Proteogenomic single cell analysis of skeletal muscle myocytes

Katherine M. Fomchenko[1,4], Rohan X. Verma[1,4], Suraj Kannan[2], Brian L. Lin[2], Xiaoping Yang[1], Tim O. Nieuwenhuis[1], Arun H. Patil[1], Karen Fox-Talbot[1], Matthew N. McCall[3], Chulan Kwon[2], David A. Kass[2], Avi Z. Rosenberg[1], Marc K. Halushka[1]*

1 Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
2 Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
3 Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA
4 These authors contributed equally to this project

Email Addresses:
Kfomche1@jhmi.edu
Rverma6@jhmi.edu
Skannan4@jhmi.edu
Blin29@jhmi.edu
Xyang15@jhmi.edu
Tnieuwe1@jhmi.edu
Ahanuma2@jhmi.edu
ktalbot@jhmi.edu
matthew_mccall@urmc.rochester.edu
ckwon13@jhmi.edu
dkass@jhmi.edu
arosen34@jhmi.edu
mhalush1@jhmi.edu

* Correspondence and address for reprints to:
Marc K. Halushka, M.D., Ph.D.
Johns Hopkins University School of Medicine
Ross Bldg. Rm 632B
720 Rutland Avenue
Baltimore, MD 21205
410-614-8138 (ph)
410-502-5862 (fax)
mhalush1@jhmi.edu

Abstract

Skeletal muscle myocytes have evolved into slow and fast-twitch types. These types are functionally distinct as a result of differential gene and protein expression. However, an understanding of the complexity of gene and protein variation between myofibers is unknown. We performed deep, whole cell, single cell RNA-seq on intact and fragments of skeletal myocytes from the mouse flexor digitorum brevis muscle. We compared the genomic expression data of 171 of these cells with two human proteomic datasets. The first was a spatial proteomics survey of mosaic patterns of protein expression utilizing the Human Protein Atlas (HPA) and the HPASubC tool. The second was a mass-spectrometry (MS) derived proteomic dataset of single human muscle fibers. Immunohistochemistry and RNA-ISH were used to understand variable expression. scRNA-seq identified three distinct clusters of myocytes (a slow/fast 2A cluster and two fast 2X clusters). Utilizing 1,605 mosaic patterned proteins from visual proteomics, and 596 differentially expressed proteins by MS methods, we explore this fast 2X division. Only 36 genes/proteins had variable expression across all three studies, of which nine are newly described as variable between fast/slow twitch myofibers. An additional 414 genes/proteins were identified as variable by two methods. Immunohistochemistry and RNA-ISH generally validated variable expression across methods presumably due to species-related differences. In this first integrated proteogenomic analysis of mature skeletal muscle myocytes we confirm the main fiber types and greatly expand the known repertoire of twitch-type specific genes/proteins. We also demonstrate the importance of integrating genomic and proteomic datasets.

Key Words: single cell RNA-sequencing; proteogenomics; skeletal muscle, twitch

62 **Introduction**

63 Skeletal muscle is a voluntary, striated muscle found throughout the body with
64 contraction regulated by nerve impulses through the neuromuscular junction (NMJ). Skeletal
65 muscles consist of different fiber types delineated by the isoform of the myosin heavy chain they
66 express, metabolic function, and other properties (1). In humans, slow fibers (type 1) and some
67 fast fibers (type 2A) exhibit oxidative metabolic properties, while fast type 2X fibers exhibit
68 glycolytic metabolic properties (2). Mice have an additional type 2B fast fiber. These fiber types
69 are variable across different muscles of the body reflecting different functional needs (2, 3).

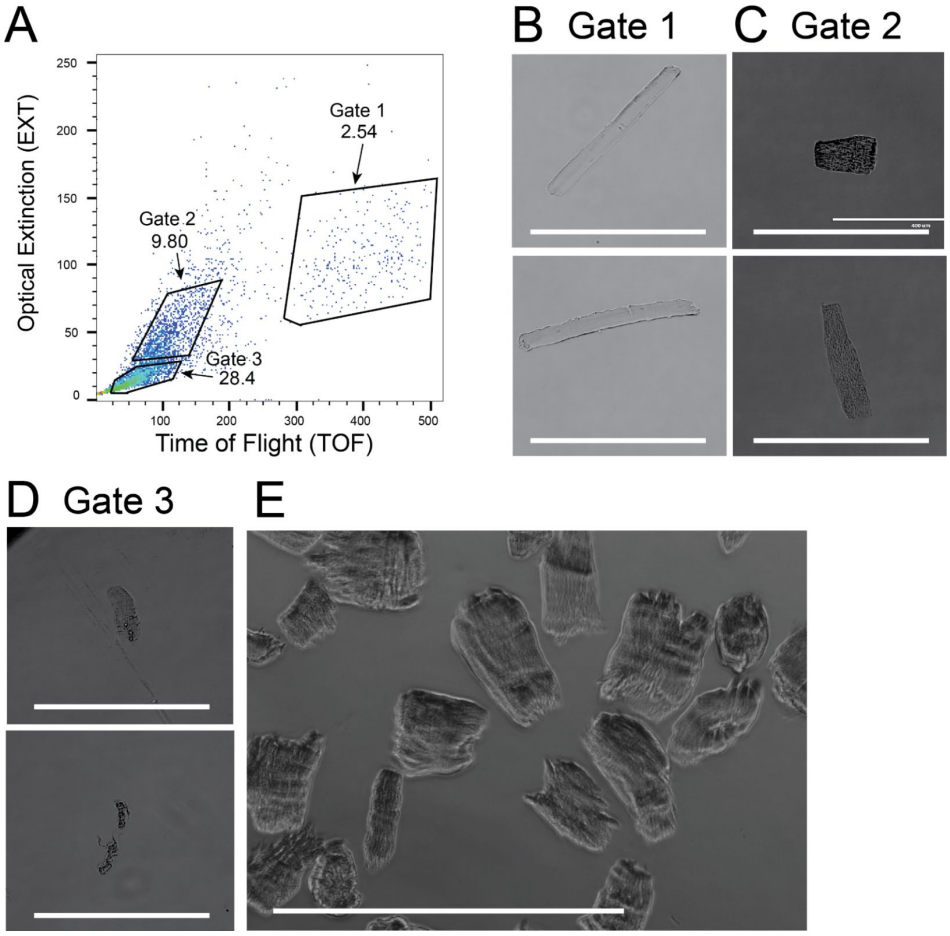
70 Multiple proteins and protein classes vary across fiber types (1, 4). These include
71 isoforms of the myosin heavy and light chains, calcium ATPase pumps, troponin T, and
72 tropomyosin proteins, as well as metabolic proteins, such as pyruvate kinase, GAP
73 dehydrogenase, and succinate dehydrogenase. Beyond these classes, there have been few efforts
74 to catalog the entirety of fast/slow twitch expression differences by proteomics or genomics.

75 Among proteins, the deepest effort, to date, has been the single fiber proteomics work of
76 the Mann laboratory (5, 6). In separate studies of mouse and human single fiber skeletal muscles,
77 1,723 and 3,585 proteins were reported, respectively, many of which were variably expressed
78 among slow and fast twitch fibers. The most comprehensive gene expression study was
79 performed in mice using DNA microarrays across ten type 1 and ten type 2B fibers (7). Single
80 cell RNA-sequencing (scRNA-seq) also has been performed in skeletal muscle and muscle
81 cultures. However, the large size of skeletal myocytes has precluded them from these datasets,
82 which are instead predominately satellite cells, and other supporting cell types (8-15). A recent
83 publication used SMART-Seq to evaluate three fast twitch mouse fibers (16). The totality of

84 these studies strongly suggests there are numerous expression differences between skeletal
85 muscle fiber types and a need for new approaches to capture this diversity.
86 The Kwon laboratory recently developed a large cell sorting method to isolate mature
87 cardiac myocytes (17). We ascertained if this method could be used to isolate the even larger
88 skeletal muscle myocytes for scRNA-seq. Our goal was to combine this genetic data with single
89 cell spatial proteomic data from the Human Protein Atlas (HPA) and an established mass
90 spectrometry human skeletal muscle proteomic dataset for a unique proteogenomic
91 characterization of skeletal muscle expression mosaicism.
92 **Results:**
93 **scRNA-seq and identification of fast/slow twitch fiber types**
94 We performed single cell RNA-seq using the established mcSCRB-seq protocol (18, 19).
95 We recovered data for 763 cells and sequenced to a median depth of 108,110 reads per cell. As
96 we were unsure of where the ideal skeletal myocytes might arise from our flow-sorting method,
97 they were taken from two different gates set on extinction (EXT) always "high" and time of
98 flight (TOF) being both high or low (Supplementary Fig. 1). Additional cells were collected
99 from a pseudo-biopsy approach with fragmented skeletal myocytes (see methods). Preliminary
100 analyses, however, indicated a distinct cluster of cells with a high percentage of mitochondrial
101 reads or otherwise low abundance reads. Notably, almost all of our pseudo-biopsy myocyte
102 fragments and many TOF-low cells fell into this category. These quality control metrics likely
103 indicated poor quality or sheared cells with loss of RNA. Thus, we excluded these cells and
104 narrowed our analysis to the best 171 cells remaining with a median read count of 239,252 per
105 cell.

