Summit: Automated Analysis of Arrayed Single-Cell Gel Electrophoresis

Julea Vlassakis1, Kevin A. Yamauchi1,2, and Amy E. Herr1

Abstract
New pipelines are required to automate the quantitation of emerging high-throughput electrophoretic (EP) assessment of DNA damage, or proteoform expression in single cells. EP cytometry consists of thousands of Western blots performed on a microscope slide-sized gel microwell array for single cells. Thus, EP cytometry images pose an analysis challenge that blends requirements for accurate and reproducible analysis encountered for both standard Western blots and protein microarrays. Here, we introduce the Summit algorithm to automate array segmentation, peak background subtraction, and Gaussian fitting for EP cytometry. The data structure storage of parameters allows users to perform quality control on identically processed data, yielding a ~6.5% difference in coefficient of quartile variation (CQV) of protein peak area under the curve (AUC) distributions measured by four users. Further, inspired by investigations of background subtraction methods to reduce technical variation in protein microarray measurements, we aimed to understand the trade-offs between EP cytometry analysis throughput and variation. We found an 11%–50% increase in protein peaks that passed quality control with a subtraction method similar to microarray “average on-boundary” versus an axial subtraction method. The background subtraction method only mildly influences AUC CQV, which varies between 1% and 4.5%. Finally, we determined that the narrow confidence interval for peak location and peak width parameters from Gaussian fitting yield minimal uncertainty in protein sizing. The AUC CQV differed by only ~1%–2% when summed over the peak width bounds versus the 95% peak width confidence interval. We expect Summit to be broadly applicable to other arrayed EP separations, or traditional Western blot analysis.

Keywords
data analysis (informatics and software), microfluidics (microtechnology), Western blot, high-throughput arrays

Introduction
Electrophoretic (EP) separations provide critical information regarding the physicochemical properties and abundance of a given protein or proteoform. Examples of physicochemical properties that separations can quantify include the protein molecular mass (as in sodium dodecyl sulfate–polyacrylamide gel electrophoresis [SDS-PAGE]), isoelectric point (determined by isoelectric focusing), or both (via 2D electrophoresis). Such separations are performed in a variety of formats, including slab gels, microchannels, and capillaries, each with specially designed analyses to quantify properties of the electropherogram. However, some analyses of EP separations still require substantial manual processing, such as the nonstandardized and poorly documented practices employed for densitometry of Western blotting. Western blotting utilizes a size-based EP separation followed by a transfer of the separated analytes to a membrane (blotting) on which proteins are detected with fluorescent or chemiluminescent antibodies. Currently, it is still typical for the user to manually outline regions of the image to quantify (e.g., selecting the protein bands) using commercially available software sold with Western blot imagers. In a random survey of 100 publications from PubMed that utilized densitometry for Western blotting quantitation, only one paper provided sufficient details for the analysis to be reproduced (e.g., how bands were selected, whether peak integrals or heights were used for quantification, and type of intensity profile background subtraction).

1Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA
2The Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Received Aug 21, 2020, and in revised form June 17, 2021. Accepted for publication June 30, 2021.

Supplemental material is available online with this article.

Corresponding Author:
Amy E. Herr, Department of Bioengineering, University of California Berkeley, 308B Stanley Hall, MC #1762, Berkeley, CA 94720, USA.
Email: aeh@berkeley.edu
Manual processing becomes impractical for high-throughput EP separations. Our group and others have introduced high-throughput EP separations with up to hundreds to thousands of separations performed in parallel across an array,11–16 or with 1300 separations in series in a capillary.17 Applications of arrayed separations include assessing cell-to-cell heterogeneity in DNA damage via single-cell comet assays,18,19 and quantifying proteoforms responsible for cancer drug resistance in breast cancer with EP cytometry.20

EP cytometry performs thousands of single-cell protein separations in a device patterned on a standard microscope slide, and protein peaks are detected with fluorescent antibodies. We previously developed quantitative algorithms for EP cytometry to determine protein expression levels by area under the curve (AUC) analysis.13,21 However, the previous algorithms were cumbersome (with 10–20+ functions) and did not include features that would enable reproducible analysis.

