

Corso di Metodi Molecolari e Bioinformatica,  
LM Biologia Evoluzionistica  
A.A. 2022-2023  
Università di Padova

**Esercitazione 6**  
***“Analisi di dati RNA-seq”***

*Padova, 14 dicembre 2022*

Docente:  
S. Bortoluzzi

Assistenti:  
Dr. Enrico Gaffo  
Dr. Alessia Buratin  
Silvia Orsi



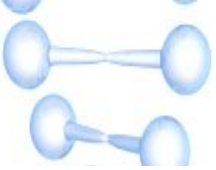
# Outline

- Visualizzazione annotazioni e allineamenti con IGV
- Allineamento al genoma di dati RNA-seq
- Quantificazione dell'espressione di geni e trascritti
- Ricostruzione del trascrittoma (con genoma di riferimento)



# Dati nel pacchetto necessari per l'analisi

- Come nell'esercitazione precedente, abbiamo:
  - Genoma in FASTA:
    - CFLAR\_HIPK3.*fa*
  - Le read sequenziate (paired-end) in FASTQ di 2 campioni:
    - SRR2923169\_CFLAR\_HIPK3\_1.fastq
    - SRR2923169\_CFLAR\_HIPK3\_2.fastq
    - SRR2923170\_CFLAR\_HIPK3\_1.fastq
    - SRR2923170\_CFLAR\_HIPK3\_2.fastq
- In più:
  - Annotazioni geniche in GTF:
    - CFLAR\_HIPK3.*gtf*



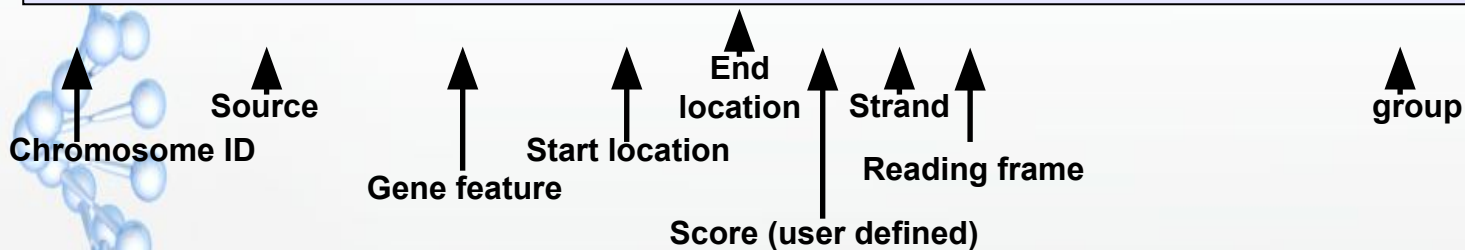
# General Feature Format (GFF)

File di testo con **9 campi** separati dal carattere **TAB**

(<https://genome.ucsc.edu/FAQ/FAQformat#format3>)

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start\_codon", "stop\_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

AB000381	Twinscan	CDS	380	401	.	+	0	gene_id "001"; transcript_id "001.1";
AB000381	Twinscan	CDS	501	650	.	+	2	gene_id "001"; transcript_id "001.1";
AB000381	Twinscan	CDS	700	707	.	+	2	gene_id "001"; transcript_id "001.1";
AB000381	Twinscan	start_codon	380	382	.	+	0	gene_id "001"; transcript_id "001.1";
AB000381	Twinscan	stop_codon	708	710	.	+	0	gene_id "001"; transcript_id "001.1";



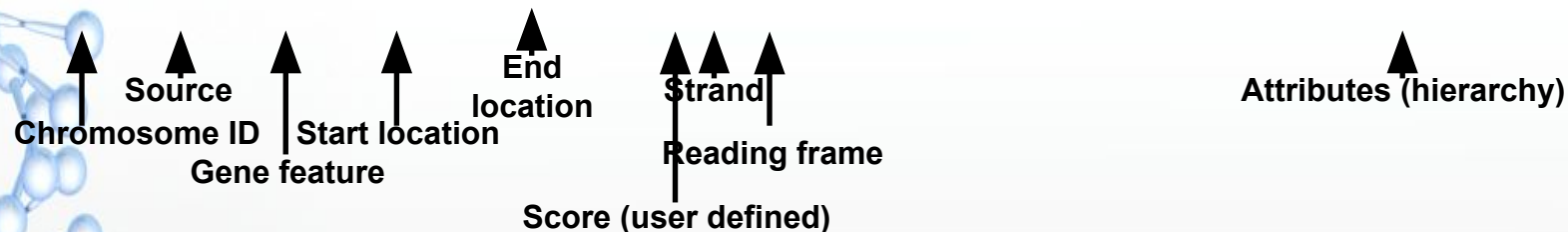
# Gene Transfer Format (GTF)

Come il formato GFF, ma il campo 9 è chiamato **attributes** e deve cominciare con due attributi:

- **gene\_id**: A globally unique identifier for the genomic source of the transcript
- **transcript\_id**: A globally unique identifier for the predicted transcript.

Vedi <http://mblab.wustl.edu/GTF2.html>

```
2   havana exon 46899275 46899314 . + . gene_id "ENSG00000228925"; transcript_id "ENST00000429761";
2   havana exon 46906884 46908678 . + . gene_id "ENSG00000228925"; transcript_id "ENST00000429761";
```






# Browser Extensible Data (BED) format

Altro formato per definire annotazioni geniche

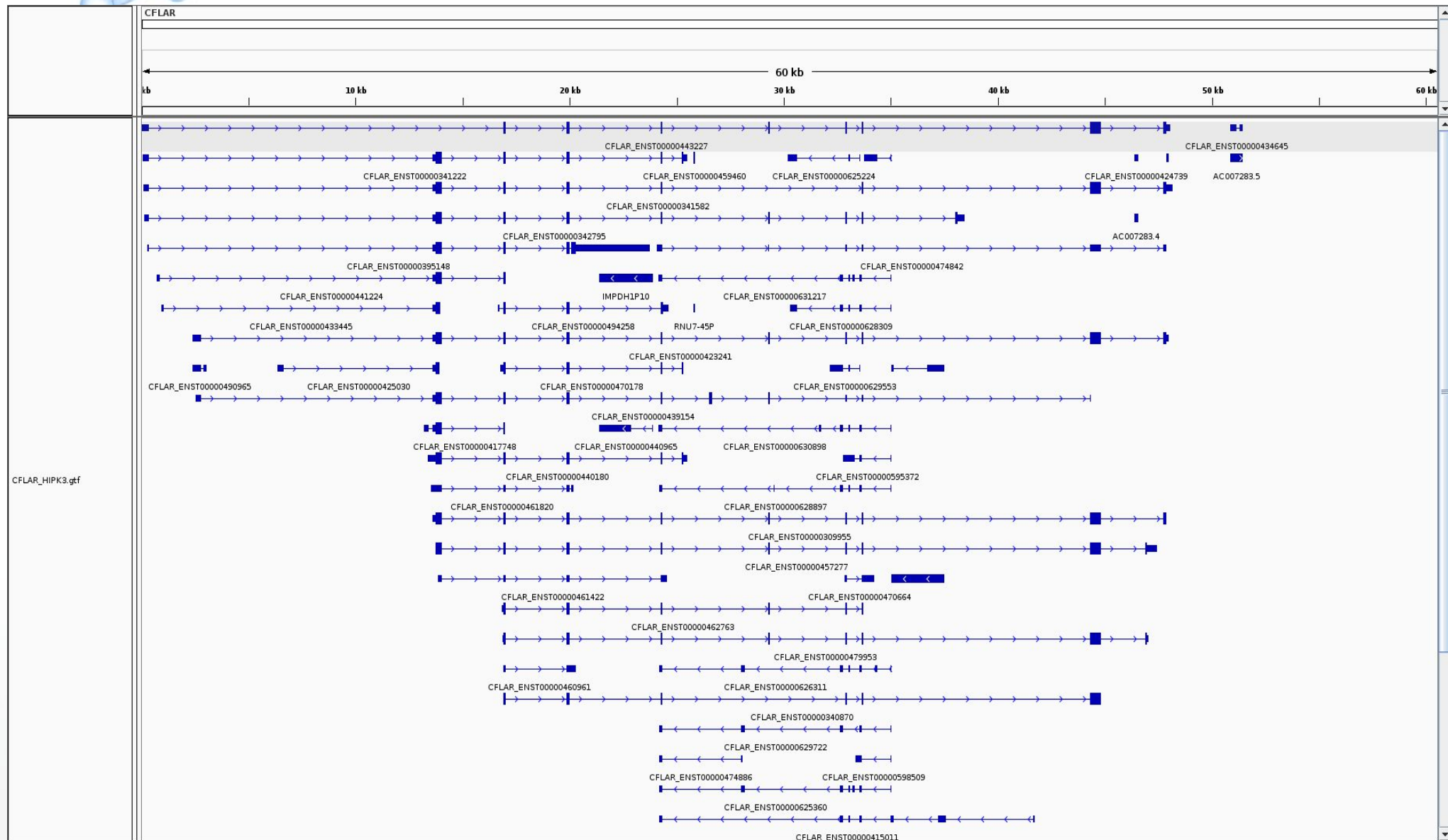
<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2\_random) or scaffold (e.g. scaffold10671).
  2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
  3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The chromEnd base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100, and span the bases numbered 0-99.
- 

