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

(Article begins on next page)

18 October 2023

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

Xing Yi Woo[1,#,§], Jessica Giordano[2,3,#], 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,§], Jeffrey A. Moscow[27,§], Livio Trusolino[2,3,§], Annette T. Byrne[14,§], Jos Jonkers[5,§], Carol J. Bult[4,§], Enzo Medico[2,3,§,*], Jeffrey H. Chuang[1,§,*], PDXNET consortium[28] & EurOPDX consortium[28]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[2] Department of Oncology, University of Torino, Candiolo, Italy
[3] Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy
[4] The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA
[5] Netherlands Cancer Institute, Amsterdam, Netherlands
[6] College of Medicine, Seoul National University, South Korea
[7] Frederick National Laboratory for Cancer Research, Frederick, MD, USA
[8] Department of Oncological Sciences, University of Utah Huntsman Cancer Institute, Salt Lake City, UT, USA
[9] Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
[10] Department of Bioinformatics and Computer Biology, The University of Texas M.D. 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

35   [14] Department of Physiology and Medical Physics, Centre for Systems Medicine, Royal College

36   of Surgeons in Ireland, Dublin, Ireland

37   [15] Center for Cancer Biology, VIB, Leuven, Belgium

38   [16] Laboratory of Translational Genetics, Department of Human Genetics, KU Leuven, Leuven,

39   Belgium

40   [17] Vall d´Hebron Institute of Oncology, Barcelona, Spain

41   [18] Department of Translational Research, Institut Curie, PSL Research University, Paris, France

42   [19] Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an,

43   Shaanxi, People's Republic of China

44   [20] Department of Life Sciences, Ewha Womans University, Seoul, Seoul, South Korea

45   [21] Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA

46   [22] Department of Surgery, University of Utah Huntsman Cancer Institute, Salt Lake City, UT,

47   USA

48   [23] Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA

49   [24] Department of Thoracic & Cardiovascular Surgery, The University of Texas M.D. Anderson

50   Cancer Center, Houston, TX, USA

51   [25] Department of Investigational Cancer Therapeutics, The University of Texas M.D. Anderson

52   Cancer Center, Houston, TX, USA

53   [26] Department of Melanoma Medical Oncology, The University of Texas M.D. Anderson Cancer

54   Center, Houston, TX, USA

55   [27] Investigational Drug Branch, National Cancer Institute, Bethesda, MD, USA

56   [28] A full list of members and their affiliations appears at the end of the paper.

57   [#] These authors contributed equally to this work.

58   [§] These authors jointly supervised this work.

59   * Correspondence should be addressed to E.M. (enzo.medico@unito.it) or J.H.C.

60   (jeff.chuang@jax.org).

61

62

63   **ABSTRACT**

64   Patient-derived xenografts (PDXs) are resected human tumors engrafted into mice for preclinical

65   studies and therapeutic testing. It has been proposed that the mouse host affects tumor evolution

66   during PDX engraftment and propagation, impacting the accuracy of PDX modeling of human

67   cancer. Here we exhaustively analyze copy number alterations (CNAs) in 1451 PDX and matched

68   patient tumor (PT) samples from 509 PDX models. CNA inferences based on DNA sequencing

69   and microarray data displayed substantially higher resolution and dynamic range than gene
70   expression-based inferences, and they also showed strong CNA conservation from PTs through
71   late-passage PDXs. CNA recurrence analysis of 130 colorectal and breast PT/PDX-early/PDX-
72   late trios confirmed high-resolution CNA retention. We observed no significant enrichment of
73   cancer-related genes in PDX-specific CNAs across models. Moreover, CNA differences between
74   patient and PDX tumors were comparable to variations in multi-region samples within patients.
75   Our study demonstrates the lack of systematic copy number evolution driven by the PDX mouse
76   host.

77

78   **MAIN**
79   Human tumors engrafted into transplant-compliant recipient mice (patient-derived xenografts,
80   PDX) have advantages over prior model systems of human cancer (e.g genetically engineered
81   mouse models[1,2] and cancer cell lines[3]) for preclinical drug efficacy studies because they allow
82   researchers to directly study human cells and tissues *in vivo*[4-7]. Comparisons of genome
83   characteristics and histopathology of primary tumors and xenografts of various cancer types[8-14]
84   have demonstrated that the biological properties of patient-derived tumors are largely preserved
85   in xenografts. A growing body of literature supports their use in cancer drug discovery and
86   development[15-17].

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

98   Here we resolve these contradicting observations by systematically evaluating CNA
99   changes and the genes they affect during engraftment and passaging in a large, internationally
100   collected set of PDX models, comparing both RNA and DNA-based approaches. The data
101   collected, as part of the U.S. National Cancer Institute (NCI) PDXNet (PDX Development and
102   Trial Centers Research Network) Consortium and EurOPDX consortium, comprises patient tumor

3

103   (PT) and PDX samples from >500 models. Our study demonstrates that prior reports of systematic

104   copy number divergence between PTs and PDXs are incorrect, and that there is high retention of

105   copy number during PDX engraftment and passaging. This work also finely enumerates the copy

106   number profiles in hundreds of publicly available models, which will enable researchers to assess

107   the suitability of each for individualized treatment studies.

108

109   **RESULTS**

110   **Catalog of copy number alterations in PDXs**

111   We have assembled copy number alteration (CNA) profiles of 1451 unique samples (324

112   PT, and 1127 PDX samples) corresponding to 509 PDX models contributed by participating

113   centers of the PDXNET, the EurOPDX consortium, and other published datasets[11,26] (see

114   METHODS, Supplementary Methods, Supplementary Table 1, Supplementary Fig. 1). We

115   estimated copy number (CN) from five data types: single nucleotide polymorphism (SNP) array,

116   whole-exome sequencing (WES), low-pass whole-genome sequencing (WGS), RNA sequencing

117   (RNA-Seq) and gene expression array data, yielding 1548 tumor datasets including samples

118   assayed on multiple platforms (see METHODS, Supplementary Methods, Supplementary Data

119   1). Paired-normal DNA and in some cases, paired normal RNA, were also obtained to calibrate

120   WES and RNA-Seq tumor samples.

121   The combined PDX data represent 16 broad tumor types derived from American,

122   European and Asian cancer patients (see METHODS), with 64% (n=324) of the models having

123   their corresponding patient tumors assayed and another 64% (n=328) having multiple PDX

124   samples of either varying passages (P0 – P21) or varying lineages from propagation into distinct

125   mice (Fig. 1a, Supplementary Table 2). The distributions of PT and PDX samples across different

126   tumor types, passages and assay platforms (Fig. 1b, Supplementary Fig. 2-12) show the wide

127   spectrum of this combined dataset, which, to the best of our knowledge, is the most

128   comprehensive copy number profiling of PDXs compiled to date (Supplementary Note 1).

129   Additionally, our data include seven patients with multiple tumors collected either from different

130   relapse time points or different metastatic sites, resulting in multiple PDX models derived from a

131   single patient.

132

133   **Comparison of CNA profiles from SNP array, WES and gene expression data**

134   To compare the CNA profiles from different platforms in a controlled fashion, we assembled a

135   dataset with matched measurements across multiple platforms (Supplementary Table 3,

136   Supplementary Fig. 13-17). Copy number calling has been reported to be noisy for several data

137　　types[27,28], and we observed that quantitative comparisons between CNA profiles are sensitive to:

138　　(1) the thresholds and baselines used to define gains and losses, (2) the dynamic range of copy

139　　number values from each platform, and (3) the differential impacts of normal cell contamination

140　　for different measurements. To control for such systematic biases, we assessed the similarity

141　　between two CNA profiles using the Pearson correlation of their $\log_2$(CN ratio) values across the

142　　genome in 100kb windows. Regions with discrepant copy number were identified as those with

143　　outlier values from the linear regression model (see METHODS).

144

145　　***CNAs from WES are consistent with CNAs from SNP array data.*** As earlier studies reported

146　　that CNA estimates from WES data have more uncertainties than those from SNP arrays[29,30], we

147　　implemented a WES-based CNA pipeline and validated it against SNP array-based estimates[31,32]

148　　for matched samples. Copy number gain/loss segments (see METHODS) from SNP arrays were

149　　of a higher resolution (Fig. 2a; median/mean segment size: 1.49/4.05 Mb for SNP, 4.70/14.6 Mb

150　　for WES, $p < 2.2e^{-16}$) and wider dynamic range (Fig. 2b; range of $\log_2$(CN ratio): –8.62 – 2.84 for

151　　SNP, –3.04 – 1.85 for WES, $p < 2.2e^{-16}$). The difference in range is apparent in the linear

152　　regressions between platforms (Supplementary Fig. 18). These observations take into account

153　　the broad factors affecting CNA estimates across platforms, such as the positional distribution of

154　　sequencing loci; the sequencing depth of WES; and the superior removal of normal cell

155　　contamination by SNP array CNA analysis workflows using SNP allele frequencies[33].

156　　　　　We observed strong agreement between SNP arrays and WES, with significantly higher

157　　Pearson correlation coefficients on matched samples than samples of different models (range:

158　　0.913 – 0.957 for matched samples, 0.0366 – 0.354 for unmatched samples, $p = 1.02e^{-06}$), with

159　　the exception of two samples that lacked CNA aberrations and were removed (Fig. 2c,

160　　Supplementary Fig. 13, 18, 19). The discordant copy number regions largely correspond to small

161　　focal events (average size 1.53Mb) detectable by SNP arrays but missed by WES

162　　(Supplementary Fig. 18, Extended Data Fig. 1a, see METHODS). Hence, CNA profiling by WES

163　　is reliable in most regions in this small dataset, with 99% of the genome locations across the

164　　samples consistent with the values from SNP arrays (Supplementary Note 2). These PT-based

165　　observations are also applicable to PDXs given that mouse DNA is absent in SNP array signal

166　　and removed from WES reads[34-36].

