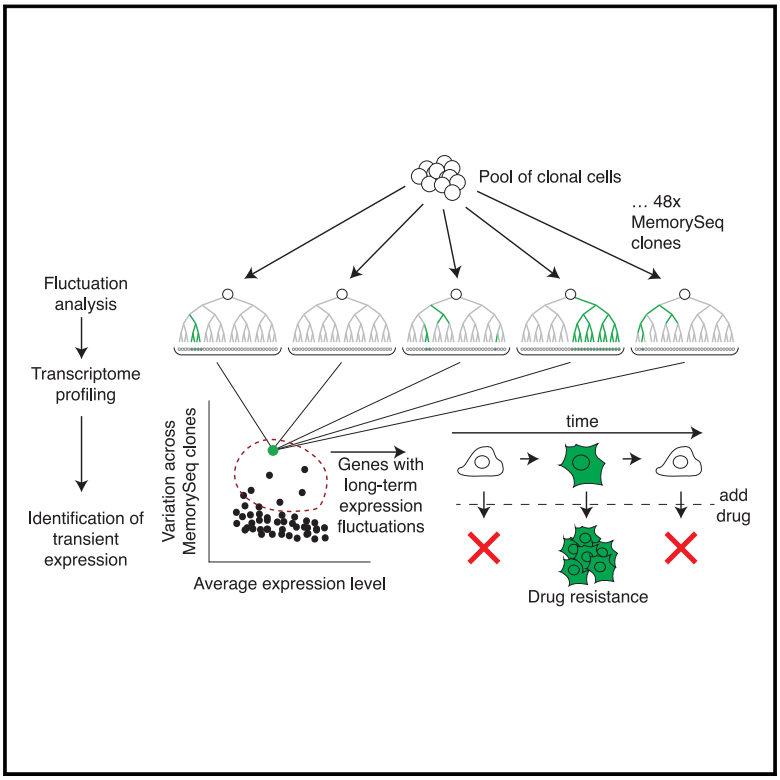
Cell

Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors

Graphical Abstract



Authors

Sydney M. Shaffer, Benjamin L. Emert, Raúl A. Reyes Hueros, ..., Abhyudai Singh, Danielle S. Bassett, Arjun Raj

Correspondence

arjunrajlab@gmail.com

In Brief

MemorySeq shows that in single cells, the expression pattern of some genes co-fluctuates slowly and is heritable, leading to subpopulations of cells with distinct behaviors, such as resistance to cancer drugs.

Highlights

- MemorySeq identifies slowly fluctuating gene expression states in rare single cells
- Slowly fluctuating states are associated with drug resistance in cancer
- MemorySeq reveals broad patterns of gene co-expression in rare cells

Shaffer et al., 2020, Cell *182*, 1–13
August 20, 2020
https://doi.org/10.1016/j.cell.2020.07.003

CellPress

Cell

CellPress

Article

Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors

Sydney M. Shaffer,[1,2,13] Benjamin L. Emert,[3,13] Raúl A. Reyes Hueros,[3,4] Christopher Cote,[2,12] Guillaume Harmange,[3,5] Dylan L. Schaff,[2] Ann E. Sizemore,[2] Rohit Gupte,[2] Eduardo Torre,[3,4] Abhyudai Singh,[6] Danielle S. Bassett,[2,7,8,9,10,11] and Arjun Raj[2,12,14,*]

[1]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, PA, USA
[3]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Biochemistry and Molecular Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5]Cell and Molecular Biology Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6]Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA
[7]Department of Physics and Astronomy, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
[8]Department of Electrical and Systems Engineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, PA, USA
[9]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[10]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[11]Santa Fe Institute, Santa Fe, NM, USA
[12]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[13]These authors contributed equally
[14]Lead Contact
*Correspondence: arjunrajlab@gmail.com
https://doi.org/10.1016/j.cell.2020.07.003

SUMMARY

Non-genetic factors can cause individual cells to fluctuate substantially in gene expression levels over time. It remains unclear whether these fluctuations can persist for much longer than the time of one cell division. Current methods for measuring gene expression in single cells mostly rely on single time point measurements, making the duration of gene expression fluctuations or cellular memory difficult to measure. Here, we combined Luria and Delbrück's fluctuation analysis with population-based RNA sequencing (MemorySeq) for identifying genes transcriptome-wide whose fluctuations persist for several divisions. MemorySeq revealed multiple gene modules that expressed together in rare cells within otherwise homogeneous clonal populations. These rare cell subpopulations were associated with biologically distinct behaviors like proliferation in the face of anti-cancer therapeutics. The identification of non-genetic, multigenerational fluctuations can reveal new forms of biological memory in single cells and suggests that non-genetic heritability of cellular state may be a quantitative property.

INTRODUCTION

Cellular memory in biology, meaning the persistence of a cellular or organismal state over time, occurs over a wide range of timescales and can be produced by a variety of mechanisms. Genetic differences are one form of memory (Ben-David et al., 2018), encoding variation between organisms on multi-generational timescales. Within an organism, mechanisms involving the regulation of gene expression encode the differences between cell types in different tissues, with cells retaining memory of their state over a large number of cell divisions (Bonasio et al., 2010). In contrast, recent measurements suggest that the expression of many genes in single cells may have very little memory, displaying highly transient fluctuations in transcription. These rapid fluctuations have been referred to as gene expression "noise" and have generally been difficult to associate with physiological distinctions between single cells (Raj and van Oudenaarden, 2008; Raj et al., 2006; Sigal et al., 2006; Symmons and Raj, 2016), although there are certainly specific examples in which such fluctuations can drive phenotype (Cohen et al., 2008; Raj et al., 2010; Spencer et al., 2009; Wernet et al., 2006).

Less well studied is memory on intermediate timescales (i.e., cellular states that may persist for several divisions but are ultimately transient) and thus are not indefinitely heritable (distinguished from the short-lived fluctuations referred to as "noise"). Such timescales would be long enough to allow for coordinated

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

fluctuations in the expression of many genes at once in individual cells, potentially resulting in biological activity within that cell that is distinct from the rest of the population. Yet it remains unclear how prevalent such longer-timescale fluctuations might be because finding the molecular markers of these longer fluctuations is difficult: current "snapshot" methods are unable to distinguish between fast and slow fluctuations because they lack any temporal component, while time lapse microscopy is laborious and difficult to scale to all genes (Hormoz et al., 2016; Phillips et al., 2019). Thus, we sought to develop a method that would enable us to find genes whose expression fluctuations would be maintained over several cell divisions. Ultimately, our goal was to use these markers of slow fluctuations to identify functionally distinct subpopulations within otherwise indistinguishable cells.

RESULTS

The methodology we developed to distinguish heritable from non-heritable fluctuations in expression levels in single cells (MemorySeq) is based on the fluctuation analysis from Luria and Delbrück's beautiful 1943 experiments on resistance to phage in bacteria, which they used to discriminate heritable from non-heritable mechanisms for resistance (Luria and Delbrück, 1943) (also used in cancer) (Shaffer et al., 2017; Tlsty et al., 1989). In our context of cellular memory, the experiment consisted of growing a number of "MemorySeq clones" (we aimed for 48 and ended up with 42–45 after losses from culture and library prep; see STAR Methods for details) of isogenic melanoma cells (WM989-A6) in individual wells, eventually growing them to around 100,000 cells per clone (Figures 1A and 1B). If a fluctuating gene transitioned in and out of the "high" expression state relatively rapidly compared to the cell division rate, then a fairly constant proportion of those 100,000 cells would be in the high expression state for that gene (with some dispersion due to Poisson sampling). This constancy occurs because the cells do not remember the state through cell division. At the opposite extreme, if the high expression state was long-lived compared to the cell division time, then if a cell occasionally moves into the high expression state early in the family tree, all of its progeny will remain in the high expression state, leading to a very high proportion of the final 100,000 cells being in the high expression state. Thus, across multiple MemorySeq clones, we would find a high variance in the proportion of cells in the high expression state in the final population, where most clones would have low expression of that gene and a few clones would have high expression, depending on exactly how far up in the family trees the cells transitioned into the high expression state. To measure variability in the proportion of cells in the high expression state for any particular gene, we used bulk RNA sequencing to measure the transcription of all genes in each expanded clone, and we then measured the variability in the expression of all genes across these clones (Figure 1B). It was important to distinguish between biological variability between clones and variability due to sequencing, sampling, and technical errors. We therefore also grew a large population of cells that we split into 48 wells containing 100,000 cells each and subjected those cells to RNA sequencing and gene expression analysis (Figure 1B, right).

We first applied MemorySeq to the melanoma cell line WM989-A6. We chose this cell line and culture system because we had already verified the presence of rare cells within the population marked by the expression of a particular subset of genes such as *EGFR*, *NGFR*, and *AXL*. These rare cells were strongly associated with resistance to the targeted melanoma drug vemurafenib (Fallahi-Sichani et al., 2017; Shaffer et al., 2017), and the independent observation that sibling cells often expressed the same genes suggested that these genes displayed some degree of memory (Shaffer et al., 2017). Thus, in this system, we have already identified several genes that are both associated with a phenotype and appear to exhibit some degree of heritability. These genes naturally serve as positive controls for the MemorySeq methodology.

Upon performing MemorySeq in this cell line, we first checked the distribution of expression levels across MemorySeq clones for a number of previously identified resistance-associated genes and non-resistance-associated genes (Figure 1C, heritable genes) (Shaffer et al., 2017). As hoped, we found that the resistance-associated genes displayed far greater variability across clones than the technical noise controls. Conversely, housekeeping genes and other genes that do not exhibit much cell-to-cell variability showed variances across MemorySeq clones that were much more similar to that of the technical noise controls (Figure 1C, non-heritable genes). Intriguingly, *MYC*, a proto-oncogene for which we have seen high levels of cell-to-cell variability, showed little increased variance across clones compared to controls, suggesting that its transcriptional memory was much lower than that observed for resistance-associated genes (Padovan-Merhar et al., 2015) (Figure 1C). We observed similar behavior for *CCNA2*, a cell-cycle gene whose expression would similarly be expected to vary from cell to cell but not exhibit much heritability due to cell cycle desynchronization (Figure 1C) (Chao et al., 2018).

Given that RNA sequencing provides expression levels across the transcriptome, MemorySeq is able to measure heritability in the expression of all genes at once. Thus, we analyzed expression variance across clones for all genes. We found that for many genes across a range of average transcription levels, variance across clones was much higher than technical noise controls, suggesting that those genes exhibited high levels of transcriptional memory (Figure 1D). We found that genes with higher expression levels typically had systematically lower variance across clones. By explicitly fitting this relationship, we could identify genes as high-memory based upon their large residuals from the fit. We generated a panel of high-memory genes with residuals in the 98th percentile or greater, and with a minimum expression level of 1.5 transcripts per million to eliminate spurious inclusion of lowly expressing genes, resulting in 227 genes identified as potentially having high heritability in WM989-A6 (alternative cutoffs and robustness analysis for this gene set provided in Figures S1A and S1B).

Our experimental design predicted that a gene whose expression exhibits high variability across MemorySeq clones would occasionally initiate high levels of expression that would persist across multiple cell divisions but would not persist indefinitely. We sought to directly confirm these expression characteristics by using time-lapse microscopy to trace the expression state

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

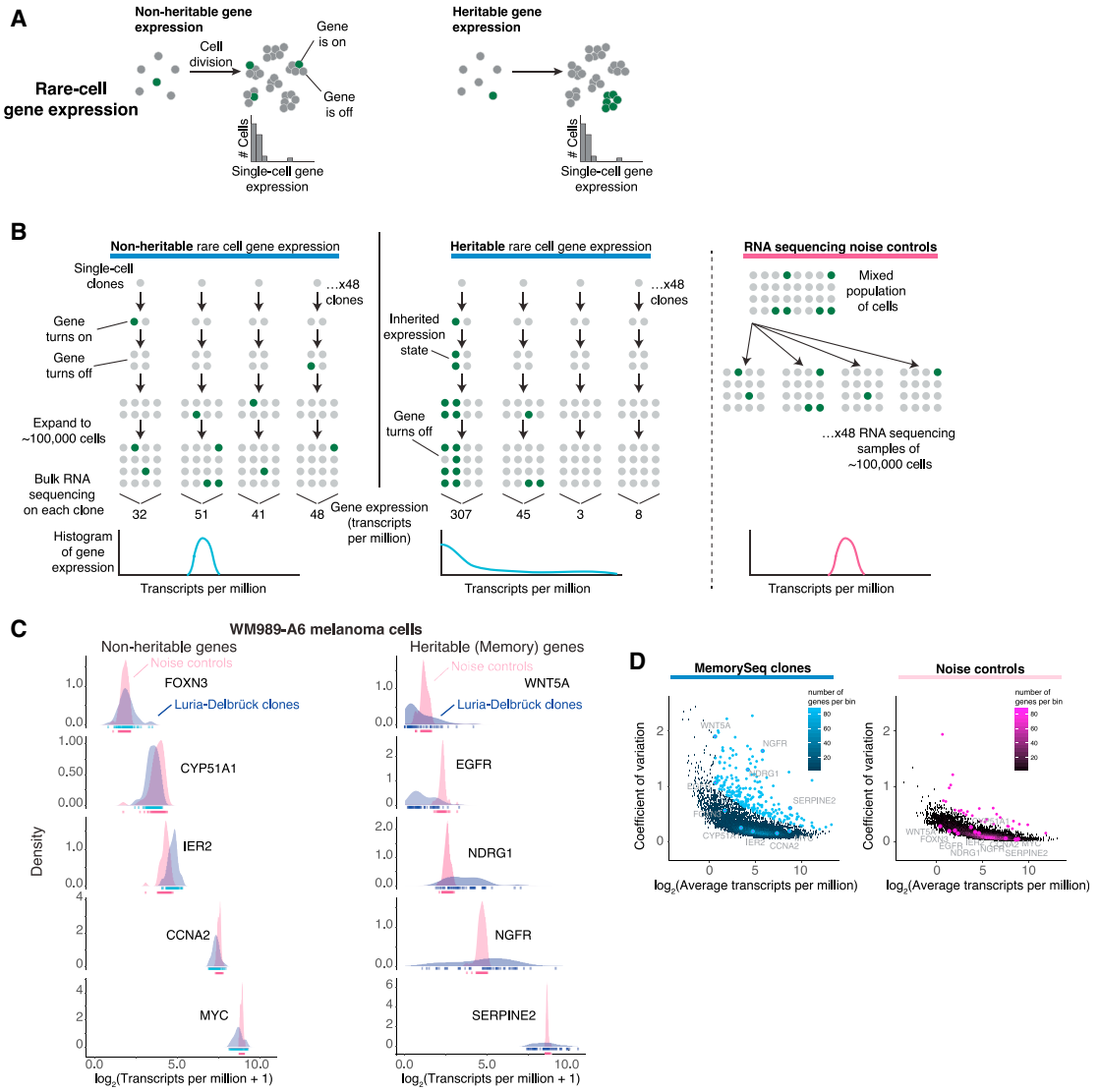


Figure 1. MemorySeq Can Identify Genes with High Transcriptional Memory

(A) Rare-cell gene expression patterns, both with and without heritability. Histograms of single-cell expression levels are unable to discriminate between these two alternatives.

(B) Schematic of MemorySeq experiment. We started with a single melanoma cell (WM989-A6), grew it to ~100 cells, then , from these, seeded single cells into 48 separate wells and allowed the cells to proliferate to around ~100,000 cells before subjecting the entire MemorySeq clone to RNA sequencing to determine expression levels. In the case of non-heritable expression, the levels of expression would not vary dramatically between MemorySeq clones, whereas in the heritable case, some clones would exhibit much higher levels of expression when a cell moved into the high expression level state early in the family tree of the clone. To determine how much variability in expression would arise for purely technical reasons, we also performed control experiments by plating around ~100,000 cells directly into individual wells and performing RNA sequencing.

(C) Expression histograms across n = 43 MemorySeq clones for genes identified as non-heritable (left) or heritable (right).