106 An average of 12,098 transcripts were identified in these cells and all had the expression
107 patterns of mature skeletal myocytes, highly expressing a myosin heavy chain isoform. Because
108 of the narrow focus of this work to delineate cell subtypes and expression variability of just
109 skeletal muscle myocytes, this isolation strategy linked to deep sequencing, proved to be
110 advantageous.

111 We performed PCA of the data, corrected the data for the top 20 PCAs and utilized the
112 top 3,000 variable genes (by +/- standard deviation) to cluster these cell types (Fig. 1a). Three



Supplementary Figure 1. Mouse skeletal muscle myocyte preparation. A) Flow cytometry showing three gated areas representing EXT-high/TOF-high, EXT-high/TOF-low and EXT-low populations of flexor digitorum brevis myocytes. B) Representative images of Gate 1 EXT-high/TOF-high. C) Representative images of Gate 2 EXT-high/TOF-low. D) Representative images of Gate 3 EXT-low. E) Representative image of pseudo-biopsy isolated myocyte fragments. Gates 1 and 2 were used for library preparation. White size bar is 400 µm.

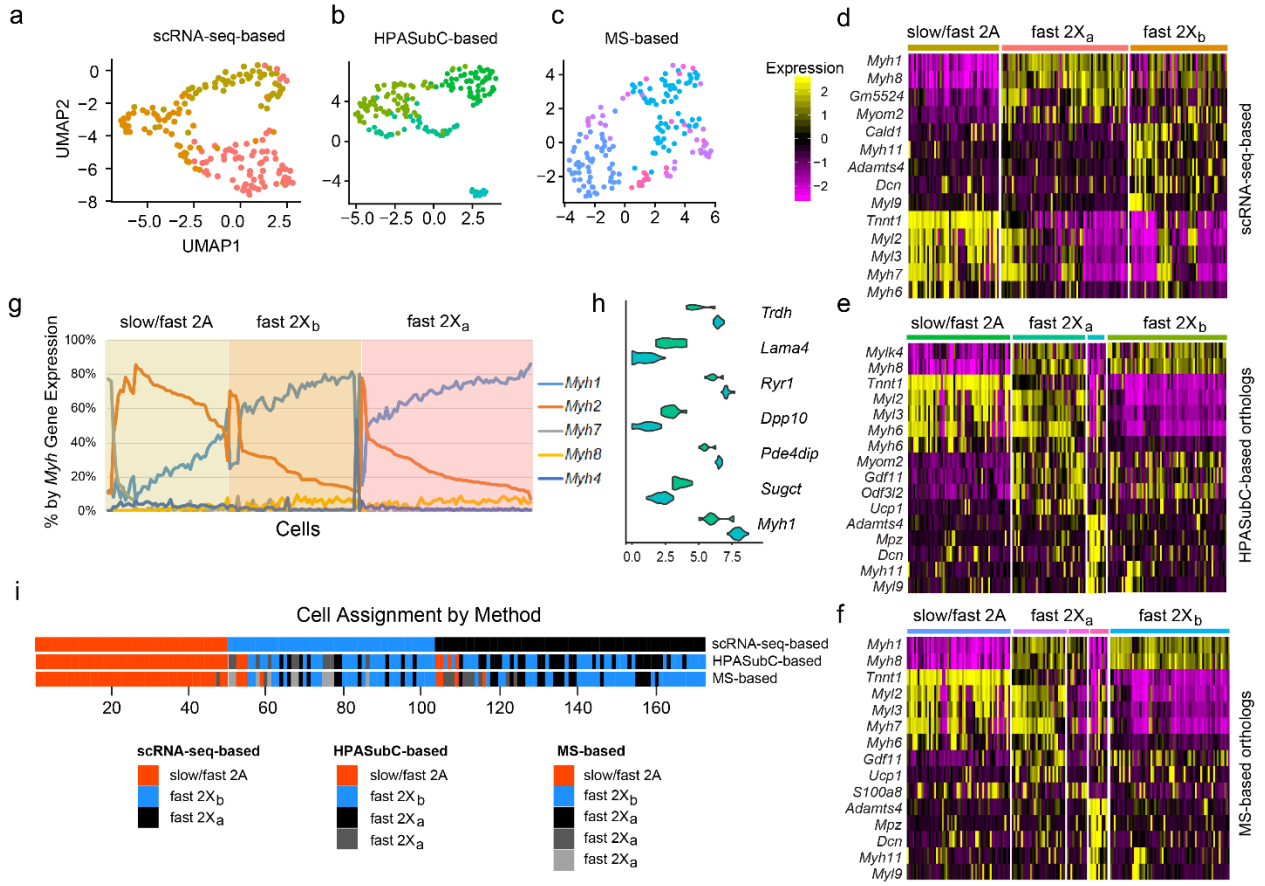


Fig. 1. a) UMAP graph of 171 skeletal muscle cells based on variable gene expression determined by scRNA-seq. **b)** UMAP graph based on mouse orthologous expression of HPASubC variable proteins. **c)** UMAP graph based on mouse orthologous expression of MS variable proteins **d-f)** Heat maps of major genes expression differences between the different fiber types based on the different datasets. **g)** Major myosin heavy chain distributions across the 171 cells as a percentage of each heavy chain. The colored areas are the assignments of each cell based on the scRNA-seq-based data **h)** Violin plots of 7 genes that varied between the two fast $2X_a$ groups in the HPASubC-based data set. **i)** Assignment of each skeletal myocyte to a fiber type across the three methods. Strong agreement existed for the slow/fast 2A cells by any method of analysis

groups were observed in a UMAP dimensionality reduction plot. The first cluster, containing 69 cells (40% of all cells) had elevated expression of *Myh1* and *Myh8* clearly identifying this group as containing fast 2X type cells and denoted as fast $2X_a$ (Fig. 1d). A second cluster (N=53 cells) had slightly more variable *Myh1* and *Myh8* differential expression, but by overall *Myh* gene expression, Fig. 1g, also appeared to be a fast 2X cell type (denoted fast $2X_b$). Of note, *Myh4*¸ a

118 myosin heavy chain associated with fiber type 2B, was elevated in a single cell in this group
119 (Fig. 1g) (3).
120 A third cluster (3) containing 49 cells (29% of the total) was defined by high expression
121 of *Tnnt1* and *Myh2.* A deeper analysis of this group showed that 12 cells had high to modestly
122 elevated *Myh7* expression (a slow-twitch marker), indicating this cluster was a combination of
123 slow-twitch cells and fast 2A fibers (Fig. 1g). The flexor digitorum brevis is a fast twitch
124 muscle, thus the overall distribution of significantly more fast (159) to slow fibers (12) is
125 consistent with expected.
126 Interestingly, the expression patterns of the main fast/slow differentiating *Myh* genes was
127 not as dichotomous as noted in protein based fiber type data (5). Here there were many more
128 cells with intermediate levels and coexpression of *Myh1* and *Myh2* suggesting higher gene
129 plasticity and more cell hybrids (Fig. 1g**)** (3).
130 **HPA-based mosaic protein discovery**
131 To complement variable gene expression data, we generated mosaic protein data by
132 performing an analysis of the IHC-based HPA dataset of skeletal muscle images using the
133 HPASubC suite of tools (20). The HPASubC tool, obtains a selected organs' images from the
134 Human Protein Atlas (HPA) and allows rapid and agnostic interrogation of images for staining
135 patterns of interest. This approach established a protein-based list of mosaically-expressed
136 proteins. Out of 50,351 images reviewed for 10,301 unique proteins, 2,164 proteins had possible
137 mosaic expression in skeletal muscle. Based on the aggregate image scores assigned to each
138 protein, they were subsetted into categories of "real" mosaicism (374 proteins), "likely"
139 mosaicism (1,231 proteins), and "unknown" probability of mosaicism (559 proteins)
140 (Supplementary Data 1, Supplementary Fig. 2). For analysis purposes, we focused on the 1,605

141 proteins that were in the “real” or “likely” categories to reduce the incidence of false positive

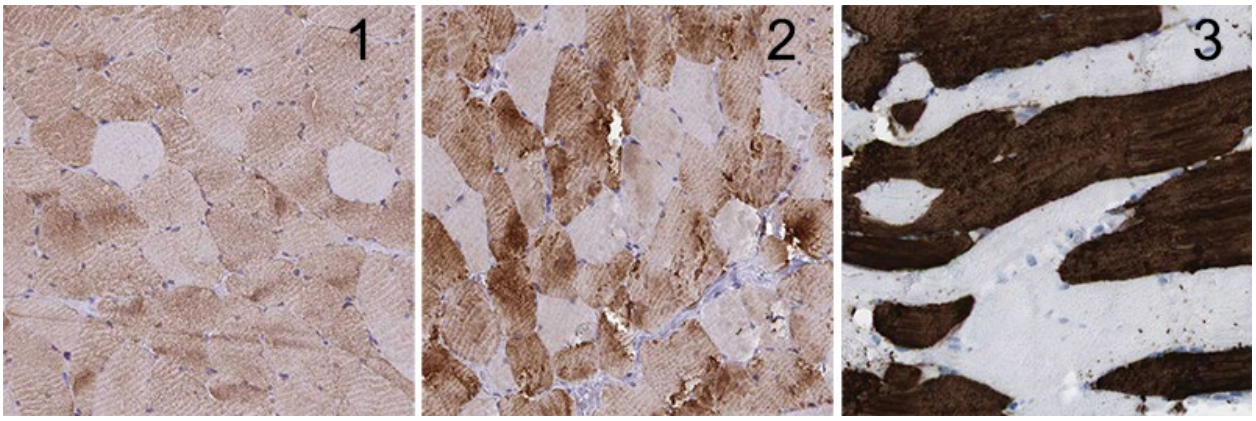
142 staining.

