NCL-SM: A Fully Annotated Dataset of Images from Human Skeletal Muscle Biopsies

Atif Khan [1,2] Conor Lawless [2*] Amy Vincent [2] Charlotte Warren [2] Valeria Di Leo [2] Tiago Gomes [2]
Stephen McGough [1]

[1]School of Computing, Newcastle University, UK; [2]Wellcome Centre for Mitochondrial Research, Newcastle University, UK; [*]Corresponding Author

Abstract

Single cell analysis of skeletal muscle (SM) tissue is a fundamental tool for understanding many neuromuscular disorders. For this analysis to be reliable and reproducible, identification of individual fibres within microscopy images (segmentation) of SM tissue should be precise. There is currently no tool or pipeline that makes automatic and precise segmentation and curation of images of SM tissue cross-sections possible. We believe that automated, precise, reproducible segmentation is possible by training ML models. However, there are currently no good quality, publicly available annotated imaging datasets available for ML model training. In this paper we release NCL-SM: a high quality bioimaging dataset of SM tissue sections that include $>$ 50k manually segmented muscle fibres (myofibres).

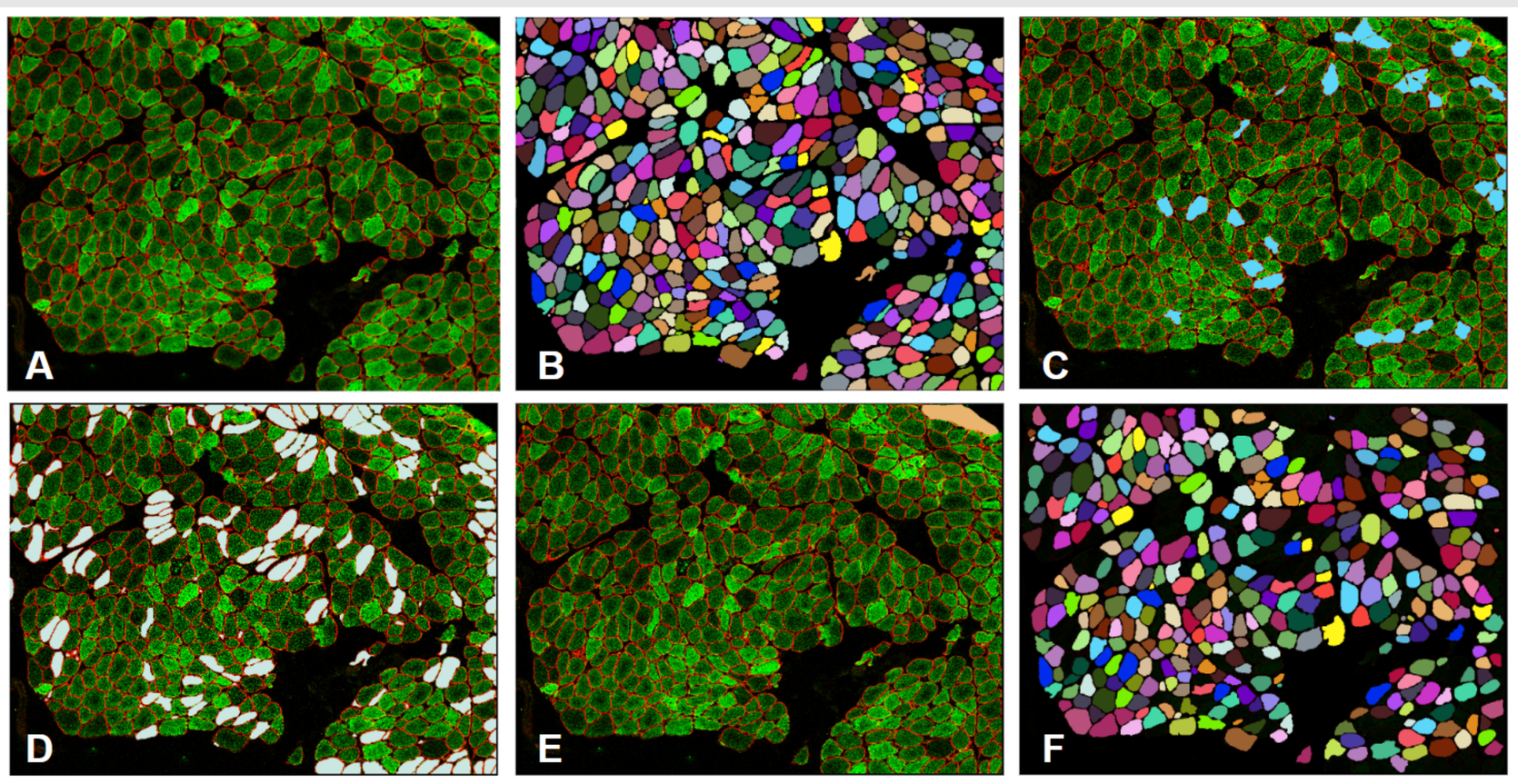


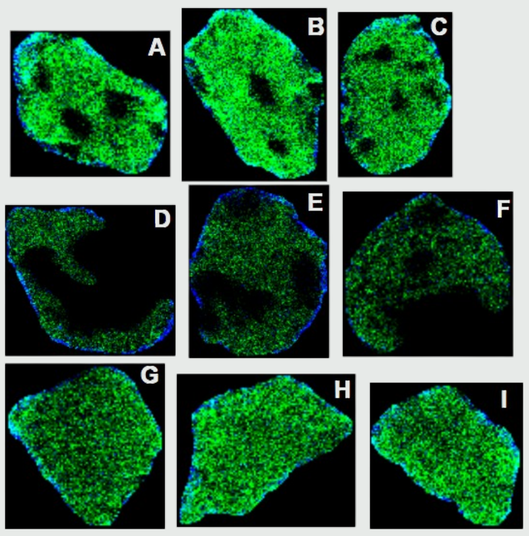
Figure 1. End-to-end SM tissue section image annotation process

Tile A: SM image made by arranging grayscale images of a cell membrane protein marker and mitochondrial mass protein marker into an RGB image where R = membrane protein marker, G = mass protein marker and B = 0; **Tile B**: Manually annotated instance segmentation mask for the image A; **Tile C**: Manually classified mask of Frozen