Title: Comprehensive analysis of RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues

Authors: Serghei Mangul[#1,2], Harry Taegyun Yang[1], Nicolas Strauli[3], Franziska Gruhl[4,5], Hagit T. Porath[6], Kevin Hsieh[1], Linus Chen[7], Timothy Daley[8], Stephanie Christenson[9], Agata Wesolowska-Andersen[10], Roberto Spreafico[2], Cydney Rios[10], Celeste Eng[11], Andrew D. Smith[8], Ryan D. Hernandez[12,13,14], Roel A. Ophoff[15,16,17], Jose Rodriguez Santana[18], Erez Y. Levanon[6], Prescott G. Woodruff[19], Esteban Burchard[9,10], Max A. Seibold*[8,19,20], Sagiv Shifman*[21], Eleazar Eskin*[1,16], Noah Zaitlen*[#9]

Affiliations:

[1]Department of Computer Science, University of California Los Angeles, Los Angeles, USA

[2]Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles, USA

[3]Biomedical Sciences Graduate Program, University of California, San Francisco, CA, USA

[4]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

[5]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[6]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

[7]Department of Bioengineering, University of California Los Angeles, Los Angeles, USA

[8]Department of Molecular and Computational Biology, University of Southern California, CA, USA

23 [9]Division of Pulmonary, Critical Care, Sleep and Allergy, Department of Medicine, and
24 Cardiovascular Research Institute, University of California, San Francisco, CA, USA
25 [10]Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA
26 [11]Department of Medicine, University of California, San Francisco, CA, USA
27 [12]Department of Bioengineering and Therapeutic Sciences, University of California, San
28 Francisco, CA, USA
29 [13]Institute for Quantitative Biosciences, University of California, San Francisco, CA, USA
30 [14]Institute for Human Genetics, University of California San Francisco, San Francisco, CA,
31 USA
32 [15]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human
33 Behavior, University California Los Angeles, Los Angeles, USA
34 [16]Department of Human Genetics, University of California Los Angeles, Los Angeles, USA
35 [17]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center
36 Utrecht, Utrecht, The Netherlands
37 [18]Pediatric Pulmonology, San Juan, Puerto Rico
38 [19]Department of Pediatrics, National Jewish Health, Denver, CO, USA
39 [20]University of Colorado School of Medicine, Denver CO, USA
40 [21]Department of Genetics, The Institute of Life Sciences, The Hebrew University of
41 Jerusalem, Jerusalem, Israel
42
43

44 # Correspondence to: Serghei Mangul smangul@ucla.edu and Noah Zaitlen
45 Noah.Zaitlen@ucsf.edu
46 *Equal contribution
47

48 **Abstract**

49 High throughput RNA sequencing technologies have provided invaluable research
50 opportunities across distinct scientific domains by producing quantitative readouts of the
51 transcriptional activity of both entire cellular populations and single cells. The majority of
52 RNA-Seq analyses begin by mapping each experimentally produced sequence (i.e., read)
53 to a set of annotated reference sequences for the organism of interest. For both biological
54 and technical reasons, a significant fraction of reads remains unmapped. In this work, we
55 develop Read Origin Protocol (ROP) to discover the source of all reads originating from
56 complex RNA molecules, recombinant T and B cell receptors, and microbial communities.
57 We applied ROP to 8,641 samples across 630 individuals from 54 tissues. A fraction of
58 RNA-Seq data (n=86) was obtained in-house; the remaining data was obtained from the
59 Genotype-Tissue Expression (GTEx v6) project. To generalize the reported number of
60 accounted reads, we also performed ROP analysis on thousands of different, randomly
61 selected, and publicly available RNA-Seq samples in the Sequence Read Archive (SRA).
62 Our approach can account for 99.9% of 1 trillion reads of various read length across the
63 merged dataset (n=10641). Using in-house RNA-Seq data, we show that immune profiles
64 of asthmatic individuals are significantly different from the profiles of control individuals,
65 with decreased average per sample T and B cell receptor diversity. We also show that

66 immune diversity is inversely correlated with microbial load. Our results demonstrate the
67 potential of ROP to exploit unmapped reads in order to better understand the functional
68 mechanisms underlying connections between the immune system, microbiome, human
69 gene expression, and disease etiology. ROP is freely available at
70 https://github.com/smangul1/rop and currently supports human and mouse RNA-Seq
71 reads.
72
73
74 **INTRODUCTION**
75 Advances in RNA sequencing (RNA-seq) technology have provided an unprecedented
76 opportunity to explore gene expression across individual, tissues, and environments
77 (Cloonan et al., 2008; Sultan et al., 2008; Tang et al., 2009) by efficiently profiling the RNA
78 sequences present in a sample of interest (Z. Wang, Gerstein, & Snyder, 2009). RNA-seq
79 experiments currently produce tens of millions of short read subsequences sampled from
80 the complete set of RNA transcripts that are provided to the sequencing platform. An
81 increasing number of bioinformatic protocols are being developed to analyze reads in
82 order to annotate and quantify the sample's transcriptome (Mihaela Pertea, 2015;
83 Nicolae, Mangul, Mandoiu, & Zelikovsky, 2011; Trapnell et al., 2010). When a reference
84 genome sequence or, preferably, a transcriptome of the sample is available, mapping-
85 based RNA-seq analysis protocols align the RNA-seq reads to the reference sequences,
86 identify novel transcripts, and quantify the abundance of expressed transcripts.
87

88 Unmapped reads, the reads falling to map to the human reference, are a large
89 and often overlooked output of standard RNA-seq analyses. Even in carefully executed
90 experiments, the *unmapped reads* can comprise a substantial fraction of the complete set
91 of reads produced; for example, approximately 9%-20% of reads are unmapped in recent
92 large human RNA-seq projects (Ardlie et al., 2015; Li, Tighe, Nicolet, Grove, Levy,
93 Farmerie, Viale, Wright, Schweitzer, Gao, Kim, et al., 2014; Seqc/Maqc-Iii Consortium,
94 2014). Unmapped reads can arise due to technical sequencing artifacts that were
95 produced by low quality and error prone copies of the nascent RNA sequence being
96 sampled (Ozsolak & Milos, 2011). A recent study by Baruzzo et al., (2017) suggests that at
97 least 10% of the reads simulated from human references remain unmapped across 14
98 contemporary state-of-the art RNA alligners. This rate may be due to shortcomings of the
99 aligner's efficient yet heuristic algorithms (Siragusa, Weese, & Reinert, 2013). Reads can
100 also remain unmapped due to unknown transcripts (Grabherr et al., 2011), recombined
101 B and T cell receptor sequences (Blachly et al., 2015; N. B. Strauli & Hernandez, 2016), A-
102 to-G mismatches from A-to-I RNA editing (Porath, Carmi, & Levanon, 2014), trans-splicing
103 (Wu et al., 2014), gene fusion (X.-S. Wang et al., 2009), circular RNAs (Jeck & Sharpless,
104 2014), and the presence of non-host RNA sequences (Kostic et al., 2011) (e.g., bacterial,
105 fungal, and viral organisms).
106
107 In this work, we report the development of a comprehensive method that can
108 characterize the origin of unmapped reads obtained by RNA-seq experiments. Analyzing
109 unmapped reads can inform future development of read mapping methods, provide