Image processing of microscopy data faces similar challenges in accessibility, scale, and reproducibility. To aid in the accessibility of the analysis, image processing tools such as CellProfiler22 and Ilastik23 have introduced graphical user interfaces (GUIs) that guide the user through the analysis. Once the image processing pipeline is established, the user can then apply the same analysis across large data sets via a batch processing mode. Further, toward improving reproducibility, CellProfiler and Ilastik allow the user to save the analysis pipeline to a file that contains all of the relevant parameters. Starfish,24 a toolkit for processing in situ transcriptomics data, approaches reproducible analysis by storing the results in data structures that automatically record the information required to reproduce the analysis (e.g., parameters used, versions of the software dependencies). However, processing of microscopy images of cells differs from the analysis needed for EP separation images, in which protein peak locations along a defined separation axis may be used to size and identify proteoforms. Existing image processing tools can quantify the integrated fluorescence (amount of protein) from segmented images, but are not readily adaptable to identify a separation lane to index protein band locations.

Inspired by advances in image processing toolkits for microscopy data, we introduce Summit for EP cytometry quantitation and assess the impact of specific algorithm design choices on measured single-cell protein and intruser variation. Functional decomposition was employed to establish the four functions of Summit. The algorithm requires minimal user interaction through the use of GUIs, making it accessible to researchers of all computational backgrounds. Design choices, including data storage in a structure, background subtraction method, and quality control thresholding, are described. Analysis reproducibility is confirmed as differences in protein AUC distributions are not statistically significant when quantified by four different Summit users.

When comparing background subtraction methods, we find that mean subtraction (similar to average on-boundary subtraction in 2D electrophoresis) yields 11%–50% more quantifiable single-cell protein peaks compared with the axial subtraction method used in our previous work. Such throughput gains come without substantial increases in variation in measured protein. We find that the appropriate background subtraction method for a given protein can be evaluated by comparing variation in the peak offset after subtraction and the AUC confidence interval with different peak bounds for summation. The guidelines presented here will aid in accurate and reproducible EP cytometry image analysis for various basic research and clinical applications.

Materials and Methods

Reagents

All reagents used are summarized in the Supplemental Information.

Cell Culture

Cell culture conditions are summarized in the Supplemental Information. All cell lines tested negative for mycoplasma and were authenticated by short tandem repeat analysis (Promega, Madison, WI) (see Suppl. Figs. S1–S3).

EP Cytometry Separations

EP cytometry was performed as described in detail elsewhere21,25 and is summarized in the Supplemental Information and Supplemental Table S1, and EP cytometry images from breast cancer tumor-derived cells are from a previously published work in which EP cytometry conditions were described.20 The example proteins characterized in this work are GFP, GAPDH, HER2, rs6, and p-rs6. GFP and GAPDH are from distinct EP cytometry experiments (different cell types, different images), while HER2, rs6, and p-rs6 are from the same EP cytometry experiment (breast cancer tissue-derived cells, distinct fluorescence channels of the same EP cytometry image). All images are available on Figsbite repositories: EP cytometry of tumor 0909 for HER2, rs6, and p-rs6: https://figshare.com/s/fc77e1e77e6976f3c59; and EP cytometry for GFP, GAPDH, and simulated skewed peak images: https://doi.org/10.6084/m9.figshare.13557482.v1.

Summit was written for MATLAB version 9.1 (2016b; MathWorks, Natick, MA) or higher. MATLAB program file dependencies of Summit include the Image Processing Toolbox (version 9.5 or higher) and the Curve Fitting Toolbox (version 3.5.4 or higher). The complete source code is available on GitHub (https://github.com/herrlabucb/summit), and instructions for installing Summit are provided in the repository readme.

Supplemental Information

All information is available in the Supplemental Information.
**Statistical Analysis**

All box plots display the median, with the box edges at the 25th and 75th percentiles and whiskers that extend to the minimum and maximum values. The Mann–Whitney U test, Kruskal–Wallis test, Dunn–S Sidak post hoc comparison test, and linear regression were performed using built-in functions in MATLAB (2019b).

**Results**

The aim of EP cytometry image processing is to both quantify the protein expression and estimate the molecular mass (when using appropriate sizing standards). We developed the Summit image processing pipeline to be modular by using functional decomposition to discretize the pipeline into four functions. The raw image in **Figure 1A** contains 1615 individual separation lanes that are processed in the four functions shown in **Figure 1B**: (1) regions of interest (ROIs) are segmented; (2) a 1D intensity profile is generated and the background subtracted; (3) the intensity profile is fit to a Gaussian curve; and (4) the user inspects the fitted peaks as a final quality control step. From analysis of the protein peak AUC, cell-to-cell heterogeneity in protein expression or correlation between proteins in known cancer
signaling networks can be quantified (Fig. 1C). Leveraging protein sizing in SDS-PAGE, Gaussian fit-determined peak locations of proteins can identify the mass of proteoforms, such as the truncated HER2 protein species t-erbB2 (~95–115 kDa) that confers trastuzumab resistance in HER2-positive breast cancer\(^{20,26}\). In this instance, the pixel intensities for t-erbB2 are summed across the width of the ROI to produce the 1D intensity profile for background subtraction and Gaussian fitting to identify the peak location.

