Title : Dumpster diving in RNA-sequencing to find the source of every last read

Authors: Serghei Mangul[#1,2], Harry Taegyun Yang[1], Nicolas Strauli[3], Franziska Gruhl[4,5], Timothy Daley[6], Stephanie Christenson[7], Agata Wesolowska-Andersen[8], Roberto Spreafico[2], Cydney Rios[8], Celeste Eng[9], Andrew D. Smith[6], Ryan D. Hernandez[10,11,12], Roel A. Ophoff[13,14,15], Jose Rodriguez Santana[16], Prescott G. Woodruff[7], Esteban Burchard[9,10], Max A. Seibold*[8,17,18], Sagiv Shifman*[19], Eleazar Eskin*[1,14], 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]Department of Molecular and Computational Biology, University of Southern California, CA, USA

[7]Division of Pulmonary, Critical Care, Sleep and Allergy, Department of Medicine, and Cardiovascular Research Institute, University of California, San Francisco, CA, USA

[8]Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA

[9]Department of Medicine, University of California, San Francisco, CA, USA

[10]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

[11]Institute for Quantitative Biosciences, University of California, San Francisco, CA, USA

[12] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA

[13] Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University California Los Angeles, Los Angeles, USA

[14]Department of Human Genetics, University of California Los Angeles, Los Angeles, USA

[15]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

[16]Pediatric Pulmonology, San Juan, Puerto Rico

[17]Department of Pediatrics, National Jewish Health, Denver, CO, USA

[18]University of Colorado School of Medicine, Denver CO, USA

[19]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

Correspondence to: Serghei Mangul smangul@ucla.edu and Noah Zaitlen Noah.Zaitlen@ucsf.edu

*Equal contribution

Abstract

High throughput RNA sequencing technologies have provided invaluable research opportunities across distinct scientific domains by producing quantitative readouts of the transcriptional activity of both entire cellular populations and single cells. The majority of RNA-Seq analyses begin by mapping each experimentally produced sequence (i.e., read) to a set of annotated reference sequences for the organism of interest. For both biological and technical reasons, a significant fraction of reads remains unmapped. In this work we develop a read origin protocol (ROP) aimed at discovering the source of all reads, originated from complex RNA molecules, recombinant antibodies and microbial communities. Our approach can account for 98.8% of all reads across poly(A) and ribo-depletion protocols. Furthermore, using ROP we show that immune profiles of asthmatic individuals are significantly different from the control individuals with decreased average per sample T-cell/B-cell receptor diversity and that immune diversity is inversely correlated with microbial load. This demonstrates the potential of ROP to exploit unmapped reads to better understand the functional mechanisms underlying the connection between immune system, microbiome, human gene expression, and disease etiology.

The ROP pipeline is freely available at https://sergheimangul.wordpress.com/rop/

Introduction

Advances in RNA sequencing (RNA-seq) technology have provided an unprecedented opportunity to explore gene expression across individuals, tissues, and environments [1–3] by efficiently profiling the RNA sequences present in a sample of interest[4]. RNA-seq experiments currently produce tens of millions of short read subsequences sampled from the complete set of RNA transcripts provided to the sequencing platform. An increasing number of bioinformatic protocols are being developed to analyze reads in order to annotate and quantify the sample's transcriptome[5–7]. When a reference genome sequence, or preferably transcriptome, of the sample is available, mapping-based RNA-seq analysis protocols attempt to align the RNA-seq reads to the reference sequences, identify novel transcripts, and quantify the abundance of expressed transcripts.

A large and often-overlooked output of standard RNA-seq analyses are the unmapped reads. Even in carefully executed experiments, these "*unmapped reads*" can comprise a substantial fraction of the complete set of reads produced (e. g., 9%-15% in recent human RNA-seq projects[8–10]). Unmapped reads can arise due to technical sequencing artifacts resulting in low quality and error prone copies of the nascent RNA sequence being sampled[11]. Even when reads are error-free, aligners can fail to map some proportion of the reads for several reasons, including the following: shortcomings of the aligner's efficient yet heuristic algorithms[12], incomplete representation of all RNA transcripts in the reference set (e.g., unknown transcripts[13], variable B/T-cell receptor

sequences[14,15], trans-splicing[16], gene fusion[17], and circular RNAs[18]), and the presence of non-host RNA sequences[19] (e.g., bacterial, fungal, and viral organisms).

