SCAN-B, and FUSCC, indicating a conserved subtype structure across populations under this framework. (Supplementary Figure S30D, Supplementary Table S15)

The Burstein classification revealed a highly significant difference in subtype distribution across cohorts (global $p < 0.01$). In IBCGA, BLIA was the predominant subtype, comprising 64.7% of TNBC cases - substantially higher than TCGA (44.7%), METABRIC (54.8%), SCAN-B (45.8%), and FUSCC (38.4%). Conversely, BLIS tumours accounted for only 14.8% of IBCGA cases, compared with markedly higher proportions in TCGA (38.2%) and SCAN-B (30.6%). The MES subtype was relatively rare in IBCGA (2.1%) compared with METABRIC (10.9%) and FUSCC (12.5%), while LAR frequencies in IBCGA (18.4%) were comparable to SCAN-B (18.7%) but lower than FUSCC (24.1%) (Supplementary Figure S30B, Supplementary Table S16).

Overall, these findings indicate that TNBC transcriptional heterogeneity within the Indian IBCGA cohort depends on the classification framework applied. While the Lehmann TNBC Type-4 model demonstrates broadly conserved subtype architecture, the TNBC Type-6 and Burstein schemes reveal significant shifts in subtype composition, particularly the marked enrichment of BLIA and reduced BLIS representation in IBCGA. These results underscore population-level differences in TNBC biology and highlight the importance of context-specific molecular stratification.



<u>Supplementary Figure S30</u>: Molecular subtype distribution across breast cancer cohorts (A) PAM50 subtype composition across IBCGA, TCGA, METABRIC, and SCAN-B. (B) Burstein TNBC subtypes across IBCGA and external cohorts. (C) TNBC Type-6 subtype patterns across cohorts. (D) TNBC Type-4 subtype patterns across cohorts. UNS, unassigned category

1. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
2. Das, S., Biswas, N. K. & Basu, A. Mapinsights: deep exploration of quality issues and error profiles in high-throughput sequence data. *Nucleic Acids Res.* **51**, E75–E75 (2023).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
5. Wang, M., Xu, Z., Wang, Z., Xu, X. & Sun, Y. Analysis of the mechanism of berberine against stomach carcinoma based on network pharmacology and experimental validation. *Transl. Cancer Res.* **13**, 4593–4607 (2024).
6. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
7. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
8. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).
9. Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1-15.9.17 (2016).
10. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, (2011).
11. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
12. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
13. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
14. Stephenson, J. D. *et al.* ProtVar: mapping and contextualizing human missense variation. *Nucleic Acids Res.* **52**, W140–W147 (2024).
15. Yang, X. *et al.* New Insights into PI3K Inhibitor Design using X-ray Structures of PI3Kα Complexed with a Potent Lead Compound. *Sci. Rep.* **7**, 1–7 (2017).
16. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
17. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications 2022 13:1* **13**, 1265- (2022).
18. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, (2021).
19. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
20. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

21. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
22. Jiang, Z. *et al.* Isthmin-1 is an adipokine that promotes glucose uptake and improves glucose tolerance and hepatic steatosis. *Cell Metab.* **33**, 1836-1852.e11 (2021).
23. Qiang, W. H. *et al.* ISM2 as a prognostic biomarker and mediator of immune infiltration in colorectal cancer: evidence from bioinformatics and experimental analysis. *J. Gastrointest. Oncol.* **16**, 2026–2043 (2025).
24. Zhang, H. Y., Chen, W. C., Fu, X. Y., Su, X. & Yang, A. K. CBX3 promotes tumor proliferation by regulating G1/S phase via p21 downregulation and associates with poor prognosis in tongue squamous cell carcinoma. *Gene* **654**, 49–56 (2018).
25. Niu, H., Chen, P., Fan, L. & Sun, B. Comprehensive pan-cancer analysis on CBX3 as a prognostic and immunological biomarker. *BMC Med. Genomics* **15**, 29 (2022).
26. Shen, Y. *et al.* Author Correction: Loss-of-function mutations in QRICH2 cause male infertility with multiple morphological abnormalities of the sperm flagella. *Nat. Commun.* **10**, (2019).
27. Saotome, K. *et al.* Structures of the Otopetrin Proton Channels Otop1 and Otop3. *Nat. Struct. Mol. Biol.* **26**, 518 (2019).
28. Gan, N., Zeng, W., Han, Y., Chen, Q. & Jiang, Y. Structural mechanism of proton conduction in otopetrin proton channel. *Nature Communications 2024 15:1* **15**, 7250- (2024).
29. Li, D. *et al.* Otopetrin 1 protects against adipose tissue wasting during cancer cachexia progression. *iScience* **29**, 116461 (2026).
30. Huang, C. H. *et al.* The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations. *Science* **318**, 1744–1748 (2007).
31. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
32. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).
33. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, (2016).
34. Kumar, R. *et al.* HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics* **19**, (2019).
35. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).
36. Lehmann, B. D. *et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* **11**, (2016).
37. Burstein, M. D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).