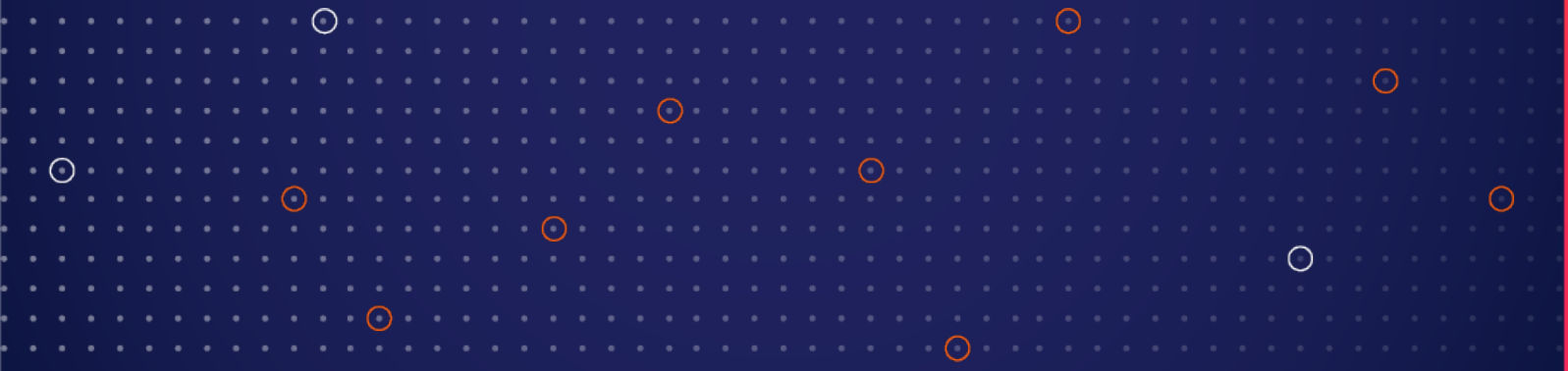
Sample Report

Capybara™ Advanced Analysis



capybio.com

Table of Contents

Executive Summary

Project Details

- Project ID: CAPY-2026-002
- Date: March 2026
- Sample(s): Six samples (Day 0, 1, 2, 3, 7, and FACS-sorted Day 14) of direct reprogramming of mouse cardiac fibroblasts to induced cardiomyocytes.
- Cells Analyzed: 30,651
- Reference Dataset: Mouse Cell Atlas: Neonatal Heart, Skin, Lung, Stomach (57 cell types).

Primary Conclusion

Capybara advanced analysis of the submitted reprogramming dataset reveals a progressive and directionally successful acquisition of cardiac-like transcriptional identity over the course of reprogramming, with atrial cardiomyocyte identity becoming the dominant population by Day 14. However, meaningful heterogeneity persists in the form of off-target populations, mixed-identity cells, and residual non-cardiac identities. Differential gene expression and pathway enrichment analyses reveal key genes and gene programs that drive each major cell type, providing actionable insights for downstream experimental work.

Key Results

- 65.4% of cells received a confident single-identity annotation across all six samples, with atrial cardiomyocytes representing 62.3% of Day 14 cells.
- Differential composition analysis confirms a statistically significant and progressive shift in cell type proportions relative to Day 0 across all timepoints (FDR < 0.05), with atrial cardiomyocytes showing the strongest enrichment over time.
- Macrophage, stromal, and smooth muscle subtypes are depleted early and remain consistently reduced relative to Day 0, consistent with clearance of the starting fibroblast-like population.
- Brown adipocyte (Cidea-high) identity emerges at Day 7 and persists through Day 14, flagging a potential off-target reprogramming outcome.
- 19.4% of cells carry mixed identities (Multi ID); by Day 14, 42.8% of Multi ID cells co-express atrial and left ventricle cardiomyocyte identity, suggesting a subset of cells in transition between cardiac subtypes.
- 22.2% of Day 14 Multi ID cells carry a cardiac + brown adipocyte mixed identity, warranting further investigation as a potential off-target population.

- Differential gene expression and pathway enrichment analyses reveal key genes and gene programs that drive each major cell type, providing insight that can be used to guide downstream experimental work such as identifying marker sets for FACS purification, qPCR analysis, functional assays, etc.

Recommended

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

- Review the Supplementary data for full cell type annotations, identity score tables, Multi ID results, and DEG and GSEA results.
- Investigate the brown adipocyte (Cidea-high) population further to determine whether it represents a stable off-target fate or a transient intermediate state and consider protocol modifications to reduce its emergence.
- Consider re-running CapyBio's Capybara Advanced Analysis with an adult cardiac reference dataset to assess whether reprogrammed cells more closely resemble mature cardiomyocyte identity, which may not be fully captured by the neonatal reference used here.
- Consider running CapyBio's CellOracle™ Differentiation Optimization Analysis to identify transcription factor perturbations that could improve reprogramming efficiency, reduce off-target populations, and potentially specify cells toward either an atrial or ventricular fate.

Background

Overview of User-Provided Data

Below, we provide an overview of your submitted dataset. This dataset consists of 30,651 cells profiled across six 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.