167

168　　***Low accuracy for gene expression-derived CNA profiles.*** To compare the suitability of gene

169　　expression for quantifying evolutionary changes in CNA, we adapted the e-karyotyping

170　　method[23,37,38] for RNA-Seq and gene expression array data (Supplementary Fig. 15, 17, see

171   METHODS). Copy number segments calibrated by non-tumor expression were of higher
172   resolution (Fig. 2a; median/mean segment size: 36.0/51.9 Mb for RNASEQ NORM, 48.2/65.3 Mb
173   for RNASEQ TUM, $p < 2.2e^{-16}$; 62.0/72.4 Mb for EXPARR NORM, 80.1/85.2 Mb for EXPARR
174   TUM, $p = 2.20e^{-07}$) and wider dynamic range (Fig. 2b; range of $\log_2$(CN ratio): –2.07 – 2.17 for
175   RNASEQ NORM, –1.79 – 1.81 for RNASEQ TUM, $p < 2.2e^{-16}$; –1.40 – 1.89 for EXPARR NORM,
176   –1.13 – 1.59 for EXPARR TUM, $p = 4.09e^{-07}$) compared to segments calculated by calibration
177   with tumor samples. These alternative expression calibrations yielded biased gain and loss
178   frequencies (Supplementary Note 3, Supplementary Fig. 20) and strong variability (Pearson
179   correlation range: 0.218 – 0.943 for RNASEQ NORM vs TUM, 0.377 – 0.869 for EXPARR NORM
180   vs TUM) in the CNA calls (Fig. 2c, Supplementary Fig. 21). This range of correlations was far
181   greater than was observed in comparisons between the DNA-based methods ($p = 9.37e^{-5}$ and p
182    $= 3.28e^{-07}$ relative to SNP vs WES). This indicates the problematic nature of RNA-based CNA
183   calling with calibration by tumor samples, which has been used when normal samples are not
184   available.
185   Furthermore, expression-based calling had segmental resolution an order of magnitude
186   worse than the DNA-based methods (Fig. 2a, Supplementary Fig. 14-17; median/mean segment
187   size: 3.45/14.0 Mb for WES, 36.0/51.9 Mb for RNASEQ NORM, $p < 2.2e^{-16}$; 1.73/ 5.18 Mb for
188   SNP, 62.0/72.4 Mb for EXPARR NORM, $p < 2.2e^{-16}$). The range of detectable copy number values
189   was also superior for DNA-based methods (Fig. 2b; range of $\log_2$(CN ratio): –6.00 – 5.33 for WES,
190   –2.07 – 2.17 for RNASEQ NORM, $p < 2.2e^{-16}$; –9.19 – 4.65 for SNP, –1.40 – 1.89 for EXPARR
191   NORM, $p < 2.2e^{-16}$). In addition, there was a lack of correlation between the expression-based
192   and DNA-based methods (range: 0.0541 – 0.942 for WES vs RNASEQ (NORM); 0.00517 – 0.921
193   for SNP vs EXPARR (NORM)) (Fig. 2c, Supplementary Fig. 22, 23). CNA estimates after tumor-
194   based expression normalization resulted in further discordance with DNA-based copy number
195   results (range: –0.182 – 0.929, $p = 0.0468$ for WES vs RNASEQ (TUM); –0.0274 – 0.847, $p =$
196    $2.20e^{-06}$ for SNP vs EXPARR (TUM)). Many focal copy number events detected by DNA-based
197   methods, as well as some larger segments, were missed by the expression-based methods
198   (Extended Data Fig. 1b-e). Representative examples illustrating the superior resolution and
199   accuracy from DNA-based estimates are given in Fig. 2d (correlations shown in Extended Data
200   Fig. 2).
201
202   **Concordance of PDXs with patient tumors and during passaging**
203   We next adopted a pan-cancer approach to elucidate potential tumor type-independent copy
204   number evolution in PDXs driven by the mouse host. We tracked the similarity of CNA profiles

205 during tumor engraftment and passaging by calculating the Pearson correlation of gene-level

206 copy-number for samples measured on the same platform (see METHODS, Extended Data Fig.

207 3, Supplementary Fig. 24-60, 62). All pairs of samples derived from the same PDX model were

208 compared – yielding 501 PT-PDX and 1257 PDX-PDX pairs (Supplementary Note 4).

209   For all DNA-based platforms we observed strong concordance between matched PT-PDX

210 and PDX-PDX pairs, significantly higher than between different models from the same tumor type

211 and the same center ($p < 2.2e^{-16}$) (Fig. 3a-c, correlation heatmaps in Supplementary Fig. 24-60).

212 We observed no significant difference in the correlation values between PT-PDX and PDX-PDX

213 pairs for SNP array data (median correlation PT-PDX = 0.950, PDX-PDX = 0.964; $p > 0.05$),

214 though there were small but statistically significant shifts for WES (PT-PDX = 0.874, PDX-PDX =

215 0.936; $p = 2.31e^{-16}$) and WGS data (PT-PDX = 0.914, PDX-PDX = 0.931; $p = 0.000299$). PT

216 samples have a smaller CNA range than their derived PDXs (median ratio PT/PDX / PDX/PDX:

217 0.832/0.982, $p = 0.000120$ for SNP; 0.626/0.996, $p < 2.2e^{-16}$ for WES; 0.667/1.00, $p < 2.2e^{-16}$ for

218 WGS; Supplementary Fig. 62b, Extended Data Fig. 4), which can be attributed to stromal DNA in

219 PT samples "diluting" the CNA signal. In PDXs, the human stromal DNA is reduced[11,13]. The

220 minimal effect for SNP array data confirms this interpretation as human stromal DNA contributions

221 can be removed from SNP arrays based on allele frequencies of germline heterozygous sites,

222 while such contributions to WES and WGS have higher uncertainties. We also performed intra-

223 model comparisons using RNA-based approaches, which showed that the expression-based

224 comparison of CNA profiles between PT and PDXs can lead to the overestimation of copy number

225 changes during engraftment and passage (Supplementary Fig. 63, Supplementary Note 5).

226

227 ***Late PDX passages maintain CNA profiles similar to early passages.*** Systematic mouse

228 environment-driven evolution, if present, should reduce CN correlations at each subsequent

229 passage. However, we observed no apparent effect during passaging on the SNP, WES, or WGS

230 platforms (Fig. 3d-f, Extended Data Fig. 5). For example, the SNP data showed no significant

231 difference between passages (Fig. 3d, Extended Data Fig. 5a). For those models having very late

232 passages, there was a small but statistically significant correlation decrease compared to models

233 with earlier passages ($p < 8.98e^{-05}$, Extended Data Fig. 6b), indicating some copy number

234 changes can occur over long-term passaging (Supplementary Fig. 35). However even at these

235 late passages, the correlations to early passages remained high (median = 0.896). In any given

236 comparison, only a small proportion of the genes were affected by copy number changes (median:

237 2.72%, range: 1.03% – 11.9%). Genes that are deleted and subsequently gained in the later

238 passages (top left quadrant of regression plots, Extended Data Fig. 6a) suggest selection of

239   preexisting minor clones as the key mechanism in these regions. For WES and WGS data, more
240   variability in the correlations can be observed (Fig. 3e, f, Extended Data Fig. 5b, c), likely due to
241   a few samples having more stromal contamination or low aberration levels (Supplementary Fig.
242   62b, Extended Data Fig. 4). However, the lack of downward trend over passaging was also
243   apparent in these sets (Supplementary Note 6).
244
245   ***PDX copy number profiles trace lineages.*** We next compared the similarity of engrafted PDXs
246   of the same model with the same passage number. Surprisingly, we discovered that these pairs
247   were not more similar than pairs of PDXs from different passage numbers (Fig. 3d-e, Extended
248   Data Fig. 5, Supplementary Note 7). Such similarity in correlations suggested that copy number
249   divergence might be associated with effects other than passaging. To further this analysis, we
250   defined, for JAX SNP array and PDMR WES datasets, samples within a lineage as those differing
251   only by consecutive serial passages, while we defined lineages as split when a tumor was divided
252   and propagated into multiple mice (Fig. 3g). For the EurOPDX CRC and BRCA WGS datasets,
253   such lineage splitting was due only to cases with initial engraftment of different fragments of the
254   PT, i.e., PDX samples of different passages were considered as different lineages if they originate
255   from different PT fragments. We observed lower correlation between PDX samples from different
256   lineages compared to within a lineage (Fig. 3h, p = 0.0233 for SNP, p = 0.00119 for WES, p =
257   0.000232 for WGS), despite a majority of these pairwise comparisons exhibiting high correlation
258   (>0.9) (Supplementary Note 8, 9). This suggests that lineage-splitting is often responsible for
259   deviations in CNAs between samples, and that copy number evolution during passaging mainly
260   arises from evolved spatial heterogeneity[24].
261   We further explored if the stability of copy number during engraftment and passaging is
262   affected by mutations in genes known to impact genome stability (see METHODS). Overall, we
263   observed that presence of mutations in such genes does not lead to increased copy number
264   changes during PDX engraftment and passaging (Supplementary Note 10, Supplementary Fig.
265   66).
266
267   **Genes with copy number alterations acquired during engraftment and passaging show no**
268   **preference for cancer or treatment-related functions.**
269   Next, we investigated which genes tend to undergo copy number changes. Genes with changes
270   during engraftment or during passaging were identified based on a residual threshold with respect
271   to the improved linear regression[39] (see METHODS, Extended Data Fig. 3). To test for functional
272   biases, we compared CNA-altered genes to gene sets with known cancer- and treatment-related

273 functions[40-43] (see METHODS). We calculated the proportion of altered genes for sample pairs

274 from each model across all platforms and tumor types. In agreement with the high maintenance

275 of CNA profiles described above, we found the proportion of altered protein-coding genes to be

276 low (median/IQR: 1.90%/ 4.11% PT-PDX, 1.25%/ 3.60% PDX-PDX pairs, Fig. 4a). Only 8.78% of

277 PT-PDX pairs and 4.53% PDX-PDX pairs showed >10% of their protein-coding genes altered.

278 We observed no significant increase ($p > 0.1$) in alterations among any of the cancer gene sets

279 compared to the background of all protein-coding genes, for either the PT-PDX or PDX-PDX

280 comparisons. This provides evidence that there is no systematic selection for CNAs in oncogenic

281 or treatment-related pathways during engraftment or passaging. We next considered tumor-type-

282 specific effects, focusing on tumor types with larger numbers of models to ensure statistical

283 power. We observed no significant increase in alterations in tumor-type-specific driver gene sets

284 significantly altered in TCGA[44-47] compared to the background ($p > 0.1$) for either PT-PDX or PDX-

285 PDX comparisons (Fig. 4b, Supplementary Note 11).

286

287 ***Low recurrence of altered genes across models.*** We observed a very low recurrent frequency

288 (Fig. 4c, see METHODS), with only 12 and 2 genes recurring at > 5% frequency for PT-PDX and

289 PDX-PDX comparisons, respectively (Supplementary Table 4). No gene had a recurrence

290 frequency higher than 8.96% (Supplementary Note 12). None of these recurrent genes

291 overlapped cancer- or treatment-related gene sets, nor did they intersect genes (n=3) reported

292 by Ben-David et al.[23] to have mouse-induced copy number changes associated with drug

293 response in the CCLE[48,49] database (Supplementary Note 12).

