

MapNext: a software tool for spliced and unspliced alignment and SNP detection of short sequence reads

1. Introduction:

MapNext offers a comprehensive and powerful tool for both spliced and unspliced alignments of the short reads and automated SNP detection from population sequences.

MapNext provides four mainly analysis:

- a) Unspliced alignment and clustering reads
- b) Spliced alignment of transcriptomic reads over intron boundaries
- c) SNP detection and estimation of minor allele frequency from population sequences
- d) Storage of result data into database to make it available for more flexible query and further analyses

2. Download and Installation:

The current version is designed for running on Linux/Unix systems.

Compiling the program requires a recent version of g++.

```
g++ -o mapnext mapnext.cpp
```

If you find bugs or have constructive suggestions to the program, please send e-mail to the author.

3. Format of input files:

a) About reference sequences:

Reference file need be a Linux format (\n for line break), not Windows or Dos format(\r\n for line break). You can use linux command: **dos2unix** to change the windows format to Linux format.

If your reference sequences are chromosomes or genomes: ('-g')

1) The reference files can be specified in two ways: as a single chromosome file or as a directory containing multiple files, one for each chromosome. Each chromosome file must only have a single FASTA sequence (i.e. only the first line of the file can start with the '>' character). You can use parameter '-g Chr1.fa' or '-g /home/genome/' for these two ways.

2) **The sequences could not contain line break '\n'** (i.e. only two lines are in the file. The first line is '>chr' and the second line is DNA sequences). You can use perl script **format_genome.pl** to remove the line breaks in sequences.

If your reference sequences are multiple cDNAs or genes: ('-c')

1) one sequence's length could not exceed **100000bp** and you input prepare the single Fasta format file (contains multiple sequences) using parameter '-c ref.fa'. This file doesn't need to remove the line breaks in sequences just like chromosome/genome sequences.

b) About Solexa reads:

MapNext can use two formats (FASTA, FASTQ) for input reads produced from Solexa sequencer. The read length must be between 24 and 50bp.

Reads file need be a Linux format (\n for line break), not Windows or Dos format(\r\n for line break). You can use linux command: **dos2unix** to change the windows format to Linux format.

The following shows examples for these two formats of the input reads files:

FASTA sequence file:

```
>4_1_798_461
TGAAGAGATAGCTTACTTATGTGTCCATGGATATT
>4_1_40_624
GATATTGGGTGAGTCTCGTGGATCGTTGCGGCGTC
```

FASTQ sequence file:

```
@4:1:876:634
GAGAGGTAGGATCAAGAAAGCAAACCTTTGAGTGGA
+4:1:876:634
hhhhhhhhhhhhhhhhhhhh_hhhehbhhhhKhSY\h
@4:1:506:759
GTCTAGCTGAAGCTGTTTCAGTAACTGTCCGAGCAT
+4:1:506:759
hhhhhhhhhhhhhhhhhhhhghhhhh[hgLXP_V
```

About Fastq's quality:

Quality score=(Character's ASCII code - 64) The range of quality score is from -5 to 40.

$Q = -10 * \log(e/(1 - e))$ Q is the quality score, e is the probability of sequencing error

So 'h' stands for quality score:40 and P(error)=0.0001

';' stands for quality score:-5 and P(error)=0.75

4. Usage:

a) Common options:

Usage: mapnext [options]

Options for input:

- a: <string> input solexa sequence file with FASTA format
- s: <string> input solexa sequence file with FASTQ format
- g: <string> input reference genome sequence file or directory(directory name+ '/')
- c: <string> input reference multiple cDNA or gene sequences file with fa format

Options for alignment:

- l: <int> read length (default 35)
- m: <int> max mismatch (default 2)
- k: <string> input splice-spanning sequences file to do spliced alignment
- t: <char> conduct de novo spliced alignment (y or n)
- w: <int> seed_size for de novo spliced alignment(12 or 11)

Options for SNP detection:

- n: <char> do snp finding (y or n)
- f: <float> minimum minor alle frequency (default 0.01)
- e: <int> minimum coverage (default 50)
- q: <int> snp site minimum quality score (default 25)
- p: <int> 4bp-flanking snp minimum quality score (default 20)

Options for output with SQL format:

-d: <char> output sql file (y or n)

b) Command lines:

alignment of reads with fasta format against chromosome sequences:

```
mapnext -s reads.fa -g chr.fa -l 36 -m 2
```

alignment of reads with fastq format against genome sequences:

```
mapnext -a reads.fq -g genome/ -l 36 -m 1
```

alignment of reads with fasta format against multiple gene sequences:

```
mapnext -s reads.fa -c genes.fa -l 42 -m 2
```

unspliced and de novo spliced alignment of transcript reads with fasta format against chromosome sequences:

```
mapnext -s reads.fa -g chr.fa -t y -w 12
```

unspliced and spliced alignment of transcript reads with fasta format against chromosome sequences using additional splice-spanning sequences file:

```
mapnext -a reads.fq -g genome.fa -k spliced.fa
```

SNP detection of population sequences with fastq format:

```
mapnext -a reads.fq -g genome.fa -n y -e 50 -f 0.02 -q 25 -p 20
```

SNP detection of population sequences with fasta format:

```
mapnext -s reads.fa -c genes.fa -n y -e 100 -f 0.05
```

c) Perl scripts:

1) format_reference.pl

Remove the line breaks in sequences.

