Jan 09, 2012

ASAP - Allele-specific alignment pipeline

(1) ASAP - Quick Reference

ASAP needs a working version of Perl and is run from the command line. Furthermore, Bowtie needs to be installed on your computer (http://bowtie-bio.sourceforge.net/index.shtml).

First you need to obtain the two reference genomes you want to use for alignments and place them in separate genome folders. Genomes can be obtained from e.g. the Ensembl website (http://www.ensembl.org/info/data/ftp/index.html/) or NCBI website (ftp://ftp.ncbi.nih.gov/genomes/). ASAP supports reference genome sequence files in FastA format, allowed file extensions are either either .fa or .fasta. Both single-entry or multiple-entry FastA files are supported.

If you want to align your NGS sequence file to two essentially identical genomes which differ only by a number of SNPs it is assumed that you have got a version of the SNP containing genome available in FastA format, too. ASAP does currently not provide a guide for the *in silico* generation of new genomes using SNP information (e.g. from dbSNP or a resequencing project), but if there is demand we could work on a solution for this in the future.

The following examples will use the file `'ASAP_test_data.fastq'` which is available for download from the ASAP homepage (it contains 10,000 reads in FastQ format, Phred33 qualities, 27 bp long reads, from a mouse hybrid ES cell line (Black6/129_SvImJ)).

(I) Running the Bowtie indexer

Before ASAP can be run both genomes need to be indexed using the Bowtie indexer (`bowtie-build`). Depending on the genome size and hardware this process can take up to several hours.

USAGE: `bowtie-build [options]* <reference_in> <ebwt_base>`

For further information please consult the Bowtie manual (http://bowtie-bio.sourceforge.net/manual.shtml#the-bowtie-build-indexer).

A typical sequential genome indexing process could look like this:

```
cd /path/to/Genomes/Mouse/NCBIM37/Black6/
bowtie-build
Mus_musculus.NCBIM37.52.dna.chromosome.10.fa,Mus_musculus.NCBIM37.52.dna.chromosome.11.fa,Mus_
musculus.NCBIM37.52.dna.chromosome.12.fa,Mus_musculus.NCBIM37.52.dna.chromosome.13.fa,Mus_musc
ulus.NCBIM37.52.dna.chromosome.14.fa,Mus_musculus.NCBIM37.52.dna.chromosome.15.fa,Mus_musculus
.NCBIM37.52.dna.chromosome.16.fa,Mus_musculus.NCBIM37.52.dna.chromosome.17.fa,Mus_musculus.NCB
```

```
IM37.52.dna.chromosome.18.fa,Mus_musculus.NCBIM37.52.dna.chromosome.19.fa,Mus_musculus.NCBIM37
.52.dna.chromosome.1.fa,Mus_musculus.NCBIM37.52.dna.chromosome.2.fa,Mus_musculus.NCBIM37.52.dn
a.chromosome.3.fa,Mus_musculus.NCBIM37.52.dna.chromosome.4.fa,Mus_musculus.NCBIM37.52.dna.chro
mosome.5.fa,Mus_musculus.NCBIM37.52.dna.chromosome.6.fa,Mus_musculus.NCBIM37.52.dna.chromosome
.7.fa,Mus_musculus.NCBIM37.52.dna.chromosome.8.fa,Mus_musculus.NCBIM37.52.dna.chromosome.9.fa,
Mus_musculus.NCBIM37.52.dna.chromosome.MT.fa,Mus_musculus.NCBIM37.52.dna.chromosome.X.fa,Mus_m
usculus.NCBIM37.52.dna.chromosome.Y.fa Black6

cd /path/to/Genomes/Mouse/NCBIM37/129_SvImJ/
bowtie-build
chr10.fa,chr11.fa,chr12.fa,chr13.fa,chr14.fa,chr15.fa,chr16.fa,chr17.fa,chr18.fa,chr19.fa,chr1
.fa,chr2.fa,chr3.fa,chr4.fa,chr5.fa,chr6.fa,chr7.fa,chr8.fa,chr9.fa,chrX.fa,chrY.fa 129_SvImJ
```

(II) Running ASAP

USAGE: `./ASAP [options] --genome_1 <genome_folder> --index_1 <basename> --genome_2 <genome_folder> --index_2 <basename> {-1 <mates1> -2 <mates2> | <singles>}`

