Schraivogel (2020) Data Documentation

Eugene Katsevich

March 16, 2022

Overview

The portion of the Schraivogel data that is currently imported is the one described in the section "TAP-seq sensitively detects gene expression changes" of their paper. It contains two separate experiments: one TAP-seq and one perturb-seq. These experiments are meant as a proof-of-concept for TAP-seq, so they contain only positive and negative control perturbations (perturbations for which the ground truth is known). Both experiments have the same experimental design; they differ only in that TAP-seq targets a small number of genes while perturb-seq targets the whole transcriptome.

The `processed` directory contains subdirectories corresponding to the two experiments:

```r
# top-level data directory
schraivogel_dir <-.get_config_path("LOCAL_SCHRAIVOGEL_2020_DATA_DIR")

# processed data directory
processed_dir <- sprintf("%sprocessed", schraivogel_dir)

# print subdirectories of `processed_dir`
list.dirs(processed_dir, full.names = FALSE, recursive = FALSE)
```

```
## [1] "ground_truth_perturbseq" "ground_truth_tapseq"
```

Each of these has subdirectories for the processed gRNA data, the processed gene expression data, and auxiliary data:

```r
# print subdirectories for one of the processed experiments
list.dirs(sprintf("%s/ground_truth_tapseq", processed_dir),
          full.names = FALSE, recursive = FALSE)
```

```
## [1] "aux"  "gene" "gRNA"
```

Experimental design

As mentioned above, the experimental design for these two ground truth experiments is the same. The file containing the experimental design is in the `aux` directory; let's take a look at its first few rows:

```r
aux_dir <- sprintf("%s/ground_truth_tapseq/aux", processed_dir)
exper_design = readRDS(sprintf("%s/experimental_design.rds", aux_dir))
exper_design %>%
  head(5) %>%
  kableExtra::kable(format = "latex",
                    booktabs = TRUE,
                    col.names = c("gRNA", "Target", "Target Type", "Known Effect"))
```

gRNA	Target	Target Type	Known Effect
CCNE2_+_95907328.23-P1P2	CCNE2-TSS	promoter	CCNE2
CCNE2_+_95907382.23-P1P2	CCNE2-TSS	promoter	CCNE2
CCNE2_+_95907406.23-P1P2	CCNE2-TSS	promoter	CCNE2
CCNE2_-_95907017.23-P1P2	CCNE2-TSS	promoter	CCNE2
CPQ_+_97657557.23-P1P2	CPQ-TSS	promoter	CPQ

The full table contains a total of 86 rows, so there are 86 gRNAs in this experiment. Breaking these down by their target,

```
exper_design %>% pull(target) %>% table()
```

```
## .
##     CCNE2-TSS       CPQ-TSS     DSCC1-TSS    FAM83A-TSS     GATA1-enh 
##             4             4             4             4             4 
##       HS2-enh    LRRCC1-TSS       MYC-enh non-targeting      OXR1-TSS 
##             4             4             4            30             4 
##   PHF20L1-TSS     RIPK2-TSS      STK3-TSS      UBR5-TSS     ZFPM2-enh 
##             4             4             4             4             4
```

we see that we have 30 non-targeting gRNAs as well as four gRNAs each for 14 targets ($30 + 4 \times 14 = 86$). Of these targets, 10 are gene TSSs and 4 are well-characterized enhancers of four genes.

TAP-seq data

Let's now take a look at the TAP-seq data:

```
# load the gRNA expression data
processed_gRNA_dir <- sprintf("%s/ground_truth_tapseq/gRNA", processed_dir)
gRNA_odm_fp <- sprintf("%s/raw_ungrouped.odm", processed_gRNA_dir)
gRNA_metadata_fp <- sprintf("%s/raw_ungrouped_metadata.rds", processed_gRNA_dir)
gRNA_expr_odm <- ondisc::read_odm(gRNA_odm_fp, gRNA_metadata_fp)
gRNA_expr_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 86 features and 21977 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis, batch.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero.
```

```
# load the gene expression data
processed_gene_dir <- sprintf("%s/ground_truth_tapseq/gene", processed_dir)
gene_odm_fp <- sprintf("%s/expression_matrix.odm", processed_gene_dir)
gene_metadata_fp <- sprintf("%s/metadata.rds", processed_gene_dir)
gene_expr_odm <- ondisc::read_odm(gene_odm_fp, gene_metadata_fp)
gene_expr_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 72 features and 21977 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis, batch.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero.
```

This experiment has 21977 cells across 2 batches. The gRNA data come in the form of expressions and are not thresholded. There are a total of 86 gRNAs, as discussed above. A total of 72 genes are measured. Based on the paper, there are supposed to be 74 genes measured: 14 that were targeted and 60 presumably unrelated genes. Of the two missing genes, one is HS2 (whose enhancer was targeted) and one is a presumably unrelated gene. We can look further into why this is the case, but perhaps it's not urgent.

Perturb-seq data

Finally, we turn to the perturb-seq data:

```r
# load the gRNA expression data
processed_gRNA_dir <- sprintf("%s/ground_truth_perturbseq/gRNA", processed_dir)
gRNA_odm_fp <- sprintf("%s/raw_ungrouped.odm", processed_gRNA_dir)
gRNA_metadata_fp <- sprintf("%s/raw_ungrouped_metadata.rds", processed_gRNA_dir)
gRNA_expr_odm <- ondisc::read_odm(gRNA_odm_fp, gRNA_metadata_fp)
gRNA_expr_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 85 features and 37918 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis, batch.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero.
```

```r
# load the gene expression data
processed_gene_dir <- sprintf("%s/ground_truth_perturbseq/gene", processed_dir)
gene_odm_fp <- sprintf("%s/expression_matrix.odm", processed_gene_dir)
gene_metadata_fp <- sprintf("%s/metadata.rds", processed_gene_dir)
gene_expr_odm <- ondisc::read_odm(gene_odm_fp, gene_metadata_fp)
gene_expr_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 17107 features and 37918 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis, batch.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero.
```

This experiment has 37918 cells across 4 batches. The gRNA data come in the form of expressions and are not thresholded. There are a total of 85 gRNAs, which is one fewer than the 86 in the experimental design. Perhaps the missing one (STK3_-_99837866.23-P1P2) got removed during QC by Schraivogel et al? I am not sure. Unlike the TAP-seq experiment, we have measured the whole transcriptome (a total of 17107 genes).