In this work, we aim to characterize the origin of every read obtained by RNA-seq experiments in order to inform future development of read mapping methods, provide access to additional biological information about each sample, and generally resolve the irksome puzzle of the origin of unmapped reads. To accomplish this objective, we develop the Read Origin Protocol (ROP) pipeline that leverages accurate alignment methods for both host and microbial sequences. The ROP pipeline contains a combination of existing tools focused on specific categories of *unmapped reads*[14,15,19–21], as well as novel procedures developed to overcome challenges of jointly fitting all reads. The comprehensive nature of the ROP pipeline prevents biases that can arise when using targeted analyses.

We apply the ROP pipeline to RNA-seq data from 53 asthmatic cases and 33 controls collected from three tissues, using both poly(A) selection and ribosomal RNA (rRNA) depleted library preparation protocols. The ROP analysis characterizes the origin of 98.8% of the reads compared to 83.8% by conventional reference-based protocols. We find that the vast majority of *unmapped reads* are human in origin and come from diverse sources, including repetitive elements, circular RNAs (circRNAs), gene fusion events, trans-splicing events, recombined B and T cell receptor (BCR and TCR) loci, and reads that are unmapped due to shortcomings of mapping algorithms. In addition to human RNA, a

large number of reads were microbial in origin, often occurring in sufficient numbers to study the taxonomic composition of microbial communities.

We found that both unmapped human reads and reads with microbial origins are useful in differentiating between types of tissue and disease status. For example, we found that the immune profiles of asthmatic individuals are significantly different from the controls with decreased average per sample immune diversity. Further, we used our method to show that immune diversity is inversely correlated with microbial load. This case study highlights the potential for novel discoveries without additional TCR/BCR or microbiome sequencing when the information in RNA-seq data is fully leveraged by incorporating the analysis of *unmapped reads*.

ROP - a computational protocol to explain unmapped reads in RNA-Sequencing

Mapping-based RNA-seq analysis protocols overlook reads, which fail to map onto the human reference sequences (i.e. *unmapped reads*). We have designed a read origin protocol (ROP) to identify the origin of both mapped and unmapped reads (illustrated in Fig. 1). The protocol first identifies human reads by mapping them onto a reference genome and transcriptome using a standard high-throughput mapping algorithm[22]. We used tophat v. 2.0.12 with ENSEMBL GRCh37 transcriptome and hg19 build, but many other mapping tools are available and have been recently reviewed by Engström, Pär G., et al. [23]. After alignment, reads are grouped into genomic (e.g., CDS, UTRs, introns) and

repetitive (e.g., SINEs, LINEs, LTRs) categories. The rest of the ROP protocol characterizes the remaining *unmapped reads*, which failed to map to the human reference sequences.

To process the *unmapped reads*, we apply six steps of the ROP protocol. First, we apply a quality control step to exclude low-quality, low-complexity, and reads matching rRNA repeat units among the unmapped reads (FASTQC[24], SEQCLEAN[25]). Next, we employ Megablast[26], a more sensitive alignment method, to search for human reads missed due to the heuristics implemented for computational speed in conventional aligners and reads with additional mismatches. These include reads with mismatches and short gaps relative to the reference set, but they can also include perfectly matched reads. We use a database of repeat sequences to identify lost repeat reads among the *unmapped reads*. Megablast, and similar sensitive alignment methods, are not designed to identify 'non-co-linear' RNA[20] reads from circRNAs, gene fusions, and trans-splicing events, which combine sequence from distant elements. For this task, we independently map 20bp read anchors onto the genome (see Supplementary Text). Similarly, reads from BCR and TCR loci, which are subject to recombination and somatic hyper-mutation (SHM), require specifically designed methods, and we use IgBlast[27] for this purpose. The remaining reads that did not map to any known human sequence and are potentially microbial in origin. We use microbial genomes and phylogenetic marker genes to identify microbial reads and assign them to corresponding taxa[28]. Microbial reads could have been introduced by contamination or the natural microbiome of the sample, which includes viral, bacterial, or other microbial species[29].

Taken together, ROP considers five classes of *unmapped reads*: (1) lost human reads, (2) lost repeat elements (3) reads from 'non-co-linear' (NCL) RNAs, (4) reads from recombinations of BCR and TCR segments (i.e. V(D)J recombination), and (5) microbial reads. While individual methods have been previously proposed to examine some of these classes [14,15,19–21], we find that performing a joint analysis, in the order described above, is critical in order to minimize misclassification of reads due to homologous sequences between the different classes. Furthermore, as shown in the Results section below, only a comprehensive pipeline allows analysis across these classes. Complete details of ROP, including all parameters and threshold used, are provided in the Supplementary Text.