(D) Coefficient of variation versus mean expression levels for all 23,669 genes that we analyzed across all MemorySeq clones. Points labeled with blue dots (on MemorySeq clones plot) or pink dots (on noise controls plot) passed the threshold for being identified as a heritable gene. These genes were identified by first fitting a Poisson regression model to the data, and then selecting genes with residuals in the top 2%. This approach identified 227 heritable genes from the MemorySeq clones, but only 30 genes passed this threshold in the noise control condition. Particular genes from the panel in (C) are labeled on both plots. See also Figure S1.

of individual cells for three genes. For one gene (*NGFR*), we were able to genetically tag the gene using a split-fluorescent protein approach (see STAR Methods) to fuse mNeonGreen2(11) to *NGFR* (producing the NGFR-mNG2 protein). We used that cell line to track NGFR-mNG2 levels by fluorescence microscopy over a period of 8.75 days (Figure 2; see Figures S2A–S2D for cell line validation; single allele tagged, which may affect the levels of variability as compared to total protein). The vast majority of cells displayed essentially no fluorescent signal, but as predicted, occasional rare cells within the population displayed high

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

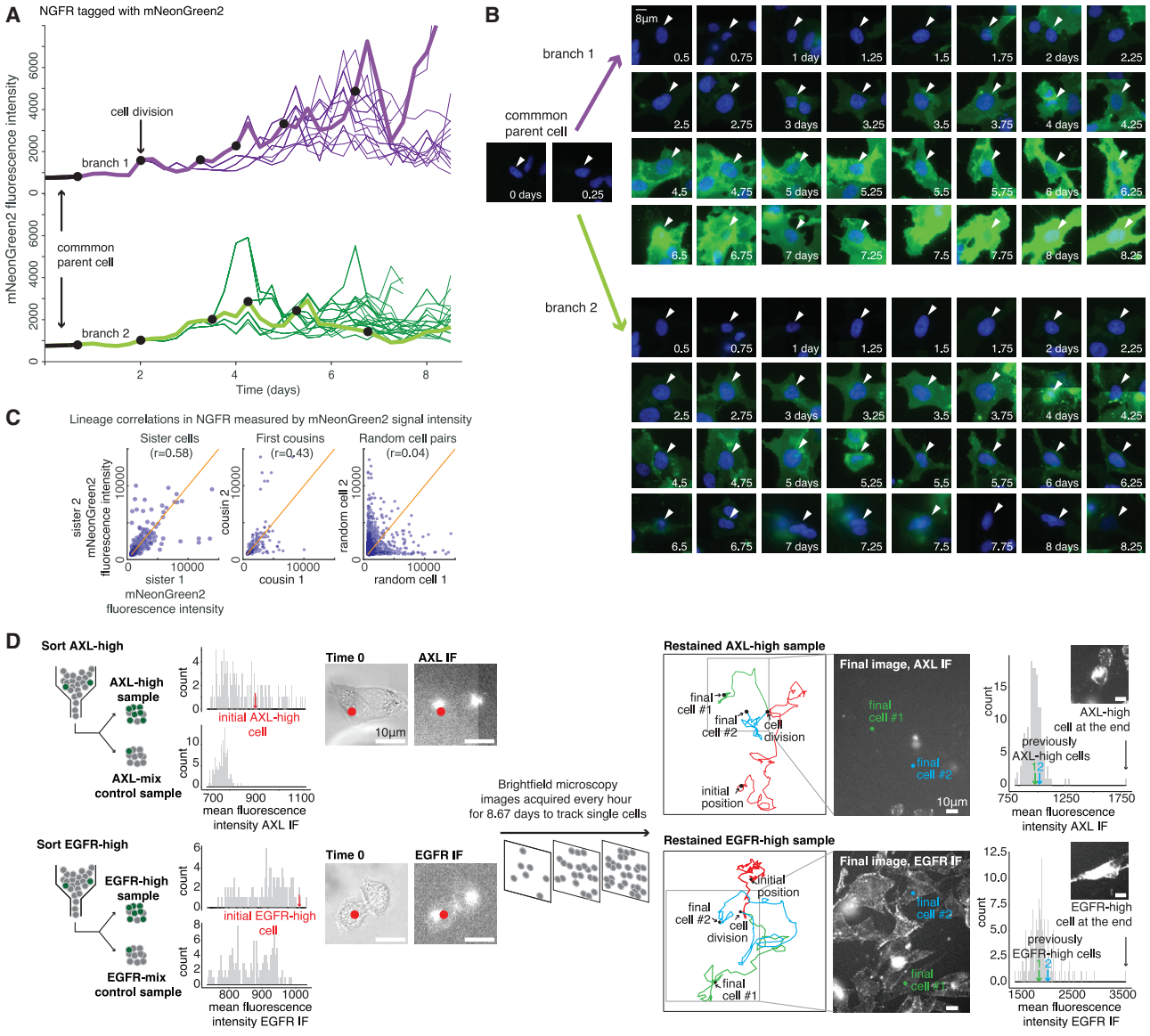


Figure 2. Time-Lapse Microscopy Verifies Rare, High-Expression States that Persist for Several Cell Divisions but Are Ultimately Transient

We generated a cell line (WM989-A6-G3 C10-C2 clone E9) that expresses a large but incomplete (and thus nonfluorescent) portion of the mNeonGreen2 fluorescent protein with the remaining piece of mNeonGreen2 fused to NGFR at the endogenous locus. When the NGFR fusion protein is expressed, the remaining portion of mNeonGreen2 binds to the NGFR fusion protein and becomes fluorescent. We then performed time-lapse microscopy imaging of the NGFR-mNG2 protein (nucleus labeled with H2b-iRFP670) at 6-h intervals for 8.75 days.

(A) We tracked cells through several cell divisions, thus building cellular lineages, and quantified fluorescence intensity for each cell. The plot shows two branches from the same parent cell with fluorescence intensity of mNeonGreen2 over time.

(B) Series of fluorescent micrographs of the two cells highlighted in (A). Images are subfield views taken from scans comprising several image tiles; boundaries in the image arise from edges between individual tiles. Scale bar, 8 μm.

(C) Correlations between sibling cells, first-cousins, and random pairs of cells (n = 486, 292, and 905, respectively).

(D) We stained cells with antibodies targeting AXL and EGFR, then sorted positive cells, plated them on a glass dish, and took images of their immunofluorescence signal. Subsequently, we acquired transmitted light images every hour for 8.67 days to facilitate tracking of cell lineages, and then we performed immunofluorescence again to measure EGFR and AXL levels at the end of the tracking period. From the time-lapse images, we tracked selected lineages that initially contained cells with high levels of EGFR and AXL. The red dots on the left images correspond to the red arrow on the histogram for an example initial cell subjected to tracking. Upon division, we colored the tracks of the sibling cells green and blue, respectively. The EGFR and AXL levels for these cells in their final state are indicated by the green and blue arrows on the histograms on the right. Scale bars, 10 μm.

See also Figure S2.

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

levels of fluorescence. We then tracked 222 cell lineages through several cell divisions (examples of positive cells in Figures 2A and 2B; Video S1). We observed that cells would occasionally initiate high levels of expression of NGFR-mNG2 (compare top branch versus bottom branch), and once initiated, that high level of expression could be maintained through multiple cell divisions, thus confirming the presence of memory. Further demonstrating the transience of this high expression state, we also observed cells transitioning from the high state to low levels of expression, with an average time in the high state of 40 h; however, a few cells showed longer fluctuations in NGFR levels ranging from 4.25–5.25 days (Figures S2E–S2G for discussion of analysis).

Another prediction of long-lived but imperfect memory is that sibling cells should show greater correlation in expression levels than cousins. Confirming this prediction, we observed that the correlation of expression between recently divided sibling cells was higher (R = 0.58) than between cousins (R = 0.43), although both values were higher than that between unrelated cells in the population (R = 0.04) (Figure 2C). Demonstrating the phenotypic significance of these fluctuations in expression levels, we further found that cells expressing high levels of NGFR-mNG2 at the time of vemurafenib addition were much more likely to continue to proliferate in drug (Figure S2L; Videos S2 and S3).

We wanted to verify that the high-expression state is transient for other genes, but it proved technically challenging to genetically tag genes such as *AXL* and *EGFR*. We thus used fluorescent antibodies to label AXL and EGFR, used fluorescence-activated cell sorting (FACS) to isolate the high population, seeded that population on an imageable surface, and performed time-lapse microscopy on the transmitted light images in order to trace lineages of these cells (Figures 2D and S2). At the end of 8.67 days, we fixed the cells and performed immunofluorescence *in situ*, followed by imaging to measure AXL and EGFR levels. We were able to track a total of 53 cells (26 for *AXL* and 27 for *EGFR*) (originating from 4; 2 each for *AXL* and *EGFR*), of which we could confidently re-identify 29 after the second round of immunofluorescence. Of these, we observed that 15/15 and 6/14 of these cells starting from those that initially had high levels of AXL and EGFR (higher than all the negative cells) (Figure 2D), respectively, eventually turned off ($<75^{th}$ percentile) within the time window of 8.67 days (Figures 2D, S2J, and S2K), demonstrating that the EGFR or AXL high state is indeed transient.

(We further verified this heritable but ultimately transient expression behavior by using FACS to isolate highly expressing cells and measuring the degree to which the expression levels in these cells reverted toward the distribution from the original population, finding a variety of timescales ranging from 5–9 days (Figures S3A–S3C, S3F, and S3G). Note that we observed an initial increase in expression for some of these genes upon sorting, which may be due to the stress associated with flow sorting or paracrine signaling in the concentrated subpopulation.)

Although time-lapse microscopy provided direct evidence of the long-lived fluctuations predicted by MemorySeq, it is difficult to perform for a panel of genes owing to the challenges associated with editing genes, especially those whose expression is very low in most cells. Thus, we sought another method to confirm these heritable fluctuations for a larger panel of genes. First, we performed experiments in fixed cells grown on culture dishes to measure heritability in gene expression by using spatial proximity as a proxy for relatedness. We seeded cells sparsely in culture dishes and then allowed them to grow for ~10 days, at which point we fixed the cells and subjected them to iterative single-molecule RNA fluorescence in situ hybridization (FISH) to measure the expression of 19 genes with a range of MemorySeq values (Figure S1C) in individual cells while preserving their spatial context (Figure 3A) (Shaffer et al., 2017). Genes with high MemorySeq signals also displayed rare-cell expression patterns as expected (Figure S4D). Our reasoning for this approach was that as cells divide, their spatial proximity would reflect their relatedness (Hormoz et al., 2016). In the case of a gene with non-heritable expression, one would expect to find no spatial correlation in which cells were deemed high expressing within the population. In contrast, for genes with heritable expression, one would expect to find the high-expressing cells to appear in patches corresponding to related neighboring cells that share a common ancestor that transitioned to the high-expression state. We found that genes identified by MemorySeq as being highly heritable (e.g., *EGFR*, *NGFR*, *NDRG1*, and *SERPINE1*) tended to show patch-like expression patterns across large numbers of cells, confirming that their expression was indeed heritable. In contrast, genes that MemorySeq would predict to not be heritable exhibited a more salt-and-pepper (variable but not heritable) expression pattern, as expected (Figures 3B, S4A, and S4C).

We wondered to what extent MemorySeq could measure differences in heritability of the high expression state for different genes. We therefore compared the degree of heritability from MemorySeq (given by the amount of skewness in the distribution of expression across MemorySeq clones) to the degree of heritability from spatial RNA FISH analysis (given by the amount of patchiness in the population). We found a strong correspondence between these two metrics (adjusted $R^2 = 0.6193$), suggesting that MemorySeq can stratify genes by the gradations in the degree of heritability that they display (Figure 3C; see Figure S4C for further analysis).

The timescales of particular genes turning between high (ON) and low (OFF) expression states should in principle be quantitatively related to the measured variability across the MemorySeq clones, with high variability corresponding to slow switching and vice versa. We thus analyzed a stochastic model of cell proliferation and switching relating these two quantities (STAR Methods) (Singh and Hespanha, 2010). Under the assumption that the ON state is relatively rare, this model yielded a direct relationship between the coefficient of variation measured by MemorySeq multiplied by the fraction of the time the cell is in the ON state and the predicted memory (number of generations cells are ON before turning OFF), with the only further parameter being the total number of divisions in the MemorySeq experiment. This equation predicted that over a relatively large range of reasonable parameters, (coefficient of variation [CV] ~0.5–2, fraction of time on ~0.01), the predicted memory was mostly in the range of 5–10 divisions, matching our experimental observations (Figures S2 and S3).

Motivated by our previous work in this cell line, we also looked for correspondences between the degree of heritability and the

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

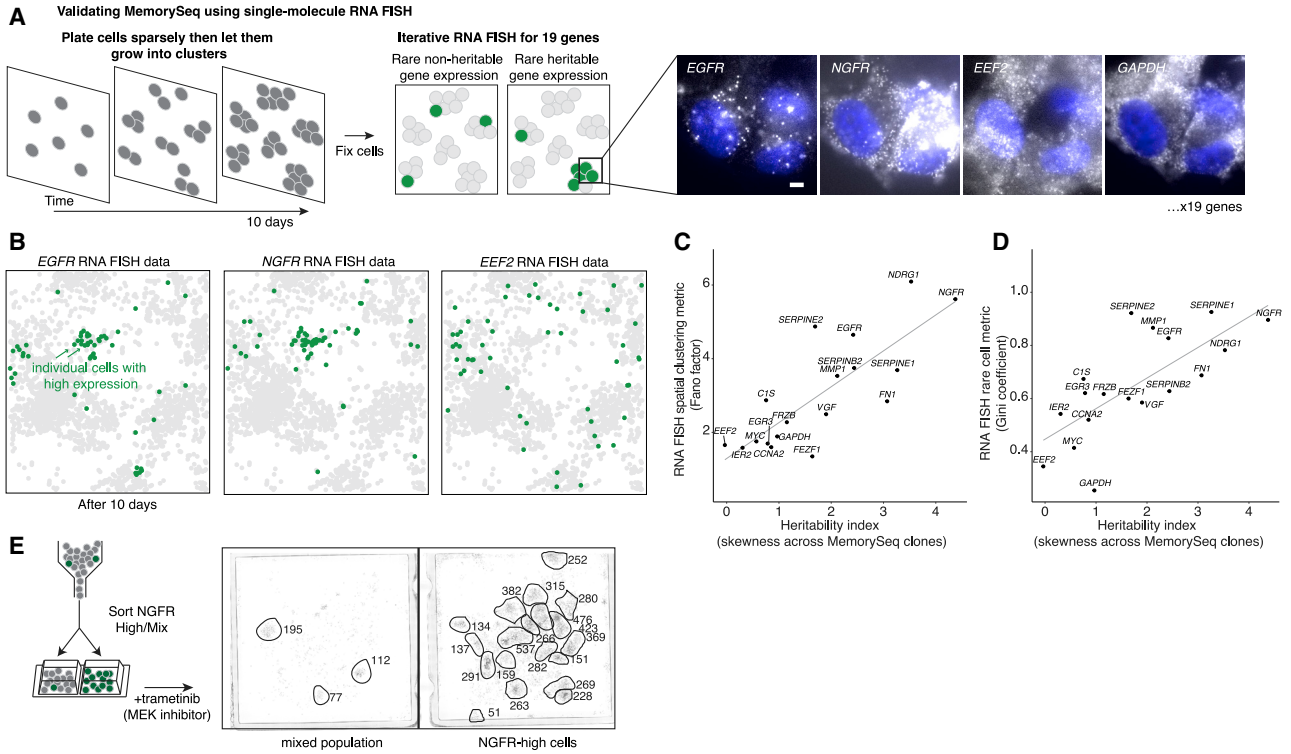


Figure 3. Single-Molecule RNA FISH Verifies the Quantitative Nature of MemorySeq for Measuring Heritability in Single Cells

(A) Schematic of spatial RNA FISH experiment. We plated WM989-A6 melanoma cells sparsely on a dish and allowed them to grow for 10 days. We then fixed the cells and performed iterative single-molecule RNA FISH to measure the expression of 19 genes. Closely related cells will remain in close proximity, thus heritable rare-cell expression would manifest as "patches" of "on" cells, whereas non-heritable rare-cell expression would display a more salt-and-pepper pattern of expression. Right: micrographs of RNA FISH for 4 genes, *EGFR* (heritable), *NGFR* (heritable), *EEF2* (housekeeping), and *GAPDH* (housekeeping).

(B) Each spot is a cell from an RNA FISH image scan of 12,192 cells (subset of 2,103 cells shown). Cells above a threshold (6 for *EGFR*, 36 for *NGFR*, and 320 for *EEF2*) were considered to be in the high expression state and colored green.

(C) Quantitative comparison of heritability as measured by MemorySeq (x axis: skewness across MemorySeq clones) and spatial RNA FISH analysis (y axis). We used the Fano factor measured for spatial bins of 20 nearest cells as a spatial clustering metric; randomly placed high-expression-state cells would display a Poisson distribution and thus give a Fano factor of 1. Cell populations with a Fano factor greater than 1 would display some degree of spatial clustering. Of note, this plot and the plot in (D) contain 18 of the 19 genes that we quantified with RNA FISH because 1 gene (*CYR61*) did not pass the minimum mean transcripts per million cutoff for analysis in the MemorySeq data.

(D) We plotted MemorySeq heritability versus the Gini coefficient (from RNA FISH). The Gini coefficient measures expression inequality, and thus indicates the rareness of expression, with 0 being completely equal and 1 being completely unequal.

(E) Rare cells within clonal WM989-A6 populations marked by high levels of NGFR protein were sorted, cultured for 8–16 h, and then subjected to trametinib treatment at 10 nM (MEK inhibitor) for 3 weeks. Image shows the number of resistant colonies (circled) along with number of cells within the resistant colony as indicated. Biological replicate available on Dropbox.

See also Figure S4.

rarity of gene expression as measured by the Gini coefficient, a metric for inequality (Jiang et al., 2016; Shaffer et al., 2017; Torre et al., 2018). We observed that indeed the two metrics were correlated (adjusted R^2 = 0.4898) (Figure 3D), suggesting that heritable genes identified by MemorySeq are more likely to express only in rare cells. This correspondence may be due to the design of the MemorySeq experiment, in which skewness can reach potentially higher levels for more rarerly expressing genes than for less rarely expressing ones.