A typical single-end analysis of a 40 bp single-end sequencing run could look like this:

```
./ASAP -n 2 -l 40 --chunkmbs 512 --genome_1
/data/Genomes/Mouse/NCBIM37/Black6/ --index_1
/data/Genomes/Mouse/NCBIM37/Black6/Black6 --genome_2
/data/Genomes/Mouse/NCBIM37/129_SvImJ/ --index_2
/data/Genomes/Mouse/NCBIM37/129_SvImJ/129_SvImJ  ASAP_test_data.fastq
```

This will produce four output files:

(1) `ASAP_test_data.fastq_g1_specific_ASAP.txt` (contains all alignments which are specific for genome 1)

(2) `ASAP_test_data.fastq_g2_specific_ASAP.txt` (contains all alignments which are specific for genome 2)

(3) `ASAP_test_data.fastq_common_alignments_ASAP.txt` (contains all alignments that align equally well to both genomes)

(4) `test_dataset.fastq_report_ASAP.txt` (contains alignment summary)

(2) ASAP - General Information

What is ASAP?

ASAP is an alignment tool to perform alignments against two reference genomes at the same time in order to determine whether a given sequence has a best alignment to one of the two references. Ungapped read alignments are carried out using the short read aligner Bowtie (Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25) and therefore it is a requirement that Bowtie is installed on your machine (see Dependencies). ASAP is written in Perl and is run from the command line.

All files associated with ASAP as well as a test data set can be downloaded from:
http://www.bioinformatics.bbsrc.ac.uk/projects/

We would like to hear your comments/suggestions about ASAP! Please email them to: felix.krueger@babraham.ac.uk

Installation notes

ASAP is written in Perl and is executed from the command line. To install ASAP simply copy the ASAP_v0.X.Y.tar.gz file into an ASAP installation folder and extract all files by typing:

```
tar xzf ASAP_v0.X.Y.tar.gz
```

Dependencies

ASAP requires a working version of Perl and Bowtie to be installed on your machine (http://bowtie-bio.sourceforge.net/index.shtml). ASAP will assume that the Bowtie executable is in your path unless the path to Bowtie is specified manually with:

```
--path_to_bowtie </../../bowtie>.
```

In order to work properly the current working directory must contain the sequence files to be analysed.

Hardware requirements

ASAP holds both reference genomes in memory and in addition to that runs two parallel instances of Bowtie. The memory usage is dependent on the size of the reference genome. For a large eukaryotic

genome (human or mouse) we experienced a typical memory usage of around 6-8 GB. We thus recommend running ASAP on a machine with 4 CPU cores and 12 GB of RAM.

Alignment speed depends largely on the read length and Bowtie parameters used. Allowing many mismatches and using a short seed length (which is the Bowtie default option, see below) tends to be fairly slow, whereas looking for near perfect matches can align up to 25 million sequences per hour.

ASAP test data set

A test data set is available for download from the ASAP homepage. It contains 10,000 single-end reads from murine hybrid ES cells (Black6/129_SvImJ) in FastQ format (ESHyb_H3K9me3, Mikkelsen et al., GEO accession GSE12241; sequence length: 27 bp; base call qualities are Sanger encoded Phred values (Phred33)).

Which kind of sequence files and/or experiments are supported?

ASAP supports the alignment of reads for the following conditions:

- sequence format either FastQ or FastA
- single-end or paired-end reads
- variable read length support
- supports similar genomes (identically sized genomes but harbouring SNPs) or dissimilar genomes (e.g. genomes of different strains or species, individual chromosomes or loci, etc.)

ASAP retains much of the flexibility of Bowtie (adjustable seed length, number of mismatches, insert size ...). For a full list of options please run `./ASAP --help` or see the Appendix at the end of this User Guide.

It should be mentioned that ASAP supports only reads in base-space, such as from the Illumina platform. There are currently no plans to extend its functionality to colour-space reads.

How does ASAP work?

Input sequences are aligned to two genomes in parallel, and the best two alignments against each of the genomes are considered. ASAP first determines whether the sequence can be mapped uniquely to each of the genomes, which is based on the number of mismatches in the read for single-end files, or the sum of mismatches from both reads for paired-end files.

Only if a sequence can be mapped unambiguously to both genomes, or to one genome but not to the other, ASAP proceeds to determine if the sequence maps better to one of the genomes or if the alignment is in common between both genomes. Sequences producing multiple "best" alignments to

either or both genomes are discarded. Un-mappable sequences can be printed out to a file using the option `--unmapped <filename>`.

