POLITECNICO DI TORINO

Bioinformatics methods and tools for identifying disease-associated B-cell populations, miRNA patterns and gene fusions in RNA-Seq data

Candidate: Giulia Paciello
Supervisors: Elisa Ficarra and Enrico Macii

Torino, 1st June 2017

POLITECNICO DI TORINO

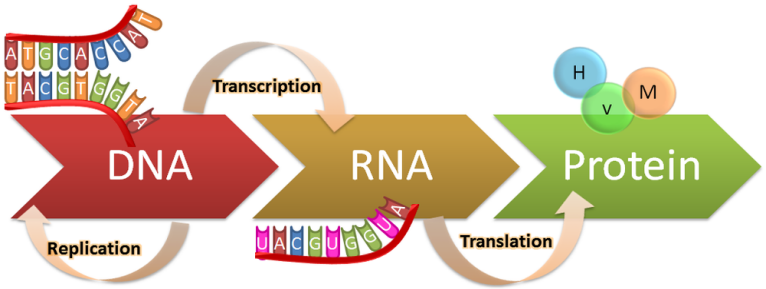
Outline

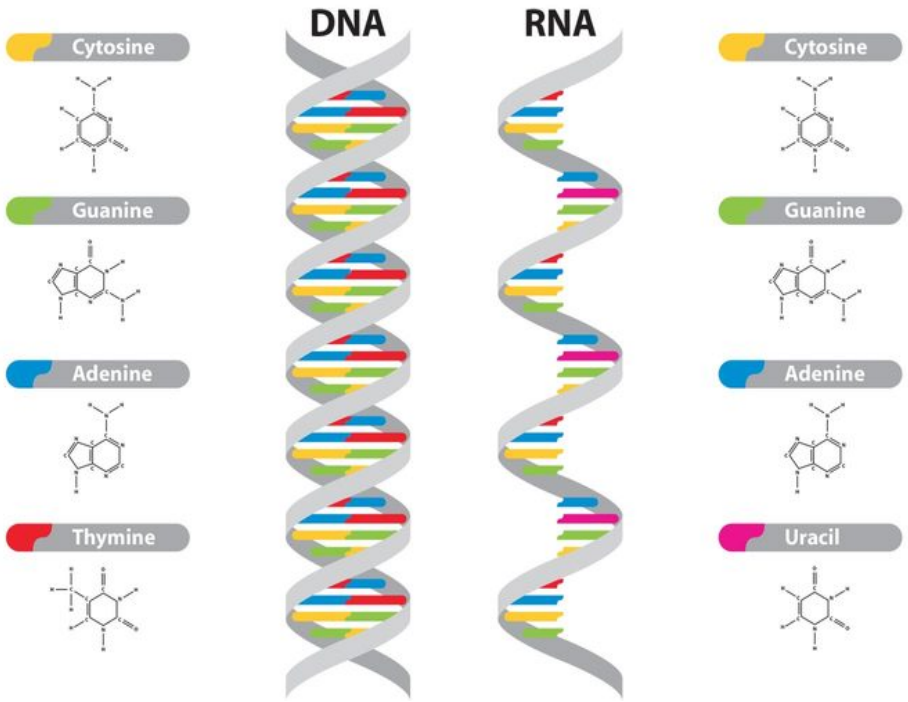
- ✓ Introduction
- ✓ Objective
- ✓ Methodologies and tools: VDJSeq-Solver, isomiR-SEA, FuGePrior
  - Biological Background
  - Motivation
  - Algorithm
  - Results
  - Remarks and Perspectives
- ✓ Conclusions

POLITECNICO DI TORINO

What does the term ‘sequencing’ mean?

The **Central Dogma of Molecular Biology**



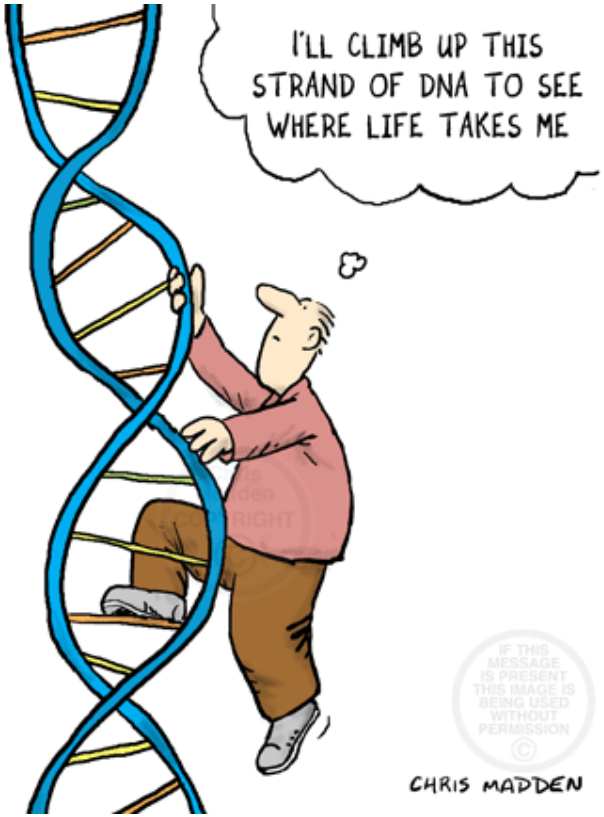


The **sequencing** is the process of determining the **sequence of nucleotide bases** in a nucleic acid molecule

POLITECNICO DI TORINO

Why is sequencing important?

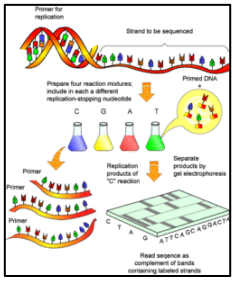
"…Decoding of the DNA that constitutes the human is a fundamental contribution toward understanding human **evolution**, the causation of **disease**, and the interplay between the environment and heredity in defining the human condition…"



Venter, J. Craig, et al. **"The sequence of the human genome."** Science 291.5507 (2001): 1304-1351

POLITECNICO DI TORINO

The sequencing time line

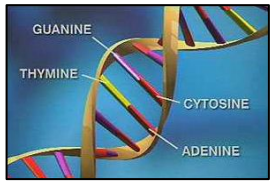


1977: Sanger-Coulson sequencing techinque

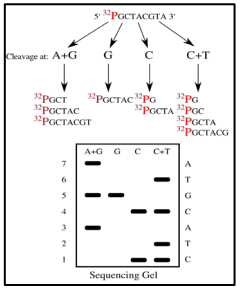
1980-1990: First Generation Sequencing

2000s: Next Generation Sequencing

1953: DNA double helix structure described

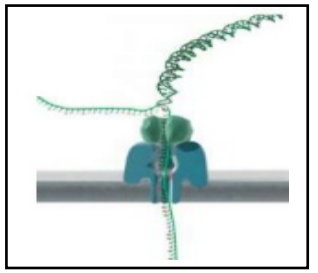


1977: Maxam-Gilbert sequencing techinque



2001: First human genome sequence

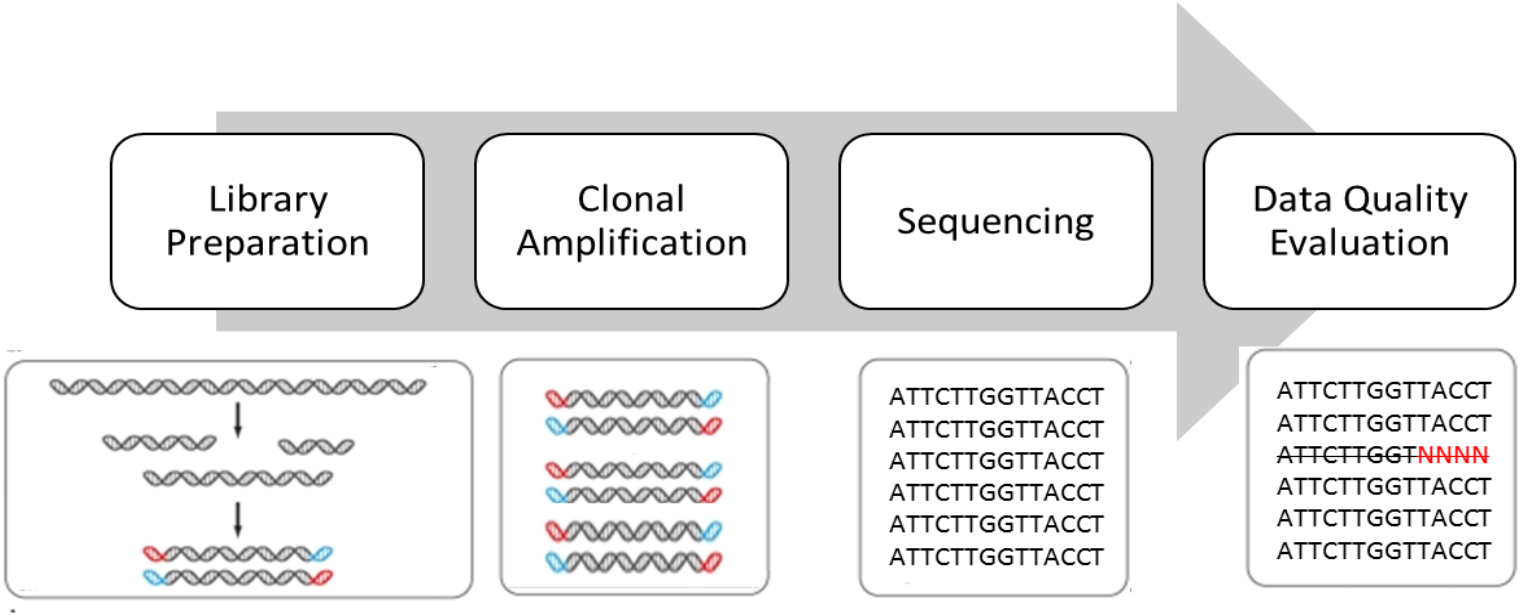
2008: Third Generation Sequencing



POLITECNICO DI TORINO

The Next Generation Sequencing

Next Generation Sequencing (NGS) technologies implement the concept of **cyclic-array sequencing**: Dense arrays of DNA are sequenced by repetitive cycles of enzymatic reaction and imaging-based data collection

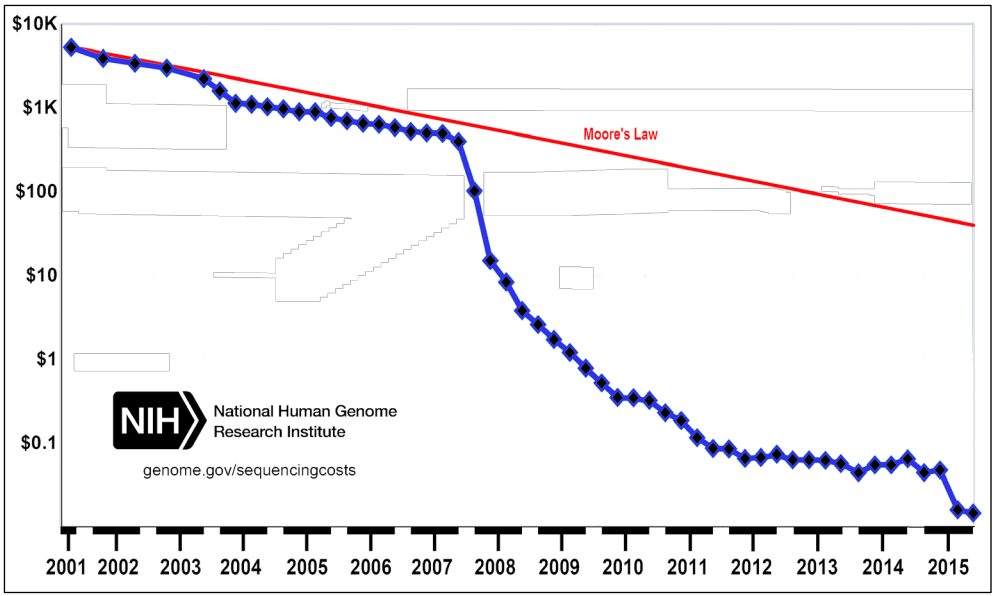


POLITECNICO DI TORINO

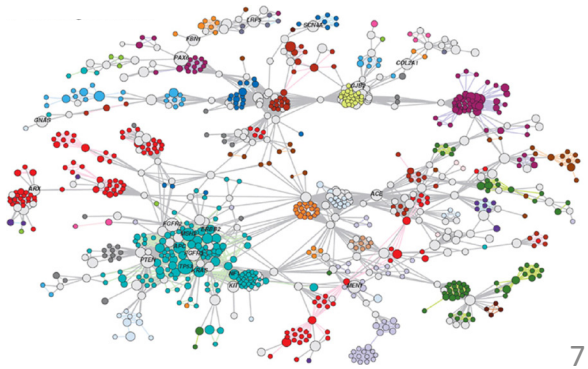
NGS data: A twofold problem

The **huge amounts** of NGS data produced worldwide need to be stored, organized and shared

Cost per Raw Megabase of DNA sequence

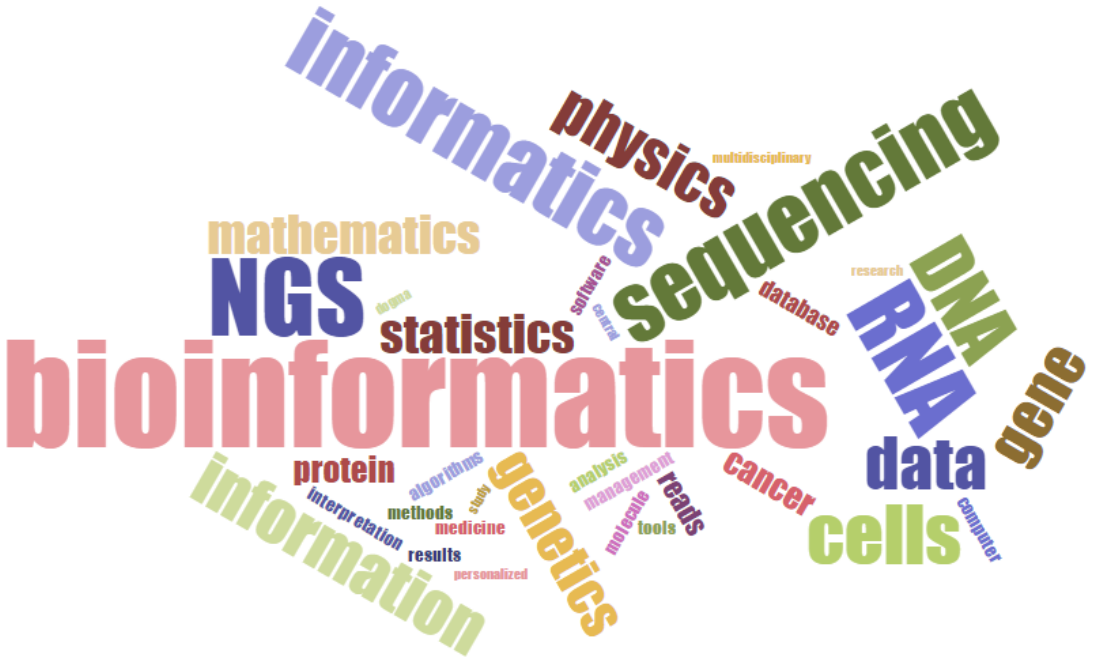


The **analysis** of NGS data deals with the **complexity** of both structure and function of biological organisms



