RESEARCH ARTICLE

Brain
Pathology

Automated and unbiased discrimination of ALS from control tissue at single cell resolution

Cathleen Hagemann[1,2] | Giulia E. Tyzack[1,3] | Doaa M. Taha[1,3] | Helen Devine[3] |
Linda Greensmith[3] | Jia Newcombe[4] | Rickie Patani[1,3] | Andrea Serio[1,2] |
Raphaëlle Luisier[5]

[1]The Francis Crick Institute, London, UK

[2]Centre for Craniofacial & Regenerative Biology, King's College London, London, UK

[3]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK

[4]NeuroResource, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, London, UK

[5]Idiap Research Institute, Martigny, Switzerland

Correspondence
Raphaëlle Luisier, Idiap Research Institute, Martigny, Switzerland.
Email: raphaelle.luisier@idiap.ch

Rickie Patani and Andrea Serio, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK.
Email: rickie.patani@crick.ac.uk (R. P.) and andrea.serio@kcl.ac.uk (A. S.)

Funding information
Medical Research Council, Grant/Award Number: FC010110 and MR/S006591/1; Cancer Research UK, Grant/Award Number: FC010110; Wellcome Trust, Grant/Award Number: 213949/Z/18/Z and FC010110

Abstract

Histopathological analysis of tissue sections is invaluable in neurodegeneration research. However, cell-to-cell variation in both the presence and severity of a given phenotype is a key limitation of this approach, reducing the signal to noise ratio and leaving unresolved the potential of single-cell scoring for a given disease attribute. Here, we tested different machine learning methods to analyse high-content microscopy measurements of hundreds of motor neurons (MNs) from amyotrophic lateral sclerosis (ALS) post-mortem tissue sections. Furthermore, we automated the identification of phenotypically distinct MN subpopulations in VCP- and SOD1-mutant transgenic mice, revealing common morphological cellular phenotypes. Additionally we established scoring metrics to rank cells and tissue samples for both disease probability and severity. By adapting this paradigm to human post-mortem tissue, we validated our core finding that morphological descriptors robustly discriminate ALS from control healthy tissue at single cell resolution. Determining disease presence, severity and unbiased phenotypes at single cell resolution might prove transformational in our understanding of ALS and neurodegeneration more broadly.

KEYWORDS
ALS, histopathology, machine learning, protein mislocalization

1 | INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a relentlessly progressive neurodegenerative disease, which remains incurable due to an incomplete understanding of the underlying molecular pathogenesis. Using the VCP mutant mouse transgenic model of ALS and human post-mortem tissue from sporadic ALS (sALS) patients, we previously showed that spinal motor neurons (MNs) exhibit reduced nuclear to cytoplasmic (N/C) ratios of FUS (Fused in Sarcoma) and SFPQ (Splicing factor Proline and Glutamine rich) proteins. Furthermore, we showed that SFPQ is also mislocalized in SOD1-mutant ALS models, while FUS is not (1,2).

Although these studies revealed novel molecular hallmarks of ALS, they also present important limitations shared by the majority of such traditional neuropathological studies. Firstly, key observations are based on manual

Rickie Patani, Andrea Serio, and Raphaëlle Luisier are contributed equally to this work.

image processing, which is time-consuming (therefore not scalable to large studies) and also introduces potential operator-dependent variability. Secondly, although hundreds of measurements can be extracted from such rich images, previous studies have typically focused on the single measurement deemed to be the most relevant. Furthermore, such analyses are usually based on averaging the signal(s) among all cells originating from identical animals or individuals. However, the degree to which individual MNs follow population-averaged trends in protein mislocalization remains unresolved. Indeed important insights into biological processes such as cellular differentiation and disease progression originate from the characterisation of heterogeneous cell populations coexisting within the same condition (3–8). These studies have been very effective when applied to assays of cultured cells and for cancer subtype classification, however there is not, to our knowledge, an analogous study on tissue sections from the nervous system (9,10). Finally these studies do not provide disease scoring at the cellular or even individual or animal level, which would be advantageous for understanding single cell phenotypic heterogeneity within a disease and, in the longer term, for developing superior patient-specific diagnostic criteria.

Recent developments using deep learning methods have enabled segmentation and object detection from complex imaging data-sets, with exciting prospects within ALS translational research (11,12). Here, by expanding on these methods, we developed a pipeline for image processing of MNs from spinal cord sections, greatly improving efficiency, whilst providing an unbiased approach to data acquisition. We then extracted hundreds of cell measurements and compared different machine learning methods to automatically identify MNs subpopulations from high-content imaging data. Using this approach we reveal that VCP- and SOD1-mutant MNs share phenotypes captured by several related morphological descriptors including Zernike moments, and we validate this finding in human ALS post-mortem tissues. By providing our fluorescence microscopy raw images together with open-source implementations of the methods, we aim to allow others to readily apply these methods in other biological contexts in order to increase the analytical power of tissue section analysis. We propose that such unbiased approaches may substantially deepen our understanding of ALS and other diseases in which we routinely use histopathological analysis.
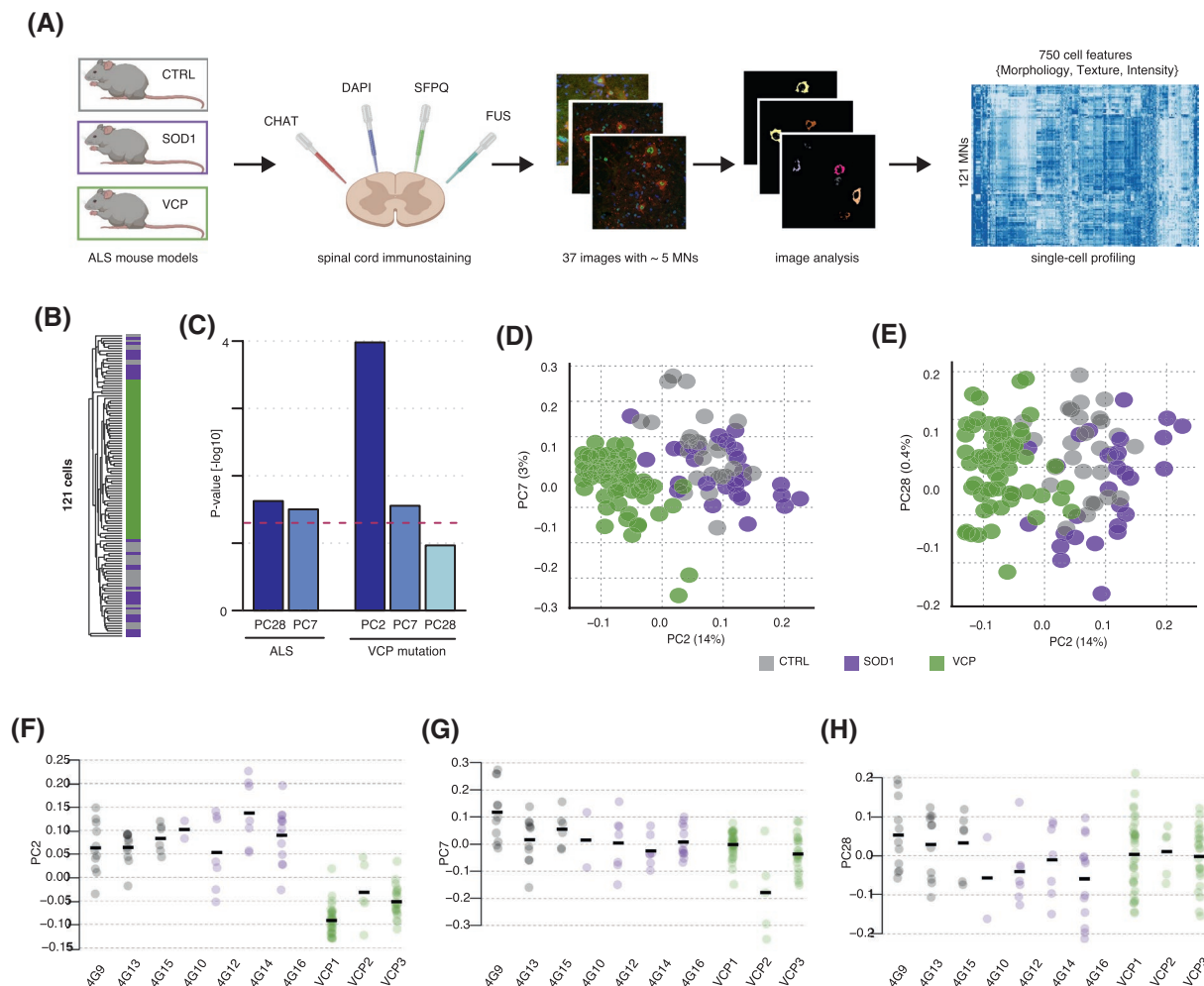
2 | RESULTS