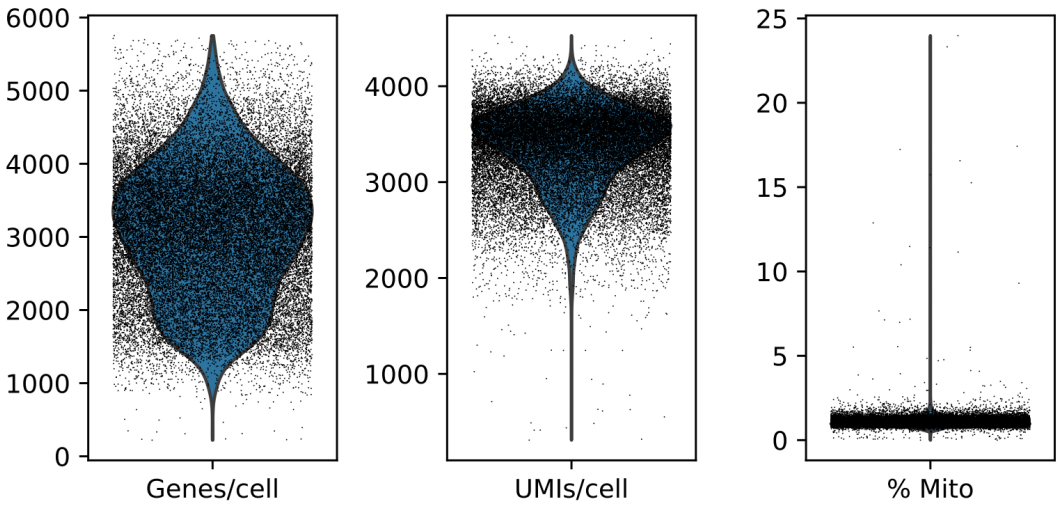


Figure 1. Quality control metrics for all cells in the dataset. Left: number of genes detected per cell. Center: total UMI (Unique Molecule Index) counts per cell. Right: percentage of reads mapping to mitochondrial genes.

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 timepoints, e.g. Day 7 and Day 14, reflecting progressive transcriptional changes during reprogramming.

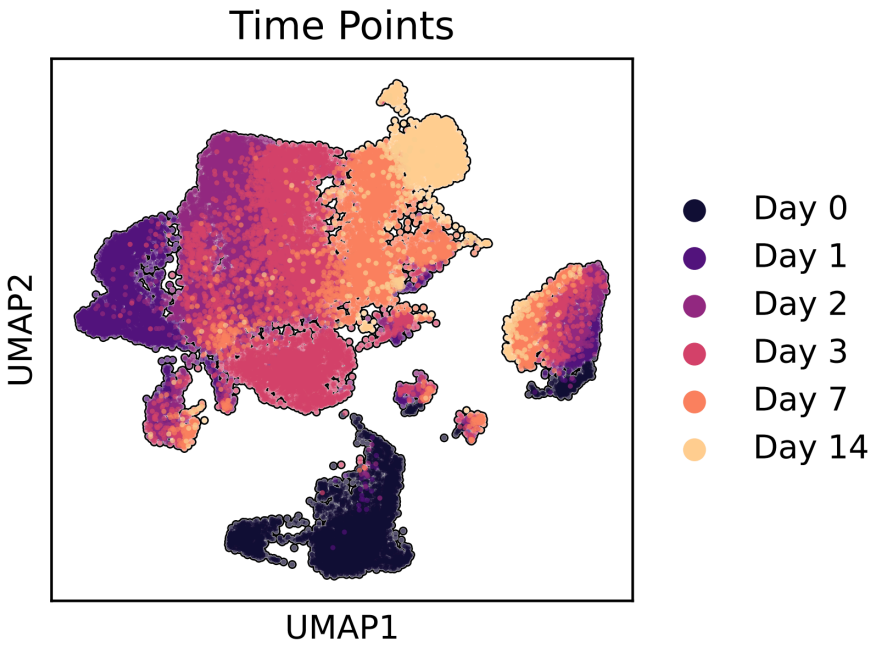


Figure 2. UMAP projection of all cells, colored by collection timepoint. Each dot represents a single cell. Cells are colored based on sample of origin.

Overview of the Cell Annotation Reference Dataset

As requested, we used a high-resolution single-cell RNA sequencing reference dataset for cell identity 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, 65.4% of cells were confidently assigned a single identity, 19.4% were assigned multiple identities, and 15.2% 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. The Unknown fraction increases during the intermediate timepoints, eventually reducing to just 2% in the Day 14 sample.

