

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

3
4 Xing Yi Woo^{1,a,#,§}, Jessica Giordano^{2,3,b,#}, Anuj Srivastava^{1,a}, Zi-Ming Zhao^{1,a}, Michael W. Lloyd^{4,a},
5 Roebi de Bruijn^{5,b}, Yun-Suhk Suh⁶, Rajesh Patidar^{7,a}, Li Chen^{7,a}, Sandra Scherer^{8,a}, Matthew
6 Bailey^{8,9,a}, Chieh-Hsiang Yang^{8,a}, Emilio Cortes-Sanchez^{8,a}, Yuanxin Xi^{10,a}, Jing Wang^{10,a},
7 Jayamanna Wickramasinghe^{11,a}, Andrew V. Kossenkov^{11,a}, Vito Rebecca^{11,a}, Hua Sun^{12,a}, R. Jay
8 Mashl^{12,a}, Sherri Davies^{12,a}, Ryan Jeon^{13,a}, Christian Frech^{13,a}, Jelena Randjelovic^{13,a}, Jacqueline
9 Rosains^{13,a}, Francesco Galimi^{2,3,b}, Andrea Bertotti^{2,3,b}, Adam Lafferty^{14,b}, Alice C. O'Farrell^{14,b},
10 Elodie Modave^{15,16}, Diether Lambrechts^{15,16}, Petra ter Brugge^{5,b}, Violeta Serra^{17,b}, Elisabetta
11 Marangoni^{18,b}, Rania El Botty^{18,b}, Hyunsoo Kim¹, Jong-Il Kim⁶, Han-Kwang Yang⁶, Charles Lee¹,
12 Dennis A. Dean, II^{12,a}, Brandi Davis-Dusenberry^{12,a}, Yvonne A. Evrard^{7,a}, James H. Doroshow¹⁹,
13 Alana L. Welm^{8,a}, Bryan E. Welm^{8,a}, Michael T. Lewis^{20,a}, Bingliang Fang^{10,a}, Jack A. Roth^{10,a},
14 Funda Meric-Bernstam^{10,a}, Meenhard Herlyn^{11,a}, Michael Davies^{10,a}, Li Ding^{12,a}, Shunqiang Li^{12,a},
15 Ramaswamy Govindan^{12,a}, Claudio Isella^{2,3,b,§}, Jeffrey A. Moscow^{21,a,§}, Livio Trusolino^{2,3,b,§},
16 Annette T. Byrne^{14,b,§}, Jos Jonkers^{5,b,§}, Carol J. Bult^{4,a,§}, Enzo Medico^{2,3,b,§,*}, Jeffrey H.
17 Chuang^{1,a,§,*}, PDXNET consortium & EurOPDX consortium

18

19 ¹ The Jackson Laboratory for Genomic Medicine, Farmington, CT

20 ² Department of Oncology, University of Torino, Candiolo, Italy

21 ³ Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy

22 ⁴ The Jackson Laboratory, Bar Harbor, ME

23 ⁵ Netherlands Cancer Institute, Amsterdam, Netherlands

24 ⁶ College of Medicine, Seoul National University, South Korea

25 ⁷ Frederick National Laboratory for Cancer Research, Frederick, MD

26 ⁸ University of Utah Huntsman Cancer Institute, Salt Lake City, UT

27 ⁹ Department of Human Genetics, University of Utah, Salt Lake City, UT

28 ¹⁰ The University of Texas M.D. Anderson Cancer Center, Houston, TX

29 ¹¹ The Wistar Institute, Philadelphia, PA

30 ¹² Washington University School of Medicine in St. Louis, St. Louis, MO

31 ¹³ Seven Bridges Genomics, Inc., Cambridge, Charlestown, MA

32 ¹⁴ Department of Physiology and Medical Physics, Centre for Systems Medicine, Royal College
33 of Surgeons in Ireland, Dublin, Ireland

34 ¹⁵ Center for Cancer Biology, VIB, Leuven, Belgium

35 ¹⁶ Laboratory of Translational Genetics, Department of Human Genetics, KU Leuven, Leuven,
36 Belgium

37 ¹⁷ Vall d'Hebron Institute of Oncology, Barcelona, Spain

38 ¹⁸ Institut Curie, Paris, France

39 ¹⁹ Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD

40 ²⁰ Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX

41 ²¹ Investigational Drug Branch, National Cancer Institute, Bethesda, MD

42 ^a PDXNET consortium

43 ^b EurOPDX consortium

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

45 [§] These authors jointly supervised this work.

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

47 (jeff.chuang@jax.org).

48

49 **PDXNET Consortium (additional members)**

50 Steven Neuhauser⁴, Javad Noorbakhsh¹, Peter Robinson¹, Tijana Borovski¹³, Vicki Chin¹³, John
51 DiGiovanna¹³, Soner Koc¹³, Tamara Stankovic¹³, Tiffany Wallace²², Lacey E. Dobrolecki²⁰, Argun
52 Akcakanat¹⁰, Kurt Evans¹⁰, Don Gibbons¹⁰, Vanessa Jensen¹⁰, Dara Keener¹⁰, Michael Kim¹⁰,
53 Scott Kopetz¹⁰, Mourad Majidi¹⁰, David Menter¹⁰, Coya Tapia¹⁰, Shannon Westin¹⁰, Timothy Yap¹⁰,
54 Jianhua Zhang¹⁰, Ran Zhang¹⁰, Xiaoshan Zhang¹⁰, Min Jin Ha¹⁰, Huiqin Chen¹⁰, Dali Li¹⁰, Xiaofeng
55 Zheng¹⁰, Erkan Yucan¹⁰, Christopher D Lanier¹⁰, Turcin Saridogan¹⁰, Bryce P Kirby¹⁰, Bingbing
56 Dai¹⁰, Gao Boning²³, John Minna²³, Hyunsil Park²³, Brenda Timmons²³, Luc Girard²³, Dylan
57 Fingerman¹¹, Qin Liu¹¹, Rajasekharan Somasundaram¹¹, Min Xiao¹¹, Mike Tetzlaff¹⁰, Vashisht G.
58 Yennu-Nanda¹⁰, Kate Nathanson²⁴, George Xu²⁵, Rebecca Aft¹², Jessica Andrews¹², Alicia
59 Asaro¹², Song Cao¹², Feng Chen¹², John DiPersio¹², Erin Dreskell¹², Ryan Fields¹², Steven Foltz¹²,
60 Katherine Fuh¹², Kian Lim¹², Jinjin Luo¹², Cynthia Ma¹², Mike McLellan¹², Tina Primeau¹²,
61 Fernanda Rodriguez¹², Dawn Ross¹², Roslynn Sims¹², Brian VanTine¹², Andrea Wang-Gillam¹²,
62 Mike Wendl¹², Cathy Wiggins¹², Yige Wu¹², Matt Wyczalkowski¹², Lijun Yao¹², Daniel Cui Zhou¹²,
63 Chong-xian Pan²⁶, Moon S. Chen, Jr²⁶, Luis Carvajal-Carmona²⁶, Matthew Ellis²⁷, Michael
64 Ittmann²⁷, Susan Hilsenbeck²⁷, Bert O'Malley²⁷, Nicholas Mitsiades²⁷, Salma Kaochar²⁷

65

66 ²² Center to Reduce Cancer Health Disparities, National Cancer Institute, Bethesda, MD

67 ²³ Hamon Center For Therapeutic Oncology, UT Southwestern Medical Center, Dallas

68 ²⁴ Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA

69 ²⁵ Department of Pathology and Laboratory Medicine, Hospital of the University of
70 Pennsylvania, Philadelphia, PA
71 ²⁶ University of California Davis, Sacramento, CA
72 ²⁷ Baylor College of Medicine, Houston, TX.
73

74 **EurOPDX Consortium (additional members)**

75 Simona Corso^{2,3}, Alessandro Fiori^{2,3}, Silvia Giordano^{2,3}, Marieke van de Ven⁵, Daniel S. Peeper⁵,
76 Ian Miller¹⁴, Cristina Bernadó¹⁷, Beatriz Morancho¹⁷, Lorena Ramírez¹⁷, Joaquín Arribas¹⁷, Héctor
77 G. Palmer¹⁷, Alejandro Piris-Gimenez¹⁷, Laura Soucek¹⁷, Ahmed Dahmani¹⁸, Elodie Montaudon¹⁸,
78 Fariba Nemati¹⁸, Virginie Dangles-Marie¹⁸, Didier Decaudin¹⁸, Sergio Roman-Roman¹⁸, Denis G.
79 Alférez²⁸, Katherine Spence²⁸, Robert B. Clarke²⁸, Mohamed Bentires-Alj²⁹, David K.
80 Chang³⁰, Andrew V. Biankin³⁰, Alejandra Bruna³¹, Steven Kupczak³¹, Martin O'Reilly³¹,
81 Magdalena Ridley³¹, Carlos Caldas³¹, Oriol Casanovas³², Eva Gonzalez-Suarez³², Purificación
82 Muñoz³², Alberto Villanueva³², Nathalie Conte³³, Jeremy Mason³³, Ross Thorne³³, Terrence F.
83 Meehan³³, Zdenka Dudova³⁴, Ales Křenek³⁴, Alexandros Sigaras³⁵, Olivier Elemento³⁵, Giorgio
84 Inghirami³⁵, Anna Golebiewska³⁶, Simone P. Niclou³⁶, G. Bea A. Wisman³⁷, Steven de Jong³⁷,
85 Petra Kralova³⁸, Radislav Sedlacek³⁸, Eleonora Leucci³⁹, Luisa Lanfrancone⁴⁰, Pier Giuseppe
86 Pelicci⁴⁰, Gunhild Mari Mælandsmo⁴¹, Jens Henrik Norum⁴¹, Emilie Vinolo⁴²

87
88 ²⁸ Manchester Cancer Research Centre, University of Manchester, UK
89 ²⁹ University of Basel, University Hospital Basel, Switzerland

90 ³⁰ University of Glasgow, Institute of Cancer Sciences, UK

91 ³¹ Cancer Research UK Cambridge Institute, Cambridge Cancer Centre, UK

92 ³² Catalan Institute of Oncology, Bellvitge Biomedical Research Institute, L'Hopistalet de
93 Llobregat, Barcelona, Spain

94 ³³ European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust
95 Genome Campus, Hinxton, Cambridge, UK

96 ³⁴ Institute of Computer Science, Mazaryk University, Brno, Czech Republic

97 ³⁵ Cornell University, Weill Cornell Medical College, New York, NY, United States

98 ³⁶ NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of
99 Health, Luxembourg, Luxembourg

100 ³⁷ University Medical Centre Groningen, Groningen, The Netherlands

101 ³⁸ Czech Center for Phenogenomics, Institute of Molecular Genetics, Vestec, Czech Republic

102 ³⁹ TRACE PDX platform, Katholieke Universiteit Leuven, Leuven, Belgium

103 ⁴⁰ European Institute of Oncology, Milan, Italy

104 ⁴¹ Oslo University Hospital, Oslo, Norway

105 ⁴² seeding science SPRL, Limelette, Belgium

106

107 ABSTRACT

108 Patient-derived xenografts (PDXs) are resected human tumors engrafted into mice for preclinical
109 studies and therapeutic testing. It has been proposed that the mouse host affects tumor evolution
110 during PDX engraftment and propagation, impacting the accuracy of PDX modeling of human
111 cancer. Here we exhaustively analyze copy number alterations (CNAs) in 1451 PDX and matched
112 patient tumor (PT) samples from 509 PDX models. CNA inferences based on DNA sequencing
113 and microarray data displayed substantially higher resolution and dynamic range than gene
114 expression-based inferences, and they also showed strong CNA conservation from PTs through
115 late-passage PDXs. CNA recurrence analysis of 130 colorectal and breast PT/PDX-early/PDX-
116 late trios confirmed high-resolution CNA retention. We observed no significant enrichment of
117 cancer-related genes in PDX-specific CNAs across models. Moreover, CNA differences between
118 patient and PDX tumors were comparable to variations in multi-region samples within patients.
119 Our study demonstrates the lack of systematic copy number evolution driven by the PDX mouse
120 host.

121

122 MAIN

123 A variety of models of human cancer have been used to study basic biological processes and
124 predict responses to treatment. For example, mouse models with genetically engineered
125 mutations in oncogenes and tumor suppressor genes have clarified the genetic and molecular
126 basis of tumor initiation and progression^{1,2}, though responses sometimes differ between human
127 and mouse³. Cell lines have also been widely used to study cancer cells, but they lack the
128 heterogeneity and microenvironment of *in vivo* tumors and have shown limitations for predicting
129 clinical response⁴. Human tumors engrafted into transplant-compliant recipient mice (patient-
130 derived xenografts, PDX) have advantages over prior systems for preclinical drug efficacy studies
131 because they allow researchers to directly study human cells and tissues *in vivo*⁵⁻⁸. Comparisons
132 of genome characteristics and histopathology of primary tumors and xenografts of human breast
133 cancer⁹⁻¹³, ovarian cancer¹⁴, colorectal cancer¹⁵ and lung cancer¹⁶⁻¹⁸, have demonstrated that the
134 biological properties of patient-derived tumors are largely preserved in xenografts. A growing body
135 of literature supports their use in cancer drug discovery and development¹⁹⁻²¹.

136 A caveat to PDX models is that intratumoral evolution can occur during engraftment and
137 passaging^{11,22-25}. Such evolution could potentially modify treatment response of PDXs with
138 respect to the patient tumors^{23,26,27}, particularly if the evolution were to systematically alter cancer-
139 related genes. This issue is related to multi-region comparisons of patient tumors²⁸⁻³¹, for which
140 local mutational and immune infiltration variations have suggested differential phenotypes among
141 multi-region samples³². However, it remains unclear how therapies should be designed with
142 respect to this variation. Comparing patient tumor-PDX evolution to the multi-region variations
143 within the patient tumor would clarify the importance of primary-PDX divergence for treatment.

144 Recently, Ben-David et al.²⁶ reported extensive PDX copy number divergence from the
145 patient tumor of origin and across passages, based mainly on large-scale assessment of CNA
146 profiles inferred from gene expression microarray data, which allowed analysis of aberrations at
147 the scale of chromosomal arms. They raised concerns about genetic evolution in PDXs as a
148 consequence of mouse-specific selective pressures, which could impact the capacity of PDXs for
149 faithful modeling of patient treatment response. Such results contrast with reports that have
150 observed genomic fidelity of PDX models with respect to the originating patient tumors and from
151 early to late passages by direct DNA measurements (DNA sequencing or SNP arrays) in several
152 dozen PDX models^{9,10,33}.

153 Here we resolve these contradicting observations by systematically evaluating CNA
154 changes and the genes they affect during engraftment and passaging in a large, internationally
155 collected set of PDX models, comparing both RNA and DNA-based approaches. The data
156 collected, as part of the U.S. National Cancer Institute (NCI) PDXNet (PDX Development and
157 Trial Centers Research Network) Consortium and EurOPDX consortium, comprises 1548 PT and
158 PDX datasets (1451 unique samples) from 509 models derived from American, European and
159 Asian cancer patients. Our study demonstrates that prior reports of systematic copy number
160 divergence between patient tumors and PDXs are incorrect, and that there is high retention of
161 copy number during PDX engraftment and passaging. This work also finely enumerates the copy
162 number profiles in hundreds of publicly available models, which will enable researchers to assess
163 the suitability of each for individualized treatment studies.

164

165 **RESULTS**

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

167 We have assembled copy number alteration (CNA) profiles of 1451 unique samples (324
168 patient tumor, PT, and 1127 PDX samples) corresponding to 509 PDX models contributed by
169 participating centers of the PDXNET, the EurOPDX consortium, and other published datasets^{9,34}

170 (see METHODS, Supplementary Table 1 and Supplementary Fig. 1). We estimated copy number
171 (CN) from five data types: single nucleotide polymorphism (SNP) array, whole-exome sequencing
172 (WES), low-pass whole-genome sequencing (WGS), RNA sequencing (RNA-Seq) and gene
173 expression array data, yielding 1548 tumor datasets including samples assayed on multiple
174 platforms. Paired-normal DNA and in some cases, paired normal RNA, were also obtained to
175 calibrate WES and RNA-Seq tumor samples. To estimate the CNA profiles for the different data
176 types, we used tools including ASCAT for SNP arrays³⁵, Sequenza for tumor-normal WES³⁶,
177 qDNAseq³⁷ and ASCAT for WGS and e-karyotyping³⁸ for gene expression (RNA-Seq and gene
178 expression array) data (see METHODS). Copy number segments for each sample were filtered
179 for measurement noise, median-centered, and intersected with gene coordinates (see
180 METHODS, Supplementary Data 1).

181 The combined PDX data represent 16 broad tumor types (see METHODS), with 64%
182 (n=324) of the models having their corresponding patient tumors assayed and another 64%
183 (n=328) having multiple PDX samples of either varying passages (ranging from P0 – P21) or
184 varying lineages from propagation into distinct mice (Fig. 1a, Supplementary Table 2). The
185 distributions of PT and PDX samples across different tumor types, passages, and assay platforms
186 (Fig. 1b, Supplementary Fig. 2-12) show the wide spectrum of this combined dataset, which is
187 the most comprehensive copy number profiling of PDXs compiled to date. Additionally, our data
188 include 7 patients with multiple tumors collected either from different relapse time points or
189 different metastatic sites, resulting in multiple PDX models derived from a single patient.

190

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

192 To compare the CNA profiles from different platforms in a controlled fashion, we assembled a
193 benchmarking dataset with matched measurements across multiple platforms (Supplementary
194 Table 3, Supplementary Fig. 13 – 17). Copy number calling has been reported to be noisy for
195 several data types^{39,40}, and we observed that quantitative comparisons between CNA profiles are
196 sensitive to: (1) the thresholds and baselines used to define gains and losses, (2) the dynamic
197 range of copy number values from each platform, and (3) the differential impacts of normal cell
198 contamination for different measurements. To control for such systematic biases, we assessed
199 the similarity between two CNA profiles using the Pearson correlation of their $\log_2(\text{CN ratio})$ values
200 across the genome in 100kb windows. Regions with discrepant copy number were identified as
201 those with outlier values from the linear regression model (see METHODS).

202

203 **CNAs from WES are consistent with CNAs from SNP array data.** While SNP arrays are widely
204 accepted for estimating tumor CNA profiles^{41,42}, CNA estimates from WES data have more
205 uncertainty^{36,43}. We implemented a WES-based CNA pipeline and benchmarked it against SNP
206 array-based estimates for matched samples, which we used as a gold standard. Copy number
207 gain or loss segments (see METHODS) from SNP arrays were of a higher resolution (Fig. 2a;
208 median/mean segment size: 1.49/4.05 Mb for SNP, 4.70/14.6 Mb for WES, $p < 2.2e^{-16}$) and wider
209 dynamic range (Fig. 2b; range of $\log_2(\text{CN ratio})$: -8.62 – 2.84 for SNP, -3.04 – 1.85 for WES, $p <$
210 $2.2e^{-16}$). The difference in range is apparent in the linear regressions between platforms
211 (Supplementary Fig. 19a). These observations take into account the broad factors affecting CNA
212 estimates across platforms, such as the positional distribution of sequencing loci; the sequencing
213 depth of WES (10 – 280X); and the superior removal of normal cell contamination by SNP array
214 CNA analysis workflows using SNP allele frequencies³⁵.

