Rilevazione di eventi di Alternative Splicing a partire da read paired-end

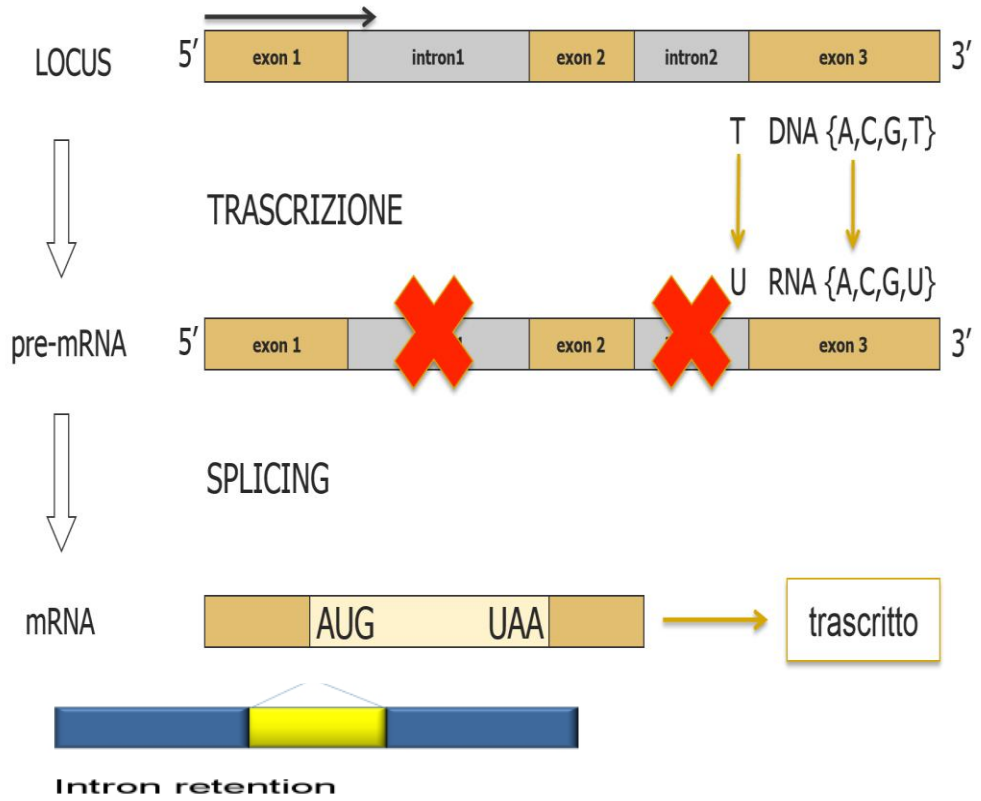
RELAZIONE FINALE DI PORTO FRANCESCO
MATRICOLA 816042
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Outline

- Concetti fondamentali (Alternative Splicing, read paired-end, ASGAL)
- Modifiche apportate ad ASGAL per supportare read paired-end
- Conclusioni e sviluppi futuri