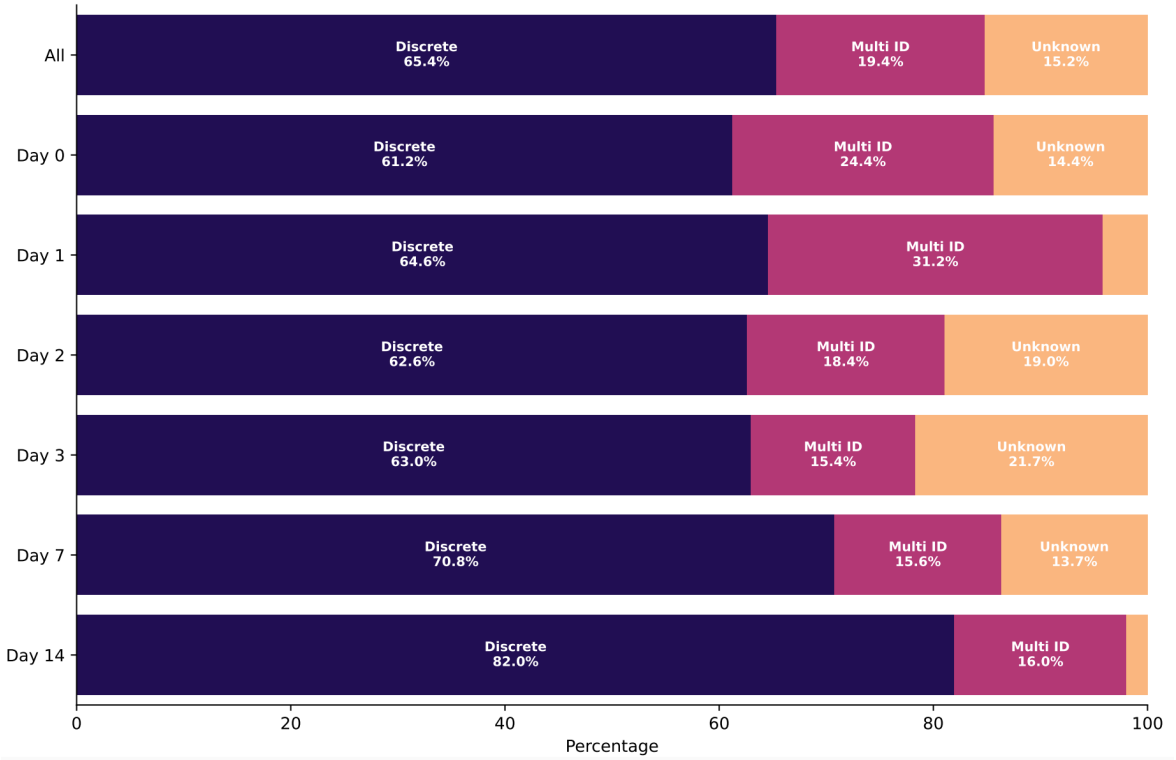


Figure 3. Overall annotation breakdown across all cells. 65.4% of cells were assigned a single cell type identity (Discrete), 19.4% matched more than one reference type (Multi ID), and 15.2% could not be confidently matched to any reference type (Unknown). Breakdown is shown for all cells combined (All), and each sample separately.

Cell Type Classification

Of the 65.4% 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 across your samples is shown in Figure 4 below. Atrial cardiomyocyte identity is the most dominant identity by Day 14 (62.3% of all Day 14 cells), while macrophage, stromal cell, muscle cell, and smooth muscle cell are predominantly found in early timepoints.

