Cancer challenge: example approach

I am not aware of a solution to this entire problem. There are a number of methods that begin to address various parts of the challenge.

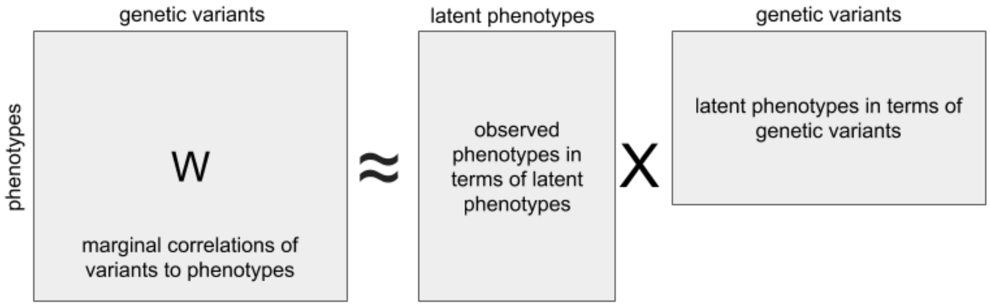
There are a number of solutions to model bulk profiles from single-cell ones. One that I think is generally well-written, includes helpful figures, and a sensible approach is "SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References" from Dong et al. (available at https://www.biorxiv.org/content/10.1101/743591v1.full)

There are also a number of solutions to identify transcriptomic signatures from bulk malignant tissues. It's important to note that only certain types of pathway changes are likely to be compatible with uncontrolled cell proliferation, so is probably not hopeless to attempt to solve this problem (as one might imagine it could be if the tumor-specific changes could be entirely random). To solve this, it will probably be necessary to model these from bulk data because, with rare pediatric cancers, we are unlikely to be able to single-cell profile enough of them to capture the full set of potential pathway states. "DeepProfile: Deep learning of cancer molecular profiles for precision medicine" by Dincer et al. (available at https://www.biorxiv.org/content/10.1101/278739v2) describes an approach to extract these states. Our own lab has examined methods that capture a low dimensional representation for their ability to capture pathway activity, evaluated using a mutation-effect prediction framing, in "Sequential compression of gene expression across dimensionalities and methods reveals no single best method or dimensionality" by Way et al. (available at https://www.biorxiv.org/content/10.1101/573782v2) and found concordant results.

The ideas in Dong et al. and Dincer et al. appear to be very nice points at which to start considering these complementary challenges.

Latent phenotypes challenge: example approaches

Two recent papers tackling this problem have both taken an approach based on matrix decomposition. Specifically, information about many genetic variants and many phenotypes can be summarized in a matrix whose i,j-th entry is the marginal correlation between phenotype i and genetic variant j. Factorizing this matrix amounts to representing each phenotype in a lower-rank space of latent phenotypes, which are themselves linear combinations of the genetic variants.


genetic variants
phenotypes
W
marginal correlations of variants to phenotypes
≈
latent phenotypes
observed phenotypes in terms of latent phenotypes
X
genetic variants
latent phenotypes in terms of genetic variants

The two approaches are as follows:

Tanigawa et. al. [1] applied singular value decomposition to a version of the above matrix W constructed from 2,138 broad phenotypes from UK biobank and 235,907 genetic variants. This approach cast a broad, agnostic net in terms of the phenotypes considered and employed the simplest possible matrix decomposition.

In contrast, Udler et. al. [2] had a specific interest in type II diabetes (T2D). To this end, they constructed the matrix W using only 94 genetic variants that were previously associated with T2D and 47 phenotypes with clear link to T2D (e.g. fasting insulin and glucose levels, BMI, waist circumference, etc.). They then processed this matrix to have non-negative entries and factorized it using Bayesian non-negative matrix factorization. They identified 5 latent phenotypes which they labeled as Beta cell, Proinsulin, Obesity, Lipodystrophy, and Liver/Lipid which they hypothesize may represent core genetic pathways underlying risk for T2D.

We leave you with two closing questions:

1. Both approaches evaluated the latent phenotypes that were discovered by checking which of the observed phenotypes they contributed to. Can you think of other ways to do this?
2. After applying these approaches, would the heritability of the “reconstructed” observed phenotypes (i.e., the appropriate linear combinations of the latent phenotypes) be greater or less than the heritability of the original observed phenotypes? What does this mean?