(3) Running ASAP

Before running ASAP we recommend spending some time on quality control of the raw sequence files using FastQC (www.bioinformatics.bbsrc.ac.uk/projects/). FastQC might be able to spot irregularities associated with your sequencing file, such as high base call error rates or contaminating sequences such as PCR primers or adapters. Many sources of error can cause the alignment efficiencies to drop or produce false alignments, so adaptive quality and/or adapter trimming might prove useful prior to running ASAP.

If no additional options are specified ASAP will use a set of default values, some of which are:

- if no specific path to Bowtie is specified it is assumed that the `bowtie` executable is in the path.
- Bowtie is run `--best` mode (it is not recommended to turn this off)
- Bowtie is run assuming a Phred33 scale for FastQ files (Sanger encoding).
- standard alignments allow up to 2 mismatches in the seed region (which is defined as the first 28 bp by default). These parameters can be modified using the options `-n` and `-l`, respectively.

Even though the user is not required to specify additional alignment options it is often advisable to do so. To see a full list of options please type `./ASAP --help` on the command line or see the Appendix at the end of this User Guide.

USAGE: `./ASAP [options] --genome_1 <genome_folder> --index_1 <basename> --genome_2 <genome_folder> --index_2 <basename> {-1 <mates1> -2 <mates2> | <singles>}`

A typical single-end analysis of a 40 bp sequencing run could look like this:

```
./ASAP -n 2 -l 40 --chunkmbs 512 --genome_1 /Genomes/Mouse/NCBIM37/Black6/
--index_1 /Genomes/Mouse/NCBIM37/Black6/Black6  --genome_2
/Genomes/Mouse/NCBIM37/129_SvImJ/ --index_2
/Genomes/Mouse/NCBIM37/129_SvImJ/129_SvImJ  ASAP_test_data.fastq
```

This will produce four output files:

(1) `ASAP_test_data.fastq_g1_specific_ASAP.txt` (contains all alignments which are specific for genome 1)

(2) `ASAP_test_data.fastq_g2_specific_ASAP.txt` (contains all alignments which are specific for genome 2)

(3) `ASAP_test_data.fastq_common_alignments_ASAP.txt` (contains all alignments that align equally well to both genomes. Note that this file will be left empty if --dissimilar has been specified, see below)

(4) `test_dataset.fastq_report_ASAP.txt` (contains alignment summary)

ASAP alignment report

Upon completion, ASAP produces a run report which contains information about the following:

- Summary of alignment parameters used

- Number of sequences analysed
- Number of sequences which could be mapped uniquely (mapping efficiency)
- Number of sequences specific for either of the two analysed genomes
- Number of sequences in common to both genomes

This alignment summary is also printed into a file called _ASAP_report.txt for your information and record keeping.

The ASAP output

For equivalent genomes (default mode) ASAP produces three comprehensive alignment output files for each input file or set of paired-end input files. The sequence basecall qualities of the input FastQ files are written out into the ASAP output file as well to allow filtering on quality thresholds. Please note that the quality values are encoded in Sanger format (Phred 33 scale), even if the input was in Phred64 or the old Solexa format.

The single-end output contains the following information (1 line per sequence, tab separated):

```
 (1) seq-ID
 (2) read sequence
 (3) specific for genome (1/2/N)
 (4) read alignment strand
 (5) alignment chromosome
 (6) alignment start position
 (7) alignment end position
 (8) genome 1 sequence
 (9) genome 1 mismatch information (blank if perfect match)
(10) genome 2 sequence
(11) genome 2 mismatch information (blank if perfect match)
(12) read quality score (Phred33 scale)
```

Single-end alignment example (a genome 2-specific alignment):

```
 (1) HWUSI-EAS611_0001:2:1:1078:13104#0/1
 (2) AGCAAGCTCTGGGGTCAGATGGGTCAGAT**G**GGTAGATCAG
 (3) 2
 (4) +
 (5) 9
 (6) 21059119
 (7) 21059158
 (8) AGCAAGCTCTGGGGTCAGATGGGTCAGAT**A**GGTAGATCAG
 (9) 29:A>G
(10) AGCAAGCTCTGGGGTCAGATGGGTCAGAT**G**GGTAGATCAG
(11)
(12) BCCCCCCCCCCCCCCCCCCCCCCB8-=?5>>>=8AA>8AAAAAA
```

