Whole genome sequencing reveals differences in the genetic landscape of interval and screen-detected breast cancer

Richard Houlston
richard.houlston@icr.ac.uk

The Institute of Cancer Research https://orcid.org/0000-0002-5268-0242

Charlie Mills
The Institute of Cancer Research https://orcid.org/0000-0002-4755-7549

Amit Sud
Dana-Farber Cancer Institute https://orcid.org/0000-0002-6133-0164

Andrew Everall
The Institute of Cancer Research

Daniel Chubb
Institute of Cancer Research

Samuel Lawrence
The Institute of Cancer Research

Ben Kinnersley
UCL Cancer Institute https://orcid.org/0000-0003-1783-6296

Alex Cornish
The Institute of Cancer Research

Robert Bentham
University College London

Article

Keywords: Genetics, breast cancer, screening, mammography

Posted Date: February 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3955962/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Additional Declarations: (Not answered)

Version of Record: A version of this preprint was published at npj Precision Oncology on May 28th, 2024. See the published version at https://doi.org/10.1038/s41698-024-00618-6.
Abstract

Interval breast cancers (IBCs) are cancers diagnosed between screening episodes. Understanding the biological differences between IBCs and screen-detected breast-cancers (SDBCs) has the potential to improve mammographic screening and patient management. We analysed and compared the genomic landscape of 288 IBCs and 473 SDBCs by whole genome sequencing of paired tumour-normal patient samples collected as part of the UK 100,000 Genomes Project. Compared to SDBCs, IBCs were more likely to be lobular, higher grade, and triple negative. A more aggressive clinical phenotype was reflected in IBCs displaying features of genomic instability including a higher mutation rate and number of chromosomal structural abnormalities, defective homologous recombination and $TP53$ mutations. We did not however, find evidence to indicate that IBCs are associated with a different immune response. While IBCs do not represent a unique molecular class of invasive breast cancer they exhibit a more aggressive phenotype, which is likely to be a consequence of the timing of tumour initiation. This information is relevant both with respect to treatment as well as defining the screening interval for mammography.

BACKGROUND

The high survival rates associated with the early detection of localised breast cancer together with the screening properties of mammography, have provided the rationale for population breast cancer screening programs\textsuperscript{1–3}. In the UK, the National Health Service (NHS) invites women aged between 50 and 70 for 3-yearly mammography, but with provision for those aged 45 and over to be screened if self-referred. Although studies suggest a mammography every 2–3 years can reduce breast cancer-specific mortality by a fifth\textsuperscript{4} around 30% of breast cancers in women attending screening are interval breast cancers (IBCs) diagnosed between screening episodes\textsuperscript{5}.

While technical and patient factors may in part explain the incidence of IBCs\textsuperscript{6,7}, studies have suggested IBCs may represent a more aggressive form of breast cancer\textsuperscript{8}. This would accord with the observation that most IBCs are not visible on the index screen in retrospective review\textsuperscript{8}. Exploring possible genetic differences between IBCs and screen-detected breast cancers (SDBCs) has the potential to inform mammographic screening and patient management. Previous analyses of IBCs have so far predominantly focused on specific genes and studies have typically been based on small patient numbers\textsuperscript{9}.

To provide the most comprehensive analysis of the genomic landscape of IBC, we analysed whole genome sequencing (WGS) data generated on 288 IBCs recruited to the UK’s 100,000 Genomes Project (100kGP). The results of our study highlight distinct clinical, genetic and evolutionary differences between IBC and SDBC (Fig. 1).

METHODS
Patients

The 100kGP project (release v14) provides WGS data on tumour-normal pairs from breast cancer patients recruited through 13 Genomic Medicine Centres across England. We restricted our WGS analysis to samples with high-quality data from PCR-free, flash-frozen fresh tumour samples (Supplementary Methods). Demographic and clinical data were obtained from Public Health England’s National Cancer Registration and Analysis Service (PHE-NCRAS), NHS Digital and the Genomic Medicine Centres. Tumour pathology information was obtained from histology reports. After QC (Supplementary Table 1), 833 of the 1,488 patients with a diagnosis of breast cancer had a documented NHS screening history. To minimise bias associated with screening indication, we confined our analysis to breast cancers diagnosed through the NHS breast screening programme of women aged between 47 and 70 years (Table 1). Considering hormone receptor status, we confined our analysis to ductal histology since no patients with triple negative (TN) tumours had lobular breast cancer.
Clinical characteristics of IBCs and SDBCs. *U* = unknown status. *P*-values were evaluated using a t-test for age and tumour size, a Cochran-Armitage trend test for grade, and Fisher’s exact test for all other variables.

