Genomic Computing
Politecnico di Milano

POLITECNICO DI MILANO

Dipartimento
di Elettronica, Informazione e
Bioingegneria





Discovering similar (epi)genomics feature patterns in multiple genome browser tracks

Piero Montanari [1], Arnaud Ceol [2], Ilaria Bartolini [1], Paolo Ciaccia [1], Marco Patella [1], Stefano Ceri [3], Marco Masseroli [3]

[1] DISI - Università di Bologna, [2] IIT@SEMM – IIT,
[3] DEIB - Politecnico di Milano

Discovering similar patterns in genomics tracks
Background

- **Next Generation Sequencing** (NGS) is opening many interesting practical and theoretical computational problems
- **Genome browsers** (UCSC Genome Browser, Integrated Genome Browser (IGB)) allow:
  - Visual inspection and identification of interesting patterns on multiple genome browser tracks
- **Pattern**: a set of (epi)genomic regions / peaks at given distances from each other in different tracks

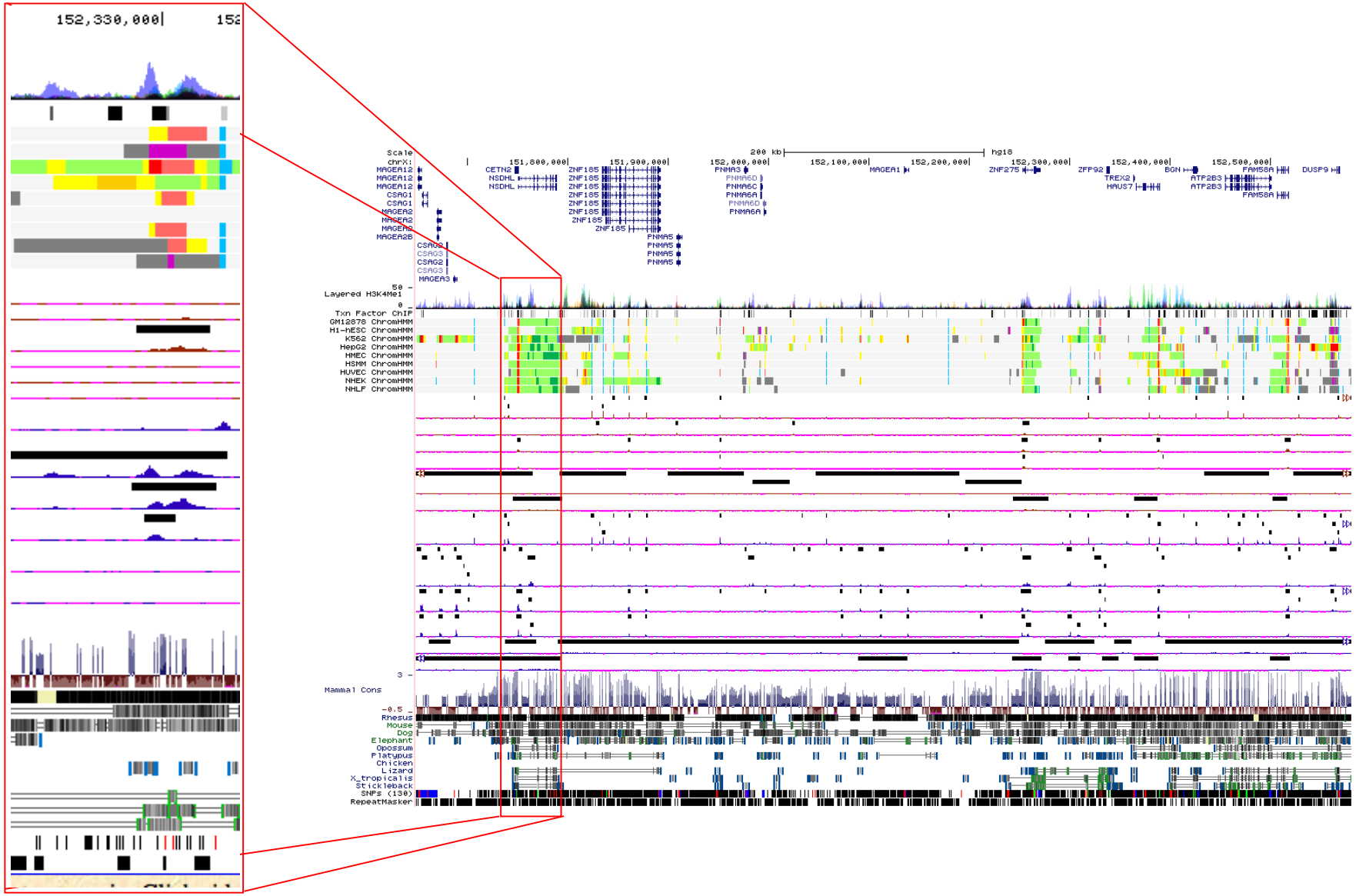
  - e.g.: gene regulatory DNA areas, which include heterogeneous (epi)genomic features (different histone modifications, TFBSs, , …)