POLITECNICO DI TORINO

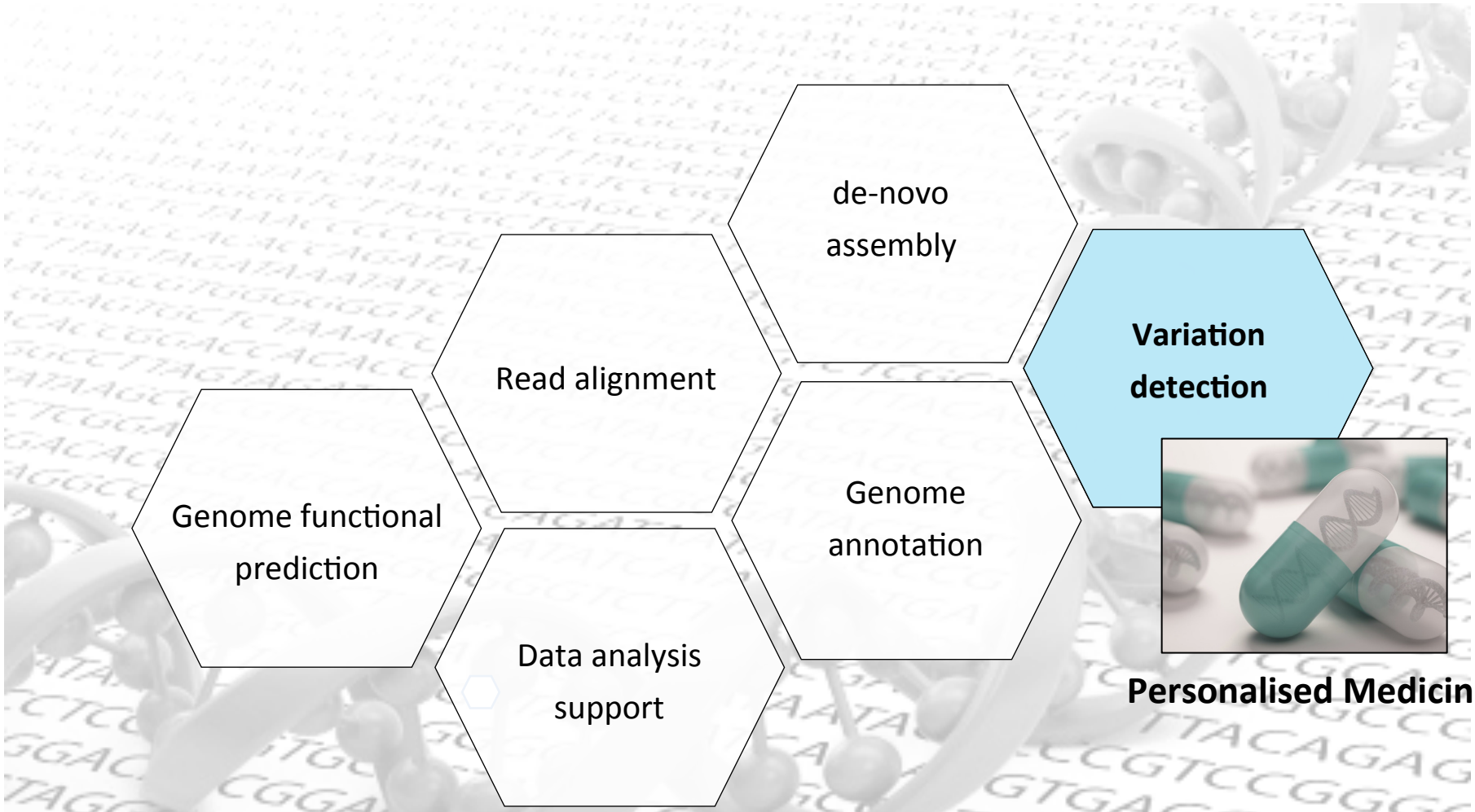
How to deal with NGS data?



The term **Bioinformatics** refers to the application of **computational techniques** to the **understanding** and **organisation** of the information associated to biological macromolecules on a large scale and comprises research fields at the interface between computer and biological sciences

POLITECNICO DI TORINO

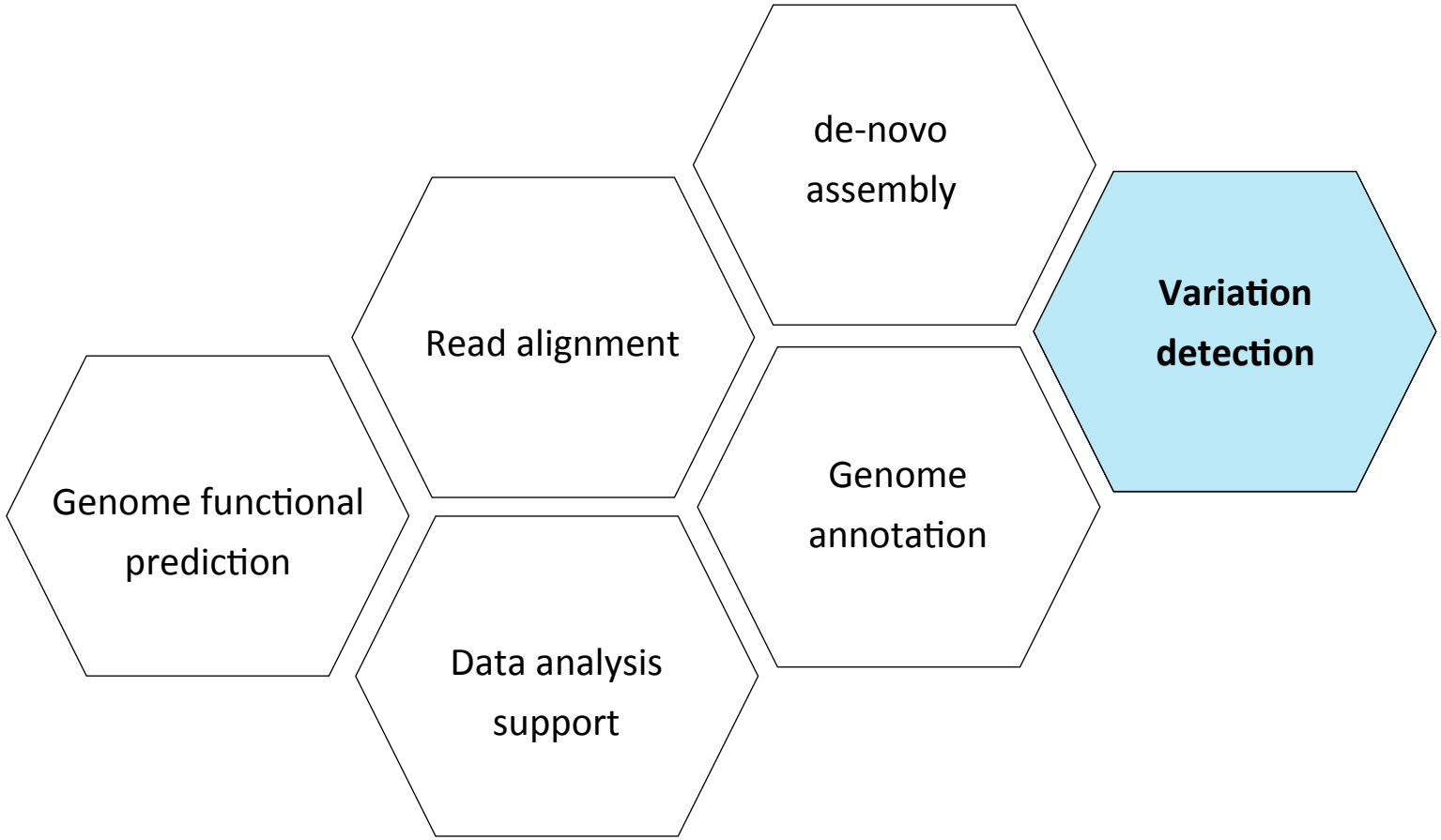
Bioinformatics revolutionised genomics



Zhang, Jun, et al. **"The impact of next-generation sequencing on genomics."** Journal of genetics and genomics 38.3 (2011): 95-109

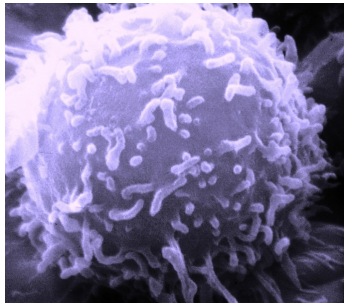
POLITECNICO DI TORINO

My contribution in this field



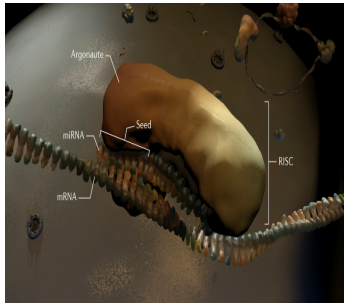
POLITECNICO DI TORINO

Which variations have I investigated?

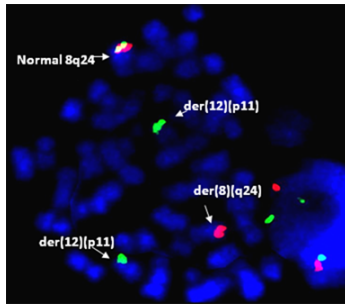


B cell clonal populations

Variation detection



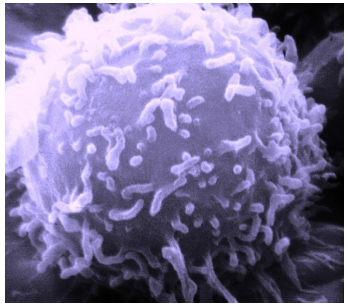
miRNA/isomiR patterns



Gene Fusions

POLITECNICO DI TORINO

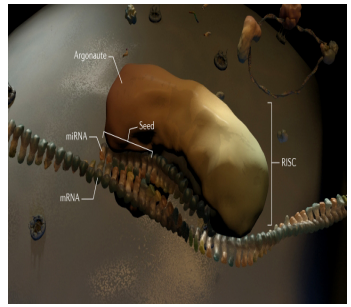
How did I investigate these variations?



VDJSeq-Solver tool

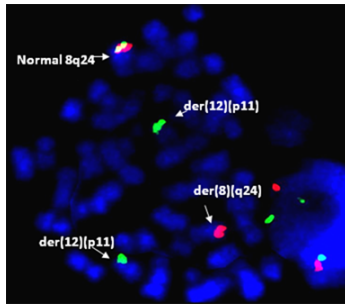
UNIVERSITÀ di VERONA

Variation detection



isomiR-SEA

ISTITUTO DI CANDIOLO-IRCCS



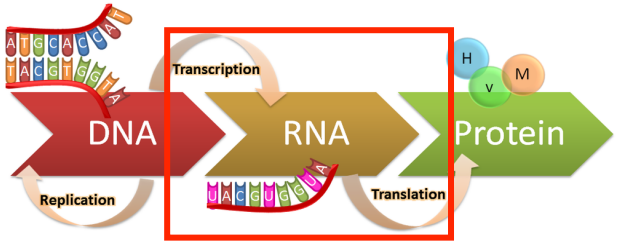
FuGePrior tool

SEVENTH FRAMEWORK PROGRAMME

NEXT GENERATION SEQUENCING for Targeted Personalized Therapy of Leukemia

POLITECNICO DI TORINO

Which features do these tools have in ~~common?~~



- ✓ **RNA-Seq** data is used as input
- ✓ Propose novel **in silico** approaches for **disease** characterization
- ✓ Designed considering latest **biological knowledge**
- ✓ Built-on-top of state-of-the-art **bioinformatics methods** and tools
- ✓ Results **confirm** previous studies or **introduce** novel genomic knowledge

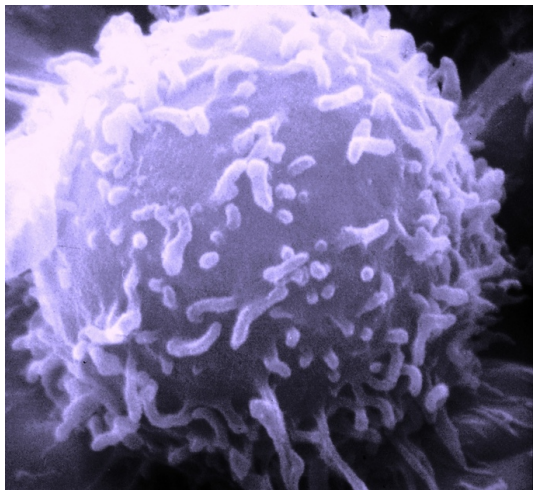
POLITECNICO DI TORINO

VDJSeq-Solver

POLITECNICO DI TORINO

Biological Background

The B cells



- ✓ One of the major effector molecules involved in **Adaptive Immune Responses**
- ✓ Originate in the **bone marrow** and undergo maturation in the **lymphoid tissues**
- ✓ Produce **antibodies**
- ✓ More than $\mathbf{10^{16}}$ different types of **B cells** ensure protection against infectious diseases

POLITECNICO DI TORINO

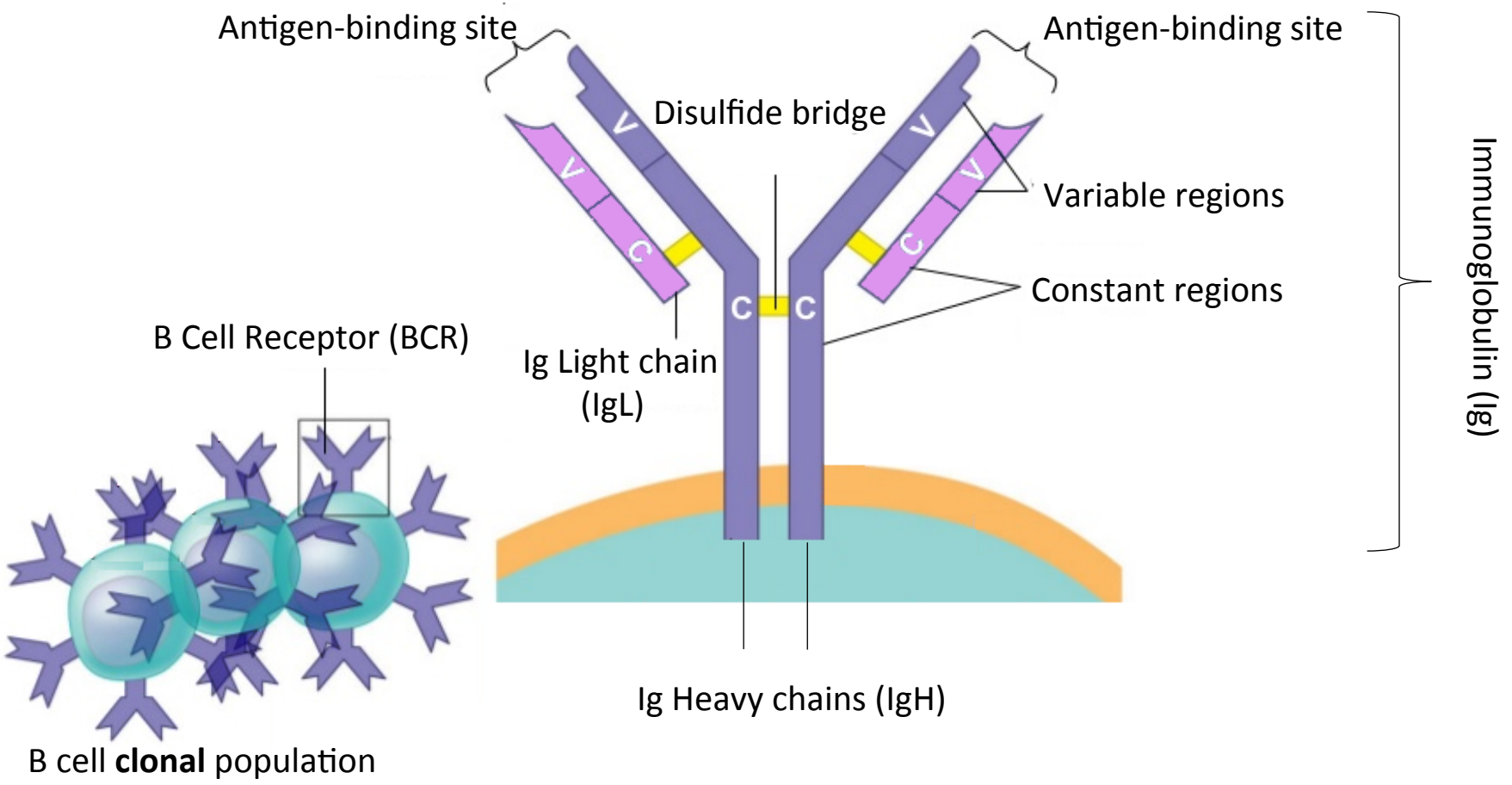
Biological Background

How do B cells work?