294

295 ***Absence of CNA shifts in 130 WGS patient tumor, early passage PDX and late passage***

296 ***PDX trios***

297 We next investigated whether recurrent CNA changes occur in PDXs in a tumor-type specific

298 fashion. To this aim, we analysed further the WGS-based CNA profiles of large metastatic

299 colorectal (CRC) and breast cancer (BRCA) series, composed of matched trios of PT, PDX at

300 early passage (PDX-early) and PDX at later passage (PDX-late). Genomic Identification of

301 Significant Targets in Cancer (GISTIC)[50,51] analysis was applied separately to identify recurrent

302 CNAs in each PT, PDX-early and PDX-late cohorts of CRC and BRCA (see METHODS,

303 Supplementary Table 6). As expected, CRCs and BRCAs generated different patterns of

304 significant CNAs, but within each tumor type GISTIC profiles of the PT, PDX-early, and PDX-late

305 cohorts were virtually indistinguishable (Fig 5a, Extended Data Fig. 7, Supplementary Note 13),

306 demonstrating no gross genomic alteration systematically acquired or lost in PDXs.

307   We then carried out gene-level analysis, where each gene was attributed the GISTIC

308   score (G-score) of the respective segment (Supplementary Table 7). In both the CRC and BRCA

309   cohorts, gene-level G-scores of the PTs were highly correlated with the respective PDX-early and

310   PDX-late cohorts (Fig. 5b, c). Moreover, PT versus PDX correlations were comparable to PDX-

311   early versus PDX-late correlations. To search for progressive shifts, we compared the change in

312   G-score (ΔG): (i) from tumor to PDX-early and (ii) from PDX-early to PDX-late. Correlations in

313   these two ΔG values were absent or even slightly negative (bottom-right panels of Fig. 5b, c,

314   Supplementary Note 13). Overall, these results confirmed the absence of systematic CNA shifts

315   in PDXs even under high resolution, gene-level analysis. To evaluate the possibility of systematic

316   copy number evolution at the pathway level in these trios, we performed Gene Set Enrichment

317   Analysis (GSEA)[52,53] using G-scores to rank genes in each cohort (See METHODS,

318   Supplementary Note 14). For both CRC and BRCA, the Normalized Enrichment Score (NES)

319   profiles for the ~8000 gene sets of PTs were highly correlated with the respective PDX-early and

320   PDX-late cohorts (Fig. 5d, e). Moreover, PT versus PDX correlations were comparable to PDX-

321   early versus PDX-late correlations. To search for progressive shifts, we calculated for each

322   significant gene set ΔNES values between PT and PDX-early, as well as between early and late

323   PDX. Similar to what was observed for the ΔG-scores, correlations were absent or at most slightly

324   negative (bottom-right panels of Fig. 5d, e), confirming the absence of systematic CNA-based

325   functional shifts in PDXs.

326

327   **CNA evolution across PDXs is no greater than variation in patient multi-region samples**

328   As a reference for the treatment relevance of PDX-specific evolution, we compared to levels of

329   copy number variation in multi-region samples of patient tumors. For this we used copy number

330   data from multi-region sampling of non-small-cell lung cancer from the TRACERx Consortium[54],

331   performing analogous CNA correlation and gene analyses between multi-region pairs

332   (Supplementary Fig. 69). We observed no significant differences in correlation ($p > 0.05$) between

333   patient multi-region and lung cancer PT-PDX pairs, while PDX-PDX pairs in fact showed

334   significantly better correlation than the multi-region pairs ($p < 0.05$, Fig. 6a), consistent across all

335   lung cancer subtypes. Cancer gene set analyses confirmed these results, with multi-region

336   samples showing greater differences than either PT-PDX or PDX-PDX comparisons, across all

337   the cancer gene sets considered ($p < 0.05$, Fig. 6b, Extended Data Fig. 8). These results show

338   that PDX-associated CNA evolution is no greater than what patients experience naturally within

339   their tumors. Our PDX collection also contains a few cases in which the patient tumor was

340   assayed at multiple time points (relapse/metastasis) or multiple metastatic sites, allowing for

341 controlled comparison of intra-patient variation versus PDX evolution (Supplementary Fig. 3, 4,

342 7). Despite a lower median in correlations among intra-patient samples, the difference compared

343 to CNA evolution during engraftment (PT-PDX) is not statistically significant ($p > 0.05$, Fig. 6c).

344 CNA profiles for these samples are shown visually in Fig. 6d.

345

346 **DISCUSSION**

347 Here we have investigated the evolutionary stability of patient-derived xenografts, an important

348 model system for which there have been prior reports of mouse-induced copy number evolution.

349 To better address this, we assembled the largest collection of CNA profiles of PDX models

350 reported to date, comprising PDX models with multiple passages and their originating patient

351 tumors. Our analysis demonstrated the reliability of copy number estimation by DNA-based

352 measurements over RNA-based inferences, which are substantially inferior in terms of resolution

353 and accuracy (Supplementary Note 15). The importance of DNA measurements is supported by

354 the inconsistent conclusions by two independent studies, Ben-David et al.[23,55] and Mer et al.[56], on

355 the same PDX expression array dataset by Gao et al.[15] Ben-David et al. concluded that drastic

356 copy number changes, driven by mouse-specific selection, often occur within a few passages. On

357 the other hand, Mer et al. reported high similarity between passages of the same PDX model

358 based on direct correlations of gene expression, consistent with our findings in large, independent

359 DNA-based datasets.

360   The CN shifts inferred by Ben-David et al. are inherently impacted by major technical

361 issues. First, the microarray signal for PT samples is diluted by introgressed human stromal cells,

362 while in PDXs mouse stromal transcripts hybridize only to a fraction of the human probes[57].

363 Consequently, PT samples with substantial stromal content would display a reduced signal

364 compared to the corresponding PDX, which can lead to an erroneous inference of systematic

365 increase in aberrations during PDX engraftment when gain/loss regions are directly compared.

366 Second, the mouse host microenvironment can affect the transcriptional profile of the PDX

367 tumor[58] and the quantity of mouse stroma can vary across passages. This can result in variability

368 in the expression signal which can be wrongly inferred as CN changes, both from the tumor itself

369 and through cross hybridization of mouse RNA to the human microarray. Although improved

370 concordance in expression between PT and PDX can be achieved with RNA sequencing with the

371 removal of mouse reads[59,60], we observed that expression-based copy number inferences still

372 have low resolution and robustness. Hence, many cancer-driving genes, which are found mainly

373 in focal events with a size of 3Mb or lower[61-64], cannot be evaluated for PDX-specific alterations.

374 These issues are further worsened by the lack of tissue-matched normal gene expression profiles

375   for calibration[37], which have been only intermittently available but can substantially impact copy
376   number inferences. Because of these considerations, the question of how much PDXs evolve as
377   a consequence of mouse-specific selective pressures cannot be adequately addressed by
378   expression data.

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

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

399   We observed no evidence for systematic mouse environment-induced selection for cancer
400   or treatment-related genes via copy number changes, though individual cases vary (see example
401   in Extended Data Fig. 6c). Moreover, only a small fraction of sample pairs (2.44%, 43 out of 1758)
402   shows large CNA discordance (see METHODS), suggesting that clonal selection out of a complex
403   population is rare. These results indicate that the variations observed in PDXs are mainly due to
404   spontaneous intratumoral evolution rather than murine pressures (Supplementary Note 17).

405   In summary, our in-depth tracking of CNAs throughout PDX engraftment and passaging
406   confirms that tumors engrafted and passaged in PDX models maintain a high degree of molecular
407   fidelity to the original patient tumors and their suitability for pre-clinical drug testing. At the same
408   time, our study does not rule out that PDXs will evolve in individual trajectories over time, and for

409 therapeutic dosing studies, the best practice is to confirm the existence of expected molecular
410 targets and obtain sequence characterizations in the cohorts used for testing as close to the time
411 of the treatment study as is practicable.
412

475

476   **AUTHOR CONTRIBUTIONS**

477 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 and jointly
478 supervised the study. X.Y.W. organized the study, collected and structured the data, and
479 designed and carried out the analyses. J.G. collected and organized the EurOPDX data and
480 carried out the analyses. X.Y.W., E.Medico and J.H.C. wrote the manuscript. J.G, C.I., Z.-M.Z.,
481 A.S., and M.W.L. contributed to the refinement of the manuscript. A.S. and M.W.L. developed the
482 workflows. A.S., Z.-M.Z., M.W.L., and Y.-S.S. assisted with the computational analyses. R.J.,
483 C.F., J.Randjelovic, D.A.D., J.Rosains and B.D. assisted with the workflow development and data
484 collection and organization on the Cancer Genomics Cloud. R.d.B. and R.E.B. contributed to
485 sample selection and processing of EurOPDX data. C.J.B., R.P., L.C., Y.A.E., J.H.D., S.S.,
486 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.,
487 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.,
488 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.-
489 K.Y., C.L., E.Medico and J.H.C. contributed the sequencing and array data. C.J.B., E.Medico and
490 J.H.C. directed the project. The named author list describes the primary contributors of data and
491 analysis to the project, though these studies were supported by consortium-wide activities. All
492 members of the PDXNet and EurOPDX Consortia participated in group discussions or supportive
493 analyses regarding the study design, data standards, sample collection, or data analysis
494 approaches.
495

496 **ETHICS COMPLIANCE**
497 All xenograft studies were completed in accordance with animal research ethics regulations. For
498 details, see METHODS and references provided for each contributing group.
499

500 **COMPETING INTERESTS**
501 A.L.W and B.E.W receive a portion of royalties if University of Utah licenses certain PDX models
502 to for-profit entities. M.T.L is a founder of, and equity stake holder in, Tvardi Therapeutics Inc., a
503 founder of, and limited partner in, StemMed Ltd., and a Manager in StemMed Holdings LLC. He
504 also receives a portion of royalties if Baylor College of Medicine licenses certain PDX models to
505 for-profit entities. J.A.R. serves as a consultant and received stocks from Genprex, Inc., and
506 receives royalties from patents issued. F.M.-B. reports receiving commercial research grants from
507 Novartis, AstraZeneca, Calithera, Aileron, Bayer, Jounce, CytoMx, eFFECTOR, Zymeworks,
508 PUMA Biotechnology, Curis, Millennium, Daiichi Sankyo, Abbvie, Guardant Health, Takeda,
509 Seattle Genetics, and GlaxoSmithKline as well as grants and travel related fees from Taiho,
510 Genentech, Debiopharm Group, and Pfizer. She also served as a consultant to Pieris, Dialectica,