# Visualizzare le annotazioni con Integrative Genomics Viewer (IGV)

<http://software.broadinstitute.org/software/igv/>





# Eseguire IGV e caricare i dati


Aprire un terminale e lanciare il comando:

```
$ /opt/IGV/igv.sh &
```

*(il carattere “&” a fine comando ci permette di eseguirlo in background)*

... attendere che si carichi l'interfaccia grafica.

Dai menu a tendina:

- Caricare il file del genoma CFLAR\_HIPK3.fa
  - Caricare il file di annotazioni CFLAR\_HIPK3.gtf
- 

# RNA-seq data analysis

- Grande quantità di informazione
- Vari passaggi di analisi
- Passaggi diversi a seconda dell'informazione che si vuole ricavare

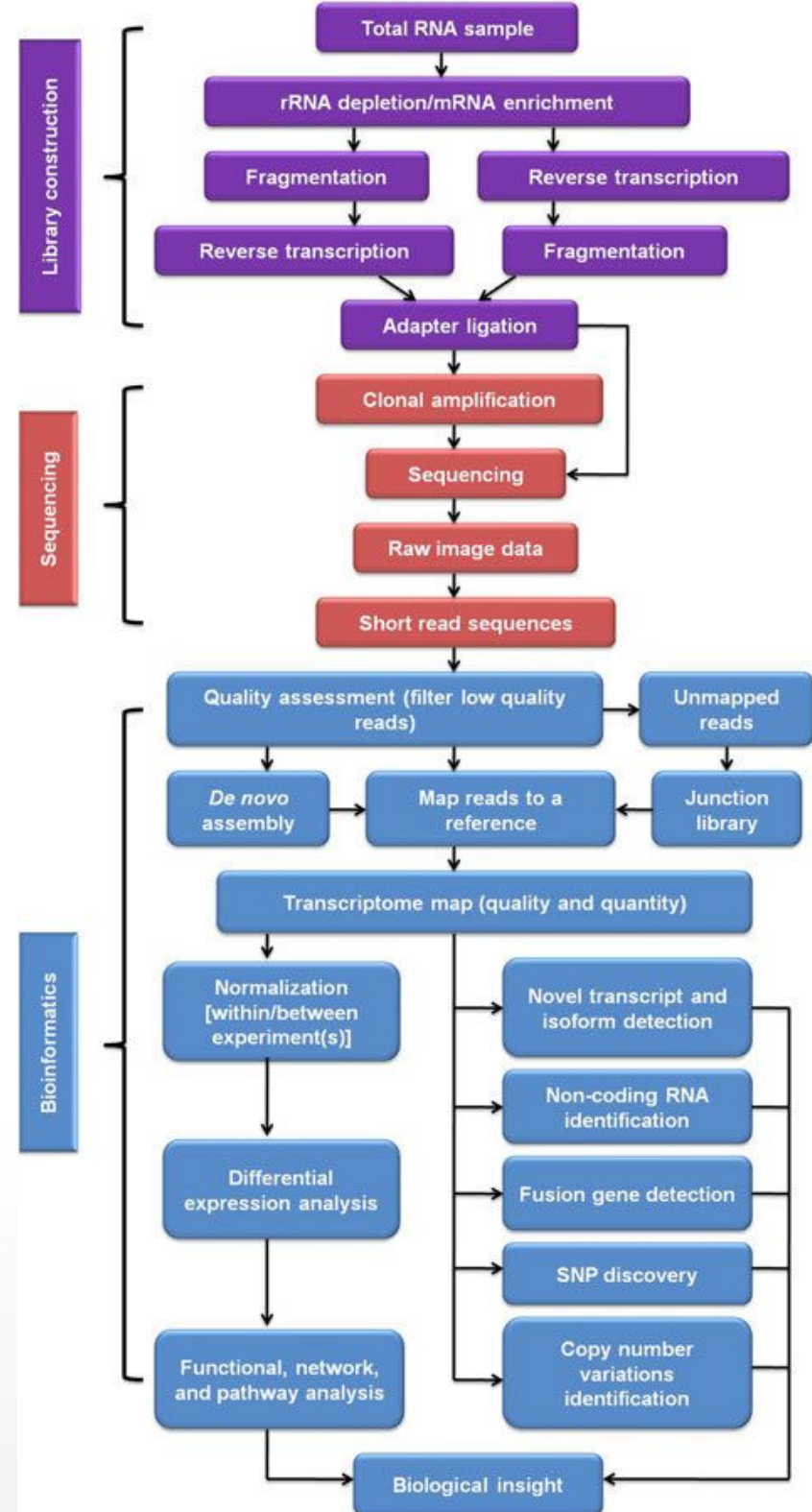
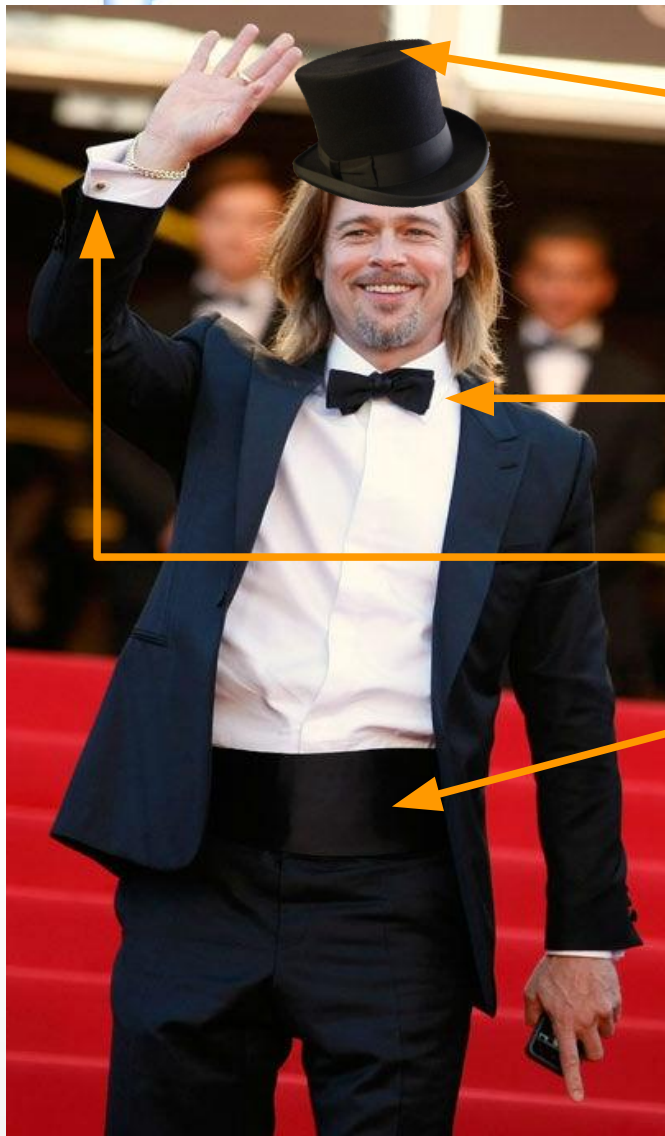


Figure from Qian et al. (2014) <https://doi.org/10.1089/omi.2013.0110>

# Quantificare l'espressione di geni/trascritti

Utilizziamo i “Tuxedo tools”<sup>[1]</sup> (<http://cole-trapnell-lab.github.io/cufflinks/>)



**TopHat**

**Bowtie**

**Cufflinks**

**CummeRbund**

[1] Trapnell, C. et al. **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. Nature Protocols 7, 562–578 (2012).

