Title
Inferring cell state from cell behavior

Author
Kimmel, Jacob Cameron

Publication Date
2018

Inferring cell state from cell behavior

by

Jacob Cameron Kimmel

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Developmental and Stem Cell Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

This page is intentionally left blank.

Acknowledgments

This work would not have been possible without the mentorship, kindness, and support of the extraordinary University of California San Francisco community. Learning in this extraordinary place among the most inspiring people has been the greatest honor of my life.

I would like to thank my thesis advisors Wallace Marshall and Andrew Brack for their ever present guidance and support throughout this thesis.

In our first meeting, I told Wallace that I would like to examine cells as if they were living artifacts from another world. We laughed together at the notion, and he provided me with the opportunity to pursue the questions that fascinated me, regardless of how far-fetched my motivations and approaches surely seemed. When neuroses caused me to doubt my own abilities, Wallace provided constant encouragement. When I felt demoralized by the challenges we inherently face in understanding the world around us, Wallace reminded me of the wonder in simple observation.

During my first days at UCSF, Andrew and I had a discussion about the biology of aging that still motivates my research questions to this day. His guidance has honed my ability to ask clear, answerable scientific questions in a manner that seemed impossible when I first began graduate school. As I random walked about the parameter space of life, attempting to find my direction, Andrew provided understanding and mentorship on the broader questions of who I would like to be, and how I should get there.

Thank you to Todd Nystul for serving on my thesis committee and for mentorship above and beyond all obligation. From you, I have learned not only to be a better scientist, but a kinder person.

Thank you to Orion Weiner for chairing my thesis committee and challenging me to be the best scientist I could be.

Thank you to Steve Altschuler, Simone Bianco, Steve Finkbeiner, Diana Laird, Alex Pollen, Julie Sneddon, Saul Villeda, Leor Weinberger, Lani Wu, Ann Zovain, and many others for the assistance in my training.

To each and every member of the Marshall and Brack labs: working alongside you has been the best learning experience I could ask for. I am proud to call each of you a colleague and friend.

To the members of the DSCB program: thank you for lighting a chreod to be followed in the early ontogeny of a scientist. You have all served as the best peers, mentors, and friends anyone could ask for.

To my sister Alexis and parents Kerry and Jeff, thank you for your support and understanding throughout this journey.

Work in this thesis has previously been published in *PLoS Computational Biology* and *bioRxiv* in the following volumes:

Kimmel JC, Chang AY, Brack AS, Marshall WF (2018) Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. PLOS Computational Biology 14(1): e1005927. https://doi.org/10.1371/journal.pcbi.1005927

Kimmel JC, Brack AS, Marshall WF (2018). Deep convolutional and recurrent neural networks for cell motility discrimination and prediction. bioRxiv 159202; doi: https://doi.org/10.1101/159202

Inferring cell state from cell behavior

Jacob C. Kimmel

Abstract

Cell populations display heterogeneous and dynamic phenotypic states at multiple scales. Similar to molecular features commonly used to explore cell heterogeneity, cell behavior is a rich phenotypic space that may allow for identification of relevant cell states. Inference of cell state from cell behavior across a time course may enable the investigation of dynamics of transitions between heterogeneous cell states, a task difficult to perform with destructive molecular observations. Cell motility is one such easily observed cell behavior with known biomedical relevance. To investigate heterogeneous cell states and their dynamics through the lens of cell behavior, we developed Heteromotility and Lanternfish, software tools to extract quantitative motility features from timelapse cell images. In mouse embryonic fibroblasts (MEFs), myoblasts, and muscle stem cells (MuSCs), Heteromotility analysis identifies multiple motility phenotypes within the population. In all three systems, the motility state identity of individual cells is dynamic. We demonstrate that using deep neural networks and heuristic machine learning methods, cell types and states can be discriminated in multiple systems. Quantification of state transitions reveals that MuSCs undergoing activation transition through progressive motility states toward the myoblast phenotype. Transition rates during MuSC activation suggest non-linear kinetics. By probability flux analysis, we find that this MuSC motility state system breaks detailed balance, while the MEF and myoblast systems do not. Balanced behavior state transitions can be captured by equilibrium formalisms, while unbalanced switching between states violates equilibrium conditions and would require an external driving force. Our data indicate that the system regulating cell behavior can be decomposed into a set of attractor states which depend on the identity of the cell, together with a set of transitions between states. These results support a conceptual view of cell populations as dynamical systems, responding to inputs from signaling pathways and generating outputs in the form of state transitions and observable motile behaviors.

Contents

List of Figures

Chapter 1

Inferring cell state from cell motility behavior reveals a dynamic state system and broken detailed balance

1.1 Introduction

Within any group of cells, each individual cell is not necessarily like its neighbors. These differences often have functional significance, with cells at different points along the phenotypic spectrum exhibiting distinct behavior [1]. This has been noted across a broad swath of cell biology, including in the cases of stem cell biology [2, 3] and cell geometry definition [4]. Regulatory decisions at one scale may reflect phenotypes at other scales, allowing identification of a broad "cellular state" based on a more limited set of observations. Effective classification of cancerous cell functionality based on morphology demonstrates this concept [5–7]. Since cell state determines cell function, state transitions may manifest as changes in cell behavior. Understanding the regulation of cell behavior will require understanding the nature of the cell state-space and the transitions that take place within it. Is the state space continuous or discrete? Are all state transitions equally likely, as in an equilibrium system, or do they tend to take place in a specific sequence, as in computing device? Answering such questions requires a framework for defining cell state in terms of observable

behaviors. Regardless of the method used to probe cell state, it must be able to measure state in living cells at multiple time points, in order to allow state transitions to be characterized.

Recent advances in single cell assays have allowed for detailed, quantitative descriptions of individual cells at the molecular level. Single cell sequencing technologies in particular have uncovered heterogeneity at both the transcriptional and epigenetic level [8, 9]. In multiple stem cell compartments, molecular analysis of heterogeneity has revealed that not all stem cells are functionally equivalent. Within the hematopoietic [10], muscle [11], epithelial, and other [12] stem cell pools, subpopulations of cells with different functionality coexist. Stem cell heterogeneity has been demonstrated in the form of lineage-bias or differences in regenerative capacity. A similar phenomena is present in malignant tumors. Within a given tumor, some subpopulations may have higher tumorogenic potential, or increased resistance to a particular therapy [13, 14]. Understanding heterogeneity in these and other contexts is essential to building an accurate picture of cell-based systems.

In addition to defining cell states to account for population heterogeneity, we also seek to understand the transitions between states, because those transitions reveal the logic of the cellular control system. In seminal work, Waddington introduced the conceptual model of an 'epigenetic landscape' governing cell phenotype decisions, akin to a potential energy landscape in a physical system [15, 16]. In this model, cells progress through phenotypic states by migrating continuously down the gradient of the landscape, eventually resting in a stable basin of attraction. This model implies that cell state transitions are governed by a 'potential energy' in each state, which can be estimated by the state's stability. A cell system governed by this model would display detailed balance in the absence of an external force or signal, and break detailed balance in the presence of such an external input [16, 17]. By directly measuring the dynamics of cell state transitions, we can produce an estimate of the landscape for cell behavioral phenotypes and determine if state transitions occur stochastically or are influenced by external inputs.

While existing molecular assays such as single-cell RNA-sequencing can provide detailed information about a cell population's heterogeneity, these assays are generally destructive and restricted to a single time point for analysis. Methods have been developed to infer cell state lineages from observations at a single time point [18, 19], but these methods are not able to quantify dynamics and

assume that proximity in the measured variable space describes a transition relationship between states. Real-time, non-destructive assays that reveal subpopulation composition over time and observe state transitions in living cells would serve as complementary approaches to investigate cellular states and state transitions.

Cells in culture have a diverse behavioral repertoire. Each cell may exhibit motion, dynamic morphology changes [4, 20], and symmetric or asymmetric divisions, all of which can be observed by simple microscopy. These cell behaviors represent a rich phenotypic space from which many quantitative features may be extracted. Behavior also has an inherent functional relevance at the single cell level, and often a functional relevance on the system level. For example, extracellular matrix remodeling and deposition by fibroblasts is dictated by their motion [21, 22], and stem cell and progenitor migration is critical for organismal development [23]. Within each cell, behavior represents a layer of abstraction above molecular phenotypes. Hence, observably different behavioral states may serve as a proxy for distinct ensembles of underlying molecular states.

Recent work from several groups has elucidated the regulatory principles governing cell shape definition [24, 25] and the relationships between morphological variables [26] utilizing quantitative, timelapse imaging. Profiling of static morphological information alone is sufficient to discriminate various states of cellular perturbation [27–30]. Considering temporal information, quantification of the dynamic transitions between morphological phenotypes using multiple Hidden Markov Model frameworks has allowed for increased precision in classifying both chemical and genetic cellular perturbations [20, 31]. Progenitor fate outcomes have also been predicted based on cell shape and behavior quantifications [32]. These remarkable results, achieved through quantitative exploration of cell morphology behaviors and the dynamic transitions between them, support the notion that quantitative analysis of cell behavior can allow for detection of cellular states and observation of state transitions. Modeling cells as dynamical systems in this manner has proven to not only elucidate fundamental biological principles, but to provide utility in applied contexts. Here, we extend this approach to the analysis of cell motility behaviors and consider alternative tools for extracting biological insight from observed state transitions.

Cell motility behaviors are inherently related to the morphological state of a cell [33–35]. Different motility behaviors are associated with different morphological states, and some morphological events

such as protrusion extension are directly causal for motility behaviors. However, motility events involving a displacement event where the cell moves along a substrate to a new location are not necessarily captured entirely through morphological descriptors. Supporting this notion, it has recently been shown that displacement motility behaviors provide a strong signal for the inference of hematopoietic stem cell differentiation decisions when used in addition to morphological descriptors [36]. Therefore, we believe it is valuable to develop means of measuring motility behaviors and considering transitions between different motility phenotypes to capture the full range of information provided by cell behavior.

Cell motility is a particularly dramatic cell behavior, but it is difficult to examine quantitatively. Traditional cell motility assays rely on a binary filtration of cells based on a functional test, such as crossing a membrane barrier. Timelapse microscopy has also been appreciated for decades as a means of tracking and quantifying cell motility at the single cell level [37–40]. These classic studies demonstrate that cell motility is predictive of cellular function and that quantitative motility analysis can elucidate underlying cell state control mechanisms [39, 40]. Recent approaches to timelapse motility analysis have expanded upon these techniques to extract multidimensional quantitative information from individual cells [32, 41–45]. However, existing methods focus largely on speed and distance metrics of cell motility on a single arbitrarily-chosen timescale, limiting the degree of heterogeneity that can be revealed within a population.

Here, we present Heteromotility, a software tool for quantitative analysis of cell motility in timelapse images with a diverse feature set. In addition to commonly calculated features such as distance traveled, turning, and speed metrics, Heteromotility provides features that allow for comparisons to models of complex motion, such as Levy flights and fractal Brownian motion, and estimate long-term dependence within a cell's displacement distribution. This feature set creates a high-dimensional space representing the possible phenotypes of cell motility. These features may be useful for multiple downstream applications, including supervised classification of different motility phenotypes, and unsupervised definition of different motility behaviors, both of which we demonstrate here. Our tool is modular, and we also provide tools to map the high-dimensional motility feature space into a low dimensional state space, tools to quantify changes in cell motility phenotypes over time as transitions in state space, and tools to consider the stochastic or directed

nature of these transitions. These tools allow the dynamics of cell behavior states to be elucidated, in addition to simply identifying heterogeneous phenotypes.

We demonstrate that Heteromotility analysis is sufficient to discriminate between simulations of several models of motion. Applied to wild-type and transformed mouse embryonic fibroblasts (MEFs), Heteromotility analysis reveals a shared set of motility states between the two systems, in which transformed cells preferentially occupy a more motile state. In mouse myoblasts and MuSCs, phenotypically distinguished motility states are revealed. MuSCs display a series of progressively more motile states, suggesting states of activation. Quantifying transition dynamics within this state system over time reveals that MuSCs transition through state space toward a progressively more activated, myoblast-like motility phenotype. Viewed through the lens of myogenic activation, we are able to follow the activation dynamics of individual MuSCs for the first time. Applying probability flux analysis, we find that the MuSC motility state system breaks detailed balance, quantitatively confirming that these state transitions occur in an ordered and predictable sequence, as in a computing device.

1.2 Results

1.2.1 Heteromotility analysis approach

Heteromotility analyzes motility features in a set of provided motion paths, as obtained through timelapse imaging, image segmentation, and tracking (see Methods). From these paths, 79 motility features are calculated to comprise a "motility fingerprint" (Fig. 1.7). These features include simple metrics of speed, total and net distance traveled, the proportion of time a cell spends moving, and the speed characteristics during that period. The linearity of motion is assessed by linear regression through all points occupied by the cell in the time series, taking Pearson's r^2 as a metric of fit. Monotonicity is also considered for the distribution of points, using Spearman's ρ^2. In some instances, cells have been proposed to have a directional bias when making turns [46]. Turn direction and magnitude features are provided that consider turns on various time intervals.

Another class of features is concerned with the directionality and persistence of motion. There are several possible ways to characterize the persistence of motion, hence we have included several

metrics that allow for reasonable consideration. A cell's "progressivity," is considered as the ratio of net distance to total distance traveled. This serves as a metric of directional persistence [47]. Mean squared displacement (MSD) is considered for each cell for a variable time lag τ, and the power law exponent α is taken as a quantification of the relationship $MSD \propto \tau^{\alpha}$. Directed motion may therefore be expected to have a larger value than undirected random-walk motion or non-motion, indicating a superdiffusive behavior [48].

The distribution of a cell's displacement steps is also informative. Motion that exhibits a displacement distribution with a heavy right tail, referred to as a Levy flight, has been demonstrated to optimize the success of a random search [49]. This property has led to the Levy flight foraging hypothesis, suggesting that biological systems may perform Levy flight-like motion when searching for resources [50–52]. To assess the Levy flight-like nature of a cell's motility behavior, the Heteromotility software provides metrics of displacement kurtosis for displacement distributions on multiple time scales. Larger values of kurtosis indicate heavier distribution tails, such that higher kurtosis may indicate more Levy flight-like motion. The Heteromotility software also considers the non-Gaussian parameter α_2, for which larger values indicate a more heavily tailed, Levy flight-like distribution [53, 54]. The second and third moments of the displacement distribution are also provided as features.

If displacements are considered as a time series, the self-similarity and long range memory may provide insight into the coordination of motility behavior [39]. The Heteromotility software calculates the autocorrelation function for displacements with variable time lags τ as a metric of self-similarity. Fractal Brownian motion (fBm) describes a Gaussian process $B_H(t)$ for a time t on the continuous interval $[0, T]$ with successive displacements that are not necessarily independent. The Hurst parameter H describes the self-similarity of a fBm process, with the interval $0 < H < 0.5$ describing a process with negatively correlated successive displacements (large displacements are more likely to follow small displacements), $H = 0.5$ describing non-correlated, independent successive displacements (Brownian motion), and $0.5 < H < 1$ describing positively correlated successive displacements (large displacements are more likely to follow large displacements)(see Supp. Methods for additional detail). The Heteromotility software estimates the Hurst parameter as a metric of long range memory using Mandelbrot's rescaled range method (see Supp. Methods) [55] [56]. As

Brownian motion displays a Hurst parameter $H = 0.5$, deviations from this value may indicate long range memory of a cell's displacement series and coordinated motility behavior. Each of the motility features described here concerns single cells, and none directly measure correlations in motion between cells as may be observed during collective migration. However, cells with similar behaviors, such as groups of "leader," and "follower," cells, may display correlations in single cell features.

1.2.2 Heteromotility features are sufficient to distinguish canonical models of motion

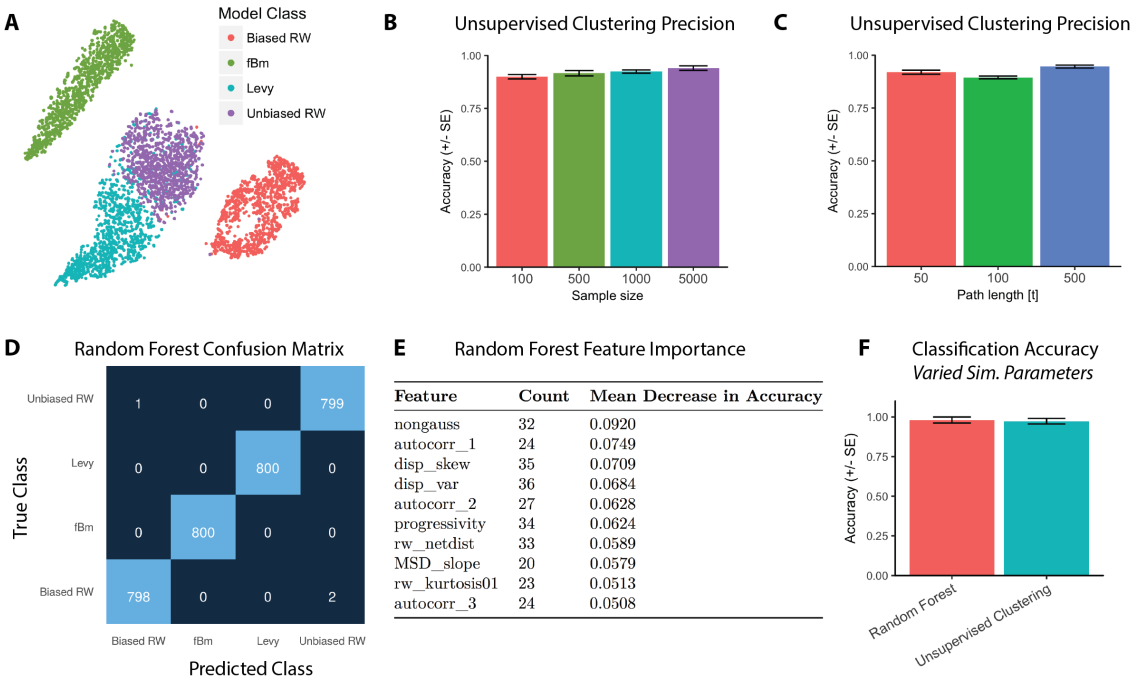
In principle one could devise an infinite number of motion descriptors. How can we determine whether a given set of motion features is sufficient to capture meaningful aspects of motion? To test whether the feature set outlined above is sufficient to distinguish biologically relevant models of motion, we simulated and analyzed cell paths generated by four distinct motion models: unbiased random walks, biased random walks, Levy flights, and fractal Brownian motion with long range memory ($H = 0.9$)(see Methods for implementations). These data were initially simulated with the same mean displacement size, such that the motion models are not trivially separable based on a parameter unrelated to their defining characteristics. This problem therefore represents the most challenging setting for separating models of these types. Large simulated data sets (n = 20,000) were generated with a range of track lengths. Samples with a range of sizes (members per class) were redrawn from these large populations three times per sample size/track length parameterization. We then analyzed the simulated data to determine if the feature set is sufficient to place each model in a distinct region of a joint feature space.

Visualization of high-dimensional data sets, such as the Heteromotility feature set, presents a fundamental challenge. Here we employ t-Stochastic Neighbor Embedding (t-SNE), a visualization method that embeds high-dimensional data into a low-dimensional map, to generate 2D projections of our high dimensional feature space [57] (see Supplemental Methods for discussion of t-SNE parameter selection). When Heteromotility feature space is visualized using t-SNE (perplexity = 50), these models of motion occupy distinct regions of feature space for a number of sample sizes

and track lengths. A sample visualization of a simulation with 1000 members per class and track lengths of 100 units is shown (Fig. 1.1A)(see further examples, Fig. 1.8). Unsupervised hierarchical clustering was performed using Ward's method [58] (see Supplemental methods for discussion). Unsupervised clustering segregates the various models with high accuracy based on Heteromotility features for a range sample sizes as small as 100 members per class, and lengths as short as 50 time steps. Accuracy tends to increase with larger sample sizes and longer track lengths (Fig. 1.1B, C). Clustering was also performed on simulations with varied fundamental parameters, and accuracy is likewise high (Fig. 1.1F). Parameters for the bias magnitude of biased random walkers, random walker displacement means, the Levy flier exponent, and fractal Brownian motion Hurst parameter were all varied. 10 groupings of the four simulations with varying parameter settings were sampled and classified as above (see Table S2 for parameter settings).

Applying a supervised Random Forest classifier [59] to each of the simulated populations, we are able to discriminate the four models of motion with high accuracy (5-fold cross validation score) across a range of sample sizes and track lengths (S2. Fig.) (see Supp. Methods for discussion on classification model selection). We present the performance of a sample Random Forest classifier, trained on simulations of 100 time steps with 1000 members per class (shown in Fig. 1.1A), in a confusion matrix. A confusion matrix describes the errors made during classification in a matrix by listing the true classes as rows and the predicted classes as columns with values of the matrix cells representing the number of observations for each true class : predicted class pair (Fig. 1.1D) [60]. In this way, values along the matrix diagonal represent correct predictions, and values off the diagonal represent incorrect predictions. As seen in the matrix, there is little confusion between Levy flights, fBm, and the random walks, but some confusion between unbiased and biased random walks.

The features important for effective classification can be determined by measuring the decrease in accuracy when each feature is removed from classification. For each of the simulated populations generated, we found the top 10 most important features based on this metric. We present the 10 features that appear most often across classifiers, and the mean decrease in accuracy associated with removal of each feature (Fig. 1.1E). We find that the non-Gaussian parameter and other metrics of the displacement distribution, as well as metrics of self-similarity, dominate this list. This supports the notion that the non-traditional features provided by Heteromotility can be useful for



Feature	Count	Mean Decrease in Accuracy
nongauss	32	0.0920
autocorr_1	24	0.0749
disp_skew	35	0.0709
disp_var	36	0.0684
autocorr_2	27	0.0628
progressivity	34	0.0624
rw_netdist	33	0.0589
MSD_slope	20	0.0579
rw_kurtosis01	23	0.0513
autocorr_3	24	0.0508

Fig 1.1. Simulated models of motion can be differentiated based on Heteromotility features. (A) Representative t-SNE visualization of Heteromotility feature space determined for simulated motion models (1000 members/class, 100 time steps)(perplexity = 50). (B) Unsupervised clustering accuracy for simulated data across a range of sample sizes and (C) track lengths. (D) The Confusion Matrix of a representative Random Forest classifier trained to distinguish four simulated models of motion shown in (A), mean accuracy of 99.8% (5-fold CV). (E) Features ranked by importance for Random Forest classification, where importance is determined as the decrease in accuracy when the feature is removed across all 36 simulated sample populations. The number of times the feature appears in the top 10 most important across the 36 populations is reflected in the 'Count' column. (F) Accuracy of classifying simulated motion models with varying parameters using hierarchical clustering and a Random Forest classifier. 10 populations of the four simulated motion models with varied parameters were generated and classified.

discriminating biologically relevant models of motion. These results demonstrate robust detection of heterogeneous motility phenotypes using a rich space of motion features. We also find high accuracy Random Forest classification of simulations with varied parameters (Fig. 1.1F). As a proof-of-concept, this analysis of simulated motion indicates that the Heteromotility feature set is sufficient to recognize different motility phenotypes where they exist.

1.2.3 Wild-type and transformed MEFs occupy a shared region of motility state-space

Utilizing the Heteromotility software, we are able to analyze behavior in live cells and map this behavior into a cell state space. We initially sought to determine if this cell behavior state space was continuous, with cell states existing along a spectrum, or discrete, with a limited set of states cells could adopt. To answer this question, we first employed the mouse embryonic fibroblast (MEF) culture system. Wild-type MEFs (WT MEFs) and MEFs transformed with *c-Myc* and *HRas-V12* overexpression constructs (MycRas MEFs) [61], which serve as a cancer model, were timelapse imaged for 8 hours to capture motility behavior. Images were segmented and tracked (see Methods), with cell centroid coordinates used as the cell location for tracking. By considering cell centroids, process extensions and retractions are considered as motile behaviors, in addition to displacement behaviors that would not be captured by an analysis of morphology alone [39]. Analysis of cell paths with the Heteromotility software revealed that WT and MycRas MEFs differentially occupy different sections of a shared motility space, as visualized by t-SNE (Fig. 1.2A). Strikingly, MycRas MEFs occupy a sub-region of wild-type motility space, rather than a unique region.

We applied hierarchical clustering (Ward's method) [58] to the motility state space generated from pooled WT and MycRas MEFs to identify heterogeneous motility phenotypes (Fig. 1.2B), considering a set of cluster validation indices to select the optimal number of clusters [62](See Supp. Methods for complete details). Clustering was performed on the first 30 principal components of the motility feature set, as this dimensionality preserves $> 95\%$ of variation for each system we study (Fig. 1.9). The Silhouette value is a metric of cluster validity on the interval [-1, 1], taking into account the similarity of samples within a cluster and the difference between clusters [63]. Higher values indicate that samples within a cluster are similar and clusters are distinct. The cluster partition displays a positive Silhouette value, indicating an appropriate cluster structure. Feature mean values between clusters are also confirmed to be significantly different by multivariate analysis of variance (MANOVA) [64]. This result suggests that the state space is continuous, but can be decomposed into a set of overlapping states with characteristic behavioral phenotypes. We define these motility states as heterogeneous based on significant differences from one other in feature

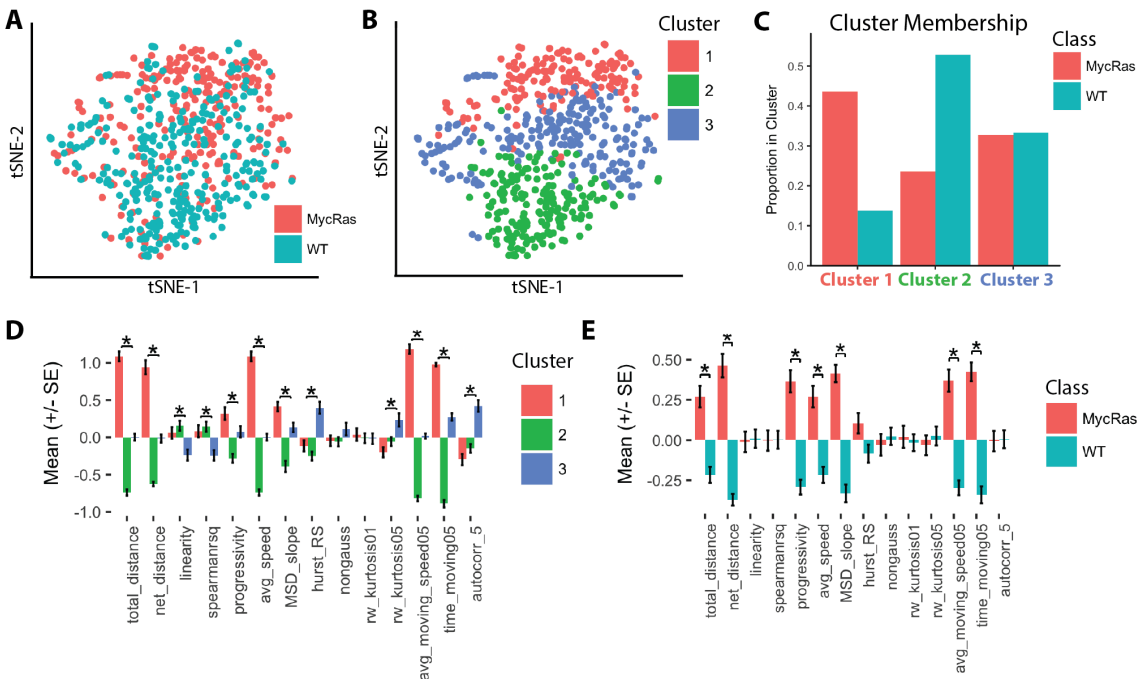


Fig 1.2. Wild-type and Myc/Ras transformed MEFs show differential occupancy within a shared motility state-space. (A) t-SNE visualization of wild-type (blue) and Myc/Ras transformed (red) MEFs in motility space (perplexity = 50). (B) Hierarchical clusters visualized with t-SNE (MANOVA, $p < 0.001$; Silhouette $S_i = 0.21$). (C) Proportion of wild-type vs. transformed cells occupying each cluster. (D) Comparison of a subset of normalized feature values between clusters (* : Holm-Bonferroni corrected $p < 0.01$ by ANOVA) and (E) wild-type and transformed MEFs (* : Holm-Bonferroni corrected $p < 0.01$ by t-test). $n > 250$ cells per condition (pooled) from six independent experiments.

space. This definition of heterogeneity is also followed for the other cell systems we investigate in this work.

