1   **Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy**

2   Xiaoyang Lan[1,2,3], David J. Jörg[4,5], Florence M. G. Cavalli[1,2], Laura M. Richards[12,13], Long V.
3   Nguyen[7], Robert J. Vanner[1,2,3], Paul Guilhamon[12,13,14], Lilian Lee[1,2], Michelle Kushida[1,2], Davide
4   Pellacani[7,8], Nicole I. Park[1,2,3], Fiona J. Coutinho[1,2,3], Heather Whetstone[1,2], Hayden J.
5   Selvadurai[1,2], Clare Che[1,2], Betty Luu[1,2], Annaick Carles[9], Michelle Moksa[9], Naghmeh
6   Rastegar[1,2], Renee Head[1,2], Sonam Dolma[1,2,11], Panagiotis Prinos[13,20], Michael D. Cusimano[17,18],
7   Sunit Das[17,18], Mark Bernstein[16,18], Cheryl H. Arrowsmith[13,20], Andrew J. Mungall[10], Richard A.
8   Moore[10], Yussanne Ma[10], Marco Gallo[19], Mathieu Lupien[12,13,14], Trevor J. Pugh[12,13], Michael D.
9   Taylor[1,2,11,15,18], Martin Hirst[9,10], Connie J. Eaves[7,8], Benjamin D. Simons[4,5,6*], and Peter B.
10   Dirks[1,2,3,15,18*]

11   [1]Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto,
12   Ontario M5G 0A4, Canada.

13   [2]The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children,
14   Toronto, Ontario M5G 0A4, Canada.

15   [3]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada.

16   [4]Cavendish Laboratory, Department of Physics, J. J. Thomson Avenue, Cambridge CB3 0HE,
17   United Kingdom.

18   [5]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, CB2
19   1QN, United Kingdom.

20   [6]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge,
21   United Kingdom.

22   [7]Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British
23   Columbia V5Z 1L3, Canada.

24   [8]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia
25   V6T 2B5, Canada.

26   [9]Centre for High-Throughput Biology, Department of Microbiology and Immunology,
27   University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada.

28   [10]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue,
29   Vancouver, British Columbia V5Z 1L3, Canada.

30   [11]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto,
31   Ontario M5G 0A4, Canada.

32   [12]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2M9,
33   Canada.

34   [13]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada.

35

36   [14]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada.

37   [15]Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario M5S 3E1, Canada.

38   [16]Division of Neurosurgery, Toronto Western Hospital, Toronto, ON M5T 2S8, Canada.

39   [17]Division of Neurosurgery, St. Michael's Hospital, Toronto, ON M5B 1W8, Canada.

40   [18]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, ON M5S
41   1A8, Canada.

42   [19]Departments of Physiology and Pharmacology, Biochemistry and Molecular Biology, Alberta
43   Children's Hospital Research Institute, Arnie Charbonneau Cancer Institute, University of
44   Calgary, Calgary, AB T2N 4N1, Canada.
45
46   [20]Structural Genomics Consortium, University of Toronto, Toronto, ON M5G 1L7, Canada.

47   *Correspondence: bds10@cam.ac.uk (B.D.S.), peter.dirks@sickkids.ca (P.B.D.)
48

66 **Summary**

67 Human glioblastomas (GBMs) harbour a subpopulation of glioblastoma stem cells (GSCs) that

68 drive tumourigenesis. However, the origin of intra-tumoural functional heterogeneity between

69 GBM cells remains poorly understood. Here we study the clonal evolution of barcoded GBM

70 cells in an unbiased way following serial xenotransplantation to define their individual fate

71 behaviours. Independent of an evolving mutational signature, we show that the growth of GBM

72 clones *in vivo* is consistent with a remarkably neutral process involving a conserved proliferative

73 hierarchy rooted in GSCs. In this model, slow-cycling stem-like cells give rise to a more rapidly

74 cycling progenitor population with extensive self-maintenance capacity, that in turn generates

75 non-proliferative cells. We also identify rare "outlier" clones that deviate from these dynamics,

76 and further show that chemotherapy facilitates the expansion of pre-existing drug-resistant

77 GSCs. Finally, we show that functionally distinct GSCs can be separately targeted using

78 epigenetic compounds, suggesting new avenues for GBM targeted therapy.

79

80

81

82

83

84

85

86

Introduction

Glioblastoma (GBM) is the most common and malignant form of adult brain tumour[1]. Central to our understanding of GBM biology is the idea that tumour initiation, maintenance, and regrowth following treatment are seeded by glioblastoma stem cells (GSCs)[2,3]. Evidence for a proliferative hierarchy in GBM has been derived from xenotransplantation of specific GBM subsets defined by surface marker expression[2], genetic lineage tracing in mouse models[3] and more recently, single-cell RNA-sequencing[4,5]. In parallel, GBMs exhibit substantial intra-tumoural genomic heterogeneity[6,7] that could theoretically be based in GSCs with variations in growth potential, treatment responsiveness, or invasiveness[8-10]. However, recent evidence from other systems demonstrate that the intrinsic growth dynamics of a functionally homogeneous population of stem cells is already sufficient to create a wide range of clonal growth behaviours[11-14]. Therefore, it is yet unclear whether the heterogeneity of human GBM clones is primarily derived from their genomic heterogeneity, or the stochastic outcome of their hierarchical mode of growth.

DNA barcoding is a methodology that enables the proliferative capacity of individual cells to be resolved within polyclonal populations, with diverse applications in stem cell and cancer biology. Recent investigations with this strategy have already provided crucial insights into the lineage potential of normal stem cells[15], the proliferative heterogeneity of their transformed counterparts[16], as well as mechanisms of cancer drug resistance[17] and metastasis[18]. Importantly, characterizations of population dynamics in a quantitative and unbiased way can be used to inform a mathematical framework to explain complex behaviours[13,17]. Here, we perform DNA barcoding of primary GBM cells in order to investigate the quantitative behaviours of GSC clones, creating a general, minimal model of GBM growth in which a high degree of intra-

110   tumoural functional complexity can be derived from a homogeneous population of stem-like

111   cells.

112

113   **Lineage tracing of human GBM cells**

114   Lineage tracing assays based on genetic mouse models have demonstrated that quiescent

115   stem-like cells promote brain tumour recurrence following chemotherapy[3,19]. However, it

116   remains unclear how these cells contribute to tumour growth in genetically heterogeneous human

117   GBM[6,7,20,21]. To identify potential differences in tumour clone-initiating potential, tolerance to

118   chemotherapy and invasion capacity, we made use of a lentiviral barcoding strategy to trace the

119   output of individual cells *in vivo* (Fig. 1a)[15,16,22]. Freshly dissociated cells from primary (GBM-

120   719, -729, -735, -743, and -754) and recurrent (GBM-742) GBMs were transduced with a library

121   of biologically neutral barcodes prior to their transplantation into the brains of NOD/SCID/IL-2γ^-

122    $^{/-}$ (NSG) mice within 24 hours of isolation, a time window below the doubling time of GSCs

123   (Extended Data Fig. 1a-c). For each tumour sample, spiked-in controls were included to estimate

124   relative clone sizes from barcode read counts (Extended Data Fig. 1d-f). Given the high library

125   diversity ($\sim 2\times 10^5$) and limiting transduction efficiency across experiments (<38%), the majority

126   of labelled cells were expected to carry unique barcodes (Extended Data Fig. 1g-h and

127   Supplementary Theory 1).

128   Exome and RNA sequencing of primary tumours identified mutations in common GBM-

129   associated genes (*TP53*, *EGFR*, *PDGFRA*) and signatures of the Classical and Proneural

130   transcriptional subgroups (Extended Data Fig. 2a-b)[20]. Histologically, xenografts resemble

131   human GBM and have abundant expression of the neural precursor marker nestin (Fig. 1b and

132 Extended Data Fig. 3a-b)[2]. Consistent with the significant inter-patient heterogeneity of human

133 GBM[20,21], tumours generated from different primary samples differed in proliferative activity,

134 apoptosis rates, growth rates and response to temozolomide (TMZ) chemotherapy (Extended

135 Data Fig. 3c-d). In the following, we focused first on GBM-719 for which the largest xenograft

136 data set was available, using xenografts from other GBMs to test for consistency in their

137 properties.

138   Growth of GBM cells *in vivo* was concomitant with expansion in both the injected

139 (ipsilateral) and non-injected (contralateral) hemispheres (Fig. 1c and Extended Data Fig. 4a-b).

140 For GBM-719, 1,532 clones (derived from ~3% of barcoded cells) expanded above the detection

141 threshold, with 475 present in both hemispheres. The sizes of these "surviving" clones were

142 broadly distributed, with the majority remaining small (Fig. 2a). A further, smaller reduction in

143 clone number was observed upon serial passaging, with a fraction becoming apparent only in the

144 second passage, indicating that some clonogenic cells did not reach the detection threshold

145 within the first passage (Fig. 1d). These observations suggest that the primary GBM population

146 contained only a subset of cells with continuous tumour-maintaining activity (GSCs). However,

147 the abundance of surviving clones and broad size distributions demonstrate that tumour growth

148 does not rely on the activity of a few tumour-initiating cells (Fig. 2a and Extended Data Fig.

149 4e)[4,5].

150

151 **GBM clones are uniformly invasive**

152   We next sought to define the invasive capacity of barcoded GBM clones by comparing

153 clonal composition between the ipsilateral and contralateral hemispheres, the latter representing

154 expansion of invasive cells (Extended Data Fig. 4a-b). In all experiments, the sizes of clones in

155 both hemispheres were either highly correlated from the first passage on, or became highly

156 correlated soon thereafter (Extended Data Fig. 4c), indicating that clonal behaviour in the

157 contralateral side reflected their behaviour in the ipsilateral side. We then asked whether clones

158 that were exclusively found in the contralateral side have a higher invasive capacity. However,

159 xenografts derived from re-injecting contralaterally-harvested cells were primarily composed of

160 clones that had been present in both hemispheres in the previous passage (Extended Data Fig.

161 4d). It follows that self-renewal and invasion capacity are coincident properties of the same

162 labelled clones within each human GBM. Spatial separation of genetically distinct clones may

163 therefore represent transient variations in local dispersal, which become amplified over

164 time[6,10,23,24].

165

166 **Neutral hierarchical growth dynamics**

167   A consistent feature of clone sizes across all passages and between hemispheres was their

168 broad distribution (Fig. 2a and Extended Data Fig. 4e). Such functional heterogeneity could

169 derive from engrained "fitness" advantages of some tumour-initiating cells over others, resulting

170 from heritable genetic or epigenetic alterations[8]. Alternatively, variation in clonal output could

171 result from "neutral" processes, reflecting the chance outcome of cell fate decisions obtained

172 within an equipotent tumour-initiating population[11,12]. To discriminate between these

173 possibilities, we looked for evidence of equipotency in the distribution of relative clone size.

174 Remarkably, the distributions were found to be consistent with a negative binomial dependence

175 — as evidenced by the exponential form of the first incomplete moment (Fig. 2b, Extended Data

176 Figs. 5-6 and Supplementary Theory 2). Some xenografts also showed a minority (<4%) of large

177 clones that lay outside this distribution (Fig. 2b and Extended Data Fig. 4g, red arrowhead), a

178 feature returned to below. With clone size distributions across all 6 patient tumour samples

179 largely characterized by just one parameter (the constant of the exponential), these observations

180 suggest that GBM intra-tumoural heterogeneity derives primarily from the growth characteristics

181 of a single equipotent cell population rather than an engrained differential fitness of subclones,

182 an unexpected finding given the inter- and intra-patient genomic diversity of GBM and the

183 ongoing genomic evolution observed in xenografts (Extended Data Fig. 5-6 and Supplementary

184 Theory 3)[6,7,20,21].

185 How could a negative binomial clone size distribution arise? Such behaviour is common

186 in population dynamics and is typically associated with processes involving the sporadic creation

187 of "individuals" —cells in this case— that, when born, undergo a stochastic process, selecting

188 with equal probability between duplication (birth) or loss (death) and supported by a slow influx

189 from another compartment (immigration) – a "critical birth-death process with immigration"

190 (Supplementary Theory 3)[25]. In the tumour context, this behaviour translates to a proliferative

191 hierarchy in which a slow-cycling stem cell-like population undergoes serial rounds of invariant

192 asymmetric cell division, giving rise to a self-sustaining, rapidly-dividing progenitor population

193 that generates short-lived non-proliferative progeny (Fig. 2c and Supplementary Theory 4).

194 But, is a mode of strictly invariant asymmetric cell division plausible? Since most

195 barcoded clones survive dilution through serial passaging (Fig. 1d), individual clones at the end

196 of the previous passage are likely to host a multiplicity of stem-like cells. Cell division must

197 therefore also lead to symmetric fate outcomes so that their numbers can accumulate in

198 individual clones. However, so long as asymmetric fate outcomes predominate, the resulting

199 clone size distributions do not depart significantly from the observed negative binomial form

200 (Supplementary Theory 4).

201   Based on a quantitative analysis of clone size, we propose that human GBM growth in

202   xenografts is defined by a minimal model involving a defined GSC hierarchy (Fig. 2c and

203   Supplementary Theory 4). To challenge the model and define the minimal set of parameters

204   governing GBM growth, we used stochastic simulations to compare the predicted clonal

205   dynamics with experimental findings (Fig. 2d-h and Supplementary Theory 5). In assessing the

206   viability of the model, we constrained the simulation using a range of biologically plausible

207   parameters based on the overall expansion of xenografts along with the proportion of actively

208   dividing and apoptotic cells (Extended Data Fig. 3d and Supplementary Theory 5). Over the

209   determined range of parameters, simulations revealed an approximately negative binomial clone

210   size distribution across all serial passages (Fig. 2e), consistent with experiment. Using the unique

211   barcoding of clones, we assessed correlations of clone size and survival likelihoods across serial

212   passages. Remarkably, the minimal model captured the range of data to a high level of accuracy

213   (Fig. 2f-h, Extended Data Fig. 4f and Supplementary Theory 6). Quantitative analysis of clone

214   size distributions for GBM-742 and GBM-754, in addition to independent analysis of mutational

215   data derived from GBM-719 xenografts, also provided strong evidence in favour of the same

216   paradigm (Extended Data Fig. 6d-i, Supplementary Theory 6-7).

217

218   **Two divergent GSC phenotypes**

219   Building on the findings above, we next sought to define the effect of TMZ

220   chemotherapy on clonal dynamics. Analysis of the TMZ-treated xenografts clearly distinguished

221   two divergent behaviours: A majority of clones were sensitive to TMZ treatment and present at

222   low abundances ("Group A" in Fig. 3a,b), while a minority were present at frequencies almost an

223   order of magnitude greater, consistent with treatment resistance ("Group B" in Fig. 3a,b).

224   Comparison of the TMZ-treated secondary xenografts with the untreated primary xenograft

225   indicated that the sizes of sensitive clones were largely uncorrelated across serial passages,

226   whereas the sizes of the resistant clones appeared to be positively correlated (Fig. 3a).

227   Interestingly, the further coincidence of distinct resistant clones in drug-treated replicate

228   xenografts (Figs. 3c,d) suggests that the resistance phenotype can be pre-existing within the

229   parental population.

230   Based on this classification, we analysed the clone size distribution within each group

231   separately. Sensitive clones maintained an approximate negative binomial dependence (Extended

232   Data Fig. 5a) suggesting that, in sharp contrast with the mouse model[3], TMZ-treatment leaves

233   the proliferative hierarchy of the majority of tumour cells unperturbed. In contrast, resistant

234   clones could not be captured by the same dynamics (Extended Data Fig. 5a, red arrowheads).

235   However, with an additional acquired resistance to apoptosis, we found that the original model

236   parameters were sufficient to explain the scale of the observed behaviours of resistant clones

237   (compare Fig. 3a to Fig. 3e and Fig. 3b to Fig. 3f, Supplementary Theory 6.5). Importantly, large

238   outlier clones can be detected even in untreated tumours across different GBM cases (Extended

239   Data Figs. 5-6). Taken together, these results demonstrate that a minority of clones in pre- and

240   post-treatment tumours conform to perturbed growth dynamics, and may constitute a key driver

241   in the clonal evolution of human GBM. We define these outliers as "Group B" clones, and the

242   majority that behave according to the negative binomial distribution as "Group A".

243

244   **Epigenetic targeting of distinct GSCs**

245   We next questioned whether the Group B phenotype exposes new therapeutic

246   vulnerabilities. Primary GSC cultures[26] established from xenografts maintained a mixture of

247   clones seen in primary, secondary and tertiary passages (Extended Data Fig. 7a-b). Moreover,

248   both cultures and xenografts derived from the same parental TMZ-treated xenograft were

249   relatively concordant in their relative clonal abundances (Extended Data Fig. 7c), suggesting that

250   GSC cultures can recapitulate their growth behaviour *in vivo*. Strikingly, Group A clones from

251   the GBM-754 primary xenograft-derived culture model $(1)_{754}$, for which the most data was

252   available, maintained a negative binomial distribution after an approximate 7-fold expansion *in*

253   *vitro*, consistent with maintenance of the proliferative hierarchy under culture conditions

254   (Extended Data Fig. 7d,e). This included the correlations of outlier clones between replicates

255   (Extended Data Fig. 7f), corroborating the previously observed presence of Group B clones in

256   untreated xenografts (Fig. 2b). Most cultures derived from other xenografts also adhered to a

257   negative binomial distribution once the largest outliers were removed (Extended Data Fig.8a-b).

258   We next combined *in vitro* drug selection of the $(1)_{754}$ culture with barcode sequencing to

259   determine whether resistance arises proportionately from each clone type (Extended Data Fig.

260   9a,b). GSC cultures analyzed by assay for transposase-accessible chromatin with high-

261   throughput sequencing (ATAC-seq) identified a shared epigenetic state, leading us to focus on

262   epigenetic targets (Extended Data Fig. 2d). Cells subjected to drug selection were allowed to

263   repopulate to a similar density as control, in order to model tumour regrowth following therapy

264   (Extended Data Fig. 9b). The drug treatments induced a range of changes to clonal dominance

265   patterns (Extended Data Fig. 9c). However, the same negative binomial distribution was

266   maintained in most cases, indicating that the underlying dynamics of Group A clones are largely

267   unperturbed (Extended Data Fig. 10a,b). Intriguingly, a Menin-Mixed Lineage Leukemia (MLL)

268   interaction inhibitor (MI-2-2)[27-29] was selective against Group B clones, as repopulation

269   following selection derived primarily from Group A clones (Fig. 3g, Extended Data Fig. 9d). By

270 the same logic, and consistent with the requirement for Enhancer of zeste homolog 2 (EZH2) in

271 GSC maintenance[30], we found that an EZH2 inhibitor (UNC1999) was instead selective against

272 Group A clones (Fig. 3g, Extended Data Fig. 9d). MI-2-2 is growth inhibitory in a polyclonal

273 context, consistent with its specificity for the highly proliferative clone type (Extended Data Fig.

274 9e). Targeting both clone types by combining MI-2-2 with an EZH2 inhibitor (UNC1999 or

275 GSK343) was uniquely sufficient to eradicate self-renewal (Fig. 3h, Extended Data Fig. 9f-h).

276 Consistent with TMZ-induced selection for Group B clones in GBM-719, MI-2-2 treatment of

277 TMZ-transformed cells eradicated self-renewal and reduced tumour growth *in vivo* (Fig. 3i-j,

278 Extended Data Fig. 9i). Efficacy of the UNC1999/MI-2-2 combination was mirrored in 4

279 additional models (G523, G549, G564, G566) even when single drug treatments did not affect

280 self-renewal, and in GBM-851 primary cells (Extended Data Fig. 9j-n). While Menin-MLL

281 inhibition is especially effective in targeting paediatric glioma that carry histone 3 variant H3.3

282 mutations[27], these findings warrant further pre-clinical studies of MI-2-2 in advanced, post-

283 treatment adult GBM.

284

285 **Discussion**

286   Efforts to define the identity and behaviour of tumour-maintaining cells in human GBM

287 have focused on genetic intra-tumoural heterogeneity[9,31]. Yet the majority of subclonal mutations

288 in cancer may be biologically neutral[14,32]. At first sight, the emergence of clonal heterogeneity

289 suggests that the evolving mutational landscape may confer a range of fitness advantages on

290 GSCs. However, quantitative analysis of clone sizes indicates that clonal heterogeneity can be

291 explained by robust features of a conserved proliferative hierarchy. In this model, heterogeneity

292 in clonal expansion does not derive from genetic diversity but, in common with other cancer

293   models[11,12], emerges as the predictable outcome of chance fate decisions made by GSCs and

294   their progeny. Given the correlation of human GBM cell transcriptomes with those of normal

295   outer radial glial cells and intermediate progenitors[33], these results suggest that the initiation of

296   human GBM may be associated with the aberrant reactivation of a surprisingly normal

297   developmental program.

298   While the majority of GSC clones adhere to neutral, hierarchical growth dynamics

299   (Group A), we identified a minority subset that showed a different growth characteristic (Group

300   B). It is currently unknown whether Group B clones share common molecular features between

301   different patient tumours. Intriguingly, however, these dominant clones are sensitive to an

302   epigenetic drug (MI-2-2) previously shown to be effective in H3.3 mutant paediatric

303   glioblastoma[27]. Together with the fact that adult GSCs can converge into an epigenetic state

304   reminiscent of paediatric GBM due to selective downregulation of H3.3 expression[29], it is

305   tempting to speculate that Group B clones in adult GBM may share additional epigenetic features

306   of H3.3 mutant paediatric GBM cells and H3.3-low adult GSCs[29]. Alternatively, Group B clones

307   may arise from Group A clones after a gradual accumulation of genetic mutations that alters their

308   mode of growth[7]. Future studies should target the origin and functional properties of these

309   clones, and assess whether they contribute disproportionately to GBM malignancy.

310

311

312

313

314

315

References

1   Stupp, R. *et al.* Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* **352**, 987-996, doi:10.1056/NEJMoa043330 (2005).

