Corso di Metodi Molecolari e Bioinformatica,
LM Biologia Evoluzionistica
A.A. 2018-2019
Università di Padova

Esercitazione 6

"Analisi di dati RNA-seq"

14 gennaio 2019

Docente: S. Bortoluzzi
Assistenti:
A. Coppe,
E. Gaffo (enrico.gaffo@unipd.it)

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

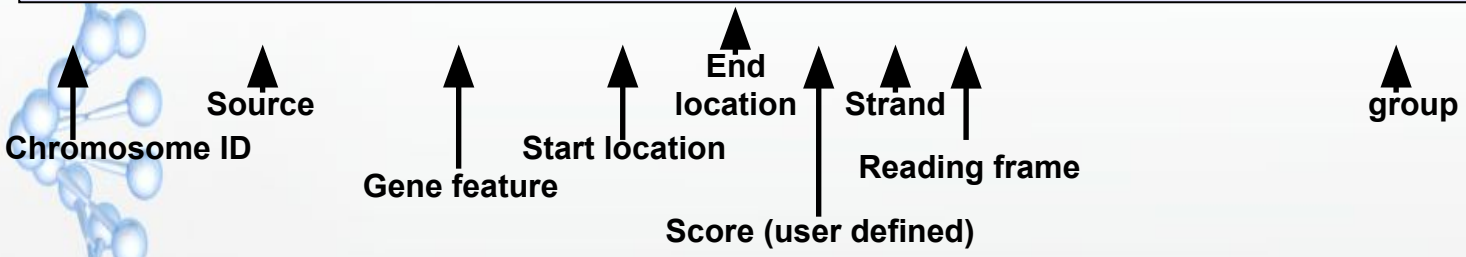
- Come nell'esercitazione precedente, abbiamo:
  - Genoma in FASTA:
    - CFLAR_HIPK3.fa
  - Le read sequenziate (paired-end) in FASTQ di 2 campioni:
    - SRR2923169_CFLAR_HIPK3_**1**.fastq
    - SRR2923169_CFLAR_HIPK3_**2**.fastq
    - SRR2923170_CFLAR_HIPK3_**1**.fastq
    - SRR2923170_CFLAR_HIPK3_**2**.fastq
- 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";
```



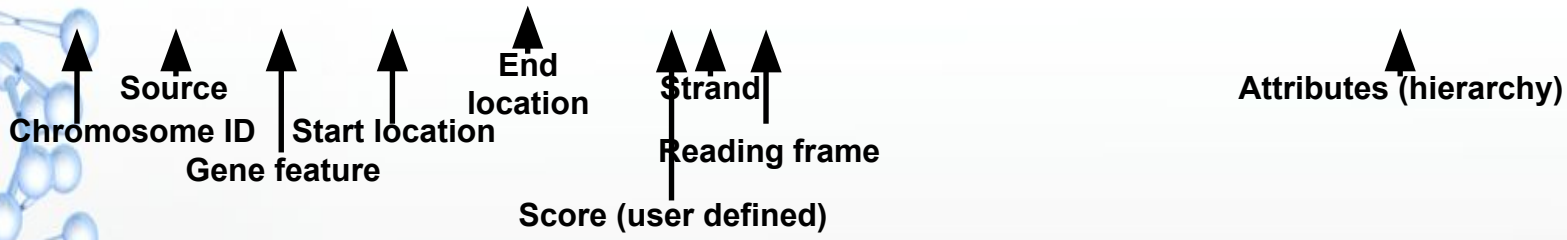
Gene Transfer Format (GTF)