Artefact Myofibres (FAMs) overlaid on the image; **Tile D**: Semi-manually classified mask of Non-Transverse Myofibres (NTMs) mask overlaid on the image; **Tile E**: Manually annotated segmentation mask of Folded tissue Regions (FRs) overlaid on the image; **Tile F**: Final instance segmentation mask of 'Analysable' myofibres made by removing C,D and E from B.

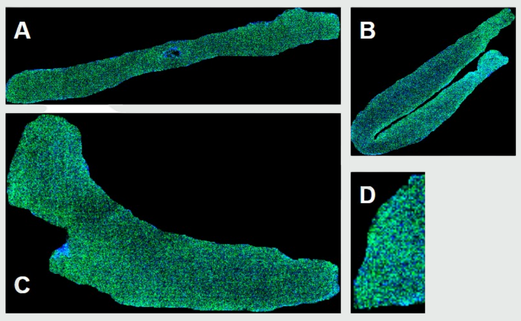
Annotation Protocols

Following are the annotation protocols for the dataset

1. **Myofibre Segmentation**: The protocol for fibre segmentation is i) include all areas within a myofibre that had mitochondrial mass signal, ii) exclude any areas within a myofibre that had myofibre membrane signal and iii) prioritise signal from within myofibre when membrane signal is weak
2. **FAMs**: The protocol for identifying myofibres with freezing artefacts is to look for myofibres described in Figure 2a



(a) **Identifying Myofibre Freezing Artefacts** On top row myofibres A , B and C are typical freezing damaged myofibres resulting in leopard spots pattern, on middle row are the partial myofibres D , E and F that are damaged by freezing, and on bottom row are myofibres G , H and I that are without freezing defects.



(b) **Typical Non-Transverse Sliced Myofibres**: A is a typical elongated myofibre, B is myofibre with unusual convexity, C is myofibre of large area and D is a partial myofibre on the border of the image

Figure 2. Non-analysable myofibres

3. **NTMs**: The protocol for identifying NTM is to look for myofibres described in Figure 2b
4. **FRs** The protocol for identifying FR is to segment areas with overlapping membrane.

