# Supplementary Information

## Supplementary Methods

### Patient cohort and selection

During September 1 2010 to March 31 2015, 408 patients were diagnosed with TNBC in the Region Skåne healthcare area based on data from the Swedish national breast cancer quality registry (NKBC). To derive this patient set the following exclusion criteria were used in the INCA technical platform in a two-step fashion:

*1: Removing non-TNBC cases*

- Cases that were not ER-negative (A100ER_Värde=2)
- Cases that were not PR-negative (A100PR_Värde=2)
- Cases that were not HER2-negative (A100HER2_Värde=2)

*2: Removing TNBC cases with unclear treatment history*

- Cases with no planned surgery (INCA parameter A050PrimOp_Värde=0) caused by
  - i) missing data for parameter A050EjOpOrsk_Värde,
    or
  - ii) other reason A050EjOpOrsk_Värde=4
    were removed
- Cases that did not have indication of planned postoperative treatment yes/no OR planned preoperative treatment yes/no were removed. This corresponded to including patients that fulfilled INCA parameters: (A120PostOpBeh2_Värde=1) or (A120PostOpBeh2_Värde=0) or (A070PreBeh_Värde=0) or (A070PreBeh_Värde=1)

Briefly, criteria 2 above excluded TNBC patients with an unclear/unknown treatment status based on registry data, irrespective of the type of treatment given. This meant that the identified and retained patients could have had neoadjuvant treatment, adjuvant systemic treatment, no treatment, or even palliative treatment due to metastatic disease already at time of diagnosis (thus including these patient categories in subsequent cohorts). Of the 408 patients, 340 provided informed written consent and were enrolled in the Sweden Cancerome Analysis – Breast (SCAN-B) study[1-3] (ClinicalTrials.gov ID NCT02306096). The final tally of 254 samples were selected into this study based on also having available quality controlled RNA sequencing (RNAseq) data from SCAN-B, sufficient DNA, and passing extensive review of available clinical data from individual patient´s files by a senior oncologist. Corresponding RNAseq data for primary cases has been deposited in GEO series GSE96058[4,5] based on a previous study (outlining quality control filters). The 254 patients were diagnosed at any of the four main hospitals in the Region Skåne healthcare region, with a catchment area of approximately 1.3 million inhabitants (year 2017).

### Tissue sampling, DNA and RNA extraction

Fresh tumor tissue samples preserved in RNAlater (Qiagen, Hilden, Germany) were obtained in conjunction with routine clinical sampling by a diagnostic pathologist in regional pathology departments (see [3] for outline). RNA and DNA were extracted using the Qiagen Allprep extraction kit (Qiagen) as described[1]. DNA from whole blood was extracted by the Labmedicin Skåne Biobank, Lund, Sweden.

## Prior germline testing and classification of *BRCA1* and *BRCA2* germline variants

49 patients had prior clinical genetic counseling involving NGS-based screening of *BRCA1* and *BRCA2*, or were enrolled in the SWEA research study (The Swedish BRCA1 and BRCA2 study collaborators (SWE-BRCA) Extended Analysis) for high-risk patients and screened for an extended panel of susceptibility genes. The inclusion criteria for the SWEA study were in line with the Swedish national clinical practice guidelines for breast cancer. Briefly, genetic testing was offered when there was at least a 10 % probability to detect a pathogenic germline variant in *BRCA1* or *BRCA2*, based on the patient's age at diagnosis, histology, and family history. Detected germline variants were classified according to the ENIGMA BRCA1/2 Gene variant Classification Criteria (2017-06-29) https://enigmaconsortium.org/library/general-documents/. Only class 5 variants were considered as pathogenic, corresponding to nine *BRCA1* and three *BRCA2* variants.

## DNA promoter methylation analysis

Bisulfite conversion of genomic DNA was performed with the column based EpiTecht Fast DNA Bisulfite kit (Qiagen GmBH, Hilden, Germany. Promoter methylation analysis was performed using a PSQ MD 96 pyrosequencing instrument (Qiagen). A fully methylated as well as an unmethylated sample was included as controls in each run. The PyroMark analysis program was used for data analysis and all electropherograms were manually checked. For *BRCA1*, analysis was performed as described[6], and included analysis of two CpG island regions. We used 7% cut-off compared to previous 10% due to cleanness of the data.

