Enhort: A Platform for Deep Analysis of Genomic Positions

Michael Menzel [1]
michael.menzel@mni.thm.de

Andreas Gogol-Döring [1]
andreas.gogol-doering@mni.thm.de

TECHNISCHE HOCHSCHULE MITTELHESSEN
UNIVERSITY OF APPLIED SCIENCES

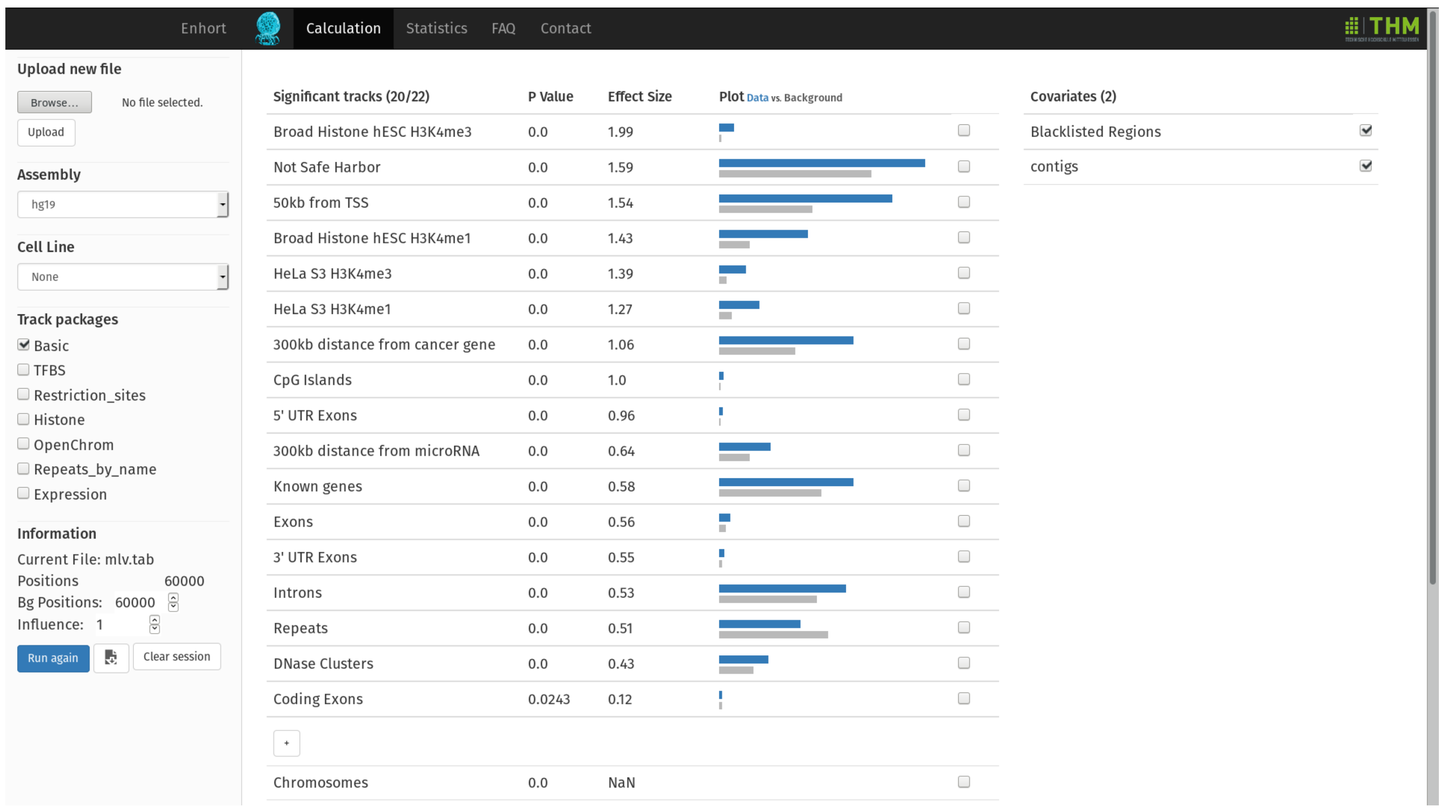
German Conference on Bioinformatics
12 – 15 September 2016 Berlin

1 Abstract

The rise of high-throughput methods in genomic research greatly expanded the amount of genomic annotations. Using annotations to characterise genomic positions, e.g. protein binding, virus integration, or differential methylation, the quantities of annotations and sites generated by high-throughput methods are way too large for a manual inspection.

Here, we present Enhort, a novel, user-friendly software tool for the deep analysis of large amounts of genomic positions. It uses a complex but easy-to-use mechanism for adjusting statistical background models according to experimental conditions or specific scientific questions.

Enhort is free and publicly available online at `www.enhort.mni.thm.de`.

