Task-1

The IGV coverage plot illustrates the mapped reads from a 10x chromium run for two genes, namely PF37_0914400-1 and PF37_0914500-1. In PF37_0914400-1, there is a noticeable increase in depth at the 5’ end, suggesting a more intricate genomic structure with numerous intronic regions. This suggests a higher level of transcriptional activity at the 5’ end due to the presence of transcription factor binding sites and regulatory elements. The heightened depth indicates an effort to capture the entire transcript length effectively.

Conversely, PF37_0914500-1 displays consistent coverage but with a lower sequencing depth compared to PF37_0914400-1. It contains only one intronic region, implying involvement in a simpler regulatory mechanism. The presence of a single intron reduces opportunities for alternative splicing, resulting in fewer isoforms and consequently lower sequencing reads and depths.

The observations from this visualization align with the anticipated genomic structures and regulatory mechanisms for these genes. The majority of exonic regions are well-covered, with some mismatches highlighted in colored bars, while noise levels are minimal. Overall, this run provides valuable insights into the transcriptional landscape and regulatory complexities of the examined genes.

Task-2

The single-cell run presents concerns regarding its metrics. The higher estimated number of reads suggests potential inaccuracies in cell count estimation or issues during Gel Bead-in-Emulsions (GEMs) generation, possibly affecting data fidelity. Low mean and median reads per cell indicate shallow sequencing depth, compromising data quality and completeness. Furthermore, a lower-than-expected median genes per cell indicates either low transcriptional diversity or sequencing/library complexity issues, potentially impacting downstream analyses. Optimizing cell capture efficiency and increasing sequencing depth are essential to improve the run's quality and ensure comprehensive transcript coverage. Despite the absence of a steep drop-off, the presence of a knee and cliff in the Barcode Rank plot implies effective separation between cell-associated and empty GEM barcodes, a positive indicator for quality control. However, rigorous evaluation and optimization are necessary to mitigate biases and ensure robust downstream analyses, emphasizing the need for thorough validation of cell identities and biological relevance.

Task-3

a)

The choice of dimension reduction using 20 principal components is made to capture a significant portion of the variance in the data while reducing computational complexity. The decision to use 10 nearest neighbours for constructing the neighbourhood graph aims to ensure local structure preservation and connectivity within the data.

Regarding cluster resolution parameters, setting it to 1 in Leiden clustering implies a moderate level of granularity in defining clusters. This value balances the desire for fine-grained clustering with the risk of over-segmentation, ensuring that clusters are meaningful and interpretable without being overly fragmented.

The rationale behind these choices suggests a balance between capturing sufficient variation in the data with dimension reduction and ensuring meaningful cluster delineation without excessive fragmentation. This approach aims to strike a balance between complexity, interpretability, and computational efficiency.

The absence of a distinct elbow in the plot indicates that there is no clear point where adding more principal components or nearest neighbours leads to diminishing returns in terms of explained variance or preservation of local structure.

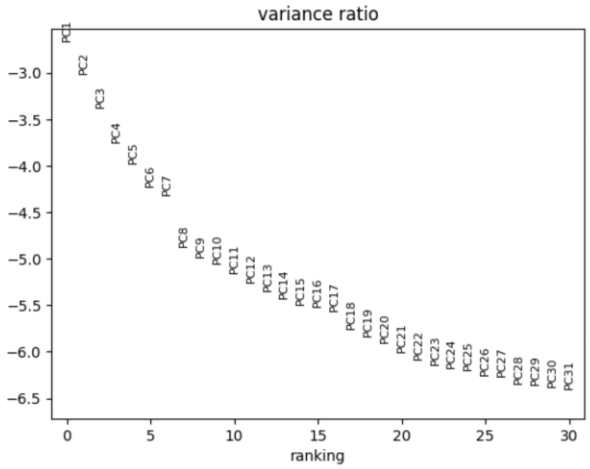

Figure 1 ElbowPlot

b)