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

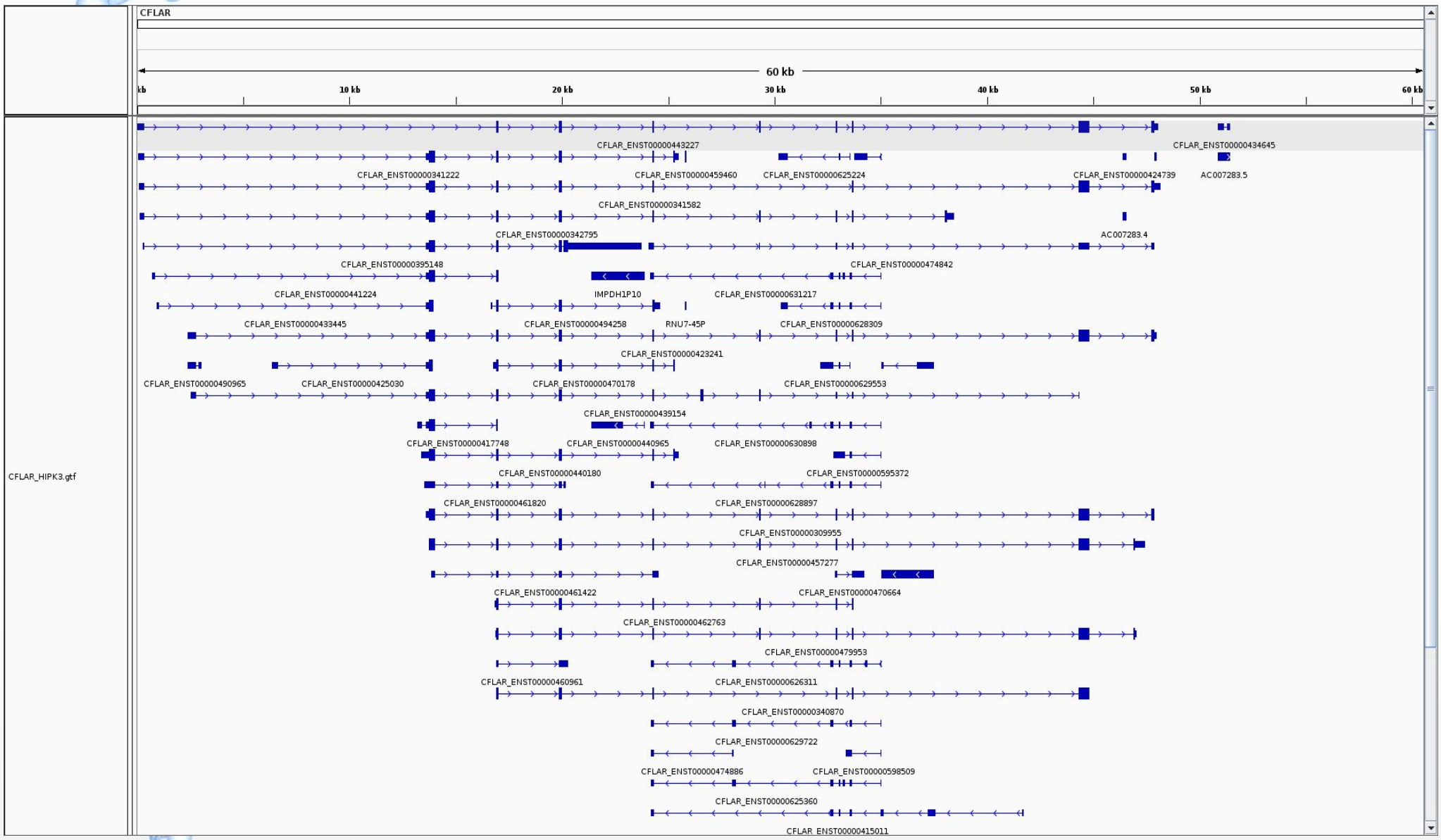
https://genome.ucsc.edu/FAQ/FAQformat.html#format1

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)

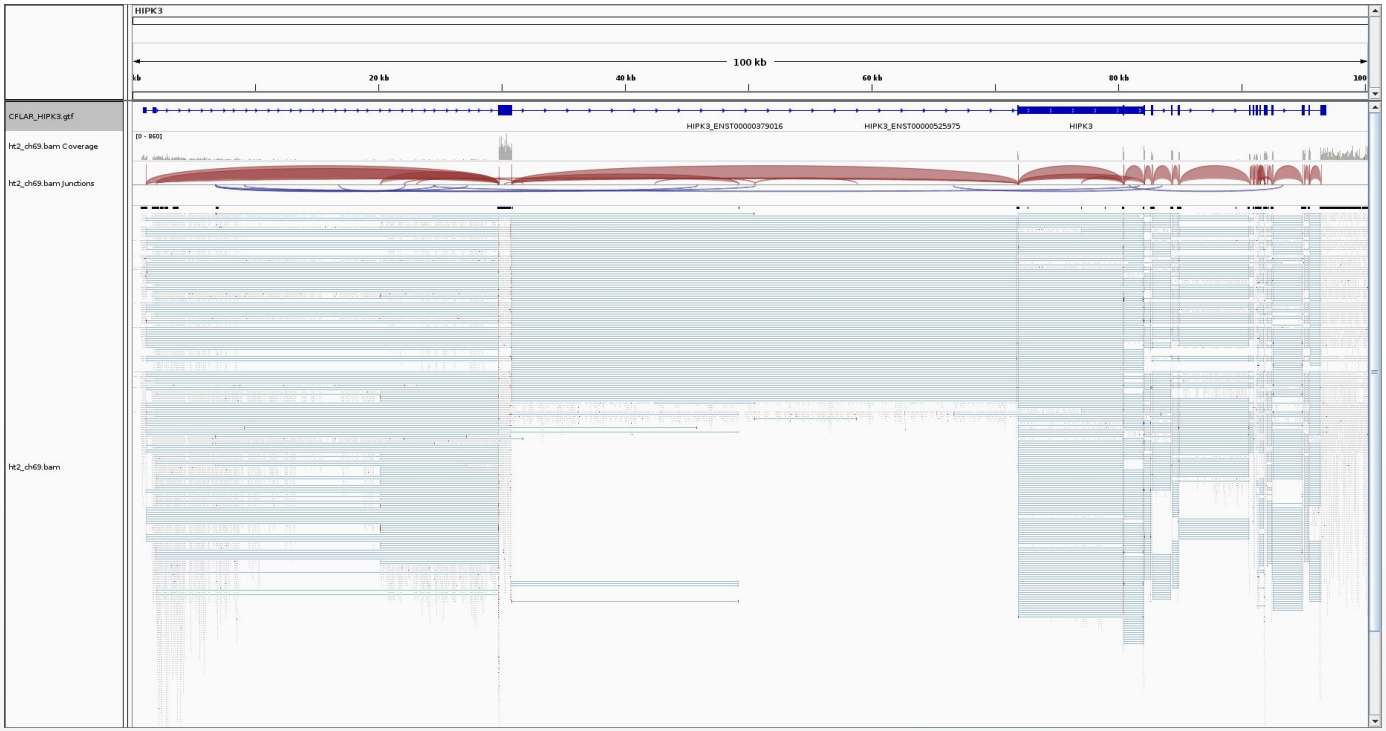
http://software.broadinstitute.org/software/igv/



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



Eseguire IGV e caricare i dati

Aprire un terminale e lanciare il comando:

- `igv.sh`

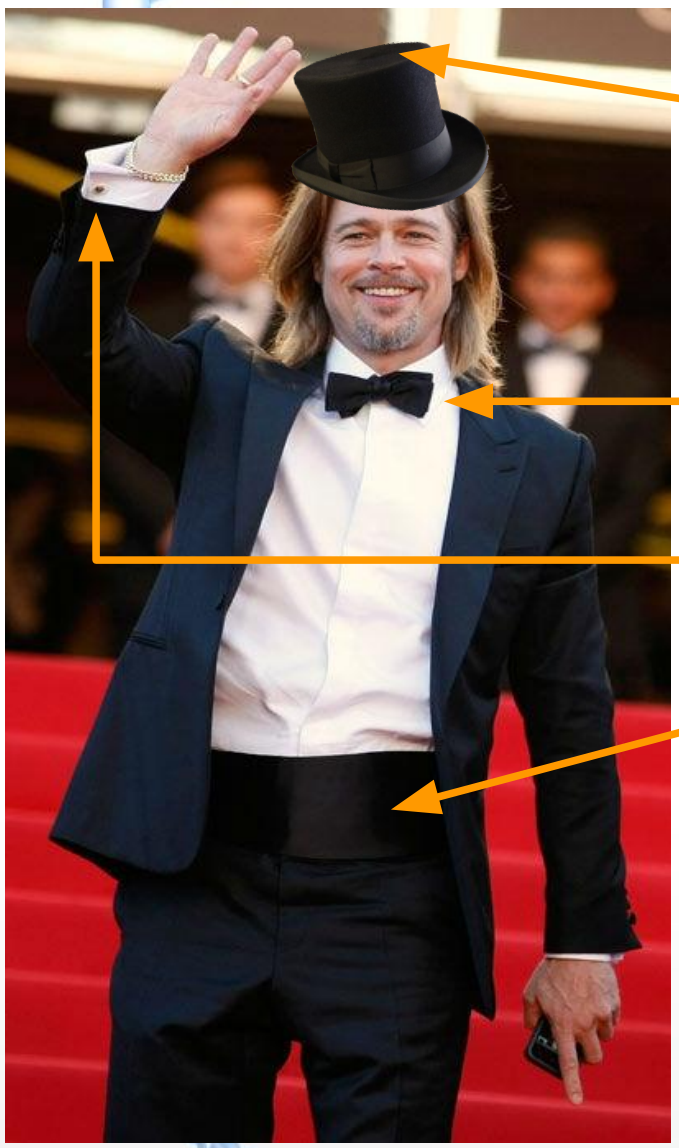
...e attendere un attimo che si carichi l'interfaccia grafica.

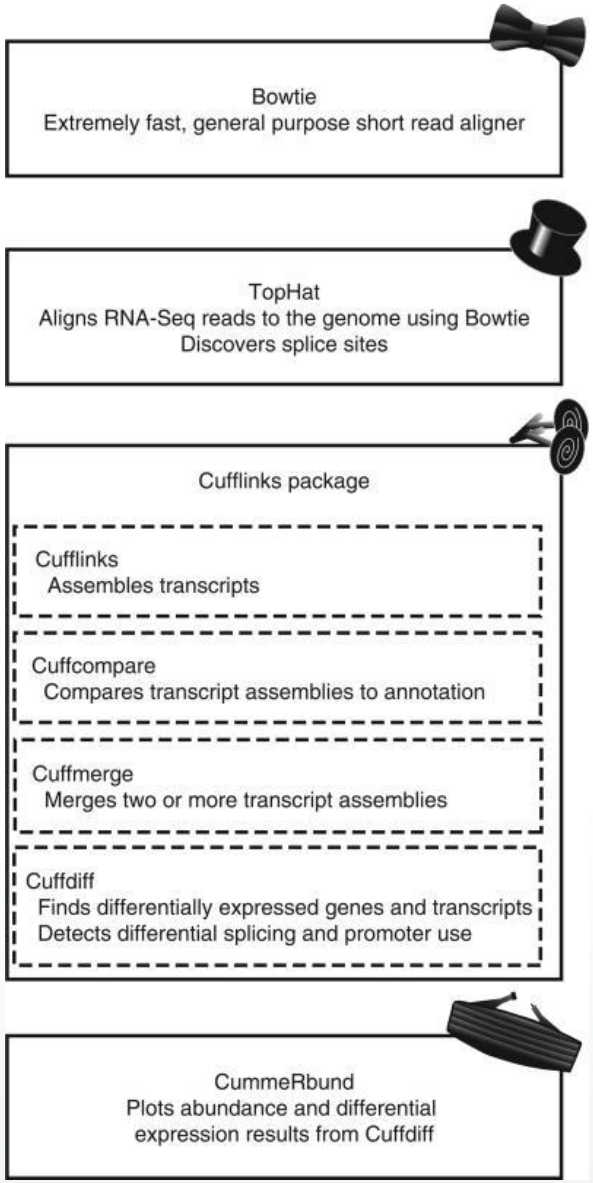
Dai menu a tendina:

- Caricare il file del genoma CFLAR_HIPK3.fa
- Caricare il file di annotazioni CFLAR_HIPK3.gtf

Quantificare l'espressione di geni/trascritti

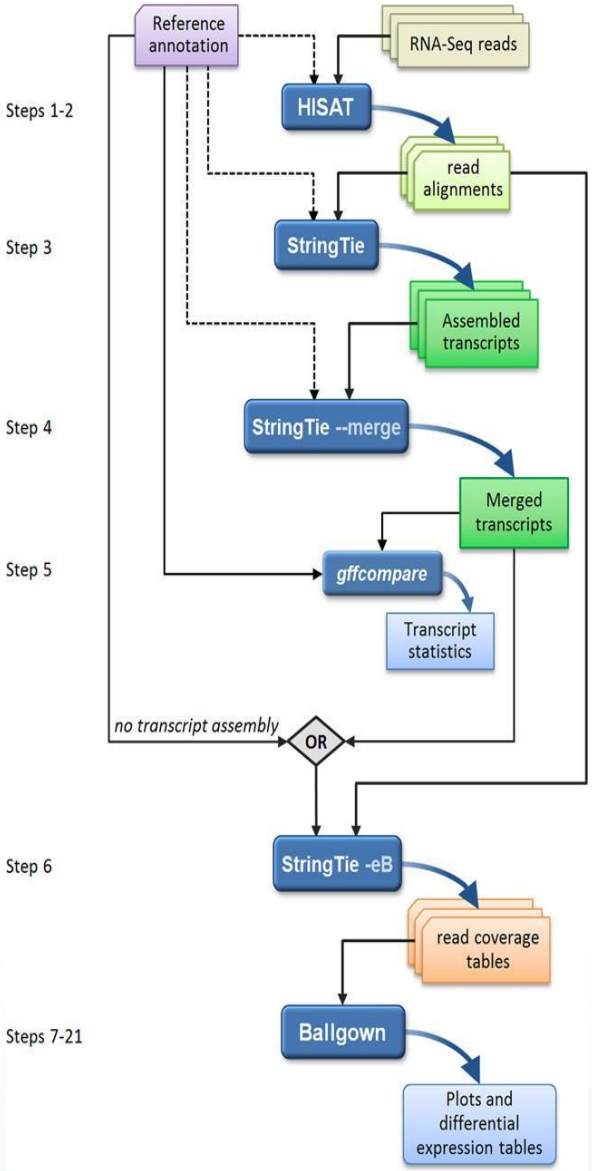
Utilizziamo i "Tuxedo tools"[1] (http://cole-trapnell-lab.github.io/cufflinks/)