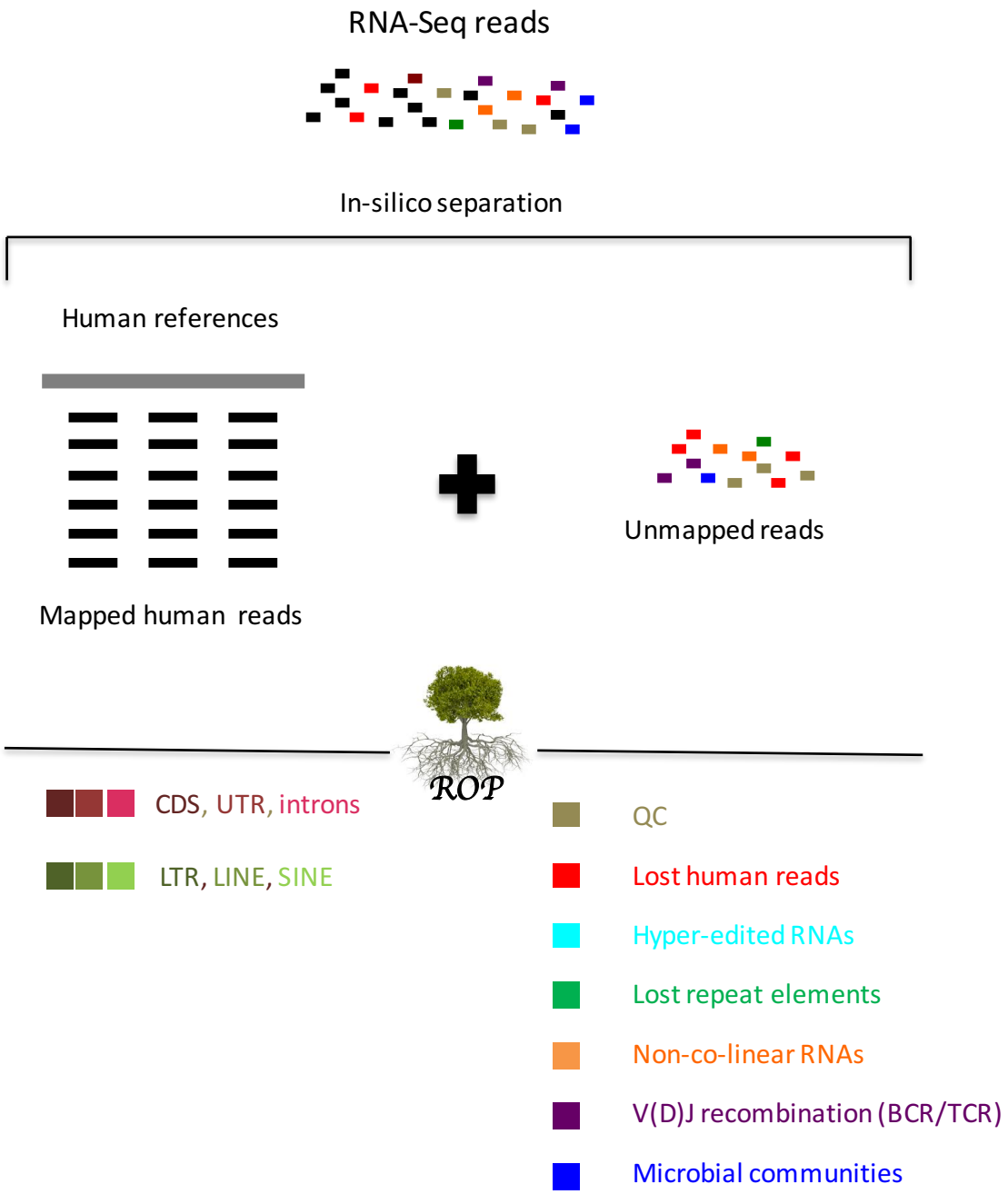
110 access to additional biological information, and resolve the irksome puzzle of the origin
111 of unmapped reads. We developed the Read Origin Protocol (ROP), a multi-step approach
112 that leverages accurate alignment methods for both host and microbial sequences. The
113 ROP tool contains a combination of novel algorithms and existing tools focused on specific
114 categories of *unmapped reads* (Blachly et al., 2015; Brown, Raeburn, & Holt, 2015; Chuang
115 et al., 2015; Kostic et al., 2011; N. Strauli & Hernandez, 2015). The comprehensive analytic
116 nature of the ROP tool prevents biases that can otherwise arise when using standard
117 targeted analyses. Currently, ROP supports human and mouse RNA-Seq data.
118
119
120 **RESULTS**
121 ***ROP – a computational protocol to explain unmapped reads in RNA-Sequencing***
122 Mapping-based RNA-seq analysis protocols overlook reads that fail to map onto the
123 human reference sequences (i.e., *unmapped reads*). We designed a read origin protocol
124 (ROP) that identifies the origin of both mapped and unmapped reads (Fig. 1). The protocol
125 first identifies human reads by mapping them onto a reference genome and
126 transcriptome using a standard high-throughput mapping algorithm (Kim et al., 2013). We
127 used tophat v. 2.0.12 with ENSEMBL GRCh37 transcriptome and hg19 build, but many
128 other mapping tools are available and have recently been reviewed by Baruzzo et al.,
129 2017) . After alignment, reads are grouped into genomic (e.g., CDS, UTRs, introns) and
130 repetitive (e.g., SINEs, LINEs, LTRs) categories. The rest of the ROP protocol characterizes
131 the remaining *unmapped reads*, which failed to map to the human reference sequences.

132

133 The ROP protocol effectively processes the *unmapped reads* in seven steps. First,
134 we apply a quality control step to exclude low-quality reads, low-complexity reads, and
135 reads that match rRNA repeat units among the unmapped reads (FASTQC (Andrews &
136 others, 2010), SEQCLEAN ("https://sourceforge.net/projects/seqclean/," n.d.)). Next, we
137 employ Megablast (Camacho et al., 2009), a more sensitive alignment method, to search
138 for human reads missed due to heuristics implemented for computational speed in
139 conventional aligners and reads with additional mismatches. These reads typically include
140 those with mismatches and short gaps relative to the reference set, but they can also
141 include perfectly matched reads. Hyper-editing pipelines recognize reads with excessive
142 ('hyper') editing, which are usually rejected by standard alignment methods due to many
143 A-to-G mismatches (Porath, Carmi, & Levanon, 2014). We use a database of repeat
144 sequences to identify lost repeat reads among the *unmapped reads*. Megablast, and
145 similar sensitive alignment methods, are not designed to identify 'non-co-linear' RNA
146 (Chuang et al., 2015) reads from circRNAs, gene fusions, and trans-splicing events, which
147 combine a sequence from distant elements. For this task, we independently map 20bp
148 read anchors onto the genome (see Supplemental Methods). Similarly, reads from BCR
149 and TCR loci, which are subject to recombination and somatic hyper-mutation (SHM),
150 require specifically designed methods. For this case, we use IgBlast (Ye, Ma, Madden, &
151 Ostell, 2013). The remaining reads that did not map to any known human sequence are
152 potentially microbial in origin. We use microbial genomes and phylogenetic marker genes
153 to identify microbial reads and assign them to corresponding taxa (Truong et al., 2015).

154 Microbial reads can be introduced by contamination or natural microbiome content in
155 the sample, such as viral, bacterial, fungi, or other microbial species (Salter et al., 2014).
156
157 Taken together, ROP considers six classes of *unmapped reads*: (1) lost human
158 reads, (2) hyper-edited reads, (3) lost repeat elements, (4) reads from 'non-co-linear'
159 (NCL) RNAs, (5) reads from the recombination of BCR and TCR segments (i.e. V(D)J
160 recombination), and (6) microbial reads. Previously proposed individual methods do
161 examine some of these classes (Blachly et al., 2015; Brown et al., 2015; Chuang et al.,
162 2015; Kostic et al., 2011; N. Strauli & Hernandez, 2015). However, we find that performing
163 a sequential analysis, in the order described above, is critical for minimizing
164 misclassification of reads due to homologous sequences between the different classes.
165 Furthermore, as shown in the Results section below, only a comprehensive analysis allows
166 comparison across these classes. We have demonstrated the robustness of the proposed
167 approach against alternating order of steps and values of the thresholds (Supplemental
168 Methods and Supplemental Methods Figure SM1). Complete details of ROP, including all
169 parameters and thresholds used, are provided in the Supplemental Methods.
170