2   Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396-401, doi:10.1038/nature03128 (2004).

3   Chen, J. *et al.* A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488**, 522-526, doi:10.1038/nature11287 (2012).

4   Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).

5   Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309-313, doi:10.1038/nature20123 (2016).

6   Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* **110**, 4009-4014, doi:10.1073/pnas.1219747110 (2013).

7   Johnson, B. E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189-193, doi:10.1126/science.1239947 (2014).

8   Greaves, M. Cancer stem cells: back to Darwin? *Semin Cancer Biol* **20**, 65-70, doi:10.1016/j.semcancer.2010.03.002 (2010).

9   Piccirillo, S. G. *et al.* Genetic and functional diversity of propagating cells in glioblastoma. *Stem Cell Reports* **4**, 7-15, doi:10.1016/j.stemcr.2014.11.003 (2015).

10   Snuderl, M. *et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* **20**, 810-817, doi:10.1016/j.ccr.2011.11.005 (2011).

11   Driessens, G., Beck, B., Caauwe, A., Simons, B. D. & Blanpain, C. Defining the mode of tumour growth by clonal analysis. *Nature* **488**, 527-530, doi:10.1038/nature11344 (2012).

12   Sanchez-Danes, A. *et al.* Defining the clonal dynamics leading to mouse skin tumour initiation. *Nature* **536**, 298-303, doi:10.1038/nature19069 (2016).

13   Rulands, S. & Simons, B. D. Tracing cellular dynamics in tissue development, maintenance and disease. *Curr Opin Cell Biol* **43**, 38-45, doi:10.1016/j.ceb.2016.07.001 (2016).

14   Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc Natl Acad Sci U S A* **113**, 128-133, doi:10.1073/pnas.1516123113 (2016).

15   Nguyen, L. V. *et al.* Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell* **14**, 253-263, doi:10.1016/j.stem.2013.12.011 (2014).

16   Nguyen, L. V. *et al.* Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. *Nature* **528**, 267-271, doi:10.1038/nature15742 (2015).

17   Bhang, H. E. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* **21**, 440-448, doi:10.1038/nm.3841 (2015).

18   Wagenblast, E. *et al.* A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis. *Nature* **520**, 358-362, doi:10.1038/nature14403 (2015).

19   Vanner, R. J. *et al.* Quiescent sox2(+) cells drive hierarchical growth and relapse in sonic hedgehog subgroup medulloblastoma. *Cancer Cell* **26**, 33-47, doi:10.1016/j.ccr.2014.05.005 (2014).

361 20 Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of
362   glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.
363   *Cancer Cell* **17**, 98-110, doi:10.1016/j.ccr.2009.12.020 (2010).
364 21 Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-
365   477, doi:10.1016/j.cell.2013.09.034 (2013).
366 22 Nguyen, L. V. *et al.* DNA barcoding reveals diverse growth kinetics of human breast
367   tumour subclones in serially passaged xenografts. *Nat Commun* **5**, 5871,
368   doi:10.1038/ncomms6871 (2014).
369 23 Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat Genet* **47**,
370   209-216, doi:10.1038/ng.3214 (2015).
371 24 Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit
372   intratumour heterogeneity. *Nature* **525**, 261-264, doi:10.1038/nature14971 (2015).
373 25 Bailey, N. T. J. *The Elements of Stochastic Processes with Applications to the Natural*
374   *Sciences*. (John Wiley & Sons, 1990).
375 26 Pollard, S. M. *et al.* Glioma stem cell lines expanded in adherent culture have tumor-
376   specific phenotypes and are suitable for chemical and genetic screens. *Cell Stem Cell* **4**,
377   568-580, doi:10.1016/j.stem.2009.03.014 (2009).
378 27 Funato, K., Major, T., Lewis, P. W., Allis, C. D. & Tabar, V. Use of human embryonic
379   stem cells to model pediatric gliomas with H3.3K27M histone mutation. *Science* **346**,
380   1529-1533, doi:10.1126/science.1253799 (2014).
381 28 Borkin, D. *et al.* Pharmacologic inhibition of the Menin-MLL interaction blocks
382   progression of MLL leukemia in vivo. *Cancer Cell* **27**, 589-602,
383   doi:10.1016/j.ccell.2015.02.016 (2015).
384 29 Gallo, M. *et al.* MLL5 Orchestrates a Cancer Self-Renewal State by Repressing the
385   Histone Variant H3.3 and Globally Reorganizing Chromatin. *Cancer Cell* **28**, 715-729,
386   doi:10.1016/j.ccell.2015.10.005 (2015).
387 30 Suva, M. L. *et al.* EZH2 is essential for glioblastoma cancer stem cell maintenance.
388   *Cancer Res* **69**, 9211-9218, doi:10.1158/0008-5472.can-09-1622 (2009).
389 31 Meyer, M. *et al.* Single cell-derived clonal analysis of human glioblastoma links
390   functional and genomic heterogeneity. *Proc Natl Acad Sci U S A* **112**, 851-856,
391   doi:10.1073/pnas.1320611111 (2015).
392 32 Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification
393   of neutral tumor evolution across cancer types. *Nat Genet* **48**, 238-244,
394   doi:10.1038/ng.3489 (2016).
395 33 Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical
396   development. *Cell* **163**, 55-67, doi:10.1016/j.cell.2015.09.004 (2015).
397
398
399
400
401
402

426 Merck, Novartis, the government of Ontario, Pfizer, Takeda, and the Wellcome Trust.

427

428 Author Contributions

429 X.L. and P.B.D. conceptualized the study and were assisted by L.V.N., R.J.V., N.I.P., F.J.C.,

430 H.J.S., M.G., and C.J.E. in experimental design. P.B.D. and B.D.S. supervised the study. X.L.

431 performed *in vivo* and *in vitro* barcoding experiments and drug validation studies. D.J.J., B.D.S.,

432 X.L., D.P., A.C., and P.B.D. analysed and interpreted barcoding results. D.J.J. and B.D.S.

433 developed the theoretical model of tumour growth, performed simulations and wrote the

434 supplemental theory section. F.M.G.C., L.M.R., M.D.T., and T.J.P. analysed WES and RNA-seq

435 results. P.G. and M.L. performed ATAC-seq and analysed results. R.J.V., L.L., M.K., N.I.P.,

436 F.J.C., H.W., C.C., B.L., N.R., R.H., and S. Dolma assisted in performing the experiments.

437 M.M., A.J.M., R.A.M., Y.M., and M.H. oversaw the generation of sequencing data. L.V.N. and

438 C.J.E. designed, generated, and validated the barcode library. P.P. and C.H.A. assisted with *in*

439 *vitro* drug assays. M.D.C., S. Das, M.B. contributed all GBM tumour samples used in the study.

440 X.L., D.J.J., C.J.E., B.D.S., and P.B.D. wrote the manuscript, all authors contributed to data

441 interpretation and approved the manuscript.

442

443 Author Information

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

445 declare no competing financial interests. Readers are welcome to comment on the online version

446 of the paper. Correspondence and requests for materials should be addressed to P.B.D.

447 (peter.dirks@sickkids.ca) and B.D.S. (bds10@cam.ac.uk).

448
449

450

451

452 **Figure Legends**

453 **Figure 1** | **Serial transplantation scheme and characterization of barcoded glioblastoma**

454 **xenografts.**

455 **a**, General transplantation scheme for barcoded xenografts derived from primary GBM tumour

456 cells (GBM-719). **b**, Staining of a secondary GBM-719 xenograft with the indicated markers,

457 scale bar = 100 μm. **c**, Tumour growth quantified as the estimated fold-change in cell number

458 between injection and harvesting for different ipsilateral derived GBM-719 xenografts. Lines

459 indicate serial transplantation trajectories. **d**, Proportional Venn diagrams depicting the number

460 of barcoded clones unique to each passage or shared between passages for the indicated

461 experiment.

462 **Figure 2** | **Clonal dynamics of GBM is consistent with a conserved proliferative hierarchy.**

463 **a**, Clone size distributions of xenografts derived from GBM-719 cells across different passages.

464 For the primary passage, distributions for the ipsilateral (blue) and contralateral sides (red) are

465 shown. For the secondary and tertiary passages, distributions for the ipsilateral side from

466 different replicate experiments are shown (shades of blue). **b**, First incomplete moment of the

467 corresponding clone size distributions shown in panel (**a**), displayed on a logarithmic scale

468 (Supplementary Theory 2). Dashed lines show exponentials as a guide for the eye. The red

469 arrowhead indicate deviations from exponential behaviour due to a small number (<4%) of

470 outlier clones. **c**, A minimal model of tumour growth based on a three-component hierarchy

471 involving transitions from a slow-cycling stem-like compartment (S) to a more rapidly cycling

472 progenitor population (P) to a non-dividing compartment (D). Following S cell divisions, a

473   fraction, ε, result in symmetric fate outcome while the remainder lead to asymmetric fate. With

474   equal probability, P cells divide symmetrically or give rise to D cells which, in turn, rapidly

475   undergo apoptosis. **d**, Representative clone size trajectories computed for the model shown in

476   (**c**). Different curves correspond to different clones across three serial passages, along with the

477   average over all trajectories, with the S cell division rate of 0.15/ day, the P cell division rate of

478   1/day, the D cell apoptosis rate of 0.5/day and $\varepsilon = 15\%$ (for details, see Supplementary Theory

479   5). **e**, First incomplete moment of the clone size distribution across passages derived from 2×10^6

480   simulated clone trajectories. The shaded areas show the regions within which 95% of the

481   respective curves fall for repeated simulations with 5×10^4 clones each. For each passage, the first

482   incomplete moment follows an approximate exponential size dependence. Parameters as in panel

483   (**d**). **f**, Clone size correlation for different passages in the model (distributions) and from

484   representative xenografts derived from GBM-719 cells (data points). Distributions show model

485   results within the biologically plausible parameter range (see Supplementary Theory, Table S2).

486   See Supplementary Theory, Figure S3 for other patients. **g**, Fraction of initially injected clones

487   growing above half of the characteristic clone frequency $n_0/2$ for the same datasets as in (**f**) (see

488   Supplementary Theory 6.3). See Supplementary Theory, Figure S2 for other patients. **h**,

489   Simulated examples of clone size correlations across successive serial passages. Parameters are

490   as in panel (**d**).

491   **Figure 3 | Chemotherapy reveals clonal transformations in GBM.**

492   **a**, Correlation of clone sizes for the primary, untreated xenograft with secondary xenografts

493   treated with TMZ (light and dark dots indicate two replicate secondary xenografts). Light dataset

494   – Group A: 1255 data points, Group B: 15 data points; dark dataset – Group A: 1228 data points,

495   Group B: 10 data points. **b**, Correlation of clone sizes for a secondary TMZ-treated xenograft

496   (light dots in panel (**a**)) with tertiary TMZ-treated xenografts, light and dark dots indicate two

497   replicate tertiary xenografts. Light dataset – Group A: 95 data points, Group B: 15 data points;

498   dark dataset – Group A: 117 data points, Group B: 15 data points. **c**, Correlation of the two

499   replicate secondary xenografts shown in (**a**) with Spearman's rho indicated. **d**, Correlation of the

500   two replicate tertiary xenografts shown in (**b**) with Spearman's rho indicated. **e-f**, Correlation of

501   clone sizes obtained from simulations with a subset of clones being resistant to cell death (blue

502   dots) and the remaining clones following unperturbed dynamics (green dots) for a primary and

503   secondary passage (**e**) and a secondary and tertiary passage (**f**) (see Supplementary Theory 6.5).

504   The S cell division rate is set at 0.1/day, the P cell division rate is 1.5/day, $\varepsilon = 10\%$, and the

505   apoptosis rate is set at 0.7/day with a 0.5% chance of each clone to show resistance to apoptosis

506   (see Supplementary Theory, Table S3). **g**, Selectivity of UNC1999 and MI-2-2 for group A and

507   B clones respectively, representative of 2 technical replicate experiments. Shown are relative

508   clone sizes after DMSO treatment, or regrowth following selection with indicate compounds.

509   The indicated values are clone sizes for groups A (black) and B (blue), lines connect the same

510   barcoded clone under different conditions. **h**, Reduction of self- renewal ability upon treatment

511   with epigenetic compounds alone and in combination as assessed by limiting dilution analysis

512   (LDA), representative of 3 independent experiments (MI-nc: inactive control for MI-2-2, M: MI-

513   2-2, C: CI-994, G: GSK591, U: UNC1999). $P = 0.0663$ for DMSO vs. CI-994, 0.132 for DMSO

514   vs. GSK591, 0.216 for DMSO vs. UNC1999, 5.74×10^{-13} for DMSO vs. MI-2-2, 4.11×10^{-18} for

515   MI-nc vs. M, 1 for M vs. M+C, 0.432 for M vs. M+G, 8.53×10^{-8} for M vs. M+U. **i**, MI-2-2

516   abrogates self-renewal in TMZ-transformed GBM-719 population, representative of 3

517   independent experiments. $P = 3.73\times10^{-3}$ for DMSO vs. UNC1999, 1.16×10^{-27} for DMSO vs MI-

518   2-2, 1.61×10^{-16} for UNC1999 vs MI-2-2. All LDA results are representative of 3 independent

519   experiments with the remaining experiments presented in Extended Data Fig. 9. Analysis of all

520   LDA results was performed using ELDA software[34], error bars represent 95% confidence

521   interval (ns $P > 0.05$, * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$). **j**, MI-2-2 inhibits tumour growth

522   in subcutaneous xenografts derived from TMZ-transformed GBM-719 cells, n = 9 tumours per

523   group, two-sided unpaired t-test. The horizontal line indicates the mean tumour weight of each

524   experimental group.

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543
544
545   **Methods**

546   No statistical methods were used to predetermine sample size. For animal studies, all animals

547   were included for the analysis. Animals from separate litters were randomly and evenly divided

548   between experimental groups to control for animal age. The investigators were not blinded to

549   group allocation during the experiments and outcome assessment.

550

551   **Processing of patient samples**

552   GBM tumour samples were obtained from consenting patients, and all procedures are approved

553   by the Research Ethics Boards at The Hospital for Sick Children (Toronto, Canada), St.

554   Michael's Hospital (Toronto, Canada) and Toronto Western Hospital (Toronto, Canada).

555   Following surgical resection, tumour specimens were immediately subjected to mechanical and

556   enzymatic dissociation in artificial cerebrospinal fluid (aCSF) containing trypsin, hyaluronidase,

557   and kyneuric acid at 37°C. GSC culture models were established as previously described[26], and

558   matched to primary GBM tumour tissue by microsatellite genotyping (The Centre for Applied

559   Genomics, Hospital for Sick Children). GSC cultures were also randomly and intermittently

560   tested for mycoplasma contamination by PCR. For barcoding experiments, primary single-cell

561   suspensions were subjected to magnetic bead depletion to remove cells expressing human CD31

562   and CD45 markers (130-091-935, 130-045-801, Miltenyi Biotech), thereby excluding endothelial

563   and hematopoietic lineages prior to lentiviral barcoding.

564

565   **Exome sequencing**

566   For the primary tumour samples, DNA was extracted from flash frozen primary tumour pieces

567   using an AllPrep DNA/RNA Mini Kit (80204, Qiagen). Genomic DNA libraries from which

568   exons are captured were constructed according to British Columbia Cancer Agency Genome

569   Sciences Centre plate-based and paired-end library protocols on a Microlab NIMBUS liquid

570   handling robot (Hamilton, USA). Briefly, 1 µg of high molecular weight genomic DNA was

571   sonicated (Covaris LE220) in 62.5 µL volume to 250-350 bp. Sonicated DNA was purified with

572   PCRClean DX magnetic beads (Aline Biosciences). The DNA fragments were end-repaired,

573   phosphorylated and bead purified in preparation for A-tailing using a custom NEB Paired-End

574   Sample Prep Premix Kit (New England Biolabs). Illumina sequencing adapters were ligated

575   overnight at 16°C and adapter ligated products bead purified and enriched with 6 cycles of PCR

576   using primers containing a hexamer index that enables library pooling. 200 ng for each of 6

577   different libraries were pooled prior to whole exome capture using Agilent SureSelect All Exon

578   V6+UTR probes. The pooled libraries were hybridized to the RNA probes at 65°C for 24 hours.

579   Following hybridization, streptavidin-coated magnetic beads (Dynal, MyOne) were used for

580   exome capture. Post-capture material was purified on MinElute columns (Qiagen) followed by

581   post-capture enrichment with 6 cycles of PCR using primers that maintain the library-specific

582   indices. The pooled libraries were sequenced on Illumina Hiseq 2500 using V4 sequencing

583   chemistry at PE125 following Illumina recommendations (Canada's Michael Smith Genome

584   Sciences Centre, BC Cancer Agency).

585

586   For the GBM-719 xenograft samples, GFP positive barcoded cells were isolated by FACS

587   (MoFlo Astrios, Beckman Coulter) from *in vitro* expanded cells (p3 TMZ TMZ) or directly from

588   dissociated tumours (all remaining samples) and subjected to DNA extraction using a PrepGEM

589   DNA extraction kit (PTI0050, ZyGEM) prior to whole-genome amplification using a REPLI-g

590   Mini kit (150023, Qiagen). 200 ng of DNA per sample was used to generate cDNA libraries

591   following Agilent SureSelect XT target enrichment kit as per protocol. 750 ng from each cDNA

592   library was then hybridized for 24 hours using the All Exon V5 capture baits from Agilent.

593   Captured, enriched libraries were size validated using the Agilent Bioanalyzer DNA high

594   sensitivity chip and library concentration was validated by qPCR (Kapa Technologies). All

595   libraries were normalized to 10 nM and diluted to 2 nM before being denatured with 0.1N

596   NaOH. Denatured library pools were diluted for a final time down to 14 pM of pooled libraries

597   and loaded onto Illumina cBot for cluster generation. The clustered flow cell was sequenced

598   paired-end 100 cycles using an Illumina HiSeq 2000 (Princess Margaret Genomics Centre,

599   University Health Network).

600

601   For the germline reference sample, DNA was extracted from the patient's whole blood using a

602   DNeasy Blood & Tissue kit (69504, Qiagen). The library was prepared using Agilent SureSelect

603   Human Exome Library Preparation V4 kit for paired end sequencing on a HiSeq 2500 platform.

604   In brief, 750 ng of genomic DNA was fragmented to 200-bp on average using a Covaris LE220

605   instrument. Sheared DNA was end-repaired and the 3' ends adenylated prior to ligation of

606   adapters with overhang-T. Genomic library was amplified by PCR using 10 cycles and

607   hybridized with biotinylated probes that target exonic regions; the enriched exome libraries were

608   amplified by an additional 8 cycles of PCR. Exome libraries were validated on a Bioanalyzer

609   2100 DNA High Sensitivity chip (Agilent Technologies) for size and by qPCR using the Kapa

610   Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems) for quantities.

611   Exome libraries were pooled and sequenced with the TruSeq SBS sequencing chemistry using a

612 V4 high throughput flowcell on a HiSeq 2500 platform following Illumina's recommended

613 protocol (The Centre for Applied Genomics, Hospital for Sick Children).

614

615 **Exome sequencing analysis of primary tumours**

616 For the primary tumour samples, Fastq files were aligned to the human reference genome hg38

617 with BWA (0.7.9a, -M option)[35]. The BAM files were further processed using MarkDuplicates

618 (Picard Tools 2.6.0), indel realignment (GATK 3.6 RealignerTargetCreator and IndelRealigner)

619 and BaseRecalibration (GATK 3.6 BaseRecalibrator and PrintReads)[36]. Samtools 1.3.1 mpileup

620 (-B, -q10 -d10000000 options)[37] was run on the processed BAM files to generate the input to

621 Varscan. Varscan (2.4.2), mpileup2cns was applied to call snp and indels in each sample (--p-

622 value 0.01 --min-var-freq 0.03, other default parameters)[38]. The calls were annotated with

623 Annovar (20160201, using refGene genes)[39]. To identify the important somatic variants, the calls

624 were further filtered to include only the following annotated events: nonsynonymous_SNV,

625 stopgain, stoploss and frameshift_deletion. In addition, calls were removed if they were in the

626 dbSNP database[40] as part of the snp147Common file downloaded from the UCSC server which

627 contains uniquely mapped variants that appear in at least 1% of the population or are 100% non-

628 reference. Therefore, the flagged SNPs (uniquely mapped variants, excluding Common SNPs,

629 that have been flagged by dbSNP as "clinically associated") were not removed. In addition, calls

630 were further filtered out if they had an AF>0.001 in ExAC (exac03, ExAC_ALL)[41] or 1000

631 Genome Project (1000g2015aug_all)[42]. Subclonal mutations with variant allele frequency < 0.2

632 were excluded.

633

634 **Exome sequencing analysis of xenografts**

635   For the GBM-719 xenograft samples, read pairs were aligned to the hg19 reference sequence

636   using the Burrows-Wheeler Aligner (v0.7.12)[35], and samples were demultiplexed using Picard

637   tools (v1.140). Data were then sorted and duplicate marked using Picard and SAMtools[37]. Local

638   realignment around insertions or deletions (indels) and base-quality score recalibration was

639   performed using the Genome Analysis toolkit (v3.4-46)[36]. QualiMap (v2.1)[43] was used to

640   evaluate resulting sequencing alignment data. To correct for coverage discrepancies between

641   Agilent V4 (germline reference sample) and V5 (xenograft samples) capture baits, an

642   intersection of common regions was performed using bedtools (v2.26.0)[44]. Common regions with

643   0X coverage in the blood or greater than 500X coverage in either reference or xenografts were

644   removed from subsequent analysis.

645

646   The MuTect (v.1.15) algorithm[45] was used for somatic variant calling and false-positive filtering.

647   Resulting variants were annotated using Oncotator (v.2.8.0)[46], including common databased

648   variants (ClinVar[47], 1000 Genomes (phase 1 variant set)[48], dbSNP (build 138)[40], COSMIC

649   (v71)[49]). Germline variants found in the 1000 Genomes Project, dbSNP build 138 were

650   excluded. Cellularity, ploidy and allele-specific copy number was estimated from normal-

651   xenograft pairs using the Sequenza algorithm (v2.1.2)[50]. Cutoffs of log 2 copy number ratios

652   between -0.35 and +0.3 were set to assign genome losses and gains, respectively.