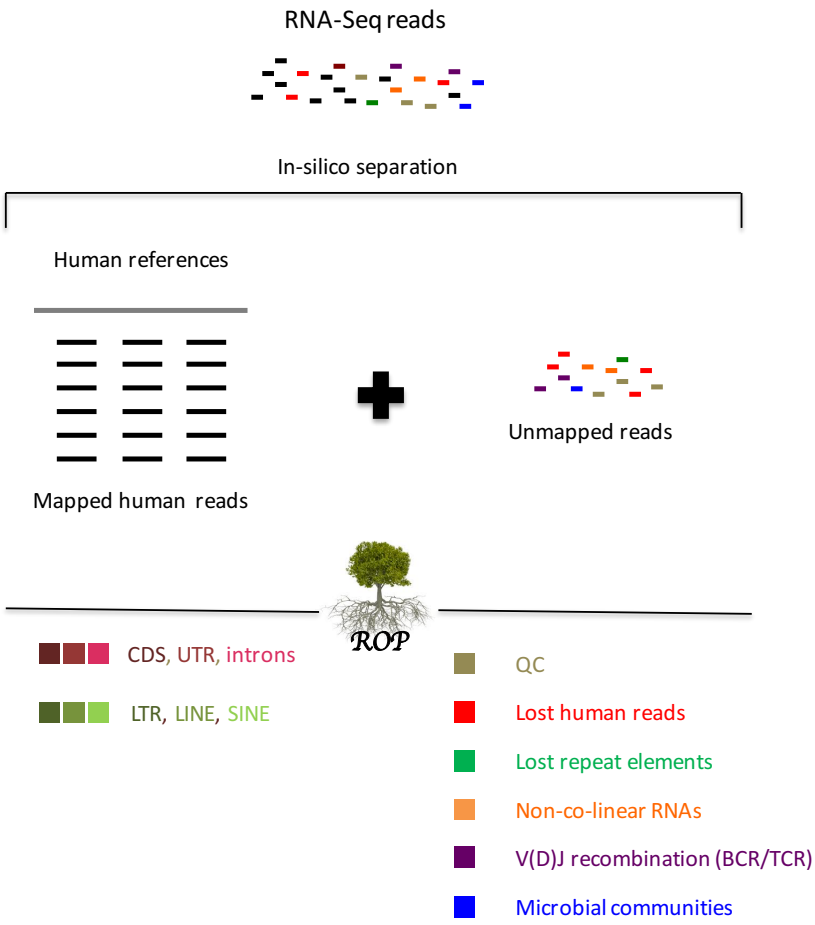

Figure 1. Schematic of the Read Origin Protocol (ROP). Human reads are identified by mapping all reads onto the reference sequences using a standard high-throughput mapping algorithm. ROP protocol categorizes mapped reads into genomic (red colors) and repetitive (green colors) categories. Unmapped reads that fail to map are extracted and further filtered to exclude low quality reads, low complexity reads, and reads from rRNA repeats (brown color). ROP protocol is able to identify unmapped reads aligned to human references using a more sensitive alignment tool (lost human reads: red color), unmapped reads aligned to the repeat sequences (lost repeat elements: green color), unmapped reads spanning sequences from distant loci (non-co-linear: orange color), unmapped reads spanning antigen receptor gene rearrangement in the variable domain (V(D)J recombination of BCR and TCR: violet color), and unmapped reads aligned to the microbial reference genomes and marker genes (microbial reads: blue color).

The ROP protocol is able to account for 98.8% of all reads

To test ROP, we applied it to RNA-Seq performed on RNA from 86 individuals (53 asthmatics and 33 controls), from three tissues: peripheral blood, nasal and large airway epithelium/ RNA-seq libraries were prepared from total RNA with two types of RNA enrichment methods: (1) Poly(A) enrichment libraries, applied to RNA from peripheral blood and nasal epithelium (n=38), and (2) ribo-depletion libraries, applied to RNA from large airway epithelium (n=49). The RNA-Seq was based on paired-end protocol with 100bp reads. In total, 3.8 billion paired-end reads (760 Gbp) were available for ROP (Table S1 and Supplementary Text).

