Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis

Benjamin D. Simons * † ‡

*Cavendish Laboratory, Department of Physics, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK,†Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK., and ‡Wellcome Trust-Medical Research Council Stem Cell Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Using deep sequencing technology, methods based on the sporadic acquisition of somatic DNA mutations in human tissues have been used to trace the clonal evolution of progenitor cells in diseased states. However, the potential of these approaches to explore cell fate behavior of normal tissues and the initiation of preneoplasia remain underexploited. Focusing on the results of a recent deep sequencing study of eyelid epidermis, we show that the quantitative analysis of mutant clone size provides a general method to resolve the pattern of normal stem cell fate, and to detect and characterize the mutational signature of rare field transformations in human tissues, with implications for the early detection of preneoplasia.

stem cells | DNA sequencing | epidermis | cancer

Advances in genetic lineage tracing in transgenic animal models have provided important insights into the proliferative potential and fate behavior of stem and progenitor cell populations in normal tissues [1, 2]. As well as providing constraints on the mechanisms that regulate stem cell self-renewal, these approaches have established a quantitative framework to address tumor initiation and progression [3, 4, 5, 6]. However, studies based on the clonal activation of oncogenes in animal models can fail to recapitulate the natural processes that lead to neoplasia in human tissues. In recent years, there has been increasing emphasis on the characterization of cancer genomes in human tissues and their potential to elucidate the pathways involved in tumor progression [7, 8, 9, 10, 11, 12, 13, 14]. Although these studies have revealed a range of cancer genes [15], the heterogeneity and evolutionary diversity of the tumor environment make the separation of driver and passenger mutations challenging.

Against the trend to focus human tumor samples, a recent study has employed ultra-deep exome sequencing to determine the mutational profile of normal human eyelid epidermis [16]. In the course of DNA replication, all dividing cells are subject to random single-nucleotide polymorphisms (SNPs). If the mutation rate is sufficiently low that their acquisition at a given locus in a cell subpopulation is typically associated with a single event, they confer a potentially unique hereditary label on cells allowing the fate of their progeny to be traced over time. By resolving the mutant allele fraction in a biopsy using deep sequencing, the relative size of mutant clones can be inferred. A similar approach based on the spontaneous acquisition of mitochondrial DNA mutation has been used to address progenitor cell fate in human airways and intestinal epithelia [17, 18]. To assess the selective growth advantage of different mutations in normal epidermis, Martincorena *et al.* compared the dN/dS ratio and average size of clones derived from mutations in genes associated with cancer drivers with those associated with synonymous mutations in non-driver genes [16]. Their analysis showed a significant increase in the abundance and average size of clones that bear mutations in NOTCH1 and TP53 when compared with the ensemble of apparently neutral mutations, while mutations in other drivers such as FAT1, NOTCH2, and NOTCH3 were not significantly increased. Based on these findings, the study reached the striking conclusion that cancer genes are under strong positive selection, even in physiologically normal skin. Yet, paradoxically, despite the apparent survival advantage, average clone sizes even in TP53 mutant clones were only a modest factor of 2 larger that the ensemble average, suggesting that the degree of clonal dominance may be limited.

At first sight, one might expect that the relative abundance of gene-specific point mutations could reveal whether they confer a selective survival advantage. However, while variations in the observed frequency of SNPs will arise from the positive/negative selection of somatic mutations, they may also be intrinsic (germline-derived) making their functional significance at different sites difficult to assess [19, 20] (Fig. S1). Equally, the value of average mutant clone sizes is diminished by their sensitivity to tails of the size distribution, which can be compromised by the resolution limit of sequencing or statistical fluctuations due to rare events. Similar effects may compromise the dN/dS ratio, a measure of the relative abundance of non-synonymous to synonymous mutations [21]. However, by analyzing the full probability distribution of mutant clone sizes, and drawing upon knowledge of adult stem cell self-renewal strategies [1, 22], we show that quantitative insights can be gained into the dynamics and fate behavior of mutant clones, providing access to both the normal state properties of tissue-maintaining cells, and their dynamics following premalignant transformation. In doing so, we offer a new perspective on the deep sequencing study of Martincorena *et al.*

Significance

The sporadic acquistion of somatic DNA mutation confers a hereditary label that can be used to trace the fate behavior of cells in normal and diseased states. Applied to human tumor samples, DNA deep sequencing methods have revealed the landscape of somatic mutations and have identified a repertoire of genes implicated in cancer. By adapting statistical methods used to analyze lineage tracing data in transgenic animal models, we use the example of epidermis to show how deep sequencing data can provide quantitative insight into the self-renewal properties of normal human tissues, and can serve as a platform to define rare non-neutral field transformations.

Reserved for Publication Footnotes

Deep sequencing as a clonal marker in human epidermis. In mammals, skin is composed of a multilayered sheet of keratinocytes interspersed with hair follicles, sebaceous glands and sweat glands [23]. Lineage tracing studies using transgenic mouse models have revealed a surprising degree of compartmentalization, with the turnover of hair follicle, sebaceous gland and interfollicular epidermis (IFE) maintained by independent stem cell populations [24]. In IFE, proliferation is confined to cells in the basal layer that adhere to an underlying basement membrane (Fig. 1A). On commitment to terminal differentiation, basal cells detach from the basement membrane and transfer into the suprabasal layers before reaching the epidermal surface from where they are shed. In homeostasis, the progenitors that maintain IFE must undergo asymmetric self-renewal so that, following division, on average one cell remains in the self-renewing compartment, while the other commits to differentiation either directly, or via a transit compartment with strictly limited proliferative potential. Such asymmetry may be invariant, enforced at the level of each and every cell division, or it may be achieved only at the level of the progenitor population (SI text).

Beginning with the work of Mackenzie and Potten, early studies of IFE maintenance in mouse placed emphasis on a stem/transit-amplifying cell paradigm in which long-lived slow cycling stem cells give rise to short-lived progenitors that undergo a limited series of symmetric division before terminal differentiation [25, 26]. Later, quantitative lineage tracing studies based on inducible genetic labeling revealed that murine epidermal maintenance relies instead upon the turnover of a basal progenitor pool that conforms to a process of "population asymmetry" in which their stochastic loss through terminal division is perfectly compensated by the duplication of neighbors [27, 28, 29, 30] (Fig. 1B). Whether the repair of murine epidermis involves a transient adjustment in the fate behavior of the progenitor pool, or is engineered by the activity of a second quiescent "reserve" stem cell population remains the subject of debate. In human, *in vitro* colony forming assays, as well as transplantation and marker based studies, point at engrained proliferative heterogeneity in the basal layer of IFE [31, 32, 33, 34]. However, in the absence of *in vivo* lineage tracing assays, the nature of stem cell self-renewal and tissue maintenance remains in question.

The resolution of cell fate behavior in mouse IFE relied upon the observation of "scaling" behavior of the clone size distribution following genetic pulse-labeling [27, 28, 29, 30, 35]. According to their stochastic fate behavior (Fig. 1B), as progenitors compete neutrally for survival, the density of clones (number per unit area) progressively diminishes, while the average size of surviving clones steadily increases (linearly with time) so that the overall number of marked cells remains constant over time (SI Text and Figs. 1C and S2). Yet, despite their continual increase in size, the chance of finding a surviving clone larger than some multiple of the average remains constant, and defined by an exponential distribution (Fig. S2). Combined with the overall conservation of labeled cell number, this phenomenon of scaling provides a robust, parameter-free signature of neutral cell competition and equipotency of the tissue-maintaining population [35]. Although the exponential size dependence is particular to epithelial (and volumnar) tissues, the phenomenon of scaling applies generically to all cycling adult tissues supported by population asymmetry (SI Text). As a result, the same general approach has been used successfully to explore stem cell fate behavior in other tissues and organisms [1].

In contrast to genetic labeling approaches, where the induction frequency can be controlled through the dose dependence of the drug-inducing agent, clonal marking by somatic mutation involves a sequence of sporadic events masking the age of individual clones (Fig. 1D). Fortunately, under conditions of neutral competition, quantitative information on the fate behavior of the self-renewing population can still be recovered. In particular, if the pattern of stochastic progenitor cell fate observed in mouse IFE (Fig. 1B) were extrapolated to human then, following the continual "induction" of clonally marked cells through the acquisition of somatic mutation, the probability of finding a mutant clone with $n > 0$ progenitors in a biopsy of a patient of age t would be independent of the (presumed unchanging) mutation rate and given by [17] (Figs. 1E,F and SI Text)

$$P_{n>0}(t) \approx \frac{1}{\ln(r\lambda t)} \frac{e^{-n/r\lambda t}}{n} \qquad [1]$$

where λ represents the division rate and $r\lambda$ denotes the loss/replacement rate of basal progenitors (Fig. 1B). The "featureless" $1/n$ divergence of the distribution at small clone sizes is simply a manifestation of neutral dynamics that results in the largest fraction of surviving clones at any instant being ones that were "induced" in the recent past (cf. Fig. 1D).

If rates of somatic mutation are sufficiently high, SNPs may arise independently at the same locus in different cells. As estimates of mutant clone size using deep sequencing are based on measurements of the variable allele fraction (VAF), the multiplicity of induction events cannot be resolved. Fortunately, would such clone "merger" events occur, they would be signaled by a breakdown of the leading $1/n$ dependence allowing their existence to be inferred indirectly (SI Text). However, while $\omega N/r\lambda \ll 1$, where N denotes the number of progenitors in a given biopsy, and ω is the mutation rate associated with the given locus, the frequency of mutant clones derived from multiple induction events can be safely neglected (SI Text). To proceed, we will assume that this condition is met and look for consistency of the data with theory.

Although Eq. (1) provides a strong prediction with which to address deep sequencing data, the nonlinear dependence of $P_n(t)$ on clone size, n, makes comparison between experiment and theory cumbersome. Fortunately, a further straightforward manipulation of the size distribution provides a more convenient representation. Specifically, defining the average mutant clone size, $\langle n(t) \rangle \equiv \sum_{n=1}^{\infty} nP_n(t) = r\lambda t/\ln(r\lambda t)$, it follows that the "first incomplete moment" [37],

$$\mu_1(n,t) = \frac{1}{\langle n(t) \rangle} \sum_{m=n}^{\infty} mP_m(t) \approx e^{-n/r\lambda t} \qquad [2]$$

acquires a simple exponential dependence on clone size, n, with a decay constant $r\lambda t$, equivalent to the average size of a surviving clone induced at birth, i.e. at the time of the first exposure to mutation (Figs. 1E-G and SI Text). Moreover, by the nature of its definition, the first incomplete moment is conveniently insensitive to the smallest clones, where the resolution of the deep sequencing approach is likely to be compromised. In the context of epidermis, it therefore follows that the corresponding distribution of clone areas, A, is given by $\mu_1(A,t) = e^{-\rho A/r\lambda t}$, with ρ the areal progenitor density.

Eq. (2) provides an objective, parameter-free prediction with which population asymmetry and neutrality of clone dynamics can be assessed. For a given array of biopsies, mutant clone sizes can be inferred from the corresponding VAFs associated with individual SNPs. Then the first incomplete moment, $\mu_1(A,t)$, can be constructed directly from the data. Departure of the inferred distribution from the predicted exponential size dependence would indicate functional heterogeneity of mutant clones and evidence of non-neutral dynamics. Convergence onto exponential would indicate that clone

dynamics is likely governed by the neutral competition of an equipotent progenitor pool. A fit to the exponent, $r\lambda t/\rho$, then provides access to the progenitor loss/replacement rate. Crucially, the exponential size dependence of $\mu_1(A,t)$ is sensitive only to the dynamics of the self-renewing (i.e. the active stem cell) population. As long as the dominant contribution to the measured mutant clone size distribution derives from clones associated with mutations that occurred on timescales in excess of the transit time through any differentiation hierarchy, the exponential size dependence would be conserved.

Neutral competition between keratinocyte progenitors. In the study of Martincorena *et al.* [16], eyelid epidermis was derived from more than 200 biopsies of sizes ranging from 0.79 to 4.71 mm^2 harvested from 4 patients aged 55 to 73 years of age. In each case, coding exons were sequenced across 74 genes implicated in skin and other cancers to an average effective coverage of 500× (SI Data). (For technical details on the sample preparation and sequencing approach, we refer to Ref. [16].) As detailed in the study, since few mutations involve change in copy number, the areal contribution of individual mutant clones can be inferred as twice the product of the VAF with the area of the biopsy. (Events involving the mutation of both alleles at a given locus are considered to occur at a negligible frequency.) For clones of a size much smaller than that of the biopsy, intersection of the clone with the boundary is statistically improbable. For larger clone sizes, the estimated clonal area in a given biopsy may represent only a fraction of the true size. However, the exponential character of the predicted first complete moment is robust to such statistical fluctuations as well as errors in the accuracy of the sequencing approach (SI Text).

To gain insight into the relative abundance of clones that "spill" outside individual biopsies, since the mutation rate is low [16], we can explore the coincidence of common point mutations found in different biopsies. Taking as an exemplar patient PD18003, for which the largest volume of data was obtained, we find that, from 1557 specific point mutations across 92 biopsies, some 102 (6.6%) are present in more than one biopsy. Of these, 90 point mutations are restricted to two biopsies, 10 span 3, 1 spans 4, and 1 spans 5. Similar frequencies of clone dispersion are found for the three other patients (Table 1), with one clone spanning no less than 12 biopsies.