Discovering similar patterns in genomics tracks

Pattern-based queries from genome browser

POLITECNICO DI MILANO

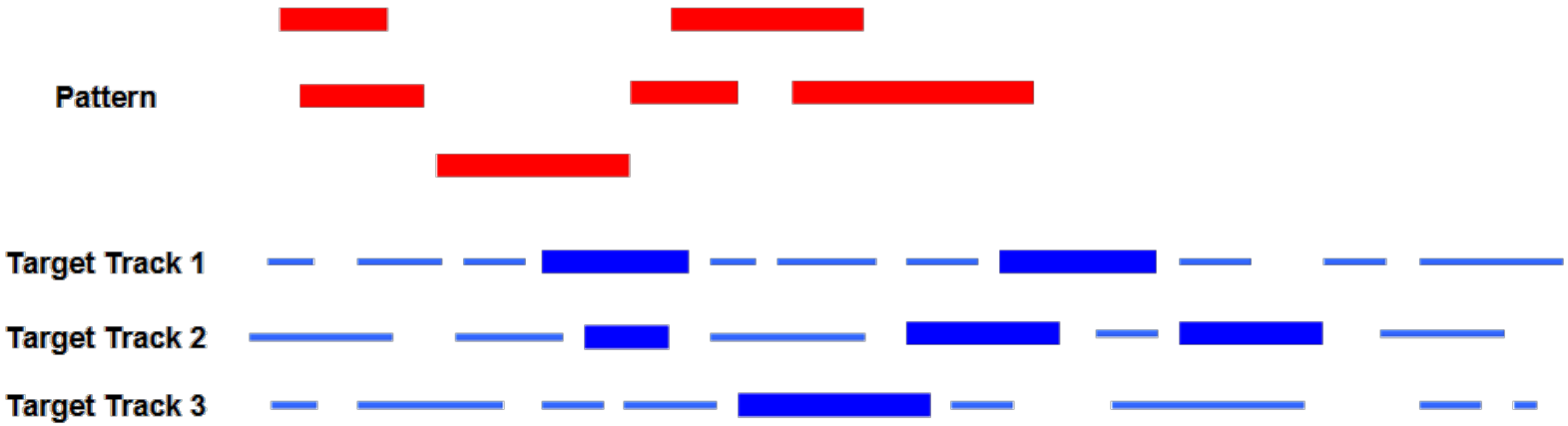
A pattern of genomic features



Discovering similar patterns in genomics tracks
Motivation

Once such patterns are visually identified in a genome section:

- **Search for pattern occurrences in whole genome**:
  - Complex computational task
  - Currently not supported
- **Their discovery in whole genome** very important for:
  - Biological interpretation of NGS experimental results
  - Comprehension of biomolecular phenomena



Discovering similar patterns in genomics tracks
Reached goal

POLITECNICO DI MILANO

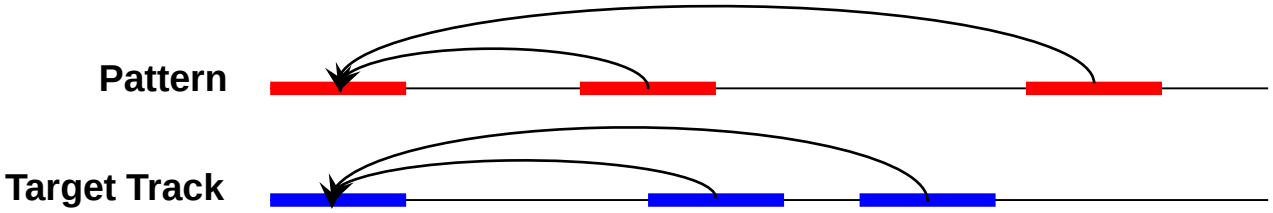
- We defined an **optimized pattern-search algorithm** to find **genomic region sets** that are **similar to a given pattern**
  - Efficiently
  - In large (epi)genomic / transcriptomic data sets
- We implemented the algorithm within an **IGB plugin**, named **SimSearch**, which allows intuitive user interaction in both:
  - **Visual selection** of an interesting **pattern** on loaded IGB tracks
  - **Visualization** of occurrences of **similar patterns** identified in the whole genome