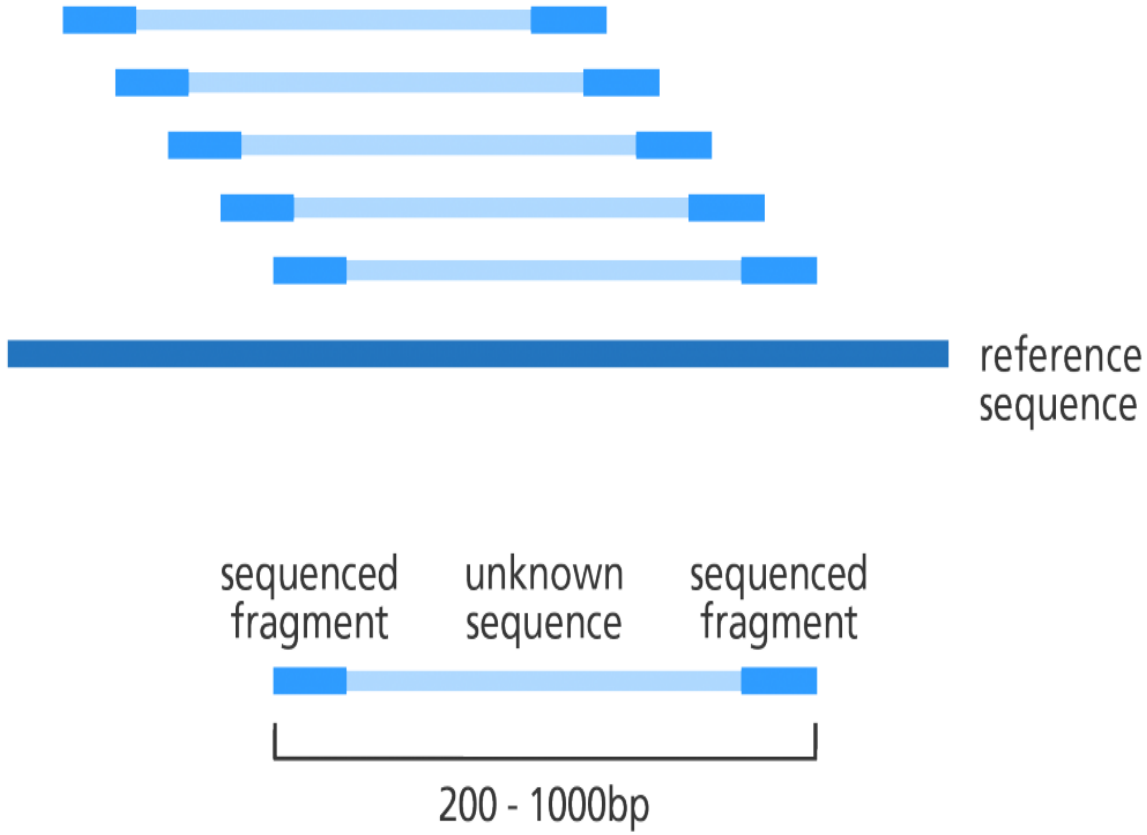
Alternative Splicing

- Meccanismo per la produzione di **proteine diverse** da uno stesso gene
- Utilizzato da oltre il **75%** dei geni umani
- Correlato a molte **malattie** (es: Alzheimer)
- Diversi tipi di eventi di Alternative Splicing