WT and MycRas MEFs are found to differentially occupy different clusters within the shared set, as may be expected from their initial distributions in the space (Fig. 1.2C). MycRas MEFs preferentially occupy Cluster 1, characterized by the highest average speed, proportion of time spent moving, and distance traveled, indicating a motile state. Conversely, WT MEFs preferentially occupy Cluster 2, characterized by the lowest average speed and time spent moving, and progressivity (net distance / total distance), indicating a less motile state (Fig. 1.2D). A proportion of both MycRas and WT MEFs occupy Cluster 3, a population characterized by high kurtosis and progressivity, indicating a Levy flight like motile state that performs "jumping" motions. Observed visually, Cluster 2 cells move relatively little, Cluster 1 cells progress smoothly, and Cluster 3 cells exhibit

erratic motion. The transformation state-dependent distribution of MEFs between behavioral clusters is reproducible across all "round-robin" groupings of 4 out of the 6 biological experiments we perform (Fig. 1.10)(Supp. Methods). The clusters are also identifiable using only a small subset of features, or even a single feature, and the relative distribution of WT and MycRas cells within them is maintained (Fig. 1.11).

To estimate whether local cell density played a role in determining a cell's motility behavior, we estimate a "local cell density index" based on the sum of inverse squares of distances between a cell and its neighbors using our cell tracking data. Constructing linear regression models between this local cell density index and each of our measured motility features, we find no meaningful relationships (Fig. 1.9). However, we note that our tracking data purposefully does not track every cell, such that the local cell density index we estimate is an imperfect representation of local cell density. We also note that our experiments are designed to minimize cell-cell contact, such that our lack of detected cell density influence may be the simple result of sparse growth conditions failing to reach a quorum sufficient to influence motility behavior.

MycRas and WT MEF cluster preferences are statistically significant by Pearson's χ^2 test of the *transformation state* $\times$ *cluster* contingency table ($p < 0.0001$). Considered as a population, the MycRas MEFs demonstrate higher progressivity, mean squared displacement, time moving, average speed, and self-similarity metrics than WT MEFs. This indicates that as a population they spend more time moving in a directed manner (Fig. 1.2E).

These results suggest that high motility and low motility states exist in both wild-type and transformed MEFs. Oncogenic transformation by *c-Myc* and *HRas-V12* may then be viewed as an input that leads a larger proportion of cells to adopt the high motility state, rather than introducing a novel phenotype unseen in WT cells. This observation is consistent with studies of motility behavior in the context of cancer, which have long suggested that increased motility is a phenotype of malignant cells [65, 66]. Functionally, the increased motility of neoplastic cells may be related to their potential for tissue invasion [67, 68].

Importantly, the motility behavior of a cancer cell population may be indicative of disease progression and outcomes [69]. We trained supervised classification models to predict if a cell was wild-type or transformed based on its motility behavior. Support vector machine (SVM) classifiers

are a set of models that learn a decision boundary between classes. SVMs have shown efficacy in many problem domains [70]. SVM classifiers were trained on the top 62% features selected by ANOVA F-value in a round-robin fashion for cross-validation, training on four experiments and testing on the remaining two iteratively. On average, round-robin trained SVM classifiers are able to classify individual cells as wild-type or MycRas transformed with ˜70% accuracy based on motility features alone (Fig. 1.9). Parameters for feature selection and the SVM classifier were identified by a grid search (see Supp. Methods for details). While clusters are identifiable with only a single feature, we find classification accuracy decreases if more than ˜62% of features are omitted (Fig. 1.11). The most important features for classification were estimated based on the decrease in round robin classification accuracy when a feature was removed, with larger decrease considered more important. The average moving speed, time spent moving, kurtosis, and autocorrelation are the most important features by this metric (Fig. 1.11).

1.2.4 Myoblasts display distinct motility states robust to perturbation

Our long-term goal is to understand the cell as a unit in a dynamical system, which entails not just identifying states, but also determining how various inputs from the extracellular environment may drive transitions between states. Mouse myogenic progenitor cells provide a well-characterized system in which chemical signals alter the behaviors of a mechanically-active cell type [71–74], and dynamic activation and lineage commitment processes can be manipulated [75, 76]. In light of this long-term goal, we applied Heteromotility analysis to the myogenic system to ask how myogenic cells occupy state space.

Myoblasts are the transit amplifying progenitor of the skeletal muscle, produced as the daughters of muscle stem cells [75, 76]. Myoblast motility has direct functional relevance, as muscle progenitors translocate along the muscle fiber to sites of injury during muscle regeneration [73, 77] and transverse fibers during development [78]. Primary myoblasts were timelapse imaged for 8 hours, with and without stimulation by the growth factor FGF2. FGF2 is a known mitogen, inhibitor of differentiation, and possible chemoattractant in myogenic cells in culture [79–81]. *In vivo*, FGF2 is released during muscle injury [82, 83], promoting expansion of the myogenic progenitor pool and possibly migration to sites of injury. As FGF2 is known to elicit different functional behaviors

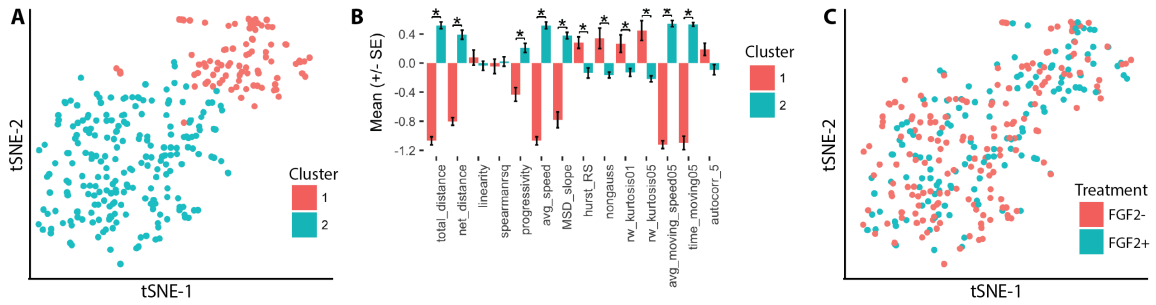


Fig 1.3. Myoblasts display distinct motility states, shared by both FGF2+ and FGF- conditions. (A) t-SNE visualization of hierarchical clusters in myoblast motility space (perplexity = 30) (MANOVA, $p < 0.001$; Silhouette $S_i = 0.31$). (B) Normalized feature means for motility state clusters (* : $p < 0.05$, Holm-Bonferroni corrected t-test). (C) t-SNE visualization of FGF2 treated (blue) and untreated (red) myoblasts in motility space. $n \approx 150$ cells per condition, taken from two separate animals, with total $n = 308$ cells.

in myogenic cells, introducing this perturbation allows us to evaluate the robustness of myoblast motility states under multiple growth conditions.

Applying hierarchical clustering, two motility clusters are detected (Fig. 1.3A). This partitioning scheme has a positive Silhouette value [63], significantly different cluster means by MANOVA, and optimizes cluster validation metrics (Table S1). Cluster 1 is a less motile state characterized by lower average moving speed, distance traveled, and a higher kurtosis. Conversely, Cluster 2 is characterized by a higher total distance, average speed, mean-squared displacement (MSD), and proportion of time spent moving. These characteristics indicate that myoblasts occupy a two state system with a consistently motile state (Cluster 2), and a less motile state that exhibits more Levy-flight like displacement behavior (Cluster 1). Observed in videos, cells in Cluster 1 are less motile, and display less directed motion than cells in Cluster 2 (S3, S4 Video). Clustergram visualizations of the myoblast clusters using multiple linkages are provided (Fig. 1.12).

Both FGF2 stimulated (FGF2+) and unstimulated (FGF2-) cells co-occupy both states, with no notable preference in state space induced by either condition (Fig. 1.3C). This is confirmed quantitatively as a lack of preferential cluster occupancy between FGF2 treated and untreated cells (χ^2 test $p > 0.05$ of *FGF2-treatment* $\times$ *cluster* contingency table). This negative result suggests that FGF2 does not induce a notable effect on myoblast motility under these conditions.

1.2.5 Muscle stem cells display motility states reflecting activation

Until this point, the cells analyzed have not been undergoing any dramatic phenotypic transitions. In contrast, stem cells are specifically designed to undergo dynamic activation and differentiation processes that radically reshape cell phenotype on an hours to days long timescale. Would it be possible to observe such phenotypic transitions through the lens of cell behavior changes? We apply Heteromotility analysis to the muscle stem cell (MuSC) system during activation from quiescence in an attempt to observe dynamic transitions in cell behavior. MuSCs undergo an exit from quiescence in cell culture conditions, entering an activated state over a roughly 48 hour window, providing a model of a dynamic process where behavior state transitions would be expected [75, 84]. Heterogeneity within the MuSC population is well appreciated [11], suggesting that motility behavior may also be heterogeneous during activation. Motility behavior is also relevant to MuSC function due to the physiological motility behavior of muscle progenitors during regeneration, as noted above.

Primary MuSCs were isolated from limb muscles by FACS ($PI^-/CD31^-/CD45^-/Sca1^-/VCAM^+/\alpha 7\text{-integrin}^+$) [85] and seeded on sarcoma-derived ECM coated well plates. After 24 hours in culture, MuSCs were timelapse imaged for 8 hours in DIC. At this stage, MuSCs have begun to activate ($MyoD^+$), but are not yet committed to differentiation ($MyoG^+$) [84]. Visualizing hierarchical clusters (Ward's method) of MuSC motility features with t-SNE, it is apparent that multiple motility subpopulations are present (Fig. 1.4A). We identify three distinct clusters with notably different phenotypes. This partitioning scheme has a positive Silhouette value [63], significantly different cluster means by MANOVA, and optimizes cluster validation metrics, as above (see Supp. Methods for complete discussion). As in myoblasts, the clusters appear to separate based on differences in total distance, average speed, and time moving. Additionally, clusters segregate based on the linearity and progressivity of motion.

Cluster 1 is characterized by the lowest total distance, average speed, time moving, progressivity, and MSD. Observed visually, cells in Cluster 1 are immotile and appear morphologically rounded, lacking any filopodia characteristic of myogenic activation and motility (S5 Video). Clusters 2 and 3 are characterized by increasing measures of total distance, average speed, and time spent moving. Cluster 3 exhibits the highest kurtosis and non-Gaussian parameter, suggesting a jumping,

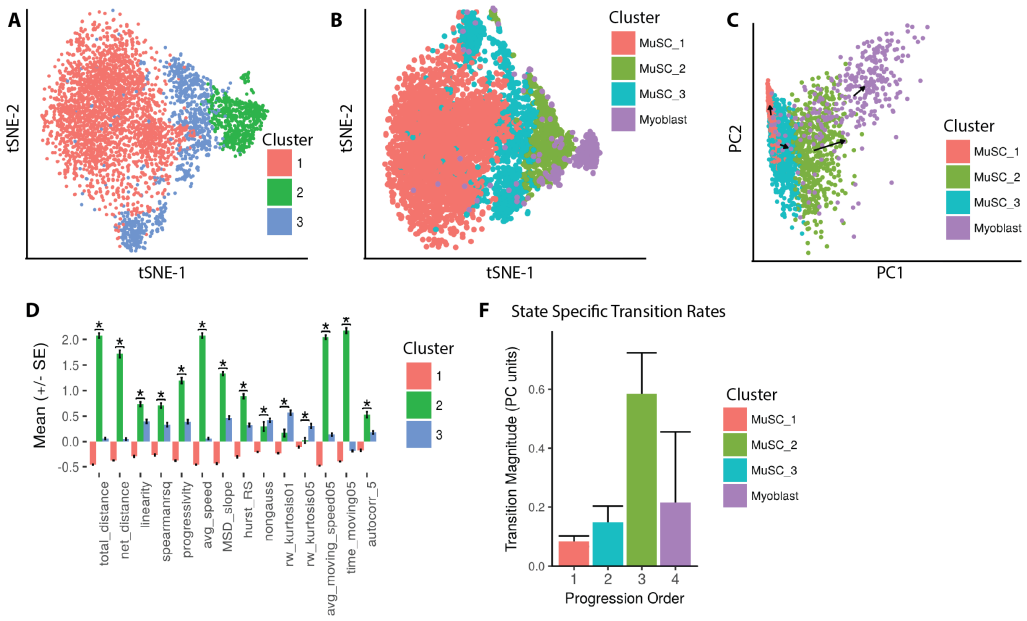


Fig 1.4. MuSC motility states reflect progressive myogenic activation. (A) t-SNE visualization of distinct motility states, as detected by hierarchical clustering (colors)(MANOVA, $p < 0.001$; Silhouette $S_i = 0.197$)(perplexity = 70). (B) t-SNE visualization of MuSC motility states and myoblasts in a shared space suggesting progressive myogenic motility states. (C) Average transition rate vectors for MuSC motility states (arrows, scaled for presentation) demonstrating transitions toward the myoblast motility phenotype over time in a shared PCA space. (D) Subset of normalized feature means for each motility state (* : p $<$ 0.05, Holm-Bonferroni corrected ANOVA). (E) The magnitude (+/- SEM) of the mean transition vector for all cells in a given state. All data represent analysis of $n > 1800$ cells per condition pooled from three animals, with total $n = 4316$ cells.

Levy-flight like state. Cluster 2 exhibits the highest time moving, moving speed, progressivity, linearity, net distance, and MSD, indicating a more directed motility state. Visually, Cluster 2 cells largely move in a single direction, while Cluster 3 cells more often turn back and retrace a portion of their previous path (S6 Video)(Fig. 1.4D)($p < 0.001$ by ANOVA for all features displayed).

MuSCs incubated with FGF2 display a similar distribution in motility space compared to untreated cells (Fig. 1.13). Quantitatively, FGF2 MuSCs display no motility state preference relative to untreated cells (Fig. 1.13)($p > 0.05$, χ^2 test of *FGF2-treatment x cluster* contingency table). On a population level, FGF2 does not appear to significantly alter any of the motility features calculated by our software. This negative result suggests that FGF2 may not play a role in

regulating motility in these conditions at this stage of myogenic activation.

Visualized in a shared t-SNE space with myoblasts, the order of MuSC motility states suggests a progressive motion toward the myoblast state space as total distance, time moving, and speed metrics increase (Fig. 1.4B). This is confirmed in a linear state space using principal component analysis (PCA), in which MuSC states and myoblasts segregate primarily along the first principal component (PC1)(Fig. 1.4C). The top 20 features contributing to PC1 in this shared space are metrics of moving speed and time spent moving, suggesting that the features which vary most between quiescent MuSCs and activated myoblasts are related to the speed and frequency of movement. These quantitative results are in line with qualitative observations, in which quiescent cells begin in a relatively immotile state and gradually increase the speed of their motility as they activate in culture. These phenotypic changes also align well with the *in vivo* role of activating MuSCs. Upon injury, MuSCs must activate from quiescence and translocate along the fiber to the site of tissue damage, necessitating more motile behavior [72, 77]. This result indicates that MuSC motility states reflect progressive states of activation toward the eventual myoblast motility state space.

We explored this possibility with pseudotime analysis, which attempts to reconstruct a dynamic cellular process from high-dimensional single cell data under the assumption that the process is ergodic and cells move in a continuous manner through feature space over time [18, 86]. Analytically, this is most commonly achieved by fitting a minimum spanning tree through the data in reduced dimensional space, interpreting the tree's longest axis as a temporal axis. Pseudotime analysis in the MuSC motility state space reflects the view that MuSC states are ordered in a progressive series (Fig. 1.13).

To further confirm that the MuSC motility states reflect states of activation, we quantified state transition rates and directions for cells in each MuSC state and myoblasts. Cell paths were divided into three equal length tracks ($\tau = 20$ frames) and Heteromotility features were extracted for each of these subpaths. The state of a cell during each time interval τ was defined in two dimensions as the cell's location along the first two principal components of a shared MuSC and myoblast PCA space. Transitions for each cell were calculated as the vector between sequential 2D state locations. The mean transition vector for a given cluster was calculated as the mean of all transitions made by

all cells assigned to that cluster.

Visualizing transition rates as vectors originating from each cluster centroid in PCA space, it is evident that cells in each MuSC state progress toward the next state in the sequence over time (Fig. 1.4C). The myoblast motility states are present at the end of the sequence, suggesting that MuSC states represent different stages of myogenic activation. Notably, the transition rates increase as a function of state progression, suggesting the kinetics of activation are non-linear (Fig. 1.4F). In this way, we provide quantitative measurements of transitions between states of myogenic activation in single cells for the first time.

1.2.6 MuSC motility states are more dynamic than MEF and myoblast motility states

We next sought to investigate the dynamics of our motility state systems. A key benefit of using behavior features to define states is that a single cell can be subjected to a state assay at multiple time points, allowing state transitions to be detected. How long does a cell reside in a particular state? Can every state transition to every other state, or are transitions restricted to exist between particular states, as would be the case in a state automaton? The answers to such questions would help clarify the computational logic underlying cell behavior.

To visualize and quantify the dynamics of the motility state systems, we applied coarse-grained probability flux analysis (cgPFA), as implemented by Battle *et. al.* [87] (Supp. Methods). In this method, the first (PC1) and second (PC2) principal components are segmented into k bins, and we define each unique combination of bins on PC1 and PC2 as a unique state. To define cell state at multiple points using motility features, cell paths are segmented into τ length subpaths and motility features are extracted from each subpath. Cell state is defined for each subpath based on its position in coarse-grained PCA space, where each binned coordinate is treated as a state. Cell states are compared from one subpath to the next to quantify the dynamics of the motility state system (see Supp. Methods).

As a validation of our implementation, we performed cgPFA on simulated cell paths that varied their model of motility on a defined interval ($\tau = 20$ time points) and compared them to invariant

simulations, using subpaths of the same length as the variable model's states ($\tau = 20$ time points). For each bin, a state transition rate is calculated as the vector mean of transitions from that state bin into neighboring states in the von Neumann neighborhood. These transition rates are visualized as arrows atop each state bin. Arrow direction represents the direction of transition rate and arrow length represents the rate magnitude. State bins with high transition consensus therefore display longer arrows, indicating that most cells transition in the same direction. As a measure of state stability, the divergence of the vector field is displayed as a heatmap. States with negative divergence may be considered metastable, with more cells entering than exiting. A Levy flight model transitioning to a random walk displays a highly directed set of state transitions (Fig. 1.5A), as compared to a simple random walk which displays minimal directionality (Fig. 1.5B).

Applying cgPFA to our biological systems with subpaths of length $\tau = 20$ time points (130 minutes), both WT MEF and MycRas MEF systems display no obvious state flux (Fig. 1.5C, D; Fig. 1.14). Topographically, MEF systems display a 'basin' of metastable states. This region has near zero divergence and low transition rates. States on the outer periphery of this metastable region have higher divergence and transition rates, indicating that these states are less stable (Fig 1.5C, D; Fig. 1.14). Visualizing state space divergence in three dimensions, the metastable states appear as a central valley, while the unstable states appear as peaks (Fig. 1.5G). Myoblast motility state systems appear similar to the MEF systems by simple observation. The topology appears to be dominated by a metastable basin at the center, surrounded by unstable states on the edge of this basin. Topology is comparable between FGF2 treated and untreated conditions (Fig. 1.5E, Fig. 1.14).

MuSC motility state systems appear more dynamic than MEF and myoblast systems. This transition directionality is apparent when viewing the transition vector fields, with many transition rates leading toward a metastable 'valley' and a metastable basin on the edge of state space (Fig. 1.5F, Fig. 1.14). Visualized in three dimensions, these metastable regions are clearly visible as valleys in state space bordered by unstable 'ridges'. We apply these cgPFA analyses with multiple temporal and course-grained resolutions, and find that the results are qualitatively similar for multiple state definition time scales and course-grained binning schemes (Fig. 1.15).

To determine how long MuSCs occupy a given state, we characterized the dwell times of each

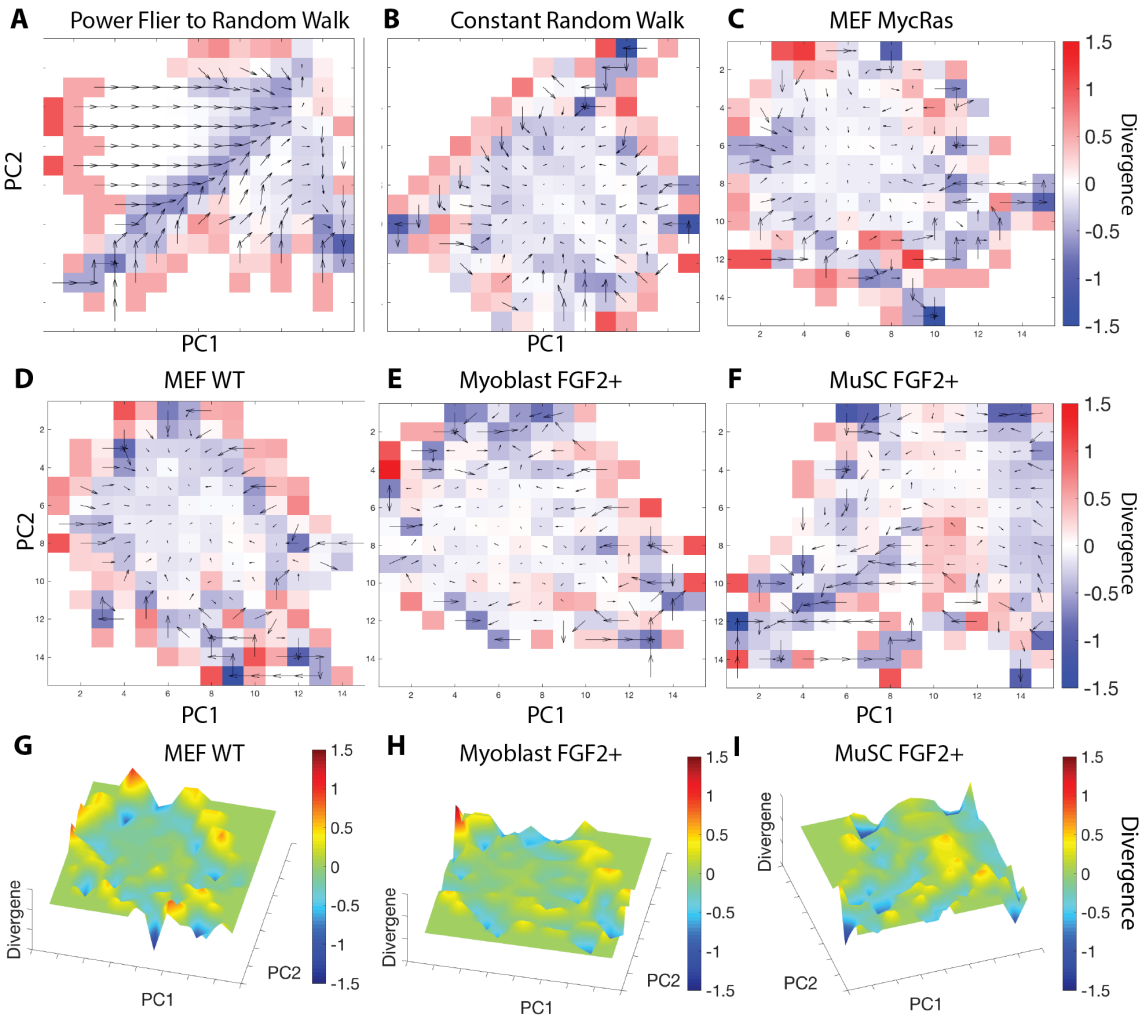


Fig 1.5. Coarse-grained probability flux analysis (cgPFA). cgPFA of (A) Levy flier to random walk simulations and (B) Random walk invariant simulations. cgPFA of (C) MycRas MEF, (D) WT MEF, (E) Myoblast FGF2+, and (F) MuSC FGF2+ motility states with subpaths of length $\tau = 20$ time points (130 minutes). Each unique combination of bins between PC1 and PC2 is considered as a unique state. Arrows represent transition rate vectors, calculated for each state bin as the vector mean of transitions into the neighboring states in the von Neumann neighborhood. Arrow direction represents the direction of these transition rate vectors, and arrow length represents transition rate vector magnitude. Underlying colors represent the vector divergence from that state as a metric of state stability. Positive divergence indicates cells are more likely to leave a state, while negative divergence indicates cells are more likely to enter a state. 3D representations of (G) MEF WT, (H) Myoblast FGF2+, and (I) MuSC FGF2+ motility state divergence.

occupied state in the course-grained PCA space by exponential decay curve fitting, providing a time constant for each state. In these course-grained state spaces, dwell times appear exponentially distributed with time constants ranging from $\tau_{TC} \approx 1$ to $\tau_{TC} \approx 3$ (Fig. 1.16). The mean dwell

times for state bins range from $\tau_{dwell} = 1$ to $\tau_{dwell} \approx 2.2$, indicating that most cells transition at least once during the time course. Dwell times appear exponentially distributed, suggesting that the state transition process is memoryless on the timescales we observe. Time constants are positively correlated with the number of cells occupying a state, supporting the notion that states of high occupancy are metastable and therefore characterized by longer dwell times (Fig. 1.16). Topology, dwell times, and transition directionality are comparable between FGF2 treated and untreated conditions (Fig. 1.5I, Fig. 1.14, Fig. 1.16). These results suggest that the MuSC motility state system is more dynamic than the MEF or myoblast systems, consistent with the dynamic MuSC activation process.

1.2.7 MuSC motility states break detailed balance

A system with a discrete set of states at equilibrium should obey the law of detailed balance, such that each individual state transition $A \rightarrow B$ occurs at the same rate as the reverse transition $B \rightarrow A$. Systems that break detailed balance transition between states in a directed manner, such that future behavior can be predicted by current state. Biological systems frequently break equilibrium when undergoing directed processes, but confirming that detailed balance is broken in a given scenario can prove challenging [87]. A system in detailed balance would be expected to exhibit equal and opposite transition vectors with no observable pattern of transitions. In our MuSC systems, visualizing transition rates by PFA suggests that detailed balance may be broken, as shown by the directed nature of transition rate vectors (Fig. 1.5, Fig. 1.14).

To confirm statistically that detailed balance is broken in these systems, we defined N-dimensional state spaces based on the first N coarse-grained principal components and performed PFA in these spaces (ND-cgPFA)(see Methods). A $2N$-dimensional matrix is generated representing every possible set of state transitions in a given N-dimensional state space. For example, a 1 dimensional state space is represented as a 2 dimensional matrix, each dimension coarse-grained into k bins (rows, columns). One dimension of this 2D space represents a cell's initial state position at one time point, and the other represents the destination state position at the next time point (as in Fig. 1.6A, B). This pattern is repeated for the construction of higher dimensional spaces. For instance, a 2D state space is represented as a 4D matrix, where the *first two* dimensions describe initial state,

and the *next two* dimensions describe final state. The number of cells that exhibit each transition are recorded as the value of the corresponding position in the matrix (see Methods for further description).

A system in detailed balance would be expected to display symmetry for each pairwise set of transitions. To determine if each of our systems was in detailed balance, we performed ND-cgPFA at several levels of dimensional ($N \in \{1, 2, 3, 4\}$) and coarse-grained (bins $k \in \{2, 3, 5, 7, 10, 15, 20\}$) resolution. At each level of resolution, we test each of the k^{2N} possible pairwise transitions for balance by the binomial test ($H_0 : p = 0.50$).

To validate this method, we performed cgPFA on simulated cell paths generated using variable and invariant models of motion. A Levy flier that transitions to a random walk shows clearly unbalanced pairwise transitions (Fig. 1.6A) at one dimension of resolution, while an invariant random walk displays symmetry about the diagonal and balanced pairwise transitions (Fig. 1.6B). A simple binomial test for pairwise transitions finds multiple unbalanced transition pairs in the variant model, but not the invariant model (Fig. 1.6A, B).

N-dimensional matrices cannot be visualized in their totality as in the 1D case. As noted above, we tested all k^{2N} possible transitions for balance at multiple dimensional and binning resolutions for each system. Here, the five transition pairs that are most unbalanced out of the possible set of k^{2N} transitions (by p value of the binomial test) in a given system and state space are presented as a 5-by-2 heatmap, where columns represent the initial and final states in a transition pair. We present the dimensional and coarse-grained resolution revealing the most asymmetry for each system, as noted above each heatmap. Visualizing ND-cgPFA pairwise transitions for MEF (Fig. 1.6C), myoblast (Fig. 1.6D), and MuSC state systems (Fig. 1.6E, F) demonstrates that transitions with some degree of unbalance exist in each of the systems. By the binomial test (Benjamini-Hochberg corrected), only the MuSC system (FGF2+ and FGF2-) displays significantly unbalanced transitions in any of the course-graining schemes tested, confirming that detailed balance is broken. Our ND-PFA test is biased toward Type II error (false negatives) rather than Type I error (false positives) for multiple reasons, providing further confidence that the detailed balance breaking we identify is valid. It is therefore possible that detailed balance is also broken in the MEF and myoblast systems and our tools are simply not sufficient to detect this asymmetry.