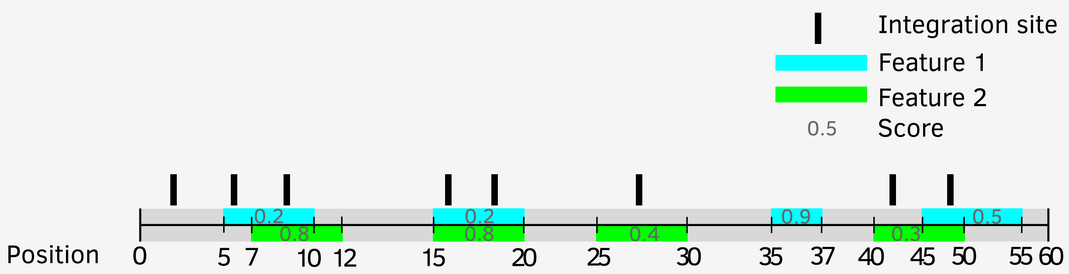


Screenshot of the application with a data set of MLV from LaFave et al. against a background model with blacklisted regions and contigs to prevent background model positions in non-sequenceable regions. The annotation data was retrieved from the UCSC Genome Browser Database (Rosenbloom et al.).

2 Usage

Calculation

Using annotation features, e.g. genes or conserved regions that define intervals with a start and end position on the genome (turquois and green) and integration sites (black bars), the application calculates the integration frequency for each feature. One of the positions is outside of both features, 5 positions are inside feature 1, 6 positions are in feature 2, and 4 of the positions are also in both features:



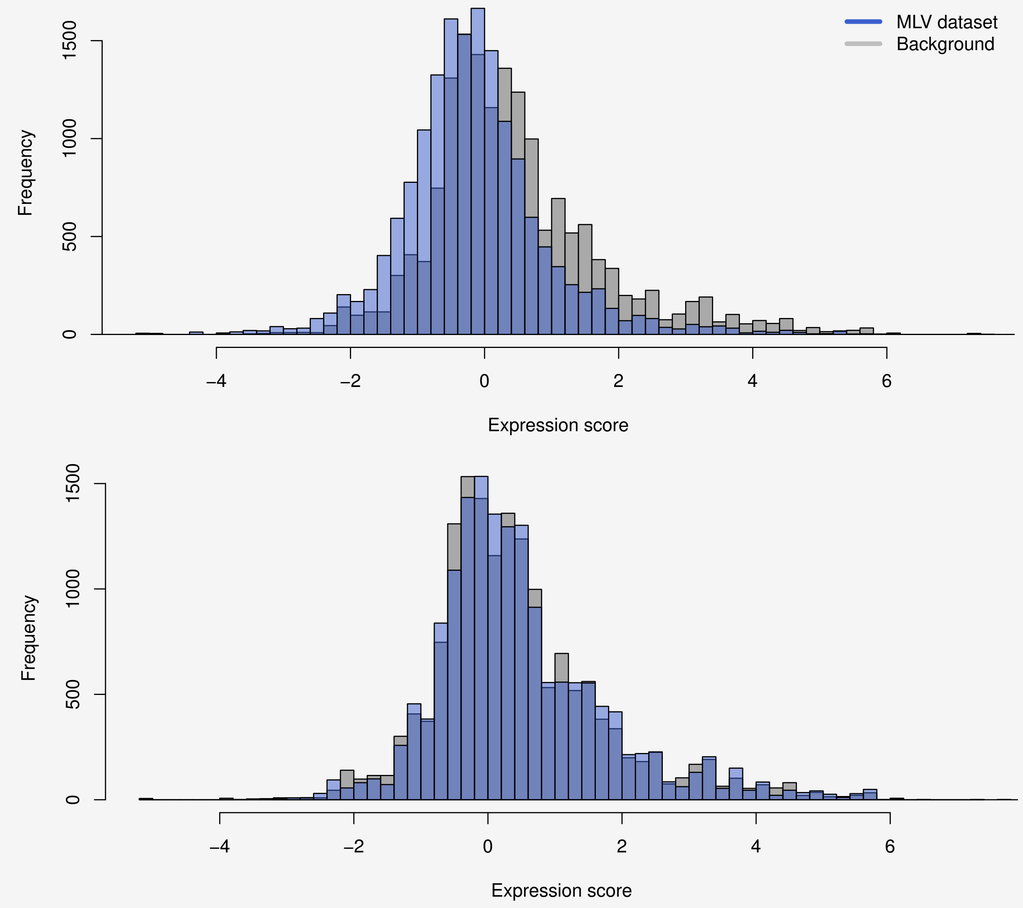
To evaluate if the integration of a data set is random across a feature, a set of random positions, called background model, is generated and also compared to the annotations. The observed data and background model are then compared using a statistical test (χ^2), resulting in a p-value used to identify tracks, on which the observed positions significantly deviate from random.

Similar results can be generated using the BEDTools suite (Quinlan et al.), however, Enhort is specialised on position analysis in reference to annotations. No manual work or programming is needed to utilise the tool.

Background Models

Background models can be manipulated by the user. E. g. virus integration sites are identified using sequencing, which rely on restriction enzymes. Large regions without cutting sites are not sequenceable. To adjust the background model, tracks of the used restriction enzymes can be selected to influence the generation of random positions so that they show the same integration frequency as the observed data for these tracks. The selected tracks are called covariates.

The following figure shows the distribution of expression scores selected by MLV integration sites from LaFave et al. against a random background with no covariate (upper) and a background model with the expression scores as covariate (lower):



References:

M. C. LaFave, G. K. Varshney, D. E. Gildea, T. G. Wolfsberg, A. D. Baxevanis, and S. M. Burgess. MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Research, 42, 2014.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 2010.

Rosenbloom, Kate R., et al. The UCSC genome browser database: 2015 update. Nucleic acids research, 43, 2015.

[1] Technische Hochschule Mittelhessen – University of Applied Sciences – www.thm.de