Bowtie  
Extremely fast, general purpose short read aligner

TopHat  
Aligns RNA-Seq reads to the genome using Bowtie  
Discovers splice sites

Cufflinks package

Cufflinks  
Assembles transcripts

Cuffcompare  
Compares transcript assemblies to annotation

Cuffmerge  
Merges two or more transcript assemblies

Cuffdiff  
Finds differentially expressed genes and transcripts  
Detects differential splicing and promoter use

CummeRbund  
Plots abundance and differential  
expression results from Cuffdiff



Quantum  
StringTie

A photograph of Brad Pitt smiling and waving on a red carpet. He is wearing a dark tuxedo jacket over a white shirt and a black bow tie. A red square box is drawn around the bow tie, and a red arrow points from the top left towards the box. A large red question mark is positioned above the box.

A woman dressed as Belle from Disney's 'Beauty and the Beast' is shown in a grand, ornate ballroom. She is wearing her iconic yellow gown and is captured in a dynamic pose, spinning or twirling, with her arms outstretched. The room features high ceilings, large columns, and chandeliers, creating a classic, elegant atmosphere. An orange arrow points to the bottom right corner of the image.

```
graph TD
    RA[Reference annotation]
    RSR[RNA-Seq reads]
    
    subgraph Steps_1_2 [Steps 1-2]
        HISAT[HISAT]
    end
    
    subgraph Step_3 [Step 3]
        StringTie[StringTie]
    end
    
    subgraph Step_4 [Step 4]
        Merge[StringTie --merge]
    end
    
    subgraph Step_5 [Step 5]
        GFF[gffcompare]
        TS[Transcript statistics]
    end
    
    subgraph Step_6 [Step 6]
        StringTie_eB[StringTie -eB]
    end
    
    subgraph Steps_7_21 [Steps 7-21]
        Ballgown[Ballgown]
        PDET[Plots and differential expression tables]
    end
    
    RA -.-> HISAT
    RA -.-> StringTie
    RA -.-> GFF
    RSR --> HISAT
    HISAT --> RA
    HISAT --> RA
    
    RA --> StringTie
    StringTie --> RA
    StringTie --> Merge
    
    Merge --> GFF
    GFF --> TS
    GFF --> OR{OR}
    
    OR -- "no transcript assembly" --> StringTie_eB
    OR --> StringTie_eB
    
    StringTie_eB --> Ballgown
    StringTie_eB --> PDET
```

The flowchart illustrates the RNA-Seq analysis pipeline across several steps:

- Steps 1-2:** RNA-Seq reads are input into HISAT, which performs alignment. The process is guided by a reference annotation.
- Step 3:** HISAT outputs read alignments, which are then processed by StringTie for initial transcript assembly. The reference annotation also guides this step.
- Step 4:** StringTie performs a merge operation to produce merged transcripts.
- Step 5:** The merged transcripts are compared against the reference annotation using *gffcompare* to generate transcript statistics.
- Decision Point:** An OR gate checks for "no transcript assembly". If this condition is met, the process proceeds to Step 6. Otherwise, it proceeds to the next step.
- Step 6:** StringTie -eB is used to generate read coverage tables.
- Steps 7-21:** Ballgown is used to generate plots and differential expression tables based on the read coverage tables.

- TopHat → HISAT
- Cufflinks → StringTie
- CummeRbund → Ballgown

[2] Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L.  
**Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.** Nat. Protocols 11, 1650–1667 (2016).

# Other expression quantification tools

[Published: 20 April 2014](#)

## **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**

[Rob Patro](#), [Stephen M Mount](#) & [Carl Kingsford](#) ✉

[Nature Biotechnology](#) **32**, 462–464 (2014) | [Cite this article](#)

[Published: 06 March 2017](#)

## **Salmon provides fast and bias-aware quantification of transcript expression**

[Rob Patro](#) ✉, [Geet Duggal](#), [Michael I Love](#), [Rafael A Irizarry](#) & [Carl Kingsford](#) ✉

[Nature Methods](#) **14**, 417–419 (2017) | [Cite this article](#)

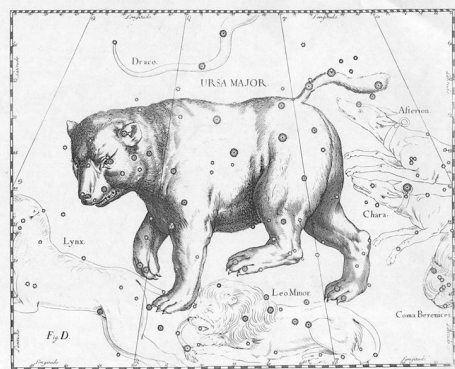
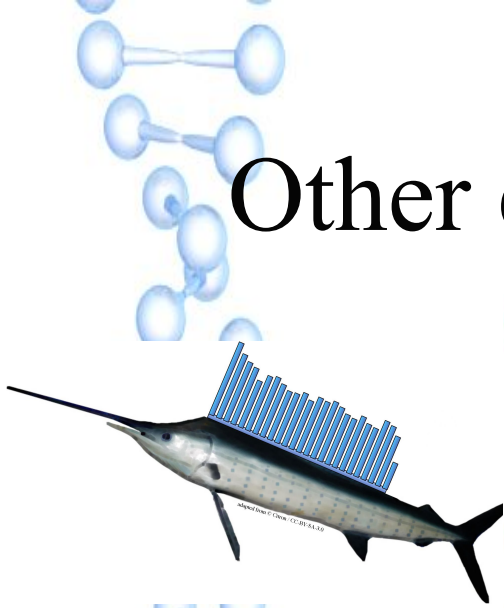
[Published: 04 April 2016](#)

## **Near-optimal probabilistic RNA-seq quantification**

[Nicolas L Bray](#), [Harold Pimentel](#), [Páll Melsted](#) & [Lior Pachter](#) ✉