143 This method identified the well-known fiber type specific proteins such as MYH1,

144 MYH2, MY4, MHY6, MHY7, and MYH8 that were categorized as both “real” or “likely” based

145 on staining patterns (Supplementary Data 1). It also identified numerous uncharacterized or

146 poorly characterized proteins, such as the zinc finger proteins ZNF213, ZNF282, ZNF343,



Supplementary Figure 2. Scoring schema for HPASubC-based skeletal muscle mosaicism. A score of 1 indicated an “unknown” mosaicism based on subtle differences in stain intensity, or inconsistent patterns. A score of 2, “likely,” was a clear distinction of staining by myofiber but the staining was not robust. A score of 3 “real” identified clear and robust staining differences by muscle cell. The score was primarily about the pattern and secondarily about the intensity of the staining difference.

147 ZNF350 and ZNF367 all of which had “real” patterns of mosaicism. A limitation of this spatial,

148 IHC-based approach is that each protein image is independent of other proteins. Thus, one

149 cannot identify co-expression patterns to assign proteins to certain fiber types.

150 We therefore investigated how these 1,605 proteins might inform on fiber type of

151 skeletal muscle cells by using this list to subset the orthologous mouse gene data from the

152 scRNA-seq experiment. Using just these orthologous mouse genes, we regenerated the UMAP

153 plot that identified four clusters (Fig. 1b). It essentially recapitulated the fast and slow fibers

154 types noted from the exclusive scRNA-seq data, despite being based on a different set of genes.

155 Uniquely, it subsetted the fast $2X_a$ cluster into two groups, one denoted by high expression of

156 *Myom2* and *Gdf11* and the other denoted by high *Ucp1* and *Adamts4* (Fig. 1e). A t-test of gene

157 expression comparing genes from just these two subsets of the fast $2X_a$ cluster identified

158 multiple genes variably expressed between them (Fig. 1h). Although the cell clustering was

159 generally similar between mouse scRNA-seq gene data and HPASubC data with regard to

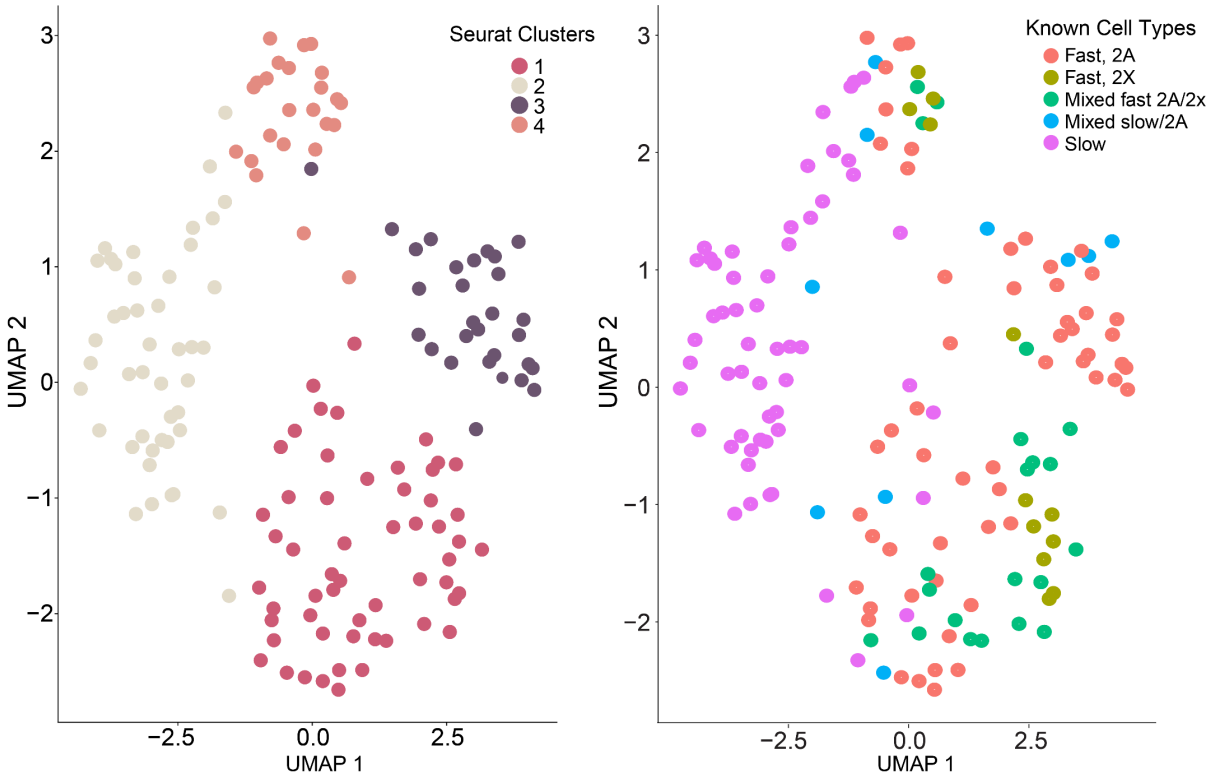
160 slow/fast 2A vs fast 2X, it was unclear which method was more representative. Therefore, we

161 obtained a public MS dataset as a third method to classify slow and fast twitch fibers.

162 **Fast/slow twitch variation by MS-based proteomics**

163 The human skeletal muscle fiber MS data in Murgia et al. is based on 152 fibers from

164 eight donors (5). This dataset had 596 proteins with >2.3 fold variation between type 1 and type



Supplementary Figure 3. UMAP of MS-based protein data by cell type. **A)** Seurat identified four cell clusters. **B)** UMAP was coloured based on cell assignments of Murgia et al. The slow type cells are generally Seurat cluster 2. Fast 2A cells are generally in Seurat cluster 3, although they are also detected in clusters 1 and 3. Fast 2X clusters are predominately in Seurat cluster 1.

165 2A fibers. We analyzed the full LFQ dataset of protein expression and constructed a UMAP plot
166 that showed four clusters (Supplementary Fig. 3). One cluster was composed primarily of slow
167 type 1 fibers and was adjacent to a second cluster with a small mixture of slow and other cell
168 types. Two other clusters were primarily a collection of fast 2X and fast 2A cell types. Similar to
169 the HPASubC approach above, we subsetted the orthologous mouse genes to these 596 proteins
170 to explore cell fiber type assignment.

171 As seen in the UMAP plot, five groups were identified (Fig. 1c). Similar to the other two
172 datasets (scRNA-seq and HPASubC), a slow/fast 2A fiber type was denoted by elevated
173 expression of several genes including *Tnnt1* and *Myl2* (Fig. 1f). One fast 2X fiber group ($2X_b$)
174 was identified by high expression of *Myh1* and *Myh8.* The second fast 2X fiber group was then
175 subdivided into three groups based on alternative elevated expression of genes that include
176 *Gdf11* and *Ucp1* (group 3), *S100A8* (group 4) and *Adamts4* and *Mpz* (group 5). Unlike the
177 protein expression level based UMAP, slow fibers and fast 2A fibers were not distinct.
178 (Supplementary Fig. 3). This difference may be a result of the higher percentage of slow fibers in
179 the MS dataset.

180 **Cross comparisons of the three approaches yield similar cell types.**

181 We identified the cluster assignment of each skeletal muscle cell based on the scRNA-
182 seq, HPASubC, and MS approaches. We then plotted this information to demonstrate the extent
183 to which there was fluidity in assignment by fiber type (Fig. 1i). All but one cell (48/49)
184 assigned to the slow/fast 2A cluster based on scRNA-seq data remained in that cluster using
185 other methods of clustering (HPASubC and MS). An additional 7-8 cells from the fast 2X groups
186 became assigned to the slow/fast 2A cluster using the other methods of cell assignment. Cells
187 moved interchangeably between the fast $2X_a$ and fast $2X_b$ clusters depending on the method used

188 to cluster. We used this information to try and understand what distinguished fast $2X_a$ and fast
189 $2X_b$ clusters.

190 **The $2X_a$ and fast $2X_b$ clusters differ by axonal genes.**

191 To understand if the two fast 2X clusters represent unique cell types, cell states, or some
192 technical division, we performed a differential expression to determine what genes drove their
193 differences. Of 5,260 genes compared, 557 genes were differentially expressed (t. test; adj. p.
194 value <0.01). A Gene Ontology (GO) analysis on the 557 genes identified an enrichment of the
195 cellular component "neuronal synapse," suggesting variability at the NMJ. A further review of
196 the top significant genes showed that >20 genes appear to have neuronal origins (*Cdh4, Cdkl5,*
197 *Cntn4, Dscam, Gabbr2, Kirrel3, Lingo2, Lrp1, L1cam, Nrcam, Ntn1, Ntrk3, Ptprt, Ptpro, Robo2,*
198 *Sdk1, Sema5a, Sema6d, Shank2, Sox5, Tnr,* and *Wwox*). Of these, NRTK3, LRP1, and ROBO2
199 were identified as mosaic in skeletal muscle cells by HPASubC. Additionally, in HPA images,
200 seven orthologous proteins of these "neuronal" genes showed moderate staining, but each of
201 these had a TPM <1 (from GTEx expression data). Only LRP1 was identified in the orthologous
202 MS-dataset. This variability made us wonder how frequently the same genes/proteins were noted
203 to be mosaic by each of the three methods.