We performed this ND-cgPFA analysis for multiple values of the timescale parameter τ and find that these results are consistent. At each value of τ, the MuSC system demonstrates broken detailed balance and the MEF and myoblast systems do not. On a very short timescale, it is unlikely that detailed balance is broken for a motility state system, even in a dynamic system like activating MuSCs. In line with this biological prior, we find that detailed balance is *not* broken in the MuSC system when we perform ND-cgPFA with the same number of $\tau = 20$ length windows, but set them to overlap with a stride $s = 1$. In this overlapping scheme, only two time steps of difference are present between the first and final temporal window, accounting for only 13 minutes in real time in our experiments (Fig. 1.17).

As additional confirmation of MuSC motility state dynamism, we performed this symmetry breaking analysis using hierarchical clustering to define cell state over the whole motility feature space, rather than coarse-grained location along the principal components (Fig. 1.18). As with ND-cgPFA, our positive control variable model breaks detailed balance by the binomial test, while our negative control invariant models do not. MuSC systems again demonstrate greater pairwise asymmetry than MEF or myoblast systems in this test (Fig. 1.18).

An important question is raised when detailed balance is broken. Is the system stationary, with the same number of cells in each state over time, or non-stationary? We evaluated the stationary nature of the MuSC systems by forming a contingency table comparing state occupancy between time points for each set of length $\tau = 20$ subpaths in the dataset and find a significant difference ($p < 0.05$, χ^2 test) for each scale where detailed balance breaking is detected. These results collectively demonstrate the more dynamic nature of the MuSC system and demonstrate that the MuSC motility state system is non-stationary and breaks detailed balance.

1.3 Discussion

Cell behavior phenotypes represent an under-exploited opportunity to explore the heterogeneity of cell systems. In contrast to existing methods based on destructive molecular assays, cell behaviors such as motility can be tracked in a single cell over time, allowing for measurement of phenotypic state transitions. Previous work to quantify morphological dynamics has allowed for quantification

of cell state transitions in this manner [20, 24, 31]. Quantification of motility behaviors in addition to morphological dynamics may allow for detection of transitions that are not revealed by morphological dynamics alone.

Our Heteromotility analysis software provides access to one portion of this rich motility feature space for analysis. Applying analysis techniques common in other single cell assays, we demonstrate that Heteromotility features are sufficient to distinguish different motility phenotypes and provide novel insight in multiple biological contexts. In addition to detecting heterogeneity within populations, Heteromotility analysis was also useful to quantitatively describe perturbations on the population level.

In our three biological systems, overlapping but characteristic motility states are present in a continuous state space. Notably, a shared set of motility states is conserved between WT and MycRas MEFs, despite the dramatic perturbation of neoplastic transformation. A similar phenomenon of conserved phenotypic states has been described for cell shapes. Despite a large number of perturbations, multiple cell systems displayed a fairly limited set of cell morphology states, with perturbation merely altering the prevalence of these states [4, 88]. These results suggest that cell systems may have a discrete set of phenotypic states despite a much greater diversity in molecular organization, with perturbations acting largely to alter the distribution of cells within these states, rather than elicit novel behavior.

Although these state definitions are robust to perturbation in our conditions, the distribution of cells among these states appears to be dynamic. This is demonstrated by the state preferences of WT and MycRas transformed MEFs. Considering the robustness of motility state definitions, state transitions may act as a mechanism of population level phenotype change. This is not necessarily in opposition to a model of motility regulation in which cell phenotypes shift within a given state. Harkening to the subsumption architectures of robotic control systems [89], a higher level state determinant, such as oncogene expression, may be viewed as inducing a preference for the selection of behavioral states, and substates within those states, by a more direct effector mechanism, in this case the cell motility machinery. The mechanisms of state transitions and substate preference may work in synchrony to specify population level phenotypes.

Within the MEF systems transformed cells preferentially occupy the more motile state, suggesting

that neoplastic transformation shifts the distribution of cells within the motility state system. This state preference is strong enough to allow for better-than-chance machine learning classification of WT and MycRas MEFs based solely on motility features. Training machine learning classifiers on quantitative motility features and ground truth patient outcomes may potentially serve as an additional cancer diagnostic metric.

Most single cell analysis methods currently used to investigate heterogeneity are not capable of assaying a single cell across a time course, but rather provide a detailed snapshot of a cell at a single time point. Real-time observation of individual cells may elucidate novel properties of a heterogeneous cell system, as pioneered by analysis of morphological dynamics. Here, we combine the real-time nature of cell behavior observations with Heteromotility analysis to investigate the dynamics of motility states.

In the context of MuSCs undergoing myogenic activation, we find a set of motility behavior states progressively more similar to myoblast motility behavior. Quantifying state transitions within each of these states, we observe that cells within each state transition toward the next state in the series over time. State transition rates increase as a function of state progression, suggesting that MuSC activation is not simply a linear process. This observation may have implications for the study of MuSC heterogeneity. If the rate of phenotypic change in MuSCs is non-linear during activation, then heterogeneity in activation kinetics between cells will be exaggerated during a 'critical period,' where the most rapidly activating cells are not only more activated at that moment, but are moving toward a fully activated state more rapidly than their less activated counterparts. This is to our knowledge the first quantification of single cell transition rates between quiescent MuSC and myoblast phenotypes during myogenic activation. It follows that similar analysis of cell behavior may elucidate transition dynamics in other contexts of cell biology.

A state system at equilibrium displays detailed balance, in which pairwise transitions between all states are equal. We check for the presence of detailed balance in our motility state systems as a method of quantitatively discerning if cells are transitioning through state space in a predictable manner. The MuSC motility state system breaks detailed balance, while the MEF and myoblast systems do not. This may reflect underlying properties of each system, as MuSCs are undergoing a dynamic activation process on the timescale of imaging, while MEFs and myoblasts are not. Broken

detailed balance in the MuSC system indicates that the current MuSC state provides predictive information about the cell's future state, as cell's transit state space in a predictable way. This supports our observation of progressive activation states discussed above. Analysis of detailed balance breaking by real-time cell state inference from cell behavior may be a useful approach to detect dynamic biological processes and predictable phenotypic patterns.

Moving forward, we believe that single-cell, quantitative motility analysis has multiple applications. Quantitative motility features may allow discrimination of cell states by supervised classification, as demonstrated by our discrimination of neoplastic and wild-type MEFs. Supervised classification in this manner may be useful in diagnostic and biosensor applications. For example, motility metrics could be added to the cell morphology metrics commonly used to diagnose neoplasia from patient biopsy samples.

Heterogeneous cellular states within a population of cells may be detected by unsupervised clustering of motility features, similar to analyses of population heterogeneity performed using single-cell molecular assays. We demonstrate this use case by identifying heterogeneous motility states in MuSCs, which correspond to heterogeneous states of stem cell activation. Unsupervised identification of heterogeneous populations is useful in understanding the behavior of cell systems, as demonstrated by findings from single-cell molecular assays. Identifying heterogeneous motility behaviors may be of particular interest in contexts where motility is directly tied to cell and tissue function, such as tissue regeneration and immunological responses.

Analysis of cell behaviors such as motility also allows for measurement of the transition rates between cell phenotypes. We demonstrate this application by measuring activation rates in the myogenic system for the first time, and identifying the presence or absence of detailed balance across biological systems. Cell state transition analysis may be applied to measure the rate of dynamic phenotypic changes, such as stem cell differentiation, immunological activation, or neoplastic transformation. Each of these applications of quantitative motility analysis may be paired with exiting approaches for quantitative analysis of morphology to provide a more complete picture of cell behavior than either approach may yield alone.

1.4 Conclusion

Our Heteromotility software provides a means of defining cell states and quantifying transitions between them by quantitative analysis of one aspect of cell behavior. We demonstrate motility behavior analysis is capable of identifying unique motility states in simulations and three biological systems. In our biological systems, these states appear to have robust definitions when perturbed, but each cell's state identify appears dynamic. Real-time, quantitative assays of single cells such as Heteromotility analysis may reveal the dynamics of heterogeneous cell systems, which cannot be accomplished by terminal molecular assays. We demonstrate this approach by showing that MuSCs transition through a progressive set of motility states toward the myoblast motility state in a predictable manner, breaking detailed balance.

1.5 Supporting Information

1.6 Materials and Methods

1.6.1 Heteromotility Analysis

DIC images were segmented using custom segmentation algorithms, optimized for each of the systems we investigated. Tracking was performed with a modified version of uTrack [90]. Code is available on the Heteromotility GitHub page. All code for the Heteromotility tool is available on the Heteromotility Github page (https://github.com/jacobkimmel/heteromotility) and in the Python Package Index (PyPI). Detailed descriptions of the algorithms are provided in the Supplemental Methods.

1.6.2 Animals

All mice were housed at the University of California San Francisco following UCSF Institutional Use and Care of Animals Committee guidelines. Adult male C57Bl/6 mice (2-4 m.o.) were used for all myogenic cell experiments. Adult female C57Bl/6 mice (2-3 m.o.) were used as mothers to derive MEFs.

1.6.3 Myoblast and Muscle Stem Cell Isolations

Mononuclear cells were isolated from limb muscles of adult mice as described [91]. Myoblasts were isolated by negative-selection against *Sca1* using the EasySep Endothelial Selection kit (Stem Cell Tech., Vancouver, BC) and subsequent 5 minute preplating. Myoblasts were maintained in myogenic growth media (F10, 20% FBS, and [5 ng/mL] FGF2) on sarcoma-derived ECM (Sigma, St. Louis, MO) coated dishes. All myoblasts used in experiments were below passage 4.

Wild-type muscle stem cells were isolated by FACS after incubating the bulk mononuclear cell population with anti-VCAM-PE (BioLegend, San Diego, CA), anti-α7-integrin-649 (R&D Systems, Minneapolis, MN), anti-CD31-PE-Cy7 (BD Pharmingen, San Jose, CA), anti-CD45-PE-Cy7 (BD Pharmingen, San Jose, CA), and anti-Sca1-APC-Cy7 (BD Pharmingen, San Jose, CA). Propidium iodide was added as a live/dead marker. Muscle stem cells were selected based on a VCAM$^+$/α7-integrin$^+$, CD31$^-$/CD45$^-$/Sca1$^-$/PI$^-$ profile, as previously described [85].

1.6.4 MEF Cell Culture

Wild-type MEFs were isolated from E13.5 embryos as described [92]. MEFs isolated from two separate mothers were used for these experiments. Transformed MEFs were generated as described [61], generously donated by the authors. All wild-type MEFs used in experiments were from passages 2 - 5. All MEFs were maintained in DMEM, 10% FBS, 1% Penicillin / Streptomycin at 37$^{\mathrm{O}}$ C and 5% CO_2.

1.6.5 Timelapse Motility Imaging

All imaging was performed on a Nikon Ti using a 20X air objective and an Oko temperature and carbon dioxide incubation unit set to 37$^{\mathrm{O}}$ C and 5% CO_2. Images were captured by a Hamamatsu C11440-22CU camera (pixel size 6.5 x 6.5 μm). Twenty-five images with 10% edge overlap covering a 3.3 mm square at the center of each 96 well were captured. Focus was maintained throughout the timelapse using the Nikon PerfectFocus system. Each XY position was segmented and analyzed independently. XY positions that experienced overexposure or out of focus events were removed from analysis by manual screening of the data.

MuSCs were seeded at 500-1000 cells/well in 20 wells of a 96 well plate immediately after FACS sorting. After 23 hours to allow for adaptation to culture, media was exchanged for the relevant experimental medium and MuSCs were incubated for 1 hour to adapt. MuSCs were imaged at 20X in DIC for 10 hours with a temporal resolution of 6.5 minutes / frame. Myoblasts were passaged and plated on sarcoma-derived ECM coated optically clear plastic 96 well plates 24 hours prior to imaging at 300-500 cells/well. Media exchanges were performed 1 hr prior to imaging, in the same manner as MuSCs. Myoblast imaging was performed at 20X in DIC for 10 hours with a temporal resolution of 6.5 minutes / frame. For FGF2 perturbation experiments, 10 wells of a 96 well plate were imaged in growth media with [5 ng/mL] FGF2, and the remaining 10 were imaged in growth media with [0 ng/mL] FGF2. MEFs were similarly plated in optically clear plastic 96 well plates (without ECM coating) 16-20 hours prior to imaging at the 300-500 cells/well and imaged in the same manner. The initial 18 frames of imaging were omitted from downstream analysis for MEF and myoblast experiments, and the initial 30 from MuSC experiments to minimize the effect of initial XY drift in our imaging apparatus.

1.6.6 Image Segmentation

Brightfield images were segmented using custom segmentation algorithms, optimized for each of the systems we investigated. Segmentation algorithms are steerable filters [93] for edge detection, top-hat filtering [94], and Otsu thresholding [95], utilizing implementations available in the Matlab Image Processing Toolbox. These standard segmentation methods were combined with contrast adjustments, noise filtration, and background subtraction as needed for each system. Images were segmented independently for each field-of-view. Cells spanning multiple fields of view were not segmented. For tracking, the centroid of segmented cells was taken to be the cell location.

1.6.7 Heteromotility

All code for the Heteromotility tool and downstream analysis is available on the Heteromotility Github page and in the Python Package Index (PyPI) (https://github.com/jacobkimmel/heteromotility).

1.6.8 Cell Tracking

Tracking was of cell centroids was performed using a global nearest neighbor optimization approach, as implemented in uTrack [90]. Specific parameters were devised for each cell type to optimize tracking, and are provided in the Heteromotility Github code base. Only cells that remained in the same field-of-view throughout the timelapse were tracked. All tracks were visually screened for accuracy, and any erroneous tracking events (tracking of non-cellular objects, tracks crossing, etc.) were removed from analysis.

1.6.9 Distance and Speed Calculations

Total distance is calculated as the sum of all displacements along a given cell's path. Net distance is the distance between the initial and final points along the path. Progressivity is defined as the ratio of net distance to total distance traveled.

Minimum speed is the minimum displacement value along a cell's path. Maximum speed is the maximum displacement value. Average speed is the mean of all displacement values. Time Moving is the proportion of time a cell spends above a given threshold speed. Average moving speed is the mean speed of a cell while above a threshold value.

1.6.10 Turning Features

To determine the magnitude and direction of a cell's turning behavior, the cell's instantaneous direction is first established by performing linear regression on a range of points size 2δ about every point p_t in the time series T where $\delta < t < T - \delta$. The cell's movement direction relative to the x axis on the interval $[p_t - \delta, p_t + \delta]$ is considered to establish directionality to the regression. This vector R is assumed to represent the cell's direction at p_t. The cell's position $p_t + \tau$ is considered for a given time lag τ. If $p_t + \tau$ falls left relative to R, the cell is said to have made a left turn, and vice versa. The magnitude of each turn is calculated as the angle between the vector R and the vector from p_t to $p_t + tau$.

1.6.11 Correlation Features

The linearity of a cell's path is calculated as Pearson's r^2 value of a linear regression through all two-dimensional positions a cell occupied during the time course. Similarly, the monotonicity of a cell path is calculated as Spearman's ρ^2 for the same distribution of two-dimensional coordinates.

1.6.12 Displacement Distribution Features

Mean Squared Displacement Coefficient

Mean squared displacement (MSD) is calculated for the two-dimensional vector of positions X and a given time lag τ as

$$MSD(\tau) = \langle x_{t+\tau} - x_t \rangle$$

Here, τ denotes the given time lag, x_t denotes the position at time t, and $\langle\rangle$ denotes the mean value of an expression. MSD is calculated for τ in the range $[1, 30]$. The relationship of MSD to τ can be expressed in a general form $MSD(\tau) \propto \tau^{\alpha}$. The exponent α in this relationship can be easily estimated by power transformation of the resulting values, such that $\log MSD \propto \alpha \log \tau$ and α not represents the slope of a linear regression through $\log MSD$ vs $\log \tau$.

Kurtosis

To calculate the kurtosis of a cell's displacements, we consider displacement sizes for a variable time lag τ, so that X_τ represents the distribution of displacements for a given time lag τ. We calculate kurtosis as the standard fourth central moment:

$$kurtosis(X_\tau) = \langle x^4 - \bar{x} \rangle - 3$$

Kurtosis values for each cell path are normalized against the kurtosis of a random walker's displacement distribution, which displays a Gaussian kurtosis of 3.

Non-Gaussian Parameter

The Non-Gaussian parameter α_2 is calculated as described [53].

$$\alpha_2 = \frac{\langle \Delta x^4 \rangle}{3\langle \Delta x^2 \rangle^2} - 1$$

Here, $\langle\rangle$ denotes the mean of an expression and Δx denotes an individual displacement in the distribution. Effectively, the α_2 parameter denotes a ratio of kurtosis to variance for a distribution, as $kurtosis(X) = \langle x^4 - \bar{x} \rangle$, and $\sigma^2 = \langle x - \bar{x} \rangle^2$. For a Gaussian distribution, $\alpha_2 = 0$. As the tails of a distribution increase in weight, α_2 also increases. As such, α_2 acts as a metric of the Non-Gaussian nature of a given distribution, with higher α_2 values indicating that the distribution is less Gaussian and heavier tailed.

1.6.13 Self-Similarity Metrics

Hurst Exponent Estimation

Fractal Brownian motion (fBm) is a more general description of Brownian motion that allows for interdependence between displacements [96]. fBm is defined as a random Gaussian process parameterized by H:

$$B_H(t) = B_H(0) + \frac{1}{\Gamma(H + \frac{1}{2})} \int_{-\infty}^{0} [(t-s)^H - \frac{1}{2} - (-s)^H - \frac{1}{2}] dB(s) + \int_0^t [(t-s)^H - \frac{1}{2}] dB(s)$$

where Γ is the standard Gamma function, H is the Hurst parameter on the interval $[0, 1]$, t is time, and s is $t + \tau$ for a given time la g τ. The Hurst parameter describes the self-similarity of a given fBm process. A classical Brownian motion process with independent displacements has $H = 0.5$. $H > 0.5$ describes a process with positively correlated displacements, while $H < 0.5$ describes a process with negatively correlated displacements.

The Heteromotility software estimates the Hurst exponent based on Mandelbrot's Rescaled Range method, as described [55]. Briefly, for a given time series X of length N, the series is divided into a set of sub-series length $n = N/2^0, N/2^1, N/2^2, ..., N/2^\alpha$ where α is the largest power of 2 less

than $\frac{1}{2}N$.

For each of these sub-series length n, the cumulative sum series Z is calculated on a mean adjusted time series $Y = X_t - \langle x \rangle$. A range $R(n) = max(Z) - min(Z)$ is calculated, along with the standard deviation $S(n) = stdev(X)$. The average rescaled range $R(n)/S(n)$ is calculated for each sub-series length n. The relationship between $R(n)/S(n)$, n, and H can be expressed generally as $E[R(n)/S(n)] = Cn^H$ where C is a constant coefficient. A simple log-log transformation is used to estimate H as the slope of a line $\log R(n)/S(n) = H \log Cn$.

Autocorrelation

Autocorrelation describes the correlation of a given process with itself across an interval of time. Expressed formally, autocorrelation R as a function of t for a given time lag τ:

$$R(t) = \frac{E[(X_t - \mu_t)(X_{t+\tau} - \mu_{t+\tau})]}{\sigma_t \sigma_{t+\tau}}$$

We estimate the autocorrelation for a cell's displacement series X of length T with a given time lag τ as:

$$R(\tau) = \frac{1}{(T-\tau)s^2} \sum_{t=1}^{T-\tau} (X_t - \langle X \rangle)(X_{t+\tau} - \langle X \rangle)$$

where s is the standard deviation of X.

1.6.14 Motion Model Simulations

Simulated data sets analyzed in this study were generated for a range of track lengths $T \in \{50, 100, 500\}$ with $n = 50000$ members per class. Data sets for analysis were resampled from these populations (with replacement) with sample sizes $n \in \{100, 500, 1000, 5000\}$ members per class, three times each per sample size/track length combination. Results for unsupervised clustering and supervised classification accuracy are reported as averages across all sampled populations for a given parameterization (i.e. sample size, track length).

Simulations were initially generated with identical mean displacement magnitudes ($d_{xy} = 5$) for each model type to prevent the trivial discrimination of different models. Simulations were

also generated with varied displacement sizes, and different values of characteristic parameters to demonstrate robustness, as detailed below. These simulations were analyzed with a path length of $t = 100$ and sample size $n = 1000$.

Random Walks

We model an unbiased random walk is modeled as described [97]. For a walk of length T with step lengths t, a displacement is made in a random direction on the xy-plane for each time step. Displacement magnitudes are sampled from a Gaussian distribution with a given mean μ and variance $\sigma = \mu/5$. The mean displacement size is set to $d_{xy} = 5$ for initial experiments, and $d_{xy} \in \{10, 20, 50\}$ for experiments with varied parameters. Biased random walks are modeled in the same manner, but directionality is biased to step in a randomly chosen direction in the xy-plane $b\%$ of the time. The bias parameter is set to $b = 0.75$ for experiments with differing path lengths, and is varied to $b \in \{0.10, 0.70, 0.90\}$ for experiments with varied parameters.

Levy flight

Levy flights are modeled in the same manner as unbiased random walks, but displacement sizes are sampled from a Levy-like power law distribution $P(l_i) \propto l_j^{-\mu}$. We set the parameter $\mu = 2$ for initial experiments, as this models an optimized random search [49]. Experiments with varied parameters set $\mu \in \{0.5, 5\}$.

Fractal Brownian Motion (fBm)

A random walk is simulated as above with displacement sizes generated by a fractal Brownian motion process with a given Hurst coefficient $H = [0, 1]$. This fBm process is simulated by the Cholesky decomposition method, as outlined [98]. Briefly, a matrix $\Gamma = (R(t_i, t_j), i, j)$ is generated for $i, j = 0...n$ where $R(t, s) = (s^{2H} + t^{2H} - |t - s|^{2H})/2$. A matrix $\Sigma = \Gamma^{\frac{1}{2}}$ is computed as the lower triangle L returned by Cholesky decomposition of Γ. A vector v of length n containing normally distributed values is generated. The vector $u = v\Sigma$ represents a series of fBm displacements. For the simulation analysis in this paper, the Hurst parameter was set to $H = 0.9$ to generate a displacement

series with long-range memory in initial experiments. Experiments with varied parameters set $H \in \{0.1, 0.5, 0.99\}$.

1.6.15 Unsupervised Clustering

Hierarchical Clustering

All data were scaled $[-1, 1]$ with unit variance prior to cluster analysis. *Ward* hierarchical clustering was performed using the standard R stats library. The appropriate number of clusters k for each system was determined using a panel of 30 cluster validation indices (implemented by the NbClust R package [**?**]), considering the relative performance of different indices [99]. These indices were determined using the first thirty principal components for each system as features to avoid colinearity while retaining >=95% of total variation. The biological relevance of cluster phenotypes for each k was also considered. All clustering partitions utilized demonstrate a positive Silhouette index, providing evidence of an appropriate cluster structure [63].

Unsupervised clustering accuracy for simulated data was determined as the percentage of samples that reside in a cluster where that sample's type (i.e. Random walk, Levy flier) is the majority sample type for the cluster. For instance, a Random walker simulation that lies in a cluster with a majority of other Random walker samples is considered an accurate clustering, while a Power flier that lies in a cluster with majority Random walkers is considered an incorrect clustering.

t-Stochastic Neighbor Embedding

t-Stochastic Neighbor Embedding was performed as described [100] using the *Rtsne* implementation [101]. The perplexity parameter of t-SNE was chosen by computing the t-SNE projection at a series of perplexity values in a range from 5 to 70 and evolving for 5000 iterations, as suggested by the algorithm's authors [57]. Between perplexity values of 30 and 70, we find that the global structure of t-SNE projections is relatively consistent. From this range, we choose the value that provides the most consistent projections for the data set, as suggested in recent work on the proper use of t-SNE [102].

Independent Component Analysis

Independent component analysis was performed as described [103] using the *ica* R package [104].

1.6.16 Pseudotiming

Pseudotiming was performed using the Monocle package [18] on all myoblast samples and a subset of 1300 randomly selected MuSC cells (MuSC data was subsampled to reduce computational expense). Monocle's Discriminative Dimensionality Reduction with Trees algorithm was used for dimensionality reduction, and the top 20 features loaded into the first principal component were used in place of "ordering genes."

1.6.17 Statistical analysis

MANOVAs comparing all feature means between clusters were performed using the first 10 principle components as measured variables for each cell. t-tests were performed on unscaled feature values, assuming a two-sided null hypothesis and unequal sample variance. One-way ANOVAs comparing feature means between clusters were performed on unscaled feature values. The binomial test was performed on transition matrices to determine if transitions were non-symmetrical using the $H_0 : P(success) = 0.5$. All tests were performed with R standard library implementations.

1.6.18 Analysis of State Transition Dynamics

Subpath Analysis

To capture motility features at multiple time points for state transition analysis, cell paths are segmented into various subpaths of length τ and features extracted from each subpath. All state transition analysis presented here was performed for states of length $\tau = 20$. Heteromotility contains features to perform this segmentation and analysis of subpaths, accessible via a command line flag. There is a lower boundary of $\tau = 20$ for subpath size, as multiple features calculated by Heteromotility require at least 20 time points for analysis.

State Specific State Transition Quantification

State transitions within each state cluster, as presented in (Fig. 1.4C) are quantified as follows. Cluster identities are assigned using hierarchical clustering of the full-length motility track, as noted above. Full length tracks are then transformed into PCA space. Paths are segmented into $\tau = 20$ length subpaths, and Heteromotility features were extracted from each subpath. Subpaths are then transformed into the PCA space defined for full length tracks, such that the principal components for full length tracks are the same for subpaths. Transitions for each cell are quantified as the vector V between each pair sequential subpaths $S_0 \rightarrow S_1$ along the first two principal components (PCs).

The mean transition vector V_c for a cluster $c \in C$, where C is the set of all clusters is defined then as:

$$V_c = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} S_i(t+1) - S_i(t)}{N(T-1)}$$

where N is the number of cells in cluster c, T is the maximum time interval considered, and $S_i(t)$ is the state position in principal component space of cell i at time interval t.

The mean of all transition vectors for all cells with a given cluster identity is taken as the mean transition vector for a given cluster. These vectors are plotted atop a PCA projection of the full-length path state space, with an origin at the cluster centroid. Vectors in plots are presented with 10X magnitude to enhance visualization. Transition vector magnitudes are calculated as the magnitude of the mean vector, plus or minus the standard error of the mean (SEM) pooled across dimensions ($SE_{xy} = \sqrt{SE_x^2 + SE_y^2}$).

1.6.19 Coarse-grained Probability Flux Analysis (cgPFA)

coarse-grained probability flux analysis (cgPFA) as presented in (Fig. 5) is implemented per the definition of Battle *et. al.* [87]. Briefly, for each system, PCA is performed on the Heteromotility feature set extracted from subpaths under consideration and the first N principal components are used to define a motility state α. Each principal component is "coarse-grained" by binning into k subsections of equal length. Combinations of bins between each of the N coarse-grained principal components define a given state α. coarse-grained PFA analysis presented here was performed for a

set of dimensionality resolutions N, where $N \in \{1, 2, 3, 4\}$ and a set of coarse-grained resolutions k for each dimensionality resolution N, where $k \in \{2, 3, 5, 7, 10, 15, 20\}$. Explicitly, cgPFA was performed in 1D, 2D, 3D, and 4D, and detailed balance was evaluated with binning schemes using 2, 3, 5, 7, 10, 15, or 20 bins. Code for all our coarse-grained PFA implementations described is available in the Heteromotility GitHub repository `https://github.com/jacobkimmel/heteromotility`.

Two-dimensional cgPFA state space representations

To produce 2D state vector plots as presented in (Fig. 1.5), each cell's state, per the above definition, is considered for the each subpath t_i, and the subsequent subpath $t_i + 1$ for $i \in [1, T - 1]$. Each cell's state transition is recorded as a two-dimensional vector $v = (x_{PC1}, y_{PC2})$ representing the distance in $N = 2$ dimensions of PC space the cell traveled in that time interval. If the cell did not travel to a directly neighboring state (a unit vector displacement), the cell's path is interpolated as a series of unit vectors between the initial and final state. Each of these interpolated displacements is recorded in an intermediary bin, representing the assumption that cells traverse state space in a linear fashion.

