Supplemental Material document

Supplemental Figures:

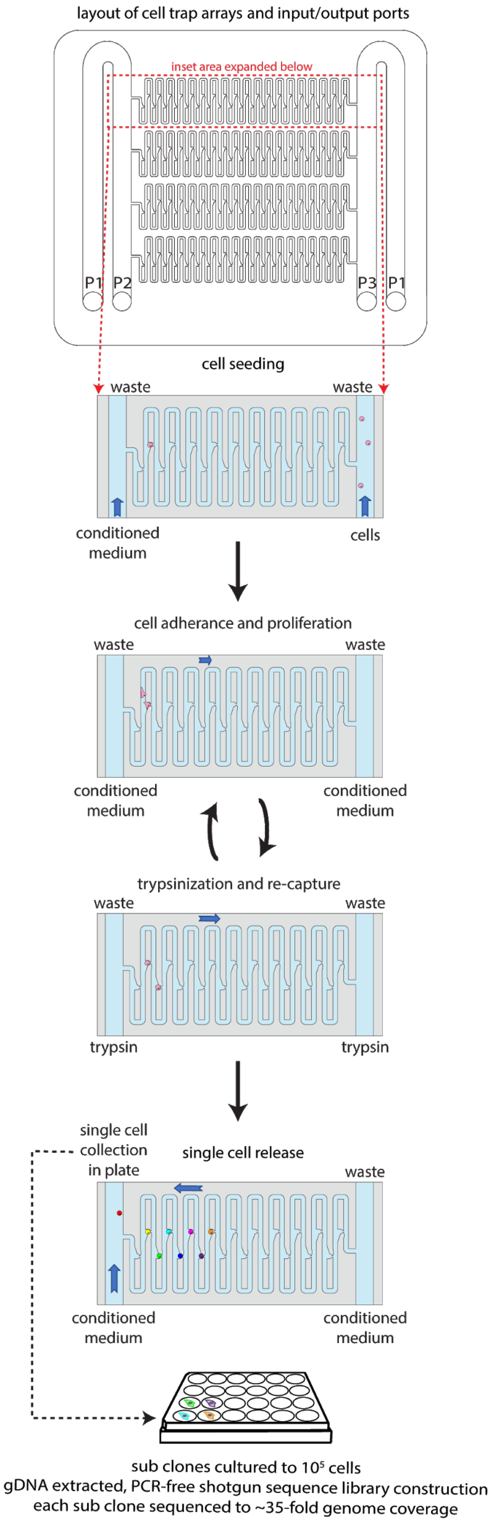

layout of cell trap arrays and input/output ports
inset area expanded below
P1
P2
P3
P1
cell seeding
waste
waste
conditioned medium
cells
cell adherance and proliferation
waste
waste
conditioned medium
conditioned medium
trypsinization and re-capture
waste
waste
trypsin
trypsin
single cell collection in plate
single cell release
waste
conditioned medium
conditioned medium
sub clones cultured to 10^5 cells
gDNA extracted, PCR-free shotgun sequence library construction
each sub clone sequenced to ~35-fold genome coverage

Figure S1: Microfluidic system workflow for seeding, culturing, and collecting cells.

Diagram of the device microarchitecture including cell trap arrays and ports used to introduce buffers and cell culture medium. Flow across the cell trap arrays is used to manipulate cell position. The cell trap arrays are flanked on either side by bypassing fluidic channels which enable rapid fluid exchange and media perfusion through the device while limiting the amount of shear force experienced by captured cells. The inset diagrams show the series of operations used to seed a cell into a trap array which allow the cell to adhere to the device, proliferate, trypsinize, distribute, re-capture cells and release cells individually from a cell trap array for collection in standard laboratory plasticware for sub-clonal culture. Together, the device features and this protocol allow for long-term culture and manipulation of cells under time lapse imaging prior to isolation of single cells for downstream culture and analysis by lineage sequencing.

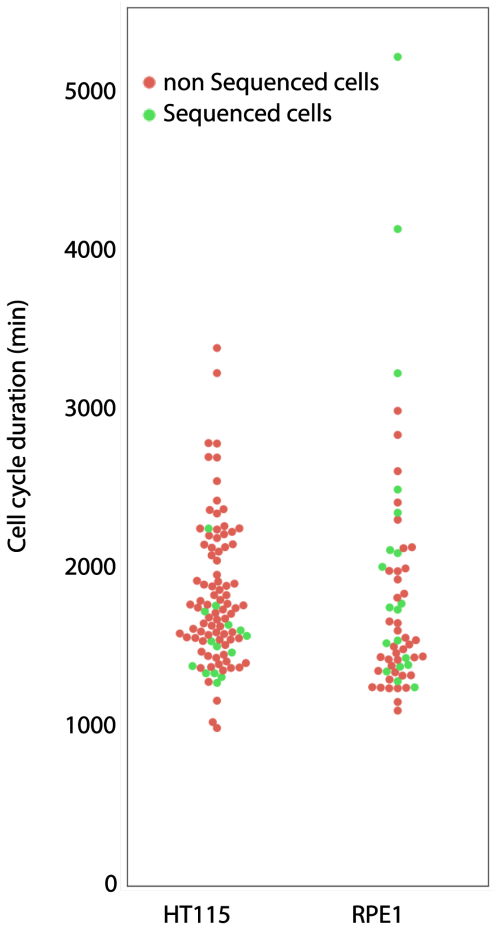


Figure S2: Interdivision time (cell cycle duration) for all tracked lineages

Plot of cell cycle duration from all tracked lineages (full cell cycles presented). The data combines several different lineages (HT115: 18 different ancestors, RPE1: 7 different ancestors). In green: all interdivision cell cycles from the cells that were sequenced and analyzed in this study. The sequenced cells show a similar distribution of interdivision times compared with the overall distribution indicating there is little or no bias in the analyzed set with respect to interdivision time.

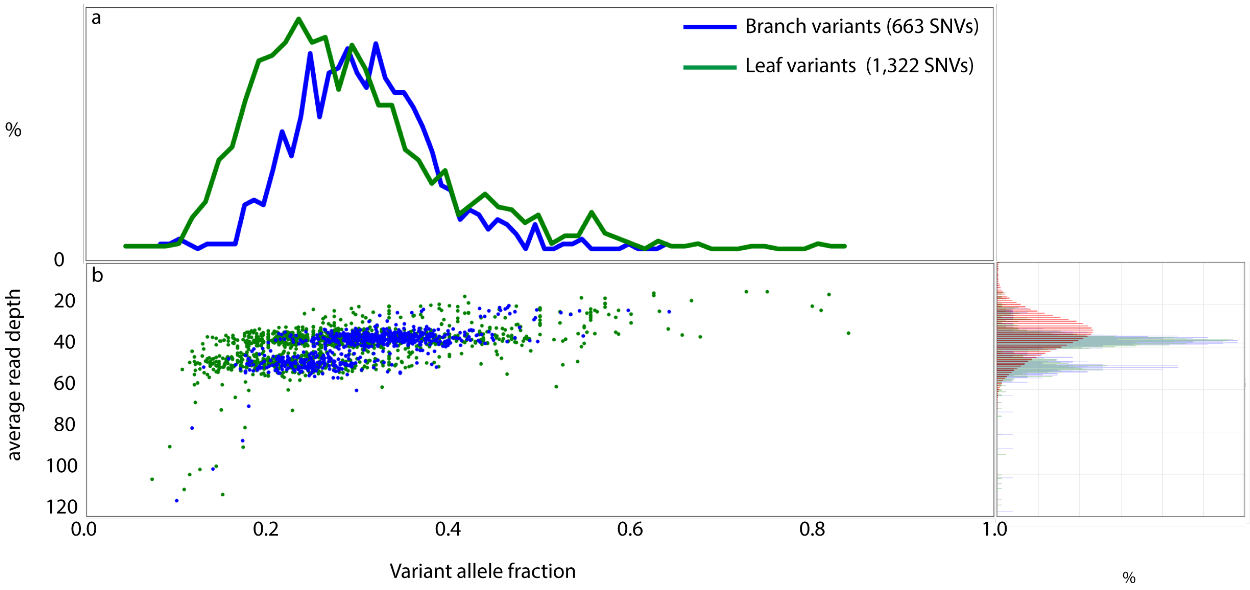


Figure S3: RPE1 branch variants mutations are clonal.

a) Histogram of allele fraction distribution for RPE1 SNVs comparing ‘branch variants’ and ‘leaf variants’. Branch variants are clonal, with the branch variant allele fraction centered near a value of 0.33 as expected for our predominantly triploid RPE1 cells (see Supplemental Figure S4). Similarly, to the HT115 cells, RPE1 leaf variants show a shift to lower allele fraction values.

b) Scatter plot of variants; average read depth versus allele fraction; branch variants (blue), and leaf variants (green). On the right panel: normalized histogram of read coverage depth for RPE1 lineage; whole genome (red), branch variants (blue), and leaf variants (green).

