Landscape of G-quadruplex DNA structural regions in breast cancer

Authors:

Robert Hänsel-Hertsch[1,2,3,4], Angela Simeone[4], Abigail Shea[4], Winnie W.I. Hui[4], Katherine G. Zyner[4], Giovanni Marsico[4], Oscar M. Rueda[4], Alejandra Bruna[4], Alistair Martin[4], Xiaoyun Zhang[5], Santosh Adhikari[5], David Tannahill[4], Carlos Caldas[4,6,7] and Shankar Balasubramanian[4,5,8]

[1]Center for Molecular Medicine Cologne, University of Cologne, Germany.
[2]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne and University Hospital Cologne, Cologne, Germany.
[3]Institute of Human Genetics, University Hospital Cologne, Cologne, Germany.
[4]Cancer Research UK Cambridge Institute and Department of Oncology, Li Ka Shing Centre, University of Cambridge, Cambridge, UK.
[5]Department of Chemistry, University of Cambridge, Cambridge, UK.
[6]Cambridge Breast Cancer Research Unit, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK.
[7]Cancer Research UK Cambridge Centre, NIHR Cambridge Biomedical Research Centre and Cambridge Experimental Cancer Medicine Centre, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK.
[8]School of Clinical Medicine, University of Cambridge, Cambridge, UK

Present addresses:

Center for Molecular Medicine Cologne, University of Cologne, Germany
Robert Hänsel-Hertsch

Cancer Research UK Cambridge Institute and Department of Oncology, Li Ka Shing Centre, University of Cambridge, Cambridge, UK.
Angela Simeone, Abigail Shea, Winnie W.I. Hui, Katherine G. Zyner, Oscar M. Rueda, David Tannahill, Alistair Martin, Carlos Caldas, Shankar Balasubramanian

Inivata, Babraham Research Campus, Cambridge, UK.
Giovanni Marsico

The Institute of Cancer Research, London, UK
Alejandra Bruna

Department of Chemistry, University of Cambridge, Cambridge, UK.
Xiaoyun Zhang, Santosh Adhikari, Shankar Balasubramanian

Corresponding author:

Shankar Balasubramanian
e-mail: sb10031@cam.ac.uk

Abstract:
Response and resistance to anticancer therapies vary due to inter- and intra-tumor heterogeneity[1]. Here, we map differentially enriched G-quadruplex (G4) DNA structure-forming regions (ΔG4Rs) in 22 breast cancer patient-derived tumor xenograft (PDTX) models. ΔG4Rs are associated with the promoter of highly amplified and expressed genes, and with somatic single-nucleotide variants. Specific ΔG4Rs reveal 7 transcription factor (TF) programs across PDTXs. ΔG4R abundance and locations stratify PDTXs into at least three G4-based subtypes. ΔG4Rs in most PDTXs (14/22) associated with more than one breast cancer subtype, which we also call an integrative cluster (IC)[2]. This suggests the frequent coexistence of multiple breast cancer states within a PDTX model; the majority displaying aggressive triple-negative IC10. Short-term cultures of PDTX models with increased ΔG4R levels are more sensitive to small molecules targeting G4 DNA. Thus, G4 landscapes reveal additional IC-related intra-tumor heterogeneity in PDTX biopsies, improving breast cancer stratification and potentially new treatment strategies.

Main:
G-quadruplexes are four-stranded secondary structures that can form in certain G-rich DNA sequences[3,4]. We previously used in vitro sequencing (G4-seq) to establish where endogenous G4s could form in the human genome[5]. Qualitative profiling of endogenous G4 DNA in chromatin by G4-ChIP-seq revealed prominence of their formation in promoters of highly expressed cancer genes[6–8]. Computational predictions of G4s in eukaryotic genomes have linked G4 motifs to genomic instability[9–11], suggesting that G4-selective helicases maintain genome stability during DNA replication and transcription[3,4]. Supporting this, we have recently reported the prevalence of endogenous DNA double-strand breakage (DSB) in G4-seq derived sequences that are found in nucleosome-depleted regions (NDRs) of highly expressed human cancer genes[12]. Fundamental mechanisms including DNA transcription and replication are endogenous sources for DSBs and genome instability[13]. Computational predictions of DNA motifs[14] have suggested that human G4s may be associated with pan-cancer somatic copy number aberrations (CNAs), which we previously confirmed by G4-seq[15]. CNA landscapes impact gene expression and shape breast cancer heterogeneity[2]. Our analysis of 2,000 primary breast cancers previously revealed 11 different subgroups, called integrative clusters (ICs)[16–18].

To establish how G4 DNA structures may relate to breast cancer biology, we developed and applied a quantitative, comparative G4-ChIP-seq (qG4-ChIP-seq) methodology to map G4 DNA structure formation in 22 breast cancer PDTX models that retain their original inter- and intra-tumor heterogeneity[17]. We adapted the ChIP-Rx approach[19] and employed *Drosophila melanogaster* chromatin as an internal reference to normalize the ChIP-seq data and reduce technical variability to enhance the characterization of true biological variation (**Fig. 1a** and **Methods**). Improvement in experimental reproducibility can be evaluated by analyzing the similarity between four repeated measurements of one PDTX sample vs. four repeated measurements acquired from a different PDTX sample; either from the same or a different PDTX model (**Fig. 1b-d**, **Extended Data Fig. 1a**). Normalization increased the reproducibility of our human cancer qG4-ChIP-seq data across technical and biological replicates. We derived

a coefficient termed improvement factor (I_F) whereby $I_F > 0$ indicates increased reproducibility, whereas $I_F < 0$ signifies decreased reproducibility after normalization (**Fig. 1b-d,** see **Methods**). We applied qG4-ChIP-seq to profile the G4 landscape in estrogen receptor-positive (ER+) or triple-negative (ER-, HER2-, PR-) PDTX models representing most integrative clusters (IC 1, 5, 8, 9 and 10)[17]. We assessed the reproducibility of qG4-ChIP-seq by processing different parts of the same tumor on a different day with different reagents, while keeping the reference chromatin batch constant (**Fig. 1d** and **Supplementary Table 1**). Overall, across all studied PDTX models, qG4-ChIP-seq identified ~26,000 reproducibly enriched regions of which 97% comprised a G4 sequence motif (**Extended Data Fig. 1b**). Comparative qG4-ChIP-seq analysis of 22 PDTX models revealed differentially enriched G4 regions (~700-17,000), hereafter called ΔG4Rs, and constant G4 regions (~100), hereafter called CG4Rs (see **Methods** for detailed description). We found that some ΔG4Rs are unique to a given PDTX (**Fig. 1e, f**) whilst others are common to more than one PDTX model (**Extended Data Fig. 1c**), suggesting that ΔG4R loci may relate to differences in intrinsic cancer biology.

