A framework for evaluating predicted sperm trajectories in crowded microscopy videos

*David Hart[1], Kylie Cashwell[2], Anita Bhandari[1], Jayath Premasinghe[1], *Cameron Schmidt[2],

1 Computer Science Department, East Carolina University, Greenville, NC, USA
2 Biology Department, East Carolina University, Greenville, NC, USA

*Shared senior authorship: hartda23@ecu.edu, schmidtc18@ecu.edu

Abstract

Since the 1980's, semi-automated sperm motility analysis of phase contrast microscopy videos has been used to measure and categorize sperm motility patterns. Motility categories are determined from various kinematic parameters such as Curvilinear Velocity (VCL) and Beat Cross Frequency (BCF). These measures ultimately rely on the quality of the tracking for each individual sperm in the microscopy video. However, common approaches to sperm tracking require sample dilution and shortening the time window of observation (less than 1 to 2 seconds) to avoid tracking errors that occur when sperm cross paths. The post-ejaculatory lifespan of sperm can exceed several hours to days in some species, and long-term adaptive changes in motility pattern may be an important distinguishing factor for predictive modeling of sperm fertilizing competence. Improving the predictive value of computer assisted semen analysis will require accurate tracking of sperm trajectories over physiologically-relevant time scales and at the high cell densities typically found in semen. In this work, we identify a framework for accurately assessing the quality of sperm trajectory tracking that is independent of standard motility measures. We utilize cell tracking metrics adapted from the more common task of tracking adherent somatic cells and propose modifications based on the unique challenges of sperm video-microscopy. We also provide a small dataset of labeled microscopy videos to allow for future comparisons and developments.

Author summary

This report develops a computer vision framework to track individual sperm cells in crowded (high cell-density) microscopy videos, with the goal of improving automated analysis of sperm motility patterns. The task of accurately analyzing sperm movement is deceptively challenging due to the high rate of cell crossovers, an issue that has long impeded long-term tracking of sperm trajectories. Here, we introduce new evaluation metrics and provide a labeled dataset to serve as a baseline for future improvements.

Introduction

Computer aided semen analysis (CASA) is used to assess male fertility as a function of motility pattern, cell count, and/or morphology for a wide range of species in clinical, field, and research applications [1]. Inherent difficulty arises when analyzing sperm phase contrast microscopy videos due to the tendency of cells to cross paths, resulting

in misclassification of assigned trajectories by common tracking algorithms. For many years, tracking systems have relied on standardized kinematic parameters [2] to measure and classify motility properties. These kinematic parameters include curvilinear and straight-line velocity, beat cross frequency, maximum amplitude of lateral head displacement, linearity, etc. These parameters, however, are ultimately derivations of underlying tracking values and become erroneous if the tracking itself is inaccurate. In order to mediate the misclassification of trajectories, it is common for automated CASA systems to simply shorten the observation time window (video frame number) and dilute cells to well below physiological densities.

The sperm of many internally fertilizing species exhibit relatively long lifespans and undergo a series of post-ejaculatory physiological maturation steps, a process collectively known as capacitation [3]. In mammals, capacitative changes are sensitive to environmental chemistry and involve coordinated motility pattern changes, that are strongly correlated with fertility competence, such as the transition from progressive to hyperactive modes [3, 4]. Despite the known importance of motility pattern transitions during capacitation, the lack of accurate tracking algorithms has not permitted direct time-series analysis of sperm behaviors, leaving a significant gap in the understanding of how changes in the motility patterns of individual sperm correspond with fertility competence during this important maturation period. Nearly all available evidence related to post-ejaculatory motility pattern changes are derived from ensemble sampling of different sperm over time with measurement windows of less than one to two seconds for trajectory analysis [5]. Importantly, recent analyses in bovine sperm that used microfluidics devices for long term monitoring suggest that phase transitions in motility pattern that oscillate over periods of several minutes may enhance search efficiency for an egg [6, 7]. These observations are further supported by biophysical models that predict that phase transitions in motility pattern from high to low persistence may optimize the statistical success of random search in confined domains such as the female reproductive tract [8, 9].

The next generation of sperm analysis will require long-term tracking of each sperm in high density microscopy video. Though some reports have presented analysis techniques for sperm [10, 11], but these reports have focused only on tracking of isolated sperm. Robust tracking techniques will need to be developed for the long-term and high-density video. Given recent advances, these methods are likely to benefit from adoption of AI methods used in other video-microscopy applications such as object identification, segmentation, tracking, and pattern classification. Though these methods are common in the adherent cell tracking and lineage tracing spaces [12–15], there is currently no established baseline for determining the effectiveness of tracking results specific to the needs of sperm motility analysis. Additionally, common metrics and analyses established for general cell tracking tasks differ from the unique considerations that need to be addressed for sperm microscopy videos. Any newly proposed metrics must account for ambiguities in sperm segmentation, possible crossover of sperm from the 2D perspective, and other unique difficulties that are not common in other cell tracking problems such as signal loss in the thin pixel-sparse flagellar structure. Additionally, there are few high-quality and fully-labeled training datasets available, necessitating over-reliance on *a priori* assumptions for evaluating different tracking approaches.

In this report, we establish a framework for evaluating tracking algorithms for sperm microscopy videos. Common cell tracking metrics are adapted to work within the unique context of sperm video-microscopy. Given predicted and ground truth trajectories (represented in a simple .csv format), our bespoke Python program computes these metrics to evaluate the performance of the tracker independently of derived kinematic outputs. Establishing this framework will allow for continued

development of advanced computer vision approaches that will dramatically improve this important task compared with current approaches that generally depend on simple algorithms and calculations that are sensitive to arbitrarily chosen parameter values (e.g., the smoothing window size in frame-by-frame path averaging) [2]. As a baseline, we provide three fully-labeled crowded microscopy videos of washed human sperm and perform standard trajectory analysis using the common Trackpy library [16, 17]. We report the associated metrics to support iterative improvement by the computer vision community in the future. Additionally, we provide several examples using our own data and the publicly available VISEM-Tracking dataset [18] to demonstrate how poor tracking assumptions leads to erroneous kinematic outputs. We also demonstrate that improving tracking metrics improves kinematic outputs in these examples.

Materials and Methods

Participants

This study was approved by the East Carolina University Institutional Review Board and all subjects provided written informed consent prior to participating in research activities. Males ages 18-40 were included in the study. Donors with self-reported sexually transmitted infections within the previous six months were excluded. Baseline fertility was assessed via pre-screening health questionnaire and qualitative semen analysis following World Health Organization Guidelines [5].

Chemicals and Reagents

All chemicals and reagents used in this study were obtained from Sigma Aldrich (St. Louis, MO).

Human Sperm Isolation

Semen samples underwent liquefaction at 37 °C for 30 minutes in a 5% CO_2 incubator. Sperm were isolated from seminal plasma by differential centrifugation in 50% isotonic percoll. Sperm were then washed by centrifugation in modified Biggers, Whitten, and Whittingham media with 3.5% bovine serum albumin. Cells were resuspended in BWW medium for analysis. Cell counts were performed using a hemocytometer. Samples were stored in a 37 °C incubator with 5% CO_2 until imaging. All imaging was performed within 90 minutes of sample collection.

Phase Contrast Video Microscopy

Samples were diluted in BWW medium for imaging. Phase contrast microscopy was performed at 37 °C on a Zeiss Axervert.A1 compound light microscope fitted with a thermal stage (Carl Zeiss AG, Jena, Germany). Cell suspensions were diluted to approximately 8 million cells per mL. Diluted suspensions were added to a $20\mu m$ glass depth chamber slide (Hamilton Thorne, Beverly, MA). Phase contrast microscopy videos were obtained at 5X (0.25NA) objective or at 10X (0.25NA). Videos were recorded continuously for two minutes at nine frames per second using Zeiss Zen Lite acquisition software with accompanying time-series data collection module. Raw video files were stored as uncompressed .avi format. Videos for analysis were compressed using h.264 encoding.

