SHOREmap: simultaneous mapping and mutation identification by deep sequencing

Korbinian Schneeberger, Stephan Ossowski, Christa Lanz, Trine Juul, Annabeth Høgh Petersen, Kåre Lehmann Nielsen, Jan-Elo Jørgensen, Detlef Weigel & Stig Uggerhøj Andersen

Supplementary figures and text:

<table>
<thead>
<tr>
<th>Supplementary Figure 1</th>
<th>Method workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Figure 2</td>
<td>Visual output from SHOREmap DENOVO</td>
</tr>
<tr>
<td>Supplementary Table 1</td>
<td>Top 10 ranked mutations from the SHOREmap ANNOTATE output</td>
</tr>
<tr>
<td>Supplementary Table 2</td>
<td>Command line programs, parameters and run time used for the computational analysis of Illumina data</td>
</tr>
<tr>
<td>Supplementary Table 3</td>
<td>Identification of additional AT4G35090 mutant alleles</td>
</tr>
<tr>
<td>Supplementary Table 4</td>
<td>Output of SHOREmap ANNOTATE using the interval based on SHOREmap DENOVO data</td>
</tr>
<tr>
<td>Supplementary Note</td>
<td>Mapping large deletions, QTLs and dominant or recessive lethal mutations.</td>
</tr>
<tr>
<td>Supplementary Methods</td>
<td>Lab protocols and computational algorithms</td>
</tr>
<tr>
<td>Supplementary Data</td>
<td>SHORE and SHOREmap example files</td>
</tr>
</tbody>
</table>
Supplementary Figure 1. Method workflow.

In the wet-lab, the Illumina library was prepared using genomic DNA extracted from a pool of recombinants. SHORE was used to align Illumina reads to the reference sequence. Based on the alignments, base counts per position and SNPs were defined. The candidate region was then delimited using SHOREmap, with (INTERVAL) or without (DENovo) marker position information. Finally, SNPs corresponding to candidate mutations were prioritized and annotated using SHOREmap ANNOTATE to allow identification of the causal mutation.
Supplementary Figure 2. Visual output from SHOREmap DENOVO.

Red lines indicate scarcity-values in a sliding window of 200 kb.
Supplementary Table 1: Top 10 ranked mutations from the SHOREmap ANNOTATE output.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (bp)</th>
<th>Reference base</th>
<th>Mutant base</th>
<th>Distance from peak</th>
<th>Core alignment support*</th>
<th>Annotation</th>
<th>Gene AGI ID</th>
<th>ORF pos. (bp)</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>16,702,262</td>
<td>C</td>
<td>T</td>
<td>4,035</td>
<td>16</td>
<td>Coding</td>
<td>AT4G35090</td>
<td>545</td>
<td>Nonsyn.</td>
</tr>
<tr>
<td>4</td>
<td>16,940,438</td>
<td>C</td>
<td>T</td>
<td>242,211</td>
<td>17</td>
<td>Intergenic</td>
<td>AT4G35900</td>
<td>383</td>
<td>Nonsyn.</td>
</tr>
<tr>
<td>4</td>
<td>17,005,131</td>
<td>C</td>
<td>T</td>
<td>306,904</td>
<td>5</td>
<td>Coding</td>
<td>AT4G36195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16,287,342</td>
<td>C</td>
<td>T</td>
<td>410,885</td>
<td>15</td>
<td>Intergenic</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16,287,340</td>
<td>C</td>
<td>T</td>
<td>410,887</td>
<td>16</td>
<td>Intergenic</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,129,041</td>
<td>C</td>
<td>T</td>
<td>430,814</td>
<td>10</td>
<td>Intrinsic</td>
<td>AT4G36520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,240,494</td>
<td>A</td>
<td>G</td>
<td>542,267</td>
<td>21</td>
<td>Intrinsic</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,245,055</td>
<td>A</td>
<td>G</td>
<td>546,828</td>
<td>15</td>
<td>3'UTR</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16,091,082</td>
<td>C</td>
<td>T</td>
<td>607,145</td>
<td>13</td>
<td>Intergenic</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,317,842</td>
<td>G</td>
<td>T</td>
<td>619,615</td>
<td>3</td>
<td>Intergenic</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Number of core read alignments (disregarding 4 bp at either end of each read) supporting the base call.
Supplementary Table 2: Command line programs, parameters and run time used for the computational analysis of Illumina data.

