Title:

Memory sequencing reveals heritable single cell gene expression programs associated with distinct cellular behaviors

Authors:
Sydney M. Shaffer[1], Benjamin L. Emert[2], Ann E. Sizemore[1], Rohit Gupte[1], Eduardo Torre[2], Danielle S. Bassett[1,4,5,6], Arjun Raj[1,3]

[1]Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania
[2]Perelman School of Medicine, University of Pennsylvania
[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania
[4]Department of Physics & Astronomy, School of Arts and Sciences, University of Pennsylvania
[5]Department of Electrical & Systems Engineering, School of Engineering and Applied Sciences, University of Pennsylvania
[6]Department of Neurology, Perelman School of Medicine, University of Pennsylvania

Abstract:
Non-genetic factors can cause individual cells to fluctuate substantially in gene expression levels over time. Yet it remains unclear whether these fluctuations can persist for much longer than the time for a single cell division. Current methods for measuring gene expression in single cells mostly rely on single time point measurements, making the time of a fluctuation of gene expression or *cellular memory* difficult to measure. Here, we report a method combining Luria and Delbrück's fluctuation analysis with population-based RNA sequencing (MemorySeq) for identifying genes transcriptome-wide whose fluctuations persist for several cell divisions. MemorySeq revealed multiple gene modules that express together in rare cells within otherwise homogeneous clonal populations. Further, we found that these rare cell subpopulations are associated with biologically distinct behaviors in multiple different cancer cell lines, for example, the ability to proliferate in the face of anti-cancer therapeutics. The identification of non-genetic, multigenerational fluctuations has the potential to reveal new forms of biological memory at the level of single cells and suggests that non-genetic heritability of cellular state may be a quantitative property.

Main text:
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 induced by a variety of mechanisms. Genetic differences are one form of memory, encoding variation between organisms on multi-generational timescales. Within an organism, epigenetic mechanisms encode the differences between cell types in different tissues, with cells retaining memory of their state over a large number of cell divisions (*1*). In contrast, recent measurements suggest that 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 has generally been difficult to associate with physiological distinctions between single cells (*2*, *3*).

Less well studied have been intermediate memory timescales; i.e., cellular states that are ultimately transient, and thus are not indefinitely heritable, but may persist for several divisions, and thus are not "noise" either. Such timescales would be long enough to allow for coordinated 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 duration 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. 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.

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 (*4*) (also used in cancer (*5*, *6*)). In our context of cellular memory, the experiment consisted of growing a number of "MemorySeq clones" (~48) of isogenic melanoma cells (WM989-A6) in individual wells, eventually growing them to around 100,000 cells per clone (Fig. 1B). If a fluctuating gene transitioned in and out of the "high" expression state relative 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 is 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 looked for variability in the expression of all genes across these clones (Fig. 1B). To measure the technical variability in our experiment, we 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; this procedure allows

us to assess the technical noise between clones due to both sequencing and sampling variability, and due to errors (Fig. 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 (*6*, *7*), and the independent observation that sister cells often expressed the same genes suggested that these genes displayed some degree of memory (*6*). 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 (*6*) resistance-associated genes and non-resistance associated genes (Fig. 1C). As hoped, we found that the resistance marker 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 far less increased variance as compared to technical noise controls. Intriguingly, *MYC*, a proto-oncogene for which we have seen high levels of cell-to-cell variability (Supp. Fig. 1), showed little increased variance across clones compared to controls, suggesting that its transcriptional memory was much lower than that observed in the marker genes (as verified by alternative means in Fig. 2C). We observed similar behavior for CCNA2, a cell cycle gene whose expression would similarly be expected to be variable but not heritable (*8*); Fig. 2C.

Given that RNA sequencing provides expression levels for all these, 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 (Fig. 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 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.

We next sought to verify by independent means that the genes we identified by MemorySeq indeed expressed in a heritable fashion. 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 around 10 days, at which point we fixed the cells and subjected them to iterative single molecule RNA FISH to measure the expression of 18 genes in individual cells while preserving their

spatial context (Fig. 2A; as per (*6*)). Our reasoning was that as cells divide, their spatial proximity would reflect their relatedness (*9*). In the case of a gene with non-heritable expression, one would expect to find no particular spatial correlations in which cells were deemed high within the population. In contrast, for genes with heritable expression, one would expect to find the high 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, 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 (Fig. 2B).

We wondered to what extent MemorySeq could measure differences in how heritable the high expression state was for different genes, so we further 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 (Fig. 2C; see Supp. Fig. 2 for further analysis).

Motivated by our previous work in this cell line, we also looked for correspondences between the degree of heritability and the rarity of gene expression as measured by the Gini coefficient, a metric for inequality (*6*, *10*, *11*). We observed that that indeed the two metrics were correlated (adjusted R^2 = 0.4898) (Fig. 2D), suggesting that heritable genes identified by MemorySeq are more likely to express only in rare cells. This may be due to the design of the MemorySeq experiment, in which skewness can reach potentially higher levels for rarer expressing genes than for less rarely expressing genes.

The MemorySeq results suggest that many genes do not have indefinite memory (i.e., remain in the high expression state indefinitely). This implies that cells could transition both in and out of the high expression state, which we sought to directly confirm using time-lapse microscopy to trace the expression state of individual cells. We used a split-fluorescent protein approach (see methods) to fuse mNeonGreen to NGFR and then tracked NGFR levels by fluorescence microscopy over a period of 8 days (Fig. 2E; see Supp. Fig. 3 for cell line validation). We found cells that transitioned both from the low state to the high state and *vice versa*.

