Single-cell tracking data aimed for big data analyses

Mónica Suárez Korsnes*[1,2] and Reinert Korsnes[2]

[1]Norwegian University of Science and Technology (NTNU), Department of Clinical and Molecular Medicine, NO-7491 Trondheim, Norway
[2]Korsnes Biocomputing (KoBio), Trondheim, Norway

May 18, 2022

Abstract

This work proposes initial refinements of data from tracking single cells in video from many experiments and at a scale assumed large enough to be meaningful for big data analyses. The present examples of data processing are for illustration only, and caution must be exercised not to consider this contribution as a stand-alone laboratory study. The main intention is to illustrate prototypes of data refinement to help initiation of fault-tolerant big data analyses to search for causal relations in possible future large amounts of data of varying quality. The authors conjecture that computer assisted comparison between cellular behavior in many and diverse experiments can constitute a new source of information. Diversity of data sources is an argument to utilize methods employing simplicity, transparency, low-cost and sensor independence as well as independence of software or alternative online services. Indicating the potential value of simple cell positional observations (tracks), can pave the way towards contributions to already established biological databases. This will presumably help to accumulate experience from a variety of data sources and facilitate big data analyses to search for new phenotypic signatures. Data can be spin-off from special analyses. Simple perturbation of the raw data can help to check for robustness of parameters derived from it. The present example data are from tracking clonal (A549) cells during several cycles while they grow in two-dimensional (2D) monolayers. The results from its processing reflect heterogeneity among the cells as well as inheritance in their response to treatments. The present illustrations include parametrization of population growth curves, aimed to simplify computerized search for similarities in large sets of single-cell tracking data. Other types of statistics, which can promote synergy between experiments, are from temporal development of cell speed in family trees with and without cell death, correlations between sister cells, development of single cell average displacements and the tendency of clustering. These are examples of parameters for further development to utilize information in large collections of data from cell behavior.

Keywords: single-cell tracking, phenotypic signature, big data, cancer diagnostic methods, daughter cells

1 Introduction

Particle tracking is developing as a tool to study cellular and intercellular processes (Meijering et al., 2012; Loeffler and Schroeder, 2019). Recent advances in single-cell research make it more interesting to follow individual cells over time to gain dynamic information from them at the individual level (Skylaki et al., 2016). Data from such observations can reflect various processes and signalling pathways inside cells and between them. Tracking single cells in video aspires to provide this type of information, which can already be lost when working on fixed dead cells. Such tracking can also contribute to characterize phenotypic states and quantify them, as permanent or temporary (Tata and Rajagopal, 2016; Gupta et al., 2019). It can provide data on lineage relationships between cells and their descendants, contributing to trace population dynamics and insight into possible pathological outcomes (Woodworth et al., 2017).

*Corresponding author: monica.s.korsnes@ntnu.no

Single-cell tracking is especially relevant to study cancer cells, which often appear to be moving targets during treatments. Cancer cells can change behavior and identity to adapt to new microenvironments by reprogramming their gene expression pool (Lüönd et al., 2021). This ability resides in their high plasticity (Lüönd et al., 2021). They can fuse during close cellular interactions, generating hybrid cancer cell subpopulations with enhance tumorigenicity and metastatic capacity (Melzer et al., 2018; Shabo et al., 2020; Hass et al., 2020). They can also display heterogeneous phenotypes within genetically identical populations as a result from expression of unique transcriptomes and proteomes (Chang et al., 2008). Heterogeneous phenotypes in cancer cells bearing epigenetic alterations are a bottleneck when guiding personalized treatments (Wakita et al., 2013; Bheda and Schneider, 2014; Shinjo and Kondo, 2015; Bintu et al., 2016).

Several authors emphasize that single-cell tracking from video has broadened the spectrum in mammalian signalling networks, drug development and cancer research (Regot et al., 2014; Suman et al., 2016; Van Valen et al., 2016; Koh et al., 2017; DuChez, 2018; Korsnes and Korsnes, 2018; Emami et al., 2020; Fujimoto et al., 2020; Fazeli et al., 2020; Ghannoum et al., 2021). Korsnes and Korsnes (2015, 2017, 2018) showed statistics from systematic single-cell tracking during several days, elucidating heterogeneous cell response and induction of cell death mechanisms. This tracking also allowed detection of inheritable traits, such as vacuolar transfer from mother to daughter cells. Inheritance may here be significant for the interpretation of observations related to autophagy signalling (Korsnes et al., 2016; Klionsky et al., 2021).

Andrei et al. (2020) pointed out different types of observables from tracking two-dimensional (2D) cell cultures and that might have biological relevance in cellular studies. 2D cultures have provided a wealth of information on fundamental biological process and diseases over the past decades (Langhans, 2018). The advantage to use these models for tracking single cells is their low cost and reproducibility as compared to three-dimensional platforms (Nishida-Aoki and Gujral, 2019; Edlund et al., 2021; Helgadottir et al., 2021). Two-dimensional models can easily integrate subsequent biochemical analyses and act as surrogate measurements for the 3D situations (Capuzzo and Vigo, 2021).

Three-dimensional (3D) models are under active development to better represent the complexity of living organisms during *in vitro* research (Yong et al., 2017; Finnberg et al., 2017; Puls et al., 2017; Fontana et al., 2021). However, they still do not recapitulate micro-environmental factors, being only reductionist of the *in vivo* counterpart (Langhans, 2018; Capuzzo and Vigo, 2021). 3D cell culture models are currently application specific and experiments with them are difficult to check for repeatability (Langhans, 2018). Current 3D platforms do not allow acquisition of cellular kinetics with a high spatial and temporal resolution over a long period of time (Wen et al., 2021; Capuzzo and Vigo, 2021). High-content screening (HCS) platforms are emerging, however visualization of 3D structures growing within complex geometrical structures remain still a big challenge mainly due to optical light scattering, light absorption and poor light penetration with prolonged imaging acquisition times (Langhans, 2018). Microfluidic devices under highly controllable environmental conditions is a well-established operation in ongoing research Schmitz et al. (2021). However, optimal nutrient supply and sufficient cell retention, especially for the long-term cultivation of slow growing cells as well as motile cells still requires a reliable cell retention concept to prevent permanent cell loss, which otherwise compromises qualitative and quantitative cell studies Schmitz et al. (2022).

This study restricts to processing data from tracking individual cells growing in two-dimensional (2D) monolayers. The intention is to show, by simple examples, the potential utility of large collections of such data, allowing users to compare their experiments with many previous similar experiments. Such collections would facilitate big data analyses, taking advantage of weak correlations in large amounts of data. The source of these data may be video recordings of various quality, assumed as by-products from experiments worldwide. The present example data therefore, for the sake of simplicity, only represent positions (tracks) of individual cells and their eventual division and dead during recording. They originate from previous work on Yessotoxin (YTX) (Korsnes and Korsnes, 2018). This small molecule compound can induce different cell death modalities (Korsnes, 2012). The broad spectrum of cellular response to YTX suits the present illustrations. The richness of responses from it may also make the compound an interesting candidate to probe cells for properties.

Data collections from single-cell tracking can be a resource for both experimental work and statistical investigations, including fault-tolerant big data analyses to search for patterns of biologic relevance. The present data processing may also have direct interest for processing videos aimed for special studies on possible emergence of rare or resistant subpopulations among cells subject to toxic agents, potential for metastasis or early screening for drug discovery. Another actual application is simply to check for the healthiness of cell populations, including testing for contamination.

The data analyses below relates cells in pedigree trees, where the initial cells are the ancestors (roots). These trees facilitate classification of cells in subpopulations according to a combined analysis of the cells in each tree. An example of such a combined analysis is to count the number of dying cells in each pedigree tree. The statistics below apply this simple idea assuming that cells in pedigree trees, with no cell death, might define a special resistant subpopulation. It reflects, for different subpopulations, variation in cell speed, correlations between sister cells as well as relocation and tendency of clustering. The authors conjecture that such data summaries can guide computerized search after patterns and causal relations in large sets of single-cell tracking data. The final proof of concept depends on access to such data sets.

A variety of relatively low-cost equipment apply to perform video-based single-cell tracking in 2D cellular models. Researchers can now in their most cost-effective way produce videos of living cells for subsequent analyses by remote (Internet/cloud based) services, as recently developed by Korsnes Biocomputing (KoBio)[1]. They may also do similar analyses/tracking using their own favorite tools, such as Image J/ TrackMate Tinevez et al. (2017). The supplementary data illustrates the potential transparency and software/equipment independence of such data production [2] facilitating sample inspection. Data perturbation can reveal if analysis results are sensitive to measurement errors. These factors make such data relevant for contribution to biological databases as proposed by Zou et al. (2015), Haniffa et al. (2021) and Osumi-Sutherland et al. (2021). The main intention here is taking advantage of opportunities to utilize data from simple and low cost recordings to create synergistic value from sharing data. The actual video recording time is therefore here assumed to be, say, less than 4-5 days without need to change media.