653
654   **RNA sequencing**

655   RNA was extracted from the same flash frozen primary tumour pieces as for exome sequencing

656   using a Qiagen AllPrep DNA/RNA Mini Kit (80204, Qiagen). Qualities of total RNA samples

657   were determined using an Agilent Bioanalyzer RNA Nanochip or Caliper RNA assay and

658   arrayed into a 96-well plate (Thermo Fisher Scientific). Polyadenylated (PolyA+) RNA was

659 purified using the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490L, NEB) from

660 500 ng total RNA normalized in 35 µL for DNase I-treatment (1 Unit, Invitrogen). DNase-

661 treated RNA was purified using RNA MagClean DX beads (Aline Biosciences, USA) on a

662 Microlab NIMBUS liquid handler (Hamilton Robotics, USA). Messenger RNA selection was

663 performed using NEBNext Oligod(T)$_{25}$ beads (NEB) with incubation at 65°C for 5 minutes

664 followed by snap-chilling at 4°C to denature RNA and facilitate binding of poly(A) mRNA to

665 the beads. mRNA was eluted in 36 µL of Tris Buffer.

666

667 First-strand cDNA was synthesized from the purified polyadenylated messenger RNA using the

668 Maxima H Minus First Strand cDNA Synthesis kit (Thermo-Fisher, USA) and random hexamer

669 primers at a concentration of 5 µM along with a final concentration of 1 µg/uL Actinomycin D,

670 followed by PCR Clean DX bead purification on a Microlab NIMBUS robot (Hamilton

671 Robotics, USA). The second strand cDNA was synthesized following the NEBNext Ultra

672 Directional Second Strand cDNA Synthesis protocol (NEB) that incorporates dUTP in the dNTP

673 mix, allowing the second strand to be digested using USER^{TM} enzyme (NEB) in the post-adapter

674 ligation reaction and thus achieving strand specificity.

675

676 cDNA was fragmented by Covaris LE220 sonication for 55 seconds at a "Duty cycle" of 20%

677 and "Intensity" of 5 to achieve 200-250 bp average fragment lengths. The paired-end sequencing

678 library was prepared following the BC Cancer Agency Genome Sciences Centre strand-specific,

679 plate-based library construction protocol on a Microlab NIMBUS robot (Hamilton Robotics,

680 USA). Briefly, the sheared cDNA was subject to end-repair and phosphorylation in a single

681 reaction using an enzyme premix (NEB) containing T4 DNA polymerase, Klenow DNA

682   Polymerase and T4 polynucleotide kinase, incubated at 20°C for 30 minutes. Repaired cDNA

683   was purified in 96-well format using PCR Clean DX beads (Aline Biosciences, USA), and 3' A-

684   tailed (adenylation) using Klenow fragment (3' to 5' exo minus) and incubation at 37°C for 30

685   minutes prior to enzyme heat inactivation. Illumina PE adapters were ligated at 20°C for 15

686   minutes. The adapter-ligated products were purified using PCR Clean DX beads, then digested

687   with USER™ enzyme (1U/µL, NEB) at 37°C for 15 minutes followed immediately by 13 cycles

688   of indexed PCR using Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) and

689   Illumina's PE primer set. PCR parameters: 98°C for 1 minute followed by 13 cycles of 98°C 15

690   seconds, 65°C 30 seconds and 72°C 30 seconds, and then 72°C 5 minutes. The PCR products

691   were purified and size-selected using a 1:1 PCR Clean DX beads-to-sample ratio (twice), and the

692   eluted DNA quality was assessed with Caliper LabChip GX for DNA samples using the High

693   Sensitivity Assay (PerkinElmer, Inc. USA) and quantified using a Quant-iT dsDNA High

694   Sensitivity Assay Kit on a Qubit fluorometer (Invitrogen) prior to library pooling and size-

695   corrected final molar concentration calculation for Illumina HiSeq 2500 sequencing with paired-

696   end 75 base reads (Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency).

697
698   **RNA sequencing analysis**

699   Fastq files were aligned with STAR (2.4.2a)[51] on the hg38 human reference genome. FPKM

700   values were computed with the DESeq2 fpkm function[52] using the raw read count per gene

701   (ReadsPerGene.out.tab file from STAR output), with size factor normalization and gene length

702   derived from the hg38 GTF files used for the alignment. Subgroup classification was done using

703   the simple GBM classifier[53]. This 32-gene classifier permits greater accuracy of GBM subgroup

704   classification when using RNA-seq data instead of gene expression microarrays, as was

705   performed in the original subgrouping study[20]. One of the 32 genes was not quantified in the

706   analysis so the classifier was run using 31 genes.

707

708   **Assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)**

709   The open chromatin profiles of 11 GSC lines were defined using ATAC-seq as described

710   previously[54] and the prepared libraries were sequenced with 50 bp single end reads. Reads were

711   mapped to hg19 using bowtie2[55] and peaks of open chromatin were called with MACS2[56]. The

712   correlation between samples was calculated as the Pearson correlation of the quantile-normalized

713   signal across the peak catalogue. Here, the peak catalogue corresponds to all peak regions

714   identified across the sample cohort, and the signal refers to the fold enrichment of the signal per

715   million reads in a sample over a modelled local background. The chronic lymphocytic leukaemia

716   (CLL) data used in this comparison was taken from a published dataset[57], and the raw signal was

717   normalized together with the GSC cohort.

718

719   **MGMT promoter methylation assay**

720   Primary tumour DNA was subjected to bisulfite conversion using the EZ DNA Methylation-

721   Gold Kit (D5005, Zymo Research), and MGMT promoter methylation status was assessed using

722   a two-step PCR protocol as previously described[58]. PCR products, including water control, were

723   visualized by electrophoresis on a 2% agarose gel along with a 100 base pair ladder (NEB).

724

725   **Lentiviral barcoding**

726   The lentiviral barcode library has been described previously[15]. For viral transduction, 1×10^6

727   primary GBM cells were seeded per plate onto 10-cm cell culture dishes that are coated with

728   poly-L-ornithine (PLO, Sigma) and laminin (Sigma). The culture media consisted of serum-free

729   Neurocult NS-A Basal (Stemcell Technologies) media, supplemented with 2 mmol/L L-

730   gutamine, N2 and B27 supplements, 75 μg bovine serum albumin, 10 ng/mL recombinant human

731   EGF (rhEGF), 10 ng/mL basic fibroblast growth factor (bFGF), and 2 μg/mL heparin (Sigma)[26].

732   Primary cells were incubated for approximately 12 hours at 37°C with lentivirus at an

733   appropriate concentration to minimize multiple integration events. The concentration of

734   lentivirus used was previously determined by titrating the library with a human fetal derived

735   neural stem cell culture (HF7450), and assessing GFP positivity by flow cytometry (LSR II, BD

736   Biosciences) 48 hours post-transduction. Barcoded cells were washed 5 times with PBS to

737   remove remaining lentivirus, and immediately harvested by accutase (Sigma) treatment for

738   orthotopic injection. A separate cell aliquot was cultured for 48 hours to allow for GFP

739   expression, and transduction efficiency was determined by flow cytometry (LSR II, BD

740   Biosciences).

741

742   **Mouse xenografts**

743   All mouse procedures were approved by The Hospital for Sick Children's Animal Care

744   Committee. For intracranial injections, animals were first anesthetized with isoflurane and given

745   Ketoprofen as an analgesic. Tumour cells were then suspended in a 2 μl volume of PBS and

746   injected in the forebrains of female NOD/SCID/IL-2γ[-/-] (NSG) mice of age 1-3 months with a

747   Hamilton syringe and stereotactic device. The coordinates for orthotropic injections are 4 mm

748   anterior of lambda, 2 mm to the right of the midline, and 3 mm deep. For secondary and tertiary

749   xenografts, 25 mg/kg TMZ (Sigma) solubilized in Cremophor or vehicle controls were

750   administered by gastric gavage for 5 consecutive days, 10 days post-injection. Mice were

751   sacrificed for further processing once neurological symptoms are observed, or at experiment

752   endpoint (6 months). Survival analysis was performed using GraphPad Prism 5 software.

753

754   **Processing of xenografts**

755   Forebrains were obtained from animals displaying neurological symptoms, and the two

756   hemispheres (ipsilateral and contralateral) were dissected for processing separately. Each

757   hemisphere was dissociated to single-cell suspensions as described in the "processing of patient

758   samples" section. Cells were subsequently subjected to magnetic bead depletion to remove

759   contaminating mouse cells (130-104-694, Miltenyi Biotech) prior to serial transplantation. Serial

760   xenografts were always established without any intermediate culturing step. Either the ipsilateral

761   or contralateral fraction from a single mouse was used to establish serial xenografts.

762   Approximately 15% of xenograft cells were used without magnetic bead depletion for PCR

763   amplification, library preparation and deep amplicon sequencing of barcodes. One xenograft per

764   experimental group was set aside for histological analysis. Splinkerette PCR according to a

765   previously published protocol[59] was performed in order to identify unique barcode vector

766   integration sites from xenografts.

767

768   **Histopathology and immunohistochemistry**

769   Mouse brains were fixed in 4% paraformaldehyde (PFA), washed in 70% ethanol and paraffin

770   embedded. 6 μm coronal sections were generated for further analysis. Haematoxylin and Eosin

771   staining was carried out according to manufacturer's instructions (MHS32-1L, Sigma-Aldrich

772   and 6766009, Thermo Scientific). Antibodies for immunohistochemistry include anti-Nestin

773   (MAB5326, Millipore used at 1:500), anti-Ki-67 (M7240, Dako used at 1:500) and anti-Cleaved

774   Caspase-3 (9661, Cell Signaling used at 1:500). A secondary anti-Mouse HRP antibody (A9044,

775   Sigma 1:500) was used for detection using 3,3'-diaminobenzidine (DAB), Alkaline Phosphatase

776   (AP) and Mouse on Mouse (M.O.M) detection kits (Vector Laboratories). Images were acquired

777   using a 3DHistech Pannoramic 250 Flash II Slide Scanner and processed using Pannoramic

778   Viewer software (3DHISTECH). Automatic detection and quantification of Ki-67 and Cleaved

779   Caspase-3 staining was performed on six representative images per sample, using TMARKER

780   software[60].

781

782   **Barcode sequencing**

783   Spiked-in controls were generated using a human fetal derived neural stem cell line (HF7450)

784   using the previously described protocol[15], and combined into single wells of a 96-well plate. For

785   the GBM-719 experiment and the first sequencing run, the cell numbers used as spiked-in

786   controls were 10, 100, 250, 500, and 5000. For all subsequent *in vivo* experiments and the second

787   sequencing run, the cell numbers used were 10, 100, and 5000. For all *in vitro* experiments and

788   the third sequencing run, the cell numbers used were 10, 100, 500, and 5000. Separate spiked-in

789   control only wells containing barcode sequences derived from 25,000 and 100,000 cells were

790   also included in the GBM-719 experiment, to test accuracy of extrapolation for larger clones.

791   The same was done in the third sequencing run for *in vitro* experiments, using a control of

792   50,000 cells. Xenograft samples were combined with spiked-in controls and subjected to DNA

793   extraction using a PrepGEM DNA extraction kit (PTI0050, ZyGEM) followed by ethanol

794   precipitation and deep amplicon sequencing as described previously[15]. Briefly, a two-step PCR

795   protocol was used to generate barcode amplicons with fault-tolerant sample indices, and

796   equimolar samples were pooled and loaded onto a single lane of a flow cell for paired-end

797 sequencing on the Illumina MiSeq platform (Canada's Michael Smith Genome Sciences Centre,

798 BC Cancer Agency).

799

800 **Barcode data analysis**

801 Barcode sequences were extracted from raw data files with custom scripts, and those with a

802 minimum base quality of 20 that matched the flanking regions (with up to 3 mismatches)

803 surrounding the barcode sequence were kept. A merging of highly similar barcodes was

804 performed in order to limit the number of false positive barcode sequences that may arise from

805 sequencing errors[61]. Specifically, a list of read counts corresponding to all unique barcode

806 sequences was generated, and read counts corresponding to sequences with up to three

807 mismatches were combined into the most abundant sequence. Barcode sequence logograms were

808 generated using the R package ggseqlogo (https://github.com/omarwagih/ggseqlogo). Spiked-in

809 controls were retrieved for defining noise thresholds and clone size estimation as described

810 previously[16,22]. We defined fractional read value (FRV) as the read count for a particular barcode

811 sequence divided by the sum read counts of all spiked-in controls in the sample. A relationship

812 was generated between FRVs and control cell number for spiked-in controls across all samples.

813 A Cook's distance of 4/n was used to define outlier controls and the relationship was generated

814 again with those outliers removed to estimate clone sizes. This step was performed to ensure that

815 outlier controls do not influence the estimation of relative clone sizes in the majority of samples

816 within a particular sequencing run. FRV thresholds were determined from spiked-in controls in

817 order to maximize the difference between the true positive rate (TPR) and false positive rate

818 (FPR), and only clones with FRVs greater than the threshold were kept. The total cell number for

819 each sample was estimated by summing up estimated cell numbers for each clone in the sample

820　that are above detection threshold. Relative clone sizes were then determined by dividing the cell

821　numbers for each clone by the total cell number calculated for each sample. Proportional Venn

822　diagrams for barcode sequences were generated with eulerAPE v3 software[62].

823

824　**Generation of xenograft-derived cultures**

825　Dissociated primary GBM xenografts were cultured as described in the "lentiviral barcoding"

826　section after depletion of contaminating mouse cells (130-104-694, Miltenyi Biotech). All short-

827　term cultures were subjected to 2 to 3 passages prior to barcode sequencing. Short-term cultures

828　were not subjected to mycoplasma testing or microsatellite genotyping, although in all cases the

829　identified barcode sequences of cultures matched those of the corresponding xenograft series.

830

831　**Cell culture assays**

832　For proliferation assays, GSCs were propagated for 11 days in triplicate under previously

833　described conditions[26]. Viable cells were counted on days 0, 2, 4, 7, 9, and 11 with a Countess

834　Automated Cell Counter (Thermo Fischer Scientific), excluding apoptotic cells that stained

835　positive for trypan blue (Thermo Fischer Scientific). Doubling times were calculated during

836　exponential growth phase (between days 4 and 11) using the formula $t/\log_{10}2 \times \log_{10}(N_{t2}/N_{t1})$,

837　where N_{t1} and N_{t2} are the number of cells on days 4 and 11 respectively and t is the elapsed time

838　in hours. For dosage response assays, GSCs were cultured with drug for 5 days with 6 technical

839　replicates per dose, without any media changes. Cell viability relative to DMSO control was then

840　assessed by AlamarBlue assay (Thermo Fisher Scientific) using a Gemini EM Fluorescence

841　Microplate Reader (Molecular Devices).

842

843   **Drug screening**

844   Primary drug screens were carried out in 96-well format on passage 2-3 cultures that were grown

845   under previously described conditions[26]. An Incucyte Zoom live-cell analysis system (Essen

846   Bioscience) was used to quantify confluency according to manufacturer's instructions. In order

847   to characterize drug responsiveness of barcoded clones, a second screen was performed where

848   cells were seeded on 6-well plates, subjected to a single round of drug selection in duplicate, and

849   harvested for barcode sequencing when the culture reached approximately the same confluency

850   as DMSO controls (~90%). In this assay, culture media was refreshed every 3 days without drug.

851   The concentrations of drugs used for screening were as follows: Rapamycin: 20 nM, Dasatinib:

852   125 nM, BIO; Daunorubicin: 1 μM, LGK-974; RO4929097; WP1066: 2 μM, Imatinib: 2.5 μM,

853   Bromosporine; CI-994; GSK591; GSK-J4; GSK-LSD1; InSolution γ-Secretase Inhibitor X;

854   IOX2; JQ-1; L-741,742; LAQ824; MI-2-2; MS023; OF-1; Olaparib; PFI-1; PNU96515E; SGC-

855   CBP30; UNC1999: 5 μM, Erlotinib: 10 μM, TMZ: 50 μM. Once ~90% confluency is reached,

856   all surviving cells were used for DNA extraction and barcode sequencing as described above.

857

858   **Limiting dilution analysis (LDA)**

859   Cells were plated onto Flat bottom 96 well plates (Sarstedt) in 100 μL of culture media, 6

860   replicates per cell dose. The culturing conditions are described previously[26], with the exception

861   that culture plates were not coated with PLO and laminin to allow for sphere formation. For

862   analysis of primary, uncultured GBM cells, two-fold dilutions from 4000 cells to 8 cells were

863   used and scored after two weeks of culture. For analysis of established GSC cultures, two-fold

864   dilutions from 2000 cells to 4 cells were used and scored after one to two weeks of culture.

865   Drugs were added only once on the first day at either 1 μM or 5 μM as indicated for each

866   experiment, with 50 μL of fresh media added to each well after the first week. Investigators were

867   blinded to the label for each plate during data collection. Data were analyzed using ELDA

868   software[34].

869

870   ***In vivo* drug assay**

871   To test the effect of MI-2-2 treatment *in vivo* on tumour growth, 200,000 $(1,1^T,1^T)_{719}$ cells were

872   transplanted subcutaneously into the flanks of NSG mice (6 mice per treatment group, total 12

873   mice) and allowed to grow for 1 week prior to drug treatment. Mice were then treated with either

874   20 mg/kg MI-2-2 (444825, Millipore) or vehicle control (15% DMSO, 25% PEG, 60% PBS) for

875   2 weeks by intraperitoneal injection. The treatment schedule was Monday, Wednesday, Friday of

876   each week for a total of 6 treatments. Mice were then monitored for tumour formation and

877   sacrificed once the control tumours reached endpoint for measurement (127 days between

878   injection and sacrifice). Flanks in which tumours were not visible were excluded from analysis.

879   Subcutaneous tumour size did not exceed the limit set by the experimental protocol with The

880   Hospital for Sick Children's Animal Care Committee (17 mm in the longest dimension).

881

882   **Stochastic simulations**

883   A standard stochastic simulation algorithm[63] was used to simulate realizations of the stochastic

884   process defined by the model shown in Fig. 2c and described fully in Supplementary Theory

885   section 5. Clone size distributions, clone size cross correlations and the ratio of surviving clones

886   were then calculated from 100,000 realizations of the system for each parameter set. To compare

887   the model with experiments, we simulated the system using 108 equidistant parameter sets

888   located in the region of biologically plausible parameters and compared the results to

889   experimental data points.

890

891   **Data availability**

892   ATAC-seq data have been deposited at the Gene Expression Omnibus

893   (http://www.ncbi.nlm.nih.gov/geo) under the accession number GSE96088. WES and RNA-seq

894   data have been deposited at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega)

895   under the accession number EGAS00001002424. All other data are available as Supplementary

896   Data Tables, Source Data, or upon reasonable request from the corresponding authors (P.B.D.

897   and B.D.S.).

898

899   **Code availability**

900   Code used throughout this study are available upon reasonable request from the corresponding

901   authors (P.B.D. and B.D.S.).

902

903

904

905

906

907

908

909

910

911

912

913

914

915

Supplementary References

916
917
918

919   34   Hu, Y. & Smyth, G. K. ELDA: extreme limiting dilution analysis for comparing depleted
920   and enriched populations in stem cell and other assays. *J Immunol Methods* **347**, 70-78,
921   doi:10.1016/j.jim.2009.06.008 (2009).
922   35   Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
923   transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
924   36   McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
925   analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303,
926   doi:10.1101/gr.107524.110 (2010).
927   37   Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
928   2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
929   38   Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery
930   in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111
931   (2012).
932   39   Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
933   variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164,
934   doi:10.1093/nar/gkq603 (2010).
935   40   Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**,
936   308-311 (2001).
937   41   Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*
938   **536**, 285-291, doi:10.1038/nature19057 (2016).
939   42   Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74,
940   doi:10.1038/nature15393 (2015).
941   43   Garcia-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment
942   data. *Bioinformatics* **28**, 2678-2679, doi:10.1093/bioinformatics/bts503 (2012).
943   44   Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
944   genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033
945   (2010).
946   45   Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and
947   heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).
948   46   Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum Mutat* **36**, E2423-
949   2429, doi:10.1002/humu.22771 (2015).
950   47   Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation
951   and human phenotype. *Nucleic Acids Res* **42**, D980-985, doi:10.1093/nar/gkt1113 (2014).
952   48   Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes.
953   *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
954   49   Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of
955   Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-950, doi:10.1093/nar/gkq929
956   (2011).