To explore whether the ΔG4Rs are coupled to the underlying PDTX biology, we performed a pairwise comparison of the ΔG4Rs in all PDTX models and stratified them according to their similarity (**Fig. 2a**). Without consideration of the annotated PDTX IC or ER status (**Extended Data Fig. 2a**), hierarchical clustering of the ΔG4R similarity alone revealed the existence of three PDTX clusters (**Fig. 2a**). To explore the relationship between CNAs and ΔG4Rs or CG4Rs, we determined CNAs in the PDTX models[20] by comparing the data from sequenced input libraries with the corresponding qG4-ChIP-seq data for each PDTX sample (see **Methods**). Examination of highly amplified (AMP), amplified (GAIN), neutral (NEUT), heterozygous deletions (HETD) and homozygous deletions (HOMD) across all PDTX models revealed a significant enrichment ($P < 0.0001$, **Fig. 2b**) of ΔG4Rs, but not CG4Rs, in AMPs relative to the other CNA categories. ΔG4Rs are also more abundant ($P < 0.0001$, **Extended Data Fig. 2b**) in amplified regions (AMP + GAIN) in comparison to the other CNA categories. Notably, the number of observed AMPs does not explain the ΔG4R abundance or enrichment in AMPs since ΔG4R and AMP levels vary independently (**Extended Data Fig. 2c**). We also explored a possible connection of G4 structure with the occurrence of single-nucleotide variants (SNVs); we previously derived SNVs for some of the PDTXs used here[17]. Notably, ΔG4Rs, but not CG4Rs, are significantly ($P < 0.0001$) enriched in SNVs of the PDTXs relative to random permutation, implying a potential role of G4 formation in the formation of breast cancer point mutations (**Fig. 2b**). In agreement with our previous observations in cell lines[6–8,21], G4 structures in the PDTXs are highly enriched in gene promoters, including 5'UTR regions (**Fig. 2c**). We find that in the tumor material derived from PDTXs, ΔG4Rs are also significantly enriched ($P < 0.0001$) in promoters of highly expressed genes when compared to medium and lowly expressed ones (**Fig. 2d**, for gene expression classification see **Methods**). Strikingly, regardless of IC or ER classification, highly expressed promoters show significantly ($P < 0.05$) greater qG4-ChIP-seq signal in highly amplified (AMP) CNAs relative to other promoters (**Fig. 2e**). Thus, G4 structures are more prevalent in promoters of highly expressed and amplified genes in a way that cannot be explained by a single IC and/or ER status. To explore whether ΔG4Rs of a particular PDTX associate with its anticipated IC gene signature,

we systematically overlapped promoter regions of the 10 different IC gene sets with the 22 different ΔG4Rs. Across all PDTXs, ΔG4Rs associate more ($P < 0.001$) with the signature gene promoter of IC10 than with IC1-9 (**Extended Data Fig. 2d**). These results suggest that the majority of PDTXs in our cohort display aggressive triple-negative IC10-related breast cancer gene activity. While the individual ΔG4Rs of the 22 PDTXs generally associate with their anticipated IC status, most (14/22) models display the existence of multiple or distinct IC-related signature genes (**Fig. 2f** and **Extended Data Fig. 2e**). For example, integrative CNA and expression profiling of '+/1/HCI005' and '-/10/VHIO179' stratifies these PDTXs as IC1 and IC10, yet their ΔG4Rs predominantly associate with at least two different IC-defining promoter sets that are highly expressed (**Fig. 2f**). This suggests ΔG4Rs provide additional information relative to CNA/expression profiling and revealed the coexistence of multiple cancer states, thus more intra-tumor heterogeneity with respect to ICs for the majority of PDTX models (**Extended Data Fig. 2e**). The analysis of 2,000 primary breast tumors revealed 45 common driver regions that are characteristic for CNA-induced gene expression alterations[2]. ΔG4Rs, but not CG4Rs, associate with the 45 common breast cancer driver regions (**Fig. 2g**), highlighting ΔG4Rs as a genomic marker of breast cancer driver regions.

While pioneering factors such as FOXA1 establish nucleosome-depleted regions (NDRs), TFs bind to NDRs, thereby mediating transcriptional activity, e.g. through promoter enhancer interactions[22]. Importantly, TFs can co-target a particular NDR via interactions with other TFs, thus they can bind DNA independently of their primary consensus binding motif[23]. As ΔG4Rs are prevalent in NDRs[6], we hypothesized that any ΔG4R association with TF binding sites (TFBS) might reveal TFs that differentially regulate breast cancer development in the PDTX models. To address this hypothesis, we extracted TF binding sites (TFBS) (see **Methods**) from breast cancer TF ChIP-seq datasets (ChIP-ATLAS)[24] and computed fold-enrichment over random in the different ΔG4Rs of all the 22 PDTX models (see **Methods**). Hierarchical clustering of ΔG4R fold-enrichments in TFBS revealed increased similarity among some PDTX models (**Extended Data Fig. 3a, b**), suggesting that some PDTX models share the same TF activities while others do not. Considering the similarity of ΔG4R fold-enrichments in TFBS across the 22 PDTX models, we identified 7 distinct TF programs that are differentially active across the PDTXs (**Fig. 3**, **Extended Data Fig. 3c**). Strikingly, ΔG4R – TFBS enrichments of 4/7 TF programs were significantly higher in either IC10/9, IC8/1, ER-negative or -positive stratified PDTX models, suggesting that these TF programs are more active in certain breast cancer subtypes. We found that differential TF expression levels of a TF program can coincide with the ΔG4R fold-enrichments in TFBS of particular PDTXs. For example, the TLE3-GATA3 TF cluster is significantly more expressed and enriched for ΔG4Rs in PDTXs that are ER-positive or IC 8/1 but not ER-negative or IC10/9 (**Fig. 3**, **Extended Data Fig. 3c**). Importantly, all TF programs are expressed (**Extended Data Fig. 3c**), suggesting that ΔG4R fold-enrichments in TFBS may infer differential TF activity in cancer tissues.