For each state α, the transition rate from that state is calculated as the vector mean of all transition vectors originating in that state. Non-transitioning cells (zero magnitude transition vectors) are not considered in calculating this transition rate. These transition rates are plotted atop the binned state space as arrows, such that the magnitude of each arrow represents the magnitude of the transition rate constant, and arrow direction the direction of the transition rate constant in a given bin. The divergence of this vector field is calculated and presented as a heatmap on the state space. The divergence at a given state serves as a metric of state stability. Metastable states display negative divergence (more cells entering than exiting), while unstable states display positive divergence (more cells exiting than entering).

The mean vector displacement is calculated as the mean magnitude of displacement for all cells in a population. The magnitude of transition directionality is quantified as the magnitude of the vector sum of all transition rates in a given state system.

Detailed Balance assessment by N-dimensional cgPFA

To provide statistical assessment of detailed balance as presented in (Fig. 1.6), a state space is represented in N dimensions using the first N principal components, and each of these N dimensions is coarse-grained into k equally sized bins, as described above. A matrix M of dimensionality $2N$ is generated with k bins in each dimension is generated, representing all possible state transition combinations in state space. The first N dimensions of the space represent a cell's initial state in the space, and the next N dimensions represent the final or destination state across a time interval. To record a state transition, the value in the corresponding bin is iterated by 1.

For example, in a $N = 3$ dimensional space coarse-grained with $k = 3$, a 6 dimensional matrix is generated with $k = 3$ bins per dimension, for a total of $k^N = 3^6$ bins. A transition from a cell that begins at location $(1, 4, 3)$ and ends at location $(2, 4, 2)$ in state space would be recorded by iterating the value of $M(1, 4, 3, 2, 4, 2)$ by 1. This process is repeated for each cell in the population, and each transition for each cell.

A state space in perfect detailed balance would be expected to have all pairwise transition rates $A \to B = B \to A$. In the matrix M, this would manifest as equal values in all pairwise sets of bins $M(a, b, c, d, e, f) = M(d, e, f, a, b, c)$ (for a $N = 6$ dimensional case). We test detailed balance by checking the pairwise balance of all such state transition sets using the binomial test with $H_0 : p = 0.5$ where p is the probability of cells falling into either bin. If this null hypothesis H_0 is rejected, it indicates that a pairwise transition set is unbalanced, and therefore that detailed balance is broken in the system. All binomial test p values are corrected using the Holm-Bonferroni method to account for multiple hypothesis testing.

For each system presented, we test detailed balance breaking in this manner for $N \in \{1, 2, 3, 4\}$ and $k \in \{2, 3, 5, 7, 10, 15, 20\}$. As $N > 2$ dimensional spaces are challenging to present in their entirety, we display the top 5 most unbalanced (lowest binomial test p values) as a heatmap, with the initial state in the left column and the destination state in the right column. We note that these tests are biased toward Type II error, as linear binning in N dimensions may not accurately capture true states and binomial tests are underpowered when testing rare state transitions.

Dwell Time Quantification

Characteristic dwell times for each state were quantified in the N dimensional course-grained PCA spaces described above. For each dimensional resolution N and course-grained resolution k, each state S_i is considered. States S_i with fewer than 10 observed dwells are omitted to avoid inaccurate time constant approximation. The dwell time in discrete time units for each cell observed in a state S_i during the series was recorded. For instance, if a cell is observed in state S_i at time points $t = 1$ and $t = 2$, then leaves the state on $t = 3$, a dwell time of two time units is recorded. If a cell has multiple dwells in state S_i, separated by a dwell in another state, multiple dwell times of observed length are recorded. The result of this procedure is a table of observed dwells for each possible discrete time length $t \in [1, T]$ where T is the total length of the observation period in discrete time units. The dwell time for each state is characterized by two means: mean estimates and fitting of exponential decay functions to estimate time constants. Dwell times in Markovian systems are exponentially distributed. Simple arithmatic means are not an appropriate maximum likelihood estimate (MLE) for exponentially distributed data, and as such alternative MLE methods and exponential decay curve fitting are employed [105, 106]. We apply the latter approach and fit exponential decay functions to the number of cells that would be observed at each discrete time point (effectively, the cumulative sum of dwell times, such that all cells are observed at $t = 1$, cells dwelling $t = 2$ units or longer are observed at $t = 2$, and so on). Curves are of the form

$$y(t) = a \exp(\frac{-t}{\tau})$$

where τ is the time constant of decay and a is a parameter coefficient, and are fit using a differential evolution method, similar to that described in [107]. Distribution of discrete dwell time data was assessed graphically relative to a binned values from a theoretical exponential distribution with the shape parameter $\lambda = \bar{X}$, where $\bar{X}$ is the mean of the sample observation distribution X [108].

Hierarchical Clustering PFA

Probability flux analysis was also performed using hierarchical clustering (Fig. 1.18) to define cell state for a given time interval, in contrast to the use of coarse-grained principal components as above. To statistically assess if detailed balance is broken using hierarchical cluster-based stated definitions, we represent state transitions as an N-by-N matrix, where N is the number of hierarchical clusters in the partitioning scheme. Rows of the matrix represent a given cell's state at the an initial time interval t_0, while columns represent the cell's state at the subsequent time interval t_1. Values in each i, j-indexed cell of the matrix represent the number of cells meeting the described state criteria (state i at t_0, state j at t_1). As in statistical assessment of one dimensional coarse-grained PFA, a system in detailed balance would display symmetry about the diagonal, and symmetrical bins about the diagonal represent pairwise state flux magnitudes. Symmetry breaking is statistically tested using a binomial test for cells falling above or below the diagonal. The same Type II error caveat applies, as this approach only tests for raw numbers of cells above and below the diagonal, and does not test for proper pairwise transition symmetry. Therefore, statistical testing of detailed balance breaking by hclust-PFA is biased toward *false negative* conclusions, rather than *false positives*. This statistical testing approach is generalizable to any arbitrary state definition scheme. We also test the equivalency of pairwise transition rates for detailed balance by the binomial test, as described for N-dimensional coarse-grained PFA above. Code for our hierarchical clustering PFA implementation is available in the Heteromotility GitHub repository.

1.6.20 Supervised Machine Learning Classification

Machine learning models utilized standard scikit-learn [109] and R implementations. All data were scaled $[-1, 1]$ with unit variance prior to model training.

Simulated Data Classification

For simulated data, Random Forests were instantiated with $n = 100$ estimators and R `randomForest` default parameters. Random Forest feature importance was determined as the decrease in classification accuracy when a feature was removed [110]. Model accuracy was estimated by performing

five-fold stratified, shuffled cross-validation. Confusion matrices are the mean predictions of a five-fold stratified split.

MEF Transformation State Classification

Classification of MycRas MEFs vs WT MEFs was performed using a Support Vector Machine. Support vector machines were chosen as the classification model based on a test of multiple model types using a simple 5-fold cross-validation scheme, in which SVMs performed best.

Classification was carried out in a "Round Robin" fashion across independent biological experiments. For the Round Robin, we train a classifier on 4 out of 6 independent experiments, and classify the remaining 2. The accuracy of classifying these 2 "held out" experiments is taken as the classification accuracy for that split. The process of training on 4 experiments, and classifying the remaining 2 is repeated for all possible training/testing combinations. The average classification accuracy across all splits is taken as the Round Robin Classification Accuracy. This Round Robin analysis represents a form of strong cross-validation, accounting for differences that may arise between individual biological experiments.

SVM parameters were optimized by a grid search using the Round Robin Classification Accuracy as the objective. Grid Search identified a radial basis function kernel, feature selection to 65% of features with top ANOVA F-values, and a regularization constant $C = 0.5$ as the optimal configuration. All classifier code is available in the Heteromotility Github repository (http://github.com/jacobkimmel/heteromotility).

1.7 Heteromotility User Guide

For users interested in applying the Heteromotility tool to their own research problems, we have outlined recommendations for experimental design and parameter selection.

1.7.1 Timelapse Imaging Experimental Design

It is critical that timelapse imaging experiments be designed with sufficient temporal resolution to capture motility behaviors of interest. For instance, if a user is interested in motility processes

that occur on the order of minutes, imaging should be performed on a higher sub-minute temporal resolution. Alternatively, if a user is interested in processes that take place on an hours or days long timescale, temporal resolution can be decreased to the order of minutes or fractions of hours. In the latter scenario, the total length of imaging should be designed to capture the whole process of interest.

1.7.2 Sample Size Determination

Depending on the downstream analysis desired, the number of cells that need to be analyzed varies. For supervised classification problems, we recommend users perform a multivariate power analysis. For unsupervised clustering, rigorous analytical tools to determine the necessary sample sizes are lacking. Due to the Curse of Dimensionality [111] principle, the number of samples necessary to define clusters increases with the dimensionality of the problem. We find in three distinct biological systems that 30 principal components are sufficient to capture the overwhelming majority of variation, and therefore we suggest it is reasonable for users to assume a <= 30 feature dimensions will be present in their unsupervised clustering problem. As a simple rule of thumb, we recommend users have at least 5 fold as many samples as features, such that at least 150 cells should be analyzed as a lower bound for unsupervised clustering. Again, we reiterate that the sample sizes necessary to elucidate structure through unsupervised clustering cannot be readily predicted *a priori*, and therefore we provide this sample size suggestion only as a lower bound for experimental planning.

1.7.3 Supervised Classification

Users interested in supervised classification of motility phenotypes will need to select a data preprocessing method and a classification model to train. As a starting point, we recommend scaling raw Heteromotility feature outputs on the range [-1, 1] and reducing dimensionality using principal component analysis (PCA) prior to training a classification model. We recommend users start with an ensemble classification method such as Random Forest Classifiers, as they are robust in a variety of classification problems [112]. For validation of efficacy, we recommend users perform at least five-fold cross-validation (five independent models, each trained on 80% of the data and

tested on the remaining 20% in a rotating fashion). If multiple biological experiments are pooled for analysis, a round-robin analysis training on one set of experiments, and predicting on an independent set of experiments, is also useful to assess the robustness of the trained classification model. We recommend users perform a Grid Search to select hyperparameters for a given classifier. A practical guide to grid search implementation can be found in [113].

1.7.4 Unsupervised Clustering

For unsupervised clustering of motility feature data, we recommend users begin by scaling raw Heteromotility features [-1, 1]. Dimensionality should be reduced using PCA prior to clustering. We recommend users retain the smallest number of PCs sufficient to explain ˜90% of variation. Clustering may be performed using hierarchical clustering. We recommend using Ward's linkage, as this yields the most intuitive results in our systems and makes assumptions about cluster shape in line with biological priors. The number of clusters k to be defined may be optimized using cluster validity indices. We recommend the R package "NbClust" to compute several of these indices. While there is no definitive index that dominates others in all contexts, we recommend that users prioritize optimization of the Silhouette value, Hubert index second derivative, and D index second derivative, as these metrics are well accepted and provide a graphical interpretation [99].

1.7.5 Feature Dimensionality

For both Supervised and Unsupervised analysis of Heteromotility feature data, the dimensionality of the feature space is effects the efficacy of analysis. The "Curse of Dimensionality" [114] reduces the efficacy of supervised learning and unsupervised clustering when many low-information dimensions are present. To avoid this problem, we recommend users reduce the dimensionality of input spaces using PCA prior to analysis, as outlined above. Users may wish to further reduce dimensionality by eliminating features prior to PCA, or using fewer principal components if they believe a Curse of Dimensionality issue is present. If a user wishes to combine Heteromotility features with other phenotypic features, the user should be mindful of the Curse of Dimensionality and take steps to reduce the dimensionality of the feature space where possible. However, the removal of features

pre-maturely may hinder analysis that relies on a discarded feature for discrimination of two samples. Therefore, we recommend users attempt analysis with PCs covering a large proportion of the population variation first, before discarding additional features.

1.7.6 t-SNE non-Linear Dimensionality Reduction Visualization

For presentation of Heteromotility feature space, we recommend the use of t-Stochastic Neighbor Embedding (t-SNE) [57]. When utilizing t-SNE, it is critical that users understand the limitation on interpretation of t-SNE visualizations, as recently described in [102]. The critical hyperparameter in t-SNE analysis is perplexity. As outlined by the algorithms authors [57], appropriate values of perplexity cannot be determined *a priori*, and vary depending on the data set. In general, larger data sets with more data points will benefit from a higher setting of perplexity. We recommend users generate multiple maps for several settings of perplexity in the range [10, 70], selecting the a perplexity value that generates qualitatively reproducible representations.

1.7.7 Motility State Transition Quantification

In quantifying motility state transitions, the temporal window hyperparameter for state definition τ must be chosen by the user. This value represents the number of time steps considered in each temporal window where the motility state of a cell is defined. Several Heteromotility features lose numerical stability if applied to series shorter than $\tau = 20$ time steps. Therefore, a setting of $\tau = 20$ represents a lower bound for the size of temporal windows for analysis.

1.8 Additional Supplemental Videos

For additional supplemental videos, please visit our lab website at `http://cellgeometry.ucsf.edu/heteromotility`.

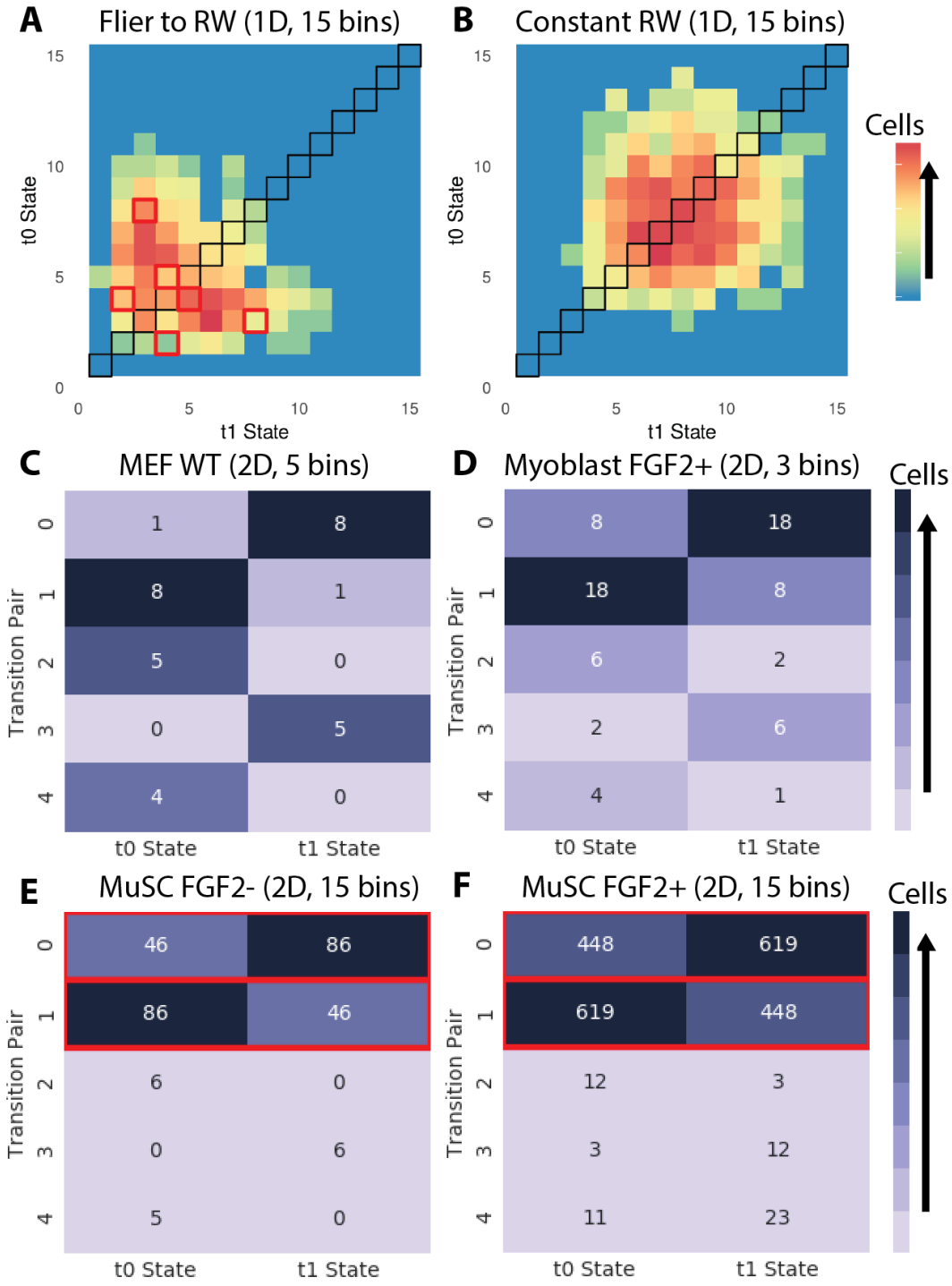
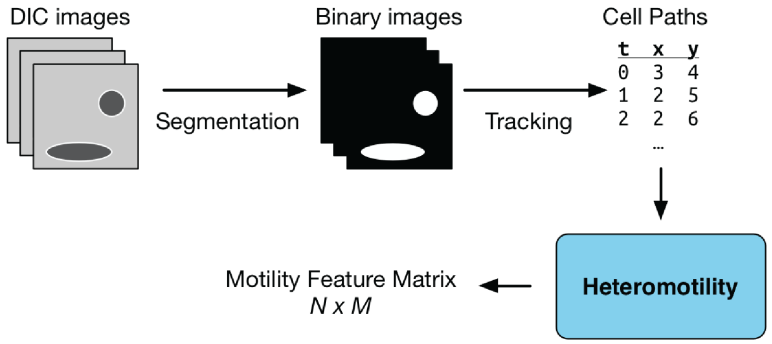


Fig 1.6. Analysis of state transition dynamics indicates MuSC motility states break detailed balance. One-dimensional coarse-grained PFA of (A) Simulated Levy flier transition to a random walk, (B) simulated random walk. Transition pairs from N-dimensional cgPFA displayed as heatmaps for (C) WT MEFs ($n = 312$), (D) Myoblasts (FGF2+) ($n = 150$), (E) MuSCs (FGF2-) ($n = 1838$), and (F) MuSCs (FGF2+) ($n = 2500$). Heatmaps show the five most unbalanced transitions in a system, with colors and numerical insets indicating the number of cells that transitioned from a given state at t_0 to a given state at t_1. Significantly unbalanced transitions are outlined in red ($p < 0.05$, Benjamini-Hochberg corrected binomial test). A system in detailed balance would display no unbalanced transitions.

A



B

Feature Name	Notation	Description
Total Distance	total_distance	total distance traveled
Net Distance	net_distance	net distance traveled
Minimum Speed	min_speed	minimum overall speed
Maximum Speed	max_speed	maximum speed
Average Speed	avg_speed	average overall speed
Time Spent Moving [1,10]	time_moving	proportion of time in motion, variable time intervals considered [1,10] frames
Average Moving Speed [1,10]	avg_moving_speed	average speed during movement, variable time intervals considered [1,10] frames
Linearity	linearity	Pearson's r^2 of regression through all cell positions
Spearman's ρ^2	spearmanrsq	Spearman's ρ^2 through all cell positions
Progressivity	progressivity	(net distance traveled / total distance traveled)
Mean Square Displacement Exponent	msd_alpha	exponent of log(τ) in a plot of log(τ) vs log(MSD)
Random Walk Linearity Comparison	rw_linearity	(cell path linearity - simulated random walk linearity)
Random Walk Net Distance Comparison	rw_net_distance	(cell path net distance - simulated random walk net distance)
Random Walk Kurtosis Comparison [1,10]	diff_kurtosis	(cell displacement kurtosis - random walk kurtosis) for variable time intervals
Hurst Exponent[1] (Rescaled-Range)	hurst_RS	Hurst exponent estimation using Mandelbrot's rescaled range methods
Autocorrelation [1,10]	autocorr	Autocorrelation of the displacement distribution for variable time lags
Non-Gaussian coefficient[2]	nongauss	Non-Gaussian coefficient (α_2) of the displacement distribution
Proportion of Right Turns [time_lag, estimation_interval]	p_turn	proportion of turns an object makes to the left, calculated for multiple time lags and direction estimation intervals
Minimum Turn Magnitude	min_theta	minimum turn angle in radians, for various time lags and direction estimation intervals
Maximum Turn Magnitude	max_theta	maximum turn angle in radians, for various time lags and direction estimation intervals
Average Turn Magnitude	mean_theta	mean turn angle in radians, for various time lags and direction estimation intervals

[1]. Mandelbrot, B. B. & Van Ness, J. W. Fractional Brownian motions, fractional noises and applications. SIAM review 10, 422,Äì437 (1968).
[2]. Rahman, A. 1964. Correlations in the Motion of Atoms in Liquid Argon. Phys. Rev. 136: A405,ÄìA411.

Fig 1.7. Heteromotility software description. (A) Heteromotility workflow diagram and (B) a table of the complete feature set.

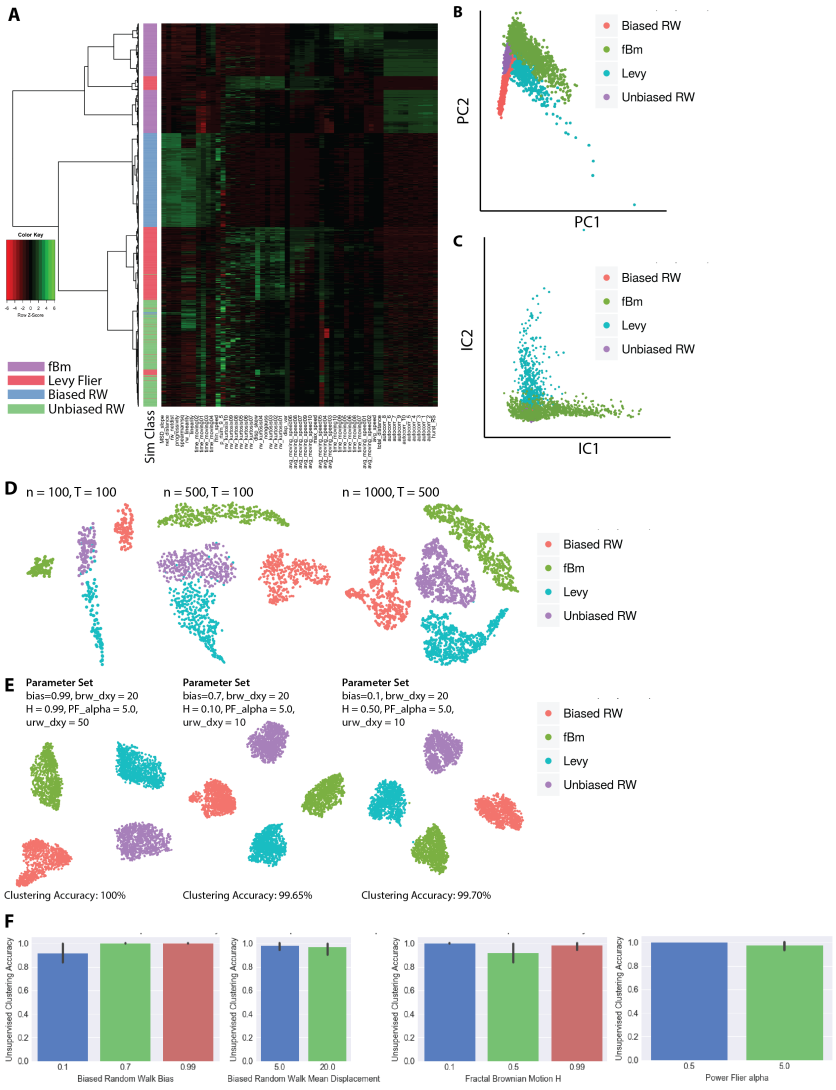


Fig 1.8. Clustering and visualization of simulated motion models with varied parameters. (A) Simulated models of motion are segregated by unsupervised hierarchical clustering, as displayed in representative a heatmap of hierarchically clustered (Ward's linkage) simulated motion paths (1000 members/class, length 100 time units) based on Heteromotility features. Color labels on the left mark a sample's True Class. Effective separation of the True Classes indicates effective detection of different phenotypes by unsupervised clustering. Two-dimensional (B) PCA and (C) ICA visualization of the simulated motion paths to provide an intuition for the linearity of motility state space and performance of traditional linear dimensionality reduction techniques. Only high (30) dimensional PCA spaces are used for analysis. ICA is not used for any downstream analysis. (D) Representative t-SNE visualizations of simulated motion models with different sample sizes and track lengths, labeled with ground truth classes. Models occupy distinct regions of state space under all sample size and track length variations. (E) Representative t-SNE visualizations of simulated motion model groups with the underlying parameters for each motion model varied. Parameters for each condition shown are displayed above the t-SNE map. (F) Unsupervised clustering accuracy (Ward's linkage) as a function of parameter variations to the underlying simulations. Performance decreases as expected when parameters are set in a manner that decreases the distinctness of the models. For example, performance is lower when the bias parameter for biased random walks is set to a low value, close to an unbiased random walk, or when the fractal Brownian motion index is set to the same index displayed by a random

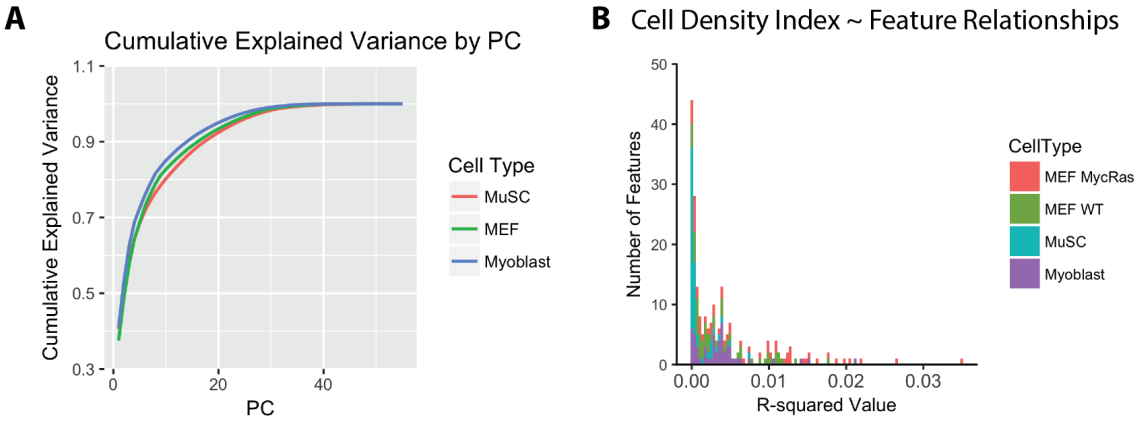
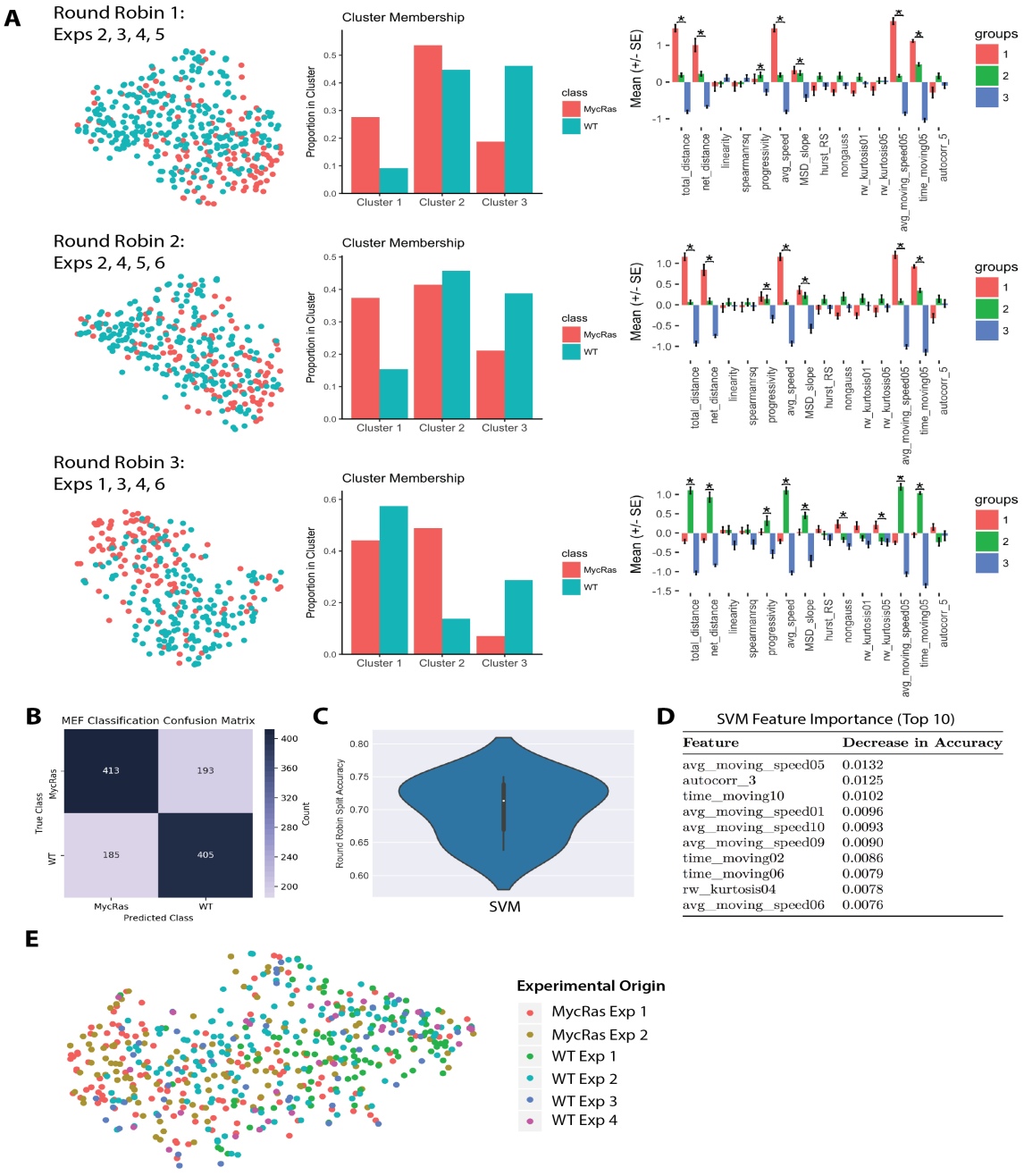


Fig 1.9. Comparison of variance dimensionality and local cell density relationships between cellular systems. (A) Cumulative variance explained for each dimensionality of principal component space across MuSC, MEF, and Myoblast systems. (B) Strength of relationships between our Local Cell Density Index and each of the Heteromotility features, displayed as overlapping histograms of Pearson's r^2 values for linear regression models fit between the Local Cell Density Index and each feature pairwise. No features in any system display a relationship with a meaningful effect size (max $r^2 \approx 0.03$).



