ARTICLES
https://doi.org/10.1038/s41588-020-00750-6
nature genetics

Check for updates

OPEN

Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts

Xing Yi Woo[1,64,65], Jessica Giordano[2,3,64], Anuj Srivastava[1], Zi-Ming Zhao[1], Michael W. Lloyd[4], Roebi de Bruijn[5], Yun-Suhk Suh[6], Rajesh Patidar[7], Li Chen[7], Sandra Scherer[8], Matthew H. Bailey[8,9], Chieh-Hsiang Yang[8], Emilio Cortes-Sanchez[8], Yuanxin Xi[10], Jing Wang[10], Jayamanna Wickramasinghe[11], Andrew V. Kossenkov[11], Vito W. Rebecca[11], Hua Sun[12], R. Jay Mashl[12], Sherri R. Davies[12], Ryan Jeon[13], Christian Frech[13], Jelena Randjelovic[13], Jacqueline Rosains[13], Francesco Galimi[2,3], Andrea Bertotti[2,3], Adam Lafferty[14], Alice C. O'Farrell[14], Elodie Modave[15,16], Diether Lambrechts[15,16], Petra ter Brugge[5], Violeta Serra[17], Elisabetta Marangoni[18], Rania El Botty[18], Hyunsoo Kim[1], Jong-Il Kim[6], Han-Kwang Yang[6], Charles Lee[1,19,20], Dennis A. Dean II[13], Brandi Davis-Dusenbery[13], Yvonne A. Evrard[7], James H. Doroshow[21], Alana L. Welm[8], Bryan E. Welm[8,22], Michael T. Lewis[23], Bingliang Fang[24], Jack A. Roth[24], Funda Meric-Bernstam[25], Meenhard Herlyn[11], Michael A. Davies[26], Li Ding[12], Shunqiang Li[12], Ramaswamy Govindan[12], Claudio Isella[2,3,65], Jeffrey A. Moscow[27,65], Livio Trusolino[2,3,65], Annette T. Byrne[14,65], Jos Jonkers[5,65], Carol J. Bult[4,65], Enzo Medico[2,3,65]✉, Jeffrey H. Chuang[1,65]✉, PDXNET Consortium* and EurOPDX Consortium*

Patient-derived xenografts (PDXs) are resected human tumors engrafted into mice for preclinical studies and therapeutic testing. It has been proposed that the mouse host affects tumor evolution during PDX engraftment and propagation, affecting the accuracy of PDX modeling of human cancer. Here, we exhaustively analyze copy number alterations (CNAs) in 1,451 PDX and matched patient tumor (PT) samples from 509 PDX models. CNA inferences based on DNA sequencing and microarray data displayed substantially higher resolution and dynamic range than gene expression-based inferences, and they also showed strong CNA conservation from PTs through late-passage PDXs. CNA recurrence analysis of 130 colorectal and breast PT/PDX-early/PDX-late trios confirmed high-resolution CNA retention. We observed no significant enrichment of cancer-related genes in PDX-specific CNAs across models. Moreover, CNA differences between patient and PDX tumors were comparable to variations in multiregion samples within patients. Our study demonstrates the lack of systematic copy number evolution driven by the PDX mouse host.

Human tumors engrafted into transplant-compliant recipient mice (patient-derived xenografts (PDXs)) have advantages over previous model systems of human cancer (for example, genetically engineered mouse models[1,2] and cancer cell lines[3]) for preclinical drug efficacy studies because they allow researchers to directly study human cells and tissues in vivo[4–7]. Comparisons of

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [2]Department of Oncology, University of Turin, Turin, Italy. [3]Candiolo Cancer Institute, FPO-IRCCS, Turin, Italy. [4]The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA. [5]Netherlands Cancer Institute, Amsterdam, the Netherlands. [6]College of Medicine, Seoul National University, Seoul, Republic of Korea. [7]Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [8]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA. [9]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [10]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [11]The Wistar Institute, Philadelphia, PA, USA. [12]Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO, USA. [13]Seven Bridges Genomics, Charlestown, MA, USA. [14]Department of Physiology and Medical Physics, Centre for Systems Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland. [15]Center for Cancer Biology, VIB, Leuven, Belgium. [16]Laboratory of Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. [17]Vall d´Hebron Institute of Oncology, Barcelona, Spain. [18]Department of Translational Research, Institut Curie, PSL Research University, Paris, France. [19]Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, People's Republic of China. [20]Department of Life Sciences, Ewha Womans University, Seoul, Republic of Korea. [21]Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA. [22]Department of Surgery, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA. [23]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA. [24]Department of Thoracic and Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [25]Department of Investigational Cancer Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [26]Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [27]Investigational Drug Branch, National Cancer Institute, Bethesda, MD, USA. [64]These authors contributed equally: Xing Yi Woo, Jessica Giordano. [65]These authors jointly supervised this work: Xing Yi Woo, Claudio Isella, Jeffrey A. Moscow, Livio Trusolino, Annette T. Byrne, Jos Jonkers, Carol J. Bult, Enzo Medico, Jeffrey H. Chuang. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: enzo.medico@unito.it; jeff.chuang@jax.org

genome characteristics and histopathology of primary tumors and xenografts of various cancer types[8–14] have demonstrated that the biological properties of patient-derived tumors are largely preserved in xenografts. A growing body of literature supports their use in cancer drug discovery and development[15–17].

A caveat to PDX models is that intratumoral evolution can occur during engraftment and passaging[18–22]. Such evolution could potentially modify treatment response of PDXs with respect to the patient tumors (PTs)[19,23,24], particularly if the evolution were to systematically alter cancer-related genes. Recently, Ben-David et al.[23] reported extensive PDX copy number divergence from the PT of origin and across passages, based mainly on large-scale assessment of copy number alteration (CNA) profiles inferred from gene expression microarray data. They raised concerns about genetic evolution in PDXs as a consequence of mouse-specific selective pressures, which could impact the capacity of PDXs to faithfully model patient treatment response. Such results contrast with reports of observations of genomic fidelity of PDX models with respect to the originating PTs and from early to late passages by direct DNA measurements in several dozen PDX models[8,11,25].

Here, we resolve these contradicting observations by systematically evaluating CNA changes and the genes they affect during engraftment and passaging in a large, internationally collected set of PDX models, comparing both RNA- and DNA-based approaches. The data collected, as part of the US National Cancer Institute (NCI) PDX Development and Trial Centers Research Network (PDXNet) Consortium and EurOPDX Consortium, comprises PT and PDX samples from >500 models. Our study demonstrates that previous reports of systematic copy number divergence between PTs and PDXs are incorrect, and that there is high retention of copy number during PDX engraftment and passaging. This work also finely enumerates the copy number profiles in hundreds of publicly available models, which will enable researchers to assess the suitability of each for individualized treatment studies.

Results

Catalog of CNAs in PDXs. We have assembled CNA profiles of 1,451 unique samples (324 PT samples and 1,127 PDX samples), corresponding to 509 PDX models contributed by participating centers of the PDXNET, the EurOPDX Consortium and other published datasets[11,26] (see Methods, Supplementary Methods, Supplementary Table 1 and Supplementary Fig. 1). We estimated the copy number from five data types (single nucleotide polymorphism (SNP) array, whole-exome sequencing (WES), low-pass whole-genome sequencing (WGS), RNA sequencing (RNA-seq) and gene expression array data), yielding 1,548 tumor datasets including samples assayed on multiple platforms (see Methods, Supplementary Methods and Supplementary Data 1). Paired normal DNA, and in some cases paired normal RNA, were also obtained to calibrate WES and RNA-seq tumor samples.

The combined PDX data represent 16 broad tumor types derived from American, European and Asian patients with cancer (see Methods), with 64% ($n=324$) of the models having their corresponding PTs assayed and another 64% ($n=328$) having multiple PDX samples of either varying passages (P0–P21) or varying lineages from propagation into distinct mice (Fig. 1a and Supplementary Table 2). The distributions of PT and PDX samples across different tumor types, passages and assay platforms (Fig. 1b and Supplementary Figs. 2–12) show the wide spectrum of this combined dataset, which, to the best of our knowledge, is the most comprehensive copy number profiling of PDXs compiled to date (Supplementary Note 1). Additionally, our data include seven patients with multiple tumors collected either from different relapse time points or different metastatic sites, resulting in multiple PDX models derived from a single patient.

Comparison of CNA profiles from SNP array, WES and gene expression data. To compare the CNA profiles from different platforms in a controlled fashion, we assembled a dataset with matched measurements across multiple platforms (Supplementary Table 3 and Supplementary Figs. 13–17). Copy number calling has been reported to be noisy for several data types[27,28], and we observed that quantitative comparisons between CNA profiles are sensitive to: (1) the thresholds and baselines used to define gains and losses; (2) the dynamic range of copy number values from each platform; and (3) the differential impacts of normal cell contamination for different measurements. To control for such systematic biases, we assessed the similarity between two CNA profiles using the Pearson correlation of their $\log_2$[copy number ratio] values across the genome in 100-kilobase (kb) windows. Regions with discrepant copy number were identified as those with outlier values from the linear regression model (see Methods).

CNAs from WES are consistent with CNAs from SNP array data. As earlier studies reported that CNA estimates from WES data have more uncertainties than those from SNP arrays[29,30], we implemented a WES-based CNA pipeline and validated it against SNP array-based estimates[31,32] for matched samples. Copy number gain/loss segments (see Methods) from SNP arrays were of a higher resolution (Fig. 2a; median and mean segment sizes = 1.49 and 4.05 megabases (Mb) for SNP and 4.70 and 14.6 Mb for WES, respectively; $P<2.2\times10^{-16}$) and wider dynamic range (Fig. 2b; range of $\log_2$[copy number ratio] = −8.62–2.84 for SNP and −3.04–1.85 for WES; $P<2.2\times10^{-16}$). The difference in range is apparent in the linear regressions between platforms (Supplementary Fig. 18). These observations take into account the broad factors affecting CNA estimates across platforms, such as the positional distribution of sequencing loci, the sequencing depth of WES and the superior removal of normal cell contamination by SNP array CNA analysis workflows using SNP allele frequencies[33].

We observed strong agreement between SNP arrays and WES, with significantly higher Pearson correlation coefficients on matched samples than samples of different models (range = 0.913–0.957 for matched samples and 0.0366–0.354 for unmatched samples; $P=1.02\times10^{-6}$), with the exception of two samples that lacked CNA aberrations and were removed (Fig. 2c and Supplementary Figs. 13, 18 and 19). The discordant copy number regions largely correspond to small focal events (average size = 1.53 Mb) detectable by SNP arrays but missed by WES (Supplementary Fig. 18 and Extended Data Fig. 1a; see Methods). Hence, CNA profiling by WES is reliable in most regions in this small dataset, with 99% of the genome locations across the samples consistent with the values from SNP arrays (Supplementary Note 2). These PT-based observations are also applicable to PDXs given that mouse DNA is absent in SNP array signal and removed from WES reads[34–36].

Low accuracy for gene expression-derived CNA profiles. To compare the suitability of gene expression for quantifying evolutionary changes in CNA, we adapted the e-karyotyping method[23,37,38] for RNA-seq and gene expression array data (Supplementary Figs. 15 and 17; see Methods). Copy number segments calibrated by non-tumor expression were of higher resolution (Fig. 2a; median and mean segment sizes = 36.0 and 51.9 Mb for RNASEQ NORM versus 48.2 and 65.3 Mb for RNASEQ TUM ($P<2.2\times10^{-16}$) and 62.0 and 72.4 Mb for EXPARR NORM versus 80.1 and 85.2 Mb for EXPARR TUM ($P=2.20\times10^{-7}$), where RNASEQ and EXPARR relate to RNA-seq and gene expression array, respectively, and NORM and TUM relate to normalization by median expression of normal and tumor samples, respectively) and a wider dynamic range (Fig. 2b; range of $\log_2$[copy number ratio] = −2.07–2.17 for RNASEQ NORM versus −1.79–1.81 for RNASEQ TUM ($P<2.2\times10^{-16}$) and −1.40–1.89 for EXPARR NORM versus −1.13–1.59 for EXPARR

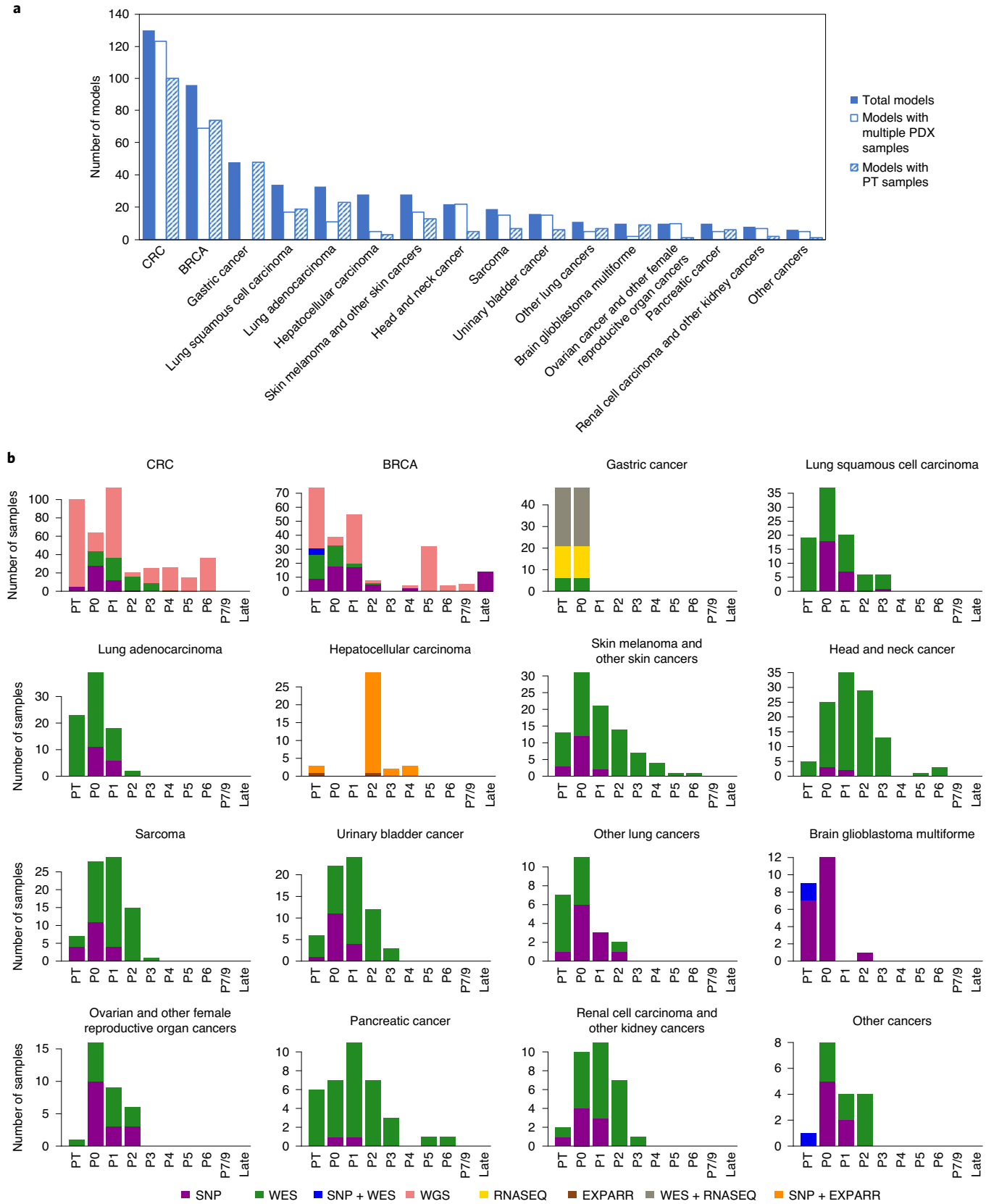


Fig. 1 | PDX datasets used for copy number profiling across 16 tumor types. a, Numbers of PDX models for each tumor type, with models also having multiple PDX samples or having matched PT samples specified. **b**, Distributions of datasets by passage number and assay platform for PTs and PDX samples, separated by tumor type. Late passages include P18, P19 and P21 samples.

TUM ($P=4.09\times10^{-7}$)) compared with segments calculated by calibration with tumor samples. These alternative expression calibrations yielded biased gain and loss frequencies (Supplementary Note 3 and Supplementary Fig. 20) and strong variability (Pearson correlation range = 0.218–0.943 for RNASEQ NORM versus TUM and 0.377–0.869 for EXPARR NORM versus TUM) in the CNA calls (Fig. 2c and Supplementary Fig. 21). This range of correlations was far greater than was observed in comparisons between the DNA-based methods ($P=9.37\times10^{-5}$ and $P=3.28\times10^{-7}$ relative to SNP versus WES). This indicates the problematic nature of RNA-based CNA calling with calibration by tumor samples, which has been used when normal samples are not available.

Furthermore, expression-based calling had segmental resolution an order of magnitude worse than the DNA-based methods (Fig. 2a and Supplementary Figs. 14–17; median and mean segment sizes = 3.45 and 14.0 Mb for WES versus 36.0 and 51.9 Mb for RNASEQ NORM ($P<2.2\times10^{-16}$) and 1.73 and 5.18 Mb for SNP versus 62.0 and 72.4 Mb for EXPARR NORM ($P<2.2\times10^{-16}$)). The range of detectable copy number values was also superior for DNA-based methods (Fig. 2b; range of $\log_2$[copy number ratio] = −6.00–5.33 for WES versus −2.07–2.17 for RNASEQ NORM ($P<2.2\times10^{-16}$) and −9.19–4.65 for SNP versus −1.40–1.89 for EXPARR NORM ($P<2.2\times10^{-16}$)). In addition, there was a lack of correlation between the expression-based and DNA-based methods (range = 0.0541–0.942 for WES versus RNASEQ NORM and 0.00517–0.921 for SNP versus EXPARR NORM) (Fig. 2c and Supplementary Figs. 22 and 23). CNA estimates after tumor-based expression normalization resulted in further discordance with DNA-based copy number results (range = −0.182–0.929 ($P=0.0468$) for WES versus RNASEQ TUM and −0.0274–0.847 ($P=2.20\times10^{-6}$) for SNP versus EXPARR TUM). Many focal copy number events detected by DNA-based methods, as well as some larger segments, were missed by the expression-based methods (Extended Data Fig. 1b–e). Representative examples illustrating the superior resolution and accuracy from DNA-based estimates are given in Fig. 2d (correlations are shown in Extended Data Fig. 2).