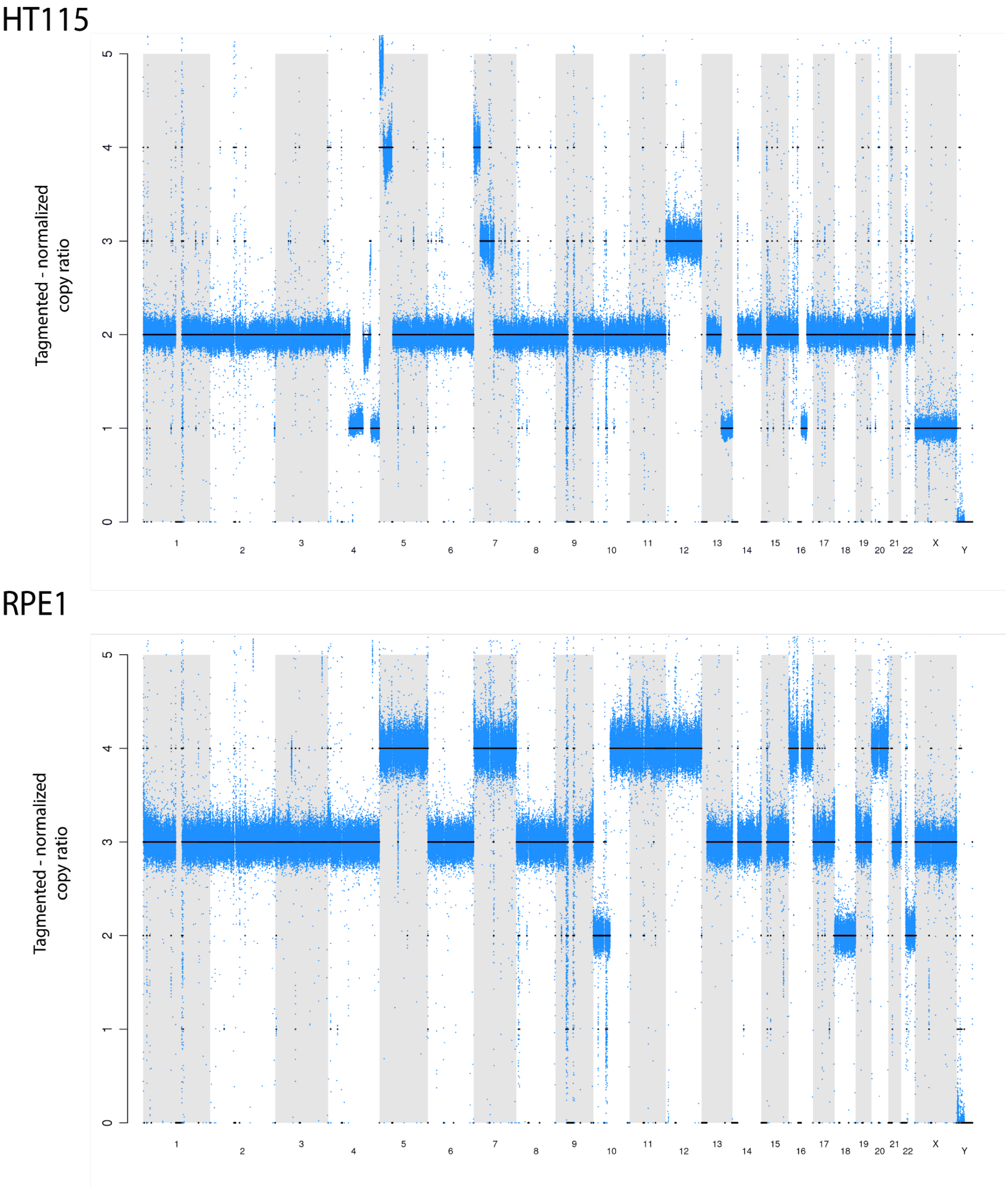


Figure S4: Genomic DNA copy number assessment

Copy number analysis for the HT115 cell line shows that the genome is mostly diploid (upper plot), while the RPE1 cell line presents mostly triploid and tetraploid genome (bottom plot). The copy numbers values were calculated for every sample using the whole genome sequencing data (WGS), and the locus-wise copy number estimate used to inform SNV analysis and genome size for mutation rate calculations.

a.

	Raw SNVs	loci with read coverage >= 5	total coincident SNVs	coincident SNV quality score >= 0.99	SNVs not associated loss of heterozygosity
HT115 Mutect1 (hg19)	64,770	64,628	18,533	3,027	2,779
HT115 Strelka (hg19)	111,784	110,871	14,712	3,277	2,810
RPE1 Mutect1 (hg19)	50,946	50,785	17,660	663	663
RPE1 Strelka (hg19)	36,762	35,787	12,868	685	679
					branch variant call set

b.

HT115

Branch	Strelka	MuTect1	
Strelka	2810	2765	98.4 %
MuTect1	2765	2779	99.5 %

Leaf	Strelka	MuTect1	
Strelka	9877	9276	93.9 %
MuTect1	9276	9883	93.9 %

RPE1

Branch	Strelka	MuTect1	
Strelka	679	659	97.1 %
MuTect1	659	663	99.4 %

Leaf	Strelka	MuTect1	
Strelka	1277	1030	80.7 %
MuTect1	1030	1322	77.9 %

c.

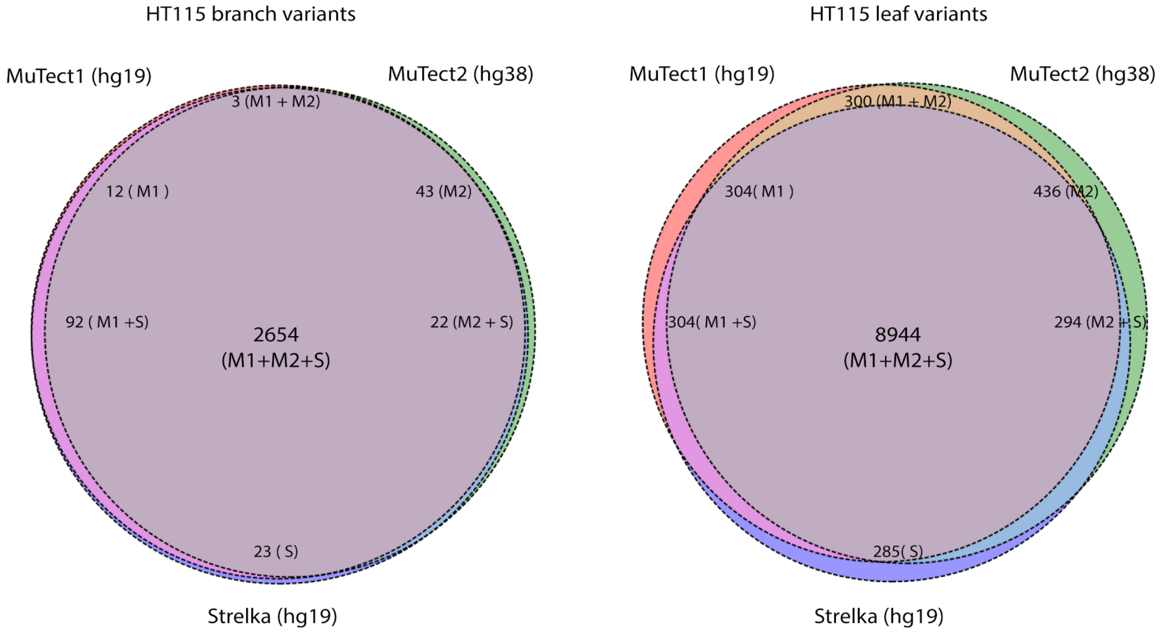


Figure S5: High consistency of branch variant calls across different genome references and variant callers

- **a)** Variant counts at different stages of the informatics filtering steps for Strelka with comparison to MuTect1 (both mapped to hg19). Many raw variants are filtered out when high quality coincident (branch) variants are selected.
- **b)** Comparison between variant calling with Strelka and MuTect1 (both mapped to hg19), showing high consistency in branch variants calling across algorithms (>97% in HT115 and RPE1 lineages) (upper tables). Comparison of leaf variants results in good but somewhat lower concordance (>93% in HT115 and >77% in RPE1 cells) (lower tables). These results reflect the high accuracy of our call sets.

c) Comparing variant calls using different reference genomes and different variant callers. Venn diagram shows shared and unique counts from different reference algorithm combinations for branch and leaf variants calls: 1) mapping to hg19 reference genome with SNVs called by Strelka, 2) mapping to hg19 reference genome with SNVs called by MuTect1 and 3) mapping to hg38 reference genome with SNVs called by MuTect2. To enable the comparison all SNVs were remapped into hg38 coordinates, using the NCBI tool https://www.ncbi.nlm.nih.gov/genome/tools/remap (whenever the remap conversion tool returned multiple possible coordinates - when converting from hg19 coordinates to hg38 coordinates - the first-listed hg38 coordinate was selected).

