

## FULL ARTICLE

# Label-free quantitative chemical imaging and classification analysis of adipogenesis using mouse embryonic stem cells

Francesco Masia<sup>1\*</sup> | Adam Glen<sup>2</sup> | Phil Stephens<sup>2</sup> | Wolfgang Langbein<sup>1</sup> | Paola Borri<sup>3</sup>

<sup>1</sup>School of Physics and Astronomy, Cardiff University, Cardiff, UK

<sup>2</sup>School of Dentistry, Cardiff University, Cardiff, UK

<sup>3</sup>School of Biosciences, Cardiff University, Cardiff, UK

**\*Correspondence**

Francesco Masia, School of Physics and Astronomy, Cardiff University, Cardiff, UK.  
Email: masiaf@cf.ac.uk

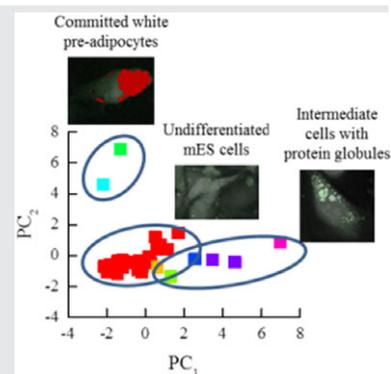
**Funding information**

Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/H045848/1, EP/M028313/1, EP/I005072/1; Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/H006575/1; Royal Society, Grant/Award Number: WM140077

Stem cells have received much attention recently for their potential utility in regenerative medicine. The identification of their differentiated progeny often requires complex staining procedures, and is challenging for intermediary stages which are a priori unknown. In this work, the ability of label-free quantitative coherent anti-Stokes Raman scattering (CARS) micro-spectroscopy to identify populations of intermediate cell states during the differentiation of murine embryonic stem cells into adipocytes is assessed. Cells were imaged at different days of differentiation by hyperspectral CARS, and images were analysed with an unsupervised factorization algorithm providing Raman-like spectra and spatially resolved maps of chemical components. Chemical decomposition combined with a statistical analysis of their spatial distributions provided a set of parameters that were used for classification analysis. The first 2 principal components of these parameters indicated 3 main groups, attributed to undifferentiated cells, cells differentiated into committed white pre-adipocytes, and differentiating cells exhibiting a distinct protein globular structure with adjacent lipid droplets. An unsupervised classification methodology was developed, separating undifferentiated cell from cells in other stages, using a novel method to estimate the optimal number of clusters. The proposed unsupervised classification pipeline of hyperspectral CARS data offers a promising new tool for automated cell sorting in lineage analysis.

**KEYWORDS**

adipogenesis, classification analysis, coherent anti-Stokes Raman Scattering, hyperspectral imaging, stem cells



## 1 | INTRODUCTION

A significant challenge when investigating biological phenomena is that biological systems exhibit complex dynamic interactions with inter-cellular and intra-cellular states fluctuating and changing to accommodate feedback from the surrounding micro-environment in order to enable tissue

homeostasis. To some extent, this complexity can be recreated via growth of cells in in-vitro environments in order to ask fundamental biological questions regarding the control of these processes [1, 2]. One class of cells which has received much attention in recent years, primarily for their potential utility in regenerative medicine applications, are stem cells. Stem cells can be derived from a variety of tissue

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 The Authors. *Journal of Biophotonics* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

types at multiple developmental stages (such as embryonic, foetal) and during adulthood (such as haematopoietic, mesenchymal stem cells and neural crest derived stem cells [3]). Each stem cell type, normally loosely defined by its source of origin, has its advantages and disadvantages. The desirable properties of stem cells, which tissue engineers aim to exploit, are their ability to self-renew in-vitro (potentially indefinitely, dependent on stem cell type) and their inherent ability to transition through several intermediate or precursor states (progenitor cells) to produce multiple specialised cell types (differentiation). Stem cells and their differentiated progeny are currently primarily identified based on a combination of in-vitro morphology and marker profiles (sometimes destructive). This represents a significant hurdle for the discovery of novel stem cell types since it requires that markers for cellular states have been pre-identified and that these markers accurately reflect the cellular phenotype [4, 5]. Furthermore, markers of intermediary stages of cellular differentiation are frequently lacking due to the heterogeneous nature of stem cell cultures, resulting from their capacity to generate multiple cell types. Also, intermediate cell states are often ill-defined as a result of current experimental approaches which, while producing large volumes of data, tend to average cell population behaviour (e.g. in microarrays, proteomics and other high-throughput marker discovery approaches). Generating well-defined markers such as antibodies typically requires well-defined populations of intermediate cell states.

Destructive cellular staining procedures utilising validated fluorescently labelled antibodies against targets of interest remains the mainstay method to visualise and assess the localisation of proteins and other cellular constituents in order to infer cellular state. Significant limitations of immunofluorescent techniques are that generation of novel antibodies is a time-consuming, technically demanding and labour intensive task that still often utilises animals and results in antibodies that are either not specific for the intended target or bind to targets which bear no relevance for the phenotype being investigated. Furthermore, fluorescence intensity readouts are not quantitative due to photobleaching. Other classification techniques based on gene expression profiling are equally invasive, as they require destruction of the samples.

Label-free techniques, such as bright field or dark field [6] and quantitative phase imaging [7], have been used for cell sorting by extracting morphological features [6, 7], for example, size, shape, granularity or phase metrics [7], for example, dry mass, but are unable to separate its chemical composition.

In order to overcome some of these limitations, we sought to develop a label-free micro-spectroscopy platform which could assess intermediary cell states through the identification of novel, non-destructive, markers of cellular phenotype by utilising chemically specific vibrational imaging

based on coherent Raman scattering (CRS) and the model of mouse embryonic stem cells differentiating towards the adipocytic lineage. In the last decade, CRS micro-spectroscopy has emerged as a powerful technique which overcomes limitations of spontaneous Raman and offers high acquisition speed, compatibility with imaging living cells, and intrinsic three dimensional spatial resolution [8]. CRS utilises the interference between two optical fields to resonantly drive molecular vibrations in the focal volume, and a third field to read out the Raman scattering. Due to the coherent driving of the vibrational excitation, Raman scattering by identical chemical bonds in the focal volume constructively interferes, generating a signal enhancement compared to spontaneous Raman.