171



172

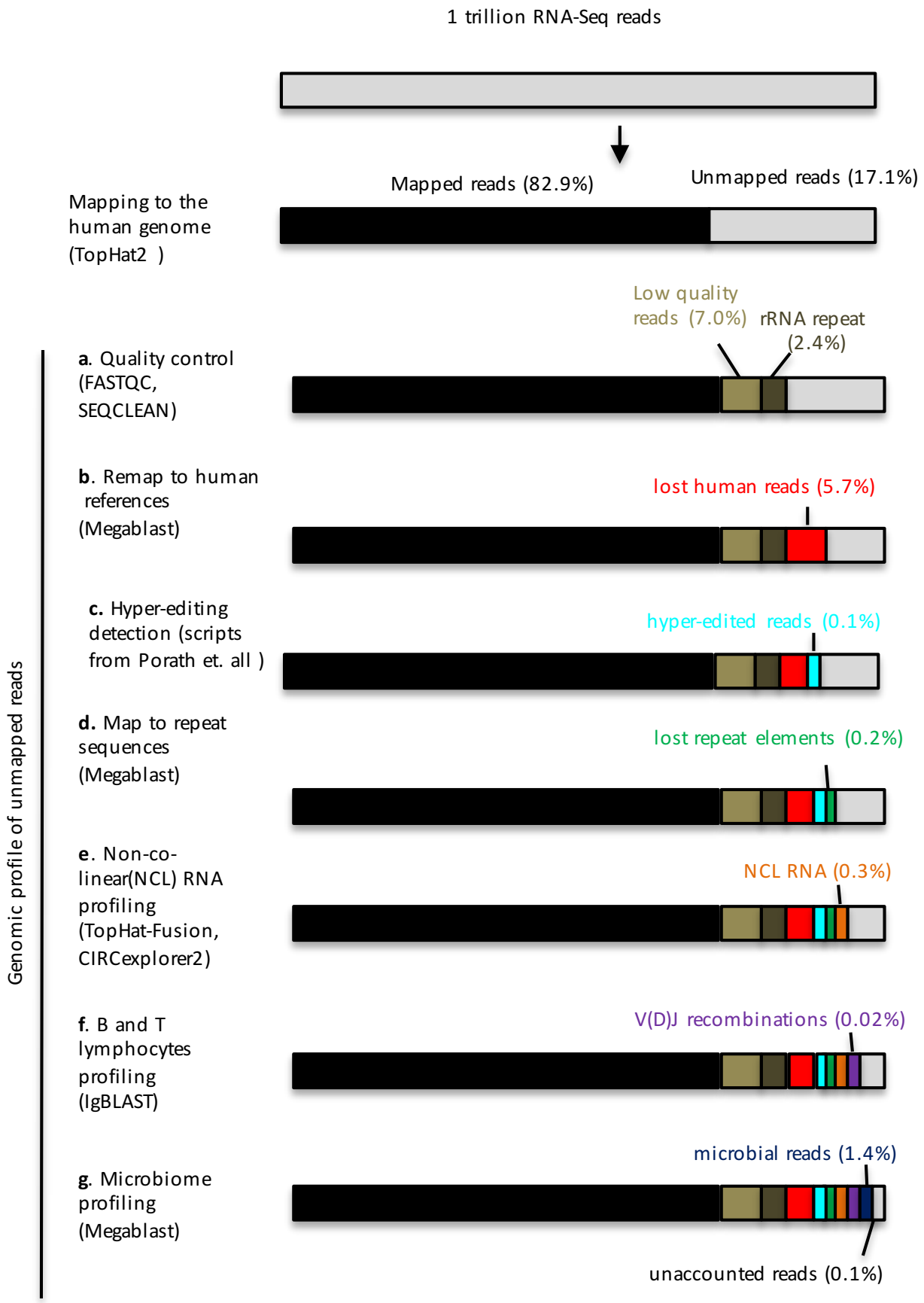
173 **Figure 1. Schematic of the Read Origin Protocol (ROP).** Human reads are identified by

174 mapping all reads onto the reference sequences using a standard high-throughput

175 mapping algorithm. ROP protocol categorizes mapped reads into genomic (red colors)
176 and repetitive (green colors) categories. Unmapped reads that fail to map are extracted
177 and further filtered to exclude low quality reads, low complexity reads, and reads from
178 rRNA repeats (brown color). ROP protocol is able to identify unmapped reads aligned to
179 human references with use of a more sensitive alignment tool (lost human reads: red
180 color), unmapped reads aligned to human references with excessive ('hyper') editing
181 (hyper-edited RNAs: cyan color), unmapped reads aligned to the repeat sequences (lost
182 repeat elements: green color), unmapped reads spanning sequences from distant loci
183 (non-co-linear: orange color), unmapped reads spanning antigen receptor gene
184 rearrangement in the variable domain (V(D)J recombination of BCR and TCR: violet color),
185 and unmapped reads aligned to the microbial reference genomes and marker genes
186 (microbial reads: blue color).
187
188 ***The ROP protocol is able to account for 99.9% of all reads***
189
190 To test ROP, we applied it to one trillion RNA-Seq reads across 54 tissues from 2630
191 individuals. The data was combined from 3 studies: (1) in-house RNA-Seq data (n=86)
192 from the peripheral blood, nasal, and large airway epithelium of asthmatic and control
193 individuals (S1); (2) multi-tissue RNA-Seq data from Genotype-Tissue Expression (GTEx v6)
194 from 53 human body sites (Consortium & others, 2015) (n=8555) (S2); (3) randomly
195 selected RNA-Seq samples from the Sequence Read Archive (SRA) (n=2000) (S3). Unless
196 otherwise noted, we reported percentage of reads averaged across 3 datasets.

197 RNA-Seq data obtained from the three sources represent a large collection of
198 tissue types and read diversity. We selected these three sources to most accurately model
199 the precision and broad applicability of ROP. The in-house RNA-Seq data was collected
200 from 53 asthmatics and 33 controls. RNA-seq libraries were prepared from total RNA with
201 two types of RNA enrichment methods: (1) Poly(A) enrichment libraries, applied to RNA
202 from peripheral blood and nasal epithelium (n=38), and (2) ribo-depletion libraries,
203 applied to RNA from large airway epithelium (n=49). The GTEx dataset was derived from
204 38 solid organ tissues, 11 brain subregions, whole blood, and three cell lines across 544
205 individuals. Randomly selected SRA RNA-Seq samples included samples from whole
206 blood, brain, various cell lines, muscle, and placenta. Length of reads from in-house data
207 was 100bp, read length in Gtex data was 76bp, read length in SRA data ranged from 36bp
208 to 100bp. In total, 1 trillion reads (97 Tbp) derived from 10641 samples were available for
209 ROP (Supplemental Table S1 and Supplementary Methods). For counting purposes, the
210 pairing information of the reads is disregarded, and each read from a pair is counted
211 separately.
212
213 We used standard read mapping procedures to obtain mapped and unmapped
214 reads from all three data sources. Read mapping for GTEx data was performed by the
215 GTEx consortium using TopHat2 (Kim et al., 2013). Following the GTEx consortium
216 practice, we used TopHat2 to map reads from in-house and SRA studies. High-throughput
217 mapping using TopHat2 (Kim et al., 2013) recovered 83.1% of all reads from three studies
218 (Fig. 2.a), with the smallest fraction of reads mapped in the SRA study (79% mapped

219 reads). From the *unmapped reads*, we first excluded low-quality/low-complexity reads
220 and reads mapping to the rRNA repeating unit, which together accounted for 7.0% and
221 2.4% of all reads, respectively (Fig. 2.b). We were then able to align unmapped reads to
222 human reference sequences (5.7% of all reads, Fig. 2.c) and identify "hyper-edited" reads
223 (0.1% of all reads Fig. 2.d). We then referenced repeat sequences (0.2% of all reads, Fig.
224 2.d), reads identified as 'non-co-linear'(NCL) RNAs (circRNAs, gene fusion or trans-
225 splicing) (0.3% of all reads, Fig. 2.e), and reads mapped to recombined B and T cell
226 receptors (0.02% off all reads, Fig. 2.f). The remaining reads were mapped to the microbial
227 sequences (1.4% off all reads, Fig. 2.g). Following the seven steps of ROP, the origins of
228 99.9% of reads were identified. Genomic profile of unmapped reads for each dataset is
229 separately reported in Table S2. Uncategorized reads from SRA samples are freely
230 available at https://smangul1.github.io/recycle-RNA-seq/. This resource allows the
231 bioinformatics community to further increase the number of reads with known origin.
232


1 trillion RNA-Seq reads
Mapping to the human genome (TopHat2)
Mapped reads (82.9%)
Unmapped reads (17.1%)
Genomic profile of unmapped reads
a. Quality control (FASTQC, SEQCLEAN)
Low quality reads (7.0%)
rRNA repeat (2.4%)
b. Remap to human references (Megablast)
lost human reads (5.7%)
c. Hyper-editing detection (scripts from Porath et. all)
hyper-edited reads (0.1%)
d. Map to repeat sequences (Megablast)
lost repeat elements (0.2%)
e. Non-co-linear(NCL) RNA profiling (TopHat-Fusion, CIRCexplorer2)
NCL RNA (0.3%)
f. B and T lymphocytes profiling (IgBLAST)
V(D)J recombinations (0.02%)
g. Microbiome profiling (Megablast)
microbial reads (1.4%)
unaccounted reads (0.1%)