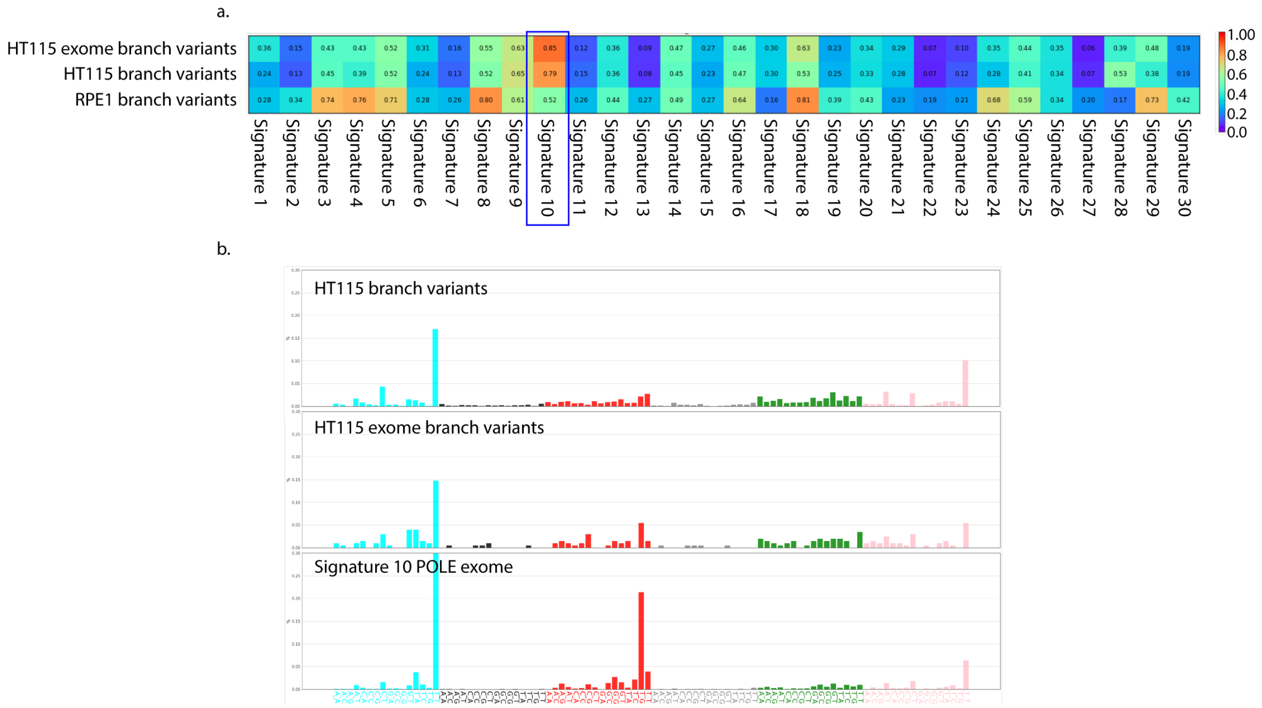


Figure S6: POLE signature of mutation process similar to HT115 mutation spectra

a) Heatmap comparing the cosine similarity scores between SNV mutation spectra of HT115 branch variants (whole-genome), the derived whole exome, and RPE1 whole genome branch variants with the COSMIC set of mutation signatures (derived from whole-exome sequencing) (Forbes et al. 2015). The blue rectangle denotes the most similar signature (number 10, *POLE*) for HT115, with the *POLE* exome signature showing cosine similarity = 0.79 (+/-0.02) to the HT115 whole genome branch variant data and 0.85 (+/-0.04) to the HT115 whole exome branch variant data.

b) Comparison of the full mutation spectra of all base substitutions observed in HT115 whole genome branch variants, the derived whole exome, and the *POLE* signature.

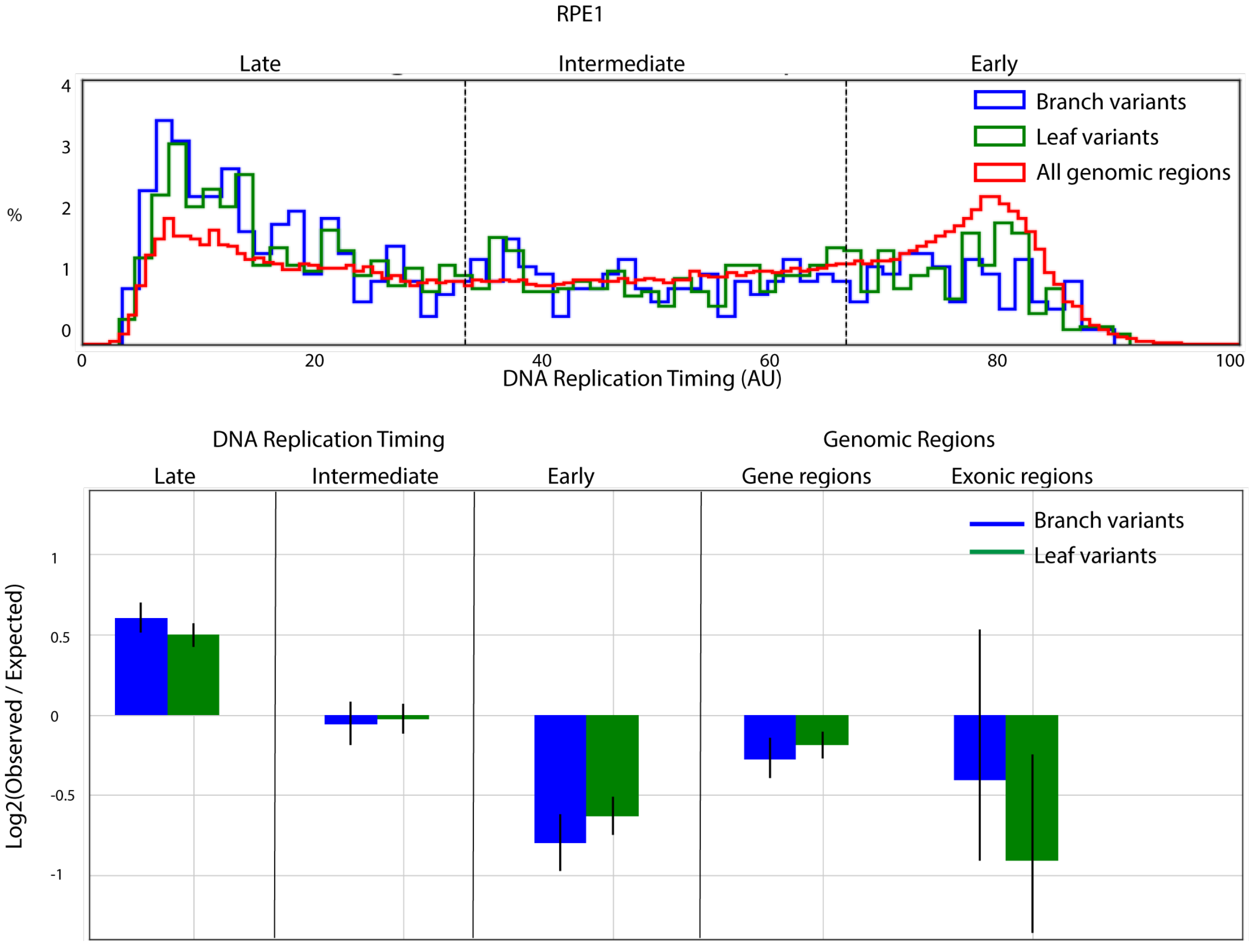


Figure S7: RPE1 cells show unequal distribution of mutation rates across the genome.

a) Distribution of the DNA replication timings for RPE1 branch variants and leaf variants (blue and green respectively) for early- and late-replicating regions ($p < 0.01$ calculated by binomial test of observed vs expected ratio).

b) Enrichment and depletion of SNVs in the indicated genomic regions, assessed as the $\log_2$ ratio versus the genome-wide average. Mutations are enriched in the late-replicating regions and depleted in the early replicating regions. Genomic regions (RefSeq gene regions – exons introns and UTRs) and also exonic regions (exons only) show depletion ($p < 0.01$) in both branch and leaf variants. Error bars were calculated by bootstrapping the data 10^4 times and represent 95% confidence intervals.

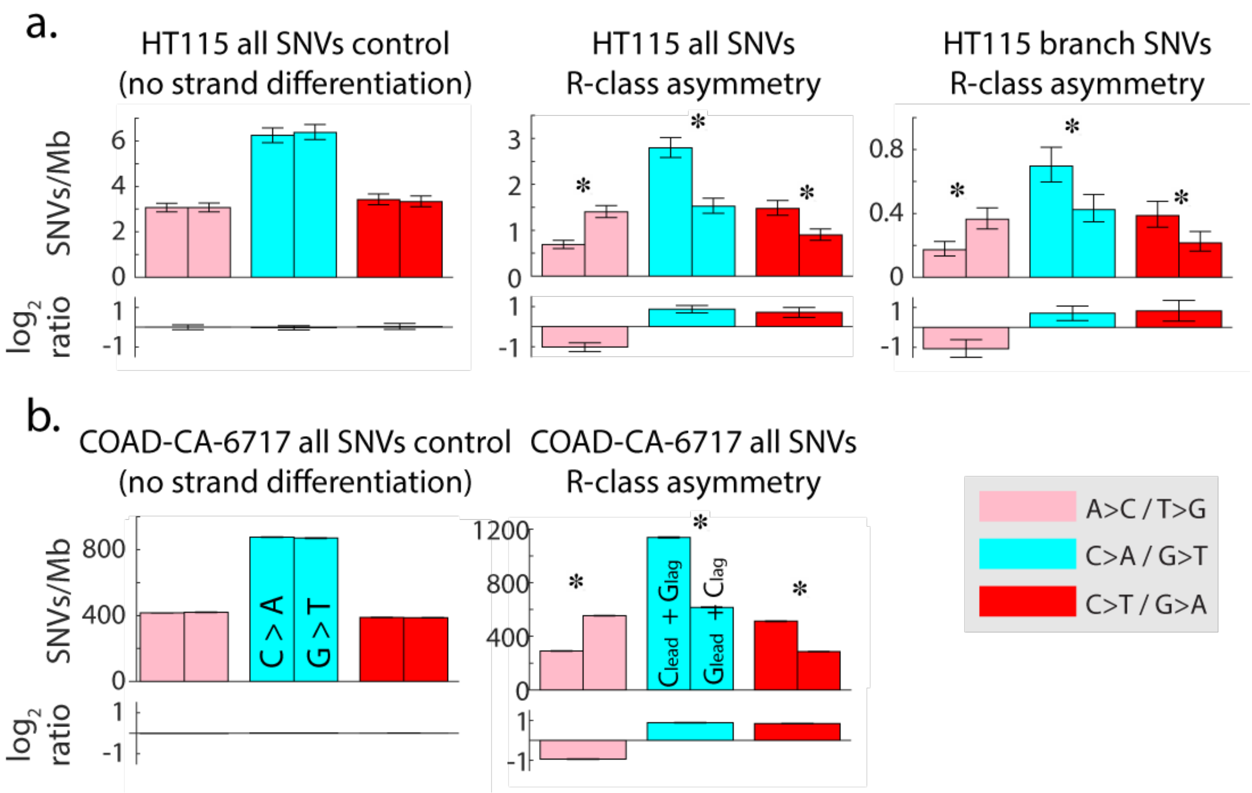


Figure S8: POLE strand asymmetries are evident in our samples:

We detected strand asymmetry associated with DNA replication (“R-class” asymmetry) in the HT115 and COAD-CA-6717 TCGA variant sets. The left plots represent the controls calculated by summing all mutations (with no differentiation of the replicated strand) and show a lack of asymmetry when the orientation of DNA replication is ignored ($P<<0.01$). R-class asymmetry consists of a skew based on the direction of DNA replication. For example, a skew toward C>A mutations in left-replicating regions where the strand synthesized by leading strand synthesis is the genomic reference. In right-replicating regions, a skew toward G>T mutations is seen where the genomic reference strand is synthesized by lagging strand synthesis. For each of three substitution groups (A>C/T>G, C>A/G>T, C>T/G>A) there are four options, for example 1) C>A in left-replicating regions 2) C>A in right-replicating regions 3) G>T in left-replicating regions and 4) G>T in right replicating regions. Left bars in each pair of bars in the top portion of plots (SNVs/Mb) show the sum of options (1) and (4) since they are both C>A with respect to the leading strand template. The right bars show the sum of (2) and (3) (since they are both G>T with respect to the leading strand template). The log$_2$ ratios of these paired bars are shown in the bottom portion of each plot and represent the effect size of the enrichment or depletion for each substitution as assessed by the R-class asymmetry. Significant R-class asymmetry signal was observed in the total HT115 SNV set, the HT115 branch SNV set, and the COAD-CA-6717 TCGA sample SNV set.

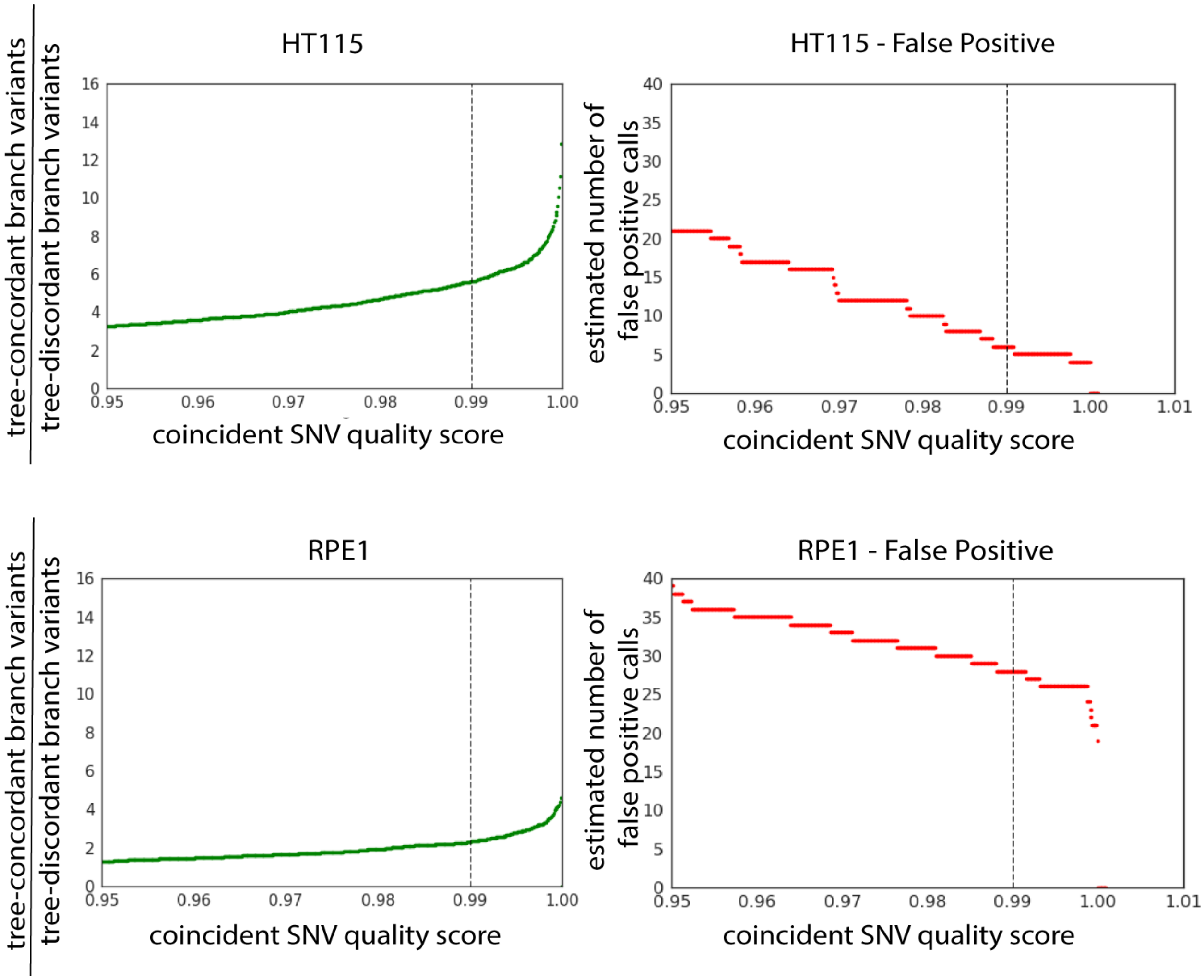


Figure S9: Lineage sequencing facilitates data quality testing without an external standard

Branch variants are called from the set of SNVs that coincide at the identical genomic locus across multiple sub clones. Left plots: ratio of called branch variants consistent with consensus lineage structure to all other branch variants that are not consistent with the tree structure for the HT115 dataset (top) and the RPE1 dataset (bottom). The ratio increases as the coincident quality score cutoff is made more stringent (a value of 0.99 was applied to call branch variants in both datasets). Right plots: Estimating false positive branch variant counts for the HT115 and RPE1 datasets by counting the "branch variants" in the sub clone label-scrambled datasets as a function of the branch variant quality score threshold.

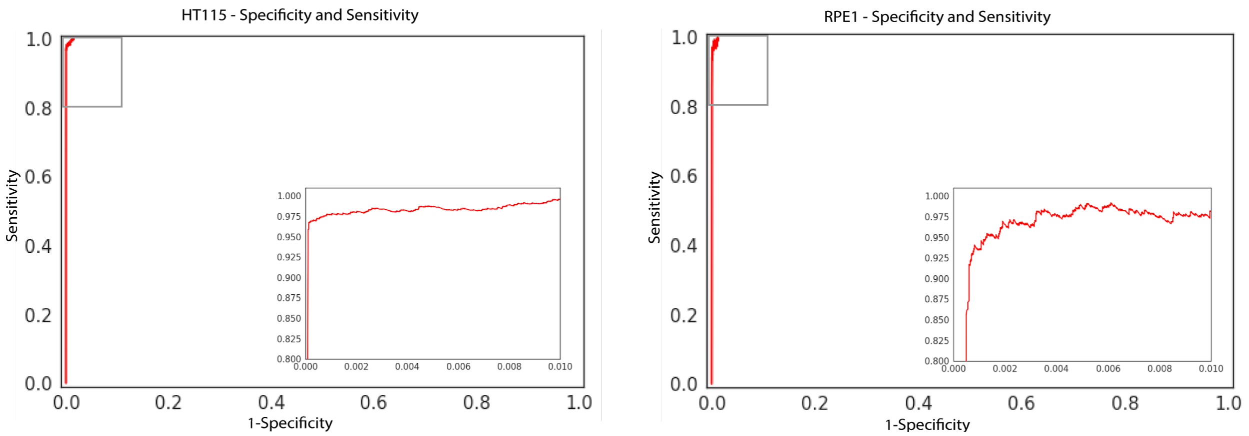


Figure S10: Estimation of specificity and sensitivity in our branch SNV calls

We estimated the sensitivity and specificity of branch variant calls for HT115 and RPE1 lineage sequencing. By using the ‘raw variants -> lineage -> called variants’ approach together with the microscopic derived lineage structure we were able to estimate the sensitivity and specificity of the lineage sequencing method. False positive calls were estimated by counting branch variant SNVs that appear as branch variants after scrambling the data, and false negative calls were estimated by counting branch variants with lower than threshold quality scores minus the expected number of false positive calls at each threshold value (methods).

