Overview

Strand NGS, our flagship next generation sequencing analysis and data management tool has been providing comprehensive workflows to investigate genomic variants, expression, regulation, and methylation studies. This tool enables biologists to perform end-to-end analyses of all their sequencing data while enhancing their research experience by application of the best computational and visualization principles.

As a part of this continued support for cutting-edge technologies in sequencing domain, Strand NGS now supports the complete analysis of expression data derived from single cell RNA sequencing (scRNA-Seq) methods. The introduction of Unique Molecular Identifiers (UMIs) handling in Strand NGS v3.1, followed by the addition of many pragmatic plot types in Strand NGS v3.2 ensures full-length support for the transcriptomic analysis at the cellular level.

This Application Note highlights the value of these new features in combination with the existing options for alignment, quality control, filter, analysis, and biological contextualization towards an integral single-cell study. By discussing a case study that addresses a complex biological question, we illustrate the technical fidelity of Strand NGS in dealing with nuanced and voluminous experimental data.

“"The pancreas is by far the most complex organ in the body."" - Patrick Soon-Shiong, Co-discoverer of Abraxane pancreatic islet cell transplant techniques.

Introduction

The generation, expansion and maturation of glucagon-secreting \( \alpha \) cells and insulin-secreting \( \beta \) cells spanning prenatal and postnatal stages\(^1\) is described as one of the most complex developmental programs in human body. Using several model organisms, researchers have been trying to gain insights into these cellular mechanisms in order to develop better therapy options for patients suffering from pancreatic disorders – especially, diabetes and cancer.
Sample Dataset

With an aim of defining the roadmap for pancreatic α and β cells, Qiu et al. performed a single cell transcriptomic study on mouse cells. Using Smart-Seq2 protocol, they profiled E17.5, P0, P3, P9, P15, P18 and P60 of β- and α- cells (except P3), as well as endocrine progenitor cells at P0, which were fluorescence-activated cell sorting (FACS) sorted from Insulin-RFP, Gcg-Cre; Rosa-RFP or Ngn3-GFP mouse strains, respectively.

This dataset is publicly available at GEO: GSE87375.

Data analysis methodology

The raw data (.fastq) corresponding to 917 cells captured using a paired-end library layout was extracted using SRA toolkit. The sample files were imported into Strand NGS v3.2 and reads were aligned against known genic regions of mm10 - RefSeq Genes and Transcripts model. A total of 1,604,744,054 reads were mapped.

Aligned samples were used for an analysis experiment creation. After post-alignment quality checks and gene body coverage assessment, all samples were taken forward for the expression profiling analysis. The data nodes depicted in experiment navigator shown in Figure 1, were obtained as follows:

- A list of cell type expression markers for both α and β cells were imported using Gene Symbol match. The entities were inspected for expression profiles and found to be in conformity.

- Proliferation markers were searched using Gene Symbol, and gene expression distribution plots for β cell development were obtained.

- A custom R-script was used to label cells (add a parameter in experimental grouping) based on the levels of proliferation markers; this new parameter was used in further interpretation creation to restrict analyses to mature and quiescent cells at P60.

- ‘Filter genes by expression’ was performed on All Entities to retain only those entities that show >0.01 normalized expression in at least all α or β cells that were quiescent at P60.

- The genes expressed above the threshold were then checked for any confounding patterns using ‘Confounding Variable Analysis (CVA).’ The corrected matrix and list of genes showing a differential expression between α and β cells at P60 were saved.

- Smooth scatter plot was used to understand the differential expression patterns.

- A t-distributed stochastic neighbor embedding (t-SNE) plot was created using the differential genes at P60 to see if they trace the varied developmental stages of both cell types.

- Natural Language Processing (NLP) based mining was used to extract relations across genes known to be involved in human pancreatic developmental stages; networks created based on the relations extracted from text and the mouse data was overlaid after Homologene-based mapping.

Figure 2: Gene Body Coverage plot showcasing uniform capture of all β cells.
Results and Discussion

Expression markers that distinguish cell types

The simplest and quickest method to validate sequencing quality is by establishing cellular identity using known marker genes. As stated by Qiu et al., the α and β cell types were checked for separately. Figure 3 shows

- **Mafb**, the immature β cell marker showing very high expression at E17.5 and a gradual decrease thereafter.
- **Ucn3**, the mature β cell marker showing a striking increase in expression from P3.
- **Pou3f4**, the immature α cell marker showing steady decrease in expression as maturation progresses.
- **Glrb**, the functional α cell marker showing a broad pattern of increase in expression as maturation progresses.

At each developmental stage of both cell types, the mean expression values did exhibit a high standard deviation indicating cellular heterogeneity.

Figure 3: Expression profiles of cell state markers. The standard deviation at each developmental stage is depicted by the brown vertical bars. The top row shows Mafb and Ucn3, β cell markers while the bottom row shows Pou3f4 and Glrb, α cell markers.