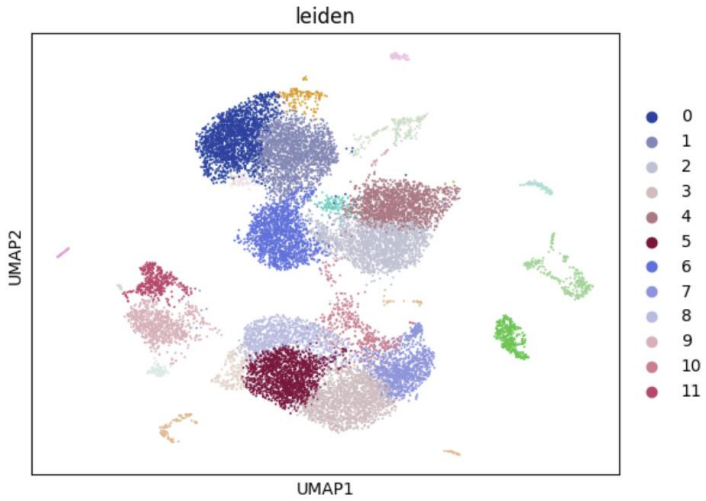

Figure 2 UMAP (Unintegrated)

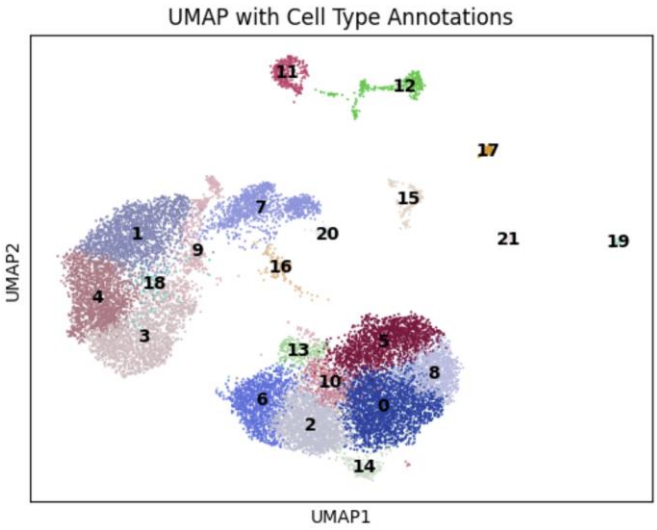

Figure 3

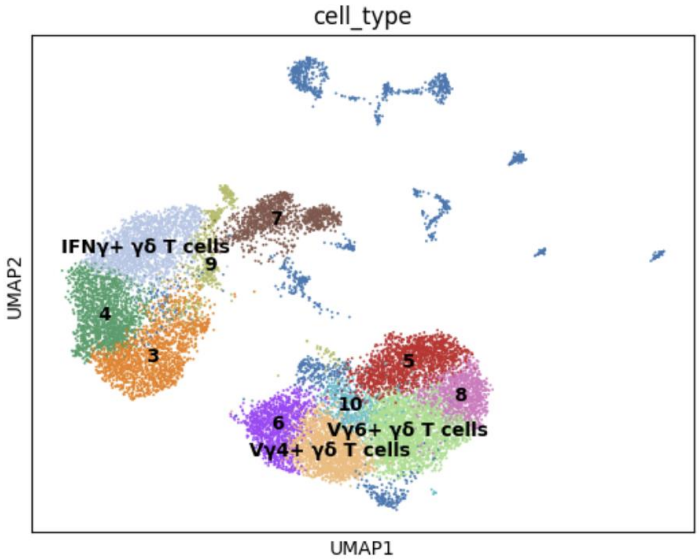

Figure 4 Annotated UMAP

c)

The frequency plot displays that in the knockout samples, Vγ6+ γδ T cells are notably abundant, constituting the largest cluster (cluster 0) which is also the cluster that corresponds to the knock-out cells. The deletion of β2 integrin gene leads to a localized increase in IL-17–producing Vγ6+ γδ T cells, particularly noticeable in the lungs, uterus, and spleen. Furthermore, it was observed that cluster 2 exhibits an abundance of markers associated with the Vγ4+ subset, including 5830411N06Rik. Additionally, the role of β2 integrins in promoting the thymic development of IFNγ-producing CD27+ Vγ4+ γδ T cells was elucidated and it is evident from the frequency plot that the markers for clusters 1 and 2 are downregulated in the knock-out cells.