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

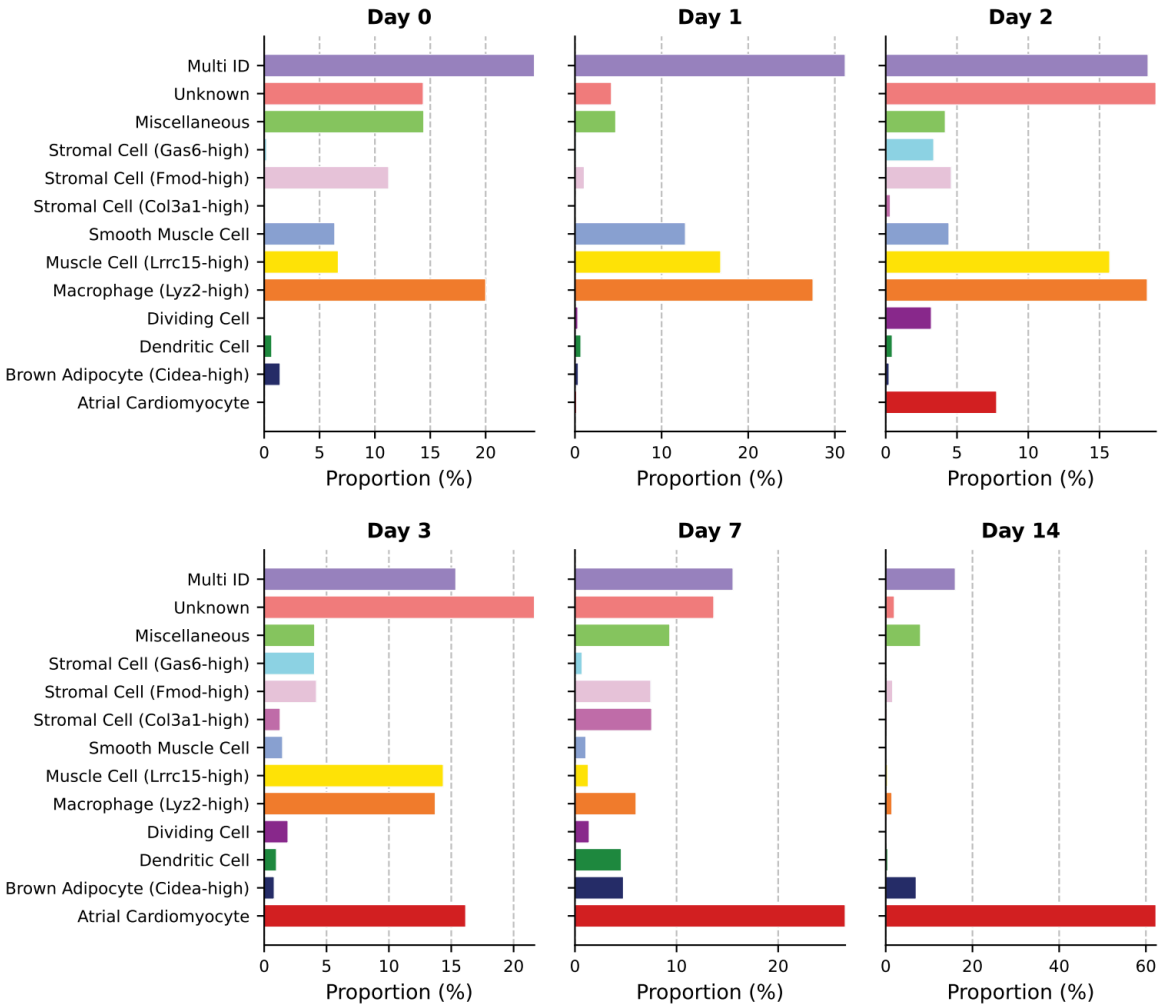


Figure 4. Bar plots of cell type proportions per sample. Top 10 cell types, alongside Multi ID, Unknown are shown. Remaining cell types are grouped into "Miscellaneous."

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 sample, allowing for a visual comparison of whether each cell type occupies distinct or overlapping regions of transcriptional space between them. Cells belonging to all other cell types are shown in gray. For clarity, we only show two samples, Day 14 and Day 0. Full results are available in the Appendix. Each annotated cell type forms a well-defined transcriptional identity providing additional confidence in our cell type annotations.

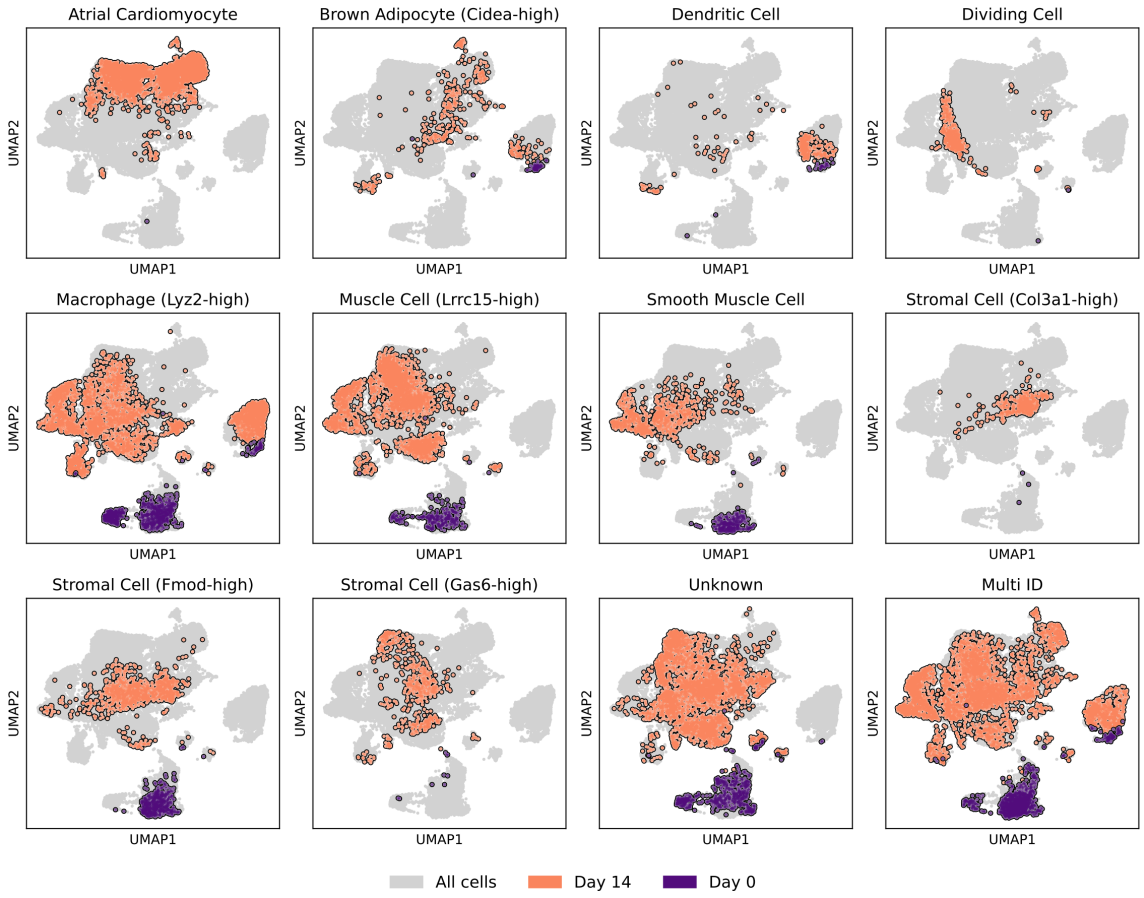


Figure 5. UMAP panels for each of the top 10 cell types, unknown, and multi ID cells. Cells are colored by time point: Day 0 (purple) and Day 14 (orange). Grey dots represent all other cells.

