Single molecule protein detection for proteomic profiling

Simon Leclerc

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE des Sciences de la Vie et de la Santé
UMR 1121 du CNRS

THÈSE présentée par :

Simon LECLERC

soutenue le : **24 Août 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**
Discipline/ Spécialité : Biologie moléculaire

Visualisation de protéines individuelles pour la quantification à haute sensibilité du lysat cellulaire

THÈSE dirigée par :
Dr Arntz Youri.....Docteur, Université de Strasbourg
Dr Taniguchi Yuichi.....Docteur, RIKEN

RAPPORTEURS :
Dr Cuisinier Frederic .....Professeur, Université Montpellier 1
Dr Csilla Gergely.....Professeur, Université Montpellier 2

AUTRES MEMBRES DU JURY :
Dr Vautier Dominique.....IR inserm, Université de Strasbourg

Acknowledgment

Firstly, I would like to express my sincere gratitude to my advisers Dr Taniguchi Yuichi and Dr Arntz Youri for the continuous support of my Ph.D study, administrative troubles and related research, for their patience, motivation, and knowledge. Their guidance helped me in all the time of research and writing of my article and this thesis. I could not have imagined having better advisers and mentors for my Ph.D study.

Besides my advisers, I would like to thank the rest of my thesis committee: Prof. Cuisinier Frederic, Prof. Csilla Gergely, and Dr. Vautier Dominique, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to Dr. YanagidaToshio who provided me an opportunity to join their team as intern, and who gave access to the laboratory and research facilities. Without they precious support it would not be possible to conduct this research.

I thank my fellow labmates Vipin and Arno for the stimulating discussions, how to recreate the world and for all the fun we have had in the last four years. I also like to thanks the other staff and labmates from QbiC, RIKEN, but also the staff of Strasbourg University

Last but not the least, I would like to thank my loving wife for the continuous support during the thesis and article redaction, and my family, parents and sisters for supporting me spiritually throughout writing this thesis and my life in general.

Table of content

Index of Tables

Index of Figure

Index of Annex

Abbreviations table

BSA	Bovine Serum Albumin
CAT	Catalase
CCB	Colloidal Coomassie Blue
CE	Coupling efficiency in Part1, Capillary Electrophoresis in Part 2
CHAPS	3-[(3-CHolamidopropyl)dimethylAmmonio]-1-PropaneSulfonate
Cy3	Cyanine 3
Da	Dalton
DIGE	DIfference Gel Electrophoresis
DMEM	Dulbecco's Modified Essential Medium
DNA	DeoxyriboNucleic Acid
DoL	Degree of Labeling
dSTORM	direct STochastic Optical Reconstruction Microscopy
DTT	DiThiolTheitol
EMCCD	Electron Mutliplying Charge-Coupled Device
EOF	ElectroOsmotic Flow
GFP	Green Fluorescent Protein
FACS	Fluorescent Activated Cell Sorting
FRET	Fluorescence Resonance Energy Transfer
LC	Liquid Chromatography
LO	Labeling Occupancy
LO_{CE}	Labeling Occupancy estimated from CE
LO_{Obs}	Labeling Occupancy estimated from single molecule observation
MEKC	Micellar ElectroKinetic Chromatography
MS	Mass-Spectroscopy
NA	Numerial Aperture
PALM	Photo-Activated Localization Microscopy
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDMS	PolyDiMethylSiloxane
PEG	PolyEthylene Glycol
PMMA	Poly(Methyl MethAcrylate)
PLA	Proximity Ligation Assay
PSA	Prostate-Specific Antigen
RNA	RiboNucleic Acid
sc-Western blot	single-cell Western blot
SDS	Sodium Dodecyl Sulfate
SNR	Signal-to-Noise Ratio
TRITC	Tetramethylrhodamine
YFP	Yellow Fluorescent Protein

Introduction

Biology : what a noisy process!

The central dogma has been first described by Francis Crick in 1958 and re-stated in 1970 and is the following (F. H. Crick 1958; F. Crick 1970):

“The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.”

The DNA replication, the transcription of the RNA and the translation of the RNA in protein is the normal way of gene expression and protein production (Figure 1). In some case involving virus, it is possible to pass from the RNA to the DNA and to replicate RNA from RNA (Temin and Mizutani 1970; Baltimore et al. 1963). In a very specific case, it is also possible to directly translate a protein from DNA (McCarthy and Holland 1965). However, it is not possible to obtain a ribonucleic sequence of the protein from the protein. The self-replication of a protein is not possible, even in the case of prions, because they do not replicate themselves, but replicate their information by modifying the spatial conformation of a strictly similar protein (Ridley 2001). A primary concern of the replication of information is the interaction probability between elements.

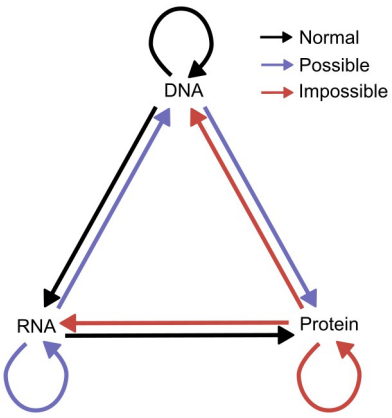


Figure 1: Central dogma in molecular biology. The information is contained on polymer of ribonucleic acid (DNA and RNA) or amino-acid (protein). It can circulate from DNA to RNA to protein in a normal cell (black line), while some special organism or condition can permit to transfer the RNA information to DNA or the DNA to the protein level (blue line). Normal cell is able to replicate their DNA, while some virus can also replicate their RNA. The end product, the protein, is not able to transfer its information to another level neither self-replicate (red lines). Reproduces from (F. Crick 1970)

The probability of two chemicals to interact together is dependent of the volume where they are evolving, with smaller volume increase this probability, other chemical presence and concentration, that can prevent interaction, and other physical factors like for example their

relative speed, orientation, and interactions methods, each one impacting the interaction probability (Milo 2013). For example, the interaction probability of two highly expressed proteins will be high, like the interaction cofilin-actin (Yeoh et al. 2002), while the interaction probability of a low express regulator with its target will be low, like the treslin (Kumagai et al. 2010). These purely chemical and physical stochastic events will lead to fluctuations in the response of a cell to a stimulus, also called ‘noise’ (Figure 2-A), creating variation in the rates of mRNA transcription and protein translation from the central dogma. Thus, a genetically identical cell population can exhibit an important heterogeneity, with different cells having different copy numbers of mRNA or protein. Such variability in gene expression can lead to diverse phenotypes and affect a wide range of cellular function, including development, homeostasis and disease progression (Eldar and Elowitz 2010). These variabilities are based on three concepts that are the burst size, the time averaging and the propagation (Figure 2-B). For the burst size, most proteins are not produced uniformly over time, but during the event of intense transcription that happens stochastically. This is both because the gene promoter can switch in ‘on’ or ‘off’ state, resulting in the sudden production of mRNA, and that each mRNA can be recognized by multiple ribosomes for translation, resulting in a sudden increase in the production of protein (Raj et al. 2006; Newman et al. 2006). The time averaging occurs when the protein lifetime is longer than the interval of time between production bursts, resulting in the accumulation of proteins over time that tends to buffer the burst expression. The propagation can be explained by some genes expression rates influenced by the levels and state of transcription factors that are themselves subject to bursting and time averaging, resulting in the propagation of the burst to the downstream genes (Sigal et al. 2006; Yu et al. 2006).

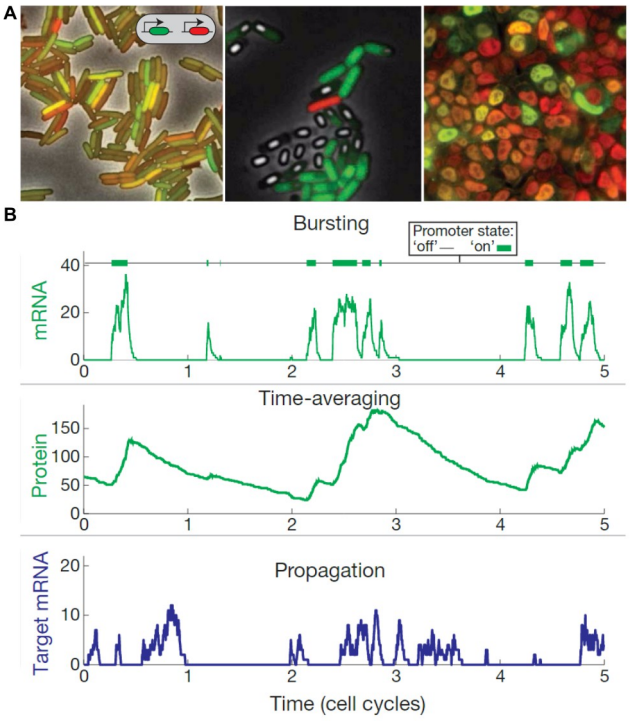


Figure 2: Stochastic gene expression process. (A) Gene expression of fluorescent genes in E.coli (left), B.subtilis (center) and EM stem cells (right) displaying change of expression across the cell population. (B) Three processes explaining the gene expression stochasticity, with a bursty mRNA production (top), and time average effect if the protein lifetime is long enough (center) and the propagation of the burst to downstream mRNA (bottom). Figure is extract from (Eldar and Elowitz 2010).

Stochastic precision

Understanding the cause and consequences of variability in gene expression requires accurate quantification of mRNA and protein at the single cell level. Counting mRNA molecules in single cells is commonly performed in laboratories thanks to the combination of single cell manipulation device and next-generation sequencing, even if low express mRNA are not captured by this technique (Kharchenko, Silberstein, and Scadden 2014). In contrast, counting protein molecules in single cells is much more challenging. Existing techniques require genetic manipulation, antibody or sophisticated devices and often yield relative levels of protein abundance (A. J. Hughes et al. 2014; Shi et al. 2012; Taniguchi et al. 2010; Albayrak et al. 2016; Darmanis et al. 2016).

Single cell protein expression analysis is routinely used by fluorescent-activated cell sorting (FACS) to analyze and sort viable cells based on six or more cell protein surface markers and is useful to purify cellular phenotypes for subsequent analysis (Herzenberg et al. 2002). In addition to this, recent techniques permit to investigate new aspects for single cell analysis (Table 1). One of them is the single cell western blot (scWesterns) that use a combination of single-cell isolation and protein migration together with protein capture to allow the sequential detection of up to twelve proteins in arrays of thousands of cells (A. J. Hughes et al. 2014). The mass cytometry (CyTOF) can assay around thirty proteins by

detecting protein using mass-tag labeled antibody (Irish et al. 2004), while microfluidic platform allows the detection of up to forty secreted proteins by the use of barcode antibody (Cai, Friedman, and Xie 2006; Shi et al. 2012). If these techniques permit to analyze the protein expression of a single cell, the quantitative aspect is at most relative and generally fails to detect lowly expressed proteins that are present in only a few copies.

Technique	Detection method	Comments
FACS	Staining with fluorophore labeled antibodies	Standard technique for cell sorting High throughput (10000 cells/min) Viable cells after analysis Up to 12 multiplex protein detection
CyTOF	Staining fixed cells with mass-tag labelled antibodies	Can analyse cytoplasmic protein Work on fixed cell and tissue Up to 30 multiplex protein detection
Single cell barcode chip	Spatially encoded antibody array for fluorescent immunoassays of secreted proteins or analytes released from lysed cells	Small sample (100-1000 cells) Up to 40 multiplex protein detection
scWesterns	Miniaturized, automated western blotting on a microchip	Small sample (100-1000 cells) Rough estimation of target protein size Up to 12 multiplex protein detection

Table 1: Single cell analysis tools for protein studies

New from old : single protein detection

The ultimate limit in sensitivity for protein concentration measurement is the detection and counting of single molecule proteins in a single cell. Different techniques permit to do this, that can be separated in three different categories, the label-free detection by UV or nanopore, specific detection by the use of antibody and detection by fusion with a fluorescent protein. We will briefly describe these techniques in the following paragraphs.

In proteins, the three aromatic amino acid residues tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe) contribute the most to the protein intrinsic fluorescence when excited in the ultraviolet region of 260-280 nm (Lakowicz 2006). By using a deep UV laser-based fluorescence lifetime microscopy system, Li et Seeger (Li and Seeger 2006) studied the native fluorescence decay and photon bursts of single β-Galactosidase protein from *Escherichia-coli* in aqueous solution by a time-correlated single-photon counting method. The average number of β-Galactosidase protein molecules in the detection volume obtained

from fluorescence correlation spectroscopy measurements proves that they observe single-molecule events, allowing single protein counting (Figure 3)

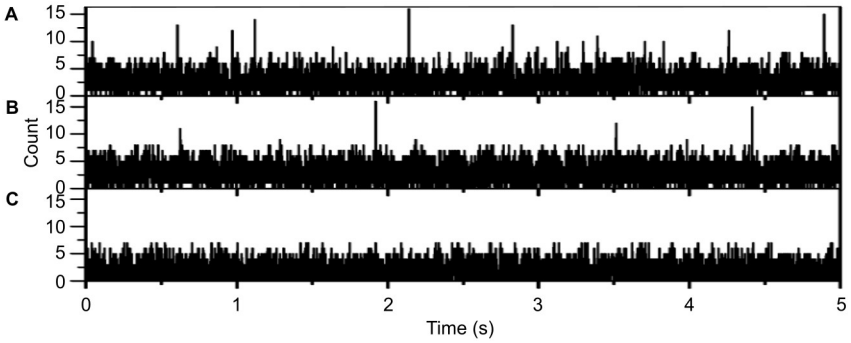


Figure 3: Single protein detection by deep UV fluorescence. (A) Fluorescence photon burst event of single β-Galactosidase in 1×10-11 mol/L solution. (B) Fluorescence photon burst event of single β-Galactosidase in 5×10-12 mol/L solution. (C) Fluorescence photon burst of PBS. Data from (Li and Seeger 2006).

The solid-state nanopore is promising in term of protein detection and identification. This technology already permits to count single protein passing through the nanopore by measuring the current blockade, even if the volume of measure is limited to the close proximity of the membrane (Figure 4) (Nelson et al. 2012). The big advantage of the nanopore is their ability to determine the sequence of the linearized protein by identifying the type of amino acid residues by measuring the nano to picoampere variation in the current intensity during the passage of the protein through the nanopore (Kolmogorov et al. 2017; Kennedy et al. 2016). This technique is theoretically able to identify up to 3000 different proteins with post-translational modification like phosphorylation and can lead to new protein identification in addition to the counting capacity.

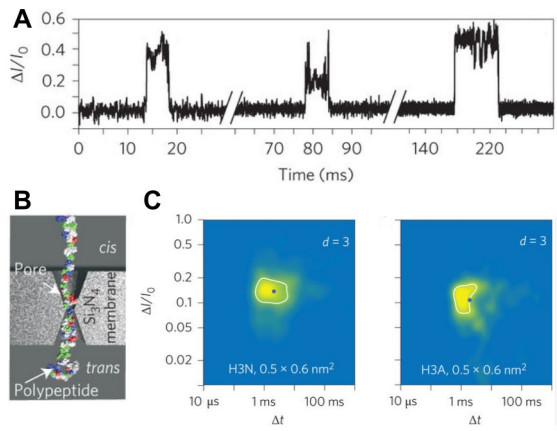


Figure 4: Single protein detection by nanopore. (A) Current blockade ratio when I0 is 217 pA. The peak show the passage of a single CCL5 protein. (B) Schematic of the translocation of a protein through a nanopore. Denaturing agents impart a uniform negative charge to the protein, resulting in a rod-like structure. (C) Heat maps that characterize the distribution of the blockades that are associated with denatured H3N and H3A. Data from (Kennedy et al. 2016).

If the previous techniques allow to indiscriminately detect single molecule protein by using physical properties, another method is to recognize and identify a specific protein by the use of a specific antibody. This method is already used for single cell analysis (Table 1), and even with the possible lack of sensitivity or specificity of antibodies, are commonly used for protein relative quantification and single protein counting.

One of the most mature and automatized technique for single protein counting is the single molecule array (SiMoA), that is commercialized by Quanterix. This technique allows the detection and counting of single molecule based on an ELISA like assay, where an antibody fixed to a bead recognized the protein of interest, the prostate-specific antigen (PSA) here, then the second biotinylated antibody, that detects the PSA on another epitope, is recognized by a streptavidin fused to the β-Galactosidase, permitting the visualization of the PSA after adding the substrate IPTG. The innovation of this method is to use femto-volume wells array that isolate each beads, allowing the reaction to occur in a very small volume, increasing the sensitivity from 10^{-12} mol/L from classical ELISA test to 10^{-15}-10^{-19} mol/L (Figure 5) (Rissin et al. 2010; Rissin and Walt 2006; Schubert et al. 2016). However, there are certain drawbacks associated with this technique like the Poisson law's limits. This law statutes that since a bead is covered by multiple antibodies, it is possible for multiple antigens to react with it, resulting in a false estimation in the number of protein. This parameter can be controlled by counting the number of protein in the array, if the count is too high, it is likely that they are more than one antigen by beads, and adjusting the dilution of the antigen permits to solve this trouble.

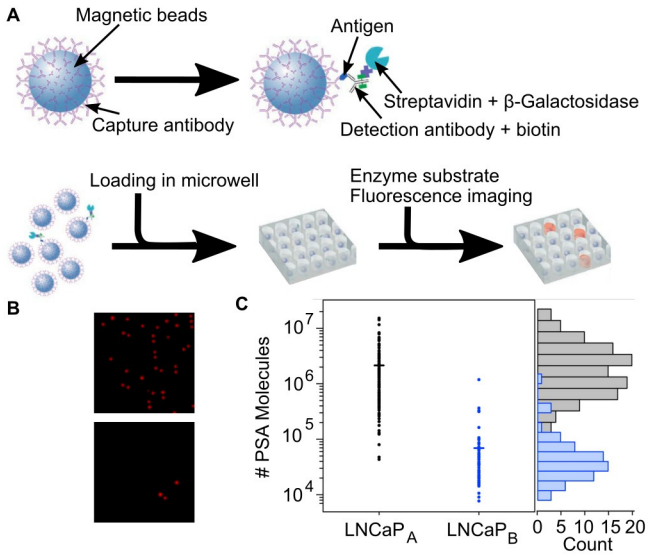


Figure 5: Single molecule array principle. (A) Principle of SiMoA assay, antigen is detected by an antibody sandwich and isolated in microwell to allow their detection. (B) Array visualization with single β-Galactosidase protein at different concentration (7×10^{-12} and 10^{-15} M). (C) PSA protein number in single cell from two cell type (left) and corresponding histogram (right). Data from (Rissin et al. 2010; Schubert et al. 2016).

Amplifying the signal given by the detection of a single molecule protein allows a more sensible identification. The proximity ligation assay (PLA) is a technique used to detect single molecule protein *in situ* that is based on antibody reaction, with the hypothesis that two proteins with DNA linker can permit a hybridization between the two DNA linker, creating a template for PCR reaction (Söderberg et al. 2006). Combined with micro-droplet technology, it has been successfully used for single protein counting in a single cell (Albayrak et al. 2016). After PLA completion, proteins are digested by a protease, and the double-stranded DNA resulting of the PLA is emulsified at dilution limit to form thousands of droplets with one or zero DNA inside. The DNA is then amplified by PCR and detected by a dye, where positive droplet corresponds theoretically to a single protein detection event. However, the PLA is not an assay that can be used for investigation of a lot of different protein, since it is necessary to use two antibodies that can recognize different epitope of the same protein that is spatially close enough to realize the DNA hybridization, making difficult to be sure about the absolute counting of a protein (Figure 6).

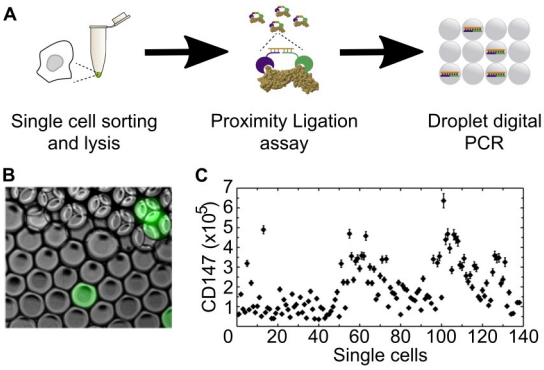


Figure 6: Proximity ligation assay. (A) Principle. A isolated single cell is lysate, PLA is realized on the target protein, then the result DNA is isolated by limited dilution droplet then detected by PCR amplification. (B) Typical droplet results. The green droplets possess the amplify DNA. (C) CD147 protein level in hundreds of individual cell. Data from (Lin and Elowitz 2016).

Manipulating a single cell can be tedious, especially when it is also necessary to perform analysis on the single cell lysate. The dilution of the cell content makes also more difficult to analyzed and detected single molecule protein, especially for low abundance one. The cell manipulation by microfluidic permits to avoid its content dilution, but also lysis, labeling of the protein of interest and electrophoresis capillary migration for protein separation (Figure 7) (Huang et al. 2007; Willison and Klug 2013). The microfluidic also allows the use of microscopy, permitting the visualization of a single fluorescent molecule for single protein counting. The single protein counting is realized by narrowing the electrophoresis channel to a volume small enough to be recorded by a wide field high sensitivity microscope, and the number of photon bursts detected is assimilated to the number of protein.

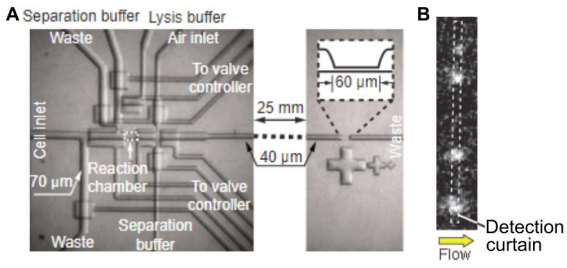


Figure 7: Single protein counting in microfluidic. (A) Microfluidic reactor photography. A single cell can be isolated, lysed and protein labeled in the reaction chamber (left), then the content migrate up to the detection area (right). (B) Detection and counting of single protein passing in front of the detector. Data from (Huang et al. 2007).

Like said previously, the majority of techniques use an antibody to detect and label the protein of interest. It is also possible to observe single protein by high sensitivity microscopy, after fusion of the protein of interest with a fluorescent protein. To count fluorescent fused protein, different methods can be used, like the photo-bleaching process, the calibration curve technique or the super-resolution microscopy (Nino et al. 2017; Coffman and Wu 2012; Verdaasdonk, Lawrimore, and Bloom 2014).

The photo-bleaching captures the irreversible photo-bleaching of fluorophores fused to the protein of interest at single-molecule resolution, with a limit in the number of steps that can be measured, ranging from 5–7 to 15 steps depending on the condition (Figure 8-A). Different kind of protein complex has been investigated by this method, including various membrane-bound channels and receptors (Coffman et al. 2011; Leake et al. 2006; Ulbrich and Isacoff 2007).

The calibration curve works by building a standard curve relating fluorescence intensity to the number of molecules through careful and consistent measurement of fluorescence intensity of one or more fluorescence standard. This method is powerful to quantify a protein in a variety of systems that do not require highly specialized equipment, even if it requires rigorous calibration for precise quantification (Figure 8-B). This method has been successfully used for the quantification of γ-tubulin ring and cytokinesis for example (Erlemann et al. 2012; Wu and Pollard 2005).

Super-resolved localization microscopy, which includes techniques such as PALM and dSTORM, can produce images of structural detail an order-of-magnitude finer than diffraction-limited techniques. The method relies on precisely localizing the spatial position of a single, fluorescent label attached to a target molecule. This typically requires the use of photo-convertible or photo-activatable fluorophores that can be induced to blink in such a way that only a random subset of the labels are visible during each frame. For a sufficiently sparse image, each diffraction-limited spot should be sufficiently well separated, and the

subset of fluorophores may be localized with a precision that scales like 1/√P, where P is the mean number of photons collected from a single blink of a fluorophore (Figure 8-C). Tens of thousands of frames are typically acquired, the spatial coordinates of each fluorophore within each frame extracted, and the resulting data from the stack rendered into a final image(Nino et al. 2017; Thompson, Lew, and Moerner 2012). This technique has permitted to resolve individual actin filaments in cells that revealed two vertically separated layers of actin networks with distinct structural organizations in sheet-like cell protrusions (Xu, Babcock, and Zhuang 2012).

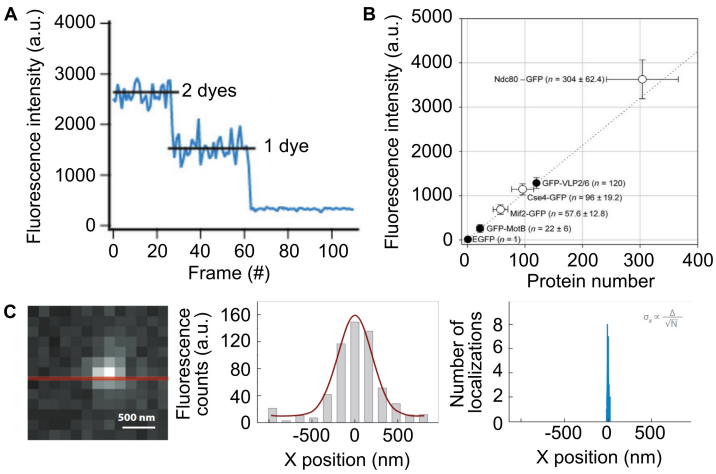


Figure 8: Fluorescent protein counting. (A) Photo-bleaching strategy. One single spot is lighted until each dye composing the spot is sequentially photobleached. (B) Calibration Curve. Spots with a known number of dye are measured to realise the calibration. (C) Super resolution microscopy. A spot is stochastically illuminated, allowing to know the number of fluorescent protein (left). Additionally, it is possible to pinpoint its position assuming that it is really a single molecule (middle and right). Image extract from (Coffman and Wu 2012).

Achieving ultimate sensitivity

If a lot of technique permits to visualize single molecule protein, visualizing a full proteome at single molecule level is a lot more challenging. Nowadays, the most popular and used technique for proteomics is the mass spectroscopy, that has the advantage to be label-free and quick to realize (Yates, Ruse, and Nakorchevsky 2009). However, despite recent improvement, the mass spectroscopy still not able to achieve single cell sensitivity, even less single molecule. To reach the ultimate sensitivity, or the molecule counting level, fluorescence microscopy is actually the most successful tool, especially by using fluorescent fused protein library that has been successfully used to achieve this objective. In Taniguchi et al., they used Yellow Fluorescent Protein (YFP) gene fused with 1018 E.coli gene in as many strain. To speed up the screening process, they used a microfluidic chip that allows the microscope observation of thousands of cell in seconds, and the single protein level was

achieved using the calibration curve technique (Taniguchi et al. 2010). Using a similar microfluidic device, yeast proteome has been investigated using a *Saccharomyces cerevisiae* Green Fluorescent Protein (GFP) library. This library is composed of 4159 genes fused with GFP in as many strains, and have been used to study the osmotic stress response (Chong et al. 2015). Additionally, it has recently been used to study the proteome dynamics of yeast and protein localization (Zhang et al. 2017). If fluorescent protein libraries are an acceptable method to investigate the whole proteome, it is lacking in the capacity for looking at the proteome of a cell at the same time, and make it difficult to visualize multiple protein expression or activation changes in the proteome.

To resolve this issue, one solution is to label the whole proteome unspecifically with a fluorescent probe that will allow single molecule detection. By doing that, we will be unable to identify the protein composing the proteome. To still have an idea about the protein identity, separating the proteins by a physical factor like their mass or their charge will allow a certain degree of discretization and generate a proteome profile. The first part of this thesis will describe the method to label the whole proteome. To investigate this, it will be necessary to develop a test that allows the measurement of the labeling efficiency at the single-molecule level. Once this measure has been fully determined, the protein labeling protocol is investigated using a purified test protein. Finally, this protocol is tested with the cell lysate protein of human cell line. The second part will be consecrate to separate by size labeled protein to obtain a fluorescent proteomic profile, that can furthermore be used for cell line identification.

Part I : Determination of the labeling efficiency

To detect the proteome of a cell lysate, they are two main methods, antibody or reactive fluorescent dye. The antibody method, that mainly uses chip or microchip, is not reliable because of the antibody specificity. In Marcon et al., they tested 1124 antibodies directed against 152 chromatin-related proteins and observed that only 452 of them recognized their target, raising questions regarding whether a given antibody even detects its intended target (Marcon et al. 2015). The other solution, the reactive fluorescent dye, is commonly used in 2D gel electrophoresis in the case of differential gel electrophoresis (DIGE) (Pretzer and Wiktorowicz 2008).

Reactive dyes can permit to recognize different parts of the proteome by changing their reactive part, while the dye can also be changed to be adapted with the microscope illumination setting. In our case, we decide to use a cyanine dye design for a 2D-DIGE experiment, Cy3. Cy3 is one of the most used dye due to its standard microscope illumination setting (same than TRITC), and is available with different classic reactive site, NHS ester, maleimide and hydrazide, that permit to detect respectively primary amine group, sulfo-group and ester group (Figure 9) (Mujumdar et al. 1993). Cy3 is specially used in 2D-DIGE, but also in single-molecule microscopy with FRET experiment in combination with Cy5 (Roy, Hohng, and Ha 2008), making this dye a very good candidate for single protein counting at the proteome level.

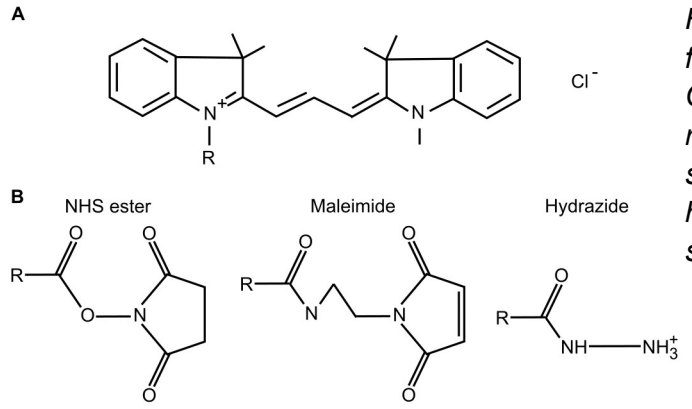


Figure 9: Cy3 and different reactive functions. (A). Chemical structure of Cy3. R indicate the site where the reactive site is bind to. (B) Chemical structure of NHS ester, maleimide and hydrazide respectively. R indicate the site where the dye is bind to.

However, since the proteome is extremely complex and composed of a lot of different protein species, each one with specific biochemical and biophysical properties, this implies that whatever is the labeling method, a portion of the proteome will be lost by precipitation or unlabelled protein (Harper and Bennett 2016). Nowadays, no study or publication focus their attention on the labeling homogeneity of the proteome at a single molecule level, even when this homogeneity is changed by the denaturation and labeling conditions. We then decide to measure the proportion of the proteome that is labeled at a single molecule sensitivity. We

have a publication in submission procedure on this subject when these lines are written (Annex I) and a patent in submission (title: ”Analytical instruments, programs, recording media, and analytical methods”, number 2017-177070, submission 2017/09/14).

Clearing the way for single molecule observation

The detection of single molecule in solid phase, in contrast to gas phase, have been realized since the 1990 with the detection of a hydrocarbon called pentacene in an organic crystal first by absorption (Moerner and Kador 1989), then by fluorescence (Orrit and Bernard 1990), opening the door for single fluorescent molecule experiment and a Nobel prize in 2014. The fluorophore improvement towards a more stable form path the way for easier and more convenient single molecule experiments (Zheng et al. 2014). The main trouble for single molecule experiment is the concentration limit imposed by extended observation volumes (Holzmeister et al. 2014).

The classic strategy to detect single molecule is the fluorescence microscopy visualization of a highly diluted fluorescent sample. To achieve this, we have at our disposition an epi-fluorescence microscope at high sensitivity that allows single molecule detection. The principle is straightforward: we diluted the fluorescent sample until we are sure that they are one molecule or less in the limit of resolution of the microscope. When one molecule is observed with a high sensitivity microscope, the resulting image will present the object deformed by the optical element and define as the point spread function, and represent the limitation of light microscopy (Shaw and Rawlins 1991). This can be troublesome when a higher resolution is required, but can now be reduced by the use of super-resolution microscopy techniques that are able to break this limitation (See Introduction - New from old: single protein detection). However, in the case of quantification or single molecule counting, the resolution is not a trouble if the signal is spread enough on the observed area, and only the sensitivity becomes important.

With the microscope sensitivity being the key for single molecule observation, it is achieved by capturing the maximum of the fluorescence emitted by the fluorophore, without catching back the laser, and then detecting all the photon that has been captured. The amount of signal capture is mainly dependent of the numerical aperture (NA) of the objective lens, that represents its light-gathering ability, with a higher NA corresponding to the collection of more light and the capture of brighter images, and is depending of both the reflection index of the immersion medium, medium separating the sample to the objective, and the angle that defines the collection cone of the lens (Ebenstein and Bentolila 2010). To avoid collecting the laser photons that can come back to the objective lens by light scattering, a dichroic mirror is

necessary, that will reflect light below a certain energy and let pass through lower energy light. The last element is a highly sensitive photon detector like an EMCCD (electron-multiplying charge-coupled device) (Michalet et al. 2007). With these materials and the condition that the fluorescent protein is scattered enough to avoid overlapping, it is relatively easy to detect single fluorescent molecule.

This scattering of the labeled protein needs to be realized on a microscope coverslip that is adapted for the objective lens, with a thickness around 0.15 mm, and can be done by various method. One of the most used actually is the passivization of coverslip using polyethylene glycol (PEG), with a controllable fraction of the PEG being PEG fused to biotin, allowing to control the proportion of protein binding site by the complex avidin/biotin (Chandradoss et al. 2014). However, simpler passivization of the coverslip can be achieved using solely BSA (Di Fiori and Meller 2009).

If the material and knowledge are available, it is quite easy to count single fluorescent molecule. The main trouble came when it is necessary to quantitatively count the fluorescently labeled protein as a method to measure the proportion of labeled protein at a single molecule level does not exist yet.

What means homogeneity at single molecule

To test the efficiency of attachment between a dye and a protein, the coupling efficiency (CE), sometimes call the degree of labeling (DoL), is the golden standard and is generally used in the labeling of fluorescent antibody. This test measures the amount of fluorescent dye over the total amount of protein in the solution by spectrophotometer. The CE then represent the number ratio of proteins to dyes and has been used for labeling optimization or control (Kim et al. 2008). However, since the CE is a bulk measurement, it is not possible to measure the labeling homogeneity, that we define like the proportion of dye by protein. To do so, it is necessary to measure the labeling at a single molecule level to be able to quantify the proportion of protein that is labeled with at least one dye, which we call labeling occupancy (*LO*) (Figure 10). The *LO* represents then a probability of protein labeling and can be used to provide an attenuation factor for estimating absolute protein numbers.

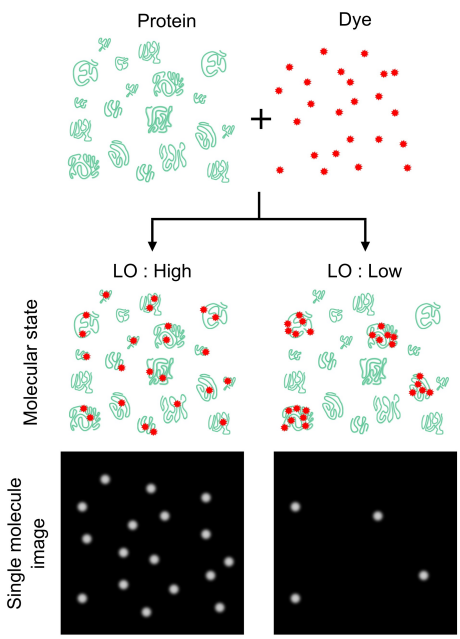


Figure 10: The effect of labeling homogeneity on protein number counting. The protein count number is highly dependent on how homogeneously protein molecules are labeled, rather than the number ratio of proteins to dyes. The homogeneity can be scored using a parameter, labeling occupancy (LO), which defines probability of labeled protein molecules against total protein molecules. This value provides the efficiency of protein counting; i.e. higher LO yields higher count numbers (left) and vice versa (right). LO = 100% is ideal, but even in LO < 100%, LO can provide an attenuation factor for estimating absolute protein numbers.

To realize the measure for *LO*, it is necessary to be able to measure the proportion of labeled protein over the total amount of protein, implying that we are able to measure the total number of protein. To realize this, we decide to evaluate the homogeneity by imaging a microscope coverslip that binds a known density of the sample protein molecules, and by characterizing the number and intensities of fluorescence spots that are imaged. In this case, homogeneously labeled proteins provide a higher number and constant intensity fluorescence spots, whereas heterogeneously labeled proteins provide fewer spots with more variable intensity. By comparing the number of spots between the measured sample and 100% labeled control sample, we can obtain the proportion of proteins yielding fluorescence spots. Fixing a certain density of protein on a glass surface has been already used (Zanacchi et al. 2017; Chandradoss et al. 2014; Funatsu et al. 1995), and in our case, we decide to use a plasma treatment to immobilize a receptor protein, the avidin, and then adding the sample protein molecules that are denatured and biotinylated in advance. The positive control is composed of Alexa-Fluor 488 biotin, to label all immobilize avidin, while the protein sample is labeled with Cy3 dye (Figure 11).

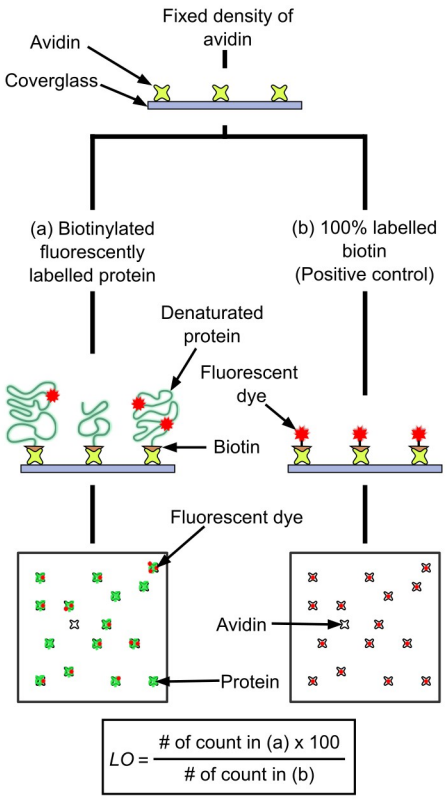


Figure 11: Assay to measure the labeling homogeneity. First, coverslips (a) and (b) are treated with a fixed density of avidin. Second, fluorescently-labeled sample proteins that are biotinylated in advance are attached on the coverslip (a) via avidin. In parallel, 100% fluorescently-labeled purified biotin is attached on the coverslip (b). Third, the coverslips are imaged by single molecule fluorescence microscopy to obtain the number and brightness distribution of fluorescence spots. And fourth, the homogeneity parameter, labeling occupancy (LO), is calculated from the ratio of spot numbers in (a) to (b).

The protein sample, BSA for the preliminary test, was labeled following a standard protocol (Nanda and Lorsch 2014). Briefly, 1 µg of BSA was mixed in 100 µl of buffer containing PBS 1x at pH 7.4, 0.1% Tween 20, 0.1% SDS and 1 mM DTT for 5 min, then 1 µl of 2 µg/ml of Cy3 NHS-ester succinimidyl was added. The labeling reaction occurs during 15 min in the dark at room temperature, allowing the dye to react with the primary amine of the protein, like the amino-terminal function or the lysine residues. After completion of this step, 400 µl of 0.8 M HEPES pH 7.4 was added to adjust the pH for the biotinylation step, then by adding 100 µl of 19 mg/ml biotin-PEG2-amine and 5 µl of 20 mg/ml EDC and by incubating in the dark at room temperature for 1 hour, allowing the biotinylation of the carboxy-terminal function of the protein. To remove unreacted labeling reagents, the protein solution was added to an ultra-filtration device that permits to remove all component with a mass inferior at 10 kDa and re-suspend in 50 mM HEPES pH 7.4. To limit the number of avidin sites on the coverslip, we decide, after plasma treatment, to spin-coated the coverslip with 200 µl of a solution of 10 ng/ml avidin and 2 mg/ml BSA in 5 mM HEPES at a speed of 1000 rpm for 30 seconds. This technique allows the formation of a very thin layer of buffer on the glass surface (few nanometers) and then reduce the number of avidin protein by area unit to a density of approximately 1 avidin by 100 μm^2, corresponding to 278 avidin proteins by microscopic frame (27889 μm^2). They are however big day-to-day variations in this density,

mainly caused by the imprecision in the dilution of avidin, making necessary to measure the avidin density for each experiment. The native BSA is present in the solution to passivated the coverslip surface and by saturating the non-specific site in order to avoid sample protein to react with them. After this avidin coating step, the sample protein, fluorescently and biotinylated labeled, is diluted if necessary and added to the glass surface and let react for 15 minutes, followed by a distilled water wash to remove free protein sample. The strong affinity between the avidin and the biotin site permit to isolate the sample protein for single-molecule fluorescence visualization.

