ARGOS: Intragenomic Sequence Similarity Patterns assist in *de novo* Genome Annotation

Niko Popitsch, Philipp Rescheneder, Miguel Gallach, **Contact: niko.popitsch@univie.ac.at**
Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, A-1030 Vienna, Austria

CIBIV MFPL
Center for Integrative Bioinformatics Vienna

Abstract

Identification of functional DNA sequence elements is a common request in comparative, evolutionary and functional genomics. This task, however, is particularly hard when *a priori* biological information is scarce, as normally happens during *de novo* genome annotation or when searching for novel regulatory elements.
Here, we show how the characterization of intragenomic sequence similarities allows the unbiased (no repeat library required) detection of (putative) functional elements. Our method combines three complementary genome-wide signals, calculated from alignment scores and positions of synthetic, overlapping reads that were extracted from and then aligned back to a genomic sequence (see workflow). From these data we calculate three base-resolution signals:

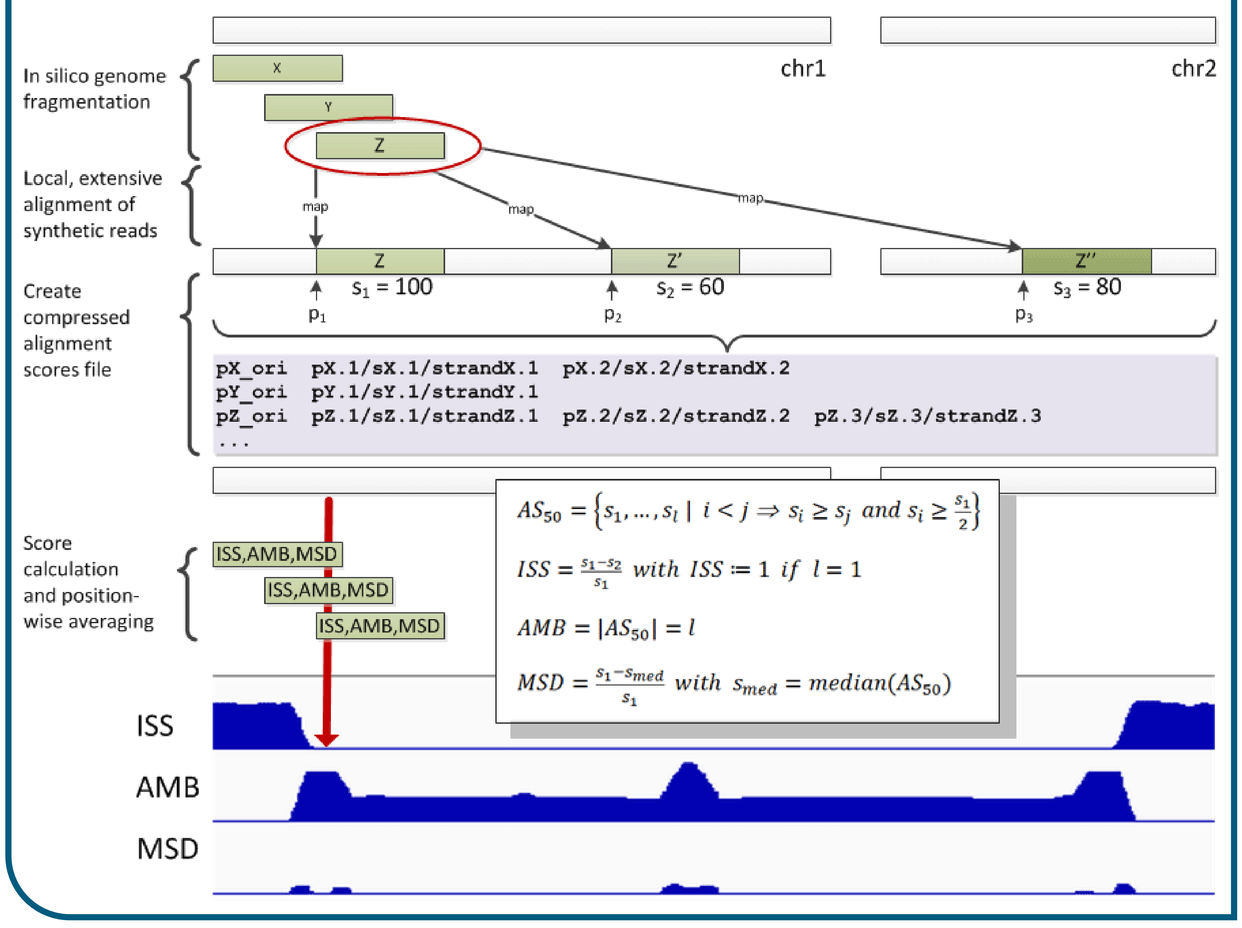
ISS: maximum similarity of a region to any other region

