High Throughput Proteomic Analysis of Heterogeneous Cell Populations with NanoDTSC

Authors

Simion Kreimer*[1,2], Aleksandra Binek[1,2], Blandine Chazarin[1,2], Jae H Cho[2], Ali Haghani[1,2], Alexandre Hutton[1-3], Eduardo Marbán[2], Mitra Mastali[1,2], Jesse G Meyer[1-3], Thassio R Ribiero Mesquita[2], Yang Song[1,2], Jennifer Van Eyk*[1,2], Sarah Parker*[1,2]

1 Advanced Clinical Biosystems Research Institute, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA
2 Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA
3 Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA
* Equal Contribution

Abstract

Characterization of rare cell types in heterogeneous cell populations extracted from tissue or organs requires profiling hundreds of individual cells. Parallelized nanoflow dual-trap single-column liquid chromatography (nanoDTSC) quantifies peptides over 11.5 out of 15 minutes of total run time using standard commercial components, thus offering the most accessible and efficient LC solution for single-cell applications. Over 1,000 proteins were quantified in individual cardiomyocytes and heterogenous aorta cells using nanoDTSC.

Main Text

Single cell transcriptomics has revealed unique cell subpopulations within a myriad of tissues[1,2] but complementary global measurements of the corresponding proteomic phenotypes via LC-MS have been hampered by the limited quantity of protein and comparatively slow analysis. Initially multiplexing by isobaric mass tag (TMT) labelling addressed these challenges by combining analysis of up to 16 samples and increasing sensitivity with a "booster" channel containing a pool of cells (SCoPE-MS).[3,4] The combined signal from proteins shared between samples in the multiplex set boosts the fragment signals into a reliably identifiable range. Complex samples such as dissociated organs or pathological specimen with substantial heterogeneity in the proteomes of phenotypically distinct cells challenge this strategy because proteins unique to a phenotype are not amplified like shared proteins and thus less likely to be identified. Reliance on a "booster" introduces bias for the dominant cell type(s) and further reduces the profiling of rare cells.[5,6] Direct analysis of individual cells is unbiased in capturing all phenotypes during proteomic analysis of heterogenous tissues and organs.

Label-free analysis of individual cells has been achieved via highly efficient ionization at ultralow flow-rates (*e.g.* 10 - 30 nL/min)[7,8] and by improved MS ion transmission and sensitivity.[9] Analysis at low flow-rates is slowed by the dwell volume of the system and the time spent loading the sample, cleaning, and equilibrating the system resulting in a low amount of data collected relative to total instrument operation time (instrument utilization, IU). Operation of two analytical columns in parallel, where a sample is loaded on one column while a second sample is analyzed elegantly increases IU, but requires a specialized ion source that accommodates two emitters and an LC with two analytical pumps.[8] Our dual trap single column (DTSC) platform optimized for the nano flowrate (500 nL/min, nanoDTSC) increases

the throughput of single cell analysis by parallelizing sensitive high ion accumulation time (166 ms) DIA-PASEF data acquisition on the timsTOF-SCP (Bruker) with resuspension, loading, and desalting of the subsequent sample.[10] NanoDTSC does not require specialized instrumentation making it widely accessible; here an Ultimate 3000 nanoRSLC (Thermo) with the standard Bruker captive spray ionization source were used.

The developed method collected 11.5 minutes of data (span of peptide elution) during 15 minutes of total run-time with negligible delay (2 s) between runs to achieve 77% IU (**Fig. 1A**). Sample loss was minimized by dispensing,[2] lysing, and digesting cells with a modified one-pot protocol[11] on the same 384-well plate in 600 nL of solvent. The generated peptides dried rapidly and were resuspended and injected by the LC autosampler (**Fig. 1B**). The 5-10 minutes required for thorough sample resuspension and loading were recovered through parallelization which increased throughput by at least 25% compared to conventional trap and elute with the same components. The platform was used in analysis of relatively homogenous heart extracted cardiomyocytes, and a heterogenous mixture of cells extracted from mouse aorta.

