

DATA-DRIVEN GENOMIC COMPUTING: MAKING SENSE OF SIGNALS FROM THE GENOME

Stefano Ceri

DEIB | Dipartimento di Elettronica, Informazione e Bioingegneria



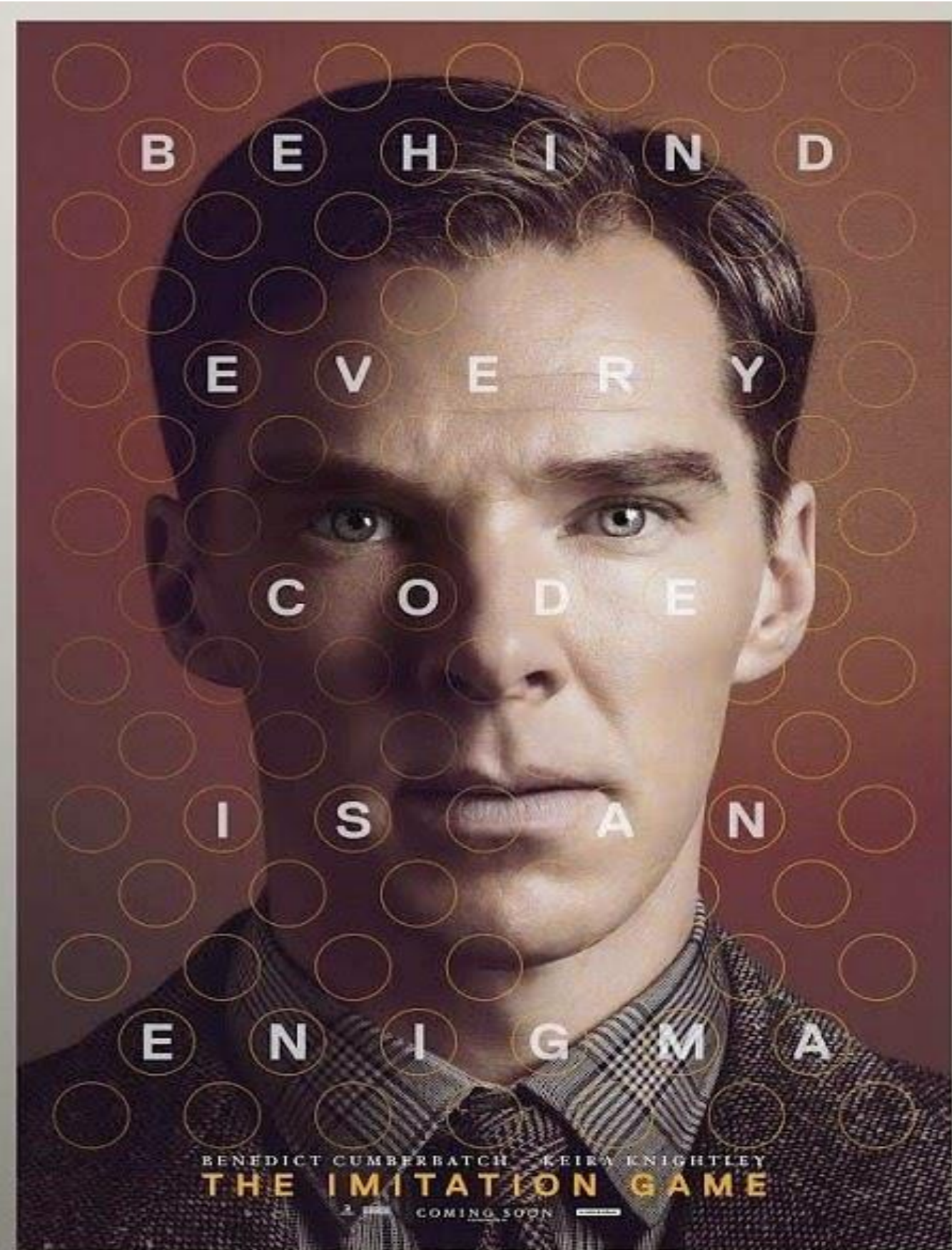
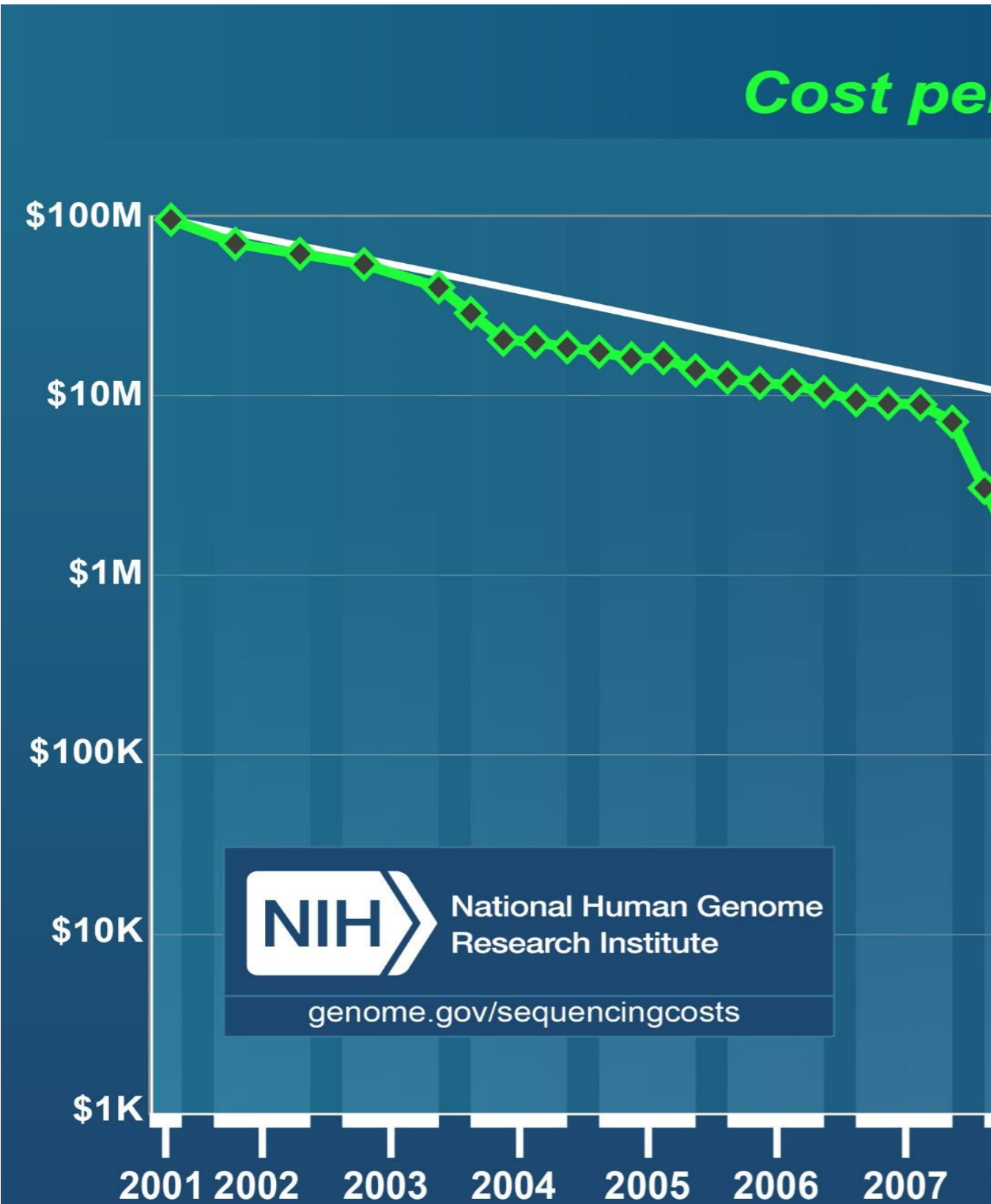
POLITECNICO
MILANO 1863

BACKGROUND

Human Genome Project



HIGH THROUGHPUT SEQUENCING COST PER GENOME, 2001-2015



Why Genomic Computing?

- Technological revolution for DNA Sequencing
- Availability of huge repositories of open data
- It is now possible to explain how DNA inheritance and replication cause/influence many diseases, leading to personalized medicine
- Many biological and clinical problems need data exploration, retrieval and analysis
- Genomic datasets are «big data»

Rough Terms and Sizes

Abstract Data

- The human DNA sequence is a string of 3.2 billions of base pairs, encoding adenine (A), cytosine (C), guanine (G), and thymine (T); size = 800Mbyte.

Raw/Aligned Data

- Data is produced as «reads», overlapping subportions of the genome, and then aligned to a reference genome, with emphasis on quality; size = 200GByte.

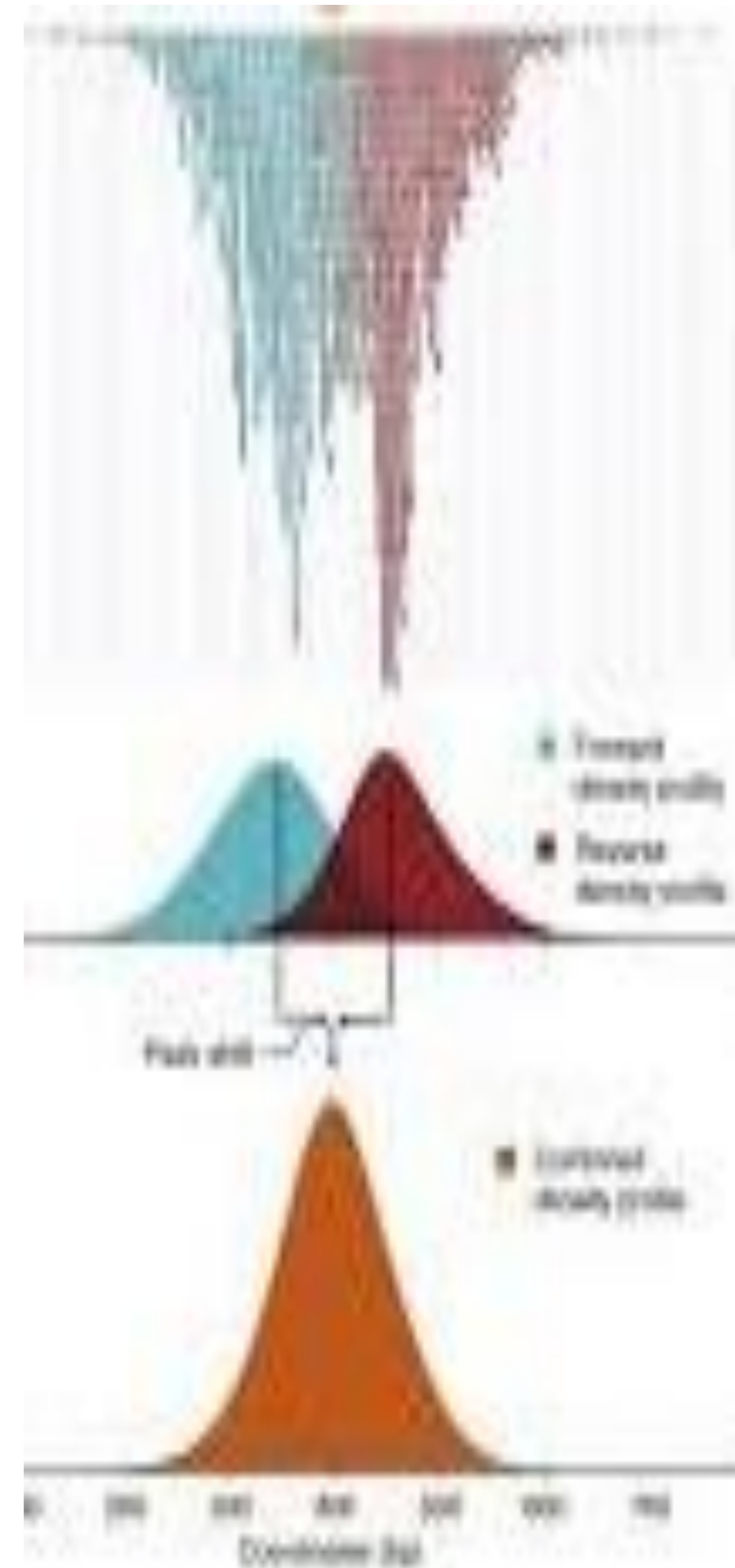
Processed Data:

- But each of us has «just» 4.1M to 5M mutations, mostly single substitutions/insertions/deletions; size = 125Mbyte

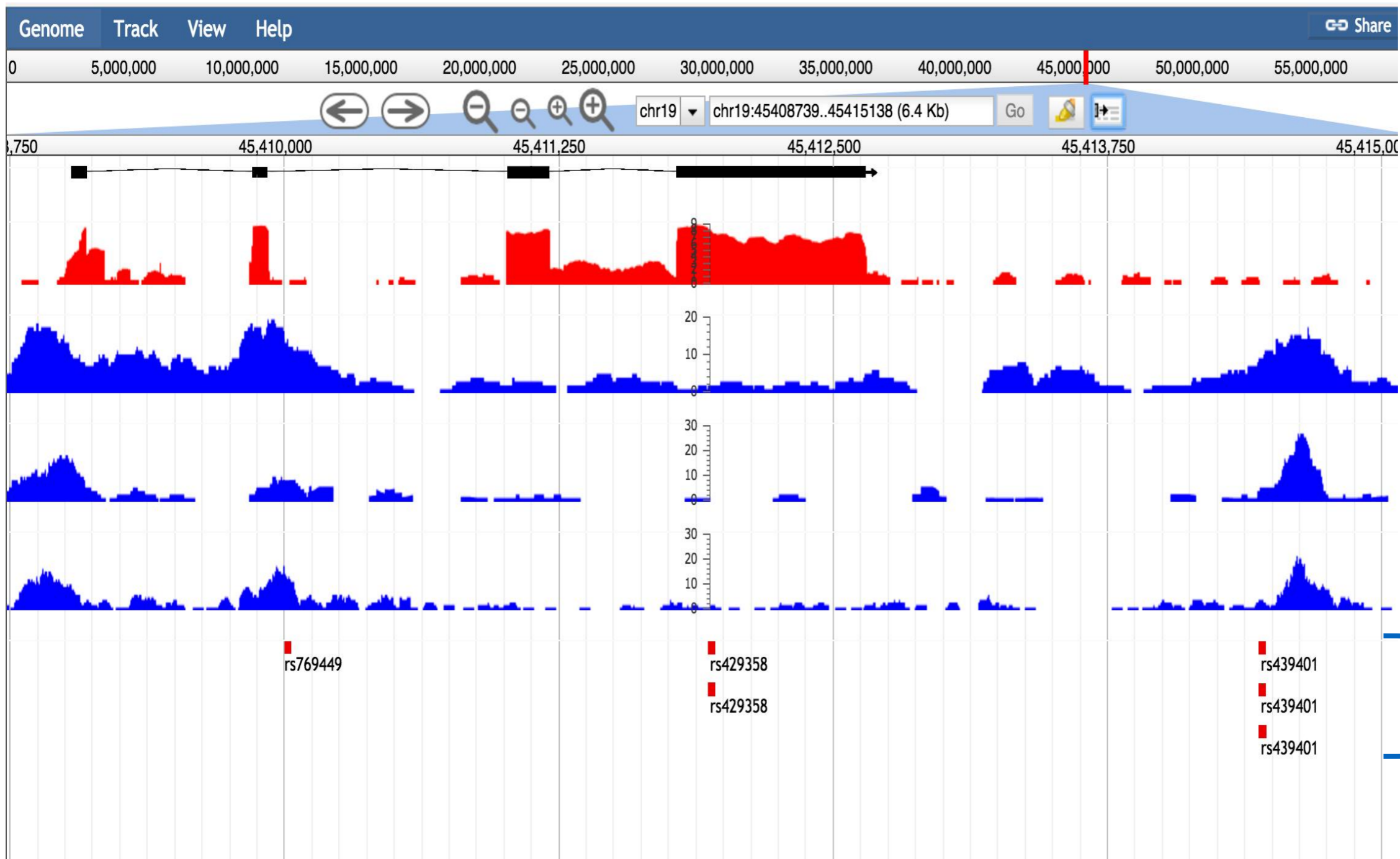
(Epi)Genomic Signals: Mutations - Expression - Peaks



	DNA	
<i>codone 5</i>	6	7
... CCT	GAG	GAG ...
... CCT	GTG	GAG ...