Concordance of PDXs with PTs and during passaging. Next, we adopted a pan-cancer approach to elucidate potential tumor type-independent copy number evolution in PDXs driven by the mouse host. We tracked the similarity of CNA profiles during tumor engraftment and passaging by calculating the Pearson correlation of gene-level copy number for samples measured on the same platform (see Methods, Extended Data Fig. 3 and Supplementary Figs. 24–60 and 62). All pairs of samples derived from the same PDX model were compared, yielding 501 PT–PDX pairs and 1,257 PDX–PDX pairs (Supplementary Note 4).

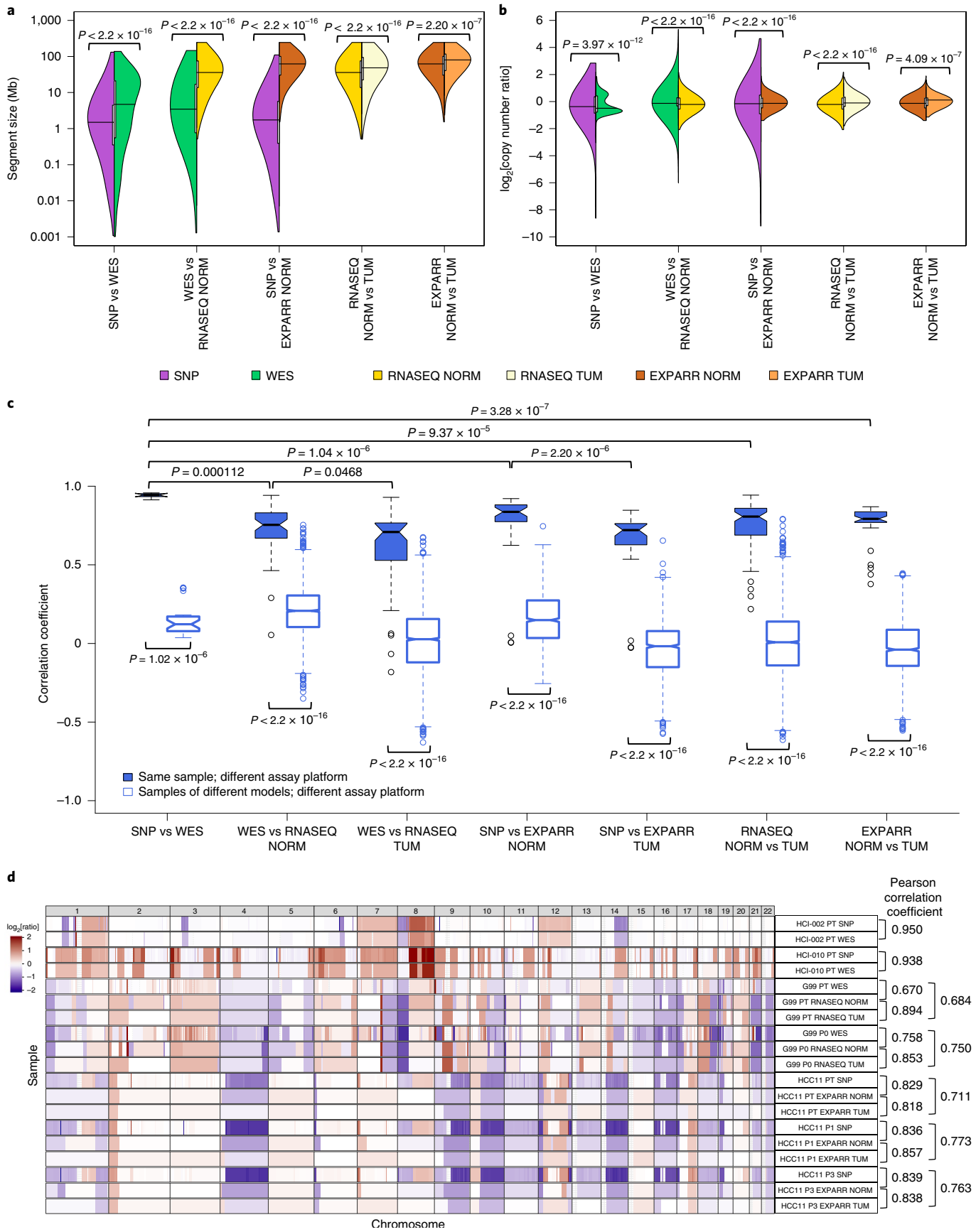
For all DNA-based platforms, we observed strong concordance between matched PT–PDX and PDX–PDX pairs, and this was significantly higher than between different models from the same tumor type and the same center ($P<2.2\times10^{-16}$) (Fig. 3a–c and correlation heatmaps in Supplementary Figs. 24–60). We observed no significant difference in the correlation values between PT–PDX and PDX–PDX pairs for SNP array data (median correlation = 0.950 for PT–PDX and 0.964 for PDX–PDX; $P>0.05$), although there were small but statistically significant shifts for WES (PT–PDX = 0.874; PDX–PDX = 0.936; $P=2.31\times10^{-16}$) and WGS data (PT–PDX = 0.914; PDX–PDX = 0.931; $P=0.000299$). PT samples have a smaller CNA range than their derived PDXs (median ratios for PT/PDX and PDX/PDX, respectively = 0.832 and 0.982 ($P=0.000120$) for SNP, 0.626 and 0.996 ($P<2.2\times10^{-16}$) for WES and 0.667 and 1.00 ($P<2.2\times10^{-16}$) for WGS; Supplementary Fig. 62b and Extended Data Fig. 4), which can be attributed to stromal DNA in PT samples diluting the CNA signal. In PDXs, the human stromal DNA is reduced[11,13]. The minimal effect for SNP array data confirms this interpretation as human stromal DNA contributions can be removed from SNP arrays based on allele frequencies of germline heterozygous sites, while such contributions to WES and WGS have higher uncertainties. We also performed intra-model comparisons using RNA-based approaches, which showed that the expression-based comparison of CNA profiles between PTs and PDXs can lead to overestimation of copy number changes during engraftment and passage (Supplementary Fig. 63 and Supplementary Note 5).

Late PDX passages maintain CNA profiles similar to early passages. Systematic mouse environment-driven evolution, if present, should reduce copy number correlations at each subsequent passage. However, we observed no apparent effect during passaging on the SNP, WES or WGS platforms (Fig. 3d–f and Extended Data Fig. 5). For example, the SNP data showed no significant difference between passages (Fig. 3d and Extended Data Fig. 5a). For those models having very late passages, there was a small but statistically significant correlation decrease compared with models with earlier passages ($P<8.98\times10^{-5}$; Extended Data Fig. 6b), indicating that some copy number changes can occur over long-term passaging (Supplementary Fig. 35). However, even at these late passages, the correlations with early passages remained high (median = 0.896). In any given comparison, only a small proportion of the genes were affected by copy number changes (median = 2.72%; range = 1.03–11.9%). Genes that are deleted and subsequently gained in the later passages (top left quadrant of regression plots; Extended Data Fig. 6a) suggest selection of pre-existing minor clones as the key mechanism in these regions. For WES and WGS data, more variability in the correlations can be observed (Fig. 3e,f and Extended Data Fig. 5b,c), probably due to a few samples having more stromal contamination or low aberration levels (Supplementary Fig. 62b and Extended Data Fig. 4). However, the lack of downward trend over passaging was also apparent in these sets (Supplementary Note 6).

PDX copy number profiles trace lineages. Next, we compared the similarity of engrafted PDXs of the same model with the same passage number. Surprisingly, we discovered that these pairs were not more similar than pairs of PDXs from different passage numbers (Fig. 3d,e, Extended Data Fig. 5 and Supplementary Note 7). Such similarity in correlations suggested that copy number divergence

Fig. 2 | Comparisons of resolution and accuracy for CNAs estimated using DNA- and expression-based methods. a, Pairwise comparisons of the distributions of CNA segment sizes as estimated using different measurement platforms in the validation dataset. CNAs are regions with ($|\log_2$[copy number ratio]$| \geq 0.1$). *P* values indicate the significance of the difference between distributions by two-sided Wilcoxon rank-sum test. vs, versus. **b**, Pairwise comparisons of the distributions of CNA segment $\log_2$[copy number ratio] values. *P* values were computed by two-sided Kolmogorov–Smirnov test. **c**, Distributions of Pearson correlation coefficients of median-centered $\log_2$[copy number ratio] values in 100-kb windows from CNA segments between pairs of samples estimated using different platforms. Samples with non-aberrant profiles in SNP array and WES data were omitted (5–95% inter-percentile range of $\log_2$[copy number ratio] < 0.3). *P* values were computed by two-sided Wilcoxon rank-sum test. In the box plots, the center line represents the median, the box limits are the upper and lower quantiles, the whiskers extend to 1.5× the interquartile range and the dots represent outliers. **d**, Examples of CNA profiles in comparisons of different platforms. Pearson correlation coefficients of CNA segments between pairs of samples are shown on the right. See Supplementary Table 3 for the number of samples per group. Examples of CNA profiles in comparisons of different platforms are shown; each sample ID is denoted by the model ID, passage number and platform used (see Supplementary Data 1).

might be associated with effects other than passaging. To further this analysis, we defined, for The Jackson Laboratory (JAX) SNP array and Patient-Derived Models Repository (PDMR) WES datasets, samples within a lineage as those differing only by consecutive serial passages, while we defined lineages as split when a tumor was divided and propagated into multiple mice (Fig. 3g). For the



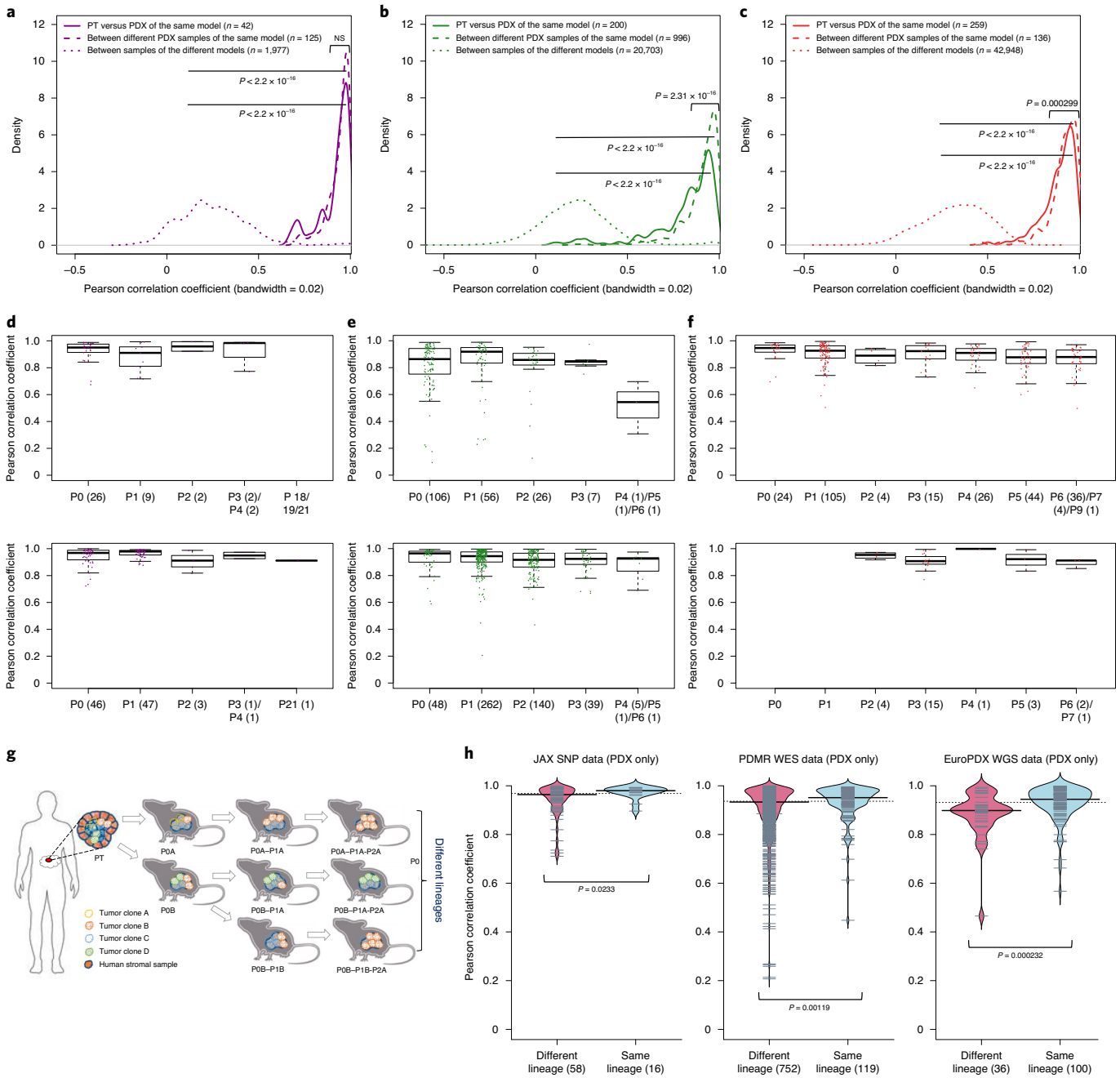


Fig. 3 | Comparisons of CNAs from PTs with early and late PDX passages. a–**c**, Distributions of Pearson correlation coefficients of gene-based copy number, estimated by SNP array (**a**), WES (**b**) and WGS (**c**) between: PT–PDX samples of the same model; PDX–PDX samples of the same model; and samples of different models from a common tumor type and contributing center. *P* values were computed by one-sided Wilcoxon rank-sum test ($P > 0.05$). Numbers of data points are indicated. NS, not significant. **d**–**f**, Distributions of Pearson correlation coefficients of gene-based copy number, estimated by SNP array (**d**), WES (**e**) and WGS (**f**) among PT and PDX passages of the same model. Comparisons relative to PT (top) and P0 (bottom) are shown (higher passages are shown in Extended Data Fig. 5). In the box plots, the center line represents the median, the box limits are the upper and lower quantiles, the whiskers extend to 1.5× the interquartile range and the dots represent all data points. **g**, Schematic of lineage splitting during passaging and expansion of tumors into multiple mice. This is a simplified illustration for passaging procedures in which different fragments of a tumor are implanted into different mice. **h**, Pearson correlation distributions for PDX sample pairs of different lineages and sample pairs within the same lineage, for (from left to right): JAX SNP array, PDMR WES and EuroPDX WGS datasets. *P* values were computed by one-sided Wilcoxon rank-sum test. For all box plots and violin plots, the numbers of pairwise correlations are indicated in the *x* axis labels.

EurOPDX colorectal cancer (CRC) and WGS breast cancer (BRCA) datasets, such lineage splitting was due only to cases with initial engraftment of different fragments of the PT (that is, PDX samples of different passages were considered as different lineages if they originated from different PT fragments). We observed lower correlation between PDX samples from different lineages compared with within a lineage (Fig. 3h; $P = 0.0233$ for SNP; $P = 0.00119$ for WES; $P = 0.000232$ for WGS), despite a majority of these pairwise comparisons exhibiting high correlation (>0.9) (Supplementary Notes 8 and 9). This suggests that lineage splitting is often responsible

for deviations in CNAs between samples, and that copy number evolution during passaging mainly arises from evolved spatial heterogeneity[24].

We further explored whether the stability of copy number during engraftment and passaging is affected by mutations in genes known to impact genome stability (see Methods). Overall, we observed that the presence of mutations in such genes does not lead to increased copy number changes during PDX engraftment and passaging (Supplementary Note 10 and Supplementary Fig. 66).

Genes with CNAs acquired during engraftment and passaging show no preference for cancer or treatment-related functions. Next, we investigated which genes tend to undergo copy number changes. Genes with changes during engraftment or during passaging were identified based on a residual threshold with respect to the improved linear regression[39] (see Methods and Extended Data Fig. 3). To test for functional biases, we compared CNA-altered genes with gene sets with known cancer- and treatment-related functions[40–43] (see Methods). We calculated the proportion of altered genes for sample pairs from each model across all platforms and tumor types. In agreement with the high maintenance of CNA profiles described above, we found the proportion of altered protein-coding genes to be low (median and IQR, respectively = 1.90 and 4.11% for PT–PDX pairs and 1.25 and 3.60% for PDX–PDX pairs; Fig. 4a). Only 8.78% of PT–PDX pairs and 4.53% of PDX–PDX pairs showed alteration of >10% of their protein-coding genes. We observed no significant increase ($P > 0.1$) in alterations among any of the cancer gene sets compared with the background of all protein-coding genes, for either the PT–PDX or PDX–PDX comparisons. This provides evidence that there is no systematic selection for CNAs in oncogenic or treatment-related pathways during engraftment or passaging. Next, we considered tumor-type-specific effects, focusing on tumor types with larger numbers of models to ensure statistical power. We observed no significant increase in alterations in tumor-type-specific driver gene sets significantly altered in TCGA[44–47] compared with the background ($P > 0.1$) for either PT–PDX or PDX–PDX comparisons (Fig. 4b and Supplementary Note 11).

Low recurrence of altered genes across models. We observed a very low recurrent frequency (Fig. 4c; see Methods), with only 12 and two genes recurring at >5% frequency for PT–PDX and PDX–PDX comparisons, respectively (Supplementary Table 4). No gene had a recurrence frequency higher than 8.96% (Supplementary Note 12). None of these recurrent genes overlapped cancer- or treatment-related gene sets, nor did they intersect genes ($n = 3$) reported by Ben-David et al.[23] to have mouse-induced copy number changes associated with drug response in the Cancer Cell Line Encyclopedia (CCLE)[48,49] database (Supplementary Note 12).

Absence of CNA shifts in 130 WGS PT, early-passage PDX and late-passage PDX trios. Next, we investigated whether recurrent CNA changes occur in PDXs in a tumor-type-specific fashion. To this aim, we analyzed further the WGS-based CNA profiles of large metastatic CRC and BRCA series, composed of matched trios of PT, PDX at early passage (PDX-early) and PDX at later passage (PDX-late). Genomic Identification of Significant Targets in Cancer (GISTIC)[50,51] analysis was applied separately to identify recurrent CNAs in each PT, PDX-early and PDX-late cohort of CRC and BRCA (see Methods and Supplementary Table 6). As expected, CRCs and BRCAs generated different patterns of significant CNAs but, within each tumor type, GISTIC profiles of the PT, PDX-early and PDX-late cohorts were virtually indistinguishable (Fig. 5a, Extended Data Fig. 7 and Supplementary Note 13), demonstrating no gross genomic alteration systematically acquired or lost in PDXs.

