Notebook

February 6, 2018

Contents

List of Figures

List of Tables

List of Codes

```
1 from ipypublish.scripts.ipynb_latex_setup import *
```

1 Identifying Repositioning Candidates for Antineoplastics

1.1 Abstract

The Cancer Genome Atlas (TCGA) has collected mutation and expression data for over 20,000 tumor samples, but most subtypes of cancer have few normal tissue samples to compare against. We uniformly computed expression data for both TCGA and The Genotype Tissue Expression Constortium (GTEx), which collected expression data from thousands of normal tissue samples, to create a large repository of cancer and normal expression data free of computational batch effects. Combined expression data was validated by identifying known cancer phenotypes for several antineoplastic drug targets and finding similar expression patterns in both TCGA and GTEx. Repositioning candidates were found by identifying cancer subtypes that share phenotypes with the positively validated targets.

```
# Set autoreload module for dev
%load_ext autoreload
%autoreload 2
%aimport rnaseq_lib
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

```
# Imports
import pandas as pd
import rnaseq_lib as r
import holoviews as hv
hv.extension('bokeh', logo=False)
```

```
<IPython.core.display.HTML object>
```

```
# Inputs
## Synapse ID: syn11515015
df_path = '/mnt/rnaseq-cancer/Objects/tcga-gtex-metadata-expression.
    ↪ tsv'
df = pd.read_csv(df_path, sep='\t', index_col=0, dtype=r.tissues.
    ↪ dtype)
```

```
# Holoviews object wrapper for dataframe
h = r.plot.Holoview(df)
```

1.2 Introduction

Previous large scale comparisons of gene expression in TCGA have been done with fewer than 5,000 samples — a majority (~90%) of the samples derived from primary tumor tissue and the rest representing 'normal' tissue, some of which is taken from the same patient carrying the tumor[? ?]. In one study, the dataset consisted of 4,043 tumor samples and 548 normal tissue samples across 21 TCGA cancer types — an average of only 26 normal samples for each cancer type compared to ~200 tumor samples[?]. While gene expression in normal tissue is more homogeneous than tumor samples, batch effects and contamination are a common problem with RNA-seq, which complicates an already noisy data source. Additionally, the small sample sizes cannot accurately reflect the general population, and therapeutics guided by results obtained from these analyses may have a higher likelihood of failing in clinical trials. Given the lack of normals in TCGA, the second largest dataset processed in the large-scale compute are non-cancerous samples collected from GTEx, which provides valuable insights into

the mechanisms of gene regulation by studying human gene expression and regulation in tissues from healthy individuals[?].

There are a myriad of antineoplastic drugs designed to fight different types of cancer, but most only target a small population within a single subtype of cancer, which gives most patient few options. By validating known expression biomarkers for antineoplastics, these same expression motifs can be used to identify candidate subtypes of cancer that may respond to treatment.

1.3 RNA-seq Datasets

The above figure, which depicts three different RNA-seq datasets containing ~20,000 samples, totals more than 110 terabytes of patient data — more data than can fit on most machines and far too much data to process efficiently on most hardware available to academics. Two of the three are cancer datasets, including The Cancer Genome Atlas (TCGA) which includes over 11,000 patients across 33 tumor types and represents the largest tumor collection of tumor data[?]. GTEx contains over 8,000 samples spanning almost every tissue in the human body, with the goal providing a homogenous set of expression data representing healthy tissue[?].

1.4 Large-scale RNA-seq Compute

In order to process this massive combined dataset in a timely and efficient manner, our lab developed *Toil*, a distributed workflow platform capable of massive scale[?]. I wrote a Toil-based RNA-seq workflow which provides results that are concordant to TCGA's RNA-seq workflow, but is an order of magnitude faster and provides quantification outputs from multiple tools **??**[?].

The workflow was run on all 20,000 samples with a throughput of 99.6% on an Amazon Web Services cluster that peaked at 32,000 cores and 60TB of memory and finished in just under 4 days with room for improvement.

1.5 GTEx as a Prior for TCGA Normals

While GTEx contains thousands of normal tissue samples, they can't be compared directly to TCGA due to differences in sequencing depth and laboratory batch effects. Unfortunately, there don't exist standard RNA-seq benchmark samples that every consortium uses to calibrate with before processing, which would likely introduce fewer batch effects that are easier to correct. Current available methods typically attempt naive distribution fitting that tend to work less effectively as the amount of samples and classes increases[? ?]. Instead, we can evaluate GTEx as a prior by normalizing for sequencing depth and dispersion using DESeq2[?], then comparing differential expression results between TCGA normals and GTEx normals.

```
sp_counts = h.sample_counts()
de_gtex = h.differential_expression_tissue_concordance(
    ↪ tissue_subset=tissues, tcga=False).relabel('GTEx')
de_tcga = h.differential_expression_tissue_concordance(
    ↪ tissue_subset=tissues, gtex=False).relabel('TCGA')
```

```
sp_counts
```

```
:Bars   [Tissue,Label]   (Count)
```

```
%%opts HeatMap [width=600]
(de_gtex + de_tcga).relabel('Differential Expressoin Gene
```

```
        ↪ Concordance')
```

```
:Layout
   .HeatMap.GTEx :HeatMap    [Tissue-Tumor/Normal,Tissue-Normal]    (
PearsonR)
   .HeatMap.TCGA :HeatMap    [Tissue-Tumor/Normal,Tissue-Normal]    (
PearsonR)
```

```
1 tissues = ['Breast', 'Colon', 'Kidney', 'Liver', 'Lung', 'Prostate'
     ↪ , 'Stomach', 'Thyroid', 'Uterus']
2 (h.differential_expression_tissue_concordance(tissue_subset=tissues
     ↪ , tcga=False).relabel('GTEx') + \
3 h.differential_expression_tissue_concordance(tissue_subset=tissues,
     ↪  gtex=False).relabel('TCGA')).relabel('Differential
     ↪ Expression Gene Concordance (PearsonR)')
```

```
:Layout
   .HeatMap.GTEx :HeatMap    [Tissue-Tumor/Normal,Tissue-Normal]    (
PearsonR)
   .HeatMap.TCGA :HeatMap    [Tissue-Tumor/Normal,Tissue-Normal]    (
PearsonR)
```

For almost every tissue, the GTEx normal counterpart is the closest approximate to the TCGA normal. Many tissues, like *Bladder*, *Esophagus*, *Pancreas*, and *Skin* have so few TCGA normals that the GTEx counterpart not being highly concordant isn't suprising. This is corroborated by GTEx being the closest approximate for tissues with larger TCGA normal sample sizes.

1.6 Validating Known Targets of Cancer Drugs

1.7 Identifying Repositionable Candidates

1.8 Discussion

2 References

The bib file biblio.bib was not found

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!

(?) !! *This reference was not found in biblio.bib* !!