15

REFERENCES

1. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Disease models & mechanisms* **1**, 78-82 (2008).
2. Walrath, J.C., Hawes, J.J., Van Dyke, T. & Reilly, K.M. Genetically engineered mouse models in cancer research. *Advances in cancer research* **106**, 113-164 (2010).
3. Hait, W.N. Anticancer drug development: the grand challenges. *Nature Reviews Drug Discovery* **9**, 253-254 (2010).
4. Shultz, L.D., Ishikawa, F. & Greiner, D.L. Humanized mice in translational biomedical research. *Nature Reviews Immunology* **7**, 118-130 (2007).
5. Brehm, M.A., Shultz, L.D. & Greiner, D.L. Humanized mouse models to study human diseases. *Current opinion in endocrinology, diabetes, and obesity* **17**, 120-125 (2010).
6. Hidalgo, M. *et al.* Patient-Derived Xenograft Models: An Emerging Platform for Translational Cancer Research. *Cancer Discovery* **4**, 998-1013 (2014).
7. Byrne, A.T. *et al.* Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nature Reviews 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 Research* **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 Research* **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. *Nature Medicine* **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. *Clinical Cancer Research* **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. *Nature Medicine* **21**, 1318-1325 (2015).
16. Hidalgo, M. *et al.* A Pilot Clinical Study of Treatment Guided by Personalized Tumorgrafts in Patients with Advanced Cancer. *Molecular Cancer Therapeutics* **10**, 1311-1316 (2011).

557 17. Tentler, J.J. *et al.* Patient-derived tumour xenografts as models for oncology drug
558   development. *Nature reviews. Clinical oncology* **9**, 338-350 (2012).
559 18. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-
560   cell resolution. *Nature* **518**, 422-426 (2014).
561 19. Cho, S.-Y. *et al.* Unstable genome and transcriptome dynamics during tumor metastasis
562   contribute to therapeutic heterogeneity in colorectal cancers. *Clinical Cancer Research*
563   **25**, 2821-2834 (2019).
564 20. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft.
565   *Nature* **464**, 999-1005 (2010).
566 21. Giessler, K.M. *et al.* Genetic subclone architecture of tumor clone-initiating cells in
567   colorectal cancer. *The Journal of Experimental Medicine* **214**, 2073-2088 (2017).
568 22. Sato, K. *et al.* Multiregion Genomic Analysis of Serially Transplanted Patient-derived
569   Xenograft Tumors. *Cancer Genomics Proteomics* **16**, 21-27 (2019).
570 23. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution.
571   *Nature Genetics* **49**, 1567-1575 (2017).
572 24. Kim, H. *et al.* High-resolution deconstruction of evolution induced by chemotherapy
573   treatments in breast cancer xenografts. *Scientific Reports* **8**, 17937 (2018).
574 25. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic
575   characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-1130 (2013).
576 26. He, S. *et al.* PDXliver: a database of liver cancer patient derived xenograft mouse models.
577   *BMC Cancer* **18**, 550 (2018).
578 27. Zare, F., Hosny, A. & Nabavi, S. Noise cancellation using total variation for copy number
579   variation detection. *BMC Bioinformatics* **19**, 361 (2018).
580 28. Wineinger, N.E. & Tiwari, H.K. The Impact of Errors in Copy Number Variation Detection
581   Algorithms on Association Results. *PLOS ONE* **7**, e32396 (2012).
582 29. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor
583   sequencing data. *Ann Oncol* **26**, 64-70 (2015).
584 30. Talevich, E., Shain, A.H., Botton, T. & Bastian, B.C. CNVkit: Genome-Wide Copy Number
585   Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**,
586   e1004873 (2016).
587 31. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**,
588   1134-1140 (2013).
589 32. Taylor, A.M. *et al.* Genomic and Functional Approaches to Understanding Cancer
590   Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).
591 33. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the
592   National Academy of Sciences* **107**, 16910-16915 (2010).
593 34. Woo, X.Y. *et al.* Genomic data analysis workflows for tumors from patient-derived
594   xenografts (PDXs): challenges and guidelines. *BMC Medical Genomics* **12**, 92 (2019).
595 35. Evrard, Y.A. *et al.* Systematic Establishment of Robustness and Standards in Patient-
596   Derived Xenograft Experiments and Analysis. *Cancer Research* **80**, 2286-2297 (2020).
597 36. Conway, T. *et al.* Xenome--a tool for classifying reads from xenograft samples.
598   *Bioinformatics* **28**, i172-i178 (2012).
599 37. Ben-David, U., Mayshar, Y. & Benvenisty, N. Virtual karyotyping of pluripotent stem cells
600   on the basis of their global gene expression profiles. *Nature Protocols* **8**, 989-997 (2013).
601 38. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse
602   models reveals driver-specific routes to tumorigenesis. *Nature Communications* **7**, 12160
603   (2016).
604 39. Motulsky, H.J. & Brown, R.E. Detecting outliers when fitting data with nonlinear regression
605   – a new method based on robust nonlinear regression and the false discovery rate. *BMC
606   Bioinformatics* **7**, 123 (2006).

607 40. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*
608   **173**, 321-337.e10 (2018).
609 41. Patterson, S.E. *et al.* The clinical trial landscape in oncology and connectivity of somatic
610   mutational profiles to targeted therapies. *Human Genomics* **10**, 4 (2016).
611 42. Patterson, S.E., Statz, C.M., Yin, T. & Mockus, S.M. Utility of the JAX Clinical
612   Knowledgebase in capture and assessment of complex genomic cancer data. *npj*
613   *Precision Oncology* **3**, 2 (2019).
614 43. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction
615   across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
616 44. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular portraits of human breast
617   tumours. *Nature* **490**, 61-70 (2012).
618 45. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular characterization of human
619   colon and rectal cancer. *Nature* **487**, 330-337 (2012).
620 46. The Cancer Genome Atlas Research, N. *et al.* Comprehensive molecular profiling of lung
621   adenocarcinoma. *Nature* **511**, 543-550 (2014).
622 47. The Cancer Genome Atlas Research, N. *et al.* Comprehensive genomic characterization
623   of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
624 48. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
625   anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
626 49. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia.
627   *Nature* **569**, 503-508 (2019).
628 50. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer:
629   Methodology and application to glioma. *Proceedings of the National Academy of Sciences*
630   **104**, 20007-20012 (2007).
631 51. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets
632   of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**, R41
633   (2011).
634 52. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
635   interpreting genome-wide expression profiles. *Proceedings of the National Academy of*
636   *Sciences* **102**, 15545-15550 (2005).
637 53. Mootha, V.K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are
638   coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273 (2003).
639 54. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New*
640   *England Journal of Medicine* **376**, 2109-2121 (2017).
641 55. Ben-David, U., Beroukhim, R. & Golub, T.R. Genomic evolution of cancer models: perils
642   and opportunities. *Nature Reviews Cancer* **19**, 97-109 (2019).
643 56. Mer, A.S. *et al.* Integrative pharmacogenomics analysis of patient-derived xenografts.
644   *Cancer Research* **79**, 4539-4550 (2019).
645 57. Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nature*
646   *Genetics* **47**, 312-319 (2015).
647 58. Park, E.S. *et al.* Cross-species hybridization of microarrays for studying tumor
648   transcriptome of brain metastasis. *Proceedings of the National Academy of Sciences* **108**,
649   17456-17461 (2011).
650 59. Liu, Y. *et al.* Gene expression differences between matched pairs of ovarian cancer patient
651   tumors and patient-derived xenografts. *Scientific Reports* **9**, 6314 (2019).
652 60. Isella, C. *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel
653   clinically relevant subtypes of colorectal cancer. *Nature Communications* **8**, 15107 (2017).
654 61. Leary, R.J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and
655   sequence alterations in breast and colorectal cancers. *Proceedings of the National*
656   *Academy of Sciences* **105**, 16224-16229 (2008).

657 62. Bierkens, M. *et al.* Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and
658   tumor suppressor in cervical carcinogenesis. *Genes, Chromosomes and Cancer* **52**, 56-
659   68 (2013).
660 63. Krijgsman, O., Carvalho, B., Meijer, G.A., Steenbergen, R.D.M. & Ylstra, B. Focal
661   chromosomal copy number aberrations in cancer—Needles in a genome haystack.
662   *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1843**, 2698-2704 (2014).
663 64. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature*
664   **463**, 893–898 (2010).
665 65. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines
666   lung cancer evolution. *Science* **346**, 251-256 (2014).
667 66. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas
668   defined by multiregion sequencing. *Nature Genetics* **46**, 225-233 (2014).
669 67. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by
670   Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892 (2012).
671 68. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature*
672   **567**, 479-485 (2019).
673

674

675 **FIGURE LEGENDS**

676 **Fig. 1: Patient derived xenograft datasets used for copy number profiling across 16 tumor**

677 **types.** (**a**) Numbers of PDX models for each tumor type, with models also having multiple PDX

678 samples or having matched patient tumor samples specified. (**b**) Distributions of datasets by

679 passage number and assay platform for patient tumors and PDX samples, separated by tumor

680 type. "Late" passages include P18, P19 and P21 samples.