Signals on the Genome Browser



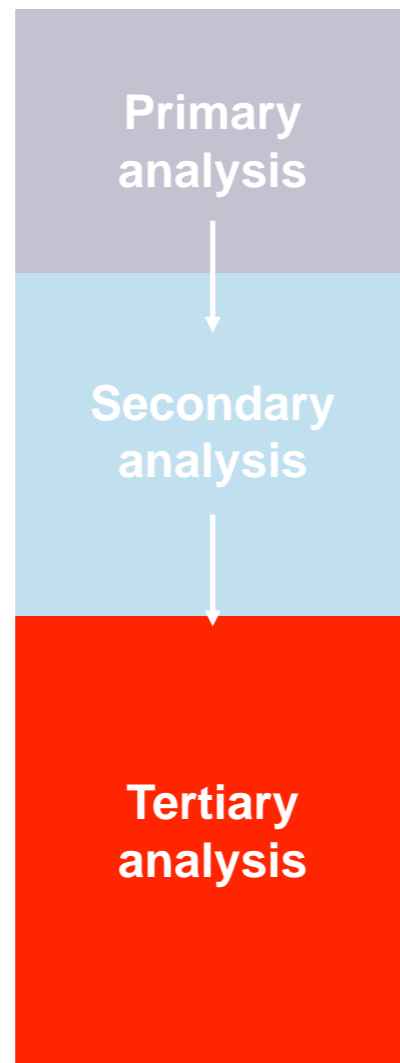
Genes

Expression
(RNA-seq)

Peaks
(ChIP-seq)

Mutations

BIG DATA ANALYSIS WITH NEXT GENERATION SEQUENCING



- Analysis of hardware generated data, machine stats, etc.
- Production of sequence reads and quality scores
- QA filtering on raw reads
- Alignment/Assembly of reads
- QA and variant calling on aligned reads
- QA/QC of variant calls
- Annotation and filtering of variants
- Data aggregation and multi-sample processing
- Association analysis
- Population structure analysis
- Genome browser driven exploratory analysis

Source: <http://blog.goldenhelix.com/grudy/a-hitchhiker%E2%80%99s-guide-to-next-generation-sequencing-part-2/>

A VIEW OF BIG DATA ANALYSIS PLAYERS



PRIMARY ANALYSIS



read counting
MOTIF finding
quality control MEME SNP BOWTIE
alignment peak calling GATK
ADAM variant calling
FASTA indel detection
SICER allele calling ^{BWA}
MACS HMMER

WordItOut

SECONDARY ANALYSIS



- **FireCloud (Broad Inst.)**
- **Paradigm4 (Spinoff)**
- **GMQL/Geco (PoliMi)**
- **DeepBlue (Blueprint)**

TERTIARY ANALYSIS

Which problems can we solve?

- Which cancer types can be explained by dysregulation of the 3-dimensional structure of the genome?
- Which co-occurring (killer) mutations cause the death of a cell in given tumors?
- Which transcription factors (dimers) always occur together?
- How can we assign predominant functions to each portion of the genome?

A Short Story of Genomic Computing in my Group at Politecnico (DEIB)

September 2012: Meeting at IEO-IIT; start of collaboration with prof. Pier Giuseppe Pelicci et. Al.

March 2013 – current: Big group with:

- **Scientists:** Daniele Braga, Alessandro Campi, Marco Masseroli, Matteo Matteucci, Giampaolo Cugola, Heiko Muller.
- **PhD students:** Anna Bernasconi, Vahid Jalili, Fernando Palluzzi, Stefano Perna, Eirini Stamoulakatou, Yuryi Vaskin, Francesco Venco.
- **Master students:** Michele Bertoni, Ilaria Buonagurio, Simone Cattani, Andrea Gulino, Luca Nanni, Ilaria Raciti.
- **Post-docs:** Arif Canakoglu, Abdulrahaman Kaitoua, Pietro Pinoli.

March 2013 – Feb. 2016: PRIN Project Gendata 2020

(with: Math@PoliMi, Sapienza, Roma3, Unibo, PoliTo, UniBg, StataleMi, UniSal, UniCal)

January 2015: first release of GMQL V1 (at Polimi and IEO-IIT)

March 2015: first accepted paper on Bioinformatics

April 2016: GMQL V2 installed at CINECA

September 2016: kick-off Advanced ERC Grant «GECO»

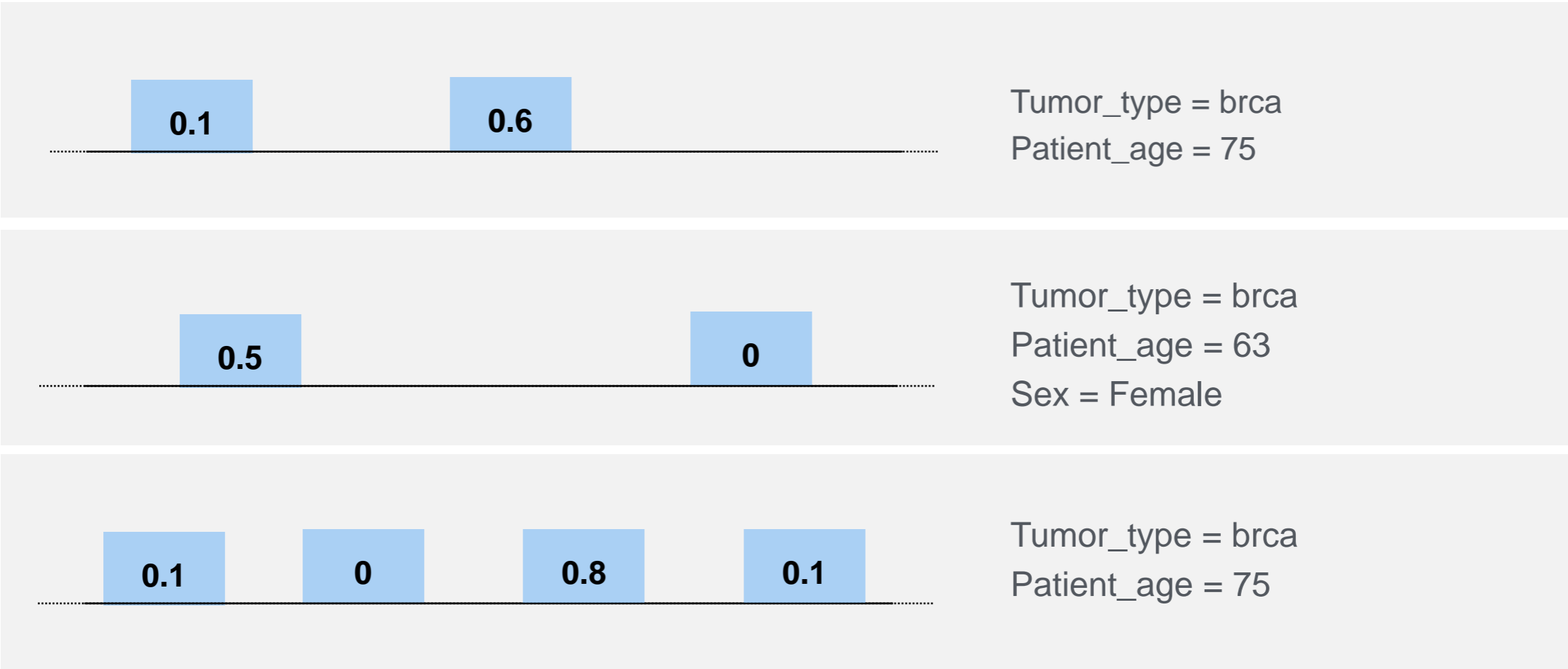
«Data-driven Genomic Computing»

DATA MODEL

GENOMIC DATA MODEL

REGIONS

METADATA



MODEL RATIONALE

REGIONS

Regions of the model describe processed data, e.g. **mutations**, **expression** or **regulation**; they have a schema, with 5 common attributes (ID, CHR, LEFT, RIGHT, STRAND) and then arbitrary typed attributes.

They achieve interoperability across a plethora of data formats

SAMPLE AND DATASET

Every sample corresponds to an «experiment», with an ID.

Every dataset is a named collection of samples with the same schema.

METADATA

Meta-data are arbitrary attribute-value pairs, independent from any standardization attempt.

They trace the **data provenance**, including **biological** and **clinical** aspects

Example of schemas and instances

Mutations (DNA-seq)

```
(id, (chr,start,stop,strand),  
(A,G,C,T,del,ins,inserted,ambig,Max,Error,A2T,A2C,A2G,C2A,C2G,C2T))  
(1, (chr1, 917179, 917180,*), (0,0,0,0,1,0,'!',',',0,0,0,0,0,0,0,0))  
(1, (chr1, 917179, 917179,*), (0,0,0,0,0,1,G,'!',0,0,0,0,0,0,0,0))
```

Expression (RNA-seq)

```
(id, ((chr,start,stop,strand), (source,type,score,frame,geneID,transcriptID,RPKM1,RPKM2,iDR))  
(1, (chr8, 101960824, 101964847,-), ('GencodeV10', 'transcript', 0.026615, NULL, 'ENSG00000164924.11',  
'ENST00000418997.1', 0.209968, 0.193078, 0.058))
```

Example of schemas and instances

Annotations

(id, (chr,start,stop,strand), (proteinID,alignID,type))

(1, (chr1, 11873, 11873, +), ('uc001aaa.3', 'uc001aaa.3', 'cds'))

(1, (chr1, 11873, 12227, +), ('uc001aaa.3', 'uc001aaa.3', 'exon'))

(1, (chr1, 12612, 12721, +), ('uc001aaa.3', 'uc001aaa.3', 'exon'))

(1, (chr1, 13220, 14409, +), ('uc001aaa.3', 'uc001aaa.3', 'exon'))

ChIA-PET

(denoting 3D genomic loops, head is assembled with coordinates, tail is in the schema)

(id,(chr,headstart,headstop,strand), (loopType, tailChr, tailStart, tailStop, PETcount, pValue, qValue))

(1, (chr1,7385626,7389841,*), ('Inter-Chromosome', chr17, 3081653, 3084755, 50, 0.0, 0.0))

QUERY LANGUAGE

```
32 <?php endwhile; wp_reset_query(); ?>
33 <div class="cleaner"></div>
34 <p><a class="button" href="/archiv/page/2">Zobrazit další příspěvky &raq
35 </div><!-- /content -->
36 <?php get_sidebar(); ?>
37 <div class="cleaner"></div>
38 <h2 class="homepage_cat_feed"><span class="white"><?php echo get_cat_name(3
39 <?php $loop = new WP_Query('posts_per_page=6&cat=312');
40 while ( $loop->have_posts() ) : $loop->the_post();
41 require ('part-homepage-cat-feed.php');
42 endwhile;
43 wp_reset_query(); ?>
44 <div class="cleaner"></div>
45 <h2 class="homepage_cat_feed"><span class="white"><?php echo get_cat_name(122)
46 <?php $loop = new WP_Query('posts_per_page=6&cat=122');
47 while ( $loop->have_posts() ) : $loop->the_post();
48 require ('part-homepage-cat-feed.php');
49 endwhile;
50 wp_reset_query(); ?>
51 <div class="cleaner"></div>
52 <h2 class="homepage_cat_feed"><span class="white"><?php echo get_cat_name(6);?</
53 <?php $loop = new WP_Query('posts_per_page=6&cat=6');
54 while ( $loop->have_posts() ) : $loop->the_post();
55 require ('part-homepage-cat-feed.php');
56 endwhile;
57 wp_reset_query(); ?>
58 <div class="cleaner"></div>
59 <h2 class="homepage_cat_feed"><span class="white"><?php echo get_cat_name(6);?</
60 <?php $loop = new WP_Query('posts_per_page=6&cat=6');
61 while ( $loop->have_posts() ) : $loop->the_post();
62 require ('part-homepage-cat-feed.php');
endwhile;
wp_reset_query(); ?>
```

Line 1, Column 1

QUERY LANGUAGE

QUERY LANGUAGE

SEQUENCE OF ALGEBRAIC OPERATIONS

High-level, declarative operations which operate both on regions and meta-data

→ each operation progressively builds the regions and meta-data of its result

Inspired by Pig Latin and targeted towards cloud computing

CLASSIC RELATIONAL OPERATIONS

SELECT	UNION
PROJECT	DIFFERENCE
GROUP	MERGE
ORDER/TOP	

DOMAIN-SPECIFIC GENOMIC OPERATIONS

COVER
GENOMETRIC JOIN
MAP

UTILITIES

LOAD, MATERIALIZE

QUERY LANGUAGE

OVERVIEW

```
PROMS = SELECT(annotationType == 'promoter') ANNOTATIONS;  
PEAKS = SELECT(dataType == 'ChipSeq') ENCODE;  
RESULT = MAP(peak_count AS COUNT) PROMS PEAKS;
```

Executed over 2,423 ENCODE samples including a total of 83,899,526 peaks mapped to 131,780 promoters producing as result 29 GB of data