Differential Composition Analysis

To better quantify how cell type composition differs across samples, we compared the proportion of each cell type across samples relative to a designated reference sample (Day 0). For each cell type, we performed a permutation test by randomly shuffling sample 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. Results are shown for all samples relative to the reference sample.

Figure 6 shows the difference in proportion for the top 10 most abundant cell types relative to Day 0 across all samples. All displayed cell types differ significantly from Day 0 at one or more timepoints (FDR < 0.05). Atrial cardiomyocytes show a progressive and marked enrichment over the time course, becoming the dominant population by Day 7. Macrophage, stromal, and smooth muscle subtypes are depleted early and remain consistently reduced relative to Day 0. Brown adipocyte (Cidea-high) identity emerges at Day 7 and persists through Day 14, suggesting off-target reprogramming of a subset of the starting cell population that becomes more pronounced at later timepoints.

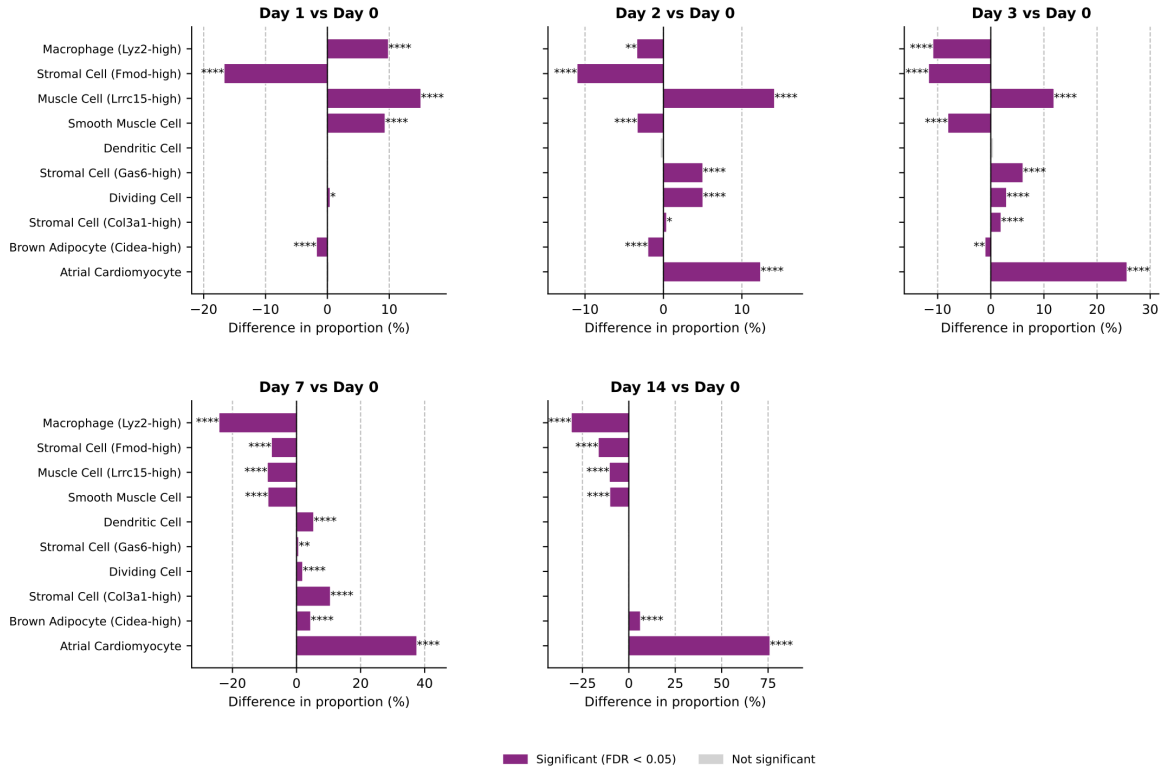


Figure 6. Difference in cell type proportion across samples, relative to Day 0. Positive values indicate enrichment, and negative values indicate depletion relative to Day 0. Significance was assessed by permutation test (10,000 permutations) with Benjamini-Hochberg correction. All shown cell types reached 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 19.4% 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 evolve across the reprogramming time course. 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 top 10 Multi ID combinations with the highest variance in proportion across timepoints. Full results are available in the Appendix.

As shown, Multi ID cells at Day 0 are predominantly non-cardiac combinations, with macrophage-stromal and macrophage-muscle pairings being the most common. These combinations progressively decline over the time course, while cardiac combinations emerge at later timepoints. By Day 14, atrial cardiomyocyte-left ventricle cardiomyocyte becomes 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, representing an unexpected mixed-identity population that may warrant further investigation as a potential off-target reprogramming outcome.

