

1 The Lair: A resource for exploratory analysis of published 2 RNA-Seq data

3
4 Harold Pimentel¹, Pascal Sturmfels², Nicolas Bray³, Páll Melsted⁴ and
5 Lior Pachter^{*1,5}
6

- 7 1. Department of Computer Science, UC Berkeley
8 2. Department of Computer Science, University of Michigan
9 3. Innovative Genomics Initiative, UC Berkeley
10 4. Faculty of Industrial Engineering, Mechanical Engineering and Computer Science,
11 University of Iceland
12 5. Departments of Mathematics and Molecular & Cell Biology, UC Berkeley

13
14 * Corresponding author: lpachter@math.berkeley.edu
15

16 **Abstract**

17 Increased emphasis on reproducibility of published research in the last few years
18 has led to the large-scale archiving of sequencing data. While this data can, in
19 theory, be used to reproduce results in papers, it is typically not easily usable in
20 practice. We introduce a series of tools for processing and analyzing RNA-Seq
21 data in the Short Read Archive, that together have allowed us to build an easily
22 extendable resource for analysis of data underlying published papers. Our
23 system makes the exploration of data easily accessible and usable without
24 technical expertise. Our database and associated tools can be accessed at The
25 Lair: <http://pachterlab.github.io/lair>

26

27 **Background**

28 The Short Read Archive (SRA) is a public repository for sequencing data that has
29 become an important archival resource for reads associated with published
30 papers. The accumulation of large amounts of data in the SRA allows for meta-

31 analyses that are possible only thanks to the centralized, open sharing of data by
32 multiple investigators [1,2]. Reads in the SRA also allow, in principle, for the
33 reproduction of results in publications [3]. However the bioinformatics difficulties
34 associated with processing and analyzing sequencing data [4] have limited the
35 utility of the SRA and have made it prohibitive for most investigators to perform
36 exploratory data analysis (EDA) on the data in the archive. The use of the SRA
37 for EDA is especially difficult for RNA-Seq data. This is because even the most
38 basic processing of RNA-Seq reads requires numerous decisions about
39 appropriate software to use, complex choices about annotations, understanding
40 of experimental design, and frequently, significant computational resources. As a
41 result, workflows designed for operating on multiple datasets in the short read
42 archive have mainly been restricted to the tasks of aligning and quantifying reads
43 [5]. Similarly, the Gene Expression Omnibus (GEO) requires an expression
44 matrix to be uploaded with every sample, however these are static and frequently
45 out of date, and fail to provide users with the complete analyses that they
46 typically seek to explore.

47

48 We have recently developed a pair of tools called kallisto [6] and sleuth for RNA-
49 Seq analysis that address a number of the challenges associated with
50 processing RNA-Seq data. The kallisto program circumvents the need for large
51 alignment files, a convenience that reduces storage needs and increases speed,
52 thus enabling the processing of large numbers of samples on modest
53 computational resources. The program sleuth utilizes bootstraps output by

54 kallisto for differential analysis, and provides a complete interactive and web-
55 compatible solution for exploring and analyzing RNA-Seq data. This has allowed
56 us to semi-automate the process of associating interactive Shiny-based [7]
57 websites for EDA of RNA-Seq data.

58

59 The processing underlying the resource we have developed can easily be rerun,
60 providing a scalable and updatable push-button system for the analysis of large
61 numbers of datasets. Thus, we have been able to create a semi-automated
62 system for analysis of archived RNA-Seq data that is much more informative
63 than mere alignment and quantification and that opens up the analyses of
64 published data.

65

66 **Construction and content**

67 The infrastructure for The Lair is based on Snakemake [8], a python-based
68 workflow system. The system is organized as shown in Figure 1 and consists of
69 three main parts: (1) an initial processing of data to produce sleuth objects that
70 are deployed in a Shiny database; (2) a Shiny server and database; and (3) a
71 website that can be constructed automatically and that links to the Shiny server.

