Sample Report

Capybara™ Benchmarking Analysis

capybio.com

CAPYBIO™

Table of Contents

Executive Summary

Project Details

- Project ID: CAPY-2026-001
- Date: March 2026
- Sample(s): Two samples (Day 0 and FACS-sorted Day 14) of direct reprogramming of mouse cardiac fibroblasts to induced cardiomyocytes.
- Cells Analyzed: 6,413
- Reference Dataset: Mouse Cell Atlas: Neonatal Heart, Skin, Lung, Stomach (57 cell types).

Primary Conclusion

Capybara benchmarking analysis reveals a directionally successful but incomplete reprogramming of mouse cardiac fibroblasts toward a cardiomyocyte-like state. Atrial cardiomyocyte identity emerges strongly at Day 14, but substantial heterogeneity and non-cardiac populations remain in the final product despite enrichment of the on-target population via FACS sorting.

Key Results

- 68.7% of cells received a confident single-identity annotation.
- Differential composition analysis confirms a statistically significant shift in cell type proportions between conditions (FDR < 0.0001), with atrial cardiomyocytes increasing by approximately 75 % by Day 14.
- Reprogrammed cells display a clear bias towards atrial identity, 62.3% of Day 14 cells are annotated as atrial cardiomyocytes while only 6.3% are annotated as ventricle cardiomyocytes.
- 21.4% of cells carry mixed identities (Multi ID), with cardiac combinations emerging at Day 14 consistent with cells transitioning toward cardiac fate.
- In the Day 14 sample, a majority of Multi ID cells (63.7%) carry an atrial + ventricle cardiomyocyte identity; further protocol optimization could specify cells toward either an atrial or ventricular fate, producing a better in vitro representation of in vivo biology.
- In the Day 14 sample, 33.1% of Multi ID cells carry a cardiac + brown adipocyte mixed identity, flagging a potential off-target population in the reprogrammed product.
- Identity scores corroborate discrete annotations, with high atrial cardiomyocyte scores concentrated around the Day 14 UMAP region.

Recommended next steps

To gain further insight into these findings, we recommend the following:

- Review the Supplementary data for full cell type annotations, identity score tables, and per-cell classification results
- Perform differential gene expression analysis across annotated cell types to identify molecular drivers of each cell state
- Consider re-running CapyBio's Capybara Cell Benchmarking Analysis with an adult cardiac reference dataset to assess whether reprogrammed cells more closely resemble adult cardiomyocyte identity.
- Consider running CapyBio's CellOracle™ Differentiation Optimization Analysis to identify transcription factor perturbations that could improve reprogramming efficiency, fully specify reprogrammed cells toward either an atrial or a ventricular fate, and/or reduce off-target populations.

Background

Overview of User-Provided Data

Below, we provide an overview of your submitted dataset. This dataset consists of 6,413 cells profiled across two time points. To assess the quality of the input data, we report several standardized data quality metrics in Figure 1. As this dataset was provided pre-processed, no additional quality filtering was applied. Overall, the dataset displays data quality metrics consistent with a high-quality single-cell RNA-seq experiment, with sufficient gene/UMI detection and low mitochondrial content across all cells.

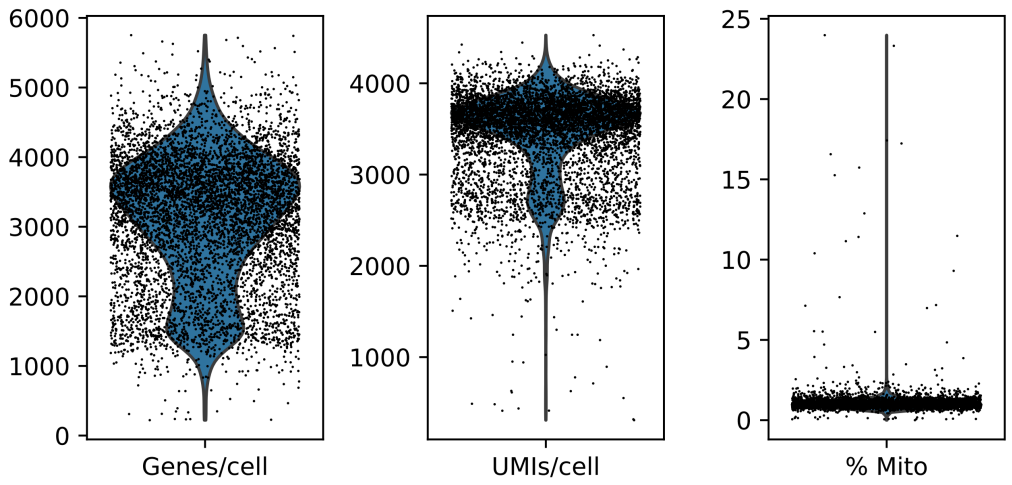

Figure 1. Quality control metrics for all cells in the dataset. Left: number of genes detected per cell. Center: total UMI (Unique Molecule Index) counts per cell. Right: percentage of reads mapping to mitochondrial genes.

