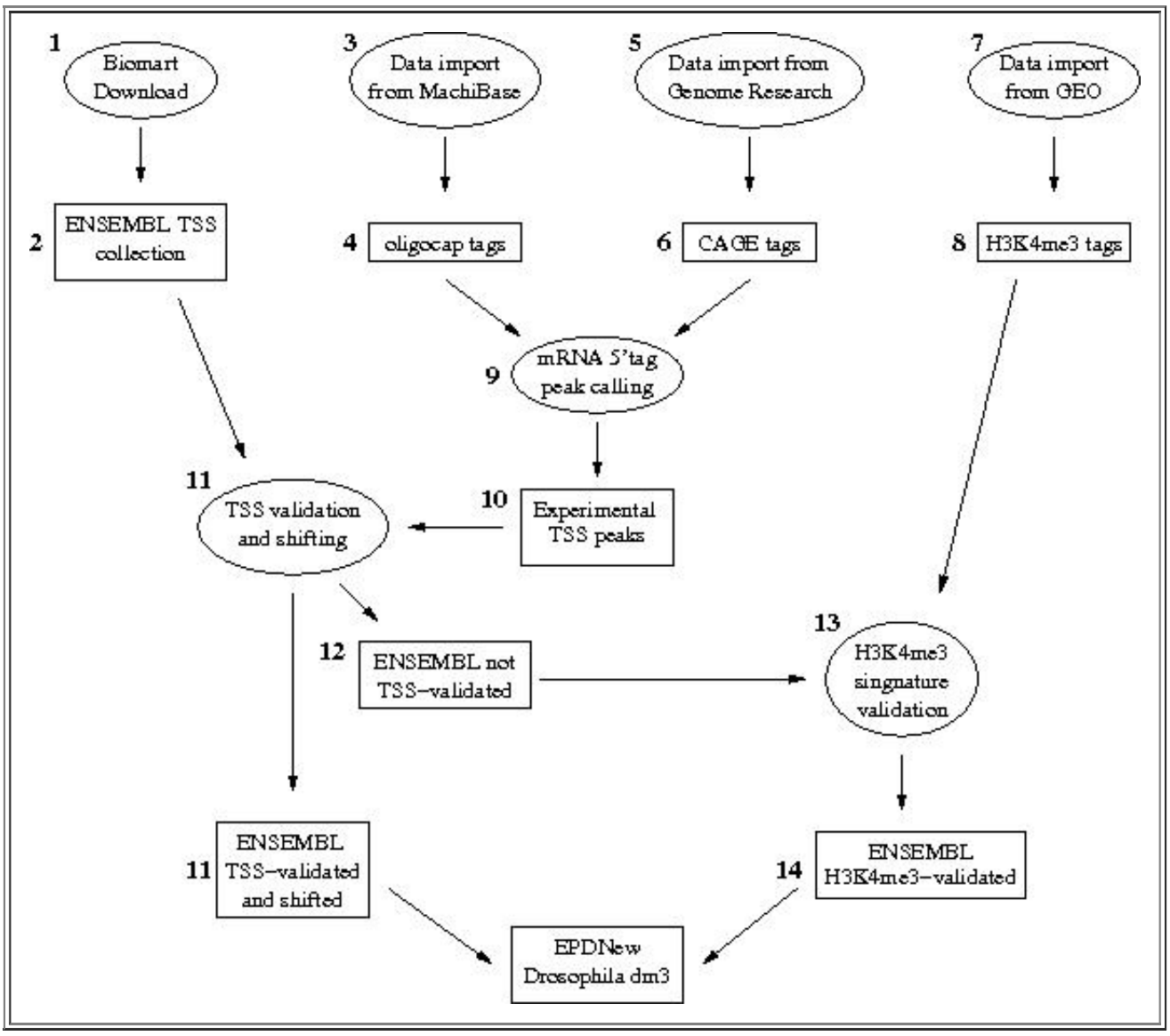
TSS assembly pipeline for Dm_EPDnew_001

Introduction

This document provides a technical description of the transcription start site assembly pipeline that was used to generate EPDnew version 001 for Drosophila melanogaster genome assembly dm3.