215 Despite the superiority of SNP arrays, we observed strong agreement between SNP
216 arrays and WES, with significantly higher Pearson correlation coefficients on matched samples
217 than samples of different models (range: 0.913 – 0.957 for matched samples, 0.0366 – 0.354 for
218 unmatched samples, $p = 1.02e^{-06}$), with the exception of 2 samples that lacked CNA aberrations
219 (Fig. 2c, Supplementary Fig. 13, 18 and 19a). Regions with discordant copy number between
220 platforms could also be identified (Supplementary Fig. 19a, see METHODS). The discordant copy
221 number regions largely correspond to small focal events (average size 1.53Mb) detectable by
222 SNP arrays but missed by WES (Supplementary Fig. 19b). Still, CNA profiling by WES is reliable
223 in most cases, with 99% of the genome locations across the samples consistent with the values
224 from SNP arrays.

225
226 **Low accuracy for gene expression-derived CNA profiles.** To compare the suitability of gene
227 expression for quantifying evolutionary changes in CNA, we adapted the e-karyotyping method
228 used in Ben-David et al.^{26,38,44} for RNA-Seq and gene expression array data. For each tumor type,
229 the expression values were calibrated relative to either median expression of non-tumor tissue
230 RNA samples, or relative to median expression of tumor samples when normal samples were not
231 available (Supplementary Fig. 15 and 17). Copy number segments calibrated by non-tumor
232 expression were of higher resolution (Fig. 2a; median/mean segment size: 36.0/51.9 Mb for
233 RNASEQ NORM, 48.2/65.3 Mb for RNASEQ TUM, $p < 2.2e^{-16}$; 62.0/72.4 Mb for EXPARR NORM,
234 80.1/85.2 Mb for EXPARR TUM, $p = 2.20e^{-07}$) and wider dynamic range (Fig. 2b; range of $\log_2(\text{CN}$
235 $\text{ratio})$: -2.07 – 2.17 for RNASEQ NORM, -1.79 – 1.81 for RNASEQ TUM, $p < 2.2e^{-16}$; -1.40 –
236 1.89 for EXPARR NORM, -1.13 – 1.59 for EXPARR TUM, $p = 4.09e^{-07}$) compared to segments

237 calculated by calibration with tumor samples. This was true for both RNA-Seq and gene
238 expression array platforms.

239 A notable problem with the expression-based calls is that the alternative expression
240 calibrations can have a major impact on called gains and losses. This is especially apparent for
241 regions frequently called as gains or losses in specific tumor types (Supplementary Fig. 20), e.g.
242 as identified in other studies⁴⁵⁻⁴⁷. Chromosomes 8q and 13 were almost exclusively identified as
243 gains and chromosomes 21 and 22 were almost exclusively as losses in the gastric cancer RNA-
244 Seq dataset when normal samples were used for calibration. Similarly, we called exclusive gains
245 in chromosomes 7q and 20 and losses in chromosomes 4q31-35, 8p,16q and 21 using normal
246 samples for calibration for the hepatocellular carcinoma expression array dataset. However,
247 changing the calibration to use tumor samples resulted in these regions being erroneously called
248 with approximately equal frequencies of gains and losses. These alternate methodologies yielded
249 strong variability in the CNA calls, and this was the case for each of the RNAseq and expression
250 array datasets (Pearson correlation range: 0.218 – 0.943 for RNASEQ NORM vs TUM, 0.377 –
251 0.869 for EXPARR NORM vs TUM, Fig. 2c and Supplementary Fig. 21). For each, this range of
252 correlations was far greater than was observed in comparisons between the DNA-based methods
253 ($p = 9.37e^{-5}$ and $p = 3.28e^{-07}$ relative to SNP vs WES). This indicates the problematic nature of
254 RNA-based CNA calling with calibration by tumor samples, which has been used when normal
255 samples are not available.

256 We observed other measures showing the limitations of RNA-based CNA calling.
257 Expression-based calling had segmental resolution an order of magnitude worse than the DNA-
258 based methods (Fig. 2a and Supplementary Fig. 14 – 17; median/mean segment size: 3.45/14.0
259 Mb for WES, 36.0/51.9 Mb for RNASEQ NORM, $p < 2.2e^{-16}$; 1.73/ 5.18 Mb for SNP, 62.0/72.4 Mb
260 for EXPARR NORM, $p < 2.2e^{-16}$). The range of detectable copy number values was also superior
261 for DNA-based methods (Fig. 2b; range of $\log_2(\text{CN ratio})$: -6.00 – 5.33 for WES, -2.07 – 2.17 for
262 RNASEQ NORM, $p < 2.2e^{-16}$; -9.19 – 4.65 for SNP, -1.40 – 1.89 for EXPARR NORM, $p < 2.2e^{-16}$). In addition, there was a lack of correlation between the expression-based and DNA-based
263 methods (range: 0.0541 – 0.942 for WES vs RNASEQ (NORM); 0.00517 – 0.921 for SNP vs
264 EXPARR (NORM)) (Fig. 2c and Supplementary Fig. 22 and 23). CNA estimates after tumor-based
265 expression normalization resulted in further discordance with DNA-based copy number results
266 (range: -0.182 – 0.929, $p = 0.0468$ for WES vs RNASEQ (TUM); -0.0274 – 0.847, $p = 2.20e^{-06}$
267 for SNP vs EXPARR (TUM)). Many focal copy number events detected by DNA-based methods,
268 as well as some larger segments, were missed by the expression-based methods (Supplementary

270 Fig. 24). Representative examples illustrating the superior resolution and accuracy from DNA-
271 based estimates are given in Fig. 2d (see also Supplementary Fig. 19a and 25).

272

273 **Concordance of PDXs with patient tumors and during passaging**

274 We tracked the similarity of CNA profiles during tumor engraftment and passaging by calculating
275 the Pearson correlation of gene-level copy-number for samples measured on the same platform
276 (see METHODS, Supplementary Fig. 26-64). All pairs of samples derived from the same PDX
277 model were compared – yielding 501 PT-PDX and 1257 PDX-PDX pairs.

278 For all DNA-based platforms we observed strong concordance between matched PT-PDX
279 and PDX-PDX pairs, significantly higher than between different models from the same tumor type
280 and the same center ($p < 2.2e^{-16}$) (Fig. 3a – c, correlation heatmaps in Supplementary Fig. 27 –
281 63). We observed no significant difference in the correlation values between PT-PDX and PDX-
282 PDX pairs for SNP array data (median correlation PT-PDX = 0.950, PDX-PDX = 0.964; $p > 0.05$),
283 though there were small but statistically significant shifts for WES (PT-PDX = 0.874, PDX-PDX =
284 0.936; $p = 2.31e^{-16}$) and WGS data (PT-PDX = 0.914, PDX-PDX = 0.931; $p = 0.000299$). PT
285 samples have a smaller CNA range than their derived PDXs (median ratio PT/PDX / PDX/PDX:
286 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
287 WGS; Supplementary Fig. 64b and 65), which can be attributed to stromal DNA in PT samples
288 “diluting” the CNA signal. In PDXs, the human stromal DNA is reduced^{9,15}. The minimal effect for
289 SNP array data confirm this interpretation – human stromal DNA contributions to CNA estimates
290 can be removed from SNP arrays based on allele frequencies of germline heterozygous sites,
291 while such contributions to WES and WGS have higher uncertainties.

292 We also performed intra-model comparisons using RNA-based approaches, but the
293 Pearson correlations between pairs of samples did not clearly reproduce the Pearson correlations
294 from DNA-based platforms for those same sample pairs (Supplementary Fig. 66a). To clarify this,
295 we considered just the highly-correlated cases (>0.8 for SNU-JAX Gastric cancer WES, >0.9 for
296 SIBS HCC SNP). We observed that the correlation values for the corresponding RNA-based
297 methods were lower and had higher variance ($p < 0.05$, Supplementary Fig. 66b). In particular,
298 the tumor-median normalization for expression data resulted in significant differences from DNA-
299 based methods.

300

301 **Late PDX passages maintain CNA profiles similar to early passages.** Next, we asked if there
302 is any systematic evolution of copy number during engraftment and passaging. Mouse
303 environment-driven evolution, if present, should reduce CN correlations relative to early samples,

304 such as the primary tumor or first engraftment (P0). However, we observed no apparent effect
305 during passaging on the SNP, WES, or WGS platforms (Fig. 3d – f, Supplementary Fig. 67). For
306 example, the SNP data showed no significant difference between passages (Fig. 3d and
307 Supplementary Fig. 67a). For those models having very late passages (14 breast cancer models,
308 P18 to P21), there was a small but statistically significant correlation decrease compared to
309 models with earlier passages ($p < 8.98e^{-05}$, Supplementary Fig. 68), indicating some copy number
310 changes can occur over long-term passaging (Supplementary Fig. 38). However even at these
311 late passages, the correlations to early passages remained high (median = 0.896). In any given
312 comparison, only a small proportion of the genes were affected by copy number changes (median:
313 2.72%, range: 1.03% – 11.9%). Genes that are deleted and subsequently gained in the later
314 passages (top left quadrant of regression plots, Supplementary Fig. 69) suggest selection of
315 preexisting minor clones as the key mechanism in these regions. For WES and WGS data, more
316 variability in the correlations can be observed (Fig. 3e and f, Supplementary Fig. 66b and c), likely
317 due to a few samples having more stromal contamination or low aberration levels (Supplementary
318 Fig. 64b and 65). However, the lack of downward trend over passaging was also apparent in
319 these sets.

320

321 **PDX copy number profiles trace lineages.** We next compared the similarity of engrafted PDXs
322 of the same model with the same passage number (i.e. all P0s, all P1s, all P2s, etc.). Surprisingly,
323 we discovered that these fragments were not more similar than PDXs from different passage
324 numbers (Fig. 3d – e and Supplementary Fig. 66b, IQR of correlation coefficient for same-
325 passages/different-passages: 0.0700/0.0619 for SNP and 0.103/0.0979 for WES). To further this
326 analysis, we defined, for JAX SNP array and PDMR WES datasets, samples within a lineage as
327 those differing only by consecutive serial passages, while we defined lineages as split when a
328 tumor was divided and propagated into multiple mice (Fig. 3g). For the EurOPDX CRC and BRCA
329 WGS datasets, such lineage splitting was due only to cases with initial engraftment of different
330 fragments of the PT, i.e., PDX samples of different passages were considered as different
331 lineages if they originate from different PT fragments. We observed lower correlation between
332 PDX samples from different lineages compared to within a lineage (Fig. 3h, $p = 0.0233$ for SNP,
333 $p = 0.00119$ for WES, $p = 0.000232$ for WGS), despite a majority of these pairwise comparisons
334 exhibiting high correlation (>0.9). A few examples of models exhibiting large drift between
335 lineages include TM01500 (Supplementary Fig. 29); 416634, 558786 and 665939
336 (Supplementary Fig. 50); 135848 and 762968 (Supplementary Fig. 51); 245127 and 959717

337 (Supplementary Fig. 52); 287954 and 594176 (Supplementary Fig. 56); 174316 and 695221
338 (Supplementary Fig. 57).

339 We next asked if the phylogenetic distance between samples could explain the observed
340 shifts in the correlations. These distance relationships are clearest for the CRC and BRCA WGS
341 sets because these models have only one lineage split occurring at the engraftment stage. We
342 compared correlation as a function of phylogenetic distance within a lineage, which in this
343 phylogeny is simply equal to the difference in passage number between the two samples.
344 Increase in passage difference did not consistently reduce the correlation between samples
345 (Supplementary Fig. 70). This suggests that lineage-splitting is often responsible for deviations in
346 CNAs between samples, and that copy number evolution during passaging mainly arises from
347 evolved spatial heterogeneity²⁷.

348

349 **Genes with copy number alterations acquired during engraftment and passaging show no
350 preference for cancer or treatment-related functions.**

351 Next, we investigated which genes tend to undergo copy number changes. Genes with changes
352 during engraftment or during passaging were identified based on a residual threshold with respect
353 to the improved linear regression⁴⁸ (see METHODS, Supplementary Fig. 26). A low copy number
354 change threshold ($|\log_2(\text{CN ratio}) \text{ change}| > 0.5$) was selected to include genes with subclonal
355 alterations. To test for functional biases, we compared CNA-altered genes to gene sets with
356 known cancer- and treatment-related functions, notably genes in TCGA oncogenic signaling
357 pathways⁴⁹; genes with copy number and expression changes associated with therapeutic
358 sensitivity, resistance or changes in drug response from the JAX Clinical Knowledgebase^{50,51}; and
359 genes with frequent amplifications or deletions in the Cancer Gene Census⁵² (Cosmic version
360 89). We calculated the proportion of altered genes for sample pairs from each model across all
361 platforms and tumor types. In agreement with the high maintenance of CNA profiles described
362 above, we found the proportion of altered protein-coding genes to be low (median/IQR: 1.90%/
363 4.11% PT-PDX, 1.25%/ 3.60% PDX-PDX pairs, Fig. 4a). Only 8.78% of PT-PDX pairs and 4.53%
364 PDX-PDX pairs showed >10% of their protein-coding genes altered. We observed no significant
365 increase ($p < 0.1$) in alterations among any of the cancer gene sets compared to the background
366 of all protein-coding genes, for either the PT-PDX or PDX-PDX comparisons. This provides
367 evidence that there is no systematic selection for CNAs in oncogenic or treatment-related
368 pathways during engraftment or passaging. We next considered tumor-type-specific effects,
369 focusing on types with larger numbers of models to ensure statistical power (breast cancer,
370 colorectal cancer, lung adenocarcinoma and lung squamous cell carcinoma). Genomic

371 Identification of Significant Targets in Cancer (GISTIC)^{53,54} analysis of TCGA tumors has
372 previously identified significantly altered genomic driver regions which can be used to differentiate
373 tumor types and subtypes⁵⁵⁻⁵⁸. We observed no significant increase in alterations in tumor-type-
374 specific GISTIC gene sets compared to the background ($p < 0.1$) for either PT-PDX or PDX-PDX
375 comparisons (Fig. 4b).

376

377 **Low recurrence of altered genes across models.** We tested if any particular genes often
378 recurred in CNAs across models. Using a stringent CNA threshold ($|\log_2(\text{CN ratio}) \text{ change}| > 1.0$
379 with respect to linear regression model) to distinguish genes with possible functional impact (see
380 METHODS), we observed a very low recurrent frequency (Fig. 4c), with only 12 and 2 genes
381 recurring at $> 5\%$ frequency for PT-PDX and PDX-PDX comparisons, respectively
382 (Supplementary Table 4). No gene had a recurrence frequency higher than 8.96%. We observed
383 that all these recurrent genes overlapped models in which one sample displayed an unusually
384 large gain or loss ($|\log_2 (\text{CN ratio})| > 1.5$). This suggests that these regions may be subject to
385 more noise in the CNA estimation procedure at these loci (Supplementary Fig. 71). None of these
386 recurrent genes overlapped cancer- or treatment-related gene sets, nor did they intersect genes
387 ($n=3$) reported by Ben-David et al.²⁶ to have mouse-induced copy number changes associated
388 with drug response in the CCLE^{59,60} database. We further queried from CCLE data whether any
389 of these recurrent genes showed evidence for copy number-related drug response (see
390 METHODS, Supplementary Table 5). For the 6 genes with sufficient data available, we found no
391 association between copy number and drug response mediated by gene expression ($q\text{-value} <$
392 1).

393

394 **Absence of CNA shifts in 130 WGS patient tumour, early passage PDX and late passage
395 PDX trios**

396 We next investigated whether recurrent CNA changes occur in PDXs in a tumor-type specific
397 fashion. To this aim, we analysed further the WGS-based CNA profiles of large metastatic
398 colorectal (CRC) and breast cancer (BRCA) series (see METHODS), respectively composed of
399 87 and 43 matched trios of patient tumour (PT), PDX at early passage (PDX-early) and PDX at
400 later passage (PDX-late). We carried out GISTIC analysis to identify recurrent CNAs by evaluating
401 the frequency and amplitude of observed events^{53,54}. GISTIC was applied separately for each PT,
402 PDX-early (P0-P1 for CRC, P0-P2 for BRCA) and PDX-late (P2-P7 for CRC, P3-P9 for BRCA)
403 cohorts of CRC and BRCA (Supplementary Table 6). As expected, CRCs and BRCAAs generated
404 different patterns of significant CNAs, with each similar to the GISTIC patterns in their respective

405 TCGA series (Supplementary Fig. 72). However, within each tumour type GISTIC profiles of the
406 PT, PDX-early, and PDX-late cohorts were virtually indistinguishable (Fig 5a and Supplementary
407 Fig. 72), demonstrating no gross genomic alteration systematically acquired or lost in PDXs.

408 To clarify these behaviors, we carried out gene-level analysis, where each gene was
409 attributed the GISTIC score (G-score) of the respective segment (Supplementary Table 7). In both
410 the CRC and BRCA cohorts, gene-level G-scores of the PTs were highly correlated with the
411 respective PDX-early and PDX-late cohorts (Fig. 5b and c). Moreover, PT versus PDX
412 correlations were comparable to PDX-early versus PDX-late correlations. To search for
413 progressive shifts, we compared the change in G-score (ΔG): (i) from tumor to PDX-early and (ii)
414 from PDX-early to PDX-late. Correlations in these two ΔG values, as shown in the bottom-right
415 panels of Fig. 5b and c, was absent or even slightly negative. These results confirmed the
416 absence of systematic CNA shifts in PDXs even under high resolution, gene-level analysis.

417

418 ***Lack of CNA-based functional shifts in trios confirms the absence of mouse-specific***
419 ***evolution.*** We then considered the possibility of systematic copy number evolution at the pathway
420 level in these triplets. To evaluate this, we performed Gene Set Enrichment Analysis (GSEA)^{61,62}
421 using G-scores to rank genes in each cohort (See METHODS). Consistent with the known
422 recurrence of cancer CNAs at driver genes, multiple gene sets displayed significant enrichment
423 in individual cohorts. To avoid spurious apparent enrichment for sets of genes with adjacent
424 chromosomal location, we implemented an additional filter based on G-score significance (see
425 METHODS and Supplementary Table 8). After applying the Normalized Enrichment Score (NES),
426 FDR q-value and G-score filters, 49 gene sets were found to be significant in at least one of the
427 three CRC cohorts, and 89 gene sets in at least one of the three BRCA cohorts (Supplementary
428 Table 9). Importantly, control gene sets composed of GISTIC hits identified in TCGA CRC and
429 BRCA datasets were all significant, confirming that the WGS cohorts used here correctly
430 recapitulate the major CNA features of these two cancer types.

