

Clonal fitness inferred from time-series modelling of single-cell cancer genomes

<https://doi.org/10.1038/s41586-021-03648-3>

Received: 8 May 2020

Accepted: 17 May 2021

Published online: 23 June 2021

 Check for updates

Sohrab Salehi^{1,8}, Farhia Kabeer^{1,2,8}, Nicholas Ceglia³, Mirela Andronescu^{1,2}, Marc J. Williams³, Kieran R. Campbell⁴, Tehmina Masud¹, Beixi Wang¹, Justina Biele¹, Jazmine Brimhall¹, David Gee¹, Hakwoo Lee¹, Jerome Ting¹, Allen W. Zhang¹, Hoa Tran¹, Ciara O'Flanagan¹, Fatemeh Dorri^{1,5}, Nicole Rusk³, Teresa Ruiz de Algara¹, So Ra Lee¹, Brian Yu Chieh Cheng¹, Peter Eirew¹, Takako Kono¹, Jenifer Pham¹, Diljot Grewal³, Daniel Lai¹, Richard Moore⁶, Andrew J. Mungall⁶, Marco A. Marra⁶, IMAXT Consortium*, Andrew McPherson³, Alexandre Bouchard-Côté⁷, Samuel Aparicio^{1,2} & Sohrab P. Shah³

Progress in defining genomic fitness landscapes in cancer, especially those defined by copy number alterations (CNAs), has been impeded by lack of time-series single-cell sampling of polyclonal populations and temporal statistical models^{1–7}. Here we generated 42,000 genomes from multi-year time-series single-cell whole-genome sequencing of breast epithelium and primary triple-negative breast cancer (TNBC) patient-derived xenografts (PDXs), revealing the nature of CNA-defined clonal fitness dynamics induced by *TP53* mutation and cisplatin chemotherapy. Using a new Wright–Fisher population genetics model^{8,9} to infer clonal fitness, we found that *TP53* mutation alters the fitness landscape, reproducibly distributing fitness over a larger number of clones associated with distinct CNAs. Furthermore, in TNBC PDX models with mutated *TP53*, inferred fitness coefficients from CNA-based genotypes accurately forecast experimentally enforced clonal competition dynamics. Drug treatment in three long-term serially passaged TNBC PDXs resulted in cisplatin-resistant clones emerging from low-fitness phylogenetic lineages in the untreated setting. Conversely, high-fitness clones from treatment-naïve controls were eradicated, signalling an inversion of the fitness landscape. Finally, upon release of drug, selection pressure dynamics were reversed, indicating a fitness cost of treatment resistance. Together, our findings define clonal fitness linked to both CNA and therapeutic resistance in polyclonal tumours.

Quantifying cellular fitness and its causal mechanisms in heterogeneous, polyclonal cancer cell populations remain unresolved problems, impeding progress in developing effective and durable therapeutic strategies^{1–7}. Despite well-documented genomic plasticity in tumours, the question of how CNA-induced changes in the genome architecture drive aetiological and drug resistance¹⁰ processes remains understudied^{11–13}. The cancer field has generally lacked serial measurements from patient-derived tissues to directly observe cancer evolution over realistic timescales with single-cell resolution^{1,2,4,14–21}. This has hindered the thorough investigation of causal factors driving selection, unlike in other biological systems²². Here we use single-genome-derived CNAs as clone-defining heritable genotypes to establish quantitative fitness attributes that serve as predictive measures of polyclonal growth potential. Our work has implications in at least three areas: predicting evolution in cancer; understanding how genomic instability processes leading to CNAs confer fitness; and parsing long-term kinetics of drug resistance in polyclonal cancer cell populations.

Modelling clonal fitness and selection

We developed an experimental and computational platform consisting of three components: time-series sampling and single-cell whole-genome sequencing (scWGS) of immortal cell lines and PDX (Extended Data Fig. 1a, b); scalable phylogenetics for single-cell genomes (sitka²³, Extended Data Fig. 1c); and a population genetics inspired (Wright–Fisher diffusion process) model of fitness (fitClone, Extended Data Fig. 1d, e, Supplementary Table 1). Using observed longitudinal clonal-abundance measurements as input, fitClone simultaneously estimates the growth trajectory (Z_i) and fitness coefficient (s_i) for each clone i in the population. The model can be used to forecast evolutionary trajectories, and its posterior probability densities can reflect evidence of positive selection in polyclonal systems. Details of fitClone, including theoretical assumptions and limitations of the model are discussed in the Supplementary Information.

¹Department of Molecular Oncology, BC Cancer, Vancouver, British Columbia, Canada. ²Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. ³Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁴Lunenfeld-Tanenbaum Research Institute Mount Sinai Hospital Joseph & Wolf Lebovic Health Complex, Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. ⁵Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. ⁶Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada. ⁷Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. ⁸These authors contributed equally: Sohrab Salehi, Farhia Kabeer. *A list of authors and their affiliations appears at the end of the paper.  e-mail: saparicio@bccrc.ca; shahs3@mskcc.org

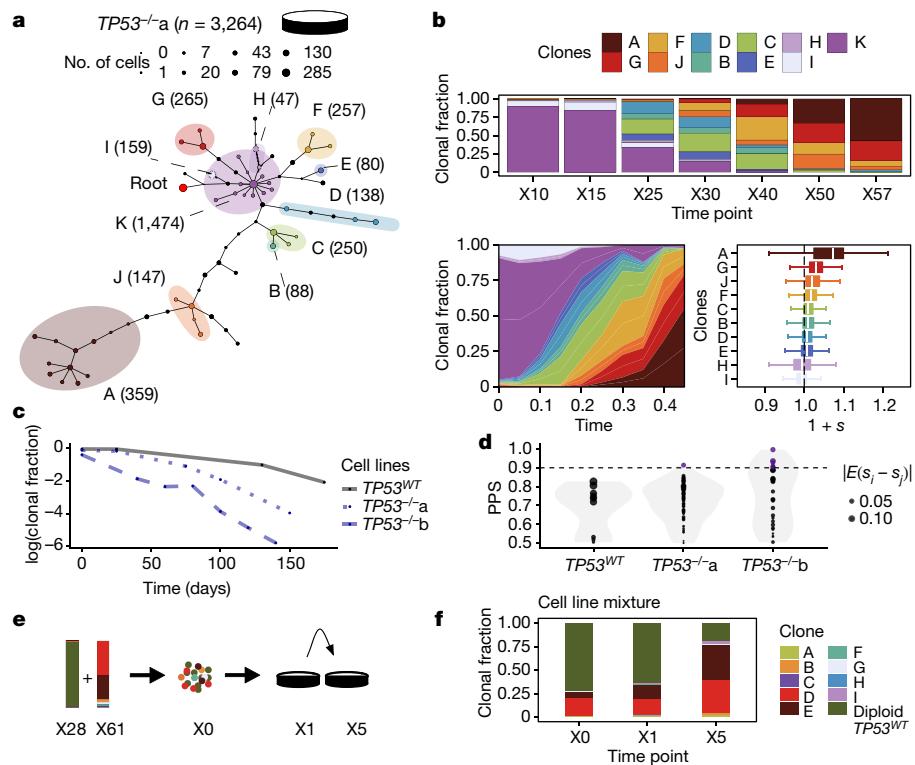


Fig. 1 | Replicate branch of p53 mutant cells and engineered mixture experiment. **a**, Phylogeny of 3,264 *TP53*^{-/-} a cells, grouped in 11 phylogenetic clades over the time series, where nodes are groups of cells with shared copy number genotype (sized proportional to the number of cells in the node) and edges represent distinct genomic breakpoints. Shaded areas represent the indicated clones, with numbers of cells in parentheses. The root of the tree is denoted by the red circle. **b**, Observed clonal fractions over time, inferred trajectories and quantiles of the posterior distributions over selection coefficients of fitClone model fits to *TP53*^{-/-} a with respect to the reference clone K. In the box plots, the white line represents the median of the

distribution, box edges show 1.5 \times the interquartile range and whiskers extend to 25th and 75th percentiles. **c**, Clonal fraction of the diploid reference over time. **d**, Distribution over the probability of positive selection (PPS) over pairs of clones computed as $\max(P(s_i > s_j), 1 - P(s_i > s_j))$. Purple dots denote a PPS above 0.9. Dot size indicates the magnitude of the expected value of the difference between selection coefficient of pairs of clones $|E(s_i - s_j)|$. **e**, Experiment in which 75% *TP53*^{WT} cells (time point X28) are cultured with 25% *TP53*^{-/-} b cells (time point X61) and examined at time points X1 and X5 relative to the start of the mixed culture. **f**, Observed clonal fractions from the mixture experiment in **e**.

CNAs and fitness in p53-deficient cells

We first applied the framework to immortalized 184-hTERT diploid breast epithelial cell lines²⁴ to measure clone-specific fitness associated with *TP53* loss of function. Known to be permissive of genomic instability, *TP53* mutations are often acquired early in breast cancer evolution^{4,25,26} and result in alteration of the CNA genome structure^{5,6,21,24,27}. We contrasted four time-series samples of wild-type *TP53* (*TP53*^{WT}) cells (60 passages over 300 days) with two isogenic null (*TP53*^{-/-}) parallel branches²⁸ (*TP53*^{-/-} a and *TP53*^{-/-} b), each passaged over 60 generations (285 and 220 days, respectively) and sampled 7 times. A median of 1,231 cells per passage was analysed by whole-genome sequencing, yielding a total of 2,713, 3,264 and 4,881 genomes for each time series, respectively (Supplementary Table 1). For each of *TP53*^{WT}, *TP53*^{-/-} a and *TP53*^{-/-} b, we inferred CNA profiles, constructed phylogenetic trees to establish clonal lineages (Methods) and measured clonal abundances over time. Phylogenetic analysis using sitka (Extended Data Fig. 2a, b) and modelling of abundances with fitClone (Extended Data Fig. 2c, Supplementary Tables 2, 3) revealed *TP53*^{WT} clonal trajectories consistent with small differences in selection coefficients amongst three major clones: E (chromosome 11q gain), D (chromosome 20 gain) and F (diploid; and used as the reference clone for fitClone modelling). By contrast, *TP53*^{-/-} a showed significant expansions of clones with aneuploid genotypes (Fig. 1a, Extended Data Fig. 2d) and higher selection coefficients, where the founder diploid population was out-competed (Fig. 1b). A second independent

TP53-mutant time series, *TP53*^{-/-} b (Extended Data Fig. 2e–g), confirmed that CNA-bearing clones confer higher fitness. *TP53*^{-/-} a and *TP53*^{-/-} b lines harboured 11 (size range 47 to 1,474 cells, median 204) and 10 (size range 158 to 997 cells, median 404) distinct clones, respectively (Supplementary Table 2). Notably, selection coefficients were highest in clones with focal amplifications of known prototypic oncogenes in breast cancer^{6,7,25,26} (Extended Data Fig. 2d, e), in some cases on a whole-genome-doubled background. Clone A, the highest-fitness clone in *TP53*^{-/-} a (57% of cells at the last time point, $1+s = 1.05 \pm 0.09$) exhibited whole-genome doubling and harboured a focal, high level amplification at the *MDM4* (1q) locus (Extended Data Fig. 2d). Clone G (27% of cells at the last time point, $1+s = 1.03 \pm 0.03$), the next highest-fitness clone in *TP53*^{-/-} a remained diploid, with the exception of a focal high-level amplification precisely at the *MYC* locus (8q) (Extended Data Fig. 2d). By contrast clone K, chosen here as the reference clone for modelling, remained entirely diploid and exhibited a monotonically decreasing trajectory (from 90% to 0% of cells over the timeseries, Fig. 1b). In *TP53*^{-/-} b, two clones exhibited large positive selection coefficients (Extended Data Fig. 2f, g): clone D (52% of cells at last time point, $1+s = 1.05 \pm 0.02$) harboured a 20q single-copy gain with an additional high level amplification at the *TSHZ2* locus; and clone E (35% of cells at the last time point, $1+s = 1.05 \pm 0.04$) harboured a chromosome 4 loss, 19pgain, 19qloss and a 20q single-copy gain (Extended Data Fig. 2e). As seen in *TP53*^{-/-} a, the ‘root’ clone I that remained diploid was systematically outcompeted, diminishing in abundance from 68% to 0% over the time series (Extended Data Fig. 2g).

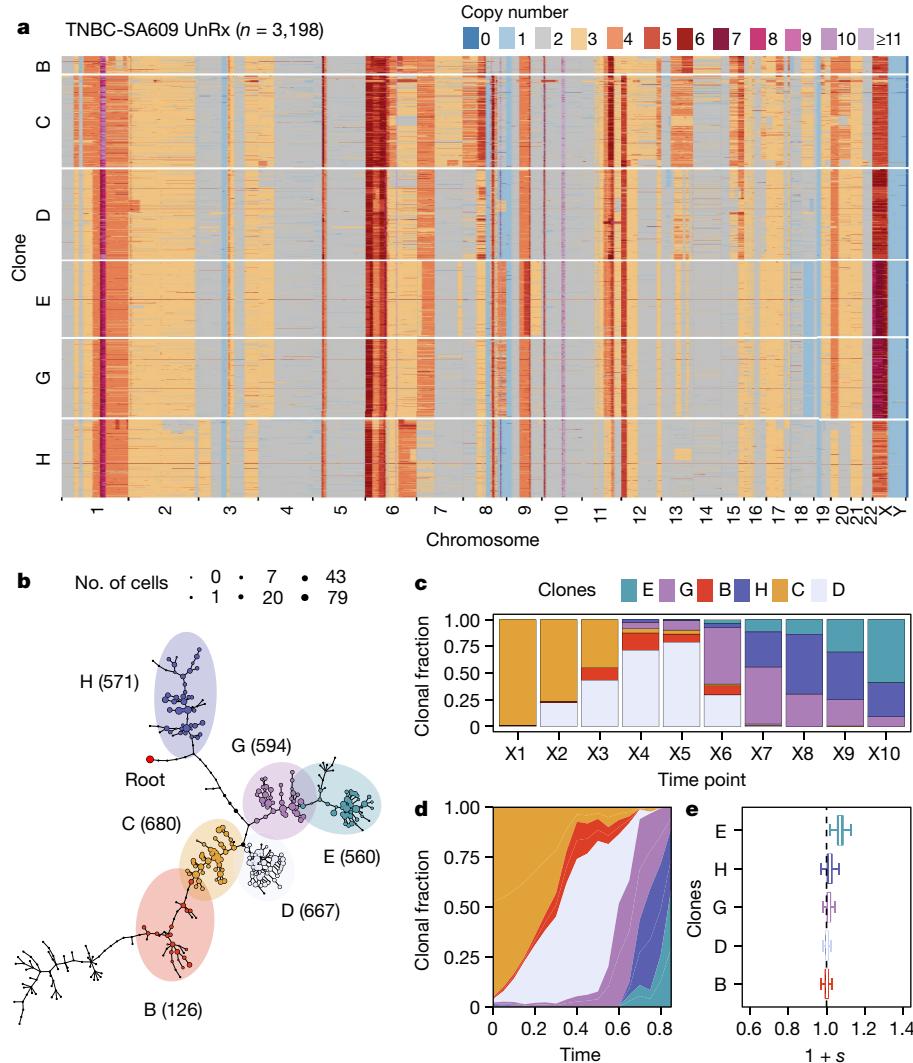


Fig. 2 | Fitness landscapes of untreated TNBC-SA609 PDX. **a**, Heat map representation of copy-number profiles of 3,198 cells from TNBC-SA609 PDX, grouped in 6 phylogenetic clades. **b–e**, Phylogeny (**b**), observed clonal

fractions (**c**), inferred fitClone trajectories (**d**) and quantiles of the selection coefficients (**e**) for TNBC-SA609 PDX. Box plots are as defined in Fig. 1b.