Since the total area of clones that occupy multiple biopsies cannot be reliably recovered, we first focused on the ensemble of clones that bear a point mutation contained within a single biopsy. Further, to assess the utility of the approach, we began by focusing on the subset of these clones that involve only synonymous mutation (Figs. 2A and S3). As such mutations leave the associated protein sequence unchanged, it is expected that the dynamics of the corresponding clones remains neutral, providing a useful control to benchmark theory. Focusing on patient PD18003, for which there were a total of 257 synonymous point mutations restricted to a single biopsy, analysis of the first incomplete moment, $\mu_1(A,t)$, reveals a remarkably exponential size dependence (Fig. 2B), consistent with neutral competition of the constituent progenitors. As well as justifying the validity of the approach, this result establishes that, under conditions of normal homeostasis, the progenitors that maintain adult human IFE conform long-term to population asymmetric self-renewal.

With the size distribution of clones associated with synonymous mutations defined, we then considered the wider class of mutant clones including both synonymous and non-synonymous (missense and nonsense) mutations. Once again, taking the 1338 clones associated with a single point mutation, the size distribution, $\mu_1(A,t)$, shows only a small departure from exponential with the divergence impacting at the largest clone sizes (Fig. 2C, arrow head). By fitting the data to the exponential clone size dependence (red curve), we can then use the predicted cumulative frequency to estimate the point of departure of the statistical distribution. Given the size of the ensemble of clones, we find that the observation of the 7 largest clones with a size in excess of 1.08 mm^2, a significant fraction of the size of the associated biopsies, would be statistically improbable within the framework of neutral dynamics, i.e. these clones would be predicted to occur with a frequency much less than 1 in 1000. Furthermore, inspection of the mutational profile of the 6 biopsies containing the 7 clones (Fig. S4) shows that 5 biopsies are associated with different point mutations (missense or frameshift deletion) in NOTCH1, while the sixth involves a missense mutation in MLL2. For the latter, 3 other mutations appear with a very similar VAF to MLL2 (Fig. 2D), suggesting that all four mutations belong to the same clone. Significantly, when these 6 biopsies are filtered out of the statistical cohort of 92, the first incomplete moment collapses onto a strikingly exponential size dependence (Figs. 2E,F). The coincidence of theory and experiment is further emphasized by comparison of the clone size distribution, $P(A,t)$, with the predicted size dependence (Fig. 2G).

Turning to patient PD13634, of the 725 discrete point mutations, 657 (89%) belong to a single biopsy with 159 of these associated with synonymous point mutations. Once again, their size distribution shows collapse onto an exponential dependence, consistent with neutral dynamics (Fig. S5A). Then, when combined with non-synonymous mutations, the size distribution of all 657 mutant clones continues to collapse onto exponential with no apparent outliers by size (Figs. S2 and S5B). For patient PD20399, of the 803 discrete point mutations, 724 (90%) belong to a single biopsy. In this case, the size distribution of all 724 point mutations as well as the 154 mutant clones that bear a synonymous mutation also collapse onto exponential with no outliers by size (Figs. S2, S5C and S5D). Finally, for patient PD21910, although the data is relatively sparse, of the 195 discrete mutations, 181 (93%) belong to a single biopsy. With just 35 of these mutant clones bearing a synonymous point mutation, the size distribution is noisy but consistent with exponential (Fig. S5E). Again, as expected, comparison of all 181 mutations also reveals a collapse onto exponential with no outliers (Figs. S2 and S5F).

Non-neutral expansion of rare mutant clones. Although these results suggest that the vast majority of point mutations leave neutral dynamics unperturbed, the statistical method also provides a quantitative scheme to identify mutant clones that lie outside the normal (exponential) size distribution. Our results suggest that very few are associated with non-neutral dynamics - in one patient (PD18003), just 6 outlier clones were identified by size, while none were found in the other patients (Figs. 2 and S5). However, so far, we have excluded from our analysis mutant clones that span multiple biopsies. Since these clones are likely to be large, one might expect that they harbor the majority of cells that have undergone non-neutral transformation. Therefore, to gain insight into the nature of these dispersive clones, we explored the mutational profile of clones that spanned more than 3 biopsies.

Starting with patient PD18003, only one mutant clone bearing a missense mutation in SCN1A (C927S) spans more than 3 biopsies (Table 1), having an aggregate size of 2.52 mm^2, more than a factor of two larger than the cut-off used to filter single-biopsy clones. Whether this outlier represents the chance expansion of a clone governed by neutral dynamics, or derives from the proliferative advantage of mutant cells

over their wildtype neighbors – the process of "field cancerization" [36] – driven by mutation of SCN1A is impossible to determine unambiguously. However, noting that the vast majority of the clone is limited to just one biopsy in which the point mutation in SCN1A is expressed with a VAF similar to that of point mutations associated with 3 other genes (Figs. 3A,B), it seems likely that expansion of this clone is driven by the chance acquisition of multiple point mutations. Indeed, by comparing the relative values of the VAFs, we can infer the likely order in which these point mutations were acquired (Fig. 3C).

Similarly, in patient PD20399, a clone bearing a missense point mutation in FGFR3 (R248C) spans no less than 12 biopsies covering an aggregate area of $7.41\,\mathrm{mm}^2$. However, as noted by Martincorena *et al.* [16], in six of the 12 biopsies, this mutation appears alongside two other missense point mutations, one in TP53 (P250L) and one in ARID1A (P929S), with all three bearing a very similar VAF (Fig. 3D). This coincidence suggests that it is the acquisition of these secondary mutations that drives non-neutral expansion of the clone. From the relative sizes of the three constituent mutations, we can infer the likely sequence of their acquisition (Fig. 3E). Interestingly, the large dispersion of the clone bearing the original mutation in FGFR3 and its irregular spatial pattern (Fig. 3D) suggests that, either mutation in FGFR3 occurred independently at the same locus, or this mutation may have a developmental origin. A second clone mutant for NOTCH2 (P426S) spans 5 biopsies, but the majority lies within just two. In this case, its net aggregate size of $0.75\,\mathrm{mm}^2$ suggests that it may belong to the ensemble of neutral mutations.

For patient PD13634, inspection of the mutational profile shows that one point mutation spans 7 biopsies, 3 span 5, and 4 span 4. Inspection of the mutational profile shows these events can be traced to the expansion of just two clones and their subclones. Comparison of the VAFs of the constituent mutant clones suggests that, in one case, a consecutive sequence of 5 independent point mutations starting with FGFR3 (*809G), followed by PPP1R3A (P967L), ARID1A (G851D), NOTCH1 (P574S) and NOTCH1 (P745), drives non-neutral expansion leading to a clone with an aggregate size in excess of $8\,\mathrm{mm}^2$ (Fig. 3F). A second independent clone, involving a synonymous mutation in MUC17 (T3292T) followed by a nonsense mutation in SPHKAP (W308*), leads to a much smaller clone with an aggregate size of only $0.77\,\mathrm{mm}^2$, well within the statistical ensemble of neutral mutations. Finally, for patient PD21910, there are no mutations that extend beyond two biopsies.

For consistency, we can further filter the ensemble of biopsies excluding those that contain the two oversized clones in patients PD13634 and PD20399. Since these clones are subject to non-neutral expansion, they may impact upon neighbors by either suppressing their expansion, or conveying them as passengers. Once removed, we find that the first incomplete moment maintains its exponential character, while the total clone size distribution falls onto the predicted size dependence (Fig. S6). Finally, to further challenge the hypothesis of neutrality, we determined the average mutant clone size across a range of genes. For all 4 patients, we found that departures of the average clone size associated with specific cancer drivers from that of the ensemble were not statistically significant (Fig. S7).

Although these findings suggest that the majority of mutations leave neutral dynamics unperturbed, it is important to consider what would emerge if the dynamics were non-neutral. If all point mutations conferred the *same* proliferative advantage, the first incomplete moment would also acquire an exponential size dependence, $\mu_1(n,t) \approx e^{-n/N(t)}$, with $N(t) = e^{\nu t}$ and ν defining the net proliferative expansion rate of mutant progenitors [37]. However, since such a size dependence would require all point mutations (synonymous and non-synonymous) to confer precisely the same proliferative advantage, its relevance to the current study is unlikely. It is, however, important to note that, while the statistical approach provides the means to define clones that lie outside the normal size distribution, we cannot rule out the existence of a further subfraction of clones associated with non-neutral transformation that lie hidden within the bulk of the neutral distribution.

Discussion. These results demonstrate how analysis of deep sequencing data provides a general framework to study stem cell self-renewal of normal cycling adult human tissues. Applied to human IFE, we find that maintenance involves the turnover of a progenitor population following population asymmetry in which their stochastic loss through differentiation is compensated by duplication of neighbors. From a fit of the data to the exponential size dependence of $\mu_1(A,t)$, the inferred ratios $r\lambda t/\rho$ are found to broadly consistent with the predicted linear increase with the age of the patient (Table 2). With an estimated basal cell density of $\rho = 10,000$ cells per mm^2 [38], and a progenitor fraction of basal cells of 1 in 3 (extrapolated from mouse [27]), a linear fit of the measured ratio suggests a loss/replacement rate of the self-renewing population of $r\lambda \approx 0.5$ per week. Although uncertainty in both the progenitor fraction and the relative frequency of divisions leading to symmetric or asymmetric fate undermine the predictive value of this rate, a loss/replacement time measured in weeks is broadly consistent with the expected proliferative activity of cycling keratinocyte progenitors in normal homeostasis which, on the basis of BrdU incorporation, points at an average cell division rate in human scalp epidermis of around 2 per week [38].

Significantly, the deep sequencing approach also provides a quantitative assay to expose rare mutant clones that have undergone field transformation and to assess their mutational profile. Application of this approach to human epidermis shows that, despite evidence for positive selection [16], population asymmetry and neutrality of epidermal progenitor cell fate may be surprisingly robust to the acquisition of somatic point mutations, even in genes associated with cancer drivers. Indeed, the multiplicity of mutations in the minority of clones (ca. 0.1% or less) that lie outside the normal size distribution suggests that proliferative advantage may typically rely on epistasis, requiring the acquisition of multiple mutations across a range of genes.

Under conditions of normal homeostasis, clonal evolution in IFE is constrained to two dimensions, with clones expanding in cohesive clusters across the basal and suprabasal layers (Fig. 1C). Applied to higher dimensional (volumnar) tissues, as well as other epithelial tissues, the same clone size dependence is predicted to apply without further revision (SI Text) [35]. However, if occupancy of the self-renewing compartment is constrained to lower dimension, or if stem cells are restricted to closed niche domains, the same general technology applies, but the predicted mutant clone size distribution must be appropriately revised (SI Text). Therefore, applied to deep sequencing studies, the current theoretical scheme provides a general method to probe stem cell fate behavior in normal cycling adult human tissues, and to identify the existence and mutational signature of rare field transformations driven by the non-neutral dynamics of mutant cells, with potential applications to the early detection of preneoplasia.

ACKNOWLEDGMENTS. We are indebted to Peter Campbell, Phil Jones and Inigo Martincorena for sharing information on the sizes of the biopsies used in their study, and for making their sequencing data publically available. We are also grateful to Trevor Graham, Philip Greulich and Anna Philpott for valuable discussions, and we acknowledge the financial support of the Wellcome Trust (grant number 098357/Z/12/Z).

1. Simons BD, Clevers H (2011) Strategies for homeostatic stem cell self-renewal in adult tissues. Cell, 145(6):851-62.
2. Van Keymeulen A, Blanpain C (2012) Tracing epithelial stem cells during development, homeostasis, and repair. The Journal of Cell Biology, 197(5):575-584.
3. Driessens G, Beck B, Caauwe A, Simons BD, Blanpain C (2012) Defining the mode of tumour growth by clonal analysis. Nature, 488(7412):527-30.
4. Blanpain C (2013) Tracing the cellular origin of cancer. Nat Cell Biol, 15(2):126-134.
5. Vermeulen L, et al. (2013) Defining stem cell dynamics in models of intestinal tumor initiation. Science, 342(6161):995-998.
6. Ellenbroek SIJ, van Rheenen J (2014) Imaging hallmarks of cancer in living mice. Nat Rev Cancer, 14(6):406-418.
7. Merlo LMF, Pepper JW, Reid BJ, Maley CC (2006) Cancer as an evolutionary and ecological process. Nat Rev Cancer, 6(12):924-935.
8. Bozic I, et al. (2010) Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci USA, 107(43):18545-18550.
9. Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature, 472(7341):90-94.
10. Gerlinger M, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine, 366(10):883-892.
11. Greaves M., Maley CC (2012) Clonal evolution in cancer. Nature, 481(7381):306-313.
12. Vogelstein B, et al. (2013) Cancer genome landscapes. Science, 339(6127):1546-1558.
13. Sottoriva A, et al. (2015) A Big Bang model of human colorectal tumor growth. Nat Genet, 47(3):209-216.
14. Martincorena I, Campbell PJ (2015) Somatic mutation in cancer and normal cells. Science, 349(6255):1483-1489.
15. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature, 458(7239):719-724.
16. Martincorena I, et al. (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. Science, 348(6237):880-887.
17. Teixeira VH, et al. (2013) Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors. eLife, 2:e00966.
18. Baker A-M, et al. (2014) Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. Cell Reports, 8:1-8.
19. Martincorena I, Seshasayee ASN, Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature, 485(7396):95-98.
20. Dees ND, et al. (2012) MuSiC: Identifying mutational significance in cancer genomes. Genome Research, 22(8):1589-1598.
21. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics, 173(4):2187-2198.
22. Krieger T, Simons BD (2015) Dynamic stem cell heterogeneity. Development, 142(8):1396-1406.
23. Hsu Y-C, Li L, Fuchs E (2014) Emerging interactions between skin stem cells and their niches. Nat Med, 20(8):847-856.
24. Page ME, Lombard P, Ng F, Göttgens B, Jensen KB (2015) The epidermis comprises autonomous compartments maintained by distinct stem cell populations. Cell Stem Cell, 13(4):471-482.
25. I C Mackenzie (1970) Relationship between mitosis and the ordered structure of the stratum corneum in mouse epidermis. Nature, 226(5246):653–655, 1970.
26. Potten CS, Kovacs L, Hamilton E (1974) Continuous labelling studies on mouse skin and intestine. Cell Proliferation, 7(3):271-283.
27. Clayton E, et al. (2007) A single type of progenitor cell maintains normal epidermis. Nature, 446(7132):185–9, mar 2007.
28. Doupe DP, Klein AM, Simons BD, Jones PH (2010) The ordered architecture of murine ear epidermis is maintained by progenitor cells with random fate. Developmental Cell, 18(2):317-323.
29. Mascre G, et al. (2012) Distinct contribution of stem and progenitor cells to epidermal maintenance. Nature, 489:257-262.
30. Lim X, et al. (2013) Interfollicular epidermal stem cells self-renew via autocrine Wnt signaling. Science, 342:1226-1230.
31. Rheinwatd JG, Green H (1975) Seria cultivation of strains of human epidemal keratinocytes: the formation keratinizin colonies from single cells. Cell, 6(3):331-343.
32. Barrandon Y, Green H (1987) Three clonal types of keratinocyte with different capacities for multiplication. Proc Natl Acad Sci USA, 84(8):2302-2306.
33. Jones PH, Watt FM (1993) Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression. Cell, 73(4):713-724.
34. Watt FM (2014) Mammalian skin cell biology: At the interface between laboratory and clinic. Science, 346:937-940.
35. Klein AM, Simons BD (2011) Universal patterns of stem cell fate in cycling adult tissues. Development, 138:3103-3111.
36. Slaughter DP, Southwick HW, Smejkal W (1953) "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. Cancer, 6(5):963-968.
37. Klein AM, Brash DE, Jones PH, Simons BD (2010) Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia Proc Natl Acad Sci USA 107(1):270-275.
38. Jones PH, Harper S, Watt FM (1995) Stem cell patterning and fate in human epidermis. Cell, 80(1):83-93.

