DRIMSeq: Dirichlet-multinomial framework for differential transcript usage and transcript usage QTL analyses in RNA-seq

Malgorzata Nowicka, **Mark D. Robinson***

Institute for Molecular Life Sciences, University of Zurich, Switzerland
SIB Swiss Institute of Bioinformatics, University of Zurich, Switzerland

*gosia.nowicka@uzh.ch

May 21, 2023

Package

DRIMSeq 1.29.0

Contents

1 Main changes in the DRIMSeq package

For the full list of changes, type:

```
news(package = "DRIMSeq")
```

Implementation of the regression framework in differential transcript usage analysis. It allows fitting more complicated than multiple group comparison experimental designs, for example, one can account for the batch effects or the fact that samples are paired or model continuous time course changes. It enables also testing of more complicated contrasts.

Transcript-level analysis based on the beta-binomial model. In this case, each transcript ratio is modeled separately assuming it follows the beta-binomial distribution which is a one-dimensional version of the Dirichlet-multinomial distribution. Based on the fact that when $(Y_1, \ldots, Y_q) \sim DM(\pi_1, \ldots, \pi_q, \gamma_0)$ then $Y_j \sim BB(\pi_j, \gamma_0)$ [1], we do not need to reestimate the beta-binomial parameters, only the likelihoods are recalculated. *DRIMSeq* returns gene-level and transcript-level p-values which can be used as input to the stage-wise testing procedure [2] as screening and confirmation p-values, respectively. Such approach provides increased power to identify transcripts which are actually differentially used in a gene detected as gene with DTU.

Usage of term 'precision' instead of 'dispersion'. In the differential analysis based on the negative-binomial model, dispersion parameter is estimated. This dispersion parameter captures all sources of the inter-library variation between replicates present in the RNA-seq data. In the DM model, we do not directly estimate dispersion but a parameter called precision which is closely linked to dispersion via the formula: $dispersion = 1/(1 + precision)$. In the previous version of [3], we used 'dispersion' as a name for the functions and variables calculating and storing, in fact, the precision estimates. Now, we use the term 'precision'.

2 Overview of the Dirichlet-multinomial model

For the statistical details about the Dirichlet-multinomial model, see the *DRIMSeq* paper [3].

In the *DRIMSeq* package we implemented a Dirichlet-multinomial framework that can be used for modeling various multivariate count data with the interest in finding the instances where the ratios of observed features are different between the experimental conditions. Such a model can be applied, for example, in differential transcript usage (DTU) analysis or in the analysis that aim in detecting SNPs that are associated with differential transcript usage (tuQTL analysis). In both cases the multivariate features of a gene are transcripts. The implementation of Dirichlet-multinomial model in *DRIMSeq* package is customized for differential transcript usage and tuQTL analyses, but the data objects used in *DRIMSeq* can contain various types of counts. Therefore, other types of multivariate differential analyses can be performed such as differential methylation analysis or differential polyA usage from polyA-seq data.

In short, the method consists of three statistical steps:

First, we use the Cox-Reid adjusted profile likelihood to estimate the precision which is inverse proportional to dispersion, i.e., the variability of transcript ratios between samples (replicates) within conditions. Dispersion is needed in order to find the significant changes in transcript ratios between conditions which should be sufficiently stronger than the changes/variability within conditions.

Second, we use maximum likelihood to estimate at the gene-level the regression coefficients in the Dirichlet-multinomial (DM) regression, the fitted transcript proportions in each sample and the full model likelihoods. For the analysis at the transcript-level we apply the beta-binomial (BB) regression to each transcript separately. In the differential transcript usage analysis, the full model is defined by the user with the design matrix. In the QTL analysis, full models are defined by the genotypes of SNPs associated with a given gene.

Finally, we fit the null model DM and BB regression and use the likelihood ratio statistics to test for the differences in transcript proportions between the full and null models at the gene and transcript level. In the differential transcript usage analysis, the null model is again defined by the user. In the QTL analysis, null models correspond to regression with intercept only.

3 Important notes

Currently, transcript-level analysis based on the BB model are implemented only in the DTU analysis (`bb_model = TRUE`).

When the model (full or null) of interest corresponds to multiple (or one) group fitting, then a shortcut algorithm called 'one way' (`one_way = TRUE`), which we adapted from the `glmFit` function in *edgeR* [4], can be used. Choosing it is equivalent to running the original *DRIMSeq* implementation. In such a case, we use maximum likelihood to estimate the transcript proportions in each group separately and then the regression coefficients are calculated using matrix operations. Otherwise, the regression coefficients are directly estimated with the maximum likelihood approach.

4 Hints for DRIMSeq pipelines

In this vignette, we present how one could perform differential transcript usage analysis and tuQTL analysis with the *DRIMSeq* package. We use small data sets so that the whole pipelines can be run within few minutes in *R* on a single core computer. In practice, the package is designed to take advantage of multicore computing for larger data sets.

In the filtering function `dmFilter`, all the parameters that are influencing transcript count filtering are set to zero. This results in a very relaxed filtering, where transcripts with zero expression in all the samples and genes with only one transcript remained are removed.

Functions `dmPrecision`, `dmFit` and `dmTest`, which perform the actual statistical analyses described above, have many other parameters available for tweaking, but they do have the default values assigned. Those values were chosen based on many real data analyses. Some of the steps are quite time consuming, especially the precision estimation, where proportions of each gene are refitted for different precision parameters. To speed up the calculations, we have parallelized many functions using *BiocParallel*. Thus, if possible, we recommend to increase the number of workers in `BPPARAM`.

In general, tuQTL analyses are more computationally intensive than differential transcript usage analysis because one needs to do the analysis for every SNP in the surrounding region of a gene. Additionally, a permutation scheme is used to compute the p-values. It is indeed feasible to perform tuQTL analysis for small chunks of genome, for example, per chromosome.

5 Differential transcript usage analysis workflow

5.1 Example data

To demonstrate the application of *DRIMSeq* in differential transcript usage analysis, we will use the *pasilla* data set produced by Brooks et al. [5]. The aim of their study was to identify exons that are regulated by pasilla protein, the Drosophila melanogaster ortholog of mammalian NOVA1 and NOVA2 (well studied transcript usage factors). In their RNA-seq experiment, the libraries were prepared from 7 biologically independent samples: 4 control samples and 3 samples in which pasilla was knocked-down. The libraries were sequenced on Illumina Genome Analyzer II using single-end and paired-end sequencing and different read lengths. The RNA-seq data can be downloaded from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE18508. In the examples below, we use a subset of *kallisto* [6] counts available in *PasillaTranscriptExpr* package, where you can find all the steps needed, for preprocessing the GEO data, to get a table with transcript counts.

5.2 Differential transcript usage analysis between two conditions

In this section, we present how to perform the DTU analysis between two conditions control and knock-down without accounting for the batch effect which is the library layout. Analysis where batch effects are included are presented in the next section.

We start the analysis by creating a *dmDSdata* object, which contains transcript counts and information about grouping samples into conditions. With each step of the pipeline, additional elements are added to this object. At the end of the analysis, the object contains results from all the steps, such as precision estimates, regression coefficients, fitted transcript ratios in each sample, likelihood ratio statistics, p-values, adjusted p-values at gene and transcript level. As new elements are added, the object also changes its name *dmDSdata* $\rightarrow$ *dmDSprecision* $\rightarrow$ *dmDSfit* $\rightarrow$ *dmDStest*, but each container inherits slots and methods available for the previous one.

5.2.1 Loading pasilla data into R

The transcript-level counts obtained from *kallisto* and metadata are saved as text files in the `extdata` directory of the *PasillaTranscriptExpr* package.

```
library(PasillaTranscriptExpr)

data_dir  <- system.file("extdata", package = "PasillaTranscriptExpr")

## Load metadata
pasilla_metadata <- read.table(file.path(data_dir, "metadata.txt"),
  header = TRUE, as.is = TRUE)

## Load counts
pasilla_counts <- read.table(file.path(data_dir, "counts.txt"),
  header = TRUE, as.is = TRUE)
```

Load the *DRIMSeq* package.

```
library(DRIMSeq)
```

To create a `dmDSdata` object, saved as variable `d`, we need to prepare a data frame containing information about samples and we will call it `pasilla_samples`. It has to have a variable called `sample_id` with unique sample names that are identical to column names in `pasilla_counts` that correspond to samples. Additionally, it has to contain other variables that the user would like to use for the further regression analysis. Here, we are interested in the differential analysis between the control and knock-down condition. This information is stored in `pasilla_metadata$condition`.