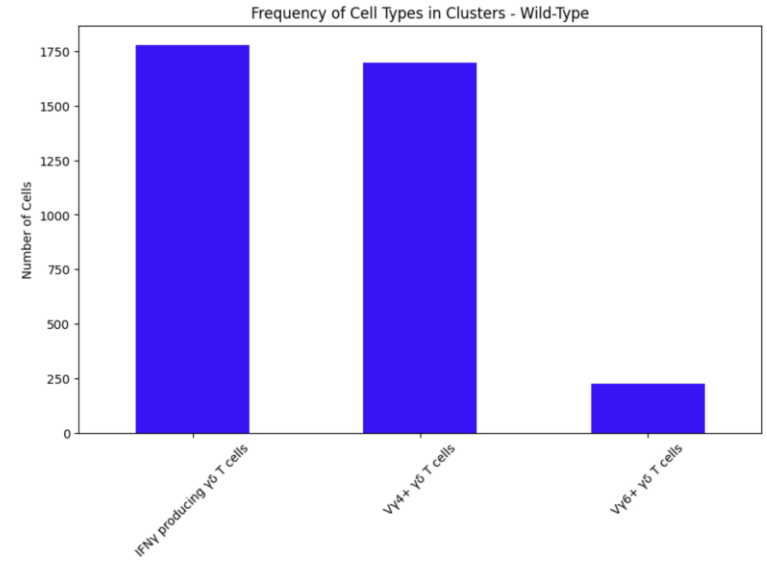

Figure 5 Frequency of cells in the largest 3 clusters in wildtype samples

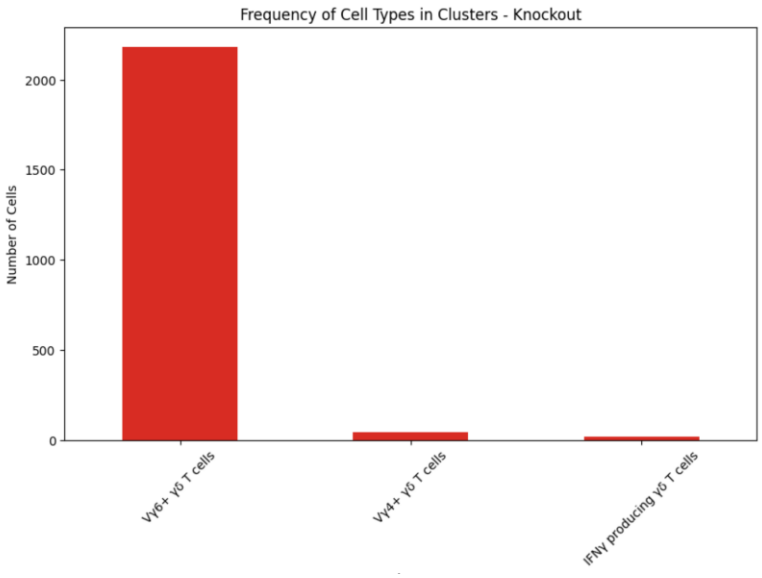

Figure 6 Frequency of cells in the largest 3 clusters in knockout samples

d)

Number of genes up-regulated in the KO with corrected p-values < 0.01 and logFC > 0.5 (using bbkn integration): 275

e)

Number of genes up-regulated in the KO with corrected p-values < 0.01 and logFC > 0.5 (using Harmony integration): 1473

f)

Since harmony explicitly models and aligns the global structure of datasets, which can lead to more comprehensive integration across samples, which was observed in the generated UMAPS for the merged samples (Figure 5, 6). By capturing the underlying biological variability, Harmony may provide a more holistic view of the data. Comparing the methods, Harmony identified a larger number

of up-regulated genes compared to BBKNN. However, it's essential to consider the trade-offs between sensitivity and specificity. Harmony might be more sensitive in detecting subtle changes but may also lead to increased false positives. BBKNN, with its lower number of up-regulated genes, might offer higher specificity but could potentially miss some true positives. Ultimately, the choice of method should consider the specific characteristics of the dataset, the desired balance between sensitivity and specificity, and the importance of controlling false discoveries.

