Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study

Johan Staaf, Dominik Glodzik, Ana Bosch, Johan Vallon-Christersson, Christel Reuterswärd, Jari Häkkinen, Andrea Degasperi, Tauanne Dias Amarante, Lao H. Saal, Cecilia Hegardt, Hilary Stobart, Anna Ehinger, Christer Larsson, Lisa Rydén, Niklas Loman, Martin Malmberg, Anders Kvist, Hans Ehrencrona, Helen R. Davies, Åke Borg, and Serena Nik-Zainal

Whole-genome sequencing (WGS) brings comprehensive insights to cancer genome interpretation. To explore the clinical value of WGS, we sequenced 254 triple-negative breast cancers (TNBCs) for which associated treatment and outcome data were collected between 2010 and 2015 via the population-based Sweden Cancerome Analysis Network–Breast (SCAN-B) project (ClinicalTrials.gov ID:NCT02306096). Applying the HRDetect mutational-signature-based algorithm to classify tumors, 59% were predicted to have homologous-recombination-repair deficiency (HRDetect-high): 67% explained by germline/ somatic mutations of BRCA1/BRCA2, BRCA1 promoter hypermethylation, RAD51C hypermethylation or biallelic loss of PALB2. A novel mechanism of BRCA1 abrogation was discovered via germline SINE-VNTR-Alu retrotansposition. HRDetect provided independent prognostic information, with HRDetect-high patients having better outcome on adjuvant chemotherapy for invasive disease-free survival (hazard ratio (HR) = 0.42; 95% confidence interval (CI) = 0.2–0.87) and distant relapse-free interval (HR = 0.31, CI = 0.13–0.76) compared to HRDetect-low, regardless of whether a genetic/epigenetic cause was identified. HRDetect-intermediate, some possessing potentially targetable biological abnormalities, had the poorest outcomes. HRDetect-low cancers also had inadequate outcomes: ~4.7% were mismatch-repair-deficient (another targetable defect, not typically sought) and they were enriched for (but not restricted to) PIK3CA/AKT1 pathway abnormalities. New treatment options need to be considered for now-discernible HRDetect-intermediate and HRDetect-low categories. This population-based study advocates for WGS of TNBC to better inform trial stratification and improve clinical decision-making.

Recent advances in sequencing technology have significantly reduced sequencing costs. Cancer whole-genome sequencing (WGS) is now feasible, with thousands of matched tumor-normal sequencing pairs successfully sequenced so far, revealing novel biological insights (International Cancer Genome Consortium: https://icgc.org). Yet, for WGS to become adopted clinically, systematic demonstrations of utility in well-characterized population-based studies and validation in clinical trials are required.

The Sweden Cancerome Analysis Network–Breast (SCAN-B; ClinicalTrials.gov identifier NCT02306096) is an ongoing population-based observational study, currently involving nine hospitals in the south of Sweden that serve nearly two million inhabitants (or ~20% of the Swedish population). All patients with suspected breast cancer are offered recruitment. Connection to national cancer quality registries ensures the availability of excellent clinical and outcome data. High inclusion rates of ~85% have resulted in the recruitment of >13,500 patients since 2010. Tissue samples are taken via standard clinical diagnostic pathways without special dispensation. Outcomes from research on SCAN-B patients thus reflect real-world population medicine (Extended Data 1).

To gauge the value of WGS in a clinical setting and capture sufficient outcome information, we defined a recruitment period of 1 September 2010 to 31 March 2015 in the Skåne healthcare region, which serves ~1.3 million inhabitants. We focused on an area of unmet clinical need: triple-negative breast cancer (TNBC) (estrogen-receptor (ER), progesterone-receptor (PR), human epidermal receptor growth factor 2/erythroblastic oncogene B (HER2/ERBB2)-negative), which is historically associated with poor clinical outcomes. A total of 4,665 patients were registered with invasive breast cancer in that period. Nine percent (n = 408) were TNBCs, consistent with national TNBC incidence (9% in 2015), and 340 had enrolled into SCAN-B. Clinical re-review and the availability of material for genomic/transcriptomic sequencing left 254 cases for matched tumor-normal WGS/RNaseq analysis (170 at 30-fold and 84 at 15-fold sequence depth) (Fig. 1). Among 254 patients, 2.4% had metastatic disease at diagnosis and 6.7% received neoadjuvant treatment.

Of these patients, 237 (93%) had WGS data of sufficient quality for comprehensive genomic profiling. The failure rate was...
influenced by sequencing depth: of cases sequenced to 30-fold depth, 3% failed, whereas 11% of cases failed when sequenced to 15-fold coverage. SCAN-B fresh-tissue procurement is fully integrated into routine clinical diagnostics across participating healthcare institutions. Failure rates of ~3% thus provide true estimates of 30-fold WGS success in a clinical context, without preselection for tumor cellularity. The 15-fold coverage for clinical WGS is unlikely to be adequate.

Predicted somatic driver mutations, pathogenic germline mutations and somatic mutational signatures were obtained, together with regions of copy number loss, gain and loss of heterozygosity (LOH). These genomic features in the SCAN-B TNBC cohort were comparable to a previously reported WGS cohort. To assess the additional benefits of WGS-based stratification, we applied a mutational-signature-based algorithm, HRDetect, designed to detect 'BRCA-ness or homologous-recombination-repair deficiency (HRD), using default breast-cancer-specific parameters. More than half of TNBCs (58.6%) were classified as HRDetect-high (exceeding a predefined score of 0.7, predictive of BRCA1/BRCA2-deficiency). A total of 35.9% were classified as HRDetect-low.

Fig. 1 | CONSORT diagram of the study. CONSORT diagram for patients identified during the period 1 September 2010 to 31 March 2015 in the Skåne healthcare region, with four participating SCAN-B sites: Lund, Malmö, Helsingborg and Kristianstad. NKBC, Swedish national breast cancer quality registry.
(score <0.2) and 5.5% fell within an HRD-intermediate category (score 0.2–0.7) (Fig. 2a).

To compare customary breast cancer stratification methods with HRDetect, we examined age, grade and gene expression phenotypes (for example, PAM50\(^{10}\), CIT\(^1\), IC10\(^{1}\), and TNBCTYPE\(^5\)) in this cohort. HRDetect-high classification was enriched in expected subgroups such as young patients (88.5% of women <50 years), high-grade tumors and a basal-like expression subtype (PAM50 basal-like\(^{10}\), CIT basal-like\(^1\), IC10 IntClust \(10\)\(^{2,14}\) and TNBCTYPE basal-like \(1^{1}\)) (Supplementary Table 1 and Fig. 2a). However, HRDetect-high scores were also observed in tumors with ER-staining (62.1% of cases with 1–10% ER-staining intensity), in middle-aged patients (58% of women aged 50–70 years were HRDetect-high) and older patients (>70 years, 36.4% HRDetect-high cases), as well as in tumors with non-basal-like gene expression profiles (Supplementary Table 1). Consequently, the HRD phenotype identified by HRDetect is enriched for, but not restricted to, typical basal-like tumors characteristic of young patients in TNBC. The corollary is also true—expression-based profiling (for example, PAM50) and integrative clusters\(^{2,14}\) are not adequately able to discriminate HRDetect groups, suggesting that HRDetect provides a novel, independent angle to TNBC stratification.