Relative to *TP53*^{WT}, rates of expansion of both *TP53*^{-/-} and *TP53*^{+/+} aneuploid clones were significantly higher, leading to rapid depletion of diploid cells (Fig. 1c, $P = 6.72 \times 10^{-4}$). Copy number breakpoints per cell increased as a function of fitness and were higher in the *TP53*-mutant lines; however, the point mutation rate remained comparatively stable (Extended Data Fig. 2h, i). In addition, both *TP53*-mutant lines exhibited higher posterior probability of positive selection (clone pairs with probability greater than 0.9) relative to the *TP53*^{WT} setting (Fig. 1d). Accordingly, we sought to experimentally confirm clonal fitness in *TP53*-mutant cells associated with increased aneuploidy. We challenged higher fitness aneuploid clones (D and E, which dominated by passage X60) from *TP53*^{-/-} with *TP53*^{WT} diploid populations in de novo population mixtures and collected samples over five generations in culture (Fig. 1e). scWGS of the mixture samples revealed that clones D and E monotonically increased from 18% to 35% and 7% to 37% of the population by passage 5, respectively, while the *TP53*^{WT} cells were out-competed, decreasing from 75% to 19% of the population at the final passage (Fig. 1f). Thus, the enforced competition resulted in the re-emergence of *TP53*-mutant aneuploid clones and relative depletion of diploid cells, supporting the original fitClone model fits. Together, these results show a broader clone fitness landscape, with overall higher fitness of clones harbouring whole genome, chromosomal

and segmental aneuploidies arising in *TP53*-mutant cells (Extended Data Fig. 2d, e). Notably, high fitness clones featured high level amplification of proto-oncogenes often seen in human breast cancer (for example, *MDM4*, *MYC* or *TSHZ2*), suggesting that p53 loss is permissive of fitness-enhancing CNAs with aetiological roles in cancer⁷.

Modelling fitness in human breast cancer

We next studied time-series CNA clonal expansions of *TP53*-mutant primary human breast cancers from four PDX transplant series. We generated serial scWGS samples from one HER2-positive (HER2⁺ SA532) and three TNBC (TNBC-SA609, TNBC-SA1035 and TNBC-SA535) PDX models, sampled over 927, 619, 381 and 353 days, respectively (Extended Data Fig. 3), yielding a median of 303 high-quality genomes per sample (9,970 in total) for downstream analysis (Supplementary Table 1). All series exhibited progressively higher tumour growth rates over time (Extended Data Fig. 4b–d) and maintained hormone receptor status from early to late passages (Supplementary Tables 4, 5). Bulk whole-genome sequencing and scWGS confirmed that all four tumours harboured *TP53* mutations with bi-allelic and truncal distribution across clones (Supplementary Table 6). Phylogenetic analysis indicated that all of the PDX models were polyclonal at the CNA level

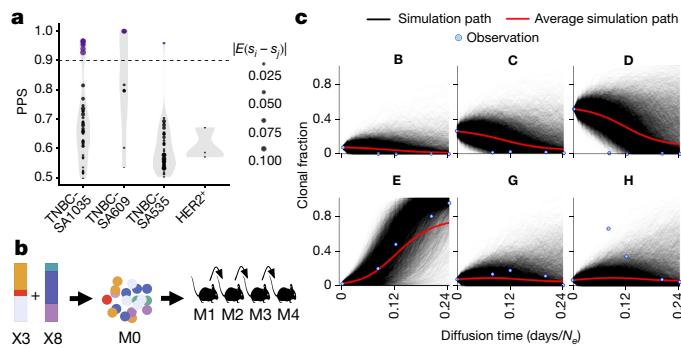


Fig. 3 | Positive selection in untreated TNBC PDX. **a**, Distribution over the PPS over pairs of clones. **b**, Clonal fractions of TNBC-SA609 line 1 at X3 and X8 were mixed to generate the initial mixture M0 for transplantation and subsequent serial passaging in mice, yielding four samples. **c**, Forward simulations from the original time series and starting population proportions in M0. Simulated trajectories are shown superimposed with mean simulation (red line) and observed clonal fractions (blue dots). The observation time is adjusted to match the simulation diffusion time. N_e , effective population size.

(Fig. 2, Extended Data Fig. 4a–i), with *in cis* gene expression effects inferred from single-cell RNA-sequencing data derived from the same single-cell suspensions (Methods, Supplementary Table 7).

In contrast to the HER2⁺SA532 model, the TNBCPDX models exhibited evidence of clonal dynamics and variation in selection coefficients consistent with positive selection and differential fitness (Fig. 3a). Eleven clones were detected for TNBC-SA1035 (Extended Data Figs. 4a, b, 5); the reference clone A had an initial prevalence of 20% but was not detectable by the last time point. Clone E, which expanded to 69% at passage X8 from minor prevalence at the initial time point ($1+s=1.06\pm 0.03$) (Extended Data Fig. 4c, d, Supplementary Table 3), formed a distinct clade, distinguished by a hemizygous deletion of the centromeric locus of 8p, gain of an extra copy of the telomeric end of 11q and a focal gain of 19q12 harbouring the *CCNE1* locus (Extended Data Figs. 4a, 5a). In TNBC-SA535, three out of ten clones propagated after the initial time point (Extended Data Fig. 4d–f). Clone G, characterized by loss of chromosome X, exhibited expansion from minor prevalence at passage X5 to 76% at passage X9 ($1+s=1.02\pm 0.01$, Extended Data Fig. 4f). For TNBC-SA609 line 1, six clones were observed (Fig. 2a, b). Clones E ($1+s=1.07\pm 0.02$) and H ($1+s=1.02\pm 0.02$) had the highest selection coefficients, with growth increasing from undetectable levels to 59% and 32%, respectively, by time point X10. Clone C contracted from nearly 100% at the initial time point to undetectable levels by X10 (Fig. 2c–e). Growth of clones E (Extended Data Fig. 6a), G and H and contraction of clone C (Extended Data Fig. 6b) was observed reproducibly in replicate transplants (Extended Data Fig. 6c–e). Notably, clones in the HER2⁺ series exhibited a maximum probability of positive selection of 0.67, suggesting overall clonal selection close to neutral (Extended Data Fig. 4i). By contrast, in all three TNBC series, at least one clone showed probability of positive selection greater than 0.9 (Fig. 3a).

Forecasting clonal trajectories

Next, we experimentally validated the fitness coefficients as indicators of positive or negative selection. We carried out forward simulations from fitClone using selection coefficients estimated from the original time series, and compared these with serially passaged physical clonal mixtures of late (X8) and early (X3) time points from TNBC-SA609 (line 1). Two mixture–retransplant–serial-passage experiments were conducted with different initial starting conditions (Fig. 3b, Extended Data Fig. 7a). In the first experiment, clone E was forecast to fixate with the highest probability (0.39) and in the second experiment,

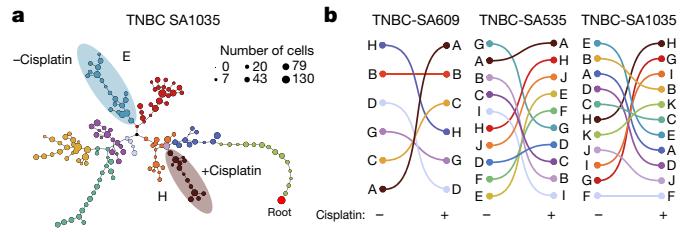


Fig. 4 | Fitness landscape reversal in early cisplatin treatment in TNBC PDX models. **a**, Phylogenetic tree for TNBC-SA1035 annotated with fittest clones with and without cisplatin treatment. **b**, Inversion of the fitness landscape. Clones are ranked according to their median selection coefficients in untreated and treated conditions, with the top-ranked clones highest.

clones E and H were forecast to fixate with high probability (0.14 and 0.04, respectively) (Fig. 3c, Extended Data Fig. 7b). The two series were then sequenced using scWGS, yielding 6,453 and 6,730 genomes, respectively. In the first mixture (Extended Data Fig. 7c), we recovered six clones from the original time series, with between 26 and 767 (median 155) cells. As anticipated by the model, clone E emerged as a high-fitness clone ($1+s=1.08\pm 0.03$), and by the last time point, clones E and H had swept through to make up 94% of cells. In the second mixture (Extended Data Fig. 7d), we recovered four clones (C, E, G and H) from the original time series. Clone E was the only clone that increased in prevalence (from 5% to 24%) and had the highest selection coefficient ($1+s=1.02\pm 0.03$). By contrast, clones C, G and H exhibited relatively stable prevalences (Extended Data Fig. 7d). Thus, both mixture experiments demonstrated expansion of the predicted highest-fitness clone (E), even when starting from low initial proportions (2% or 5% of cells).

The fitness cost of platinum resistance

Using CNA clone-specific fitness measurements, we next tested how drug treatment with cisplatin (standard therapy for primary TNBC) perturbs the fitness landscape of the three PDX series. For each time series, we propagated a separate branch treated with cisplatin (Methods, Extended Data Figs. 1b, 3) to induce gradual onset of resistance to platinum-based chemotherapy (platinum resistance), physically confirmed with a progressive reduction in tumour growth inhibition²⁹ (per cent tumour growth inhibition from first to last cycle: TNBC-SA609, 77% to 4.7%; TNBC-SA1035, 76% to 15%; TNBC-SA535, 58% to 16%; Extended Data Fig. 3b–d). For TNBC-SA609 a total of five independent transplant lineages were resurveyed with technical replicates for lines 1 and 2 (Methods). In each series, emergent clones following treatment were distinct in phylogenetic origin from those with high fitness in the untreated setting, indicating an inversion of the clone fitness landscapes (Fig. 4a, Extended Data Fig. 8a). Suppression of high-fitness clones that dominated in the absence of treatment and expansion of low-fitness and/or previously unobserved genotypes led to a substantially altered rank order of selection coefficients in treated samples relative to untreated samples (Fig. 4b, Extended Data Fig. 8b, c). Samples were coded U or T for previous and present passages depending on whether they had received no treatment or treatment, respectively, during that passage. For TNBC-SA609 line 2, growth dynamics at X3 (U), X4 (UT), X5 (UTT), X6 (UTT) and X7 (UTTT) showed expansion of clone B and its derivative clones (A and R), from a starting population comprising primarily clones C, D and B (Extended Data Fig. 9) in three replicate transplants. Notably, resistant clones in all three replicate treated lines were phylogenetically distinct from clone H—the highest-fitness clone in the treatment-naïve setting (Extended Data Figs. 8a–c (left), 9a, e). The other two TNBC series also supported a fitness inversion, exhibiting monotonically decreasing prevalence of treatment-naïve high-fitness clones and increasing prevalence of low-fitness clones. Of

note, in both TNBC-SA535 and TNBC-SA1035, both high- and low-fitness clones were observed in initial conditions, ruling out sampling bias as a strict determinant of selection dynamics (Extended Data Figs. 5, 10). Specifically, the numbers of low-fitness clones in the untreated series all increased, with the fittest clones under no treatment decreasing to near zero prevalence (for example, TNBC-SA535 clone G (Extended Data Fig. 10a) and TNBC-SA1035 clone E). The probability of positive selection increased in the treatment series (Extended Data Fig. 8d), indicating that there were more clones under positive selection in the cisplatin setting, and selection coefficients exhibited a wider variance between clones.

Finally, we tested the effect of lifting the drug selective pressure at each time point, with drug holiday replicate transplants (Extended Data Figs. 1b, 3a–e). In TNBC-SA609, inverted fitness was reversible in a short interval (Extended Data Fig. 9). In the first drug holiday, X5-UTU, clonal composition reverted to consist predominantly of precursor clone B with 90% abundance, and only 10% abundance from clone A (Extended Data Fig. 9c–e). However, clone A comprised more than 99% of X6-UTTU and X7-UTTTU, similar to their on-treatment analogues, and no reversion was detected. Thus, when clonal competition was possible in the absence of drug, cells derived from the precursor B clade outcompeted clone A, indicating that clone-specific cisplatin resistance has a fitness cost. Moreover, the specificity of reversion between X4-UT and X5-UTU reflects selection of predefined clones with differential fitness. The TNBC-SA1035 series exhibited more moderate reversibility. Clone G growth was attenuated from 10% at X5-UT to 9% at X6-UTU, compared to 20% at X6-UTT (Extended Data Fig. 5b–e). Similarly, in TNBC-SA535 (Extended Data Fig. 10c–e) growth attenuation of the highest-fitness clone (A) (Extended Data Fig. 10b) in the treatment setting was observed in the holiday setting, and clone E exhibited clonal fractions similar to treated time points at X7-UTU, X8-UTTU, and X9-UTTTU. However, clone E increased to 32% of cells in X10-UTTTTU from 10% in X9-UTTTT. Thus, in all series, treatment selective pressure was reversible with drug holidays, consistent with a fitness cost of platinum resistance.

