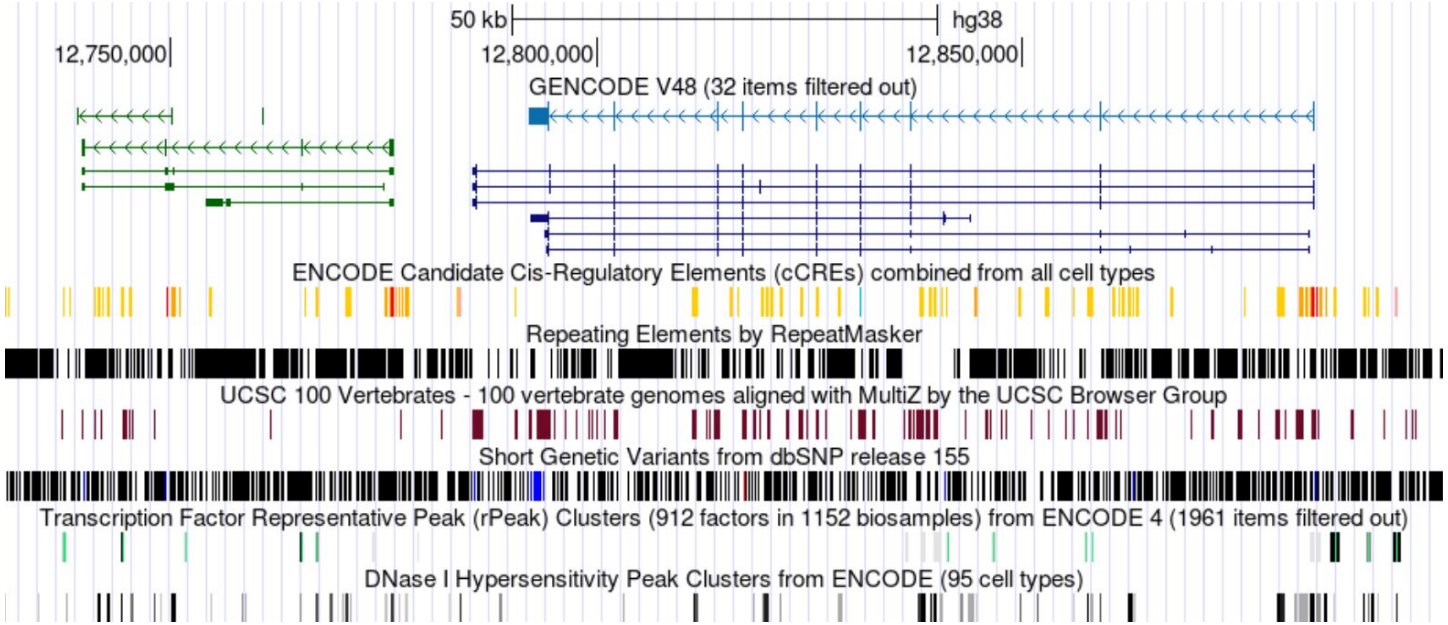
Efficient Analysis of Genome Annotation Colocalization

Askar Gafurov[1,3], Tomáš Vinař[1], Paul Medvedev[2], Broňa Brejová[1]

[1] Comenius University in Bratislava
[2] The Pennsylvania State University
[3] LIRMM, Montpellier

Genome annotations

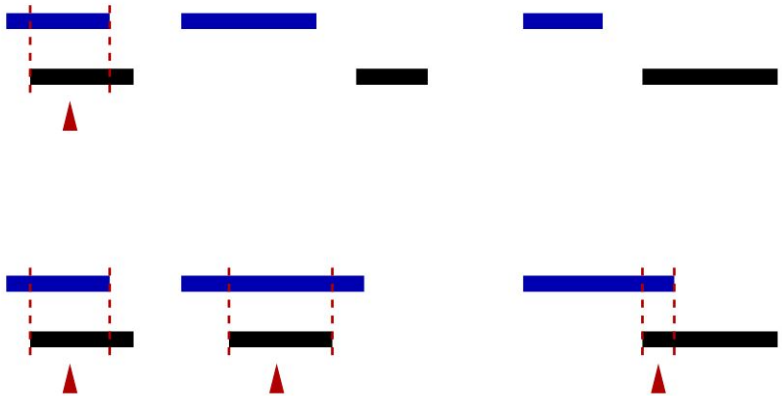
Annotation = set of disjoint intervals on a chromosome



genome.ucsc.edu, hg38, chr18:12,730,000-12,900,000

Comparing two annotations

- If two annotations overlap often, they are possibly related
- **Example:** H3K4me3 histone modifications coincide with promoters, because H3K4me3 has a role in regulation of transcription initiation
- **Goal:** Is intersection of two annotations surprisingly large?
  In other words: Compute p-value, statistical significance



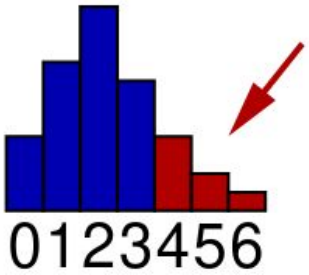
Ingredients of a statistical test for colocalization