234 **Figure 2. Genomic profile of unmapped reads across 10641 samples and 54 tissues.**
235 Percentage of unmapped reads for each category is calculated as a fraction from the total
236 number of reads. Bars of the plot are not scaled. Human reads (black color) mapped to
237 reference genome and transcriptome via TopHat2. (a) Low quality/low-complexity (light
238 brown) and reads matching rRNA repeating unit (dark brown) were excluded. (b) Hyper-
239 edited reads are captured by hyper-editing pipeline proposed in (Porath et al., 2014). (c)
240 ROP identifies lost human reads (red color) from unmapped reads using a more sensitive
241 alignment. (d) ROP identifies lost repeat sequences (green color) by mapping unmapped
242 reads onto the reference repeat sequences. (e) Reads arising from trans-spicing, gene
243 fusion and circRNA events (orange color) are captured by a TopHat-Fusion and
244 CIRCexplorer2 tools. (f) IgBlast is used to identify reads spanning B and T cell receptor
245 gene rearrangement in the variable domain (V(D)J recombinations) (violet color). (g)
246 Microbial reads (blue color) are captured by mapping the reads onto the microbial
247 reference genomes.
248
249 ***The ROP protocol identifies lost human reads***
250 Some human reads may remain unmapped due to the heuristic nature of high throughput
251 aligners (Baruzzo et al., 2016; Siragusa et al., 2013). As shown by Baruzzo et al.,
252 even the best performing RNA-Seq aligners fail to map at least 10% of reads simulated
253 from the human references. To prevent misclassification of reads derived from human
254 genome into other downstream ROP categories, we used the slower and more sensitive

255 Megablast aligner on this subset of unmapped reads. This method allows us to filter an
256 additional 5.7% of human reads.
257
258 We investigated the impact of mapping parameters and RNA-Seq aligners on the
259 number of unmapped reads. We additionally used STAR (Dobin et al., 2013) and added
260 results for sensitive and very sensitive mapping settings of each of the tools
261 (Supplemental Methods and Supplemental Table S4). We observe that an alternative
262 aligner and a more sensitive mapping setting has no substantial effect on the number of
263 mapped reads (Supplementary Table S5). This is in line with Baruzzo et al., 2016, which
264 have shown that optimizing the parameters of RNA-Seq aligner is a non-trivial task and
265 methods with good performance for the default setting is a preferred choice.
266
267 Using both mapped and unmapped reads across the studies, we classified on
268 average 7.5% of the RNA-Seq reads as repetitive sequences originated from various
269 repeat classes and families (Supplemental Fig. S2). We observe Alu elements to have 33%
270 relative abundance, which was the highest among all the repeat classes. Among DNA
271 repeats, hAT-Charlie was the most abundant element with 50% relative abundance
272 (Supplemental Fig. S3). Among SVA retrotransposons, SVA-D was the most abundant
273 element with 50% relative abundance (Supplemental Fig. S4). Consistent with repEnrich
274 (Criscione, Zhang, Thompson, Sedivy, & Neretti, 2014), when using in-house data we
275 observe the differences in proportions of L1 and Alu elements between poly(A) and ribo-
276 depletion libraries. Among the repeat reads, poly(A) samples have the highest fraction of

277 reads mapped to Alu elements, and ribo-depleted samples have the highest fraction
278 mapped to L1 elements (Supplemental Fig. S5). Among the GTEx tissues, testis showed
279 significantly higher expression of SVA F retrotransposons compared to other GTEx tissues
280 ($p = 2.46\times10^{-33}$) (Supplemental Figure S6). Furthermore, we observe high co-
281 expression of *Alu* elements and L1 elements across GTEx tissues ($R^2 = 0.7615$)
282 (Supplemental Figure 7).
283
284 ***ROP identifies hyper edited reads***
285 Using standard read mapping approaches, some human reads may remain unmapped due
286 to "hyper editing." An extremely common post-transcriptional modification of RNA
287 transcripts in human is A-to-I RNA editing (Bazak et al., 2014). Adenosine deaminases
288 acting on RNA (ADARs) proteins can modify a genetically encoded adenosine (A) into an
289 inosine (I). Inosine is read by the cellular machinery as a guanosine (G), and, in turn,
290 sequencing of inosine results in G where the corresponding DNA sequencing reads
291 A. Current methods to detect A-to-I editing sites are based on the alignment of RNA-Seq
292 reads to the genome to identify such A-to-G mismatches. Reads with excessive ('hyper')
293 editing are usually rejected by standard alignment methods. In this case, many A-to-G
294 mismatches obscure their genomic origin.
295 We have identified hyper-edited reads by using the pipeline proposed in (Porath,
296 Carmi, & Levanon, 2014). This hyper-editing pipeline transforms all As into Gs, in both
297 the unmapped reads and the reference genome, and the pipeline realigns the

298 transformed RNA-Seq reads and the transformed reference genome. The method then
299 recovers original sequences and searches for dense clusters of A-to-G mismatches.
300 A total of 201,676,069 hyper-edited reads were identified across all samples
301 from the three studies. As a control for the detection, we calculated the prevalence of all
302 6 possible nucleotide substitutions and found that 79.9% (201,676,069/252,376,867) of
303 the detected reads were A-to-G mismatches (Supplemental Fig. S8). In comparison, the
304 in-house RNA-Seq samples have a 96.1% rate of A-to-G mismatches. This massive over-
305 representation of mismatches strongly suggests that these reads resulted from ADAR
306 mediated RNA editing. However, additional experiments are required to confirm the
307 nature of these edits. In addition, we found that the nucleotide sequence context of the
308 detected editing sites complies with the typical sequence motif of ADAR targets
309 (Supplemental Fig. S9).
310
311 ***The ROP protocol complements transcriptome profiling by non-co-linear RNAs***
312 The ROP protocol is able to detect 'non-co-linear' reads via Tophat-Fusion (Kim & Salzberg,
313 2011) and CIRCexplorer2 (Zhang et al., 2016) tools from three classes of events: (1) reads
314 spliced distantly on the same chromosome supporting trans-splicing events; (2) reads
315 spliced across different chromosomes supporting gene fusion events; and (3) reads
316 spliced in a head-to-tail configuration supporting circRNAs. On average, we observed 816
317 trans-splicing events, 7510 fusion events, and 930 circular events per individual sample
318 supported by more than one read. Over 90% of non-co-linear events were supported by
319 fewer than 10 samples (Supplemental Fig. S10). We used a liberal threshold, based on

320 number of reads and individuals, because our interest is mapping all reads. However, a
321 more stringent cut off is recommended for confident identification of non-co-linear
322 events, specially in the clinical settings.

323 Based on the in-house RNA-Seq data, we observe that the library preparation
324 technique strongly affects the capture rate of non-co-linear transcripts. To compare the
325 number of NCL events, we sub-sampled unmapped reads to 4,985,914 for each sample,
326 which corresponded to the sample with the smallest number of unmapped reads among
327 in-house RNA-Seq samples. We observed an average increase of 92% of circRNAs in
328 samples prepared by ribo-depletion compared to poly(A) protocol (p-value = 3×10^{-12})
329 (Supplemental Fig. S11). At the same time, we observed an average 43% decrease of
330 trans-splicing and fusion events in samples prepared by ribo-depletion compared to
331 poly(A) protocol (p-value < 8×10^{-4}) (Supplemental Fig. S11). However, because the
332 tissues differed between protocols (e.g., nasal versus large airway epithelium), this effect
333 might be due in part to tissue differences in NCL events. We view the tissue differences
334 effect to be unlikely. We previously showed that gene expression profiles of nasal airway
335 tissue largely recapitulate expression profiles in the large airway epithelium tissue (Poole
336 et al., 2014).