Run time is estimated on one core of an Intel Xeon processor with 2.3 GHz except for MAPFLOWCELL. Memory requirements depend on genome size and range from 1 to 16 GB.

<table>
<thead>
<tr>
<th>Program</th>
<th>Parameters</th>
<th>Estimated run time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SHORE</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 1 PREPROCESS | -f TAIR8.fa
               -i IndexFolder
               -r
               -l 142                                                     | 6 h                 |
| 2 ILLUMINA2FLAT | (the -c parameter requires that the GAPipeline was run with the --with-sig2 flag) | 2 h |
| 3 MAPFLOWCELL | -n 4
               -o Run_01
               -i IndexFolder/TAIR8.fa.shore                                | 6 h (8 CPUs)        |
| 4 MERGE      | -p Run_01
               -d AlignmentFolder                                              | 30 min              |
| 5 CONSENSUS  | -n Mutant
               -f IndexFolder/TAIR8.fa.shore
               -o AnalysisFolder
               -i map.list
               -v
               -r                                                               | 1 h                 |
| **SHOREmap**|                                                                            |                     |
| 6 INTERVAL  | --consensus=consensus_summary.txt
               --marker=ler-1.marker_pos.txt
               --chrsizes=At.chrsizes.txt
               --referrors=At.ref.errors.txt                               | 1 h                 |
| 7 ANNOTATE  | --snp=homozygous_snps.txt
               --dist=SHOREmap_INTERVAL.output.txt
               --chrom=4
               --start=15,000,000
               --end=18,000,000                                              | 3 min               |
| 8 DENOVO     | --snp=minor_allele_frequency.txt
               --refseq=reference.txt
               --chrsizes=At.chrsizes.txt
               --support=4 (recommended 2 to 4)
               --freq=0.15 (recommended 0.1 to 0.2)                         | 1 h                 |
**Supplementary Table 3:**
**Identification of additional AT4G35090 mutant alleles.**

<table>
<thead>
<tr>
<th>Allele</th>
<th>CDS position</th>
<th>Codon change</th>
<th>AA change</th>
<th>Identified by</th>
</tr>
</thead>
<tbody>
<tr>
<td>at4g35090-3</td>
<td>545</td>
<td>AGT -&gt; AAT</td>
<td>Ser -&gt; Asn</td>
<td>Illumina sequencing</td>
</tr>
<tr>
<td>at4g35090-4</td>
<td>751</td>
<td>GCG -&gt; ACG</td>
<td>Ala -&gt; Thr</td>
<td>dideoxy sequencing</td>
</tr>
<tr>
<td>at4g35090-5</td>
<td>801</td>
<td>TGG -&gt; TGA</td>
<td>Trp -&gt; Stop</td>
<td>dideoxy sequencing</td>
</tr>
<tr>
<td>at4g35090-6</td>
<td>879</td>
<td>TGG -&gt; TGA</td>
<td>Trp -&gt; Stop</td>
<td>dideoxy sequencing</td>
</tr>
<tr>
<td>at4g35090-7</td>
<td>1350</td>
<td>TGG -&gt; TGA</td>
<td>Trp -&gt; Stop</td>
<td>dideoxy sequencing</td>
</tr>
</tbody>
</table>

The AT4G35090 open reading frame was sequenced in four mutants allelic to that used for Illumina sequencing. This revealed one single base pair change within the coding sequence (CDS) of each mutant. Three of the changes cause premature stop codons, and one results in an amino acid substitution.

Primers used for PCR-amplification and sequencing:
AT4G35090_F1: 5’ TTCGTCCAGAACATGCAGAC 3’
AT4G35090_R1: 5’ TTCAAAACCGTGCTTTCTCC 3’
AT4G35090_F2: 5’ GAATCCCTCATGGTGTTGT 3’
AT4G35090_R2: 5’ ACCAACTCTGGTGCTCTGT 3’

T-DNA insertion mutants displaying reduced growth rate and light green leaves have been described for AT4G35090 [7, 8].
**Supplementary Table 4:**
*Output of SHOREmap ANNOTATE using the interval based on SHOREmap DENOVO data.*