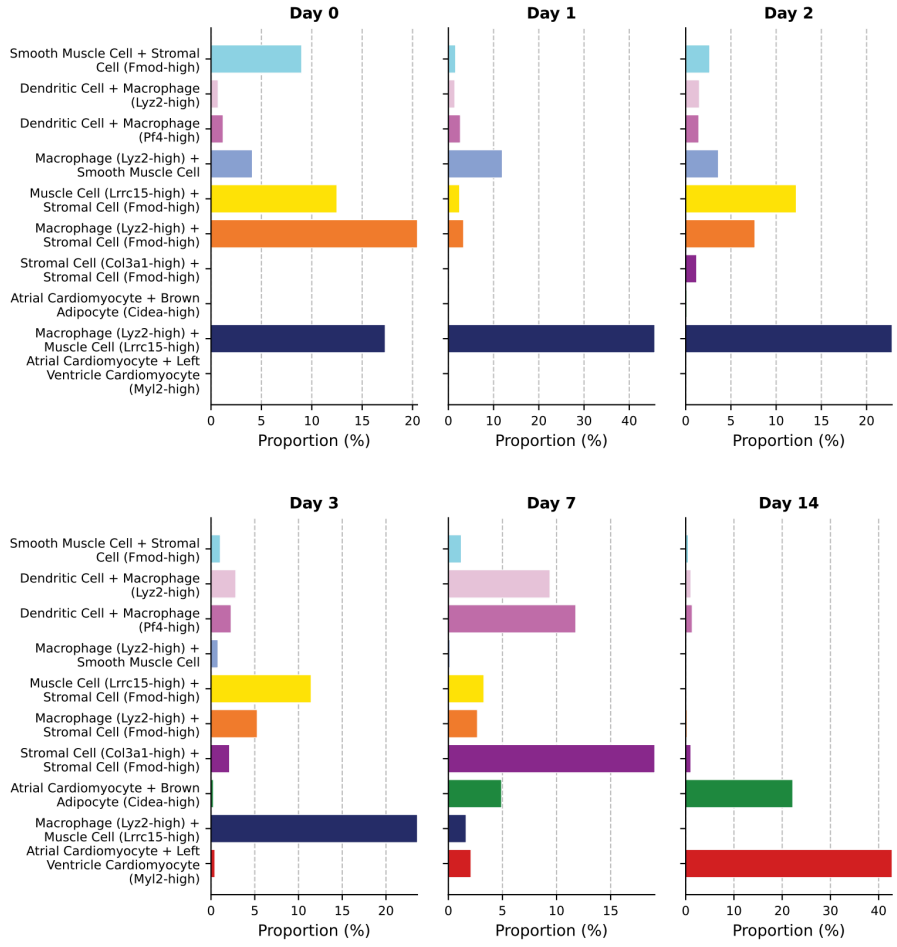


Figure 7. Top "Multi ID" combinations across all samples. Top 10 most variable multi-ID combinations across samples are shown.

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.

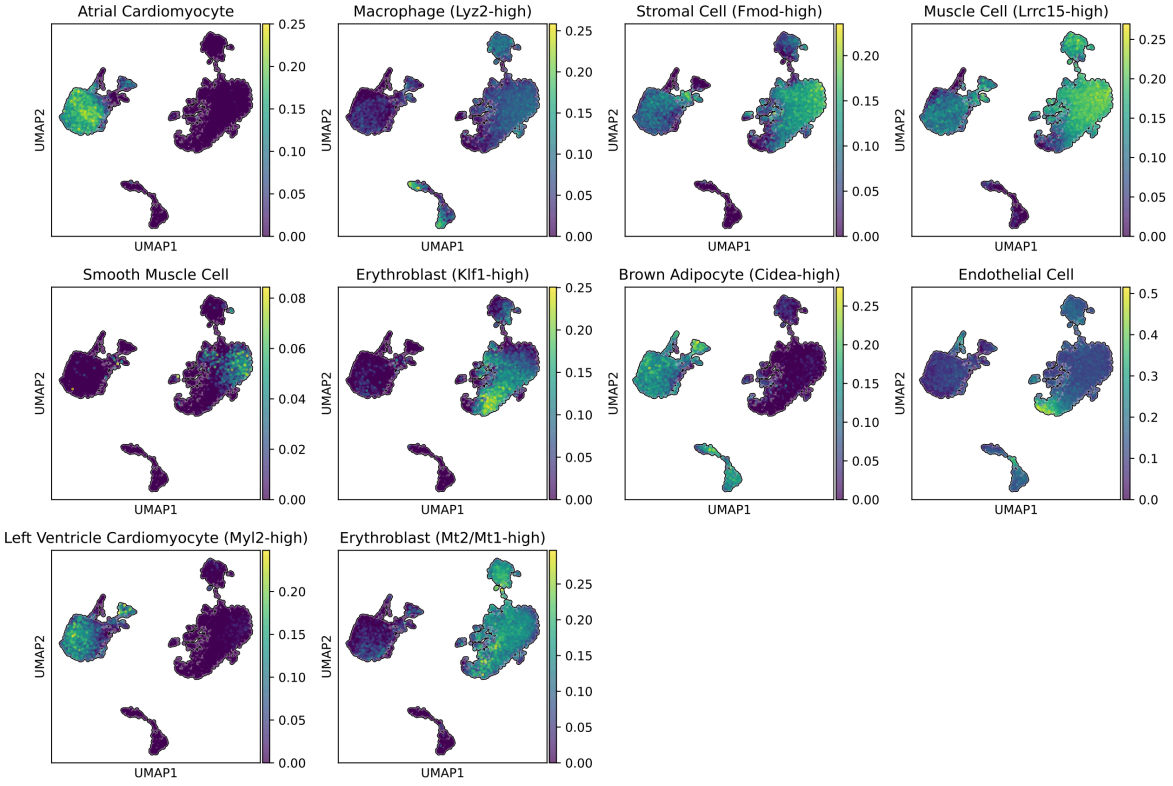


Figure 8. Continuous identity scores projected onto the UMAP for the top 10 most abundant cell types. Each panel shows one reference type, with color intensity (dark to bright) indicating how closely each cell matches that type.