AMB: number of similar (min 50% of the maximum alignment score) regions

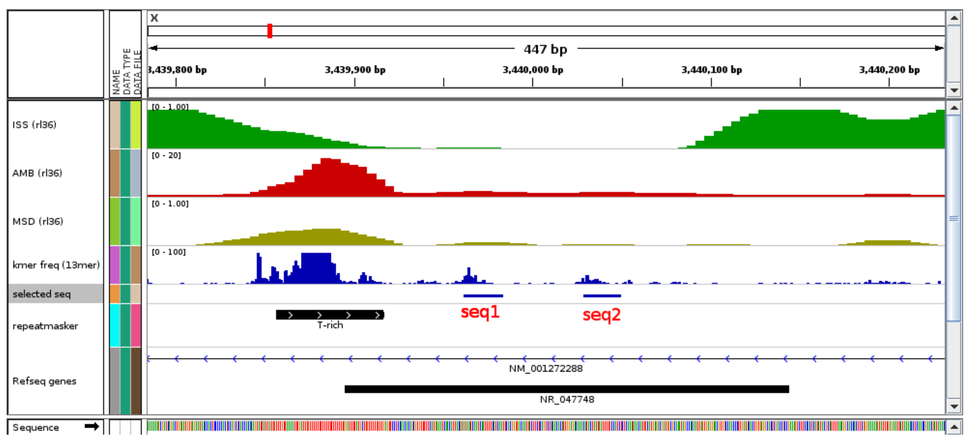
MSD: A similarity measure of the majority of these related regions

Additional we calculate all signals based on a restricted genomic context (a genomic window) to describe local effects and also calculate the k-mer frequency per genomic position.

Workflow



Example 1: D. melanogaster ncRNA FBtr0307366



This example shows the non-coding RNA FBtr0307366 in *d. melanogaster*. A large portion of its sequence is non-unique (low ISS score) and it contains a T-rich subsequence at its 3′ end that is found several times in the genome (AMB/MSD peaks, also found by repeatmasker). The k-mer frequency track shows two short motifs with increased frequency (seq1 and seq2) which we BLASTed and found to be overrepresented in 3′UTRs and underrepresented in exonic regions.