The paired-end output looks like this (1 line per sequence pair, tab separated):

```
 (1) seq-ID
 (2) read 1 sequence
 (3) read 2 sequence
 (4) specific for genome (1/2/N)
 (5) read 1 alignment strand
 (6) alignment chromosome
 (7) alignment start position
 (8) alignment end position
 (9) genome 1 sequence 1
(10) genome 1 read 1 mismatch information
(11) genome 1 sequence 2
(12) genome 1 read 2 mismatch information
(13) genome 2 sequence 1
(14) genome 2 read 1 mismatch information
(15) genome 2 sequence 2
(16) genome 2 read 2 mismatch information
(17) read 1 quality score (Phred33 scale)
(18) read 2 quality score (Phred33 scale)
```

Paired-end alignment example (a perfect alignment in common to both genomes):

```
 (1) HWUSI-EAS611:5:1:22:830#0
 (2) TACGGCCCCACACCTCCTCCTTACCTTCCTCCACGCAC
 (3) AGATCTGGTGCGTGGAGGAAGGTAAGGAGGAGGTGTG
 (4) N
 (5) -
 (6) 2
 (7) 156190737
 (8) 156190781
 (9) TACGGCCCCACACCTCCTCCTTACCTTCCTCCACGCAC
(10)
(11) AGATCTGGTGCGTGGAGGAAGGTAAGGAGGAGGTGTG
```

```
(12)
(13) TACGGCCCCACACCTCCTCCTTACCTTCCTCCACGCAC
(14)
(15) AGATCTGGTGCGTGGAGGAAGGTAAGGAGGAGGTGTG
(16)
(17) 9@>>=A>@6A@?9A13?>72@@969@//9;69?/>8@
(18) BBAB@BBBBBBB>BBBBBBBBBBB@BAA=ABBBBB@BB
```

If you get stuck at any point or have any questions/comments please contact me via email: felix.krueger@babraham.ac.uk

(4) APPENDIX - Full list of options

A full list of options can also be viewed by typing: `./ASAP --help`

USAGE: `./ASAP [options] --genome_1 <genome_folder> --index_1 <basename> --genome_2 <genome_folder> --index_2 <basename> {-1 <mates1> -2 <mates2> | <singles>}`

ARGUMENTS (required)

--genome_1 <>	The full path to the folder containing reference genome 1. ASAP expects one or more FastA files in this folder (file extension: .fa or .fasta).
--genome_2 <>	The full path to the folder containing reference genome 2. ASAP expects one or more FastA files in this folder (file extension: .fa or .fasta).
--index_1 <>	The full path to the bowtie index basename of genome 1 (e.g. /data/genomes/mouse/mus_musculus/C57BL6). The basename is the name of any of the index files up to but not including the final .1.ebwt / .rev.1.ebwt / etc.
--index_2 <>	The full path to the bowtie index basename of genome 2 (e.g. /data/genomes/mouse/mus_musculus/castaneus). The basename is the name of any of the index files up to but not including the final .1.ebwt / .rev.1.ebwt / etc.
-1 <mates1>	Comma-separated list of files containing the #1 mates (filename usually includes "_1"), e.g. flyA_1.fq,flyB_1.fq). Sequences specified with this option must correspond file-for-file and read-for-read with those specified in <mates2>. Reads may be a mix of different lengths. ASAP will produce three mapping result and one report file per paired-end input file pair.
-2 <mates2>	Comma-separated list of files containing the #2 mates (filename usually includes "_2"), e.g. flyA_1.fq,flyB_1.fq). Sequences specified with this option must correspond file-for-file and read-for-read with those specified in <mates1>. Reads may be a mix of different lengths.
<singles>	A comma-separated list of files containing the reads to be aligned (e.g. lane1.fq, lane2.fastq,lane3.txt). Reads may be a mix of different lengths. ASAP will produce three mapping result and one report file per input file.