For *RAD51C* we first performed a correlation analysis based on TCGA breast cancers used in a previous study[7] similar to the study by Polak et al.[8]. This analysis identified seven CpGs present on the Illumina 450K methylation beadchip array <1000bp upstream of the gene. Four CpGs with a Pearson correlation less than -0.44 (cg05214530, cg27221688, cg02118635, cg10487724) were selected, from which primers for *RAD51C* were adapted from Hansmann et al.[9].

Final RAD51C primers were:
RAD51C_PCR_F 5'-NNATGGTGTATAAGTGTGAAAATTTATAAG*A-3'*
RAD51C_PCR_R 5'-biotin-CCTCTAAAAATTCCTCAACAATCTAAA-3'

RAD51C_SEQ_1 5'-ATTGAGTAAAGTTGTAAGGT-3'
RAD51C_SEQ_2 5'-GGGGTTAGTAGGTGAGTTTG-3'

In the final analysis, two different primer sets of four (BRCA1 and RAD51C) different CpG sites were used for each gene. CpG allele methylation percentage was averaged across each primer set and next merged to the mean of the two sets. Cut-offs were applied (BRCA1 7%, RAD51C 9%) for making a call on methylation or otherwise. The cut-off was set higher for RAD51C as this assay generated slightly higher background variability between primer sets. The cut-off was verified (and supported) by RNA sequencing data for *RAD51C*. Two cases failed RAD51C methylation analysis.

For *PALB2* and *RAD51* promoter hypermethylation analysis primers described by Wanatabe et al.[10] were used. A similar cut-off (9%) as for RAD51C was used to call methylated cases (none observed).

## IHC validation of mismatch repair deficient (MMRd) cases

Suspected MMRd cases identified by whole genome sequencing were stained for MLH1, PMS2, MSH2 and MSH6 as outlined originally in Joost et al.[11]. Stained slides were evaluated for protein expression in both tumor cells and non-malignant cells.

## Gene expression analyses

Gene expression data was available from Gene Expression Omnibus[4], series GSE96058 reported elsewhere[5]. FPKM data for specific genes were extracted and log2 transformed. For TNBCtype, IC10, CIT classification, 228 cases were available for analysis based on GSE96058[5] (primary tumors only). For remaining cases, these were included separately and subtyped only using AIMS[12] (as this is a single sample predictor of molecular subtype) and analyzed for individual FPKM gene expression.

Classification according to different molecular subgroups in breast cancer was performed as follows, after *i*) an offset of 1 was added to all FPKM values, *ii*) log2 transformation:

- *PAM50*. PAM50 subtypes were obtained using the AIMS single sample classifier[12], based on the aims R package. All samples were classified.
- *TNBCtype*[13,14]. For TNBCtype classification the entire GSE96058 data set was used. Data was mean-centered across all samples for each gene, TNBC cases were extracted and uploaded as a separate data set into the web-based classifier[14]. For a few cases the web-based application called the samples as not being ER-negative. These samples were removed from the TNBC data set (inferring missing values) and remaining samples were again uploaded to the web-based application for subtyping.
- *IC10*[15]. For IC10 classification the entire GSE96058 data set was used. Data was mean-centered across all samples for each gene. IC10 subgroups were obtained through the ic10 R package using default processing.
- *CIT*[16]. CIT subtypes were obtained through the citbcmst R package, using pearson correlation as distance method and gene symbol as matching entity.

Calculation of six gene expression metagenes representing different biological functions in breast cancer was performed as described[17], using the GSE96058 data set for gene-centering across samples.

*Unsupervised clustering*

All unsupervised analysis was performed in R [18] using the ConsensusClusterPlus R-package [19]. Two input formats were used, FPKM data and PCA components.

When using FPKM data as input this was first offset by +0.1, log2 transformed, and mean-centered across samples for each RefSeq associated gene. A filter step based on standard deviation of expression was used as defined in result presentations. In the clustering we used Pearson correlation as distance metric and ward.D2 linkage. Additional parameters were pItem=0.8, pFeature=0.8, number of iterations = 2000.