204 **There is limited overlap of shared expression information**

205 We compared the 3,000 most variable genes, the 1,605 HPASubC proteins, and the 596
206 MS proteins for shared patterns of mosaicism (Fig. 2). Only 23 genes/proteins were mosaic by
207 all three approaches using the Seurat analysis method (Fig. 2a). An additional 300
208 genes/proteins were shared across two methods, with the most overlap identified between the
209 two datasets with the most genes/proteins. Thus, we reasoned the abundance of genes/proteins
210 by method was a major driver of overlap leading us to focus on the 3,052 transcripts shared by

211 all three approaches, regardless of their mosaic/variable status. This resulted in 157

212 genes/proteins shared across any two methods with the most overlap between the two protein

213 datasets (77 proteins) (Fig. 2c).

214 As so few genes were shared with the protein sets, we wondered if the computational

215 approaches of the Seurat method limited the discovery of the correct variable genes. Therefore,

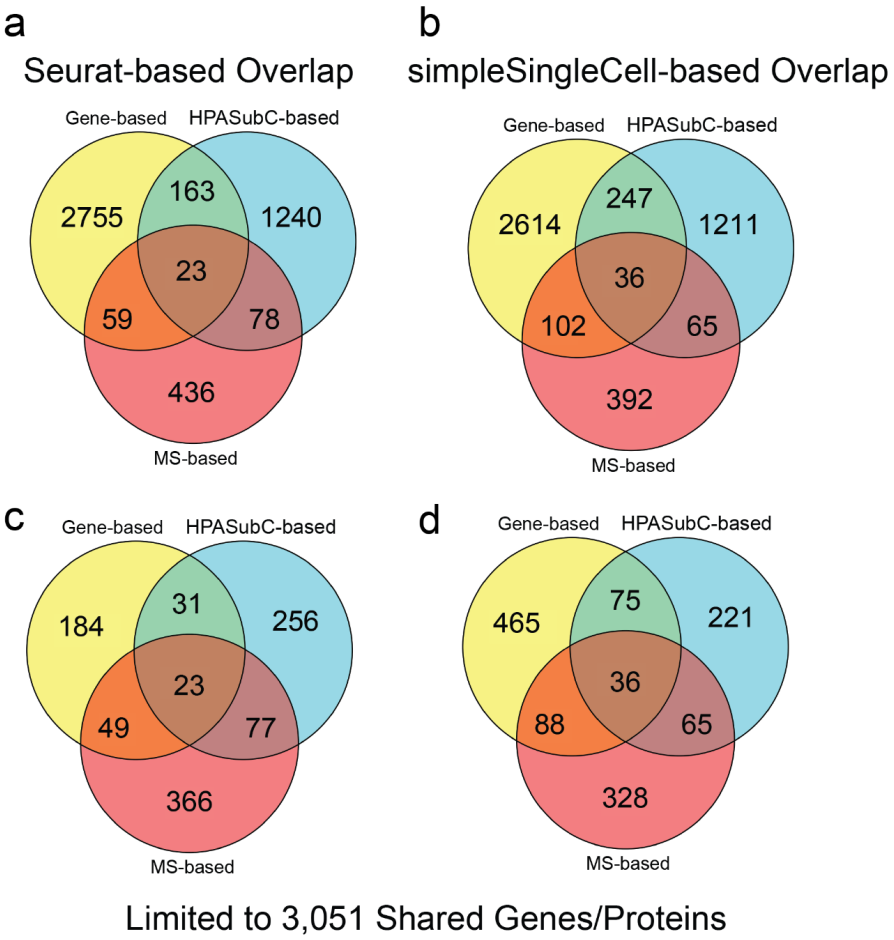


Fig. 2. Venn diagrams comparing the three methods and two analysis types with the full datasets (top) and the limited datasets (below). **a)** A Seurat-based overlap including all mosaic genes/proteins. **b)** A simpleSingleCell-based overlap including all mosaic genes/proteins. **c** & **d)** Seurat and simpleSingleCell-based methods limited to the 3,051 genes/proteins shared across the three studies.

216 we tried a second analysis approach, simpleSingleCell, to identify variable genes (21). By this

217 method, there was an increase (N=36) in overlap of genes/proteins being identified by all three

218 methods and more genes/proteins being identified by two methods (414) (Fig. 2b). Interestingly,
219 comparisons limited to the shared gene/protein list resulted in the highest overlap between the
220 MS- and gene-based datasets (Fig. 2d). A third method of using differential expression on the
221 scRNA-seq data to compare the subset of 12 slow-twitch cells to all fast twitch (2X and 2A) or
222 just fast 2X cells gave equivalent data to the simpleSingleCell approach.

223 **Shared, abundant transcripts by cell type**

224 We then wondered about the extent to which highly abundant proteins/genes were driving
225 our ability to detect mosaic proteins/genes. By normalized read counts of the scRNA-seq data,
226 we determined the 50 most abundant transcripts by the average of each cell type in the three
227 clusters determined by Seurat (Supplementary Data 2). Not surprisingly, the overall most
228 abundant transcripts were *Ttn*, *Acta1* and *mt-Rnr2*. Of the 23 mosaic genes/proteins found by all
229 three methods (using Seurat analysis), only *Myh1* and *Tnnt1* were on the list. Adding the mosaic
230 genes from the simpleSingleCell analysis, seven additional genes (*Mylpf*, *Tnnt3*, *Tmp1*, *Tnni2*,
231 *Eno3*, *Atp2a1* and *Pfkm*) were noted. This overall indicates that most abundant genes (≥41/50)
232 are not consistently mosaic in skeletal myocytes.

233 **Species dichotomy in protein expression patterns**

234 The generally low amount of overlap across the methods was unexpected. We wondered
235 if this discrepancy particularly between the gene and protein data was the result of species
236 differences in twitch type expression. To address this, we investigated staining patterns for three
237 proteins. Two (DCAF11, ENO3) were selected as they had clear mosaic staining by human
238 HPASubC images and no gene variation by Seurat analysis of the scRNA-seq. PVALB was
239 selected for showing variation by the mouse scRNA-seq data, but no variation by HPASubC.

240 DCAF11 was robustly mosaic in human but non-mosaic in mouse. ENO3 was mosaic in
241 both and PVALB was weakly mosaic in human but robustly mosaic in the mouse tissue (Fig. 3).
242 This data suggested that discrepancies may relate to differences in mosaic protein expression
243 between species (DCAF11) and possible technical causes (PVALB). Because ENO3 was mosaic
244 in the mouse skeletal muscle, but not mosaic by Seurat gene expression analysis, we explored if
245 a posttranscriptional form of regulation was occurring.
246

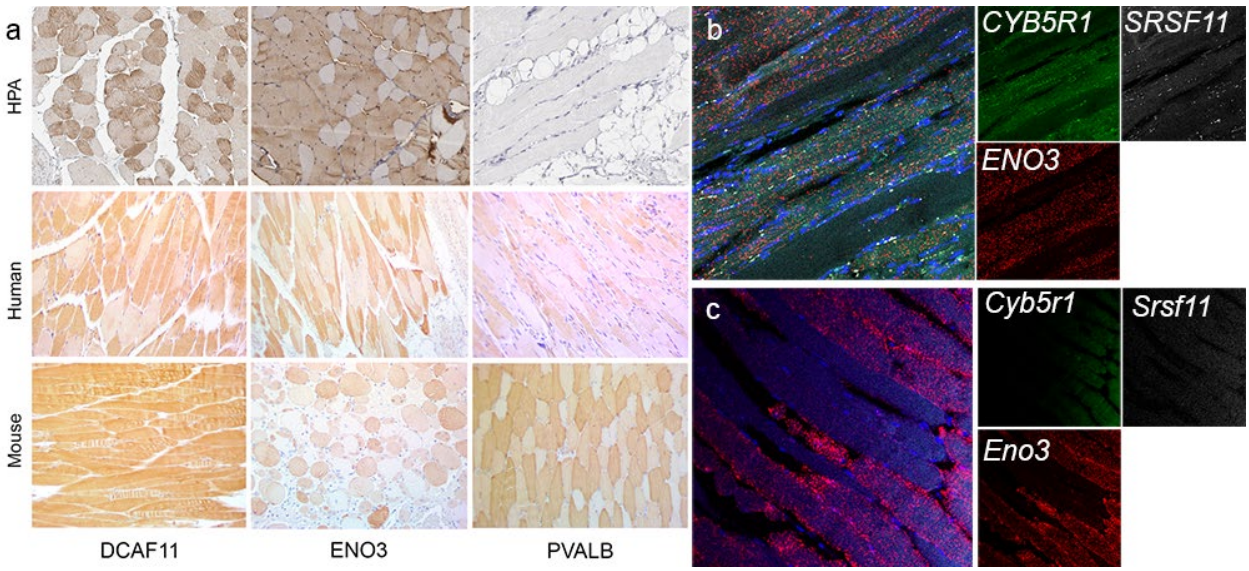


Fig. 3. Representative IHC and RNA-ISH of discrepant proteins and genes. a) HPA images (top row) are mosaic for DCAF11 and ENO3 and negative for PVALB staining. Follow up staining validated the DCAF11 and ENO3 staining while suggesting a subtle mosaicism of PVALB in humans. In mice, ENO3 and PVALB are clearly mosaic, while DCAF11 is not. **b)** RNA-ISH demonstrates co-expression of *CYB5R1* and *ENO3* in a mosaic pattern. **c)** Only *Eno3* was observed (in a mosaic pattern) in mouse muscle by RNA-ISH.