Our initial high-throughput mapping using tophat2[22] recovered 83.8% of all reads (Fig. 2a). From the *unmapped reads*, we first excluded low-quality/low-complexity reads and reads mapping to the rRNA repeating unit, which together accounted for 4.8% and 3.6% of all reads respectively (Fig. 2b). We were then able to align unmapped reads to human reference sequences (6.1% of all reads, Fig. 2c), reference repeat sequences (0.02% of all reads, Fig. 2d), reads identified as 'non-co-linear'(NCL) RNAs (circRNAs, gene fusion or trans-splicing) (0.1% of all reads, Fig. 2e), and reads mapped to recombined sequences of BCR and TCR loci (0.01% off all reads, Fig. 2f). The remaining reads were mapped to the microbial sequences (0.3% off all reads, Fig. 2g). Following the six steps of ROP, the origin of 98.8% of reads was identified.

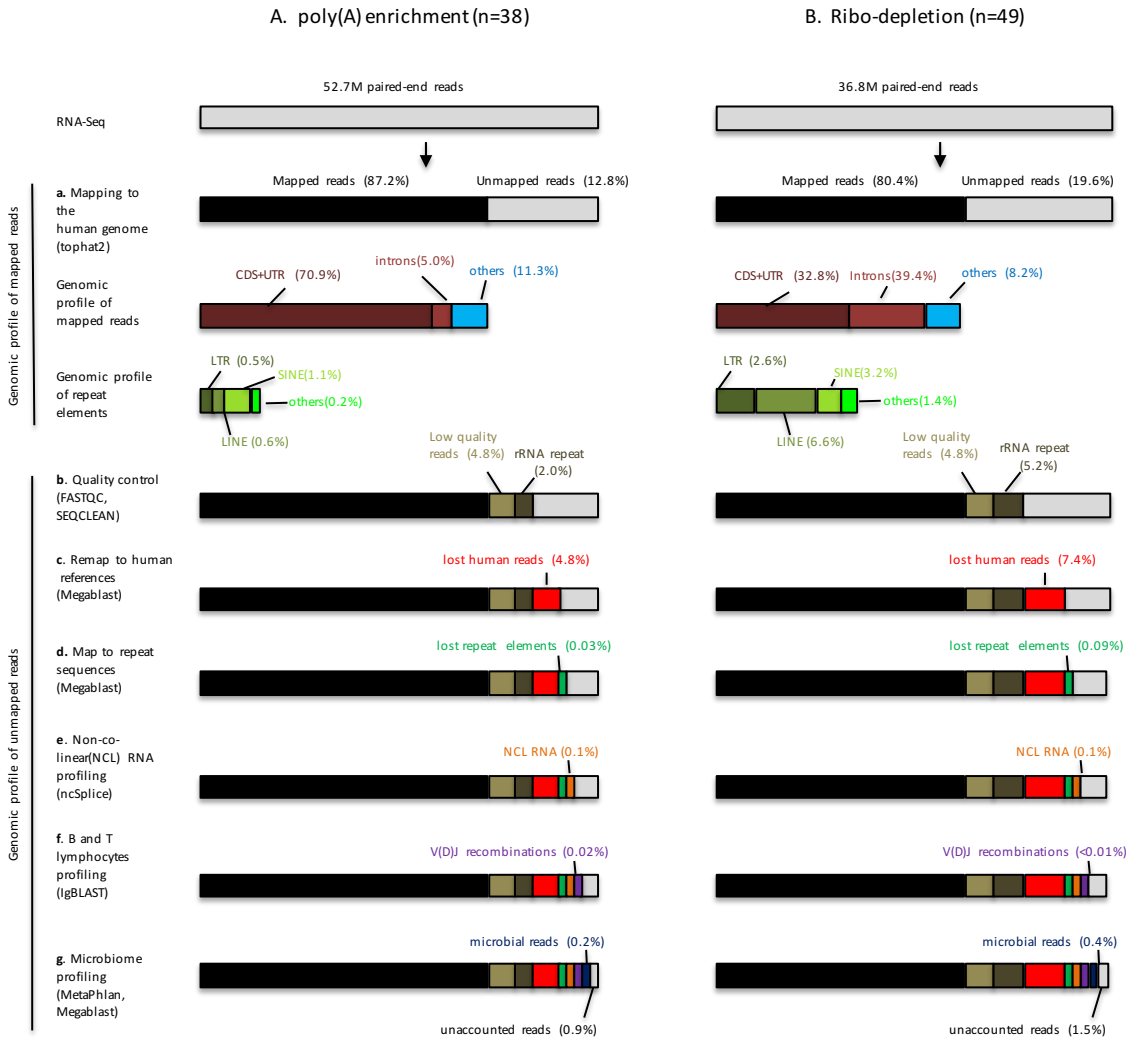

Figure 2. The percentage of reads mapping to various classes for poly(A) enrichment and ribo-depletion libraries. Percentages are calculated as a fraction from the total number of reads. Bars of the plot are not scaled. RNA-Seq libraries were prepared from total RNA using poly(A) enrichment (n=38) and ribo-depletion (n=49) protocols. (a) Human reads (black color) mapped to reference genome and transcriptome via tophat2. Mapped reads are categorized into genomic (CDS, UTR, introns) and repetitive (SINEs, LINEs, LTRs) categories. Unmapped reads are shown in grey. (b) Low quality/low-complexity (light brown) and reads matching rRNA repeating unit (dark brown) were excluded. (c) ROP identifies lost human reads (red color) from unmapped reads using a more sensitive alignment (d) ROP identifies lost repeat sequences (green color) by mapping unmapped reads onto the reference repeat sequences (e) Reads arising from trans-

spicing, gene fusion and circRNA events (orange color) are captured by a custom pipeline (nclSplice) (f) IgBlast is used to identify reads spanning BCR/TCR receptor gene rearrangement in the variable domain (V(D)J recombinations) (violet color). (g) Microbial reads (blue color) are captured by mapping the reads onto the microbial reference genomes and phylogenetic marker genes.

The ROP protocol can identify lost human reads and complement transcriptome profiling by non-co-linear RNAs

The heuristic nature of high throughput aligners limits their ability to map all the human reads onto the reference genome[12,31]. We use the slower yet more sensitive Megablast aligner, which allows extraction of an additional 6.1% of human reads. One-fourth of the lost human reads are within the tophat2 threshold (edit distance less or equal to 2). Other reads missed by tophat2 contained additional mismatches and/or short gaps (Fig. S1).