957 50   Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor
958   sequencing data. *Ann Oncol* **26**, 64-70, doi:10.1093/annonc/mdu479 (2015).
959 51   Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21,
960   doi:10.1093/bioinformatics/bts635 (2013).
961 52   Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
962   for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8
963   (2014).
964 53   Crisman, T. J. *et al.* Identification of an Efficient Gene Expression Panel for
965   Glioblastoma Classification. *PLoS One* **11**, e0164649, doi:10.1371/journal.pone.0164649
966   (2016).
967 54   Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J.
968   Transposition of native chromatin for fast and sensitive epigenomic profiling of open
969   chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218,
970   doi:10.1038/nmeth.2688 (2013).
971 55   Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*
972   **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
973 56   Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137,
974   doi:10.1186/gb-2008-9-9-r137 (2008).
975 57   Rendeiro, A. F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia
976   identify subtype-specific epigenome signatures and transcription regulatory networks.
977   *Nat Commun* **7**, 11938, doi:10.1038/ncomms11938 (2016).
978 58   Palmisano, W. A. *et al.* Predicting lung cancer by detecting aberrant promoter
979   methylation in sputum. *Cancer Res* **60**, 5954-5958 (2000).
980 59   Uren, A. G. *et al.* A high-throughput splinkerette-PCR method for the isolation and
981   sequencing of retroviral insertion sites. *Nat Protoc* **4**, 789-798,
982   doi:10.1038/nprot.2009.64 (2009).
983 60   Schuffler, P. J. *et al.* TMARKER: A free software toolkit for histopathological cell
984   counting and staining estimation. *J Pathol Inform* **4**, S2, doi:10.4103/2153-3539.109804
985   (2013).
986 61   Thielecke, L. *et al.* Limitations and challenges of genetic barcode quantification. *Sci Rep*
987   **7**, 43249, doi:10.1038/srep43249 (2017).
988 62   Micallef, L. & Rodgers, P. eulerAPE: drawing area-proportional 3-Venn diagrams using
989   ellipses. *PLoS One* **9**, e101717, doi:10.1371/journal.pone.0101717 (2014).
990 63   Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of*
991   *Physical Chemistry* **81**, 2340-2361, doi:10.1021/j100540a008 (1977).
992

993

994

995

996

997

998

999

1000

1001

1002 **Extended Data Figure Legends**
1003

1004 **Extended data figure 1 | Barcode data processing.**

1005 **a**, Summary of GBM models used for barcoding experiments indicating TCGA subgroups as

1006 determined by RNA-Seq[20], self-renewing frequency as assessed by primary limiting dilution

1007 analysis (LDA), the number of primary xenografts successfully established and the cell dose

1008 used for primary xenografts (n.d: not done, n.s: no spheres). **b**, Proliferation kinetics of GSC

1009 cultures *in vitro*. Data are shown as mean ± sd of 3 technical replicates. **c**, Cell doubling times of

1010 GSCs grown in culture calculated using the data in (**b**). Data are shown as mean ± sd of 3

1011 technical replicates, horizontal line marks 24 hours. **d-f**, Relationship between fractional read

1012 value (FRV) and input cell numbers in spiked-in controls for the three sequencing runs. The

1013 highly influential data points (Cook's distance > 4/n) are grayed out and not used for regression

1014 analysis to estimate relative clone sizes. The black line is the line of best fit, and the grey box

1015 indicates sequencing noise threshold. **g**, Analysis of barcode sequence saturation across six *in*

1016 *vivo* experiments. **h**, Position weight matrices depicting the representation of variable nucleotides

1017 in the barcode library, the $(1)_{719}$ ipsilateral sample, as well as the largest and smallest 100 clones

1018 in that sample. The height of nucleotides at each position represents its relative frequency, with

1019 the most frequently occurring nucleotide shown in the top position. **i**, Summary of unique

1020 barcode integration sites identified by splinkerette PCR.

1021

1022 **Extended data figure 2 | Molecular characterization of GBMs and GBM xenografts.**

1023 **a**, Oncoprint plot of mutations identified in primary GBM tissue samples that are of the top 200

1024 recurrently mutated genes in the provisional TCGA dataset[21]. **b**, Multidimensional scaling plot

1025 for the 32-gene simple GBM classification method using RNA-Seq[53]. Shown are the TCGA

1026 samples with RNA-Seq data and 5 patient samples used in the current study. TCGA samples are

1027 labelled and coloured according to their original subgroup as determined from microarray

1028 expression analysis[20]. **c**, Methylation-specific PCR assay for the MGMT promoter in 6 primary

1029 GBMs. L: ladder, -ve: water only control, U: unmethylated PCR product, M: methylated PCR

1030 product. Specific ladder marker sizes are shown in base pairs. **d**, Pairwise correlation of ATAC-

1031 Seq peak intensities across GSC culture models and compared with a chronic lymphocytic

1032 leukaemia (CLL) control[57]. Black outline highlights correlations for GSC cultures derived from

1033 the GBMs used for the *in vivo* barcoding study (G719, G729, G754). **e**, Summary of somatic

1034 mutations identified using exome sequencing from representative GBM-719 barcoded

1035 xenografts, grouped according to type. p2 Veh: passage 2; treated with vehicle, p2 TMZ: passage

1036 2; treated with TMZ, p3 Veh Veh: passage 3; treated with vehicle at passages 2 and 3, p3 TMZ

1037 TMZ: passage 3; treated with TMZ at passages 2 and 3 and briefly expanded *in vitro* prior to

1038 sequencing. **f**, Heat map representing relative copy number profiles from whole exome

1039 sequencing of GBM-719 xenograft samples. Segments of gains (red) or deletions (blue) are

1040 colour-coded based on log2 copy number ratios. Frequent loss of chromosome 10 is a common

1041 observation in GBM. **g**, Summary of patient characteristics for all tumour samples used

1042 throughout the study, and the experiment(s) that each sample is used for.

1043

1044 **Extended data figure 3 | Functional characterization of GBMs and GBM xenografts.**

1045 **a**, H&E and human-specific nestin staining in primary glioblastoma specimens, scale bar = 100

1046 μm. **b**, H&E and human-specific nestin staining for representative GBM xenografts, scale bar =

1047 100 μm. **c**, Survival analysis of xenografts derived from the indicated GBM model and treatment

1048 conditions. All survival analyses were performed using a log-rank test (n = 4 mice per group

1049   with the exception of the GBM-754 experiment, Vehicle – Vehicle group which contains 3

1050   mice). **d**, Quantification of percentage proliferative activity in serial xenografts by Ki-67 staining

1051   and percentage apoptosis by cleaved Caspase-3 staining, mean ± sd of 6 representative sections

1052   from the same xenograft sample.

1053

1054   **Extended data figure 4 | GSCs are able to invade contralaterally and have heterogeneous**

1055   **clonal outputs.**

1056   **a**, Human-specific nestin staining in representative xenografts between ipsilateral and

1057   contralateral hemispheres (scale bar = 1mm, Ipsi: ipsilateral hemisphere, Contra: contralateral

1058   hemisphere). **b**, Comparison of cell numbers recovered from xenografts between the ipsilateral

1059   and contralateral fractions, two-sided paired t-tests. Single data points are overlaid over the box

1060   plot, the horizontal line represents the median, and the lower and upper hinges represent the 25th

1061   and 75th quartiles respectively. The lower and upper whiskers extend from the hinge to the

1062   lowest and highest values within 1.5 times the inter-quartile range (IQR). **c**, Plot of Pearson

1063   correlation coefficients comparing relative clone sizes between two hemispheres, for the

1064   indicated sample groups. The box-plots are displayed as with panel (**b**). **d**, Clonal composition

1065   of tumours generated serially from contralateral fractions, grouped according to the geographical

1066   distribution of each detected clone in the previous (primary) passage. **e**, Clone size distributions

1067   for representative xenograft samples. All data shown are from ipsilateral hemispheres, not treated

1068   with TMZ, and generated from ipsilateral-derived cells from the previous passage (in the case of

1069   secondary and tertiary xenografts). Fits to a negative binomial distribution (curve) are included

1070   for patients with rich data sets (GBM-719, GBM-742, and GBM-754), used for quantitative

1071   analyses. Plot titles identify the respective sequence of serial passages by the nomenclature

1072   introduced in the Supplementary Theory. **f**, Representative correlation of clone size between

1073   successive serial passages of GBM-719 untreated xenografts with Pearson's *r* indicated. P1:

1074   primary passage, P2: secondary passage, P3: tertiary passage. **g**, Representative correlations of

1075   clone size between different secondary passage replicate experiments derived from the same

1076   primary xenograft as panel (**f**), with Pearson's *r* indicated. The red arrowhead shows deviations

1077   from a linear correlation due to large outliers. R1: replicate 1, R2: replicate 2, R3: replicate 3.

1078

1079   **Extended data figure 5 | First incomplete moment of clone size distributions for GBM-719,**

1080   **-729, and -735 xenografts.**

1081   **a-c**, First incomplete moments of the clone size distributions for all xenograft samples derived

1082   from patient tumours GBM-719 (**a**), GBM-729 (**b**), and GBM-735 (**c**). Samples are named

1083   according to the sequence of samples injected, V: vehicle treated, T: TMZ treated, C: generated

1084   from the contralateral fraction of the previous passage. For illustrative purposes, GBM-719

1085   xenografts (**a**) that are TMZ-treated are marked with a red arrowhead where the distribution

1086   appears to deviate from the negative binomial. The indicated fit parameter n_0 describe a

1087   characteristic clone size of the population (Supplementary Theory 2-3). Where Group B clones

1088   (large outliers) were removed to generate a more accurate fit, the number of clones removed is

1089   indicated and the re-calculated first incomplete moment distributions with outliers removed are

1090   plotted in grey. **d**, Schematic describing how a sequence of treatments resulting in a particular

1091   xenograft sample is incorporated into the sample nomenclatures.

1092

1093   **Extended data figure 6 | First incomplete moment of clone size distributions for GBM-742,**

1094   **-743, and -754 xenografts and variant allele frequencies (VAFs) for GBM-719 xenografts.**

1095   **a-c**, First incomplete moments of the clone size distributions for all xenografts derived from the

1096   tumours GBM-742 (**a**), GBM-743 (**b**), and GBM-754 (**c**). Sample and plot annotations are as

1097   described for Extended data figure 5. **d**, Distribution of variant allele frequencies (VAFs) across

1098   GBM-719 xenograft samples. Mutations with a VAF of 0.5 likely corresponds to variants in the

1099   clonal population (found in all cells within the tumour), while less prevalent mutations

1100   correspond to subclonal populations defined by recent mutational events found only in a subset

1101   of cells. **e**, Comparison of VAF values for mutations in paired secondary and tertiary passages. **f**,

1102   First incomplete moments show a negative binomial distribution for VAF values below 0.5

1103   across xenograft samples. The dashed line shows a fit to the exponential and the vertical line

1104   marks a VAF of 0.5. **g**, First incomplete moments for mutations that are newly detected in the

1105   tertiary vehicle- and TMZ-treated passage. **h**, Same as panel (**f**) after filtering out mutations that

1106   do not occur in diploid regions of the genome. **i**, Same as panel (**g**) after filtering out mutations

1107   that do not occur in diploid regions of the genome.

1108

1109   **Extended data figure 7 | Barcode analysis of xenograft derived cultures.**

1110   **a**, Proportional Venn diagrams depicting the number of unique and shared barcoded clones as

1111   defined by the *in vivo* passages (primary, secondary, or tertiary), that are also detectable within

1112   the specified xenograft-derived cultures. **b**, Comparison of clone sizes between paired primary

1113   xenografts and primary xenograft-derived GSC cultures. **c**, Correlation of clone sizes between

1114   TMZ-treated GBM-719 xenografts, and cultures derived from these xenografts. A select cluster

1115   of clones that become outcompeted after secondary xenografts are outlined in blue, and

1116   Spearman's rho coefficients are as indicated. **d**, First incomplete moments of the full clone size

1117   distributions for GBM-754 primary xenograft cultures at different times throughout culture

1118   expansion. **e**, First incomplete moments of the clone size distributions used in panel (**d**), with the

1119   14 largest outlier clones removed from each sample. **f**, Pairwise clone size comparisons between

1120   replicate cultures in (**d**), with Spearman's rho indicated.

1121

1122   **Extended data figure 8 | First incomplete moment of clone size distributions for remaining**

1123   **GBM xenograft derived cultures.**

1124   **a**, Plots of first incomplete moment for cultures derived from the indicated GBM xenografts. **b**,

1125   Same as (**a**), with the indicated number of large outlier clones removed from the analysis.

1126

1127   **Extended data figure 9 | Epigenetic drug screening of GBM-754 primary xenograft culture.**

1128   **a**, Primary drug screen of GBM-754 primary xenograft-derived culture, with growth assessed as

1129   culture density relative to DMSO control. Compounds highlighted in blue were used in

1130   subsequent experiments. **b**, Strategy to identify clonal differences in drug response. Cells are

1131   treated in duplicate with each compound, and allowed to repopulate to the same density as

1132   DMSO controls prior to barcode sequencing. **c**, Summary of results from drug repopulation

1133   experiments. The top plot shows the ratio between sum relative clone sizes of Group B and

1134   Group A, technical replicates are denoted as 1, 2, or 3. The horizontal line marks the mean

1135   Group B/Group A ratio for DMSO treated cultures. The bottom plot shows the number of reads

1136   obtained from each sample after repopulation, relative to DMSO. The horizontal line marks the

1137   mean number of reads for DMSO samples. **d**, Additional technical replicate experiments related

1138   to Fig. 3g, demonstrating selectivity of UNC1999 and MI-2-2 on Group A and B clones

1139   respectively. **e**, Dose response assays for the indicated GSC culture models upon UNC1999 and

1140   MI-2-2 treatment, mean ± sd of 6 technical replicates. **f**, Two additional independent experiments

1141   related to Fig. 3h. *P* values for the left and right replicates respectively are 6.95×10^{-4}; 0.148 for

1142   DMSO vs. CI-994, 0.338; 0.55 for DMSO vs. GSK591, 3.31×10^{-3}; 0.0177 for DMSO vs.

1143   UNC1999, 2.15×10^{-11}; 1.59×10^{-7} for DMSO vs. MI-2-2, 1.49×10^{-10}; 3.7×10^{-12} for MI-nc vs.

1144   M, 0.963; 0.408 for M vs. M + C, 0.355; 0.408 for M vs. M + G, 2.68×10^{-9}; 6.06×10^{-8} for M

1145   vs. M + U. **g**, Combined effect of GSK343 and MI-2-2 on self-renewal. $P = 4.42 \times 10^{-6}$ for

1146   DMSO vs GSK343, 2.96×10^{-12} for DMSO vs MI-2-2, 3.62×10^{-6} for GSK343 vs M + G, 0.0125

1147   for MI-2-2 vs M + G. **h**, Combined effect of UNC1999 and MI-2-2 on self-renewal when used at

1148   1 μM, representative of 3 independent experiments. $P = 0.147$ for DMSO vs. UNC1999, 0.129

1149   for DMSO vs MI-2-2, 9.84×10^{-4} for DMSO vs. M + U. **i**, Two additional independent

1150   experiments related to Fig. 3i. *P* values for the left and right replicates respectively are $4.59 \times 10^{-}$

1151 ^{5}; 4.81×10^{-15} for DMSO vs. UNC1999, 3.28×10^{-25}; 1.13×10^{-31} for DMSO vs MI-2-2, $1.86 \times$

1152 10^{-11}; 3.61×10^{-6} for UNC1999 vs MI-2-2. **j-m**, Combined effect of UNC1999 and MI-2-2 on

1153   self-renewal in the indicated GSC culture models. *P* values for the G523, G549, G564, G566

1154   experiments respectively are 1.9×10^{-5}; 1; 0.758; 0.799 for DMSO vs UNC1999, 8.14×10^{-18};

1155 2.14×10^{-4}; 0.503; 6.12×10^{-4} for DMSO vs MI-2-2, 2.72×10^{-12}; 3.28×10^{-30}; 1.15×10^{-21}; 2.54

1156 $\times 10^{-8}$ for UNC1999 vs M + U, 7.69×10^{-3}; 1.26×10^{-15}; 2.61×10^{-18}; 8.82×10^{-3} for MI-2-2 vs

1157   M + U. **n**, Combined effect of UNC1999 and MI-2-2 on self-renewal of uncultured GBM-851

1158   cells. $P = 3.01 \times 10^{-3}$ for DMSO vs UNC1999, 1.36×10^{-4} for DMSO vs MI-2-2, 3.11×10^{-3} for

1159   UNC1999 vs M + U, 0.0276 for MI-2-2 vs M + U. Analysis of LDA results was performed using

1160   ELDA software[34], error bars represent 95% confidence interval (ns $P > 0.05$, * $P \leq 0.05$, ** $P \leq$

1161   0.01, *** $P \leq 0.001$).

1162

1163   **Extended data figure 10 | First incomplete moment of the clone size distributions for drug-**

1164 **treated GBM-754 primary xenograft cultures.**

1165 **a**, First incomplete moments of the full clone size distributions of GBM-754 primary xenograft

1166 cultures treated with different drugs. **b**, First incomplete moments of the clone size distributions

1167 used in panel (**a**), with 5 group B clones removed.

1168

1169

a Primary passage / Secondary passage / Tertiary passage

Primary injection → Tumour growth → Ipsilateral / Contralateral → Harvesting → ...TCA ...TCG ...CTG ...AAA... Sequencing

Re-injection: +TMZ +Veh +Veh → Harvesting → ...TCA ...TCG ...CTG ...AAA... Sequencing

Re-injection: +TMZ +Veh +Veh → Harvesting → ...TCA ...TCG ...CTG ...AAA... Sequencing

b H&E · Nestin · Ki-67 · Cleaved Caspase-3

c Fold change in cell number vs Passage (1, 2, 3)

d GBM-719 · GBM-729 · GBM-735 · GBM-742 · GBM-743 · GBM-754

Primary (blue) · Secondary (red) · Tertiary (green)

GBM-719: 994, 18, 154, 98, 355, 286, 165
GBM-729: 89, 8, 6, 8, 4, 9
GBM-735: 85, 35
GBM-742: 266, 116, 74
GBM-743: 27, 10, 1
GBM-754: 694, 15, 53, 27, 153, 85

a

GBM-719 primary passage

Ipsilateral (1227 clones)
Contralateral (780 clones)

Bin probability / Relative clone size

GBM-719 secondary passage

Replicate 3 (243 clones)
Replicate 2 (467 clones)
Replicate 1 (402 clones)

Relative clone size

GBM-719 tertiary passage

Replicate 3 (200 clones)
Replicate 2 (347 clones)
Replicate 1 (353 clones)

Relative clone size

b

First incomplete moment

Contralateral
Ipsilateral

Relative clone size

Replicate 3
Replicate 1
Replicate 2

Relative clone size

Replicate 3
Replicate 1
Replicate 2

Relative clone size

c

Stem-like cells

S → S S ε
S → S P $1 - \varepsilon$

Progenitors

P → P P   50%
P → D   50%

Differentiating progeny

D → ∅

d

Clone size (cell number)

Primary   Secondary   Tertiary

Mean clone size

Time (days)

40   80   120   160   200

e

First incomplete moment

Primary
Secondary
Tertiary

Clone size (cell number)

0   1000   2000

f

Clone size cross correlation

Pairs of passages: 1,2   1,3   2,3

g

Clone survival probability $\beta_{1/2}$ (%)

Passage: 1   2   3

h

P2 clone size (×10^3)

P1 clone size (×10^3)

P3 clone size (×10^3)

P2 clone size (×10^3)

a

Relative clone size (P2 + TMZ) vs Relative clone size (P1). Group B, Group A.

b

Relative clone size (P3 + TMZ) vs Relative clone size (P2 + TMZ). Group B, Group A.

c

$\rho = 0.257$. Relative clone size (P2 + TMZ, R2) vs Relative clone size (P2 + TMZ, R1).

d

$\rho = 0.361$. Relative clone size (P3 + TMZ, R2) vs Relative clone size (P3 + TMZ, R1).

e

Simulation. Cell number in P2 ($\times 10^4$) vs Cell number in P1 ($\times 10^4$).

f

Simulation. Cell number in P3 ($\times 10^4$) vs Cell number in P2 ($\times 10^4$).

g

$(1)_{754}$ culture 5 µM

Group A / Group B

29.2	13.3	
70.8	86.7	

DMSO — UNC1999

| 29.2 | 82.5 |
| 70.8 | 17.5 |

DMSO — MI-2-2

Percentage clone size

h

$(1)_{754}$ culture 5 µM

Sphere forming frequency. DMSO, CI-994, GSK591, UNC1999, MI-2-2. ns, ns, ***, ***

MI-nc, M, M + C, M + G, M + U. ***, ns, ns, ***

i

$(1, 1^T, 1^T)_{719}$ culture 5 µM

Sphere forming frequency. DMSO, UNC1999, MI-2-2. **, ***

j

$(1, 1^T, 1^T)_{719}$ derived xenografts

$P = 0.0305$. Tumour weight (mg). Vehicle, MI-2-2.

a

Sample	Sample type	Subgroup	Primary LDA	Primary xenografts	Cell dose (×10^4)
GBM-719	Primary GBM	Classical	1 in 202	4 of 4	12.5
GBM-729	Primary GBM	Proneural	1 in 50	4 of 4	12.5
GBM-732	Primary GBM	n.d.	n.s.	0 of 4	10.0
GBM-735	Primary GBM	n.d.	1 in 92	2 of 4	3.0
GBM-738	Primary GBM	n.d.	n.s.	0 of 4	2.0
GBM-742	Secondary GBM	Proneural	n.s.	3 of 3	2.4
GBM-743	Primary GBM	Proneural	1 in 1600	4 of 4	8.0
GBM-753	Primary GBM	n.d.	1 in 12	0 of 4	1.5
GBM-754	Primary GBM	Classical	1 in 15	4 of 4	12.5

b

c

d

y = 2.1966x - 8.1237
R^2 = 0.952
FRV threshold: 0.000625 (173.6 cells)
87.9% sensitivity
100.0% specificity

e

y = 1.0266x - 3.802
R^2 = 0.915
FRV threshold: 0.00300 (17.6 cells)
91.4% sensitivity
100.0% specificity

f

y = 0.9499x - 3.5995
R^2 = 0.805
FRV threshold: 0.0325 (166.9 cells)
84.9% sensitivity
98.1% specificity

g

h

Barcode library (167,746 sequences)

All clones detected in (1)$_{719}$ ipsilateral (1,227 sequences)

100 largest clones in (1)$_{719}$ ipsilateral

100 smallest clones in (1)$_{719}$ ipsilateral

i

Experiment	Chromosome	Locus	Genomic Features
GBM-719	Chr3+	17,402,188	TBC1D5 (intron)
	Chr12+	120,894,622	SPPL3 (intron)
	Chr17-	75,353,667	GRB2 (intron)
	Chr7+	157,240,766	UBE3C (intron)
	Chr16+	57,606,779	GPR114 (5'); GPR56 (3')
	Chr2+	206,017,591	INO80D (intron)
GBM-729	Chr3-	68,192,047	FAM19A1 (intron)
	Chr10+	94,551,172	HELLS (intron)
GBM-735	ChrX-	155,521,554	ENOX2 (5'); SLC25A14 (3')
	Chr16+	3,370,841	OR2C1 (5'); MTRN2L4 (3')
	Chr12-	57,600,347	PIP3K2C (exon)
GBM-743	ChrX-	149,328,429	IDS (5'); AFF2 (3')
	Chr19-	58,410,082	ZNF584 (exon)
	ChrX+	99,394,659	XRCC6P5 (5'); DIAPH2 (3')
GBM-754	Chr19-	11,712,987	ZNF823 (5'); ZNF833P (3')
	Chr6+	101,906,549	GRIK2 (intron)
	Chr1+	11,863,070	KIAA2013 (5'); NPPB (3')
	Chr7-	69,341	FAM20C (5')
	Chr6-	131,054,618	EPB41L2 (intron)
	Chr6-	159,780,213	TCP1 (intron)