Visualization of User-Provided Data

To visualize the overall structure of your data, we reduced the high-dimensional gene expression profiles to two dimensions using UMAP (Uniform Manifold Approximation and Projection). In this plot, cells with similar expression patterns appear closer together, forming clusters that often correspond to distinct cell types. This projection is shown in Figure 2. Cells collected at the earlier timepoint, Day 0, occupy distinct regions from those collected at the later timepoint, Day 14, reflecting progressive transcriptional changes during reprogramming.

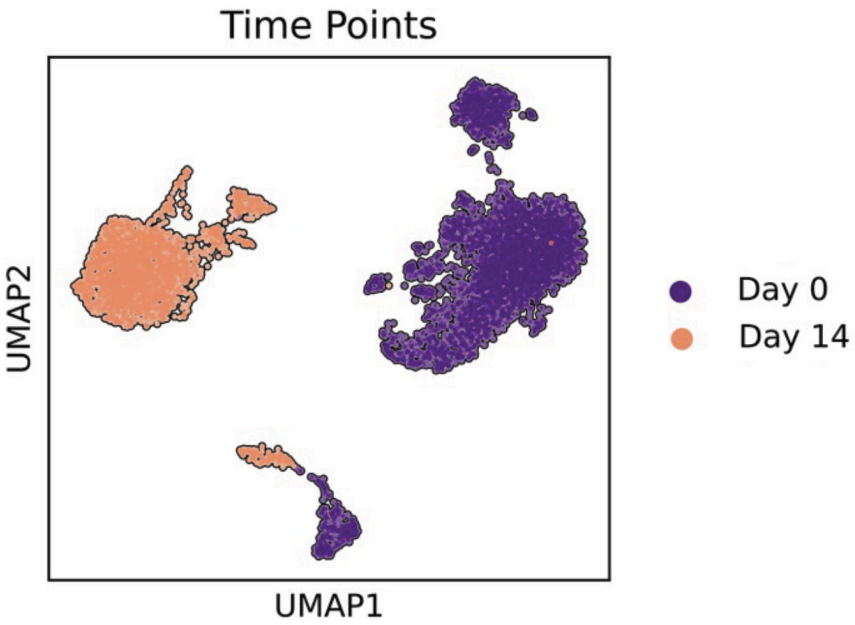

Figure 2. UMAP projection of all cells, colored by collection timepoint. Each dot represents a single cell; Day 0 cells are in purple and Day 14 cells are in light orange.

Overview of the Cell Annotation Reference Dataset

As requested, we used a high-resolution single-cell RNA sequencing reference dataset for cell identify analysis that spans four neonatal tissues: skin, heart, lung, and stomach. This reference dataset includes 57 distinct cell types, each annotated by domain experts using well-established biological markers. The cell types and their broader groupings are listed below:

Atrial:

Atrial Cardiomyocyte, Atrial Cardiomyocyte (cta2 high)

Ventricular:

Left ventricle cardiomyocyte (Myl2 high), Ventricle cardiomyocyte (Kcnj8 high)

Blood:

Dendritic cell, Erythroblast, Erythroblast (Car2 high), Erythroblast (Hba.x high), Erythroblast (Hbb.bs high), Erythroblast (Klf1 high), Erythroblast (Mt2 high), Erythroblast (Mt2, Mt1 high), Erythroblast (Snca high), Macrophage, Macrophage (Lyz2 high), Macrophage (Pf4 high), Mast cell, Neutrophil, Neutrophil (Gm5483 high), Neutrophil (Ngp high), Neutrophil (S100a8 high)

Muscle:

Cardiac muscle cell, Muscle cell, Muscle cell (Actc1 high), Muscle cell (Lrrc15 high), Smooth muscle cell, Smooth muscle cell (Acta2 high)

Stromal Cell:

Stromal cell (Akr1c18 high), Stromal cell (Ankfy1 high), Stromal cell (Cdkn1c high), Stromal cell (Col3a1 high), Stromal cell (Dcn high), Stromal cell (Fmod high), Stromal cell (Gas6 high), Stromal cell (Ptn high)

Other:

Acinar cell (Ctrb1 high), Adipocyte, Brown adipose tissue (Cidea high), Dividing cell, Endothelial cell, Endothelial cell (Igfbp5 high), Epithelial cell, Epithelial cell (Aldh1a2 high), Keratinocyte, Melanocyte, Neuron, Osteoblast (Ppic high), Vascular endothelial cell, Acinar cell (Spp1 high), Endocrine cell, Endocrine progenitor cell, Endothelial cell (Enpp2 high), Epithelial cell (Sftpc high), Immunocyte (Lyz2 high), Osteoblast (Dlk1 high), Progenitor cell, Stomach cell (Kazald1 high).