To characterize pharmacogenomic correlations and enable strategies for precision medicine, we established short-term cultures of PDTX-derived tumor cells (PDTC). Importantly, PDTCs preserve the intra-tumor heterogeneity of the PDTX models[17]. Our high-throughput drug screens, deposited in the Breast Cancer PDTX Encyclopedia (BCaPE), revealed substantial

differences in PDTC response, importantly, even among PDTCs derived from models stratified into the same integrative cluster (**Extended Data Fig. 4a**). This suggests the need to consider additional approaches to decode pharmacogenomic correlations. We previously demonstrated that human immortalized keratinocytes displayed ~7-fold more G4 regions than normal keratinocytes by G4-ChIP-seq[6] and exhibited a corresponding increase in sensitivity (~7-fold) to G4-ligand treatment by pyridostatin (PDS)[25]. This led us to hypothesize that models with higher ΔG4R levels would respond better to G4-ligand treatment, because they are a quantitative measure of differences in the number of G4 regions. To explore this, we evaluated G4 ligand-sensitivity in PDTC derived from models with qG4-ChIP-seq data. We evaluated two established, yet structurally distinct, small molecules with high G4 DNA selectivity; PDS[25] and CX-5461[26]. As a negative control, we synthesized an isomer of PDS (i-PDS) that shows substantially reduced G4 affinity (**Extended Data Fig. 4b**, **Methods** and **Supplementary Data 1**). We found that PDTCs with an increased level of ΔG4Rs were significantly ($P < 0.05$, $r = 0.5$-0.8) more sensitive to PDS and CX-5461 but not control i-PDS G4-ligand treatments (**Fig. 4**). Since PDTX ΔG4Rs are highly enriched in amplified CNAs, we asked whether CNA amplification level alone was sufficient to predict G4 sensitivity. Notably, we found that amplified CNA levels lacked a positive correlation with PDTC responses to all G4-ligands (**Fig. 4**, **Extended Data Fig. 4c**). Taken together, these findings highlight the potential of ΔG4R mapping as a predictive biomarker for G4-ligand therapy[26].

By developing quantitative G4-ChIP-seq, we have obtained G4 DNA maps in chromatin from patient-derived models, which substantially advances previous qualitative methods using established cell lines in 2D culture or tissue immunohistochemistry[6,27]. We have generated G4-DNA maps for 22 PDTX breast cancer models and revealed how they reflect the underlying breast cancer biology, such as the relationship with TF occupancy and highly expressed driver genes in amplified CNAs. Our matched integrative analysis of PDTX-derived somatic mutations, CNAs and SNVs, and endogenous G4 DNA landscapes highlight a link between cancer genome instability and G4 structure formation. Overall, we discovered that G4 DNA regions are highly associated to critical drivers of triple-negative breast cancer models and/or IC9-10 relative to ER+/IC1-8 PDTX models. While strategies are currently under development to identify cancers that respond to G4 ligand treatment based upon their BRCA1/2 status[26,28,29], our results indicate that G4 profiling alone can identify sensitive cancers, which may or may not be related to their BRCA status. By integrating PDTX ΔG4Rs with established gene signatures of 10 different breast cancer subtypes (IC), we discovered that the majority of the 22 PDTX models have G4 patterns that associate with more than one IC, providing an added layer of intra-tumor heterogeneity. Our interrogation of breast cancer TF ChIP-seq profiles with the ΔG4Rs has highlighted the existence of at least seven distinct TF programs that are mostly dominant in either ER+/IC1-8 or triple-negative/IC9-10 breast cancers. This supports that many TFs, instead of a single, defined TF or TF-complex, co-target and -regulate breast cancer gene activity. Quantitative profiling of G4 structures adds information to conventional copy number aberration and expression profiling, potentially increasing resolution up to ~1000-fold (~100-500 bp vs. ~100 kb), hence helping in the identification of specific drivers within large amplicons. We also provide evidence that ΔG4Rs, in combination with established knowledge on subtypes, can refine the genomic, transcriptomic and regulatory classification of

breast cancer. Finally, G4 levels in cancer models are sufficient to predict response to treatment by small molecules that target G4 DNA structures, highlighting G4s as genomic features with potential for future diagnostics and therapeutics.

Acknowledgements

The authors would like to thank the staff in the Genomics and the Compliance and Biobanking Core Facilities at Cancer Research UK Cambridge Institute. We acknowledge support from University of Cambridge and Cancer Research UK program. We kindly thank Dr. Ben Czech of the Hannon laboratory for providing S2 *D. melanogaster* cells. The Caldas and Balasubramanian laboratories are supported by core funding from Cancer Research UK (C14303/A17197). The Balasubramanian laboratory is supported by Program grant funding from Cancer Research UK (C9681/A18618 and C9681/A29214) and a Wellcome Trust Investigator Award (209441/z/17/z). We acknowledge Dr. Marco Di Antonio for conceptualizing the design of i-PDS. Prior to the revision of this study work by Dr. Robert Hänsel-Hertsch was supported by the Balasubramanian group, afterwards additionally supported by core funding of the Center for Molecular Medicine Cologne (CMMC).

Author Contributions

R.H.H., C.C. and S.B. conceived this study. R.H.H. developed quantitative G4-ChIP-seq. R.H.H., A.B., O.M.R. and C.C. designed the PDTX model panel for this study. R.H.H. processed all the PDTX tissues and prepared the chromatin samples for G4-ChIP-seq. R.H.H., W.H. and K.G.Z. performed G4-ChIP-seq. A.M., A.B., O.M.R. and C.C. performed and interpreted genomic and transcriptomic characterization of all PDTX models. A. Shea performed the G4-ligand treatment assay, which was analyzed by O.M.R.. R.H.H. and G.M. implemented a computational pipeline to measure normalization performance, which was refined by A. Simeone. X.Z. synthesized i-PDS with support of S.A., and performed G4-ligand *in vitro* experiments and analysis. R.H.H. and A. Simeone performed all the G4-ChIP-seq related computational analysis. R.H.H., C.C., and S.B. interpreted the results with input from all authors. R.H.H. prepared the figures. R.H.H., C.C., and S.B. wrote the manuscript with contributions from all authors.

Competing interests

S.B. is an advisor and shareholder of Cambridge Epigenetix Ltd.