Differential Gene Expression Analysis

To characterize the transcriptional identity of each annotated cell type, we performed differential gene expression (DEG) analysis comparing each cell type against all remaining cells within the same sample. Genes were ranked by adjusted p-value using a Wilcoxon rank-sum test, and significance was defined as an adjusted p-value below 0.05 with an absolute log fold change exceeding 0.5. Figure 9 shows volcano plots for the six most abundant discrete cell types, with upregulated genes shown in red and downregulated genes in blue. The top differentially expressed genes by fold change are labeled for each cell type. Full DEG results for all cell types are available in the Appendix.

As shown, each annotated cell type exhibits a distinct pattern of differential gene expression. Atrial cardiomyocytes show enrichment of genes such as Actc1 alongside additional differentially expressed features including Pon2. Dendritic cells are characterized by genes such as Fth1, Ftl1, and Ctss. Macrophage (Lyz2-high) cells exhibit differential expression of genes including Fabp5 and Tagln2, and muscle cells (Lrrc15-high) show enrichment of Acta2. Overall, these patterns capture statistically significant differences across cell populations.

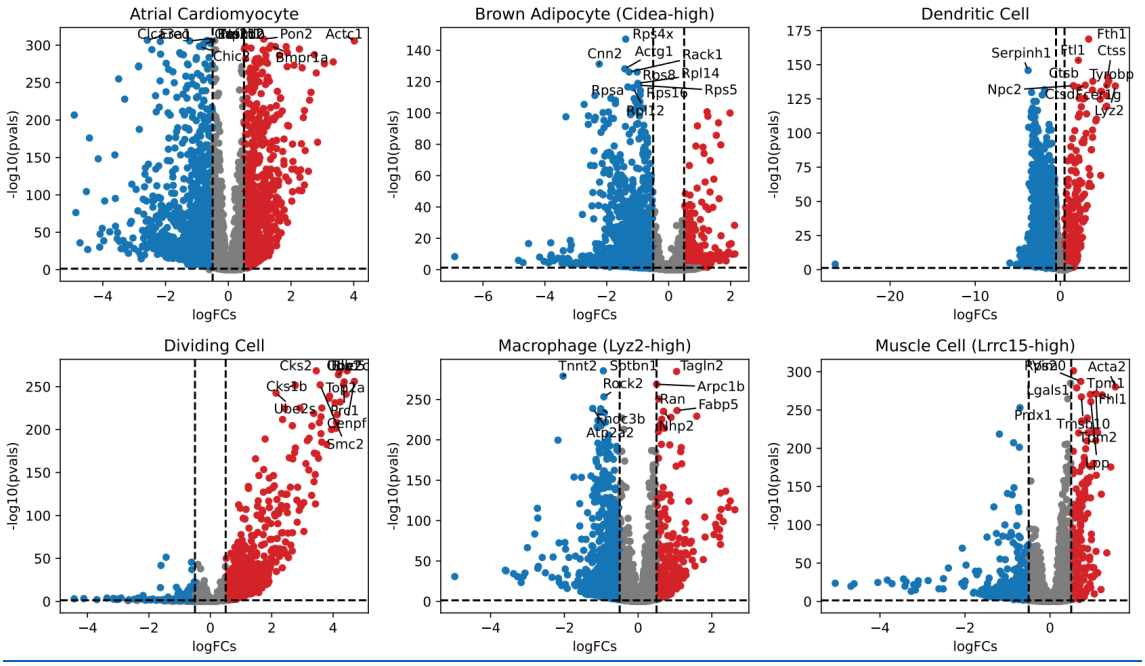


Figure 9. Volcano plots showing differential gene expression results for top 6 cell types. Volcano plots show log fold change (x-axis) against -log10 adjusted p-value (y-axis) for each cell type, comparing each cell type against all remaining cells. Significantly upregulated genes (log fold change > 0.5, adjusted p-value < 0.05) are shown in red, downregulated genes in blue, and remaining genes in gray. The top differentially expressed genes are labeled. Dashed lines indicate significance thresholds.

Gene Set Enrichment Analysis

To identify the biological processes underlying each annotated cell type, we performed gene set enrichment analysis (GSEA) using the MSigDB Hallmark gene sets. For each cell type, genes were ranked by Wilcoxon test statistic and enrichment was assessed against 50 curated hallmark pathways. Figure 10 shows the top 8 most significantly enriched pathways per cell type. Dot size reflects statistical significance and color reflects enrichment direction, with red indicating positive enrichment and blue indicating depletion.