72

73 The specification for datasets is stored on GitHub at
74 https://github.com/pachterlab/bears_analyses. Each dataset requires a
75 config.json file that provides information about the dataset that will be used

76 during its processing and deployment. For example, the config.json file for a
77 recent paper on a HOXA1 knockdown transcriptome survey in human [9] is:

78

```
79 {  
80     "species": "homo_sapiens",  
81     "use_paired_end": true,  
82     "directory": "Trapnell_2013_10.1038_nbt.2450",  
83     "design_file" : "Trapnell_2013_10.1038_nbt.2450/SraRunTable.txt",  
84     "full_model": "~transfection_s",  
85     "reduced_model": "~1",  
86     "DOI": "10.1038/nbt.2450",  
87     "kmer-size": 21,  
88     "bootstrap_samples": 30,  
89     "bias": true,  
90     "analysis": "http://lair.berkeley.edu/trapnell/"  
91 }  
92
```

93 The config.json file specifies the species name for the RNA-Seq analysis, thus
94 allowing for automatic downloading and indexing of the appropriate Ensembl [10]
95 transcriptome for the analysis, the type of reads (single or paired), the design
96 matrix and type of testing to perform, as well as the DOI of the paper and
97 parameters for the kallisto processing. Along with the config.json file, a design
98 matrix must be included which specifies the structure underlying the samples.
99 The design matrix can be downloaded from the SRA but sometimes requires
100 manual curation due to inconsistencies in SRA formatting. The entire workflow
101 begins with only these two files, from which all the relevant information is
102 extracted for processing.

103

104 The organization of The Lair allows for updating of all the analyses at the push of
105 a button. This is useful in the case of updates to the component programs and

106 the emphasis on speed of the constituent software allows for frequent updating.
107 In fact, the current main bottleneck with The Bair's Lair is downloading of the
108 SRA data after which the entire workflow processes individual samples in
109 minutes [6].

110

111 The website is a static page built using Jekyll [11] and Bootstrap v3.3.6 [12]. The
112 Analyses section of the website contains a dynamically generated table of paper
113 with corresponding live analyses. The table is powered by the JQuery plug-in
114 DataTables [13] and features filtering and sorting by Authors, Title, Journal and
115 Date. The title of each paper links to the original paper published in the stated
116 journal. If the paper was published in print, the date given is the paper's print
117 date; otherwise, the date is the paper's online publication date. The analysis
118 button links to an in-browser analysis of the experiment, made possible by
119 sleuth's efficient use of statistical bootstrapping and RStudio's Shiny plug-in; the
120 analysis for each paper is generated automatically by the above build system.

121

122 The table is populated automatically by data from the bears analyses
123 (https://github.com/pachterlab/bears_analyses) Github repository. This means
124 that anyone can submit additional datasets for processing via Git pull requests
125 which once accepted will become part of the website.

126

127 **Utility and discussion**

128 To demonstrate the utility of The Bair's Lair we examined the results from our
129 analysis of the Trapnell *et al.* [9] data. In that paper, an RNA-Seq differential
130 analysis was performed on lung fibroblasts responding to the knockout of the
131 developmental transcription factor HOXA1. First, it is easy to confirm that The
132 Lair analysis replicates the main results of the paper. Figure 2 shows a principal
133 component analysis of the data, confirming high quality data with substantial
134 separation between the two conditions.

135

136 Specific results about individual genes that are discussed in the Trapnell *et al.*
137 paper are easily confirmed. For example, Figure 5 shows transcript abundance
138 changes in response to the knockdown, providing examples of some key genes
139 of interest. The T-box DNA binding domain TBX3 displays an increase in exactly
140 one out of three isoforms (Figure 5d). The differential isoform,
141 ENST00000349155, is displayed via the "transcript view" feature in the Shiny app
142 as shown in Figure 3. The associated q-value in the sleuth test is $q = 3.15e-05$.
143 The two remaining isoforms of the gene are not significantly differential, in
144 concordance with the Trapnell *et al.* paper (for ENST00000257566, $q = 0.148$
145 and in the case of ENST00000613550, the isoform did not pass the requisite
146 filters to be tested).

147

148 While the reproducibility of results is reassuring, the innovation in The Lair
149 resource is the ability to go beyond the limited view of the data provided by the
150 authors. In the Trapnell *et al.* example there are thousands of significantly

151 differential transcripts, and The Lair allows for viewing the raw quantifications
152 underlying each of them. Another advantage is the ability to examine results of
153 different papers analyzed with the same framework. For example, the Ng *et al.*
154 [14] paper is immediately established to have much higher variance in the
155 estimates due to having fewer replicates in each condition (two instead of three),
156 and the common framework underlying its analysis provides a quantification of
157 that assessment.

158

159 **Conclusion**

160 RNA-Seq technology provides rich and complex data for analysis in projects
161 where expression dynamics are of interest. While investigators are eager to
162 squeeze every bit of information out of their data, there are a number of reasons
163 why they are unlikely to be able to do so at the time of publication of their work:
164 analysis methods and tools improve over time and data may be revealed to be
165 useful for applications not considered at the time of acquisition.