References:

1. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
2. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
3. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **43**, 8627–37 (2015).
4. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol.* (2020). doi:10.1038/s41580-020-0236-x
5. Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple

species. *Nucleic Acids Res.* **47**, 3862–3874 (2019).
6. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–72 (2016).
7. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **13**, 551–564 (2018).
8. Hänsel-Hertsch, R., Antonio, M. Di & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Biol.* **18**, 279–284 (2017).
9. Paeschke, K., Capra, J. A. & Zakian, V. A. DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678–91 (2011).
10. Cheung, I., Schertzer, M., Rose, A. & Lansdorp, P. M. Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nat. Genet.* **31**, 405–409 (2002).
11. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 1264–1271 (2018).
12. Lensing, S. V. *et al.* DSBCapture: In situ capture and sequencing of DNA breaks. *Nat. Methods* **13**, (2016).
13. Bouwman, B. A. M. & Crosetto, N. Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling. *Genes (Basel).* **9**, (2018).
14. De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.* **18**, 950–955 (2011).
15. Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 1–7 (2015).
16. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
17. Bruna, A. *et al.* A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* **167**, 260-274.e22 (2016).
18. Rueda, O. M. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 (2019).
19. Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170 (2014).
20. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* **24**, 2022–32 (2014).
21. Mao, S.-Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.* **25**, 951–957 (2018).
22. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
23. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol. Cell* **52**, 25–36 (2013).
24. Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255 (2018).
25. Rodriguez, R. *et al.* A Novel Small Molecule That Alters Shelterin Integrity and Triggers a DNA-Damage Response at Telomeres. *J. Am. Chem. Soc.* **130**, 15758–15759 (2008).
26. Xu, H. *et al.* CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat. Commun.* **8**, 14432 (2017).
27. Biffi, G., Tannahill, D., Miller, J., Howat, W. J. & Balasubramanian, S. Elevated Levels of G-Quadruplex Formation in Human Stomach and Liver Cancer Tissues. *PLoS One* **9**, e102711 (2014).
28. McLuckie, K. I. E. *et al.* G-Quadruplex DNA as a Molecular Target for Induced Synthetic Lethality in Cancer Cells. *J. Am. Chem. Soc.* **135**, 9640–9643 (2013).
29. Zimmer, J. *et al.* Targeting BRCA1 and BRCA2 Deficiencies with G-Quadruplex-Interacting Compounds. *Mol. Cell* **61**, 449–460 (2016).

Figure 1 | Quantitative G4-ChIP-seq of PDTX reveals differentially enriched G4 DNA regions. a, Quantitative G4-ChIP-seq (qG4-ChIP-seq), exemplified with two different PDTX models (brackets: estrogen receptor status/integrative cluster status/PDTX model name). Blue and red indicates chromatin isolated from two different PDTX models, which is combined with *D. melanogaster* (reference) chromatin (black). **b**, Normalization strategy: technical and biological replicates of the same condition (nodes of the same color) get closer in space after normalization and samples of different PDTX models become more separated in space. **c**, Estimation of normalization factors using reference read coverage. To derive normalization factors, either all (Total recovery) reference sequencing reads are considered or only the ones in a predefined set of enriched regions (G4 region). Subsequent rescaling of the cumulative human cancer signal by the normalization factors is done across all experiments (see **Methods** section). **d**, Barplot of the improvement factors (I_F) quantifying normalization performance for all 22 PDTX models considering the reads in the enriched (G4 regions) and all recovered reads (Total recovery). Improvement factor evaluates the level of increased similarity (positive values) between technical replicates (black) and biological replicates (grey) (see **Methods**). **e**, Heatmap of human cancer normalized (reference normalized and input subtracted counts per million) qG4-ChIP-seq data for ΔG4Rs of two PDTX models -/10/AB863M (red) and +/8/STG143 (blue). **f**, Example genome browser views showing ΔG4Rs and normalized qG4-ChIP-seq track intensities of four technical replicates qG4-ChIP-seq for two PDTX models (red -/10/AB863M, blue +/8/STG143). PDTX annotation: estrogen receptor status/integrative cluster status/PDTX model name.

Figure 2 | G4 DNA prevalence in the genomic and transcriptomic architecture of PDTX breast cancer models. a, Heatmap visualizing similarity of ΔG4Rs from 22 different PDTX models. Hierarchical clustering is shown (Euclidean distance, ward.d2); color intensity and the size of the circle are proportional to the correlation coefficients. **b**, Left: Distribution of PDTX median fold-enrichments for ΔG4Rs and CG4Rs (common or unchanged qG4-ChIP-seq regions) in copy number aberrations (CNAs) relative to random permutation (n = 10 permutations for each of the 23 independent ΔG4R and CG4R maps); AMP = highly amplified regions, GAIN = amplified, NEUT = unchanged or neutral regions, HETD = heterozygous deletions, HOMD = homozygous deletions. Right: Distribution of ΔG4R and CG4R enrichments for single-nucleotide variants (SNVs) within the PDTX samples relative to random permutation (n = 10 permutations for each of the 16 independent ΔG4R and CG4R maps). Significances were calculated using a *t* test (Mann-Whitney) **** $P < 0.0001$ (exact, two-tailed). **c**, Genome annotation of ~26,000 PDTX G4-ChIP-seq regions. Black bars: proportion of G4 regions in particular genomic annotation, red bars: fold-enrichment over random (n = 5 permutations) genomic regions. Data are presented as mean values ± SD. **d**, Y-axis: high (blue), medium (red) and low (black) expressed genes. X-axis: For each PDTX, percentage of ΔG4Rs in the expressed promoters. Brackets indicate significant differences of ΔG4Rs spanning promoters of different expression levels. N = 20 PDTX ΔG4Rs were associated with n = 20 PDTX promoter expression levels. Significant differences were calculated using a Tukey multiple comparison test **** $P < 0.0001$ (adjusted *P* value). **e**, Distribution of the integral of ΔG4R signal intensities (median of cpm) in high, medium or low

expressed gene promoters (± 1 kb TSS) that are in AMP, GAIN, NEUT or HETD regions. N = 22 PDTX ΔG4R ChIP intensities measured in n = 22 PDTX promoter expression levels in the different CNA categories. Significances were calculated using a *t* test (Mann-Whitney) ** $P < 0.01$, * $P < 0.05$ (*P* values are exact, two-tailed). **f**, Scatter plots of four individual PDTX. Y-axis: Overlap of gene promoters (%) for distinct gene signatures of the 10 different integrative clusters[2] with ΔG4Rs. X-axis: the significance of the overlap relative to chance (Fisher). The expected IC classification for each PDTX model is highlighted in red. **g**, Fold-enrichment over random (bar plot) of ΔG4Rs or CG4Rs in 45 common breast cancer driver regions[2] relative to chance. Significance (color code): empirical *P* value (exact, two-tailed) obtained with 1,000 randomizations. Red dashed line indicates threshold of fold-enrichment over random. Box plot elements: center line, median; box limits, lower and upper quartiles; whiskers, lowest and highest value.

