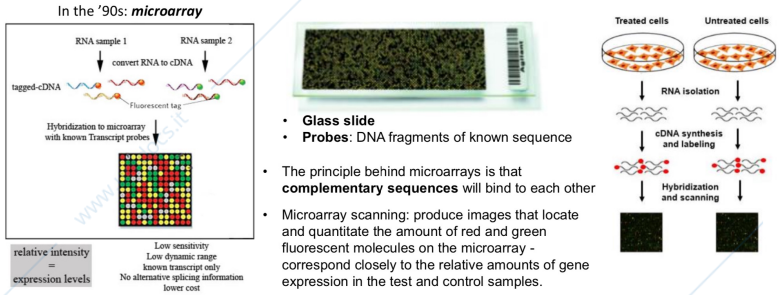
Single-cell omics: single-cell RNA-sequencing ✓

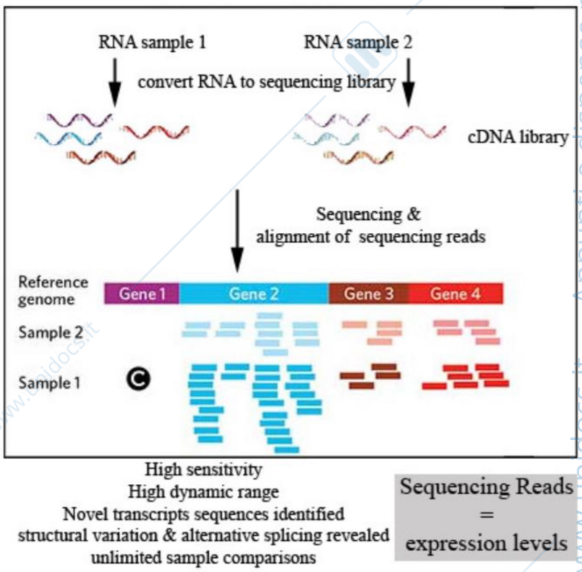
La biologia è molto complessa infatti nel nostro corpo ci sono 37,2 trilioni di cellule e ognuna è caratterizzata da un preciso profilo trascrizionale (insieme di geni espressi nella stessa cellula), che è direttamente associato al fenotipo e dunque alla funzionalità della cellula.

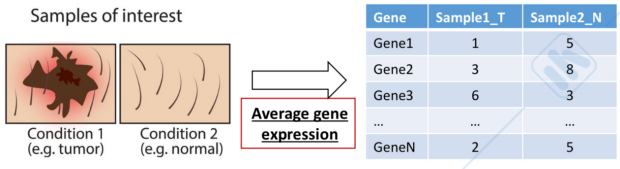
Come si è arrivati alla single-cell RNA-sequencing
Innanzitutto bisogna parlare di trascrittomica, ovvero, lo studio della trascrizione genica. Nella trascrittomica c'è stata una vera evoluzione in termini delle tecnologie usate e delle scoperte fatte.

1. Microarray: si tratta di un vetrino su cui sono depositate delle probes, ovvero, dei frammenti di DNA di sequenza nota. Immaginiamo di avere due condizioni ex: una linea cellulare trattata con un farmaco e una non trattata con un farmaco e si vuole studiare a livello dell'espressione genica cosa accade. Innanzitutto si effettua l'isolamento dell'RNA, la conversione in cDNA e il marcare questi frammenti con delle sonde fluorescenti caratterizzate da due colori diversi, poi avviene l'ibridazione, ovvero, il DNA target viene messo a contatto con frammenti di DNA presenti sui microarray, tale per cui le sequenze complementari si legano tra loro, infine avviene lo scanning per localizzare e quantificare le molecole fluorescenti rosse e verdi che corrispondono alla quantità relativa di espressione genica nei due campioni. Il vantaggio principale è il basso costo, tuttavia ha dei limiti: si possono studiare solo quei geni tali per cui abbiamo una sonda corrispondente sul microarray.