[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).

Quantificare l'espressione di geni/trascritti: tool suite più recente

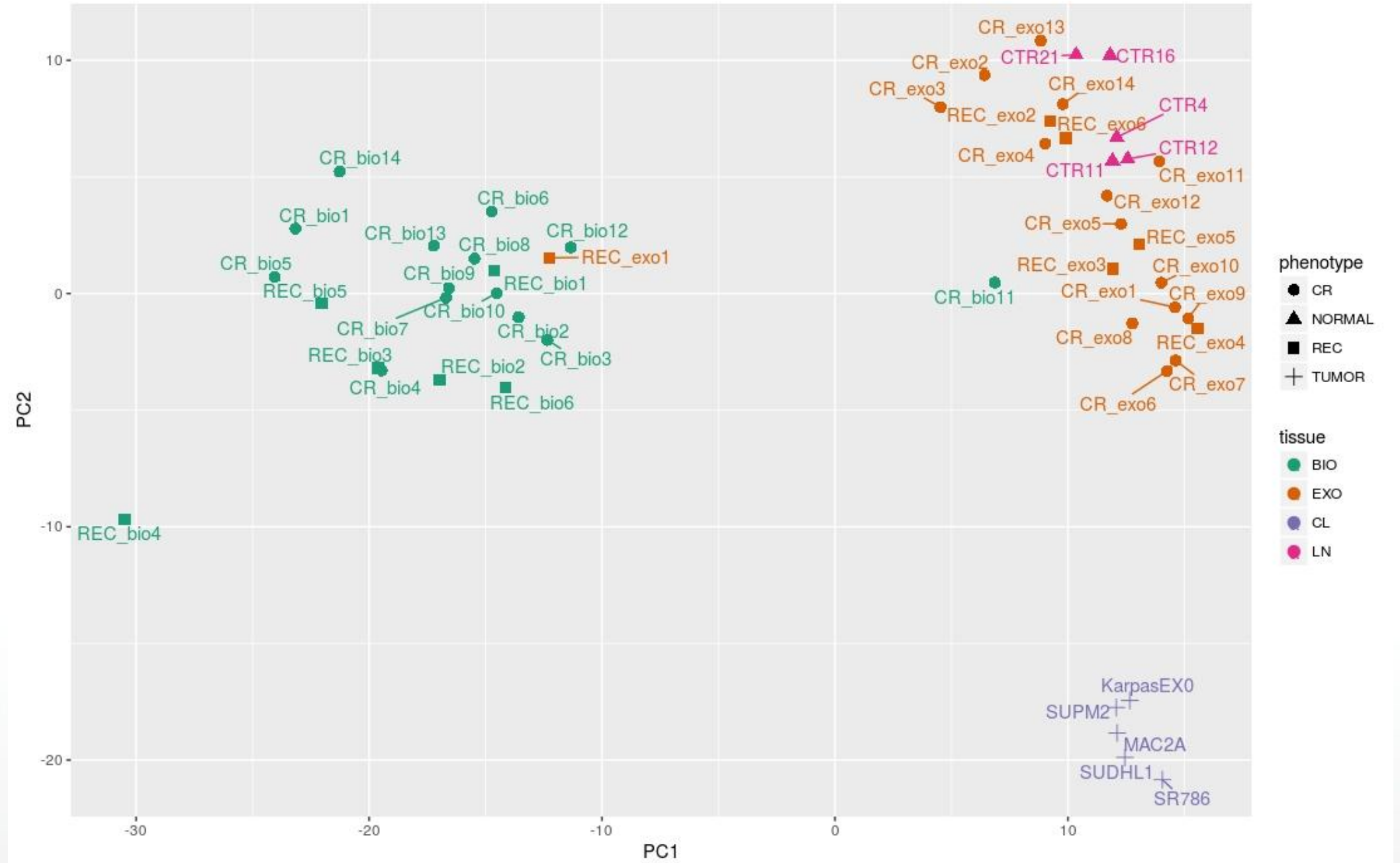


Updates of the tool suite[2]:

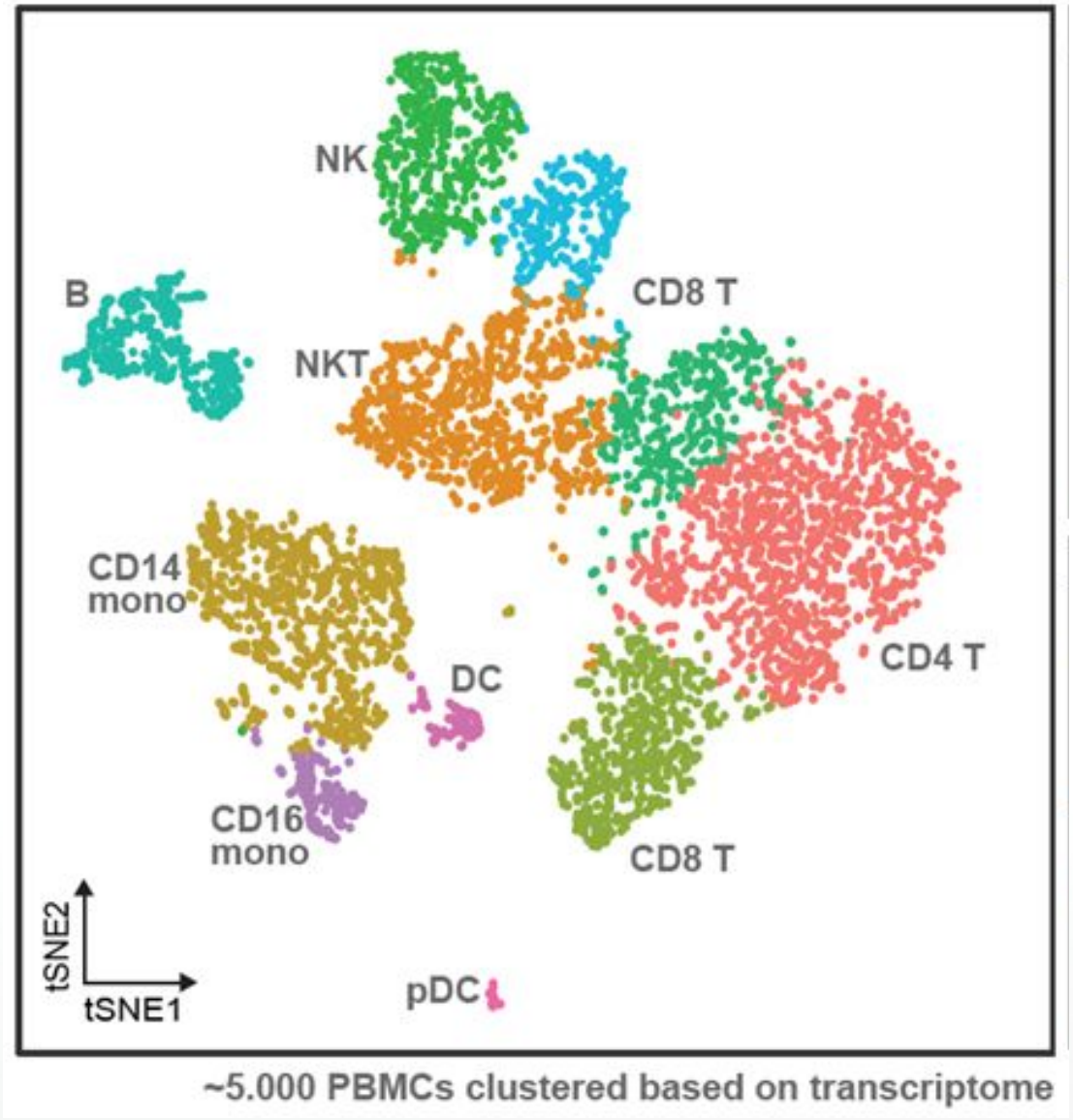
- TopHat → **HISAT**
- Cufflinks → **StringTie**
- CummeRbund→**Ballgown**

[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).