References

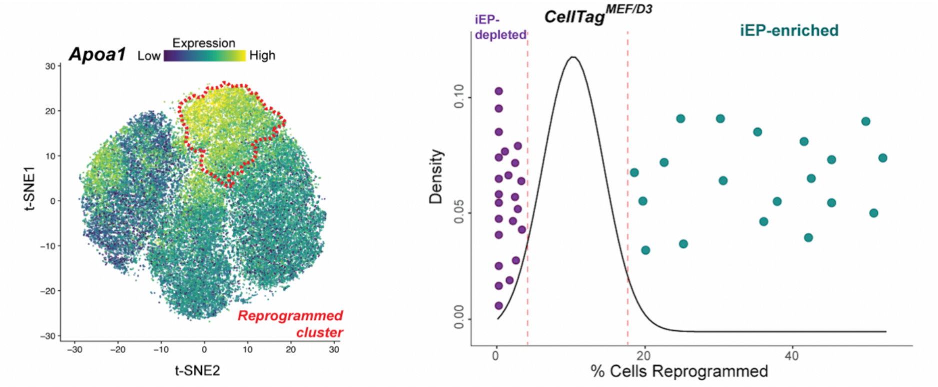
1. Tanigawa Y, Li J, Justesen JM, Horn H, Aguirre M, DeBoever C, et al. Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. Nature Communications. 2019. doi:10.1038/s41467-019-11953-9

2. Udler MS, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018;15: e1002654.

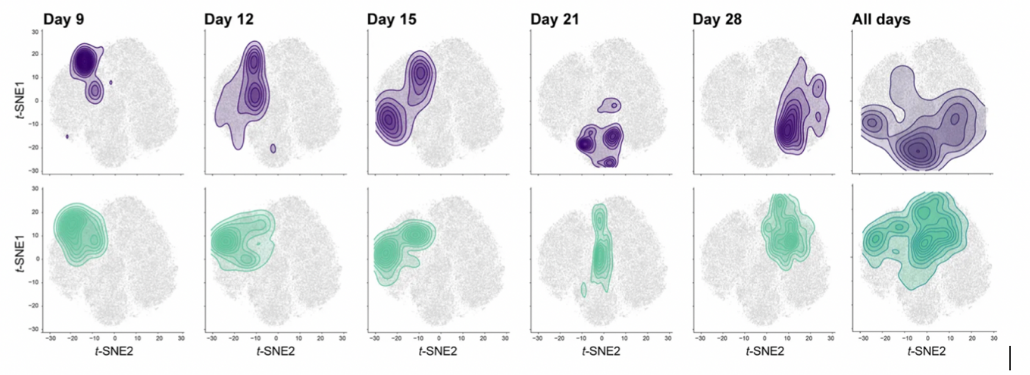
Differentiation trajectories challenge: example approach

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

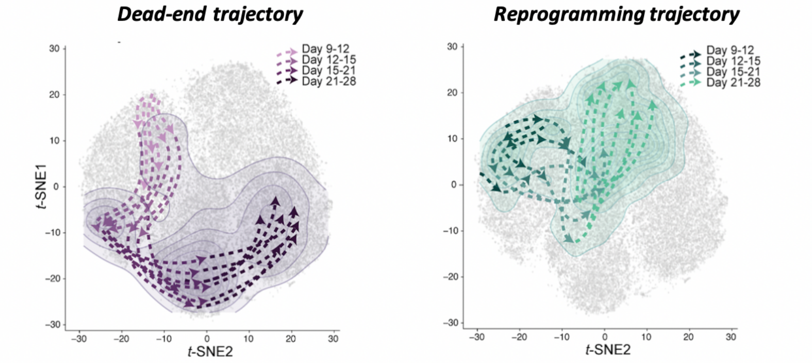
We present an approach to solving "Wrinkle 3: from thousands of clones and lineages identified across a time course, which of these will be most informative to reconstruct differentiation trajectories?"



Solution: We used a randomized test to identify clones and lineages enriched for, or depleted of reprogrammed cells (red outline on above *t*-SNE plot, left). Right: Density plot of the mean proportion of reprogrammed cells for groups of randomly selected cells (defined by reprogrammed cluster occupancy; *n* = 59 groups, 10,259 cells). Randomized testing of 59 CellTag clones (≥35 cells per clone, *n* = 10,259 cells) identifies iEP-enriched clones (*n* = 20 clones, 6,128 cells; *P* < 0.05) and iEP-depleted clones (*n* = 24 clones, 3,177 cells; *P* < 0.05)



We then project the cell distribution of these reprogrammed cell depleted clones (top row) and reprogrammed cell enriched clones (bottom row) onto the *t*-SNE plot, for each day of the reprogramming process. Using this contour plotting approach, we can identify the areas of highest clonal density for each group of clones, for each day of reprogramming.