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, likely involving deviations in the expression of several genes simultaneously, as opposed to a transient (and, as we hypothesized, likely 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 genes that mark rare cells identified by MemorySeq overlapped the set of genes associated with resistance. We found that most MemorySeq heritable genes (162 out of 227) overlapped with resistance markers (as determined by Shaffer et al., Supp. Fig. 4). Moreover, by sorting EGFR-high cells out of the total untreated cell population and performing RNA sequencing, we were able to measure the transcriptional profile of the rare cells alone. We again found a strong overlap between the profiles of the EGFR-high cells and the MemorySeq heritable genes (Fig. 3A, p-value $< 10^{-45}$); furthermore, genes whose expression levels correlated with EGFR/NGFR across MemorySeq clones also tended to correlate with EGFR-high/NGFR-high RNA sequencing results (Supp. Fig. 5). (Note that we did not expect a perfect overlap because it is possible that some genes could have expression levels that correlated with EGFR but are not strongly heritable, and there are strongly heritable genes whose expression may not correlate with EGFR, as we will demonstrate later.) Together, 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 sorted cells by either NGFR or EGFR expression and subjected them to a combination of drugs used to treat melanoma (dabrafenib and trametinib; $BRAF^{V600E}$ and MAPK signaling inhibitors, respectively). On unsorted populations, upon treatment with these drugs, 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 combination, showing that this subpopulation is highly enriched for pre-resistant cells (Fig. 3B; Supp. Fig. 6). This result demonstrated that MemorySeq revealed the same subset of cells that we had previous 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 fluctuations in single cells 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 (Taxol) is a drug used to combat 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 and continue to proliferate, thus leading to drug resistance (Fig. 4A). However, prospective markers to isolate the subpopulation of cells that are resistant to drug have remained elusive (*12*), 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 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, Fig. 4B). The properties of variability between the MemorySeq clones across the transcriptome was comparable to that identified by MemorySeq for the WM989-A6 cell line. Interestingly, however, the overlap between the gene sets was relatively small (Supp. Fig. 7), 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 we had also performed for the melanoma lines (Supp. Fig. 8, compare to Fig. 2C).)

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 Taxol. Amongst the genes identified by MemorySeq was CA9, a surface marker known to be negatively associated with breast cancer chemosensitivity (*13*, *14*), 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 positive cells (2-4% percent, along with CA9 low and mixed subpopulations; validation of sorting in Supp. Fig. 9), after which we added taxol to both and grew the cells for 5 days. We found that, when treated with 1nM taxol, the CA9 positive cells were more likely to be resistant than either the CA9 low or mixed subpopulations. 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 that these rare cells 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 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 (Fig. 5B). To validate that the programs so identified by MemorySeq corresponded to single cell correlations, we compared the correlations between MemorySeq and single cell RNA FISH on a panel of genes across two separate clusters. We found a strong 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 (Fig. 5C). We observed similar clustering in another melanoma cell line (WM983B-E9; see Supp. Fig. 10) and MDA-MB-231-D4, although the specific genes were typically different (Supp. Fig. 7).

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 (*15*–*17*) to identify such groups of genes (Fig. 5D) (see 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 (*6*). We verified this correspondence by comparing the transcriptomes of sorted subpopulations of EGFR-high cells to the clusters of 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 (Fig. 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 one included *DDX58* (RIG-I), *IFIT1*, and *OASL*, which may be related to type 1 interferon signaling (*18*), and was notable because it also appeared to some extent in MDA-MB-231-D4 and WM983B-E9 (Supp. Fig. 11). 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 (Fig. 5E). Another community in WM989-A6 cells was somewhat less coherent, and included genes such as VGF. VGF also shows strong heritability (Fig. 2C) but its expression levels appeared not to correlate with the other pre-resistance genes at the single cell level (Fig. 5E), and was not associated with the drug resistance phenotype (*6*).

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 that propagates to daughters through partitioning upon cell division. Labeling sites of nascent transcription for EGFR (*19*) revealed that transcriptional activity 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 (Supp. Fig. 12).

In sum, we believe MemorySeq is a simple but powerful method for identifying rare, heritable cell 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 (*6*, *20–23*). 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 iPS cells (*24*).

It is also interesting that MemorySeq was quantitative in the sense that it was not just able to report that a gene's expression is heritable, but was also able to provide a relative sense of how heritable that expression is, meaning how many cell divisions does it require 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 reevaluation as a quantitative, rather than qualitative, property of some cellular states.

Materials and Methods:

Cell lines and culture: We used 3 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 (Wistar Institute, a kind gift of Meenhard Herlyn); and MDA-MB-231-D4, a subclone of MDA-MB-231 (ATCC HTB-26). We verified all cell lines by DNA fingerprinting (WM989-A6 and WM983B-E9 performed at the Wistar Institute by DNA STR Microsatellite and MDA-MB-231 by ATCC human STR profiling cell line authentication services). We cultured all melanoma cell lines in the recommended medium, which is TU2% for the melanoma lines (containing 80% MCDB, 10% Leibovitz's L-15, 2% FBS, and 2.4mM $CaCl_2$) and DMEM with 10% FBS for MDA-MB-231.