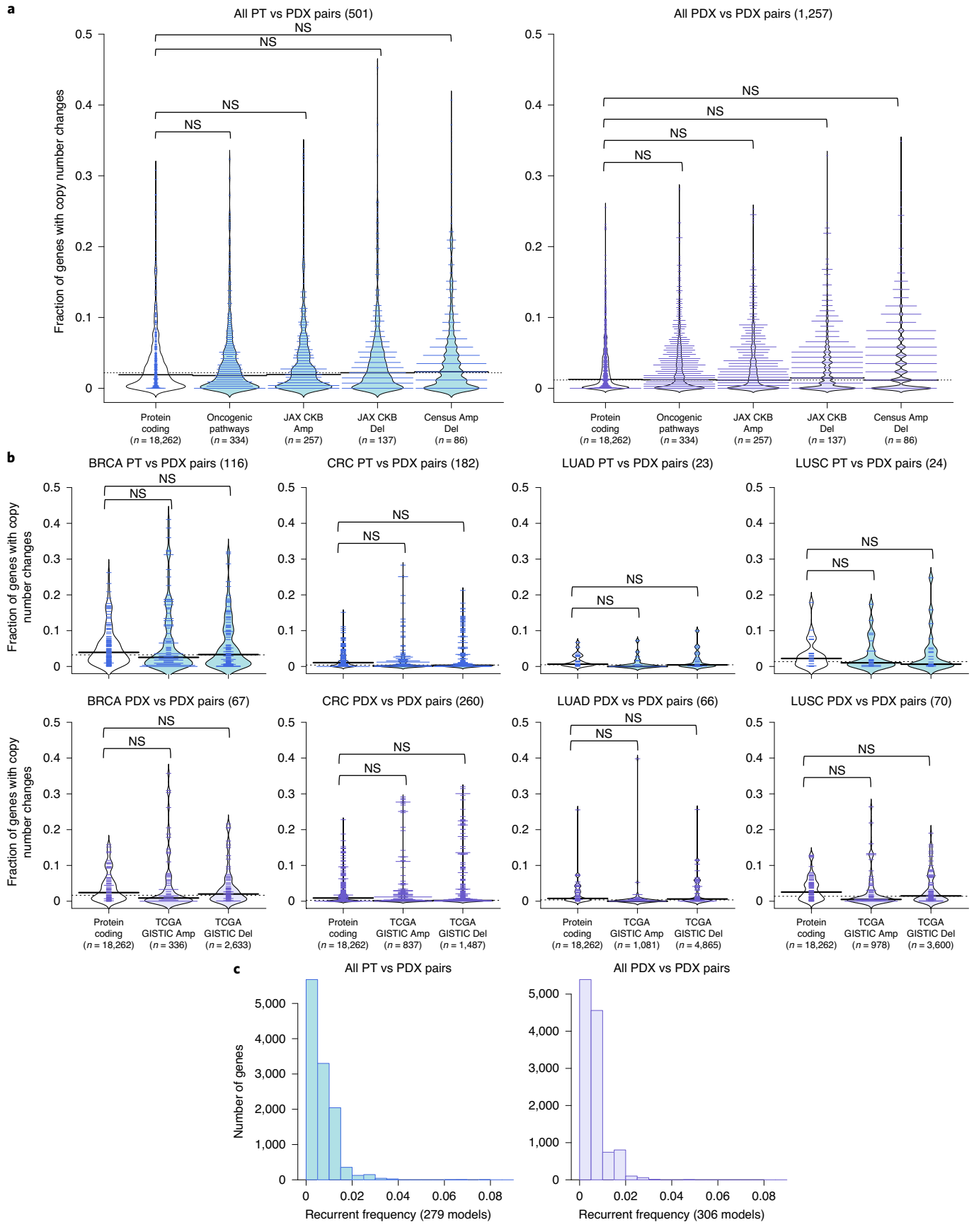
We then carried out gene-level analysis, where each gene was attributed the GISTIC score (G score) of the respective segment (Supplementary Table 7). In both the CRC and BRCA cohorts, gene-level G scores of the PTs were highly correlated with the respective PDX-early and PDX-late cohorts (Fig. 5b,c). Moreover, PT versus PDX correlations were comparable to PDX-early versus PDX-late correlations. To search for progressive shifts, we compared the change in G score (ΔG): (1) from tumor to PDX-early; and (2) from PDX-early to PDX-late. Correlations in these two ΔG values were absent or even slightly negative (bottom-right panels of Fig. 5b,c and Supplementary Note 13). Overall, these results confirmed the absence of systematic CNA shifts in PDXs, even under high-resolution gene-level analysis. To evaluate the possibility of systematic copy number evolution at the pathway level in these trios, we performed gene set enrichment analysis (GSEA)[52,53] using G scores to rank genes in each cohort (see Methods and Supplementary Note 14). For both CRC and BRCA, the normalized enrichment score (NES) profiles for the ~8,000 gene sets of PTs were highly correlated with the respective PDX-early and PDX-late cohorts (Fig. 5d,e). Moreover, PT versus PDX correlations were comparable to PDX-early versus PDX-late correlations. To search for progressive shifts, we calculated for each significant gene set ΔNES values between PT and PDX-early, as well as between PDX-early and PDX-late. Similar to what was observed for ΔG, correlations were absent or at most slightly negative (bottom-right panels of Fig. 5d,e), confirming the absence of systematic CNA-based functional shifts in PDXs.

CNA evolution across PDXs is no greater than variation in patient multiregion samples. As a reference for the treatment relevance of PDX-specific evolution, we compared this with the levels of copy number variation in multiregion samples of PTs. For this, we used copy number data from multiregion sampling of non-small-cell lung cancer from the TRACERx Consortium[54], performing analogous CNA correlation and gene analyses between multiregion pairs (Supplementary Fig. 69). We observed no significant differences in correlation ($P > 0.05$) between patient multiregion and lung cancer PT–PDX pairs, while PDX–PDX pairs in fact showed significantly better correlation than the multiregion pairs ($P < 0.05$; Fig. 6a), consistent across all lung cancer subtypes. Cancer gene set analyses confirmed these results, with multiregion samples showing greater differences than either PT–PDX or PDX–PDX comparisons, across

Fig. 4 | Cancer gene set analysis for copy number–altered genes during engraftment and passaging. a, Distribution of the proportion of altered genes between pairwise PT–PDX (left) and PDX–PDX comparisons (right) of the same model in various gene sets. Along the *x* axes from left to right are: protein-coding genes annotated by Ensembl; genes in oncogenic signaling pathways identified by TCGA; genes with copy number gain or overexpression (Amp) and genes with copy number loss or underexpression (Del) associated with therapeutic sensitivity or resistance or changes in drug response identified by JAX CKB; and genes from the Cancer Gene Census frequently altered by amplifications or deletions. CNA genes were identified by |residual| > 0.5 from a linear regression model. **b**, Distribution of the proportion of altered genes between pairwise PT–PDX (top) PDX–PDX comparisons (bottom) of the same model in various gene sets within BRCA, CRC, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) models. Along the *x* axes from left to right are: protein-coding genes annotated by Ensembl, followed by significantly amplified and deleted genes from TCGA GISTIC analysis for the corresponding tumor type. For all violin plots, *P* values were computed by one-sided Wilcoxon rank-sum test ($P > 0.1$). The numbers of pairwise comparisons are indicated above each plot, whereas the numbers of genes per gene set are indicated in the *x* axis labels. **c**, Recurrence frequencies of protein-coding genes with CNAs, |residual| > 1, across all models in PT–PDX (left) and PDX–PDX comparisons (right). Number of models are indicated in the *x* axis labels.

all cancer gene sets considered ($P < 0.05$; Fig. 6b and Extended Data Fig. 8). These results show that PDX-associated CNA evolution is no greater than what patients experience naturally within their tumors.

Our PDX collection also contains a few cases in which the PT was assayed at multiple time points (relapse/metastasis) or multiple metastatic sites, allowing for controlled comparison of intra-patient



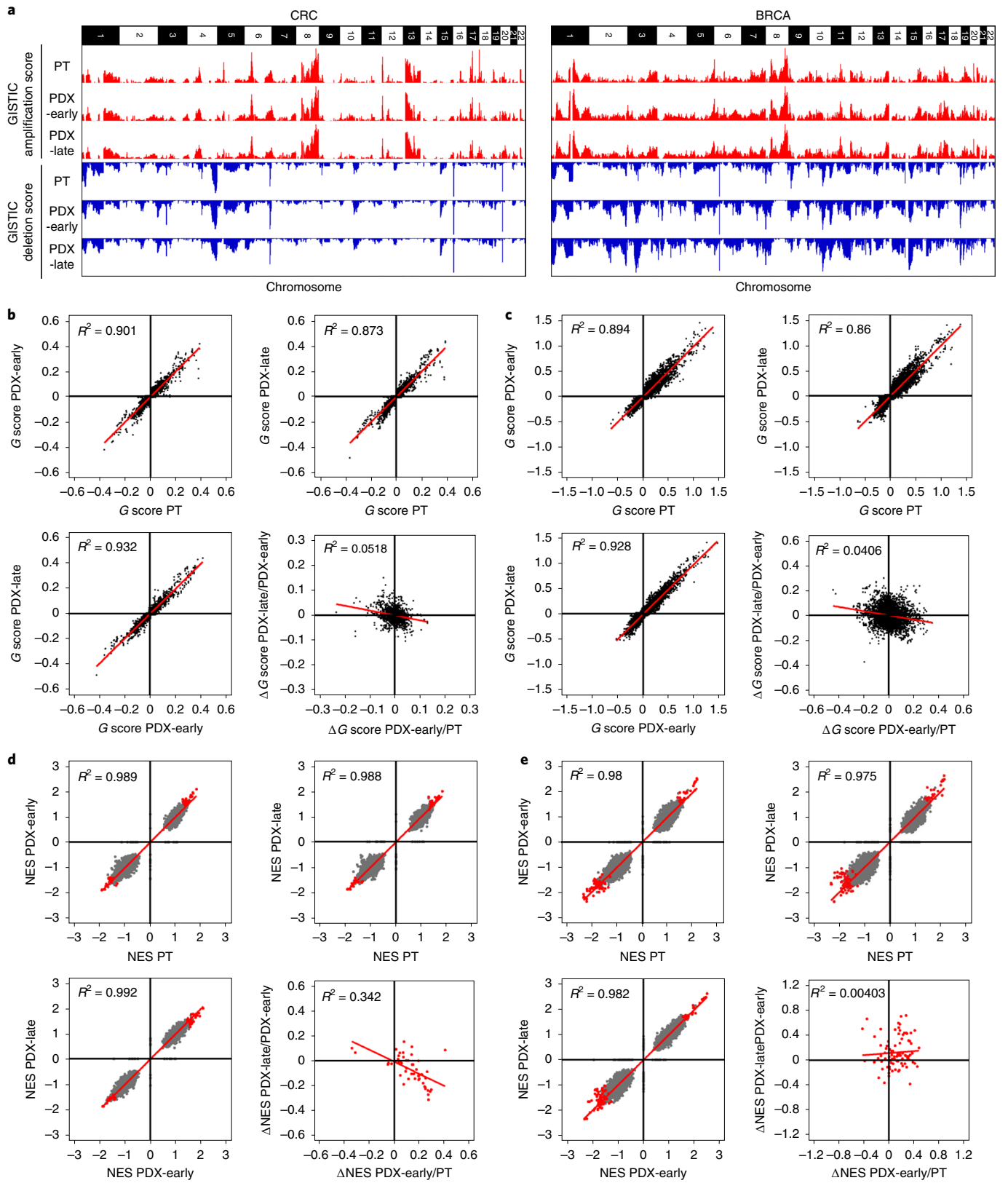


Fig. 5 | Absence of mouse-driven recurrent CNAs during engraftment and propagation of CRC and BRCA PDXs. a, Bar charts representing genome-wide *G* scores for amplifications and deletions in each of the three cohorts of CRC (left; 87 trios) and BRCA (right; 43 trios): PT, PDX-early (P0–P1 for CRC; P0–P2 for BRCA) and PDX-late (P2–P7 for CRC; P3–P9 for BRCA). **b**,**c**, Scatter plots comparing gene-level *G* scores between each of the three cohorts for CRC (**b**) and BRCA (**c**). The bottom-right panels of **b** and **c** show scatter plots comparing Δ*G* values from PT to PDX-early and from PDX-early to PDX-late. **d**,**e**, Scatter plots comparing GSEA NESs for gene sets between each of the three cohorts for CRC (**d**) and BRCA (**e**). The bottom-right panels of **d** and **e** show scatter plots comparing ΔNES from PT to PDX-early and from PDX-early to PDX-late. Gray data points represent all gene sets, whereas red data points represent gene sets significantly enriched in at least one of the three cohorts (that is, PT, PDX-early or PDX-late).

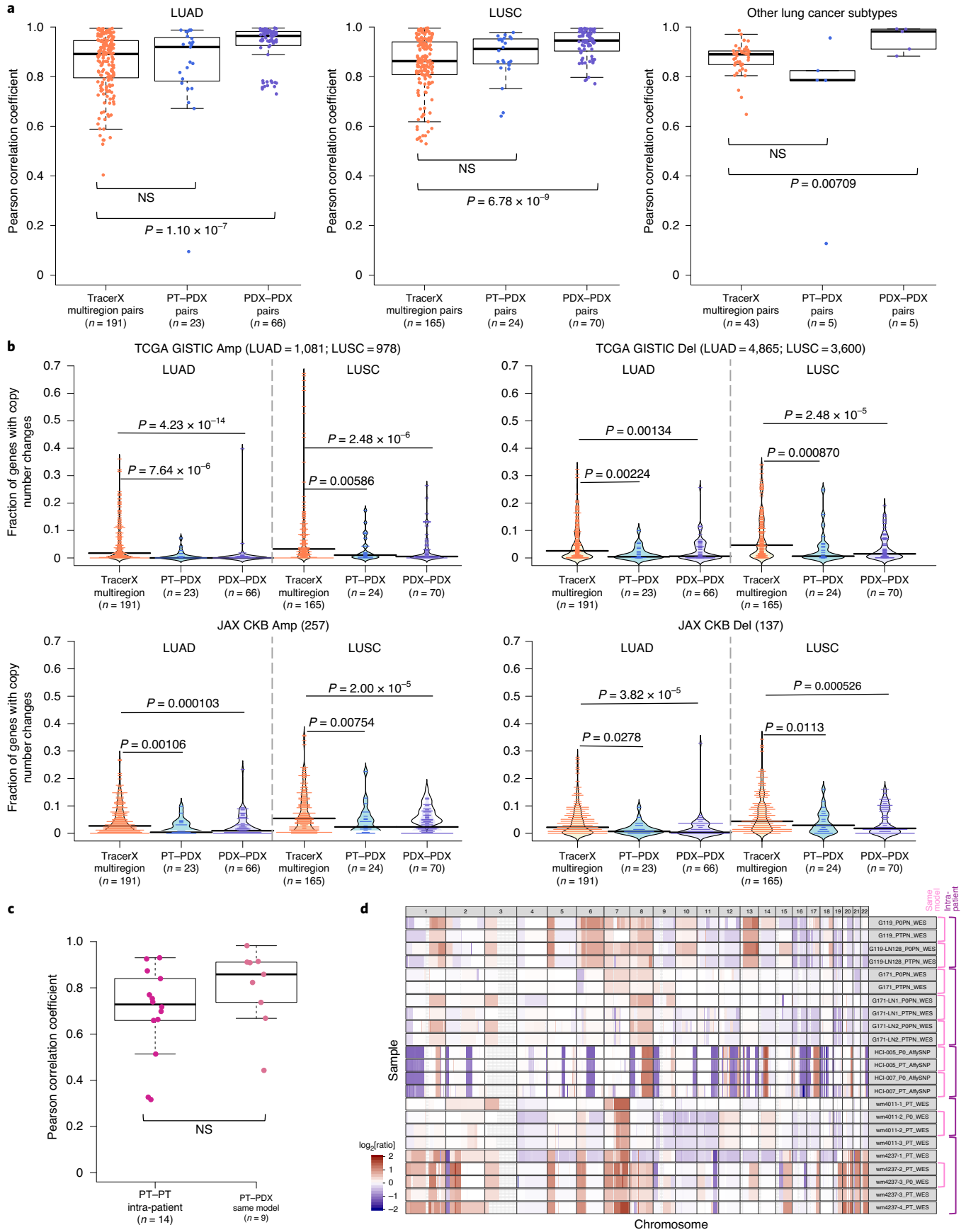

a
LUAD
LUSC
Other lung cancer subtypes
Pearson correlation coefficient
NS
P = 1.10 × 10−7
P = 6.78 × 10−9
P = 0.00709
TracerX multiregion pairs (n = 191)
PT–PDX pairs (n = 23)
PDX–PDX pairs (n = 66)
TracerX multiregion pairs (n = 165)
PT–PDX pairs (n = 24)
PDX–PDX pairs (n = 70)
TracerX multiregion pairs (n = 43)
PT–PDX pairs (n = 5)
PDX–PDX pairs (n = 5)
b
TCGA GISTIC Amp (LUAD = 1,081; LUSC = 978)
TCGA GISTIC Del (LUAD = 4,865; LUSC = 3,600)
JAX CKB Amp (257)
JAX CKB Del (137)
Fraction of genes with copy number changes
P = 4.23 × 10−14
P = 7.64 × 10−6
P = 2.48 × 10−6
P = 0.00586
P = 0.00134
P = 0.00224
P = 2.48 × 10−5
P = 0.000870
P = 0.000103
P = 0.00106
P = 2.00 × 10−5
P = 0.00754
P = 3.82 × 10−5
P = 0.0278
P = 0.000526
P = 0.0113
TracerX multiregion (n = 191)
PT–PDX (n = 23)
PDX–PDX (n = 66)
TracerX multiregion (n = 165)
PT–PDX (n = 24)
PDX–PDX (n = 70)
c
Pearson correlation coefficient
NS
PT–PT intra-patient (n = 14)
PT–PDX same model (n = 9)
d
Sample
Chromosome
log2[ratio]
Same model
Intra-patient
G119_P0PN_WES
G119_PTPN_WES
G119-LN128_P0PN_WES
G119-LN128_PTPN_WES
G171_P0PN_WES
G171_PTPN_WES
G171-LN1_P0PN_WES
G171-LN1_PTPN_WES
G171-LN2_P0PN_WES
G171-LN2_PTPN_WES
HCI-005_P0_AffySNP
HCI-005_PT_AffySNP
HCI-007_P0_AffySNP
HCI-007_PT_AffySNP
wm4011-1_PT_WES
wm4011-2_P0_WES
wm4011-2_PT_WES
wm4011-3_PT_WES
wm4237-1_PT_WES
wm4237-2_PT_WES
wm4237-3_P0_WES
wm4237-3_PT_WES
wm4237-4_PT_WES