The setting to visualize single molecule fluorescence is using an inverted epi-fluorescence high sensitivity microscope with a 60x oil-immersion objective lens (Walter et al. 2008). Fluorescence was induced with wide-field illumination by a 488-nm Argon ion laser or 560-nm fiber laser with a power density at the observation area set to 117 W/cm^2. The fluorescent biotin was imaged through a 495 nm dichroic mirror and an emission filter of 520(±14) nm to filtrate out scattering light. The protein sample labeled with Cy3 was imaged through a 561 nm dichroic mirror and an emission filter of 617(±31) nm (Figure 12). Each and every image was acquired with an emission time of 100 ms. For each condition, more than 100 images were taken for statistical analysis. In addition, every experiment was realized with the imaging of at least one negative control, that is an avidin-free glass, and the positive control, avidin Alexa-Fluor 488 biotin reaction, to confirm the unspecific binding and the maximum density respectively.

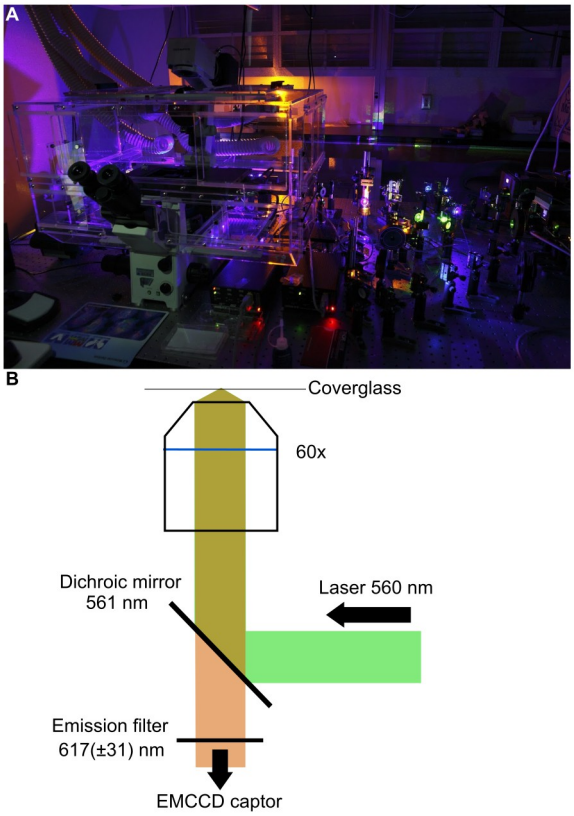


Figure 12: Microscope setting for single molecule visualization. (A). Photography of the microscope. The laser line, where all different laser focus on the same point, can be see at the right of the image. Laser beams are expanded and homogenized in the back of the microscope before being directed inside for the sample illumination. (B) Setting for Cy3. After entry in the microscope, the 560 nm laser is reflected against a dichroic mirror and focus by a 60x objective lens at 0.6 µm of the coverslip surface. Fluorescence from the sample is catched back by the objective lens and pass through the dichroic mirror, then is filtered to remove scattering light and raman effect before photon count on a photo multiplicator

The obtained images were first processed with a laser illumination heterogeneity compensation to uniform fluorescence counts over the entire imaging area (Taniguchi et al. 2010). Then the images are processed with the rolling ball algorithm with a ball radius of 50 pixels to subtract a background (Sternberg 1983), and a bandpass filter to extract spots of 1-20 pixels. The image was binarized using the Triangle threshold, and processed in an image filter that removes spots smaller than 2 pixels square. Then, the spot number was calculated from each image, which was used for the *LO* calculation following the formula in Figure 11. Also, the histogram of fluorescence counts for spots was analyzed by re-binned based on peaks observed in the histogram, in order to represent the number of dyes in spots. n_{dye} was calculated from the mean of this distribution, and the histogram integrated 100 - *LO* at zero position to obtain a probability density function of the number of dyes in a protein.

We first need to check if the avidin is well fixed on the coverslip and that the avidin-biotin reaction can occur. To test this, we visualize the positive control, or condition (b) of Figure 11, that is composed of Alexa-Fluor 488 biotin in large excess. Since the avidin possesses 4 biotin binding sites (Livnah et al. 1993), and that the fluorescent biotin is a small molecule of 3.4 kDa limiting the steric hindrance, we expect that the avidin can bind a maximum of 4 fluorescent biotins. When looking at the image data, we can observe different intensity spots, representing the recognized avidin active site population (Figure 13). This population is

better visualized by looking at the histogram of spot intensity, where it is possible to see up to 4 peaks, each one representing an event binding. At this stage, we suppose that all the avidins are detected by at least one fluorescent biotin, and since the link between the avidin and biotin is very strong, no free fluorescent biotin is expected to be present (Green 1963). For confirmation, the incubation time was elongated from 15 min to 1 hour, with no change in count number or spot intensity. In addition, if the effect of multiple biotin binding to the avidin can be visualized in this condition, we expect that in the case of a biotinylated and fluorescently labeled protein sample the steric hindrance will lead to the fixation of only one protein sample to each avidin. This has been confirmed in another experiment where a single dye is bound to the protein sample, leading to a homogeneous intensity spot population (Figure 26). The negative control permits to confirm that the biotin does not react with the plasma treated coverslip and that the excess is efficiently washed away.

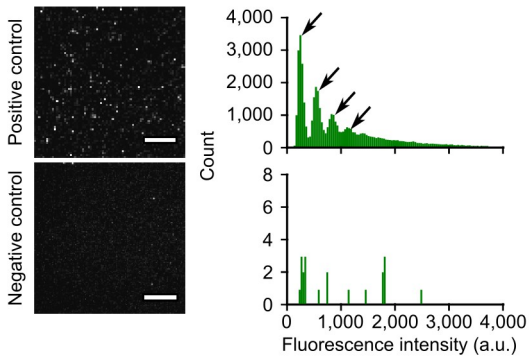


Figure 13: Controls for avidin-biotin reaction. The positive control is composed of avidin treated coverslip and fluorescent biotin, with left is the image data and right is the histogram of spots intensity, with the arrow indicating the number of avidin active binding sites recognize. The negative control is similar, with a coverslip not treated by avidin, and displaying the minimum fluorescence spots count.

The second point to check is if the *LO* assay displays the same behavior than the traditional *CE* technique. Since the *CE* is a bulk fluorometer measurement, the ratio of the dye to the protein in the sample solution is measured to be 0.91 dye/protein in the original labeled BSA. It is necessary to convert this *CE* value, that is a ratio, to a value comparable to the *LO*. If we suppose that the labeling process of a protein follows a Poisson law, it is possible to calculate the proportion of labeled protein, that we call LO_{CE}, by the equation:

$$LO_{CE} = 1 - \left(\sum_{\mu=0}^{\infty} \left(CE^{\mu} * e^{-CE} \right) / \mu ! \right)$$

where CE is the coupling efficiency and μ is the number of dye. By applying this equation with μ = 0, we can calculate the proportion of unlabeled protein, then deduce the labeled proportion of labeled protein, corresponding to 55.8% for the labeled BSA.

The labeled BSA was mixed at different ratios with unlabeled BSA, ranging from 1:3 to 1:0, to artificially change the proportion of labeled protein and measure with the bulk experiment

and the *LO* assay. The single molecule images of the *LO* assay and spots histogram display a heterogeneous population of labeled BSA (Figure 14-A) with a *LO* (LO_{Obs}) of 33.1%, significantly lower than the LO_{CE} value, suggesting that the molecular labeling event is more heterogeneous than the Poisson process. By increasing the amount of unlabeled BSA, we observe that both the LO_{Obs} and LO_{CE} diminish in a linear relationship (Figure 14-B).

Changing the ratio of labeled/unlabeled BSA impact the bulk measurement of the LO_{CE}, but will normally not impact the number of dye by protein detected at a single molecule level. To calculate this, the number of dyes binding to proteins in spots (n_{dye}) can be obtained from the ordinal number of the peak in the fluorescence intensity that the spot has. The averages of n_{dye} among all the spots ($\overline{n}_{dye}$) were found to be mostly constant at any mixing ratios (Figure 14-C), that is coherent with the fact that we dilute the labeled protein, not the labeling, and that detecting at a single protein sensitivity permits to bypass the dilution since we observe solely the labeled portion. This also suggests that only one biotinylated labeled protein binds to the avidin. We also found that the probability density function of the number of the dye binding to a protein exhibited a higher probability of proteins binding either zero or multiple dyes than expected if it were a Poisson process. The principal explanation is that some protein molecules in the sample have deficient reactivity with the label, due to a reason such as incomplete denaturation.

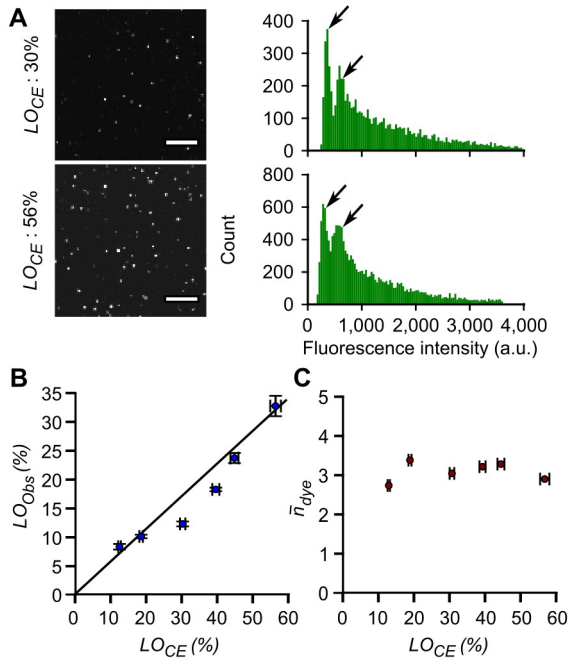


Figure 14: Evaluation of labeling homogeneity. (A) Raw data. Image data (left) and histograms of fluorescence intensities of spots (right) when measuring a mixture of fluorescently labeled and unlabeled BSA at different ratios are shown. Arrow represent conjugation of different discrete numbers of fluorescent dyes. Scale bar = 10 µm. (B) Comparison between the estimated LO (LO_{CE}) and observed LO (LO_{Obs}). The black line shows a linear regression (R2 = 0.95). (C) Comparison between LO_{CE} and the average numbers of the dyes binding to proteins in spots ($\overline{n}_{dye}$), which were estimated from the fluorescence intensity histograms.

Control experiment and robustness test

We observed that the LO_{Obs} follows the expected comportment. However, we still need to confirm that the dots that we are counting are single-molecules, that we do not miss-count the number of dots by a bad threshold method, that the LO_{Obs} is not sensitive to the protein sample or the dye concentration, that we sample an area large enough to avoid local variation in the avidin coating and that day-to-day variation in the *LO* measurements is minimal.

To confirm that we really observe single-molecules, we realized a photo-bleaching experiment like describe in the introduction. To do this, we took one hundred images of the same position of the protein sample BSA with a 100 ms acquisition time, then look at individual spot intensity evolution (Figure 15-A). The majority of the spots are turned off after 3 seconds of exposure in a single step like for spot 1 and 2 in Figure 15, signifying that the protein sample is single labeled by the Cy3 dye. Some spots also display double photo-bleaching steps (spot 3 and 4) indicating a double-labeled sample protein, with a second photo-bleaching step around 6 seconds. They are also few spots that are still present after the 10 seconds exposure that can be highly labeled protein, with three or more dyes, but representing a small proportion of the total spots. The different photo-bleaching steps are easily identified when looking at the plot intensity (Figure 15-B), confirming the single protein isolation on the coverslip and allowing the counting of the sample protein.

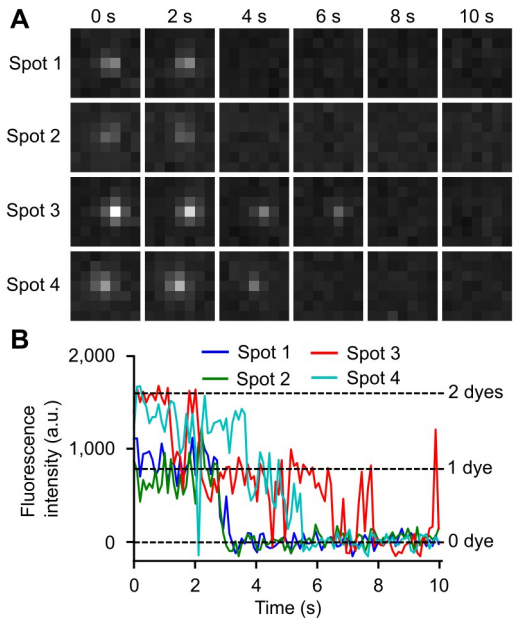


Figure 15: Single molecule photobleaching experiment. (A) Spot evolution during photobleaching. One position was acquired during 100 ms for 10 s, and an 7x7 pixel sub-image corresponding to a spot is extracted. (B) Intensity profile. The average intensity of each sub-image is calculated for each time to display the change in fluorescence intensity during the photobleaching process.

On the acquisition side, one point to confirm is that our counting method is not over-counting or miss-counting the number of labeled protein during the image threshold setting. The threshold method to count dots is accurate if the dots are in low density on the observed area and that the signal-to-noise ratio (SNR) is large enough. In our case, the microscope allows the observation of single-molecule, even with a single dye on them clearly, as observed in the photo-bleaching experiment (Figure 15). This is realized by sectioning the histogram intensities of pixels identified as 'signal' and compared to the same number of 'background' pixel (Figure 16). We can observe that the background pixels present a very narrow peak, around 3,000 fluorescence intensity, while the signal pixels present a wider range, coherent with the different number of fluorescent dye that can participate to the global signal. We also observe that the two histograms do not overlap and that the separation between the two of them is large enough to accurately place the threshold value, and those small variations in this value will have only a minor impact on the global count, and then the *LO* estimation. In addition, the low density in spots of our assay can be estimated by using the positive control and is around 10^{-2} dot/pixel, that corresponds to a low density. This low density reduce the probability that two avidin proteins are so close together that we cannot distinguish themselves, resulting in an under-count. The combination of the clear separation between background and signal and the low density of single-molecule permits the utilization of the threshold method for single-molecule counting, and then *LO* determination.

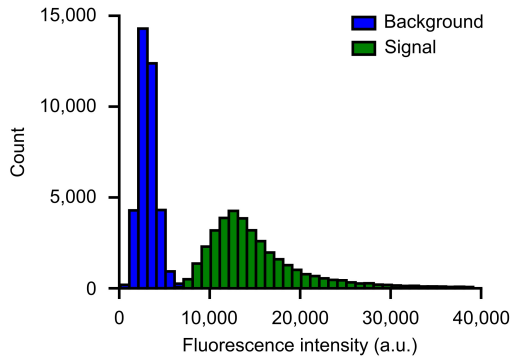


Figure 16: Pixel values from the background and detected spots. 36,000 pixels values from detected spots as signal are compared to values of 36,000 pixels randomly chosen from background.

Another point to confirm is that not only the fluorescent biotin but also the biotinylated and fluorescently-labeled protein sample is specifically reacting with the fixed avidin and not simply adsorbed on the glass surface by a hydrophobic reaction. We applied a different concentration of protein sample, biotinylated and labeled BSA, to measure if we reach a constant value in the number of counted spots even if we keep increasing the sample concentration. In Figure 17, we observe that a concentration higher than 0.76 fmol of BSA

leads to the stabilization in the number of counted spots, signifying that the *LO* assay is insensitive to the sample concentration. Below this value, the ratio of the number of avidin sites to the number of sample protein is probably too low for saturation in a short amount of time. An amount of 76 amol of BSA correspond to approximately 4.6×10^7 sample protein on the coverslip, and the total number of avidin is estimated to be around 10^7 by using the positive control count as a reference, meaning that the active sites are hardly saturated and will correspond to an intermediate saturation effect.

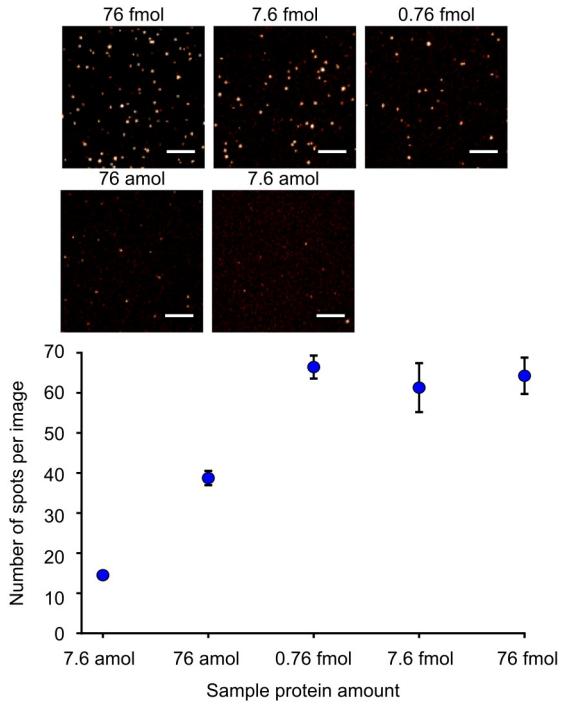


Figure 17: Insensitivity of the sample protein amount on the output spot number in the assay system. In top is shown a representative image when measuring different quantities of fluorescently labeled BSA were analysed. Scale bar = 10 µm. In bottom is shown the mean numbers of spots per image. Insensitivity of the spot number was observed at quantities greater than 0.76 fmol. Error bars represent standard error of the mean.

A similar point to control is to confirm that the ratio of reactive dye to protein sample in the labeling reaction is in excess, and verifying that the increase of this ratio results in a further increase in the *LO*. To realize this, we realized measurement of the *LO* with a variable ratio of dye/protein, ranging from 0.2 to 50. We observe that the labeling efficiency follow a Poisson law at a low number of dye/protein, then stabilize at a higher ratio (Figure 18). This experiment confirms that the labeling reaction is saturating to measure the *LO*. In addition, we fit the data point a curve following the expression $LO = 1 - e^{-ratio*efficiency}$, corresponding to a mathematical resolution of the *LO* equation simplification in the case or $\mu = 0$. In this equation, we need to add an additional factor, called 'efficiency', that is the variable parameter that is optimized for the fitting curve. This parameter corresponds probably to the proportion of the dye that can really label the protein and is equal to 22.5%. This means that for the BSA labeling, in these conditions, only 22.5% of the available dye reacts with the

protein sample. This number can be explained by the strong basic condition (pH12) of the labeling reaction buffer, that will quickly hydrolyze and inactivate the NHS-ester function of the dye.

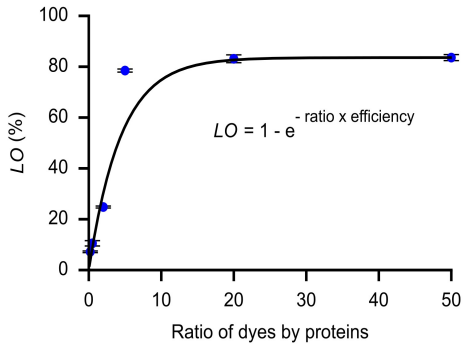


Figure 18: Effect of concentration ratio of the dyes to proteins during the labeling reaction on the LO. Data for labeled BSA sample denatured using Condition J (Table 2) is shown. The data shows the measured LO was saturated at greater than 5 dyes/protein, which includes the range of our labeling conditions (20 dyes/protein). The fitting curve is given by: 1 - $e^{-ratio*efficiency}$, where 'ratio' represents the ratio of dyes to proteins, whilst 'efficiency' represents the proportion of the dye that can bind to the protein and is optimized into 22.5%.

Another important to check is the minimum number of images to acquire and analyze, since the avidin coating of the coverslip can present a heterogeneity, to obtain statistically strong results. To realize this test, we used the data of a sample presented in Figure 26, that have 100 images to calculate the *LO* (Figure 19). The data are presented as a number of dots count by image, and the *LO* is generally estimated by using the average of this value. For this experiment, we randomly selected a given number of image 10,000 times and calculated the mean value and the standard variation of this bootstrapping. Like expected, the result displays a similar average for all number of image, but with decreasing standard error when the number of sampling increase. In our case, the minimum sampling realized was 20, corresponding to an error of ±4.6% in the *LO* measurement, while the median measurement corresponds to 52 images, so ±2.3%.

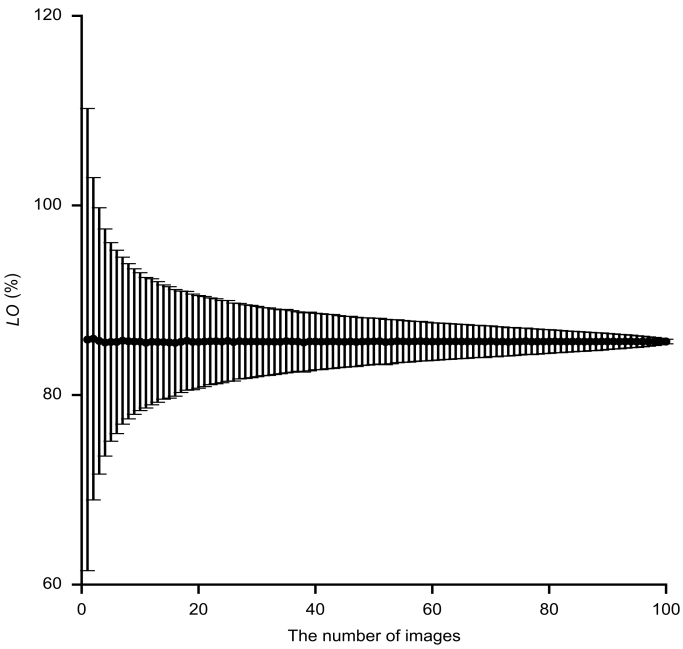


Figure 19: Effect of the image number on the LO accuracy. The accuracy is evaluated by standard deviations of LO from 10,000 times bootstrap sampling from 100 images of the 33-42 kDa sample presents in Figure 26.

Finally, we want to confirm that independent experiment using the same sample realized on different day permits to obtain similar *LO* value (Figure 20). We measure the *LO* of two samples from Figure 26 to confirm this, and see minimal variations, indicating that the *LO* measurement is replicable.

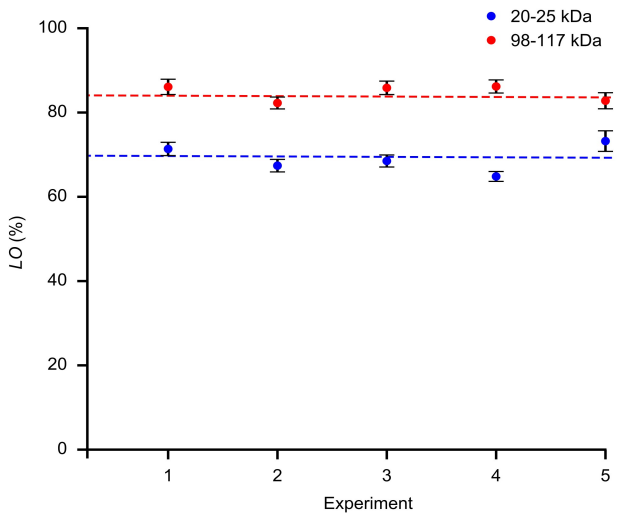


Figure 20: Reproducibility of LO measurements. LO was measured for 5 independently prepared samples for the 20-25 and 98-117 kDa proteome samples. Data are mean ± s.e. 51, 56, 75, 48 and 152 images were analyzed for the 5 replicates of the 20-25 kDa sample and 74, 101, 79, 102 and 120 for the 5 replicates of the 98-117 kDa sample from Figure 26.

We confirmed that the dots that we measure and count by microscopy were single-molecules, labeled by one or multiple dyes (Figure 15) and that the threshold method used for the counting algorithm is justified (Figure 16). We have shown that the sample protein needs to be in excess compared to the number of avidin sites and that the *LO* is

independent of the excess value (Figure 17). We also have shown that the labeling reaction needs to present an excess of dye by protein, at least 5 to label available protein, and by security, we choose 20 dyes/protein (Figure 18). Finally, we tested the robustness of our measurement by investigating the sampling effect (Figure 19) and the day-to-day variations (Figure 20) and observed that the *LO* presented small variations. This permits to conclude that the *LO* is a real parameter, and that may be modulated to obtain the desired protein labeling homogeneity.

Improving the labeling

Using classical protocol, the test protein BSA is lowly and heterogeneously labeled. However, the BSA amino-acid composition reveals that it possesses 59 lysines plus the amino-terminal function, for a total of 60 actives sites by protein. We do not include the arginine, asparagine or glutamine residues in the active site count since the primary amine need to be protonated for its reaction with the NHS-ester dye (Cline and Hanna 1988). However, we observe that the BSA labeled portion have around 3 dyes by protein, a lot less than the number of actives sites. This low number can be caused by the accessibility of the protein lysine, and this accessibility can also explain the labeling heterogeneity, with a population of protein in a state that permits a better accessibility to some lysine. They also have the effect of the dye on the protein stability, since the Cy3 dye is hydrophobic, its binding in a large amount can impact the solubility of the target protein (Zanetti-Domingues et al. 2013; L. D. Hughes, Rawle, and Boxer 2014). In addition, too many dyes in the same volume can present interferences, leading to a quenching effect. Altogether, we want to define the condition that permits, for the test protein BSA, to be homogeneously labeled with a constant number of dye by protein low enough to not impact the protein solubility. One obvious parameter to change is the denaturation state of the protein: if the all the proteins are in the same denaturation state, the labeling will likely be homogeneous, and controlling this state can allow limiting the number of dye by protein.

The denaturation of a protein is the loss of the quaternary structure, tertiary structure and secondary structure which is present in the native state of the protein and can be caused by extreme pH condition, inorganic salt concentration, organic solvent, pressure or heat. In our case, we decide to use high pH for protein denaturation, allowing at the same time the protonation of the lysine residue of the protein, that possesses pKa of 10.07. This high pH has the drawback to quickly hydrolysate the reactive site of the dye for the labeling, and imply to add a higher concentration of dye and that a long incubation time is useless to improve the labeling efficiency. The presence of detergent and surfactant is classic in a denaturation buffer, with the SDS acting like a surfactant and a denaturation agent by

linearizing the protein, while the Tween 20 improve the solubilization of the protein. When the pH of the environment of the protein change from the cellular pH (7.4), protein precipitation can quickly occur, and to mitigate this effect, 3-[(3-Cholamidopropyl)dimethylammonio]propanesulfonate (CHAPS) is classically used to increase the solubility and stability of proteins, especially the membrane one. We decided to test different condition with high pH and combination of buffer and detergent, and the different buffer compositions are shown in Table 2 and compare to the original condition (condition A). The buffer concentration is kept low to allow a quick pH adjustment for the following step of biotinylation, that necessitate a slightly acidic or neutral pH to occur.

Condition	Buffer	DTT	SDS	Tween 20	CHAPS
A	1x PBS pH 7.4	1 mM	0.1 %	0.1 %	0 %
B	50 mM borate pH 12	1 mM	0.1 %	0.1 %	0 %
C	50 mM borate pH 12	1 mM	0.1 %	1 %	0 %
D	50 mM borate pH 12	1 mM	1 %	0.1 %	0 %
E	50 mM borate pH 12	1 mM	1 %	1 %	0 %
F	50 mM borate pH 12	10 mM	0.1 %	0.1 %	0 %
G	50 mM borate pH 12	10 mM	0.1 %	1 %	0 %
H	50 mM borate pH 12	10 mM	1 %	0.1 %	0 %
I	50 mM borate pH 12	10 mM	1 %	1 %	0 %
J	50 mM borate pH 12	10 mM	1 %	1 %	2 %

Table 2: Labeling buffer composition at each condition.

Bulk measuring the LO_{CE} of the three different conditions leads to similar results around 40%, meaning that the labeling is either completely similar with inefficient denaturation condition, or more heterogeneous with a minor population presenting high dye/protein ratio, or a more homogeneous with a major population presenting the same dye/protein ratio. In contrast, LO_{Obs} significantly increased in the new conditions compared to the original condition (29.5%) (Fig. 3B), indicating that the labeling homogeneity can be improved by the labeling conditions. Addition of our tested four supplements (DTT, SDS, Tween 20 and CHAPS) respectively provided positive effects on LO_{Obs}, which reached 82% when adding all four supplements together. Consistent with this, histograms of fluorescence intensities of each spot were less spread and less skewed to smaller values in the new conditions, suggesting that less and constant numbers of dyes tend to bind to single proteins (Figure 21). We can see up to three peaks, which correspond to 1, 2 and 3 dyes by protein. The good separation between the peaks can signify that they are a very limited quenching effect of the dye in the condition I and J compared to A, leading to a better appreciation of the dye number by protein.

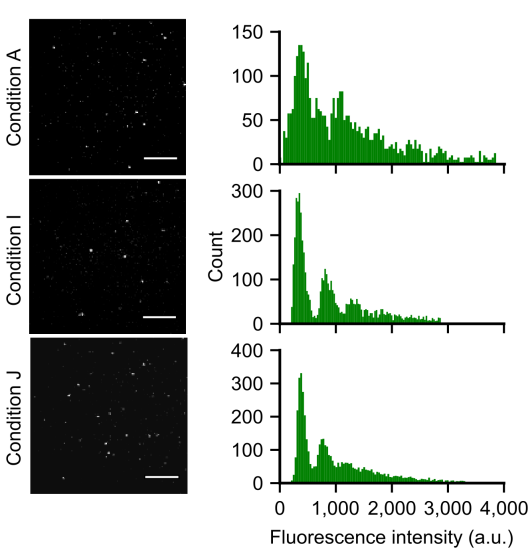


Figure 21: Raw data of different labeling conditions of BSA. Imaging data (left) and spot fluorescence intensity histograms (right) for the three differents protein denaturation buffer described in Table 2.

The probability density functions of the number of dye by protein can be estimated from the discretization of the histogram(Mutch et al. 2007). We realize it by binning the histogram based on the width of the first peak, starting from 1 dye up to an arbitrary limit of 7 dyes. It was possible to add the zero dye by a cross product between the total number of counted spot and the LO_{Obs} to obtain the total amount of spots in the image, and from this, the number of protein unlabeled. From this discretized histogram, the $\bar{n}_{dye}$ and the derived Poisson law can be calculated (Figure 22), and display that the all-or-none dye binding fashion from condition A is reduced, but instead, a constant-number type of binding dominates in other condition. We expect that such constant-number binding is caused because reactive labeling sites or lysine residues in proteins are filled by dyes.

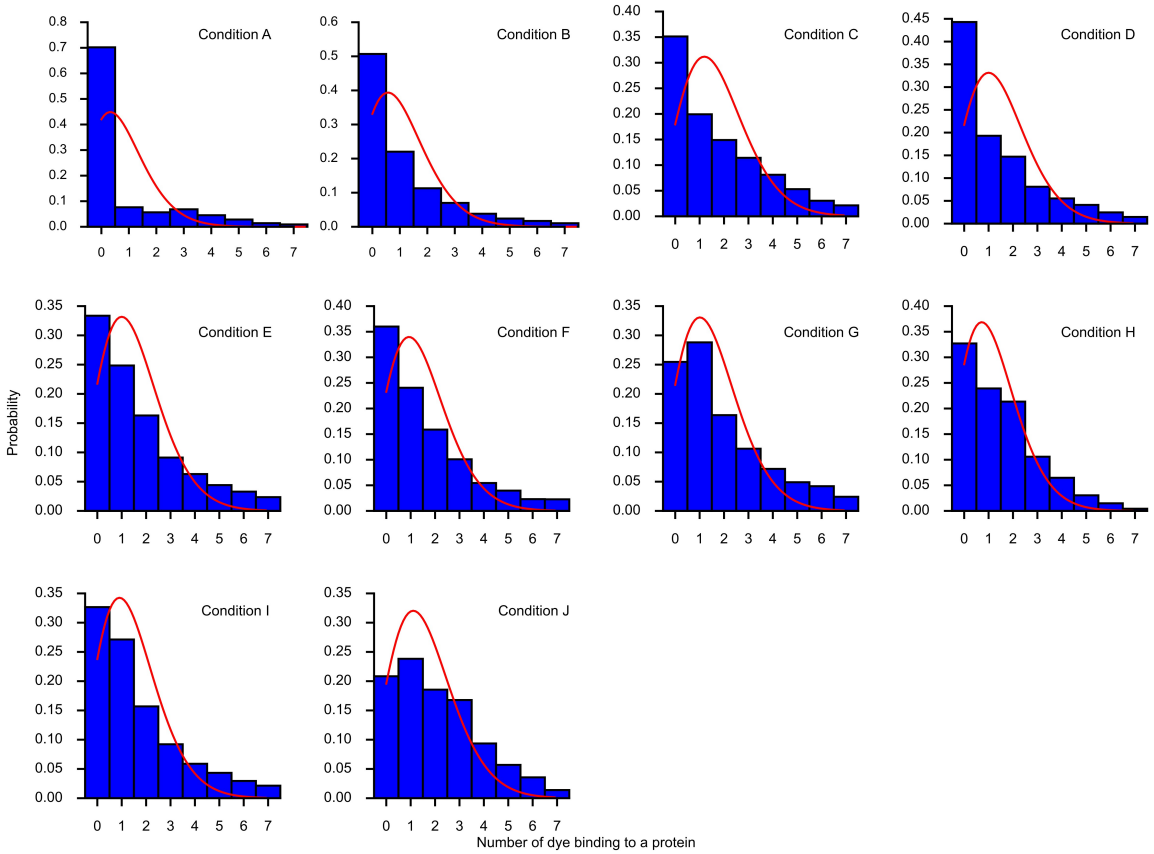


Figure 22: Probability density functions of numbers of the dye binding to proteins of labeled BSA at different labeling conditions. The probability density function (blue) and the Poisson distribution curve with a mean of the function (red) are shown.

In conclusion, the condition J, allowing the best denaturation and solubilization of protein, permits to obtain a better proportion and a more homogeneous population of labeled BSA at a single protein level, with approximately 75 % of labeled BSA with 2 dyes (Figure 23).

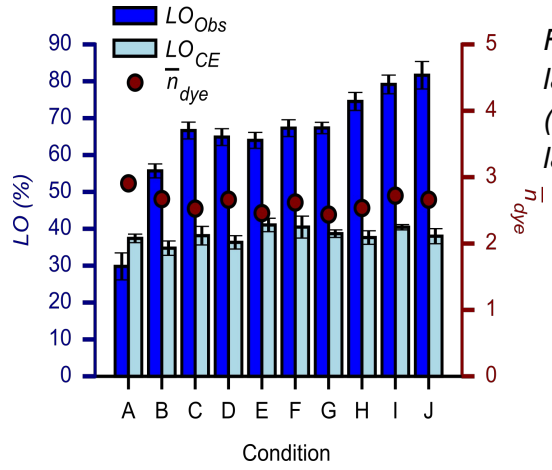


Figure 23: Labeling homogeneity at different labeling conditions of BSA. LOObs (blue), LOCE (light blue) and $\overline{n}$ dye (dark red) at different labeling conditions.

Proteome scale labeling

Single molecule protein counting and spatial visualization is done since 1995 (Funatsu et al. 1995) and is today commonly realized with the fluorescent labeling of antibody. However, labeling the whole proteome of a cell lysate is another scale and it is important to keep in mind that it is not possible to label every kind of protein. They are too many of them with a too important range in term of biophysical and biochemical properties, implying that a portion of them can potentially precipitate. In addition, we increased the challenge a little more by realizing a one-step buffer, where the cells are lysed, protein denatured and then labeled in a single step, without any intermediate step like protein extraction.

Since the dye detected primary amine, every component of the cell with a free amine will be recognized by the dye, including small molecules like single amino-acid or short peptide, multiplying the potential number of target and necessitating to separate them from the proteome sample. By keeping the current protocol, that realizes a size based purification, we can remove all these small molecules that we are not interested in. We choose the condition J (Table 2) as the one-step buffer, since the presence of Tween 20 permits to break the cell wall and the high pH, SDS and DTT to denature the protein content.

We first need to confirm that it is possible to visualize the proteome of a cell lysate. Hela cells are one of the most know and classical cell line and have been used for the cell lysate. Cells were classically growth in DMEM supplemented with fetal bovine serum and antibiotic mix until reaching 70 %, and exponentially growing cells were collected by treatment with 0.1% trypsin, washed with 1x PBS, and re-suspended in 1x PBS at 10^5 cells/ml. This wash is critical to remove the trypsin and other protein that can be labeled. 1 µl of the cultured cells, corresponding to 1000 cells, were lysed and denatured by mixing with 100 µl of buffer condition A or J for 5 min, then labeled by adding 1 µl of 2 µg/ml Cy3 NHS-ester dye and by incubating for 15 min in the dark at room temperature. A classical SDS-PAGE was realized using 15 µl of labeled cell lysate, with the particularity of not boiling the sample prior migration to avoid degradation further degradation of the cell lysate proteome, and migration occurred until the front of migration exited the gel. The gel was then first visualized using the Cy3 setting of a gel viewer for the fluorescence channel, then stained using colloidal Coomassie blue (CCB) to visualize the whole proteome (Figure 24).

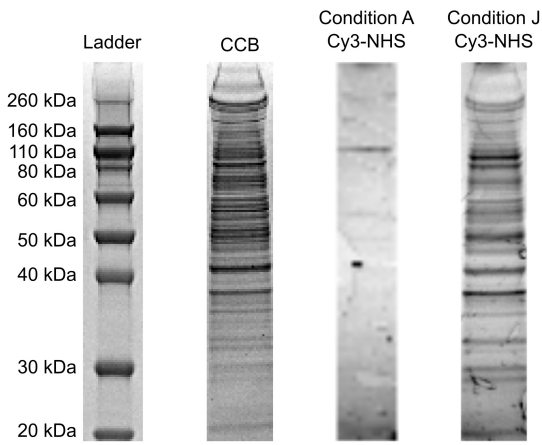


Figure 24: Comparison of proteome staining methods. Images of SDS-PAGE gels stained by CCB, by Cy3-NHS in condition A and J and compare to the ladder image. Condition J provides more efficient labeling of proteome than condition A. The heterogeneous pattern in condition J compared to CCB would be caused because nlys are highly varied depending on the protein species rather than molecular weights.

We observe that the condition A fluorescent labeling is very poor, with only a few bands of protein labeled, implying that the labeling is extremely heterogeneous. It is clearly evident when comparing the CCB and the condition A channel. In the other hand, the condition J detect clearly more protein types, with a comparable pattern than the CCB. We can, however, note small differences, especially in the low molecular weight protein, lower than the 40 kDa. The comparison between the different profile reveal some similarities between the Cy3 labeling and the CCB staining, even if their labeling principle is different, with a direct and covalent reaction for the fluorescent staining and a non-covalent based staining that binds to proteins by a combination of hydrophobic interactions and hetero-polar bonding with basic amino acids (Georgiou et al. 2008).

The direct comparison between the CCB and the condition J migration profile is based on the fact that it is theoretically the same proteome profile, and then permits the same analysis. However, the migration and image acquisition leads to some artifacts that need to be removed for a strict analysis. The art of baseline correction is complex, and we decided to use a simple original approach. We first wish to remove global trends in the signal, since we expect that protein peaks are limited in their width. We realized this by making the rolling mean on the data after filtration by a local Otsu filter. This baseline was furthermore smooth using a Savgol filter (Savitzky and Golay 1964), allowing us to subtract this baseline (Annex III).