Sample	Ki-67 $\pm$ SD	Cleaved Caspase-3 $\pm$ SD
$(2)_{719}$	53.1 ± 7.2	36.5 ± 10.6
$(1,4^V)_{719}$	31.5 ± 3.7	0.6 ± 0.3
$(1,3^V,3^V)_{719}$	53.7 ± 3.8	13.9 ± 7.9
$(2)_{729}$	40.8 ± 2.6	0.9 ± 0.3
$(1,3^V)_{729}$	40.1 ± 3.5	0.9 ± 0.6
$(1,1^V,3^V)_{729}$	37.0 ± 5.0	0.9 ± 0.3
$(2)_{754}$	41.1 ± 4.1	0.7 ± 0.3
$(1,3^V)_{754}$	36.4 ± 6.6	0.3 ± 0.2
$(1,2^T)_{754}$	36.1 ± 4.0	0.1 ± 0.1
$(1,2^V,3^V)_{754}$	36.9 ± 6.9	0.4 ± 0.2
$(1,2^T,3^T)_{754}$	29.9 ± 5.8	0.5 ± 0.4

a

$(1)_{719}$ ipsilateral	$(1)_{719}$ contralateral	$(1,1^V)_{719}$ ipsilateral	$(1,1^V)_{719}$ contralateral	$(1,2^V)_{719}$ ipsilateral	$(1,2^V)_{719}$ contralateral

1227 clones, $n_0 = 0.00133$

780 clones, $n_0 = 0.00209$

402 clones, $n_0 = 0.00363$

327 clones, $n_0 = 0.00409$

467 clones, $n_0 = 0.00378$

339 clones, $n_0 = 0.00557$

$(1,3^V)_{719}$ ipsilateral — 243 clones, $n_0 = 0.00999$, Outliers removed: 9

$(1,3^V)_{719}$ contralateral — 215 clones, $n_0 = 0.0098$

$(1,1^T)_{719}$ ipsilateral — 105 clones, $n_0 = 0.0784$

$(1,1^T)_{719}$ contralateral — 63 clones, $n_0 = 0.0665$

$(1,2^T)_{719}$ ipsilateral — 40 clones, $n_0 = 0.122$, Outliers removed: 9

$(1,2^T)_{719}$ contralateral — 49 clones, $n_0 = 0.0468$, Outliers removed: 12

$(1^C,1)_{719}$ ipsilateral — 127 clones, $n_0 = 0.0129$, Outliers removed: 4

$(1^C,1)_{719}$ contralateral — 87 clones, $n_0 = 0.0169$

$(1^C,2)_{719}$ ipsilateral — 183 clones, $n_0 = 0.0095$

$(1^C,2)_{719}$ contralateral — 126 clones, $n_0 = 0.0106$

$(1,2^V,1)_{719}$ ipsilateral — 353 clones, $n_0 = 0.00497$

$(1,2^V,1)_{719}$ contralateral — 316 clones, $n_0 = 0.00484$

$(1,2^V,2)_{719}$ ipsilateral — 347 clones, $n_0 = 0.00584$

$(1,2^V,2)_{719}$ contralateral — 223 clones, $n_0 = 0.00701$

$(1,2^V,3)_{719}$ ipsilateral — 200 clones, $n_0 = 0.00994$

$(1,2^V,3)_{719}$ contralateral — 146 clones, $n_0 = 0.00982$

$(1,3^V,1^V)_{719}$ ipsilateral — 170 clones, $n_0 = 0.0205$, Outliers removed: 9

$(1,3^V,1^V)_{719}$ contralateral — 149 clones, $n_0 = 0.0194$, Outliers removed: 9

$(1,3^V,2^V)_{719}$ ipsilateral — 183 clones, $n_0 = 0.0143$, Outliers removed: 9

$(1,3^V,2^V)_{719}$ contralateral — 145 clones, $n_0 = 0.0132$, Outliers removed: 9

$(1,3^V,1^T)_{719}$ ipsilateral — 48 clones, $n_0 = 0.0719$

$(1,3^V,1^T)_{719}$ contralateral — 41 clones, $n_0 = 0.0323$

$(1,1^T,1^V)_{719}$ ipsilateral — 80 clones, $n_0 = 0.0733$

$(1,1^T,1^V)_{719}$ contralateral — 57 clones, $n_0 = 0.0737$

$(1,1^T,2^V)_{719}$ ipsilateral — 64 clones, $n_0 = 0.0552$

$(1,1^T,2^V)_{719}$ contralateral — 45 clones, $n_0 = 0.0413$

$(1,1^T,1^T)_{719}$ ipsilateral — 25 clones, $n_0 = 0.141$

$(1,1^T,1^T)_{719}$ contralateral — 18 clones, $n_0 = 0.134$

$(1,1^T,2^T)_{719}$ ipsilateral — 59 clones, $n_0 = 0.0241$, Outliers removed: 3

$(1,1^T,2^T)_{719}$ contralateral — 23 clones, $n_0 = 0.14$

$(1,1^T,3^T)_{719}$ ipsilateral — 25 clones, $n_0 = 0.0609$, Outliers removed: 3

$(1,1^T,3^T)_{719}$ contralateral — 98 clones, $n_0 = 0.0201$, Outliers removed: 3

b

$(1)_{729}$ ipsilateral — 89 clones, $n_0 = 0.616$

$(1)_{729}$ contralateral — 50 clones, $n_0 = 0.0668$

$(1,1^V)_{729}$ ipsilateral — 8 clones, $n_0 = 0.29$

$(1,1^V)_{729}$ contralateral — 13 clones, $n_0 = 0.29$

$(1,2^V)_{729}$ ipsilateral — 16 clones, $n_0 = 0.547$

$(1,2^V)_{729}$ contralateral — 11 clones, $n_0 = 0.517$

$(1^C,1)_{729}$ ipsilateral — 14 clones, $n_0 = 1.82$

$(1^C,1)_{729}$ contralateral — 7 clones, $n_0 = 1.81$

$(1^C,2)_{729}$ ipsilateral — 7 clones, $n_0 = 0.875$

$(1^C,2)_{729}$ contralateral — 5 clones, $n_0 = 0.528$

$(1,1^V,1^V)_{729}$ ipsilateral — 16 clones, $n_0 = 0.472$

$(1,1^V,1^V)_{729}$ contralateral — 6 clones, $n_0 = 0.713$

$(1,1^V,2^V)_{729}$ ipsilateral — 15 clones, $n_0 = 0.977$

$(1,1^V,2^V)_{729}$ contralateral — 5 clones, $n_0 = 0.543$

c

$(1)_{735}$ ipsilateral — 108 clones, $n_0 = 0.0349$

$(1)_{735}$ contralateral — 53 clones, $n_0 = 0.301$

$(1,1^V)_{735}$ ipsilateral — 34 clones, $n_0 = 0.0668$

$(1,1^V)_{735}$ contralateral — 32 clones, $n_0 = 0.0505$

$(1,2^V)_{735}$ ipsilateral — 9 clones, $n_0 = 0.245$

$(1,2^V)_{735}$ contralateral — 8 clones, $n_0 = 0.199$

d

GBM-**719**

Ipsilateral / Contralateral

Mouse 1 → Mouse **3** Vehicle treated → Mouse 1 **T**MZ treated → $(1,3^V,1^T)_{719}$ Contralateral — Sequencing

Mouse 1 → $(1^C,1)_{719}$ Ipsilateral — Sequencing

First incomplete moment (y-axis)

Relative clone size (x-axis)

■ Primary ipsilateral ■ Primary contralateral ■ Secondary ipsilateral ■ Secondary contralateral ■ Tertiary ipsilateral ■ Tertiary contralateral ■ Outliers removed

a

(1)$_{742}$ ipsilateral — 318 clones, $n_0 = 0.0623$
(1)$_{742}$ contralateral — 176 clones, $n_0 = 0.0236$
(1,1^V)$_{742}$ ipsilateral — 46 clones, $n_0 = 0.148$
(1,1^V)$_{742}$ contralateral — 29 clones, $n_0 = 0.19$
(1,2^V)$_{742}$ ipsilateral — 55 clones, $n_0 = 0.137$
(1,2^V)$_{742}$ contralateral — 44 clones, $n_0 = 0.168$

(1,3^V)$_{742}$ ipsilateral — 66 clones, $n_0 = 0.0819$
(1,3^V)$_{742}$ contralateral — 36 clones, $n_0 = 0.107$
(1,1^T)$_{742}$ ipsilateral — 108 clones, $n_0 = 0.0902$
(1,1^T)$_{742}$ contralateral — 33 clones, $n_0 = 0.106$

b

(1)$_{743}$ ipsilateral — 36 clones, $n_0 = 0.11$
(1)$_{743}$ contralateral — 10 clones, $n_0 = 0.362$

(1,1^V)$_{743}$ ipsilateral — 9 clones, $n_0 = 0.327$
(1,1^V)$_{743}$ contralateral — 6 clones, $n_0 = 0.549$
(1,2^V)$_{743}$ ipsilateral — 3 clones, $n_0 = 0.892$
(1,2^V)$_{743}$ contralateral — 7 clones, $n_0 = 0.861$
(1^C,1)$_{743}$ ipsilateral — 7 clones, $n_0 = 0.655$
(1^C,1)$_{743}$ contralateral — 6 clones, $n_0 = 0.473$

c

(1)$_{754}$ ipsilateral — 789 clones, $n_0 = 0.0171$, Outliers removed: 3
(1)$_{754}$ contralateral — 457 clones, $n_0 = 0.0209$, Outliers removed: 1
(1,1^V)$_{754}$ ipsilateral — 136 clones, $n_0 = 0.087$, Outliers removed: 1
(1,1^V)$_{754}$ contralateral — 110 clones, $n_0 = 0.0877$, Outliers removed: 1
(1,2^V)$_{754}$ ipsilateral — 92 clones, $n_0 = 0.326$, Outliers removed: 1
(1,2^V)$_{754}$ contralateral — 80 clones, $n_0 = 0.0731$, Outliers removed: 1

(1,1^T)$_{754}$ ipsilateral — 172 clones, $n_0 = 0.0409$, Outliers removed: 1
(1,1^T)$_{754}$ contralateral — 182 clones, $n_0 = 0.0813$
(1^C,1)$_{754}$ ipsilateral — 147 clones, $n_0 = 0.187$
(1^C,1)$_{754}$ contralateral — 59 clones, $n_0 = 0.246$, Outliers removed: 1
(1,2^V,1^V)$_{754}$ ipsilateral — 100 clones, $n_0 = 0.504$, Outliers removed: 1
(1,2^V,1^V)$_{754}$ contralateral — 49 clones, $n_0 = 1.12$

(1,2^V,2^V)$_{754}$ ipsilateral — 47 clones, $n_0 = 0.396$, Outliers removed: 1
(1,2^V,2^V)$_{754}$ contralateral — 20 clones, $n_0 = 0.543$
(1,2^V,1^T)$_{754}$ ipsilateral — 66 clones, $n_0 = 0.595$
(1,2^V,1^T)$_{754}$ contralateral — 40 clones, $n_0 = 0.693$
(1,2^V,2^T)$_{754}$ ipsilateral — 80 clones, $n_0 = 0.53$
(1,2^V,2^T)$_{754}$ contralateral — 46 clones, $n_0 = 0.266$, Outliers removed: 1

(1,1^T,1^V)$_{754}$ ipsilateral — 78 clones, $n_0 = 0.695$
(1,1^T,1^V)$_{754}$ contralateral — 25 clones, $n_0 = 0.921$
(1,1^T,2^V)$_{754}$ ipsilateral — 27 clones, $n_0 = 0.245$
(1,1^T,2^V)$_{754}$ contralateral — 26 clones, $n_0 = 0.305$, Outliers removed: 1
(1,1^T,3^V)$_{754}$ ipsilateral — 122 clones, $n_0 = 0.585$
(1,1^T,3^V)$_{754}$ contralateral — 87 clones, $n_0 = 0.371$

(1,1^T,1^T)$_{754}$ ipsilateral — 115 clones, $n_0 = 0.213$
(1,1^T,1^T)$_{754}$ contralateral — 12 clones, $n_0 = 0.229$
(1,1^T,2^T)$_{754}$ ipsilateral — 99 clones, $n_0 = 0.295$
(1,1^T,2^T)$_{754}$ contralateral — 52 clones, $n_0 = 0.142$
(1,1^T,3^T)$_{754}$ ipsilateral — 99 clones, $n_0 = 0.725$
(1,1^T,3^T)$_{754}$ contralateral — 26 clones, $n_0 = 0.11$, Outliers removed: 1

Y-axis (c): First incomplete moment
X-axis: Relative clone size

Legend: ■ Primary ipsilateral ■ Primary contralateral ■ Secondary ipsilateral ■ Secondary contralateral ■ Tertiary ipsilateral ■ Tertiary contralateral ■ Outliers removed

d

p2 Veh — 546 mutations
p3 Veh Veh — 112 mutations
p2 TMZ — 122 mutations
p3 TMZ TMZ — 995 mutations

Y-axis: Number of mutations
X-axis: Variant allele frequency (VAF)

e

p3 Veh Veh VAF vs p2 Veh VAF
p3 TMZ TMZ VAF vs p2 TMZ VAF

f

p2 Veh
p3 Veh Veh
p2 TMZ
p3 TMZ TMZ

Y-axis: First incomplete moment
X-axis: Variant allele frequency (VAF)

g

Veh - new mutations in p3 — 44 mutations
TMZ - new mutations in p3 — 939 mutations

h

p2 Veh — 407 mutations
p3 Veh Veh — 43 mutations
p2 TMZ — 96 mutations
p3 TMZ TMZ — 271 mutations

Y-axis: First incomplete moment
X-axis: Variant allele frequency (VAF)

i

Veh - new mutations in p3 — 13 mutations
TMZ - new mutations in p3 — 260 mutations

X-axis: Variant allele frequency (VAF)

a

GBM-719 primary derived cultures

GBM-719 secondary derived cultures

GBM-719 tertiary derived cultures

GBM-729 primary derived culture

GBM-729 secondary derived culture

GBM-729 tertiary derived culture

GBM-754 primary derived culture

GBM-754 secondary derived cultures

GBM-754 tertiary derived cultures

Primary Secondary Tertiary

b

$(1)_{719}$ ipsilateral culture vs $(1)_{719}$ ipsilateral xenograft

$(1)_{729}$ ipsilateral culture vs $(1)_{729}$ ipsilateral xenograft

$(1)_{754}$ ipsilateral culture vs $(1)_{754}$ ipsilateral xenograft

c

$(1, 1^T)_{719}$ ipsilateral culture vs $(1, 1^T)_{719}$ ipsilateral xenograft, $\rho = -0.089$

$(1, 1^T)_{719}$ ipsilateral culture vs $(1, 1^T, 1^T)_{719}$ ipsilateral xenograft, $\rho = 0.371$

$(1, 1^T, 1^T)_{719}$ ipsilateral culture vs $(1, 1^T, 1^T)_{719}$ ipsilateral xenograft, $\rho = 0.325$

d

First incomplete moment vs Relative clone size

$(1)_{754}$ ipsilateral culture, day 1 A — 489 clones
$(1)_{754}$ ipsilateral culture, day 1 B — 548 clones
$(1)_{754}$ ipsilateral culture, day 3 A — 510 clones
$(1)_{754}$ ipsilateral culture, day 3 B — 492 clones
$(1)_{754}$ ipsilateral culture, day 5 A — 463 clones
$(1)_{754}$ ipsilateral culture, day 5 B — 462 clones
$(1)_{754}$ ipsilateral culture, day 7 A — 427 clones
$(1)_{754}$ ipsilateral culture, day 7 B — 446 clones

e

$(1)_{754}$ ipsilateral culture, day 1 A — 475 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 1 B — 534 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 3 A — 496 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 3 B — 478 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 5 A — 449 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 5 B — 448 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 7 A — 413 clones (14 Outliers removed)
$(1)_{754}$ ipsilateral culture, day 7 B — 432 clones (14 Outliers removed)

f

Replicate B vs Replicate A

Day 1, $\rho = 0.556$
Day 3, $\rho = 0.549$
Day 5, $\rho = 0.499$
Day 7, $\rho = 0.510$

a

(1)₇₁₉ ipsilateral culture — 137 clones

(1)₇₁₉ contralateral culture — 329 clones

(1,2ᵛ)₇₁₉ ipsilateral culture — 381 clones

(1,1ᵀ)₇₁₉ ipsilateral culture — 76 clones

(1,3ᵛ,1ᵛ)₇₁₉ ipsilateral culture — 159 clones

(1,1ᵀ,1ᵀ)₇₁₉ ipsilateral culture — 48 clones

(1)₇₂₉ ipsilateral culture — 105 clones

(1,1ᵛ)₇₂₉ ipsilateral culture — 63 clones

(1,1ᵛ,2ᵛ)₇₂₉ ipsilateral culture — 48 clones

(1,1ᵛ)₇₅₄ ipsilateral culture — 139 clones

(1,1ᵀ)₇₅₄ ipsilateral culture — 183 clones

(1,2ᵛ,1ᵛ)₇₅₄ ipsilateral culture — 88 clones

(1,1ᵀ,1ᵀ)₇₅₄ ipsilateral culture — 128 clones

b

(1)₇₁₉ ipsilateral culture — 127 clones, Outliers removed: 10

(1)₇₁₉ contralateral culture — 325 clones, Outliers removed: 4

(1,2ᵛ)₇₁₉ ipsilateral culture — 379 clones, Outliers removed: 2

(1,1ᵀ)₇₁₉ ipsilateral culture — 67 clones, Outliers removed: 9

(1,3ᵛ,1ᵛ)₇₁₉ ipsilateral culture — 152 clones, Outliers removed: 7

(1,1ᵀ,1ᵀ)₇₁₉ ipsilateral culture — 40 clones, Outliers removed: 8

(1)₇₂₉ ipsilateral culture — 103 clones, Outliers removed: 2

(1,1ᵛ)₇₂₉ ipsilateral culture — 59 clones, Outliers removed: 4

(1,1ᵛ,2ᵛ)₇₂₉ ipsilateral culture — 43 clones, Outliers removed: 5

(1,1ᵛ)₇₅₄ ipsilateral culture — 134 clones, Outliers removed: 5

(1,1ᵀ)₇₅₄ ipsilateral culture — 178 clones, Outliers removed: 5

(1,2ᵛ,1ᵛ)₇₅₄ ipsilateral culture — 83 clones, Outliers removed: 5

(1,1ᵀ,1ᵀ)₇₅₄ ipsilateral culture — 121 clones, Outliers removed: 7

First incomplete moment (y-axis) — Relative clone size (x-axis)

Legend:
■ Primary ipsilateral culture   ■ Primary contralateral culture   ■ Secondary ipsilateral culture   ■ Tertiary ipsilateral culture

a   **b**

First incomplete moment

Relative clone size

■ Control   ■ Drug treated

Supplemental Theory for

Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy

Xiaoyang Lan[1,2,3], David J. Jörg[4,5], Florence M. G. Cavalli[1,2], Laura M. Richards[12,13], Long V. Nguyen[7], Robert J. Vanner[1,2,3], Paul Guilhamon[12,13,14], Lilian Lee[1,2], Michelle M. Kushida[1,2], Davide Pellacani[7,8], Nicole I. Park[1,2,3], Fiona J. Coutinho[1,2,3], Heather Whetstone[1,2], Hayden J. Selvadurai[1,2], Clare Che[1,2], Betty Luu[1,2], Annaick Carles[9], Michelle Moksa[9], Naghmeh Rastegar[1,2], Renee Head[1,2], Sonam Dolma[1,2,11], Panagiotis Prinos[13,20], Michael D. Cusimano[17,18], Sunit Das[17,18], Mark Bernstein[16,18], Cheryl H. Arrowsmith[13,20], Andrew J. Mungall[10], Richard A. Moore[10], Yussane Ma[10], Marco Gallo[19], Mathieu Lupien[12,13,14], Trevor J. Pugh[12,13], Michael D. Taylor[1,2,11,15,18], Martin Hirst[9,10], Connie J. Eaves[7,8], Benjamin D. Simons[4,5,6*], and Peter B. Dirks[1,2,3,15,18*]

[1]Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [2]The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [3]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [4]Cavendish Laboratory, Department of Physics, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom. [5]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, CB2 1QN, United Kingdom. [6]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge, United Kingdom. [7]Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [8]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [9]Centre for High-Throughput Biology, Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [10]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [11]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario M5G 0A4, Canada. [12]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2M9, Canada. [13]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada. [14]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [15]Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario M5S 3E1, Canada. [16]Division of Neurosurgery, Toronto Western Hospital, Toronto, ON M5T 2S8, Canada. [17]Division of Neurosurgery, St. Michael's Hospital, Toronto, ON M5B 1W8, Canada. [18]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, ON M5S 1A8, Canada. [19]Departments of Physiology and Pharmacology, Biochemistry and Molecular Biology, Alberta Children's Hospital Research Institute, Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, AB T2N 4N1, Canada. [20]Structural Genomics Consortium, University of Toronto, Toronto, ON M5G 1L7, Canada.
*Correspondence: bds10@cam.ac.uk (B.D.S.), peter.dirks@sickkids.ca (P.B.D.).