Feature	Decrease in Accuracy
avg_moving_speed05	0.0132
autocorr_3	0.0125
time_moving10	0.0102
avg_moving_speed01	0.0096
avg_moving_speed10	0.0093
avg_moving_speed09	0.0090
time_moving02	0.0086
time_moving06	0.0079
rw_kurtosis04	0.0078
avg_moving_speed06	0.0076

Fig 1.10. Round Robin analysis of MEF motility state distribution and classification performance. (A) Representative t-SNE visualizations, MycRas and wild-type cluster distributions, and selected cluster feature values for three "splits" from a Round Robin analysis. In the Round Robin analysis, analyses were performed on 4 experiments (1 MycRas, 3 wild-type) leaving 2 experiments out in an iterative fashion for each possible combination of experiments. MEF cluster distributions and properties are reproducible across all combinations. (B) Aggregate confusion matrix for SVM classifiers trained in a Round Robin fashion, such that 4 experiments were used for training and 2 for evaluation for all combinations of 4 experiments. SVMs classify MycRas and wild-type cells with 70% accuracy and do not demonstrate a prediction bias for one class over the other. (C) Distribution of Round Robin Classification Accuracies for SVMs trained on each Round Robin split. (D) Top 10 most important features for Round Robin Classification Accuracy using SVM classifiers. Features importance is determined as the decrease in classification accuracy when an SVM is retrained without a given feature as input. The Top 10 features were selected based on the mean decrease in Round Robin Classification Accuracy across Round Robin splits. (E) t-SNE visualization of all MEF cells analyzed, labeled with their experimental origin.

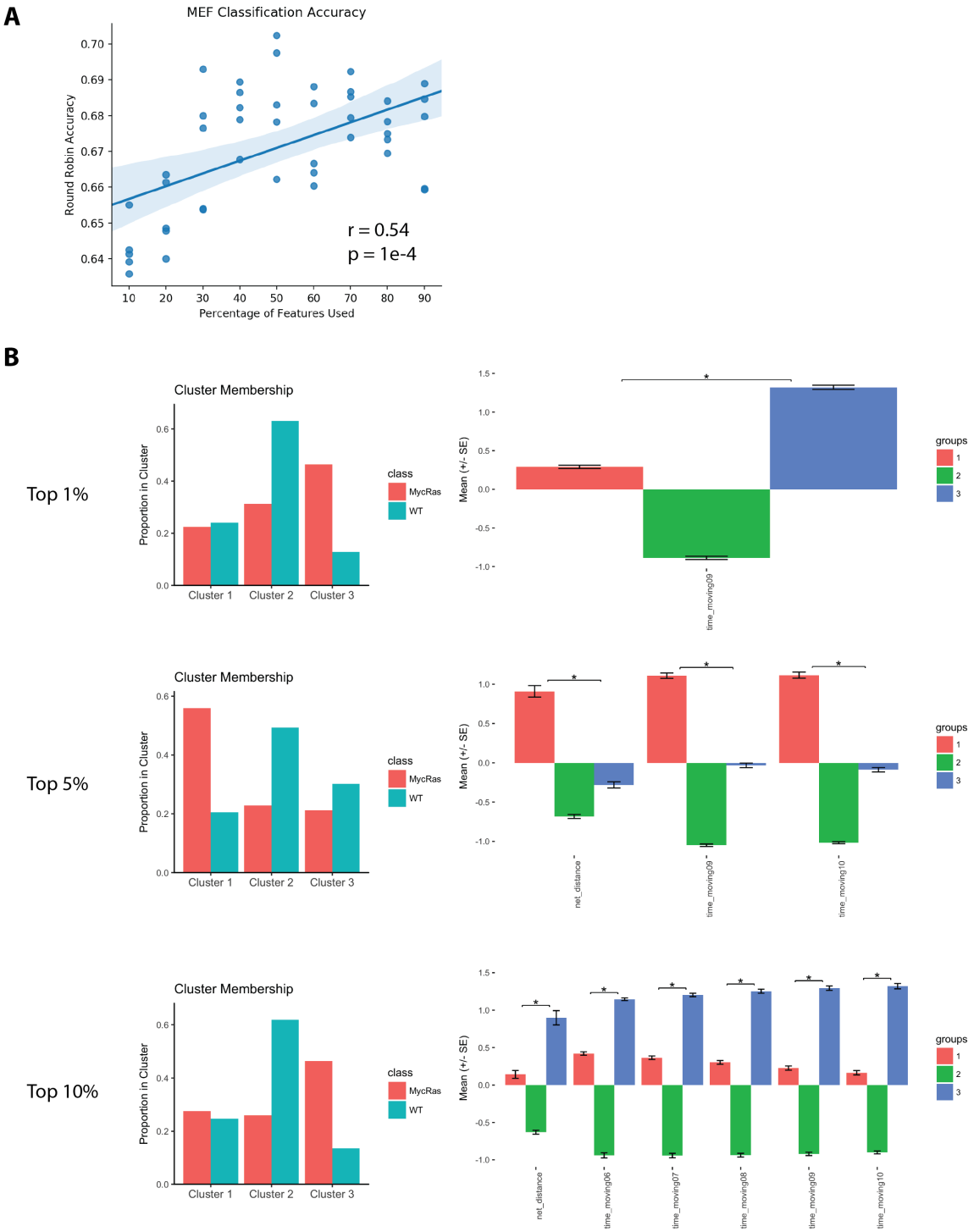


Fig 1.11. MEF analysis with reduced feature sets. (A) Round Robin classification accuracy is significantly, positively correlated to the proportion of features utilized. Points represent mean Round Robin Classification Accuracy for a given parameter set of N% of features and a value for the SVM bias parameter C. Five bias parameter values were tested in a linear distribution in the range $[0.4, 0.6]$ around the bias parameter $C \approx 0.5$ we found for the optimal SVM by Grid Search. Reduced feature sets were selected using only the top N% of features based on ANOVA F-scores. (B) Cluster distributions and cluster feature values for clustering of MEFs with reduced feature sets. Feature sets were reduced to the top N% based on ANOVA F-scores. Behavioral clusters are still identifiable and MycRas/wild-type dependent state distribution is preserved with a single feature.

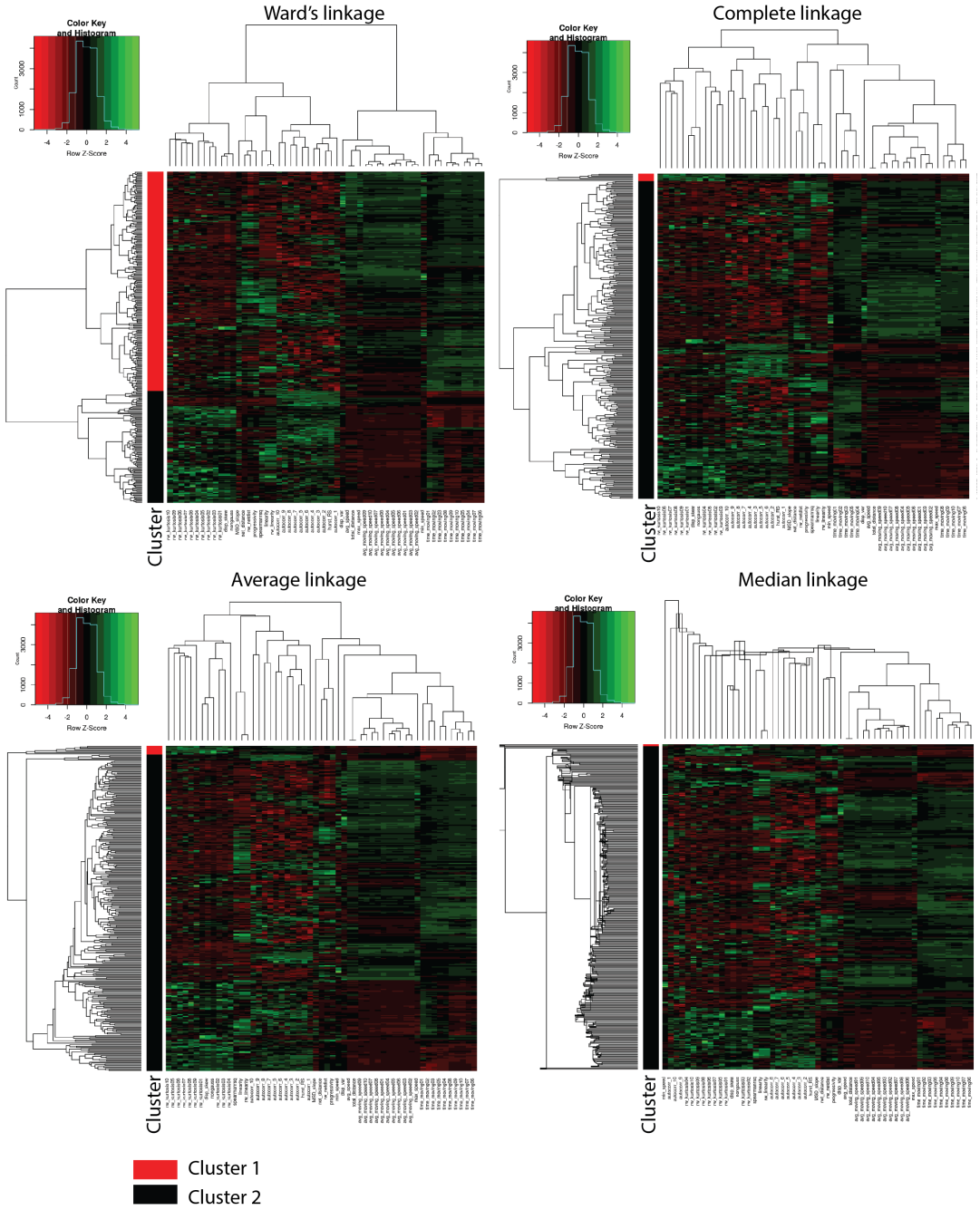


Fig 1.12. Myoblast clustergram visualizations. Clustergrams of myoblast feature space using several hierarchical clustering linkages. The assigned cluster label for each linkage map is displayed in a color coded column on the left hand side of each heatmap. Ward's linkage was used for downstream analysis.

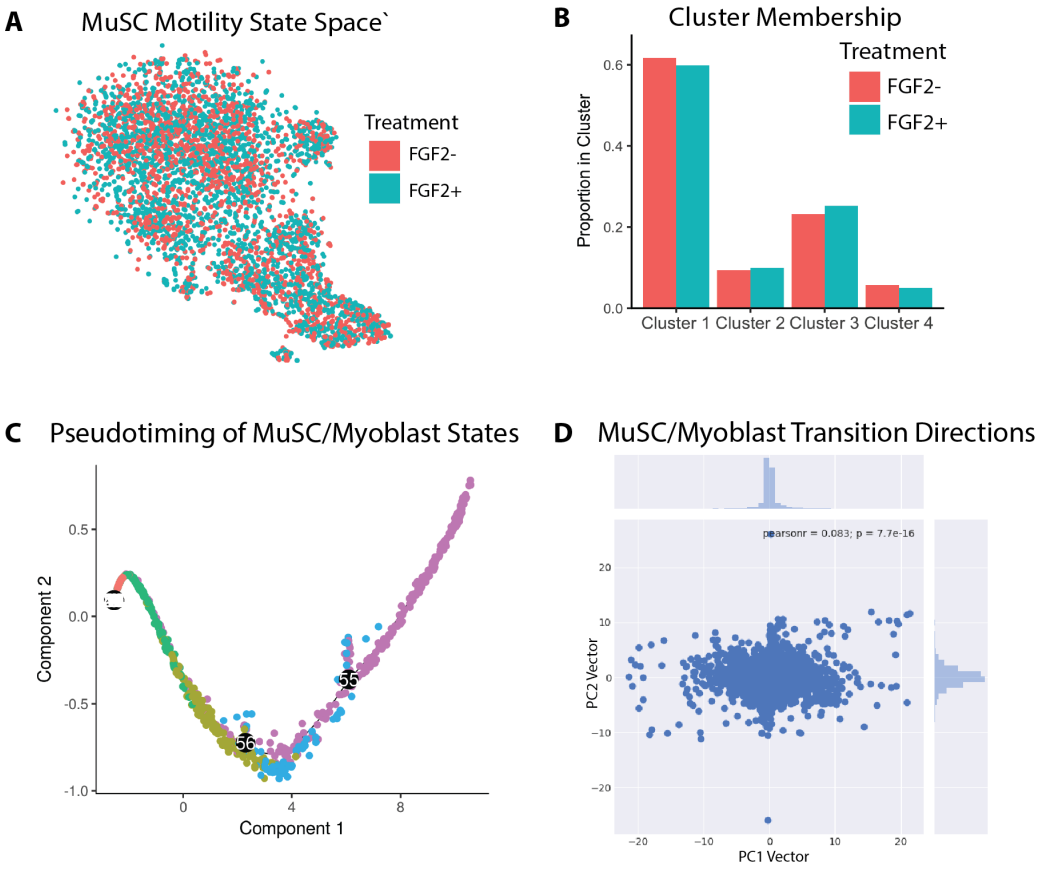


Fig 1.13. MuSCs display multiple motility states, reflecting states of activation. FGF2 does not influence MuSC motility phenotypes, and MuSC motility states reflect progressive states of activation. (A) t-SNE visualization of MuSC motility space with FGF2 treated and untreated color labels. (B) Occupancy of MuSC motility states in FGF2 treated and untreated conditions. (C) Pseudotime analysis displaying a reduced dimensional representation of MuSC and myoblast motility space (DDRT) with a minimum spanning tree plotted to mark the pseudotime axis. Pseudotime analysis attempts to find a temporal axis through an ergodic process observed at a single time point by fitting a minimum spanning tree (MST) to a reduced dimensional representation of multidimensional data. The longest axis of the MST is assumed to represent the temporal axis of the ergodic process. MuSC states are clearly ordered in a progressive sequence, moving toward the myoblast phenotype over pseudotime. (D) Scatterplot of MuSC/Myoblast transition vectors, demonstrating that transition vectors are primarily along the first principal component and slightly skewed in the positive direction.

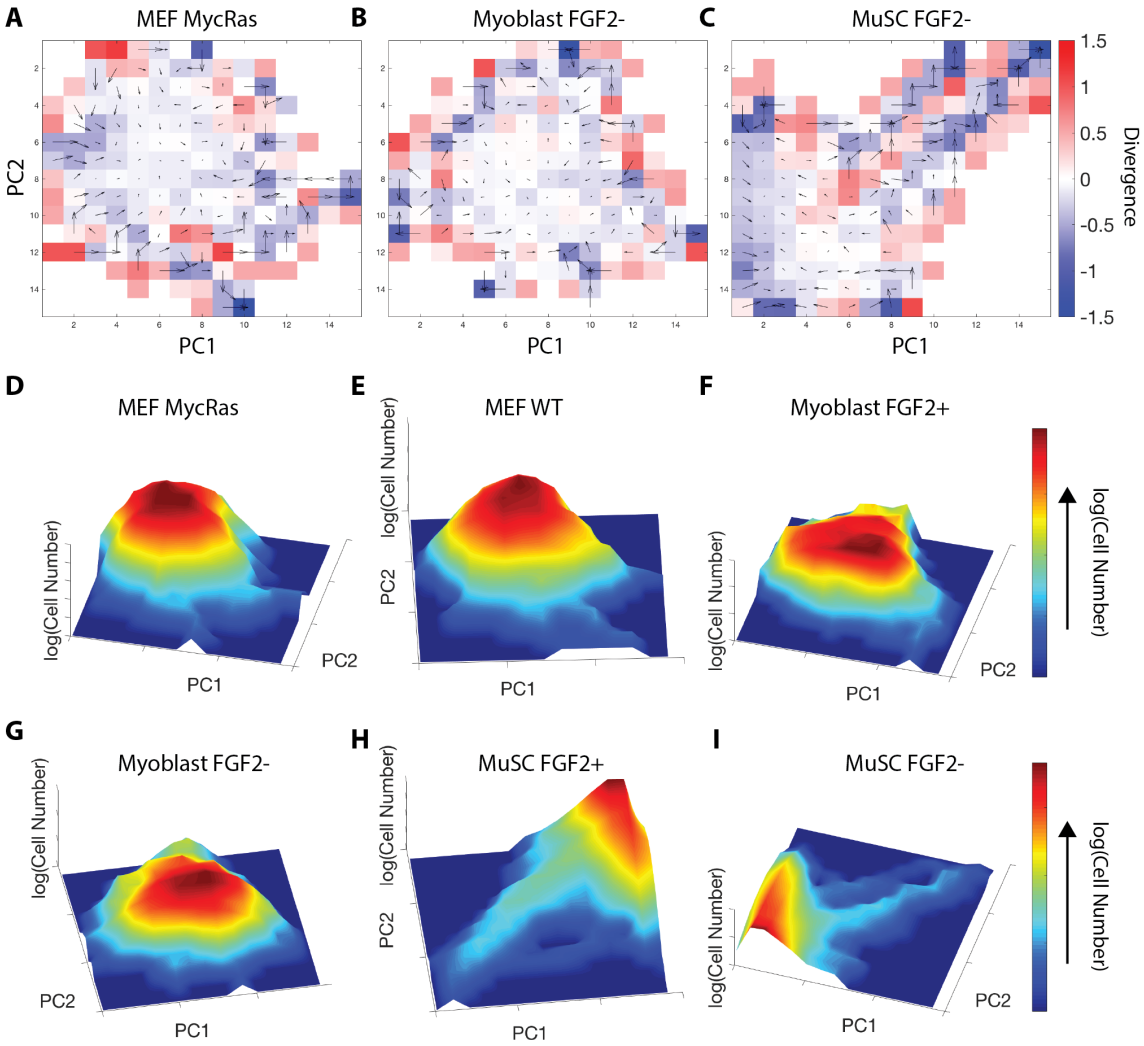


Fig 1.14. Course-grained probability flux analysis of motility state spaces. (A) Three-dimensional representation of the MycRas MEF state divergence surface as measured by cgPFA using $tau = 20$ and 15 course-grained bins. Course-grained probability flux analysis (cgPFA) of (B) myoblast (FGF2-), and (C) MuSC (FGF2+) motility states with subpaths of length $\tau = 20$ time points (130 minutes) and 15 course-grained bins per dimension. Each unique combination of bins between PC1 and PC2 is considered as a unique state. Arrows represent transition rate vectors, calculated for each state bin as the vector mean of transitions into the neighboring states in the von Neumann neighborhood. Arrow direction represents the direction of these transition rate vectors, and arrow length represents transition rate vector magnitude. Underlying colors represent the vector divergence from that state as a metric of state stability. Positive divergence indicates cells are more likely to leave a state, while negative divergence indicates cells are more likely to enter a state. (D-I) State occupancy visualizations of the same course-grained PCA presented for cgPFA analysis. The number of cells that occupy a given state for at least one time unit is represented in the third dimension of the landscape and by the heatmap colors.

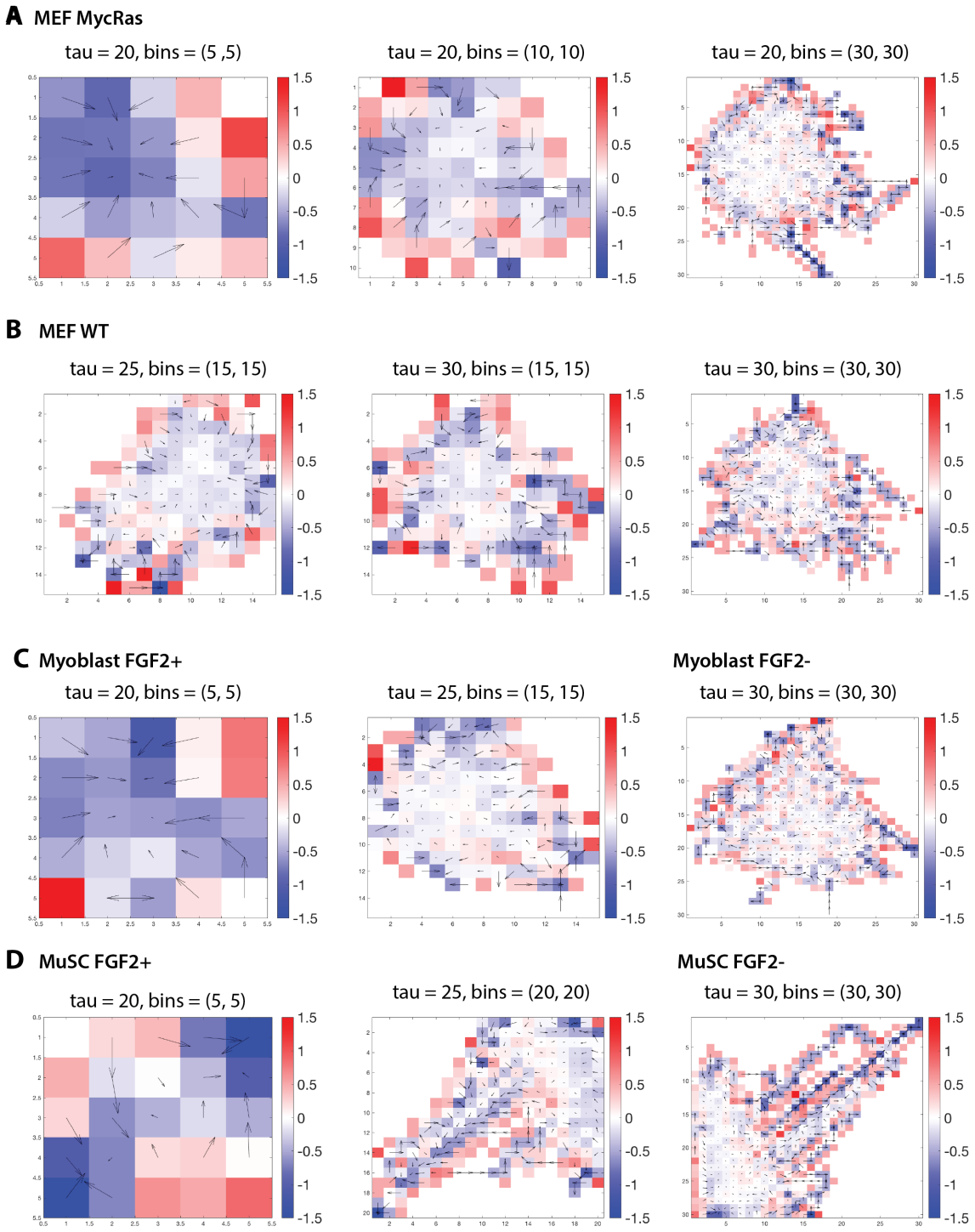


Fig 1.15. Course-grained probability flux analysis of motility state spaces on multiple time scales and binning resolutions. Course-grained PFA analysis as demonstrated in Fig. 1.5 and Fig. 1.15 was performed for all parameter combinations of the temporal window size $tau \in \{20, 25, 30\}$ and binning resolution $k \in \{5, 10, 15, 20, 30\}$ across all cellular systems. Representative visualizations across these parameter ranges are presented. Both (A) MycRas and (B) wild-type MEFs retain the qualitative metastable 'basin' appearance across time scales. As binning resolution decreases below $k = 10$, the structure of the state space is obscured. At higher resolutions of k, more bins little net divergence are present. (C) Myoblast cgPFA state spaces likewise retain a metastable 'basin' appearance across time scales. (D) MuSC state spaces retain a metastable 'valley' surrounded by unstable ridges across time scales. At higher binning resolutions of k, an unstable ridge within the metastable valley becomes more apparent.

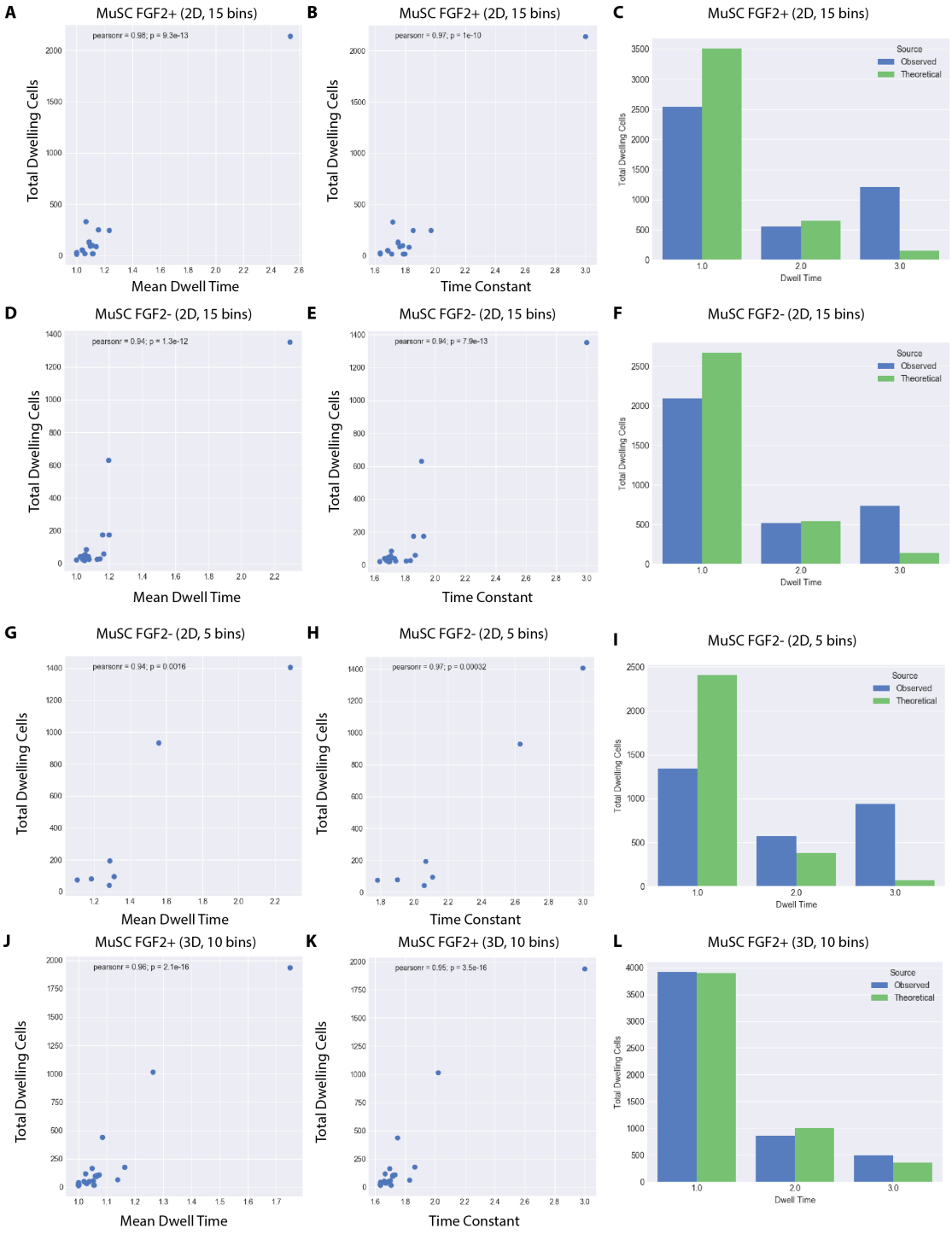


Fig 1.16. Dwell time analysis of MuSC motility states. MuSC motility state dwell time analysis reveals rapid transitions, longer dwell times in higher occupancy states, and roughly exponentially distributed dwell times. Dwell times vs. total number of observed cells for each occupied state in course-grained PCA space for (A, B) FGF2- MuSCs, (D,E) FGF2+ MuSCs in two-dimensional PCA space, and (G, H) FGF2- MuSCs and (J, K) FGF2+ MuSCs in PCA spaces where detailed balance is broken. Dwell time distributions relative to the binned samples from a fitted exponential distribution for (C) FGF2- MuSCs and (F) FGF2+ MuSCs in two-dimensional PCA space, and (I) FGF2- MuSCs and (L) FGF2+ MuSCs in PCA space where detailed balance is broken.

