DSgen: a tool to generate SOLiD and ILLUMINA synthetic NGS datasets to test mapping tools behavior

Beccuti M[1], Donatelli S[1], Calogero RA[2], Cordero F[1,2]

Motivation

Next Generation Sequencing (NGS) technologies are having a deep impact on modern biology. The recent introduction of instruments capable of producing millions of DNA sequence reads in a single run is rapidly changing the landscape of genetics, providing the ability to answer questions with unimaginable speed. These technologies will provide a genome-wide sequence readout as an endpoint to applications ranging from chromatin immunoprecipitation, mutation/polymorphism to non-coding RNA discovery. Furthermore the NGS technology offers the possibility to move from comparative analysis of gene expression, i.e. microarray based studies, to absolute measurement of RNA species present in a biological sample, i.e RNA-seq studies. However, the sequencing power reachable with NGS technologies open various issues on the experimental design of RNA-seq studies as well as on the definition of application specific analysis pipelines (e.g. non-coding RNA, whole transcriptome analysis, etc). It is mandatory to efficiently evaluate the effect of any step introduced in an analysis pipeline for RNA-seq analysis (e.g. filters, primary mapping tools, etc.). The first step in the pipeline definition is the choice of the best mapping algorithm for the application of interest. In the last two years the number of available mapping has increased steadily, and the choice of the best algorithm may require a deep understanding of the peculiarity of each of them. To investigate their peculiarities it is important to have a common dataset of comparison. We have therefore developed DSgen a tool that allows the GENeration of controlled SOLiD and Illumina Data Set starting from a set of sequences provided by a user forcing mismatches of different sizes.

Methods

When building a dataset for a reference, DSgen takes in input a set of Specific Sequences (SS), a set of Background Sequences (BS) and a maximum number of mismatches (MAX). SS have to be present in the reference while BS should not be part of the reference. DSgen produces totally 24 million of sequences by inserting up to MAX mismatches in a randomly chosen subset of SS and BS. This produces a dataset with a known number of sequences that should not match and a known number that should match up to MAX mismatches. The user is free to define the

[1] Department of Computer Science, University of Torino, Torino, Italy [2] Department of Clinical and Biological Sciences, University of Torino, Torino, Italy

size of the reads (from 35 to 100 nts). The resulting data set encodes in the name of the reads all the information on the mismatches that have been inserted in the read, if any. For example, >SequenceID 2*19*25 means 2 mismatches in positions 19 and 25. The data set is produced in an unique output file. Depending on the request of the user the sequences in the file can be ordered, for example or requiring that all those generated from SS precede those generated by BS, or requiring a random order. The tool generates two types of output: FASTQ file, including quality values, for the Illumina platform, and two files for the SOLiD platform, one for reads and one for quality both in fasta format. DSgen is very efficient in producing such datasets (~ 11' for a 23.5 million tags in color space).

Results

Using this tool we have generated two data sets, base and color space, composed by 24 million of 35 nts reads with up to 4 MMs. We have compared the mapping abilities, over these data sets, of five of the most known NGS aligners PerM, SOCS, SHRiMP, MAQ, BOWTIE. We have selected these mapping algorithms that share a certain similarity in the approach but whose performance differ significantly w.r.t.: main data structure, output sensibility, algorithms to investigate similarity, optimizations to speed-up comparison. These peculiarity can massively impact on computational time and sensibility especially when increasing MAX. In term of the ability to correctly detect the alignment up to three mismatches SHRiMP is the most efficient. PerM is faster but its sensibility decrease beyond an acceptable level for MAX>=3. All tools show an higher mapping capability in base space, with the exclusion of SOCS that is designed only for color space. We have also generated datasets for the detection of mature microRNAs, in which the length of the microRNA sequence is approximately 18-25 nts and therefore reads of 35 are encompassing also part of linker, as background we have used other ncRNAs. Using this dataset we are optimizing a pipe-line for efficient quantification of microRNA in RNA-seq studies. The described tool will be soon implemented as web service.

Contact e-mail

fcordero@di.unito.it