Contents

In this Supplementary Text, we describe the quantitative analysis of clonal data obtained from serial transplantation experiments of human glioblastoma (GBM) xenografts involving lentiviral barcoding. Our strategy is to analyse the features of barcode frequency distributions to infer the underlying cell fate dynamics giving rise to the heterogeneity of clonal behaviour observed in experiments. This heterogeneity could either be (i) a consequence of differential engrained or evolving fitness advantages of cells or (ii) reflecting stochastic fate choices of equipotent progenitor pools (Clayton et al., 2007; Blanpain and Simons, 2013). Here we show that the experimental data is consistent with the latter scenario and that the key features of barcode frequency distributions and correlations can be explained by a simple proliferative hierarchy with glioblastoma stem-like cells at the apex.

In Section 1, we address statistical properties of lentiviral barcoding and give estimates of the amount of uniquely labelled cells. In Section 2, we show that the experimentally obtained barcode frequencies follow a negative binomial distribution. This behaviour is characteristic of a specific class of proliferative hierarchies—in Section 3, we show how such a distribution generically arises. Based on these observations, we develop a minimal model of tumour growth in Section 4 and study its predictions on tumour expansion and composition. In Section 5, we use our model to develop a simulation of the serial transplantation experiments which permits a direct comparison of our model with experiments. In Section 6, we infer plausible parameter ranges for our model on biological grounds and compare the model results of our theory with experiments. In Section 7, we use the experimentally obtained data from exome deep sequencing to probe the mutational heterogeneity of the parent tumour and as an independent window on the clonal dynamics of GBM cells.

1   Statistical properties of lentiviral barcoding

Lentiviral barcoding relies on the random infection of cells. While it entails the possibility to uniquely identify clone lineages, the randomness of the barcoding procedure may lead to the same cell acquiring multiple barcodes or to the same barcode being present in more than one cell. Since this can affect the statistical properties of the derived barcode frequency distributions, we here give an estimate for the relative amount of multiply labelled cells and barcodes present in multiple cells.

For a library consisting of N_B unique barcodes with a barcoding event occurring with probability p_B, the number n_B of barcodes acquired by a single cell follows the binomial distribution $Q(n_B) = P_{\text{Binomial}}(n_B | N_B, p_B)$, where $P_{\text{Binomial}}(n | N, p) = \binom{N}{n} p^n (1 - p)^{N-n}$. For large N_B, the distribution Q can be approximated by a

Patient	N_C (10^4)	η	Q_0	Q_1	$Q_{>1}$	R_0	R_1	$R_{>1}$	$\mathcal{Q}$
GBM719	12.5	37.8%	62.2%	29.5%	8.3%	74.3%	22.1%	3.6%	69.6%
GBM729	12.5	21.6%	78.4%	19.1%	2.5%	85.9%	13.1%	1.0%	84.4%
GBM735	3	37.6%	62.4%	29.4%	8.2%	93.2%	6.6%	0.2%	91.6%
GBM742	2.4	28.7%	71.3%	24.1%	4.6%	96.0%	3.9%	0.1%	95.3%
GBM743	8	17.3%	82.7%	15.7%	1.6%	92.7%	7.0%	0.3%	92.0%
GBM754	12.5	33.2%	66.8%	27.0%	6.2%	77.7%	19.6%	2.7%	74.1%

Table S1 Probabilities characterising the statistical properties of lentiviral barcoding with a library of $N_B = 2 \times 10^5$ barcodes (L. V. Nguyen, M. Makarem, *et al.*, 2014).

Poisson distribution,

$$Q(n_B) \simeq \frac{\nu^{n_B}}{n_B!} e^{-\nu} , \tag{1}$$

where $\nu = p_B N_B$. Using Eq. (1), the relative amount of unlabelled cells, $Q_0 = Q(0)$, the relative amount of cells labelled with one barcode, $Q_1 = Q(1)$, and the relative amount of cells carrying more than one barcode, $Q_{>1} = \sum_{n_B>1} Q(n_B)$, are obtained as

$$Q_0 = e^{-\nu} , \qquad Q_1 = \nu e^{-\nu} , \qquad Q_{>1} = 1 - (1 + \nu)e^{-\nu} . \tag{2}$$

The parameter ν characterizing the distribution of barcodes can be obtained from the labelling efficiency η, which denotes the relative amount of cells that bear at least one barcode, by requiring $1 - Q_0 = \eta$. This yields

$$\nu = -\ln(1 - \eta) . \tag{3}$$

Conversely, we can ask how likely it is that the same barcode appears in multiple cells. Out of a total of N_C cells prepared for barcoding, the number n_C of cells acquiring the same barcode is distributed according to $R(n_C) = P_{\text{Binomial}}(n_C | N_C, p_B)$. Again, for a large number of cells N_C, this can be approximated by a Poisson distribution,

$$R(n_C) \simeq \frac{\kappa^{n_C}}{n_C!} e^{-\kappa} , \tag{4}$$

where $\kappa = p_B N_C = \nu N_C / N_B$. Analogous to Eqs. (2), we obtain the relative amount of barcodes that are present in no cell, $R_0 = R(0) = e^{-\kappa}$, the relative amount of barcodes present in exactly one cell, $R_1 = R(1) = \kappa e^{-\kappa}$ and the relative amount of barcodes that have been acquired by more than one cell, $R_{>1} = \sum_{n_C>1} R(n_C) = 1 - (1 + \kappa)e^{-\kappa}$.

Multiple barcoding of the same cell is unproblematic for the quantitative analysis of barcode frequency distributions—it generates copies of clones which are however subject to the same distribution of barcode frequencies. On the other hand, barcodes distributed to multiple cells lead to an effective merging of the sizes of derived clones and thus may alter the statistical properties of the barcode frequency distribution. Among the labelled cells, the relative amount of uniquely labelled cells, i.e., cells with a unique combination of one or more barcodes, is given by

$$Q = \frac{1}{1 - Q_0} \sum_{n_{\rm B}=1}^{\infty} Q(n_{\rm B}) \left[(1 - p_{\rm B})^{N_{\rm C}-1} \right]^{n_{\rm B}} , \tag{5}$$

which, for $N_{\rm C} \gg 1$ and $p_{\rm B} \ll 1$, can be approximated in terms of the probabilities Q_0 and R_0 as

$$Q \simeq \frac{1 - Q_0^{-R_0}}{1 - Q_0^{-1}} . \tag{6}$$

Table S1 summarizes the respective probabilities for all xenografts used in this study; a large majority of labelled cells carries a unique combination of barcodes in all xenografts.

2 Barcode frequencies follow a negative binomial distribution

To obtain a quantitative understanding of tumour growth, we analyse the distribution of barcode frequencies obtained from serial transplantation experiments. Here, we show that the distributions $p(n)$ of barcode frequencies above the detection threshold for all passages and replicate experiments follow a negative binomial distribution,

$$p(n) = \frac{1}{\mathcal{N}_0} \frac{e^{-n/n_0}}{n} , \tag{7}$$

where n_0 is a characteristic barcode frequency of the respective population and $\mathcal{N}_0$ is a normalisation constant. A robust method to detect negative binomial distributions is to obtain the first incomplete moment of the distribution p, defined by

$$\mu(n) = \frac{1}{\langle n \rangle} \sum_{n'=n}^{\infty} n' p(n') , \tag{8}$$

where $\langle n \rangle = \sum_n n \, p(n)$ is the average barcode frequency. By definition, $\mu(n)$ is the relative average barcode frequency of all barcode frequencies larger than n. If the

barcode frequency distribution $p(n)$ has the negative binomial form Eq. (7), the first incomplete moment acquires an exponential dependence on the barcode frequency,

$$\mu(n) = \frac{1}{\mathcal{N}_1} e^{-n/n_0} , \tag{9}$$

where $\mathcal{N}_1$ is another normalization constant. Since the first incomplete moment, together with the average barcode frequency $\langle n \rangle$, carries the same information as the original barcode frequency distribution[1], an exponential dependence of μ is completely equivalent to a negative binomial barcode frequency distribution.

Fig. 2b in the main text and Extended Data Figs. 5-6 show all first incomplete moments of the experimental barcode frequency distributions. They clearly exhibit an exponential behaviour (linear on a logarithmic scale) over many decades of barcode frequencies, indicating negative binomial distributions across different patients, xenografts, passages, and replicate experiments. By definition of the first incomplete moment, data points with large barcode frequencies outside the negative binomial distribution show up as a strong deviation from the exponential behaviour (see, e.g., red arrowhead in Fig. 2b). This is caused by the barcode frequency entering as a multiplicative term in the definition Eq. (8). Importantly, this does not affect its ability to detect negative binomial distributions for small barcode frequencies.

3 Emergence of negative binomial distributions

What can the barcode frequency distribution tell us about the proliferative dynamics underlying tumour growth? A generic mechanism giving rise to a negative binomial distribution is a process long-known in population dynamics, termed 'critical birth-death process with immigration' (Bailey, 1990; Simons, 2016). Translated into the language of cell population dynamics, such a process can be realized by a population of cells that stochastically divide ('birth') and differentiate ('death') with equal probability ('critical'), with a slow influx of cells from another cell compartment ('immigration') through differentiation. In the tumour context, such a process could naturally arise if there is (i) a slowly cycling glioblastoma stem cell (GSC) compartment at the apex of a proliferative hierarchy that sporadically gives rise to progenitor cells by asymmetric division and (ii) the resulting progenitor population undergoes division and differentiation that are balanced on the population level. Schematically, the dynamics of stem cells (S) and progenitors (P) can be expressed as

$$\tag{10}$$

---
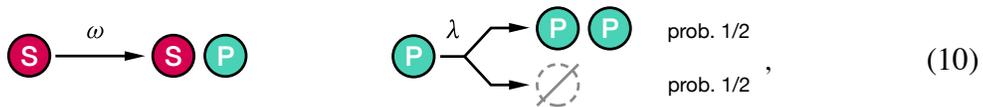
[1]The barcode frequency distribution can be retrieved from the first incomplete moment via the relation $p(n) = \langle n \rangle [\mu(n) - \mu(n-1)]/n$.

where λ is the loss-and-replacement rate of the progenitors and ω is the asymmetric division rate of the stem cells, also called 'immigration rate' since it describes the rate at which cells enter the progenitor compartment. If the immigration rate ω is small compared to the loss-and-replacement rate λ, uniquely barcoded stem-like cells produce clones with a negative binomial barcode frequency distribution, Eq. (7). To show this, we describe the cell fate dynamics shown in scheme (10) as independent Poisson processes with rates ω and λ, respectively. Formally, the corresponding master equation that governs the dynamics of the probability $P = P(n, t)$ to find n progenitor cells in a clone derived from a single uniquely labelled S-type cell,

$$\frac{\partial P}{\partial t} = \left\{ \omega(\hat{E}^- - 1) + \frac{\lambda}{2}(n-1)\hat{E}^- + \frac{\lambda}{2}(n+1)\hat{E}^+ - \lambda n \right\} P \, , \tag{11}$$

where we have introduced the ladder operators $\hat{E}^\pm$, defined by $\hat{E}^\pm P(n, t) = P(n \pm 1, t)$. The first term in brackets describes the asymmetric division of a single S-type cell whereas the remaining three terms describe symmetric division and death of P-type cells[2]. Note that asymmetric division of S-type cells leaves the number of S-type cells unchanged so that it is sufficient to only describe the number n of P-type cells.

The master equation (11) describes the dynamics of S-type and P-type cells shown in the scheme (10) as independent Poisson processes. An analytical solution can readily be obtained by standard methods (Walczak et al., 2012). For initially no progenitor cells being labelled, $P(n, 0) = \delta_{n,0}$, the exact solution to the master equation (11) is given by the negative binomial distribution

$$P(n, t) = \frac{1}{n!} \frac{\Gamma(\zeta + n)}{\Gamma(\zeta)} \left(\frac{n_0(t)}{1 + n_0(t)} \right)^n \left(1 - \frac{n_0(t)}{1 + n_0(t)} \right)^\zeta \, , \tag{12}$$

where $n_0(t) = \lambda t/2$, the dimensionless parameter $\zeta = 2\omega/\lambda$ is the ratio of immigration rate and progenitor loss-and-replacement rate, and $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \, du$

---

[2]The structure of the master equation (11) can be understood by considering, for instance, a reduced dynamics that only describes asymmetric divisions of the S-type cell. This amounts to setting $\lambda = 0$ in Eq. (11) which yields the reduced equation $\partial P/\partial t = \omega P(n-1, t) - \omega P(n, t)$, where we have used the definition of the ladder operator $\hat{E}^-$. This equation describes the rate of change of the probability $P(n, t)$ to find n P-type cells. A state with n P-type cells can only be reached if there are already $n-1$ P-type cells and an asymmetric division of an S-type cell occurs, giving rise to another P-type cell. The corresponding contribution $\omega P(n-1, t)$ to the rate of change $\partial P/\partial t$ is given by the probability $P(n-1, t)$ to find the system in the state $n-1$ multiplied by the rate ω of asymmetric divisions. Conversely, the state with n P-type cells is left if another asymmetric division of the S-type cell occurs, raising the number of P-type cells to $n+1$. The analogous contribution $-\omega P(n, t)$ enters with a negative sign as it describes the process of leaving the state n. The other terms in the full master equation (11) follow the same logic. For more detailed reviews on general master equations and birth-death processes, we refer the reader to standard textbooks (Gardiner, 2009; Bailey, 1990)

is the Gamma function. On average, barcode frequencies grow linearly in time,

$$\langle n(t) \rangle = 1 + \omega t , \tag{13}$$

where $\langle n(t) \rangle = 1 + \sum_n n P(n,t)$ is the average barcode frequency with the first term accounting for the stem cell. For small immigration rates ω, the barcode frequency distribution of cell populations with at least one progenitor cell, given by $P_*(n,t) = P(n,t)/(1 - P(0,t))$, acquires the form Eq. (7),

$$P_*(n,t) = \frac{1}{\ln(1 + n_0)} \frac{1}{n} \left(\frac{n_0(t)}{1 + n_0(t)} \right)^n + \mathcal{O}(\omega)$$
$$\approx \frac{1}{\mathcal{N}_0(t)} \frac{e^{-n/n_0(t)}}{n} , \tag{14}$$

where $\mathcal{N}_0(t) = \ln n_0(t)$. For non-vanishing but small immigration rates ω, the resulting barcode frequency distribution is still well-approximated by Eq. (14). Thus, the dynamics (10) generically give rise to negative binomial barcode frequency distributions and hence are the starting point for our quantitative analysis.

Comparison with clone size distributions emerging from engrained proliferative heterogeneity

Could a negative binomial barcode frequency distribution also be caused by engrained proliferative heterogeneity instead of equipotency? To address this question, let us consider a large population of clones in which the cell of each clone i undergoes loss and replacement with clone-specific probabilities. For concreteness, we consider the following cell fate dynamics in which each cell undergoes loss and replacement with different probabilities,

$$\tag{15}$$

The parameter δ_i determines whether cell i is primed for proliferation ($\delta_i > 0$) or loss ($\delta_i < 0$). The average size of a clone derived from cell i evolves according to $\langle n_i(t) \rangle = e^{2\delta_i \lambda t}$ and on average, clones will thus either grow exponentially ($\delta_i > 0$) or die out ($\delta_i < 0$). In this picture, engrained proliferative heterogeneity is captured by a broad distribution of the δ_i, so that some clones expand faster than others while some clones die. For a clone with a given δ_i, the resulting surviving clone size distribution at large times is exponential (Bailey, 1990),

$$p(n|\alpha_i) \simeq \alpha_i e^{-\alpha_i n} . \tag{16}$$

with an exponent α_i that depends on the proliferative potential and on time. Hence, a distribution in engrained proliferative advantages δ_i entails a distribution in the

shape parameter α_i of the clone size probabilities for the different clones. As an example, let us consider the distribution of α_i at a fixed time $t = t_0$. For simplicity, we here consider a Gamma distribution[3] for α, which ensures that $\alpha > 0$,

$$\bar{p}(\alpha) = \frac{\alpha^{m-1} e^{-m\alpha/\alpha_0}}{(\alpha_0/m)^m \Gamma(m)} \; . \tag{17}$$

The clone size distribution resulting from this distribution of clone size scales is given by

$$p(n) = \int_0^\infty p(n|\alpha)\, \bar{p}(\alpha)\, d\alpha = \frac{\alpha_0}{(\alpha_0 n/m + 1)^{m+1}} \; , \tag{18}$$

which asymptotically has the power law behaviour $n^{-(m+1)}$ and is therefore distinctly different from the negative binomial form $e^{-n/n_0}/n$. Which distribution of proliferative potentials would be needed to generate a negative binomial clone size distribution under these circumstances? In fact, a negative binomial form can only be obtained under very artificial conditions: the distribution for α would have to take the non-normalizable discontinuous form $\bar{p}(\alpha) \propto \alpha^{-1}\Theta(\alpha - \alpha_0)$ where Θ is the Heaviside step function; in this case, the clone size distribution would sensitively depend on the position α_0 of the step as it determines the characteristic scale of clone sizes, $p(n) = e^{-\alpha_0 n}/n$. While being simplistic, this minimal model of engrained proliferative heterogeneity illustrates that negative binomial clone size distributions do not generically arise from a mere loss and replacement of clones—rather, the cell fate dynamics have to display certain distinctive features, such as the minimal hierarchy of the type (10), which robustly leads to such clone size distributions.

4   Theoretical model of tumour growth

In Section 3, we have shown how a negative binomical barcode frequency distribution can arise from a single uniquely labelled stem cell at the apex of a critical birth-death process with immigration. However, there are several reasons why growth of glioblastoma as observed in serial transplantation experiments warrant a more comprehensive model: First, the model (10) only considers strictly asymetrically dividing stem cells, leading to linear growth of barcode frequencies on average. However, there is no reason to a priori rule out *symmetric* stem cell divisions, which potentially provide a considerable contribution to tumour growth. Second, in the model (10), loss of the stem cell leads to a remaining progenitor cell population that will not grow on average and will eventually die out (Clayton et al., 2007). In

---

[3]The Gamma distribution as defined in Eq. (17) has mean α_0 and variance α_0^2/m; in the limit of large m, it is approximately equal to a normal distribution with the same mean and variance. For $m = 1$, the Gamma distribution reduces to an exponential distribution.

the serial transplantation experiments, only small fractions of a harvested tumour ($\sim 5\%$) are chosen for reinjection. If clones were indeed maintained by a single stem cell, it would thus be likely that the stem cell is lost upon reinjection, giving rise to a massive loss of barcodes across passages which is not observed in experiments. Third, the model (10) neglects the potential presence of a non-proliferating compartment undergoing apoptosis that may affect the tumour size and composition. This non-proliferating compartment may be the differentiating progeny of the progenitor population or a quiescent progenitor population.

Therefore, in this section, we now formulate a more comprehensive model of glioblastoma growth and study its predictions on tumour growth and composition. Our model makes falsifiable predictions and to compare it with experiments, we introduce a simulation procedure that combines the clonal dynamics with harvesting and reinjection scheme to mimic the experimental procedure (Section 5). Subsequently, we compare our model to experimental data and show that it captures the key features of clonal dynamics (Section 6).

4.1 Stochastic dynamics of cell division and differentiation

Our model of tumour growth describes the dynamics of three cell compartments: a stem-like cell compartment (S), a progenitor compartment (P), and a non-proliferating compartment that may account for differentiating progeny (D). In our model, stem-like cells divide symmetrically with a probability ε and asymmetrically with probability $1 - \varepsilon$. Progenitor cells either divide symmetrically or differentiate into their progeny, both with probability $1/2$, so that division and differentiation are balanced on the population level. The differentiating compartment has a finite lifetime and constitutes the lowest level of the differentiation hierarchy in our model. Schematically, the model can be expressed as
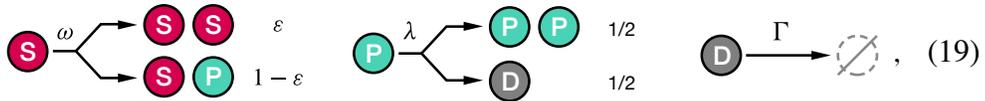
where ω and λ are the division rates of stem cells and progenitors, respectively, and Γ is the apoptosis rate of the differentiating progeny. Defining $P(n^S, n^P, n^D, t)$ as the probability to find n^S stem cells, n^P progenitor cells, and n^D differentiated cells at time t within a clone, we write down a master equation governing the stochastic dynamics in the same spirit as in the previous section,

$$
\begin{aligned}
\frac{\partial P}{\partial t} = &\left\{ \varepsilon\omega \left(n^S - 1\right) \hat{E}_S^- + (1 - \varepsilon)\omega n^S \hat{E}_P^- - \omega n^S + \frac{\lambda}{2} \left(n^P - 1\right) \hat{E}_P^- \right. \\
&\left. + \frac{\lambda}{2} n^P \hat{E}_P^+ \hat{E}_D^- - \lambda n^P + \Gamma \left(n^D + 1\right) \hat{E}_D^+ - \Gamma n^D \right\} P \;,
\end{aligned}
$$

(20)

where we have again used ladder operators defined by $\hat{E}_S^{\pm} P(n^S, n^P, n^D, t) = P(n^S \pm 1, n^P, n^D, t)$ and analogously for the other cell compartments P and D. Together with an initial condition $P(n^S, n^P, n^D, 0)$ that characterizes the initially barcoded population, Eq. (20) permits to compute the clone composition and barcode frequency distribution of our model at any later time. The distribution $p(n, t)$ of total barcode frequencies $n = n^S + n^P + n^D$ is obtained from the joint distribution P by summing over all barcode frequency configurations that lead to a total size n,

$$p(n, t) = \sum_{n'=0}^{n} \sum_{n''=0}^{n-n'} P(n', n'', n - n' - n'', t) . \tag{21}$$

While the full clonal dynamics of our model can only be explored by means of numerical simulations, several important insights about growth and composition of the tumour can be drawn from analytical arguments.

