A somatic mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers

Authors

Dominik Glodzik[1], Sandro Morganella[1], Helen Davies[1], Peter T. Simpson[2], Yilong Li[1], Xueqing Zou[1], Javier Diez-Perez[1], Johan Staaf[3], Ludmil B. Alexandrov[1,4,5], Marcel Smid[6], Arie B Brinkman[7], Inga Hansine Rye[8,9], Hege Russnes[8,9], Keiran Raine[1], Colin A. Purdie[10], Sunil R Lakhani[2,11], Alastair M. Thompson[10,12], Ewan Birney[13], Hendrik G Stunnenberg[6], Marc J van de Vijver[14], John W.M. Martens[6], Anne-Lise Børresen-Dale[8,9], Andrea L. Richardson[15,16], Gu Kong[17], Alain Viari[18,19], Douglas Easton[20], Gerard Evan[21], Peter J Campbell[1], Michael R. Stratton[1] and Serena Nik-Zainal[1,22]

Affiliations

1 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom
2 University of Queensland: Centre for Clinical Research and School of Medicine, Brisbane, Australia
3 Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
4 Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America
5 Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America
6 Erasmus MC Cancer Institute and Cancer Genomics Netherlands, Erasmus University Medical Center, Department of Medical Oncology, Rotterdam, The Netherlands
7 Radboud University, Department of Molecular Biology, Faculties of Science and Medicine, Nijmegen, Netherlands
8 Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital The Norwegian Radiumhospital
9 K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo, Norway
10 Department of Pathology, Ninewells Hospital & Medical School, Dundee DD1 9SY, UK
11 Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane, Australia
12 Department of Breast Surgical Oncology, University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, Texas 77030

13 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD
14 Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands
15 Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115 USA
16 Dana-Farber Cancer Institute, Boston, MA 02215 USA
17 Department of Pathology, College of Medicine, Hanyang University, Seoul, South Korea
18 Equipe Erable, INRIA Grenoble-Rhône-Alpes, 655, Av. de l'Europe, 38330 Montbonnot-Saint Martin, France
19 Synergie Lyon Cancer, Centre Léon Bérard, 28 rue Laënnec, Lyon Cedex 08, France
20 Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, United Kingdom.
21 Department of Biochemistry, University of Cambridge CB2 1GA, United Kingdom.
22 East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 9NB, UK

Corresponding authors:

Serena Nik-Zainal (snz@sanger.ac.uk)

Summary

Somatic rearrangements contribute to the mutagenized landscape of cancer genomes. Here, we systematically interrogated rearrangements in 560 breast cancers using a piecewise-constant fitting approach. We highlight 33 hotspots of large (>100kb) tandem duplications, a mutational signature associated with Homologous Recombinational repair deficiency. Remarkably, these tandem duplication hotspots are enriched for breast cancer germline susceptibility loci (OR 4.28) and breast-specific "super-enhancer" regulatory elements (OR 3.54). They could represent sites of selective susceptibility to double-strand break damage due to high transcriptional activity, or through incrementally increasing copy number, represent sites of secondary selective pressure. Transcriptomic consequences range from strong individual oncogene effects through to weak but quantifiable multigene expression effects. We thus present a somatic rearrangement mutational process exerting its influence through coding sequences and non-coding regulatory elements, contributing a continuum of driver consequences, from modest through to strong effects, supporting a polygenic model of cancer development.

Introduction

Whole genome sequencing (WGS) has permitted unrestricted access to the human cancer genome, triggering the hunt for driver mutations that could confer selective advantage in all parts of human DNA. Recurrent somatic mutations in coding sequences are often interpreted as driver mutations particularly when supported by transcriptomic changes or functional evidence. However, recurrent somatic mutations in non-coding sequences are less straightforward to interpret. Although *TERT* promoter mutations in malignant melanoma[1,2] and *NOTCH1* 3' region mutations in chronic lymphocytic leukaemia[3] have been successfully demonstrated as driver mutations, multiple non-coding loci have been

highlighted as recurrently mutated but evidence supporting these as true drivers remains lacking. Indeed, in a recent exploration of 560 breast cancer whole genomes[4], the largest cohort of WGS cancers to date, statistically significant recurrently mutated non-coding sites (by substitutions and insertions/deletions (indels)) were identified but alternative explanations for localized elevation in mutability such as a propensity to form secondary DNA structures were observed[4].

These efforts have been focused on recurrent substitutions and indels and an exercise seeking sites that are recurrently mutated through rearrangements has not been formally performed. Such sites could be indicative of driver loci under selective pressure (such as amplifications of *ERBB2* and *CCND1*) or could represent highly mutable sites that are simply prone to double-strand break (DSB) damage. Sites that are under selective pressure generally have a high incidence in a particular tissue-type, are highly complex and comprise multiple classes of rearrangement including deletions, inversions, tandem duplications and translocations. By contrast, sites that are simply breakable may show a low frequency of occurrence and demonstrate a preponderance of a particular class of rearrangement, a harbinger of susceptibility to a specific mutational process.

An anecdotal observation in the cohort of 560 breast cancers was of sites in the genome that appeared to be rearranged recurrently, albeit at a low frequency, and by a very specific rearrangement class of tandem duplications. Rarely, tandem duplications recurred at approximately the same locus in the same cancer resulting in the appearance of nested tandem duplications. No explanation was provided for this observation. Here, we have taken a novel approach to systematically seek sites in the human cancer genome that are recurrently mutagenized by rearrangements, specifically tandem duplications, in order to fully understand the prevalence and the impact of these sites of recurrent tandem duplications in this cohort of breast cancers.

In all, 77,695 rearrangements including 59,900 intra-chromosomal (17,564 deletions, 18,463 inversions and 23,873 tandem duplications) and 17,795 inter-

chromosomal translocations were identified in this cohort previously. The distribution of rearrangements within each cancer was complex (Figure 1A-D); some had few rearrangements without distinctive patterns, some had collections of focally occurring rearrangements such as amplifications, whereas many had rearrangements distributed throughout the genome - indicative of very different set of underpinning mutational processes.