Fig. 6 | Comparison of CNA variation during PDX engraftment and passaging with CNA variation among patient multiregion, tumor relapse and metastasis samples. a, Distributions of Pearson correlation coefficients of gene-based copy number for lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and other lung cancer subtypes, comparing different datasets. From left to right on the *x* axis, these include: multiregion tumor samples of the same patient from TRACERx ($n = 92$ PTs; $n = 295$ multiregion samples); PT–PDX samples of the same model; and PDX–PDX samples of the same model. *P* values were computed by two-sided Wilcoxon rank-sum test ($P > 0.05$). **b**, Distributions of the proportion of altered genes between multiregion tumor pairs from TRACERx, as well as PT–PDX and PDX–PDX pairs, for various gene sets for LUAD and LUSC. The gene sets and CNA thresholds are the same as in Fig. 4. TCGA GISTIC Amp/Del and JAX CKB Amp Del gene sets are shown (other gene sets are shown in Extended Data Fig. 8). *P* values were computed by one-sided Wilcoxon rank-sum test. The numbers of genes per gene set are indicated above each plot. **c**, Distributions of Pearson correlation coefficients of gene-based copy number between intra-patient PT pairs ($n = 14$; primary, relapse or metastasis) from the same patient ($n = 5$) and corresponding PT–PDX pairs (derived from the same model; a different PT sample from the same patient generates a different model) for the same set of patients. *P* values were computed by two-sided Wilcoxon rank-sum test ($P > 0.05$). For all box and violin plots, the numbers of pairwise comparisons are indicated in the *x* axis labels. In all box plots the center line represents the median, the box limits are the upper and lower quantiles, the whiskers extend to 1.5× the interquartile range and the dots represent all data points. **d**, CNA profiles of PT and PDX samples from patients with PDX models derived from multiple PT collections (primary, relapse and metastasis). Each sample ID is denoted by the model ID, passage number and platform used (see Supplementary Data 1).

variation versus PDX evolution (Supplementary Figs. 3, 4 and 7). Despite a lower median in correlations among intra-patient samples, the difference compared with CNA evolution during engraftment (PT–PDX) was not statistically significant ($P > 0.05$; Fig. 6c). CNA profiles for these samples are shown visually in Fig. 6d.

Discussion

Here, we have investigated the evolutionary stability of PDXs—an important model system for which there have been previous reports of mouse-induced copy number evolution. To better address this, we assembled a collection of CNA profiles of PDX models, comprising PDX models with multiple passages and their originating PTs. Our analysis showed the reliability of copy number estimation by DNA-based measurements over RNA-based inferences, which are substantially inferior in terms of resolution and accuracy (Supplementary Note 15). The importance of DNA measurements is supported by the inconsistent conclusions by two independent studies (Ben-David et al.[23,55] and Mer et al.[56]) on the same PDX expression array dataset by Gao et al.[15]. Ben-David et al. concluded that drastic copy number changes, driven by mouse-specific selection, often occur within a few passages. In contrast, Mer et al. reported high similarity between passages of the same PDX model based on direct correlations of gene expression, consistent with our findings in large, independent DNA-based datasets.

The copy number shifts inferred by Ben-David et al. were inherently impacted by major technical issues. First, the microarray signal for PT samples is diluted by introgressed human stromal cells, while in PDXs mouse stromal transcripts only hybridize to a fraction of the human probes[57]. Consequently, PT samples with substantial stromal content would display a reduced signal compared with the corresponding PDX, which can lead to an erroneous inference of systematic increase in aberrations during PDX engraftment when gain/loss regions are directly compared. Second, the mouse host microenvironment can affect the transcriptional profile of the PDX tumor[58] and the quantity of mouse stroma can vary across passages. This can result in variability in the expression signal, which can be wrongly inferred as copy number changes, both from the tumor itself and through cross-hybridization of mouse RNA to the human microarray. Although improved concordance in expression between PT and PDX can be achieved with RNA-seq with the removal of mouse reads[59,60], we observed that expression-based copy number inferences still have low resolution and robustness. Hence, many cancer-driving genes, which are found mainly in focal events with a size of 3 Mb or lower[61–64], cannot be evaluated for PDX-specific alterations. These issues are further worsened by the lack of tissue-matched normal gene expression profiles for calibration[37], which have been only intermittently available but can substantially impact copy number inferences. Because of these considerations, the question of how much PDXs evolve as a consequence of mouse-specific selective pressures cannot be adequately addressed by expression data.

The studies we have presented here take into account the above issues by the use of DNA data, as well as by assessing copy number changes by pairwise correlation/residual analysis to control for systematic biases, and they overall confirm the high retention of CNA profiles from PDX engraftment to passaging. We did observe larger deviations between PT–PDX than in PDX–PDX comparisons, although this was probably due to dilution of the PT signal by human stromal cells. Interestingly, we found that a major contributor to the differences between PDX samples is lineage-specific drift associated with the splitting of tumors into fragments during PDX propagation. This spatial evolution within tumors appears to affect sample comparisons more than time or the number of passages. This suggests that PDX expansion and passaging is the bottleneck of copy number evolution in PDXs, reflecting stochasticity in sampling within spatially heterogeneous tumors (Supplementary Note 16).

A challenge for evaluating any model system is that there is no clear threshold for genomic change that determines whether the model will still reflect patient response. Genetic variation among multiregion samples within a patient can shed light on this point[54,65–68] since the goal of a successful treatment would be to eradicate all of the multiple regions of the tumor. We found that the copy number differences between PT and PDX are no greater than the variations among multiregion tumor samples or intra-patient samples. Thus, concerns about the genetic stability of the PDX system are likely to be less important than the spatial heterogeneity of solid tumors themselves. This result is consistent with our results on lineage effects during passaging, which indicate that intratumoral spatial evolution is the major reason for genetic drift.

We observed no evidence for systematic mouse environment-induced selection for cancer- or treatment-related genes via copy number changes, although individual cases vary (see example in Extended Data Fig. 6c). Moreover, only a small fraction of sample pairs (2.44%; 43 out of 1,758) showed large CNA discordance (see Methods), suggesting that clonal selection out of a complex population is rare. These results indicate that the variations observed in PDXs are mainly due to spontaneous intratumoral evolution, rather than murine pressures (Supplementary Note 17).

In summary, our in-depth tracking of CNAs throughout PDX engraftment and passaging confirms that tumors engrafted and passaged in PDX models maintain a high degree of molecular fidelity to the original PTs, thus verifying their suitability for preclinical drug testing. At the same time, our study does not rule out that PDXs

will evolve in individual trajectories over time; thus, for therapeutic dosing studies, the best practice is to confirm the existence of expected molecular targets and obtain sequence characterizations in the cohorts used for testing as close to the time of the treatment study as is practical.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00750-6.

Received: 30 November 2019; Accepted: 18 November 2020;
Published online: 7 January 2021

References

1. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Dis. Models Mech.* **1**, 78–82 (2008).
2. Walrath, J. C., Hawes, J. J., Van Dyke, T. & Reilly, K. M. Genetically engineered mouse models in cancer research. *Adv. Cancer Res.* **106**, 113–164 (2010).
3. Hait, W. N. Anticancer drug development: the grand challenges. *Nat. Rev. Drug Discov.* **9**, 253–254 (2010).
4. Shultz, L. D., Ishikawa, F. & Greiner, D. L. Humanized mice in translational biomedical research. *Nat. Rev. Immunol.* **7**, 118–130 (2007).
5. Brehm, M. A., Shultz, L. D. & Greiner, D. L. Humanized mouse models to study human diseases. *Curr. Opin. Endocrinol. Diabetes Obes.* **17**, 120–125 (2010).
6. Hidalgo, M. et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* **4**, 998–1013 (2014).
7. Byrne, A. T. et al. Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat. Rev. Cancer* **17**, 254–268 (2017).
8. 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).
9. Reyal, F. et al. Molecular profiling of patient-derived breast cancer xenografts. *Breast Cancer Res.* **14**, R11 (2012).
10. Landis, M. D., Lehmann, B. D., Pietenpol, J. A. & Chang, J. C. Patient-derived breast tumor xenografts facilitating personalized cancer therapy. *Breast Cancer Res.* **15**, 201 (2013).
11. DeRose, Y. S. et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17**, 1514–1520 (2011).
12. Bankert, R. B. et al. Humanized mouse model of ovarian cancer recapitulates patient solid tumor progression, ascites formation, and metastasis. *PLoS ONE* **6**, e24420 (2011).
13. Julien, S. et al. Characterization of a large panel of patient-derived tumor xenografts representing the clinical heterogeneity of human colorectal cancer. *Clin. Cancer Res.* **18**, 5314–5328 (2012).
14. Lee, H. W. et al. Patient-derived xenografts from non-small cell lung cancer brain metastases are valuable translational platforms for the development of personalized targeted therapy. *Clin. Cancer Res.* **21**, 1172–1182 (2015).
15. Gao, H. et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
16. Hidalgo, M. et al. A pilot clinical study of treatment guided by personalized tumorgrafts in patients with advanced cancer. *Mol. Cancer Ther.* **10**, 1311–1316 (2011).
17. Tentler, J. J. et al. Patient-derived tumour xenografts as models for oncology drug development. *Nat. Rev. Clin. Oncol.* **9**, 338–350 (2012).
18. Eirew, P. et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2014).
19. Cho, S.-Y. et al. Unstable genome and transcriptome dynamics during tumor metastasis contribute to therapeutic heterogeneity in colorectal cancers. *Clin. Cancer Res.* **25**, 2821–2834 (2019).
20. Ding, L. et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
21. Giessler, K. M. et al. Genetic subclone architecture of tumor clone-initiating cells in colorectal cancer. *J. Exp. Med.* **214**, 2073–2088 (2017).
22. Sato, K. et al. Multiregion genomic analysis of serially transplanted patient-derived xenograft tumors. *Cancer Genom. Proteom.* **16**, 21–27 (2019).
23. Ben-David, U. et al. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* **49**, 1567–1575 (2017).
24. Kim, H. et al. High-resolution deconstruction of evolution induced by chemotherapy treatments in breast cancer xenografts. *Sci. Rep.* **8**, 17937 (2018).
25. Li, S. et al. Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4**, 1116–1130 (2013).
26. He, S. et al. PDXliver: a database of liver cancer patient derived xenograft mouse models. *BMC Cancer* **18**, 550 (2018).
27. Zare, F., Hosny, A. & Nabavi, S. Noise cancellation using total variation for copy number variation detection. *BMC Bioinformatics* **19**, 361 (2018).
28. Wineinger, N. E. & Tiwari, H. K. The impact of errors in copy number variation detection algorithms on association results. *PLoS ONE* **7**, e32396 (2012).
29. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
30. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
31. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
32. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
33. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
34. Woo, X. Y. et al. Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): challenges and guidelines. *BMC Med. Genet.* **12**, 92 (2019).
35. Evrard, Y. A. et al. Systematic establishment of robustness and standards in patient-derived xenograft experiments and analysis. *Cancer Res.* **80**, 2286–2297 (2020).
36. Conway, T. et al. Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, i172–i178 (2012).
37. Ben-David, U., Mayshar, Y. & Benvenisty, N. Virtual karyotyping of pluripotent stem cells on the basis of their global gene expression profiles. *Nat. Protoc.* **8**, 989–997 (2013).
38. Ben-David, U. et al. The landscape of chromosomal aberrations in breast cancer mouse models reveals driver-specific routes to tumorigenesis. *Nat. Commun.* **7**, 12160 (2016).
39. Motulsky, H. J. & Brown, R. E. Detecting outliers when fitting data with nonlinear regression—a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics* **7**, 123 (2006).
40. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).
41. Patterson, S. E. et al. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum. Genomics* **10**, 4 (2016).
42. Patterson, S. E., Statz, C. M., Yin, T. & Mockus, S. M. Utility of the JAX Clinical Knowledgebase in capture and assessment of complex genomic cancer data. *NPJ Precis. Oncol.* **3**, 2 (2019).
43. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
44. The Cancer Genome Atlas Networket al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
45. The Cancer Genome Atlas Networket al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
46. The Cancer Genome Atlas Research Networket al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
47. The Cancer Genome Atlas Research Networket al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
48. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
49. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
50. Beroukhim, R. et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012 (2007).
51. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
52. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
53. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
54. Jamal-Hanjani, M. et al. Tracking the evolution of non–small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
55. Ben-David, U., Beroukhim, R. & Golub, T. R. Genomic evolution of cancer models: perils and opportunities. *Nat. Rev. Cancer* **19**, 97–109 (2019).

56. Mer, A. S. et al. Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res.* **79**, 4539–4550 (2019).
57. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
58. Park, E. S. et al. Cross-species hybridization of microarrays for studying tumor transcriptome of brain metastasis. *Proc. Natl Acad. Sci. USA* **108**, 17456–17461 (2011).
59. Liu, Y. et al. Gene expression differences between matched pairs of ovarian cancer patient tumors and patient-derived xenografts. *Sci. Rep.* **9**, 6314 (2019).
60. Isella, C. et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat. Commun.* **8**, 15107 (2017).
61. Leary, R. J. et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA* **105**, 16224–16229 (2008).
62. Bierkens, M. et al. Focal aberrations indicate *EYA2* and *hsa-miR-375* as oncogene and tumor suppressor in cervical carcinogenesis. *Genes Chromosomes Cancer* **52**, 56–68 (2013).
63. Krijgsman, O., Carvalho, B., Meijer, G. A., Steenbergen, R. D. M. & Ylstra, B. Focal chromosomal copy number aberrations in cancer—needles in a genome haystack. *Biochim. Biophys. Acta Mol. Cell Res.* **1843**, 2698–2704 (2014).
64. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
65. De Bruin, E. C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
66. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
67. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
68. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).

PDXNET Consortium

Xing Yi Woo[1,64,65], Anuj Srivastava[1], Zi-Ming Zhao[1], Michael W. Lloyd[4], Rajesh Patidar[7], Li Chen[7], Sandra Scherer[8], Matthew H. Bailey[8,9], Chieh-Hsiang Yang[8], Emilio Cortes-Sanchez[8], Yuanxin Xi[10], Jing Wang[10], Jayamanna Wickramasinghe[11], Andrew V. Kossenkov[11], Vito W. Rebecca[11], Hua Sun[12], R. Jay Mashl[12], Sherri R. Davies[12], Ryan Jeon[13], Christian Frech[13], Jelena Randjelovic[13], Jacqueline Rosains[13], Dennis A. Dean II[13], Brandi Davis-Dusenbery[13], Yvonne A. Evrard[7], James H. Doroshow[21], Alana L. Welm[8], Bryan E. Welm[8,22], Michael T. Lewis[23], Bingliang Fang[24], Jack A. Roth[24], Funda Meric-Bernstam[25], Meenhard Herlyn[11], Michael A. Davies[26], Li Ding[12], Shunqiang Li[12], Ramaswamy Govindan[12], Jeffrey A. Moscow[27,65], Carol J. Bult[4,65], Jeffrey H. Chuang[1,65], Peter N. Robinson[1], Brian J. Sanderson[1], Steven B. Neuhauser[4], Lacey E. Dobrolecki[23], Xiaofeng Zheng[10], Mourad Majidi[24], Ran Zhang[24], Xiaoshan Zhang[24], Argun Akcakanat[25], Kurt W. Evans[25], Timothy A. Yap[25], Dali Li[25], Erkan Yucan[25], Christopher D. Lanier[25], Turcin Saridogan[25], Bryce P. Kirby[25], Min Jin Ha[28], Huiqin Chen[28], Scott Kopetz[29], David G. Menter[29], Jianhua Zhang[30], Shannon N. Westin[31], Michael P. Kim[32], Bingbing Dai[32], Don L. Gibbons[33], Coya Tapia[34], Vanessa B. Jensen[35], Gao Boning[36], John D. Minna[36], Hyunsil Park[36], Brenda C. Timmons[36], Luc Girard[36], Dylan Fingerman[11], Qin Liu[11], Rajasekharan Somasundaram[11], Min Xiao[11], Vashisht G. Yennu-Nanda[26], Michael T. Tetzlaff[37], Xiaowei Xu[37], Katherine L. Nathanson[38], Song Cao[12], Feng Chen[12], John F. DiPersio[12], Kian H. Lim[12], Cynthia X. Ma[12], Fernanda M. Rodriguez[12], Brian A. Van Tine[12], Andrea Wang-Gillam[12], Michael C. Wendl[12], Yige Wu[12], Matthew A. Wyczalkowski[12], Lijun Yao[12], Reyka Jayasinghe[12], Rebecca L. Aft[39], Ryan C. Fields[39], Jingqin Luo[39], Katherine C. Fuh[40], Vicki Chin[13], John DiGiovanna[13], Jeffrey Grover[13], Soner Koc[13], Sara Seepo[13], Tiffany Wallace[41], Chong-Xian Pan[42], Moon S. Chen Jr[42], Luis G. Carvajal-Carmona[43], Amanda R. Kirane[44], May Cho[44], David R. Gandara[44], Jonathan W. Riess[44], Tiffany Le[44], Ralph W. deVere White[44], Clifford G. Tepper[44], Hongyong Zhang[45], Nicole B. Coggins[45], Paul Lott[45], Ana Estrada[45], Ted Toal[45], Alexa Morales Arana[45], Guadalupe Polanco-Echeverry[45], Sienna Rocha[45],

Ai-Hong Ma[43], Nicholas Mitsiades[46,47], Salma Kaochar[46], Bert W. O'Malley[47], Matthew J. Ellis[23], Susan G. Hilsenbeck[23] and Michael Ittmann[48]

[28]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [29]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [30]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [31]Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [32]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [33]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [34]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [35]Department of Veterinary Medicine and Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [36]Hamon Center For Therapeutic Oncology, UT Southwestern Medical Center, Dallas, TX, USA. [37]Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA, USA. [38]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA. [39]Department of Surgery, Washington University School of Medicine in St. Louis, St. Louis, MO, USA. [40]Division of Gynecologic Oncology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA. [41]Center to Reduce Cancer Health Disparities, National Cancer Institute, Bethesda, MD, USA. [42]Department of Internal Medicine, Division of Hematology and Oncology, University of California, Davis, Sacramento, CA, USA. [43]Department of Biochemistry and Molecular Medicine, University of California, Davis, Sacramento, CA, USA. [44]UC Davis Comprehensive Cancer Center, University of California, Davis, Sacramento, CA, USA. [45]UC Davis Genome Center, University of California, Davis, Sacramento, CA, USA. [46]Department of Medicine, Baylor College of Medicine, Houston, TX, USA. [47]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [48]Department of Pathology, Baylor College of Medicine, Houston, TX, USA.