Figure 3 | G4 DNA regions reveal the activity of distinct transcription factor programs. Transcription factor correlation matrix heatmap (134 × 134). obtained by starting from the matrix of fold-enrichments over random of 22 ΔG4Rs at 134 breast cancer ChIP-seq transcription factor binding regions (TFBS) from ChIP-ATLAS database. Hierarchical clustering (ward.d2) of the correlations identifies TF sub-groups with similar correlation values across the 22 PDTX models; TF subgroups are highlighted by dashed line. Color intensity and the size of the circle are proportional to the correlation coefficients. Name of each TF subgroup relates to first and last TF within each subgroup. For each subgroup, there are boxplots of ΔG4R/TFBS fold-enrichments (blue) and of TF expression levels (TPM, red) stratified by various classifications of the PDTX models (ER+, ER-, membership to IC8/1, membership to IC10/9). N = 22 PDTX models were used to derive 22 ΔG4R maps and ChIP-ATLAS fold-enrichment values. Significances illustrated in box plots were calculated using the Mann-Whitney test **** $P < 0.0001$, ** $P < 0.01$, * $P < 0.05$ (exact *P* values, two-tailed). Box plot elements: center line, median; box limits, lower and upper quartiles; whiskers, lowest and highest value.

Figure 4 | G4 DNA levels predict response to G4-ligands. Scatterplots of ΔG4Rs (left-top, left-bottom, right-top) or highly amplified regions (AMP, right-bottom) levels (x-axis) against PDTC response (Area under the curve; AUC, y-axis) of PDTC models to G4-ligands with enhanced (PDS, CX-5461) and reduced (i-PDS) G4 affinities, see also **Methods**. Error bars reflect mean, upper and lower limit AUCs. N = 9 PDTC samples. Additionally, N = 3 PDTC samples were independently investigated. Spearman correlation (*r*) and significance (exact two-tailed *P* value for nonparametric correlation) are shown.

Methods:

Quantitative G4-ChIP-seq (qG4-ChIP-seq).

G4-ChIP-seq was performed as previously described[7] with the following adaptions for PDTX tissue. Briefly, *D. melanogaster* S2 cells were cultured in Schneider's *Drosophila* Medium (Thermo Fisher Scientific, cat no. R69007) containing 10% fetal bovine serum (FBS) Medium (Thermo Fisher Scientific, cat no. 10500064). To prepare spike-in *Drosophila* chromatin, 100 million cells were i) harvested by centrifugation, ii) fixed for 10 min in a solution of media

containing 10 % FBS, 1% formaldehyde (Thermo Fisher Scientific, cat no. 28908) and iii) quenched for 5 min by addition of 125 mM glycine (Fisher Scientific, cat no. 11545005). The cell pellet was washed with 10 ml PBS, pelleted by centrifugation and subsequently stored on ice for the lysis procedure. The 2-step chromatin lysis procedure was performed according to the Chromatrap procedure ("Spin column ChIP kit for qPCR v6.4"). 500 µl intact chromatin was sonicated into 100-500 bp fragments using a Bioruptor Plus (Diagenode cat. no. B02010003 with cooling) at 4°C. Sonicated chromatin was diluted with 1.5 ml lysis buffer (Chromatrap cat no. 100005) before snap-freezing as 25 µl aliquots. PDTX chromatin was prepared essentially as described in Schmidt *et al*. (Methods 2009)[30]. Briefly, a snap-frozen PDTX biopsy, ~1 cm^3, was transferred into a 50 ml falcon tube, on dry-ice, and crushed into smaller chunks on dry-ice using a scalpel followed by fixation for 20 min in 30 ml solution A, containing 1% formaldehyde, and then quenched for 5 min by adding 125 mM glycine. The supernatant of the pelleted tissue was discarded, and the pellet washed twice with 10 ml ice-cold PBS before resuspending in 1 ml PBS and transferred to a 1 ml glass Douncer (Fisher scientific, cat no. 11591295). 10 strokes were employed for each douncing step with a loose and then tight pestle, and the remaining tissue slurry was transferred to a 15 ml tube, centrifuged for 5 min at 2,500×g and subjected to lysis according to Schmidt et al.[30]. Briefly, after the 10 ml LB2 treatment and nuclei pelleting step, the pellet was resuspended in 500 µl LB3 and LB3-chromatin solution split into two Bioruptor TBX (Diagenode, cat no. C30010010-300) sonication tubes. The samples were sonicated until the desired fragment length (100-500 bp) was achieved. Finally, 50 µl of a 10% Triton X-100 LB3 solution was mixed with the sonicated solution and aliquoted into 50 µl aliquots before snap-freezing in liquid nitrogen. 5 µl of PDTX chromatin was quantified by Qubit using the "broad range kit" (Thermo Fisher Scientific, cat. no. Q32853). In each qG4-ChIP-seq reaction, 225 ng of PDTX chromatin, 102 ng of spike-in *Drosophila* chromatin and 2% RNaseA (Invitrogen, cat. no. AM2271) in blocking buffer (25 mM HEPES, pH 7.5, 10.5 mM NaCl, 110 mM KCl, 1 mM $MgCl_2$ and 1% BSA (Merck, cat. no. A7030) in Milli-Q water were mixed and incubated at 37°C for 30 min at 800 rpm. All PDTX chromatin batches containing a different concentration than 30 ng/µl were balanced to the same level, either by dilution with LB3 containing 1% Triton X-100 or by up-scaling the ChIP reaction. For PDTX chromatin with a concentration of 30 ng/µl, 7.5 µl of the PDTX chromatin was added to a solution containing 270 µl blocking buffer including 2% RNase A and 7.5 µl spike-in *Drosophila* chromatin. After RNaseA treatment, 15 µl of 2 µM BG4, prepared as described previously[31], was added to each qG4-ChIP-seq reaction and the reaction mixture shaken at 1,400 rpm at 16°C for 1 hour. Meanwhile, 65 µl of anti-FLAG magnetic beads (Sigma-Aldrich, cat. no. M8823) were washed three times with 650 µl of blocking buffer and resuspended in 1,300 µl blocking buffer. The pre-washed beads were incubated at 16°C at 1,400 rpm and 300 µl of pre-washed beads added to the reaction mixture after BG4 incubation. The reaction mixture with beads was incubated at 16°C for 1 hour at 1,400 rpm. Then, the beads were washed four times in 400 µl cold wash buffer (100 mM KCl, 0.1% Tween 20 and 10 mM Tris, pH 7.4 in Milli-Q water) in the cold room and twice at 37°C for 15 min at 1,400 rpm, followed by one cold wash on magnetic stand. The enriched chromatin on beads was resuspended in 75 µl TE buffer and 1 µl Proteinase K (Invitrogen, cat. no. AM2546) added. 6 µl Proteinase K was added to input sample which refers

to a qG4-ChIP-seq reaction mixture without BG4 and beads. The reaction mixture was incubated at 65°C for 3 hours at 1,400 rpm and purified using QIAGEN MinElute Kit (QIAGEN, cat. no. 28206).