Chr: Chromosome. Position: Chromosomal location in bp. Ref base: base in the reference sequence. Mut base: base in the sequenced recombinants. Peak distance: Distance to the maximum *scarcity*-value in base pairs. Support: Number of core read alignments (disregarding 4 bp at either end of each read) supporting the base call. Annotation: features containing the mutation. Gene: Arabidopsis Genome Initiative identifiers. ORF pos.: position of the mutation within the open reading frame. Type: non-synonymous (Nonsyn) or synonymous (Syn) base change.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Ref base</th>
<th>Mut base</th>
<th>Peak distance</th>
<th>Support</th>
<th>Annotation</th>
<th>Gene</th>
<th>ORF pos</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>16,702,262</td>
<td>C</td>
<td>T</td>
<td>112,262</td>
<td>16</td>
<td>Coding</td>
<td>AT4G35090</td>
<td>545</td>
<td>Nonsyn</td>
</tr>
<tr>
<td>4</td>
<td>16,287,342</td>
<td>C</td>
<td>T</td>
<td>302,658</td>
<td>15</td>
<td>Intergenic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16,287,340</td>
<td>C</td>
<td>T</td>
<td>302,660</td>
<td>16</td>
<td>Intergenic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16,940,438</td>
<td>C</td>
<td>T</td>
<td>350,438</td>
<td>17</td>
<td>Intergenic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,005,131</td>
<td>C</td>
<td>T</td>
<td>415,131</td>
<td>5</td>
<td>Coding</td>
<td>AT4G35900</td>
<td>383</td>
<td>Nonsyn</td>
</tr>
<tr>
<td>4</td>
<td>16,091,082</td>
<td>C</td>
<td>T</td>
<td>498,918</td>
<td>13</td>
<td>Intergenic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,129,041</td>
<td>C</td>
<td>T</td>
<td>539,041</td>
<td>10</td>
<td>Intronic</td>
<td>AT4G36195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,240,494</td>
<td>A</td>
<td>G</td>
<td>650,494</td>
<td>21</td>
<td>Intronic</td>
<td>AT4G36520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17,245,055</td>
<td>A</td>
<td>G</td>
<td>655,055</td>
<td>15</td>
<td>3'UTR</td>
<td>AT4G36540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>15,889,791</td>
<td>C</td>
<td>T</td>
<td>700,209</td>
<td>13</td>
<td>Intronic</td>
<td>AT4G32920</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Note

Mapping large deletions

Small deletions (1-3 bp) are annotated in the SHOREmap ANNOTATE output table. In case larger deletions are expected, for example after fast neutron mutagenesis, they should be considered as potential causal mutations within the candidate region.

Since the number of large deletions within the candidate region will usually be relatively small, the relevant entries in the unsequenced.txt file from the standard SHORE output can easily be inspected manually. Further information about the file format can be found in the SHORE manual (http://1001genomes.org/downloads/shore.html).

Mapping recessive lethal or dominant mutations

Mapping of fully-penetrant gain-of-function mutations and recessive lethal mutations are two very interesting possible applications of our method. We propose to use a similar approach for these two cases.

Consider two polymorphic strains, “A” and “B”. The mutant allele resides in the “A” background and is designated “a”.

For dominant mutations, the mapping cross could be:

aA x BB → aB aB AB AB or
aa x BB → aB aB aB aB

Select individuals presenting the mutant phenotype (aB) and self:
aB x aB → aa aB aB BB

Select individuals presenting mutant phenotype (aB and aB) and sequence pooled DNA. Since BB individuals are discarded, 2/3 of the alleles in the candidate region will be of “A” origin - a significant overrepresentation. Candidate causal mutations would be identified as positions where 2/3 of the bases at a given position do not match the “A” or “B” reference sequence.

For lethal recessive mutations, the mapping cross would be:

aA x BB → aB aB AB AB
Select aB individuals, based on segregation of dead aa offspring, and self:
aB x aB → aa aB aB BB
Select viable individuals (aB and BB) and sequence pooled DNA. Since aa individuals are discarded, 2/3 of the alleles in the candidate region will be of “B” origin. Candidate causal mutations would be identified as positions where 1/3 of the bases at a given position do not match the “A” or “B” reference sequence.

We expect that high genome coverage, at least 22x, will be necessary for these approaches to allow accurate determination of the candidate region and reliable identification of candidate mutations.
**QTL mapping**

We consider SHOREmap well suited for QTL mapping, but its success will depend on genetic architecture (how much variation is explained by one QTL) and, perhaps even more importantly, the number of sequence changes in the QTL region relative to the reference genome.