(Please note that P3 data was unavailable for α cells.)

Proliferation markers used to establish cell stage

As established by many developmental studies and emphasized by Qiu et al., pancreatic cell maturation is a coordination of maturation and proliferation. As a result, expression patterns of a given cell type at any stage are highly heterogeneous. Klochendler et al., have remarkably noted that replicating cells revert to a less mature state as they are repressed for genes that are responsible for designated physiological functions. This makes it essential for us to distinguish quiescent cells from those that are proliferating at each developmental stage. In order to see if indeed the data set shows such a myriad of cell cycle stages, we used the gene expression distribution plots as shown in Figure 4.

Figure 4: Gene expression distribution plots at various developmental stages of β cells for the three genes – (a) DNA methyltransferase, Dnmt1 (b) Histone methyltransferase, Ezh2 and (c) Histone acetyltransferase, Hat1. The colors of the rainbow depict from embryonic stage (red) to P60 (violet). Note that all three show a bi-modal distribution indicating a mix of proliferative and quiescent cells.
**Cell stage identification**

To ensure that the study clearly identifies the developmental roadmap, we decided to restrict our further analyses to quiescent cells. As a high expression of *Dnmt1*, *Ezh2*, and *Hat1* is a mark of proliferative stage, we labelled all cells with >100 reads mapped to any of the three genes as ‘proliferative’ and the others as ‘quiescent’ using a custom script. This feature allows one to define cut-offs and facilitates the use of marker gene expression values in defining cellular identities. The flexible approach further makes trial and error feasible in exploratory studies dealing with uncharted and complex tissues.

We then included ‘Cell stage’ parameter in our new interpretation along with ‘Source’ (α and β) and ‘Developmental Stage.’

**Correcting for confounding effects**

In order to find genes that are differentially expressed between α and β cells at P60, we used the f-test that is a part of the CVA feature in Strand NGS. Given that the cells were collected by two different collection methods and from various strains, we wanted to ensure that they did not bias the expression patterns. The CVA facilitates the correction of expression matrix when affected by confounders and then performs a statistical test on the corrected matrix to give a list of entities that are truly differential due to biological parameters alone and not due to any technical factors that crept in during experiment set-up.

![Gene expression patterns](image)

*Figure 5: Gene expression patterns.*
Differentially expressed genes

A total of 1007 genes were found to be significantly differentially expressed between quiescent α, and β cells at P60. Figure 6 shows the expression of these genes as a smooth scatter, the central green line represents the no change line, while the flanking lines on both sides indicate a fold change of 5. Most of the genes are slightly down-regulated in β cells as compared to α cells; as indicated by the density. The expression levels of 3 genes – Ins2, Ins1, and lapp stand out (highlighted in red) in β cells.

Islet amyloid polypeptide (lapp) encodes a member of the calcitonin family of peptide hormones. This hormone is released from pancreatic beta cells following food intake to regulate blood glucose levels and acts as a satiation signal. Ins1 encodes insulin, a peptide hormone that plays a vital role in the regulation of carbohydrate and lipid metabolism. lapp and Ins1 are co-secreted by β cells. Figure 7 shows that their levels are higher in each of the β cells as compared to the α cells.

Figure 7: A covariate regression plot depicting that β cells (shown in blue) always show a higher level of lapp and Ins1 as compared to α cells (shown in red).

We then applied t-SNE on the list of 1007 genes as the dimensionality reduction technique to find the local and global relationships between α and β cells across all developmental stages.

Figure 8: A 3D t-SNE plot showing the clear segregation of the developmental paths of α and β cells. Colors of the rainbow are used to depict the various developmental stages – red (embryonic) to violet (mature).

(Please note that P3 data was unavailable for α cells.)
Pathway analysis

Pagliuca et al summarized the steps involved in glucose-stimulated insulin secretion (GSIS) pathway in humans. We used ‘Extract relation via NLP’ to mine this text and then construct a network that depicts the steps.

The expression data from the current experiment was then overlaid on the NLP-derived pathway. Mouse data reaffirmed the observations made in human subjects indicating the value of mouse as a model organism for all pancreatic development related studies. The quilts next to matched entities in Figure 9 show expression levels across three (rows) cell types and seven (columns) developmental stages.

System specifications

- Alignment and Quantification were performed on a machine with 32 core processor, 64 GB RAM and 12 TB of storage space.
- Analysis was performed on a laptop with 4 core processor, 8 GB RAM and 350 GB of storage space.

Conclusions

This Application Note describes the suitability of Strand NGS for single cell transcriptomic studies by

- Providing an end-to-end workflow for alignment and analysis
- Offering a multitude of algorithmic and visualization features that facilitate exploratory analysis
- Making large-scale data management an inherent characteristic, thereby ensuring hassle-free analysis experience for the researcher.

References