2.1 | Automated phenotyping of spinal cord MNs from transgenic ALS mouse models

Large variability in phenotype between transgenic ALS mouse models is well described in the literature (13). We previously showed that VCP- and SOD1-mutant ALS mouse models exhibit distinct pathological phenotypes in terms of FUS and/or SFPQ mislocalization (1,2). In these studies, we analysed spinal cord sections from control, SOD1- and VCP-mutant mouse models immuno-labeled for FUS and SFPQ, where nuclear and cytoplasmic compartments were manually identified with DAPI and ChAT (1,2). Using these same images (Table S1), we first aimed to test whether spinal cord MNs from these ALS mouse models exhibit common cellular phenotypes which were not captured by our previous analyses. In order to achieve this, we moved beyond single-protein localisation analysis at the population average level to comprehensive single cell analyses by developing a pipeline for automated image segmentation and morphological profiling combining the open-source Ilastik and CellProfiler softwares (11,14) (Figure 1A and Figure S1A). This allowed for the rapid, automated identification of distinct cellular compartments, and extraction of 750 morphological features for every MN identified (n = 121), thereby greatly improving processing speed and throughput compared to manual processing. Features included fluorescent intensity data and signal distribution (for FUS, SFPQ, ChAT and DAPI), shape and morphometric descriptors (size, shape, perimeter and texture of subcellular structures including the nucleus and cytoplasm).

We next examined whether phenotypic states either associated with ALS-mutant cells or control cells are accompanied by reproducible changes in morphological descriptors. Here we hypothesise that information related to the disease state is self-contained in high-content microscopy data obtained from nuclear and cytoplasmic compartments, thus providing a high-resolution catalogue of specific features of cellular state. Unsupervised hierarchical clustering (Euclidean distance and average clustering) of the 121 MNs using the 750 measured morphological features segregated cells based on the presence or absence of the VCP-mutation rather than segregating SOD1- and VCP-mutant cells together (Figure 1B). This suggests that the ALS phenotype associated with the VCP mutation, hereafter referred to as *vcpALS* phenotype, dominates a common ALS phenotype shared between SOD1 and VCP mutant MNs, hereafter referred to as *comALS* phenotype. Hierarchical clustering is, however, often dominated by a single morphological profile and thus can fail to reveal subtle but relevant features in the data (15). In contrast to this, singular value decomposition (SVD) permits deconvolution and ranking of orthogonal morphological profiles (16). Indeed each principal component obtained from the SVD analysis is the weighted sum of all ~800 available microscopy measurements, which capture -in the case of this study- a characteristic cell phenotype. Here we find that the information contained in the data is well distributed among the 121 components derived from SVD analysis (Shannon Entropy = 0.649), and that the

FIGURE 1 High-content MNs profiling captures cell features that discriminate the ALS phenotype in mice spinal cord sections. (A) Workflow for generating morphological profiles of MNs in ALS pathological spinal cord sections. Spinal cord sections from SOD1^{G93A} mutant (n = 4 mice), mutant human VCP^{A232E} (n = 3 mice), and wild-type (n = 3 control mice) mice were co-stained against ChAT, DAPI, SFPQ and FUS and then imaged. A combination of *Ilastik* (11) and *CellProfiler* (14) enables automatic identification and feature extraction from cytoplasmic and nuclear compartments. (B) Unsupervised hierarchical clustering of 121 cells using 750 scaled measurements shows evidence for VCP-mutation dominant signal over SOD1-mutation in ALS MNs populations. Grey bars = control MNs; magenta bars = SOD1 MNs; green bars = VCP MNs. Euclidean distance and average clustering method. (C) Barplots showing the association between principal components and either common ALS (comALS) or VCP (vcpALS) phenotype. Linear mixed effects analysis of the relationship between comALS or vcpALS phenotypes and each of the 31 first principal components accounting for idiosyncratic variations due to animals shows significant association of PC7 and PC28 with comALS phenotype and PC2 only with vcpALS phenotype. (D, E) SVD performed on 750 standardized measurements across 121 cells. Cells are plotted by their coordinates along PC2 (14% of variance), PC7 (3% of variance), and PC28 (0.4% of variance). Colors of data points indicate similar mouse genetic background: control (grey), SOD1 mutant mice (magenta), and VCP mutant mice (green). (F–H) Individual cell coordinates along PC2, PC7 and PC8 grouped by animal and colored according to genetic background

first 31 components capture 90% of the variance in the data (Figure S2). Using linear mixed modelling we find significant association between principal component (PC) 7 (3% of variance) and PC28 (0.4% of variance) with *comALS* phenotype (SOD1 and VCP mutation), while PC2 (14% of variance) is strictly associated with *vcpALS* phenotype (Figure 1C). While this analysis confirms the dominance of morphological changes associated with the VCP mutation, as further visualised in the scatter plots (Figure 1D–F), it indicates that SOD1- and VCP-mutant cells together exhibit subtle and consistent morphological characteristics

likely to reflect common ALS attributes and that are captured by the aforementioned 800 measurements (Figure 1G,H).
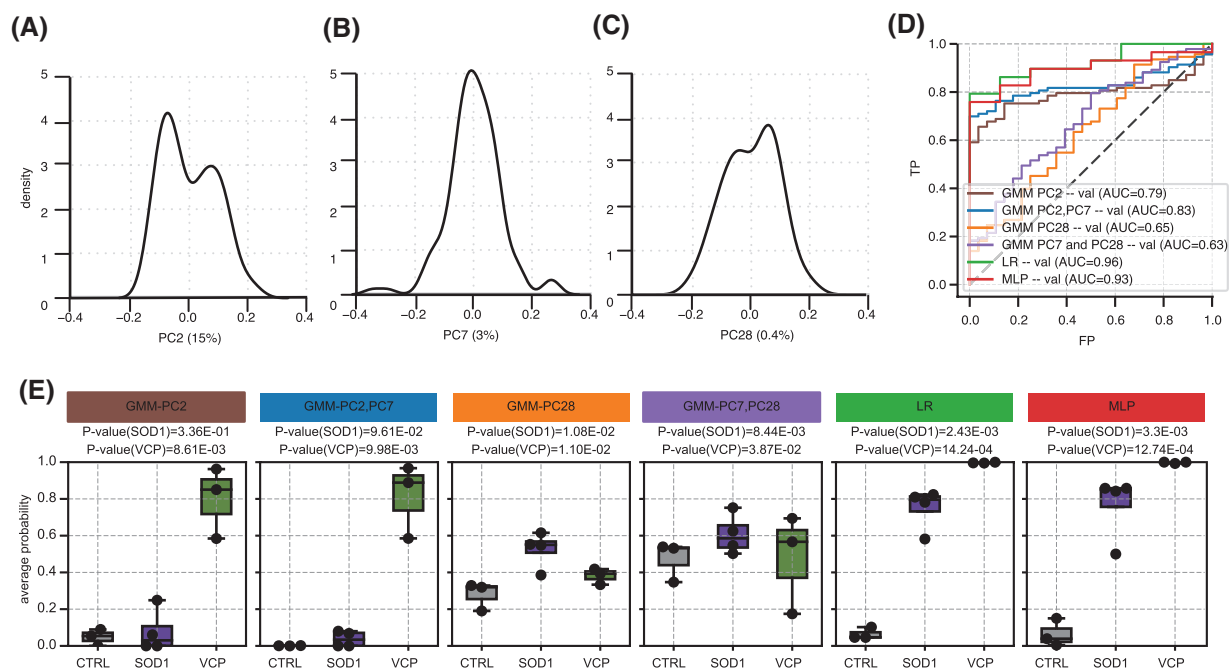
2.2 | Machine learning methods can identify common phenotypes between SOD1 and VCP mutant ALS mouse models