Thus, large, focal collections of "clustered" rearrangements were first separated from rearrangements that were widely distributed or "dispersed" in each cancer, then distinguished by class (inversion, deletion, tandem duplication or translocation) and size (1-10kb, 10-100kb, 100kb-1Mb, 1-10Mb, more than 10Mb)[4], before a mathematical method for extracting mutational signatures was applied[5]. Six rearrangement signatures were extracted (RS1-RS6) representing discrete rearrangement mutational processes in breast cancer[4]. Two distinctive mutational processes in particular were associated with dispersed tandem duplications. RS1 and RS3 are mostly characterized by large (>100kb) and small (< 10kb) tandem duplications, respectively (Figure 1E). Although both are associated with tumors that are deficient in homologous recombination (HR) repair[4,6-9], RS3 is specifically associated with inactivation of *BRCA1*. Thus, because they represent distinct biological defects in human cells, we have chosen to proceed with a systematic analysis of sites of recurrent mutagenesis of these two mutational signatures as independent processes.

Previously, tumors have been described as having a large degree of genomic instability[10,11] and even a tandem duplicator phenotype[12-14] but did not have the resolution to distinguish different tandem duplication signatures. Here, we show the importance of taking a mutational signatures approach, highlighting differences in behavior between short (<10kb) and long (>100kb) tandem duplications.

We identified a surprising number of rearrangement hotspots dominated by the RS1 mutational process characterized by long (>100kb) tandem duplications[4]. Intuitively, a hotspot of mutagenesis that is enriched for a particular mutational

signature implies a propensity to DNA double-strand break (DSB) damage and specific recombination-based repair mutational mechanisms that could explain these tandem duplication hotspots. However, we find additional intriguing features associated with these hotspots that challenge current perceptions in cancer biology, explained below.

Results

Identification of rearrangement hotspots

In order to systematically identify hotspots of tandem duplications through the genome, we first considered the background distribution of rearrangements that is known to be non-uniform. A regression analysis was performed to detect and quantify the associations between the distribution of rearrangements and a variety of genomic landmarks including replication time domains, gene-rich regions, background copy number, chromatin state and repetitive sequences (Online Methods Section 3 and Supplementary Figure S1). The associations learned were taken into consideration creating an adjusted background model and were also applied during simulations, these steps being critical to the following phase of hotspot detection. Adjusted background models and simulated distributions were calculated for RS1 and RS3 tandem duplication signatures separately because of vastly differing numbers of rearrangements in each signature of 5,944 and 13,498 respectively, which could bias the detection of hotspots for the different signatures.

We next employed the principle of intermutation distance[15] (IMD)- the distance from one breakpoint to the one immediately preceding it in the reference genome and used a piece-wise constant fitting (PCF) approach[16,17], a method of segmentation of sequential data that is frequently utilized in analyses of copy number data. PCF was applied to the IMD of RS1 and RS3 separately, seeking segments of the breast cancer genomes where groups of rearrangements exhibited short IMD, indicative of "hotspots" that are more frequently

rearranged than the adjusted background model (Figure 2, Supplementary Materials). The parameters used for the PCF algorithm were optimized against simulated data (Online Methods Section 5 and Supplementary Figure S2). We aimed to detect a conservative number of hotspots while minimising the number of false positive hotspots. Note that all highly clustered rearrangements such as those causing driver amplicons had been previously identified in each sample and removed, and thus do not contribute to these hotspots. However, to ensure that a hotspot did not comprise only a few samples with multiple breakpoints each, a minimum of eight samples was required to contribute to each hotspot. Of note, this method negates the use of genomic bins and permits detection of hotspots of varying genomic size.

Thus, the PCF method was applied to RS1 and RS3 rearrangements separately, seeking loci that have a rearrangement density exceeding twice the local adjusted background density for each signature and involving a minimum of eight samples. Interestingly, 0.5% of 13,498 short RS3 tandem duplications contributed towards four RS3 hotspots. By contrast, 10% of 5,944 long RS1 tandem duplications formed 33 hotspots demonstrating that long RS1 tandem duplications are 20 times more likely to form a rearrangement hotspot than short RS3 tandem duplications. Indeed, these were visible as punctuated collections of rearrangements in genome-wide plots of rearrangement breakpoints (Figure 2C and Supplementary Table S1).

Contrasting RS3 hotspots to RS1 hotspots

RS3 hotspots had different characteristics to that of RS1 hotspots. The four RS3 hotspots were highly focused, occurred in small genomic windows and exhibited very high rearrangement densities (range 61.8 to 658.3 breakpoints per Mb (Figure 3B). In contrast, the 33 RS1 hotspots had densities between 7.6 and 83.2 breakpoints per Mb and demonstrated other striking characteristics (Figure 3A). In several RS1 hotspots, duplicated segments showed genomic overlap between patients, even when most patients had only one tandem duplication, as depicted in a cumulative plot of duplicated segments for samples contributing

rearrangements to a hotspot (Figure 3C, Supplementary Figure S3). Interestingly, the nested tandem duplications that were observed incidentally in the past[4], were a particular characteristic of RS1 hotspots. The hotspots of RS1 and RS3 were distinct from one another apart from one locus where two lncRNAs *NEAT1* and *MALAT1* reside (discussed in Section 3 of Supplementary Note).

Assessing the potential genomic consequences of RS1 and RS3 tandem duplications on functional components of the genome[12], RS1 rearrangements were observed to duplicate important driver genes and regulatory elements while RS3 rearrangements were found to mainly transect them (Figure 4, Online Methods Section 8, Supplementary Table S2). This is likely to be related to the size of tandem duplications in these signatures. Short (<10kb) RS3 tandem duplications are more likely to duplicate very small regions, with the effect equivalent of disrupting genes or regulatory elements. In contrast, RS1 tandem duplications are long (>100kb), and would be more likely to duplicate whole genes or regulatory elements.

Strikingly, the effects were strongest for tandem duplications that contributed to hotspots of RS1 and RS3 than they were for tandem duplications that were not in hotspots or that were simulated. Thus, although the likelihood of transection/duplication may be governed by the size of tandem duplications, the particular enrichment for hotspots must carry important biological implications.

The enrichment of disruption of tumor suppressor genes by RS3 hotspots (OR 167, $P=9.4 \times 10^{-41}$ by Fisher's exact test) and is relatively simple to understand - these are likely to be under selective pressure. Accordingly, two of the four RS3 hotspots occurred within well-known tumor suppressors, *PTEN* and *RB1*. Other rearrangement classes are also enriched in these genes in-keeping with being driver events (Online Methods Section 7, Supplementary Table S3). Furthermore, these sites were identified as putative driver loci in an independent analysis seeking driver rearrangements through gene-based methods[4].