681

682 **Fig. 2: Comparisons of resolution and accuracy for copy number alterations estimated by**

683 **DNA-based and expression-based methods.** (**a**) Pairwise comparisons of distributions of

684 segment size (Mb) of CNAs estimated by different measurement platforms in the validation

685 dataset. CNAs are regions with ($|\log_2(CN\ ratio)| \geq 0.1$). P-values indicate significance of difference

686 between distributions by two-sided Wilcoxon rank sum test. (**b**) Pairwise comparisons of

687 distributions of $\log_2(CN\ ratio)$ of CNA segments. P-values were computed by two-sided

688 Kolmogorov-Smirnov test. (**c**) Distributions of Pearson correlation coefficient of median-centered

689 $\log_2(CN\ ratio)$ in 100-kb windows from CNA segments between pairs of samples estimated by

690 different platforms. Samples with non-aberrant profiles in SNP array and WES data are omitted

691 (5-95% inter-percentile range of $\log_2(CN\ ratio) < 0.3$). P-values were computed by two-sided

692 Wilcoxon rank sum test. In the boxplots, the center line is the median, box limits are the upper

693 and lower quantiles, whiskers extend 1.5 × the interquartile range, dots represent the outliers. (**d**)

694 Examples of CNA profiles in comparisons of different platforms. Pearson correlation coefficients

695 of CNA segments between pairs of samples are shown on the right. In all the plots, SNP: SNP

696 array, WES: whole-exome sequencing, RNASEQ: RNA sequencing, EXPARR: gene expression
697 array, NORM: normalization by median expression of normal samples, TUM: normalization by
698 median expression of tumor samples, see Supplementary Table 3 for number of samples per
699 group.
700

701 **Fig. 3: Comparisons of copy number alterations from patient tumor to early and late PDX**
702 **passages.** (**a-c**) Distributions of Pearson correlation coefficient of gene-based copy number,
703 estimated by (**a**) SNP array, (**b**) WES, and (**c**) WGS, between: PT-PDX samples from the same
704 model; PDX-PDX samples of the same model; samples of different models from a common tumor
705 type and contributing center. P-values were computed by one-sided Wilcoxon rank sum test (ns:
706 not significant, $p > 0.05$). Number of data points are indicated in the legend. (**d-f**) Distributions of
707 Pearson correlation coefficients of gene-based copy number, estimated by (**d**) SNP array, (**e**)
708 WES, and (**f**) WGS, among patient tumor and PDX passages of the same model. Comparisons
709 relative to PT and P0 are shown (higher passages are shown in Extended Data Fig. 5). In the
710 boxplots, the center line is the median, box limits are the upper and lower quantiles, whiskers
711 extend 1.5 × the interquartile range, dots represent the all data points. (**g**) Schematic of lineage
712 splitting during passaging and expansion of tumors into multiple mice. This is a simplified
713 illustration for passaging procedures in which different fragments of a tumor are implanted into
714 different mice. (**h**) Pearson correlation distributions for PDX sample pairs of different lineages and
715 sample pairs within the same lineage: for JAX SNP array, PDMR WES, and EuroPDX WGS
716 datasets. P-values were computed by one-sided Wilcoxon rank sum test. For all boxplots and
717 violin plots, number of pairwise correlations are indicated in the horizontal axis labels.
718

719 **Fig. 4: Cancer gene sets analysis for copy number altered genes during engraftment and**
720 **passaging.** (**a**) Distribution of proportion of altered genes between pairwise PT-PDX or PDX-PDX
721 comparisons of the same model in various gene sets. Protein-coding: protein-coding genes
722 annotated by Ensembl; Oncogenic pathways: genes in oncogenic signaling pathways identified
723 by TCGA; JAX CKB Amp/Del: genes with copy number gain or over-expression / copy number
724 loss or under-expression associated with therapeutic sensitivity or resistance or changes in drug
725 response; Census Amp Del: genes from Cancer Gene Census frequently altered by amplifications
726 or deletions. CNA genes were identified by |residual| > 0.5 from linear regression model. (**b**)
727 Distribution of proportion of altered genes between pairwise PT-PDX or PDX-PDX comparisons
728 of the same model in various gene sets within breast cancer, colorectal cancer, lung
729 adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) models. TCGA Gistic

730 Amp/Del: significantly amplified/deleted genes from TCGA GISTIC analysis for the corresponding

731 tumor type. For all violin plots, P-values were computed by one-sided Wilcoxon rank sum test (ns:

732 not significant, $p > 0.1$); number of pairwise comparisons are indicated in the plot title, number of

733 genes per gene set is indicated in the horizontal axis labels. (**c**) Recurrence frequency of protein

734 coding genes with copy number alterations, |residual| > 1, across all models in PT-PDX and PDX-

735 PDX comparisons. Number of models is indicated in the horizontal axis labels.

736

737 **Fig. 5: Absence of mouse-driven recurrent CNAs during engraftment and propagation of**

738 **colorectal (CRC) and breast cancer (BRCA) PDXs.** (**a**) Bar charts representing genome-wide

739 GISTIC G-score for amplifications and deletions in each of the three cohorts of CRC (87 trios)

740 and BRCA (43 trios): PT, PDX-early (P0-P1 for CRC, P0-P2 for BRCA), PDX-late (P2-P7 for CRC,

741 P3-P9 for BRCA). (**b-c**) Scatter plots comparing gene-level GISTIC G-score between each of the

742 three cohorts, for (**b**) CRC and (**c**) BRCA. Bottom-right panels of (**b**) and (**c**): scatter plots

743 comparing ΔG-scores from PT to PDX-early and from PDX early to PDX-late. (**d-e**) Scatter plots

744 comparing GSEA Normalized Enrichment Score (NES) for gene sets between each of the three

745 cohorts, for (**d**) CRC (**e**) and BRCA. Bottom-right panels of (**d**) and (**e**): scatter plots comparing

746 ΔNES from PT to PDX-early and from PDX-early to PDX-late.

747

748 **Fig. 6: Comparison of CNA variation during PDX engraftment and passaging to CNA**

749 **variation among patient multi-region, tumor relapse, and metastasis samples.** (**a**)

750 Distributions of Pearson correlation coefficients of gene-based copy number for lung

751 adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and other lung cancer

752 subtypes, comparing different datasets. TracerX multiregion: multi-region tumor samples of the

753 same patient from TRACERx (92 patient tumors, 295 multi-region samples); PT-PDX samples of

754 the same model; PDX-PDX samples of the same model. P-values were computed by two-sided

755 Wilcoxon rank sum test (ns: not significant, $p > 0.05$). (**b**) Distributions of proportion of altered

756 genes between multi-region tumor pairs from TRACERx, and PT-PDX and PDX-PDX pairs for

757 various gene sets for LUAD and LUSC. Gene sets and CNA thresholds are the same as Fig. 4.

758 TCGA Gistic Amp/Del and JAX CKB Amp Del gene sets are shown (other gene sets are shown

759 in Extended Data Fig. 8). P-values were computed by one-sided Wilcoxon rank sum test. Number

760 of genes per gene set are indicated in the plot title. (**c**) Distributions of Pearson correlation

761 coefficients of gene-based copy number between intra-patient PT (primary/relapse/metastasis)

762 pairs from the same patient and corresponding PT-PDX (derived from the same model; a different

763 PT sample from the same patient generates a different model) pairs for the same set of patients.

764 P-values were computed by two-sided Wilcoxon rank sum test (ns: not significant, $p > 0.05$).

765 Number of patients and models are indicated in the plot title. For all box plots and violin plots,

766 number of pairwise comparisons are indicated in the horizontal axis labels. In all boxplots, the

767 center line is the median, box limits are the upper and lower quantiles, whiskers extend 1.5 × the

768 interquartile range, dots represent the all data points. (**d**) CNA profiles of PT and PDX samples

769 from patients with PDX models derived from multiple PT collection (primary/relapse/metastasis).

770

771

772 **METHODS**

773

774 **Experimental details for sample collection, PDX engraftment and passaging, and array or**

775 **sequencing**

776 See Supplementary Methods.

777

778 **Consolidating tumor types from different datasets**

779 As the terminology of tumor types/subtypes by the different contributing centers were not

780 consistent, we used the Disease Ontology database[69] (http://disease-ontology.org/), cancer types

781 listed in NCI website (https://www.cancer.gov/types) and in TCGA publications[70,71] to unify and

782 group the tumor types/subtypes under broader terms as shown in Fig.1 and Supplementary Table

783 2.

784

785 **Copy number alteration (CNA) estimation methods**

786 ***SNP array.*** The estimation of CNA profiles from SNP array were detailed previously[34]. In short,

787 for Affymetrix Human SNP 6.0 arrays, PennCNV-Affy and Affymetrix Power Tools[72] were used to

788 extract the B-allele frequency (BAF) and Log R Ratio (LRR) from the CEL files. Due to the

789 absence of paired-normal samples, the allele-specific signal intensity for each PDX tumor were

790 normalized relative to 300 randomly selected sex-matched Affymetrix Human SNP 6.0 array CEL

791 files obtained from the International HapMap project[73]. For Illumina Infinium Omni2.5Exome-8

792 SNP arrays (v1.3 and v1.4 kit), the Illumina GenomeStudio software was used to extract the B-

793 allele frequency (BAF) and Log R Ratio (LRR) from the signal intensity of each probe. The single

794 sample mode of the Illumina GenomeStudio was used, which normalizes the signal intensities of

795 the probes with an Illumina in-house dataset. The single tumor version of ASCAT[33] (v2.4.3 for

796 JAX SNP data, v2.5.1 for SIBS SNP data) was used for GC correction, predictions of the

797 heterozygous germline SNPs based on the SNP array platform, and estimation of ploidy, tumor

798   content and allele-specific copy number segments. The resultant copy number segments were

799   annotated with $\log_2$ ratio of total copy number relative to predicted ploidy from ASCAT.

800   ***Whole-exome sequencing (WES) data.*** Aligned bams (See Supplementary Methods) were

801   subset to target region by GATK 4.0.5.1, and SAMTools[74] v0.1.18 was used to generate the pileup

802   for each sample. Pileup data were used for CNA estimation as calculated with Sequenza[29] v2.1.2.

803   Both tumor and normal data, that utilized the same capture array, were used as input. pileup2seqz

804   and GC-windows (-w 50) modules from sequenza-utils.py utility were used to create the native

805   seqz format file for Sequenza and compute the average GC content in sliding windows from hg38

806   genome, respectively. Finally, we ran the three Sequenza modules with these modified

807   parameters (sequenza.extract: assembly = "hg38", sequenza.fit: chromosome.list = 1:23, and

808   sequenza.results: chromosome.list = 1:23) to estimate the segments of copy number

809   gains/losses. Finally, segments lacking read counts, in which ≥50% of the segment with zero read

810   coverage, were removed. A reference implementation of this workflow (Supplementary Fig. 71)

811   is developed and deployed in the Cancer Genomics Cloud by Seven Bridges

812   (https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-

813   workflow/, https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-

814   xenome-tumor-normal-workflow/).