To create the NGFR-mNG2 WM989-A6-G3 cell line, we used the split mNeonGreen2(mNG2) system described in (*25*). 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 on Dropbox in https://www.dropbox.com/s/3orb3hi1fex32nq/FlucSeq_supplementalSequences.xlsx?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 can be found at https://www.dropbox.com/sh/qwcbljyuurmfmzr/AAAW_MjU1_AWUDwiT6k4L4Zya?dl=0) We further confirmed that mNG2 fluorescence correlates with NGFR mRNA abundance at the single-cell level by RNA FISH and validated that the NGFR-mNG2 WM989-A6-G3 cell line recapitulates phenomenology described in Shaffer et al. 2017, showing increased resistance to vemurafenib in the mNG2-high cells (Supp. Fig. 3; experiments described in RNA FISH, FACS, and drug treatment methods below). To facilitate cell tracking in the timelapse images, we transduced NGFR-mNG2 tagged cell lines with lentivirus encoding an 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 on Dropbox here https://www.dropbox.com/sh/z6y5bl66nix48u6/AABNPWeLnT3DFGP872XfWfqMa?dl=0.

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. Of the 100 starting clones, we used

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 (Qiagen 217004) to isolate RNA from each clone, followed by library preparation using the NEBNext Poly(A) Magnetic Isolation Module (NEB 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). We aligned the reads using STAR and enumerated read counts per gene using HTseq (code available at https://bitbucket.org/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 MDA-MB-231, and 230 heritable genes in WM983B. We generated correlation matrices from the pairwise Pearson correlations for heritable genes across all clones. Full computational analysis of RNA-sequencing data is available on the dropbox here https://www.dropbox.com/sh/rihjom7dt5cct08/AAAfwADapO9NINmCAG3sw7Jua?dl=0. Lists of heritable genes are in Supplementary Table 1.

Fluorescence Activated Cell Sorting (FACS): We stained WM989 A6-G3 melanoma cells for fluorescence assisted 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 in 1x PBS, and incubated for 1 hour at 4℃ with 1:200 mouse anti-EGFR antibody, clone 225 (Millipore, MABF120) in 0.1% BSA PBS. Next, we washed with 0.1% BSA PBS and then incubated for 30 minutes at 4℃ with 1:500 donkey anti-mouse IgG labeled with Alexa Fluor 488 (Jackson Laboratories, 715-545-150). We washed the samples again with 0.1% BSA PBS and then incubated for 10 minutes at 4ºC with 1:250 anti-NGFR antibody conjugated directly to PE-Cy7 (Biolegend, 400126, clone ME20.4) 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 (Biolegend, 400126). After gating for live cells, we collected either the top 0.02-0.2% $EGFR^+$ cells or the top 0.5% $NGFR^+$ 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.

For testing the response of CA9-high cells to paclitaxol, we trypsinzized MDA-MB-231 cells, washed with 0.1% BSA-PBS, then stained cells with anti-carbonic anhydrase 9 antibody conjugated to PE (Miltenyi Biotech clone REA658, 130-110-057) at a dilution of 1:11 for 30 minutes at 4C. We gated for live cells, and then we sorted the top 2-3% percent of CA9 stained cells (CA9-high), the bottom 2-3% of CA9 cells (CA9-low), and a mixed population using only the live cell gates (CA9-mix). We performed single-molecule RNA FISH for CA9 confirming that sorting with CA9 antibody enriched for CA9-high expressing cells (Supp. Fig. 9).

To validate that NGFR-high cells tagged by the mNG2 fluorescent protein are more resistant to vemurafenib, we trypsinized and pelleted the NGFR-mNG2 WM989-A6-G3 cell line, washed cells once with 1x PBS containing 2mM EDTA, then resuspended in 1xPBS 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 the dropbox here https://www.dropbox.com/sh/08unb7zuazi729g/AABfkRnsetuOdoCoj3dKWD6Oa?dl=0.

Drug treatment experiments: We made stock solutions of 4mM trametinib (GSK1120212, Seleckchem, S2673), 4mM vemurafenib (PLX4032, Selleckchem, S1267), and 4mM paclitaxol (LifeTechnologies, P3456). 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 paclitaxol. For trametinib treatment experiments, we cultured NGFR FACS sorted WM989-A6 (sorting is described in the FACS section of methods) for 2-3 weeks. For vemurafenib experiments, we cultured the FACS sorted NGFR-mNG2 WM989-A6-G3 in vemurafenib for 21 days. For paclitaxol experiments, we cultured CA9 FACS sorted MDA-MB-231 cells in paclitaxol 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.

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.

Time-lapse imaging: Our timelapse images were taken on 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 9 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.

For analysis of timelapse images we used custom MATLAB scripts available at https://www.dropbox.com/sh/9caovkv858r8gcn/AADHy_KkovcrqCpDD_kgYR1Ja?dl=0. For each timepoint, this pipeline stitches together images from each scan, adjusts image contrast then overlays mNG2 and iRFP670 images for the final video (Supplementary video 1). From this video, we identified three cells that first increased and then decreased *NGFR* expression during the course of the timelapse. We manually tracked positions of these cells using the Manual Tracking FIJI plugin, and we then used these tracks along with custom MATLAB scripts to create cropped and annotated videos (Supplementary videos 2-7).

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 (probe sequences available in Supplementary Table 2). We pooled 15-32 oligonucleotides targeting each gene and coupled each set of probes to either Cy3 (GE Healthcare), Alexa 594 (Life Technologies), Atto647N or Atto 700 (Atto-Tec). We performed single molecule RNA FISH as described in Raj et al. 2008 for multiple cycles of hybridization. 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 37ºC and then performed two cycles of 30 minute washes at 37ºC 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.