431 However, differences associated with PDX engraftment and passage were negligible. For
432 both CRC and BRCA, the NES profiles for the ~8000 gene sets of PTs were highly correlated
433 with the respective PDX-early and PDX-late cohorts (Fig. 5d and e). Moreover, PT versus PDX
434 correlations were comparable to PDX-early versus PDX-late correlations. To search for
435 progressive shifts, we calculated for each significant gene set Δ NES values between PT and
436 PDX-early, as well as between early and late PDX. Similarly to what was observed for the ΔG -
437 scores, as shown in the bottom-right panels of Fig. 5d and e, correlations were absent or at most
438 slightly negative, confirming the absence of systematic CNA-based functional shifts in PDXs.

439
440 **CNA evolution across PDXs is no greater than variation in patient multi-region samples**
441 As a reference for the treatment relevance of PDX-specific evolution, we compared to levels of
442 copy number variation in multi-region samples of patient tumors. For this we used copy number
443 data from multi-region sampling of non-small-cell lung cancer (92 patient tumors, 295 multi-region
444 samples) from the TRACERx Consortium³¹, performing analogous CNA correlation and gene
445 analyses ($|residual| > 0.5$) between multi-region pairs (Supplementary Fig. 73). We observed no
446 significant differences in correlation ($p > 0.05$) between patient multi-region and lung cancer PT-
447 PDX pairs, while PDX-PDX pairs in fact showed significantly better correlation than the multi-
448 region pairs ($p < 0.05$, Fig. 6a). These findings were consistent when tumors were grouped as
449 adenocarcinomas, squamous cell carcinomas, or others. Cancer gene set analyses confirmed
450 these results, with multi-region samples showing greater differences than either PT-PDX or PDX-
451 PDX comparisons, across all the cancer gene sets considered ($p < 0.05$, Fig. 6b and
452 Supplementary Fig. 74). These results show that PDX-associated CNA evolution is no greater
453 than what patients experience naturally within their tumors. Our PDX collection also contains a
454 few cases in which the patient tumor was assayed at multiple time points (relapse/metastasis) or
455 multiple metastatic sites, allowing for controlled comparison of intra-patient variation versus PDX
456 evolution (Supplementary Fig. 3, 4 and 7). We observed no significant difference between the
457 CNA evolution during engraftment and passaging compared to the intra-patient samples (Fig. 6c).
458 CNA profiles for these samples are shown visually in Fig. 6d.

459
460 **DISCUSSION**
461 Here we have investigated the evolutionary stability of patient-derived xenografts, an important
462 model system for which there have been prior reports of mouse-induced copy number evolution.
463 To better address this, we assembled the largest collection of CNA profiles of PDX models
464 reported to date, comprising over 1500 datasets from PDX samples of multiple passages and
465 their originating patient tumors from more than 500 PDX models across a variety of tumor types.
466 Our analysis demonstrated the reliability of copy number estimation by DNA-based
467 measurements over RNA-based inferences, which are substantially inferior in terms of resolution
468 and accuracy. The importance of DNA measurements is supported by the inconsistent
469 conclusions by two independent studies on the same PDX expression array datasets by Gao et
470 al.¹⁹ Ben-David et al.^{26,63} concluded that drastic copy number changes, driven by mouse-specific
471 selection, often occur within a few passages. On the other hand, Mer et al.⁶⁴ reported high

472 similarity between passages of the same PDX model based on direct correlations of gene
473 expression, consistent with our findings in large, independent DNA-based datasets.

474 To understand this, we note that the CN shifts inferred by Ben-David et al. are inherently
475 impacted by major technical issues. First, the microarray signal for PT samples is diluted by
476 introgressed human stromal cells, while in PDXs mouse stromal transcripts hybridize only to a
477 fraction of the human probes⁶⁵. As a consequence, PT samples with substantial stromal content
478 would display a reduced signal compared to the corresponding PDX, which can lead to an
479 erroneous inference of systematic increase in aberrations during PDX engraftment. Second, the
480 mouse host microenvironment can affect the transcriptional profile of the PDX tumor⁶⁶ and the
481 quantity of mouse stroma can vary across passages. This can result in variability in the expression
482 signal which can be wrongly inferred as CN changes, both from the tumor itself and through cross
483 hybridization of mouse RNA to the human microarray. Although improved concordance in
484 expression between PT and PDX can be achieved with RNA sequencing with the removal of
485 mouse reads^{67,68}, we observed that expression-based copy number inferences still have low
486 resolution and robustness. Hence, many cancer-driving genes, which are found mainly in focal
487 events with a size of 3Mb or lower⁶⁹⁻⁷², cannot be evaluated for PDX-specific alterations. These
488 issues are further worsened by the lack of tissue-matched normal gene expression profiles for
489 calibration³⁸, which have been only intermittently available but can substantially impact copy
490 number inferences. Because of these considerations, the question of how much PDXs evolve as
491 a consequence of mouse-specific selective pressures cannot be adequately addressed by
492 expression data.

493 The studies we have presented here take into account the above issues by use of DNA
494 data, as well as by assessing copy number changes by pairwise correlation/residual analysis to
495 control for systematic biases, and they overall confirm the high retention of CNA profiles from
496 PDX engraftment to passaging. We do observe larger deviations between PT-PDX than in PDX-
497 PDX comparisons, though this is likely due to dilution of PT signal by human stromal cells.
498 Interestingly, we found that a major contributor to the differences between PDX samples is
499 lineage-specific drift associated with splitting of tumors into fragments during PDX propagation.
500 This spatial evolution within tumors appears to affect sample comparisons more than time or the
501 number of passages.

502 A challenge for evaluating any model system is that there is no clear threshold for genomic
503 change that determines whether the model will still reflect patient response. Genetic variation
504 among multi-region samples within a patient can shed light on this point, since the goal of a
505 successful treatment would be to eradicate all of the multiple regions of the tumor. We found that

506 the copy number differences between PT and PDX are no greater than the variations among
507 multi-region tumor samples or intra-patient samples. Thus concerns about the genetic stability of
508 the PDX system are likely to be less important than the spatial heterogeneity of solid tumors
509 themselves. This result is consistent with our results on lineage effects during passaging, which
510 indicate that intratumoral spatial evolution is the major reason for genetic drift.

511 We observed no evidence for systematic mouse environment-induced selection for cancer
512 or treatment-related genes via copy number changes, though individual cases vary (see example
513 in Supplementary Fig. 75). Moreover, only a small fraction of sample pairs (2.44%, 43 out of 1758)
514 shows large CNA discordance (see METHODS), suggesting that clonal selection out of a complex
515 population is rare. These results indicate that the variations observed in PDXs are mainly due to
516 spontaneous intratumoral evolution rather than murine pressures. The extreme cases (see
517 Supplementary Fig. 76 for examples with same lineage) may be informative for future studies of
518 the evolutionary process, especially through consideration of repeated spatial sampling. It may
519 be informative to compare such examples to those reported by Eirew et al.²², who described a
520 variety of clonal selection dynamics during engraftment and passaging for breast cancer PDXs,
521 as well as by Ding et al.¹¹, who demonstrated the possibility of cellular selection during xenograft
522 formation similar to that during metastasis. While such cases are uncommon in our study, further
523 subclonal analysis may be useful for clarifying potential selection pressures.

524 In summary, our in-depth tracking of CNAs throughout PDX engraftment and passaging
525 confirms that tumors engrafted and passaged in PDX models maintain a high degree of molecular
526 fidelity to the original patient tumors and their suitability for pre-clinical drug testing. Overall, we
527 find that PDX are highly concordant with the originating patient tumor and stable through multiple
528 passaging, and that differences are no greater than those observed spatially within patient solid
529 tumors. At the same time, our study does not rule out that PDXs will evolve in individual
530 trajectories over time, and for therapeutic dosing studies, the best practice is to confirm the
531 existence of expected molecular targets and obtain sequence characterizations in the cohorts
532 used for testing as close to the time of the treatment study as is practicable.

533

534 **ACKNOWLEDGEMENTS**

535 Support for the PDXNET consortium included funding provided by the NIH to the PDXNet Data
536 Commons and Coordination Center (NCI U24-CA224067), to the PDX Development and Trial
537 Centers (NCI U54-CA224083, NCI U54-CA224070, NCI U54-CA224065, NCI U54-CA224076,
538 NCI U54-CA233223, and NCI U54-CA233306) and to the National Cancer Institute Cancer
539 Genomics Cloud (HHSN261201400008C and HHSN261201500003I). The Jackson Laboratory

540 (JAX) PDX resource data were supported by the National Cancer Institute of the National
541 Institutes of Health under the JAX Cancer Center NCI Grant (Award Number P30CA034196). The
542 content is solely the responsibility of the authors and does not necessarily represent the official
543 views of the National Institutes of Health. The genomic data for JAX PDX tumors used in this work
544 were generated by JAX Genome Technologies and Single Cell Biology Scientific Service. The
545 development of PDX models and generation of data from Seoul National University, in
546 collaboration with The Jackson Laboratory, was supported by the Korean Healthcare Technology
547 R&D project through the Korean Health Industry Development Institute, funded by the Ministry of
548 Health & Welfare, Republic of Korea (grant number: HI13C2148). Sample procurement and next
549 generation sequencing at Huntsman Cancer Institute was performed at the Genomics and
550 Bioinformatics Analysis and Biorepository and Molecular Pathology Shared Resources,
551 respectively, supported by NCI P30CA042014. SNP arrays were performed at the University of
552 Utah Health Sciences Center Genomics Core. We are grateful to Michael P. Klein for assistance
553 with SNP array data. M.H.B. is funded by the National Institutes of Health under Ruth L.
554 Kirschstein National Research Service Award T32HG008962 from the National Human Genome
555 Research Institute. M.T.L. is supported by a P30 Cancer Center Support Grant CA125123 and a
556 Core Facility Support Grant from the Cancer Research and Prevention Initiative of Texas
557 RP170691. PDX generation and whole exome sequencing at the University of Texas MD
558 Anderson Cancer Center were supported by the University of Texas MD Anderson Cancer Center
559 Moon Shots Program, Specialized Program of Research Excellence (SPORE) grant CA070907.
560 The development of PDX models and generation of data from Wistar Institute was supported by
561 National Cancer Institute, National Institutes of Health (NCI R50-CA211199). PDMR data has
562 been funded in whole or in part with federal funds from the National Cancer Institute, National
563 Institutes of Health (Contract Number HHSN261200800001E). The content of this publication
564 does not necessarily reflect the views or policies of the Department of Health and Human
565 Services, nor does mention of trade names, commercial products, or organizations imply
566 endorsement by the U.S. Government. The breast cancer PDX models from Washington
567 University used for this study were developed in part through the support from The Breast Cancer
568 Research Foundation and Fashion Footwear Charitable Foundation of New York, Inc.. The
569 pancreatic cancer PDX models from Washington University used in this study were developed
570 with the support of NCI grants P50 CA196510, P30 CA091842 and The Foundation for Barnes-
571 Jewish Hospital's Cancer Frontier Fund through the Siteman Cancer Center Investment Program.
572 The data for these models was provided by U54 CA224083. Support for the EurOPDX consortium
573 included funding provided by Fondazione AIRC under 5 per Mille 2018 - ID. 21091 program

574 (E.M., A.B., L.T.), AIRC Investigator Grants 18532 (L.T.) and 20697 (A.B.), AIRC/CRUK/FC
575 AECC Accelerator Award 22795 (L.T.), EU H2020 Research and Innovation Programme, grant
576 agreement no. 731105 "EDIReX" (E.M., A.B., L.T., A.T.B., V.S., J.J.), Fondazione Piemontese
577 per la Ricerca sul Cancro-ONLUS 5 per mille Ministero della Salute 2015 (E.M., L.T.), 2014
578 and 2016 (L.T.), My First AIRC Grant (MFAG) 19047 (C.I.), EU H2020 Research and Innovation
579 Programme, Grant Agreement No. 754923 "COLOSSUS" (A.T.B., D.L., L.T.), European Research
580 Council Consolidator Grant 724748 – BEAT (A.B.), Science Foundation Ireland under grant
581 13/CDA/2183 "COLOFORETELL" (A.T.B.), Irish Health Research Board grant ILP-POR-2019-066
582 (A.T.B.), ISCIII - Miguel Servet program CP14/00228 and GHD-Pink/FERO Foundation grant
583 (V.S.), Netherlands Organization for Scientific Research (NWO) Vici grant 91814643 (J.J.),
584 European Research Council (ERC) Synergy project CombatCancer (J.J.), Oncode Institute (J.J.,
585 R.d.B.) and Dutch Cancer Society (J.J., R.d.B.) , NCI grant U24 CA204781 (J.H.C., T.F.M.). The
586 authors thank the Breast Cancer Group from VHO for providing study materials. We thank Debbie
587 M. Krupke from The Jackson Laboratory for assistance with organizing the tumor type information.
588

589 **AUTHOR CONTRIBUTIONS**

590 X.Y.W., C.J.B., J.J., A.T.B., L.T, J.A.M., C.I., E.M. and J.H.C. conceived and jointly supervised
591 the study. X.Y.W. organized the study, collected and structured the data, and designed and
592 carried out the analyses. J.G. collected and organized the EurOPDX data and carried out the
593 analyses. X.Y.W., E.M. and J.H.C. wrote the manuscript. J.G, C.I., Z.-M.Z., A.S., and M.W.L.
594 contributed to the refinement of the manuscript. A.S. and M.W.L developed the workflows. A.S.,
595 Z.-M.Z, M.W.L, and Y.-S.S assisted with the computational analyses. R.J., C.F., J.R., D.A.D, J.R.
596 and B.D. assisted with the workflow development and data collection and organization on the
597 Cancer Genomics Cloud. R.E.B. and R.d.B. contributed to sample selection and processing of
598 EurOPDX data. C.J.B., R.P., L.C., Y.A.E., J.H.D., S.S., M.B., C.-H.Y., E.C.-S., A.L.W, B.E.W.,
599 M.T.L., Y.X., J.W., B.F., J.R., F.M.-B., J.W., A.V.K., V.R., M.H., H.S., R.J.M., S.D., L.D., F.G.,
600 A.B., L.T., A.L., A.C.O., A.T.B., E.M., D.L., R.d.B., P.t.B., J.J., V.S., E. Marangoni, H.K., J.-I.K.,
601 H.-K.Y., C.L., E.M. and J.H.C. contributed the sequencing and array data. C.J.B., E.M. and J.H.C.
602 directed the project. The named author list describes the primary contributors of data and analysis
603 to the project, though these studies were supported by consortium-wide activities. All members
604 of the PDXNet and EurOPDX Consortia participated in group discussions or supportive analyses
605 regarding the study design, data standards, sample collection, or data analysis approaches.
606

607 **COMPETING FINANCIAL INTERESTS**

608 A.L.W and B.E.W receive a portion of royalties if University of Utah licenses certain PDX models
609 to for-profit entities. M.T.L is a founder of, and equity stake holder in, Tvardi Therapeutics Inc., a
610 founder of, and limited partner in, StemMed Ltd., and a Manager in StemMed Holdings LLC. He
611 also receives a portion of royalties if Baylor College of Medicine licenses certain PDX models to
612 for-profit entities. F.M.-B. reports receiving commercial research grants from Novartis,
613 AstraZeneca, Calithera, Aileron, Bayer, Jounce, CytoMx, eFFECTOR, Zymeworks, PUMA
614 Biotechnology, Curis, Millennium, Daiichi Sankyo, Abbvie, Guardant Health, Takeda, Seattle
615 Genetics, and GlaxoSmithKline as well as grants and travel related fees from Taiho, Genentech,
616 Debiopharm Group, and Pfizer. She also served as a consultant to Pieris, Dialectica, Sumitomo
617 Dainippon, Samsung Bioepis, Aduro, OrigIMed, Xencor, The Jackson Laboratory, Zymeworks,
618 Kolon Life Science, and Parexel International, and advisor to Inflection Biosciences, GRAIL,
619 Darwin Health, Spectrum, Mersana, and Seattle Genetics. L.T. reports receiving research grants
620 from Symphogen, Servier, Pfizer, and Merus, and he is in the speakers' bureau of Eli Lilly,
621 AstraZeneca, and Merck KGaA. J.J. reports receiving funding for collaborative research from
622 Artios Pharma. He also serves as SAB member of Artios Pharma. The other authors declare no
623 competing financial interests.

624

625 FIGURE LEGENDS

626 **Figure 1. PDXNet and EurOPDX patient derived xenograft datasets used for copy number
627 profiling across 16 tumor types.** (a) Numbers of PDX models for each tumor type, with models
628 also having multiple PDX samples or having matched patient tumor samples specified. (b)
629 Distributions of datasets by passage number and assay platform for patient tumors and PDX
630 samples, separated by tumor type. “Late” passages include P18, P19 and P21 samples.

631

632 **Figure 2. Comparisons of resolution and accuracy for copy number alterations estimated
633 by DNA-based and expression-based methods.** (a) Pairwise comparisons of distributions of
634 segment size (Mb) of CNAs estimated by different measurement platforms in the benchmarking
635 dataset (see Supplementary Table 3). CNAs are regions with $(|\log_2(\text{CN ratio})| \geq 0.1)$. P-values
636 indicate significance of difference between distributions by Wilcoxon rank sum test. (b) Pairwise
637 comparisons of distributions of $\log_2(\text{CN ratio})$ of CNA segments. P-values were computed by
638 Kolmogorov-Smirnov test. (c) Distributions of Pearson correlation coefficient of median-centered
639 $\log_2(\text{CN ratio})$ in 100-kb windows from CNA segments between pairs of samples estimated by
640 different platforms (see Supplementary Table 3). Samples with non-aberrant profiles in SNP array
641 and WES data are omitted (Range (5-95 percentile) of $\log_2(\text{CN ratio}) < 0.3$). P-values indicate

642 Wilcoxon rank sum test. **(d)** Examples of CNA profiles in comparisons of different platforms.
643 Pearson correlation coefficients of CNA segments between pairs of samples are shown on the
644 right. (SNP: SNP array, WES: whole-exome sequencing, RNASEQ: RNA sequencing, EXPARR:
645 gene expression array, NORM: normalization by median expression of normal samples, TUM:
646 normalization by median expression of tumor samples)

647

648 **Figure 3. Comparisons of copy number alterations from patient tumor to early and late PDX**
649 **passages. (a-c)** Distributions of Pearson correlation coefficient of gene-based copy number,
650 estimated by **(a)** SNP array, **(b)** WES, and **(c)** WGS, between: PT-PDX samples from the same
651 model; PDX-PDX samples of the same model; and samples of different models from a common
652 tumor type and contributing center. P-values were computed by Wilcoxon rank sum test (ns: not
653 significant p-value > 0.05). **(d-f)** Distributions of Pearson correlation coefficients of gene-based
654 copy number, estimated by **(d)** SNP array, **(e)** WES, and **(f)** WGS, among patient tumor and PDX
655 passages of the same model. Comparisons relative to PT and P0 are shown (higher passages
656 are shown in Supplementary Fig. 66). **(g)** Schematic of lineage splitting during passaging and
657 expansion of tumors into multiple mice. This is a simplified illustration for passaging procedures
658 in which different fragments of a tumor are implanted into different mice. **(h)** Pearson correlation
659 distributions for PDX sample pairs of different lineages and sample pairs within the same lineage:
660 for JAX SNP array, PDMR WES, and EuroPDX WGS datasets. P-values were computed by
661 Wilcoxon rank sum test. The numbers in the parentheses represent the number of pairwise
662 correlations.