815   ***Low-pass whole-genome sequencing (WGS) data.*** For EuroPDX CRC liver metastasis data,

816   raw copy number profiles for each sample were estimated by QDNAseq[75] R package v1.20 by

817   dividing the human reference genome in non-overlapping 50 kb windows and counting the

818   number of reads (See Supplementary Methods) in each bin. Bins in problematic regions were

819   removed[76]. Read counts were corrected for GC content and mappability by a LOESS regression,

820   median-normalized and $\log_2$-transformed. Values below −1000 in each chromosome were floored

821   to the first value greater than −1000 in the same chromosome. Raw $\log_2$ ratio values were then

822   segmented using the ASCAT[33] algorithm implemented in the ASCAT R package v2.0.7. For

823   EuroPDX BRCA tumors, raw copy number profiles were estimated for each sample by dividing

824   the human reference genome in non-overlapping 20 kb windows and counting the number of

825   reads (See Supplementary Methods) in each bin. Only reads with at least mapping quality 37

826   were considered. Bins within problematic regions (i.e. multimapper regions) were excluded.

827   Downstream analysis to estimate copy number was conducted as described above.

828   ***RNA-sequencing (RNA-Seq) and gene expression microarray (EXPARR) data.*** For

829   expression-based copy number inference, we referred to the previous protocols for e-karyotyping

830   and CGH-Explorer[37,38,77,78]. For each cancer type, expression values (see Supplementary

831   Methods) of tumor and corresponding normal samples were merged in a single table, and gene

23

832   identifiers were annotated with chromosomal nucleotide positions. Genes located on sex

833   chromosomes were excluded. Genes which values below 1 TPM (RNAseq) or probeset $\log_2$-

834   values below 6 (microarray) in more than 20% of the analyzed dataset were removed. Remaining

835   gene expression values below the thresholds were respectively raised to 1 TPM or $\log_2$-value of

836   6. In the case of multiple transcripts (RNA-seq) or probesets (microarray) per gene, the one with

837   the highest median value across the entire dataset was selected. According to the e-karyotyping

838   protocol, the sum of squares of the expression values relative to their median expression across

839   all samples was calculated for each gene, and 10% most highly variable genes were removed.

840   For each gene, the median $\log_2$ expression value in normal samples was subtracted from the log2

841   expression value in each tumor sample and subsequently input in CGH-explorer. For tumor-only

842   datasets, the median $\log_2$ expression value in the same set of tumor samples was instead

843   subtracted. The preprocessed expression profiles of each sample were individually analyzed

844   using CGH-Explorer (http://heim.ifi.uio.no/bioinf/Projects/CGHExplorer/). CGH-PCF analysis was

845   carried out to call copy number according to parameters previously reported[23]: least allowed

846   deviation = 0.25; least allowed aberration size = 30; winsorize at quantile = 0.001; penalty = 12;

847   threshold = 0.01.

848

849   **Statistical Methods**

850   All statistical analysis for data comparison were performed using either one-tailed or two-tailed

851   Wilcoxon rank sum test, two-tailed Kolmogorov–Smirnov test, or one-tailed Wilcoxon signed rank

852   test.

853

854   **Filtering and gene annotation of copy number segments**

855   Copy number (CN) segments with $\log_2$ copy number ratio estimated from the various platforms

856   were processed in the following steps (Extended Data Fig. 3). Segments <1kb were filtered based

857   on the definition of CNA[79]. In addition, SNP array segments had to be covered by >10 probes,

858   with an average probe density of 1 probe per 5kb. The copy number segments were then binned

859   into 10kb windows to derive the median $\log_2$(CN ratio), which was subsequently used to re-center

860   the copy number segments. Median-centered copy number segments were visualized using IGV[80]

861   v2.4.13 and GenVisR[81] v1.16.1. Median-centered copy number of genes were calculated by

862   intersecting the genome coordinates of copy number segments with the genome coordinates of

863   genes (Ensembl Genes 93 for human genome assembly GRCh38, Ensembl Genes 96 for human

864   genome assembly GRCh37). In the case where a gene overlaps multiple segments, the most

865 conservative (lowest) estimate of copy number was used to represent the copy number of the
866 entire intact gene.
867
868 **Comparison of CN gains and losses**
869 For the comparison of resolution, range of CN values and frequency of gains and losses between
870 different platforms and analysis methods, we defined copy number gain or loss segments as –
871 Gain: $\log_2$(CNratio) > 0.1; Loss: $\log_2$(CN ratio) < -0.1.
872
873 **Correlation of CNA profiles**
874 The overall workflow to compare CNA profiles is shown in Extended Data Fig. 3. PDX samples
875 without passage information were omitted in the following downstream analysis. The copy number
876 segments were binned into 100kb-windows or smaller using Bedtools[82] v2.26.0, and the variance
877 of $\log_2$(CN ratio) and 5-95% inter-percentile range of $\log_2$(CN ratio) values across all the bins were
878 calculated as a measure of degree of aberration for each CNA profile. A non-aberrant profile
879 results in a low variance or range. While variance can be biased for CNA profiles with small
880 segments of extreme gains or losses, we preferred the use of 5-95% inter-percentile range of
881 $\log_2$(CN ratio) to identify samples with low degree of aberration, such that a narrow range indicates
882 ≥90% of the genome has very low-level gains and losses. The similarity of two CNA profiles is
883 quantified by the Pearson correlation coefficient of $\log_2$(CN ratio) of 100kb-windows binned from
884 segments or genes between 2 samples. Gene-based and segment-based (100kb windows)
885 correlations were highly similar (data not shown). Using correlation avoided the issue of making
886 copy number gain and loss calls based on thresholds. Sample-based variations in baseline due
887 to median-normalization and range in copy number values could introduce further inconsistencies
888 gain and loss calls between samples. Such variations are further impacted by sample-specific
889 variation in human stromal contamination or sensitivity of copy number detection by different
890 platforms. As median-centering of each CNA profile approximates normalization by the sample
891 ploidy, we confirmed that in general ploidy (estimated from ASCAT analysis of SNP array
892 samples) had no association with the copy number correlation values (Pearson's product-moment
893 correlation, p > 0.05, cor = 0.0248). One caveat of our approach, however, is that it cannot
894 distinguish genome-wide multiplication of ploidy between samples, as the correlation statistic is
895 invariant to such genome-wide transformations. As such we cannot assess whether ploidy
896 changes occur between samples of a given model.
897 ***Comparison of CNA profiles between different platforms.*** The copy number segments of each
898 pair of data were intersected and binned into 100kb-windows or smaller using Bedtools. The

25

899 Pearson correlation coefficient and linear regression model was calculated for the $\log_2$(CN ratio)

900 of the windows. Windows with discrepant copy number were identified by outliers of the linear

901 regression model defined by |studentized residual| > 3. These outlier windows were mapped to

902 their corresponding segments to identify the size of CNA events that were discordant between

903 the different copy number estimation methods. The proportion of the genome discordant CNA

904 was calculated from the summation of the outlier windows.

905 ***Identification of genes with CNA between different samples of the same model.*** To compare

906 the CNA profiles between different samples (PT or PDX) of the same model, the Pearson

907 correlation coefficient and linear regression model was calculated for the $\log_2$(CN ratio) of the

908 genes for each pair of data. Prior to that, deleted genes with $\log_2$(CN ratio) < -3 were rescaled to

909 -3 to avoid large shifts in the correlation coefficient and linear regression model due to extremely

910 negative values on the log scale. Extreme outliers of the linear regression model defined by

911 |studentized residual| > 3 were removed to derive an improved linear regression model[39] not

912 biased by few extreme values. Genes with copy number changes between the samples were

913 identified by the difference in $\log_2$(CN ratio) relative to the improved linear regression model of

914 |standard residual| < 0.5. We also removed some samples with low correlation due to sample

915 mislabeling as they displayed high correlation with samples from other models. We also omitted

916 samples with low correlation values (<0.6) which resulted from non-aberrant CNA profiles in

917 genomically stable tumors (5-95% inter-percentile range of $\log_2$(CN ratio) < 0.3, Supplementary

918 Fig. 62).

919 ***Identification of aberrant sample pairs with highly discordant CNA profiles.*** Aberrant CNA

920 profiles were identified based on the 5-95% inter-percentile range of $\log_2$(CN ratio) >0.5, for both

921 samples. Sample pairs with Pearson correlation <0.6 were selected as highly discordant CNA

922 profiles between them.

923 ***Association of mutations with copy number correlations.*** Mutational calls for each WES

924 sample used in this study were obtained using a tumor-normal variant calling workflow developed

925 for patient tumor and PDXs[35]. Subsequently, genes with either germline and somatic variants that

926 pass through the quality filters (FILTER = PASS or germline) and IMPACT = MODERATE or

927 HIGH by SnpEff (v4.3) annotation are labeled as mutated, and wildtype if otherwise. For SNP

928 array and WGS data, we collected the mutational status (wildtype or mutated) of TP53, BRCA1

929 and BRCA2 per model where available, which may or may not be obtained from the exact same

930 tumor samples used in this study. For the JAX SNP array dataset, variant calls (tumor-only) were

931 made from various targeted sequencing approaches (TruSeq Amplicon Cancel Panel, JAX

932 Cancer Treatment Profile panel and whole exome). The workflow and filtering criteria to call

933   mutations is described elsewhere[34]. For the HCI SNP array data, mutations were obtained from

934   whole exome sequencing (unpublished data) and were filtered for frameshift, inframe, missense,

935   and nonsense and splice-site mutations. For BCM SNP array data, mutational status were

936   obtained from clinical samples by immunohistochemistry or Sequenom[83] (unpublished data). For

937   WGS data, mutations were obtained from whole exome or targeted panel sequencing[84]

938   (unpublished data) and high-quality and likely functional mutations were retained. For each

939   sample pair with copy number correlations, mutational status of *TP53* or *BRCA* was obtained for

940   each individual sample for WES data, while the mutational status was available on a per model

941   basis for SNP and WGS data. *BRCA* is labelled as mutated when either *BRCA1* or *BRCA2* is

942   mutated. For mutations in DNA repair genes[85] from the WES data, each pair of samples was

943   classified as mutated if any DNA repair gene was reported to be mutated in either sample.

944

945   **Annotation with gene sets with known cancer or treatment-related functions**

946   A low copy number change threshold ($|\log_2$(CN ratio) change$| > 0.5$) was selected to include

947   genes with subclonal alterations. Copy number altered genes ($|$residual$| > 0.5$) were annotated

948   by various gene sets with cancer or treatment-related functions gathered from various databases

949   and publications (Extended Data Fig. 3):

950   1. Genes in 10 oncogenic signaling pathways curated by TCGA and were found to be frequently

951   altered in different cancer types[40].