In addition to this baseline correction, we also estimated the theoretical protein abundance by using the protein expression database paxdb[4]. This database possesses a compilation of quantitative proteomics over a wide range of cell types and organisms, comprising Hela cell, even if the coverage is approximately 40%. Using the protein ladder, it is possible to calculate the interval of protein molecular weight theoretically present in one pixel, or one

point, of the profile. The normalized sum of the abundance of each protein inside this interval permits to obtain the theoretical migration profile, and is then compared to the real data from the CCB and the fluorescence condition J. When looking at the protein abundance database of Hela cells, it becomes obvious that the majority of the most expressed protein has a low molecular weight, like the cofilin-1 at 18.5 kDa, with the top 10 presenting an average molecular weight of 17.6 kDa, lower than the limit in our SDS-PAGE at 20 kDa (Table 3). The high molecular weight protein is generally less expressed, with the moesin being the only big protein, at 67.8 kDa, highly expressed, but still 10 times less than the cofilin. This implies that a lot of protein is even not considered in the analysis and that it is difficult to directly compare the estimated to the experimental profile. In addition, the experimental data protein bands tend to be wider than one-pixel width (see the ladder in Figure 24), making the estimated data looking a lot sharper than the experimental data.

ID	Protein Name	MW (Da)	Abundance (ppm)
P23528	COF1 Cofilin-1	18502.90	33062168.10
A8MX94	Glutathione S-transferase P	19481.03	22732898.15
P37802	TAGL2 Transgelin-2	22392.08	20103702.92
P05387	RLA2 60S acidic ribosomal protein P2	11665.29	17892707.42
A8MUS3	60S ribosomal protein L23a	21916.70	17887495.21
Q71DI3	H32 Histone H3.2	15388.47	11330735.76
P68431	H31 Histone H3.1	15404.54	11250129.19
Q6NXT2	H3C Histone H3.3C	15214.20	10677335.17
A6NIW5	Peroxiredoxin 2, isoform CRA_a	15138.75	10658257.18
B5MCP9	40S ribosomal protein S7	21312.36	9360266.57

Table 3: Top 10 of protein abundance in Hela cell.

After data processing, the proteome profile obtained from the database, the CCB and the fluorescent reactive dye can be compared to each other (Figure 25). We can observe that the CCB and the reactive dye profile are quite similar, presenting same peaks sharpness and position through the whole protein molecular weight range, even if the reactive dye profile look more blurry in the original image (Figure 24). The biggest difference consists in the relative intensity, with the reactive dye displaying a higher intensity than the CCB in low molecular weight, from 20 to 40 kDa, and the CCB presenting a higher intensity than the reactive in high molecular weight protein, more than 100 kDa. The estimated profile, however, found a lot more abundant protein at a low molecular weight, and only a few proteins species of high molecular weight present spikes big enough to be visualized. When

transforming the data in a log scale, it is possible to observe a general trend where higher is the molecular weight of the protein, lower its abundance, especially from 100 kDa.

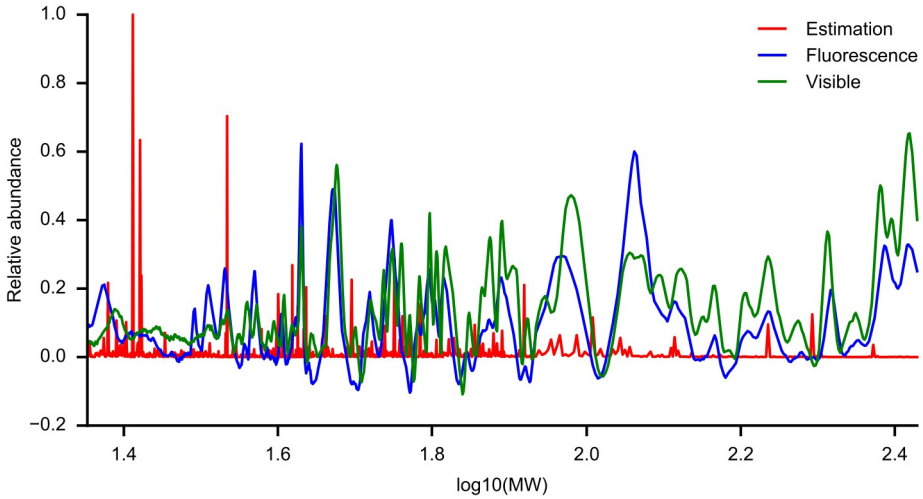


Figure 25: Protein abundance compared to SDS-PAGE migration profile. The fluorescence profile from the condition C is compared to the CCB profile after baseline correction and normalization. The estimated abundance of the protein is calculated from the database paxdb[4] with the Hela proteome.

These results confirm that the labeling protocol using the condition J permits to label a representative portion of the proteome without visible bias. We then repeat this SDS-PAGE experiment but focus our attention on 4 different visible fractions from bands after migration and the whole migration length in a sequential order. From these bands, the proteins are extracted by electro-elution for 2 hours in a migration buffer, then biotinylated as previously described to realize the *LO* measurement of these fractions. In parallel, the *LO* of the whole cell lysate was measured for comparison (Figure 26). The molecular weights of the fractions were chosen based on the previous experiment to try to represent the different molecular weight comportment with the 16 and 23 kDa band for the low molecular weight protein, the 55 kDa for the middle molecular weight while the 120 kDa is for the high molecular weight. These molecular weights are an obvious approximation, and we estimated that we capture in reality 2% around the target, meaning many protein species, and we reflect this aspect in the bio-informatics analysis by selecting, for each band, the proteins in the molecular weight range.

When looking at the raw images, we can observe that the whole cell lysate *LO* looks noisier than the corresponding BSA image (Figure 21). This is explained by the difference in the sample purity, with the cell lysate being not purified, they are a higher chance that some cell waste or cell fragments are present on the glass surface, leading to more noise. However,

the image acquisition and analysis can permit to not take into account these kinds of artifacts. In the other hands, these kinds of artifacts are extremely unlikely in the case of the electro-eluted proteins since they are extracted from the gel, implying cleaner data images. In the histograms present in Figure 26, we can observe that the whole cell lysate presents the multi-peaks behavior already visualized in the case of the BSA condition J, and a LO of 71.6%, slightly lower than the BSA, but in the expected value considered the proteome complexity. One of the most interesting parts is the 16 kDa band histogram, that reveals a unique peak, signifying that the proteins are mostly labeled by one and a unique fluorescent dye. This has been already discussed in a previous part, but this implies that only one protein can bind to one avidin. The fact that this protein is the smallest of the tested one is incredibly convenient by signifying that it is probably the case for the other bigger protein. The *LO* measured for this band is lower compared to the whole cell lysate, at 49.6%. Another interesting effect is the tail signal that we observe in the three others selected bands, especially heavy in the 23 kDa. When multiple dyes bind to the same protein, it can present a partial quenching caused by a local dye saturation or too close proximity with tryptophan residues (Vaiana et al. 2003), interfering with the fluorescence intensity and then making the peaks more diffuse, causing the observed tail signal. The *LO* measure for these tails fractions is high, going up to 91.2% for the 120 kDa fraction.

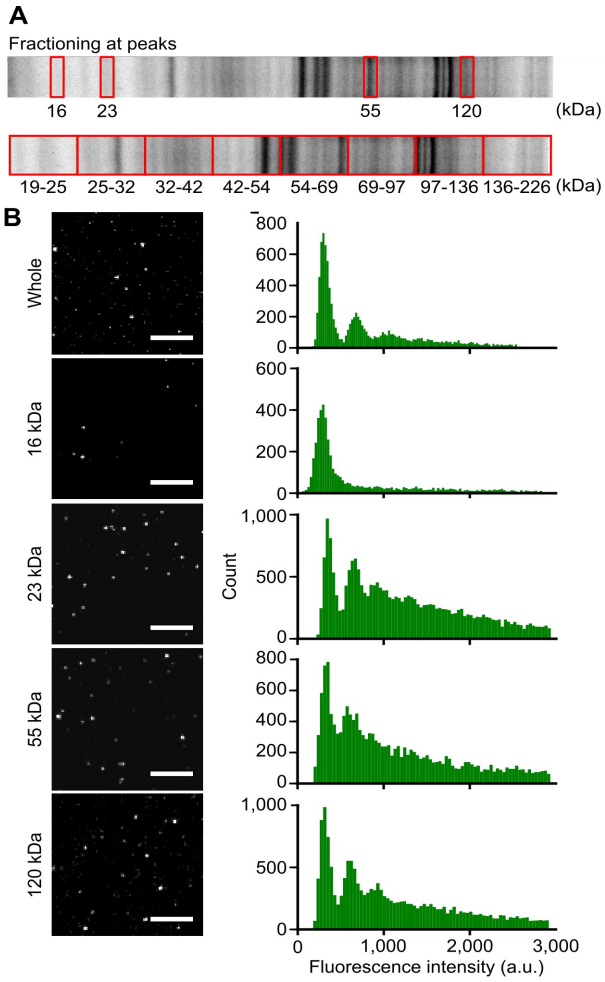


Figure 26: Raw data of labeling the whole proteome. (A) Proteome profile of the cell lysate on SDS-PAGE gel, with the red rectangle indicating the excised region that sustain the electroelution for LO analysis. (B) Fluorescence spot images (left) and histograms of spot intensities (right) at different molecular weight fractions. The scale bar is 10 µm.

As described previously, the number of dye by protein can be calculated from the histogram and checked if the labeling follows a Poisson law (Figure 27). As said previously, since the 16 kDa fraction presents only 50% of labeling efficiency, the majority of the labeled protein presents 1 dye, and then the process is more homogeneous than a Poisson law. On the other hand, long tail fraction presents a big variability in the number of dye by protein, with the unlabeled fraction being very small, and possibly signifying a labeling phenomenon that does not follow a Poisson law, especially for the 55 kDa fraction. In the case of the whole lysate analysis, we can observe that the 1 dye portion represents the biggest proportion of the signal, with a sudden drop followed by a tail. This further confirms that the labeling process is not a simple stochastic process following a Poisson law, but a more complex process.

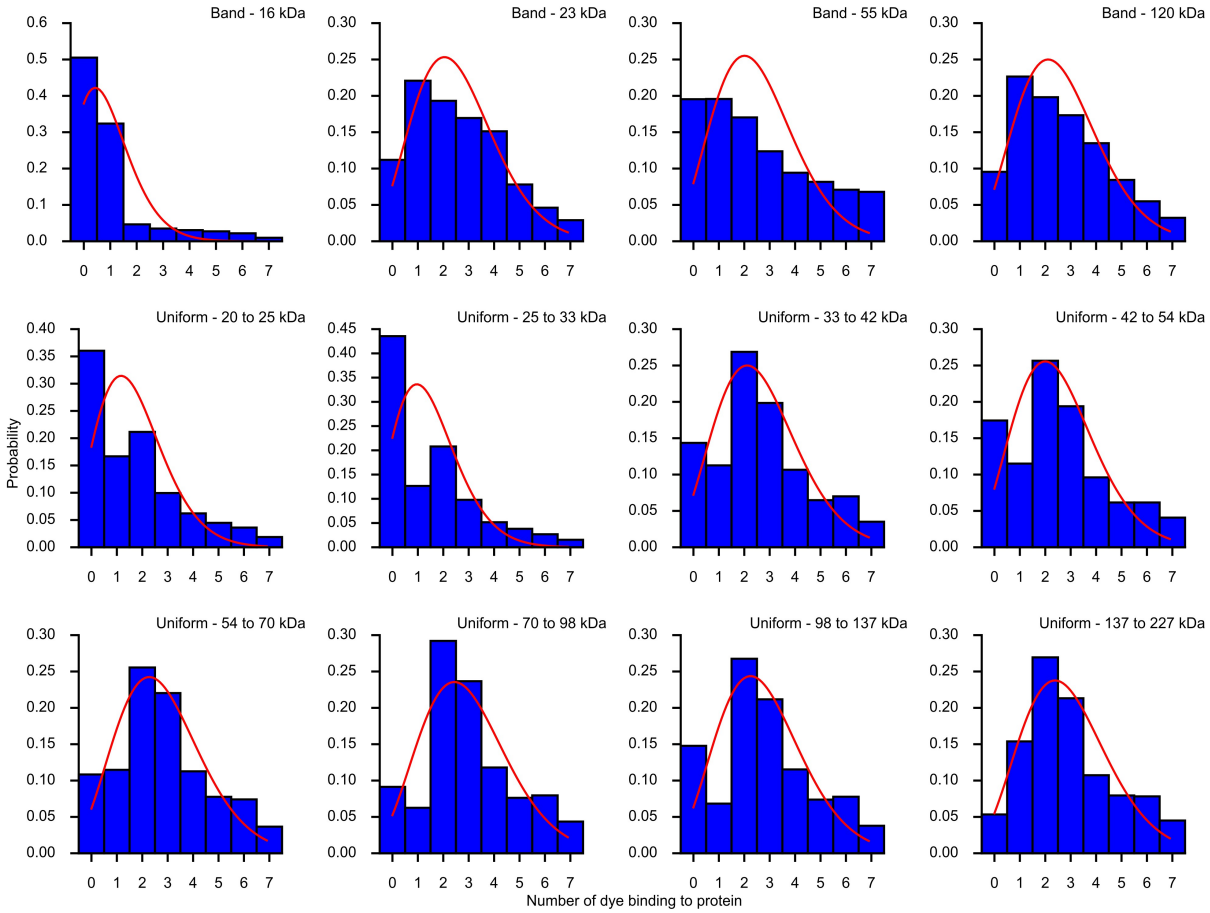


Figure 27: Probability density functions of numbers of the dye binding to proteins of the labeled proteome sample at different molecular weight fractions. The probability density function (blue) and the Poisson distribution curve with a mean of the function (red) are shown.

When looking at the number of dye by fraction, we observe that this number increase with the increase of the protein molecular weight up to around 3 dyes by protein (Figure 28-A). One interesting behavior is the fraction at 55 kDa, that present the highest number of dye by protein, but surprisingly a small diminution in the *LO*, at 82%, compared to its neighbors. This increase in the number of dye can be explained by an increase in the number of lysine in the fraction, which we name n_{lys}. The lysine number of each protein has been calculated from the protein composition in the interval of 2% around the molecular weight of each fraction, then the n_{lys} was obtained as the weighted average based on the protein abundance. These values are highly dependent on the protein species that have the highest abundance in the fractions, which are V-type proton ATPase 16 kDa proteolipid subunit (ATP6V0C), histone H1.5 (HIST1H1B), pre-mRNA-processing factor 19 (PRPF19) and desmoglein-2 (DSG2), respectively. Typically, the high n_{lys} value at 23 kDa is due to a large number of lysine residues of protein histone H1.5 (66 in total), but we have no confirmation that we correctly labeled these proteins. In case of the uniform and continuous gel cutting,

the size interval was used to calculate n_{lys} similarly to the peak fractionating. When we plotted *LO* values as a function of n_{lys} and fitted them with an exponential curve, we obtained the relationship: $LO = 100 - 71.5\exp(-0.039 n_{lys})$ (Figure 28-B). The relationship describes that the *LO* is proportional to n_{lys} at a small n_{lys}, but reaches a plateau at 100% at a large n_{lys}, which supports the fact that the succinimidyl ester labeling reaction occurs at lysine residues in a protein. In addition, the equation indicates that there is a positive offset of the *LO* at 28.5% regardless of the n_{lys} value, which is consistent with the fact that the succinimidyl ester reaction also occurs at the amino-terminal of each protein when accessible. We also analyzed the relationship between $\overline{n}_{dye}$ and n_{lys}, which is given by the relation: $\overline{n}_{dye} = 3.3 - 2.6\exp(-0.08 n_{lys})$ (Figure 28-C). The relationship describes that $\overline{n}_{dye}$ is proportional to n_{lys} at small n_{lys}, but reaches to a plateau at 3.3 at large n_{lys}, which would be caused by exclusion effects of dyes binding to the same protein, or quenching effects between neighbor dyes on the same protein. The $\overline{n}_{dye}$ also has a positive offset around 0.75, supporting the reaction at the NH_2 terminal end of proteins.

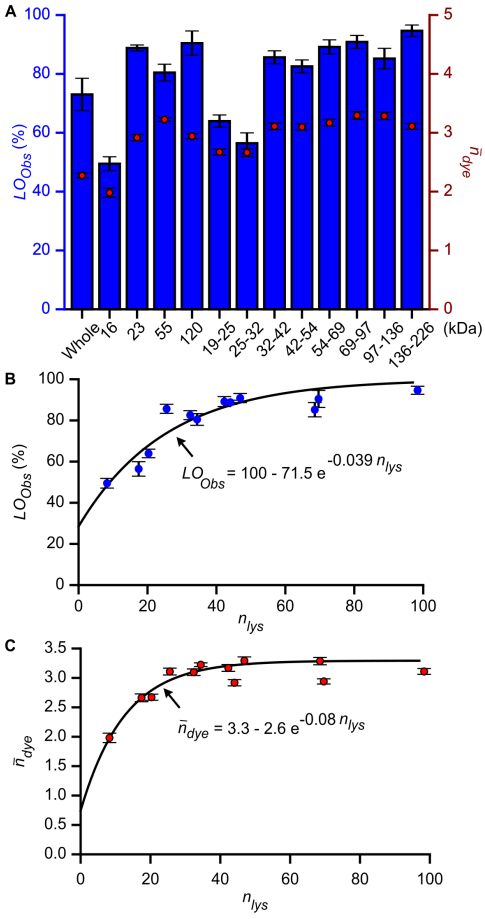


*Figure 28: Labeling homogeneity of proteome sample. (A) LO (blue) and $\overline{n}_{dye}$ (dark red) at different molecular weight fractions. (B) LO as a function of estimated number of lysine (n_{lys}). The line is a fitting curve to the equation: LO = 100 - A*exp(-B*n_{lys}), where A and B are the fitting parameters. (C) $\overline{n}_{dye}$ as a function of n_{lys}. The line is a fitting curve to the equation: $\overline{n}_{dye}$ = A - B*exp(-C*n_{lys}), where A, B and C are the fitting parameters.*

Possible improvements

Determining the labeling homogeneity of the whole proteome is obviously impossible. Like said previously, they are proteins species that will be unlabeled, or worst, not present at all due to their precipitation. Further improving the protein solubility is difficult, since changing the pH or the temperature can improve the solubility of some of them, but can also reduce it for other protein (Pelegrine and Gasparetto 2005). In addition, because of the nature of the proteome, it will not be possible to know, based on this kind of experiment, the solubility effect. One experiment to realize will be the analysis by mass spectrometry of the labeled whole cell lysate. This experiment will be expensive in term of time since it will necessitate to purify and concentrate the proteins to identify a significant portion of the proteome.

Another discussion point is about the coverslip passivisation. In our case, we used a simple BSA passivization, but for single molecule FRET experiment, they tend to realize a passivization using PEG layer (Chandradoss et al. 2014). This passivization is well adapted to avoid unspecific protein binding to the glass. In addition, controlling the ratio of PEG-biotin over PEG permits more precision on a day to day variation than the avidin glass coating that we realized. However, they need to use three steps, or sandwich, to attach their protein of interest on the glass, under the form PEG-biotin+avidin+biotin-protein. In this case, our protocol is simpler, with a single step avidin+biotin-protein, and also safer. Indeed, they need to use pirana acid, a strong oxidative acid, while we use simpler air plasma treatment to remove organic material from the glass surface. Overall, our technique is simpler and safer while giving good count data, with the drawback to increasing some unspecific reaction between some protein or cell fragment, and it can be interesting to compare the counting data with this other passivization protocol.

One uncertainty about our technique is the stoichiometry of the avidin/biotin reaction. Like stated previously, one avidin can bind up to for biotins, and we observed this with the positive control composed of fluorescent biotin, that give up to 4 peaks (Figure 13). Even if we have shown that proteins with only one fluorescent make only one peak in the intensity histogram (Figure 26-B), this experiment was not designed for this measure, and then we cannot be sure that they are only one protein binding event by avidin. One way to test this out is to make possible only one protein binding site by avidin. To realize this, one quick solution is to make a competition between unlabeled biotin and labeled protein, in a ratio big enough to be sure that they are one or less protein by avidin. Badly, the freedom of the unlabeled biotin is a lot higher compared to the protein, leading to uncertainty in the measure. Another way is to use a monovalent version of the avidin(Howarth et al. 2006). In this case, the test is a lot simpler and permit to be sure that they are one protein, labeled or

not, fixed to the avidin (or streptavidin). Badly, this variant of the avidin is not yet available commercially, and necessitate a complex procedure to produce and purify.

Conclusion

We successfully measured the labeling homogeneity at the single molecule level of the whole proteome. The results show that a high, 50-90% of *LO* can be achieved by high pH and solubility conditions, which importantly indicates that the application of single molecule counting to proteome analysis is realistic. Our method can provide attenuation factors to fill the gap between the actual protein numbers and the labeled, countable protein numbers, in the form of *LO*.

Our results also provide an important indication that labeling homogeneity is significantly changed by labeling conditions, beyond the simple Poisson process. The data suggest mechanisms that make the labeling more heterogeneous and homogeneous (Figure 29). Heterogeneity is considered to be caused by all-or-none mechanisms resulting in the existence of non-reactive proteins due to incomplete solubilization or less affinity. In contrast, the homogeneity is explained by the fulfillment of reactive lysine residues in a protein with dyes. These mechanisms highlight that single molecule phenomena can significantly deviate from simple theoretical views supposed by ensemble-averaged bulk experiments, and suggests the necessity of single-molecule experiments.

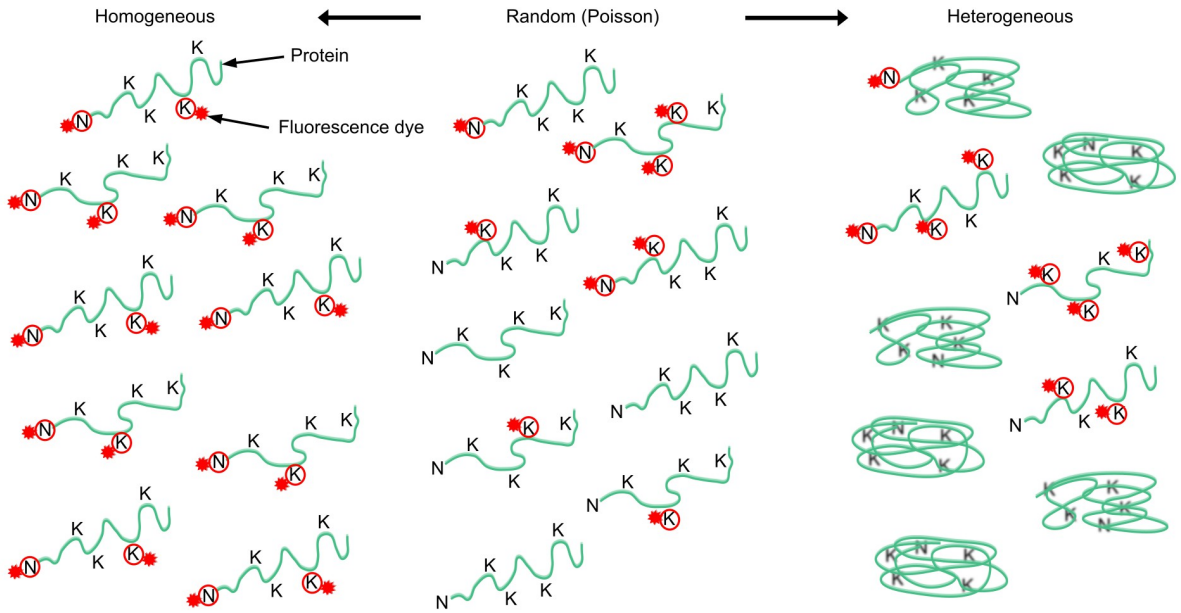


Figure 29: Mechanism of labeling homogeneity and heterogeneity. The letter, 'K' and 'N' denotes a lysine residue and NH2 terminal, respectively.

Part II : Proteome profiling at single molecule level

The proteome profiling, also called protein expression profile, is the information about all protein present in a biologic sample at a certain time. This profile can be used in a purpose of sample identification or to understand a particular cellular or molecular mechanism. These profiles can be easily realized by any protein separation technique, like SDS-PAGE or isoelectric-focusing, with the most famous technique being the 2D-DIGE, that allow the separation of two proteomes on two dimensions for relative comparison (Bouvet et al. 2014; Rodrigues et al. 2015). 2D-DIGE is still the most routinely performed despite other attractive alternatives thanks to its low cost and inexpensive infrastructure needed for acquisition and analysis, that resume the purchase of fluorescent dye and a fluorescent gel scanner. One alternative is the mass spectrometry (MS), that permits to identify peptides by their mass/charge profile. The increase of protein and peptide identification by the MS brings an association between the 2D-DIGE and the MS, with the interesting protein spots detected by 2D-DIGE then identified by MS (Bouvet et al. 2014). Another method to obtain a proteome profile is by the combination of liquid chromatography (LC) and MS, with the analysis of human serum able to identify cancer (Petricoin et al. 2002). Since the beginning of 2000, the proteome profiling have been massively realized by MS or LC-MS, and combine with the advance in the genome sequencing project, allows the identification of up to 8000 different proteins, even if they are still some portion of the proteome, roughly 45%, that is unidentified (Sidoli, Kulej, and Garcia 2017).

However, when the question asked is more about the comprehension of a particular process, like the identification of the interaction protein/ligand, a shotgun analysis is not adapted since it is containing too much information. To reduce the analyzed scope, simplifying the observed proteome profile is necessary, and different techniques permit to obtain this kind of results, all of them implying the stability/solubility of proteins (Franken et al. 2015; Savitski et al. 2014; Lomenick et al. 2009; Strickland et al. 2013). In addition to this simplification, an increase of the sensitivity of detection, up to the single cell level, is actively research (Nie et al. 2016; Rose et al. 2012), with the final goal being the yocto-molar sensitivity, more exactly the 1.66 yocto-molar, or the number of moles equal to the inverse of Avogadro’s number, implying a single molecule detection. Since we extensively discuss this single molecule detection level, or ultimate detection by fluorescence microscopy in the first part, let discuss here what it means to obtain a proteome profile at a single molecule level.

The first obvious point the single molecule add is the sensitivity, that will allow the detection of single protein in the profile and permits to achieve an absolute quantification of the proteome. However, to reach this absolute goal, the protein needs to be resolved on a perfect profile, where each protein species are separated from each other, without any

overlapping. Like in the case of the labeling efficiency, because of the protein biophysical and biochemical properties, they are actually no technique that permits to completely separate protein from another protein at this resolution. However, even with an imperfect protein resolution, the proteome can still reveal a lot of information like new protein discovery, protein pathway knowledge as well as cell identification, especially used for the identification of cancer cell (Petricoin et al. 2002).

The SDS-PAGE is the most used protein separation technique for proteome profiling, realized on an everyday basis in bio-molecular laboratories using as separation factor the size/mass of the protein through an inert polymer, generally composed of acrylamide. The SDS-PAGE will separate the proteins based on their denatured form, with the electrical field permitting to separate the protein, using SDS to neutralize and change the native protein charge to be proportional to their size. Another historical separation method is the isoelectric focusing of the protein, where the protein migrates under an electric field based on their native charge, inside a pH gradient. Since the global charge of a protein is determined by the side function of the amino-acid, and these functions are sensitive to pH, a protein will stop to migrate once all its charges are neutralized. Both techniques are usually combined to make 2D gel electrophoresis, that allowed the identification of protein and mechanism, despite its lack of reproducibility between gels and its difficulty to compare different samples (Rodrigues et al. 2015). To bypass this, 2D-DIGE label two or three different samples using different fluorescent dyes, and migrate all samples on the same gel, allowing exactly similar migration conditions, and enable the direct comparison between sample. Proteome profile can also be obtained from liquid chromatography, with the diversity of column allowing to separate proteins permits a larger biochemical or biophysical combination, like size, hydrophobicity or affinity to the matrix. These kinds of profiles are mainly detected by labeling free method like UV detection, then the apparition of the electrospray allows the use of the MS or tandem MS to add a second dimension allowing the identification of the protein in addition to the proteome profile (Petricoin et al. 2002). The single molecule or even single cell profile detection still need to be done, and we tried two different techniques to obtain these kinds of profiles, first, the capillary electrophoresis (CE), and second the SDS-PAGE, both of them adapted to theoretically allow single molecule detection using high sensitivity microscopy.

Capillary electrophoresis for profiling

Capillary electrophoresis is a family of electrokinetic separation methods performed in sub-millimeter diameter capillaries and in microfluidic channels. This part will be focused on the

capillary zone electrophoresis, referred as CE here, and not on other methods of the same family. CE is mainly used for DNA sequencing through automated multi-capillary electrophoresis systems using fluorophore labeling with multispectral imaging, even if the next sequencer generation do not use them (Karger and Guttman 2009) and another application is the protein separation (Burgi and Smith 2001).

CE is an electrophoretic technique that separate proteins based upon their size and ionic properties determine by the amino-acid composition of the given protein. When applying an electrical current to the CE, the protein will move toward the electrode of opposite charge. A buffer flow is also generated inside the column when the electric field is applied from the cathode to the anode electrode that is called electroosmotic flow (EOF). Since the EOF of the buffer solution is greater than the electrophoretic mobility of the protein, the protein is carried along with the buffer solution toward the cathode, whatever is its charge. In the case of native protein, the migration order seen by the detector is determined by the protein global charge and size, with small multiple positively charged proteins migrate quickly and small multiply negatively charged proteins are retained strongly (Figure 30-A). However, when the proteins are in presence of SDS, the global charge of the protein become negative, with the amount of charge being proportional to the protein size, allowing a protein separation solely on the size (Figure 30-B). To reach a single molecule level, we need to observe the CE under a high sensitivity microscope and to realize this the microfluidic CE is the best suited (Huang et al. 2007).

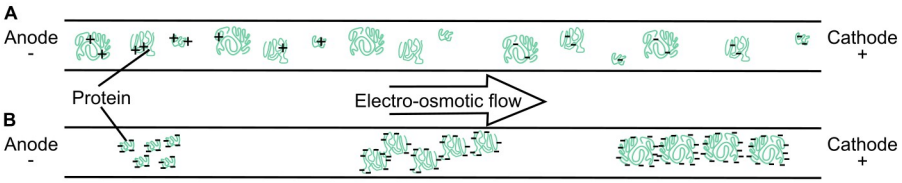


Figure 30: Protein separation in capillary electrophoresis. (A) Native proteins are separated depending of their global charge and size. (B) Charge uniformized protein (by SDS) are only separated by their size, with the most charged protein migrating quicker. + and - represent the global charge of the protein.

The material used to create the microfluidic is one of the most important factors for a successful protein separation. In our case, we used the cheap and very flexible elastomers polydimethylsiloxane (PDMS). The PDMS has the particularity to be a viscoelastic material, acting like honey, and allowing to replicate a pattern with high precision (Ren, Zhou, and Wu 2013). This pattern called the master is realized in the laboratory through soft lithography by the use of epoxy-based negative photoresist, in this case, SU-8 from microchem. Depending

on the viscosity of the SU-8 solution, a different thickness can be achieved by spinning this solution over a silicon disc. We used SU-8 3005 that permit to obtain a layer of 6 µm after spinning at 2000 rpm for the CE layer, while we used SU-8 3025 and a spinning speed of 2000 rpm to achieve a 40 µm thick layer for the valves layer. Then, after evaporation of the solvent at 90 degree Celsius, we initiate the polymerization of the SU-8 layer by UV exposure using a mask to create the master pattern. The polymerization of the SU-8 is then continued by baking at 90 degree Celsius, then developed in 1-methoxy-2-propanol acetate. The final result possesses a high stability to chemicals and radiation damage and is used as the master for PDMS casting. The PDMS is very well fitted for microscopy with its reflexion index close to water (1.38), high transparency and low auto-fluorescence while being adapted for biological samples with the capacity to transmit gas. PDMS have however a big drawback: it is hydrophobic, not the comportment excepted for a microfluidic material. Surface plasma oxidation permits to make the PDMS surface hydrophilic, creating a cross-link with glass, useful for making sealed chip, but this oxidation is provisory, and after a certain time, become hydrophobic again and then absorb hydrophobic compounds present in the solution, in our case, some protein species (Hillborg et al. 2000).

We first developed a simple design based on the double T, allowing the migration of 0.5 nl of a sample (Figure 31). This simple design also accommodates the presence of valves on the second layer of PDMS. Since the PDMS is an elastic polymer, its deformation can easily be realized using pressure, and adding a control layer incorporating valves allows an easy control of the flow. In this case, when the sample is injected, the anode and cathode valves are closed, forcing the sample to fill the loading area and exit to the waste. For the migration, sample and waste channel are closed and anode and cathode valves are open. Since we do not apply high pressure for the injection, no leakages are observed despite the simple design of the valves.

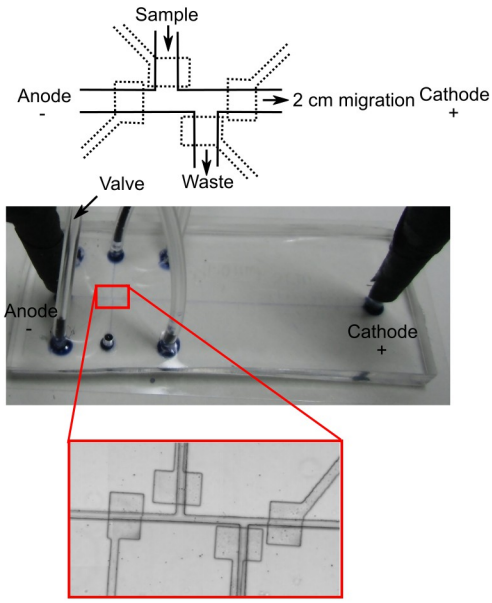


Figure 31: Double-T design of microfluidic for protein separation. The top is the schematical representation of the double T with the valves represented with dashed lines to control the flow. In the middle is the actual microchip, with visible dye to see the channel. At the extremity is the electrode with buffer reservoir. At the bottom is the photography reconstruction of the loading area with the valves system, without dye. The dimension of the channel is 50 µm wide for 6 µm height and 500 µm, for a total volume of approximately 0.5 nl.

The injection procedure was first realized using fluorescence dye, Cy3. Like shown in Figure 9, Cy3 possesses a unique positive charge at a nitrogen, making this dye naturally moving toward the anode. Since Cy3 is a small molecule, it will move quickly, but the EOF will still direct it again toward the cathode. Since EOF need time to initiate and is dependent on the voltage, we expect that the dye will move first toward the anode before being directed to the cathode. To confirm this, we acquire a time-lapse at the second T, the closer to the cathode, using the fluorescence microscope, an acquisition time of 100 ms and a different voltage to see the impact on the bandwidth (Figure 32-A). As expected, we observe that the dye seems to concentrate to the anode before the EOF bring it back to the cathode, sometimes with a fluorescence saturating the camera setting. When looking at the fluorescence profile intensity, we can observe a half peak width of 0.6 seconds at 0.1 kV, while at 3kV, this value is about 0.3 second, and appear sooner (Figure 32-B). If the migration at 3kV permit to reduce the bandwidth, it has the disadvantage to heat the capillary by the Joule effect. If for small migration time it does not create troubles, we expect troubles for longer migration time, especially since the protein solubility is depending on the temperature, motivating our choice to use the 0.1kV migration setting.

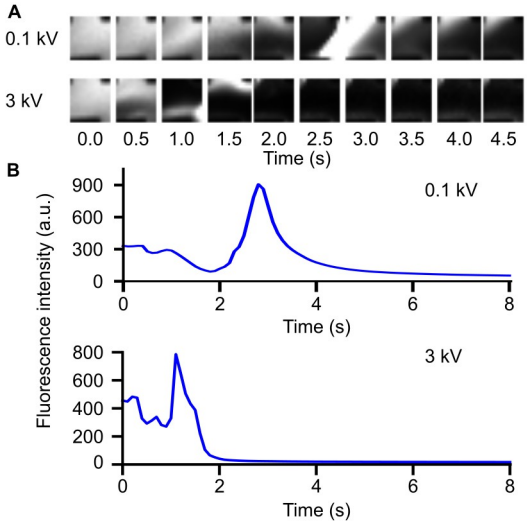


Figure 32: Dye injection process. (A). Timelapse of the double T section with dye migration. The dye, Cy3, is first directed toward the anode because of the global charge, then the EOF move the dye toward the cathode. It is possible to see a concentration effect. (B) Fluorescent profile intensity of the previous images. We can see a wider and later peaks at low voltage compare to high voltage.

The migration profile of pure labeled protein was measured using the double T microchip with a voltage of 0.5 kV, with the microchip and the buffer tank filled with a buffer composed of Tris-HCl pH 7.4 and SDS at 0.1%, allowing the protein separation solely on their size. The migration was realized using labeled protein inside the denaturation and labeling buffer, without any purification step, and we expect at least two peaks, the excess of non-reactive dye making a major peak, and the labeled protein making another peak. In Figure 33, we observe on the top the three different conditions, with the dye alone, the labeled lysozyme and the labeled BSA, with a single peak at 2.1 seconds, 3.5-3.8 seconds and 1.8-2.0 seconds of migration time respectively. These retention times were conserved between experiment, with the most precise being the dye. In the case of the lysozyme, we can observe a long tail, that can be the result of either a slight absorption to the PDMS surface or a leakage at the double T junction. It is interesting to note that the migration is very quick, with a complete migration in less than 5 seconds, even with the low molecular weight of the lysozyme at 14.3 kDa. However, the dye, even if presenting a sharp peak, is disrupting the analysis since it is the most important peak in intensity, covering other peaks. We decided to remove the excess of dye by realizing a size exclusion column on a mixture of labeled BSA and labeled lysozyme (ratio 1:1), then perform again the protein mixture migration. This result is presented in Figure 33, bottom, and display a two peaks pattern, with the first peak at 1.9 seconds and the second peak at 3 seconds. The first peak is very likely the BSA, with corresponding time compared to the previous experiment. The second peak can only be the lysozyme, however, the retention time is inferior to the previous experiment. This can signify that the lysozyme reacts with the BSA or the PDMS wall for example, in a different way that when it migrates alone. This suggests that the peak identification of a complex sample will

not be possible because of this uncontrollable reaction of some proteins. In addition, the peak resolution is weak in this system, with a too wide peak that does not allow the realization of a protein profile with enough resolution.

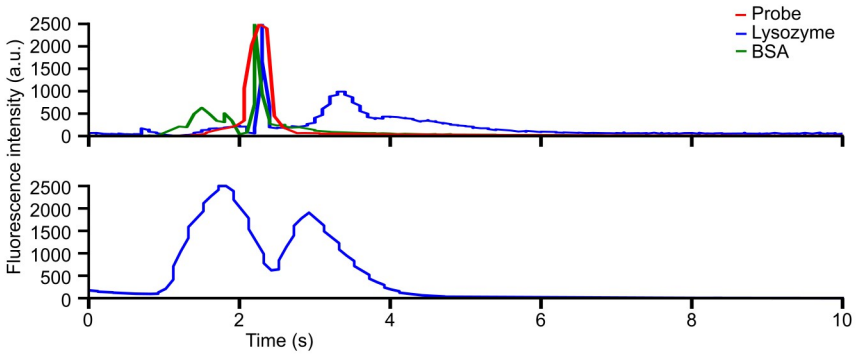


Figure 33: Migration profile of proteins in capillary electrophoresis. On the top is displayed the migration profile of pure probe, labeled lysozyme and labeled BSA. On the bottom is displayed the migration profile of the mixture labeled BSA and labeled lysozyme, without the dye excess.