Es. di analisi dei dati di espressione: PCA



Es. 2: clustering di dati single-cell sequencing



~5.000 PBMCs clustered based on transcriptome

Allineamento read RNA-seq con HISAT2

https://ccb.jhu.edu/software/hisat2/index.shtml

"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 il BAM

- salvare in un BAM ordinato, per visualizzarle successivamente:
  - **samtools view** -bhS ht2_ch69.sam > ht2_ch69.bam
  - **samtools sort** -o ht2_ch69.bam asd > ht2_ch69.sorted.bam
- creare l'indice del BAM:
  - **samtools index** ht2_ch69.sorted.bam

Procedura “one-line”

- In alternativa, usiamo l’operatore di **pipe** per redirezionare gli output ai programmi samotools ed evitare di scrivere file intermedi:

  - **hisat2** -x CFLAR_HIPK3 -1 SRR2923169_CFLAR_HIPK3_1.fastq -2 SRR2923169_CFLAR_HIPK3_2.fastq --dta-cufflinks | **samtools view** -uhS - | **samtools sort** -o - asd > ht2_ch69.sorted.bam

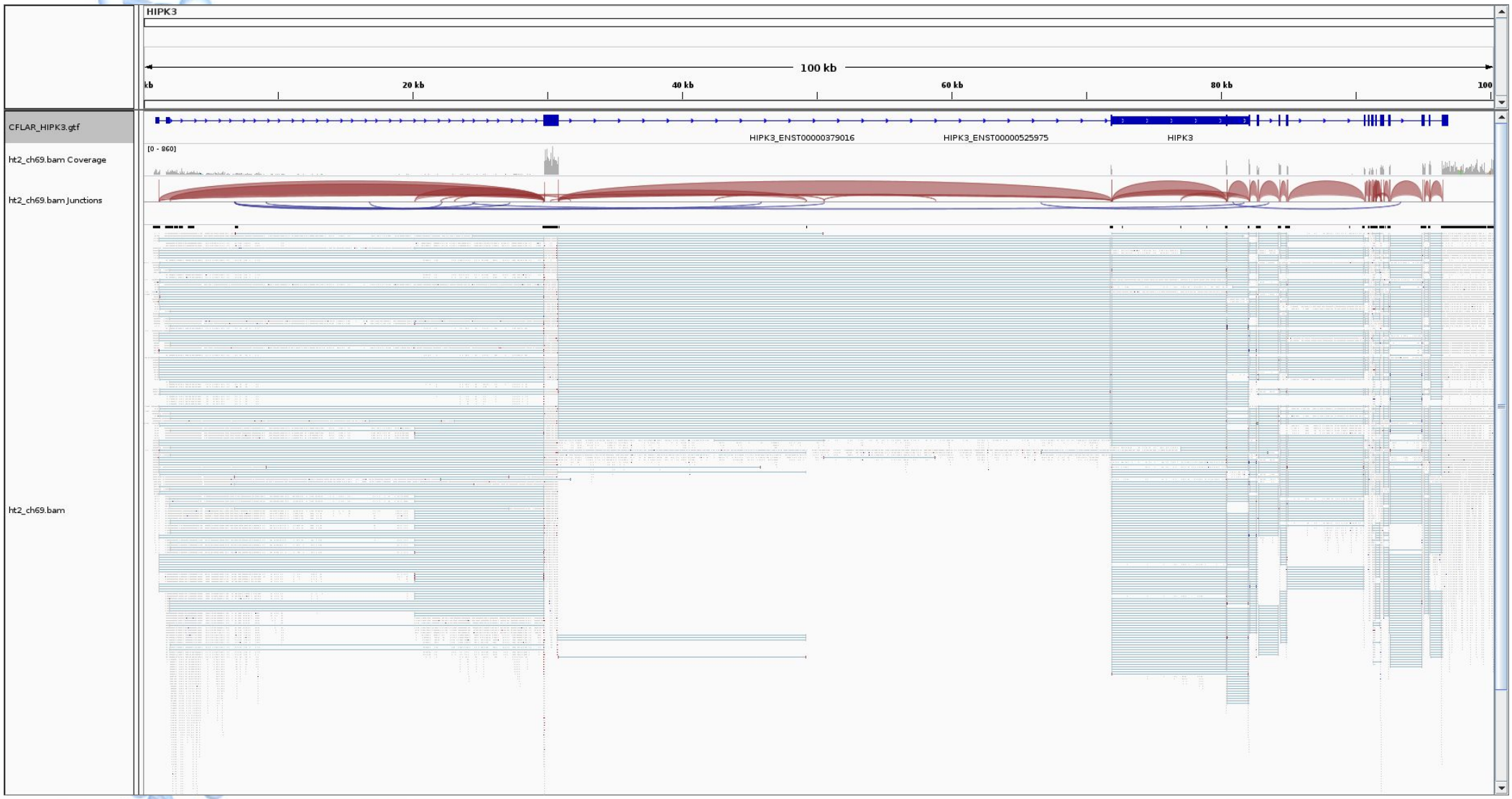
N.B: scrivere tutto in un’unica riga!