Having established that high-content MNs profiling data from fluorescence microscopy can capture subtle but consistent phenotypes across VCP and SOD1 genetic

backgrounds, we next aimed to train and compare different machine learning methods to automatically identify MNs exhibiting this *comALS* phenotype in spinal cord sections. Visual inspection of the distribution of the principal components associated with either *comALS* (PC7 and PC28) or *vcpALS* (PC2) phenotype revealed that PC2 and PC28, but not PC7, exhibit bimodal distributions (Figure 2A–C). This observation suggests the presence of at least two distinct subpopulations among the 121 cells. Here we propose that MNs populations can be described in terms of a mixture of two phenotypically distinct subpopulations, namely "healthy" and "sick (diseased)" cells (hereafter referred to as **disMNs**). Given that healthy cells can exist in spinal cord sections of VCP/SOD1 mutant mice, we selected the probabilistic label-independent Gaussian Mixture Model (GMM) classifier, and compared it with two label-dependent classifiers, namely Logistic Regression (LR) and Multi-Layer Perceptron (MLP). We tested GMM with the following combinations of principal components selected based on their association with either *comALS* or *vcpALS* phenotype: *GMM-PC2*, *GMM-PC2,PC7*, *GMM-PC7,PC28*, and *GMM-PC28*. Importantly the a priori distinction between healthy and sick cells is only required for the label-dependent methods (LR and NN) and is based on tissue origin, for example MNs were labeled as sick when they originated from SOD1- or FUS-mutant mice, and healthy otherwise.
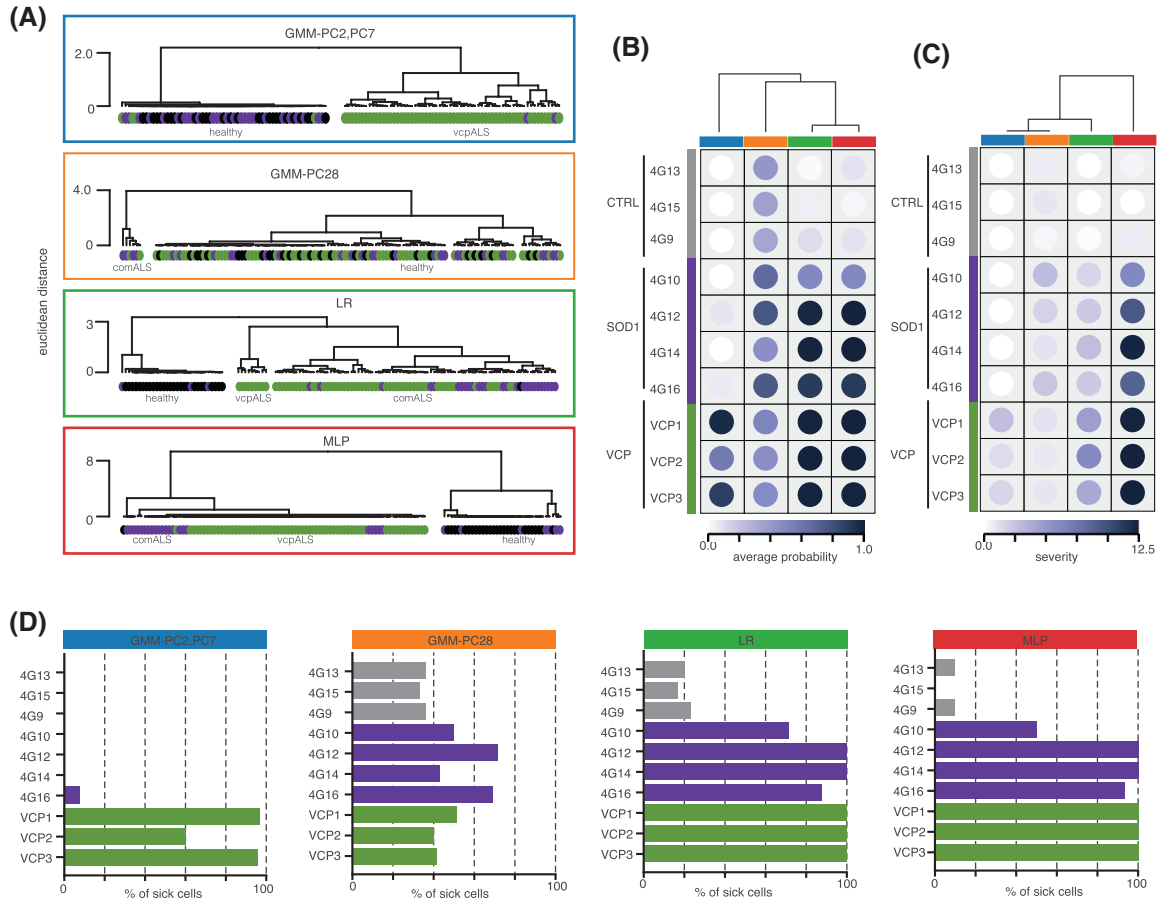
When considering accuracy in the ability of the classifiers to discriminate VCP- and SOD1-mutant cells from control cells, LR and MLP performed equally well (AUC_{LR} = 0.96 and AUC_{MLP} = 0.93), while GMM-derived methods resulted in lower scores overall ($AUC_{GMM-PC2,PC7}$ = 0.83 > $AUC_{GMM-PC2}$ = 0.79 > $AUC_{GMM-PC28}$ = 0.65 > $AUC_{GMM-PC7,PC28}$ = 0.63; Figure 2B). AUC (Area Under the Curve) and other related scores are widely used to rank classifiers, however these need to be treated with caution in the current context given the existence of healthy cells in ALS-derived tissue sections (and indeed theoretical disMNs in control mice). We therefore assigned each animal a probability to be sick using the mean probability of their respective cells, as provided by each classifier, and tested whether the VCP and/or SOD1-mutant group exhibited significantly higher probability to be sick compared to the control group. Comparison of the classifiers' ability to discriminate control from ALS-mutant animals using disease probability is expected to better reflect the biological significance of these classifiers. Using this method we showed that only the VCP-mutant group exhibited significantly higher disease probability than the control group when probabilities were derived from either *GMM-PC2* or *GMM-PC2,PC7* ($P_{GMM-PC2|VCP}$ = 8.61e-03, $P_{GMM-PC2|SOD1}$ = 3.4e-01,

FIGURE 2 Machine-learning based methods enable automated discrimination of VCP- and SOD1-mutant groups from control groups. (A–C) Density distributions of MNs contributions on PC2 (15% of variance), PC7 (3% of variance) and PC28 (0.4% of variance) respectively. (D) Performance analysis of each classifier in their ability to discriminate ALS-mutant cells from control cells using receiver operating characteristic (ROC) curves and areas under the curves (AUC). (E) The ability of each clustering algorithm to detect SOD-1 or VCP-mutation effect is assessed by comparing the disease probabilities of either the SOD1- or VCP-mutant groups with those of the control group. The disease probability of each animal is obtained by using the mean probabilities of its cells to be sick according to (1) *GMM-PC2,PC7*, (2) *GMM-PC28*, (3) Logistic regression or (4) Multilayer perceptron classifiers. Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. Dots are the animal disease profile. *p*-values obtained from Welch's *t* test