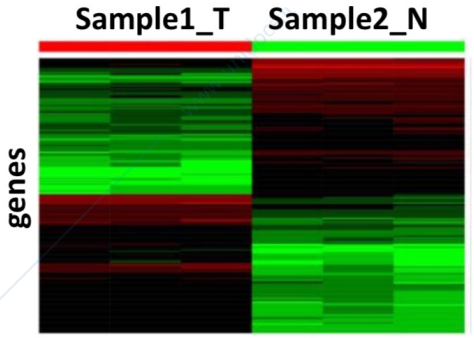


2. Bulk RNA-sequencing: si vuole studiare la trascrittomica di due situazioni. Si isola l'RNA, si converte in cDNA e poi si fa sequenziamento. Dalla macchina di sequenziamento si ottengono le sequencing reads, ovvero, migliaia di sequenze di DNA (FASTQ files). Il primo step è l'allineamento delle reads al genoma di riferimento. Il numero di reads allineato a ciascun gene è una rappresentazione del livello di espressione di quel gene, quindi più reads allineate al gene 2 vuol dire che il livello di espressione di quel gene è elevato. Si potrà poi fare un confronto tra campione 1 e 2e vedere i geni che vengono deregolati e quindi differenzialmente espressi tra le due condizioni. Ex: abbiamo due condizioni: una normale e una tumorale, abbiamo fatto il sequenziamento bulk abbiamo ottenuto le reads, le allineiamo al genoma e otteniamo una matrice di questo tipo: sulle righe tutti i geni e sulle colonne i campioni, i numeri corrispondono al numero di reads associate al gene nel campione. In questo caso, a differenza della single-cell, stiamo misurando un'espressione genica media che deriva da tutte le cellule presenti nel campione. Generalmente l'analisi che viene fatta è l'analisi di espressione differenziale in cui lo scopo è quello di andare a identificare i geni che sono differenzialmente regolati nelle due condizioni, in cui vediamo rappresentati sulle righe i geni e. Si vedono delle differenze di colore, che rappresenta l'intensità del gene e il rosso rappresenta una maggiore intensità quindi vediamo che i geni sopra sono più espressi nel tessuto sano quindi sono più down-regolati nel tessuto tumorale, mentre i geni sotto sono più espressi nel tessuto tumorale rispetto al sano. Questa analisi ci consente di fare delle ipotesi in particolare sui pathways biologici che vengono alterati.





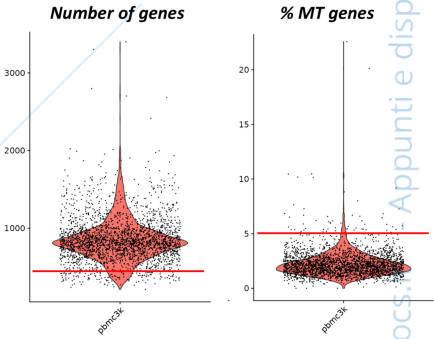
Gene	Sample1_T	Sample2_N
Gene1	1	5
Gene2	3	8
Gene3	6	3
...	...	...
GeneN	2	5



3. Per risolvere gli svantaggi relativi al bulk RNA-sequencing in cui si ha una misurazione media dell'espressione genica, si è iniziato a parlare di single-cell sequencing, dove non misuriamo più un'espressione genica media di tutte le cellule, ma misuriamo l'espressione genica nelle singole cellule. Quindi a partire dal tessuto si effettua una dissociazione delle cellule, dalle singole cellule si misura l'espressione genica. Quindi la single-cell sequencing è una tecnologia che ci permette di misurare l'espressione genica a risoluzione di singola cellula. Le applicazioni sono tantissime: ci permette di studiare l'eterogeneità cellulare e l'identificazione di popolazioni cellulari rare. I dati che si ottengono sono molto complessi per cui entra in gioco la bioinformatica. Quando si parla di scRNA-seq non si intende un solo metodo ma una collezione di protocolli ognuno dei quali on dei punti di forza e dei limiti. Nel 2009 è stato pubblicato un primo studio in cui si è dimostrata la possibiità di misurare la gene expression nelle singole cellule. 10x Genomics è la piattaforma più usata. Ciascun protocollo comprende 3 step principali:
    - Isolamento delle singole cellule
    - Preparazione della library
    - Sequenziamento

Steps principali dell'analisi dei dati: l'output della macchina di sequenziamento sono milioni di reads.