4.2 Composition of the tumour

Using the master equation (20), we can obtain insights into the composition of the tumour in our model, i.e., its relative content of stem cells, progenitors and differentiating progeny. The time evolution of the mean cell numbers is given by

$$\langle \dot{n}^S \rangle = \varepsilon \omega \langle n^S \rangle ,$$
$$\langle \dot{n}^P \rangle = (1 - \varepsilon) \omega \langle n^S \rangle , \tag{22}$$
$$\langle \dot{n}^D \rangle = \tfrac{1}{2} \lambda \langle n^P \rangle - \Gamma \langle n^D \rangle ,$$

where the dot denotes the time derivative. In particular, the evolving clone, while steadily growing, acquires a steady-state composition characterized by a constant relative amount of stem-like cells, progenitor cells, and differentiated cells: Defining the relative cell contents $\phi^S = \langle n^S \rangle / \langle n \rangle$, $\phi^P = \langle n^P \rangle / \langle n \rangle$, and $\phi^D = \langle n^D \rangle / \langle n \rangle$ where $\langle n \rangle = \langle n^S \rangle + \langle n^P \rangle + \langle n^D \rangle$ is the total barcode frequency, this stationary composition satisfies $\dot{\phi}^S = \dot{\phi}^P = \dot{\phi}^D = 0$ and is given by

$$\phi^S = \varepsilon \Omega^{-1} , \qquad \phi^P = (1 - \varepsilon) \Omega^{-1} , \qquad \phi^D = 1 - \Omega^{-1} , \tag{23a,b,c}$$

with Ω being a dimensionless parameter given by

$$\Omega = 1 + \frac{\lambda}{2} \frac{1 - \varepsilon}{\Gamma + \varepsilon \omega} . \tag{24}$$

Eqs. (23a–c) show that the probability ε for symmetric stem cell division determines the relative fraction of stem-like and progenitor cells while the composite parameter Ω determines the relative fraction of the differentiating progeny and the remaining two compartments. Note that in general the ratio of averages does not correspond to the average of the ratio, $\langle n^X \rangle / \langle n \rangle \neq \langle n^X / n \rangle$ for $X = S, P, D$. However, simulations show that Eqs. (23a–c) are excellent approximations for the averages $\langle n^X / n \rangle$ in the considered parameter ranges.

4.3 Tumour expansion

The average growth of a clone (and thus the tumour) can be determined from Eqs. (22) as well. Defining the fold-change in cell number compared to the initial barcode frequency, $\gamma(t) = \langle n(t) \rangle / \langle n(0) \rangle$, we obtain

$$\gamma(t) = e^{\varepsilon \omega t} , \tag{25}$$

given that, from the outset, the tumour has the stationary composition given by Eqs. (23). Hence, the tumour expands exponentially with the growth speed given by the rate $\varepsilon \omega$ of symmetric stem cell divisions.

5 Simulation of transplantation experiments

To capture the dynamics of the serial transplantation experiments, we develop a simulation of the clonal dynamics involving the repeated procedure of injection, unperturbed growth, and harvesting of the tumour. To this end, we use a stochastic simulation algorithm to compute many realizations of the clonal dynamics (Gillespie, 1977). The simulation consists of (i) the injection of a single uniquely labelled cell, (ii) unperturbed clonal dynamics according to the process (19), and (iii) subsequent harvesting of cells for sequencing and reinjection. Key observables such as barcode frequency distributions, numbers of surviving barcodes, and clonal growth are then obtained by performing statistics over the computed realizations.

5.1 Primary injection

To mimic the experimental procedure in our simulation, we start the primary passage by injecting a single labelled S or P cell, each with a probability that reflects the steady-state fractions given in Eqs. (23a,b). Differentiating progeny (represented by the D compartment in our model) are unlikely to survive the process of serial transplantation. The corresponding initial condition for the probability P is thus given by

$$P(n^S, n^P, n^D, 0) = \frac{\phi^S \delta_{n^S,1} \delta_{n^P,0} \delta_{n^D,0} + \phi^P \delta_{n^S,0} \delta_{n^P,1} \delta_{n^D,0}}{\phi^S + \phi^P} . \tag{26}$$

5.2 Tumour growth

After the injection, the clone is subject to unperturbed growth according to Eq. (20) for the duration τ_i of the corresponding passage i.

5.3 Harvesting and reinjection

After each passage, the next passage i is initiated by reinjecting cells harvested from the previous passage $i - 1$. This amounts to setting a new initial condition for the

probability P at the injection time t_i^{inj}, which coincides with the harvesting time $t_{i-1}^{\text{harv}} = \sum_{j=1}^{i-1} \tau_j$ of the previous passage, where τ_i is the passage duration of passage i. Again, assuming that it is unlikely for differentiating cells (D) to survive the process of serial transplantation, only stem-like cells (S) and progenitors (P) are reinjected, each such a cell with a probability p_i^{inj}. The probability p_i^{inj} is determined by requiring that on average, the number n_i^{inj} of injected cells matches the number in the corresponding experiment. The probability p_i^{inj} can be calculated as follows. From Eqs. (22), the average growth of a clone can be calculated for any initial composition of the clone. If only S and P cells are injected, with cell numbers that reflect the stationary composition given by Eqs. (23a,b), the fold change $\gamma(t) = \langle n(t) \rangle / \langle n(0) \rangle$ in cell number is given by

$$\gamma(t) = \Omega e^{\varepsilon \omega t} - (\Omega - 1)e^{-\Gamma t} . \tag{27}$$

with Ω defined in Eq. (24). Hence, the total tumour size after passage i is given by $n^{\text{inj}}\gamma(\tau)$ where n^{inj} is the number of injected cells and τ is the passage duration. Since the composition of the tumour quickly acquires the stationary composition given by Eqs. (23a,b,c) during the passage, the total number of S and P cells upon harvesting is given by $(\phi^S + \phi^P)n^{\text{inj}}\gamma(\tau)$. Therefore, to inject an average of n_i^{inj} cells at the beginning of passage i, the probability p_i^{inj} must be chosen as

$$p_i^{\text{inj}} = \frac{n_i^{\text{inj}}}{(\phi^S + \phi^P)n_{i-1}^{\text{inj}}\gamma(\tau_{i-1})} . \tag{28}$$

Then, the system again evolves according to Eq. (20) till t_{i+1}^{harv} and the same procedure is repeated for the next passage.

5.4 Example

Fig. S1 and Fig. 2d in the main text show numerical examples of the simulation. The upper panel displays different trajectories of barcode frequencies across three passages. Because of stochastic cell fate decisions, clones stochastically grow or shrink during a passage. Therefore, individual trajectories may emerge above and drop below a detection threshold (shaded area in Fig. S1) several times over the course of time (see yellow trajectory in Fig. S1 for an example). While the majority of clones are lost, a few clones grow very large by chance, acquiring several hundreds of cells. After each passage, all barcode frequencies abruptly drop due to harvesting and reinjection of a small sample of the tumour ($\sim 5\%$). From many realizations of the system, statistical properties of the clones such as barcode frequency distributions and correlations can be obtained: the lower panel of Fig. S1 shows, e.g., the average barcode frequency. Note that the average barcode frequency is strongly affected by the majority of clones becoming extinct very quickly while only a few
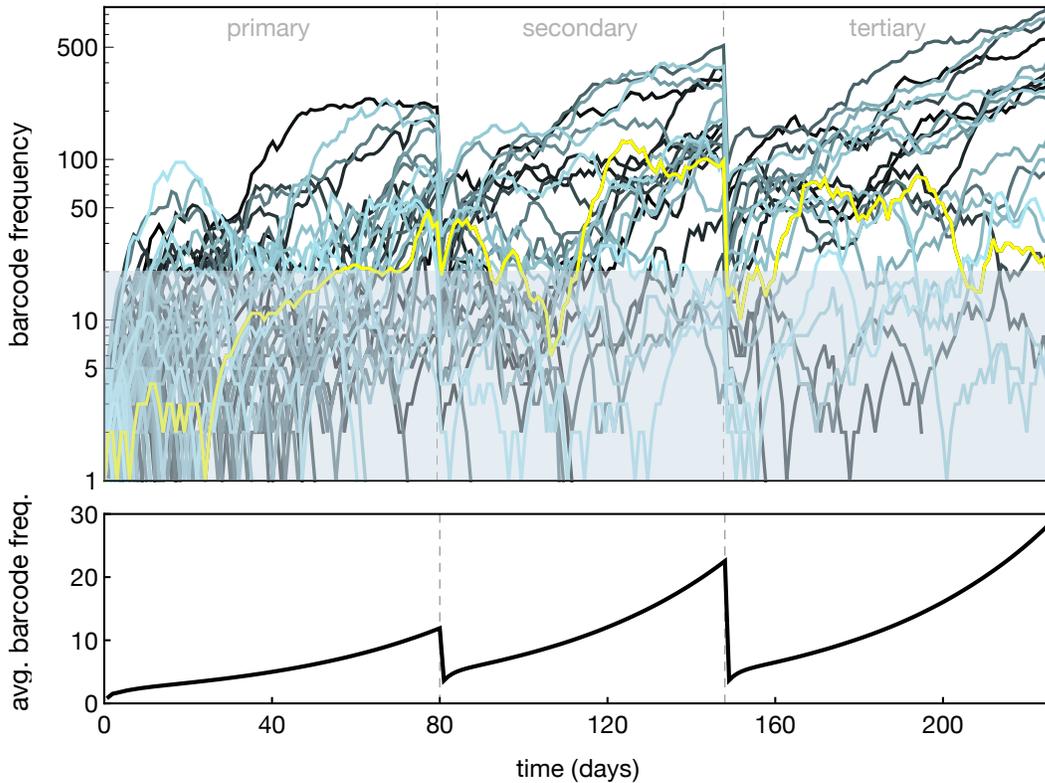
Figure S1 Numerical examples of barcode frequency trajectories across three serial passages on a logarithmic scale. The shaded area indicates an example detection threshold. The yellow curve shows a clone that emerges above and drops below the detection threshold several times. The lower plot shows the average over all trajectories. Parameters are given in Table S3.

clones become large. Fig. 2e in the main text shows the first incomplete moment of the barcode frequency distribution, revealing a negative binomial distribution over many decades as discussed in Section 2. We now use these simulations to systematically compare experimental data with our theory.

6   Comparison of theory and experiments

We now compare our theory with experiments. First, we discuss biologically sensible parameter ranges for our model. We then compare barcode frequency distributions and number of barcodes that survive the serial transplantations with experiments, highlighting that many qualitative key features of our theory are actually independent of the specific choice of parameters.

6.1 Parameter estimates

Can all experiments be characterized by the same set of parameters? Experimental data show a considerable degree of variation in the growth of different xenografts: for instance, referring to Table S4, the tertiary xenograft of the transplantation series labelled $(1, 2, 1)_{719}$ grows by 42-fold over a duration of 55 days, while the tertiary xenograft of $(1, 2, 3)_{719}$ grows by only 26-fold over the longer duration of 78 days, with both xenografts having been derived from the same secondary xenograft $(1, 2)_{719}$. While there are many potential sources for these variations among replicate experiments, this example already indicates that it is not possible to characterize all experiments with a single set of parameters. Rather, it suggests a corresponding degree of variation for the proliferation and differentiation rates of stem-like cells and progenitors as well as the apoptosis rate of the differentiating progeny. Here we aim at constraining plausible parameter ranges using experimental data.

An estimate for the apoptosis rate Γ of the differentiating progeny can be inferred from the steady-state composition of the tumour: we used Ki67 staining of xenograft samples to determine the relative amount of proliferating cells as 50% on average, see Extended Data Fig. 3d. Based on this estimate we fixed the relative amount of progenitor cells among the progenitor population, $\psi = n^{\mathrm{P}}/(n^{\mathrm{D}} + n^{\mathrm{P}})$, as $\psi \approx 0.5$. Using Eqs. (23) and (24), the apoptosis rate Γ can be expressed in terms of ψ and the other parameters as

$$\Gamma = \frac{\lambda}{2} \frac{1}{\psi^{-1} - 1} - \varepsilon\omega \ . \tag{29}$$

Hence, given numerical values for the other parameters ω, ε, and λ, this fixes the value of Γ. For the loss-and-replacement rate λ of the progenitors, we choose an upper bound of $\lambda = 1.5/\mathrm{day}$, motivated by the fact that in mammalian cells, the typical S phase duration is already 5 to 6 hours which constrains the cell cycle speed. In Section 3, we have seen that progenitors have to divide much faster than stem-like cells ($\lambda \gg \omega$) in order to generate the characteristic negative binomial form of the barcode frequency distribution. Therefore, we restrict the stem cell division rate ω to values of at least an order of magnitude less, $\omega \lesssim 0.3/\mathrm{day}$. In our model, overall growth of the tumour crucially depends on the rate $\varepsilon\omega$ of symmetric stem cell divisions (see Sections **??** and 5.3). Considering fast death of the differentiating progeny ($\Gamma \gg \varepsilon\omega$) and a small ratio of symmetric divisions ($\varepsilon \ll 1$), Eq. (27) enables us to estimate the symmetric division rate of the stem cells as $\varepsilon\omega \approx \tau^{-1} \ln \gamma\psi$, where τ is the passage duration, γ is the fold-change in cell number from injection to harvesting and ψ is the amount of proliferating cells introduced above. Given the range of values for γ and τ shown in Table S4, we obtain an estimate for the range of $\varepsilon\omega$ of $0.02 \ldots 0.06/\mathrm{day}$. Since $\varepsilon < 1$, this automatically yields a lower bound of $\omega \gtrsim 0.02/\mathrm{day}$ for the stem cell division rate. In our model, the ratio ε of symmetric

Param.	Range	Description
ω	$0.02\ldots0.3\,\mathrm{d}^{-1}$	stem cell division rate
ε	$<1\ldots20\%$	probability of symmetric stem cell division
λ	$0.5\ldots1.5\,\mathrm{d}^{-1}$	progenitor loss-and-replacement rate
Γ	$0.2\ldots1.5\,\mathrm{d}^{-1}$	death rate of the differentiating progeny

Table S2 Parameter ranges for the model of tumour growth, Eq. (19).

stem cell divisions sets the relative size of the stem cell pool and the progenitor pool, see Eq. (23). Assuming that the stem-like cells form a minority population, we here restrict $\varepsilon \lesssim 20\%$.

A summary of the thus inferred parameter ranges is given in Table S2. To show that these estimates for the parameter ranges are consistent with the clonal behaviour observed in experiments, we now compare numerical solutions of the model with experimental data.

6.2 Barcode frequency distributions

A direct quantitative comparison of barcode frequency distributions is currently not possible because of limitations in experimentally determining absolute barcode frequencies. However, the characteristic functional shape of the barcode frequency distributions is independent of absolute barcode frequencies and can be compared with experiments. To assess the barcode frequency distributions generated by our model, we obtain their first incomplete moment μ as defined in Eq. (8) from Eq. (21). Fig. 2e in the main text shows examples for μ for each passage, obtained from a numerical simulation of 2×10^6 realizations of the system. The linear behaviour over many decades of barcode frequencies indicates a negative binomial size distribution as discussed in Section 2. In fact, we find these negative binomial distributions within a large range of parameters. This linear behaviour is preceded by a short non-linear transient behaviour for very small barcode frequencies that are likely below the experimental detection threshold.

6.3 Barcode survival

The survival of barcodes is reflected by the number of detected barcodes across passages. In experiments, the number of detected barcodes depends on the detection threshold and the fraction of sequenced cells. To obtain a measure for clone survival that is independent of these experimental constraints, we make use of the fact that barcode frequency distributions have the negative binomial form Eq. (7), which entails a characteristic barcode frequency n_0. This enables us to define the number

of clones that exceed a specified fraction θ of the characteristic barcode frequency n_0 as $\sum_{n>\theta n_0} h(n)$, where $h(n)$ is the number of clones with size n. The ratio of clones derived from initially injected barcoded cells that exceed the size θn_0 at a given passage therefore serves as a measure for barcode survival,

$$\beta_\theta = \frac{1}{N_{\mathrm{B}}} \sum_{n>\theta n_0} h(n) , \tag{30}$$

where N_{B} is the number of uniquely barcoded cells injected before the first passage, given by $N_{\mathrm{B}} = \eta n_1^{\mathrm{inj}}$ with η being the labelling efficiency and n_1^{inj} being the number of cells injected.

Fig. S2 shows the results from simulations[4] covering the parameter ranges indicated in Table S2, along with the corresponding experimental data[5]. Density bars show the distribution of values for $\beta_{1/2}$, dots show experimental data points[6]. Clearly, most of the values obtained in the biologically plausible parameter range also capture the experimentally obtained values. Moreover, simulations show a systematic decline of the barcode survival probability with increasing passage number.

6.4 Correlations of barcode frequencies across passages

We now make use of the fact that unique barcoding enables us to identify clones throughout different passages and replicate experiments. A characteristic feature of the clonal dynamics that includes this longitudinal data is the correlation of the size of a uniquely labelled clone across passages, see Extended Data Fig. 4f and Fig. 2h in the main text. To quantify these correlations, we define the normalized

---

[4]A total of 108 parameter sets equally distributed in the parameter ranges for ω, ε, and λ indicated in Table S2 have been used to sample the parameter space. The parameter Γ was fixed according to Eq. (29). Each simulation consists of 100 000 realizations of clones using the passage times and number of injected cells reported in Table S4.

[5]We obtain β_θ from experimental data as follows. Barcode frequency distributions $h(x)$ with x being the relative barcode frequency are generated by binning the experimentally obtained barcode frequencies with a bin size of $(x_{\max} - x_{\min})/100$ where $x_{\max}$ and $x_{\min}$ are the largest and smallest relative barcode frequencyies, respectively. We then fitted the resulting barcode frequency distributions using the negative binomial form $p(x) = \mathcal{N}_0^{-1} e^{-x/x_0}/x$ with n_0 and the normalisation constant $\mathcal{N}_0$ as fit parameters. Since the detection threshold from sequencing may distort the distributions for small barcode frequencies, we truncate the barcode frequency distributions from below (within the first 20 data points) such that the coefficient of determination R^2 of the fit is maximized. This yields the characteristic barcode frequency x_0 and β_θ is readily obtained as $\beta_\theta = N_{\mathrm{B}}^{-1} \sum_{x>\theta x_0} h(x)$. The standard error σ_{x_0} on x_0 obtained from the fit is used to calculate positive and negative errors for β_θ as $\sigma_\beta^\pm = N_{\mathrm{B}}^{-1} \sum_{x>\theta(x_0\pm\sigma_{x_0})} h(x)$.

[6]The value $\theta = 1/2$ was chosen because the corresponding threshold $n_0/2$ lies well above the detection threshold from sequencing and at the same time takes into account most of the acquired data.
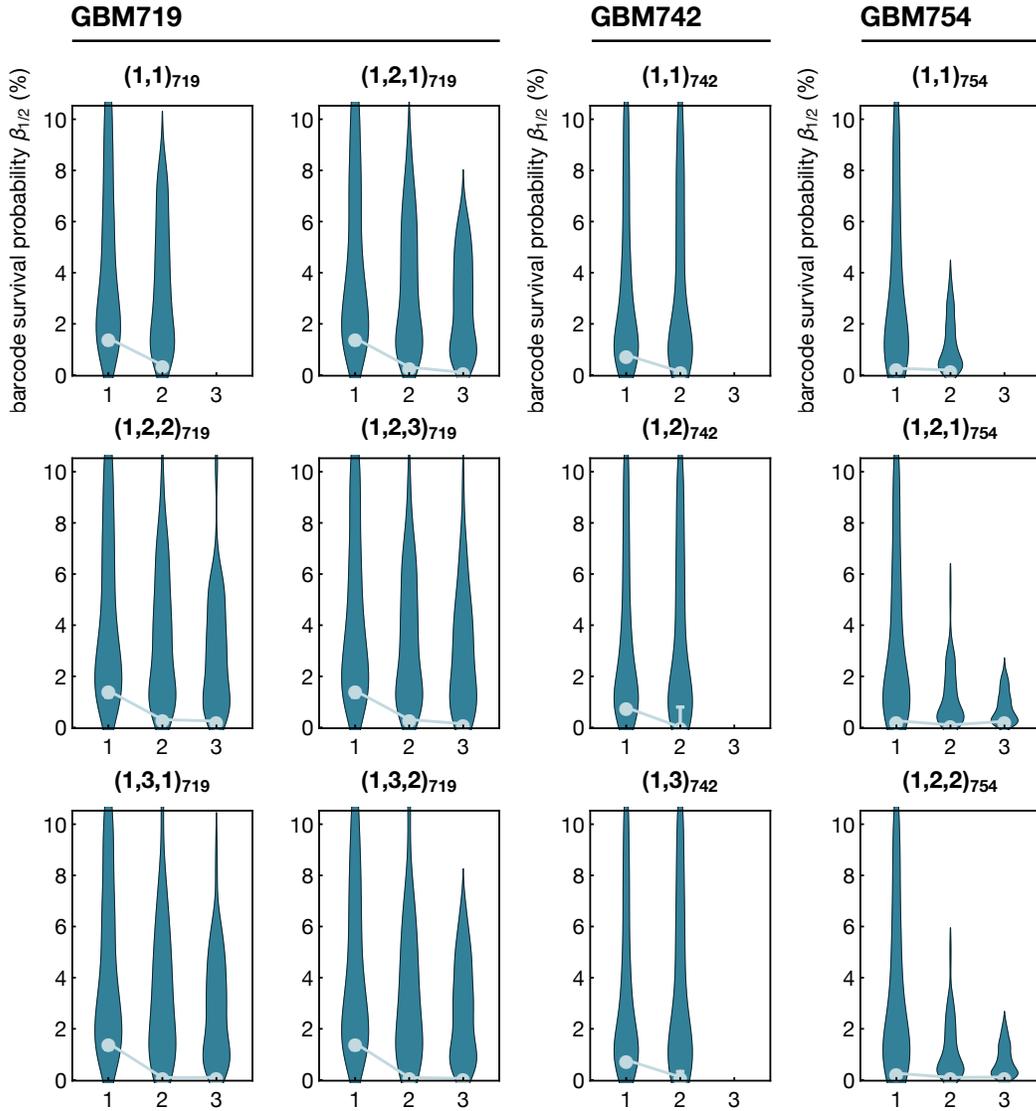
Figure S2 Fraction $\beta_{1/2}$ of initially injected barcodes growing above half of the characteristic barcode frequency $n_0/2$ as defined in Eq. (30) for all experimental trajectories given in Table S4. Density bars show the distribution of simulation results pooled over the parameter ranges indicated in Table S2. Dots show experimental data. The plot titles indicate the experimental trajectory as given in the first column of Table S4.

cross correlation of the barcode frequency for passages i and j as

$$C_{ij} = \frac{\langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle}{\sqrt{\langle n_i^2 \rangle - \langle n_i \rangle^2} \sqrt{\langle n_j^2 \rangle - \langle n_j \rangle^2}} \, , \tag{31}$$

	ω	ε	λ	Γ	φ
Figs. 2d,e,h and S1, S4	$0.15\,\mathrm{d}^{-1}$	15%	$1\,\mathrm{d}^{-1}$	$0.48\,\mathrm{d}^{-1}$	0%
Figs. 3e,f	$0.1\,\mathrm{d}^{-1}$	10%	$1.5\,\mathrm{d}^{-1}$	$0.74\,\mathrm{d}^{-1}$	0.5%

Table S3 Parameter values used for the numerical examples in Figs. S1 and S4 and Figs. 2 and 3 in the main text. The parameters ω, ε, λ, and Γ are introduced in Sec. 4.1; the parameter φ is introduced in Sec. 6.5. These parameter sets are used to illustrate the model behaviour and have therefore been chosen to be located in the center of the biologically plausible parameter ranges indicated in Table S2.

where $n_i = n(t_i^{\text{harv}})$ is the barcode frequency after passage i. The normalized cross correlation C_{ij} takes values between -1 and 1, where $C_{ij} = 1$ indicates perfect correlation of barcode frequencies (i.e., small/large clones in passage i correspond to small/large clones in passage j), $C_{ij} = 0$ indicates that barcode frequencies are completely uncorrelated, and $C_{ij} = -1$ indicates perfect anticorrelation (i.e., large clones in passage i correspond to small clones in passage j and vice versa).