EurOPDX Consortium

Jessica Giordano[2,3,64], Roebi de Bruijn[5], Francesco Galimi[2,3], Andrea Bertotti[2,3], Adam Lafferty[14], Alice C. O'Farrell[14], Elodie Modave[15,16], Diether Lambrechts[15,16], Petra ter Brugge[5], Violeta Serra[17], Elisabetta Marangoni[18], Rania El Botty[18], Claudio Isella[2,3,65], Livio Trusolino[2,3,65], Annette T. Byrne[14,65], Jos Jonkers[5,65], Enzo Medico[2,3,65], Simona Corso[2,3], Alessandro Fiori[2,3], Silvia Giordano[2,3], Marieke van de Ven[5], Daniel S. Peeper[5], Ian Miller[14], Cristina Bernadó[17], Beatriz Morancho[17], Lorena Ramírez[17], Joaquín Arribas[17], Héctor G. Palmer[17], Alejandro Piris-Gimenez[17], Laura Soucek[17], Ahmed Dahmani[18], Elodie Montaudon[18], Fariba Nemati[18], Virginie Dangles-Marie[18], Didier Decaudin[18], Sergio Roman-Roman[18], Denis G. Alférez[49], Katherine Spence[49], Robert B. Clarke[49], Mohamed Bentires-Alj[50], David K. Chang[51], Andrew V. Biankin[51], Alejandra Bruna[52], Martin O'Reilly[52], Carlos Caldas[52], Oriol Casanovas[53], Eva Gonzalez-Suarez[53], Purificacíon Muñoz[53], Alberto Villanueva[53], Nathalie Conte[54], Jeremy Mason[54], Ross Thorne[54], Terrence F. Meehan[54], Helen Parkinson[54], Zdenka Dudova[55], Ales Křenek[55], Dalibor Stuchlík[55], Olivier Elemento[56], Giorgio Inghirami[56], Anna Golebiewska[57], Simone P. Niclou[57], G. Bea A. Wisman[58], Steven de Jong[58], Petra Kralova[59], Radislav Sedlacek[59], Elisa Claeys[60], Eleonora Leucci[60], Massimiliano Borsani[61], Luisa Lanfrancone[61], Pier Giuseppe Pelicci[61], Gunhild Mari Mælandsmo[62], Jens Henrik Norum[62] and Emilie Vinolo[63]

[49]Manchester Breast Centre, Division of Cancer Sciences, University of Manchester, Manchester, UK. [50]University Hospital of Basel, University of Basel, Basel, Switzerland. [51]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK. [52]Cancer Research UK Cambridge Institute, Cambridge Cancer Centre, Cambridge, UK. [53]Catalan Institute of Oncology, L'Hospitalet de Llobregat, Barcelona, Spain. [54]European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, UK. [55]Institute of Computer Science, Masaryk University, Brno, Czech Republic. [56]Weill Cornell Medical College, Cornell University, New York, NY, USA. [57]NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of Health, Luxembourg, Luxembourg. [58]University Medical Centre Groningen, Groningen, the Netherlands. [59]Czech Center for Phenogenomics, Institute of Molecular Genetics, Prague, Czech Republic. [60]TRACE PDX Platform, Katholieke Universiteit Leuven, Leuven, Belgium. [61]European Institute of Oncology, Milan, Italy. [62]Oslo University Hospital, Oslo, Norway. [63]Seeding Science SPRL, Limelette, Belgium.

Methods

Experimental details for sample collection, PDX engraftment and passaging, and array or sequencing. For details of sample collection, abbreviations of PDX model sources, PDX engraftment and passaging, and array/sequencing, see the Supplementary Methods.

Consolidating tumor types from different datasets. As the terminology of tumor types/subtypes by the different contributing centers was not consistent, we used the Disease Ontology database[69] (http://disease-ontology.org/), along with cancer types listed on the NCI website (https://www.cancer.gov/types) and in TCGA publications[70,71], to unify and group the tumor types/subtypes under broader terms, as shown in Fig. 1 and Supplementary Table 2.

CNA estimation methods. *SNP array.* The estimation of CNA profiles from SNP array was detailed previously[34]. In short, for Affymetrix Human SNP 6.0 arrays, PennCNV-Affy and Affymetrix Power Tools[72] were used to extract the B-allele frequency and log[R ratio] from the CEL files. Due to the absence of paired normal samples, the allele-specific signal intensity for each PDX tumor was normalized relative to 300 randomly selected sex-matched Affymetrix Human SNP 6.0 array CEL files obtained from the International HapMap Project[73]. For Illumina Infinium Omni2.5Exome-8 SNP arrays (version 1.3 and version 1.4 kits), the Illumina GenomeStudio software was used to extract the B-allele frequency and log[R ratio] from the signal intensity of each probe. The single sample mode of the Illumina GenomeStudio was used, which normalizes the signal intensities of the probes with an Illumina in-house dataset. The single tumor version of ASCAT[33] (version 2.4.3 for JAX SNP data and version 2.5.1 for SIBS SNP data) was used for GC correction, predictions of the heterozygous germline SNPs based on the SNP array platform, and estimation of ploidy, tumor content and allele-specific copy number segments. The resultant copy number segments were annotated with the $\log_2$[ratio of the total copy number relative to the predicted ploidy from ASCAT].

WES data. Aligned BAMs (see Supplementary Methods) were subset to the target region by GATK 4.0.5.1, and SAMTools[74] version 0.1.18 was used to generate the pileup for each sample. Pileup data were used for CNA estimation, as calculated with Sequenza[29] version 2.1.2. Both tumor and normal data, which utilized the same capture array, were used as input. pileup2seqz and GC-windows (-w 50) modules from sequenza-utils.py utility were used to create the native seqz format file for Sequenza and to compute the average GC content in sliding windows from the hg38 genome, respectively. We ran the three Sequenza modules with these modified parameters (sequenza.extract: assembly = 'hg38', sequenza.fit: chromosome.list = 1:23 and sequenza.results: chromosome.list = 1:23) to estimate the segments of copy number gains/losses. Finally, segments lacking read counts, in which ≥50% of the segment had zero read coverage, were removed. A reference implementation of this workflow (Supplementary Fig. 71) was developed and deployed in the Cancer Genomics Cloud by Seven Bridges (https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-workflow/ and https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-xenome-tumor-normal-workflow/).

Low-pass WGS data. For EuroPDX CRC liver metastasis data, raw copy number profiles for each sample were estimated using the QDNAseq[75] R package (version 1.20) by dividing the human reference genome into non-overlapping 50-kb windows and counting the number of reads (see Supplementary Methods) in each bin. Bins in problematic regions were removed[76]. Read counts were corrected for GC content and mappability using a LOESS regression, median normalized and $\log_2$ transformed. Values below −1,000 in each chromosome were floored to the first value greater than −1,000 in the same chromosome. Raw $\log_2$[ratio] values were then segmented using the ASCAT[33] algorithm implemented in the ASCAT R package (version 2.0.7). For EuroPDX BRCA tumors, raw copy number profiles were estimated for each sample by dividing the human reference genome into non-overlapping 20-kb windows and counting the number of reads (see Supplementary Methods) in each bin. Only reads with a mapping quality of at least 37 were considered. Bins within problematic regions (that is, multimapper regions) were excluded. Downstream analysis to estimate copy number was conducted as described above.

RNA-seq and gene expression microarray data. For expression-based copy number inference, we referred to the previous protocols for e-karyotyping and CGH-Explorer[37,38,77,78]. For each cancer type, expression values (see Supplementary Methods) of tumor samples and corresponding normal samples were merged in a single table, and gene identifiers were annotated with chromosomal nucleotide positions. Genes located on sex chromosomes were excluded. Genes with values below one transcript per million (TPM) (RNA-seq) or probeset $\log_2$ values below 6 (microarray) in more than 20% of the analyzed dataset were removed. Remaining gene expression values below the thresholds were respectively raised to 1 TPM or a $\log_2$ value of 6. In the case of multiple transcripts (RNA-seq) or probesets (microarray) per gene, the one with the highest median value across the entire dataset was selected. According to the e-karyotyping protocol, the sum of squares of the expression values relative to their median expression across all samples was calculated for each gene, and 10% of the most highly variable genes were removed. For each gene, the median $\log_2$[expression] value in normal samples was subtracted from the $\log_2$[expression] value in each tumor sample and subsequently input into CGH-Explorer. For tumor-only datasets, the median $\log_2$[expression] value in the same set of tumor samples was instead subtracted. The preprocessed expression profiles of each sample were individually analyzed using CGH-Explorer (http://heim.ifi.uio.no/bioinf/Projects/CGHExplorer/). Piecewise constant fit analysis was carried out to call copy number according to parameters previously reported[23]: least allowed deviation = 0.25; least allowed aberration size = 30; winsorize at quantile = 0.001; penalty = 12; and threshold = 0.01.

Statistical methods. All statistical analyses for data comparison were performed using either a one- or two-tailed Wilcoxon rank-sum test, a two-tailed Kolmogorov–Smirnov test or a one-tailed Wilcoxon signed-rank test.

Filtering and gene annotation of copy number segments. Copy number segments with a $\log_2$[copy number ratio] estimated from the various platforms were processed in the following steps (Extended Data Fig. 3). Segments <1 kb were filtered based on the definition of CNA[79]. In addition, SNP array segments had to be covered by more than ten probes, with an average probe density of one probe per 5 kb. The copy number segments were then binned into 10-kb windows to derive the median $\log_2$[copy number ratio], which was subsequently used to re-center the copy number segments. Median-centered copy number segments were visualized using IGV[80] version 2.4.13 and GenVisR[81] version 1.16.1. The median-centered copy number of genes was calculated by intersecting the genome coordinates of copy number segments with the genome coordinates of genes (Ensembl Genes 93 for human genome assembly GRCh38 and Ensembl Genes 96 for human genome assembly GRCh37). In the case where a gene overlapped multiple segments, the most conservative (lowest) estimate of copy number was used to represent the copy number of the entire intact gene.

Comparison of copy number gains and losses. For the comparison of resolution, the range of copy number values and the frequency of gains and losses between different platforms and analysis methods, we defined copy number gain or loss segments as $\log_2$[copy number ratio] > 0.1 (for gain) and $\log_2$[copy number ratio] < −0.1 (for loss).

Correlation of CNA profiles. The overall workflow to compare CNA profiles is shown in Extended Data Fig. 3. PDX samples without passage information were omitted in the following downstream analysis. The copy number segments were binned into 100-kb windows or smaller using BEDTools[82] version 2.26.0, and the variance of the $\log_2$[copy number ratio] and 5–95% inter-percentile range of the $\log_2$[copy number ratio] values across all of the bins were calculated as a measure of the degree of aberration for each CNA profile. A non-aberrant profile results in a low variance or range. While variance can be biased for CNA profiles with small segments of extreme gains or losses, we preferred use of the 5–95% inter-percentile range of $\log_2$[copy number ratio] to identify samples with a low degree of aberration, such that a narrow range indicates that ≥90% of the genome has very low-level gains and losses. The similarity of two CNA profiles is quantified by the Pearson correlation coefficient of the $\log_2$[copy number ratio] of 100-kb windows binned from segments or genes between two samples. Gene-based and segment-based (100-kb-window) correlations were highly similar (data not shown). Using correlation avoided the issue of making copy number gain and loss calls based on thresholds. Sample-based variations in the baseline due to median normalization and the range in copy number values could introduce further inconsistencies in gain and loss calls between samples. Such variations are further impacted by sample-specific variation in human stromal contamination or the sensitivity of copy number detection by different platforms. As median centering of each CNA profile approximates normalization by the sample ploidy, we confirmed that, in general, ploidy (estimated from ASCAT analysis of SNP array samples) had no association with the copy number correlation values (Pearson's product moment correlation = 0.0248; $P > 0.05$). However, one caveat of our approach is that it cannot distinguish genome-wide multiplication of ploidy between samples, as the correlation statistic is invariant to such genome-wide transformations. As such, we cannot assess whether ploidy changes occur between samples of a given model.

Comparison of CNA profiles between different platforms. The copy number segments of each pair of data were intersected and binned into 100-kb windows or smaller using BEDTools. The Pearson correlation coefficient and linear regression model were calculated for the $\log_2$[copy number ratio] of the windows. Windows with discrepant copy numbers were identified by outliers of the linear regression model defined by |studentized residual| > 3. These outlier windows were mapped to their corresponding segments to identify the size of CNA events that were discordant between the different copy number estimation methods. The proportion of the genome-discordant CNA was calculated from the summation of the outlier windows.

Identification of genes with CNA between different samples of the same model. To compare the CNA profiles between different samples (PT or PDX) of the same

model, the Pearson correlation coefficient and linear regression model were calculated for the $\log_2$[copy number ratio] of the genes for each pair of data. Before that, deleted genes with a $\log_2$[copy number ratio] of <-3 were rescaled to -3 to avoid large shifts in the correlation coefficient and linear regression model due to extremely negative values on the log scale. Extreme outliers of the linear regression model defined by |studentized residual| > 3 were removed to derive an improved linear regression model[39] not biased by a few extreme values. Genes with copy number changes between the samples were identified by the difference in $\log_2$[copy number ratio] relative to the improved linear regression model of |standard residual| < 0.5. We also removed some samples with low correlation due to sample mislabeling as they displayed high correlation with samples from other models. We also omitted samples with low correlation values (<0.6), which resulted from non-aberrant CNA profiles in genomically stable tumors (5–95% inter-percentile range of $\log_2$[copy number ratio] < 0.3; Supplementary Fig. 62).

Identification of aberrant sample pairs with highly discordant CNA profiles. Aberrant CNA profiles were identified based on the 5–95% inter-percentile range of $\log_2$[copy number ratio] > 0.5, for both samples. Sample pairs with a Pearson correlation of < 0.6 were selected as having highly discordant CNA profiles between them.

Association of mutations with copy number correlations. Mutational calls for each WES sample used in this study were obtained using a tumor normal variant calling workflow developed for PTs and PDXs[35]. Subsequently, genes with either germline or somatic variants that passed through the quality filters (FILTER = PASS or germline) and IMPACT = MODERATE or HIGH by SnpEff (version 4.3) annotation were labeled as mutated. Otherwise, they were labeled as wild type. For SNP array and WGS data, we collected the mutational status (wild type or mutated) of *TP53*, *BRCA1* and *BRCA2* per model where available, which may or may not have been obtained from the exact same tumor samples used in this study. For the JAX SNP array dataset, variant calls (tumor only) were made from various targeted sequencing approaches (TruSeq Amplicon Cancel Panel, JAX Cancer Treatment Profile panel and WES). The workflow and filtering criteria to call mutations is described elsewhere[34]. For the HCI SNP array data, mutations were obtained from WES (unpublished data) and were filtered for frameshift, inframe, missense, nonsense and splice-site mutations. For the BCM SNP array data, mutational status was obtained from clinical samples by immunohistochemistry or Sequenom[83] (unpublished data). For the WGS data, mutations were obtained from WES or targeted panel sequencing[84] (unpublished data), and high-quality and probable functional mutations were retained. For each sample pair with copy number correlations, the mutational status of *TP53* or *BRCA* was obtained for each individual sample for the WES data, while the mutational status was available on a per-model basis for the SNP and WGS data. *BRCA* was labeled as mutated when either *BRCA1* or *BRCA2* was mutated. For mutations in DNA repair genes[85] from the WES data, each pair of samples was classified as mutated if any DNA repair gene was reported to be mutated in either sample.

Annotation with gene sets with known cancer- or treatment-related functions. A low copy number change threshold (|$\log_2$[copy number ratio] change| > 0.5) was selected to include genes with subclonal alterations. Copy number–altered genes (|residual| > 0.5) were annotated by various gene sets with cancer- or treatment-related functions gathered from various databases and publications (Extended Data Fig. 3):

(1) Genes in ten oncogenic signaling pathways curated by TCGA that were found to be frequently altered in different cancer types[40];
(2) Genes with a gain in copy number or expression or a loss in copy number or expression that conferred therapeutic sensitivity, resistance or an increase/decrease in drug response from the JAX Clinical Knowledgebase (CKB)[41,42] (based on literature curation (https://ckbhome.jax.org/; as of 18 June 2019).
(3) Genes with evidence of promoting oncogenic transformation by amplification or deletion from the Cancer Gene Census[43] (COSMIC version 89); and
(4) Significantly amplified or deleted genes in TCGA cohorts of BRCA[44], CRC[45], lung adenocarcinoma[46] and lung squamous cell carcinoma[47] by GISTIC analysis, which identified significantly altered genomic driver regions that can be used to differentiate between tumor types and subtypes.

Identification of genes with recurrent copy number changes. A stringent CNA threshold (|$\log_2$[copy number ratio] change| > 1.0 with respect to the linear regression model) was selected to distinguish genes with a possible functional impact. Genes with |residual| > 1.0 with respect to the improved regression linear model (without discriminating gain or loss) were selected for each pairwise comparison between different samples of the same model. Pairwise cases in which genes were deleted in both samples ($\log_2$[copy number ratio] ≤ -3) were omitted. The recurrent frequency for each gene across all models was calculated on a model basis such that genes with a copy number between multiple pairs of the same model were counted once. This avoided bias towards models with many samples of similar copy number changes between the different pairs.

