Cancer single-cell/bulk challenge: wrinkles

Refresher: When we refer to bulk gene expression profiling, we're discussing a process whereby RNA from throughout that tissue is profiled together. Bulk profiling can't easily tell us what's going on within each individual cell, but we do get to observe the extent to which each gene is expressed in that section of tissue. Single-cell profiling involves a step where bulk tissue is separated into individual cells, and then gene expression levels in those cells are measured separately for each cell. Using profiles from a small set of single-cell profiled tumors, we seek to computationally estimate the proportion of cell types in existing bulk datasets for which single-cell measurements are unavailable. This would be a straight forward problem of modeling linear mixtures were it not for at least two major wrinkles.

Wrinkle 1: We don't necessarily measure every cell type. Dissociation is a physical process. If certain cell types are preferentially lost during dissociation then our approach to estimate the mixture would lead to biased estimates (Figure 1). How would you determine if there are likely to be missing cell types? How would you estimate what the expression profiles of those cell types would be?

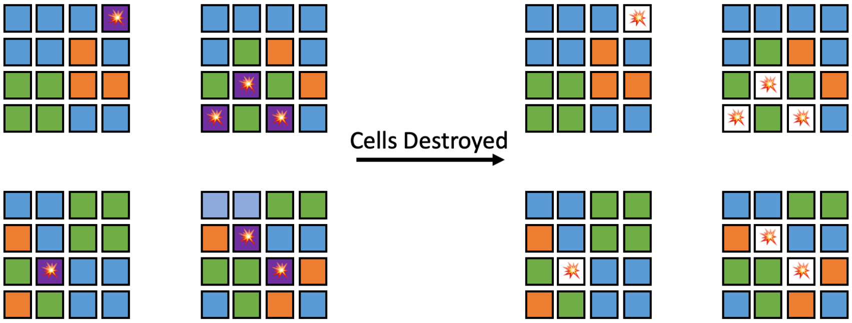


Figure 1: Cells may not be comprehensively measured and those of certain types may be more likely to be unobserved than others due to physical destruction during the profiling process. In this figure we have tumors comprised of four cell types (blue, green, orange, and purple). The purple type is destroyed during profiling, leading to influences in bulk data from cell types that are not observed with single-cell profiling.

Wrinkle 2: Cancer cells may have different expression profiles and may differ from tumor to tumor. "An emerging theme from recent scRNA-seq studies is that malignant cells tend to cluster in their expression profiles primarily by patient sample, and non-malignant cells cluster in their expression profiles by cell type, somewhat independently of the patient of origin (Suva and Tirosh 2019 *Molecular Cell*). This indicates that (1) inter-tumor heterogeneity is typically larger for malignant cells than for any particular type of non-malignant cells. (2) For malignant cells, inter-tumor heterogeneity is much larger than intra-tumor heterogeneity. Put more simply: cancer cells from different people's tumors can be quite different in gene expression from each other (Figure 2). This means that the comparable expression profiles for cancer cells, which will often comprise a large fraction of the tumor, may often be unavailable in the single-cell profiles and may need to be inferred.

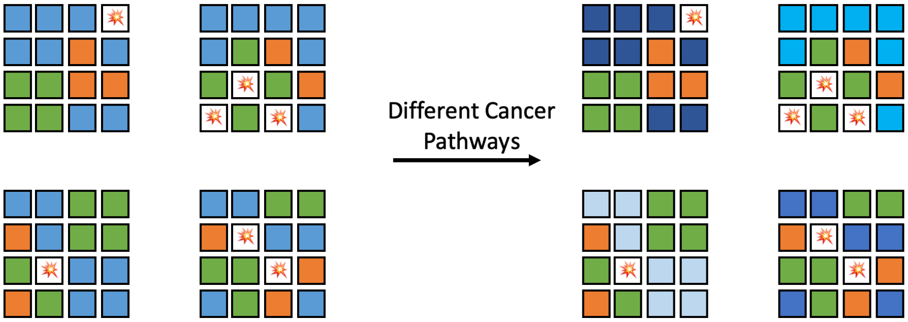


Figure 2: Cancer cells may exhibit expression profiles that are quite different between people. In this figure we have the same tumors from the previous figure, except we've now made the analysis more complex by denoting that cancer cells (hypothetically blue cell types) may in fact express different pathways from tumor to tumor (shown with different shades of blue).