FIGURE 3 Comparisons of the different classifiers for predicted percentage of sick cells, disease probability, and disease severity. (A) Comparison of the disease profile similarity among the ALS-mutant cells versus control cells as obtained from *GMM-PC2-PC7* (blue rectangle), *GMM-PC28* (orange rectangle), LR (green rectangle) or MLP (red rectangle) classifiers using unsupervised hierarchical clustering of the 121 cells. Grey circles = control MNs; magenta circles = SOD1-mutant MNs; green circles = VCP-mutant MNs. (B and C) Heatmap of the animal-level disease probabilities and disease severity scores as predicted by *GMM-PC2-PC7* (blue bar), *GMM-PC28* (orange bar), LR (green bar) or MLP (red bar) classifiers. The disease probability and disease scores of each animal are obtained by using the mean probabilities to be sick and severity scores respectively of its cells. The different classifiers are hierarchically clustered using the average algorithm on the Euclidean distance between the disease probabilities or disease severity scores across the 10 animals. (D) Bar plots displaying the animal-level percentage of sick cells as predicted by *GMM-PC2,PC7*, *GMM-PC28*, LR or MLP classifiers. Grey bars = control mice. Magenta bars = SOD1-mutant mice. Green bars = VCP-mutant mice

$P_{\text{GMM-PC2,PC7|VCP}} = 9.9\text{e-}03$, and $P_{\text{GMM-PC2,PC7|SOD1}} = 9.6\text{e-}02$; Figure 2E). Furthermore, *GMM-PC28* resulted in a significant difference between control and SOD1-mutant groups, and to some extent between VCP-mutant and control groups ($P_{\text{GMM-PC28|VCP}} = 1.1\text{e-}02$, $P_{\text{GMM-PC28|SOD1}} = 1.0\text{e-}02$). In summary, only *GMM-PC28* enabled recovery of *comALS* effect, however at higher significance for SOD1 group. In contrast, both LR and MLP classifiers resulted in large significant differences between control and ALS mutant groups irrespective of their genetic backgrounds ($P_{\text{LR|VCP}} = 1.4\text{e-}04$, $P_{\text{ML|VCP}} = 1.2\text{e-}04$, $P_{\text{LR|SOD1}} = 2.4\text{e-}03$ and $P_{\text{MLP|SOD1}} = 3.3\text{e-}03$). These results suggest that PC2 and PC7 capture a cellular phenotype characteristic of *vcpALS*. Furthermore, MLP, LR and to some extent *GMM-PC28* are able to capture a unifying phenotype amongst VCP and SOD1 mice, i.e., *comALS*. In the following sections we focus on the three *comALS* classifiers, namely LR, MLP, *GMM-PC28,* and compare these with the *vcpALS* classifier *GMM-PC2,PC7* to study their underlying differences.

2.3 | LR classifier best captures the broader ALS phenotype at single-cell level

To describe the disease status of an animal or a tissue section, it is important to acknowledge that two cells (or indeed tissues/animals and patients) with similar disease probability can also exhibit different degrees of aberrant phenotype(s). Here we associated each MN with the following two metrics together forming a per-cell 'disease profile': (1) the disease probability P given the observed phenotype as outputted by the classifiers; (2) a 'severity' score S directly derived from P and expected to reflect on the possibility for a population of cells with identical P to exhibit different degrees of

disease phenotypes (see Materials and Methods, and Figure S3).

We next performed hierarchical clustering (Euclidean distance and average clustering) of the MNs using the 121 disease profiles to test the ability of each classifier to group cells into disease versus healthy groups. We found that the *vcpALS* classifier *GMM-PC2,PC7* segregated cells in two groups according to the VCP genetic background (Figure 3A). As expected, some MNs derived from VCP-mutant mice cluster with cells derived from either control or SOD1-mutant group; these are potentially healthy cells present in mutant mice reflecting the cellular heterogeneity in tissue. *GMM-PC28* classifier segregated MNs in two genetically heterogeneous groups, the smaller group being exclusively composed of cells derived from ALS-mutant mice (4 VCP-mutant cells and 2 SOD1 mutant cells), and therefore called *comALS* group. LR and MLP similarly clusters cells in three large groups, one group almost exclusively composed of cells derived from control mice (and some SOD1-mutant cells), called the healthy group, one group almost exclusively composed of VCP-mutant cells, called the vcpALS group, and the third group being a mixture of SOD1- and VCP-mutant cells, i.e., the comALS group. This analysis confirms that LR, MLP, and to some extent *GMM-PC28*, capture the *comALS* phenotype. This is distinct from the *vcpALS* classifier *GMM-PC2,PC7* that discriminates VCP-mutant cells from the control and SOD1-mutant cells. Notably, using LR we can clearly separate the control group from the vcpALS/comALS groups with the highest degree of confidence (i.e., separation between the groups), whilst MLP groups cells in three clusters, healthy, vcpALS and comALS. This indicates that the MLP classifier first captures vcpALS phenotype whilst LR, and to some extent *GMM-PC28*, first capture the comALS phenotype.

From the two per-cell disease score metrics, we next derived for each animal (1) the animal probability to be sick, (2) the animal disease severity, and (3) the animal percentage of disMNs, and used them to quantify the similarity between the four classifiers *GMM-PC2,PC7*, *GMM-PC28*, LR, and MLP. Using hierarchical clustering based on the per-animal probability to be sick further confirmed the largest differences between the vcpALS classifier *GMM-PC2,PC7* and the three comALS classifiers *GMM-PC28,* LR, MLP are similar (Figure 3B). When the per-animal disease severity was used to compare the classifiers however, MLP was the most divergent classifier, exhibiting large severity for all ALS mutant mice irrespective of their genetic background (Figure 3B). This indicates that although MLP predicts large differences in disease profiles between vcpALS and comALS groups of cells, when looking at the per-animal severity, both VCP- and SOD1-mutant animals exhibit large severity.