To date, only a few studies of stem cell differentiation using CRS micro-spectroscopy have been reported [9–13]. In Ref. [9], an increase of the protein:RNA ratio was found after the differentiation of murine embryonic stem cells (mESc), similar to what was observed in spontaneous Raman studies [14]. The ability of CRS micro-spectroscopy to detect lipid content and volume concentrations within lipid droplets formed during adipogenesis was demonstrated for adipose-derived stem cells in Refs. [10, 12]. Hofemeier and co-workers used Raman and CARS micro-spectroscopy to identify early signs of calcium deposition during osteogenic differentiation of human stem cells [13]. A hyperspectral CARS investigation of adipogenic and osteogenic differentiation of human mesenchymal stromal cells (MSCs) in Ref. [11] showed an increase of the lipid (for the adipogenesis) and mineral (for the osteogenesis) content during differentiation. Although promising, these works either studied differentiation into well-defined cell lineages with little heterogeneity (as is the case for MSCs), or lacked the ability to classify intermediate cell types in the absence of a priori knowledge.

In the present work, we have investigated the heterogeneous differentiation of mouse embryonic stem cells using hyperspectral coherent anti-Stokes Raman scattering (CARS) micro-spectroscopy combined with an advanced quantitative data analysis tool for unsupervised classification analysis. Cells were imaged at 14, 16, 18, 20, 23 and 29 days of differentiation by hyperspectral CARS, and images were analysed with an unsupervised factorization algorithm providing Raman-like spectra and spatially resolved maps of chemical components in vol:vol concentration units [15–17]. Chemical decomposition into protein, lipid and aqueous components combined with a statistical analysis of their spatial patterns enabled us to extract a set of parameters that were used for classification analysis. We found 3 main clusters that were attributed to undifferentiated cells, cells differentiated into committed white pre-adipocytes (as discussed below) and intermediate differentiating cells exhibiting a distinct protein globular structure with adjacent lipid droplets.

## 2 | MATERIALS AND METHODS

### 2.1 | Cell growth, differentiation and staining

IMT11 mouse embryonic stem cells [18] were maintained and differentiated towards an adipogenic lineage as previously described [15, 19] on gelatin (0.1% v/v) coated glass slides. Samples were fixed in 4% (w/v) paraformaldehyde at 14, 16, 18, 20, 23 and 29 days of differentiation and stored in Phosphate buffered saline (PBS)/50 units/ml penicillin/streptomycin prior to immunostaining. All time points were immunostained concurrently with an anti Fatty Acid Binding Protein 4 (rabbit polyclonal IgG anti-FABP4, ab66682, abcam, Cambridge, UK), a well described marker of adipogenesis [20] or with a secondary only control (fluorescein isothiocyanate isomer 1 [FITC]-conjugated swine anti-rabbit immunoglobulins, F0205, Dako, Cambridgeshire, UK) where omission of the primary antibody acts as a control for non-specific binding of the secondary antibody to the sample. Immunostaining was achieved via washing samples 3 times in PBS, blocked with blocking buffer comprising of 1% (w/v) BSA in PBS/tween (0.1% v/v) for 15 minutes and staining with or without primary antibody (1:200, 0.7 mg/ml) for 1 hour at 4°C. Samples were washed again 3 times in PBS, secondary antibody diluted in blocking buffer (1:200) and samples stained with secondary antibody overnight at 4°C. Samples were subsequently washed 3 times in PBS prior to CARS analysis.

### 2.2 | CARS micro-spectroscopy

CARS is a non-linear process, where a third-order signal is emitted when two electromagnetic fields of different frequency, historically called pump and Stokes, coherently excite a molecular vibration resonant at their frequency difference. In our experiment, CARS hyperspectral images have been acquired on a home-built multi-modal laser-scanning microscope based on an inverted Nikon Ti-U. A description of the set-up can be found in [21]. Briefly, pump and Stokes beams for CARS excitation are obtained by splitting a broadband (660-970 nm) laser beam from a 5 fs Ti:Sa laser into the wavelength ranges of 660 to 730 nm and 730 to 900 nm, respectively. Hyperspectral imaging is achieved by spectral focussing [22–25]. In this technique, the pump and Stokes pulses have equal linear chirp, so that their frequency difference is constant. By changing the delay between the two pulses, the frequency difference can be tuned to record a CARS spectrum. The CARS signal is collected in forward direction, discriminated by a pair of band-pass filters (Semrock FF01-562/40) and detected by a photomultiplier (Hamamatsu H7422-40). The vibrational frequencies which can be addressed in our set-up are in the range (1200-3800)  $\text{cm}^{-1}$  with a spectral resolution of 10  $\text{cm}^{-1}$ . The data discussed in this paper were taken over a (2600-3700)  $\text{cm}^{-1}$  range with a  $20 \times 0.75$  NA dry objective

and a 0.72 NA dry condenser. The excitation beam fill factor  $w_0/(f\text{NA})$  was 0.55, where  $w_0$  is the Gaussian waist parameter of the excitation beam at the objective entrance and  $f$  is the objective focal length [17]. The data were acquired by scanning the spatial positions for each vibrational frequency step, with a dwell time for each pixel of 10  $\mu\text{s}$ . The measured spatial resolutions for the CARS intensity (full-width at half-maximum [FWHM]) are 0.6 (1.1)  $\mu\text{m}$  in the lateral (axial) direction. The measured spatial resolutions for the retrieved CARS susceptibility (FWHM) are 1.0 (4.4)  $\mu\text{m}$  in the lateral (axial) direction [17].

### 2.3 | Hyperspectral image analysis

The CARS hyperspectral images were analysed using the methodology described in [15, 16]. Briefly, the CARS intensity  $I_C$  data are first de-noised using an unbiased singular value decomposition approach [15]. The data are then normalised using the CARS intensity spectrum  $I_{\text{ref}}$  measured in a reference material (glass) which does not show vibrational resonances in the frequency range investigated [15, 17]. This approach allows to correct for the spectrally dependent transduction coefficient of the set-up and to reference the signal to a known response. The resulting CARS ratio  $\bar{I}_C = I_C/I_{\text{ref}}$  is given by the absolute square  $|\bar{\chi}|^2$  of the complex susceptibility  $\bar{\chi}$  normalised to the reference material susceptibility. In order to obtain  $\bar{\chi}$ , which is proportional to the concentration of the chemical species present in the focal volume, we retrieve  $\bar{\chi}$  in amplitude and phase using a phase-corrected Kramers-Kronig method [15]. The obtained spectra  $\mathfrak{S}(\bar{\chi})$  resemble spontaneous Raman spectra and can be described as a linear combination of chemical components with spatially dependent concentrations. To infer the chemical components and their concentration maps, we use the FSC<sup>3</sup> algorithm (Factorization into Spectra and Concentrations of Chemical Components) on the retrieved hyperspectral susceptibility  $\bar{\chi}(\omega)$  [15, 16]. We refer to the resulting components as FSC<sup>3</sup> components, and to their spectra as FSC<sup>3</sup> spectra.