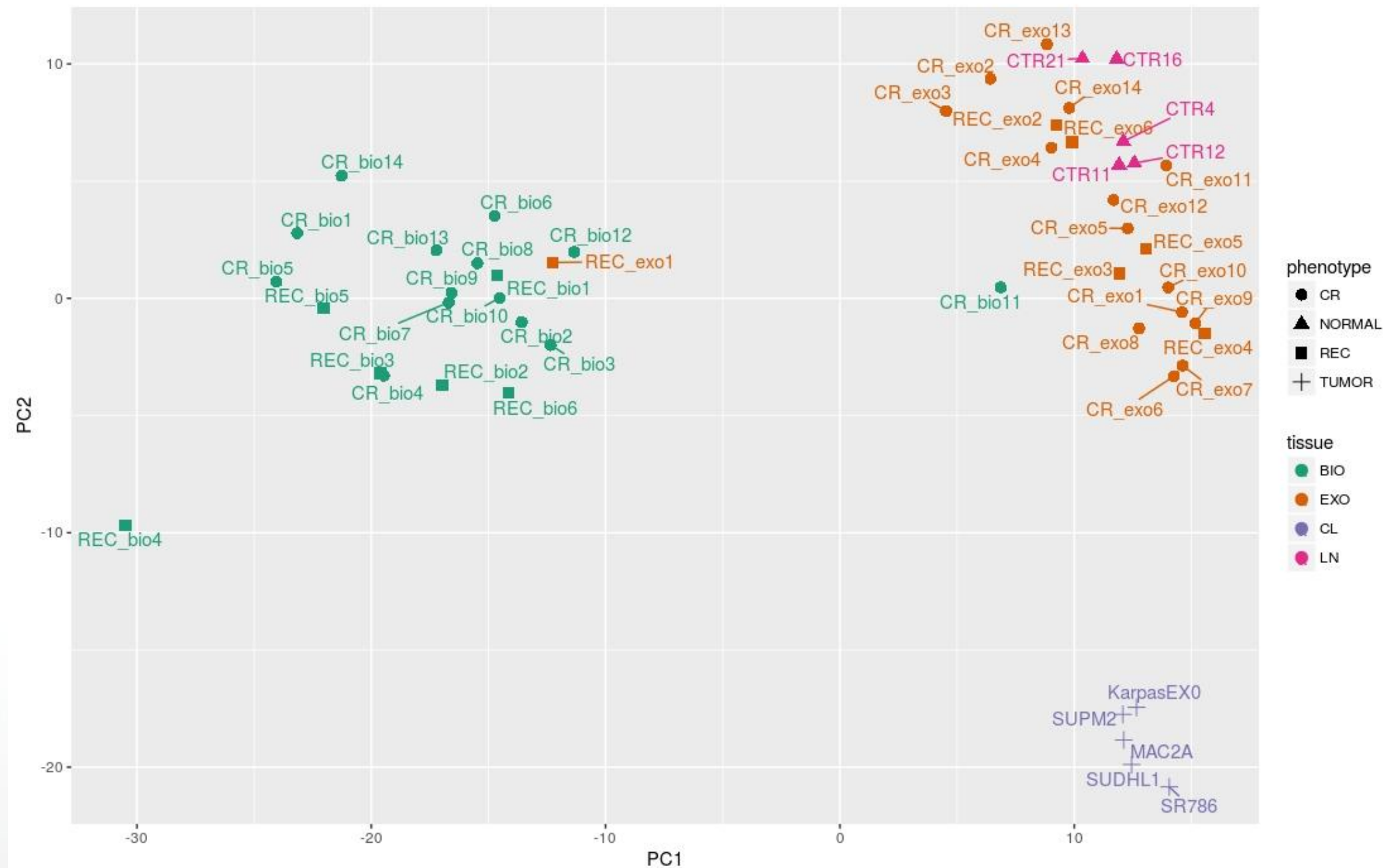
[Nature Biotechnology](#) **34**, 525–527 (2016) | [Cite this article](#)