We first compared the composition of mapped and lost human reads across libraries by categorizing human reads into genomic categories (e.g., CDS, UTR, intons). Similar to Li, S. *et al*[32] we observed that library preparation has a strong effect on the fraction of both mapped and lost human reads mapping to CDS and intronic regions (Fig. S2). ROP also identifies and categorizes repetitive sequences among the mapped and unmapped reads. Consistent with repEnrich[35], we observe the differences in proportions of L1 and Alu elements between poly(A) and ribo-depletion libraries. Among the repeat reads, poly(A) samples have the highest fraction of reads mapped to Alu elements, ribo-depleted samples have the highest fraction mapped to L1 elements. We consistently observed this effect in the unmapped reads (Fig. S3). However, the tissue type (whole

blood versus nasal epithelium) appears to have no effect on the genomic profiles and repeat profiles (Tables S2, S3).

The ROP protocol is able to detect 'non-co-linear' reads from three classes of events: reads spliced distantly on the same chromosome supporting trans-splicing events; reads spliced across different chromosomes supporting gene fusion events; and reads spliced in a head-to-tail configuration supporting circRNAs. We observed 471 trans-splicing, 1732 fusion, and 1268 circular events on average per individual sample supported by more than one read. Over 90% of non-co-linear events were supported by fewer than 10 samples (Fig. S4). We used a liberal threshold because our interest is mapping all reads. However, a more stringent cut off, as has been previously used[36], is recommended for confident identification of non-co-linear events. The library preparation technique had a strong effect on capture rate of non-co-linear transcripts. To compare number of NCL events, we sub-sampled unmapped reads to 4,985,914 for each sample, which corresponded to the sample with the smallest number of unmapped reads. We observe an average increase of 46% of 'non-co-linear' (NCL) events detected in samples prepared by ribo-depletion compared to poly(A) protocol ($p\text{-value} < 8 \times 10^{-3}$) (Fig. S5). However, because the tissues differed between protocols (nasal versus large airway epithelium), this effect might be due in part to tissue differences in NCL events. We view this possibility as unlikely given the differences in RNA sampled from each protocol. There were no statistically significant differences between NCL events in cases and controls.

ROP can identify microbial and immune reads and differentiate tissue types and disease status

Both mapped and unmapped reads were used to survey the human antibody repertoire in health and disease. We first used the mapped reads to extract reads entirely aligned to BCR and TCR genes. Using IgBlast[27], we identified unmapped reads with extensive somatic hyper mutations (SHM) and reads arising from V(D)J recombination. After we identified the reads with human origin, we found microbial reads by mapping the remaining reads onto the microbial reference genomes and phylogenetic marker genes. Here, the total number of microbial reads obtained from the sample is used to estimate microbial load. We use MetaPhlAn2[28] to assign reads on microbial marker genes and determine the taxonomic composition of the microbial communities.