663

664 **Figure 4. Cancer pathway analysis for copy number altered genes during engraftment and**
665 **passaging. (a)** Distribution of proportion of altered genes for pairwise comparisons of PDX
666 samples for various gene sets: Protein-coding genes annotated by Ensembl; Oncogenic signaling
667 pathways identified by TCGA⁴⁹; JAX CKB^{50,51} Amp indicates genes with copy number gain or
668 over-expression associated with therapeutic sensitivity or resistance; JAX CKB Del indicates
669 genes with copy number loss or under-expression associated with therapeutic sensitivity or
670 resistance; Census Amp Del indicates genes with frequent amplifications or deletions in the
671 Cancer Gene Census⁵². CNA genes were identified by $|\text{residual}| > 0.5$ from linear regression
672 model. P-values were computed by Wilcoxon rank sum test (ns: not significant, p > 0.1). **(b)**
673 Distribution of proportion of altered genes for pairwise comparisons within breast cancer,
674 colorectal cancer, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC)
675 models. Prevalence of alterations in significantly amplified (TCGA Gistic Amp) or deleted (TCGA

676 Gistic Del) genes for the corresponding tumor type are shown. P-values were computed by
677 Wilcoxon rank sum test (ns: not significant, $p > 0.1$). The numbers in the parentheses in the
678 horizontal axis represent the number of genes, and those in the plot title represent the number of
679 pairwise correlations. (c) Recurrence frequency of protein coding genes with copy number
680 alterations, $|\text{residual}| > 1$, across all models in PT-PDX and PDX-PDX comparisons.

681

682 **Figure 5. Absence of mouse-driven recurrent CNAs during engraftment and propagation**
683 **of colorectal and breast cancer PDXs.** (a) Bar charts representing genome-wide GISTIC G-
684 score for amplifications (red) and deletions (blue) in each of the three cohorts (PT, PDX-early,
685 PDX-late) for CRC and BRCA. (b-c) Scatter plots comparing gene-level GISTIC G-score between
686 each of the three cohorts, for (b) CRC and (c) BRCA. Bottom-right panels of (b) and (c): scatter
687 plots comparing ΔG -scores from PT to PDX-early and from PDX early to PDX-late. (d-e) Scatter
688 plots comparing GSEA Normalized Enrichment Score (NES) for gene sets between each of the
689 three cohorts, for (d) CRC (e) and BRCA. Grey dots: all gene sets; red dots: gene sets significantly
690 enriched in at least one among the PT, PDX-early, PDX-late cohorts. Bottom-right panels of (d)
691 and (e): scatter plots comparing Δ NES from PT to PDX-early and from PDX-early to PDX-late.

692

693 **Figure 6. Comparison of CNA variation during PDX engraftment and passaging to CNA**
694 **variation among patient multi-region, tumor relapse, and metastasis samples.** (a) Distributions of Pearson correlation coefficients of gene-based copy number for lung
695 adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and other lung cancer
696 subtypes. Columns compare: multi-region tumor samples from TRACERx³¹; PT-PDX samples
697 from the same model; and PDX-PDX samples from the same model. P-values indicate Wilcoxon
698 rank sum test (ns: p -value > 0.05). (b) Distributions of proportion of altered genes between multi-
700 region tumor pairs from TRACERx, and PT-PDX and PDX-PDX pairs for various gene sets for
701 LUAD and LUSC. Gene sets are the same as in Fig. 4. TCGA Gistic and JAX CKB gene sets are
702 shown (other gene sets are shown in Supplementary Fig. 76). (c) Distributions of Pearson
703 correlation coefficients of gene-based copy number between intra-patient PT
704 (primary/relapse/metastasis) pairs from the same patient and corresponding PT-PDX/PDX-PDX
705 (derived from the same model; a different PT sample from the same patient generates a different
706 model) pairs from the same set of patients. P-values were computed by Wilcoxon rank sum test
707 (ns: p -value > 0.05). (d) CNA profiles of PT and PDX samples from patients with multiple PDX
708 models from multiple PT collection (primary/relapse/metastasis).

709

710 **REFERENCES**

- 711 1. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer
712 therapeutics. *Disease models & mechanisms* **1**, 78-82 (2008).
- 713 2. Walrath, J.C., Hawes, J.J., Van Dyke, T. & Reilly, K.M. Genetically engineered mouse
714 models in cancer research. *Advances in cancer research* **106**, 113-164 (2010).
- 715 3. Dennis, C. Off by a whisker. *Nature* **442**, 739-741 (2006).
- 716 4. Hait, W.N. Anticancer drug development: the grand challenges. *Nature Reviews Drug
717 Discovery* **9**, 253 (2010).
- 718 5. Shultz, L.D., Ishikawa, F. & Greiner, D.L. Humanized mice in translational biomedical
719 research. *Nature Reviews Immunology* **7**, 118 (2007).
- 720 6. Brehm, M.A., Shultz, L.D. & Greiner, D.L. Humanized mouse models to study human
721 diseases. *Current opinion in endocrinology, diabetes, and obesity* **17**, 120-125 (2010).
- 722 7. Hidalgo, M. *et al.* Patient-Derived Xenograft Models: An Emerging Platform for
723 Translational Cancer Research. *Cancer Discovery* **4**, 998 (2014).
- 724 8. Byrne, A.T. *et al.* Interrogating open issues in cancer precision medicine with patient-
725 derived xenografts. *Nature Reviews Cancer* **17**, 254 (2017).
- 726 9. DeRose, Y.S. *et al.* Tumor grafts derived from women with breast cancer authentically
727 reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Medicine* **17**,
728 1514 (2011).
- 729 10. Bruna, A. *et al.* A Biobank of Breast Cancer Explants with Preserved Intra-tumor
730 Heterogeneity to Screen Anticancer Compounds. *Cell* **167**, 260-274.e22 (2016).
- 731 11. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft.
732 *Nature* **464**, 999 (2010).
- 733 12. Landis, M.D., Lehmann, B.D., Pietenpol, J.A. & Chang, J.C. Patient-derived breast tumor
734 xenografts facilitating personalized cancer therapy. *Breast Cancer Research* **15**, 201
735 (2013).
- 736 13. Reyal, F. *et al.* Molecular profiling of patient-derived breast cancer xenografts. *Breast
737 Cancer Research* **14**, R11 (2012).
- 738 14. Bankert, R.B. *et al.* Humanized Mouse Model of Ovarian Cancer Recapitulates Patient
739 Solid Tumor Progression, Ascites Formation, and Metastasis. *PLOS ONE* **6**, e24420 (2011).
- 740 15. Julien, S. *et al.* Characterization of a Large Panel of Patient-Derived Tumor Xenografts
741 Representing the Clinical Heterogeneity of Human Colorectal Cancer. *Clinical Cancer
742 Research* **18**, 5314-5328 (2012).
- 743 16. Reddy, S., Piccione, D., Takita, H. & Bankert, R.B. Human Lung Tumor Growth Established
744 in the Lung and Subcutaneous Tissue of Mice with Severe Combined Immunodeficiency.
745 *Cancer Research* **47**, 2456-2460 (1987).
- 746 17. Simpson-Abelson, M.R. *et al.* Long-Term Engraftment and Expansion of Tumor-Derived
747 Memory T Cells Following the Implantation of Non-Disrupted Pieces of Human Lung
748 Tumor into NOD-scid IL2R γ ^{null} Mice. *The Journal of Immunology* **180**, 7009-
749 7018 (2008).
- 750 18. Sugiyama, Y. *et al.* Human Inflammatory Cells Within the Tumor Microenvironment of
751 Lung Tumor Xenografts Mediate Tumor Growth Suppression in Situ that Depends on and
752 Is Augmented by Interleukin-12. *Journal of Immunotherapy* **24**, 37-45 (2001).

- 753 19. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to
754 predict clinical trial drug response. *Nature Medicine* **21**, 1318 (2015).
- 755 20. Hidalgo, M. *et al.* A Pilot Clinical Study of Treatment Guided by Personalized Tumorgrafts
756 in Patients with Advanced Cancer. *Molecular Cancer Therapeutics* **10**, 1311-1316 (2011).
- 757 21. Tentler, J.J. *et al.* Patient-derived tumour xenografts as models for oncology drug
758 development. *Nature reviews. Clinical oncology* **9**, 338-350 (2012).
- 759 22. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-
760 cell resolution. *Nature* **518**, 422 (2014).
- 761 23. Cho, S.-Y. *et al.* Unstable genome and transcriptome dynamics during tumor metastasis
762 contribute to therapeutic heterogeneity in colorectal cancers. *Clinical Cancer Research*,
763 clincanres.3460.2018 (2019).
- 764 24. Giessler, K.M. *et al.* Genetic subclone architecture of tumor clone-initiating cells in
765 colorectal cancer. *The Journal of Experimental Medicine* **214**, 2073 (2017).
- 766 25. Sato, K. *et al.* Multiregion Genomic Analysis of Serially Transplanted Patient-derived
767 Xenograft Tumors. *Cancer Genomics - Proteomics* **16**, 21-27 (2019).
- 768 26. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution.
769 *Nature Genetics* **49**, 1567 (2017).
- 770 27. Kim, H. *et al.* High-resolution deconstruction of evolution induced by chemotherapy
771 treatments in breast cancer xenografts. *Scientific Reports* **8**, 17937 (2018).
- 772 28. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines
773 lung cancer evolution. *Science* **346**, 251-256 (2014).
- 774 29. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas
775 defined by multiregion sequencing. *Nature Genetics* **46**, 225 (2014).
- 776 30. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by
777 Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892 (2012).
- 778 31. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *New*
779 *England Journal of Medicine* **376**, 2109-2121 (2017).
- 780 32. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature*
781 **567**, 479-485 (2019).
- 782 33. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic
783 characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-30 (2013).
- 784 34. He, S. *et al.* PDXliver: a database of liver cancer patient derived xenograft mouse models.
785 *BMC Cancer* **18**, 550 (2018).
- 786 35. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*
787 **107**, 16910-5 (2010).
- 788 36. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor
789 sequencing data. *Ann Oncol* **26**, 64-70 (2015).
- 790 37. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by
791 shallow whole-genome sequencing with identification and exclusion of problematic
792 regions in the genome assembly. *Genome research* **24**, 2022-2032 (2014).
- 793 38. Ben-David, U., Mayshar, Y. & Benvenisty, N. Virtual karyotyping of pluripotent stem cells
794 on the basis of their global gene expression profiles. *Nature Protocols* **8**, 989 (2013).
- 795 39. Zare, F., Hosny, A. & Nabavi, S. Noise cancellation using total variation for copy number
796 variation detection. *BMC Bioinformatics* **19**, 361 (2018).

- 797 40. Wineinger, N.E. & Tiwari, H.K. The Impact of Errors in Copy Number Variation Detection
798 Algorithms on Association Results. *PLOS ONE* **7**, e32396 (2012).
- 799 41. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**,
800 1134-40 (2013).
- 801 42. Taylor, A.M. *et al.* Genomic and Functional Approaches to Understanding Cancer
802 Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).
- 803 43. Talevich, E., Shain, A.H., Botton, T. & Bastian, B.C. CNVkit: Genome-Wide Copy Number
804 Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**,
805 e1004873 (2016).
- 806 44. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse
807 models reveals driver-specific routes to tumorigenesis. *Nature Communications* **7**, 12160
808 (2016).
- 809 45. Schumacher, S.E. *et al.* Somatic copy number alterations in gastric adenocarcinomas
810 among Asian and Western patients. *PLOS ONE* **12**, e0176045 (2017).
- 811 46. The Cancer Genome Atlas Research, N. *et al.* Comprehensive molecular characterization
812 of gastric adenocarcinoma. *Nature* **513**, 202 (2014).
- 813 47. Ally, A. *et al.* Comprehensive and Integrative Genomic Characterization of Hepatocellular
814 Carcinoma. *Cell* **169**, 1327-1341.e23 (2017).
- 815 48. Motulsky, H.J. & Brown, R.E. Detecting outliers when fitting data with nonlinear
816 regression – a new method based on robust nonlinear regression and the false discovery
817 rate. *BMC Bioinformatics* **7**, 123 (2006).
- 818 49. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*
819 **173**, 321-337.e10 (2018).
- 820 50. Patterson, S.E. *et al.* The clinical trial landscape in oncology and connectivity of somatic
821 mutational profiles to targeted therapies. *Human Genomics* **10**, 4 (2016).
- 822 51. Patterson, S.E., Statz, C.M., Yin, T. & Mockus, S.M. Utility of the JAX Clinical
823 Knowledgebase in capture and assessment of complex genomic cancer data. *npj Precision
824 Oncology* **3**, 2 (2019).
- 825 52. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across
826 all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
- 827 53. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer:
828 Methodology and application to glioma. *Proceedings of the National Academy of Sciences*
829 **104**, 20007 (2007).
- 830 54. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets
831 of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**, R41
832 (2011).
- 833 55. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular portraits of human breast
834 tumours. *Nature* **490**, 61 (2012).
- 835 56. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular characterization of human
836 colon and rectal cancer. *Nature* **487**, 330 (2012).
- 837 57. The Cancer Genome Atlas Research, N. *et al.* Comprehensive molecular profiling of lung
838 adenocarcinoma. *Nature* **511**, 543 (2014).
- 839 58. The Cancer Genome Atlas Research, N. *et al.* Comprehensive genomic characterization of
840 squamous cell lung cancers. *Nature* **489**, 519 (2012).

- 841 59. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
842 anticancer drug sensitivity. *Nature* **483**, 603 (2012).
- 843 60. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia.
844 *Nature* **569**, 503-508 (2019).
- 845 61. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
846 interpreting genome-wide expression profiles. *Proceedings of the National Academy of
847 Sciences* **102**, 15545 (2005).
- 848 62. Mootha, V.K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are
849 coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273 (2003).
- 850 63. Ben-David, U., Beroukhim, R. & Golub, T.R. Genomic evolution of cancer models: perils
851 and opportunities. *Nature Reviews Cancer* **19**, 97-109 (2019).
- 852 64. Mer, A.S. *et al.* Integrative pharmacogenomics analysis of patient-derived xenografts.
853 *Cancer Research*, canres.0349.2019 (2019).
- 854 65. Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nature
855 Genetics* **47**, 312 (2015).
- 856 66. Park, E.S. *et al.* Cross-species hybridization of microarrays for studying tumor
857 transcriptome of brain metastasis. *Proceedings of the National Academy of Sciences* **108**,
858 17456 (2011).
- 859 67. Liu, Y. *et al.* Gene expression differences between matched pairs of ovarian cancer patient
860 tumors and patient-derived xenografts. *Scientific Reports* **9**, 6314 (2019).
- 861 68. Isella, C. *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel
862 clinically relevant subtypes of colorectal cancer. *Nature Communications* **8**, 15107 (2017).
- 863 69. Leary, R.J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and
864 sequence alterations in breast and colorectal cancers. *Proceedings of the National
865 Academy of Sciences* **105**, 16224 (2008).
- 866 70. Bierkens, M. *et al.* Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and
867 tumor suppressor in cervical carcinogenesis. *Genes, Chromosomes and Cancer* **52**, 56-68
868 (2013).
- 869 71. Krijgsman, O., Carvalho, B., Meijer, G.A., Steenbergen, R.D.M. & Ylstra, B. Focal
870 chromosomal copy number aberrations in cancer—Needles in a genome haystack.
871 *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1843**, 2698-2704 (2014).
- 872 72. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**,
873 893 (2010).
- 874
- 875