Discussion

In population genetics, the repeated observation of dominance or decline in clones, defined here by CNA genotypes, implies that either the genotype or a factor heritably linked to the genotype (such as single-nucleotide variants or epigenetic states) is a determinant of fitness. We expect that additional variation due to single-nucleotide variants, structural variations from genomic rearrangements or rare CNAs beyond first-approximation estimates will additively affect fitness. However, deeper population sampling would be required to appropriately capture these effects. Our results, decoded by measuring single-cell time series, suggest that fitness linked to CNAs may be under-appreciated. This has implications for interpreting aetiological processes of tumour-suppressor-driven cancers, shown by inducing *TP53* loss, where rates of structural variation acquisition and deviation away from diploid configurations conferred fitness advantages. Over successive generations in vitro with *TP53* mutational perturbation and in three TNBC PDX lines in the context of cisplatin drug treatment, emergent CNA measurably contributed to the fitness landscape, consistent with a continuously diversifying mechanism that induces competitive clonal advantages.

Our results demonstrate that time-series fitness mapping is a realistic initial approach for studying how the effect of driver mutations inducing genomic instability leads to clonal expansions and evolutionary selection. The ability to genetically manipulate the systems we describe provides a future path to mechanistically dissecting fitness effects of individual CNA regions. Furthermore, as the impact of drug intervention on CNA-driven cancer evolution is a key determinant of patient outcomes across all human cancers^{20,21,30–32}, forecasting the

trajectories of cancer clones is of immediate importance to understanding therapeutic response in cancer, and for deploying adaptive approaches¹⁰. We suggest that the presence of resistant genotypes that have a chemotherapeutic fitness cost may define time windows within which clonal competition could be exploited. Future investigations in patients with time-series tumour or cell-free DNA-based population genetics modelling may therefore enable evolution-informed approaches to clinical management³³.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-021-03648-3>

- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
- Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
- Salichos, L., Meyerson, W., Warrell, J. & Gerstein, M. Estimating growth patterns and driver effects in tumor evolution from individual samples. *Nat. Commun.* **11**, 732 (2020).
- Shah, S. P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
- Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
- Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
- Wright, S. The distribution of gene frequencies in populations. *Proc. Natl Acad. Sci. USA* **23**, 307–320 (1937).
- Tataru, P., Simonsen, M., Bataillon, T. & Hobolth, A. Statistical inference in the Wright–Fisher model using allele frequency data. *Syst. Biol.* **66**, e30–e46 (2017).
- Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. *Nature* **575**, 299–309 (2019).
- Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).
- Sunshine, A. B. et al. The fitness consequences of aneuploidy are driven by condition-dependent gene effects. *PLoS Biol.* **13**, e1002155 (2015).
- Sheltzer, J. M., Torres, E. M., Dunham, M. J. & Amon, A. Transcriptional consequences of aneuploidy. *Proc. Natl Acad. Sci. USA* **109**, 12644–12649 (2012).
- Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
- Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
- Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- Khan, K. H. et al. Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial. *Cancer Discov.* **8**, 1270–1285 (2018).
- Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
- Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
- López, S. et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
- McPherson, A. et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48**, 758–767 (2016).
- Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
- Dorri, F. et al. Efficient Bayesian inference of phylogenetic trees from large scale, low-depth genome-wide single-cell data. Preprint at <https://doi.org/10.1101/2020.05.06.058180> (2020).
- Burleigh, A. et al. A co-culture genome-wide RNAi screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. *Breast Cancer Res.* **17**, 4 (2015).
- de la Vega, M. R., Chapman, E. & Zhang, D. D. NRF2 and the Hallmarks of cancer. *Cancer Cell* **34**, 21–43 (2018).
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Patch, A.-M. et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
- Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221 (2019).
- Hather, G. et al. Growth rate analysis and efficient experimental design for tumor xenograft studies. *Cancer Inform.* **13**, 65–72 (2014).
- Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
- Bakhour, S. F. et al. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).

Article

32. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
33. Acar, A. et al. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer. *Nat. Commun.* **11**, 1923 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

IMAXT Consortium

Gregory J. Hannon⁹, Giorgia Battistoni⁹, Dario Bressan⁹, Ian Gordon Cannell⁹, Hannah Casbolt⁹, Atefeh Fatemi⁹, Cristina Jauset⁹, Tatjana Kovačević⁹, Claire M. Mulvey⁹, Fiona Nugent⁹, Marta Paez Ribes⁹, Isabella Pearsall¹⁰, Fatime Qosaj⁹, Kirsty Sawicka⁹, Sophia A. Wild⁹, Elena Williams⁹, Samuel Aparicio¹², Emma Laks¹², Yangguang Li¹, Ciara H. O'Flanagan¹, Austin Smith¹, Teresa Ruiz¹, Daniel Lai^{1,2}, Andrew Roth^{2,4}, Shankar Balasubramanian^{9,10,11}, Maximillian Lee^{9,10}, Bernd Bodenmiller¹², Marcel Burger¹², Laura Kuett¹², Sandra Tietscher¹², Jonas Windhager¹², Edward S. Boyden¹³, Shahar Alon¹³, Yi Cui¹³, Amauche Emenari¹³, Dan Goodwin¹³, Emmanouil D. Karagiannis¹³, Anubhav Sinha¹³, Asmamaw T. Wassie¹³, Carlos Caldas¹⁴, Alejandra Bruna¹⁴, Maurizio Callari⁹,

Wendy Greenwood⁹, Giulia Lerda⁹, Yaniv Eyal-Lubling¹⁴, Oscar M. Rueda¹⁴, Abigail Shea¹⁴, Owen Harris¹⁵, Robby Becker¹⁵, Flaminia Grimaldi¹⁵, Sudi Harris¹⁵, Sara Lisa Vogl¹⁵, Joanna Weselak¹⁵, Johanna A. Joyce¹⁶, Spencer S. Watson¹⁶, Sohrab P. Shah³, Andrew McPherson³, Ignacio Vázquez-García³, Simon Tavaré^{9,17,18}, Khanh N. Dinh¹⁷, Eyal Fisher⁹, Russell Kunes¹⁷, Nicholas A. Walton¹⁹, Mohammad Al Sa'd¹⁹, Nick Chornay¹⁹, Ali Dariush¹⁹, Eduardo A. González-Solares¹⁹, Carlos González-Fernández¹⁹, Aybüke Küpcü Yoldas¹⁹, Neil Millar¹⁹, Tristan Whitmarsh¹⁹, Xiaowei Zhuang^{20,21,22}, Jean Fan^{20,21,22}, Hsuan Lee^{20,21,22}, Leonardo A. Sepúlveda^{20,21,22}, Chenglong Xia^{20,21,22} & Pu Zheng^{20,21,22}

⁹Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. ¹⁰Department of Chemistry, University of Cambridge, Cambridge, UK.

¹¹School of Clinical Medicine, University of Cambridge, Cambridge, UK. ¹²Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. ¹³McGovern Institute, Departments of Biological Engineering and Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹⁴Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. ¹⁵Súil Interactive Ltd, Dublin, Ireland. ¹⁶Department of Oncology and Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland. ¹⁷Herbert and Florence Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA. ¹⁸New York Genome Center, New York, NY, USA. ¹⁹Institute of Astronomy, University of Cambridge, Cambridge, UK. ²⁰Howard Hughes Medical Institute, Harvard University, Cambridge, MA, USA. ²¹Department of Physics, Harvard University, Cambridge, MA, USA. ²²Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

All methods are detailed in the Supplementary Information. We studied normal human breast epithelial cells²⁴ in vitro and in breast cancer PDX, sequencing >129,500 whole genomes from single cells over interval passaging (scWGS DLP⁺ method²⁸; Extended Data Fig. 1 a, b, Supplementary Table 1). After read-coverage based quality control and omission of replicating cells, we retained >42,000 genomes from 113 libraries across cell lines and PDX transplants for analysis (average 995,000 reads/cell, 0.022x coverage). We calculated phylogenetic trees over cells to identify genotypic clones and their relative abundances as a function of time.

Human mammary cell lines and serial passaging and mixtures

The human mammary epithelial cell line 184-hTERT wild type and isogenic 184-hTERT-P53 KO cell line, generated from 184hTERT WT-L9, were grown as previously described^{24,28}. *TP53* was knocked out by using CRISPR–Cas9 technology and one clone (99.25) was serially passaged to further subdivided at passage 10 into ‘branch a’ and ‘branch b’ parallel knockout branches (NM_000546(TP53):c.[156delA];[156delA]), p.(Gln52Hisfs*71), and the absence of P53 protein was confirmed with western blot (Supplementary information). Two branches of 184-hTERT-*TP53*^{-/-} (clone 95.22) along with the counterpart wild-type branch were serially passaged over ~55–60 generations, by seeding ~1 million cells into a new 10 cm tissue culture treated dish (Falcon CABD353003) and cryopreserving every fifth passage. Mammary Epithelial Cell SingleQuot Kit Supplements (MEGTM), Growth Factors (Lonza CC-4136), with 5 µg ml⁻¹ transferrin (Sigma) and 2.5 µg ml⁻¹ isoproterenol (Sigma) were used as a growth medium as previously described²⁴. Cells were grown to around 85–90% confluence, trypsinized for 2 min (Trypsin/EDTA 0.25%, VWR CA45000-664), re-suspended in cryopreservation medium (10% DMSO-Sigma-D2650, 40% FBS-GE Healthcare SH30088.03, 50% media) and frozen to -80 °C at a rate of -1 °C min⁻¹. Cells were cultured continuously from passage 10 (post initial cloning²⁴) to passage 60 for 184-hTERT WT and upto passage 57 and passage 55 for the *TP53*^{-/-} branches a and b, respectively, from initial cloning/isolation. Genome sequencing was undertaken at passages 25, 30, 51 and 60 from the wild type branch, passages 10, 15, 25, 30, 40, 50 and 57 from *TP53*^{-/-} branch a and passages 20, 30, 35, 40, 45, 50 and 55 from *TP53*^{-/-} branch b. Also, the transcriptome sequencing was carried out on passages 11 and 57 of *TP53*^{-/-} branch a and passages 15, 30 and 50 of *TP53*^{-/-} branch b. All cell line cultures were tested negative for mycoplasma by PCR Mycoplasma contamination detection test (Genetica Cell Line Testing). The initial conditions of the mixtures were biased in favour of diploid cell populations in a 3:1 ratio of 184-hTERT WT p28 (SA039) and *TP53*^{-/-} clone 95.22 (SA906b) passage 61. Prior to plating in the culture, an aliquot was analysed by DLP⁺ to measure the baseline clonal composition labelled as X0, and 2,701 single cell genomes were generated. At 80% confluence on the plate, cells from X1 were harvested and serially passaged up to the 20th passage. scWGS data from three time points X0, X1 and X5 with a median of 898 cells per time point were collected.

Establishment and serial passaging of patient derived xenografts

The Ethics Committees at the University of British Columbia approved all the experiments using human resources. Donors in Vancouver, British Columbia were recruited, and samples were collected under the tumour tissue repository (TTR-H06-00289, H16-01625) and transplanted in mice under the animal resource centre (ARC) bioethics protocol (A19-0298-A001) approved by the animal care committee (University of British Columbia) and BC Cancer Research Ethics Board

H20-00170 and H18-01113 protocols. After informed consent, tumour fragments from patients undergoing excision or diagnostic core biopsy were collected. Tumour materials were processed as described in ref.³⁴ and transplanted in 8- to 12-week-old female mice approved by the animal care committee. In brief, tumour fragments were chopped finely with scalpels and mechanically disaggregated for 1 min using a Stomacher 80 Biomaster (Seward) 1 ml to 2 ml cold DMEM/F-12 with glucose, L-glutamine and HEPES (Lonza 12-719F). An aliquot of 200 µl of medium (containing cells/clumps) from the resulting suspension was used equally for 4 transplantations in mice. Tumours were transplanted subcutaneously in mice as previously described³⁴ in accordance with SOP BCCRC 009.

Serial passaging of PDX

Tumours were serially passaged as previously described³⁴. In brief, for serial passaging of PDX, xenograft-bearing mice were euthanized when the size of the tumours approached 1,000 mm³ in volume (combining together the sizes of individual tumours when more than one was present). The tumour material was excised aseptically, and processed as described for primary tumour. In brief, the tumour was harvested and minced finely with scalpels then mechanically disaggregated for one minute using a Stomacher 80 Biomaster (Seward) in 1 ml to 2 ml cold DMEM-F12 medium with glucose, L-glutamine and HEPES. Aliquots from the resulting suspension of cells and fragments were used for xenotransplants in the next generation of mice and cryopreserved. Serially transplanted aliquots represented approximately 0.1–0.3% of the original tumour volume. HER⁺ SA532 and TNBC-SA609 PDX were passaged upto 10 generations and scDNaseq was carried out at each time point. The other three untreated and treated PDX time series were generated in the same way for 4–5 passages.

TNBC PDX tumour-mixing experiments

Frozen vials from the untreated TNBC PDX passages three (X3) and eight (X8), were thawed and physically remixed in two different volumetric proportions of X3:X8 by tumour weight. The ratio of approximately 1:1 and 1:0.4, labelled as mixture branch a and branch b, respectively. From each of different dilutions, 200 µl of aliquot was transplanted in two mice each using the same protocol described above. Before transplantation, a small proportion of the physical mixture of cells from the 1:1 ratio, was subjected to whole genome single cell sequencing to measure the baseline clonal composition labelled as M0 and its subsequent PDX as M1. The tumour cell mixture was then serially passaged over 4 generations for branch a and 5 generations for branch b, designating the transplants as M1–M4 and M1–M5, respectively. Tumours from serial passages (X3:X8) from both mixtures branches were collected and analysed with scWGS (DLP⁺) as for other samples.