When using PCA components as input, these were derived from a PCA analysis of all 19000 RefSeq genes available using the *prcomp()* function in R. All components were then used in consensus clustering with pItem=0.8, pFeature=0.98, number of iterations = 2000.

*Machine learning*
All machine learning was performed in R [18] using the Caret package. Model performance was assessed using ROC analysis and the area under the curve (AUC) estimate. We used 70% of samples for training and 30% for internal validation. Division into training and internal validation was performed using the *createDataPartition*() Caret function, balancing the sample splitting for grade (1,2,3), lymph node status (node-negative, node-positive), and age (<50, >=50 years). For FPKM data as input this was offset by +0.1, log2 transformed and mean-centered across samples in respective cohort (training and test). For machine learning using PCA components (see above for extracting these) all components were used directly in training.

We evaluated seven machine-learning methods listed below by their names in Caret:

- *svmLinear* (linear support vector machine)
- *gbm*
- *pam*
- *rf* (random forest)
- *glm*
- *glmboost*
- *knn* (k nearest neighbor)

In the training we used the *trainControl*() Caret function with parameters:
method=repeatedcv
number=4
repeats=10
classProbs=TRUE
summaryFunction=twoClassSummary
savePredictions=TRUE

For the actual training we used the *train*() Caret function with ROC as metric, the defined trainControl parameters above, and tune.length=10. Trained models were saved as R data objects. We applied the models to the internal validation set for each evaluation group using the *predict*() Caret function. ROC values were calculated using the *roc*() function in the pROC R package.

The entire process described above was repeated 10 times with different splits of training and internal validation sets to reduce the possibility of bias in the results based on sample splitting. This generated 10 different AUC estimates from 10 potentially different predictors for each model.

## Survival analyses

*Definition of clinical endpoints:*

- Overall survival was obtained from national registries, calculated as the time from diagnosis to death of any cause.

- Invasive disease-free survival (IDFS) was defined according to STEEP guidelines[20], as the time from diagnosis to either death of any cause or invasive breast-cancer related events (loco-regional and distant recurrence).
- Distant relapse-free interval (DRFI) was defined according to STEEP guidelines as the time from surgery to diagnosis a distant relapse (event) or to last day of follow-up (censoring). Events include patients that first developed a loco-regional relapse, and then a distant relapse. For these patients the day of the distant relapse was used.

*Exclusion criteria for outcome analyses:*

- Neoadjuvant treatment
- Metastatic disease at time of diagnosis (including microinvasive disease).
- Metastatic disease identified immediately prior to, or during adjuvant chemotherapy.
- Patients not managed in an adjuvant setting (irrespective if adjuvant treatment or not provided later).
- Bilateral breast cancer.
- Lost to follow-up before start of systemic treatment.
- Unclear histological type (one case).
- For DRFI, patients with a relapse or death from a malignancy of uncertain origin were excluded. These patients were however included in OS and IDFS analyses.

*Multivariable analyses*

Analysis was performed using the coxph R function from the survival R package. Covariates in multivariable Cox regression were patient age (<50, ≥50 years), lymph node status (N0/N+), tumor size (≤20, >20mm), and tumor grade (1,2,3). Data for lymph node status, and tumor size were obtained from NKBC data. Tumor grade was obtained from clinical review of individual patient's files.

# Whole Genome Sequencing Analysis

## Whole genome sequencing and alignment

150 base pair paired-end sequencing was performed on Illumina X10 machines following standard library preparation according to the manufacturer's protocols. Target insert size was 450 bp. The target coverage for tumour-normal pairs was 30x for patients that received adjuvant chemotherapy (based on NKBC registry data, see above), and 15x otherwise. Paired-end reads were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner, BWA (v0.7.15).