The strategy would be to use bulk segregant analysis and sequence two pools of recombinants at each extreme of the phenotypic distribution. SHOREmap analysis of these data would then provide accurate information about the relative representation of parental alleles at each genomic locus as well as data on the differences between parents in candidate regions.

This approach is similar to eXtreme array mapping implemented by Borevitz, Chory et al., which suffered from peaks being very broad. We expect that the digital signal from sequencing combined with SHOREmap analysis will lead to an improved signal to noise ratio and much sharper peaks.
Supplementary Methods

Nuclear DNA extraction protocol

The following protocol is based on combining two previously described methods for isolating nuclei [1, 2] and then using the CTAB method [3] for extracting DNA from the nuclei.

(01) Recombinants were grown (16 h light/8 h dark) for 20 days and in darkness for 4 days. Shoots from 500 individuals displaying the mutant phenotype were collected and ground in liquid nitrogen. 2 g of powder was used as starting material.

(02) Transfer powder to 10 ml of ice-cold 1xHB plus 0.15% β-mercaptoethanol in a 50 ml blue-capped Falcon tube and redissolve by gentle swirling on ice.

(03) Filter solution through 70 µm cell strainer sieve (Falcon) and take out 300 µl (sample#1).

(04) Centrifuge filtered solution at 57 g at 4°C for 2 minutes to remove intact cells and tissue residues in a swinging bucket rotor, 15 ml Falcon tube. Transfer supernatant to a fresh tube by filtering through a 70 µm cell strainer sieve (Falcon).

(05) Add 1/20 volume 1xHB + 10% Triton X-100 and mix gently to lyse chloroplasts.

(06) Pellet nuclei by centrifugation in a swinging bucket rotor at 1800 x g for 15 min at 4°C. Take out 300 µl of the supernatant (sample#2).

(07) Resuspend pellet gently using a small paintbrush soaked in ice-cold wash buffer.

(08) Wash pellet 2 times by resuspending in wash buffer followed by centrifugation at 1800 g at 4°C for 15 min in a swinging bucket rotor.

(09) Resuspend pellet in 200 µl 1 x gradient buffer and store nuclei on ice. Take out 5 µl nuclei prep (sample#3) and store on ice.

(10) Precipitate samples# 1 and 2 with 210 µl isopropanol (15 min, max speed, 25°C), wash pellet with 500 µl 70% ethanol, air dry, resuspend in 20 µl TE.

(11) Place 250 µl of 30% Percoll into the bottom of a 2 ml round-bottomed microcentrifuge tube and underlay 250 µl of the 80% Percoll solution.

(12) Gently pipette the nuclei preparation onto the top of the 30% Percoll layer.

(13) Spin in an angled microcentrifuge at 1,000 x g (ca. 3,500 r.p.m. in a standard microcentrifuge) for 30 min at 4 °C. After centrifugation, the nuclei will be present as a white band at the interface between the 30% and 80% fractions.
(14) Remove the extraction buffer, draw nuclei from the interface with a 200 µl pipette and move to a 1.5 ml microfuge tube.

(15) Add 1 x gradient buffer to the enriched nuclei fraction to a total volume of 0.5 ml, and then underlay the solution with 250 µl of 30% Percoll.

(16) Centrifuge at 1,000g (ca. 3,500 r.p.m. in a standard microcentrifuge) for 10 min at 4 °C and remove the supernatant containing gradient buffer and Percoll.

(17) Resuspend pellet in 1 x gradient buffer (about 500µl) and transfer to a microcentrifuge tube. Centrifuge gently (<1,000g, 2000 r.p.m.) to pellet the nuclei. Remove supernatant. Take out a small sample, 1-2 µl, of the nuclei (sample#4).

(18) Add 500 µl of CTAB buffer preheated to 60°C

(19) Incubate at 60°C for 30 min, mix contents regularly during incubation.

(20) Add 500 µl chloroform:isoamylalcohol (24:1) and mix by turning tube until uniform emulsion is formed.

(21) Centrifuge 10 min, 7,000 x g, 4°C

(22) Transfer water phase (top phase) to a new tube - avoid interphase.

(23) Add 5 µl RNase (10 mg/ml) and incubate at 37°C for 30 min.

(24) Add 0.6 volumes ice-cold isopropanol and turn tube gently to mix.

(25) Incubate at -20 °C for 1 hour or overnight.