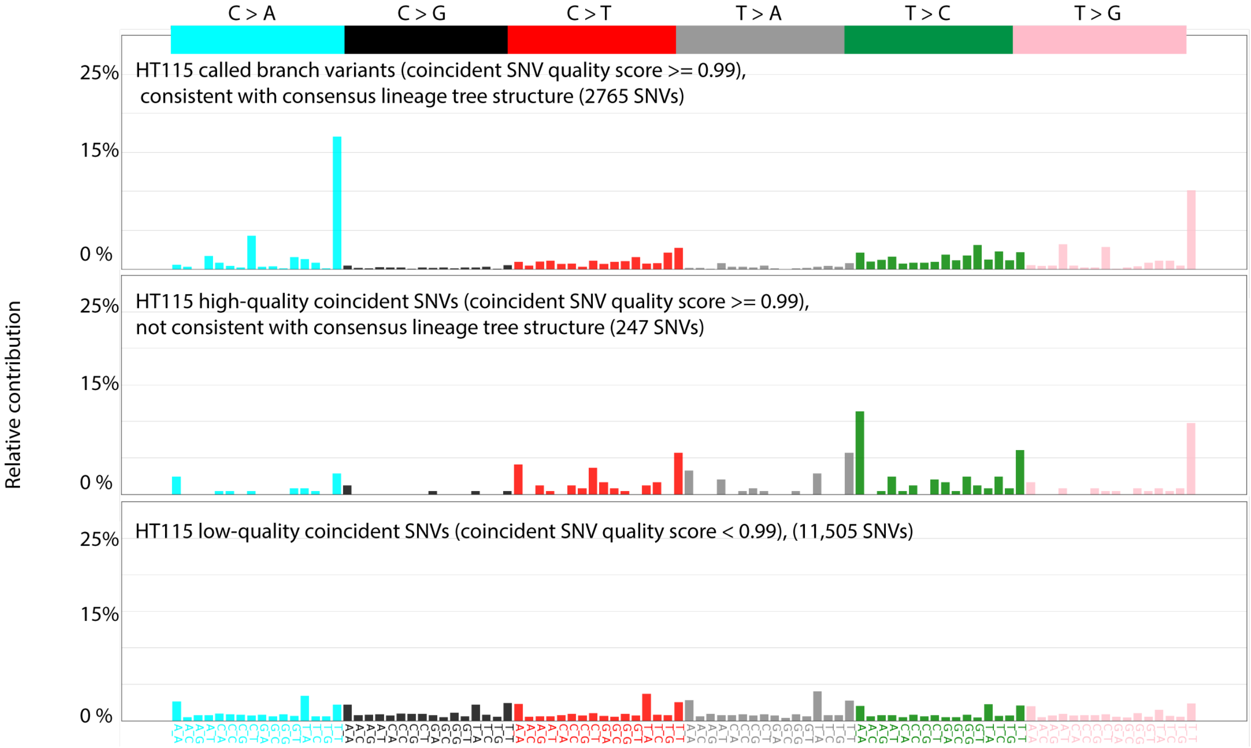


Figure S11: mutation spectra of filtered coincident SNVs differ from spectrum of called branch variants.

HT115 called branch variants (top plot, coincident SNV quality score >= 0.99 and consistent with the prior lineage estimate) show a mutation spectrum that matches reference *POLE* mutant whole-genome spectrum (not shown here). This called branch variant spectrum differs from the mutation spectrum of high quality coincident SNVs (middle plot, quality score < 0.99 not consistent with prior lineage estimate) and also the mutation spectrum of low-quality coincident variants (bottom plot, quality score < 0.99).

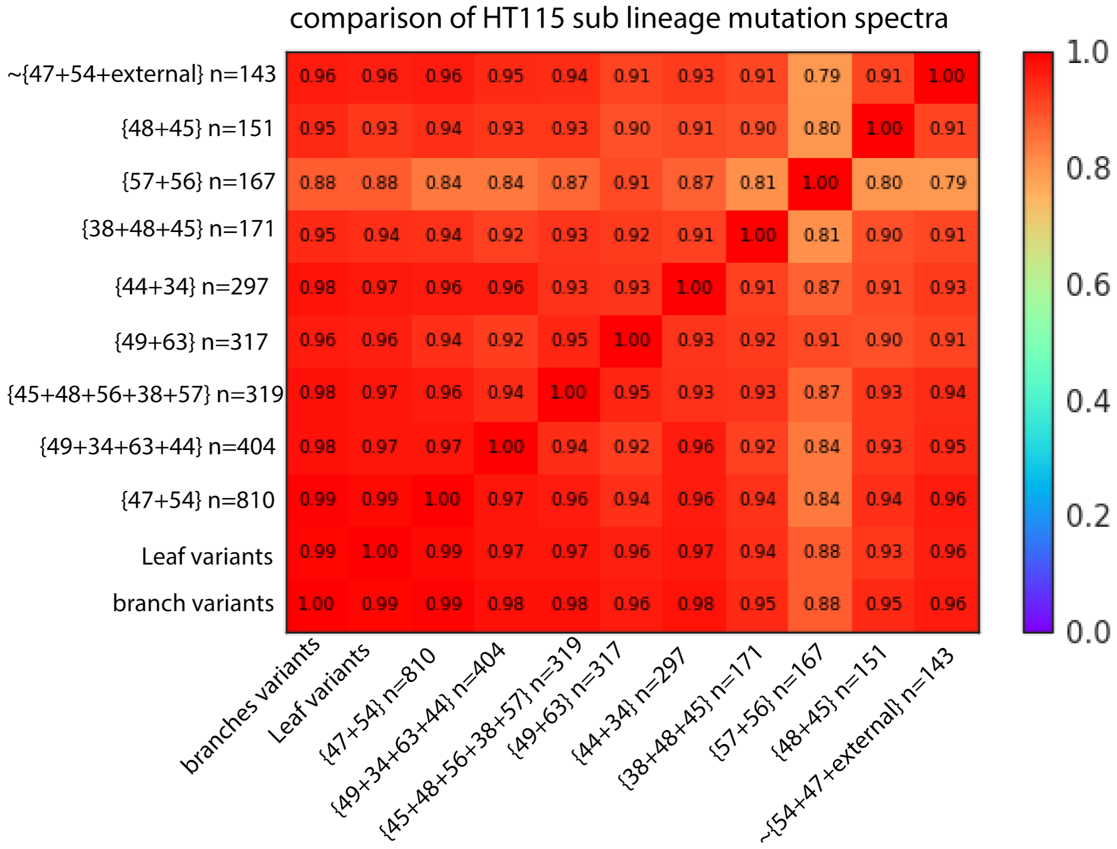


Figure S12: HT115 lineages show similar branch variant mutation spectra

Heatmap showing the cosine similarity score matrix across the mutation spectra of each sublineage in the HT115 experiment (which has high enough variant counts to support this analysis) and the complete branch and leaf variant call sets. The high similarity across the score matrix indicates the high specificity of these data and the consistency of the mutation spectrum across the HT115 lineage.

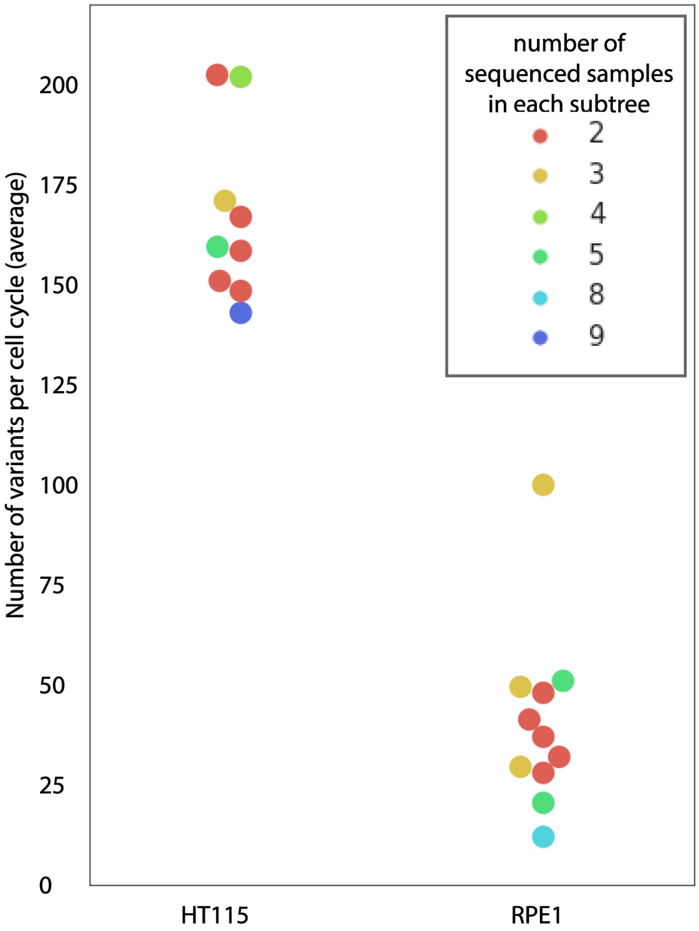
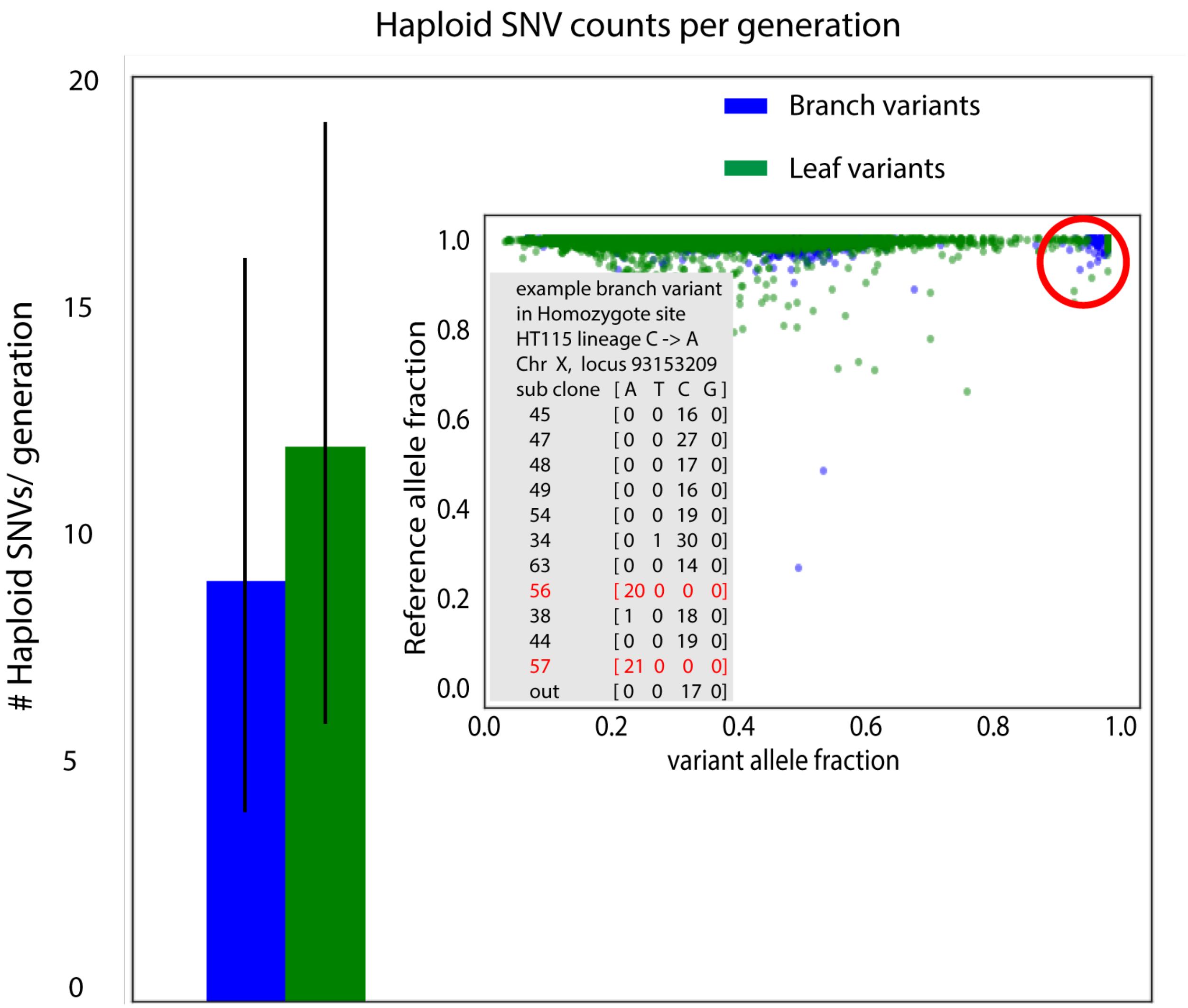