By contrast, the enrichment of oncogene duplication by RS1 hotspots (OR 1.49, $P=4.1 \times 10^{-3}$ by Fisher's exact test) was apparent[12], although not as strong as the enrichment of transections of cancer genes by RS3 hotspots. More notably, the enrichment of other putative regulatory features was also observed. Indeed, we observed that susceptibility loci associated with breast cancer[18,19] were 4.28 times more frequent in an RS1 hotspot than in the rest of the tandem duplicated genome ($P=3.4 \times 10^{-4}$ in Poisson test, Supplementary Figure S4A, Supplementary Figure S5, Supplementary Figure S6). Additionally, 18 of 33 (54.5%) RS1 tandem duplication hotspots contained at least one breast super-enhancer. The density of breast super-enhancers was 3.54 times higher in a hotspot compared to the rest of the tandem duplicated genome ($P=7.0 \times 10^{-16}$ in Poisson test, Supplementary Figure S4B, Supplementary Figure S5, Supplementary Figure S6). This effect was much stronger than for non-breast tissue super-enhancers (OR 1.62) or enhancers in general (OR 1.02, Supplementary Table S4). This gradient reinforces how the relationship between tandem duplication hotspots and regulatory elements deemed as super-enhancer, is tissue-specific.

The reason underlying these observations in RS1 hotspots however is a little less clear. Single or nested tandem duplications in RS1 hotspots effectively increase the number of copies of a genomic region but only incrementally. The enrichment of breast cancer specific susceptibility loci, super-enhancers and oncogenes at hotspots of a very particular mutational signature could reflect an increased likelihood of damage and thus susceptibility to a passenger mutational signature that occurs because of the high transcriptional activity associated with such regions. However, it is also intriguing to consider that the resulting copy number increase could confer some more modest selective advantage and contribute to the driver landscape. To investigate the latter possibility, we explored the impact of RS1 tandem duplications on gene expression.

Impact of RS1 hotspots on expression

Several RS1 hotspots involved validated breast cancer genes[12] (e.g. *ESR1*, *ZNF217,* Supplementary Figures S7,S8) and could conceivably contribute to the driver landscape through increasing the number of copies of a gene - even if by only a single copy.

ESR1 is an example of a breast cancer gene that is a target of an RS1 hotspot. In the vicinity of *ESR1* is a breast tissue specific super-enhancer and a breast cancer susceptibility locus. Fourteen samples contribute to this hotspot, of which ten have only a single tandem duplication or simple nested tandem duplications of this site. Six samples had expression data and all showed significantly elevated levels of *ESR1* despite modest copy number increase (Supplementary Figure S7a). Four samples have a small number of rearrangements (< 30) yet have a highly specific tandem duplication of *ESR1*, suggestive of selection (Supplementary Figure S9). Most other samples with rearrangements in the other 32 hotspots were triple negative tumors. By contrast, samples with rearrangements in the *ESR1* hotspot showed a different preponderance – eleven of fourteen were estrogen receptor positive tumors. Thus we propose that the duplications in the *ESR1* hotspot are putative drivers that would not have been detected using customary copy number approaches previously, but are likely to be important to identify because of the associated risk of developing resistance to anti-estrogen chemotherapeutics[20,21].

c-MYC encodes a transcription factor that coordinates a diverse set of cellular programs and is deregulated in many different cancer types[22,23]. 30 patients contributed to the RS1 hotspot at the *c-MYC* locus with modest copy number gains. A spectrum of genomic outcomes was observed including single or nested tandem duplications, flanking (16 samples) or wholly duplicating the gene body of *c-MYC* (14 samples) (Figure 5A). Notably, a breast tissue super-enhancer and two germline susceptibility loci lie in the vicinity of *c-MYC* [24] [19](Figure 5B). We had a larger number of samples with corresponding RNA-seq data and thus modeled the expression levels of *c-MYC* taking breast cancer subtype, background copy number (whole chromosome arm gain is common for chr 8) and sought whether tandem duplicating a gene was associated with increased

transcription. We find that tandem duplications in the RS1 hotspot were associated with a doubling of the expression level of *c-MYC* (0.99 s.e. 0.28 log2 FPKM, $P=4.4 \times 10^{-4}$ in t-test) (Supplementary Table S5, Supplementary Figure S10).

The expression–related consequences of tandem duplications of putative regulatory elements however, is more difficult to assess because of the uncertainty of the downstream targets of these regulatory elements. We have thus taken a global gene expression approach and applied a mixed effects model to understand the contribution of tandem duplications of these elements, controlling for breast cancer subtype and background copy number. We find that tandem duplications involving a super-enhancer or breast cancer susceptibility locus are associated with an increase in levels of global gene expression even when the gene itself is not duplicated. The effect is strongest on oncogenes (0.30 +- 0.20 log2 FPKM, P=0.12 in likelihood ratio test) than for other genes (0.16 s.e. 0.04 log2 FPKM, $P=1.8 \times 10^{-4}$) within RS1 hotspots or for genes in the rest of the genome (Supplementary Table S4).

Thus, tandem duplications of cancer genes demonstrate strong expression effects in individual genes (e.g. *ESR1* and *c-MYC*) while tandem duplications of putative regulatory elements demonstrate modest but quantifiable global gene expression effects. The spectrum of functional consequences at these loci could thus range from insignificance, through mild enhancement, to strong selective advantage – consequences of the same somatic rearrangement mutational process.

Long tandem duplication hotspots are present and distinct in other cancers

We additionally explored other cancer cohorts where sequence files were available. Two cancer types are known to exhibit tandem duplications, particularly pancreatic and ovarian cancers. Raw sequence files were parsed through our mutation-calling algorithms and rearrangement signatures

extracted as for breast cancers. Adjusted background models and simulations were performed on these new datasets separately. The total numbers of available samples (73 ovarian and 96 pancreatic)[10,11] were much smaller than the breast cancer cohort, which is currently the largest cohort of WGS cancers of a single cancer type in the world. Thus power for detecting hotspots was substantially reduced particularly for pancreatic cancer (Supplementary Figure 11 for power calculation). Nevertheless, in ovarian tumors 2,923 RS1 rearrangements were found and seven RS1 hotspots identified (Supplementary Table S6), of which six were distinct from breast cancer RS1 hotspots. A marked enrichment for ovarian cancer specific super-enhancers (11 super-enhancers over 20.2 Mb, OR 2.9, $P=1.9 \times 10^{-3}$ in Poisson test) was also noted for these hotspots. *MUC1*, a validated oncogene in ovarian cancer was the focus at one of the hotspots. Thus, although we require larger cohorts of WGS cancers in the future to be definitive, the presentiment is that different cancer-types could have different RS1 hotspots that are focused at highly transcribed sites specific to different tissues.