TNBC PDX time-series treatment with cisplatin

Female NRG mice of 8–12 weeks of age and genotype were used for randomized controlled transplantation treatment experiments. Drug treatment with cisplatin, an analogue of platinum salts was commenced when the tumour size reached approximately 300–400 mm³. Cisplatin (Accord DIN: 02355183) was administered intraperitoneally at 2 mg kg⁻¹ every third day for 8 doses maximum (Q3Dx8). The dosage schedule was adjusted to 50% the dose used in the literature^{35,36} and around one third of the maximum tolerated dose (MTD) calculated in immunodeficient female mice of 8–12 weeks of age (Supplementary Information). Low-dose cisplatin pulse and tumour collection timings were optimized to achieve the experimental aims of tumour resistance. The aim was to collect the tumour at around 50% shrinkage (from the starting tumour at the time treatment started) in size when measured with a caliper. Cisplatin (1 mg ml⁻¹) was diluted in 0.9% NaCl to obtain concentrations 200 µl per 20 g of mouse weight and kept in glass vials at room temperature. Quality control drug samples were prepared freshly on each day before the dosing. For all three TNBC PDX, 8 female mice at initial passage were

Article

transplanted in parallel for the treatment per treatment holiday study group. The choice of sample size was made to get at least 3 mice in the treatment group. Half of the mice were treated with cisplatin when tumours exhibited ~50% shrinkage, the residual tumour was collected as above and re-transplanted for the next passage in the group of eight mice. Again, half of the mice at X5 were randomly kept untreated while the other half were exposed to cisplatin following the same dosing strategy. Four cycles of cisplatin treatment were generated, with a parallel drug holiday group at each passage. Cisplatin treated tumours were coded as UT, UTT, UTTT, UTTTT for each of the four cycles of drug respectively, while the tumours on drug holiday were labelled as UTU, UTTU and UTTTU for the three time points. The number of Ts in the coded label shows the number of cycles of drug exposure. scWGS and single-cell RNA sequencing (scRNA-seq) was carried out on each tumour during the time-series treatment with counterpart drug holiday and untreated controls (Extended Data Fig. 3a). In particular, TNBC-SA609 PDX was processed to establish 5 independent lines to explore the biological and technical replicate tumours as well as treatment series. All five lines from TNBC-SA609 were passaged identically after initial establishment. Line 1 untreated samples were seeded (X3 to X4) from a freshly dissociated tumour, whereas 4 other lines (all treated and line 2 untreated) were seeded (X3 to X4) from a frozen vial of tumour. Technical replicates were collected and sequenced for lines 1 and 2.

PDX tumour growth measurement curves

NRG mice received sub-cutaneous inoculation of tumour cells (150 µl) on day 0. The tumours were allowed to grow to palpable solid nodules. Around 7–9 days after they were palpable, their size was measured with calipers every 3rd day. Tumours were measured in two dimensions using a digital caliper and expressed as tumour volume in mm³; defined as: [volume = 0.52 × (length) × (width) × (width)]. Record of patient derived xenografts 10 generations time series, HER2⁺ SA532 and TNBC-SA609 exhibited progressively higher tumour growth rates in later passages (Extended Data Fig. 3b–d) and took fewer days to reach the humane end point. Tumour growth inhibition percentage range²⁹ was defined as: [1 – (mean volume of treated tumours)/(mean volume of control tumours) × 100%].

scWGS and library construction with DLP⁺

All libraries, including metrics on number of cells, average number of reads per cell and quality control metrics are listed in Supplementary Table 1. Tumour fragments from PDX samples were incubated with collagenase/hyaluronidase, 1:10 (10X) enzyme mix (Stem Cell Technologies, 07912) in 5 ml DMEM/F-12 with glucose, L-glutamine and HEPES (Lonza 12-719F) and 1% BSA (Sigma) at 37 °C. Intermittent gentle pipetting up and down was done every 30 min for 40–60 s, during the first hour with a wide bore pipette tip, and every 15–20 min for the second hour, followed by centrifugation (1,100 rpm, 5 min) and supernatant removal. The tissue pellet was resuspended in 1 ml of 0.25% trypsin-EDTA (VWR CA45000-664) for 1 min, superadded by 1 ml of DNase/dispase (100 µl/900 µl), (StemCell 07900,000082462) pipetted up and down for 2 min, followed by neutralization with 2% FBS in HBSS with 10 mM HEPES (Stem Cell Technologies, 37150). Undigested tissue was removed by passing through a 70 µM filter and centrifuged for 5 min at 1,100 rpm after topping it up to 5 ml with HBSS. The single cell pellet was resuspended in 0.04% BSA (Sigma) and PBS to achieve approximately 1 million cells per 500 µl for robot spotting for DLP⁺.

Robot spotting of single cells into the nanolitre wells and library construction

DLP⁺ library construction was carried out as described²⁸. In brief, single cell suspensions from cell lines and PDX were fluorescently stained using CellTrace CFSE (Life Technologies) and LIVE/DEAD Fixable Red Dead Cell Stain (ThermoFisher) in a PBS solution containing 0.04% BSA (Miltenyi Biotec 130-091-376) incubated at 37 °C for 20 min. Cells

were subsequently centrifuged to remove stain, and resuspended in fresh PBS with 0.04% BSA. This single cell suspension was loaded into a contactless piezoelectric dispenser (sciFLEXARRAYER S3, Scienion) and spotted into the open nanowell arrays (SmartChip, TakaraBio) preprinted with unique dual index sequencing primer pairs. Occupancy and cell state were confirmed by fluorescent imaging and wells were selected for single cell CN profiling using the DLP⁺ method²⁸. In brief, cell dispensing was followed by enzymatic and heat lysis. After lysis, fragmentation mix (14.335 nL TD Buffer, 3.5 nL TDE1, and 0.165 nL 10% Tween-20) in PCR water were dispensed into each well followed by incubation and neutralization. Final recovery and purification of single cell libraries was done after 8 cycles of PCR. Cleaned up pooled single-cell libraries were analysed using the Agilent Bioanalyzer 2100 HS kit. Libraries were sequenced at UBC Biomedical Research Centre (BRC) in Vancouver, British Columbia on the Illumina NextSeq 550 (mid- or high-output, paired-end 150-bp reads), or at the GSC on Illumina HiSeq2500 (paired-end 125-bp reads) and Illumina HiSeqX (paired-end 150-bp reads). The data was then processed to a quantification and statistical analysis pipeline²⁸.

Processing of cell lines and PDXs for scRNA-seq data

All libraries generated using 10x scRNA-seq are listed in Supplementary Table 8. Suspensions of 184-hTERT TP53^{WT} and TP53^{-/-} cells were fixed with 100% ice-cold methanol before preparation for scRNA-seq. Single cell suspensions were loaded onto the 10x Genomics single cell controller and libraries prepared according to the Chromium Single Cell 3' Reagent Chemistry kit standard protocol. Libraries were then sequenced on an Illumina Nextseq500/550 with 42bp paired-end reads, or a HiSeq2500 v4 with 125 bp paired-end reads. 10x Genomics Cell-Ranger, V3.0.2 (V3 chemistry), was used to perform demultiplexing, alignment and counting.

Viable frozen tumour clumps and fragments were incubated with digestion enzymes as with DLP⁺ single cells preparation (as above) and the cells were resuspended in 0.04% BSA in PBS. Dead cells were removed using the Miltenyi MACS Dead Cell Removal kit and cells were processed as previously described³⁷. To avoid processing artefacts and dissociation methods, the timings were tightly controlled between the samples. Library construction of the samples at the same time point was performed on the same chips. Library construction sample batch groupings are listed in Supplementary Table 7.

Phylogenetic tree inference, clone determination and clonal abundance measurements

We developed a single cell Bayesian tree reconstruction method based on copy number change point binary variables called sitka²³ to fit phylogenetic trees to the copy number profiles. In the output of sitka, cells are the terminal leaf nodes of the phylogenetic topology. The inferred trees were post-processed to identify clonal populations from major clades. With clonal populations defined, their abundances were counted as a function of time series and these were used for fitness inference (see below). Clones were constructed by identifying connected components (each a clade or a paraphyly) in the phylogenetic tree reconstruction. The tree was 'cut' into discrete populations according to the following procedure. The inputs to the algorithm are the rooted phylogenetic tree and the copy number states of its cells and the minimum and maximum allowed clone sizes. A clone is defined as connected components (each a clade or a paraphyly) in the graph tree composed of cell of sufficient genomic homogeneity. The degree of homogeneity can be tuned by limiting the number of loci and the difference in copy number of sub-clades in a clone. The algorithm works by first finding the coarse structure, that is dividing the tree into major clades and then looking for fine structures within each clade by traversing the tree in a bottom up manner and merging loci that are sufficiently similar. The remaining loci constitute the roots of detected clades. See the Supplementary Information for more details.

For the cell lines datasets, namely $TP53^{WT}$ and $TP53^{-/-}a$ and $TP53^{-/-}b$, we opted to also split clades by the ploidy of their constituent cells, where ploidy is defined as the most recurrent copy number state in the cell. Once clones are identified, we set the abundance of each clone at a specific time point as the fraction of cells in that clone from that time point. We note that for the data from whole-genome bulk sequencing³⁴, we used the following procedure to estimate clonal fractions: (1) let ν denote the mutational cellular prevalence (rows) estimated over multiple time points (columns) using the multi-sample PyClone³⁸ model; (2) define β as the genotype matrix (which mutation-cluster (rows) is present in which clones (columns)); (3) then we set $\beta\gamma = \nu$, where $\gamma = \beta^{-1}\nu$ are the clonal fractions over time; and (4) we solve for γ using QR-decomposition.

Fitness modelling

We describe in this section a Bayesian state-space model (fitClone) based on the Wright–Fisher⁸ diffusion with selection. For simulation studies see the Supplementary Information.

fitClone: a Bayesian fitness model for time-series data

We developed a Bayesian model and associated inference algorithm based on a diffusion approximation to K-allele Wright–Fisher model with selection.

We start with time-series clonal abundance measurements over a fixed number of clones and estimate two key unknown parameters of interest: fitness coefficients s_i for clone i which represents a quantitative measure of the growth potential of a given clone; and distributions over continuous-time trajectories, a latent (unobserved) population structure trajectory in ‘generational’ time. After briefly reviewing and setting notation for Wright–Fisher diffusions with selection, we introduce the Bayesian model we used to infer quantitative fitness of clones from time-series data. We then describe our posterior inference method and ancillary methods for effective population size estimation, and reference clone selection. A key difference of fitClone with methods that use a transformation of allele fractions to infer the existence of clones and focus on attempting to infer dynamics from bulk sequencing and single time points (for example, the method in ref.²) is that the inputs and outputs are fundamentally different, addressing fundamentally non-overlapping analytical problems. In particular, (1) fitClone models explicitly defined clones and their timecourse data, and (2) fitClone is a generative model which allows for forecasting and prediction. See ref.⁹ for more background on the Wright–Fisher model and refs.^{39–47} for previous work on inference algorithms for Wright–Fisher models.

Wright–Fisher diffusions with selection

Let K denote the number of clones obtained using the tree cutting procedure described above, and denote by $Z_t = (Z_t^1, \dots, Z_t^K)$ the relative abundance of each of the K clones at time t in the population. The process Z_t satisfies, for all t , the constraints $\sum_{i=1}^K Z_t^i = 1$ and $Z_t^i \geq 0$ for $i \in \{1, \dots, K\}$. We would like to model the process Z_t using a Wright–Fisher diffusion with selection.

A Wright–Fisher diffusion can be written in stochastic calculus notation as

$$dZ_t = \mu^{s, N_e}(Z_t)dt + \sigma(Z_t)dW_t, \quad (1)$$

where $\{W_t\}$ is a K -dimensional Brownian motion, and the functions μ and σ , defined below, respectively control the deterministic and stochastic aspects of the dynamics.

For $z = (z^1, z^2, \dots, z^K)$, the vector-valued function $\mu^{s, N_e} : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is defined as

$$\mu^{s, N_e}(z) = (\mu_1^{s, N_e}(z), \dots, \mu_K^{s, N_e}(z))$$

$$\mu_i^{s, N_e}(z) = N_e z^i (s_i - \langle s, z \rangle),$$

where $\langle x, y \rangle$ is the inner product of vectors x and y , N_e the effective population size, discussed in more detail below, and the parameters $s = (s_1, s_2, \dots, s_K)$ are called fitness coefficients. The interpretation of the fitness parameters is that if $s_i > s_j$, then subpopulation i has higher growth potential compared to subpopulation j . The matrix-valued function $\sigma : \mathbb{R}^K \rightarrow \mathbb{R}^{K^2}$ is defined as

$$\sigma^2(z) = \sigma_{i,j}^2(z)_{i,j \in \{1, \dots, K\}}$$

$$\sigma_{i,j}^2(z) = z^i (\delta_{i,j} - z^j),$$

where $\delta_{i,j}$ is the Kronecker delta. Given an initial value z , we denote the marginal distribution of the process at time t by $Z_t \sim WF(s, N_e, t, z)$.

The fitClone model

Given as input time-series data measuring the relative abundances of K populations at a finite number of time points, the output of the fitClone model is a posterior distribution over the unknown parameters of interest: the fitness parameters s described in the previous section, and the continuous-time trajectories interpolating and extrapolating the discrete set of observations.

To do this, fitClone places a prior on the fitness parameters s , and uses a state space model in which the latent Markov chain is distributed according to a Wright–Fisher diffusion, and the observation model encodes noisy sampling from the population at a discrete set of time points.

Each component of the fitness parameter, now a random variable S_i , is endowed with a uniform prior over a prior range I ,

$$S_k \sim \text{Uniform}(I), k > 1,$$

where we set $S_1 = 0$ to make the model identifiable. We used $I = (-10, 10)$ in our experiments. Note that the posterior is contained far from the boundaries of this prior range in all experiments.

The initial distribution, that is, the distribution of the value of the process at time zero, is endowed a Dirichlet distribution with hyper-parameter $(1, 1, \dots, 1)$,

$$Z_0 \sim \text{Dirichlet}(1, 1, \dots, 1).$$

This can equivalently be seen as a uniform distribution over the K -simplex.

Let $t_1 < t_2 < \dots < t_{T-1} < t_T$ denote a set of process times at which measurements are available. Ideally, we would like the latent transition kernels to be given by the marginal transitions of the Wright–Fisher diffusion from last section,

$$Z_{t_m} | Z_{t_{m-1}}, S \sim WF(S, N_e, t_m - t_{m-1}, Z_{t_{m-1}}), \quad (2)$$

where N_e is estimated as a pre-processing step. In practice we resort to approximating the distribution in equation (2) via an Euler–Maruyama scheme.