We compare immunological and microbial profiles across asthmatics and unaffected controls for the three tissues. A total of 30 bacterial taxa were assigned with Metaphlan2[28]. Consistent with previous studies, we observe the nasal epithelium is dominated by *Actinobacteria phyla* (particularly *Propionibacterium genus*)[37], and the large airway epithelium is dominated by *Proteobacteria phyla* [38] (Table S4).

As expected, analysis of blood tissue produced the highest fraction of immune reads (blood: 5075 immune reads per million reads (RPM); nasal: 114 RPM; and large airway: 17 RPM) (Table S5). The overall immune profiles suggest differences in fraction of BCR and TCR reads across tissues. Blood samples show a larger fraction of reads mapped

to BCR genes, while nasal and large airway epithelium produce a larger fraction of reads mapped to TCR genes (Fig. S6). Consistent with previous findings, we observe an increased fraction of reads mapped to IgM (produced by B cells) in blood relative to other tissues[39,40] (Fig. S7).

We assess combinatorial diversity of the antibody repertoire by looking at the recombinations of the of VJ gene segments of BCR and TCR loci. The blood yields increased number of combinations of gene segments, with 191 combinations, on average, per sample for immunoglobulin kappa locus (IGK) (Fig. 3.a). Similar to NCL analysis, unmapped reads were sub-sampled to 4,985,914 for each sample corresponding to a sample with smallest number of unmapped reads. We used per sample alpha diversity (Shannon entropy) to incorporate the total number of VJ combinations and their relative proportions into a single diversity metric. We observed a mean alpha diversity of 4.2 for blood, 2.5 for nasal, and 1.0 for large airway (Fig. S8). Decreased alpha diversity in large airway samples could correspond to an overall decrease in percentage of immune reads. This effect can be attributed to the ribo-depletion protocol not enriching for polyadenylated antibody transcripts. Alternatively, it may result from clonal expansion of certain clonotypes responding to the cognate antigen.

Our comprehensive ROP protocol presents several advantages over previous methods examining features of unmapped reads. First, our method is able to interrogate relationships between features. To explore the interaction between the immune system

and microbiome, we compared immune diversity against microbial load. Microbes trigger immune responses, eliciting proliferation of antigen-specific lymphocytes. Such a dramatic expansion skews the antigen receptor repertoire in favor of a few dominant clonotypes and decreases immune diversity[41]. Therefore, we reasoned that, in the presence of microbial insults, antigen receptor diversity should shrink. In line with our expectation, we observed that combinatorial immune diversity of IGK locus was negatively correlated with the viral load (Pearson coefficient r=-0.55, p-value =2.4 x 10^{-6}), consistent also for bacteria and eukaryotic pathogens across BCR and TCR loci.

We also compared the combinatorial immune diversity of asthmatic individuals (n=9) and healthy controls (n=10). The combinatorial profiles of antibody receptors in blood and large airway tissue provide no differentiation between case control statuses. Among nasal samples, we observed decreased alpha diversity for asthmatic individuals relative to healthy controls (p-value=10^{-3}) (Fig. 3.b). Additionally, we used beta diversity (Sørensen–Dice index) to measure compositional similarities between samples, including gain or loss of VJ combinations of IGK locus. We observed higher beta diversity corresponding to a lower level of similarity across the nasal samples of asthmatic individuals in comparison to samples from unaffected controls (Fig. 3.c, p-value<3.7x10^{-13}). Moreover, nasal samples of unaffected controls are significantly more similar to each other than to samples from the asthmatic individuals (Fig. 3.c, p-value<2.5x10^{-9}). Recombination profiles of immunoglobulin lambda locus (IGL) and T cell receptor beta and gamma (TCRB and TCRG) loci yielded a similar pattern of decreased beta diversity

across nasal samples of asthmatic individuals (Fig. S9-S11). Together the results demonstrate the ability of ROP to interrogate additional features of the immune system without the addition of expensive TCR/BCR sequencing.