Results

Tracking Quality's Effect on Standard Sperm Analysis Outputs

Semen characteristics are typically analyzed using a small set of standard kinematic parameters [5]. These outputs come from Computer-Aided Semen Analysis (CASA), typically performed using commercial hardware/software systems. CASA is often used to define and classify sperm motility and movement patterns [2]. CASA parameters are calculated based on the observed x and y coordinates of each sperm's nuclear centroid position in each video frame. Some common examples of kinematic parameters include:

- Curvilinear Velocity (VCL): Measures the total distance traveled along the sperm's actual curvilinear trajectory, expressed in microns per second.
- Straight-Line Velocity (VSL): Calculates the shortest distance between the starting and ending points of the trajectory, divided by the total time, also expressed in microns per second.
- Average Path Velocity (VAP): Represents the average velocity along a smoothed path of the sperm's trajectory, expressed in microns per second.

Sperm velocity distributions can reveal important insights into sperm movement patterns and functionality. For example, a bimodal distribution in a 5x magnification suggests the presence of distinct subpopulations, such as progressive, non-progressive, and hyperactivated sperm, while an exponential distribution at a 10x magnification can indicate that most sperm move slowly, with only a few reaching high velocities. Factors like sperm concentration, medium viscosity, and magnification can shift these various distributions and affect the estimated proportions of motile sperm. Analyzing and understanding these distributions helps explain how sperm movement is statistically distributed among cell populations and how selection pressures may favor certain motility traits. Kinematic parameters and their associated motility pattern classifications can be used in more complex analyses to determine if sperm are efficiently navigating their environment as a measure of fertilizing potential.

Calculating these distributions, however, ultimately depends on the reliability of the underlying tracking software. Poor tracking may give erroneous results in the calculated kinematic parameters. To demonstrate common tracking issues that may occur, we first define a uniform baseline tracking method and dataset for comparisons. We then provide three experiments that illustrate poor tracking assumptions and the effects on the calculated kinematic parameters.

Baseline Tracker and Dataset

For our baseline tracking pipeline, we use the commonly used Trackpy library [17]. For a baseline detector, we use the particle detector *locate* function from the Trackpy library [17]. For all experiments, we use default parameters of an 11-pixel diameter and 500 minmass. For a baseline tracker, we use particle tracker *link* function from the Trackpy library [17]. Additinally, we use default parameters of 25 for the search range and 3 for the frame memory. We empirically determined that this configuration gave the best overall performance (data not shown).

For the dataset, sperm samples were provided by healthy human donors. The samples were washed in BWW medium and diluted before recording the negative phase contrast microscopy videos at a 10x magnification. Additional details are outlined in the Materials and Methods section of this report.

To create the fully-labeled sperm trajectory data, the sperm microscopy videos were first fed through the baseline Trackpy workflow. The resulting tracks ultimately had

errors that needed to be corrected. To correct the mistakes, a tool was built using OpenCV [19] in Python that allows a human labeler to manually modify track estimates. Simple mouse and keyboard commands for adding, deleting, and merging trajectory data were included. The labeler program is included as part of our code repository provided at `github.com/CAS-ReproLab/Sperm_Object_Tracking`. A small dataset of three 30-second videos from different donors was hand labeled for our experiments.

Experiment 1: A Poorly Configured Tracker

Many tracking algorithms have default configurations that are used that control properties such as maximum linking distance between frames. The default parameters that are used are unlikely to be effective across the many types of sperm microscopy videos that are captured. However, researchers and technicians may not have access to meaningful ways to tune these parameters based on their specific data depending on the acquisition and analysis software configuration.

For this first experiment, we demonstrate the effect of a poorly tuned tracker on one of our microscopy videos. For illustrative purposes, we intentionally handicap our baseline approach using the same Trackpy detect and tracking methods, but using a search range of 7 pixels (the default parameter) instead of 25 which we have previously determined to be optimal. This greatly limits the algorithm's ability to track fast-moving sperm. Many inaccuracies in the tracking may occur from just one poorly optimized hyperparameter. In particular, because of the frequently dropped fast-moving sperm, the velocity distributions become highly skewed, drastically misrepresenting the true underlying behavior of the sperm population. An example of the difference between the poorly predicted velocity distribution and the actual velocity distribution is shown in Figure 1. This example is taken from video 3 in our dataset.

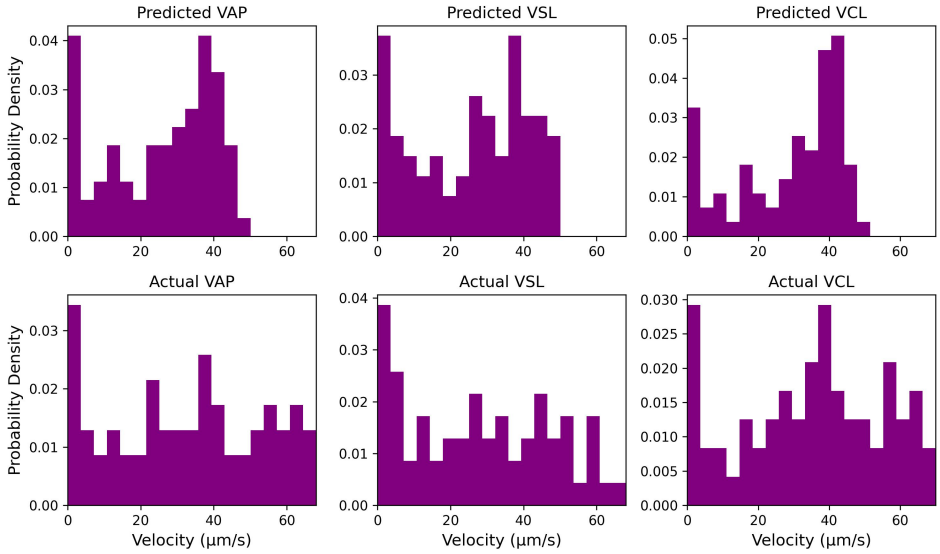


Fig 1. Velocity Distribution Comparison Example distributions for the Average Path Velocity (VAP), Straight Line Velocity (VSL), and Cuvilinear Velocity (VCL). If a particular tracking algorithm is unsuited to capture high-velocity or crowded trajectories (top), it will misrepresent the motility properties of the sperm population (bottom).

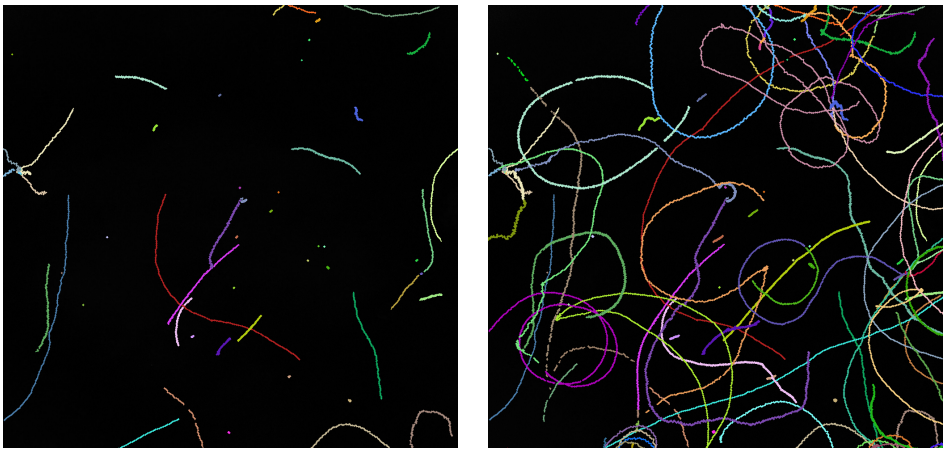
To quantify this difference, we use Earth Mover's Distance (EMD) between the two distributions. For this example, the EMD in VAP, VSL, and VCL between the poor prediction and the actual labeled data is 8.20, 5.43, and 7.97 respectively. In

comparison, the properly configured baseline has EMD values of 1.73, 1.48, and 1.89 respectively. This experiment demonstrates the need for more effective tools and metrics for evaluating the quality of the underlying tracking and improving confidence in predicted kinematic outputs.