Latent phenotypes challenge: wrinkle

We have been discussing the correlation structure of phenotypes in the UK biobank. However, phenotypes are the result of both environment and genetics, which can sometimes have opposing consequences.

Consider the relationship between smoking and body mass index (BMI). A recent study in the UK biobank found that being a current smoker is correlated with lower BMI, presumably because smoking is an appetite suppressant [1].

Subsequent UK biobank studies have found that there is a significant overlap between genetic variants that increase smoking risk and those that increase BMI [2]. This suggests that there might be a latent phenotype -- perhaps related to addiction? -- that acts to increase both BMI and smoking susceptibility, *despite* the fact that the phenotypes are negatively correlated. How could such a latent phenotype be learned from UK biobank data?

References

1. Dare S, Mackay DF, Pell JP. Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. PLoS One. 2015;10: e0123579.

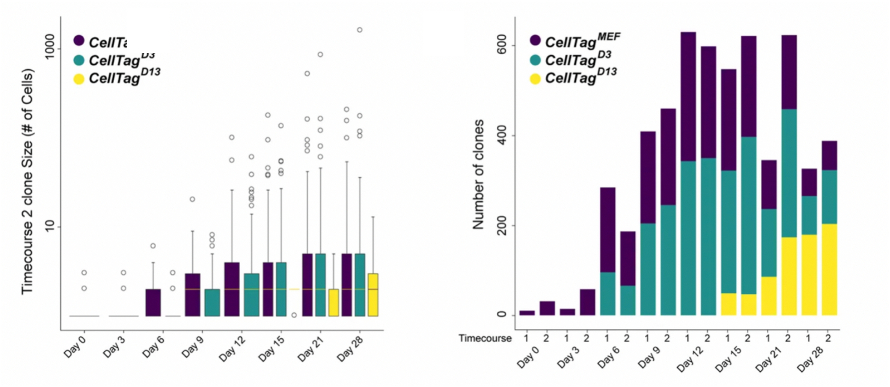
2. Wills AG, Hopfer C. Phenotypic and genetic relationship between BMI and cigarette smoking in a sample of UK adults. Addictive Behaviors. 2019. pp. 98–103. doi:10.1016/j.addbeh.2018.09.025

Differentiation trajectories challenge: wrinkles

(Data from Biddy et al., *Nature* 2018; for color versions of figures please see the online version of this handout)

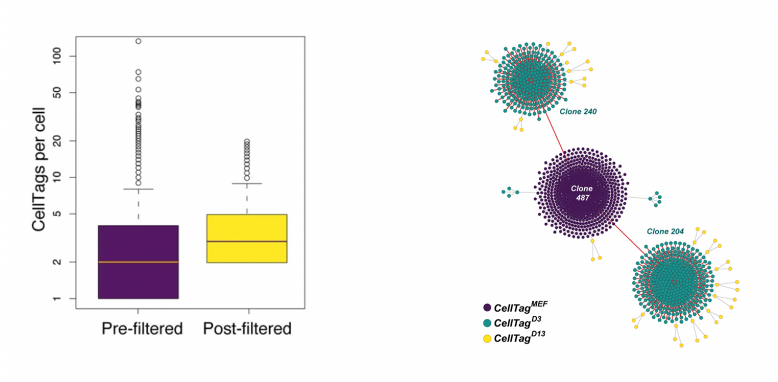
Wrinkle 1a: For our current analyses, we focus on the largest clones containing 35 or more cells. However, most of the data is comprised of minor clones and lineages (n < 4 cells per clone). How can we leverage these small clones?

Wrinkle 1b: The detection of clones relies on clonal expansion and capture. The probability that a particular clone is detected in the earliest stages of the timecourse is low. How can we trace lineages to their origins?