Method reproducibility and robustness were evaluated using 3 ng of bulk digested mouse heart tissue as the system suitability sample (SSS). Analysis of 56 technical replicates quantified on average 1,145 mouse cardiac proteins with 1,063 at high precision (coefficient of variance, CV < 20%). After almost 1,700 injections using the same trapping and analytical columns a second batch of 40 SSS injections was analyzed and an average of 1,025 proteins were quantified, with 925 proteins under 20% CV demonstrating a minor loss in performance after 3 weeks of continuous operation. Longitudinal system performance was also monitored by analysis of 4 ng of Pierce HeLa standard once on each trap every 50 single cell runs. On average 17,290 precursor ions and 2,942 proteins were identified without significant deviation (< 10% deviation from average) through the course of analysis. Peak widths and ion current MS1 and MS2 intensities also remained stable (< 20% deviation from average, **Fig. 1C**).

Analysis of 44 individual wild-type mouse cardiomyocytes found 96.9% proteome overlap with the SSS and similar numbers of identified precursors and proteins. The reproducibility and data completeness were lower between individual mouse cardiomyocytes than the SSS replicates due to biological heterogeneity. Human cardiomyocytes isolated from the endomyocardium (n = 30), midmyocardium (n = 13), and epimyocardium (n = 49) of the left ventricle of one donor heart had similar protein identification numbers and data completeness but even higher quantitative heterogeneity (**Fig. 1D**). This variance could be attributed to the wide cell size distribution (CellenOne diameter: 47 - 114 µm), as it exceeds the technical imprecision baseline established from SSS and mouse cardiomyocyte analysis. The rank vs. log intensity plot of heart tissue and cardiomyocytes shows highly similar protein identifications and a distribution spanning roughly 5 orders of magnitude with high data completeness that drops off for lower abundance proteins (**Fig. 2A**). Dimension reduction of protein quantities from human cardiomyocytes with Uniform manifold approximation and projection (UMAP) found that the number of identified proteins was the biggest driver of clustering. Separate clustering of the high identification (>1100 proteins) and low identification groups (<1100) still did not identify clear subpopulations or differences in source tissue layers (**Fig. S1**). The ability to identify cardiomyocyte subpopulations by UMAP is hindered by the low cell number (92) and the quantitative variance resulting from the wide cell size distribution.

Conventional normalization strategies such as adjustment based on median or total protein that are typically used to reduce variance due to differences in bulk protein content may not be appropriate for cells spanning a range of sizes, lifecycle stages, and sub-types. We attempted a ratio-based analysis where each protein is divided by every other protein within each cell (limited to proteins quantified in all cells, 255 for this set) and the CV of each ratio across the cell population is used to infer correlations in protein expression. In human cardiomyocytes, 4,010 ratios from 154 proteins had a CV under 20%. These proteins were from mitochondrial complexes which is not surprising since their relative abundance is independent of the number of mitochondria within a cell and tightly regulated. Proteins from the sarcomeric complex appeared in the 11,000 ratios from 204 proteins within the 20-30% CV range (**Fig. S2**). In agreement with known sarcomeric protein complex composition,[12] the average proportion of Myosin light chain (MYL3) to heavy chain (MYH7) was 0.5 (25% CV). Additionally, myosin light chain 2 (MYL2) to myosin light chain 3 (MYL3) was a stable 1:1 ratio with (CV 30%) and the proportion of tropomyosin 1 (TPM 1) to tropomyosin 2 (TPM2) was also a stable average of 8.4 (26% CV) (**Fig. 2B**). Ratio-based analysis of mouse smooth muscle cells (SMC) extracted from aorta also found high regulation of mitochondrial protein ratios and reproducible ratios of tropomyosin 1 to tropomyosin 2 (1.3, CV 16%). Interestingly, SMC heavy (MYH11) to regulatory light chain (MYL9) ratio was on average 7.5 with substantial heterogeneity between SMC subtypes (CV 31%, subtype annotation described below) (**Fig. S3**). This analysis reflects fundamental biology and circumvents the challenge of individual protein intensity heterogeneity due to cell size distribution, but it should be noted that ratios combine quantitative imprecision of each protein and are inflated by outlier cells. Furthermore, the coefficient of variance is not an adequate metric in cell populations where the protein quantities are regulated only in some of the cells or are expressed in two or more ratios. More sophisticated statistical methods that identify outliers and multiple distributions are needed in future work.