337 Furthermore, many NCL events will not be captured by poly-A selection.
338 Therefore, we expect systematic differences in NCL abundance between capture
339 methods. There were no statistically significant differences (p-value > 5×10^{-3}) between
340 NCL events in cases and controls. We have compared number of NCL reads across GTEx
341 tissue, and we observe the highest fraction of NCL reads across pancreas samples with

342 0.75% of reads classified as NCL reads. In all other tissue types, ROP classified
343 approximately .3% reads as NCL reads (Supplemental Figure S12).
344
345 ***ROP identifies microbial and immune reads and differentiate tissue types and disease***
346 ***status***
347 Reads mapped to B and T cell receptor loci and unmapped reads were used to survey the
348 human adaptive immune repertoires in health and disease. We first used the mapped
349 reads to extract reads entirely aligned to BCR and TCR genes. Using IgBlast (Ye et al.,
350 2013), we identified unmapped reads with extensive somatic hyper mutations (SHM) and
351 reads arising from V(D)J recombination. After we identified all the reads with the human
352 origin, we detected microbial reads by mapping the remaining reads onto microbial
353 reference genomes and phylogenetic marker genes. Here, the total number of microbial
354 reads obtained from the sample is used to estimate microbial load. We use MetaPhlAn2
355 (Truong et al., 2015) to assign reads on microbial marker genes and determine the
356 taxonomic composition of the microbial communities.
357 Using in-house RNA-Seq data, we compare immunological and microbial profiles
358 across asthmatics and unaffected controls for the peripheral blood, nasal, and large
359 airway epithelium tissues. A total of 339 bacterial taxa were assigned with Metaphlan2
360 (Truong et al., 2015) across all studies and are freely available at
361 https://smangul1.github.io/recycle-RNA-seq/.
362 Using Metaphlan2, we detected bacterial reads in all GTEx tissues except testis,
363 adrenal gland, heart, brain, and nerve. We also observe no bacteria reads in the following

364 cell lines: EBV-transformed lymphocytes(LCLs), Cells-Leukemia (CML), and Cells-
365 Transformed fibroblasts cell lines. On average, we observe 1.43 +-0.43 phyla assigned per
366 sample. All samples were dominated by Proteobacteria (relative genomic abundance of
367 73%+-28%). Other phyla detected included Acidobacteria, Actinobacteria, Bacteroidetes,
368 Cyanobacteria, Fusobacteria, and Firmicutes. Consistent with previous studies, we
369 observe the nasal epithelium is dominated by Actinobacteria phyla (particularly the
370 *Propionibacterium* genus) (Yan et al., 2013), and the large airway epithelium is dominated
371 by Proteobacteria phyla (Beck, Young, & Huffnagle, 2012) (Supplemental Table S3). As a
372 positive control for virus detection, we used GTEx samples from EBV-transformed
373 lymphoblastoid cell lines (LCLs). ROP detected EBV virus across all LCLs samples. An
374 example of a coverage profile of EBV virus for one of the LCLs samples is presented in
375 Supplemental Fig. S13.

376 We assess combinatorial diversity of the B and T cell receptor repertoires by
377 examining the recombination of the of Variable (V) and Joining (J) gene segments from
378 the variable region of BCR and TCR loci. We used per sample alpha diversity (Shannon
379 entropy) to incorporate the total number of VJ combinations and their relative
380 proportions into a single diversity metric. We observed a mean alpha diversity of .7 among
381 all the samples for immunoglobulin kappa chain (IGK). Spleen, minor salivary gland, and
382 small intestine (terminal ileum) were the most immune diverse tissue, with corresponding
383 IGK alpha diversity of 86.9, 52.05, and 43.96, respectively (Supplemental Fig. S14-S15).
384 Across all the tissues and samples, we obtained a total of 312 VJ recombinations for IGK
385 chains and 194 VJ recombinations for IGL chains. Inferred recombinations are freely

386 available at https://smangul1.github.io/recycle-RNA-seq/.

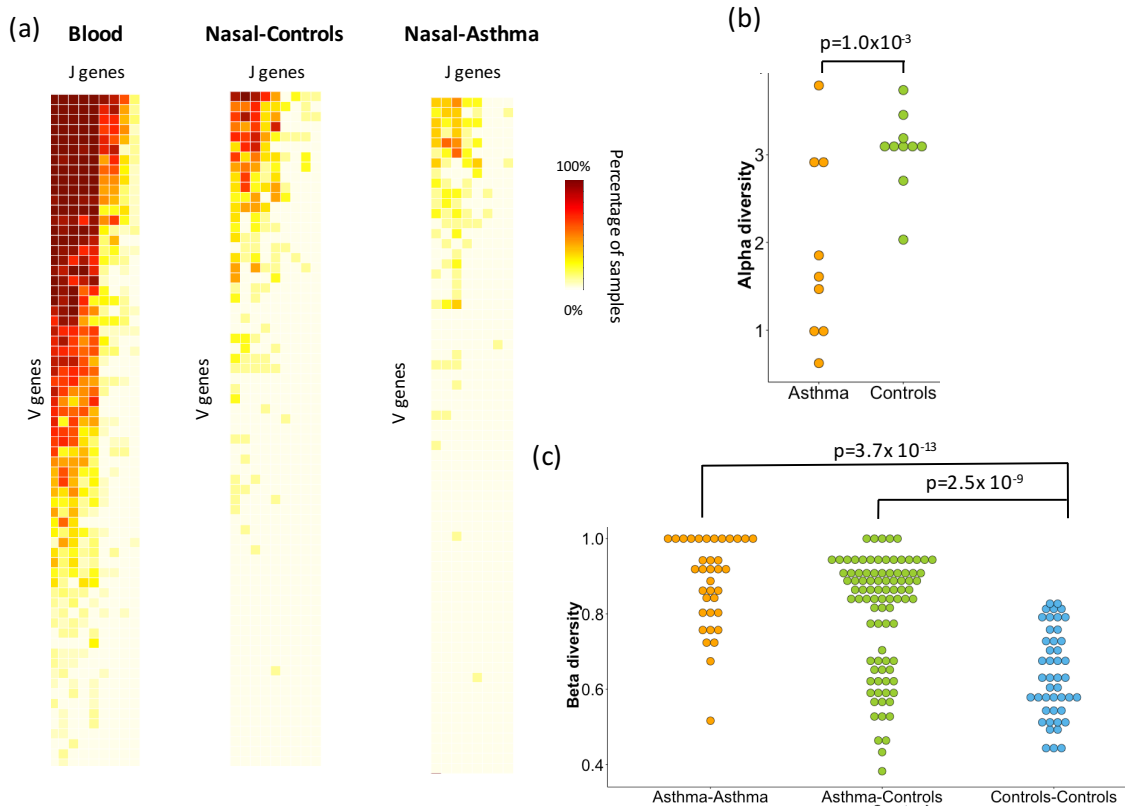
387

388 Using in-house data, we investigated the effect of different library preparation
389 techniques over the ability to detect B and T cell receptor transcripts. We compared the
390 alpha diversity in large airway samples to nasal samples (Supplemental Fig. S16).
391 Decreased alpha diversity in large airway samples compared to nasal (2.5 for nasal versus
392 1.0 for large airway) could correspond to an overall decrease in percentage of immune
393 reads. This effect can be attributed to the ribo-depletion protocol not enriching for
394 polyadenylated antibody transcripts. Alternatively, it may result from clonal expansion of
395 certain clonotypes responding to the cognate antigen.

396

397 Our comprehensive ROP protocol presents several advantages over previous
398 methods designed to examine features of unmapped reads. First, our method
399 interrogates relationships between features. To explore interactions between the
400 immune system and microbiome, we compared immune diversity against microbial load.
401 Microbes trigger immune responses, eliciting proliferation of antigen-specific
402 lymphocytes. This dramatic expansion skews the antigen receptor repertoire in favor of a
403 few dominant clonotypes and decreases immune diversity (Spreafico et al.,
404 2016). Therefore, we reasoned that antigen receptor diversity in the presence of
405 microbial insults should shrink. In line with our expectation, we observed that
406 combinatorial immune diversity of IGK locus was negatively correlated with the viral load

407 (Pearson coefficient r = -0.55, p-value = 2.4 x 10^{-6}), consistent also for bacteria and
408 eukaryotic pathogens across BCR and TCR loci (Supplemental Fig. S17).
409
410 Using in-house data, we compared alpha diversity of asthmatic individuals (n = 9)
411 and healthy controls (n = 10). The combinatorial profiles of B and T cell receptors in blood
412 and large airway tissue provide no differentiation between case control statuses. Among
413 nasal samples, we observed decreased alpha diversity for asthmatic individuals relative
414 to healthy controls (p-value = 10^{-3}) (Fig. 2.b). Additionally, we used beta diversity
415 (Sørensen–Dice index) to measure compositional similarities between samples, including
416 gain or loss of VJ combinations of IGK locus. We observed higher beta diversity
417 corresponding to a lower level of similarity across the nasal samples of asthmatic
418 individuals in comparison to samples from unaffected controls (Fig. 2.c, p-value < $3.7 \times 10^{-}$
419 ^{13}). Moreover, nasal samples of unaffected controls are significantly more similar than
420 samples from the asthmatic individuals (Fig. 2.c, p-value < 2.5×10^{-9}). Recombination
421 profiles of immunoglobulin lambda locus (IGL) and T cell receptor beta and gamma (TCRB
422 and TCRG) loci yielded a similar pattern of decreased beta diversity across nasal samples
423 of asthmatic individuals (Supplemental Fig. S18-S20). Together the results demonstrate
424 the ability of ROP to interrogate additional features of the immune system without the
425 expense of additional TCR/BCR sequencing.
426
427