Fig. 1. Dynamics of clones in mammalian IFE. (A) Schematic depicting the cellular organization of human IFE. (B) In the paradigm of population asymmetry, IFE is maintained by basal progenitors that, following division, choose stochastically between symmetric and asymmetric fate with the probability of symmetric proliferation, $0 < r \leq 1/2$, perfectly balanced by terminal division. (C) When lineage labeled, the clonal progeny of marked progenitors may expand through the replacement of neighbors (yellow clone), or may become lost through differentiation (green clone). (D) Following the chance acquisition of somatic mutation, clones (with colors denoting different SNPs) compete for survival. Over time, some clones become lost through differentiation while others expand. Clones that bear one mutation may acquire further mutations. (E) According to the model depicted in B, the average size of mutant clones (orange line) is predicted to increase as $\omega N t$, where ω denotes the mutation rate per progenitor at a given locus, N is the size of the progenitor pool, and t is the age of the patient. The average mutant clone size is predicted to increase as $r\lambda t/\ln(r\lambda t)$ (black line). (F) The corresponding mutant clone size distribution, $P_n(t)$, and (G) first incomplete moment, $\mu_1(n,t)$ predicted by Eqs. (1) and (2) respectively (black lines). In E-G points show the results of stochastic simulation of the birth-death process B where, in F and G, $\lambda t = 2^6$ (orange), 2^8 (brown), 2^{10} (grey), and 2^{12} (black) with $\omega N/\lambda = 1$ and $r = 1/2$.

Fig. 2. Size distribution of mutant clones provides evidence of neutral dynamics. (A) Mutant clone sizes inferred from the analysis of the variable allele fractions of all (N=1557) point mutations for patient PD18003 (black points). Synonymous mutations are marked as red, and mutations belonging to the 6 biopsies considered outliers by size are shown as brown. (B) First incomplete moment, $\mu_1(A,t)$, associated with N=257 synonymous mutations that belong to single biopsies (points) shows collapse onto the predicted exponential size dependence (Eq. (2), red line and Table 2). (C) First incomplete moment associated with all (N=1338) point mutations that belong to single biopsies. The departure of $\mu_1(A,t)$ (arrow head) from exponential (red line) reflects contributions from 7 mutant clones contained within 6 biopsies. (D) Clone sizes of one biopsy showing one of the outliers associated with mutation in MLL2 (**). Note that three other point mutations have clone sizes comparable to this outlier suggesting that they belong to the same subclone. Non-point mutations are indicated as grey. When the 6 biopsies are excluded from the statistical ensemble, the distribution $\mu_1(A,t)$ (points) shown in (E,F), collapses onto the predicted exponential form (red line and Table 2). (G) The corresponding clone size distribution, $P(A,t)$, with points (black) showing data and the line (red) showing the theoretical prediction of Eq. (1). The departure of theory from experiment at clone sizes below of $0.05\,\mathrm{mm}^2$ and below indicates the resolution limit of the sequencing approach. With a basal cell density of 10,000 cells per mm^2, this suggests that a resolution limit of around 500 basal cells. Error bars denote s.e.m.

Fig. 3. Field transformation and non-neutral clone dynamics driven by multiple mutations. (A) Coordinates of biopsies of eyelid epidermis sampled for patient PD18003 with the size of points scaled by biopsy area. Colored points contained within single biopsies depict the 6 biopsies containing the 7 mutant clones considered outliers by size. Green points mark 5 biopsies sharing the same point mutation in SCN1A. Size of colored dots scaled by the size of the outlier clones or the shared mutant clone. (B) Size of mutant clones belonging to one of the 5 biopsies with SCN1A mutation. Comparison of VAFs across all 5 biopsies suggest the clonal structure depicted in (C). (D) Coordinates of biopsies from patient PD20399 with the size of dots scaled by the area of the biopsy. The 12 biopsies colored light or dark green share the same point mutation in FGFR3 while the 6 biopsies marked in dark green also share point mutations in TP53 and ARID1A. Green biopsies scaled by the size of the majority mutant clone. (E) Clonal organization implied by the mutational landscape of the clone depicted in D. (F) Analysis of mutational profile of mutant clones that span more than 3 biopsies in patient PD13634 reveal an oversized clone with the given clonal architecture.

Table 1. Frequency of point mutations shared by multiple biopsies.

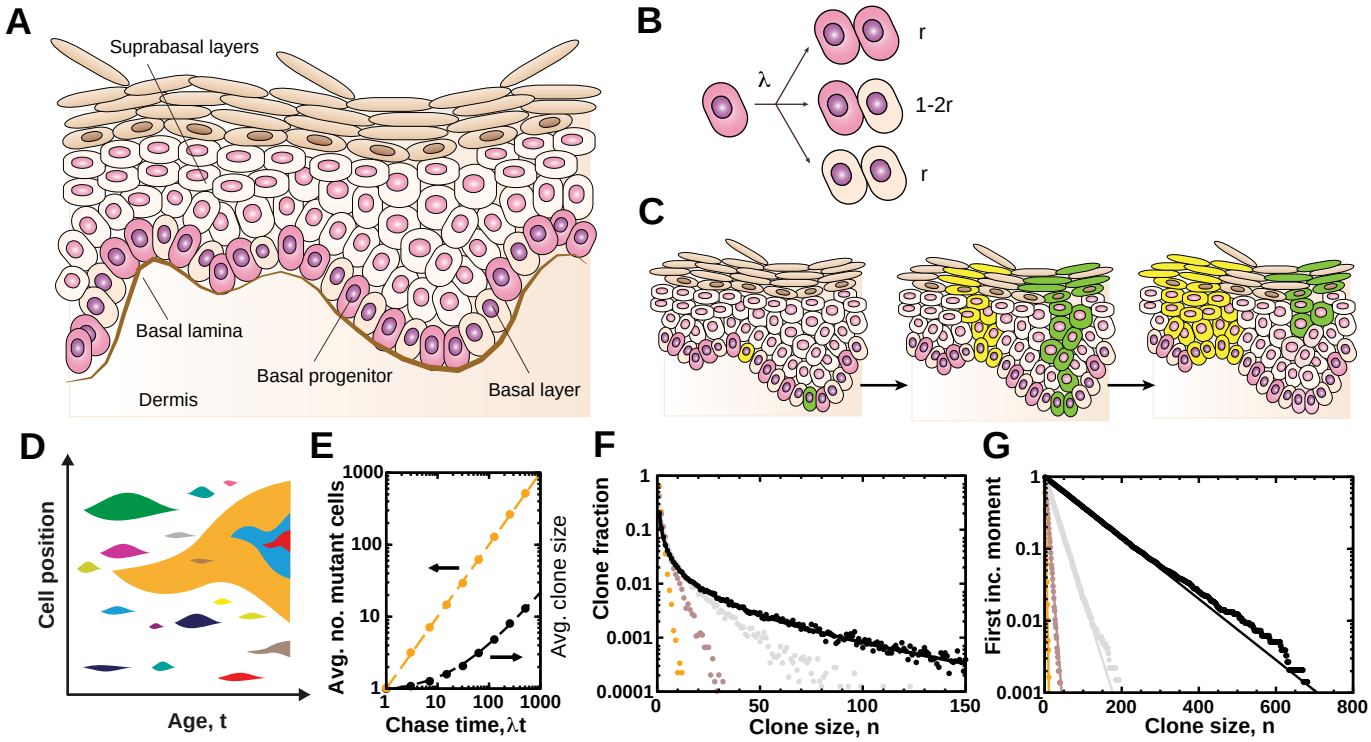
Patient ID	Number of biopsies*							
	1	2	3	4	5	6	7	12
PD13634	647	63	7	4	3		1	
PD18003	1338	90	10	1	1			
PD20399	724	64	10	1		3		1
PD21910	181	14						

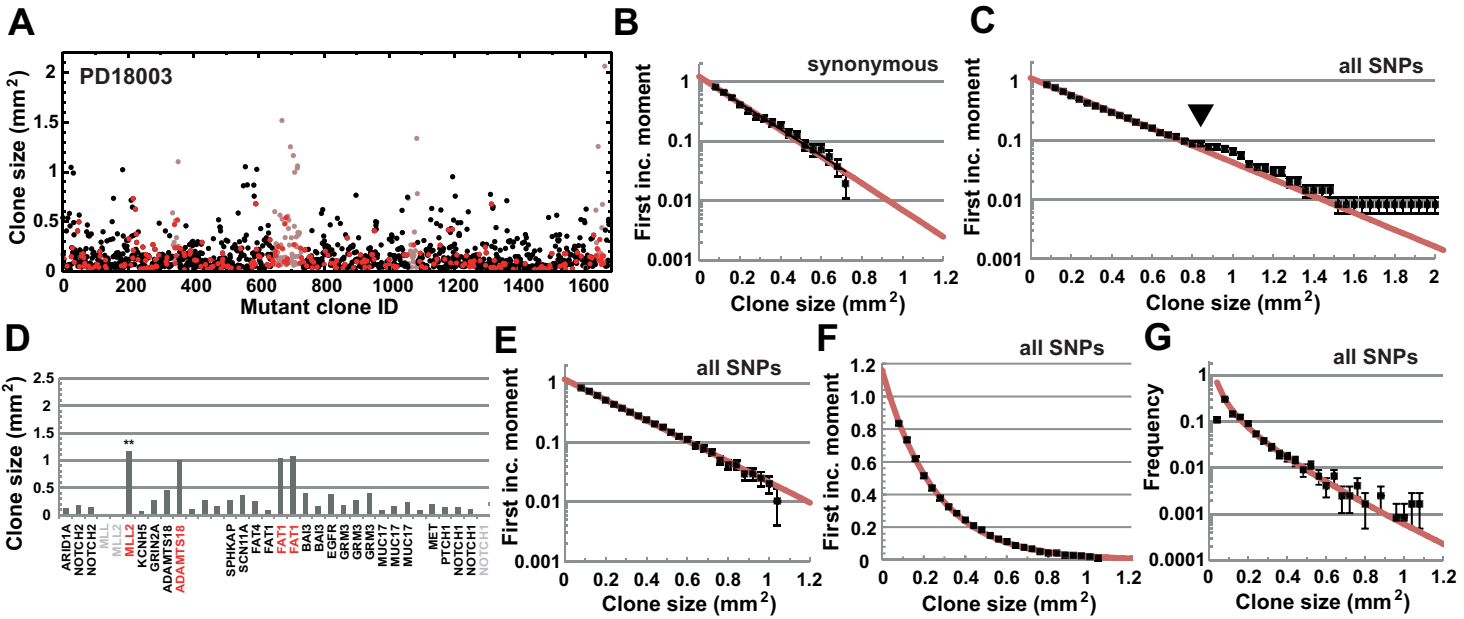
*Multiplicity of point mutations that span multiple biopsies. For example, in patient PD13634, 647 point mutations are found in only one biopsy, 63 and found in 2 biopsies, etc.