OPTIONS:

Input:

-q/--fastq	The query input files (specified as <mate1>,<mate2> or <singles> are FASTQ files (usually having extension .fg or .fastq). This is the default. See also --solexa-quals.
-f/--fasta	The query input files (specified as <mate1>,<mate2> or <singles> are FASTA files (usually havin extension .fa, .mfa, .fna or similar). All quality values are assumed to be 40 on the Phred scale.
-s/--skip <int>	Skip (i.e. do not align) the first <int> reads or read pairs from the input.
-u/--qupto <int>	Only aligns the first <int> reads or read pairs from the input. Default: no limit.
--phred33-quals	FASTQ qualities are ASCII chars equal to the Phred quality plus 33. Default: on.
--phred64-quals	FASTQ qualities are ASCII chars equal to the Phred quality plus 64. Default: off.
--solexa-quals	Convert FASTQ qualities from solexa-scaled (which can be negative) to phred-scaled (which can't). The formula for conversion is: phred-qual = 10 * log(1 + 10 ** (solexa-qual/10.0)) / log(10). Used with -q. This is usually the right option for use with (unconverted) reads emitted by the GA Pipeline versions prior to 1.3. Default: off.
--solexa1.3-quals	Same as --phred64-quals. This is usually the right option for use with (unconverted) reads emitted by GA Pipeline version 1.3 or later. Default: off.
--path_to_bowtie	The full path </../../> to the Bowtie installation on your system. If not specified it will be assumed that Bowtie is in the path.
--dissimilar	Specifying this option will inform ASAP that the two genomes are not essentially the same except for SNPs (which is the default), but that they are dissimilar (e.g. genome 1 could be the Black6 mouse genome, and genome 2 could be just one chromosome from a different mouse strain which can potentially include SNPs and/or chromosomal rearrangements). In such a case, ASAP will not attempt to extract the genomic sequence at the corresponding position in the second genome, but will write out the first best alignment to the second genome instead (if appplicable; if there was no best alignment genome 2 fields will be left blank). This option will not write

any sequences to an "alignments in common" output file as the concept of homologous sequences does not apply to this scenario.

Alignment:

-n/--seedmms <int> The maximum number of mismatches permitted in the "seed", which is the first 20 base pairs of the read by default (see -l/--seedlen). This may be 0, 1, 2 or 3.

-l/--seedlen The "seed length"; i.e., the number of bases of the high quality end of the read to which the -n ceiling applies. The default is 28.

-e/--maqerr <int> Maximum permitted total of quality values at all mismatched read positions throughout the entire alignment, not just in the "seed". The default is 70. Like Maq, bowtie rounds quality values to the nearest 10 and saturates at 30.

--chunkmbs <int> The number of megabytes of memory a given thread is given to store path descriptors in --best mode. Best-first search must keep track of many paths at once to ensure it is always extending the path with the lowest cumulative cost. Bowtie tries to minimize the memory impact of the descriptors, but they can still grow very large in some cases. If you receive an error message saying that chunk memory has been exhausted in --best mode, try adjusting this parameter up to dedicate more memory to the descriptors. Default: 512.

-I/--minins <int> The minimum insert size for valid paired-end alignments. E.g. if -I 60 is specified and a paired-end alignment consists of two 20-bp alignments in the appropriate orientation with a 20-bp gap between them, that alignment is considered valid (as long as -X is also satisfied). A 19-bp gap would not be valid in that case. Default: 0.

-X/--maxins <int> The maximum insert size for valid paired-end alignments. E.g. if -X 100 is specified and a paired-end alignment consists of two 20-bp alignments in the proper orientation with a 60-bp gap between them, that alignment is considered valid (as long as -I is also satisfied). A 61-bp gap would not be valid in that case. Default: 250.

Reporting:

-k <2> Due to the way ASAP works Bowtie will report up to 2 valid alignments. This option will be used by default and cannot be changed.

--best Make Bowtie guarantee that reported singleton alignments are "best" in terms of stratum (i.e. number of mismatches, or mismatches in the seed in the case if -n mode) and in terms of the quality; e.g. a 1-mismatch alignment where the mismatch position has Phred quality 40 is preferred over a 2-mismatch alignment where the mismatched positions both have Phred quality 10. When --best is not specified, Bowtie may report alignments that are sub-optimal in terms of stratum and/or quality (though an effort is made to report the best alignment). --best mode also removes all strand bias. Note that --best does not affect which alignments are considered "valid" by Bowtie, only which valid alignments are reported by Bowtie. Bowtie is about 1-2.5 times slower when --best is specified. Default: on.

--no_best Disables the --best option which is on by default. This can speed up the alignment process, e.g. for testing purposes, but for credible results it is not recommended to disable --best.

--quiet Print nothing besides alignments.

--unmapped <filename> Instructs ASAP to write out all sequences which did not yield a unique alignment (either not mappable or ambiguously mapping sequences) to <filename> in the same format as the inputfile. For paired-end alignments, two files (_1 and_2) will be generated.

Other:

-h/--help Displays this help file.

-v/--version Displays version information.