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

- ricordarsi di creare l’indice del BAM:

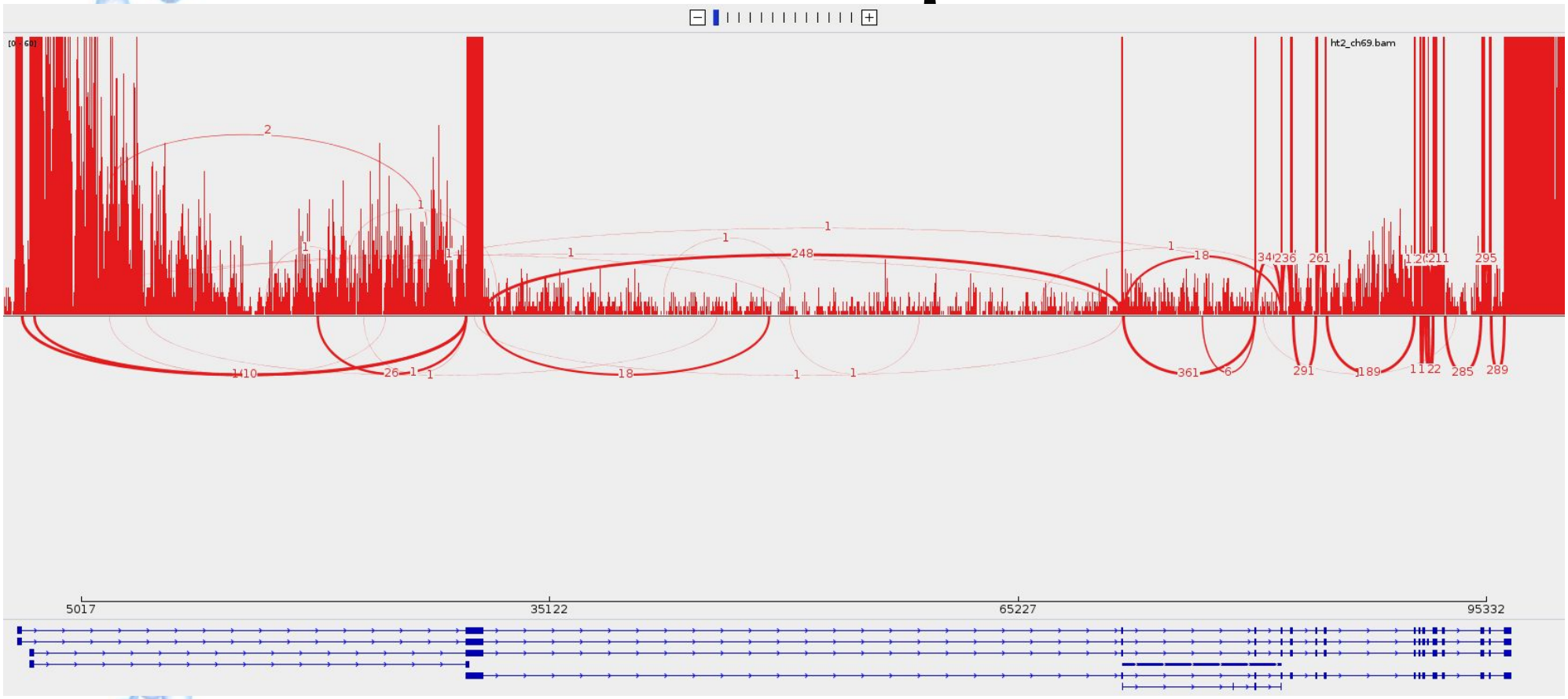
  - **samtools index** ht2_ch69.sorted.bam

Visualizzare 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….

Quantificare l'espressione dei geni/trascritti: comandi

- `mkdir geneexp`
- **`cufflinks`** `-G CFLAR_HIPK3.gtf -o geneexp/ht2_ch69.sorted.bam`

Otteniamo due tabelle con i valori di espressione di:

1. Geni → **genes.fpkm_tracking**
2. Trascritti → **isoforms.fpkm_tracking**

Li possiamo aprire con programmi per fogli di calcolo (es. Excel, Libreoffice Calc, ...)