Drug response analysis using CCLE data. We developed a pipeline to evaluate gene copy number effects on drug sensitivity[86,87] by using CCLE[48,88] cell line genomic and drug response data (Cancer Therapeutics Response Portal version 2). We downloaded the CCLE drug response data from the Cancer Therapeutics Response Portal (www.broadinstitute.org/ctrp) and CCLE gene-level CNA and gene expression data from the DepMap data portal (public_19Q1_gene_cn.csv and CCLE_depMap_19Q1_TPM.csv; https://depmap.org/portal/download/). For CCLE drug response data, we used the area-under-the-concentration-response curve (AUC) sensitivity scores for each cancer cell line and each drug. In total, we collected gene-level $\log_2$[copy number ratio] data derived from the Affymetrix SNP 6.0 platform from 668 pan-cancer CCLE cell lines, with a total of 545 cancer drugs tested. With the CCLE gene-level CNA and AUC drug sensitivity scores, we performed gene–drug response association analyses for genes with recurrent copy number changes. Pearson correlation *P* values between each gene's $\log_2$[copy number ratio] and each drug's AUC score across all cell lines were calculated, and *q* values were calculated by multiple-testing Bonferroni correction. Significant gene CNA–drug associations were kept (*q* value < 0.1) to further evaluate gene expression and drug response associations. If a gene's expression was also significantly correlated with AUC drug sensitivity scores, particularly in the same direction (either positively or negatively correlated) as the gene CNA–drug association, that gene would be considered as significantly correlated with drug response based on both its CNA and gene expression.

GISTIC analysis of WGS data. We carried out GISTIC analysis to identify recurrent CNAs by evaluating the frequency and amplitude of observed events. To obtain perfectly matching and comparable PT–PDX cohorts for GISTIC analysis, CRC trios in which at least one sample displayed non-aberrant CNA profiles were excluded from the analysis, resulting in a total of 87 triplets. The GISTIC[51] algorithm (GISTIC 2.0 version 6.15.28) was applied on the segmented profiles using the GISTIC GenePattern module (https://cloud.genepattern.org/), with default parameters and the genome reference files Human_Hg19.mat for the EuroPDX CRC data and hg38.UCSC.add_miR.160920.refgene.mat for the EuroPDX BRCA data. For each dataset, GISTIC provides separate results (including segments, *G* scores and false discovery rate *q* values) separately for recurrent amplifications and recurrent deletions. Deletion *G* scores were assigned negative values for visualization. We observed that the *G* score range was systematically lower in PT cohorts, which was probably the result of the dilution of CNA by normal stromal DNA. In contrast, human stromal DNA in PDX samples was lower or negligible. To account for this difference in gene-level *G* scores, PDXs at early and late passages were scaled with respect to PT gene-level *G* score values using global linear regression, separately for amplification and deletion outputs.

GSEA of WGS data. To assess the biological functions associated with the recurrent alterations detected by the GISTIC analysis, we performed GSEAPreranked analysis[52,53] (GSEA version 3.0) on gene-level *G* score profiles for both amplifications and deletions. In particular, we applied the algorithm with 1,000 permutations on various gene set collections from the Molecular Signatures Database[89,90] (MSigDB version 6.2): (1) hallmark; (2) curated (chemical and genetic perturbations and canonical pathways); (3) Gene Ontology (biological processes, molecular functions and cellular components); and (4) oncogenic signatures. These collections were composed of 50, 4,762, 5,917 and 189 gene sets, respectively. We also included gene sets with known cancer- or treatment-related functions, as described above. We noted that multiple genes with contiguous chromosomal locations—typically in recurrent amplicons—generated spurious enrichment for gene sets consisting of multiple genes of adjacent positions, while very few or none of them had a significant *G* score. To avoid this confounding issue, we only considered the leading-edge genes (that is, those genes with an increasing NES up to its maximum value that contribute to the GSEA significance for a given gene set). The leading-edge subset can be interpreted as the core that accounts for the gene set's enrichment signal (http://software.broadinstitute.org/gsea). We included a requirement that the leading-edge genes passing the *G* score significance thresholds based on a GISTIC *q* value of 0.25 (Supplementary Table 8 and Extended Data Fig. 7) make up at least 20% of the gene set. This 20% threshold was chosen as the minimum threshold at which gene sets assembled from TCGA-generated lists of genes with recurrent CNAs in CRC or BRCA were identified as significant in GSEA (see Supplementary Table 9). Finally, gene sets with a NES of >1.5 and a false discovery rate *q* value of <0.05 that passed the leading-edge criteria were considered significantly enriched in genes affected by recurrent CNAs.

Ethics. All of the xenograft studies were completed in accordance with animal research ethics regulations. For details, see the Supplementary Methods.

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

Data availability

Copy number calls from all datasets are available in Supplementary Data 1, and these were used for all of the figures. Raw sequence data for these calls are a

combination of previously described sources (notably, the publicly available NCI Patient-Derived Models Repository; pdmr.cancer.gov) and newly sequenced data. New sequence data from PDXNet are being shared as part of the NCI Cancer Moonshot initiative through the Cancer Data Service. For further details, contact the corresponding authors. The SNP array data generated by The Jackson Laboratory can be requested via the Mouse Models of Human Cancer Database (tumor.informatics.jax.org). The WGS data generated by EurOPDX can be made available by directly contacting the EurOPDX Consortium (dataportal.europdx.eu or e-mail to E. Medico). Other publicly available data used in the analyses include those deposited to the Gene Expression Omnibus (GSE90653, GSE3526 and GSE33006) and ArrayExpress (E-MTAB-1503-3), as well as CCLE cell line genomic and drug response data (Cancer Therapeutics Response Portal version 2), and MSigDB version 6.2 and TRACERx non-small cell lung cancer data (https://doi.org/10.1056/NEJMoa1616288).

Code availability

We have used well-established computational sequence analysis and statistical analysis techniques, so no code is provided. Full descriptions of all of the analysis techniques are provided in the Methods. The implementation of the copy number estimation workflow from WES data is deployed in the cancer genomics cloud at SevenBridges (https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-workflow/ and https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-xenome-tumor-normal-workflow/).

References

69. Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2018).
70. The Cancer Genome Atlas Networket al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
71. Abeshouse, A. et al. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171**, 950–965.e28 (2017).
72. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
73. International HapMap Consortium The International HapMap Project. *Nature* **426**, 789–796 (2003).
74. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
75. 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**, 2022–2032 (2014).
76. Desmedt, C. et al. Uncovering the genomic heterogeneity of multifocal breast cancer. *J. Pathol.* **236**, 457–466 (2015).
77. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nat. Commun.* **7**, 12144 (2016).
78. Lingjaerde, O. C., Baumbusch, L. O., Liestol, K., Glad, I. K. & Borresen-Dale, A. L. CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 821–822 (2005).
79. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
80. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
81. Skidmore, Z. L. et al. GenVisR: genomic visualizations in R. *Bioinformatics* **32**, 3012–3014 (2016).
82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
83. Zhang, X. M. et al. A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73**, 4885–4897 (2013).
84. Coussy, F. et al. A large collection of integrated genomically characterized patient-derived xenografts highlighting the heterogeneity of triple-negative breast cancer. *Int. J. Cancer* **145**, 1902–1912 (2019).
85. Riaz, N. et al. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun.* **8**, 857 (2017).
86. Adams, D. J. et al. NAMPT is the cellular target of STF-31-like small-molecule probes. *ACS Chem. Biol.* **9**, 2247–2254 (2014).
87. Viswanathan, V. S. et al. Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* **547**, 453–457 (2017).
88. Stransky, N. et al. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
89. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
90. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

Acknowledgements

Support for the PDXNET consortium included funding provided by the National Institutes of Health (NIH) to the PDXNet Data Commons and Coordination Center (NCI U24-CA224067), the PDX Development and Trial Centers (NCI U54-CA224083, NCI U54-CA224070, NCI U54-CA224065, NCI U54-CA224076, NCI U54-CA233223 and NCI U54-CA233306) and the NCI Cancer Genomics Cloud (HHSN261201400008C and HHSN261201500003I). JAX PDX resource data were supported by the NCI of the NIH under the JAX Cancer Center NCI Grant (award number P30CA034196). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The genomic data for JAX PDX tumors used in this work were generated by JAX Genome Technologies and the Single Cell Biology Scientific Service. The development of PDX models and the generation of data from Seoul National University, in collaboration with JAX, was supported by the Korean Healthcare Technology R&D project through the Korean Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number HI13C2148). C.L. is supported in part by operational funds from The First Affiliated Hospital of Xi'an Jiaotong University. C.L. was a distinguished Ewha Womans University Professor, supported in part by the Ewha Womans University Research grant of 2018–2019. Sample procurement and next-generation sequencing at the Huntsman Cancer Institute were performed at the Genomics and Bioinformatics Analysis and Biorepository and Molecular Pathology shared resources, respectively, supported by NCI P30CA042014. SNP arrays were performed at the University of Utah Health Sciences Center Genomics Core. We are grateful to M. P. Klein for assistance with the SNP array data. M.H.B. is funded by the NIH under Ruth L. Kirschstein National Research Service Award Institutional Training Grant 5T32HG008962-05. M.T.L. is supported by a P30 Cancer Center Support Grant (CA125123) and a Core Facility Support Grant from the Cancer Research and Prevention Initiative of Texas (RP170691). PDX generation and WES at the University of Texas MD Anderson Cancer Center were supported by the University of Texas MD Anderson Cancer Center Moon Shots Program, funded by Specialized Program of Research Excellence grant CA-070907. J.A.R. is supported in part by the NIH/NCI through The University of Texas MD Anderson Cancer Center's Cancer Center Support Grant CA-016672—the Lung Program and Shared Core Facilities, the Specialized Program of Research Excellence grant CA-070907 and the Lung Cancer Moon Shot Program. The development of PDX models and the generation of data from The Wistar Institute was supported by the NCI, NIH (NCI R50-CA211199). Patient-Derived Models Repository data have been funded in whole or in part with federal funds from the NCI, NIH (contract number HHSN261200800001E). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government. The BRCA PDX models from Washington University used for this study were developed in part through support from the Breast Cancer Research Foundation and the Fashion Footwear Charitable Foundation of New York. The pancreatic cancer PDX models from Washington University used in this study were developed with the support of NCI grants P50 CA196510 and P30 CA091842 and The Foundation for Barnes-Jewish Hospital's Cancer Frontier Fund through the Siteman Cancer Center Investment Program. The data for these models were provided by U54-CA224083. Support for the EurOPDX consortium included funding provided by Fondazione AIRC under the 5 per Mille 2018 (ID. 21091) program (E. Medico, A.B. and L.T.), AIRC Investigator Grants 18532 (L.T.) and 20697 (A.B.), AIRC/CRUK/FC AECC Accelerator Award 22795 (L.T.), EU Horizon 2020 Research and Innovation Programme grant agreement number 731105 'EDIReX' (E. Medico, A.B., L.T., A.T.B., V.S. and J.J.), Fondazione Piemontese per la Ricerca sul Cancro-ONLUS 5 per Mille Ministero della Salute 2015 (E. Medico and L.T.), 2014 and 2016 (L.T.), and 2017 (E. Medico), My First AIRC Grant 19047 (C.I.), EU Horizon 2020 Research and Innovation Programme grant agreement number 754923 'COLOSSUS' (A.T.B., D.L. and L.T.), European Research Council Consolidator Grant 724748 'BEAT' (A.B.), Science Foundation Ireland grant 13/CDA/2183 'COLOFORETELL' (A.T.B.), Irish Health Research Board grant ILP-POR-2019-066 (A.T.B.), ISCIII Miguel Servet program CP14/00228 and the GHD-Pink/FERO Foundation grant (V.S.), Netherlands Organisation for Scientific Research (NWO) Vici grant 91814643 (J.J.), European Research Council Synergy project CombatCancer (J.J.), the Oncode Institute (J.J. and R.d.B.), the Dutch Cancer Society (J.J. and R.d.B.) and NCI grant U24 CA204781 (J.H.C. and T.F.M.). The EurOPDX consortium members thank C. Saura from the Breast Cancer and Melanoma Group (VHIO) and J. Balmaña from the Hereditary Cancer Genetics Group (VHIO) for providing study samples. We thank D. Krupke from JAX for assistance with organizing the tumor type information.

Author contributions

X.Y.W., C.J.B., J.J., A.T.B., L.T., J.A.M., C.I., E. Medico and J.H.C. conceived of and jointly supervised the study. X.Y.W. organized the study, collected and structured the data and designed and carried out the analyses. J.G. collected and organized the EurOPDX data and carried out the analyses. X.Y.W., E. Medico and J.H.C. wrote the manuscript. J.G., C.I., Z.-M.Z., A.S. and M.W.L. contributed to the refinement of the manuscript. A.S. and M.W.L. developed the workflows. A.S., Z.-M.Z., M.W.L. and Y.-S.S. assisted with the computational analyses. R.J., C.F., J. Randjelovic, D.A.D., J. Rosains and B.D.-D. assisted with the workflow development and data collection and organization on the Cancer Genomics Cloud. R.d.B. and R.E.B. contributed to sample selection and the processing of EurOPDX data. C.J.B., R.P., L.C., Y.A.E., J.H.D., S.S., M.H.B., C.-H.Y., E.C.-S., A.L.W, B.E.W., M.T.L., Y.X., J. Wang, B.F., J.A.R., F.M.-B., J. Wickramasinghe, A.V.K., V.W.R.,

M.H., M.A.D., H.S., R.J.M., S.R.D., L.D., S.L., R.G., F.G., A.B., L.T., A.L., A.C.O., A.T.B., E. Modave, D.L., P.t.B., J.J., V.S., E. Marangoni, H.K., J.-I.K., H.-K.Y., C.L., E. Medico and J.H.C. contributed the sequencing and array data. C.J.B., E. Medico and J.H.C. directed the project. The named author list describes the primary contributors of data and analysis to the project, but these studies were supported by consortium-wide activities. All members of the PDXNet and EurOPDX consortia participated in group discussions or supportive analyses regarding the study design, data standards, sample collection or data analysis approaches.

Competing interests

A.L.W. and B.E.W. receive a portion of royalties if the University of Utah licenses certain PDX models to for-profit entities. M.T.L. is a founder of, and equity stake holder in, Tvardi Therapeutics, a founder of, and limited partner in, StemMed and a manager in StemMed Holdings. He also receives a portion of royalties if the Baylor College of Medicine licenses certain PDX models to for-profit entities. J.A.R. serves as a consultant and received stocks from Genprex, and receives royalties from patents issued. F.M.-B. reports receiving commercial research grants from Novartis, AstraZeneca, Calithera, Aileron, Bayer, Jounce, CytomX, eFFECTOR, Zymeworks, PUMA Biotechnology, Curis, Millennium, Daiichi Sankyo, Abbvie, Guardant Health, Takeda, Seattle Genetics and GlaxoSmithKline, as well as grants and travel-related fees from Taiho, Genentech, Debiopharm Group and Pfizer. She also served as a consultant to Pieris, Dialectica, Sumitomo Dainippon, Samsung Bioepis, Aduro, OrigiMed, Xencor, The Jackson Laboratory, Zymeworks, Kolon Life Science and Parexel International, and an advisor to Inflection Biosciences, GRAIL, DarwinHealth, Spectrum, Mersana and Seattle Genetics. L.T. reports receiving research grants from Symphogen, Servier, Pfizer and Merus, and he is in the speakers' bureau of Eli Lilly, AstraZeneca and Merck. J.J. reports receiving funding for collaborative research from Artios Pharma. He also serves as a Scientific Advisory Board member of Artios Pharma. The other authors declare no competing interests.

Additional information

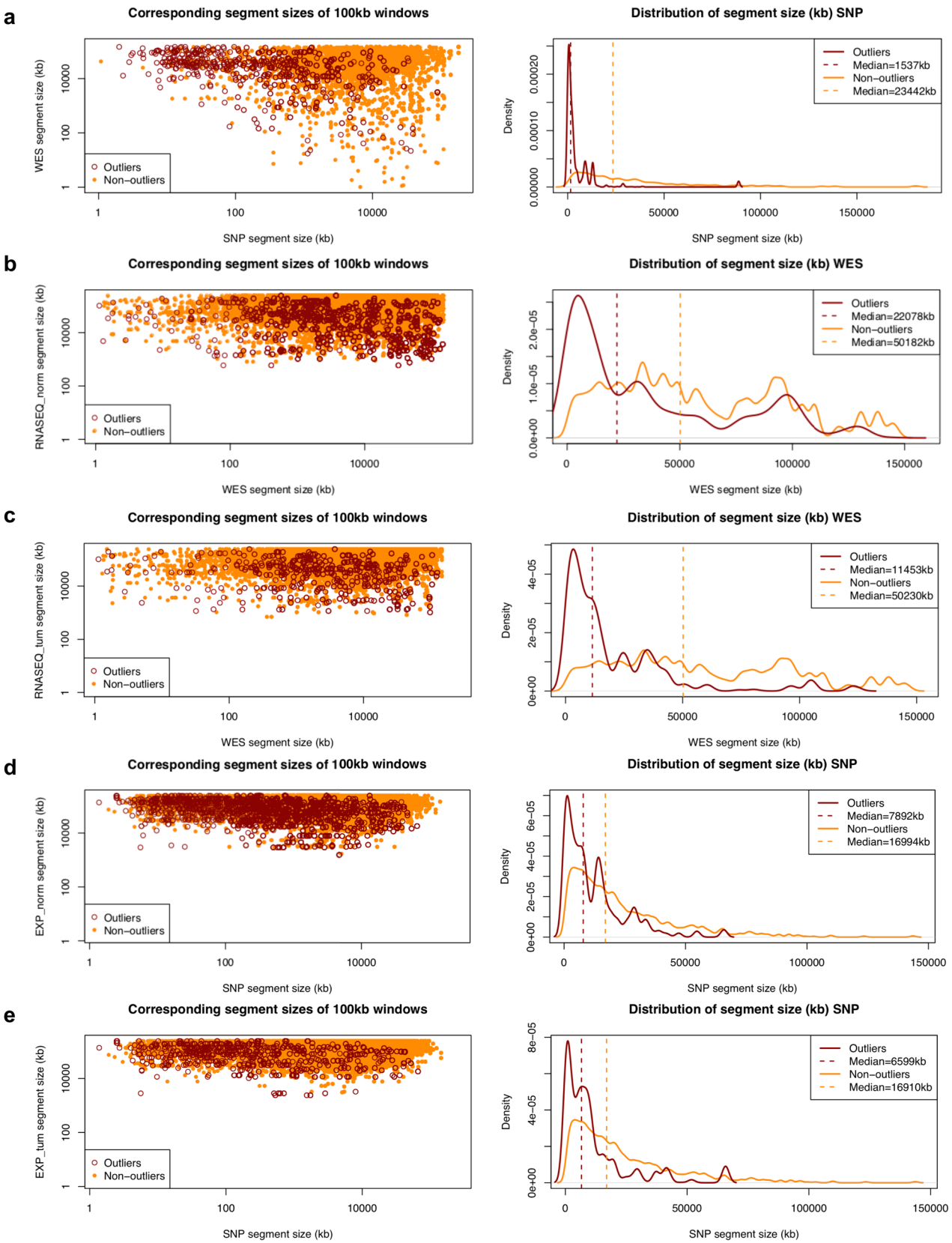
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00750-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00750-6.

