SHOREmap Manual

Version 1.0

Stephan Ossowski
Korbinian Schneeberger

@ Max Planck Institute for Developmental Biology
Tübingen, Weigelworld, 26.03.2009

Contents

SHOREmap is an analysis pipeline for mapping and mutant identification in one step. See "Next-generation genetics: mapping and mutant identification in one step by deep sequencing", Schneeberger, Ossowski et al., for further information about the application of SHOREmap.

In order to run SHOREmap you need to provide the output of a Next Generation Sequencing read analysis in the right format. SHORE, a short read analysis tool, provides correctly formatted files. However, any alignment tool can be used, if its output is formatted in the right way for SHOREmap.

Find a detailed description of how to use SHORE in the SHORE-README file on http://1001genomes.org/downloads.

Korbinian Schneeberger and Stephan Ossowski, Tübingen, March 2009

1 Prerequisites

1.1 Download and Installation

Download SHORE and SHOREmap from http://1001genomes.org/downloads/shore. After storing the SHOREmap_release_1.0.tar.gz somewhere on your harddrive, use the terminal application and change the working directory to the directory where you stored SHOREmap. Typing

```
tar zxf SHOREmap_release_1.0.tar.gz
```

will unpack SHOREmap. There is no further need for installing.

1.2 Installation of R

SHORE and SHOREmap make use of R for data visualization. Where R is not necessary for SHORE, it becomes necessary for SHOREmap. R has to be installed and the installation path has to be added to the $PATH environmental variable.

1.3 Installation of PERL

SHOREmap is written in PERL. To run SHOREmap you need to have PERL installed on your computer.

2 SHOREmap Guide

2.1 SHOREmap_interval.pl

INTERVAL generates a visual output allowing the user to define a mapping interval. By default INTERVAL prints 10 different plots of all chromosomes, each with a different sliding window size. After inspecting these plots, the user selects one of them and uses the text file associated with that plot as input for ANNOTATE. It is not critical to select the smallest possible interval, the only requirement is that the highest peak must be included within the selected region. This peak is used by ANNOTATE to prioritize the mutations within the interval. If the first set of plots is not satisfying, e.g. too flattened or too jagged, we recommend re-running INTERVAL with a new set of sliding window sizes.

Example:

```
> SHOREmap_interval.pl --consensus=consensus_summary.txt
--marker=marker.txt --chrsizes=chrsizes.txt --referrors=ref_error.txt
```

-consensus describes the base count per position. -marker lists the marker positions. -chrsizes indicates the length of each reference sequence, which is necessary for a uniform plotting of the chromosomes. -referrors describes all the positions that will be disgarded by INTERVAL.

Not used in this example: -windowstep describes the number of bp between the reported data points. Usually you do not have to change this. -windowsize might be adjusted once the first round of interval finding has been analyzed.

2.2 SHOREmap_annotate.pl

ANNOTATE outputs a list of mutations prioiritized by their distance to the highest peak in the user-defined interval (either generated by INTERVAL or by DENOVO). If additional information about gene annotations is provided (in GFF format) the effect of the mutations will be added to the output.

Example:

```
> SHOREmap_annotate.pl --snp=homozygous_snps.txt
--dist=SHOREmap.output.txt --chrom=4 --start=16000000
--end=17000000 --genome=TAIR8.v1.fa --gff=TAIR8.gff
--referr=ref_error.txt
```

-snp describes all mutations called, in SHORE format. -dist is the output file of either INTERVAL or DENOVO. -chrom, -start and -end indicate the interval location. -genome is the reference sequence in fasta format. -gff is a GFF formatted description of the reference annotation. -referr describes all the positions, which will be disgarded in the final output.

Not used in this example: -del and -ins describe small indels found in the read mapping which could of course also be causal for the selected phenotype. Files are in SHORE format.

2.3 SHOREmap_denovo.pl

Like INTERVAL, DENOVO helps defining the mapping interval, but does not rely on knowledge of marker postions. Instead markers are defined on the fly as heterozygous positions in the short read data. DENOVO detects the heterozygous positions based on the read alignments and calculates their frequencies and the distance of each postion to the next marker. This is based on the assumption that the homozygous character of the genomic region harboring the causal mutation reduces the density of markers and thus increases the average distance to the nearest marker.