2 Materials and Methods

2.1 Toxin

Yessotoxin (YTX) was obtained from the Cawthron institute (Nelson, New Zealand). YTX was dissolved in methanol as a 50 µM stock solution. The stock solution was after diluted in RPMI medium (Lonza, Norway), achieving a final concentration of 2 µM YTX in 0.2 % methanol. Treated cells were incubated with 200 nM, 500 nM and 1000 nM YTX and control cells were incubated with 0.2 % methanol as vehicle.

2.2 Cell culture

A549 cell lines were provided by Dr. Yvonne Andersson and Dr. Gunhild Mari Mœlandsmo from the Institute of Cancer Research at the Norwegian Radium Hospital. Cells were cultured in RPMI 1640 (Lonza, Norway), supplemented with 9 % heat inactivated fetal calf serum (FCS, Bionordika, Norway), 0.02 M Hepes buffer 1M in 0.85 % NaCl (Cambrex no 0750, #BE17-737G) and 10 ml 1X Glutamax (100X, Gibco #35050-038), 5 ml in 500 ml medium. Cells were maintained at 37 °C in a humidified 5 % CO_2 atmosphere.

2.3 Single live-cell imaging and tracking

A549 cells were plated onto 96 multi-well black microplates (Greiner Bio-One GmbH, Germany) for time-lapse imaging. Cells were imaged into Cytation 5 Cell Imaging Reader (Biotek, USA), with temperature and gas control set to 37 °C and 5 % CO_2 atmosphere, respectively. Sequential imaging of each well was taken using 10× objective.

The bright and the phase contrast imaging channel was used for image recording. Two times, two partly overlapping images were stitched together to form images of appropriate size. A continuous kinetic procedure was chosen where imaging was carried out with each designated well within an interval of 6 min for an 94 h incubation period. Exposed cells were recorded simultaneously subject to three different concentrations of YTX 200 nM, 500 nM and 1000 nM.

The single-cell tracking in this work was performed using the in-house computer program *Kobio_Celltrack*[3]. The present data derives from previous work on YTX (Korsnes and Korsnes, 2018).

[1] https://www.korsnesbiocomputing.no/

[2] Supplementary data are available via https://user.korsnesbiocomputing.no (user *iniref_2022*, password korsnes1)

[3] https://www.korsnesbiocomputing.no/

3 Results

3.1 Single-cell tracking

The snapshots of Figure 1 illustrate production of input data for the present analyses. These data represent individual cell positions, their division, and dead during 94 h of recording. The process and tools for tracking are outside the scope of this study, which in principle could rely on data from any functional tracking system.

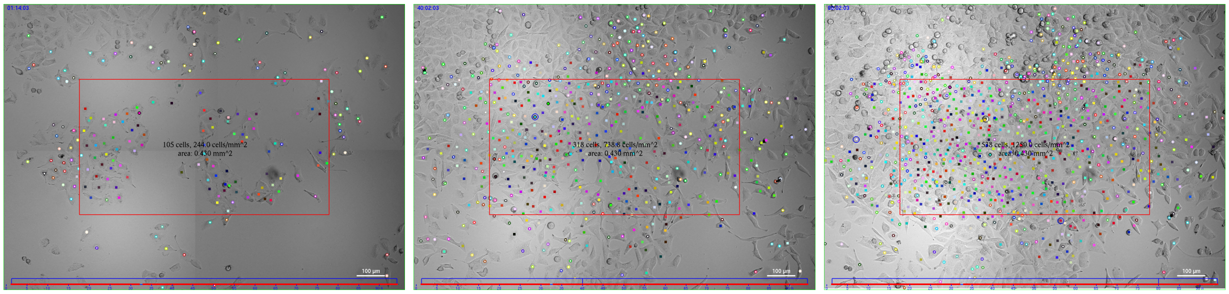


Figure 1: Snapshots illustrating tracking individual cells from video of A549 lung cancer cells. Left: at start, center: after 40 h, right: 80 h. Cells were in this case exposed to 200 nM YTX. The actual recording instrument was Cytation 5 with 10× magnification. Each image consists of 2 × 2 stitched (approximately) simultaneous images. The supplementary data shows video demonstrations of the actual tracking[2].

Figure 2 shows visual representations of a lineage derived from the present tracking of cells. The right part of this figure illustrates the positions of the actual cells during recording. The horizontal positions (x-y coordinates) here represent spatial location and the vertical axis (z-coordinate) represents time. The red frame in the figure is just large enough to contain 100 root cells at the start of recording. The present examples of statistical analyses are for the cells belonging to these pedigree trees (starting inside the red frame).

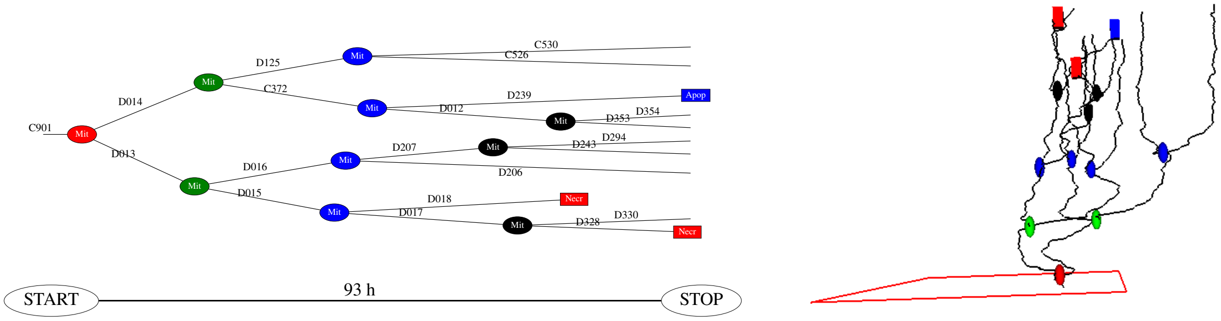
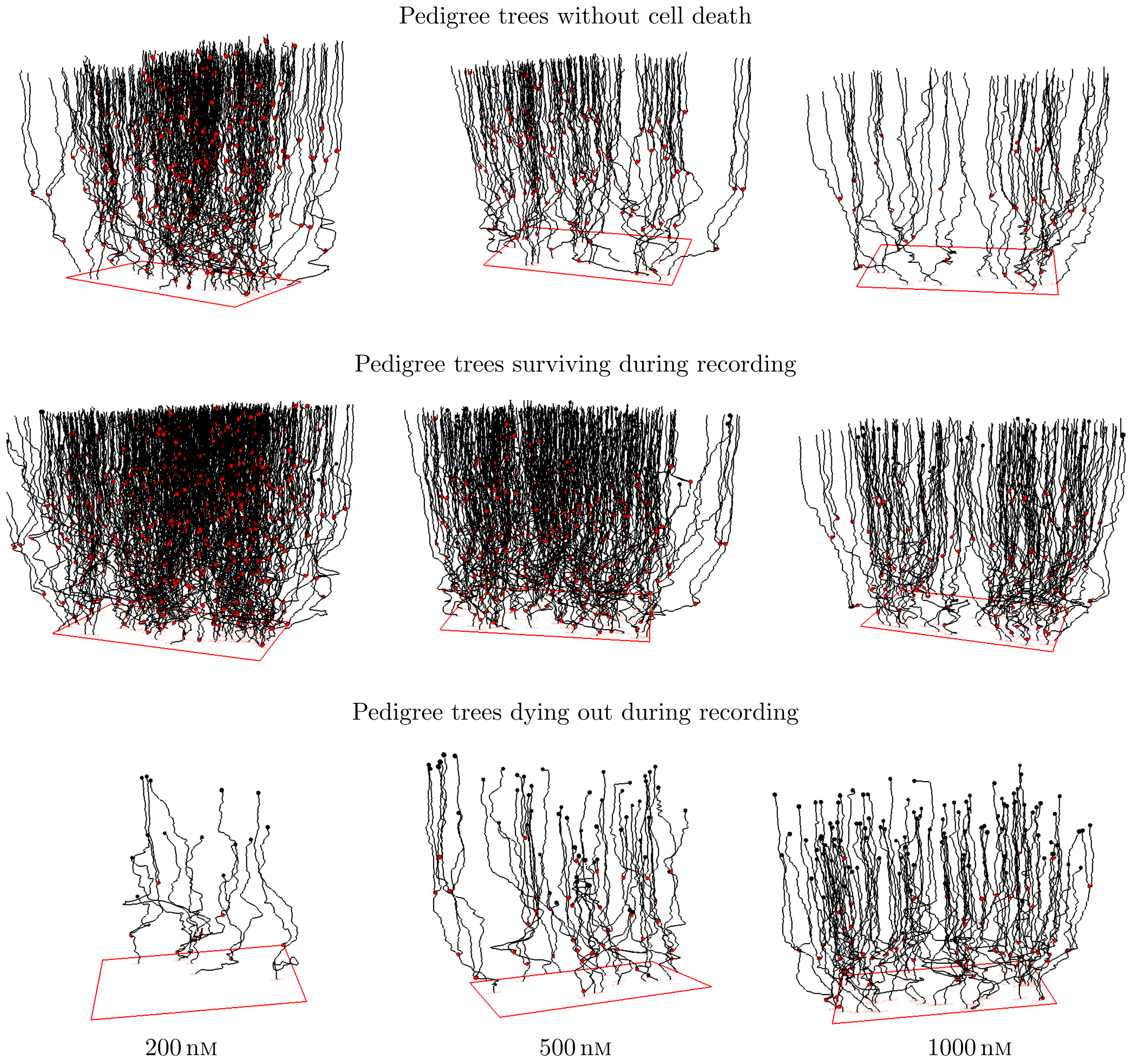


Figure 2: Example of data from tracking a cell and its descendants during about 94 h. Left: "flat" temporal representation of a pedigree tree showing cell tags/names for reference in communications. Cell division appears as ovals, where their color depends on generation. Rectangles represent cell death (blue: apoptosis-like, red: necrosis-like). Right: 3D illustration of the same pedigree tree, providing information on motion of the cells. The start position of the initial cell is inside the red frame, which in the present analyses is large enough to contain 99 other equivalent root cells (in total 100 root cells).