User interaction is limited to the selection of array boundaries in the ROI segmentation function, the selection of peak boundaries in the curve-fitting function, and the quality control function. There are options to input thresholding parameters to assist with quality control, or the user may choose to manually inspect each separation lane, as described in more detail in the quality control section. Here, the reproducibility and accuracy of quantitation using Summit are explored in detail.

### Assessing Analysis Reproducibility

To ensure that the results of an analysis performed with Summit are reproducible, we specified the output of the Summit analysis pipeline to include all parameters necessary to perform the analysis. By selecting a data structure for algorithm output variable storage, variables of different data types can be saved within a single file (Suppl. Table S2) and accessed by the field in the command line. Output variables allow the analysis to be reproduced by another user if they have access to the data structure. In order to achieve analysis reproducibility, parameters for EP cytometry image segmentation to ROIs must be logged and ROIs should be able to be duplicated (either by other users or by the same user analyzing multiple images from the same EP cytometry device).

We leverage the arrayed format of most EP cytometry separation devices in the image segmentation process to generate the ROIs for each separation lane (Fig. 2) using the roiGeneration function. In order to establish array boundaries and align the separation axes with the vertical axis of the image, a GUI first prompts the user to select microwells within a row approximately 10 microwells apart (Fig. 2A, B). The necessary angle of rotation to straighten or align the image is calculated from the two points. The rotation is performed with a built-in MATLAB function, imrotate(), which takes inputs of an image and an angle. The function first translates the image centroid to the origin (the upper left corner of the image), and then the image matrix is rotated around the origin by the angle calculated from the dot product of two direction vectors between the two selected microwell coordinates. Following rotation, the matrix is translated back to the original centroid location. After rotation, the user is instructed to select four microwells along the row and column boundaries of the array (Fig. 2C). The function informs the user if they have selected incorrect boundaries (e.g., if the coordinate of the “top-most” well of the array is lower than the coordinate of the “bottom-most”). As the gel is attached to a glass slide, image warping does not hinder the use of array boundaries and well spacings to segment the image. The aligned image is next segmented into ROIs based on the user input well-to-well spacing (Fig. 2D). The length \(L\) and width \(W\) of each lane set the ROI dimensions. The ROIs are then stored in the data structure as an \(L \times W \times N\) array, where \(N\) is the number of separation lanes. As an optional input variable of roiGeneration(), the user may provide a data structure with parameters such as the user-selected array boundaries, and angle of rotation, to apply the same ROIs to another EP cytometry image.

---

**Figure 2.** The EP cytometry image is rotated to align the array before segmentation into thousands of individual ROIs. (A) Process workflow. (B) False-colored fluorescence micrograph showing two rows of microwells with user selection of two microwells indicated with red plus signs for image alignment with the MATLAB function imrotate(). Scale bar is 0.1 mm. (C) False-colored fluorescence micrograph with example of user-selected array boundaries (red lines indicating rows or columns of microwells at the edges of the array). The numbers indicate the required order of selection. Scale bar = 1 mm. Inset is a false-colored fluorescence micrograph in the region of the first array bound (scale bar = 0.1 mm). (D) False-colored fluorescence micrograph of a subset of the array segmented into ROIs with a length \(L\) and width \(W\) input by the user. ROIs with protein signal from cells are shaded green.
When four users, users 1–4 (including two developers, users 1–2), independently performed EP cytometry analysis with Summit (using identical ROIs), they did not find substantial differences between the measured protein AUC distributions. We utilize Gaussian curve fitting to determine protein expression via AUC analysis. The Gaussian fit function, \( f(x) \), employed is

\[
f(x) = ae^{-\left(\frac{x - \mu}{b}\right)^2}
\]

where \( a \) is the amplitude, \( \mu \) is the peak center, and \( b \) is the peak width. The protein peaks assume a Gaussian distribution owing to protein diffusion during the EP separation. The AUC may be determined by summing over \( \mu \pm 2\sigma \), where \( \sigma = \frac{b}{\sqrt{2}} \). We choose the summation bounds that correspond with the 95% confidence interval for a Gaussian distribution. As shown in Figure 3A, we cannot reject the null hypothesis, and thus the AUC distributions quantified by the four users can be considered as drawn from a single distribution (Kruskal–Wallis \( p > 0.05 \)).