166

167 The Lair resource we have developed opens up large volumes of RNA-Seq data
168 for both general and targeted exploration. The modular and automated
169 construction of our system will allow us to upgrade it over time, adding
170 functionality and analyses as we improve and expand the kallisto and sleuth
171 methods and programs. An added benefit of our holistic analysis of SRA data is
172 that our use of the same tools to process diverse datasets also allows for

173 comparison of results across studies. Future plans for The Lair include facilitating
174 such cross-study comparisons.

175

176 While we have focused The Lair on RNA-Seq data, the ideas and tools
177 developed in this work should be adaptable to other data types. Hopefully such
178 work will establish tools that create symbiotic rather than parasitic relationships
179 between data generators and data analyzers.

180

181

182 **Methods**

183 **Data handling**

184 There are two data bottlenecks in processing of short read archive RNA-Seq
185 data: first, the downloading and storing of large read files, an issue we do not
186 address in this paper but that can be ameliorated with compression schemes.
187 Second, sleuth utilizes statistical bootstraps generated by kallisto as part of its
188 differential analysis and these can be time consuming to transmit via the web. To
189 make sleuth usable via The Lair, the code was refactored to pre-compute
190 variance and quantile data that are sufficient and necessary statistics to generate
191 the plots in the online visualization. The individual bootstraps estimates can then
192 be discarded. This reduced the size of the analysis objects by orders of
193 magnitude and allowed sleuth analyses to be shared online and loaded in
194 standard web browsers.

195

196 **The Snakemake workflow**

197 The Snakefile used to generate the analysis requires two input parameters: a
198 json configuration file and a design matrix file. The configuration file has the
199 following required parameters:

200

- 201 • species: species used in the experiment.
- 202 • used_paired_end: true if the experiment was paired-end, false otherwise.
- 203 • directory: the directory to put the results into.
- 204 • design_file: the name of the required design matrix file.
- 205 • full_model: formula which describes the full or alternative model used in
206 differential analysis.
- 207 • reduced_model: formula which describes the reduced or null model
- 208 • DOI: the digital object identifier of the publication.

209

210 The configuration file also accepts the following optional parameters:

- 211 • kmer-size: the k-mer length used to build the kallisto index (defaults to 31)
- 212 • bias: perform sequence-specific bias correction during quantification
213 (defaults to True)

214

215 The design file must be a .tsv file, one column of which is titled 'run' or 'Run_s'
216 and contains the SRR accessions for each run in the experiment. Furthermore,
217 the full_model and reduced_model in the config file must match column names in
218 the design file for the differential analysis to work correctly.

219

220 The build system checks whether the FASTA reference file for the input species
221 is already locally accessible. If not, it downloads it to the FASTA file from
222 Ensembl. It then uses this FASTA file to build a kallisto index given the input k-
223 mer size if the index does not already exist, and makes this index locally
224 accessible.

225

226 Then the build system uses the 'run' or 'Run_s' column of the design matrix to
227 download the raw SRA data and quantifies the raw data using kallisto and the
228 index for the specified species. Once kallisto finishes quantification, sleuth is run
229 with the likelihood ratio test using the specified full_model and reduced_model
230 parameters. The build system then deploys the resulting sleuth analysis onto the
231 server, where it is available to explore online.

232

233 **Availability of data and materials**

234 The main user website is at <http://pachterlab.github.io/lair>. The workflow software
235 is at https://github.com/pachterlab/bears_analyses and the website code is at
236 <https://github.com/pachterlab/lair>.

237

238 **Acknowledgements**

239 We thank the members of the Pachter Lab for contributing design and feature
240 ideas and for suggesting the initial datasets with which to prototype The Lair. HP

241 and LP were partially supported by NIH grants R01 HG006129, R01 DK094699
242 and R01 HG008164.

243

244 **References**

245 1. Frazee AC, Langmead B, Leek JT. ReCount: A multi-experiment resource of
246 analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*. 2011;12:449.

247 2. Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Pritt J, Morton J,
248 et al. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *bioRxiv*.
249 2015;019067.

250 3. Stodden V, Leisch F, Peng RD. *Implementing Reproducible Research*. CRC
251 Press; 2014.

252 4. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson
253 A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*.
254 2016;17:13.