247

248 **RNA-ISH indicates variable mosaicism**

249 We performed RNA-ISH in both mouse and human skeletal muscles for *Eno3*, *Srsf11* and
250 *Cyb5r1*. All of their protein products were mosaic by HPASubC and MS protein expression and
251 had high or reasonably abundant gene expression (6552.8, 19.3, 201.5 pTPM respectively,

252 HPA). None of these genes were variably mosaic in the mouse gene data. We found mosaic co-
253 expression of all three genes in human skeletal muscle (Fig. 3). Whereas *ENO3* and *CYB5R1*
254 RNA was diffusely present across human skeletal myocytes, *SRSF11* was localized to sub-cell
255 membrane areas. In mouse muscle, *Eno3* was variably expressed, but neither *Cyb5r1* or *Srsf11*
256 were identified, although their levels of expression (~1,000x lower than *Eno3* in mouse) may be
257 too low to be seen by this method.

258 **Many highly-supported variably expressed proteins were not previously identified**

259 Thirty-six gene/proteins were variably expressed based on the simpleSingleCell,
260 HPASubC and MS based analyses (Fig. 2d, Table 1). Of these, based on an extensive literature
261 search, nine functionally diverse proteins are uniquely reported here as mosaic. Of the full 36,
262 22 were present in fast twitch myocytes and 14 in slow twitch myocytes based on the MS data.
263 In addition to these 36, another 414 genes/proteins were identified by two complementary
264 methods (Fig. 2). This includes well-known type specific proteins TNNC1 and TNNI1 (present
265 in the HPASubC and simpleSingleCell datasets, but not variably expressed in the MS dataset).

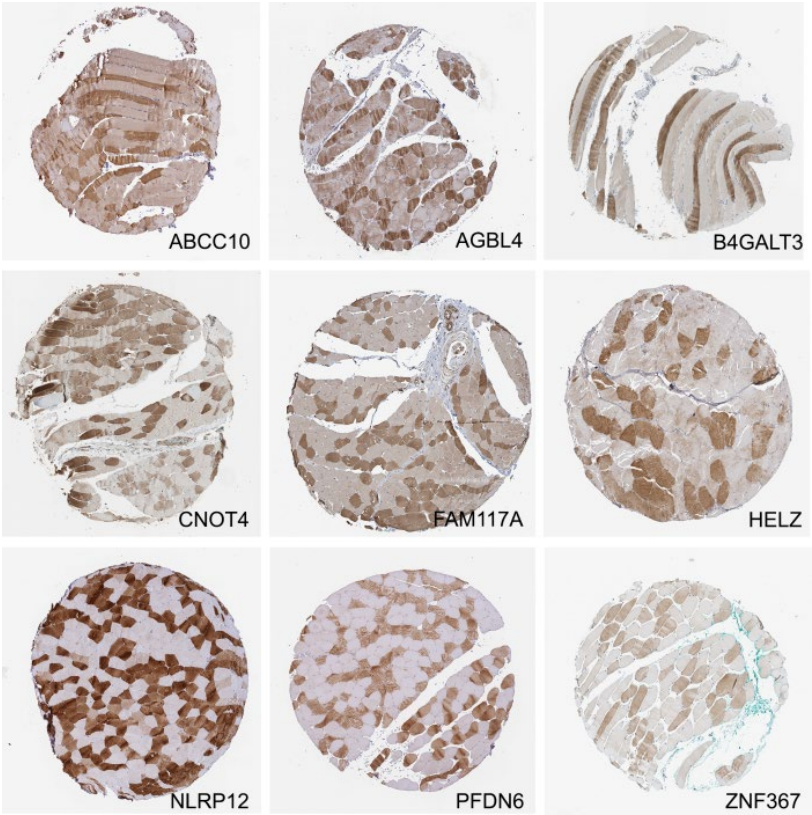
266 Finally, another 4,217 genes/proteins were variably expressed by one method. Of this
267 group, 1,211 were detected by HPASubC and 270 of these proteins were scored as "real" with
268 clear patterns of mosaicism (Supplementary Fig. 4).

Gene Name	Gene Symbol	Muscle Type- MS based	HPASubC Confidence	Seurat Norm. Counts	Mosaic Status
ArfGAP With Coiled-Coil, Ankyrin Repeat And PH Domains 2	Acap2	Fast	Likely	1.79	Unknown
Adenylate Kinase 1	Ak1	Fast	Likely	5.75	Known
Aldolase, Fructose-Bisphosphate A	Aldoa	Slow	Likely	7.98	Known
ATPase Sarcoplasmic/Endoplasmic Reticulum Ca2+ Transporting 1	Atp2a1	Fast	Real	6.68	Known
ATPase Sarcoplasmic/Endoplasmic Reticulum Ca2+ Transporting 2	Atp2a2	Slow	Real	2.08	Known
Calsequestrin 2	Casq2	Slow	Likely	1.61	Known

CD36 Molecule	Cd36	Slow	Likely	2.71	Known
Creatine Kinase, Mitochondrial 2	Ckmt2	Slow	Likely	4.54	Known
DnaJ Heat Shock Protein Family (Hsp40) Member C3	Dnajc3	Fast	Likely	2.2	Unknown
Enolase 3	Eno3	Fast	Real	6.97	Known
ELKS/RAB6-interacting/CAST Family Member 1	Erc1	Fast	Likely	3.14	Unknown
Glyceraldehyde-3-Phosphate Dehydrogenase	Gapdh	Fast	Likely	5.91	Known
Glyoxalase Domain Containing 4	Glod4	Slow	Likely	1.79	Unknown
Glycerol-3-Phosphate Dehydrogenase 1	Gpd1	Fast	Likely	3.22	Known
Kinesin Family Member 5B	Kif5b	Fast	Likely	3.83	Unknown
Lactate Dehydrogenase A	Ldha	Fast	Real	5.81	Known
Lactate Dehydrogenase B	Ldhb	Slow	Real	4.5	Known
Myosin Binding Protein C, Fast Type	Mybpc2	Fast	Real	5.2	Known
Myosin Heavy Chain 1	Myh1	Fast	Real	7.91	Known
Myosin Heavy Chain 6	Myh6	Slow	Likely	1.79	Known
Myosin Heavy Chain 8	Myh8	Fast	Likely	5.41	Known
Myosin Light Chain, Phosphorylatable, Fast Skeletal Muscle	Mylpf	Fast	Real	3.47	Known
Myosin Light Chain 3	Myl3	Slow	Real	1.95	Known
Myozenin 2	Myoz2	Slow	Real	8.06	Known
PDZ And LIM Domain 1	Pdlim1	Slow	Likely	3.43	Known
Peroxisomal Biogenesis Factor 19	Pex19	Fast	Likely	2.41	Unknown
Phosphofructokinase, Muscle	Pfkm	Fast	Likely	2.2	Known
Phosphoglycerate Kinase 1	Pgk1	Fast	Likely	3.67	Known
Ribosomal Protein S15a	Rps15a	Slow	Likely	4.94	Unknown
Thymosin Beta 4 X-Linked	Tmsb4x	Fast	Likely	3.58	Unknown
Troponin C2, Fast Skeletal Type	Tnnc2	Fast	Real	1.1	Known
Troponin I2, Fast Skeletal Type	Tnni2	Fast	Likely	5.19	Known
Troponin T1, Slow Skeletal Type	Tnnt1	Slow	Real	7.47	Known
Troponin T3, Fast Skeletal Type	Tnnt3	Fast	Real	6.31	Known
Topomyosin 1	Tpm1	Fast	Likely	7.97	Known
UDP-Glucose 6-Dehydrogenase	Ugdh	Slow	Likely	8.01	Unknown

Table 1. 37 Genes/Proteins identified as mosaic by all three methods based on simpleSingleCell analysis.

Discussion



Supplementary Figure 4. Nine representative images of 270 proteins scored as real mosaicism using HPASubC, but not identified by other methods. All images from HPA.

We describe the first proteogenomic analysis of skeletal muscle single fiber types using combined scRNA-seq, spatial proteomics, and MS proteomics. Because delineations of skeletal muscle fiber types are known and this project was exclusive to this one cell type, our study is a useful model system to evaluate combining and synthesizing gene and protein data into a coherent description of a cell. Also, by utilizing a deep sequencing approach and fewer cells, we were not limited to just classifying a cell, but rather had sufficient data to delve into full gene expression. Our data identifies common themes across the methods, but also significant differences and complexities in gene/protein assignments.