To assess differences between the four users in variation in AUC (which will not be normally distributed), we quantified the CQV, which is recommended over the coefficient of variation for skewed distributions.

\[
CQV = \frac{Q_3 - Q_1}{Q_3 + Q_1}
\]

where \( Q_i \) and \( Q_i \) are the third and first quartile of the distribution. The CQVs only differ by \( \pm 6.5\% \) between the distributions measured by the four users, who each analyzed 102, 107, or 108 peaks. A GUI displays each intensity profile and the corresponding Gaussian fit in a grid and requests that the user select the peak profiles to be thrown out. In total, there were 11 intensity profiles out of 121 manually inspected that all four users did not agree upon. To determine intrauser variation, user 2 repeated analysis of the GFP EP cytometry image days after the initial analysis (Fig. 3B). User 2 identified 108 peaks passing quality control initially and 112 on repeat analysis. To assess the similarity between the initial and repeat analysis sets of peaks, we calculated the Jaccard index to be 0.96 (the size of the intersection divided by the size of the union of the two sets), where 1 is the maximum. For the 108 peaks in common between initial and repeat analyses, the difference between the AUC distributions is not drawn from a distribution around zero (Wilcoxon signed rank test \( p < 0.0001 \)). While \( p < 0.0001 \), the median AUCs are \( <1\% \) different (user 2: median AUC = 8273; analysis 2, median AUC = 8223) and the distributions overlap substantially. The agreement in the AUC variation and the manual inspection of the separation lanes between users demonstrates the user-to-user reproducibility of Summit, while intrauser variation is also minimal.

As protein peaks may assume non-Gaussian shapes, particularly in chromatography data, users may opt to apply a built-in skewed peak function for fitting non-Gaussian protein peaks. We provide an example of simulated skewed protein peaks in Figure 3D, and the function required to generate skewed peak data is available on the Summit Github repository. The fit function utilizes a skewed Gaussian function given by

\[
f(x) = \beta e^{-\frac{1}{2} \left(\frac{x - \xi}{\omega}\right)^2} \times \frac{1}{2} \text{erfc} \left( -\frac{\alpha (x - \xi)}{\sqrt{2}} \right)
\]

where \( \beta \) is the amplitude, \( \xi \) is a location parameter, \( \omega \) is a scale parameter, \( \text{erfc} \) is the complementary error function, and \( \alpha \) is a shape parameter, such that \( \alpha > 0 \) produces a right-skewed peak, \( \alpha < 0 \) gives a left-skewed peak, and \( \alpha = 0 \) returns a nonskewed Gaussian peak. We calculate the AUC by summing between \( \xi \pm 2 \frac{\omega}{\sqrt{2}} \). With 400 simulated skewed Gaussian peaks (with hard-coded background and random noise) and user input random variation in peak location and width, we find an average fit \( R^2 \) value of 0.98 ± 0.01 (standard deviation).

To assess user-to-user variation in analysis of the simulated skewed Gaussian image, the four users performed Summit analysis with skewed Gaussian peak fitting (Fig. 3E). Here, all lanes of the simulated image pass quality control, so variation between users represents the impact of steps such as ROI generation, peak boundary selection (which aids in parameter estimation for the fits), and how robustly the peak fitting identifies bounds for AUC integration. The CQVs varied by \( <0.2\% \) and the Kruskal–Wallis \( p \) value was 0.98, indicating that skew peak analysis is highly reproducible across users.