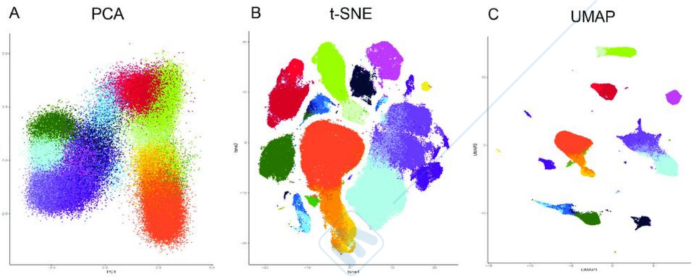
- Si genera la count matrix (tabella)
- Controllo qualità per rimuovere dai dati le cellule con bassa qualità che possono derivare da cellule che si sono rotte o morte durante il protocollo sperimentale. Come si identificano? Vengono usate alcune metriche che sono il numero di geni misurati in ciascuna cellula, perchè cellule di bassa qualità hanno molto spesso un numero molto più basso di geni rispetto ala media delle altre cellule, un'altra metrica che si usa è la % di geni mappati al genoma mitocondriale, perchè le cellule morte o di bassa qualità spesso esibiscono una contaminazione mitocondriale molto alta, questo è dovuto al fatto che nella cellula rotta si ha una fuoriuscita dell'RNA citoplasmatico e quindi si misura l'RNA mitocondriale. Si ottengono i cosiddetti grafici a violino. Nel grafico a sinistra i pallini rappresentano una cellula e sull'asse delle y si ha numero di geni. Nella maggior parte delle cellule si vede che c'è un numero di geni intorno a 1000. Nel grafico a destra i pallini sono le cellule e sull'asse delle y vediamo la % di geni mappati nel genoma mitocondriale, nella maggior parte delle cellule vi è una % di geni mitocondriali molto bassa (sotto 5%). Mettiamo delle soglie per dire che tutte le cellule che sono sotto (grafico sx) o sopra la soglia (grafico dx) sono di bassa qualità che vogliamo eliminare.



- Essendo questi dati di dimensioni elevate, per lavorarci c'è bisogno di due step:
    - Feature selection: il data set è filtrato per tenere solo quei geni informativi della variabilità dei dati, ex: immaginiamo di avere un gene con stesso livello di espressione in tutte le cellule, di fatto questo gene non mi dà alcuna informazione sulla variabilità delle cellule, quindi si può rimuovere dal data set. Questo metodo permette di velocizzare i dati.
    - Dimensionality reduction: proiettare i dati in uno spazio di dimensioni inferiori mantenendo le proprietà chiave dei dati originali. Lo scopo è di vedere i risultati su dei grafici di due dimensioni x e y. Vi sono dei punti colorati ce son delle cellule. Questi punti formano dei gruppetti, e questo indica il fatto che le cellule si raggruppano in clusters. Come è già stato detto come primo step si il clustering: le cellule si raggruppano in cluster e questo è il primo risultato intermedio di un analisi di single cell. Perchè le cellule verdi stanno solo con le versi e non con le rosse? Perchè il clustering è fatto sulla base della somiglianza dei profili di espressione genica delle cellule, questo vuol dire che tutte le cellule che stanno in un cluster avranno un profilo di espressione più simili tra loro rispetto alle cellule di un altro cluster.