Results

Cell Annotation

Using the cell annotation reference dataset described above, each cell in the provided sample (or collective samples) was compared against all 57 cell types to determine its identity. Each cell receives an “identity score” for every reference type and is then assigned an identity label based on its closest match.

Most cells match clearly to a single cell type. These are labeled as “Discrete” cells. However, some cells closely resemble more than one reference type and are labeled as "Multi ID.” Multi ID cells can reflect cells that are transitioning between states or share features of multiple cell types. Cells that do not closely resemble any of the 57 reference types are labeled as "Unknown."

In the provided sample, 68.7% of cells were confidently assigned a single identity, 21.4% were assigned multiple identities, and 9.9% could not be confidently matched to any reference type. It is not uncommon to see a portion of the cells labeled as “Unknown.” These cells may reflect transitional or intermediate states that are not well-represented in the reference dataset, rather than low-quality cells. The per-sample breakdown is shown in Figure 3. Notably, the Discrete population increases from Day 0 to Day 14, and the Unknown fraction is substantially lower at Day 14 (2.0%) compared to Day 0 (14.4%).

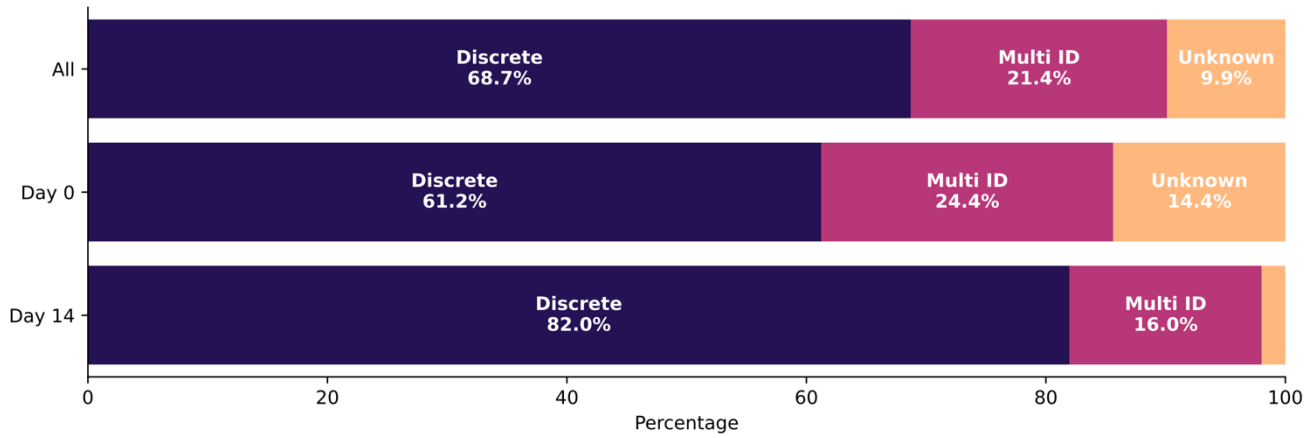

Figure 3. Overall annotation breakdown across all cells. 68.7% of cells were assigned a single cell type identity (Discrete), 21.4% matched more than one reference type (Multi ID), and 9.9% could not be confidently matched to any reference type (Unknown). Breakdown is shown for all cells combined (All), Day 0, and Day 14 separately.

Cell Type Classification

Of the 68.7% of cells that were confidently assigned a single identity, we identified 30 distinct cell types from across the 57 reference types. The distribution of the top 10 cell types in your two samples is shown in Figure 4 below. Atrial cardiomyocyte identity is almost exclusively present at Day 14, while macrophage, stromal cell, muscle cell, and smooth muscle cell are predominantly found at Day 0. By Day 14, we see 62.3% of cells identified as atrial cardiomyocytes.

Results for all cell types may be found in the Appendix.