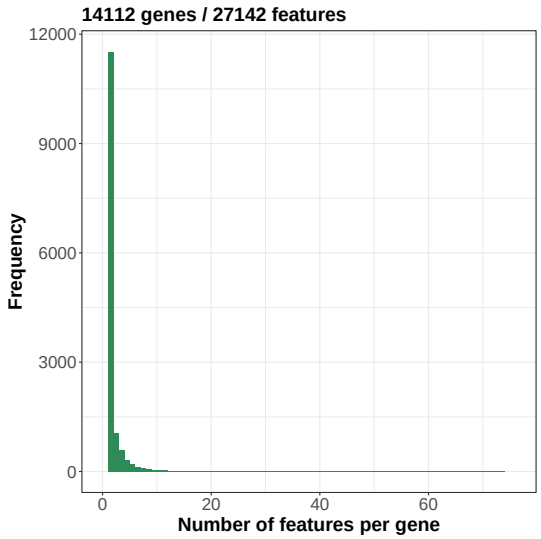
The data frame with counts called `pasilla_counts` is already formatted in the right way. It contains variables `feature_id` with unique transcript names and `gene_id` with unique gene IDs and columns with counts have the same names as `sample_id` in `pasilla_samples`.

When printing variable `d`, you can see its class, size (number of genes and samples) and which accessor methods can be applied. For `dmDSdata` object, there are two methods that return data frames with counts and samples.

```
pasilla_samples <- data.frame(sample_id = pasilla_metadata$SampleName,
  group = pasilla_metadata$condition)
levels(pasilla_samples$group)

## NULL

d <- dmDSdata(counts = pasilla_counts, samples = pasilla_samples)
d

## An object of class dmDSdata
## with 14112 genes and 7 samples
## * data accessors: counts(), samples()

head(counts(d), 3)

##       gene_id  feature_id GSM461176 GSM461177 GSM461178 GSM461179 GSM461180
## 1 FBgn0031208 FBtr0300689   0.00000         0   0.00000  27.04866         0
## 2 FBgn0031208 FBtr0300690   5.01688         4   1.00518   0.00000         2
## 3 FBgn0031208 FBtr0330654   0.00000         0   0.00000   0.00000         0
##   GSM461181 GSM461182
## 1   0.00000         0
## 2   2.00464         0
## 3   0.00000         0

head(samples(d), 3)

##   sample_id group
## 1 GSM461176   CTL
## 2 GSM461177   CTL
## 3 GSM461178   CTL
```

You can also make a data summary plot, which is a histogram of the number of transcripts per gene.

```
plotData(d)
```



To make the analysis runnable within this vignette, we want to keep only a small subset of genes, which is defined in the following file.

```
gene_id_subset <- readLines(file.path(data_dir, "gene_id_subset.txt"))
d <- d[names(d) %in% gene_id_subset, ]
d
```

```
## An object of class dmDSdata
## with 42 genes and 7 samples
## * data accessors: counts(), samples()
```

5.2.2 Filtering

Genes may have many transcripts that are lowly expressed or not expressed at all. You can remove them using the `dmFilter` function. Filtering of lowly expressed transcripts can be done at two levels: minimal *expression* using `min_samps_feature_expr` and `min_feature_expr` parameters or minimal *proportion* with `min_samps_feature_prop` and `min_feature_prop`.

In the *pasilla* experiment we use a filtering based only on the transcript absolute expression and parameters are adjusted according to the number of replicates per condition. Since we have 3 knock-down and 4 control samples, we set `min_samps_feature_expr` equal to 3. In this way, we allow a situation where a transcript is expressed in one condition but not in another, which is a case of differential transcript usage. The level of transcript expression is controlled by `min_feature_expr`. We set it to the value of 10, which means that only the transcripts that have at least 10 estimated counts in at least 3 samples are kept for the downstream analysis.

Filtering at the gene level ensures that the observed transcript ratios have some minimal reliability. Although, Dirichlet-multinomial model works on feature counts, and not on feature ratios, which means that it gives more confidence to the ratios based on 100 versus 500 reads than 1 versus 5, minimal filtering based on gene expression removes the genes with mostly zero counts and reduces the number of tests in multiple test correction. For the *pasilla* data, we want that genes have at least 10 counts in all the samples: `min_samps_gene_expr = 7` and `min_gene_expr = 10`.

```
# Check what is the minimal number of replicates per condition
table(samples(d)$group)
```

```
##
## CTL  KD
##   4   3
```

```
d <- dmFilter(d, min_samps_gene_expr = 7, min_samps_feature_expr = 3,
  min_gene_expr = 10, min_feature_expr = 10)
```

5.2.3 Precision estimation

Ideally, we would like to get accurate precision estimates for every gene, which is problematic when analyzing small data sets because precision estimates become inaccurate when the sample size decreases, especially for lowly expressed genes. As an alternative, we could assume that all the genes have the same precision and based on all the data, we could calculate a common precision, but we expect this to be too strong of an assumption. Moderated precision is a trade-off between gene-wise and common precision. The moderated estimates originate from a weighted likelihood which is a combination of common and individual likelihoods. We recommend this approach when analyzing small sample size data sets.

At this step, three values may be calculated: mean expression of genes, common precision and gene-wise precisions. In the default setting, all of them are computed and common precision is used as an initial value in the grid approach to estimate gene-wise precisions, which are shrunk toward the trended precision. The grid approach is adapted from the `estimateDisp` function in *edgeR* [4].

By default, to estimate the common precision, we use 10% percent (`prec_subset = 0.1`) of randomly selected genes. That is due to the fact that common precision is used only as an initial value, and estimating it based on all the genes takes a substantial part of time. To ensure that the analysis are reproducible, the user should define a random seed `set.seed()` before running the `dmPrecision()` function. Thank to that, each time the same subset of genes is selected.

To estimate precision parameters, the user has to define a design matrix with the full model of interest, which will be also used later in the proportion estimation. Here, the full model is defined by a formula $\sim group$ which indicates that samples come from two conditions.

This step of our pipeline is the most time consuming. Thus, for real data analysis, consider using `BPPARAM = BiocParallel::MulticoreParam()` with more than one worker.

```
## Create the design matrix
design_full <- model.matrix(~ group, data = samples(d))
design_full
```

```
##   (Intercept) groupKD
## 1           1       0
## 2           1       0
## 3           1       0
## 4           1       1
## 5           1       1
## 6           1       1
## 7           1       0
## attr(,"assign")
```

```
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

```
## To make the analysis reproducible
set.seed(123)
## Calculate precision
d <- dmPrecision(d, design = design_full)
```

```
## !  Using a subset of 0.1 genes to estimate common precision !
```

```
## !  Using common_precision = 44.3452 as prec_init !
```

```
## !  Using 0 as a shrinkage factor !
```

```
d
```

```
## An object of class dmDSprecision
## with 26 genes and 7 samples
## * data accessors: counts(), samples()
##   design()
##   mean_expression(), common_precision(), genewise_precision()
```

```
head(mean_expression(d), 3)
```

```
##       gene_id mean_expression
## 1 FBgn0000256        2622.286
## 2 FBgn0020309       13217.714
## 3 FBgn0259735       11992.903
```

```
common_precision(d)
```

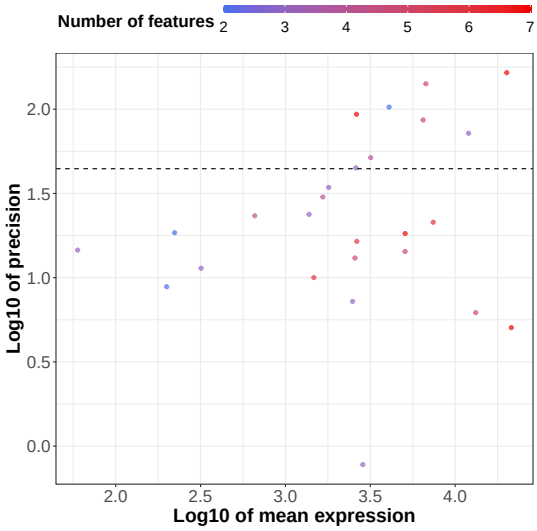
```
## [1] 44.34515
```