<table>
<thead>
<tr>
<th></th>
<th>Interval breast cancer</th>
<th>Screen detected breast cancer</th>
<th><em>P</em>-value</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>288</td>
<td>473</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ductal</td>
<td>230 (79.9%)</td>
<td>409 (86.5%)</td>
<td>0.019</td>
<td>0.62 (0.41, 0.94)</td>
</tr>
<tr>
<td>Lobular</td>
<td>58 (20.1%)</td>
<td>64 (13.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (years)</td>
<td>60</td>
<td>60</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Range (years)</td>
<td>47–70</td>
<td>47–70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumour size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (mm)</td>
<td>24</td>
<td>20</td>
<td>1.9 x 10^-5</td>
<td></td>
</tr>
<tr>
<td>Range (mm)</td>
<td>9–95</td>
<td>3.5–115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laterality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>136 (47.2%)</td>
<td>245 (51.8%)</td>
<td>0.23</td>
<td>0.83 (0.61, 1.13)</td>
</tr>
<tr>
<td>Left</td>
<td>152 (52.8%)</td>
<td>228 (48.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>20 (6.9%)</td>
<td>64 (13.5%)</td>
<td>3.8 x 10^-7</td>
<td>1.92 (1.51–2.47)</td>
</tr>
<tr>
<td>G2</td>
<td>140 (48.6%)</td>
<td>284 (60.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>127 (44.1%)</td>
<td>123 (26.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GX</td>
<td>&lt; 5</td>
<td>&lt; 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>193 (67.0%)</td>
<td>356 (75.3%)</td>
<td>1.11 x 10^-3</td>
<td>0.43 (0.26, 0.73)</td>
</tr>
<tr>
<td>Negative</td>
<td>40 (13.9%)</td>
<td>32 (6.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>55 (19.1%)</td>
<td>85 (17.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>106 (36.8%)</td>
<td>222 (46.9%)</td>
<td>4.45 x 10^-5</td>
<td>0.40 (0.25, 0.63)</td>
</tr>
<tr>
<td>Negative</td>
<td>60 (20.8%)</td>
<td>49 (10.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>122 (42.4%)</td>
<td>202 (42.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HER2 status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>26 (9.0%)</td>
<td>32 (6.8%)</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Borderline</td>
<td>26 (9.0%)</td>
<td>55 (11.6%)</td>
<td>1.44 (0.80, 2.57)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>211 (73.3%)</td>
<td>366 (77.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>25 (8.7%)</td>
<td>20 (4.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interval breast cancer</td>
<td>Screen detected breast cancer</td>
<td>P-value</td>
<td>Odds ratio (95% CI)</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------</td>
<td>------------------------------</td>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Triple Negative</td>
<td>22 (7.6%)</td>
<td>20 (4.2%)</td>
<td>0.03</td>
<td>2.01 (1.02, 3.98)</td>
</tr>
</tbody>
</table>

**Calling of somatic variants**

In addition to using variant calls from the 100kGP analysis pipeline we: (i) removed alignment bias introduced by ISAAC soft clipping of semi-aligned reads; (ii) called tumour copy number using Battenberg; (iii) called structural variants (SVs) from a consensus of Manta, LUMPY, and DELLY; (iv) removed insertions and deletions within 10bps of a common germline indel (Supplementary Methods).

**Driver gene identification**

To compare the frequency and clonality of driver gene mutations in IBCs and SDBCs we considered all driver genes identified across the 100kGP cohort of 1,488 invasive primary breast cancers. We identified cancer driver genes in these 1,488 cases using IntOGen, which combines seven computational methods to detect signals of positive mutational selection of missense mutations in coding regions of the genome (Supplementary Methods). Details of pre-processing of mutations, combining driver gene identification methodologies, post-processing and annotation of driver gene mutations are provided in Supplementary Methods.

**Signature analysis**

*De novo* extraction of single-base-substitution (SBS), doublet-base-substitution (DBS), insertion and deletion (ID) signatures, copy number (CN) signatures and structural variant (SV) signatures including decomposition to known COSMIC signatures (v3.3), was performed using SigProfilerExtractor. We used the extraction performed by Everall et al. and complemented this analysis by screening for evidence of homologous recombination deficiency (dHR) in tumours using HRDetect.