Regardless of the method and genes/proteins used to cluster, we found general agreement on the major types of skeletal muscle myocytes. We identified a small group of slow twitch cells

281 that clustered with fast 2A cells. These groups were consistently clustered away from two
282 clusters of fast 2X cells. The differences between these two fast 2X groups, described herein as
283 $2X_a$ and $2X_b$, are open to interpretation. The simplest explanation is that some axonal material
284 remained variably adherent to skeletal muscle cells through the NMJ, and these 20+ genes
285 resulted in the separation observed by UMAP (Fig. 1a). This would imply a technical cause of
286 the two fast 2X cell subtypes as a result of myocyte isolation. Adherent cell fragments are likely
287 to be a global issue for some cell type isolation, although it would not impact nuclear scRNA-seq
288 studies. A more interesting explanation is variable neuronal transfer of mRNAs across the NMJ
289 into the skeletal muscles via extracellular vesicles (22, 23). This would imply a real state-
290 difference in these cells, notable only by the deep sequencing strategy employed. Regardless, of
291 which is accurate, this division is unlikely to indicate true separate fast 2X subtypes. In fact, the
292 cross-referenced proteomic data was useful in demonstrating the arbitrary nature of this
293 delineation (Fig. 1i).

294 The extent of overlap of mosaic genes/proteins across the methods was surprisingly low.
295 Only 36 genes/proteins were cross-validated across all three approaches using the
296 simpleSingleCell method (Table 1). This list included well-known, fiber-type specific proteins
297 such as MYH1 and MYH6 and newly described mosaic proteins like DNAJC3 and GLOD4. The
298 lack of agreement across methods has made it difficult to confidently state how many
299 proteins/genes are variable by twitch pattern and further demonstrates the challenge of relying on
300 a single method. If a gene or protein is mosaic by two methods, this number climbs to 450. If all
301 mosaic genes and proteins are included, this increased to >4,500 genes/proteins. Over 1,600
302 proteins appear to be mosaic by the HPASubC method alone (Supplementary Data 1 and
303 Supplementary Fig. 4).

304 The reason for the variability in mosaic genes/proteins is certainly multifactorial. One
305 potential major difference is the comparison across two separate species (mouse and human). As
306 we noted with the DCAF11 IHC, this protein was mosaic in human muscle but did not appear to
307 be mosaic in mouse. Secondly, some genes have markedly variable expression levels between
308 the two species. While *CYB5R1* and *SRSF11* are robustly expressed in human muscle at 201.5
309 and 19.3 pTPM (in HPA), they were only 17.7 and 1.9 FPKM in our mouse scRNA-seq. It is
310 also possible that post-transcriptional regulation leads to more extreme expression variation in
311 proteins than genes. As described above, extreme expression dichotomy in *Myh* genes was less
312 than in similar MYH protein data (Fig. 1d) (5).
313 Our study represents the first use of LP-FACS to isolate single myofibers for scRNA-seq.
314 As skeletal myocytes are often long, stretching across the length of a muscle, isolation
315 techniques (particularly from human samples) may rely on the use of biopsies or otherwise
316 fragmented myocytes. To test the effect of myocyte fragmentation on scRNA-seq data quality,
317 we used a liberal gating strategy of our dissociated myocytes (including both EXT-high/TOF-
318 low and EXT-high/TOF-high populations) as well as directly sequencing fragmented myocytes
319 generated through a pseudo-biopsy approach. Disappointingly, we found that a large portion of
320 our sequenced myocytes were of poor quality, including those from our pseudo-biopsy approach.
321 By contrast, the highest quality data likely came from fully intact myocytes, in particular the
322 EXT-high/TOF-high population. Because this population is almost completely enriched for
323 intact myocytes, we believe that future experiments using LP-FACS to isolate skeletal myocytes
324 should focus solely on the EXT-high/TOF-high population. We are confident that this will allow
325 for a much higher percentage of good quality scRNA-seq libraries, akin to what we have
326 observed previously with LP-FACS isolation of cardiac myocytes (17). These results also mean

327 that more work must be done to identify better isolation methods for human skeletal muscle.
328 Current methods of human skeletal muscle biopsying from the quadriceps only obtains muscle
329 fragments and thus more creative methods to obtain full length fibers or non-damaged fibers
330 must be considered.

331 Technical factors also impact our ability to detect mosaicism on all platforms. Discovery
332 mass spectrometry is challenged to identify low abundance proteins. Having low input from
333 single fibers was further limiting and reduced the ability to computationally distinguish
334 expression differences in low abundance proteins. Most fibers had between 500-700 proteins
335 identified. As we have stated repeatedly, IHC in the HPA is subject to false positive staining
336 from shared epitopes (20, 24-26). It also incurs false negative staining for failed antibodies or
337 antibodies with staining parameters designed for other tissues. Further, some genes/proteins
338 observed in the other datasets were missing from the HPA data. The gene data was also limited
339 in the number of total cells analyzed (171) and the rarity of slow twitch cells from this muscle.
340 Cross-cell contamination, may have also stunted the differences between cell types (27).

341 In conclusion, we have created the first proteogenomic analysis of gene/protein
342 mosaicism in skeletal muscle. We replicated the known fiber types of slow, fast 2A, and fast 2X,
343 as well as greatly expanded our understanding of genes and with variable expression across these
344 cell types.

345 **Methods**

346 **Isolation and Sequencing of Adult Skeletal Myocytes**

347

348 Experiments were performed using C57BL/6J mice greater than 3 months of age. To isolate
349 skeletal myocytes, we performed collagenase-based digestion of the flexor digitorum brevis

350 (FDB), a short muscle of the hind feet, as per previously established protocols (28). We tested
351 two separate approaches to isolating myocytes. In the first approach, we dissected the FDB from
352 tendon to tendon prior to digestion, enabling isolation of fully intact myocytes. In the second
353 approach, we cut small portions of the FDB muscle using scissors. We reasoned that the latter
354 approach would broadly mimic skeletal muscle biopsy as might be done, for example, from a
355 human patient sample. The FDB was transferred to a dish containing DMEM with 1%
356 penicillin/streptomycin, 1% fetal bovine serum, and 2mg/mL Collagenase Type II
357 (Worthington). Muscle was digested for 1.5 hours in a 37C cell incubator with 5% CO_2.
358 Subsequently, the muscle was transferred to a dish containing media without collagenase, and
359 gently triturated to release single myocytes. Large undigested chunks and tendons were removed
360 with tweezers prior to single cell isolation.
361 We subsequently isolated single myocytes through large particle fluorescent-activated cell
362 sorting (LP-FACS), using a flow channel size of 500 µm. The COPAS SELECT Flow Pilot
363 Platform (Union Biometrica) was employed. Using time-of-flight (TOF, measuring axial length)
364 and optical extinction (EXT, measuring optical density) parameters, we found that skeletal
365 myocytes separated into three populations – an EXT-low population, EXT-high/TOF-low
366 population, and EXT-high/TOF-high population (Supplementary Fig. 1A). The EXT-high/TOF-
367 high population was comprised almost entirely of intact myofibers with lengths > 400 µm,
368 suggesting successful sorting of large myocytes (Supplementary Fig. 1B). Interestingly, the
369 EXT-high/TOF-low population was composed of what appeared to be rod-shaped fragments that
370 maintained sarcomeric proteins, albeit disrupted (Supplementary Fig. 1C). The EXT-low
371 population was comprised mostly of debris and dead cells, as previously observed with cardiac
372 myocytes (Supplementary Fig. 1D). The EXT-high/TOF-low population qualitatively resembled

373 our pseudo-biopsy isolated myocyte fragments (Supplementary Fig. 1E), which also shared
374 similar TOF and EXT parameters (not shown). To our knowledge, this is the first FACS-based
375 single cell RNA-seq study of skeletal myocytes; thus, we adopted a broad gating strategy for
376 isolation of single cells. We sorted 700 EXT-high myocytes (comprised of both TOF-high and
377 TOF-low populations) as well as 100 myocyte fragments isolated through the pseudo-biopsy
378 method.
379 These sorted cells were placed individually into 96-well plates. Capture plate wells contained 5
380 µl of capture solution (1:500 Phusion High-Fidelity Reaction Buffer, New England Biolabs;
381 1:250 RnaseOUT Ribonuclease Inhibitor, Invitrogen). Single cell libraries were then prepared
382 using the previously described mcSCRB-seq protocol (18, 19). Briefly, cells were subjected to
383 proteinase K treatment followed by RNA desiccation to reduce the reaction volume. RNA was
384 subsequently reverse transcribed using a custom template-switching primer as well as a barcoded
385 adapter primer. The customized mcSCRB-seq barcode primers contain a unique 6 base pair cell-
386 specific barcode as well as a 10 base pair unique molecular identifier (UMI). Transcribed
387 products were pooled and concentrated, with unincorporated barcode primers subsequently
388 digested using Exonuclease I treatment. cDNA was PCR-amplified using Terra PCR Direct
389 Polymerase (Takara Bio). Final libraries were prepared using 1ng of cDNA per library with the
390 Nextera XT kit (Illumina) using a custom P5 primer as previously described.
391
392 **scRNA-seq sequencing and analysis**
393 Pooled libraries were sequenced on two high-output lanes of the Illumina NextSeq500 with a 16
394 base pair barcode read, 8 base pair i7 index read, and a 66 base pair cDNA read design. To
395 analyze sequencing data, reads were mapped and counted using zUMIs 2.2.3 with default

396 settings and barcodes provided as a list (29). zUMIs utilizes STAR (2.5.4b) (30) to map reads to
397 an input reference genome and featureCounts through Rsubread (1.28.1) to tabulate counts and
398 UMI tables (30, 31). Reads were mapped to the mm10 version of the mouse genome. We used
399 GRCm38 from Ensembl concatenated with ERCC spike-in references for the reference genome
400 and gene annotations. Dimensionality reduction and cluster analysis were performed with Seurat
401 (2.3.4) (32).