The aorta contains a heterogeneous mixture of smooth muscle cells (SMC), adipocytes, endothelial cells, fibroblasts, macrophages, and other immune cells. Cells from dissociated mouse aorta were analyzed by nanoDTSC and putatively annotated and clustered based on the presence of known markers. Dimensionality of quantitative single cell protein data was reduced by UMAP and clustered with the Leiden algorithm. Putative marker-based annotations were projected onto the points in the dimension reduced space. For SMCs the annotation mapped unto four unbiased clusters (#0, 4, 5, and 8) (**Fig. 2C**). Cluster 8 was close to 0 and only a few cells in size, thus we grouped SMCs into 3 proteome-based clusters. Notably, transgelin (SM22alpha) showed the most profound difference in abundance in cluster 5 (SMC3) which is consistent with modified, proliferative SMC phenotype that is biologically significant in disease states (**Fig. 2D**).[13] Fibroblast markers mostly mapped unto a single cluster (#3), which also contained un-annotated cells, suggesting that the proteins used for annotation (CYGB and ADD3) do not represent all fibroblasts and these unannotated cells were assigned as fibroblasts due to overall proteome similarity. Adipocytes, endothelial, and macrophage markers did not map neatly unto clusters, although some patterns could be discerned. Overall, unsupervised clustering is a viable initial strategy for definition of putative cell types in complex mixtures without *a priori* knowledge, and could be improved with higher total cell numbers that include more instances of the less abundant cell-types.

NanoDTSC reproducibly quantified the proteomes of individual cardiomyocytes to a depth of over 1,000 protein groups in 15 minutes of total run time, with meaningful data acquired for 77% of the time. Longitudinal quality control showed that the system remained robust over 1,700 samples without replacement of components. The platform proved suitable for analysis of heterogeneous cell mixtures

extracted from an organ and the combination of cell marker-based annotation and unsupervised clustering was able to define putative cell types. The proposed ratio-based analysis circumvented the challenge of wide cell-size distributions and confirmed the expected stability of mitochondrial and sarcomeric protein complexes. NanoDTSC neatly integrated with automated isolation and tryptic digestion of cells by resuspending the generated peptides from the same 384-well plate unto which the living cells were dispensed thus minimizing sample loss. While 15 minute/cell is a decent throughput for most applications it can be increased further with a lower volume analytical column (**Fig. S4)** and non-isobaric labeled multiplexing (plexDIA).[14] The presented data is shared through MassIVE (MSV000090903, Proteome Xchange PXD038828, https://doi.org/doi:10.25345/C5Z60C65X, password: dualtrapscp).