Having validated that MemorySeq was accurately identifying genes displaying transcriptional memory, we then asked what these genes were and what their expression in rare cells signified. The underlying hypothesis was that these slow fluctuations are more likely to be associated with distinct cellular behaviors in those cell subpopulations than fast fluctuations. Our reasoning was that a distinct cellular behavior would likely require a persistently different gene expression pattern, involving deviations in the expression of several genes simultaneously, as opposed to a transient (and, as we hypothesized, probably inconsequential) fluctuation. In this melanoma cell line, we have previously shown that rare cells have high levels of expression of certain genes associated with therapy resistance (including *EGFR*, *NGFR*, and *AXL*), and that these rare cells are much more likely to survive the initial application of drug to develop into resistant clusters. Thus, we first wondered whether the set of genes identified by MemorySeq that mark rare cells overlapped with the set of genes associated with resistance. We found that most MemorySeq heritable genes (162 out of 227) were also markers of

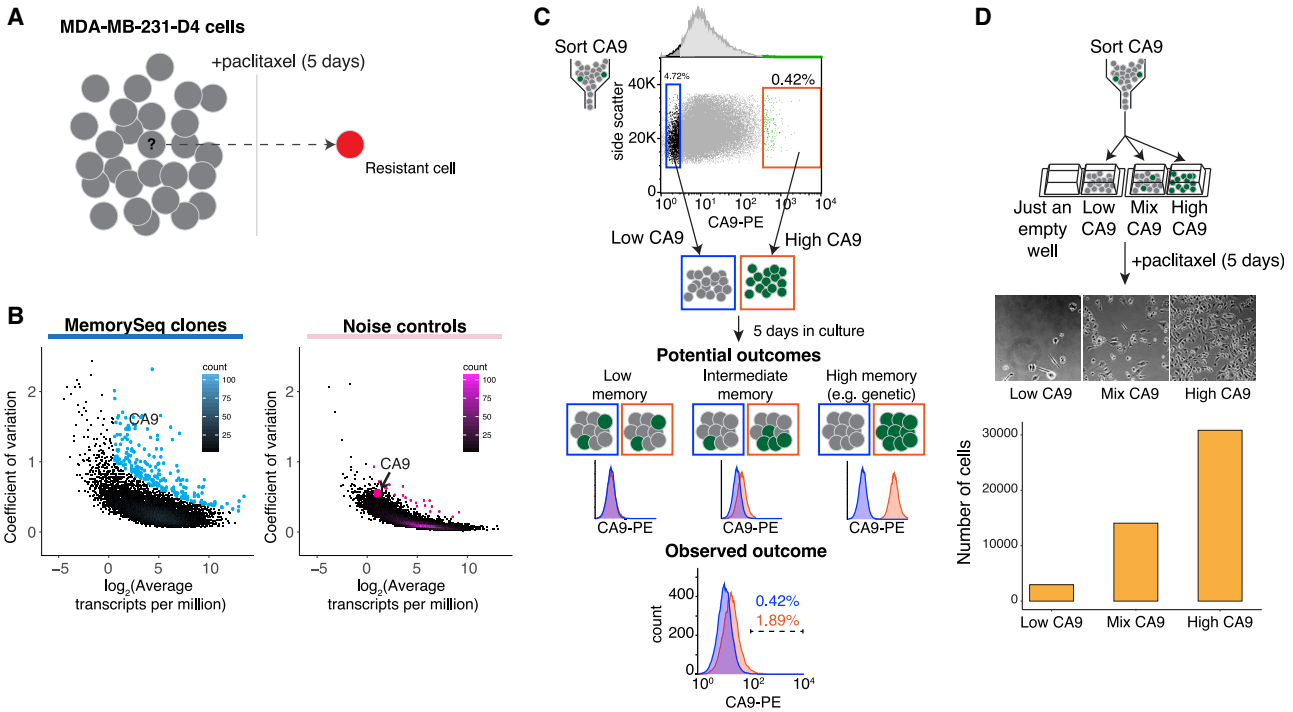


Figure 4. MemorySeq Reveals a Rare Subpopulation of MDA-MD-231-D4 Cells Associated with Drug Resistance

(A) Most MDA-MD-231-D4 cells die upon treatment with paclitaxel for 5 days, but a small subpopulation of cells (cell marked with ""?) survive and become resistant (red cell).

(B) We performed MemorySeq analysis on MDA-MD-231-D4 cells (n = 39 clones, left; n = 46 control clones, right). The blue colored dots correspond to genes that we statistically identified as being highly heritable by fitting a Poisson regression model and selecting genes with residuals in the top 2%, as was done with WM989.

(C) We stained cells with antibody targeting the CA9 surface marker and then sorted out the top 0.5% of cells and the lowest 5%. After culturing for 5 days, we re-stained the cells and measured CA9 levels by flow-cytometry. Potential outcomes for the levels of CA9 staining depending on the degree of gene expression memory. Observed outcome below for CA9 staining 5 days after initial sorting.

(D) We stained cells with antibody targeting the CA9 surface marker and then sorted out the top 2%–4% of cells, the lowest 2%–4% of cells, and the total "mix" population into chamber wells, after which we applied paclitaxel 1 day after sorting for 5 days. Transmitted light micrographs show the number of cells remaining after drug treatment for the different populations, and the quantification of the number of cells was performed using cell counting based on nuclear identification by imaging the DAPI nuclear stain and identifying computational techniques. All scale bars are 50 μm long.

See also Figure S3.

resistance (as determined by Shaffer et al., [2017]) (Figures S4I and S4J). These results suggest that the genes identified by MemorySeq are expressed in cell populations that are phenotypically distinct from the bulk of the population.

To verify the phenotypic differences of these cells, we used FACS to isolate cells by either high levels of *NGFR* or *EGFR* expression, and then we subjected them to treatment with a targeted inhibitor of MEK (trametinib) used to treat melanoma. On unsorted populations, upon treatment with this drug, a small percentage of cells will continue to grow and form colonies, mimicking the acquisition of drug resistance. In congruence with previous results, the EGFR/NGFR-high subpopulations resulted in far more resistant colonies after application of the drug, showing that this subpopulation is highly enriched for pre-resistant cells (Figure 3E). This result demonstrated that MemorySeq revealed the same subset of cells that we had previously determined to be highly enriched for drug resistant cells.

Our results thus far highlight MemorySeq's ability to prospectively reveal functionally distinct subpopulations within clonal populations of apparently homogeneous cells. In the case of the WM989-A6 melanoma cell line, we had already established the existence of such a subpopulation, but for most cell lines, there is little to no information about single-cell fluctuations that exhibit memory and thus may also be associated with distinct phenotypes. We thus set about testing MemorySeq in another cell line, MDA-MB-231-D4, which is a clonal derivative of a triple negative breast cancer cell line (does not express *HER2*, estrogen, or progesterone). Paclitaxel is a drug used to treat such breast cancers, but while it is able to kill most MDA-MB-231-D4 cells, some cells in the population are still able to survive the drug, thus leading to drug resistance (Figure 4A). However, prospective markers to isolate the subpopulation of cells that are resistant to drug have remained elusive (Gao et al., 2017), and we hypothesized that MemorySeq might be able to reveal genes whose expression was associated with this single-cell phenotype.

To test this hypothesis, we performed MemorySeq on the MDA-MB-231-D4 cell line by growing each of 48 subclones to

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

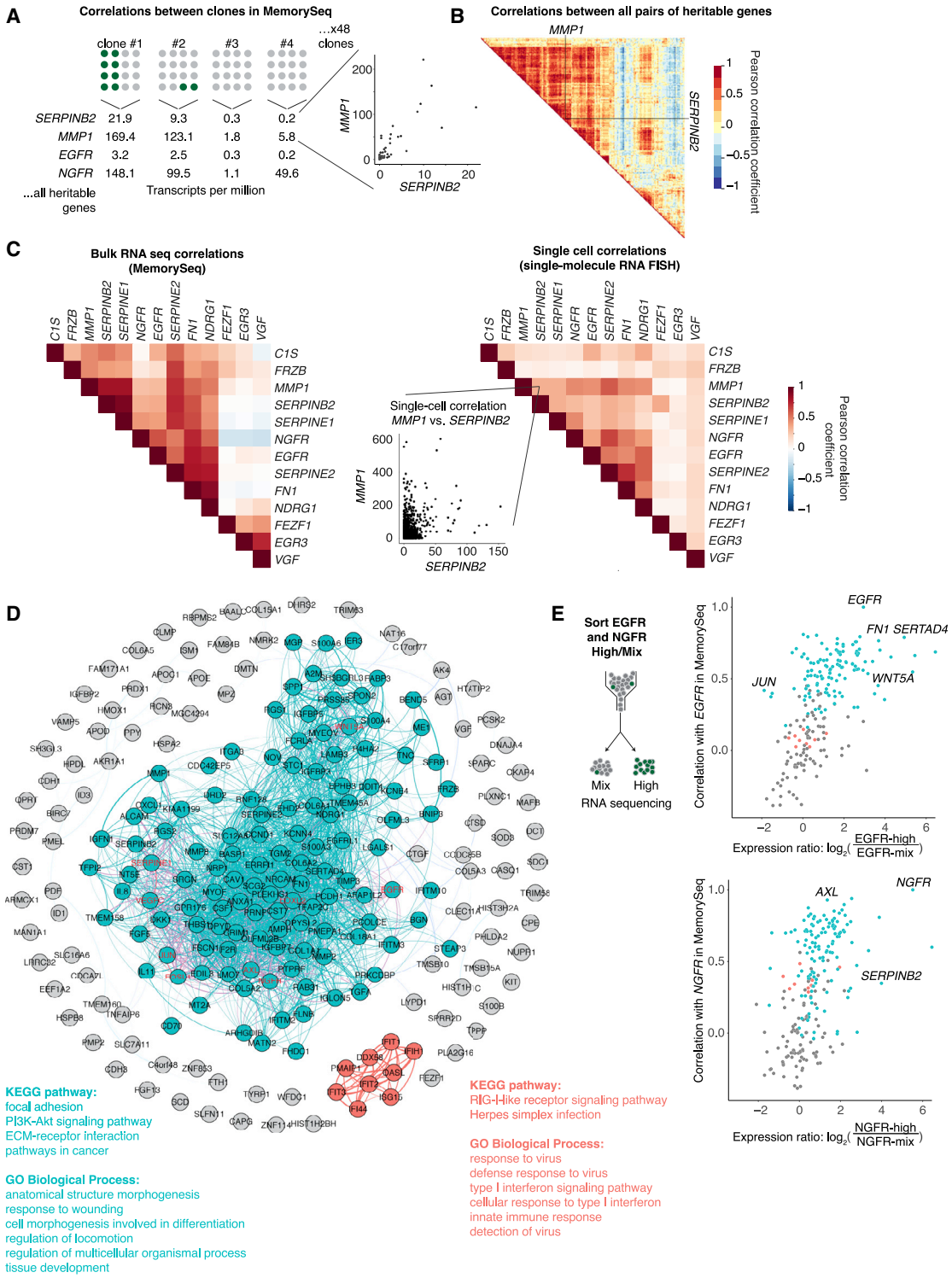


Figure 5. MemorySeq Enables the Identification of Coordinated Rare-Cell Expression Programs

(A) We measured correlations between genes across MemorySeq clones derived from WM989-A6 melanoma cells. Shown is an example correlation between *MMP1* and *SERPINB2* across 43 MemorySeq clones.

(B) Correlations between all pairs of genes exhibiting heritability as determined by the threshold described in Figure 1. Cook's distance analysis to test for outliers is given in Figure S1D.

(legend continued on next page)

around 100,000 cells, after which we performed RNA sequencing and quantification as described for WM989-A6. As with the melanoma cell lines, MemorySeq revealed a large number of genes in the MDA-MB-231-D4 cell line with putative heritable expression patterns (230 genes, Figure 4B). The range of variability in expression levels for all genes across the MemorySeq clones (most genes having low variability, but a few showing high variability) was comparable to that identified by MemorySeq for the WM989-A6 cell line. Interestingly, however, the overlap between the gene sets was relatively small (Figure S1L; Table S1), suggesting that different cell lines may have distinct sets of ''memory genes'' and that there is no universal ''memory gene'' expression program (with a potential exception noted below). As with WM989, we confirmed both the rarity and heritability of the expression pattern by using RNA FISH on cells initially seeded sparsely and then allowed to grow in place for 10 days, as done for the melanoma lines (Figure S1F, compare to Figure 3C).

Given the existence of these rare, slowly fluctuating subpopulations in the MDA-MB-231-D4 cells, we next asked whether these newly identified subpopulations were associated with phenotypic differences such as differential sensitivity to paclitaxel. Among the genes identified by MemorySeq was *CA9*, a surface marker known to be negatively associated with breast cancer chemosensitivity (Aomatsu et al., 2014; Span et al., 2003), but for which there was no reason to suspect that its expression at the single-cell level would be indicative of which cells specifically survived upon drug treatment. We thus immunolabeled the MDA-MB-231-D4 cells using antibodies targeting CA9 and then used FACS to isolate a rare population of CA9-high cells (top 2%–4%, along with CA9-low and mixed subpopulations; validation of sorting in Figure S1M), after which we added paclitaxel to both and grew the cells for 5 days (Figure 4D). We found that, when treated with 1 nM of paclitaxel, the CA9-high cells were more likely to be resistant than either the CA9-low or mixed subpopulations (Figure S1N). Furthermore, the sorted subpopulations reverted to the population average, demonstrating that the CA9-high state is ultimately transient (Figures 4C, S3H, and S3I). These results, in a cancer cell line of a completely different type involving a drug with a completely different mechanism of action, demonstrate that MemorySeq is able to identify *de novo* heritable, rare-cell expression states, and these states are phenotypically distinct from the others in the population.

Behavioral differences such as drug resistance are typically associated with the differential expression of many genes at once. We thus further hypothesized that the long timescale of these single-cell fluctuations could allow for significant co-fluctuation; that is, if a cell expresses a high level of one high-memory gene for a sufficiently long time period, it could also have a higher probability of expressing another slowly fluctuating gene simultaneously. Indeed, should such a phenomenon be prevalent, it would allow us to organize these high memory genes into characteristic modules of genes that co-fluctuate in single cells.

To isolate such modules, we calculated the correlation coefficient between the expression of all pairs of heritable genes across the MemorySeq clones derived from WM989-A6. We reasoned that if a particular clone had a high abundance of a particular transcript, then the abundance of transcripts of co-fluctuating genes would also be high in that particular clone. We saw large blocks of genes whose expression appeared to correlate strongly with each other, suggesting that they co-fluctuate at the single-cell level (Figure 5B). To validate that the programs so identified by MemorySeq corresponded to single-cell correlations, we compared the correlations between MemorySeq and RNA FISH in single cells on a panel of genes across two separate clusters. We found a general correspondence between these two assays, suggesting that MemorySeq is able to identify groups of genes *de novo* that co-fluctuate in rare-cell expression programs (Figures 5C, S4G, and S4H). (The expression of some genes appeared to mismatch between MemorySeq and single-molecule RNA FISH, potentially due to mutually exclusive expression patterns in rare cells or due to short-lived transition states.) We observed similar clustering in another melanoma cell line (WM983B-E9) (Figure S1I) and MDA-MB-231-D4, although the specific genes were typically different (Figure S1E). We also performed MemorySeq on the lung cancer cell line PC-9, which showed a total of 240 heritable genes, including 8 genes that were heritable in all of the other cell lines on which we performed MemorySeq (WM989, WM983B, and MDA-MB-231) (Figures S1G, S1H, and S1L).

The appearance of distinct clusters of slowly co-fluctuating genes led us to use community detection algorithms for network data to demarcate these groups of genes for further analysis. We used a weighted version of *k*-clique community detection (Derényi et al., 2005; Jonsson et al., 2006; Palla et al., 2005) to identify such groups of genes (Figure 5D) (see STAR Methods for more information). We chose *k*-clique community detection because it allows for nodes to be in multiple communities at once, echoing the ability of a protein to simultaneously play multiple roles within the cell. In WM989-A6 cells, one large community of genes overlapped very strongly with the vemurafenib pre-resistance marker gene set that we identified earlier (Shaffer et al., 2017). We verified this correspondence by comparing the transcriptomes of sorted subpopulations of EGFR-high cells to the clusters of

(C) Comparison of coherence between MemorySeq bulk RNA sequencing analysis and single-cell correlations as measured by single-molecule RNA FISH. We performed RNA FISH on 20 genes in WM989-A6 cells, keeping for further analysis genes whose RNA FISH Gini coefficient was greater than 0.6 (13 genes remaining). The correlation between bulk MemorySeq RNA sequencing levels is on the left, RNA FISH on the right. Callout shows raw RNA FISH counts for 12,192 cells between *MMP1* and *SERPINB2* in single cells.

(D) Community detection within the network defined by the correlation matrix of co-expression patterns among the heritable genes. Gray circles indicate genes that did not comprise a network community. Green and red indicate the two communities detected; KEGG pathway and Gene Ontology (GO) biological process analysis results shown for both communities.

(E) Comparison of rare-cell expression programs identified by MemorySeq and those identified by sorting EGFR-high (top) or NGFR-high (bottom) cells (using fluorescent antibody labeling followed by RNA sequencing on the high versus mix populations).