POLITECNICO DI TORINO

Biological Background

How can B cells recognize different ~~antigens?~~



Different B Cell Receptors bind different antigens

POLITECNICO DI TORINO

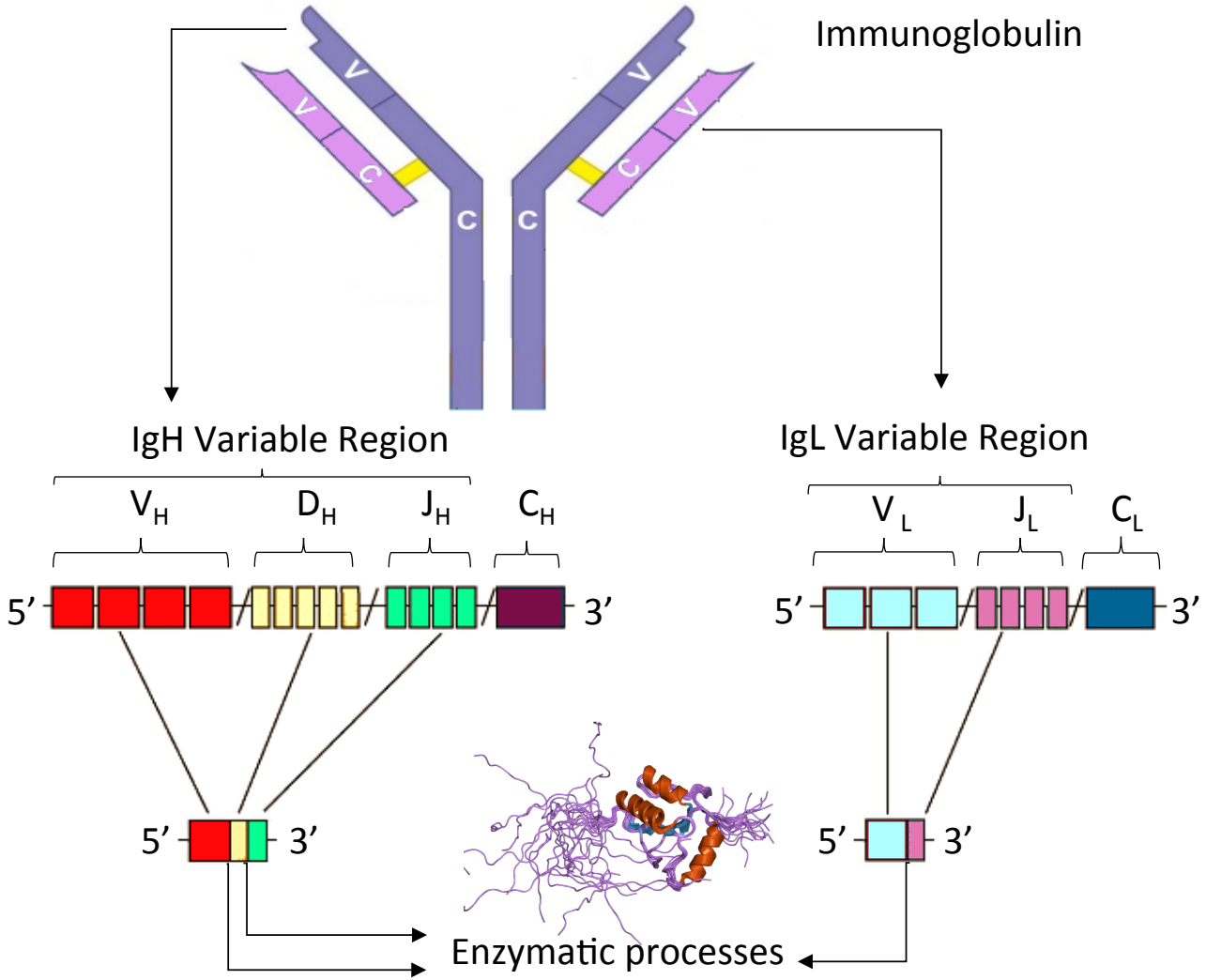
Biological Background

V(D)J Recombination accounts for BCR ~~diversity~~

Protein Level

DNA Level

RNA Level

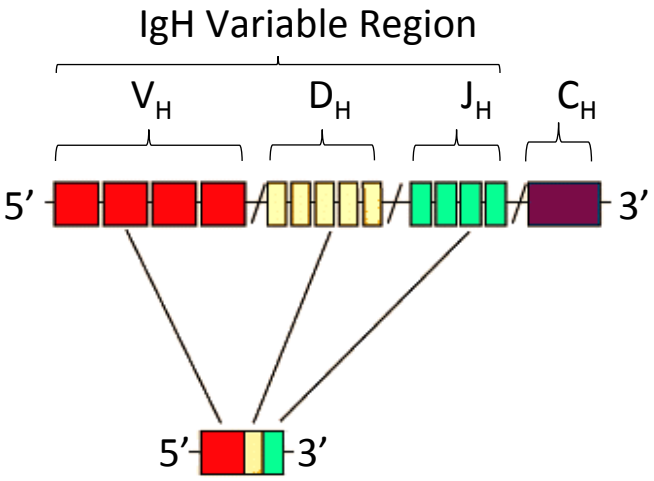


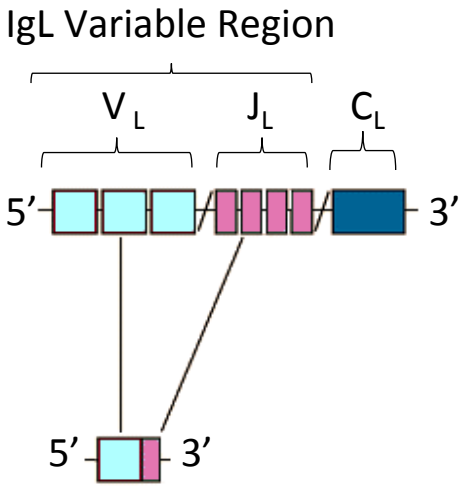
POLITECNICO DI TORINO

Motivation

Why is B cell clone identification ~~important~~?

- ✓ Several B cell diseases are characterized by **massive expansion** of a single B cell clone
- ✓ B cell clonal expansion is generally interpreted by clinicians as **evidence of cancer**
- ✓ There are **biases** in IgH and IgL **recombined genes** in different pathologies and **correlations** between the clinical course of the disease and Ig gene usage



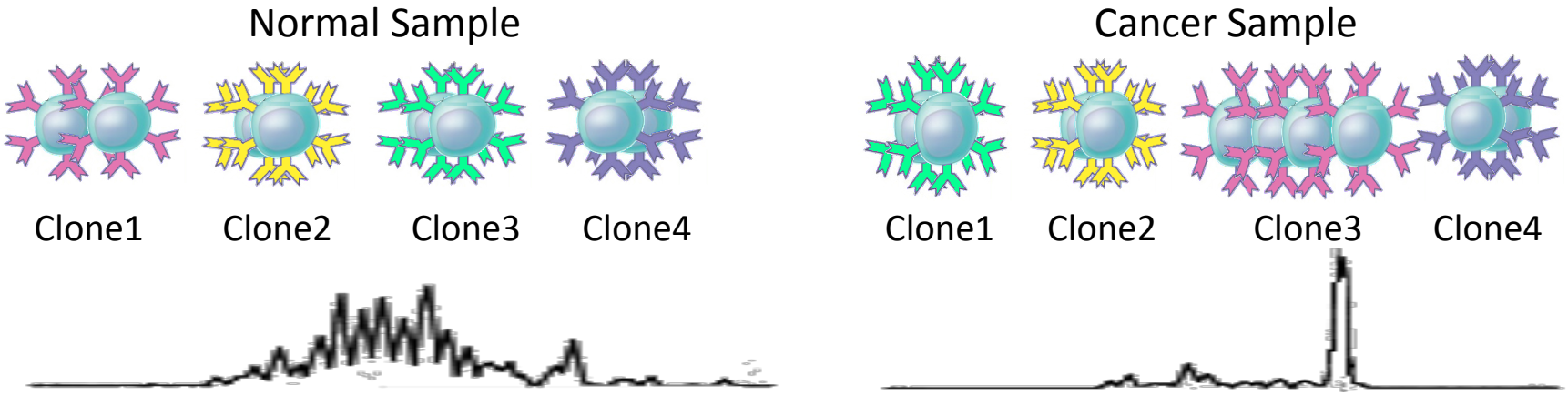


Capello, D., et al. **"Evidence of biased immunoglobulin variable gene usage in highly stable B-cell chronic lymphocytic leukemia."** Leukemia 18.12 (2004): 1941-1947

POLITECNICO DI TORINO

Motivation

B cell clone identification approaches

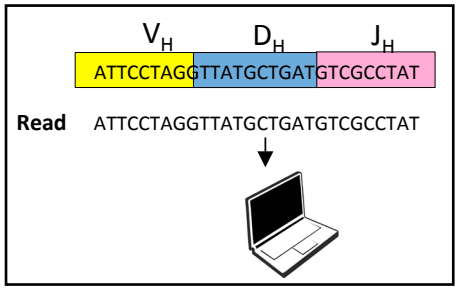


PCR Clonality tests

✓ **Limited sensitivity** associated to the normal polyclonal background

NGS analyses

✓ Rely on the usage of **Ig amplified** DNA/RNA molecules

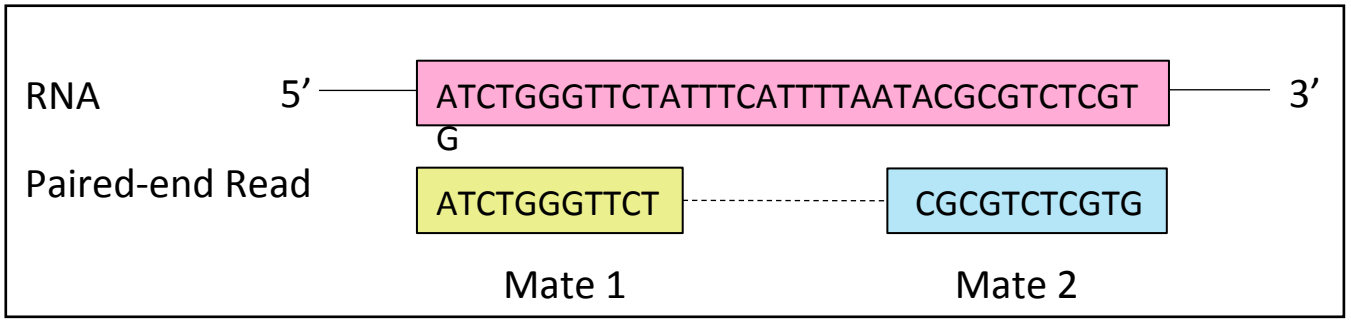


Gazzola, Anna, et al. **"The evolution of clonality testing in the diagnosis and monitoring of hematological malignancies."** Therapeutic advances in hematology 5.2 (2014): 35-47.

POLITECNICO DI TORINO

Motivation

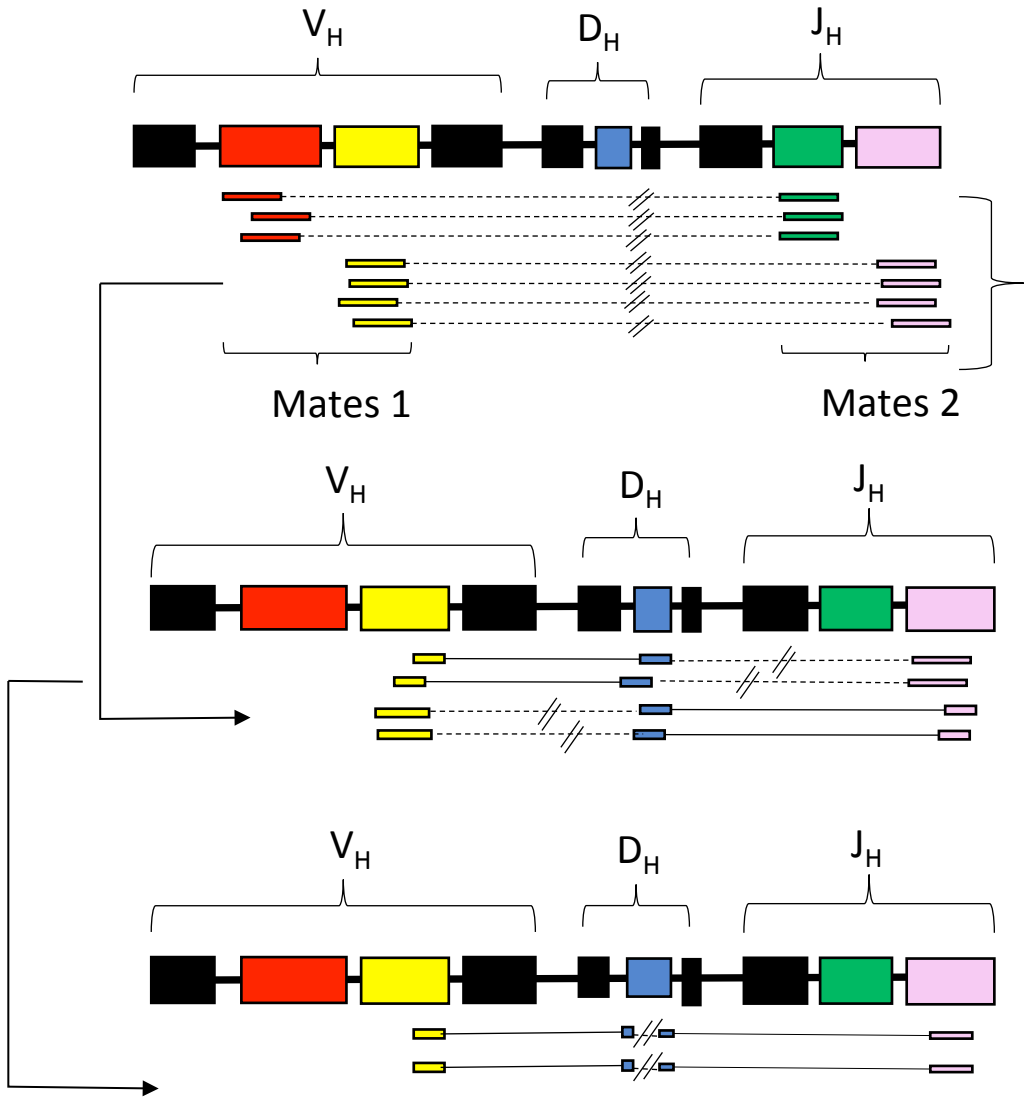
VDJSeq-Solver



POLITECNICO DI TORINO

Algorithm

B cell clone identification



VJ Encompassing reads allow to identify VJ clones in both IgHs and IgLs

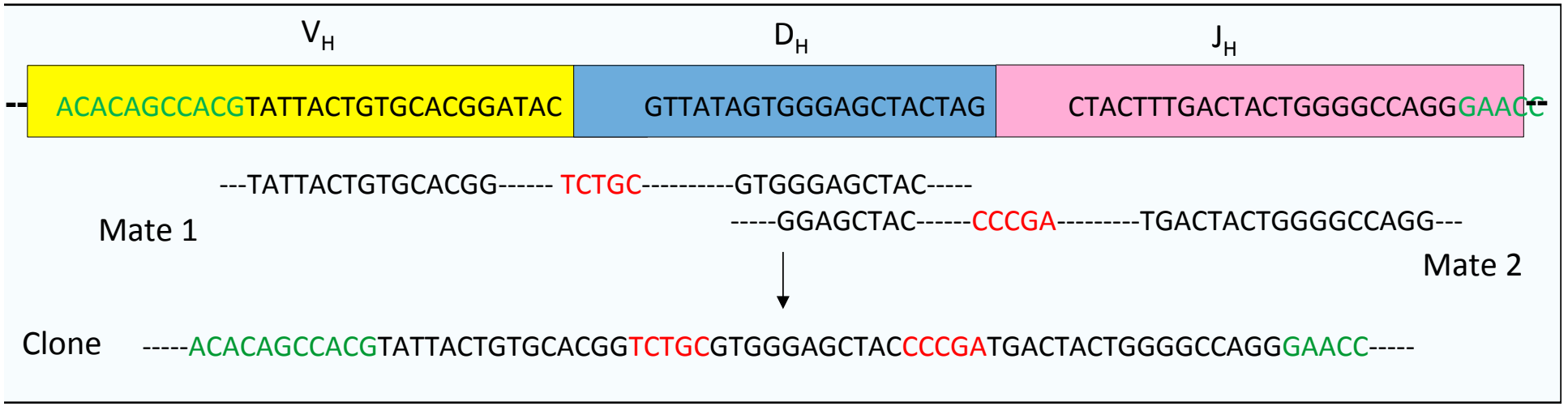
VJ Encompassing reads having **one or both** mates partially mapped onto a D gene identify IgH V(D)J clones