Previously, signature 3 and specific patterns of copy number aberrations (CNAs, ‘genomic scars’) have been used to infer an HRD phenotype\(^{14,16}\). However, choosing a signature 3 cutoff is challenging, as this signature has a featureless, flat profile, where mutations are often mis-assigned. The CNA-based ‘HRD assay’ has a designated cutoff of 42 (ref. 14). When compared with HRDetect as reference, the ‘HRD assay’ has a false negative rate of 13%, and 16% of CNA-based HRD-high cases were false positives. HRDetect is therefore more specific than current substitution signature and CNA approaches, while extending identification of HRD to a wider set of samples and revealing tumors that are likely inactivated by ways other than classical mechanisms (Fig. 2b).

Next, to comprehensively understand the causes of HRDetect-high scores, we examined germline and somatic mutation status of BRCA1, BRCA2, a set of 163 additional HR-related or breast cancer susceptibility genes (Supplementary Data Table), as well as the status of the remaining wild-type parental allele in all samples. Systematic promoter hypermethylation by pyrosequencing of BRCA1 and RAD51C was also performed.

Of 139 HRDetect-high cases, 29 (21%) had confirmed biallelic loss of BRCA1 or BRCA2 (that is, 20 germline and 9 somatic with LOH of the wild-type parental allele) and 55 (40%) had BRCA1 hypermethylation with loss of the other parental allele in ~90% of cases (Fig. 2a, note one case with concurrent BRCA2 biallelic alteration and BRCA1 hypermethylation). There were no instances of a dominantly inherited 5′ untranslated region (UTR) variant causing methylation-associated BRCA1 silencing\(^\text{17}\).

Five tumors had pathogenic germline PALB2 variants, three cases with a c.509_510delAA variant, one c.1039G>T variant. Four had wild-type allele inactivation through somatic pathogenic mutation in three cases and LOH in one instance. Five RAD51C hypermethylation cases were also observed with concomitant marked downregulation of RAD51C mRNA expression (Wilcoxon-test \(P = 0.0001\)). Intriguingly, the four biallelic PALB2 and five RAD51C cases (together comprising 6.5% of HRDetect-high cases) consistently showed a BRCA2-null phenotype\(^\text{11}\) including elevated substitution signature 3, elevated rearrangement signatures 2 and 5, and no rearrangement signature 3 (Figs. 2a and 3a–d and Supplementary Data Table), evidenced also by principal component analysis (PCA) of HRDetect components (Fig. 3e). PALB2, RAD51C and BRCA2 are involved in a complex that stimulates strand invasion of the RAD51 nucleoprotein filament\(^{16,17}\), a critical step in HR-related repair. Thus, their BRCA2-like profiles may be explained by these molecular relationships.

The causes for HRDetect-high scores in the remaining 46 (33%) samples were unclear. Six were monoallelic for BRCA1/BRCA2 and had low tumor cellularity possibly preventing a correct call of LOH for the parental allele. RAD51 and PALB2 pyrosequencing did not reveal positive findings. Mobile element analysis for 11 HR genes (ATM, BARD1, BRCA1, BRCA2, BRIP1, CHEK2, MRE11, NBN, PALB2, RAD51C, RAD51D) identified a germline SINE-VNTR-Alu (SVA) retrotransposon 1.8 kb downstream of the first coding exon of BRCA1 in one patient, PD535958a. This specific mobile element has not been reported in the 1,000 genomes dataset. Moreover, wild-type allele LOH was noted and BRCA1 expression was markedly reduced in this tumor. SVA elements have been reported as disease-causing\(^\text{18}\), but not in BRCA1. Thus, this observation may indicate a potential novel mechanism of germline BRCA1 abrogation.

Fourteen new pathogenic germline BRCA1/BRCA2 variants were identified, in addition to the 12 hitherto known variants, raising clinical genetic counselling implications. In Sweden, re-contacting patients/families based on BRCA1/BRCA2 incidental findings has been perceived positively\(^\text{12}\) supporting the added value that tumor-directed WGS brings to family counselling.

To seek distinguishing features between different HRDetect groups, we examined driver alterations. HRDetect-high and HRDetect-intermediate cases tended toward more driver amplifications (for example, MYC and MCL1) than HRDetect-low (Extended Data 3). Eight substitution/indel driver genes were enriched, albeit non-discriminatory between groups (TP53, PTEN, ARID1B, MLL4 for HRDetect-high, PIK3CA and AKTI for HRDetect-low and RB1 and FBXW7 for HRDetect-high and HRDetect-intermediate) (chi-squared test, \(P < 0.05\), Extended Data 3). Of interest, PTEN driver mutations and activating PIK3CA and AKTI mutations are differently enriched between HRDetect-high and HRDetect-low groups: 29% versus 14% for PTEN, 2.2% versus 25% for PIK3CA, and 0.7% versus 7.1% for AKTI for HRDetect-high versus HRDetect-low, respectively. Thus, PIK3CA/AKT1/PTEN pathway dysregulation is not restricted to a particular HRDetect group. This has potential implications for clinical genetic counselling and tumour-directed WGS.

**Fig. 2** | HRDetect classification and genomic characteristics in population-based TNBC. a, Bar plot of HRDetect probability obtained in 237 TNBCs together with clinical and genomic characteristics obtained from WGS and RNAseq. Annotation tracks for samples include (from top to bottom) ER immunohistochemistry (IHC) scoring, patient age, the basal-like phenotype from Prediction Analysis of Microarray 50 (PAM50) classification\(^\text{16}\) and genetic alterations in homologous recombination-associated genes (BRCA1, BRCA2, PALB2, RAD51C). Further, proportions of mutational and rearrangement signatures and indel patterns are shown as proportional bar plots. Mutations and copy number amplifications in key oncogenes and tumor suppressors are represented for individual samples. Molecular subtype proportions in HRDetect-high and HRDetect-low cases for PAM50, Cartes d’Identité des Tumeurs (CIT) subtypes\(^5\), integrative cluster 10 (IC10) and TNBCTYPE are represented by pie charts. Intermediary samples are excluded due to low numbers. CIT subtypes: mApo, molecular apocrine. IC10 subtypes\(^1\): cl (IntClust) \(10\) corresponding to basal-like tumors by other subtyping schemes. TNBCTYPE subtypes\(^5\): BL1, basal-like 1; BL2, basal-like 2; IM, immunomodulatory; M, mesenchymal; MSL, mesenchymal stem-like; LAR, luminal androgen receptor; UNS, uncertain. b, Proportions of mutational signature 3 (in tumors with >20 events) and copy number–derived HRD scores according to ref. 15 across subgroups defined first by HRDetect class (-low, -intermediate and -high), where the HRDetect-high subgroup is further divided into whether BRCA1/BRCA2 was inactivated by a germline mutation, somatic mutation or promoter hypermethylation, or no mutation was identified. Right axes in boxplots show the number of patients in each group. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.
implications for patient selection in clinical trials using PI3K-AKT-mTOR pathway agents, as mis-stratifying patients based on single mutations may affect clinical trial success.

