SRmapper – User Guide

Release 0.1.2

INTRODUCTION

SRmapper is a tool for aligning short reads obtained from next-generation sequencing experiments to a reference genome. It is written by Paul G. Gontarz and Jennifer Berger in the laboratory of Chung F. Wong at the University of Missouri-Saint Louis.

This early release demonstrates that a genome-hashing alignment tool can have speed comparable to or faster than alignment tools based on the Burrow-Wheeler Transform such as the BWA package and have similar sensitivity. In addition, SRmapper was designed to have a memory footprint small enough that it can be fully functional on a computer with as little as 4GB of memory for genomes the size of human’s. Before performing alignment to a reference sequence, SRmapper requires a one-time indexing of the reference sequence (buildindex). This index is written to disk and used in the alignment (align) of short reads from a fastq file. SRmapper can align tens of billions of nucleotides per computer day and stores the results in the SAM file format.

buildindex command

```
buildindex { file1.fa file2.fa ... fileN.fa } <index.sqn> [options]

                    -N      Treat nonstandard nucleotides as random nucleotides [off]
```

align command

```
align <index.sqn> { file1.fastq file2.fastq ... fileN.fastq } alignment.sam [options]
```

-a int	Maximum number of equal quality alignments to store per quarter index [5]
-f str	Write unmatched reads to new fastq file [disabled if –f is not included]
-g ull	Manually define genome length.
-I int	Maximum insert size between two mates in a pair end alignment. (Pair end alignment only) [1000]
-m int	Maximum number of mismatches allowed per alignment. If allowing m mismatches results in an alignment score lower than q for a certain read length, the maximum number of mismatches allowed for that length will be such that pHred(m)=q. [off by default]
-p int	Print a maximum of p optimal alignments [1]
-P	Pair end alignment mode. Input format for fastq files is { file1.1.fq file1.2.fq ... fileN.1.fq fileN.2.fq }
-q int	Only search for alignments with a quality of q or higher [3]
-r int	Maximum length for each read in the fastq file [1000]
-s int	Only search each bucket for s keys. Use –s -1 to disable [100]

EXAMPLES:

Ex1: Use the individual chromosomes from a human genome to build the index files for the human genome:

```
seqaln buildindex { chr1.fa chr2.fa chr3.fa chr4.fa chr5.fa
chr6.fa chr7.fa chr8.fa chr9.fa chr10.fa chr11.fa chr12.fa
chr13.fa chr14.fa chr15.fa chr16.fa chr17.fa chr18.fa chr19.fa
chr20.fa chr21.fa chr22.fa chrX.fa chrY.fa } human.sqn
```

Ex2: Align reads from `foo.fq` to the human genome with default settings, and store the alignments in the file `bar.sam` and write unaligned reads to the file `foobar.fq` within the users home directory:

```
seqaln align human.sqn { foo.fq } bar.sam –f /usr/foobar.fq
```

Ex 3: Perform pair end alignment of the pair mate files foo.1.fq and foo.2.fq to the human genome allowing a maximum insert size of 500bp and only considering alignments where the alignment for each pair has a pHred score of at least 5:

```
Seqaln align human.sqn { foo.1.fq foo.2.fq } bar.sam –P –q 5
```

REPORTING BUGS

Report bugs to pmg2m9@mail.umsl.edu