- **Statistic:** # of overlapping intervals, # of shared bases, ...
- **Null hypothesis:** random process for "scrambling" one annotation
- **P-value computation:** Pr[random ≥ observed]

Low P-value suggests possibly related annotations

observed = 4

Pr[random ≥ observed]:



Many existing approaches to colocalization

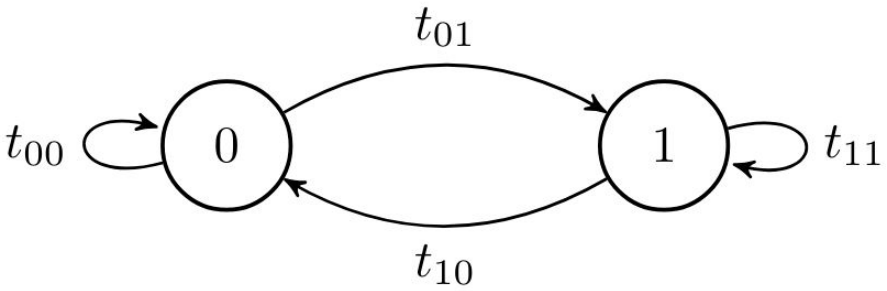
- **Statistic:**
  # of overlapping intervals (e.g. Layer et al. 2013)
  # of shared bases (e.g. Zarrei et al. 2015)
- **Null hypothesis:**
  independent positions (e.g. Dozmorov et al. 2016)
  permutational (e.g. Coarfa et al. 2014)
- **P-value computation:**
  exact (e.g. McLean et al. 2010)
  sampling (e.g. Yu, Wang, and He 2015)

Review Kanduri et al. 2019

Our approach: Markov chain null hypothesis

Blue annotation fixed, black generated by a Markov chain

Two states: 0 (outside), 1 (inside)



State sequence 0,0,**1,1,1**,0,**1,1**,0,0 represents annotation {[2, 4], [6, 7]}

Transition probabilities estimated from the black annotation

Markov chain null hypothesis

Faithfulness between independent positions and permutations

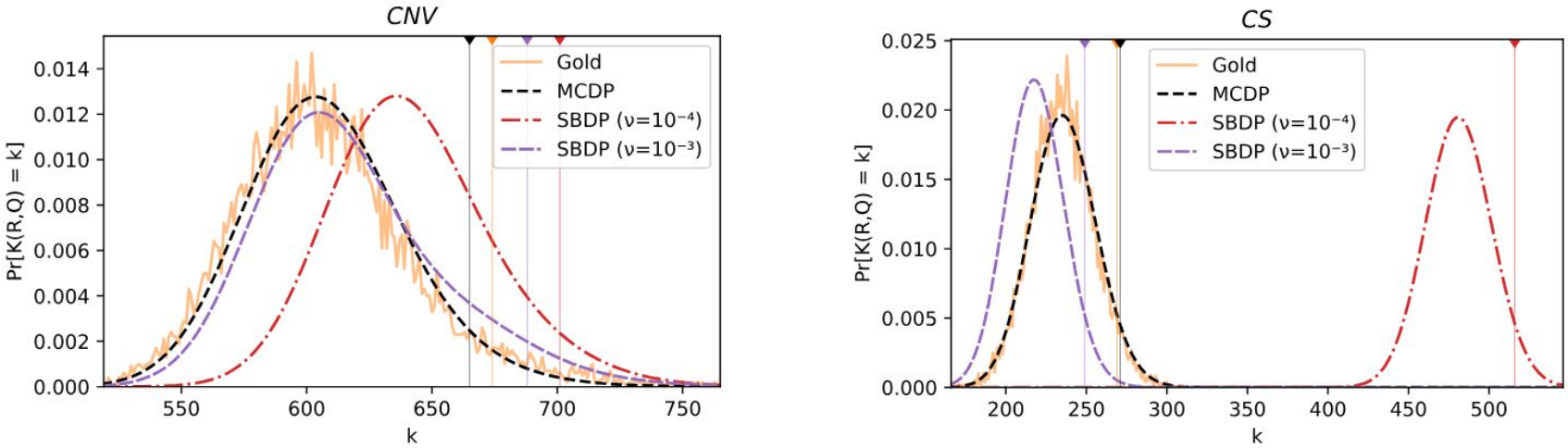
- Captures mean lengths of intervals and gaps between them

Allows fast computation of P-values:

- **Full probability distribution**, exact p-value in $O(n^2)$
  Dynamic prog., O(1) computation of transition matrix powers
- **Exact μ and σ**, normal approx. of p-values in O(n) or O(n log n)
  Conditional variance in intervals combined by the law of total variance

Implemented in MCDP2 tool https://github.com/fmfi-compbio/mcdp2

Comparison of model accuracy



Comparison of Markov chains (MCDP) to permuting intervals (gold) and genome scaling algorithm by Sarmashghi and Bafna 2019 (SBDB)