In the factorization, the frequency range was reduced to (2675-3200)  $\text{cm}^{-1}$  to limit the influence of the water component which is dominating  $\mathfrak{S}(\bar{\chi})$  for frequencies above 3100  $\text{cm}^{-1}$ . All shown data have been factorised together to provide a common spectral basis, and 5 components were used. The factorization is unsupervised and starts from random spectra and concentrations. No pure known compounds are used to train the algorithm. To provide sufficient repeatability of the FSC<sup>3</sup>, the knock-out approach [16] was used, where multiple runs of the algorithm are performed and the resulting factorization with the smallest error is selected. In Ref. [15], we demonstrated that FSC<sup>3</sup> is able to retrieve the spectra and the concentration of chemical components in a model system (lipid mixture) without pre-knowledge.

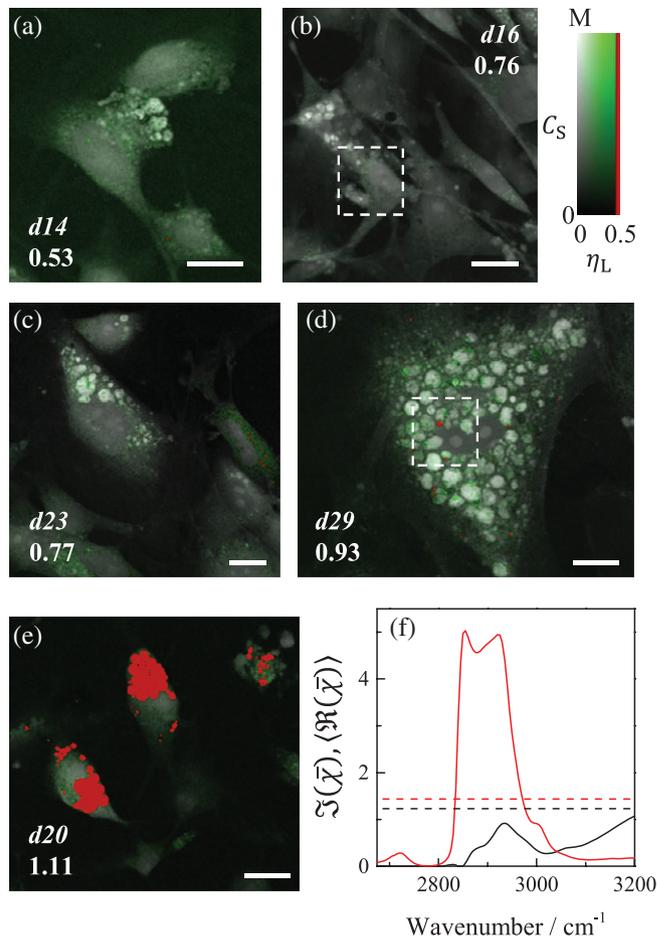
### 3 | RESULTS AND DISCUSSION

#### 3.1 | Hyperspectral CARS imaging and factorization

Representative results from hyperspectral CARS acquisition and analysis (see also section 2) for cells at different stages of differentiation are shown in Figure 1. The data analysis pipeline retrieves the imaginary part of the CARS susceptibility  $\Im(\chi)$ , providing a Raman-like spectrum [8], which is then factorised into the superposition of chemical components with spatially resolved concentration maps (see section 2.3). Three major components corresponding to water, protein and lipid were distinguished (see also in File S1 Figure S1, Supporting Information). The concentration maps  $C_L$  of lipids and  $C_P$  of proteins are shown in Figure 1 using the sum concentration  $C_S = C_L + C_P$  as brightness and the lipid fraction  $\eta_L = C_L/C_S$  as saturation with a green hue. In this scale, regions with high sum concentration of proteins and lipids (i.e. dry mass) appear bright. The lipid fraction is instead given by the colour saturation, that is, the colour changes from white to green as  $\eta_L$  increases from 0 to 0.5. Regions with  $\eta_L > 0.5$ , indicative of lipid droplets, are red.

The correlation between morphology and the number of days in the differentiation medium was found to be not strict. Even within a single plate, the morphology of the cells was heterogeneous (more data are shown in the Figure S2 in File S1). We also observed a low efficiency of the differentiation protocol, with less than 1% of the cells showing pre-adipocyte morphology containing large lipid droplets. These findings are expected for pluripotent stem cells undergoing differentiation and reflect the known limitations of differentiation protocols [26, 27]. Samples prepared under the same differentiation induction agents showed clear expression of adiponectin (a marker for differentiation into adipocytes). In Figure 1, we ordered the cells according to the morphology and lipid concentration  $C_L$  to approximate the development through the sequential stages. At early stages of differentiation (see Figure 1a,b), we observe a small lipid concentration  $C_L < 15\%$  and a lipid fraction  $\eta_L < 70\%$ . At later stages, the lipid concentration increases as small lipid droplets with  $\eta_L$  reaching unity form in the proximity of globular structures (see, eg, Figure 1c,d) dominated by proteic material.

For pre-adipocytes, identified by the presence of large lipid droplets as shown in Figure 1e, such proteic structures are not observed and  $C_P$  is homogeneous in the cytosol. The large size and the limited number of lipid droplets observed in those cells indicate that they are pre-adipocytes committed to differentiation into white adipocytes [28]. This assignment is supported by the comparison between spontaneous Raman spectra (see Figure S4 in File S1) acquired on mES cells-derived adipocytes and on white fat tissue. In both cases, the fingerprint region spectra show the absence of typical features observed in brown fat tissues, that is,



**FIGURE 1** Mouse embryonic stem cells at different stages of differentiation imaged using hyperspectral CARS and FSC<sup>3</sup> factorization. (a–e) sum concentration  $C_S$  of protein and lipid (0 –  $M$ ) as brightness, and the lipid fraction  $\eta_L$  (0–0.5) as saturation of green, with  $\eta_L > 0.5$  shown in red (see colour bar). The italic label indicates the number of days in the differentiation medium before fixation. The roman labels indicate  $M$ . The scale bars indicate 20  $\mu\text{m}$ . The dashed regions in (d) and (b) are used in Figure 4 and Figure S15 in File S1, respectively. (f) FSC<sup>3</sup> susceptibility spectra for the protein component  $C_P$  (black) and the lipid component  $C_L$  (red). The solid (dashed) lines refer to the imaginary part (mean value of the real part over the spectral range, giving the non-resonant contribution to the susceptibility), respectively

peaks between 1500 and 1600  $\text{cm}^{-1}$  [29]. Cells showing protein vesicles are unlikely to be beige [30] or brown (pre)-adipocytes, since first, they do not show a significant concentration of lipids in smaller lipid droplets, and second, both beige and brown adipocytes are characterised by the presence of a large concentration of mitochondria. From measurements comparing CARS and fluorescence images of human cells stained with MitoTracker, such large density of mitochondria gives rise to small micron-sized structures visible in the protein channel, much smaller than the vesicles observed here.

