Deconvolution of subcellular protrusion heterogeneity and the underlying actin regulator dynamics from live cell imaging

Chuangqi Wang[1,*], Hee June Choi[1,*], Sung-Jin Kim[1], Yongho Bae[2], Kwonmoo Lee[1]
[1]Department of Biomedical Engineering, Worcester Polytechnic Institute, Massachusetts, 01609, USA
[2]Department of Pathology and Anatomical Sciences, University at Buffalo, New York, 14214, USA
[*]These authors equally contributed to this work.
Corresponding Author: Kwonmoo Lee, Email: klee@wpi.edu

Abstract

Cell protrusion is morphodynamically heterogeneous at the subcellular level. However, the mechanistic understanding of protrusion activities is usually based on the ensemble average of actin regulator dynamics at the cellular or population levels. Here, we establish a machine learning-based computational framework called HACKS (deconvolution of Heterogeneous Activity Coordination in cytosKeleton at a Subcellular level) to deconvolve the subcellular heterogeneity of lamellipodial protrusion. HACKS identifies distinct subcellular protrusion phenotypes hidden in highly heterogeneous protrusion activities and reveals their underlying actin regulator dynamics. The association between each protrusion phenotype and the actin regulator dynamics is further corroborated by predicting the protrusion phenotype based on actin regulator dynamics. Using our method, we discovered the hidden rare 'accelerating' protrusion phenotype in addition to 'fluctuating' and 'periodic' protrusions. Intriguingly, the accelerating protrusion was driven predominantly by VASP-mediated actin elongation rather than by Arp2/3-mediated actin nucleation. Our analyses also suggested that VASP controls protrusion velocity more directly than Arp2/3 complex, thereby playing a significant role in the accelerating protrusion phenotype. Taken

together, we have demonstrated that HACKS allows us to discover the fine differential coordination of molecular dynamics underlying subcellular protrusion heterogeneity via a machine learning analysis of live cell imaging data.

Introduction

Cell protrusion is driven by spatiotemporally fluctuating actin assembly processes and is morphodynamically highly heterogeneous at the subcellular level[1-3]. Because protrusion determines the directionality and persistence of cell movements or facilitates the exploration of the surrounding environment[4], elucidating the underlying molecular dynamics associated with subcellular protrusion heterogeneity is crucial to understanding the biology of cellular movement. Recent studies of the vital roles of cell protrusion in regeneration[5,6], cancer invasiveness and metastasis[7-9], and the environmental exploration of leukocytes[10] further emphasize the physiological and pathophysiological implication of understanding the fine molecular details of protrusion mechanisms. Although there have been considerable progresses in analyzing individual functions of actin regulators, it remains unclear as to how actin regulators are spatiotemporally coordinated to collectively organize cell protrusion *in situ*. Moreover, it is a formidable task to dissect the actin regulator dynamics involved with cell protrusion because those dynamics are highly heterogeneous and fluctuate on both the micron length scale and the minute time scale[11-13].

Although advances in computational image analysis on live cell movies have allowed us to study the dynamic aspects of molecular and cellular events at the subcellular level, the significant degree of heterogeneity in molecular and subcellular dynamics complicates the extraction of useful information from complex cellular behavior. The current method of characterizing molecular dynamics involves averaging molecular activities at the cellular level, which significantly conceals

the fine differential subcellular coordination of dynamics among actin regulators. For example, our previous study employing the ensemble average method with the event registration showed that the leading edge dynamics of Arp2/3 and VASP were almost indistinguishable[13]. Over the past decade, hidden variable cellular phenotypes in heterogeneous cell populations have been uncovered by applying machine learning analyses[14,15]. However, these analyses primarily focused on static datasets acquired at the single-cell level, such as immunofluorescence[16], mass cytometry[17], and single-cell RNA-Seq[18] datasets. Although some studies have examined the cellular heterogeneity of the migratory mode[19,20], subcellular protrusion heterogeneity has not yet been addressed. Moreover, elucidating the molecular mechanisms that generate each subcellular phenotype has been experimentally limited because it is a challenging task to manipulate particular subclasses of molecules at the subcellular level with fine spatiotemporal resolution, even with the advent of optogenetics.

To address this challenge, we developed a machine learning-based computational analysis pipeline called HACKS (deconvolution of Heterogeneous Activity Coordination in cytosKeleton at a Subcellular level) (Fig. 1) for live cell imaging data by combining unsupervised and supervised machine learning approaches with our local sampling and registration method[13]. HACKS allowed us to deconvolve the subcellular heterogeneity of protrusion phenotypes and statistically link them to the dynamics of actin regulators at the leading edge of migrating cells. Based on our method, we quantitatively characterized subcellular protrusion phenotypes, including fluctuating, periodic and accelerating protrusions from migrating PtK1 cells, under unperturbed conditions. Furthermore, each protrusion phenotype was demonstrated to be associated with the specific temporal coordination of actin regulator, Arp2/3 and VASP at the leading edge. Intriguingly, VASP was shown to control the protrusion velocity more directly than the Arp2/3 complex, and

VASP-mediated actin elongation played an unexpected role in promoting persistently accelerating protrusion. These results demonstrated that our machine learning framework HACKS provides an effective method to elucidate the fine differential molecular dynamics that underlie subcellular protrusion heterogeneity.

Results

A computational framework to deconvolve subcellular protrusion heterogeneity and the underlying molecular dynamics

To deconvolve the heterogeneity of the subcellular protrusion activity and their regulatory proteins at fine spatiotemporal resolution, we developed a computational analysis pipeline, HACKS (Fig. 1), which combines unsupervised and supervised machine learning methods. HACKS allowed us to (1) identify distinct subcellular protrusion phenotypes based on a time series clustering analysis of heterogeneous subcellular protrusion velocities extracted from live cell movies using an unsupervised machine learning approach (Fig. 1a-e), (2) associate each protrusion phenotype with pertinent actin regulator dynamics by comparing the average temporal patterns of protrusion velocities with those of actin regulators (Fig. 1f), (3) perform highly specified correlation analyses between the protrusion velocities and actin regulator dynamics of protrusion phenotypes to establish their association with fine mechanistic details (Fig. 1g) and (4) evaluate the predictability of the protrusion phenotype from the actin regulator dynamics to corroborate their correlation using a supervised machine learning approach (Fig. 1h). The framework provided mechanistic insight into how the differential coordination of actin regulator dynamics organizes various protrusion phenotypes.

Identification of subcellular protrusion phenotypes by a time series clustering analysis

Sample videos for the analysis were prepared by taking time-lapse movies of PtK1 cells expressing fluorescently tagged actin, Arp3, VASP and a cytoplasmic marker, HaloTag, with a spinning disk confocal microscope for approximately 200 frames at 5 sec/frame[11] (Fig. 1a). After segmenting the leading edge of each cell by multiple probing windows with a size of 500 by 500 nm[13] (Fig. 1b), time series of velocities[11] and fluorescence intensities of the tagged molecules[12,13] acquired from each probing window were quantified (Fig. 1b). After registering protrusion onset at time zero (t=0), the time- series were aligned using the protrusion onset as a temporal fiduciary[13] (Fig. 1c). To ensure a uniform time length of the data for the subsequent clustering analysis, we selected the first 50 frames (250 seconds) of protrusion segments, which is about the average protrusion duration[13] (see Methods), from the pooled velocity time series (Fig. 1d). These protrusion segments were subjected to a further time series cluster analysis, whereas the fluorescence intensity data were later used to correlate the protrusion phenotypes and their underlying molecular dynamics.

The selected time series of the registered protrusion velocity contained a substantial amount of intrinsic fluctuations, hindering the identification of distinct clusters. Therefore, we first denoised the time series velocity profile using Empirical Mode Decomposition[21] and discretized the data using SAX (Symbolic Aggregate approximation, see Methods)[22] to reduce the dimensionality and complexity of the data (Fig. 1e). We then extracted distinct patterns from fluctuating velocity time series by combining the autocorrelation distance measure with the Density Peak clustering[23] (Fig. 1e). The distance measures between different time series were calculated using the squared Euclidean distances between the corresponding autocorrelation functions of each discretized time series. This autocorrelation distance partitioned the fluctuating time series of similar underlying patterns into the same cluster, enabling us to identify clusters

with clear dynamic patterns. Following the autocorrelation distance measure, we applied the Density Peak clustering algorithm, which has been shown to be superior to conventional K-means in partitioning data with complex cluster shapes[23]. As a result, the density graph in Fig. 2a revealed four distinct clusters of protrusion phenotypes. Of note, when we tested different sets of algorithms, such as conventional Euclidean distance and K-means, identified clusters were much less distinguishable (Extended Data Fig. 1 and 2, Supplementary Text), demonstrating the capability of our clustering workflow. Furthermore, we could not identify substantial differences among the velocity cluster profiles in each molecule (actin, Arp3, VASP, HaloTag) (protrusion velocities in Extended Data Fig. 3), confirming that our clustering results are not skewed by the specific data set. The number of cells and probed windows used in the time series clustering analysis are presented in Extended Data Table 1.

Distinct subcellular protrusion phenotypes: the fluctuating, periodic and accelerating protrusion

The visual inspection of the average velocity profiles of the identified clusters (Fig. 2b-e) demonstrated that the overall differences among the protrusion phenotypes originated from differences in the timing and number of peaks the velocity reached. Whereas Cluster I did not exhibit dramatic changes in protrusion velocities after reaching its peak at the earlier part of the protrusion segment (Fig. 2b), the remaining clusters exhibited substantial acceleration or deceleration in the protrusion velocities with varying timing and number. Clusters II-1 (Fig. 2c) and II-2 (Fig. 2e) exhibited periodic fluctuation in the acceleration and deceleration of protrusion velocities with differential timing, reaching their velocity peak more than twice. Conversely, Cluster III (Fig. 2d) demonstrated persistently accelerating behavior where protrusion velocities continued to increase until the late phase of the protrusion. Because Cluster II-1 and II-2 exhibited

essentially similar periodic behaviors with different time intervals, we merged them into a single cluster. Thus, we interpolated the time series data in Cluster II-2 and pooled them into Cluster II-1 (Fig. 2g) based on the aligned average velocity profiles generated by the Dynamic Time Warping algorithm[24] (Fig. 2f), thereby finalizing the identified protrusion phenotypes into three distinct clusters. Clusters I, II and III comprised 50.1%, 39.2%, and 10.7% of the entire sample, respectively (Fig. 2l), and individual cells generally exhibited similar tendencies. Nevertheless, cell-to-cell variability in cluster distribution persisted (Extended Data Fig. 4d), suggesting that the clusters may reflect individual cellular responses to differential cellular contexts or microenvironments.

The validity of our clustering result was confirmed by visually inspecting the velocity activity map (Fig. 2m-p). Of the ensemble global pattern (Fig. 2m), Clusters II (Fig. 2o) and III (Fig. 2p) exhibited clearly distinguishable patterns, whereas Cluster I (Fig. 2n) contained fluctuating velocity profiles. The multidimensional scaling (Fig. 2h), t-SNE (Fig. 2i), silhouette (Fig. 2j), and order distance plots (Fig. 2k) of the clustering results (Extended Data Fig. 4a-c before merging Clusters II-1 and II-2) further confirmed the stability and tightness of Clusters II and III but suggested residual heterogeneity in Cluster I, which is in agreement with the velocity activity map.

Notably, this procedure revealed the differential subcellular protrusion phenotypes with distinct velocity profiles using our time series clustering framework. These variabilities were previously hidden when the entire time series was ensemble averaged without taking into account the subcellular protrusion heterogeneity[13]. The visualization of the edge evolution and kymographs of the exemplified edges assigned to each cluster representatively manifested the morphodynamic features of each protrusion phenotype (Fig. 3a, Supplementary Movie 1-3). Because Cluster II clearly exhibited periodic edge evolution, we refer to Cluster II as ‘periodic protrusion’.

Conversely, Cluster III showed accelerated edge evolution, and therefore we refer to Cluster III as 'accelerating protrusion'. As in summarized in Extended Data Table 2, this phenotype was found only by the specific combination of the ACF dissimilarity measure and Density Peak clustering (Supplementary Text). To the best of our knowledge, this study is the first to quantitatively characterize persistently 'accelerating' cell protrusion, whereas previous studies described persistent protrusion based on the protrusion distance over longer time scales[11,13,25,26]. Finally, Cluster I was named the 'fluctuating protrusion' cluster because of the irregularity of its velocity profiles.

Differential molecular dynamics of actin regulators underlying the accelerating protrusion phenotype

We hypothesized that the distinctive subcellular protrusion phenotypes arise from the differential spatiotemporal regulation of actin regulators. Therefore, we next investigated the relationship between the velocity profiles of each protrusion phenotype and the fluctuation of the signal intensities of actin and selected actin regulators for each protrusion phenotype. We selected a set of fluorescently tagged molecules to be expressed and monitored; SNAP-tag-actin, HaloTag-Arp3 (tagged on the C-terminus), which represented the Arp2/3 complex involved in actin nucleation, and HaloTag-VASP or GFP-VASP, which represented actin elongation. A diffuse fluorescent marker, HaloTag labeled with tetramethylrhodamine (TMR) ligands[27], was used as a control signal. The fluorescence intensities of each tagged molecule were acquired from each probing window along with the protrusion velocities (Fig. 1b). The time-series of the fluorescence intensities of each molecule were then grouped and averaged according to the assigned protrusion phenotypes (Fig. 1f and Fig. 3c-f, h-k, m-p, r-u, Extended Data Fig. 3).