# Kallisto

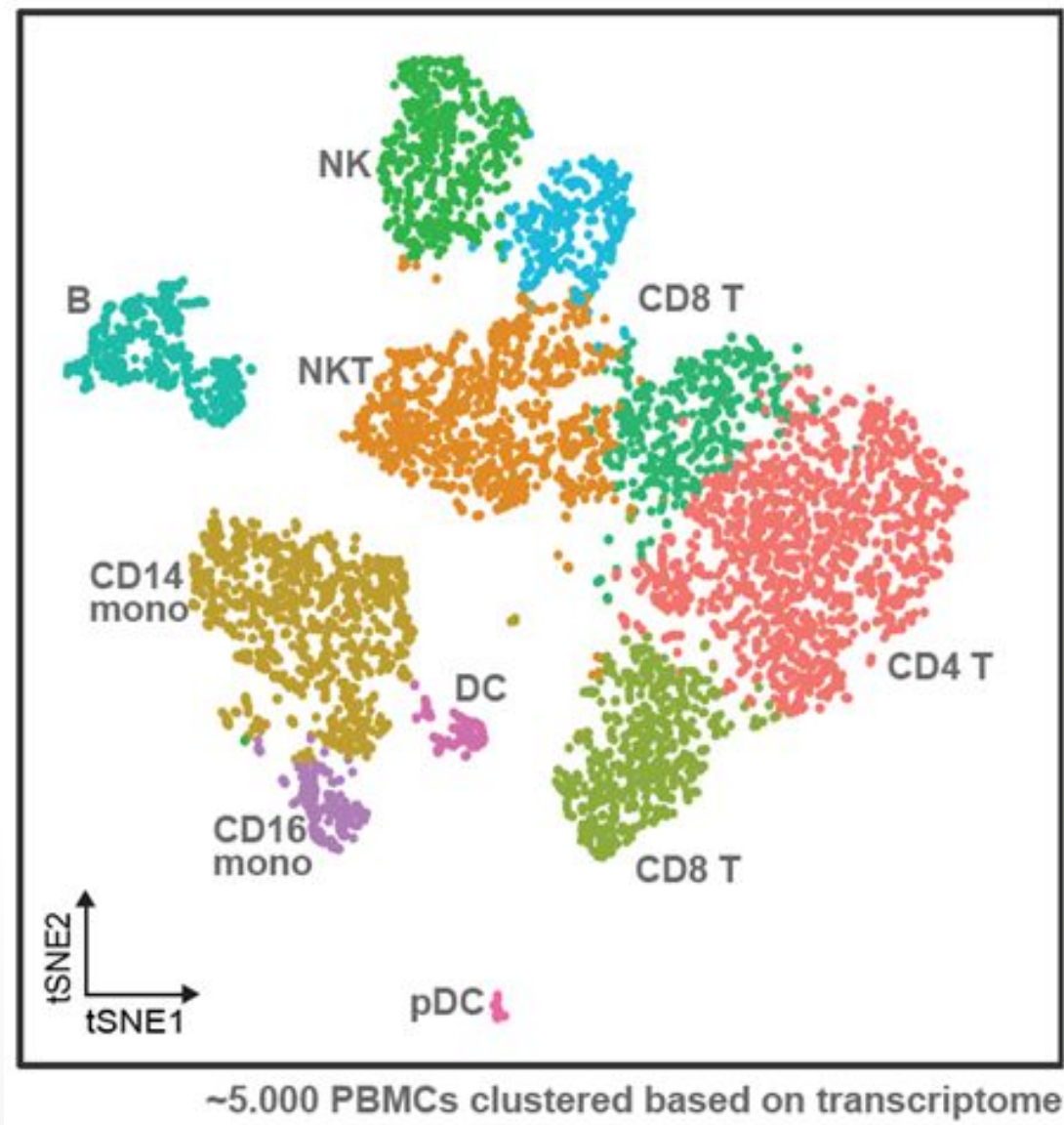




# Es. di analisi dei dati di espressione: PCA

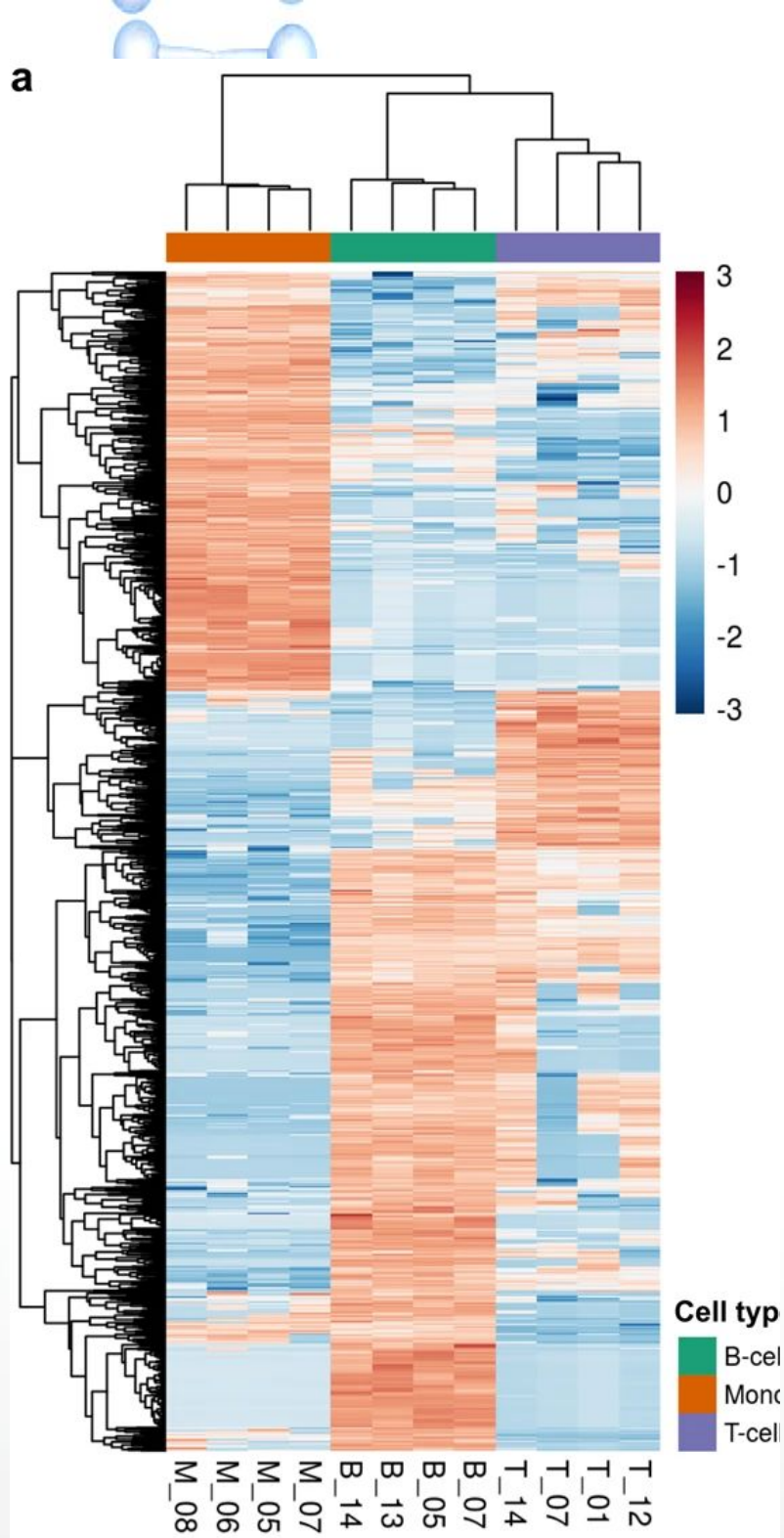


## Es. 2: clustering di dati single-cell sequencing

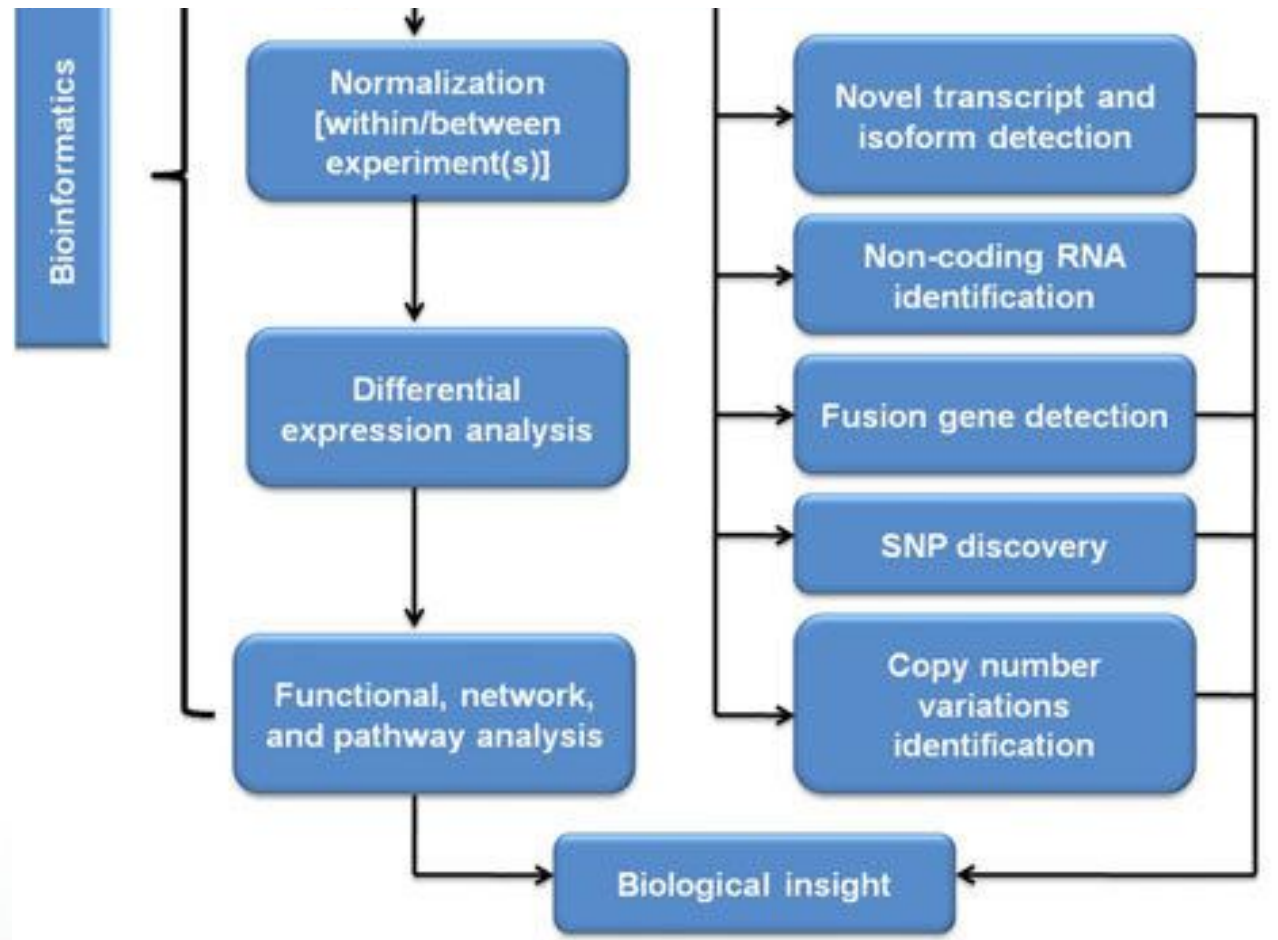


# Analisi comparative dell'espressione

Identificazione di geni differenzialmente espressi tra cellule, tessuti o condizioni sperimentali diverse



...e molto altro





# Allineamento read RNA-seq con HISAT2

<https://daehwankimlab.github.io/hisat2/>

*“HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA)”*

(La procedura è simile a quanto fatto nella lezione precedente)

- Creare l'indice del genoma:

```
$ hisat2-build CFLAR_HIPK3.fa CFLAR_HIPK3
```

- Allineare le read:

```
$ hisat2 -x CFLAR_HIPK3 -1  
SRR2923169_CFLAR_HIPK3_1.fastq -2  
SRR2923169_CFLAR_HIPK3_2.fastq --dta-cufflinks >  
ht2_ch69.sam
```



# Preparare gli allineamenti

- salvare in un BAM ordinato, per visualizzarle successivamente:

```
$ samtools view -bh ht2_ch69.sam > ht2_ch69.bam
```

- Controlliamo i parametri da usare con **samtools sort**

```
$ samtools sort
```

```
$ samtools sort ht2_ch69.bam > ht2_ch69.sorted.bam
```