Discussion: Selective susceptibility or selective pressure?

Rearrangement signatures may, in principle, be mere passenger read-outs of the stochastic mayhem in cancer cells. However, mutational signatures recurring at specific genomic sites, which also coincide with distinct genomic features, suggest a more directed nature – a sign of either selective susceptibility or selective pressure.

Perhaps it is an attribute of being more highly active or transcribed (e.g. super-enhancers) or some other as yet unknown quality (e.g. germline SNP sites and other hotspots with no discerning features), these hotspots exemplify loci that are rendered more available for DSB damage and more dependent on repair that generates large tandem duplications[6,25-27]. They signify genomic sites that are innately more susceptible to the HR-deficient tandem duplication mutational process – sites of selective susceptibility.

An alternative argument could also hold true: It could be that the likelihood of damage/repair relating to this mutational process is similar throughout the genome. However, through incrementally increasing the number of copies of coding genes that drive tissue proliferation, survival and invasion (*ESR1*, *ZNF217*) or non-coding regions that have minor or intermediate modifying effects in cancer such as germline susceptibility loci or super-enhancer elements, long tandem duplications (unlike other classes of rearrangements) could specifically enhance the overall likelihood of carcinogenesis. The profound implication is that these loci do come under a degree of selective pressure, and that this HR-deficient tandem duplication mutational process is in fact a novel mechanism of generating secondary somatic drivers.

Functional activity related to being a super-enhancer or SNP site could underlie primary susceptibility to mutagenesis of a given locus, but it requires a repair process that generates large tandem duplications to confer selective advantage (Figure 5C). Tandem duplication mutagenesis is associated with DSB repair in the context of HR deficiency and is a potentially important mutagenic mechanism driving genetic diversity in evolving cancers by increasing copy number of portions of coding and non-coding genome. It could directly increase the number of copies of an oncogene or alter non-coding sites where super-enhancers/risk loci[28] are situated. It could therefore produce a spectrum of driver consequences[29,30], ranging from strong effects in coding sequences to weaker effects in the coding and non-coding genome, profoundly, supporting a polygenic model of cancer development.

Conclusions

Structural mutability in the genome is not uniform. It is influenced by forces of selection and by mutational mechanisms, with recombination-based repair playing a critical role in specific genomic regions. Mutational processes may however not simply be passive contrivances. Some are possibly more harmful than others. We suggest that mutation signatures that confer a high degree of

genome-wide variability are potentially more deleterious for somatic cells and thus more clinically relevant. Translational efforts should be focused on identifying and managing these adverse mutational processes in human cancer.

Data availability statement

We used breast cancer data from a previous publication[4]. Our group previously submitted raw data for breast to the European-Genome Phenome Archive under the overarching accession number EGAS00001001178. Somatic variants in breast cancer genomes had been deposited at the International Cancer Genome Consortium Data Portal (https://dcc.icgc.org/).
Whole genome datasets from pancreatic and ovarian cancers had been described previously[10,11].

Code availability

The computer code used to identify hotspots of rearrangements can be accessed from the URL: https://github.com/DominikGlodzik/hotspots/tree/glodzik2016.

Acknowledgements

Data used in this analysis was funded through the ICGC Breast Cancer Working group by the Breast Cancer Somatic Genetics Study (BASIS), a European research project funded by the European Community's Seventh Framework Programme (FP7/2010-2014) under the grant agreement number 242006; the Triple Negative project funded by the Wellcome Trust (grant reference 077012/Z/05/Z) and the HER2+ project funded by Institut National du Cancer (INCa) in France (Grants N° 226-2009, 02-2011, 41-2012, 144-2008, 06-2012). JWMM received funding for this project through an ERC Advanced grant (no. 322737). The ICGC Asian Breast Cancer Project was funded through a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (A111218-SC01).

Funding: DG is supported by the EU-FP7-SUPPRESSTEM project. SN-Z is funded by a Wellcome Trust Intermediate Fellowship (WT100183MA) and is a Wellcome Beit Fellow.

Author contributions

D.G. and S.N-Z designed the study, analysed data and wrote the manuscript.

M.R.S., P.J.C., D.E, G.E., contributed towards idea development.

D.G. and S.M. performed all statistical analyses.

H.R.D., S.M., J.D.P., J.S., M.S. and X.Z. performed curation and contributed towards analyses.

M.S., contributed towards curation and analysis of transcriptomic data.

Y.L., L.B.A. contributed towards analysis.

C.P., P.T.S., S.R.L., I.H.R., H.R., contributed pathology assessment and/or samples and FISH analyses.

K.R. contributed IT expertise.

All authors discussed the results and commented on the manuscript.

Conflicts of interest

DG and SNZ are inventors on a patent application relating to the use of hotspots as breast and ovarian cancer diagnostics.

References

1. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).

2. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185 (2013).
3. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-24 (2015).
4. Nik-Zainal, S. Landscape of somatic mutations in 560 whole-genome sequenced breast cancers. (2016).
5. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
6. Mehta, A. & Haber, J.E. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol* **6**, a016428 (2014).
7. Ceccaldi, R., Rondinelli, B. & D'Andrea, A.D. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol* **26**, 52-64 (2016).
8. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nature communications* **7**(2016).
9. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-98 (2014).
10. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501 (2015).
11. Patch, A.M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-94 (2015).
12. Menghi, F. *et al.* The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci U S A* **113**, E2373-82 (2016).
13. McBride, D.J. *et al.* Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *J Pathol* **227**, 446-55 (2012).
14. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10 (2009).
15. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
16. Nilsson, B., Johansson, M., Heyden, A., Nelander, S. & Fioretos, T. An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biol* **9**, R13 (2008).
17. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
18. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**, 392-8, 398e1-2 (2013).
19. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
20. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-30 (2013).
21. Robinson, D.R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat Genet* **45**, 1446-51 (2013).
22. Soucek, L. *et al.* Modelling Myc inhibition as a cancer therapy. *Nature* **455**, 679-83 (2008).

23. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev* **27**, 2648-62 (2013).
24. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-82 (2016).
25. Costantino, L. *et al.* Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88-91 (2014).
26. Willis, N.A., Rass, E. & Scully, R. Deciphering the Code of the Cancer Genome: Mechanisms of Chromosome Rearrangement. *Trends Cancer* **1**, 217-230 (2015).
27. Saini, N. *et al.* Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature* **502**, 389-92 (2013).
28. Sloan, C.A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-32 (2016).
29. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer* **15**, 680-5 (2015).
30. Roy, A. *et al.* Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat Commun* **6**, 8891 (2015).

Figure legends

Figure 1: Spectrum of distribution of rearrangements in human breast cancers. Circos plots depicting somatic rearrangements with chromosomal ideogram on the outermost right and lines representing rearrangements (green=tandem duplications, orange=deletions, blue=inversions and gray=interchromosomal events). **A,** quiescent tumor, **B,** tumor with focal "clustered" rearrangements, **C,** tumor with mainly tandem duplications distributed throughout the genome ("dispersed" rearrangements) **D,** tumor with a mixed pattern of dispersed rearrangements and clustered rearrangements. **E,** Rearrangement Signatures 1 and 3 comprise mainly tandem duplications but are characterized predominantly by tandem duplications of different lengths (>100kb and <10kb respectively).

Figure 2: Identifying hotspots of rearrangements. A, A schematic of dispersed rearrangements in the genomes of 5 hypothetical patients, with regions that are identified as hotspots by the PCF algorithm highlighted in beige. Note the differing sizes of each putative hotspot permitted through this method that negates the use of bins. **B,** Workflow of PCF application to rearrangement signatures. **C,** Rainfall plots of chromosome 8 rearrangements for tandem duplication signatures RS1 (>100kb) top panel and RS3 (<10kb) bottom panel. Inter-rearrangement distance is plotted on a log-scale on the y-axis. Black lines demonstrate PCF-defined hotspots. RS1 is 20 times more likely to form hotspots than RS3 and these are visible as punctuated collections of breakpoints in these plots.

Figure 3: Hotspots of dispersed rearrangements: A large (>100kb) tandem duplication mutational process shows distinctive genomic overlap between patients and coincides with germline susceptibility loci and super-enhancer regulatory elements

A, A summary of 33 hotspots of long tandem duplications (RS1) and, **B**, 4 hotspots of short tandem duplications (RS3). Higher panel shows density of rearrangement breakpoints within hotspots, and their positions on chromosomes. The black horizontal lines denote the expected breakpoint density according to the background model. Lower panel shows frequency of each hotspot in the cohort of 560 patients. Hotspots that contain breast cancer susceptibility SNPs are marked with blue circles, and breast-specific super-enhancers marked with red triangles. Genes that may be relevant are highlighted although their true significance is uncertain. **C,** Two different hotspots of RS1:

left panel (chr12:11.8Mb-12.8Mb) coincides with two breast tissue specific super-enhancers and right panel (chr8:116.6Mb-117.7Mb) coincides with a germline susceptibility locus of breast cancer. Nearby cancer genes are annotated, although relevance of these genes is uncertain. Next six panels depict genomic rearrangements for six individual patients at each locus. Copy number (y-axis) depicted as black dots (10kb bins). Green lines present tandem duplication breakpoints. Note the precise genomic overlap between patients. Lowermost panel presents cumulative number of samples with a rearrangement involving this genomic region, emphasizing at its peak, the region of critical genomic overlap between samples. Thick red lines represent breast-tissue specific super enhancers. Blue vertical line represents position of germline susceptibility locus of breast cancer. Relevant SNP rsID is provided.

Figure 4: Genomic consequences of the tandem duplication signatures

Tandem duplications can transect or duplicate genomic features like regulatory elements or genes. **A,** tandem duplications attributed to rearrangement signature RS1 often duplicate genomic regions containing breast cancer predisposition SNPs, breast tissue super-enhancers and oncogenes. RS1 rearrangements in hotspots show a particular enrichment when compared to RS1 rearrangements that occur in other regions and when compared to simulated rearrangements. There are 524 RS1 duplications in hotspots, and 4,916 duplications outside of hotspots. **B,** tandem duplications attributed to RS3 in hotspots are enriched for transecting cancer genes more than in the rest of genome, or in simulated data. There are 57 RS3 duplications in hotspots, and 10,967 RS3 duplications outside of hotspots. Asterisks highlight statistically

significant enrichment of any particular genomic feature within hotspots compared to outside hotspots, as calculated by two-sided Fisher's exact test. Four asterisks **** denote p-value P<= 0.0001, ** P<=0.01, * P<=0.05 . Error bars show the standard deviation across ten different simulated datasets.

Figure 5: From selective susceptibility to selective pressure

A, The spectrum of genomic structural variation at a single locus: *c-MYC*. Copy number (y-axis) depicted as black dots (10kb bins). Lines represent rearrangement breakpoints (green= tandem duplications, pink=deletions, blue=inversions and purple=interchromosomal events). Genes other than c-MYC were marked as black lines at the top of the panel. **B,** Cumulative number of samples with dispersed rearrangements within the *c-MYC*-related tandem duplication hotspot. A peak is observed very close to *c-MYC* but also flanking *c-MYC* where two germline susceptibility loci are observed. A large super-enhancer is also situated upstream of *c-MYC*. **C,** Putative model of cascade of events underlying the RS1-enriched hotspots in breast cancer. Sites enriched for super-enhancers (SENH) may be more highly transcribed and thus exposed to damage including DSB damage. Long tandem duplications are particularly at risk of copying whole genes in contrast to other rearrangement classes. Thus although other rearrangement classes may be found (in low numbers in the same region), an enrichment of long tandem duplications is observed because of a small degree of selection in action.