To investigate whether HRDetect groups were simply genetic portraits of traditional transcriptional TNBC subgroups, we performed unsupervised group discovery and machine learning based
supervised classification using matched RNAseq data (Extended Data 4). Two approaches of unsupervised consensus clustering were used. Neither was able to find distinct transcriptional patterns distinguishing HRDetect-high and HRDetect-low groups (Extended Data 4a–e). Similarly, exhaustive machine learning based exploration could also not achieve high prediction accuracy for HRDetect-high and HRDetect-low subgroups (Extended Data 4f,g). Furthermore, tumor-infiltrating lymphocytes are increasingly implicated as a predictor of relapse in breast cancer. We examined CD8/CD3/CD4 and CD247 infiltration using matched transcriptomic data for all patients in this cohort. None was differentially expressed between HRDetect groups (Kruskal–Wallis, P > 0.05, Extended Data 4h). Together, this implies that the mutational signature based algorithm HRDetect captures distinctive, pathogenicomic phenotypes of TNBC that are not apparent at the bulk tissue transcriptional level.

HRDetect categorizes tumors differently to customary TNBC classifiers. We thus evaluated whether HRDetect's signature-based stratification had prognostic potential. TNBC patients that did not receive adjuvant treatment due to age and/or poor performance status were considered a particular subgroup and assessed separately (Supplementary Table 2). For these patients, there were no differences observed in overall survival (OS), invasive disease-free survival (IDFS) or distant relapse-free interval (DRFI) between HRDetect-high and HRDetect-low categories (log-rank P > 0.05).

In distinct analyses of patients that received standard-of-care adjuvant chemotherapy (ACT) (typically fluorouracil, epirubicin and cyclophosphamide (VEC) + docetaxel, Supplementary Data Table), patients with HRDetect-high tumors had significantly improved IDFS and DRFI (log-rank P = 0.009 and P = 0.01) compared to those with HRDetect-low tumors (Fig. 4a,b). An improved OS was also observed (log-rank P = 0.06, Fig. 4c), albeit non-significant, likely because of the limited follow-up interval. This suggests that TNBC patients with HRDetect-high scores have a higher degree of chemosensitivity than HRDetect-low cases and are worth identifying.

Strikingly, chemotherapy-treated HRDetect-low patients had a similar IDFS to patients that did not receive ACT, in spite of superior fitness and lower age of diagnosis (85%, ≤70 years) (Fig. 4d,e). Soberingly, this suggests that HRDetect-low cases are deriving limited benefit from the current standard of care, warranting a reappraisal of systemic therapies for this now-identifiable subgroup with poor outcome.

To test the independent prognostic value of HRDetect in patients receiving adjuvant chemotherapy, we performed multivariable Cox regression, adjusting for tumor size (≤20 mm, >20 mm), patient age (<50, ≥50 years), tumor grade (1, 2, 3) and lymph node status (N0, N+). We found that HRDetect classification among chemotherapy-treated patients provided independent prognostic information favoring a better outcome in HRDetect-high cases than in HRDetect-low cases, for IDFS (HR = 0.42, 95% CI = 0.20–0.87) and DRFI (HR = 0.31, CI = 0.13–0.76). Although not significant for OS (HR = 0.46, CI = 0.19–1.13), this implies a potential effect with longer follow-up time (~60% of the chemotherapy-treated cases had ≤5 years of follow-up).

Notably, within the HRDetect-high group, we observed no difference in patient outcomes (OS, IDFS and DRFI) between cases with confirmed BRCA1/BRCA2 loss and cases where it was not possible to confirm genetic/epigenetic abrogation of these genes (log-rank \( P = 0.79, 0.51 \) and 0.67, respectively, Fig. 4f). Lending further credence, survival analysis excluding known BRCA1/BRCA2 cases showed that HRDetect-high classification remained significantly associated with improved IDFS in remaining patients (log-rank \( P = 0.008, \) multivariable Cox regression HR = 0.41, CI = 0.19–0.91, \( P = 0.03)\). BRCA1/BRCA2 abrogated status was next added as a covariate to the multivariable Cox regression. HRDetect classification persisted as a meaningful independent positive prognostic indicator for IDFS (HR = 0.38, CI = 0.17–0.85, \( P = 0.02)\). This strengthens the argument that, even in the absence of confirmation of the genetic/epigenetic aetiology, the mutational signature-based approach is capable of predicting potential clinical benefit for ACT. Critically, it expands the number of TNBC patients eligible for treatment strategies targeting DNA repair mechanisms such as poly (ADP-ribose) polymerase (PARP) inhibitors.

We note that four HRDetect-low patients (4.7%) had mismatch-repair deficient (MMRd) tumors (Extended Data 5). One occurred in association with HRD. Despite a lack of genetic/epigenetic confirmation of MMR abrogation in these cases, independent analyses by protein IHC confirmed tumor-cell-specific loss of MLH1 and PMS2 expression in all cases (Extended Data 5). This is of particular interest, as checkpoint inhibitors have been FDA-approved for MMRd in the metastatic setting, irrespective of tumor origin, emphasizing the therapeutically valuable incidental insights that can be obtained via a WGS-based approach.

Finally, we explored the HRDetect-intermediate cohort (scores 0.2–0.7, \( n = 13, 5.5\% )\). We noted that 83.5% of HRDetect-high tumors had scores exceeding 0.99, and nearly all were BRCA1/BRCA2-null tumors. Additionally, based on mutational, indel and rearrangement signature patterns, cases with HRDetect-intermediate scores had different characteristics (Fig. 2a). We thus broadened the HRDetect-intermediate category to scores between 0.1 and 0.9 (\( n = 32)\). Driver alterations and mutational signatures for the redefined groups were examined in a new analysis (Extended Data 6). The broadened HRDetect-intermediate group had a high prevalence of driver amplifications (75%, 24/32 cases) and harbored 47% of all CCNE1 amplifications. CCNE1 was the most enriched amplification, constituting 25% of all amplifications in this new intermediate group. This is interesting, as CCNE1 is implicated in oncogene-induced replicative stress\(^2\)\(^2\)\(^5\). The intermediate category is also enriched for hypermutators of a rearrangement signature of long tandem-duplications (RS1) (where RS1 is predominant, exceeding 100 RS1 rearrangements)\(^2\)\(^5\). This pattern has been shown in prostate cancer\(^2\)\(^7\)\(^,\)\(^2\)\(^8\) and also in ovarian cancer\(^7\)\(^,\)\(^2\)\(^8\) associated with CDK12 mutations. CCNE1 and CDK12 over-expression has been shown to deregulate cell cycle progression and disrupt DNA replication during S phase in vitro\(^9\)\(^,\)\(^10\). Recently, inhibitors of
replication stress response such as Wee1 kinase inhibitors, ATR inhibitors and CHK1-inhibitors were developed to target tumors with hallmarks of replication stress31, and CCNE1 overexpression was reported to sensitize TNBCs to these compounds32. When assessing outcomes, the broadened intermediate group showed poorer IDFS, regardless of whether the patients received ACT or
not, although numbers are limited (Extended Data 6). Therefore, the HRDetect-intermediate group is a subset of tumors that are difficult to distinguish using customary genomic scar approaches or individual substitution signatures, but are important to recognize because their idiosyncratic tumor biology is a harbinger of potentially poor outcome and may be differently targetable in terms of therapeutics.