Figure 3 shows spatially located pedigree trees for cells exposed to YTX at three different concentrations. Cells in pedigree trees surviving at the highest YTX concentration (1000 nM) may appear to behave similar to cells subject to the lowest YTX (200 nM) concentration. It might reflect a resistant subpopulation.



Pedigree trees without cell death

Pedigree trees surviving during recording

Pedigree trees dying out during recording

200 nM

500 nM

1000 nM

Figure 3: "Forest" of pedigree trees from tracking A549 cells exposed to YTX at concentrations 200 nM, 500 nM and 1000 nM. The upper row shows trajectories for cells in lineages without death ("resistant cells"). The middle row is for trajectories of cells in lineages where at least one cell live at the end of recording ("surviving pedigree trees"). The lower row shows trajectories for cells in lineages dying out during recording. Red and black dots represent cell division and cell death respectively. Note that single-cell tracking can provide more precise information on cell viability as compared to traditional bulk assays. These type of measurements are prone to overestimate cell survival due to prior apoptotic cell clearance and disintegration.

3.2 Single-cell viability

Figure 4 shows the change in size of various cell subpopulations during video recording. The graphs show the development of number of cells in pedigree trees with roots (initial ancestors) inside a frame centered in the video and just large enough to contain 100 cells at the start of recording (see Figure 3). The population of cells belonging to the largest pedigree trees naturally grows faster than the total population. These cells potentially dominate in number after some time, if they inherit their tendency of cell division and survival. Correlations between proliferation and survival of descendants of sister cells (see Figure 5) can indicate such inheritance. The lower row in Figure 4 illustrates that cell "viability analyses" based on single-cell tracking can provide information beyond results from traditional bulk analyses.

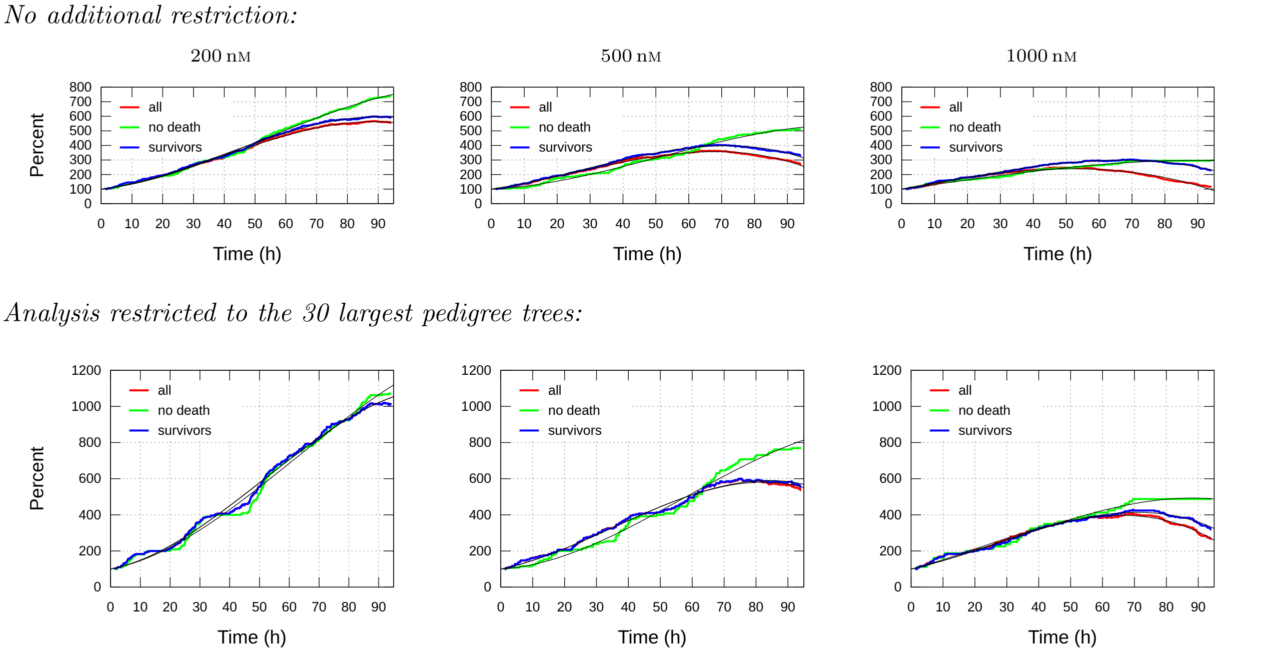


Figure 4: Growth of initial cell subpopulations starting inside the red frame just large enough to contain 100 individuals. "all": all cells; "no death": cells in pedigree trees with no death; "survivors": cells in pedigree trees where at least one cell lives at the end of recording. The upper row is for no additional restriction, whereas the lower row is for the 30 largest pedigree trees. The sum of recorded lifespan durations for cells in a pedigree tree here defines its "size". Note the smoothness of the graphs, enabling possible simple parametrization ("data compression").

The black solid lines in Figure 4 represent a third degree polynomial model fit to the data:

$$P_3(t) = 100 + at + bt^2 + ct^3 \tag{1}$$

where a, b and c are the (model) parameters and t represents time. Polynomials (or Taylor expansions) are generally a convenient way to represent smooth ("simple") functions and to compress data (representing it by three parameters). Parameters from fitting a complex biologically justified model may not necessarily represent more biological relevant information if they are less effective to compress data.

Assume fitting a Taylor model (Equation 1) to data as above (see Figure 4). Consider the resulting parameters as a point, $\boldsymbol{P} = (a, b, c)$, in the three-dimensional parameters space. Similar parameters from various experiments will give a set of points in the parameter space. If these points spread out close to for example a 2D structure (embedded in the 3D space), then there should, intuitively, be hope for finding statistical models with two parameters (instead of three) providing a biological interpretation/understanding. Voids in the parameter space can also represent knowledge.

Figure 6 shows percentage development of number of cells in pedigree trees as a function of time after the first cell division. The left part of the figure is for the all 100 pedigree tress (initiating in the red frame as

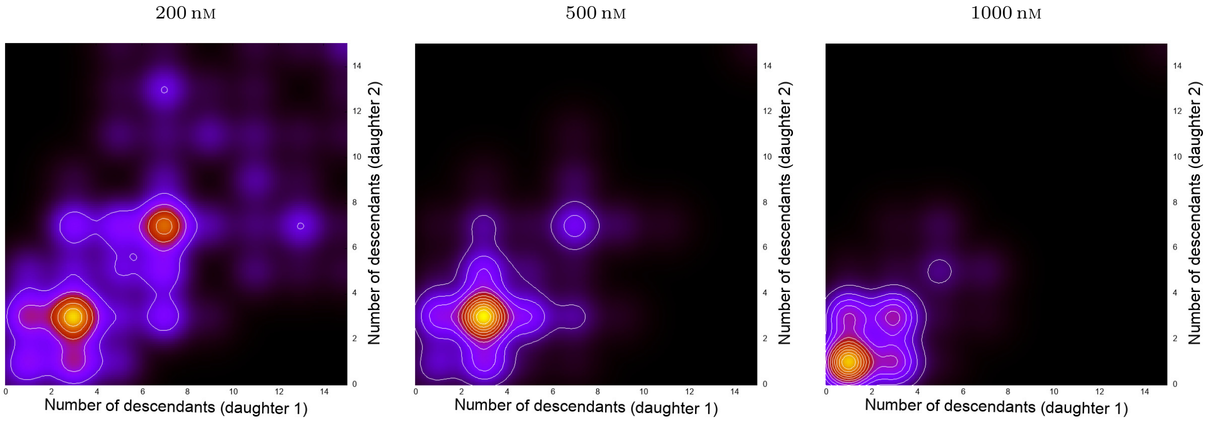


Figure 5: Results from kernel smoothing (bandwidth 1.5) of stack plots for number of descendants of first generation (A549) sister cells within 70 h after their birth. The cells are exposed to YTX at concentrations 200 nM, 500 nM and 1000 nM. The plots are only for cells born within 20 h after start of recording. Note the two apparent clusters in the plot for cells exposed to 200 nM YTX indicating that sister cells here inherit their survival and tendency of division from their common mother cell (at least within the actual time frame of 70 h after their birth).