Experiment 2: Short vs. Long Timescale Trajectories

It is a common practice to collect or aggregate sperm microscopy tracking data over very short time intervals (∼1-3 seconds) [2]. This practice is necessary to eliminate tracking errors that occur due to sperm moving near or over each other in the 2D image projection. Though this practice may alleviate some tracking errors, it comes at the cost of being unable to capture more complex, long timescale behaviors of sperm [6, 7]. Here, we analyze our labeled dataset which includes longer videos in comparison with standard CASA analysis (i.e., 30 seconds), enabling comparative analysis of both the short and long term movement behaviors among the same group of sperm.

Two second subsets of the thirty second videos were compared and kinematic parameters were calculated. The 30-second video provides much richer information for understanding more complex sperm movement behaviors. For example, on the short timescale, progressively moving sperm look very similar, however, on longer timescales oscillating patterns are observed with various phases and symmetries. As expected, the 2-second subsets fail to capture the underlying long-term behavior of the sperm. A visual comparison between short-term and long-term trajectories is provided in Figure 2.



2-second Trajectories 30-second Trajectories

Fig 2. Sperm Tracking Example Example visualization of sperm paths for a single video. In the short term, sperm trajectories present similar information (left). In contrast, long-term tracking highlights evolving differences in motility behaviors (right).

Additionally, we compared the output parameter distributions between the short-term and long-term trajectory ensembles in Figure 3. As indicated, shorter videos may not capture long-term behavior of sperm that would more accurately reflect their motility properties. This effect highlights limitations in the restricted timescales required by tracking issues in standard CASA, as well as challenges associated with the kinematic parameter calculations themselves. Because VAP, VSL, and VCL are generally represented using time-averages, the values tend to converge despite the visually apparent differences in motility behaviors. Though some useful additional information may be gleaned from analysis of kinematic parameter variance, this information is not part of the standard CASA analysis framework and is rarely reported [5].

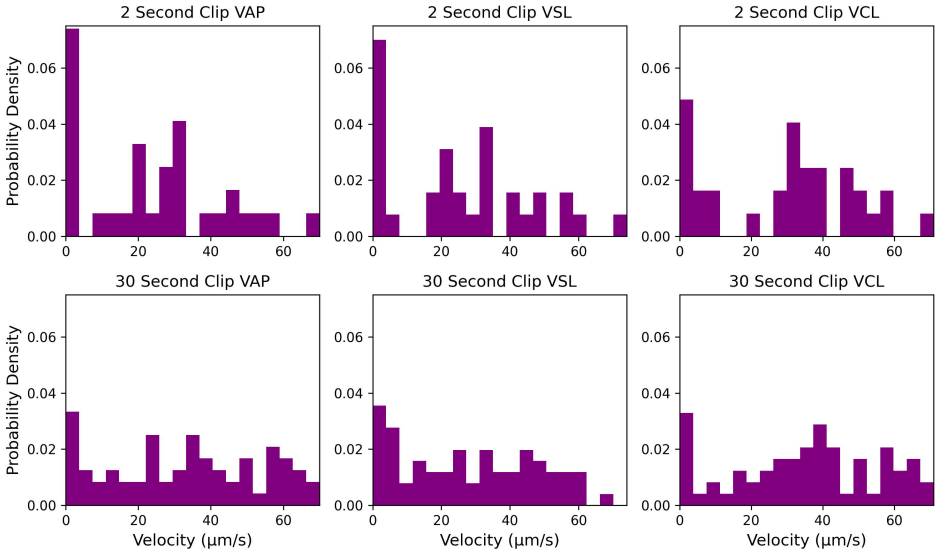


Fig 3. 2-second vs 30-second trajectory distributions. Short-term trajectory analysis tends to underestimate convergent parameter distributions that result from longer time-averaging.

Experiment 3: Bounding Boxes vs. Centroids

Multi-object tracking has seen a large amount of interest in the computer vision field [20, 21] and many AI based approaches are revolutionizing videomicroscopy analysis [15, 22, 23]. AI based methods are in the early stages of application in solving the sperm tracking problem [11, 24]. Several rely on the publicly available VISEM-Tracking dataset [18] which used manual annotation to define bounding boxes around sperm nuclei, similar to the YOLO frameworks [25] used frequently in other areas of computer vision.

As more of these bounding box trackers are used, an important issue arises. Since the bounding boxes are only approximately centered on the sperm nuclear centroid position, the path of the bounding boxes may not accurately portray the movement of the sperm, resulting in misleading positional data for sperm kinematic parameter calculations. A visualization of the bounding box movement and the sperm nuclear centroid movement is shown in Figure 4.

For this experiment, we used the VISEM-Tracking dataset and extracted head centroid placement in each bounding box. This was done using the baseline Gaussian Blob method provided by Trackpy followed by taking the presented x, y coordinates that fell within each bounding box. If zero or multiple coordinates were present in a box, the bounding box center was taken to avoid ambiguity, allowing for correction of nuclear centroid placement throughout the entire modified VISEM-Tracking dataset.

As expected, the bounding box center vs the head centroids give different velocity distributions. An example is given in Figure 5. The average EMD for each kinematic parameter across the 20 video dataset was also computed. The VSL EMD remained small with an average value of 1.1. This was expected because VSL is simply the time-averaged Euclidean distance from the starting location and ending location in the video segment. In this case, each sperm still approximately started and ended at the same positions. The VAP and VCL, however, saw dramatic changes between the corrected and non-corrected versions, with average EMD values of 11.8 and 53.0 respectively. These kinematic parameters were expected to be more susceptible to

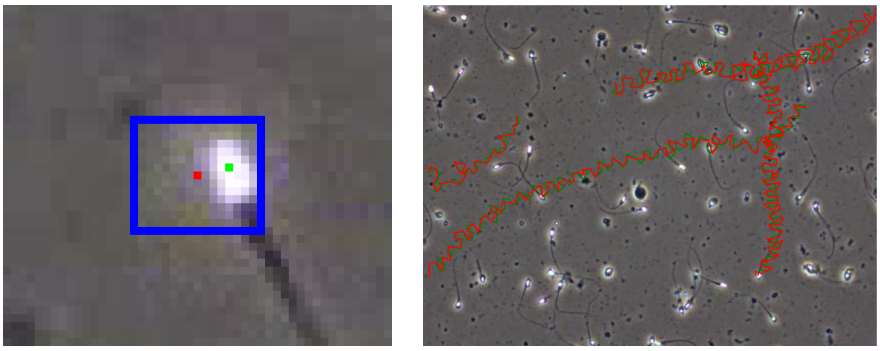


Fig 4. VISEM-Tracking corrections The publicly available and commonly used VISEM-Tracking dataset uses manually applied bounding boxes to indicate the location of sperm nuclei. However, taking the center of these bounding boxes (shown in red) does not necessarily correspond to the centroid of the sperm heads (shown in green), calculated using the Gaussian Blob method from the Python Trackpy library [17]. Trackpy was used to correct the nuclear centroid locations given by VISEM-Tracking, making the measured motility behavior much more accurate.

tracking method because even small inaccuracies in position accumulate as noise over time.

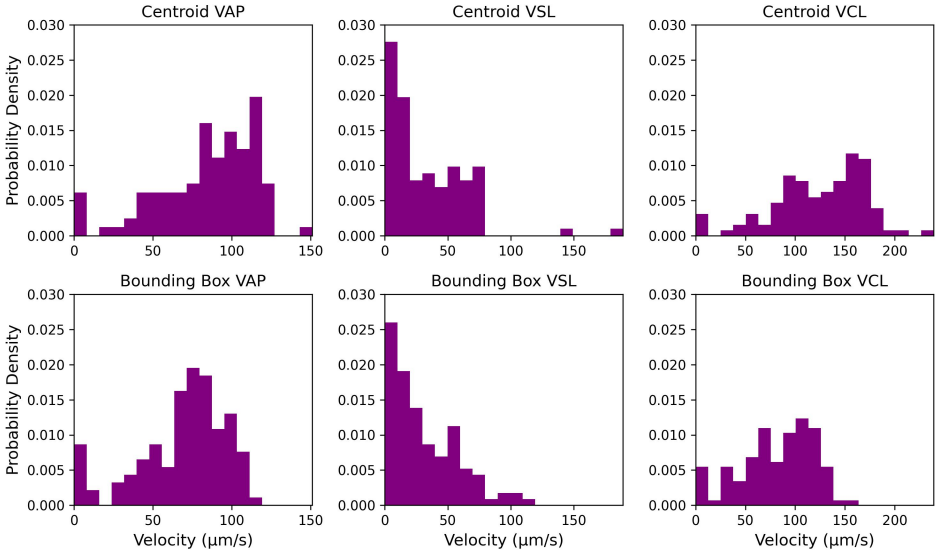


Fig 5. Velocity Distribution Comparison between Bounding Box Centers vs Trackpy Centeroids Using Trackpy to correct the centers of the bounding boxes leads to differences in the resulting kinematic parameter distributions.