**Estimation of clonality**

The clonal state of driver mutations were estimated by MutationTimeR. Sample odds ratios (ORs) were calculated for both early/late and clonal/subclonal driver mutations with associated *P*-values calculated using Fisher's exact test. The ratio of subclonal to clonal mutations was used as a proxy for intratumor heterogeneity.

**Immune evasion**
Neoantigens were predicted using pVAC-Seq\textsuperscript{22} to predict the binding affinities of epitopes that arise as a result of non-synonymous mutations, based on the HLA-alleles typed by POLYSOLVER\textsuperscript{23}. We investigated three immune escape mechanisms, specifically: (i) a non-synonymous mutation in any of the three (HLA-A,-B,-C) HLA Class I genes; (ii) loss of heterozygosity (LOH) in any of the three HLA-I genes or (iii) any inactivating mutation in one of the 22 antigen-presenting genes (APGs) involved in the IFN-γ pathway, the PF-L1 receptor, the CD58 receptor, and epigenetic escape via \textit{SETDB1} (\textit{APLNR, B2M, CANX, CALR, CD274, CD58, CIITA, ERAP1, ERAP2, IFR2, IFNGR1, IFNGR2, JAK1, JAK2, NLRC5, PDIA3, RFX5, SETDB1, STAT1, TAPBP, TAP1, TAP2})\textsuperscript{24,25}. We considered a tumour sample with positive immune escape status if it exhibited any one of (i)-(iii). HLA gene mutations were found using POLYSOLVER and LOH at HLA was predicted using LOHHLA\textsuperscript{26}. We estimated numbers of tumour infiltrating lymphocytes based on the somatic copy number in conjunction with read depth of the T-cell receptor-α gene\textsuperscript{27}.

**Germline variants and polygenic scores**

We examined the germline of the patients for pathogenic variants in any of the 14 well-established hereditary breast ovarian cancer (HBOC) susceptibility genes (\textit{ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, MSH6, PALB2, PTEN, RAD51C, RAD51D, STK11 and TP53}). Assignment of pathogenicity was based on a CADD\textsuperscript{28} score > 30 and ClinVar\textsuperscript{29}. Genomic positions of canonical gene transcripts were retrieved from the Ensembl database (EnsDb.Hsapiens.v86)\textsuperscript{30} and were referenced to build GRCh38.

To generate polygenic scores (PGS) we used genome-wide association studies (GWAS) summary statistics estimated in European populations for breast cancer risk reported by Mavaddat \textit{et al}\textsuperscript{31}. For the well-established modifiable risk factors for breast cancer, we used results from the following resources: GSCAN consortium meta-analysis of smoking initiation (ever vs never status)\textsuperscript{32} UK biobank (UKBB) meta-analysis of body mass index (BMI)\textsuperscript{33}, summary statistics relating to breast density reported by Chen \textit{et al}\textsuperscript{34}, and those relating to diabetes, such as fasting glucose and fasting insulin, were obtained from UKBB studies\textsuperscript{35}.

**Statistical analysis**

The relationship between categorical variables was assessed using either Chi-square or Fisher exact tests. Linear and logistic regression were used in analysis of continuous traits. The Kolmogorov–Smirnov test was performed to compare cumulative differences in PGS profiles. All statistical analysis we performed using R Version 4.2 and we considered a two-sided $P$-value < 0.05 as statistically significant.

**RESULTS**

**Clinico-pathological differences**
After curation we based our analysis on 288 IBCs and 473 SDBCs. The clinical details of the 761 cases are summarised in Table 1. Overall, 70.4% (n = 548) of the cases were self-reported to be White British and there was no difference in the frequency of ethnicities between IBCs and SDBCs. Interval breast cancers were more likely to be larger tumours than SDBCs (median size 24mm and 20mm respectively, \( P = 1.9 \times 10^{-5} \)), be lobular (OR = 1.61, 95% confidence interval (CI) 1.07–2.42, \( P = 0.02 \)), be higher grade tumours (OR = 1.83, 95% CI 1.45–2.31, \( P = 3.8 \times 10^{-7} \)), be oestrogen receptor (ER) negative (OR = 0.43, 95% CI 0.26–0.73, \( P = 1.11 \times 10^{-3} \)), be progesterone receptor (PR) negative (OR = 0.40, 95% CI 0.25–0.65, \( P = 4.45 \times 10^{-5} \)), be TN (OR = 2.01, 95% CI: 1.02–3.98, \( P = 0.03 \)) and patients were more likely to be diagnosed with nodal disease at presentation (\( P = 0.046 \)) (Table 1).