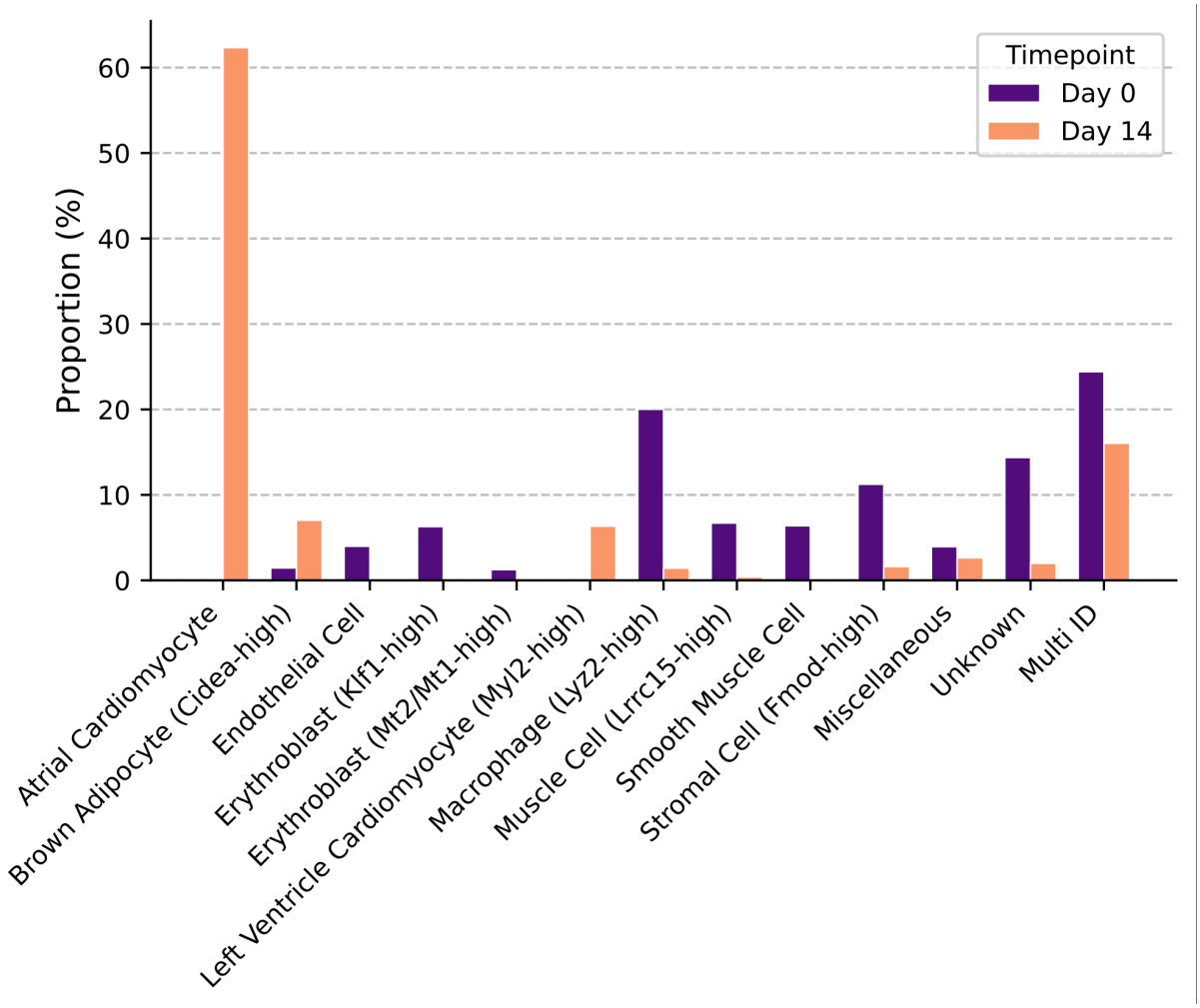

Figure 4. Cell type proportions across the two conditions. Day 0 is in purple and Day 14 in light orange.

Top Cell Types on UMAP

To visualize how the most common cell types are distributed across your dataset, we projected all cells onto a two-dimensional UMAP (Uniform Manifold Approximation and Projection). In this plot, cells with similar expression patterns appear closer together, forming clusters that often correspond to distinct cell types and well-defined transcriptional identities. Figure 5 shows the UMAPs for the top 10 cell types, unknown, and multi ID cells. The cells are colored by timepoint, allowing for a visual comparison of whether each cell type occupies distinct or overlapping regions of transcriptional space between Day 0 and Day 14. Cells belonging to all other cell types are shown in gray. Each annotated cell type, forms a well-defined transcriptional identity providing additional confidence in our cell type annotations.