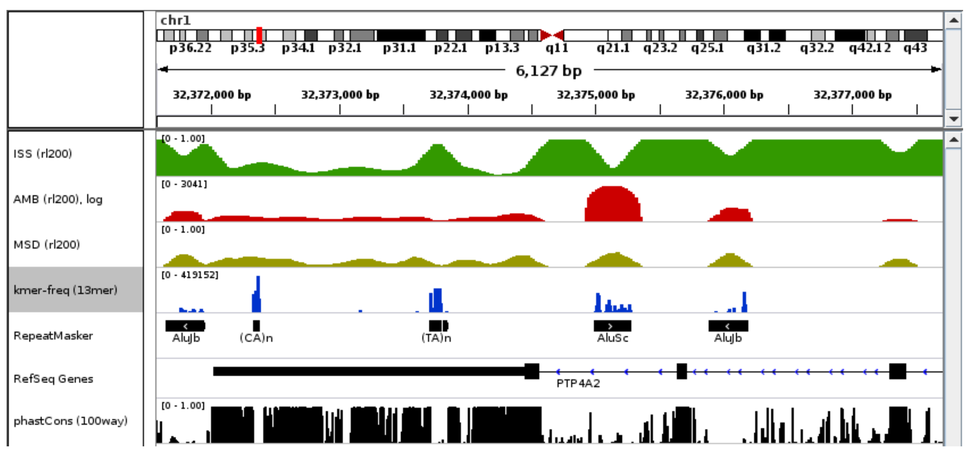
Example 1: Motif analysis

Seq1: TATAATCAATCAATTTTCATCAA
Seq2: ACATACTATATATTTTGCCATC

Type	Seq1	Seq2	Total OBS	EXP	[OBS-EXP]/EXP
ncRNA	4	3	7	-	-
3′UTR	10	7	17	7.7	120.78%
5′UTR	1	1	2	4.9	-59.18%
exonic	1	3	4	30.3	-86.80%
intronic	52	31	83	66.0	25.76%
intergenic	29	22	51	48.1	6.03%

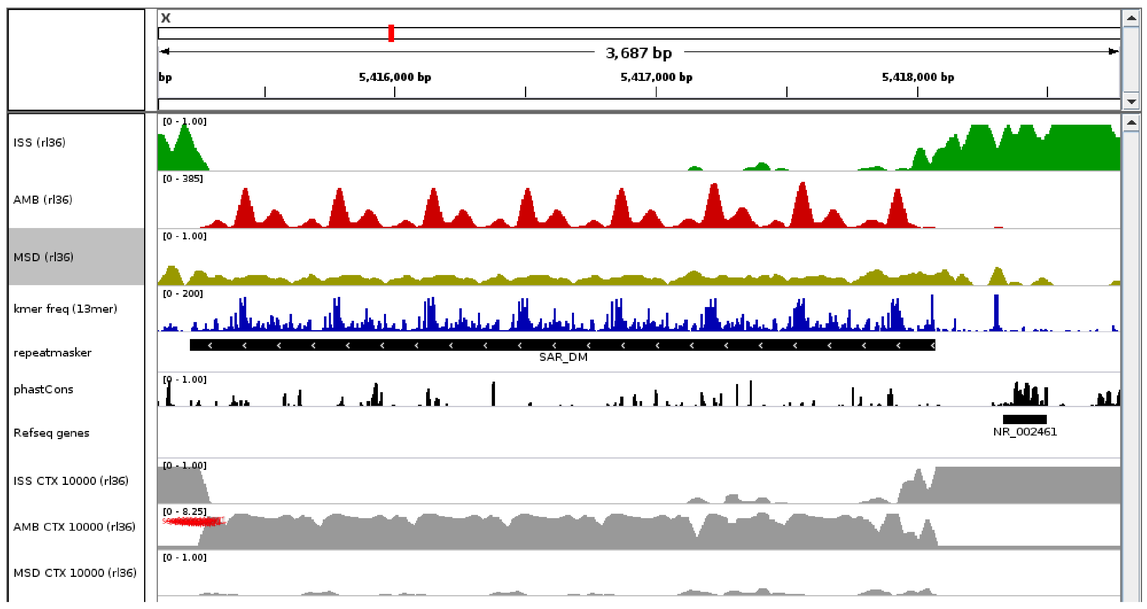
$\chi^2 = 40.42; df = 4; p - value = 3.54e^{-8}$

