TSS assembly pipeline for Hv_EPDnew_000

Introduction

This document provides a technical description of the transcription start site assembly pipeline that was used to generate the EPDnew version 000 for *H. vulgare*.

Source Data

Gene annotation resource:

Name		Genome Assembly		Promoters	Genes	PMID	Access data		
UCSC/RefSeq annotation	gene	Apr MorexV3	2021	21193	20502	37784172	SOURCE	DOC	DATA1 DATA2

Experimental data:

Name	Type	Samples	Tags	PMID	Access data		
pavlu23	CAGE	3	151'754'815	38173877	SOURCE	DOC	DATA

Description of procedures and intermediate data files

1. Genome Annototion Download

The "RefSeq gene predictions from NCBI" track for the MorexV3_pseudomolecules_assembly Apr. 2021 (GCF_904849725.1) was downloaded from UCSC, file:

GCF_904849725.1_MorexV3_pseudomolecules_assembly.ncbiRefSeq.bb.

2. UCSC/RefSeq TSS and codon start collection

TSS positions and CDS stast positions of complete open reading frames were extracted from RefSeq gene annotation and refrmatted into sga format

ucsc_transcriptStart_list.sga.gz
ucsc_CDSstart_list.sga.gz

The six fields of these files contain the following information:

- NCBI/RefSeq chromosome id
- "TSS" or CDSstart
- position
- strand ("+" or "-")
- "1"
- RefSeqID..geneName

Note that the second and fifth field are invariant in both files.

3. Rawdata download and tag mapping to barley genome

Raw CAGE data were downloaded from SRA in FASTQ format, using SRX identifiers provided in GEO entry GSE227219. The sequence tags were subsequently mapped to the Barly MorexV3 genome using Bowtie2 v1.2.2. SAM output files were reformatted to SGA format.

4. CAGE tag peak calling

All six CAGE samples were merged into a single file. Candidate TSS were selected in two stages using the programs *chippeak* and *chipscore* from ChIP-Seq v. 1.5.5.

chippeak was used with the following options and parameters:

- window width = 1
- vicinity range = 200
- count cutoff = 9999999
- threshold = 5

This selects candidate peak summit position, which have at least 5 CAGE tag mapped to it, and which constitute a maximum within a range of ±200 bp. This preliminary list, together with the CAGE tag input file, was subsequently processed with *chipscore* in order to select peak summits which are covered by at least 50 tags within a surrounding range of ±50 bp. The peaks from step 1 were used as reference features for *chipscore* , and the merged CAGE tags from all CAGE samples as target features.

5. TSS validation and attribution to gene

Candidate TSS of annotated genes where then selected from the preliminary list obtained in the previous step using proximity mapping: All peak summits were retained, which are located either

between 50 bp upstream and 200 bp downstream from a RefSEQ annotated TSS

or

no more than 300 bp upstreamd of a RefSeq annotated CDS start site.

These ranges were empirically determined by analyzing the input CAGE tag distributions around RefSeq annotated TSS and CSD start sites.

6. Final EPDnew collection

The 21193 experimentally validated promoters were stored in the EPDnew database, which can be downloaded from our ftp site. Scientists are welcome to use our other tools ChIP-Seq (for correlation analysis) and SSA (for motif analysis around promoters) to analyze the EPDnew database.