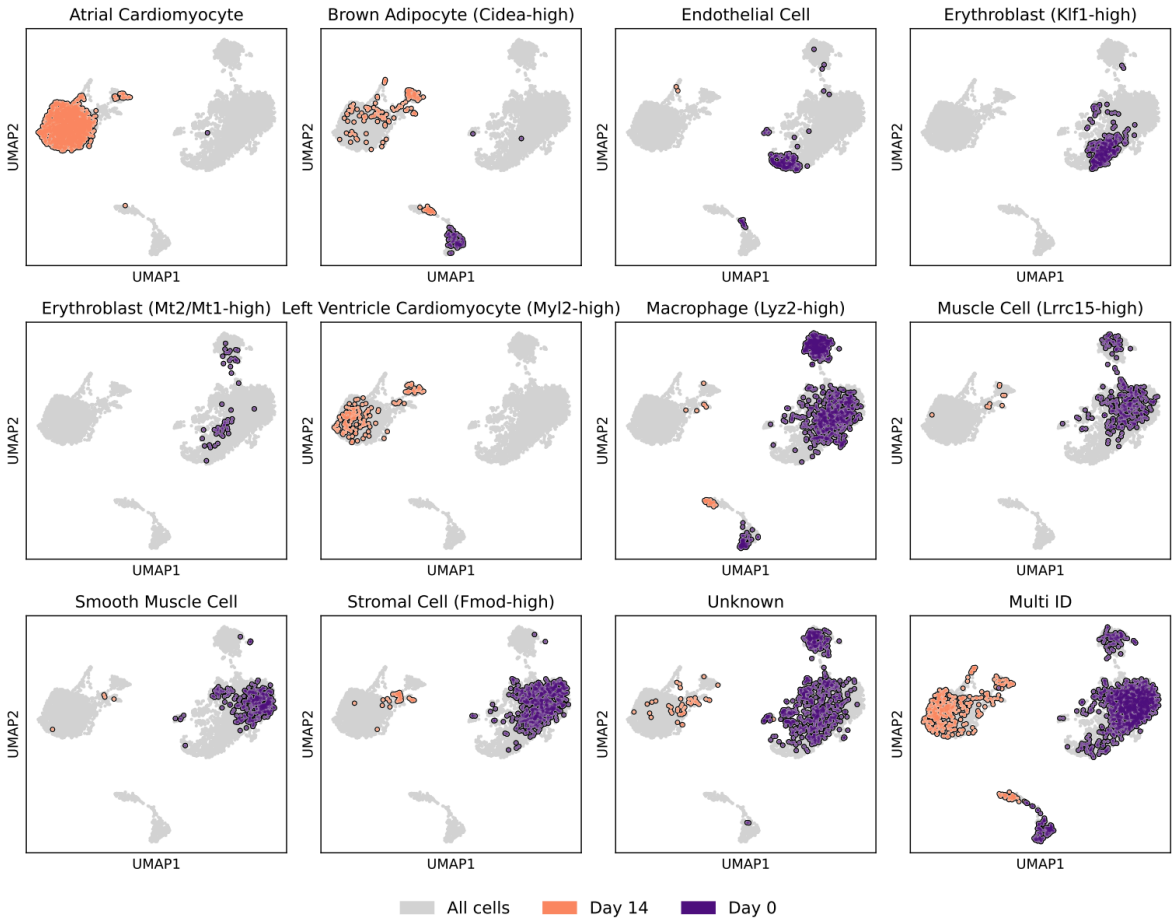

Figure 5. UMAP panels for each of the top 10 cell types, unknown, and multi ID cells. Cells are colored by time point: Day 0 (purple) and Day 14 (orange). Grey dots represent all other cells.

Differential Composition Analysis

To better quantify how cell type composition differs between Day 0 and Day 14, we compared the relative proportion of each cell type across the two conditions. For each cell type, we performed a permutation test by randomly shuffling condition labels across all cells 10,000 times and computing the difference in proportion for each permutation. The observed difference was then compared against this null distribution to obtain a p-value. To account for testing across multiple cell types, we applied a Benjamini-Hochberg correction to control the false discovery rate.

Figure 6 shows the difference in proportion for the top 10 most abundant cell types, all of which differ significantly between Day 0 and Day 14 (FDR < 0.05). Atrial cardiomyocytes are the most strongly enriched at Day 14, while macrophage, stromal, and muscle subtypes are markedly depleted, consistent with a progressive shift toward cardiac identity over the reprogramming time course. We also observe the emergence of Brown Adipocyte (Cidea-high) identity on Day 14, indicate possible off-target reprogramming of a subset of the starting cell population.