```
head(genewise_precision(d))
```

```
##       gene_id genewise_precision
## 1 FBgn0000256          93.388660
## 2 FBgn0020309           6.196615
## 3 FBgn0259735          72.027232
## 4 FBgn0032785          11.380309
## 5 FBgn0040297          13.085074
## 6 FBgn0032979         103.009092
```

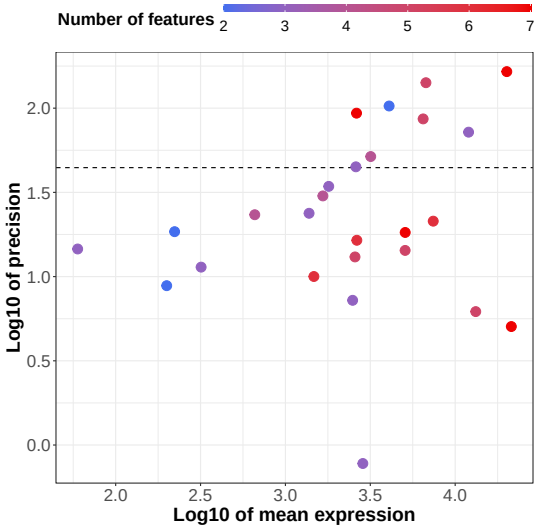
To inspect the behavior of precision estimates, you can plot them against the mean gene expression. Normally in the differential analysis based on RNA-seq data, such plot has dispersion parameter plotted on the y-axis. Here, the y-axis represents precision since in the Dirichlet-multinomial model this is the parameter that is directly estimated. It is important to keep in mind that the precision parameter is inverse proportional to dispersion: $dispersion = 1/(1 + precision)$. In RNA-seq data, we can typically observe a trend where the dispersion decreases (precision increases) for genes with higher mean expression.

```
plotPrecision(d)
```



All of the plotting functions from *DRIMSeq* package, return a *ggplot* object which can be further modified using *ggplot2* package. For example, the user can increase the size of points.

```
library(ggplot2)
ggp <- plotPrecision(d)
ggp + geom_point(size = 4)
```



5.2.4 Proportion estimation

In this step, we estimate the full model regression coefficients, fitted proportions and likelihoods. We use the same design matrix as in the precision estimation step.

By default, `one_way = TRUE` which means that whenever the design corresponds to multiple group fitting, the 'one way' shortcut algorithm will be used, which in fact corresponds to the first implementation of *DRIMSeq*. Transcript proportions are estimated for each condition separately and then the regression coefficients are calculated using matrix operations. The 'one way' algorithm is adapted from the `glmFit` function in *edgeR* [4]. By setting `verbose = 1`, we can see that the one way approach is used with the current design.

When `bb_model = TRUE` (the default), additionally to the gene-level Dirichlet-multinomial estimates the transcript-level beta-binomial results will be computed.

```
d <- dmFit(d, design = design_full, verbose = 1)

## * Fitting the DM model..

##    Using the one way approach.

## Took 0.1379 seconds.

## * Fitting the BB model..

##    Using the one way approach.

## Took 0.0818 seconds.

d

## An object of class dmDSfit
## with 26 genes and 7 samples
## * data accessors: counts(), samples()
##   design()
##   mean_expression(), common_precision(), genewise_precision()
##   proportions(), coefficients()

## Get fitted proportions
head(proportions(d))

##       gene_id  feature_id   GSM461176   GSM461177   GSM461178   GSM461179
## 1 FBgn0000256 FBtr0290077 0.357375706 0.357375706 0.357375706 0.076553229
## 2 FBgn0000256 FBtr0290078 0.043422163 0.043422163 0.043422163 0.270030880
## 3 FBgn0000256 FBtr0290082 0.006049622 0.006049622 0.006049622 0.001885836
## 4 FBgn0000256 FBtr0077511 0.286678646 0.286678646 0.286678646 0.222608942
## 5 FBgn0000256 FBtr0290081 0.016852677 0.016852677 0.016852677 0.036981836
## 6 FBgn0000256 FBtr0077513 0.280207783 0.280207783 0.280207783 0.378704927
##     GSM461180   GSM461181   GSM461182
## 1 0.076553229 0.076553229 0.357375706
## 2 0.270030880 0.270030880 0.043422163
## 3 0.001885836 0.001885836 0.006049622
## 4 0.222608942 0.222608942 0.286678646
## 5 0.036981836 0.036981836 0.016852677
## 6 0.378704927 0.378704927 0.280207783

## Get the DM regression coefficients (gene-level)
head(coefficients(d))

##       gene_id  feature_id X.Intercept.     groupKD
## 1 FBgn0000256 FBtr0290077    3.6366530 -1.88148245
## 2 FBgn0000256 FBtr0290078    1.5288353  1.48688523
## 3 FBgn0000256 FBtr0290082   -0.4421389 -1.50630555
## 4 FBgn0000256 FBtr0077511    3.4162273 -0.59362640
## 5 FBgn0000256 FBtr0290081    0.5823749  0.44523627
## 6 FBgn0000256 FBtr0077513    3.3933968 -0.03945519

## Get the BB regression coefficients (feature-level)
head(coefficients(d), level = "feature")
```

```
##       gene_id  feature_id X.Intercept.     groupKD
## 1 FBgn0000256 FBtr0290077    3.6366530 -1.88148245
## 2 FBgn0000256 FBtr0290078    1.5288353  1.48688523
## 3 FBgn0000256 FBtr0290082   -0.4421389 -1.50630555
## 4 FBgn0000256 FBtr0077511    3.4162273 -0.59362640
## 5 FBgn0000256 FBtr0290081    0.5823749  0.44523627
## 6 FBgn0000256 FBtr0077513    3.3933968 -0.03945519
```

5.2.5 Testing for differential transcript usage

Calling the `dmTest` function results in two calculations. First, null model is fitted. This null model can be defined by the user via `coef`, `design` or `contrast` parameters. Second, likelihood ratio statistics are used to test for the difference between the full and null model. Both steps are done at the gene and transcript level when `bb_model = TRUE`.

In our example, we would like to test whether there are differences in transcript usage between control (CTL) and knock-down (KD). We can achieve that by using the `coef` parameter which should indicate which columns of the full design should be removed to get the null design. We define it equal to `"groupKD"`. Then the null design has only an intercept column which means that all the samples are treated as if they came from one condition. Note that `one_way = TRUE` and the one way approach is used.

```
d <- dmTest(d, coef = "groupKD", verbose = 1)
```

```
## * Fitting the DM model..
```

```
##    Using the one way approach.
```

```
## Took 0.0826 seconds.
```

```
## * Calculating likelihood ratio statistics..
```

```
## Took 6e-04 seconds.
```

```
## * Fitting the BB model..
```

```
##    Using the one way approach.
```

```
## Took 0.0342 seconds.
```

```
## * Calculating likelihood ratio statistics..
```

```
## Took 4e-04 seconds.
```

```
design(d)
```

```
##   (Intercept)
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
## 7           1
```

```
head(results(d), 3)
```

```
##       gene_id          lr df       pvalue   adj_pvalue
## 1 FBgn0000256 146.1896729  6 4.942434e-29 1.285033e-27
## 2 FBgn0020309  17.9269288  4 1.275344e-03 4.467073e-03
## 3 FBgn0259735   0.9374648  2 6.257950e-01 7.747938e-01
```

The same can be achieved by directly defining the null design matrix with the `design` parameter.

```
design_null <- model.matrix(~ 1, data = samples(d))
design_null
```

```
##   (Intercept)
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
## 7           1
## attr(,"assign")
## [1] 0
```

```
d <- dmTest(d, design = design_null)
head(results(d), 3)
```

```
##       gene_id          lr df       pvalue   adj_pvalue
## 1 FBgn0000256 146.1896729  6 4.942434e-29 1.285033e-27
## 2 FBgn0020309  17.9269288  4 1.275344e-03 4.467073e-03
## 3 FBgn0259735   0.9374648  2 6.257950e-01 7.747938e-01
```

Or by using the `contrast` parameter. The null design is calculated using the approach from the `glmLRT` function in *edgeR* [4].