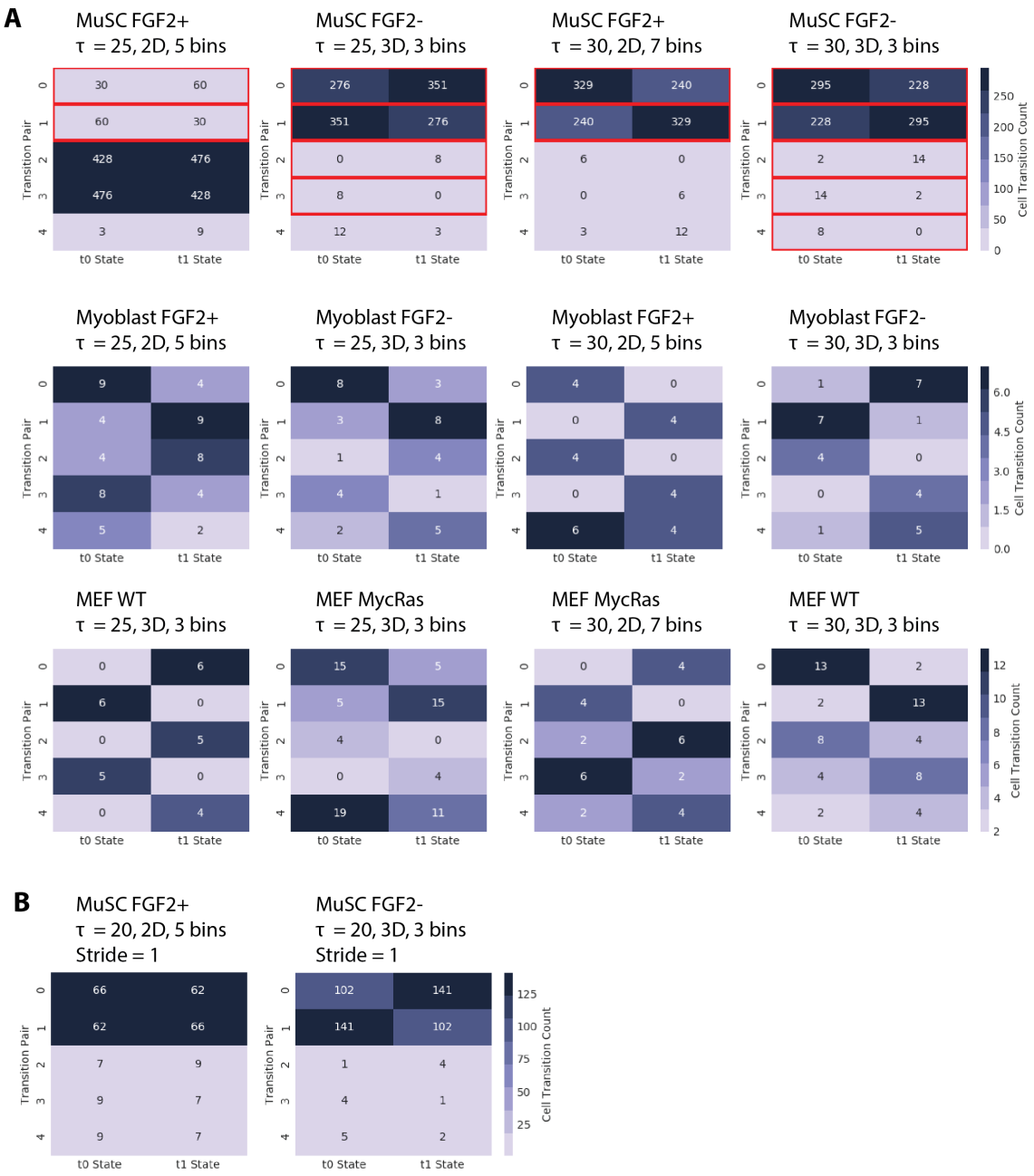


Fig 1.17. *N*-dimensional course-grained probability flux analysis across multiple time scales. ND-cgPFA as presented in Fig. 1.6 was repeated for values of the temporal window size parameter $\tau \in \{20, 25, 30\}$. (A) The results of detailed balance breaking are robust across settings of this time scale parameter. At each time scale, the MuSC system breaks detailed balance, while the MEF and myoblast systems do not. Heatmaps display the five most unbalanced transitions for each defined cgPFA space. *tau*, course-grained bin, and stride parameters are listed above each heat map. (B) To demonstrate that detailed balance is present in the MuSC system on short time scales, we performed ND-cgPFA using the same number of temporal windows size $tau = 20$, but overlapped them with a single unit stride of $s = 1$. In this scheme, each window is only 1 time unit different than it's neighbor, such that only 2 time units of difference are present between the initial and final time window. On this short time scale, MuSC systems do not break detailed balance.

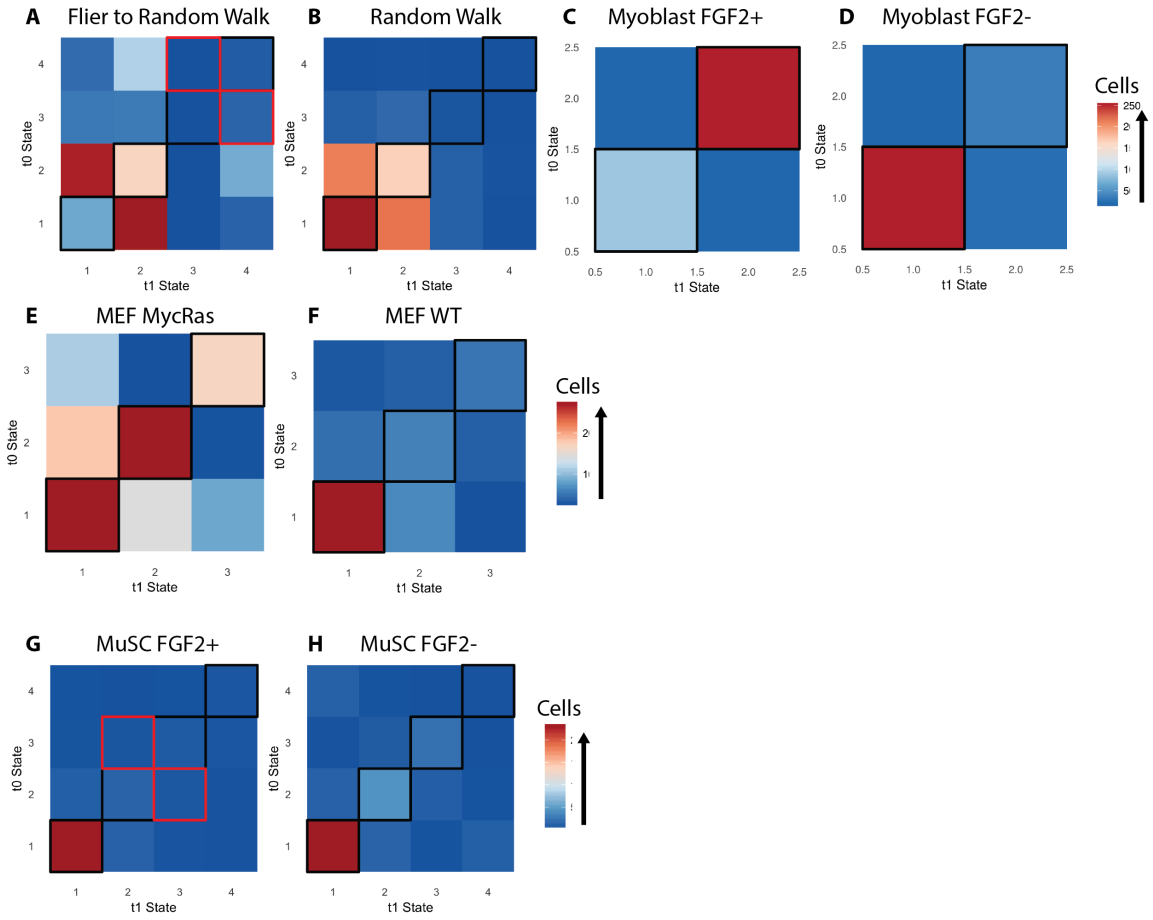


Fig 1.18. Probability flux analysis between states defined by hierarchical clustering. Hierarchical clustering based probability flux analysis of (A) a Levy flight simulation transitioning to a random walk, (B) an invariant random walk simulation, (C) myoblasts (FGF2+), (D) myoblasts (FGF2-), (E) MycRas MEFs, (F) WT MEFS, (G) MuSCs (FGF2+), and (H) MuSCs (FGF2-). The matrix displays transitions in state space as values in a matrix. Rows of the matrix correspond to an initial cell state (t0 state) and columns correspond to a destination state (t1 state). The value of each bin represents the number of times a state transition was observed. Symmetrical bins about the diagonal represent reciprocal pairwise transitions, with one 'forward' transition and one 'reverse' in each pair. The identity line represents "self" or non-transitions. MuSCs show a less balanced distribution than either MEFs or myoblasts by the binomial test for pairwise transition balance. Pairwise transitions breaking detailed balance are outlined in red.

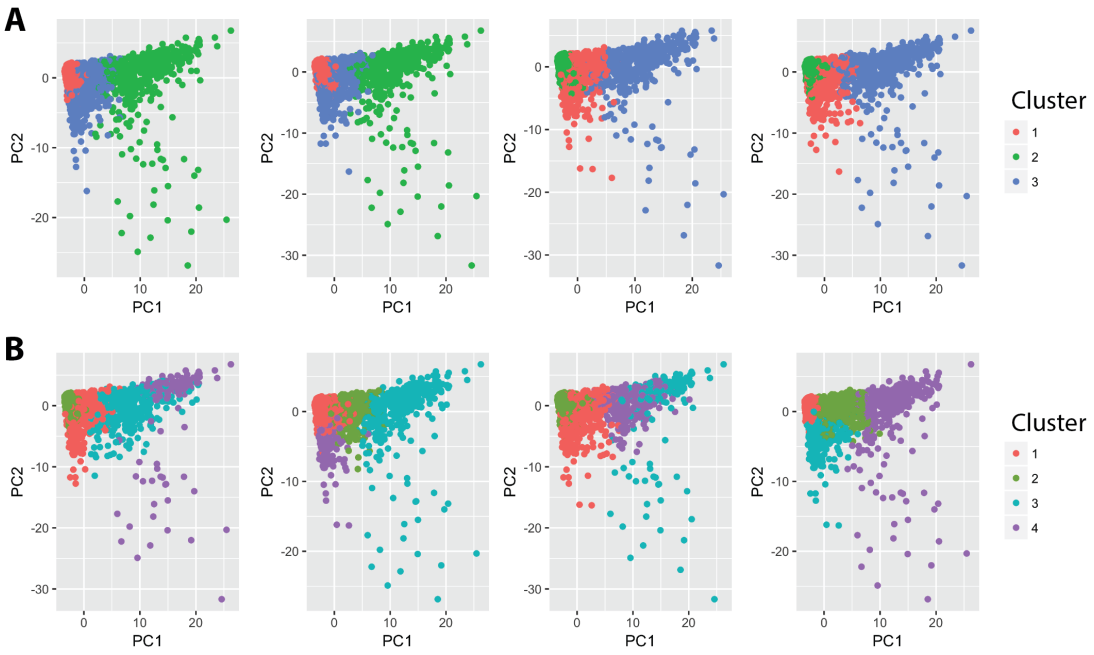


Fig 1.19. Resampling analysis of MuSC cluster partitions. (A) Representative random samples of 80% of MuSCs with a 3 cluster partition (Ward's linkage) applied. Note cluster separation along a common axis, robust to resampling. (B) Representative random samples of 80% of MuSCs with a 4 cluster partition (Ward's linkage) applied. Note separation of clusters along multiple axes in some samples, and a single axis in others. The 4 cluster partition is not robust to resampling.

Chapter 2

Deep convolutional and recurrent neural networks for cell motility discrimination and prediction

2.1 Introduction

Cell motility is an emergent property of living matter that spans the nanomolecular and macroscopic length scales, involving a complex regulatory network and dynamic reorganization of the cell's geometry [34, 115]. Cells can display a diverse set of motility behaviors, and these behaviors can provide a useful window for inference of a cell's functional state. Neoplastic transformation has long been appreciated to alter cell motility behaviors, increasing the migration rate of various models in culture and serving as a mechanism for metastasis [65, 116–119]. The motility behaviors of cancer cells in culture can even be predictive of broader tumor progression [69].

Likewise, the migration of progenitor cells is critical in early development and tissue regeneration [23]. Skeletal muscle stem cells (MuSCs) provide an accessible platform to study stem cell motility phenotypes *in vitro* by timelapse imaging. During embryonic development, MuSC precursors must migrate from early stage developmental structures (somites) to their adult location along the edge of muscle fibers in the trunk and limbs [78, 120]. In the adult, motility continues to play a critical

role, as MuSCs migrate along muscle fibers *in vivo* to sites of injury to initiate tissue repair [71, 121]. Motility behaviors are heterogeneous between MuSCs and change during stem cell activation [72, 122]. Heterogeneous fitness for regeneration within the MuSC pool is well appreciated [11], and analysis of heterogeneous motility behaviors may provide an additional lens through which to decompose different MuSC phenotypes.

Given the biological importance of motility phenotypes, classification of cells based on motility behaviors has useful applications in research and diagnostics. Similarly, exploration of heterogeneity within the motility behaviors of a cell population may provide biological insights. However, it is often difficult to determine which features of motility behavior will be predictive of a phenotype of interest, or allow for discrimination of heterogeneous behavior. Different phenotype classification tasks and cell populations may require distinct feature sets to extract valuable biological information. A method to algorithmically determine relevant features of cell motility for a given classification or discrimination task is therefore advantageous.

2.1.1 Related Work

To date, a number of tools have been proposed that rely upon a set of handcrafted features to quantify cell motility behaviors, providing some remarkable results [32, 41–43, 45]. Neural progenitor cells were discriminated by morphology and motility behavior alone [32], and genes that affect motility have been identified solely from timelapse imaging data [43]. We have recently demonstrated that rates of cell state transitions and the ordered or random nature of these transitions may also be inferred from motility alone [122]. These dramatic results demonstrate the potential insights that may be gathered from more extensive analysis of cell motility. However, these methods rely upon engineering of a hand-crafted feature set, and have thus far focused largely on features of speed and directional persistence. It is possible that more complex features may allow for improved discrimination of cell motility phenotypes, but it is difficult to predict what these features may be in each context.

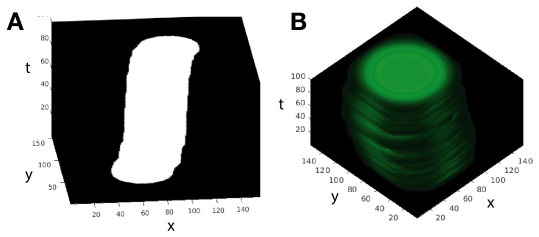
Convolutional neural networks provide an approach to learn relevant features from data, rather than handcrafting features based on a "best guess" of which features are relevant. In the field of computer vision, convolutional neural networks (CNNs) have recently made rapid advancements,

demonstrating state-of-the-art performance on a variety of image classification tasks [123–126]. CNNs utilize a set of parameterized kernels to extract spatial features, allowing distinct feature kernels to be learned for a given classification task [127]. In this way, CNNs are able to learn a "representation" of the problem's feature space. Feature space representations may also be learned in an unsupervised manner by training CNN autoencoder architectures to encode and decode [128,129]. This approach may be useful for learning relevant motility features where an explicit classification task is not present.

While CNNs are most commonly applied to tasks involving analysis in two-dimensional images at a single time-point, convolution is a natural analytical tool for any input information with spatial dimensions. CNNs have been successfully applied to a diverse set of non-imaging domains, including natural language processing [130], bird song segmentation [131], and EEG recordings [132]. Perhaps most clearly mirroring our challenge of motion classification, CNNs have performed well in the classification of video recordings [133–136]. These successful implementations have simply extended CNNs to consider three-dimensional images as inputs, where one axis is time. If the spatial nature of cell motility data is represented explicitly as a 3D image, in the same manner used for video classification, CNNs may allow for motility phenotype classification and unsupervised feature learning, without *a priori* definition of handcrafted features.

Deep neural networks have also been extensively applied to the analysis of sequential inputs, such as natural language sentences and biological polymer sequences [130, 137, 138]. While simple 1D CNNs that consider raw sequence inputs can be effective, the introduction of recurrent units such as long- short-term memory (LSTM) units to learn temporal relationships within the input sequence can improve performance and effectively learn long-term dependencies [139]. Cell motility data can be represented as a two-channel, 1D sequence, where each channel contains position values for an axis in physical space. In this representation, 1D CNN models with recurrent units may also allow for motility phenotype analysis without handcrafted features.

We investigated whether CNNs could be effectively applied to the problem of cell motility phenotype classification utilizing either of these two representation schemes. Cell motility is inherently 3D spatial data, where one dimension is time, such that either explicit representation of the temporal dimension in an image, or learning the temporal dimension relationships with a



[!t]

Fig 2.1. Representative motility traces using different markers of location. (A) Disk structuring element markers, (B) Gaussian distribution markers.

recurrent unit, may allow for effective analysis.

Here, we present *Lanternfish*, a tool to represent motility paths explicitly as 3D images or implicitly as multi-channel time series, classify different motility behaviors, learn motility features in an unsupervised fashion using deep neural networks, and predict future cell motility from past behavior. *Lanternfish* represents cell motility using a set of positional markers in a 3D volume, with the depth axis representing time, or as a simple multi-channel time series, where time series values are Cartesian coordinates. We demonstrate that standard CNN architectures are sufficient to accurately distinguish experimentally observed cell motility phenotypes represented using either method. Autoencoder architectures based on these models can be trained successfully on motility representations for use as unsupervised feature extractors. Additionally, we show that our RNN model can be adapted to predict cell motility in subsequent frames.

2.2 Methods

All implementations for work presented here are available on Github at https://github.com/jacobkimmel/lanternfish.

2.2.1 Explicit Spatial Representations of Motility Paths

Motion data in a two-dimensional plane is inherently three dimensional, with two dimensions in physical space (x and y) and a single time dimension t. Each of these dimensions has relevant spatial meaning, and spatial relationships are required to fully represent the motion of an object. This spatial nature makes motion an ideal candidate for the application of convolutional neural networks, which specialize in learning representations of spatial data.

One manner in which motility may be presented for analysis by CNNs is in the form of a static 3D image. Representing time as a spatial axis has allowed for successful time-series analysis by CNNs in multiple other problem domains [133–136]. Cell motility is typically recorded as the position of the cell centroid at each time point. To represent this time series of positions as a 3D image, we first produce a simple 3D representation of an (x, y) path by placing a 1 pixel (px) binary marker on the location of the object at each time point t in a single slice of a cube with dimensions (X, Y, T), leaving all other values at 0, where $x \in X$, $y \in Y$, and $t \in T$. Viewed one plane at a time along the t dimension, this cube is simply a video of the 2D path representing the object's location with a 1 px marker. However, this trivial representation presents a very sparse feature space, and intuitively may not allow for efficient learning of convolutional kernels.

In expectation of this sparsity problem, we produced tools to build representations that mark an object's location in each (X, Y, T) plane with a binary disk of arbitrary size or Gaussian distribution of arbitrary variance, instead of a single 1 px point. Gaussian distributions are scaled [0, 1] for each σ value. The resulting representation resembles a "stack of dinner plates" (Fig. 1). These representations contain information about the objects location at more (x, y) coordinates within a plane than the 1 px representations, so we reasoned that they may aide learning of 3D convolutional kernels.

Further information can be encoded by setting the amplitude of the disk or distribution in each t plane based on some real valued measurement. For instance, instantaneous speed or object size could be encoded as amplitudes. Compression of motion paths may be necessary due to GPU memory constraints. For all experiments performed here, motion paths were compressed 4- or 6-fold in the (x, y) dimensions by simple integer division of (X, Y) coordinates.

2.2.2 Multi-channel Time Series Representation of Motility Paths

Cell motility is also naturally represented as a two-channel time series, where the values of each channel are the Cartesian coordinates X and Y. This representation is obviously extensible to the 3D motion case with the simple addition of a third channel for the Z axis. CNNs may be applied with 1D convolutional filters to these multi-channel time series. Multiple problem domains have shown success in applying CNNs to multi-channel time series data in this manner [140, 141].

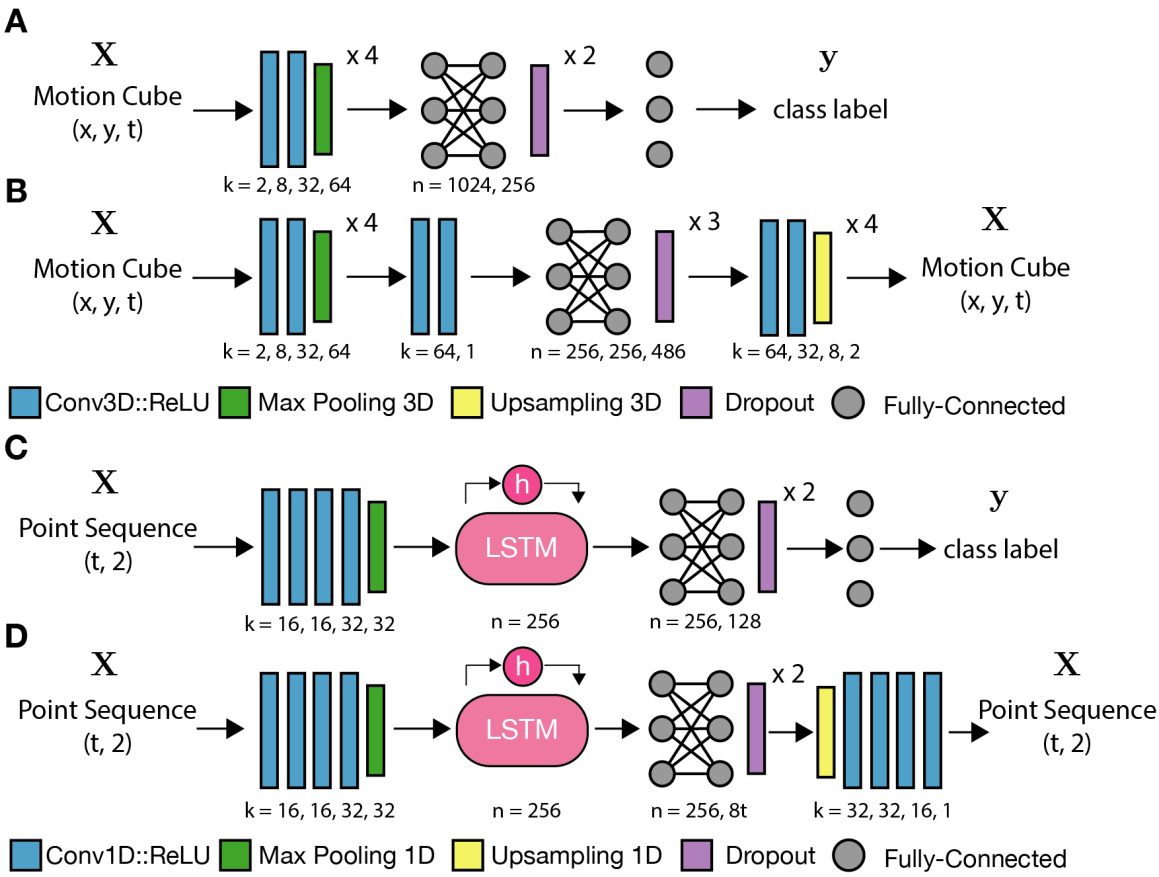


Fig 2.2. Cell motility classification and autoencoder architecture overview. (A) 3D CNN classification and (B) autoencoder architectures, where k is the number of parameterized kernels used by 3D convolutional layers in each block and n is the number of nodes in a fully-connected layer. (C) 1D recurrent neural network classification and (D) autoencoder architectures, where k is the number of parameterized kernels used by each 1D convolutional layer and n is the number of nodes in a fully-connected layer or LSTM unit. Convolutional layers and paired with a rectified linear unit activation. Pooling and upsampling layers operate with isotropic kernels of size 2 and stride of 2. Zero padding is performed as needed in autoencoder models to match input size.

This approach is inherently simpler than the explicit representation described above. However, 1D representation requires a model to implicitly learn kernels to preserve the relationship between X and Y dimensions in higher layers. We reasoned that there may be cases where this implicit representation is inferior to an explicit representation based on this principle.

2.2.3 3D Classification architecture

For classification of different types of motion represented as 3D images, we apply a standard CNN architecture utilizing 3D convolutional and max pooling layers, diagrammed in (Fig. 2.2A). 3D Convolutional layers convolve the 3D motion cube inputs with a set of parameterized kernels, passing the convolutional outputs to the layers above. The max pooling layers perform a max operation for voxels in an 3D-window, reducing the input size, and returns the resulting output to the layer above. This architecture is similar to well known 2D classification architectures [123, 142]. All convolutional layers are paired with a rectified linear unit (ReLU) activation ($\max(0, x)$) [143], utilize unit strides $s = 1$, and convolve with $(3, 3, 3)$ kernels. Convolutional layers pad input images by 1 px by reflecting edge values to avoid reduction of input size by convolution. Max pooling layers use kernels of size $(2, 2)$ and a corresponding stride of $s = 2$.

Fully connected layers are the same as in a traditional neural network, in which each perceptron unit considers input from all units in the previous layer, and outputs to all units in the next layer [144]. Dropout layers eliminate a random proportion p of fully connected units from a fully connected layer during each forward pass, reducing reliance upon individual units and preventing overfitting [145]. Two fully connected layers with dropout ($p = 0.3$, where p is the proportion of neurons dropped per epoch) and ReLU activations are utilized at the bottom of the network. Final class outputs are returned by a fully connected layer with a number of neurons equal to the number of classes and a *softmax* activation (Fig. 2.2A).

3D CNN classification networks were trained using stochastic gradient descent with momentum ($\mu = 0.5$). Categorical crossentropy was used as a loss function. We find that training is sensitive to the learning rate ϵ, and thus utilize a low initial learning rate $\epsilon_0 = 0.005$ with a rapid decay function

$$\epsilon_i = \epsilon_0 d^i$$

where $i \in [0, N]$ is the training epoch, ϵ_0 is the initial learning rate, and d is a decay coefficient, set to $d = 0.8$ for our experiments.

2.2.4 1D Recurrent Classification Architecture

Recurrent classification networks follow a similar standard architecture, utilizing 1D convolutional layers with size 3 kernels at the base followed by a max pooling layer with kernel size 2 and stride $s = 2$. The center of the network contains an LSTM layer with $n = 256$ units. Following the LSTM layer are two fully-connected layers with Dropout (as above) and a final *softmax* classification layer (Fig. 2.2C).

2.2.5 3D Autoencoder Architecture

The 3D CNN autoencoder architecture is similar to the classification network, employing stacked 3D convolutional and max pooling layers at the bottom of the network to encode the input, followed by fully-connected layers to reduce dimensionality and subsequent stacked 3D convolutions and upsampling layers to decode the input (Fig. 2.2B). As in the classification architecture, all convolutional layers are paired with a ReLU activation. 3D CNN autoencoder networks were trained with the *Adadelta* optimization algorithm [146], utilizing crossentropy or mean-squared error as the loss function for binary and Gaussian representations respectively. This architecture resembles others in the literature [127, 128].