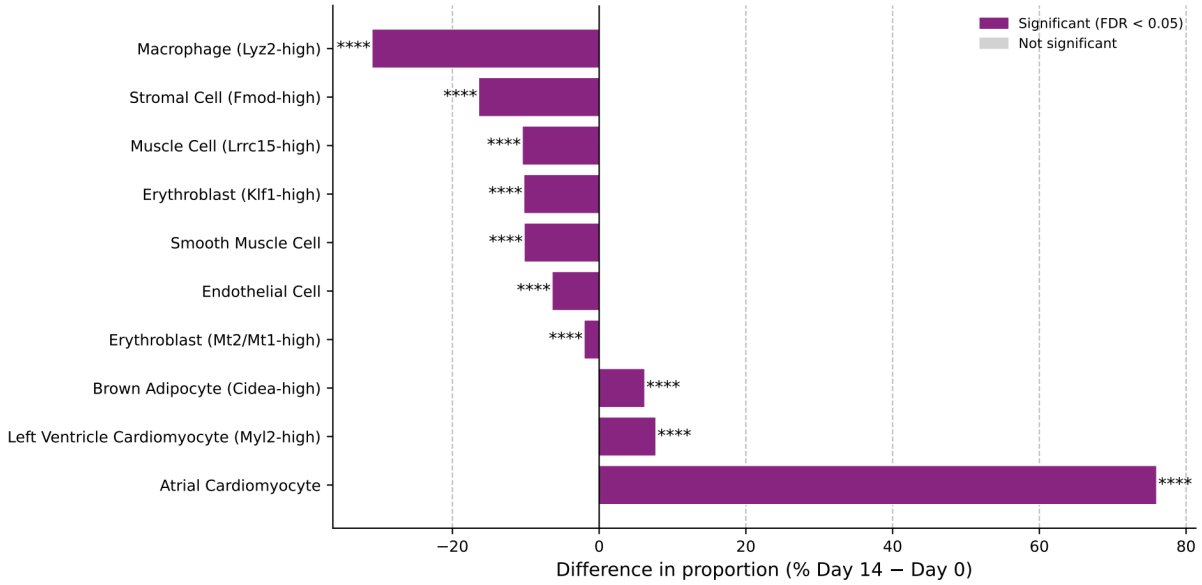

Figure 6. Difference in cell type proportion between Day 0 and Day 14. Positive values indicate enrichment at Day 14; negative values indicate enrichment at Day 0. Significance was assessed by permutation test (10,000 permutations) with Benjamini-Hochberg correction. All shown cell types reached statistical significance (FDR < 0.05). *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001, ****p ≤ 0.0001; ns = not significant. Note: in the absence of biological replicates, these differences reflect the two submitted samples and may not generalize.

Multi ID Analysis

A total of 21.4% of cells were assigned multiple identities (multi ID cells in Figure 3). In this section, we further analyzed these cells to characterize which identity combinations are most prevalent and how they are distributed across the two timepoints. Multi ID cells arise when a cell's transcriptional profile confidently matches more than one reference type, and this can reflect transitional states, shared transcriptional programs between related cell types, or cells that have not yet committed to a single identity. Figure 7 shows the distribution of the top 10 "Multi ID" annotations across the two samples. Full results are available in the Appendix.

As shown, multi ID cells at Day 0 are predominantly combinations of non-cardiac types, with macrophage-stromal and macrophage-muscle pairings being the most common. At Day 14, the multi ID landscape shifts markedly toward cardiac combinations, with atrial cardiomyocyte- left ventricle cardiomyocyte emerging as the dominant pairing, suggesting that a subset of reprogrammed cells is transitioning between or co-expressing atrial and ventricular cardiac programs. Notably, a subset of Day 14 cells co-express atrial cardiomyocyte and brown adipocyte (Cidea-high) identity scores, representing an unexpected mixed-identity population that may warrant further investigation as a potential off-target reprogramming outcome.