(26) Spin 6 min, 13,000 x g, 4°C. A pellet will be visible.

(27) Discard supernatant and wash with 70% EtOH. Spin 6 min, 13000 x g, 4°C.

(28) Remove ethanol and let pellet air dry.

(29) Resuspend in 55 µl TE.

(30)Optional. Compare chloroplast/genomic DNA ratios in samples 1-4 to the final sample by performing a quantitative PCR reaction using the primers detailed below.

The protocol described above yielded 55 µl of 40 ng/µl DNA, 2 µg DNA in total, from 2 grams of ground powder. Judging from quantitative PCR data, the protocol led to a ca. 3-fold increase in the genomic/chloroplast DNA ratio when compared to a CTAB DNA preparation performed on the same starting material without any steps aimed at enriching for nuclear DNA.
Buffers - sucrose nuclei isolation:
10x HB stock (homogenization buffer)
0.1M Trizma base, 0.8M KCl, 0.1M EDTA, 10mM Spermine, 10 mM spermidine. pH 9.4-9.5. Store at 4°C.

1x HB plus sucrose to 0.5 M. Can be stored at 4°C.
β-mercaptoethanol to 0.15% is added just before use.

Wash buffer:
1x HB including sucrose but without mercaptoethanol plus 0.5% Triton X-100.

Buffers - Percoll gradient centrifugation
For 1 ml 5x gradient buffer: 321 µl hexylene glycol (2-Methyl-2,4-pentanediol), 50 µl 500 mM PIPES-KOH (pH 7.0), 20 µl 1 M MgCl2, 50 µl Triton X-100, 559 µl water.

500 µl 80% Percoll: 400 µl Percoll, 100 µl 5X gradient buffer.
1000 µl 30% Percoll: 300 µl Percoll, 200 µl 5X Gradient buffer, 500 µl water.

Buffers - CTAB
For 200 ml CTAB buffer: 4.0 g CTAB (Cetyl-trimethyl-ammonium-bromide), 16.36 g NaCl, 8.0 ml 0.5M EDTA pH 8, 20 ml 1 M Tris-HCl pH 8, water to 200 ml.

Primers – Chloroplast/genomic DNA
For genomic DNA amplification: 5’ggcctgatggtcaaggatcttc3’ and 5’gcacctgcaaatccaatcaca3’.
For chloroplast DNA amplification: 5’taccggtctgtctgtcacca3’ and 5’acgccttgctctgcaacca3’.

Illumina library preparation
The library was prepared using the Illumina kit FC-102-1001 SINGLE-READ SEQ DNA PREP KIT according to the manufacturer’s instructions using 1 µg of DNA as starting material. Insert size was 150-200. Final concentration was 16.6 ng/µl. Total yield was 800 ng. Cloning and control sequencing (n=46) indicated 11% organelle contamination in the library.

Genome Analyzer data analysis
Image analysis and base calling of the Genome Analyzer output of 11 lanes was performed with the IPAR and Bustard programs from the GAPipeline software package (version 1.3) as provided by Illumina. The Col-0 reference sequence [4] (ftp://ftp.arabidopsis.org/home/tair/Sequences/ (version TAIR8)) was indexed with the SHORE software package according to its manual (Supplementary table 2 record #1). Read filtering, masking and trimming was also performed with the SHORE software package (Supplementary table 2 record #2). Best-hit read alignments to the reference sequence were performed with GenomeMapper, considering alignments of up to four mismatches including gaps (Supplementary table 2 record #3). Final base calling on the combined set of alignments produced the input files for SHOREmap (Supplementary table 2 record #4-5). Note that SHORE masks low quality bases by default.

The SHORE output file for homozygous SNPs and base count per position together with the positions of 82,127 Arabidopsis Ler-1 SNPs [5] and 1219 positions of known reference errors [6] (http://1001genomes.org/cgi-bin/col0db/col0db_app.cgi) were used as input for SHOREmap INTERVAL (Supplementary table 2 record #6). Based on the SHOREmap INTERVAL visualization (Fig. 1A) an interval of 3 MB was selected.
(Supplementary table 2 record #7). The base changes found in the interval were prioritized according to their distance to the highest \( r \)-value (Supplementary table 2 record #7).