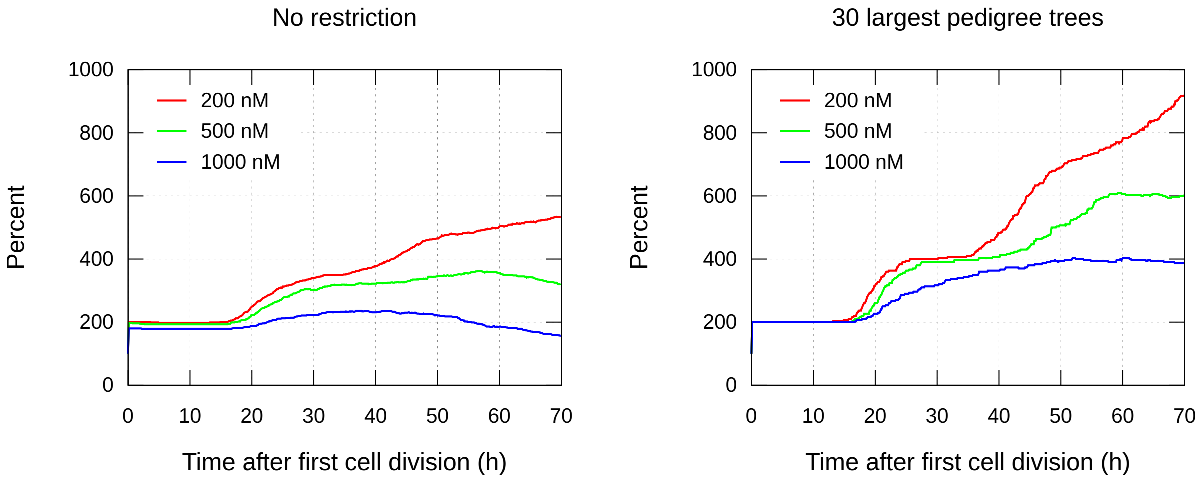


Figure 6: Percentage development of number of cells in pedigree trees as a function of time after first division (doubling at start). The graphs are for A549 cells exposed to YTX at concentrations 200 nM, 500 nM and 1000 nM. The graphs start at 200 percent, reflecting the doubling of number of cells immediately after the first cell division. The sum of recorded lifespan duration for cells in a pedigree tree here defines its "size".

explained above), and the right part is for the 30 largest pedigree trees. The figure shows that cells exposed to the lowest concentration of YTX (200 nM) tend to follow a regular timing for cell division, as opposed to those subject to the highest concentration (1000 nM). This tendency is most expressed for the largest pedigree trees (right part of the figure).

3.3 Speed

Track length for a cell during a period of time t (divided by t) can intuitively define its average *speed* during that period. However, track length is not in practice directly available nor be it well-defined for imprecise and irregular positional data, where measures of length can depend on resolution. *Cell speed* could (ad hoc) refer to movements of a given defined point in a cell (for example, the mean point of the nucleus/nuclei). However, it may principally be looked at as a spatio-temporally localized (statistical) property of a cell. Future work may assume an ”uncertainty principle” where a positional data point is considered as a random selection from a set of possible positions depending on the tracking method. An alternative approach is to increase the level of sophistication and replace the concept of ”cell speed” with temporal change in the (segmented) set of points covered by an actual cell.

Estimates of positions are, for any definition, imprecise for low quality imagery data. This work therefore, for the sake of simplicity, demonstrates Gaussian kernel smoothing and interpolation (Wand and Jones, 1994) to define speed. The actual bandwidth is 15 min. Perturbations of estimates of cell positions may help to reveal how final results are sensitive to this choice of bandwidth. The authors leave out this exercise for a separate study. Note that big data approaches may in principle automatically sort out useful definitions of speed.

Figure 7 shows distributions of the eight hours centered moving generalized mean speed for cells in lineages with and without dead during recording. The upper and third row are for the regular mean, whereas the second and lower row similarly show the fourth power mean for the same data. This example illustrates a possible data product that presumably could provide information to big data analyses. The power mean M_p is increasingly more sensitive to the highest speeds for increased values of p. The distribution for M_4, for example, seems to be more sensitive to cell death in lineages as compared to lineages with no cell death. One can expect that the power mean M_p for $p = 1, 2, \ldots, n$ will in a compact way reflect the distribution of speed for a restricted value of n.

3.4 Correlation between descendants of sister cells

Figure 8 shows joint distributions for the total track length of first generation sister cells and their descendants 60 h after the birth of these (initial) sister cells. These statistics restrict to sister cells born within 30 h after start of recording. The estimates result from using the algorithm scipy.stats.gaussian_kde from SciPy[4] with defaults settings (i.e., the ’scott’ method defines the estimator bandwidth). Section 3.3 outlines the present estimation of length from imprecise positional data (applying Gaussian kernel smoothing).

The joint distributions of Figure 8 show positive correlations and hence reflect inheritance from mother cells to their daughters. The authors will not further speculate on the biological significance of these statistics, since they only reflect results from one experiment. However, a main finding here is that such distributions are sensitive to cell treatment. One may therefore suspect such data summaries to be relevant for big data analyses. The regularity of such distributions enables effective parametrization (or data compression) to help search in large databases.

[4] https://scipy.org/

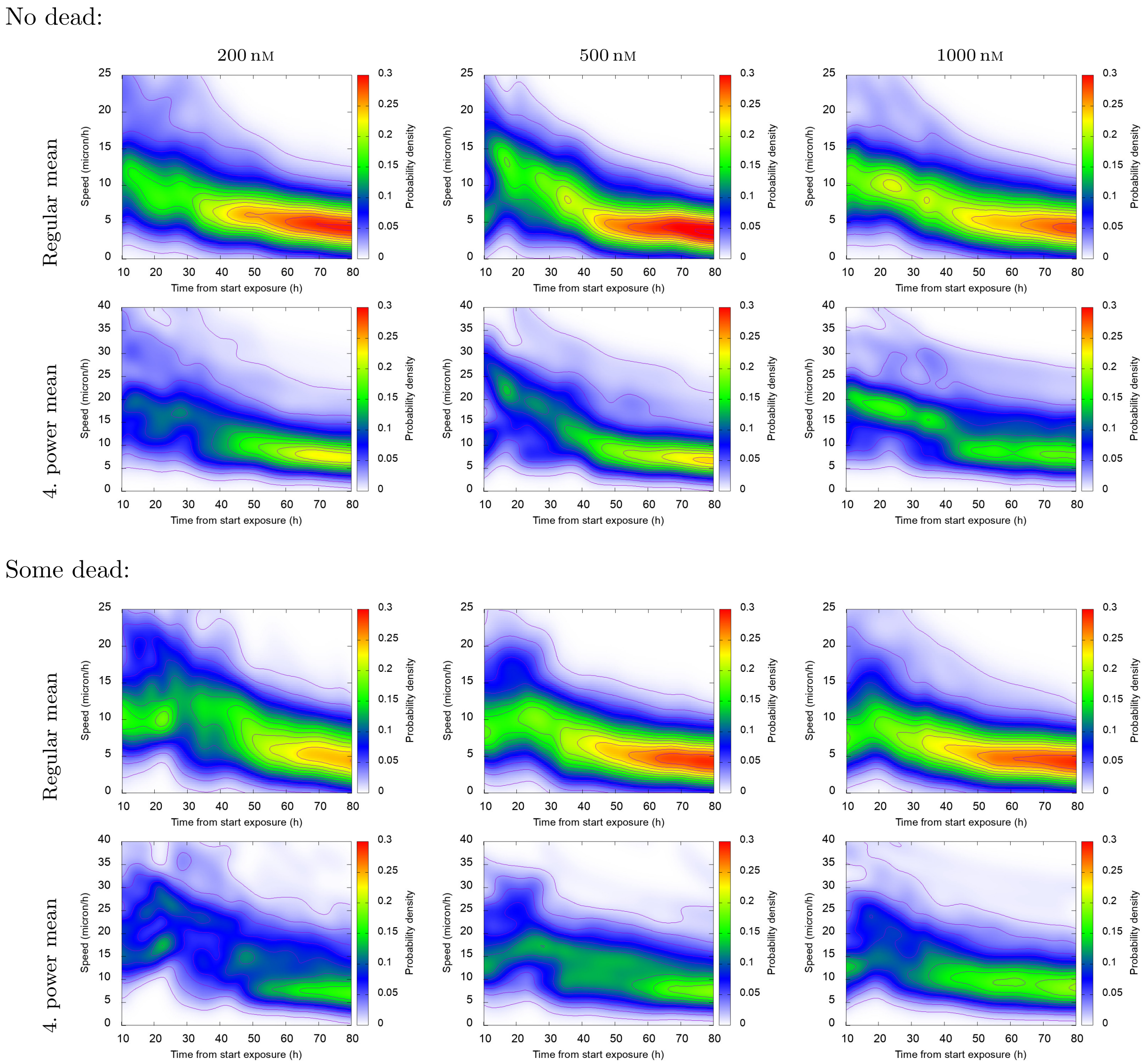
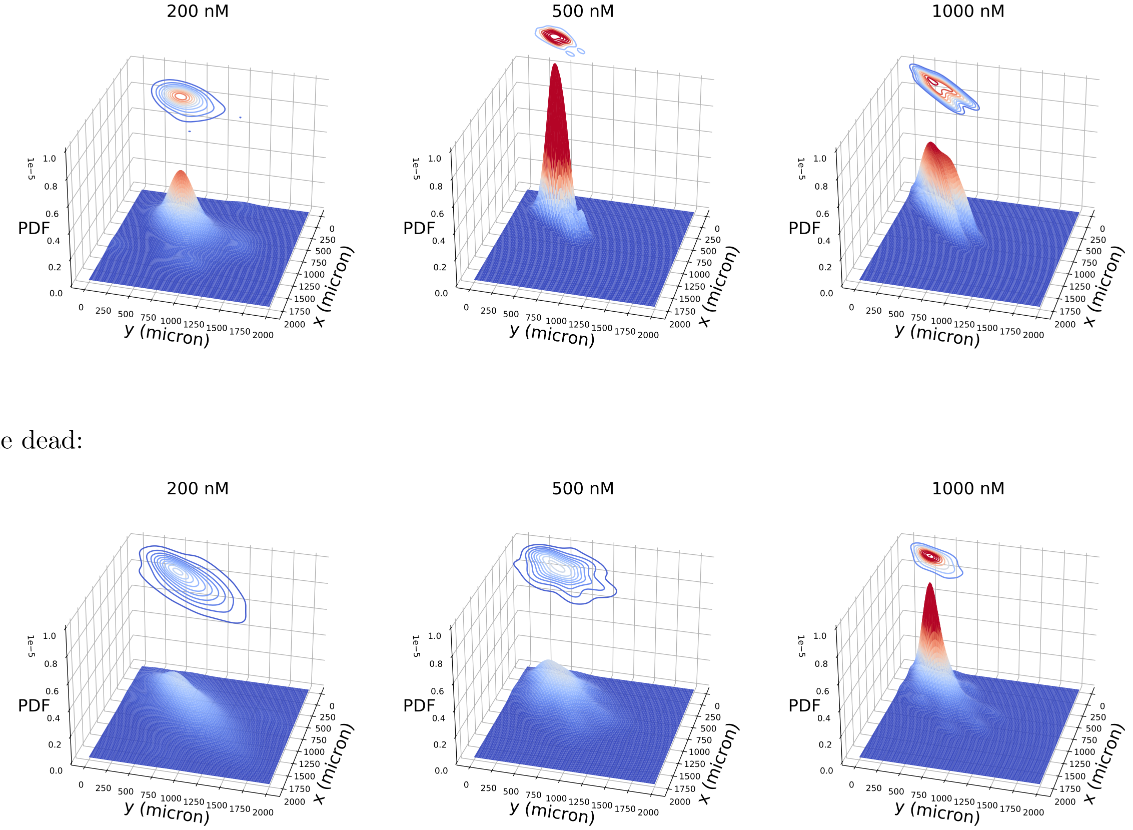


