Overview: These are the things we have/need to have that are in a chronological order. They are all focused around answering the question: how comparable is the data from thawed samples when compared to fresh.

For the purpose of avoiding Pierre’s FA2 v FA3 paper, this paper will treat FA2 and FA3 as 2 hiPSC cell lines (not focusing on the isogenic part or the good vs bad endoderm stuff). Our biological question is differentiation from D0 to D4.

What we have Bioinformatic need to do Lit search need to do Other need to do

1. **Cell culture**

-Optimisation:

-DMSO vs Cryostore (non-fixation storage)

-Accutase vs Trypsin

-Have datapoints D0, D1, D4 from Countess- Can get more + flow data relatively easily

1. **Differentiation data**

-Viability/counts for differentiation- cells from 10X run

-Images from run

-If we need to prove differentiation/justify further- can get staining if needed. Have biomark data on differentiationg cells though this is Pierre’s unpublished data…

1. **Experimental design**

-Here we use schematics to show both TC experimental layout and Chip layout on the 10X.

-Do we talk about our Chip control? Can we explain it/normalise it bioinformatically or do we just ignore it?

-Can we at least justify it based on lit searches (e.g. there are length of freeze time conditions in this paper: <https://www.nature.com/articles/s41598-019-46932-z>

1. **10X Quality Control obtained during run**

-We have TapeStation and Qubit results

-Only relevant if I can see that large amounts of noise in the QC are related to noise in the sequencing data: NEED TO CHECK A v B (might actually be something here) and Fresh v Thaw

1. **Data QC before cleaning**

*It is in this section that we want to very comprehensively show how similar and dissimilar the raw data we generated is between fresh and frozen. The idea is that when someone gets back their reads after running their thawed samples, they have a rough idea how much noise and how many counts/genes/cells to expect in a thawed rather than fresh sample. This is before all this noise is removed for further analysis.*

**Raw data:**

- Alignment scores (FvT)

-No of reads (after sequencing) -> probably supp. But can link back to QC (part 4)

-Cell counts FvT (+ min, max average cell counts etc.)

-No of genes vs total counts graph that Simon has already generated- this shows sample quality and also noise- cells with lots of total reads but only few genes

-This is the graph where we are asking if the data we have generated is useful at a single cell level? i.e. is every single cell distributed nicely in terms of total UMI to unique UMI? Or are there single cells that are garbage? Is this the same between conditions?

-Simon, can we overlay these graphs with different colours between fresh and thawed for the same condition? I would like to see the size of the tails on the same axis fresh v thawed.

-Is there any other way to quantify noise at this stage of processing? At this point, we really want to show that the amount of “background” is/is not affected by thawing. Basically, is the tail due to freezing or is it more related to cell line/ day of differentiation etc.

**Correlation:**

-Gene level as log10(gene\_total\_counts+1): Simon already has these

-This is good to put here because it lets us compare conditions before we start playing around with thresholds and different packages + is directly comparable to what 10X publishes.

-Basically, here we are asking can we use the raw thawed data as a substitute for the raw fresh data.

-Simon, should we be generating these graphs with mean gene counts rather than total? My feeling is that the samples with higher number of cells will skew this data?

-Removing mitochondrial and ribosomal genes

-Does the literature show that a higher number of MT genes reflect worse samples quality? If so this can be a figure. If not, a footnote on our methodology.

1. **Setting threshold and looking at the data after processing with Seurat**

*Here we want to tell people what values we need to be setting to make our data usable. Basically, we have acknowledged any differences between the raw data in section 5. However, no one uses unfiltered data. Here, we want to explore what thresholds need to be set so the clean data between Fresh and Thawed is the “same”.*

-Justify why we are going to use Seurat + some sort of bioinformatic overview

-Thresholding experiments: Seurat allows us to pretty easily change min cells, min features. We want to know at what values the total/usable genes expressed are most correlated between fresh and thawed samples. This means we need to slide our min features at the same time as min cells and make a heat map that shows at what point there is best correlation between samples. To do so:

-Min cells: Naz is going from 0 to 10 (steps of 1)

-Min features: Naz is going 0 to 1000? Or 500? In steps of 10 or 5 (I feel like even 250 is going to be too high but shall see)

-maybe to ease people into this we first generate/present a frequency distribution graph of genes per cell (i.e. min features) when we pick a default min cell (5 or 3). We display it with fresh and thawed from the same condition overlayed so it is clear that the level at which “noise” is filtered out is different. Therefore, we present Naz’s heatmap monstrosity and choose the best thresholds.

-Once we have thresholds, we need to regenerate our No. of genes vs total counts (single cell) graph and our correlation graph. With these (hopefully) we can see that now we have similar distributions with no noise and our total gene expressions are well correlated

-Are we just going to rely on whatever normalisation Seurat does? Should we compare this to another method?

-The conclusion at this part is that we know what values need to be set for people who are using frozen data if they want it to be at the standard of fresh data.

-I want to know what that number would have been without the frozen data? Is it different?

-“That number” means what is the value for fresh on average where we lose our noise (i.e. first peak)- is it different?

-To do so, I think we can run the same analysis but find the correlation for FA2vFA3 Fresh and FA2vFA3 Thaw and see if the cutoff is lower or higher

1. **Biological Differences:** *this one is still a bit vague*

**-**We cluster with Seurat

-Are the clusters preserved between the conditions? Correlation graph for genes in cluster

-We find diff expressed genes overall and find the related pathways

-We check DNA damage pathways/genes + cell cycle

-We then need a way of assessing the effectiveness of our differentiation analysis in fresh and if it represented by frozen.

-Align to in vivo mouse single cell data set? Or make pseudotimeline? This is the bit where we ask how good the thawed data is at answering the D1 vs D4 biological question.

-Finally, if there is a super obvious difference in a particular pathway (say ox phos), we might have to freeze and thaw some more cells and do some verifying tests in the lab. This would depend on what pathways we find that are significant.