Library preparation and sequencing. For 40 µl library preparation reaction, 3-5 ng of the ChIP or input DNA (Qubit high sensitivity kit, Thermo Fisher Scientific, cat. no. Q32854), 20 µl 2× tagmentation buffer (Illumina, cat. no. 15027866), 1.25 µl Tn5 enzyme (Illumina, cat no. 18027865) and nuclease-free water was incubated at 37°C for 20 min at 800 rpm. The reaction mixture was purified using QIAGEN MinElute Kit (QIAGEN, cat. no. 28206) according to the manufacturer's instruction and eluted in 20 µl EB buffer. To amplify the library, 20 µl of the DNA was then mixed with 25 µl NEB Next High Fidelity 2× PCR Master Mix (New England Biolabs, cat. no. N0541S), 2.5 µl Nextera index kit i5 primer (Illumina, cat. no. 15055290) and 2.5 µl Nextera index kit i7 primer (Illumina, cat. no. 15055290). The PCR program was as follows: 72°C for 5 minutes, 98°C for 30 seconds, followed by 8 cycles of 98°C for 10 seconds, 63°C for 30 seconds and 72°C for 1 minute. Libraries were quantified using a Bioanalyzer (Agilent) to estimate the average library size and concentration determined via Qubit HS. The library concentration was corrected for the library size using the following relationship: 1 ng/µl = 3nM = 500 bp. Samples were subjected to single-end sequencing with a read length of 75 bp on an Illumina NextSeq instrument.

Mapping, peak calling and peak processing. Fastq files were trimmed from adapters using *cutadapt* (options: -q 20 -O 3 http://dx.doi.org/10.14806/ej.17.1.200, ver: 1.16) and aligned[32] to a combined genome consisting of hg19 (*Homo sapiens*), dm6 (*D. melanogaster*) and mm10 (*Mus musculus*) with *bwa-mem* (ver. 0.7.17-r1188). Bam files were generated from the alignment with *samtools view* (options: -Sb -F2308 -q 10, ver: 1.8) and subsequently split by organisms to obtain 3 bam files for each sample. Duplicated reads were marked and removed using *picard MarkDuplicates* (ver: 2.20.3). For all organisms, the total sequencing coverage (total recovery) was quantified as the total number of unique reads aligning to the respective genome. Standard peak calling was performed for each sample using *MACS2* (ver. 2.1.2) with default options. For each human PDTX model, peak regions were considered positive if confirmed in 2 out of 4 technical replicates (multi2) with *bedtools v2.27.1 multiinter* (see **Supplementary Table 2**). All human confirmed G4-ChIP-seq peak files (multi2) of the 22 models were merged (*bedtools merge*) and regions more than 99 bp long retained to generate a single G4 DNA consensus of 26,103 G4 regions. Finally, the coverage of the samples was quantified using a consensus human set (*bedtools coverage*).

Reference normalization factor estimation and human ChIP signal normalization. For each PDTX biopsy, four technical qG4-ChIP experiments were performed and sequenced alongside one input chromatin (control), see also Life Sciences Reporting Summary. In each experiment, a similar amount of reference (*D. melanogaster*) chromatin from the same batch was added. To estimate PDTX normalization factors, reference coverage was determined at a pre-defined consensus consisting of 1,367 intervals (see **Supplementary Data 2**). The reference consensus set was defined from, and covers, G4-enriched regions observed in more than 110 pull-down

experiments. The normalization factor of each ChIP sample has been defined as the ratio between the maximum observed coverage (across all ChIP samples) and the individual sample coverage. Note that only ChIP experiments were used for this step (i.e. inputs are excluded and forced to 1). In turn, the outcome of the normalization approaches were tested using either the total recovery or the recovery at the G4 reference consensus regions. The normalization factors were then exported and used as input for a customized R script performing the normalization of the human signal. For each G4-ChIP-seq experiment, human signal (i.e. read coverage within human G4 consensus) was quantified by performing input subtraction and normalization with their respective reference reads and human library sizes. To assess if the normalization step has globally improved the experimental reproducibility, a quantitative parameter, the Improvement Factor I_F, was devised that measures both the increase (i.e. improvement) in data similarity between experiments corresponding to the same technical and biological samples and the increase dissimilarity between different samples. Specifically, the improvement factor of each biological sample has been estimated as:

$$I_F = \sum_{i}^{N} D_{ratio_{inter}} - \sum_{i}^{N} D_{ratio_{intra}}$$

$$D_{ratio} = \frac{Eucl\_dist\_after\_dm\_norm}{Eucl\_dist\_before\_dm\_norm}$$

Where:

- N: ChIP samples
- $Eucl\_dist\_after\_dm\_norm$: Euclidean similarity matrix computed on input subtracted, library size adjusted, drosophila normalized data and rescaled to its maximum value;
- $Eucl\_dist\_before\_dm\_norm$: Euclidean similarity matrix computed on input subtracted, library size adjusted data and rescaled to its maximum value;
- $D_{ratio_{inter}}$: similarity values among samples belonging to the same technical or biological group;
- $D_{ratio_{intra}}$: similarity values among samples not belonging to the same technical or biological group;

135 individual samples (ChIP + Input) were processed from 22 different PDTX models. Some PDTX models have more than one biological sample (**Supplementary Table 1**).

Guidelines to normalize G4-ChIP-seq data. During the optimization of the normalization procedure, we identified some general empirical criteria that can guide in deciding whether the reference (*D. melanogaster*) G4-ChIP-seq data can be used to normalize the human G4-ChIP-seq data, and whether it reduces technical noise and therefore has a beneficial outcome for the reproducibility of the replicated experiments.

1. Sequencing depth of the reference data per G4-ChIP-seq library should be around 5 M reads (after alignment and duplicate removal).

2. The number of detected peaks in the reference data of the 4 G4-ChIP-seq replicates (merge of confirmed peaks) should be in the range of several hundreds - 1,000. If no peaks are detected, it is not reliable to use the reference signal for normalization purposes.
3. The fraction of reads in the consensus reference peaks should exceed 0.5% of the total, ideally 1%. Consensus reference peaks are high-confidence regions that were consistently detected across many experiments (> 100) and are provided with this study.
4. Each G4-ChIP-seq library must have a fraction of reads at the consensus reference peaks at least 2x greater than the respective input library.
5. Technical and biological IF (average) should be positive, which indicates that reference normalization has improved the experimental reproducibility of the human G4-ChIP-seq replicates.