RNA FISH image analysis: For analysis of gridded image scans, we used custom MATLAB software designed for the analysis in (*6*). 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 extract 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 the bitbucket here https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Home. 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 (*10*)) for each gene using the "ineq" package. The code for this analysis is on the dropbox here https://www.dropbox.com/s/kn6otdws01gcgi5/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 for MDA-MB-231-D4 if they 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 recomputed 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/xsa46euxi2n1jub/AABgY2wVyx3qWt59f2v-2-wQa?dl=0).

Network community identification: For each pair of significantly heritable genes (from gene lists described in MemorySeq RNA sequencing analysis section of 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 edges between nodes assigned the value of the correlation coefficient. We sought tightly connected groups of nodes within this network known as *network communities*. We performed *k*-clique community detection (*15*–*17*) with k=4 on binarized networks created by thresholding the original weighted network at decreasing values (*26*, *27*). 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 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 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.

Acknowledgements:
Thank Siyu Feng, David Brown, and Bo Huang for constructs pSFFV_mNG2(1-10). We thank Caroline Bartman, Yogesh Goyal, and other members of the Raj lab 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. SMS acknowledges support from NIH F30 AI114475. BLE acknowledges support from NIH training grants T32GM007170 and T32HG000046. AES and DSB acknowledge support from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, and the ISI Foundation. AR acknowledges NIH/NCI PSOC award number U54 CA193417, NSF CAREER 1350601, P30 CA016520, SPORE P50 CA174523, NIH U01 CA227550, NIH 4DN U01 HL129998, NIH Center for Photogenomics (RM1 HG007743), and the Tara Miller Foundation.

Figure legends:

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 seeded 48 wells from those cells 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. We also performed control experiments for technical noise by taking a population of cells, plating around ~100,000 cells directly into individual wells and performed RNA sequencing to determine how much variability in expression would arise for purely technical reasons. **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 panel C noted as indicated on both plots.

Figure 2: Single cell RNA FISH and time-lapse microscopy verify quantitative nature of MemorySeq for measuring heritability. A. Schematic of space-FISH experiment. We plated WM989-A6 melanoma cells sparsely on a dish and allowed them to grow for 10 days, which is approximately 10 cell divisions. We then fixed the cells and performed iterative RNA FISH to measure the expression of 20 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. **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.** 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 expresses, the remaining portion of mNeonGreen2 binds to the NGFR fusion protein and becomes fluorescent. We then performed time-lapse microscopy imaging the NGFR protein (nucleus labeled with H2b-iRFP670) at 6 hour intervals for 8 days. Here, we show a series of fluorescent micrographs of a single cell tracked over time having gone through 2 divisions over 72 hours, with the red arrow indicating the tracked cell. Scale bars are all 5μm long.

Figure 3: High memory genes overlap with pre-resistance markers. A. We compared the set of genes identified as high memory by MemorySeq to the set of genes associated with pre-resistance (*6*). We identified genes associated with pre-resistance by performing RNA sequencing followed by differential expression analysis on cells sorted by EGFR levels. Genes with increased expression in EGFR-high cells have a minimum of 5 reads per million in at least one sample and log2 fold change (EGFR-high/EGFR-Mix) > 1.5. The p-value displayed reflects the statistical significance of the overlap between the sets of genes in the universe indicated. **B.** Rare cells within clonal WM989-A6 populations marked by high levels of NGFR protein were sorted, cultured for 8-16 hours and then subjected to trametinib treatment at 10nM (MEK inhibitor) for 3 weeks. Image shows number of resistant colonies (circled) along with number of cells within the resistant colony as indicated. Biological replicate in Supp. Fig. 6.

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% (same approach as applied to WM989 data). **C.** 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 computationally identifying cells. All scale bars are 50μm long.

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 Fig. 1. **C.** Comparison of coherence between MemorySeq bulk RNA-seq 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-seq levels is on the left, RNA FISH on the right. Callout shows raw single cell RNA FISH counts for 12,192 cells between *MMP1* and *SERPINB2*. **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 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 (left) or NGFR (right) high cells (using fluorescent antibody labeling followed by RNA-seq on the High versus Mix populations).

References:

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

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

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

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

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

6. S. M. Shaffer *et al.*, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*. **546**, 431–435 (2017).

7. M. Fallahi-Sichani *et al.*, Adaptive resistance of melanoma cells to RAF inhibition via reversible induction of a slowly dividing de-differentiated state. *Mol. Syst. Biol.* **13**, 905 (2017).

8. H. X. Chao *et al.*, Evidence that the cell cycle is a series of uncoupled, memoryless phases. *bioRxiv* (2018), p. 283614.

9. S. Hormoz *et al.*, Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst*. **3**, 419–433.e8 (2016).

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

11. E. Torre *et al.*, Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. *Cell Syst* (2018), doi:10.1016/j.cels.2018.01.014.

12. R. Gao *et al.*, Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8**, 228 (2017).

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

14. N. Aomatsu *et al.*, Carbonic anhydrase 9 is associated with chemosensitivity and prognosis in breast cancer patients treated with taxane and anthracycline. *BMC Cancer*. **14**, 400 (2014).

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

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

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

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

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

20. S. V. Sharma *et al.*, A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*. **141**, 69–80 (2010).

21. S. Spencer, S. Gaudet, J. Albeck, J. Burke, P. Sorger, Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* (2009), doi:10.1038/nature08012.