402 **Seurat and simpleSingleCell**

403 Analysis was performed using the Seurat R toolkit V3.1.1 for this dataset (32). Initial filtering
404 removed lower quality cells (read count <5000 RNAs detected or >20% mitochondrial genes)
405 before sctransform normalization (33). A standard Seurat workflow was initially used for data
406 analysis. This workflow identifies a subset of genes with high cell-to-cell variation within the
407 scRNA-seq data. This subset is subsequently used as input to principal component analysis as
408 well as downstream nonlinear dimensionality reduction methods such as Uniform Manifold
409 Approximation and Projection (UMAP). Additionally, Seurat also allows for use of custom gene
410 lists as input to downstream analysis. This allowed us to use two custom gene lists, specifically
411 those derived from orthologous genes to mosaic proteins in the visual (HPASubC) dataset (20) or
412 the differentially expressed proteins in the MS proteomic dataset (5). Thus each of our three gene
413 lists, one produced by Seurat's workflow, another visual proteomic-based gene list, and a final
414 mass spectrometry-based gene list defining known muscle cell types, were used one at a time to
415 subset our initial data set and generate principal components for downstream analysis.
416 After determining clustering via these three approaches, UMAPs were generated alongside with
417 heat maps representing the top genes in clusters as determined by each gene set used for PCA.
418 Overlapping genes between the HPAsubC data, MS data, and significant genes determined by

419 Seurat were also examined for overlaps. Gene expression for *Trdh, Lama4, Ryr1, Dpp10,*

420 *Pde4dip, Sugct,* and *Myh1* was plotted across two fast $2X_a$ clusters based on the HPASubC data.

421

422 **Simple Single Cell and Scran**

423 Simple single cell 1.8.0 workflow was followed using scran 1.12.1 for normalization of raw

424 counts and fitting a mean-dependent trend to the gene-specific variances in single-cell RNA-seq

425 data (21). In line with this, we decomposed the gene-specific variance into biological and

426 technical components and selected the top 3000 genes for comparisons.

427

428 **RNA-ISH**

429 Mouse and human skeletal muscles were obtained at necropsy (>3 month old) and rapid autopsy

430 (66 year old male), the latter under an IRB-approved protocol. Tissues were immediately fixed in

431 formalin and paraffin-embedded blocks were created, from which 5 micron slides were made.

432 Custom probes for RNA *in situ* hybridization (RISH) were obtained from RNAscope (ACDBio).

433 These probes were designed to detect human and mouse forms of the following genes: *ENO3*

434 (GenBank accession nm_001976.5), *CYB5R1* (nm_016243.3), *SRSF11* (nm_004768.5), *Eno3*

435 (nm_007933.3), *Cyb5r1* (nm_028057.3), and *Srsf11* (nm_001093753.2). Each probe set targeted

436 all validated NCBI refseq transcript variants of the gene.

437 The Multiplex Fluorescent Reagent Kit v2 (ACDBio) was used following the manufacturer's

438 instructions. Briefly, FFPE tissue slides were baked for one hour at 60°C. The slides were

439 subsequently deparaffinized with xylene, rinsed with 100% ethanol and air-dried. After

440 application of hydrogen peroxide and washing, slides were treated with target retrieval reagent in

441 a steamer (>99°C) for 20 minutes. Then, the tissue was permeabilized using a protease.

442 Hybridization of the probes to the targeted mRNAs was performed by incubation in a 40°C oven
443 for 2 hours. After washes, the slides were processed for the standard signal amplification and
444 application of fluorescent dye (Opal dye 520, 570 and 620, AKOYA Biosciences) steps. Finally,
445 the slides were counterstained with DIPA, mounted with Prolong Gold Antifade Mounting
446 solution (Invitrogen) and stored in a 4°C room. The fluorescent images were obtained in the
447 Johns Hopkins Microscope Core Facility using a Zeiss LSM700 Laser scanning confocal
448 microscope.
449
450 **Immunohistochemistry**
451 The same tissues described above were used for standard immunohistochemistry. Antibodies
452 were obtained for WDR23/DCAF11 (bs-8388R, Bioss Antibodies), PVALB (A2781, Abclonal),
453 and ENO3 (ARP48203_T100, Aviva Systems Biology) that were reported to cross react to both
454 human and mouse. Immunohistochemistry was performed as described previously (25, 34).
455
456 **HPA and HPASubC**
457 The HPA is a comprehensive repository of IHC stained tissue microarrays for numerous
458 tissues, including skeletal muscle (35, 36). The HPASubC tool can rapidly and agnostically
459 interrogate images of the HPA to characterize specific staining patterns in organs (20, 24, 26).
460 HPASubC v1.2.4 was used to download 50,351 skeletal muscle tissue microarray images
461 covering 10,301 unique proteins from the HPA website (v18). The images were individually
462 reviewed using HPASubC by K.M.F to evaluate the presence of a mosaic pattern of protein
463 expression based on IHC staining. The classification of mosaicism was based on a pre-study
464 training set of 300 images from HPA reviewed collaboratively (K.M.F and M.K.H). Mosaicism

465 was defined as a dispersed pattern of differential staining in which a significant number of non-
466 adjacent muscle fibers had a higher staining intensity than the surrounding fibers, preferably
467 persisting across the entire microarray. All positive selections made by the trainee were reviewed
468 and rescored, as needed, by a board-certified pathologist (M.K.H.).

469 After an initial fast review of the images, a re-review to score the images was performed.
470 A three-tiered classification system was used indicating increasing certainty of mosaicism: 0
471 indicated the absence of mosaic staining; 1 indicated unknown mosaic staining; 2 indicated
472 likely mosaic staining; 3 indicated real mosaic staining. Scoring evaluation was based on the
473 quality of the mosaic pattern, including stain intensity differential between fibers, the presence of
474 "blush"/incomplete staining within cells, and the consistency and completeness of the fiber
475 staining pattern throughout the sample. HPASubC was used on an Apple MacBook Pro running
476 macOS Sierra v10.12.6 with 8 GB RAM and 3.1 GHz CPU and a Dell Precision Tower 3620
477 running Windows 10 with 16 GB RMA and a 3.7 GHz CPU.

478

479 **Conversion of gene and protein symbols**

480 To identify orthologs across human and mouse genes/proteins we had to synchronize
481 across gene/protein names and across the species. We used the David Gene ID Conversion Tool
482 (https://david.ncifcrf.gov/conversion.jsp), BioMart at Ensembl
483 (http://useast.ensembl.org/biomart/martview/e8a4fba4cb5c0be7a30841471b55674d), UniProt
484 Retreive/ID mapping (https://www.uniprot.org/uploadlists/) and direct searches at both UniProt
485 and GeneCards (https://www.genecards.org/), to cross integrate the human protein symbols,
486 mouse gene symbols, human gene symbols and ENSG IDs (37-39).

487

488 **Gene Ontology (GO) Validation**

489 GO was performed on the 557 most variable genes between two fast 2X clusters ($2X_a$ and
490 $2X_b$) using the Gene Ontology resource (http://geneontology.org/) and selecting for cellular
491 component.

492

493 **Mass Spectrometry (MS) Data Set**

494 We utilized the Murgia et al. human skeletal muscle fiber MS-based proteomic dataset
495 (5). This contained information from 3,585 proteins across 152 fibers from 8 donors (5). The
496 ratio of expression of proteins between Type 1 and Type 2A cells were determined using Table
497 S6 of Murgia et al. Five hundred and ninety-six proteins with >2.3 fold differences between cell
498 types were selected. Label-free quantification (LFQ) data, from Supplemental Table S4, for the
499 154 human single muscle fiber proteomics was obtained. The log2 transformed LFQ data was
500 converted to raw values and only proteins expressed across all fiber types (n=94) were
501 considered for plotting UMAP as described (5). Functions of the R-package Seurat (Version
502 3.1.1) were executed sequentially to derive a UMAP along with its dependency library
503 "uwot (Version 0.1.4)" in R (Version 3.6.1) (40, 41). A Seurat object of the data matrix was
504 created using 'CreateSeuratObject' with default parameters. This data was normalized using the
505 'NormalizeData' function and outlier proteins were identified using the 'FindVariableFeatures.'
506 Proteins across the fiber types were scaled and centered to create a PCA object using 'ScaleData'
507 and 'RunPCA' respectively. Further, k-nearest neighbors and shared nearest neighbor for each
508 fiber type were generated on the Seurat object using 'FindNeighbors' and 'FindClusters'to plot
509 UMAP using 'RunUMAP'. All of these functions were executed using default parameters. The

510 clustering obtained with UMAP was overlaid with the classification of muscle fiber types based
511 on Murgia et al. using ggplot2 (Version 3.2.1).