Figure 7: Distribution of 8 hours centered running generalized mean of speed of cells during recording. The top and third row show regular (first order) mean, and the second and fourth row show fourth power mean (M_p, $p = 4$). Note the difference between the distributions, especially at the first part of the recording.

No dead:



Some dead:

Figure 8: Joint distribution of total track length (x and y) for the first generation sister cells and their descendants 60 h after the birth of these (initial) sister cells. The cells are subject to YTX exposure at concentrations 200 nM, 500 nM and 1000 nM. The upper row shows distributions for the pedigree trees with no cell death, and the lower one shows pedigree trees with some cell death.

3.5 Mean square displacement (MSD) of first generation daughter cells

Figure 9 shows mean square displacement (MSD) of first generation daughter cells as function of time from their birth. The figure is for cells in pedigree trees, with and without cell death during recording. The upper row here shows the tendency of cells to need extra time to start drifting from their place of birth. Processing of more data may reveal if this extra time can be considered a “phenotype”.

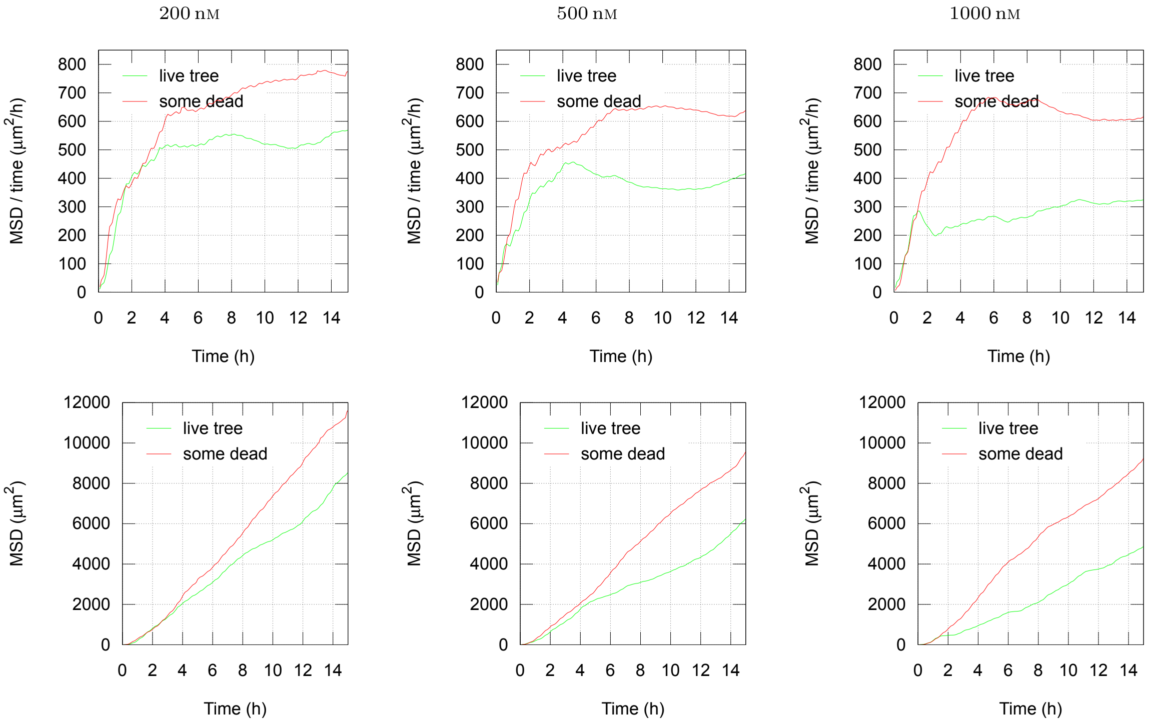


Figure 9: Upper row: Mean square displacement (MSD) of first generation daughter cells as a function of time t from their birth (divided by t). Lower row: MSD of first generation daughter cells as a function of time from their birth. “live tree”: for cells in pedigree trees with no cell dead (during recording period). “some dead”: for cells in pedigree trees with some cell dead (during recording period). The cells are subject to exposure by YTX at concentrations 200 nM, 500 nM and 1000 nM.

Note that cells in lineages with dying cells here tend to move faster from their initial position as compared to cells with no observed cell death. A possible hypothesis is that cells with the strongest (inheritable) tendencies to move, are more vulnerable to the actual toxin (YTX) as compared to the others. One may also relate the observation to the concept of “fight-or-flight” reaction, where many types of cells respond to a variety of stressors in a reasonably standardized fashion, which allows them to combat the offending stimulus or escape from it Goligorsky (2001).

A “memory-less” (Brownian type) motion for movement would give graphs as straight horizontal lines for the upper row in Figure 9 (and straight upward tilting lines for the lower row). The present graphs reflect that the direction of movement tend to be independent of the direction about 4 h to 6 h earlier. The period up to about 4 h is “memory time” reflecting how long cells tend to keep their direction. It can partly correlate with cell shape, assuming elongated cells move in their longitudinal direction.

Assume the vector $\boldsymbol{r}(t)$ represents the relocation of a cell t time units after its birth. The vector dot (inner) product

$$\boldsymbol{c}(t) = \boldsymbol{r}(t) \cdot \boldsymbol{r}(t) \tag{2}$$

then gives this distance squared (equal $|\boldsymbol{r}(t)|^2$). Figure 9 shows average values for $c(t)$ for two subsets of cells where t ranges from 0 to 15 h. A tempting idea is slightly to modify this elaboration and check for average value of

$$\boldsymbol{c}_{1,2}(t) = \boldsymbol{r}_1(t) \cdot \boldsymbol{r}_2(t) \tag{3}$$

where $\boldsymbol{r}_1(t)$ and $\boldsymbol{r}_2(t)$ each represent the positions of a couple of siblings (sister cells) t time units after their birth. Figure 10 shows an example of results from such a numerical experiment. A motivation for this test is the conceptual simplicity and the pure formal similarity between the Equations 2 and 3. The authors have no specific biological interpretations of these graphs, except that they reflect the tendency for sister cells to follow each other after their birth. This tendency seems to depend on exposure.