Finally, to benchmark Summit’s quantitative accuracy in assessing peak AUC variation, we carried out Summit and CellProfiler analysis on the simulated skewed Gaussian peak image. Among various batch image processing tools, we selected CellProfiler because it includes an example pipeline for skewed electrophoresis peak analysis from Comet assay gels (in which skewed DNA peaks randomly distributed across the gel report the degree of DNA damage from each single cell). Importantly, CellProfiler does not report metrics that readily allow protein sizing (peak locations relative to a common separation axis origin, e.g., microwell), but the image segmentation utilized in CellProfiler does report integrated fluorescence intensities for each background-subtracted peak in the gel. The
calculated AUCs from Summit and CellProfiler are not directly comparable because of the 1D averaging of the ROI performed in Summit. Thus, we aimed to assess normalized AUC distributions (AUC for each peak divided by the mean AUC for all 400 simulated peaks) calculated by Summit and CellProfiler (Fig. 3F). The normalized AUC distributions are not statistically significantly different (Mann–Whitney $p = 0.82$). However, CellProfiler reports slightly lower CQVs in the normalized AUC (2.34% vs 3.01%). Thus, Summit quantifies AUC variation from Gaussian fitting of 1D intensity profiles that is in agreement with established batch image processing tools, while Summit is equipped to
report separation metrics (e.g., estimated protein molecular mass or separation resolution between two peaks).

Comparing Efficacy of Mean versus Axial Background Subtraction

We investigated two local approaches for background subtraction for each individual ROI: (1) mean background subtraction (which closely resembles average on-boundary background subtraction in 2D electrophoresis\(^3\)) and (2) axial background subtraction (employed in our previous work\(^2\)). The background subtraction methods are compared with EP cytometry of four example proteins: GFP and GAPDH (from MCF7 breast cancer cell lines) and HER2 and p-rs6 (from patient-derived breast cancer tumor tissue; data made publicly available previously\(^2\)) in Figure 4. Once each ROI is segmented, we generate 1D intensity profiles from each 2D EP cytometry image. This is achieved by averaging each pixel along the width of the ROI, which collapses the 2D image to a 1D raw profile (Fig. 4A). Both subtraction methods take “gutter” regions directly adjacent to the separation lane as the background region (Suppl. Fig. S4). For mean background subtraction, all pixels within the two gutters are averaged (Fig. 4B) and the average value is subtracted from the 1D profile (Fig. 4B). Axial background subtraction calculates an average gutter background at every position along the length of the separation lane (Fig. 4A). For example, for a typical 5-pixel gutter width,
10 total pixels are averaged and that average value is subtracted from the corresponding intensity value at that location along the separation axis (Fig. 4B).

In order to assess the efficacy of background subtraction, we introduce a method to quantify how well each method brings the protein peak baseline to zero. We quantified the median of an offset region of the intensity profile outside of the peak area (i.e., 2 to 3σ from the peak center; outlined in gold in Fig. 4B). The offset intensity is lower when using mean background subtraction for GFP (median offset, 218 vs 258; Mann–Whitney p = 0.001; Cohen’s d < 0.01) and p-rs6 (median offset, 69 vs 159; Mann–Whitney p = 0.015; Cohen’s d < 0.01, Fig. 4C). Corroborating this result, the number of peaks that pass quality control is ~11%–50% higher for mean instead of axial subtraction.

Automated Gaussian Curve Fitting for Thousands of Peak Intensity Profiles

In order to improve the quality of the Gaussian fit, the region of the intensity profile to be fit must be selected, and initial guesses for each of the fit parameters must be provided. The manual peak bound selection with an overlaid plot of all electropherograms was previously described for high-throughput microfluidic capillary electrophoresis17 and is employed here for thousands of separations (Suppl. Fig. S5). The constraints for the parameters are 0 < a < a_max (where a_max is the maximum intensity across all of the intensity profiles), left bound < μ < right bound, and 0 < b < right bound – left bound. MATLAB uses a least-squares curve-fitting algorithm for Gaussian fitting of each curve. If an estimated signal-to-noise (SNR) threshold is provided, only intensity profiles with an estimated SNR ratio >3 are fit to the Gaussian function. Goodness-of-fit metrics such as the R^2 value and confidence intervals for each fit parameter are stored in the results data structure.

Increasing the goodness of the Gaussian fit improves the measurement of molecular mass and quantity of a detected protein. The R^2 values of the Gaussian fit were lower with axial subtraction than with mean subtraction, except for the HER2 peaks (Fig. 5 A, B). The largest difference in R^2 values between the two background subtraction methods was with the skewed GAPDH peaks (median R^2 = 0.71 and 0.81 for axial and mean subtraction, respectively). GAPDH and HER2 had statistically significantly lower R^2 values than GFP or p-rs6 with mean subtraction (Kruskal–Wallis p < 0.0001; Dunn–Sidak multiple comparison test p < 0.0001 for each pairwise comparison except for GAPDH and HER2, p = 0.14; all sample sizes shown in Fig. 5B). The median R^2 values for GAPDH and HER2 were 0.71 and 0.88, respectively. For a more specific assessment of fit performance, we quantified the uncertainty in the fit parameters. We calculated the percent variation between each fit parameter and the lower bound of the parameter’s 95% confidence interval and found that the uncertainty decreases with increasing R^2 value, as anticipated (Fig. 5B). For example, the percent variation in peak width is

\[
\% \text{ variation } \sigma = \frac{\sigma - \sigma_{ci}}{\sigma} \times 100
\]