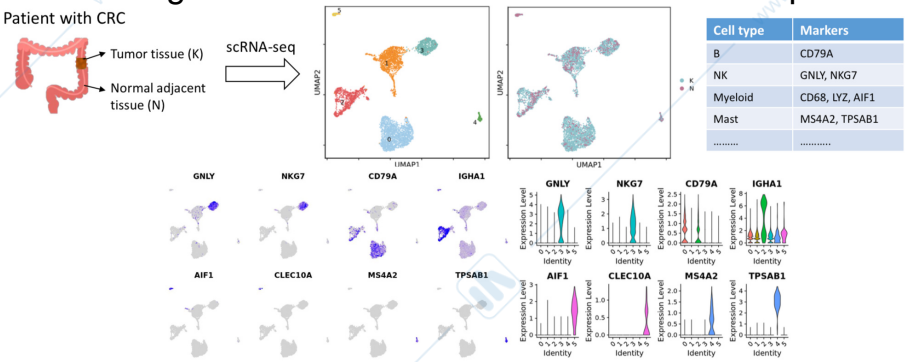
Quindi i vari raggruppamenti non sono altro che popolazioni cellulari diverse.

- PCA: analisi dei contenuti principali.
- t-SNE
- UMAP: più usato



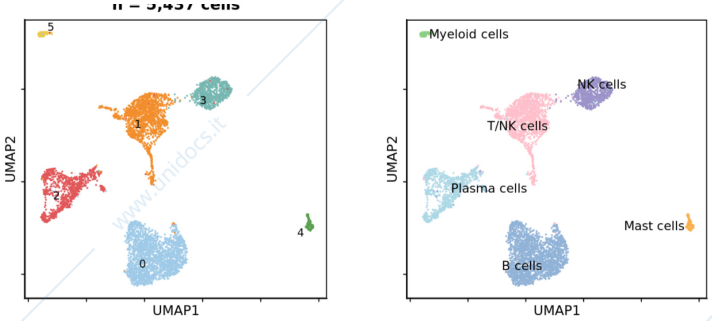
- Si usano degli algoritmi di clustering. Raggruppare le cellule in popolazioni cellulari. In base all'algoritmo si studiano i cluster per andare a vedere i geni caratteristici chiamati markers.

- Cell annotation, per denominare le cellule (cellule dendritiche, cellule T…). Si va a vedere il livello di espressione dei markers noti in letteratura per marcare alcune popolazioni cellulari. Ex: studiare dei campioni derivati da pazienti con tumore al colon, in particolare campione tumorale e campione sano proveniente dal tessuto sano adiacente è stato fatto un scRNA-seq e poi UMAP. Si va a vedere il livello di espressione dei marcatori per identificare le popolazioni cellulari. Si considera una lista di geni marcatori e si va a vedere la loro espressione nelle cellule. Si colorano le cellule dove il gene è maggiormente espresso. I risultati possono essere riportati anche sottoforma di grafici a violino dove x=clusters e y=livello di espressione del gene.

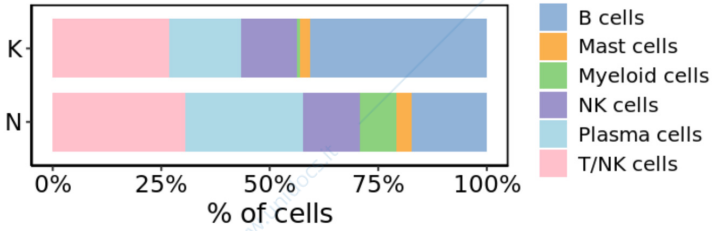


Cell type	Markers
B	CD79A
NK	GNLY, NKG7
Myeloid	CD68, LYZ, AIF1
Mast	MS4A2, TPSAB1
.....	.....