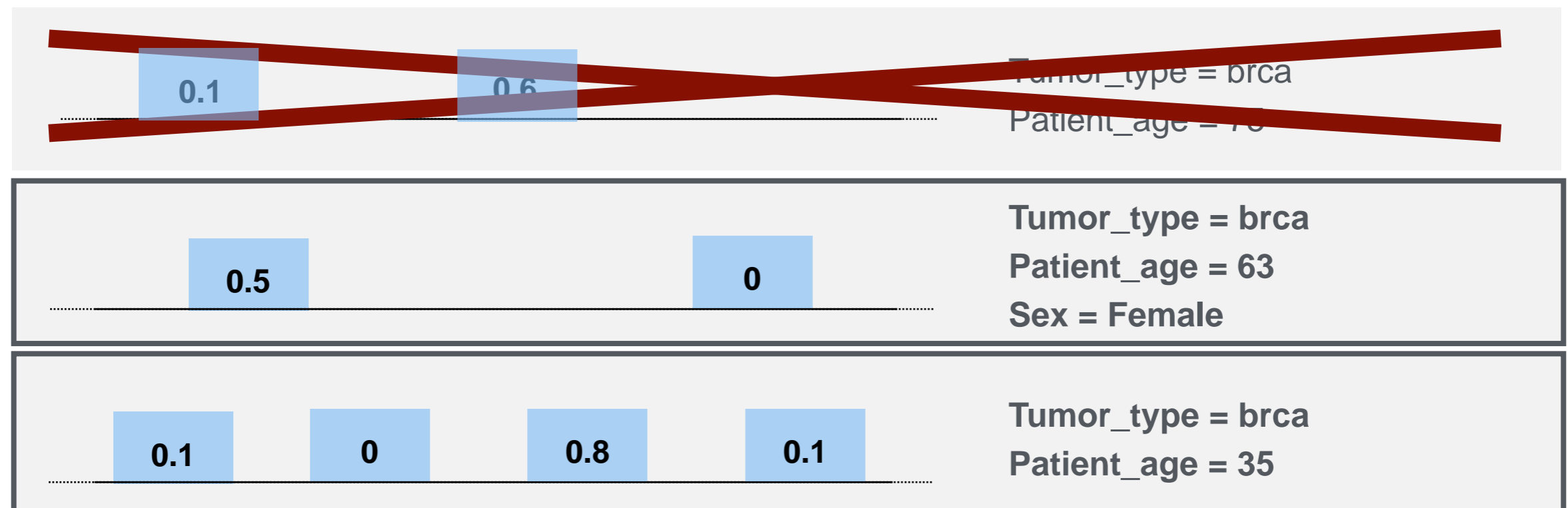
ID	ATTRIBUTE	VALUE
131	order	1
131	antibody	RBBP5
131	cell	H1-hESC
131	count	32028
133	order	2
133	antibody	SIRT6
133	cell	H1-hESC
133	count	30945
113	order	3
113	antibody	H2AFZ
113	cell	H1-hESC
113	count	30825

# Samples	# Regions	Join(dist <0)	Map(COUNT)	Cover
10	~1.9 M	14.66 sec.	20.29 sec.	19.25 sec.
50	~8.8 M	23.86 sec.	43.08 sec	46.34 sec.
100	~17.4 M	35.38 sec	74.43 sec.	79.02 sec.
1000	~60 M	120.98 sec	473.39 sec	235.22 sec.

METADATA SELECTION

Selection of the samples

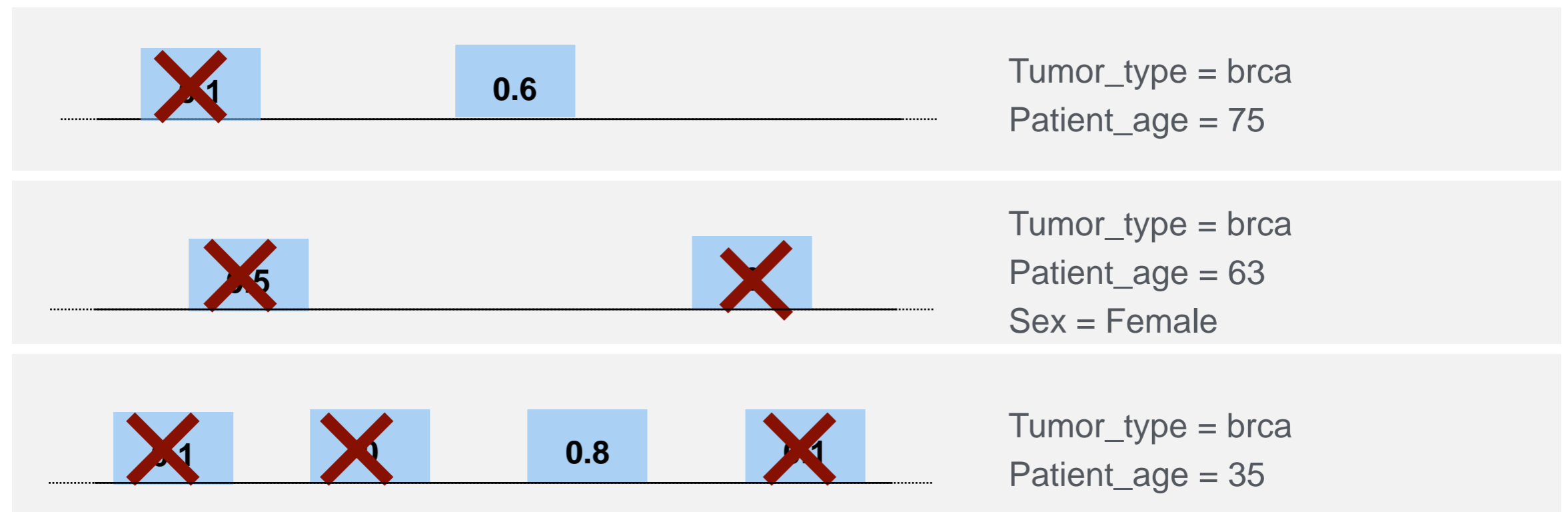
e.g. select patients younger than 70 years



REGION SELECTION

Selection of the regions

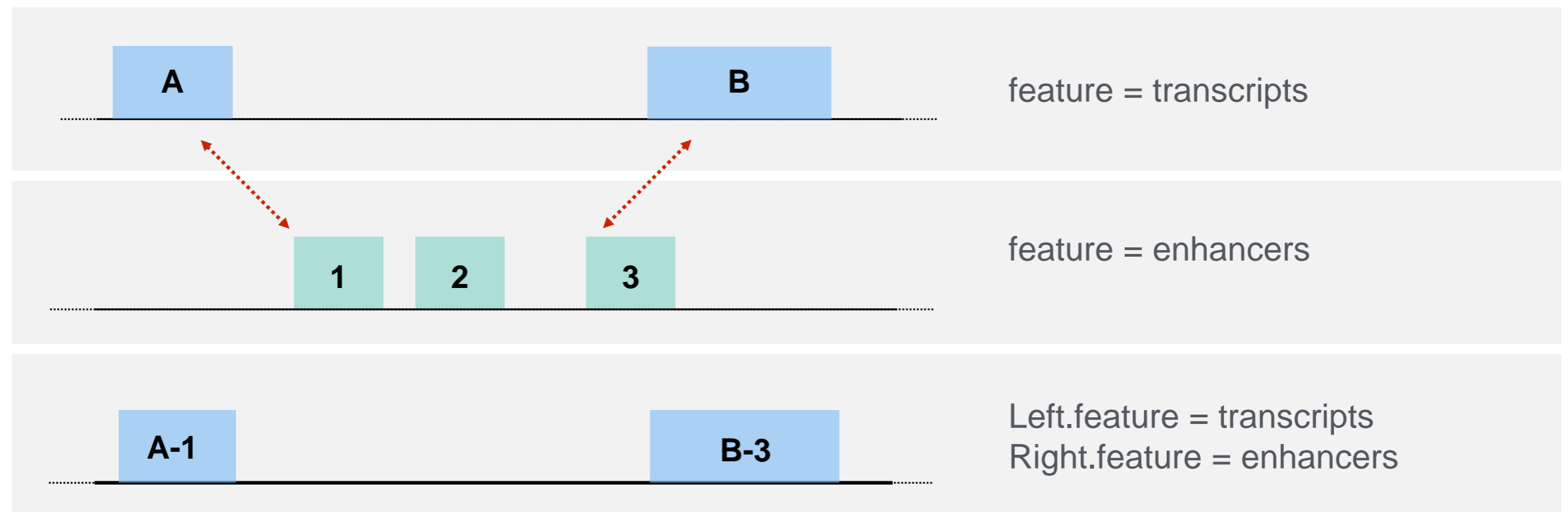
e.g. select those regions which have a score greater than 0.5)



REGION JOIN (GenoMetric)

Join at min-distance:

Associate each region in the former dataset with the closest in the latter.

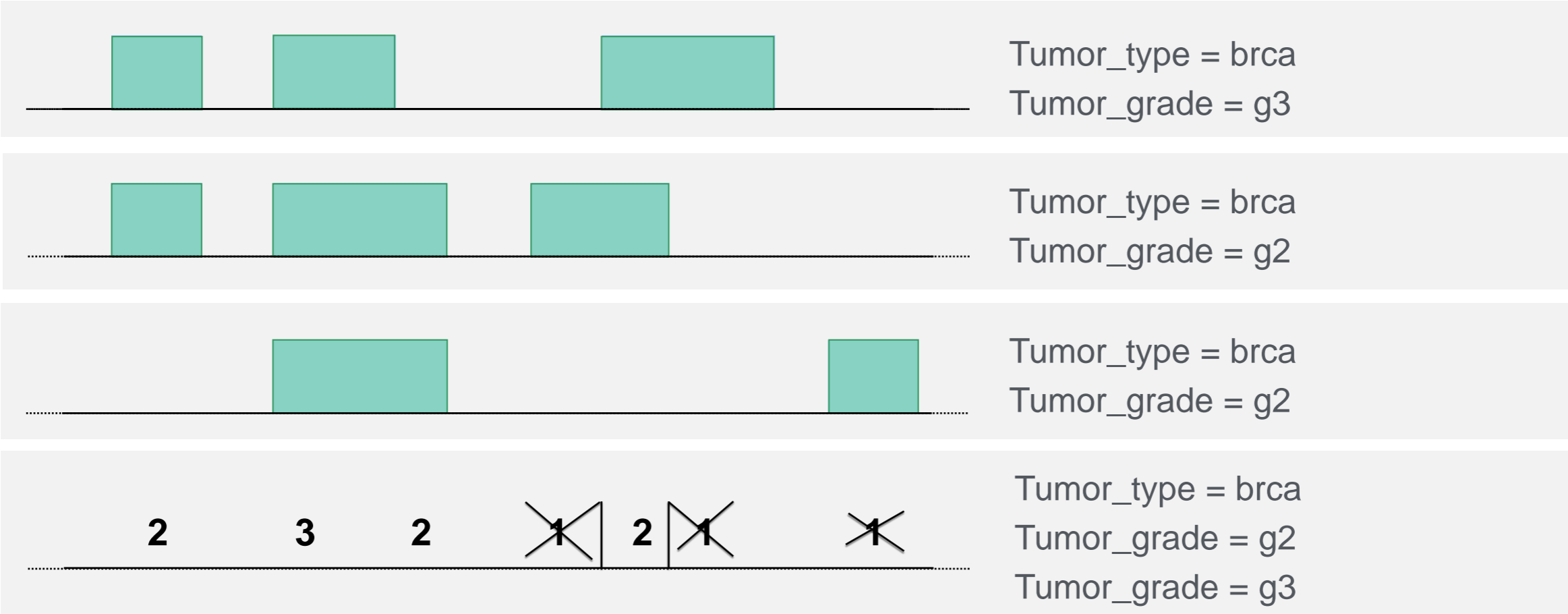


QUERY LANGUAGE

COVER

Cover(2,ANY)

Find portions of the genome that are covered by at least two regions

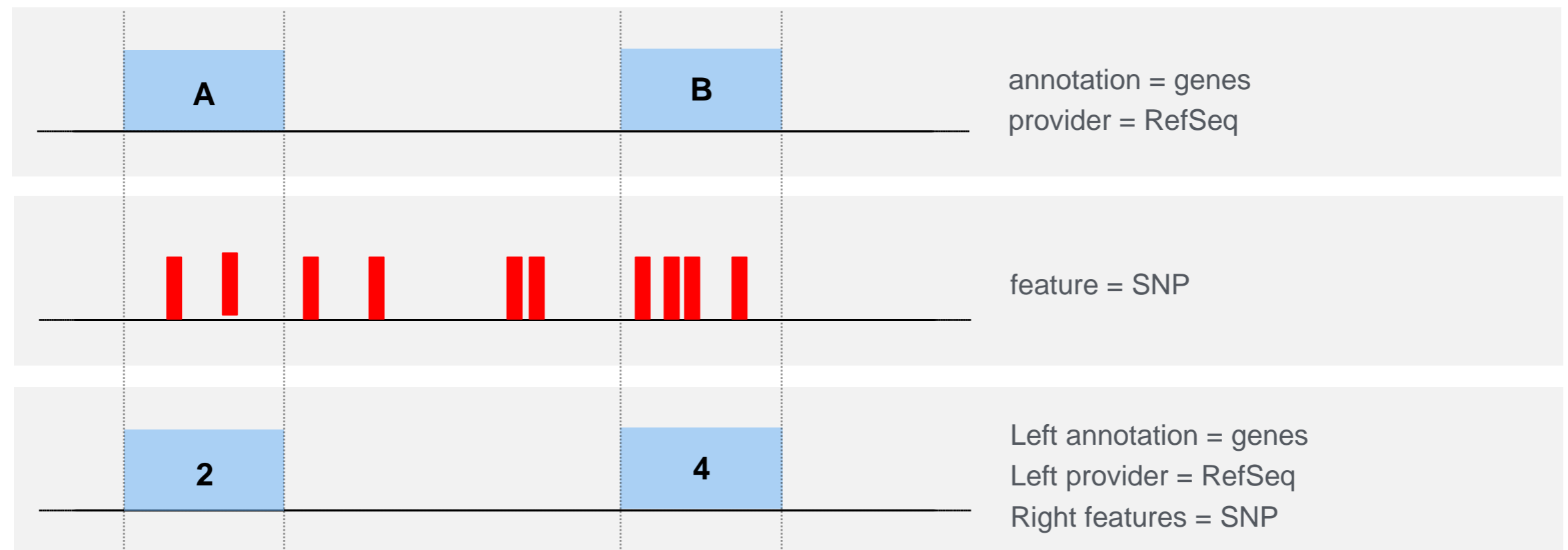


QUERY LANGUAGE

MAP

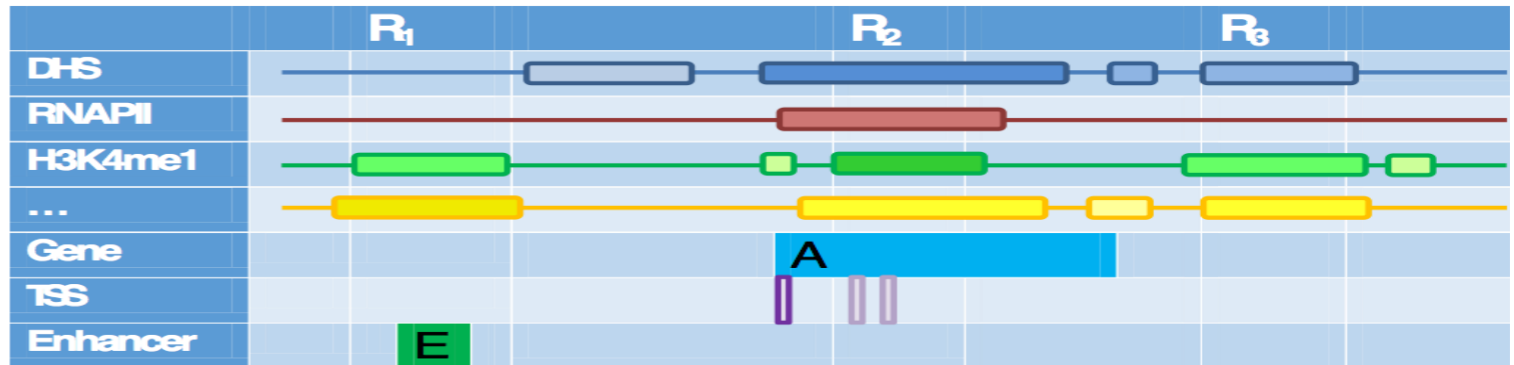
Region map

Compute an aggregate function (e.g. COUNT) on all the regions intersecting the reference