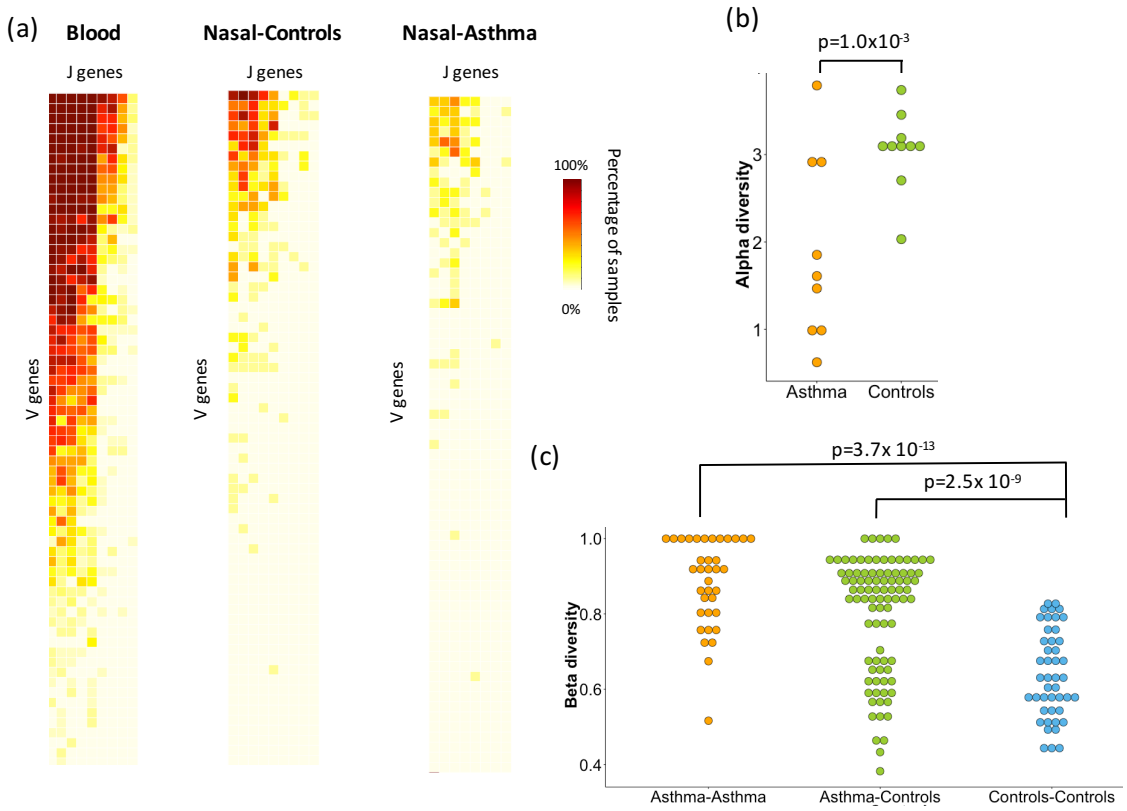

Figure 3. Combinatorial diversity of immunoglobulin kappa locus (IGK) locus differentiates disease status. (a) Heatmap depicting the percentage of RNA-Seq samples supporting of particular VJ combination for whole blood (n=19), nasal epithelium of healthy controls (n=10) and asthmatic individuals (n=9). Each row corresponds to a V gene, and each column correspond to a J gene. (b) Alpha diversity of nasal samples is measured using the Shannon entropy and incorporates total number of VJ combinations and their relative proportions. Nasal epithelium of asthmatic individuals exhibits decreased combinatorial diversity of IGK

locus compared to healthy controls (p-value=$1x10^{-3}$) (c) Compositional similarities between the nasal samples in terms of gain or loss of VJ combinations of IGK locus are measured across pairs of sample from the same group (Asthma, Controls) and pairs of sample from different groups (Asthma versus Controls) using Sørensen–Dice index. Lower level of similarity is observed between nasal samples of asthmatic individuals compared to unaffected controls (p-value<7.3×10^{-13}). Nasal samples of unaffected controls are more similar to each other than to the asthmatic individuals (p-value<2.5×10^{-9}).

Discussion

Our study is the first that systematically accounts for almost all reads in RNA-seq studies. We demonstrate the value of analyzing unmapped reads present in the RNA-seq data to study the non-co-linear, immunological and microbiome profiles of a tissue of interest. We developed a new ROP pipeline that leverages accurate alignment methods and accounts for 98.8% of the reads, compared to the 83.8% rate produced by conventional reference-based protocols. The 'dumpster diving' profile of unmapped reads output by our method is not limited to RNA-Seq technology and may apply to whole-exome and whole-genome sequencing. We anticipate that 'dumpster diving' profiling will find broad future applications in studies involving different tissue and disease types.