See also Figure S5.

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

genes identified by MemorySeq, showing that the expression levels of genes specifically in this cluster correlated well with that of genes that correlated with EGFR expression (Figure 5E).

We also found other communities within the set of heritable genes in WM989-A6 cells, suggesting the existence of multiple independent heritable gene expression programs. One prominent program included *DDX58* (*RIG-I*), *IFIT1*, *PMAIP1*, and *OASL*, which may be related to type 1 interferon signaling (Loo and Gale, 2011), and was notable because it also appeared to some extent in MDA-MB-231-D4 and WM983B-E9 (Table S1). We verified that this cluster expressed in a distinct rare-cell subpopulation from that containing *EGFR* and *AXL* by performing RNA FISH for *PMAIP1*, *DDX58*, *AXL*, and *EGFR* simultaneously. This analysis showed that *PMAIP1* and *DDX58* expression exhibited very strongly correlated rare-cell expression, but neither correlated to much extent with either *AXL* or *EGFR* expression, and both genes' expression also exhibited memory (Figures S4E–S4H). This cluster appeared relatively distinct from the primary community associated with drug resistance and *EGFR* expression, and indeed, did not show any association with the EGFR-high transcriptome (Figure 5E). Another community in WM989-A6 cells was somewhat less coherent and included genes such as *VGF*. *VGF* expression also showed strong heritability (Figure 3C), but its expression appeared not to correlate with the other pre-resistance genes at the single-cell level (Figure 5E) and was not associated with the drug resistance phenotype (Shaffer et al., 2017).

The fact that multiple genes appeared to coordinate their expression across multiple chromosomes suggested that the mechanism for maintaining memory occurs in *trans* i.e., through the regulatory milieu rather than just a short but intense pulse of transcription from a single allele that is maintained through cell division. Labeling sites of nascent transcription for *EGFR* and *AXL* (Levesque and Raj, 2013) revealed that transcriptional bursting also occurred in patches (as opposed to just transcript abundance), further suggesting that memory of the high expression state was due to sustained transcriptional activity as opposed to a single sporadic, large, transient burst of transcription in a precursor cell (Figures S3D and S3E). Moreover, we often observed multiple active transcription sites active within a single cell, confirming that the fluctuations in expression were not the result of a fluctuation at just one allele (Figures S5J and S5K). We also performed allele-specific expression analysis from the RNA sequencing data across MemorySeq clones to confirm these results across a broader swath of MemorySeq genes. We found that fluctuations appeared to occur across both alleles simultaneously, further suggesting that fluctuations are driven by transactivation rather than pulses at a single allele of a given gene (Figure S5H).

We sought to identify regulatory factors that may be responsible for the unique behavior of MemorySeq genes. We first utilized ATAC-seq profiling on WM989-A6 cells (Shaffer et al., 2017) to reveal regions of DNA with increased chromatin accessibility in the vicinity of the genes identified by MemorySeq. We searched these regions for sequence motifs matching known transcription factor binding sites, and found enrichment for motifs corresponding to SOX10 and the FOS/JUN and NFAT families of transcription factors compared to accessible regions surrounding randomly chosen expression-matched control genes (Figure S5E). To validate these associations, we knocked out a total of 6 transcription factors from these families and performed RNA sequencing to measure changes in expression (Torre et al., 2019). As suggested by our ATAC-seq analysis, knockout of 4 of these transcription factors showed strong effects on MemorySeq gene transcription (FOSL1, JUNB, NFAT5, and SOX10) and 2 showed small effects (JUN and NFAC2) (Figure S5F). The manner by which such factors may act to change the expression of these genes is unknown, but we did look for specific signatures of histone marks around MemorySeq genes. We observed depletion of H3K27 acetylation (H3K27ac) marks typically associated with active transcription and gain of H3K27 trimethylation (H3K27me3) marks typically associated with repression (again, compared to expression-matched control genes). These associations suggest that the fluctuations may result from an inability to sustain ongoing expression due to this distinct pattern of histone acetylation and methylation (Figures S5A–S5D).

DISCUSSION

In sum, we believe MemorySeq is a simple but powerful method for identifying rare, heritable expression patterns in cells. We have shown here that such rare, heritable expression programs may be related to non-genetic mechanisms of therapy resistance in cancer (Pisco and Huang, 2015; Roesch et al., 2010; Shaffer et al., 2017; Sharma et al., 2010; Spencer et al., 2009). However, they may also be important in other contexts in which rare cells behave differently than the rest of the population, both in cancer (such as metastasis), but also in otherwise healthy tissues or in cellular reprogramming contexts like the induction of induced pluripotent stem cells (iPSCs) (Hanna et al., 2009).

The mechanisms underlying these fluctuations remain mysterious. We identified some transcription factors that can affect expression of these genes and found patterns of histone acetylation that are associated with these genes as well. Further work will be required to test other potential models of gene regulation that could lead to fluctuations, such as methylation patterns or other regulatory mechanisms that may operate on the relevant timescales (Meir et al., 2020). Other potential mechanisms include autocrine or paracrine signaling, by which cells are able to sustain longer fluctuations by signaling feedback to themselves or nearby relatives.

It is also interesting that MemorySeq is quantitative in the sense that it is not just able to report that a gene's expression is heritable but is also able to provide a relative sense of how heritable that expression is, meaning how many cell divisions are required before cells begin to forget the rare-cell expression state. By revealing such intermediate timescales of fluctuation, the results of MemorySeq suggest that the classic use of the concept of somatic epigenetics (non-genetic heritability) may require re-evaluation as a quantitative, rather than qualitative, property of some cellular states.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT AVAILABILITY
  - Cell lines and culture
- METHOD DETAILS
  - MemorySeq
  - Generation of NGFR-mNG2 reporter cell line
  - Fluorescence-Activated Cell Sorting (FACS)
  - Live-cell immunofluorescence
  - Drug treatment experiments
  - Time-lapse imaging
  - RNA FISH
  - RNA FISH imaging
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Cell quantification
  - Time-lapse analysis
  - RNA FISH image analysis
  - Spatial analysis
  - ATAC sequencing analysis
  - RNA sequencing analysis of WM989-A6-G3 knockouts
  - Allele-specific expression (ASE) analysis
  - ChIP sequencing analysis
  - Network community identification
  - Inferring timescale of expression heritability from MemorySeq
- STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.07.003.

ACKNOWLEDGMENTS

We thank Siyu Feng, David Brown, and Bo Huang for constructs pSFFV_mNG2(1-10) and Christian Meyer for discussions about sort-relaxation experiments. We thank Caroline Bartman, Yogesh Goyal, Rodrigo Gier, Margaret Dunagin and other members of the Raj and Shaffer labs for many useful suggestions. We thank the Flow Cytometry Core Laboratory at the Children's Hospital of Philadelphia Research Institute for assistance in designing and performing FACS. S.M.S. acknowledges support from NIH (F30 AI114475 and DP5 OD028144). B.L.E. acknowledges support from NIH (F30 CA236129, T32 GM007170, and T32 HG000046). R.A.R.H. acknowledges support from the NSF Graduate Research Fellowship (DGE-1845298). A.S. acknowledges support by NIH (5R01GM124446 and 5R01GM126557). A.E.S. and D.S.B. acknowledge support from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, and the ISI Foundation. A.R. acknowledges NIH/NCI PSOC (U54 CA193417), NSF (CAREER 1350601), SPORE (P50 CA174523), NIH (P30 CA016520, U01 CA227550, TR01 GM137425, and R01 CA232256), NIH 4DN (U01 HL129998), NIH Center for Photogenomics (RM1 HG007743), and the Tara Miller Foundation.

AUTHOR CONTRIBUTIONS

S.M.S., A.R., and B.L.E. conceived of study and designed experiments. S.M.S., B.L.E., and E.T. performed experiments. S.M.S., B.L.E., and G.H. performed failed experiments. S.M.S., A.R., B.L.E., R.A.R.H., C.C., G.H., D.L.S., A.E.S., R.G., and A.S. performed analyses of data. D.S.B. supervised network analyses. S.M.S., A.R., and B.L.E. wrote the manuscript. A.R. supervised the research.

DECLARATION OF INTERESTS

A.R. receives consulting income from Biosearch Technologies. A.R. and S.M.S. receive royalties related to Stellaris RNA FISH probes.

Received: December 23, 2019
Revised: May 4, 2020
Accepted: July 1, 2020
Published: July 30, 2020

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Aomatsu, N., Yashiro, M., Kashiwagi, S., Kawajiri, H., Takashima, T., Ohsawa, M., Wakasa, K., and Hirakawa, K. (2014). Carbonic anhydrase 9 is associated with chemosensitivity and prognosis in breast cancer patients treated with taxane and anthracycline. BMC Cancer *14*, 400.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. ICWSM *8*, 361–362.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C.A., Dempster, J., Lyons, N.J., Burns, R., et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. Nature *560*, 325–330.

Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. Science *330*, 612–616.

Breese, M.R., and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics *29*, 494–496.

Buonato, Janine M., and Lazzara, Matthew J. (2014). ERK1/2 Blockade Prevents Epithelial–Mesenchymal Transition in Lung Cancer Cells and Promotes Their Sensitivity to EGFR Inhibition. Cancer Research *74*, 309–319.

Castel, S.E., Mohammadi, P., Chung, W.K., Shen, Y., and Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat. Commun. *7*, 12817.

Chao, H.X., Fakhreddin, R.I., Shimerov, H.K., Kumar, R.J., Gupta, G.P., and Purvis, J.E. (2018). Evidence that the cell cycle is a series of uncoupled, memoryless phases. bioRxiv. https://doi.org/10.1101/283614.

Cohen, A.A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., et al. (2008). Dynamic proteomics of individual cancer cells in response to a drug. Science *322*, 1511–1516.

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. Phys. Rev. Lett. *94*, 160202.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Fallahi-Sichani, M., Becker, V., Izar, B., Baker, G.J., Lin, J.-R., Boswell, S.A., Shah, P., Rotem, A., Garraway, L.A., and Sorger, P.K. (2017). Adaptive resistance of melanoma cells to RAF inhibition via reversible induction of a slowly dividing de-differentiated state. Mol. Syst. Biol. *13*, 905.

Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M.D., and Huang, B. (2017). Improved split fluorescent proteins for endogenous protein labeling. Nat. Commun. *8*, 370.

Franco, H.L., Nagari, A., Malladi, V.S., Li, W., Xi, Y., Richardson, D., Allton, K.L., Tanaka, K., Li, J., Murakami, S., et al. (2018). Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res. *28*, 159–170.

Gao, R., Kim, C., Sei, E., Foukakis, T., Crosetto, N., Chan, L.-K., Srinivasan, M., Zhang, H., Meric-Bernstam, F., and Navin, N. (2017). Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. Nat. Commun. *8*, 228.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, arXiv:1207.3907.

Giusti, C., Pastalkova, E., Curto, C., and Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. Proc. Natl. Acad. Sci. USA *112*, 13455–13460.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595–601.

Hormoz, S., Singer, Z.S., Linton, J.M., Antebi, Y.E., Shraiman, B.I., and Elowitz, M.B. (2016). Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Syst. *3*, 419–433.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144.

Jonsson, P.F., Cavanna, T., Zicha, D., and Bates, P.A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics *7*, 2.

Krepler, Clemens, Sproesser, Katrin, Brafford, Patricia, Beqiri, Marilda, Garman, Bradley, Xiao, Min, Shannan, Batool, et al. (2017). A Comprehensive Patient-Derived Xenograft Collection Representing the Heterogeneity of Melanoma. Cell Reports *21*, 1953–1967.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46* (D1), D252–D259.

Levesque, M.J., and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat. Methods *10*, 246–248.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Loo, Y.-M., and Gale, M., Jr. (2011). Immune signaling by RIG-I-like receptors. Immunity *34*, 680–692.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Luria, S.E., and Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics *28*, 491–511.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A., and Tanay, A. (2020). Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. Nature Genetics *52*, 709–718.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics *32*, 1749–1751.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature *435*, 814–818.

Phillips, N.E., Mandic, A., Omidi, S., Naef, F., and Suter, D.M. (2019). Memory and relatedness of transcriptional activity in mammalian cell lineages. Nat. Commun. *10*, 1208.

Pisco, A.O., and Huang, S. (2015). Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. Br. J. Cancer *112*, 1725–1732.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol. *4*, e309.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913–918.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44* (W1), W160-5.

Rieck, B., Fugacci, U., Lukasczyk, J., and Leitte, H. (2018). Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks. IEEE Trans. Vis. Comput. Graph. *24*, 822–831.

Roesch, A., Fukunaga-Kalabis, M., Schmidt, E.C., Zabierowski, S.E., Brafford, P.A., Vultur, A., Basu, D., Gimotty, P., Vogt, T., and Herlyn, M. (2010). A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth. Cell *141*, 583–594.

Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435.

Sharma, S.V., Lee, D.Y., Li, B., Quinlan, M.P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M.A., et al. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell *141*, 69–80.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. Nature *444*, 643–646.

Singh, A., and Hespanha, J.P. (2010). Stochastic hybrid systems for studying biochemical processes. Philos. Trans.- Royal Soc., Math. Phys. Eng. Sci. *368*, 4995–5011.

Span, P.N., Bussink, J., Manders, P., Beex, L.V.A.M., and Sweep, C.G.J. (2003). Carbonic anhydrase-9 expression levels and prognosis in human breast cancer: association with treatment outcome. Br. J. Cancer *89*, 271–276.

Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., and Sorger, P.K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature *459*, 428–432.

Symmons, O., and Raj, A. (2016). What's Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism. Mol. Cell *62*, 788–802.

Tlsty, T.D., Margolin, B.H., and Lum, K. (1989). Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbrück fluctuation analysis. Proc. Natl. Acad. Sci. USA *86*, 9441–9445.

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. Cell Syst. *6*, 171–179.

Torre, E.A., Arai, E., Bayatpour, S., Beck, L.E., Emert, B.L., Shaffer, S.M., Mellis, I.A., Fane, M., Alicea, G., Budinich, K.A., et al. (2019). Genetic screening for single-cell variability modulators driving therapy resistance. bioRxiv. https://doi.org/10.1101/638809.

Verfaillie, A., Imrichova, H., Atak, Z.K., Dewaele, M., Rambow, F., Hulselmans, G., Christiaens, V., Svetlichnyy, D., Luciani, F., Van den Mooter, L., et al. (2015). Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat. Commun. *6*, 6683.

Wernet, M.F., Mazzoni, E.O., Çelik, A., Duncan, D.M., Duncan, I., and Desplan, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. Nature *440*, 174–180.

Xi, Y., Shi, J., Li, W., Tanaka, K., Allton, K.L., Richardson, D., Li, J., Franco, H.L., Nagari, A., Malladi, V.S., et al. (2018). Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. BMC Genomics *19*, 150.