952   2. Genes with gain in copy number or expression, or loss in copy number or expression that

953   conferred therapeutic sensitivity, resistance or increase/decrease in drug response from the JAX

954   Clinical Knowledgebase[41,42] (JAX-CKB) based on literature curation (https://ckbhome.jax.org/, as

955   of 06-18-2019).

956   3. Genes with evidence of promoting oncogenic transformation by amplification or deletion from

957   the Cancer Gene Census[43] (COSMIC v89).

958   4. Significantly amplified or deleted genes in TCGA cohorts of breast cancer[44], colorectal cancer[45],

959   lung adenocarcinoma[46] and lung squamous cell carcinoma[47] by GISTIC analysis, which identified

960   significantly altered genomic driver regions which can be used to differentiate tumor types and

961   subtypes.

962

963   **Identification of genes with recurrent copy number changes**

964   A stringent CNA threshold ($|\log2$(CN ratio) change$| > 1.0$ with respect to linear regression model)

965   was selected to distinguish genes with possible functional impact. Genes with $|$residual$| > 1.0$ with

966   respect to the improved regression linear model (without discriminating gain or loss) were

967 selected for each pairwise comparison between different samples of the same model. Pairwise
968 cases in which genes are deleted in both samples ($\log_2$(CN ratio) ≤ -3) are omitted. Recurrent
969 frequency for each gene across all models was calculated on a model basis such that genes with
970 copy number between multiple pairs of the same model was counted as once. This avoided the
971 bias towards models with many samples of similar copy number changes between the different
972 pairs.

973

974 **Drug response analysis using CCLE data**
975 We developed a pipeline to evaluate gene copy number effects on drug sensitivity[86,87] by using
976 the Cancer Cell Line Encyclopedia[48,88] (CCLE) cell line genomic and drug response data (CTRP
977 v2). We downloaded the CCLE drug response data from Cancer Therapeutics Response Portal
978 (www.broadinstitute.org/ctrp), and CCLE gene-level CNA and gene expression data from depMap
979 data portal ('public_19Q1_gene_cn.csv' and 'CCLE_depMap_19Q1_TPM.csv',
980 https://depmap.org/portal/download/). For CCLE drug response data, we used the area-under-
981 concentration-response curve (AUC) sensitivity scores for each cancer cell line and each drug. In
982 total, we collected gene-level $\log_2$ copy number ratio data derived from the Affymetrix SNP 6.0
983 platform from 668 pan-cancer CCLE cell lines, with a total of 545 cancer drugs tested. With the
984 CCLE gene-level CNA and AUC drug sensitivity scores, we performed gene-drug response
985 association analyses for genes with recurrent copy number changes. Pearson correlation p-
986 values between each gene's $\log_2$ (CN ratio) and each drug's AUC score across all cell lines were
987 calculated, and q-values were calculated by multiple testing Bonferroni correction. Significant
988 gene-CNA and drug associations were kept (q-value < 0.1) to further evaluate gene-expression
989 and drug response associations. If a gene's expression was also significantly correlated with AUC
990 drug sensitivity scores, particularly in the same direction (either positively or negatively correlated)
991 as the gene-CNA and drug association, that gene would be considered as significantly correlated
992 with drug response based on both its CNA and gene expression.

993

994 **Genomic Identification of Significant Targets in Cancer (GISTIC) analysis of WGS data**
995 We carried out GISTIC analysis to identify recurrent CNAs by evaluating the frequency and
996 amplitude of observed events. To obtain perfectly matching and comparable PT–PDX cohorts,
997 for GISTIC analysis, CRC trios in which at least one sample displaying non-aberrant CNA profiles
998 were excluded from the analysis resulting in a total of 87 triplets. The GISTIC[51] algorithm (GISTIC
999 2.0 v6.15.28) was applied on the segmented profiles using the GISTIC GenePattern module
1000 (https://cloud.genepattern.org/), with default parameters and genome reference files

28

1001 Human_Hg19.mat for EuroPDX CRC data and hg38.UCSC.add_miR.160920.refgene.mat for

1002 EuroPDX BRCA data. For each dataset, GISTIC provides separate results (including segments,

1003 G-scores and FDR q-values) separately for recurrent amplifications and recurrent deletions.

1004 Deletion G-scores were assigned negative values for visualization. We observed that the G-Score

1005 range was systematically lower in PT cohorts, which is likely the result of the dilution of CNA by

1006 normal stromal DNA. In contrast, human stromal DNA in PDX samples were lower or negligible.

1007 To account for this difference in gene-level G-scores, PDXs at early and late passages were

1008 scaled with respect to PT gene-level G-score values using global linear regression, separately for

1009 amplification and deletion outputs.

1010

1011 **Gene set enrichment analysis (GSEA) of WGS data**

1012 To assess the biological functions associated with the recurrent alterations detected by the

1013 GISTIC analysis, we performed GSEAPreranked analysis[52,53] (GSEA v3.0) on gene-level GISTIC

1014 G-score profiles, for both amplifications and deletions. In particular, we applied the algorithm with

1015 1000 permutations on various gene set collections from the Molecular Signatures Database[89,90]

1016 (MSigDB v6.2): H (Hallmark), C2 (Curated : CGP chemical and genetic perturbations, CP

1017 canonical pathways), C5 (Gene Ontology: BP biological process, MF molecular function, CC

1018 cellular component) and C6 (Oncogenic Signatures) composed of gene sets respectively. We

1019 also included gene sets with known cancer or treatment-related functions described in an earlier

1020 section. We noted that multiple genes with contiguous chromosomal locations, typically in

1021 recurrent amplicons, generated spurious enrichment for gene sets which consists of multiple

1022 genes of adjacent positions, while very few or none of them had a significant GISTIC G-score. To

1023 avoid this confounding issue, we only considered the "leading edge genes", i.e. those genes with

1024 increasing Normalized Enrichment Score (NES) up to its maximum value, that contribute to the

1025 GSEA significance for a given gene set. The leading-edge subset can be interpreted as the core

1026 that accounts for the gene set's enrichment signal (http://software.broadinstitute.org/gsea). We

1027 included a requirement that the leading edge genes passing the GISTIC G-score significant

1028 thresholds based on GISTIC q-value 0.25 (Supplementary Table 8 and Extended Fig. 7) make up

1029 at least 20% of the gene set. This 20% threshold was chosen as the minimal threshold at which

1030 gene sets assembled from TCGA-generated lists of genes with recurrent CNA in CRC or BRCA

1031 were identified as significant in GSEA (see Supplementary Table 9). Finally, gene sets with a NES

1032 greater than 1.5 and a FDR q-value of less than 0.05, which passed the leading edge criteria,

1033 were considered significantly enriched in genes affected by recurrent CNAs.

1034

29

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, CCLE cell line genomic and drug response data (CTRP v2), MSigDB v6.2 and TRACERx NSCLC data (DOI: 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 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/).

CONSORTIA

PDXNET Consortium (full list of members)

Xing Yi Woo[1,#,§], Anuj Srivastava[1], Zi-Ming Zhao[1], Michael W. Lloyd[4], Rajesh Patidar[7], Li Chen[7], Sandra Scherer[8], Matthew 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 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 Davies[26], Li Ding[12], Shunqiang Li[12], Ramaswamy Govindan[12], Jeffrey A. Moscow[27,§], Carol J. Bult[4,§], Jeffrey H. Chuang[1,§,*], Peter N.

1069 Robinson[1], Brian J. Sanderson[1], Steven B. Neuhauser[4], Lacey E. Dobrolecki[23], Xiaofeng Zheng[10],

1070 Mourad Majidi[24], Ran Zhang[24], Xiaoshan Zhang[24], Argun Akcakanat[25], Kurt W. Evans[25], Timothy

1071 A. Yap[25], Dali Li[25], Erkan Yucan[25], Christopher D. Lanier[25], Turcin Saridogan[25], Bryce P. Kirby[25],

1072 Min Jin Ha[29], Huiqin Chen[29], Scott Kopetz[30], David G. Menter[30], Jianhua Zhang[31], Shannon N.

1073 Westin[32], Michael P. Kim[33], Bingbing Dai[33], Don L. Gibbons[34], Coya Tapia[35], Vanessa B.

1074 Jensen[36], Gao Boning[37], John D. Minna[37], Hyunsil Park[37], Brenda C. Timmons[37], Luc Girard[37],

1075 Dylan Fingerman[11], Qin Liu[11], Rajasekharan Somasundaram[11], Min Xiao[11], Vashisht G. Yennu-

1076 Nanda[26], Michael T. Tetzlaff[38], Xiaowei Xu[38], Katherine L. Nathanson[39], Song Cao[12], Feng Chen[12],

1077 John F. DiPersio[12], Kian H. Lim[12], Cynthia X. Ma[12], Fernanda M. Rodriguez[12], Brian A. Van Tine[12],

1078 Andrea Wang-Gillam[12], Michael C. Wendl[12], Yige Wu[12], Matthew A. Wyczalkowski[12], Lijun Yao[12],

1079 Reyka Jayasinghe[12], Rebecca L. Aft[40], Ryan C. Fields[40], Jingqin Luo[12,40], Katherine C. Fuh[31], Vicki

1080 Chin[13], John DiGiovanna[13], Jeffrey Grover[13], Soner Koc[13], Sara Seepo[13], Tiffany Wallace[42],

1081 Chong-Xian Pan[43], Moon S. Chen, Jr[43], Luis G. Carvajal-Carmona[44], Amanda R. Kirane[45], May

1082 Cho[45], David R. Gandara[45], Jonathan W. Riess[45], Tiffany Le[45], Ralph W. deVere White[45], Clifford

1083 G. Tepper[45], Hongyong Zhang[46], Nicole B. Coggins[46], Paul Lott[46], Ana Estrada[46], Ted Toal[46],

1084 Alexa Morales Arana[46], Guadalupe Polanco-Echeverry[46], Sienna Rocha[46], Ai-Hong Ma[44],

1085 Nicholas Mitsiades[47,48], Salma Kaochar[47], Bert W. O'Malley[48], Matthew J. Ellis[23], Susan G.

1086 Hilsenbeck[23], Michael Ittmann[49]

1087

EurOPDX Consortium (full list of members)