Online Methods

1. Dataset

The primary dataset was obtained from another publication [4]. Briefly, 560 matched tumor and normal DNAs were sequenced using Illumina sequencing technology, aligned to the reference genome and mutations called using a suite of somatic mutation calling algorithms as defined previously. In particular, somatic rearrangements were called via BRASS (Breakpoint AnalySiS) (https://github.com/cancerit/BRASS) using discordantly mapping paired-end reads for the discovery phase. Clipped reads were not used to inform discovery. Primary discovery somatic rearrangements were filtered against the germline copy number variants (CNV) in the matched normal, as well as a panel of fifty normal samples from unrelated samples to reduce the likelihood of calling germline CNVs and to reduce the likelihood of calling false positives.

In silico and /or PCR-based validation were performed in a subset of samples [4]. Primers were custom-designed and potential rearrangements were PCR-amplified and identified as putatively somatic if a band observed on gel electrophoresis was seen in the tumour and not in the normal, in duplicate. Putative somatic rearrangements were then verified through capillary-sequencing. Amplicons that were successfully sequenced were aligned back to the reference genome using Blat, in order to identify breakpoints to basepair resolution. Alternatively, an *in silico* analysis was performed using local reassembly. Discordantly mapping read pairs that were likely to span breakpoints as well as a selection of nearby properly paired reads, were grouped for each region of interest. Using the Velvet de novo assembler [31], reads were locally assembled within each of these regions to produce a contiguous consensus sequence of each region. Rearrangements, represented by reads from

the rearranged derivative as well as the corresponding non-rearranged allele were instantly recognisable from a particular pattern of five vertices in the de Bruijn graph (a mathematical method used in de novo assembly of (short) read sequences) of component of Velvet. Exact coordinates and features of junction sequence (e.g. microhomology or non-templated sequence) were derived from this, following aligning to the reference genome, as though they were split reads.

Only rearrangements that passed the validation stage were used in these analyses. Furthermore, additional post-hoc filters were included to remove library-related artefacts (creating an excess of inversions in affected samples).

2. Rearrangement signatures

Previously, we had classified rearrangements as mutational signatures as extracted using the Non-Negative Marrix Factorization framework.

Briefly, we first separated rearrangements that were focally clustered from widely dispersed rearrangements because we reasoned that the underlying biological processes that generates these different rearrangement distributions are likely to be distinct. A piecewise constant fitting (PCF) approach was applied in order to distinguish focally clustered rearrangements from dispersed ones. For each sample, both breakpoints of each rearrangement were considered separately from one another and all breakpoints were ordered by chromosomal position. The inter-rearrangement distance, defined as the number of base pairs from one rearrangement breakpoint to the one immediately preceding it in the reference genome, was calculated. Putative regions of clustered rearrangements were identified as having an average inter-rearrangement distance that was at least 10 times greater than the whole genome average for the individual sample. PCF parameters used were $\gamma = 25$ and $k\text{min} = 10$. The respective partner breakpoint of all breakpoints involved in a clustered region are likely to have arisen at the same mechanistic instant and so were considered as being involved in the cluster even if located at a distant chromosomal site.

In both classes of rearrangements, clustered and non-clustered, rearrangements were subclassified into deletions, inversions and tandem duplications, and then further subclassified according to size of the rearranged segment (1-10kb, 10kb-100kb, 100kb-1Mb, 1Mb-10Mb, more than 10Mb). The final category in both groups was interchromosomal translocations. The classification produces a matrix of 32 distinct categories of structural variants across 544 breast cancer genomes. This matrix was decomposed using the previously developed approach for deciphering mutational signatures by searching for the optimal number of mutational signatures that best explains the data [5]. A rearrangement was attributed to a signature if the posterior probability for the rearrangement to be generated by the signature in a given sample exceeded 0.5[8].

In all, six different rearrangement signatures were identified. Rearrangement Signatures 1 and 3 were two signatures that were particularly characterised by tandem duplications (Main Figure 1E).

Rearrangement signature 1 (RS1) is characterized mainly by large tandem duplications (>100kb) while rearrangement signature 3 (RS3) is characterised mainly by short tandem duplications. There is good reason to believe that these signatures are biologically distinct entities as RS3 is very strongly associated with BRCA1 abrogation (germline or somatic mutation or promoter hypermethylation with concurrent loss of the wild-type allele). BRCA1 tumours also contain moderate numbers of RS1, but there are also samples with a larger excess of RS1 rearrangements that do not carry a specific genetic abnormality [4].

In order to perform a systematic survey of tandem duplication hotspots, we focused on these two rearrangements signatures. However, tandem duplications (and other rearrangements) are also not uniformly distributed through the genome. Thus, the following sections describe how we detect hotspots of tandem duplications of RS1 and RS3, after correcting for genomic biases.

3. Modelling the background distribution of rearrangements

Rearrangements are known to have an uneven distribution in the genome. There have been numerous descriptions linking genomic features such as replication timing with the non-uniform distribution of rearrangements. Thus, any analysis that seeks to detect regions of higher mutability than expected must take the genomic features that influence this non-uniform distribution into account in its background model. In order to formally detect and quantify associations between genomic features and somatic rearrangements in breast cancer, we conducted a multi-variate genome-wide regression analysis. Please see Section 1 in Supplementary Note for details.

4. Simulations of rearrangements

Simulations consisted of as many rearrangements as was observed for each sample in the dataset, preserving the type of rearrangement (tandem duplication, inversion, deletion or translocation), the length of each rearrangement (distance between partner breakpoints) and ensuring that both breakpoints fell within mappable/callable regions in our pipeline. Simulations also took into account the genomic bias of rearrangements that were identified above. Please see Section 2 in Supplementary Note for details.

5. Optimization of the PCF algorithm

The PCF (Piecewise-Constant-Fitting) algorithm is a method of segmentation of sequential data. We used PCF to find segments of the genome that had a much higher rearrangement density than the neighbouring genomic regions, and higher than expected according the background model. We show the significance of the identified hotspots by applying the same method to simulated data

(Section 4) that follows the known genomic biases of rearrangements like replication time domains, transcription and background copy number status.

Each rearrangement has two breakpoints and these breakpoints were treated independently of each other. Breakpoints were sorted according to reference genome coordinates and an intermutation distance (IMD) between two genome-sorted breakpoints was calculated for each breakpoint, then log-transformed to base 10. Log 10 IMD were fed into the PCF algorithm.