## Identification of somatic mutations

Paired tumour-normal bam files were interrogated for somatic mutations using the following algorithms:

- Caveman (1.11.0, 1.11.5) for identification of somatic point mutations https://github.com/cancerit/CaVEMan
  Jones (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data.[21]

- Pindel (2.1.0, 2.2.4) for identification of somatic small insertions and deletions https://github.com/cancerit/cgpPindel
  Raine (2016). cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing.[22].
- Brass (5.3.3, 6.0.5) for identification of somatic rearrangements https://github.com/cancerit/BRASS
- ASCAT (4.0.0, 4.0.1) for identification of somatic copy number changes https://github.com/cancerit/ascatNgs
- Battenberg (3.0.1, 3.2.2) for identification of subclonal copy number changes https://github.com/cancerit/cgpBattenberg

To ensure that the final dataset had high specificity for signature analysis:

- Point mutations were additionally filtered by CLPM (median number of soft clipped bases in variant supporting reads) and ASMD (median alignment score of reads showing the variant allele) filters (CLPM=0 and ASMD>=140).
- Indels were filtered by QUAL (variant quality score assigned by Pindel) and REP (change repeat count within range) (QUAL>=250 and REP<10) to ensure high specificity.
- Both point mutations and indels were additionally annotated with respect to presence in databases of common human variation using Annovar (1000 genomes August 2015, ExAC 20151129, dbSNP147).
  Wang K. et al. (2010). ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data.[23]
- All rearrangements had to pass the stage of re-assembly in Brass, and were further filtered by size (at least 1kb between breakpoints). A known library-preparation artefact was removed by filtering out inversions that were shorter than 5kb and reported by 5 or fewer reads.

For seeking cancer driver mutations, we applied less stringency in specificity.

## Criteria for sample exclusion from analysis

DNA was extracted from diagnostic cancer samples in the population-based study as described above (section ***Tissue sampling, DNA and RNA extraction***). For some patients it is possible that the amount of tumor DNA present in the sample was insufficient for whole-genome sequencing and interpretation of tumour-specific somatic mutations.

Samples were excluded from analysis if:

- the library preparation failed to yield sufficient output
- for samples with coverage below 20x, Battenberg algorithm failed to find a copy-number solution
- for samples with coverage over 20x, Battenberg failed to find a copy-number solution and the the number of somatic point mutations identified was less than 500.

## Identification and curation of germline and somatic mutations in genes related to DNA repair

We interrogated germline and somatic mutations in genes related to DNA repair. The list of 165 genes included genes tested in the SWEA study (see above in section ***Prior germline testing and classification of BRCA1 and BRCA2 germline variants***) and genes included in the INSIGNIA study (experimental study to explore mutagenesis in isogenic cell systems and in patients with inherited DNA repair defects, **www.mutationsignatures.org**). Screened genes are listed in the Supplementary Data file.

In order to prioritise germline mutations which are likely pathogenic, the variants were filtered to ensure the mutations were real (supported by at least 5 reads and not in a region of repetitive sequence, by curation), and further filtered according to prior information on particular mutations (protein consequence, frequency in population through the ExAC database less than 1 in 100, and if considered pathogenic in the ClinVar database, dated 20170501). Nonsense, frameshift or essential splice-site mutations were considered pathogenic, unless they were known to be exceptions (eg. p.K3326* BRCA2 nonsense).

Germline copy number in most commonly-mutated DNA-repair genes (*BRCA1*, *BRCA2*, *PALB2*) were assessed through visual inspection of coverage in the bam files.

Among somatic mutations in the DNA repair genes, as pathogenic we considered those that were either nonsense, frameshift or essential splice site. Somatic rearrangements from BRASS algorithm were also included if a breakpoint disrupted the gene body.

Finally, somatic copy number at the DNA-repair genes was inferred from the genome-wide copy number profiles from the ASCAT algorithm. When total copy number is zero in the tumour, the gene was considered lost (homozygous deletion), and when it was one, it was flagged as loss of heterozygosity.

## Identification of cancer drivers

Among point mutations and small insertions and deletions, only coding and essential splice site mutations were considered as potential drivers. To be classified as putative drivers, missense or in-frame variants in dominant genes had to either be recurrent in literature (e.g. COSMIC database) or be deleterious to previously reported recessive cancer genes. Mutations in 740 genes identified in literature as cancer drivers were considered, together with classification regarding dominant or recessive status of each gene.

