# Examining lost reads to survey the microbiome and immune components of the human body across 43 human sites from 175 individuals

Serghei Mangul[1], Nicolas Strauli[2], Ryan Hernandez[2], Roel Ophoff[3], Eleazar Eleazar Eskin[1,3], Noah Zaitlen[4]

[1]UCLA, Computer Science, Los Angeles, CA, [2]UCSF, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, [3]UCLA, Human Genetics, Los Angeles, CA, [4]UCSF, Department of Medicine, San Francisco, CA

Contact: smangul@ucla.edu

Advances in RNA sequencing technology and the ability to generate deep coverage data in the form of millions of reads provide an unprecedented opportunity to probe the universe of gene expression. Standard RNA-seq analysis protocols map reads against a host reference genome to determine the placement of the reads on the genome. Mapping-based protocols are complemented by assembly procedures to accurately profile the origin of reads condensed into isoform transcripts. Many reads are discarded by these protocols and the possibility that reads originate outside of the extant genome is usually ignored. In this work we aim to profile the origin of every last read delivered by RNA sequencing, in order to identify shortcomings of existing technologies as well as identify novel uses of RNA-Seq data. Our study reveals that the vast majority of unmapped reads are human reads discarded by the mapping protocol. Many unmapped human reads correspond to novel exon junctions from previously unknown isoforms. Another significant source of discarded human reads are sequences originating from the recombined Ig locus of B and T lymphocytes (BCR and TCR sequences). In addition to human DNA, the human body harbors a diverse microbial ecosystem, and we identified a substantial number of reads mapping to non-human sequence. Careful analysis of the BCR and TCR sequences in conjunction with the microbial communities provides an opportunity to profile immune system function across tissues directly from RNA-seq data.

We use 1641 RNA-Seq samples corresponding to 175 individuals and 43 sites from GTEx project: 29 solid organ tissues, 11 brain subregions, whole blood, and two cell lines, LCL and cultured fibroblasts from skin. Illumina Hiseq 2000 platform was used to produce Illumina RNA-Seq data sets. RNA-Seq libraries were prepared from total RNA using poly(A) enrichment of the mRNA. We use the unmapped reads to obtain a detailed profile of the microbial and immune components of the human body (unmapped reads were extracted from the .bam files download from the gtex storage). Wee obtained 6.77 +-1.60 million 76bp unmapped reads per sample. First we filtered out 54.68%+7.28% of the unmapped reads, which were low-quality and/or low complexity (using FASTX and SEQCLEAN). We attempted realignment of remaining reads to the human reference sequences using the bowtie2 aligner (up to 10 mismatches were allowed). Bowtie2 was able to identify 33.18%+-4.82% of the reads compatible with human genome and transcriptome reference (ENSEMBL hg19 build, ENSEMBL GRCh37 transcripome).

The remaining high-quality unique reads are used to perform a survey of the microbiome and immune components. We used phylogenetic marker genes to assign candidate microbial reads to the bacterial and archeal taxa. We use Phylosift to perform taxonomic classification of the samples and compare it across the tissues. The Phylosift approach uses hypervariable taxa-specific gene families to provide the precise resolution for the bacteria and archaea community assemblages. Hyper-variable regions from gene families are previously identified to be nearly universal and present in a single copy allowing differentiating between species and taxa. Reads are also mapped to a reference database of viral (n = 1,401), bacterial (n = 1,980) and fungal (n = 32) genomes downloaded from NCBI. To study the distribution of of B and T cells cross individuals and tissues we use reads mapped to the V(D)J regions of the Ig loci. Those recombinations correspond to early stages of T and B cell maturation.

A total of 713 taxa were assigned with Phylosift, with 8 taxa on the phylum level. Most of the taxa we observe are bacterial and a smaller portion is archeal .We observed no evidence of the presence of nonhuman eukaryotes. We observe all tissues to be dominated by Proteobacteria. No microbial organisms were observed in heart, pituitary and adrenal gland. All other tissues contain at least onebacterial or archeal phyla (0.79+-0.55 phyla per sample ). We observe two viruses harbored in multiple tissues. EBV virus is present in 20% of the skin samples and 66% of the liver samples and it is not present in any of the brain samples. Enterobacteria phage phiX174 virus is present in 20% of the skin samples and is not present in liver and brain tissues.

Examining immune and microbial genes in GTEx can help define typical profiles for a healthy tissue. It is essential to monitor microbial and immune diversity, and this work may eventually help diagnose immune and microbiome imbalance in a tissue specific manner.