255 5. Bhuvaneshwar K, Sulakhe D, Gauba R, Rodriguez A, Madduri R, Dave U, et al.
256 A case study for cloud based high throughput analysis of NGS data using the
257 globus genomics system. *Comput. Struct. Biotechnol. J*. 2015;13:64–74.

258 6. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-
259 seq quantification. *Nat. Biotechnol*. 2016;34:525–7.

260 7. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, RStudio, et al. shiny: Web
261 Application Framework for R [Internet]. 2016 [cited 2016 May 25]. Available from:
262 <https://cran.r-project.org/web/packages/shiny/index.html>

263 8. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine.
264 *Bioinformatics*. 2012;28:2520–2.

265 9. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L.
266 Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat*.
267 *Biotechnol*. 2013;31:46–53.

268 10. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014.
269 *Nucleic Acids Res*. 2014;42:D749–55.

270 11. Preston-Werner T. Jekyll [Internet]. 2016. Available from: <https://jekyllrb.com/>

271 12. Otto M, Thornton J, Rebert C, Thilo J, XhmikosR, Fenkart H, et al. Bootstrap
272 [Internet]. 2016. Available from: <http://getbootstrap.com/>

273 13. SpryMedia. DataTables [Internet]. 2016. Available from:
274 <https://datatables.net/>

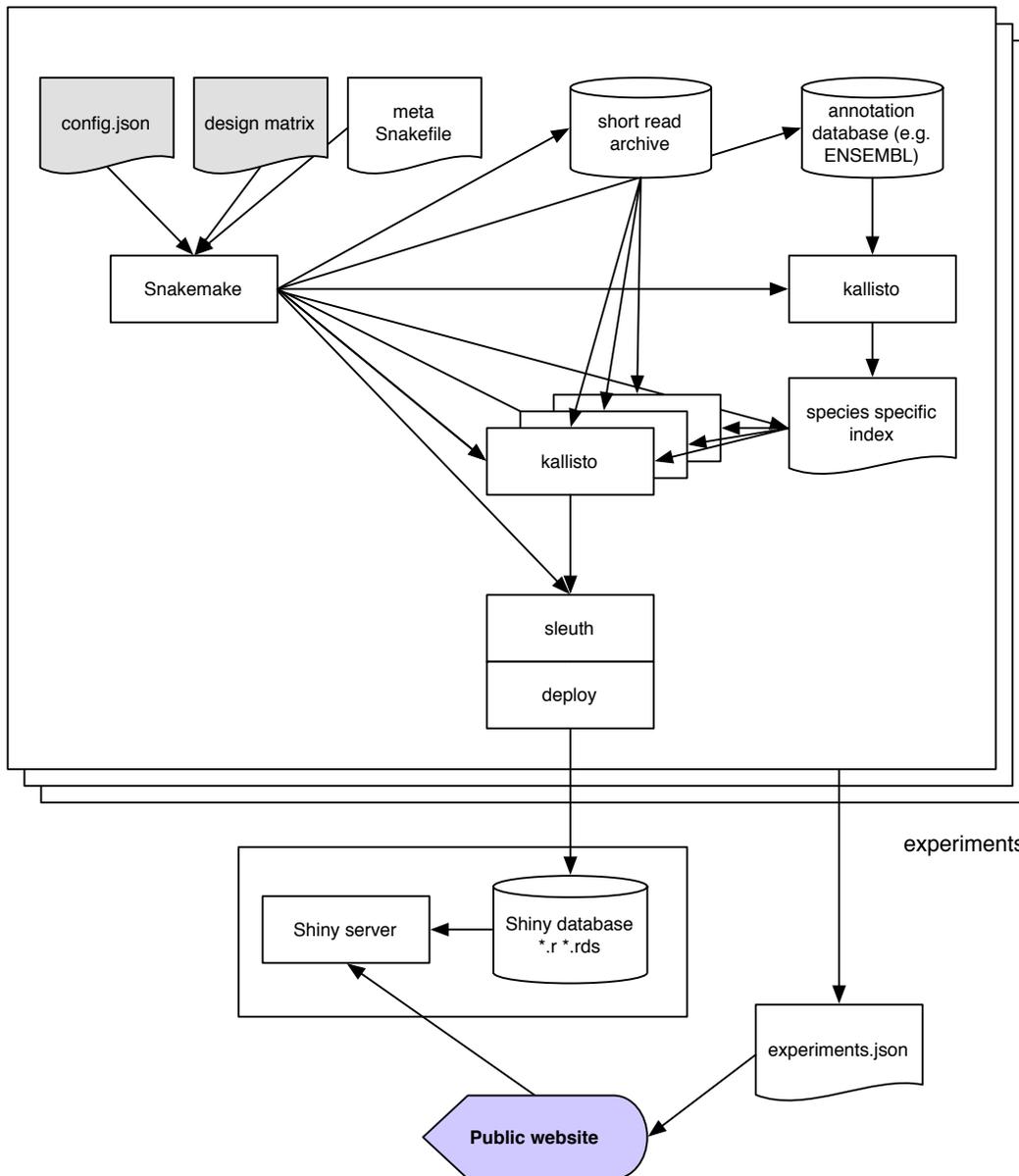
275 14. Ng S-Y, Soh BS, Rodriguez-Muela N, Hendrickson DG, Price F, Rinn JL, et
276 al. Genome-wide RNA-Seq of Human Motor Neurons Implicates Selective ER
277 Stress Activation in Spinal Muscular Atrophy. *Cell Stem Cell*. 2015;17:569–84.