Genomic Space Abstraction

Map operations, through reference regions R , extract and standardize genomic features



MAP

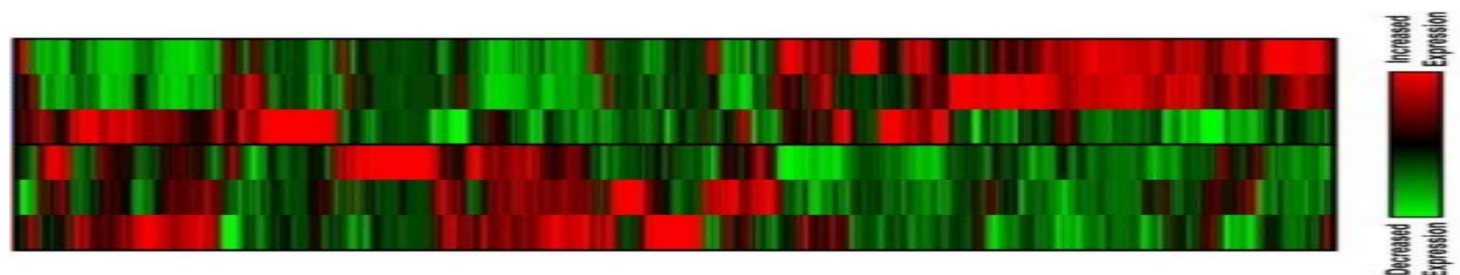
GENOMIC SPACE:

Simplified structured outcome, ideal format for data analysis

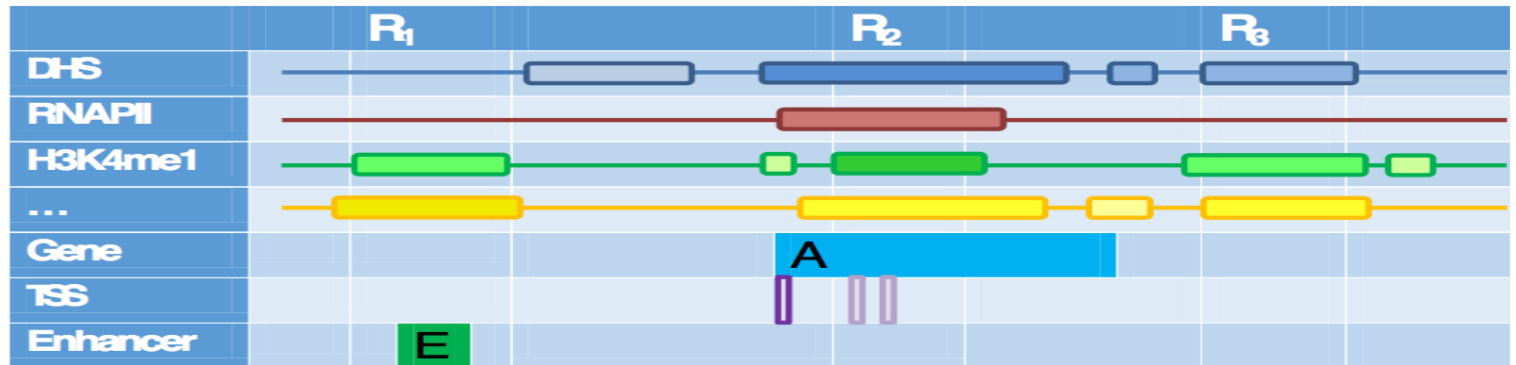
Genomic Space								
	E_1	E_2	E_3	...	E_m	E_{m+1}	E_{m+2}	E_{m+3}
R_1	10	3	2		12	0	1	0
R_2	1	48	12		3	0	0	0
R_3	11	9	10		0	1	0	1
...			
R_{n-1}	56	47	1		6	1	0	1
R_n	46	3	21		13	1	0	0

HEAT MAP:

Visualization of the genome space using intensity of colors



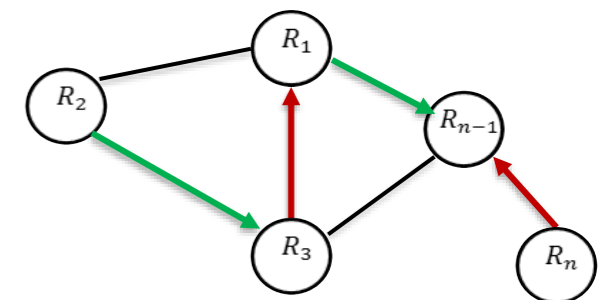
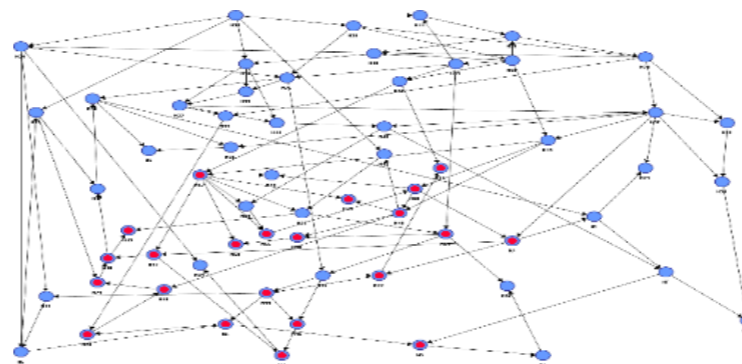
Genomic Space Abstraction



GENOMIC SPACE
represents adjacency
matrices, i.e. networks

Genomic Space								
	E_1	E_2	E_3	...	E_m	E_{m+1}	E_{m+2}	E_{m+3}
R_1	10	3	2	...	12	0	1	0
R_2	1	48	12		3	0	0	0
R_3	11	9	10		0	1	0	1
...			
R_{n-1}	56	47	1		6	1	0	1
R_n	46	3	21		13	1	0	0

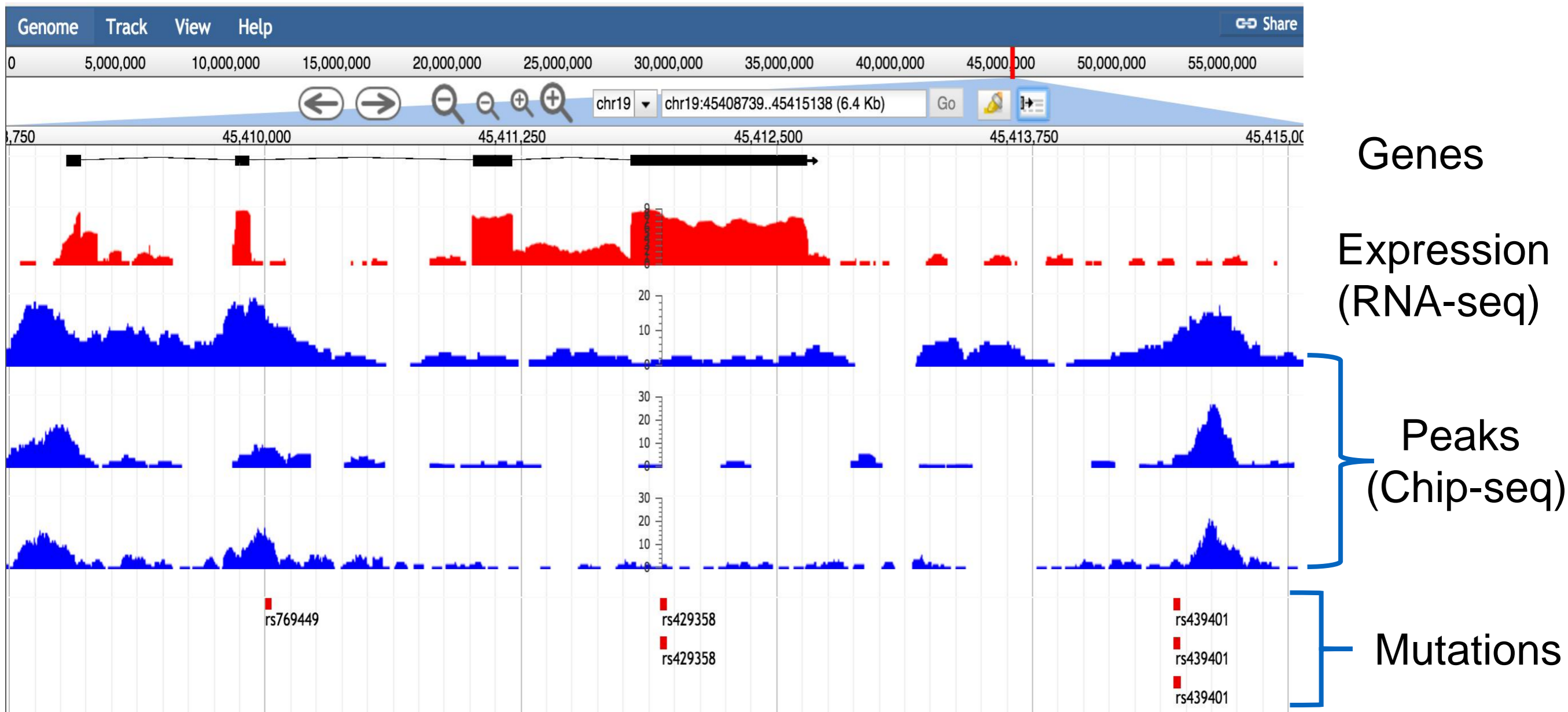
**Network analysis
methods** (e.g. page
rank, hub/authority,
community detection,...)



APPLICATIONS

Example of biological query

Given three replicas of a Chip.Seq experiment, extract high-confidence regions into one sample, identify which of these regions overlap with given genes, and for each resulting region count ICG mutations and select regions with at least one mutation.



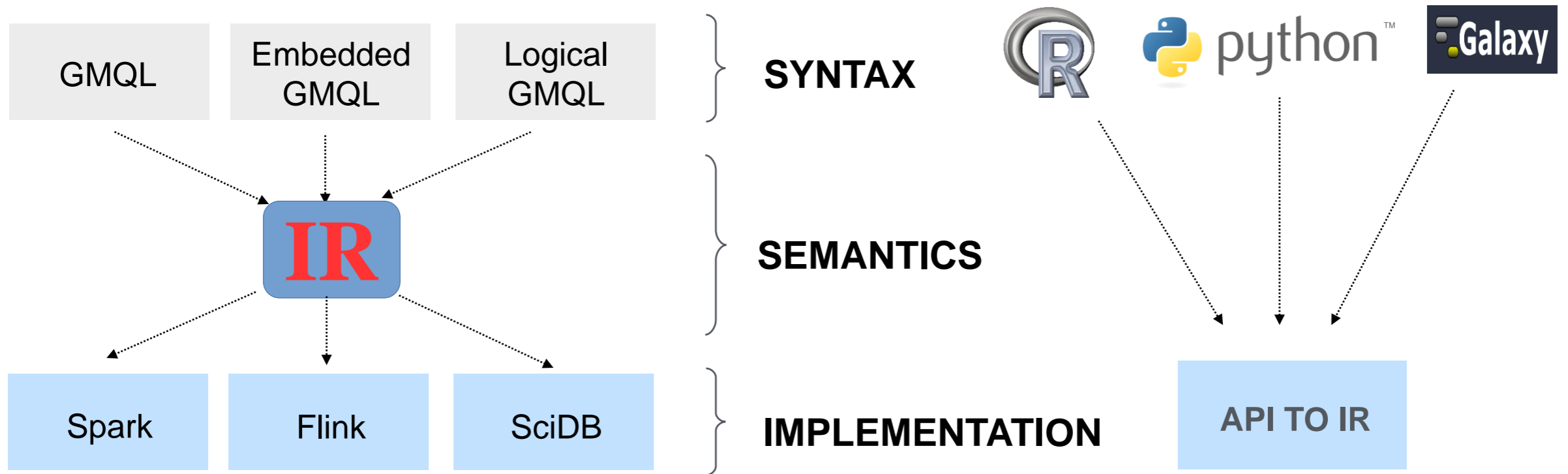
Video

<http://www.bioinformatics.deib.polimi.it/geco/?video>

ARCHITECTURE

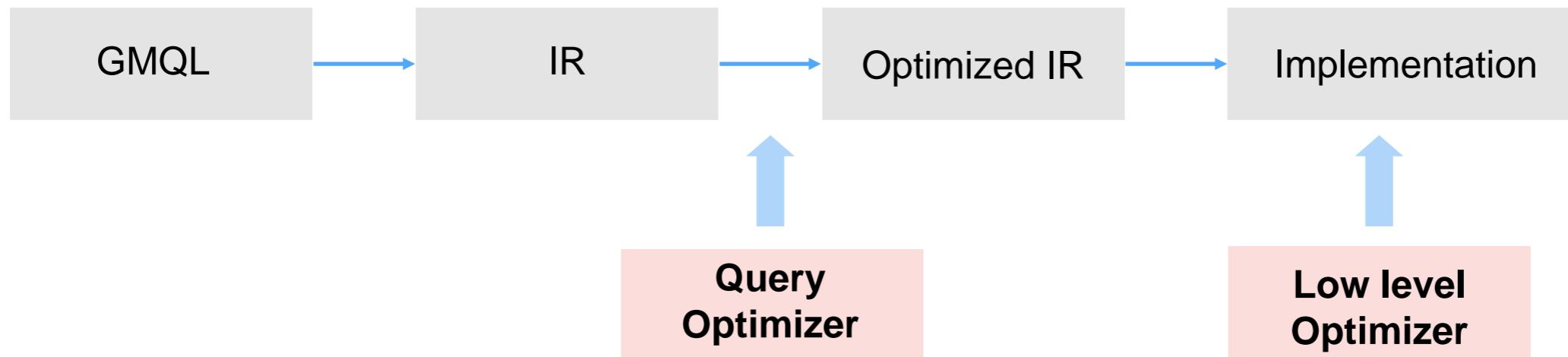
GMQL Implementation / V2

- A different approach, with language-independent intermediate representation
- Targeting also usability from within R and Galaxy