In order to call a segment of a genome that has a higher rearrangement density as a "hotspot", a number of parameters had to be determined. The smoothness of segmentation is determined by the **gamma (γ)** parameter of the PCF analysis. A segment of genome was only considered a peak if it had a sufficient number of mutations, as specified by $\mathbf{k_{min}}$. The average inter-mutation distance in the segment had to exceed an **inter-mutation distance factor (*i*)**, which is the threshold when comparing breakpoint density in a segment to genome-wide density of breakpoints:

$$\frac{d_{seg}}{d_{bg}} > i$$

where:

d_{seg} is the density of breakpoints in a segment defined as:

$d_{seg} =$ (number of breakpoints in segment)/(length in bp of a segment)

d_{bg} is the expected density of breakpoints in the segment, given the background model from Section 3, which includes the genomic covariates of the segment.

More specifically,

$d_{bg} = (\sum_{i=1}^{n} b_i)/(\mathrm{n} * \mathrm{s})$, where b_i is the expected number of breakpoints in the bins overlapping the segment, n is the number of overlapping bins, and s is bin size (0.5 Mb).

The choice of parameters k_{min}, γ and i for the PCF algorithm was based on training on the observed data and comparing the outcomes with that of the simulated data.

Combinations of γ and i were explored to determine the optimal parameters for detection of hotspots where the sensitivity of detection of every hotspot in observed data was balanced against the detection of false positive hotspots in simulated datasets (Supplementary Figure S2). This was quantified according to the false discovery rate.

Based on the number of detected hotspots on observed and simulated data, we used the $\gamma=8$ and $i=2$ in the final analyses which results in 33 hotspots of RS1 and 4 of RS3. In further 1000 simulated datasets the same parameters resulted on average in 3.3 (standard deviation 1.9) and 0.1 (standard deviation 0.3) hotspots respectively.

A dataset that is not "clean" and that contains a lot of false positive rearrangements could result in the identification of hotspots of false positives. Thus, it is imperative to have a set of high quality, highly curated rearrangement data – with a better specificity than sensitivity – in order to avoid calling loci where algorithms have a tendency to miscall rearrangements, as hotspots.

6. Workflow

Six rearrangement signatures were extracted from this dataset of 560 breast tumours as previously described (Section 2). Each rearrangement was probabilistically assigned to each rearrangement signature given the six rearrangement signatures and the estimated contribution of each signature to each sample [4].

To define hotspots of rearrangements in RS1 and RS3, the PCF algorithm was applied to the log10 IMD of RS1 or RS3 breakpoints separately using the

following parameters: γ= 8, k_{min} = 8 and i = 2. Each locus was required to be represented by 8 or more samples. The section below describes the hotspots that were identified by this method.

7. Identifying hotspots for individual rearrangement signatures

To explore hotspots associated with signatures of tandem duplications, we first separated rearrangements associated with the two signatures that are strongly characterised by tandem duplications (RS1 and RS3) (Main Figure 2). PCF was performed on each of these two categories. 33 hotspots of long RS1 tandem duplications were identified and 4 hotspots of short RS3 tandem duplications were seen, and they are listed and annotated in Supplementary Table S1.

RS3 characterised by short tandem duplications also demonstrated 4 hotspots, two were likely drivers (*PTEN*, *RB1*) and the significance of the other two are less clear (*CDK6* and *NEAT1/MALAT1*). The interpretation of duplications at the *NEAT1/MALAT1* locus is provided in Supplementary Note Section 3.

For hotspots of remaining rearrangement signatures please see the Supplementary Note Section 4.

8. Genomic consequences of tandem duplications

We assessed the potential genomic consequences of the two rearrangement signatures associated with tandem duplications on gene function and on regulatory elements.

Rearrangements associated with the RS1 signature are usually long tandem duplications (>100kb). These are more likely to duplicate whole genes and whole super-enhancer regulatory elements. In contrast, rearrangements associated with the RS3 signature are usually short tandem duplications

(<10kb), and therefore more likely to duplicate smaller regions which could have an effect equivalent of transecting genes or regulatory elements.

To formally assess the potential genomic consequences of RS1 and RS3 tandem duplications on gene function and on regulatory elements, we explored the following genomic elements:

- breast cancer susceptibility SNPs
- breast-tissue specific super-enhancer regulatory elements
- oncogenes (if a duplications covers both a super-enhancer and an oncogene, it will be counted in both categories)
- tumour suppressor genes
- all genes

An element was considered as wholly duplicated by a tandem duplication if the element was completely between the two breakpoints. An element was considered as transected by a tandem duplication if one or both breakpoints lay within the element.
We did not consider the events where only one breakpoint of duplication was within an element, as the effect of such events on genes and other elements is unclear.

We counted the number of times each of the five elements noted above was duplicated or transected for RS1 and RS3 respectively for:

- RS1 or RS3 tandem duplications in hotspots (counted only once per sample – even if there are multiple tandem duplications affecting the same locus in the same person),
- RS1 or RS3 tandem duplications that are not within hotspots,
- RS1 and RS3 tandem duplications that have been simulated correcting for all the characteristics described above.

Strikingly, RS1 hotspots are clearly enriched for duplicating whole oncogenes and whole super-enhancers, compared to RS1 rearrangements that are not

within hotspots and simulated RS1 rearrangements (Main Figure 4, Supplementary Table 3). This enrichment is not observed for RS3 hotspots. Furthermore, RS1 hotspot tandem duplications hardly ever transect genes or regulatory elements. In contrast, RS3 hotspots are strongly enriched for gene transections in-keeping with being driver loci.

Thus here we provide evidence for different genomic consequences – whole gene/regulatory element duplications versus transections - given hotspots generated through different types of rearrangements, long or short tandem duplications.

9. Germline loci of susceptibility to breast cancer

The list of breast cancer germline susceptibility alleles was derived from the literature [18,32-40]. This analysis is aimed at trying to determine whether there is an enrichment for breast cancer susceptibility SNP alleles in breast cancer, to quantify this relationship and provide a measure of statistical significance.

We performed an analysis that compares the density of SNPs in the genomic footprint of RS1 hotspots against the genomic footprint of other RS1 rearrangements in general (instead of simply to the rest of genome) – this controls for the unevenness in the distribution of tandem duplications. RS1 hotspots encompass 58Mb of the genome while other segments of the genome covered by (at least) one tandem duplication encompasses 2,106Mb.