Command lines:

```
perl format_reference.pl Chr1.fa (format of single file)
```

```
perl format_reference.pl genome/ (format multiple files in directory of genome)
```

2) get_splice_seq.pl

Command lines:

```
perl get_splice_seq.pl NC_003070.gbk 70
```

```
perl get_splice_seq.pl genebank_file length_of_splice_spanning_sequence
```

This script creates the file spliced.fa consisting of splice-spanning sequences.

And the file spliced.fa can be used as reference to find splicing alignment. (-k spliced.fa)

The input file should be genebank format and length should be a even number.

output file example:

spliced.fa:

```
>NC_003070|1|(3881,3913)(3996,4028)
ACATCTGTAGCTACGATCCTTGGAACCTTGCCTTCCAGTCAAAGTA
CAAATCGAGAGATGCTATGT
>NC_003070|2|(4244,4276)(4486,4518)
GAGTTCCACTACGACCTCTTACCAGAACATCAGAGGACATATGTCA
TCTGCAGACTTGAGTACAAG
```

3) get_splice_posi.pl

Command lines:

```
perl get_splice_posi.pl outfile
```

This script reads the file `adition_map` (when using `'-k spliced.fa'`, MapNext will output the file `adition_map`) which contains the position relative to each splice-spanning sequence.

This `get_splice_posi.pl` script change the relative mapping position to absolute mapping position on chromosome. The output file contains the position relative to chromosome.

5. Output files and format

Output files:

genome_map: unspliced alignment of reads on genome/chromosome sequences

genome_cluster: clustering of reads based on alignment

cDNA_map: unspliced alignment of reads on multiple cDNA/gene sequences

cDNA_cluster: clustering of reads based on alignment

additional_map: alignment of reads on additional splice-spanning sequences

spliced_map: de novo spliced alignment of reads on genome/chromosome sequences

candidate_snp: candidate SNPs detected from population sequences based on cDNA mapping

genome_candidate_snp: candidate SNPs detected from population sequences based on genome mapping

One line for One hit. The columns are separated by `'\t'`.

Map file:

1)read name: name of query read

2)reference sequence name: name of reference sequence

3)mapping location: location of query read on the reference sequence, counted from 1

4)+/-: alignment on the direct(+) or reverse(-) chain of the reference

5)mismatch number: number of mismatches

6)unique mapping: unique mapping(1) or multiple mapping(0)

example:

```
@4:1:518:715 NC_003070 76483..76517 + 1 1
```

Cluster file:

cluster NO. (starting position...ending position)

read_name read_sequence reference_name mapping_location strand mismatch

example:

```
cluster1 (3761...3915)
```

read289	TGGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAAC	NC_003070.fa	3761	+	0
read305	GGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAACG	NC_003070.fa	3762	+	0
cluster2 (3994...4276)					
read163	CTTCCAGTCAAAGTACAAATCGAGAGATGCTATGT	NC_003070.fa	3994	+	2
read196	TTCCAGTCAAAGTACAAATCGAGAGATGCTATGTG	NC_003070.fa	3995	+	1

Spliced map file:

- 1)read name: name of query read
- 2)reference sequence name: name of reference sequence
- 3)mapping location: location of query read on the reference sequence(two part)
- 4)+/-: alignment on the direct(+) or reverse(-) chain of the reference

example:

@4:1:234:544 NC_003071 4264...4276 4486...4507 +

SNP file:

- 1)reference sequence name: name of reference sequence
- 2)location: location of SNP on the reference sequence
- 3)major and minor allele: nucleotide of major and minor allele
- 4)minor allele frequency: the frequency of minor allele
- 5)number of A: number of nucleotide A at SNP site
- 6)number of C: number of nucleotide C at SNP site
- 7)number of G: number of nucleotide G at SNP site
- 8)number of T: number of nucleotide T at SNP site

example:

AT1G01070 712 C/A 0.222 131 458 3 0

Author: Hua Bao, Hui Guo and Yuanyan Xiong
 State Key Laboratory of Biocontrol, School of Life Sciences,
 Sun Yat-Sen (Zhongshan) University, Guangzhou, China
 Hua bao: baohua@mail2.sysu.edu.cn

This document was created with Win2PDF available at <http://www.win2pdf.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.