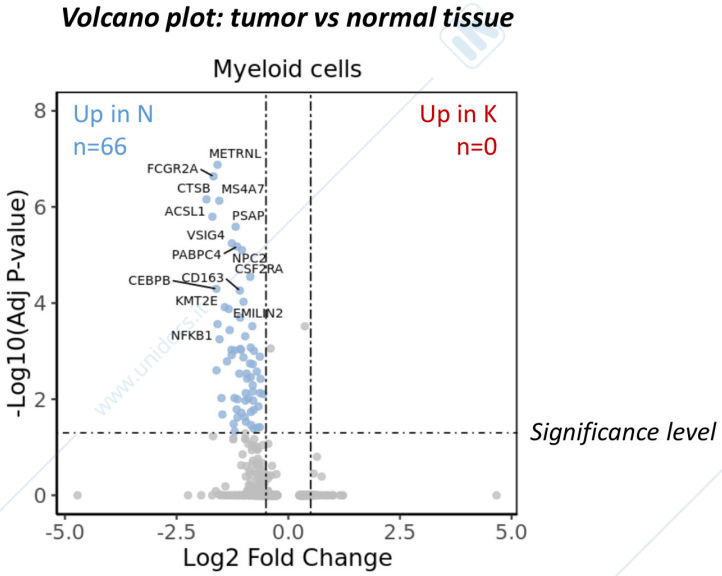
Attraverso questo studio si arriva a identificare le diverse popolazioni cellulari:



Si possono poi fare diverse analisi downstream sulla base delle domande biologiche.
Ex: vedere se c'è un arricchimento o una deplezione di alcune popolazioni cellulari nel tumore rispetto al sano, andando a determinare la % delle varie tipologie di cellule nei due campioni e si è visto che vi è un arricchimento delle cellule B nel tessuto tumorale in confronto al sano, mentre le cellule mieloidi sono deplete nel tessuto tumorale.

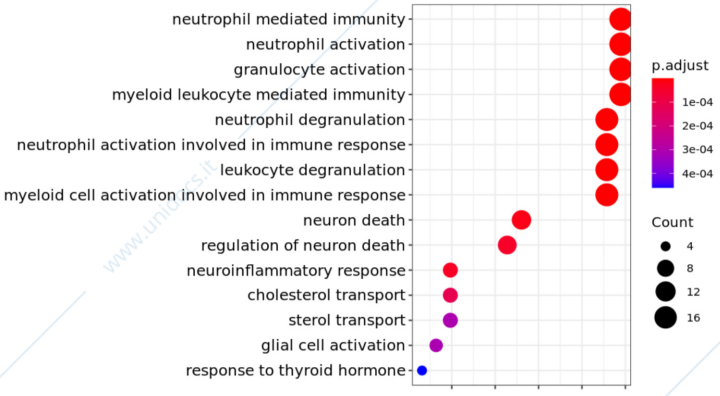


Ma a livello di espressione genica c'è una variazione nelle cellule mieloidi tra tessuto sano e tumorale? Per rispondere si fa analisi di espressione differenziale in cui confronto due condizioni per vedere se c'è una deregolazione nell'espressione genica tra le due condizioni. Per farlo si usa un metodo statistico, in particolare il volcano plot, y = significatività, cioè, -log10(P value corretto), il chè significa che il gene più si trova nella parte alta del grafico più il gene è significativamente deregolato tra le due condizioni. Solitamente si usa un livello di significatività p value corretto = 0.05. Si vede che tutti i geni (pallini) al di sotto di questo livello sono colorati in grigio e questo sta a indicare che questi geni non sono significativamente deregolati tra le due condizioni. L'asse x = log2 del fold change che rappresenta l'intensità della deregolazione, ovvero, quanto il gene è deregolato tra le due condizioni e più il gene sta all'estremità e più il gene è deregolato.

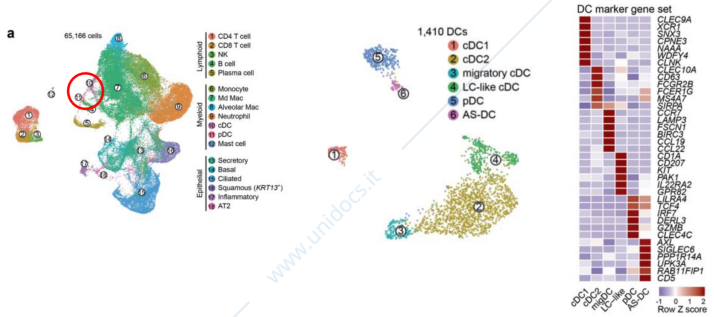


A destra vediamo che non ci sono pallini il che vuol dire che nessun gene è significativamente upregolato nel tumore. Dall'altra parte invece ci sono 66 geni upregolati nel sano e di conseguenza downregolati nel tumore.