The lower number of up-regulated genes identified in the pseudobulk (pb) samples compared to the single-cell (sc) analysis could be due to several factors. In pseudobulk analysis, the expression values of individual cells within each cluster are averaged, which can lead to a loss of information, especially for genes with heterogeneous expression patterns across cells within a cluster. This averaging effect may reduce the observed fold changes and significance of differential expression. Pseudobulk analysis collapses the cellular heterogeneity within clusters into a single representative profile, potentially obscuring subtle changes in gene expression present in only a subset of cells. This loss of cell-to-cell variability can result in fewer genes being identified as differentially expressed. The pseudobulk analysis on the bbknn integrated data only identified 2 genes while for harmony integrated data it only identified 24 genes.

Thus, for this dataset, harmony proved to be the best method in all aspects including batch effect removal by integration and identifying the upregulated genes correctly.

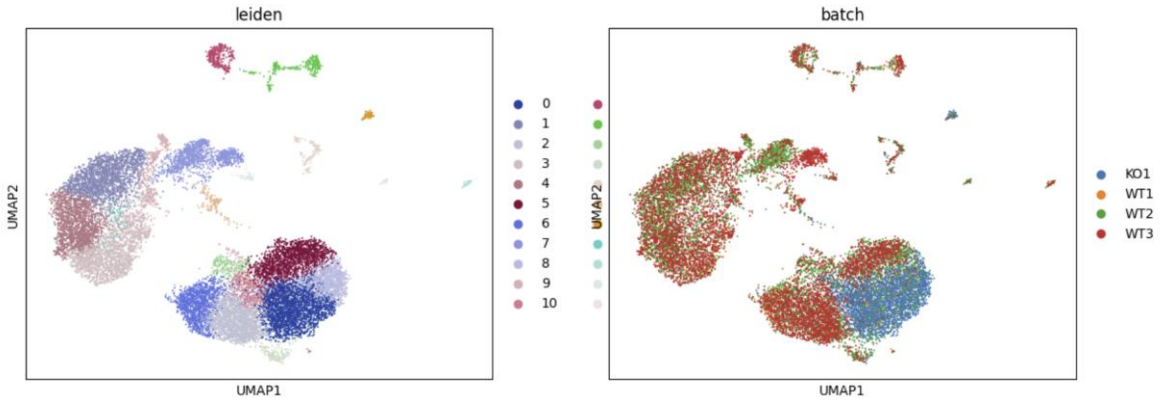

Figure 7 UMAP after BBKNN integration

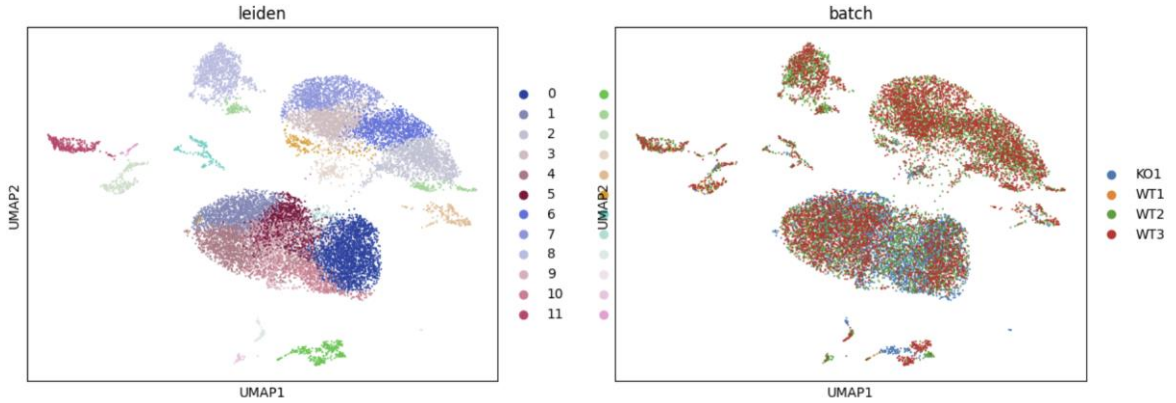

Figure 8 UMAP after harmony integration

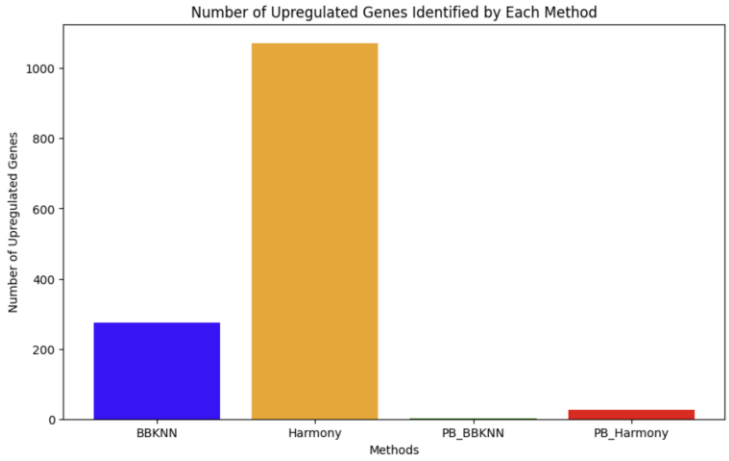

Figure 9 Summary plot for the number of upregulated genes after differential analysis via different methods

	Gene	scores	pval_adj	lfc	cluster
0	Klrk1	13.996684	0.0	3.077076	KO1
1	Ddx5	12.683079	0.0	2.119973	KO1
2	Rpl10	10.747263	0.0	1.456836	KO1
3	mt-Nd4	10.701342	0.0	1.242422	KO1
4	Hnrnpk	10.67269	0.0	2.588415	KO1

Figure 10 Top 5 upregulated genes in KO after bbknn integration