Evaluating Sperm Detection and Tracking

As we have demonstrated with these simple examples, there is a substantive need to establish standardized metrics that can quantitatively compare and evaluate the performance of sperm tracking algorithms. We now turn to discussion of some candidate metrics and justify their use for the unique challenges associated with sperm video-microscopy.

A microscopy video is made up of individual image frames. The task of effectively tracking sperm throughout the entirety of the video can be separated into two distinct subtasks. First, the sperm must be detected in each individual frame. Second, individual sperm must be correctly identified and linked to corresponding sperm objects in subsequent frames. This process has many similarities to the general cell tracking problem in video-microscopy of adherent somatic cells. Thus, many of the metrics used to evaluate the effectiveness of cell tracking may be adapted for sperm tracking.

Of particular interest are metrics from two common benchmarks: The Multi-object Tracking Benchmark [13] and the Cell Tracking Challenge [12]. General multi-object tracking has context outside of the biological realm and well-established metrics have been used such as Multi Object Tracking Accuracy (MOTA) [26], Identification F1 Score (IDF1) [27], and Higher Order Tracking Accuracy (HOTA) [28]. These general benchmarks are calculated based on both the accuracy of the detections (true positives, false positives, and false negatives) and the number of correctly and incorrectly linked objects between frames. For cell tracking, metrics are often defined based on the spatio-temporal graph representation of the trajectories [29]. Each metric can be thought of as the number of steps needed to reconstruct the ground truth graph given the predicted graph. Metrics for the detection (DET) and linking (LNK) focus on accurate node and edge placement respectively, with these two being combined to determine the overall tracking performance (TRA). Another metric of interest is track fractions (TF) which describes the average length of trajectory that is correctly identified by a predicted trajectory. Additional details providing the mathematical definitions of each metric are provided in Appendix S1.

These common cell tracking metrics provide a foundation for evaluating and comparing the quality of sperm tracking algorithms. Notably, sperm tracking does present some unique differences compared with standard cell tracking that must be considered and properly accounted for. We will describe specific considerations for each of these differences in the following subsections.

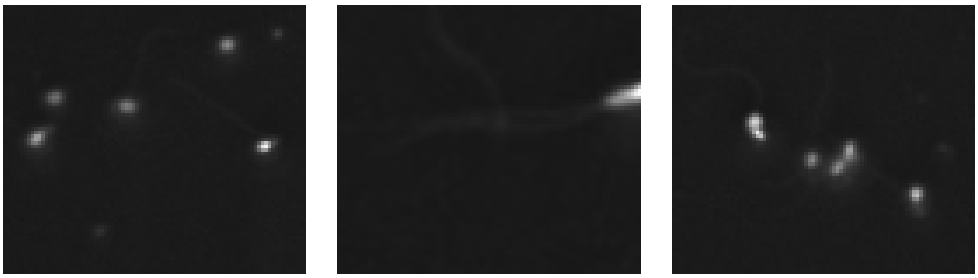
Unique Considerations for Evaluating Sperm Detection

Unlike most cells, sperm are incredibly thin, typically ranging from 0.8 to 0.01 μm along the flagellar length. Even at high magnifications and resolutions, the flagellar width may only occupy a few pixels at its width, but may be many pixels long. In most somatic cell-tracking detection metrics, it is assumed that a segmentation or a bounding box is provided during the detection process. This is needed to accurately compute the Jaccard similarity scores (Intersection over Union) and match predicted trajectory IDs to the reference IDs [12]. However, relying on these common similarity scores for sperm trajectories will likely lead to inconsistent association scores due to the ambiguity in labeling of the sperm pixels. Additionally, there are multiple confounding factors that can mislead the detection process when compared to somatic cell-tracking algorithms:

- Variations in brightness of pixels within a video can make some sperm more visible than others and make sperm outlines unclear.
- Dust, dead cells, and other debris in the video may appear like sperm, but ultimately do not move within the video.
- Sperm may enter and exit the frame, and sperm near boundaries of the video may be partially missing in shape.
- Sperm pixels can overlap as they swim above or below each other, making them appear as one or more sperm of irregular shape.

- Labeling programs may be inconsistent in the use of points, bounding boxes, or segmentations for identifying sperm, as well as differences in the labelers that are used to isolate the whole cells, flagellum, or nucleus only.

Examples of the unique computer vision challenges associated with phase contrast microscopy videos of sperm in Fig. 6



Brightness Variations | Border Cutoff | Overlap

Fig 6. Sperm Detection Issues Example issues with detection of sperm. Variations in brightness and dead cells in view (left), partially missing sperm near boundaries (middle), overlap and irregular shapes (right).

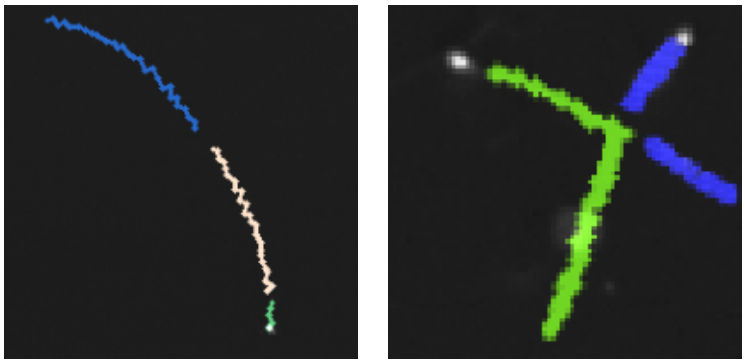
To combat these issues and establish correct predicted-to-actual ID linking for all types of sperm trajectory videos, it is reasonable to represent sperm detection using a single point rather than a segmentation or bounding box. The x,y coordinate is approximately centered on the head of the sperm which we call the sperm *centroid* (since most of the mass is located in the sperm nucleus). Localization between the actual and predicted point can be quickly computed using the Hungarian algorithm with a distance cutoff. The standard cell tracking metrics can be then used in the same fashion that we outlined in preceding sections.

Unique Consideration for Evaluating Sperm Tracking

Once sperm have been identified in all frames of the video, they must be properly linked together to make a complete tracking of individual sperm. Some approaches may also rely on detections in previous frames to update detections in future frames and link in a forward fashion. Regardless of the approach, the goal is to assign each sperm with a unique identifier that it keeps throughout the whole microscopy video. Doing so enables drawing the trajectory path of each sperm and calculating the kinematic parameters.

Tracking sperm presents unique challenges compared to most cell tracking problems. First, many sperm move very quickly compared to spatial resolution of most microscopy videos. Even during high FPS recordings, some sperm may move multiple pixels per frame. For tracking algorithms, this can cause frequent dropout and relabeling of sperm in subsequent frames. Second, as mentioned previously, sperm paths can cross over each other as one sperm swims above another. This crossover can cause an incorrect path labeling, where the two sperm swap trajectory labels in subsequent frames. These two issues are illustrated in Figure 7.