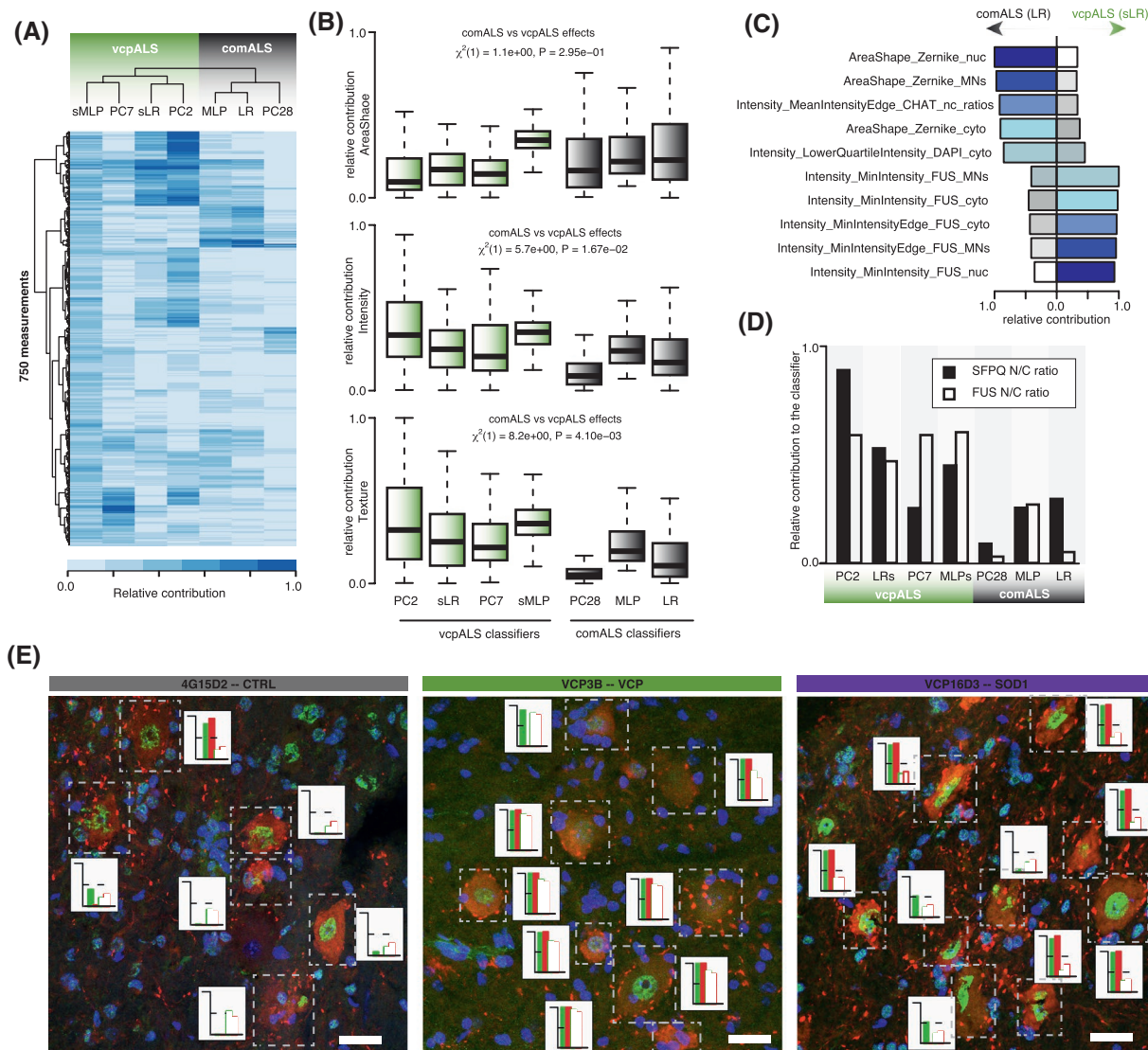Finally we looked at the predicted percentage of disMNs obtained by each classifier and found that the vcpALS classifier *GMM-PC2,PC7* predicted disMNs only in VCP-mutant animals (Figure 3C), while the three comALS classifiers (LR, MLP and *GMM-PC28)* predicted disMNs in ALS-mutant animals (Figure 3A). Notably *GMM-PC28* predicted disMNs across all animals although with higher percentages in SOD1-mutant mice and VCP-mutant mice.

In summary, this analysis reveals that MLP, LR and to some extent GMM-PC28 deliver similar results in terms of their ability to discriminate ALS-mutant mice from control animals (similar per-animal scoring metrics). However, for individual cell scoring metrics, LR is the classifier which best captures measurements associated with comALS phenotype rather than dominant vcpALS phenotype, as evidenced by the large component driven by the VCP mutation in MLP classification.

2.4 | Morphometric descriptors are top classifiers for ALS phenotypes

The finding that LR and MLP best capture *comALS* cell phenotype prompted us to study what combinations of measurements actually carry the relevant information for ALS disease. We therefore looked into which measurements contribute most to the *vcpALS* versus *comALS* classifiers. In order to exclude the possibility that divergence in the top contributors between the *vcpALS* versus *comALS* classifiers could stem from the different mathematical foundations (GMM, LR and MLP), we trained two additional classifiers (one LR and one MLP hereafter called sLR and sMLP) with a subset of the data composed of control and VCP-mutant cells only, to specifically learn the *vcpALS* phenotype using these methods (Figure S4A,B). Hierarchical clustering of the classifiers using the predicted disease probability scores across the 121 MNs confirmed the similarity between *GMM-PC2,PC7*, sLR and sMLP in identifying *vcpALS* phenotype, compared with MLP, LR and *GMM-PC28* in identifying *comALS* phenotype (Figure S4C). Remarkably, although the *comALS* or *vcpALS* classifiers arise from diverse mathematical foundations, the relative contribution of the 750 measurements are very similar among these two groups of classifiers (Figure 4A). This result indicates that the cell measurements which lead to the grouping in *comALS* or *vcpALS* groups are robust enough to be independent of the modeling approach.

We next analysed how each of the three categories of measurements, namely area-shape, texture and intensity, contributed to these classifiers using LMM (see Material and Method). We found that area-shape related measurements exhibited higher (although non significant) relative contribution to the *comALS* classifiers, while texture- and intensity-related measurements contributed more significantly to the *vcpALS* classifiers (Figure 4B). Extraction of the five top contributors to either LR or

FIGURE 4 Area-shape related measurements contribute to the identification of ALS-mutant cells. (A) Heatmaps of the relative contribution of the 750 cell measurements of the different classifiers when applied or to the distinct principal components (PCs) to which GMM classifiers are fitted. The different classifiers/PCs are hierarchically clustered using the average algorithm of the euclidean distance between the relative contributions across the 750 measurements. vcpALS classifiers (green) discriminate VCP-mutant from SOD1-mutant and control cells. comALS classifiers (grey) discriminate ALS-mutant from control cells. (B) Boxplots showing the distributions of the relative contributions of area-shape (top), texture (middle) and intensity (bottom) related measurements in each classifier or PCs when applied. Linear mixed effects analysis of the relationship between comALS and vcpALS classifiers and the relative contribution of each measurement category to account for idiosyncratic variation due to classifiers. Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. Green = VCP classifiers; grey = ALS classifiers. (C) Barplots showing the relative contribution of the top five measurements in LR (comALS classifier) and sLR (vcpALS classifiers). Zernike moments either in the nucleus or the whole MNs contribute largely to comALS but not vcpALS classifier. Bars are color-coded according to the ranking in contribution for the given classifier, from dark-blue to white for high to low ranking. (D) Barplots showing SFPQ (black bars) and FUS (white bars) N/C ratios relative contribution to the different classifiers or PCs when applied. (E) Examples of motor neurons in the ventral spinal cord of wild-type, VCP^{A232E} and SOD1^{G93A} mice. Images were acquired as confocal z-stacks using a Zeiss 710 confocal microscope with a z-step of 1 μm, and processed to obtain a maximum intensity projection. MN cytoplasm is identified by ChAT immunofluorescence, nuclei are counterstained with DAPI, and SFPQ proteins are localised with immunofluorescence (green channel). For each MNs the disease probability is obtained either with comALS classifiers (green/red filled bars for LR/MLP classifiers respectively) or vcpALS classifiers (green/red empty bars for sLR and sMLP respectively). Scale bar = 26 μm

sLR further revealed that the area-shape-related Zernike moments are indeed in the top five contributors of the *comALS* but not *vcpALS* classifiers (Figure 4C), as further confirmed in MLP versus sMLP top five contributors (Figure S4D).