Intriguingly, Cluster III exhibited distinctive differential molecular dynamics among the molecules and in relation to velocity profiles (Fig. 3r-t), whereas the molecular dynamics of actin, Arp3 and VASP all exhibited patterns similar to those of the velocity profiles in Clusters I and II (Fig. 3h-j, m-o). More specifically, whereas the protrusion velocity continued to increase until the late stages of the protrusion segment in the accelerating protrusion cluster (Cluster III) (Fig. 3q), the actin fluorescence intensity soon reached its maximum in the early phase and remained constant (Fig. 3r). This pattern indicates that edge movement during accelerating protrusion is mediated by the elongation of existing actin filaments rather than *de novo* actin nucleation. Conversely, Clusters I and II exhibited increased actin intensity at the leading edge along with increased protrusion velocity (Fig. 3h and m), indicating that *de novo* actin nucleation mediates protrusion. In accordance with the plateaued actin intensities in Cluster III (Fig. 3r), the VASP intensities began to increase at protrusion onset and continued to increase (Fig. 3t), whereas the Arp3 intensity continued to decrease until the later phase after reaching its peak (Fig. 3s).

These findings suggest that actin elongation by VASP plays a crucial role in driving accelerating protrusion, whereas actin nucleation by Arp2/3 plays a minor role. This finding is surprising because the Arp2/3 complex has been considered a major actin nucleator that drives lamellipodial protrusion. Nevertheless, Arp2/3 seemed to play a role in the earlier part of the protrusion segment. Approximately 80 seconds after protrusion onset, the Arp3 intensity reached its peak (Fig. 3s), and velocity acceleration temporarily stopped (Fig. 3q). Notably, the Arp3 intensities began to increase 50 seconds prior to the protrusion onset in Cluster III (Fig. 3s), whereas they began to increase at the onset of protrusion in Clusters I and II (Fig. 3i and n). These findings imply that the Arp2/3 complex maintains its role in the early phase, and VASP then takes over the role of the Arp2/3 complex to drive the later stages of accelerating protrusion.

The specificity of the relationship between the protrusion phenotypes and the underlying molecular dynamics was further validated with a control experiment using HaloTag-TMR (Fig. 3f, k, p and u). Diffuse cytoplasmic fluorescence did not exhibit a cluster-specific pattern. Instead, it inversely correlated with the protrusion velocity, suggesting that the edges becomes thinner as the protrusion velocity increases[13]. Notably, the differential dynamics of Arp3 and VASP were not observed when the entire time series dataset was ensemble averaged[13] (Fig. 3d and e). This finding demonstrates the power of our computational framework in revealing the hidden differential subcellular dynamics of actin regulators involved in the generation of heterogeneous morphodynamic phenotypes.

VASP recruitment strongly correlates with protrusion velocity

To quantitatively assess the coordination between protrusion velocities and the dynamics of actin regulators, we performed a time correlation analysis by calculating Pearson's correlation coefficients between protrusion velocities and actin regulator intensities with varying time lags in the same windows and averaged over different sampling windows (Figs. 1g and 4a). For actin, significant correlations were identified between the protrusion velocity and actin intensities in all clusters (Fig. 4c). However, whereas Clusters I and II both exhibited 10-20 second time lags in the peak correlation, the peak correlation in Cluster III did not exhibit a time lag. A previous study suggested that the time lag is attributable to reinforced actin assembly driven by Arp2/3[13,28]. The absence of a time lag in Cluster III suggests that the actin dynamics in Cluster III and those in Clusters I and II are driven by different mechanisms. Consistent with this idea, the Arp3 intensities did not significantly correlate with the protrusion velocities in Cluster III, whereas they exhibited similar correlation peaks in Clusters I and II (Fig. 4d). Conversely, the maximum correlation of VASP in Cluster III was significantly stronger than those in Clusters I and II (Fig. 4e). These

findings suggest that VASP plays a more important role in promoting the accelerating protrusion (Cluster III), whereas Arp2/3 complex plays a role in non-accelerating protrusions (Clusters I and II). Furthermore, a comparison of the maximum correlations in each cluster revealed that VASP exhibited significantly stronger correlations than the Arp2/3 complex in all clusters (Fig. 4g, Extended Data Table 3).

Although the above-described conventional time correlation analysis effectively demonstrated the overall correlation between molecular dynamics and the protrusion velocity, its ability to reveal changes in this correlation over time as the protrusion progresses is limited. In other words, the correlation between the protrusion velocities and the fluorescence intensities for each specific time point was not examined in the previous analyses. Therefore, we performed sample-based correlation analyses whereby calculating pairwise Pearson correlation coefficients, $c(\{V\}_{t_i}, \{I\}_{t_j})$, between the sample of the protrusion velocity, $\{V\}_{t_i}$, at the registered time, t_i, and the sample of the actin regulator intensity $\{I\}_{t_i}$, at the registered time, t_j, over the entire probing window population (Fig. 1g and Fig. 4b)[13].

As expected, the pairwise time correlation analysis between the actin intensities and protrusion velocities (Fig. 5h, l, p, t, x and B) further supported the proposition that accelerating protrusions are mediated by the elongation of pre-existing actin filaments, whereas actin nucleation is responsible for non-accelerating protrusions. The instantaneous positive correlations between the actin intensities and protrusion velocities at the leading edge found in Clusters I and II (Fig. 4h, l, t and x) were absent in Cluster III (Fig. 4p and B). Notably, in the previous time lag correlation analysis, the correlation for actin persisted in Cluster III (Fig. 4c). This finding suggests that pairwise correlations at specific time points can effectively and more precisely reveal the various aspects of the coordination between protrusion velocities and the underlying molecular dynamics.

Intriguingly, we did not identify a significant instantaneous correlation between the protrusion velocity and Arp3 in any cluster (Fig. 4i, m, q, u, y and C). Instead, the protrusion velocities in 50-100 s and Arp3 in 100-250 s exhibited a significant negative correlation in Cluster II (Fig. 4m and y), which may indicate that the high protrusion velocity at the first peak resulted in a more significant decrease in subsequent Arp2/3 recruitment. This finding is consistent with a previous report showing that a high protrusion velocity can inhibit Arp2/3-dependent nucleation caused by strong membrane tension[29], suggesting that the role of the Arp2/3 complex in supporting sustained protrusion may be limited.

In contrast to actin and Arp2/3, we identified a significant instantaneous correlation between the VASP intensities and protrusion velocities in all clusters in the time-specific correlation analysis (Fig. 4j, n, r, v, z and D). This is consistent with the previous study where the edge velocity and lamellipodial VASP intensity were highly correlated when the leading edges of B16 melonoma cells had uniform rate of protrusion[30]. This result suggests that VASP, not Arp2/3, plays a more direct role in controlling the protrusion velocity at the leading edge in all protrusion phenotypes. In Cluster I and II, VASP-dependent actin elongation likely tightly coordinates with Arp2/3 complex-mediated actin nucleation because actin exhibited a strong instantaneous correlation with protrusion velocity. Conversely, the significant and strong instantaneous correlation between VASP and the protrusion velocity, which starts to appear 50 seconds after protrusion onset (Fig. 4r and D), and the weak correlation between actin and the protrusion velocity in Cluster III (Fig. 4p and B) suggest that actin elongation by VASP plays a more independent role in accelerating protrusion. The previous visual observation that the Arp3 intensity decreased at approximately this time point (Fig. 3s) further supports this notion.

The specificity of the pairwise time correlation was again validated using HaloTag-TMR. Consistent with the previous visual observation (Fig. 3f, k, p and u), HaloTag-TMR significantly and negatively correlated with the protrusion velocity (Fig. 4k, o, s, w, A and E). Interestingly, this correlation was weak in the accelerating protrusion cluster (Fig. 4s and E) because the protrusion velocities were gradually changing. Moreover, multiple vertical regions of correlation were also observed, indicating that cytoplasmic fluorescence can depend on the protrusion velocities at specific times in each cluster.

Notably, both the strong correlation between VASP and the protrusion velocity observed in all clusters and the postulated mode of VASP in regulating accelerating protrusions suggest that VASP, not the Arp2/3 complex, plays a more important role in generating differential protrusion phenotypes. The differences in how VASP and Arp2/3 polymerize actin further validate our interpretation; VASP facilitates actin filament elongation by binding to the barbed ends of actin filaments at the leading edge[31-33], whereas Arp2/3 binds to the sides of the mother filaments and initiates actin nucleation; thus, the ability of Arp2/3 to directly control bared end elongation is limited [34]. Because actin elongation at the barbed end pushes the plasma membrane and generates protrusion velocity, the strong correlation between VASP activity and protrusion velocity at the leading edge is plausible.

The accelerated protrusion phenotype is predicted by VASP dynamics

Taken together, the above data suggest that actin elongation by VASP plays a far more important role in accelerating protrusions (Cluster III) than Arp2/3-mediated actin nucleation. Thus, we next investigated whether VASP recruitment better predicts the accelerating protrusion phenotype than actin or Arp3 (Fig. 1h). First, a visual inspection of the normalized intensities of fluorescence signals of each molecule demonstrated that the temporal patterns of intensity

fluctuation in VASP (Fig. 5c), but not Arp3 (Fig. 5a and b), could distinguish accelerating protrusions (Cluster III) from non-accelerating protrusions (Cluster I/II). Consistently, the t-SNE plot[35], which projected the distance relationship of high-dimensional data of the time series of fluorescence intensities on a 2-dimensional space, demonstrated that the time series of VASP intensities in Cluster III mapped to the localized space, which separated them from Cluster I/II (Fig. 5f). Compared to VASP, the time series of actin and Arp3 intensities of Cluster III were uniformly distributed (Fig. 5d and e).

Furthermore, we applied supervised learning approaches to further validate that VASP is a better predictor of the acceleration phenotype than actin and Arp3. Using support vector machine (SVM), deep neural network (DNN), and random forest (RF), we built classifiers from the normalized intensities of actin, Arp3 and VASP to distinguish the non-accelerating (Clusters I/II) and accelerating (Cluster III) protrusion phenotypes. As a result, classifiers that were trained using VASP intensities could distinguish accelerating protrusions from non-accelerating protrusions with a significantly higher accuracy (Fig. 5h) and MCC (Matthews correlation coefficient) (Fig. 5j) than the classifiers trained using the actin and Arp3 intensities (p-values in Extended Data Table 4a). This difference suggests not only that VASP correlated with protrusion velocity but also that the correlation is sufficiently strong to predict the accelerating protrusion phenotype. Conversely, Arp3 did not provide much information about accelerating protrusions, which is consistent with the idea that VASP predominantly mediates the accelerating protrusions. In contrast, when we performed the same classification analyses using only non-accelerating protrusion phenotypes (Clusters I and II), all proteins have a significant ability to predict Clusters I and II (Fig. 5g and i, p-values in Extended Data Table 4b), suggesting that Arp2/3-mediated actin nucleation assisted by VASP is the mechanism involved in non-accelerating protrusion phenotypes.

Therefore, the predictability assessed by the classification analyses can reveal the differential coordination of actin regulators. This result further highlights the unique role of VASP in the accelerating protrusion, whereas Arp3 was not predictive.

In summary, our framework effectively demonstrated that heterogeneous edge movements could be deconvolved into variable protrusion phenotypes to reveal the underlying differential regulation of actin molecular dynamics, which could not be identified with conventional ensemble average method.

Discussion

We have demonstrated that our computational framework HACKS could effectively deconvolve highly heterogeneous subcellular protrusion activities into distinct protrusion phenotypes and establish an association between each protrusion phenotype and the underlying differential actin regulator dynamics. Although previous studies have examined the spatiotemporal patterning of cell edge dynamics[11,36-38], our study is the first to propose an effective framework to analyze the heterogeneity in protrusion activities at the subcellular level. Using our framework, we have identified the 'fluctuating', 'periodic' and 'accelerating' protrusion phenotypes of distinct temporal patterns in protrusion velocity. Although previous studies also described persistent protrusion based on protrusion distance on a longer time scale[11, 13, 23, 24], our study is the first to further dissect protrusion phenotypes at a fine spatiotemporal scale and quantitatively characterize persistently accelerating protrusions. Intriguingly, accelerating protrusion was later shown to be regulated by differential mechanisms, although they accounted for only a minute portion of entire sampled protrusions. This finding indicates that identifying even a small subset of phenotypes is crucial to fully understand the biology underlying heterogeneous cellular behavior. Notably, other

combinations of clustering methods failed to clearly distinguish the accelerating protrusion phenotype, suggesting that this phenotype would not have been identified without our effective time series clustering.

We were also able to quantitatively measure how the underlying molecular dynamics are coordinated with protrusion phenotypes, thereby revealing the hidden variability of molecular regulatory mechanisms. Elucidating precise differential regulatory mechanisms related to protrusion heterogeneity has been difficult partly because it remains challenging to experimentally perturb a subset of molecules involved with specific subcellular phenotypes *in situ*. To address this challenge, our framework employed highly specific correlation and classification analyses. The result of the correlation analyses provided novel and detailed information about the differential coordination between molecular dynamics and the protrusion phenotype at the subcellular level. The classification analysis further established the association between each protrusion phenotype and the actin regulators by predicting the protrusion phenotypes using actin regulator dynamics. Our results suggested that VASP is a key molecule that drives the variability of lamellipodial protrusion phenotypes. To date, the Arp2/3 complex has been widely accepted as a master organizer of branched actin networks in lamellipodia that acts by nucleating actin[39], whereas VASP has been thought to be a mere elongator of actin filaments or anti-capper of the barbed end[31,32,40,41]. In this study, Arp3 and VASP exhibited similar recruitment dynamics in Clusters I and II, which is consistent with the fact that both Arp2/3 and VASP are known to work in collaboration to regulate protrusion activity. However, the distinct recruitment dynamics identified in the accelerating protrusion phenotype (Cluster III) suggest that Arp2/3-dependent actin nucleation only provides a branched structural foundation for protrusion activity, and VASP-mediated actin elongation subsequently takes over to persistently accelerate protrusions. The

ability of VASP dynamics to predict the accelerating protrusion phenotype further corroborated our finding. Moreover, VASP recruitment was shown to directly correlate with the protrusion velocity, resulting in protrusion variability. Notably, VASP was reported to increase cell protrusion activities[25,26,42] and has been implicated in cancer invasion and migration[25,43,44]. Thus, VASP may regulate the plasticity of protrusion phenotypes, and the functional deregulation of the VASP or its isoforms in cancer may promote cellular migratory behaviors by promoting the accelerated protrusion. Therefore, a mechanical understanding of the connection between short-term protrusion and long-term migration may allow us to understand how distinctive migratory behavior arises among populations and shed light on how a subset of cancer cells acquire metastatic ability.