This population-based study of TNBC in a routine diagnostic setting demonstrates what can be revealed by WGS. We surmise that it is valuable to identify HRDetect status, whether high, intermediate or low. All groups are informative. Combinations of targeted-sequencing, microsatellite instability (MSI) assays and CNA approaches may be used increasingly. However, the value of holistic WGS as a single assay is reinforced when we consider three matters. First, patients may be mis-classified based on individual mutations. For example, we show that PIK3CA/AKT1/PTEN mutations identified through targeted sequencing are differentially enriched in HRDetect categories, with different survival likelihoods. Using mutations alone to stratify patients should thus be carefully considered in clinical trials. Second, it is possible to identify poor responders to the current standard of care that cannot be detected by any other method. The HRDetect-low category has many more patients than would be detected by binary PIK3CA/AKT1/PTEN targeted assays alone. These limited assays will also not identify the HRDetect-intermediate category, an interesting now-detectable subset in which to explore alternative therapeutic strategies. Third, limited sequencing assays will miss the substantial proportion of tumors with HRD signatures that do not have genetic/epigenetic drivers but are predicted to have good outcomes. Using mutations alone to stratify patients should thus be carefully considered in clinical trials. Second, it is possible to identify poor responders to the current standard of care that cannot be detected by any other method. The HRDetect-low category has many more patients than would be detected by binary PIK3CA/AKT1/PTEN targeted assays alone. These limited assays will also not identify the HRDetect-intermediate category, an interesting now-detectable subset in which to explore alternative therapeutic strategies. Third, limited sequencing assays will miss the substantial proportion of tumors with HRD signatures that do not have genetic/epigenetic drivers but are predicted to have good outcomes. In short, this study argues for WGS to improve TNBC patient stratification.

Fig. 4 | Association of HRDetect classification with clinical outcomes in an unselected population-based TNBC cohort. a–c, Kaplan-Meier analysis of association with outcome for HRDetect classification (HRDetect-high versus HRDetect-low; HRDetect-intermediate omitted due to low patient numbers) in TNBC patients treated with standard-of-care adjuvant chemotherapy with IDFS as clinical endpoint (a), DRFI as clinical endpoint (b) and OS as clinical endpoint (c). d, IDFS as clinical endpoint, showing both adjuvantly treated and untreated patients stratified by HRDetect status. e, Distribution of patient age between HRDetect-high and -low groups stratified by treatment and eligibility for IDFS analysis. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. The right axis shows the number of patients in each group. f, Kaplan–Meier analysis of association with IDFS of the HRDetect-high group, demonstrating no significant difference between subjects where BRCA alterations (mutations in BRCA1/2 or BRCA1 hypermethylation) were and were not identified. All P values in a–d and f were calculated using the log-rank test and are two-sided.
Online content Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0582-4.

Received: 22 May 2019; Accepted: 15 August 2019; Published online: 30 September 2019

References
1. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).
2. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979–993 (2012).
3. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).
4. Byrd, 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 (2018).
5. 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 (2015).
6. Haffly, B. G. et al. Lecoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. J. Clin. Oncol. 24, 5652–5657 (2006).
7. Liedtke, C. et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J. Clin. Oncol. 26, 1275–1281 (2008).
8. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
9. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. 23, 517–527 (2017).
10. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular subtype. J. Natl Cancer Inst. 107, 357 (2014).
11. Guedj, M. et al. A refined molecular taxonomy of breast cancer. Oncogene 31, 1196–1206 (2012).
12. Ali, H. R. et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol. 15, 431 (2014).
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 (2011).
14. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast cancers. Nature 486, 346–352 (2012).
15. Telli, M. L. et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. Clin. Cancer Res. 22, 3764–3773 (2016).
16. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. 49, 1476–1486 (2017).
17. Evans, D. G. R. et al. A dominantly inherited 5′ UTR variant causing methylation-associated silencing of BRCA1 as a cause of breast and ovarian cancer. Am. J. Hum. Genet. 103, 213–220 (2018).
18. Park, J. Y. et al. Breast cancer-associated missense mutants of the PALB2 WD40 domain, which directly binds RAD51C, RAD51 and BRCA2, disrupt DNA repair. Oncogene 33, 4803–4812 (2014).
19. Park, J. Y., Zhang, F. & Andreassen, P. R. PALB2: the hub of a network of tumor suppressors involved in DNA damage responses. Biochim. Biophys. Acta 1846, 263–275 (2014).
20. Qian, Y. et al. Identification of pathogenic retrotransposon insertions in cancer predisposition genes. Cancer Genet. 216–217, 159–169 (2017).
21. Nilsson, M. P. et al. Germline mutations in BRCA1 and BRCA2 incidentally revealed in a biobank research study: experiences from re-contacting mutation carriers and relatives. J. Commun. Genet. 9, 201–208 (2018).
22. Sharma, P. Biology and management of patients with triple-negative breast cancer. Oncologist 21, 1050–1062 (2016).
23. Davies, H. et al. Whole-genome sequencing reveals breast cancers with mismatch repair deficiency. Cancer Res. 77, 4755–4762 (2017).
24. Mazouzi, A., Velimezi, G. & Loizou, J. J. DNA replication stress: causes, resolution and disease. Exp. Cell Res. 329, 85–93 (2014).
25. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. Science 319, 1352–1355 (2008).
26. Glodzik, D. et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. Nat. Genet. 49, 341–348 (2017).
27. Popova, T. et al. Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. Cancer Res. 76, 1882–1891 (2016).
28. Hillman, R. T., Chisholm, G. B., Lu, K. H. & Futreal, P. A. Genomic rearrangement signatures and clinical outcomes in high-grade serous ovarian cancer. J. Natl Cancer Inst. 110, 265–272 (2018).
29. Bartkova, J. et al. Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. Nature 444, 633–637 (2006).
30. Di Micco, R. et al. Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. Nature 444, 638–642 (2006).
31. Forment, J. V. & O’Connor, M. J. Targeting the replication stress response in cancer. Pharmacol. Ther. 188, 155–167 (2018).
32. Chen, X. et al. Cyclin E overexpression sensitizes triple-negative breast cancer to weel kinase inhibition. Clin. Cancer Res. 24, 6594–6610 (2018).