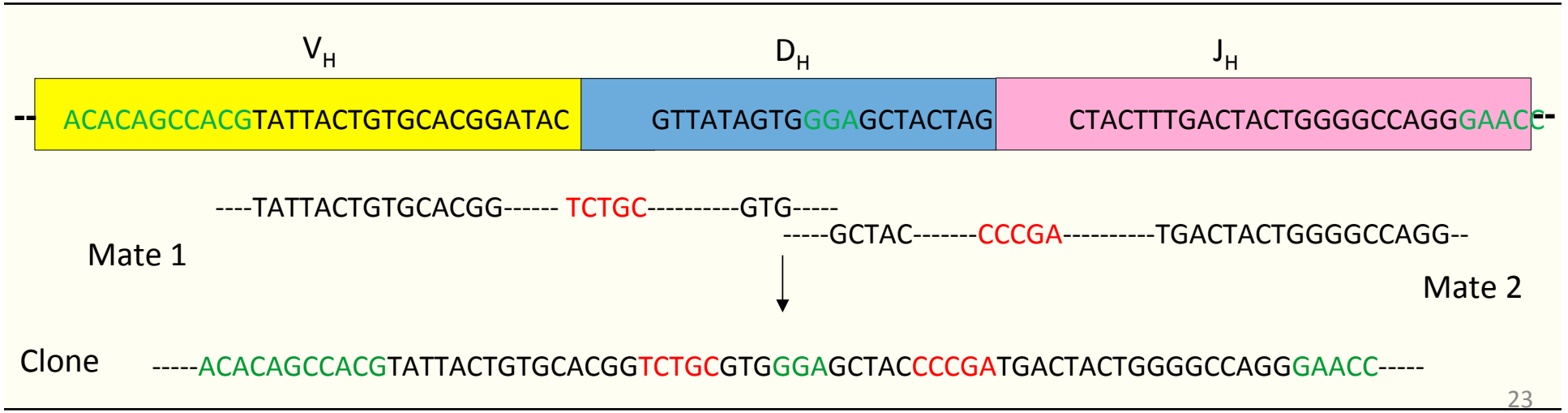
VJ Encompassing reads having **both** mates mapped onto a D gene allow to reconstruct IgH V(D)J clone sequences

POLITECNICO DI TORINO

Algorithm

B cell clone sequence reconstruction

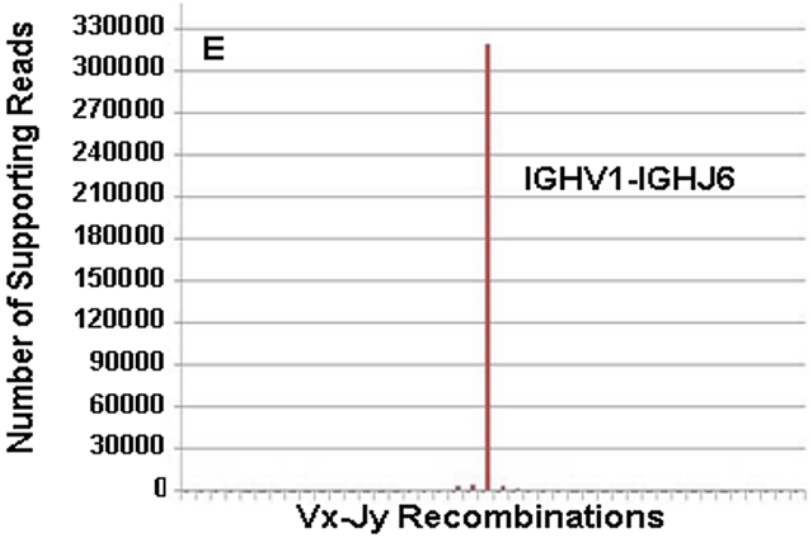




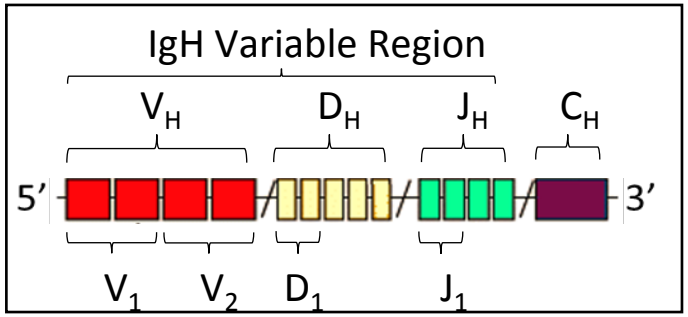
POLITECNICO DI TORINO

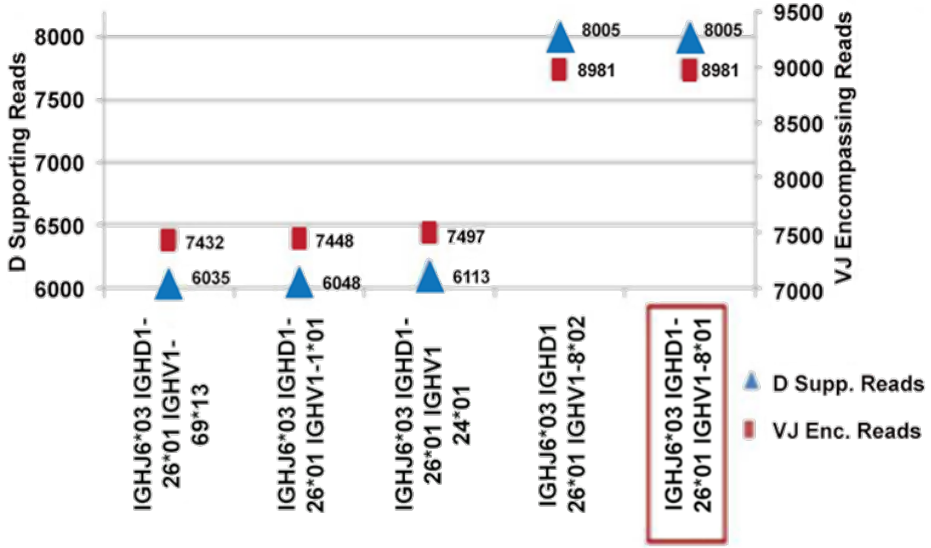
Results

IgH main clone identification in MCL samples



#Reads: 44M





VDJSeq-Solver correctly **identified IgH main clone** in MCL and DLBCL samples

POLITECNICO DI TORINO

Results

IgH main clone sequence in MCL samples

VDJSeq-Solver IgH main clone sequences are **highly similar** to those obtained via wet-lab experiments, maximum 1.56% error

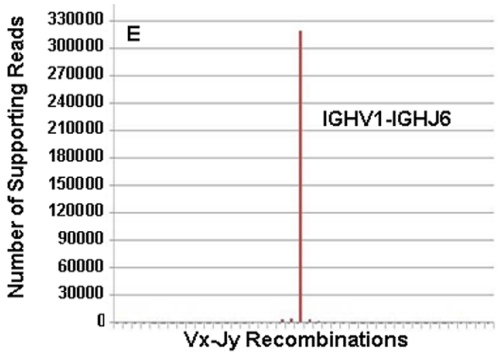
The neglectable impact of the identified mismatches is confirmed by the analysis performed adopting **state-of-the-art tools** for Ig gene assignment

Tool	V_H		D_H		J_H	
	Sanger	**VDJSeq-Solver**	**Sanger**	**VDJSeq-Solver**	**Sanger**	**VDJSeq-Solver**
IMGT/V-QUEST	IGHV1-8*01	IGHV1-8*01	IGHD1-26*01	IGHD1-26*01	IGHJ6*03	IGHJ6*03
JOINSOLVER	IGHV1-8*01	IGHV1-8*01	IGHD1-26*01	IGHD1-26*01	IGHJ6*03	IGHJ6*03
VDJsolver	IGHV1-8*01	IGHV1-8*01	IGHD1-26*01	IGHD1-26*01	IGHJ6*03	IGHJ6*03
SoDA	IGHV1-8*01	IGHV1-8*01	IGHD1-26*01	IGHD1-26*01	IGHJ6*03	IGHJ6*03
iHMMune-align	IGHV1-8*01	IGHV1-8*01	IGHD1-26*01	IGHD1-26*01	IGHJ6*03	IGHJ6*03

POLITECNICO DI TORINO

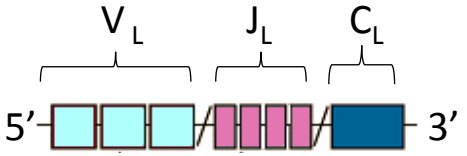
Remarks and Perspectives

To summarise



VDJSeq-Solver **identified** the IgH main clonal population in MCL and DLBCL samples and **reconstructed** its sequence in MCL samples

Preliminary results on **IgLs** in MCL samples are promising



Future work:

- T Cell Receptor (TCR) analysis
- Clone Collapsing
- Parallelisation

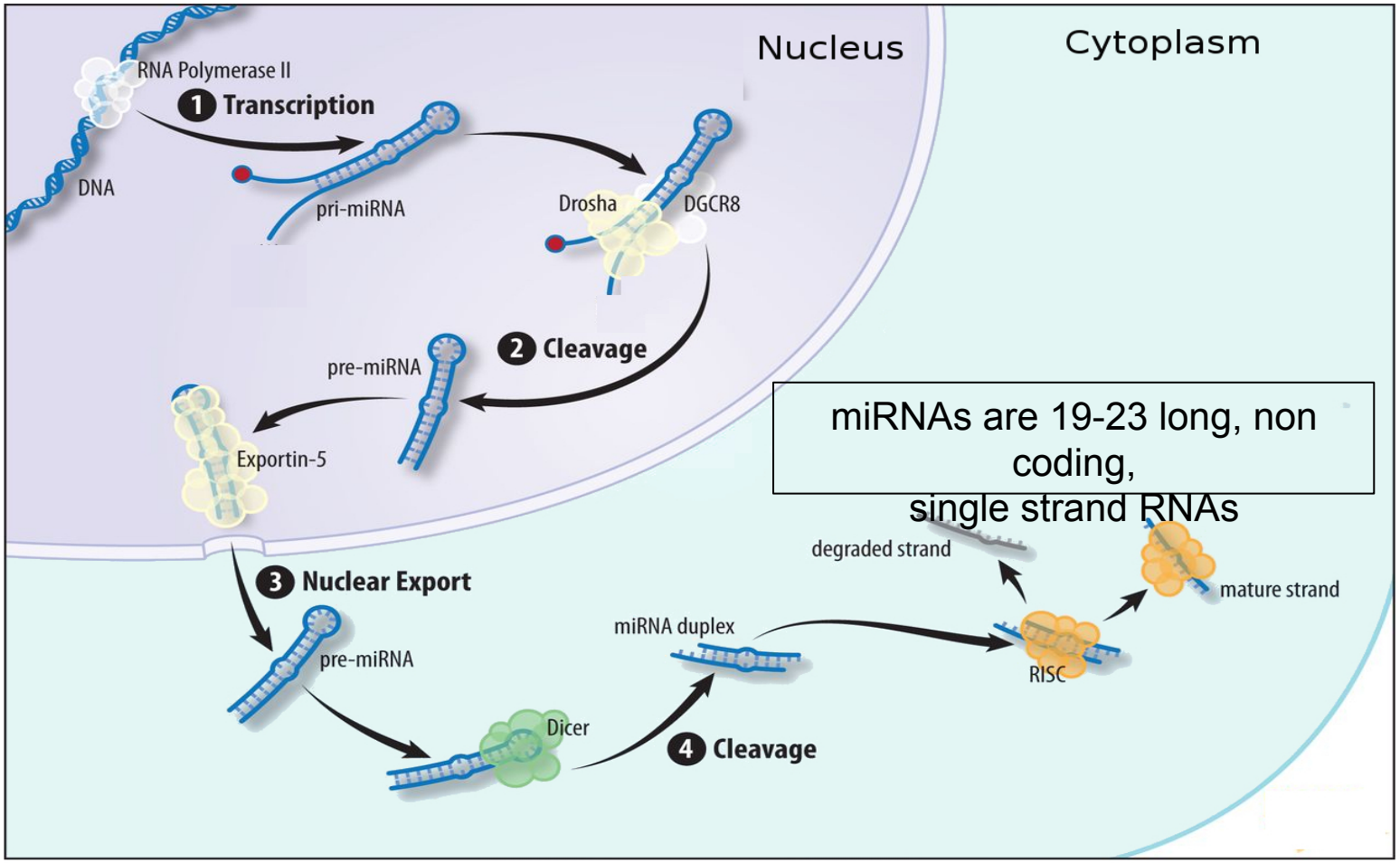
POLITECNICO DI TORINO

isomiR-SEA

POLITECNICO DI TORINO

Biological Background

The miRNAs

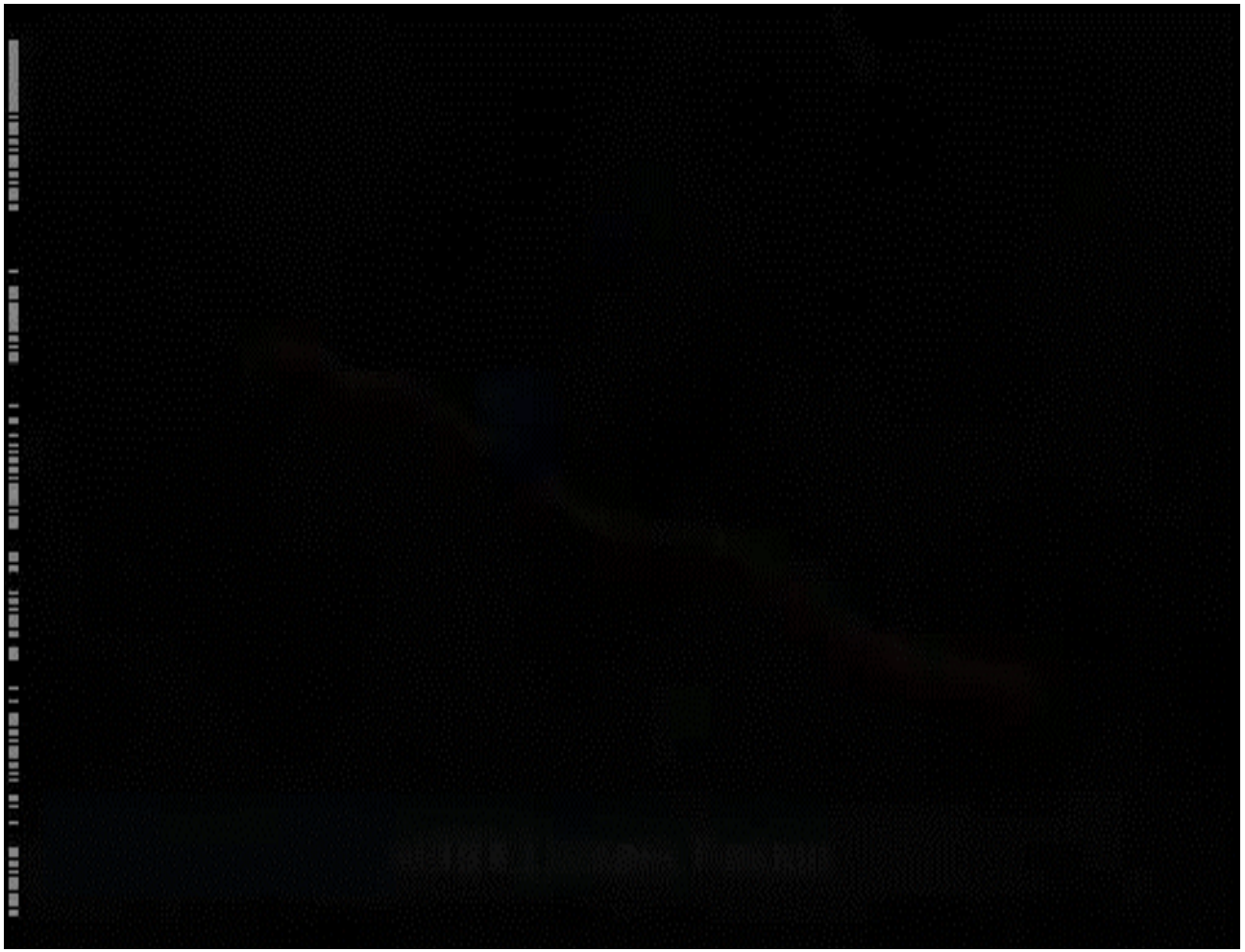


miRNAs are 19-23 long, non coding, single strand RNAs

POLITECNICO DI TORINO

Biological Background

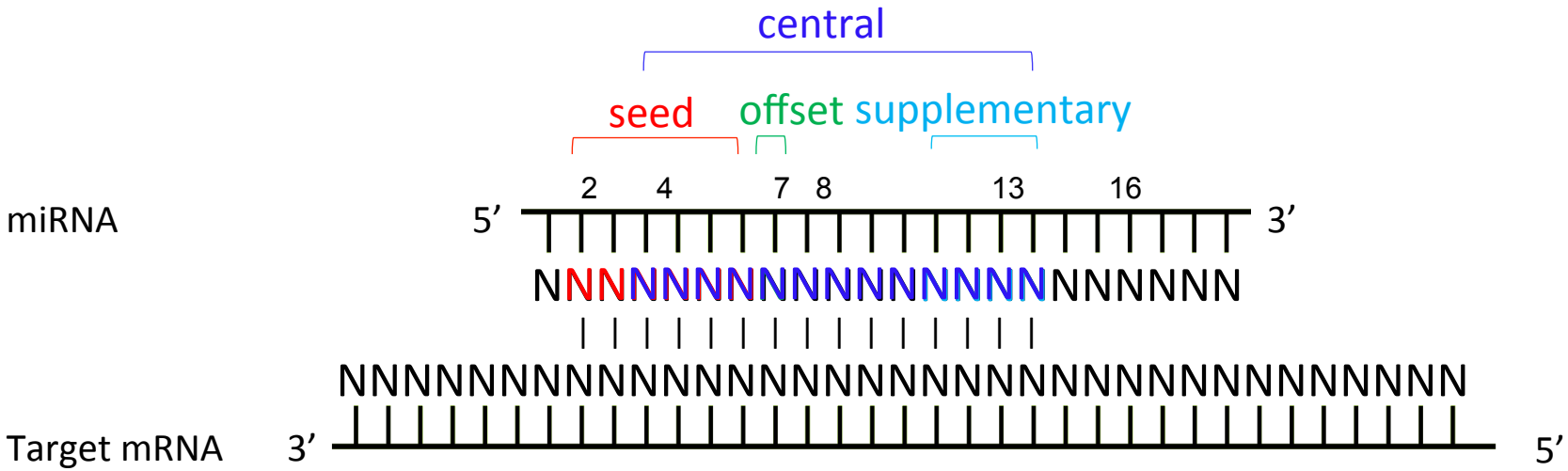
How do miRNAs work?



POLITECNICO DI TORINO

Biological Background

miRNA:mRNA interaction



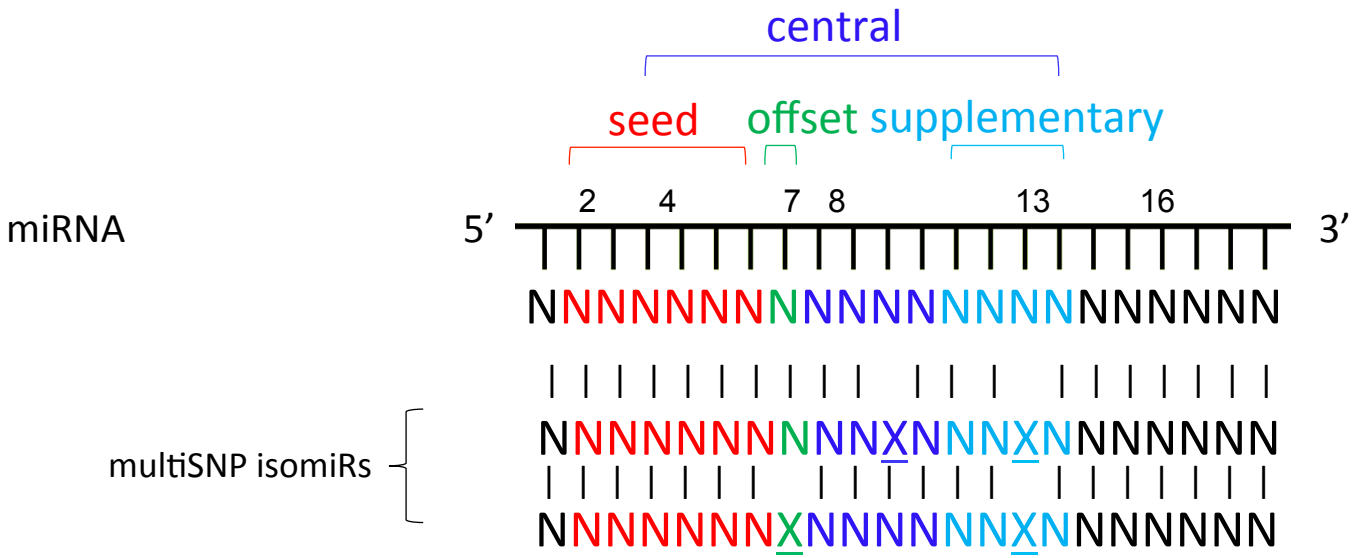
miRNAs belonging to the same **family** have identical **seed** sequences

Bartel, David P. **"MicroRNAs: target recognition and regulatory functions."** Cell 136.2 (2009): 215-233.

POLITECNICO DI TORINO

Biological Background

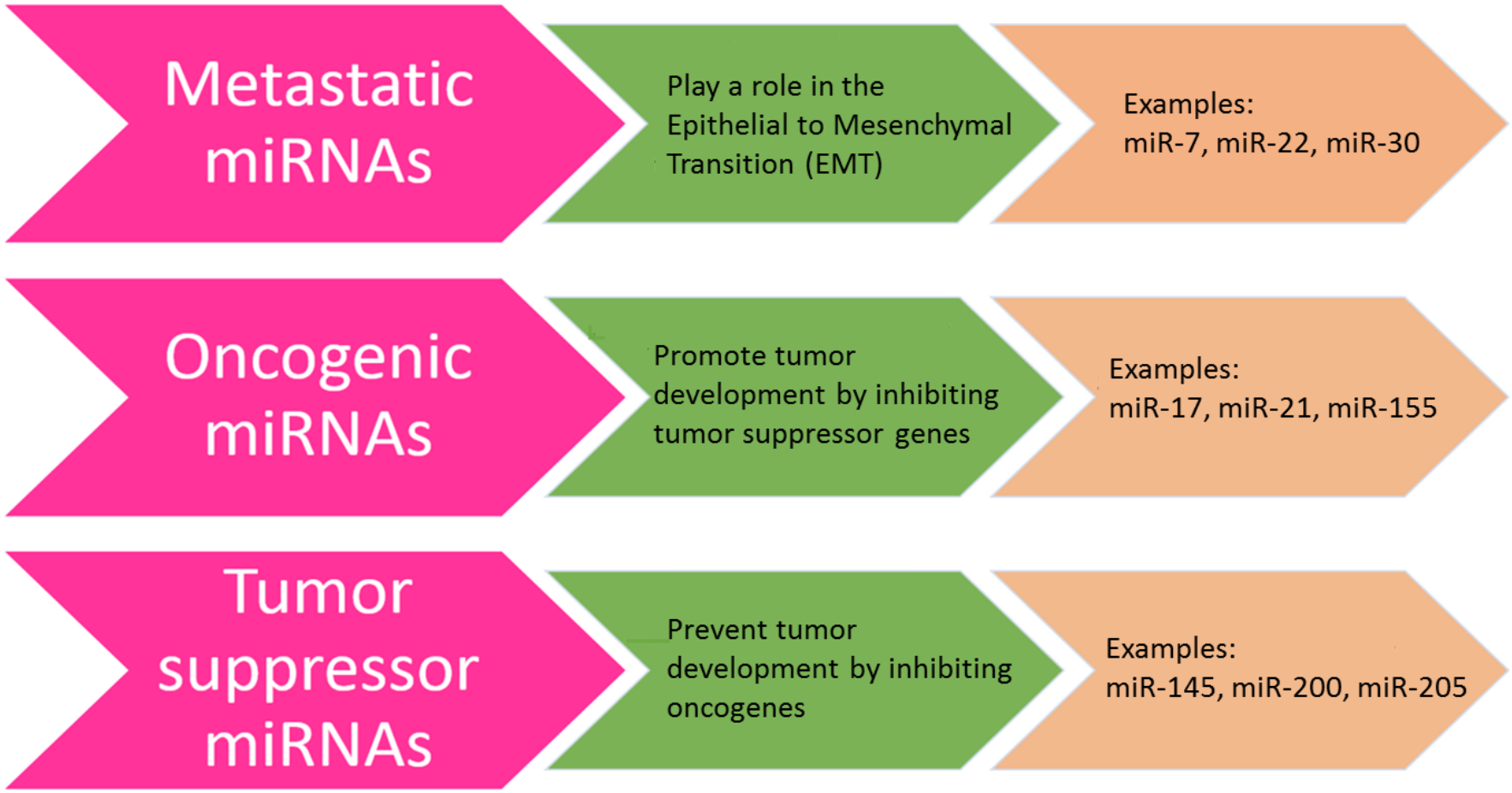
miRNA variants: The isomiRs



POLITECNICO DI TORINO

Motivation

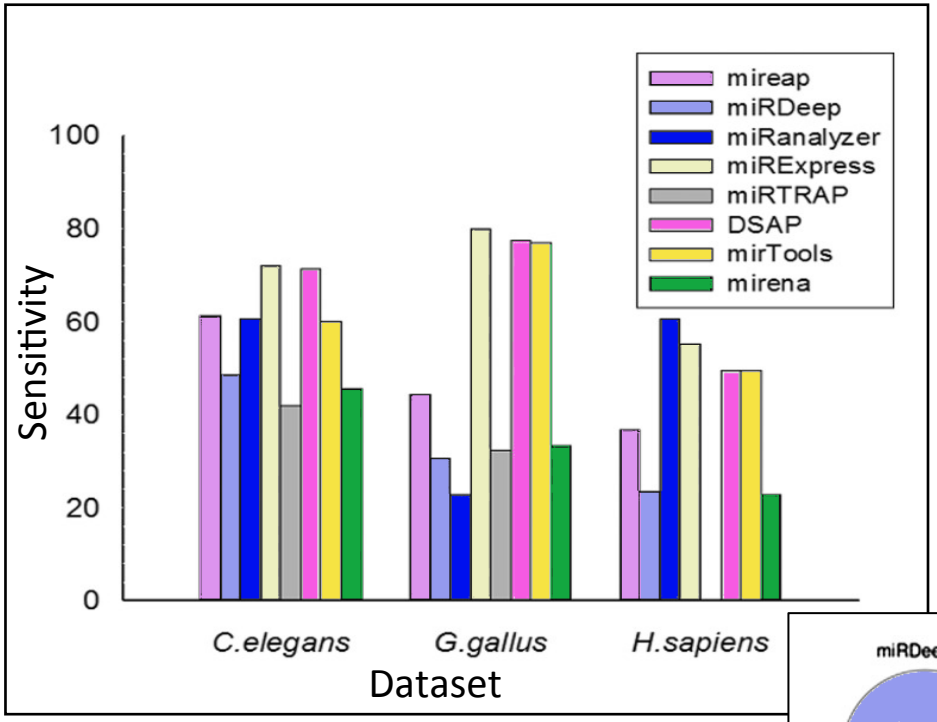
Why is miRNA/isomiR identification ~~important~~?



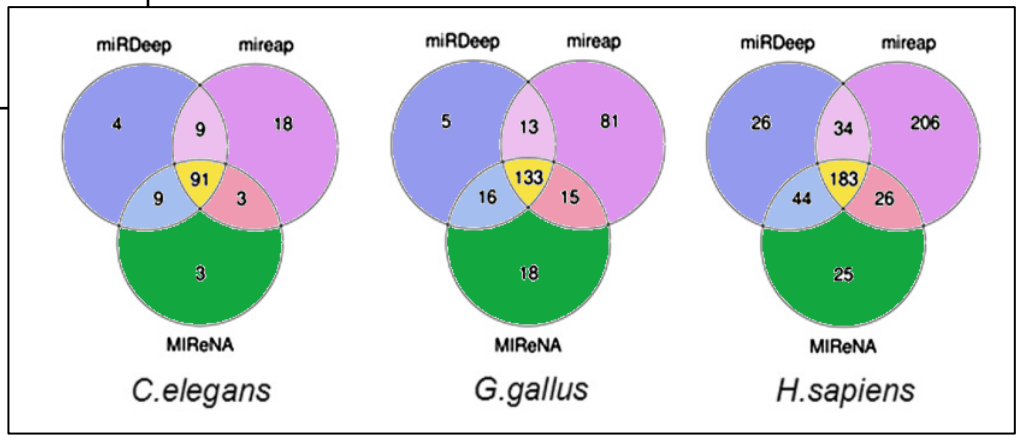
POLITECNICO DI TORINO

Motivation

State-of-the-art miRNA detection tools



Li, Yue, et al. **"Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis."** Nucleic Acids Research (2012): gks043