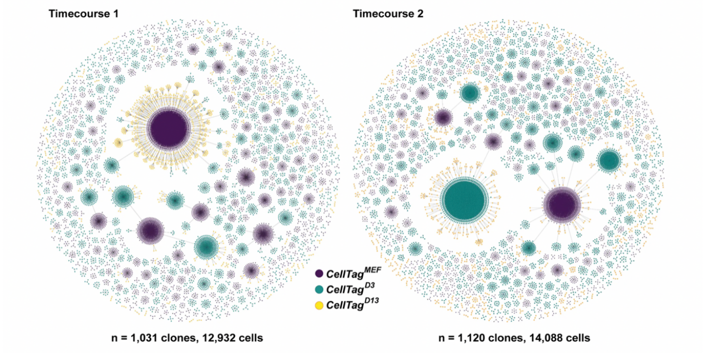


Wrinkle 2a: 99% of cells express CellTags, but around 30% of cells are not tracked as they express ≤ 1 CellTags (there is a high false-positive rate in clone-calling below this threshold). Can we leverage this 'pre-filtered' dataset (below, left) for analysis?

Wrinkle 2b: Not all lineages are complete, i.e., not every cell tagged in the first round will go on to pick up additional tags in subsequent rounds of labeling (below, right). Can we fill in these gaps?



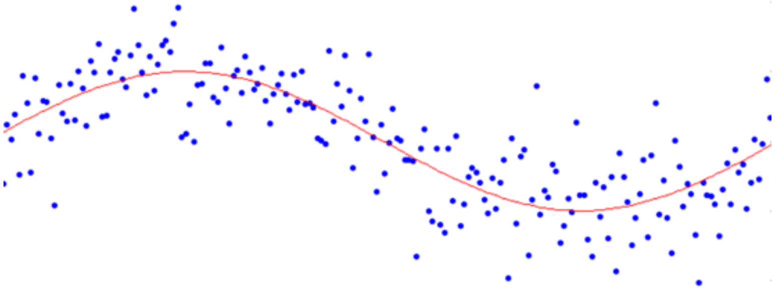
Wrinkle 3: From thousands of clones and lineages identified across a time course, which of these will be most informative to reconstruct differentiation trajectories?



Generative challenge: wrinkles

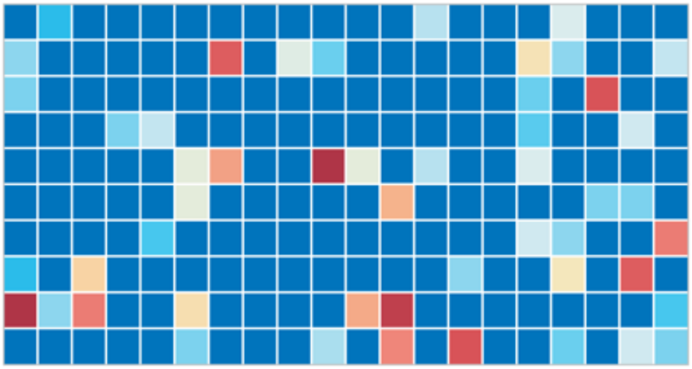
Wrinkle 1: Single cell RNA-sequencing data suffer from noise and dropout

Noise:



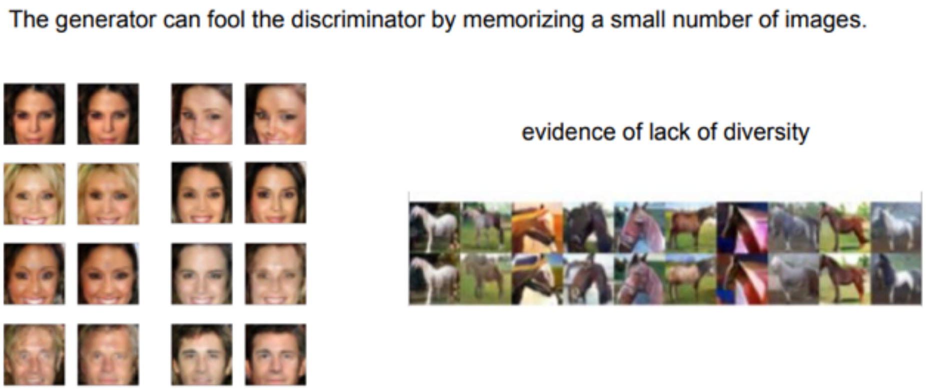
Most biological measurements have a lot of noise due to background processes like bursting and errors in measurement devices.

Dropout:



Undercounting of transcripts due to low capture efficiency. Only about 5% of the transcript is captured. Note this is different from having MISSING VALUES.

Wrinkle 2: GANs suffer from mode collapse



The generator can fool the discriminator by memorizing a small number of images.