Getting started with recount2 and accessing it via R

Collado-Torres L[1,2,*], Nellore A[3,4,5], Jaffe AE[1,2,6,7]

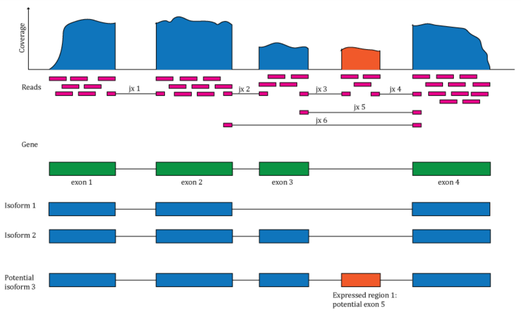
[1]Lieber Institute for Brain Development, [2]CCB JHU, [3]BME [4]Surgery and [5]CBP OHSU, [6]Biostatistics and [7]Mental Health JHU,

[*]leo.collado@libd.org

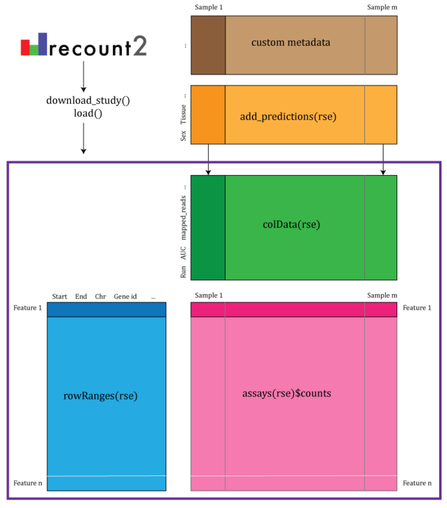
ABSTRACT

The recount2 resource is composed of over 70,000 uniformly processed human RNA-seq samples spanning TCGA and SRA, including GTEx. The processed data can be accessed via the recount2 website https://jhubiostatistics.shinyapps.io/recount/ and the recount Bioconductor package http://bioconductor.org/packages/recount. Here we describe the recount2 resource starting from how the coverage count matrices were computed in recount2 as well as different ways of obtaining public metadata, which can facilitate downstream analyses. We showcase how to use the recount package and how to integrate it with other Bioconductor packages. We illustrate step-by-step directions that show how to do a gene-level differential expression analysis, visualize base-level genome coverage data, and perform an analyses at multiple feature levels. The associated workflow at https://f1000research.com/articles/6-1558/v1 provides further information to understand the data in recount2 and a compendium of R code to use the data.

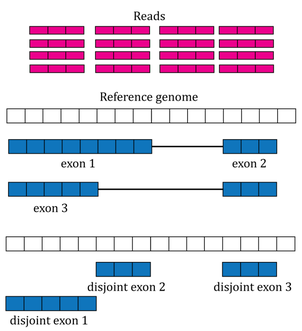
DATA IN RECOUNT2

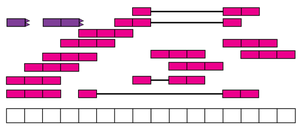


Gene, exon, exon-exon junction, and expressed region data is available from recount2. The data is provided as RangedSummarizedExperiment objects that can be easily downloaded and loaded in an R session.

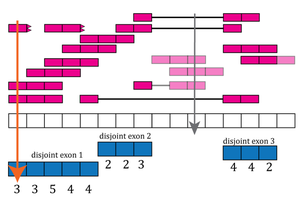


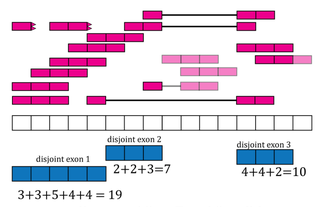
COUNTS PROVIDED



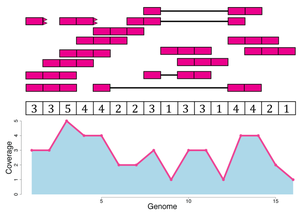


RNA-seq reads were aligned to the reference genome using Rail-RNA which can soft-clip reads. Using the annotation information from Gencode v25 we identified the *disjoint* exonic fragments of the transcriptome so that we could quantify them.