The density of breast cancer susceptibility SNPs outside of RS1 hotspots was 0.036 per Mb. Within RS1 hotspots, there were 9 breast cancer susceptibility SNPs or 0.16 SNPs per Mb. Thus, the odds ratio (OR) of finding a breast cancer susceptibility SNP in RS1 hotspots compared to tandem duplicated regions outside of RS1 hotspots is 4.28 ($P=3.4\times 10^{-4}$ Poisson one-sided).

The Poisson test was used in order to compare rates of events between genomic regions of different sizes, and to account for uncertainty that comes from low number of events (9 SNPs) falling into the hotspots.

10. *Enrichment for regulatory elements*

The super-enhancer dataset was obtained from Super-Enhancer Archive (SEA)[40]. This archive uses publicly available H3K27ac Chip-seq datasets and published super-enhancers lists to produce a comprehensive list of super-enhancers in multiple cell types/tissues. From this list (containing 2,282 unique super-enhancers for 15 human cell types/tissues), we extracted the super-enhancers active in breast cancer (755 elements) and the super-enhancers active in the other cell types/tissues (1,528 elements). Regulatory elements were mutually exclusive to each list to ensure that each super-enhancer was analyzed only in one category, and a super-enhancer was placed in the breast cancer category where there was experimental evidence for multiple activations.

The list of general enhancers was obtained from Ensembl Regulatory Build (GRCh37)[41]. We used the "Multicell" list containing 139,204 elements active in 17 different cell lines. From this list, we filtered out the enhancers that overlapped with super-enhancers, and we obtained a final list composed of 136,858 regulatory elements.

As described in the previous section, we divided the genome into RS1 hotspots (58Mb), and other segments of the genome covered by a minimum of a single tandem duplication (2,106Mb). We compared the density of super-enhancers within RS1 hotspot segments and outside of the hotspots (Supplementary Figure S4).

Method 1:

The OR of finding a super-enhancer active in breast tissue in RS1 hotspots, compared to regions of the genome rarely covered by RS1 duplications is 3.54 (Poisson one-sided test $P=7.0 \times 10^{-16}$). The OR for observing a super-enhancers that is not associated with breast tissue is lower at 1.62, with $P=6.4\times 10^{-4}$. The OR for finding any enhancer in an RS1 hotspots is 1.02, with a p-value of 0.12.

Method 2:

The assumption made in the above analysis is that super-enhancers follow a Poisson distribution, which could be violated by clusters of super-enhancer elements that exist in the genome. We thus performed a set of simulations that do not depend on these assumptions.

In order to assess the likelihood of observing 59 super-enhancers within the regions of RS1 hotspots, the same number of regions of equivalent sizes was sampled from the genome. Similarly as in the previous analysis, the random segments of the genome were drawn from genomic regions representative of non-hotspot tandem duplications (2,106Mb). The procedure was repeated 10,000 times and super-enhancers falling into the simulated segments were counted.

The observed overlap with 59 or more super-enhancers occurred zero times in 10,000 simulation rounds, by which we estimate the p-value of the observation to be $P<10^{-4}$. Figure S4C shows the empirical distribution observed in the simulations.

11. Analysis of gene expression

RNA expression levels of genes in the samples were obtained from RNA-seq data as reported by another publication [4]. We set out to assess whether tandem duplications in the hotspots are associated with increased expression of affected

genes. Statistical methods and results are presented in Supplementary Note Section 5.

12. Hotpots of RS1 in other tumours

In addition to breast cancer, tumours of other tissue types sometimes show excess of tandem duplications in their genomes. In order to investigate whether the rearrangements in other tumor types also accumulate in hotspots, we utilized previously published sequences of ovarian and pancreatic cancer genomes. Please see Supplementary Note Section 6.

13. Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomised and the investigators were not blinded to allocation during experiments as this was not relevant to the study.

REFERENCES

1. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).
2. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185 (2013).
3. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-24 (2015).
4. Nik-Zainal, S. Landscape of somatic mutations in 560 whole-genome sequenced breast cancers. (2016).
5. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
6. Mehta, A. & Haber, J.E. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol* **6**, a016428 (2014).
7. Ceccaldi, R., Rondinelli, B. & D'Andrea, A.D. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol* **26**, 52-64 (2016).
8. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nature communications* **7**(2016).

9. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-98 (2014).
10. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501 (2015).
11. Patch, A.M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-94 (2015).
12. Menghi, F. *et al.* The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci U S A* **113**, E2373-82 (2016).
13. McBride, D.J. *et al.* Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *J Pathol* **227**, 446-55 (2012).
14. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10 (2009).
15. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
16. Nilsson, B., Johansson, M., Heyden, A., Nelander, S. & Fioretos, T. An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biol* **9**, R13 (2008).
17. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
18. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**, 392-8, 398e1-2 (2013).
19. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
20. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-30 (2013).
21. Robinson, D.R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat Genet* **45**, 1446-51 (2013).
22. Soucek, L. *et al.* Modelling Myc inhibition as a cancer therapy. *Nature* **455**, 679-83 (2008).
23. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev* **27**, 2648-62 (2013).
24. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-82 (2016).
25. Costantino, L. *et al.* Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88-91 (2014).
26. Willis, N.A., Rass, E. & Scully, R. Deciphering the Code of the Cancer Genome: Mechanisms of Chromosome Rearrangement. *Trends Cancer* **1**, 217-230 (2015).
27. Saini, N. *et al.* Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature* **502**, 389-92 (2013).
28. Sloan, C.A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-32 (2016).
29. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer* **15**, 680-5 (2015).
30. Roy, A. *et al.* Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat Commun* **6**, 8891 (2015).

31. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
32. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* **39**, 352-8 (2007).
33. Easton, D.F. *et al.* A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet* **81**, 873-83 (2007).
34. Ahmed, S. *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* **41**, 585-90 (2009).
35. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-80 (2015).
36. Siddiq, A. *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum Mol Genet* **21**, 5373-84 (2012).
37. Stacey, S.N. *et al.* Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **40**, 703-6 (2008).
38. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**, 579-84 (2009).
39. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**, 504-7 (2010).
40. Wei, Y. *et al.* SEA: a super-enhancer archive. *Nucleic Acids Res* **44**, D172-9 (2016).
41. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. & Flicek, P.R. The ensembl regulatory build. *Genome Biol* **16**, 56 (2015).