428

429

430 **Figure 3. Combinatorial diversity of immunoglobulin kappa locus (IGK) locus**

431 **differentiates disease status.** (a) Heatmap depicting the percentage of RNA-Seq samples

432 supporting of particular VJ combination for whole blood (n = 19), nasal epithelium of

433 healthy controls (n = 10), and asthmatic individuals (n = 9). Each row corresponds to a V

434 gene, and each column correspond to a J gene. (b) Alpha diversity of nasal samples is

435 measured using the Shannon entropy and incorporates total number of VJ combinations

436 and their relative proportions. Nasal epithelium of asthmatic individuals exhibits

437 decreased combinatorial diversity of IGK locus compared to healthy controls (p-value = 1

438 x 10^{-3}). (c) Compositional similarities between the nasal samples in terms of gain or loss

439 of VJ combinations of IGK locus are measured across paired samples from the same group

440 (Asthma, Controls) and paired samples from different groups (Asthma versus Controls)
441 using Sørensen–Dice index. Lower level of similarity is observed between nasal samples
442 of asthmatic individuals compared to unaffected controls (p-value < 7.3×10^{-13}). Nasal
443 samples of unaffected controls are more similar to each other than to the asthmatic
444 individuals (p-value < 2.5×10^{-9}).
445
446 **DISCUSSION**
447 Our study is the first that systematically accounts for almost all reads, totaling one trillion,
448 available via three RNA-seq datasets. We demonstrate the value of analyzing unmapped
449 reads present in the RNA-seq data to study the non-co-linear, RNA editing,
450 immunological, and microbiome profiles of a tissue. We developed a new tool (ROP) that
451 accounts for 99.9% of the reads, a substantial increase compared over the 82.2% of reads
452 account for using conventional protocols. We found that the majority of *unmapped reads*
453 are human in origin and from diverse sources, including repetitive elements, A-to-I RNA
454 editing, circular RNAs, gene fusions, trans-splicing, and recombined B and T cell receptor
455 sequences. In addition to those derived from human RNA, many reads were microbial in
456 origin and often occurred in numbers sufficiently large to study the taxonomic
457 composition of microbial communities in the tissue type represented by the sample.
458
459 We found that both unmapped human reads and reads with microbial origins are
460 useful for differentiating between type of tissue and status of disease. For example, we
461 found that the immune profiles of asthmatic individuals have decreased immune diversity

462 when compared to those of controls. Further, we used our method to show that immune
463 diversity is inversely correlated with microbial load. This case study highlights the
464 potential for producing novel discoveries, when the information in RNA-seq data is fully
465 leveraged by incorporating the analysis of unmapped reads, without need for additional
466 TCR/BCR or microbiome sequencing. The ROP profile of unmapped reads output by our
467 method is not limited to RNA-Seq technology and may apply to whole-exome and whole-
468 genome sequencing. We anticipate that ROP profiling will have broad future applications
469 in studies involving different tissue and disease types.

470 We observed large effects when using different library preparation protocols on
471 non-co-linear, immunological, and microbial profiles. For example, the poly(A) protocol
472 better captures antibody transcripts by enriching for polyadenylated transcripts, while
473 ribo-depletion protocols capture more circRNAs. The results presented here suggest that
474 selection of a protocol impacts quality of analysis results, and our study may guide the
475 choice of protocol depending on the features of interest.

476 The ROP protocol facilitates a simultaneous study of immune systems and
477 microbial communities, and this novel method advances our understanding of the
478 functional, interrelated mechanisms driving the immune system, microbiome, human
479 gene expression, and disease etiology. We hope that future efforts will provide a
480 quantitative and qualitative assessment of the immune and microbial components of
481 disease across various tissues. Recent increase in read length and sequencing efficiency
482 provides opportunity for studying individual microbial species and full TCR/BCR
483 sequencing.

484 **METHODS**

485 **In-house RNA-Seq data**

486 For poly(A) selected samples (n=38), we used a subset of Puerto Rican Islanders recruited
487 as part of the on-going Genes-environments & Admixture in Latino Americans study
488 (GALA II) (Anders, Pyl, & Huber, 2014; Jin, Tam, Paniagua, & Hammell, 2015; Melé et al.,
489 2015; Tarailo-Graovac & Chen, 2009). Nasal epithelial cells were collected from behind
490 the inferior turbinate with a cytology brush using a nasal illuminator. Whole blood was
491 collected using PAXgene RNA blood tubes. RNA was isolated using PAXgene RNA blood
492 extraction kits. For ribo-depleted samples (n=49), we recruited adults aged 18-70 years
493 to undergo research bronchoscopy. During bronchoscopy airway epithelial brushings,
494 samples were obtained from 3rd-4th generation bronchi. RNA was extracted from the
495 epithelial brushing samples using the Qiagen RNeasy mini-kit.

496

497 Poly(A) selected RNA-seq libraries (n=38) were constructed using 500 ng of blood and
498 nasal airway epithelial total RNA from 9 atopic asthmatics and 10 non-atopic controls.
499 Libraries were constructed and barcoded with the Illumina TruSeq RNA Sample
500 Preparation v2 protocol. Barcoded nasal airway RNA-seq libraries from each of the 19
501 subjects were pooled and sequenced as 2 x 100bp paired-end reads across two flow cells
502 of an Illumina HiSeq 2000. Barcoded blood RNA-seq libraries from each of the 19 subjects
503 were pooled and sequenced as 2 x 100bp paired end reads across 4 lanes of an Illumina
504 Hiseq 2000 flow cell. Ribo-depleted RNA-seq libraries (n=38), were constructed using
505 100ng of isolated RNA of large airway epithelium total RNA from 61 samples. Libraries

506 were constructed and barcoded with the TruSeq Stranded Total RNA using a Ribo-Zero
507 Human/Mouse/Rat library preparation kit, per manufacturer's protocol. Barcoded
508 bronchial epithelial RNA-seq libraries were multiplexed and sequenced as 2 x 100bp
509 paired end reads on an Illumina HiSeq 2500. We excluded 12 samples from further
510 analyses due to high ribosomal RNA read counts (library preparation failure), leaving a
511 total of 49 samples suitable for further analyses.

512 **GTEx RNA-Seq data**

513 We used RNA-Sequencing data from Genotype-Tissue Expression study (GTEx Consortium
514 v.6) corresponding to 8,555 samples collected from 544 individuals from 53 tissues
515 obtained from Genotype-Tissue Expression study (GTEx v6). RNA-Seq data is from Illumina
516 HiSeq sequencing of 75 bp paired-end reads. The data was derived from 38 solid organ
517 tissues, 11 brain subregions, whole blood, and three cell lines of postmortem donors. The
518 collected samples are from adults matched for age across males and females. We
519 downloaded the mapped and unmapped reads in BAM format from dbGap
520 (http://www.ncbi.nlm.nih.gov/gap).

521

522 **SRA RNA-Seq data**

523

524 Samples (n=2000) were randomly selected using SQLite database from R/Bioconductor
525 package SRAdb (https://bioconductor.org/packages/release/bioc/html/SRAdb.html). We
526 have used a script from
527 https://github.com/nellore/runs/blob/master/sra/define_and_get_fields_SRA.R to

528 select run_accessions from the sra table with platform = 'ILLUMINA', library_strategy =
529 'RNA-Seq', and taxon_id = 9606 (human).
530

531 **Workflow to categorize mapped reads**

532 We mapped reads onto the human transcriptome (Ensembl GRCh37) and genome
533 reference (Ensembl hg19) using TopHat2 (v 2.0.13) with the default parameters. Tophat2
534 was supplied with a set of known transcripts (as a GTF formatted file, Ensembl GRCh37)
535 using –G option. The mapped reads of each sample are stored in a binary format (.bam).
536 ROP (gprofile.py) categorizes the reads into genomic categories (junction read, CDS,
537 intron, UTR3, UTR5, introns, inter-genic read, deep a deep inter-genic read, mitochondrial
538 read, and multi-mapped read) based on their compatibility with the features defined by
539 Ensembl (GRCh37) gene annotations. ROP (rprofile.py) categorizes reads into repeat
540 elements (classes and families) based on their compatibility with repeat instances defined
541 by RepeatMasker annotation (Repeatmasker v3.3, Repeat Library 20120124). We count
542 the number of reads overlapping variable(V), diversity (D), joining (J), and constant (C)
543 gene segments of B cell receptor (BCR) and T cell receptor (TCR) loci using htseq-count
544 (HTSeq v0.6.1).
545