Two real datasets

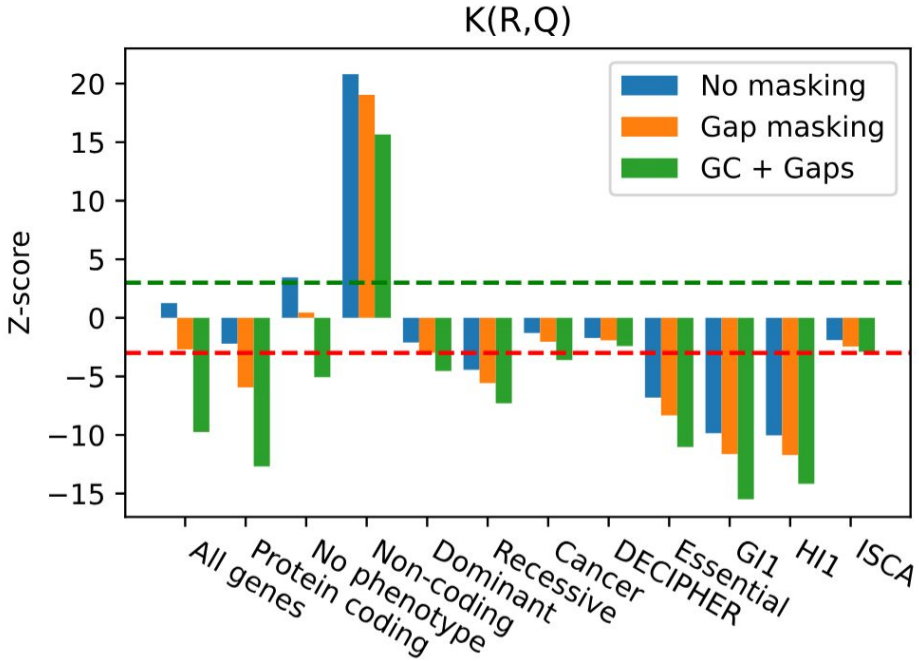
Impact of context-based Markov chains

Data Zarrei et al. 2015

Enrichment / depletion of CNV losses wrt. exons of different types

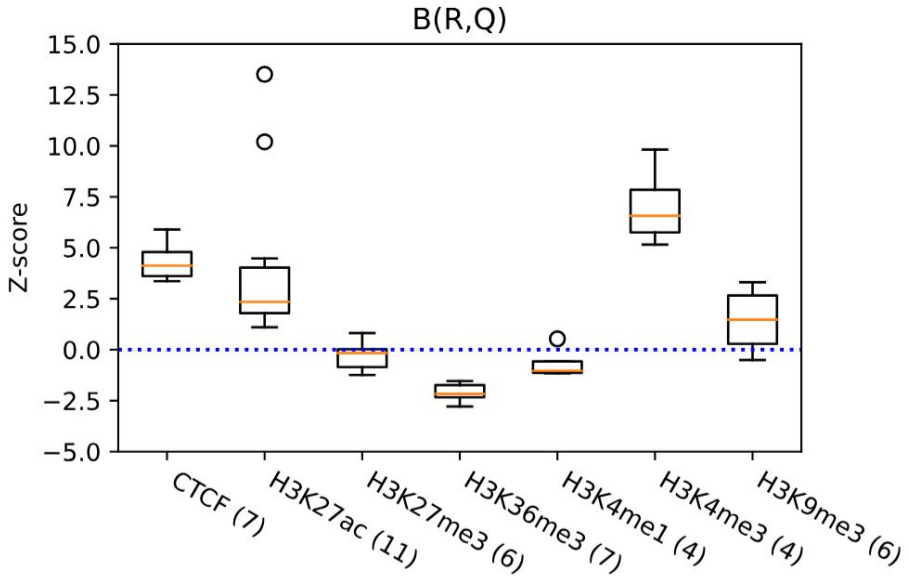
Separate Markov chains for different contexts:
assembly gaps, 4 GC content levels

All genes slightly enriched without contexts, but significantly depleted with them



T2T human genome analysis

- Gershman et al. 2022 studied TARs, telomere-associated repeats
- Enrichment of epigenetic marks in non-telomeric compared to sub-telomeric TARs
- Significance of H3K27ac and H3K4me3 confirmed, CTCF also enriched



Summary

- MCDP2 uses **Markov chains** to capture annotation properties in multiple genomic **contexts**
- It compares two annotations efficiently, considering the number of overlapping intervals and the number of bases
- It computes Z-scores and **p-values of enrichment or depletion**
- **Future extensions:** fast computation in genomic windows, multi-state Markov chains for length modeling

Publications, tool and funding

- Markov chains improve the significance computation of overlapping genome annotations, ISMB 2022
- Fast Context-Aware Analysis of Genome Annotation Colocalization, RECOMB 2024, JCB 2024
- **Use our MCDP2 tool** https://github.com/fmfi-compbio/mcdp2

Funding APVV-22-0144, VEGA 1/0538/22 and 1/0140/25, PANGAIA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 872539

Thank you for your attention