SSAHA2 Manual

February 10, 2010
Version 0.2

Abstract

SSAHA2 maps DNA sequencing reads onto a genomic reference sequence using a combination of word hashing and dynamic programming. Reads from most types of sequencing platforms are supported including paired-end sequencing reads.

The package consists of two separate executables for building the hash index and read mapping. A third executable is included for SNP (Single Nucleotide Polymorphism) calling from Sanger capillary reads. A genotype call of the consensus sequence can be produced using the *ssaha_pileup* package which is distributed separately.

1 Synopsis

ssaha2Build [**-rtype** *platform*] [**-kmer** *wordlen*] [**-skip** *stepsiz*] **-save** *HASH-NAME SUBJECT-FILE*

ssaha2 [**-rtype** *platform*] [**-kmer** *wordlen*] [**-skip** *stepsiz*] [**MAPPING-OPTIONS**] **-save** *HASH-NAME QUERY-FILE* [*MATE-FILE*]
ssaha2 [**-rtype** *platform*] [**-kmer** *wordlen*] [**-skip** *stepsiz*] [**MAPPING-OPTIONS**] *SUBJECT-FILE QUERY-FILE* [*MATE-FILE*]

ssahaSNP [**-rtype** *platform*] [**-kmer** *wordlen*] [**-skip** *stepsiz*] [**MAPPING-OPTIONS**] [**SNP-OPTIONS**] **-save** *HASH-NAME QUERY-FILE*
ssahaSNP [**-rtype** *platform*] [**-kmer** *wordlen*] [**-skip** *stepsiz*] [**MAPPING-OPTIONS**] [**SNP-OPTIONS**] *SUBJECT-FILE QUERY-FILE*

2 Description

ssaha2 uses a hash table of words of fixed length *wordlen* sampled along the genomic reference sequence at equidistant steps of *stepsiz* nucleotide letters. The input file *SUBJECT-FILE* contains the genomic reference sequences in FASTA format. The sequencing reads in the FASTQ file *QUERY-FILE* are

then mapped against the genomic reference sequences one-by-one. First, exactly matching seeds are identified in the reference sequences by looking up the k-mer words of the read in the hash table. Based on the number of k-mer word seeds, segment pairs are selected that potentially match. These candidate pairs are then aligned using a banded Smith-Waterman algorithm.

ssaha2Build builds a hash table for genomic sequences stored in the file *SUBJECT-FILE* in FASTA format and saves the table to the disk. Five files are created all beginning with *HASH-NAME* followed by the extensions *.base, *.body, *.head, *.name and *.size.

ssaha2 maps the reads in the FASTQ file *QUERY-FILE* against the reference sequences. A pre-computed hash table can be read from disk with the **-save** *HASH-NAME* option. Without this option the hash table is created on-the-fly.

ssaha2SNP is a polymorphism detection tool. It detects homozygous SNPs and indels by aligning shotgun reads to the finished genome sequence. From the best alignment, SNP candidates are screened, taking into account the quality value of the bases with variation as well as the quality values in the neighbouring bases, using neighbourhood quality standard (NQS).

3 Options

3.1 GENERAL OPTIONS

-h , -help Print on-line help.

-v , -version Print version information.

-c , -cookbook Print command line examples for common tasks.

-rtype *readtyp* Tune command line parameters for reads from particular sequencing platforms (see SEQUENCING PLATFORMS below). Those settings are overwritten by options explicitly specified on the command line. *readtyp* an be one of 3 strings solexa, 454 and abi.
Example: **-rtype** 454 implies **-skip** 3 which can be overwritten with the options **-rtype** 454 **-skip** 4.

-kmer *wordlen* Sets the length of the hashed words *wordlen* <= 15.

-skip *stepsiz* Sets the number of nucleotide letters between the starting letter of successive words. I.e. With the option **-skip** 1 every word is hashed, with **-skip** 2 every second word, with **-skip** 3 very third etc.

-save *HASH-NAME* Specifies the root name of the hash table files (see FILES below).

3.2 MAPPING OPTIONS

3.2.1 Paired-end reads

-pair *lo,hi* This option indicates that paired-end reads are to be mapped. The interval [*lo*,*hi*] specifies the expected range of the insert sizes. Mate pairs are classified as 'proper' or 'abnormal' depending on whether or not the mates are mapped at distances within the specified interval. I.e. a 'proper' pair would have its mates mapped within *hi* bases from each other but at least *lo* bases apart (*lo*<=*hi*).
This option expects first and second mates of the mate pairs in FASTA or FASTQ format in two separate input files *QUERY-FILE*and *MATE-FILE*.