GMQL implementation, V2

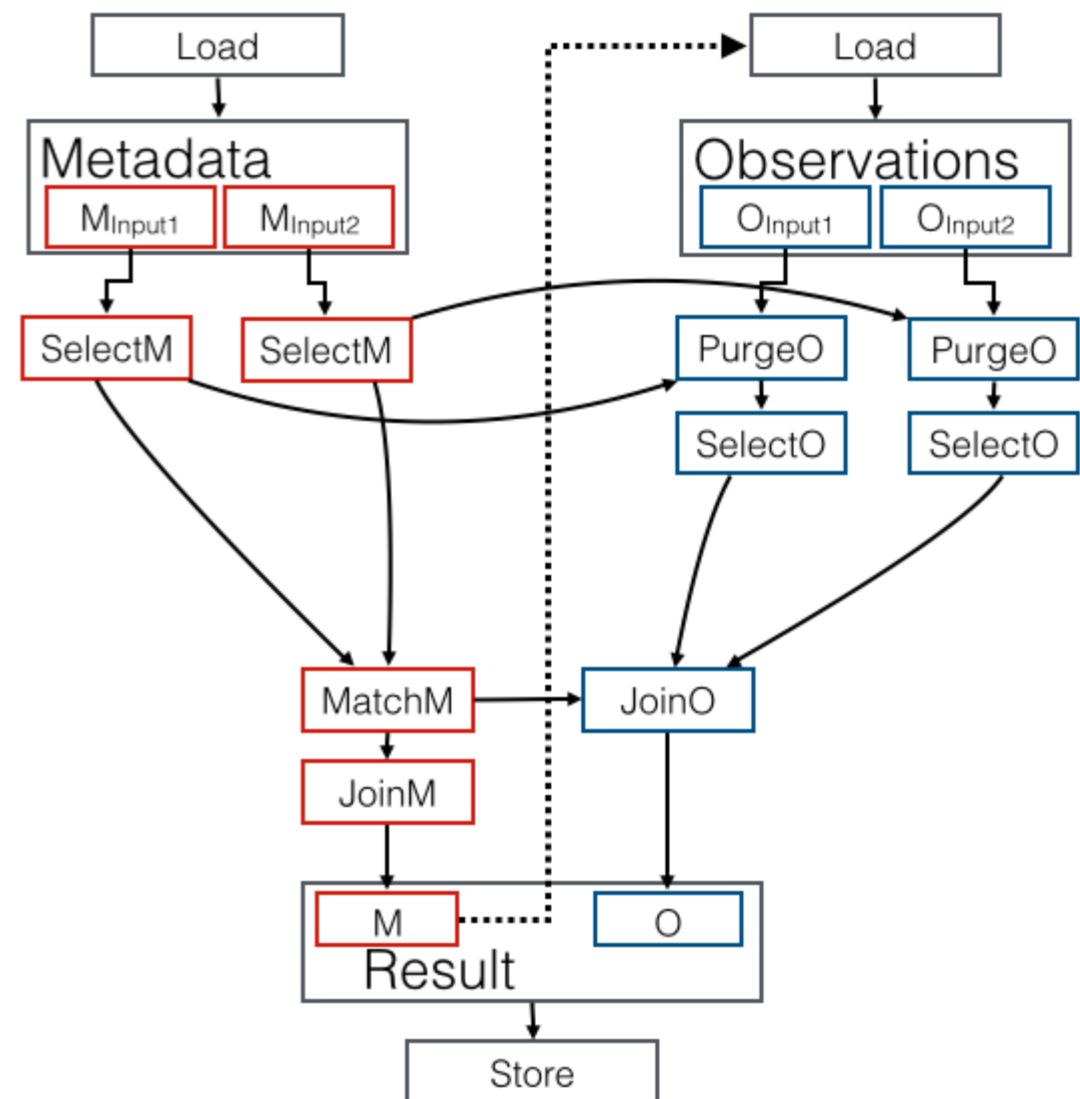
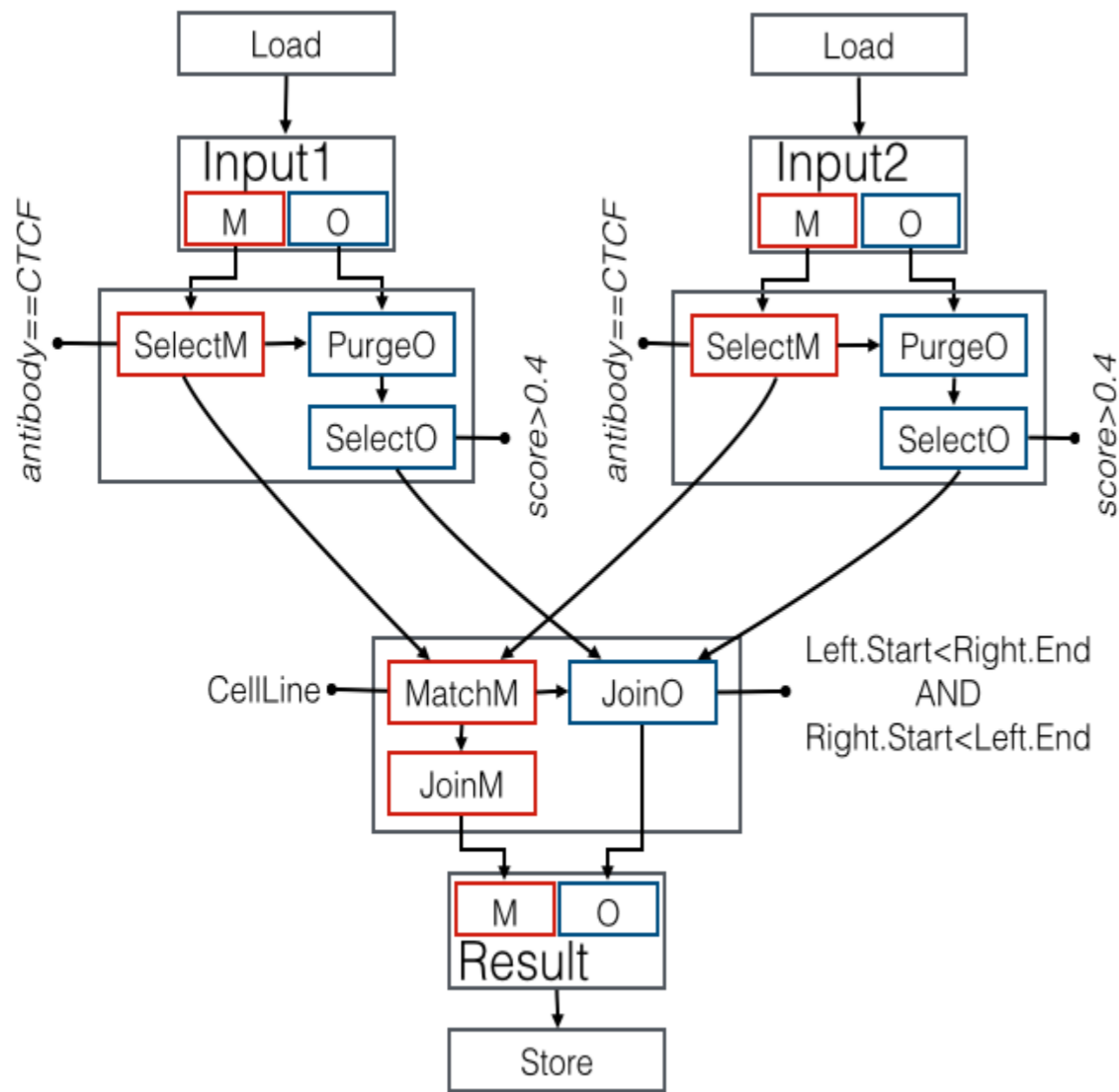
New optimization options



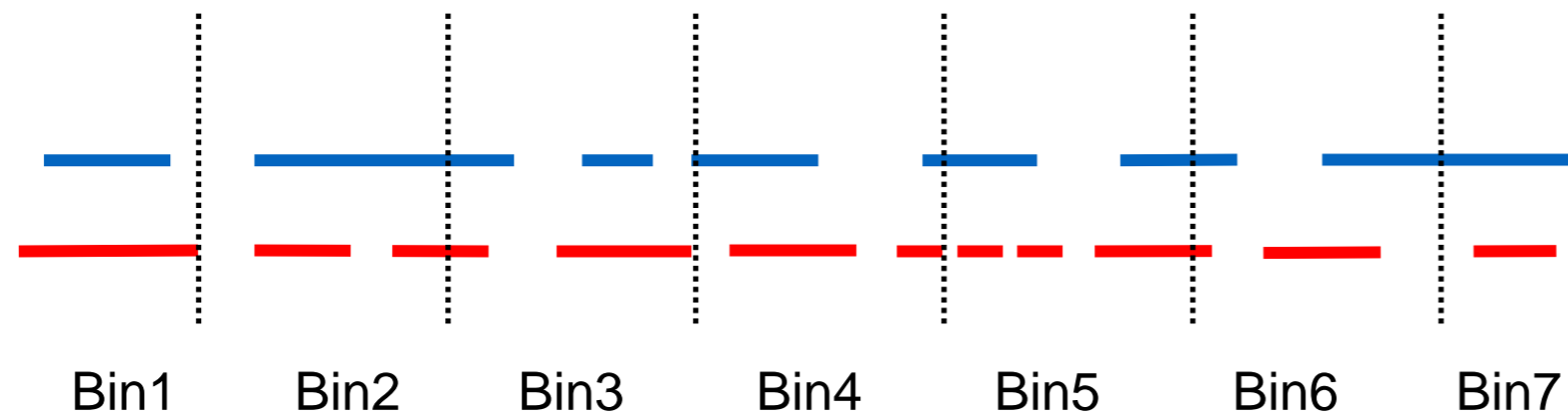
- 1) Node reordering / deletion
- 2) Select condition refinement

- 1) Alternative algorithms
- 2) Parallelism tuning
- 3) Data partitioning
- 4) Caching

Meta-first



Binning strategy



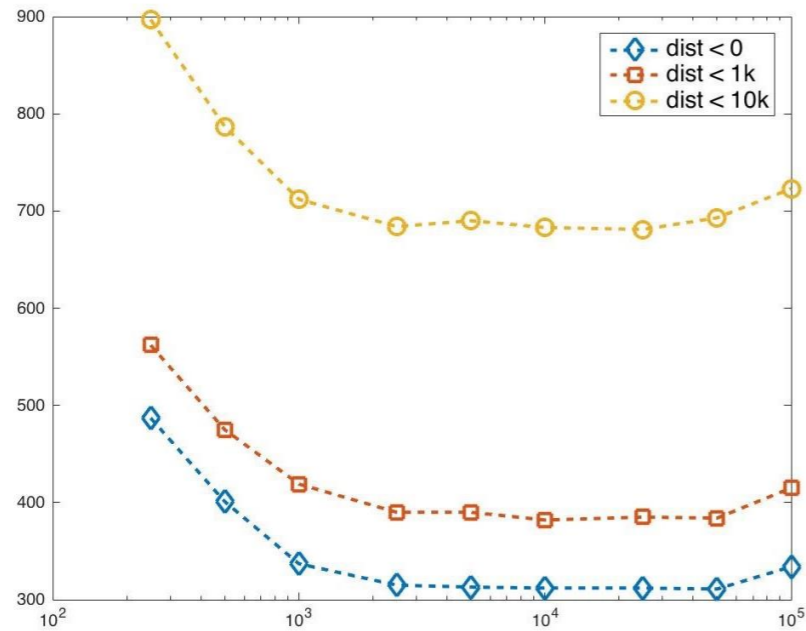
Strategy for intersection:

1. Partition the genome in bins
2. Assign each region to all the bins it overlaps
3. Search for intersections within each bin

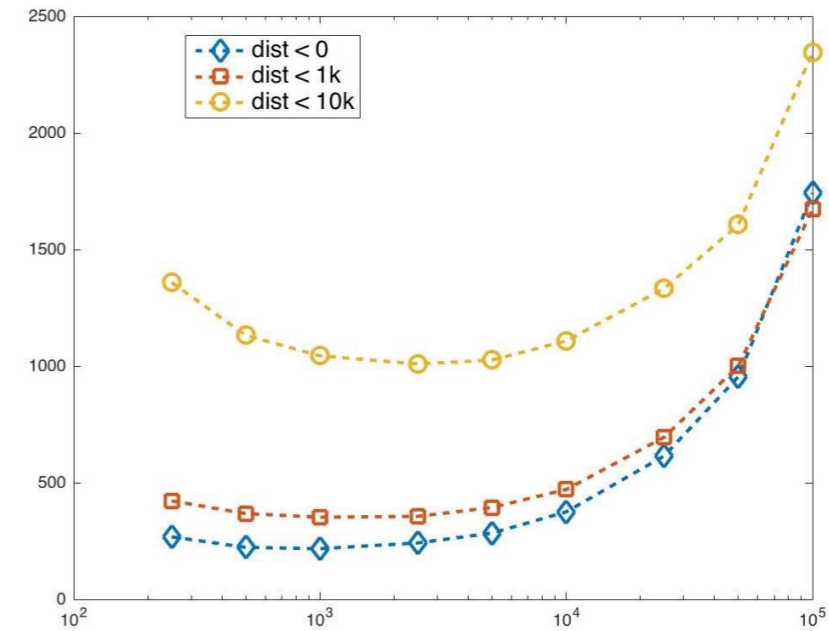
In the case of more complex operations, we change the way in which the regions are assigned to the bins

IEEE BIG DATA CONFERENCE , IEEE TS ON COMPUTERS, IEEE – TS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

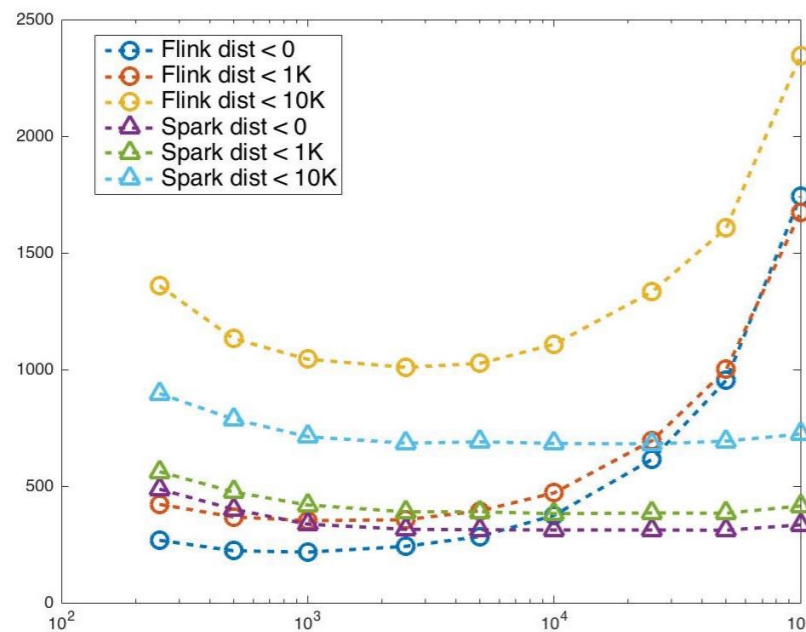
FLINK - JOIN



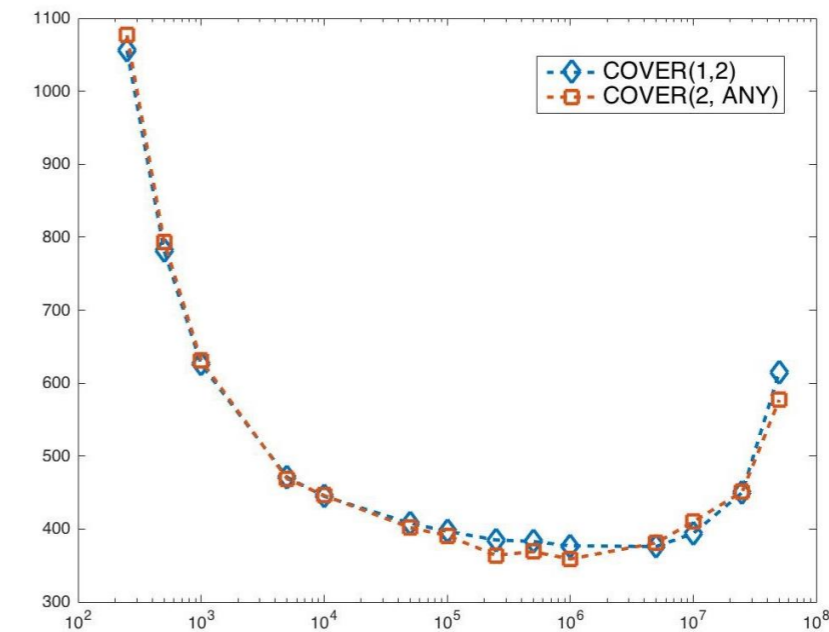
SPARK - JOIN



FLINK, SPARK - JOIN



FLINK - COVER



REPOSITORY

REPOSITORY

Stores experimental datasets and annotations collected from external databases:

- **ENCODE** (more than 4000 processed datasets for humans and mice, relevant to epigenomic research)
- **Epigenomics Roadmap** (about 1000 human epigenomic datasets for stem cells and ex-vivo tissues)
- **TCGA (The Cancer Genome Atlas)**, providing more than 50,000 processed datasets for more than 30 cancer types, including mutations, copy number variations, gene and miRNA expressions, methylations)

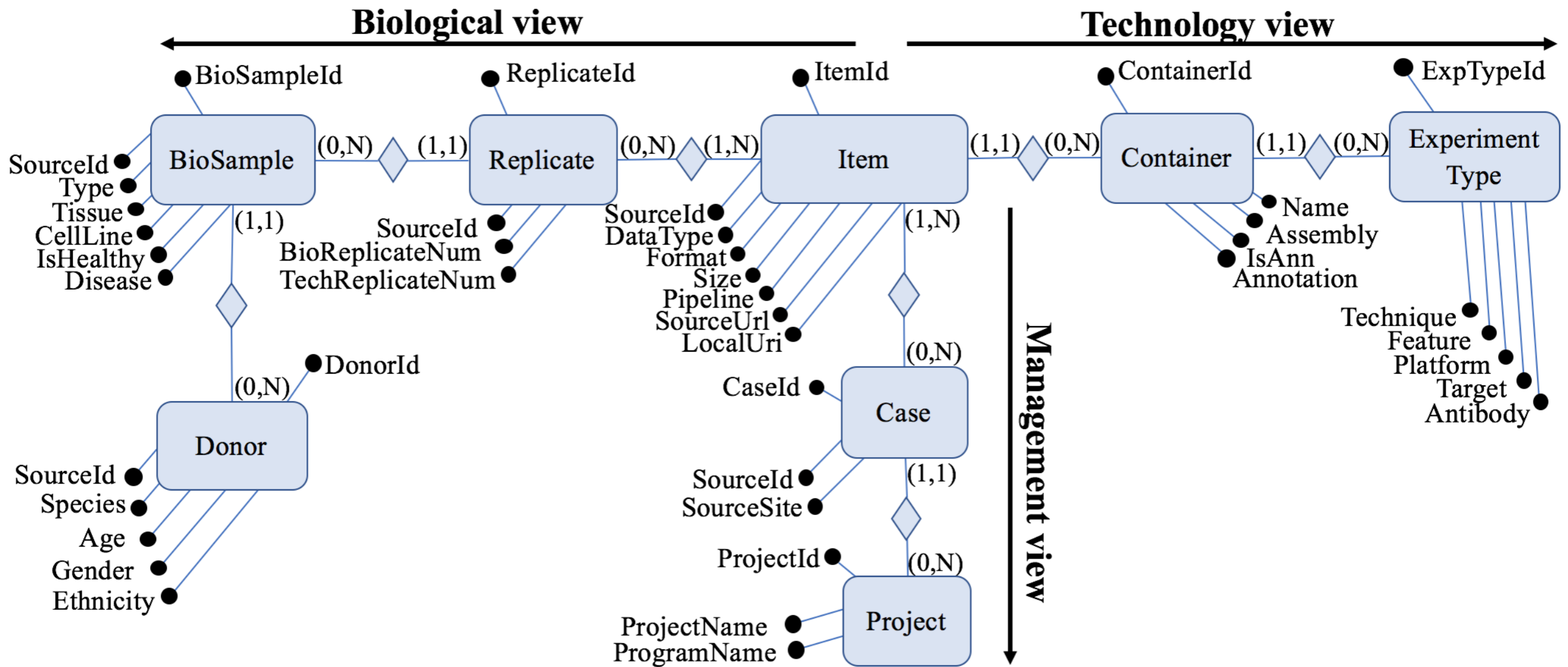
REPOSITORY

REPOSITORY

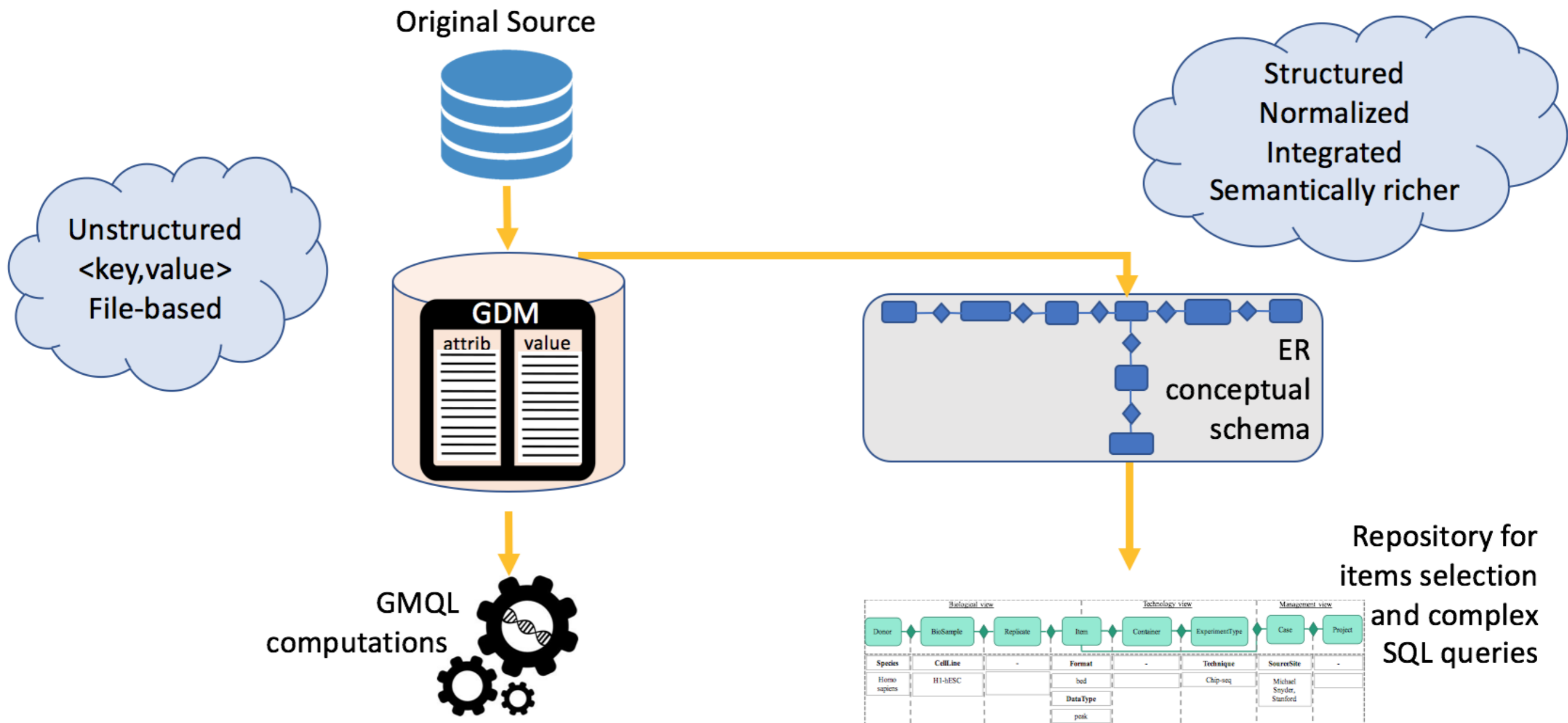
CONTENT

Consortium	Imported datasets	# of samples	File size (MB)
ENCODE	GRCh38_ENCODE_BROAD	366	2,776
	GRCh38_ENCODE_NARROW	10,542	113,646
	HG19_ENCODE_BROAD	2,136	24,423
	HG19_ENCODE_NARROW	11,468	107,291
EPIGENOMICS ROADMAP	HG19_EPIGENOMICS_ROADMAP_BED	78	595
	HG19_EPIGENOMICS_ROADMAP_BROAD	979	23,244
TCGA	HG19_TCGA_cnv	22,632	759
	HG19_TCGA_dnamethylation	12,860	236265
	HG19_TCGA_dnaseq	6,914	272
	HG19_TCGA_mirnaseq_isoform	9,909	4011
	HG19_TCGA_mirnaseq_mirna	9,909	711
	HG19_TCGA_rnaseq_exon	3,675	45459
	HG19_TCGA_rnaseq_gene	3,675	5080
	HG19_TCGA_rnaseq_spljxn	3,675	42320
	HG19_TCGA_rnaseqv2_exon	9,825	118583
	HG19_TCGA_rnaseqv2_gene	9,825	20848
	HG19_TCGA_rnaseqv2_isoform	9,825	50622
HG19_TCGA_rnaseqv2_spljxn	9,825	109756	
Grand total	18 datasets	138,118	906,661

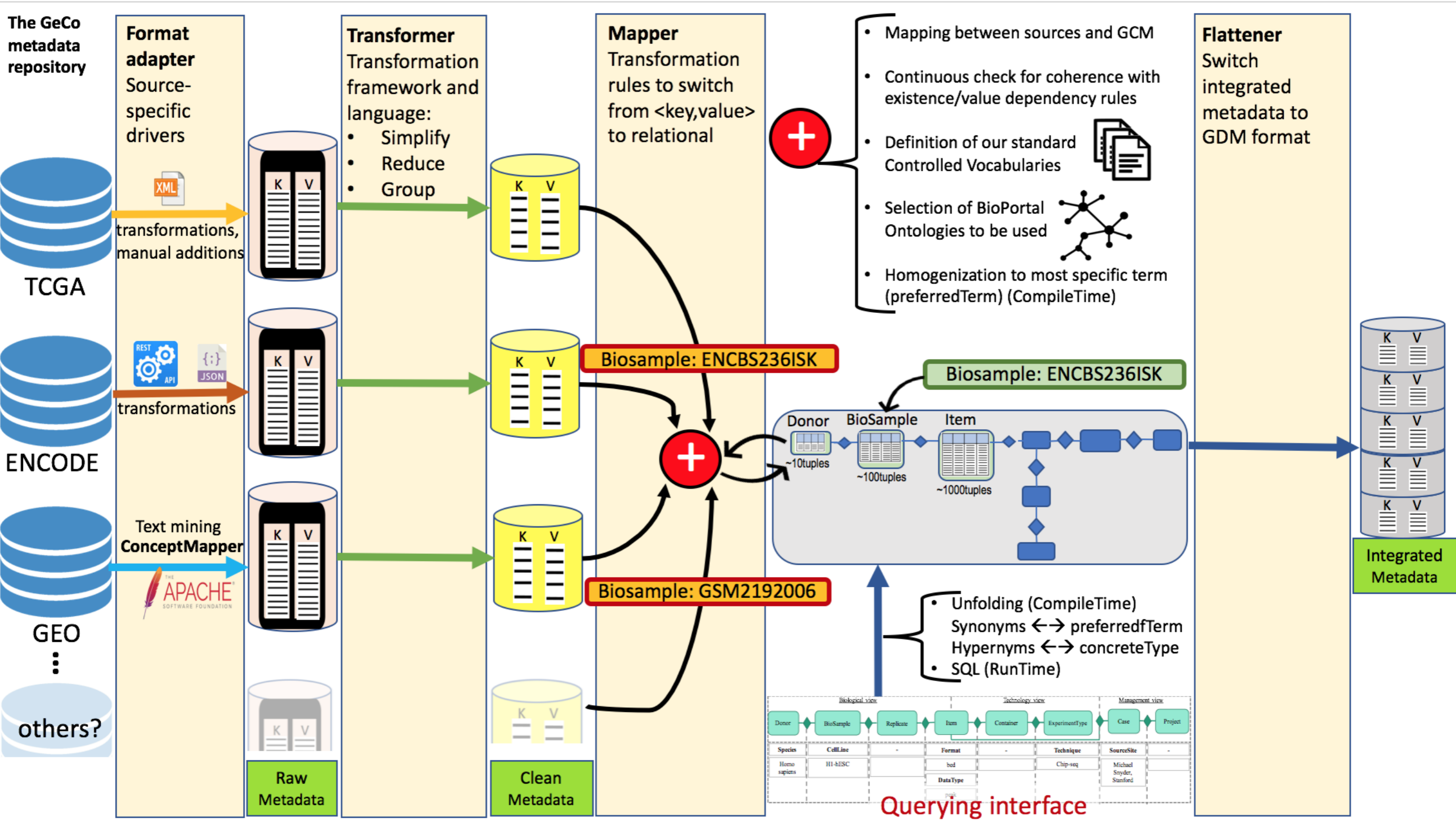
GeCo Conceptual Model



Extraction of Common Metadata



GeCo Repository Pipeline



COMPUTER SCIENCE COLLABORATIONS

joint work with:

SAPIENZA ROME (Prof.
Lenzerini)

UNIV. ROMA3 (Prof. Atzeni)

UNIV. BOLOGNA (Prof.
Ciaccia)

TU BERLIN (Prof. Markl)

PATTERN-BASED QUERIES

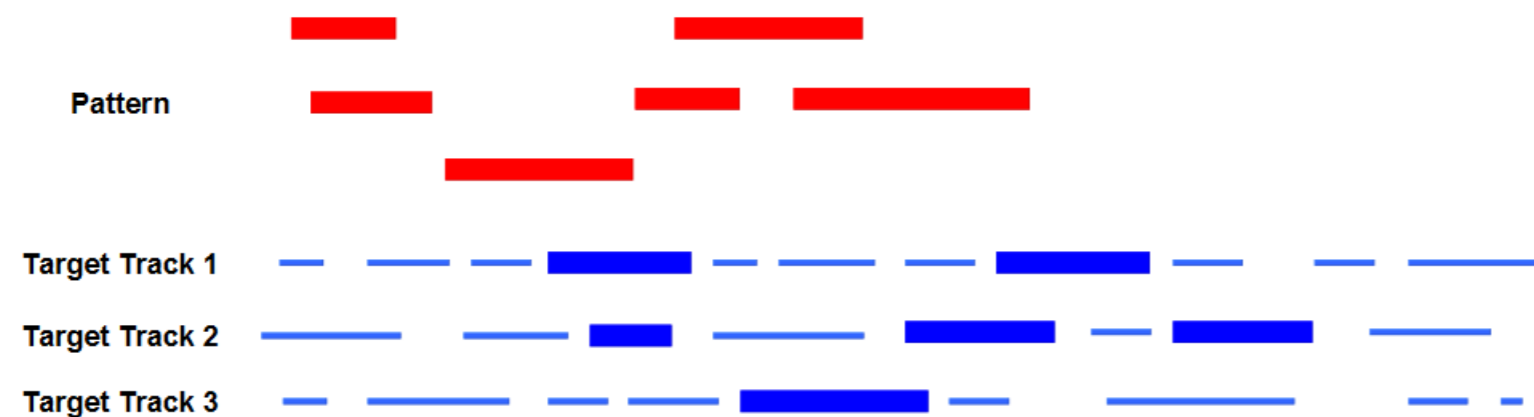
Search for matching regions, driven by patterns which are drawn on the screen using the Integrated Genome Browser; the method is based on dynamic programming.