546 **Workflow to categorize unmapped reads**

547 We first converted the unmapped reads saved by TopHat2 from a BAM file into a FASTQ
548 file (using samtools with bam2fq option). The FASTQ file of unmapped reads contains full

549 read pairs (both ends of a read pair were unmapped) and discordant read pairs (one read
550 end was mapped while the other end was unmapped). We disregarded the pairing
551 information of the unmapped reads and categorized unmapped reads using the
552 protocol's seven steps. Reads identified at each step are filtered out.
553
554 **A. Quality Control.** Low quality reads, defined as reads that have quality lower than 30 in
555 at least 75% of their base pairs, were identified by FASTX (v 0.0.13). Low complexity reads,
556 defined as reads with sequences of consecutive repetitive nucleotides, were identified by
557 SEQCLEAN. As a part of the quality control, we also excluded unmapped reads mapped
558 onto the rRNA repeat sequence (HSU13369 Human ribosomal DNA complete repeating
559 unit) (BLAST+ 2.2.30).
560
561 **B. Remap to human references**. We remapped the remaining unmapped reads to the
562 human reference genome (hg19) and transcriptome (known transcripts, Ensembl
563 GRCh37) using Megablast (BLAST+ 2.2.30). ROP step 3.
564
565 **C. Hyper-editing detection.** We used a hyper-editing pipeline (HE-pipeline
566 http://levanonlab.ls.biu.ac.il/resources/zip), which is capable of identifying hyper-
567 edited reads.
568
569 **D. Map to repeat sequences.** The remaining unmapped reads were mapped to the
570 reference repeat sequences using Megablast (BLAST+ 2.2.30). The reference repeat

571 sequences were downloaded from Repbase v20.07 (http://www.girinst.org/repbase/).
572 Human repeat elements (humrep.ref and humsub.ref) were merged into a single
573 reference.
574
575 **E. Non-co-linear (NCL) RNA profiling**. NCL events include three classes of events: reads
576 supporting trans-splicing events that are spliced distantly on the same chromosome;
577 reads supporting gene fusion events that are spliced across different chromosomes; and
578 reads supporting circRNAs that are spliced in a head-to-tail configuration. To distinguish
579 between these three categories, we use circExplorer2 (v2.0.13). CircExplorer2 relies on
580 Tophat-Fusion (v2.0.13, bowtie1 v0.12.) and allows simultaneous monitoring of NCL
581 events in the same run. To extract trans-spicing and gene fusion events from the TopHat-
582 Fusion output, we ran a ruby custom script, which is part of the ROP pipeline (NCL.rb).
583
584 **D. B and T lymphocytes profiling.** We used IgBLAST (v. 1.4.0) with a stringent e-value
585 threshold (e-value < 10^{-20}) to map the remaining unmapped reads onto the V(D)J gene
586 segments of the of the B cell receptor (BCR) and T cell receptor (TCR) loci. Gene segments
587 of B cell receptors (BCR) and T cell receptors (TCR) were imported from IMGT
588 (International ImMunoGeneTics information system). IMGT database contains: variable
589 (V) gene segments; diversity (D) gene segments; and joining (J) gene segments.
590
591 **E. Microbiome profiling.** We used Megablast (BLAST+ 2.2.30) to align remaining
592 unmapped reads onto the collection of bacterial, viral, and eukaryotic reference

593 genomes. Bacterial and viral genomes were downloaded from NCBI
594 (ftp://ftp.ncbi.nih.gov/). Genomes of eukaryotic pathogens were downloaded from
595 EuPathDB (http://eupathdb.org/eupathdb/). We used MetaPhlAn2 (Metagenomic
596 Phylogenetic Analysis, v 2.0) to obtain the taxonomic profile of microbial communities
597 present in the sample.
598

599 **Reference databases**

600 A detailed description of reference databases used by ROP is provided in Supplemental
601 Materials.

602 **Comparing diversity across groups**

603 First, we sub-sampled unmapped reads by only including reads corresponding to a sample
604 with the smallest number of unmapped reads. Diversity within a sample was assessed
605 using the richness and alpha diversity indices. Richness was defined as a total number of
606 distinct events in a sample. We used Shannon Index (SI), incorporating richness and
607 evenness components, to compute alpha diversity, which is calculated as follows:

608 $$\mathrm{SI} = -\sum (p \times \log_2(p))$$

609 We used beta diversity (Sørensen–Dice index) to measure compositional similarities
610 between the samples in terms of gain or loss in events. We calculated the beta diversity
611 for each combination of the samples, and we produced a matrix of all pairwise sample
612 dissimilarities. The Sørensen–Dice beta diversity index is measured as $1 - \frac{2\mathrm{J}}{\mathrm{A}+\mathrm{B}}$, where J is

613 the number of shared events, while A and B are the total number of events for each
614 sample, respectively.
615
616
617 **The robustness of the ROP results against changing the thresholds for each of the ROP**
618 **steps**
619 We have performed the robustness analysis to investigate the impact of the thresholds
620 used in each step of the ROP approach. For each ROP step, we reported the number of
621 reads identified under different thresholds. The results are presented as cumulative
622 frequency plots (Supplemental Methods Figure SM1).
623
624 **The impact of mapping parameters and RNA-Seq aligners on the number of unmapped**
625 **reads**
626 We investigated the impact of alternative aligners (STAR,
627 https://github.com/alexdobin/STAR) and carefully adjusted the mapping setting to
628 achieve sensitive and very sensitive settings (Supplemental Table S4). The average
629 runtime on Hoffman2 Cluster for Tophat per million reads was 2.5 hours; STAR, 0.13
630 hours; and Novoalign, 9.1 hours. Novoalign was not considered in the analysis due to its
631 substantially longer running time, which made it infeasible for the protocol.
632
633
634 **Software availability**

635 The ROP software is publicly available at https://github.com/smangul1/rop. The source
636 code for ROP v1.0.4 is also available as Supplemental Material. Custom scripts
637 necessary to reproduce the results and reference files are distributed with Supplemental
638 Material, and are also available with ROP software. A tutorial with detailed instructions
639 on how to run ROP is freely available at https://github.com/smangul1/rop/wiki. For a
640 quick start, an example with 2508 unmapped reads is distributed with the ROP package.
641 Reads are randomly selected from a publically available normal skin (SRR1146076) RNA-
642 Seq sample and might not represent the typical reads of RNA-Seq experiment. The reads
643 are provided for demonstration purposes and are distributed with ROP software.
644 Additional details of ROP, including all parameters and thresholds used, are provided in
645 the Supplemental Methods.
646
647
648 **ACKNOWLEDGEMENT**
649
650 **DISCLOSURE**
651 The authors declare no competing interests.
652
653 **DATA ACCESS**
654
655 Portions of the nasal airway epithelial whole transcriptome data were published in a
656 previous manuscript (Tarailo-Graovac & Chen, 2009)

657

658

659 **REFERENCES**

660 Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D. S., Busby, M. A., Berlin, A. M., …
661 others. (2013). Comparative analysis of RNA sequencing methods for degraded or
662 low-input samples. *Nature Methods*, *10*(7), 623–629. article.

663 Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq--A Python framework to work with high-
664 throughput sequencing data. *Bioinformatics*, btu638. article.

665 Andrews, S., & others. (2010). FastQC: A quality control tool for high throughput sequence
666 data. *Reference Source*. article.

667 Ardlie, K. G., Deluca, D. S., Segrè, A. V, Sullivan, T. J., Young, T. R., Gelfand, E. T., … others.
668 (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene
669 regulation in humans. *Science*, *348*(6235), 648–660. article.

670 Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2016).
671 Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature*
672 *Methods*. article.

673 Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., … others. (2014). A-to-
674 I RNA editing occurs at over a hundred million genomic sites, located in a majority of
675 human genes. *Genome Research*, *24*(3), 365–376. article.