We then simply 'join-the-dots' between these areas of highest clonal density across the reprogramming timecourse, revealing two distinct cell differentiation trajectories. Using differential expression analysis, we revealed significant differences in gene expression between these trajectories early in the reprogramming process.

Generative challenge: example approaches

Example 1: Generating drug perturbation response data from an autoencoder. This approach learns a vector in the latent space corresponding to perturbations, and then applies this vector before decoding on new baseline cell measurements.

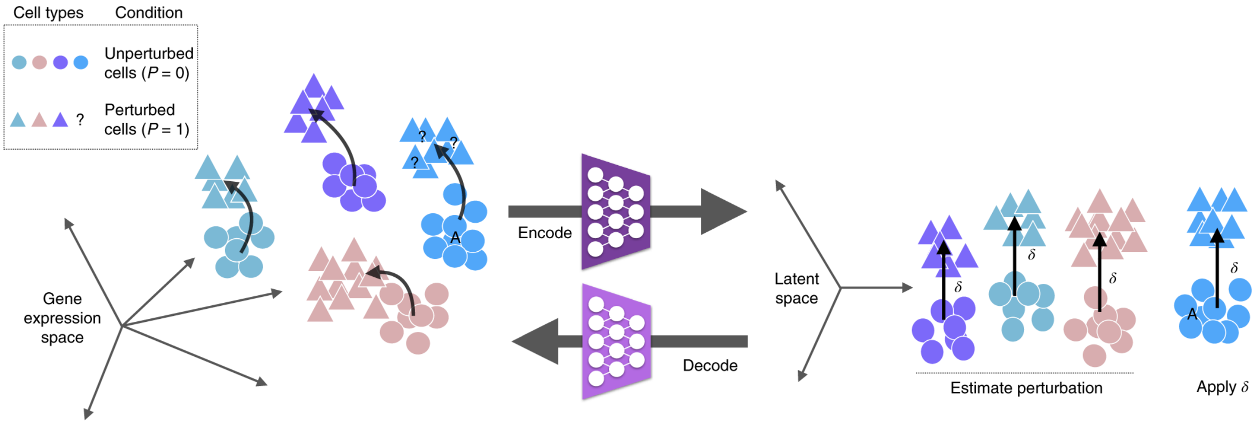


Figure from Loftallahi et al. Nature Methods 2019.

Example 2: Generates a different measurement modality from a baseline measurement. This could be two different batches of the same datatype (immune cells from two patients), or two different measurement modalities on the same system. This uses a cycle-GAN framework with an extra loss termed "correspondence loss" to make sure that the generated measurement type has features that match.

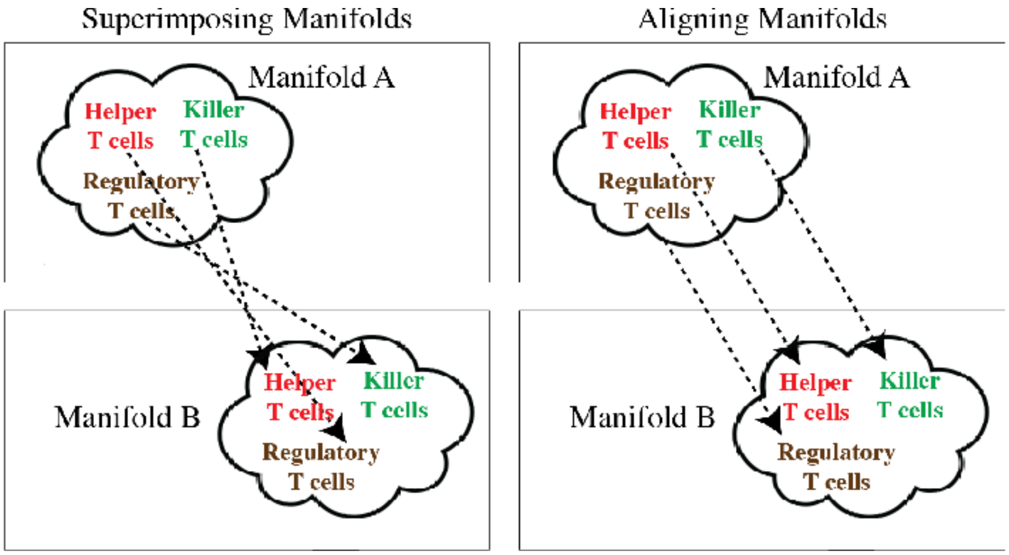


Figure from Amodio et al. ICML 2018.