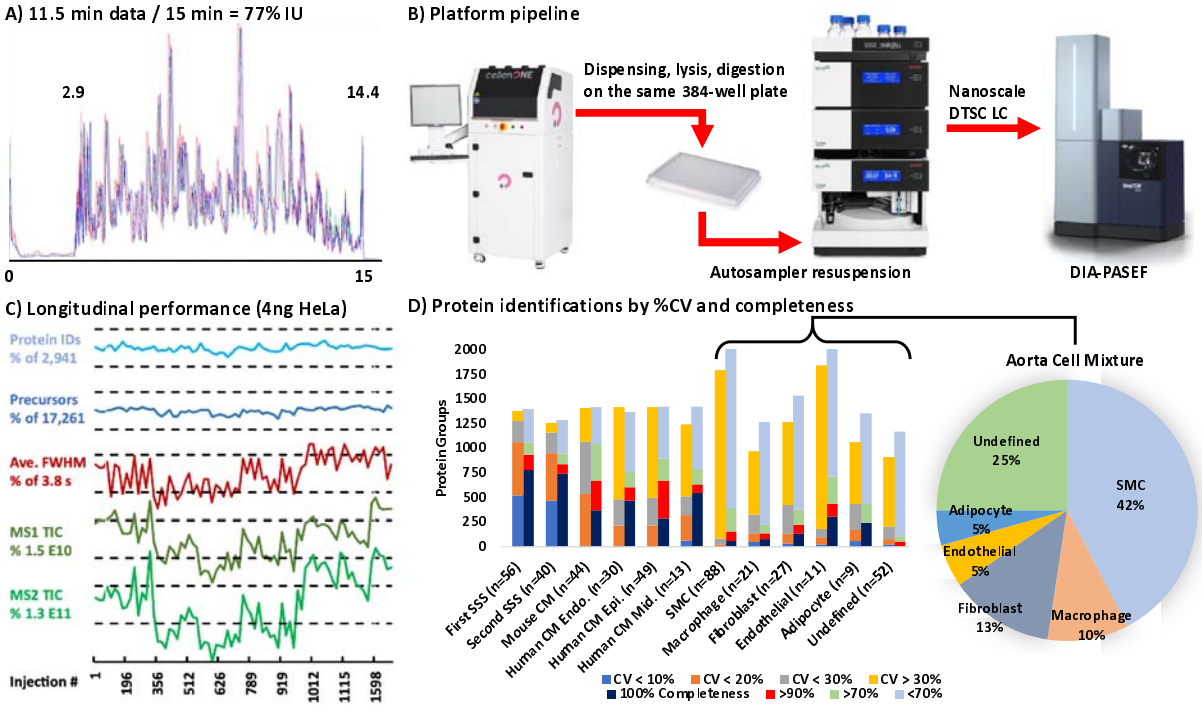


Figure 1. A – Overlay of 4-SSS runs, peptides elution 2.9-14.4 minutes (77% instrument utilization). B – Workflow overview: Individual cells are dispensed by CellenOne into 384-well plate, digested, resuspended by the autosampler, and analyzed by DIA-PASEF. C – Consistent system performance over 1,700 injections monitored by HeLa QC, dashed lines represent 10% deviation from average. D – Reproducibility and data completeness for the various sample types analyzed.

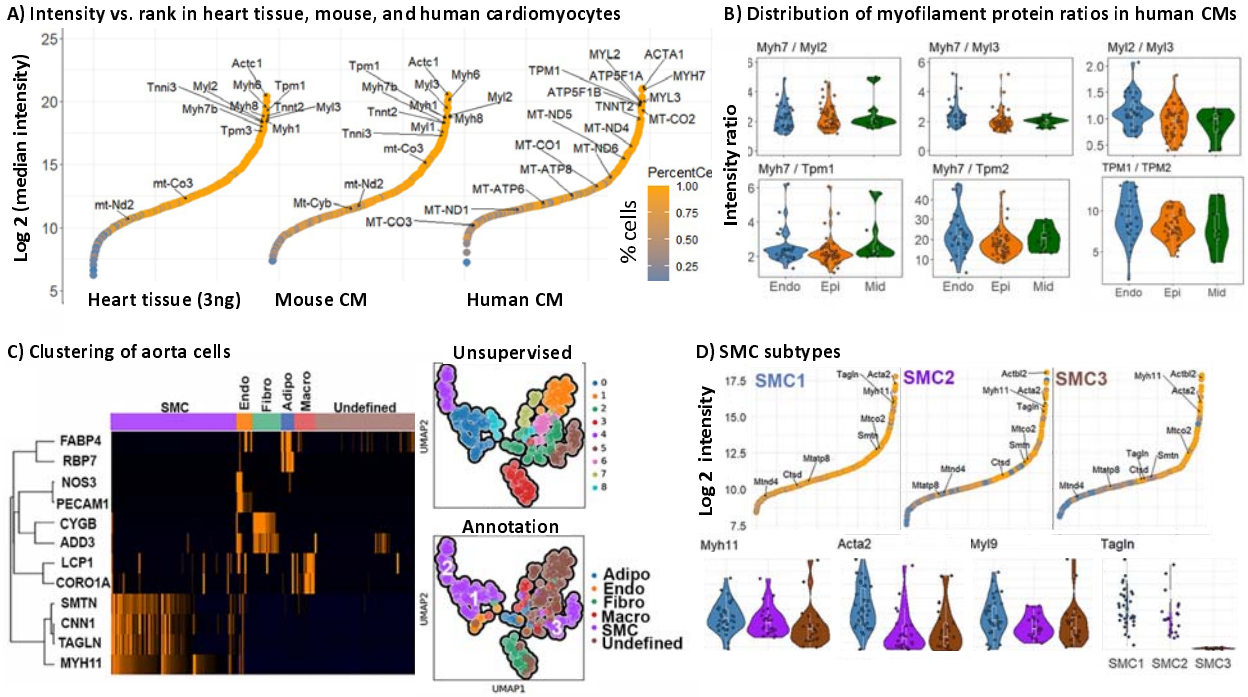


Figure 2. A – Log intensity vs rank curves, data completeness is indicated by color. B – Ratios of sarcomeric proteins in human CMs by heart region. C – Clustering of aorta cells by cell-type markers, UMAP clustering by entire proteome and projection of cell types unto the UMAP plot. D – Rank-intensity curves of SMC sub-types and distributions of key contractile markers.