Finally, for each $t \in \{t_1, t_2, \dots, t_T\}$, let $Y_t = (Y_t^1, \dots, Y_t^K)$ denote a noisy observation of the population prevalences at process time t . In the single-cell context, this is obtained by counting, for each clone, the number of cells coming from each passage, and normalizing by the number of cells sequenced in that passage. For simplicity, in both cases we use a normal observation model, that is, $Y_t^i | Z_t^i \sim \mathcal{N}(Z_t^i, \sigma_{\text{obs}}^2)$, where $\sigma_{\text{obs}}^2 = np_i(1-p_i)$ and $n = \sum_j Y_t^j$ and $p_i = Y_t^i/n$.

Estimating the effective population size

Following ref.⁴⁸, we use F'_s an unbiased moment-based estimator of the N_e where $N_e = \frac{1}{F'_s}$; and t is the number of generations between each passage.

$$F'_s = (1/t) \frac{F_s(1 - 1/(2\tilde{n})) - 1/\tilde{n}}{(1 + F_s/4)(1 - 1/n_y)}, \quad (3)$$

where $F_s = \frac{(x-y)^2}{z(1-z)}$, $z = (x+y)/2$ and $\tilde{n} = \frac{2n_y n_x}{n_y + n_x}$, the harmonic mean of the sample size (initial population size at the passage) n_x and n_y , at the two time points. Here x and y are the minor allele frequencies at the two time points.

In the multi-allelic case, we have:

$$F_s = \frac{1}{K} \sum_{i=1}^K \frac{(x_i - y_i)^2}{z_i(1 - z_i)}.$$

This is equivalent to plan 2 in ref.⁴⁸, sampling before reproduction and without replacement.

We used the sum of clone sizes as the approximate initial population size at each time point/passage. Supplementary Table 3 lists the resulting N_e estimates. Since fitClone is robust to the choice of N_e in this range, we set $N_e = 500.0$ for all datasets analysed in this paper. We note that in our model we assume that the effective population size remains constant over all time points.

Probability of positive selection

To infer evidence of positive selection, we computed a posterior distribution over the difference in selection coefficients between pairs of clones. Here, higher probability reflects the posterior density that one clone has higher fitness than another. As such, the higher the mass of this distribution, the more likely positive selection is operating over the time series.

Distribution of the probability of positive selection over pairs of clones was computed as $\max(P(s_i > s_j), 1 - P(s_i > s_j))$ for all pairs of clones i, j such that $i > j$. Let $s_{1:M} = (s_1, s_2, \dots, s_M)$ be the M post burn-in MCMC samples for the selection coefficients where $s_m = \{s_{m,1}, s_{m,2}, \dots, s_{m,K-1}\}$ are the sampled selection coefficients of clones 1 to $K-1$ at iteration m . Define $P(s_i > s_j) = \sum_{m=1}^M \mathbb{I}(s_{m,i} > s_{m,j})$ for $i, j \in \{1, \dots, K-1\}$ be the posterior probability of clone i having a larger coefficient than clone j . We computed the effect size as the absolute value of the expected difference between the selection coefficients of clones i, j , that is $|E(s_i - s_j)| = |\sum_{m=1}^M (s_{m,i} - s_{m,j})|$.

Selecting the reference clone

In our formulation of the Wright–Fisher diffusion one reference clone with selection coefficient of zero has to be chosen. The selection coefficient of the other clones are reported relative to this value. For instance, if the fittest clone is chosen as reference, the other clones will have negative selection coefficients. We chose to set the reference to a clone with an approximately monotonically decreasing trajectory (clonal abundance over time). This choice was motivated by a desire to infer a non-negative value for the fittest clones. The model is robust to the choice of the reference clone. We run the inference procedure over the same data set multiple times, each time changing the reference. The posterior ordering of clones over different choices of clones remained mostly identical.

Forecasting clonal trajectories

We forward-simulated trajectories from fitClone using the sample median of the estimated selection coefficients in TNBC-SA609 (line 1) ($B = 1.00 \pm 0.01$, $D = 1.00 \pm 0.01$, $G = 1.01 \pm 0.01$, $H = 1.02 \pm 0.02$, $E = 1.07 \pm 0.02$). We compared two independent starting clonal proportions of ($B = 0.08$, $C = 0.25$, $D = 0.51$, $E = 0.02$, $G = 0.08$, $H = 0.07$) and ($C = 0.02$, $D = 0.00$, $E = 0.05$, $G = 0.06$, $H = 0.87$), derived by physically mixing cells from a late (X8) and an early (X3) passage of the TNBC-SA609 (line 1) series in mixture-retransplant-serial passage experiments (Fig. 3b).

Reporting summary

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

Data availability

Raw sequencing data for DLP⁺ and 10x scRNA-seq are available from the European Genome-Phenome under study ID EGAS00001004448. Single-cell data from this report may be visualized in an instance of our scWGExploration platform, Alhena, available at <https://www.cellmine.org>. Source data are provided with this paper.

Code availability

The software implementation of fitClone is available at <https://github.com/UBC-Stat-ML/fitclone>.

34. Eirew, P. et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
35. Li, D. et al. Enhanced tumor suppression by adenoviral PTEN gene therapy combined with cisplatin chemotherapy in small-cell lung cancer. *Cancer Gene Ther.* **20**, 251–259 (2013).
36. Wang, Y. et al. Klotho sensitizes human lung cancer cell line to cisplatin via PI3k/Akt pathway. *PLoS ONE* **8**, e57391 (2013).
37. O’Flanagan, C. H. et al. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 210 (2019).
38. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
39. Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. & Robert, C. P. Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990 (2009).
40. Foll, M. et al. Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genet.* **10**, e1004185 (2014).
41. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).
42. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**, 599–607 (2012).
43. Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D. & Wegmann, D. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. *Genetics* **203**, 831–846 (2016).
44. Beskos, A. et al. Exact simulation of diffusions. *Ann. Appl. Probab.* **15**, 2422–2444 (2005).
45. Pollock, M. et al. On the exact and ε-strong simulation of (jump) diffusions. *Bernoulli* **22**, 794–856 (2016).
46. Jenkins, P. A. et al. Exact simulation of the Wright–Fisher diffusion. *Ann. Appl. Probab.* **27**, 1478–1509 (2017).
47. Blanchet, J. Exact simulation for multivariate Itô diffusions. *Adv. Appl. Probab.* **52**, 1003–1 (2020).
48. Jorde, P., Palm, S. & Ryman, N. Estimating genetic drift and effective population size from temporal shifts in dominant gene marker frequencies. *Mol. Ecol.* **8**, 1171–1178 (1999).

Acknowledgements This project was generously supported by the BC Cancer Foundation at BC Cancer and Cycle for Survival supporting Memorial Sloan Kettering Cancer Center. S.P.S. holds the Nicholls Biondi Chair in Computational Oncology and is a Susan G. Komen Scholar (GC233085). S.A. holds the Nan and Lorraine Robertson Chair in Breast Cancer and is a Canada Research Chair in Molecular Oncology (950-230610). Additional funding provided by the Terry Fox Research Institute Grant 1082, Canadian Cancer Society Research Institute Impact program Grant 705617, CIHR Grant FDN-148429, Breast Cancer Research Foundation award (BCRF-18-180, BCRF-19-180 and BCRF-20-180), MSK Cancer Center Support Grant/Core Grant (P30 CA008748), National Institutes of Health Grant (1RM1 HG011014-01), CCSR Grant (705636), the Cancer Research UK Grand Challenge Program, Canada Foundation for Innovation (40044) to S.A., S.P.S. and A.B.-C. We thank S. P. Otto, E. Laks, D. Min and E. Zaikova for their contribution to the project.

Author contributions S.P.S. and S.A.: project conception and oversight, manuscript writing, senior responsible authors; A.B.-C.: statistical inference method development and oversight; S.S.: computational method development, data analysis, manuscript writing; F.K.: mouse modelling, tissue procurement, data generation, manuscript writing; N.C., M.A., M.J.W., K.R.C., A.W.Z., F.D., J.P., D. Gee, D.L., A.M.: computational biology, data analysis; D. Grewal, C.O., T.M., B.W., J. Brimhall, J. Biele, J.T., H.L., T.R.d.A., S.R.L., B.Y.C.C., P.E., T.K.: tissue procurement, biological substrates and data generation; R.M., A.J.M., M.A.M.: genome sequencing; N.R.: manuscript editing.

Competing interests S.P.S. and S.A. are shareholders and consultants of Canexia Health Inc.

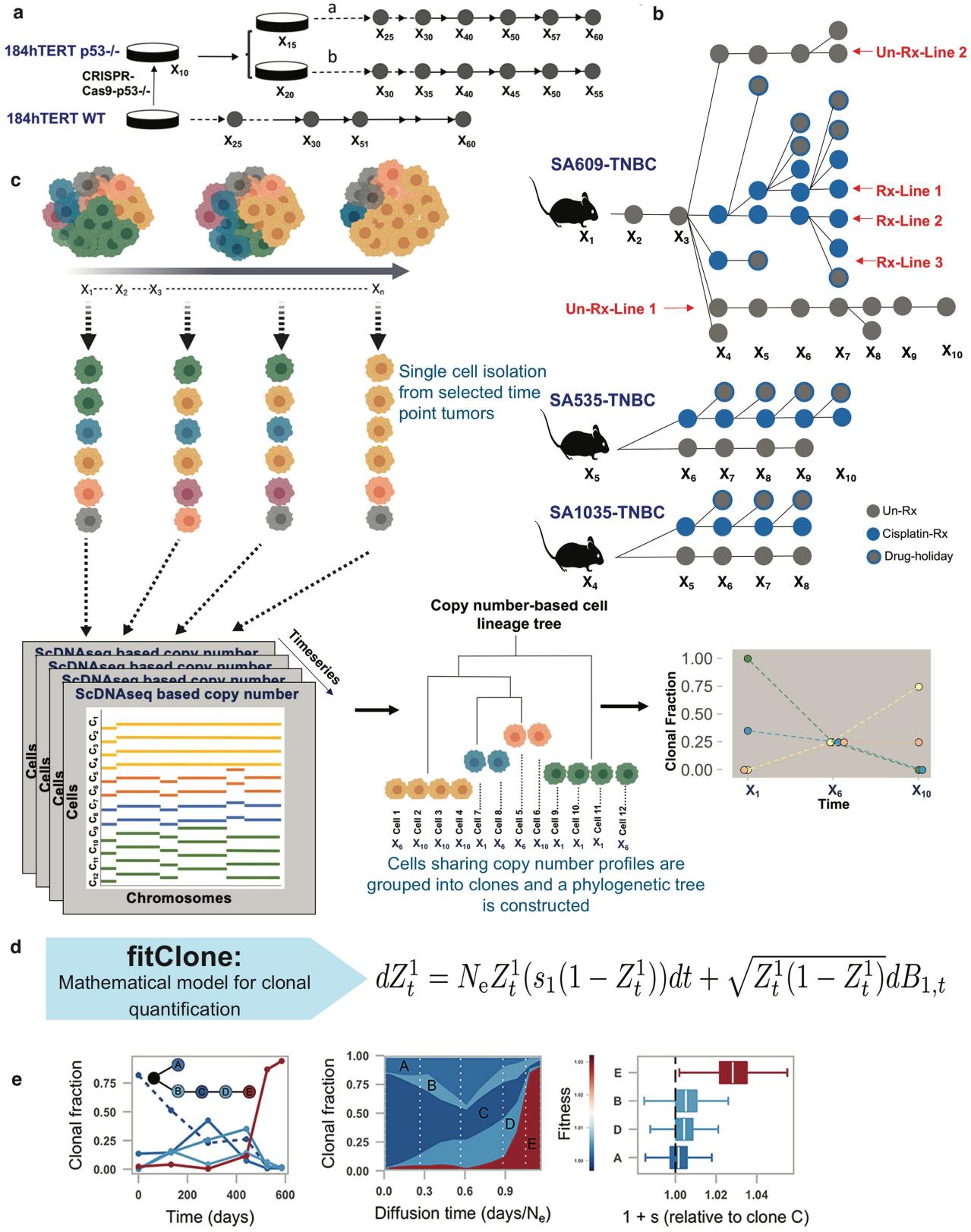
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-021-03648-3>.

Correspondence and requests for materials should be addressed to S.A. or S.P.S.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

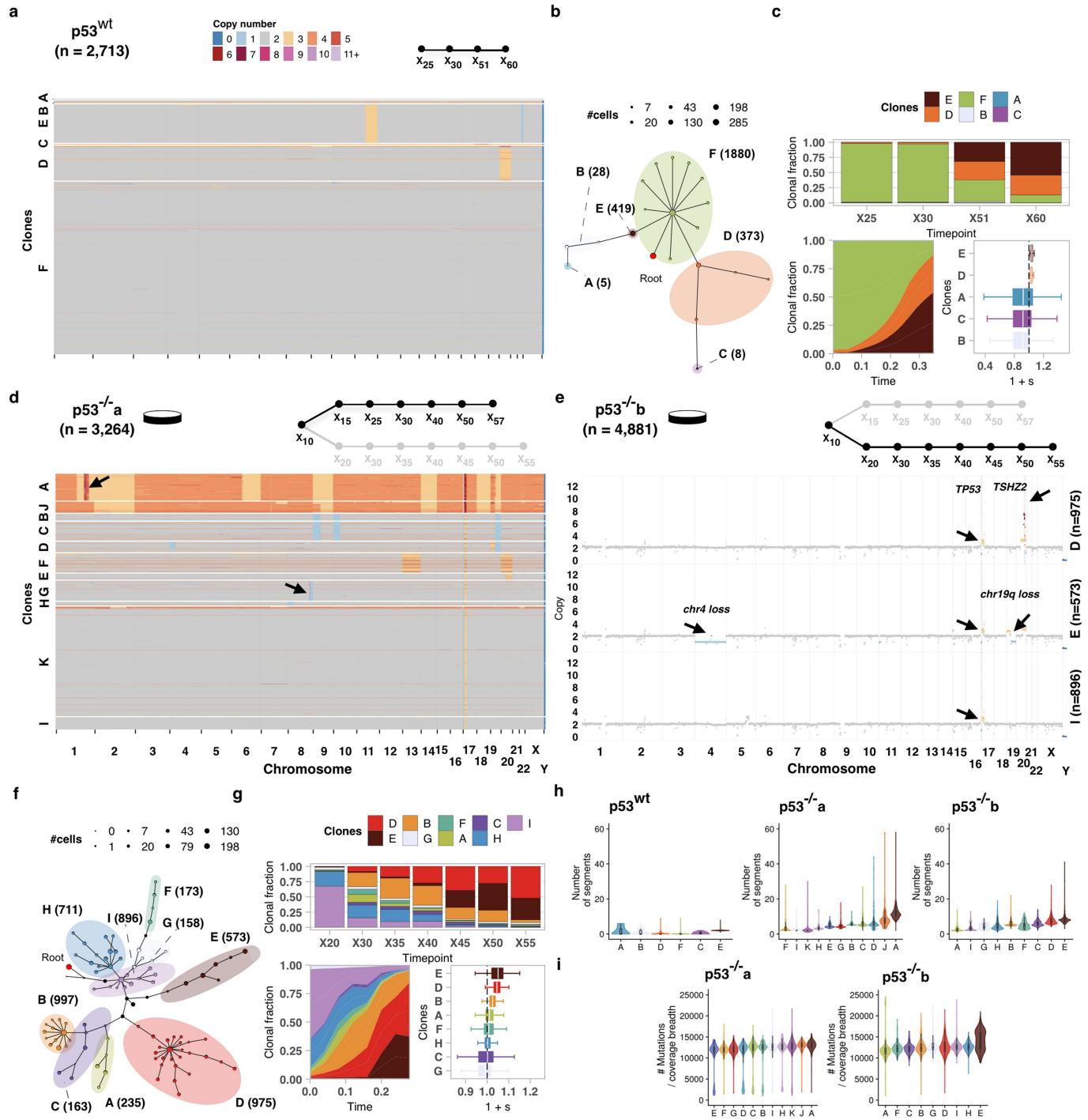
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Schematic overview of experimental design for quantitatively modelling clone-specific fitness. **a, b**, Time-series sampling from in vitro (**a**) and PDX (**b**) systems. Grey circles represent un-treated, blue represents cisplatin treated and grey with a blue outline denotes drug-holiday samples. **c**, Clonal dynamics of cell populations observed over time. Whole genome single cell sequencing of time-series samples gives copy number (left)