We observed large effects of library preparation protocol on non-co-linear, immunological, and microbial profiles. For example, the poly(A) protocol can better capture antibody transcripts by enriching for polyadenylated transcript. Ribo-depletion

protocol is able to capture more non-co-liner events. The results presented here may further guide the choice of protocol depending on the features of interest.

The ROP protocol identified additional human reads missed by the high-throughput aligners. Alternative mapping algorithms may produce small differences in the number of mapped reads[23], but assessing their relative performances is a task beyond the scope of this work. The majority of the lost human reads contained additional mismatches that do not easily align to the genome and are important for RNA editing and screening[42]. This suggests that additional power can be gained by improving the mapping accuracy of current alignment algorithms.. Our tool provides a genomic profile of the mapped reads, which can be used to benchmark different sequencing platform and library preparation methods, as well as assess the efficiency of rRNA depletion and level of the sample degradation[33].

The ROP protocol facilitates a simultaneous study of immune systems and microbial communities that advances our understanding of the functional, interrelated mechanisms driving the immune system, microbiome, human gene expression, and disease etiology. In particular, we hope that these future efforts will provide a quantitative and qualitative assessment of the immune and microbial components of disease across various tissues. With an increase in length of reads and efficiency of sequencing, there is also the potential for studying individual microbial species and full TCR/BCR sequencing.

References

1. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (80-.).* **321,** 956–960 (2008).

2. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5,** 613–619 (2008).

3. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).

4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

5. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

6. Nicolae, M., Mangul, S., Mandoiu, I. I. & Zelikovsky, A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.* **6,** 9 (2011).

7. Mihaela Pertea, J. T. M. S. L. S. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).

8. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

9. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32,** 915–925 (2014).

10. Seqc/Maqc-Iii Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control

Consortium. *Nat. Biotechnol.* **32,** 903–914 (2014).

11. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98 (2011).

12. Siragusa, E., Weese, D. & Reinert, K. Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Res.* **41,** e78–e78 (2013).

13. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).

14. Blachly, J. S. *et al.* Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci.* **112,** 4322–4327 (2015).

15. Strauli, N. & Hernandez, R. Statistical Inference of a Convergent Antibody Repertoire Response to Influenza Vaccine. *bioRxiv* 25098 (2015).

16. Wu, C.-S. *et al.* Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res.* **24,** 25–36 (2014).

17. Wang, X.-S. *et al.* An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat. Biotechnol.* **27,** 1005–11 (2009).

18. Jeck, W. R. & Sharpless, N. E. Detecting and characterizing circular RNAs. *Nat. Biotechnol.* **32,** 453–61 (2014).

19. Kostic, A. D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29,** 393–396 (2011).

20. Chuang, T.-J. *et al.* NCLscan: accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Res.* gkv1013 (2015).

21. Brown, S. D., Raeburn, L. A. & Holt, R. A. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Med.* **7,** 1–8 (2015).

22. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

23. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).

24. Andrews, S. & others. FastQC: A quality control tool for high throughput sequence data. *Ref. Source* (2010).

25. https://sourceforge.net/projects/seqclean/.

26. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

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

28. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (2015).

29. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).

30. Poole, A. *et al.* Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *J. Allergy Clin. Immunol.* **133,** 670–678

(2014).

31. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7,** 576–577 (2010).

32. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32,** 915–925 (2014).

33. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10,** 623–629 (2013).

34. Sultan, M. *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15,** 675 (2014).

35. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15,** 583 (2014).

36. Carrara, M. *et al.* State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC Bioinformatics* **14,** 1 (2013).

37. Yan, M. *et al.* Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and S. aureus carriage. *Cell Host Microbe* **14,** 631–640 (2013).

38. Beck, J. M., Young, V. B. & Huffnagle, G. B. The microbiome of the lung. *Transl. Res.* **160,** 258–66 (2012).

39. Perucheon, S. *et al.* Tissue-specific B-cell dysfunction and generalized memory B-cell loss during acute SIV infection. *PLoS One* **4,** e5966 (2009).

40. Inman, C. F., Murray, T. Z., Bailey, M. & Cose, S. Most B cells in non-lymphoid tissues are naïve. *Immunol. Cell Biol.* **90,** 235–242 (2012).

41. Spreafico, R. *et al.* A circulating reservoir of pathogenic-like CD4+ T cells shares a genetic and phenotypic signature with the inflamed synovial micro-environment. *Ann. Rheum. Dis.* **75,** 459–465 (2016).

42. Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* **5,** 4726 (2014).