As described in Figure 1, the Gerstein lab has developed a sophisticated toolkit called exceRpt that preprocesses, aligns, and quantitates small and long RNA-seq datasets. As an input, exceRpt accepts “.fq” and “.sra” files. In the first step, the toolkit maps the raw reads to UniVec contaminants and to 45S, 5S, and mitochondrial ribosomal RNAs. The unmapped reads are then aligned (STAR alignments) to the human and mouse genome, as well as to small RNA databases such as miRbase, gtRNAdb, and RNAdb. After the raw reads are mapped onto endogenous genomes, repetitive elements, and endogenous and exogenous small RNAs, the remaining unmapped reads are mapped to all genomes in Ensembl and NCBI. These genomes include vertebrates, fungi, bacteria, and viruses. The exceRpt output reports both the discovered exogenous species and the respective sequences found in the sample.



**Figure 1. Analysis of non-human RNA across different tissues and individuals.** exceRpt -a mapping pipeline that has been developed by the Gerstein lab- considers reads in FASTQ format. After filtering for contaminants and low-quality reads, exceRpt consequently maps raw reads across different categories of mitochondrial, small, long, endogenous or exogenous RNA, including all genomes in ensemble and NCBI.

To explore the detection capabilities of our pipeline, we applied exceRpt on the EN-TEx dataset. EN-TEx consists of 91 total RNA-seq assays from 31 human tissues and four individuals. Overall, after filtering for contaminants and low-quality reads, around 5% of raw reads remain unmapped to the human genome and small RNA libraries. From these reads, about 0.5% of the total reads would consequently map to exogenous genomes. Considering that our input raw reads consist of about 50,000,000 reads, the exogenous genome reads correspond to ~250,000 reads per sample. Using two pair-end fastq files for two technical replicates, exogenous reads correspond to about 1,000,000 reads per tissue, per individual. Similar to other publications, our preliminary results indicate that RNA microbial sequences can be detected in RNA samples. For stomach tissues, our preliminary analysis identified ~380 different species, whereas for intestine tissues we identified 402 species. Additionally, we identified the presence of ~60 viruses. As an example, in the stomach tissue of a 37-year-old male individual, we identified 248 raw RNA sequences of *Human herpesvirus 7*.

To test the LDA-link algorithm preliminarily, we analyzed samples collected from human bodies. LDA-link identified connections between genes and microbes reported elsewhere in the literature as well as novel observations. A bipartite graph summarizing a subset of the connections between genes and microbes shows that in most cases several genes are linked to each microbe. Several of the observed links showed strong literature precedence. For example, the gene lactotransferrin was linked to Aeromonas (1), Burkholderia was linked to gene MUC6 (2), Haemophilus was linked to NFKB Inhibitor Zeta (3), and Pasteurella was linked to IL1B (4). To the best of our knowledge, several novel links were also found, such as between Haemophilus and IL1B, and between Candida and GCSAML. In addition to gene-microbe pairs, we layered on pathway and cell deconvolution data to identify larger-scale effects of microbes.

Microbes were linked to genes that are enriched in pathways relating to auto-immunity and inflammation as well as cytokine receptors and their interactions. The microbes associated with cytokine pathways included Synechococcus, Lactococcus, Dialister, Psychrobacter, Moraxella, Brenneria, Proteus, Haemophilus, and Pasteurella. In addition, we related the cell-type signatures to identify the immune cell types that are related to each microbe. We observed the Haemophilus-IL1B link in monocytes and mast cells. Samples containing Haemophilus triggered more activated mast cells according to its cell fraction (5-8). Similarly, the fungal genus Candida was linked to the gene GCSAML, which was highly expressed by eosinophils. The presence of Candida was associated with increased numbers of Eosinophils in the airway.

**REFERENCE**

1. Ascencio F, Ljungh A, Wadstrom T. Characterization of lactoferrin binding by Aeromonas hydrophila. Appl Environ Microbiol. 1992;58(1):42-7. Epub 1992/01/01. PubMed PMID: 1311545; PMCID: PMC195170.

2. Sajjan U, Keshavjee S, Forstner J. Responses of well-differentiated airway epithelial cell cultures from healthy donors and patients with cystic fibrosis to Burkholderia cenocepacia infection. Infect Immun. 2004;72(7):4188-99. Epub 2004/06/24. doi: 10.1128/IAI.72.7.4188-4199.2004. PubMed PMID: 15213163; PMCID: PMC427436.

3. Park CY, Heo JN, Suk K, Lee WH. Sodium azide suppresses LPS-induced expression MCP-1 through regulating IkappaBzeta and STAT1 activities in macrophages. Cell Immunol. 2017;315:64-70. Epub 2017/04/11. doi: 10.1016/j.cellimm.2017.02.007. PubMed PMID: 28391993.

4. Hildebrand D, Bode KA, Riess D, Cerny D, Waldhuber A, Rommler F, Strack J, Korten S, Orth JH, Miethke T, Heeg K, Kubatzky KF. Granzyme A produces bioactive IL-1beta through a nonapoptotic inflammasome-independent pathway. Cell Rep. 2014;9(3):910-7. Epub 2014/12/02. doi: 10.1016/j.celrep.2014.10.003. PubMed PMID: 25437548.

5. Chapman SJ, Khor CC, Vannberg FO, Rautanen A, Segal S, Moore CE, Davies RJ, Day NP, Peshu N, Crook DW, Berkley JA, Williams TN, Scott JA, Hill AV. NFKBIZ polymorphisms and susceptibility to pneumococcal disease in European and African populations. Genes Immun. 2010;11(4):319-25. Epub 2009/10/03. doi: 10.1038/gene.2009.76. PubMed PMID: 19798075; PMCID: PMC3051152.

6. Baldwin AS, Jr. The NF-kappa B and I kappa B proteins: new discoveries and insights. Annu Rev Immunol. 1996;14:649-83. Epub 1996/01/01. doi: 10.1146/annurev.immunol.14.1.649. PubMed PMID: 8717528.

7. Motoyama M, Yamazaki S, Eto-Kimura A, Takeshige K, Muta T. Positive and negative regulation of nuclear factor-kappaB-mediated transcription by IkappaB-zeta, an inducible nuclear protein. J Biol Chem. 2005;280(9):7444-51. Epub 2004/12/25. doi: 10.1074/jbc.M412738200. PubMed PMID: 15618216.

8. Yamazaki S, Muta T, Matsuo S, Takeshige K. Stimulus-specific induction of a novel nuclear factor-kappaB regulator, IkappaB-zeta, via Toll/Interleukin-1 receptor is mediated by mRNA stabilization. J Biol Chem. 2005;280(2):1678-87. Epub 2004/11/04. doi: 10.1074/jbc.M409983200. PubMed PMID: 15522867.