Fig. S3 shows a comparison of the correlations for the same simulations and experimental data sets as in Fig. S2. Density bars show the distribution of values for the cross correlations C_{ij}, dots show experimental data points. Without a fine tuning of the parameters, the theoretically computed cross correlations not only cover the experimentally obtained values in most cases but also clearly capture the correct trend between different pairwise comparisons within a particular injection series. In the case of the GBM754 experiment, deviations from experimental results is likely due to the comparably small number of detected clones which makes the cross correlation a less reliable measure; nevertheless, that the trend of correlations is largest between the secondary and tertiary passage is correctly captured.

6.5 Effects of chemotherapy

In the main text, we observed that the clonal behaviour after treatment of xenografts with temozolomide (TMZ) can be characterized by two distinctive groups of small and large clones (termed Group A and Group B, respectively), see Fig. 3a–d. There, we hypothesized that such a behaviour is consistent with a subset of clones exhibiting a resistance to apoptosis. To assess whether our results support this scenario, we modified the simulation such that with a certain probability φ, a clone's differentiating progeny does not die off during the second passage ($\Gamma = 0$ for the respective clones). Fig. 3e,f in the main text shows the resulting correlations of barcode frequencies for the clones resisting apoptosis (blue dots, $\Gamma = 0$) and clones following the unperturbed dynamics (green dots, $\Gamma \neq 0$) for an example simulation with parameters given in Table S3. Indeed, the resulting behaviour recapitulates the experi-
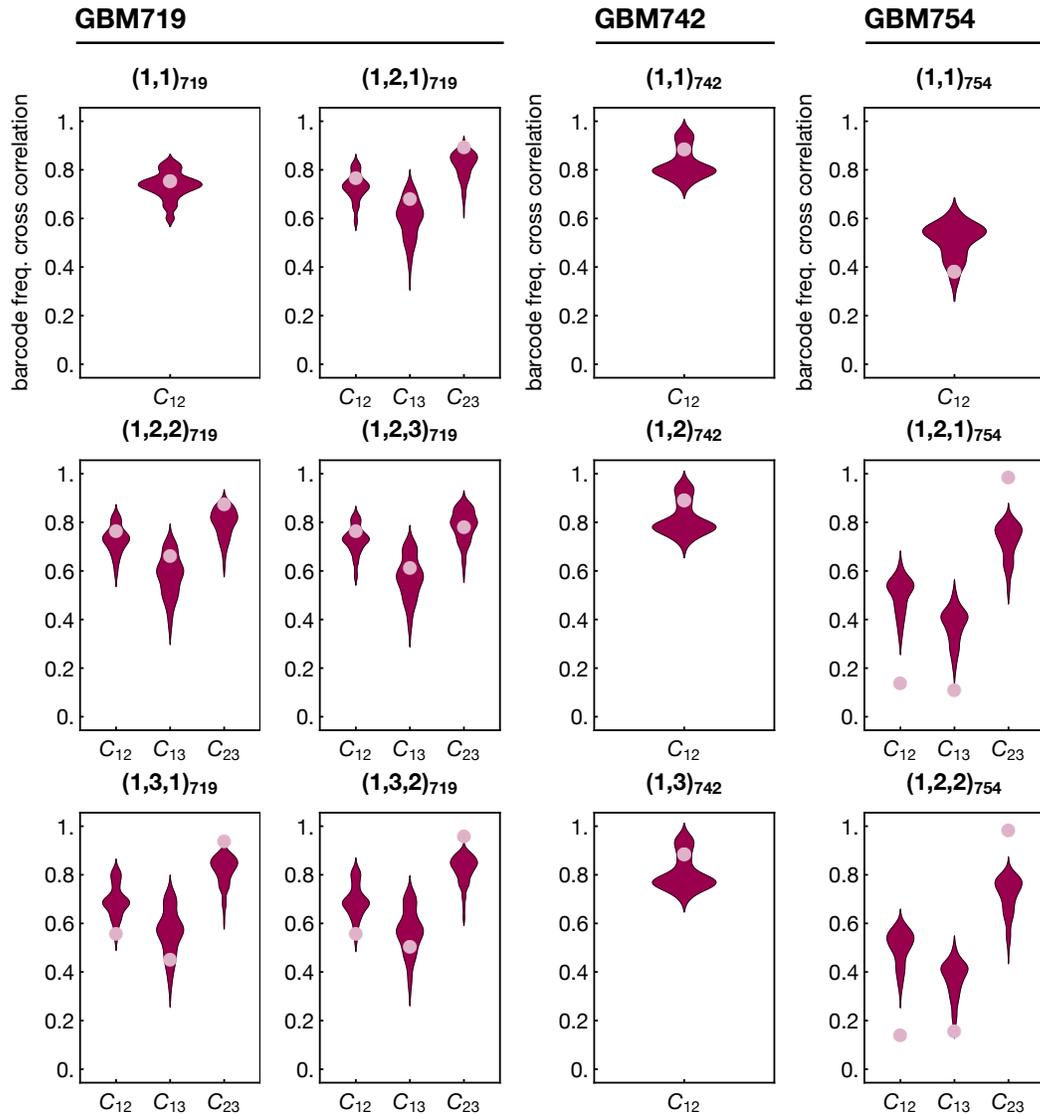
Figure S3 Barcode frequency cross correlations C_{ij} as defined in Eq. (31) for all experimental trajectories given in Table S4. Density bars show the distribution of simulation results pooled over the parameter ranges indicated in Table S2. Dots show experimental data. The plot titles indicate the experimental trajectory as given in the first column of Table S4.

mental findings: two clusters of small and large clones, respectively, with the size of large clones being positively correlated between subsequent passages, see Figs. 3a,b in the main text. The qualitative features of these correlations robustly appear without fine-tuning and within a large range of parameters, supporting that resistance to apoptosis of a subset of clones generically leads to the observed behaviour.

7 Exome deep sequencing as a window on the mutational heterogeneity and clonal dynamics of GBM cells

To probe the mutational heterogeneity of the parent tumour and its evolution over time, we applied exome sequencing to xenografts from GBM719, focusing first on passage (p)2 and 3 of the untreated system. This analysis identified 546 mutations at p2 with variable allele frequencies (VAFs) that were above the threshold of detection, and 112 at p3. Analysis of the distribution of VAFs revealed a wide variation, with the majority clustered around the threshold value while some appeared to be clonally fixed within the population with VAFs of 0.5 or more. Note that copy number variation can amplify VAFs above the value of 0.5, the value expected for a heterozygous point mutation that has become clonally fixed across of the population. Comparison of the mutational signature between p2 and p3 identified 68 mutant clones that were shared by both groups and therefore likely to be present in the parental tumour, emphasizing the mutational heterogeneity of both the parent tumour sample, and its conservation in the xenograft model.

As well as indicating the mutational heterogeneity of the tumour sample, the VAF also carries quantitative information on the relative abundance of point mutations within a sample and therefore carries information about the relative size of host mutant clones. Indeed, such data sets can often be used to identify cancer drivers and, in some cases, the phylogeny of mutations that drive non-neutral transformation (Williams et al., 2016; Eirew et al., 2015). However, in the present context, the current barcoding study indicates "neutral" competition between growing mutant clones suggesting that the vast majority of heterozygous point mutations, even when they occur in cancer genes, may leave the fate behaviour of tumour cells largely unperturbed. In this case, we can instead use point mutations as a surrogate clonal mark from which information on clonal dynamics of tumour cells can be inferred. However, in contrast to cellular barcoding, where the clonal mark is created at a given instant in time, mutations occur sporadically leading to modified "clonal" distributions. As a result, the VAFs obtained from exome sequencing represent a product of both the underlying fate dynamics of the mutant cells within the sample and the mutational dynamics (Simons, 2016), involving the ongoing acquisition of new point mutations and copy number variations. Nevertheless, when copy number variation is low, such approaches can be used to quantify cell fate behaviour, as exemplified by a recent study of stem cell dynamics in physiological normal human epidermis obtained from punch biopsies of eyelid epidermis (Martincorena et al., 2015; Simons, 2016).

To develop a similar approach here, we reasoned that biopsies from primary tumours are likely to contain geographically restricted mutations (Johnson et al., 2014), further compounding the potential complexity of the VAF distribution. How-

ever, since normal cells are unlikely to survive passaging through the xenograft, we reasoned that VAFs obtained at p2 and p3 were likely to be rooted in the tumour-maintaining population. To address this data, we first considered the qualitative behaviour of the raw VAF distributions in both control (untreated) samples from p2 and p3. If, for a given locus, mutations of both alleles occur at a negligible rate, a VAF of 0.5 indicates a mutation present in the entire cell population and is therefore fixed across the population. Indeed, the VAF distributions in both samples (Extended Data Fig. 6d) exhibit an abundance of small clones as well as a smaller peak at VAFs of around 0.5, which likely corresponds to mutations that have already become fixed in the population after the respective passage.

Examining the correlations of VAFs between passages in xenografts (Extended Data Fig. 6e), we found a population of larger clones that are present after both passages, as expected for mutations that have become fixed (or almost fixed) at the end of p2. Alongside these clones, we also found both (i) clones that became extinct (or, more accurately, fell below the threshold of the deep sequencing) during repopulation and expansion in p3 as well as (ii) new clones that emerge during p3. If we assume that these new mutant clones arise from new mutations acquired during p3 (rather than from pre-existing clones that grew above the detection threshold), we can use the dynamics inferred from the barcoding to derive expected features of the VAF distribution of these newly-generated clones.

To predict the large-scale dependence of the VAF distribution, we adapted our simulation to take into account random "induction" of clones through mutations during the tertiary passage[7]. Model simulations suggested that the resulting VAF distribution again approximates a negative binomial form, or, equivalently, acquires an approximately exponential first incomplete moment consistent with experiment (Extended Data Fig. 6f-g, Fig. S4). Remarkably, focusing on the first incomplete moment of the 44 clones that emerge during the tertiary passage, we find that the first incomplete moment of the VAF distribution again reveals an exponential distribution (Extended Data Fig. 6g), in accordance with expectations from the barcoding study. By comparison, the TMZ-treated samples show (i) a much larger number of newly acquired mutations during p3 (Extended Data Fig. 6d) and (ii) a broad distribution

---

[7]Considering a constant mutation rate for each locus in each cell (Simons, 2016), the probability for a mutation to occur is proportional to the instantaneous number of cells in the tumour. Therefore, knowing that in our model specified by Eq. (20), the time-dependent fold-change in cell number is given by $\gamma(t) = e^{\varepsilon \omega t}$, see Eq. (25), we reasoned that the time-dependent probability distribution for a mutation to have occured during the tertiary passage is given by $p_{\text{ind}}(t) = \gamma(t) / \int_0^{\tau_3} \gamma(t) \, dt$ with $0 \leq t \leq \tau_3$ where $t = 0$ refers to the start of the passage and where τ_3 is the passage duration of the tertiary passage. Hence, for each clone, we drew a time t_{ind} from the distribution p_{ind} and simulated the respective clone for the time $\tau_3 - t_{\text{ind}}$, i.e., the remaining time from induction during the tertiary passage to the end of the passage. We then obtained the clone size distribution and first incomplete moment from the resulting clone population.
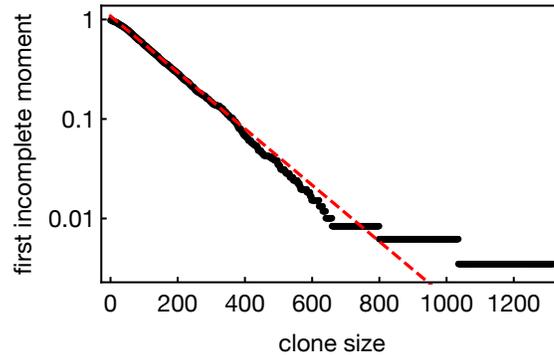
Figure S4 First incomplete moment of the clone size distribution obtained from a simulation of 10^5 clones with random induction times during the tertiary passage[7]. Parameters are given in the first row of Table S3. The red dashed line shows an exponential fit of the first incomplete moment.

of VAFs after p3 with a considerable subset of clones displaying VAFs larger than 0.5 (Extended Data Fig. 6d), both pointing at a treatment-induced higher genomic variability.

Although the agreement between the theoretical prediction based on the barcoding data and experiment is encouraging, we must also exercise some caution. While correction of VAFs to account for copy number variation (CNV) is already challenging in the parent tumour, with new mutations, the challenge is even greater. When CNV occurs before the mutation, the VAF provides a faithful read-out of clone size; where it occurs afterwards, the VAF is corrupted by the amplification. The correlation between VAFs associated with shared mutations between p2 and p3 of the control xenograft suggests that CNV may be rather infrequent as compared to the clonal dynamics, consistent with the systematic behaviour of the measured clone size distribution as predicted by a conserved proliferative hierarchy. In addition, we repeated the same analysis only taking genomic regions that are predicted to be diploid within each sample based on exome sequencing. After filtering, the VAF distributions continue to conform to the negative binomial (Extended Data Fig. 6h-i). However, a more detailed quantitative analysis would require a comprehensive investigation and understanding of the interplay between tumour growth, mutational dynamics and, indeed, chemotherapy-induced mutation (Johnson et al., 2014), which are beyond the scope of the current study.

8 Remarks

Here we have introduced a theoretical model of human glioblastoma (GBM) growth based on a critical birth-death process with immigration, describing the stochastic cell fate dynamics of a proliferative hierarchy with glioblastoma stem cells (GSCs) at the apex. Our model is able to robustly capture key features of the clonal dynamics assessed experimentally: importantly, it explains the characteristic negative binomial barcode frequency distributions across all serial passages observed in experiments. Moreover, comparison of (i) the number of surviving barcodes across serial passages and (ii) correlations of barcode frequencies between serial passages show that the inferred parameter range covers the observed behaviour in the overwhelming majority of cases.

Note that the model presented here is still a minimal model in the sense that more complex alterations and refinements are conceivable. These may include a slight imbalance between loss and replacement of progenitors as well as multiple progeny compartments. Also, small amounts of cell death may occur in the stem cell and progenitor compartments. However, if the death rate was of comparable size (or larger) than the rate of symmetric proliferation, we would expect a massive loss of clones. If, on the contrary, cell death only represents a small contribution relative to the symmetric proliferation rate, it could be accounted for by an effective adjustment of the other model parameters that, e.g., determine the net tumour growth and would only be visible in subtle changes of the barcode frequency distributions that are impossible to detect in the experimentally given distributions. Importantly, these alterations do not change the basic characteristics of our model. Moreover, we have neglected the spatial aspect of tumour growth and potential ongoing driver gene mutations (Michor et al., 2006; Waclaw et al., 2015), assuming that cell division and loss-and-replacement occur at constant rates as the tumour expands within the brain. Despite its simplicity, the fact that our model is able to capture the main features of the clonal dynamics indicates a remarkably simple proliferative behaviour of human GBM despite the genomic variability of GBM cells.

References

Bailey, N. T. J. (1990). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. A Wiley publication in applied statistics. Wiley.

Blanpain, C. and Simons, B. D. (2013). Unravelling stem cell dynamics by lineage tracing. *Nat. Rev. Mol. Cell Biol.*, 14:489–502.

Clayton, E., Doupé, D. P., Klein, A. M., Winton, D. J., Simons, B. D., and Jones, P. H. (2007). A single type of progenitor cell maintains normal epidermis. *Nature*, 446:185–189.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J. L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A. J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C. J., Hansen, C., Marra, M. A., Caldas, C., Shah, S. P., and Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426.

Gardiner, C. (2009). *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer.

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.*, 93555(1):2340–2361.

Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C. Y., Fouse, S. D., Yamamoto, S., Ueda, H., Tatsuno, K., Asthana, S., Jalbert, L. E., Nelson, S. J., Bollen, A. W., Gustafson, W. C., Charron, E., Weiss, W. A., Smirnov, I. V., Song, J. S., Olshen, A. B., Cha, S., Zhao, Y., Moore, R. A., Mungall, A. J., Jones, S. J. M., Hirst, M., Marra, M. A., Saito, N., Aburatani, H., Mukasa, A., Berger, M. S., Chang, S. M., Taylor, B. S., and Costello, F. F. (2014). Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science*, 343(6167):189–193.

L. V. Nguyen, M. Makarem, *et al*. (2014). Clonal Analysis via Barcoding Reveals Diverse Growth and Differentiation of Transplanted Mouse and Human Mammary Stem Cells. *Cell Stem Cell*, 14:253–263.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

Michor, F., Iwasa, Y., and Nowak, M. A. (2006). The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proc. Natl. Acad. Sci. USA*, 103(40):14931–14934.

Simons, B. D. (2016). Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl. Acad. Sci. USA*, 113:128–133.

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A. (2015). A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264.

Walczak, A. M., Mugler, A., and Wiggins, C. H. (2012). Analytic methods for modeling stochastic regulatory networks. *Meth. Mol. Biol.*, 880(Chapter 13):273–322.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, 48(3):238–244.

ID	Passage	inj. cells n^{inj}	τ	surv. prob. $\beta_{1/2}$	growth γ
		GBM719			
$(1)_{719}$	Primary	1.25×10^5	79 d	$1.44^{+0.}_{-0.26}\%$	40
$(1,1)_{719}$	— Secondary	3×10^5	64 d	$0.32^{+0.06}_{-0.07}\%$	7.7
$(1,2)_{719}$	— Secondary	3×10^5	68 d	$0.31^{+0.08}_{-0.07}\%$	18
$(1,2,1)_{719}$	—— Tertiary	3×10^5	55 d	$0.13^{+0.07}_{-0.03}\%$	42
$(1,2,2)_{719}$	—— Tertiary	3×10^5	70 d	$0.25^{+0.04}_{-0.03}\%$	43.3
$(1,2,3)_{719}$	—— Tertiary	3×10^5	78 d	$0.18^{+0.04}_{-0.03}\%$	26
$(1,3)_{719}$	— Secondary	3×10^5	89 d	$0.03^{+0.05}_{-0.01}\%$	16.3
$(1,3,1)_{719}$	—— Tertiary	3×10^5	66 d	$0.13^{+0.03}_{-0.}\%$	15.1
$(1,3,2)_{719}$	—— Tertiary	3×10^5	62 d	$0.07^{+0.03}_{-0.01}\%$	47.7
		GBM754			
$(1)_{754}$	Primary	1.25×10^5	99 d	$0.26^{+0.}_{-0.}\%$	26
$(1,1)_{754}$	— Secondary	6×10^4	79 d	$0.19^{+0.}_{-0.}\%$	42.7
$(1,2)_{754}$	— Secondary	6×10^4	86 d	$0.1^{+0.}_{-0.}\%$	98.3
$(1,2,1)_{754}$	—— Tertiary	6×10^4	72 d	$0.24^{+0.}_{-0.}\%$	56
$(1,2,2)_{754}$	—— Tertiary	6×10^4	73 d	$0.11^{+0.}_{-0.}\%$	31.7
		GBM742			
$(1)_{742}$	Primary	2.4×10^4	78 d	$0.78^{+0.}_{-0.}\%$	530
$(1,1)_{742}$	— Secondary	3×10^5	43 d	$0.13^{+0.06}_{-0.}\%$	5.7
$(1,2)_{742}$	— Secondary	3×10^5	47 d	$0.01^{+0.78}_{-0.01}\%$	1.8
$(1,3)_{742}$	— Secondary	3×10^5	50 d	$0.12^{+0.22}_{-0.03}\%$	8.5

Table S4 Experimental data sets used to compare with theory. Here, n^{inj} is the number of injected cells, τ is the passage duration, s is the fraction of cells sequenced, $\beta_{1/2}$ is the fraction of initially injected barcodes growing above half of the characteristic barcode frequency $n_0/2$, as defined in Eq. (30), and γ is the estimated fold-change in cell number between injection and harvesting, which quantifies tumour growth. In all cases, cells were harvested and injected from the ipsilateral side.