Copy number profiles of the cancer genomes were scrutinised for changes to common cancer genes. The ASCAT algorithm provided estimates of overall tumour ploidy, cellularity and copy number profiles. Copy number changes were considered putative drivers when they coincided with frequently altered genes[24]. Specifically, copy number changes were deemed as an amplifications when total copy number in the tumour exceeded (p*2+1), where p stands for overall tumour ploidy. Homozygous deletions were declared when total copy number in tumour was zero. Loss of heterozygosity was determined when total copy number of a gene was one and minor copy number was zero.

Finally, previous analysis[24] revealed increased frequencies of somatic rearrangements in gene bodies of several recessive cancer genes, compared to the surrounding chromosomal regions. We declared some somatic rearrangements as putative drivers when they disrupted gene bodies, with the exception of instances when both breakpoints of rearrangements hit the same intron. The list of recessive cancer genes included: *ARID1B*, *CDKN2A*, *FBXW7*, *MAP2K4*, *MAP3K1*, *PTEN*, *RB1*, *TP53*, and *MLLT4*.

## Application of the HRDetect algorithm

A high-specificity somatic mutation dataset was first compiled for all cancer genomes in the study as described in above (section ***Identification of somatic mutations***).

Cellularity and ploidy estimates were obtained from the output of ASCAT.

Point mutations were annotated with the sequence context (and reverse-complemented when necessary to put them into pyrimidine context). Signatures of point mutations were assigned with the SigFit R package using the 30 signatures from the COSMIC signature database (https://cancer.sanger.ac.uk/cosmic/signatures). Signatures of rearrangements were also analysed by SigFit using the six rearrangement signatures from Nik-Zainal et al.[24]. Small deletions were classified into the following categories: micro-homology mediated, repeat-mediated and others. HRD score was calculated from ASCAT copy number profiles as described previously[24].

Exposures of signatures, HRDetect score, and proportion of indels at microhomology were tabulated for the entire cohort. The features were normalised as described in the original report describing HRDetect[25]. The HRDetect weights were applied to the normalised features, to obtain HRDetect probability for every patient.

## Principal-component analysis of components of HRDetect

HRDetect assigns a single probability of HR deficiency to every sample. However, we have previously demonstrated that *BRCA1*-null and *BRCA2*-null tumors were distinguishable. To explore the sub-structure of cancers with HR deficiency, we conducted a Principal-component analysis (PCA) of components of HRDetect.

Specifically, the products of normalised features and HRDetect weights were tabulated as matrix N x 6, where N is the number of samples in the study and 6 represents the 6 non-zero weights of HRDetect. Each entry in the matrix is a product of HRDetect weight and the respective normalised sample feature. Prior to the PCA analysis, the variables in the new matrix were not scaled, so that the dimensions with highest weights of HRDetect exhibited highest variance.

Two-dimensional visualisations of the PCA results are presented in Figure 3.

## Mobile element (MELT) analysis

Mobile element (ME) insertions not present in the GRCh37 reference genome were called and genotyped using the Mobile Element Locator Tool (MELT) version 2.1.5 [26]. The ME reference sequences provided with MELT were used for discovery of three major classes of MEs: *Alu*, SVA (SINE-VNTR- *Alu*) and LINE1 (long interspersed nucleotide element-1). Default parameters were used except -z (set to 50000) and read length, coverage and insert length, which were determined for each bam file. Tumor and normal samples from the same individual were analyzed together to increase sensitivity (MELT GroupAnalysis and MakeVCF). ME insertions previously discovered as part of the 1000 Genomes Projects, Phase III, were used as priors [27,28]. Supporting aligned read data for ME insertions discovered in or near (within 5 kb) 11 genes with potential impact on HR (*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *BRIP1*, *CHEK2*, *MRE11*, *NBN*, *PALB2*, *RAD51C*, *RAD51D*) were visually inspected in Integrative Genomics Viewer (IGV [29]).

On average, 1302 *Alu*, 186 LINE1 and 90 SVA insertions were detected in each pair of tumor and normal sample. Of these, 90.5% of the *Alu*, 79.5% of the LINE1 and 86.3% of the SVA insertions were previously seen in the 1000 Genomes Projects. Most insertions were detected in both the normal and tumor data (93.1% of *Alu*, 89.3% of LINE1 and 86.3% of SVA elements).