Our study is the first to successfully dissect the subcellular heterogeneity of dynamic cellular behavior, which requires the deconvolution of temporal patterns in a highly fluctuating dynamic dataset. We do not consider HACKS to be limited to the analysis of subcellular protrusion heterogeneity: it can be expanded to study the morphodynamic heterogeneity of other types of cytoskeletal structures and membrane-bound organelles. Together with the further development of correlation and classification analyses with an increased repertoire of molecular dynamics, we expect our machine learning framework for live cell imaging data to accelerate the mechanistic understanding of heterogeneous cellular and subcellular behaviors.

Supplementary Information is available in the online version of the paper.

Acknowledgements: We thank Seungeun Oh for the critical reading of the manuscript. We thank NVIDIA for providing us with TITAN X GPU cards (NVIDIA Hardware Grant Program) and Microsoft for providing us with Azure cloud computing resources (Microsoft Azure Research Award). This work is supported by the WPI Start-up Fund for new faculty.

Author Contributions C. W. initiated the project, designed the algorithm of the time series clustering, performed the correlation analysis and wrote the final version of the manuscript and supplement. H. C. performed the fluorescence live cell imaging and wrote the final version of the manuscript and supplement; S. K. performed the classification analyses; Y. B. contributed to the writing of the manuscript; K. L. coordinated the study and wrote the final version of the manuscript and supplement. All authors discussed the results of the study.

Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L. (klee@wpi.edu).

References

1 Small, J. V., Stradal, T., Vignal, E. & Rottner, K. The lamellipodium: where motility begins. *Trends in cell biology* **12**, 112-120 (2002).

2 Pankov, R. *et al.* A Rac switch regulates random versus directionally persistent cell migration. *The Journal of cell biology* **170**, 793-802, doi:10.1083/jcb.200503152 (2005).

3 Lauffenburger, D. A. & Horwitz, A. F. Cell migration: a physically integrated molecular process. *Cell* **84**, 359-369 (1996).

4 Guirguis, R., Margulies, I., Taraboletti, G., Schiffmann, E. & Liotta, L. Cytokine-induced pseudopodial protrusion is coupled to tumour cell migration. *Nature* **329**, 261-263, doi:10.1038/329261a0 (1987).

5 Morikawa, Y. *et al.* Actin cytoskeletal remodeling with protrusion formation is essential for heart regeneration in Hippo-deficient mice. *Science signaling* **8**, ra41, doi:10.1126/scisignal.2005781 (2015).

6 Antonello, Z. A., Reiff, T., Ballesta-Illan, E. & Dominguez, M. Robust intestinal homeostasis relies on cellular plasticity in enteroblasts mediated by miR-8-Escargot switch. *The EMBO journal* **34**, 2025-2041, doi:10.15252/embj.201591517 (2015).

7 Liu, Y. H. *et al.* Protrusion-localized STAT3 mRNA promotes metastasis of highly metastatic hepatocellular carcinoma cells in vitro. *Acta pharmacologica Sinica* **37**, 805-813, doi:10.1038/aps.2015.166 (2016).

8 Taniuchi, K., Furihata, M., Hanazaki, K., Saito, M. & Saibara, T. IGF2BP3-mediated translation in cell protrusions promotes cell invasiveness and metastasis of pancreatic cancer. *Oncotarget* **5**, 6832-6845, doi:10.18632/oncotarget.2257 (2014).

9 Ioannou, M. S. *et al.* DENND2B activates Rab13 at the leading edge of migrating cells and promotes metastatic behavior. *The Journal of cell biology* **208**, 629-648, doi:10.1083/jcb.201407068 (2015).

10 Leithner, A. *et al.* Diversified actin protrusions promote environmental exploration but are dispensable for locomotion of leukocytes. *Nat Cell Biol* **18**, 1253-1259, doi:10.1038/ncb3426 (2016).

11 Machacek, M. & Danuser, G. Morphodynamic profiling of protrusion phenotypes. *Biophysical journal* **90**, 1439-1452, doi:10.1529/biophysj.105.070383 (2006).
12 Machacek, M. *et al.* Coordination of Rho GTPase activities during cell protrusion. *Nature* **461**, 99-103, doi:10.1038/nature08242 (2009).
13 Lee, K. *et al.* Functional hierarchy of redundant actin assembly factors revealed by fine-grained registration of intrinsic image fluctuations. *Cell systems* **1**, 37-50, doi:10.1016/j.cels.2015.07.001 (2015).
14 Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559-563, doi:10.1016/j.cell.2010.04.033 (2010).
15 Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226, doi:10.1016/j.cell.2008.09.050 (2008).
16 Slack, M. D., Martinez, E. D., Wu, L. F. & Altschuler, S. J. Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19306-19311, doi:10.1073/pnas.0807038105 (2008).
17 Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197, doi:10.1016/j.cell.2015.05.047 (2015).
18 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).
19 Shafqat-Abbasi, H. *et al.* An analysis toolbox to explore mesenchymal migration heterogeneity reveals adaptive switching between distinct modes. *Elife* **5**, e11384, doi:10.7554/eLife.11384 (2016).
20 Sailem, H., Bousgouni, V., Cooper, S. & Bakal, C. Cross-talk between Rho and Rac GTPases drives deterministic exploration of cellular shape space and morphological heterogeneity. *Open Biol* **4**, 130132, doi:10.1098/rsob.130132 (2014).
21 Huang, N. E. *et al.* The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **454** (1998).
22 Keogh, E., Lin, J. & Fu, A. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. *In Proc. of the 5th IEEE International Conference on Data Mining* 226 - 233 (2005).
23 Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496, doi:10.1126/science.1242072 (2014).
24 Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. *AAAI Workshop on Knowledge Discovery in Databases*, 229-248 (1994).
25 Bae, Y. H. *et al.* Profilin1 regulates PI(3,4)P2 and lamellipodin accumulation at the leading edge thus influencing motility of MDA-MB-231 cells. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21547-21552, doi:10.1073/pnas.1002309107 (2010).
26 Barnhart, E. L., Allard, J., Lou, S. S., Theriot, J. A. & Mogilner, A. Adhesion-Dependent Wave Generation in Crawling Cells. *Curr Biol* **27**, 27-38, doi:10.1016/j.cub.2016.11.011 (2017).
27 Los, G. V. *et al.* HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS Chem Biol* **3**, 373-382, doi:10.1021/cb800025k (2008).
28 Ji, L., Lim, J. & Danuser, G. Fluctuations of intracellular forces during cell protrusion. *Nat Cell Biol* **10**, 1393-1400, doi:10.1038/ncb1797 (2008).

29 Houk, A. R. *et al.* Membrane tension maintains cell polarity by confining signals to the leading edge during neutrophil migration. *Cell* **148**, 175-188, doi:10.1016/j.cell.2011.10.050 (2012).

30 Rottner, K., Behrendt, B., Small, J. V. & Wehland, J. VASP dynamics during lamellipodia protrusion. *Nat Cell Biol* **1**, 321-322, doi:10.1038/13040 (1999).

31 Barzik, M. *et al.* Ena/VASP proteins enhance actin polymerization in the presence of barbed end capping proteins. *J Biol Chem* **280**, 28653-28662, doi:10.1074/jbc.M503957200 (2005).

32 Breitsprecher, D. *et al.* Clustering of VASP actively drives processive, WH2 domain-mediated actin filament elongation. *The EMBO journal* **27**, 2943-2954, doi:10.1038/emboj.2008.211 (2008).

33 Hansen, S. D. & Mullins, R. D. Lamellipodin promotes actin assembly by clustering Ena/VASP proteins and tethering them to actin filaments. *Elife* **4**, doi:10.7554/eLife.06585 (2015).

34 Machesky, L. M. *et al.* Scar, a WASp-related protein, activates nucleation of actin filaments by the Arp2/3 complex. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 3739-3744 (1999).

35 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J Mach Learn Res* **9**, 2579-2605 (2008).

36 Martin, K. *et al.* Spatio-temporal co-ordination of RhoA, Rac1 and Cdc42 activation during prototypical edge protrusion and retraction dynamics. *Sci Rep* **6**, 21901, doi:10.1038/srep21901 (2016).

37 Verkhovsky, A. B. The mechanisms of spatial and temporal patterning of cell-edge dynamics. *Curr Opin Cell Biol* **36**, 113-121, doi:10.1016/j.ceb.2015.09.001 (2015).

38 Dobereiner, H. G. *et al.* Lateral membrane waves constitute a universal dynamic pattern of motile cells. *Phys Rev Lett* **97**, 038102, doi:10.1103/PhysRevLett.97.038102 (2006).

39 Pollard, T. D. & Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112**, 453-465 (2003).

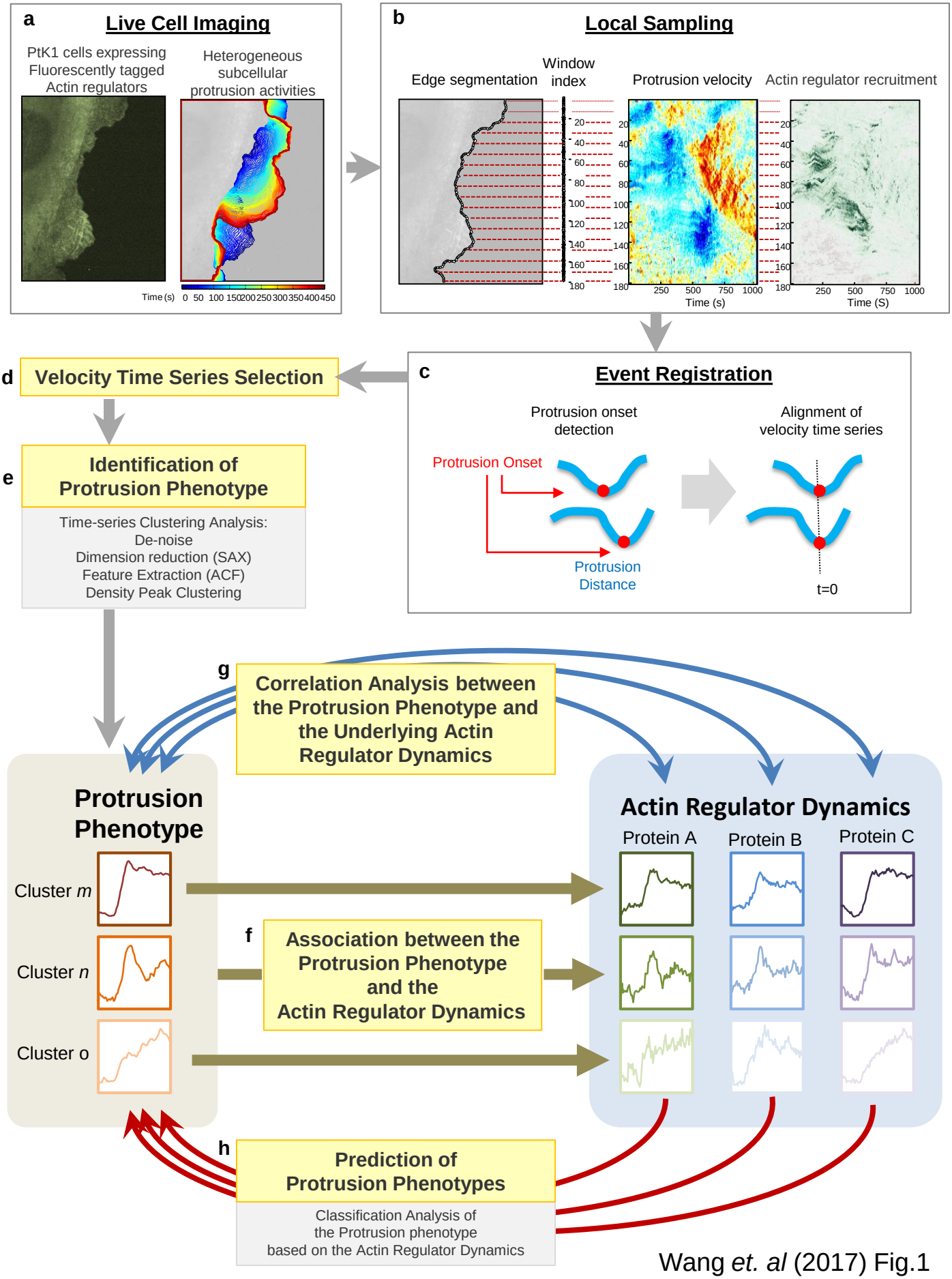
40 Hansen, S. D. & Mullins, R. D. VASP is a processive actin polymerase that requires monomeric actin for barbed end association. *The Journal of cell biology* **191**, 571-584, doi:10.1083/jcb.201003014 (2010).

41 Rotty, J. D. *et al.* Profilin-1 serves as a gatekeeper for actin assembly by Arp2/3-dependent and -independent pathways. *Dev Cell* **32**, 54-67, doi:10.1016/j.devcel.2014.10.026 (2015).