```
contrast <- c(0, 1)
d <- dmTest(d, contrast = contrast)
design(d)
```

```
##   [,1]
## 1   -1
## 2   -1
## 3   -1
## 4   -1
## 5   -1
## 6   -1
## 7   -1
```

```
head(results(d), 3)
```

```
##       gene_id          lr df       pvalue   adj_pvalue
## 1 FBgn0000256 146.1896729  6 4.942434e-29 1.285033e-27
## 2 FBgn0020309  17.9269288  4 1.275344e-03 4.467073e-03
## 3 FBgn0259735   0.9374648  2 6.257950e-01 7.747938e-01
```

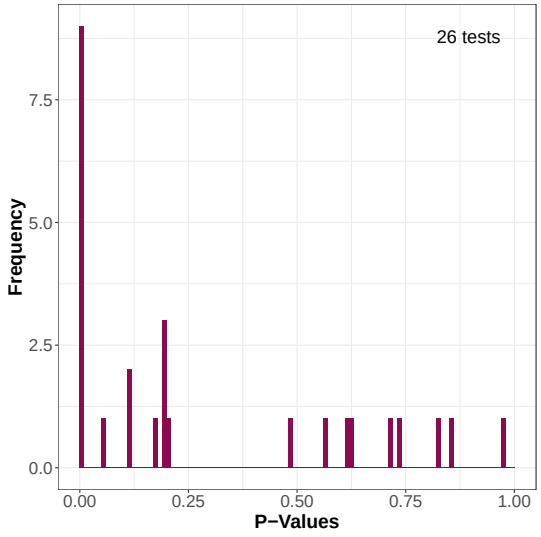
To obtain the results of likelihood ratio tests, you have to call the function `results`, which returns a data frame with likelihood ratio statistics, degrees of freedom, p-values and Benjamini and Hochberg (BH) adjusted p-values for each gene by default and for each transcript when `level = "feature"`.

```
head(results(d, level = "feature"), 3)
```

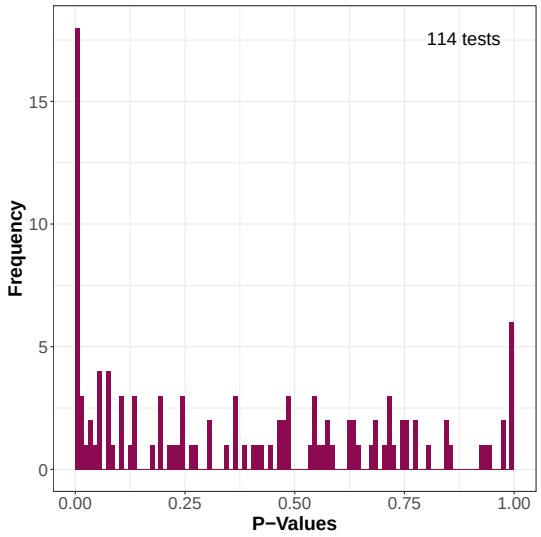
```
##       gene_id  feature_id        lr df       pvalue   adj_pvalue
## 1 FBgn0000256 FBtr0290077 87.619933  1 7.932120e-21 4.521309e-19
## 2 FBgn0000256 FBtr0290078 72.329719  1 1.820846e-17 6.919213e-16
## 3 FBgn0000256 FBtr0290082  1.341238  1 2.468158e-01 5.542627e-01
```

You can plot a histogram of gene-level and transcript-level p-values.

```
plotPValues(d)
```

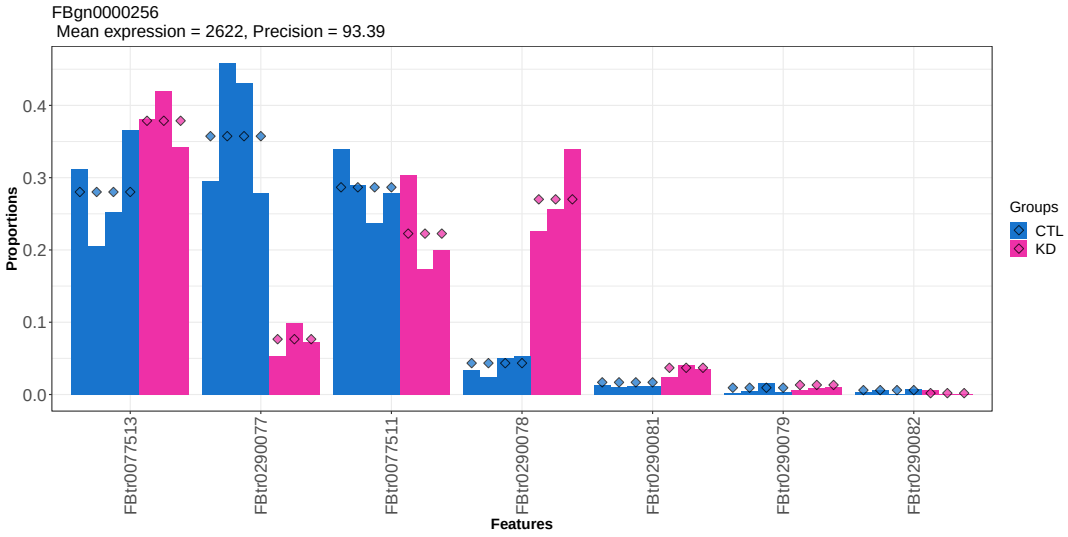



```
plotPValues(d, level = "feature")
```

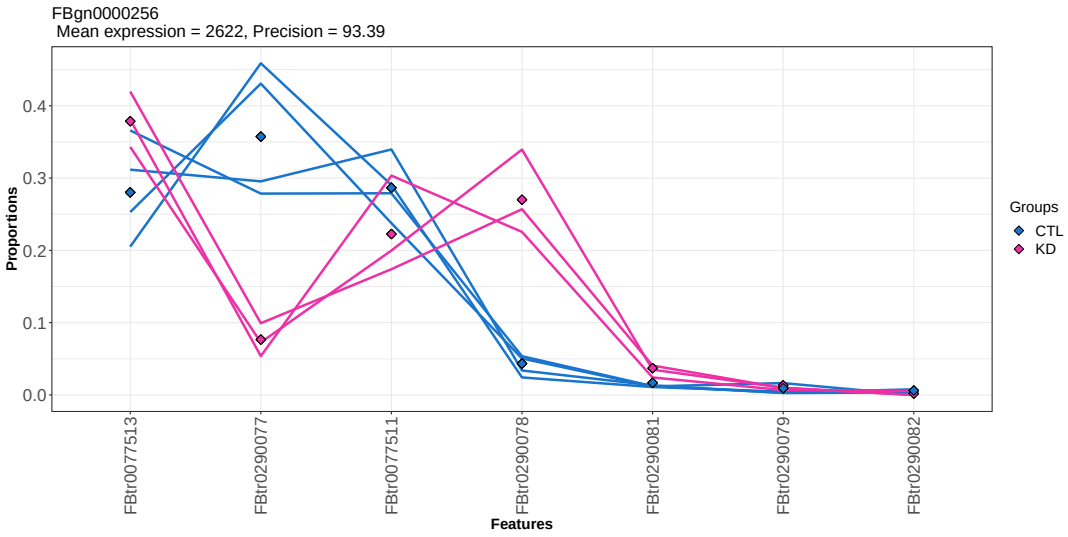


For genes of interest, you can make plots (bar plots, line plots, box plots, ribbon plots) of observed and estimated with Dirichlet-multinomial model transcript ratios. You have to define the `group_variable` parameter which should indicate a variable from `samples(d)`. Currently, plots can be done only for categorical variables. We choose the `"group"` column since it corresponds to the comparison of our interest. Estimated proportions are marked with diamond shapes. As an example, we plot the top significant gene.