Gene = 19 + 7 + 10 = 36

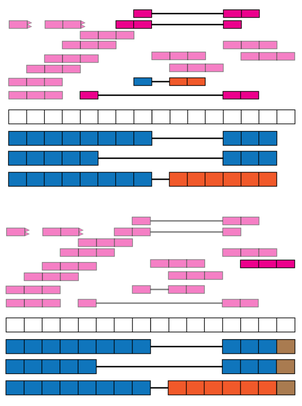


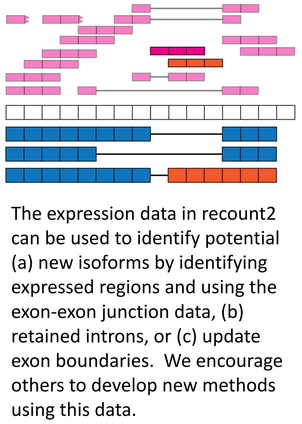
AUC = area under coverage = 45

We compute the base-pair coverage for all disjoint exons, and sum them to obtain the disjoint exon counts. We then sum the exon counts to obtain the gene counts. One could then scale the counts using equation (1) assuming equal read length. Due to soft-clipping, not all reads contribute equally so we use the area under the coverage as shown in equation (2) to scale the counts to a common library size.

$$\frac{\sum_i^n \text{coverage}_i}{\text{Read Length}} * \frac{\text{target}}{\text{mapped}} = \text{scaled read counts} \quad (1)$$

$$\frac{\sum_i^n \text{coverage}_i}{\text{AUC}} * \text{target} = \text{scaled read counts} \quad (2)$$





The expression data in recount2 can be used to identify potential (a) new isoforms by identifying expressed regions and using the exon-exon junction data, (b) retained introns, or (c) update exon boundaries. We encourage others to develop new methods using this data.

EXAMPLE ANALYSIS

Download a dataset of interest

```
library("recount")

## Find the project ID by searching abstracts of studies
abstract_search("human brain development by age")
```

```
##      number_samples species
## 1296             72   human
##                                                                  abstract
## 1296 RNAseq data of 36 samples across human brain development by age group from LIBD
##        project
## 1296 SRP045638
```

```
## Download the data if it is not there
if(!file.exists(file.path("SRP045638", "rse_gene.Rdata"))) {
    download_study("SRP045638", type = "rse-gene")
}
```

```
## 2017-10-17 08:02:34 downloading file rse_gene.Rdata to SRP045638
```

```
## Check that the file was downloaded
file.exists(file.path("SRP045638", "rse_gene.Rdata"))
```

```
## [1] TRUE
```

```
## Load the data
load(file.path("SRP045638", "rse_gene.Rdata"))
```

Add phenotype predictions* and metadata from SRA

* https://www.biorxiv.org/content/early/2017/06/03/145656

```
## Add the predictions
rse_gene <- add_predictions(rse_gene)
```

```
## Append the variables of interest
colData(rse_gene) <- cbind(colData(rse_gene), sra[, sra_vars])

## Final dimensions
dim(colData(rse_gene))
```