-mthresh *mapscor* Threshold in the mapping score for mate pairs to be reported if they map outside the distance range (see option **-outfile**). *mapscor* is an integer between 0 and 50 (default: 30). This option requires the **-pair** option to be set.

-multi *randseed* Flag modifying the way paired-end reads with multiple (repetitive) mappings are reported. If *randseed* == 0: skip such pairs entirely. If *randseed* > 0: select one pair at random. The integer *randseed* seeds the random number generator. This option requires the **-pair** option to be set.

3.2.2 Memory

The following option influence the memory footprint.

-array *nint* Memory to be allocated for k-mer word hits as the number of 32-bit integers. The default is *nint* = 4,000,000.

-disk *dswitch* This switch determines how much of the hash table is kept in memory.

- **0:** Map all files into memory including the set of query sequences in file *QUERY-FILE* (and, with the option **-pair**, also *MATE-FILE*).
- **1:** Only the head of the hash table (hash index) is kept in memory. The hash table body, the genomic reference sequences and query sequences are left on disk.
- **2:** Like 1 but subject sequences are kept in memory along with the hash index.

-memory *memsiz* Memory to be allocated for the alignment matrix in Megabytes. Default is *memsiz* = 200.

3.2.3 Segment filtering

Options concerning the selection of segment pairs to be passed on to the dynamic programming step.

-cut *num* Seeds with more than *num* hits in the hash table are ignored.

-depth *nseg* At most the *nseg* segment pairs with the highest number of seeds are passed on to the dynamic programming step.

-seeds *seednum* Segment pairs with a number of seeds below *seednum* are filtered out and not passed on to the dynamic programming step.

-weight *w* w is an integer representing a weighting factor for rare k-mer words. The default is $w = 0$ (no weighting).

3.2.4 Smith-Waterman alignment

The following options influence the dynamic programming step. Segment pairs are seeded before alignment in order to determine the alignment band.

-ckmer *idxlen* Specifies the size of the hash index as the length of the indexed part of the seed.

-cmatch *matchlen* Sets the seed length (*matchlen* >= *idxlen*, see (see option **-ckmer** *idxlen*).

3.2.5 Output filtering

The following options influence how many of the alignments are reported.

-best 0|1 If this flag is set (1) report only the best (by Smith-Waterman score) mapping for each read. If there are multiple best mappings with the same Smith-Waterman score, report them all. This option has no effect if paired-end reads are mapped (with the option **-pair** *a,b*).

-diff *swscor* Only report mappings with Smith-Waterman scores within *swscor* of the best mapping.

-udiff *uswscor* Do not report the best (by Smith-Waterman score) alignment if the second best alignment is within *uswscor* of the best.

-identity *seqid* Report only mappings with a pairwise sequence identity of at least *seqid*. Other mappings are skipped. *seqid* specified the pairwise sequence identity as a percentage of the alignment length.

-score *swatscor* Report only mappings with a Smith-Waterman score of at least *swatscor*. Mappings with lower scores are not reported.

3.2.6 Output formats

-output *format* Sets the format in which the mappings are reported. *ssaha2* writes the mappings to standard output except when the **-outfile** *filename* option is specified. Supported output formats are all ASCII text based. For a detailed description of output formats see `http://www.sanger.ac.uk/Software/analysis/SSAHA2/formats.shtml`.

format can be one of the following strings:

aln Tony Cox' alignment output format.

cigar Compact Idiosyncratic Gapped Alignment Report.

gff `http://www.sanger.ac.uk/Software/formats/GFF/`

psl Tab delimited format similar to BLT `http://genome.ucsc.edu/goldenPath/help/customTrack.html`

pslx psl format complemented with sequence output.

sam SAM format `http://samtools.sourceforge.net`. With the **-outfile** *filename* option SAM output is written to the file *filename*. Only the aligned segment of the reads sequence is reported (hard clipping).

sam_soft SAM format using soft clipping. The sequence of the entire read is output.

ssaha2 *ssaha2* native output format (default).

sugar Simple UnGapped Alignment Report.

vulgar Verbose Useful Labelled Gapped Alignment Report

-output ssaha2 cigar Alternates between ssaha2 and cigar format lines.