Discovering similar patterns in genomics tracks
Method

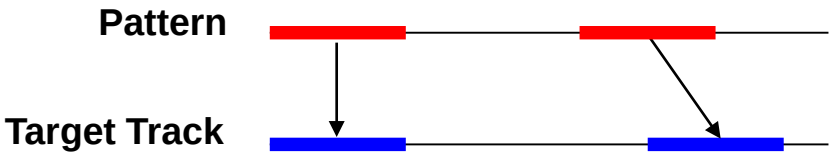
- **Best-Matching Problem** (BMP) is suspected to be NP-hard
- For the considered application:
  - **Aligned genomic data** -> strictly increasing region sequences
  - **M << N** (M and N: number of elements in a pattern and in the target tracks to be compared, respectively)
- **Our proposal**: a **Root-element approach** (R-BMP) and **Windowed Dynamic Programming algorithm** (WDP-BMP) to lower complexity:
  - Best matching for each element of the pattern in the target track through a **binary search**
  - Complexity only of O(N*log(N))
  - Applicable also to (very) large data problems

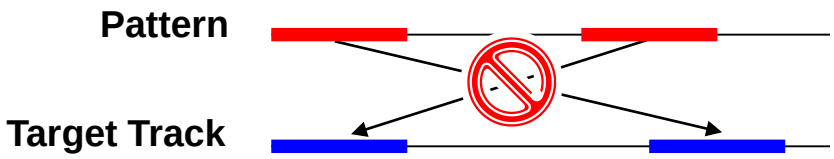
Discovering similar patterns in genomics tracks
Method – Base problem / model

- Pattern = single track
- Regions = points, identified by their linear genomic coordinate
  - Patterns and tracks = sequences of integers (region coordinates)
- Only relative distances between regions/points are relevant



- A **matching** is a strictly increasing function *f* that assigns to each pattern (query) element a (target) track element
  - Preserving element order





Discovering similar patterns in genomics tracks
Method – Base problem / model

- **Pattern matching** typically solved by a cost based approach, where lower cost implies higher similarity
- **Root-element approach** (R-BMP): The cost C of a **matching** *f* is the sum of squared distances relative to the first (root) matching pair

$$C_f(Q,T) = \sum_i \left(T_{f(i)} - Q_i - \left(T_{f(1)} - Q_1\right)\right)^2$$

  - Q and T: sequences of query (pattern) and target elements
  - $T_{f(1)} - Q_1$: root-distance (relative distance between root elements)

- **Goal**: find the matching with minimum cost

Example

- Pattern (query) Q = <1,7,10>
- Target track T = <3,5,9,11,13,14,18,21>
- A possible matching: {(1->3),(7->9),(10->21)}
  - Root-distance = 3-1 = 2
  - Cost = $(3-1-2)^2 + (9-7-2)^2 + (21-10-2)^2 = 81$
- Best matching {(1->5),(7->11),(10->14)}
  - Root-distance = 5-1 = 4
  - Cost = $(5-1-4)^2 + (11-7-4)^2 + (14-10-4)^2 = 0$

Discovering similar patterns in genomics tracks
Method – Base problem / model

Windowed Dynamic Programming algorithm (WDP-BMP)

- **Main result**: A matching is optimal if and only if all its partial matchings have minimum cost
- For all possible root positions in T:
  - They are |T|-|Q|+1
- For all regions Q_i in Q:
  - Find (with binary search) the best match for Q_i in T
  - Possibly, widen the "window" to avoid conflicts (regions in T that are best match for multiple regions in Q)
- For all possible matches:
  - Compute the cost using the partial cost obtained till then
  - Abandon the current solution if partial cost ≥ best cost till then
- Overall complexity: O(|T||Q|(log|T|+|Q|)
  - But **O(|T|log|T|)** if |Q|<<|T| (usually |Q|=1÷10, |T|~10^5÷10^6)

Discovering similar patterns in genomics tracks
Method – Extended model

- **Multi-track patterns**
  - Same approach is repeated for each pattern track
- **Negative matching tracks**
  - Pattern tracks with regions that should not appear in the solution
  - Removed from the solution search space (before search start)
- **Partial matching tracks**
  - Pattern tracks with regions that might be missing in the results
  - A cost is considered for those regions that remain unmatched
- **Region features**
  - The cost function includes the (dis-)similarity between features (attributes) of query and target track regions
- **Regions as intervals**
  - Region size modeled as a region feature
- **Top-K distinct matchings**
  - To consider diverse results, we require that the best K results have no matched region in common