The parameter with the smallest percent variation is the peak center (Suppl. Fig. S6), which largely differed by <5%. In contrast, for the majority of peaks with a fit R^2 value <0.8, the percent variation in the peak width is >10% (Fig. 5B).

Dependence of AUC on Peak Width Parameter Confidence Interval

We quantified AUC variation with summation of the background-subtracted intensity profile over μ ± 2σ or μ ± 2σ_ci (which represents a 95% confidence interval for the AUC based on the uncertainty of the peak width fit parameter) as shown in Figure 5A. As shown in Figure 5C and Table 1, the measured distributions vary minimally based on the background subtraction employed. For all proteins studied here except GFP, we fail to reject the null hypothesis that protein levels measured with mean or axial subtraction are drawn from a single distribution (Mann–Whitney test p > 0.05). The CQV of AUC differed by only a few percent between mean and axial background subtraction for each protein target investigated here. Mean subtraction slightly increases AUC CQV by ~2% and ~4.5%, respectively, for HER2 and p-rs6, while decreasing CQV by ~3.5% and 1% for GFP and GAPDH, respectively. Here, protein variation for GFP, GAPDH, HER2, and p-rs6 is relatively accurately reported when quantified from AUCs with the typical μ ± 2σ summation because CQVs differ by only ~1%–2% when calculated with μ ± 2σ_ci summation bounds instead.

Discussion

We designed Summit as an automated platform for carrying out analysis of EP cytometry images, storing critical analysis parameters needed for reproducibility and often neglected in Western blot densitometry. Further, we sought to understand how algorithm design decisions, such as a method of background subtraction, affect uncertainty and variation in protein expression. Consequently, we introduce offset intensity as a metric for assessing the efficacy of background subtraction. Additionally, we evaluate the confidence intervals of Gaussian fit parameters and AUCs of protein peaks. While parameters and design choices may need to be tuned for each protein analyzed, we offer guidelines based on increasing the number of quantifiable separations per image and minimizing variation induced by the analysis itself.
Figure 5. Quantitative assessment of goodness of Gaussian fitting and AUC dependence on mean versus axial background subtraction. (A) Representative intensity profiles (axial subtraction, black trace; mean subtraction, red trace) and Gaussian fit (blue trace) of GFP peaks that passed quality control with the minimum (min), median, or maximum (max) $R^2$ value for the axial subtraction set. Black dashed lines indicate the summation bounds for calculating the AUC from the peak width, $\sigma$ ($\pm 2\sigma$ or $\pm 2\sigma_{ci}$, where $ci$ corresponds to the 95% confidence interval for the peak width fit parameter). (B) Percent variation between $\sigma$ and $\sigma_{ci}$ as a function-of-fit $R^2$ value for mean (red points) and axial (black points) subtraction for the analysis of EP cytometry images of GFP, GAPDH, HER2, and p-rs6. Number of data points and the Mann–Whitney $p$ value for the $R^2$ value are shown. (C) Box plot quantification of the AUC with the indicated summation bounds. Red indicates mean background subtraction and black indicates axial background subtraction. The numbers of lanes that passed quality control with both background subtraction methods are $n = 523$ (GFP), $n = 19$ (GAPDH), $n = 104$ (HER2), and $n = 45$ (p-rs6). Mann–Whitney $p$ value significance levels: ns = not significant, *** $p < 0.0001$. 
Table 1. Comparison of the CQV (×100%) for Mean versus Axial Background Subtraction and Summation Over ±2σ or ±2σσ
(Representing the 95% Confidence Interval for the AUC).