To increase the resolution, we can used two techniques, either increase the migration voltage or increase the length of the capillary. In our case, we decide to increase the length of the capillary, to a length of up 35 cm. In addition, we also decide to incorporate a special observation area, that possess a thickness of 0.5 µm, to be sure to catch all the signal with the microscope. Water based buffer flow very easily inside the microchip by capillarity, but the sudden neck in the channel result in an increase pressure at the cathode. To resolve this trouble, we installed a derivation system, controlled by a valve, that permit to quickly change the buffer in the migration channel, even between two runs (Figure 34).

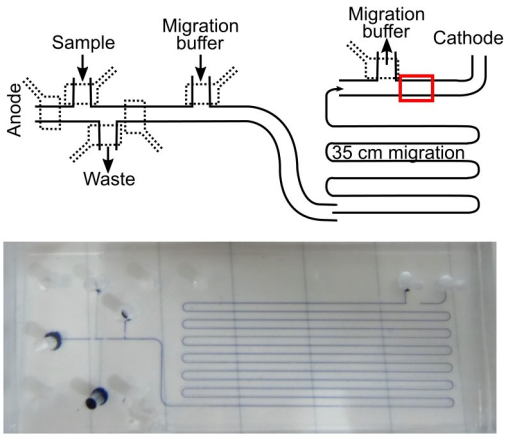


Figure 34: 35 cm capillary electrophoresis microchip. On top is the schematic microchip, with the valve layer in dots line. The red square represents the observation area that is 0.5 µm thick channel. At the bottom is the actual microchip in PDMS. The microscopic observation area is just after the migration buffer outlet, invisible even with the dye.

To realize the two different heights resulting in the observation area, the first layer was made of SU8-2000.5 spin coated at 3000 rpm, resulting in a layer of 0.5 µm, bake to evaporate the solvent, then UV expose with the full pattern. Then, instead of a hard bake, another SU8 3025 is spin coated at 2000 rpm to create the second layer of 6 µm. It is then baked to evaporate the solvent, then the second mask, where the observation area is lacking, is aligned using targets before UV exposition. The two layers are then baked for the final polymerization of the SU-8, resulting in a two layers master, from where PDMS can easily be cast. However, when running this microchip design, we were unable to obtain a migration over the 35 cm, whatever was the voltage applied, the running buffer at the extremity was boiling because of the Joule effect. We tried different design and buffer composition, unsuccessful.

In conclusion, we were able to separate two labeled proteins with success on a short microcapillary chip, with a strong limitation in the potential number of separated proteins, and the strategy to increase the length of the capillary was unsuccessful. Instead of increasing the length of the capillary, realizing a gel capillary was probably a better solution for the separation with a higher resolution of the protein. Another idea will be to use a 2-dimensional capillary microchip, where the protein is first separated by size, with a fraction of this size being transferred to conduct a micellar electrokinetic electrophoresis (MEKC), a second method that permits to separate protein based on their hydrophobicity (Shadpour and Soper 2006). Separating protein based on their hydrophobicity is difficult in PDMS chip, and is generally better conduct in Poly(methyl methacrylate) (PMMA), or necessitate specific treatment of the surface of the PDMS, all expertise that we do not have in the laboratory.

Since we were unable to obtain proper proteome profile using capillary electrophoresis, we decided to adopt another technique that is used for the proteome profile since the beginning of protein analysis: SDS-PAGE.

Profiling on thin SDS-PAGE

One of the oldest technique that permits to realize a proteomic profile is the SDS-PAGE, combined with a protein visualization method like the Coomassie blue or the silver nitrate staining (Neuhoff et al. 1988; Poehling and Neuhoff 1981). If these techniques permit the detection of proteins, their sensitivity or their specificity is limited, resulting in an incapacity to be used in single molecule studies. Like we have already shown in Part 1, labeling a big portion of the proteome is possible, and then it can be possible to detect a single molecule in the acrylamide gel after migration, increasing the sensitivity up to the point to realize the proteome profiling of a single cell.

Single cell western blot in capillary or array is already a reality, realized first in 2012 by Hughes and Herr (A. J. Hughes et al. 2014; A. J. Hughes and Herr 2012), where they used a UV activable acrylamide that permits to capture the protein after their migration, and therefore limit their diffusion during their detection by an antibody. If this technique permits to analyze thousands of single cell, the multiplex is limited to the detection of 11 different proteins, with the incapacity to detect discrete size change due to the migration length limitation. Since it is a western blot technique, it used an antibody for the detection of the protein of interest, with the antibody affinity to its target protein being a limitation for the detection of the single molecule. In addition, the UV polymerization of the gel in order to trap the protein will destroy the fluorescence of fluorescent fusion protein, rendering single molecule detection impossible.

The idea to use a thin gel for protein separation is not new, and displays advantage with its easy fabrication and quicker steps, especially for the drying compared to traditional size gel (Heukeshoven and Dernick 1992). The advantages to using a thin SDS-PAGE gel are the easy fixation of the gel to the microscope coverslip, a small volume of gel to observe under the microscope and a quick gel drying that will limit the protein diffusion. To realize this, we first need to bind the polyacrylamide gel to the glass surface. They are a lot of methods to do this, and we decide to use the couple Bind/Repel Silane glass treatment (Tegelström and Wyöni 1986).

Since the Bind/Repel Silane is a treatment for glass surface, the master need to be realized on a 1 mm thick glass plasma treated glass, with the plasma treatment allowing the SU-8 to adhere to the glass surface for a thickness of 204 µm. The coverslip glass treatment by bind silane, also known as 3-methacryloxypropyltrimethoxysilane, permits to covalently bind a polyacrylamide gel to the glass surface. This is realized by the silanization of the glass by the bind silane, that also possesses a propen-2-ol function at the other extremity, permitting

the reaction of the acrylamide. The support glass, that displays the mold for the acrylamide gel, need to be treated with repel silane, also known as dimethyldichlorosilane, that renders the glass surface hydrophobic and avoids the polyacrylamide binding.

Acrylamide polymerization of a thin gel needs to be perfectly done, mainly because a polymerization default has an impact much more important compared to classic gel. Since oxygen inhibits the acrylamide polymerization, bubbles presence between the two glasses needs to be avoided, with a small bubble being able to inhibit the polymerization of few millimeters of gel around it. In addition, the gel homogeneity is also a critical point and can be realized with a slow polymerization, that sadly has the effect to increase the oxygen diffusion in the gel. To reduce this effect, we vacuum for 15 to 20 minutes the acrylamide solution to remove soluble oxygen (20% acrylamide and 7 mM Tris-HCl pH 7.4) and add 0.1% of riboflavin, known to react with oxygen and then slow down the oxygen inhibition (Massey 2000).

The gel design was the most tedious part. We need a gel big enough to obtain a nice proteome profile, but not too big to avoid too long acquisition time with the microscope, and with a standard coverslip size that fit under the microscope. After a lot of trial and error, we finally decide to use a very simple master that fit a 50 by 75 mm glass (Figure 35). A 20% polyacrylamide gel was realized with this master, polymerized for 1h30 at room temperature with a weight of 5 kg, then store in a humid box at 4 degree Celsius before utilization. The day of use, the two glasses was delicately separated, and the coverslip with the gel was extensively washed in running buffer (20 mM Tris HCl, 0.2 M glycine and 0.1% SDS) to remove unreacted components that can interfere with the migration. After putting 1 µl of sample in each well, migration occurs first at 50 V for 15 minutes, then 500 V for 3 hours, on a horizontal plate cool down at 20 degree Celsius to avoid heating by the Joule effect. After migration, the gel was quickly washed with distilled water to remove salt, the excess of water was carefully removed using an absorption paper then the gel was dried for 30 min at 60 degree Celsius on a hot plate. The gel was pre-visualised using a gel viewer with a Cy3 illumination setting to determine if the migration quality was enough to justify the microscopic scanning.

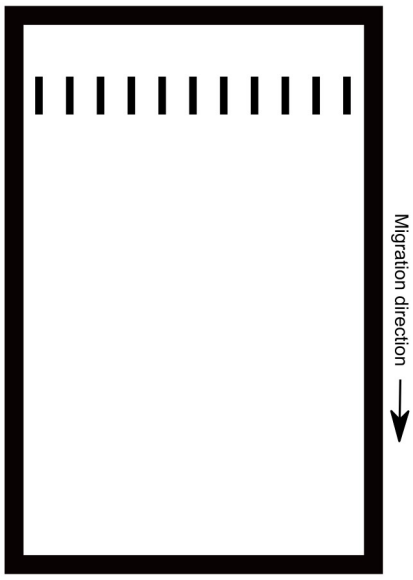


Figure 35: Design for SDS-PAGE. The dark area represent the polymerized area of SU-8. The total size is 50 x 75 mm (scale 1:1) with a thickness of 0.2 mm. The sample loading area are on the top with a size of 1 x 5 mm, allowing up to 1 µl of sample.

All preliminary experiments and tests were realized using a commercial fluorescent ladder from Benchmark, composed of 7 proteins at 155, 98, 63, 40, 32, 21 and 11 kDa. Once the migration conditions previously explained were fixed, we realized a time course of ladder migration to determine the optimal duration (Figure 36). Since we expect that the most important part of the proteome profiling is in the high molecular weight, and since the gel that we use for the protein separation is uniform, we want a long migration time to fully separate them, justifying the 3 hours of migration even if the band shape are not fully straight anymore. Another justification for this long migration is that the labeling efficiency of the proteome is higher when the molecular weight of the protein is bigger, implying that the fluorescence signal is likely to be stronger.

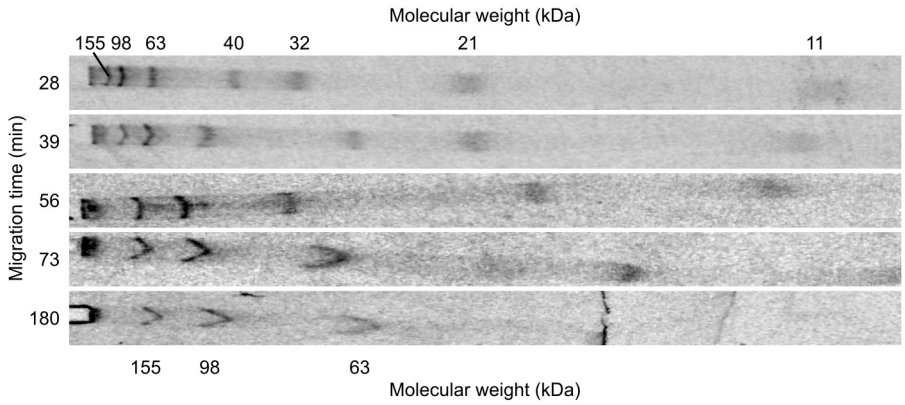


Figure 36: Time course of a migration. 1 µl of fluorescent ladder, composed of 7 different proteins, is migrated at 500 V and imaged at different time point using the gel viewer.

The gel observation at the microscope was a little tedious, mainly because we want to observe a big area through a small aperture. If the gel viewer permits to give us an idea about the gel and migration aspect, it does not have the sensitivity of the microscope, with a band detection of approximately 3 ng of protein (Figure 37-A). To globally observe the gel with the microscope, we need to scan the area of interest, and for this purpose, we develop a macro in Metamorph, where we manually select the center of each well to scan and the length to scan, then the program will scan around the well (2 mm in total width) and on the desired length. To keep the focus, the glass is mounted on an homemade support that presses the glass between two metallic plates, then screw to the stage, and the auto-focus is realized using an infrared laser every 10 frames. After the acquisition, each image was treated using ImageJ macro for the laser correction and background subtraction, then the gel image was reconstructed. Using this method, we are able to visualize a protein band, that can span through different images and then not be easily visible by looking at the individual image. The sensitivity of this method is then investigated by diluting the fluorescent protein ladder, with each protein present at a concentration of 0.1 µg/µl in the pure solution. The visualization of one band using the high sensitivity microscope does not enable single-molecule observation, mainly because of the fluorescence background of the gel surface, but still permits to achieve a sensitivity of 100 of picograms of protein (Figure 37-B). It is important to keep in mind that the protein ladder is a commercial ladder, with theoretically 100% of the protein being labeled by a fluorescent dye.

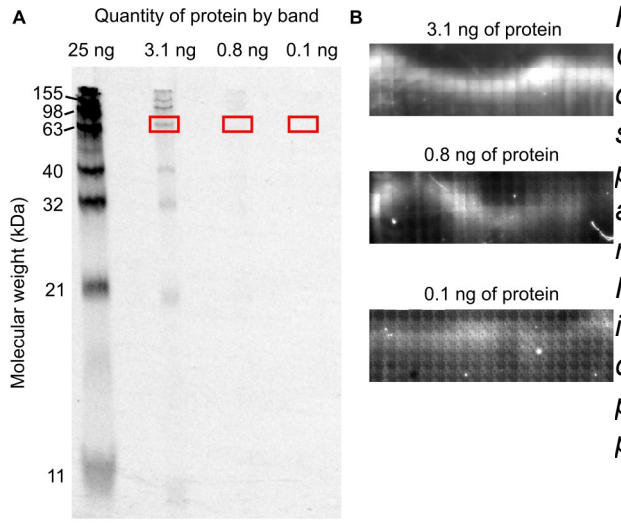


Figure 37: Limit of sensitivity of the gel. (A) Gel viewer limitation. Protein migration observed at the gel viewer with the Cy3 setting, permit to visualize up to 3 ng of protein by band, but fail to detect with accuracy 0.8 ng. The red box design the region scanned using the microscope. (B) Microscope observation of the gel. The image is a reconstruction composed of dozens of microscope image, and it is possible to detect up to 100 of picograms of protein.

Using the labeling protocol that we describe in Part 1, we labeled BSA and catalase protein, mixed them together, and then migrated 0.01 µg of each of them on the gel. BSA has a molecular weight of 66 kDa, while catalase is a heavy protein of 250 kDa, composed of 4

subunits with a molecular weight of 60 kDa each(Sund, Weber, and Molbert 1967). We expect the catalase to be partially or totally denatured, and then detecting two to three bands, one at 60 kDa, another at 66 kDa and an optional one at 250 kDa. Since the very high molecular weight of the catalase protein, there is a possibility that this protein is not entering the gel, or that the migration distance is too small for the protein to penetrate enough the gel for visualization. When scanning the gel with the microscope, we can observe the two bands around the middle of the gel. Independent sample on a different gel, with the similar migration parameters (2h45 versus 2h55) and 2 days apart give the same peaks and general profile (Figure 38). The thickness of the line for the day 2 is not caused by the profile, but by a bad tiling effect or illumination correction, causing the fluorescence intensity at the edge of one image being smaller than the edge of the next image.

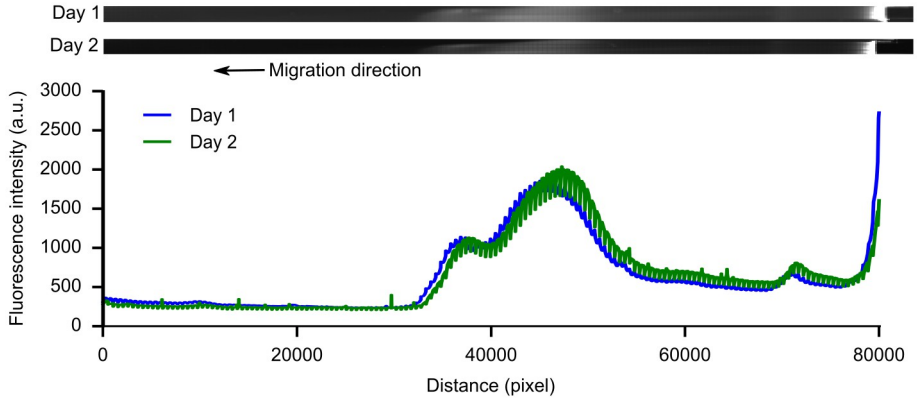


Figure 38: Migration profile of BSA and catalase. With similar labeling and migration parameters, BSA and catalase give the same profile.

The proteome profiling possesses a large variety of information, however, it will be extremely uncertain to try a protein identification solely based on the molecular weight approximation. Then the proteome profile from one cell line does not have a lot of interest, we decide to compare different cell lines from the breast cancer. Since we can detect as little as 100 pg of proteins by bands, and that a mammalian cell has around 50 to 75 picograms of proteins, we decide to analyze 100 cells, for an estimated total amount of protein of 7.5 ng. We expect around 20 bands in the profile for this number of cell, expecting that it will be enough for a cell line identification. The cell lines chosen was ordered from ATCC, and three strains were from epithelial tissue (HTB 132, MDA-MB 231 and MCF 7) while one strain was from melanoma tissue (HTB 129). These cells are interesting because they can realize the epithelial to mesenchymal transition, that is a phenomenon that occurs in wound healing, organ fibrosis and the initiation of metastasis for cancer progression (Kalluri and Weinberg 2009; Lamouille, Xu, and Derynck 2014). Being able to differentiate the proteome profile of

the two different tissues can permit to identify new biomarkers, better understand the transition or lead to a quicker diagnostic.

Each cell line was cultivated in Eagle's Minimum Essential Medium (EMEM) supplemented by 0.1% of insulin and 10% of fetal bovine serum, at 37 degree Celsius in a CO_2 incubator. The cell culture was passed by the trypsin method, wash in PBS 1x, pH 7.4, then re-suspend at a concentration of $5x10^5$ cells by ml. Taking 0.2 µl of this solution permits to sample 100 cells, that are put in a PCR tube. Immediately after, the cells are lysed by adding 1 µl of lysing buffer, incubated 5 min at RT, then the proteins are labeled by adding 0.2 µl of 2 µg/ml Cy3 NHS-ester dye for 15 min at RT. This protein sample can then be stored for several weeks at -80 degree Celsius. Before use, proteins are unfrozen on ice, then 1 µl of the labeled proteome was put on a freshly washed SDS-PAGE gel and migrated at 500V for 3 hours. After migration, the gel was washed and pre-visualized with the gel viewer. The quality control needs to be strict and was composed of 3 points. The first point is the global quality of the gel and glass, looking at breakpoints, cuts, scratch, dust that can interfere with the microscope acquisition. The second point looks at the global migration quality, with the fluorescent ladder migration that needs to be well separated and with a similar aspect of the gel. The third point is the sample migration quality, with all sample presenting a straight migration profile being chosen. They were less than 70% of gel that passes the first point, diminishing to 40% for the second point. The last point is the most severe, with only 1 gel on 4 that pass, resulting in total 1 gel on 10 that success to pass the quality check, but it was necessary to assure replicative results.

The acquisition is realized by scanning with the microscope using a 30x silicone oil immersion lens around the well on a length of 4.4 cm, corresponding to 160 images of 274 x 274 µm, with the width varying from 5 to 10 images, depending on how straight was the migration, allowing to visualize the protein with a molecular weight comprised between 200 to 50 kDa for a 3 hours long migration. The image treatment was done using a homemade python program, that realized the tiling quicker than imageJ and without any resolution loss. In addition, for each column of pixel (corresponding to the width), outlier pixels were removed using the 1.5 x interquartile range, where all pixel lower the value of interquartile 1 x 1.5 or higher the value of interquartile 3 x 1.5 were designed as outliers and replace by the closest not outlier pixel value. This operation normally permits to remove the stronger dust or hot pixel. After this step, each column was averaged and the proteome profile was then saved as a text file for further treatment. In our case, we suspect that the baseline is composed of proteins that we were unable to resolve, and if this signal probably possesses information, it also makes the visualization of the data more delicate. We then decide to

remove the global baseline using the exact same function already used and described in Part 1. Even after this treatment, we still observe some speaks, mainly representing dust, and the image to image fluorescence intensity variation. To remove then, we then applied an Otsu filtering to remove the speaks followed by a Savitzky-Golay smoothing. This permits to generate clean and make the proteome profile looking continuous and easier to compare to each other (Figure 39).

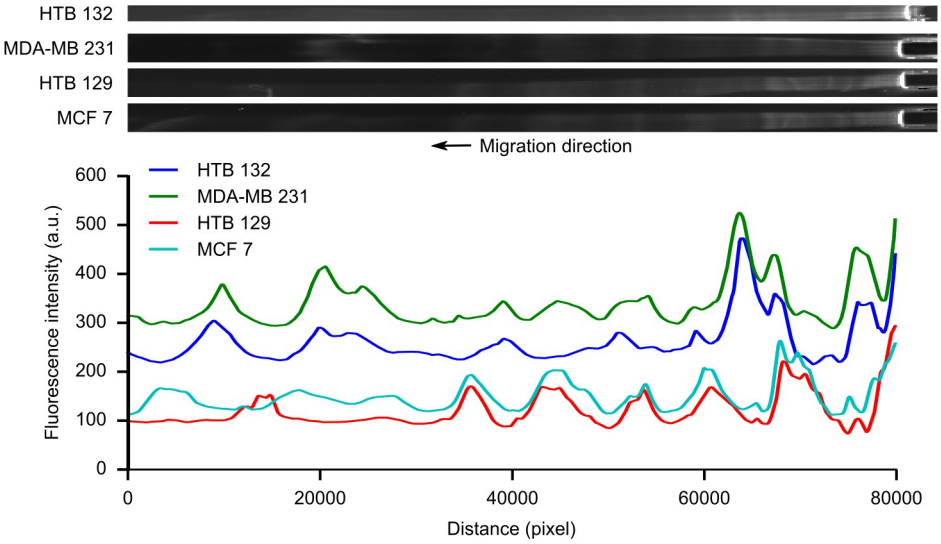


Figure 39: Proteome profile of cell line. The proteome profile was obtained after the migration of 100 cells and microscope scanning. Following the profile extraction, profile processing permits to remove speaks and image to image fluorescence intensity variation. Intensity value level were modified to group together similar profile.

Very similar profile are grouped together, with the HTB 132 and MDA-MB 231 presenting a lot of similar peaks, especially at the beginning and the end of the spectrum, but with some variations in the middle. On the other hand, HTB 129 and MCF 7 display a similar profile at the high molecular weight (right part), with more important variations at the low one. Globally, each proteome profile is different from each other, with around 10 to 15 peaks, but at different position and intensity. To measure more precisely the difference between two profiles, we realized a Pearson correlation matrix (Figure 40). When comparing the whole profile, we show that HTB 132 and MDA-MB 231 are very similar, and the same tendency can be observed in the first tier and last tier of the profile. In contrast, HTB 129 and MCF 7 are similar in the second and last tier of the profile, but show no correlation in the first tier, confirming the visual observation of the global profile. Badly, if this Pearson correlation matrix permits to put a number on the difference or similitude between profile, it does not

clearly separate the epithelial to the melanoma cell lines, even if HTB 129 (melanoma) seems to be apart compared to the others.

Whole profile (0 to 80000 pixels)

	HTB 132	MDA-MB 231	HTB 129	MCF 7
HTB 132	1.0	0.99	0.92	0.9
MDA-MB 231		1.0	0.93	0.92
HTB 129			1.0	0.96

0 to 30000 pixels

	HTB 132	MDA-MB 231	HTB 129	MCF 7
HTB 132	1.0	0.9	-0.02	-0.4
MDA-MB 231		1.0	-0.16	-0.19
HTB 129			1.0	-0.2

30000 to 55000 pixels

	HTB 132	MDA-MB 231	HTB 129	MCF 7
HTB 132	1.0	0.91	0.6	0.64
MDA-MB 231		1.0	0.82	0.86
HTB 129			1.0	0.97

55000 to 80000 pixels

	HTB 132	MDA-MB 231	HTB 129	MCF 7
HTB 132	1.0	0.97	0.81	0.77
MDA-MB 231		1.0	0.83	0.78
HTB 129			1.0	0.92

Figure 40: Pearson correlation matrix of proteome profile. Pearson correlation scores were calculated for the whole profile (left), or approximately the tier of the profile that seems to present similar or different parts.

In conclusion, we succeed to obtain the fluorescent proteome profile with 100 cells of 4 different cell types and see some difference between each profile. However, the difference is more important in the low molecular weight part of the profile, suggesting that the protein profile at an inferior molecular weight can display more drastic differences. The next step will be to reduce the migration time of 3 hours to 30 minutes to be able to see the migration profile of protein from approximately 100 to 20 kDa. This will have the advantage to lead to less variation between the gels and hopefully results in an easier data generation. At longer terms, it can be possible to increase the sensitivity by photo-bleaching the gel before the migration, in order to destroy the native auto-fluorescence or shift to low auto-fluorescence acrylamide gel. Biologically, comparing epithelial and melanoma tissue is not new, even if working with 100 cells will permit to reduce the sample amount for this identification. It will be a lot more interesting to studying the transition epithelial to mesenchymal, that can be realized using these cell lines and other that we have in the laboratory (KPL 4 and ZR 75-1), and looking at the proteome profile during the transition to identify potential biomarkers.

Discussion and Conclusion

We have previously shown that we are able to label the majority of the cellular proteome with a fluorescent dye and to observe it at a single molecule level using high sensitivity microscopy. We also succeed to measure the labeling homogeneity through the proteome and see a correlation between the homogeneity and the molecular weight of the labeled protein. The idea was then to obtain a proteome profile at a single molecule sensitivity. Badly, the two methods that we try, capillary electrophoresis and SDS-PAGE did not allow the single molecule detection in the tested conditions. In case of the capillary electrophoresis, it is probably because of the molecule migration during the acquisition, that dilutes the signal in a bigger volume and renders it more difficult to detect. For the SDS-PAGE case, the use of a smaller NA objective lens (30x, silicon oil immersion, NA 1.05) can be one explanation to the inability to detect a single-molecule, while another explanation can be the gel autofluorescence, or the gel thickness, too important. Even though we were unable to visualize single-molecule, we still improve the sensitivity of the proteome profile measurement with the SDS-PAGE, by being able to measure and differentiate different cell lines based on 100 cells.

Like defined previously, the proteome is the ensemble of protein expressed at a given time by a cell. From UniProt, the human proteome is composed of 92179 proteins, counting for the splicing variant, but not the nearly unlimited number of post-transcriptional modifications that proteins can undergo. It is important to precise that all proteins are not express in a given cell, only a portion of it, and is estimated to be around 15000 proteins. However, a portion of these proteins are invisible to classical method, mainly because some of them are transmembrane proteins that tend to precipitate when the cell is lysate, or that is already in a 3D conformation that does not allow its analysis. These proteins can account for up to 54% of eukaryote or virus proteome and are referred as the “dark proteome” (Perdigão et al. 2015). The protocol developed in part 1 for protein labeling obviously do not take into account this portion of the proteome. In our case, the labeled portion of the proteome is suspected to label a majority of accessible protein. It will be very interesting to variate the cell lysis condition as well as the labeling conditions to observe the evolution in the labeling efficiency and homogeneity. For example, we can imagine a condition that will extract more efficiently acidic or basic protein by adjusting the pH of the solution. It is also possible to sub-

fractionate the cell to label only some cellular compartment like the nucleus or the Golgi for example (Drissi, Dubois, and Boisvert 2013).

Checking which part of the proteome is labeled also lead to the question of how the labeling really occurs. In the current setting, the cyanine dye is linked to an NHS function, and we have shown that this function links the dye to lysine residue as well as the NH2 terminal function of the protein. We can, however, argue about what happened about the other residue, more precisely about the arginine or histidine that are both parts of the same type of amino-acid than the lysine, and both possess amine function. This NHS function is also known to react less specifically with other amino-residue like the OH function of the tyrosine (W.-C. Yang et al. 2006), leading to some unspecificity. It is then interesting to realize a similar study to the one present in Figure 24, where the LO was compared to the average number of lysine present in each proteome fraction. It is possible to count each and every amino-acid occurrence in each protein (Figure 41), and we can observe in this case that most of the amino-acid residues do not present a fitting with the LO. The lysine (K) stay the best match, confirming that the reaction mainly occurs on the side-chain amine, but the Proline (P) and the Alanine (A) also show a good fitting, even if both of them are not known to react with the NHS-ester. In addition, we also test the fitting with the total number of amino-acid in the protein, or so the size, and do not see a good fitting. This information support that the labeling reaction occurs mainly between the dye and the lysine residues, even if more work is necessary to rule out the possibility of nonspecific reaction of the dye with other amino-acid.

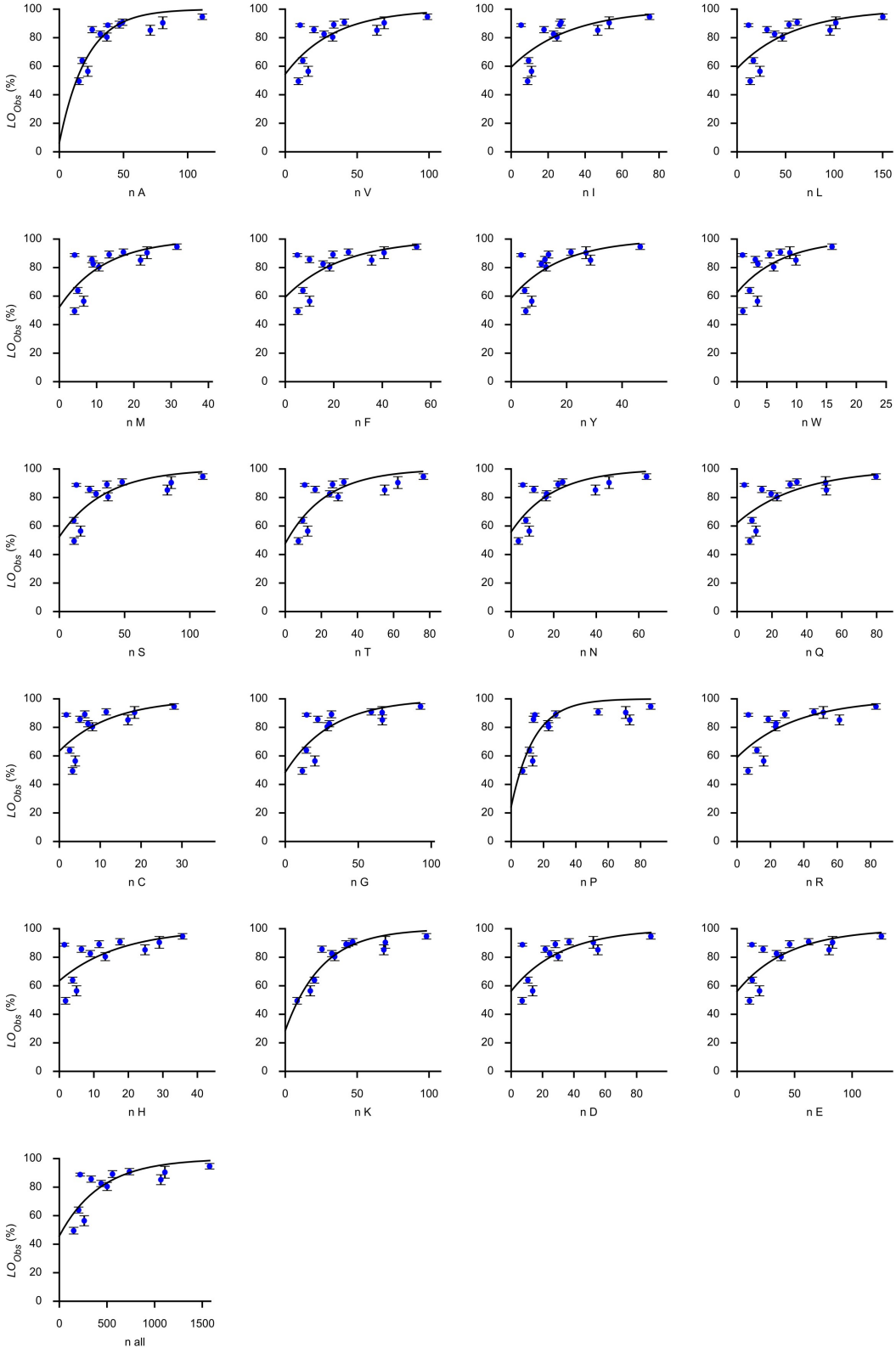


Figure 41: LO fitting with all different amino-acid residues. The LO values used are the same as in the Figure 24, with the amino-acid counting identify here by there single letter code. In addition, the total protein size is also plot with the denomination nall, for counting all amino-acid.

Going further a little more, we can also argue about the biotinylation process used for protein immobilization, that is critical for the *LO* determination. The biotinylation protocol used here is a very classical protocol that is daily used in the laboratory (Elia 2010). However, we didn't quantify this biotinylation efficiency, mainly because we suspect that labeled proteins are fully accessible for biotinylation. This can also mean that the *LO* is being overestimated because unlabeled protein is less likely to be biotinylated. On the other hand, we also suspect that over-labeled protein is more likely to be unstable due to the change in their hydrophobicity caused by the dye, leading to an underestimation of the *LO*. One interesting solution will be to double purify the labeled protein. The first purification will be to remove unreacted reagent, like the unreacted biotin and dye by using a size exclusion column for example. The second purification will be an affinity chromatography using immobilized avidin or streptavidin column (Bayer and Wilchek 1990). By comparing the CE value before and after the second column, it may be possible to estimate which proportion of protein is lost. However, it will not inform if this loss is homogeneous through all type of protein, or favorize some specific protein type.

Another point to take in consideration is the immobilization of the protein on the coverslip. Since the labeled protein is biotinylated, and the coverslip is cover by avidin protein, we expect that the reaction biotin-avidin to be strong enough to tightly bind the labeled protein to the glass surface during the observation time. This force is one of the strongest non-covalent force but can be moved out during the coverslip washing. One experiment that can confirm the washing force will be to compare the avidin density before washing using the exact number of fluorescent biotin. Another way is to increase the force between the biotin and the anchor by using the streptavidin, that requires a stronger force before separation (Lo et al. 1999).

Finally, another to point that can be changed is the way to check the avidin density. In our current protocol, a separate coverslip treated exactly like the sample coverslip, is used with fluorescent biotin to determine the maximum density, and then the positive control. This control main flaw is that it is assumed that the treatment is exactly similar. The replicability experiment suggests this is the case, but a more sophisticated control can think about, like

for example to add after the sample reaction and coverslip washing, a fluorescently labeled biotin that is spectrally different from the protein labeling. By imaging the two channels at the same time, it will be possible to estimate the number of unoccupied sites and then correct the *LO* from underestimation. It will also provide a control of saturation of the sample compared to the number of active sites. One of the best control will be to have already labeled avidin or streptavidin (already commercially available), again with a different excitation/emission system to the labeled protein, and then to be able to determine the *LO* directly for each image, since one color will be the total number of sites, while the other color will be the labeled protein.

Quantitative proteomics is a growing field, with the main trouble is to identify and quantify at the same time the proteome of a biological sample. If mass spectroscopy is the main method to identify proteins, its quantification is generally relative, with the absolute quantification necessitating bias correction depending on the detected analyte (Nikolov, Schmidt, and Urlaub 2012). At the other hand, gel-based proteomic, through 2D-DIGE, is able to obtain a reliable relative quantification and with rigorous control, an absolute quantification (Von Bergen et al. 2011). In this case, however, the protein identification is generally lacking, and necessitate either immuno-identification by using antibodies or mass-spectroscopy based identification by extracting the protein of interest from the gel. However, accessing the protein identification is not always necessary, with the profile containing enough information for the objective, for example, cell identification or biomarker detection (Lawrie, Fothergill, and Murray 2001). With this idea in mind, increasing the sensitivity of the analysis can allow the same kind of identification with fewer cells, or better to see the cellular heterogeneity allowed by single-cell analysis, with keeping the quantification of the proteome rigorous. Actual techniques that permit to realize single cell analysis still focus more on the identification than quantification (A. J. Hughes et al. 2014; Tentori, Yamauchi, and Herr 2016), but the next step will be to add and improve an absolute quantification at their analysis. This quantification will either be by using fluorescently labeled antibodies, like actually or can concentrate on the whole proteome by using a nonspecific labeling method. However, to enable an absolute quantification, that necessitate reaching the single molecule

level to analyze low express protein in a single cell, knowing how the protein is homogeneously labeled can permit to furnish a correction factor to improve the quantification. In this case, the correction factor may be the *LO*.

It is actually possible to test that the *LO* is the correction factor by using data already generated, especially those from the migration of labeled cell lysate experiment, visualized by fluorescence or by CCB, that theoretically allows the visualization of all the protein. Proteins can be separated by their size, but also by their abundance to create a heatmap of 30 by 30 units on their log10 scale axis, allowing a greater readability. The first parameter to look at is the lysine number by protein (Figure 42 A), where it is obvious that big protein has on average a higher number of lysine. Little more unusual is the abundance of repartition, with the high number of lysine protein presenting a low abundance, especially at a high molecular weight. It is also possible to make the heat-map of the *LO* value, creating an obvious column pattern of different and mainly separated molecular weight with different *LO* value (Figure 42 B). The cyanine 3 and CCB molecular weight mapping is using the same system described in the results, with the value in the heat-map being the normalized intensity value. The low resolution of the heatmap does not allow to capture all the bands of the different protein at a different size that we can see on the gel, but instead mix them, even if it is still possible to see some bands, especially on the cyanine 3 heatmap (Figure 42 C and D). Following the hypothesis that the lysine is the most important factor to allow a homogeneous labeling and then enable a better quantification, the lysine heatmap represent the wished results that we want to obtain after correction of the fluorescence intensity by the *LO*. Ideally, this correction follows the formula established in part 1:

$$LO = 100 - 67.6 \times \exp(-0.036 \times n_{lys})$$

where we want to obtain the parameter n_{lys}, so with the modified following formula:

$$n_{lys} = \log((LO-100)/-71.5) \times (\text{Cyanine 3 intensity}/-0.039)$$

Making the fluorescence intensity being corrected by the *LO* to inform about the number of lysines, number from which the dynamic can be compared with the original assumption (Figure 42 E). We can then observe that the number of lysines is globally underestimated, but the dynamic between the theoretical and the calculated number is quite similar. From this point, it is also possible to calculate and visualize the number of fluorescent dye that can be

fixed on the protein (Figure 42 F). In this case, without surprise, the number of dye quickly saturate to the maximum, equal to 3.3.

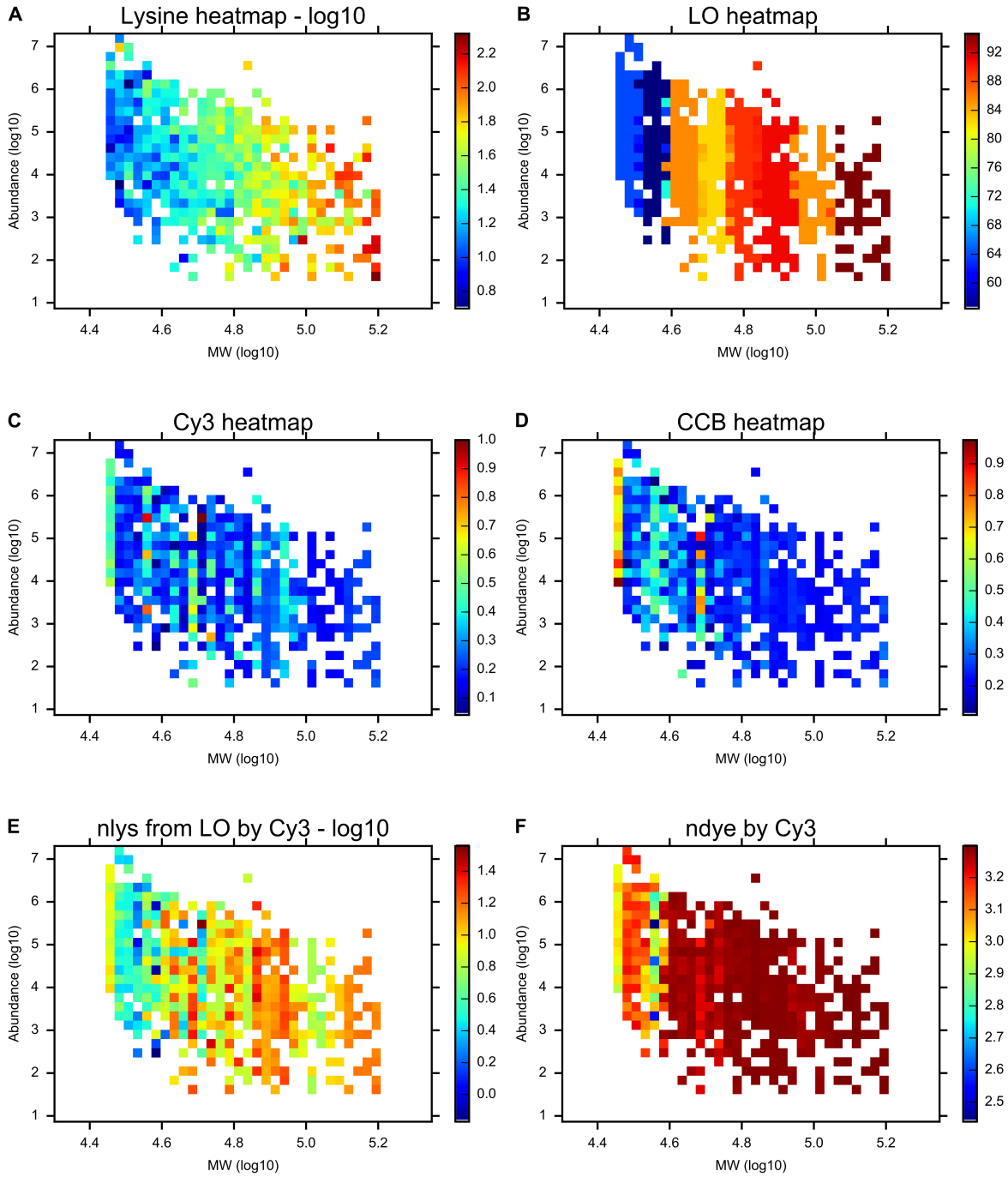


Figure 42: Heatmap to determine the correction factor efficiency. Both axis on all heatmap are on log10 scale. A. Theoretical average lysine number heatmap. B. LO heatmap. C. Cyanine 3 fluorescence intensity extract from cellular profile on SDS-PAGE gel. D. CCB intensity extract from cellular profile on SDS-PAGE gel. E. Correction factor apply using the formula nlys = log((LO-100)/-71.5)×(Cyanine 3 intensity/-0.039). F. Average number of dye by protein.