Patients with lobular IBCs were more likely to be older at diagnosis than those with SDBCs (mean ages 61.1 years and 58.5 years respectively; \( P = 0.03 \)). While patients with ductal TN IBCs were older than those with SDBCs (mean age 61.1 years and 56.6 years respectively; \( P = 0.03 \)), patients with ductal ER-positive breast IBCs cancers tended to be younger than those with SDBC with the same histology (mean ages 58.2 years and 59.8 years respectively; \( P = 0.02 \)).

### Germline variation and polygenic scores

We identified 17 patients with IBCs who were carriers of a pathogenic mutation in one of the HBOC genes (13 BRCA2, 2 CHEK2, 1 MSH6 and 1 BRCA1) and 27 patients with SDBC who were HBOC gene carriers (12 BRCA2, 7 ATM, 3 CHEK2, 1 BARD1, 1 BRIP1, 1 BRCA1, 1 MSH6, 1 PALB2) (Supplementary Table 2). While the frequency of patients with BRCA-gene mutations was marginally, but not statistically significantly, higher in IBCs compared to SBDCs (OR = 1.80, 95% 0.77–4.23, \( P = 0.16 \)). Considering all HBOC genes, there was also no significant difference in carrier frequency (5.9% versus 5.7%).

We generated polygenic scores to investigate the relationship between genetically predicted BMI and breast density as well as modifiable breast cancer risk factors in IBC and SDBC. In a case-only analysis, while SDBCs had a higher PGS for overall breast cancer than IBC, this was not significant (\( P = 0.68 \)). Similarly, there was no difference in genetically predicted breast density or BMI between IBCs and SDBCs (\( P \)-values 0.09 and 0.65 respectively; Supplementary Table 3).

### Somatic alterations

The frequency of somatic mutations (single nucleotide variants and indels) was significantly higher in all IBCs compared to all SDBCs (1.47/Mb (0.18–18.55) and 1.06/Mb (0.08–19.05) respectively, \( P = 1.43 \times 10^{-3} \)). This significant difference was also shown stratified by tumour subtype - all ductal (\( P = 6.32 \times 10^{-4} \)), ductal ER-positive (\( P = 5.23 \times 10^{-3} \)) and ductal non-TN tumours (\( P = 1.81 \times 10^{-3} \)) (Fig. 2). After adjustment for tumour grade the mutational burden remained significantly higher in all IBCs compared to all SDBCs (\( P = 0.03 \)).

In the 67 putative invasive breast cancer drivers identified in the 100kGP data, we identified 539 predicted oncogenic mutations in IBCs and 931 predicted oncogenic mutations in SDBCs\(^{36} \). As expected
from previously published studies of breast cancer in both IBCs and SDBCs tumours the most commonly mutated breast cancer driver genes included PIK3CA, TP53, KMT2C, GATA3, CDH1, MAP3K1 and PTEN (Supplementary Table 4). Furthermore, mutational frequencies in drivers differed significantly between ductal and lobular breast cancer histologies irrespective of screen detection status (Supplementary Fig. 1). There were, however, significant differences in the frequency of mutations in these driver genes between all IBCs and all SDBCs and when tumours were stratified by histological subtype (Fig. 3).

The frequency of TP53 mutations was higher in IBCs compared with SDBCs (36% vs 24%, OR = 1.80, 95% CI: 1.29–2.51, P = 3.64 x 10\(^{-4}\)), while the frequency of PIK3CA (33% vs 45%, OR: 0.59, 95% CI: 0.43–0.80, P = 6.11 x 10\(^{-4}\)) and MAP3K1 mutations was lower (IBC 8% vs SDBC 17%, OR = 0.43, 95% CI: 0.25–0.71, P = 4.34 x 10\(^{-4}\)) (Fig. 3). These differences in mutational frequency remained significant when the analysis was confined to ductal histology and non-TN breast cancers. No significant difference in the frequency of mutations in oncogenic driver genes was shown between lobular IBC and SDBC, likely a consequence of limited statistical power. When considering the clonality of driver mutations, only TP53 mutations were significantly more likely to be clonal in IBCs (P = 0.04, OR = 3.43, 95% CI 1.02–14.96).