512 **Data availability**

513 Mouse skeletal muscle sequencing was deposited at the Sequence Read Archive (SRA –
514 SRP241908) and the Gene Expression Omnibus (GSE143636).

515 **Code availability**

516 All analysis scripts are available at GitHub
517 (https://github.com/mhalushka/Skeletal_muscle_mosaicism).

518

519 **Acknowledgements:**

520 The authors thank Efrain Ribeiro for his helpful comments on the project. M.K.H. was supported
521 by grants 1R01HL137811, R01GM130564, and P30CA006973 from the National Institutes of
522 Health and 17GRNT33670405 from the American Heart Association. T.O.N. was supported by
523 grant R01GM130564. M.N.M. was supported by R01HL137811 and the University of Rochester
524 CTSA award number UL1TR002001. A.Z.R was supported by R01GM130564. C.K. and S.K.
525 were supported by NIH R01HD086026, TEDCO 2019-MSCRFD-5044, and the JHU Discovery
526 Award. S.K. was supported by fellowship 20PRE35200028 from the American Heart
527 Association.

528 **Contributions** K.M.F helped conceive the project and generated proteomic data. S.K. and
529 B.L.L. generated the skeletal muscle sequencing library. X.Y and K.F-T. performed IHC and
530 RISH. R.X.V., S.K., T.O.N, A.H.P. and M.N.M. performed analysis. C.K. and D.A.K. oversaw
531 the library preparation. A.Z.R helped develop the project. M.K.H. conceived the project,

532 performed analyses and wrote the manuscript. All authors contributed toward revisions of the
533 manuscript.

534 **Conflicts of interest**

535 The authors declare no conflicts of interest.

536 **References**

537 1. Okumura N, Hashida-Okumura A, Kita K, Matsubae M, Matsubara T, Takao T, et al.
538 Proteomic analysis of slow- and fast-twitch skeletal muscles. Proteomics. 2005;5(11):2896-906.
539 2. Gonzalez-Freire M, Semba RD, Ubaida-Mohien C, Fabbri E, Scalzo P, Hojlund K, et al.
540 The Human Skeletal Muscle Proteome Project: a reappraisal of the current literature. J Cachexia
541 Sarcopenia Muscle. 2017;8(1):5-18.
542 3. Schiaffino S, Reggiani C. Fiber types in mammalian skeletal muscles. Physiological
543 reviews. 2011;91(4):1447-531.
544 4. Drexler HC, Ruhs A, Konzer A, Mendler L, Bruckskotten M, Looso M, et al. On
545 marathons and Sprints: an integrated quantitative proteomics and transcriptomics analysis of
546 differences between slow and fast muscle fibers. Molecular & cellular proteomics : MCP.
547 2012;11(6):M111 010801.
548 5. Murgia M, Toniolo L, Nagaraj N, Ciciliot S, Vindigni V, Schiaffino S, et al. Single
549 Muscle Fiber Proteomics Reveals Fiber-Type-Specific Features of Human Muscle Aging. Cell
550 reports. 2017;19(11):2396-409.
551 6. Murgia M, Nagaraj N, Deshmukh AS, Zeiler M, Cancellara P, Moretti I, et al. Single
552 muscle fiber proteomics reveals unexpected mitochondrial specialization. EMBO Rep.
553 2015;16(3):387-95.
554 7. Chemello F, Bean C, Cancellara P, Laveder P, Reggiani C, Lanfranchi G. Microgenomic
555 analysis in skeletal muscle: expression signatures of individual fast and slow myofibers. PLoS
556 One. 2011;6(2):e16807.
557 8. Giordani L, He GJ, Negroni E, Sakai H, Law JYC, Siu MM, et al. High-Dimensional
558 Single-Cell Cartography Reveals Novel Skeletal Muscle-Resident Cell Populations. Molecular
559 cell. 2019;74(3):609-21 e6.
560 9. Porpiglia E, Samusik N, Ho ATV, Cosgrove BD, Mai T, Davis KL, et al. High-resolution
561 myogenic lineage mapping by single-cell mass cytometry. Nature cell biology. 2017;19(5):558-
562 67.
563 10. Dell'Orso S, Juan AH, Ko KD, Naz F, Perovanovic J, Gutierrez-Cruz G, et al. Single cell
564 analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions.
565 Development. 2019;146(12).
566 11. Cho DS, Doles JD. Single cell transcriptome analysis of muscle satellite cells reveals
567 widespread transcriptional heterogeneity. Gene. 2017;636:54-63.
568 12. Cornelison DD, Wold BJ. Single-cell analysis of regulatory gene expression in quiescent
569 and activated mouse skeletal muscle satellite cells. Developmental biology. 1997;191(2):270-83.
570 13. Cacchiarelli D, Qiu X, Srivatsan S, Manfredi A, Ziller M, Overbey E, et al. Aligning
571 Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants
572 of Myogenic Reprogramming Outcome. Cell Syst. 2018;7(3):258-68 e3.

573 14. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics
574 and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.
575 Nature biotechnology. 2014;32(4):381-6.
576 15. Rubenstein AB, Smith GR, Raue U, Begue G, Minchev K, Ruf-Zamojski F, et al. Single-
577 cell transcriptional profiles in human skeletal muscle. Scientific reports. 2020;10(1):229.
578 16. Blackburn DM, Lazure F, Corchado AH, Perkins TJ, Najafabadi HS, Soleimani VD.
579 High-Resolution Genome-Wide Expression Analysis of Single Myofibers Using SMART-Seq. J
580 Biol Chem. 2019.
581 17. Kannan S, Miyamoto M, Lin BL, Zhu R, Murphy S, Kass DA, et al. Large Particle
582 Fluorescence-Activated Cell Sorting Enables High-Quality Single-Cell RNA Sequencing and
583 Functional Analysis of Adult Cardiomyocytes. Circ Res. 2019;125(5):567-9.
584 18. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS.
585 Characterization of directed differentiation by high-throughput single-cell RNA-Seq. BioRxiv.
586 2014.
587 19. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al.
588 Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular cell.
589 2017;65(4):631-43 e4.
590 20. Cornish TC, Chakravarti A, Kapoor A, Halushka MK. HPASubC: A suite of tools for
591 user subclassification of human protein atlas tissue images. Journal of pathology informatics.
592 2015;6:36.
593 21. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of
594 single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5:2122.
595 22. Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like Gag
596 Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell. 2018;172(1-2):262-74 e11.
597 23. Korkut C, Ataman B, Ramachandran P, Ashley J, Barria R, Gherbesi N, et al. Trans-
598 synaptic transmission of vesicular Wnt signals through Evi/Wntless. Cell. 2009;139(2):393-404.
599 24. Anene DF, Rosenberg AZ, Kleiner DE, Cornish TC, Halushka MK. Utilization of
600 HPASubC for the identification of sinusoid-specific proteins in the liver. Journal of proteome
601 research. 2016;15(5):1623-9.
602 25. Wang TY, Lee D, Fox-Talbot K, Arking DE, Chakravarti A, Halushka MK.
603 Cardiomyocytes have mosaic patterns of protein expression. Cardiovasc Pathol. 2018;34:50-7.
604 26. Cheah JX, Nieuwenhuis TO, Halushka MK. An expanded proteome of cardiac t-tubules.
605 Cardiovasc Pathol. 2019;42:15-20.
606 27. Nieuwenhuis TO, Yang S, Verma RX, Pillalamarri V, Arking DE, Rosenberg AZ, et al.
607 Basal Contamination of Sequencing: Lessons from the GTEx dataset. BioRxiv. 2019.
608 28. Shefer G, Yablonka-Reuveni Z. Isolation and culture of skeletal muscle myofibers as a
609 means to analyze satellite cells. Methods Mol Biol. 2005;290:281-304.
610 29. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible
611 pipeline to process RNA sequencing data with UMIs. Gigascience. 2018;7(6).
612 30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
613 universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
614 31. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping
615 by seed-and-vote. Nucleic Acids Res. 2013;41(10):e108.
616 32. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell
617 transcriptomic data across different conditions, technologies, and species. Nature biotechnology.
618 2018;36(5):411-20.

619 33. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq
620 data using regularized negative binomial regression. BioRxiv. 2019.
621 34. Wang TY, Arking DE, Maleszewski JJ, Fox-Talbot K, Nieuwenhuis TO, Santhanam L, et
622 al. Human cardiac myosin light chain 4 (MYL4) mosaic expression patterns vary by sex.
623 Scientific reports. 2019;9(1):12681.
624 35. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al.
625 Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
626 36. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards
627 a knowledge-based Human Protein Atlas. Nature biotechnology. 2010;28(12):1248-50.
628 37. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl
629 2016. Nucleic Acids Res. 2016;44(D1):D710-6.
630 38. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res.
631 2017;45(D1):D158-D69.
632 39. Fishilevich S, Zimmerman S, Kohn A, Iny Stein T, Olender T, Kolker E, et al. Genic
633 insights from integrated human proteomics in GeneCards. Database : the journal of biological
634 databases and curation. 2016;2016.
635 40. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al.
636 Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902 e21.
637 41. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality
638 reduction for visualizing single-cell data using UMAP. Nature biotechnology. 2018.
639