In addition to these challenges, immotile sperm and debris may be of interest for automated counting, but are not of interest for more computationally intensive motility analysis and may mislead performance benchmarking of the tracking algorithm. Immotile sperm are easier to identify in each subsequent frame compared with motile sperm and may disproportionally inflate tracking metrics for a given algorithm. Additionally, some nonmoving dust and debris may look like sperm, and be incorrectly labeled in reference datasets. Finally, failure to capture long-timescale movement



Dropout Crossover

Fig 7. Sperm Tracking Issues Example visualization of sperm paths issues. A single sperm may appear to dropout for a few frames, causing the algorithm to think that new sperm are appearing in and out of view (left). Two sperm that intersect may cause the path of one sperm to be assigned to another, causing a straightline path to be incorrectly marked as jagged (right).

patterns is another major challenge for sperm motility analysis. Improved analysis of long-timescale behavior is essential to better understand the complex process of capacitation, because it involves heterogeneous context-dependent changes in motility pattern that are either not captured, or are mislead, by current motility analysis algorithms [3, 4, 30, 31].

Effective tracking algorithms must be able to account for these confounding factors and capture objects moving at variable speeds over longer timescales. When labeled ground truth data is present, the standard tracking metrics are used in this report. However, we also propose an additional filtered version of these metrics which removes non-moving objects and very-short trajectories. For both the ground truth and predicted paths, a pre-processing step was used to remove all trajectories that do not spatially extend past a small radius of (ϵ). This removes all trajectories for nonmoving or nearly nonmoving sperm, giving more weight on the metric to faster moving sperm and longer-term trajectories.

Evaluation on the Dataset

With the evaluation process established in the previous section, the detection and tracking metrics are presented for a baseline tracking method with a small dataset of three microscopy videos that were collected from separate healthy human sperm donors. For comparison, we also present results from a re-analysis of the publicly available VISEM-Tracking dataset [18]. In the following subsections, we present the outputted metrics for the baseline Trackpy method. Additionally, we demonstrate how the metrics can indicate the reliability of standard motility outputs.

Metric Computations

For our dataset containing videos with manually corrected trajectories, we use the baseline Trackpy library and report the common cell tracking metrics for three videos. The ID association process described previously was used. We compared results with and without filtering the non-moving objects in the calculations. The results are shown in Table 1. The results for VISEM-Tracking are given in Table 2 and Table 3. For all metrics, a higher score indicates better performance and 1.0 is the highest score. The outputted metrics serve as a baseline for future tracking algorithms to compare against.

Table 1. Detection and Tracking Results

	Video 1		Video 2		Video 3	
Metrics	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
DET	0.970	0.960	0.979	0.976	0.967	0.924
LNK	0.674	0.642	0.679	0.845	0.635	0.607
TRA	0.932	0.918	0.940	0.959	0.924	0.882
TF	0.891	0.881	0.923	0.920	0.841	0.804
MOTA	0.789	0.737	0.822	0.790	0.840	0.818
IDF1	0.816	0.782	0.867	0.846	0.839	0.795
HOTA	0.831	0.801	0.878	0.860	0.856	0.819

Table notes: Common multi-object and cell-tracking metrics computed for the three videos. For all metrics, a higher score indicates better performance and 1.0 is the highest score. The baseline Trackpy approach is measured against the human-labeled and corrected data. Additionally, the results are also compared before and after filtering the predictions and labels to exclude nonmoving objects, better capturing the trackers ability to follow fast moving sperm.

Table 2. Detection and Tracking Results - VISEM-Tracking (Bbox)

Video	DET	LNK	TRA	TF	MOTA	IDF1	HOTA
11	0.766	0.881	0.781	0.605	-0.366	0.472	0.528
12	0.773	0.748	0.769	0.392	0.583	0.526	0.555
13	0.854	0.855	0.854	0.570	0.779	0.770	0.760
14	0.642	0.974	0.685	0.748	-2.374	0.331	0.425
15	0.668	0.882	0.696	0.827	-1.269	0.422	0.499
19	0.358	0.846	0.421	0.449	-4.240	0.176	0.282
21	0.808	0.918	0.822	0.653	-0.251	0.534	0.574
22	0.681	0.958	0.717	0.764	-1.835	0.334	0.434
23	0.133	0.867	0.228	0.880	-6.469	0.210	0.342
24	0.821	0.844	0.824	0.511	0.481	0.599	0.633
29	0.507	0.779	0.543	0.545	-2.087	0.310	0.411
30	0.801	0.918	0.816	0.716	-0.289	0.514	0.571
35	0.669	0.798	0.686	0.330	-0.958	0.215	0.301
36	0.810	0.897	0.821	0.711	-0.057	0.440	0.511
38	0.906	0.937	0.910	0.693	0.569	0.710	0.744
47	0.534	0.841	0.574	0.458	-2.505	0.295	0.374
52	0.693	0.874	0.716	0.739	-0.992	0.402	0.476
54	0.787	0.921	0.804	0.676	-0.569	0.486	0.550
60	0.780	0.849	0.789	0.761	0.089	0.617	0.640
82	0.832	0.880	0.838	0.782	0.313	0.653	0.685

Table notes: Resulting metrics of our baseline approach when run on the VISEM-Tracking dataset using the original bounding box centers as reference.

Relating Tracking to Motility Outputs

The presented metrics can determine the effectiveness of the tracking, thereby improving on the metrics also improves confidence in the calculated motility parameters. To demonstrate this, we conduct the following analysis.

First, we present the detection and tracking metrics, comparing the suboptimal

Table 3. Detection and Tracking Results - VISEM-Tracking (Centroid)

Video	DET	LNK	TRA	TF	MOTA	IDF1	HOTA
11	0.852	0.907	0.859	0.718	-0.275	0.496	0.558
12	0.950	0.866	0.939	0.722	0.819	0.599	0.669
13	0.965	0.900	0.957	0.818	0.931	0.835	0.868
14	0.645	0.977	0.688	0.855	-2.373	0.343	0.427
15	0.758	0.976	0.786	0.851	-1.311	0.432	0.507
19	0.477	0.884	0.530	0.555	-4.095	0.195	0.306
21	0.855	0.934	0.865	0.780	-0.229	0.532	0.588
22	0.705	0.974	0.740	0.824	-1.883	0.335	0.438
23	0.152	0.992	0.261	0.842	-7.420	0.211	0.343
24	0.901	0.728	0.878	0.678	0.589	0.645	0.689
29	0.727	0.993	0.762	0.918	-1.683	0.401	0.498
30	0.857	0.965	0.871	0.884	-0.285	0.526	0.597
35	0.799	0.855	0.806	0.530	-0.727	0.265	0.368
36	0.861	0.806	0.854	0.729	-0.012	0.441	0.521
38	0.950	0.968	0.953	0.855	0.624	0.720	0.767
47	0.636	0.956	0.678	0.825	-2.417	0.317	0.413
52	0.760	0.873	0.774	0.793	-0.917	0.419	0.493
54	0.843	0.962	0.858	0.805	-0.502	0.499	0.569
60	0.907	0.977	0.916	0.926	0.231	0.667	0.695
82	0.919	0.923	0.920	0.867	0.397	0.684	0.722

Table notes: Resulting metrics of our baseline approach when run on the VISEM-Tracking dataset using the corrected centroids as reference.

(default) configuration to the optimal configuration for the Trackpy approach after filtering. These results are presented in Table 4. The data support the conclusion that the default configuration significantly under-performs in every metric, indicating that the metrics adequately represent the ability to track fast-moving objects.

Table 4. Detection and Tracking Ablation

	Video 1		Video 2		Video 3	
Metrics	Suboptimal	Optimal	Suboptimal	Optimal	Suboptimal	Optimal
DET	0.624	0.960	0.634	0.976	0.713	0.924
LNK	0.414	0.642	0.585	0.845	0.483	0.607
TRA	0.596	0.918	0.627	0.959	0.683	0.882
TF	0.514	0.881	0.588	0.920	0.638	0.804
MOTA	0.509	0.737	0.444	0.790	0.633	0.818
IDF1	0.564	0.782	0.648	0.846	0.669	0.795
HOTA	0.618	0.801	0.686	0.860	0.707	0.819

Table notes: Comparison of tracking and detection metrics between the suboptimal configuration and the optimal configuration on the filtered data. The suboptimal configuration is the default configuration for Trackpy parameters which fails in tracking fast-moving objects, showcasing the metrics effectiveness in the sperm tracking task.

Second, we present the computed motility outputs for the predictions from the suboptimal and optimal configurations, as well as the actual outputs from the hand labeled data. These distributions are shown in Figure 8. As can be seen, these motility

outputs are drastically affected by the tracking parameters. Specifically, the suboptimal approach can not track the trajectories of fast moving sperm, greatly skewing the resulting distributions, ultimately affecting the predicted motility categories.