# Procedura “one-line”

- In alternativa, usiamo l'operatore di **pipe** per redirezionare gli output ai programmi samtools ed evitare di scrivere file intermedi (N.B: scrivere tutto in un'unica riga!):

```
$ hisat2 -x CFLAR_HIPK3 -1 SRR2923169_CFLAR_HIPK3_1.fastq -2  
SRR2923169_CFLAR_HIPK3_2.fastq --dta-cufflinks | samtools view  
-uh - | samtools sort - > ht2_ch69.sorted.bam
```



il carattere - (“meno” o “trattino”) solitamente indica che l'input proviene dallo standard input

- ricordarsi di creare l'indice del BAM:

```
$ samtools index ht2_ch69.sorted.bam
```

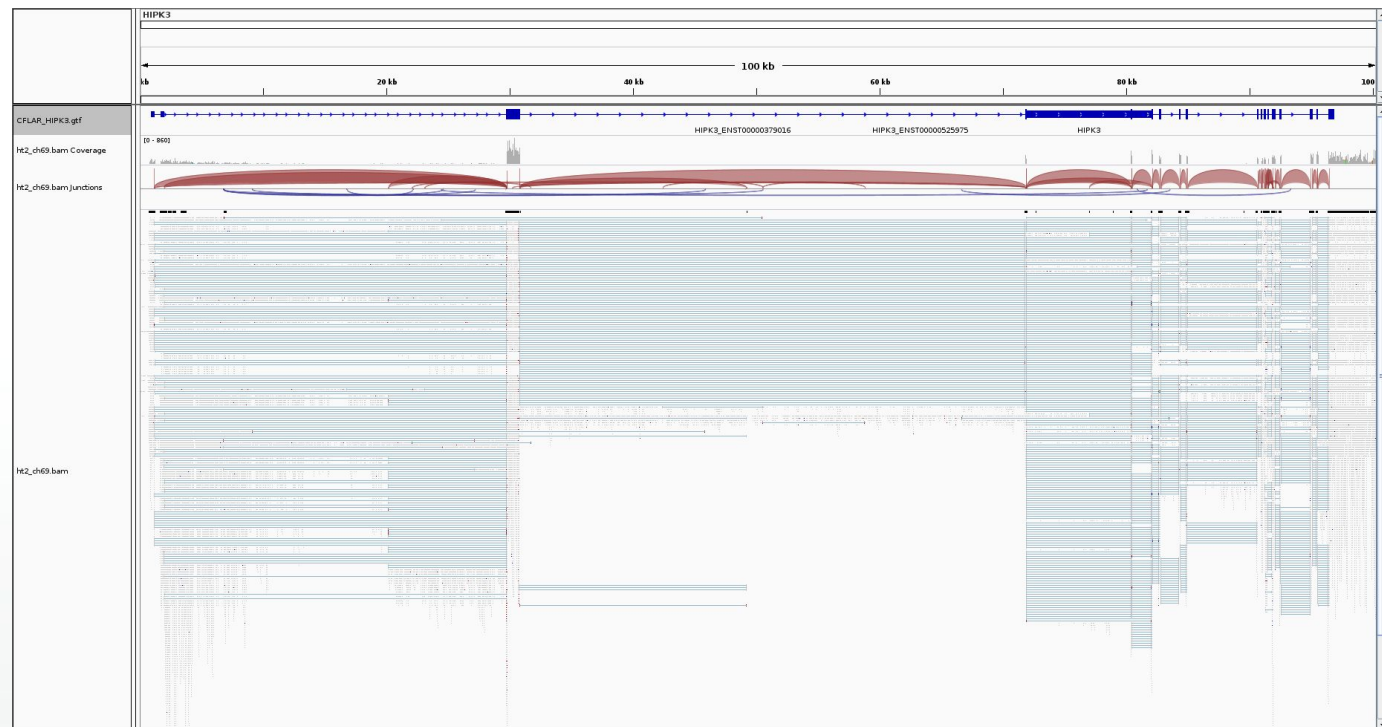


# Visualizzare allineamenti con IGV

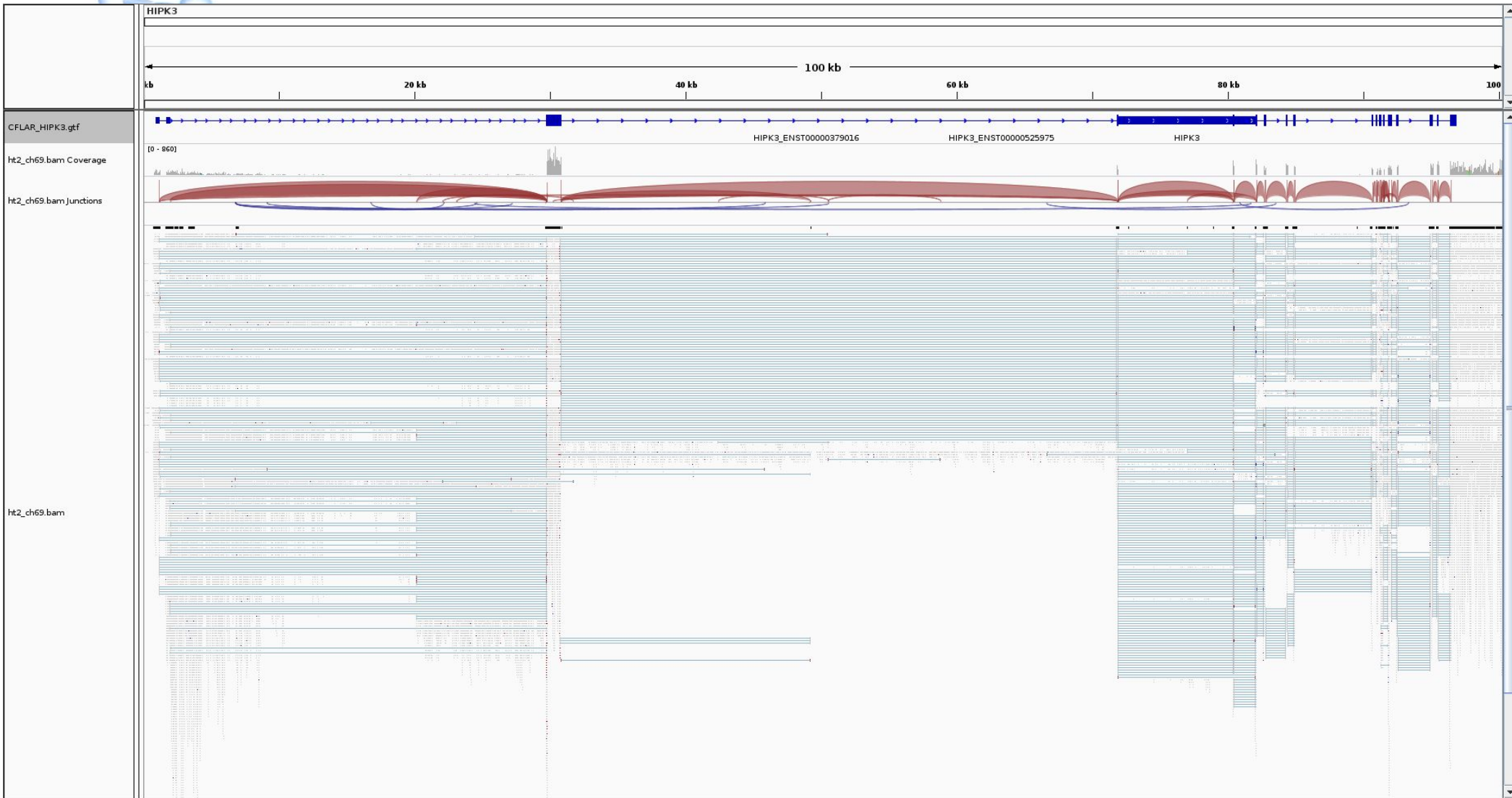
Con IGV possiamo visualizzare anche  
allineamenti di read NGS, come avevamo  
visto con la visualizzazione da terminale di