<table>
<thead>
<tr>
<th></th>
<th>GFP (n = 523)</th>
<th>GAPDH (n = 19)</th>
<th>HER2 (n = 104)</th>
<th>p-rs6 (n = 45)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean subtraction</td>
<td>40.23%</td>
<td>22.35%</td>
<td>46.27%</td>
<td>67.72%</td>
</tr>
<tr>
<td>Axial subtraction (σσ) mean subtraction sum)</td>
<td>43.76%</td>
<td>23.20%</td>
<td>44.33%</td>
<td>63.15%</td>
</tr>
<tr>
<td>Mean subtraction (σσ) sum)</td>
<td>44.31%</td>
<td>22.18%</td>
<td>46.16%</td>
<td>62.92%</td>
</tr>
</tbody>
</table>

User-to-User Variation in Quantitation Is Minimal Owing to Accessibility of Parameters to Reproduce Analysis and Automation

Emerging guidelines for computational research to improve scientific reproducibility suggest that individual steps of code that affect the final analysis should be logged. Thus, the choice of how to store output of both the analysis of interest and the variables used to execute the functions is crucial for making a given analysis reproducible.

Even using the same ROIs to segment an EP cytometry image, user-to-user-variation in analysis output can occur owing to manual quality control with optional quantitative thresholds. The optional thresholds include a minimum Gaussian fit R² value (previously recommended at ~0.7) and an SNR ratio estimate >3. After thresholding, a user may eliminate a given intensity profile for various reasons: (1) punctate fluorescence (e.g., dust or fluorescent antibody aggregates) in the peak region, (2) gel damage in the peak or background region, (3) large fluorescent debris (e.g., fluorescent cell debris) in the peak or background region, (4) unexplained injection dispersion causing peak skew, (5) poor Gaussian fitting resulting in inaccurate peak boundary identification, or (6) peak SNR ratio that is too low. Similar manual inspection and elimination of protein spots for analysis has been applied to regions of reverse-phase protein arrays, prompting concern that user-to-user variation in quality control could yield inconsistent results.

Choice of Background Subtraction Method Impacts Assay Throughput

Background subtraction of protein signal remains a challenge in the quantitative analysis of protein electrophoresis and microarray images. The choice of background subtraction method strongly influences the measured quantity of protein and variance. Common background subtraction approaches for electropherograms include rolling ball, baseline, and “on-boundary” (e.g., lowest or average) subtraction. The rolling ball method has been shown to yield densitometry protein measurements that are less correlated with a radioimmunoassay than if no background subtraction were performed.

Average on-boundary subtraction (the average of a trace just beyond the protein band boundary) in 2D electrophoresis was shown to achieve the lowest overall protein coefficient of variation in acidic 2D electrophoresis. A similar “neighborhood correction” strategy applied to reverse-phase protein microarrays that suffer from nonuniformities across the array also yielded a lower protein spot coefficient of variation.

Quantifying variation in offset intensity, CQV dependence on subtraction method, and summation bounds, and assessing if higher n reveals otherwise unmeasured cell subpopulations, can aid in the choice of background subtraction best suited for a given protein target and research question. When mean subtraction reduces or only slightly increases the CQV relative to axial subtraction, mean subtraction may be the preferred background subtraction method, as it likely increases n.

Maximizing the fraction of analyzed samples that yield useful data (i.e., pass quality control) is critical in single-cell analysis in order to identify rare cell subpopulations. We hypothesize that mean subtraction yields higher n because the effect of any “hot” pixels in the background gutter region is averaged out with mean subtraction. In the
case of cells derived from precious patient tumor samples, maximizing assay quality yield is critical to assess biomarker heterogeneity (like the HER2 and p-rs6 analyzed here) that may dictate whether a chemotherapy drug is effective or not.\textsuperscript{20} Thus, further refining the background subtraction process to maximize $n$ is an important future direction.

As new data types may benefit from alternative background subtraction methods tailored to the features of the background, Summit includes an optional input to the intProf function in which the user may supply their own background subtraction function. Such custom functions will be especially useful for low-abundance proteins, for which low peak SNR ratios may result from averaging to generate the 1D intensity profile. Additionally, low SNR ratio peaks may benefit from denoising algorithms employed in 2D electrophoresis,\textsuperscript{40} such as nonlinear filtering.\textsuperscript{41}

\textbf{Narrow Confidence Intervals of Fit Parameters Indicate Analysis Contributes Minimally to Technical Variation of Quantified Peak AUC or Location}

While offset intensity is different between mean and axial subtraction, we only expect differences in variation of offset intensity, not magnitude, to result in altered AUC variation. Thus, to truly understand whether either axial or mean background subtraction is associated with higher protein variation, we also directly compared the AUCs with different subtraction methods while evaluating the accuracy of AUC determination. We hypothesize that mean and axial subtraction yield different GFP AUC distributions because of the bleed-through of GFP signal into the background region seen in Figure 4A.

