Kallisto-0.42.3-INDEX-QUANT-PE in the Discovery Environment

Introduction and Overview

kallisto is a program for quantifying abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads. It is based on the novel idea of *pseudoalignment* for rapidly determining the compatibility of reads with targets, without the need for alignment. On benchmarks with standard RNA-Seq data, kallistocan quantify 30 million human reads in less than 3 minutes on a Mac desktop computer using only the read sequences and a transcriptome index that itself takes less than 10 minutes to build. Pseudoalignment of reads preserves the key information needed for quantification, and kallisto is therefore not only fast, but also as accurate than existing quantification tools. In fact, because the pseudoalignment procedure is robust to errors in the reads, in many benchmarks kallistosignificantly outperforms existing tools. kallisto quantified RNA-Seq can be analyzed with sleuth.

This tutorial is for using Kallisto workflow that includes index and quantification. (Please visit http://pachterlab.github.io/kallisto/manual.html for the manual.)

Input Data:

Data taken from the "Cuffdiff2 paper"

> *Differential analysis of gene regulation at transcript resolution with RNA-seq by Cole Trapnell, David G Henderickson, Martin Savageau, Loyal Goff, John L Rinn and Lior Pachter, Nature Biotechnology 31, 46–53 (2013).*

The human fibroblast RNA-Seq data for the paper is available on GEO at accession GSE37704. The samples to be analyzed are the six samples LFB_scramble_hiseq_repA, LFB_scramble_hiseq_repB, LFB_scramble_hiseq_repC, LFB_HOXA1KD_hiseq_repA, LFB_HOXA1KD_hiseq_repA, and LFB_HOXA1KD_hiseq_repC. These are three biological replicates in each of two conditions (scramble and HoxA1 knockdown) that will be compared with sleuth.

HOXA1 is a critical regulator of embryonic development and body patterning, in maintaining adult cells.HOXA1 knockdown perturbs the expression of thousands of genes

run_accession	experiment_accession	spots	condition	sequencer	sample
SRR493366	SRX145662	15117833	scramble	hiseq	A
SRR493367	SRX145663	17433672	scramble	hiseq	B
SRR493368	SRX145664	21830449	scramble	hiseq	C
SRR493369	SRX145665	17916102	HOXA1KD	hiseq	A
SRR493370	SRX145666	20141813	HOXA1KD	hiseq	B
SRR493371	SRX145667	23544153	HOXA1KD	hiseq	C

Kallisto RNA seq analysis using workflow

The kallisto workflow is quite simple.

Open Kallisto-0.42.3-INDEX-QUANT-PE (Apps > Public Apps > Kallisto-0.42.3-INDEX-QUANT-PE)

1. **Index name:** Enter a Index name (**"human_trans"**)
2. **Fasta file:** Load your Reference genome in fasta format (Community Data -> iplantcollaborative -> example_data -> kallisto -> Human -> Index -> ***"Homo_sapiens.GRCh38.cdna.all.fa")***
3. Optional argument (**k-mer (odd) length**): For now you can leave the default k-mer size (k=31). Note, that if you have very short reads (e.g. 35bp), you should change k to something smaller (e.g. `-k 21`).
4. **Output directory:** Enter the name of the output directory (default is **"myoutput"**)
5. **Input Read1&Read2 fastq files:** Select fastq files - Community Data -> iplantcollaborative -> example_data -> kallisto -> Human -> Reads -> ***"SRR493366.sra_1.fastq SRR493366.sra_2.fastq"***
6. Optional arguments:
    a. Estimated average fragment length: Leave this blank
    b. Number of bootstrap samples: Set the number of bootstraps too **100**.
    c. Number of threads to use for bootstraping: Enter the number of threads to use for bootstrapping purposes (**default is 5)**
7. Once you filled in with all the details, click "**Launch Analysis**"
8. Once the analysis is completed (approx. 30 min. with the sample data), click on "Analysis," and then click on the analysis "Name" to open the output folder.
9. Output:
   After quantification, you will get a number of files in the output directory.

- `run_info.json` - some high-level information about the run, including the command and versions of kallisto used to generate the output
- `abundance.tsv` - a plain text file with transcript level abundance estimates. This file can be read into R or any other statistical language easily (e.g. `read.table('abundance.tsv')`)
- `abundance.h5` - a HDF5 file containing all of the quantification information including bootstraps and other auxiliary information from the run. This file is read by sleuth

The parameters are pretty minimal. You must supply an index, an output location, and a set of reads. There is also one other important parameter: the number of bootstrap iterations. By default, kallisto runs zero bootstrap iterations. If you do not plan to run sleuth for differential expression analysis, this is okay. But if you plan to run sleuth, you must provide a nonzero number of bootstraps. In general, this number should be at least 30. In your human RNA seq data example we will set it to 100.