that in turn is used to infer a phylogenetic tree (middle), and clonal fractions over time (right). **d**, fitClone: mathematical modelling of fitness with diffusion approximation to the K-type Wright–Fisher model. **e**, fitClone inputs of clonal dynamics measured over time series (left), and inferred trajectories (middle) and posterior distributions of fitness coefficients (right). Box plots are as defined in Fig. 1b.

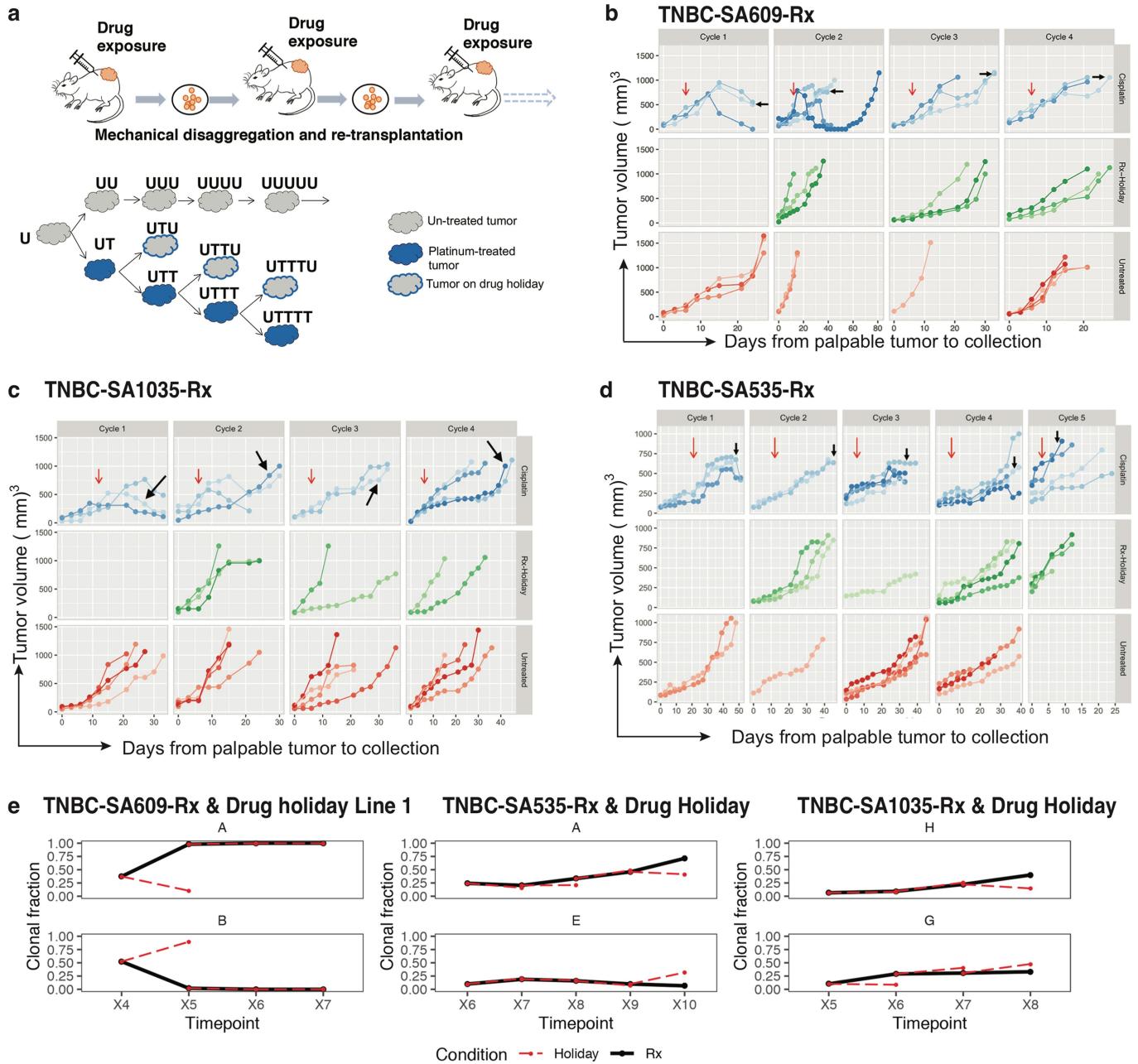
Article



Extended Data Fig. 2 | Impact of p53 mutation on fitness in 184hTERT cells.

a, Heatmap representation of copy number profiles of 2,713 p53^{WT} cells, grouped in 6 phylogenetic clades. **b**, Phylogeny of cells over the time series p53^{WT} where nodes are groups of cells (scaled in size by number) with shared copy number genotype and edges represent distinct genomic breakpoints. Shaded areas represent clones. Tree root is denoted by the red circle. **c**, Observed clonal fractions over time, inferred trajectories and quantiles of the posterior distributions over selection coefficients of fitClone model fits to

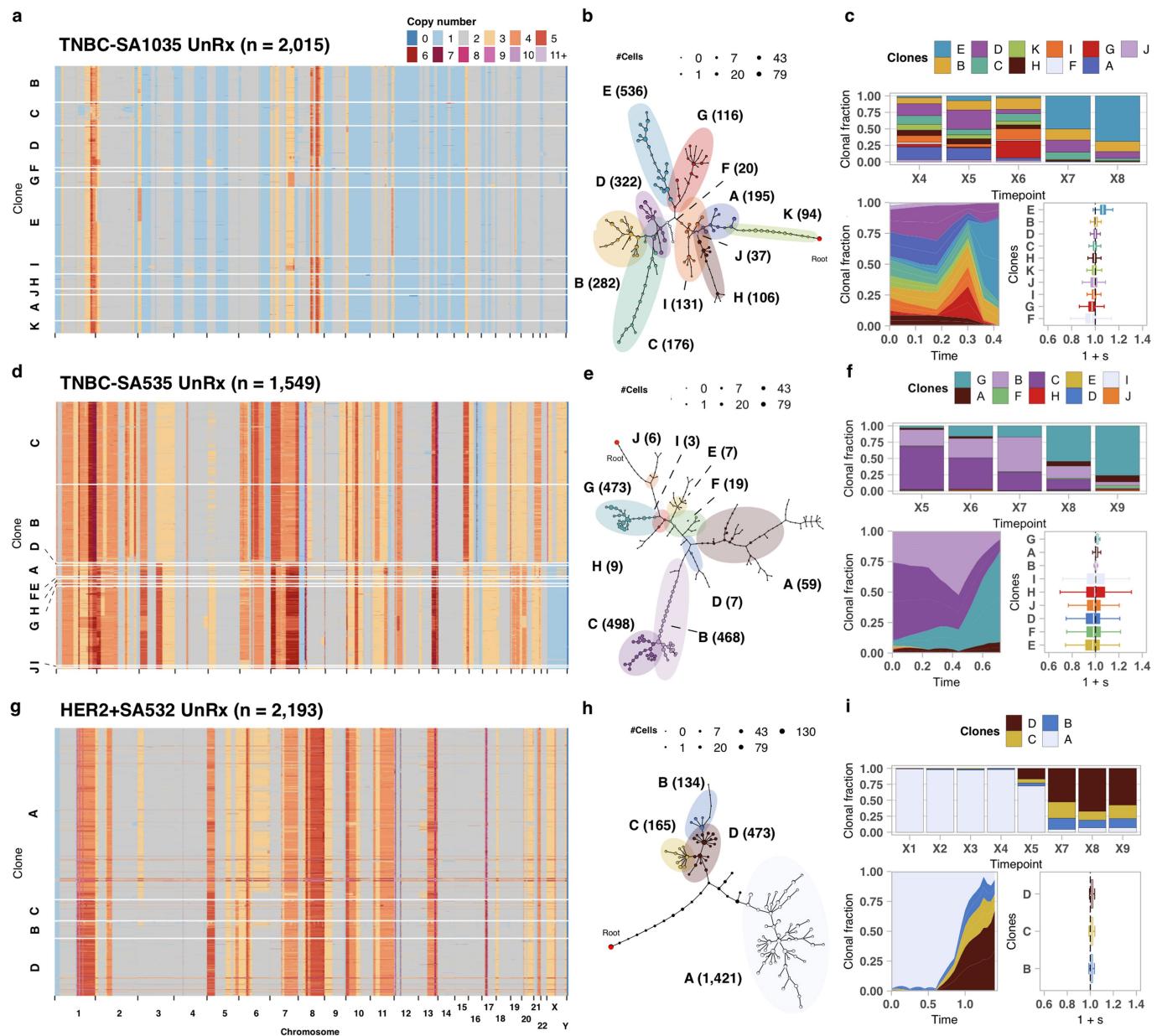
p53^{WT} with respect to the reference clone F. **d**, Analogous to **a** but for p53^{-/-} a (n = 3,264 p53^{-/-} a cells). **e**, Clonal genotypes of three representative clones for p53^{-/-} b showing high level amplification of TSHZ2 in clone D, chromosome 4 loss in clone E. Reference diploid clone I is shown for comparison. **f,g**, Analogous to **b,c**, but for p53^{-/-} b (n = 4,881 p53^{-/-} b cells; reference clone I). **h**, Number of segments per clone in hTERT WT and p53^{-/-} a and p53^{-/-} b branches. **i**, Number of mutations in p53^{-/-} a and p53^{-/-} b branches. Box plots are as defined in Fig. 1b.



Extended Data Fig. 3 | PDX tumour growth and clonal dynamics with cisplatin. **a**, Experimental design of cisplatin treatment in PDX. The solid blue colour representing cisplatin treated tumours (UT, UTT, UTTT, UTTTU); blue outlined in grey as drug holiday (UTU, UTTU, UTTTU); grey as untreated series. **b-d**, Tumour response curves in TNBC-SA609, TNBC-SA535 and TNBC-SA1035 treated with Cisplatin (blue), in drug Holiday (green) and untreated (red) where each tumour replicate is shown in a different shade. The vertical axis on the right denotes the status of tumours and on the left denotes the tumour volumes. The top horizontal axis represents number of cisplatin cycles and at

the bottom days from palpable tumours to collection. The red arrows indicate the start of treatment and the black arrows indicate the tumour sampled for scDNaseq. The bottom horizontal axis shows the tumour passage number. Each line in the big box is an individual tumour showing the growth over time. **e**, Top, clonal trajectories of the clone with the highest inferred selection coefficient in the treatment regime (solid black line) and the drug holiday counterpart (dashed red line) at each time point, in the three TNBC PDX time series; bottom, as the top row, but for a clone that grows back in the holiday regime.