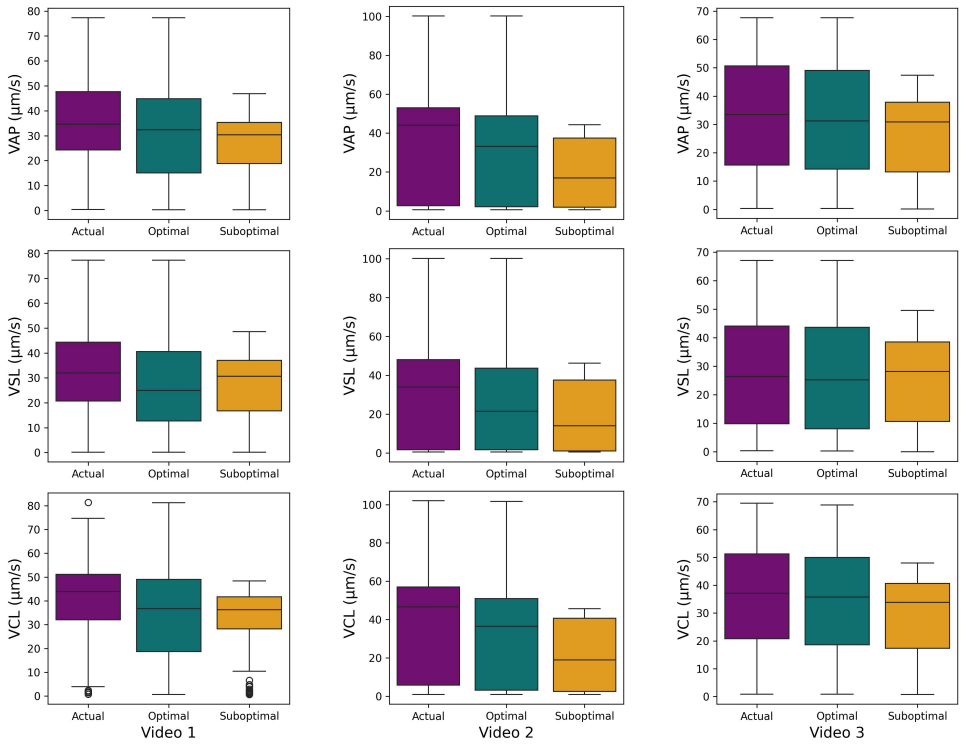


Fig 8. Motility Outputs The outputs for VAP, VSL, and VCL on the suboptimal configuration, optimal configuration, and actual hand labeled and filtered data for Video 1, Video 2, and Video 3.

We also output Earth Mover's Distance (EMD) scores for each kinematic parameter for the suboptimal and optimal approach when compared to the ground truth. These scores are provide in Table 5. As expected, the handicapped approach has larger EMD scores, indicating its inability to provide accurate kinematic parameters. Overall, this indicates the wide variation in possible tracking consistency between algorithms and configurations and the need for open tracking metrics to evaluate the reliability of these motility values.

Table 5. Earth Mover's Distance

	Video 1		Video 2		Video 3	
Parameter	Suboptimal	Optimal	Suboptimal	Optimal	Suboptimal	Optimal
VAP	4.14	1.57	4.91	1.09	2.23	0.47
VSL	2.79	1.30	3.56	0.83	1.48	0.41
VCL	4.41	1.76	5.09	1.15	2.17	0.52

Table notes: Comparison of Earth Mover's Distance between the suboptimal and optimal approach when compared to the actual hand labeled data.

Discussion

Recent advances in computer vision are revolutionizing fields that rely on automated image analysis such as healthcare, self-driving vehicles, and manufacturing. Adaptation of computer vision algorithms and machine learning models from these fields has dramatically benefited microscopy applications in cell biology due to improved cell location, segmentation, and tracking but has been largely restricted to adherent somatic cells in subculture. Though CASA has proven extremely useful in basic research and agricultural breeding programs, its use in clinical andrology laboratories remains limited because it does not yet outperform manual semen analysis by trained lab technicians. Here, we identified several key issues with current sperm tracking algorithms and kinematic calculations employed in CASA. One major barrier to improvement is a lack of standardized metrics to compare tracking performance that conform to the unique challenges inherent in phase contrast microscopy of motile sperm. We suggest several methods, borrowed from other applications in computer vision, to facilitate quantitative comparison of sperm tracking program performance. Additionally, we highlight important issues that arise from inaccurate labeling methods in common training datasets, such as the bounding box method used in the publicly available VISEM-tracking dataset which unintentionally resulted in significant short timescale noise. [18]. Our report highlights two major issues: 1) CASA could be much more powerful than it currently is but requires specialized metrics to compare performance among algorithms and models, and 2) there is a conspicuous need for high quality training datasets with accurate labels.

Enhanced CASA holds significant potential for improving fertility analysis, largely because of its automation and standardization, which remains problematic for analyses that rely on the subjective performance of laboratory technicians [32, 33]. Despite its potential, CASA has historically underperformed in its predictive value for assisted reproductive therapies and diagnostic tests, often necessitating additional functional assays [34]. Currently, CASA is most useful for relatively extreme phenotypes such as low sperm counts (oligozoospermia), completely immotile sperm (asthenozoospermia), or obviously malformed sperm (teratozoospermia) but lacks discriminatory ability when sperm function or morphology is subtle or context dependent [35]. For example, sperm motility patterns in non-viscous fluids differ significantly from more physiologically relevant viscous or visco-elastic fluids [6]. Sperm that appear to move normally in the non-viscous fluids recommended in the WHO standard, may fail to move appropriately in visco-elastic fluids characteristic of the female reproductive tract or the vestments of the egg.

Another important limitation to CASA analysis is that default tracker configurations may significantly alter results, as demonstrated in this report in experiment 1. Sperm are typically classed into motility types (e.g., progressive, hyperactive) for fertility analysis, but apparent motility profile of a sperm sample can change substantially depending on algorithmic parameters such as search radius, maximum frame gap for object loss, or trajectory linking rules [17]. These choices determine whether a moving object is linked into a continuous track or fragmented into multiple short tracks, which in turn alters the derived kinematic parameters and even whether a sperm trajectory is accurately assigned. Tracking program defaults can be a significant strength of CASA, because they enable far more rigorous standardization than lab technicians can provide. However, if applied without quantitative measures that quantify program performance, two laboratories analyzing the same dataset may arrive at different results simply due to differences in tracker configuration [36]. A general lack of transparency regarding algorithmic configuration limits CASA reproducibility and undermines comparisons across studies. Addressing this issue will require not only standardized reporting of tracker settings but also the development of

sperm-specific benchmarks that define biologically appropriate parameter ranges, analogous to hyperparameter optimization in broader machine learning practice [37].

Another important source of variability in CASA output is the timescale of analysis. Most CASA implementations rely on relatively short tracks, often spanning less than one to two seconds of observation [5]. This introduces two major problems: 1) sperm motility is summarized via ensemble averaging despite observation over a period of time that represents only a miniscule fraction of the time required to fertilize an egg, and 2) this practice produces noisy estimates of kinematic parameters because the variance is computed for the sperm sample, rather than individual trajectory time-series [2]. Additionally, short timescale observations may be computationally convenient, but are biologically misleading, since sperm motility is highly dynamic and characterized by transitions between progressive movement, periods of quiescence, and hyperactivation that unfold over much longer timescales [6, 7]. Capturing the biological variation in these transitions may yield more stable kinematic measures, but doing so also presents practical challenges: sperm frequently cross paths or move out of focus in dense samples leading to limited track continuity. As a result, CASA systems often prioritize feasibility over biological fidelity by diluting sperm samples or by tracking for shorter periods of time - compromises that likely contribute to the limited predictive power for more subtle fertility outcomes. These timescale dependent tracking limitations are closely intertwined with the methods used for cell detection, since reliable long-term tracking ultimately depends on the accuracy and stability of the underlying detection step. Improved methods are an absolute necessity for CASA to reach it full predictive potential.

