

### 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

<u>http://bioconductor.org/packages/recount</u>. 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



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.



# Getting started with recount2 and accessing it via R

Collado-Torres L<sup>1,2,\*</sup>, Nellore A<sup>3,4,5</sup>, Jaffe AE<sup>1,2,6,7</sup> <sup>1</sup>Lieber Institute for Brain Development, <sup>2</sup>CCB JHU, <sup>3</sup>BME <sup>4</sup>Surgery and <sup>5</sup>CBP OHSU, <sup>6</sup>Biostatistics and <sup>7</sup>Mental Health JHU, \*leo.collado@libd.org

## **COUNTS PROVIDED**



 $\sum_{i=1}^{n} \text{coverage}_{i}$ \* <u>target</u> = scaled read counts Read Length mappe

 $\frac{\sum_{i}^{n} \text{coverage}_{i}}{\text{AUC}}$  $\frac{r_{l}}{r}$  \* target = scaled read counts (2)





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.



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 (1) area under the coverage as shown in equation (2) to scale the counts to a common library size.



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.

	D	ownlo	ad a d	lataset	of inte	erest	
ry("recou	nt")						
nd the pro act_searc	oject ID by h("human br	v searchin ain devel	g abstra opment b	cts of stud y age")	dies		
number_s	amples spec 72 hi	cies uman					abetrae
RNAseq d projec SRP04563	lata of 36 s t 88	samples ac	ross hun	nan brain d	evelopme	nt by age	group from LIB
wnload the ile.exist ownload_s	e data if i s(file.path tudy("SRP04	t is not n("SRP0456 5638", ty	there 38", "rs pe = "rs	e_gene.Rdat e-gene")	ta"))) {		
17-10-17	08:02:34 dc	ownloading	file rs	e_gene.Rdat	ta to SRF	P045638	
eck that exists(fi	the file wa le.path("SR	s downloa P045638",	ded "rse_ge	ne.Rdata"))	)		
] TRUE							
ad the da file.path	ta ("SRP045638	8", "rse_g	ene.Rdat	a"))			
pheno tps://w	otype pre <u>vww.bio</u> r	ediction <u>xiv.org</u>	s* anc <mark>/conte</mark>	l metada nt/early	ata froi /2017/	m SRA / <u>06/03</u> /	<u>/145656</u>
d the prec ene <- ado	dictions d_predictio	ns(rse_ge	ne)				
a(rse_ger al dimens	ie) <- cbin sions e_gene))	d(colData)	(rse_gene	e), sra[, s	ra_vars]	)	
72 40							
olore resu a(rse_ger	ult ne)[, 34:nc	ol(colData	a(rse_ger	ne))]			
aEramo wi	ith 72 rows	and 7 co					
arraile wi	sex sex <factor> <f< td=""><td>race actor&gt; <nu< td=""><td>RIN meric&gt;</td><td>age</td><td>isolate</td><td>disease <factor></factor></td><td>tissue <factor></factor></td></nu<></td></f<></factor>	race actor> <nu< td=""><td>RIN meric&gt;</td><td>age</td><td>isolate</td><td>disease <factor></factor></td><td>tissue <factor></factor></td></nu<>	RIN meric>	age	isolate	disease <factor></factor>	tissue <factor></factor>
2071341	female		8.3 8.4	67.7800 40.4200		Control	DLPFC
2071346	male female	AA AA	8.7 5.3	41.5800	R2869 R3098	control control	DLPFC
2071348	female 	AA 	9.6	-0.3836	R3452	control	DLPFC
1554541 1554554	male female	AA AA	5.7 8.1	-0.3836 0.3041	R3485 R3669	control control	DLPFC DLPFC
1554535 1554558	male female	AA CAUC	8.7 9.1	41.5800 16.7000	R2869 R4028	control control	DLPFC DLPFC
1554553	male	CAUC	8.4	0.3918	R3652	control	DLPFC
tify dif	ferential	ly expre	essed g	genes, ex	plore	results	and perforr
e ontolo	ogy enric	chment	analys	ses.			
ecify our 1 <- with odel.matr	design mat (colData(rs ix(~ sex +	rix e_gene_sca RIN + pre	aled), natal))				
					A processing		
· · · ·	•. • •			obromese			p.adjust
		• • •	RNA sp	licing, via transesterificat	ion reactions -		0.003
				mRNA splicing, via	spliceosome -		0.002
			fication reactions	with bulged adenosine as	s nucleophile -		
				DN	IA replication -		Count
							● 100
				cell cycl	le checkpoint -	•	• 100 • 200
-5				cell cycl oxidative pho	le checkpoint -	•	<ul> <li>100</li> <li>200</li> <li>300</li> </ul>
	Id the project RNAseq of project RNAseq of project RNAseq of project RNAseq of project RNAseq of project SRP04563 vnload the ile.exist wnload_s L7-10-17 eck that exists(find) L7-10-17 eck that exists(find) TRUE ad the da file.path phencet tors://w l the predent aframe with 2071341 2071341 2071345 2071346 2071346 2071347 2071348 1554554 1554554 1554553 tify dif contolous ecify our secify our se	Id the project ID by act_search("human br number_samples sper 72 hr RNAseq data of 36 s project SRP045638 vnload the data if i ile.exists(file.path wnload_study("SRP04 17-10-17 08:02:34 do eck that the file wa exists(file.path("SR 17-10-17 08:02:34 do eck that the file wa exists(file.path("SR 10-10-17 08:02:34 do eck that the file wa exists(file.path("SR 10-17 08:02:34 do eck that the file wa exists(file.path("SR 10-10-17 