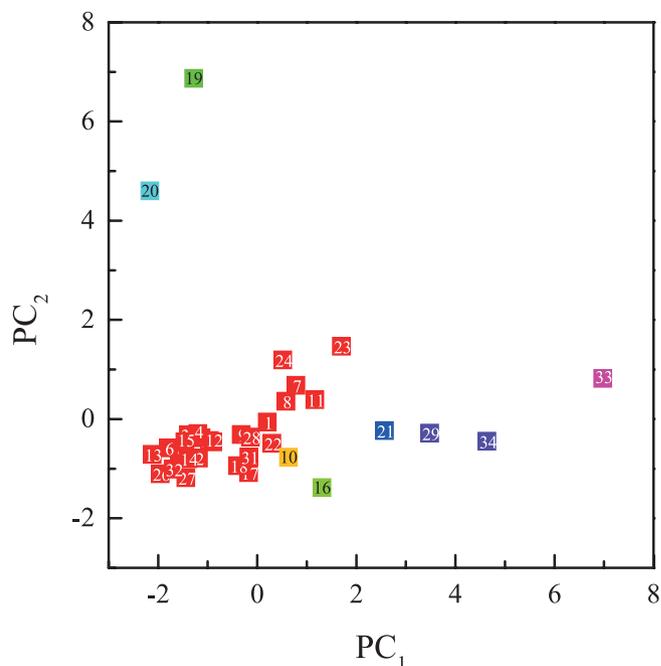
#### 3.2 | Cell classification

Fatty acid binding protein 4 (FABP4) is thought to be responsible for the formation of adipocytes [31]. We

therefore investigated FABP4 levels during the differentiation of embryonic stem cells, and the extent to which they correlate with the chemical composition measured by FSC<sup>3</sup>. To this end, we immunostained cell samples at different stages of differentiation and imaged them using widefield fluorescence microscopy prior to CARS imaging (see Supporting Information). We found that immunofluorescence can be detected only for pre-adipocytes where large lipid droplets are already formed, while the signal observed in undifferentiated cells seems to be dominated by autofluorescence and unspecific aggregation of the stain. We correlated the immunofluorescence observed in the pre-adipocytes to the chemical components individuated by FSC<sup>3</sup> and found no significant correlation, indicating that the expression level of FABP4 is below the CARS detection limit.

Driven by the need to identify label-free markers for early and intermediate stages of differentiation, we studied the spatial distribution of the chemical components obtained by the FSC<sup>3</sup> analysis, and developed a methodology to use these for classification analysis as detailed in the following. Cells which were contained in the field of view and not overlapping were segmented as individual objects using the image analysis software CellProfiler. Statistical features not explicitly depending on the cell size were calculated from the concentration maps of the lipid and protein component for each of those cells (see a list in Table S1 in File S1). The distribution of each feature across the cell ensemble was offset and normalised to have zero mean and unit variance, in order to allow relative comparison between the features (a common procedure in classification analysis [32]). To reduce the dimensionality, we applied principal component analysis (PCA) to the normalised features and we retained the principal components (PCs) with the largest variance, carrying 90% of the total variance (in the present case this resulted in 5 PCs, see also the dependence of the variance as a function of the number of PCs shown in Figure S11 in File S1). Note that the PCA is applied here on the statistical features of the chemical concentration distribution, and not directly on the spectral domain data as often done in Raman spectroscopy [33]. Figure 2 shows the analysed cells vs the first two PCs. The first PC contains all features except the average lipid concentration and its standard deviation (SD), while in the second component, those two features are dominant (see Table S1 in File S1). A visual inspection of Figure 2 shows most cells in the lower left part. Two cells (indexes 19 and 20) are well separated at large PC2—they are committed white pre-adipocytes, having a large PC2 due to their large lipid content. A row of cells towards larger PC1 are also distinguishable from the main distribution. These contain the differentiating cells with large protein globular structures (cells 33, 34 and 29).

To go beyond this visual inspection of the first two PCs, we have developed an unsupervised classification method based on hierarchical cluster analysis (HCA). We use the

