Pseudo-Alignment with kallisto

Efficient Probabilistic Quantification of RNA-Seq Reads

Nima Hejazi
Division of Biostatistics
University of California, Berkeley
nh@nimahejazi.org

1 Pseudo-Alignment of RNA-Seq Reads

The present project concerns the re-analysis of transcriptomic data from the study described in the paper "Developmental regulation of human cortex transcription and its clinical relevance at base resolution", Jaffe *et al.*, *Nature Neuroscience*. In order to quantify RNA-Seq reads, alignment against a reference transcriptome must be performed, a procedure which results in tables of read counts for use in downstream statistical analysis. Here, we take advantage of **pseudo-alignment**, a novel development in sequencing algorithms, to probabilistically align reads. Below, we describe pseudo-alignment and the results of its application to the Jaffe *et al.* data.

1.1 The Pseudo-Alignment Process

Pseudo-alignment is a novel process for quantifying a set of samples of RNA-Seq reads by performing partial matching against a reference transcriptome. The novel pseudo-alignment process, implemented in the command line tool **kallisto**, takes into account all of the information contained in a set of reads while reducing the computational burden imposed by more traditional alignment techniques. The **kallisto** tool provides results similar to that produced by other alignment software (*e.g.*, **bowtie**), while taking only a fraction of the time. For a complete description of pseudo-alignment, consult the paper "Near-optimal probabilistic RNA-seq quantification", Bray *et al.*, *Nature Biotechnology*.

1.2 The Results of Pseudo-Alignment

The pseudo-alignment procedure was implemented on the Jaffe *et al.* data through the use of the **kallisto** command line tool. Using a publicly available trascriptome assembled from the GRCh38 (hg19) *Homo sapiens* genome, sets of paired-end RNA-Seq reads for each of the 12 subjects involved in the study were pseudo-aligned, resulting in count tables mapping each set of reads to **173,259** transcriptomic objects. Tables of counts are produced for each set of paired-end RNA-Seq reads for each subject in a tab-separated file format; these files are suitable for concatenation into a count table for all subjects, which can be subjected to statistical analysis after appropriate data cleaning. For reference, alongside this document, a sample file (in JSON format) indicating the results of running **kallisto** on data from a single subject is provided.