We previously showed that MNs from VCP-mutant mice exhibit reduced N/C ratios of FUS and SFPQ proteins and that SFPQ is also mislocalized in SOD1-mutant ALS models, while FUS is not (1,2). Specifically looking at the relative contribution of SFPQ and FUS

FIGURE 5 The Zernike moments capture aberrant cell behaviour in human ALS pathological post-mortem tissues. (A) Workflow for generating morphological profiles of MNs in ALS pathological spinal cord sections. Spinal cord sections from healthy donors and sporadic ALS patients were co-stained against ChAT, DAPI, SFPQ or FUS and then imaged (see Table S2 for metadata details). A combination of *Ilastik* (11) and *Cell Profiler* (14) enables automatic identification and feature extraction from cytoplasmic and nuclear compartments. (B) Unsupervised hierarchical clustering using 800 scaled measurements of the 66 cells co-immunolabeled with SFPQ or of the 366 cells co-immunolabeled with FUS do not show evidence for sALS phenotype. Grey bars = control MNs; green bars = sALS MNs. Euclidean distance and average clustering method. (C and D) The ability for LR and MLP clustering algorithms to detect the sALS effect in either SFPQ-stained data (C) or FUS-stained data (D) is assessed by comparing the disease probabilities of either the sALS group with those of the control group. The disease probability of each individual is obtained by using the mean probabilities of their cells to be sick according to (1) LR or (2) MLP classifiers. Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. Dots are the individual disease profile. *p*-values obtained from Welch's t test. (E) Barplots showing SFPQ (*left*) and FUS (*right*) N/C ratios relative contribution to the different classifiers or PCs when applied. (F) Boxplot showing the SFPQ nuclear/cytoplasmic (N/C) ratio per cell from three healthy donors and three sALS patients either grouped according to the patient-based labeling (*left*) or grouped according to the LR classifiers (*right*). Grey dots = control MNs; green dots = sALS-derived MNs. *p*-value obtained using Welch's t test. (G) Boxplot showing the FUS nuclear/cytoplasmic (N/C) ratio per cell from eight healthy donors and ten sALS patients either grouped according to the patient-based labeling (*left*) or grouped according to the LR classifiers (*right*). Grey dots = control MNs; green dots = sALS-derived MNs. *p*-value obtained using Welch's t test. (H and I) Boxplots showing the distributions of the relative contributions of area-shape (light grey), texture (grey) and intensity (dark grey) related measurements in LR and MLP classifiers either using SFPQ data (H) or FUS data (I). Z-scores obtained by a permutation test to assess the significance of the measurement category to each classifier are shown above each boxplot. Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. (J) Barplots showing the relative contribution of the top five measurements in LR in either SFPQ or FUS data. Zernike moments either in the nucleus or the whole MNs contribute largely to both LR classifiers. Bars are color-coded according to the ranking in contribution for a given classifier, from dark blue to white for high to low ranking

N/C ratios' to the classifiers showed a stronger impact on the vcpALS classifiers compared to the comALS classifiers (Figure 4D and Figure S4E,F). This corroborates our previous findings that showed increased protein mislocalization in VCP-mutant compared to SOD1-mutant ALS mouse models (1,2). Visual inspection of randomly chosen tissue sections from example animals further confirmed that aberrant SFPQ protein mislocalization in VCP mutant MNs can easily be captured by eye and that SOD1- and VCP- mutant cells share features which are more challenging to pick up by inspection, hence reinforcing the importance of machine learning methods which can identify subtle changes from high-content data (Figure 4E).

2.5 | Morphometric descriptors also capture phenotypes in human tissue sections from sporadic ALS cases

We next sought to validate our findings in human post-mortem tissue (PMT) from ALS patients. We previously analysed spinal cord sections from healthy and sporadic ALS (sALS) PMTs, immuno-labeled for either FUS or SFPQ, where nuclear and cytoplasmic compartments were manually identified with DAPI and ChAT (1,2). Using these same images (Tables S2 and S3) and a combination of Ilastic and CellProfiler (see Materials and Methods and Figure S1B), we automatically identified 432 MNs (66 stained for SFPQ and 366 stained for FUS) from which we extracted 850 morphological features (Figure 5A). Unsupervised hierarchical clustering (euclidean distance and average clustering) using the morphological profiles across either the 66 MNs stained for SFPQ or the 366 MNs stained for FUS did not lead to clear segregation of the cells based on disease status of the human samples (Figure 5B), suggesting that ALS phenotype is more subtle in human PMT compared to ALS mouse models. Nevertheless, SVD analysis of either the SFPQ data or FUS data, both rich-content data as shown by the high Shannon Entropies (Figure S5A), confirmed the presence of disease-related phenotype with the association of three principal components in either SFPQ or FUS data (Figure S5B).

We next trained a combination of GMM with sALS related components, as well as LR and MLP for sALS automated identification. While the different methods showed very similar accuracy for the SFPQ data (Figure S5C), LR and MLP led to the highest accuracy for FUS data (Figure S5D). This would suggest that the sALS phenotype is supported by weighted linear combinations of all measurements in FUS data and that a restricted list of morphological profiles captured by two principal components is sufficient to capture cellular characteristics associated with sALS in SFPQ data. Further comparison of the abilities of each classifier to discriminate

control from sALS groups using the per-individual probability to be sick confirmed our prior finding in mice that LR and MLP are the most efficient classifiers to capture sALS phenotype in high-content imaging data (Figure 5C,D and Figure S5E,F). We next looked at the relative contribution of either SFPQ or FUS N/C ratio in these classifiers. We found that SFPQ localisation has comparatively more weight than FUS in discriminating sALS from control cells in LR and ML (Figure 5E). This result is in line with our previous findings showing that, while both SFPQ and FUS proteins exhibit significant mislocalization in ALS, the extent of mislocalization is larger for SFPQ than for FUS (1,2).

We next sought to assess the effect of MNs re-labelling on the significance of SFPQ and FUS protein mislocalization. We therefore performed Welsh's t-test comparing the protein N/C ratios between two groups, either using the patient-based labels or the healthy versus disMNs labeling derived from the classifiers. Remarkably LR-based re-labeling decreased p-values in both data-sets, confirming the increase in signal-to-noise ratios upon re-classification of the cells (Figure 5F,G).

Using a permutation test, we next tested the relationship between the relative contribution of each category of measurements (area shape, texture and intensity) in the classifiers that best discriminate sALS from control cells, namely LR and MLP. We found that area-shape related measurements significantly contribute to these models in both SFPQ and FUS data (Figure 5H,I). Further looking at the five top relative contributors in LR in either data-sets revealed that area-shape-related Zernike moments are again among the top five measurements that contribute to the LR classifier in both SFPQ and FUS data (Figure 5J), as well as in MLP classifiers (Figure S5G). Cumulatively, these findings suggest that common cellular disease phenotypes occur in human and mouse ALS models and are captured by the Zernike moments.

3 | DISCUSSION

3.1 | Novel pipeline for rapid processing and automated analysis of pathological sections

Our previous studies and those of others have relied on manual segmentation of pathological sections followed by analysis of single measurements. Here we developed a pipeline that couples fluorescence microscopy with highly quantifiable, automated and reproducible morphological profiling of MNs to allow fast and unbiased recovery of single-cell morphology profiles. By applying SVD analysis to the hundreds of cellular measurements derived from this platform we uncovered independent complex cellular phenotypes that associate with ALS MN populations. By showing that ALS phenotypes naturally emerge from such

multichannel fluorescence microscopy high-dimensional data, we confirm the richness of the data over single-cell measurements and its relevance for future investigations.

Unbiased profiling at the single cell level represents a versatile and powerful readout for many cell states capturing the mechanistic details of a wide range of bioactivities (17–19). ALS-related molecular events (e.g. mislocalization of an RBP from the nucleus to cytoplasm) may trigger events that lead to the divergence of outcomes amongst single cells. A recent study has indeed shown that there is substantial intrinsic heterogeneity at the single cell level among ALS cells (20), which suggests that looking at the tissue as a whole might miss valuable insights into the underlying ALS pathobiology. Thus studies such as these highlight not only the importance of evaluating molecular phenotypes at the single cell level, but also the need to have appropriate tools and pipelines for the task. Here we show that machine learning methods such as LR and MLP capture combinations of subtle morphological alterations, without focusing on specific cell measurements, to accurately classify MNs in ALS and control groups. Remarkably the re-classification of MNs into healthy and sick cells using these label-independent classifiers, as opposed to patient-dependent labeling, led to an increase in signal-to-noise ratio for specific markers as well as an increase in significance in protein mislocalization. We also propose a combination of two single-cell disease scoring metrics derived from these classifiers that will be instrumental for future investigations of disease development using histopathological tissue sections.