Attenzione alla codifica: i numeri decimali sono definiti all'inglese, cioè con ".", non ","

Transcriptome reconstruction (con genoma di riferimento)

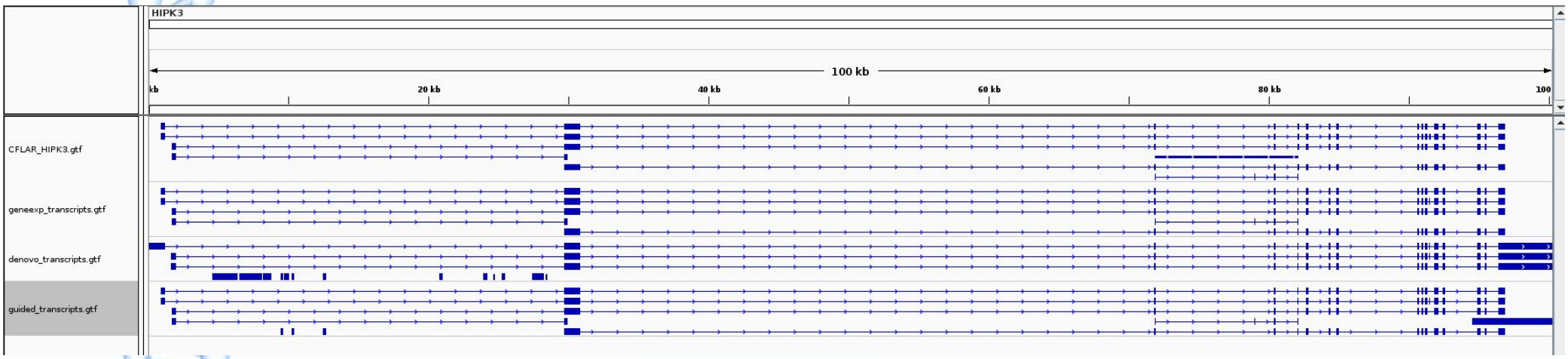
- ***senza annotazione***:
  - `mkdir noanno`
  - `cufflinks -o noanno/ht2_ch69.bam`

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

- **Guidato** (supportato da annotazione esistente):
  - `mkdir guided`
  - `cufflinks -g CFLAR_HIPK3.gtf -o guided/ ht2_ch69.bam`

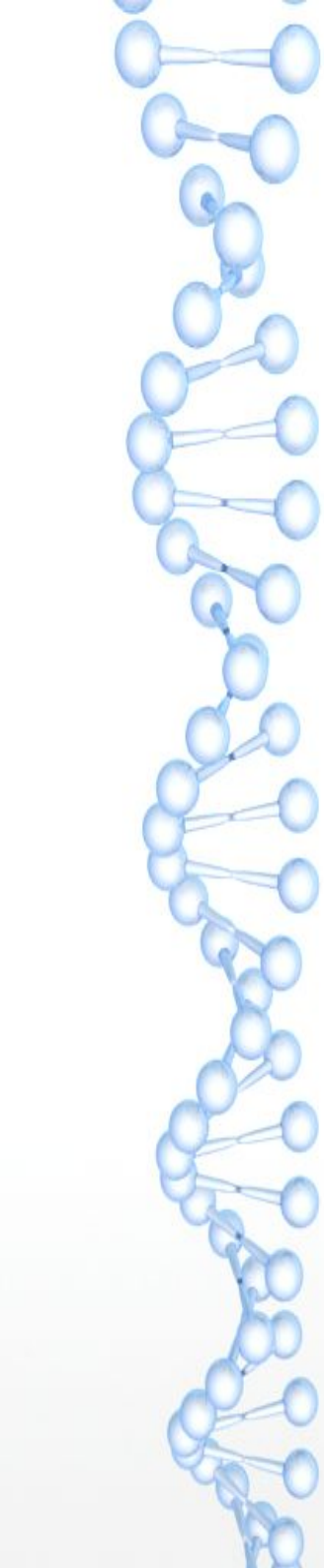
Confrontiamo i trascrittomi definiti con le varie opzioni, visualizzandoli con IGV

- geneexp/transcripts.gtf
- noanno/transcripts.gtf
- guided/transcripts.gtf



Esercizi

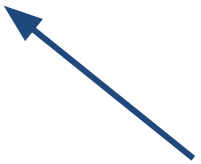
- quantificare l’espressione dei geni definiti nel trascrittoma ricostruito:
  - quali sono gli input?
- ripetere le analisi anche per il secondo campione



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 samotools sort):
  - `bwa mem CFLAR_HIPK3.fa -1 SRR2923169_CFLAR_HIPK3_1.fastq SRR2923169_CFLAR_HIPK3_2.fastq | samtools sort -O bam -T chch > ch69sorted.bam`
- Visualizziamo gli allineamenti con IGV:
  - Creare l'indice degli allineamenti prima di caricare il file in IGV:
    - `samtools index ch69sorted.bam`
- Diamo uno sguardo alle statistiche degli allineamenti:
  - `samtools flagstat ch69sorted.bam`



samtools >= v1.0

Altri possibili processamenti degli allineamenti

- Diamo uno sguardo alle statistiche degli allineamenti:
  - `samtools flagstat ht2_ch69.bam`
- Se non volessimo gli allineamenti secondari?
  - `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	skipped region from the reference
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