22. A. O. Pisco, S. Huang, 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 (2015).

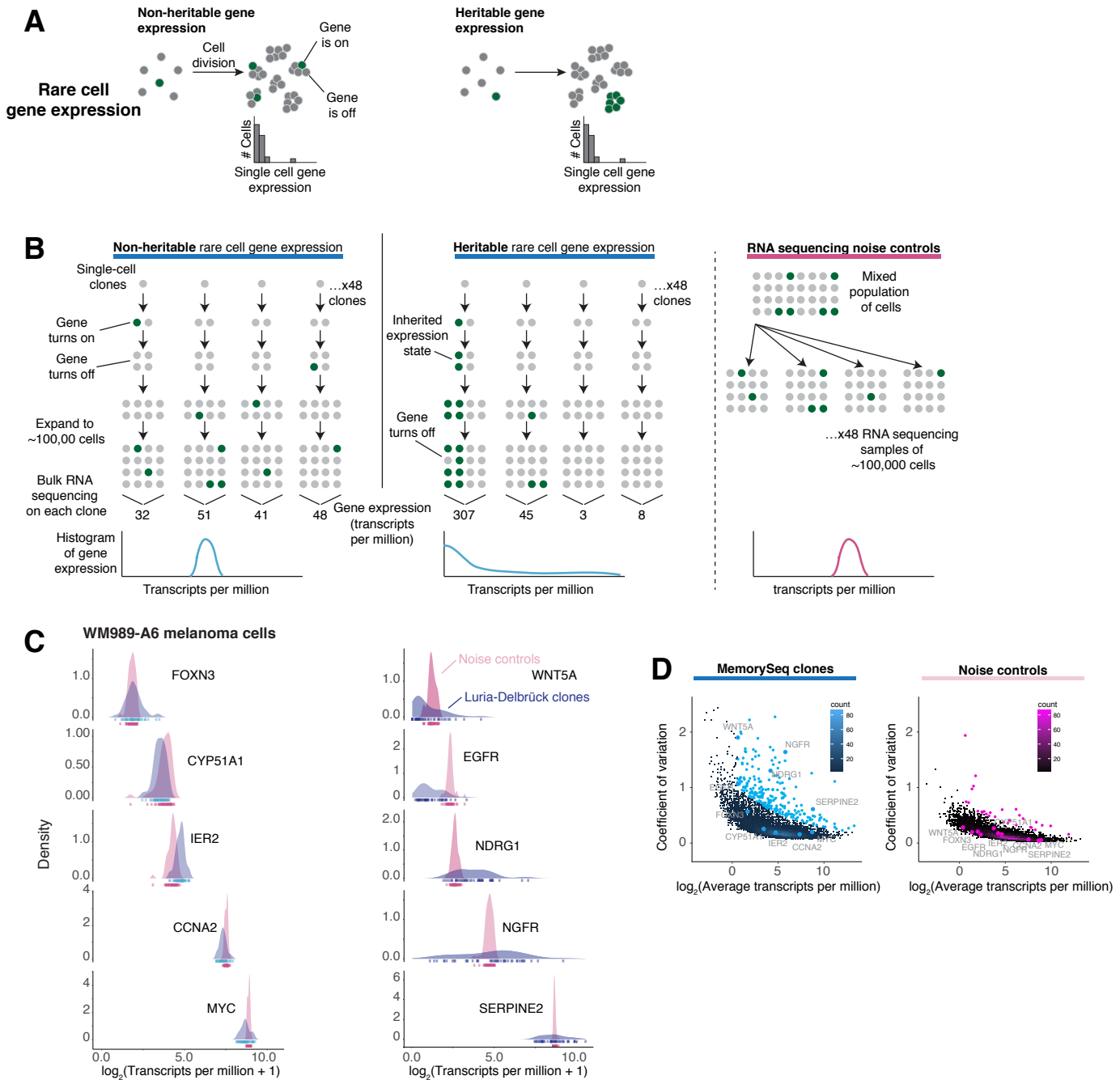
23. A. Roesch *et al.*, A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth. *Cell*. **141**, 583–594 (2010).