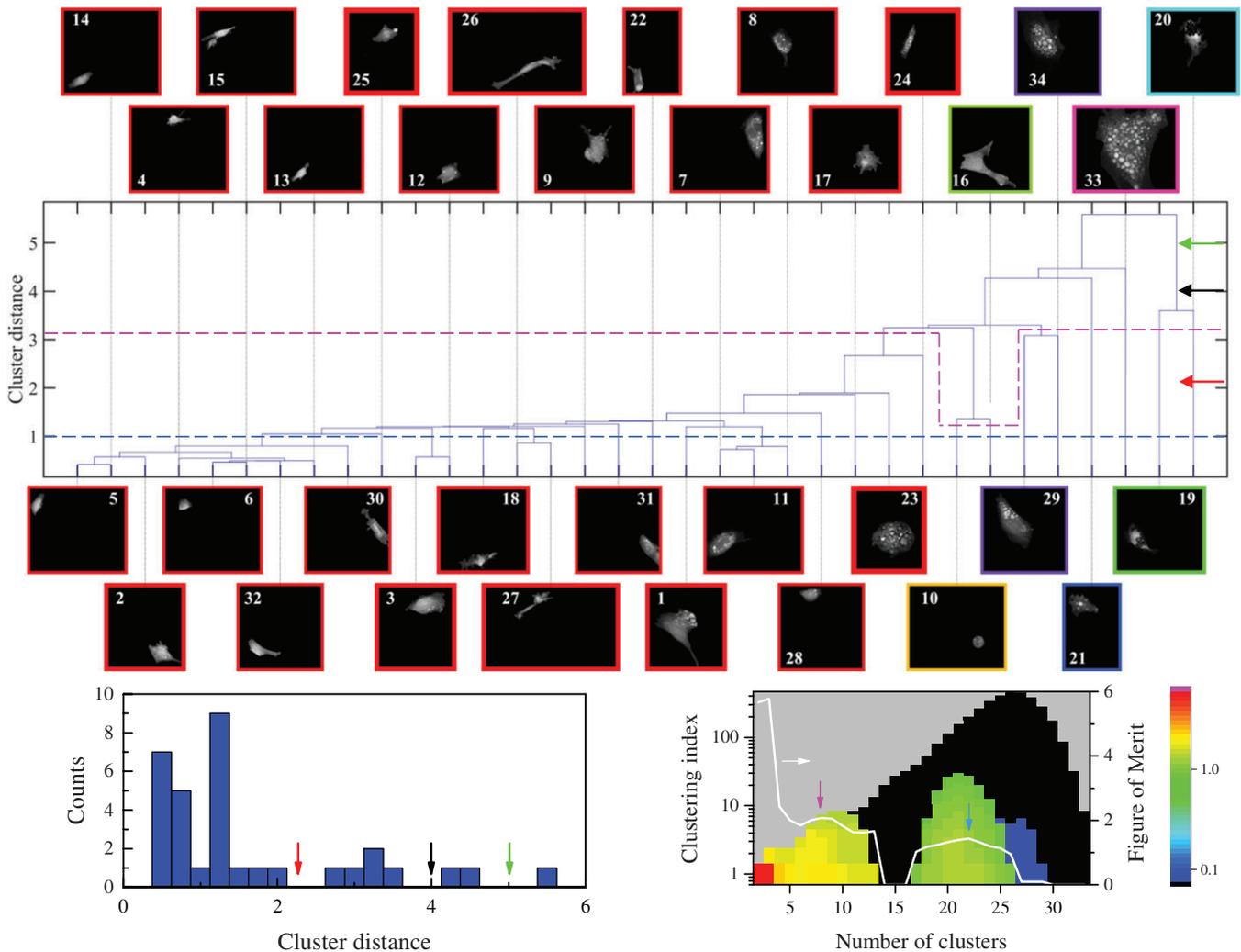


**FIGURE 2** Scatter plot of the first two PCs. The numbers inside the symbols refer to the cell index of Figure 3. The symbol colour reflects the affinity to a particular cluster (same coding as Figure 3)

first five PCs which are capturing the relevant information as discussed above, and an Euclidean metric to calculate the distance between pairs of cells in the PC space. Each cell is represented by the PC vector describing its statistical features, and the HCA clusters the cells by linking them, starting with the pair of shortest distance. The obtained dendrogram is shown in Figure 3. A dendrogram consists of U-shaped lines that connect clusters in a hierarchical tree. The height of each connection is given by the minimum distance between the corresponding cluster pair.

The dendrogram shows that the majority of cells have similar features and can be grouped in a single cluster. A few cells show distinctive properties and are progressively separated in individual clusters. The estimation of the optimal number of clusters is still a challenge in cluster analysis. Several models have been proposed, either based on measuring the intra- and inter-cluster distance [34, 35] or by estimating the stability of the clustering method against random sampling of the population [36, 37]. Considering the limited number of cells in our investigation and the tendency of HCA to isolate cells, the stability-based validation methods were not found suitable for our data.

An indication of the number of clusters present in the ensemble can be taken from the histogram of the distances. The minima in this histogram identify distances for which the clustering is most stable against perturbations of the features, that is, a slight change in the statistical quantities extracted from the cells will not affect the clustering. The corresponding histogram of cluster distances in Figure 3 shows three minima. The first one corresponding to the green arrows in the top of Figure 3 and in the bottom left of

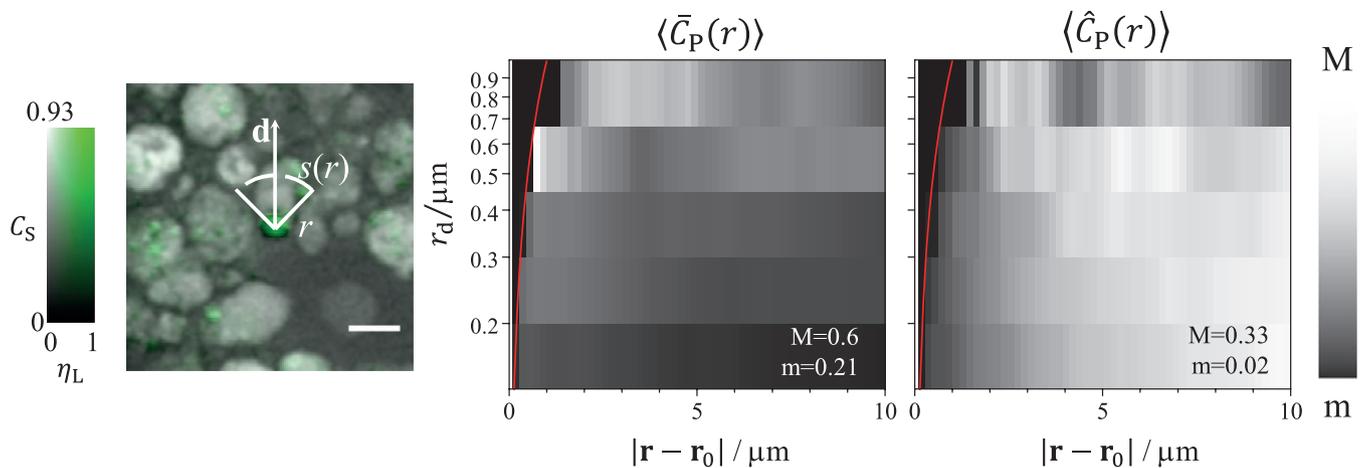


**FIGURE 3** Results of the classification analysis. Top: Dendrogram obtained from HCA. The cells associated with the PCs of the extracted features are shown as label of the dendrogram. The coloured frame indicates the affinity to the clusters corresponding to the combination with the highest calculated FOM when 8 total clusters are considered (also indicated by the dashed magenta line). The other horizontal dashed line correspond to the combination with the highest calculated FOM when 22 total clusters are considered. The coloured arrows indicate the minima of the histogram of cluster distances. Bottom left: histogram of the frequency counts of cluster distance. Bottom right: the coloured map indicates the clustering FOM as function of the number of clusters, and the cluster combination indexed with descending FOM. The value of the FOM is given on a logarithmic colour scale as indicated, with grey representing no data and black representing zero FOM. The white line (axis on the right) shows the highest FOM (i.e., the FOM of the cluster combination with index 1) as function of the number of clusters

Figure 3, separates 2 committed white pre-adipocytes from all other cells (ie, classifies cells into 2 clusters). The second one (black arrows in the top and bottom left of Figure 3, respectively) separates the group into 4 clusters: the committed white pre-adipocytes (1 cluster), 2 late-stage differentiating cells (each forming 1 cluster) with large lipid peak concentration and a large number of protein globules and all other cells. The third minimum (red arrows in the top and bottom left of Figure 3) separates the group into 9 clusters: the committed white pre-adipocytes (now, each forming 1 cluster), the 2 late-stage differentiation cells (each forming 1 cluster), intermediate differentiating cells with fewer protein globular structures and lower lipid concentration (forming 4 clusters mostly as individual cells) and all other undifferentiated cells.