ΔG4Rs and CG4Rs. After the normalization step, differential G4-binding analysis was employed to identify differentially enriched G4 regions (ΔG4Rs), as described[6,7]. Both normalization and differential analysis are integrated into our workflow (see https://github.com/sblab-bioinformatics/qG4-ChIP-seq-of-breast-cancer-PDTX/).
Differential G4-binding was carried out with edgeR[33]. Initially, library size and *Drosophila*-normalized (human) read coverage within human G4 consensus regions were computed. Then, a generalized linear model with default parameters (negative binomial log-linear distribution of read counts) were used to assess regions with differential binding signal.
Specifically, the differential binding analysis compared each PDTX model to all the others. For each comparison, differential DNA G4 regions ΔG4R (i.e., regions specifically present in a given PDTX model) were defined as those satisfying the following criteria: $\log_2(CPM) \geq 0.6$ and FDR < 0.05. Constant G4 regions CG4Rs were defined as those that did not show any significant differential binding in any model i.e. regions that did not pass the filter in any of the individual comparisons.

PDTX gene expression data. Gene expression profiling of the individual PDTX models, except for STG316 for which part of the primary, patient-derived, tumor was used, was acquired via RNA-seq. For the AB863M model, a PDTX and a primary tumor sample were separately processed to generate RNA-seq data. Normalized TPMs have been quantified as explained in Georgopoulou & Callari et al., in preparation (EGA accession: EGAS00001001913) in all PDTX models except for PAR1006, PAR1022. For each model, expressed genes were stratified in 3 groups: high-, medium- and low-expression if they were belonging to the top, middle or bottom expression tertile, respectively.

ΔG4R PDTX stratification. Similarity across all the 22 PDTX ΔG4Rs was estimated using *bedtools jaccard*. Jaccard indexes of all pairwise comparisons resulted into a 22 × 22 matrix that have been explored with the Shiny-based web application https://asntech.shinyapps.io/intervene[34]. After loading the data, a pairwise intersection heatmap has been generated with the following settings: plot type: corrplot; correlation coefficient: Spearman; Agglomerative method: ward.d2; N. of cluster:3; distance matrix computation: Euclidean.

Somatic copy number aberration regions identification. Copy number segmentation was performed using the R package QDNAseq[20,35] on input (genomic background of qG4-ChIP-seq) BAM files sub-set to 5 million reads. A customized R script binned the genome into 100-kb windows, extracted the read-counts (*binReadCounts*), applied the QDNAseq filters, calculated (*estimateCorrection*), applied GC correction (*correctBins*), and then normalized and smoothed outliers. Finally, the copy-number profile of each PDTX model was segmented and exported. The copy number alterations regions were classified according to the following filtering criteria:

- highly amplified regions AMP: $\log_2$(fold ratio) > 0.75;
- amplified regions GAIN : $0.25 < \log_2$(fold ratio) ≤ 0.75;
- neutral regions NEUT : $-0.3 < \log_2$(fold ratio) ≤ 0.25
- heterozygous deletions HETD: $-1.4 < \log_2$(fold ratio) ≤ -0.3
- homozygous deletions HOMD: $\log_2$(fold ratio) ≤ -1.4.

ΔG4R and CG4R enrichment in CNA regions relative to random. The fold-enrichment of ΔG4R and CG4R was empirically estimated over randomly permutated genomic regions. First ΔG4Rs and CG4Rs were 10 times randomly shuffled across the genome (*bedtools shuffle*); then the number of PDTX ΔG4Rs and CG4Rs overlapping each of the CNA type was counted in the actual case and in the randomized case. For all CNA types in each PDTX model, the fold-enrichments were estimated as the ratio of the actual case over each of the ten random cases, see **Supplementary Table 3**. The distribution of all PDTX models' median fold-enrichments were then visualized in all individual CNA regions as a combined boxplot (**Fig. 2b**).

ΔG4R and CG4R enrichment for single nucleotide variants SNV relative to random. As in the case of CNA, the fold-enrichment of SNV at ΔG4R and CG4R was empirically estimated. After 5 random shufflings of the ΔG4R across the genome, we computed the fold-enrichment as the actual number of overlaps of G4 regions with SNVs over the average random case (i.e. average of number of overlaps obtained in each randomization) (see **Supplementary Table 4**). For all PDTX models, the analysis was conducted by comparing the model specific ΔG4R and SNVs maps. CG4Rs were compared to all PDTX SNV individually.

Genomic and G4-motif annotation and enrichment analysis of PDTX qG4-ChIP-seq peaks. PAVIS[36] was used to annotate 26,103 PDTX G4 human consensus regions. Fold-enrichment analysis was performed as described[6]. The consensus peaks were randomly shuffled across the genome 5 times. Fold-enrichments were computed as the ratio between the fraction of overlaps with each genomic feature in the actual case versus the corresponding average random fractions. G4 motifs were predicted and the presence in the PDTX G4 human consensus regions measured as previously reported[6].

Promoter - ΔG4R - gene expression. Promoter transcription start site (TSS) coordinates, 1 kb (±) from TSS, were generated for 22,483 genes using hg19[6]

https://www.gencodegenes.org/human/release_19.html. The fraction of ΔG4Rs overlapping high-, medium-, low-expression gene promoters was estimated. Significance was tested using the Tukey multiple comparisons test (GaphPad Prism7).

Promoter - G4 intensity - Gene expression - CNA analysis. For this, the human G4 drosophila-normalized intensity at ΔG4R overlapping promoters was considered. The distribution of this signal was visualized after stratifying promoter by CNA alteration (promoters overlapping to: AMP, GAIN, NEUT, HETD) and gene-expression groups (promoters belonging to: High-, Med.- Low- expression group).

ΔG4R and CG4R - Association to upregulated genes from integrative cluster signature IC. Promoter coordinates of differentially upregulated genes (Adjusted *P* value < 0.05; $\log_2$(fold-change) > 0.6) of each integrative cluster IC1-10[2] were extracted. For each PDTX model the association of ΔG4Rs and CG4Rs to upregulated promoters was quantified by computing the corresponding *P* value (-$\log_{10}$ *P* value) from the fisher test (intervene pairwise option fisher). For high significant associations, resulting in *P* values of 0, -$\log_{10}$ *P* value was set to 300. In addition, the fraction of the IC promoters having a ΔG4R overlapping was estimated (intervene pairwise option fraction). Fractions were transformed into percentage overlap and visualized together with *P* values as scatter plots (**Fig. 2f, Extended Data Fig. 2e**).