42 Lacayo, C. I. *et al.* Emergence of large-scale cell morphology and movement from local actin filament growth dynamics. *PLoS Biol* **5**, e233, doi:10.1371/journal.pbio.0050233 (2007).

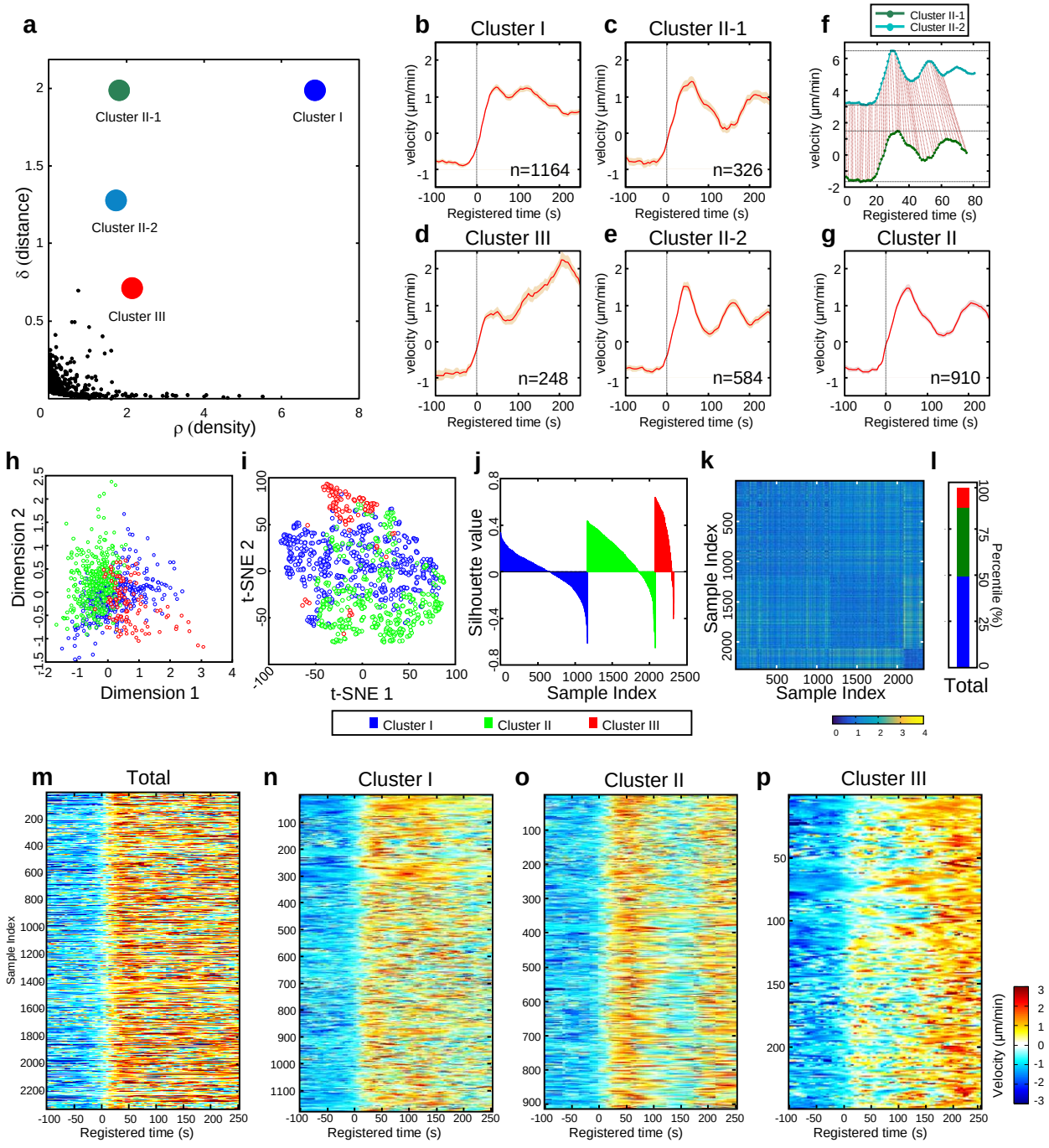
43 Carmona, G. *et al.* Lamellipodin promotes invasive 3D cancer cell migration via regulated interactions with Ena/VASP and SCAR/WAVE. *Oncogene* **35**, 5155-5169, doi:10.1038/onc.2016.47 (2016).

44 Philippar, U. *et al.* A Mena invasion isoform potentiates EGF-induced carcinoma cell invasion and metastasis. *Dev Cell* **15**, 813-828, doi:10.1016/j.devcel.2008.09.003 (2008).



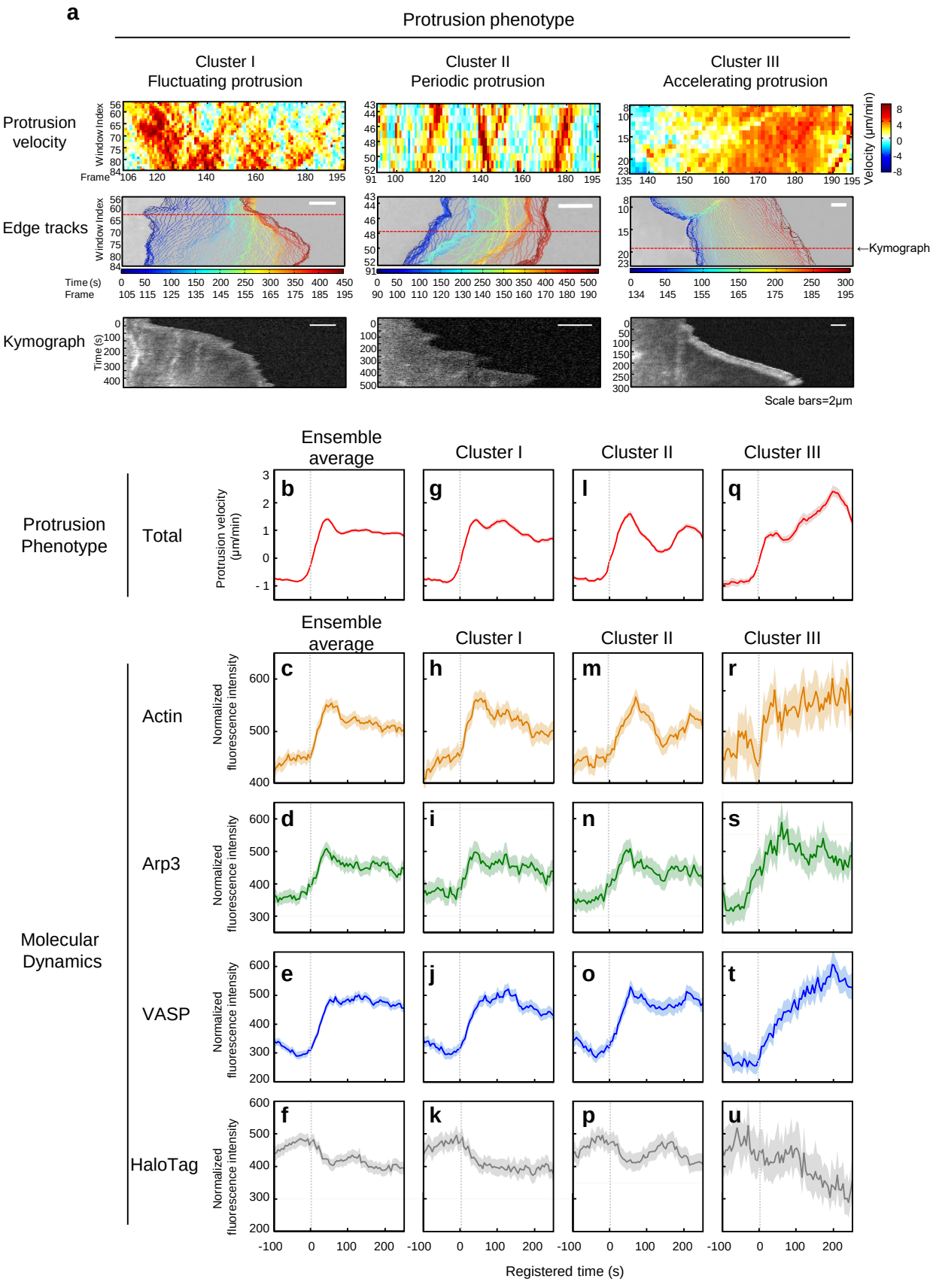
Wang *et. al* (2017) Fig.1

Figure 1. Schematic Representation of Analytical Steps of HACKS (a) Fluorescence time-lapse movies of the leading edges of a migrating PtK1 cell expressing Arp3-HaloTag were taken at 5 seconds/frame. **(b)** Probing windows (500 by 500 nm) were generated to track the cell edge movement and quantify protrusion velocities and fluorescence intensities. **(c)** The protrusion distance was registered with respect to protrusion onsets (t = 0). Time series of protrusion velocities were then aligned. **(d)** Velocity time series lasting longer than 250 seconds were selected. **(e)** The protrusion phenotypes were identified with a time series clustering analysis. **(f)** Association between the protrusion phenotypes and the actin regulator dynamics. **(g)** Correlation analysis between time series of the protrusion velocities and fluorescence intensities. **(h)** The classification analysis of fluorescence time series for protrusion phenotypes.



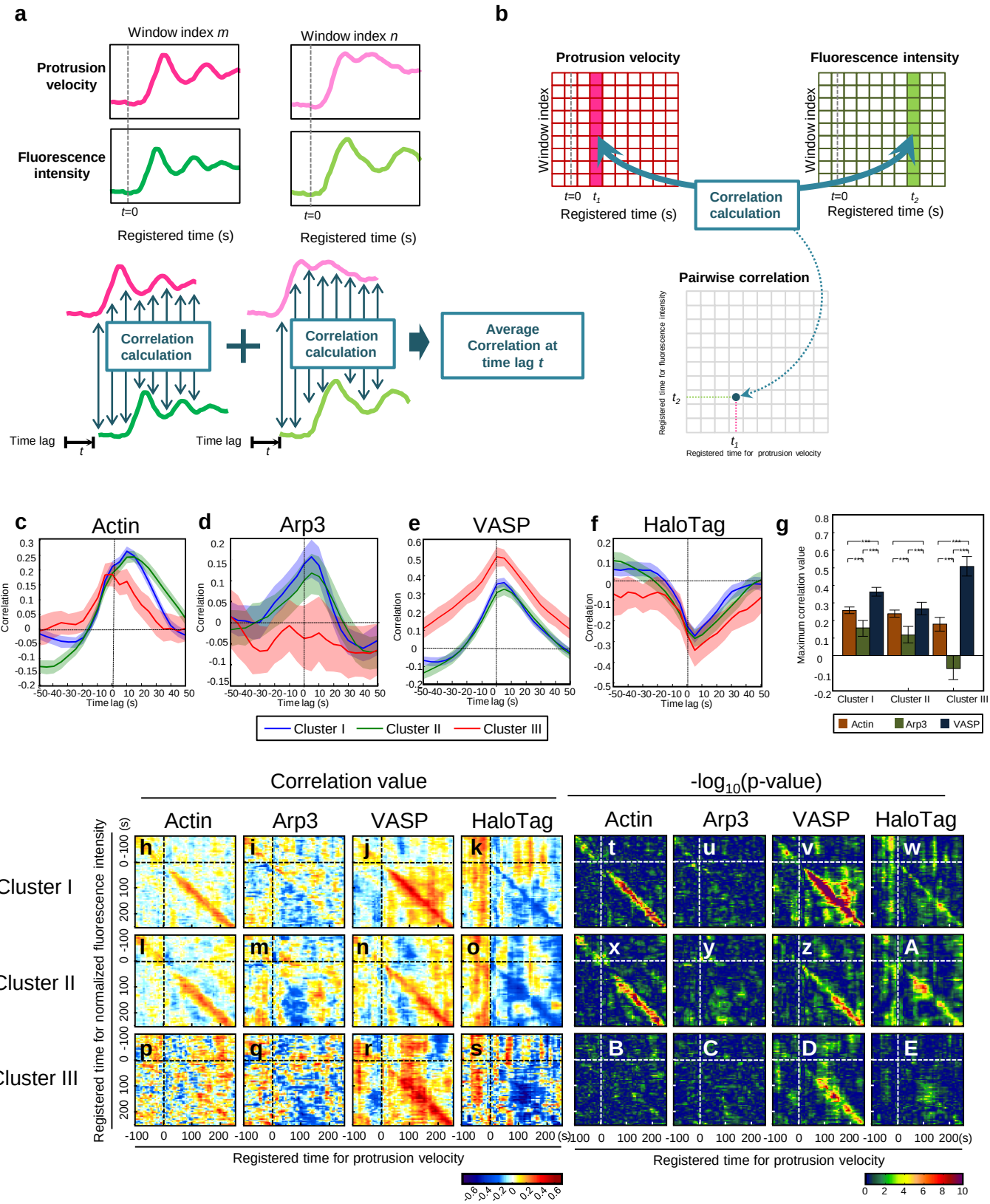
Wang *et. al* (2017) Fig.2

Figure 2. Subcellular Protrusion Phenotypes Revealed by a Time Series Clustering Analysis (a) Decision graph of the Density Peak clustering analysis of protrusion velocities. **(b-e)** Average time series of protrusion velocity registered at protrusion onsets (t=0) in each cluster. **(f)** Alignment of average velocity profiles of Cluster II-1 and Cluster II-2. **(g)** Final cluster result from the Dynamic Time Warping with Clusters II-1 and II-2. Solid lines indicate population averages. Shaded error bands about the population averages indicate 95% confidence intervals computed by bootstrap sampling. n is the number of sampled velocity time series pooled from multiple cells. **(h)** Multidimensional scaling plot of the clusters. **(i)** t-SNE plot of the clusters. **(j)** Silhouette plot of the clusters. **(k)** Distance map of each velocity time series after the clustering. **(l)** Proportions of each cluster. **(m-p)** Raw velocity maps for whole samples pooled from multiple cells (m), Cluster I (n) Cluster II (o) Cluster III (p).