-outfile *filename* This option is used in connection with the SAM output format (options **-output** sam and **-output** sam_soft) or when paired-end reads are mapped (option **-pair** *lo,hi*).

For all output formats except the SAM format mate-pairs mapped at distances in the interval [*lo*,*hi*] (i.e. 'proper' pairs) are reported to standard output. Mate pairs mapped outside the interval are written to the file *filename* or, if the option **-outfile** *filename* is not given, are suppressed.

For SAM format all mappings are either written to the file *filename* or to standard output (if the option **-outfile** *filename* is not specified).

-tags 0|1 If this flag is set (1) output lines reporting a mapping are preceded by a keyword or tag to aid parsing. The tag depends on the output format used. (Default is 1).

-align 0|1 If this flag is set (1), a graphic representation of the sequence alignment is output. (default is 0).

-name 0|1 Flag that modifies option '-output cigar' such that read name and length are also reported when there was no hit found. Default is 0.

3.3 SNP OPTIONS

The following options are specific to *ssahaSNP*

-fix 0|1 If this flag is set (1) fix -edge, -seeds, -score so that they are not updated according to read length in *ssahaSNP*. The default is 0.

-NQS 0|1 Flag indicating the use Neighborhood Quality Standard (NQS) to filter SNPs (1), otherwise output all candidates (0). The default is 1.

-quality *qval* Sets the quality value to use for the variation base when NQS is used. The default is *qval* = 23.

4 Sequencing Platforms

The following options tune parameters for particular sequencing platforms.

-rtype *platform* *platform* can be one of the following strings:

solexa Tunes for Illumina-Solexa reads. This implies the following options **-kmer** 13 **-skip** 2 **-seeds** 2 **-score** 12 **-cmatch** 9 **-ckmer** 6. If any of these implied options occur together with the option **-rtype** solexa on the ssaha2 command line they overwrite the implied value. I.e. **-rtype** solexa **-kmer 11** would use 11-mer words in the hash.

454 Tunes for Roche 454 FLX reads. This implies the following options **-kmer** 13 **-skip** 3 **-seeds** 2 **-score** 30 **-cmatch** 10 **-ckmer** 6.

abi Tunes for Sanger capillary reads. This is the default.

-solexa tunes for Illumina-Solexa reads like **-rtype** solexa.

-454 tunes for Roche 454 FLX like **-rtype** 454.

5 Files

`HASH-NAME.base` Compressed set of reference sequences for which the hash table of k-mer words was generated.

`HASH-NAME.body` Body of the hash table with k-mer word positions.

`HASH-NAME.head` Head of the hash table (hash index).

`HASH-NAME.name` Names of the reference sequences.

`HASH-NAME.size` Lengths of the reference sequences.

6 Memory Requirements

How much memory *SSAHA2* occupies is determined largely by the size of the **head** (the actual index) and the **body**(the word locations) of the hash table.

The size of the **head** depends on the length k of the hashed k-mer words and is approx. $4^{(k+1)}$ bytes. The word length is specified with the **-kmer** *k* option.

The size of the **body** depends on the total length N of the hashed reference sequences and the skip step s specified with the option **-skip** *s*. The body size is approx. 4*N/s bytes.

For example, a hash table built for the human genome (N = $3x10^9$ bases) with the options **-kmer** 12 **-skip** 2 would have a header of 67 MB and a body of 6 GB. With parameters **-kmer** 13 **-skip** 3 the header would occupy 268 MB, the body 4 GB.

6.1 Reducing the memory requirements

The option **-disk** 1 leaves the body of the table on disk, but the program will then be slower because it has to access the disk frequently.

The header is always kept in memory.

7 Examples

7.1 454 reads

Map Roche-454 reads against the human genome.
Assume you have a FASTA file `NCBI36.fa` with the chromosome sequences. Generate the hash index and store it on the disk:

```
ssaha2Build -454 -save htab NCBI36.fa
```

This comman will requires 7.6 GB to run. It generates the hash table files `hs36htab.base`, `hs36htab.body`, `hs36htab.head`, `hs36htab.name`, `hs36htab.size`.