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-EGFR antibody	Millipore	Cat#MABF120; clone 225; RRID: AB_11205738
donkey anti-mouse IgG AlexaFluor488	Jackson Laboratories	Cat#715-545-150; RRID: AB_2340846
anti-NGFR PE-Cy7 antibody	Biolegend	Cat#400126; clone ME20.4; RRID: AB_326448
anti-AXL antibody	Novus	Cat#AF154; lot DMG0516031; RRID: AB_2861155
bovine anti-goat IgG AlexaFluor647	Jackson Laboratories	Cat#805-605-180; RRID: AB_2340885
anti-CA9 PE antibody	Miltenyi Biotec	Cat#130-110-057; clone REA658; RRID: AB_2651325
Bacterial and Virus Strains		
Endura electrocompetent *E. coli*	Lucigen	Cat#60242
One Shot STBL3 chemically competent E. Coli	Thermo Fisher	Cat#C737303
Chemicals, Peptides, and Recombinant Proteins		
DAPI, dilactate	Thermo Fisher	Cat#D3571
CellTraceViolet	Invitrogen	Cat#C34557
7-AAD	Biolegend	Cat#420404
Trametinib (GSK1120212)	Selleckchem	Cat#S2673
Vemurafenib (PLX4032)	Selleckchem	Cat#S1267
Paclitaxel	Life Technologies	Cat#P3456
Cy3 Mono-reactive dye pack	GE healthcare	Cat#PA23001
AlexaFluor 594 NHS Ester	Invitrogen	Cat#A37572
Atto 647N NHS Ester	ATTO-TEC	Cat#AD 647N-31
Atto 700 NHS Ester	ATTO-TEC	Cat#AD 700-31
EnGen Cas9 NLS, S. pyogenes	NEB	Cat#M0646M
Agencourt AMPure XP	Beckman Coulter	Cat#A63881
MCDB 153	Sigma; prepared by University of Pennsylvania Cell Center	Cat#M7403
Leibovitz L-15 Medium	Thermo Fisher	Cat#11415064
DMEM, high glucose, GlutaMAX, pyruvate	Thermo Fisher	Cat#10569010
RPMI 1640	Thermo Fisher	Cat#11875085
Fetal Bovine Serum	Fisher scientific	Cat#16000044
Propidium Iodide	BD Biosciences	Cat#556463
Penicillin/Streptomycin	Thermo Fisher	Cat#15140148
Critical Commercial Assays		
miRNAeasy RNA extraction mini kit	QIAGEN	Cat#217004
NEBNext Poly(A) Magnetic Isolation Module	NEB	Cat#E7490L
NEBNext Ultra RNA sequencing library prep kit for Illumina	NEB	Cat#E7530L
EnGen sgRNA Synthesis Kit, S. pyogenes	NEB	Cat#E3322S
NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1)	NEB	Cat#E7600S

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
RNA sequencing data	This paper	GEO: GSE151375
RNA sequencing data	Shaffer et al., 2017	GEO: GSE97679
ATAC sequencing data	Shaffer et al., 2017	GEO: GSE97680
RNA sequencing data	Torre et al., 2019	GEO: GSE151825 and https://www.dropbox.com/sh/0olkoxmaobssl44/AAAKPLdADB1_sB-j8lhhP7Eja?dl=0
ChIP-seq data	Verfaillie et al., 2015	GEO: GSE60666 and http://ucsctracks.aertslab.org/papers/melanoma_paper/
RNA sequencing data	Verfaillie et al., 2015	GEO: GSE60666
ChIP-seq data	Franco et al., 2018; Xi et al., 2018	GEO: GSE85158
RNA sequencing data	Franco et al., 2018; Xi et al., 2018	SRA:SRP102239
Experimental Models: Cell Lines		
WM989	Laboratory of Meenhard Herlyn	Krepler et al., 2017; Shaffer et al., 2017
WM983b	Laboratory of Meenhard Herlyn	Krepler et al., 2017; Shaffer et al., 2017
MDA-MB-231	ATCC	ATCC HTB-26
PC-9	Laboratory of Matthew Lazzara	Buonato and Lazzara, 2014
Oligonucleotides		
RNA FISH probes	This paper	Table S2
mNG2(11) HDR template	This paper	Table S3
NGFR sgRNA	This paper	Table S3
Recombinant DNA		
mNG2(1-10) lentivirus plasmid	Laboratory of Bo Huang	Feng et al., 2017
H2b-iRFP lentivirus plasmid	This paper	N/A
Software and Algorithms		
STAR version 2.3.0e	Dobin et al., 2013	https://github.com/alexdobin/STAR
HTSeq version 0.6.1	Anders et al., 2015	https://github.com/htseq/htseq
Samtools version 0.1.19	Li et al., 2009	https://github.com/samtools/samtools
NGSUtils	Breese and Liu. 2013	https://github.com/ngsutils/ngsutils
RajLabImageTools	This paper	https://github.com/arjunrajlaboratory/rajlabimagetools
R version 3.6	R core team	https://www.R-project.org/
MATLAB	Mathworks	https://www.mathworks.com/products/matlab.html
Deeptools version 3.3.0	Ramírez et al., 2016	https://deeptools.readthedocs.io/en/develop/
Freebayes version 1.3.2	Garrison and Marth. 2012	https://github.com/ekg/freebayes
bcftools version 1.9	Narasimhan et al., 2016	http://samtools.github.io/bcftools/
phASER	Castel et al., 2016	https://github.com/secastel/phaser
Analysis of Motif Enrichment (AME)	McLeay and Bailey. 2010	http://meme-suite.org/index.html
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Gephi	Bastian et al., 2009	https://gephi.org/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Arjun Raj (arjunrajlab@gmail.com).

Materials Availability

All materials and reagents generated as part of this study will be made available upon installment of a material transfer agreement (MTA).

Data and Code Availability

Sequencing data generated as part of this study are available on GEO under the accession number GEO: GSE151375. In addition, these data and code used for their analysis are available on Dropbox at https://www.dropbox.com/sh/f90osuwxak1f9vj/AAAOWsxAudr77aT1bhzLNb69a?dl=0.

EXPERIMENTAL MODEL AND SUBJECT AVAILABILITY

Cell lines and culture

We used 4 cell lines in our study: WM989-A6, which is a subclone of the melanoma line WM989 (Wistar Institute, a kind gift of Meenhard Herlyn); WM983B-E9, a subclone of WM983B (a kind gift of Meenhard Herlyn, Wistar Institute); MDA-MB-231-D4, a subclone of MDA-MB-231 (ATCC HTB-26) and PC-9-D11, a subclone of PC-9 (a kind gift from Matthew Lazzara, University of Virginia). We verified all cell lines by DNA fingerprinting: WM989-A6 and WM983B-E9 were performed at the Wistar Institute by DNA STR Microsatellite testing and MDA-MB-231 and PC-9 were performed by ATCC human STR profiling cell line authentication services. We cultured the melanoma cell lines in TU2% (containing 80% MCDB 153, 10% Leibovitz's L-15, 2% FBS, 2.4mM $CaCl_2$ and 50 U/mL penicillin, and 50 μg/mL streptomycin), the MDA-MB-231 cell lines in DMEM10% (DMEM with glutamax, 10% FBS and 50 U/mL penicillin, and 50 μg/mL streptomycin) and the PC-9 cell lines in RPMI10% (RPMI 1640, 10% FBS, 2mM glutamax, and 50 U/mL penicillin, and 50 μg/mL streptomycin) .

METHOD DETAILS

MemorySeq

Our experiment roughly followed the design of Luria and Delbrück's original fluctuation analysis, but with RNA sequencing as the terminal readout instead of the number of resistant colonies. From the parental cell line (WM989-A6, WM983B-E9, MDA-MB-231-D4), we isolated a single cell, let it proliferate until reaching roughly 100-200 cells, then plated these cells into a 96 well plate at a dilution in which we expected roughly 0.5 cells per well. From these wells, we isolated ~100 clones for further expansion, excluding wells that were seeded with more than 1 cell. Of the 100 starting clones, we aimed for 48 clones from each cell line for downstream analysis. We grew the clones until they reached a minimum of around 100,000 cells, with some reaching as high as roughly 200,000 cells. At that point, we used miRNAeasy RNA extraction kit to isolate RNA from each clone, followed by library preparation using the NEBNext Poly(A) Magnetic Isolation Module and NEBNext Ultra RNA sequencing library prep kit for Illumina. At the time of RNA isolation for the clones, we also isolated 48 separate samples of 100,000 cells from the parental line and prepared these samples for RNA sequencing as controls. For each cell line, we sequenced a total of 96 samples, including 48 clones and 48 controls from a mixture of the parental cell line. We sequenced to a depth of at least 500,000 reads per RNA sequencing library (with a typical depth of around 4 million reads) on a NextSeq500 (Illumina). While we targeted 48 clones and 48 controls for each cell line, we had a few samples with poor RNA quality and occasionally lost samples when preparing the libraries. Therefore, after culturing, extracting RNA, and preparing libraries, we ended up with 39-46 clones total for each cell line for our analysis (43 clones for WM989, 46 clones for WM983B, 39 clones for MDA-MB-231, and 42 clones for PC-9). We aligned the reads using STAR and enumerated uniquely mapped read counts per gene using HTseq (Anders et al., 2015; Breese and Liu, 2013; Dobin et al., 2013; Li et al., 2009); pipeline available at https://github.com/arjunrajlaboratory/RajLabSeqTools).

For computational analysis of the Luria-Delbruck RNA sequencing data, we calculated the transcripts per million of every gene in each sample. We then calculated metrics of the variation across the different 48 clonal samples, including the coefficient of variation, skewness, and kurtosis. We also compared these metrics in the clonal samples to those observed in the mixed controls. We found that the relationship between the coefficient of variation and the transcripts per million for every gene could be fit by a Poisson regression model. We fit this model for each cell line and then defined the panel of heritable genes as those with residuals greater than the 98th percentile. We also set a minimum level of expression for heritable genes as 1.5 transcripts per million for WM989 and MDA-MB-231 and 1.5 transcripts per million for WM983B. This approach yielded 227 heritable genes in WM989, 230 heritable genes in WM983B, 230 heritable genes in MDA-MB-231, and 240 heritable genes in PC-9 (listed in Table S1). We generated correlation matrices from the pairwise Pearson correlation coefficients for heritable genes across all clones. We calculated the Cook's distance for the pairwise correlations to determine sensitivity to outliers (Figure S1D). For a few pairs, the correlation coefficient was deemed sensitive to the presence of outliers. This sensitivity is to be expected because of the experimental design; as per Luria and Delbrück's fluctuation analysis, there will be rare outlier cultures. Given that we did not observe such outlier cultures in our technical controls, we believe that these outlier cultures reflect true biological variability. Computational analysis of RNA sequencing data is available on the Dropbox here https://www.dropbox.com/sh/7z0n6ixshghdvlo/AADb7cY9ZGFMF1CRbLJpnYQSa?dl=0 and https://www.dropbox.com/sh/vykm0gu0afy39dm/AAAjPMibwX5LfaPWHnpD5k67a?dl=0.

Generation of NGFR-mNG2 reporter cell line

To create the NGFR-mNG2(11) WM989-A6-G3 cell line, we used the split mNeonGreen2(mNG2) system described in Feng et al. (2017). In brief, we first transduced WM989-A6-G3 with 10/11ths of mNG2, which is non-fluorescent without the remaining 1/11th of the protein. We then electroporated cells with Cas9 RNP containing a guide RNA specific to the C terminus of *NGFR* and a single-stranded DNA template encoding the remaining 1/11th of the protein flanked by sequences homologous to the targeted locus (sequences available in Table S2 and at https://www.dropbox.com/sh/pqjropgxr1u5xd3/AACEgMjetkmINp_l0Df8ujSUa?dl=0). We then isolated fluorescent cells by FACS and generated clonal cell lines by serial dilution. To verify in-frame integration of the mNG2(11) construct, we PCR amplified the C terminus *NGFR* locus from cell lysates and cloned the amplicon into a recipient plasmid. Half of the resulting plasmids contained the in-frame mNG2(11) sequence and the remaining half contained the unedited *NGFR* sequence without substitutions, insertions or deletions. Sanger sequencing traces are located here: https://www.dropbox.com/sh/65rvfs3pka5zii4/AAAwBZOynkB4rU9NJvCo_RJha?dl=0. We further confirmed that mNG2 fluorescence correlates with *NGFR* mRNA abundance at the single-cell level by single-molecule RNA FISH and validated that the NGFR-mNG2(11) WM989-A6-G3 cell line recapitulates phenomenology described in Shaffer et al. (2017), showing increased resistance to vemurafenib in the mNG2-high cells (Figure S2; experiments described in RNA FISH, FACS, and drug treatment methods below). We also demonstrated that the localization of the protein was identical to that of the endogenous protein by comparing the NGFR-mNG2 fluorescence to that of the signal produced by immunofluorescent labeling of NGFR (Figure S2). To facilitate cell tracking in the time-lapse images, we transduced cells lines in which we tagged NGFR with mNG2(11) with lentivirus encoding H2B-iRFP670 which localizes to the nucleus, thus enabling us to track cell nuclei. Following transduction, we derived clonal cell lines by serial dilution before imaging. All plasmid sequences are available here: https://www.dropbox.com/sh/wpgyfup6gwiyaeb/AAC0gRgg0dmBJoUlypvSDWnIa?dl=0.

To validate that cells with high levels of NGFR-mNG2 (as measured by fluorescence) were more resistant to vemurafenib, we trypsinized and pelleted the NGFR-mNG2(11) WM989-A6-G3 cell line, washed cells once with PBS containing 2mM EDTA, then resuspended in PBS containing 2mM EDTA and 100ng/mL propidium iodide or 200ng/mL 7-AAD and proceeded with sorting. Using a MoFlo Astrios (Beckman Coulter) or FACSJazz (BD Biosciences), we isolated the top 0.5%–1% of mNG2 fluorescent cells and equal numbers of the bulk population gated only for live cells. We then treated these samples with vemurafenib as described in the drug treatment methods below. All flow cytometry data is available on Dropbox here: https://www.dropbox.com/sh/9bq1eg0k5o0q452/AACunY5g1xtp5lxSIox1OqOka?dl=0.

Fluorescence-Activated Cell Sorting (FACS)

We stained WM989-A6-G3 melanoma cells for fluorescence-activated cell sorting using antibodies for EGFR and NGFR. We note that while we stained for both proteins, we did not isolate enough EGFR-high cells for testing trametinib resistance in all three replicates. First, for EGFR staining, we trypsinized 40-50 million cells, washed once with 0.1% BSA-PBS, and incubated for 1 hour at 4°C with 1:200 mouse anti-EGFR antibody in 0.1% BSA-PBS. Next, we washed with 0.1% BSA-PBS and then incubated for 30 minutes at 4°C with 1:500 donkey anti-mouse IgG labeled with Alexa Fluor 488. We washed the samples again with 0.1% BSA-PBS and then incubated for 10 minutes at 4°C with 1:250 anti-NGFR antibody conjugated directly to PE-Cy7 in 0.5% BSA-PBS with 2mM EDTA. Finally, we washed the samples with 0.5% BSA-PBS containing 2mM EDTA, then resuspended in 1% BSA-PBS containing 100ng/mL propidium iodide or 200ng/mL 7-AAD and proceeded with sorting using a MoFlo Astrios (Beckman Coulter) or FACSJazz (BD Biosciences). To aid with gating, we incubated control samples without the anti-EGFR primary antibody and with a PE/Cy7 mouse IgG1 isotype control. After gating for live cells, we collected either the top 0.02%–0.2% EGFR-high cells or the top 0.5% NGFR-high cells. We also collected equal numbers of the bulk population by using the same gating for live cells, but without gating on either the EGFR or NGFR stains.

To monitor the dynamics of *AXL* expression, we stained WM989-A6-G3 melanoma cells for fluorescence-activated cell sorting by trypsinizing 40-50 million cells, washing once with 1% BSA-PBS, and incubating for 30 minutes at 4°C with 1:50 goat anti-AXL antibody in 1% BSA-PBS. Next, we washed the cells twice with 1% BSA-PBS and then incubated for 30 minutes at 4°C with 1:85 bovine anti-goat IgG labeled with Alexa Fluor 647. Finally, we washed the samples with 1% BSA-PBS, then resuspended in 1% BSA-PBS containing 100ng/mL propidium iodide or 50ng/mL DAPI and proceeded with sorting. After gating for live cells, we collected the top 0.5%–1% AXL-high cells and equal numbers of the lowest 80%–95% AXL-low cells, then plated cells onto glass-bottom chamber plates. After 1, 3, 6 or 9 days in culture, we fixed the sorted cells for RNA FISH as described below. To account for cell growth and changes in cell density, we plated fewer cells for later time points. We performed a similar set of experiments for *EGFR* and *NGFR* using the staining procedure described above. For *NGFR,* we re-stained the sorted population after 7 days in culture (following the same procedure described above) and assessed NGFR intensity by flow cytometry using the same instrument as the initial sort.

To monitor the dynamics of *CA9* expression in MDA-MB-231 cells, we trypsinized cells, washed once with 0.1% BSA-PBS then stained with anti-CA9 antibody conjugated to phycoerythrin (PE) at a dilution of 1:11 in 0.1% BSA-PBS for 30 minutes at 4°C. After staining we washed the cells twice with 0.1% and resuspended in 0.1% BSA-PBS containing 200ng/mL 7-AAD and proceeded with sorting. After gating for live cells, we sorted the top 0.5%–2% CA9-high and the bottom 5%–10% CA9-low cells. For 2 of 3 replicates, we stained the sorted cells with 2.5-5μM CellTraceViolet in PBS at 37°C for 20 minutes, followed by 2 washes with media. We then culture the cells for 5 days before re-staining the cells and measuring CA9-PE and CellTraceViolet intensities by flow cytometry.

For testing the response of CA9-high cells to paclitaxel, we stained MDA-MB-231 cells as described above then sorted the top 2%–3% of CA9-high, the bottom 2%–3% CA9-low, and a mixed population using only the live cell gates (CA9-mix). After allowing

the sorted cells to adhere overnight, we began treatment with 1nM paclitaxel as described below. We performed single-molecule RNA FISH for *CA9* mRNA confirming that sorting with CA9 antibody enriched for CA9-high expressing cells (Figure S1M).

Live-cell immunofluorescence

We stained and sorted AXL-high and EGFR-high cells as described above, then proceeded with live-cell imaging on a Nikon Ti-E inverted microscope enclosed in a temperature-controlled and humidified chamber at 5% CO_2. We acquired an initial set of images measuring Alexa Fluor 488 and Alexa Fluor 647 fluorescence to verify that the cells indeed had high protein levels, then proceeded to scan the slide (20x magnification) in brightfield every hour for 8 days and 16 hours. We then incubated the live cells with 1:80 goat anti-AXL antibody or 1:200 mouse anti-EGFR antibody in TU2% for 1 hour at 4°C followed by two washes with TU2%, secondary incubation with 1:250 bovine anti-goat Alexa647 or 1:250 donkey anti-mouse Alexa488, and two final washes with TU2%. After this re-staining, we imaged these live cells for Alexa Fluor 488 and Alexa Fluor 647 fluorescence at 20x magnification and quantified fluorescence intensity using rajlabimagetools available: https://github.com/arjunrajlaboratory/rajlabimagetools and custom MATLAB scripts available at https://www.dropbox.com/sh/zrc0g9vtxewcctl/AABVYCaXHYyvWFPOx4ipXLJGa?dl=0.