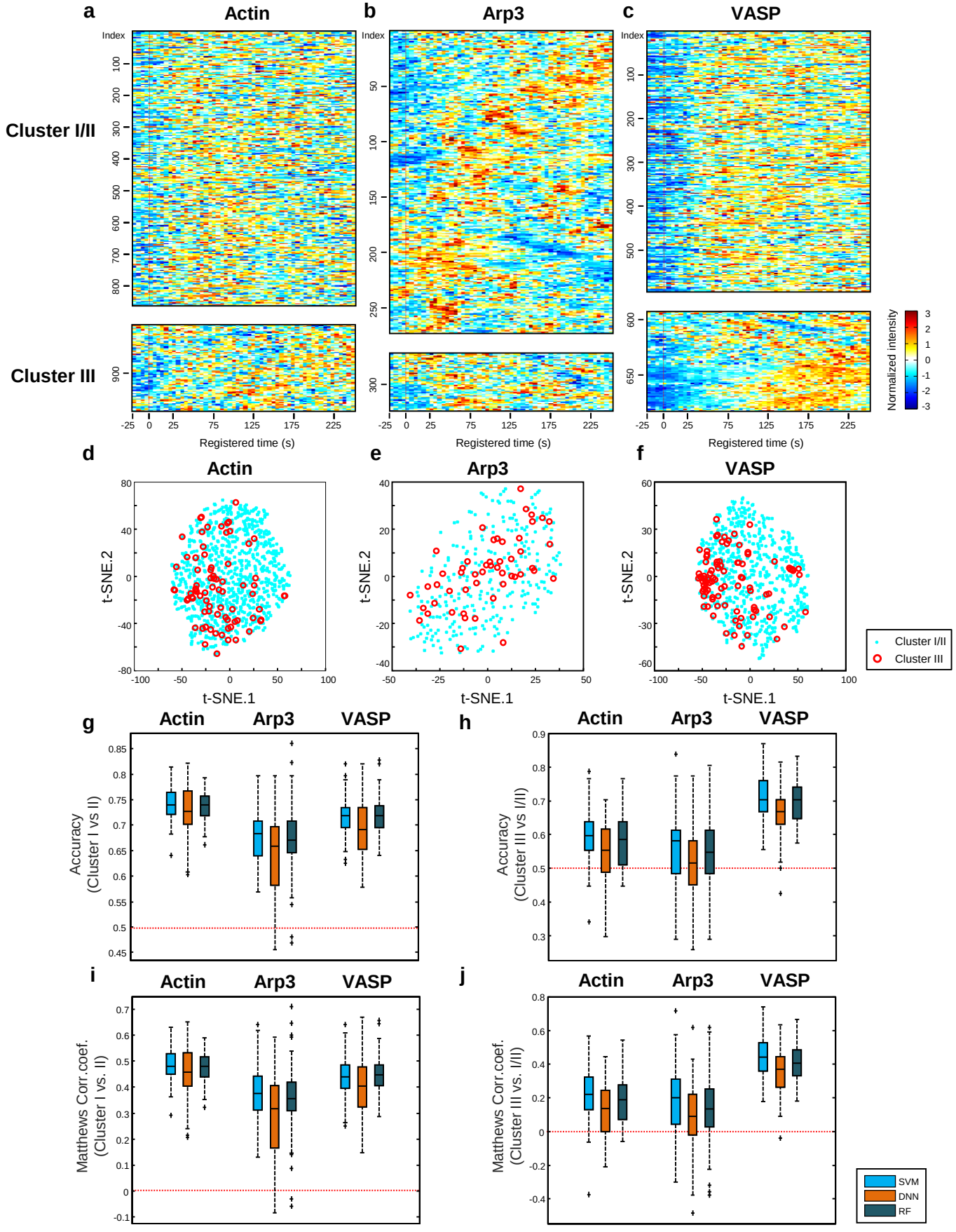
Wang *et. al* (2017) Fig.3

Figure 3. Identified Distinctive Subcellular Protrusion Phenotypes and Their Differential Underlying Molecular Dynamics (a) Representative examples of each cluster with raw velocity maps, edge evolution, and edge kymographs (Cluster I,III: HaloTag-TMR-VASP, Cluster II: HaloTag-TMR). (**b-u)** Protrusion velocity and normalized fluorescence intensity time series of registered with respect to protrusion onset for ensemble averages of entire samples (b-f), Cluster I (g-k), Cluster II (l-p) and Cluster III (q-u). Solid lines indicate population averages. Shaded error bands about the population averages indicate 95% confidence intervals computed by bootstrap sampling.



Wang *et. al* (2017) Fig.4

Figure 4. Cluster-Specific Correlation between Protrusion Velocity and Actin Regulator Dynamics (a-b) Schematic diagrams of time lag (a) and time-specific correlation analysis (b). **(c-f)** Pearson's cross-correlation of edge velocity and actin (c), Arp2/3 (d), VASP (e), and HaloTag (f) as a function of the time lag between the two time series. Solid lines indicate population averages. Shaded error bands about the population averages indicate 95% confidence intervals computed by bootstrap sampling. **(g)** Comparison and statistical testing of maximum correlation coefficients from (c-e) in each cluster. *** indicates the statistical significance by two-sample Kolmogorov-Smirnov (KS) test. The p-values are listed in Extended Data Table 3. **(h-E)** Pairwise Pearson's correlation coefficients (h-s) and their p-values ($-\log_{10}$ (p value)) (t-E) of protrusion velocity and fluorescence intensity time series registered relative to protrusion onset.



Wang *et. al* (2017) Fig.5

Figure 5. Prediction of Protrusion Phenotypes based on Actin Regulator Dynamics **(a-c)** Normalized fluorescence time series of actin (a), Arp2/3 (b) and VASP (c) in each protrusion cluster. Time series data were randomly sampled to better visualize the fluorescence time series for each protein. Full time series are presented in Extended Data Fig. 5. **(d-f)** t-SNE plots of the normalized fluorescence time series of actin (d), Arp2/3 (e) and VASP (f). Red dots indicate the time series from the accelerating protrusion cluster (Cluster III). **(g-j)** Accuracy (g) and Matthews correlation coefficients (i) of the classification analysis of Cluster I against Cluster II. Accuracy (h) and Matthews correlation coefficients (j) of the classification analysis of Cluster III against Clusters I and II.

a

	Global		Cluster I		Cluster II		Cluster III	
	# video	# window	# video	# window	# video	# window	# video	# window
Velocity	33	2322	33	1164	33	910	28	248
Actin	10	934	10	454	10	403	9	77
Arp3	8	323	8	141	8	131	8	61
VASP	9	682	9	381	9	212	6	89
Halo	6	383	6	188	6	164	5	31

b

	Cluster II-1		Cluster II-2	
	# video	# window	# video	# window
Velocity	31	326	33	584
Actin	9	143	10	260
Arp3	8	46	8	85
VASP	9	88	9	124
Halo	5	49	6	115

Extended Data Table 1. Number of Cells and Probed Windows Used in the Time Series Clustering Analysis. (a) The number of cell videos and windows from which protrusion segments were sampled. **(b)** The number of cell videos and windows assigned for Cluster II-1 and II-2.

Method Combination				Persistent Pattern	Oscillating Pattern
DV	SAX	Distance	Clustering method		
Yes	No	ED	DP	No	No
Yes	No	ACF	DP	Yes	No
Yes	Yes	app_ED	DP	No	No
Yes	Yes	ACF	Kmeans	No	Yes
Yes	Yes	ACF	DP	Yes	Yes

DV: de-noised velocity
ED: Euclidean Distance
app_ED: approximate Euclidean Distance in SAX
ACF: Autocorrelation
SAX: Symbolic Aggregate Approximation
Kmeans: K-nearest mean method
DP: Density Peaks

Extended Data Table 2. Summary of Different Combinations of Algorithms Tested

	Cluster I			Cluster II			Cluster III		
Protein pair	actin-Arp3	actin-VASP	Arp3-VASP	actin-Arp3	actin-VASP	Arp3-VASP	actin-Arp3	actin-VASP	Arp3-VASP
P-value	1.16×10^{-6}	2.60×10^{-11}	4.62×10^{-15}	1.95×10^{-7}	0.16	5.76×10^{-7}	3.08×10^{-10}	8.10×10^{-17}	5.04×10^{-26}

Extended Data Table 3. Statistical Analyses of the Maximum Correlation Coefficients in Figure 4g. All p-values were calculated by two-sample Kolmogorov-Smirnov (KS) test.

a

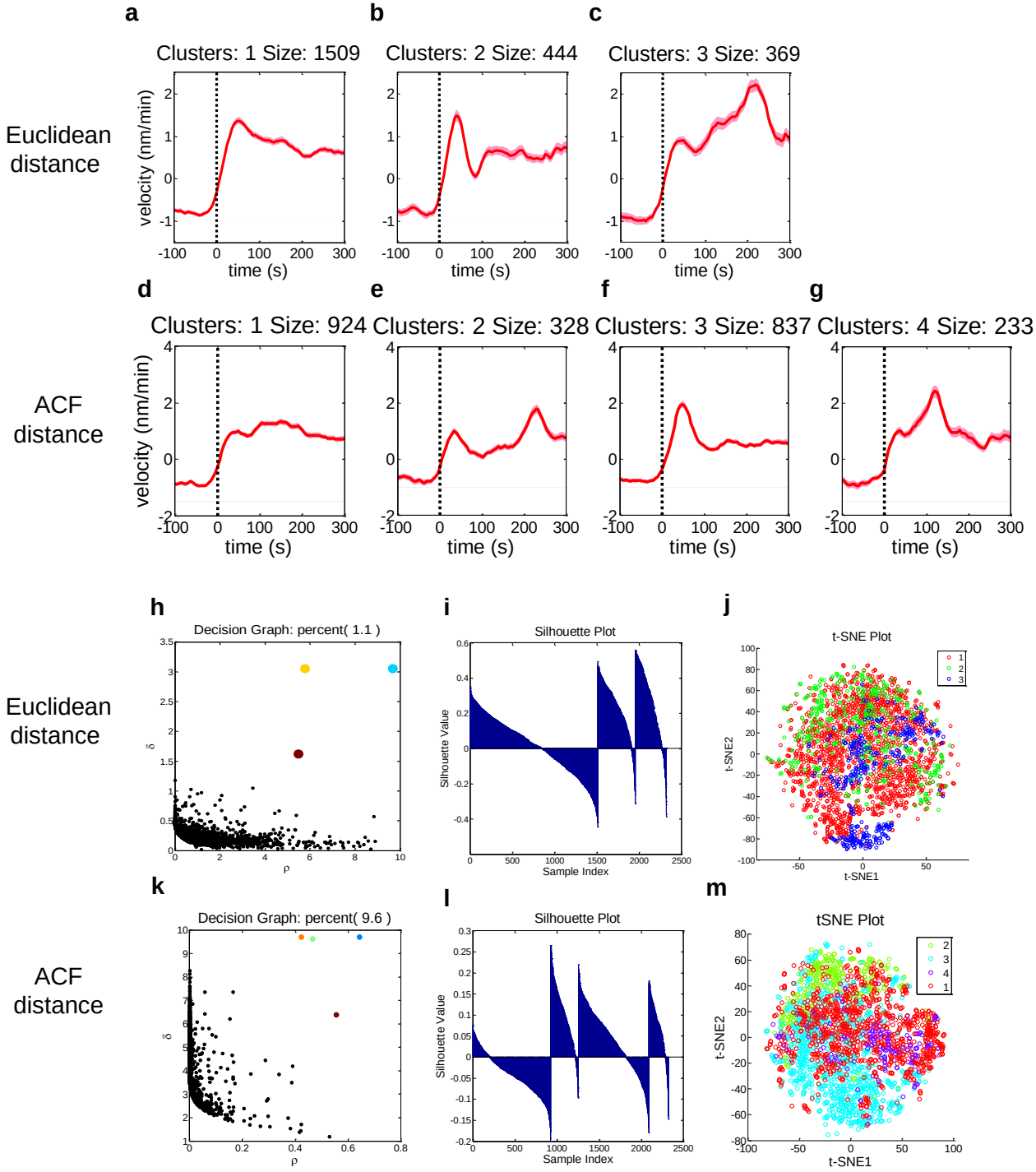
	SVM			DNN			RF		
Protein pair	actin-Arp3	actin-VASP	Arp3-VASP	actin-Arp3	actin-VASP	Arp3-VASP	actin-Arp3	actin-VASP	Arp3-VASP
Accuracy	9.12×10^{-09}	3.96×10^{-16}	8.44×10^{-26}	2.85×10^{-06}	2.95×10^{-11}	8.24×10^{-24}	9.25×10^{-05}	4.52×10^{-14}	7.17×10^{-28}
MCC	4.81×10^{-05}	3.70×10^{-12}	1.12×10^{-20}	1.74×10^{-04}	1.06×10^{-11}	1.75×10^{-19}	1.00×10^{-03}	4.26×10^{-13}	1.12×10^{-20}