Figure S13: SNV detection sensitivity is not detectably affected by lineage position

The estimated number of branch variants per cell cycle is shown for every lineage segment in the two datasets, colored according to the number of sub clones exhibiting the variant. Variants “high” in the lineage anchor larger subtrees containing more variant sub clones. The lack of apparent bias in branch variant count according to subtree position supports the idea that we have comparable sensitivity for detecting branch SNVs across different lineage segments.



example branch variant
in Homozygote site
HT115 lineage C -> A
Chr X, locus 93153209

sub clone	[A T C G]
45	[0 0 16 0]
47	[0 0 27 0]
48	[0 0 17 0]
49	[0 0 16 0]
54	[0 0 19 0]
34	[0 1 30 0]
63	[0 0 14 0]
56	[20 0 0 0]
38	[1 0 18 0]
44	[0 0 19 0]
57	[21 0 0 0]
out	[0 0 17 0]

Figure S14: Mutation rate at haploid loci can be measured using leaf variants

Branch variants in the predominantly diploid HT115 cell line cluster at alternative allele fractions 0.5 and 1.0. Loci with both reference allele fraction and alternative allele fraction near 1.0 represent haploid sites. At these sites (red circle), there is enhanced confidence that the branch variants *and the leaf variants* are highly enriched for true variants. This is because the range of noise sources that produce errors in the leaf variant call set are unlikely to produce such a high alternative allele fraction. An example of one such specific leaf variant at one site on the haploid X chromosome with both reference allele fraction and alternative allele fraction near 1.0 is shown. This example shows a C to A mutation that appears only in sub clones 56 and 57. The accuracy of this leaf variant subset enables estimation of the mutation rate at haploid sites using leaf variants despite the fact leaf variants do not benefit from the suppression of false-positive variant calls in the branch variant call set resulting from the use of a prior lineage estimate. The number of branch and leaf variant SNVs is shown in the bar plot (error bars represent 95% confidence intervals calculated using the Poisson-distributed count approach).

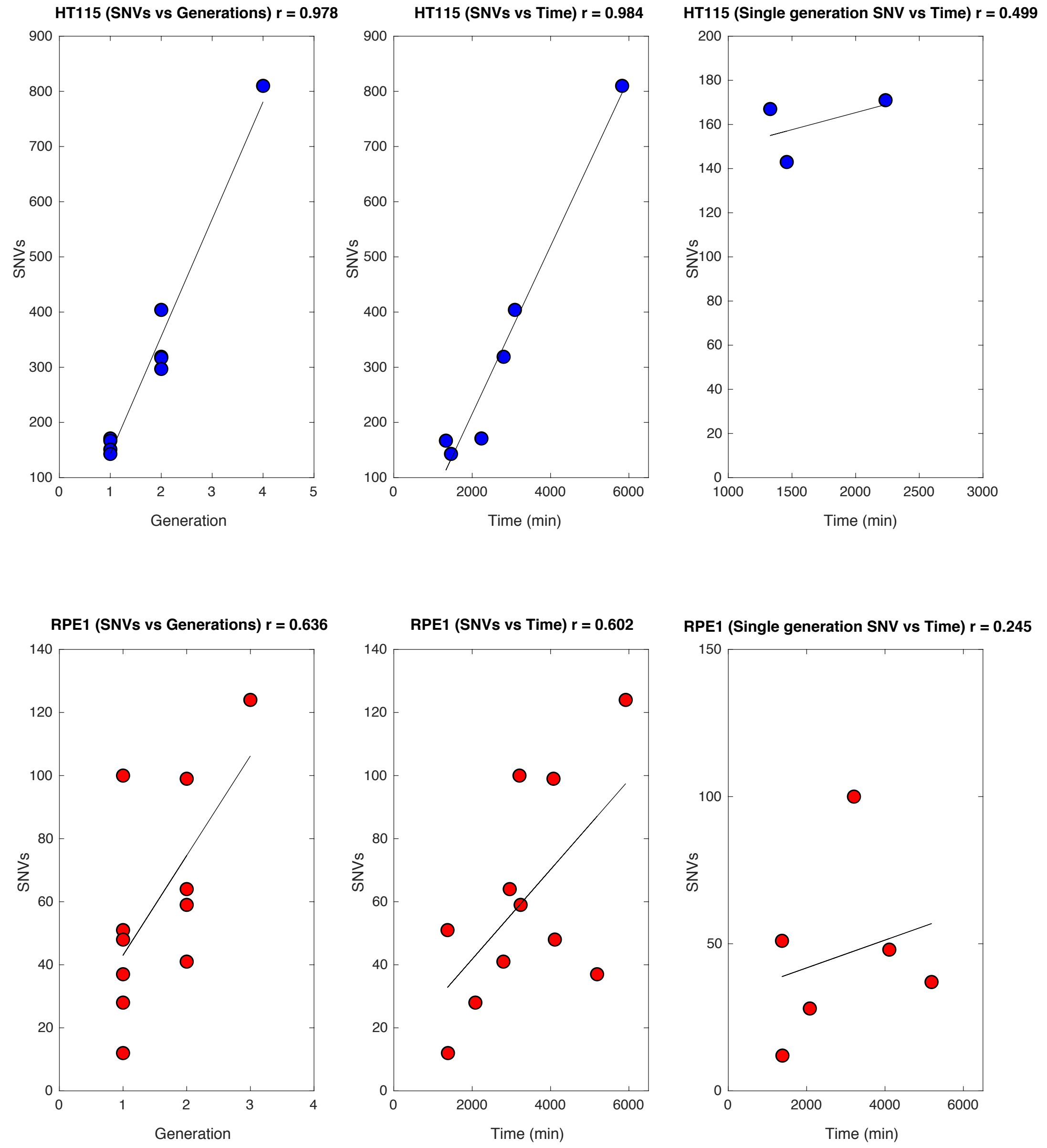


Figure S15: Correlation of mutation count versus elapsed cell divisions and time to division

Upper row presents results for HT115 cells; lower row presents results for RPE1 cells. Left, branch variant SNVs versus number of generations (or cell divisions); center, versus time for all non-overlapping lineage segments; right, versus time for single generation segments only. RPE1 (lower line) shows lower mutation counts and correlation coefficients for mutation counts versus time to division and generation number: ((ρ(mut,generations) = 0.636) and for times ((ρ(mut,times) =0.602). HT115 cells have higher mutation counts and better-synchronized replication, resulting in higher correlation coefficients versus time to division and generation number: (ρ(mut,generations) = 0.978), (ρ(mut,times) = 0.984).