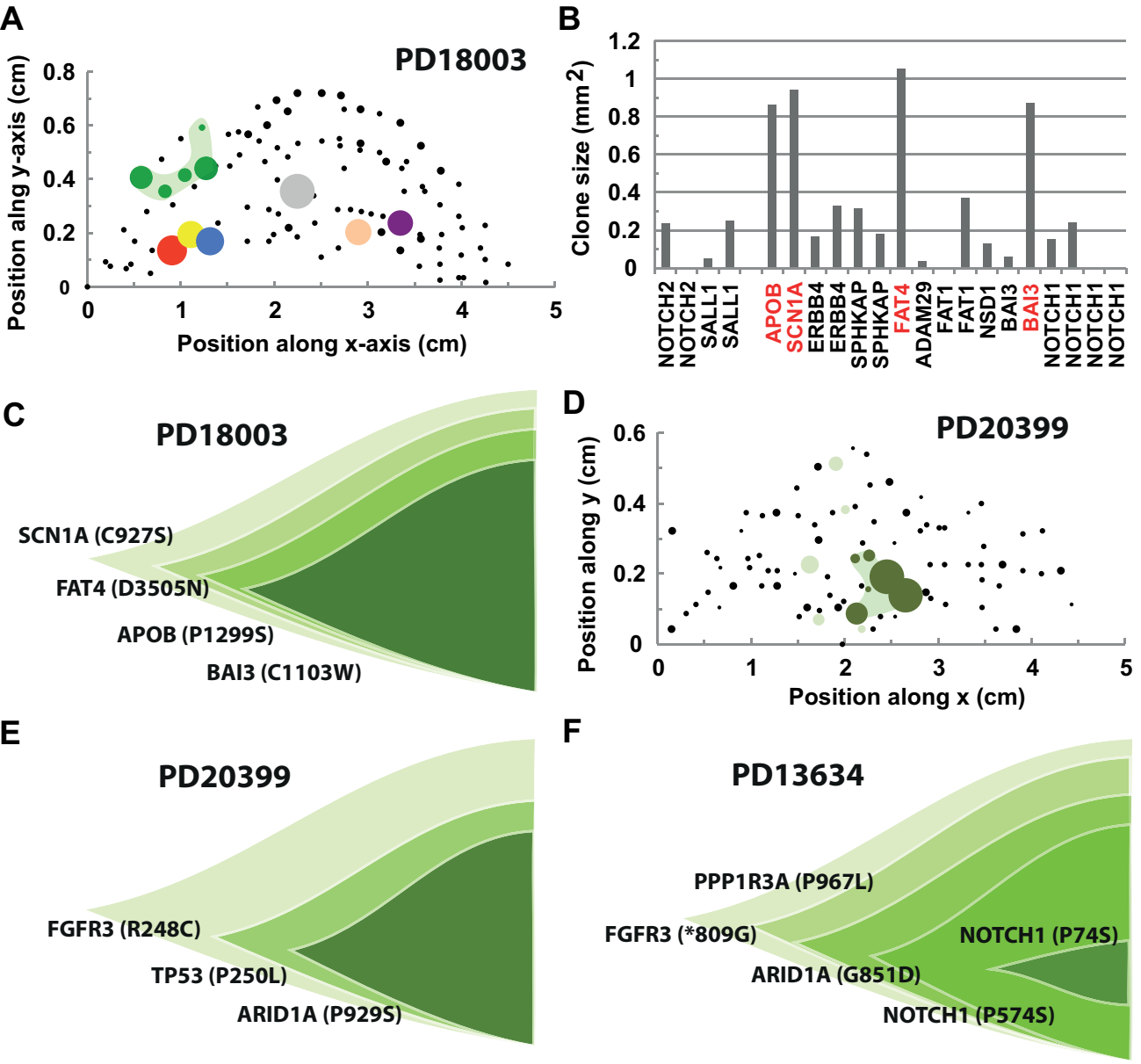
Table 2. Fit of the first incomplete moment to the predicted exponential size dependence.

	Patient*			
	PD18003	PD13634	PD20399	PD21910
Age/years (sex)	65 (F)	73 (M)	55 (F)	58 (F)
$r\lambda t/\rho$ ($\mathrm{mm}^2{}_{R^2}$) (syn.)	$0.20_{0.98}$	$0.22_{0.98}$	$0.12_{0.96}$	$0.31_{0.87}$
$r\lambda t/\rho$ ($\mathrm{mm}^2{}_{R^2}$) (all)	$0.25_{0.99}$	$0.24_{0.94}$	$0.14_{0.98}$	$0.32_{0.96}$
$r\lambda/\rho$ ($\mathrm{mm}^2\,\mathrm{yr}^{-1}$)	0.0039	0.0032	0.0025	0.0054

*Ratios $r\lambda t/\rho$ inferred from fits to $\mu_1(n,t)$ (red lines in Figs. 2 and S5E and S6)


A
Suprabasal layers
Basal lamina
Basal progenitor
Basal layer
Dermis
B
λ
r
1-2r
r
C
D
Cell position
Age, t
E
Avg. no. mutant cells
Avg. clone size
Chase time,λt
1000
100
10
1
1
10
100
1000
F
Clone fraction
1
0.1
0.01
0.001
0.0001
0
50
100
150
Clone size, n
G
First inc. moment
1
0.1
0.01
0.001
0
200
400
600
800
Clone size, n


A
PD18003
Clone size (mm2)
Mutant clone ID
B
synonymous
First inc. moment
Clone size (mm2)
C
all SNPs
First inc. moment
Clone size (mm2)
D
Clone size (mm2)
**
ARID1A
NOTCH2
NOTCH2
MLL
MLL2
MLL2
KCNH5
GRIN2A
ADAMTS8
ADAMTS18
SPHKAP
SCN11A
FAT4
FAT1
FAT1
FAT1
BAI3
BAI3
EGFR
GRM3
GRM3
GRM3
MUC17
MUC17
MUC17
MET
PTCH1
NOTCH1
NOTCH1
NOTCH1
E
all SNPs
First inc. moment
Clone size (mm2)
F
all SNPs
First inc. moment
Clone size (mm2)
G
all SNPs
Frequency
Clone size (mm2)


A
PD18003
Position alng y-axis (cm)
Position along x-axis (cm)
B
Clone size (mm2)
NOTCH2
NOTCH2
SALL1
SALL1
APOB
SCN1A
ERBB4
ERBB4
SPHKAP
SPHKAP
FAT4
ADAM29
FAT1
FAT1
NSD1
BAI3
BAI3
NOTCH1
NOTCH1
NOTCH1
NOTCH1
C
PD18003
SCN1A (C927S)
FAT4 (D3505N)
APOB (P1299S)
BAI3 (C1103W)
D
PD20399
Position along y (cm)
Position along x (cm)
E
PD20399
FGFR3 (R248C)
TP53 (P250L)
ARID1A (P929S)
F
PD13634
PPP1R3A (P967L)
FGFR3 (*809G)
ARID1A (G851D)
NOTCH1 (P74S)
NOTCH1 (P574S)

Supplementary Information

The quantitative analysis of the probability distribution of mutant clone sizes described in the main text relies upon a robust and generic model of stem cell self-renewal in adult cycling tissues. In the following, we expand upon the theoretical basis of the modeling scheme, and the derivation of the size distribution of mutant clones defined by Eq. (1) in the main text. Furthermore, we discuss generalizations of the theoretical scheme to different tissue architectures.

Background: As a starting point, it is necessary to review the constraints that restrict the possible fate behavior of progenitors that maintain cycling adult tissues. For clarity, we refer to this cycling and self-renewing population as *progenitors* rather than stem cells noting that, in the context of epidermis as well as other tissues, these cells may be underpinned by a second quiescent "stem cell" population. To achieve long-term homeostasis, the maintenance of a cycling tissue must ultimately rely on the turnover of a single equipotent progenitor pool[1] that divides asymmetrically so that, on average, one daughter cell stays in the self-renewing compartment while the other commits to differentiation, either directly or through a transit compartment with a strictly limited proliferative potential [35].

Although fate asymmetry may be invariant, enforced at the level of individual cell divisions, it may also follow from a stochastic pattern of behavior in which chance progenitor cell loss through differentiation is perfectly compensated by the duplication of neighboring cells (cf. Fig. 1B). In this mode of "population asymmetric" self-renewal, neutral competition between progenitors leads to a gradual consolidation of clonal diversity, while the size of surviving clones continually increases (cf. Figs. 1C and S2).

On this background, consider the impact of a neutral hereditary label (such as a synony-

[1] Note that the equipotency of progenitor cell behavior over long times does not rule out potential short-term heterogeneity in which progenitors transfer reversibly between states biased for duplication or differentiation [22]

mous point mutation) that marks a progenitor cell. In systems characterized by invariant asymmetric self-renewal, individual progenitor cells are long-lived and, once acquired, a marked cell would persist indefinitely. By contrast, in systems defined by population asymmetric self-renewal, through neutral competition between neighboring progenitors, a marked cell and its progeny - a clone - may, by chance, be purged from the population, or the clone may expand (cf. Fig. 1C). In a tissue defined by an ensemble of closed niche domains, such as the crypt organization of the intestinal epithelium, this process of "neutral drift" and clonal consolidation continues until the clone is lost, or cells in the given niche domain drift to monoclonality and the hereditary label (mutation) becomes locally fixed. By contrast, in systems defined by an open or facultative niche, such as the interfollicular epidermis or testis, the neutral dynamics of clones would continue unabated.

To address the quantitative dynamics of clones in an open or "facultative" niche subject to population asymmetric self-renewal, it is important to discriminate between different patterns of regulation [35]. If the balance between proliferation and differentiation follows from intrinsic (cell autonomous) regulation, long-term clone dynamics of the progenitor pool becomes indistinguishable from that of the critical birth-death process,

$$P \xrightarrow{2r\lambda} \begin{cases} P+P & \text{Pr. } 1/2 \\ \emptyset & \text{Pr. } 1/2 \end{cases} .$$

where $r\lambda$ denotes the effective loss/replacement rate of progenitors P. Here, we have chosen our definition of the loss/replacement rate to align with the particular model of progenitor cell self-renewal in IFE, as depicted in Fig. 1B. Furthermore, we have suppressed progenitor cell divisions that lead to asymmetric fate outcome as these leave the progenitor cell number - and therefore the progenitor clone size - unchanged.

Then, if we start with a progenitor population of total size $N \gg r\lambda t$, the chance of finding a surviving clone of size n progenitors after a time t is given by [39]

$$p_n(t) = \left(1 + \frac{1}{r\lambda t}\right)^{-(n+1)} \times \begin{cases} 1 & n = 0 \\ \dfrac{1}{(r\lambda t)^2} & n > 0 \end{cases} .$$

In particular, at times $r\lambda t \gg 1$, the distribution of "surviving" clones, i.e. those containing at least one progenitor, converges onto a hallmark exponential clone size distribution, $p_n(t)/(1 - p_0(t)) = (1/r\lambda t)\exp[-n/r\lambda t]$. From this result, it follows that the cumulative distribution, defined as the chance of finding a surviving clone with a size greater than n

progenitors, is given by $\exp[-n/\langle n(t)\rangle]$, where $\langle n(t)\rangle \simeq r\lambda t$ defines the average size of surviving clones, i.e. while the average size of surviving clones increases, the chance of finding a clone with a size given by some mulitple of the average remains constant over time, and defined by an exponential distribution. The speed with which the size distribution convergences to this "scaling" limit is illustrated by the results of a stochastic simulation shown in Fig. S2.

Conversely, if the balance between proliferation and differentiation follows from extrinsic regulation (such as neutral competition for limited niche access), the long-term clonal dynamics depends on the local coordination of neighboring progenitor cells, and therefore the effective "dimensionality" of the niche [35]. In dimensions of two (epithelial) and above (volumnar), the clone size distribution converges onto the same exponential size dependence as that defined above for intrinsic regulation. However, in quasi one-dimensional (ductal) tissues, for $\lambda t \gg 1$, the clone size distribution converges onto the form

$$p_n(t) \simeq \begin{cases} 1 - \dfrac{1}{\sqrt{\pi\lambda t}} & n = 0 \\ \dfrac{1}{\sqrt{\pi\lambda t}} \dfrac{n}{2\lambda t} \exp\left[-\dfrac{n^2}{4\lambda t}\right] & n > 0 \end{cases},$$

where λ defines the loss/replacement rate of neighboring progenitors. In this case, the surviving clone size distribution takes the form $p_n(t)/(1 - p_0(t)) = (n/2\lambda t)\exp[-n^2/4\lambda t]$. Once again, the chance of finding a clone with a size larger than n progenitors is given by the scaling form, $\exp[-(\pi/4)(n/\langle n(t)\rangle)^2]$, where $\langle n(t)\rangle \simeq \sqrt{\pi\lambda t}$ defines the average size of surviving clones.

Mutant clone dynamics in interfollicular epidermis

With these preliminaries, let us now consider mutant clone dynamics following the acquisition of a neutral somatic point mutation in the IFE. In particular, let us suppose that cells belonging to the self-renewing progenitor pool acquire sporadic point mutations at a low rate of ω per base per progenitor cell, where ω may vary substantially with the specific locus along the genome. Later we will define what we mean by a "low mutation rate". Each point mutation then serves as a hereditary clonal marker identifying mutant cells and their progeny. In the course of turnover, through (neutral) competition with neighbors, the clonal progeny of these mutated cells may survive and expand or they may become extinct. Formally, in this case, the dynamics is equivalent to a critical birth-death process with immigration [39]. If each progenitor cell simply persisted without loss and replacement, i.e. $r\lambda = 0$, then for

$\omega t \ll 1$, where t denotes the time during which the population is exposed to mutations (i.e. the age of patient), the multiplicity of marked (mutated) cells, M, at any given locus would be given by a Poisson distribution,

$$Z_M(t) = \frac{N!}{(N-M)!M!}(\omega t)^M(1-\omega t)^{N-M} \simeq \frac{(\omega t N)^M}{M!}e^{-\omega t N}.$$

Over time, the field of cells without mutation at this locus must steadily decrease, leading to a small adjustment of this probability. However, providing the fraction of mutated cells at a given locus remains small, $M/N \ll 1$, the adjustment of $Z_M(t)$ can be safely neglected. In this case, the average number of cells with a mutation at the given locus is given by $\langle M \rangle = \omega N t$.