b

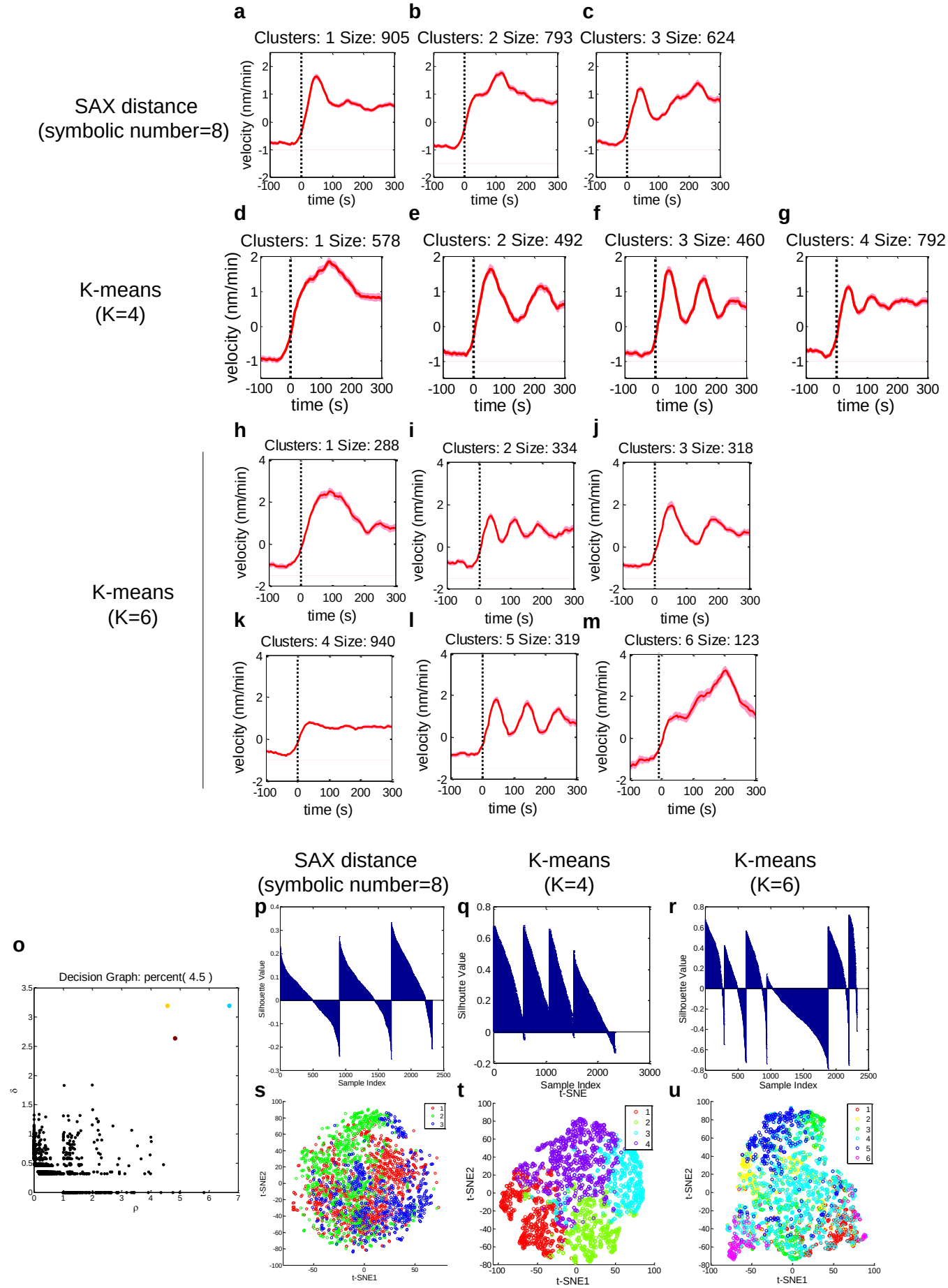
	SVM			DNN			RF		
Protein	actin	Arp3	VASP	actin	Arp3	VASP	actin	Arp3	VASP
Accuracy	1.44×10^{-34}	1.15×10^{-16}	5.00×10^{-03}	3.59×10^{-27}	1.15×10^{-16}	5.00×10^{-03}	1.44×10^{-34}	2.77×10^{-21}	2.21×10^{-08}
MCC	9.30×10^{-31}	4.52×10^{-14}	6.91×10^{-02}	3.97×10^{-25}	8.08×10^{-11}	4.7×10^{-02}	5.06×10^{-30}	2.49×10^{-18}	2.45×10^{-05}

SVM: Support Vector Machine
DNN: Deep Neural Network
RF: Random Forest
MCC: Matthews Correlation Coefficient

Extended Data Table 4. Statistical Analyses of Accuracy and Matthews Correlation Coefficients of the Classification Analysis in Figure 5. (a) The p-values of the differences of accuracy and Matthews correlation coefficients between each protein pair for the classification analysis of Cluster III against Clusters I/II (Fig. 5h and j). **(b)** The p-values of the differences of accuracy and Matthews correlation coefficients of each protein between the classification analyses of Cluster II against Clusters I (Fig. 5g and i) and Cluster III against Clusters I/II (Fig. 5h and j). All p-values were calculated by two-sample Kolmogorov-Smirnov (KS) test.

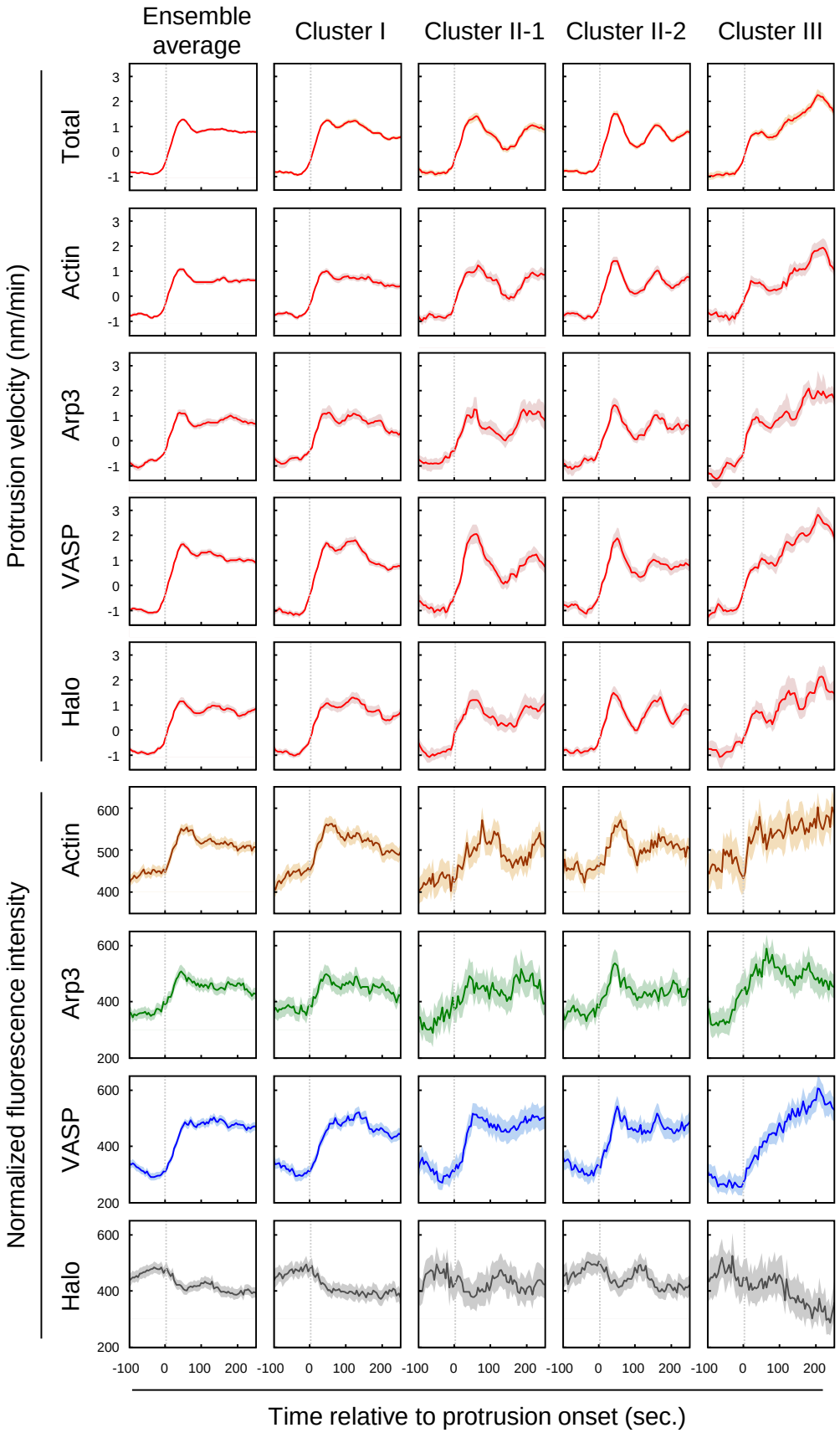


Extended Data Figure 1. Comparison of Time Series Clustering Results between the Euclidean Distance and ACF Distance (a-c) Clustering result after employing ACF distance and **(d-g)** Euclidean distance. **(h)** Decision graph after employing ACF distance and the validation of the clustering result using a **(i)** silhouette plot and **(j)** t-SNE plot. **(k)** Decision graph after employing Euclidean distance and validation of the clustering result using a **(l)** silhouette plot and **(m)** t-SNE plot.