```
res <- results(d)
res <- res[order(res$pvalue, decreasing = FALSE), ]
top_gene_id <- res$gene_id[1]

plotProportions(d, gene_id = top_gene_id, group_variable = "group")
```

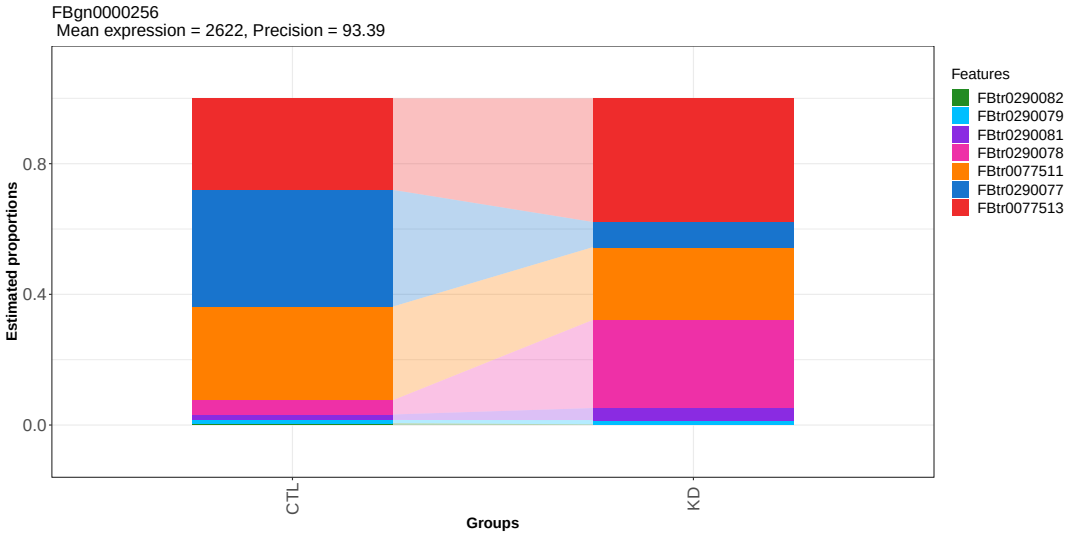



```
plotProportions(d, gene_id = top_gene_id, group_variable = "group",
  plot_type = "lineplot")
```



```
plotProportions(d, gene_id = top_gene_id, group_variable = "group",
  plot_type = "ribbonplot")
```

```
## Coordinate system already present.  Adding new coordinate system, which will
## replace the existing one.
```



5.2.6 Two-stage test

DRIMSeq returns gene and transcript level p-values which can be used as an input to the stage-wise analysis [2] implemented in the `stageR` package, currently available on github https://github.com/statOmics/stageR. As pointed by the authors of `stageR`, interpreting both gene-level and transcript-level adjusted p-values does not provide appropriate FDR control and should be avoided. However, applying a stage-wise testing provides a useful biological interpretation of these results and improved statistical performance.

In short, the procedure consists of a screening stage and a confirmation stage. In the screening stage, gene-level BH-adjusted p-values are screened to detect genes for which the hypothesis of interest is significantly rejected. Only those genes are further considered in the confirmation stage, where for each gene separately, transcript-level p-values are adjusted to control the FWER and BH-adjusted significance level of the screening stage.

It is important to note that transcript-level stage-wise adjusted p-values for genes that do not pass the screening stage are set to `NA`. Also the stage-wise adjusted p-values can not be compared to significance level other than chosen in the stage-wise analysis. If that is of interest, one has to rerun this analysis with the new significance level.

The following code chunk is not evaluated by this vignette and to run it, user has to make sure that the `stageR` package is installed. It shows how one can use the *DRIMSeq* output in the stage-wise analysis.

```
library(stageR)

## Assign gene-level pvalues to the screening stage
pScreen <- results(d)$pvalue
names(pScreen) <- results(d)$gene_id
```

```
## Assign transcript-level pvalues to the confirmation stage
pConfirmation <- matrix(results(d, level = "feature")$pvalue, ncol = 1)
rownames(pConfirmation) <- results(d, level = "feature")$feature_id

## Create the gene-transcript mapping
tx2gene <- results(d, level = "feature")[, c("feature_id", "gene_id")]

## Create the stageRTx object and perform the stage-wise analysis
stageRObj <- stageRTx(pScreen = pScreen, pConfirmation = pConfirmation,
  pScreenAdjusted = FALSE, tx2gene = tx2gene)

stageRObj <- stageWiseAdjustment(object = stageRObj, method = "dtu",
  alpha = 0.05)

getSignificantGenes(stageRObj)

getSignificantTx(stageRObj)

padj <- getAdjustedPValues(stageRObj, order = TRUE,
  onlySignificantGenes = FALSE)

head(padj)
```

5.3 Differential transcript usage analysis between two conditions with accounting for the batch effects

The regression framework implemented in *DRIMSeq* allows to account for the batch effects. Here, this would be the library layout stored in `pasilla_metadata$LibraryLayout`. The steps of this analysis are the same as described above. The only difference is that we have to include the library layout variable in the `sample` slot in the `dmDSdata` object and define a full model that contains the batch effect.

```
pasilla_samples2 <- data.frame(sample_id = pasilla_metadata$SampleName,
  group = pasilla_metadata$condition,
  library_layout = pasilla_metadata$LibraryLayout)

d2 <- dmDSdata(counts = pasilla_counts, samples = pasilla_samples2)

## Subsetting to a vignette runnable size
d2 <- d2[names(d2) %in% gene_id_subset, ]

## Filtering
d2 <- dmFilter(d2, min_samps_gene_expr = 7, min_samps_feature_expr = 3,
  min_gene_expr = 10, min_feature_expr = 10)

## Create the design matrix
design_full2 <- model.matrix(~ group + library_layout, data = samples(d2))
design_full2
```

```
##   (Intercept) groupKD library_layoutPAIRED
```

```
## 1           1       0                   0
## 2           1       0                   1
## 3           1       0                   1
## 4           1       1                   0
## 5           1       1                   1
## 6           1       1                   1
## 7           1       0                   0
## attr(,"assign")
## [1] 0 1 2
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
##
## attr(,"contrasts")$library_layout
## [1] "contr.treatment"
```

```
## To make the analysis reproducible
set.seed(123)

## Calculate precision
d2 <- dmPrecision(d2, design = design_full2)
```

```
## !  Using a subset of 0.1 genes to estimate common precision !
```

```
## !  Using common_precision = 70.5773 as prec_init !
```

```
## !  Using 0 as a shrinkage factor !
```

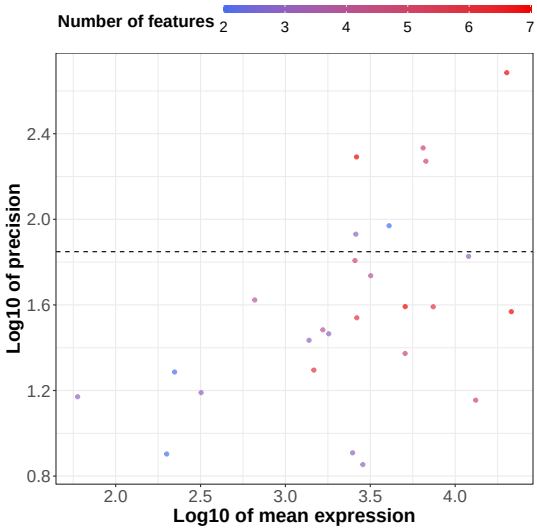
```
common_precision(d2)
```

```
## [1] 70.57734
```

```
head(genewise_precision(d2))
```

```
##       gene_id genewise_precision
## 1 FBgn0000256          195.62755
## 2 FBgn0020309           14.28474
## 3 FBgn0259735           67.13490
## 4 FBgn0032785           15.48549
## 5 FBgn0040297           64.17574
## 6 FBgn0032979           93.30566
```

```
plotPrecision(d2)
```



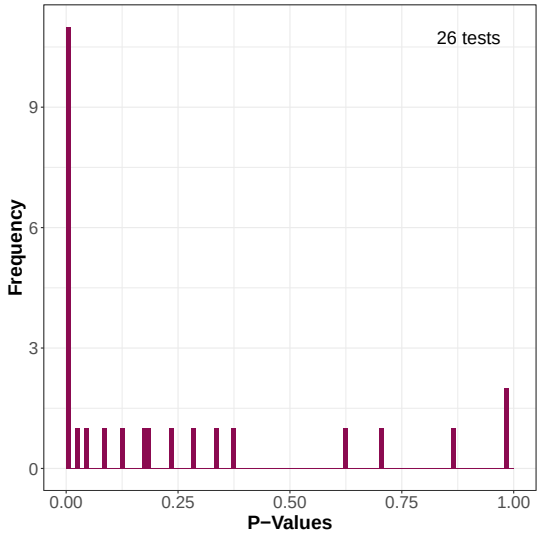

```
## Fit proportions
d2 <- dmFit(d2, design = design_full2, verbose = 1)

## * Fitting the DM model..

##    Using the regression approach.

## Took 0.3649 seconds.

## * Fitting the BB model..

##    Using the regression approach.

## Took 0.0571 seconds.

## Test for DTU
d2 <- dmTest(d2, coef = "groupKD", verbose = 1)

## * Fitting the DM model..

##    Using the one way approach.

## Took 0.1269 seconds.

## * Calculating likelihood ratio statistics..

## Took 2e-04 seconds.

## * Fitting the BB model..

##    Using the one way approach.

## Took 0.0563 seconds.

## * Calculating likelihood ratio statistics..

## Took 3e-04 seconds.

design(d2)

