1st information letter

EPIC - Eukaryotic Promoter and transcription Initiation site prediction Challenge

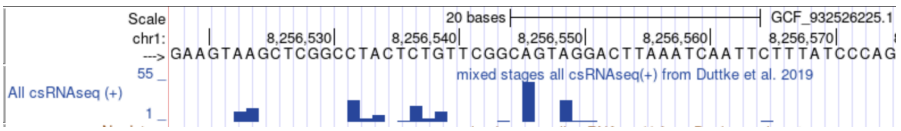
Organizers: Philipp Bucher, Ivan V. Kulakovskiy, Sascha Duttke,
Vsevolod Makeev, Ivo Grosse

Purpose of the challenge

We aim to assemble and evaluate a community-driven suite of computational methods for predicting genome-wide transcription initiation profiles in non-model eukaryotic species, in the hope that such a research infrastructure would advance understanding of promoter architectures and the molecular mechanisms of transcriptional initiation site selection across eukaryotes.

The challenge in a Nutshell

Transcription initiates at specific positions in the genome at different rates. There are many established and emerging methods to measure transcription initiation rates genome-wide and at single-base resolution. The resulting TSS data have the structure of a count histogram along the genome; see genome browser snapshot below:



Example: csRNA-seq data for *Nematostella vectensis* chr1:8,256,517-8,256,575 (GEO: **GSE135498**).

The challenge goal is to predict TSS counts at single-base resolution from the genome sequence. Participants will receive TSS counts for a part of the genome, the training chromosomes or contigs, and are requested to submit counts for the remaining test chromosomes.

The challenge data

Unpublished csRNA-seq data for five metazoan species are provided by Sascha Duttke's lab. The challenge will be using the data from 5 eukaryotic species, for which transcription initiation has never been characterized before. csRNA-seq is a technology that measures transcription initiation events in the nucleus at single-base resolution. Genome sequences, gene annotations, and RepeatMasker files for the corresponding species are already available from GenArk, both for download and genome browser viewing. The csRNA-seq data will be provided at the challenge kick-off date for training chromosomes only.

Challenge design

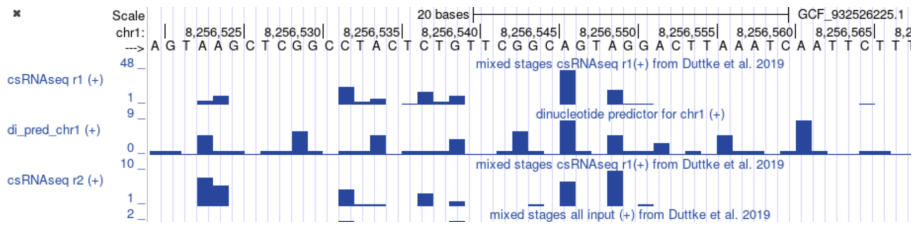
The IBIS challenge https://ibis.autosome.org/home partly inspired the EPIC design. During the leaderboard phase, participants can upload predictions for online scoring, and team rankings will be automatically posted on the challenge website. By the challenge closing date, participants will have submitted their final predictions, along with a method write-up. It is not allowed to use any other inputs than the genome sequence, although pretrained genomics models (such as AlphaGenome or Evo2) can be used. During the post-challenge analysis, top-scoring teams will be invited to participate in the writing of a post-challenge paper. Notably, they will be requested to submit the predictor software (via a GitHub repo) as an easily portable tool that scans a FASTA sequence file and returns predictions in the challenge format, and a reproducible code to train the predictor using the challenge data. All challenge participants will be co-authors via the EPIC consortium members list.

Challenge submissions, evaluation and performance metrics

The evaluation procedure is illustrated with a simple dinucleotide predictor, which assigns a count value to the second position of each dinucleotide, wherever it occurs in the genome:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	1	0	9	1	6	1	1	3	0	1	5	1	4	0

The counts returned by this predictor are illustrated in the genome browser screenshot below:



And here are the corresponding numerical data:

```
rep1  0 0 0 6 13 0 0 0 0 0 0 0 25 4 9 0 1 18 5 12 0 0 0 0 0 0 48 0 0 20 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
dinuc 1 1 0 5  1 1 0 1 1 6 1 0  1 1 5 0 1  1 1  4 0 0 1 6 1 0  9 1 0  5 1 1 3 0 1 0 5 1 1 0 1 9 1 0 0 1 1 0
rep2  0 0 0 8  6 0 0 0 0 0 0 0  5 1 1 0 0  4 0  2 0 0 0 0 1 0  7 0 0 10 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

In the example, we use non-negative integer arrays; in the challenge, the ground truth data will be kept in its raw form (integer read counts), and the predictions should be supplied as fixed-precision real values (0.xxxx).

As performance measures, we will use the area under precision-recall curve (AUPRC) for classification, and Spearman's correlation for regression.

Prospective timeline

Presentation and kick-off on Sep 4, 2026 at the EPD 40th anniversary symposium; see https://epd.expasy.org/EPD_symposium2026/