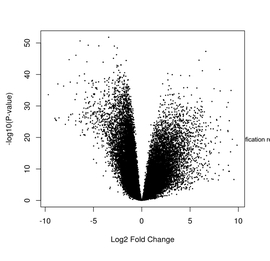
```
## [1] 72 40
```

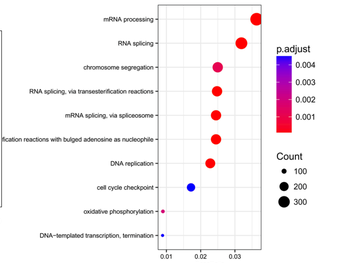
```
## Explore result
colData(rse_gene)[, 34:ncol(colData(rse_gene))]
```

```
## DataFrame with 72 rows and 7 columns
##                sex     race       RIN       age  isolate  disease   tissue
##           <factor> <factor> <numeric> <numeric> <factor> <factor> <factor>
## SRR2071341   female       AA       8.3   67.7800    DLPFC  Control    DLPFC
## SRR2071345     male       AA       8.4   40.4200    DLPFC  Control    DLPFC
## SRR2071346     male       AA       8.7   41.5800    R2869  control    DLPFC
## SRR2071347   female       AA       5.3   44.1700    R3098  control    DLPFC
## SRR2071348   female       AA       9.6   -0.3836    R3452  control    DLPFC
## ...             ...      ...       ...       ...      ...      ...      ...
## SRR1554541     male       AA       5.7   -0.3836    R3485  control    DLPFC
## SRR1554554   female       AA       8.1    0.3041    R3669  control    DLPFC
## SRR1554535     male       AA       8.7   41.5800    R2869  control    DLPFC
## SRR1554558   female     CAUC       9.1   16.7000    R4028  control    DLPFC
## SRR1554553     male     CAUC       8.4    0.3918    R3652  control    DLPFC
```

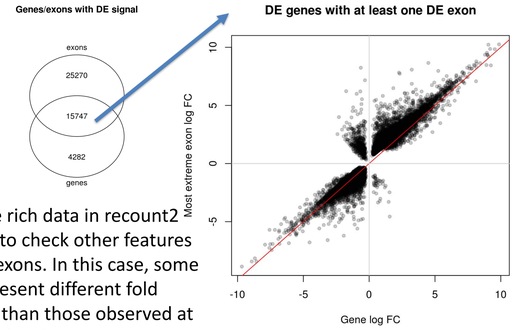
Identify differentially expressed genes, explore results and perform gene ontology enrichment analyses.

```
## Specify our design matrix
design <- with(colData(rse_gene_scaled),
    model.matrix(~ sex + RIN + prenatal))
```



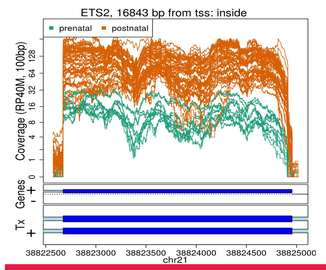


BEYOND GENES



With the rich data in recount2 it's easy to check other features such as exons. In this case, some exons present different fold changes than those observed at the gene level.

```
## Visualize DER #2
plotRegionCoverage(regions = regions_resized, regionCoverage = regionCov,
    groupInfo = colData(rse_er_scaled)$prenatal,
    nearestAnnotation = nearest_ann,
    annotatedRegions = regions_ann,
    txdb = gencode_v25_hg38_txdb,
    scalefac = 1, ylab = "Coverage (RP40M, 100bp)",
    ask = FALSE, verbose = FALSE, whichRegions = 2)
```



We can also identify expressed regions using derfinder and then determine if any of them are differentially expressed. The differentially expressed regions might match known exons such as the one shown here, but do not necessarily overlap annotated features.

SUMMARY

We described in detail the available data in recount2, how the coverage count matrices were computed, the metadata included in recount2 and how to get new phenotypic information from other sources. We showed how to perform a DE analysis at the gene and exon levels as well as use an annotation-agnostic approach. Finally, we explained how to visualize the base-pair information for a given set of regions. This work constitutes a strong basis to leverage the recount2 data for human RNA-seq analyses.

ACKNOWLEDGEMENTS

LCT and AEJ were supported by the National Institutes of Health (grant R21 MH109956-01). LCT and AN were supported by the National Institutes of Health (grant R01 GM105705).

We would like to acknowledge the members of Andrew Jaffe (Lieber Institute for Brain Development, Johns Hopkins Medical Campus) and Alexis Battle (Department of Computer Science, Whiting School of Engineering at Johns Hopkins University) labs for feedback on the explanatory figures.

recount2 is hosted on SciServer, a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by, the Institute for Data Intensive Engineering and Science (IDIES) at Johns Hopkins University. SciServer is funded by the National Science Foundation Award ACI-1261715.