Acknowledgements
We thank patients and clinicians participating in the SCAN-B study, the staff at the central SCAN-B laboratory at the Division of Oncology and Pathology, Lund University, the Swedish national breast cancer quality registry (NRKB), RCC Syd, the South Sweden Breast Cancer Group (SSBCG), the CASM IT and Wellcome Sanger Institute sequencing team for support, R. Harris for administrative, technical and coordination support, K. Lövgren and E. Rambech for technical assistance with MMRd cases and P.-O. Bendahl for statistical comments. Financial support for this study was provided by the Swedish Cancer Society (CAN 2016/659, CAN 2018/685 and Senior Investigator Award SIA190013), the Mrs Berta Kamprad Foundation (FBKS-2018-3-166 and FBKS-2018-4-146), the Crafoord Foundation (20180543), the Swedish Research Council, the Lund-Lausanne L2-Bridge/Biltema Foundation (F 2016/1330), the Mats Paulsson Foundation (JACD 2017), the Gustav V’s Jubilee Foundation (174271) and Governmental Funding of Clinical Research within the National Health Service (ALF) (2018/40612). Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0582-4. Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0582-4. Peer review information
Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.
Reprints and permissions information is available at www.nature.com/reprints. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Letters

Methods

Ethics approval and consent to participate. The SCAN-B study was approved by the Regional Ethical Review Board in Lund, Sweden (applicable registration numbers 2009/658, 2015/277, 2016/742, 2018/267 and 2019/01232 for this study). All patients provided written informed consent prior to enrolment.

Patient cohort. In Sweden, the definition of TNBC is a tumor with ≤10% of cells with IHC-staining for ER and PR (thus including tumors with 1–10% stained cells) and an IHC HER2-staining score of <2, or for patients with IHC 2+, a non-amplified ISH status. During the period 1 September 2010 to 31 March 2015, 408 patients were diagnosed with TNBC (localized or advanced disease with specified treatment status) in the Skåne healthcare region in southern Sweden, based on data from the Swedish national breast cancer quality registry, NKBC (Fig. 1). In total, 340 of these patients were enrolled in the SCAN-B study4,5 (ClinicalTrials.gov ID NCT02306096), a prospective, observational, population-based cohort study, from which 254 patients with concurrent RNAseq were selected for extensive clinical review and WGS. Of the reviewed cases, 153 (60%) patients were eligible for OS/IDFS survival analysis after standard-of-care adjuvant chemotherapy (FEC-based (combination of 5-fluorouracil, epirubicin and cyclophosphamide) ± a taxane in 96% of cases) according to national guidelines. Of these (irrespective of clinical outcome status), 41% had ≥5 years of follow-up, 25% 4–5 years, 31% 2–4 years and 4% <2 years of follow-up. Of the 153 patients, 148 (97%) were eligible for relapse analysis, of which 20% developed a relapse of some type (loco-regional or distant). Remaining cases received either neoadjuvant treatment, no adjuvant treatment or were not treated in an adjuvant context (for example, metastatic disease at diagnosis). As part of routine oncogenetic clinical screening, 49 of 254 recruited patients were previously screened for pathogenic germline variants in BRCA1 and BRCA2, with 12 positive findings (nine BRCA1 and three BRCA2 carriers). Patient cohort characteristics, enrolled SCAN-B patients, WGS analyzed SCAN-B patients and WGS analyzed SCAN-B treatment subsets are described in Supplementary Table 2. Individual patient characteristics are provided in the Supplementary Data Table.

Tissue sampling, DNA and RNA extraction. Fresh tumor samples preserved in RNAlater (Qiagen) were obtained in conjunction with routine clinical sampling by a diagnostic pathologist in regional pathology departments7. RNA and DNA were extracted using the Qiagen Allprep extraction kit (Qiagen) as described in ref. 8. DNA from whole blood was extracted by the Labmeden Skåne Biobank.

Whole-genome sequencing. WGS of TNBCs were performed using Illumina sequencing technology to achieve average coverage of 15–30-fold depth as previously described in matched tumor-normal samples8. Each patient was sequenced only once. Patients that received ACT were primarily selected for 30-fold coverage, whereas untreated patients were sequenced to 15-fold depth. WGS data quality, basic data analysis, variant calling, mobile element analysis and HRD classification by the HRDetect algorithm were performed as outlined in refs. 9 and the Supplementary Information. HRDetect classification was verified for known BRCA1/BRCA2-deficient cases versus tumor cellularity by WGS or pathology estimation for both 30-fold and 15-fold sequencing depth (Extended Data Fig. 7) to demonstrate that there was no systematic bias as a result of sequencing coverage or tumor cellularity.

DNA promoter methylation analysis. DNA promoter hypermethylation analysis of bisulfite-treated DNA for specific CpG promoter sites in BRCA1, RAD51C, RAD51 and PALB2 was performed as described in the Supplementary Information.

Gene expression analyses. Gene expression profiling of all TNBCs was performed using RNA sequencing as described in ref. 1, and data have been reported elsewhere13, Molecular subtype classification according to PAM50 (by AIMS)10, IC1012, CIT11 and reported TNBC subtypes (TNBCtype)13, unsupervised clustering and machine learning based supervised classification were performed as described in the Supplementary Information.

Statistical analyses. Survival analyses were performed in R (version 3.3.0) using the survival package with OS, IDFS or DRFI as endpoints, defined according to the STEEP criteria15 (see Supplementary Information for endpoint definitions and analysis exclusion criteria). Survival curves were compared using Kaplan–Meier estimates and the log-rank test. Hazard ratios were calculated through univariable or multivariable Cox regression using the coxph R function. Harrell’s C-index was computed using the dpred R package. Statistical comparisons between groups were performed using Wilcoxon’s or Kruskal–Wallis tests for numerical values or the chi-squared test for ordinal values. All P values reported from statistical tests are two-sided if not otherwise specified.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Somatic mutational data are available at https://data.mendeley.com/datasets/2mn4ctdpkp/1.

Raw sequence data may be obtained by contacting the Swedish corresponding authors with a request that is compliant with Swedish regulations on data protection, ethical permissions and patient consent.