676 Beck, J. M., Young, V. B., & Huffnagle, G. B. (2012). The microbiome of the lung.
677 *Translational Research : The Journal of Laboratory and Clinical Medicine*, *160*(4),
678 258–66. http://doi.org/10.1016/j.trsl.2012.02.005

679 Blachly, J. S., Ruppert, A. S., Zhao, W., Long, S., Flynn, J., Flinn, I., … others. (2015).
680 Immunoglobulin transcript sequence and somatic hypermutation computation from
681 unselected RNA-seq reads in chronic lymphocytic leukemia. *Proceedings of the*
682 *National Academy of Sciences*, *112*(14), 4322–4327. article.
683 Brown, S. D., Raeburn, L. A., & Holt, R. A. (2015). Profiling tissue-resident T cell repertoires
684 by RNA sequencing. *Genome Medicine*, *7*(1), 1–8.
685 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden,
686 T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.
687 article.
688 Chuang, T.-J., Wu, C.-S., Chen, C.-Y., Hung, L.-Y., Chiang, T.-W., & Yang, M.-Y. (2015).
689 NCLscan: accurate identification of non-co-linear transcripts (fusion, trans-splicing
690 and circular RNA) with a good balance between sensitivity and precision. *Nucleic*
691 *Acids Research*, gkv1013. article.
692 Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., …
693 others. (2008). Stem cell transcriptome profiling via massive-scale mRNA
694 sequencing. *Nature Methods*, *5*(7), 613–619. article.
695 Consortium, Gte., & others. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis:
696 Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. article.
697 Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M., & Neretti, N. (2014).
698 Transcriptional landscape of repetitive elements in normal and cancer human cells.
699 *BMC Genomics*, *15*(1), 583. http://doi.org/10.1186/1471-2164-15-583
700 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R.

701 (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.
702 article.
703 Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., … others. (2013).
704 Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature*
705 *Methods*, *10*(12), 1185–1191. article.
706 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev,
707 A. (2011). Full-length transcriptome assembly from RNA-Seq data without a
708 reference genome. *Nature Biotechnology*, *29*(7), 644–52.
709 http://doi.org/10.1038/nbt.1883
710 https://sourceforge.net/projects/seqclean/. (n.d.). article.
711 Inman, C. F., Murray, T. Z., Bailey, M., & Cose, S. (2012). Most B cells in non-lymphoid
712 tissues are naïve. *Immunology and Cell Biology*, *90*(2), 235–242.
713 http://doi.org/10.1038/icb.2011.35
714 Jeck, W. R., & Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nature*
715 *Biotechnology*, *32*(5), 453–61. http://doi.org/10.1038/nbt.2890
716 Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). TEtranscripts: a package for
717 including transposable elements in differential expression analysis of RNA-seq
718 datasets. *Bioinformatics*, btv422. article.
719 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2:
720 accurate alignment of transcriptomes in the presence of insertions, deletions and
721 gene fusions. *Genome Biology*, *14*(4), R36. http://doi.org/10.1186/gb-2013-14-4-r36
722 Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion

723 transcripts. *Genome Biol*, *12*(8), R72. article.

724 Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G. W., Getz, G., &
725 Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep
726 sequencing of human tissue. *Nature Biotechnology*, *29*(5), 393–396. article.

727 Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., … Mason, C. E. (2014).
728 Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF
729 next-generation sequencing study. *Nature Biotechnology*, *32*(9), 915–925.
730 http://doi.org/10.1038/nbt.2972

731 Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., … others. (2014). Multi-
732 platform assessment of transcriptome profiling using RNA-seq in the ABRF next-
733 generation sequencing study. *Nature Biotechnology*, *32*(9), 915–925. article.

734 Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., … others.
735 (2015). The human transcriptome across tissues and individuals. *Science*, *348*(6235),
736 660–665. article.

737 Mihaela Pertea, J. T. M. S. L. S. (2015). StringTie enables improved reconstruction of a
738 transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*, 290–295.
739 http://doi.org/10.1038/nbt.3122

740 Nicolae, M., Mangul, S., Mandoiu, I. I., & Zelikovsky, A. (2011). Estimation of alternative
741 splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*,
742 *6*(1), 9.

743 Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and
744 opportunities. *Nature Reviews. Genetics*, *12*(2), 87–98.

745 http://doi.org/10.1038/nrg2934

746 Perucheon, S., Chaoul, N., Burelout, C., Delache, B., Brochard, P., Laurent, P., … Richard,
747 Y. (2009). Tissue-specific B-cell dysfunction and generalized memory B-cell loss
748 during acute SIV infection. *PLoS ONE*, *4*(6), e5966.
749 http://doi.org/10.1371/journal.pone.0005966

750 Poole, A., Urbanek, C., Eng, C., Schageman, J., Jacobson, S., O'Connor, B. P., … others.
751 (2014). Dissecting childhood asthma with nasal transcriptomics distinguishes
752 subphenotypes of disease. *Journal of Allergy and Clinical Immunology*, *133*(3), 670–
753 678. article.

754 Porath, H. T., Carmi, S., & Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA
755 reveals numerous new sites. *Nature Communications*, *5*, 4726.
756 http://doi.org/10.1038/ncomms5726

757 Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., … Walker,
758 A. W. (2014). Reagent and laboratory contamination can critically impact sequence-
759 based microbiome analyses. *BMC Biology*, *12*(1), 87. article.

760 Seqc/Maqc-Iii Consortium. (2014). A comprehensive assessment of RNA-seq accuracy,
761 reproducibility and information content by the Sequencing Quality Control
762 Consortium. *Nature Biotechnology*, *32*(9), 903–914.
763 http://doi.org/10.1038/nbt.2957

764 Siragusa, E., Weese, D., & Reinert, K. (2013). Fast and accurate read mapping with
765 approximate seeds and multiple backtracking. *Nucleic Acids Research*, *41*(7), e78--
766 e78. article.

767 Spreafico, R., Rossetti, M., van Loosdregt, J., Wallace, C. A., Massa, M., Magni-Manzoni,
768 S., … Albani, S. (2016). A circulating reservoir of pathogenic-like CD4+ T cells shares
769 a genetic and phenotypic signature with the inflamed synovial micro-environment.
770 *Annals of the Rheumatic Diseases*, *75*(2), 459–465. article.
771 Strauli, N. B., & Hernandez, R. D. (2016). Statistical inference of a convergent antibody
772 repertoire response to influenza vaccine. *Genome Medicine*, *8*(1), 1. article.
773 Strauli, N., & Hernandez, R. (2015). Statistical Inference of a Convergent Antibody
774 Repertoire Response to Influenza Vaccine. *bioRxiv*, 25098. article.
775 Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., … others.
776 (2008). A global view of gene activity and alternative splicing by deep sequencing of
777 the human transcriptome. *Science*, *321*(5891), 956–960. article.
778 Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … others. (2009). mRNA-
779 Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382.
780 article.
781 Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive
782 elements in genomic sequences. *Current Protocols in Bioinformatics*, 4–10. article.
783 Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter,
784 L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated
785 transcripts and isoform switching during cell differentiation. *Nature Biotechnology*,
786 *28*(5), 511–515. http://doi.org/10.1038/nbt.1621
787 Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., … Segata,
788 N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature*

789 *Methods*, *12*(10), 902–903. article.

790 Wang, X.-S., Prensner, J. R., Chen, G., Cao, Q., Han, B., Dhanasekaran, S. M., … Chinnaiyan,
791 A. M. (2009). An integrative approach to reveal driver gene fusions from paired-end
792 sequencing data in cancer. *Nature Biotechnology*, *27*(11), 1005–11.
793 http://doi.org/10.1038/nbt.1584

794 Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for
795 transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. article.

796 Wu, C.-S., Yu, C.-Y., Chuang, C.-Y., Hsiao, M., Kao, C.-F., Kuo, H.-C., & Chuang, T.-J. (2014).
797 Integrative transcriptome sequencing identifies trans-splicing events with important
798 roles in human embryonic stem cell pluripotency. *Genome Research*, *24*(1), 25–36.

799 Yan, M., Pamp, S. J., Fukuyama, J., Hwang, P. H., Cho, D. Y., Holmes, S., & Relman, D. a.
800 (2013). Nasal microenvironments and interspecific interactions influence nasal
801 microbiota complexity and S. aureus carriage. *Cell Host and Microbe*, *14*(6), 631–640.
802 http://doi.org/10.1016/j.chom.2013.11.005

803 Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable
804 domain sequence analysis tool. *Nucleic Acids Research*, gkt382. article.

805 Zhang, X.-O., Dong, R., Zhang, Y., Zhang, J.-L., Luo, Z., Zhang, J., … Yang, L. (2016). Diverse
806 alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome*
807 *Research* . http://doi.org/10.1101/gr.202895.115

808

809

810