While the profile of somatic copy number alterations was broadly consistent between all IBCs and SDBCs (Fig. 4), the overall frequency was significantly higher (P = 2.56 x 10\(^{-4}\)). This difference remained significant when restricted to ductal (P = 9 x 10\(^{-6}\)) and ER-positive (P = 4.67 x 10\(^{-3}\)) subtypes (Fig. 5). Whole genome duplication (WGD) was also a significantly more frequent event for all (P = 8.51 x 10\(^{-3}\), OR = 1.59, 95% CI 1.11–2.27), ductal (P = 3.14 x 10\(^{-4}\), OR = 1.98, 95% CI 1.35–2.91) and non-TN IBCs (P = 0.01, OR = 1.76, 95% CI 1.11–2.79) compared to corresponding SDBCs. However, after adjusting for grade, the fraction of genome altered was no longer significantly different between IBCs and SDBCs (P = 0.27), reflecting the observation that IBCs are more likely to be higher grade.

Structural variants were a feature of almost all cancers (99% IBC vs 98% SDBC) and classification of SVs revealed that both balanced (P = 5.87 x 10\(^{-3}\), OR = 1.74, 95% CI 1.16–2.63) and unbalanced (P = 8.79 x 10\(^{-3}\), OR = 1.49, 95% CI 1.10–2.02) translocation events were enriched in IBCs. However, an adjustment for tumour grade greatly attenuated the effect such that only unbalanced translocations remained significantly enriched in IBCs (P = 0.03, OR = 1.03, 95% CI 1.00-1.06).

**Mutational signatures**

To examine if the mutational processes in IBC and SDBC are different we utilised signature activities that have been previously extracted de novo and related to known COSMIC signatures (Supplementary Table 5). Across all cancers the most common signatures were those resulting from clock-like mutagenic processes (SBS1 and SBS5). As expected, signatures indicative of dHR were strongly associated with all TN breast cancers. Across both IBC and SDBC the aetiological basis of dHR was only identifiable in 16% cases on the basis of biallelic inactivation of BRCA1, BRCA2, PALB2, BRIP1 or RAD51B through germline and somatic mutations. Hence a majority of the remaining cases may be
caused by promoter methylation; however, this data is not available for 100kGP samples. dHR was significantly more common in IBC (OR = 2.55, 95% CI: 1.33–4.98, \( P = 3.38 \times 10^{-3} \)) and this association was primarily driven by ductal ER+ tumours (OR = 4.90, 95% CI: 1.79–14.81, \( P = 6.95 \times 10^{-4} \)). Additionally, signatures associated with dHR had significantly higher activities in IBC compared to SDBC: SBS3 (\( P = 0.02 \)), ID6 (\( P = 1.79 \times 10^{-4} \)) and CN17 (\( P = 2.65 \times 10^{-3} \)).

**Immune evasion**

We predicted the presence of 12,785 neoantigens across both IBCs and SDBCs. There was no significant difference in neoantigen burden between IBCs and SDBCs (\( P = 0.26 \)). The frequency of somatic mutations and LOH of HLA class I genes was not significantly different between IBCs and SDBCs. There was also no evidence to support preferential inactivation of an APG in IBC. Finally, there was no significant difference in the TCRA T-cell fraction between IBCs and SDBCs (\( P = 0.83 \)).

**DISCUSSION**

In the 100kGP cohort, IBCs were characterised by more aggressive tumour characteristics than SDBCs - a higher histological grade, larger tumour size, more frequently lymph node-positive at diagnosis, a higher proportion of ER and PR negativity, and more likely to be TN. These observations are consistent with several published studies\(^7,38\)–\(^55\). While not statistically significant, in keeping with a recent study, IBC patients were more likely to be carriers of BRCA-mutations than those with SDBC (4.5% vs 2.5%)\(^56\).