References

33. 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 Prec. Oncol. 2, 1–18 (2018).
34. Chen, X. et al. TNBCtype: a subtyping tool for triple-negative breast cancer. Cancer Inf. 11, 147–156 (2012).
35. Hudis, C. A. et al. Proposal for standardized definitions for efficacy endpoints in adenocancer breast cancer trials: the STEEP system. J. Clin. Oncol. 25, 2127–2132 (2007).
Extended Data Fig. 1 | Sweden Cancerome Analysis Network – Breast (SCAN-B). In the Skåne healthcare region (Region Skåne) four main hospitals are participating in the SCAN-B study: Lund, Malmö, Helsingborg and Kristianstad. a, SCAN-B overall enrolment rate at all participating hospitals, including Skåne healthcare region, during the period 1 September 2010 to 31 March 2015, corresponding to the same time period from which the TNBC cases in the current study were selected. The statistics are restricted to the seven hospitals where enrolment was operational from the start in 2010. b, Overall accrual rate per quarter of a year (Q1–Q4) for the SCAN-B study since the start in 2010 Q4 up until 2018 Q1. The red line corresponds to the cumulative number of enrolled patients, reaching nearly 12,000 in 2018 Q1. c, Illustration of the population-based nature of the SCAN-B study for primary resectable breast cancer. Based on data from the national breast cancer quality registry in Sweden (NKBC), a background population of primary resectable breast cancers from the entire SCAN-B catchment region during the period 1 September 2010 to 31 March 2015 was identified (same time period from which the TNBC cases in the current study were selected), comprising 8,587 patients. Of these 8,587 patients, 5,417 were enrolled in SCAN-B, with 3,520 patients having RNA sequencing data passing basic quality criteria. The lower panels demonstrate the clinicopathological characteristics of the different subgroups in the consort diagram, demonstrating the representativity of the end RNA sequencing cohort compared to all enrolled SCAN-B patients and the total patient population in the catchment region. Note that the RNA sequencing cohort has a slightly lower inclusion of smaller tumors, due to the fact that the SCAN-B tissue sampling is performed by a pathologist after enough tissue has been secured for routine diagnostics. 

d, Demonstration of the year to year representativity of molecular subtypes in breast cancer (PAM50, top panel) and administered treatments based on data from the NKBC (lower panel) for patients identified in c. The bars show patients in the RNA sequencing cohort from c, stratified by year of diagnosis (all patients diagnosed in a particular year are included). PAM50 subtyping was performed using the AIMS method (Paquet et al. J. Natl Cancer Inst. 107, 357, 2014) (as for the TNBC cases in the current study) as this classifier is a single sample classifier that does not rely on a mean centering of gene expression data across a cohort (thus is not sensitive to, for example, potential bias in year to year inclusion). ACT, adjuvant chemotherapy.
Extended Data Fig. 2 | Similar genomic characteristics of SCAN-B TNBC cases compared to previously reported WGS-analysed TNBCs. 

**a**, Comparison of copy number alterations (CNA) as defined by Nik-Zainal et al. *(Nature, 534, 47-54, 2016)* in the 237 SCAN-B TNBC cases versus 162 TNBC cases from Nik-Zainal et al. A frequency below 0 means frequency of copy number loss. 

**b**, Comparison of frequency of LOH defined as in Nik-Zainal et al. *(Nature, 2016)* between the same SCAN-B cases and Nik-Zainal et al. *(Nature, 2016)* TNBC cases. 

**c**, Comparison of copy number neutral (cnn) LOH defined as in Nik-Zainal et al. *(Nature, 2016)* between the same set of samples. 

**d**, Comparison of the frequency of driver gene amplifications between the same set of samples. Only amplifications matched in both cohorts are displayed. Driver gene list was obtained from Nik-Zainal et al. *(Nature, 2016)*. 

**e**, Comparison of the frequency of homozygous deletions based on ASCAT data, as described in Nik-Zainal et al. *(Nature, 2016)*, for the same set of samples. Only deletions matched in both cohorts are shown. 

**f**, Frequency of somatic substitutions and indels for driver genes from Nik-Zainal et al. *(Nature, 2016)* in the two cohorts. Only genes with >1% mutation frequency in Nik-Zainal et al. are displayed. 

**g**, Exposure to mutation substitution signatures as defined in Nik-Zainal et al. *(Nature, 2016)* for the same set of samples. The line corresponds to a 1:1 relationship. 