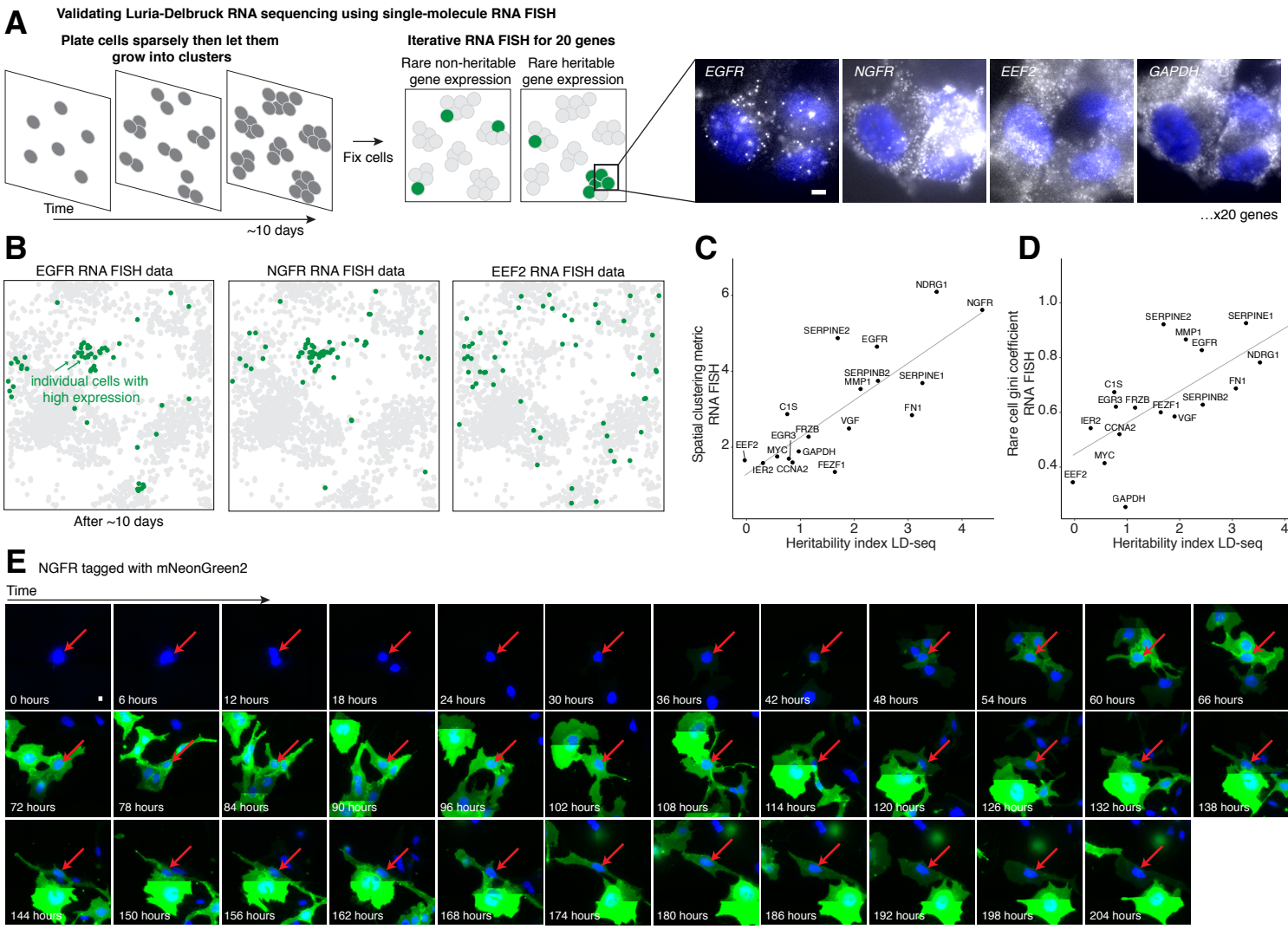
24. J. Hanna *et al.*, Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*. **462**, 595–601 (2009).

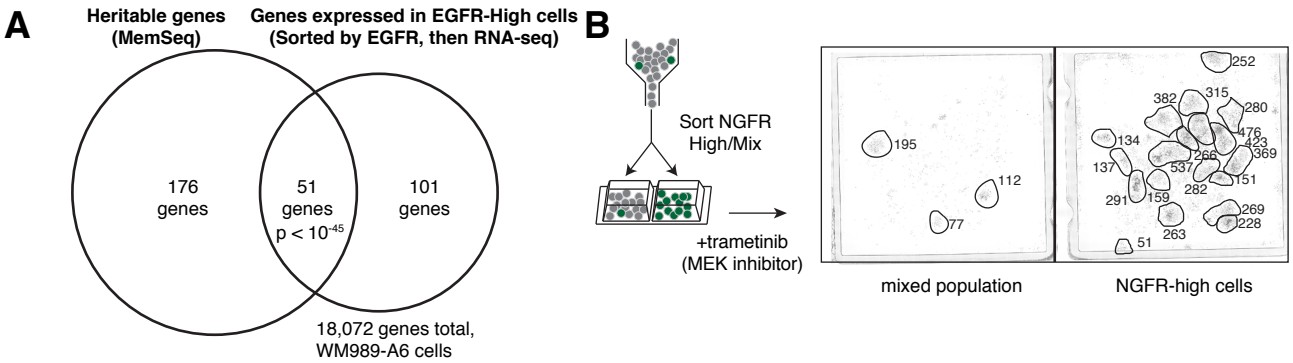
25. S. Feng *et al.*, Improved split fluorescent proteins for endogenous protein labeling. *Nat. Commun.* **8**, 370 (2017).