Finding enhancer regions

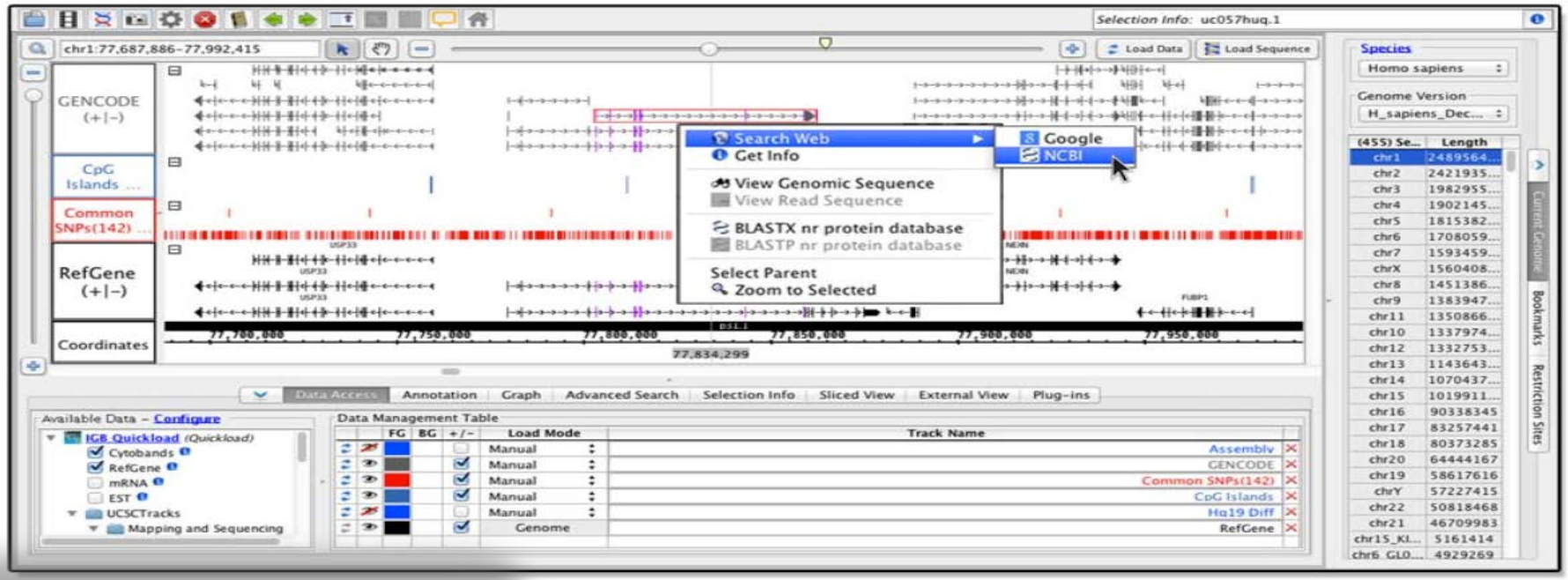
- Complex pattern
  - 2 single region positive tracks (H3K4me1 and H3K27ac)
  - 2 single region negative tracks (TSS and H3K4me3)
  - 4 single region partial tracks (CTCF, DHS, P300 and Pol2)
- Target data tracks about K562 cell line (AML) from ENCODE
- Top-100 results visually inspected by an expert
  - **100% precision** (of the resulting matchings to the pattern)
- All (1,651) results compared with "*enhancer*" chromatin state regions from ChromHMM tool (Broad Institute), data available in ENCODE
  - **85.46% precision** (but ChromHMM uses more histone marks)

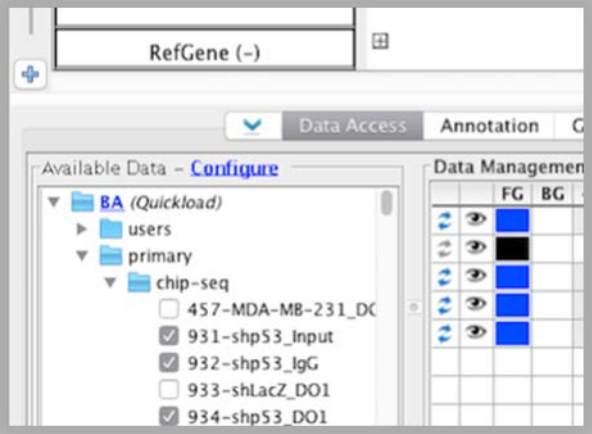
Discovering similar patterns in genomics tracks