With this definition, we may now construct the size distribution of mutant cells at a given locus in the population. More precisely, the chance, $Q_n(t)$, of finding n mutant progenitor cells at age t is given by

$$Q_n(t) = \sum_{M=0}^{\infty} Z_M(t) \int_0^t \frac{dt_1}{t} \cdots \int_0^t \frac{dt_M}{t} \sum_{n_1,\cdots n_M=0}^{\infty} \delta_{n,n_1+\cdots n_M}\, p_{n_1}(t_1)\cdots p_{n_M}(t_M).$$

Formally, the first component describes the multiplicity of mutated cells. Each of these mutations could have occurred at any time between 0 and t. Each will produce a clone with n progenitor cells with probability $p_n(t)$ integrated over time t. Summing the total number of marked cells, n, for each induction event gives the total number of progenitor cells with a mutation at that locus in the genome.

Then, if we suppose that the balance between proliferation and differentiation follows from intrinsic regulation, using the results above, and setting $\delta_{n,n_1+\cdots n_M} = \int_0^{2\pi} \frac{d\phi}{2\pi} \exp[-i\phi(n_1 + \cdots n_M - n)]$, all time integrals and cell number summations can be performed. In particular, making use of the identity,

$$\frac{1}{\lambda t}\int_0^t \lambda dt' p_n(t') = \left(1 - \frac{1}{r\lambda t}\ln(1+r\lambda t)\right)\delta_{n,0} + \frac{1}{r\lambda t}\frac{1}{n}\left(1+\frac{1}{r\lambda t}\right)^{-n},$$

it follows that

$$\begin{aligned}\sum_{n=0}^{\infty} e^{-in\phi}\frac{1}{r\lambda t}\int_0^t \lambda dt' p_n(t') &= 1 - \frac{1}{r\lambda t}\ln(1+r\lambda t) + \frac{1}{r\lambda t}\ln\left[\frac{1+r\lambda t}{1+r\lambda t - r\lambda t e^{-i\phi}}\right] \\ &= 1 - \frac{1}{r\lambda t}\ln\left(1 + r\lambda t - r\lambda t e^{-i\phi}\right).\end{aligned}$$

Therefore, making use of this result, we have that

$$\sum_{M=0}^{N} Z_M(t) \left[1 - \frac{1}{r\lambda t} \ln\left(1 + r\lambda t - r\lambda t e^{-i\phi}\right)\right]^M = \left[1 - \frac{\omega}{r\lambda} \ln\left(1 + r\lambda t - r\lambda t e^{-i\phi}\right)\right]^N$$

from which we obtain the formal expression for the probability distribution function,

$$Q_n(t) = \int_0^{2\pi} \frac{d\phi}{2\pi} e^{in\phi} \left[1 - \frac{\omega}{r\lambda} \ln\left(1 + r\lambda t - r\lambda t e^{-i\phi}\right)\right]^N.$$

In the limit $r\lambda t \gg 1$, this expression can be simplified to the form,

$$Q_n(t) = \int_0^{2\pi} \frac{d\phi}{2\pi} e^{i\phi n} \left[1 + r\lambda t \left(1 - e^{-i\phi}\right)\right]^{-\zeta}, \tag{1}$$

where we have defined the parameter $\zeta = \omega N / r\lambda$ which indexes the relative frequency of mutations to progenitor cell loss/replacement events. In the particular case that $n = 0$ (i.e. when no cells bear a mutation at a given locus), the integral over ϕ can be performed explicitly and leads to the probability, $Q_0(t) = (1 + r\lambda t)^{-\zeta}$. For $n > 0$, the integral cannot be performed exactly, but does admit to useful limits. Specifically, when the mutation rate is low as compared to the loss/replacement rate, $\zeta \ll 1$, the integrand can be expanded and the full distribution converges to the form,

$$Q_n(t) = \left[1 - \zeta \ln(r\lambda t)\right] \delta_{n,0} + \zeta \frac{e^{-n/r\lambda t}}{n} (1 - \delta_{n,0}) + O(\zeta^2).$$

Note that, in this limit, the frequency of cells that have escaped mutation at a given locus declines only logarithmically with time. This slow (logarithmic) dependence reflects the fact that the majority of cells that acquire a mutation at a specific locus will most likely become lost through neutral competition with non-mutated neighbors.

From this result, we find that the size distribution of surviving mutant clones, $P_n(t) = Q_n(t)/(1 - Q_0(t))$, i.e. clones that contain at least one mutated progenitor cell,

$$P_n(t) = \frac{1}{\ln(r\lambda t)} \frac{e^{-n/r\lambda t}}{n}, \tag{2}$$

is independent of (and therefore insensitive to genomic variations in) the mutation rate, ω. This result (Eq. (1) of the main text) is easy to understand: In the limit $\zeta \ll 1$, there is typically at most only one clone that contributes to the frequency of mutations at the given locus. In this case, the distribution should converge to the time-integrated form of $p_n(t)$,

which gives rise to the $e^{-n/r\lambda t}/n$ dependence.

More generally, when the rate of mutation is larger, individual point mutations may not confer a unique clonal label. Instead, contributions from independent clones associated with mutations at the same locus will contribute to a net variable allele fraction. Qualitatively, such contributions are evidenced by the acquisition of a peak in the mutant clone size distribution. To determine an estimate for $Q_n(t)$, we can make use of a stationary phase approximation to evaluate the integral in Eq. (1). Varying the exponent of the integrand with respect to ϕ, we obtain the stationary phase solution,

$$\bar{\phi} = i \ln \left[\frac{n(1 + r\lambda t)}{r\lambda t(n + z)} \right].$$

Then, taking $r\lambda t \gg 1$, $n \gg \zeta \gtrsim 1$, and making use of the approximation,

$$e^{i\bar{\phi}n}(1 + r\lambda t(1 - e^{-i\bar{\phi}}))^{-\zeta} \simeq \exp\left[-\frac{n}{r\lambda t} - \zeta \ln\left(\frac{r\lambda t}{n} \right) \right].$$

Then, expanding the integrand in fluctuations $\eta = \phi - \bar{\phi}$, and integrating, we obtain

$$Q_{n>0}(t) \simeq e^{i\bar{\phi}n}(1 + r\lambda t(1 - e^{-i\bar{\phi}}))^{-\zeta} \int_{-\infty}^{\infty} \frac{d\eta}{2\pi} \exp\left[-\frac{n^2}{2\zeta}\eta^2 \right] \simeq \left(\frac{\zeta}{2\pi} \right)^{1/2} \left(\frac{r\lambda t}{n} \right)^{-\zeta} \frac{e^{-n/r\lambda t}}{n}.$$

Finally, using this result to construct the surviving clone size distribution, and normalizing, we obtain the general result,

$$P_n(t) \simeq \frac{1}{\Gamma[\zeta]} \left(\frac{r\lambda t}{n} \right)^{-\zeta} \frac{e^{-n/r\lambda t}}{n}, \tag{3}$$

where $\Gamma[\zeta]$ denotes the Gamma function.

Although Eqs. (2) and (3) provide a prediction with which to address deep sequencing data, the non-trivial dependence of the size distribution $P_n(t)$ on n makes a direct comparison with theory cumbersome. Fortunately, the size distribution is easily manipulated into a form where it translates to a simple exponential dependence. In particular, for $\zeta \ll 1$, if we define the average mutant clone size $\langle n(t) \rangle = r\lambda t / \ln(r\lambda t)$, the first incomplete moment (cf. Ref. [37])

$$\mu_1(n, t) \equiv \frac{1}{\langle n(t) \rangle} \sum_{m=n}^{\infty} m P_m(t) \approx e^{-n/r\lambda t}$$

acquires an exponential dependence on clone size n (Eq. (2) of the main text). In this form, the clone size dependence may be straightforwardly compared with experimental data. Conversely, in the limit when $\zeta \gtrsim 1$, if we define

$$\langle n^{1-\zeta}(t)\rangle \equiv \sum_{n=1}^{\infty} n^{1-\zeta} P_n(t) \simeq \frac{(r\lambda t)^{1-\zeta}}{\Gamma[\zeta]},$$

the generalized incomplete moment,

$$\mu_{1-\zeta}(n,t) \equiv \frac{1}{\langle n^{1-\zeta}(t)\rangle} \sum_{m=n}^{\infty} n^{1-\zeta} P_n(t) \approx e^{-n/r\lambda t},$$

also acquires a simply exponential form. Operationally, when the ratio of the mutation and loss/replacement rate, ζ, is unknown - which would typically be the case - it would be necessary to continuously adjust ζ until the incomplete moment $\mu_{1-\zeta}(n,t)$ acquires an exponential form.

Resolution limit of deep sequencing

The practical implementation of this analytical scheme relies on the ability to faithfully reconstruct the mutant clone size distribution from measurements of the variable allele fraction using deep sequencing. In practice, the reliability of this approach will be comprised by several factors, some of which have been addressed in the main text. However, one significant effect which we must consider is the impact of the resolution limit of the sequencing approach which will typically place a lower limit on the size of the smallest clones that can be resolved. However, if we adjust the size distribution to accommodate a cut-off, n_0, excluding mutant clones of size $n \leq n_0 \ll r\lambda t$, the generalized incomplete moment takes the form

$$\mu_{1-\zeta}(n,t) = e^{-(n-n_0)/r\lambda t},$$

for all ζ including $\zeta \to 0$. Therefore, it follows that the potential limitation of the deep sequencing approach in capturing the smallest clones does not change the exponential character of the size dependence of the generalized incomplete moment. Instead, it imposes an overall constant prefactor. Therefore, while the exponential character of the generalized incomplete moment can be assessed directly, a fit to the experimental data requires the adjustment of both $r\lambda t$ and the unknown size cut-off n_0. It is this procedure that we use to fit the data sets in Figs. 2, S5 and S6.

Alongside the resolution limit, the estimate of clone sizes from the measured variable allele fractions are subject to additional sources of error and uncertainty. First, clones may extend outside the boundaries of individual biopsies, leading to an underestimate of mutant clone size. Second, errors due to variations in the read depth of the sequencing will also compromise the accuracy of the method. Fortunately, the exponential character of the predicted clone size distribution serves to mitigate against the impact of such errors. In particular, suppose that the sequencing approach leads to a sampling error that translates to a normal distribution of clone sizes, i.e. clones of "true" size n_{T} are recorded at a frequency of

$$G(n, n_{\mathrm{T}}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(n-n_{\mathrm{T}})^2/2\sigma^2}.$$

where σ denotes the standard deviation of the error. In this case, the first incomplete moment would be given by

$$\begin{aligned}
\mu_1(n,t) &\approx \int_{-\infty}^{\infty} dm_{\mathrm{T}} \int_n^{\infty} dm\, G(m, m_{\mathrm{T}})\, \frac{m}{m_{\mathrm{T}}} e^{-m_{\mathrm{T}}/r\lambda t} \\
&= e^{-n/r\lambda t} + \frac{\sigma^2}{(r\lambda t)^2} \int_n^{\infty} \frac{dm}{r\lambda t}\, e^{-m/r\lambda t}\, \frac{m^2 + 2mr\lambda t + 2(r\lambda t)^2}{2m^2} + O(\sigma/r\lambda t)^4 \\
&= e^{-n/r\lambda t} + \frac{\sigma^2}{(r\lambda t)^2} \left(1 + \frac{r\lambda t}{n}\right) e^{-n/r\lambda t} + O(\sigma/r\lambda t)^4.
\end{aligned}$$

Therefore, providing $\sigma \ll \sqrt{n\, r\lambda t}$, the correction of the first incomplete moment due to fluctuations will be small. For the smallest clone sizes n, this condition cannot be met. However, such contributions are in any case below the resolution limit of the sequencing approach. Conversely, for $n = O(r\lambda t)$, consistent with the typical clone sizes examined here, this condition will be safely met.

Generalizations of the analytical approach

Until now, we have focussed our analysis on the dynamics of cycling adult tissues in which the self-renewing progenitor cell population conforms to a pattern of population asymmetry in which stochastic fate behavior follows from intrinsic (cell-autonomous) regulation. As discussed above, for systems in which population asymmetric follows from extrinsic regulation (e.g. when stem cells compete neutrally for limited niche access), in dimensions of two (epithelia) and higher (volumar), the mutant clone size distribution (and the corresponding incomplete moments) are predicted to assume the same functional dependence, i.e. as with a pulse-labeling clonal assay, in dimensions of two and above, the mechanism of regula-

tion (intrinsic vs. extrinsic) cannot be inferred from the analysis of the mutant clone size distribution alone.

