Next Generation Sequencing

- Riguarda la quantificazione dei trascritti
- I trascritti possono avere diverse isoforme determinati dal processo di splicing
- L'analisi di sequenze trascrittomiche si divide in 3 parti:
  - **Mapping/Alignment**: Si cercano di mappare le reads ai trascrittomi/genomi di riferimento
  - **Reconstruction**: Si vanno a ricostruire i vari trascritti tramite assembly
  - **Quantification**: Si fanno delle stime di quantita' dei vari trascritti

Mapping

E' il task piu' basilare, e consiste nell'andare a mappare le sequenze di reads rispetto ad un trascrittoma/genoma di riferimento. Si dividono in due tipologie principali: *unspliced aligners* e *spliced aligners*.

UnSpliced Aligners

Cercano di mappare carattere per carattere sul genoma di riferimento, senza permettere gap di grandi dimensioni. Due tipologie principali:

- **Seed-Based**: Dividono le reads in diverse sequenze dette *seeds* che vengono matchati sul genoma, poi le parti matchate vengono estese con algoritmi piu' precisi tipo Smith Waterman.
- **Burrows Wheeler-Based**: Utilizzano una struttura a grafo per massimizzare lo score (quindi ridurre al minimo i gaps/mismatches e massimizzare i match)

Il problema di questo tipo di allineatori e' che non tengono conto delle spanning reads, per cui le performance, sopratutto con BWT-based decrementa esponenzialmente in relazione al numero di gaps/mismatchesiu

Spliced Aligners

Cercano di tenere conto dello splicing, e si dividono in due approcci principali:

- **Exon First**: vanno ad applicare un allineatore unspliced tipo BWT, poi prendono tutte le reads che non sono state matchate, le dividono, e riprovano a matcharle. Le regioni genomiche che circondano le reads che sono state mappate, vengono poi ricercate per una possibile connessione di reads spanning.
- **Seed Extended**: dividono le reads in segmenti di *k-mers* di dimensione molto piccola, per poi provare ad estendere le zone che sono state mappate con delle reads con degli algoritmi piu' precisi come Smith-Waterman

Transciptome Reconstruction

Sono metodi per risalire all'isoforma del trascrittoma letto. Per poter fare cio', e' necessario assemblare tutte le reads lette in diversi trascritti. Si dividono in due metodologie principali:

- **Genome Guided**
- **Genome Indipendent**

Genome Guided

Utilizzano un genoma di riferimento per mappare inizialmente tutte le reads possibili, e poi assemblare tutte le reads che si sovrappongono nei trascritti. Si dividono a loro volta in 2 tipologie:

- **Exon Identification**: Prima mappano le reads rispetto al genoma, per cui tutte le reads che fanno match vengono categorizzate come *"isole esoniche"*, successivamente vanno a prendere tutte le reads che non sono state allineate (e quindi reads di tipo *spanning*) e provano ad assemblare queste isole mediante l'overlap di tali reads. Sono stati sviluppati quando i macchinari trattavano reads di ~50bp
- **Genome-guided Assembly**: trasformano direttamente in genoma in una struttura a grafo che rappresenta tutte le possibili connessioni di basi nel trascrittoma sia quando regioni intronicheesoniche sono vicine, che quando sono divise.

Genome Indipendent

Al posto di mappare preliminarmente le reads ad un trascrittoma/genoma di riferimento, utilizzano le reads per costruire direttamente dei trascritti di consenso. Questi trascritti vengono poi mappati sul genoma oppure su sequenze prote-geno/miche in modo da essere catalogate.

In generale, utilizzano dei grafi di DeBruijn per rappresentare tutte le sottosequenze di *k-mers* delle reads che fanno overlap. I percorsi all'interno di questo grafo costituiscono tutte le possibili sequenze che possono essere costruire con i *k-mers*