Adjusting for grade, IBCs displayed a higher mutation rate while copy number abnormalities were not significantly different. IBCs additionally showed salient differences in driver mutational profile notably with a lower frequency of PIK3CA mutations and higher frequency of TP53 mutations. These \( TP53 \) mutations in IBCs were more likely to be clonal, supporting the notion that IBCs have differing evolutionary histories when compared to SDBCs\(^11,21\). The IBCs were also more likely to display dHR, which in part may reflect a higher rate of germline \( BRCA \) mutations. This is in keeping with published studies reporting a higher rate of germline HBOC predisposition in IBC compared to SDBC, as well as elevated rates of interval cancers in HBOC mutation carriers\(^9,56\)–\(^62\). We did not, however, find evidence to indicate that women with IBC are at a lower PGS-defined breast cancer risk which has been reported by another analysis\(^53\). Similarly, while it has been proposed that IBC may be associated with a different immune response to SDBC\(^64\), our analysis of genetically mediated immune evasion or T-cell tumour infiltration provides no support for such an assertion.

While a number of our findings are consistent with a recently published study on IBCs from the Lifepool cohort\(^9\), there are a number of notable differences. In a smaller cohort of 46 IBCs and 64 SDBCs, except for dHR, no difference in somatic aberrations between IBCs and SDBCs was found. Additionally, Cheasley et al. report cancers in low mammographic density breasts (\( n = 142 \)) had a higher frequency of \( TP53 \) mutations, dHR, higher fraction of the genome altered, more copy number gains and were more likely to be interval breast cancers when compared to cancers in high mammographic density breasts (\( n \)
Whilst sample sizes and case ascertainment may explain this discrepancy, the IBCs in 100kGP appear to have molecular features common to the breast cancers in low mammographic density breasts in Lifepool. Since mammographic sensitivity for breast cancer detection increases with decreasing breast density, this provides further evidence that the breast cancers in the 100kGP represent rapidly growing breast cancers rather than breast cancers missed at screening due to technical factors.

The major strengths of our study are having access to a unique dataset with high-quality WGS data, linkage to high quality clinicopathological information features on patients screened in the context of population-wide screening, thus avoiding biases related to screening indication. Furthermore, rather than focusing on a restricted set of genes we have been able to undertake a systematic analysis of the genetic landscape of IBCs. This has allowed us to assess tumour intrinsic mechanisms contributing to IBC as well as surrogates for the tumour microenvironment (TME). We do however, acknowledge a number of limitations, notably the lack of expression data and other classifiers of breast cancer histology such as the Gallen subtypes or PAM50. Moreover, we made use of indirect measurements of the TME and are unable to assess the impact of immune dysfunction in the earliest stages of tumour development. Finally, in assessing the role of breast density and other modifiable factors we have relied on PGS, which inevitably affords limited power to demonstrate a relationship since these only capture a small proportion of the phenotypic variation of each risk factor.

As well as informing on the biology of IBCs, the findings of our study may inform on the detection and management of breast cancer. The higher number of TP53 and copy number alterations as well as an increased mutation rate and dHR, suggests genomic instability has a greater role in IBCs when compared to SDBC. Genomic instability may be exploited therapeutically through synthetic lethality, such as sensitivity to PARP in the context of BRCA-deficiency. Furthermore, given 30% of breast cancers in women regularly attending screening are IBCs, additional screening modalities have been advocated as a means to advance the early detection of breast cancer. Whilst early in development, noninvasive detection of chromosomal instability in plasma circulating cell-free DNA has shown promise and may serve as an adjunct to breast screening mammography.

In conclusion our findings indicate that while IBCs may not represent a distinct molecular subtype of breast cancer they are characterised by a more aggressive phenotype, in part likely to be a consequence of the timing of tumour initiation. As well as being relevant to informing patient management our findings provide a strong rationale for the UK to transition from three-yearly mammography to a program based on biannual screening, which has commonly been adopted by many countries.

Declarations

ACKNOWLEDGEMENTS

Funding was provided by the Wellcome Trust (214388), Cancer Research UK (C1298/A8362) and the Medical Research Council. A.S. is in receipt of a National Institute for Health Research (NIHR) Academic
Clinical Lectureship, funding from the Royal Marsden Biomedical Research Centre, a starter grant for clinical lecturers from the Academy of Medical Sciences and is a Wellcome Trust Early Career Award recipient. This is a summary of independent research supported by the NIHR Biomedical Research Centre at the Royal Marsden NHS Foundation Trust and the Institute of Cancer Research.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

AUTHOR CONTRIBUTIONS

A.S., D.C. and R.S.H. designed the study. A.S., A.J.C. and D.C. performed sample curation, C.M., A.S., A.E., D.C., S.L., B.K., A.J.C. and R.B. performed bioinformatic and statistical analysis. C.M. A.S., A.E. and R.S.H. drafted the manuscript; all authors reviewed, read, and approved the final manuscript.