SimSearch: a plugin for IGB

POLITECNICO DI MILANO

Available at: https://arnaudceol.github.io/simsearch/





Bed/BigBed, BigWig, GTF …

Galaxy

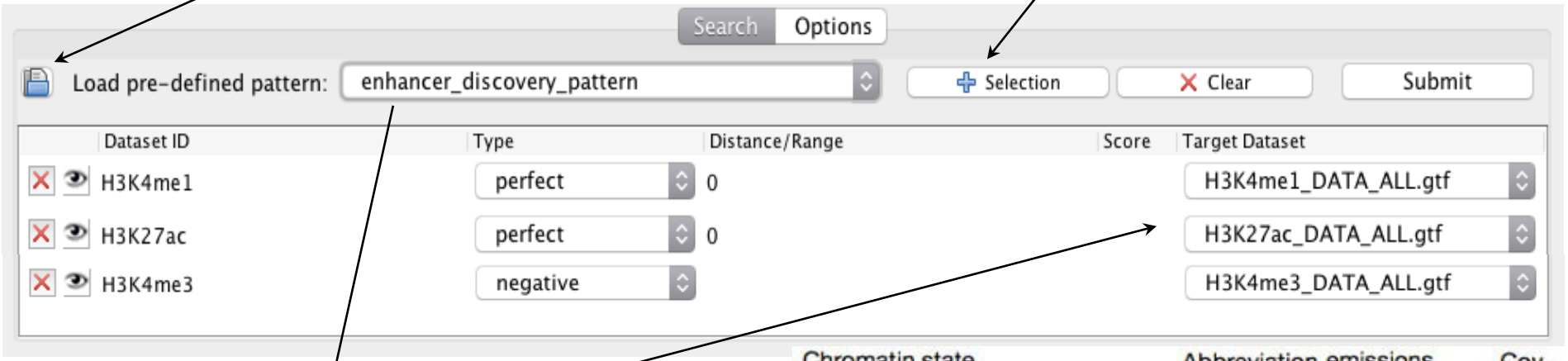
HTS-flow (*http://arnaudceol.github.io/htsflow/*)

The **Integrated Genome Browser** (*http://bioviz.org/*)

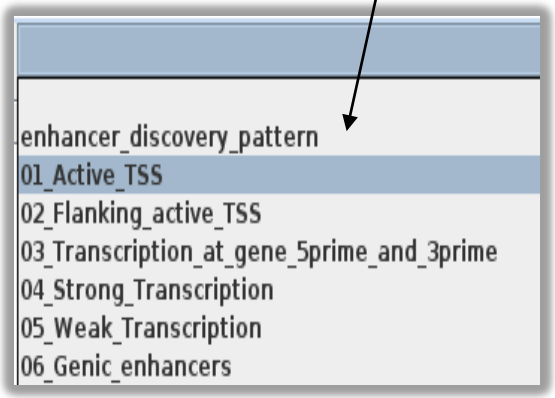
Discovering similar patterns in genomics tracks

SimSearch: define the pattern to search for

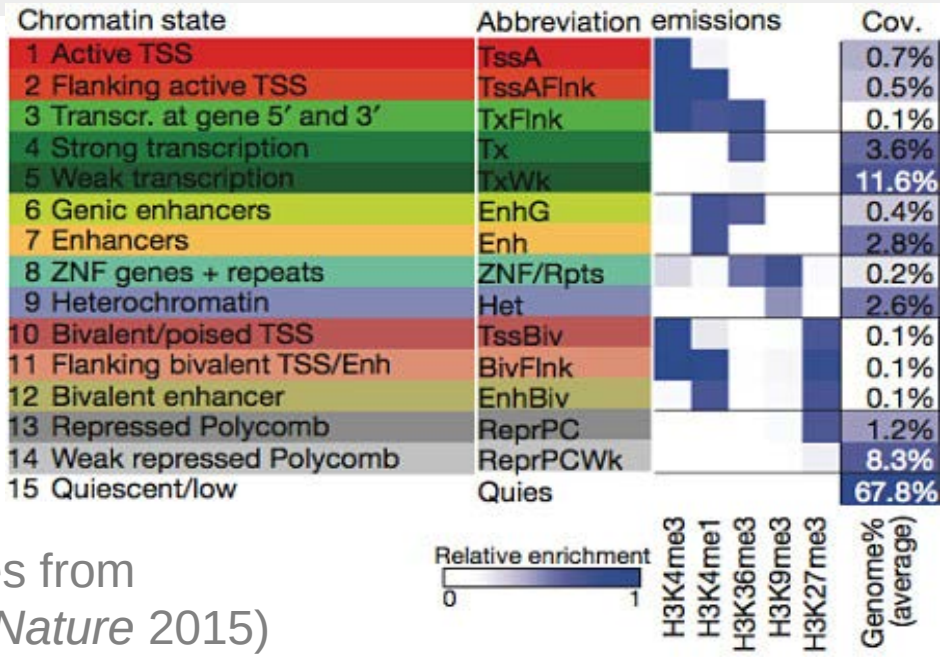
Load pattern from file **Create** pattern by selecting IGB tracks / regions