The proposed correction is validating the original model that was assuming that the number of lysines is important to obtain a high labeling homogeneity. It also proves that it is possible to correct results from a bulk experiment to obtain more quantitative results, enabling a wider use of the *LO* for protein labeling efficiency measure or for proteome quantitative measurement. However, this correction factor by itself is not a guarantee of quantification, and if it is validated on an ensemble of protein like a cell lysate, it is very likely that this correction factor needs to be measured to allow a better accuracy in the case especially in the case of specific protein. Indeed, during the protocol development, several purified proteins have been tested for the *LO*, like the catalase, avidin or lysozyme. In case of the lysozyme, the *LO* was at 74%, but we lose around 90% of the protein by precipitation, and then only measure the *LO* of the solubilize part. We do not know of the labeled proportion of protein is identical between the precipitate and solubilize protein, meaning that we do not know if the *LO*, and then the correction factor is correct in this case. It is then easy to make the parallel between this example and some specific proteins. For example, histone proteins are basic proteins, and classical protocol for their extraction require an acid extraction (Shechter et al. 2007), signifying that our current protocol, that works at a basic pH, is unlikely to solubilize and label them. This parameter, the protein solubility in solution, is still an open question in biochemistry.

The next step will be to enable the single molecule visualization of a proteome profile. This will allow absolute quantification, especially with the addition of the labeling occupancy with its correction factor. Single molecule sensitivity will also open the door to the single cell proteome profiling, enable the characterization of mechanisms like cell differentiation or cell communication, especially for low express proteins that are only present in a few copies in a single cell, and allow a better understanding of the central dogma at the single cell level.
When trying to separate protein in a microfluidic channel by microchip, the separation resolution was not enough to enable the analysis of a cellular proteome. However, capillary electrophoresis (CE) is already able to analyze a proteome profile with its capacity to resolve thousands of protein size (Wilson et al. 2015), justifying the use of the CE. Protein detection using a CE is realized using a laser-induced fluorescence (LIF) detector, that lack sensitivity,

limiting the minimum amount of sample to few thousands of cells. The easy solution will be to replace the LIF detector by single-molecule microscopy. The main down point of this is that the capillary glass thickness is around 150 µm, with the inside diameter at 50 µm, posing trouble for the working distance of single molecule objective lens (around 120-150 µm). In addition, the volume observed is a fraction of the volume of the capillary, resulting in a lower sensitivity. To solve these trouble, treating the outside capillary wall by hydrofluoric acid permit to obtain thinner capillary wall of 40-50 µm, and the observation of a cross-section of the capillary can be realized by a light sheet microscope (Figure 43 A), allowing the observation of the large portion of the migration volume, and then the counting of single molecule.

We design a proof of concept experiment with this kind of setting, and successfully observe single molecule signal of a fluorescently labeled BSA diffusing in the CE using a light-sheet microscope (Figure 43 B). This project will continue with the migration of purified protein first in a 1-meter length CE before direct observation with the light-sheet microscope, followed by labeled cell lysate then single cell. The background is still too high to enable single molecule detection and counting with a high confidence, even if a brighter single molecule can be observed, but can be improved by a better control of the component of the migration. This experiment can open the door to single molecule proteomic, and in the future, a possible combination between the microscope fluorescence quantification and the mass-spectroscopy protein identification.

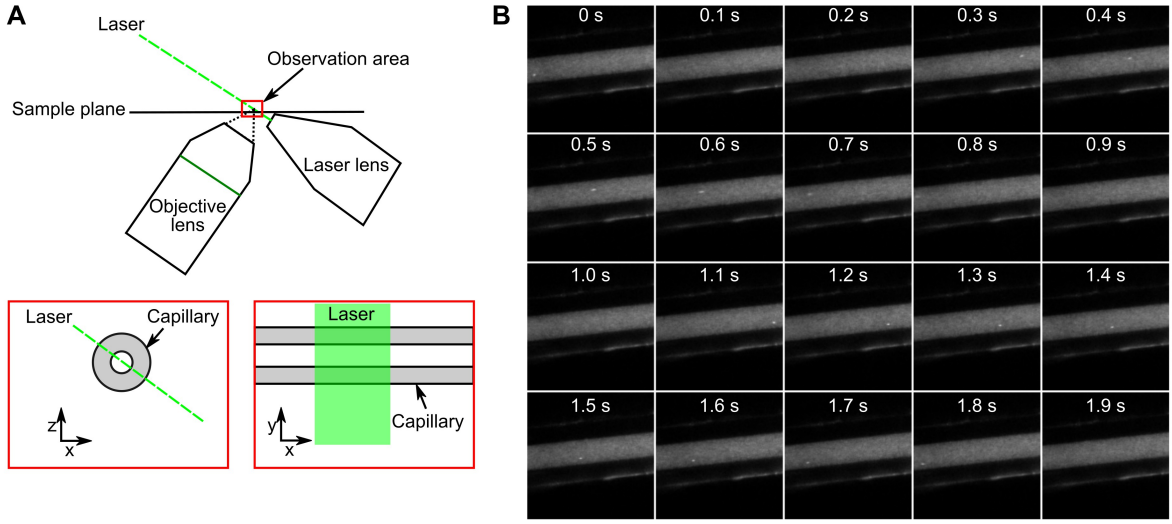


Figure 43: Light-sheet microscopy for capillary electrophoresis. A. Side view of the handmade light-sheet microscope, with detailed view of the observation area with the capillary. B. Timelapse images of single molecule labeled BSA migration.

On the first part of this thesis, the labeling homogeneity proposes a correction factor allowing a more rigorous fluorescence quantification. Like previously discussed, this can be applied outside the single molecule domain, enlarging the possible application to classical biochemistry with 2D-DIGE. However, the actual proteome profiling is mainly using mass-spectroscopy and shotgun analysis, less quantitative, but more qualitative. The fluorescent proteome profiling, even if containing a lot of information, is not satisfying enough, being unable to identify a or a set of proteins of interest, making the *LO* measure irrelevant. One of the domains that can profit the more about the *LO* measurement is the antibody labeling. A lot of commercially available antibodies are fluorescently labeled with a dye, especially the secondary antibody. However, the main companies selling such antibodies generally do not inform about their labeling efficiency, CE or DoL. For an application like super-resolution microscopy, where the number and position of a single molecule are important, antibody information is not found at single molecule level, but at best as a CE measure that has been optimized to be equal to 1. In this case, the *LO* can permit to calculate the probability that a single molecule signal is overlapping with itself because of a double dye on a single antibody and to correct the total number of single molecule counted. Another application with an antibody that is more global is the use of the correction factor when protein is quantified by

fluorescence WB. If fluorescence WB can be directly used for relative quantification, absolute quantification needs the correction factor, or at worst the CE.

The interest for the proteome profiling resides in the multiple possibilities of utilization. In research, it can be used to found new biomarkers, even if it does not allow a strict identification. Since biomarkers are mainly used for cell identification, their identification and characterization are an important task. In the biomedical field, a quick and cheap method to acquire the proteome profile can permit to determine the action of a drug on a given cell, and possibly discover new action mechanism. The field that can have the higher merit in the application of proteome profiling will be the clinical field. The quick identification of biomarker in a sample will permit a quick diagnostic. In addition, the high sensitivity will permit to reduce the sample amount used for this analysis, allowing the use of other technique for cross-validation or a smaller starting sample from the initial biopsy. The sample can not only be a cell or a tissue sample, but can also be any other form of liquid sample like the blood plasma, the cerebrospinal fluid or the eye vitreous body for example, for the identification of biomarker that can reveal a disease (Geyer et al. 2016; Guldbrandsen et al. 2014; Murthy et al. 2014).

Résumé étendu en Français

Introduction

L'expression des protéines d'une cellule est un phénomène qui se décompose en deux étapes principales, la transcription et la traduction. Certaines gènes sont fortement régulés de part l'activité de leur promoteur, en résultant des piques de transcription lorsque le promoteur est activé. L'ARN messager en résultant possède également une durée de vie, qui peut être plus ou moins courte, avant d'être finalement traduit par le ribosome. Tous ces phénomènes peuvent produire des piques de production de protéine aléatoire, ce qui produit une hétérogénéité dans la population cellulaire. De telle variabilité dans l'expression de protéine peut amener à divers phénotype cellulaire et affecter une large gamme de fonctions cellulaires, incluant le développement, l'homéostasie et la progression de maladie.

Comprendre les causes et les conséquences de la variabilité dans l'expression génétique nécessite une quantification précise des ARNm et des protéines au niveau de la cellule individuelle. Compter des molécules d'ARNm dans une cellule individuelle est actuellement réalisé grâce à la combinaison d'appareil manipulant des cellules individuelles et au séquençage de nouvelle génération. En revanche, compter des protéines est plus difficiles, et nécessite la manipulation génétique, des anticorps ou de techniques plus sophistiquées. Des techniques ont ainsi été développées afin de pouvoir mesurer plusieurs protéines sur ou dans la même cellule individuelle, comme la cytométrie ou le Western-Blot à cellule individuelle. Cependant, ces techniques fournissent au mieux une quantification relative et échouent généralement à détecter des protéines faiblement exprimées qui ne sont présentes qu'en quelques copies, qui peuvent être de facteurs de transcription ou des régulateurs de l'état cellulaire, protéines pouvant changées le phénotype cellulaire.

Afin de détecter de telles protéines, il est nécessaire d'augmenter la sensibilité des tests, jusqu'au niveau de pouvoir compter les protéines individuellement d'une cellule individuelle. Il est actuellement déjà possible de compter des protéines d'une cellule unique. Une des techniques les plus prometteuse est le nanopore, qui permet non seulement de détecter quand une protéine passe à travers le nanopore, mais également de les identifier de part leur séquence. Cette technique est cependant toujours immature, et les anticorps marquées de manière fluorescente ou chimique sont actuellement les plus utilisés. Dans ce cas, les anticorps sont isolées de manières à pouvoir les compter, soit dans des micro-puits, des micro-billes ou sur une micro-puce. Les méthodes reposant sur les anticorps sont cependant limitées dans leur possibilité d'analyser le protéome, ou l'ensemble des protéines, d'une cellule unique, de part la disponibilitée de ces anticorps. Afin d'analyser tout un protéome, la

spectroscopie de masse est l'outil le plus populaire, mais ne permet pas une telle sensibilité. Une autre méthode est de marquer de manière non spécifique le protéome cellulaire avec une sonde fluorescente, suivi d'une séparation des protéines. La thèse se décompose en deux parties, la première sur la méthode développée pour marquer le protéome et comment cette méthode a été validée, et la deuxième partie est consacrée à la séparation du protéome cellulaire marqué afin d'obtenir un profil protéomique.

Résultats

Détermination de l'homogénéité de marquage

Les marqueurs fluorescent réactifs sont communément utilisés dans le cadre d'électrophorèse sur gel, en particulier dans le cas d'électrophorèse différentielle sur gel en deux dimensions (2D-DIGE). Dans ce cas, les marqueurs fluorescents les plus utilisés sont de type cyanine, avec Cy3 étant le marqueur le plus utilisé de part ses propriétés d'illumination similaire au TRITC. Ce marqueur est utilisé pour la détection de molécule individuelle, incluant des protéines, et peut également posséder un site réactif lui permettant de reconnaître des protéines. Il est actuellement non réalisable de marquer tout le protéome avec ce marqueur, il y aura toujours une portion des protéines qui ne seront pas solubles ou certaines protéines non marquées parce que le site de réaction ne sera pas accessible. Il est donc nécessaire de connaître cette proportion avant de pouvoir quantifier les protéines, spécialement dans le cas d'un comptage de protéine individuelle.

Actuellement, la manière la plus commune de l'efficacité de marquage est la mesure de l'efficacité de couplage (CE), aussi appelé degré de marquage (DoL), et consiste en un ratio de la quantité de marqueur et de la quantité de protéine au spectrophotomètre. Parce que le CE est une mesure en masse, il n'est pas possible de mesurer l'homogénéité du marquage, que l'on définie comme étant la proportion de marqueur par protéine. pour pouvoir réaliser cette mesure, il est nécessaire d'être capable de mesurer l'homogénéité du marquage au niveau de protéine individuelle. Nous appelons la proportion de protéine marqué par au moins un marqueur l'occupation de marquage ou LO. Le LO représente ainsi la probabilité de marquer la protéine et peut être utilisé comme facteur de correction afin d'estimer le nombre absolu de protéines.

Pour mesurer le LO, il est nécessaire de pouvoir compter la totalité des protéines, puis uniquement celle qui sont marquées. Vu qu'il n'est pas possible de compter quelque chose qui n'est pas marqué, on a décidé de fixer par un point d'ancrage un nombre fixe de protéines sur une lamelle de microscopie, qu'elles soit marquées ou non. En connaissant la

densité de point d'ancrage, ici l'avidine, il est possible de connaître le nombre total de protéine individuelle, puis de mesurer la proportion de protéines marquées (Figure 44A).

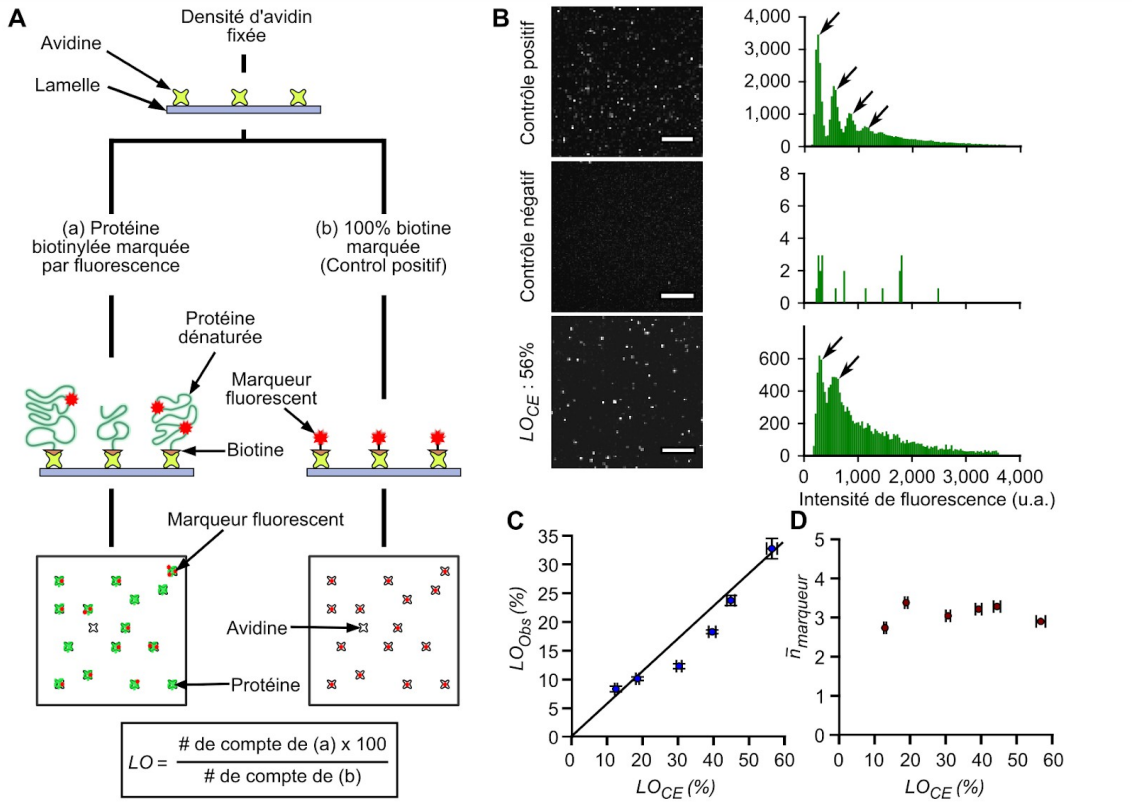


Figure 44: Principe et validation de la mesure du LO. A. Une densité fixe d'avidine est fixée à une lamelle de microscopie après un traitement au plasma, et l'échantillon de protéine biotinylée marquée par fluorescence est déposée sur la lamelle (a), où sont compter les signaux fluorescent. Ce nombre est ensuite comparer avec la mesure du nombre d'avidine, visualisé sur une seconde lamelle à l'aide de biotine marquée (b). B. Contrôles pour la réaction avidine-biotine, avec les flèches indiquant le nombre de sites actifs de l'avidine reconnus, la dernière image étant une mesure du LO avec de la BSA. C. Comparaison entre le LO estimé à partir du CE (LOCE) et observé par protéine individuelle (LOObs). La régression linéaire est montrée par la ligne noire (R2=0.95). D. Intensité moyenne du signal d'une protéine unique à travers différente dilution de la protéine marquée.

Avant de pouvoir utiliser ce test, des étapes de validation sont nécessaire. La réaction avidine-biotine utilisé ici est très spécifique, et avec 4 sites réactifs par avidin, on peut avoir jusqu'à quatre biotines par ancre dans le cas du contrôle positif (Figure 44B), même si la majorité des avidines ont entre 1 ou 2 biotines. Le contrôle négatif montre que la biotine n'est quasiment pas absorbée sur la lamelle de verre, et la dernière condition montre de la BSA marquée par fluorescence, avec deux pics, qui représente probablement une protéine fixée à l'avidine marquée par une ou deux marqueurs fluorescent. Pour vérifier si les deux pics de la BSA est bien le résultat de la fixation d'une seule protéine, une solution est de

diluer cette protéine marquée dans de la protéine non marquée, et mesurer par le CE (LO_{CE}) et comparer aux résultats par protéine individuelle (LO_{Obs}), où l'on observe une diminution linéaire du signal dans les deux cas (Figure 44C). De plus, le nombre de marqueur moyen mesurer par protéine individuelle reste constant quelque soit la dilution de la protéine marquée, signifiant qu'il n'y a qu'une seule protéine de BSA par point d'ancrage, ou d'avidine (Figure 44D).

Afin d'avoir des résultats comparables, il est nécessaire que l'échantillon soit en excès vis-à-vis du nombre de site d'avidine présent sur la lamelle. De plus, lors de l'étape de marquage, le marqueur fluorescent doit également être en excès, jusqu'à 20 fois. Enfin, la lamelle peut présenter des hétérogénéité dans la densité d'avidine fixée, et il est nécessaire d'acquérir au moins une cinquantaine d'images à différentes positions afin d'obtenir une mesure précise. En respectant ces paramètres, il est possible d'avoir une grande reproductibilité avec un minimum de variations entre différentes mesure.

L'efficacité de marquage est dépendante des conditions d'accessibilité du marqueur fluorescent avec la protéine, et plus précisément avec les résidus de lysine. Cependant, tous les résidus ne sont pas identiquement accessible, sans compter les variations de structure de la protéine. La solution la plus simple est de complètement dénaturer la protéine d'intérêt afin qu'un maximum de site réactif soit accessible et ait la même structure. L'homogénéité du marquage de la BSA est le paramètre que l'on souhaite moduler, et pour ce faire, la dénaturation peut être réalisée par une augmentation du pH, une réduction de la protéin et la présence de détergent et de tensioactif. Différente composition de ces paramètres ont été testée, et l'on observe qu'un protocol de marquage classique est d'environ 30%, alors que rajouter des éléments ou d'augmenter le pH permet de la hausse du LO_{Obs}, et donc de l'homogénéité du marquage (Figure 45).

Condition	Tampon	DTT	SDS	Tween 20	CHAPS
A	1x PBS pH 7.4	1 mM	0.1 %	0.1 %	0 %
B	50 mM borate pH 12	1 mM	0.1 %	0.1 %	0 %
C	50 mM borate pH 12	1 mM	0.1 %	1 %	0 %
D	50 mM borate pH 12	1 mM	1 %	0.1 %	0 %
E	50 mM borate pH 12	1 mM	1 %	1 %	0 %
F	50 mM borate pH 12	10 mM	0.1 %	0.1 %	0 %
G	50 mM borate pH 12	10 mM	0.1 %	1 %	0 %
H	50 mM borate pH 12	10 mM	1 %	0.1 %	0 %
I	50 mM borate pH 12	10 mM	1 %	1 %	0 %
J	50 mM borate pH 12	10 mM	1 %	1 %	2 %

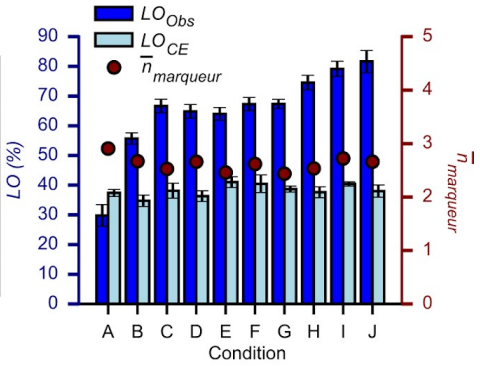


Figure 45: Homogénéité du marquage de la BSA à différente condition. La condition A est la condition native, alors que les conditions B à J sont de plus en plus dénaturantes. Le LOObs est en bleu, le LOCE est en bleu clair et le nombre moyen de marqueur est représenté en rouge foncé.

Marquer des protéines purifiées est une pratique courante, spécialement dans le cadre du marquage des anticorps. Il est cependant peu fréquent de marquer un protéome en entier, principalement car des pertes sont inévitables, avec des protéines peu ou pas soluble dans la condition testée, même si cela est réalisé dans le cadre de DIGE. Cependant, réaliser le marquage du protéome et mesurer l'homogénéité du marquage peut permettre de corriger des biais, spécialement en comptant au niveau de protéine individuelles. La méthode la plus rapide consiste à faire migrer des protéines déjà marquées par fluorescence sur un gel de SDS-PAGE afin de les séparer par la taille, puis de les marquer de manière non spécifique avec un autre marqueur, comme le bleu de Coomassie (Figure 46A) afin de comparer les profils de migration. Dans cette expérience, on peut observer un profile de migration similaire, même si l'intensité de certaines bandes sont différentes. Par la suite, des bandes du gel, représentant des protéine de différent poids moléculaire ont été excisées, les protéines extraites et finalement le LO mesuré (Figure 46B). On peut dans ce cas observer que les protéines de faibles poids moléculaire ont un LO plus faible, donc une plus grande hétérogénéité dans le marquage, comparé à des protéines de plus haut poids moléculaire (Figure 46C). L'explication la plus simple est que le nombre de site actif susceptible de réagir avec le marqueur fluorescent, la lysine, est présent en plus grand nombre sur des protéines de grand poids moléculaire.

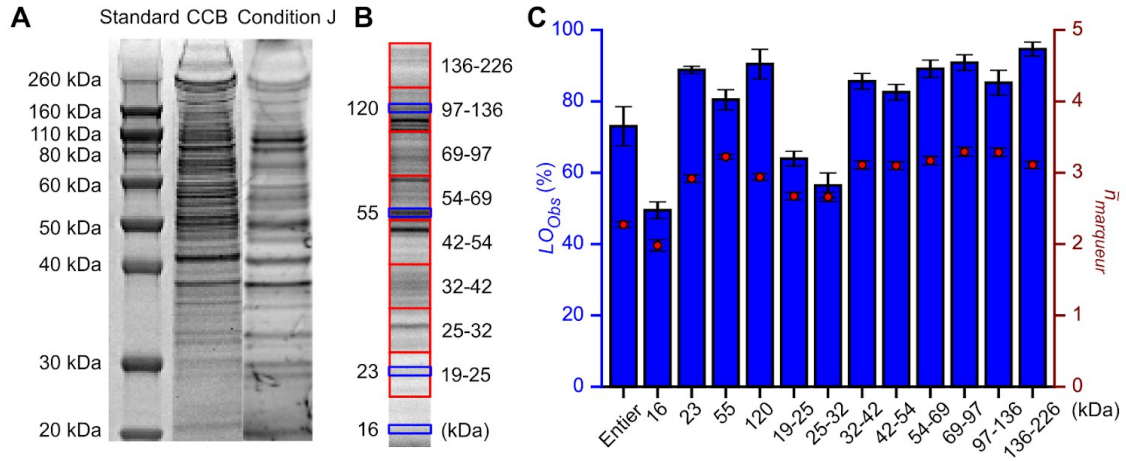


Figure 46: Homogénéité du marquage à travers le protéome. A. Comparaison entre le marquage au bleu de Coomassie et fluorescent. B. En bleu, les protéines extraites de bandes et en rouge les protéines extraites de régions. C. Mesure du LOOBs du protéome entier, des bandes et des régions spécifique du protéome en bleu, et en rouge foncé le nombre moyen de marqueur par condition.

Durant ma thèse, j'ai développé une méthode pour mesurer l'hétérogénéité de marquage d'une echantillon, que cet echantillon soit une protéine purifiée ou un une complexe mixture

de protéine. Il faut cependant prendre en compte certains point avant d'appliquer cette mesure d'hétérogénéité.

Le premier point est sur la passivation de la lamelle de microscopie. Afin de fixer la protéine d'avidine sur la lamelle, il est nécessaire de complètement enlever toute poussière qui pourrait contribuer à faire du bruit, mais également pouvoir fixer la protéine. Ce nettoyage est réalisé avec un traitement au plasma, qui permet en même temps l'activation de la lamelle autorisant la fixation de l'avidine. C'est un traitement extremement simple, mais qui est également sujet à des hétérogénéités, ce qui peut amener à des erreurs dans le comptage. Ces erreurs sont normalement réduite par le grand nombre d'images utilisé pour déterminer le LO, mais il peut également être possible d'améliorer le protocol de passivation et de réduire le nombre de faux positif en réutilisant des protocoles utilisés pour l'analyse de protéine unique par FRET. Dans ces protocols, le nettoyage est réalisé par une solution d'hydroxide de potassium, et la passivation par des molécules de PEG. Parmis ces PEG, certains ont été modifiés pour exhiber une biotine à une extrémitée. En contrôlant la proportion de PEG-biotine, il devient possible de contrôler la densité de molécule unique que l'on peut compter de manière plus précise que l'actuelle protocol. Cela permettrai également réduire le nombre de protéines qui sont absorbées sur la surface en verre, et donc le nombre de faux positif:

Il est possible d'estimer le nombre de lysine de chaque bande ou régions extraites du gel. La base de données UniProt informe sur la séquence et le poids moléculaire de chaque protéines, alors que la base de donnée Paxdb fournit l'abondance de protéines dans certains type cellulaire. Le lysat cellulaire ayant été obtenu à partir de cellule humaine Hela, il est possible de calculer une moyenne pondérée entre l'abondance et le nombre de lysine pour chaque protéine exprimée dans les cellules Hela, puis de les séparer par poids moléculaire correspondant aux bandes ou régions extraites du gel. On observe dans ce cas une relation entre le nombre du lysine et le LO ou le nombre moyen de marqueur, que l'on peut synthétiser avec les équations suivantes : $LO = 100 - 71.5\exp(-0.039n_{lys})$ et $n_{marqueur} = 3.3 - 2.6\exp(-0.08n_{lys})$. Si le LO se rapproche de 100% lorsque le nombre de lysine augmente, le nombre de marqueur par protéine sature rapidement aux alentours de 3 marqueurs par protéines.

Le protocol développé pour mesurer l'hétérogénéité de marquage d'une protéine à permis d'optimiser le protocol de marquage afin d'augmenter la proportion de protéines marquées et de diminuer l'hétérogénéité de marquage, mais a également permis de mesurer la proportion de protéines marquées d'un lysat cellulaire. En fractionnant ce lysat cellulaire, on a pu mettre en évidence que le nombre de lysine de la protéine est le facteur critique afin de

permettre à la fois un marquage homogène et efficace. Ce protocol a également permis de mettre en évidence qu'il est possible de marquer de manière non spécifique une partie représentative des protéines composant le protéome cellulaire, mais qu'il est aussi nécessaire de séparer ce protéome afin de pouvoir en extraire des informations.

Séparation de protéomes pour profilage

La profilage de protéome est l'information concernant toutes les protéines présente dans un échantillon biologique à un moment donné. Ce profil peut être utilisé à des fins d'identifications d'échantillon ou afin de découvrir et comprendre des mécanismes cellulaires ou moléculaires. Ce genre de profil peut facilement être réalisé par techniques de séparation des protéines comme le SDS-PAGE ou le focus isoélectrique, ou plus récemment avec la spectroscopie de masse. Une augmentation de la sensibilité de ces techniques est recherchée, principalement afin d'augmenter la détection de protéines de faible abondance ou de diminuer la quantité d'échantillon. Dans cette optique, atteindre une sensibilité jusqu'à une molécule individuelle serait utile et permettrait de révéler de nouvelles informations, particulièrement sur l'identification de cellule cancéreuse.

La séparation de protéines d'un échantillon aussi complexe qu'un protéome de cellule humaine est principalement basé sur une séparation sur la taille des protéines, comme décrit dans la partie 1. Cette séparation peut être réalisée par différente méthode, comme le SDS-PAGE, l'électrophorèse en capillaire ou par chromatographie d'exclusion stérique. Chaque technique dispose de ces avantages et inconvénients, mais on va surtout se concentrer sur le SDS-PAGE et l'électrophorèse en capillaire qui sont les plus à même d'atteindre la sensibilité désirée.

L'électrophorèse en capillaire (CE) est une famille de méthode de séparation par electro-kinetic réalisé dans des capillaire ou des micro-canaux d'un diamètre de quelques micromètres. CE est une technique qui permet de séparer les protéines par leur taille et leur propriétés ioniques, qui sont déterminé par la composition en amino-acide de la-dite protéine. Lorsqu'une protéine est soumise à un champs électrique, elle va se déplacer vers l'électrode ayant une charge opposée à sa propre charge. Les ions du buffer composant le milieux liquide réagissent également de la même façon, créant un flux électro-osmotique. Vu que le flux électro-osmotique est plus grand que le flux des protéines, les protéines font se faire emmener par ce flux dominant. Ceci va résulter en la séparation des protéines, avec les protéines de grosses tailles fortement chargées migrant rapidement alors que les protéines de petites tailles fortement chargées passeront devant le détecteur en dernier. Ce phénomène est d'autant plus accentué si les protéines sont en présence de SDS, ce qui va uniformiser leur charge électrique dépendant de leur taille.

Nous avons tenté de séparer des protéines dans une puce microfluidique à base de PDMS, doté volontairement d'une conception simple, appelé double T, qui permet la concentration d'un petit volume d'échantillon, ici 0.5nl (Figure 47A), et d'un capillaire d'une longueur totale de 2 cm. Dans un premier temps, nous avons marqué avec un marqueur fluorescent deux protéines, la BSA et le lysozyme, que nous avons ensuite injecté séparément, puis nous avons observé la fluorescence à la fin du capillaire à l'aide d'un microscope à haute sensibilité. La migration a été réalisée à 500V, soit 250V/cm, et permet la migration des protéines en moins de 5 secondes, ce qui est consistant avec la littérature. Lorsque les deux types de protéines sont mélangés, on observe bien l'apparition de deux pics, qui représente la séparation des protéines (Figure 47B). Cette séparation n'est cependant pas très importante, surtout sachant que le lysozyme est une petite protéine d'environ 16 kDa alors que la BSA est une protéine de poids moléculaire moyen de 56 kDa. Si comme on l'attendait, la BSA est bien séparée en premier, la vitesse de migration est telle qu'elle ne permet pas une séparation bien net.

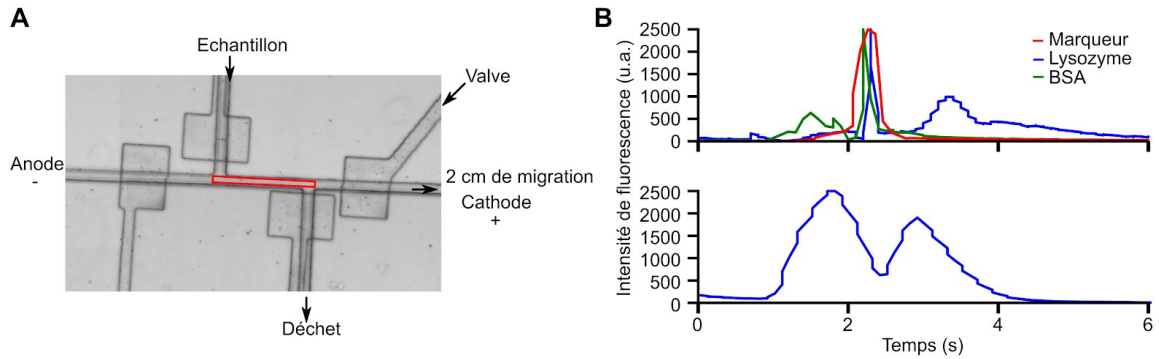


Figure 47: Séparation de protéines par puce microfluidique. A. Agrandissement du double T. Un système de valve permet de contrôler l'injection de l'échantillon et la migration. Le rectangle rouge correspond au volume d'échantillon qui va migré, soit 0.5 nl. B. Profil de migration de protéines purifiées ou mélangées. En haut, temps de migration de protéines purifiées ou du marqueur seul. En bas, profil de migration de la BSA et du lysozyme après purification.

Afin d'améliorer la séparation des protéines, et donc la résolution de la micropuce, on a essayé la même séparation avec un capillaire d'une longueur de 35 cm, ainsi que l'incorporation d'un polymère normalement utilisé en capillaire en verre. Malheureusement, appliqué un voltage de 250V/cm résulte en l'évaporation du tampon dans le capillaire par l'effet Joule, et réduire ce voltage résulte en une absence complète de migration.

En conclusion de cet essai, il est possible de séparer quelques espèces de protéines dans une micropuces, cependant, ces puces ne disposent pas de la résolution pour séparer les dizaines, voir les centaines de protéines majeures composants le protéome. De plus, les

quelques essais pour séparer des protéines par microfluidique reposent principalement sur des puces en acrylique (PMMA) qui est plus inerte que le PDMS vis-à-vis des protéines, et augmente leur résolution en séparant les protéines sur deux dimensions, généralement la taille et l'hydrophobicité.

Une autre possibilité pour séparer les protéines est d'utiliser une des plus vieille technique de séparation : le SDS-PAGE. Afin de pouvoir observer un gel de SDS-PAGE sous le microscope afin de détecter des molécules individuelles, il est nécessaire de le modifier afin de réduire l'autofluorescence et d'utiliser un gel plus fin afin de concentrer le signal. Enfin, il est également nécessaire que le gel soit fixé sur un support permettant sa visualisation par un microscope. Afin de réunir toute ces conditions, le gel d'acrylamide est fixé sur une grande lamelle de microscopie de 50 sur 75 mm en rendant la surface du verre réactif après réaction avec du Bind-Silane. Le moule est lui réalisé en utilisant de la lithographie, et présente une épaisseur de 0.2 mm, avec 11 puits de 1 sur 5 mm chacun, permettant l'injection de 1 microlitre d'échantillon par puit, et traité avec un réactif hydrophobe afin d'éviter la polymérisation du gel sur sa surface.

Le temps de migration a été déterminé expérimentalement en faisant migrer un standard de taille fluorescent, visualiser par un observateur fluorescent de gel (Figure 48A) et a été fixé à 180 minutes afin d'optimiser la séparation de protéines de haut poids moléculaire, qui présente généralement un marquage fluorescent plus fort et plus efficace. Le gel a ensuite été observé au microscope en réalisant un quadrillage d'observation afin de déterminer la sensibilité de la technique (Figure 48B). Avec le microscope, on ne peut observer que jusqu'à 100 picogrammes de protéines, et non compter le nombre de protéines individuelles, principalement parce que la surface du gel est auto-fluorescent. Cette auto-fluorescence peut être causée par de la diffusion de lumière ou par la présence de cristaux ou de poussières. En utilisant ce système, il est possible de séparer deux protéines de poids moléculaire proches, comme la BSA (66 kDa) et la sous-unité de la catalase (50 kDa) et d'obtenir un profile répétable sur différent gel (Figure 48C), permettant ainsi de valider le gel.

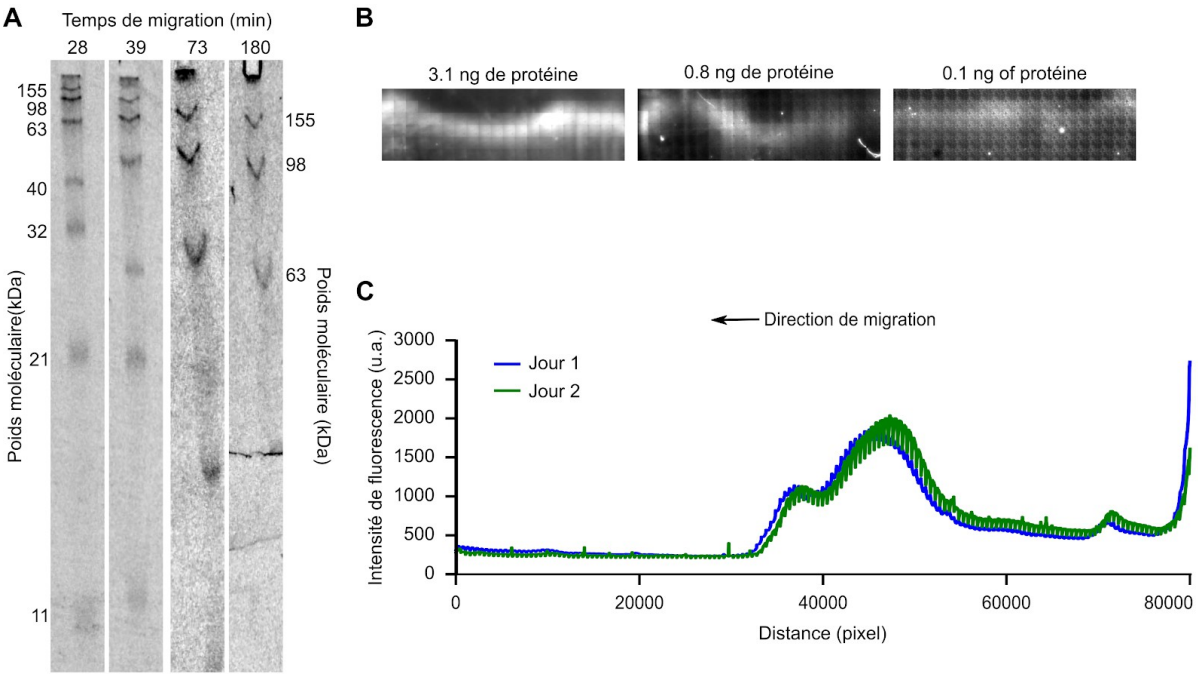
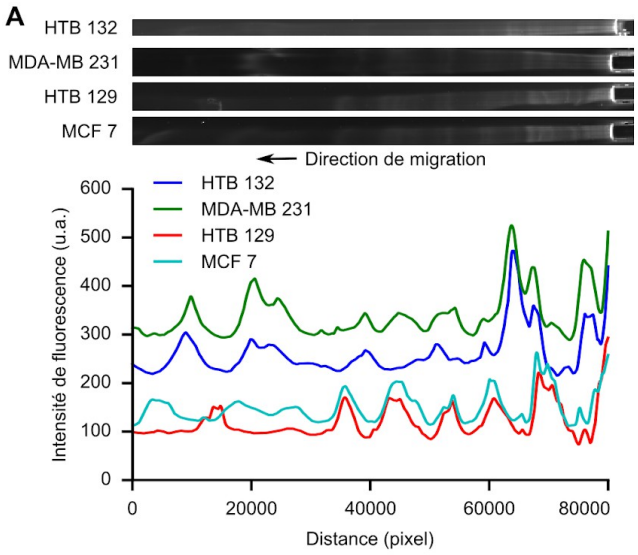


Figure 48: Paramètres du gel. A. Temps de migration d'un standard de taille fluorescent pour protéine. B. Observation de différentes quantité de bandes de protéines à l'aide du microscope. C. Profil de migration d'un mélange de BSA (66 kDa) et de sous-unité de catalase (50 kDa) marqué de manière à l'aide d'un marqueur fluorescent sur deux gels différents. La reconstruction du second gel n'est pas optimal, et nécessite un lissage.