1089 Jessica Giordano[2,3,#], Roebi de Bruijn[5], Francesco Galimi[2,3], Andrea Bertotti[2,3], Adam Lafferty[14],

1090 Alice C. O'Farrell[14], Elodie Modave[15,16], Diether Lambrechts[15,16], Petra ter Brugge[5], Violeta

1091 Serra[17], Elisabetta Marangoni[18], Rania El Botty[18], Claudio Isella[2,3,§], Livio Trusolino[2,3,§], Annette

1092 T. Byrne[14,§], Jos Jonkers[5,§], Enzo Medico[2,3,§,*], Simona Corso[2,3], Alessandro Fiori[2,3], Silvia

1093 Giordano[2,3], Marieke van de Ven[5], Daniel S. Peeper[5], Ian Miller[14], Cristina Bernadó[17], Beatriz

1094 Morancho[17], Lorena Ramírez[17], Joaquín Arribas[17], Héctor G. Palmer[17], Alejandro Piris-Gimenez[17],

1095 Laura Soucek[17], Ahmed Dahmani[18], Elodie Montaudon[18], Fariba Nemati[18], Virginie Dangles-

1096 Marie[18], Didier Decaudin[18], Sergio Roman-Roman[18], Denis G. Alférez[50], Katherine Spence[50],

1097 Robert B. Clarke[50], Mohamed Bentires-Alj[51], David K. Chang[52], Andrew V. Biankin[52], Alejandra

1098 Bruna[53], Martin O'Reilly[53], Carlos Caldas[53], Oriol Casanovas[54], Eva Gonzalez-Suarez[54],

1099 Purificacíon Muñoz[54], Alberto Villanueva[54], Nathalie Conte[55], Jeremy Mason[55], Ross Thorne[55],

1100 Terrence F. Meehan[55], Helen Parkinson[55], Zdenka Dudova[56], Ales Křenek[56], Dalibor Stuchlík[56],

1101 Olivier Elemento[57], Giorgio Inghirami[57], Anna Golebiewska[58], Simone P. Niclou[58], G. Bea A.

1102 Wisman[59], Steven de Jong[59], Petra Kralova[60], Radislav Sedlacek[60], Elisa Claeys[61], Eleonora

Leucci[61], Massimiliano Borsani[62], Luisa Lanfrancone[62], Pier Giuseppe Pelicci[62], Gunhild Mari Mælandsmo[63], Jens Henrik Norum[63], Emilie Vinolo[64]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

[2] Department of Oncology, University of Torino, Candiolo, Italy

[3] Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy

[4] The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA

[5] Netherlands Cancer Institute, Amsterdam, Netherlands

[6] College of Medicine, Seoul National University, South Korea

[7] Frederick National Laboratory for Cancer Research, Frederick, MD, USA

[8] Department of Oncological Sciences, University of Utah Huntsman Cancer Institute, Salt Lake City, UT, USA

[9] Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

[10] Department of Bioinformatics and Computer Biology, The University of Texas M.D. 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, Shaanxi, People's Republic of China

[20] Department of Life Sciences, Ewha Womans University, Seoul, Seoul, South Korea

[21] Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA

[22] Department of Surgery, University of Utah Huntsman Cancer Institute, Salt Lake City, UT, USA

[23] Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA

1136   [24] Department of Thoracic & Cardiovascular Surgery, The University of Texas M.D. Anderson

1137   Cancer Center, Houston, TX, USA

1138   [25] Department of Investigational Cancer Therapeutics, The University of Texas M.D. Anderson

1139   Cancer Center, Houston, TX, USA

1140   [26] Department of Melanoma Medical Oncology, The University of Texas M.D. Anderson Cancer

1141   Center, Houston, TX, USA

1142   [27] Investigational Drug Branch, National Cancer Institute, Bethesda, MD, USA

1143   [28] A list of members and affiliations appears in the Supplementary Information.

1144   [29] Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston,

1145   TX, USA

1146   [30] Department of Gastrointestinal Medical Oncology, The University of Texas M.D. Anderson

1147   Cancer Center, Houston, TX, USA

1148   [31] Department of Genomic Medicine, The University of Texas M.D. Anderson Cancer Center,

1149   Houston, TX, USA

1150   [32] Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas

1151   M.D. Anderson Cancer Center, Houston, TX, USA

1152   [33] Department of Surgical Oncology, The University of Texas M.D. Anderson Cancer Center,

1153   Houston, TX, USA

1154   [34] Department of Thoracic/Head & Neck Medical Oncology, The University of Texas M.D.

1155   Anderson Cancer Center, Houston, TX, USA

1156   [35] Department of Translational Molecular Pathology, The University of Texas M.D. Anderson

1157   Cancer Center, Houston, TX, USA

1158   [36] Department of Veterinary Medicine & Surgery, The University of Texas M.D. Anderson

1159   Cancer Center, Houston, TX, USA

1160   [37] Hamon Center For Therapeutic Oncology, UT Southwestern Medical Center, Dallas, TX, USA

1161   [38] Department of Pathology and Laboratory Medicine, Hospital of the University of

1162   Pennsylvania, Philadelphia, PA, USA

1163   [39] Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA

1164   [40] Department of Surgery, Washington University School of Medicine in St. Louis, St. Louis, MO,

1165   USA

1166   [41] Department of Obstetrics and Gynecology, Washington University School of Medicine in St.

1167   Louis, St. Louis, MO, USA

1168   [42] Center to Reduce Cancer Health Disparities, National Cancer Institute, Bethesda, MD, USA

1169   [43] Department of Internal Medicine, Division of Hematology/Oncology, University of California

1170   Davis, Sacramento, CA, USA

1171   [44] Department of Biochemistry and Molecular Medicine, University of California Davis,

1172   Sacramento, CA, USA

1173   [45] UC Davis Comprehensive Cancer Center, University of California Davis, Sacramento, CA,

1174   USA

1175   [46] UC Davis Genome Center, University of California Davis, Sacramento, CA, USA

1176   [47] Department of Medicine, Baylor College of Medicine, Houston, TX, USA

1177   [48] Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA

1178   [49] Department of Pathology, Baylor College of Medicine, Houston, TX, USA

1179   [50] Manchester Breast Centre, Division of Cancer Sciences, University of Manchester, UK

1180   [51] University Hospital Basel, University of Basel, Switzerland

1181   [52] Institute of Cancer Sciences, University of Glasgow, UK

1182   [53] Cancer Research UK Cambridge Institute, Cambridge Cancer Centre, UK

1183   [54] Catalan Institute of Oncology, L'Hopistalet de Llobregat, Barcelona, Spain

1184   [55] European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust

1185   Genome Campus, Hinxton, Cambridge, UK

1186   [56] Institute of Computer Science, MazarykUniversity, Brno, Czech Republic

1187   [57] Weill Cornell Medical College, Cornell University, New York, NY, USA

1188   [58] NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of

1189   Health, Luxembourg, Luxembourg

1190   [59] University Medical Centre Groningen, Groningen, The Netherlands

1191   [60] Czech Center for Phenogenomics, Institute of Molecular Genetics, Vestec, Czech Republic

1192   [61] TRACE PDX platform, Katholieke Universiteit Leuven, Leuven, Belgium

1193   [62] European Institute of Oncology, Milan, Italy

1194   [63] Oslo University Hospital, Oslo, Norway

1195   [64] seeding science SPRL, Limelette, Belgium

1196   [#] These authors contributed equally to this work.

1197   [§] These authors jointly supervised this work.

1198   * Correspondence should be addressed to E.M. (enzo.medico@unito.it) or J.H.C.

1199   (jeff.chuang@jax.org).

1200

1201

34

1202 69. Schriml, L.M. *et al.* Human Disease Ontology 2018 update: classification, content and
1203   workflow expansion. *Nucleic Acids Research* **47**, D955-D962 (2018).
1204 70. The Cancer Genome Atlas, N. *et al.* Comprehensive genomic characterization of head
1205   and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
1206 71. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult
1207   Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).
1208 72. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-
1209   resolution copy number variation detection in whole-genome SNP genotyping data.
1210   *Genome Res* **17**, 1665-1674 (2007).
1211 73. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-796 (2003).
1212 74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-
1213   2079 (2009).
1214 75. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by
1215   shallow whole-genome sequencing with identification and exclusion of problematic
1216   regions in the genome assembly. *Genome research* **24**, 2022-2032 (2014).
1217 76. Desmedt, C. *et al.* Uncovering the genomic heterogeneity of multifocal breast cancer.
1218   *Journal of Pathology* **236**, 457-466 (2015).
1219 77. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal
1220   aberrations and recombination by allelic bias in RNA-Seq. *Nature Communications* **7**,
1221   12144 (2016).
1222 78. Lingjaerde, O.C., Baumbusch, L.O., Liestol, K., Glad, I.K. & Borresen-Dale, A.L. CGH-
1223   Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 821-822 (2005).
1224 79. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-
1225   454 (2006).
1226 80. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV):
1227   high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*
1228   **14**, 178-192 (2013).
1229 81. Skidmore, Z.L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012-3014
1230   (2016).
1231 82. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
1232   features. *Bioinformatics* **26**, 841-842 (2010).
1233 83. Zhang, X.M. *et al.* A renewable tissue resource of phenotypically stable, biologically and
1234   ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer
1235   Research* **73**, 4885-4897 (2013).
1236 84. Coussy, F. *et al.* A large collection of integrated genomically characterized patient-derived
1237   xenografts highlighting the heterogeneity of triple-negative breast cancer. *International
1238   Journal of Cancer* **145**, 1902-1912 (2019).
1239 85. Riaz, N. *et al.* Pan-cancer analysis of bi-allelic alterations in homologous recombination
1240   DNA repair genes. *Nature Communications* **8**, 857 (2017).
1241 86. Adams, D.J. *et al.* NAMPT Is the Cellular Target of STF-31-Like Small-Molecule Probes.
1242   *ACS Chemical Biology* **9**, 2247-2254 (2014).
1243 87. Viswanathan, V.S. *et al.* Dependency of a therapy-resistant state of cancer cells on a lipid
1244   peroxidase pathway. *Nature* **547**, 453-457 (2017).
1245 88. Stransky, N. *et al.* Pharmacogenomic agreement between two cancer cell line data sets.
1246   *Nature* **528**, 84-87 (2015).
1247 89. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-
1248   1740 (2011).
1249 90. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
1250   collection. *Cell Syst* **1**, 417-425 (2015).
1251
1252