Source Data

Assembly pipeline overview



Description of procedures and intermediate data files

1. Biomart Download

Data was downloaded from sep2011.archive.ensembl.org/biomart/martview/ selecting the following attributes:

1. Ensembl Gene ID
2. Ensembl Transcript ID
3. Chromosome Name
4. Strand
5. Transcript Start (bp)
6. Transcript End (bp)
7. Gene Start (bp)
8. Gene End (bp)
9. Status (transcript)
10. Status (gene)
11. Associated Gene Name

Then, transcrips have been filtered according to the following rules:

1. Transcripts of protein coding genes only
2. Transcript length > 0 [Transcript Start different from Transcript End]
3. Transcript lies on full chromosomes
4. Gene must have a 5' UTR [Transcript Start different from Gene Start]
5. Genes must be annotated [Associated Gene Name present]
6. Gene and transcripts status known

Gene names were taken from the field "Associated Gene Name". Since the EPD format doesn't allow gene names longer than 18 characters, we checked whether the names repsected this limitation.
Transcripts with the same TSS position were merged under a common ID. As a conseguence of this, from the 23850 transcrips originally present in the ENSEMBL database, 5953 were merged, leaving 17897 uniquely mapped promoters in the input list.

2. EMBL TSS collection

The ENSEMBL TSS collection is stored as a tab-deliminated text file conforming to the SGA format under the name:

filename.

The six field contain the following kinds of information:

- NCBI/RefSeq chromosome id
- "TSS"
- position
- strand ("+" or "-")
- "1"
- gene name.

Note that the second and forth fields are invariant.

3. Data import from MachiBase

MachiBase data were generated with the oligo-capping technology. The source data were downloaded from:

http://download.utgenome.org/pub/machibase/tssExp.tar.gz

According to the readme file included in the tar archive, the 5' end tags were mapped to the Drosophila genome using BLAT as alignment tool allowing for up to three mismatches.

4. oligocap tags

The compressed version of this file is available from the MGA archive (see above) under the name:

all_oligocap.sga.gz.

5. Data import from Genome Research

Mapped sequence tags were extracted from Supplementary Data File 1 available from Genome Research at:

http://genome.cshlp.org/content/21/2/182/suppl/DC1

The downloaded source file is in SAM format and has been generated with the tag mapping program StatMap as described in the article cyted above. We extracted all

tags with mapping quality scores greater or equal to 30.

6. CAGE tags

The compressed version of this file is available from our ftp site (see above link) with the name:

embryo_cage.sga.gz

7. Data import from GEO

BED files for the GEO serie GSE19325 were downloaded from GEO ftp site and converted into SGA file using in house software.

8. H3K4me3 tags

The compressed version of this file is available from our ftp site (see above link) with the name:

GSM480156_dm3-S2-H3K4me3.sga.gz

9. mRNA 5' tags peak calling

CAGE tags and oligocap tags have been merged into a unique file with the following command:

sort -m -s -k1,1 -k3,3n -k4,4 embryo_cage.sga all_oligocap.sga > all_data.sga

Peak calling for the merged file has been carried out using ChIP-Peak on-line tool with the following parameters:

- Window width = 100
- Vicinity range = 200
- Peak refine = Y
- Count cutoff = 9999999
- Threshold = 8

The sga file containing the list of peaks can be downloaded here.

11. TSS validation and shifting

The source data (mRNA peaks and ENSEMBL TSS) was then processed in a pipeline aiming at validating transcription start sites with mRNA peaks. An ENSEMBL TSS was experimentally confirmed if an mRNA peak lied in a window of 100 bp around it. The validated TSS was then shifted to the nearest base with the higher tag density. After this step, the total number of validated promoters was 10389.

The list of validated and shifted promoters can be downloaded here.

12. ENSEMBL not-validated TSS

The total number of non mRNA validated TSS was 7508. These promoters were not discarded but were the subject of the following validation step.

13. H3K4me3 signature validation

H3K4me3 histone mark was used as a marker for promoter validation. The 7508 promoters that were not validated by mRNA signature were scored according to the presence of H3K4me3 histon mark near the TSS. This procedure recover more than 2000 promoters bringing the total number of validated promoters to 12435.

15. EPDnew collection

The 12435 experimentally validated promoter were stored in the EPDnew database that can be downloaded from our ftp site. Scientist are wellcome to use our other tools ChIP-Seq (for correlation analysis) and SSA (for motifs analysis around promoters) to analyse EPDnew database.