Joint work with UniBo (Bartolini, Ciaccia, Montanari, Patella) published on **IEEE/ACM TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**

Pattern example



Pattern found



SEMANTIC UNDERSTANDING OF ENCODE METADATA

Search for approximate matching with Encode metadata by using the Unified Medical Language System (UMLS). Technically, we build the completion of the ontology w.r.t encode metadata, using forward chaining.

Joint work with Sapienza (Fernandez, Lenzerini), published on **IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS**

```

<doc>
  <field name="dccAccession"> wgEncodeEH001865</field>
  <field name="version"> ...
  <field name="creation_date"> ...
</doc>
  
```

ENCODE metadata document

```

<doc>
  <field name="attribute">cell_description </field>
  <field name="value">leukemia, "The continuous cell line K-562 was
  established by Lozzio and Lozzio from the pleural effusion of a 53-year-old
  female with chronic myelogenous leukemia in conceptual blast crises." -
  ATCC </field>
  <doc>
    <concepts count="13">
      <position pos="0_8">
        <concept id="C0023418">Leukaemia</concept>
      </position>
      ....
      <position pos="118_3">
        <concept id="C0580836">Old</concept>
      </position>
      <position pos="122_6">
        <concept id="C0015780">Female</concept>
        <concept id="C0043210">female</concept>
        <concept id="C0086287">female</concept>
        <concept id="C1705497">Female</concept>
        <concept id="C1705498">Female</concept>
      </position>
      ...
      <position pos="175_12">
        <concept id="C0005699">Blast Crises</concept>
      </position>
      ...
    </concepts>
  </doc>
</doc>
  
```

Enriched metadata pair document A_1 A_2 A_n

```

<doc>
  <conceptBase id="C0005699">Blast Crises</conceptBase>
  <parentConcept id="C0023473" level="1" ontology="[CHV, MSH, MTH, NCI, NDFRT, PDQ]">
    Chronic Granulocytic Leukemias</parentConcept>
  <parentConcept id="C0280251" level="1" ontology="[CHV, MSH, MTH, NCI, NDFRT, PDQ]">
    stage, chronic myelogenous leukemia</parentConcept>
  ...
  <parentConcept id="C0023470" level="2" ontology="[CHV, CSP, ICD9CM, MSH, MTH, NCI, NDFRT, PDQ]">
    Myeloid Leukemias</parentConcept>
  <parentConcept id="C0006826" level="2" ontology="[CHV, MEDLINEPLUS, MSH, MTH, NCI, NDFRT, PDQ]">
    Cancers</parentConcept>
  ...
  <parentConcept id="C0023418" level="3" ontology="[CHV, CSP, ICD9CM, MSH, MTH, NCI, NDFRT, PDQ]">
    Leukemias</parentConcept>
  ...
</doc>
  
```

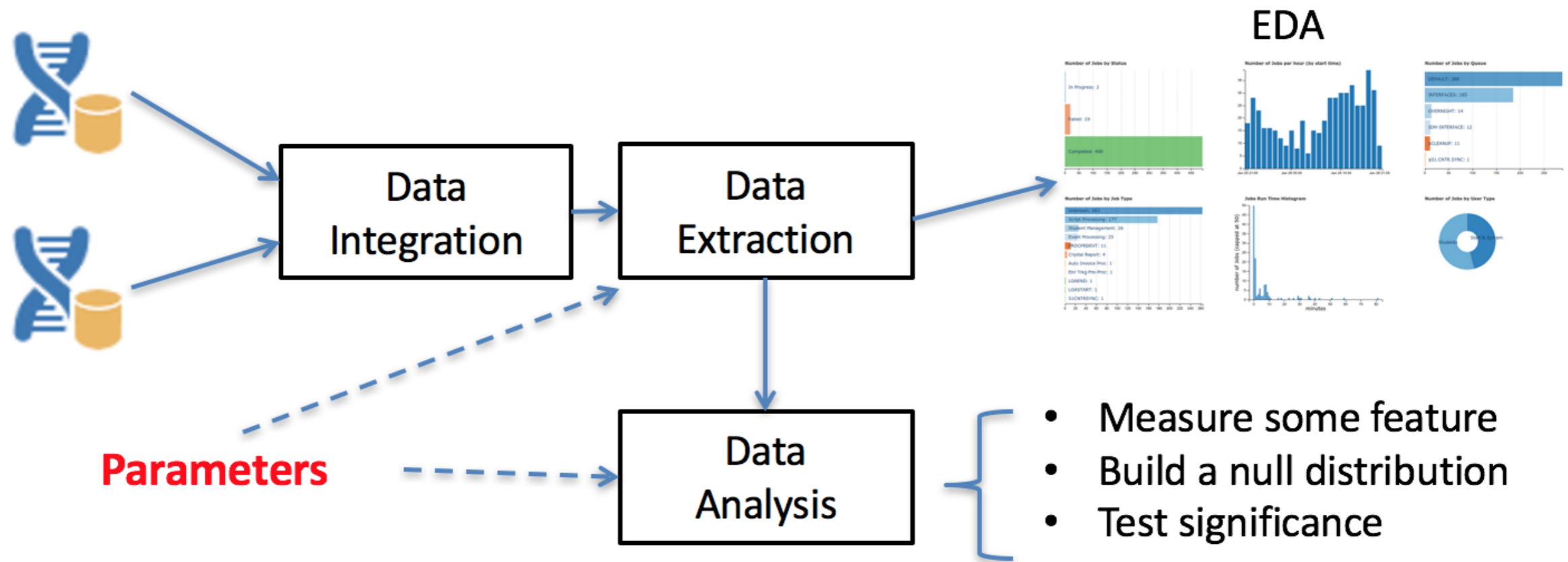
concept document C_1 C_2 C_j

BIOLOGICAL RESEARCH

joint work with
IEO (Pier Giuseppe
Pelicci's group),
NUS SINGAPORE
(Prof. Ken Wing and
Limsoon Wong)
HARVARD IACS (Prof.
Pavlos Protopapas)

Principled software design of analysis pipeline

- Transcription factor dimers
- Synthetic lethal mutations
- Downregulation at Tad boundaries
- Characterization of Tads



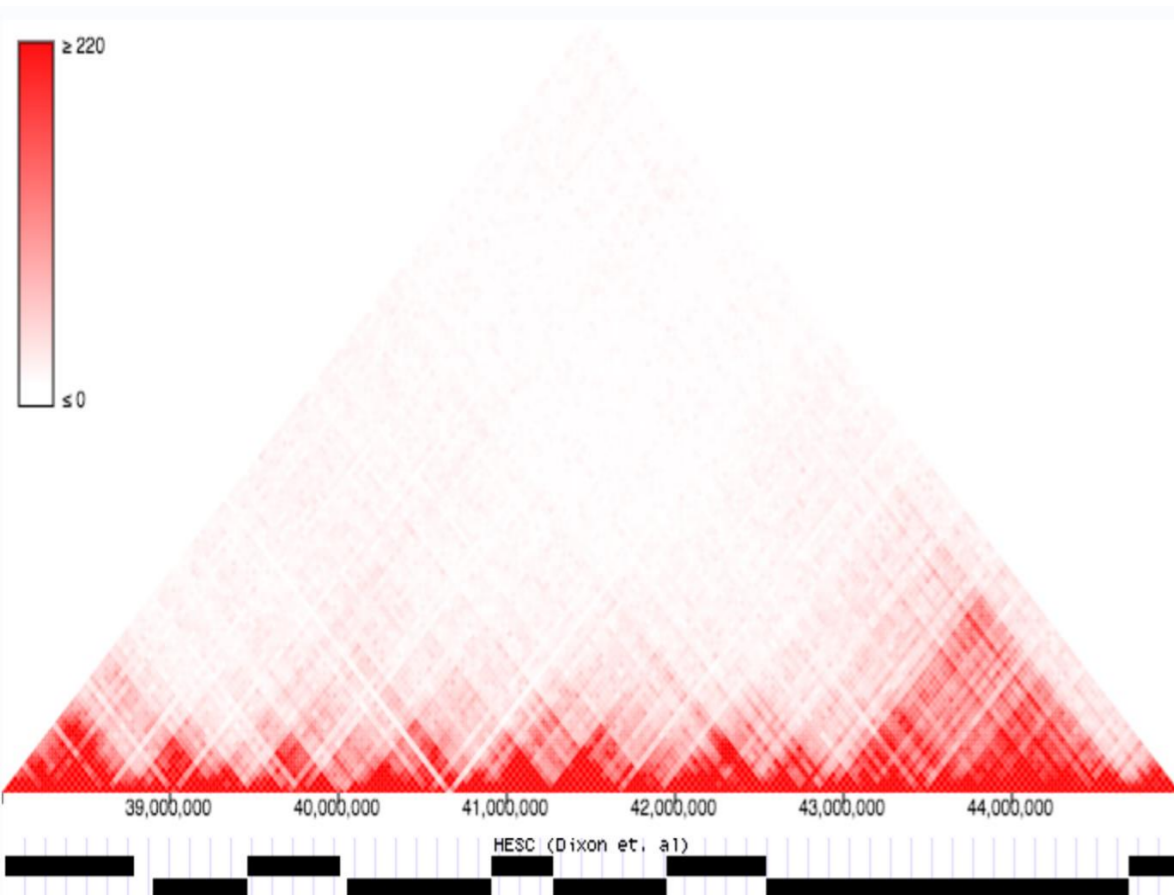
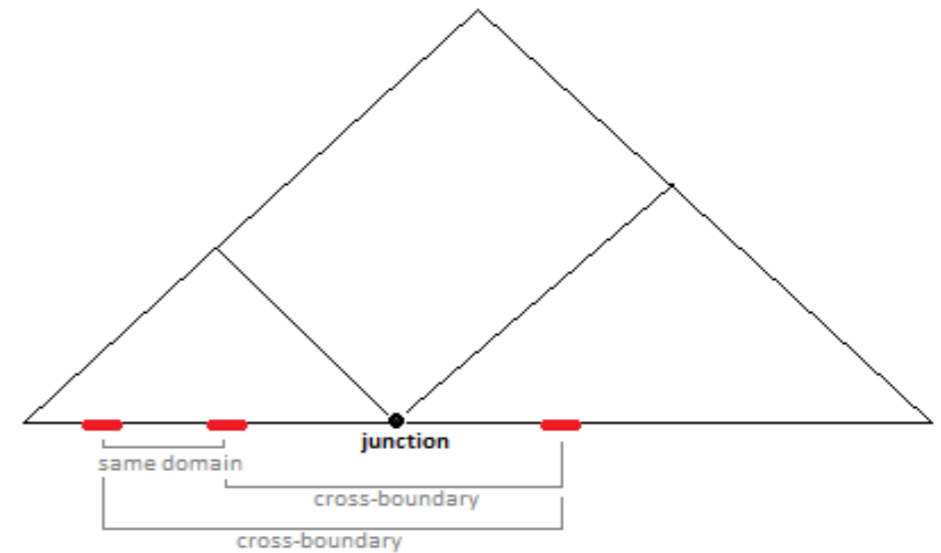
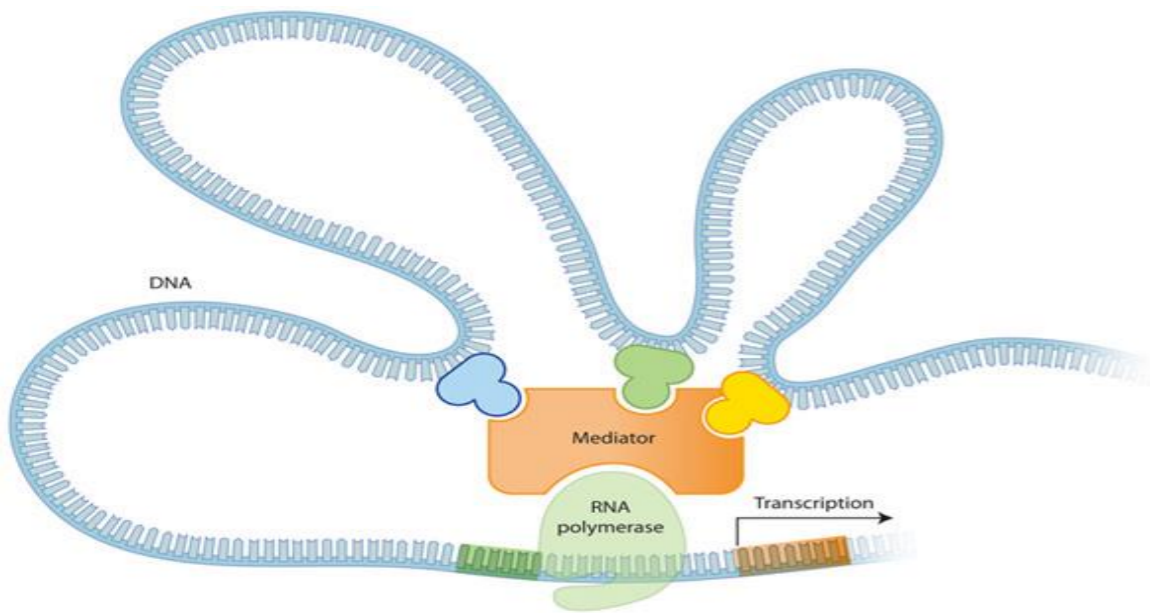
Challenges:

- Reproducibility
- Domain specific operations
- Flexibility



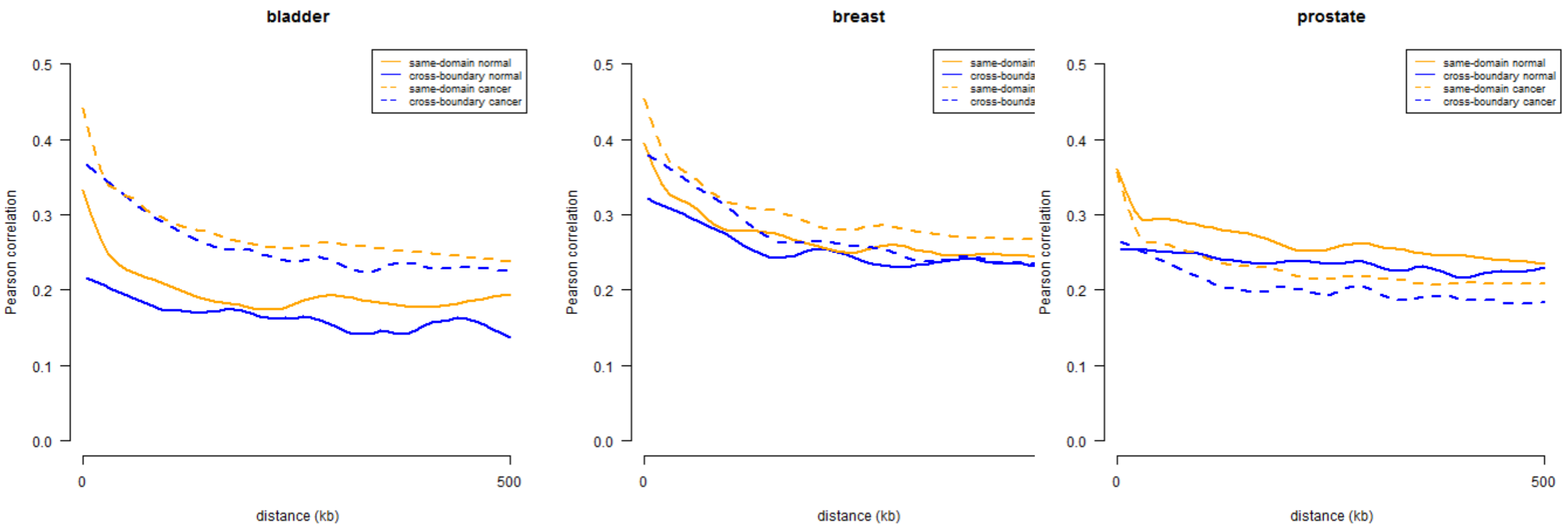
GMQL + Python + Jupyter notebook

3D Structure and tumors

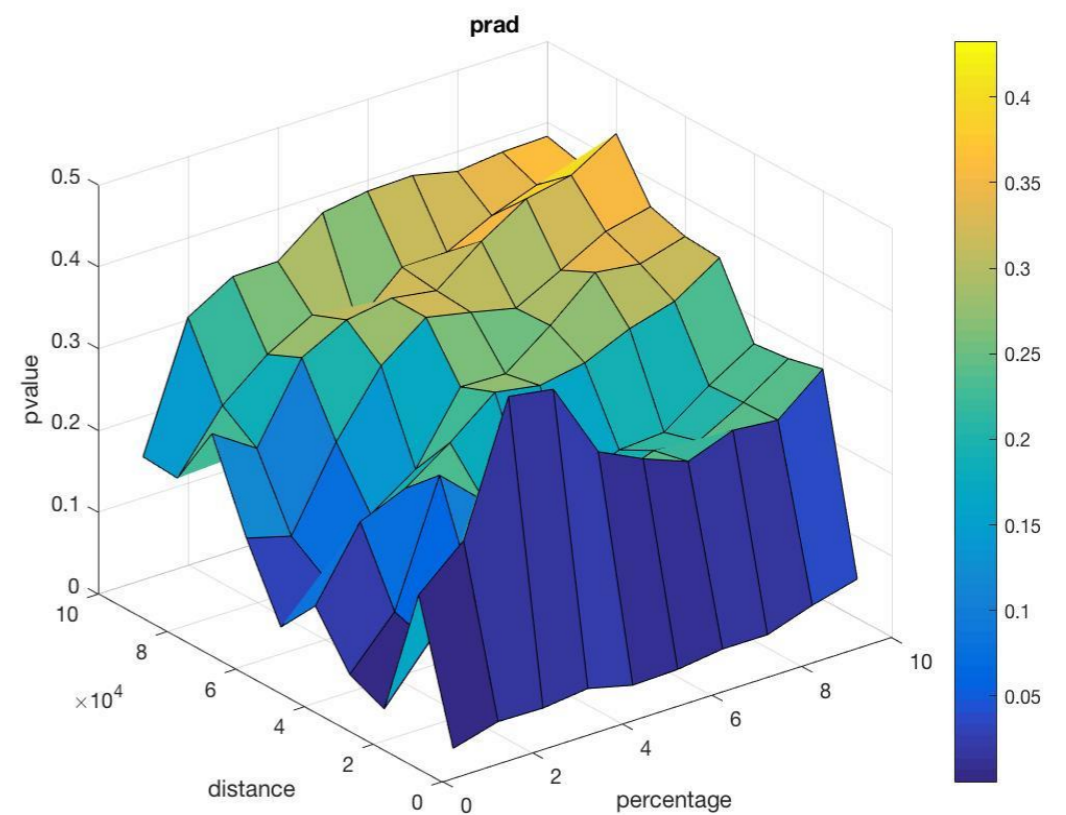
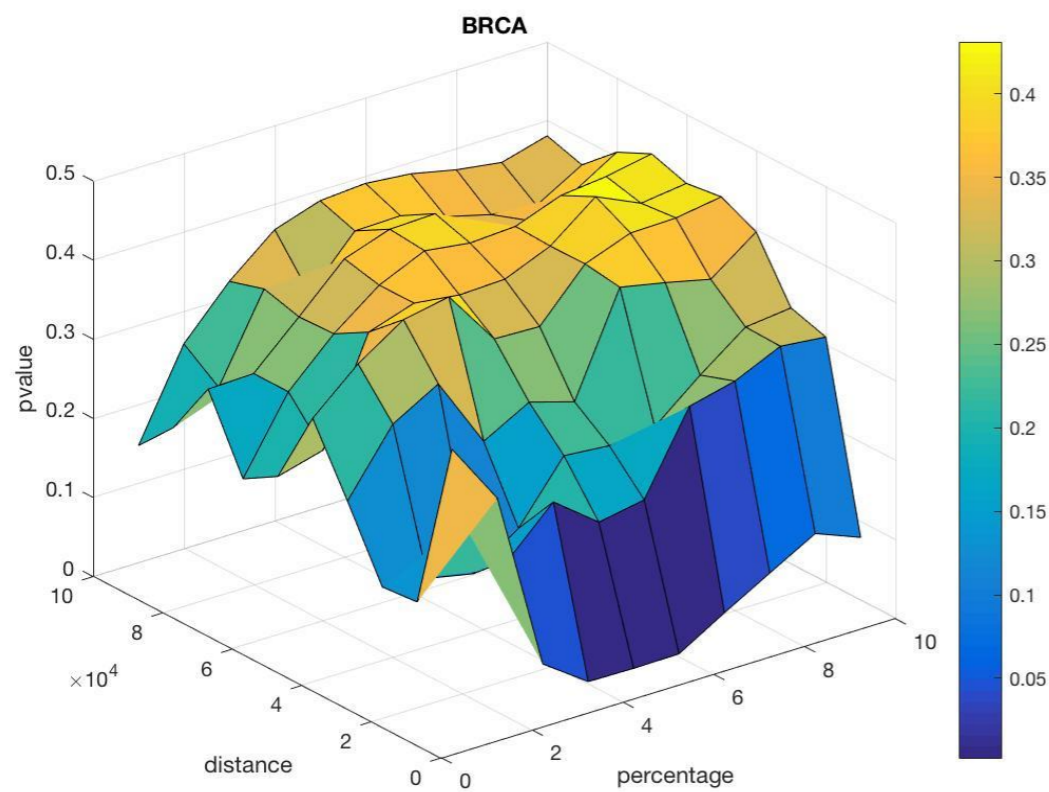


```
GENES = SELECT(tissue=="cortex") GTEx;  
J1 = SELECT(cell=="imr90") TADs;  
J2 = SELECT(cell=="gm12878") TADs;  
PAIRS = JOIN(distance < 500000; CONTIG)  
GENES GENES;  
MAPPING_t = MAP() PAIRS J1;  
MAPPING = MAP() MAPPING_t J2;  
SAME = SELECT(left.id < right.id and L1.count == 0  
and L2.count == 0) MAPPING;  
CROSS = SELECT(left.id < right.id and L1.count > 1  
and L2.count > 1) MAPPING;
```

Same/cross gene activity correlations in normal vs tumor cells



Significantly disregulated junctions in tumors



Other Biological Problems

- Dimers: pairs of TFs that co-regulate genes in rigid and compact pairs [with Limsoon Wong (NUS Singapore)].
- Super-TADs: clusters of topological domains [with Ken Wing and Limsoon Wong (NUS Singapore)]
- Killer Mutations: pairs of mutations: when both present they cause the death of the cell [with Limsoon Wong (NUS Singapore) and S. Srihari (EMBL-Australia)].
- Identification of TFs that co-occur with TEAD4 binding sites [with Stefano Campaner (IEO-IIT)].
- Detect DNA areas where multiple TFs bind (dense TF binding regions) [with Stefano Campaner (IEO-IIT)].
- DNA Sequencing of Microbioma in Cystic Fibrosis patients who are colonized with mycobacterium abscessus [with N. Segata (UniTn), L. Cariatì et Al. (Policlinico Milano), G. Porta (U. Insubria), J. LiPuma (U. Michigan)].



VISION





Short-Term Goals

DESCRIPTIVE STATISTICS

Provide automatic summarization describing result samples; integrate classic significance or regression tests within the query capabilities.

METADATA TRACING

Develop methods and tools supporting users in explaining observed query outputs. The study of data causality is based on determining data lineage (or provenance), especially relevant with queries over multiple sources

PATTERN-BASED REGION EXTRACTION

Define complex patterns of genomic features enabling the formulation of similarity queries (e.g., distal patterns, or using the notions of similar/dense/sparse genomic regions).



Mid-Term Goals

INTEGRATED REPOSITORY

Produce an integrated repository with semantically well-defined and compatible metadata, by integrating GDM with ENCODE, TCGA, 1000 Genomes and Roadmap Epigenomics (and possibly other sources).

WEB SERVICES

Use GMQL for building several public web services for solving general-purpose biological problems, supporting powerful statistics to indicate the significance of query results.

INTERACTION NETWORKS, MACHINE LEARNING, DEEP LEARNING

Provide automatic interpretation of query results as interaction networks or build tight integration with data analysis methods, e.g., based upon machine learning or deep learning.



Long-Term Goals

SEMANTIC AND FEATURE-BASED SEARCH

Develop semantic metadata search with semantic query expansion (leveraging on available ontologies e.g., OBO, UMLS) and region-based search patterns. Provide results in ranking order (as in classic search engines).

GENOMIC RECOMMENDERS

Trace query histories and build recommending systems for the “best” ways of solving genomic problems.

INTERNET OF GENOMES

Use GMQL as a basis for simple interaction protocols for:

- **Requesting information** about remote datasets, using both metadata and region schemas
- **Sending a query** and obtain data about its compilation, (including also estimates of the data sizes)
- **Launching execution** and then controlling the staging resources and of communication load

Website:

<http://www.bioinformatics.deib.polimi.it/geco/>

[Home](#)[Scenario](#)[Approach](#)[Video](#)[TryGMQL](#)[Workplan](#)[Projects](#)[Team](#)[Collaborations](#)[Events](#)[Publications](#)[OpenCalls](#)

TRY GMQL

WEB Interface

Includes:

- Web interface for browsing datasets and building GMQL queries
- Processed data from ENCODE, Roadmap Epigenomic and TCGA public sources

REST APIs

Includes:

- REST APIs for programmatic access to GMQL repository and query execution engines

GitHub Site

Includes:

- GeCo group GitHub repository (contains all the open source files of the GeCo project)

Downloads

Includes:

- Local mode or MapReduce mode (over Hadoop, or Hadoop YARN) for GNU/Linux systems
- Quick start - Install GMQL and get started

Documentation

Includes:

- GMQL Introduction to the language [pdf](#)
- GMQL Examples (Draft) [pdf](#)

Contact us

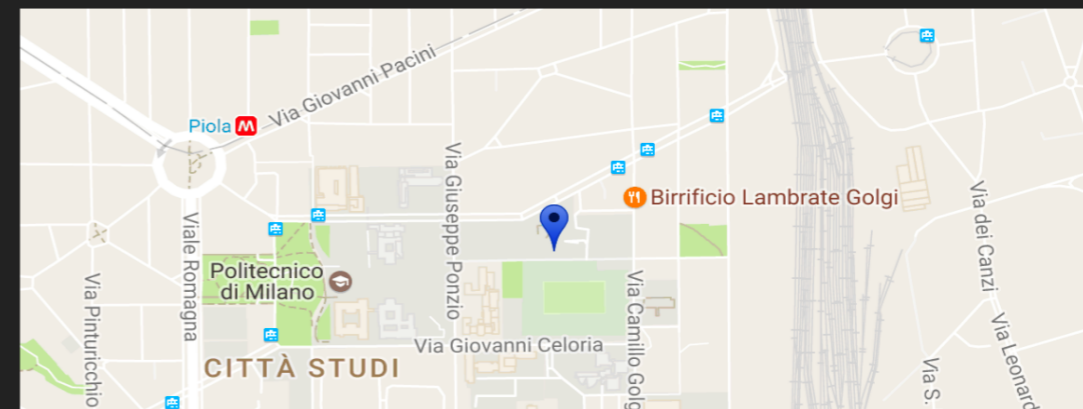
Politecnico di Milano
DEIB | Dipartimento di Elettronica Informazione e Bioingegneria

📍 Via Ponzio 34/5, Milan - Italy

✉ Stefano Ceri: stefano.ceri@polimi.it

✉ Marco Masseroli: marco.masseroli@polimi.it

☎ +39 02 23993400



Resources & Websites @ CINECA

We opened a link to CINECA supporting:

- **web interface** where bio-informaticians can browse the datasets of genomic features and biological/clinical metadata and build GMQL queries upon them
- **processed data from TCGA, ENCODE, Roadmap Epigenomics public sources** (open, anonymized data for secondary use)

<http://genomic.elet.polimi.it/gmql-rest/>

Examples of biological questions/exercises

Given expression data of all genes (as from TCGA, RNAseq gene expression datasets for tumor/normal cells):

- Build a classifier that can guess the label of an experiment
- Extract the 100 genes for each tumor that are the best tumor predictors.
- Look for the transcription factors (as from Encode ChIPseq) that most correlate with those genes (use region intersection and aggregation)
- Rank them by relevance according to various criteria (number of bindings, bindings weighed by score, bindings weighed by shortest distance to the gene's promoter)

... and soon you start to be very involved in genomics!