Determining the clusters by drawing a horizontal line in the dendrogram corresponds to setting a lower limit for the

distance of the clusters, but in general it does not give information on how well the clusters are separated. We therefore introduce a new figure of merit (FOM) which quantifies how well all clusters are separated from each other. This FOM is calculated using the separation of each pair of clusters by the hyperplane determined by a support vector machine (SVM) classifier [38], see section S5 in File S1 for details. For a given number of clusters, we evaluate the FOM for each cluster selection from the HCA dendrogram which contains all cells. Figure 3 bottom right shows the resulting FOM as a function of the number of clusters and the specific combination of clusters which is identified by an index which increases with decreasing FOM. Pixels above the maximum index are given in grey, while the black pixels indicate cluster combinations which could not be separated, resulting in a FOM of zero. The method showed that the best classification occurs at 3 clusters,



**FIGURE 4** Distribution of the protein concentration around lipid droplets. Left: sketch of the integration contours used to calculate  $\bar{C}_P(r)$  and  $\hat{C}_P(r)$  around the lipid droplets, with the quadrant  $s(r)$  covering  $\pm \pi/4$  around the direction  $\mathbf{d}$  (see the supplementary information for details on the estimation of  $\mathbf{d}$ ). The image is a zoom of the areas indicated in Figure 1d. The colour scale is shown on the left, with the brightness proportional to  $C_S$  and the saturation of the green hue equal to  $\eta_L$ . The scale bar represents 5  $\mu\text{m}$ . Center and right: radial distribution of the protein concentration  $\langle \bar{C}_P(r) \rangle$  (center) and its relative SD  $\langle \hat{C}_P(r) \rangle$  (right) around LDs as function of the effective LD radius  $r_d$ . The red lines show  $r_d = |\mathbf{r} - \mathbf{r}_0|$ . Linear grayscale from minimum  $m$  to maximum  $M$  as indicated

2 formed by the 2 committed white pre-adipocytes and all other cells grouped in the third cluster. Increasing the number of clusters, a second maximum in the FOM is found at 8 clusters (shown by the dashed magenta line in the top of Figure 3). The corresponding clustering is shown as coloured frames of the images used as the labels of the dendrogram. In this configuration, the majority of the cells at an early stage of differentiation have been grouped in a single cluster, while cells at later stage of development (including the 2 committed white pre-adipocytes) form mostly individual clusters. We have tested our FOM against common distance-based indexes and we found some similarity in the results (see section S7 in File S1 for details). We also verified that the defined FOM is robust against increasing the number of PCs included in the analysis (see section S8 in File S1 for details). A third maximum is found for 22 clusters. The corresponding configuration is shown by the blue horizontal dashed line in Figure 3. In this case, the majority of cells occupy individual clusters, but still a large part of cells in an early stage of differentiation is grouped in a single cluster.

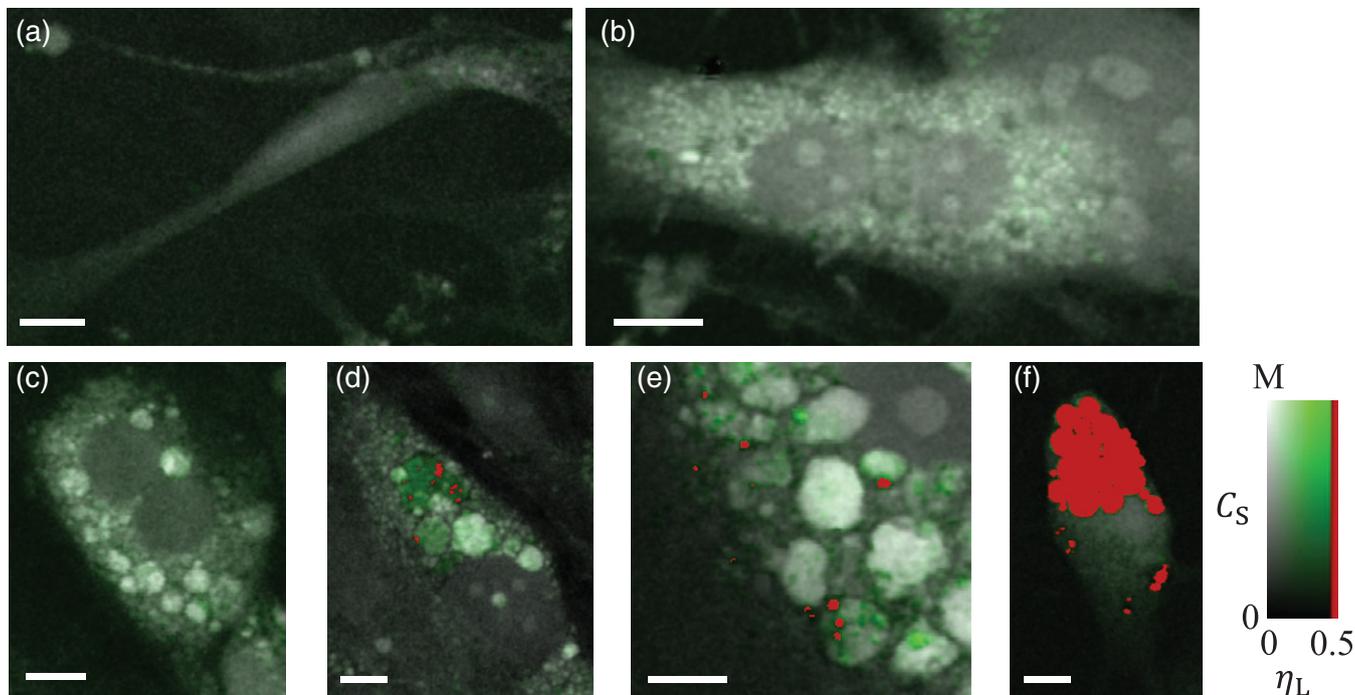
To put these results into context, we assigned the symbol colour in Figure 2 according to the cluster membership as given in Figure 3. A plot using as colour coding the 22 cluster configuration, indicated by the blue dashed line in Figure 3, is shown in Figure S8 in File S1.

Summarising, we find that the unsupervised classification on the basis of the spatial distribution of lipid and protein chemical components separates the cell ensemble into undifferentiated cells of low protein and lipid content, and differentiating cells of different types, with committed white pre-adipocytes with large lipid droplets and low protein content being most separated, followed by cells exhibiting large protein globular structures.