Acknowledgements

JGM and Alexandre H were supported by the NIGMS (R35GM142502). SP was supported by OWHR/NHLBI R01HL165471

1. Molenaar, B. *et al.* Single-cell transcriptomics following ischemic injury identifies a role for B2M in cardiac repair. doi:10.1038/s42003-020-01636-3.

2. Ctortecka, C. *et al.* An automated workflow for multiplexed single-cell proteomics sample preparation at unprecedented sensitivity. doi:10.1101/2021.04.14.439828.

3. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol* **19**, 1–12 (2018).

4. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol* **22**, 1–27 (2021).

5. Karp, N. A. *et al.* Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular and Cellular Proteomics* **9**, 1885–1897 (2010).

6. Pappireddi, N., Martin, L. & Wühr, M. A Review on Quantitative Multiplexed Proteomics. *Chembiochem* **20**, 1210–1224 (2019).

7. Mellors, J. S., Jorabchi, K., Smith, L. M. & Ramsey, J. M. Integrated microfluidic device for automated single cell analysis using electrophoretic separation and electrospray ionization mass spectrometry. *Anal Chem* **82**, 967–973 (2010).

8. Webber, K. G. I. *et al.* Label-Free Profiling of up to 200 Single-Cell Proteomes per Day Using a Dual-Column Nanoflow Liquid Chromatography Platform. *Anal Chem* **94**, 6017–6025 (2022).

9. Brunner, A. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol Syst Biol* **18**, 1–15 (2022).

10. Kreimer, S. *et al.* Parallelization with Dual-Trap Single-Column Configuration Maximizes Throughput of Proteomic Analysis. *Anal Chem* (2022) doi:10.1021/acs.analchem.2c02609.

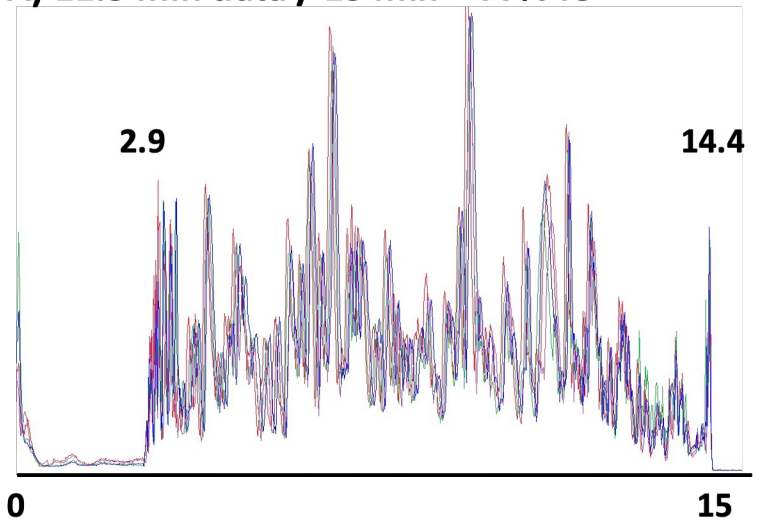
11. Tsai, C. F. *et al.* Surfactant-assisted one-pot sample preparation for label-free single-cell proteomics. *Commun Biol* **4**, 265 (2021).

12. Hamdani, N. *et al.* Sarcomeric dysfunction in heart failure. *Cardiovascular Research* vol. 77 649–658 Preprint at https://doi.org/10.1093/cvr/cvm079 (2008).