	Gene	scores	pval_adj	lfc	cluster
0	Klrk1	29.995029	0.0	2.727787	KO1
1	AC163354.1	29.754721	0.0	3.184084	KO1
2	Crip1	29.648893	0.0	1.736999	KO1
3	Itgb1	24.578897	0.0	3.420709	KO1
4	Cd47	24.577991	0.0	1.767219	KO1

Figure 11 Top 5 Upregulated genes in KO after harmony integration

Task-4

The findings from the trajectory analysis using the PAGA package provide valuable insights into the composition and relationships among different subsets of gamma delta (γδ) T cells in the studied dataset, particularly in the context of beta 2 integrin deficiency.

Cluster 0, characterized by the upregulation of Cd163l1, is identified as a subset of γδ T cells enriched in Vg6+ cells. The presence of these cells suggests a specific response associated with beta 2 integrin deficiency, as their accumulation is known to occur in this condition. (Figure 12)

Cluster 3, marked by CD27 expression, also appears to be influenced by beta 2 integrin deficiency, as CD27+ γδ T cells are upregulated in such conditions. The strong connection between clusters 0 and 3

further supports the idea of a shared regulatory mechanism or functional relationship between these subsets in the context of beta 2 integrin deficiency. (Figure 9)

Cluster 4 stands out for its enrichment in ribosomal gene markers, indicating a subset of γδ T cells with enhanced protein synthesis capabilities. This suggests a potential role for these cells in rapidly responding to stimuli or in sustaining prolonged immune responses.

Cluster 5, characterized by Ramp1 and Capg gene expression, is associated with Vg6+ cells, further supporting the notion of a specific response linked to this γδ T cell subset.