Example 2: H. Sapiens PTP4A2



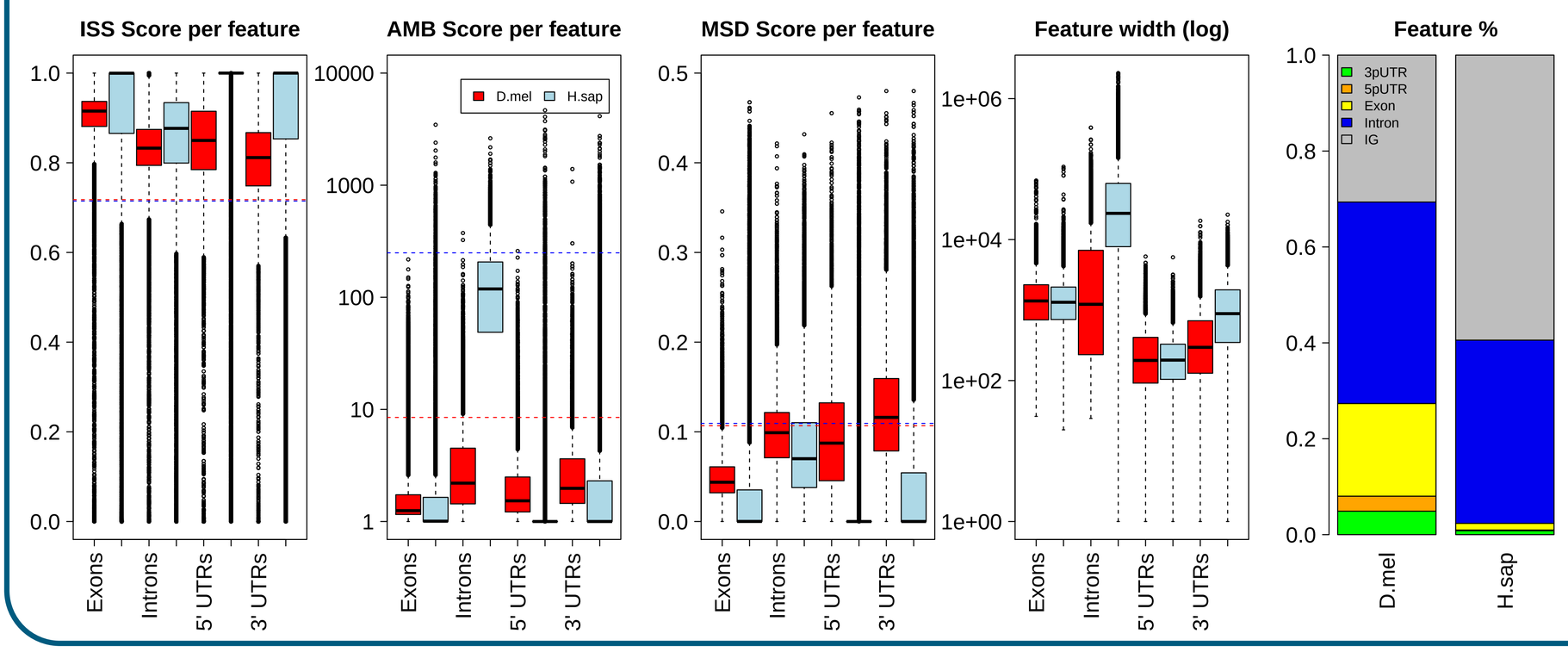
This example shows the 3′ end of PTP4A2, a protein tyrosine phosphatase. The high abundance of repetitive elements (such as Alu-elements) in human intronic and intergenic regions are responsible for the high overall average AMB score of these features (see box below). ISS, AMB and MSD reveal that the rightmost exon of this gene seems to be duplicated in the genome (which was confirmed by a BLAST hits on chr11 and chr17).

Example 3: D. melanogaster satellite DNA



This example shows a SAR_DM satellite DNA element on the X chromosome of D. melanogaster. The ISS confirms that the sequence is not unique, the AMB and k-mer tracks reveal the internal segmentation of this genomic feature, the contextualized AMB (grey tracks on bottom) shows the locality of the effect and reveals the number of segments (see the track scaling).

Score distributions



Summary

ARGOS is a pipeline for the efficient (hg19: 6 days, d.mel: 1 hour) and exact (full Smith-Waterman alignments) calculation of genome-wide scores that characterize intragenomic sequence similarity patterns. Some research questions that can be addressed with ARGOS:

- What parts of my genome/gene/exon are unique?
- How many similar regions are there?
- What is the basic unit of a repeat?
- Are there local duplication events?
- Are there enriched short (exact) sequence motifs/k-mers?

This makes ARGOS a highly valuable tool for the annotation and analysis of known and newly assembled genomic sequences.