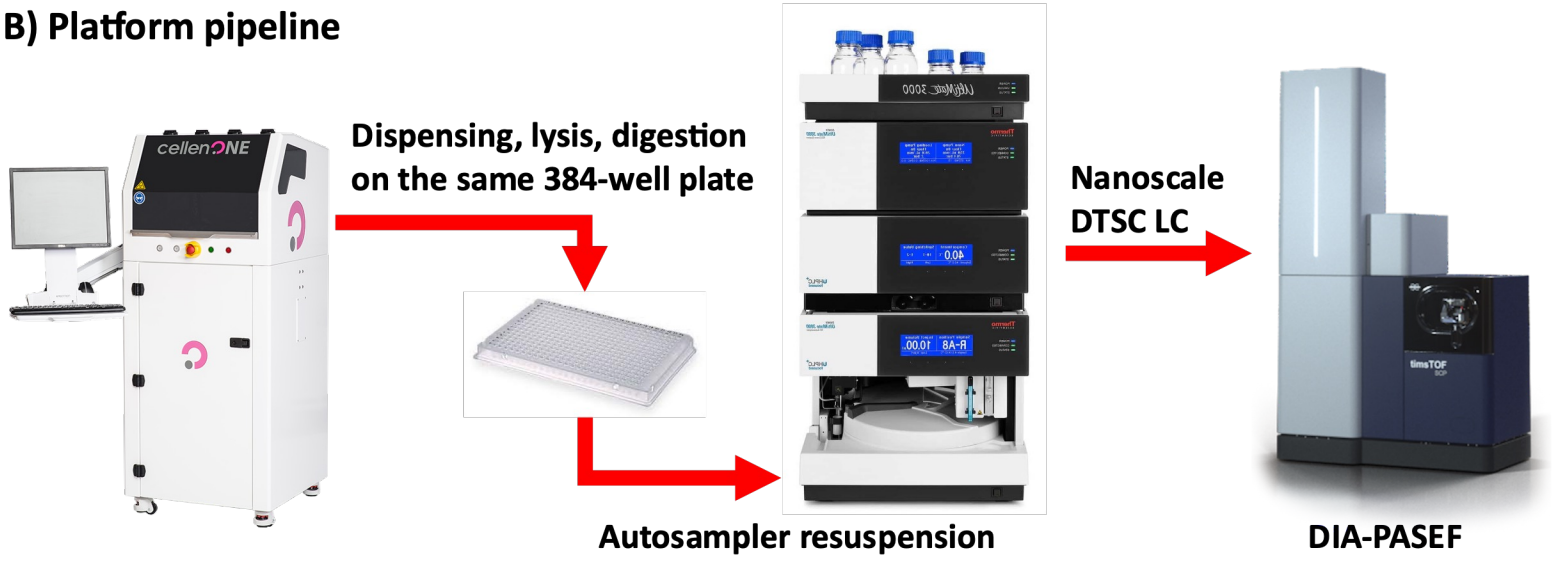
13. Feil, S., Hofmann, F. & Feil, R. SM22α Modulates Vascular Smooth Muscle Cell Phenotype during Atherogenesis. *Circ Res* **94**, 863–865 (2004).

14. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nat Biotechnol* (2022) doi:10.1038/s41587-022-01389-w.