Accurate quantitation of peak width and location parameters is critical for the determination of protein size,\textsuperscript{42} separation performance (e.g., separation resolution), and summation bounds for the AUC. Here, AUC is determined from the background-subtracted intensity profile,\textsuperscript{43} as AUC quantification directly from the Gaussian fit is preferable for peaks with substantial overlap. We hypothesized that higher uncertainty of the peak width parameter may reduce accuracy of the AUC for lower $R^2$ fits. We found that GAPDH and HER2 had the lowest median $R^2$ values, which results from the lower SNR ratio of the GAPDH peaks and notable injection dispersion for HER2 shown in Figure 4A. However, the uncertainty of the peak width parameter for these lower $R^2$ peaks did not lead to statistically significant differences in the AUCs when summing over $\mu \pm 2\sigma$ versus $\mu \pm 2\sigma$ (Mann–Whitney $p > 0.05$). Thus, Summit quantifies these examples of peaks with lower SNR ratio or dispersion without higher uncertainty in the AUC.

Quantitative analysis of EP separation data can reveal the presence and abundance of proteoforms responsible for biological processes such as cancer progression.\textsuperscript{44–47} Summit provides reproducible quantitation by storing all critical parameters and output variables in the results data structure. Further, automating most aspects of the analysis and quality control reduces user-to-user variation in the analysis results. We found that the differences between quantified protein peak AUC are not statistically significant between Summit users. Minimization of assay technical variation introduced by the algorithm may be achieved by examining variation in background subtraction offset, and CQV variation within the 95\% confidence interval of peak bounds. In some instances, there may be a trade-off between the number of analyzed peaks and variation, depending on the background subtraction employed.

For both background subtraction methods explored here, punctate noise in the vicinity of the peak or background “gutter region” is a common feature that prevents certain peaks from passing quality control. We are investigating 2D segmentation methods that may improve yield by isolating the antibody signal from the punctate nonspecific signals on the periphery of the separation lane. Improving assay quality yield would enhance the ability of EP separation methods to detect rare cell populations.

Owing to the modular design of Summit, we anticipate components such as the ROI generation and intensity profile background subtraction could be applied to other arrayed separations and detection of different biomolecules. To date, the algorithm has been applied to both size-based separations in custom lab-on-a-disk devices\textsuperscript{48} and single-cell isoelectric focusing assays.\textsuperscript{16} For the latter, geometric differences between the 1-mm-long EP cytometry separation lanes and 9-mm-long focusing zones were readily modified, as the parameters defining the geometry of the separation array are Summit input variables. Given the ease of adapting the analysis to other geometries, Summit could likely also be used for standard Western blotting densitometry. We anticipate that the skewed Gaussian function included in Summit could be applied to commonly skewed data types such as the arrayed comet assay for DNA damage.\textsuperscript{15} Further, addition of an alternate skewed fit function, such as the Lorentzian-modified Gaussian,\textsuperscript{49,50} could allow application of Summit to chromatography data.

In the future, Summit-extracted peak shape parameters may shed insight on proteins that are poorly solubilized in EP cytometry (e.g., proteins originating from circulating tumor cells\textsuperscript{51}). Metrics of EP injection dispersion may guide optimization of sample preparation. For example, we envision identifying effective denaturation conditions for EP cytometry tuned to different classes of protein domain structures\textsuperscript{52} based on peak skew, variance, and quantity of signal entrapped at the microwell interface. Such quantitative analysis will
broaden the applicability of EP cytometry to new sample types and protein targets.

Acknowledgments

We are grateful to current and former Herr lab members, and students from the Cold Spring Harbor Labs Single-Cell Analysis Course for user feedback on Summit. We acknowledge Ms. Anjali Gopal and Mr. Andoni Mordoukoutas for their contributions to reproducibility analyses.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors acknowledge competing financial interest(s) as follows: J.V., K.A.Y., and A.E.H. are inventors of EP cytometry intellectual property and may benefit from licensing royalties.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Society of Lab Automation and Screening Graduate Education Fellowship (J.V.), NSF Graduate Fellowship DGE1106400 (J.V., K.A.Y.), NSF CAREER CBET1056035 (A.E.H.), and NIH R01CA203018 (A.E.H.).

ORCID iDs

Julea Vlassakis https://orcid.org/0000-0001-5456-6366
Amy E. Herr https://orcid.org/0000-0002-6906-2985

References