Where to search



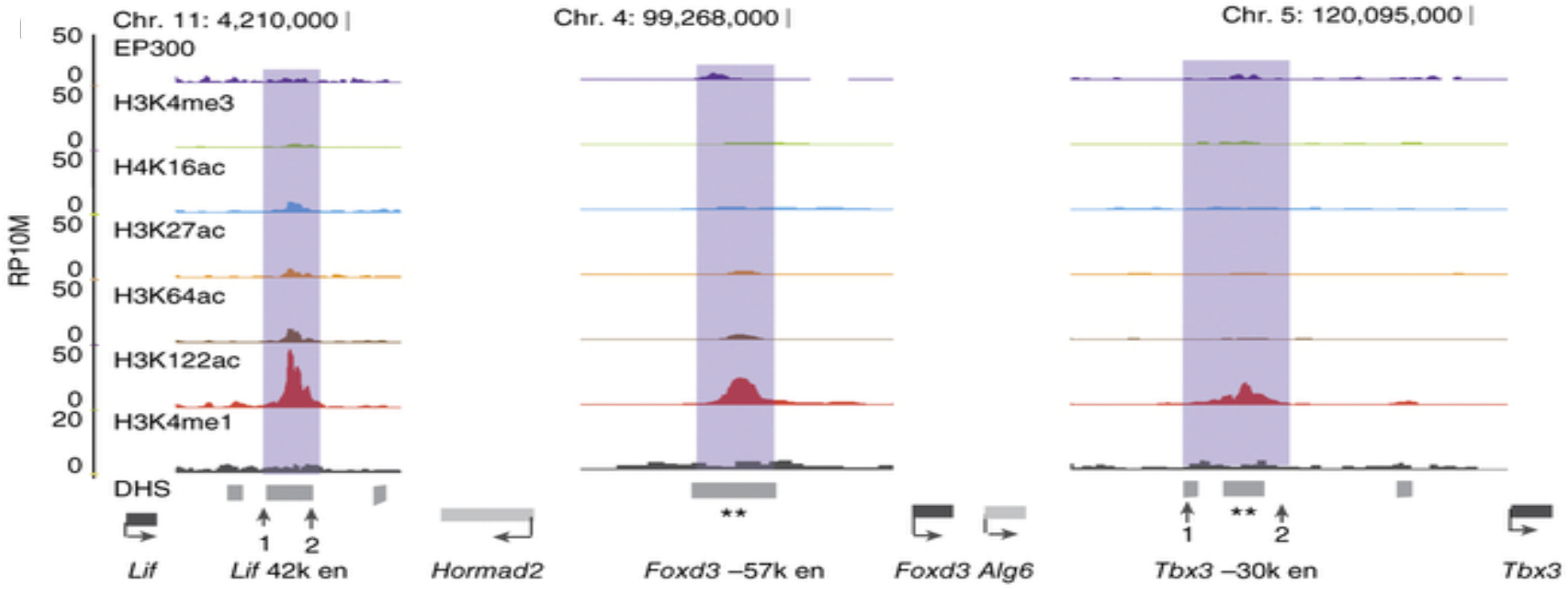
Load pre-defined pattern



(e.g. CHromHMM chromatin states from **Roadmap Epigenomics Consortium**, *Nature* 2015)

Discovering similar patterns in genomics tracks
SimSearch: searching for new enhancers

Pradeepa et al. (*Nature Genetics* 2016) identified a new class of **enhancer** (with **H3K122ac** and absence of **H3K27ac**) in mouse



Several examples validated in mouse

- Homozygous deletion of group 2 putative enhancer 42 kb **downstream** of ***Leukemia inhibitory factor*** (*Lif*) gene led to **reduced expression** of ***Lif***, but not of flanking ***Hormad2*** gene
- Deletion of one allele of putative enhancer 30 kb **upstream *Tbx3*** led to **downregulation** of ***Tbx3***

Discovering similar patterns in genomics tracks
SimSearch: searching for new enhancers