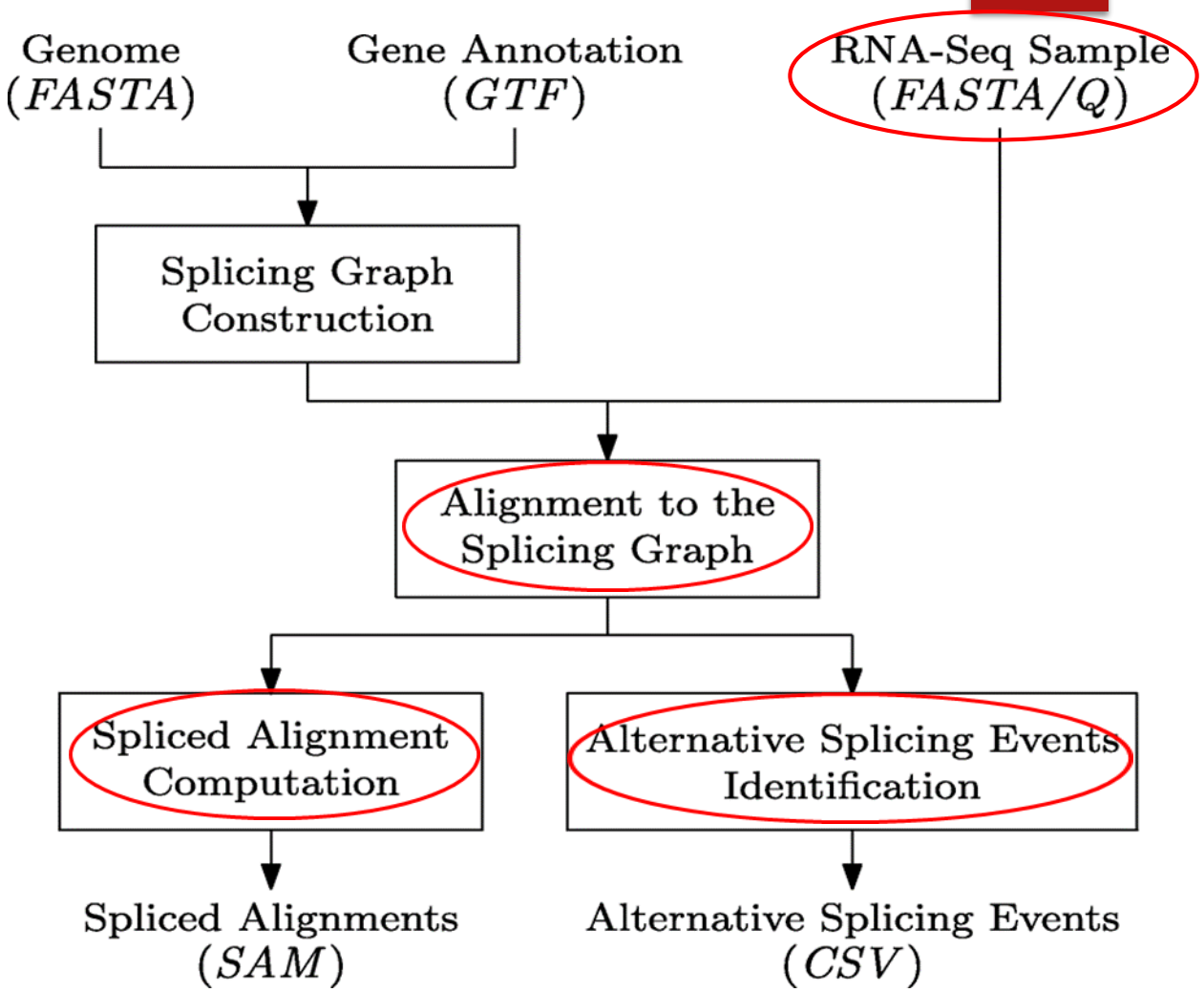
Read paired-end

- **Due letture** (“*mates*”) + **distanza** tra di esse
- Prodotte da sequenziatori **NGS** (*Next Generation Sequencing*)
- Risultati più accurati rispetto a read single-end