##   (Intercept) library_layoutPAIRED
## 1           1                    0
## 2           1                    1
## 3           1                    1
```

```
## 4           1                   0
## 5           1                   1
## 6           1                   1
## 7           1                   0
```

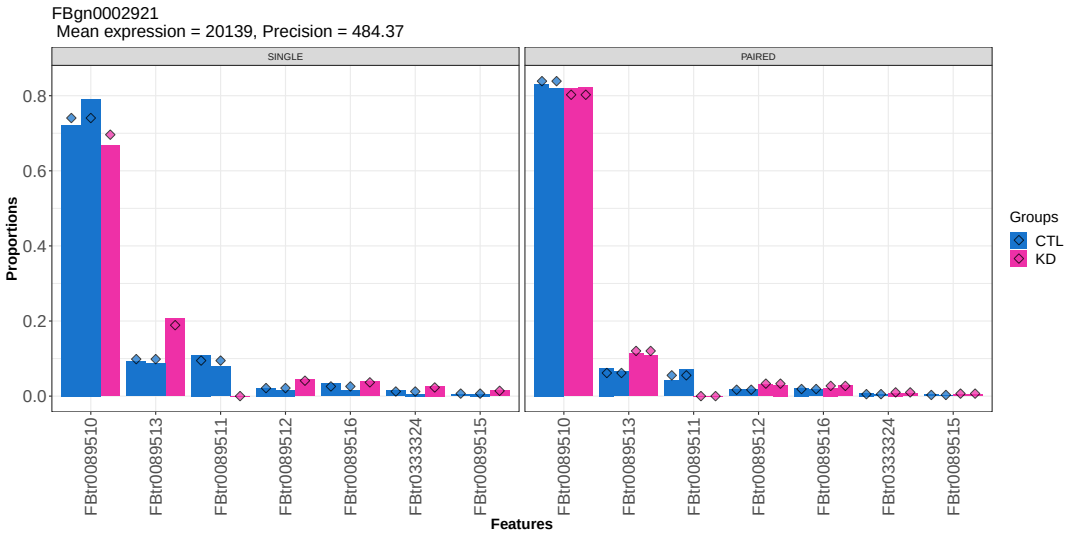
```
head(results(d2), 3)
```

```
##       gene_id          lr df       pvalue   adj_pvalue
## 1 FBgn0000256 297.3671486  6 2.998346e-61 3.897850e-60
## 2 FBgn0020309  18.6783136  4 9.089518e-04 2.625861e-03
## 3 FBgn0259735   0.9258417  2 6.294424e-01 7.438865e-01
```

```
## Plot p-value distribution
plotPValues(d2)
```




```
## Plot the top significant gene
res2 <- results(d2)
res2 <- res2[order(res2$pvalue, decreasing = FALSE), ]
top_gene_id2 <- res2$gene_id[1]
ggp <- plotProportions(d2, gene_id = top_gene_id2, group_variable = "group")
ggp + facet_wrap(~ library_layout)
```



6 tuQTL analysis workflow

In the transcript usage QTL analysis, we want to identify genetic variants (here, bi-allelic SNPs) that are associated with changes in transcript usage. Such SNPs are then called transcript usage quantitative trait locies (tuQTLs).

Ideally, we would like to test associations of every SNP with every gene. However, such an approach would be very costly computationally and in terms of multiple testing correction. Under the assumption that SNPs that directly affect transcript usage are likely to be placed in the close surrounding of genes, we test only the SNPs that are located within the gene body and within some range upstream and downstream of the gene.

6.1 Example data

To demonstrate the tuQTL analysis with the *DRIMSeq* package, we use data from the GEUVADIS project [7], where 462 RNA-Seq samples from lymphoblastoid cell lines were obtained. The genome sequencing data of the same individuals is provided by the 1000 Genomes Project. The samples in this project come from five populations: CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI). We use transcript quantification (expected counts from FluxCapacitor) and genotypes available on the GEUVADIS project website http://www.ebi.ac.uk/Tools/geuvadis-das/, and the Gencode v12 gene annotation is available at http://www.gencodegenes.org/releases/12.html.

In order to make this vignette runnable, we perform the analysis on subsets of bi-allelic SNPs and transcript expected counts for CEPH population (91 individuals) that correspond to 50 randomly selected genes from chromosome 19. The full dataset can be accessed from *GeuvadisTranscriptExpr* package along with the description of preprocessing steps.

6.2 tuQTL analysis with the DRIMSeq package

Assuming you have gene annotation, feature counts and bi-allelic genotypes that are expressed in terms of the number of alleles different from the reference, the *DRIMSeq* workflow for tuQTL analysis is analogous to the one for differential transcript usage.

First, we have to create a *dmSQTLdata* object, which contains feature counts, sample information and genotypes. Similarly as in the differential transcript usage pipeline, results from every step are added to this object and at the end of the analysis, it contains precision estimates, proportions estimates, likelihood ratio statistics, p-values, adjusted p-values. As new elements are added, the object also changes its name *dmSQTLdata* → *dmSQTLprecision* → *dmSQTLfit* → *dmSQTLtest*. For each object, slots and methods are inherited from the previous one.

6.2.1 Loading GEUVADIS data into R

We use the subsets of data defined in the *GeuvadisTranscriptExpr* package.

```
library(GeuvadisTranscriptExpr)

geuv_counts <- GeuvadisTranscriptExpr::counts
geuv_genotypes <- GeuvadisTranscriptExpr::genotypes
geuv_gene_ranges <- GeuvadisTranscriptExpr::gene_ranges
geuv_snp_ranges <- GeuvadisTranscriptExpr::snp_ranges
```

Load the *DRIMSeq* package.

```
library(DRIMSeq)
```

In the tuQTL analysis, an initial data object `d` is of `dmSQTLdata` class and, additionally to feature counts and sample information, it contains genotypes of SNPs that are in some surrounding of genes. This surrounding is defined with the parameter `window`. In order to find out which SNPs should be tested with which genes, the `dmSQTLdata` functions requires as an input the location of genes (`gene_ranges`) and SNPs (`snp_ranges`) stored as *GRanges* objects. Variables with transcript IDs and gene IDs in the `counts` data frame must have names `feature_id` and `gene_id`, respectively. In the `genotypes` data frame, the variable with SNP IDs must have name `snp_id`.

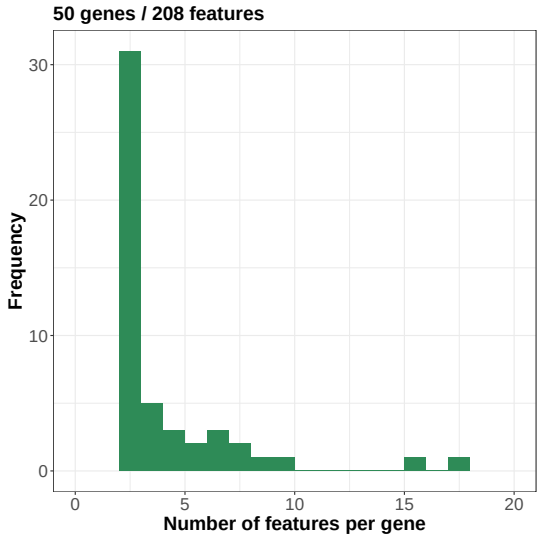
```
colnames(geuv_counts)[c(1,2)] <- c("feature_id", "gene_id")
colnames(geuv_genotypes)[4] <- "snp_id"
geuv_samples <- data.frame(sample_id = colnames(geuv_counts)[-c(1,2)])

d <- dmSQTLdata(counts = geuv_counts, gene_ranges = geuv_gene_ranges,
  genotypes = geuv_genotypes, snp_ranges = geuv_snp_ranges,
  samples = geuv_samples, window = 5e3)
d
```