**h**, Exposure to rearrangement signatures (RS1-RS6) as defined in Nik-Zainal et al. *(Nature, 2016)* for the same set of samples. The line corresponds to a 1:1 relationship.
Extended Data Fig. 3 | see figure caption on next page.
Extended Data Fig. 3 | Clinicopathological and genomic characteristics of HRDetect groups. a, Expression of checkpoint proliferation (left), steroid (center) and basal (right) metagenes from Fredlund et al. (Breast Cancer Research, 2012) across HRDetect groups stratified by BRCA status. HRDetect-inter, intermediate subgroup; BRCA1pm, BRCA1 promoter hypermethylated; BRCAgerm, BRCA1/2 germline carriers; BRCAson, BRCA1/2 somatic cases. b, Distribution of patient age (left), Ki67 staining (%) center) and clinical grade (right) across the same groups (same set of patient numbers). c, Distribution of number of detected substitutions (left), indels (center) and rearrangements (right) for the same groups limited to cases with 30-fold sequence coverage. Two-sided P values were calculated using the Kruskal–Wallis test. d, Frequency of the genome altered by copy number gain and loss (CN-FGA, left), LOH (LOH-FGA, center) and copy number neutral LOH (cnnLOH-FGA, right) defined as in Nik-Zainal et al. (Nature, 2016). e, Frequency of copy number gain (above zero centerline) and copy number loss across the genome for HRDetect-high tumors versus HRDetect-low tumors defined as in Nik-Zainal et al. (Nature, 2016). HRDetect-intermediate tumors were omitted due to small numbers. f, Frequency of amplification of driver genes from Nik-Zainal et al. (Nature, 2016) across HRDetect groups (left) and putative homozygous deletions (HD) called using ASCAT (right) as defined in Nik-Zainal et al. g, Comparison of somatic mutation frequency (substitutions, indels and curated rearrangements) for driver genes from Nik-Zainal et al. (Nature, 2016) versus HRDetect groups. Two-sided P values calculated using the chi-squared test. h, Violin plot of the distribution of rearrangement signature (RS) proportions per sample defined in Nik-Zainal et al. (Nature, 2016) versus HRDetect groups for patients with at least 20 called rearrangements. Violin plot line elements: center line, median; thick limits, upper and lower quartiles; whiskers, 1.5x interquartile range. In all boxplots, the top axis shows the number of patients in each group. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. Kruskal, Kruskal–Wallis test; ChiSq, chi-squares test. All calculated P values are two-sided.
Extended Data Fig. 4 | see figure caption on next page.
Extended Data Fig. 4 | Unsupervised and supervised gene expression analyses versus HRDetect groups. In all analyses, raw expression data (FPKM) have been offset by the addition of +0.1, followed by log, transformation before further analyses. Only RefSeq annotated genes were used. A total of 232 cases with gene expression were included in all analyses. In all consensus cluster analyses, clustering was performed using Pearson correlation and ward.d2 linkage, with 2,000 repetitions using the R ConsensusClusterPlus package. For PCA analyses pitem = 0.8 and pFeature = 0.98 were used in the consensus cluster function. For non-PCA analyses, the corresponding values were 0.8 and 0.8. a, Consensus clustering of PCA components from PCA analysis of 19,102 genes using a two-group solution. The heatmap to the left shows consensus, with blue indicating that samples often cluster together across repetitions (rows = samples; columns). Bars to the right show the proportion of HRDetect groups in different consensus clusters according to the legend. PCA captures all variation in the data in the different principal components on which clustering was performed. b, Same as in a, but for a three-group consensus solution. c, Same as in a but for a four-group solution. d, Consensus clustering performed on 16,364 genes with mean-centered log2 data as input (that is, no PCA). HRDetect-high implies probabilities >0.7, HRDetect-low probabilities <0.2, that is, according to the main text definitions. Heatmaps show the percentage of samples for a group in respective consensus clusters (x axis), across different cluster solutions (y axis). For example, for HRDetect-high cases (left heatmap) using a k = 2 solution, >70% of these tumors are located in cluster 1, together with 40–70% of HRDetect-low samples (as seen in the right heatmap). e, Same visualization as in d, but now for 6,776 genes with a standard deviation of >0.6 in expression across samples. f, Supervised prediction of HRDetect-high (P > 0.7) and HRDetect-low (P < 0.2) according to main text definitions based on the top 10,000 varying RefSeq genes across all 232 cases using seven different types of machine learning method. FPKM values were offset by +0.1 and then log2 transformed. The 10,000 most varying genes across all relevant cases were selected. For each method, cases were divided into training (70% of cohort) and test (30%), balanced for age, lymph node status and grade. HRDetect-intermediate cases were omitted. Training and test cohorts were individually mean-centered. ROC was used as the optimization metric and fourfold cross-validation was repeated 10 times for training using the training cohort. The optimized model was applied to the test set. The entire procedure was repeated 10 times through an outer loop, with a different division of samples in the training and test set in each loop to assure that sample selection was not skewing results. For each model this generated, for example, 10 ROC metrics as each outer loop iteration created a (potentially) new model. The summarized results are shown to the left. For all methods, bar height corresponds to the average metric across the 10 iterations with one standard deviation range shown in red and individual values in orange. All analyses were performed using the Caret R-package using the classifier names indicated in the plot and with the tuneLength variable set to 10. g, The same analysis as in f, but instead using PCA components as input data for machine learning. PCA components were derived originally in a to capture all variation in the data and are now used as input for supervised prediction using the same set-up and parameters as in f, h, Gene expression (log2(FPKM + offset)) of prototypical immunomarkers versus HRDetect groups. Two-sided P values were calculated using the Kruskal–Wallis test. s.d., standard deviation. In all boxplots the top axis shows the number of patients in each group. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.
Extended Data Fig. 5 | MMRd SCAN-B tumors. Note, unlike in colorectal cancer, mismatch repair deficient (MMRd) tumors are also able to carry signs of chromosomal or genomic instability as seen in PD31144a (BRCA1 promoter hypermethylated case) and PD31040a. Thus the mutational processes driving these two features are not mutually exclusive in breast cancer.
Extended Data Fig. 6 | see figure caption on next page.
Extended Data Fig. 6 | Characteristics of expanded HRDetect-intermediate cases. a. Comparison of driver amplifications from Nik-Zainal et al. (Nature, 534, 47–54, 2016) between HRDetect groups defined from a broadened intermediate group (0.1–0.9 in HRDetect score). HRDetect (0.9–1) = 127 cases; HRDetect (0.1–0.9) = 32 cases; HRDetect (0–0.1) = 78 cases. b. Comparison of somatic driver mutations (substitutions, indels) for driver genes defined in Nik-Zainal et al. (Nature, 2016). For the specific set of genes curated for rearrangements in Nik-Zainal et al. (for example, RB1 and PTEN) these are included as events in the analysis (that is, RB1 includes both mutations and rearrangements). c. Distribution of mutational signature exposure for signatures S3 (e.3) and S5 (e.5) defined in Nik-Zainal et al. (Nature, 2016), and a copy number–based HRD score defined by Telli et al. (Clin. Cancer Res., 22, 3764–3773, 2016) (originally based on SNP arrays, ‘genomic scars’) across HRDetect subgroups defined by a broadened intermediate group. In all boxplots the top axis shows the number of patients in each group. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. d, Distribution of total number of detected substitutions, indels and rearrangements for 30-fold sequenced cases across HRDetect subgroups defined by a broadened intermediate group. In all boxplots, the top axis shows the number of patients in each group. Boxplot elements: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. e, Outcome analysis for original HRDetect-groups (left panels) and new division with a broadened HRDetect-intermediate group (right panels) stratified by treatment status using invasive disease-free survival (IDFS) as clinical endpoint. The top two panels show IDFS for patients receiving adjuvant chemotherapy (ACT) and the bottom two panels show IDFS for untreated patients according to division by HRDetect score. Log-rank P values are two-sided. g, Distribution of different molecular subtypes in the broadened HRDetect-intermediate group based on 232 cases with gene expression data. mApo, molecular apocrine; BL1, basal-like 1; BL 2, basal-like 2; IM, immunomodulatory; M, mesenchymal; MSL, mesenchymal stem-like; LAR, luminal androgen receptor; UNS, uncertain.
Extended Data Fig. 7 | Tumor cellularity versus HRDetect probability scores and characteristic rearrangement signature proportions for BRCA1-null (biallelic alteration or promoter hypermethylation) and BRCA2-null (biallelic alterations) tumors. a, HRDetect probabilities versus WGS estimated tumor cell content based on the ASCAT algorithm (n = 84 cases). b, HRDetect probabilities versus a pathological assessment of the invasive cancer proportion from a section adjacent to the extracted tumor piece (n = 67 cases). Tumors are further stratified by their intended sequencing depth (30-fold or 15-fold) in a and b. c, Proportions of rearrangement signature 3 (Nik-Zainal et al., Nature, 534, 47–54, 2016) for BRCA1-null cases. Two of the 237 cases do not have a value for the signature. d, Proportions of rearrangement signature 5 for BRCA2-null cases. One outlier exists, corresponding to a tumor with concurrent BRCA1 hypermethylation that has a genetic phenotype very similar to a BRCA1-null tumor rather than a BRCA2-null tumor. Two of the 237 cases do not have a value for the signature. In all boxplots the top axis shows the number of patients in each group. Box-plot elements: center line, median; box limits, upper and lower quartiles; whiskers, =1.5x interquartile range.
**Statistical parameters**

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

*Our web collection on statistics for biologists may be useful.*

**Software and code**

**Policy information about availability of computer code**

**Data collection**

| Whole genome sequencing data from Illumina X10 machines generated 150 base pair paired-end reads. Paired-end reads were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner, BWA (v0.7.15) to generate raw sequencing data. |
| DNA methylation data from pyrosequencing was analyzed using the PyroMark software included with the pyrosequencing instrument (Qiagen). |
| Collection and basic processing of RNA sequencing data from SCAN-B has been previously described (Saal et al. Genome Medicine 2015). |

**Data analysis**

| All statistical tests generated two-sided p-values if not otherwise specified. Statistical comparisons between groups were performed using Wilcoxon’s or Kruskal-Wallis tests for numerical values, or the Chi-square test for ordinal values. |
| Whole genome sequencing: |
| Identification of somatic mutations (see references for actual software used): |
| 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 |
| - Pindel (2.1.0, 2.2.4) for identification of somatic small insertions and deletions |
| https://github.com/cancerit/cgpPindel |
Upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

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

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. The ME reference sequences provided with MELT were used for discovery of three major classes of Mobile element (ME) insertions not present in the GRCh37 reference genome were called and genotyped using the Mobile Element Locator Tool (MELT) analysis.