7.1.1 SAM output

single reads Assuming the reads are in the FASTQ file `hs454.fq` run the mapper with

```
ssaha2 -454 -output sam -outfile mapped.sam
       -save htab hs454.fq
```

This will write mappings to the files `mapped.sam` in SAM format.

paired-end reads Assuming the mates of the read pairs are in the FASTQ files `hs454_1.fq` and `hs454_2.fq`. Run the mapper with the options

```
ssaha2 -454 -pair 200,3000 -output sam
       -outfile mapped.sam
       -save htab hs454_1.fq hs454_2.fq
```

This will write mapped and unmapped reads to the files `mapped.sam` in SAM format.

7.1.2 CIGAR output

For output formats other than SAM, the behaviour for the **-outfile** *filnam* flag differs.

single reads The command

```
ssaha2 -454 -output cigar -save htab hs454.fq
```

writes mappings to standard output in CIGAR format. CIGAR lines are preceded by the *cigar* keyword. If the **-outfile** option is present it is ignored.

paired-end reads The command

```
ssaha2 -454 -pair 200,3000 -output cigar
       -save htab hs454_1.fq hs454_2.fq
```

writes mappings for proper pairs to standard output in CIGAR format, i.e. pairs that are mapped at most 3000 but no fewer than 200 bases apart. 'Abnormal' mappings outside that interval can be retrieved in a separate output file `abnorm.cig` with the command line:

```
ssaha2 -454 -pair 200,3000 -outfile abnorm.cig
       -output cigar
       -save htab hs454_1.fq hs454_2.fq
```

7.1.3 Reduced memory footprint

Running *ssaha2* with the above commands will require 5-6 GB. The memory footprint will also depend on the size of the query read file(s). To run the mapper on machines with less memory one can use the **-disk** 1 flag:

```
ssaha2 -454 -pair 200,3000 -output sam
       -outfile mapped.sam -disk 1
       -save htab hs454_1.fq hs454_2.fq
```

This will require less than 1 GB regardless of the size of the query read files `hs454_1.fq`,`hs454_2.fq`. However, the run time will be considerably slower.

7.2 Illumina-Solexa reads

7.2.1 Paired-end reads against human genome

Map paired-end Illumina-Solexa reads against the human genome (in FASTA file `NCBI36.fa`) producing SAM output.

Short reads For example, for read pairs of 2x36 bp length and mean insert size 200bp in FASTQ files `hs36l200i_1.fq`,`hs36l200i_2.fq` you might use:

```
ssaha2Build -solexa -save hs36k13s2 NCBI36.fa
ssaha2 -solexa -pair 20,400 -outfile mapped.sam -output sam
       -save hsk13s2 hs36l_1.fq hs36l_2.fq
```

This will write all SAM output to the file `mapped.sam`.

Longer reads For read pairs of, say, 2x75 bp length one might be able (depending on the sensitivity required) to increase the step size and therefore speed with reduced memory demands:

```
ssaha2Build -solexa -skip 6 -save hs36k13s6 NCBI36.fa
ssaha2 -solexa -skip 6 -pair 20,400 -outfile mapped.sam
       -output sam
       -save hs36k13s6 hs75l_1.fq hs75l_2.fq
```

7.2.2 Paired-end reads against contigs

Map paired-end reads against contigs of a partial assembly in file `contigs.fa` with the hash index calculated on-the-fly:

```
ssaha2 -solexa -pair 20,400 contigs.fa reads_1.fq reads_2.fq
```

This will output mappings in the default output format, each line preceded by the keyword *ALIGNMENT*.

7.3 Sanger capillary reads

Map single capillary reads against contigs of a partial assembly in file `contigs.fa` with the hash index calculated on-the-fly and PSL output format :

```
ssaha2 -output psl contigs.fa reads.fq
```

8 Version

Version: 0.2 of February 10, 2010.

9 License and Copyright

License Binaries compiled for a range of platforms are available free of charge from `http://www.sanger.ac.uk/Software/analysis/SSAHA2/`.
If you are working in academia and require the SSAHA2 source code you first must obtain a license for *phrap/cross_match* from Phil Green at the University of Washington `http://www.phrap.org`.

10 Authors

SSAHA2 was written at the Wellcome Trust Sanger Institute, Cambridge, UK, originally by Adam Spargo and Zemin Ning with contributions by James Mullikin, Tony Cox and Nikolai Ivanov. It is currently being developed by Hannes Ponstingl [`hp3@sanger.ac.uk`].