As shown, each cell type displays a biologically coherent enrichment profile. Dividing cells are strongly enriched for proliferative programs including G2M checkpoint, E2F targets, and mitotic spindle, consistent with their actively cycling state. Dendritic cells show enrichment of interferon gamma and alpha response pathways alongside complement, reflecting their innate immune identity. Macrophage (Lyz2-high) cells are enriched for MYC targets and mTORC1 signaling, consistent with an activated metabolic state. Brown adipocyte (Cidea-high) cells show the expected adipogenesis signature alongside apical junction, reflecting their lipid-handling identity. Muscle cells (Lrrc15-high) display enrichment of reactive oxygen species and apical junction pathways, consistent with metabolically active contractile cells. Atrial cardiomyocytes show depletion of inflammatory and TNF signaling programs, suggesting transcriptional suppression of immune-related pathways in the dominant cardiac population, which may reflect the relatively quiescent inflammatory state expected of maturing cardiomyocytes.

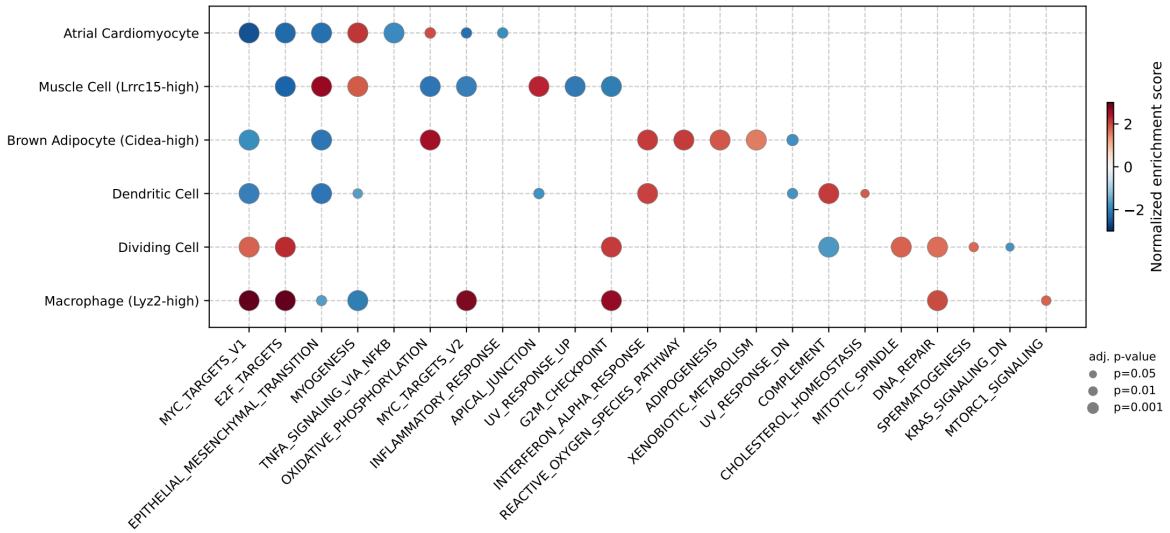


Figure 10. Gene set enrichment analysis for the six most abundant discrete cell types. Dot size reflects statistical significance (-log10 adjusted p-value) and dot color reflects the normalized enrichment score, with red indicating positive enrichment and blue indicating depletion. Only pathways with adjusted p-value < 0.05 are shown.

Conclusions

The Capybara Advanced Analysis of the submitted reprogramming dataset reveals a progressive but incomplete acquisition of cardiac-like transcriptional identity over the course of reprogramming. Atrial cardiomyocyte identity emerges as the dominant cell type by Day 14, representing a marked shift from the fibroblast-like starting population at Day 0, and differential composition analysis confirms this shift is statistically significant across all sampled timepoints. However, meaningful heterogeneity remains, including residual non-cardiac populations and the emergence of brown adipocyte (Cidea-high) identity at Day 7 and Day 14. Multi ID analysis reveals that mixed-identity cells transition from predominantly non-cardiac combinations at Day 0 toward an atrial - left ventricle cardiomyocyte pairing by Day 14, consistent with cells traversing between cardiac subtypes. A secondary mixed-identity population co-expressing cardiac and brown adipocyte programs is present at Day 14 and may represent an off-target reprogramming outcome warranting further investigation. Differential gene expression and gene set enrichment analyses reveal key genes and gene programs that drive each major cell type, providing actionable insights for downstream experimental work. 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.

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, DEG results, and per-cell classification results.
- Investigate the brown adipocyte (Cidea-high) population further to determine whether it represents a stable off-target fate or a transient intermediate state and consider protocol modifications to reduce its emergence.
- Consider re-running CapyBio's Capybara Advanced Analysis with an adult cardiac reference dataset to assess whether reprogrammed cells more closely resemble mature cardiomyocyte identity, which may not be fully captured by the neonatal reference used here.
- Consider running CapyBio's CellOracle Differentiation Optimization Analysis to identify transcription factor perturbations that could improve reprogramming efficiency, reduce off-target populations, and potentially specify cells toward either an atrial or ventricular fate

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.

Differential gene expression analysis. Differential gene expression analysis was performed comparing each cell type against all remaining cells. Analysis was restricted to cell types with more than 10 cells and genes expressed in at least 10% of all cells. Genes were ranked using a Wilcoxon rank-sum test as implemented in scanpy.

Gene set enrichment analysis. Gene set enrichment analysis was performed using the MSigDB Hallmark gene sets for mouse (v2026.1). For each cell type, genes were ranked by Wilcoxon test statistic and restricted to the top 3,000 highly variable genes prior to testing. Enrichment was assessed against 50 curated hallmark pathways using 10,000 permutations as implemented in decoupler.

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