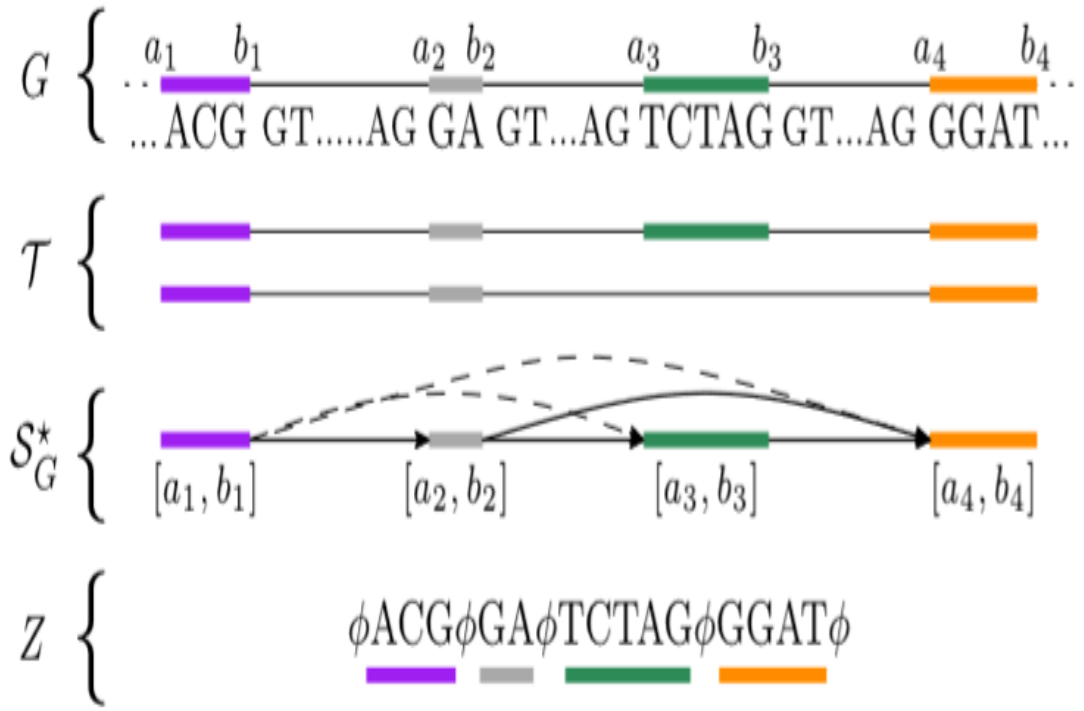


ASGAL

- **ASGAL** sta per "Alternative Splicing Graph ALigner"
- Rileva eventi di Alternative Splicing a partire da campioni di **RNA**
- **<u>Obiettivo stage</u>**: estendere ASGAL per supportare read paired-end, incrementandone le capacità di rilevazione



Costruzione Splicing Graph & Allineamento Splice-Aware



- Un **MEM (Maximal Exact Match)** è una tripla (t,p,l) tale che:
  - t = posizione di partenza su Z
  - p = posizione di partenza sulla read
  - l = lunghezza della sottostringa comune massimale
- 1 allineamento <-> 1 o più MEM
- Allineamenti primari e secondari
- Necessario allineare i due mate contemporaneamente

Introduzione read unmapped & placeholder

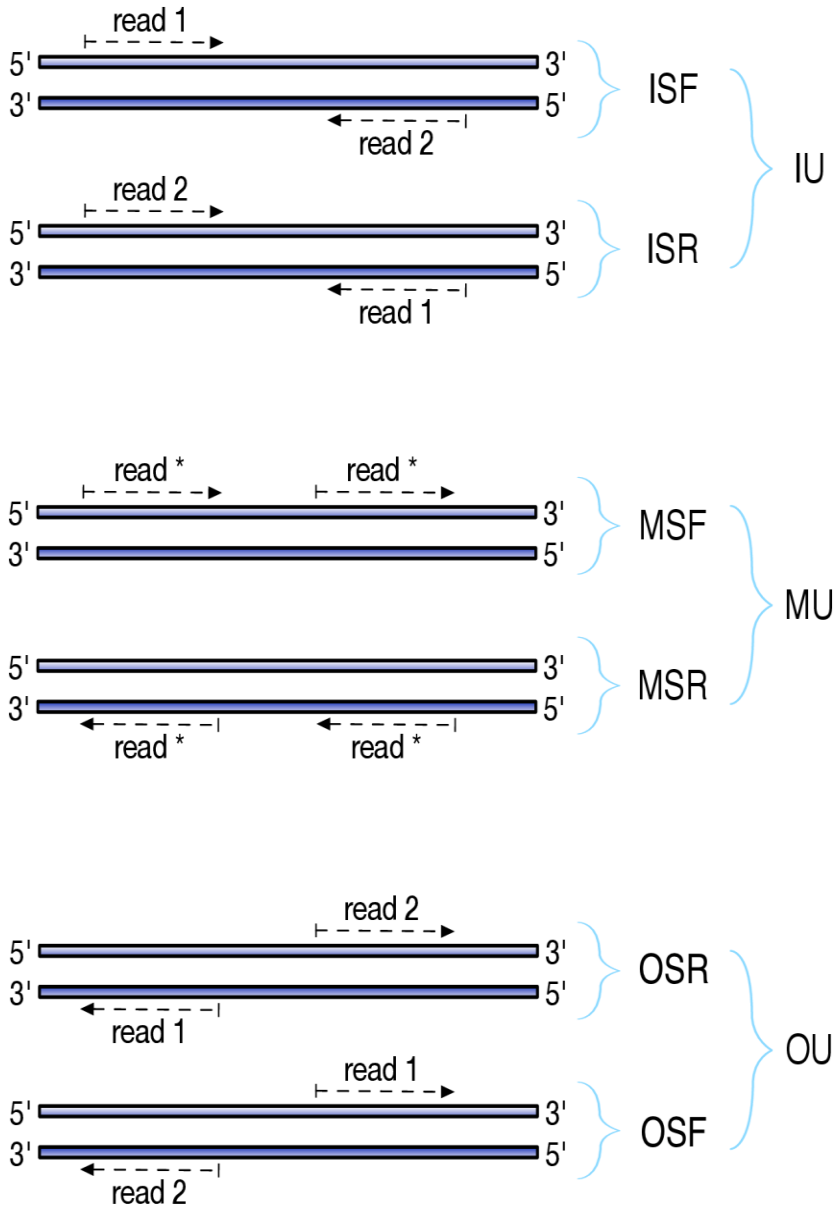
Consideriamo le due estremità («mate») di una read paired-end:

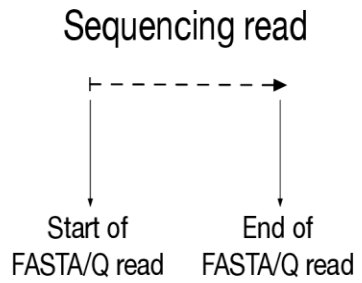
- Problema #1: Non sempre entrambe le estremità vengono allineate
  - Introduzione di **read unmapped**
- Problema #2: Non sempre entrambe le estremità sono coinvolte nello stesso numero di allineamenti secondari
  - Introduzione di **read placeholder**

```
MAPPED + ENST00000623047_e_4_21_6632291 0 (5940,1,53) (8956,53,48)
GGATTTGCCCCATCGCATATCTGGAGTTCGGGGTCTTAGAAAGCTTTCTTGCCCTATTTCTTTAGCAGAATGAGTGTCGTACATTTCCCAGGACTGTTTT
UNMAPPED ENST00000623047_e_90_21_6638170
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATGAGCCACTGTACTGCGCTGTGCCTACTTCAAAGGACTGAAAATAAAAAATAAATA
PLACEHOLDER ENST00000623047_e_4_21_6632291
```

Fragment Library Types

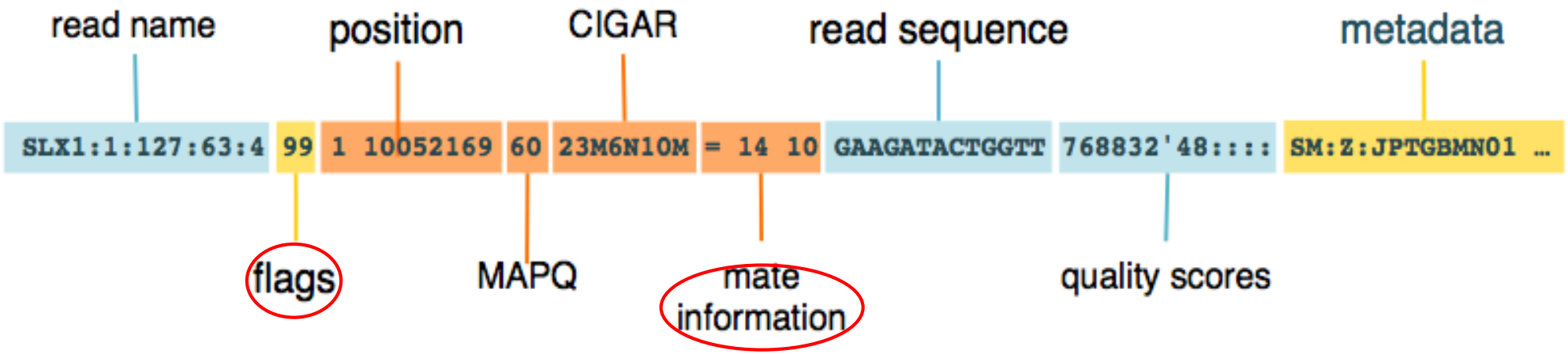
- Descrivono il formato di read paired-end
- **Posizione reciproca** delle read + **orientamento prima read** (se disponibile)
- Aumento dell'efficienza in fase di allineamento fino al 50%