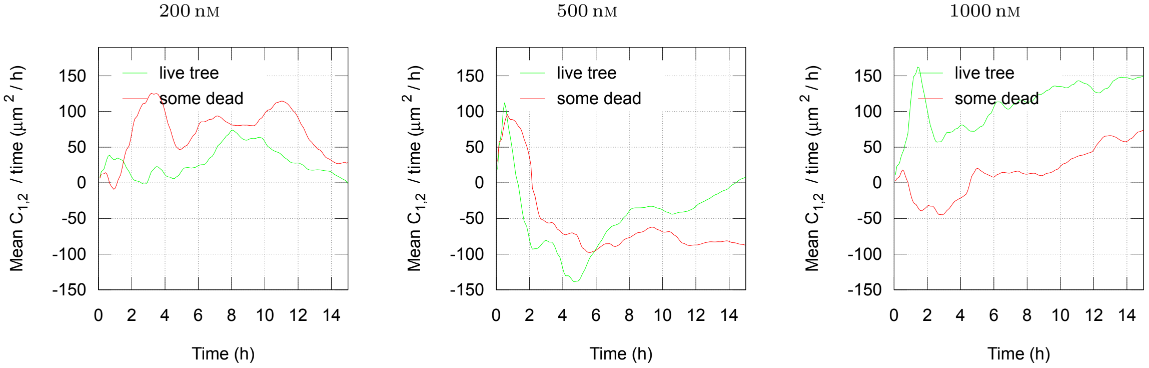


Figure 10: Average values of $\boldsymbol{c}_{1,2} = \boldsymbol{r}_1 \cdot \boldsymbol{r}_2$ for cells in pedigree trees with and without cell death. These examples are for cells exposed to YTX at concentrations 200 nM, 500 nM and 1000 nM.

3.6 Material exchange and trait inheritance

Moving cells can remain physically close for periods. This can indicate intercellular communication or exchange of material and which can subsequently affect them. Figure 11 illustrates identification of such events where cells stay close for periods. This type of data may have special interest for co-culture or studies on differentiation where interactions are crucial. Cells can interact through physical contact, surface receptor-ligand interaction, cellular junctions, and secreted stimulus (Nishida-Aoki and Gujral, 2019). Understanding these type of interactions can contribute to decipher the complex network of interaction between cells, helping to improve therapeutics (Nishida-Aoki and Gujral, 2019).

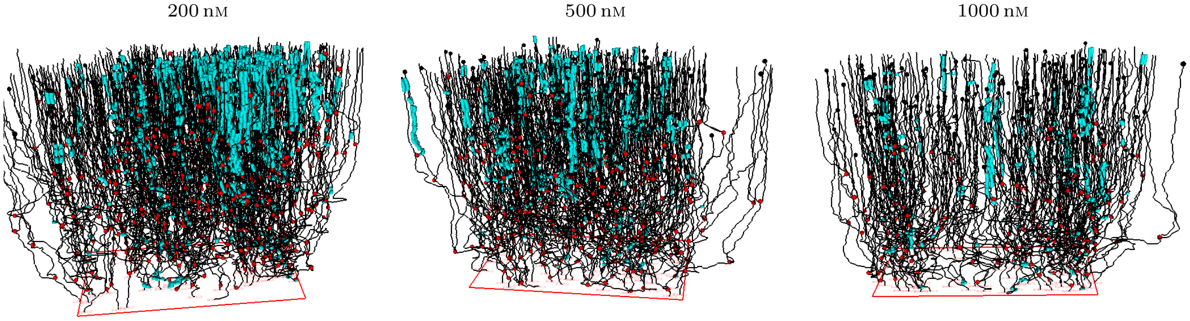


Figure 11: Forest of pedigree trees including identification of events where cells stay at the vicinity of each other for at least 4 hours 2 hours apart from their birth (cell division). The cells were subject to YTX exposure at concentrations 200 nM, 500 nM and 1000 nM.

Analyses of “forests” of pedigree trees can reflect effects from events where cells absorb debris from dead cells and transfer it to their descendants. Figure 12 shows an example of such behavior where a cell includes an apoptotic body from a neighboring dying cell. Such apoptotic bodies can subsequently appear as vacuoles

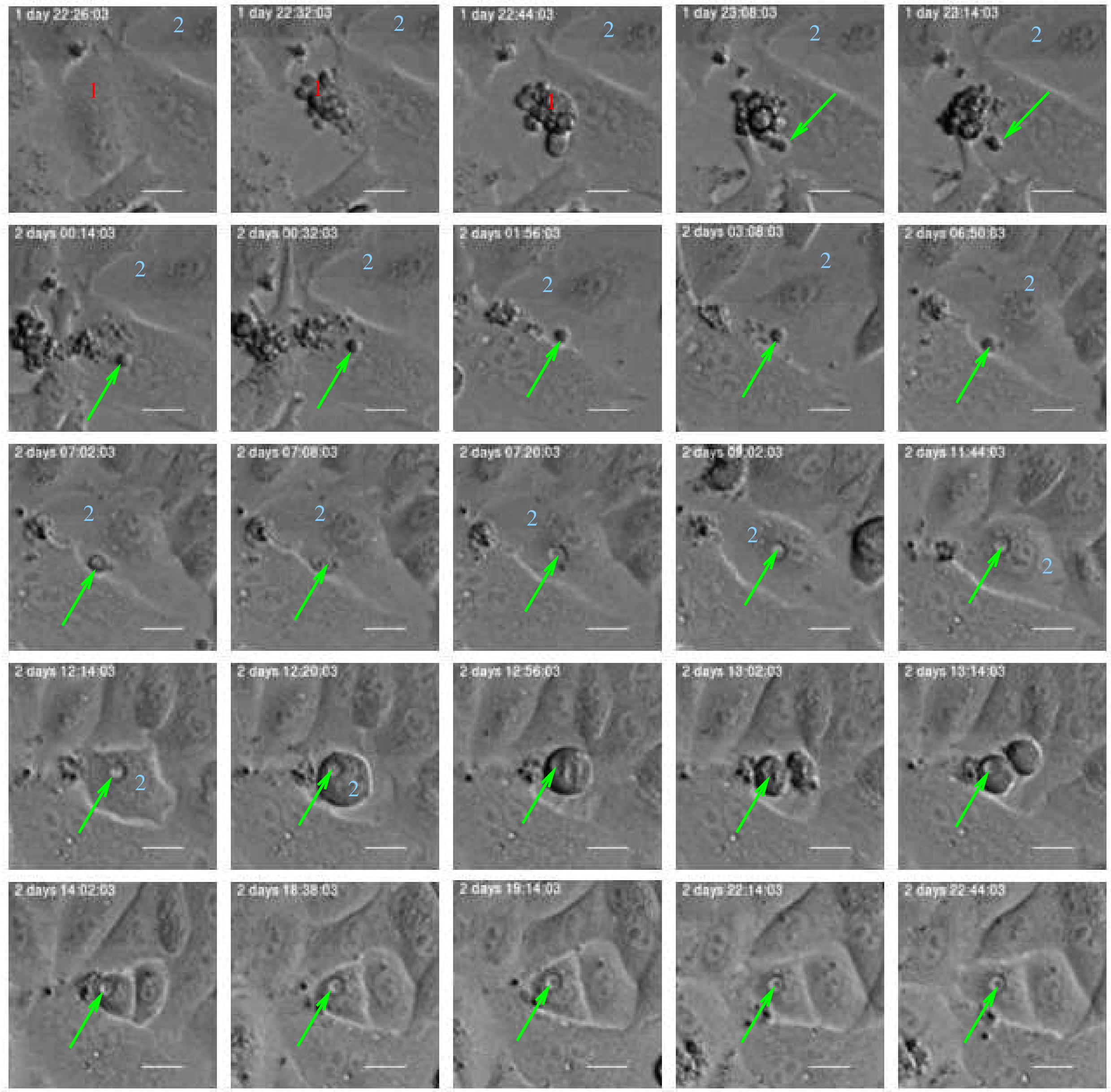


Figure 12: Example where an apoptotic body (green arrow) from a dying cell (1) ends up as a vacuole in a neighboring cell (2) which subsequently divides, and the vacuole ends up in one of the daughter cells. Detailed inspection of many cells in video can reveal such rare events and shed light on epigenetic heritage and generally signalling downstream pedigree trees. Such signalling is an argument to study lineages as independent entities and for example apply information on lineage relations when for example classifying cells.

in the absorbing cell. Sets of such vacuoles in a cell are traceable throughout cell division by comparing their size and number.

4 Discussion

This work illustrates a number of possible methods to refine (or compress) data from video-based single-cell tracking. The intention is to provide relevant input for big data analyses (or machine learning in general) to identify biomarkers for better diagnosis and prognosis. Well proven fault-tolerant computerized methods are here available to search for causal relations in large data sets (Theodoridis, 2015; Kotsiantis et al., 2007; Linero, 2017). The principle of Occam's razor (Domingos, 1998, 1999) can guide the search, favouring simplifications and approximations. It can be considered as a contradiction to anticipate the exact result from trying big data analysis methods, nor can one expect to anticipate what refinement methods are most effective. Successful big data analyses are (similar to data mining) assumed beyond the reach of human brains. However, their result may finally be understood by humans.