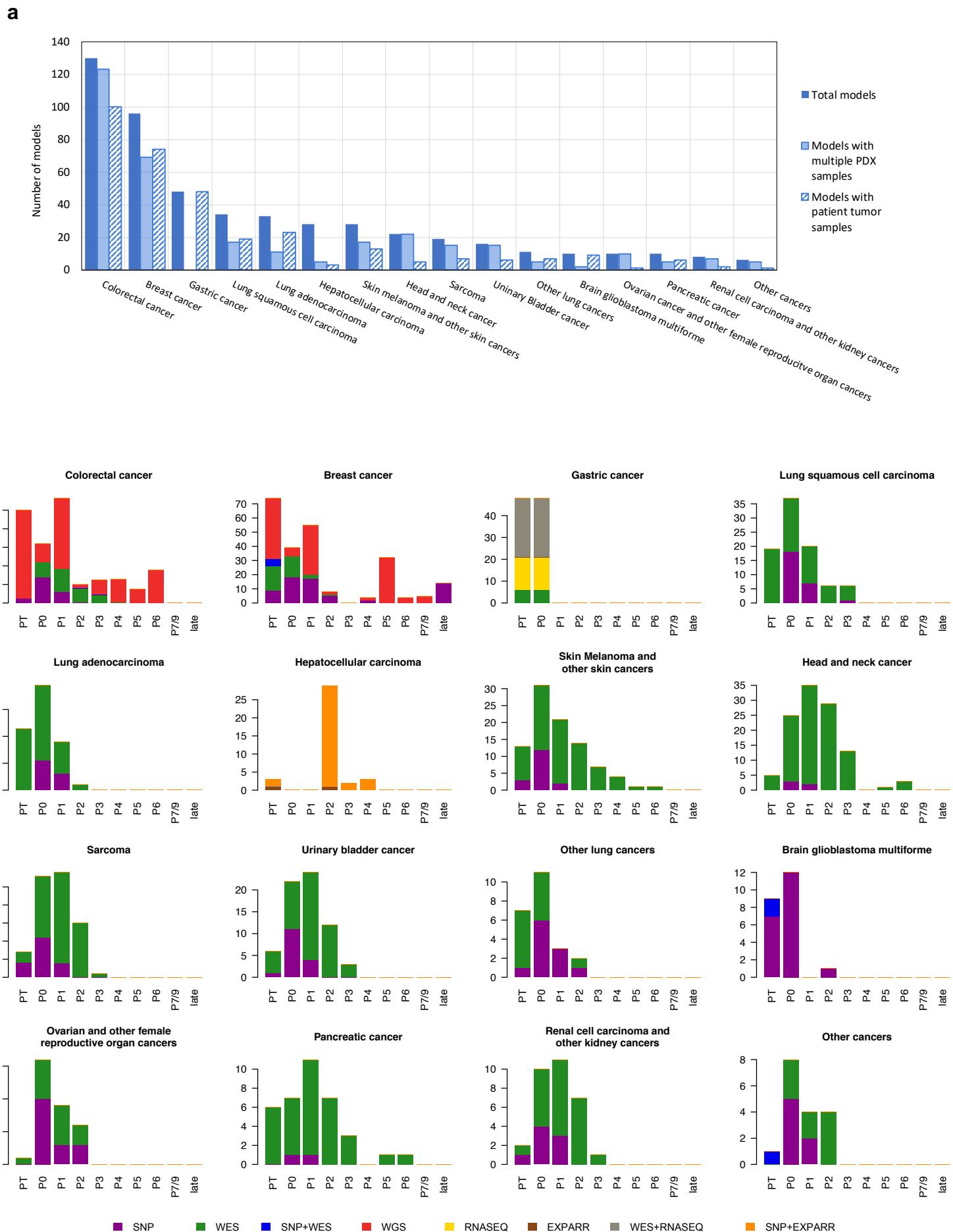
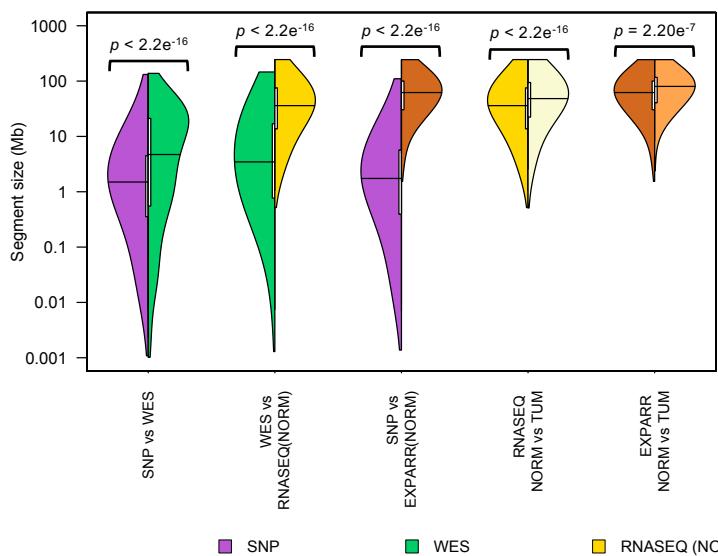
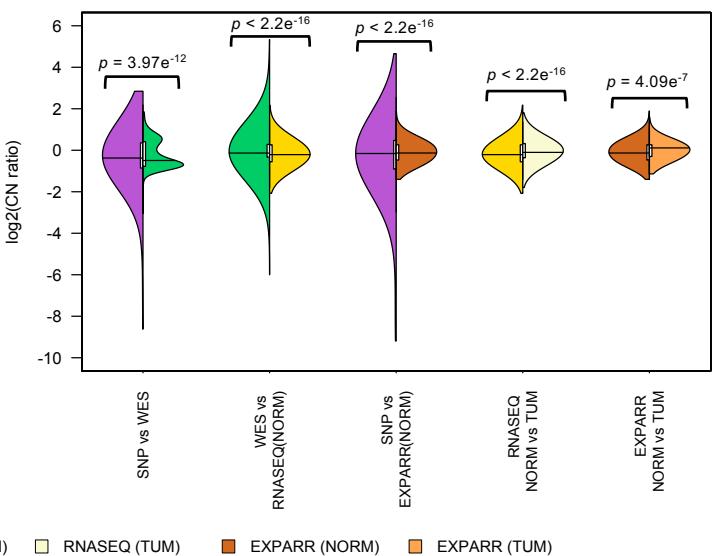


Figure 1.

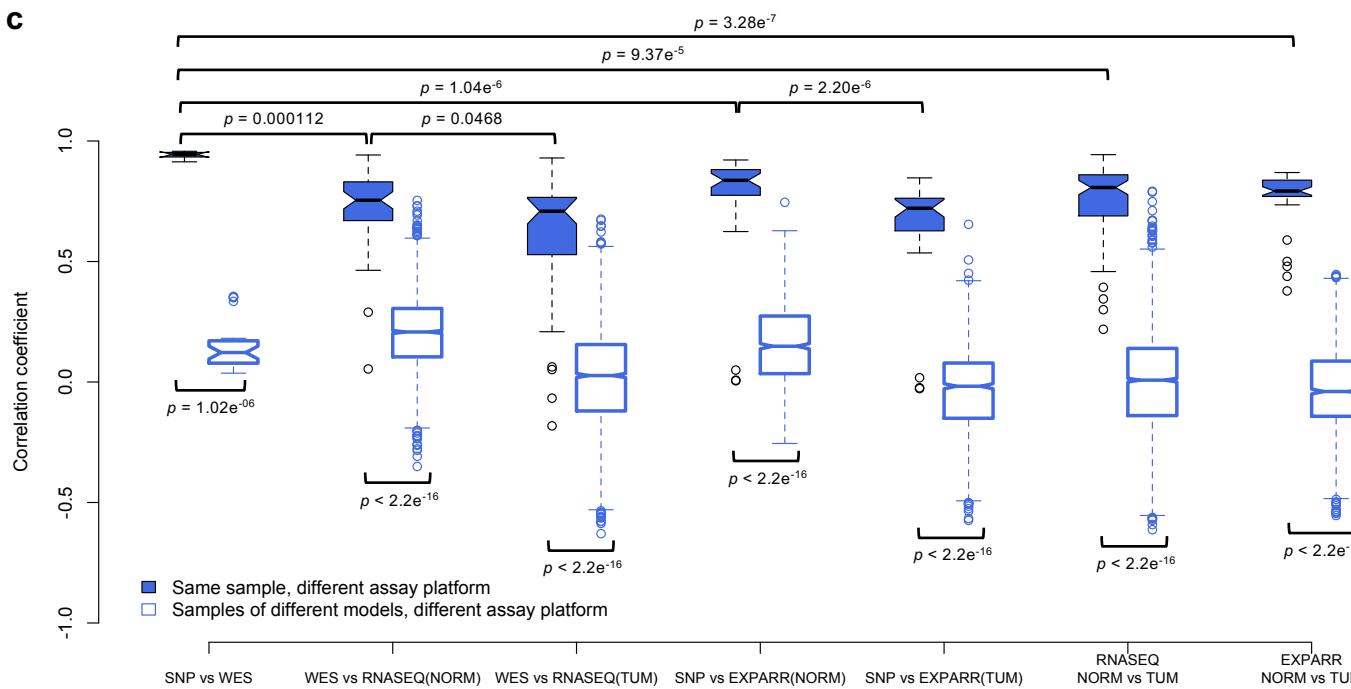
a



b



c



d

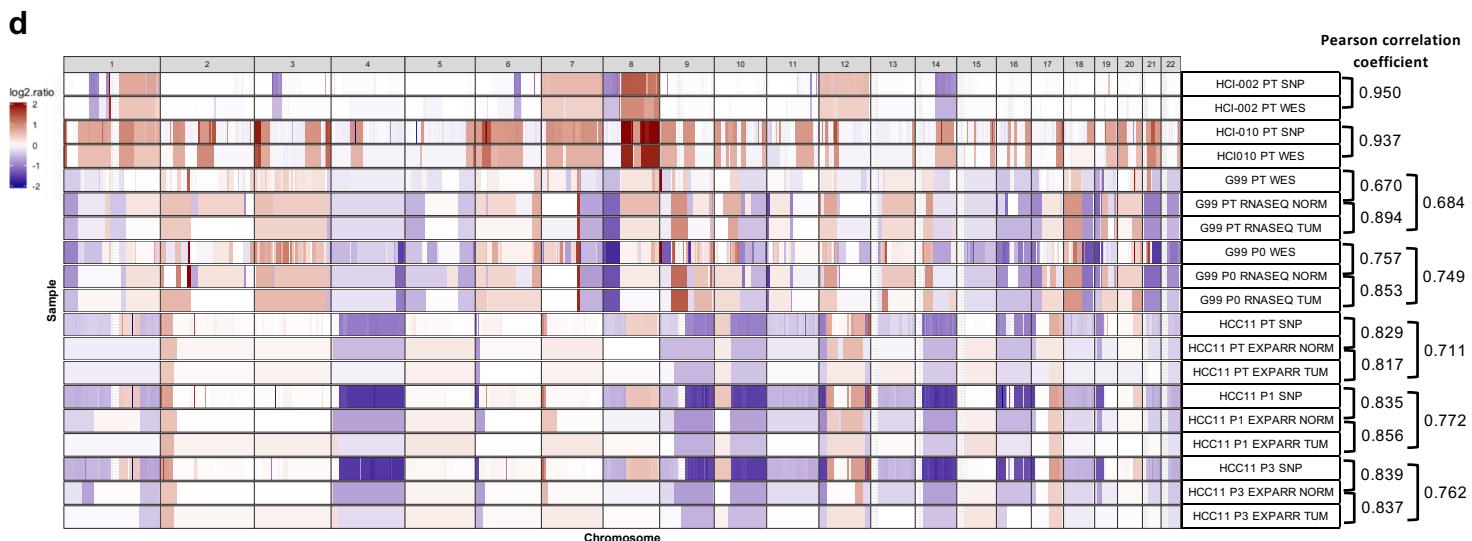


Figure 2.

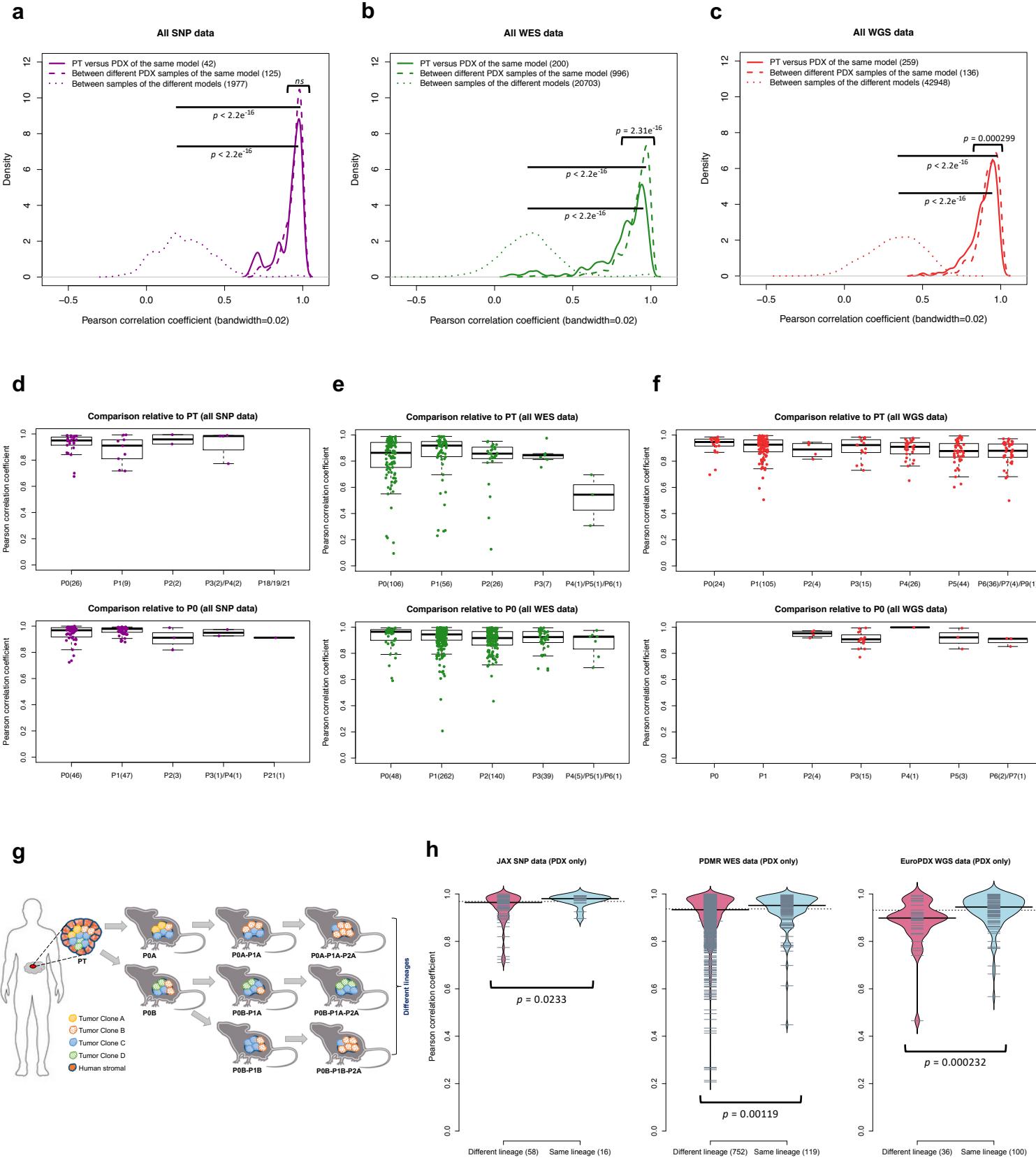


Figure 3.

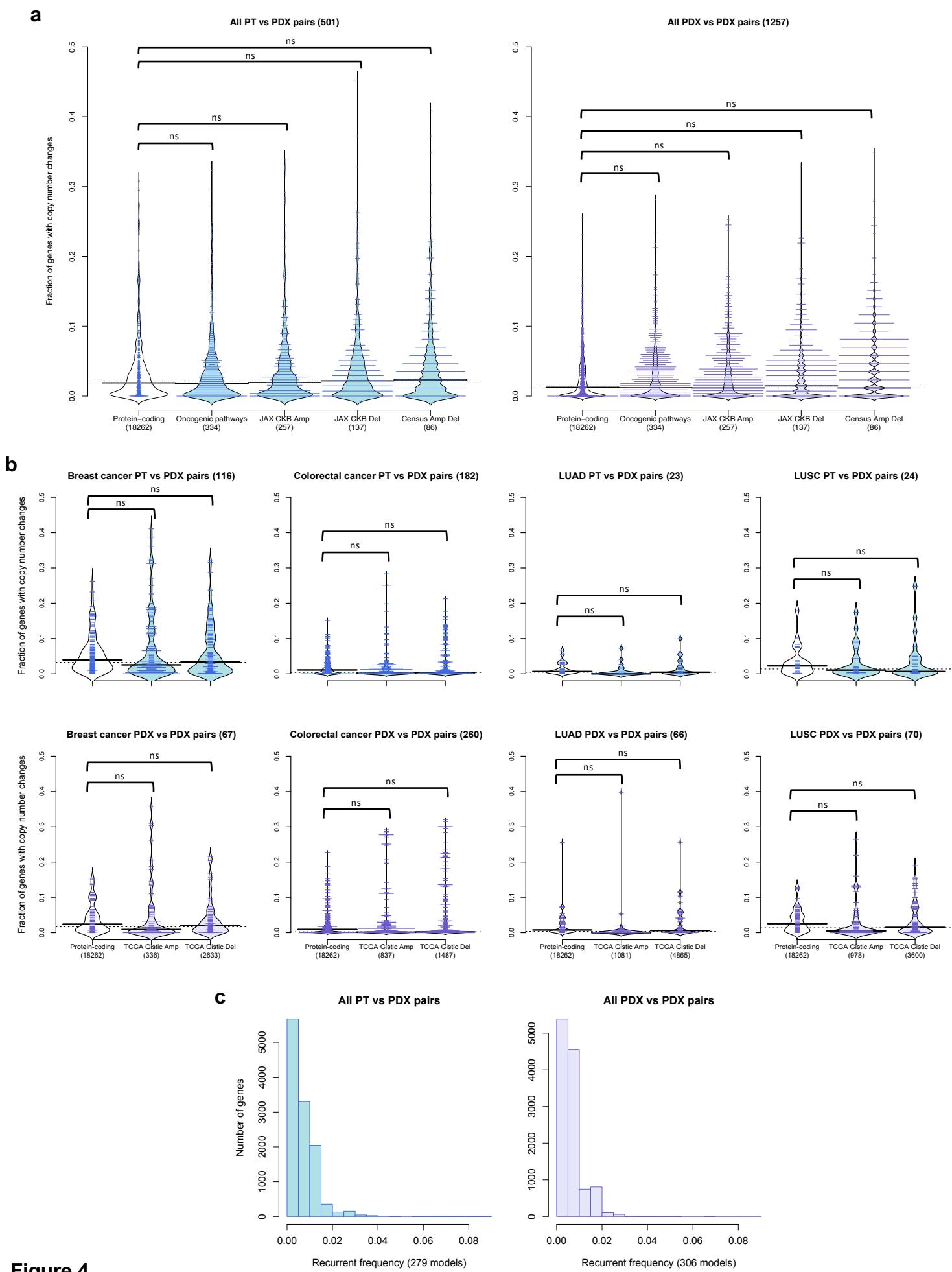


Figure 4.