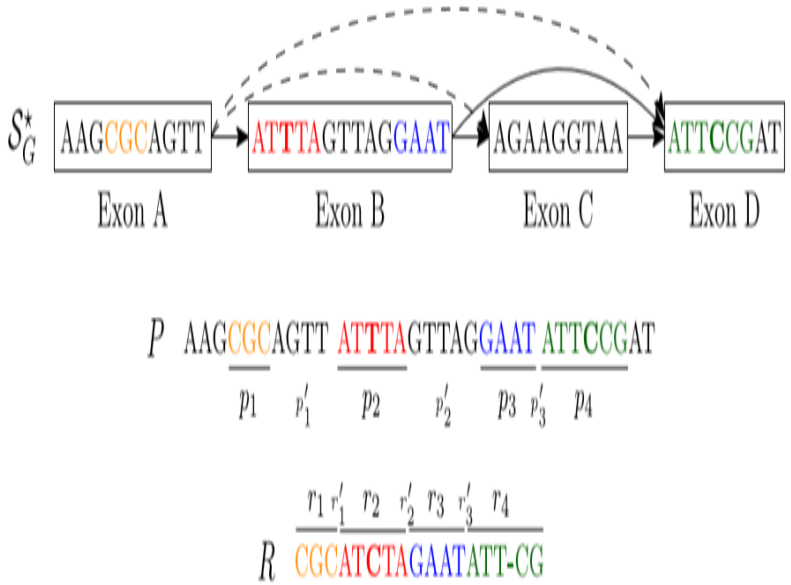
Computazione degli allineamenti Spliced (Formattazione SAM)

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



Rilevazione di eventi di Alternative Splicing

- Fase di **analisi** dei MEM prodotti dall'Allineatore Splice-Aware per rilevare nuovi introni
- Viene effettuato un confronto tra introni *noti* e introni *dedotti* dai MEM
  - **Differenze** tra i due <-> **eventi di Alternative Splicing**
- E' stata effettuata una **merge degli introni** dedotti da entrambi i mates




```
Type,Start,End,Support,Transcripts
ES,6634769,6670521,7,ENST00000625185/ENST00000624965
A3,6630571,6634603,22,ENST00000623047
A3,6668244,6670521,11,ENST00000623324
A3,6668255,6670521,32,ENST00000623313
read-ENSG00000280145.events.csv (END)
```

IDMP & TIDMP, Sviluppi Futuri

- **IDMP** (Inner Distance between Mate Pairs) distanza sul genoma tra read allineate
  - **Problema #1 :** Distanza tra read non sempre disponibile, possibile calcolarla via allineamento con altri allineatori
  - **Problema #2 :** Non tutti gli allineatori hanno la stessa nozione di «distanza tra read» + politiche di allineamento diverse portano a risultati diversi
- **TIDMP** (Transcript-based IDMP) distanza sui trascritti tra read allineate
  - **Problema #1:** Non è ancora stato calcolato su trascritti non consecutivi
  - **Problema #2:** Valida alternativa a IDMP?

IDMP & TIDMP, Sviluppi Futuri (2)

- **Scartare le read** allineate con IDMP/TIDMP eccessivo
  - Potrebbe peggiorare le capacità di rilevazione di alcuni eventi (es: Exon Skipping)
- **Riallineamento** in caso di IDMP/TIDMP eccessivo
  - Difficile da implementare
  - Possibile peggioramento delle capacità di rilevazione

Competenze acquisite

- Utilizzo di **dati biologici** in formati diversi (fasta, gtf, SAM, ecc.) e creazione di **algoritmi** che li manipolano
- Utilizzo di **strumenti** di natura bioinformatica: SAMTools per la validazione dei file SAM, IGV per la visualizzazione degli eventi sul genoma, ecc.
- Utilizzo di **community di esperti** di bioinformatica: è stato usato **Biostar** (forum di riferimento per la bioinformatica) per chiarire alcuni dubbi sull'allineamento di read paired-end (https://www.biostars.org/p/376192/)
- Approfondimento dei linguaggi di programmazione **Python** e **C++** e utilizzo di **librerie specifiche** per la bioinformatica (es: kseq)
- Utilizzo di **Linux**

Conclusioni

- ASGAL è ora in grado di:
  - **allineare** correttamente read paired-end con lo Splicing Graph
  - **salvare** gli allineamenti ottenuti nel formato SAM (rispettando le specifiche per quanto riguarda read paired-end)
  - **rilevare** eventi di Alternative Splicing a partire da read paired-end
- Rimangono da investigare i possibili utilizzi di IDMP e TIDMP per migliorare la qualità della rilevazione
- Codice disponibile su: https://github.com/HopedWall/galig