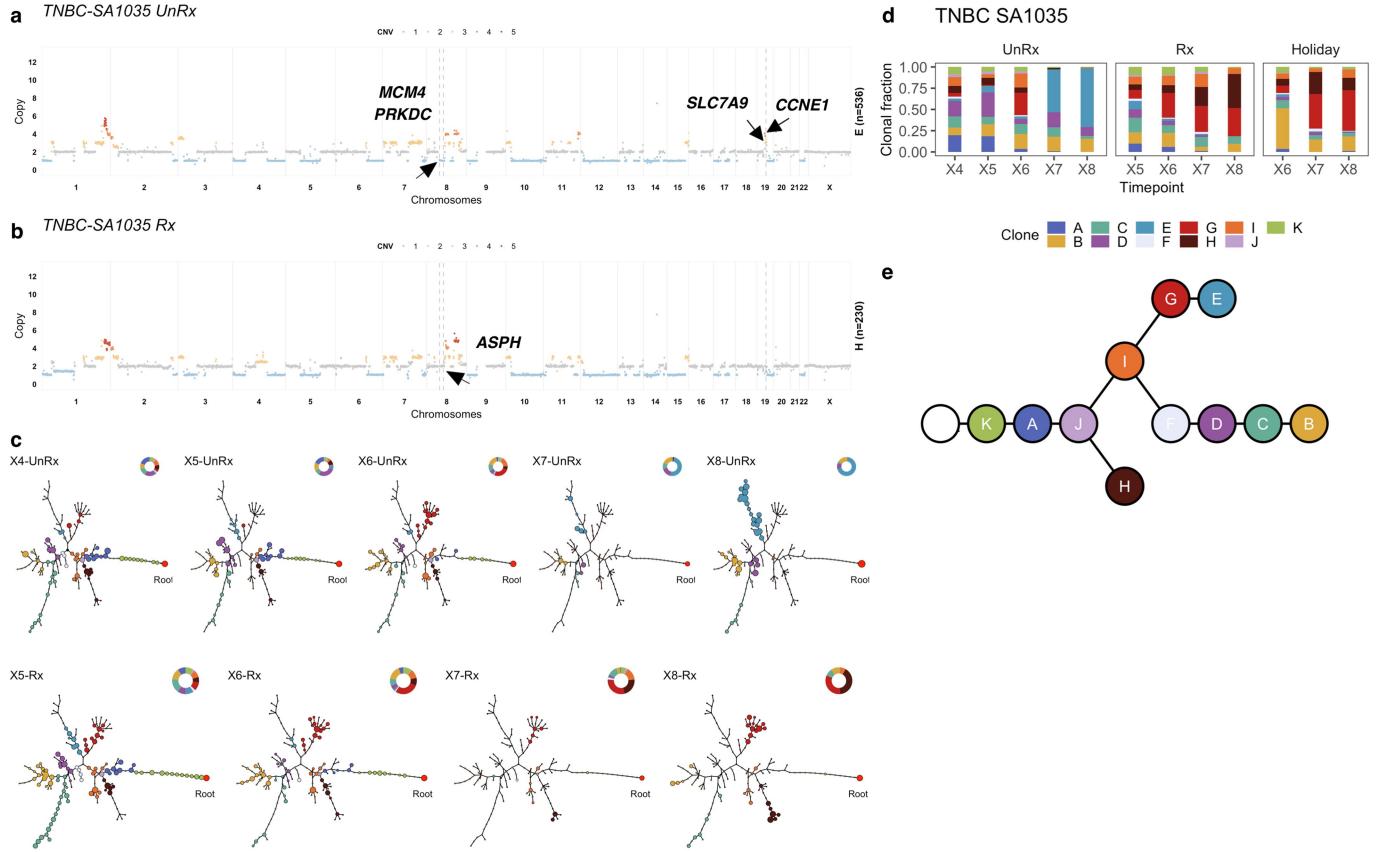
Article



Extended Data Fig. 4 | Comparison of fitness landscapes of breast cancer PDX models.

a, Heatmap representation of copy number profiles of 2,015 cells from TNBC-SA1035, grouped in 11 phylogenetic clades. **b**, Phylogeny for TNBC-SA1035. **c**, Observed clonal fractions, inferred fitClone trajectories and

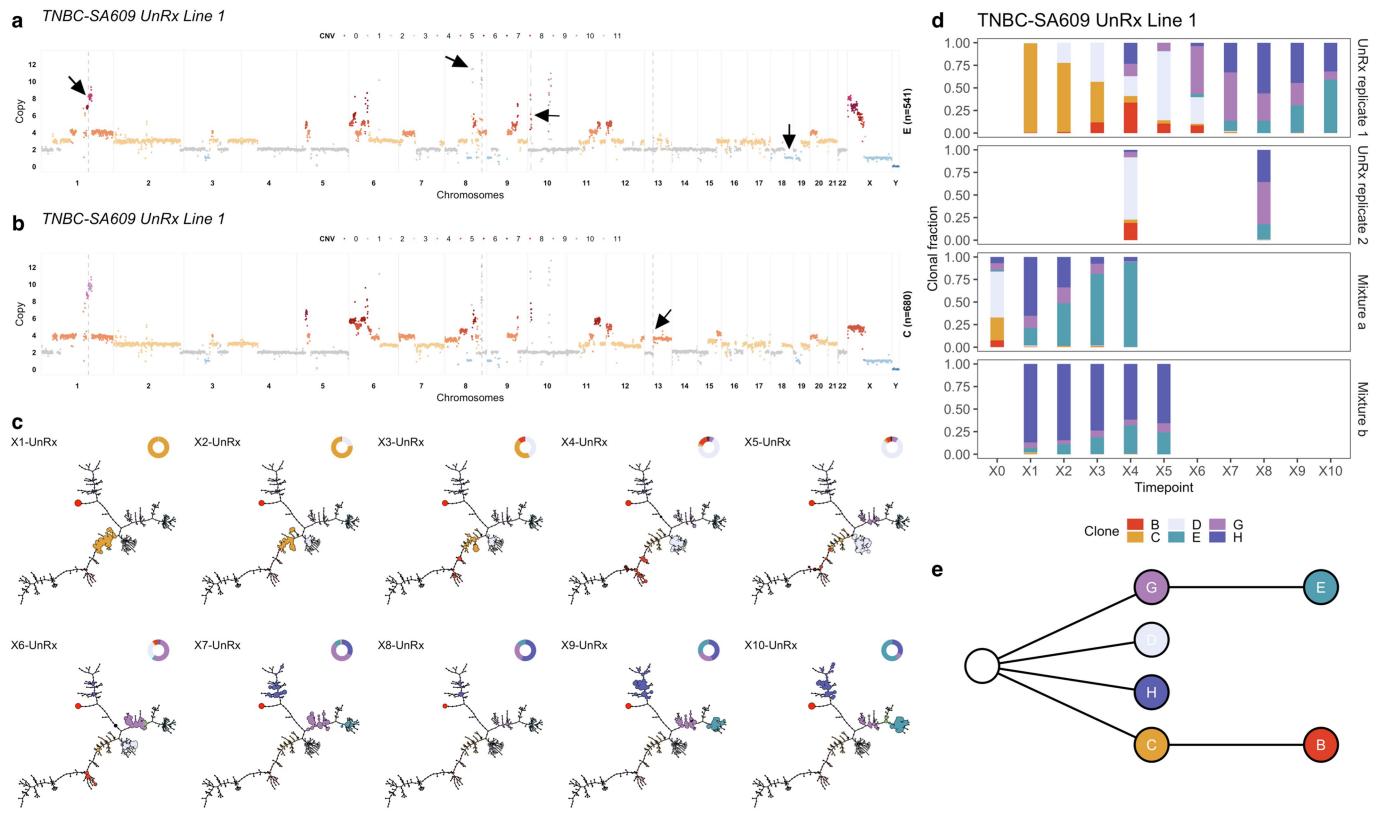
quantiles of the selection coefficients with respect to the reference clone A for the TNBC-SA1035 UnRx model. **d–f**, Analogous to **a–c** but for HER2⁺SA535 (n=1,549 cells; reference clone C). **g–i**, Analogous to **a–c** but for HER2⁺SA532 (n=2,193 cells; reference clone A). Box plots are as defined in Fig. 1b.



Extended Data Fig. 5 | Impact of pharmacologic perturbation with cisplatin on fitness landscapes in TNBC-SA1035. **a**, Copy number genotype of clone E from the untreated time series. **b**, Copy number genotype of clone H from treated time series (arrows indicate differences to clone E). **c**, Evolution in

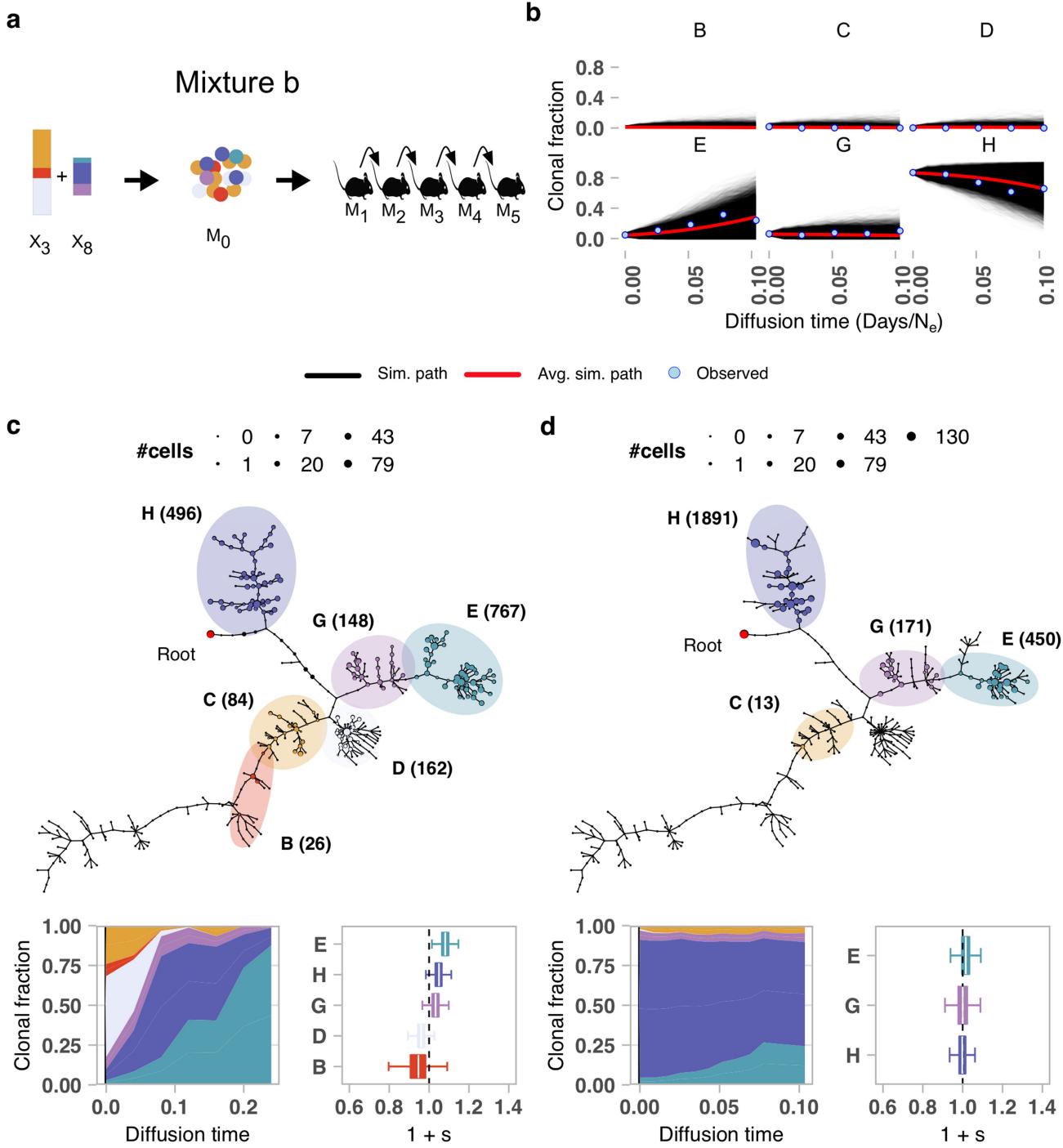
absence of treatment and as a function of drug treatment. For each sample, the phylogeny with clonal abundance from DLP⁺ is shown, reflecting selection. **d, e**, The observed clonal abundances (**d**) and the summarized clonal phylogenetic tree (**e**).

Article



Extended Data Fig. 6 | Tumour evolution in absence of pharmacologic perturbation in TNBC-SA609 line 1. **a, b,** Copy number genotype of clone E (**a**) and copy number genotype of clone C, the reference clone (arrows indicate differences to clone E) (**b**). **c,** Evolution in absence of treatment. For each

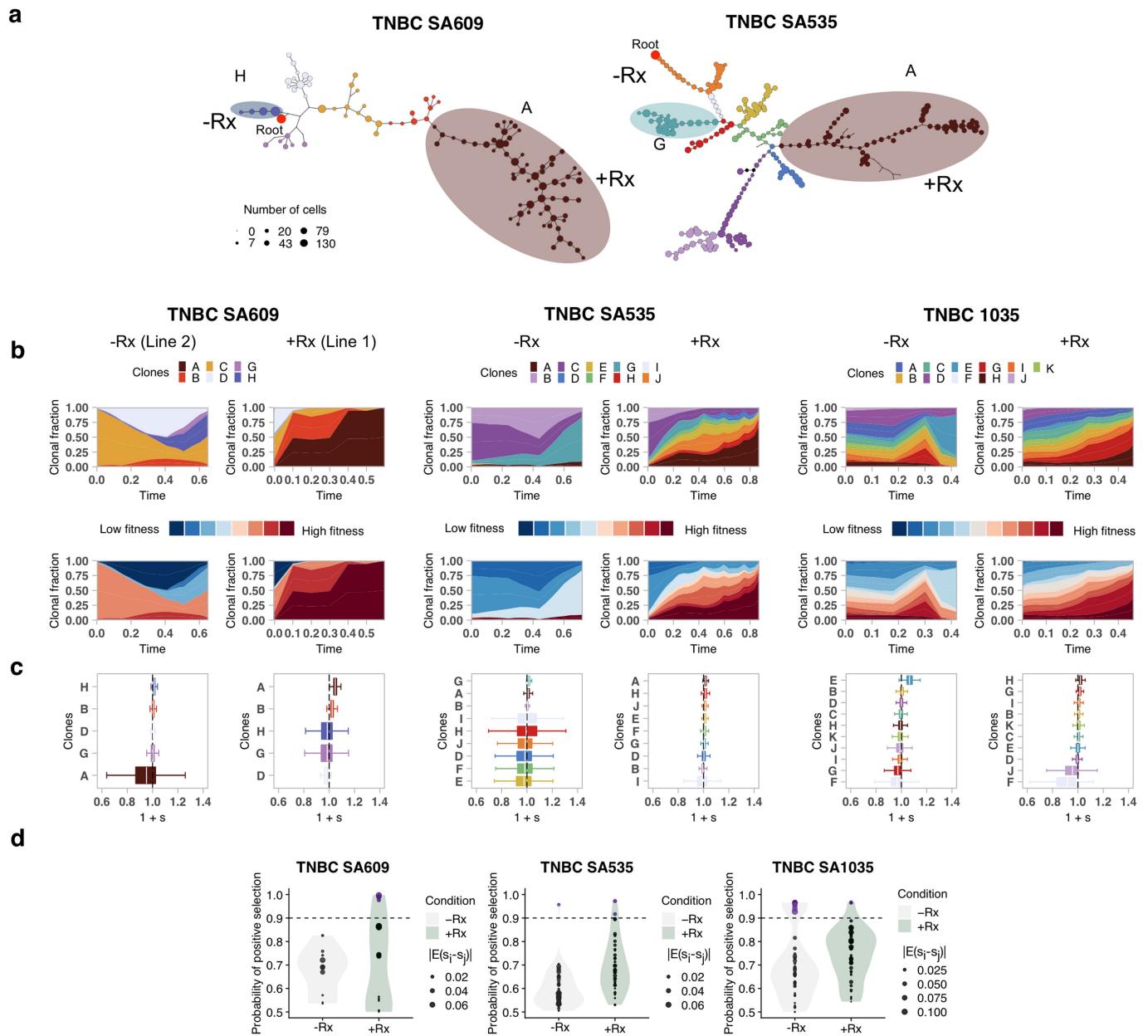
sample, the phylogeny with clonal abundance from DLP⁺ is shown, reflecting selection. **d, e,** The observed clonal abundances (**d**) and the summarized clonal phylogenetic tree (**e**).