08:02:34 do eck that the file wa exists(file.path("SR 10-17 08:02:34 do eck that the file wa exists(file.path("	number_samples species 72 human RNAseq data of 36 samples ac project sRP045638 vnload the data if it is not ile.exists(file.path("SRP045638", ty 17-10-17 08:02:34 downloading eck that the file was downloa exists(file.path("SRP045638", ty 17-10-17 08:02:34 downloading eck that the file was downloa exists(file.path("SRP045638", "rse_g phenotype prediction tps://www.biorxiv.org, 1 the predictions ene <- add_predictions(rse_gen end the variables of interest a(rse_gene) <- cbind(colData( al dimensions lData(rse_gene)) 72 40 1 ore result a(rse_gene)[, 34:ncol(colData( al dimensions lData(rse_gene)) 72 40 1 ore result a(rse_gene)[, 34:ncol(colData( al dimensions lData(rse_gene)) 72 40 1 ore result a(rse_gene)[, 34:ncol(colData( al fine AA 2071345 male AA 2071346 male AA 2071347 female AA 2071348 female AA 2071347 female AA 2071348 female AA 2071349 male AA 2071349 male AA 2071340 female AA 2071340 female AA 2071340 female AA 2071341 female AA 2071342 female AA 2071343 male AA 2071343 male AA 2071344 female AA 2071345 male AA 2071345 male AA 2071345 male AA 2071346 female AA 2071347 female AA 2071347 female AA 2071348 female AA 2071348 female AA 2071349 female AA 2071340 female AA 2071340 female AA 2071340 female AA 2071340 female AA 2071340 female AA 2071341 female AA 2071341 female AA 2071345 female AA 2071346 female AA 2071346 female AA 2071347 female AA 2071347 female AA 2071348 female AA 2071348 female AA 2071348 female AA 2071349 female AA 2071340 female AA 2	<pre>d the project ID by searching abstra act_search("human brain development b</pre>	<pre>d the project 10 by searching abstracts of stur act_search("human brain development by age") number_samples species 72 human RNAseq data of 36 samples across human brain d project sRP045638 wnload the data if it is not there [le.exists(file.path("SRP045638", "rse_gene.Rdat wnload_study("SRP045638", type = "rse-gene") //-10-17 08:02:34 downloading file rse_gene.Rdata")) //-10-17 08:02:34 downloaded exists(file.path("SRP045638", "rse_gene.Rdata")) // TRUE // TRU</pre>	d the project ID by searching abstracts or studies ict_search("human brain development by age") number_samples species 72 human RNAseq data of 36 samples across human brain developme project srP045638 mload the data if it is not there ile.exists(file.path("srP045638", "rse_gene.Rdata"))) { wmload_study("SRP045638", type = "rse-gene") iz-10-17 08:02:34 downloading file rse_gene.Rdata to SRI sck that the file was downloaded exists(file.path("SRP045638", "rse_gene.Rdata")) I TRUE id the data file.path("SRP045638", "rse_gene.Rdata")) phenotype predictions* and metadata from tps://www.biorxiv.org/content/early/2017/ I the predictions ine <- add_predictions(rse_gene) end the variables of interest a(rse_gene) <- cbind(colData(rse_gene), sra[, sra_vars] al dimensions lbata(rse_gene)) 72 40 iore result a(rse_gene)[, 34:ncol(colData(rse_gene))] 72 40 iore result a(rse_gene)[, 34:ncol(colData(rse_gene)]] aframe with 72 rows and 7 columns sex race RIN age isolate c <factory cfactory="" cfactory<br="">2071346 male AA 8.7 41.5800 R3689 2071347 female AA 9.6 -0.3836 R3452 1554541 male AA 5.7 -0.3836 R3452 1554551 male AA 8.7 41.5800 R3689 2051346 female AA 9.6 -0.3836 R3452 1554551 male AA 8.7 41.5800 R3689 2051346 female AA 9.6 -0.3836 R3452 1554551 male AA 0.7 0.8368 R3452</factory>	<pre>id the project ID by searching abstracts of studies (cc_search("human brain development by age") number_samples species 72 human RNAseq data of 36 samples across human brain development by age project skp045638 mload the data if it is not there [le.exists(file.path("skp045638", "rse_gene.Rdata"))) { wnload_study("skp045638", type = "rse-gene.Rdata")) { /7-10-17 08:02:34 downloading file rse_gene.Rdata to Skp045638 ck that the file was downloaded wists(file.path("skp045638", "rse_gene.Rdata")) i tRue d the data "ile.path("skp045638", "rse_gene.Rdata")) phenotype predictions* and metadata from SRA tps://www.biorxiv.org/content/early/2017/06/03/ i the predictions me &lt;- add_predictions(rse_gene) end the variables of interest a(rse_gene) &lt;- cbind(colData(rse_gene), sra[, sra_vars]) al dimensions lotat(rse_gene)] 72 40 lore result a(rse_gene)], 34:ncol(colData(rse_gene))] aframe with 72 rows and 7 columns</pre>

### **BEYOND GENES**



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 <u>SciServer</u>, a collaborative research environment for largescale 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.