← `$ samtools tview`

Read alignments  
displayed by IGV



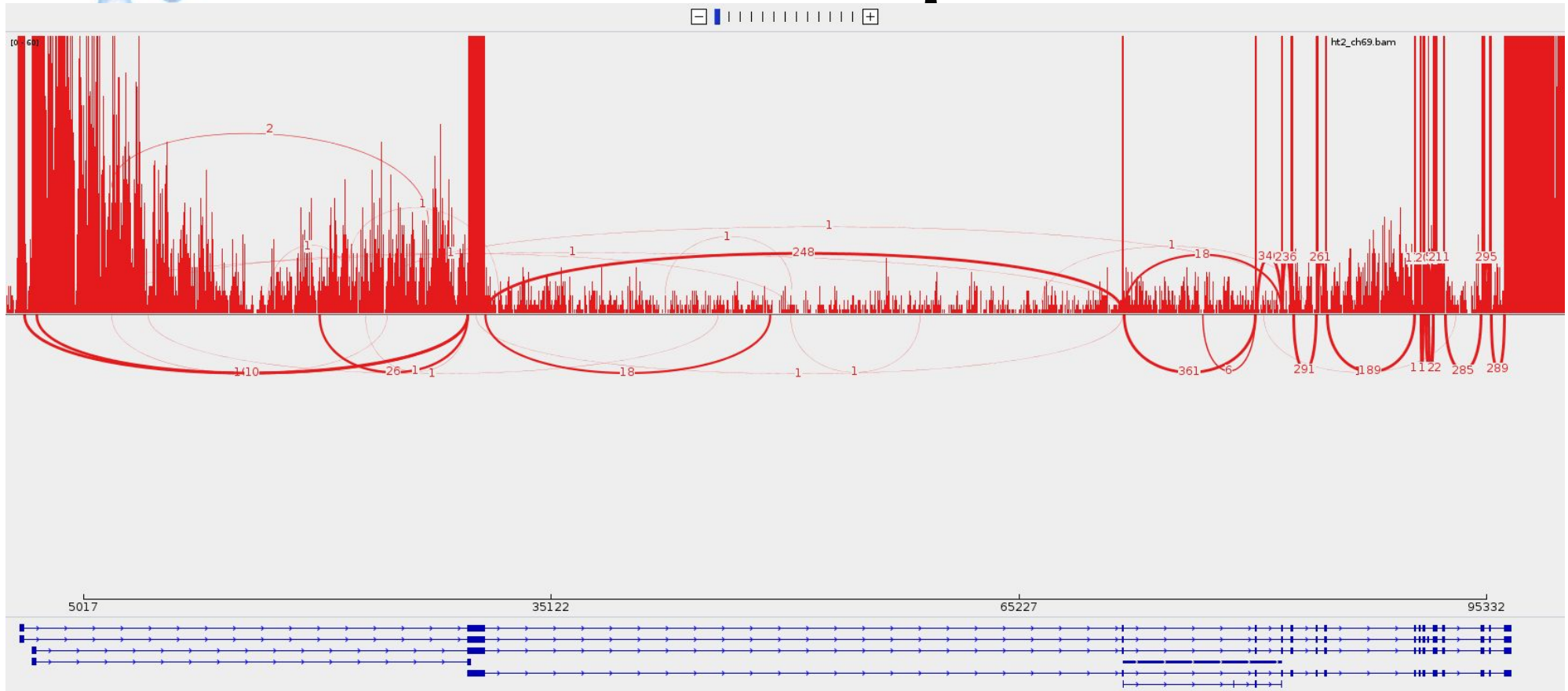
# Visualizzare le giunzioni di splicing



Le linee rappresentano il collegamento tra i frammenti delle read che sono state allineate in modo “spliced”.

La “junction track” dà un’idea della quantità degli eventi di splicing identificati nel mappaggio e delle regioni interessate.

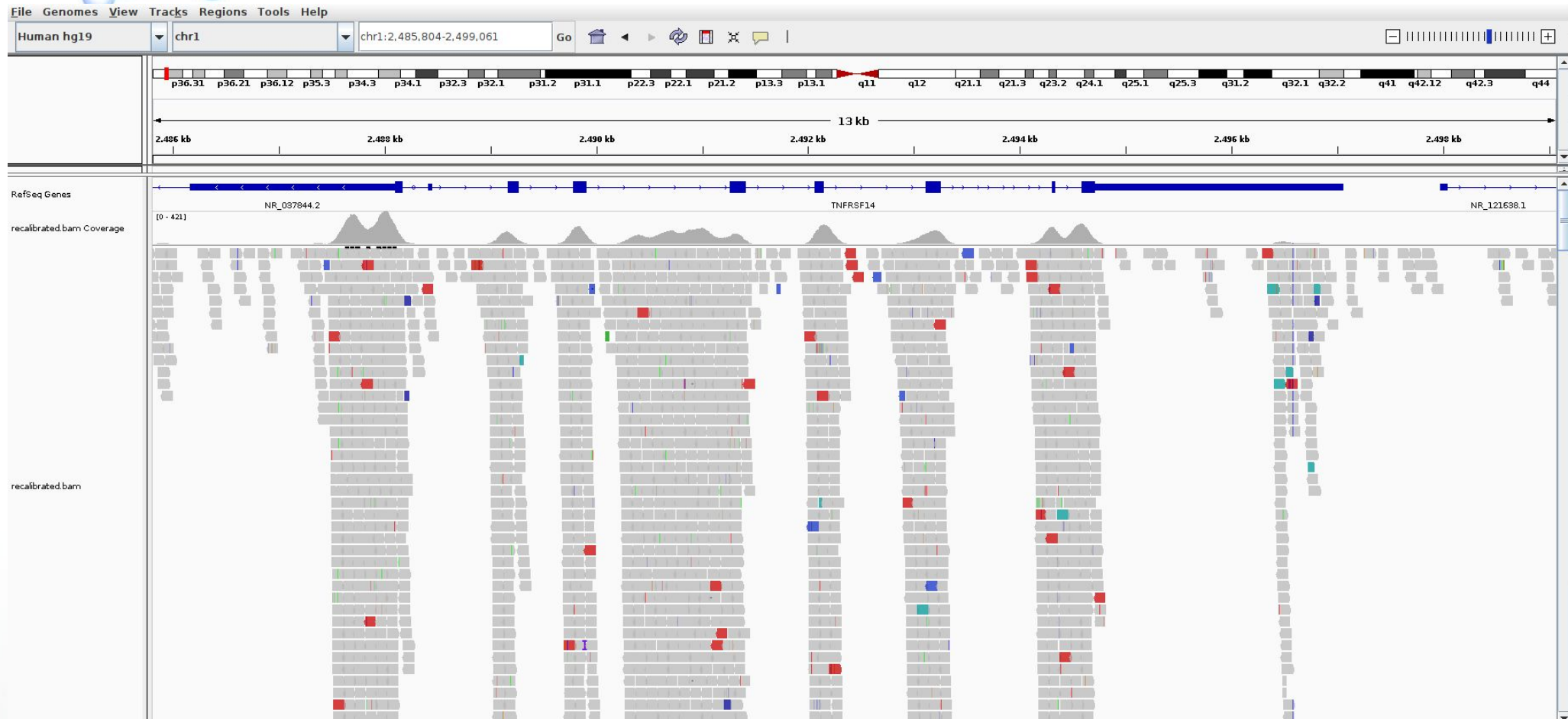
# Sashimi plot



Con un po' di fantasia....