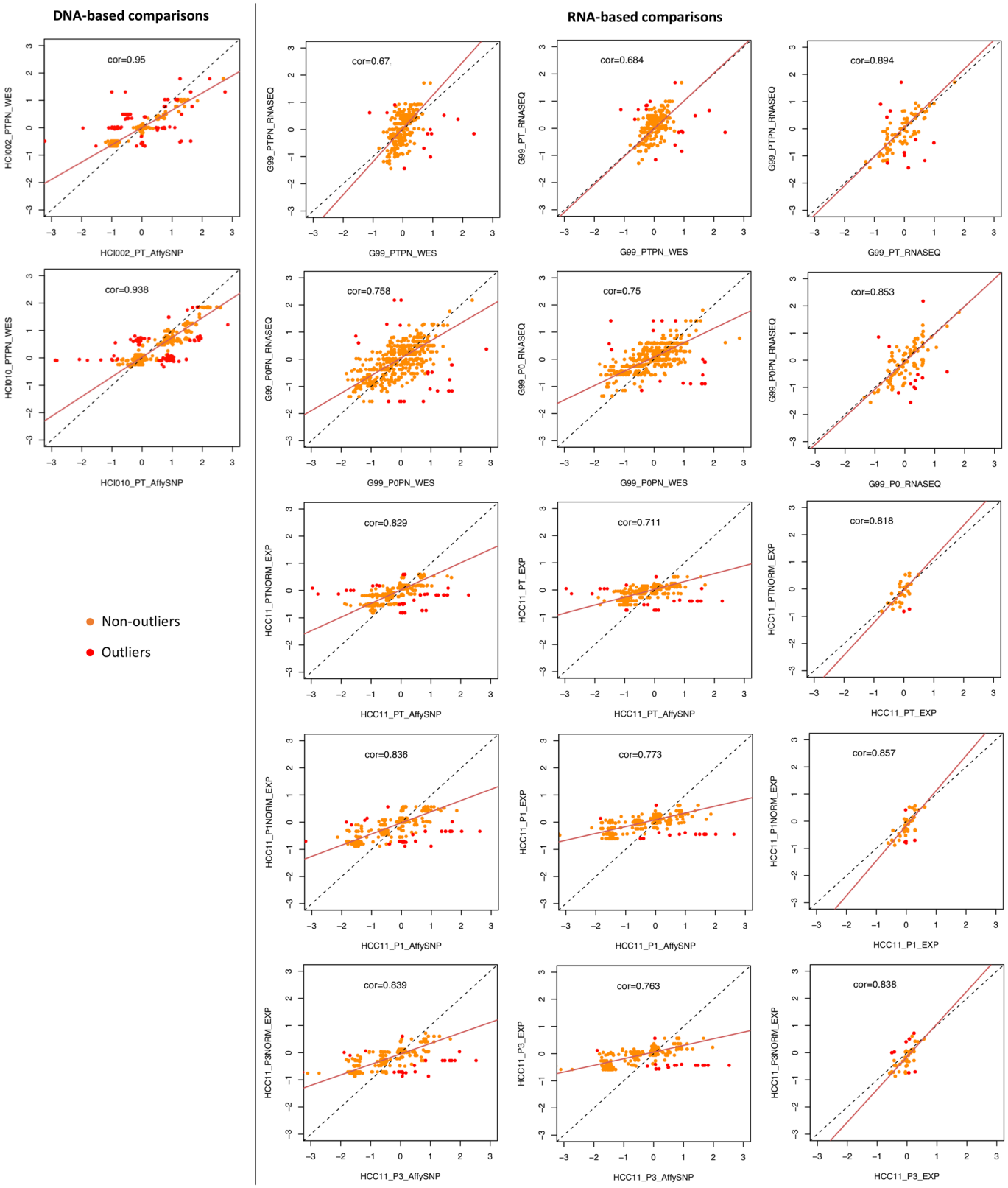
Correspondence and requests for materials should be addressed to E. Medico or J.H.C.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

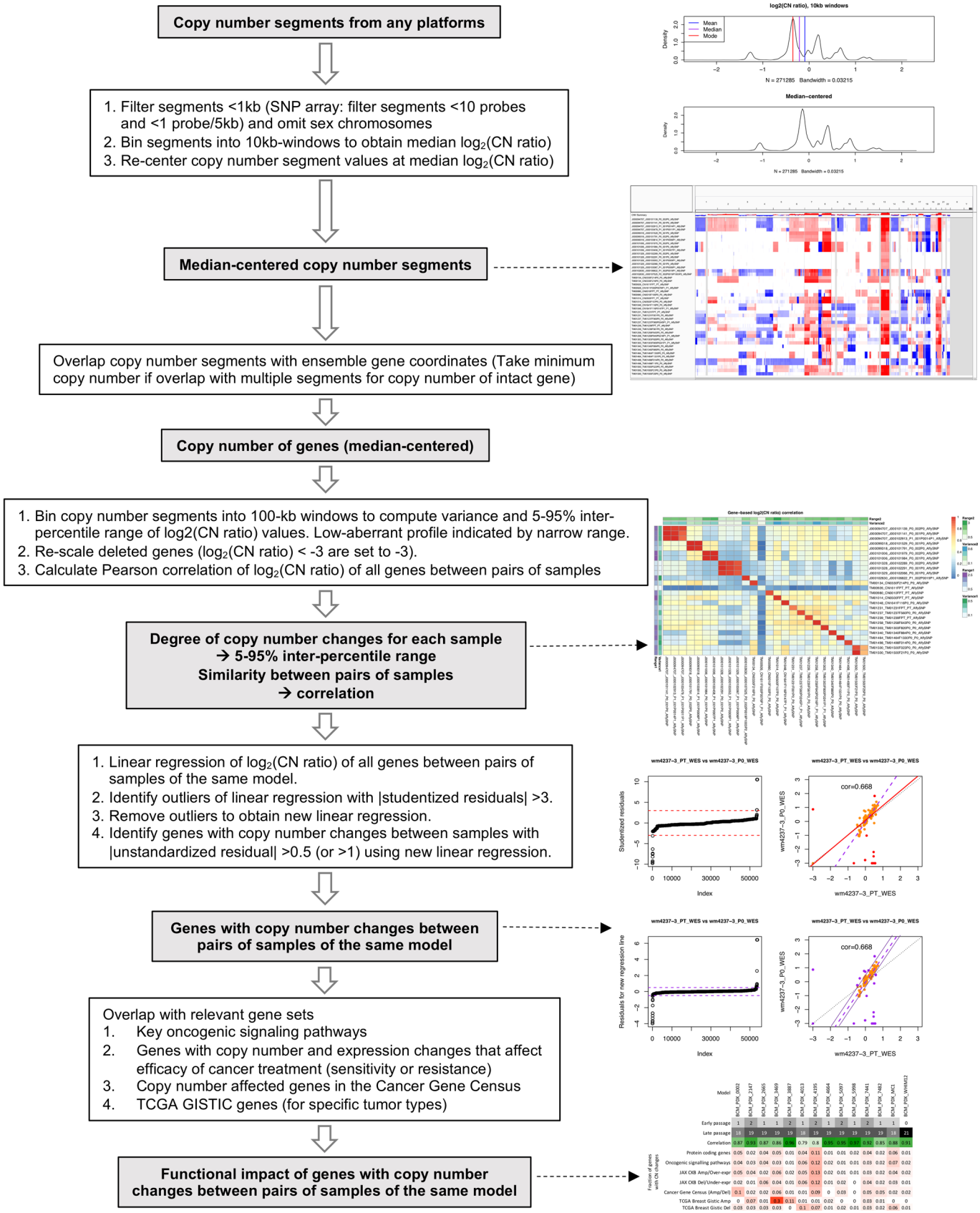
Reprints and permissions information is available at www.nature.com/reprints.



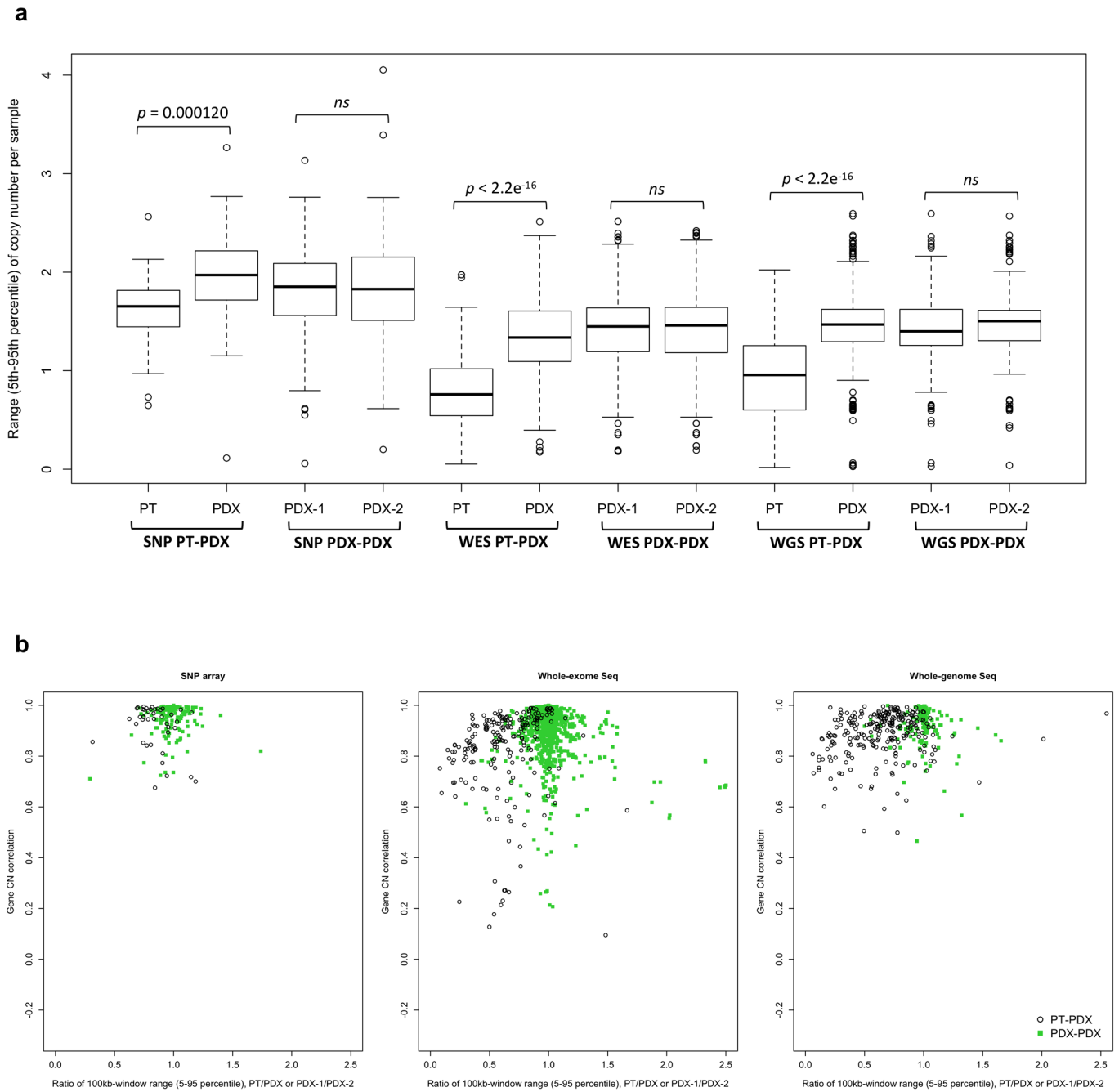
Extended Data Fig. 1 | Comparison of segment sizes between different platforms. The left panel compares the combined corresponding segment sizes of outlier and non-outliers from the linear regression of the $\log_2$(CN ratio) of 100-kb windows binned from copy number segments between matched samples estimated from two different platforms or methods combined. Outliers of the linear regression are identified by studentized residuals > 3 and < -3. **a**, SNP vs. WES. **b**, WES vs. RNASEQ (NORM). **c**, WES vs. RNASEQ (TUM). **d**, SNP vs. EXPARR (NORM). **e**, SNP vs. EXPARR (TUM) (see Supplementary Table 3). The right panel compares the distribution of the segment sizes of outliers and non-outliers for the platform or method of higher resolution.



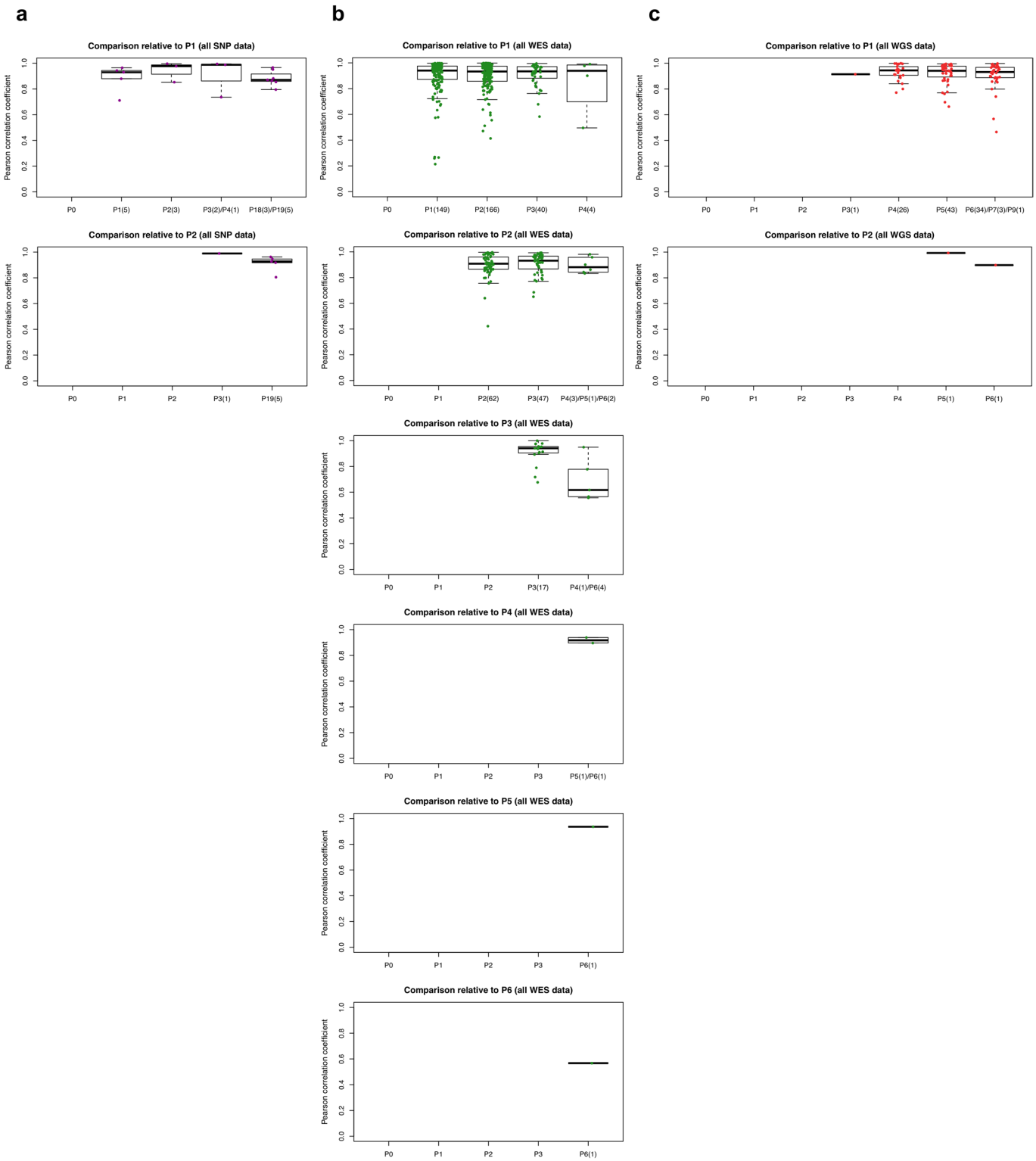
Extended Data Fig. 2 | Comparison of copy number between different platforms. Pearson correlation and linear regression of the $\log_2$(CN ratio) of 100-kb windows binned from copy number segments of CNA profiles between matched patient tumor samples estimated from different platforms or analysis methods for examples shown in Fig. 2d. Outliers of the linear regression are identified by studentized residuals > 3 and < -3. RNA-seq and expression array samples denoted with 'PN' or 'NORM' are normalized by the median expression of normal samples.