Drug treatment experiments

We made stock solutions of 4mM trametinib, 4mM vemurafenib, and 4mM paclitaxel. For drug treatment experiments, we diluted the stock solutions in culture medium to a final concentration of 10nM for trametinib, 1μM for vemurafenib, and 1nM for paclitaxel. For trametinib treatment experiments, we sorted WM989-A6 subclone G3 by NGFR levels as described in the FACS section of Methods and then cultured them for 2-3 weeks. For vemurafenib experiments, we cultured the FACS sorted NGFR-mNG2 WM989-A6-G3 in vemurafenib for 21 days. For paclitaxel experiments, we cultured CA9 FACS sorted MDA-MB-231 cells in paclitaxel for 5 days. At the end of all drug treatment regimens, we fixed each sample in 4% formaldehyde for 10 minutes, permeabilized the sample in 70% ethanol, and then performed cell quantification.

Time-lapse imaging

We acquired time-lapse images of the NGFR-mNG2 WM989-A6-G3 cell line using two different imaging platforms. First, for data of the WM989 NGFR-mNG2 cell line growing without drug, we used a Nikon Ti-E microscope encased in a plexiglass chamber ventilated with heated air and CO_2. We took images at 60x magnification of mNG2 fluorescence every 6 hours and images of the iRFP nuclear reporter (H2B-iRFP670) every hour for 8.75 days. We chose these time intervals based on pilot experiments we performed to minimize overt signs of phototoxicity (cell death, growth inhibition, nuclear morphology changes) and enable the tracking of cell lineages. Second, for data in which we cultured these cells and then treated with vemurafenib, we used an IncuCyte S3 Live Cell Imaging Analysis System (Sartorius). We cultured the NGFR-mNG2 cell line on a 96-well plate inside the IncuCyte, which is fully contained within an incubator for long-term culture and time-lapse microscopy. With this system, we acquired images in green, red, and brighfield using a 10X objective at intervals of 2 hours over a total of 14.8 days. We added 500nM of vemurafenib after 6 days and 4 hours in culture, and then changed the media with vemurafenib every 3 days. We used these two different imaging platforms for their distinct advantages. The high magnification Nikon system allowed for the most accurate quantification of the mNeonGreen2 signal, allowing us to measure the length of time of the NGFR fluctuations. Meanwhile, the IncuCyte is a more stable environment for time-lapse imaging that therefore induced less stress on the cells, and thus we used this platform for the longer experiments, particularly involving the additional stress of drug treatment.

RNA FISH

We designed custom oligonucleotide probe sets complementary to our genes of interest using custom probe design software written in MATLAB (code freely available for non-commercial use here https://flintbox.com/public/project/50547/) and ordered them with an amine group on the 3′ end from Biosearch Technologies (probe sequences available in Table S2). We pooled 15-32 oligonucleotides targeting each gene and coupled each set of probes to either Cy3, Alexa Fluor 594, Atto 647N or Atto 700. We performed single-molecule RNA FISH as described in Raj and van Oudenaarden (2008) for multiple cycles of hybridization (Shaffer et al., 2017). We fixed cells in 4% formaldehyde solution for 10 minutes at room temperature, permeabilized in 70% ethanol, and stored samples at 4C. For hybridization, we first washed samples with washing buffer containing 10% formamide and 2x SSC. We then applied hybridization buffer containing custom RNA FISH probes and 10% formamide, 2x SSC, and 10% dextran sulfate. We hybridized samples overnight at 37°C and then performed two cycles of 30 minute washes at 37°C with washing buffer. For imaging, we first DAPI stained the cells and then transferred them to 2x SSC.

RNA FISH imaging

We imaged RNA FISH samples on an inverted Nikon TI-E microscope with a 60x Plan-Apo or a 100x Plan-Apo using filter sets for DAPI, Cy3, Atto647N, Alexa594, and Atto700. We took images in either z stacks of 30 planes at 0.3μm intervals using custom journals built in Metamorph or tiled grids of single-plane images using Metamorph Scan Slide Application. We used a Nikon Perfect Focus system to maintain focus across the imaging area.

Please cite this article in press as: Shaffer et al., Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors, Cell (2020), https://doi.org/10.1016/j.cell.2020.07.003

QUANTIFICATION AND STATISTICAL ANALYSIS

Cell quantification

We quantified cell numbers for drug treatment experiments by fixing the cells, staining with DAPI, then imaging across the majority of the well via image scanning at 20x magnification. After scanning, we computationally stitched the images together, after which we used custom software written in MATLAB to identify nuclei, which is publicly available here: https://github.com/arjunrajlaboratory/colonycounting_v2.

Time-lapse analysis

For tracking cell lineages and quantifying fluorescence signal of the NGFR-mNG2 WM989-A6-G3 cell line, we developed a set of publicly available tools for tracking cells in time-lapse images (https://github.com/arjunrajlaboratory/timelapseGUI). First, this pipeline uses an automated algorithm for nuclear segmentation to identify all the cells in each image. We then used a combination of automatic assignment of parents along with human-supervised annotation to fix errors to obtain lineage information. Next, we quantify the fluorescence signal from the NGFR-mNG2 by using the nuclear segment from each cell and calculating mean and median fluorescence intensity across these segments (Figure S2). Of note, for the time-lapses of AXL-high and EGFR-high sorted cells that lacked the H2B-iRFP670 nuclear marker, we had to manually mark all the cells that we wanted to analyze because the lack of nuclear markers precluded automatic segmentation. This same method of analysis was applied to all data acquired on the IncuCyte platform. Our subsequent analysis consisted of building a custom data structure in MATLAB to contain each lineage and a series of plotting functions to allow us to plot the fluorescence intensity from an entire lineage (or part of a lineage) over the length of these experiments. For the lineages derived from AXL-high and EGFR-high sorted cells, the final frames were manually registered to images acquired after repeated immunofluorescence staining of live cells. The code for all the downstream processing is available on the Dropbox here: https://www.dropbox.com/sh/q2kbatojibljg8j/AABiFvNvsQm288tC3n-4qn5Za?dl=0 and https://www.dropbox.com/sh/zrc0g9vtxewcctl/AABVYCaXHYyvWFPOx4ipXLJGa?dl=0.

RNA FISH image analysis

For analysis of gridded image scans, we used custom MATLAB software designed for the analysis in (Shaffer et al., 2017). This pipeline consists of first segmenting the nuclei of individual cells using DAPI images. Next, the software calculates regional maxima for all RNA FISH dyes and then the user specifies a global threshold for calling individual spots. Through a GUI interface the user then reviews the high expressing cells and uses editing tools to remove artifacts or autofluorescent debris. Lastly, we constructed tables containing RNA FISH spot counts for each gene in individual cells.

For image z stacks, we used custom MATLAB software to count spots for each cell. Briefly, this image analysis pipeline includes manual segmentation of cell boundaries, thresholding of each channel in each cell to identify individual spots, and then extraction of spot counts for each gene in each cell. The software for analysis of RNA FISH images is available on bitbucket here https://github.com/arjunrajlaboratory/rajlabimagetools. After extracting spot counts from either data format, we performed the remainder of the analysis of mRNA distributions in R. We calculated Gini coefficients (as described in Jiang et al., 2016) for each gene using the "ineq" package. The code for this analysis is on the Dropbox here https://www.dropbox.com/s/n8cppr17b3bgssf/gini_coef_analysis.R?dl=0.

Spatial analysis

We used spatial single-cell analysis to enable us to independently measure the heritability of high expression states. We sparsely plated cells (WM989-A6, MDA-MB-231-D4, WM983B) on a 2-well chambered coverglass, and then we allowed the cells to grow for 10 days (sometimes fewer days for MDA-MB-231-D4 if the cells grew faster), at which point we performed iterative RNA FISH, image analysis, and thresholding for high expression as described above. Intuitively, the stronger the heritability (i.e., over several generations), the larger the clusters of high-expressing cells we would find. To quantify clustering, we created, for each cell in the dataset, a "bucket" consisting of the 20 (or 50, 100, 200) closest cells and then kept track of the number of high-expressing cells in that bucket. We then computed the variance and the mean of this number across all buckets, allowing us to then calculate the heritability index, which is the Fano factor (defined as the variance divided by the mean). In the case of complete spatial randomness, the distribution would be Poisson, and the heritability index would be 1. To verify this null distribution, we permuted the label of cells as jackpots or non-jackpots uniformly at random 1000 times, and re-computed the heritability index for each permutation. This approach allowed us to compute 95% confidence intervals for the null distribution given our particular spatial configuration of cells; the data for null distributions is not shown, but is available online here: https://www.dropbox.com/sh/m1rdxjrfeo6be3m/AADphkUBGT0zgpppssP_LbSa?dl=0.

ATAC sequencing analysis

Using previously published ATAC sequencing data from EGFR-high sorted WM989-A6 cells, we identified peaks within each of the 227 WM989 MemorySeq genes (including 10 kb upstream and 10 kb downstream) (Shaffer et al., 2017). We included peaks present in 2 of 2 biological replicates and merged overlapping regions with a point-source peak to peak distance of less than 500 bp. For comparison, we used these same data to identify peaks within TPM-matched, control gene sets. We generated these control gene sets by

taking the expression of each MemorySeq gene (average TPM across MemorySeq clones) and randomly sampling a gene with similar expression levels (+/− 20%). We included MemorySeq genes in the sampling set and sampled without replacement to obtain gene sets with equal numbers of unique genes. We repeated this sampling procedure 100 times to generate 100 TPM-matched, control gene sets.

We then used the Analysis of Motif Enrichment (AME) tool from the MEME suite to identify enriched motifs underlying peaks near MemorySeq genes compared to each of the control gene sets (McLeay and Bailey, 2010). For AME, we used the average odds score and Fisher's exact test to measure motif enrichment and the HOCOMOCO Human v11 CORE database match enriched motifs to transcription factors (Kulakovskiy et al., 2018). We then summarized and plotted results using R (v3.6). All datasets are available at http://www.dropbox.com/login?cont=https%3A%2F%2Fwww.dropbox.com%2Fwork%2FPapers%2Fcancerpaper_public%2Fdata%2FATACseq and all scripts are available at http://www.dropbox.com/login?cont=https%3A%2F%2Fwww.dropbox.com%2Fwork%2FPapers%2Fmemoryseq%2Fcode%2FATACanalysis.

RNA sequencing analysis of WM989-A6-G3 knockouts

We accessed RNA sequencing count data for 266 CRISPR/Cas9 knockouts of WM989-A6-G3 (Torre et al., 2019). These data included knockouts for 84 genes (2-6 replicates using distinct sgRNAs for each) and 10 control samples transduced with non-targeting sgRNAs. For each gene we used DESeq2 to calculate $\log_2$ fold change in expression compared to the non-targeting control samples(Love et al., 2014). The RNA sequencing count data can be found https://www.dropbox.com/sh/zold8icu30h6njc/AAC4BJOlz-uWGU0DU5SOcDzYa?dl=0 and scripts used for analysis can be found https://www.dropbox.com/sh/wtbk6j0umxgovwi/AABbzGg7sIuqh2I_umU4nRpSa?dl=0 and https://www.dropbox.com/s/7ly259ou1fm8lsk/plotKOenrichment.R?dl=0

Allele-specific expression (ASE) analysis

To identify variants in MemorySeq data we first aligned the RNA sequencing reads to hg38 (reference available at https://www.dropbox.com/sh/dpxfu9vowd9rgxw/AABUd3d1_Eg3VgWEtCU45CCja?dl=0) using STAR (v2.7.1a), then filtered unique alignments, and marked duplicate alignments (Dobin et al., 2013). Next, we combined alignments from all MemorySeq clones and used freebayes (v1.3.2) to identify biallelic variants with hg38 as the reference (Garrison and Marth, 2012). We repeated this procedure separately for the 48 control clones (see MemorySeq Methods above and Figure 1). Using bcftools, we filtered heterozygous variants with a minimum quality score of 30 and minimum depth of 5 reads then took the intersection of variants identified in both MemorySeq clones and control clones to remove mutations present in a small subset of clones (Narasimhan et al., 2016). All scripts for this analysis are available at https://github.com/arjunrajlaboratory/memSeqASEanalysis and the final list of variants is available on dropbox at https://www.dropbox.com/sh/yinthy6f2u7pwwu/AACqf_4WSFow6njp8xzzTRYca?dl=0 (under {cell line}/freebayes/{cell line}.singleRG.-freebayes.intersect.Q30.DP5.het/0002.*vcf*.gz). With these lists of variants, we used phASER to count uniquely mapped, allele specific reads for each individual clone (Castel et al., 2016). We then summarized and plotted these data using custom R scripts available at https://www.dropbox.com/sh/l607yt9vu4dlr3n/AAAofMByQ4KYTdQWm6zlxj_Va?dl=0.

ChIP sequencing analysis

We downloaded H3K27 acetylation (H3K27Ac) and tri-methylation (H3K27me3) ChIP-sequencing coverage data and matched RNA sequencing data from 11 melanoma cell lines from (Verfaillie et al., 2015). With the RNA sequencing data, we generated TPM-matched control gene sets by randomly sampling genes whose expression matched each MemorySeq gene (+/− 20%). MemorySeq genes were included in the sampling set and we sampled without replacement to obtain gene sets with equal numbers of unique genes. For both WM989 and WM983b MemorySeq genes we separately generated 100 control gene sets using each of the 11 melanoma samples. For each of the ChIP-sequencing datasets, we then used deepTools to calculate the ChIP-sequencing coverage flanking the transcription start site (+/− 2,000 bp in 10 bp bins) for MemorySeq genes and TPM-matched control gene sets (Ramírez et al., 2016).

For MDA-MB231 cells, we downloaded H3K4 trimethylation (H3K4me3), H3K9 acetylation (H3K9Ac), H3K79 dimethylation (H3K79me2), H3K27Ac and H3K27me3 ChIP sequencing coverage data and matched RNA sequencing data from (Franco et al., 2018; Xi et al., 2018). As described for the Memory Seq analysis, we aligned the RNA sequencing data to hg19 using STAR and counted alignments using HTSeq. We generated TPM-matched control gene sets and calculated ChIP-sequencing coverage flanking the transcription start sites of MemorySeq and control gene sets as described above. We then summarized and plotted these data using R. All datasets are available at https://www.dropbox.com/sh/mjw5n5zeau9lpwb/AAD6uCZ4fFU8c36izOJ35B9aa?dl=0 and analysis scripts are available at https://www.dropbox.com/sh/zr6f9rggepmhx8y/AACzfQzAVvWFiKgGQ7BOatzxa?dl=0 and https://www.dropbox.com/sh/mwmq7kegnw6h4yq/AAD0f1YZZip0tyy7p9R2eE8va?dl=0.

We used the RefSeq hg19 annotations for determining transcription start sites rather than the Ensembl annotations used for RNA sequencing as we observed better aligned coverage profiles across genes using RefSeq annotations. However, ~7% of MemorySeq genes (11/227 for WM989, 16/230 for WM983b and 15/230 for MDA-MB231) were not uniquely identified in the RefSeq annotation and were excluded from this analysis, including the initial sampling procedure. We observed lower levels of activating chromatin marks and slightly elevated levels of repressive marks near MemorySeq genes using either RefSeq or Ensembl transcription start site annotations. Both reference files are available at https://www.dropbox.com/sh/sf93jf48p7u1q0e/AABt6Od4AeOMABFb6pl9f881a?dl=0.

Network community identification

For each pair of significantly heritable genes (from gene lists described in the MemorySeq RNA sequencing analysis section of the Methods), we calculated the Pearson correlation coefficient between their expression across clones. This procedure resulted in a symmetric weighted matrix of size 227 genes x 227 genes in WM989 (and 230 genes x 230 genes in WM983B, as well as MDA-MB-231-D4). We represent these matrices as undirected weighted networks with nodes corresponding to genes, and with the weight of edges between nodes corresponding to the value of the correlation coefficient. Within this network, we sought tightly connected groups of nodes within this network known as *network communities*. We performed *k*-clique community detection (Derényi et al., 2005; Jonsson et al., 2006; Palla et al., 2005) with k = 4 on binarized networks created by thresholding the original weighted network at decreasing values (Giusti et al., 2015; Rieck et al., 2018). More specifically, a *k*-clique is a collection of *k* nodes that are all-to-all connected and a *k*-clique community is a collection of *k*-cliques that are connected through adjacent *k*-cliques (two *k*-cliques are adjacent if they share *k-1* nodes). Repeatedly thresholding the network at decreasing values of the edge weight creates a sequence of binary graphs, each of which is included in the next. Since the addition of edges to a binary graph can only enlarge or merge *k*-clique communities, we can track communities from one binary network to the next in a well-defined manner. This mapping allows us to both observe which nodes were included in the community at slightly lower threshold values and to qualitatively assign statistical significance to communities based on the range of threshold values for which they stay isolated from the rest of the network. All networks shown in figures were visualized with Gephi (Bastian et al., 2009).