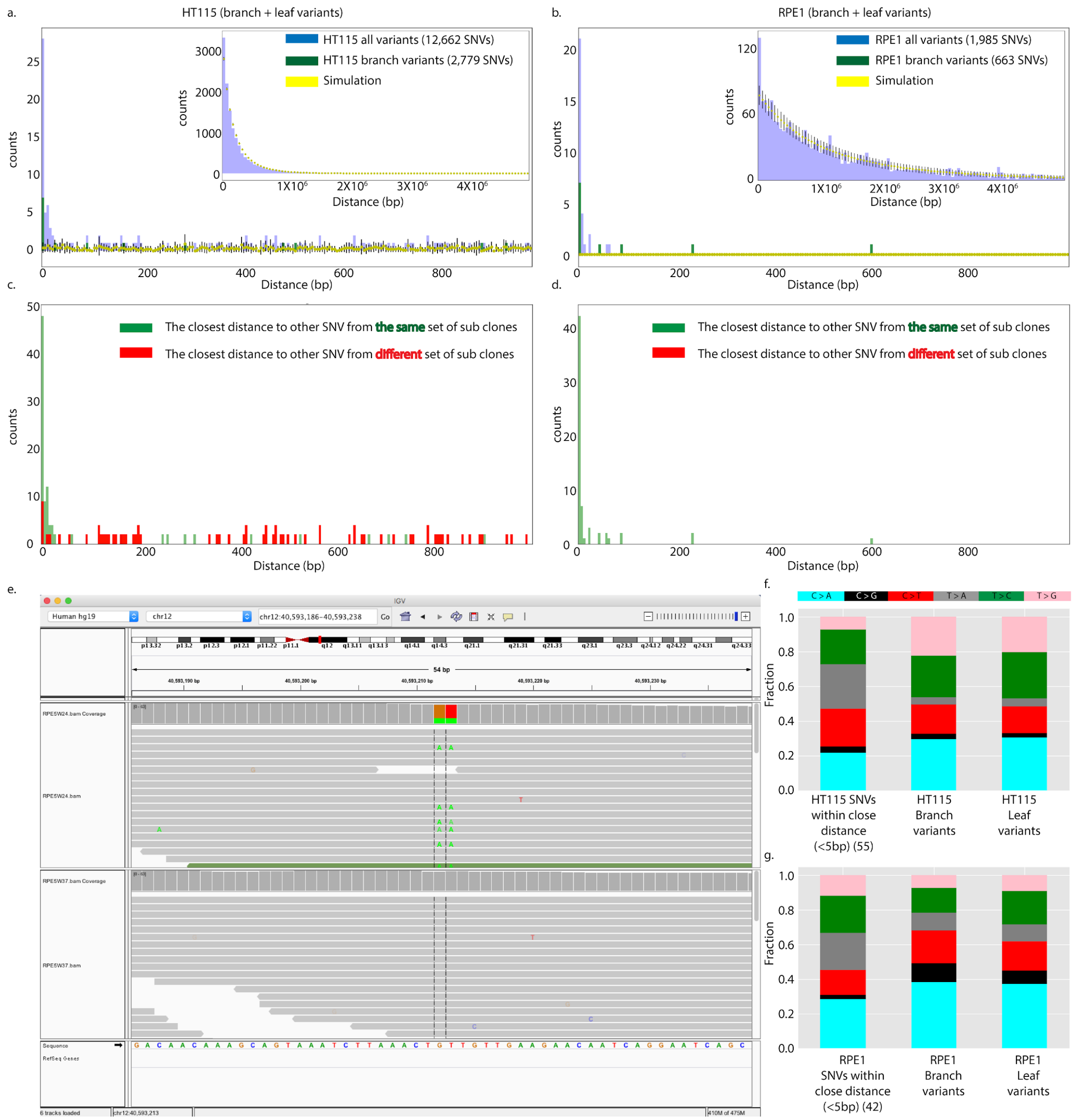


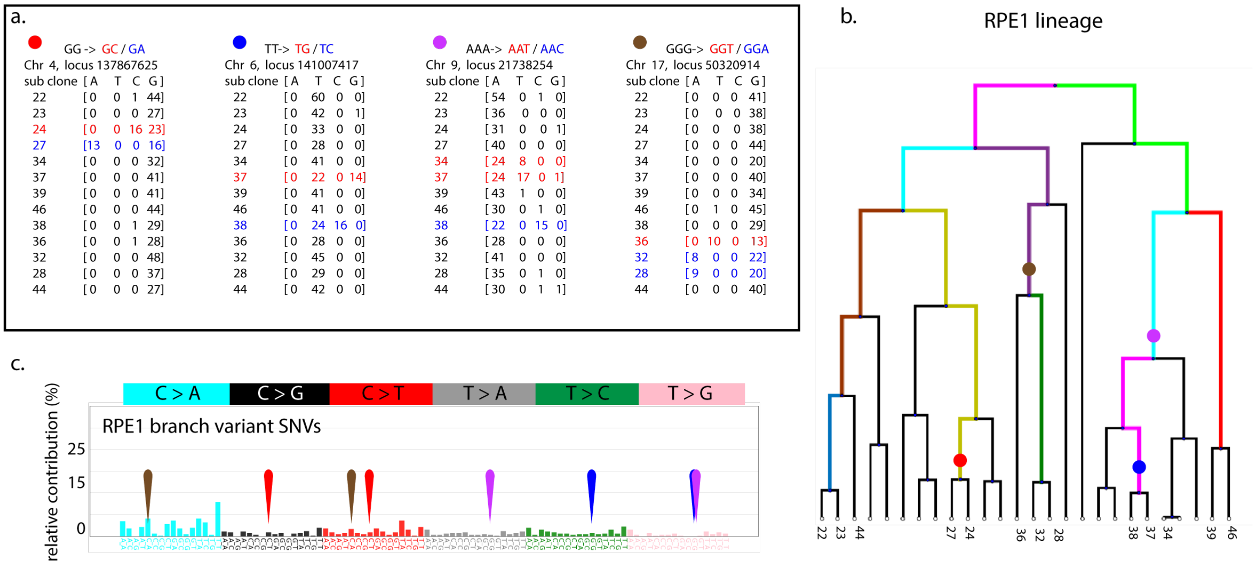
Figure S16: multi-nucleotide variants

a+b) Distribution of the observed distances between all detected variants (branch and leaf variants, in blue) or only branch variants (in green) compared with simulation of the expected distribution of distances for independent (uniform) mutations (simulated 10^3 times; error bars represent standard deviation across simulation runs). Main plot shows short distance scales (every bar represents 5bp) where strong enrichment in variants (total and branch variants only) less than 25 bp apart is evident versus the simulation. Similar enrichment was detected in HT115 (a) and RPE1 (b). Insets show expanded view (every bar represents 50,000 bp) of the distribution of distances where observed values closely match simulated values on longer length scales.

c+d) The closest distance between every variant was measured and plotted with small scale (every bar is 5bp), revealing a higher fraction of closely linked variants that share a sub lineage (green). Fewer closely linked variants are observed that span different sub lineages (red). This suggests that this relation between variants are linked not only in the genome, but also in time.

e) Screenshot from Interactive Genomics Viewer (IGV), showing one example from single generation resolved (one cell cycle) sub-lineage, suggesting that this multi-nucleotide variant mutation occurred during a single cell cycle.

f+g) Multi-nucleotide variants show a different mutation spectrum than the branch and leaf variant call sets, suggesting the operation of a different mutation process.



● GG -> GC / GA

Chr 4, locus 137867625

sub clone	[A	T	C	G]
22	[0	0	1	44]
23	[0	0	0	27]
24	[0	0	16	23]
27	[13	0	0	16]
34	[0	0	0	32]
37	[0	0	0	41]
39	[0	0	0	41]
46	[0	0	0	44]
38	[0	0	1	29]
36	[0	0	1	28]
32	[0	0	0	48]
28	[0	0	0	37]
44	[0	0	0	27]

● TT-> TG / TC

Chr 6, locus 141007417

sub clone	[A	T	C	G]
22	[0	60	0	0]
23	[0	42	0	1]
24	[0	33	0	0]
27	[0	28	0	0]
34	[0	41	0	0]
37	[0	22	0	14]
39	[0	41	0	0]
46	[0	41	0	0]
38	[0	24	16	0]
36	[0	28	0	0]
32	[0	45	0	0]
28	[0	29	0	0]
44	[0	42	0	0]

● AAA-> AAT / AAC

Chr 9, locus 21738254

sub clone	[A	T	C	G]
22	[54	0	1	0]
23	[36	0	0	0]
24	[31	0	0	1]
27	[40	0	0	0]
34	[24	8	0	0]
37	[24	17	0	1]
39	[43	1	0	0]
46	[30	0	1	0]
38	[22	0	15	0]
36	[28	0	0	0]
32	[41	0	0	0]
28	[35	0	1	0]
44	[30	0	1	1]

● GGG-> GGT / GGA

Chr 17, locus 50320914

sub clone	[A	T	C	G]
22	[0	0	0	41]
23	[0	0	0	38]
24	[0	0	0	38]
27	[0	0	0	44]
34	[0	0	0	20]
37	[0	0	0	40]
39	[0	0	0	34]
46	[0	1	0	45]
38	[0	0	0	29]
36	[0	10	0	13]
32	[8	0	0	22]
28	[9	0	0	20]
44	[0	0	0	40]

Figure S17: multiple mutation events, RPE1 cells.

- **a)** The four multiple mutation events detected in RPE1 cells: the genomic location of each SNV and allele counts. The colored dot on the upper left side corresponds to the colored markers in parts b) and c).
- **b)** The lineage segment for each SNV pair is indicated on the lineage tree structure with resolved segments shown.
- **c)** Plot marking spectral location of RPE1 multiple mutation events in the context of our overall mutation spectrum for RPE1.