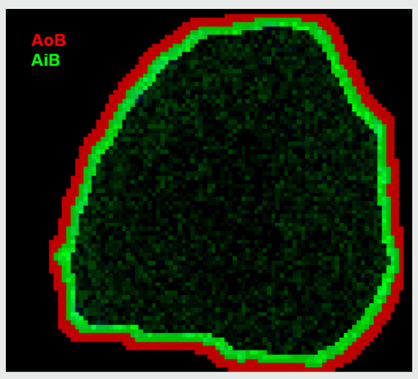
NCL-SM Counts

Imaging Technique	TS Count	Myofibre Count	AM Count	NTM Count	FAM Count	FR Count
IMC	27	22,979	14,841	7,358	780	84
IF	19	27,455	15,953	10,744	758	321
Total	46	50,434	30,794	18,102	1,538	405

Table 1. All annotation counts in NCL-SM dataset. In the table TS, AM, NTM,FAM, FR, IMC and IF stands for Tissue Section, Analysable Myofibre, Non-Transverse Myofibre, Freezing Artefact Myofibre, Folded regions, Imaging Mass Cytometry and Immunofluorescence respectively.

Quality Evaluation Metrics

(1) Myofibre Mass Missed Correlation (r_{AoB}) is the Pearson correlation of myofibre mass pixels missed in Area outside the myofibre Border (AoB) between two duplicate annotations across all myofibres assessed.



$$r_{AoB} = \frac{\sum_{i=1}^{n}(x_{AoBi}-\overline{x_{AoB}})(y_{AoBi}-\overline{y_{AoB}})}{\sqrt{\sum_{i=1}^{n}(x_{AoBi}-\overline{x_{AoB}})^2 \sum_{i=1}^{n}(y_{AoBi}-\overline{y_{AoB}})^2}}$$

$$r_{AiB} = \frac{\sum_{i=1}^{n}(x_{AiBi}-\overline{x_{AiB}})(y_{AiBi}-\overline{y_{AiB}})}{\sqrt{\sum_{i=1}^{n}(x_{AiBi}-\overline{x_{AiB}})^2 \sum_{i=1}^{n}(y_{AiBi}-\overline{y_{AiB}})^2}}$$

Figure 3. Image illustrating Area outside the Border (AoB) and Area inside the Border(AiB) on either side of myofibre annotated border identified by eroding and dilating the border.

(2) Myofibre Membrane Included Correlation (r_{AiB}) is the Pearson correlation of myofibre membrane pixels included in Area inside the myofibre border (AiB) between two duplicate annotations across all myofibres assessed.

(3) IoU (IoU_i) is intersection of overlapping pixels divided by union of all pixels between two annotations of myofibre i. $\overline{IoU}$ is the mean across all n myofibres assessed.

Results: NCL-SM Annotation Quality

Annotations	MF-A	r_{AoB}	r_{AiB}	$\overline{IoU}$	A%(IoU >0.80)	A%(IoU >0.90)	A%(IoU >0.95)
QA-IMC	53	0.99	0.77	0.96	100	100	77.4
QA-IF	23	0.92	0.94	0.96	100	100	74

Table 2. Annotation quality metrics' values for Quality Assurance(QA) human-to-human annotation comparison. In the table MF-A, r_{AoB}, r_{AiB}, $\overline{IoU}$, (A%(IoU >0.80), A%(IoU >0.90), A%(IoU >0.95)), QA-IMC, QA-IF stands for 'Myofibres Assessed', 'Myofibre Mass Missed Correlation', 'Myofibre Membrane Included Correlation', 'Mean IoU', ('Accuracy in terms of % of myofibres meeting IoU thershold of 0.8, 0.9 and 0.95'), 'QA for IMC images' and 'QA for IF images' respectively.

Limitations and Future Work

(1) The **FR annotations in NCL-SM are limited** to 405, which might not be enough for deep learning training. To address this we shall expand NCL-SM to include more high quality annotated data not only from our centre but establish a process in place by which others can add their data which meet the quality standards of NCL-SM.
(2) We will build an automatic ML pipeline to address the challenges highlighted in this paper utilising the NCL-SM.