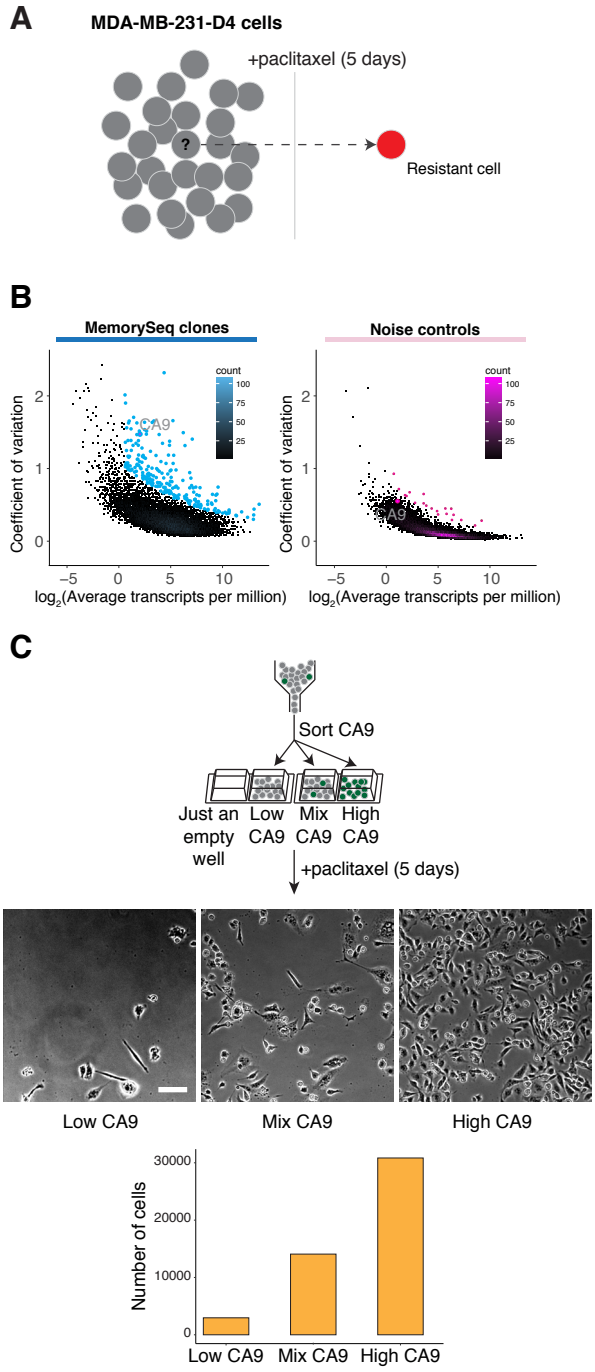
26. C. Giusti, E. Pastalkova, C. Curto, V. Itskov, Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13455–13460 (2015).

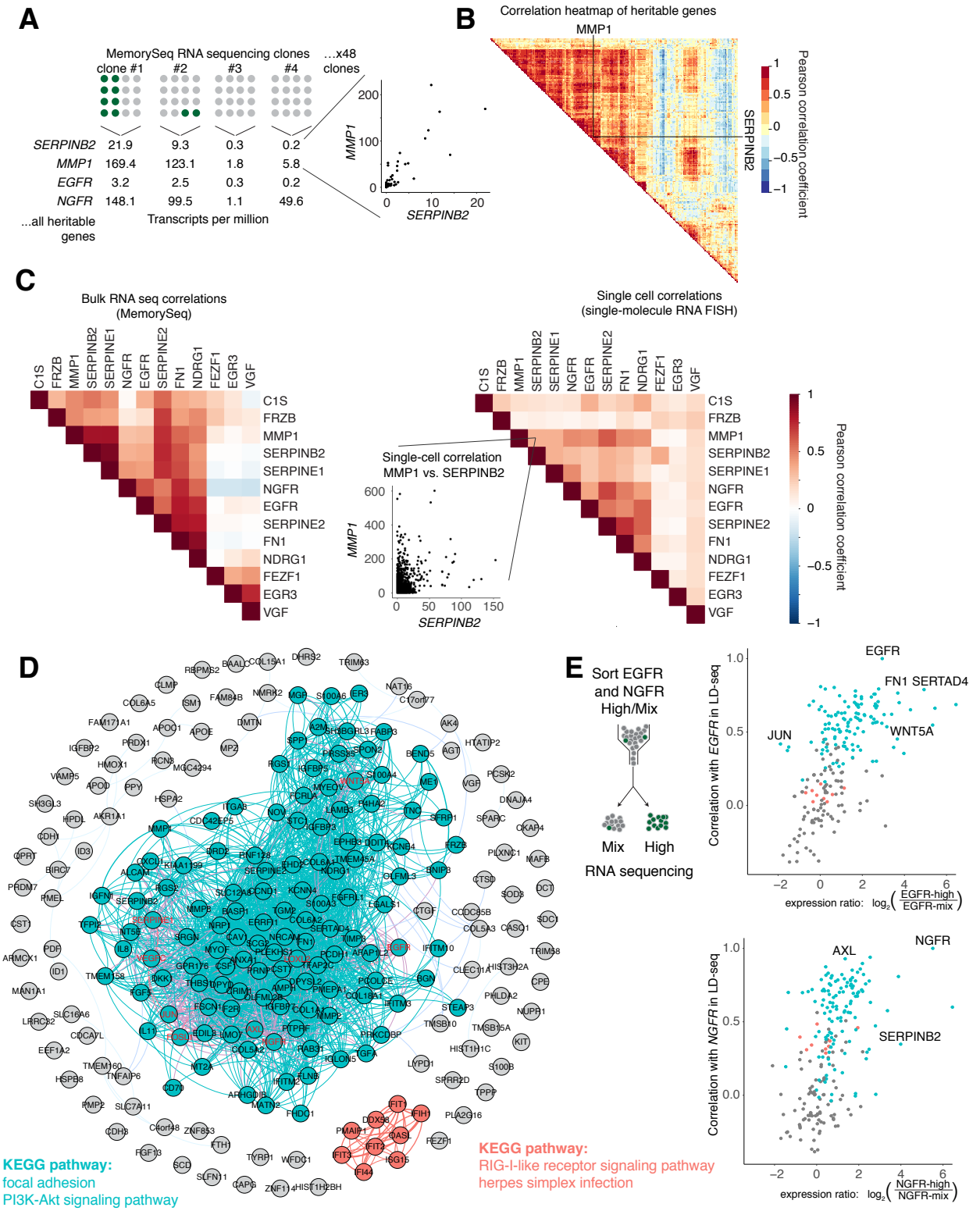
27. B. Rieck, U. Fugacci, J. Lukasczyk, H. Leitte, Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks. *IEEE Trans. Vis. Comput. Graph.* **24**, 822–831 (2018).