2.2.6 1D Recurrent Autoencoder Architecture

As with the 3D CNN autoencoders, our RNN autoencoder architecture resembles the corresponding classification network. Following the fully-connected layers in the classification architecture, the RNN autoencoder appends a 1D upsampling layer and mirror 1D convolutional layers to return the input back to the original size (Fig. 2.2D). Mean-squared error (MSE) against the input sequence was utilized as a loss function for training. RNN autoencoders were trained with the *Adam* optimizer with a learning rate of $\epsilon = 0.001$.

2.2.7 1D Recurrent Motility Prediction Architecture

We adapt our RNN autoencoder architecture to a prediction architecture by removing the max pooling, fully-connected, and dropout layers. Sequences are convolved by four 1D convolutional

layers, as in the autoencoder, before being passed to a linearized LSTM and convolved by four more 1D convolutional layers. The final convolutional layer uses a linear activation function rather than a ReLU. Input sequences length τ_{in} are provided in the same multi-channel time series format as our other RNN architectures, and output sequences are multi-channel time series of length τ_{out}. The number of LSTM units is adjusted to $n = 2\tau_{out}$ depending on the length of desired output sequences. Mean squared error between the predicted path and the ground truth path was used as the loss function and the *Adam* optimizer was used with learning rate $\epsilon = 0.001$ (Fig. 2.6C).

2.2.8 Baseline Motility Classification

As a baseline motility classifier, a heuristic feature extractor is paired with a Random Forest (RF) classifier [59]. The feature extractor calculates four parameters of motion: (1) mean displacement, (2) displacement variance, (3) total distance traveled, and (4) net distance traveled. These four heuristics are commonly employed in the quantitative cell motility literature [26, 32, 47]. The RF classifier utilizes 10 estimators and scikit-learn default parameter settings. Code for the baseline classifier is available on Github.

2.2.9 Baseline Kinematic Motion Prediction

A linear kinematic model is used for baseline motility predictions. The kinematic model calculates the mean velocity

$$\vec{v} = \frac{1}{\tau} \sum_{i=1}^{\tau} \vec{dv}_i$$

across the last τ time steps in the preceding track and projects the object by $\vec{v}$ for each predicted time step. The temporal window τ is optimized by parameter search.

2.2.10 Cell Culture

Mouse embryonic fibroblasts, muscle stem cells, and myoblasts were cultured as previously described [122]. Neoplastic MEFs were generated as described and generously donated by the authors of [61].

2.2.11 Timelapse Cell Imaging

Timelapse cell imaging, cell segmentation, and cell tracking was performed as described [122]. Briefly, cells were imaged for 10 hours in DIC at 6.5 minute intervals using a stage-top incubator at 37°C and 5% CO_2. Images were segmented using common heuristic techniques and tracking was performed using a modified version of uTrack [147]. Cell tracking data is available on the "Heteromotility" Github repository https://github.com/cellgeometry/heteromotility.

2.3 Experimental Results

2.3.1 Motility Simulations

To determine if CNNs could discriminate between different types of motion under ideal conditions, we trained both 3D CNN and RNN classification networks on simulated data from 3 distinct, biologically relevant models of motion, namely random walking, Levy flights, and fractional Brownian motion. Random walking is motion with normally distributed random step sizes and directionality. Random walking is observed in freely diffusing biomolecules [97]. Levy flights similarly display random directionality, but step sizes are instead chosen from a long-tailed Levy distribution. Levy flights are observed in multiple biological systems and optimize path finding [49–51, 148]. Fractal Brownian motion models a random walk with long term dependence, similarly relevant as a representation of regulated motion in biology [96, 149]. By starting with simulated data we can optimize parameters using large sample sizes that would be difficult to obtain with living cells.

Random walks, Levy flights, and fractal Brownian motion were simulated for classification, each with a mean displacement of 5 (x, y) units per time step. Simulations were carried out for 100 time steps and restricted to a $(2048, 2048)$ pixel plane, representing the field-of-view that might be expected using a standard 4 megapixel microscopy camera.

For 3D CNN classification with explicit representations, 4-fold compression and cropping were performed to meet GPU memory constraints. Compressed tracks traveling more than 156 px from the origin in any direction were removed from analysis, to prevent dilution of the representation space by a few outlier tracks. Remaining tracks were represented in $(156, 156, 101)$ vx cubes. Nvidia

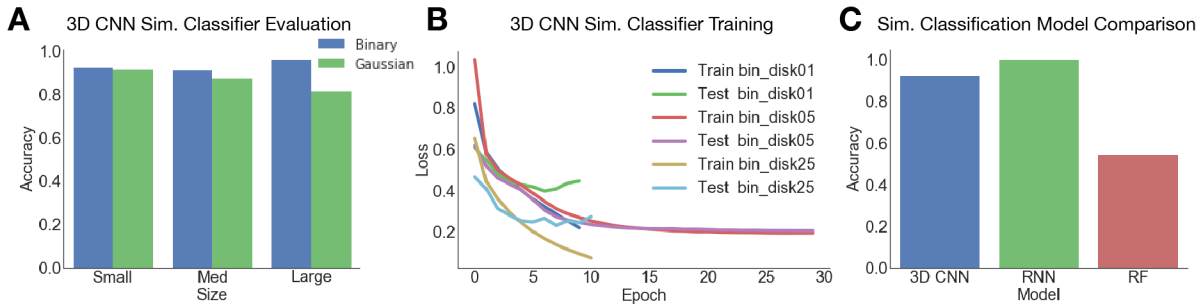


Fig 2.3. 3D CNN and RNN classifiers effectively distinguish simulated models of motion. (A) 3D CNN classifier performance using different marker sizes. (B) 3D CNN classifier training progress with different binary marker sizes. (C) RNN and 3D CNN (binary marker, 25 px) classifier performance.

GTX 1080 (Pascal) and Titan Xp GPUs were used for all experiments.

2.3.2 CNNs accurately classify simulated motility behaviors using both representations

Experiments to classify explicit 3D representations were performed using binary disks of three diameters $d \in \{1, 5, 25\}$ and broad Gaussian distributions of different variance $\sigma \in \{3, 10, 20\}$ as place markers. 12,000 samples per class were used for training, 1,500 for stopping criteria, and 1,500 for final validation. Early stopping was performed in all models after the testing loss failed to improve for 3 consecutive epochs [150]. Models were evaluated based on the prediction accuracy on the validation set.

The largest binary disk representations achieved ˜95% validation accuracy after 30 epochs, and the largest Gaussian representations of the same data yielded ˜81% validation after 30 epochs of training (Fig. 2.3A). Using binary distributions, accuracy increased as the marker sized increased. At all marker sizes, binary markers perform better than Gaussian markers. Both binary and Gaussian representations appear to overfit in later epochs, as evidenced by the divergence of the training performance from validation performance (Fig. 2.3B). A baseline random forest (RF) classifier model utilizing heuristic motility features (see Methods) reached ˜54% (mean of 5 random samplings). The 3D CNN approach therefore represents a 75% improvement over the heuristic RF baseline.

These results suggest that 3D CNNs are sufficient to distinguish different classes of motion

represented as 3D images, that multiple representation schemes can be effective, and that 3D CNNs can beat baseline heuristic classification methods by a wide margin. Large binary representation schemes appear to be the most effective representation scheme we tested. Therefore, we utilize large binary representations for all further 3D CNN experiments with live cell data.

Experiments to classify 1D multi-channel time series representations were performed using the same train, test, validation data split described above for 3D CNN experiments. Convolutional recurrent neural network (RNN) models train two orders of magnitude more rapidly than comparable 3D CNN models, and the reduced memory requirements allow for much larger batch sizes than in the 3D case. RNN classifiers achieve ˜99% validation accuracy, demonstrating superior performance to the 3D CNN models on simulated data. These results indicate that RNN models are sufficient to distinguish different models of motion represented as multi-channel time series, and that this classification scheme is superior to both a baseline heuristic approach and the 3D CNN approach for this task.

2.3.3 CNNs accurately discriminate cell types by motility behavior

After validating that CNNs were sufficient to distinguish simulated classes of motion, we applied the same classification networks to distinguish different types of experimentally measured cell motility. Cell motility was tracked in three different cell types by timelapse imaging for 10 hours, followed by segmentation and tracking by standard methods. Mouse embryonic fibroblasts (MEFs) are commonly used for *in vitro* cell culture assays, and neoplastic transformation of these cells has been demonstrated to alter their motility behaviors [122]. We tracked both wild-type and neoplastic (*c-Myc* overexpression, *HRas-V12*) MEFs to compare their motility behaviors. Muscle stem cells (MuSCs) are the obligate stem cell of the skeletal muscle, and their motility is known to be effected by their activation state [72]. Activated MuSCs commit to become myoblasts, a transit amplifying myogenic progenitor cell. We tracked both MuSCs and myoblasts to compare motility between these states of myogenic commitment (see Methods for culture details).

To determine if 3D CNNs could distinguish cell types based on experimentally measured motility, we trained a 3D CNN to discriminate between MEF and MuSC motility, represented using large binary disks (*diameter* = 25 px) in 3D space as described above. RNN models were trained

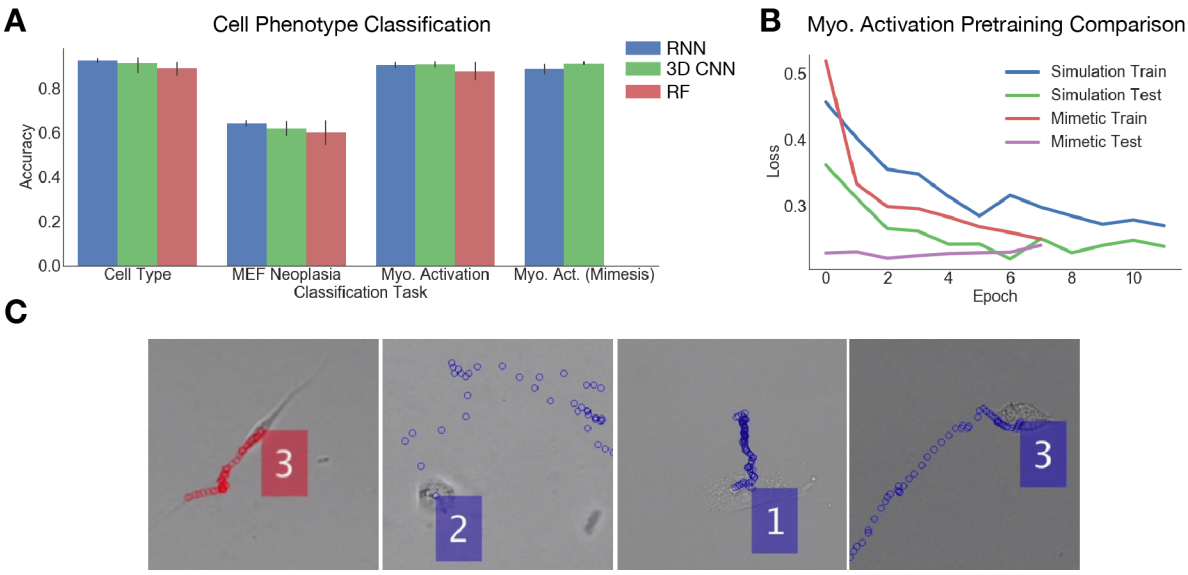


Fig 2.4. CNNs can discriminate between different cell types and cell states based on motility. (A) Validation prediction accuracy for 3D CNN and RNN models on cell type classification, MEF neoplastic state classification, and myogenic activation state classification with either simulated or "mimetic" pretraining. (B) Representative training progress for 3D CNN classifiers pretrained with either generic simulation classification weights or mimetic simulation classification weights. (C) Representative images of different cell types, from left to right: MuSC, myoblast, neoplastic MEF, and wild-type MEF. Colored markers indicate the cell's path along the substrate over time.

as on simulated models of motion above. Both networks were initialized with weights from the corresponding trained simulation classifier.

The 3D CNN classification network was trained for 30 epochs and RNN classifiers for 1000 epochs. Early stopping as above was performed with a 3 epoch and 100 epoch patience period for 3D CNNs and RNNs, respectively. Training was performed on MuSC motility traces and MEF motility traces ($n = 405$ per class). Testing and validation were each performed with $n = 50$ samples per class.

Each network type was trained five separate times to account for variability in stochastic optimization. Mean validation accuracy on this cell type classification task was ˜91.2% for 3D CNN models and ˜92.6% for RNN models (Fig. 2.4A). Mean validation accuracy for our heuristic RF baseline model was ˜88.8%. These results indicate that even with a small data set such as this, both 3D CNNs and RNNs can be effectively trained to discriminate different types of cell motility (Fig. 2.4A).

2.3.4 CNNs provide discriminative power between stem cell activation states

To determine if CNNs can distinguish between more nuanced differences in cell state, both 3D CNN and RNN classifiers were trained to discriminate between myogenic activation states. Training was performed on $n = 200$ MuSCs and myoblasts per class, with testing on $n = 50$ samples and validation on $n = 27$ samples. Dynamic data augmentation was utilized in the 3D CNNs due to the small available sample size. Motion cubes were horizontally and vertically flipped to increase training set diversity without perturbing the representation of motility. No augmentation was used for RNN models.

As above, five independent training experiments were performed for each network. Transfer learning was employed, taking advantage of weights learned from simulated data classification. Mean validation accuracy reached ˜90.7% for 3D CNN classifiers, ˜90.4% for RNN classifiers, and ˜87.7% for our heuristic baseline (Fig. 2.4A). These results demonstrate that both 3D CNN and RNN models can discriminate between stem cell activation states based on motility alone, even with small data sets.

Classifiers were also trained in the same manner to discriminate between wild-type and neoplastic MEFs with transfer learning from the simulated motion classifier. Training was performed on $n = 160$ samples per class, with testing and validation on $n = 30$ samples per class. Classification failed to achieve validation accuracy >64% for either 3D CNN or RNN models (Fig. 2.4A). The baseline heuristic model performed at 60% validation accuracy. The more nuanced phenotypic difference between wild-type and neoplastic MEFs may be an inherently more challenging classification problem. The small available sample size likely compounds this difficulty and exacerbates the classifiers' poor performance.

2.3.5 Cell mimetic pretraining

Given the success of pre-training by classification of simulated models of motion, we next attempted to generate simulated data that more accurately reflected real cell motility to enhance pre-training efficacy. For a set of real cell motility data, we measure the displacements and turning behavior of each cell. Displacements are measured simply as the Euclidean distance between each set of

sequential timepoints. The turning direction at a point t_i is determined as the angle between the vectors that connect points t_{i-1} to t_i and t_i to t_{i+1}.

Cells are decomposed into a set of k clusters by k-means clustering on a set of parameters measured from these displacement and turn angle distributions. The number of clusters $k = 5$ was chosen empirically to capture the diversity of the cell phenotypes while still leaving non-trivial numbers of cells in each cluster. For each cluster, a bounded Johnson distribution is fit to the aggregate distribution of displacements and the aggregate distribution of turn angles. Simulated samples are generated by randomly sampling displacement magnitudes and turn angles from the fitted Johnson distributions for T time steps. To represent a population of cells, the proportion of simulations generated from each cluster is equivalent to the cluster's prevalence in the original cell data. This approach may be conceptually likened to the bag-of-words model [151], in which k-means clustering is used to decompose features into a representative "vocabulary." By sampling from each of k clusters proportionally, we aim to capture and simulate heterogeneous phenotypes within a cell population, rather than simply reproducing a single averaged phenotype that may not be representative of any true cell phenotype.

We generated "cell mimetic" simulations for MuSCs and myoblasts by the above method, with $n = 15,000$ simulated samples for each of the two activation states. 3D CNN and RNN classifiers were pretrained by classifying between the two simulated data sets, reaching ˜97% validation accuracy for the 3D CNN classifier and ˜99% for the RNN classifier. The weights from this pretrained network were used to initialize classifiers for the myogenic activation task outlined above.

Training speed appeared to increase for 3D CNNs and remain unchanged for RNN classifiers (Fig. 2.4B). Mean validation accuracies were effectively unchanged at ˜91.1% for 3D CNN classifiers and ˜88.8% for RNN classifiers. These results indicate that "mimetic" pretraining may aide training speed for 3D CNN motility classification networks, and has little effect on final classification accuracy.

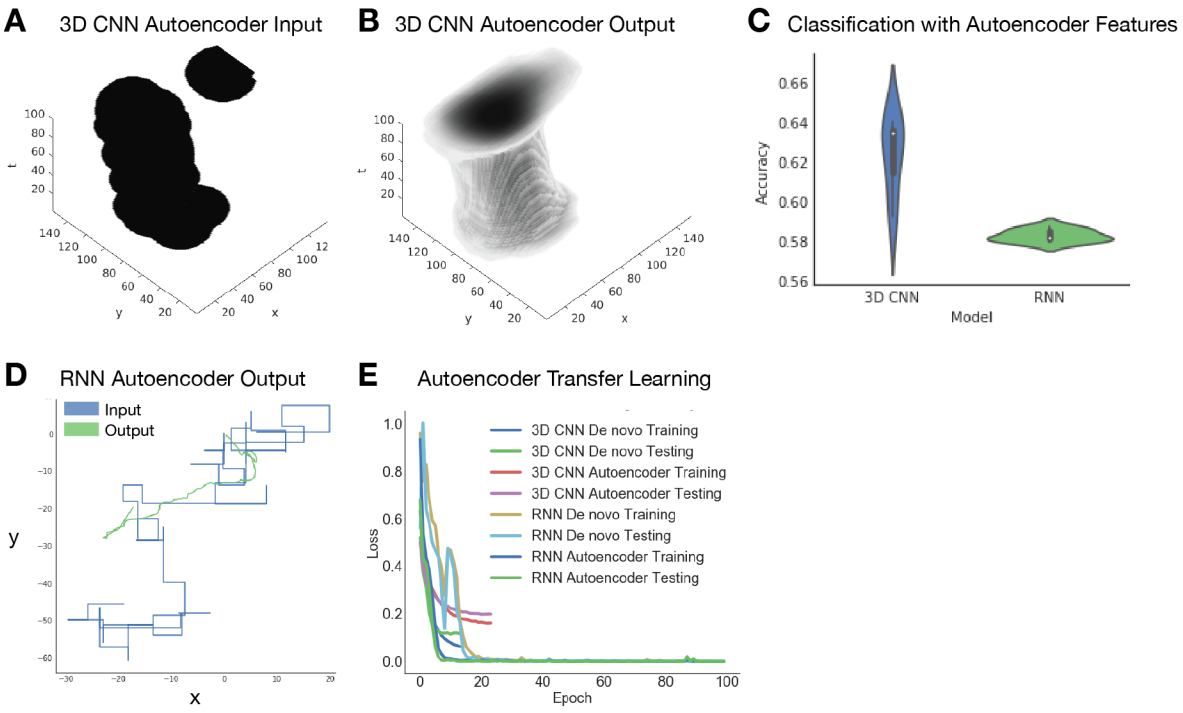


Fig 2.5. CNN autoencoders can learn representations of motion feature space in an unsupervised manner. (A) Sample 3D CNN autoencoder input and (B) output. (C) Classification accuracies of Random Forest classifiers trained on 3D CNN or RNN autoencoder features. (D) RNN autoencoder sample input and output. (E) Training comparison of de novo trained simulated motility classifiers to classifiers with autoencoder transfer learning.

2.3.6 Autoencoders allow unsupervised learning of representations in motion feature space

Results up to this point indicate that supervised classification of different cell motility phenotypes using both 3D CNN and RNN models is effective. However, in the analysis of motility data, supervised classification data is not always available. For instance, to explore the heterogeneity of types in a given population, there is no obvious method to generate supervised classification data that may be used to learn relevant feature kernels by optimization of a standard classification loss function. This would also be an issue in the identification of heterogeneous motility behaviors in patient biopsy samples, in which the distinguishing features are not known *a priori*. Training CNNs as autoencoders in an unsupervised fashion has been used in other contexts to learn relevant feature kernels where no obvious classification problem is present [128, 129]. We next attempted to train autoencoders on our 3D representations of cell motility and multi-channel time series to learn

relevant feature kernels in the absence of a supervised classification problem.

A 3D CNN autoencoder architecture was formulated using stacked convolutions, followed by fully-connected layers, upsampling, and stacked convolutional layers (Fig. 2.2B). Similarly, an RNN autoencoder was formulated by appending upsampling and convolutional layers following the fully-connected layers in our classification architecture (Fig. 2.2D).

Both autoencoders were trained on $n = 12,000$ samples of each class for three types of simulated motion (random walk, Levy flight, fractal Brownian motion), with training and validation each on $n = 1,500$ samples per class. Large binary disk representations were used for 3D CNN autoencoder experiments. Binary crossentropy was used as a loss function with 3D CNN models, while mean squared error was used to as a loss function for RNN models. All autoencoders successfully reduced loss over several training epochs. When visually inspected, 3D CNN autoencoder outputs appear to accurately reflect input motility representations (Fig. 2.5A, B). However, RNN autoencoder outputs consistently fail to capture the full extent of a cell's motility, though some degree of path shape is preserved (Fig. 2.5D).

To determine if 3D CNN or RNN autoencoders trained on motility representations could be employed as feature extractors, we utilized the output of the autoencoders' central layer (the encoded representation) as features. To quantify the amount of class information preserved by the encoded representations, we trained a Random Forest classifier to distinguish the simulation classes using either 3D CNN autoencoder features or RNN autoencoder features. Random Forests trained on 3D CNN autoencoder features achieved ˜62.3% and RNN autoencoder features achieved ˜58.2% accuracy on this 3 class problem. Both Random Forests trained on autoencoder features are notably more effective than our heuristic RF baseline, which achieved ˜54.4% accuracy on the same task. These results indicate that both 3D CNN and RNN autoencoders are able to learn meaningful motility features in an unsupervised manner. In this context, 3D CNN autoencoder features appear to be marginally more predictive than RNN autoencoder counterparts.

Autoencoders are often used for unsupervised pre-training prior to a classification problem. To determine if autoencoder features could effectively aide motility classification by transfer learning, we initialized and 3D CNN and RNN models with autoencoder weights and trained these classifiers to distinguish the simulated motion models. Transfer learning with autoencoder features appears

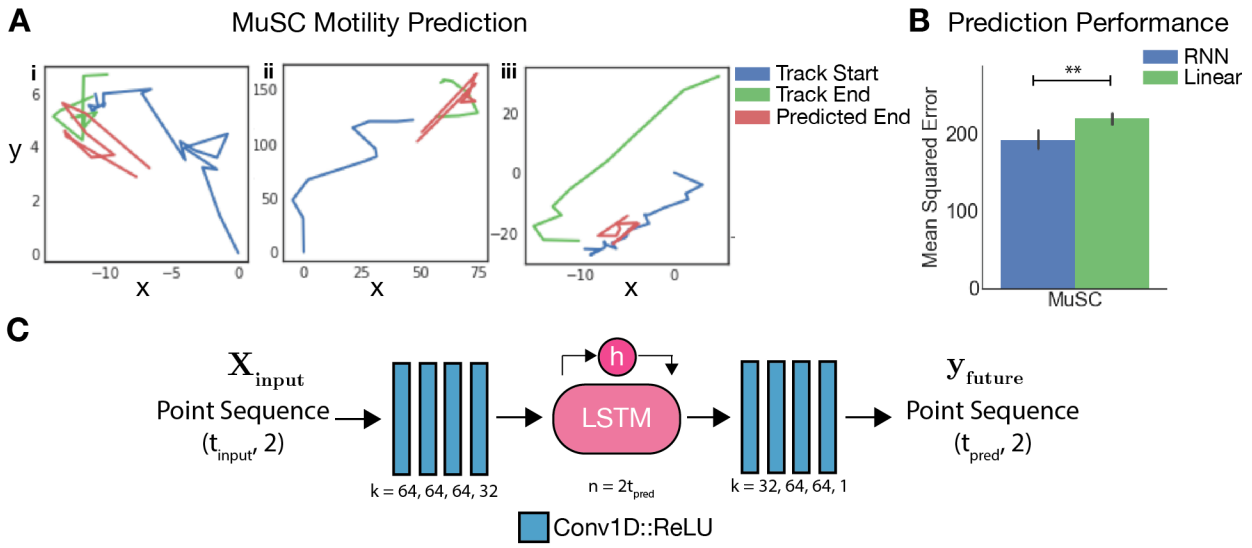


Fig 2.6. RNN models predict MuSC motility more effectively than linear kinematics. (A) Representative samples of MuSC tracks used for prediction with predicted track endings and true track endings. (B) Performance of the RNN motion prediction model relative to a linear kinematic baseline model, determined as the mean squared error between ground truth and predicted track endings. (**t-test $p < 0.001$) (C) RNN motility prediction architecture, where k is the number of kernels in each 1D convolutional layer and n is the number of units in the LSTM. All convolutional layers except the final layer are paired with a ReLU activation.

to increase RNN training speed, but decrease 3D CNN training speed (Fig. 2.5E). This result indicates that autoencoding motility data may be effective pretraining for our RNN classification architectures, but not their 3D CNN counterparts.

2.3.7 RNNs predict muscle stem cell motility

Tracking individual cells in timelapse microscopy experiments is a difficult multi-object tracking problem [152]. Popular tracking methods utilize a motion model to predict cell motility in advance of the next frame to improve tracking performance [153]. This motion prediction is especially useful in the event of "missed detections," where a cell is not detected or segmented for a given set of frames but is detected later on. The most common motion models employed are based on linear kinematics, with Kalman filters serving as a popular choice [147]. However, cell motion does not adhere to kinematic assumptions in all cell types, with myogenic cells being an excellent example of such a system. A motion model specifically tailored to the cell type of interest may therefore be useful to improve tracking performance, but such specific tailoring would require a prior knowledge

of the very motion features that the live cell experiment is designed to analyze. Some way to tailor prediction models on the fly could help solve this problem.

Recurrent neural networks have been effectively utilized for sequence prediction in multiple fields [154–157]. We adapted the convolutional RNN autoencoder model to a sequence prediction model by removing the pooling layers and fully-connected layers and altering the number of nodes in the central LSTM layer (Fig. 2.6C). As a prediction task, we trained the RNN prediction model on $\tau_{in} = 20$ time steps of motion and predicted $\tau_{out} = 10$ time steps into the future. As a data set, we split MuSC motion paths into subpaths of length $\tau_{total} = 30$ for a total of $n = 8,676$ paths. A validation set of 10% of all tracks was held out, and the remaining tracks were split with 80% used for training and 20% used for testing.

As a baseline for comparison, we performed a simple kinematic prediction of MuSC paths that assumes persistence of the velocity from preceding time points. The velocity for prediction was obtained by averaging instantaneous velocity for $\tau = 15$ time points prior to the track terminus, where τ was optimized by parameter search. This baseline model leads to an average mean squared error (MSE) (30 train/test splits) of ˜220. The RNN prediction model by comparison produces a significantly lower MSE of ˜192 (t-test $p < 0.001$), indicating that the RNN model is a superior motion predictor in the MuSC context (Fig. 2.6B). Representative track endings (length $\tau_{out} = 10$) produced by the RNN prediction model are displayed alongside the preceding track beginnings (length $\tau_{in} = 20$) and the ground truth track endings (Fig. 2.6A). In most cases the motion prediction reasonably approximates the cell's ground truth direction, but does not closely mirror the exact path (Fig. 2.6A, inset **i** and **ii**). In some events, the RNN model fails to predict even the correct direction of motion (Fig. 2.6A, inset **iii**). We performed the same experiment with mimetic myoblast simulations using $n = 10^5$ total samples, holding $n = 5000$ samples for validation. Similar to the MuSC results, RNN motion predictors achieved a markedly lower MSE of ˜1195, relative to the baseline kinematic model MSE of ˜9797 (t-test $p < 0.001$).

These results indicate that convolutional RNN models can be effective cell motility prediction models and are superior to simple linear kinematic approaches in some real world circumstances. RNN motility prediction models may therefore offer a scalable way to fit a uniquely tailored motion model to specific cell biology contexts. These cell-context specific RNN motility predictors may be

useful to improve multi-cell tracking performance, as outlined above.