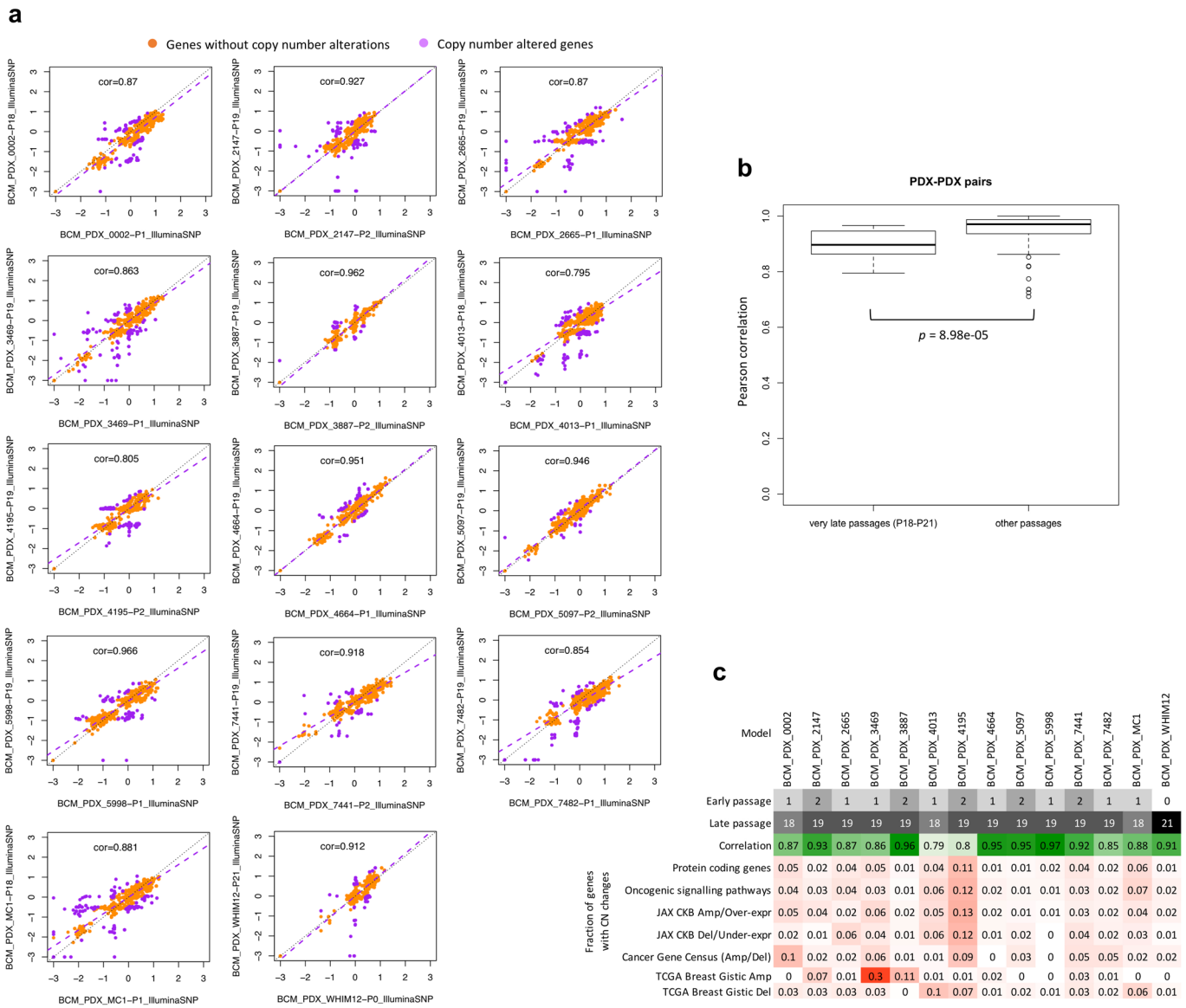
Extended Data Fig. 3 | Analysis workflow to compare CNA between two samples of the same PDX model. A correlation and robust regression approach to quantify similarity of CNA profiles and identify genes with copy number changes between two samples.



Extended Data Fig. 4 | Correlations between PT-PDX and PDX-PDX pairs. a, The 5-95% inter-percentile range of CNA profiles between PT-PDX or PDX-PDX sample pairs from the same model on different platforms as shown in Fig. 3a–c. The 5-95% inter-percentile range of $\log_2$(CN ratio) values were calculated across all 100-kb windows per sample. *P*-values were computed by one-sided Wilcoxon rank sum test (ns: non-significant, $P > 0.05$). In the boxplots, the center line is the median, box limits are the upper and lower quantiles, whiskers extend 1.5× the interquartile range, and dots represent the outliers. **b**, Pearson correlation of the samples versus the ratio of 5-95% inter-percentile range between two samples (PT/PDX or PDX-1/PDX-2). Samples pairs with ratio of range much greater or less than 1 (that is one sample is much less aberrant than the other) tend to have lower correlations. PDX-1, lower passage PDX; PDX-2, later passage PDX or same passage PDX of different lineage.

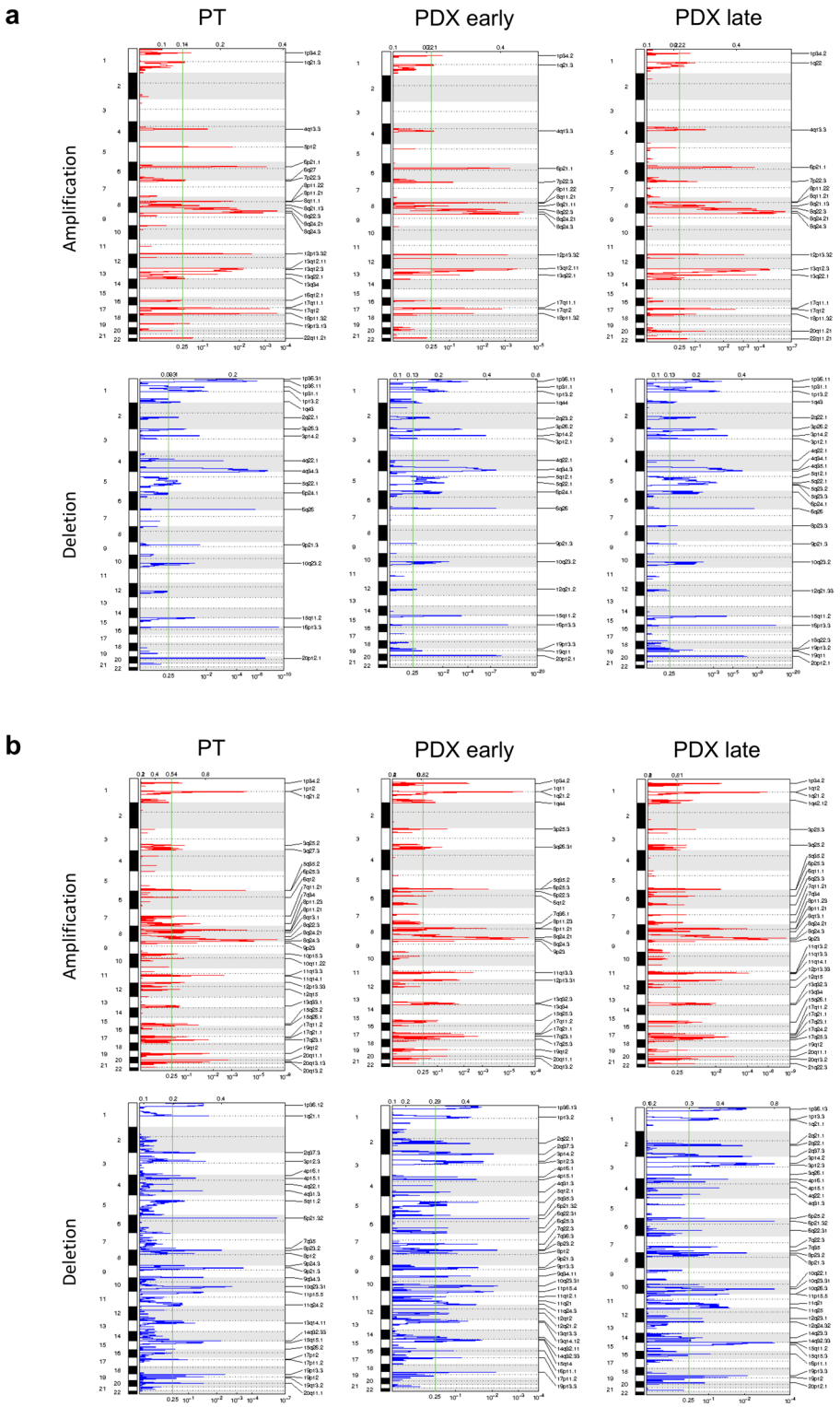


Extended Data Fig. 5 | Distribution of Pearson correlation coefficients of gene-based copy number. a-c, Estimated by SNP array (**a**), WES (**b**), and WGS (**c**) between different combinations of patient tumor and PDX passages of the same model. Comparisons relative to passages P1 or later passages (refer to Fig. 3d–f for comparisons with PT and P0). In the boxplots, the center line is the median, box limits are the upper and lower quantiles, whiskers extend 1.5× the interquartile range, and dots represent all data points.

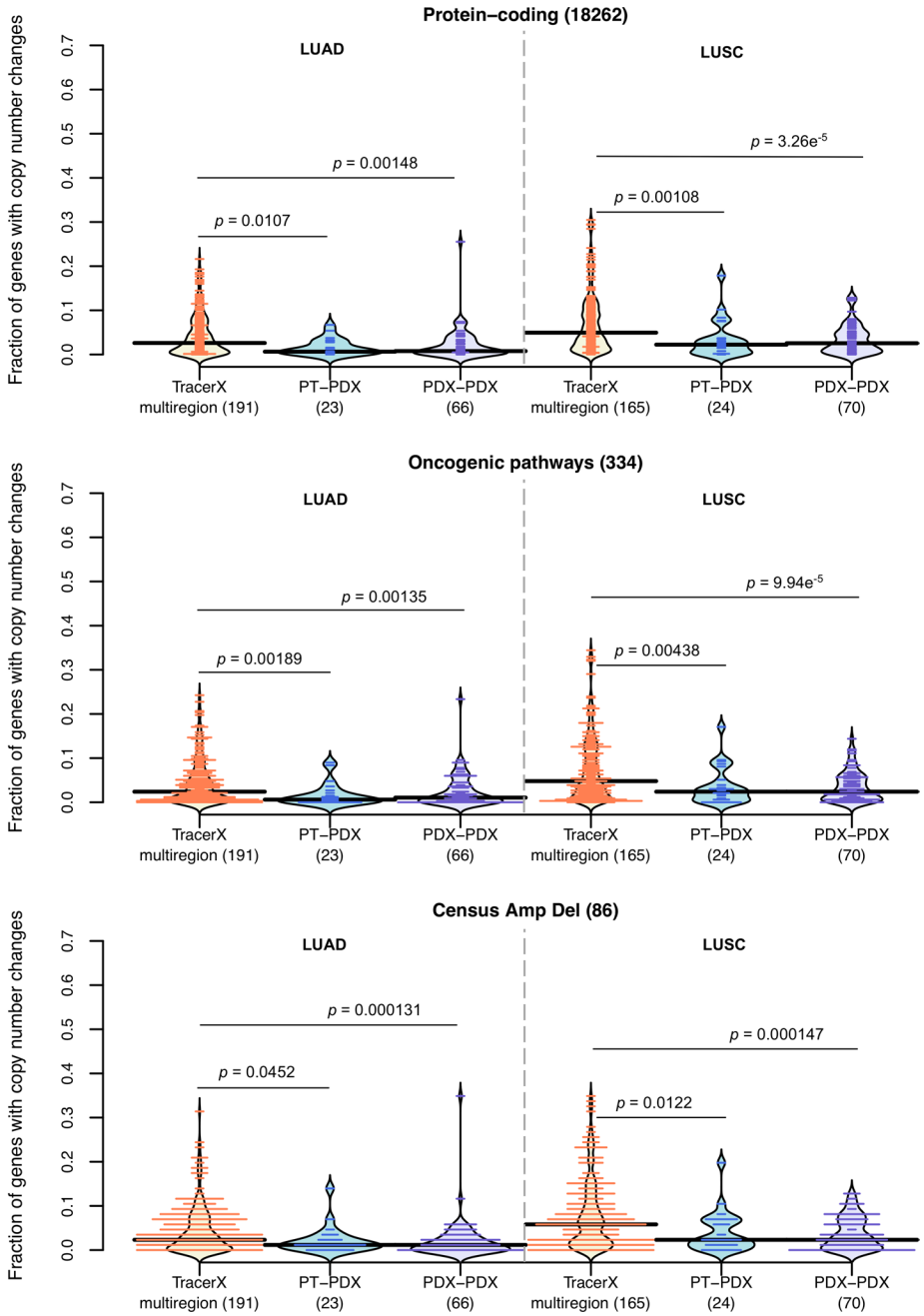


Extended Data Fig. 6 | Comparison of CNA between early and very-late passages. In the BCM SNP array breast cancer dataset. **a**, Correlation and robust regression of gene-based copy number between early (P0-P2) and very-late passages (P18-P21) of the same model. Genes with copy number changes between the passages are identified by |residual| > 0.5. Some genes show signs of complete deletion (log_2(CN ratio) < -2) but then reappear in later passages. This can only be explained by the early and late passages being dominated by different pre-existing subclones. **b**, Distribution of Pearson correlation coefficients of gene-based copy number between early and very-late passages of the same model (14 models/pairwise correlations) compared to correlation coefficients between lower passages denoted as 'other passages' (< P4). Correlation for 'other passages' are based on models from all other non-BCM SNP array datasets (111 pairwise correlations). *P*-values were computed by one-sided Wilcoxon rank sum test. In all boxplots, the center line is the median, box limits are the upper and lower quantiles, whiskers extend 1.5× the interquartile range, and dots represent outliers. **c**, Summary of passage numbers, copy number correlation, and fraction of genes of different gene sets with copy number changes (|residual| > 0.5) between passages of each breast cancer model.

Model	BCM_PDX_0002	BCM_PDX_2147	BCM_PDX_2665	BCM_PDX_3469	BCM_PDX_3887	BCM_PDX_4013	BCM_PDX_4195	BCM_PDX_4664	BCM_PDX_5097	BCM_PDX_5998	BCM_PDX_7441	BCM_PDX_7482	BCM_PDX_MC1	BCM_PDX_WHIM12
Early passage	1	2	1	1	2	1	2	1	2	1	2	1	1	0
Late passage	18	19	19	19	19	18	19	19	19	19	19	19	18	21
Correlation	0.87	0.93	0.87	0.86	0.96	0.79	0.8	0.95	0.95	0.97	0.92	0.85	0.88	0.91
Protein coding genes	0.05	0.02	0.04	0.05	0.01	0.04	0.11	0.01	0.01	0.02	0.04	0.02	0.06	0.01
Oncogenic signalling pathways	0.04	0.03	0.04	0.03	0.01	0.06	0.12	0.02	0.01	0.01	0.03	0.02	0.07	0.02
JAX CKB Amp/Over-expr	0.05	0.04	0.02	0.06	0.02	0.05	0.13	0.02	0.01	0.01	0.03	0.02	0.04	0.02
JAX CKB Del/Under-expr	0.02	0.01	0.06	0.04	0.01	0.06	0.12	0.01	0.02	0	0.04	0.02	0.03	0.01
Cancer Gene Census (Amp/Del)	0.1	0.02	0.02	0.06	0.01	0.01	0.09	0	0.03	0	0.05	0.05	0.02	0.02
TCGA Breast Gistic Amp	0	0.07	0.01	0.3	0.11	0.01	0.01	0.02	0	0	0.03	0.01	0	0
TCGA Breast Gistic Del	0.03	0.03	0.03	0.03	0	0.1	0.07	0.01	0.02	0.01	0.03	0.02	0.06	0.01



Extended Data Fig. 7 | GISTIC analysis of recurrent CNAs. a,**b**, GISTIC plots showing amplified and deleted regions in the EurOPDX WGS of trios of PTs and derived PDXs, at early and late passages, of colorectal cancer (**a**, 87 trios) and breast cancer (**b**, 43 trios). For each GISTIC plot, the top axis reports the G-score and the bottom axis the *q*-value.



Extended Data Fig. 8 | Distribution of proportion of altered genes for lung cancer samples. Comparison between multi-region tumor pairs from TRACERx, and PT-PDX and PDX-PDX pairs for various gene sets for LUAD and LUSC. Gene sets and CNA thresholds are the same as Fig. 4, other gene sets are shown in Fig. 6b. *P*-values were computed by one-sided Wilcoxon rank sum test. Numbers of genes per gene set are indicated in the plot title, and number of pairwise comparisons are indicated in the horizontal axis labels.

nature research

Corresponding author(s): Enzo Medico, Jeffrey Chuang

Last updated by author(s): Oct 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specialized software was used for data collection.
Data analysis	We have used well-established computational sequence analysis and statistical analysis techniques, so no code is provided. Full descriptions of all analysis techniques are provided in the Methods. The implementation of the copy number estimation workflow from whole-exome sequencing data is deployed in the cancer genomics cloud at SevenBridges (https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-workflow/, https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-xenome-tumor-normal-workflow/). Publicly available algorithm/software used in the analyses include PennCNV-Affy, Affymetrix Power Tools v1.15.0, Illumina GenomeStudio, ASCAT (v2.0.7, v2.4.3, v2.5.1), cut-adapt v1.15, BWA (v0.7.12, v0.7.15), Xenome v1.0.0, Picard (v1.43 , v2.8.1), GATK 4.0.5.1, SnpEff v4.3, SAMTools v0.1.18, XenofilteR, Sequenza v2.1.2, QDNAseq v1.20, RSEM v1.3.1, CGH-Explorer, IGV v2.4.13, GenVisR v1.16.1, Bedtools v2.26.0, GISTIC 2 v6.15.28 and GSEA v3.0.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Copy number calls from all datasets are available in Supplementary Data 1, and these are used for all figures. Raw sequence data for these calls are a combination of previously described sources (notably the publicly available NCI Patient Derived Models Repository, pdmr.cancer.gov) and newly sequenced data. New sequence

data from the PDXNet are being shared as part of the NCI Cancer Moonshot initiative through the Cancer Data Service. For further details, contact the authors. The SNP array data generated by The Jackson Laboratory can be requested via the Mouse Models of Human Cancer Database (tumor.informatics.jax.org). The whole genome sequencing data generated by EurOPDX can be made available by directly contacting the EurOPDX consortium (dataportal.europdx.eu). Other publicly available data used in the analyses include GSE90653, GSE3526, GSE33006 and E-MTAB-1503-3, MSigDB v6.2 and TRACERx NSCLC data (DOI: 10.1056/NEJMoa1616288).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	This is an extensive meta-analysis of evolutionary behaviors using >1500 samples to consider a variety of hypotheses. Descriptions of sample size, data exclusions, and replicability are provided throughout the manuscript.
Data exclusions	See above
Replication	See above
Randomization	This is not a case/control study and randomization is not suited to the project design.
Blinding	This is not a case/control study and blinding is not suited to the project design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Details on mouse strains used for xenografting studies are provided in the Methods and the references therein.
Wild animals	No wild animals
Field-collected samples	No field collected samples.
Ethics oversight	This study is a meta-analysis of data collected across a large number of consortium sites and does not follow a single study protocol. All studies were conducted in compliance with ethics regulations, as detailed in the Methods and the references therein.

Note that full information on the approval of the study protocol must also be provided in the manuscript.