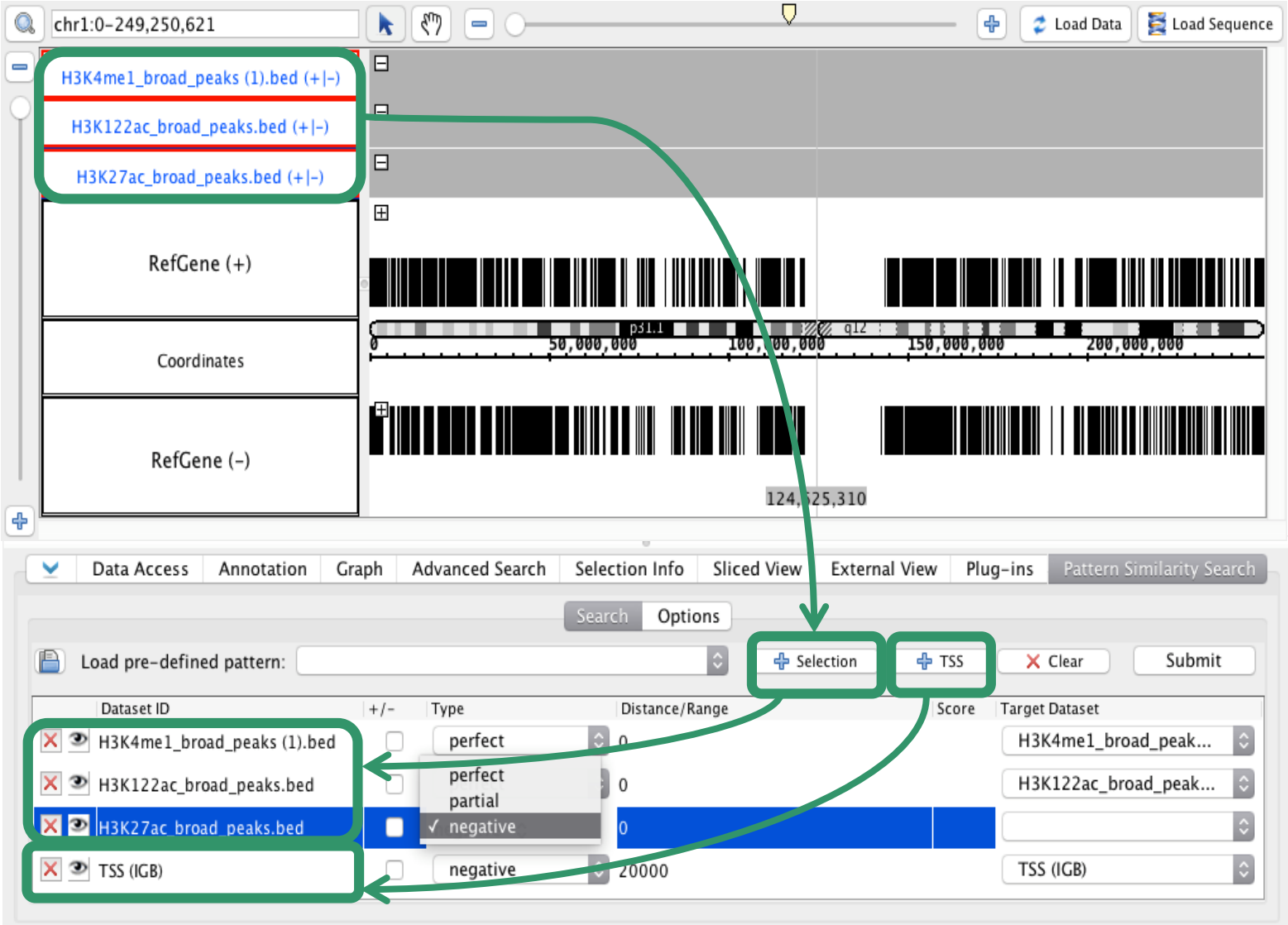
POLITECNICO DI MILANO

Look for **this new pattern** in human samples:

1. **Load** tracks of human data

2. **Add** tracks to the pattern

3. **Add** TSS (negative)