Le profile de protéome peut révéler des informations précieuses, dont une des principales est la capacité de différencier différent types cellulaire ou cancéreux. Vu que le gel développé permet la détection d'environ 100 picogrammes de protéines par bande, et qu'une cellule de mammifère contient entre 50 et 75 picogrammes de protéine, il a été décidé d'utiliser 100 cellules par analyse. En réalisant une rapide analyse bio-informatique, on s'attend à observer une vingtaine de bandes de protéines dans l'intervalle 50-200 kDa qui présente une abondance suffisante pour franchir la limite des 100 picogrammes. A partir de ces vingt bandes, on espère pouvoir identifier 4 différents type cellulaire de cellules en culture provenant tous de tissue cancéreux du sein. Trois de ces lignées sont issues de tissu épithélial (HTB 132, MDA-MB 231 and MCF 7) et la dernière d'un mélanome(HTB 129), et elles permettent toutes l'étude de la transition épithélio-mésenchymateuse, phénomène observé au cours des processus de cicatrisation et de fibrose et au cours du développement tumoral. Le balayage de la lamelle de microscopie est réalisé en utilisant un microscope haute sensibilité, un système d'auto-focus et un objectif à immersion 30x disposant d'une grande ouverture numérique, autorisant un pas de 274 µm, ce qui nécessite un total de 160 images pour scanner 4.3 mm de long, et une largeur variant de 5 à 10 images, soit 1 à 2.7 mm. Les images sont automatiquement traitées par un programme réalisant sur Python, qui permet la reconstitution de toute la bande de migration, ainsi que la suppression de tous les

pixels aux valeurs aberrantes, qui proviennent généralement de poussières présent à la surface du gel. Le profile de migration est ensuite réalisé en faisant la moyenne des pixels par colonne, et le profile lissé en utilisant le filtre de Savitzky-Golay (Figure 49A). On observe que certains profils sont plus similaires que d'autre, et pour pouvoir les comparer de manières plus rigoureuses, une matrice de Pearson est calculée sur une partie ou la totalité du profil (Figure 49B). On peut dans ce cas observé que les cellules provenant du mélanome (HTB 129) semble être séparées des autres, même si le résultat le plus évident qui en ressort est que HTB 132 et MDA-MB 231 présentent des profils extrêmement similaires.



B

Profil entier (0 à 80000 pixels)

	MDA-MB 231	HTB 129	MCF 7
HTB 132	0.99	0.92	0.9
MDA-MB 231	1.0	0.93	0.92
HTB 129		1.0	0.96

0 à 30000 pixels

	MDA-MB 231	HTB 129	MCF 7
HTB 132	0.9	-0.02	-0.4
MDA-MB 231	1.0	-0.16	-0.19
HTB 129		1.0	-0.2

30000 à 55000 pixels

	MDA-MB 231	HTB 129	MCF 7
HTB 132	0.91	0.6	0.64
MDA-MB 231	1.0	0.82	0.86
HTB 129		1.0	0.97

Figure 49: Comparaison de profils cellulaire. A. Profil cellulaire de 100 cellules sur gel. Les profils sont le résultat de la moyenne de triplicata. B. Matrice de corrélation de Pearson sur le profil de migration entier, le premier ou le second tier. Le dernier tier n'est pas montré car toute les corrélations sont fortes.

Discussion

Le marquage du protéome est complexe, à cause de la nature des protéines d'une part, et à cause des conditions nécessaire pour leur solubilisation d'autre part. Cette incapacité de dissoudre et resuspendre la totalité des protéines composant le protéome induit un biais dans tout analyse, surtout lorsque l'on sait qu'environ la moitié du protéome est invisible au analyse classique. Le protocol développé afin de mesurer l'homogénéité de marquage du protéome ne prend pas en compte ces protéines dites "cachées", mais se concentre sur un marquage des protéines accessibles le plus homogène possible. Il serait très intéressant de faire varier les conditions de lyse des cellules et de dénaturation des protéines afin de

concentrer sur des types de protéines en particulier, comme une meilleur extraction de protéines basique ou acide en ajustant le pH. Il serait également intéressant de fractionner la cellule en fonction de ses compartiments cellulaires afin d'observer comment change l'efficacité de marquage. Le marquage est également influencé par la fonction chimique utilisée, dans ce cas NHS. NHS est capable de reconnaître les fonctions amines primaires, tel que ceux présent sur la lysine, l'arginine et la fonction terminale de la protéine. Cette fonction est également capable de reconnaître de manière non spécifique d'autre amino-acides, comme la fonction OH de la tyrosine. Si nos analyses montrent de façon clair que la lysine est l'amino-acide qui possède la plus grande corrélation avec le LO, d'autre amino-acides comme la proline et l'alanine donne également de bonne corrélation, même si ils sont censé être inerte au NHS. Ceci indique qu'il est nécessaire de vérifier si la fonction NHS peut réagir de manière non spécifique avec d'autres amino-acides.

Les protéines sont immobilisées par une réaction avidine-biotine afin de permettre leur comptage. Cette étape requiert la biotinylation de la protéine, et n'a pas été vérifiée. On fait la supposition que les protéines déjà marquées seront facilement biotinylé car elles sont accessible, induisant une sur-représentation des protéines marquées dans la mesure du LO. De l'autre côté, des protéines trop marquées par les marqueurs fluorescent sont suspectées d'être instable, de part l'augmentation de leur hydrophobicité, ce qui pourrait cette fois éliminer de l'analyse les protéines trop marquées. Il est possible de tester cette hypothèse en réalisant une double purification des protéines, avec comme première purification l'élimination des marqueurs n'ayant pas réagi avec la protéine, et la seconde purification permettant de retenir les protéines biotinylées par une chromatographie d'affinité en utilisant de l'avidine immobilisée. Il est alors possible de mesurer le CE avant et après la seconde colonne de filtration, permettant d'estimer quelle est la proportion de protéine non biotinylée.

Du point de vue protéomique, la technique de référence est la spectroscopie de masse, qui ne permet pas facilement une quantification absolue, mais permet en contrepartie une identification des protéines. La protéomique sur gel est l'inverse, avec le 2D-DIGE qui permet une quantification relative et/ou absolue, mais peu de moyen d'identifier la protéine d'intérêt. Dans ce contexte, il est possible de permettre une meilleure quantification des protéines en mesurant l'efficacité de marquage ainsi que son homogénéité à travers le protéine. AVoir cette mesure permettra de corriger les biais pouvant être causé par un marquage hétérogène ou incomplet, sans avoir le besoin de réaliser une mesure au niveau des molécules individuelles. Avec les données déjà dans la thèse, il est possible de démontrer l'utilité de ce facteur de correction, en utilisant les gels de SDS-PAGE, fluorescent ou non. En appliquant la formule de correction suivante sur le profile fluorescent

n_{lys} = log((*LO*-100)/-71.5)×(Intensité de Cyanine 3/-0.039)

Il est possible d'estimer le nombre de lysine à travers le protéome, et ainsi de comparer aux valeurs théoriques. On peut alors observer une grande ressemblance entre la fluorescence du gel corrigée par le facteur de correction et les données théoriques, permettant ainsi de valider le modèle. Cette analyse permet également de prouver qu'il est possible d'utiliser le facteur de correction, calculé à partir de molécules individuelles, pour une expérience classique, et ainsi d'obtenir des résultats plus quantitatifs. Il est cependant important de rappeler que le facteur de correction est valable pour une expérience sur le protéome, mais qu'il doit être mesuré lors de changement de conditions ou dans le cas de protéines spécifiques. Le facteur de correction est très intéressant dans le cas de mesures de protéines individuelles, surtout dans le cas des anticorps marquées par une marqueur fluorescent. L'homogénéité de marquage est important à vérifier pour de tels anticorps, surtout qu'ils sont généralement utilisés afin de réaliser des mesures quantitatives, aussi dans le cas de WB fluorescent, mais également dans le cas de microscopie à haute résolution comme le STORM. Dans les deux, les informations de l'homogénéité de marquage est important afin de pouvoir réaliser une mesure précise, au niveau de la quantification pour le WB fluorescent ou au niveau de la localisation et comptage dans le cas du STROM, ou une même protéine peut être compté plusieurs fois.

Conclusion

Nous avons montré qu'il est possible de mesurer l'hétérogénéité de marquage fluorescent d'une protéine au niveau des molécules individuelles. Cette hétérogénéité de marquage peut être modifié en variant les conditions de lyse et de marquage, en jouant sur le pH de la solution ainsi que sur la solubilité et la dénaturation de la protéine à marquer. Après optimisation du protocol de marquage en utilisant une protéine test, on a démontré qu'il est également possible de marquer un lysat cellulaire complet ou fractionné par poids moléculaire. Ce fractionnement a permis de montrer que l'homogénéité de marquage évolue à travers le nombre moyen de lysine dans la fraction analysée, avec un nombre de lysine important résultant en un marquage plus homogène. On a ensuite essayé de séparer le protéome par taille afin de le visualiser directement au niveau des protéines individuelles, principalement afin de permettre un comptage des protéines et l'identification de lignées cellulaires. Si les puces microfluidiques ne permettent pas une séparation satisfaisante des protéines, les gels de SDS ultra fin permettent une réduction significative du nombre de cellule à utiliser afin d'obtenir un profil, permettant également d'identifier différentes lignées

cellulaires, même si ils ne permettent pas de pouvoir compter des protéines individuelles à cause d'un bruit de fond trop important.

References

Aalderen, Michiel C. van, Maartje van den Biggelaar, Ester B. M. Remmerswaal, Floris P. J. van Alphen, Alexander B. Meijer, Ineke J. M. Ten Berge, and René A. W. van Lier. 2017. “Label-Free Analysis of CD8(+) T Cell Subset Proteomes Supports a Progressive Differentiation Model of Human-Virus-Specific T Cells.” *Cell Reports* 19 (5): 1068–79.

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” *Nature* 422 (6928): 198–207.

Albayrak, Cem, Christian A. Jordi, Christoph Zechner, Jing Lin, Colette A. Bichsel, Mustafa Khammash, and Savaş Tay. 2016. “Digital Quantification of Proteins and mRNA in Single Mammalian Cells.” *Molecular Cell* 61 (6): 914–24.

Altelaar, A. F. Maarten, Javier Munoz, and Albert J. R. Heck. 2013. “Next-Generation Proteomics: Towards an Integrative View of Proteome Dynamics.” *Nature Reviews. Genetics* 14 (1): 35–48.

Baltimore, D., H. J. Eggers, R. M. Franklin, and I. Tamm. 1963. “Poliovirus-Induced RNA Polymerase and the Effects of Virus-Specific Inhibitors on Its Production.” *Proceedings of the National Academy of Sciences of the United States of America* 49 (June): 843–49.

Bayer, E. A., and M. Wilchek. 1990. “Avidin Column as a Highly Efficient and Stable Alternative for Immobilization of Ligands for Affinity Chromatography.” *Journal of Molecular Recognition: JMR* 3 (3): 102–7.

Berggren, K., E. Chernokalskaya, T. H. Steinberg, C. Kemper, M. F. Lopez, Z. Diwu, R. P. Haugland, and W. F. Patton. 2000. “Background-Free, High Sensitivity Staining of Proteins in One- and Two-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gels Using a Luminescent Ruthenium Complex.” *Electrophoresis* 21 (12): 2509–21.

“Bind-Silane Working Solution.” 2011. *Cold Spring Harbor Protocols* 2011 (3): rec12413.

Bouvet, Marion, Annie Turkieh, Adelina E. Acosta-Martin, Maggy Chwastyniak, Olivia Beseme, Philippe Amouyel, and Florence Pinet. 2014. “Proteomic Profiling of Macrophages by 2D Electrophoresis.” *Journal of Visualized Experiments: JoVE*, no. 93 (November): e52219.

Burgi, D., and A. J. Smith. 2001. “Capillary Electrophoresis of Proteins and Peptides.” *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]* Chapter 10 (May): Unit 10.9.

Cai, Long, Nir Friedman, and X. Sunney Xie. 2006. “Stochastic Protein Expression in Individual Cells at the Single Molecule Level.” *Nature* 440 (7082): 358–62.

Chandradoss, Stanley D., Anna C. Haagsma, Young Kwang Lee, Jae-Ho Hwang, Jwa-Min Nam, and Chirlmin Joo. 2014. “Surface Passivation for Single-Molecule Protein Studies.” *Journal of Visualized Experiments: JoVE*, no. 86 (April). https://doi.org/10.3791/50549.

Chong, Yolanda T., Judice L. Y. Koh, Helena Friesen, Supipi Kaluarachchi Duffy, Kaluarachchi Duffy, Michael J. Cox, Alan Moses, Jason Moffat, Charles Boone, and Brenda J. Andrews. 2015. “Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis.” *Cell* 161 (6): 1413–24.

Cline, Gary W., and Samir B. Hanna. 1988. “Kinetics and Mechanisms of the Aminolysis of N-Hydroxysuccinimide Esters in Aqueous Buffers.” *The Journal of Organic Chemistry* 53 (15). American Chemical Society: 3583–86.

Coffman, Valerie C., and Jian-Qiu Wu. 2012. “Counting Protein Molecules Using Quantitative Fluorescence Microscopy.” *Trends in Biochemical Sciences* 37 (11): 499–506.

Coffman, Valerie C., Pengcheng Wu, Mark R. Parthun, and Jian-Qiu Wu. 2011. “CENP-A Exceeds Microtubule Attachment Sites in Centromere Clusters of Both Budding and Fission Yeast.” *The Journal of Cell Biology* 195 (4): 563–72.

Crick, F. H. 1958. “On Protein Synthesis.” *Symposia of the Society for Experimental Biology* 12: 138–63.

Crick, Francis. 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258): 561–63.

Darmanis, Spyros, Caroline Julie Gallant, Voichita Dana Marinescu, Mia Niklasson, Anna Segerman, Georgios Flamourakis, Simon Fredriksson, et al. 2016. “Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells.” *Cell Reports* 14 (2): 380–89.

Di Fiori, Nicolas, and Amit Meller. 2009. “Automated System for Single Molecule Fluorescence Measurements of Surface-Immobilized Biomolecules.” *Journal of Visualized Experiments: JoVE*, no. 33 (November). https://doi.org/10.3791/1542.

Drissi, Romain, Marie-Line Dubois, and François-Michel Boisvert. 2013. “Proteomics Methods for Subcellular Proteome Analysis.” *The FEBS Journal* 280 (22): 5626–34.

Ebenstein, Yuval, and Laurent A. Bentolila. 2010. “Single-Molecule Detection: Focusing on the Objective.” *Nature Nanotechnology* 5 (2): 99–100.

Eldar, Avigdor, and Michael B. Elowitz. 2010. “Functional Roles for Noise in Genetic Circuits.” *Nature* 467 (7312): 167–73.

Elia, Giuliano. 2010. “Protein Biotinylation.” In *Current Protocols in Protein Science*, 3.6.1–3.6.21.

Erlemann, Sarah, Annett Neuner, Linda Gombos, Romain Gibeaux, Claude Antony, and Elmar Schiebel. 2012. “An Extended γ-Tubulin Ring Functions as a Stable Platform in Microtubule Nucleation.” *The Journal of Cell Biology* 197 (1): 59–74.

Fabre, Bertrand, Dagmara Korona, Arnoud Groen, Jakob Vowinckel, Laurent Gatto, Michael J. Deery, Markus Ralser, Steven Russell, and Kathryn S. Lilley. 2016. “Analysis of Drosophila Melanogaster Proteome Dynamics during Embryonic Development by a Combination of Label-Free Proteomics Approaches.” *Proteomics* 16 (15-16): 2068–80.

Franken, Holger, Toby Mathieson, Dorothee Childs, Gavain M. A. Sweetman, Thilo Werner, Ina Tögel, Carola Doce, et al. 2015. “Thermal Proteome Profiling for Unbiased Identification of Direct and Indirect Drug Targets Using Multiplexed Quantitative Mass Spectrometry.” *Nature Protocols* 10 (10): 1567–93.

Funatsu, T., Y. Harada, M. Tokunaga, K. Saito, and T. Yanagida. 1995. “Imaging of Single Fluorescent Molecules and Individual ATP Turnovers by Single Myosin Molecules in Aqueous Solution.” *Nature* 374 (6522): 555–59.

Georgiou, Christos D., Konstantinos Grintzalis, George Zervoudakis, and Ioannis Papapostolou. 2008. “Mechanism of Coomassie Brilliant Blue G-250 Binding to Proteins: A Hydrophobic Assay for Nanogram Quantities of Proteins.” *Analytical and Bioanalytical Chemistry* 391 (1): 391–403.

Geyer, Philipp E., Nils A. Kulak, Garwin Pichler, Lesca M. Holdt, Daniel Teupser, and Matthias Mann. 2016. “Plasma Proteome Profiling to Assess Human Health and Disease.” *Cell Systems* 2 (3): 185–95.

Green, N. M. 1963. “AVIDIN. 1. THE USE OF (14-C)BIOTIN FOR KINETIC STUDIES AND FOR ASSAY.” *Biochemical Journal* 89 (December): 585–91.

Guldbrandsen, Astrid, Heidrun Vethe, Yehia Farag, Eystein Oveland, Hilde Garberg, Magnus Berle, Kjell-Morten Myhr, Jill A. Opsahl, Harald Barsnes, and Frode S. Berven. 2014. “In-Depth Characterization of the Cerebrospinal Fluid (CSF) Proteome Displayed through the CSF Proteome Resource (CSF-PR).” *Molecular & Cellular Proteomics: MCP* 13 (11): 3152–63.

Harper, J. Wade, and Eric J. Bennett. 2016. “Proteome Complexity and the Forces That Drive Proteome Imbalance.” *Nature* 537 (7620): 328–38.

Herzenberg, Leonard A., David Parks, Bita Sahaf, Omar Perez, Mario Roederer, and Leonore A. Herzenberg. 2002. “The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford.” *Clinical Chemistry* 48 (10): 1819–27.

Heukeshoven, Jochen, and Rudolf Dernick. 1992. “Native Horizontal Ultrathin Polyacrylamide Gel Electrophoresis of Proteins under Basic and Acidic Conditions.” *Electrophoresis* 13 (1): 654–59.

Hillborg, H., J. F. Ankner, U. W. Gedde, G. D. Smith, H. K. Yasuda, and K. Wikström. 2000. “Crosslinked Polydimethylsiloxane Exposed to Oxygen Plasma Studied by Neutron Reflectometry and Other Surface Specific Techniques.” *Polymer* 41 (18): 6851–63.

Holzmeister, Phil, Guillermo P. Acuna, Dina Grohmann, and Philip Tinnefeld. 2014. “Breaking the Concentration Limit of Optical Single-Molecule Detection.” *Chemical Society Reviews* 43 (4): 1014–28.

Howarth, Mark, Daniel J-F Chinnapen, Kimberly Gerrow, Pieter C. Dorrestein, Melanie R. Grandy, Neil L. Kelleher, Alaa El-Husseini, and Alice Y. Ting. 2006. “A Monovalent Streptavidin with a Single Femtomolar Biotin Binding Site.” *Nature Methods* 3 (4): 267–73.

“H.sapiens - Whole Organism (Integrated) in PaxDb.” n.d. Accessed July 12, 2017. http://pax-db.org/dataset/9606/29.

Huang, Bo, Hongkai Wu, Devaki Bhaya, Arthur Grossman, Sebastien Granier, Brian K. Kobilka, and Richard N. Zare. 2007. “Counting Low-Copy Number Proteins in a Single Cell.” *Science* 315 (5808): 81–84.

Hughes, Alex J., and Amy E. Herr. 2012. “Microfluidic Western Blotting.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21450–55.

Hughes, Alex J., Dawn P. Spelke, Zhuchen Xu, Chi-Chih Kang, David V. Schaffer, and Amy E. Herr. 2014. “Single-Cell Western Blotting.” *Nature Methods* 11 (7): 749–55.

Hughes, Laura D., Robert J. Rawle, and Steven G. Boxer. 2014. “Choose Your Label Wisely: Water-Soluble Fluorophores Often Interact with Lipid Bilayers.” *PloS One* 9 (2): e87649.

Irish, Jonathan M., Randi Hovland, Peter O. Krutzik, Omar D. Perez, Øystein Bruserud, Bjørn T. Gjertsen, and Garry P. Nolan. 2004. “Single Cell Profiling of Potentiated Phospho-Protein Networks in Cancer Cells.” *Cell* 118 (2): 217–28.

Jain, Saumya, Joshua R. Wheeler, Robert W. Walters, Anurag Agrawal, Anthony Barsic, and Roy Parker. 2016. “ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure.” *Cell* 164 (3): 487–98.

Kalluri, Raghu, and Robert A. Weinberg. 2009. “The Basics of Epithelial-Mesenchymal Transition.” *The Journal of Clinical Investigation* 119 (6): 1420–28.

Karger, Barry L., and András Guttman. 2009. “DNA Sequencing by CE.” *Electrophoresis* 30 Suppl 1 (June): S196–202.

Kennedy, Eamonn, Zhuxin Dong, Clare Tennant, and Gregory Timp. 2016. “Reading the Primary Structure of a Protein with 0.07 nm(3) Resolution Using a Subnanometre-Diameter Pore.” *Nature Nanotechnology* 11 (11): 968–76.

Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. 2014. “Bayesian Approach to Single-Cell Differential Expression Analysis.” *Nature Methods* 11 (7): 740–42.

Kim, Younggyu, Sam O. Ho, Natalie R. Gassman, You Korlann, Elizabeth V. Landorf, Frank R. Collart, and Shimon Weiss. 2008. “Efficient Site-Specific Labeling of Proteins via Cysteines.” *Bioconjugate Chemistry* 19 (3): 786–91.

Kolmogorov, Mikhail, Eamonn Kennedy, Zhuxin Dong, Gregory Timp, and Pavel A. Pevzner. 2017. “Single-Molecule Protein Identification by Sub-Nanopore Sensors.” *PLoS Computational Biology* 13 (5): e1005356.

Kumagai, Akiko, Anna Shevchenko, Andrej Shevchenko, and William G. Dunphy. 2010. “Treslin Collaborates with TopBP1 in Triggering the Initiation of DNA Replication.” *Cell* 140 (3): 349–59.

Lakowicz, Joseph R., ed. 2006. *Principles of Fluorescence Spectroscopy*. Boston, MA: Springer US.

Lamouille, Samy, Jian Xu, and Rik Derynck. 2014. “Molecular Mechanisms of Epithelial-Mesenchymal Transition.” *Nature Reviews. Molecular Cell Biology* 15 (3): 178–96.

Lawrie, Laura C., John E. Fothergill, and Graeme I. Murray. 2001. “Spot the Differences: Proteomics in Cancer Research.” *The Lancet Oncology* 2 (5): 270–77.

Leake, Mark C., Jennifer H. Chandler, George H. Wadhams, Fan Bai, Richard M. Berry, and Judith P. Armitage. 2006. “Stoichiometry and Turnover in Single, Functioning Membrane Protein Complexes.” *Nature* 443 (7109): 355–58.

Le Bihan, Marie-Catherine, Inigo Barrio-Hernandez, Tenna Pavia Mortensen, Jeanette Henningsen, Søren Skov Jensen, Anne Bigot, Blagoy Blagoev, Gillian Butler-Browne, and Irina Kratchmarova. 2015. “Cellular Proteome Dynamics during Differentiation of Human Primary Myoblasts.” *Journal of Proteome Research* 14 (8): 3348–61.

Lin, Yihan, and Michael B. Elowitz. 2016. “Central Dogma Goes Digital.” *Molecular Cell* 61 (6): 791–92.

Li, Qiang, and Stefan Seeger. 2006. “Label-Free Detection of Single Protein Molecules Using Deep UV Fluorescence Lifetime Microscopy.” *Analytical Chemistry* 78 (8): 2732–37.

Livnah, O., E. A. Bayer, M. Wilchek, and J. L. Sussman. 1993. “Three-Dimensional Structures of Avidin and the Avidin-Biotin Complex.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (11): 5076–80.

Lomenick, Brett, Rui Hao, Nao Jonai, Randall M. Chin, Mariam Aghajan, Sarah Warburton, Jianing Wang, et al. 2009. “Target Identification Using Drug Affinity Responsive Target Stability (DARTS).” *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21984–89.

Lo, Yu-Shiu, Neil D. Huefner, Winter S. Chan, Forrest Stevens, Joel M. Harris, and Thomas P. Beebe. 1999. “Specific Interactions between Biotin and Avidin Studied by Atomic Force Microscopy Using the Poisson Statistical Analysis Method.” *Langmuir: The ACS Journal of Surfaces and Colloids* 15 (4): 1373–82.

Marcon, Edyta, Harshika Jain, Anandi Bhattacharya, Hongbo Guo, Sadhna Phanse, Shuye Pu, Gregory Byram, et al. 2015. “Assessment of a Method to Characterize Antibody Selectivity and Specificity for Use in Immunoprecipitation.” *Nature Methods* 12 (8): 725–31.

Massey, V. 2000. “The Chemical and Biological Versatility of Riboflavin.” *Biochemical Society Transactions* 28 (4): 283–96.

McCarthy, B. J., and J. J. Holland. 1965. “Denatured DNA as a Direct Template for in Vitro Protein Synthesis.” *Proceedings of the National Academy of Sciences of the United States of America* 54 (3): 880–86.

Michalet, X., O. H. W. Siegmund, J. V. Vallerga, P. Jelinsky, J. E. Millaud, and S. Weiss. 2007. “Detectors for Single-Molecule Fluorescence Imaging and Spectroscopy.” *Journal of Modern Optics* 54 (2-3): 239.

Milo, Ron. 2013. “What Is the Total Number of Protein Molecules per Cell Volume? A Call to Rethink Some Published Values.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35 (12): 1050–55.

Moerner, W. E., and L. Kador. 1989. “Optical Detection and Spectroscopy of Single Molecules in a Solid.” *Physical Review Letters* 62 (21): 2535–38.

Mujumdar, R. B., L. A. Ernst, S. R. Mujumdar, C. J. Lewis, and A. S. Waggoner. 1993. “Cyanine Dye Labeling Reagents: Sulfoindocyanine Succinimidyl Esters.” *Bioconjugate Chemistry* 4 (2): 105–11.

Munkholm, Christiane, Don R. Parkinson, and David R. Walt. 1990. “Intramolecular Fluorescence Self-Quenching of Fluoresceinamine.” *Journal of the American Chemical Society* 112 (7). American Chemical Society: 2608–12.

Murthy, Krishna R., Renu Goel, Yashwanth Subbannayya, Harrys Kc Jacob, Praveen R. Murthy, Srikanth Srinivas Manda, Arun H. Patil, et al. 2014. “Proteomic Analysis of Human Vitreous Humor.” *Clinical Proteomics* 11 (1): 29.

Mutch, Sarah A., Bryant S. Fujimoto, Christopher L. Kuyper, Jason S. Kuo, Sandra M. Bajjalieh, and Daniel T. Chiu. 2007. “Deconvolving Single-Molecule Intensity Distributions for Quantitative Microscopy Measurements.” *Biophysical Journal* 92 (8): 2926–43.

Nanda, Jagpreet S., and Jon R. Lorsch. 2014. “Labeling a Protein with Fluorophores Using NHS Ester Derivitization.” *Methods in Enzymology* 536: 87–94.

Nelson, Edward M., Volker Kurz, Jiwook Shim, Winston Timp, and Gregory Timp. 2012. “Using a Nanopore for Single Molecule Detection and Single Cell Transfection.” *The Analyst* 137 (13): 3020–27.

Neuhoff, Volker, Norbert Arold, Dieter Taube, and Wolfgang Ehrhardt. 1988. “Improved Staining of Proteins in Polyacrylamide Gels Including Isoelectric Focusing Gels with Clear Background at Nanogram Sensitivity Using Coomassie Brilliant Blue G-250 and R-250.” *Electrophoresis* 9 (6). Wiley Subscription Services, Inc., A Wiley Company: 255–62.

Newman, John R. S., Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. 2006. “Single-Cell Proteomic Analysis of S. Cerevisiae Reveals the Architecture of Biological Noise.” *Nature* 441 (7095): 840–46.

Nie, Litong, Mingrui Zhu, Shengnan Sun, Linhui Zhai, Zhixiang Wu, Lili Qian, and Minjia Tan. 2016. “An Optimization of the LC-MS/MS Workflow for Deep Proteome Profiling on an Orbitrap Fusion.” *Analytical Methods* 8 (2): 425–34.

Nikolov, Miroslav, Carla Schmidt, and Henning Urlaub. 2012. “Quantitative Mass Spectrometry-Based Proteomics: An Overview.” *Methods in Molecular Biology* 893: 85–100.

Nino, Daniel, Nafiseh Rafiei, Yong Wang, Anton Zilman, and Joshua N. Milstein. 2017. “Molecular Counting with Localization Microscopy: A Bayesian Estimate Based on Fluorophore Statistics.” *Biophysical Journal* 112 (9): 1777–85.

Olson, Bradley J. S. C., and John Markwell. 2007. “Assays for Determination of Protein Concentration.” *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]* Chapter 3 (May): Unit 3.4.

Orrit, M., and J. Bernard. 1990. “Single Pentacene Molecules Detected by Fluorescence Excitation in a P-Terphenyl Crystal.” *Physical Review Letters* 65 (21): 2716–19.

Park, Yun-Jong, Jin Koh, Jin Teak Kwon, Yong-Seok Park, Lijun Yang, and Seunghee Cha. 2017. “Uncovering Stem Cell Differentiation Factors for Salivary Gland Regeneration by Quantitative Analysis of Differential Proteomes.” *PloS One* 12 (2). Public Library of Science: e0169677.

Peck, K., L. Stryer, A. N. Glazer, and R. A. Mathies. 1989. “Single-Molecule Fluorescence Detection: Autocorrelation Criterion and Experimental Realization with Phycoerythrin.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (11): 4087–91.

Pelegrine, D. H. G., and C. A. Gasparetto. 2005. “Whey Proteins Solubility as Function of Temperature and pH.” *LWT - Food Science and Technology* 38 (1): 77–80.

Perdigão, Nelson, Julian Heinrich, Christian Stolte, Kenneth S. Sabir, Michael J. Buckley, Bruce Tabor, Beth Signal, et al. 2015. “Unexpected Features of the Dark Proteome.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (52): 15898–903.

Petricoin, Emanuel F., Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, et al. 2002. “Use of Proteomic Patterns in Serum to Identify Ovarian Cancer.” *The Lancet* 359 (9306): 572–77.

Picotti, Paola, Bernd Bodenmiller, Lukas N. Mueller, Bruno Domon, and Ruedi Aebersold. 2009. “Full Dynamic Range Proteome Analysis of S. Cerevisiae by Targeted Proteomics.” *Cell* 138 (4): 795–806.

Poehling, Hans-Michael, and Volker Neuhoff. 1981. “Visualization of Proteins with a Silver ‘stain’: A Critical Analysis.” *Electrophoresis* 2 (3). Wiley Subscription Services, Inc., A Wiley Company: 141–47.

Pretzer, Elizabeth, and John E. Wiktorowicz. 2008. “Saturation Fluorescence Labeling of Proteins for Proteomic Analyses.” *Analytical Biochemistry* 374 (2): 250–62.

Raj, Arjun, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. 2006. “Stochastic mRNA Synthesis in Mammalian Cells.” *PLoS Biology* 4 (10): e309.

Ren, Kangning, Jianhua Zhou, and Hongkai Wu. 2013. “Materials for Microfluidic Chip Fabrication.” *Accounts of Chemical Research* 46 (11): 2396–2406.

Ridley, R. M. 2001. “What Would Thomas Henry Huxley Have Made of Prion Diseases?” *Methods in Molecular Medicine* 59: 1–16.

Rissin, David M., Cheuk W. Kan, Todd G. Campbell, Stuart C. Howes, David R. Fournier, Linan Song, Tomasz Piech, et al. 2010. “Single-Molecule Enzyme-Linked Immunosorbent Assay Detects Serum Proteins at Subfemtomolar Concentrations.” *Nature Biotechnology* 28 (6): 595–99.

Rissin, David M., and David R. Walt. 2006. “Digital Concentration Readout of Single Enzyme Molecules Using Femtoliter Arrays and Poisson Statistics.” *Nano Letters* 6 (3): 520–23.

Rodrigues, Anderson Messias, Paula H. Kubitschek-Barreira, Geisa Ferreira Fernandes, Sandro Rogério de Almeida, Leila M. Lopes-Bezerra, and Zoilo Pires de Camargo. 2015. “Immunoproteomic Analysis Reveals a Convergent Humoral Response Signature in the Sporothrix Schenckii Complex.” *Journal of Proteomics* 115 (February): 8–22.

Rose, G., A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus. 1985. “Hydrophobicity of Amino Acid Residues in Globular Proteins.” *Science* 229 (4716): 834–38.

Rose, Rebecca J., Eugen Damoc, Eduard Denisov, Alexander Makarov, and Albert J. R. Heck. 2012. “High-Sensitivity Orbitrap Mass Analysis of Intact Macromolecular Assemblies.” *Nature Methods* 9 (11): 1084–86.

Roy, Rahul, Sungchul Hohng, and Taekjip Ha. 2008. “A Practical Guide to Single-Molecule FRET.” *Nature Methods* 5 (6): 507–16.

Savitski, Mikhail M., Friedrich B. M. Reinhard, Holger Franken, Thilo Werner, Maria Fälth Savitski, Dirk Eberhard, Daniel Martinez Molina, et al. 2014. “Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome.” *Science* 346 (6205): 1255784.

Savitzky, Abraham, and M. J. E. Golay. 1964. “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.” *Analytical Chemistry* 36 (8): 1627–39.

Schubert, Stephanie M., Stephanie R. Walter, Mael Manesse, and David R. Walt. 2016. “Protein Counting in Single Cancer Cells.” *Analytical Chemistry* 88 (5): 2952–57.

Shadpour, Hamed, and Steven A. Soper. 2006. “Two-Dimensional Electrophoretic Separation of Proteins Using Poly(methyl Methacrylate) Microchips.” *Analytical Chemistry* 78 (11): 3519–27.

Shaw, Peter J., and David J. Rawlins. 1991. “The Point-Spread Function of a Confocal Microscope: Its Measurement and Use in Deconvolution of 3-D Data.” *Journal of Microscopy* 163 (2): 151–65.

Shechter, David, Holger L. Dormann, C. David Allis, and Sandra B. Hake. 2007. “Extraction, Purification and Analysis of Histones.” *Nature Protocols* 2 (6): 1445–57.

Sheng, Rongsheng, Fan Ni, and Therese M. Cotton. 1991. “Determination of Purine Bases by Reversed-Phase High-Performance Liquid Chromatography Using Real-Time Surface-Enhanced Raman Spectroscopy.” *Analytical Chemistry* 63 (5). American Chemical Society: 437–42.

Shi, Qihui, Lidong Qin, Wei Wei, Feng Geng, Rong Fan, Young Shik Shin, Deliang Guo, Leroy Hood, Paul S. Mischel, and James R. Heath. 2012. “Single-Cell Proteomic Chip for Profiling Intracellular Signaling Pathways in Single Tumor Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (2): 419–24.

Sidoli, Simone, Katarzyna Kulej, and Benjamin A. Garcia. 2017. “Why Proteomics Is Not the New Genomics and the Future of Mass Spectrometry in Cell Biology.” *The Journal of Cell Biology* 216 (1): 21–24.

Sigal, Alex, Ron Milo, Ariel Cohen, Naama Geva-Zatorsky, Yael Klein, Yuvalal Liron, Nitzan Rosenfeld, Tamar Danon, Natalie Perzov, and Uri Alon. 2006. “Variability and Memory of Protein Levels in Human Cells.” *Nature* 444 (7119): 643–46.

Snapp, Erik. 2005. “Design and Use of Fluorescent Fusion Proteins in Cell Biology.” *Current Protocols in Cell Biology / Editorial Board, Juan S. Bonifacino ... [et Al.]* Chapter 21 (July): Unit 21.4.

Söderberg, Ola, Mats Gullberg, Malin Jarvius, Karin Ridderstråle, Karl-Johan Leuchowius, Jonas Jarvius, Kenneth Wester, et al. 2006. “Direct Observation of Individual Endogenous Protein Complexes in Situ by Proximity Ligation.” *Nature Methods* 3 (12): 995–1000.

Sternberg. 1983. “Biomedical Image Processing.” *Computer* 16 (1): 22–34.

Strickland, Erin C., M. Ariel Geer, Duc T. Tran, Jagat Adhikari, Graham M. West, Patrick D. DeArmond, Ying Xu, and Michael C. Fitzgerald. 2013. “Thermodynamic Analysis of Protein-Ligand Binding Interactions in Complex Biological Mixtures Using the Stability of Proteins from Rates of Oxidation.” *Nature Protocols* 8 (1): 148–61.

Sund, H., K. Weber, and E. Molbert. 1967. “Dissoziation Der Rinderleber-Katalase in Ihre Untereinheiten.” *European Journal of Biochemistry / FEBS* 1 (4): 400–410.

Taniguchi, Yuichi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. 2010. “Quantifying E. Coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells.” *Science* 329 (5991): 533–38.

Tegelström, Håkan, and Per-Ivan Wyöni. 1986. “Silanization of Supporting Glass Plates Avoiding Fixation of Polyacrylamide Gels to Glass Cover Plates.” *Electrophoresis* 7 (2): 99–99.

Temin, Howard M., and Satoshi Mizutani. 1970. “Viral RNA-Dependent DNA Polymerase: RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus.” *Nature* 226 (5252): 1211–13.

Tentori, Augusto M., Kevin A. Yamauchi, and Amy E. Herr. 2016. “Detection of Isoforms Differing by a Single Charge Unit in Individual Cells.” *Angewandte Chemie* 55 (40): 12431–35.

Thompson, Michael A., Matthew D. Lew, and W. E. Moerner. 2012. “Extending Microscopic Resolution with Single-Molecule Imaging and Active Control.” *Annual Review of Biophysics* 41: 321–42.

Tonge, R., J. Shaw, B. Middleton, R. Rowlinson, S. Rayner, J. Young, F. Pognan, E. Hawkins, I. Currie, and M. Davison. 2001. “Validation and Development of Fluorescence Two-Dimensional Differential Gel Electrophoresis Proteomics Technology.” *Proteomics* 1 (3): 377–96.

Ulbrich, Maximilian H., and Ehud Y. Isacoff. 2007. “Subunit Counting in Membrane-Bound Proteins.” *Nature Methods* 4 (4): 319–21.