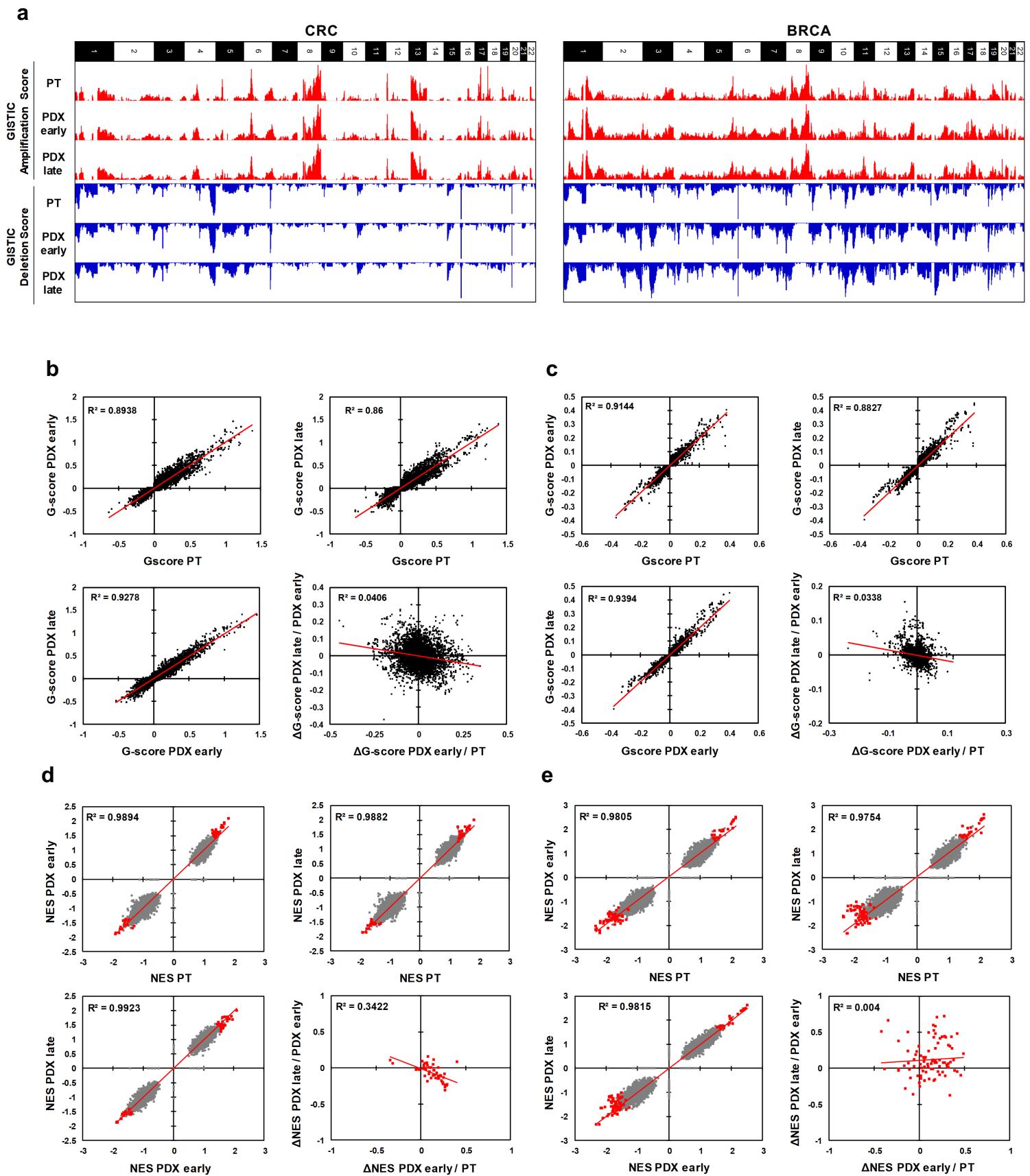


Figure 5.

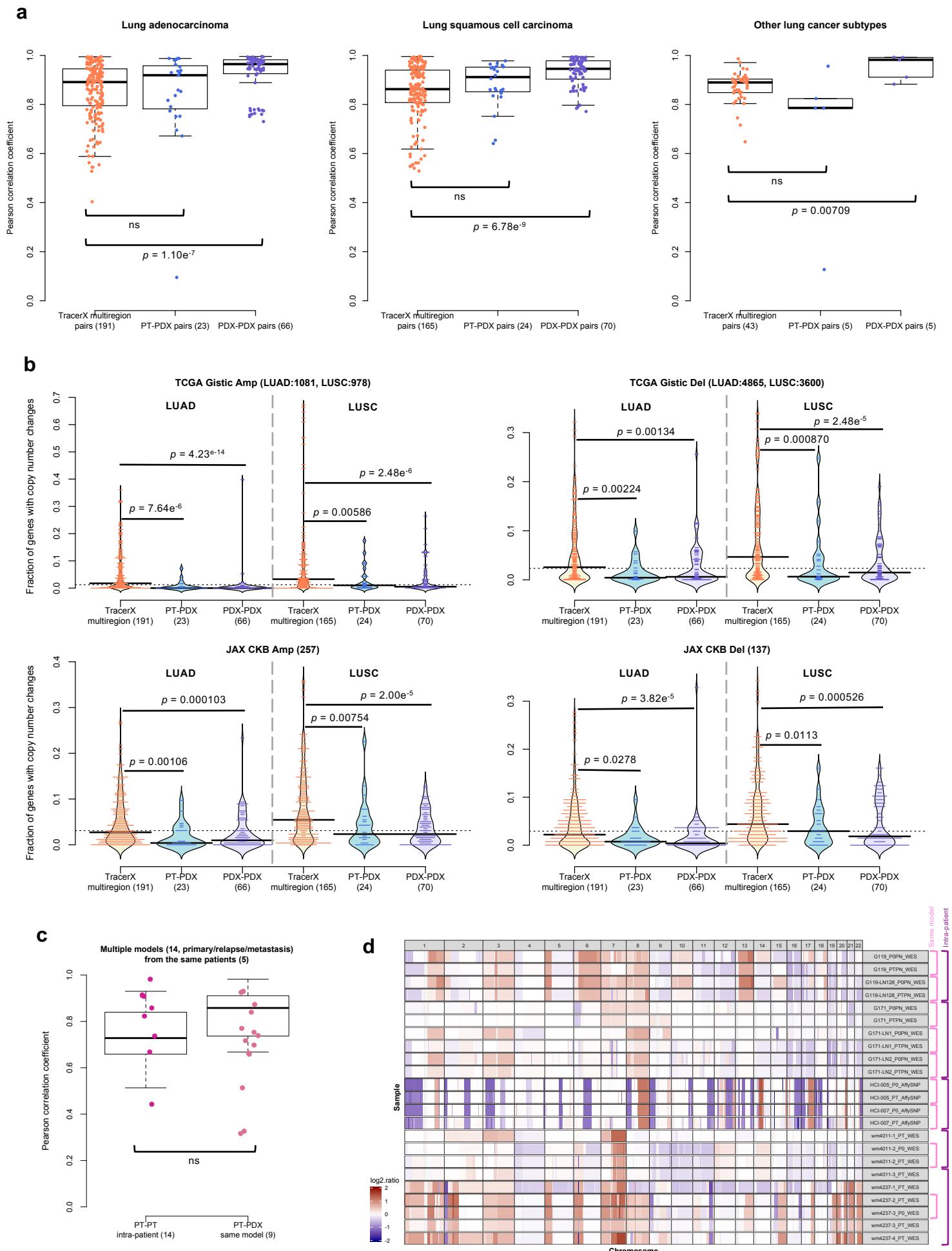


Figure 6.

1 **ONLINE METHODS**

2

3 **Experimental details for sample collection, PDX engraftment and passaging, and array or**
4 **sequencing.**

5 The tumor types and patient tumor (PT) and patient derived xenograft (PDX) samples contributed
6 by various centers are summarized in Supplementary Fig. 1-12 and Supplementary Table 1. The
7 sample collection, PDX engraftment and passaging, and array and sequencing methodologies by
8 the various centers are described below.

9 ***The Jackson Laboratory (JAX).*** Patient tumor engraftment and PDX passaging of various tumor
10 types were performed as previously described¹⁻³. Detailed information of the PDX models can be
11 found in the PDX model search form in Mouse Tumor Biology Database (MTB,
12 <http://tumor.informatics.jax.org/mtbwi/pdxSearch.do>). SNP array samples were genotyped with
13 the Affymetrix Genome-Wide Human SNP Array 6.0 as described in Woo et al³. Whole-exome
14 sequencing were processed as follows: DNA was isolated from tumor and blood samples using
15 the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's protocols.
16 DNA quality was assessed using an E-Gel General Purpose Agarose Gel, 0.8% (Invitrogen) and
17 Nanodrop 2000 spectrophotometer (Thermo Scientific). DNA concentration was determined using
18 a Qubit dsDNA BR Assay Kit (Thermo Scientific). Libraries were prepared by the Genome
19 Technologies core facility at The Jackson Laboratory using SureSelectXT Reagents and
20 SureSelectXT Human All Exon V4 Target Enrichment System (Agilent Technologies), according
21 to the manufacturer's instructions. Briefly, the protocol entails shearing the DNA using the Covaris
22 E220 Focused-ultrasonicator (Covaris), ligating Illumina specific adapters, and PCR amplification.
23 Amplified DNA libraries are then hybridized to the Human All Exon probes, amplified using
24 indexed primers, and checked for quality and concentration using the DNA High-Sensitivity
25 LabChip assay (Agilent Technologies) and quantitative PCR (KAPA Biosystems), according to
26 the manufacturers' instructions. Libraries were sequenced on a HiSeq 2500 100bp paired-end
27 flow cell using TruSeq Rapid SBS reagents (Illumina). Average coverage for normal samples was
28 154.38x (115.13 min – 212.31 max), and was 232.10x for tumor samples (161.48 min – 280.65
29 max).

30 ***Seoul National University-Jackson Laboratory (SNU-JAX).*** Gastric cancer tissues, paired
31 normal gastric tissues, and blood samples were obtained from individuals who underwent
32 gastrectomies at the Hospital of Seoul National University from 2014 to 2016. All samples were
33 obtained with informed consent at the Hospital of Seoul National University, and the institutional
34 review board approved the study per the Declaration of Helsinki. These samples were stored into

35 RPMI media with 1% penicillin/streptomycin immediately after resected from patients and shipped
36 using specimen ice box to the laboratory within half an hour. Gastric cancer samples were divided
37 into several small pieces (2mm × 2mm) and used to generate PDX models and for genomic
38 analysis. Mice were cared for according to institutional guidelines of the Institutional Animal Care
39 and Use Committee of the Seoul National University (no. 14-0016-C0A0). For PDX models,
40 surgically resected tissues were minced into pieces approximately ~2 mm in size and injected
41 into the subcutaneous area in the flanks of 6-week-old NOD/SCID/IL-2γ-receptor null female mice
42 (NSG™ mice, Jackson Laboratory, Bar Harbor, ME). The volume of tumors and body weight of
43 mice were checked once or twice a week. The volume was calculated as (tumor length x tumor
44 width²) / 2. When a tumor reached >700~1000 mm³, the mouse was sacrificed, and tumor tissues
45 were stored. Tumor tissues were divided and stored for several purposes: (1) Tumor tissues were
46 cryopreserved in liquid nitrogen and stored at -80 °C for generating next passage PDXs. (2)
47 Tumor tissues were frozen in liquid nitrogen for genomic analysis. Whole-exome sequencing was
48 conducted as follows: Genomic DNA (gDNA) was extracted from blood and tissues using DNeasy
49 blood and tissue kit (QIAGEN) and checked for purity, concentration, and integrity by OD260/280
50 ratio using NanoDrop Instruments (NanoDrop Technologies, Wilmington, DE, USA) and agarose
51 gel electrophoresis. DNA was sheared by fragmentation by Bioruptor (Diagenode, Inc., Denville,
52 NJ, USA) and purified using Agencourt AMPure XP beads (Beckman Coulter, Fullerton, CA,
53 USA). DNA samples were then tested for size distribution and concentration using an Agilent
54 Bioanalyzer 2100. Standard protocols were utilized for adaptor ligation, indexing, high-fidelity
55 PCR amplification. Subsequently, exome enrichment was performed by hybrid capture with the
56 All Exon v5 capture library. Capture libraries were amplified, pooled, and submitted to the
57 commercial sequencing company (Macrogen) for 100bp paired-end, multiplex sequencing on a
58 HiSeq 2000 sequencing system. Average coverage for normal samples was 62.67x (38.97 min –
59 108.77 max), and was 102.35x for tumor samples (36.02 min – 150.49 max). RNA-Sequencing
60 data was generated as follows: RNA was extracted from tissues using the RNeasy Mini Kit
61 (Qiagen, Valencia, CA, USA). RNA-Sequencing libraries were prepared from 1 µg total RNA using
62 the TruSeq RNA Sample Preparation v2 Kit (Illumina, San Diego, CA) according to the
63 manufacturer's protocol. Libraries were submitted to the commercial sequencing company
64 (Macrogen) for 100bp paired-end, multiplex sequencing on a HiSeq 2000 sequencer.

65 **Huntsman Cancer Institute (HCI).** Patient tumor engraftment and PDX passaging of breast
66 cancer samples were performed as previously described^{4,5}. SNP array samples were genotyped
67 by the Affymetrix SNP 6.0 array for profiling. These samples were processed, according to
68 DeRose et al⁵. Additionally, some samples, were also processed using the Illumina Infinium Omni

69 2.5 Exome-8 v1.3 Beadchip array. Hybridized arrays were scanned using an Illumina iScan
70 instrument following the Illumina Infinium LCG Assay Manual Protocol and processed using
71 GenomeStudio. When samples had both Affymetrix and Illumina chips, we deferred to Illumina
72 intensity values for copy number calling. Whole-exome sequencing was conducted as follows:
73 Agilent SureSelectXT Human All Exon V6+COSMIC or Agilent Human All Exon 50Mb library
74 preparation protocols were used with inputs of 100-3000ng sheared genomic DNA (Covaris).
75 Library construction was performed using the Agilent Technologies SureSelectXT Reagent Kit.
76 The concentration of the amplified library was measured using a Qubit dsDNA HS Assay Kit
77 (ThermoFisher Scientific). Amplified libraries (750 ng) were enriched for exonic regions using
78 either the Agilent Technologies SureSelectXT Human All Exon v6+COSMIC or Agilent Human All
79 Exon 50Mb kits and PCR amplified. Enriched libraries were qualified on an Agilent Technologies
80 2200 TapeStation using a High Sensitivity D1000 ScreenTape assay and the molarity of adapter-
81 modified molecules was defined by quantitative PCR using the Kapa Biosystems Kapa Library
82 Quant Kit. The molarity of individual libraries was normalized to 5 nM, and equal volumes were
83 pooled in preparation for Illumina sequence analysis. Sequencing libraries (25 pM) were
84 chemically denatured and applied to an Illumina HiSeq v4 paired-end flow cell using an Illumina
85 cBot. Hybridized molecules were clonally amplified and annealed to sequencing primers with
86 reagents from an Illumina HiSeq PE Cluster Kit v4-cBot (PE-401-4001). Following the transfer of
87 the flowcell to an Illumina HiSeq 2500 instrument (HCS v2.2.38 and RTA v1.18.61), a 125-cycle
88 paired-end sequence run was performed using HiSeq SBS Kit v4 sequencing reagents (FC-401-
89 4003). Average coverage for normal samples was 90.22x (15.28 min – 131.69 max), and was
90 96.66x for tumor samples (10.65 min – 166.06 max).

91 **Baylor College of Medicine (BCM).** Patient tumor engraftment and PDX passaging of breast
92 cancer samples were performed as previously described^{6,7}. SNP array samples were genotyped
93 at Huntsman Cancer Institute using the Illumina Infinium Omni 2.5Exome-8 v1.4 Beadchip array
94 by the procedures provided in the HCl section above.

95 **The University of Texas MD Anderson Cancer Center (MDACC).** Fresh non-small-cell lung
96 carcinoma tumor samples were collected from surgically resected specimens with the informed
97 consent of the patients. Generation and passaging of PDXs, and histological analysis and DNA
98 fingerprint assay for PDXs and their primary tumor tissues were performed as previously
99 described⁸. The protocols for the use of clinical specimens and data in this study were approved
100 by the Institutional Review Board at The University of Texas MD Anderson Cancer Center. All
101 animal studies were carried out in accordance with the Guidelines for the Care and Use of
102 Laboratory Animals (National Institutes of Health Publication 85-23) and the institutional

103 guidelines of MDACC. Whole-exome sequencing was conducted at the Sequencing and
104 Microarray Core Facility at MD Anderson Cancer Center as follows: Genomic DNA was quantified
105 and quality was assessed using Picogreen (Invitrogen) and Genomic DNA Tape for the 2200
106 Tapestation (Agilent), respectively. DNA from each sample (100-500 ng of genomic DNA) was
107 sheared by sonication and then used for library preparation by using KAPA library preparation kit
108 (KAPA) following manufacturer's instruction. Equimolar amounts of DNA were pooled (2-6
109 samples per pool) and whole exome regions were captured by using biotin labeled probes from
110 Roche Nimblegen (Exome V3) followed manufacture's protocol. The captured libraries were
111 sequenced on a HiSeq 2000 with 100bp paired-end (Illumina Inc., San Diego, CA, USA) on a
112 paired-end flowcell. Average coverage for normal samples was 85.61x (40.80 min – 228.41 max),
113 and was 125.79x for tumor samples (25.12 min – 251.53 max).

114 ***The WISTAR Institute (WISTAR)***. Tumor biopsy samples were collected according to IRB-
115 approved protocol with the informed written consent of the patients. Collected fresh tumor pieces
116 were snap frozen and stored at -80 °C. Subcutaneous implantation into NSG SCID mice were
117 used to create PDX models. BRAF inhibitor treatment (PLX) was administered as PLX4720
118 200ppm chemical additive diet chow (Research Diets, New Brunswick, NJ). Whole exome
119 sequencing was conducted as follows: Genome DNA extraction was done using Qiagen DNeasy
120 Blood & Tissue Kit, and libraries for whole exome sequencing were performed using Nextera DNA
121 exome kit. Capture libraries were amplified, pooled, and then sequenced on an Illumina HiSeq
122 2500 76bp paired-end run. Average coverage for normal samples was 97.50x (71.46 min – 124.64
123 max), and was 208.27x for tumor samples (146.88 min – 281.20 max).

124 ***National Cancer Institute Patient-Derived Models Repository (PDMR)***. For engraftments,
125 tumor material plus a drop of Matrigel (BD BioSciences, Bedford, MA) were implanted
126 subcutaneously in NSG™ mouse model NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ. Mice were housed in
127 sterile, filter-capped polycarbonate cages, maintained in a barrier facility on a 12-hour light/dark
128 cycle, and were provided sterilized food and water, ad libitum. Animals were monitored weekly
129 for tumor growth. The initial passage of material was grown to approximately 1000-2000mm³
130 calculated using the following formula: weight (mg) = (tumor length x [tumor width]²) / 2. Tumor
131 material was then harvested, a portion cryopreserved, and the remainder implanted into NSG
132 host mice. Every PDX tumor harvested and cryopreserved also has 2-3 fragments snap frozen
133 for next generation sequence analysis and short tandem repeat validation and a piece is fixed in
134 neutral buffered formalin and then embedded in paraffin for histological assessment. Related
135 patient data, clinical history, representative histology and short-tandem repeat profiles for the PDX
136 models can be found at <https://pdmr.cancer.gov>. Full PDMR standard operating procedures for

137 tumor engraftment and PDX passaging are available at <https://pdmr.cancer.gov/sops>. Whole-
138 exome sequencing data were generated with the Agilent SureSelect capture kit, and sequenced
139 with 125bp pair-end Illumina HiSeq 2500 runs following standard operating procedures available
140 here: <https://pdmr.cancer.gov/sops>. Average coverage for normal samples was 148.47x (50.95
141 min – 242.24 max), and was 174.77x for tumor samples (81.41 min – 403.22 max).