CONFLICTS OF INTERESTS

The authors declare no competing financial interests.

DATA AVAILABILITY

The data supporting the findings of this study are available within the Genomics England Research Environment, a secure cloud workspace. An example for details on how to access data for this publication can be found at https://re-docs.genomicsengland.co.uk/pan_cancer_pub/. Additional processed aggregated data supporting the findings presented in this manuscript can be found in the Supplementary Tables. To access genomic and clinical data within this Research Environment, researchers must first apply to become a member of either the Genomics England Research Network (https://www.genomicsengland.co.uk/research/academic) or the Discovery Forum (industry partners https://www.genomicsengland.co.uk/research/research-environment). The process for joining the network is described at https://www.genomicsengland.co.uk/research/academic/join-gecip and consists of the following steps:

1. Your institution will need to sign a participation agreement available athttps://files.genomicsengland.co.uk/documents/Genomics-England-GeCIP-Participation-Agreement-v2.0.pdf and email the signed version to gecip-help@genomicsengland.co.uk.

2. Once you have confirmed your institution is registered and have found a domain of interest, you can apply through the online form athttps://www.genomicsengland.co.uk/research/academic/join-
geeip. Once your Research Portal account is created you will be able to login and track your application.

3. Your application will be reviewed within 10 working days.

4. Your institution will validate your affiliation.

5. You will complete online Information Governance training and will be granted access to the Research Environment within 2 days of passing the online training.

Data that has been made available to registered users include: alignments in BAM or CRAM format, annotated variant calls in VCF format, signatures assignment, tumour mutation burden, sequencing quality metrics, summary of findings that is shared with Genomic Lab Hubs, secondary clinical data as described in this paper. Further details of the types of data available (for example, mortality, hospital episode statistics and treatment data) can be found at https://re-docs.genomicsengland.co.uk/data_overview/. Germline variants can be explored in Interactive Variant Analysis Browser (see description at https://re-docs.genomicsengland.co.uk/iva_variant/). Cancer patients cohort and longitudinal clinical information on treatment and mortality can be explored with Participant Explorer (see description at https://re-docs.genomicsengland.co.uk/pxa/).

**Funding**

was provided by the Wellcome Trust (214388), Cancer Research UK (C1298/A8362) and the Medical Research Council. A.S. is in receipt of a National Institute for Health Research (NIHR) Academic Clinical Lectureship, funding from the Royal Marsden Biomedical Research Centre, a starter grant for clinical lecturers from the Academy of Medical Sciences and is a Wellcome Trust Early Career Award recipient. This is a summary of independent research supported by the NIHR Biomedical Research Centre at the Royal Marsden NHS Foundation Trust and the Institute of Cancer Research.

**References**


**Figures**

![Diagram](image-url)

**Figure 1**
**Study overview.** WGS, whole genome sequencing. Created with BioRender.

**Figure 2**

**Mutational burden in IBC and SDBC.** The mutational burden distribution for IBCs and SDBCs in selected cohorts. The median of the distribution is indicated by the horizontal line and $P$-values calculated using a Wilcoxon rank sum test.
Figure 3

Frequency and clonality of driver gene mutations in IBC and SDBC. The mutational frequency of driver genes for: (a) all IBCs and SDBCs; (b) ductal IBCs and SDBCs (c) lobular IBCs and SDBCs. The colour scale corresponds to the total fraction of mutations that are clonal. ** significant difference in frequency imposing a Bonferroni adjusted $P$-value of $7.46 \times 10^{-4}$, * a significant difference at an unadjusted $P$-value of 0.05.
Figure 4

Copy number profile of IBC and SDBC. (a) IBC, (b) SDBC. Loss of heterozygosity is indicated in light blue, while other copy number losses are displayed in dark blue. Copy number gains are shown in red while large amplifications are indicated in orange. The regions with focal alterations, as defined by GISTIC2, are annotated with grey bars.
Figure 5

Proportion of the genome altered in IBC and SDBC. The distribution of the fraction of genome altered for IBCs and SDBCs. A region of the genome is considered to be altered if there is any deviation from a single copy of the major and minor allele. The median of the distribution is indicated by the horizontal line and $p$ values are calculated using a Wilcoxon rank sum test.

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMethods.docx
- SupplementaryTables.xlsx
- Supplementaryfigures.docx