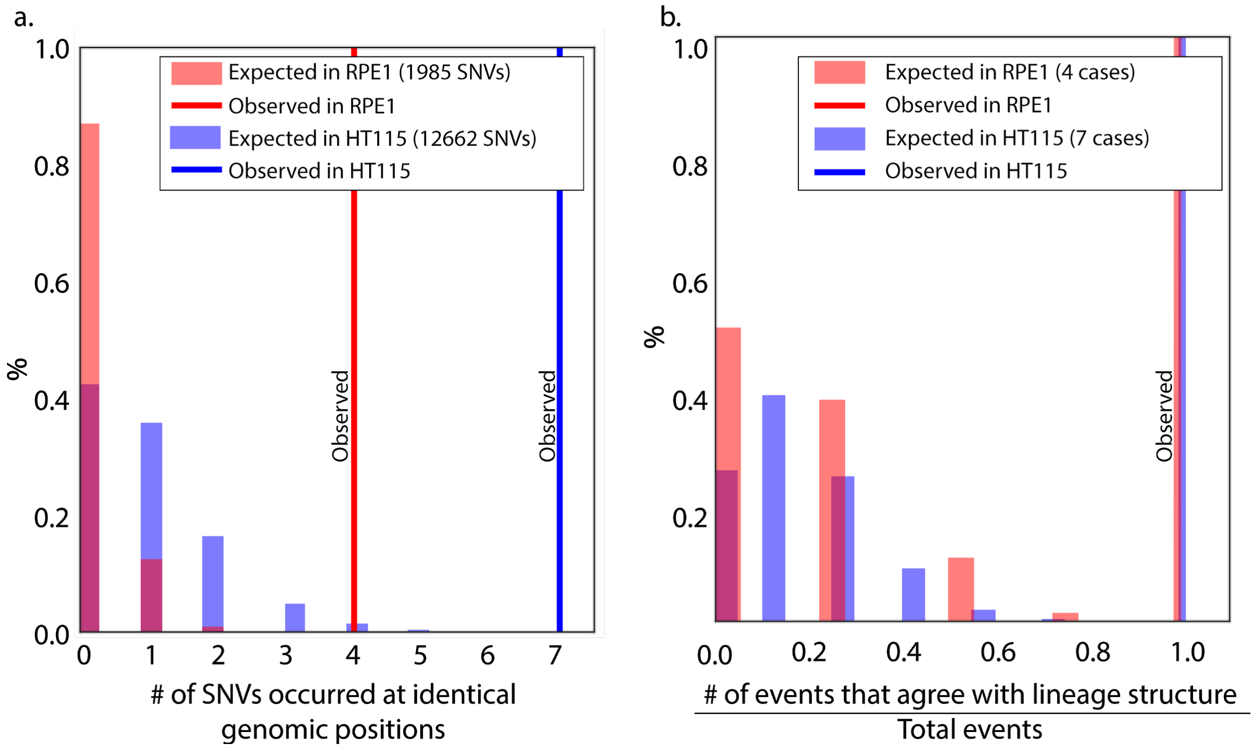


Figure S18: Coincident mutations are dependent events

a) In our dataset, it is unlikely that the observed sets of two different variants coincide at the same genomic locus by chance. By using the calculated probability to get two different variants coinciding at a given genomic locus by chance (for HT115; $p < 6.9\times10^{-5}$), we resampled the data 10^5 times to calculate the expected number of coincident mutations. The histogram of expected number of coincident mutations in HT115 (red) out of total number of 12,662 SNVs the branch and leaf variants, and in RPE1 data set (blue). In comparison with the observed number of coincident mutations, $p<1\times10^{-5}$ (zero events out of 10^5) for the lineage experiments with both cell lines.

b) We observed that all the detected pairs of different SNVs which occurred at identical genomic positions were found in related sub-lineages that agreed with the lineage structure (marked as 1.0 in the X-axis). We tested the probability that a random set of variant pairs would occur in related sub-lineages. We simulated two independent events at a time using the observed probability of finding mutations in each specific subset of samples. We simulated 5×10^5 instances and for each chose randomly two possible branch or leaf variants (4 times for RPE1, and 7 times for HT115). For every set of pairs, we measured the fraction of pairs that were congruent with the lineage structure and plotted the resulting distribution of observed fractions for RPE1 (red) and for HT115 (blue). In the simulation, three out of half million cases gave this result for HT115, indicating a p-value less than ($P<6\times10^{-6}$; 7/7 events). Significance was similarly achieved for RPE1 ($P<6.6\times10^{-4}$; 4/4 events).

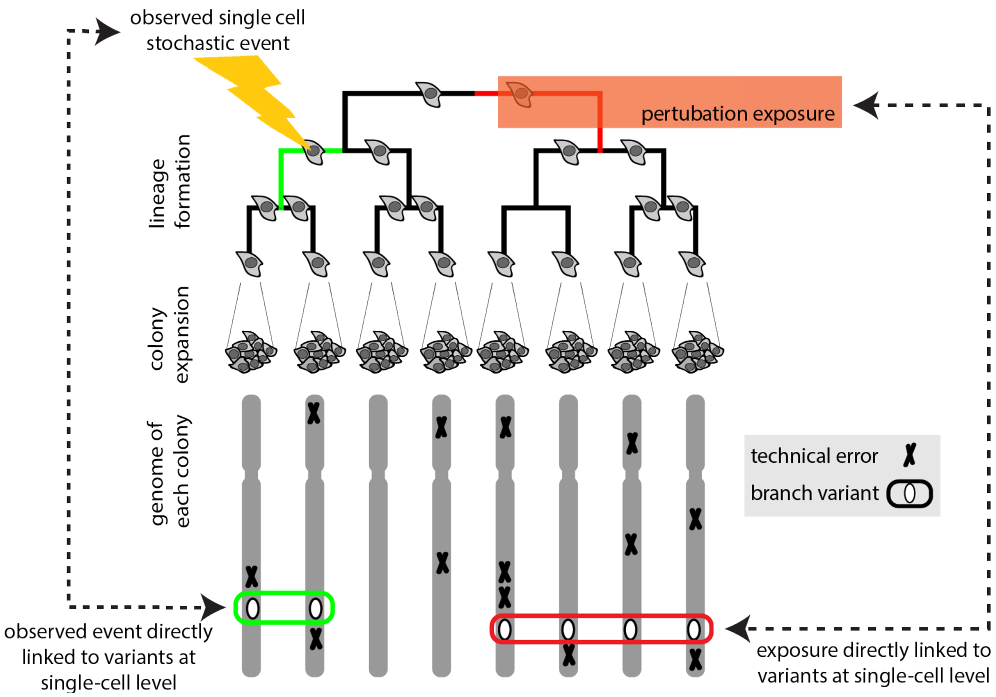

observed single cell stochastic event
pertubation exposure
lineage formation
colony expansion
genome of each colony
technical error
branch variant
observed event directly linked to variants at single-cell level
exposure directly linked to variants at single-cell level

Figure S19: Scheme for directly associating single-cell events and exposures to single-cell sequence variation

When the lineage is cultured under observation specific events occurring in single cells can be noted. Similarly, if particular cells in the lineage are exposed to perturbations such as chemical exposures, pathogens, radiation, or different culture conditions, the specific cells perturbed can be noted. Because lineage sequencing can assign novel somatic variants to lineage segments that correspond to individual cell lifetimes, we expect it will be possible to directly associate observed single-cell phenotypes and exposures with single-cell resolved genomic variants using the lineage sequencing approach.

Supplemental files Legends:

Supplemental Movie S1: ‘Supplement_Movie_S1.mp4’

The movie presents the cell handling steps in our implementation of ‘Lineage Sequencing’ (as described in figure S1). 1) Seeding the cells into the device: the movie presents a diagram of the complete system with four trap arrays (Kimmerling et al. 2016), input and output ports, tubes and vials. Each tube is controlled by air pressure that creates pressure gradients along the channels and drives fluid flow through the device. The animation shows how changes in pressure gradients change the flow in the system. A single cell was loaded into each lane of the trap array. 2) The bypass channels were then flushed with conditioned cell growth media and the cells were pushed to the opposite side of the trap array with the pressures set for long term culture such that conditioned media was continuously perfused along the bypass channels. 3) The cells were cultured for multiple generations on-chip under continuous observation. Several generations of HT115 cell population growth from a single founding cell are depicted in the microfluidic trap array. Imaging occurred in time-lapse every 3 min, while trypsinization and re-capture stages were continually imaged at 24 Hz. 4) Example showing cell release (trypsin) and manipulation of individual cells to the bypass channel with flow direction reversal (P3 > P2) until a cell reached the bypass channel. Then, flow into the traps was re-established (P2 > P3) to ensure that no other cells were released. Each cell was retrieved by 30 seconds of bypass flow to ensure it was flushed from the device. Released cells were collected separately and sub-cultured.

Supplemental Table S1: TCGA POLE mutant colorectal and uterine corpus endometrial cancer sample annotations.

This study examines the cosine similarity between HT115 branch variants mutation spectra with mutation spectra derived from tumor-normal whole genome sequencing of patients with previously identified POLE proofreading deficiency generated by the TCGA Research Network. Cosine similarity was calculated with 95% intervals by bootstrapping the HT115 spectra 10^4 times. This table lists the germline coding mutations in the POLE gene associated with proofreading defects.

Supplemental Table S2:

This table lists the complete SNV lists calculated by the ‘optical tracking -> lineage -> called variants’ approach. For HT115 lineage branch variants (2,779 SNVs) and leaf variants (9,883 SNVs). Coordinates are in hg19.

Supplemental Table S3:

The complete SNV lists calculated by the ‘optical tracking -> lineage -> called variants’ approach. For RPE1 lineage, branch variants (663 SNVs) and leaf variants (1,322 SNVs). Coordinates are in hg19.

Supplemental Table S4:

DNA replication timing for HT115 cell line. Mat file contains the raw (unsmoothed) and the smoothed data, coordinates are in hg19, and DNA replication timing values.

Supplemental File S5:

The custom python notebook scripts for analyzing 'Lineage sequencing'. The scripts were written in Python 2.7 on Jupyter notebook, and demonstrate step-by-step analysis of the data. Available also in GitHub https://github.com/yehudabrody/Lineage-sequencing---proof-of-concept.