Big data methods go beyond assuming linear association between variables. The present examples therefore restrict to visual/intuitive illustrations of data refinement left for further processing. Existence of several local maxima in joint distributions (clustering) may, as an example, reflect significant biological information. The left part of Figure 5 illustrates this point. It shows two main maxima of the joint distribution of number of descendants of sister cells. This may indicate inheritance of robustness/viability making it likely for the most robust cells finally to dominate in number (which could be relevant for prognoses in cancer).

The present examples of refinement methods typically show different behaviour of cells in pedigree trees with cell death as compared to the behaviour of cells in pedigree trees without cell death (during recording). Some of these examples also show correlations between sister cells or descendants of sister cells. This is an argument to treat whole pedigree trees as individual entities in the initial data refinement.

Successful application of big data analyses can, in addition to sort out causal relations, give the possibility to search for similarities between the behavior of cells in many experiments. Methods to compare experiments can in general be an important part of a collective knowledge base of cell behavior.

Recent progress in techniques for sparse representations, compressive sensing and machine learning (see e.g. Papyan et al., 2018; Mousavi et al., 2019) give a perspective of direct automatic identification of actual biomarkers directly from video of cells. The present work contributes to this development by demonstrating initial refinement of data from single-cell tracking. These data summaries may also be of direct biological or medical interest in the conceptual frame of standalone experiments. They may in addition help the development of formal mathematical methods applying concepts from statistical physics (Banigan, 2013). However, note that machine search for causality in data may utilize weak correlations without any immediate intuitive meaning.

This work illustrates derivation of the following parameters from single-cell tracking data which represent positions of individual live cells, their division, and dead during several cell cycles:

- Number of cells in various classes of pedigree trees during video recording (Section 3.2). It may reflect that some pedigree trees consists of specially viable and resilient cells. This property seems to be already written into the root (ancestor) cell. Intrusive single cell analyses after tracking, while preserving track identities, may clarify corresponding mechanisms behind this resilience.

- Parameters from (representations of) speed distributions for various subsets of cells during tracking (Figure 7). The regularity of these distributions allow representations by few parameters (so-called sparse representation).

- Parameters from joint distributions of the size of (pedigree) subtrees for the first generation sister cells where they are roots cells (Figure 8). Such distributions can be parameterized by correlation coefficients, covariance, and shape parameters (or various sparse representations).

- "Memory" time of trajectories for cells in various subpopulations. Figure 9 reveals that cell trajectories can have a tendency to keep their direction, typically during 2 h to 4 h. This tendency can reflect cell shape.

- Tendency for cells to stay close to each other for periods. Figure 11 visualizes an example where cells tend to stay close for periods of time. Such events can potentially reflect intercellular communication and material exchange (see Figure 12). The tendency may have special interest for studies where communication between different cell types play a role. Tracking of cells in co-culture can in this case help to reveal how to affect such behavior.

An intention behind the present work is, as pointed out above, to promote ideas for better and easier comparison between different experiments. This would promote securing reproducibility of observations, which has emerged as a main concern in life science research in recent years (Hirsch and Schildknecht, 2019). Easy exchange of raw and refined data is paramount in such quality assurance. Experiments on cells can include video recording of them under standard (common) conditions, and statistics from tracking the cells can reveal differences between experiments and which can affect their reproducibility. Tracking under standard conditions may in general reveal effects on cells and which otherwise may pass under the radar using bulk assays. This is an example of direct use of the present type of statistics.

Large-scale sharing of data from tracking single cells in video will naturally raise questions on robustness of results from initial analyses of them. Cells in different experiments may never be treated exactly the same way. Cells can be sensitive to photo-toxicity as well as possible molecular probes. Type of extracellular matrices and their proteins can also affect cellular behavior in test wells Vigilante et al. (2019). Data analyses can reveal to what degree comparison of data from them still apply. It will be important to identify ranges of conditions for cells in which they will behave in comparable ways. It will also be important to identify conditions/treatments where cellular behavior is sensitive to small and uncontrollable perturbations. Data analyses may also reveal possible probabilistic views of results from observing cellular behavior.

Further development of sensors and software will extend the above restriction to data on cell positions, division, and dead. This will advance exploitation of its potential utility, as indicated by several authors (Van Valen et al., 2016; Tsai et al., 2019; Moen et al., 2019; Liu et al., 2020; Ghannoum et al., 2021). Single-cell tracking from high quality imagery allows collecting data on phenotypical changes, otherwise difficult to measure from an end-point measurement such as single-cell RNA-sequencing (scRNA-seq) (Liu et al., 2020). Furthermore, epigenetic states, protein expression and enzyme activity, can not only be inferred from changes in gene expression (Liu et al., 2020; Zhang et al., 2020). Integrating single-cell tracking with RNA-seq analyses can therefore complement characterization of biological process by combining analyses of cellular phenotypes with gene expression profiles (Lane et al., 2017; Yuan et al., 2018). These analyses allow overlaying phenotypic cell identity with genetic lineage information for a more comprehensive view of clonal relationships, since gene expression alone is not sufficient to classify cell states (Woodworth et al., 2017; Gerbin et al., 2021). Integrating such analyses into cell ontology can help to discover a large variety of novel cell populations (Osumi-Sutherland et al., 2021). Tracking individual cells can therefore complement current cell ontology efforts.

Authors contribution

M.S.K. conceived the study and conducted the laboratory experiments, R.K. made the computer programming and statistic analyses; both authors analyzed the results and wrote the manuscript.

Funding

Olav Raagholt and Gerd Meidel Raagholts legacy supported this study.

Acknowledgements

This work was supported by The Norwegian University of Science and technology (NTNU), Department of Clinical and Molecular Medicine.

Conflicts of interest

Mónica Suárez Korsnes and Reinert Korsnes declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest. Mónica Suárez Korsnes is the owner of the upstart firm Korsnes Biocomputing (KoBio) aimed to participate in research and development of methods for single-cell analysis.

References

Andrei, L., Kasas, S., Garrido, I. O., Stanković, T., Korsnes, M. S., Vaclavikova, R., et al. (2020). Advanced technological tools to study multidrug resistance in cancer. *Drug Resistance Updates* 48, 100658

Banigan, E. J. (2013). *Statistical physical models of cellular motility* (University of Pennsylvania)

Bheda, P. and Schneider, R. (2014). Epigenetics reloaded: the single-cell revolution. *Trends in cell biology* 24, 712–723

Bintu, L., Yong, J., Antebi, Y. E., McCue, K., Kazuki, Y., Uno, N., et al. (2016). Dynamics of epigenetic regulation at the single-cell level. *Science* 351, 720–724

Capuzzo, A. M. and Vigo, D. (2021). Microfluidic live-imaging technology to perform research activities in 3d models. *bioRxiv*

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 453, 544–547

Domingos, P. (1998). Occam's two razors: The sharp and the blunt. In *KDD*. 37–43

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data mining and knowledge discovery* 3, 409–425

DuChez, B. J. (2018). Automated tracking of cell migration with rapid data analysis. *Current protocols in cell biology* 76, 12–12

Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., et al. (2021). Livecell—a large-scale dataset for label-free live cell segmentation. *Nature methods* , 1–8

Emami, N., Sedaei, Z., and Ferdousi, R. (2020). Computerized cell tracking: Current methods, tools and challenges. *Visual Informatics*

Fazeli, E., Roy, N. H., Follain, G., Laine, R. F., von Chamier, L., Hänninen, P. E., et al. (2020). Automated cell tracking using stardist and trackmate. *F1000Research* 9

Finnberg, N. K., Gokare, P., Lev, A., Grivennikov, S. I., MacFarlane IV, A. W., Campbell, K. S., et al. (2017). Application of 3d tumoroid systems to define immune and cytotoxic therapeutic responses based on tumoroid and tissue slice culture molecular signatures. *Oncotarget* 8, 66747

Fontana, F., Marzagalli, M., Sommariva, M., Gagliano, N., and Limonta, P. (2021). In vitro 3d cultures to model the tumor microenvironment. *Cancers* 13, 2970

Fujimoto, K., Seno, S., Shigeta, H., Mashita, T., Ishii, M., and Matsuda, H. (2020). Tracking and analysis of fucci-labeled cells based on particle filters and time-to-event analysis. *IJBBB*

Gerbin, K. A., Grancharova, T., Donovan-Maiye, R. M., Hendershott, M. C., Anderson, H. G., Brown, J. M., et al. (2021). Cell states beyond transcriptomics: integrating structural organization and gene expression in hipsc-derived cardiomyocytes. *Cell Systems*

Ghannoum, S., Antos, K., Leoncio Netto, W., Gomes, C., Köhn-Luque, A., and Farhan, H. (2021). Cellmaptracer: A user-friendly tracking tool for long-term migratory and proliferating cells associated with fucci systems. *Cells* 10, 469

Goligorsky, M. S. (2001). The concept of cellular "fight-or-flight" reaction to stress. *American Journal of Physiology-Renal Physiology*