ΔG4R and CG4R – Association to 45 common driver regions. The genomic coordinates of 45 common breast cancer driver regions were taken from Curtis et al.[2] and lifted to hg19 (UCSC *liftover* tool) for assessment of the fold-enrichments of each PDTX ΔG4Rs and CG4Rs at those genomic locations using the Genomic Association Tester (GAT, https://gat.readthedocs.io/en/latest/contents.html).

Transcription factor binding site (TFBS) - ΔG4R enrichment analysis. The genomic fold-enrichment of each ΔG4R over transcription factors binding profiles from breast cancer, and breast immortalized cells, was determined using the ChIP-ATLAS enrichment web tool (https://chip-atlas.org/enrichment_analysis ,with the following parameters: Antigen class: TFs and others; Cell type Class: Breast; Threshold for Significance: 500; Select your data: individual ΔG4R in bed format; Select permutation to be compared: 100 random permutations). 22 result tables were obtained with each containing 12 tab-separated columns from which the following parameters selected: #2 Antigen name, #9 LogPvalue; #11 FoldEnrichment (FE). Rows were excluded where FE was "Inf". Selected enrichments with LogPvalue < -3 for each "antigen name" were averaged by their fold enrichments to give a table with 2 columns: "antigen name" and "relative averaged fold-enrichment". The 22 tables were then entered into a FE matrix with 134 TFs on the rows (where a TF has a FE value in at least 1/22 cases) and 22 columns representing FE in each of the 22 PDTX models. Next, we computed (a) the Spearman correlation of FE matrix 134 x 134 to assess the similarity between TF FE and (b) the Spearman correlation on the transposed FE matrix (22 × 22) to assess the similarity between PDTXs. The correlation matrix (a) for TF was additionally analyzed via hierarchical clustering (ward.d2). Seven subgroups of TFs were identified. For each subgroup of TF, boxplots were

generated for fold-enrichments and TF expression levels stratified by IC classification and/or ER status of the PDTX models (ER+, ER-, membership to IC8/, membership to IC10/9).

PDTX prepared into cell suspension (PDTC) and high throughput G4-ligand screen. PDTCs were prepared from cryopreserved xenograft fragments using a Tumour Dissociation Kit (MACS Miltenyi Biotec, cat no. 130-095-929) following the protocol for tough tumors. PDTCs were filtered through a 40 μm strainer and washed by centrifugation with complete growth media: RPMI-1640, supplemented with serum-free B27, EGF (20 ng/ml), FGF (20 ng/ml), Penicillin-Streptomycin (50 U/ml) and Gentamicin (5 μg/ml). Cells were plated to approximately 1.5 million cells/ml in 384-well plates. PDTCs were cultured for 24 hours and the PDTC compound screen was performed as described by Bruna et al.[17]. 9 different PDTX models (AB521M, HCI005, HCI009, STG139M, STG143, STG201, STG316, STG331, VHIO098) were treated with different concentrations of 3 different small molecules (i-PDS, PDS, CX-5461) for 14 days. 3 technical replicates were performed, and 3 models were analyzed within an independent screen. i-PDS and PDS were employed at 10, 3, 1, 0.3, 0.1, 0.03 and 0.01 μM. Due to solubility, CX-5461 treatments were 100x lower in comparison to i-PDS/PDS. Cell viability was assessed at day 0 and after 14 days of G4-ligand treatment using CellTiter-Glo 3D (Promega, cat no. G968). To correlate G4 ligand PDTC response with qG4-ChIP-seq signatures, area under the curves (AUC) were extracted from PDTC G4-ligand dose-response curves, fitted using isotonic regression, and scattered against ΔG4R/CNA or CG4R/CNA signatures.

Fluorescence quench equilibrium dissociation binding assay for PDS and i-PDS. The assay was performed as reported elsewhere[37]. The chemical synthesis of i-PDS is described in the supplementary information (**Supplementary Data 1**). Cy5-labelled oligonucleotides were analyzed as previously described[37] (see **Supplementary Table 5**).

Animal experiments and human research participants.
The research was done with the appropriate approval by the National Research Ethics Service, Cambridgeshire 2 REC (REC reference number: 08/H0308/178), which were all obtained under the appropriate Institutional Review Boards and transferred to Cambridge under Materials Transfer Agreements. All animal experiments were conducted in compliance with the rigorous Home Office framework of regulations (Project License 707679). Full names of the ethics committee: Revd. Dr. Derek Fraser, Mrs. Beth Midgley, Mr. Adam Garretty. The mouse strain NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ was used as PDTX avatar. Sex of mice: female. Age of mice: 3 month. Housing conditions for mice: 21 °C, Humidity: 55% ± 10%, light/dark cycle 12 h on, 12 h off. All patients were women with breast cancer. Patients were recruited by the Cambridge Cancer Centre. The covariate-relevant population characteristics of the breast cancer patients from the Cambridge Cancer Center (e.g. age, genotypic information) are reported in Supplementary Table 1.

Data Availability

The qG4-ChIP-seq data reported in this paper are available at GEO (NCBI repository), accession number GSE152216. Gene expression (RNA-seq) data of the PDTX models are available at the European Genome-Phenome Archive, accession number EGAS00001001913.

Code availability

Sample sheets describing the detailed experimental design are available at https://github.com/sblab-bioinformatics/qG4-ChIP-seq-of-breast-cancer-PDTX/. Details of data analysis have been deposited at https://github.com/sblab-bioinformatics/qG4-ChIP-seq-of-breast-cancer-PDTX/. An overview of all software tools for the processing of sequencing data is available (see **Supplementary Table 6**).

Methods-only References:

30. Schmidt, D. *et al.* ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48**, 240–248 (2009).
31. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–6 (2013).
32. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
33. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
34. Khan, A. & Mathelier, A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **18**, 287 (2017).
35. Chin, S.-F. *et al.* Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers. *Exp. Mol. Pathol.* **104**, 161–169 (2018).
36. Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D. & Li, L. PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics* **29**, 3097–3099 (2013).
37. Le, D. D., Di Antonio, M., Chan, L. K. M. & Balasubramanian, S. G-quadruplex ligands exhibit differential G-tetrad selectivity. *Chem. Commun.* **51**, 8048–8050 (2015).