2.4 Conclusion

Convolutional neural networks enable representation learning, or learning of features relevant for the description of a feature space. By representing cell motility as a 3D image, we show that 3D CNNs may be applied as an effective analytical tool. Using RNNs, motility may also be analyzed in the native multi-channel time series representation. Our results demonstrate that both approaches are capable of discriminating between simulated models of motion and multiple types of experimentally measured cell motility behaviors and are superior to a baseline heuristic model in some circumstances. In our experimentally measured cell motility data, we find that both CNN models effectively discriminate between different cell types, and different states of myogenic progenitor activation. We also find that CNN autoencoders can be trained effectively on either motion representation in an unsupervised fashion. Adapting the convolutional RNN autoencoder for motility prediction, we find that the RNN model is more effective at predicting MuSC motility than a kinematic model. Such prediction models may be useful for cell tracking. While we apply the methods described here to cell biology, there is no conceptual limitation that prevents application to other fields where discrimination based on motion recordings is desired. In the field of cell biology, analysis of motility with CNNs may allow for useful insights to be gathered in contexts where relevant features are non-obvious or laborious to construct.

Bibliography

1. Altschuler SJ, Wu LF. Cellular Heterogeneity: Do Differences Make a Difference? Cell. 2010;141(4):559–563. doi:10.1016/j.cell.2010.04.033.

2. Rodgers JT, King KY, Brett JO, Cromie MJ, Charville GW, Maguire KK, et al. mTORC1 controls the adaptive transition of quiescent stem cells from G0 to GAlert. Nature. 2014;510(7505):393–396. doi:10.1038/nature13255.

3. Wang YH, Israelsen WJ, Lee D, Yu VWC, Jeanson NT, Clish CB, et al. Cell-State-Specific Metabolic Dependency in Hematopoiesis and Leukemogenesis. Cell. 2014;158(6):1309–1323. doi:10.1016/j.cell.2014.07.048.

4. Yin Z, Sadok A, Sailem H, McCarthy A, Xia X, Li F, et al. A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol. 2013;15(7):860–871. doi:10.1038/ncb2764.

5. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science translational medicine. 2011;3(108):1–11. doi:10.1126/scitranslmed.3002564.

6. Tabesh A, Teverovskiy M, Pang HY, Kumar VP, Verbel D, Kotsianti A, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE transactions on medical imaging. 2007;26(10):1366–1378. doi:10.1109/TMI.2007.898536.

7. Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Medical Imaging. 2006;6(1):13–11. doi:10.1186/1471-2342-6-14.

8. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology. 2015;33(2):155–160. doi:10.1038/nbt.3102.

9. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature Biotechnology. 2015;33(11):1165–1172. doi:10.1038/nbt.3383.

10. Copley MR, Beer PA, Eaves CJ. Hematopoietic Stem Cell Heterogeneity Takes Center Stage. Cell Stem Cell. 2012;10(6):690–697. doi:10.1016/j.stem.2012.05.006.

11. Tierney MT, Sacco A. Satellite Cell Heterogeneity in Skeletal Muscle Homeostasis. Trends Cell Biol. 2016;26(6):434–444. doi:10.1016/j.tcb.2016.02.004.

12. Goodell MA, Nguyen H, Shroyer N. Somatic stem cell heterogeneity: diversity in the blood, skin and intestinal stem cell compartments. Nat Rev Mol Cell Biol. 2015;16(5):299–309. doi:10.1038/nrm3980.

13. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338–345. doi:10.1038/nature12625.

14. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501(7467):328–337. doi:10.1038/nature12624.

15. Waddington CH. An Introduction to Modern Genetics. 3rd ed. Oxon (England); New York: Routledge; 1939. Available from: `http://dx.doi.org/10.4324/9781315665412`.

16. Huang S. The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? Bioessays. 2011;34(2):149–157. doi:10.1002/bies.201100031.

17. Egolf DA. Equilibrium regained: from nonequilibrium chaos to statistical mechanics. Science. 2000;287(5450):101–104. doi:10.1126/science.287.5450.101.

18. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014;32(4):381–386. doi:10.1038/nbt.2859.

19. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nature methods. 2016;13(6):493–496. doi:10.1038/nmeth.3863.

20. Held M, Schmitz MHA, Fischer B, Walter T, Neumann B, Olma MH, et al. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. Nature methods. 2010;7(9):747–754. doi:10.1038/nmeth.1486.

21. Halfter W, Liverani D, Vigny M, Monard D. Deposition of extracellular matrix along the pathways of migrating fibroblasts. Cell and tissue research. 1990;262(3):467–481.

22. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. Nat Rev Mol Cell Biol. 2014;15(12):786–801. doi:10.1038/nrm3904.

23. Laird DJ, von Andrian UH, Wagers AJ. Stem Cell Trafficking in Tissue Development, Growth, and Disease. Cell. 2008;132(4):612–630. doi:10.1016/j.cell.2008.01.041.

24. Johnson GR, Buck TE, Sullivan DP, Rohde GK, Murphy RF. Joint modeling of cell and nuclear shape variation. Molecular Biology of the Cell. 2015;26(22):4046–4056. doi:10.1091/mbc.E15-06-0370.

25. Cooper S, Sadok A, Bousgouni V, Bakal C. Apolar and polar transitions drive the conversion between amoeboid and mesenchymal shapes in melanoma cells. Molecular Biology of the Cell. 2015;26(22):4163–4170. doi:10.1091/mbc.E15-06-0382.

26. Lock JG, Mamaghani MJ, Shafqat-Abbasi H, Gong X, Tyrcha J, Strömblad S. Plasticity in the macromolecular-scale causal networks of cell migration. PLoS One. 2014;9(2):e90593. doi:10.1371/journal.pone.0090593.

27. Kang J, Hsu CH, Wu Q, Liu S, Coster AD, Posner BA, et al. Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. Nature Biotechnology. 2015; p. 1–11. doi:10.1038/nbt.3419.

28. Loo LH, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nature methods. 2007;4(5):445–453. doi:10.1038/nmeth1032.

29. Rajaram S, Pavie B, Wu LF, Altschuler SJ. PhenoRipper: software for rapidly profiling microscopy images. Nature methods. 2012;9(7):635–637. doi:10.1038/nmeth.2097.

30. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7(10):R100. doi:10.1186/gb-2006-7-10-r100.

31. Gordonov S, Hwang MK, Wells A, Gertler FB, Lauffenburger DA, Bathe M. Time series modeling of live-cell shape dynamics for image-based phenotypic profiling. Integrative Biology. 2016;8(1):73–90. doi:10.1039/C5IB00283D.

32. Cohen AR, Gomes FLAF, Roysam B, Cayouette M. Computational prediction of neural progenitor cell fates. Nature methods. 2010;7(3):213–218. doi:10.1038/nmeth.1424.

33. Keren K, Pincus Z, Allen GM, Barnhart EL, Marriott G, Mogilner A, et al. Mechanism of shape determination in motile cells. Nature. 2008;453(7194):475–480. doi:10.1038/nature06952.

34. Mitchison TJ, Cramer LP. Actin-based cell motility and cell locomotion. Cell. 1996;84(3):371–379.

35. Cooper S, Sadok A, Bousgouni V, Bakal C. Apolar and polar transitions drive the conversion between amoeboid and mesenchymal shapes in melanoma cells. Molecular Biology of the Cell. 2015;26(22):4163–4170. doi:10.1091/mbc.E15-06-0382.

36. Buggenthin F, Buettner F, Hoppe PS, Endele M, Kroiss M, Strasser M, et al. Prospective identification of hematopoietic lineage choice by deep learning. Nature methods. 2017;14(4):1–7. doi:10.1038/nmeth.4182.

37. Thurston G, Jaggi B, Palcic B. Measurement of cell motility and morphology with an automated microscope system. Cytometry Part A. 1988;9(5):411–417. doi:10.1002/cyto.990090502.

38. Gail MH, Boone CW. The Locomotion of Mouse Fibroblasts in Tissue Culture. Biophysical journal. 1970;10(10):980–993. doi:10.1016/S0006-3495(70)86347-0.

39. Dunn GA, Brown AF. A unified approach to analysing cell motility. Journal of cell science Supplement. 1987;8:81–102.

40. Partin AW, Schoeniger JS, Mohler JL, Coffey DS. Fourier analysis of cell motility: correlation of motility with metastatic potential. Proceedings of the National Academy of Sciences of the United States of America. 1989;86(4):1254–1258.

41. Martens L, Monsieur G, Ampe C, Gevaert K, Vandekerckhove J. BMC Bioinformatics. BMC bioinformatics. 2006;7(1):289–6. doi:10.1186/1471-2105-7-289.

42. Sacan A, Ferhatosmanoglu H, Coskun H. CellTrack: an open-source software for cell tracking and motility analysis. Bioinformatics. 2008;24(14):1647–1649. doi:10.1093/bioinformatics/btn247.

43. Sebag AS, Plancade S, Raulet-Tomkiewicz C, Barouki R, Vert JP, Walter T. A generic methodological framework for studying single cell motility in high-throughput time-lapse data. Bioinformatics. 2015;31(12):320–328. doi:10.1093/bioinformatics/btv225.

44. Cordelières FP, Petit V, Kumasaka M, Debeir O, Letort V, Gallagher SJ, et al. Automated Cell Tracking and Analysis in Phase-Contrast Videos (iTrack4U): Development of Java Software Based on Combined Mean-Shift Processes. PLoS One. 2013;8(11):e81266–10. doi:10.1371/journal.pone.0081266.

45. Bray MA, Carpenter AE. CellProfiler Tracer: exploring and validating high-throughput, time-lapse microscopy image data. BMC bioinformatics. 2015;16(1):368. doi:10.1186/s12859-015-0759-x.

46. Xu J, Van Keymeulen A, Wakida NM, Carlton P, Berns MW, Bourne HR. Polarity reveals intrinsic cell chirality. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(22):9296–9300. doi:10.1073/pnas.0703153104.

47. Gorelik R, Gautreau A. Quantitative and unbiased analysis of directional persistence in cell migration. Nature Protocols. 2014;9(8):1931–1943. doi:10.1038/nprot.2014.131.

48. Saxton MJ, Jacobson K. Single-particle tracking: applications to membrane dynamics. Annual review of biophysics and biomolecular structure. 1997;26(1):373–399. doi:10.1146/annurev.biophys.26.1.373.

49. Viswanathan GM, Buldyrev SV, Havlin S, da Luz MG, Raposo EP, Stanley HE. Optimizing the success of random searches. Nature. 1999;401(6756):911–914. doi:10.1038/44831.

50. Viswanathan GM, Raposo EP, da Luz MGE. Levy flights and superdiffusion in the context of biological encounters and random searches. Physics of Life Reviews. 2008;5(3):133–150. doi:10.1016/j.plrev.2008.03.002.

51. Viswanathan GM, Afanasyev V, Buldyrev SV, Murphy EJ, Prince PA, Stanley HE. Levy flight search patterns of wandering albatrosses. Nature. 1996;381(6581):413–415. doi:10.1038/381413a0.

52. Humphries NE, Queiroz N, Dyer JRM, Pade NG, Musyl MK, Schaefer KM, et al. Environmental context explains Levy and Brownian movement patterns of marine predators. Nature. 2010;465(7301):1066–1069. doi:10.1038/nature09116.

53. Rahman A. Correlations in the Motion of Atoms in Liquid Argon. Physical Review. 1964;136(2A):A405–A411. doi:10.1103/PhysRev.136.A405.

54. Parry BR, Surovtsev IV, Cabeen MT, O'Hern CS, Dufresne ER, Jacobs-Wagner C. The Bacterial Cytoplasm Has Glass-like Properties and Is Fluidized by Metabolic Activity. Cell. 2014;156(1-2):183–194. doi:10.1016/j.cell.2013.11.028.

55. Mandelbrot BB, Wallis JR. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resources Research. 1969;5(5):967–988. doi:10.1029/WR005i005p00967.

56. Bryce RM, Sprague KB. Revisiting detrended fluctuation analysis. Sci Rep. 2012;2:1–6. doi:10.1038/srep00315.

57. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9:2579–2605.

58. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. 2014;31(3):274–295. doi:10.1007/s00357-014-9161-z.

59. Ho TK. Random decision forests. In: 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press; 1995. p. 278–282. Available from: `http://ieeexplore.ieee.org/document/598994/`.

60. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.

61. Truitt ML, Conn CS, Shi Z, Pang X, Tokuyasu T, Coady AM, et al. Differential Requirements for eIF4E Dose in Normal Development and Cancer. Cell. 2015;162(1):59–71. doi:10.1016/j.cell.2015.05.049.

62. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust package for determining the number of clusters in a dataset. Journal of Statistical Software. 2014;61(6):1–36.

63. Rousseeuw PJ. Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.

64. Rencher AC, Christensen WF. Methods of Multivariate Analysis. John Wiley & Sons; 2013.

65. Enterline HT, Coman DR. The ameboid motility of human and animal neoplastic cells. Cancer. 1950;3(6):1033–1038.

66. Wells A, Chao YL, Grahovac J, Wu Q. Cell motility in carcinoma metastasis as modulated by switching between epithelial and mesenchymal phenotypes. Frontiers in Bioscience. 2014;16(1):815. doi:10.2741/3722.

67. Verschueren H, Vandertalen I, Dewit J, Debraekeleer J, Debaetselier P. Metastatic Competence of Bw5147 T-Lymphoma Cell-Lines Is Correlated with in-Vitro Invasiveness, Motility and F-Actin Content. Journal of Leukocyte Biology. 1994;55(4):552–556.

68. Doyle GM, Sharief Y, Mohler JL. Prediction of Metastatic Potential by Cancer Cell Motility in the Dunning R-3327 Prostatic Adenocarcinoma In Vivo Model. The Journal of Urology. 1992;147(2):514–518. doi:10.1016/S0022-5347(17)37291-9.

69. Smith CL, Kilic O, Schiapparelli P, Guerrero-Cazares H, Kim DH, Sedora-Roman NI, et al. Migration Phenotype of Brain-Cancer Cells Predicts Patient Outcomes. Cell Reports. 2016;15(12):2616–2624. doi:10.1016/j.celrep.2016.05.042.

70. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research. 2002;2(2):265–292.

71. Bentzinger CF, von Maltzahn J, Dumont NA, Stark DA, Wang YX, Nhan K, et al. Wnt7a stimulates myogenic stem cell motility and engraftment resulting in improved muscle strength. J Cell Biol. 2014;205(1):97–111. doi:10.1083/jcb.201310035.

72. Siegel AL, Atchison K, Fisher KE, Davis GE, Cornelison DDW. 3D Timelapse Analysis of Muscle Satellite Cell Motility. Stem Cells. 2009;27(10):2527–2538. doi:10.1002/stem.178.

73. Webster MT, Fan CM. c-MET Regulates Myoblast Motility and Myocyte Fusion during Adult Skeletal Muscle Regeneration. PLoS One. 2013;8(11):e81757–16. doi:10.1371/journal.pone.0081757.

74. Bae GU, Gaio U, Yang YJ, Lee HJ, Kang JS, Krauss RS. Regulation of myoblast motility and fusion by the CXCR4-associated sialomucin, CD164. J Biol Chem. 2008;283(13):8301–8309. doi:10.1074/jbc.M706730200.

75. Brack AS, Rando TA. Tissue-specific stem cells: lessons from the skeletal muscle satellite cell. Cell Stem Cell. 2012;10(5):504–514. doi:10.1016/j.stem.2012.04.001.

76. Wang YX, Rudnicki MA. Satellite cells, the engines of muscle repair. Nat Rev Mol Cell Biol. 2011;13(2):127–133. doi:10.1038/nrm3265.

77. Schultz E, Jaryszak DL, Valliere CR. Response of satellite cells to focal skeletal muscle injury. Muscle & nerve. 1985;8(3):217–222. doi:10.1002/mus.880080307.

78. Hughes SM, Blau HM. Migration of myoblasts across basal lamina during skeletal muscle development. Nature. 1990;345(6273):350–353. doi:10.1038/345350a0.

79. Florini JR, Magri KA. Effects of growth factors on myogenic differentiation. The American journal of physiology. 1989;256(4 Pt 1):C701–11.

80. Sheehan SM, Allen RE. Skeletal muscle satellite cell proliferation in response to members of the fibroblast growth factor family and hepatocyte growth factor. Journal of cellular physiology. 1999;181(3):499–506.

81. Bischoff R. Chemotaxis of skeletal muscle satellite cells. Developmental dynamics. 1997;208(4):505–515.

82. Do MKQ, Suzuki T, Gerelt B, Sato Y, Mizunoya W, Nakamura M, et al. Time-coordinated prevalence of extracellular HGF, FGF2 and TGF-β3 in crush-injured skeletal muscle. Animal Science Journal. 2012;83(10):712–717. doi:10.1111/j.1740-0929.2012.01057.x.

83. Bischoff R. A satellite cell mitogen from crushed adult muscle. Dev Biol. 1986;115(1):140–147. doi:10.1016/0012-1606(86)90235-6.

84. Cornelison DD, Wold BJ. Single-cell analysis of regulatory gene expression in quiescent and activated mouse skeletal muscle satellite cells. Dev Biol. 1997;191(2):270–283. doi:10.1006/dbio.1997.8721.

85. Chakkalakal JV, Christensen J, Xiang W, Tierney MT, Boscolo FS, Sacco A, et al. Early forming label-retaining muscle stem cells require p27kip1 for maintenance of the primitive state. Development. 2014;141(8):1649–1659. doi:10.1242/dev.100842.

86. Trapnell C. Defining cell types and states with single-cell genomics. Genome Research. 2015;25(10):1491–1498. doi:10.1101/gr.190595.115.

87. Battle C, Broedersz CP, Fakhri N, Geyer VF, Howard J, Schmidt CF, et al. Broken detailed balance at mesoscopic scales in active biological systems. Science (New York, NY). 2016;352(6285):604–607. doi:10.1126/science.aac8167.

88. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, Grenier JK, et al. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(6):1826–1831. doi:10.1073/pnas.0808843106.

89. Brooks RA. Intelligence without representation. Artificial Intelligence. 1991;47(1-3):139–159. doi:10.1016/0004-3702(91)90053-M.

90. Jaqaman K, Loerke D, Mettlen M, Kuwata H, Grinstein S, Schmid SL, et al. Robust single-particle tracking in live-cell time-lapse sequences. Nature methods. 2008;5(8):695–702. doi:10.1038/nmeth.1237.

91. Shea KL, Xiang W, LaPorta VS, Licht JD, Keller C, Basson MA, et al. Sprouty1 regulates reversible quiescence of a self-renewing adult muscle stem cell pool during regeneration. Cell Stem Cell. 2010;6(2):117–129. doi:10.1016/j.stem.2009.12.015.

92. E Michalska A. Isolation and propagation of mouse embryonic fibroblasts and preparation of mouse embryonic feeder layer cells. Current protocols in stem cell biology. 2007;Chapter 1:Unit1C.3. doi:10.1002/9780470151808.sc01c03s3.

93. Jacob M, Unser M. Design of steerable filters for feature detection using Canny-like criteria. IEEE transactions on pattern analysis and machine intelligence. 2004;26(8):1007–1019. doi:10.1109/TPAMI.2004.44.

94. Serra J. *Image Analysis and Mathematical Morphology*; 1982. Academic Press. Available from: `http://cds.cern.ch/record/235415`.

95. Otsu N. A threshold selection method from gray-level histograms. Automatica. 1975;.

96. Mandelbrot BB, Van Ness JW. Fractional Brownian motions, fractional noises and applications. SIAM review. 1968;10(4):422–437. doi:10.1137/1010093.

97. Berg HC. Random Walks in Biology. Princeton University Press; 1993. Available from: `https://books.google.com/books/about/Random_Walks_in_Biology.html?id=DjdgXGLoJY8C`.

98. Dieker T. Simulation of fractional Brownian motion. Amsterdam; 2004. Available from: `http://www.columbia.edu/~ad3217/fbm/thesisold.pdf`.

99. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013;46(1):243–256. doi:10.1016/j.patcog.2012.07.021.

100. van der Maaten L. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research. 2014;15(1):3221–3245.

101. Krijthe J. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation (R package version 0.10). Computer Software; 2015. Available from: `http://scholar.google.com/scholar?q=related:PqRLA2L8ch4J:scholar.google.com/&hl=en&num=20&as_sdt=0,5`.

102. Wattenberg M, Viégas F, Johnson I. How to Use t-SNE Effectively. Distill. 2016;1(10). doi:10.23915/distill.00002.

103. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. Neural Networks. 2000;13(4-5):411–430. doi:10.1016/S0893-6080(00)00026-5.

104. Helwig NE. Independent Component Analysis. 2015;.

105. Milescu LS, Yildiz A, Selvin PR, Sachs F. Maximum likelihood estimation of molecular motor kinetics from staircase dwell-time sequences. Biophysj. 2006;91(4):1156–1168. doi:10.1529/biophysj.105.079541.

106. Landowne D, Yuan B, Magleby KL. Exponential Sum-Fitting of Dwell-Time Distributions without Specifying Starting Parameters. Biophysj. 2013;104(11):2383–2391. doi:10.1016/j.bpj.2013.04.030.

107. Zou W, Larson RG. A mesoscopic simulation method for predicting the rheology of semi-dilute wormlike micellar solutions. Journal of Rheology. 2014;58(3):681–721. doi:10.1122/1.4868875.

108. Lilliefors HW. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association. 1969;64(325):387–389. doi:10.1080/01621459.1969.10500983.

109. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825–2830.

110. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.

111. Domingos P. A Few Useful Things to Know About Machine Learning. Communications of the Acm. 2012;55(10):78–87. doi:10.1145/2347736.2347755.

112. Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms. In: International Conference on Machine Learning; 2006. p. 1–8.

113. Hsu CW, Chang CC, Lin CJ. A Practical Guide to Support Vector Classification. NTU. 2003; p. 1–16.

114. Steinbach M, Ertoz L, Kumar V. New Directions in Statistical Physics. In: The Challenges of Clustering High Dimensional Data; 2004. p. 279–303.

115. Lauffenburger DA, Horwitz AF. Cell migration: A physically integrated molecular process. Cell. 1996;84(3):359–369.

116. Coman DR. Mechanism of the Invasiveness of Cancer. Science. 1947;105(2727):347–348. doi:10.1126/science.105.2727.347.

117. Sahai E, Olson MF, Marshall CJ. Cross-talk between Ras and Rho signalling pathways in transformation favours proliferation and increased motility. The EMBO Journal. 2001;20(4):755–766. doi:10.1093/emboj/20.4.755.

118. Pollock CB, Shirasawa S, Sasazuki T, Kolch W, Dhillon AS. Oncogenic K-RAS is required to maintain changes in cytoskeletal organization, adhesion, and motility in colon cancer cells. Cancer Res. 2005;65(4):1244–1250. doi:10.1158/0008-5472.CAN-04-1911.

119. Nieto MA. Epithelial plasticity: a common theme in embryonic and cancer cells. Science (New York, NY). 2013;342(6159):1234850–1234850. doi:10.1126/science.1234850.

120. Buckingham M, Bajard L, Chang T, Daubas P, Hadchouel J, Meilhac S, et al. The formation of skeletal muscle: from somite to limb. Journal of Anatomy. 2003;202(1):59–68.

121. Alfaro LAS, Dick SA, Siegel AL, Anonuevo AS, McNagny KM, Megeney LA, et al. CD34 Promotes Satellite Cell Motility and Entry into Proliferation to Facilitate Efficient Skeletal Muscle Regeneration. Stem Cells. 2011;29(12):2030–2041. doi:10.1002/stem.759.

122. Kimmel JC, Chang AY, Brack AS, Marshall WF. Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. PLoS computational biology. 2018;14(1):e1005927–29. doi:10.1371/journal.pcbi.1005927.

123. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. 2012; p. 1097–1105.

124. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular Systems Biology. 2016;12(7):878. doi:10.15252/msb.20156651.

125. He K. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Microsoft Research, Redmond, United States; 2016. p. 770–778. Available from: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=84986274465&origin=inward.

126. Van Valen DA, Van Valen DA, Kudo T, Kudo T, Lane KM, Lane KM, et al. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLoS computational biology. 2016;12(11):e1005177. doi:10.1371/journal.pcbi.1005177.

127. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

128. Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: ICLR; 2013.Available from: http://arxiv.org/abs/1312.6114.

129. Bengio Y, Lamblin P, Popovici D. Greedy layer-wise training of deep networks. NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems. 2007;.

130. Conneau A, Schwenk H. Very deep convolutional networks for text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2016. p. 1107–1116. Available from: https://pdfs.semanticscholar.org/f797/fd44b9ddd5845611eb7a705ca9464a8819d1.pdf.

131. Neal L, Briggs F, Raich R. Time-frequency segmentation of bird song in noisy acoustic environments. Acoustics. 2011; p. 2012–2015. doi:10.1109/ICASSP.2011.5946906.

132. Bashivan P, Rish I, Yeasin M, Codella N. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. In: ICLR; 2015.Available from: https://arxiv.org/abs/1511.06448.

133. Molchanov P, Gupta S, Kim K, Kautz J. Hand Gesture Recognition With 3D Convolutional Neural Networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 1–7. Available

from: http://www.cv-foundation.org/openaccess/content_cvpr_workshops_2015/W15/html/Molchanov_Hand_Gesture_Recognition_2015_CVPR_paper.html.

134. Wu D, Pigou L, Kindermans PJ, Le NDH, Shao L, Dambre J, et al. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. IEEE transactions on pattern analysis and machine intelligence. 2016;38(8):1583–1597. doi:10.1109/TPAMI.2016.2537340.

135. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond Short Snippets: Deep Networks for Video Classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015; p. 4694–4702. doi:10.1109/CVPR.2015.7299101.

136. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale Video Classification with Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014; p. 1725–1732. doi:10.1109/CVPR.2014.223.

137. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research. 2016;44(11):e107–e107. doi:10.1093/nar/gkw226.

138. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology. 2015;33(8):831–838. doi:10.1038/nbt.3300.

139. Lipton ZC, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv. 2015;.

140. Cui Z, Chen W, Chen Y. Multi-Scale Convolutional Neural Networks for Time Series Classification. arXiv. 2016;.

141. Wang Z, Yan W, Oates T. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. arXiv. 2016;.

142. Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE; 2014. p. 730–734. Available from: http://ieeexplore.ieee.org/document/7486599/.

143. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning; 2010. p. 807–814. Available from: http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_NairH10.pdf.

144. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989;2(5):359–366. doi:10.1016/0893-6080(89)90020-8.

145. Srivastava N, Hinton GE, Krizhevsky A. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 2014;15:1929–1958.

146. Zeiler MD. ADADELTA: An Adaptive Learning Rate Method. arXiv. 2012;doi:arxiv:1212.5701.

147. Jaqaman K, Loerke D, Mettlen M, Kuwata H, Grinstein S, Schmid SL, et al. Robust single-particle tracking in live-cell time-lapse sequences. Nature methods. 2008;5(8):695–702. doi:10.1038/nmeth.1237.

148. Lomholt MA, Tal K, Metzler R, Joseph K. Lévy strategies in intermittent search processes are advantageous. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(32):11055–11059. doi:10.1073/pnas.0803117105.

149. Mandelbrot BB. A Fast Fractional Gaussian Noise Generator. Water Resources Research. 1971;7(3):543–553. doi:10.1029/WR007i003p00543.

150. Prechelt L. Early Stopping — But When? In: Montavon G, Orr GB, Müller KR, editors. Neural Networks: Tricks of the Trade: Second Edition. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 53–67. Available from: https://doi.org/10.1007/978-3-642-35289-8_5.

151. Sivic J, Zisserman A. Efficient visual search of videos cast as text retrieval. IEEE transactions on pattern analysis and machine intelligence. 2009;31(4):591–606. doi:10.1109/TPAMI.2008.111.

152. Skylaki S, Hilsenbeck O, Schroeder T. Challenges in long-term imaging and quantification of single-cell dynamics. Nature Biotechnology. 2016;34(11):1137–1144. doi:10.1038/nbt.3713.

153. Luo W, Xing J, Milan A, Zhang X, Liu W, Zhao X, et al. Multiple Object Tracking: A Literature Review. arXiv. 2014;.

154. Bütepage J, Black M, Kragic D, Kjellström H. Deep representation learning for human motion prediction and classification. arXiv. 2017;.

155. Martinez J, Black MJ, Romero J. On human motion prediction using recurrent neural networks. arXiv. 2017;.

156. Ma X, Tao Z, Wang Y, Yu H, Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C. 2015;54(C):187–197. doi:10.1016/j.trc.2015.03.014.

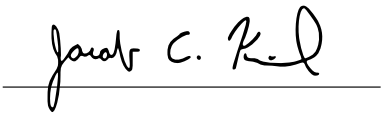
157. Shi X, Chen Z, Wang H, Yeung DY, Wong Wk, Woo Wc. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv. 2015;.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  **Date** 2018-04-26