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. Supporting aligned read data for ME insertions discovered in or near the normal and tumor samples were visually inspected in Integrative Genomics Viewer (IGV).

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. 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.

Survival analyses:
Survival analyses were performed in R (ver 3.3.0) using the survival package with overall survival (OS), invasive disease-free survival (IDFS), or distant relapse-free interval (DRFI), as endpoints defined according with the STEEP criteria (see Supplementary Information for endpoint definitions and analysis exclusion criteria). Survival curves were compared using Kaplan-Meier estimates and the log-rank test. Hazard ratios were calculated through univariable or multivariable Cox regression using the coxph R function.

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 guidelines20, as the time from diagnosis to either death of any cause or invasive breast-cancer related events (local-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 Cox regression 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.

Mobile element (ME) 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. 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. 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).

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. 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.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data will be deposited in the European Genome-phenome Archive (EGA) based on compliance with Swedish and EU law and patient data protection rules. Until this has been accomplished raw sequence data may be obtained by contacting corresponding Swedish authors based on request that is compliant with Swedish regulations on data protection, study ethical permissions, and patient consent.

Field-specific reporting

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

Life sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size

No specific sample size calculation (power calculation) was performed before the study, as the expected group proportions and actual effect size with respect to, e.g., treatment outcome for the whole genome sequencing based analysis (HRDetect) were unknown in a general TNBC study population prior to this study. The study had a population-based angle meaning that as many patients as possible during a defined inclusion period in a specific Swedish healthcare region was eligible for sequencing, after a stepwise filtering as outlined in the CONSORT diagram and Supplementary Methods. The final number of sequenced patients were restricted by: a) available and completed chart review, b) access to matched RNA sequencing data generated from the ongoing SCAN-B study, c) sufficient tumor DNA, and d) cost for whole genome sequencing. Notably, 254 of 340 enrolled SCAN-B patients were whole genome sequenced, irrespective of their tumor cellularity (i.e., no pre-selection of tumor tissue with a specific lower tumor cell content by a pathologist were performed). Importantly, not all 340 enrolled patients were available for sequencing due to: a) no tumor tissue sampled by the diagnostic pathologist (e.g. a very small tumor), and b) no available RNA sequencing data (insufficient RNA extracted or degraded). In the end, approximately 62% of ALL TNBC patients in the catchment region, and approximately 75% of all enrolled SCAN-B patients during the specified inclusion time period were whole genome sequenced.

Data exclusions

Initial definition of the patient study population (TNBC with information about treatment status) was based on data from the national Swedish cancer registry (NKBC through the INCA platform) provided to the SCAN-B study, forming a population-based collection of TNBC patients during a specific time period in a single healthcare region. Next, patients enrolled in the SCAN-B study underwent an independent chart review by a senior breast cancer oncologist to assure that registry data was in line with patient chart data and to acquire detailed follow-up and treatment data. Patients passing these filters had their tumor DNA, collected through the SCAN-B study, pre-screened for genotype consistency with matched blood to assure that no mismatches existed (for instance introduced through handling at the two different tissue biobanks [blood and tissue] involved in the study). The pre-screening was performed at the Wellcome Trust Sanger Institute, UK, sequencing facility. 254 patients were finally selected for whole genome sequencing restricted by sequencing cost and that patients should also have available complementary RNA sequencing data from the SCAN-B study. After quality control of whole genome sequencing data, 237 cases remained. Exclusion criteria for whole genome sequencing data were:
- the library preparation failed to yield sufficient output.
- for samples with sequence coverage below 20x, the Battenberg algorithm failed to find a copy-number solution.
- for samples with sequence coverage over 20x, the Battenberg algorithm failed to find a copy-number solution and the the number of somatic point mutations identified was less than 500.

The above exclusion criteria was used to ensure that samples with insufficient tumor cellularity, yielding insecure predictions by HRDetect, were excluded.

Replication

Replication in the form of repeated whole genome sequencing of the same DNA from a patient was not performed. All patients (tumor - normal blood sample pair) were analyzed once by whole genome sequencing. If sequencing failed, the patient was excluded. Thus, no technical replicates exist.

Randomization

Randomization was not applicable for this study. No pre-determined groups to test genetic/genomic differences between existed. Groups were formed based on data driven analysis (the HRDetect algorithm) of whole genome sequencing data for which correlative analyses and survival analyses were then performed.

Blinding

Investigators were blinded to patient identity during the entire study. The experimental analyses generating the study subgroup populations (performed by the HRDetect data algorithm) did not have access to clinicopathological data from patients at any point. Only after the data driven study subpopulations had been formed (using the pre-specified parameters and cut-off for a HRDetect-high classification defined in e.g. the original HRDetect study by Davies and Gol dizk et al.) were correlative analyses of HRDetect classification versus for instance clinicopathological and outcome data performed.
Breast cancer patients were recruited to the ongoing observational population-based SCAN-B study (ClinicalTrials.gov ID NCT02306096) currently active at 9 participating hospitals and covering about 25% of the national Swedish population. SCAN-B invites all patients with confirmed or suspected breast cancer to be included (women or men, young or old) after written informed consent. Study enrollment and tumor sampling is integrated into routine clinical management and performed by healthcare professionals (e.g. breast cancer surgeons and breast cancer pathologists) at the participating hospitals. Based on the high inclusion rate (>85% per year since start 2010) the SCAN-B study represents a truly population-based cohort. Currently, the written informed consent is in Swedish only. All study subpopulations were compared to the original cancer population in the catchment region during the inclusion period to demonstrate the population-based aspect of the study.

Patients are enrolled after written informed consent in a population-based manner, inviting all patients based on the discretion of the treating clinician and ability to understand the written informed consent. The SCAN-B study has been approved by the Regional Ethical Review Board in Lund, Sweden (applicable registration numbers: 2009/658, 2009/659, 2012/58, 2014/521, 2016/742, and 2018/267, and 2019/01252).