### 3.3 | FSC<sup>3</sup> lipid-protein correlation

In order to quantify in a statistically significant way the observation of lipid droplets forming next to protein-rich globules in differentiating cells, we calculated the average protein distribution as a function of the distance  $r = |\mathbf{r} - \mathbf{r}_0|$  from the centre of lipid droplets  $\mathbf{r}_0 = (x_0, y_0)$  (zooms of the regions corresponding to the areas indicated in Figure 1b,d, are presented in Figure S15 in File S1 and Figure 4, respectively). For each droplet, the distribution of the protein concentration around the LD is analysed (see sketch in Figure 4) by evaluating the mean value of the protein concentration  $\bar{C}_P(r)$  and its SD  $\hat{C}_P(r)$  over the contour of the quadrant  $s(r)$  of radius  $r$  centred at  $\mathbf{r}_0$  (for details see section S9 in File S1). The LD are then sorted according to their effective radius  $r_d = \sqrt{A_L/\pi}$ , where the lipid area  $A_L$  was obtained by spatially integrating the lipid concentration over the LD mask.  $r_d$  corresponds to the radius of a LD with the same total lipid area but made of pure lipid. Lastly, the mean values  $\langle \bar{C}_P(r) \rangle$  and  $\langle \hat{C}_P(r) \rangle$  outside of the droplet, that is, for  $r > r_d$ , were determined over the droplet ensemble in logarithmically spaced ranges of  $r_d$ , and are given in Figure 4.

The protein concentration around lipid droplets is larger for short distances to the droplet centre and reduces as the distance increases, which confirms the hypothesis that the LDs are localised in the proximity of the protein-rich globular structure. Interestingly, the concentration of protein in the nearby globules increases with the effective size (i.e., the mass) of the lipid droplet. The average normalised SD increases with the distance from the lipid droplet, showing a more homogeneous protein environment in proximity of the droplet than further away (similar results were obtained by performing a full circular integration of the protein



**FIGURE 5** Sequence of chemical composition images of stem-cells during differentiation, including undifferentiated cells (a) and committed pre-adipocytes (f). (b)-(e) show cells with increasing size of protein globules of increasing lipid content, and formation of lipid droplets in (d,e). The images have  $C_S$  as brightness, and  $\eta_L$  (0–0.5) as saturation of green, with  $\eta_L > 0.5$  shown in red (see colour bar). The scale bar represents 10  $\mu\text{m}$

concentration around the LDs instead of using a directional quadrant, see section S10 and Figure S15 in File S1). The results thus indicate that the position of LDs correlates with the boundaries of protein-rich structures.

To understand the possible role of these proteic globules, we speculate in Figure 5 that they rise and decay during differentiation and might be responsible for LD formation. Starting with undifferentiated cells (Figure 5a), which show a rather homogeneous lipid-protein composition of low protein concentration, protein-rich regions develop outside of the nucleus, as exemplified in Figure 5b. As the differentiation progresses the protein structures are organised in globules (Figure 5c), which are rich in protein. Notably, the lipid concentration inside these globules increases, a sign of lipid assimilation and/or synthesis. Eventually a sufficiently high lipid concentration is reached such that lipid droplets emerge from these globules (Figure 5d,e). The lipid droplets then fuse together and the cell acquires the committed white pre-adipocyte morphology (Figure 5f) showing large lipid droplets. These observations suggest that protein rich globules might be responsible for the initial formation of lipid droplets during differentiation into adipocytes, and that their presence, alongside that of adjacent lipid droplets, might be used a label-free marker of differentiation. Our interpretation suggests a model of lipid droplet formation specific to the differentiation phase towards adipocytes. It is however different to the prevailing model [39, 40] that LDs form in the endoplasmic reticulum. In the absence of a time course study on the same live cell, this interpretation remains a hypothesis.

#### 4 | CONCLUSIONS

We investigated mESc undergoing heterogeneous differentiation towards adipocytes using label-free chemically specific hyperspectral CARS micro-spectroscopy and our latest advances in quantitative data analysis. Chemical decomposition into protein and lipid components provided spatially resolved concentration maps. These maps were analysed on a cell-by-cell basis to provide 11 descriptors for each cell. PCA of the normalised descriptors captured 90% of the variance in 5 PCs. Visual inspection of the first 2 PCs allowed to identify undifferentiated cells, cells differentiated into committed white pre-adipocytes, and a second type of differentiating cells exhibiting significant protein globular structures. Small lipid droplets were found to colocalize with the protein structures in these cells. The first 5 PCs were used for HCA. To select the depth along the hierarchical tree, we introduced a novel FOM for the cluster separability based on the distance between clusters determined by a SVM analysis. The resulting clusters split the undifferentiated cells from the differentiated ones, first separating the pre-adipocytes, and then cells with protein globular structures. These results suggest an analysis pipeline for automated cell sorting, generally applicable to heterogeneous samples in the absence of a priori knowledge of their cell types. Furthermore, research in adipogenesis has reached worldwide biomedical importance, after obesity, and associated diseases have become a modern epidemic. Thus, our observation of a distinct chemical and morphological phenotype of differentiating cells exhibiting small lipid droplets next to large proteic globular structures

shall stimulate further studies into the quantitative understanding of such complex process from embryonic stem cells. The data presented in this work are available from the Cardiff University data archive [41].

#### ACKNOWLEDGMENTS

This work was funded by the UK EPSRC Research Council (grant no. EP/H45848/1). The CARS microscope development was funded by the UK BBSRC Research Council (grant no. BB/H006575/1) and by the UK EPSRC Research Council (grant n. EP/M028313/1). P.B. acknowledges the UK EPSRC Research Council for her Leadership Fellowship Award (grant no. EP/I005072/1) and the Royal Society for her Wolfson Research Merit Award (grant no. WM140077). We thank Iestyn Pope for support in running the microscope.

#### AUTHOR BIOGRAPHIES

Please see Supporting Information online.