The accuracy of CASA output ultimately depends on the method of detection, since tracking quality cannot exceed the fidelity of the initial object identification. Most current systems and training datasets, such as VISEM-tracking, rely on bounding-box based approaches [18]. While computationally efficient, these methods introduce substantial error: bounding boxes are sensitive to orientation changes and background noise, and inaccurately placed bounding box centers creates artifactual noise in kinematic parameter calculations as demonstrated in experiment 3 of this report. These limitations are particularly problematic for the phase-contrast imaging typical of CASA due to 'halo' effects which complicate bounding box placement. In dense samples, overlapping trajectories further exacerbate misidentification, leading to spurious tracks and inflated motility metrics. Overcoming these challenges will require more sophisticated detection methods, potentially combining deep learning–based segmentation with physics-informed constraints on trajectory smoothness and flagellar-beat periodicity. This approach has been applied previously, but remains relatively low throughput and lacks an open-source codebase limiting its application [38]. Framing detection, tracking, and timescale of analysis as interdependent problems will provide a clear path toward next-generation CASA systems with high predictive value for use in clinical andrology and elsewhere.

Several limitations of this study should be acknowledged. First, our labeled dataset is relatively small and is simply intended as a proof-of-principle to highlight the importance of long timescale tracking. The ground truth trajectories were derived by manual correction of Trackpy outputs, an approach that introduces potential bias from the initial algorithm and emphasizes head centroids rather than full cell morphology. Imaging constraints, including relatively low frame rates (necessitated by tracking duration), low magnification, and phase-contrast artifacts also contribute noise that propagates into both kinematic parameters and tracking metrics. In addition, while we adapted well-established multi-object tracking metrics and distributional comparisons, these measures may not fully capture sperm-specific features such as flagellar beat periodicity or long-term motility pattern phase transitions linked to capacitation.

Despite these limitations, the framework presented here provides a transparent and extensible foundation highlighting the importance of releasing open data and code in standardized formats, our aim is to enable iterative improvements by the community and to encourage future studies that address these challenges through larger, more diverse datasets, multimodal imaging, and metrics tailored specifically to the unique challenges of sperm video-microscopy.

Conclusion

Taken together, these results point toward the need for a new framework in CASA that treats detection, tracking configuration, and timescale of analysis in an interconnected way. The limitations highlighted in this report such as variability in tracker configuration, short observation windows, and imprecise detection methods are fundamental barriers to reproducibility and predictive value of CASA. Overcoming them will require three key advances: 1) the establishment of standardized benchmarks and biologically relevant metrics to compare algorithms, 2) the creation of high-quality accurately labeled training datasets that reflect the unique challenges of sperm morphology and imaging artifacts, and 3) the development of integrated pipelines that combine machine learning approaches with physics-informed constraints on sperm motion. Such a framework would transform CASA from a descriptive tool that reports kinematic averages into a predictive system capable of informing clinical decision-making and improving outcomes in assisted reproduction. By adopting open-source, transparent codebases and fostering reproducible methods, CASA could finally deliver on its promise to standardize semen analysis in the same manner that computer vision has revolutionized diagnostics in radiology and other fields.

With this goal in mind, we make our code publicly available for others to use and run their methods on. The source code can be found at `github.com/CAS-ReproLab/Sperm_Object_Tracking`. For a given video, trajectory data can be stored in a single comma-separated-values file (.csv). This .csv file contains x,y coordinates, a sperm label, and frame in each row. This provides an efficient, human-readable, and unambiguous way of storing predictions and/or ground truth data. Once the prediction data is provided in a standard format, the metrics can be computed to evaluate its performance as previously described.

In future work, we aim to expand our comparisons to additional datasets such as MIaMIA-SVDS [39]. Each dataset stores the trajectory information in a different way, so tools will need to be develop to convert between representations. Additionally, though the segmentation is not ultimately needed for trajectory tracking, obtaining full sperm outlines can be helpful for biological outputs such as flagellar waveform analysis. Additional baselines and metrics can be established to properly account for these use cases.

Supporting information

S1 Appendix. Common Cell Tracking Metrics

For convenience of the reader, we provide a description of the common multi-object and cell tracking metrics. Relevant works are cited to provided sources for additional details.

Association Accuracy and Target Effectiveness

Association Accuracy (AA) and Target Effectiveness (TE) have been a standard metric within the cell tracking literature. AA is a measure of the number of correctly predicted cell associations between frames divided by the total number of associations in the

ground truth data. TE is a measure describing the number of frames a ground truth track is correctly followed by the predicted tracks. These two metrics are still used frequently to compare modern cell tracking approaches [40–42].

MOTA, IDF1, HOTA, and CHOTA

Multi-object tracking has many applications outside the realm of cell tracking. Standard metrics have been used for comparing methods on various tasks, such as in the Multi-object Tracking Benchmark [13]. These metrics have also been adopted as cell tracking for comparison, such as in the Cell Tracking With Mitosis Detection Dataset Challenge (CMTC) [43].

MOTA is a measure of the total detection coverage, and is calculated as

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \tag{1}$$

where t is given frame of the video, GT is the number of objects in the ground truth of the frame, FN is the number of false negatives (missed detections), FP is the number of false positives (spurious detections), and $IDSW$ is number of correct detections that are labeled with a different ID than the track they were associated with in the previous frame. The best MOTA score is 1.0 and scores can go into the negatives.

IDF1 is a measure of the consistency of labels across frames of a video. It uses the Hungarian algorithm to optimize how long each ground truth track was connected to a single ID trajectory. After this matching, the metric is calculated as

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{2}$$

where IDTP is the total length and number of correctly labeled tracks for all frames, IDFP is the number of predicted tracks that don't match any ground truth, and IDFN is ground truth tracks that weren't predicted.

Recently, a new metric has been proposed called HOTA that better balances detection and association pieces [28]. A variant of this metric has been proposed specifically for cell tracking called CHOTA [44], but CHOTA was designed specifically to measure mitosis splitting events which do not occur in sperm.

DET, LNK, TRA, and TF

The Cell Tracking Challenge [12] has adopted standard evaluation methodology that can be used across many types of cells and data types. Many of these metrics rely on the Acyclic Oriented Graph Matching (AOGM) measure that is presented in [29]. This measure determines the number of steps needed to correct a graph representation of the predicted tracks to match the graph representation of the ground truth tracks. It is calculated as

$$AOGM = w_{NS}NS + w_{FN}FN + w_{FP}FP + w_{ED}ED + w_{EA}EA + w_{EC} \tag{3}$$

In this formulation, FN and FP represent false negatives and false positive in detections. ED, EA, and EC represent different kinds of incorrect tracking associations (edges in the graph). w represents a weight applied to each measure based on how difficult it would be to correct an error through a manual modification. The standard values are $w_{NS} = 5, w_{FN} = 10, w_{FP} = 1, w_{ED} = 1, w_{EA} = 1.5, w_{EC} = 1$. NS represents mitosis events where cell splits, so this term can be ignored for sperm tracking.

The DET, LNK, and TRA metrics derive from this measure. As an example, TRA is calculated as

$$TRA = 1 - min(AOGM, AOGM_0)/AOGM_0 \quad (4)$$

where $AOGM$ is the value of the metric for the predicted tracks, and $AOGM_0$ is the value of the metric for an empty reference (i.e. creating all tracks from scratch). This forces the measure to be between 0-1. DET is calculated in the same way but with no weight assigned to the edge calculations ($w_{ED} = w_{EA} = w_{EA} = 0$). Likewise, LNK is calculated with no weight assigned to the node calculations ($w_{NS} = w_{FN} = w_{FP} = 0$).

The Cell Tracking Challenge [12] also presents other biological plausibility metrics, many related to cell mitosis. Track Fractions (TF) is of particular interest for the sperm tracking problem. For each ground truth track, the predicted track that matches for the most frames is determined. Then, the number of frames of that are correctly matched is divided by the total frames for the ground truth track. These are then averaged for all ground truth sperm trajectories.