Like with INTERVAL, multiple plots are generated and the user selects a plot and an interval based on peak appearance. The text file associated with the selected plot is then used as input for ANNOTATE.

Example:

```
> SHOREmap_denovo.pl --snp=minor_allele_frequency.txt
--refseq=reference.txt --chrsizes=chrsizes.txt
```

-snp describes the base counts of the short read data (not to be confused with the snp file in ANNOTATE). -refseq lists the positions with significant support for the reference base. This information is used as a measure for the callability of regions. -chrsizes indicates the length of each of the reference sequences, which is necessary for a uniform plotting of the chromosomes.

Not used in this example: -support minimum read support for both of the alleles at a single position to be considered as a het call. This depends on the coverage, usually 2 to 4 should work. -freq same as -support though this time independent of the absolute read alignments at this position. -winsize has to be adjusted and probably run multiple times with different settings in order to get a broad view of the outcome of different window sizes, also in combination with different adjustments of -freq and -support.

3 File formats

All columns described for all file formats have to be tab delimited. The number in the begining of a column description indicates the column number within the respective file format. Since we use the standard SHORE output as input for SHOREmap, some file formats feature columns not used in SHOREmap.

3.1 Input files

3.1.1 Consensus file (INTERVAL)

Summary of the read alignments. c̈onsensus_summary.txtïn the SHORE output. We have to admit that it is unfortunate that the columns used are scattered throughout the file. However, by simply using SHORE as the short read analysis tool the user will not have any file format issues.

1: <chromosome> Chromosome identifier
2: <position> Position of a base in the reference sequence
4: <coverage> Number of alignments overlapping this position
5: <A> Number of A overlapping this position
6: <C> Number of C overlapping this position
7: <G> Number of G overlapping this position
8: <T> Number of T overlapping this position
61: <ref base> Base in the reference sequence

3.1.2 Marker file (INTERVAL)

1: <chromosome> Chromosome identifier
2: <position> Position of a base in the reference sequence
4: <SNP base> Marker allele different to the reference base

3.1.3 Chromosome sizes (INTERVAL, DENOVO)

1: <chromosome> Chromosome identifier
2: <size> Chromosome size

3.1.4 Reference errors (INTERVAL, ANNOTATE)

1: <chromosome> Chromosome identifier
2: <position> Position of a reference error

3.1.5 Homozygous SNPs (ANNOTATE)

See the SHORE manual for descrition for ḧomozygous_snps.txt:.

3.1.6 Genome (ANNOTATE)

Fasta file of the reference sequence.

3.1.7 Annotation GFF (ANNOTATE)

See the general description of GFF files.

3.1.8 Minor alleles (DONOVO)

See the SHORE manual for description of m̈inor_allele_positions.txt:

3.1.9 Reference calls (DONOVO)

See the SHORE manual for description of r̈eference.txt:

3.2 Output files

3.2.1 Prioritized SNP list (ANNOTATE)

1: <chromosome> Chromosome identifier
2: <position> Position of a base in the reference sequence
3: <ref base> Base in the reference sequence
4: <mut base> Mutation
5: <Distance> Number of bp to the peak in the interval
6: <Support> Number of reads supporting the base change
7: <Concordance> Concordance of the short reads overlapping the position
8: <Quality> Highest base quality supporting the base change
9: <Type> Either NEWSNP or REFERR
10: <SeqType> Type of the DNA which is affected of the change
11: <ID> Gene identifier
12: <Isoform> Isoform of the gene
13: <Codon pos> Coding sequence position of the change
14: <Codon pos> Codon position of the change
15: <AA type> Either Syn or Nonsyn
16: <Ref AA> Amino acid of the reference
17: <Mut AA> Amino acid after base change

We appreciate any kind of feedback. Please do not hesitate to contact us in case of any problems or questions.

Korbinian Schneeberger <korbinian.schneeberger@tuebingen.mpg.de>
Stephan Ossowski <stephan.ossowski@tuebingen.mpg.de>