142 **Washington University in St. Louis (WUSTL).** All human tissues acquired for these
143 experiments were processed in compliance with NIH regulations and institutional guidelines,
144 approved by the Institutional Review Board at Washington University. Tumors from all patients
145 were obtained via core needle biopsy, skin punch biopsy, or surgical resection after informed
146 consent. All animal procedures were reviewed and approved by the Institutional Animal Care and
147 Use Committee at Washington University in St. Louis. Pancreatic cancer models were derived
148 from tissue fragments implanted subcutaneously into dorsal flank regions of non-humanized,
149 female NOD/SCID/γ mice (Jackson Laboratory, Bar Harbor, ME) using Matrigel. The sample
150 tissues for these PDX models were obtained from archived, cryopreserved PDX harvests. Final
151 tumor passages in mice were kept cold and harvested into RPMI-1640 with antibiotic and
152 antimycotic additives. Pieces of each tumor were processed into the following: flash frozen tissue
153 fragments, OCT blocks and matched Haemotoxylin and Eosin (H&E) slides, formalin fixed paraffin
154 blocks and matched H&E slides, RNAlater tissue storage, and cryopreserved fragments (FBS +
155 10% DMSO). A minimum of 250 mg of flash frozen material was submitted to the Siteman Cancer
156 Center's Proteomics Core. The tissues were cryo-pulverized and subsequently divided for DNA
157 and RNA preparation, and long-term storage. Patient tumors were obtained directly from
158 operating rooms and placed into sterile collection media (RPMI-1640 with antibiotic and
159 antimycotic additives). Pieces of each tumor were processed into the following: flash frozen tissue
160 fragments, OCT blocks and matched H&E slides, formalin fixed paraffin blocks and matched H&E
161 slides, and cryopreserved fragments (FBS + 10% DMSO). Parental genomic DNA was prepared
162 from OCT blocks if available, and if not available, paraffin blocks were utilized. In addition,
163 genomic DNA for sequencing control was prepped from peripheral blood mononuclear cells that
164 were both procured and processed at time of surgery. Breast cancer models were derived from
165 tissue fragments implanted subcutaneously into dorsal flank regions of non-humanized,
166 NOD/SCID/γ mice (Jackson Laboratories, Bar Harbor, ME) as previously described^{7,9}. Whole-
167 exome sequencing was conducted as follows: Libraries were constructed using unamplified
168 genomic DNA (minimum 100 ng) from blood (normal), tumor, and xenograft samples. Exons were
169 captured via IDT Exome library kit followed by high-throughput sequencing on an Illumina
170 NovaSeq S4 platform (Illumina Inc., San Diego, CA) using 150bp paired-end reads. Details of

171 whole exome library construction have been given elsewhere (Fisher, Barry et al. 2011). Average
172 coverage for normal pancreatic cancer samples was 85.73x (55.65 min – 108.91 max), and was
173 124.01x (49.68 min – 242.35 max) for tumor pancreatic cancer samples. Average coverage for
174 normal breast cancer samples was 58.33x (45.37 min – 70.30 max), and was 89.90x (17.24 min
175 – 149.53 max) for tumor breast cancer samples.

176 **Shanghai Institute for Biological Sciences (SIBS).** Gene expression and copy number data,
177 generated by the Affymetrix Human Genome U133 Plus 2.0 Array and Affymetrix Human SNP
178 6.0 platforms respectively, of hepatocellular carcinoma (HCC) PDX models were retrieved from
179 the Gene Expression Omnibus (GEO) accession ID GSE90653¹⁰. Expression microarray data
180 generated by the Affymetrix Human Genome U133 Plus 2.0 Array for normal liver were
181 downloaded from GEO and ArrayExpress: GSE3526¹¹, GSE33006¹² and E-MTAB-1503-3¹³.

182 **EurOPDX colorectal cancer (EuroPDX CRC).** Liver-metastatic colorectal cancer samples were
183 obtained from surgical resection of liver metastases at the Candiolo Cancer Institute, the
184 Mauriziano Umberto I Hospital, and the San Giovanni Battista Hospital. Informed consent for
185 research use was obtained from all patients at the enrolling institution before tissue banking, and
186 study approval was obtained from the ethics committees of the three centers. Tissue from hepatic
187 metastasectomy in affected individuals was fragmented and either frozen or prepared for
188 implantation as described previously^{14,15}. Non-obese diabetic/severe combined immunodeficient
189 (NOD/SCID) female mice (4–6 weeks old) were used for tumor implantation. Snap-frozen aliquots
190 were obtained from surgical specimens and corresponding tumor grafts at different passages.
191 Whole genome sequencing was conducted as follows: DNA was extracted using Maxwell RSC
192 Blood DNA kit (Promega AS1400) from colorectal cancer liver metastasis and corresponding
193 tumor grafts at different passages. Genomic DNA was fragmented and used for Illumina TruSeq
194 library construction (Illumina) according to the manufacturer's instructions. Libraries were then
195 purified with Qiagen MinElute column purification kit and eluted in 17 µl of 70°C EB to obtain 15
196 µl of DNA library. The libraries were sequenced on HiSeq4000 (Illumina) with single-end reads of
197 51bp at low coverage (~0.1x genome coverage on average).

198 **EurOPDX breast cancer (EuroPDX BRCA).** Human breast tumors were obtained from surgical
199 resections at the Netherland Cancer Institute (NKI), Institut Curie (IC) and Vall d'Hebron Institute
200 of Oncology (VHIO). Engraftment was conducted with different procedures at each center. NKI:
201 Small tumor fragments (2mm diameter) were implanted into the 4th mammary fat pad of 8-week-
202 old Swiss female nude mice. Mice were checked for tumor appearance once a week, and
203 supplemented with estrogen, if the tumor was ER positive. After palpable tumor detection, tumor
204 size was measured twice a week. When tumors reached a size of 700-1000 mm³, animals were

205 sacrificed and tumors were explanted and subdivided in fragments for serial transplantation as
206 described above, or for frozen vital storage in liquid nitrogen. IC: Breast cancer fragments were
207 obtained from patients at the time of surgery, with informed written patient consent. Fragments of
208 30 to 60 mm³ were grafted into the interscapular fat pad of 8 to 12-week-old female Swiss nude
209 mice. Mice were supplemented with estrogen. Xenografts appeared at the graft site 2 to 8 months
210 after grafting. When tumors were close to 1500 mm³, they were subsequently transplanted from
211 mouse to mouse and stocked frozen in DMSO-fetal calf serum (FCS) solution or frozen dried in
212 nitrogen. Fragment fixed tissues in phosphate buffered saline (PBS) 10% formol for histologic
213 studies were also stored. The experimental protocol and animal housing were in accordance with
214 institutional guidelines as proposed by the French Ethics Committee (Agreement B75-05-18,
215 France). VHIO: Fresh tumor samples from patients with breast cancer were collected for
216 implantation following an institutional IRB-approved protocol and the associated informed
217 consent, or by the National Research Ethics Service, Cambridgeshire 2 REC (REC reference
218 number: 08/H0308/178). Experiments were conducted following the European Union's animal
219 care directive (2010/63/EU) and were approved by the Ethical Committee of Animal
220 Experimentation of the Vall d'Hebron Research Institute. Surgical or biopsy specimens from
221 primary tumors or metastatic lesions were immediately implanted in mice. Fragments of 30 to 60
222 mm³ were implanted into the mammary fat pad (surgery samples) or the lower flank (metastatic
223 samples) of 6-week-old female athymic HsdCpb:NMRI-Foxn1nu mice (Harlan Laboratories).
224 Animals were continuously supplemented with estradiol. Upon growth of the engrafted tumors,
225 the model was perpetuated by serial transplantation onto the lower flank. Tumor growth was
226 measured with caliper bi-weekly. In all experiments, mouse weight was recorded twice weekly.
227 When tumors reached 1500 mm³, mice were euthanized and tumors were explanted. Whole
228 genome sequencing was conducted as follows: genomic DNA was extracted from breast cancers
229 and corresponding PDXs using (i) QIAamp DNA Mini Kit s(50) (#51304, Qiagen) (IC) or (ii)
230 according to Laird PW's protocol¹⁶ (NKI and VHIO). The amount of double stranded DNA in the
231 genomic DNA samples was quantified by using the Qubit® dsDNA HS Assay Kit (Invitrogen, cat
232 no Q32851). Up to 2000 ng of double stranded genomic DNA were fragmented by Covaris
233 shearing to obtain fragment sizes of 160-180bp. Samples were purified using 1.6X Agencourt
234 AMPure XP PCR Purification beads according to manufacturer's instructions (Beckman Coulter,
235 cat no A63881). The sheared DNA samples were quantified and qualified on a BioAnalyzer
236 system using the DNA7500 assay kit (Agilent Technologies cat no. 5067-1506). With an input of
237 maximum 1 µg sheared DNA, library preparation for Illumina sequencing was performed using
238 the KAPA HTP Library Preparation Kit (KAPA Biosystems, KK8234). During library enrichment,

239 4-6 PCR cycles were used to obtain enough yield for sequencing. After library preparation the
240 libraries were cleaned up using 1X AMPure XP beads. All DNA libraries were analyzed on the GX
241 Caliper (a PerkinElmer company) using the HT DNA High Sensitivity LabChip, for determining the
242 molarity. Up to two pools of 24 uniquely indexed samples and one pool of 81 uniquely indexed
243 samples were mixed together by equimolar pooling in a final concentration of 10nM, and
244 subjected to sequencing on an Illumina HiSeq2500 machine in a total of 12 lanes of a single read
245 65bp run at low coverage (~0.4x genome coverage on average), according to manufacturer's
246 instructions.

247

248 **Consolidating tumor types from different datasets**

249 As the terminology of tumor types/subtypes by the different contributing centers were not
250 consistent, we used the Disease Ontology database¹⁷ (<http://disease-ontology.org/>), cancer types
251 listed in NCI website (<https://www.cancer.gov/types>) and in TCGA publications^{18,19} to unify and
252 group the tumor types/subtypes under broader terms as shown in Fig.1 and Supplementary Table
253 2.

254

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

256 **SNP array.** The estimation of CNA profiles from SNP array were detailed previously³. In short, for
257 Affymetrix Human SNP 6.0 arrays, PennCNV-Affy and Affymetrix Power Tools²⁰ were used to
258 extract the B-allele frequency (BAF) and Log R Ratio (LRR) from the CEL files. Due to the
259 absence of paired-normal samples, the allele-specific signal intensity for each PDX tumor were
260 normalized relative to 300 randomly selected sex-matched Affymetrix Human SNP 6.0 array CEL
261 files obtained from the International HapMap project²¹. For Illumina Infinium Omni2.5Exome-8
262 SNP arrays (v1.3 and v1.4 kit), the Illumina GenomeStudio software was used to extract the B-
263 allele frequency (BAF) and Log R Ratio (LRR) from the signal intensity of each probe. The single
264 sample mode of the Illumina GenomeStudio was used, which normalizes the signal intensities of
265 the probes with an Illumina in-house dataset. The single tumor version of ASCAT²² (v2.4.3 for
266 JAX SNP data, v2.5.1 for SIBS SNP data) was used for GC correction, predictions of the
267 heterozygous germline SNPs based on the SNP array platform, and estimation of ploidy, tumor
268 content and allele-specific copy number segments. The resultant copy number segments were
269 annotated with log₂ ratio of total copy number relative to predicted ploidy from ASCAT.

270 **Whole-exome sequencing (WES) data.** All the samples were subjected to quality control
271 (filtering and trimming of poor-quality reads and bases) using in-house QC script with the cut-off
272 that half of the read length should be ≥20 in base quality at phred scale. We further removed the

273 known adaptors using cut-adapt²³ v1.15 11 at -m 36. Afterward, we aligned the reads to the
274 human genome (GRCh38.p5) using bwakit²⁴ v0.7.15. Engrafted tumor samples were subjected
275 to the additional step of mouse read removal using Xenome²⁵ v1.0.0, with default parameters.
276 The alignment was converted to BAM format using Picard SortSam v2.8.1
277 (<https://broadinstitute.github.io/picard/>), and duplicates were removed by Picard MarkDuplicates
278 utility. BaseRecalibrator from the Genome Analysis Tool Kit^{26,27} (GATK) v4.0.5.1 was used to
279 adjust the quality of raw reads. Training files for the base quality scale recalibration were
280 Mills_and_1000G_gold_standard.indels.hg38.vcf.gz,
281 Homo_sapiens_assembly38.known_indels.vcf.gz, and dbSNP v151. Mean target coverage was
282 determined for each sample by Picard CollectHsMetrics. Aligned bams were subset to target
283 region by GATK and SAMTools²⁸ v0.1.18 was used to generate the pileup for each sample. Pileup
284 data were used for CNA estimation as calculated with Sequenza²⁹ v2.1.2. Both tumor and normal
285 data, that utilized the same capture array, were used as input. pileup2seqz and GC-windows (-w
286 50) modules from sequenza-utils.py utility were used to create the native seqz format file for
287 Sequenza and compute the average GC content in sliding windows from hg38 genome,
288 respectively. Finally, we ran the three Sequenza modules with these modified parameters
289 (sequenza.extract: assembly = "hg38", sequenza.fit: chromosome.list = 1:23, and
290 sequenza.results: chromosome.list = 1:23) to estimate the segments of copy number
291 gains/losses. Finally, segments lacking read counts, in which ≥50% of the segment with zero read
292 coverage, were removed. A reference implementation of this workflow (Supplementary Fig. 77)
293 is developed and deployed in the cancer genomics cloud at SevenBridges
294 (<https://cgc.sbggenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-workflow/>,
295 <https://cgc.sbggenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-xenome-tumor-normal-workflow/>).

297 **Low-pass whole-genome sequencing (WGS) data.** Whole-genome sequence reads from
298 EuroPDX CRC liver metastasis and corresponding tumor grafts at different passages were
299 mapped to the reference human genome (GRCh37) using Burrows-Wheeler Aligner²⁴ (BWA)
300 v0.7.12. SAMTools²⁸ v0.1.18 was used to convert SAM files into BAM files and Picard v1.43 to
301 remove PCR duplicates (<http://broadinstitute.github.io/picard/>). Raw copy number profiles for
302 each sample were estimated by QDNAseq³⁰ R package v1.20 by dividing the human reference
303 genome in non-overlapping 50 kb windows and counting the number of reads in each bin. Bins in
304 problematic regions were removed³¹. Read counts were corrected for GC content and mappability
305 by a LOESS regression, median-normalized and log₂-transformed. Values below -1000 in each
306 chromosome were floored to the first value greater than -1000 in the same chromosome. Raw

307 \log_2 ratio values were then segmented using the ASCAT²² algorithm implemented in the ASCAT
308 R package v2.0.7. Whole-genome sequence reads from EuroPDX BRCA tumors and
309 corresponding tumor grafts at different passages were mapped to the reference human genome
310 (GRCh38) and mouse genome (GRCm38/mm10, Ensembl 76) using Burrows-Wheeler Aligner
311 (BWA) v0.7.15. Subsequently, mouse reads were excluded with Xenofilter³². Other processing
312 steps are similar as described above. Raw copy number profiles were estimated for each sample
313 by dividing the human reference genome in non-overlapping 20 kb windows and counting the
314 number of reads in each bin. Only reads with at least mapping quality 37 were considered. Bins
315 within problematic regions (i.e. multimapper regions) were excluded. Downstream analysis to
316 estimate copy number was conducted as described above.

317 **RNA-sequencing (RNA-Seq) and gene expression microarray (EXPARR) data.** For SNU-JAX
318 RNA-Seq data, Simultaneous read alignment was performed to both mouse (mm10) and human
319 genome (GRCh38.p5) and only human specific reads were used for the expression quantification.
320 Expression of mRNA was quantified as Transcripts Per Million (TPM) for downstream analysis
321 using RNA-Seq by Expectation Maximization³³ (RSEM) with ensemble GTF reference
322 GRCh38.92. Gene expression microarray data for SIBS HCC and normal liver samples from GEO
323 and ArrayExpress databases were profiled as follows. After initial quality control and outlier
324 removal, CEL files were normalized according to RMA algorithm and probesets were annotated
325 according to Affymetrix annotation file for HG-U133 Plus 2, released on 2016-03-15 build 36. For
326 expression-based copy number inference, we referred to the previous protocols for e-karyotyping
327 and CGH-Explorer³⁴⁻³⁷. For each cancer type, expression values of tumor and corresponding
328 normal samples were merged in a single table, and gene identifiers were annotated with
329 chromosomal nucleotide positions. Genes located on sex chromosomes were excluded. Genes
330 which values below 1 TPM (RNAseq) or probeset \log_2 -values below 6 (microarray) in more than
331 20% of the analyzed dataset were removed. Remaining gene expression values below the
332 thresholds were respectively raised to 1 TPM or \log_2 -value of 6. In the case of multiple transcripts
333 (RNA-seq) or probesets (microarray) per gene, the one with the highest median value across the
334 entire dataset was selected. According to the e-karyotyping protocol, the sum of squares of the
335 expression values relative to their median expression across all samples was calculated for each
336 gene, and 10% most highly variable genes were removed. For each gene, the median \log_2
337 expression value in normal samples was subtracted from the \log_2 expression value in each tumor
338 sample and subsequently input in CGH-explorer. For tumor-only datasets, the median \log_2
339 expression value in the same set of tumor samples was instead subtracted. The preprocessed
340 expression profiles of each sample were individually analyzed using CGH-Explorer

341 (<http://heim.ifi.uio.no/bioinf/Projects/CGHExplorer/>). CGH-PCF analysis was carried out to call
342 copy number according to parameters previously reported³⁸: least allowed deviation = 0.25; least
343 allowed aberration size = 30; winsorize at quantile = 0.001; penalty = 12; threshold = 0.01.

344

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

346 Copy number (CN) segments with \log_2 copy number ratio estimated from the various platforms
347 were processed in the following steps (Supplementary Fig. 26). Segments <1kb were filtered
348 based on the definition of CNA³⁹. In addition, SNP array segments had to be covered by >10
349 probes, with an average probe density of 1 probe per 5kb. The copy number segments were then
350 binned into 10kb windows to derive the median \log_2 (CN ratio), which was subsequently used to
351 re-center the copy number segments. Median-centered copy number segments were visualized
352 using IGV⁴⁰ v2.4.13 and GenVisR⁴¹ v1.16.1. Median-centered copy number of genes were
353 calculated by intersecting the genome coordinates of copy number segments with the genome
354 coordinates of genes (Ensembl Genes 93 for human genome assembly GRCh38, Ensembl
355 Genes 96 for human genome assembly GRCh37). In the case where a gene overlaps multiple
356 segments, the most conservative (lowest) estimate of copy number was used to represent the
357 copy number of the entire intact gene.