Ora voglio capire in quali pathway biologici sono implicati questi geni o lo si fa ricercando nella letteratura scientifica, oppure attraverso dei metodi di analisi. Quest'analisi si chiama analisi dei pathway. Vi è sempre un valore di significatività (p value corretto), più il valore è piccolo più il risultato è significativo quindi più è rosso e più significativo e poi abbiamo un pallino di dimensioni variabili che rappresenta il numero di geni che sono coinvolti nei pathway. Si capisce che questi geni sono coinvolti in pathway immunologici principalmente. I dati si analizzano con R o Python.



Molto spesso viene effettuato un re-clustering: si vuole studiare le sottopopolazioni di un cluster, ex: cellule dendritiche, quindi riutilizziamo gli algoritmi di clustering solo sulle cellule dendritiche.



ALTRE SINGLE-CELL OMICS

Negli ultimi anni c'è stato un forte sviluppo delle tecnologie single-cell per studiare diversi aspetti della biologia di una cellula ex: modificazioni di istoni, accessibilità della cromatina, metilazione di DNA.

- CITE-seq: permette di studiare la trascrittomica e l'espressione di proteine di superficie
- scATAC-seq: tecnologia che studia l'accessibilità della cromatina

CITE-seq

È un metodo che consente di misurare simultaneamente sia il sequenziamento dell'RNA e dà informazione quantitativa sull'espressione di proteine di superficie sempre a livello di singole cellule grazie all'uso di anticorpi.
Sull'asse delle x non abbiamo i geni come nella sc-RNA seq ma abbiamo le proteine.

	Cell 1	Cell 2
Protein X	100	1000
Protein Y	400	30

Può essere utile perchè in alcuni casi vediamo che c'è corrispondenza tra l'espressione dell'RNA e della proteina in altri però non è così. Ex: l'espressione dell'RNA di CD8 è oltre debole mentre quello della proteine è più forte e in questo modo risulta essere più facile identificare una popolazione di proteine.

scATAC-seq

È importante studiare l'accessibilità della cromatina perchè l'identità cellulare è influenzata dall'epigenetica. Ci fornisce i profili delle regioni aperte e accessibili della cromatica che sono indici di una regione regolatoria attiva e i attiva trascrizione.
Sulle righe non abbiamo ne geni ne proteine ma i picchi (una regione del DNA) nella matrice. A partire dalla matrice si fanno diverse analisi:

- Visualizzazione con UMAP
- Clustering: raggruppamento dipende dalla somiglianza dei profili di accessibilità della cromatina delle cellule
- Analisi di accessibilità differenziale: vediamo com varia l'accessibilità a queste regioni tra due popolazioni cellulari