Extended Data Fig. 7 | Mixture experiment in TNBC-SA609 PDX Line 1.
a, Clonal proportions of TNBC-SA609 Line 1 X3 and X8 used to generate the initial mixture M0 and subsequent serial passaging, yielding 5 samples for mixture experiment b (mixture b). **b**, Forward simulations from the original time series and starting population proportions in the initial experimental mixture b. Simulated trajectories are shown superimposed with mean

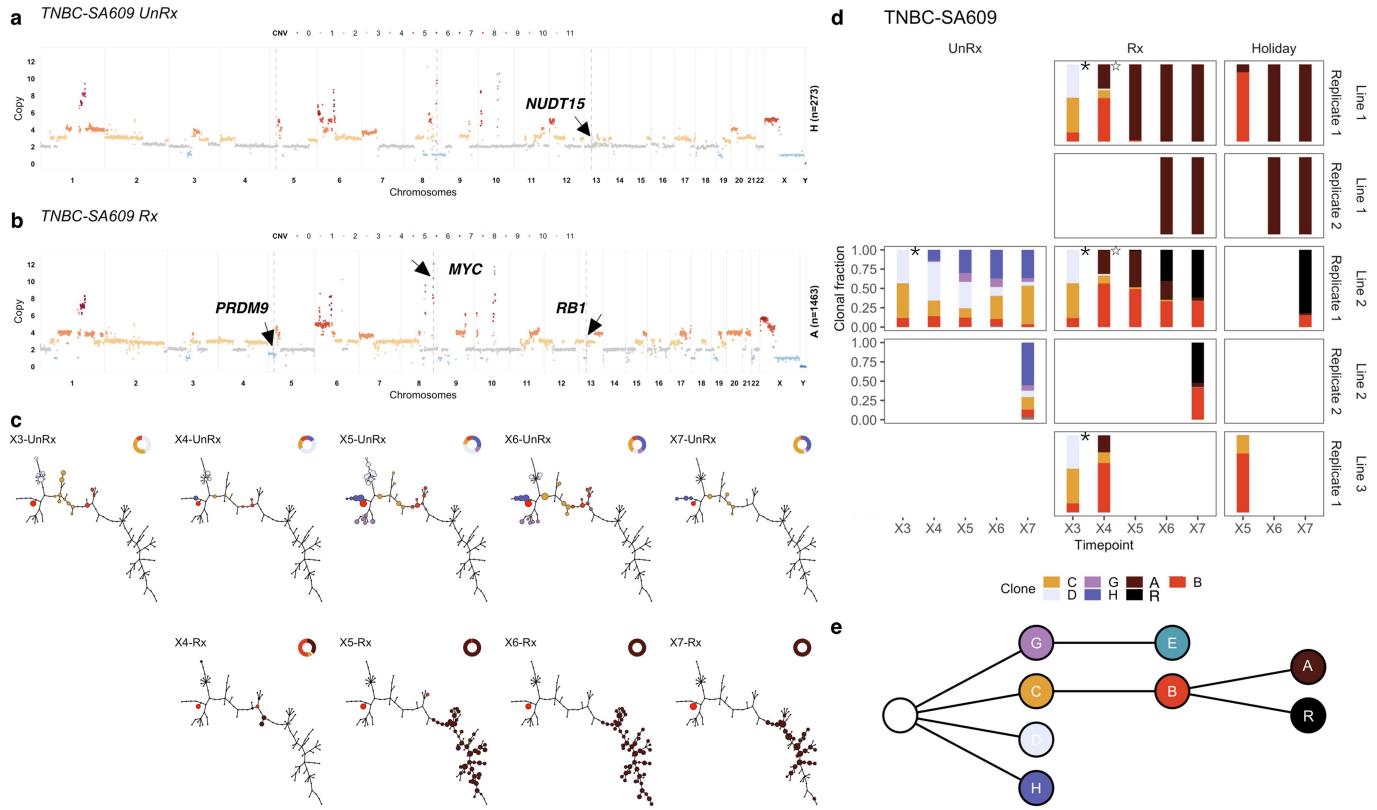
simulation (red line) and observed clonal fractions (blue dots). The observation time is adjusted to match the simulation diffusion time. **c**, Summary phylogenetic tree, inferred trajectories and fitness coefficients (relative to reference clone C) for mixture a. **d**, As in c but for mixture b (relative to reference clone C). Box plots are as defined in Fig. 1b.

Article



Extended Data Fig. 8 | Fitness landscape reversal in early cisplatin treatment in TNBC PDX models. In each column, the left and right sub-panels are from the untreated and treated branches respectively. **a**, Phylogenetic trees annotated with fittest clones in -Rx and Rx. **b, c**, Inferred trajectories, first coloured by clonal assignment, and then coloured by fitness rank (**b**), and

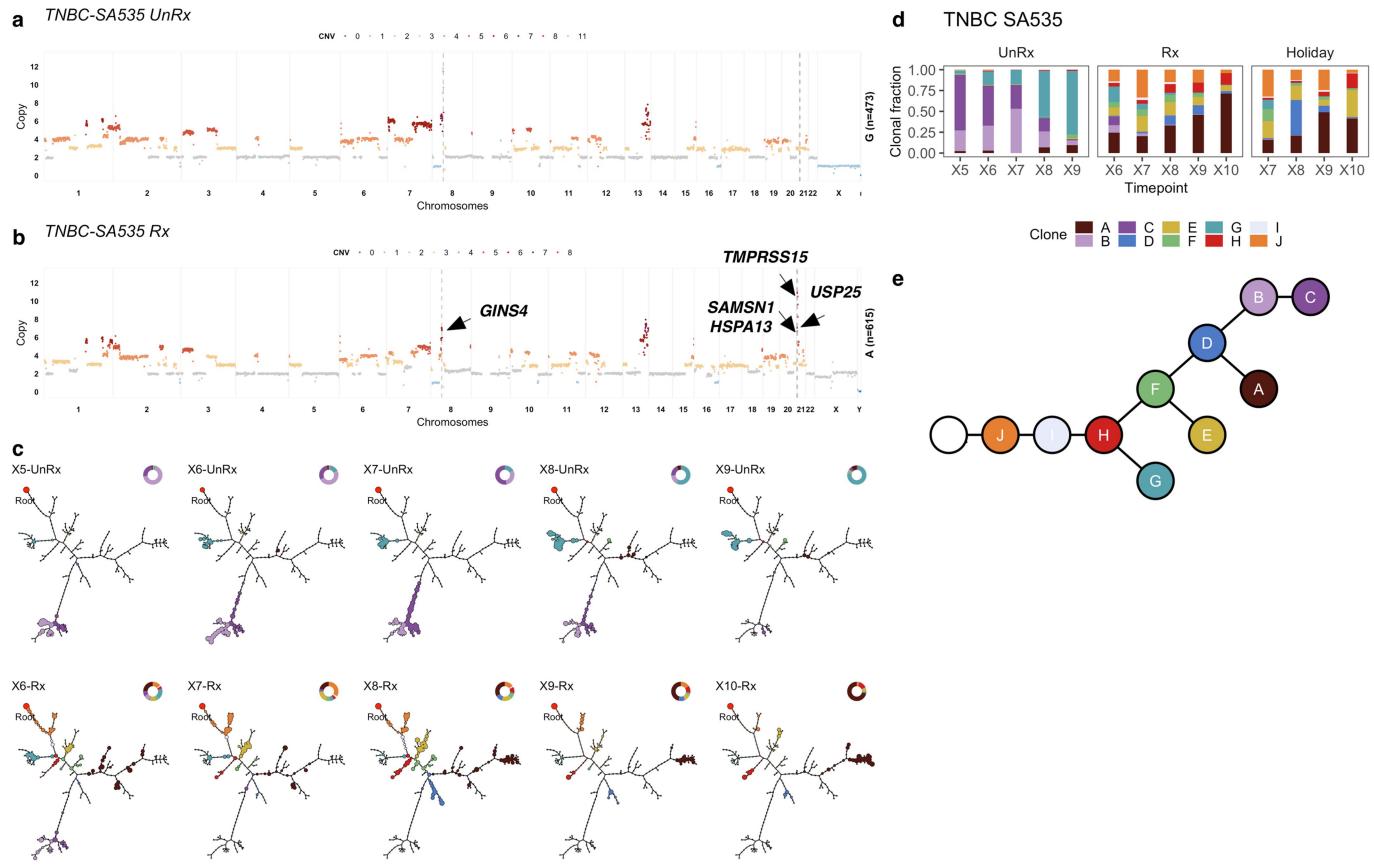
quantiles of selection coefficients of fitClone model fits to each branch with respect to the reference Clone C in TNBC-SA609, Clone C in TNBC-SA535, and clone A in TNBC-SA1035 (**c**). **d**, Distribution over the probability of positive selection over pairs of clones for each series. Box plots are as defined in Fig. 1b.



Extended Data Fig. 9 | Impact of pharmacologic perturbation with cisplatin on fitness landscapes in TNBC-SA609. **a**, Copy number genotype of clone H from untreated time series. **b**, Copy number genotype of clone A from the treated time series (arrows indicate differences to clone H). **c**, Evolution in absence of treatment (top) and as a function of treatment (bottom). For each

sample, the phylogeny with clonal abundance from DLP⁺ is shown, reflecting selection. **d**, The observed clonal abundances. Starred time points are identical and reproduced to denote the identical starting point. **e**, Summarized clonal phylogenetic tree.

Article



Extended Data Fig. 10 | Impact of pharmacologic perturbation with cisplatin on fitness landscapes in TNBC-SA535. **a**, Copy number genotype of clone G from untreated time series. **b**, Copy number genotype of clone A from treated time series (arrows indicate differences to clone E). **c**, Evolution in

absence of treatment and as a function of drug treatment. For each sample, the phylogeny with clonal abundance from DLP⁺ is shown, reflecting selection. **d, e**, The observed clonal abundances (**d**) and the summarized clonal phylogenetic tree (**e**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

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

n/a Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

All commercial, open source and custom code software is detailed in the Supplementary Information. Custom software that is part of this study is available as specified in the Methods section: Software and source code implementing analytical and statistical methods available on github: [<https://github.com/UBC-Stat-ML/fitclone>]

Data analysis

All commercial, open source and custom code software is detailed in the Supplementary Information. Custom software that is part of this study is available as specified in the Methods section: Software and source code implementing analytical and statistical methods available on github: [<https://github.com/UBC-Stat-ML/fitclone>]

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

Raw sequencing data for DLP+ and 10x scRNASeq will be available from the European Genome-Phenome archive prior to publication.

Field-specific reporting

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

Life sciences

Behavioural & social sciences

Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size	For each timeseries sample, we estimated power to detect clones to 2% for choosing the number of cells to sequence. For number of samples per timeseries, this was determined due to practical constraints, sampling tumours over a 2.5yr interval. As this was an exploratory study, no calculations were used to determine the number of samples/timeseries. We note that these series spanned 100s of days (or cell division generations) in the cell lines, and up to 10 serial passages in the PDX models.
Data exclusions	For all samples, low quality sequenced cells were excluded as per the predefined criteria reported in the Supplementary Information.
Replication	In cell line and PDX experiments, unperturbed timeseries were compared to perturbed timeseries. Replicates were performed in the cell lines and the TNBC-SA609 Line 1 untreated PDX, via one and two 'mixing' experiment respectively to reproduce the observation of fitness clones. Furthermore, multiple replicate observations were generated in the PDX models in two ways: parallel replicate timeseries with cisplatin treatment, and duplicate experiments of specific time points.
Randomization	Randomization was not applicable in this study as all experiments were controlled perturbation experiments, generating timeseries with or without isogenic TP53 mutation, or cisplatin dosing as described in the Supplementary Information.
Blinding	Blinding was not relevant due to the controlled perturbation nature of the experimental design.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

See Supplementary Table 4 in the Supplementary Information

Validation

See Section 2.1 Histopathology of PDX tumours in the Supplementary Information.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Section 1 of the Supplementary Information provides this in detail: "The human mammary epithelial cell line wild type 184-hTERT and isogenic 184-hTERT-P53 KO cell line, generated from 184hTERT WT-L9, were grown as previously described" References [1, 2].

Authentication

These cell lines were used previously and reported in previous papers from our laboratory:
 1. Laks, E. et al. Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell* 179, 1207–1221 (2019).
 2. Burleigh, A. et al. A co-culture genome-wide rnai screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. *Breast Cancer Research* 17, 4 (2015).

Mycoplasma contamination

Mycoplasma testing was negative.

Commonly misidentified lines
(See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Animals and other organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

The details on mouse strains for xenografting tumours is detailed in Section 2 of the Supplementary Information.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

The Ethics Committees at the University of British Columbia approved all the experiments using human resources. Patients in Vancouver, British Columbia were recruited, and samples were collected under the tumour tissue repository (TTRH06-00289) protocol and transplanted in mice under the Neoadjuvant PDX (University of British Columbia BC Cancer Research Ethics Board H20-00170) protocols.

After informed consent, tumour fragments from patients undergoing excision or diagnostic core biopsy, were collected. Tumour materials were processed as described in [3] and transplanted in mice under the animal resource centre (ARC) bioethics protocol (A19-0298-A001) approved by the animal care committee.

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

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

The Ethics Committees at the University of British Columbia approved all the experiments using human resources. Patients in Vancouver, British Columbia were recruited, and samples were collected under the tumour tissue repository (TTRH06-00289) protocol and transplanted in mice under the Neoadjuvant PDX (University of British Columbia BC Cancer Research Ethics Board H20-00170) protocols.

Recruitment

Bias and collection was not relevant here as this was not a population based study, nor a clinical trial. Breast cancer patients in Vancouver, British Columbia were recruited, and samples were collected under the tumour tissue repository (TTRH06-00289) protocol.

Ethics oversight

The Ethics Committees at the University of British Columbia

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