358

359 **Comparison of CN gains and losses**

360 For the comparison of resolution, range of CN values and frequency of gains and losses between
361 different platforms and analysis methods, we defined copy number gain or loss segments as –
362 Gain: $\log_2(\text{CNratio}) > 0.1$; Loss: $\log_2(\text{CN ratio}) < -0.1$.

363

364 **Correlation of CNA profiles**

365 The overall workflow to compare CNA profiles is shown in Supplementary Fig. 26. PDX samples
366 without passage information were omitted in the following downstream analysis. The copy number
367 segments were binned into 10kb-windows or smaller using Bedtools⁴² v2.26.0, and the variance
368 of $\log_2(\text{CN ratio})$ and range (difference) of $\log_2(\text{CN ratio})$ between 5th to 95th percentile across all
369 the bins were calculated as a measure of degree of aberration for each CNA profile. A non-
370 aberrant profile results in a low variance or range. While variance can be biased for CNA profiles
371 with small segments of extreme gains or losses, we preferred the use of 5th to 95th percentile
372 range to identify samples with low degree of aberration, such that a narrow range indicates ≥90%
373 of the genome has very low-level gains and losses. The similarity of two CNA profiles is quantified
374 by the Pearson correlation coefficient of $\log_2(\text{CN ratio})$ of 100kb-windows binned from segments

375 or genes between 2 samples. Gene-based and segment-based (100kb windows) correlations
376 were highly similar (data not shown). Using correlation avoided the issue of making copy number
377 gain and loss calls based on thresholds, though it can be inconsistent due to different baseline
378 and range in copy number values. Such variations are impacted by sample-specific variation in
379 human stromal contamination or sensitivity copy number detection by different platforms.

380 **Comparison of CNA profiles between different platforms.** The copy number segments of each
381 pair of data were intersected and binned into 100kb-windows or smaller using Bedtools. The
382 Pearson correlation coefficient and linear regression model was calculated for the $\log_2(\text{CN ratio})$
383 of the windows. Windows with discrepant copy number were identified by outliers of the linear
384 regression model defined by $|\text{studentized residual}| > 3$. These outlier windows were mapped to
385 their corresponding segments to identify the size of CNA events that were discordant between
386 the different copy number estimation methods. The proportion of the genome discordant CNA
387 was calculated from the summation of the outlier windows.

388 **Identification of genes with CNA between different samples of the same model.** To compare
389 the CNA profiles between different samples (PT or PDX) of the same model, the Pearson
390 correlation coefficient and linear regression model was calculated for the $\log_2(\text{CN ratio})$ of the
391 genes for each pair of data. Prior to that, deleted genes with $\log_2(\text{CN ratio}) < -3$ were rescaled to
392 -3 to avoid large shifts in the correlation coefficient and linear regression model due to extremely
393 negative values on the log scale. Extreme outliers of the linear regression model defined by
394 $|\text{studentized residual}| > 3$ were removed to derive an improved linear regression model⁴³ not
395 biased by few extreme values. Genes with copy number changes between the samples were
396 identified by the difference in $\log_2(\text{CN ratio})$ relative to the improved linear regression model of
397 $|\text{standard residual}| < 0.5$. We also removed some samples with low correlation due to sample
398 mislabeling as they displayed high correlation with samples from other models. We also omit
399 samples with low correlation values (<0.6) which resulted from non-aberrant CNA profiles in
400 genetically stable tumors (5th to 95th percentile range < 0.3, Supplementary Fig. 64).

401 **Identification of aberrant sample pairs with highly discordant CNA profiles.** Aberrant CNA
402 profiles were identified based on the 100kb-window copy number range (5th to 95th percentile)
403 >0.5, for both samples. Sample pairs with Pearson correlation <0.6 were selected as highly
404 discordant CNA profiles between them.

405

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

407 Copy number altered genes ($|\text{residual}| < 0.5$) were annotated by various gene sets with cancer
408 or treatment-related functions gathered from various databases and publications (Supplementary
409 Fig. 26):

410 1. Genes in 10 oncogenic signaling pathways curated by TCGA and were found to be frequently
411 altered in different cancer types⁴⁴.

412 2. Genes with gain in copy number or expression, or loss in copy number or expression that
413 conferred therapeutic sensitivity, resistance or increase/decrease in drug response from the JAX
414 Clinical Knowledgebase^{45,46} (JAX-CKB) based on literature curation (<https://ckbhome.jax.org/>, as
415 of 06-18-2019).

416 3. Genes with evidence of promoting oncogenic transformation by amplification or deletion from
417 the Cancer Gene Census⁴⁷ (COSMIC v89).

418 4. Significantly amplified or deleted genes in TCGA cohorts of breast cancer⁴⁸, colorectal cancer⁴⁹,
419 lung adenocarcinoma⁵⁰ and lung squamous cell carcinoma⁵¹ by GISTIC analysis.

420

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

422 Genes with a more stringent threshold of $|\text{residual}| > 1.0$ with respect to the improved regression
423 linear model (without discriminating gain or loss) were selected for each pairwise comparison
424 between different samples of the same model. Pairwise cases in which genes are deleted in both
425 samples ($\log_2(\text{CN ratio}) \leq -3$) are omitted. Recurrent frequency for each gene across all models
426 was calculated on a model basis such that genes with copy number between multiple pairs of the
427 same model was counted as once. This avoided the bias towards models with many samples of
428 similar copy number changes between the different pairs.

429

430 **Drug response analysis using CCLE data**

431 We developed a pipeline to evaluate gene copy number effects on drug sensitivity^{52,53} by using
432 the Cancer Cell Line Encyclopedia^{54,55} (CCLE) cell line genomic and drug response data (CTRP
433 v2). We downloaded the CCLE drug response data from Cancer Therapeutics Response Portal
434 (www.broadinstitute.org/ctrp), and CCLE gene-level CNA and gene expression data from depMap
435 data portal ('public_19Q1_gene_cn.csv' and 'CCLE_depMap_19Q1 TPM.csv',
436 <https://depmap.org/portal/download/>). For CCLE drug response data, we used the area-under-
437 concentration-response curve (AUC) sensitivity scores for each cancer cell line and each drug. In
438 total, we collected gene-level \log_2 copy number ratio data derived from the Affymetrix SNP 6.0
439 platform from 668 pan-cancer CCLE cell lines, with a total of 545 cancer drugs tested. With the

440 CCLE gene-level CNA and AUC drug sensitivity scores, we performed gene-drug response
441 association analyses for genes with recurrent copy number changes. Pearson correlation p-
442 values between each gene's \log_2 (CN ratio) and each drug's AUC score across all cell lines were
443 calculated, and q-values were calculated by multiple testing Bonferroni correction. Significant
444 gene-CNA and drug associations were kept (q-value < 0.1) to further evaluate gene-expression
445 and drug response associations. If a gene's expression was also significantly correlated with AUC
446 drug sensitivity scores, particularly in the same direction (either positively or negatively correlated)
447 as the gene-CNA and drug association, that gene would be considered as significantly correlated
448 with drug response based on both its CNA and gene expression.

449

450 **GISTIC analysis of WGS data**

451 To obtain perfectly matching and comparable PT-PDX cohorts, for GISTIC analysis, CRC trios in
452 which at least one sample did not display significant CNAs were excluded from the analysis
453 resulting in a total of 87 triplets. The GISTIC⁵⁶ algorithm (GISTIC 2 v6.15.28) was applied on the
454 segmented profiles using the GISTIC GenePattern module (<https://cloud.genepattern.org/>), with
455 default parameters and genome reference files Human_Hg19.mat for EuroPDX CRC data and
456 hg38.UCSC.add_miR.160920.refgene.mat for EuroPDX BRCA data. For each dataset, GISTIC
457 provides separate results (including segments, G-scores and FDR q-values) separately for
458 recurrent amplifications and recurrent deletions. Deletion G-scores were assigned negative
459 values for visualization. We observed that the G-Score range was systematically lower in PT
460 cohorts, which is likely the result of the dilution of CNA by normal stromal DNA. In contrast, human
461 stromal DNA in PDX samples were lower or negligible. To account for this difference in gene-
462 level G-scores, PDXs at early and late passages were scaled with respect to PT gene-level G-
463 score values using global linear regression, separately for amplification and deletion outputs.

464

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

466 To assess the biological functions associated with the recurrent alterations detected by the
467 GISTIC analysis, we performed GSEAPreranked analysis^{57,58} on gene-level GISTIC G-score
468 profiles, for both amplifications and deletions. In particular, we applied the algorithm with 1000
469 permutations on various gene set collections from the Molecular Signatures Database^{59,60}
470 (MSigDB): H (Hallmark), C2 (Curated : CGP chemical and genetic perturbations, CP canonical
471 pathways), C5 (Gene Ontology: BP biological process, MF molecular function, CC cellular
472 component) and C6 (Oncogenic Signatures) composed of 50, 4762, 5917 and 189 gene sets
473 respectively. We also included gene sets with known cancer or treatment-related functions

474 described in an earlier section. We noted that multiple genes with contiguous chromosomal
475 locations, typically in recurrent amplicons, generated spurious enrichment for gene sets which
476 consists of multiple genes of adjacent positions, while very few or none of them had a significant
477 GISTIC G-score. To avoid this confounding issue, we only considered the “leading edge genes”,
478 i.e. those genes with increasing Normalized Enrichment Score (NES) up to its maximum value,
479 that contribute to the GSEA significance for a given gene set. The leading-edge subset can be
480 interpreted as the core that accounts for the gene set’s enrichment signal
481 (<http://software.broadinstitute.org/gsea>). We included a requirement that the leading edge genes
482 passing the GISTIC G-score significant thresholds based on GISTIC q-value 0.25 (Supplementary
483 Table 8 and Fig. 73) make up at least 20% of the gene set. This 20% threshold was chosen as
484 the minimal threshold at which gene sets assembled from TCGA-generated lists of genes with
485 recurrent CNA in CRC or BRCA were identified as significant in GSEA (see Supplementary Table
486 9). Finally, gene sets with a NES greater than 1.5 and a FDR q-value of less than 0.05, which
487 passed the leading edge criteria, were considered significantly enriched in genes affected by
488 recurrent CNAs.

489

490 REFERENCES

- 491 1. Krupke, D.M. *et al.* The Mouse Tumor Biology Database: A Comprehensive Resource
492 for Mouse Models of Human Cancer. *Cancer Res* **77**, e67-e70 (2017).
- 493 2. Conte, N. *et al.* PDX Finder: A portal for patient-derived tumor xenograft model
494 discovery. *Nucleic Acids Res* **47**, D1073-D1079 (2019).
- 495 3. Woo, X.Y. *et al.* Genomic data analysis workflows for tumors from patient-derived
496 xenografts (PDXs): challenges and guidelines. *BMC Medical Genomics* **12**, 92 (2019).
- 497 4. DeRose, Y.S. *et al.* Patient-derived models of human breast cancer: protocols for in vitro
498 and in vivo applications in tumor biology and translational medicine. *Curr Protoc
499 Pharmacol Chapter 14*, Unit14 23 (2013).
- 500 5. DeRose, Y.S. *et al.* Tumor grafts derived from women with breast cancer authentically
501 reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Medicine* **17**,
502 1514 (2011).
- 503 6. Zhang, X.M. *et al.* A renewable tissue resource of phenotypically stable, biologically and
504 ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer
505 Research* **73**, 4885-4897 (2013).
- 506 7. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic
507 characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-30 (2013).
- 508 8. Chen, Y. *et al.* Tumor characteristics associated with engraftment of patient-derived non-
509 small cell lung cancer xenografts in immunocompromised mice. *Cancer* **0**(2019).
- 510 9. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and
511 xenograft. *Nature* **464**, 999 (2010).
- 512 10. He, S. *et al.* PDXliver: a database of liver cancer patient derived xenograft mouse
513 models. *BMC Cancer* **18**, 550 (2018).

- 514 11. Roth, R.B. *et al.* Gene expression analyses reveal molecular relationships among 20
515 regions of the human CNS. *Neurogenetics* **7**, 67-80 (2006).
- 516 12. Huang, Y. *et al.* Identification of a two-layer regulatory network of proliferation-related
517 microRNAs in hepatoma cells. *Nucleic Acids Research* **40**, 10478-10493 (2012).
- 518 13. Malouf, G.G. *et al.* Transcriptional profiling of pure fibrolamellar hepatocellular
519 carcinoma reveals an endocrine signature. *Hepatology* **59**, 2228-2237 (2014).
- 520 14. Bertotti, A. *et al.* A Molecularly Annotated Platform of Patient-Derived Xenografts
521 ("Xenopatients") Identifies HER2 as an Effective Therapeutic Target in Cetuximab-
522 Resistant Colorectal Cancer. *Cancer Discovery* **1**, 508-523 (2011).
- 523 15. Galimi, F. *et al.* Genetic and expression analysis of MET, MACC1, and HGF in
524 metastatic colorectal cancer: response to met inhibition in patient xenografts and
525 pathologic correlations. *Clin Cancer Res* **17**, 3146-56 (2011).
- 526 16. Laird, P.W. *et al.* Simplified mammalian DNA isolation procedure. *Nucleic Acids Res* **19**,
527 4293 (1991).
- 528 17. Schriml, L.M. *et al.* Human Disease Ontology 2018 update: classification, content and
529 workflow expansion. *Nucleic Acids Research* **47**, D955-D962 (2018).
- 530 18. Lawrence, M.S. *et al.* Comprehensive genomic characterization of head and neck
531 squamous cell carcinomas. *Nature* **517**, 576-582 (2015).
- 532 19. Lazar, A.J. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft
533 Tissue Sarcomas. *Cell* **171**, 950-+ (2017).
- 534 20. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-
535 resolution copy number variation detection in whole-genome SNP genotyping data.
536 *Genome Res* **17**, 1665-74 (2007).
- 537 21. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-96 (2003).
- 538 22. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U
539 S A* **107**, 16910-5 (2010).
- 540 23. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
541 *2011* **17**, 3 (2011).
- 542 24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
543 transform. *Bioinformatics* **25**, 1754-60 (2009).
- 544 25. Conway, T. *et al.* Xenome--a tool for classifying reads from xenograft samples.
545 *Bioinformatics* **28**, i172-8 (2012).
- 546 26. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
547 analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
- 548 27. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-
549 generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
- 550 28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
551 2078-9 (2009).
- 552 29. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor
553 sequencing data. *Ann Oncol* **26**, 64-70 (2015).
- 554 30. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by
555 shallow whole-genome sequencing with identification and exclusion of problematic
556 regions in the genome assembly. *Genome research* **24**, 2022-2032 (2014).
- 557 31. Desmedt, C. *et al.* Uncovering the genomic heterogeneity of multifocal breast cancer.
558 *Journal of Pathology* **236**, 457-466 (2015).

- 559 32. Kluin, R.J.C. *et al.* XenofilteR: computational deconvolution of mouse and human reads
560 in tumor xenograft sequence data. *Bmc Bioinformatics* **19**(2018).
- 561 33. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with
562 or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- 563 34. Ben-David, U., Mayshar, Y. & Benvenisty, N. Virtual karyotyping of pluripotent stem
564 cells on the basis of their global gene expression profiles. *Nature Protocols* **8**, 989
565 (2013).
- 566 35. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal
567 aberrations and recombination by allelic bias in RNA-Seq. *Nat Commun* **7**, 12144 (2016).
- 568 36. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse
569 models reveals driver-specific routes to tumorigenesis. *Nature Communications* **7**, 12160
570 (2016).
- 571 37. Lingjaerde, O.C., Baumbusch, L.O., Liestol, K., Glad, I.K. & Borresen-Dale, A.L. CGH-
572 Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 821-2 (2005).
- 573 38. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution.
574 *Nature Genetics* **49**, 1567 (2017).
- 575 39. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-
576 454 (2006).
- 577 40. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV):
578 high-performance genomics data visualization and exploration. *Briefings in
579 Bioinformatics* **14**, 178-192 (2013).
- 580 41. Skidmore, Z.L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012-
581 3014 (2016).
- 582 42. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
583 features. *Bioinformatics* **26**, 841-842 (2010).
- 584 43. Motulsky, H.J. & Brown, R.E. Detecting outliers when fitting data with nonlinear
585 regression – a new method based on robust nonlinear regression and the false discovery
586 rate. *BMC Bioinformatics* **7**, 123 (2006).
- 587 44. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas.
588 *Cell* **173**, 321-337.e10 (2018).
- 589 45. Patterson, S.E., Statz, C.M., Yin, T. & Mockus, S.M. Utility of the JAX Clinical
590 Knowledgebase in capture and assessment of complex genomic cancer data. *npj
591 Precision Oncology* **3**, 2 (2019).
- 592 46. Statz, C.M., Patterson, S.E., Yin, T. & Mockus, S.M. The JAX Clinical Knowledgebase:
593 A Valuable Resource for Rapid Extraction of Cancer-Related Data. *Journal of Molecular
594 Diagnostics* **18**, 1037-1038 (2016).
- 595 47. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction
596 across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
- 597 48. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular portraits of human breast
598 tumours. *Nature* **490**, 61 (2012).
- 599 49. The Cancer Genome Atlas, N. *et al.* Comprehensive molecular characterization of human
600 colon and rectal cancer. *Nature* **487**, 330 (2012).
- 601 50. The Cancer Genome Atlas Research, N. *et al.* Comprehensive molecular profiling of lung
602 adenocarcinoma. *Nature* **511**, 543 (2014).
- 603 51. The Cancer Genome Atlas Research, N. *et al.* Comprehensive genomic characterization
604 of squamous cell lung cancers. *Nature* **489**, 519 (2012).

- 605 52. Adams, D.J. *et al.* NAMPT Is the Cellular Target of STF-31-Like Small-Molecule
606 Probes. *AcS Chemical Biology* **9**, 2247-2254 (2014).
- 607 53. Viswanathan, V.S. *et al.* Dependency of a therapy-resistant state of cancer cells on a lipid
608 peroxidase pathway. *Nature* **547**, 453-+ (2017).
- 609 54. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
610 anticancer drug sensitivity. *Nature* **483**, 603 (2012).
- 611 55. Stransky, N. *et al.* Pharmacogenomic agreement between two cancer cell line data sets.
612 *Nature* **528**, 84-+ (2015).
- 613 56. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the
614 targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**,
615 R41 (2011).
- 616 57. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
617 interpreting genome-wide expression profiles. *Proceedings of the National Academy of
618 Sciences* **102**, 15545 (2005).
- 619 58. Mootha, V.K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are
620 coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273 (2003).
- 621 59. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**,
622 1739-1740 (2011).
- 623 60. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
624 collection. *Cell Syst* **1**, 417-425 (2015).
- 625