Two ME insertions were found in the 11 HR related genes. One was a known *Alu* insertion in *BRCA2* at chr13: 32922439 found in two sample pairs (ID: ALU_umary_ALU_9673, detected in both normal and tumor). This *Alu* insertion has an allele frequency of 4.2% in the 1000 genomes project, phase III dataset (ref 3). The other was an SVA element inserted 1.8kb downstream of the first coding exon in *BRCA1* at chr17:41274217. This ME was not seen in the 1000 genomes project, phase III dataset.

## References

1 Saal, L. H. *et al.* The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med* **7**, 20, doi:10.1186/s13073-015-0131-9 (2015).

2 *SCAN-B*, <http://www.med.lu.se/klinvetlund/canceromics/konsortium/scan_b> (

3 Ryden, L. *et al.* Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. *Br J Surg* **105**, e158-e168, doi:10.1002/bjs.10741 (2018).

4 *Gene Expression Omnibus*, <http://www.ncbi.nlm.nih.gov/geo/> (

5 Brueffer, C. *et al.* Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative. *JCO Precision Oncology*, 1-18, doi:10.1200/po.17.00135 (2018).

6 Jonsson, G. *et al.* The retinoblastoma gene undergoes rearrangements in BRCA1-deficient basal-like breast cancer. *Cancer research* **72**, 4028-4036, doi:10.1158/0008-5472.CAN-12-0097 (2012).

7 Holm, K. *et al.* An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Res* **18**, 27, doi:10.1186/s13058-016-0685-5 (2016).

8 Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature genetics*, doi:10.1038/ng.3934 (2017).

9 Hansmann, T. *et al.* Constitutive promoter methylation of BRCA1 and RAD51C in patients with familial ovarian cancer and early-onset sporadic breast cancer. *Human molecular genetics* **21**, 4669-4679, doi:10.1093/hmg/dds308 (2012).

10 Watanabe, Y. *et al.* Aberrant DNA methylation status of DNA repair genes in breast cancer treated with neoadjuvant chemotherapy. *Genes Cells* **18**, 1120-1130, doi:10.1111/gtc.12100 (2013).

11 Joost, P., Bendahl, P. O., Halvarsson, B., Rambech, E. & Nilbert, M. Efficient and reproducible identification of mismatch repair deficient colon cancer: validation of the MMR index and comparison with other predictive models. *BMC Clin Pathol* **13**, 33, doi:10.1186/1472-6890-13-33 (2013).

12 Paquet, E. R. & Hallett, M. T. Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype. *Journal of the National Cancer Institute* **107**, doi:10.1093/jnci/dju357 (2015).

13 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, doi:10.1172/JCI45014 (2011).

14 Chen, X. *et al.* TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. *Cancer Inform* **11**, 147-156, doi:10.4137/CIN.S9983 (2012).

15 Ali, H. R. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology* **15**, 431, doi:10.1186/s13059-014-0431-1 (2014).

16 Guedj, M. *et al.* A refined molecular taxonomy of breast cancer. *Oncogene* **31**, 1196-1206, doi:10.1038/onc.2011.301 (2012).

17 Fredlund, E. *et al.* The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. *Breast Cancer Res* **14**, R113, doi:10.1186/bcr3236 (2012).

18 *The R Project for Statistical Computing*, <http://www.r-project.org> (

19 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573, doi:btq170 [pii]
10.1093/bioinformatics/btq170 (2010).

20 Hudis, C. A. *et al.* Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. *J Clin Oncol* **25**, 2127-2132, doi:10.1200/JCO.2006.10.3523 (2007).

21 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15 10 11-15 10 18, doi:10.1002/cpbi.20 (2016).

22 Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15 17 11-12, doi:10.1002/0471250953.bi1507s52 (2015).

23 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

24 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).

25 Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature medicine* **23**, 517-525, doi:10.1038/nm.4292 (2017).

26 Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome research* **27**, 1916-1929, doi:10.1101/gr.218032.116 (2017).