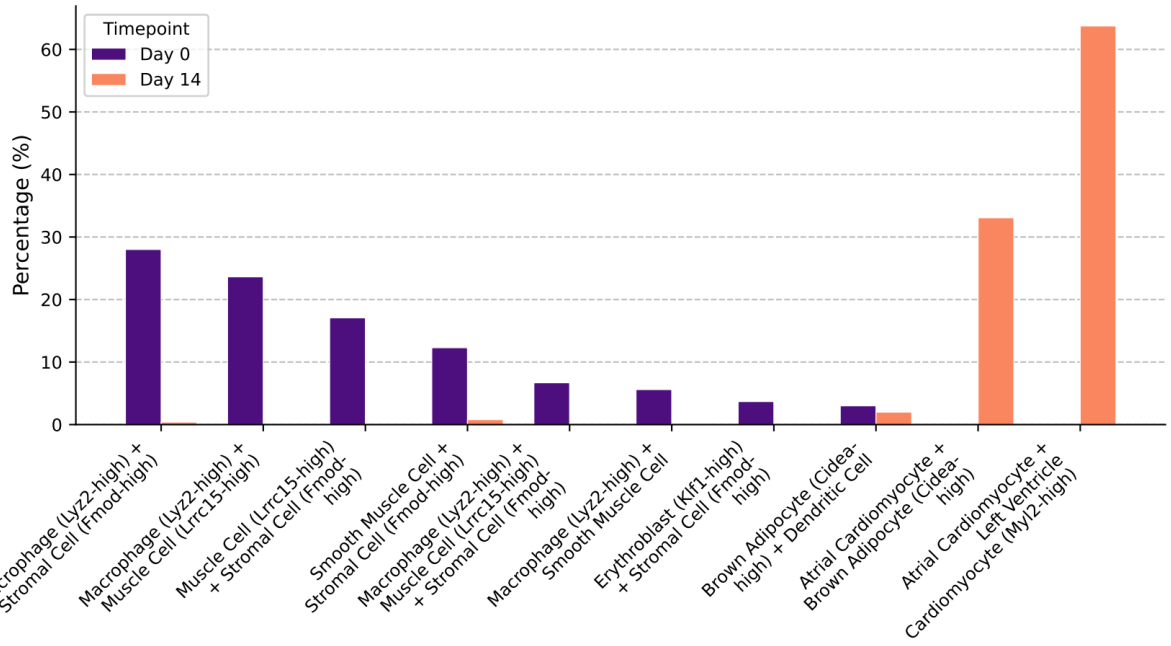

Figure 7. Distribution of top 10 "Multi ID" combinations across the two conditions. Day 0 is in purple and Day 14 in light orange.

Identity Score Distribution

The cell type annotations shown earlier are based on underlying identity scores, a continuous measure of how closely each cell's gene expression matches each reference cell type. Figure 8 visualizes those scores, projected onto the UMAP for the 10 most abundant cell types. Brighter color indicates a stronger match. This serves two purposes: first, it allows you to verify that annotations are supported by strong underlying signals. Second, because these scores are computed on a consistent scale against the same reference, they can be compared across experiments, allowing one to track whether identity strength changes between batches, protocols, or conditions.

As shown, identity score distributions mirror the discrete cell type annotations in Figure 5. For example, regions of high atrial cardiomyocyte score align with where atrial cardiac cells are annotated, while non-cardiac scores concentrate in other regions of the UMAP. Together, this confirms that the transcriptional shift observed across conditions reflects a genuine change in underlying cell identity. Top 10 identities are shown below with full results in Appendix.

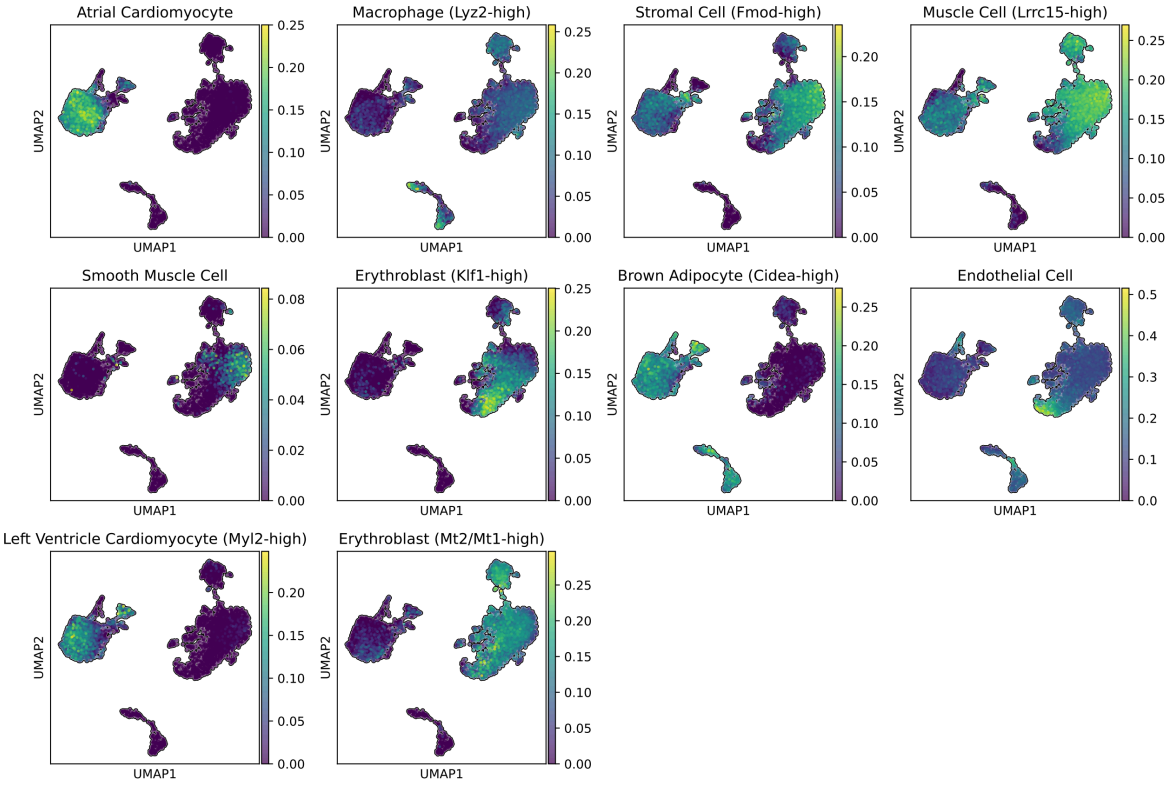

Figure 8. Continuous identity scores projected onto the UMAP for the top 10 most abundant cell types. Each panel shows one reference type, with color intensity (dark to bright) indicating how closely each cell matches that type.