Thus our study represents a significant advance in ALS histopathology as it presents an integrated pipeline for automatic identification and classification of MNs in pathological sections from raw image processing to the figures presented here. To facilitate the ready application and future development of methods for automatic identifications and scoring of ALS cells, we provide our complete raw image data sets, as well as open-source implementations of the various methods including the open-source workflow to carry out the segmentation, profiling and MNs automated classification. Importantly given reasonable tissue and immunolabeling quality, and a sufficient number of cells, our approach is theoretically applicable to any type of cell irrespective of their size or other characteristics. Given increasing recognition of non-cell autonomous effects in ALS, we are currently adapting this method to analyse ALS astrocytes which have been shown play a central role in the disease progression (21,22).

3.2 | Morphometric descriptors capture common cellular phenotypes across human and mouse ALS models

Most studies related to ALS development report on molecular mechanisms that include aberrant protein homeostasis (ER stress and autophagy) and/or changes in RNA processing (23–26). Although cell morphology is intimately related to the physiological state of cells and intracellular mechanisms (27–29), there is not, to our knowledge, a comprehensive description of cell morphological changes in ALS. Here we show that area-shape related measurements, such as Zernike moments, are key discriminators of aberrant disease phenotypes across human and mouse ALS models. We also show that FUS and SFPQ N/C ratios have varying but lower weights in ALS MNs classifiers compared to the cellular characteristics described above. Zernike moments are the mapping of an image onto a set of complex orthogonal Zernike polynomials (30) and have been proposed to provide a complete basis function representation of cell shape that preserves all the information about the shape of the cell (31). These quantitative single cell shape descriptors have been shown to be reliable indicators of various cancers (28,32,33). Our study suggests that area-shape measurements including Zernike moments may provide a means to read out the phenotypic state of ALS cells which cannot be easily captured by visual microscopic examination, hence the advantage of combining high-content microscopy data with machine learning methods to identify ALS sick cells. An important future avenue of this work will investigate the relationship between these area-shape measurements and intracellular molecular phenotypes.

3.3 | Future perspectives

Our work provides a framework for future studies with larger sample sizes, greater numbers of patients, and higher numbers of markers to rigorously test questions about the nature of heterogeneity between (i) cells within a tissue and (ii) patients. We envision that the ability to identify and interpret information contained in patterns of cellular heterogeneity in pathological sections will provide insights into physiology and disease that may be missed by traditional population-averaged or small-cell number studies. In particular, there are two specific areas where we envisage this method to be particularly informative for ALS, as well for other disorders. The first one is the study of the disease spatio-temporal evolution of the disease at the single cell level. In addition to progression in its severity, ALS also spreads spatially across the rostro-caudal axis (34). At present we do not have good data on what is the progression of changes that occur at the cell level, which could inform on the molecular pathways involved in spreading across anatomical regions. To address this we will be able to use the pipeline we developed here for a future study of serial sections along the spinal cord of individual animals to investigate spatio-temporal progression at different disease stages, as well as comparing the same anatomical regions of the spinal cord across the disease spectrum.

This will offer a unique and crucial dataset on ALS progression. Another area where this method might be used, is in associating molecular phenotypes with cellular phenotypes at the single cell level. Specifically, by being able to reliably and automatically discriminate between cells that present a disease phenotype and cells that are still healthy (even within the same patient) we will be able to effectively create a framework for "digital molecular pathology" for ALS. Utilisation of these automated image analysis techniques can inform on which cells are (i) most likely to be affected at any given time in individual patients, and therefore, (ii) help to determine both the precise molecular pathomechanisms and potentially therapeutically treatable targets. Finally adapting this method to other relevant cell types may shed light on previously unrecognized glial phenotypes in ALS pathology.

AUTHOR CONTRIBUTIONS
Conceptualization, R.L., A.S., R.P.; Formal Analysis, C.H., A.S., R.L.; Investigation, C.H., G.E.T., R.L., A.S. ; Writing – Original Draft, R.L., A.S., R.P.; Writing – Review & Editing, R.L., A.S., R.P., C.H., G.E.T., J.N., H.D. ; Human sample provision, J.N.; Resources, R.P.; Visualization, R.L.; Funding Acquisition, R.L., A.S., R.P.; Supervision, R.L., A.S., R.P.

DATA AVAILABILITY STATEMENT
We provide raw images and complete source code (which is not a software but rather a compilation of R and python) to readily reproduce figures, tables, and other results that involve computation in order to facilitate the development and evaluation of additional profiling methods. We also provide the measurements of each of the ~600 cells whose origins are annotated. The raw images, metadata and single-cell measurements provided as comma-delimited files have been deposited Zenodo under the accession number 3985099, together with the image processing pipelines. The scripts for automated detection of MNS subpopulation can be freely accessed on Github in the following repository: https://github.com/RLuisier/ALSdisMNs.

ORCID
Cathleen Hagemann https://orcid.org/0000-0002-9900-3556
Giulia E. Tyzack https://orcid.org/0000-0001-5730-2112
Helen Devine https://orcid.org/0000-0002-0744-6270
Linda Greensmith https://orcid.org/0000-0002-7839-5052
Jia Newcombe https://orcid.org/0000-0001-9609-7975
Rickie Patani https://orcid.org/0000-0002-3825-7675
Andrea Serio https://orcid.org/0000-0002-7271-1878
Raphaëlle Luisier https://orcid.org/0000-0002-5657-2943

REFERENCES
1. Luisier R, Tyzack GE, Hall CE, Mitchell JS, Devine H, Taha DM, et al. Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. Nat Commun. 2018;9:2010.
2. Tyzack GE, Luisier R, Taha DM, Neeves J, Modic M, Mitchell JS, et al. Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. Brain. 2019;142:2572–80.
3. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453:544–7.
4. Ho R, Workman MJ, Mathkar P, Wu K, Kim KJ, O'Rourke JG, et al. Single-cell RNA-seq analysis of human iPSC-derived motor neurons resolves early and predictive ALS signatures. bioRxiv. 2020:2020.04.27.064584.
5. Loo L-H, Lin H-J, Singh DK, Lyons KM, Altschuler SJ, Wu LF. Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. J Cell Biol. 2009;187:375–84.
6. Rubin H. The significance of biological heterogeneity. Cancer Metastasis Rev. 1990;9:1–20.
7. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc Natl Acad Sci U S A. 2008;105:19306–11.
8. Snijder B, Pelkmans L. Origins of regulated cell-to-cell variability. Nat Rev Mol Cell Biol. 2011;12:119–25.
9. Ščupáková K, Dewez F, Walch AK, Heeren RMA, Balluff B. Morphometric cell classification for single-cell MALDI-MSI. Angew Chem. 2020;132(40):17600–3. https://doi.org/10.1002/ange.202007315.
10. Yamamoto Y, Saito A, Tateishi A, Shimojo H, Kanno H, Tsuchiya S, et al. Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach. Sci Rep. 2017;7:46732.
11. Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, et al. ilastik: interactive machine learning for (bio)image analysis. Nat Methods. 2019;16:1226–32.
12. Von Chamier L, Jukkala J, Spahn C, Lerche M. ZeroCostDL4Mic: an open platform to simplify access and use of Deep-Learning in Microscopy. BioRxiv. 2020. https://doi.org/10.1101/2020.03.20.000133v1.abstract.
13. De Giorgio F, Maduro C, Fisher EMC, Acevedo-Arozena A. Transgenic and physiological mouse models give insights into different aspects of amyotrophic lateral sclerosis. Dis Model Mech. 2019;12(1):dmm037424. https://doi.org/10.1242/dmm.037424.
14. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. Cell Profiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7:R100.
15. Ronan T, Qi Z, Naegle KM. Avoiding common pitfalls when clustering biological data. Sci Signal. 2016;9:re6.
16. Luisier R, Unterberger EB, Goodman JI, Schwarz M, Moggs J, Terranova R, et al. Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion. Nucleic Acids Res. 2014;gkt1415.
17. Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico JA. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. Nat Rev Drug Discov. 2009;8:567–78.
18. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S, et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J Biomol Screen. 2013;18:1321–9.

19. Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc Natl Acad Sci U S A. 2014;111:10911–6.

20. Tam OH, Rozhkov NV, Shaw R, Kim D, Hubbard I, Fennessey S, et al. Postmortem cortex samples identify distinct molecular subtypes of ALS: retrotransposon activation, oxidative stress, and activated glia. Cell Rep. 2019;29(5):1164–77. https://doi.org/10.1101/574509.

21. Boillée S, Vande Velde C, Cleveland DW. ALS: a disease of motor neurons and their nonneuronal neighbors. Neuron. 2006;52:39–59.

22. Valori CF, Brambilla L, Martorana F, Rossi D. The multifaceted role of glial cells in amyotrophic lateral sclerosis. Cell Mol Life Sci. 2014;71:287–97.

23. Hall CE, Yao Z, Choi M, Tyzack GE, Serio A, Luisier R, et al. Progressive motor neuron pathology and the role of astrocytes in a human stem cell model of VCP-related ALS. Cell Rep. 2017;19:1739–49.

24. Mackenzie IRA, Bigio EH, Ince PG. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. Ann Neurol. 2007;61(5):427–34. https://doi.org/10.1002/ana.21147.

25. Serio A, Patani R. Concise review: the cellular conspiracy of amyotrophic lateral sclerosis. Stem Cells. 2018;36:293–303.

26. Zhang K, Donnelly CJ, Haeusler AR, Grima JC, Machamer JB, Steinwald P, et al. The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature. 2015;525:56–61.

27. Gomez JM, Chumakova L, Bulgakova NA, Brown NH. Microtubule organization is determined by the shape of epithelial cells. Nat Commun. 2016;7:13172.

28. Pincus Z, Theriot JA. Comparison of quantitative methods for cell-shape analysis. J Microsc. 2007;227:140–56.

29. Ramkumar N, Baum B. Coupling changes in cell shape to chromosome segregation. Nat Rev Mol Cell Biol. 2016;17:511–21.

30. Hwang S-K, Kim W-Y. A novel approach to the fast computation of Zernike moments. Pattern Recognit. 2006;39:2065–76.

31. Mahi H, Isabaten H, Serief C. Zernike moments and svm for shape classification in very high resolution satellite images. Int Arab J Inf Technol. 2014;11:43–51.

32. Alizadeh E, Lyons SM, Castle JM, Prasad A. Measuring systematic changes in invasive cancer cell shape using Zernike moments. Integr Biol. 2016;8:1183–93.

33. Tahmasbi A, Saki F, Shokouhi SB. Classification of benign and malignant masses based on Zernike moments. Comput Biol Med. 2011;41:726–35.

34. Wong JCT, Concha L, Beaulieu C, Johnston W, Allen PS, Kalra S. Spatial profiling of the corticospinal tract in amyotrophic lateral sclerosis using diffusion tensor imaging. J Neuroimaging. 2007;17:234–40.

35. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671–5.

36. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. Nat Methods. 2017;14:849–63.

37. Loo L-H, Lin H-J, Steininger RJ 3rd, Wang Y, Wu LF, Altschuler SJ. An approach for extensibly profiling the molecular states of cellular subpopulations. Nat Methods. 2009;6:759–65.

38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

Supplementary Material

FIGURE S1 (A) Diagram representing the image processing workflow for mouse tissue immunolabeled for FUS, SFPQ, ChAT and counterstained with DAPI. To facilitate the MN segmentation, we first applied contrast enhancement, background correction and application of gaussian blur and median filters to ChAT stained images using ImageJ (35). Next we used a subset of these preprocessed ChAT images to train a pixel classification algorithm in Ilastik (11) for automated identification of artifacts and MNs. In parallel, automated nuclear segmentation was trained in Ilastik on randomly selected subsets of overlaid ChAT and DAPI images. Finally, all generated segmentations were added to the original dataset as additional channels, and used to identify the cytoplasm of MNs and perform automated densitometry and morphometric measurements in each compartment (cytoplasm, nucleus, whole MNs) in CellProfiler (14). (B) Diagram representing the image processing workflow for human post-mortem tissue immunolabeled for ChAT, FUS or SFPQ, and counterstained with DAPI. Individual channels were preprocessed in ImageJ for contrast enhancement and histogram equalisation. Next automated MN segmentation was trained using a subset of the original images using IIastik based on the three channels (SFPQ, DAPI, ChAT). Next DAPI channel and the masked MNs were provided to CellProfiler for automated segmentation of nuclei and cytoplasm followed by automatic acquisition of single-cell measurements in each compartment

FIGURE S2 Fraction of explained variance captured by the first 31 principal components that captures 90% of the signal. Shannon Entropy of 0.65 indicates that the information in the data is well distributed among the principal components

FIGURE S3 Comparison between disease probability P and disease severity S scores showing how similarly high disease probability can exhibit large differences in disease severity

FIGURE S4 (A, B) MNs predicted probability distribution (*left*), per-animal percentage of sick cells (*centre*), and per-animal disease probability (*right*) as obtained by LR classifier (A) and MLP classifier (B) trained on data censored for SOD1-mutant cells. (C) Heatmap showing the predicted disease probability for the 121 cells. Classifiers are hierarchically clustered using average alog on euclidean distances between disease probability profiles across the 121 cells. Green = *vcpALS* classifiers. Grey = *comALS* classifiers. (D) Barplots showing the relative contribution of the top five measurements in MLP and MLPs i.e. *comALS* versus *vcpALS* classifiers. Zernike moments either in the nucleus or the whole MNs contribute largely to ALS but not VCP classifier. Bars are color-coded according to the ranking in contribution for the given classifier, from dark blue to white for high to low ranking. (E,F) Boxplots showing the relative contribution of FUS (E) and SFPQ (F) intensity related measurements in *vcpALS* versus *comALS* classifiers.

Linear mixed effects analysis of the relationship between the type of classifiers (*comALS* versus *vcpALS*) and the relative contribution of the measurement categories to account for idiosyncratic variation due to classifiers. Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. Green = *vcpALS* classifiers; grey = *comALS* classifiers

Figure S5 (A) Fraction of explained variance captured by the first 24 and 40 principal components that captures 90% of the signal in SFPQ and FUS data respectively. (B) Barplots showing the association between principal components and ALS in SFPQ (*left*) and FUS (*right*) data. Linear mixed effects analysis of the relationship between ALS phenotype and each of the 24 and 40 first principal components to account for idiosyncratic variation due to individuals shows significant association of PC1, PC4 and PC7 and ALS in SFPQ data, and association between PC4, PC15 and PC17 with ALS in FUS data. (C, D) Performance analysis of each classifier in SFPQ data (C) or FUS data (D) in their ability to discriminate sALS MNs from healthy MNs using receiver operating characteristic (ROC) curves and aurea under the curves (AUC). (E, F) The ability for each clustering algorithm to detect sALS effect is assessed by comparing the disease probabilities of sALS group with those of the control group. The disease probability of each individual is obtained by using the mean probabilities of its cells to be sick according to individual classifiers in SFPQ data (B) or FUS data (F). Data shown as box plots in which the centre line is the median, limits are the interquartile range and whiskers are the minimum and maximum. Dots are the individual disease profile. *p*-values obtained from Welch's t test. (G) Barplots showing the relative contribution of the top five measurements in MLP in either SFPQ or FUS post-mortem tissue data. Zernike moments either in the nucleus or the whole MNs contribute largely to both MLP classifiers. Bars are color-coded according to the ranking in contribution for the given classifier, from dark blue to white for high to low ranking

TABLE S1 List of images used for FUS and SFPQ cellular localisation in (1,2); mouse data

TABLE S2 List of images used for FUS and SFPQ cellular localisation in (1,2); human data

TABLE S3 Description of the donors from which PMTs were obtained