27 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

28 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

29 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

**Supplementary Table S1**. **HRD frequency in subgroups of population-based TNBC for successfully analysed WGS cases**.

| | HRDetect-low (prob 0-0.2) | HRDetect-intermediate (prob 0.2-0.7) | HRDetect-high (prob >0.7) |
|---|---|---|---|
| General cohort (n=237) | 35.9% | 5.5% | 58.6% |
| Primary disease only (n=231)[A] | 35.5% | 5.2% | 59.3% |
| BRCA-status | | | |
| *BRCA1* biallelic inactivated (n=22) † | 0% | 0% | 100% |
| *BRCA1* monoallelic alteration (n=4) | 25.0% | 0% | 75.0% |
| *BRCA1* hypermethylated (n=57) † | 1.8% | 0% | 98.2% |
| *BRCA2* biallelic inactivated (n=7) † | 0% | 0% | 100% |
| *BRCA2* monoallelic alteration (n=6) † | 16.7% | 0% | 83.3% |
| Neoadjuvant therapy (n=16) | 31.2% | 6.2% | 62.5% |
| Adjuvant therapy | | | |
| Chemotherapy (n=149, OS/IDFS) † | 30.9% | 2.7% | 66.4% |
| Untreated (n=50) † | 54.0% | 10.0% | 36.0% |
| ER-staining positivity | | | |
| <1% (n=206) † | 36.9% | 5.3% | 57.8% |
| 1-10% (n=29) † | 31.0% | 6.9% | 62.1% |
| Grade | | | |
| 2 (n=27) † | 77.8% | 3.7% | 18.5% |
| 3 (n=207) † | 30.0% | 5.8% | 64.3% |
| Age | | | |
| <50 years (n=52) † | 9.6% | 1.9% | 88.5% |
| 50-70 years (n=119) † | 38.7% | 3.4% | 58.0% |
| >70 years (n=66) † | 51.5% | 12.1% | 36.4% |
| Lymph node status | | | |
| N0 (n=153) † | 37.3% | 3.9% | 58.8% |
| N+ (n=81) † | 34.6% | 8.6% | 56.8% |
| Tumor size | | | |
| ≤20mm (n=120) † | 34.2% | 2.5% | 63.3% |
| >20mm (n=117) † | 37.6% | 8.5% | 53.8% |
| PAM50 subtypes [11] | | | |
| Basal-like (n=183) † | 25.1% | 4.9% | 69.9% |
| HER2-enriched (n=31) † | 67.7% | 12.9% | 19.4% |
| Normal-like (n=22) † | 77.3% | 0.0% | 22.7% |
| TNBC molecular subtypes [36] | | | |
| Basal-like 1 (BL1, n=46) † | 15.2% | 2.2% | 82.6% |
| Basal-like 2 (BL2, n=23) | 43.5% | 17.4% | 39.1% |
| Immunomodulatory (IM, n=46) † | 30.4% | 4.3% | 65.2% |
| Luminal androgen receptor (LAR, n=30) † | 86.7% | 6.7% | 6.7% |
| Mesenchymal (M, n=41) † | 19.5% | 4.9% | 75.6% |
| Mesenchymal stem-like (MSL, n=14) † | 42.9% | 0% | 57.1% |
| IC10 [15] | | | |
| Cluster 10 (n=148) † | 21.6% | 4.7% | 73.6% |
| Cluster 9 (n=13) † | 61.5% | 0.0% | 38.5% |
| Cluster 4 (n=57) † | 64.9% | 7.0% | 28.1% |
| CIT [12] | | | |
| Basal-like (basL, n=175) † | 24.6% | 3.4% | 72.0% |
| Molecular apocrine (mApo, n=46) † | 76.1% | 10.9% | 13.0% |

Proportions calculated excluding missing data.

A: Excluding patients with metastatic disease at diagnosis or micro/macro residual disease after surgery.

†: Chi-square test for given probabilities p-value < 0.05, 2 degrees of freedom.