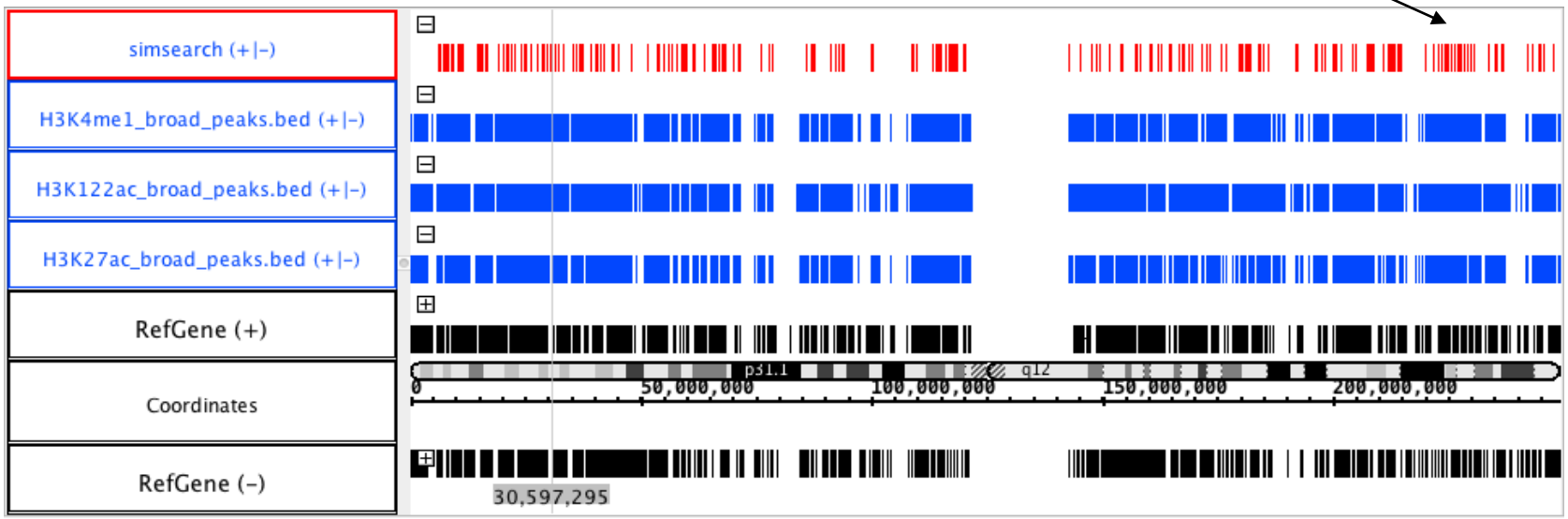


Discovering similar patterns in genomics tracks

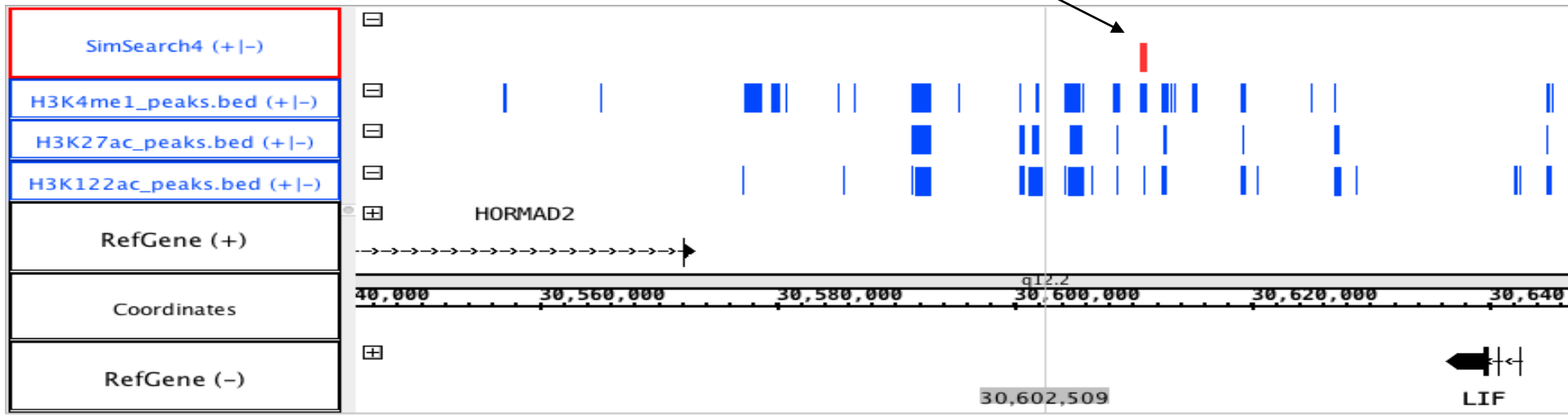
SimSearch: editing the pattern

POLITECNICO DI MILANO

In a new track (red), see all active regions (matches) found



Similar match found in human cells



Discovering similar patterns in genomics tracks
Conclusions

- Definition of a method for **finding patterns in genomic sequences** and of a dynamic programming algorithm for efficient solution
  - Experimental evaluation found it accurate and efficient
- Algorithm implemented as a plugin for Integrated Genome Browser
- Applicable to any (epi)genomic / transcriptomic region data, e.g.:
  - Histone Modifications (HM), Transcription Factor Binding Sites (TFBS), Single Nucleotide Polymorphisms (SNP), Differentially Expressed Genes (DEG), DNase I Hypersensitive Sites (DHS), Transcription Start Sites (TSS), or any other annotations
- Support for understanding biomolecular mechanisms
  - Response to treatments, pathology onset/development, …
- Thanks to the GenData 2020 PRIN project and GeCo ERC project!

Thank you for your attention!

Any question?