Conversely, clusters 1, 6, 8, 7, and 2 are identified as subsets enriched in Vg4+ genes, indicating distinct populations of γδ T cells compared to those in clusters 0, 3, 4, and 5.

Overall, these findings suggest that beta 2 integrin deficiency influences the composition and dynamics of γδ T cell subsets, particularly those enriched in Vg6+ cells, and that these subsets may play specific roles in the immune response associated with this deficiency. Additionally, the identification of clusters with distinct gene expression profiles underscores the heterogeneity and functional diversity within the γδ T cell population. (Sarah C et al, 2021)

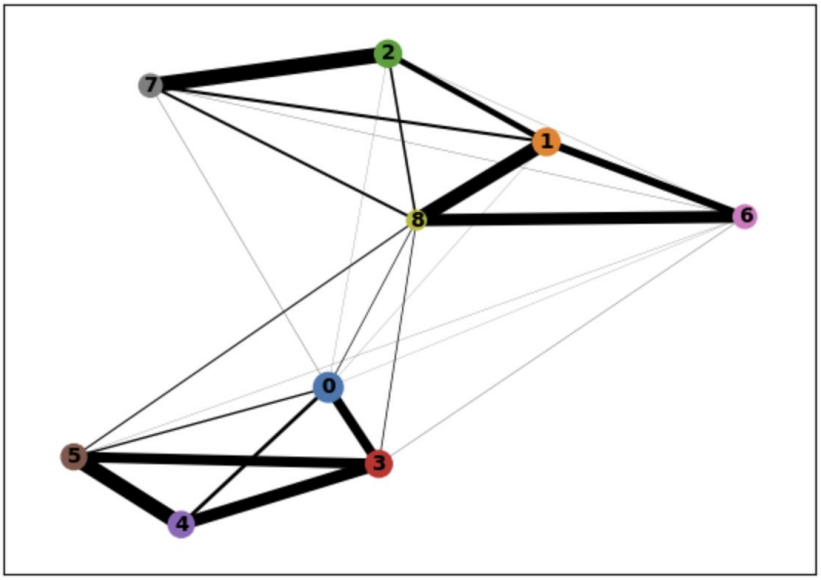

Figure 12 Trajectory plot for the main clusters

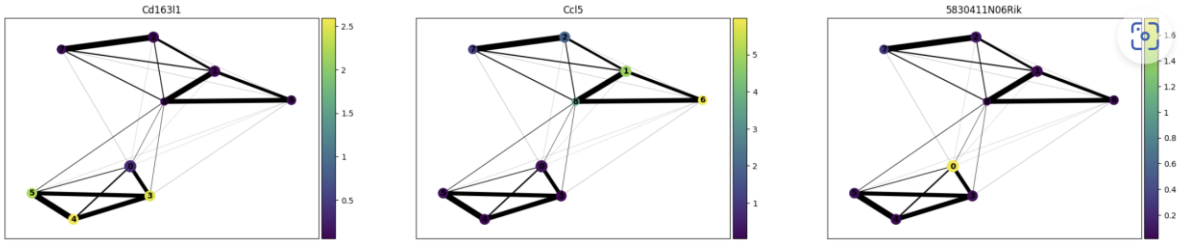

Figure 13

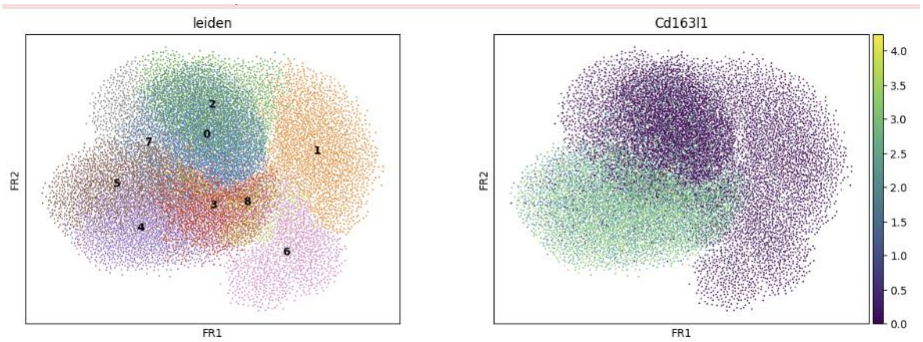

Figure 14

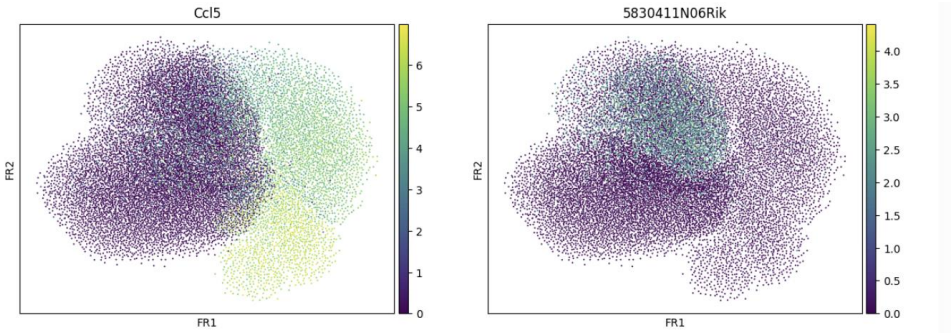

Figure 15

Task-5

The γδ T cell experiment demonstrates a methodical approach to data processing and analysis, covering essential steps from initial preprocessing to final clustering. However, to enhance the experiment's robustness and impact, several improvements can be implemented. Firstly, refining the experimental design by incorporating biological replicates and providing detailed metadata on sample conditions would increase the reliability and interpretability of the results. Secondly, employing more advanced statistical methods, such as generalized linear models and pathway enrichment analyses, could uncover subtle yet biologically significant gene expression changes. Additionally, validation through experimental assays and comparison with independent datasets would strengthen the validity and generalizability of the findings. Comprehensive annotation of cell clusters and integration of complementary data modalities would further enhance the interpretability and depth of biological insights.