**Supplementary Table S2**. **Patient characteristics and clinicopathological variables of the TNBC study cohorts**.

| | Background (healthcare region) [A] | SCAN-B enrolled patients | SCAN-B WGS patients | SCAN-B WGS adjuvant chemotherapy outcome | SCAN-B WGS Untreated outcome |
|---|---|---|---|---|---|
| N | 408 | 340 | 237 | 149 | 50 |
| Cases with 30X sequence depth (%) | - | - | 69.6% | 100% | 0% |
| Cases with 15X sequence depth (%) | - | - | 30.4% | 0% | 100% |
| ER IHC staining <1% | 86% | 85% | 87.7% | 88.5% | 84.0% |
| ER IHC staining 1-10% | 14% | 15% | 12.3% | 11.5% | 16.0% |
| Tumor size ≤20mm | 49% | 48% | 50.6% | 55.0% | 38.0% |
| Tumor size >20mm | 51% | 52% | 49.4% | 45.0% | 62.0% |
| Grade 1 | 1% | 1% | 0% | 0% | 0% |
| Grade 2 | 16% | 16% | 12.1% | 7.4% | 20.0% |
| Grade 3 | 83% | 83% | 87.9% | 92.6% | 80.0% |
| Age <50 years | 23% | 21% | 21.9% | 28.9% | 0% |
| Age 50-70 years | 45% | 46% | 45.6% | 57.0% | 14.0% |
| Age >70 years | 32% | 32% | 32.5% | 14.1% | 86.0% |
| Node negative (N0) | 68% | 69% | 65.4% | 66.9% | 70.0% |
| Node positive (N+) | 32% | 31% | 34.6% | 33.1% | 30.0% |
| BRCA and RAD51C status* | | | | | |
| *BRCA1* germline carrier | - | - | 8% | 9.4% | 2.0% |
| *BRCA2* germline carrier | - | - | 3.0% | 2.7% | 2.0% |
| *BRCA1* hypermethylation | - | - | 24.1% | 28.9% | 12.0% |
| *RAD51C* hypermethylation | - | - | 2.1% | 2.7% | 0% |
| Adjuvant chemotherapy | | | | | |
| FEC alone | - | - | 8.0% | 12.8% | 0% |
| FEC + taxane | - | - | 60.8% | 83.2% | 0% |
| Other combination/Not specified** | - | 67% | 3.0% | 4.0% | 0% |
| Outcome | | | | | |
| Death events (%) | - | 24% | 26.6% | 14.8% | 40.0% |
| Relapse, all types (%) | - | - | 23.6% | 18.1% | 26.0% |
| Distant metastases (%) | - | - | 20.7% | 14.8% | 24.0% |
| PAM50 subtypes*** | | | | | |
| Basal-like | - | - | 77.2% | 84.6% | 54.0% |
| HER2-enriched | - | - | 13.1% | 7.4% | 30.0% |
| Lum A | - | - | 0% | 0% | 0% |
| Lum B | - | - | 0.4% | 0% | 2.0% |
| Normal-like | - | - | 9.3% | 8.1% | 14.0% |
| TNBC subtypes*** | | | | | |
| Basal-like 1 (BL1) | - | - | 20.3% | 22.1% | 12.2% |
| Basal-like 2 (BL2) | - | - | 10.1% | 11.4% | 4.1% |
| Immunomodulatory (IM) | - | - | 20.3% | 22.8% | 14.3% |
| Luminal androgen receptor (LAR) | - | - | 13.2% | 6.7% | 34.7% |
| Mesenchymal (M) | - | - | 18.1% | 17.4% | 18.4% |
| Mesenchymal stem-like (MSL) | - | - | 6.2% | 6.7% | 2.0% |
| Uncertain | - | - | 11.9% | 12.8% | 14.3% |

Proportions calculated excluding missing data. For all enrolled and background population numbers are presented at a general level.

[A]: Identified in the Swedish national breast cancer quality registry.

[B]: In Sweden ER-negativity is defined as ≤10% of cells with IHC-staining for ER.

*: Mutation calls based on WGS analysis.

**: For SCAN-B enrolled patients specified value indicates % of patients having planned or started chemotherapy (including neoadjuvant) treatment according to available registry data.

***: RNAseq data is not available for all enrolled SCAN-B patients, and thus not provided.