#### REFERENCES

- [1] K. H. Benam, S. Dauth, B. Hassell, A. Herland, A. Jain, K.-J. Jang, K. Karalis, H. J. Kim, L. MacQueen, R. Mahmoodian, S. Musah, Y. Suke Torisawa, A. D. van der Meer, R. Villenave, M. Yadid, K. K. Parker, D. E. Ingber, *Annu. Rev. Pathol. Mech.* **2015**, *10*, 195.
- [2] N. Alépée, A. Bahinski, M. Daneshian, B. D. Wever, E. Fritsche, A. Goldberg, J. Hansmann, T. Hartung, J. Haycock, H. T. Hogberg, L. Hoelting, J. M. Kelm, S. Kadereit, E. McVey, R. Landsiedel, M. Leist, M. Lübberstedt, F. Noor, C. Pellevoisin, D. Petersohn, U. Pfannenbecker, K. Reisinger, T. Ramirez, B. Rothen-Rutishauser, M. Schäfer-Korting, K. Zeilinger, M.-G. Zurich, *ALTEX* **31**, 441–477 (2014).
- [3] S. Hauser, D. Widera, F. Qunneis, J. Müller, C. Zander, J. Greiner, C. Strauss, P. Lüningschrör, P. Heimann, H. Schwarze, J. Ebmeyer, H. Sudhoff, M. J. Arauzo-Bravo, B. Greber, H. Zaehres, H. Schöler, C. Kaltschmidt, B. Kaltschmid, *Stem. Cell. Dev.* **2012**, *21*, 742.
- [4] R. A. Young, *Cell* **2011**, *144*, 940.
- [5] B. R. Sousa, R. C. Parreira, E. A. Fonseca, M. J. Amaya, F. M. P. Tonelli, S. M. S. N. Lacerda, P. Lalwani, A. K. Santos, K. N. Gomes, H. Ulrich, A. H. Kihara, R. R. Resende, *Cytometry A* **2014**, *85*, 43.
- [6] T. Blasi, H. Hennig, H. D. Summers, F. J. Theis, J. Cerveira, J. O. Patterson, D. Davies, A. Filby, A. E. Carpenter, P. Rees, *Nat. Commun.* **2016**, *7*, 10256.
- [7] R. Kasprzowicz, R. Sumana, P. O'Toole, *Int. J. Biochem. Cell Biol.* **2017**, *84*, 89.
- [8] A. Zumbusch, W. Langbein, P. Borri, *Prog. Lipid Res.* **2013**, *52*, 615.
- [9] S. O. Konorov, C. H. Glover, J. M. Piret, J. Bryan, H. G. Schulze, M. W. Blades, R. F. B. Turner, *Anal. Chem.* **2007**, *79*, 7221.
- [10] R. Mouras, P. O. Bagnaninchi, A. R. Downes, A. P. D. Elfick, *J. Biomed. Opt.* **2012**, *17*, 116011.
- [11] Y. J. Lee, S. L. Vega, P. J. Patel, K. A. Aamer, P. V. Moghe, M. T. Cicerone, *Tissue Eng. C* **2014**, *20*, 562.
- [12] C. Di Napoli, I. Pope, F. Masia, W. Langbein, P. Watson, P. Borri, *Anal. Chem.* **2016**, *88*, 3677.
- [13] A. D. Hofemeier, H. Hachmeister, C. Pilger, M. Schrmann, J. F. W. Greiner, L. Nolte, H. Sudho, C. Kaltschmidt, T. Huser, B. Kaltschmidt, *Sci. Rep.* **2016**, *6*, 26716.
- [14] I. Notinger, I. Bisson, A. E. Bishop, W. L. Randle, J. M. P. Polak, L. L. Henchã, *Anal. Chem.* **2004**, *76*, 3185.
- [15] F. Masia, A. Glen, P. Stephens, P. Borri, W. Langbein, *Anal. Chem.* **2013**, *85*, 10820.
- [16] F. Masia, A. Karuna, P. Borri, W. Langbein, *J. Raman Spectrosc.* **2015**, *46*, 727.
- [17] A. Karuna, F. Masia, P. Borri, W. Langbein, *J. Raman Spectrosc.* **2016**, *47*, 1167.
- [18] F. C. Mansergh, C. S. Daly, A. L. Hurley, M. A. Wride, S. M. Hunter, M. J. Evans, *BMC Dev. Biol.* **2009**, *9*, 5.
- [19] C. Dani, A. Smith, S. Dessolin, P. Leroy, L. Staccini, P. Villageois, C. Darimont, G. Ailhaud, *J. Cell Sci.* **1997**, *110*, 1279.
- [20] T. Shan, W. Liu, S. Kuang, *FASEB J.* **2013**, *27*, 277.
- [21] I. Pope, W. Langbein, P. Watson, P. Borri, *Opt. Express* **2013**, *21*, 7096.
- [22] T. Hellere, A. M. Enejder, A. Zumbusch, *Appl. Phys. Lett.* **2004**, *85*, 25.
- [23] I. Rocha-Mendoza, W. Langbein, P. Borri, *Appl. Phys. Lett.* **2008**, *93*, 201103.
- [24] W. Langbein, I. Rocha-Mendoza, P. Borri, *J. Raman Spectrosc.* **2009**, *40*, 800.
- [25] W. Langbein, I. Rocha-Mendoza, P. Borri, *Appl. Phys. Lett.* **2009**, *95*, 081109.
- [26] K. Osafune, L. Caron, M. Borowiak, R. J. Martinez, C. S. Fitz-Gerald, Y. Sato, C. A. Cowan, K. R. Chien, D. A. Melton, *Nat. Biotech.* **2008**, *26*, 313.
- [27] P. Cahan, G. Q. Daley, *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 357.
- [28] M.-J. Obregon, *Am. J. Physiol. Cell Physiol.* **2014**, *306*, C425.
- [29] M. Troyanova-Wood, C. Gobbell, Z. Meng, A. A. Gashev, V. V. Yakovlev, *J. Biophotonics* **2017**, *1*, 9.
- [30] M. Harms, P. Seale, *Nat. Med.* **2013**, *19*, 1252.
- [31] D. Mosefi, A. Regassa, W.-K. Kim, *Int. J. Mol. Sci.* **2016**, *17*, 124.
- [32] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, *BMC Genomics* **2006**, *7*, 142.
- [33] X. Li, T. Yang, S. Li, D. Wang, Y. Song, S. Zhang, *Laser Phys.* **2016**, *26*, 035702.
- [34] G. W. R. Tibshirani, T. Hastie, *J. R. Statist. Soc. B* **2001**, *63*, 411.
- [35] P. Rousseeuw, *J. Comput. Appl. Math.* **1987**, *20*(1987), 53.
- [36] J. S. C. W. Hu, S. M. Kornblau, A. A. Qutub, *Sci. Rep.* **2015**, *5*, 12894.
- [37] S. Dudoit, J. Fridlyand, *Genome Biol.* **2002**, *3*, 0036.1.
- [38] N. Cristianini, J. C. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, **2000**.
- [39] S. Martin, R. G. Parton, *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 373.
- [40] D. J. Murphy, J. Vance, *Trends Biochem. Sci.* **1999**, *24*, 109.
- [41] Cardiff University data archive: <http://dx.doi.org/https://doi.org/10.17035/d.2017.0009096459>.

#### SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

#### File S1 Supporting Information

**How to cite this article:** Masia F, Glen A, Stephens P, Langbein W, Borri P. Label-free quantitative chemical imaging and classification analysis of adipogenesis using mouse embryonic stem cells. *J. Biophotonics*. 2018;e201700219. <https://doi.org/10.1002/jbio.201700219>