**SHOREmap INTERVAL algorithm**

By combining information from adjacent markers in a sliding window approach, one can effectively interrogate many more recombinant chromosomes than represented at individual positions. For example, with a density of 160 markers/200 kb, and 22-fold coverage, over 3,000 chromosomes are analyzed within a 200-kb window, since the short Illumina reads constitute random sampling of independent chromosomes. Thus, the great majority of the 1,000 recombinant chromosomes represented in our DNA sample is predicted to contribute to the definition of the mapping interval. In conventional mapping of Arabidopsis mutations with individual recombinants, the final mapping interval from 1,000 chromosomes would be 0.1 cM, or on average 20 kb.

We developed the software package SHOREmap to allow simultaneous genetic mapping and identification of causal mutations, based on SHORE output of aligned reads from pools of recombinants. A map of 82,127 high-quality Col-0/Ler-1 SNPs is available (almost 1 SNP/kb in non-repetitive regions), as is a set of 1,219 errors in the reference sequence [5, 6] (http://1001genomes.org/cgi-bin/col0db/col0db_app.cgi). The SHORE base count per position, the positions of Col-0 and Ler-1 SNPs, and reference errors were used as input for SHOREmap INTERVAL (Supplementary Fig. 1, Supplementary table 2), which determines base frequencies of the Col-0 and Ler-1 alleles at each marker position and plots the parameter \( r \):

\[
\begin{align*}
    r &= \begin{cases} 
    0, & \text{if } \text{ler} = \text{col} \\
    -\text{ler}, & \text{if } \text{col} = 0 \\
    \text{col}, & \text{if } \text{ler} = 0 \\
    \text{col}/\text{ler}, & \text{if } \text{col} > \text{ler} \\
    -(\text{ler}/\text{col}), & \text{if } \text{ler} > \text{col}
    \end{cases}
\end{align*}
\]

where \( \text{ler} \) and \( \text{col} \) are the sums of reads supporting the Ler-1 and Col-0 alleles at each marker position. Ten plots with different sliding window parameters are automatically generated by SHOREmap INTERVAL, based on normalized \( r \)-values.

**SHOREmap ANNOTATE**

SHOREmap ANNOTATE ranks base pair substitutions according to their distance to the highest \( r \)-value (the peak in Fig. 1A-B), and outputs the effect of base changes according to feature annotation (Supplementary Fig. 1, Supplementary Table 1). Any General Feature Format (GFF) file can be applied; in this case, Arabidopsis TAIR8 annotation (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/) was used.

We note that bulk segregant analysis makes the method robust to occasional mis-phenotyping of recombinant individuals, which can be a considerable source of error in conventional mapping. If Ler-1 reads do not disappear near the peak of \( r \)-values, this would indicate the presence of mis-phenotyped individuals in the mapping population. This information can in turn be taken into account when identifying potential causative mutations.
**SHOREmap DENOVO algorithm**

SHOREmap DENOVO determines marker positions based on the base count from the SHORE output. Each position featuring four or more reads (22-fold coverage) or two or more reads (11-fold coverage) from two alleles and a frequency of at least 15% for both of these alleles will be recorded as a marker position. The number of such marker positions in and near the candidate region will be reduced due to the homozygous nature of this interval. SHOREmap DENOVO determines the scarcity within a sliding window as the average of the sum of position-wise distance to the nearest marker (dist\(_m\)) multiplied with the inverse sum of frequencies of the Ler-1 allele at the predicted marker positions (ler):

\[
\text{scarcity} = \frac{\sum \text{dist}_m \times \frac{1}{\sum \text{ler} \times \text{ref}^2}}{\text{window size}},
\]

where

\[
\text{ref} = \frac{\text{ref calls}}{\text{window size}}
\]

is the percentage of reference calls within the sliding window used to normalize for the accessibility of the genomic region to base calling (mostly influenced by repetitiveness).

The output plot from SHOREmap DENOVO is shown in Figure 1B and Supplementary Figure 2. The commands used can be found in Supplementary table 2 record #8. A target interval of 3 Mb was chosen based on the SHOREmap DENOVO visualization for which all included base changes were ranked using SHOREmap ANNOTATE. Supplementary table 4 lists the prioritized mutations along with their effects on genes. The causal mutation was ranked number one for both tested datasets (22-fold and 11-fold).
Supplementary Data

Files from SHORE and SHOREmap analysis

Output files from the data analysis are available from:
http://1001genomes.org/downloads/shore
Supplementary References