Conclusions

The Capybara Cell Benchmarking Analysis of the submitted reprogramming dataset reveals a progressive but incomplete acquisition of cardiac-like transcriptional identity over the course of reprogramming. At Day 14, atrial cardiomyocyte annotations emerge as the dominant cell type, representing a marked shift from the fibroblast-like starting population at Day 0. Differential composition analysis confirms that this shift is statistically significant, with the atrial cardiomyocyte proportion increasing by approximately 75 % between conditions (FDR < 0.0001). However, heterogeneity remains on Day 14 in the form of non-cardiac identities, including macrophage, stromal, and muscle subtypes, and the emergence of brown adipocyte (Cidea-high) identity. Multi ID analysis further reveals that mixed-identity cells at Day 0 are predominantly non-cardiac combinations, while at Day 14 the dominant pairing shifts to atrial cardiomyocyte-left ventricle cardiomyocyte, suggesting a subset of cells transitioning between cardiac subtypes. A secondary mixed-identity population co-expressing atrial cardiomyocyte and brown adipocyte programs is also present at Day 14 and may represent an off-target reprogramming outcome warranting further investigation. It is worth noting that cell identities were assigned using a neonatal reference dataset; cells annotated as cardiomyocytes therefore reflect resemblance to neonatal cardiac identity, which may differ from mature adult cardiomyocyte programs. Together, these findings indicate that reprogramming is directionally successful but incomplete, with some off-target and residual populations persisting in the final product.

Recommended Next Steps

To gain further insight into these findings, we recommend the following:

- Review the Supplementary data for full cell type annotations, identity score tables, and per-cell classification results
- Perform differential gene expression analysis across annotated cell types to identify molecular drivers of each cell state
- Consider re-running CapyBio's Capybara Benchmarking Analysis with an adult cardiac reference dataset to assess whether reprogrammed cells more closely resemble mature cardiomyocyte identity
- Consider running CapyBio's CellOracle Differentiation Optimization Analysis to identify transcription factor perturbations that could improve reprogramming efficiency and reduce off-target populations.

This sample report uses publicly available data that was originally analyzed as part of our scientific publication on Capybara[1].

[1] Kong, Wenjun, Yuheng C. Fu, Emily M. Holloway, et al. "Capybara: A Computational Tool to Measure Cell Identity and Fate Transitions." Cell Stem Cell 29, no. 4 (2022): 635-649.e11. https://doi.org/10.1016/j.stem.2022.03.001.

Methods

Cell identity scoring. Cell identity was quantified using Capybara, CapyBio's proprietary cell identity scoring platform. Unlike conventional annotation approaches that assign a single label based on a handful of marker genes, Capybara compares each cell's full transcriptional profile against a curated reference dataset to produce a continuous identity score for every reference cell type. This provides a richer, more complete picture of cell identity, capturing not just what a cell most closely resembles, but how confidently it matches that type and whether it shares features with other types. Scores range from 0 to 1, with higher values indicating stronger resemblance. The reference was constructed from the Mouse Cell Atlas neonatal dataset spanning 57 cell types across heart, skin, lung, and stomach.

Reference dataset. Cell identity scoring was performed against a high-resolution single-cell RNA sequencing reference derived from the Mouse Cell Atlas, spanning four neonatal tissues: heart, skin, lung, and stomach. The reference comprises 57 expert-annotated cell types. A pseudobulk reference profile was constructed for each cell type by aggregating 90 cells, and query and reference expression matrices were intersected to a common gene set, library-size normalized, and log-transformed prior to scoring.

Cell type annotation. Each identity score was evaluated for statistical significance by comparison against empirical background distributions. Cells confidently matching a single type were labeled Discrete. Cells significantly matching more than one type were labeled Multi ID, and may reflect transitional states or shared transcriptional programs between related cell types. Cells not significantly matching any reference type were labeled Unknown.

Differential composition analysis. To test whether cell type proportions differ between conditions, a permutation test was performed for each cell type by randomly shuffling condition labels across all cells 10,000 times. Benjamini-Hochberg correction was applied to control the false discovery rate. In the absence of biological replicates, results reflect differences between the two submitted samples and may not generalize.

Multi ID analysis. Co-occurring identity combinations among Multi ID cells were tabulated and normalized per condition to account for differences in total cell counts between timepoints. The top 10 most frequent combinations are reported.

Report generated by CapyBio Inc
Analysis powered by Capybara
Questions: **info@capybio.com**