POLITECNICO DI TORINO

Motivation

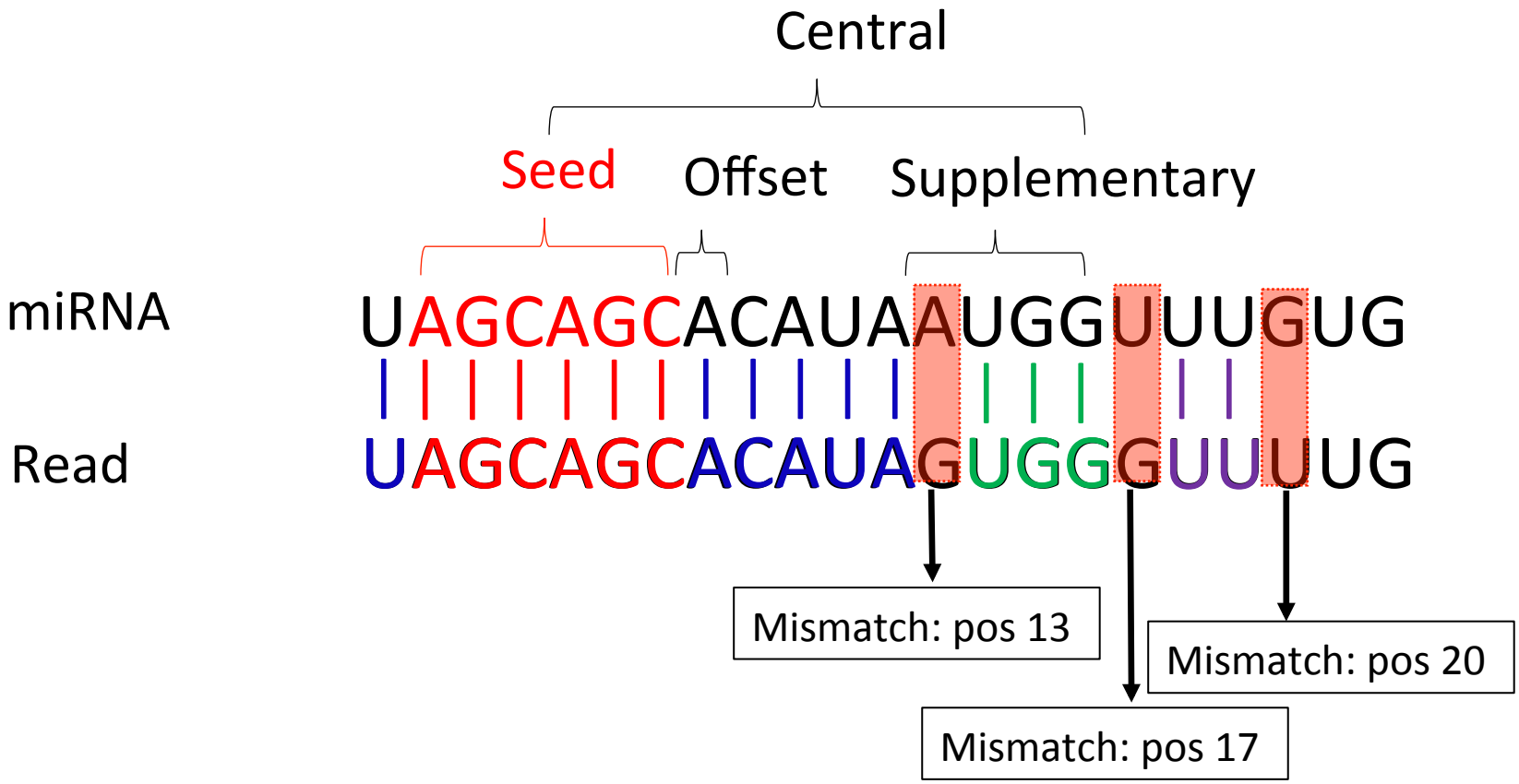
isomiR-SEA

- ✓ Implements a **miRNA-specific** alignment procedure
- ✓ Identifies **miRNA**, **isomiR** and **interaction site** expression profiles
- ✓ Uses **Seqan** library
- ✓ **Freely available** → http://eda.polito.it/isomir-sea/
- ✓ Users can easily **configure** the run

POLITECNICO DI TORINO

Algorithm

The alignment procedure

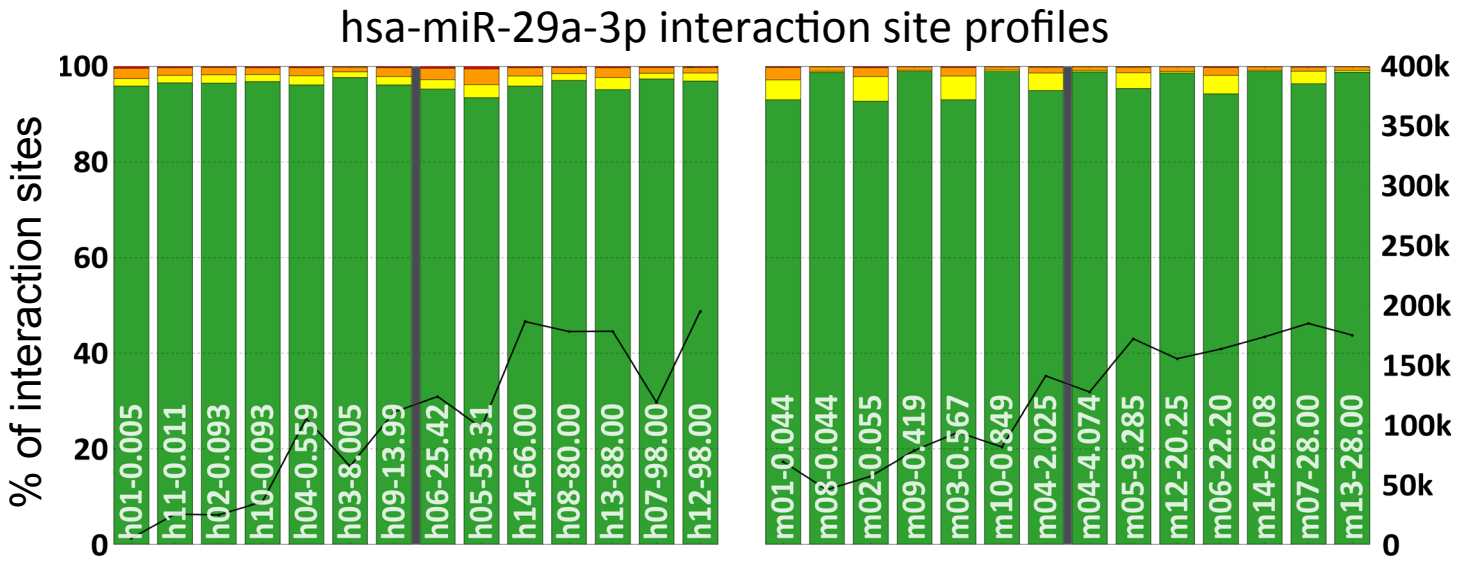


By evaluating the **positions** of the encountered **mismatches**, isomiR-SEA can distinguish among miRNAs/isomiRs and assess the conservation of miRNA interaction sites

POLITECNICO DI TORINO

Results

miRNA/isomiR profiles in Somel dataset



Avg#Reads: 5M

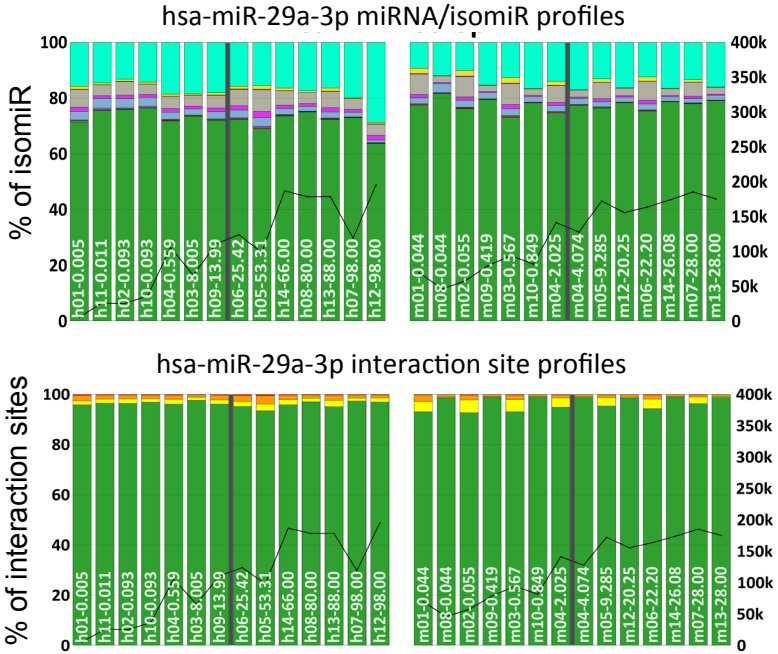
		Conserved nt		
Col.	Interaction site Type	nt 8	nt 9-11	nt 12-16
	offset-suppl-central	X	X	X
	offset-suppl	X		X
	offset-only	X		
	suppl-only			X

Somel, Mehmet, et al. **"MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain."** Genome Research 20.9 (2010): 1207-1218

POLITECNICO DI TORINO

Remarks and Perspectives

To summarise



isomiR-SEA identified **miRNA/isomiR** and **conserved interaction site** expression levels in normal and cancer samples

isomiR-SEA results allow to explain miRNA/isomiR **targeting activity**

Future work:

- pre-miRNA evaluation
- read quality value evaluation
- long non-coding RNA identification

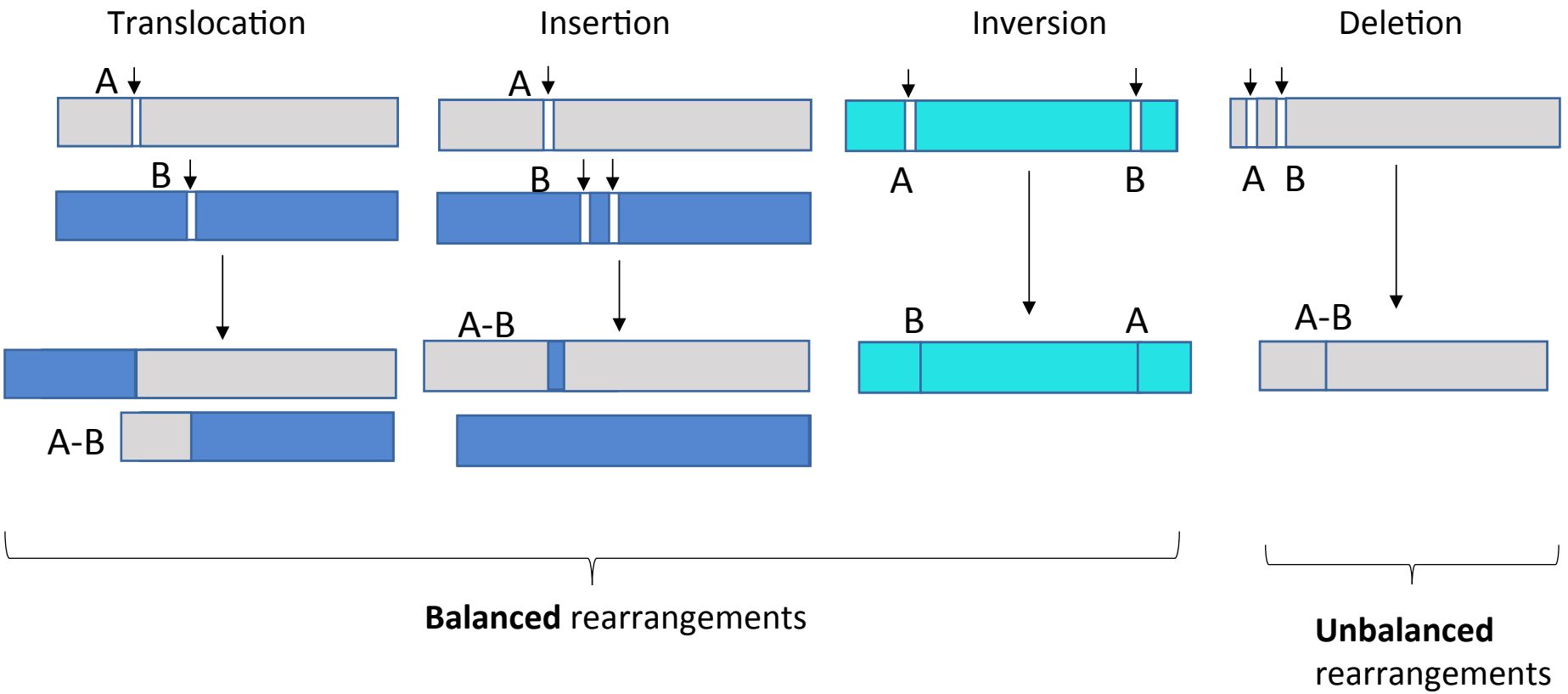
POLITECNICO DI TORINO

FuGePrior

POLITECNICO DI TORINO

Biological Background

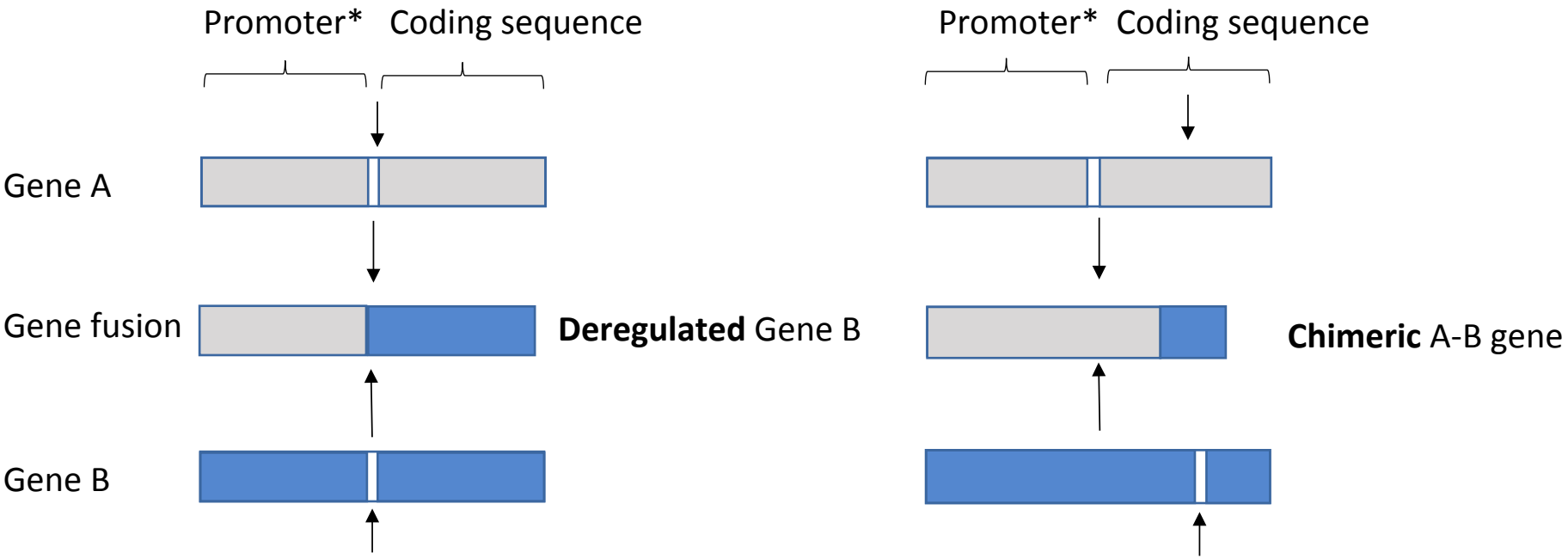
The gene fusions



POLITECNICO DI TORINO

Biological Background

Which is the effect of gene fusions?

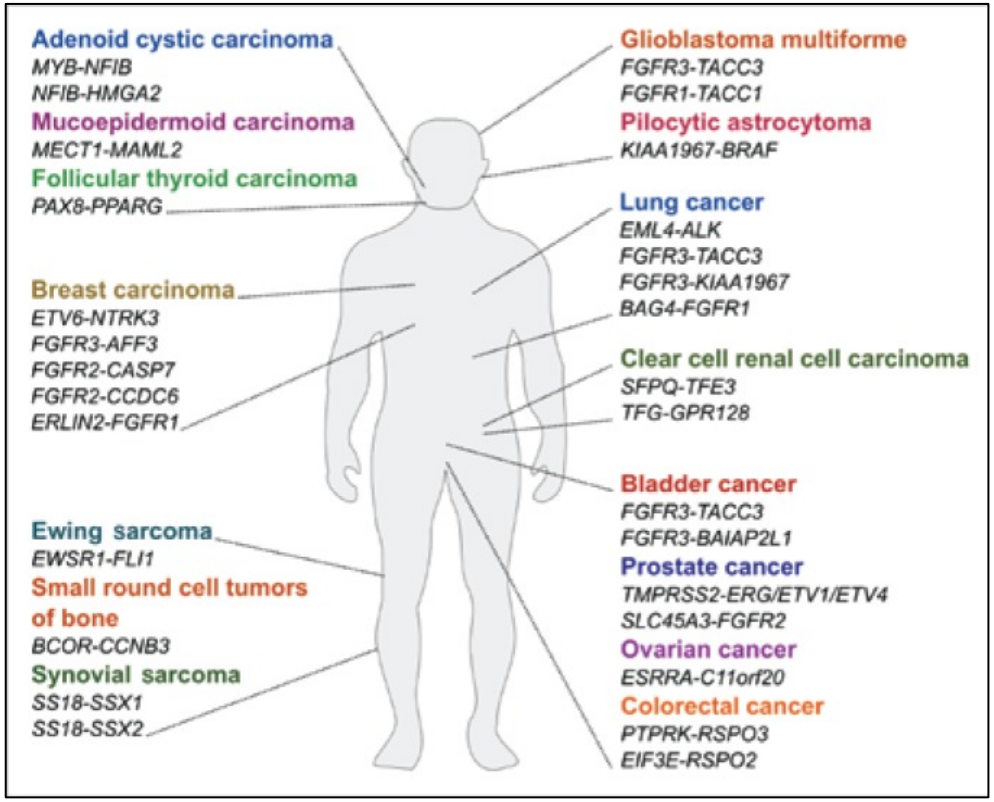


***Promoter:** a region of DNA that initiates transcription of a particular gene

POLITECNICO DI TORINO

Motivation

Why is gene fusion identification ~~important~~?