Vaiana, Andrea C., Hannes Neuweiler, Andreas Schulz, Jürgen Wolfrum, Markus Sauer, and Jeremy C. Smith. 2003. “Fluorescence Quenching of Dyes by Tryptophan: Interactions at Atomic Detail from Combination of Experiment and Computer Simulation.” *Journal of the American Chemical Society* 125 (47): 14564–72.

Verdaasdonk, Jolien Suzanne, Josh Lawrimore, and Kerry Bloom. 2014. “Determining Absolute Protein Numbers by Quantitative Fluorescence Microscopy.” *Methods in Cell Biology* 123: 347–65.

Von Bergen, Martin, Franziska Dautel, Stefan Kalkhof, Saskia Trump, Irina Lehmann, and Andreas Beyer. 2011. “Large-Scale 2-D DIGE Studies - Guidelines to Overcome Pitfalls and Challenges along the Experimental Procedure.” *Journal of Integrated OMICS* 1 (1). https://doi.org/10.5584/jiomics.v1i1.50.

Walter, Nils G., Cheng-Yen Huang, Anthony J. Manzo, and Mohamed A. Sobhy. 2008. “Do-It-Yourself Guide: How to Use the Modern Single-Molecule Toolkit.” *Nature Methods* 5 (6): 475–89.

Willison, Keith R., and David R. Klug. 2013. “Quantitative Single Cell and Single Molecule Proteomics for Clinical Studies.” *Current Opinion in Biotechnology* 24 (4): 745–51.

Wilson, Steven Ray, Tore Vehus, Henriette Sjaanes Berg, and Elsa Lundanes. 2015. “Nano-LC in Proteomics: Recent Advances and Approaches.” *Bioanalysis* 7 (14): 1799–1815.

Wu, Jian-Qiu, and Thomas D. Pollard. 2005. “Counting Cytokinesis Proteins Globally and Locally in Fission Yeast.” *Science* 310 (5746): 310–14.

Xu, Ke, Hazen P. Babcock, and Xiaowei Zhuang. 2012. “Dual-Objective STORM Reveals Three-Dimensional Filament Organization in the Actin Cytoskeleton.” *Nature Methods* 9 (2): 185–88.

Yang, Pengyi, Ellis Patrick, Shi-Xiong Tan, Daniel J. Fazakerley, James Burchfield, Christopher Gribben, Matthew J. Prior, David E. James, and Yee Hwa Yang. 2014. “Direction Pathway Analysis of Large-Scale Proteomics Data Reveals Novel Features of the Insulin Action Pathway.” *Bioinformatics* 30 (6): 808–14.

Yang, Wen-Chu, Hamid Mirzaei, Xiuping Liu, and Fred E. Regnier. 2006. “Enhancement of Amino Acid Detection and Quantification by Electrospray Ionization Mass Spectrometry.” *Analytical Chemistry* 78 (13): 4702–8.

Yates, John R., Cristian I. Ruse, and Aleksey Nakorchevsky. 2009. “Proteomics by Mass Spectrometry: Approaches, Advances, and Applications.” *Annual Review of Biomedical Engineering* 11: 49–79.

Yeoh, Sharon, Brian Pope, Hans G. Mannherz, and Alan Weeds. 2002. “Determining the Differences in Actin Binding by Human ADF and Cofilin.” *Journal of Molecular Biology* 315 (4): 911–25.

Yu, Ji, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X. Sunney Xie. 2006. “Probing Gene Expression in Live Cells, One Protein Molecule at a Time.” *Science* 311 (5767). American Association for the Advancement of Science: 1600–1603.

Zanacchi, Francesca Cella, Carlo Manzo, Angel S. Alvarez, Nathan D. Derr, Maria F. Garcia-Parajo, and Melike Lakadamyali. 2017. “A DNA Origami Platform for Quantifying Protein Copy Number in Super-Resolution.” *Nature Methods* 14 (8): 789–92.

Zanetti-Domingues, Laura C., Christopher J. Tynan, Daniel J. Rolfe, David T. Clarke, and Marisa Martin-Fernandez. 2013. “Hydrophobic Fluorescent Probes Introduce Artifacts into Single Molecule Tracking Experiments due to Non-Specific Binding.” *PloS One* 8 (9): e74200.

Zhang, Rongfei, Haiyu Yuan, Shujing Wang, Qi Ouyang, Yong Chen, Nan Hao, and Chunxiong Luo. 2017. “High-Throughput Single-Cell Analysis for the Proteomic

Dynamics Study of the Yeast Osmotic Stress Response." *Scientific Reports* 7 (February): 42200.

Zheng, Qinsi, Manuel F. Juette, Steffen Jockusch, Michael R. Wasserman, Zhou Zhou, Roger B. Altman, and Scott C. Blanchard. 2014. "Ultra-Stable Organic Fluorophores for Single-Molecule Research." *Chemical Society Reviews* 43 (4): 1044–56.

Annex

Annex I : Extending single molecule imaging to proteome analysis by quantitation of fluorescent labeling homogeneity in complex protein samples

Authors: Simon LECLERC[1,2], Youri ARNTZ[2], Yuichi TANIGUCHI[1,3*]

[1] Laboratory for Single Cell Gene Dynamics, Quantitative Biology Center, RIKEN

Address: 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan

[2] BioMatériaux et BioIngénierie, INSERM UMR 1121

Address: Etage 7, 11 Rue Humann, 67000 Strasbourg, France

[3] PRESTO, Japan Science and Technology Agency

Address: 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan

* Author to whom correspondence should be addressed; E-Mail: taniguchi@riken.jp;

Tel.: +81-6-6155-0114; Fax: +81-6-6155-0112.

Abstract

Fluorescence-based electrophoresis has been widely used for proteome analysis, in which every protein species in cells are labeled with a fluorescent dye, separated by electric migration and are quantified using fluorescence detection. The ultimate limit of sensitivity for this approach could be reached by single-molecule fluorescence imaging and counting individual proteins, requiring exhaustive fluorescent labeling of proteins across molecular populations and species. However, it remains unclear how homogeneous is the fluorescence labeling of individual protein molecules of each species across the proteome. To address this question, we developed a method to measure the labeling homogeneity based on a single-molecule fluorescence counting assay. Our results reveal that the proportion of proteins labeled with at least one dye, called labeling occupancy (*LO*), was 35% for BSA fluorescently labeled using existing protocols. We then found that the *LO* could be improved to 82% under high pH and surfactant-rich conditions. Furthermore, when a proteome sample from a human cell lysate was analyzed, the total *LO* was 71%, whereby the values varied between 50 and 90% for low and high molecular weight proteome fractions, respectively. The results prove

that single molecule protein counting across a proteome is a practical and useful approach. Overall, our method using the *LO* parameter provides a system for estimating protein amounts from single molecule counting assays.

Introduction

Currently, there is a large and growing need for quantitative profiling of proteins in complex mixtures. Protein detection and visualization based on the use of fluorescent compounds has the advantages of high sensitivity, possibility for quantitation, and compatibility with mass spectrometry. Indeed, fluorescence-based techniques have been widely applied for sensitive proteome analysis in the cell.[1,2] The conventional separation methods for such analyses are polyacrylamide gel electrophoresis (PAGE), capillary electrophoresis, and 2D gel electrophoresis. Overall, these methods rely on the detection of every protein species composing the proteome separated by electrophoresis, and therefore require nonspecific labeling of all proteins. One strategy for such global protein labeling is to use dyes that can be noncovalently bound to proteins by electrostatic and hydrophobic interactions, such as Coomassie Blue and Sypro-Ruby.[3,4] An alternative strategy is to use dyes that are able to bind covalently through specific groups such as N-hydroxysuccinimide (NHS, or succinimidyl) ester, which targets primary amines, or maleimide, which binds thiols.[5,6] While the covalent labeling method allows higher sensitivity and specificity in protein detection and minimizes background noise, the proteins need to be labeled prior to separation on a gel.

The ultimate limit in sensitivity of this analysis is defined by detecting and counting single molecules. In the bioimaging field, such single molecule detection has been achieved by measuring fluorescently-labeled proteins under laser-based specific illumination and high-sensitivity fluorescence imaging.[7–12] Such single molecule detection has often been applied to individual species of proteins to characterize their *in vitro* and *in vivo* dynamics in biological events,[13–19] such as stepwise molecular motor movements [7,20–22] and stochastic gene expression events in single cells.[23,24] Meanwhile, we expect that the single molecule imaging sensitivity can be extended to the whole proteome by using nonspecific labeling with fluorescent dyes. However, it remains unclear how uniformly or homogeneously different proteins, as well as each species of individual protein molecules among a population, can be nonspecifically labeled with fluorescent dyes (Fig. 1), due to varying composition of amino acids and biochemical properties such as solubility and electrical charges,[25] as well as the

statistical probability of labeling of identical species. Conventionally, coupling efficiency,[6] has been described as the molar ratio of fluorophores to proteins in bulk solution, but this estimate is only accurate under the assumption that all fluorophores and proteins are randomly and uniformly bound. In contrast, in this work we propose a new technique to directly evaluate the labeling homogeneity across the proteome by counting the labeled protein proportion at the single molecule level. Our results show that protein molecules in a human cell proteome can be labeled with a high efficiency of up to 90%, which supports the possibility of using nonspecific labeling with fluorescent dyes and single molecule counting in proteome analysis.

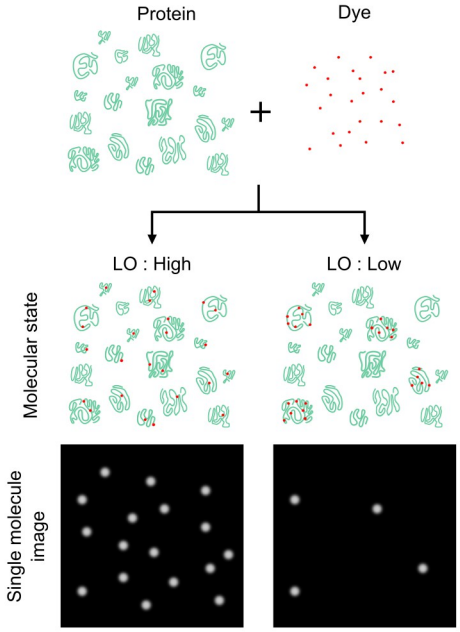


Figure 1. Effect of proteome labeling homogeneity on protein number counting. The protein count number is highly dependent on how strongly and homogeneously protein molecules are labeled, rather than the number ratio of proteins to dyes. The homogeneity can be scored using a parameter termed labeling occupancy (*LO*), which defines the probability of labeled protein molecules against total protein molecules. This value provides the efficiency of protein counting; i.e. higher *LO* yields higher count numbers (left) and vice versa (right). While the *LO* value of 100% is ideal, *LO* of < 100% can provide an attenuation factor for estimating absolute protein numbers.

Results

Assay for evaluation of labeling homogeneity. First, we developed an assay to evaluate labeling homogeneity of a fluorescently labeled protein sample at the single molecule level (Fig. 2A). As such, we evaluated the homogeneity by imaging a microscope coverslip that binds a known density of the sample protein molecules, and by characterizing the number and intensities of fluorescence spots that are observed on the same coverslip. In this case, homogeneously-distributed labeling should provide a higher number of fluorescence spots of constant intensity, whereas a heterogeneously labeled proteome would provide fewer spots of more variable intensity, whereby some proteins are unlabeled and others are bound to

multiple fluorescent dyes. By comparing the number of spots between the measured sample and the 100%-labeled, homogeneous control sample, we can obtain the percentage of proteins yielding fluorescence spots, which we call labeling occupancy (*LO*). The microscope coverslip was prepared by plasma-mediated oxidation followed by coating with a fixed density of avidin, and by binding the sample protein molecules, which have been denatured and biotinylated, to the immobilized avidin. We found that a fixed number of the protein sample molecules can be bound to the coverslip when the amount of the applied protein is greater than 0.76 fmol (Supplementary Fig. S1). Imaging was performed using a wide-field single molecule fluorescence microscope over a large area of the coverslip ($\simeq 4$ mm^2) to obtain better statistics. We also confirmed that few fluorescent spots were observed when performing the experiment without avidin coating (Fig. 2B, negative control), probably due to a low-affinity interaction of labeled proteins with the coverslip.

To test this assay system, we analyzed labeling homogeneity of BSA samples that were fluorescently labeled using the succinimidyl ester of Cy3 under a standard protocol (Fig. 2B). For the subsequent analysis, BSA samples were prepared by mixing different ratios of labeled and unlabeled protein, ranging from 1:3 to 1:0, respectively. In a bulk fluorometer experiment, the ratio of the dye to the protein in the sample solution (coupling efficiency, *CE*), which is the conventional parameter for characterizing labeling efficiency, was measured to be 0.47 for the labeled BSA sample before mixing with the unlabeled protein. Based on this, *LO* of the labeled BSA was predicted to be 38% assuming that the labeling event follows the Poisson law (see Materials & Methods), whereby it can also be assumed that such *LO* estimated from *CE* (LO_{CE}) decreases proportionally with a decrease of the proportion of labeled BSA in the mixed samples. In contrast, the actual *LO* measured by our method (LO_{Obs}) for the labeled BSA before mixing was smaller (33 ± 1.8%) than LO_{CE} (Fig. 2C), suggesting that the molecular labeling event is more biased and heterogeneous than the predicted Poisson process. In addition, LO_{Obs} was found to be proportional to the mixing ratio and to LO_{CE} (Fig. 2C). Furthermore, the distribution of fluorescence intensities was multi-modal among the fluorescence spots (Fig. 2B), suggesting that one protein could bind multiple dyes. We confirmed that the smallest peak is generated by a single fluorescent dye by observing one-step photobleaching under continuous laser irradiation (Supplementary Fig. S2). Therefore, the number of dyes binding to proteins within the fluorescence spots (n_{dye}) could be obtained from the fluorescence intensity of the spots. The average n_{dye} among all the

spots (n_{dye}) were found to be mostly constant at any mixing ratio (Fig. 2D), which confirms that dye labeling is stable after protein dilution and no dye migration occurs between different protein molecules. We also found that the probability density function for labeled BSA exhibited a higher probability of proteins binding either zero or multiple dyes than expected if the labeling were a Poisson process (Supplementary Fig. S3). This greater than Poisson labeling heterogeneity can be explained by the cooperativity of labeling, whereby binding of one dye to a protein induces binding of further dye molecules to the same protein. Another explanation is that some protein molecules in the sample have reduced reactivity toward the label due to incomplete denaturation, solubility problems, or inefficient mixing during the reaction.

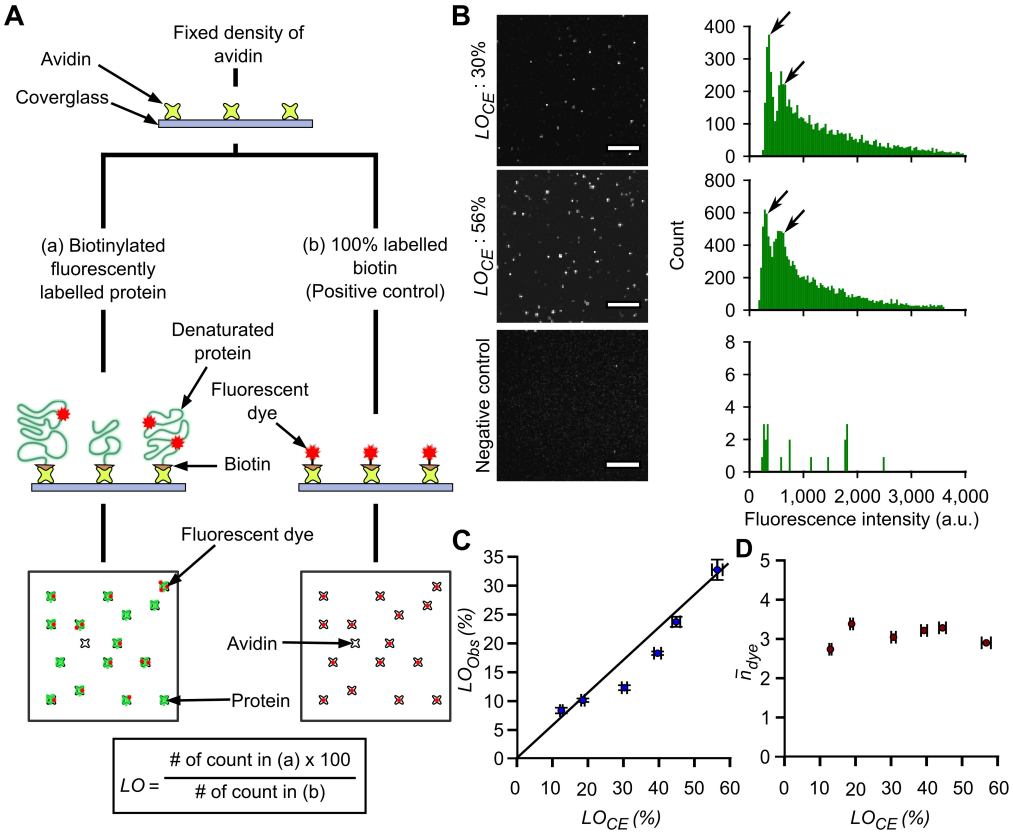


Figure 2. Evaluation of labeling homogeneity. (A) Assay workflow. First, coverslips (a) and (b) are treated with avidin to achieve a fixed density. Second, fluorescently-labeled sample proteins were biotinylated and attached to the coverslip (a) via the avidin-biotin interaction. In parallel, 100% fluorescently-labeled purified biotin is immobilized on the coverslip (b). Third, the coverslips are imaged by single-molecule fluorescence microscopy to obtain the

number and brightness distribution of fluorescence spots. Finally, the homogeneity parameter, *LO*, is calculated from the ratio of spot numbers in (a) to that in (b). (B) Raw data. Image data (left) and histograms of fluorescence intensities of spots (right) obtained from mixtures of fluorescently-labeled and unlabeled purified BSA at different ratios (top and second from top) are shown. Additionally, data obtained from 100% labeled fluorescent biotin bound to the avidin coating (third from top, positive control), and from labeled BSA loaded and washed from the coverslips without avidin coating (bottom, negative control) are shown. Peaks in the histograms represent conjugation of different discrete numbers of biotinylated fluorescent dyes. The scale bar is 10 μm. (C) Comparison of the estimated *LO* (LO_{CE}) and the observed *LO* (LO_{Obs}) values. LO_{CE} was obtained from the ratio of the dye to the protein, measured by a bulk fluorometer analysis, and assuming a random, Poisson labeling process. The black line shows a linear regression (R_2 = 0.95), suggesting that LO_{Obs} is proportional to LO_{CE}. Data are expressed as mean ± s.e. 44, 45, 50, 41, 39 and 38 images were analyzed for 8, 12, 20, 26, 30 and 37% values of LO_{CE} respectively. (D) Comparison between LO_{CE} and the average number of dyes binding to proteins within the fluorescence spots (n_{dye}), which were estimated from the fluorescence intensity histograms.

Influence of different labeling protocols on labeling homogeneity. We next tried to improve the *LO* of BSA samples by testing different labeling conditions. These included combinations of increased DTT (1 to 10 mM), SDS (0.1 to 1%) and Tween 20 (0.1 to 1%) concentrations compared to conventional protocols,[26–32] which are known to enhance protein reduction, denaturation, and solubilization, respectively (see Table 1). We also tested the influence of addition of 3-[(3-cholamidopropyl)dimethylammonio]propanesulfonate (CHAPS), a zwitterionic detergent that enhances solubility of proteins and thus their accessibility to the dye.[33,34] These new conditions were set to a higher pH (from pH 7.4 to 12) to improve the affinity of the reactive amine groups, those of lysine residues and/or the N-terminal of the protein, for the succinimidyl ester moiety in the dye. Similar to the original reaction conditions (Condition A in Fig. 3A and 3B), LO_{CE} values obtained under the newly tested conditions (Conditions B to J) were found to be around 40%. In contrast, LO_{Obs} values increased 1.4 times under these new conditions compared to that of condition A (29.5%) (Fig. 3B), indicating that the labeling homogeneity can be improved by the labeling conditions. Addition of our four tested reagents (DTT, SDS, Tween 20 or CHAPS) provided positive

effects on LO_{Obs}, which reached 82% when adding all the four additives together. Consistent with this, histograms of fluorescence intensities of each spot were less spread and less skewed to smaller values under the new conditions, suggesting that fewer and constant numbers of dyes tended to bind to the same single proteins (Fig. 3A). Also, under the new conditions, LO_{Obs} values increased compared to those of LO_{CE} (Fig. 3B), and the probability density functions of numbers of dyes binding to proteins exhibited smaller or no peak at the zero molecules data point (Supplementary Fig. S4), suggesting the all-or-none dye binding outcome was reduced, but instead a constant-number of dyes per protein type of labeling dominated. We suggest that such constant-number binding resulted from the exposure of the reactive labeling sites in proteins, which are then occupied by the dye molecules.

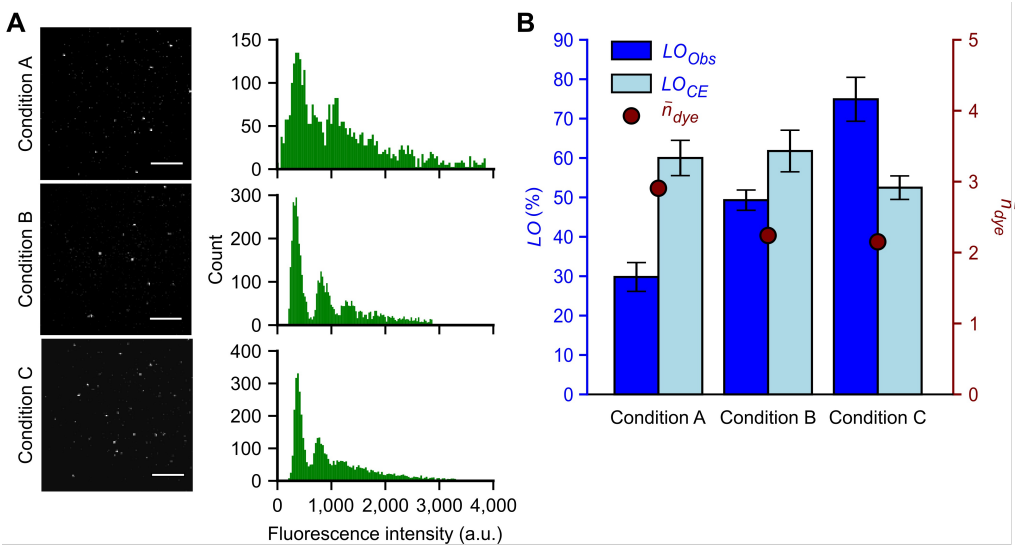


Figure 3. Labeling homogeneity of BSA under different labeling conditions. (A) Imaging data (left) and spot fluorescence intensity histograms (right). The relevant conditions are described in Table 1. Condition A is the original condition before the optimization, and Condition J is the most optimized condition consisting of higher concentrations of DTT, SDS, Tween 20 and CHAPS. The scale bar is 10 µm. (B) LO_{Obs} (blue), LO_{CE} (light blue) and n_{dye} (dark red) under different labeling conditions. Data are expressed as mean ± s.e. 34, 60, 49, 96, 116, 81, 53, 71, 94 and 114 images were analyzed for Condition A, B, C, D, E, F, G, H, I and J, respectively.

Condition	Buffer	DTT	SDS	Tween 20	CHAPS
A	1x PBS pH 7.4	1 mM	0.1 %	0.1 %	0 %
B	50 mM borate pH 12	1 mM	0.1 %	0.1 %	0 %
C	50 mM borate pH 12	1 mM	0.1 %	1 %	0 %
D	50 mM borate pH 12	1 mM	1 %	0.1 %	0 %
E	50 mM borate pH 12	1 mM	1 %	1 %	0 %
F	50 mM borate pH 12	10 mM	0.1 %	0.1 %	0 %
G	50 mM borate pH 12	10 mM	0.1 %	1 %	0 %
H	50 mM borate pH 12	10 mM	1 %	0.1 %	0 %
I	50 mM borate pH 12	10 mM	1 %	1 %	0 %
J	50 mM borate pH 12	10 mM	1 %	1 %	2 %

Table 1. Labeling buffer conditions tested.

Measuring labeling homogeneity of proteome samples. Following the optimization of the labeling protocol, we analyzed labeling homogeneity of a proteome sample that was obtained from a lysate of HeLa cells. As a proteome sample contains many protein species with widely differing molecular weights, we measured fractions of the proteome samples at different specific (16, 23, 55 and 120 kDa) or ranges of (19-25, 25-32, 32-42, 42-54, 54-69, 69-97, 97-136 and 136-226 kDa) molecular weights, which were obtained by the separation with SDS-PAGE (Fig. 4A and Supplementary Fig. S5). The specific molecular weights we focused on correspond to major bands in SDS-PAGE analysis for the cell lysate, whereas the molecular weight ranges correspond to sequential fractions covering most of the proteome. In this assay, we observed higher *LO* as molecular weights increased, ranging from 50% at 16 kDa to 90% at 120 kDa, and an intermediate *LO* for the whole proteome sample without separation (72%) (Figs. 4B and 4C). Correspondingly, the spot fluorescence intensities tend to be higher as molecular weights increase (Fig. 4C). These increasing tendencies are considered to be due to the dye labeling frequency, which depends on the number of lysine residues in a protein molecule. Consistent with this, probability density functions of numbers of dyes binding to proteins showed no strong peaks at zero molecules data point except for the 16, 20-25 and 25-33 kDa fractions of the proteome (Supplementary Fig. S6), suggesting that the higher molecular weight proteins have enough numbers of reactive lysine residues to retain a stable number of dye molecules.

Then, we considered the average numbers of lysine residues across protein species dominating each molecular weight fraction, which we termed n_{lys}. We estimated n_{lys} by

extracting a list of protein species falling into each molecular weight range from uniProt,[35] and by calculating an average number of lysine residues weighted using protein abundance stored in the paxdb4 database (Supplementary Table S1).[36] In the specific molecular weight fractions, n_{lys} values were estimated to be 8.26, 44.07, 34.39, 69.63 for 16, 23, 55 and 120 kDa fractions, respectively. Note that the estimated numbers are not simply proportional to the molecular weights, because of the protein species that have the highest abundance in the fractions. For instance, the high n_{lys} value at 23 kDa is due to the large number of lysine residues of protein histone H1.5 (66 in total). When we plotted *LO* values as a function of n_{lys} and fitted them with an exponential curve, we obtained the relationship: $LO = 100 - 71.5 \times exp(-0.039 \times n_{lys})$ (Fig. 4D). This relationship indicates that the *LO* is proportional to n_{lys} at a small n_{lys}, but reaches a plateau of 100% at large n_{lys} values, supporting the fact that the succinimidyl ester labeling reaction occurs predominantly at lysine residues in a protein. In addition, the equation indicates that there is a positive offset of the *LO* at 29% regardless of the n_{lys} value, which is consistent with the fact that the succinimidyl ester reaction also occurs at the NH_2 terminal of a protein. We also analyzed the relationship between n_{dye} and n_{lys}, which is given by: $n_{dye} = 3.3 - 2.6 \times exp(-0.08 \times n_{lys})$ (Fig. 4E). This relationship describes that n_{dye} is proportional to n_{lys} at small n_{lys}, but reaches a plateau of 3.3 at large n_{lys}, which would be caused by exclusion effects of dyes binding to the same protein, or quenching effects between neighboring dyes on the same protein.[37] The n_{dye} also has a positive offset around 0.75, supporting the reaction at the NH_2 terminal end of proteins. These derived equations can provide an estimate of the *LO* and single molecule brightness for any protein species in the proteome, which will be useful for proteome quantification by single molecule counting.

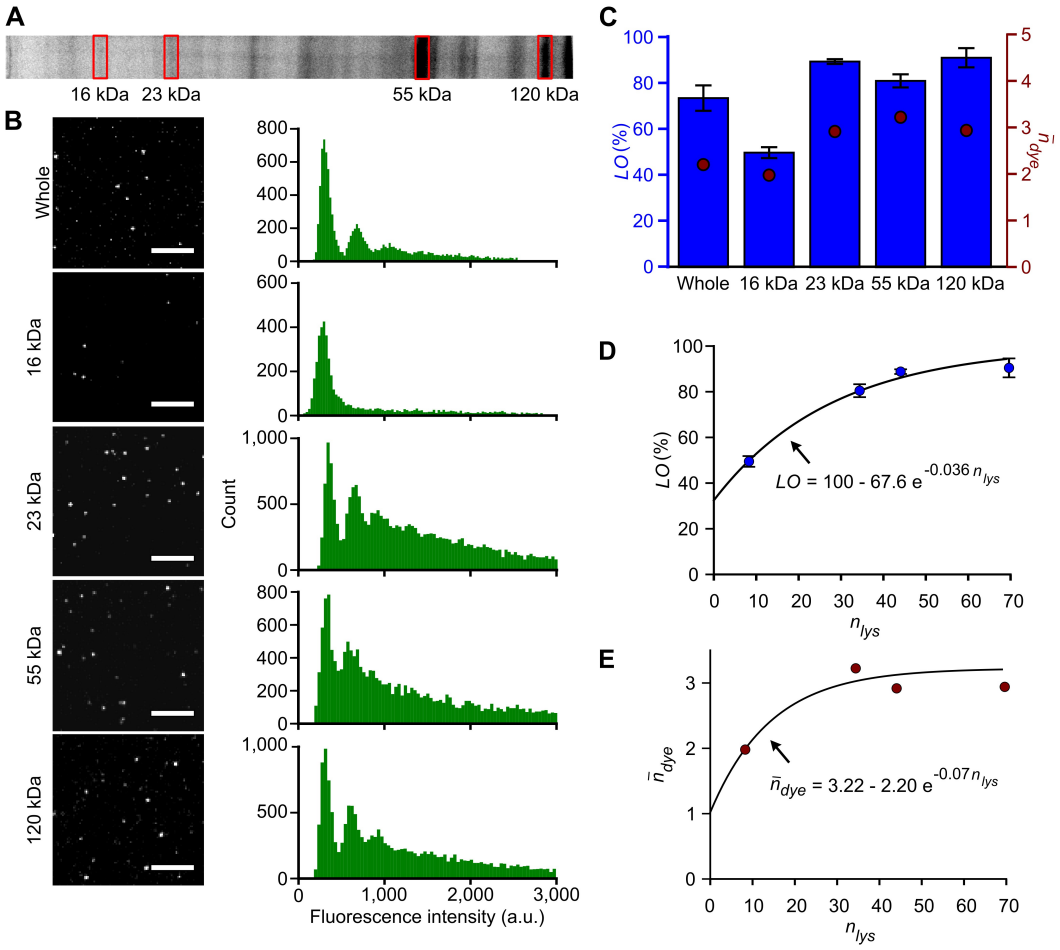


Figure 4. Labeling homogeneity of proteome samples. (A) Separation of the proteome sample from human cancer cells with SDS-PAGE. The red boxes represent the regions that are extracted for labeling homogeneity analysis. Full-length gel images are shown in Supplementary Figure S7. (B) Fluorescence spot images (left) and histograms of spot intensities (right) obtained from different molecular weight proteome fractions. The scale bar is 10 μm. (C) *LO* (blue) and n_{dye} (dark red) obtained from different molecular weight fractions. Data are expressed as mean ± s.e. 98, 46, 27, 77, 44, 39, 62, 101, 65, 62, 82, 61 and 70 images were analyzed for the whole cell lysate, 16, 23, 55, 120, 19-25, 25-32, 32-42, 42-54, 54-69, 69-97, 97-136 and 136-226 kDa fractions, respectively. (D) *LO* as a function of estimated number of lysine (n_{lys}). The line is a fitting curve to the equation: $LO = 100 - A \times exp(-B \times n_{lys})$, where *A* and *B* are the fitting parameters. Data are expressed as mean ± s.e. (E) n_{dye} as a function of n_{lys}. The line is a fitting curve to the equation: $n_{dye} = A - B \times \exp(-C \times n_{lys})$, where *A*, *B* and *C* are the fitting parameters. Data are expressed as mean ± s.e.

Discussion

In this work, we established an imaging assay for measuring the homogeneity of covalent fluorescent labeling of protein or proteome samples at the single-molecule level, and we introduced a new term to describe the labeling homogeneity, *LO*. Moreover, we were able to achieve a high *LO* (50-90%) by optimizing the labeling conditions, which included a high pH and enhanced solubility. Our results confirm the possibility of application of single-molecule counting to proteome analysis. Our method can also provide attenuation factors to estimate the actual protein numbers from the labeled, countable protein numbers, in the form of the *LO*. Furthermore, our derived equations can provide an estimation of *LO* depending on protein species even without experiments.

While it might be ideal to use monovalent avidin [38] to bind proteins on the coverslip surface as part of the assay, we argue that standard tetrameric avidin can still provide reasonable values. When we studied 100%-labeled biotin using the standard avidin surface (Fig. 2B, positive control), we observed that the majority (55%) of avidin molecules on the coverslip bound only one or two equivalents of fluorescently-labeled biotin, although a small fraction of avidin bound up to 4 molecules, consistent with its tetrameric nature.[39] Considering that biotin is small, larger protein molecules should bind to avidin in fewer numbers due to steric effects. Accordingly, we observed that the number of fluorescent dyes bound to avidin was almost constant at all mixing ratios of labeled and unlabeled BSA (Fig. 2D), suggesting that only one protein was able to bind to each avidin.

Our results also provide an important indication that labeling homogeneity is significantly affected by the labeling reaction conditions, and that this reaction is not a simple Poisson process. We suggest mechanisms that make labeling more heterogeneous or homogeneous (Fig. 5). Heterogeneity is considered to be caused by the so-called all-or-none mechanism, brought about by processes such as cooperative binding of dyes to proteins or the existence of a fraction of non-reactive proteins due to incomplete solubilization or lowered affinity for the dye. In contrast, homogeneity can be achieved by the complete binding of reactive lysine residues within a protein to the dye. The labeling is however never observed to be 100% complete, and there are several reasons for this. Firstly, we reason that changes in protein hydrophobicity can explain reaching the steady state of binding, whereby changes in

hydrophobicity of proteins upon binding to dyes can inhibit further binding to the label.[40] Another possibility for reaching an apparent maximum dye binding is the intramolecular dye quenching that occurs when many dyes exist in the same protein molecule.[41,42] These mechanisms highlight that single-molecule phenomena can deviate significantly from simple theoretical models consisting of ensemble-averaged bulk experiments, suggesting the necessity for single-molecule measurements.

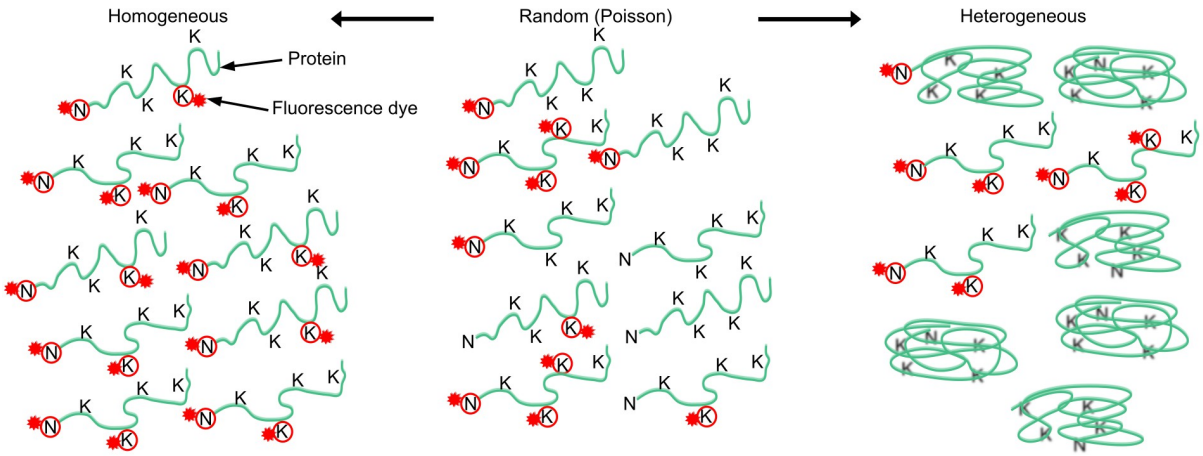


Figure 5. An explanatory model of labeling homogeneity and heterogeneity. The labeling homogeneity can be described as the complete binding of reactive lysine residues in a protein to the fluorescent label. Meanwhile, the heterogeneity can be rationalized by the presence of a fraction of proteins that are non-reactive due to incomplete solubilization or reduced affinity. The letters 'K' and 'N' denote reactive lysine residues and the NH_2 terminals, respectively.

We expect that our single molecule proteome-level analysis can be extended to further in-depth analyses in straightforward ways. One possibility may be to perform a native PAGE analysis to determine the molecular stoichiometry for complexes composed of multiple proteins in the cell. Another direction is to combine the single cell isolation and single molecule visualization in one microchip to allow single cell proteome analyses.[43–45] Overall, our methods and results will open the door for single molecule sensitivity-proteomics that may be ultimately used for precise cell properties determination and disease diagnosis.

Experimental methods

Labeling of purified proteins. Purified BSA (A9547, Sigma-Aldrich) was dissolved in 1,000 µl of distilled water to a concentration of 1 mg/ml. 1 µl of this BSA solution was denatured in the buffer indicated in Table 1 (Tween 20 (Sigma-Aldrich), SDS (Wako, Japan), DTT (Nacalai-tesque, Japan), borate buffer (Nacalai-tesque), CHAPS (Dojindo, Japan)). The denatured BSA was labeled by adding 1 µl of 2 µg/ml freshly diluted Cy3 NHS-ester dye (PA13101, GE Healthcare) in DMF (Nacalai-tesque) then by incubating for 15 min in the dark at room temperature with gentle agitation. In this protocol, the molar concentration ratio of the dyes to the proteins is 20, where the measured *LO* value is confirmed to be saturated (Supplementary Fig. S8).

Labeling of the proteome molecules from cell culture. HeLa cell line was grown at 37 ^{o}C under 5% CO_2 in Dulbecco's modified Eagle's medium (Sigma-Aldrich) containing 10% fetal bovine serum (Gibco) and 100 U/mL penicillin, 100 µg/mL streptomycin, and 0.25 µg/mL Amphotericin B (Gibco). Exponentially growing cells were collected by centrifugation after treatment with 0.1% trypsin (Gibco), washed with 1x PBS (pH 7.4), and resuspended in 1x PBS (pH 7.4) at 10_6 cells/ml. 1 µl of the cultured cells, corresponding to 1000 cells, were lysed and denatured by mixing with 100 µl of buffer condition J. Protein contents in the lysed cells, corresponding to 75 ng of protein, were labeled by mixing with 1 µl of 2 µg/ml Cy3 NHS-ester dye and by incubating for 15 min in the dark at room temperature with gentle agitation.

Protein biotinylation. The Cy3-labeled protein solution was adjusted to pH 7.4 by adding 400 µl of 0.8 M HEPES. The protein was biotinylated by adding 100 µl of 19 mg/ml biotin-PEG2-amine (Thermo Scientific) and 5 µl of 20 mg/ml EDC (Nacalai-tesque) and by incubating at room temperature for 1 hour with gentle agitation. To remove unreacted labeling reagents, protein solution was added to an ultrafiltration device (Amicon Ultra, NMWL, 10 kDa, Merck Millipore) and washed 5 times following the manufacturer's instructions.

Coverslip preparation and attachment of labeled proteins. A 22 x 22 mm sized and 0.17 mm thick coverslip (VWR) was treated with air plasma for one minute in a plasma cleaner (Diener Electronic). The coverslip was spin-coated with 200 µl of Avidin buffer (5 mM HEPES, 10 ng/ml avidin (Nacalai-tesque), and 2 mg/ml BSA) for 5 seconds at 500 rpm, followed by 30 seconds at 1,000 rpm. This avidin concentration results in 164 (± 35.5) spots

per image when measuring 100%-labeled biotin. The avidin-coated coverslip was then dried for 15 minutes at room temperature. 200 µl of the fluorescently labeled and biotinylated protein that was diluted 100 times in 5 mM HEPES were loaded on the coverslip for 15 minutes. Concentrations of the protein or biotin samples are in excess compared to the estimated number of avidin molecules (Supplementary Fig. S1). The coverslip was then washed 5 times with distilled water for 5 min each. A fresh, plasma-cleaned coverslip was placed on the washed coverslip to avoid drying. Alternatively, in the case of the positive control, 10 ng/ml biotin labeled with Alexa Fluor 488 (Nanocs, USA) was loaded instead of the sample protein. In the case of the negative control, fluorescently labeled and biotinylated BSA was loaded on the coverslip coated with Avidin free buffer (5 mM HEPES and 2 mg/ml BSA).