Inferring timescale of expression heritability from MemorySeq

We modeled gene expression as a binary switch, where individual cells were either in an high (ON) or low (OFF) expression state. We assumed that cells proliferated exponentially at a rate k_x (i.e., a generation time of $1/k_x$), and that each MemorySeq clone began as a single cell, ultimately growing into the final population. We made two simplifying assumptions: i) The proliferation rate of a cell was the same irrespective of its ON/OFF state; and ii) The population remained in the exponential growth phase during the time span of the experiment. Further, cells in the OFF state turned ON with rate k_{ON}, and reverted back to the OFF state with rate k_{OFF}. We defined $f = k_{ON}/(k_{ON} + k_{OFF})$ as the fraction of ON cells in the original population, and further assumed (as indicated by our experimental data) that the ON state was rare, i.e., $f << 1$.

To model Luria-Delbrück fluctuation analysis, we sampled a single cell from the original population, which was ON with probability f, and OFF with probability $1 - f$. Starting with this initial condition, we defined the random variables $\boldsymbol{x}(\boldsymbol{t})$and $\boldsymbol{y}(\boldsymbol{t})$ to be the number of cells in the ON and OFF states, respectively, at time t. The time evolution of integer-valued random processes $\boldsymbol{x}(\boldsymbol{t})$and $\boldsymbol{y}(\boldsymbol{t})$is governed by proliferation (with probabilities $k_x xdt$ and $k_x ydt$) and switching between the ON and OFF states (with probabilities $k_{OFF} xdt$ and $k_{ON} ydt$). The ratio $\boldsymbol{x}(t)/(\boldsymbol{x}(\boldsymbol{t}) + \boldsymbol{y}(\boldsymbol{t}))$ represented the fraction of ON cells at time t, and our goal was to quantify fluctuations in this ratio across MemorySeq clones. To do this, we derived the time evolution of the first two statistical moments of $\boldsymbol{x}(\boldsymbol{t})$and $\boldsymbol{y}(\boldsymbol{t})$ (Singh and Hespanha, 2010). Assuming $f << 1$, the coefficient of variation squared (CV^2) of the ratio $\boldsymbol{x}(\boldsymbol{t})/(\boldsymbol{x}(\boldsymbol{t}) + \boldsymbol{y}(\boldsymbol{t}))$ is given by Equation (1):

$$CV^2 \times f = \frac{2T_{ON} e^{\left(T - \frac{2T}{T_{ON}}\right)} - 2 - T_{ON}}{(2e^T - 1)(T_{ON} - 2)},$$

where $T = tk_x$ is the duration of the experiment (normalized to the generation time), and $T_{ON} = k_x/k_{OFF}$is the time spent in the ON state (normalized to the generation time). This equation quantifies the noise measured in MemorySeq, and as expected, CV^2 is a monotonically increasing function ofT_{ON}; i.e., slower switching results in higher variation across MemorySeq clones. It turns out that the product $CV^2 \times f$ is independent of f, and it is convenient to look at this product as a function of T_{ON}. Given *a priori* knowledge of f and a measurement of the noise level CV^2, T_{ON} can be estimated via an inverse transformation of Equation (1). In our experiments, f was typically around 1% or less, as measured using RNA FISH, and each MemorySeq clone started as a single cell and grew to around 10^5 cells, for which the number of doublings T = 17. Using these values, we estimated T_{ON} of 5-10 cell divisions for several MemorySeq genes based on the observed CV values across MemorySeq clones. Further discussion of this model can be found at https://www.dropbox.com/s/n34mz3ctk1w5xu8/SupplementalNote1.pdf?dl=0.

STATISTICAL ANALYSIS

For each experiment, the sample sizes, number of replicates, and randomization procedure (when relevant) are indicated in the main text and figure legends. Statistical tests for calculating P values and thresholds for judging significance are indicated in the figure legends.

CellPress

Supplemental Figures

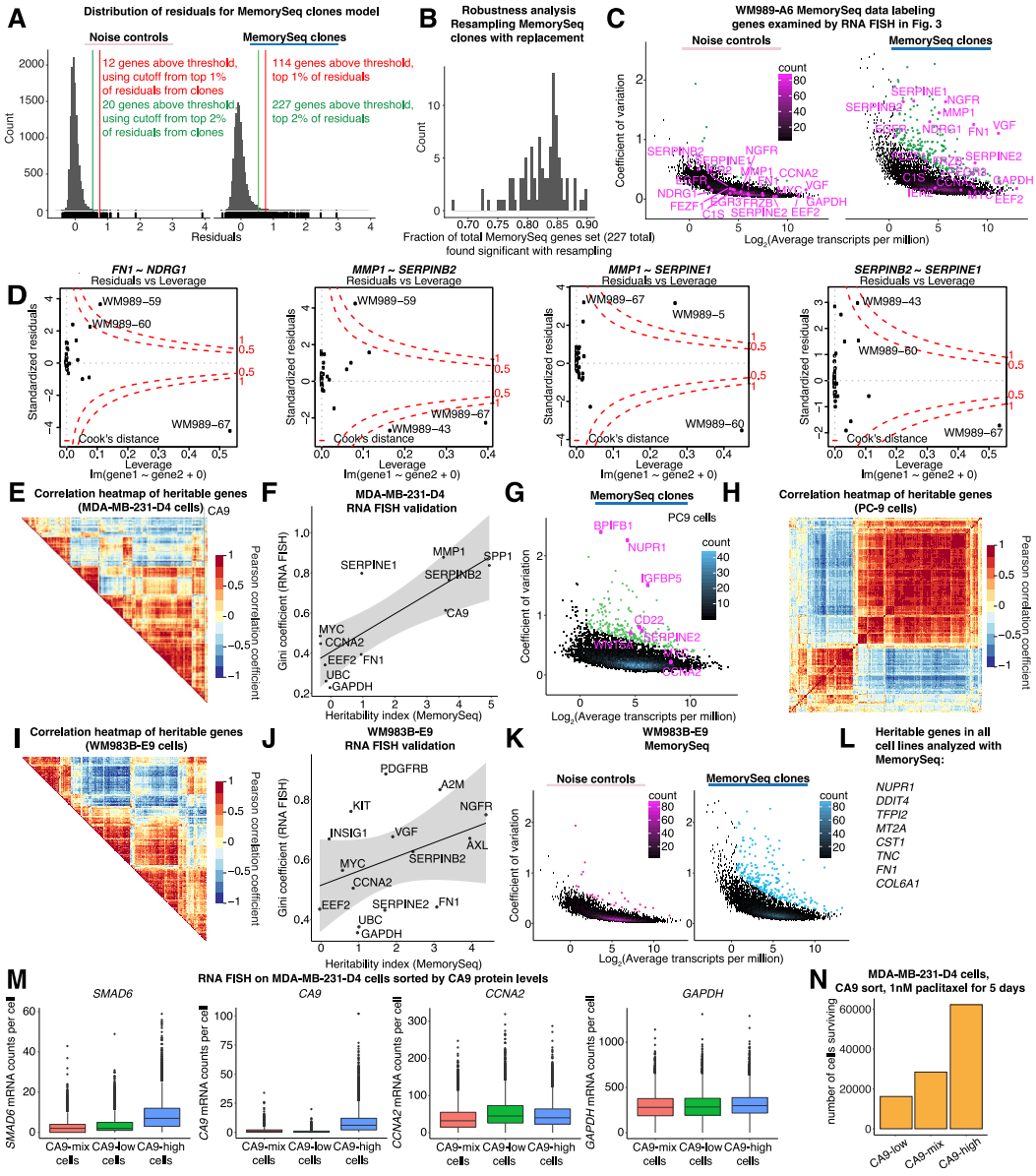


Figure S1. MemorySeq Identifies Heritable Genes across Multiple Cell Lines, Related to Figure 1

A. Histograms showing the overall distribution of the residuals for MemorySeq clones and Noise controls in WM989-A6 cells. B. Robustness analysis showing the fraction of MemorySeq genes discovered by resampling the clones with replacement over 100 iterations. The plot shows the histogram of the fraction of genes overlapping with the core set of 227 through each iteration. C. Same plot as Figure 1D, except the genes labeled in pink are the set selected for RNA FISH in Figure 3. D. Cook's distance plots for 4 gene pairs showing the strongest correlations in the MemorySeq data. Each plot depicts the residuals versus leverage for the regression with the gene pair. Red dotted lines delineate the cut-offs for Cook's distance > 0.5 and > 1. E. Correlation heatmap for all pairs of heritable genes in MDA-MB-231-D4 MemorySeq. F. Plot of heritability index from MemorySeq versus Gini coefficient from RNA FISH for MDA-MB-231-D4 (replicate RNA FISH data available on Dropbox). G. Plot of coefficient of variation versus the $\log_2$ mean expression for MemorySeq on PC-9 cells. H. Correlation heatmap for all pairs of heritable genes in PC-9 MemorySeq. I. Correlation heatmap for all pairs of heritable genes in WM983B-E9 MemorySeq. (J). Plot of heritability index from MemorySeq versus Gini coefficient from RNA FISH for WM983B-E9 (replicate RNA FISH data available on Dropbox). K. Plot of coefficient of variation versus the log2 mean expression for WM983B-E9 cells with Noise controls on left and MemorySeq clones on right. L. List of genes found to be heritable in MemorySeq across WM989-A6, WM983B-E9, MDA-MB-231-D4, and PC-9 cell lines. M. We sorted CA9-stained cells into high- and low-expressing subpopulations (as determined by antibody staining) as well as the unsorted mixed population and performed RNA FISH for *SMAD6*, *CA9*, *CCNA2*, and *GAPDH*. Boxplots show counts per cell for each gene from each of the samples sorted based on CA9 protein labels. **N.** Replicate bar graph for data shown in Figure 4D.

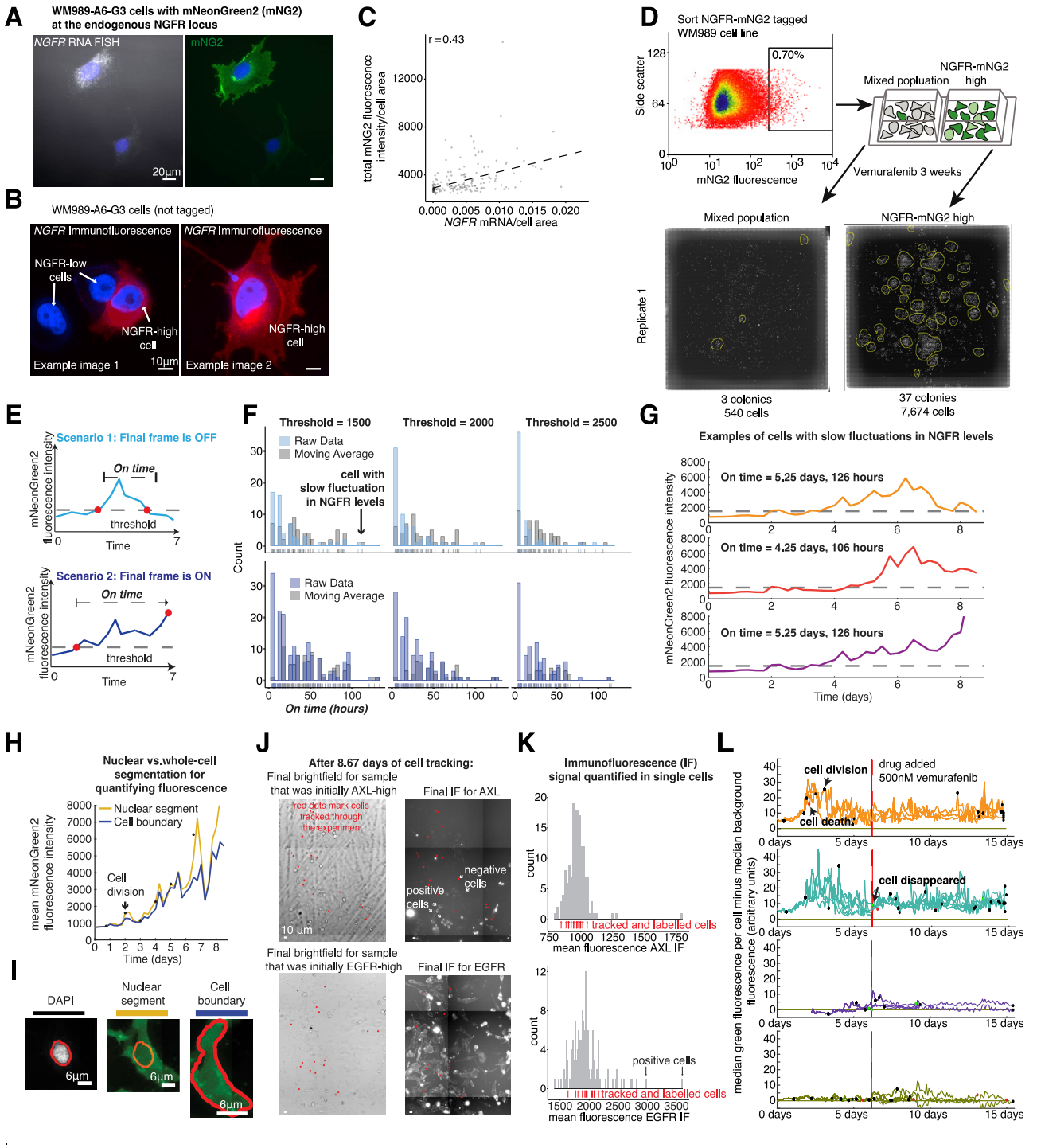


Figure S2. Time-Lapse Imaging Directly Shows that WM989 Cells Have Long-Lived Fluctuations in Expression of Genes Found with MemorySeq, Related to Figure 2

A-C. RNA FISH and immunofluorescence of endogenous *NGFR* mRNA and protein in the NGFR-mNeonGreen2 (NGFR-mNG2) tagged WM989-A6-G3 cell line showing that mNG2 fluorescence correlates with *NGFR* expression and protein localization. D. We sorted cells based on mNG2 fluorescence from the indicated gates, treated cell populations with vemurafenib for 3 weeks, then imaged the wells to quantify resistant colonies. Stitched whole well images from 3 additional replicates are available on Dropbox. E. We performed time-lapse imaging of WM989-A6-G3 cells tagged with NGFR-mNG2 and tracked cells to quantify the length of time that cells remain in the NGFR-high state. Cartoon shows two possible outcomes from cells that become NGFR-high. F. Histograms of the length of time that individual cells remain in the NGFR-high state using three different thresholds for determining when a cell is NGFR-high. The blue (both dark and light

(legend continued on next page)

blue) show the raw data for the on time, while the gray shows the on times that result from using a moving average of the time trace using a 5 point median filter. At a threshold of 1500, the average on time is 40 hours for scenario 1 (light blue) and 52 hours for scenario 2 (dark blue). G. Plots of mNG2 fluorescence intensity over time for three example cells that enter the NGFR-high state and remain high for 4.25 to 5.25 days, thereby demonstrating memory. H-I. Comparison of nuclear segmentation versus whole segmentation for quantifying mNG2 fluorescence. Images are subfield views taken from scans comprising several image tiles; boundaries in the image arise from edges between individual tiles. J-K. Tiled brightfield and immunofluorescence image scans of WM989-A6-G3 after tracking by timelapse for 8.67 days. Histograms show the immunofluorescence signal intensity in 159 cells (top; AXL IF) and 155 cells (bottom; EGFR IF) after timelapse. Descendents of the AXL-high or EGFR-high cells in Figure 2D are labeled with red points in the images and their immunofluorescence levels are indicated by red arrows in the histogram. L. Time-lapse imaging of the WM989-A6-G3 NGFR-mNG2 cell line treated with 500nM vemurafenib. Plot of fluorescence intensity versus time for four lineages tracked through the experiment. Black dots label cell division, red dots label cell death, and the green dots label when a cell disappears from the field of view. Related to Videos S2 and S3.