```
## An object of class dmSQTLdata
## with 50 genes and 91 samples
## * data accessors: counts(), samples()
```

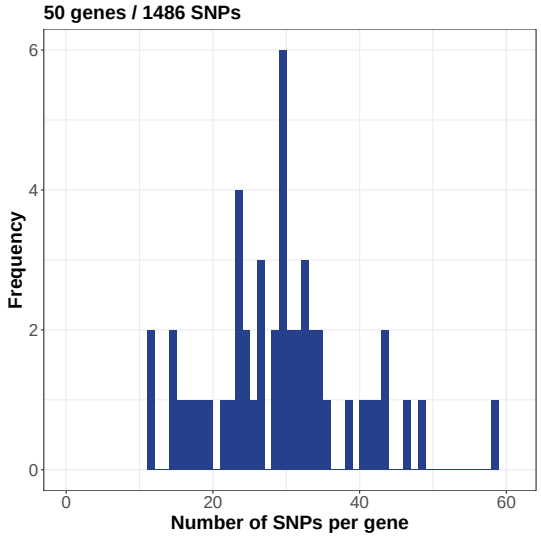
In our tuQTL analysis, we do not repeat tests for the SNPs that define the same grouping of samples (genotype). We identify SNPs with identical genotypes across the samples and assign them to blocks. Estimation and testing are done at the block level, but the returned results are extended to a SNP level by repeating the block statistics for each SNP that belongs to a given block.

The data summary plot `plotData` produces three histograms: the number of features per gene, the number of SNPs per gene and the number of blocks per gene.

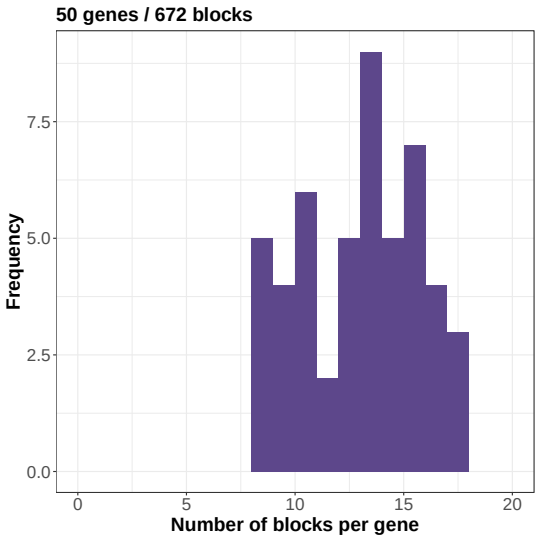
```
plotData(d, plot_type = "features")
```




```
plotData(d, plot_type = "snps")
```




```
plotData(d, plot_type = "blocks")
```



6.2.2 Filtering

The filtering step eliminates genes and features with low expression, as in the differential transcript usage analysis (see section 5.2.2). Additionally, it filters out the SNPs/blocks that do not define at least two genotypes where each of them is present in at least `minor_allele_freq` individuals. Usually, `minor_allele_freq` is equal to roughly 5% of the total sample size.

Ideally, we would like that genes were expressed at some minimal level in all samples because this would lead to better estimates of feature ratios. However, for some genes, missing values may be present in the counts data, or genes may be lowly expressed in some samples. Setting up `min_samps_gene_expr` to 91 may exclude too many genes from the analysis. We can be slightly less stringent by taking, for example, `min_samps_gene_expr = 70`.

```
d <- dmFilter(d, min_samps_gene_expr = 70, min_samps_feature_expr = 5,
  minor_allele_freq = 5, min_gene_expr = 10, min_feature_expr = 10)
```

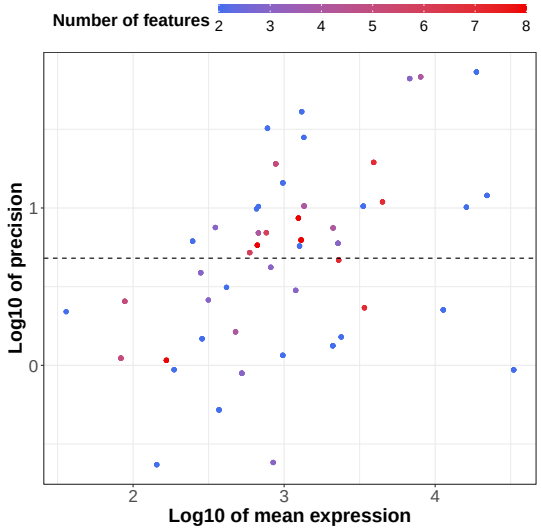
6.2.3 Precision estimation

In the DTU analysis (section 5.2.3), the full model used in precision estimation has to be defined by the user. Here, full models are defined by genotypes. For a given SNP, genotype can have numeric values of 0, 1, and 2. When `one_way = TRUE`, multiple group fitting is performed. When `one_way = FALSE`, a regression framework is used with the design matrix defined by a formula $\sim group$ where $group$ is a continuous (not categorical) variable with genotype values 0, 1, and 2.

For the tuQTL analysis, it has an additional parameter called `speed`. If `speed = FALSE`, gene-wise precisions are calculated for each gene-block. This calculation may take a long time, since there can be hundreds of SNPs/blocks per gene. If `speed` is set to `TRUE`, there will be only one precision calculated per gene (assuming a null model, i.e., model with intercept only), and it will be assigned to all the blocks matched to this gene. In the default setting, `speed = TRUE` and common precision is used as an initial value in the grid approach to estimate gene-wise precisions with NO moderation, since the sample size is quite large.

Again, this step of the pipeline is one of the most time consuming. Thus consider using `BPPARAM = BiocParallel::MulticoreParam()` with more than one worker when performing real data analysis.

```
## To make the analysis reproducible
set.seed(123)
## Calculate precision
d <- dmPrecision(d)

## !  Using a subset of 0.1 genes to estimate common precision !

## !  Using common_precision = 4.7973 as prec_init !

plotPrecision(d)
```



6.2.4 Proportion estimation

Dirichlet-multinomial full model proportions/coefficients and likelihoods are estimated for each gene-block pair. Currently, no transcript-level analysis are implemented in the tuQTL workflow.

```
d <- dmFit(d)
```

6.2.5 Testing for tuQTLs

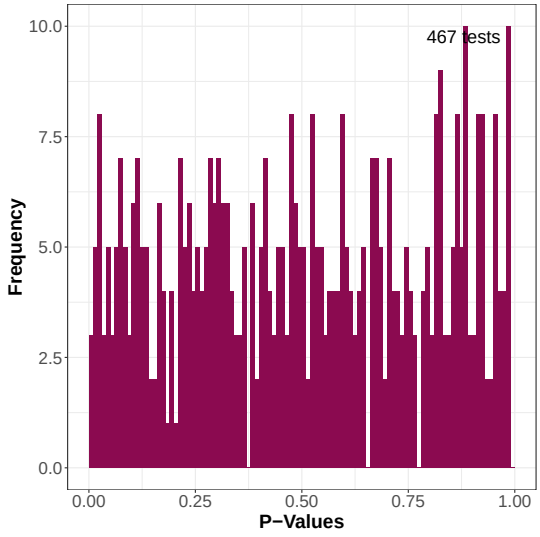
`dmTest` function estimates gene-level null model proportions/coefficients and likelihoods and performs the likelihood ratio test. The null models equal to models with intercept only.

In contrast to the DTU analysis, there are some additional challenges that have to handled in the tuQTL analysis. They include a large number of tests per gene with highly variable allele frequencies (models) and linkage disequilibrium. As in other sQTL studies, we apply a permutation approach to empirically assess the null distribution of associations and use it for the adjustment of nominal p-values.

There are two permutation schemes available. When `permutation_mode` equals to `"all_genes"`, the null p-value distribution is calculated from all the genes. When `permutation_mode = "per_gene"`, null distribution of p-values is calculated for each gene separately based on permutations of this individual gene. The latter approach may take a lot of computational time. We suggest using the first option, which is also the default one.

The function `results` returns a data frame with likelihood ratio statistics, degrees of freedom, p-values and Benjamini and Hochberg adjusted p-values for each gene-block/SNP pair.