However, as emphasized above, if cell dynamics follows from extrinsic regulation, in the one-dimensional or quasi one-dimensional geometry (viz. ductal or tubular tissues), the mutant clone size distribution must be revised. In particular, when the mutation rate, ω, is low as compared to the rate of stem cell loss and replacement, the mutant clone size distribution takes the form,

$$Q_n(t) \simeq \begin{cases} 1 - \dfrac{\omega N}{\lambda}\sqrt{\dfrac{4\lambda t}{\pi}} & n = 0 \\ \dfrac{\omega N}{\lambda}\mathrm{Erfc}\left[\dfrac{n}{2\sqrt{\lambda t}}\right] & n > 0 \end{cases}$$

where $\mathrm{Erfc}[x] = \frac{2}{\sqrt{\pi}}\int_x^\infty dz\, e^{-z^2}$ denotes the complementary error function. In this case, the frequency of non-mutated clones diminishes more rapidly, as the clonal loss rate due to neutral competition scales only as $1/\sqrt{t}$. As a result, the size distribution of surviving mutant clones is predicted to take the form,

$$\begin{aligned} P_n(t) = \frac{Q_n(t)}{1 - Q_0(t)} &= \sqrt{\frac{\pi}{4\lambda t}}\mathrm{Erfc}\left[\frac{n}{2\sqrt{\lambda t}}\right] \\ &\simeq \begin{cases} \sqrt{\dfrac{\pi}{4\lambda t}}\left(1 - \dfrac{n}{\sqrt{\pi t}} + O(n/\sqrt{t})^3\right) & n^2/\lambda t \ll 1 \\ \dfrac{1}{n}\exp\left[-\dfrac{n^2}{4\lambda t}\right]\left(1 + O(4\lambda t/n^2)\right) & n^2/\lambda t \gg 1 \end{cases} . \end{aligned}$$

Then, with the average clone size given by $\langle n(t)\rangle = \sum_{n=1}^\infty nP_n(t) \simeq \frac{\sqrt{\pi\lambda t}}{2}$, the first incomplete moment takes the form

$$\mu_1(n,t) = \frac{1}{\langle n(t)\rangle}\sum_{m=n}^\infty mP_m(t) \simeq \left(1 - \frac{n^2}{2\lambda t}\right)\mathrm{Erfc}\left[\frac{n}{2\sqrt{\lambda t}}\right] + \frac{n}{\sqrt{\pi\lambda t}}\exp\left[-\frac{n^2}{4\lambda t}\right]$$

In particular, in the limit $n^2/\lambda t \gg 1$, when $\mathrm{Erfc}[n/2\sqrt{\lambda t}] \simeq \sqrt{\frac{4\lambda t}{\pi}}\frac{e^{-n^2/4\lambda t}}{n}$, we have

$$\mu_1(n,t) \simeq \mathrm{Erfc}\left[\frac{n}{2\sqrt{\lambda t}}\right] \simeq \sqrt{\frac{4\lambda t}{\pi}}\frac{e^{-n^2/4\lambda t}}{n}.$$

Finally, if tissues are divided into an ensemble of isolated niche domains, such as the

glandular crypt structures found in the intestinal tract and stomach, the mutant clone size distribution must be revised again. In this case, the chance induction of stem cells through somatic mutation may, through neutral competition, lead to the clonal fixation of individual glands in which all constituent cells become monoclonal. With N_S defining the (effective) stem cell number per gland, we expect the frequency of monoclonal glands associated with a specific mutation to grow as $\omega N_g t$, where N_g denotes the total number of glands in the biopsy sample, and ω denotes the locus-specific mutation rate per stem cell. Alongside this growing fraction of monoclonal glands, at any given instant in time, we expect to find a time-independent distribution of partially mutated glands corresponding to clones whose fate (extinction through neutral competition or fixation) has yet to resolve. In the quasi one-dimensional arrangement of stem cells in the intestinal crypt of the colon or small intestine, the size distribution of these partially labeled crypts is given simply by

$$P_{N_S>n\geq 1}(t) = \frac{2(N_S - n)}{N_S(N_S - 1)},$$

independent of time, t.

References

39. N. T. J. Bailey, The Elements of Stochastic Processes with Applications to the Natural Sciences. New York City: John Wiley and Sons, Inc. (1964).

40. M. Q. Zhang, *Statistical features of human exons and their flanking regions*, Hum. Mol. Gen. **7**, 919-932 (1998).

Supplementary Figure S1. Abundance of gene-specific mutant clones.
(A) Scatter plot of the exome length in the genome against the total number of mutant clones associated with that exome (coverage obtained from AceView, NBI). The points show all genes considered in the deep sequencing study of Martincorena *et al.* In particular, we have picked out the abundance of specific genes that emerge as outliers from the otherwise neutral size distribution of mutant clones, as discussed in the main text. The results reveal no obvious correlation of exome length with clone number, emphasizing the degree of intrinsic variation of the mutation rate along the genome. (B, left) In common with previous studies of translated exons [40], the distribution of sequenced exon lengths (bars) is consistent with a log-normal distribution, $\exp[-(\mathrm{Log}_2(\mathrm{Length}) - \mathrm{Log}_2 a)^2/2\sigma]$ centered on $a = 2^{7.0}$ bp with a width $\sigma = 2.9$ (red line). (B, right) Although there is no obvious correlation between exon length and mutant clone abundance, the corresponding distribution of mutant clone number for the chosen exons is also broadly consistent with a log-normal distribution centerd on $a = 2^{5.2}$ bp with the same width, σ.

Supplementary Figure S2. Neutral dynamics and scaling of the clone size distribution.
(A) According to the paradigm of balanced stochastic fate choice (Fig. 1B), following pulse-labeling, the average clone size (progenitor number) is predicted to increase as $1 + r\lambda t$ (black line, SI Text), while the clone survival probability falls as $1/(1 + r\lambda t)$ (orange line), so that the overall average number of labeled progenitors per clone (brown line) remains constant at unity. (B) The corresponding cumulative clone size distribution converges onto an exponential dependence, a manifestation of scaling behavior (black line, SI Text). Points in A and B represent the results of stochastic simulation of the corresponding critical birth-death process, with chase times in B of $\lambda t = 2^6$ (orange), 2^8 (brown), 2^{10} (grey), and 2^{12} with $r = 1/2$ (black).

Supplementary Figure S3. Distribution of mutant clone sizes.
Sizes of the $N = 1557$, 725, 803 and 195 clones associated with (synonymous and non-synonymous) point mutations that are contained within single biopsies for, respectively, patient PD18003, PD13634, PD20399 and PD21910 before filtering.

Supplementary Figure S4. Mutational profile of biopsies containing outlier clones in patient PD18003.
Mutational profile of the 7 mutant clones (denoted by **) in patient PD18003 considered as outliers by size and contained within 6 biopsies with the specific mutant genes marked

in red. Large mutant clones with similar sizes (also marked in red) are likely to represent subclones of the outlier mutant clone. Non-point mutations are marked in grey.

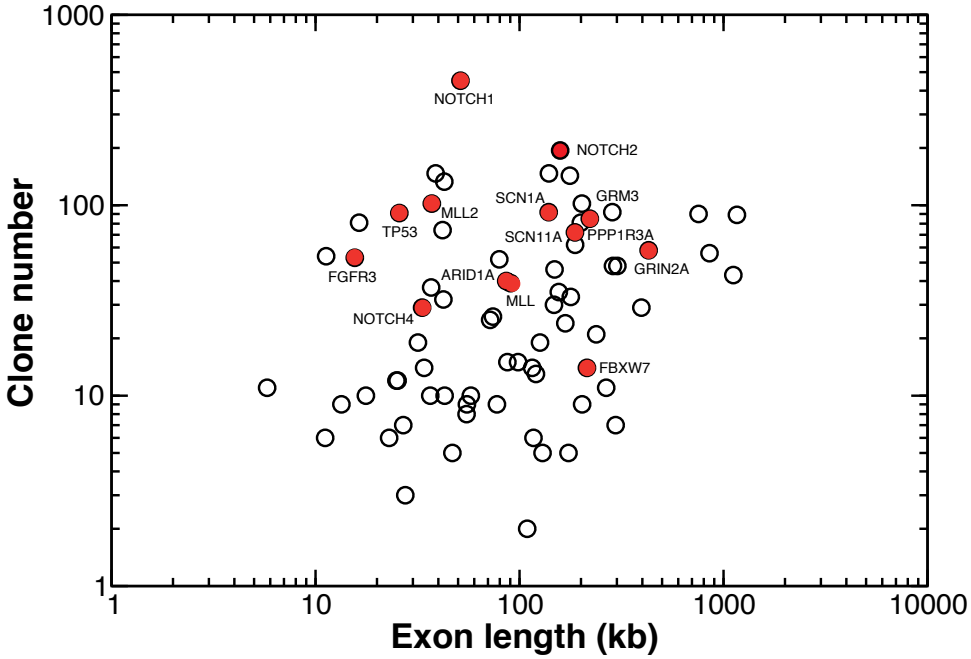
Supplementary Figure S5. Neutral progenitor cell dynamics is conserved across patients.
First incomplete moment, $\mu_1(n,t)$, associated with $N = 159$, 154 and 25 synonymous mutations that belong to single biopsies (points) of, respectively, patient (A) PD13634, (C) PD20399 and (E) PD21910 show collapse onto the predicted exponential size dependence (Eq. (2), red line and Table 2). First incomplete moment associated with all $N = 647$, 724 and 181 point mutations that belong to single biopsies (points) of, respectively, patient (B) PD13634, (D) PD20399 and (F) PD21910 also show collapse onto the predicted exponential size dependence (Eq. (2), red line and Table 2). Error bars denote s.e.m.

Supplementary Figure S6. Filtered clone size distribution confirms neutral clone dynamics.
(A-C) and (D-F) show the first incomplete moment, $\mu_1(n,t)$, and clone size distribution, $P_n(t)$, for patients PD13634 and PD20399, respectively, following the removal of biopsies containing the two clones that span multiple biopsies depicted in Fig. 3. Points show data and the lines show a fit to the predicted distributions Eqs. (1) and (2). Error bars denote s.e.m.

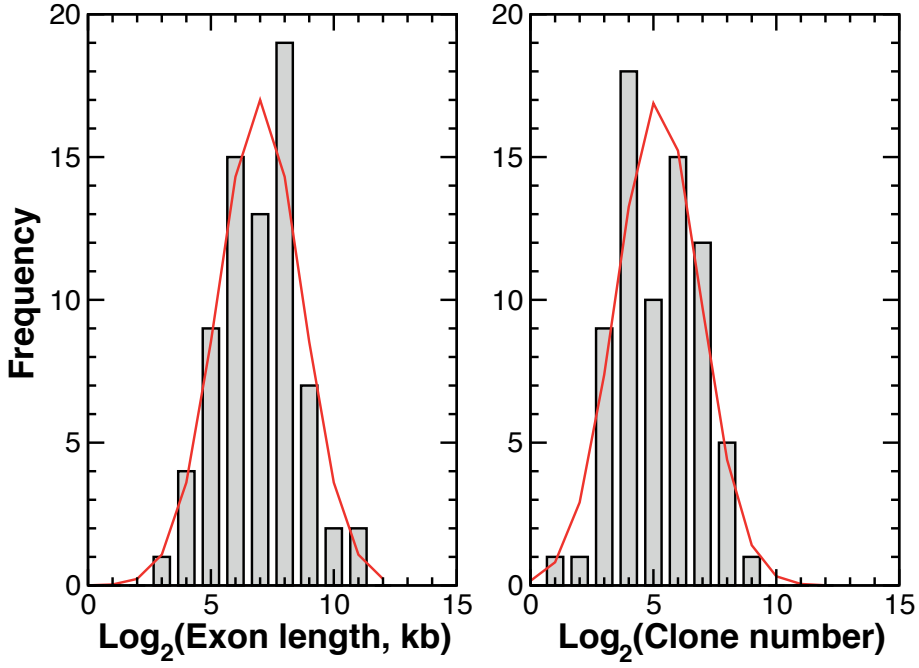
Supplementary Figure S7. Average clone size for typical driver genes is consistent with neutral dynamics.
Average size of mutant clones belonging to different cancer drivers (black) compared by the ensemble average of all genes (red) for patient (A) PD18003, (B) PD13634, (C) PD20399 and (D) PD21910. Note that, for the majority of genes and patients, the departure of the average clone size for cancer drivers from the ensemble average is not statistically significant. Error bars denote s.e.m. To test for inequivalence of specific genes from the ensemble average, we have used a Kolmogorov-Smirnov test to compare mutant clone size distributions. The p-values from the four patients, PD18003, PD13634, PD20399 and PD21910 are, respectively, given by NOTCH1: 0.050, 0.18, 0.025 and 0.25; NOTCH2: 0.081, 0.31, 0.97 and 0.25; NOTCH3: 0.34, 0.15, 0.66 and 0.19; NOTCH4: 0.43, 0.14, 0.99 and N/A; FGFR3: 0.16, 0.54, 0.012 and N/A; ARID1A: 0.51, 0.81, 0.043 and 0.92; PPP1R3A: 1.00, 0.17, 0.27 and 0.072; TP53: 0.53, 0.06, 0.33 and 0.37; FAT1: 0.46, 0.96, 0.92 and 0.044; GRM3: 0.95, 0.18, 0.64 and 0.92; SCN1A: 0.51, 0.65, 0.078 and N/A; FBXW7: 0.84, N/A,

