**Supplementary Information**

Homotools: a suite of genomic tools for homologous retrieval and comparison

Liu et al.

Scripts in this Supplementary Information are examples for showing how to run homotools modules. Example data can be downloaded from <https://github.com/liu3zhenlab/homotools>. Sampled subset datasets (1 bp to 10 Mb in chromosome 1) were extracted from genomic data of five maize inbred lines, namely, B73, A188, CML333, Ki3, and P39. Genomic data include genomic DNA sequence and annotations. Five directories were created in the directory of “1\_data”. We assume that the current directory is under the same parental directory as “1\_data”. In all scripts here, <homotools\_path> needs to be modified to your own path to access homotools.

**1. A geneseq example**

We first extract the B73 genomic sequence of the gene Zm00001eb001720 and its gene annotation.

|  |
| --- |
| # input data source\_ref=../1\_data/B73/B73.fasta  source\_gtf=../1\_data/B73/B73.gtf prefix=B73 gene=Zm00001eb001720  # gene extraction perl <homotools\_path>/geneseq \  --fas $source\_ref \  --gtf $source\_gtf \  --gene $gene \  --prefix $prefix |

The FASTA sequence of the gene and annotation were extracted from the reference and GTF files:

|  |
| --- |
| B73.1.Zm00001eb001720.fasta B73.3.Zm00001eb001720.original.bed B73.3.Zm00001eb001720.original.gtf |

In addition, two directories were produced. One including GTF and BED files with new coordinates adjusted to the extracted genomic sequence of the gene. The other contains genomic sequences with exon sequences capitalized and coding sequences bolded. The coding sequencing was implemented with the Markdown format with "\*\*". Therefore, the file needs to be opened as a Markdown file to see bold highlights.

|  |
| --- |
| B73.4.pos.adjusted.gtf.bed # directory  Zm00001eb001720\_T001.adjusted.bed  Zm00001eb001720\_T001.adjusted.gtf  Zm00001eb001720\_T002.adjusted.bed  Zm00001eb001720\_T002.adjusted.gtf  Zm00001eb001720\_T003.adjusted.bed  Zm00001eb001720\_T003.adjusted.gtf  B73.5.cdna.highlighted # directory  Zm00001eb001720\_T001.highlighted.fasta  Zm00001eb001720\_T002.highlighted.fasta  Zm00001eb001720\_T003.highlighted.fasta |

**2. A homocomp example**

**2.1 Homocomp with the A188 genome**

Now we have the genomic DNA sequence of the gene. We can search the homologous sequence in the A188 genome using module homocomp.

|  |
| --- |
| query=B73/B73.1.Zm00001eb001720.fasta ref\_db=../1\_data/A188/A188.fasta prefix=A188comp  perl <homotools\_path>/homocomp \  --prefix $prefix \  --query $query \  --db $ref\_db |

Here are outputs:

|  |
| --- |
| A188comp.1.blastn A188comp.2.blastn.filt A188comp.3.target.bed A188comp.4.target.fas A188comp.5a.delta A188comp.5b.delta.txt A188comp.5c.filt.delta.txt A188comp.5d.uniq.delta A188comp.5e.uniq.filt.delta A188comp.5f.uniq.filt.delta.txt A188comp.o1.alnplot.pdf A188comp.o2.filt.alnplot.pdf A188comp.o3.uniq.filt.alnplot.pdf A188comp.o4.dotplot.pdf A188comp.o5.filt.dotplot.pdf A188comp.o6.uniq.filt.dotplot.pdf |

Files suffixed with “txt” are outputs of NUCMER alignments between two sequences. xxx.5b.delta.txt includes raw alignments; xxx.5c.filt.delta.txt includes alignments after filtering based on the input criteria (identity and coverage); xxx.5f.uniq.filt.delta.txt includes unique alignments after filtering based on the input criteria. These alignments are used to produce alignment plots (alnplots) and dotplots.

**2.2 Homocomp with the other 3 genomes**

Similar to homocomp analysis using the A188 genome, the genomes of CML333, Ki3, P39 can be used for the identification of the best-hit homolog and visualization of alignments.

|  |
| --- |
| #!/bin/bash genolist=" CML333 Ki3 P39 " query=B73/B73.1.Zm00001eb001720.fasta for geno in $genolist; do  ref\_db=../1\_data/${geno}/${geno}.fasta  prefix=${geno}comp  perl <homotools\_path>/homocomp \  --prefix $prefix \  --query $query \  --db $ref\_db done |

**3. A homomine example**

Homomine can search the best homolog of Zm00001eb001720 (B73) in the A188 genome, identify and annotate polymorphisms, including SNP, INDEL, and structural variation (SV). To run homomine, databases are required to be organized in a certain format. First, a directory storing genomic data of B73 as the query includes:

* + a FASTA genome sequence, required (B73.fasta)
  + a BLAST+ database of B73.fasta, required (B73.fasta.n\*)
  + a GTF gene annotation file, required (B73.gtf)
  + A TE annotation file, optional (B73.TE.gff3)

Secondly, a directory storing genomic data of A188 as the target includes:

* + a FASTA genome sequence, required (A188.fasta)
  + a BLAST+ database of A188.fasta, required (a set of A188.fasta.n\*)
  + a GTF gene annotation file, required (A188.gtf)
  + A TE annotation file, optional (A188.TE.gff3)

Note: all file suffixes in the databases are fixed. “B73 '' must be specified using the parameter “qrybase”, and “A188” must be specified using the parameter “tgtbase”. The BLAST+ databases should be created by using “makeblastdb” and the same version of BLAST+ should be used as the one installed in homotools. Here is an example code to create a BLAST+ database.

|  |
| --- |
| makeblastdb -in B73.fasta -dbtype nucl |

Once the data directories and files are prepared, the search of the B73 gene Zm00001eb001720 in the A188 genome is as simple as the following script.

|  |
| --- |
| qrygene=Zm00001eb001720 datadir=../1\_data perl <homotools\_path>/homomine \  --qrygene $qrygene \  --qrydir $datadir/B73 \  --qrybase B73 \  --tgtdir $datadir/A188 \  --tgtbase A188 \  --prefix B73vsA188HM |

Major outputs

|  |
| --- |
| B73vsA188HM.homomine.report.html  Zm00001eb001720\_A188.homo.txt |

“Zm00001eb001720\_A188.homo.txt” showed the information of the best hit in the A188 genome as below.

|  |
| --- |
| chr start end strand gene genome chr1 5075499 5081601 + Zm00001eb001720 B73 chr1 5211165 5220046 + Zm00056a000187 A188 |

“B73vsA188HM.homomine.report.html” was provided in **Supplementary Data**.

**4. A homostack example**

We have identified the best-hit homolog in the four non-B73 genomes with the Zm00001eb001720 region of B73. Alignments of these homologous sequences and Zm00001eb001720 of B73 can be sequentially stacked through homostack. The B73 annotation of the gene from the geneseq output (suffixed with adjusted.bed) can be used to highlight exon and coding regions.

|  |
| --- |
| perl <homotools\_path>/homostack \  --seq B73/B73.1.Zm00001eb001720.fasta \  --annot B73/B73.4.pos.adjusted.gtf.bed/Zm00001eb001720\_T001.adjusted.bed \  --seq CML333comp/CML333comp.4.target.fas --annot none \  --seq Ki3comp/Ki3comp.4.target.fas --annot none \  --seq A188comp/A188comp.4.target.fas --annot none \  --seq P39comp/P39comp.4.target.fas --annot none \  --prefix hsOut |

Major outputs

|  |
| --- |
| hsOut.2.nucmer.0.txt  hsOut.2.nucmer.1.txt  hsOut.2.nucmer.2.txt  hsOut.2.nucmer.3.txt  hsOut.3.alnstack.pdf |

“hsOut.2.nucmer\*” provides the detailed alignments for a pair. In total, four pairs of alignments were conducted.

“hsOut.3.alnstack.pdf” visualizes these alignments.

**5. A homograph example**

The homologous sequences used in homostack analysis can be used for homograph analysis, which clusters homologs, conducts multiple sequence alignment (MSA) and visualizes MSA.

We performed homograph analysis using B73 Zm00001eb001720 as the reference. Its annotation from geneseq was input using the parameter “genebed”. Prior to run homograph, the other non-B73 homologs were copied to a directory. Each homologous sequence was saved in a FASTA file. Here the names of the homologs were simplified to a better visualization.

|  |
| --- |
| #!/bin/bash datadir=4i\_fas if [ ! -d $datadir ]; then  mkdir $datadir fi  pushd $datadir cp ../\*comp/\*4.target.fas .  # simplify sequence names: for fas in \*fas; do  sed -i 's/\_.\*//g' $fas done popd  perl <homotools\_path>/homograph \  --ref B73/B73.1.Zm00001eb001720.fasta \  --genebed B73/B73.4.pos.adjusted.gtf.bed/Zm00001eb001720\_T001.adjusted.bed \  --fasdir $datadir \  --prefix hg |

Major outputs

|  |
| --- |
| hg.2.msa  hg.6.msa.pdf  hg.7.block.haplotype.fasta  hg.8.block.genotype.txt |

“hg.2.msa” is the MSA output in the FASTA format.

“hg.6.msa.pdf” visualizes MSA and color-codes polymorphic levels per block of each individual/taxon.

“hg.7.block.haplotype.fasta” provides sequences of every haplotype of each block.

“Hg.8.block.genotype.txt” are genotyping results of each block in each taxon.