Finally, adopting transparent and reproducible practices, including sharing raw data and analysis code openly, would promote scientific rigor and facilitate validation and extension of the findings by the broader research community. Overall, by addressing these technical aspects, the experiment can significantly advance our understanding of γδ T cell biology and its implications in health and disease.

The exploration of γδ T cells holds immense promise in the realm of cancer research, with potential implications for both understanding tumor immunology and developing novel therapeutic strategies. Understanding the intricate interactions between γδ T cells and tumors can provide valuable insights into the complex dynamics of the tumor microenvironment and immune evasion mechanisms employed by cancer cells.

γδ T cells possess unique features that make them particularly intriguing for cancer immunotherapy. They exhibit potent cytotoxicity against tumor cells, produce various cytokines that modulate the immune response, and demonstrate a certain degree of antigen recognition independent of major histocompatibility complex (MHC) molecules. Moreover, γδ T cells are found in various tissues, including epithelial tissues where many cancers arise, suggesting their involvement in tumor surveillance and immune responses at barrier sites.

REFERENCES

1. *Sarah C. Edwards, Ann Hedley, Wilma H. M. Hoevenaar, Teresa Glauner, Robert Wiesheu, Anna Kilbey, Robin Shaw, Katerina Boufea, Nizar Batada, Karen Blyth, Crispin Miller, Kristina Kirschner, Seth B. Coffelt doi: https://doi.org/10.1101/2021.07.04.451035*
2. *McIntyre CL, Monin L, Rop JC, Otto TD, Goodyear CS, Hayday AC, Morrison VL. β2 Integrins differentially regulate γδ T cell subset thymic development and peripheral maintenance. Proc Natl Acad Sci U S A. 2020 Sep 8;117(36):22367-22377. doi: 10.1073/pnas.1921930117. Epub 2020 Aug 26. PMID: 32848068; PMCID: PMC7486781.*