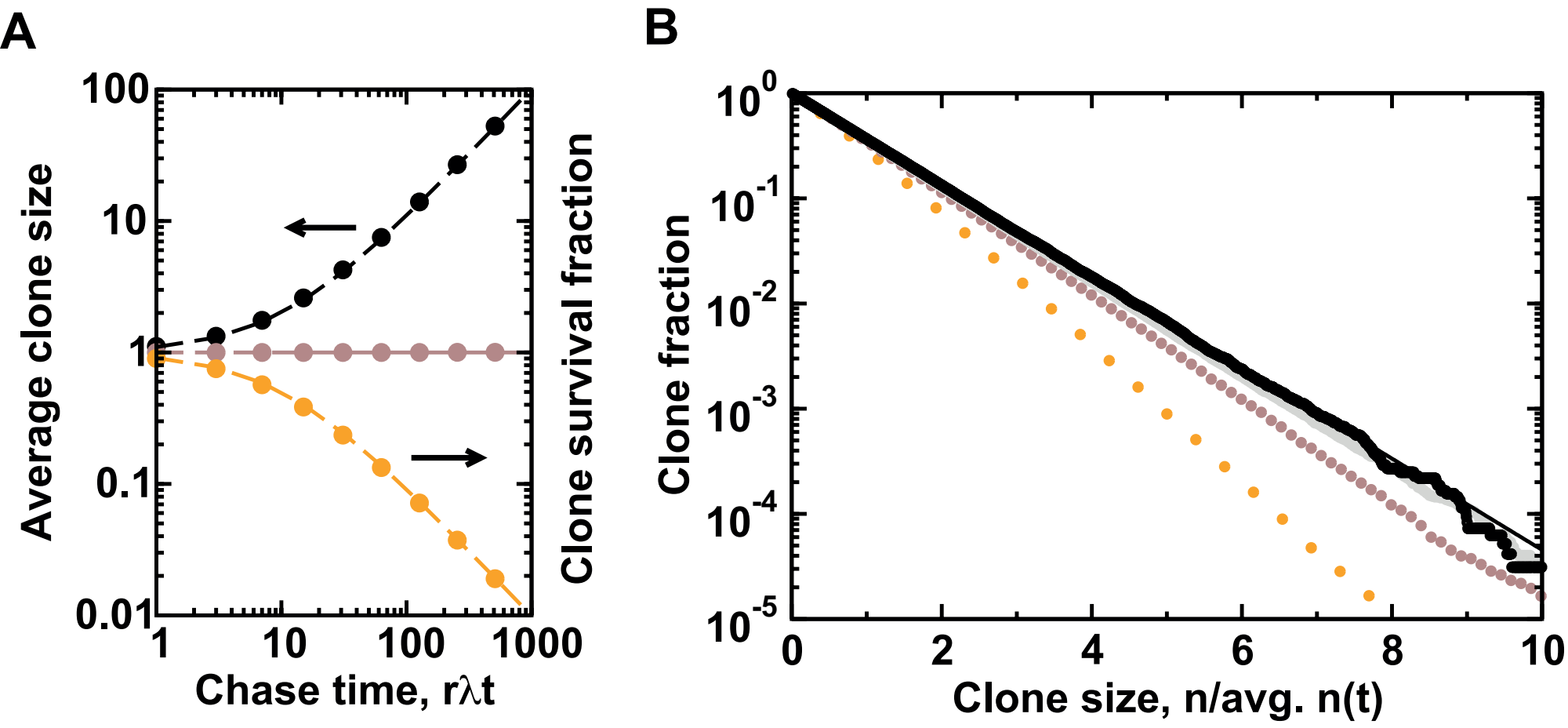
0.65 and N/A. These results show that, for all but three entries, the statistical equivalence of the distributions cannot be ruled out (at a significance level of 0.05).

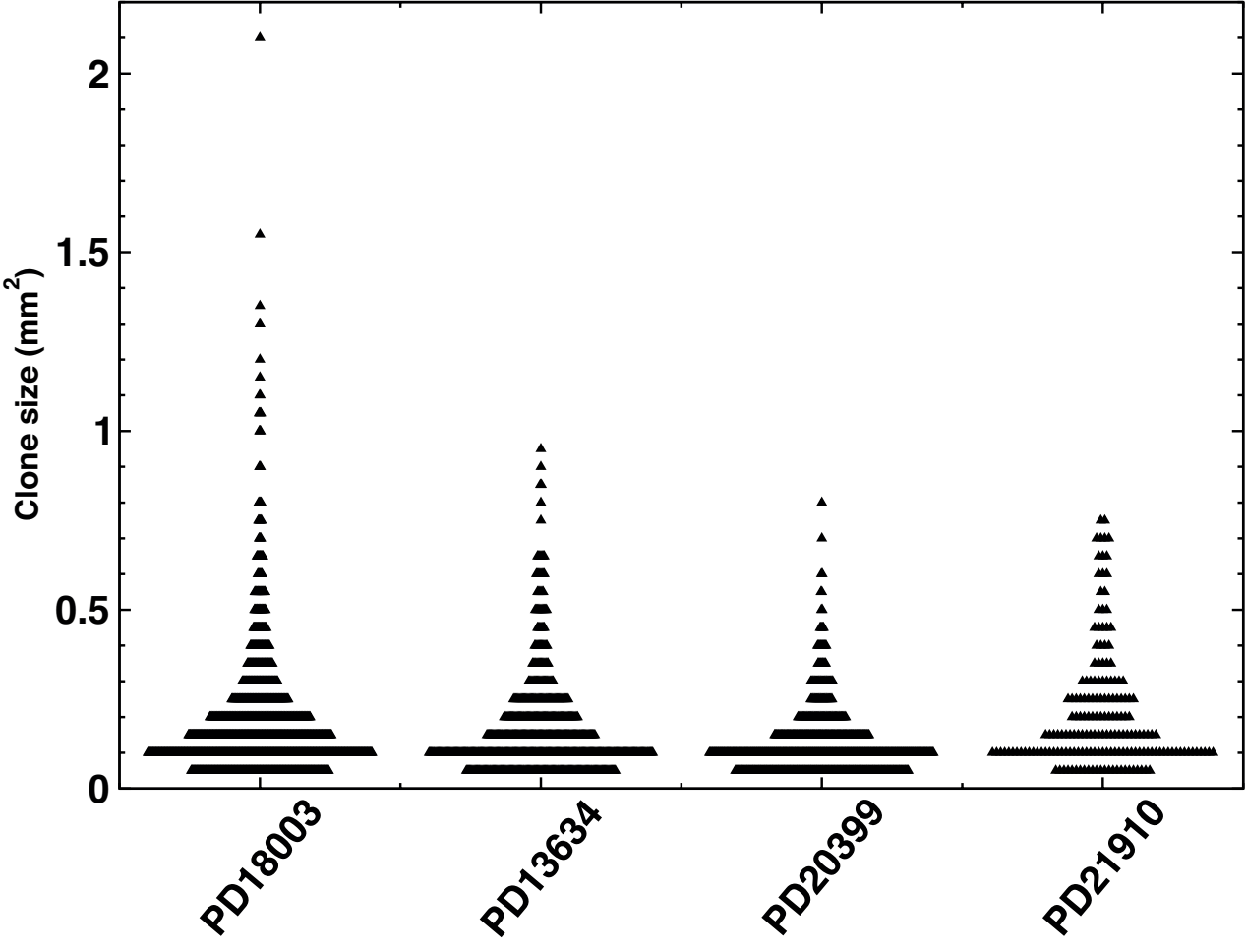
A

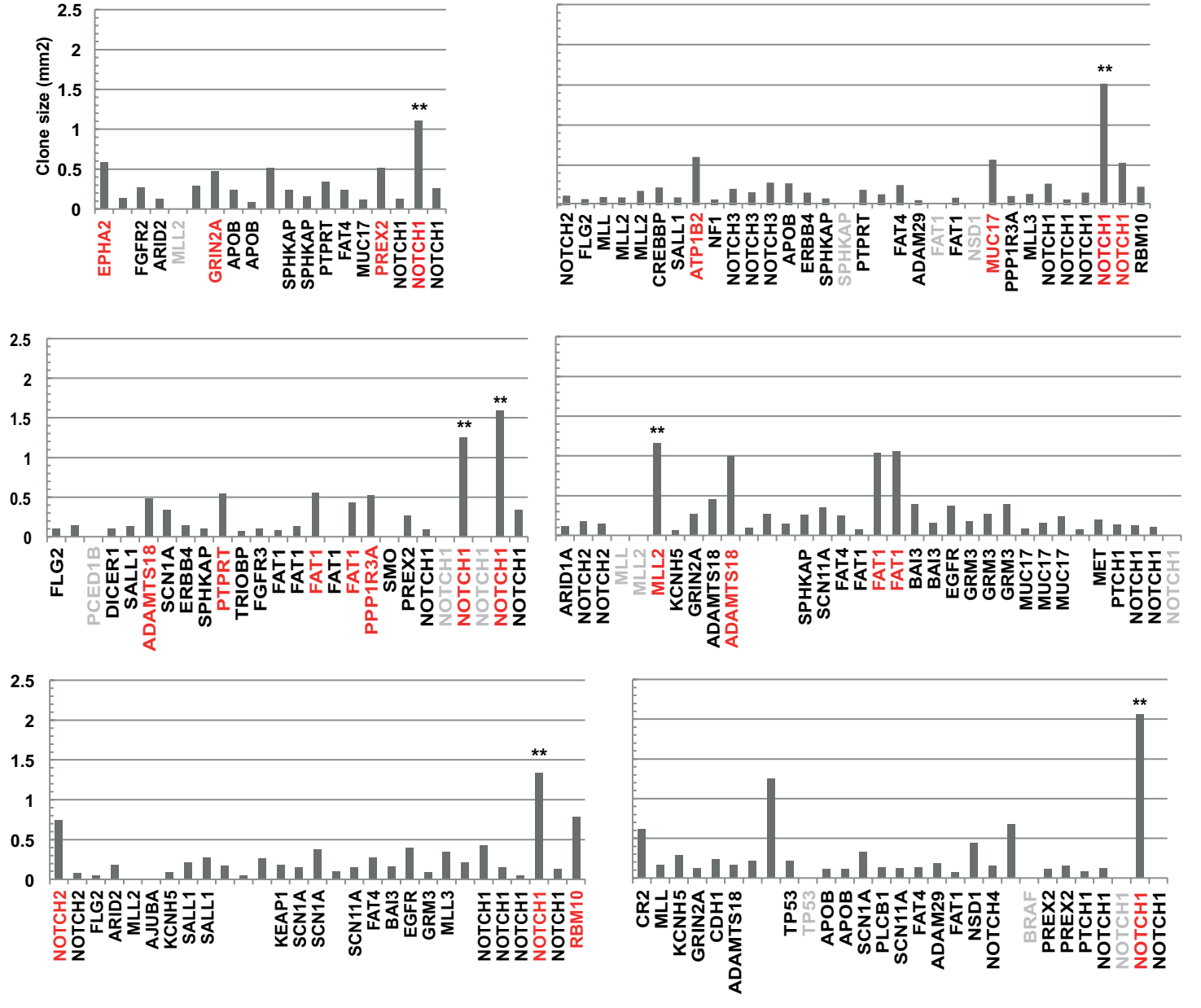

1000
100
10
1
Clone number
1
10
100
1000
10000
Exon length (kb)
NOTCH1
NOTCH2
SCN1A
GRM3
MLL2
TP53
SCN11A
PPP1R3A
GRIN2A
FGFR3
ARID1A
MLL
NOTCH4
FBXW7