Image acquisition. The coated coverslips were imaged using an inverted epi-fluorescence microscope (IX81, Olympus) with an oil-immersion objective lens (PLAPON 60x, NA 1.42, Olympus) equipped with an EMCCD camera (iXon 897, Andor). Fluorescence was induced with wide-field illumination by a 488-nm Argon ion laser (Innova 70C, Coherent) or 560-nm fiber laser (F-04306-2, MPB Communications). The laser power density at the observation area was set to 117 W/cm2. Alexa Fluor 488 was imaged through a dichroic mirror FF495-Di03 (Semrock) and an emission filter FF02-520/28 (Semrock), whereas Cy3 was imaged through a dichroic mirror Di02-R561 (Semrock) and an emission filter FF02-617/73 (Semrock). The acquisition time was set to 100 ms. Commercial software (Metamorph, Molecular Devices) was used for the hardware control. For each condition, more than 100 images were taken to obtain enough values for statistical significance.

Image analysis. The obtained images were first processed using the laser illumination heterogeneity compensation [23] to uniformize fluorescence counts over the entire imaging area. Then the images were processed with the rolling ball algorithm with the ball radius of 50 pixel to subtract the background, and a band pass filter to extract spots of 1-20 pixels. The image was binarized using the Triangle threshold, and processed with the image filter that removes spots smaller than 2 pixels square. Then, the spot number was calculated from each image, which was used for the *LO* calculation (Fig. 2A). Also, histogram of fluorescence counts for spots was analyzed. Then, the histogram was re-binned based on peaks observed in the histogram, in order to represent the number of dyes in spots. n_{dye} was calculated from the mean of this distribution, and the histogram integrated 100 - *LO* at zero position to obtain a

probability density function of the number of dyes in a protein (Supplementary Figs. S3, S4 and S6). These processes were performed using the ImageJ software (v1.51n) or Python (v2.7).

Data reproducibility and errors. We checked the reproducibility of *LO* measurements by confirming that similar values were obtained from independent samples (Supplementary Fig. S9). We have also analyzed the accuracy of *LO* measurements depending on the number of acquired images. We observed that the statistical error ratio of the measured *LO* decreased from 4.6% to 0.2% for 20 and 100 images, respectively (Supplementary Fig. S10). As we have analyzed at least 20 images for every conditions (median: 52 images), our data has less than 5% errors.

Coupling efficiency calculation. The coupling efficiency, *CE*, is given by the equation:

$$CE = (A_{max} \times \varepsilon_{protein}) \,/\, ((A_{280} - A_{max} \times CF) \times \varepsilon_{dye}) \quad (1)$$

where *CF* is the correction factor for the dye (0.11 for Cy3), A_{max} is the absorbance of the dye at its maximum absorption wavelength (560 nm), A_{280} is the absorbance of the protein at 280 nm, ε_{dye} is the extinction coefficient of the dye at its maximum absorption wavelength (150,000 $M^{-1}.cm^{-1}$), and $\varepsilon_{protein}$ is the extinction coefficient of the protein at 280 nm.[6]

The estimated labeling occupancy from the coupling efficiency, LO_{CE}, was calculated under the assumption that a dye binds to a protein in a manner that follows the Poisson law, described by the following equation:

$$LO_{CE} = 1 - \left(\sum_{\mu=0}^{\infty} \left(CE^{\mu} * e^{-CE} \right) / \mu ! \right) \quad (2)$$

where *CE* is the coupling efficiency and μ is the number of dye molecules.

Calculation of the average number of lysine residues. The average number of lysine residues of proteins in a proteome fraction, n_{lys}, was obtained as the weighted average based on the protein abundance for a group of proteins having similar molecular weights with an error of ± 2%.

SDS-PAGE. SDS-PAGE was performed following the standard procedure using 20% acrylamide gel. To avoid a damage to the dye, the protein sample was not heated after adding the SDS sample buffer. The gel was imaged in a gel viewer (ImageQuant LAS 4000, GE Healthcare Life Science) with 180 seconds acquisition time using a filter set for Cy3. After imaging of gel, gel was cut using a sharp blade at the band level or uniformly, and proteins

were extracted from the gel using a dialyzer by electroelution (D-tube, 6-8 kDa, Merck Millipore) following the manufacturer's instructions. The electroelution was done for 3 hours under 100 V.

Data Availability. The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Associated content

The Supporting Information is available free of charge on the ACS Publications website.

Additional details on sensitivity limit, single molecule photobleaching, probability density function of the dye binding to the sample, full SDS-PAGE migration profile, effect of the dye concentration on the labeling, day-to-day reproducibility and accuracy.

Author informations

Corresponding author
E-Mail: taniguchi@riken.jp, Tel.: +81-6-6155-0114; Fax: +81-6-6155-0112.

ORCID
Taniguchi Yuichi : 0000-0001-5677-8901
Leclerc Simon : 0000-0003-1611-7572

Notes
RIKEN has filed a patent application on these results with S.L. and Y.T. named as co-inventors. Y.A. declares having no competing interests.

Acknowledgments

The authors thank Masae Ohno and Yamato Yoshida for critically reading the manuscript; Vipin Kumar for helpful discussions; Kazuya Nishimura for experimental assistance; and David G. Priest for helpful discussions and editorial assistance. This work was supported by PRESTO (JPMJPR15F7), Japan Science and Technology Agency, by grants-in-aid for Young Scientists (A) (24687022), Challenging Exploratory Research (26650055) and Scientific Research on Innovative Areas (23115005), Japan Society for the Promotion of Science, and by grants from the Takeda Science Foundation and the Mochida Memorial Foundation for Medical and Pharmaceutical Research. S.L. acknowledges support from RIKEN International Program Associate (IPA) program.

Abbreviations

PAGE : polyacrylamide gel electrophoresis; BSA : bovine serum albumin; *LO* : labeling occupancy; *CE* : coupling efficiency; *DoL* : degree of labeling; LO_{Obs} : single molecule measured *LO*; LO_{CE} : *LO* estimated from *CE*; CHAPS : 3-[(3-cholamidopropyl)dimethylammonio]propanesulfonate; DTT : dithiothreitol; SDS : sodium dodecyl sulfate

References

(1) Kondo, T.; Hirohashi, S. Application of Highly Sensitive Fluorescent Dyes (CyDye DIGE Fluor Saturation Dyes) to Laser Microdissection and Two-Dimensional Difference Gel Electrophoresis (2D-DIGE) for Cancer Proteomics. *Nat. Protoc.* **2006**, *1* (6), 2940–2956.

(2) Tannu, N. S.; Hemby, S. E. Two-Dimensional Fluorescence Difference Gel Electrophoresis for Comparative Proteomics Profiling. *Nat. Protoc.* **2006**, *1* (4), 1732–1742.

(3) Neuhoff, V.; Arold, N.; Taube, D.; Ehrhardt, W. Improved Staining of Proteins in Polyacrylamide Gels Including Isoelectric Focusing Gels with Clear Background at Nanogram Sensitivity Using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis* **1988**, *9* (6), 255–262.

(4) Berggren, K.; Chernokalskaya, E.; Steinberg, T. H.; Kemper, C.; Lopez, M. F.; Diwu, Z.; Haugland, R. P.; Patton, W. F. Background-Free, High Sensitivity Staining of Proteins in One- and Two-Dimensional Sodium Dodecyl Sulfate-Polyacrylamide Gels Using a Luminescent Ruthenium Complex. *Electrophoresis* **2000**, *21* (12), 2509–2521.

(5) Nanda, J. S.; Lorsch, J. R. Labeling a Protein with Fluorophores Using NHS Ester Derivitization. *Methods Enzymol.* **2014**, *536*, 87–94.

(6) Kim, Y.; Ho, S. O.; Gassman, N. R.; Korlann, Y.; Landorf, E. V.; Collart, F. R.; Weiss, S. Efficient Site-Specific Labeling of Proteins via Cysteines. *Bioconjug. Chem.* **2008**, *19* (3), 786–791.

(7) Funatsu, T.; Harada, Y.; Tokunaga, M.; Saito, K.; Yanagida, T. Imaging of Single Fluorescent Molecules and Individual ATP Turnovers by Single Myosin Molecules in Aqueous Solution. *Nature* **1995**, *374* (6522), 555–559.

(8) Peck, K.; Stryer, L.; Glazer, A. N.; Mathies, R. A. Single-Molecule Fluorescence Detection: Autocorrelation Criterion and Experimental Realization with Phycoerythrin. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86* (11), 4087–4091.

(9) Sheng, R.; Ni, F.; Cotton, T. M. Determination of Purine Bases by Reversed-Phase High-Performance Liquid Chromatography Using Real-Time Surface-Enhanced Raman Spectroscopy. *Anal. Chem.* **1991**, *63* (5), 437–442.

(10) Mutch, S. A.; Fujimoto, B. S.; Kuyper, C. L.; Kuo, J. S.; Bajjalieh, S. M.; Chiu, D. T. Deconvolving Single-Molecule Intensity Distributions for Quantitative Microscopy Measurements. *Biophys. J.* **2007**, *92* (8), 2926–2943.

(11) Betzig, E.; Patterson, G. H.; Sougrat, R.; Lindwasser, O. W.; Olenych, S.; Bonifacino, J. S.; Davidson, M. W.; Lippincott-Schwartz, J.; Hess, H. F. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **2006**, *313* (5793), 1642–1645.

(12) Lee, S.-H.; Shin, J. Y.; Lee, A.; Bustamante, C. Counting Single Photoactivatable Fluorescent Molecules by Photoactivated Localization Microscopy (PALM). *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (43), 17436–17441.

(13) Dickson, R. M.; Cubitt, A. B.; Tsien, R. Y.; Moerner, W. E. On/off Blinking and Switching Behaviour of Single Molecules of Green Fluorescent Protein. *Nature* **1997**, *388* (6640), 355–358.

(14) Penna, A.; Demuro, A.; Yeromin, A. V.; Zhang, S. L.; Safrina, O.; Parker, I.; Cahalan, M. D. The CRAC Channel Consists of a Tetramer Formed by Stim-Induced Dimerization of Orai Dimers. *Nature* **2008**, *456* (7218), 116–120.

(15) Ji, W.; Xu, P.; Li, Z.; Lu, J.; Liu, L.; Zhan, Y.; Chen, Y.; Hille, B.; Xu, T.; Chen, L. Functional Stoichiometry of the Unitary Calcium-Release-Activated Calcium Channel. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (36), 13668–13673.

(16) Ulbrich, M. H.; Isacoff, E. Y. Subunit Counting in Membrane-Bound Proteins. *Nat. Methods* **2007**, *4* (4), 319–321.

(17) Sugiyama, Y.; Kawabata, I.; Sobue, K.; Okabe, S. Determination of Absolute Protein Numbers in Single Synapses by a GFP-Based Calibration Technique. *Nat. Methods* **2005**, *2* (9), 677–684.

(18) Puchner, E. M.; Walter, J. M.; Kasper, R.; Huang, B.; Lim, W. A. Counting Molecules in Single Organelles with Superresolution Microscopy Allows Tracking of the Endosome Maturation Trajectory. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (40), 16015–16020.

(19) Grassart, A.; Cheng, A. T.; Hong, S. H.; Zhang, F.; Zenzer, N.; Feng, Y.; Briner, D. M.; Davis, G. D.; Malkov, D.; Drubin, D. G. Actin and dynamin2 Dynamics and Interplay during Clathrin-Mediated Endocytosis. *J. Cell Biol.* **2014**, *205* (5), 721–735.

(20) Finer, J. T.; Simmons, R. M.; Spudich, J. A. Single Myosin Molecule Mechanics: Piconewton Forces and Nanometre Steps. *Nature* **1994**, *368* (6467), 113–119.

(21) Svoboda, K.; Schmidt, C. F.; Schnapp, B. J.; Block, S. M. Direct Observation of Kinesin Stepping by Optical Trapping Interferometry. *Nature* **1993**, *365* (6448), 721–727.

(22) Ishijima, A.; Kojima, H.; Funatsu, T.; Tokunaga, M.; Higuchi, H.; Tanaka, H.; Yanagida, T. Simultaneous Observation of Individual ATPase and Mechanical Events by a Single Myosin Molecule during Interaction with Actin. *Cell* **1998**, *92* (2), 161–171.

(23) Taniguchi, Y.; Choi, P. J.; Li, G.-W.; Chen, H.; Babu, M.; Hearn, J.; Emili, A.; Xie, X. S. Quantifying E. Coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* **2010**, *329* (5991), 533–538.

(24) Yu, J.; Xiao, J.; Ren, X.; Lao, K.; Sunney Xie, X. Probing Gene Expression in Live Cells, One Protein Molecule at a Time. *Science* **2006**, *311* (5767), 1600–1603.

(25) Rose, G.; Geselowitz, A.; Lesser, G.; Lee, R.; Zehfus, M. Hydrophobicity of Amino Acid Residues in Globular Proteins. *Science* **1985**, *229* (4716), 834–838.

(26) Bhuyan, A. K. On the Mechanism of SDS-Induced Protein Denaturation. *Biopolymers* **2010**, *93* (2), 186–199.

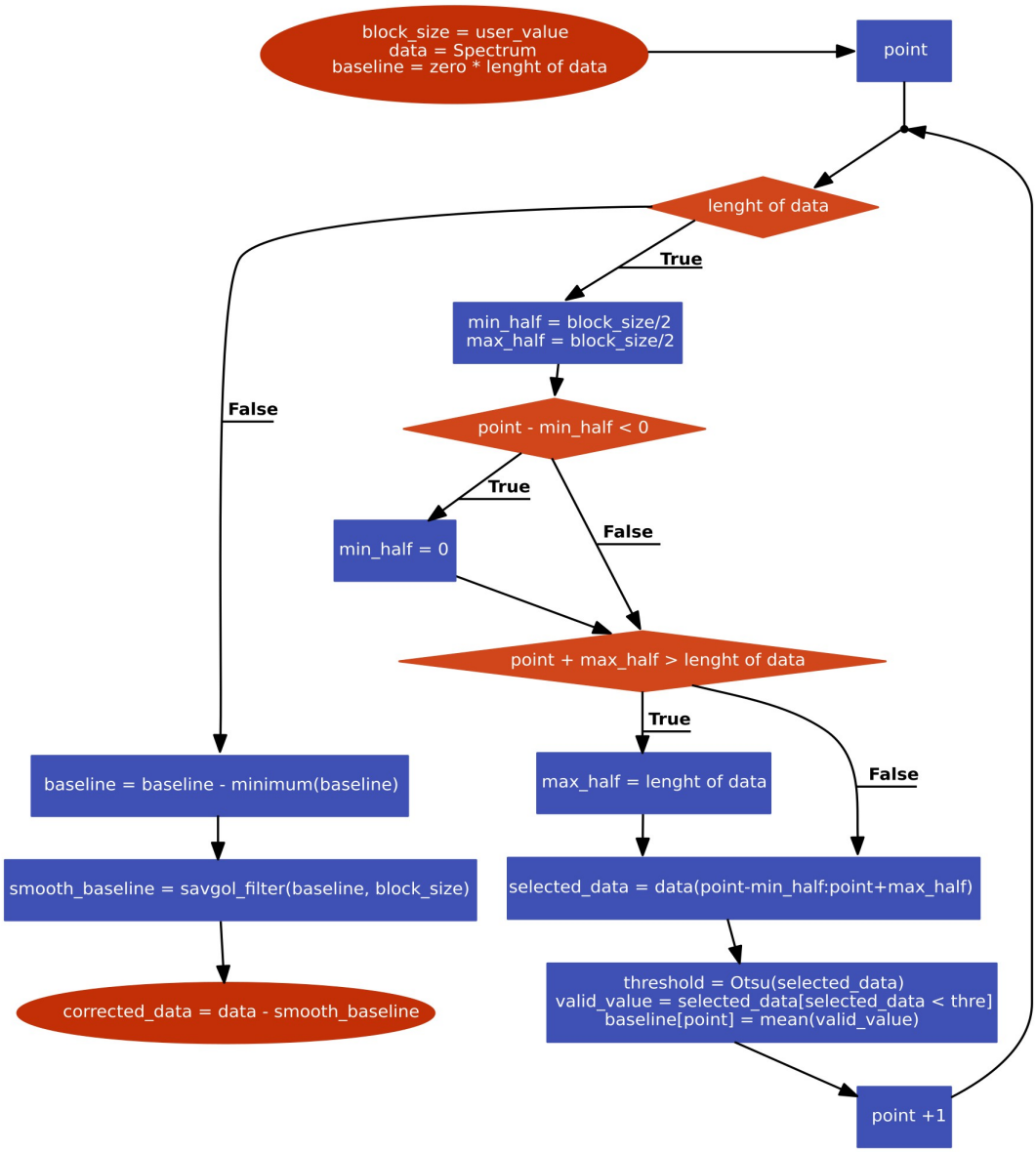
(27) Kato, A.; Tsutsui, N.; Matsudomi, N.; Kobayashi, K.; Nakai, S. Effects of Partial Denaturation on Surface Properties of Ovalbumin and Lysozyme. *Agric. Biol. Chem.* **1981**, *45* (12), 2755–2760.

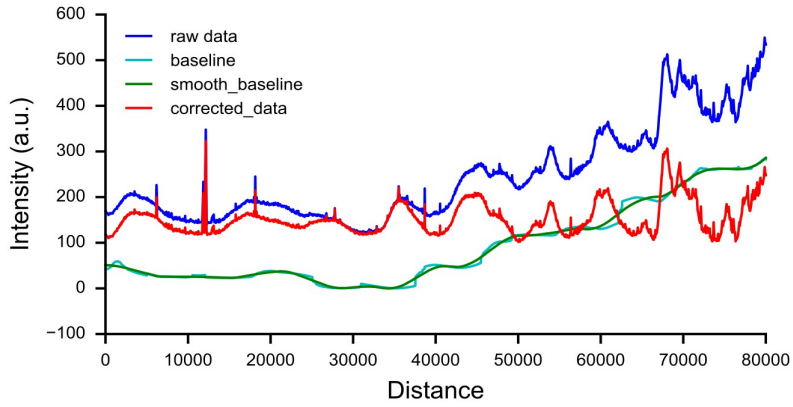
(28) Nelson, C. A. The Binding of Detergents to Proteins. I. The Maximum Amount of Dodecyl Sulfate Bound to Proteins and the Resistance to Binding of Several Proteins. *J. Biol. Chem.* **1971**, *246* (12), 3895–3901.

(29) Björck, L.; Kronvall, G. Purification and Some Properties of Streptococcal Protein G, a Novel IgG-Binding Reagent. *J. Immunol.* **1984**, *133* (2), 969–974.
(30) Bhattacharyya, B.; Volff, J. Membrane-Bound Tubulin in Brain and Thyroid Tissue. *J. Biol. Chem.* **1975**, *250* (19), 7639–7646.
(31) Konigsberg, W. [13] Reduction of Disulfide Bonds in Proteins with Dithiothreitol. In *Enzyme Structure, Part B*; Methods in Enzymology; Elsevier, 1972; Vol. 25, pp 185–188.
(32) Cai, H.; Wang, C. C.; Tsou, C. L. Chaperone-like Activity of Protein Disulfide Isomerase in the Refolding of a Protein with No Disulfide Bonds. *J. Biol. Chem.* **1994**, *269* (40), 24550–24552.
(33) Natarajan, S.; Xu, C.; Caperna, T. J.; Garrett, W. M. Comparison of Protein Solubilization Methods Suitable for Proteomic Analysis of Soybean Seed Proteins. *Anal. Biochem.* **2005**, *342* (2), 214–220.
(34) Rabilloud, T.; Adessi, C.; Giraudel, A.; Lunardi, J. Improvement of the Solubilization of Proteins in Two-Dimensional Electrophoresis with Immobilized pH Gradients. *Electrophoresis* **1997**, *18* (3-4), 307–316.
(35) Breuza, L.; Poux, S.; Estreicher, A.; Famiglietti, M. L.; Magrane, M.; Tognolli, M.; Bridge, A.; Baratin, D.; Redaschi, N.; UniProt Consortium. The UniProtKB Guide to the Human Proteome. *Database* **2016**, *2016*.
(36) Wang, M.; Herrmann, C. J.; Simonovic, M.; Szklarczyk, D.; von Mering, C. Version 4.0 of PaxDb: Protein Abundance Data, Integrated across Model Organisms, Tissues, and Cell-Lines. *Proteomics* **2015**, *15* (18), 3163–3168.
(37) Munkholm, C.; Parkinson, D. R.; Walt, D. R. Intramolecular Fluorescence Self-Quenching of Fluoresceinamine. *J. Am. Chem. Soc.* **1990**, *112* (7), 2608–2612.
(38) Zhang, M.; Biswas, S.; Deng, W.; Yu, H. The Crystal Structure of Monovalent Streptavidin. *Sci. Rep.* **2016**, *6*, 35915.
(39) Livnah, O.; Bayer, E. A.; Wilchek, M.; Sussman, J. L. Three-Dimensional Structures of Avidin and the Avidin-Biotin Complex. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (11), 5076–5080.
(40) Unlü, M.; Morgan, M. E.; Minden, J. S. Difference Gel Electrophoresis: A Single Gel Method for Detecting Changes in Protein Extracts. *Electrophoresis* **1997**, *18* (11), 2071–2077.
(41) Chen, H.; Ahsan, S. S.; Santiago-Berrios, M. B.; Abruña, H. D.; Webb, W. W. Mechanisms of Quenching of Alexa Fluorophores by Natural Amino Acids. *J. Am. Chem. Soc.* **2010**, *132* (21), 7244–7245.
(42) Zhuang, X.; Ha, T.; Kim, H. D.; Centner, T.; Labeit, S.; Chu, S. Fluorescence Quenching: A Tool for Single-Molecule Protein-Folding Study. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (26), 14241–14244.
(43) Johnson, A. C.; Bowser, M. T. High-Speed, Comprehensive, Two Dimensional Separations of Peptides and Small Molecule Biological Amines Using Capillary Electrophoresis Coupled with Micro Free Flow Electrophoresis. *Anal. Chem.* **2017**, *89* (3), 1665–1673.
(44) Huang, B.; Wu, H.; Bhaya, D.; Grossman, A.; Granier, S.; Kobilka, B. K.; Zare, R. N. Counting Low-Copy Number Proteins in a Single Cell. *Science* **2007**, *315* (5808), 81–84.

(45) Hughes, A. J.; Spelke, D. P.; Xu, Z.; Kang, C.-C.; Schaffer, D. V.; Herr, A. E. Single-Cell Western Blotting. *Nat. Methods* **2014**, *11* (7), 749–755.

Annex II : Baseline correction pseudo-code





Annex III : Detailed protocol

Sample labelling

The popular Cy3 dye with NHS-ester modification can be use to label the amino-terminal function of the protein either with the lysine residue. By fixing strong denaturant conditions, the labeling is expected to be more uniform and increase the proportion of labeled protein.

Material

- Lysis/Denaturation buffer:
    - 20 mM Borate buffer
    - 1 % Tween 20 (v/v)
    - 1 % SDS (w/v)
    - Adjust at pH 12 using NaOH
- DTT (Dithiothreitol)
- CHAPS (3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate)
- Cy3 NHS ester at 0.2 µg/ml in DMF (Dimethyl Sulfide)

Protocol

- For 100 µl of fresh Lysis/Denaturation buffer, add 2 µl of DTT 0.5M and 4 µl of CHAPS 50% (w/v)
- In a low protein binding tube, add 1 µl of Lysis/Denaturation buffer
- Add 0.2 µl of sample, then incubate 5 min at RT
- Add 0.2 µl Cy3 dye
- Incubate 15 min in dark at RT
- Quickly use the label sample or stock freeze at -80 degree Celsius

Single molecule labeling visualization

This technique permit to isolate labeled single protein on a glass surface to allow visualization/counting of single molecule. Protein need to be previously labeled using a NHS-ester dye. The principle is to fixed a biotin function to the target protein, then put this fusion on a avidin treated glass at a given density. The density of avidin is controlled by the avidin concentration, and can be necessary to optimize. The density is calculated by using a positive control composed of fluorescent biotin, and is then compared to the sample.

Annex-Annex III : Detailed protocol

Material

- Microscope coverslip, 22 x 22 mm, thickness 1 (0.12 to 0.17 mm)
- Hepes buffer, 50mM, pH7.5
- Avidin 1 mg/ml in 50 mM Hepes
- BSA 2 mg/ml in 50 mM Hepes
- PEG2 biotin amine 19mg/ml
- EDC 20 mg/ml
- Fluorescently labeled biotin

Protocol

Step 1: sample biotinylation

Add 20 µl of biotin amine PEG2 and 1 µl of EDC for every 100 µl of reaction. Incubate 1 hour at RT.

It is necessary to remove biotin excess for the labeling occupancy measurement using a size exclusion column (10 KDa):

- Put the sample on the column, complete to 500 µl using Hepes buffer, pH 7.5

- Centrifugate at 14000 g for 5 min

- Repeat the step 1 and 2 three times, then collect the sample

Step 2 : Single protein isolation on glass surface

- Prepare a solution at 10^{-5} g/L of avidin and 2 g/L of BSA in Hepes buffer
- Plasma treated 30 seconds the coverslip, one by condition
- Add 200 µl of BSA/avidin solution on the coverslip
- Spin-coated 5 seconds at 500 rpm, then 30 seconds at 1000 rpm
- Dry the glass for 15 minutes, protect from dust
- Add the biotinylated sample (dilution adjustment) with a final volume of 100 µl
- In parallel, make the controls (positive composed of fluorescent biotin at 10^{-3} mg/ml)
- Incubate 15 min protect from light and dusts

- Wash at least 3 times 5 minutes using distilled water
- Add a second plasma treated coverslip

Step 3 : Single molecule visualization

Focus at 0.6 µm of the first glass, then acquire a least 50 differents position of good images, since the avidin density can have some variability, with an acquisition time of 100 ms.

Glass Mold Fabrication for Micro-Gel

To create an ultra-thin SDS PAGE, it is necessary to first generate a mold. This mold can be generate by using different technique (3D print, layering, cut…). We used SU-8, negative photoresist to realize this mold on a glass substrate.

Material

- Plasma Cleaner
- Hot plate
- Spin Coater
- Chemical Hood
- SU-8 3025 solution
- UV lamp
- Glass press
- SU-8 developer : Propylene glycol monomethyl ether acetate
- Isopropanol
- Mask of the mold
- Support glass of 50x70 mm, thickness 1mm

Protocol

- Design of the mold print on transparent plastic with transparent part becoming hard
- Plasma treat 1 min a 50 x 70 mm glass, 1mm thickness
- Glass is heat at 100 degrees Celsius on a hot plate for 5 min
- Cool down the glass at RT, then place in the spin coater
- Add 3 ml of SU-8 3025 evenly on the surface of the glass

- Spin coat at 500 rpm for 15 seconds, then 1000 rpm for 30 seconds
- Soft bake 90 degrees Celsius until the solvent completely evaporate
- Repeat the 4 last steps two times, up to have 68 µm x 3 = 204 µm total thick
- Mont the glass and SU-8 with the plastic mask, and align for the edges on a press
- UV exposure 15 seconds (250 Joule/cm_2)
- Soft bake at 60 degrees Celsius for 10 min. The pattern need to be visible.
- Increase the temperature of the hot plate of 5 degrees Celsius every minute to reach a final temperature of 90 degrees Celsius
- Incubate 10 min at 90 degrees Celsius
- Cool down the glass at RT, then develop in a bath of SU-8 developer 5 min
- Make a second bath to remove all non polymerized SU-8 in SU-8 developer
- Wash the mold in distilled water, then in isopropanol
- Dry the mold, hard baked at 150 degrees Celsius for 5 min
- Slowly let the mold cool down at room temperature
- UV exposure of 5 min the mold to completely hardened the SU-8 layer

Micro SDS-PAGE fabrication

The fabrication of micro SDS-PAGE gel with a thickness of 0.2 mm is delicate and need to take in consideration the polymerization speed as well as the different inhibition factors, like oxygen, that have a bigger effect. The binding of the gel to a coverslip is also a condition for the microscope observation.

Material

- Previously made glass mold (50x70 mm)
- Coverslip (50x70mm, thickness 1 - 0.12-0.17 mm)
- Plasma cleaner
- Vacuum line or box
- Bind Silane solution in ethanol
    - 5 % acetic acid (v/v)
    - 0.3 % Bind Silane (v/v) (3 methacryloxypropyl trimethoxysilane)

- Repel Silane (dimethyldichlorosilane)
- Acrylamide gel solution
    - Acrylamide 20% (v/v)
    - SDS 0.1% (w/v)
    - Tris HCl pH 8.6 75 mM
- APS 10% (w/v) in water
- TEMED 10% (v/v) in water
- Riboflavin 1% (w/v) in water

Protocol

- Vacuum 15 min the acrylamide solution
- Plasma clean 4 coverslip 1 min
- Clean and dry the glass mold using isopropanol then drying
- Put 250 µl of Repel Silane uniformly on the glass mold surface
- Put 500 µl of Bind Silane solution uniformly on the coverslip surface
- Let Bind and Repel Silane react with the glass surface for 5 min, protect from evaporation by cover them using petri dish lid
- Dry both glass using an air gunner and realize a 100% ethanol wash on the Bind Silane glass
- Add 4.5 µl of TEMED 10% and Riboflavin 1%, then 1.5 µl of APS 10% for 700 µl of acrylamide solution
- Put the gel solution at the base of the cover-glass treated face
- Gently and slowly apply the mold (Repel treated face on gel contact) using tweezer
- Put a 5 kg pressure on the top of the glass for a flat polymerization
- Let polymerized at least 1.5 hours at room temperature
- Store the gel sandwich at 4 degree, maximum one week
- Before use, warm the gel sandwich at room temperature, then delicately separate glasses using thin tweezer or razor blade

Micro-Gel Migration

Migration of sample on very thin gel need to take some extra precaution, mainly because the gel thickness that cause the gel to quickly dry, motivating the use of a cool plate for the migration and lower voltage.

Material

- Horizontal migration device with cooling unit
- Hot Plate
- Gel viewer
- Running Buffer 10x
    - 144 g Glycine
    - 30.30 g Tris
    - 10 g SDS

Protocol

- Wash the gel in Running buffer 1x 2 min twice
- Remove the excess of liquid by gravity, by tilting the gel at 45 degree 30 min in a humid box
- Remove excess of liquid in the well by using a soft tissue
- Put some drop of kerosene on the cooling plate (20 degrees Celsius), then the gel
- Put the wet bridge on each extremity of the gel, well parallele
- Fill the tank using Running buffer 1x
- Add dry silicate beads close to the gel to absorb excess of humidity
- Add 1 µl of sample by well
- Migrate at 50V 15 min then 500V 2h45
- After migration, 2 times 2 min wash using big volume of distilled water
- Remove excess of liquid of the gel surface using soft tissue
- Dry the gel 30 min at 65 degree Celsius on a hot plate
- Pre-visualize the gel using the gel viewer, setting Cy3, 300s acquisition time

Gel Observation

The visualization of a fluorescent signal in the gel is achieved using a high sensitive motorized microscope, that possess the filter and dichroic setting for the visualization of the fluorescent dye Cy3. The lens used is a 30x objective with silicone oil immersion, having the advantages of a big numerical aperture (1.05 NA), a big scanning step of 274 micrometers thanks to the low magnification, quite small, and the use of the ZDC system that permit to realize a automatic and quick auto focus.

Material

- High sensitivity motorized microscope
- Laser and filter set to observe Cy3 (look exactly)
- Auto focus module ZDC
- Handmade glass holder
- Metamorph (or any microscope controller software)
- Macro in Metamorph for automatic scanning

Protocol

- Fixation of gel on the glass holder is realized by using a press system and screw (image)
- Add silicon oil to the glass side of the gel and on top of the objective
- Screw the glass holder to the stage

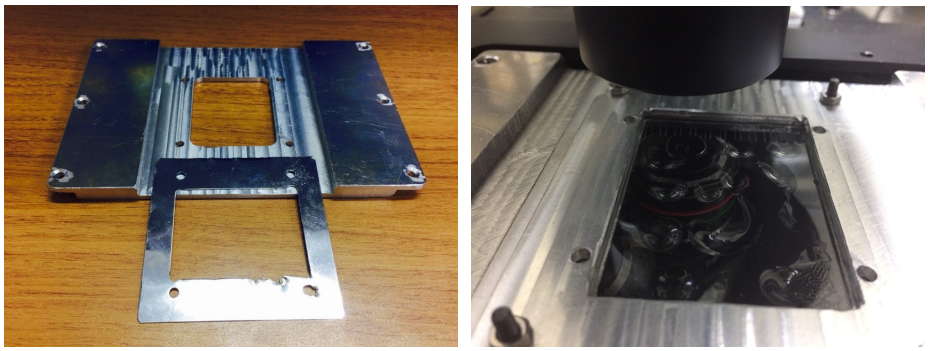
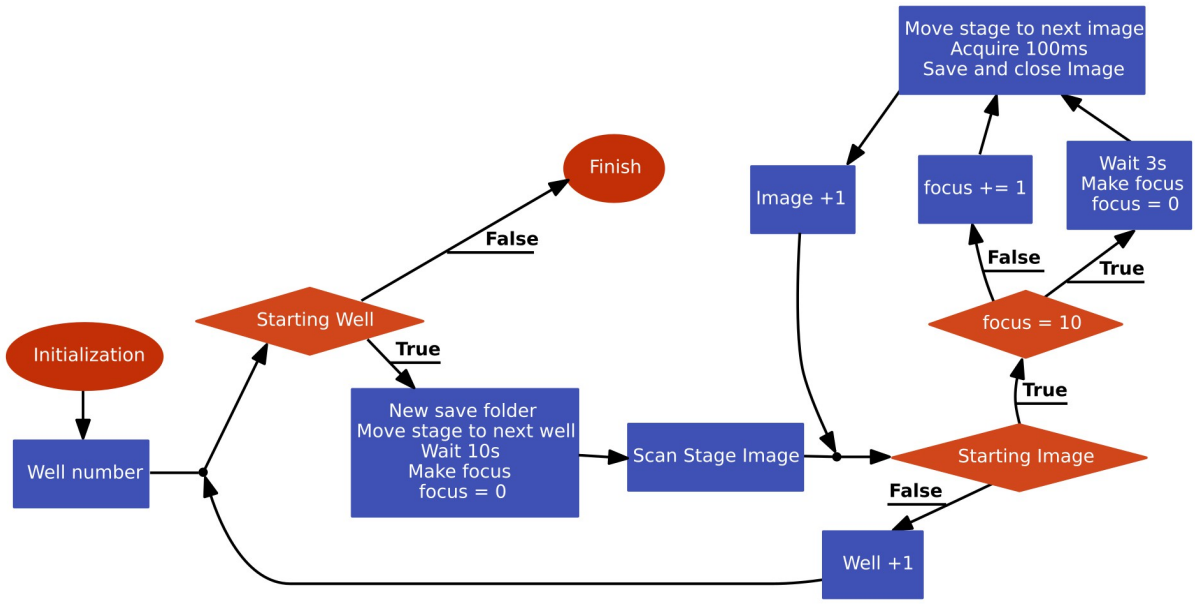


Photo of the glass holder empty (left) and mount on the microscope (right)

- Focus using the ZDC at 1 µm of the glass, in the gel
- Setup the speed of the stage to be at 100 µm/s
- Initialize the Macro for automatic scanning by manually detecting the center of the well to scan

- Enter the wish lenght and width area for the scanning of each well
- Launch the Macro for automatic scanning.
- The macro pseudo-code is the following:



Classic SDS-PAGE

SDS PAGE permits to separate proteins on a gel by their size, and is composed of polyacrylamide mesh. The proteins are separated by their mass, thanks to the SDS that neutralize their charge. The smallest protein exit the gel the quickest.

Material

- SDS
- polyacrylamide and bis-acrylamide
- APS
- TEMED
- Tris buffer 0.25 m at pH 6.8 and 0.75 M at pH 8.6
- Gel cassette
- Migration unit
- Running Buffer 10x
    - 144 g Glycine

- 30.30 g Tris
- 10 g SDS

Protocol

- Seal the gel cassette sides and bottom by using a pressure system (clamp)
- Prepare the following solution:

Stacking gel (4 ml)	Components	Resolving gel (12ml)
600 µl	Acrylamide 30%	8 ml
800 µl 0.25M pH 6.8	Tris Hcl buffer	2 ml 0.75 M pH 8.6
80 µl	SDS 10%	200 µl
2.5 ml	dH2O	1.8 ml
20 µl	APS 10%	50 µl
4 µl	TEMED	10 µl

- Pour the Resolving gel in the gel cassette, and equalize the level by gently adding 0.8 ml of distilled water
- After polymerization (around 30 min), remove the distilled water
- Pour the Stacking gel, then put immediately the well spacer
- After polymerization, mount the gel cassette on the migration unit
- Fill the migration tank and the gel cassette tank with running buffer 1x
- Remove the gel spacer, clean the well and inject the sample
- Migrate at 160V constant for 1h30 or until the bromophenol blue exit the gel

Simon LECLERC

Single molecule protein detection for proteomic profiling

Résumé
La quantification du protéome à très haute sensibilitée n'est actuellement pas réalisable, et est un problème pour l'analyse de cellule isolée, d'échantillon rare ou pour la détection de protéine de faible abondance. Afin d'améliorer la sensibilité, une idée est d'utiliser un microscope capable de détecter des protéines individuellement. Il faut pour cela dans un premier temps mesurer la proportion du protéome actuellement marquée par une sonde fluorescente afin de pouvoir faire des mesures quantitatives. Avec des conditions dénaturantes et la détection des amines, on arrive à marquer jusqu'à 75% du protéome d'un lysat cellulaire, avec un marquage plus efficace quand la protéine est de grande taille. Dans un deuxième temps, il faut séparer par la taille le protéome afin de réaliser un profil protéique. Si la puce microfluidique ne permet pas la réalisation d'un profil avec une résolution, le micro SDS-PAGE en est capable en permettant également l'observation du profil par microscopie, autorisant la détection jusqu'à 10 ng de protéines par bande et ainsi permettant d'obtenir un profil à partir de seulement 100 cellules. Cette sensibilité a permis l'identification de quatre lignées cellulaires de cancer du sein, avec un fort potentiel pour une application pour le diagnostic de cellule cancéreuse provenant de petite biopsie, plus facile pour le patient.

Mots-clés : molécule individuelle, marquage du protéome, protéomique, microscope à haute sensibilité

Abstract
Proteomic quantification at very high sensitivity is not achieved yet, even if they are a need to realize this quantification for the analysis of uncommon samples at a single cell level, or for the detection of low abundance protein. To improve this sensitivity, one way is to use a microscope able to detect single-molecule. In this optic, the first step to enable precise quantification is to measure the proportion of the proteome that is labeled by a fluorescent probe. When using strong denaturant conditions combined with a probe able to detect the amine of the protein, we are able to label up to 75% of the proteome from a cell lysate, with an increase in the labeling efficiency when the protein is bigger. The second step necessitates the protein separation by size in order to realize a proteome profile. Two technics were used for that, the microfluidic chip and the micro SDS-PAGE. The second one enables the possibility to scan the profile by microscopy, allowing the detection of up 10 ng of protein and then permits the analysis of only 100 cells. This sensitivity enables the differentiation of 4 different proteome profiles from cell lines originated from breast cancer, with a potential in the diagnostic of cancer cell from a smaller biopsy, allowing a less painful experience for the patient.
Keywords: single-molecule, protein labeling, proteomic, high sensibility microscopy