278

279

280

281



282

283 Figure 1: Workflow of The Lair system for distributing analysis of short read
284 archive data. The inputs to the system are sets of two files: `config.json` file that
285 specifies parameters to be used during the processing of each experiment and a
286 design matrix for each experiment that specifies its structure. A master
287 Snakemake workflow organizes a series of computations starting with
288 downloading of data to the short read archive and ending with deployment of a

289 sleuth analyses to a Shiny server. Finally, a website generated from information
290 in the config.json files links to objects in the Shiny server thus providing access to
291 the processed experiments.

292

293

294

295

296

297

298

299

300

301

302

303

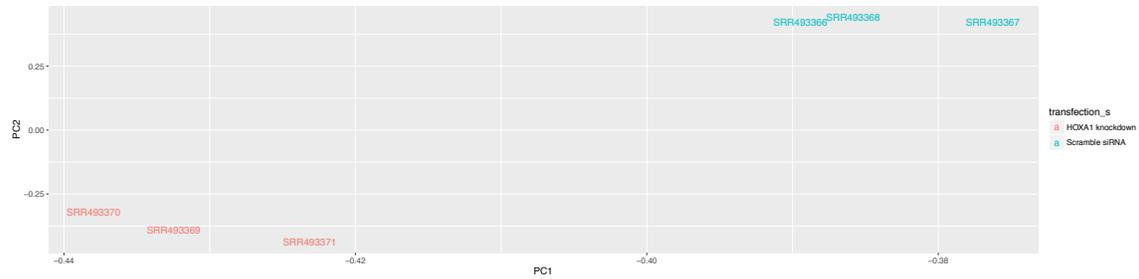
304

305

306

307

308



311 Figure 2: Principal Components Analysis of the Trapnell *et al.* HOXA1
312 knockdown RNA-Seq data. The Lair allows for plotting projections with respect to
313 any pair of principal components, and also identifies the transcripts constituting
314 the loadings of each dimension.

315

316

317

318

319

320

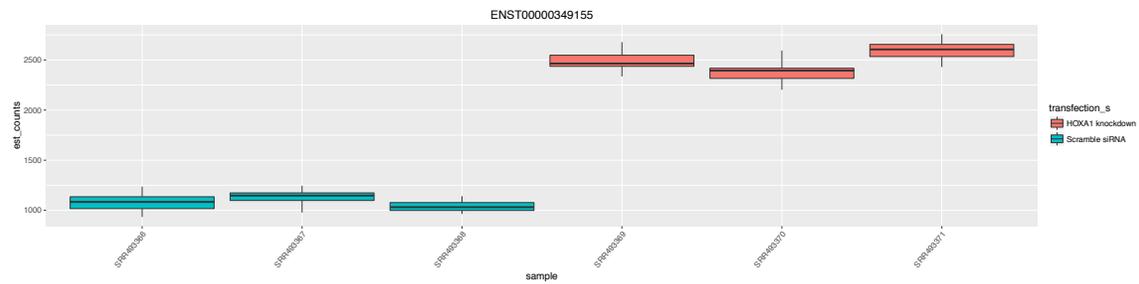
321

322

323

324

325



326

327

328 Figure 3: Transcript abundances for the differential isoform of the TBX3 gene in
329 the Trapnell *et al.* data. The error bars on each quantification are produced via
330 the bootstrap feature of kallisto, which establishes the inferential variance
331 associated with quantification. The Lair provides an interactive template for
332 viewing such plots for any transcript.