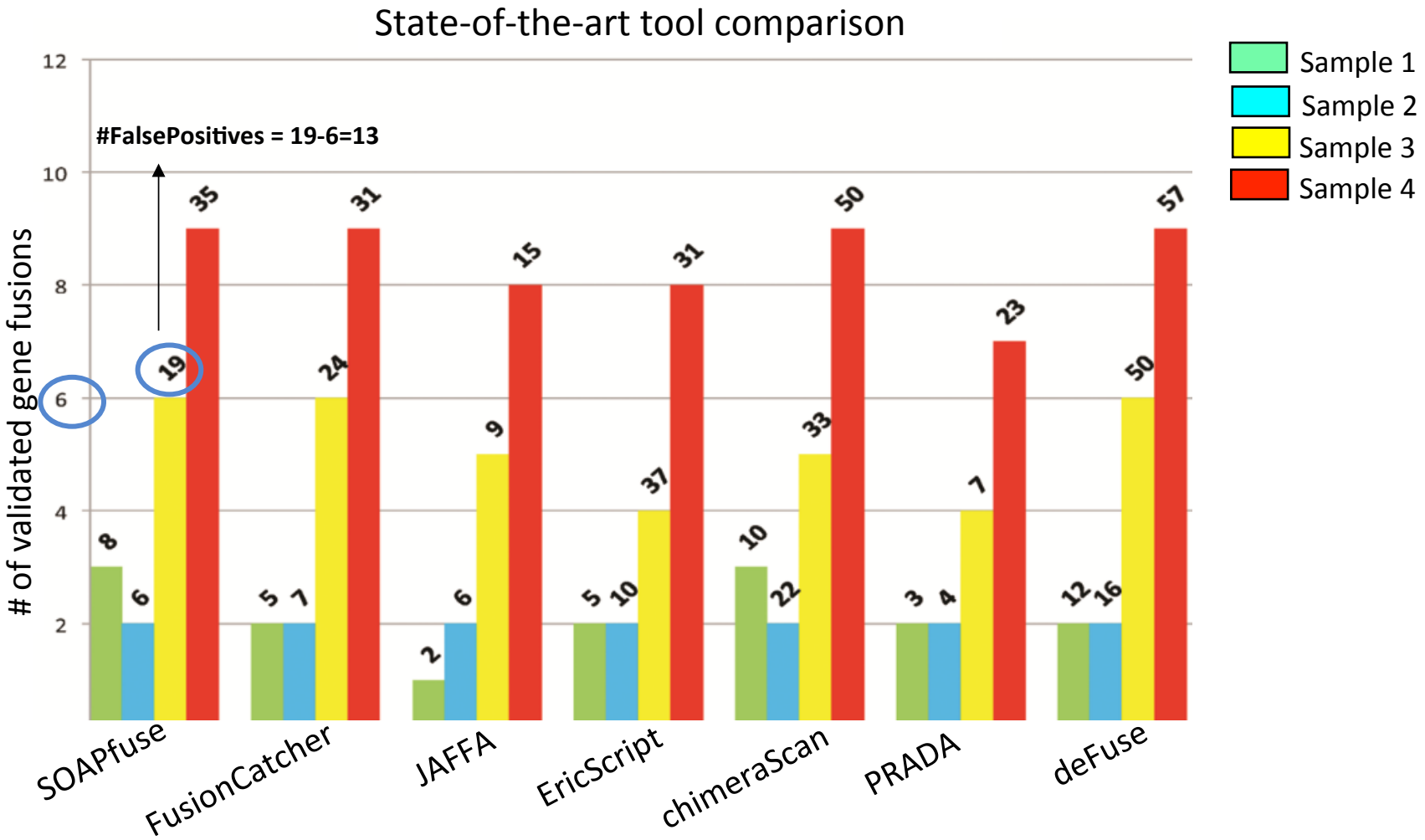


Gene fusions are **cancer-specific** and their identification is considered fundamental for the **diagnosis** and **prognosis** of several cancers

POLITECNICO DI TORINO

Motivation

State-of-the-art gene fusion detection tools



Kumar, Shailesh, et al. **"Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data."** *Scientific Reports* 6 (2016)

POLITECNICO DI TORINO

Motivation

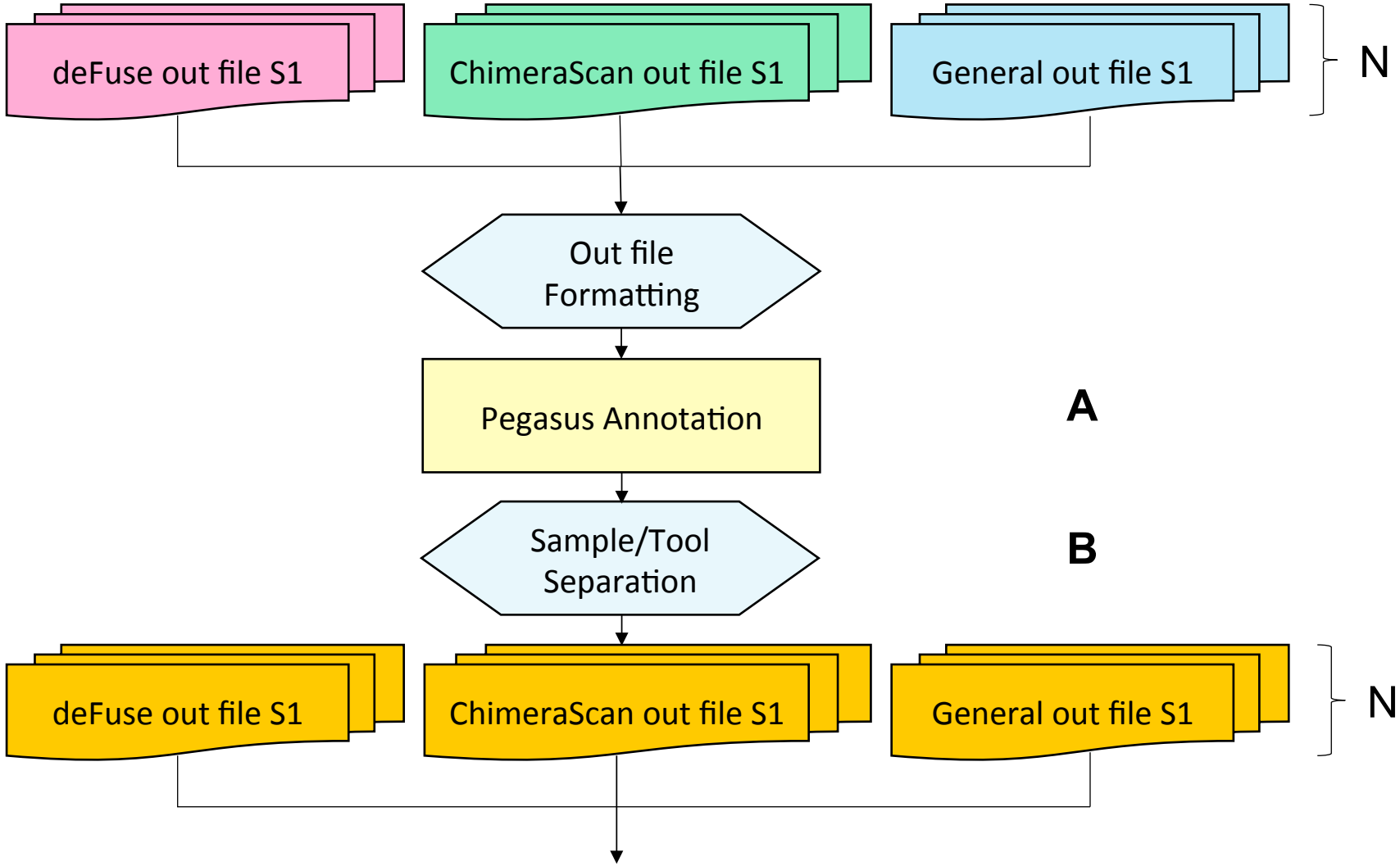
FuGePrior

- ✓ **Prioritizes** fusions from gene fusion discovery tools
- ✓ Implements a set of **filtering** and **processing** stages
- ✓ Integrates the driver scores from 2 **Machine Learning** algorithms
- ✓ **Freely available** → https://philae.polito.it/paciello/FuGePrior/
- ✓ Users can easily **configure** the run

POLITECNICO DI TORINO

Algorithm

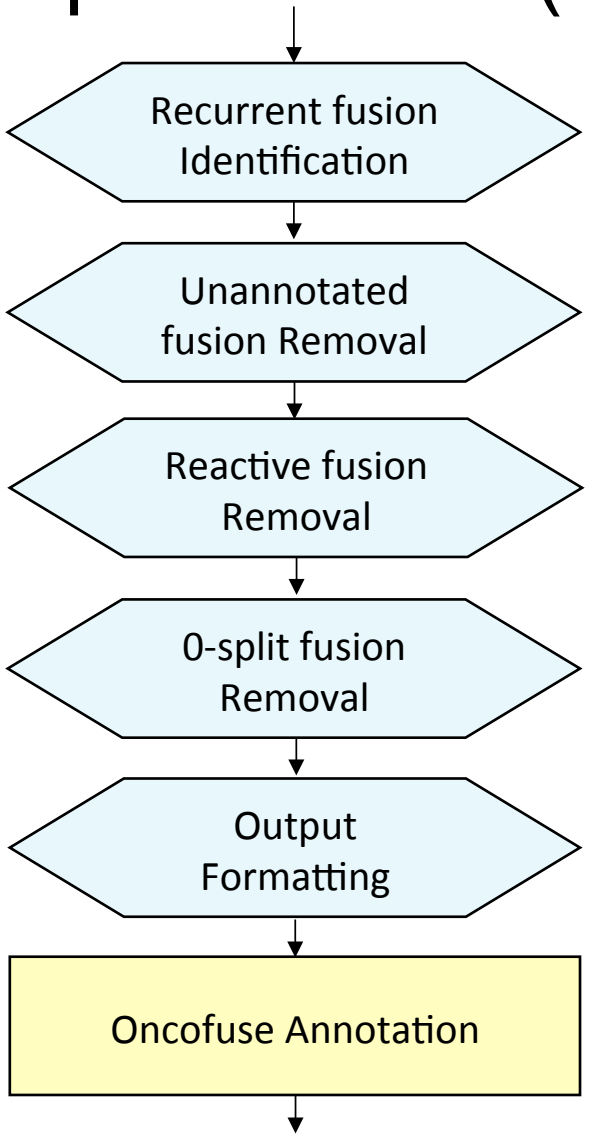
How are fusions prioritized? (1)



POLITECNICO DI TORINO

Algorithm

How are fusions prioritized? (2)



C

D

E

F

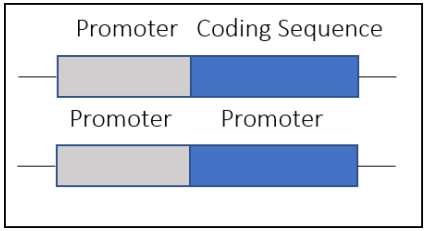
G

H

POLITECNICO DI TORINO

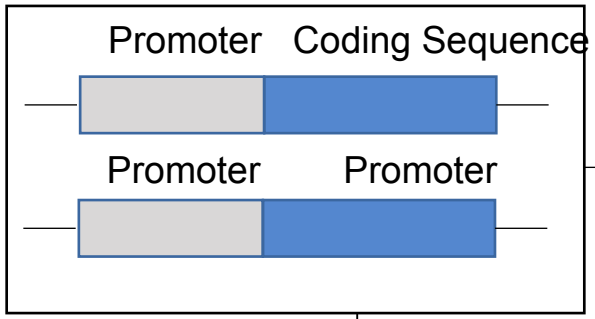
Algorithm

How are fusions prioritized? (3)

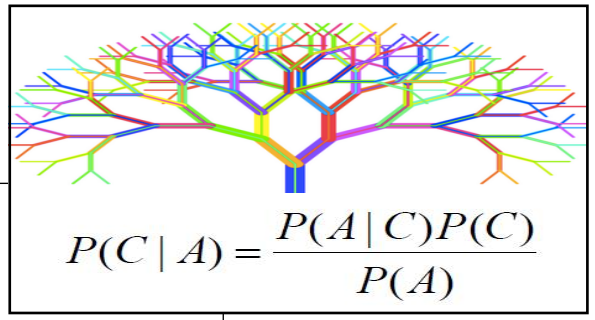


Fusion mechanism Analysis

I



Fusion Selection



L

Output file S1

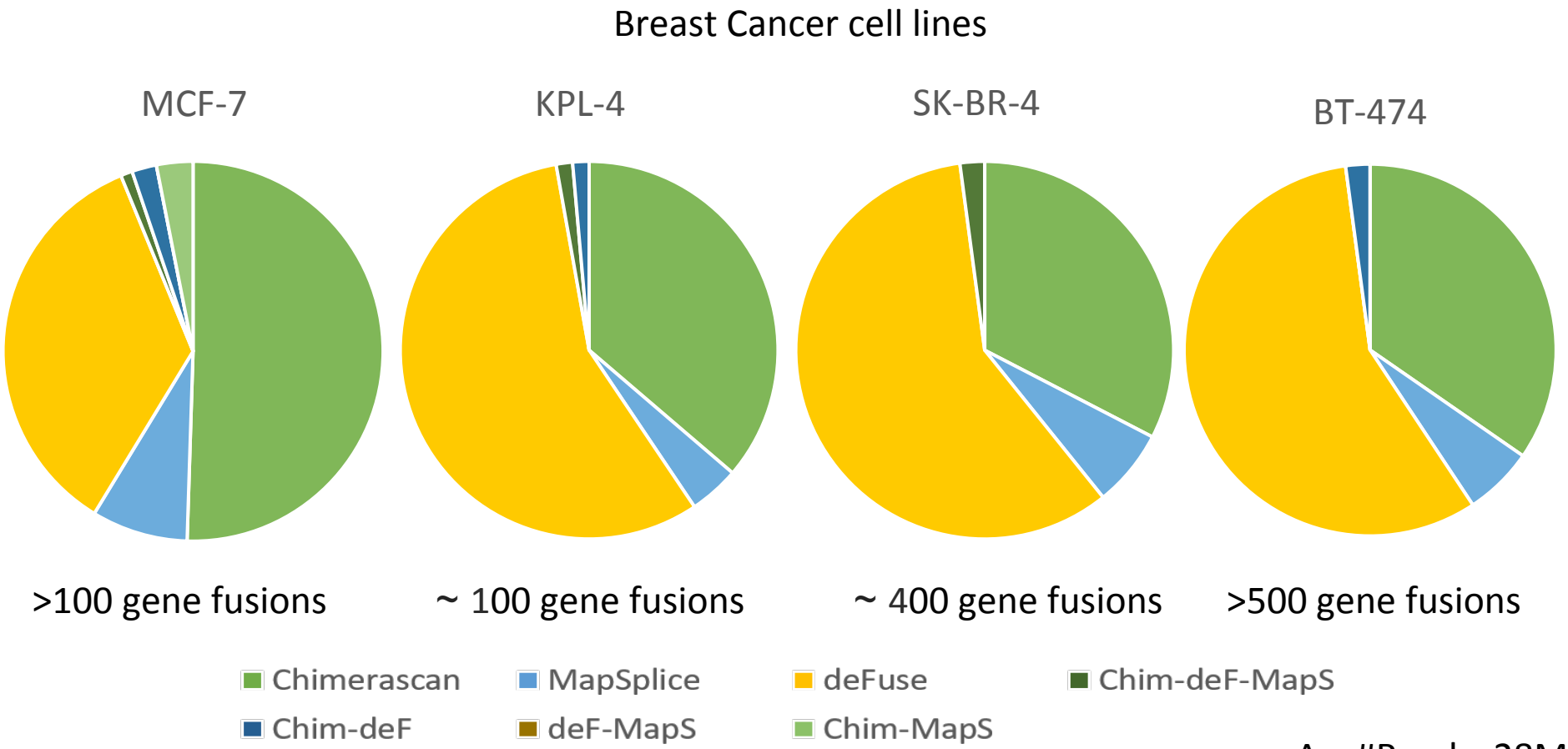
Output file S2

Output file SN

POLITECNICO DI TORINO

Results

Do we really need to prioritize fusions?