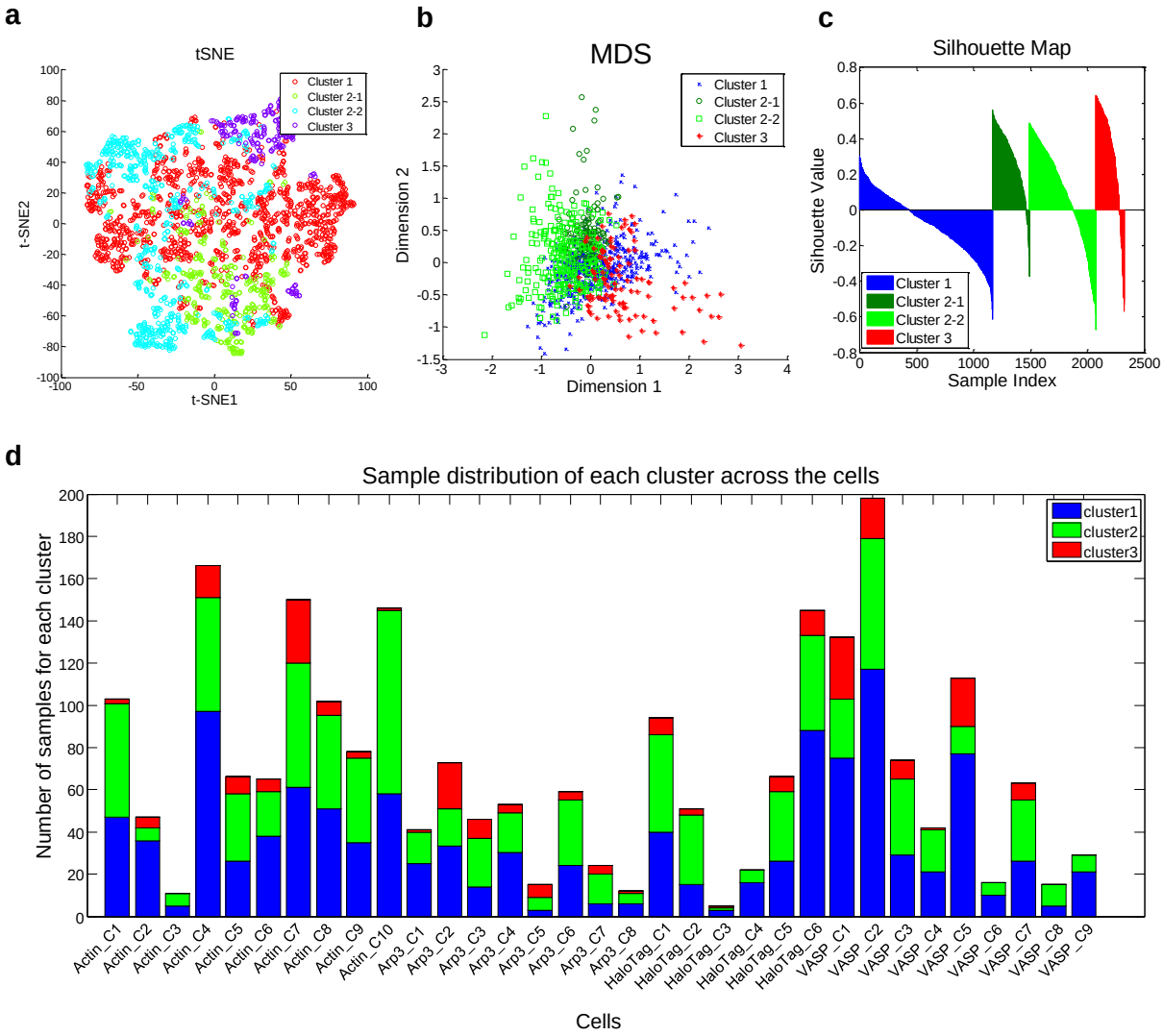


Wang *et. al* (2017)) Extended Data Fig. 2

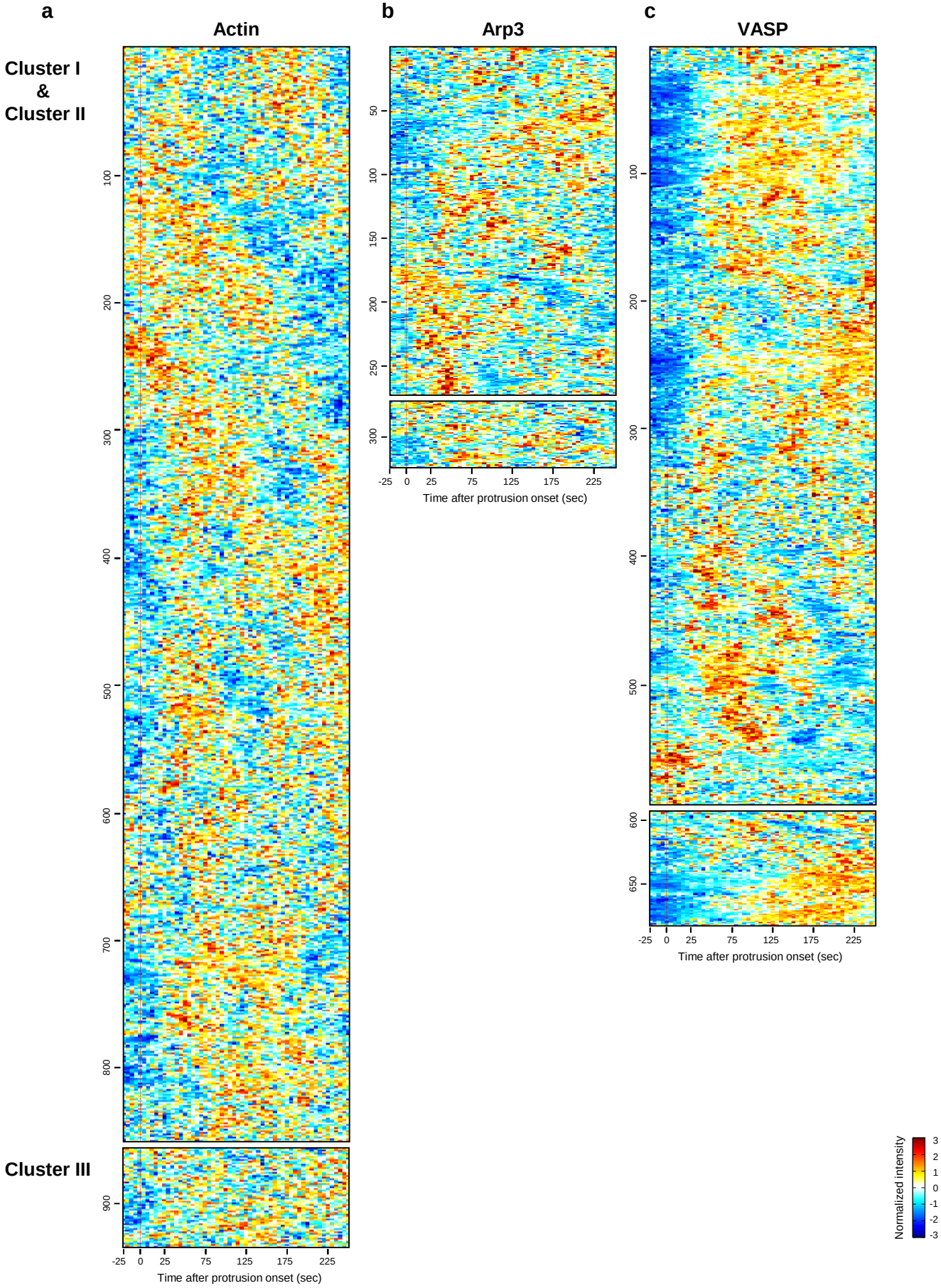
Extended Data Figure 2. Comparison of Time Series Clustering Results between SAX Distance and K-means (a-c) Clustering result after employing SAX distance. 8 was selected as a symbolic number because 4 did not produce useful results. **(d-g)** Clustering result after employing K-means (K=4) **(h-m)** Clustering result after employing K-means (K=6). **(o)** Density plot. **(p-r)** Silhouette plot for the clustering result after employing SAX distance (symbolic number=8) (p), K-means (K=4) (q) or K-means (K=8) (r) **(s-u)** t-SNE plot for the clustering result after employing SAX distance (symbolic number=8) (s), K-means (K=4) (t) or K-means (K=8) (u).



Extended Data Figure 3. Velocity Profiles and Fluorescence Intensity Profiles of Actin, Arp3, VASP and HaloTag in Each Cluster before Merging Clusters II-1 and II-2.



Extended Data Figure 4. Validation of Clustering Result and Sample Distribution of Clusters. (a) t-SNE plot and **(b)** MDS and **(c)** Silhouette map for the validation of the clustering result before merging Cluster II-1 and Cluster II-2 **(d)** Sample distribution of clusters for individual monitored cells.



Extended Data Figure 5. Normalized Fluorescence Time Series of Actin, Arp3 and VASP

SUPPLEMENTARY TEXT

Evaluating the role of each component of our time series clustering.

Our time series clustering primarily consisted of SAX (dimensional reduction), ACF (dissimilarity measure) and Density Peak clustering. To identify the role of each component in our algorithm, we systematically tested the performance of the clustering results by replacing each component with other methods. First, without SAX for dimensional reduction, two different distance dissimilarity measures, ACF and Euclidean distances, were tested. When the ACF distance was used without SAX, we were still able to identify accelerated protrusions (Extended Data Fig. 1c) from the noisy samples, but the periodic protrusion pattern was severely degraded (Extended Data Fig. 1b). Without the dimension reduction provided by SAX, the temporal fluctuation of the edge velocity influenced the calculation of autocorrelation coefficients in the periodic velocity patterns. Therefore, SAX effectively reduced the local influence of data fluctuation and allowed general patterns to be easily distinguished. When Euclidean distance was used without SAX, we were unable to identify any periodic or accelerated protrusions (Extended Data Fig. 1d-g), despite the fact that the Density Peak clustering clearly suggests four distinct clusters (Extended Data Fig. 1k). Furthermore, when we tested the clustering using SAX and its Euclidean distance (the lower bound of Euclidean distance in SAX), clusters similar to those identified using Euclidean distance were found, as shown in Extended Data Fig. 2a-c. From these analyses, we concluded that the ACF distance is the important factor for extracting the periodic and accelerating protrusion patterns. Although SAX reduced the local fluctuation of the data, ACF played a major role in revealing the distinct patterns out of noisy datasets.

To identify the role of Density Peak clustering, we replaced it with conventional K-means. When the number of clusters, K, was set to 4, only periodic clusters with different time intervals, not accelerating protrusions, were identified (Extended Data Fig. 2d-g). When K was increased to 6 (Extended Data Fig. 2m), we were able to identify accelerating protrusions (Extended Data Fig. 2h-m). However, the

drawback of K-means is that the optimal number of clusters, K, is not easily determined. To delineate both periodic and accelerating protrusions, a large K should be used, resulting in small numbers of samples in each cluster. Density Peaks not only provides information to determine the number of clusters but also effectively discovers both periodic and accelerating protrusions with a relatively small number of clusters.

As summarized in Extended Data Table 2, the combination of SAX (dimensional reduction), ACF (dissimilarity measure) and Density Peaks clustering allowed us to identify distinct protrusion clusters within the highly fluctuating velocity data.

SUPPLEMENTARY MOVIE LEGENDS

Supplementary Movie 1. Representative Movie of HaloTag-TMR-VASP in a PtK1 cell for Cluster I. Scale bar: 5 μm. Original movie captured by spinning disc confocal microscopy at a 5s frame interval; replay at 30 frames/second.

Supplementary Movie 2. Representative Movie of HaloTag-TMR in a PtK1 cell for Cluster II. Scale bar: 5 μm. Original movie captured by spinning disc confocal microscopy at a 5s frame interval; replay at 30 frames/second.

Supplementary Movie 3. Representative Movie of HaloTag-TMR-VASP in a PtK1 cell for Cluster III. Scale bar: 5 μm. Original movie captured by spinning disc confocal microscopy at a 5s frame interval; replay at 30 frames/second.

METHODS

Local sampling and event registration.

Using a custom-built software package written in Matlab (Mathworks, MA, USA), we performed the following computational procedures. The software is explained in detail elsewhere[1]. The threshold-based method was used to segment cell edges in the fluorescence images, and the cell edge velocity was calculated by tracking the cell edges using a mechanical model (Machacek et al, 2006). The software generated probing windows whose initial size was 500 nm by 500 nm along the cell boundary to locally sample the protrusion velocity and fluorescence intensity. The number of probing windows then maintained constant throughout the movie. The local protrusion velocity and fluorescence intensity were quantified by averaging the values within probing windows. By repeating this procedure in each frame of the time-lapse movies, we acquired the time series of protrusion velocities and fluorescence intensities.

We then identified significant protrusion events on a per-window basis. To reduce the effects of random fluctuations in the protrusion velocity time series, we obtained an edge displacement time series for a particular window by integrating the protrusion velocity over time. The noise of the time series was removed with a smoothing spline filter using the Matlab function *csaps*() and a smoothing parameter of 0.01. Small protrusion and retraction events considered insignificant in terms of the overall cell edge movement were then further eliminated as follows. First, we identified local maxima/minima (protrusion/retraction onsets) at the edge displacement time series using the Matlab function *findpeaks*() and calculated the net protrusion/retraction distances for each event. A previous study using the same PtK1 cells showed that the distribution of distances can be decomposed into two exponential distributions, indicating small fluctuations and large movement during protrusion and retraction events[1]. Thus, small events whose protrusion distances were less than 720 nm (10 pixels in length) were discarded from the analysis. In addition, we eliminated short-term switches between the protrusion and retraction

phases within 50 seconds. After these insignificant events were removed[1], the remaining protrusion onsets were used for event registration.

The protrusion velocity and fluorescence intensities over time in individual windows were registered by aligning the protrusion onset at t = 0. After the registration, the negative time indicates the retraction phase, and the positive time indicates the protrusion phase. Time series in negative time were limited by the preceding retraction onset, and time series in positive time were limited by the subsequent retraction onset.

Treatment of missing values.

Because of image noise, the software can produce abnormal data in a rare case. In this case, values could be missing from a time series, and the following strategy was applied to treat these missing values: For each edge velocity or fluorescence intensity time sample, the entire time series was discarded if the length of continuous missing values is longer than a threshold (here, we used the value 8). Otherwise, the average of four values before and after the missing value was used to individually estimate the value for this location.

De-noising the distance samples by Empirical Mode Decomposition-De-trended Fluctuation Analysis (EMD-DFA).

For each registered sample, the edge displacement was calculated from the edge velocity using the Matlab function *trapz()*. Empirical Mode Decomposition (EMD)[2] was then applied to the transformed protrusion edge displacement to remove noise. Finally, the denoised velocity was calculated from the denoised displacement using the Matlab function *diff()*.

Cell movement is highly non-stationary, and Empirical Mode Decomposition (EMD)[2] is a local and data-driven de-noising method to decompose non-stationary signals into a series of intrinsic components. The general procedure of EMD can be described as follows:

1) Identify all the extremes (minima and maxima) of sample *d(t)*;
2) Connect the local maximum points and local minimum points separately using an interpolation method to generate the envelope, *e(t)*;
3) Compute the average of envelopes, $avg(t) = (e_{min}(t) + e_{max}(t))/2$;
4) Eliminate the average signal of the envelope from the sample d(t) to obtain the residual: $m(t) = d(t) - avg(t)$; and
5) Iterate from steps 1) to 4) on the residual $m(t)$ until the $avg(t)$ equals zero.

After EMD, the original signals can be decomposed into intrinsic mode functions (IMF) without any loss of information, and the trend is called the residue. For each component, a de-trended fluctuation analysis was used to measure the self-affinity as the fractal scaling index (alpha), which estimates the fractal-like autocorrelation properties. The value of alpha inversely correlated with the possibility that the component originated from noise. In our procedure, the code was obtained from a previous publication[3], and the value of alpha was empirically set to 0.33 to balance the maintaining information and trimming noise.

Determining the time interval for the clustering analysis.

The duration of cell protrusion is heterogeneous, and some protrusion events were not completely recorded because of the finite length of the movies. Our clustering analysis focuses on the equal length of time series data. Therefore, time series shorter than a certain threshold were discarded from the analysis. The number of samples used inversely correlates with the information that remains in each sample because the threshold should be smaller. The best scenario is that the threshold will be selected as the optimized solution of maximizing the multiplication of these two factors. After optimizing these two factors with equal weight, the best threshold is approximately 50. After this step, more than 60% of samples remain for further analysis. The exact threshold, which is near 50, is finally decided based on convenience for further analysis. Considering the ambiguity of detected protrusion onsets, 5 frames before the protrusion onset were also included for the analysis. Therefore, the time series for the analysis

consisted of 56 frames, which included the previous 5 frames before the protrusion onset and 51 frames after the protrusion onset.

Symbolic Aggregate approximation (SAX) representation.

The dimensionality of times series samples consisting of 56 frames remained high, and we were interested in the general trend over large time scales. Therefore, the dimensionality of time series data should be reduced. To this end, we applied SAX (Symbolic Aggregate approXimation)[4] to our time series dataset to reduce dimensionality and discretize the data. The general procedure of SAX is summarized as follows:

1) Manually determine the reduced dimension, N, and the symbolic number, M (the number of discretization levels)
2) For each normalized sample, the time series data over the entire time range are pooled together and fitted to a Gaussian distribution. The entire time series is divided into M ranges with equal probabilities of the fitted Gaussian distribution. In each range, one symbol is assigned accordingly.
3) Subsequently, the time series is divided into N intervals along over time. The average value is calculated in each interval and combined to represent the raw time series data.
4) Finally, the average value in each interval is represented by the symbol defined in step 2.
5) Iterate from step 2) and 3) until all samples are represented.

After the SAX representation, all time series data are reduced into low-dimensional (N) symbolic series data , which are used in further analyses. In the current experiment, M and N were both empirically set to 4. Here, 4 symbols that range from 0 to 3 will be used to calculate the autocorrelation coefficients. In addition, the average process in SAX also removes local noise.

Dissimilarity Calculation.