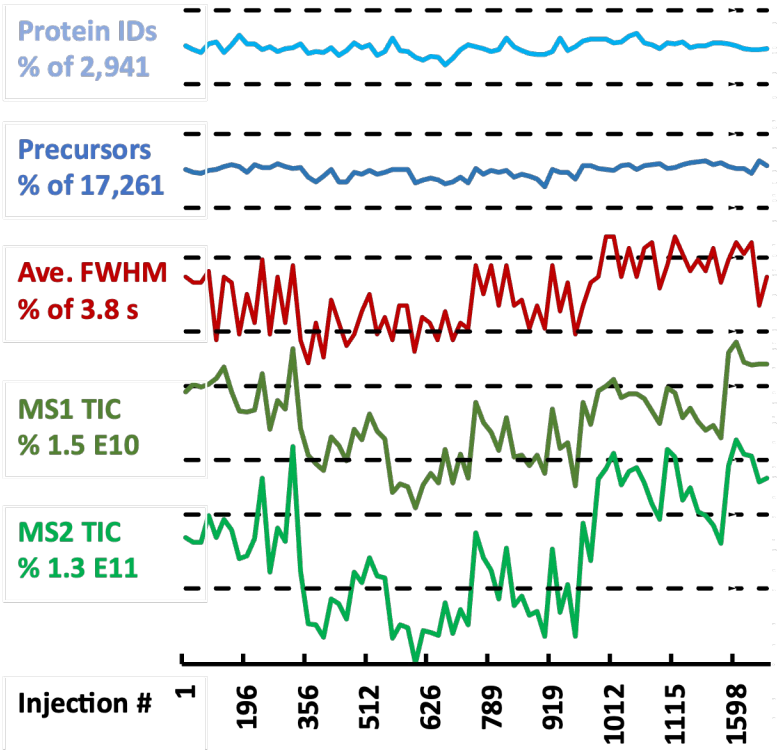
A) 11.5 min data / 15 min = 77% IU



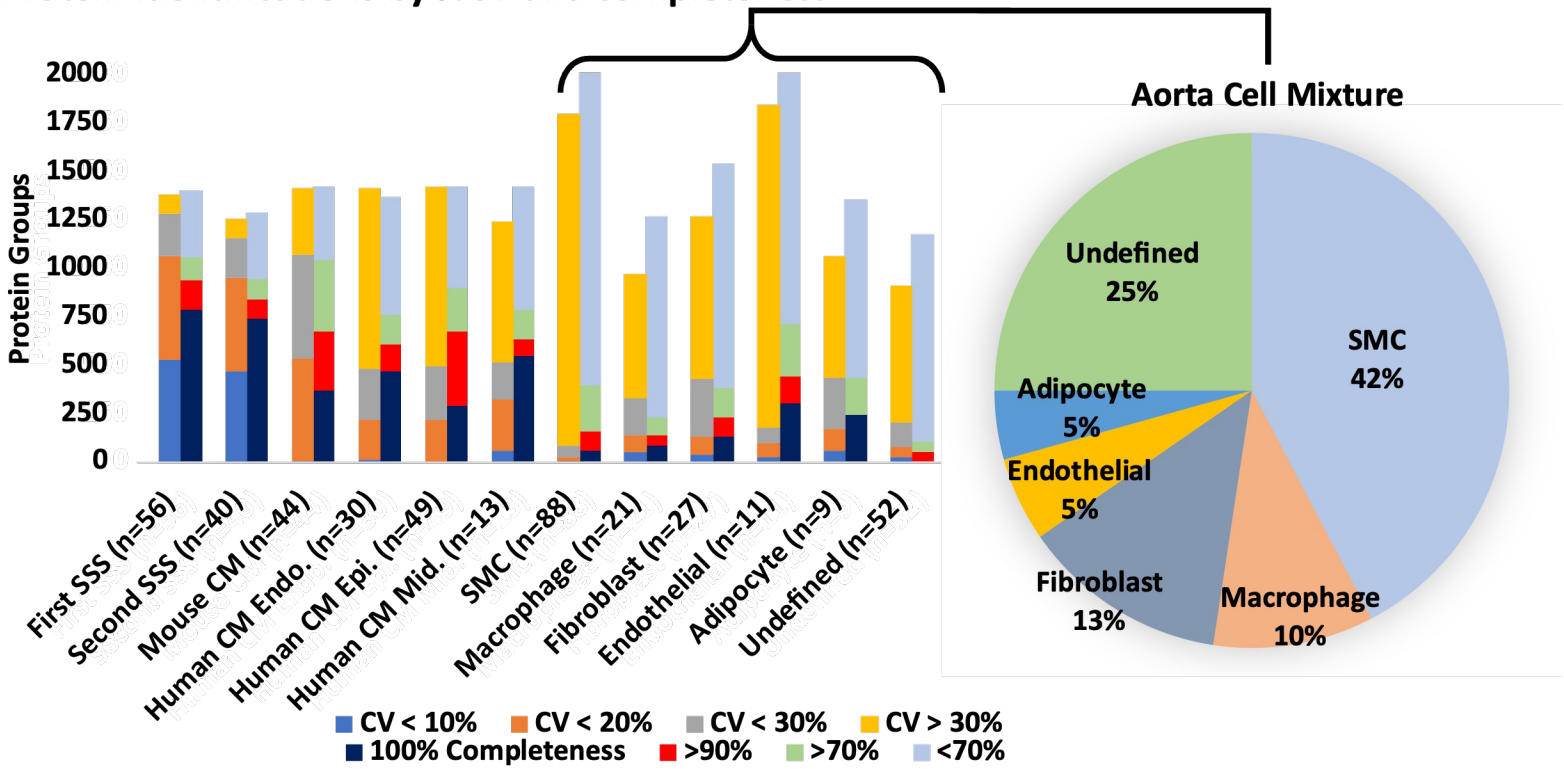
B) Platform pipeline

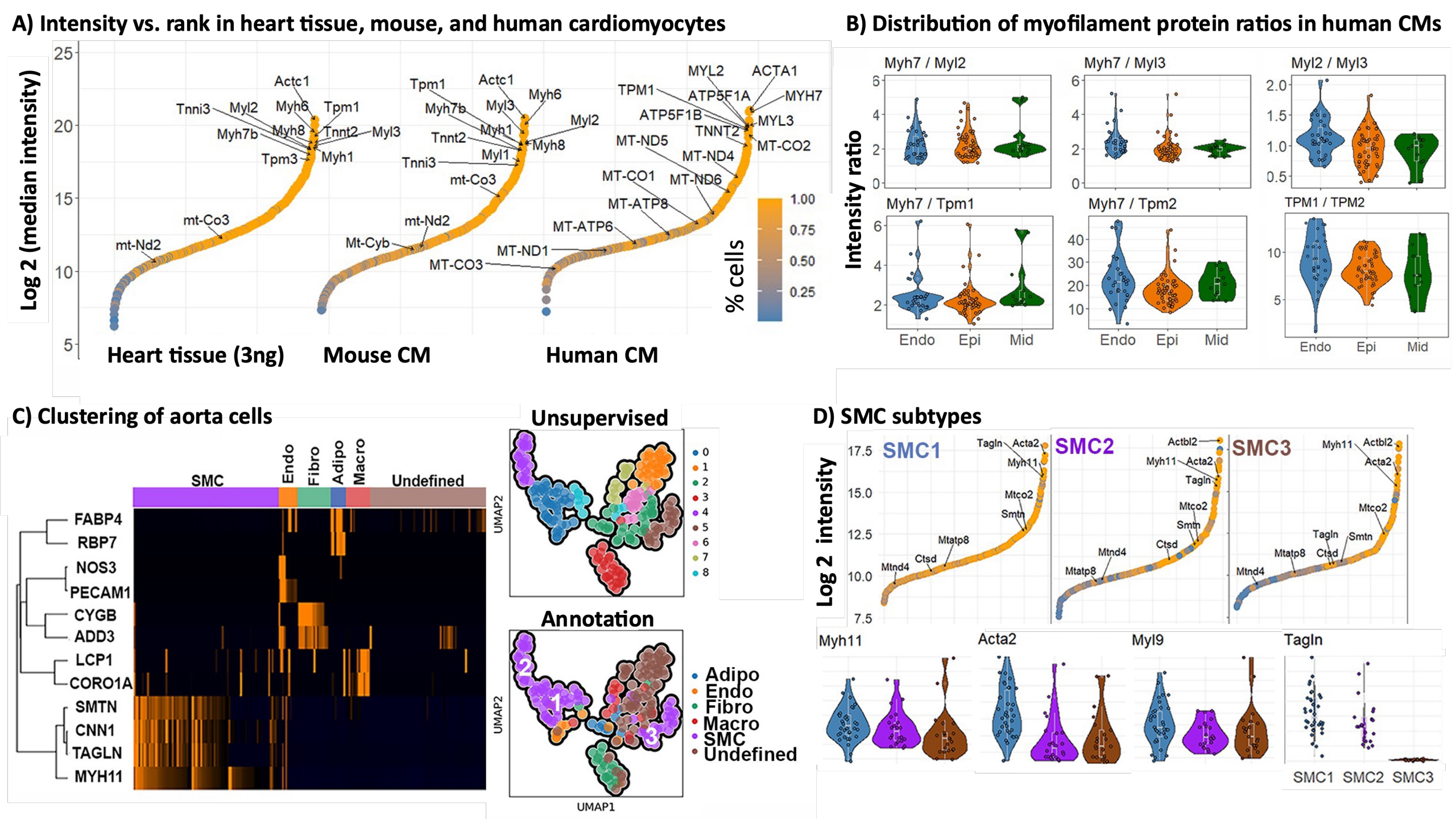


C) Longitudinal performance (4ng HeLa)



D) Protein identifications by %CV and completeness




A) Intensity vs. rank in heart tissue, mouse, and human cardiomyocytes
Log 2 (median intensity)
Heart tissue (3ng)
Mouse CM
Human CM
% cells
B) Distribution of myofilament protein ratios in human CMs
Intensity ratio
Myh7 / Myl2
Myh7 / Myl3
Myl2 / Myl3
Myh7 / Tpm1
Myh7 / Tpm2
TPM1 / TPM2
Endo
Epi
Mid
C) Clustering of aorta cells
SMC
Endo
Fibro
Adipo
Macro
Undefined
FABP4
RBP7
NOS3
PECAM1
CYGB
ADD3
LCP1
CORO1A
SMTN
CNN1
TAGLN
MYH11
Unsupervised
Annotation
UMAP1
UMAP2
Adipo
Endo
Fibro
Macro
SMC
Undefined
D) SMC subtypes
Log 2 intensity
SMC1
SMC2
SMC3
Myh11
Acta2
Myl9
Tagln
SMC1
SMC2
SMC3