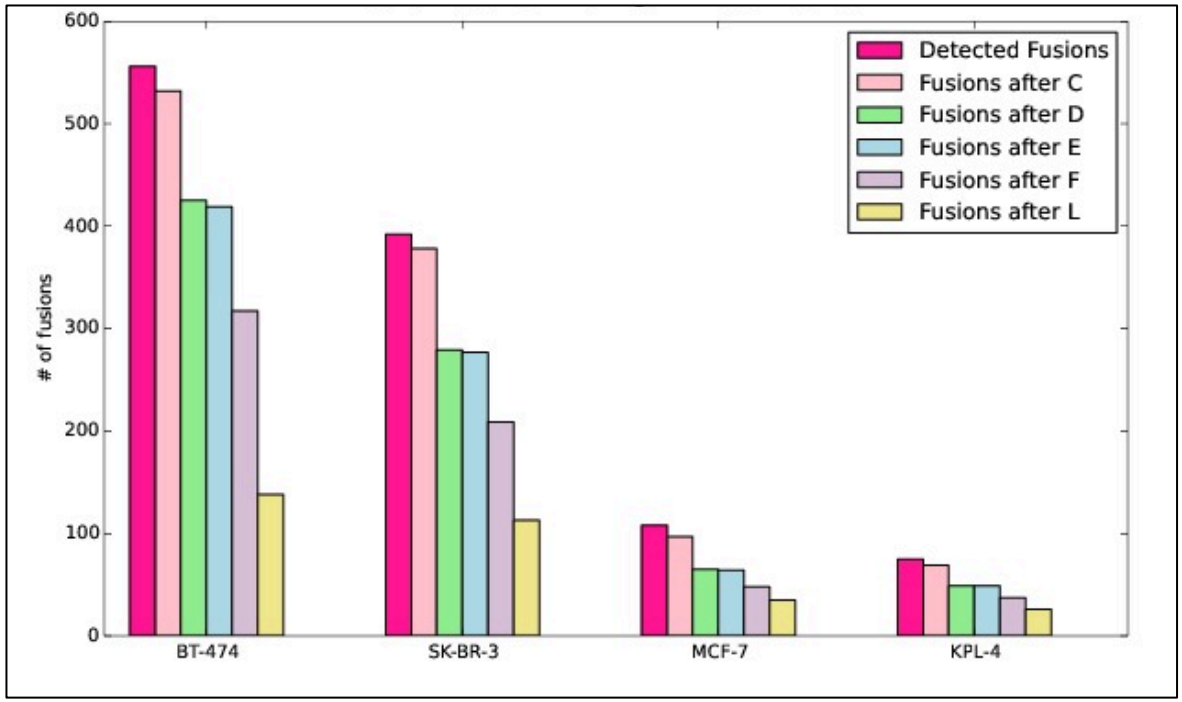
Av. #Reads: 28M

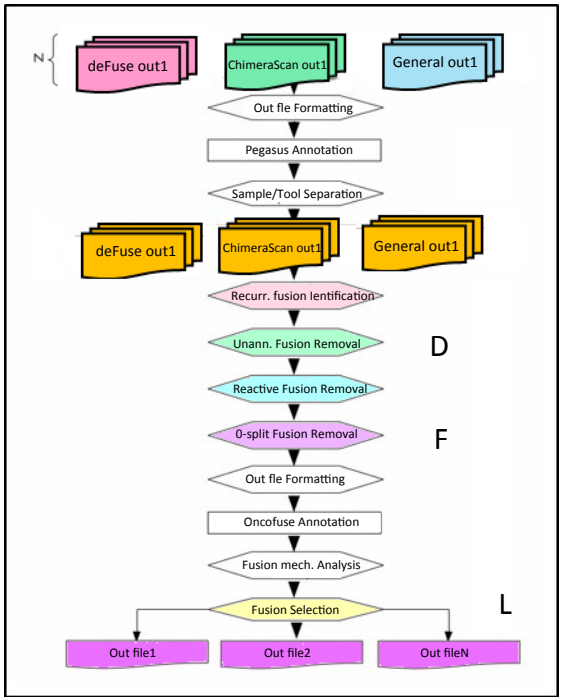
Edgren, Henrik, et al. **"Identification of fusion genes in breast cancer by paired-end RNA-sequencing."** Genome Biology 12.1 (2011): R6

POLITECNICO DI TORINO

Results

Priority fusions in breast cancer cell lines



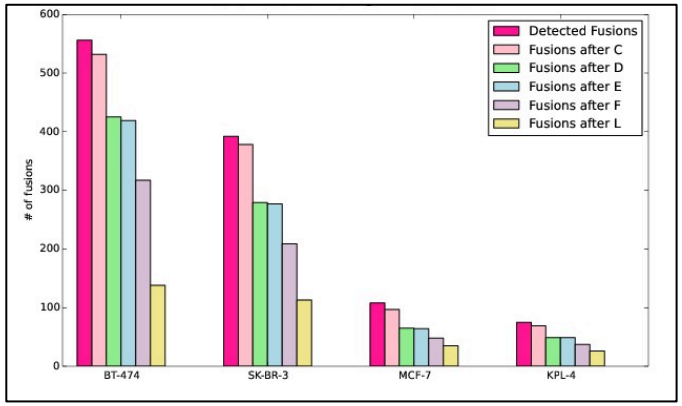


- ✓ **Reduction** in the number of fusions ranging from 65% to 75%
- ✓ 25 out of 26 **validated fusions** reported in output

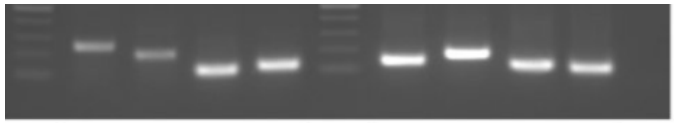
POLITECNICO DI TORINO

Remarks and Perspectives

To summarise



FuGePrior **prioritized** fusions from state-of-the-art gene fusion discovery tools in breast, prostate, AML and Burkitt's lymphoma cancer samples



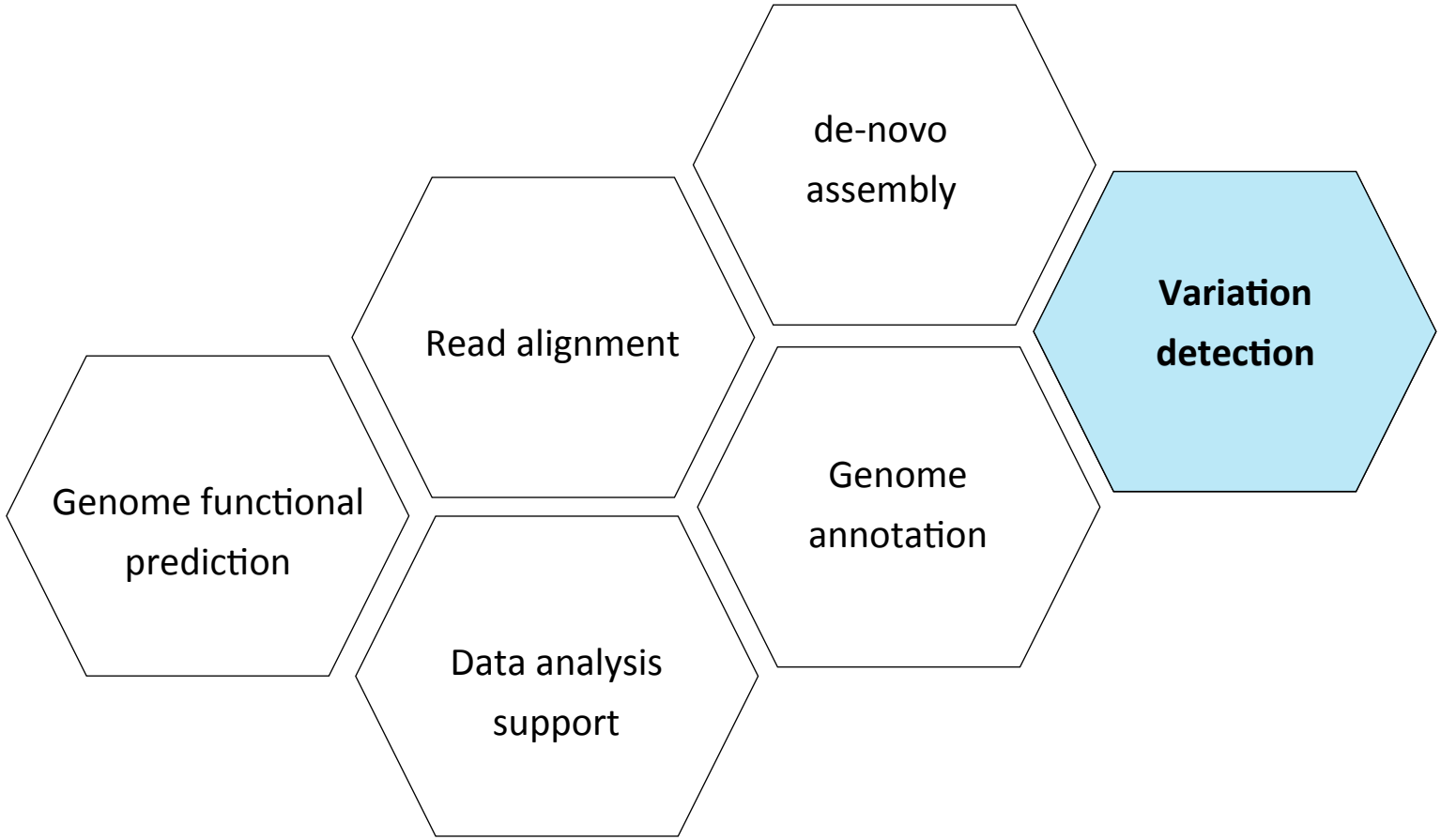
FuGePrior **priority** fusions were confirmed by both previous researches or by PCR experiments

Future work:

- additional gene fusion discovery tools
- additional filtering stages
- extensive ALL sample analysis

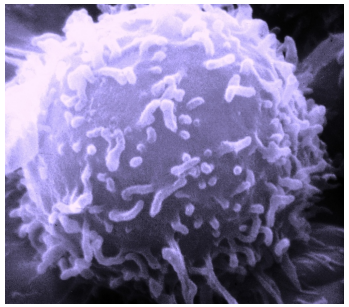
POLITECNICO DI TORINO

Conclusions



POLITECNICO DI TORINO

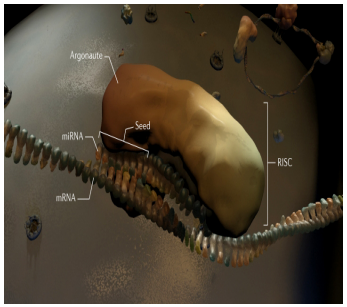
Conclusions



VDJSeq-Solver tool

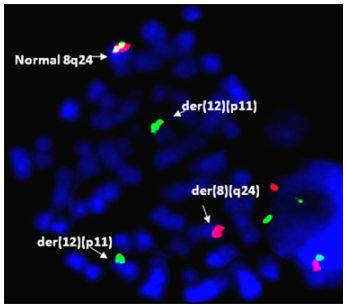
UNIVERSITÀ di VERONA

Variation detection



isomiR-SEA

ISTITUTO DI CANDIOLO-IRCCS



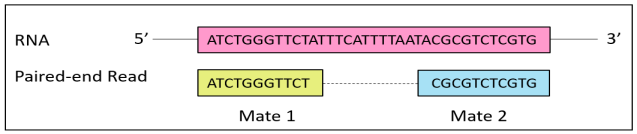
FuGePrior tool

SEVENTH FRAMEWORK PROGRAMME

NEXT GENERATION SEQUENCING for Targeted Personalized Therapy of Leukemia

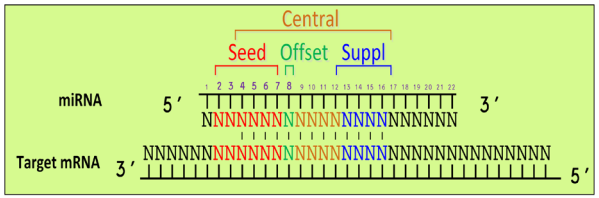
POLITECNICO DI TORINO

Conclusions



- ✓ **No Ig-Seq** experiments
- ✓ **Multiple** analyses on the same datasets

Variation detection



- ✓ miRNA reads have a **complete seed**
- ✓ **isomiR** profiles
- ✓ Evaluation of **miRNA/isomiR targeting** activity

- ✓ **Reduction** of false positive predictions
- ✓ FuGePrior fusions are highly **reliable**

POLITECNICO DI TORINO

Publications

Paciello, Giulia, et al. **"A novel pipeline for V (D) J junction identification using RNA-Seq paired-end reads."** In BIOSTEC 2013: , International Conference on Bioinformatics Models, Methods and Algorithms, Barcelona, 11-14 February 2013

Paciello, Giulia, et al. **"VDJSeq-Solver: in silico V (D) J recombination detection tool."** PloS one 10.3 (2015): e0118192

Paciello, Giulia, et al. **"Detection of rearranged light chain sequences by RNA-Seq in B-cell lymphomas and reactive lymphadenopathies."** In: EAHP 2016: 18th Meeting of the European Association for Hemathopathology, Basilea, 3-8 Sept. 2016

Urgese, Gianvito, Paciello, Giulia, et al. **"miR-SEA: miRNA Seed Extension based Aligner Pipeline for NGS Expression Level Extraction."** In IWBBIO 2014: 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, 7-9 Apr, 2014

Urgese, Gianvito, Paciello, Giulia, et al. **"On the relevance of a complete characterisation of miRNAs, isomiRs and miRNA-mRNA interaction sites through miRNA-specific alignment tools**." In RNAi 2015: Short & Long Non-coding RNAs, 10th International Conference & Exhibition, Oxford, 24-26 March, 2015

Urgese, Gianvito, Paciello, Giulia, et al. **"isomiR-SEA: an RNA-Seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation.**" BMC bioinformatics 17.1 (2016): 148

Padella, Antonella, Paciello, Giulia, et al**. "Next-Generation Sequencing Analysis Revealed That BCL11B Chromosomal Translocation Cooperates with Point Mutations in the Pathogenesis of Acute Myeloid Leukemia.**" Blood 124.21 (2014): 2352-2352. Proceedings of AHS Annual Meeting

Padella, Antonella, Paciello Giulia, et al**. "RNA Sequencing Reveals Novel and Rare Fusion Transcripts in Acute Myeloid Leukemia."** Blood 126.23 (2015): 3627-3627. Proceedings of AHS Annual Meeting

Padella, Antonella, Paciello, Giulia, et al**. "Novel fusion transcripts identified by RNAseq cooperate with somatic mutations in the pathogenesis of acute myeloid leukemia."** Cancer Res (76) (2016). Proceedings of AACR 107th Annual Meeting