Gupta, P. B., Pastushenko, I., Skibinski, A., Blanpain, C., and Kuperwasser, C. (2019). Phenotypic plasticity: driver of cancer initiation, progression, and therapy resistance. *Cell stem cell* 24, 65–78

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., et al. (2021). A roadmap for the human developmental cell atlas. *Nature* 597, 196–205

Hass, R., von der Ohe, J., and Ungefroren, H. (2020). Impact of the tumor microenvironment on tumor heterogeneity and consequences for cancer cell plasticity and stemness. *Cancers* 12, 3716

Helgadottir, S., Midtvedt, B., Pineda, J., Sabirsh, A., B. Adiels, C., Romeo, S., et al. (2021). Extracting quantitative biological information from bright-field cell images using deep learning. *Biophysics Reviews* 2, 031401

Hirsch, C. and Schildknecht, S. (2019). In vitro research reproducibility: Keeping up high standards. *Frontiers in pharmacology* 10, 1484

Klionsky, D. J., Abdel-Aziz, A. K., Abdelfatah, S., Abdellatif, M., Abdoli, A., Abel, S., et al. (2021). Guidelines for the use and interpretation of assays for monitoring autophagy. *autophagy* 17, 1–382

Koh, S.-B., Mascalchi, P., Rodriguez, E., Lin, Y., Jodrell, D. I., Richards, F. M., et al. (2017). A quantitative fastfucci assay defines cell cycle dynamics at a single-cell level. *Journal of cell science* 130, 512–520

Korsnes, M. S. (2012). Yessotoxin as a tool to study induction of multiple cell death pathways. *Toxins* 4, 568–579

Korsnes, M. S., Kolstad, H., Kleiveland, C. R., Korsnes, R., and Ørmen, E. (2016). Autophagic activity in bc3h1 cells exposed to yessotoxin. *Toxicology in Vitro* 32, 166–180

Korsnes, M. S. and Korsnes, R. (2015). Lifetime distributions from tracking individual bc3h1 cells subjected to yessotoxin. *Frontiers in bioengineering and biotechnology* 3, 166

Korsnes, M. S. and Korsnes, R. (2017). Mitotic catastrophe in bc3h1 cells following yessotoxin exposure. *Frontiers in cell and developmental biology* 5, 30

Korsnes, M. S. and Korsnes, R. (2018). Single-cell tracking of a549 lung cancer cells exposed to a marine toxin reveals correlations in pedigree tree profiles. *Frontiers in oncology* 8, 260

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* 160, 3–24

Lane, K., Van Valen, D., DeFelice, M. M., Macklin, D. N., Kudo, T., Jaimovich, A., et al. (2017). Measuring signaling and rna-seq in the same cell links gene expression to dynamic patterns of nf-κb activation. *Cell systems* 4, 458–469

Langhans, S. A. (2018). Three-dimensional in vitro cell culture models in drug discovery and drug repositioning. *Frontiers in pharmacology* 9, 6

Linero, A. R. (2017). A review of tree-based bayesian methods. *Communications for Statistical Applications and Methods* 24, 543–559

Liu, Z., Yuan, J., Lasorella, A., Iavarone, A., Bruce, J. N., Canoll, P., et al. (2020). Integrating single-cell rna-seq and imaging with scope-seq2. *Scientific reports* 10, 1–15

Loeffler, D. and Schroeder, T. (2019). Understanding cell fate control by continuous single-cell quantification. *Blood, The Journal of the American Society of Hematology* 133, 1406–1414

Lüönd, F., Tiede, S., and Christofori, G. (2021). Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer* , 1–12

Meijering, E., Dzyubachyk, O., and Smal, I. (2012). Methods for cell and particle tracking. *Methods in enzymology* 504, 183–200

Melzer, C., von der Ohe, J., and Hass, R. (2018). Enhanced metastatic capacity of breast cancer cells after interaction and hybrid formation with mesenchymal stroma/stem cells (msc). *Cell Communication and Signaling* 16, 1–15

Moen, E., Borba, E., Miller, G., Schwartz, M., Bannon, D., Koe, N., et al. (2019). Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. *bioRxiv* , 803205

Mousavi, A., Rezaee, M., and Ayanzadeh, R. (2019). A survey on compressive sensing: Classical results and recent advancements. *arXiv preprint arXiv:1908.01014*

Nishida-Aoki, N. and Gujral, T. S. (2019). Emerging approaches to study cell–cell interactions in tumor microenvironment. *Oncotarget* 10, 785

Osumi-Sutherland, D., Xu, C., Keays, M., Kharchenko, P. V., Regev, A., Lein, E., et al. (2021). Cell types and ontologies of the human cell atlas. *arXiv preprint arXiv:2106.14443*

Papyan, V., Romano, Y., Sulam, J., and Elad, M. (2018). Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. *IEEE Signal Processing Magazine* 35, 72–89

Puls, T., Tan, X., Whittington, C. F., and Voytik-Harbin, S. L. (2017). 3d collagen fibrillar microstructure guides pancreatic cancer cell phenotype and serves as a critical design parameter for phenotypic models of emt. *PloS one* 12, e0188870

Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S., and Covert, M. W. (2014). High-sensitivity measurements of multiple kinase activities in live single cells. *Cell* 157, 1724–1734

Schmitz, J., Stute, B., Taeuber, S., Kohlheyer, D., von Lieres, E., and Grünberger, A. (2022). Reliable cell retention of mammalian suspension cells in microfluidic cultivation chambers. *bioRxiv*

Schmitz, J., Täuber, S., Westerwalbesloh, C., von Lieres, E., Noll, T., and Grünberger, A. (2021). Development and application of a cultivation platform for mammalian suspension cell lines with single-cell resolution. *Biotechnology and bioengineering* 118, 992–1005

Shabo, I., Svanvik, J., Lindström, A., Lechertier, T., Trabulo, S., Hulit, J., et al. (2020). Roles of cell fusion, hybridization and polyploid cell formation in cancer metastasis. *World journal of clinical oncology* 11, 121

Shinjo, K. and Kondo, Y. (2015). Targeting cancer epigenetics: Linking basic biology to clinical medicine. *Advanced drug delivery reviews* 95, 56–64

Skylaki, S., Hilsenbeck, O., and Schroeder, T. (2016). Challenges in long-term imaging and quantification of single-cell dynamics. *Nature Biotechnology* 34, 1137–1144

Suman, R., Smith, G., Hazel, K. E., Kasprowicz, R., Coles, M., O'Toole, P., et al. (2016). Label-free imaging to study phenotypic behavioural traits of cells in complex co-cultures. *Scientific reports* 6, 1–6

Tata, P. R. and Rajagopal, J. (2016). Cellular plasticity: 1712 to the present day. *Current opinion in cell biology* 43, 46–54

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective* (Academic press)

Tinevez, J.-Y., Perry, N., Schindelin, J., Hoopes, G. M., Reynolds, G. D., Laplantine, E., et al. (2017). Trackmate: An open and extensible platform for single-particle tracking. *Methods* 115, 80–90

Tsai, H.-F., Gajda, J., Sloan, T. F., Rares, A., and Shen, A. Q. (2019). Usiigaci: Instance-aware cell tracking in stain-free phase contrast microscopy enabled by machine learning. *SoftwareX* 9, 230–237

Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., et al. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS computational biology* 12, e1005177

Vigilante, A., Laddach, A., Moens, N., Meleckyte, R., Leha, A., Ghahramani, A., et al. (2019). Identifying extrinsic versus intrinsic drivers of variation in cell behavior in human ipsc lines from healthy donors. *Cell reports* 26, 2078–2087

Wakita, S., Yamaguchi, H., Omori, I., Terada, K., Ueda, T., Manabe, E., et al. (2013). Mutations of the epigenetics-modifying gene (dnmt3a, tet2, idh1/2) at diagnosis may induce flt3-itd at relapse in de novo acute myeloid leukemia. *Leukemia* 27, 1044–1052

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing* (CRC press)

Wen, C., Miura, T., Voleti, V., Yamaguchi, K., Tsutsumi, M., Yamamoto, K., et al. (2021). 3deecelltracker, a deep learning-based pipeline for segmenting and tracking cells in 3d time lapse images. *Elife* 10, e59187

Woodworth, M. B., Girskis, K. M., and Walsh, C. A. (2017). Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nature Reviews Genetics* 18, 230–244

Yong, K. M. A., Li, Z., Merajver, S. D., and Fu, J. (2017). Tracking the tumor invasion front using long-term fluidic tumoroid culture. *Scientific reports* 7, 1–7

Yuan, J., Sheng, J., and Sims, P. A. (2018). Scope-seq: a scalable technology for linking live cell imaging and single-cell rna sequencing. *Genome biology* 19, 1–5

Zhang, J. Q., Siltanen, C. A., Liu, L., Chang, K.-C., Gartner, Z. J., and Abate, A. R. (2020). Linked optical and gene expression profiling of single cells at high-throughput. *Genome biology* 21, 1–11

Zou, D., Ma, L., Yu, J., and Zhang, Z. (2015). Biological databases for human research. *Genomics, proteomics & bioinformatics* 13, 55–63