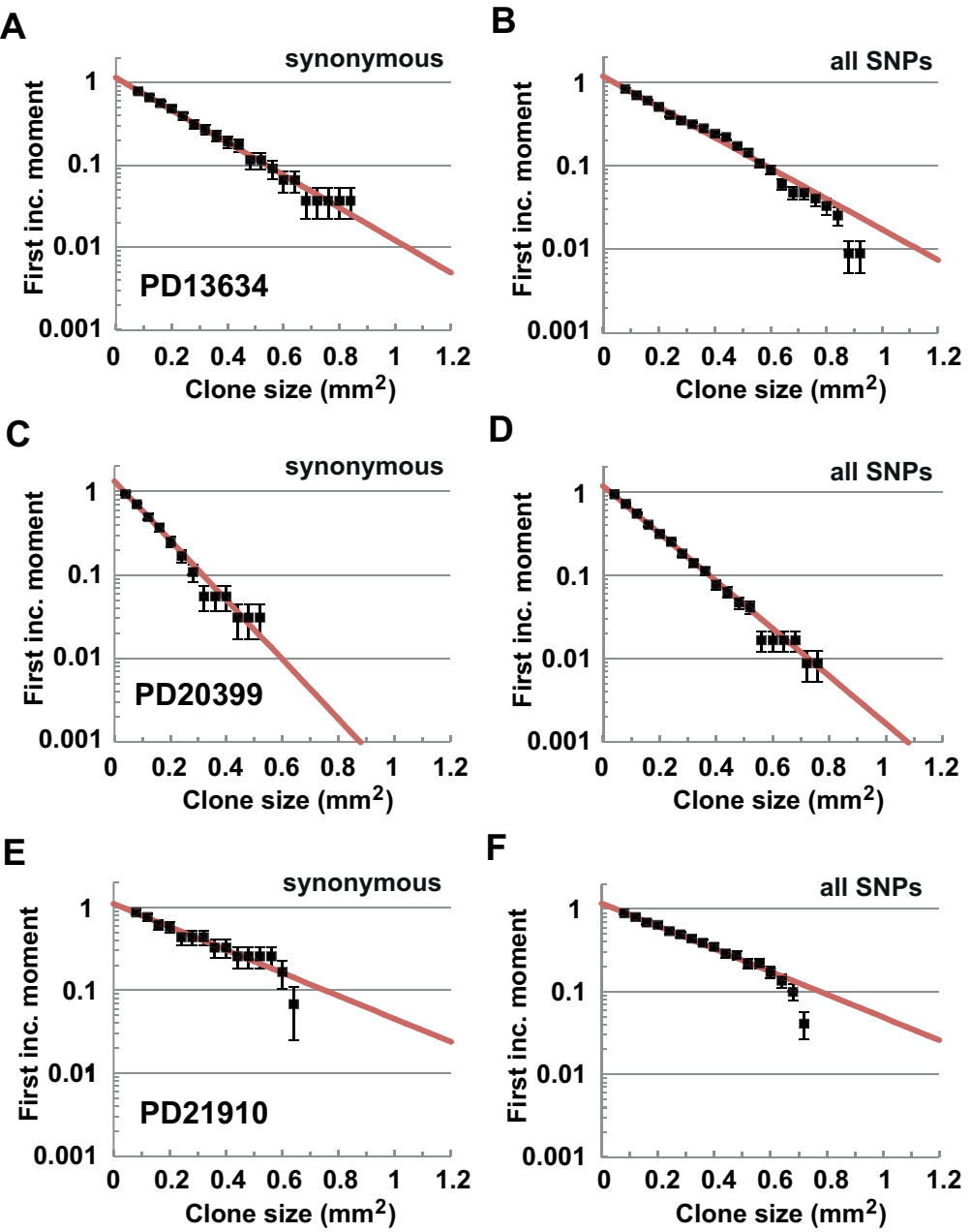
B

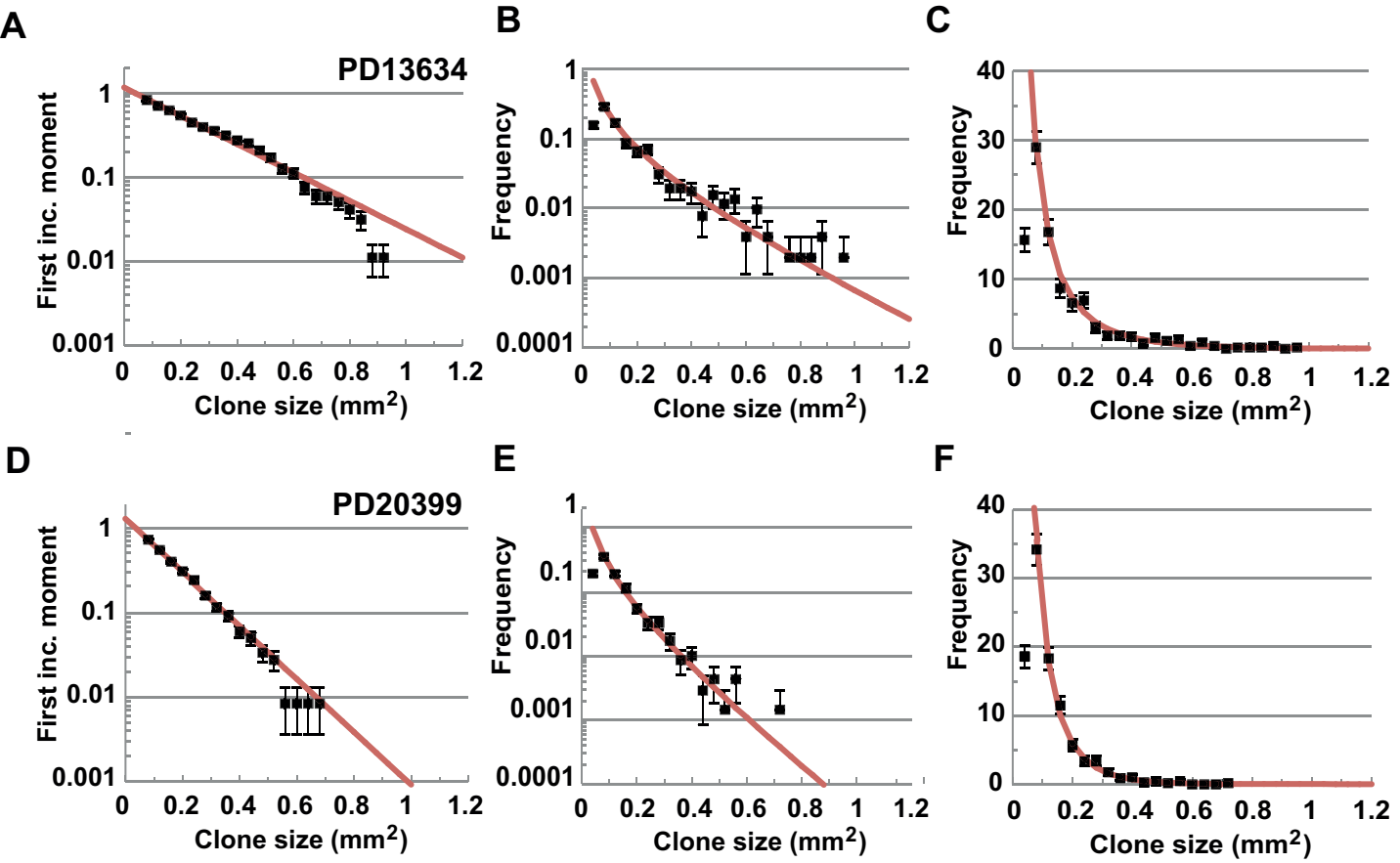

20
15
10
5
0
Frequency
0
5
10
15
Log2(Exon length, kb)
Log2(Clone number)

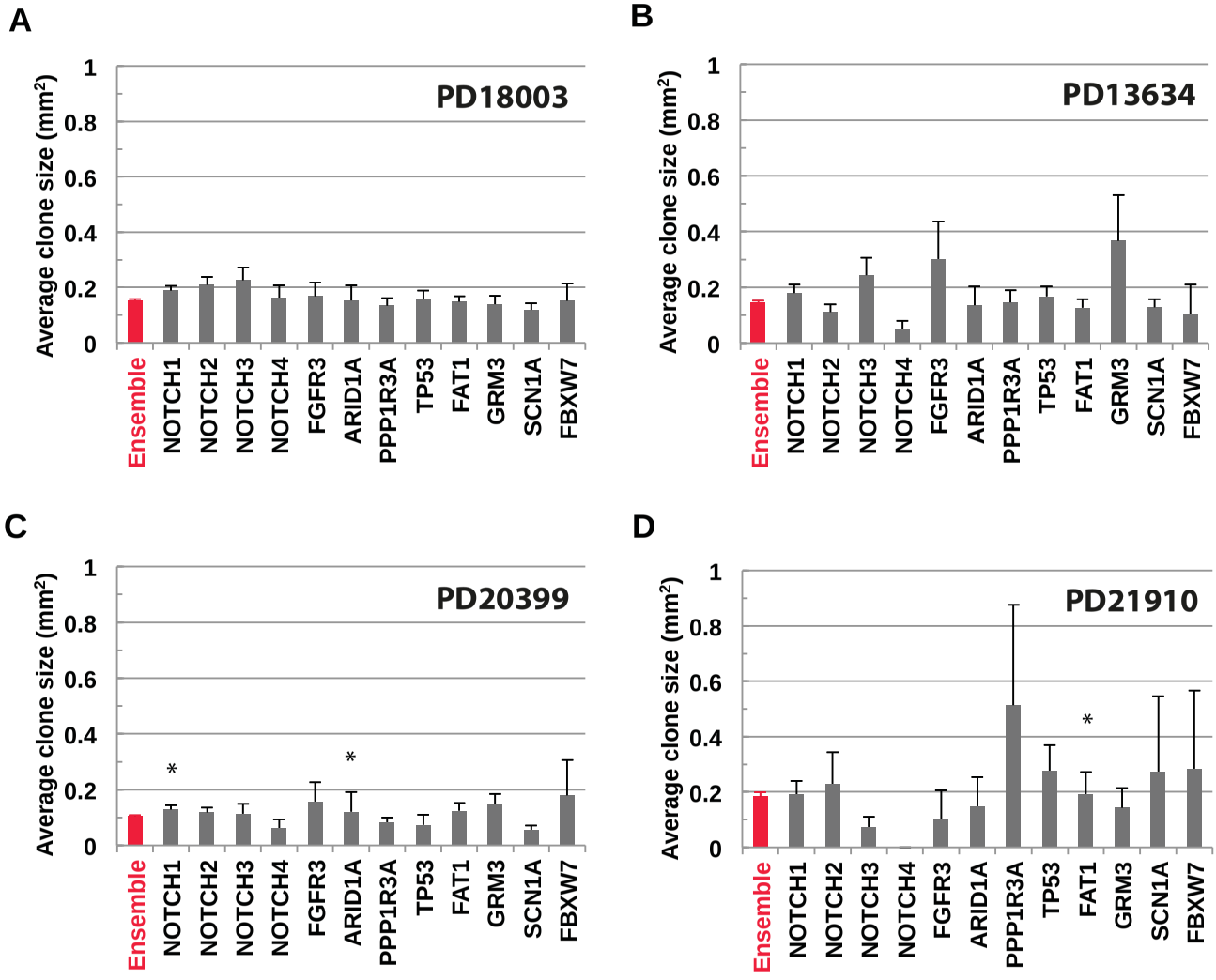

A
B
Average clone size
Clone survival fraction
Chase time, rλt
Clone fraction
Clone size, n/avg. n(t)
100
10
1
0.1
0.01
1
10
100
1000
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
2
4
6
8
10


Clone size (mm2)
2
1.5
1
0.5
0
PD18003
PD13634
PD20399
PD21910


Clone size (mm2)
2.5
2
1.5
1
0.5
0
**
EPHA2 FGFR2 ARID2 MLL2 GRIN2A APOB APOB SPHKAP SPHKAP PTPRT FAT4 MUC17 PREX2 NOTCH1 NOTCH1 NOTCH1
**
NOTCH2 FLG2 MLL MLL2 MLL2 CREBBP SALL1 ATP1B2 NF1 NOTCH3 NOTCH3 NOTCH3 APOB ERBB4 SPHKAP SPHKAP PTPRT FAT4 ADAM29 FAT1 FAT1 NSD1 MUC17 PPP1R3A MLL3 NOTCH1 NOTCH1 NOTCH1 NOTCH1 NOTCH1 RBM10
2.5
2
1.5
1
0.5
0
** **
FLG2 PCED1B DICER1 SALL1 ADAMTS18 SCN1A ERBB4 SPHKAP PTPRT TRIOBP FGFR3 FAT1 FAT1 FAT1 FAT1 FAT1 PPP1R3A SMO PREX2 NOTCH1 NOTCH1 NOTCH1 NOTCH1 NOTCH1
**
ARID1A NOTCH2 NOTCH2 MLL MLL2 MLL2 KCNH5 GRIN2A ADAMTS18 ADAMTS18 SPHKAP SCN11A FAT4 FAT1 FAT1 FAT1 BAI3 BAI3 EGFR GRM3 GRM3 GRM3 MUC17 MUC17 MUC17 MET PTCH1 NOTCH1 NOTCH1 NOTCH1
2.5
2
1.5
1
0.5
0
**
NOTCH2 NOTCH2 FLG2 ARID2 MLL2 AJUBA KCNH5 SALL1 SALL1 KEAP1 SCN1A SCN1A SCN11A FAT4 BAI3 EGFR GRM3 MLL3 NOTCH1 NOTCH1 NOTCH1 NOTCH1 NOTCH1 RBM10
**
CR2 MLL KCNH5 GRIN2A CDH1 ADAMTS18 TP53 TP53 APOB APOB SCN1A PLCB1 SCN11A FAT4 ADAM29 FAT1 NSD1 NOTCH4 BRAF PREX2 PREX2 PTCH1 NOTCH1 NOTCH1 NOTCH1 NOTCH1


A
synonymous
PD13634
First inc. moment
1
0.1
0.01
0.001
0
0.2
0.4
0.6
0.8
1
1.2
Clone size (mm2)
B
all SNPs
First inc. moment
Clone size (mm2)
C
synonymous
PD20399
First inc. moment
Clone size (mm2)
D
all SNPs
First inc. moment
Clone size (mm2)
E
synonymous
PD21910
First inc. moment
Clone size (mm2)
F
all SNPs
First inc. moment
Clone size (mm2)


A
PD13634
First inc. moment
1
0.1
0.01
0.001
0
0.2
0.4
0.6
0.8
1
1.2
Clone size (mm2)
B
Frequency
1
0.1
0.01
0.001
0.0001
Clone size (mm2)
C
Frequency
40
30
20
10
0
Clone size (mm2)
D
PD20399
First inc. moment
Clone size (mm2)
E
Frequency
Clone size (mm2)
F
Frequency
Clone size (mm2)


A
PD18003
Average clone size (mm²)
1
0.8
0.6
0.4
0.2
0
Ensemble
NOTCH1
NOTCH2
NOTCH3
NOTCH4
FGFR3
ARID1A
PPP1R3A
TP53
FAT1
GRM3
SCN1A
FBXW7
B
PD13634
Average clone size (mm²)
1
0.8
0.6
0.4
0.2
0
Ensemble
NOTCH1
NOTCH2
NOTCH3
NOTCH4
FGFR3
ARID1A
PPP1R3A
TP53
FAT1
GRM3
SCN1A
FBXW7
C
PD20399
Average clone size (mm²)
1
0.8
0.6
0.4
0.2
0
*
*
Ensemble
NOTCH1
NOTCH2
NOTCH3
NOTCH4
FGFR3
ARID1A
PPP1R3A
TP53
FAT1
GRM3
SCN1A
FBXW7
D
PD21910
Average clone size (mm²)
1
0.8
0.6
0.4
0.2
0
*
Ensemble
NOTCH1
NOTCH2
NOTCH3
NOTCH4
FGFR3
ARID1A
PPP1R3A
TP53
FAT1
GRM3
SCN1A
FBXW7