Jaccard Similarity/IoU/SEG

Whether as a bounding box or a pixel segmentation, the accuracy of the overlap between predicted and ground-truth regions is usually calculated using Jaccard Similarity. This is also known as Intersection over Union (IoU). This is calculated as

$$IoU = \frac{R \cap S}{R \cup S} \quad (5)$$

where R is the ground truth segmentation and S is the predicted segmentation. The SEG measure from the Cell Tracking Challenge [12] is also based-on this score. In this work, we do not provide segmentation results do due the difficulty of manually labeling each pixel. Thus, this metric is not presented in this work.

References

1. Gallagher MT, Smith D, Kirkman-Brown J. CASA: tracking the past and plotting the future. Reproduction, Fertility and Development. 2018;30(6):867–874.
2. Alquézar-Baeta C, Gimeno-Martos S, Miguel-Jiménez S, Santolaria P, Yániz J, Palacín I, et al. OpenCASA: A new open-source and scalable tool for sperm quality analysis. PLoS computational biology. 2019;15(1):e1006691.
3. Gervasi MG, Visconti PE. Chang's meaning of capacitation: A molecular perspective. Molecular Reproduction and Development. 2016;83(10):860–874.
4. Suarez SS. Control of hyperactivation in sperm. Human reproduction update. 2008;14(6):647–657.
5. Organization WH, et al. WHO laboratory manual for the examination and processing of human semen. World Health Organization; 2021.
6. Zaferani M, Abbaspourrad A. Biphasic Chemokinesis of Mammalian Sperm. Physical Review Letters. 2023;130(24). doi:10.1103/PhysRevLett.130.248401.
7. Zaferani M, Baouche Y, Lago-Alvarez Y, Pandya A, Cheong SH, Petry S, et al. Sperm hyperactivation drives a circling-and-wandering migration strategy. bioRxiv. 2025; p. 2025–07.
8. Brisard BM, Cashwell KD, Stewart SM, Harrison LM, Charles AC, Dennis CV, et al. Modeling diffusive search by non-adaptive sperm: Empirical and computational insights. PLOS Computational Biology. 2025;21(4):e1012865.

9. Bier M, Majka M, Schmidt C. A Two-State Random Walk Model of Sperm Search on Confined Domains. Entropy. 2025;27(5):539.

10. Hansen JN, Rassmann S, Jikeli JF, Wachten D. SpermQ–A Simple Analysis Software to Comprehensively Study Flagellar Beating and Sperm Steering. Cells. 2019;8(1). doi:10.3390/cells8010010.

11. Fujii T, Nakagawa H, Takeshima T, Yumura Y, Hamagami T. Automated Sperm Assessment Framework and Neural Network Specialized for Sperm Video Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024. p. 7675–7684.

12. Maška M, Ulman V, Delgado-Rodriguez P, Gómez de Mariscal E, Necasova T, Guerrero Peña FA, et al. The Cell Tracking Challenge: 10 years of objective benchmarking. Nature Methods. 2023;20:1–11. doi:10.1038/s41592-023-01879-y.

13. Dendorfer P, Osep A, Milan A, Schindler K, Cremers D, Reid I, et al. MOTChallenge: A Benchmark for Single-camera Multiple Target Tracking. International Journal of Computer Vision. 2020; p. 1–37.

14. Bragantini J, Lange M, Royer L. Large-Scale Multi-Hypotheses Cell Tracking Using Ultrametric Contours Maps. 2023;.

15. Bragantini J, Theodoro I, Zhao X, Huijben TA, Hirata-Miyasaki E, VijayKumar S, et al. Ultrack: pushing the limits of cell tracking across biological scales. bioRxiv. 2024; p. 2024–09.

16. Crocker JC, Grier DG. Methods of digital video microscopy for colloidal studies. Journal of colloid and interface science. 1996;179(1):298–310.

17. Allan DB, Caswell T, Keim NC, van der Wel CM, Verweij RW. soft-matter/trackpy: v0.6.4; 2024. Available from: `https://doi.org/10.5281/zenodo.12708864`.

18. Thambawita V, Hicks SA, Storås AM, Nguyen T, Andersen JM, Witczak O, et al. VISEM-Tracking, a human spermatozoa tracking dataset. Scientific Data. 2023;10(1):260.

19. Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000;.

20. Cheng HK, Wug Oh S, Price B, Schwing A, Lee JY. Tracking Anything with Decoupled Video Segmentation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 1316–1326.

21. Cheng HK, Oh SW, Price B, Lee JY, Schwing A. Putting the Object Back into Video Object Segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. p. 3151–3161.

22. Stringer C, Wang M, Michaelos N, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nature Methods. 2021;18:113–116.

23. O'Connor OM, Dunlop MJ. Cell-TRACTR: A transformer-based model for end-to-end segmentation and tracking of cells. PLOS Computational Biology. 2025;.

24. Zhang C, Zhang Y, Chang Z, Li C. Sperm YOLOv8E-TrackEVD: A Novel Approach for Sperm Detection and Tracking. Sensors. 2024;24(11):3493. doi:10.3390/s24113493.

25. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 779–788.

26. Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing. 2008;2008:1–10.

27. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer; 2016. p. 17–35.

28. Luiten J, Osep A, Dendorfer P, Torr P, Geiger A, Leal-Taixé L, et al. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision. 2021;129:548–578.

29. Matula P, Maška M, Sorokin DV, Matula P, Ortiz-de Solórzano C, Kozubek M. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. PloS one. 2015;10(12):e0144959.

30. Eisenbach M. Sperm navigation in humans: a concerted action of multiple means. Communications Biology. 2025;8(1):923.

31. Darszon A, Nishigaki T, López-González I, Visconti PE, Treviño CL. Differences and similarities: the richness of comparative sperm physiology. Physiology. 2020;35(3):196–208.

32. Jørgensen N, Auger J, Giwercman A, Irvine D, Jensen T, Jouannet P, et al. Semen analysis performed by different laboratory teams: an intervariation study. International journal of andrology. 1997;20(4):201–208.

33. Barratt CL, Wang C, Baldi E, Toskin I, Kiarie J, Lamb DJ, et al. What advances may the future bring to the diagnosis, treatment, and care of male sexual and reproductive health? Fertility and sterility. 2022;117(2):258–267.

34. Wang C, Swerdloff RS. Limitations of semen analysis as a test of male fertility and anticipated needs from newer tests. Fertility and sterility. 2014;102(6):1502–1507.

35. Finelli R, Leisegang K, Tumallapalli S, Henkel R, Agarwal A. The validity and reliability of computer-aided semen analyzers in performing semen analysis: a systematic review. Translational Andrology and Urology. 2021;10(7):3069.

36. Hu Y, Lu J, Shao Y, Huang Y, Lü N. Comparison of the semen analysis results obtained from two branded computer-aided sperm analysis systems. Andrologia. 2013;45(5):315–318.

37. Franceschi L, Donini M, Perrone V, Klein A, Archambeau C, Seeger M, et al. Hyperparameter optimization in machine learning. arXiv preprint arXiv:241022854. 2024;.

38. Gallagher MT, Cupples G, Ooi EH, Kirkman-Brown J, Smith D. Rapid sperm capture: high-throughput flagellar waveform analysis. Human Reproduction. 2019;34(7):1173–1185.

39. Group M. MIaMIA-SVDS. 2021;doi:10.6084/m9.figshare.15074253.v1.

40. Hayashida J, Nishimura T. MPM: Joint Representation of Motion and Position Map for Cell Tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;.

41. Nishimura K, Hayashida J, Wang C, Ker DFE, Bise R. Weakly-Supervised Cell Tracking via Backward-and-Forward Propagation. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII. Berlin, Heidelberg: Springer-Verlag; 2020. p. 104–121. Available from: https://doi.org/10.1007/978-3-030-58610-2_7.

42. Ben-Haim T, Riklin-Raviv T. Graph Neural Network for Cell Tracking in Microscopy Videos. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022.

43. Anjum S, Gurari D. CTMC: Cell Tracking With Mitosis Detection Dataset Challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2020.

44. Kaiser T, Ulman V, Rosenhahn B. CHOTA: A Higher Order Accuracy Metric for Cell Tracking. In: European Conference on Computer Vision Workshops (ECCVW). Springer; 2024.