```
d <- dmTest(d)
plotPValues(d)
```



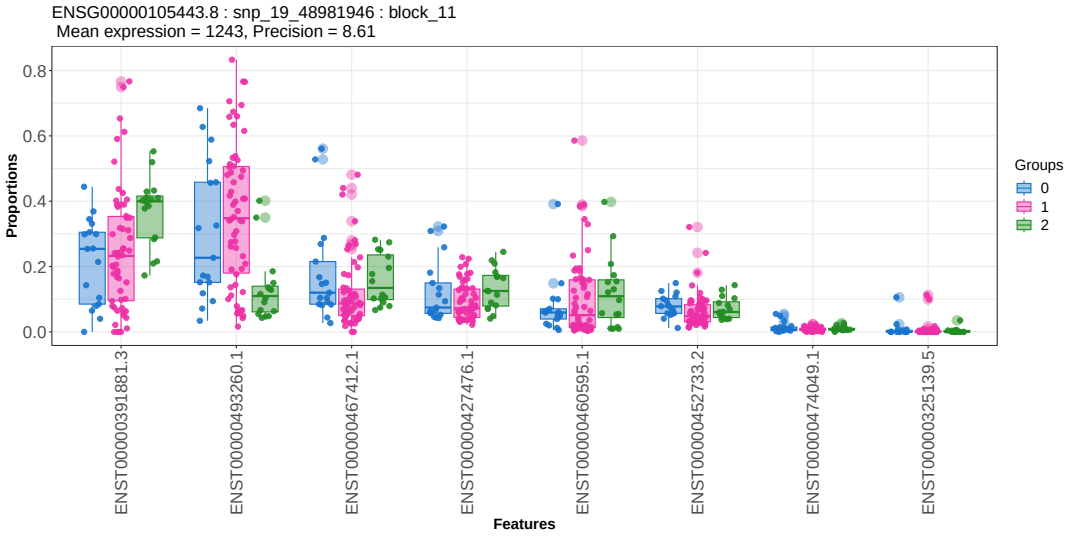

```
head(results(d))
```

```
##             gene_id block_id          snp_id       lr df    pvalue
## 1 ENSG00000221983.2  block_3 snp_19_18678970 1.257939  2 0.4124666
## 2 ENSG00000221983.2  block_3 snp_19_18685964 1.257939  2 0.4124666
## 3 ENSG00000221983.2  block_3 snp_19_18687175 1.257939  2 0.4124666
## 4 ENSG00000221983.2  block_3 snp_19_18689904 1.257939  2 0.4124666
## 5 ENSG00000221983.2  block_4 snp_19_18679155 1.072372  2 0.4503189
## 6 ENSG00000221983.2  block_4 snp_19_18679379 1.072372  2 0.4503189
##   adj_pvalue
## 1  0.9888912
## 2  0.9888912
## 3  0.9888912
## 4  0.9888912
## 5  0.9888912
## 6  0.9888912
```

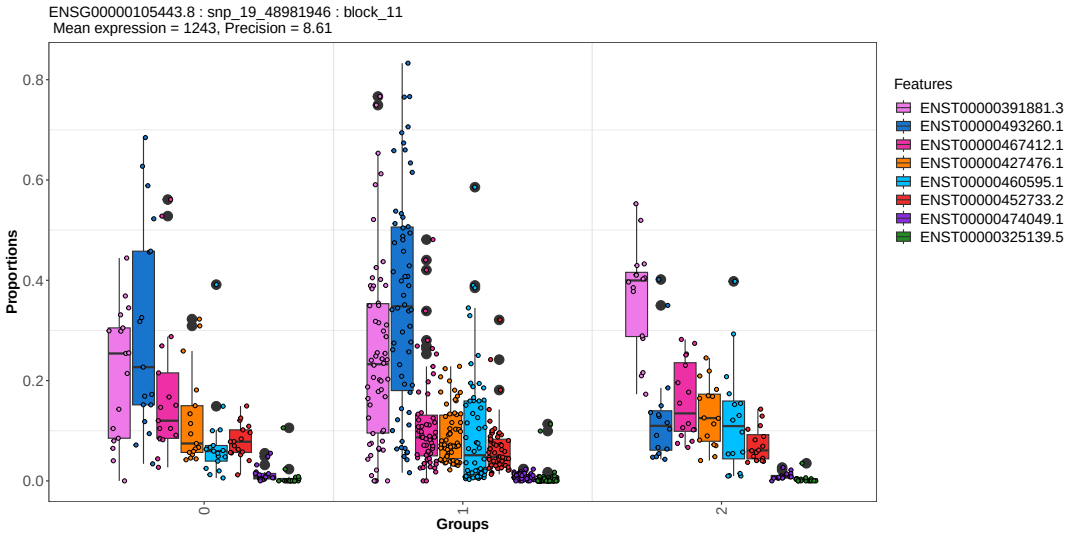
You can plot the observed transcript ratios for the tuQTLs of interest. Plotting the fitted values is not possible as we do not return this estimates due to their size. When the sample size is large, we recommend using box plots as a `plot_type`. We plot a tuQTL with the lowest p-value.

```
res <- results(d)
res <- res[order(res$pvalue, decreasing = FALSE), ]
```

```
top_gene_id <- res$gene_id[1]
top_snp_id <- res$snp_id[1]

plotProportions(d, gene_id = top_gene_id, snp_id = top_snp_id)
```




```
plotProportions(d, gene_id = top_gene_id, snp_id = top_snp_id,
  plot_type = "boxplot2")
```



7 Session information

```
sessionInfo()

## R version 4.3.0 RC (2023-04-18 r84287)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.2 LTS
```

```
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.18-bioc/R/lib/libRblas.so
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB              LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/New_York
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] GeuvadisTranscriptExpr_1.29.0 ggplot2_3.4.2
## [3] DRIMSeq_1.29.0                PasillaTranscriptExpr_1.29.0
## [5] knitr_1.42
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.3              generics_0.1.3          bitops_1.0-7
##  [4] stringi_1.7.12          lattice_0.21-8          digest_0.6.31
##  [7] magrittr_2.0.3          evaluate_0.21           grid_4.3.0
## [10] fastmap_1.1.1           plyr_1.8.8              GenomeInfoDb_1.37.1
## [13] limma_3.57.1            BiocManager_1.30.20     fansi_1.0.4
## [16] scales_1.2.1            codetools_0.2-19        cli_3.6.1
## [19] rlang_1.1.1             BiocStyle_2.29.0        XVector_0.41.1
## [22] munsell_0.5.0           withr_2.5.0             yaml_2.3.7
## [25] tools_4.3.0             parallel_4.3.0          reshape2_1.4.4
## [28] BiocParallel_1.35.1     dplyr_1.1.2             colorspace_2.1-0
## [31] locfit_1.5-9.7          GenomeInfoDbData_1.2.10 BiocGenerics_0.47.0
## [34] vctrs_0.6.2             R6_2.5.1                stats4_4.3.0
## [37] lifecycle_1.0.3         zlibbioc_1.47.0         stringr_1.5.0
## [40] S4Vectors_0.39.1        edgeR_3.43.2            IRanges_2.35.1
## [43] MASS_7.3-60             pkgconfig_2.0.3         pillar_1.9.0
## [46] gtable_0.3.3            glue_1.6.2              Rcpp_1.0.10
## [49] xfun_0.39               tibble_3.2.1            GenomicRanges_1.53.1
## [52] tidyselect_1.2.0        highr_0.10              farver_2.1.1
## [55] htmltools_0.5.5         labeling_0.4.2          rmarkdown_2.21
## [58] compiler_4.3.0          RCurl_1.98-1.12
```

References

[1] Peter J Danaher. Parameter estimation for the dirichlet-multinomial distribution using supplementary beta-binomial data. *Communications in Statistics - Theory and Methods*, 17(6):1777–1788, jan 1988. URL: http://dx.doi.org/10.1080/03610928808829713, doi:10.1080/03610928808829713.

[2] Koen Van den Berge, Charlotte Soneson, Mark D Robinson, and Lieven Clement. A general and powerful stage-wise testing procedure for differential expression and differential transcript usage. *bioRxiv*, feb 2017. URL: http://biorxiv.org/content/early/2017/02/16/109082.abstract.

[3] M Nowicka and M D Robinson. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; referees: 2 approved]. *F1000Research*, 5(1356), 2016. URL: https://f1000research.com/articles/5-1356/v2, doi:10.12688/f1000research.8900.2.

[4] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

[5] Angela N Brooks, Li Yang, Michael O Duff, Kasper D Hansen, Jung W Park, Sandrine Dudoit, Steven E Brenner, and Brenton R Graveley. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome research*, 21(2):193–202, 2011.

[6] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, advance on, apr 2016. URL: http://dx.doi.org/10.1038/nbt.3519http://10.0.4.14/nbt.3519http://www.nature.com/nbt/journal/vaop/ncurrent/abs/nbt.3519.html{#}supplementary-information.

[7] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3918453{&}tool=pmcentrez{&}rendertype=abstract.