# Differenza tra gli allineamenti RNA-seq e DNA-seq?



# Quantificare l'espressione dei geni/trascritti: StringTie



Published: 18 February 2015

## **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**

Mihaela Perte, Geo M Perte, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell & Steven L Salzberg ✉

Nature Biotechnology **33**, 290–295 (2015) | Cite this article

<https://ccb.jhu.edu/software/stringtie>

```
$ stringtie -e -G CFLAR_HIPK3.gtf -o trx.gft -A geneabund.tab  
ht2_ch69.sorted.bam
```

## Otteniamo

- un file GTF ([trx.gft](#)) con i valori di espressione dei trascritti nel nuovo tag degli attributi “cov”, “FPKM”, “TPM”
- un file tabellare ([geneabund.tab](#)) con l'espressione dei geni



# Analisi di espressione dei geni/trascritti

- Possiamo aprire il file tabellare con un foglio di calcolo.
- Il file GTF invece, meglio elaborarlo con gli strumenti della shell che abbiamo visto nelle lezioni precedenti

---

Usare ulteriori strumenti di analisi post-quantificazione.

Ad es si potrebbe usare il pacchetto `tximport` su R  
(<https://bioconductor.org/packages/release/bioc/html/tximport.html>)

```
$ stringtie -e -G CFLAR_HIPK3.gtf -o trx.gft -A geneabund.tab  
-b tabfiles ht2_ch69.sorted.bam
```

```
$ less tabfiles/t_data.ctab
```





# Transcriptome reconstruction (con genoma di riferimento)

**Ricostruzione guidata** (supportata) da annotazione

```
$ stringtie -G CFLAR_HIPK3.gtf -o annotrx.gtf ht2_ch69.bam
```

```
$ less -S annotrx.gtf
```





# Transcriptome reconstruction (con genoma di riferimento)

## Ricostruzione senza annotazione:

!! Da non confondere con la ricostruzione *de novo*, cioè senza di genoma di riferimento !!

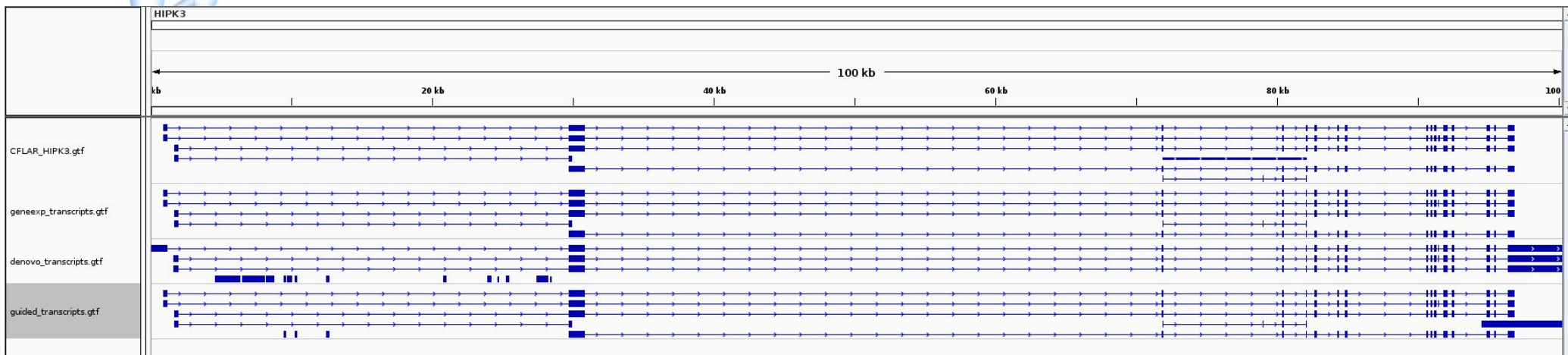
Altri strumenti vengono usati per quello, ad esempio Trinity o Velvet/Oases

```
$ stringtie -o noannotrx.gtf ht2_ch69.bam
```

```
$ less -S annotrx.gtf
```

# Confrontiamo i trascrittomi definiti con le varie opzioni, visualizzandoli con IGV

- CFLAR\_HIPK3.gtf
- annotrx.gtf
- noannotrx.gtf

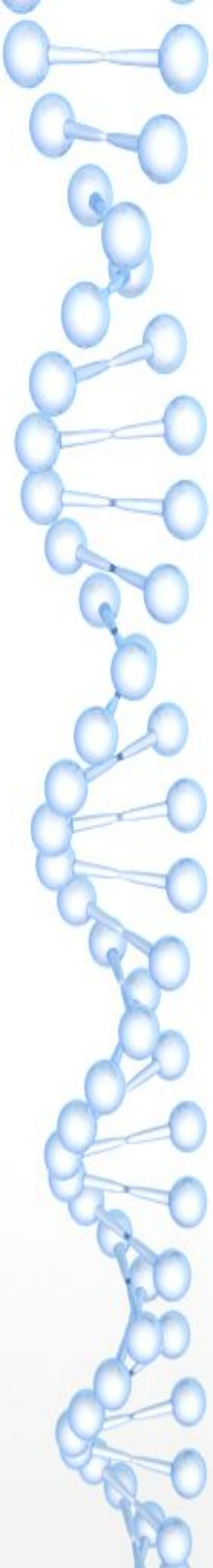


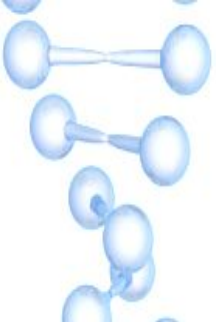


# Esercizi

- quantificare l'espressione dei geni definiti nel trascrittoma ricostruito:
  - quali sono gli input?
- ripetere le analisi anche per il secondo campione

# Appendice opzionale





# Allineamento con BWA

- Creare l'indice del genoma:

```
$ bwa index CFLAR_HIPK3.fa
```


- Allineare le read e salvare in un BAM già ordinato (N.B: usiamo l'operatore di pipe per redirezionare direttamente l'output al programma samtools sort):

```
$ bwa mem CFLAR_HIPK3.fa -1 SRR2923169_CFLAR_HIPK3_1.fastq  
SRR2923169_CFLAR_HIPK3_2.fastq | samtools sort - >  
ch69sorted.bam
```

- Visualizziamo gli allineamenti con IGV:

- Creare l'indice degli allineamenti prima di caricare il file in IGV:

```
$ samtools index ch69sorted.bam
```



# Altri possibili processamenti degli allineamenti

- Diamo uno sguardo alle statistiche degli allineamenti:

```
$ samtools flagstat ht2_ch69.bam
```

- Se non volessimo gli allineamenti secondari?
  - samtools offre vari strumenti per filtrare gli allineamenti

```
$ samtools view -h -F 256 ht2_ch69.bam | samtools flagstat -
```






# Allineamenti “spliced” nei campi SAM

La CIGAR string riporta i gap degli allineamenti spliced

```
$ samtools view ht2_ch69.bam | cut -f 6 | grep N | less
```

6. CIGAR: CIGAR string. The CIGAR operations are given in the following table (set ‘\*’ if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	<u>skipped region from the reference</u>
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- H can only be present as the first and/or last operation.
  - S may only have H operations between them and the ends of the CIGAR string.
  - For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- 



# Multi-mapped reads nei campi SAM

Quali sono le read mappate in più loci (a.k.a. multi-mapped reads)?

Lo possiamo ricavare facilmente se è stato riportato il tag **NH** nei campi opzionali

(**NH**: Number of reported alignments that contains the query in the current record)

```
$ samtools view -F 4 ht2_ch69.bam | grep -v -w "NH:i:1" | less -S
```

Notare l'uso dell'opzione -v di grep che inverte la condizione di output. Senza tale opzione otterremmo le read allineate univocamente (uniquely aligned reads)

Per altri tag opzionali usati come standard <http://samtools.github.io/hts-specs/SAMtags.pdf>

Tag	Type	Description
AM	i	The smallest template-independent mapping quality of segments in the rest
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence
BQ	Z	Offset to base alignment quality (BAQ)
CC	Z	Reference name of the next hit
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read base qualities
CS	Z	Color read sequence
CT	Z	Complete read annotation tag, used for consensus annotation dummy features.
E2	Z	The 2nd most likely base calls
FI	i	The index of segment in the template