To measure the dissimilarity of two time series, the original description of SAX representation proposed an approximate Euclidean distance of SAX as a dissimilarity measure. Instead, we used the dissimilarity measure based on estimated autocorrelation functions (ACF) [5]. We compared the clustering performance using different dissimilarity measures in the Extended Data Fig. 1. First, the estimated autocorrelation vector was calculated and the square Euclidean distance between the autocorrelation coefficients was then used to measure the dissimilarity of two samples as follows:

$$\mathrm{d}_{\mathrm{ACF}}(X, Y) = \sum_{i=1}^{L} (X_i - Y_i)^2$$

In our implementation, the ACF distance was calculated using the *TSdist* R package[6].

Density Peak Clustering.

After we calculated the pairwise dissimilarity of the time series, we performed a clustering analysis using the Density Peak clustering algorithm[7]. Here, cluster centers are the samples with local density maxima and are far away from other clustering centers. The number of cluster centers can be easily identified using the density map generated by the algorithm, where the x-axis denotes the local density, and the y-axis shows the minimum distance from another sample with higher density. The samples located at the upper-right region of the density map can be treated as the cluster centers, and the number of samples in that region can be estimated to be the cluster number, K. Based on the local density value of each sample, a hierarchical tree structure is built by linking one sample to the sample with a higher density value and the lowest distance. After the number of clusters, K, is determined by visual inspection, the hierarchical tree can then be divided into several disconnected sub-trees as the clusters. In our implementation, the density around each sample was determined by calculating sum of distances with the Gaussian Kernel of the manually selected radius as follows:

$$\rho(\mathrm{S_i}) = \sum_{k=1, i \neq k}^{N} e^{(-\frac{d(S_i, S_k)}{dc})^2}$$

In this calculation, dc is manually defined, and the number of clusters is also manually selected by visualizing the density-distance map. The details of implementation can be found in the supplementary materials of a previous paper[7].

Validation of clustering results.

We used the following methods to validate our clustering results.

- Ordered Dissimilarity Map: After clustering, the distances between samples within the same clusters should be smaller than those between samples in different clusters. Therefore, after the samples are grouped together by cluster indices and ordered by distance, the distance map can be visualized as blocks along the diagonal. In addition, because the input of the Density Peak clustering method is only a distance matrix, the ordered dissimilarity map will be suitable to evaluate clustering results to show several blocks along the diagonal.
- MDS: Classical multidimensional scaling (MDS)[8] is a method to visualize the similarity of individual samples in a dataset based on the distance dissimilarity matrix. The MDS algorithm aims to place each sample in a lower dimensional space under the constraint that the between-sample distances are preserved as much as possible. Here, the Classical Multidimensional Scaling method in Matlab, *cmdscale()*, was used.
- t-SNE: t-SNE (t-distribution Stochastic Neighboring Embedding) is an advanced dimensionality reduction technique and a particularly suitable visualization method for high-dimensional datasets. In t-SNE, the similarity of two samples is the conditional probability density to measure the neighborhood under the t-distribution centered at each sample. The ultimate goal is to minimize the total mismatch between the conditional probability of piecewise samples under the t-distribution determined by calculating the sum of Kullback-Leibler (K-L) divergences over all samples. Here, the parameters (final_dimension, initial-dimension, perplexity) of t-SNE were (2,

10, 20), indicating that the dataset was first reduced to 10 dimensions and then mapped to 2 dimension by optimizing the K-L divergences. The details can be found in elsewhere[9].

- Silhouette Plot: Silhouette plots[10] are used to validate the consistency within clustered data. For each sample, $a(i)$ is represents the average dissimilarity within the same cluster, whereas $b(i)$ represents the lowest average dissimilarity within any other cluster. Finally, the silhouette value of the sample, i, is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

The range of $s(i)$ is $[-1, 1]$, and larger values indicate a better clustering performance.

Merging two sub-clusters by Dynamic Time Warping.

After Density Peak clustering was applied to our data, 4 distinct clusters were extracted. The average patterns of 2 clusters (Clusters II-1 and II-2) exhibited a similar periodic shape over different time scales. To simplify our subsequent analysis, these two clusters were merged using a Dynamic Time Warping strategy. Cluster II-2, which featured a smaller interval between peaks, was mapped to Cluster II-1 to generate data with a larger time interval. The general procedures can be described as follows:

1) Using the average time series of these two clusters, the mapping path between piecewise points of these two average time series was calculated by Dynamic Time Warping from the Machine Learning toolbox proposed by Jyh-Shing Roger Jang[11].
2) Using this mapping path, each sample in Cluster II-2 was mapped to generate a sample with a larger interval, and all mapped samples were then collected as mapped Cluster II-2.
3) Finally, mapped Cluster II-2 was combined with raw Cluster II-1 to generate the final Cluster II.

Normalization of Actin, Arp2/3 and VASP fluorescence signals to determine average protein dynamic patterns (Fig. 3).

Because differences in the expression levels of fluorescent proteins and their endogenous non-fluorescent proteins are not known, we cannot average the registered time series of raw fluorescence intensity. Moreover, we aimed to determine the recruitment pattern for a fluorescent protein independent of the absolute level. Therefore, before a protrusion event was registered, we separately normalized the intensity time series of each window by min-max scaling as follows:

$$I_{norm}(w,t) = \frac{I(w,t) - \min\left(I(w,t)\right)}{\max(I(w,t)) - \min(\mathrm{I}(w,t))} * 1000$$

Correlation analysis of two time series.

The time lag correlation analysis between two time series was previously described elsewhere (Machacek et al., 2009). Pearson's correlation coefficients were calculated using the time series of only protrusion segments (after protrusion onsets). The 95% confidence intervals for the average correlation were calculated by bootstrap resampling (Matlab function *bootci()*).

The time-specific correlation analysis between two time series was previously described elsewhere[1]. After the protrusion velocity and fluorescence intensities were registered with respect to protrusion onset at t = 0, Pearson's correlation coefficients (Matlab function *corrcoef*()) between the fluorescence intensity at t_1 and protrusion velocity at t_2 across the samples were calculated across the time points, where t_1 and t_2 were measured relative to the protrusion onset. A permutation t-test was used to test the significance of the correlation.

Evaluation of different time series clustering methods.

To show the effectiveness of our time series clustering, three main components, SAX for dimensional reduction, ACF for dissimilarity measure and Density Peak for clustering, were evaluated by replacing them with different methods as follows.

1) Evaluating the role of SAX: Without using SAX, ACF was directly applied to the denoised velocity dataset to calculate the ACF distances, implemented in the TSdist R package[6]. The Density Peak method was then used for clustering with a parameter equal to 1.1.

2) Evaluating the role of ACF distance: The dissimilarity measure was changed from the ACF distance to the distance metric proposed by SAX, which is the lower bound of the true Euclidean distance (Lin, 2003). Here, 8 was empirically selected as the number of symbols for SAX, and 8 symbols ranging from 0 to 7 were used to calculate the distance. The parameter of the Density Peak clustering, 4.5, was manually selected.

3) Evaluating the role of the combination of SAX and ACF. Without dimensional reduction by SAX, the denoised velocity dataset was directly used to calculate the dissimilarity using Euclidean distance from the TSdist R package[6]. The Density Peak method was then used for clustering, and the selected parameter for the Density Peak method was 9.6.

4) Evaluating the effect of Density Peak clustering. Instead of Density Peak clustering, a conventional clustering method, K-means, was used for comparison while all other steps remained unchanged. The number of clusters for K-means was set to 4 or 6.

Classification analyses of actin regulator intensities.

To further investigate the role of VASP in accelerating cell protrusions, we applied the classification approach to the fluorescence intensity time series with their corresponding protrusion clusters. For this purpose, we focused on the classification between the non-accelerating protrusion class (Clusters I/II) and accelerating protrusion class (Cluster III). First, the fluorescence intensity time series was normalized to have a mean of 0 and a standard deviation of 1 for each cell and then normalized again in the same manner for each window. We used three different classification algorithms to measure the performance of the classification, including random forest (RF)[12], support vector machine (SVM)[13], and deep neural networks (DNN)[14]. The inputs of the classifiers were the normalized fluorescence intensities (195..251), and the output was the corresponding protrusion class (non-accelerating vs accelerating protrusion). The

supervised learning was performed using the Python Scikit-Learn toolkit for RF and SVM[15] and Keras with Theano engines in Python for DNN. Because the number of time series in the non-accelerating protrusion class (Clusters I/II) was larger than those of the time series in the accelerating protrusion class (Cluster III), we under-sampled the accelerating protrusion class so that the number of data points in two classes was have the same. To obtain reproducible results, random under-sampling was applied ten times to the non-accelerating protrusion class using the Imbalanced-learn package[16]. A cross-validation was performed with 67% and 33% splitting of each sample dataset for training and testing. Moreover, the cross-validations for each sample were repeated ten times after randomly shuffling the data in each iteration. Hence, we performed the training procedures for each fluorescence intensity dataset for a total of 100 times. The hyperparameters we used in the evaluation are summarized in Table S1, and the hyperparameters were determined using a grid search approach. To assess the performance of the classification, we used accuracy (N_c/N), where N is the number of total sequences and N_c is the number of matched sequences between the original and prediction and Matthews correlation coefficient (MCC) defined as

$$\frac{N_{tp}N_{tn} - N_{fp}N_{fn}}{\sqrt{(N_{tp} + N_{fp})(N_{tp} + N_{fn})(N_{tn} + N_{fp})(N_{tn} + N_{fn})}}$$

Here, N_{tp}, N_{tn}, N_{fp}, and N_{fn} are the numbers of true positives, true negatives, false positives and false negatives, respectively. The accuracy and MCC were calculated using the Python Scikit-Learn toolkit.

Table S1. Hyperparameters used in the classification methods for each protein case. We followed the notation of Scikit-Learn and Keras for the variable names of each parameter[15]. Furthermore, we set the values of non-specified parameters to the default values in Scikit-Learn and Keras.

Methods	Hyperparameters	VASP	Arp2/3	Actin
RF	N_estimators	100		
	OOB_score	True		
SVC	C	1.93		
	γ	0.037	0.014	0.037
DNN	1st layer	Convolution, 3 filters with 3 taps, ReLU activation		

	2nd layer	Dense, 30 taps, 20% dropout, ReLU activation
	3rd layer	Dense, 10 taps, 20% dropout, ReLU activation

General Statistical Methods.

The 95% confidence interval of the velocity and normalized intensity was calculated using the bootstrap Matlab function *bootci(),* and the number of bootstrap samples was set to 1000.

For hypothesis testing, a two-sample K-S (Kolmogorov-Smirnov) test implemented with the Matlab function *kstest2()* was applied here because the K-S test makes no assumptions about the underlying data distribution.

Cell culture and imaging.

The cell culture and live cell imaging procedures are explained in detail elsewhere[1]. Briefly, PtK1 cells were cultured in Ham's F12 medium (Invitrogen) supplemented with 10% FBS, 0.1 mg/ml streptomycin, and 100 U/ml penicillin. The cells were transfected by electroporation using the Neon transfection system (Invitrogen) according to manufacturer's instructions (1 pulse, 1400 V, 20 ms) and grown on acid-washed glass #1.5 coverslips for 2 days before imaging. Prior to imaging, expressed HaloTag or SNAP-tag fusion proteins were labeled with HaloTag-TMR ligand (Promega) or SNAP-tag-TMR (New England BioLabs) ligand according to the manufacturer's instructions. Imaging was performed in imaging medium (Leibovitz's L-15 without phenol red, Invitrogen) supplemented with 10% FBS, 0.1 mg/ml streptomycin, 100 U/ml penicillin, 0.45% glucose, and 1.0 U/ml Oxyrase (Oxyrase Inc.). The PtK1 cells were then imaged at 5 s intervals for 1000 s using a 60x, 1.4 NA Plan Apochromat objective with 1.5x optovar for spinning disk confocal microscopy (see Lee et al. 2015[1] for light microscopy setup)

Plasmid construction.

Mouse VASP was subcloned into pFN21A vector (Promega) containing an N-terminal fusion to HaloTag. Human Arp3 was subcloned into the pFC14K vector (Promega) containing a C-terminal fusion to

HaloTag according to the manufacturer's instructions. A SNAP-tag-actin in C1-vector with a truncated CMV promoter (kindly provided by Martin Schwartz) was used.

Supplementary References

1. Lee, K. *et al.* Functional hierarchy of redundant actin assembly factors revealed by fine-grained registration of intrinsic image fluctuations. *Cell Syst* **1**, 37-50 (2015).
2. Huang, N.E. *et al.* The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **454**, 903–995 (1998).
3. Sundar, A., Pahwa, V., Das, C., Deshmukh, M. & Robinson, N. A Comprehensive Assessment of the Performance of Modern Algorithms for Enhancement of Digital Volume Pulse Signals *International Journal of Pharma Medicine and Biological Sciences* **5**, 91-98 (2016).
4. Lin, J., Keogh, E., Lonardi, S. & Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM*, 2-11 (2003).
5. Pierpaolo, D. & Maharaj, E.A. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems* **160**, 3565-3589 (2009).
6. Mori, U., Mendiburu, A. & Lozano, J.A. Distance Measures for Time Series in R: The TSdist Package. (2016).
7. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496 (2014).
8. Wickelmaier, F. An introduction to MDS. *Sound Quality Research Unit, Aalborg University, Denmark*, 46 (2003).
9. Laurens van der, M. & G., H. Visualizing data using t-SNE *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
10. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).
11. Jang, J.-S.R. Machine Learning Toolbox. *available at "http://mirlab.org/jang/matlab/toolbox/machineLearning".*
12. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844 (1998).
13. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).
14. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).
15. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830, (2011).
16. Lematre, G., Nogueira, F. & Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **7**, 1-5 (2016).