A
Non-heritable gene expression
Rare cell gene expression
Cell division
Gene is on
Gene is off
Cells
Single cell gene expression
Heritable gene expression
Cells
Single cell gene expression
B
Non-heritable rare cell gene expression
Single-cell clones
…x48 clones
Gene turns on
Gene turns off
Expand to ~100,00 cells
Bulk RNA sequencing on each clone
32
51
41
48
Histogram of gene expression
Transcripts per million
Heritable rare cell gene expression
Inherited expression state
Gene turns off
307
45
3
8
Transcripts per million
Gene expression (transcripts per million)
RNA sequencing noise controls
Mixed population of cells
…x48 RNA sequencing samples of ~100,000 cells
transcripts per million
C
WM989-A6 melanoma cells
FOXN3
CYP51A1
IER2
CCNA2
MYC
Density
log2(Transcripts per million + 1)
Noise controls
WNT5A
Luria-Delbrück clones
EGFR
NDRG1
NGFR
SERPINE2
log2(Transcripts per million + 1)
D
MemorySeq clones
Noise controls
Coefficient of variation
count
log2(Average transcripts per million)


A
Validating Luria-Delbruck RNA sequencing using single-molecule RNA FISH
Plate cells sparsely then let them grow into clusters
Time
~10 days
Fix cells
Iterative RNA FISH for 20 genes
Rare non-heritable gene expression
Rare heritable gene expression
EGFR
NGFR
EEF2
GAPDH
…x20 genes
B
EGFR RNA FISH data
NGFR RNA FISH data
EEF2 RNA FISH data
individual cells with high expression
After ~10 days
C
Spatial clustering metric RNA FISH
Heritability index LD-seq
NDRG1
NGFR
SERPINE2
EGFR
SERPINB2
MMP1
SERPINE1
C1S
FN1
VGF
FRZB
EGR3
EEF2
MYC
GAPDH
IER2
CCNA2
FEZF1
D
Rare cell gini coefficient RNA FISH
Heritability index LD-seq
SERPINE2
SERPINE1
MMP1
EGFR
NGFR
NDRG1
C1S
FN1
EGR3
FRZB
FEZF1
SERPINB2
VGF
IER2
CCNA2
MYC
EEF2
GAPDH
E
NGFR tagged with mNeonGreen2
Time
0 hours
6 hours
12 hours
18 hours
24 hours
30 hours
36 hours
42 hours
48 hours
54 hours
60 hours
66 hours
72 hours
78 hours
84 hours
90 hours
96 hours
102 hours
108 hours
114 hours
120 hours
126 hours
132 hours
138 hours
144 hours
150 hours
156 hours
162 hours
168 hours
174 hours
180 hours
186 hours
192 hours
198 hours
204 hours


A
Heritable genes
(MemSeq)
Genes expressed in EGFR-High cells
(Sorted by EGFR, then RNA-seq)
176
genes
51
genes
p < 10^-45
101
genes
18,072 genes total,
WM989-A6 cells
B
Sort NGFR
High/Mix
+trametinib
(MEK inhibitor)
195
112
77
252
315
382
280
476
423
369
134
266
537
137
151
282
291
159
269
228
263
51
mixed population
NGFR-high cells


A
MDA-MB-231-D4 cells
+paclitaxel (5 days)
?
Resistant cell
B
MemorySeq clones
Noise controls
count
100
75
50
25
CA9
Coefficient of variation
log2(Average transcripts per million)
C
Sort CA9
Just an empty well
Low CA9
Mix CA9
High CA9
+paclitaxel (5 days)
Low CA9
Mix CA9
High CA9
Number of cells
30000
20000
10000
0


A
MemorySeq RNA sequencing clones
clone #1
#2
#3
#4
...x48 clones
SERPINB2 21.9 9.3 0.3 0.2
MMP1 169.4 123.1 1.8 5.8
EGFR 3.2 2.5 0.3 0.2
NGFR 148.1 99.5 1.1 49.6
...all heritable genes
Transcripts per million
MMP1
SERPINB2
B
Correlation heatmap of heritable genes
MMP1
SERPINB2
Pearson correlation coefficient
C
Bulk RNA seq correlations (MemorySeq)
Single cell correlations (single-molecule RNA FISH)
Single-cell correlation MMP1 vs. SERPINB2
C1S
FRZB
MMP1
SERPINB2
SERPINE1
NGFR
EGFR
SERPINE2
FN1
NDRG1
FEZF1
EGR3
VGF
Pearson correlation coefficient
D
KEGG pathway:
focal adhesion
PI3K-Akt signaling pathway
ECM-receptor interaction
pathways in cancer
GO Biological Process:
anatomical structure morphogenesis
response to wounding
cell morphogenesis involved in differentiation
regulation of locomotion
regulation of multicellular organismal process
tissue development
E
Sort EGFR and NGFR High/Mix
Mix
High
RNA sequencing
EGFR
FN1 SERTAD4
JUN
WNT5A
Correlation with EGFR in LD-seq
expression ratio: log2 (EGFR-high / EGFR-mix)
NGFR
AXL
SERPINB2
Correlation with NGFR in LD-seq
expression ratio: log2 (NGFR-high / NGFR-mix)
KEGG pathway:
RIG-I-like receptor signaling pathway
herpes simplex infection
GO Biological Process:
response to virus
defense response to virus
type I interferon signaling pathway
cellular response to type I interferon
innate immune response
detection of virus