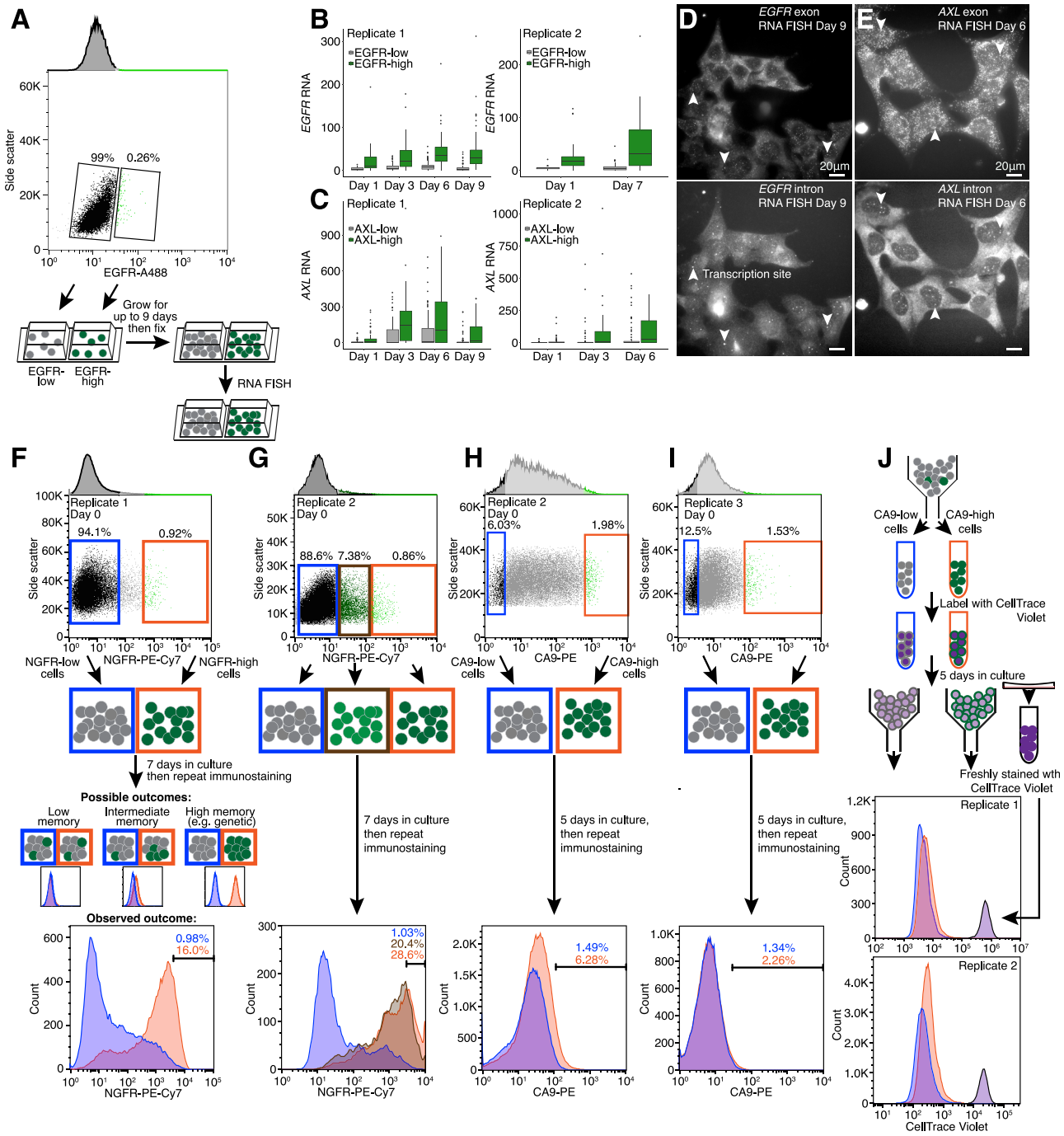


Figure S3. Expression of MemorySeq Genes Is Maintained over Multiple Cell Divisions but Ultimately Reversible, Related to Figure 4

A-C. We sorted AXL-high and EGFR-high WM989-A6-G3 cells into multiple 2-well plates then fixed the separate plates at multiple time points up to 9 days after sorting. We then performed RNA FISH to measure single-cell expression of *AXL* and *EGFR*. Boxplots show the single-cell expression of *AXL* and *EGFR* in the indicated sorted populations. D-E. Representative RNA FISH images showing clusters of cells expressing high levels of *EGFR* or *AXL*. Arrowheads mark transcription sites identified by colocalization of exon-targeting and intron-targeting RNA FISH probes. The presence of multiple transcription sites in these cells suggested that the observed gene expression memory is due to persistence of active transcription rather than slower cell division or RNA degradation rates. F-G. We sorted equal numbers of NGFR-high and NGFR-low W989-A6-G3 cells from the indicated gates then cultured the cells for 7 days before re-labeling with an antibody targeting NGFR and measuring immunofluorescence intensity by flow cytometry. At day 7, the distribution of immunofluorescence intensities is higher in the NGFR-high samples relative to the NGFR-low samples, although the distributions are closer than at day 0 (2 of 2 replicates shown). H-I. Two additional

(legend continued on next page)

biological replicates of the data presented in Figure 4C. J. We monitored cell division in the CA9-sorted MDA-MB-231-D4 by staining cells with CellTrace Violet (CTV) immediately after sorting, then measuring CTV fluorescence by flow after 5 days in culture. For reference, we freshly stained equal numbers of cells on day 5 to flow in parallel. As shown, both CA9-high and CA9-low sorted samples showed lower CTV fluorescence than freshly stained cells consistent with multiple cell divisions (also noted by seeing clusters of cells in culture). We note that the CA9-high sample retained slightly more of the CTV dye than the CA9-low sample, which may reflect small differences in cell division, cell size or protein turnover.

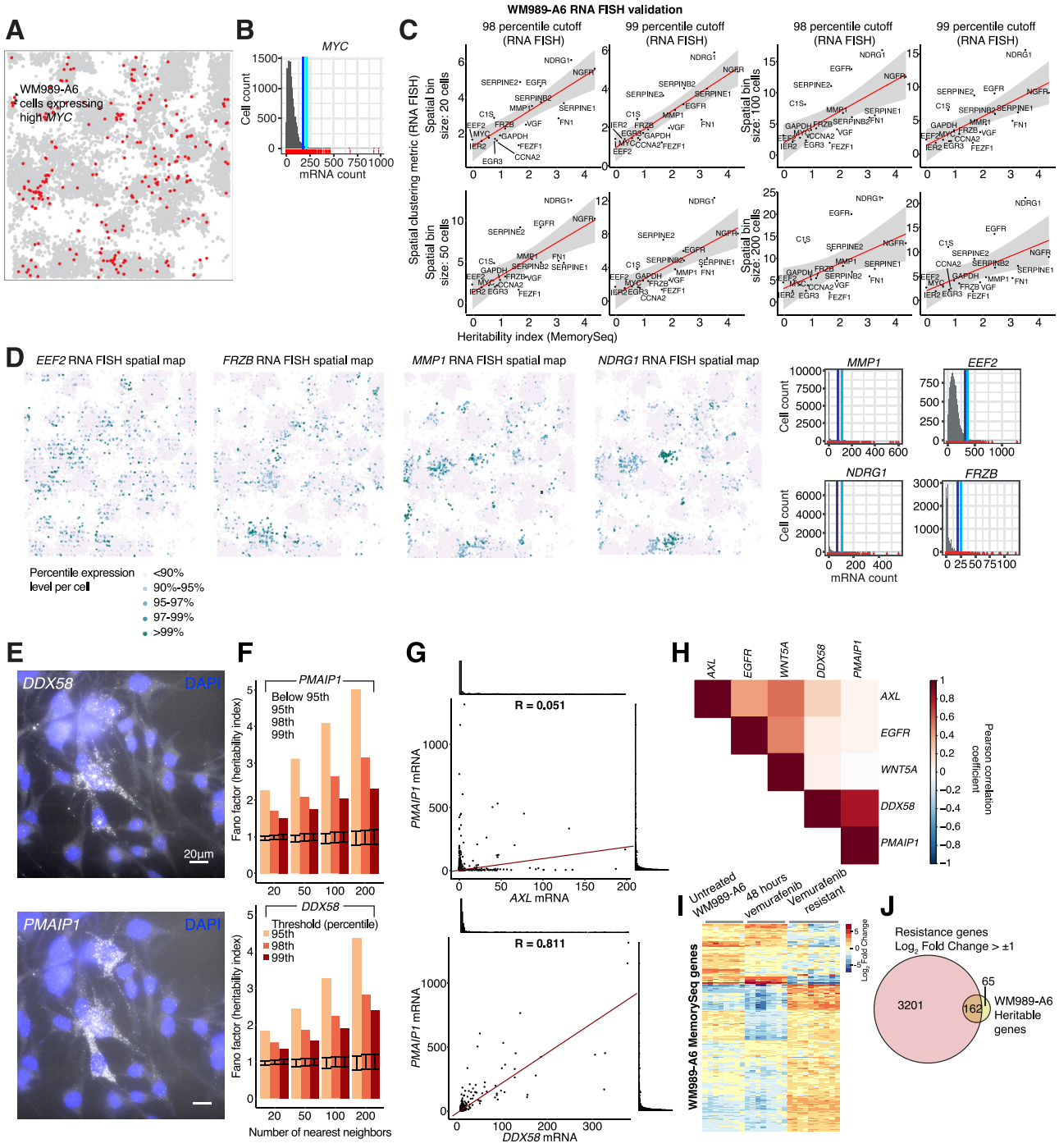


Figure S4. RNA FISH Reveals Spatial Patterns of Gene Expression Memory, Related to Figure 3

A. We measured expression of *MYC* in WM989-A6 cells grown on culture dishes for 10 days. Spatial position of each cell in culture with the top 2% of *MYC* expressing cells in red and the rest of cells labeled in gray. B. Histogram of *MYC* mRNA levels across individual cells, showing the large amount of variability in MYC expression. The rightmost line (light blue) marks the top 1% of cells, the left line (dark blue) marks the top 2%. C. Related to Figure 3C. We analyzed different cutoffs for cells to be considered to be in the high-expression-level state (either the top 1 percent of cells or the top 2 percent of cells). We also increased the bin size for each neighborhood for analysis, which is quantified as the number of neighboring cells included in each bin. We quantified the skewness across MemorySeq clones versus the spatial clustering metric, which is the Fano factor (variance/mean) in the number of positive cells per bin across all bins of indicated size. D. Spatial maps and histograms for RNA FISH experiments in Figure 3A. Data is shown for 4 example genes (equivalent plots for all measured genes are available on Dropbox). E. Using RNA FISH in WM989-A6-G3, we find that *DDX58* and *PMAIP1* expression is correlated in single cells, but both genes are far less correlated with the expression of *AXL*, *EGFR* or *WNT5A*. Shown are representative images (maximum Z-projection) of a group of cells co-expressing high levels of

(legend continued on next page)

DDX58 and *PMAIP1*. F. We observe spatial clustering of cells expressing *DDX58* and *PMAIP1* as calculated by Fano factor. Error bars correspond to the 95% confidence interval of the randomly permuted dataset. Spatial maps as in D are available on Dropbox. G. Scatterplots demonstrating correlation between *PMAIP1* and *DDX58* expression, as well as minimal correlation between *PMAIP1* and *AXL* expression. Additional pairwise comparisons including *EGFR* and *WNT5A* are available on Dropbox. H. Heatmap of the measured Pearson correlation values from RNA FISH across 362 cells. I. Heatmap shows the differential expression of MemorySeq genes (rows) in untreated WM989-A6 cells versus cells treated with vemurafenib. Color indicates the $\log_2$ fold change in reads per million (rpm + 0.1) versus the mean reads per million across all conditions. J. Venn diagram shows the overlap between heritable genes identified in WM989-A6 using MemorySeq and differentially expressed genes in vemurafenib-resistant versus drug-naive WM989-A6 cells. Replicate data for A-D are available on Dropbox.

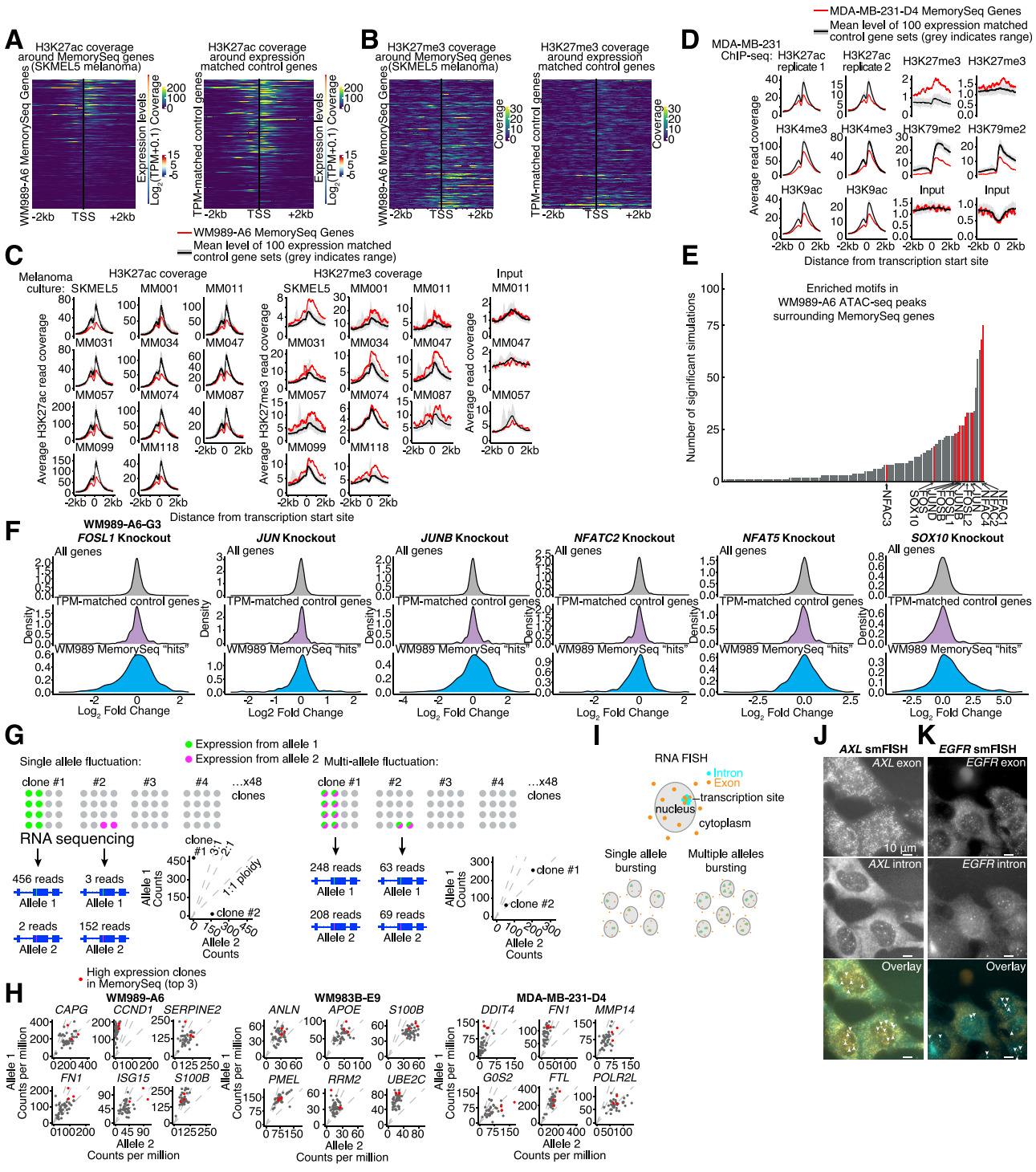


Figure S5. MemorySeq Gene Fluctuations Are Associated with Lower Levels of Activating Chromatin Marks and Simultaneous Transcription from Multiple Alleles, and Can Be Regulated by FOS/JUN, NFAC, and SOX10 Transcription Factors, Related to Figure 5

A-B. H3K27 acetylation (H3K27Ac) and tri-methylation (H3K27me3) ChIP sequencing coverage surrounding the transcription start site (TSS ± 2,000 bp) for WM989-A6 MemorySeq genes and expression matched control genes. Data for SKMEL5 cell line. C. Average H3K27Ac and H3K27me3 ChIP sequencing coverage surrounding the TSS for WM989-A6 MemorySeq genes and 100 expression-matched control gene sets in 11 different melanoma cultures. D. Average ChIP sequencing coverage surrounding the TSS of MDA-MB-231-D4 MemorySeq genes and 100 expression-matched control gene sets for 5 histone marks assayed in MDA-MB-231. Similar plots for 6 additional marks available on Dropbox. E. Comparison of WM989-A6 ATAC sequencing peaks across MemorySeq genes (+/− 10,000 bp) to peaks across 100 separate expression-matched gene sets identified 148 transcription factors enriched in at least one comparison

(legend continued on next page)

(adjusted p value $\leq$ 0.025). Plot shows the number of comparisons enriched for each motif. Motifs with corresponding transcription factor knockout data in F are colored red. F. Log_2 fold change in gene expression of MemorySeq genes (blue), a randomly sampled, control gene set (purple), and all genes included in the MemorySeq analysis (gray; mean RPM $\geq$ 1 across MemorySeq clones) in the indicated WM989-A6-G3 CRISPR/Cas9-knockouts measured by RNA sequencing. G. Model for MemorySeq gene expression fluctuations driven by transcription from a single allele versus multiples alleles. H. We identified heterozygous variants in 202 MemorySeq genes for allele-specific expression analysis (54 in WM989-A6, 79 in WM983B-E9 and 69 in MDA-MB-231-D4). Plotted are the normalized, allele-specific read counts across MemorySeq clones for the 6 MemorySeq genes in each cell line with the highest RNA sequencing coverage of heterozygous variants. Comparison of allelic expression showed that the highest expressing MemorySeq clones (red; based on total expression from corresponding genes) tended to have higher expression from both alleles relative to the majority of other clones, consistent with regulation by a *trans*-acting factor. Equivalent plots for all MemorySeq genes with heterozygous variants are available on Dropbox. I. Single-cell model for MemorySeq gene expression fluctuations driven by transcription from a single allele versus multiples alleles. J-K. Images of RNA FISH for *AXL* and *EGFR* using both intron and exon probes labeled with distinguishable fluorescent dyes. Co-localization of FISH signals demonstrates multiple transcription sites in single cells. These cells are the same as AXL-high and EGFR-high sorted WM989-A6-G3 cells presented in Figure S3.