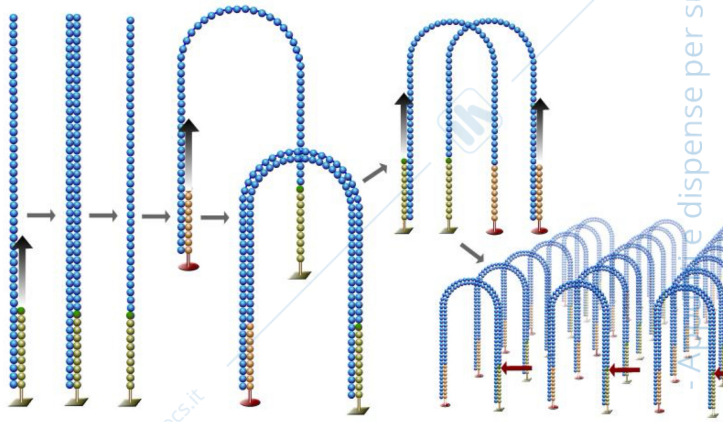
Lezione 9 ottobre

Analisi trascrittomica

Trascrittomica: Misura della quantità di trascritti
Sì è passati dalla RT-PCR a PCR quantitativa con cui si è in grado di misurare in modo preciso la quantità di trascritti in cellule bersaglio. Ha un'alta sensibilità quindi rileva anche piccole quantità di trascritto.
Svantaggio: non si può utilizzare per quantificare i trascritti di tanti geni.

Per analizzare l'intero panorama dei trascritti nella cellula target sono nati gli array di cDNA -> in un solo esperimenti possiamo analizzare migliaia di trascritti. Limiti: c'è bisogna di una grande quantità di trascritto di partenza e non ci danno informazioni su splicing o su sequenza dei singoli nucleotidi presenti sui trascritti.
Ad oggi si usano le tecniche di RNAseq basate su NGS.

Tecnologia di illumina: si usa una cella a flusso-> cella di vetro scavata da canali che attraversano per tutta la lunghezza la flowcell e hanno un foro di entrata e di uscita. Nei canali la superficie interna è rivestita da due olidonucleotidi diversi.
Immaginiamo di voler sequenziare un intero genoma umano: prendere cellule di interesse si estrae DNA e lo frammentiamo in modo casuale (ex: sonicazione). Poi leghiamo due adattatori diversi alle due estremità dei frammenti e otteniamo la library di DNA genomico. Prendo la libreria la denaturo e carico i filamenti di DNA nei canali della flowcell, gli adattatori sono complementari ai primer della cella di flusso. Per cui l'adattatore lega un primer sulla flowcell, e con una polimerasi si sintetizza il filamento complementare. Poi il filamento originale (legato in maniera non covalente alla flowcell) viene lavato via e quello che rimane fa interagire il suo adattatore rimasto libero con il primer formando un ponte e poi la polimerasi sintetizza il filamento complementare. Si ottiene un double-strand bridge che viene denaturato ottenendo due filamenti singoli di DNA attaccati al flowcell. In questo modo ho ottenuto due copie dello stesso frammento. Questo processo viene ripetuto tante volte per ottenere a partire da un singolo filamento un cluster di DNA. Viene fatto al posto della PCR perché lo scopo oltre ad essere quello di amplificare è anche quello di mantenere fisicamente distinti i cluster.



Effettuo questa amplificazione per aumentare il segnale emesso e ottenere un dato facilmente leggibile.
A questo punto sequenzio utilizzando due tecniche diverse:

- Single-read: sequenzio in una sola direzione
- Paired-end: sequenzio in entrambe le direzioni (ho più dati)

Sequenziamento:
Utilizzo primer di innesco e solo nucleotidi marcati con fluorocromo e bloccati al 3'.
Introdotto il primo nucleotide la reazione finisce, colpisco la cella con luce laser e registro l'emissione di luce da ognuno dei cluster presenti e in base alla luce emessa posso capire qual è il nucleotide che è stato incorporato in ciascun cluster. A questo punto devo rimuovere il blocco e togliere il segnale di fluorescenza e viene internalizzato il secondo nucleotide.
È vantaggioso perché possiamo sequenziare parallelamente un numero enorme di molecole di DNA.

La NGS posso usarla anche per quantificare l'RNA in una cellula e per farlo prima lo retrotrascrivo a cDNA, ottenerne un double strand ed effettuare questa tecnica. Se io ho n RNA dell' actina di conseguenza ho n cluster che mappano su actina. E questo possono farlo in parallelo per più geni umani. Ottengo una serie di reads (ogni read si riferisce ad un cluster) che mappano su di un gene.

Immaginiamo di avere le cellule A1 (cellule tumorali) e B1 (controllo) e di voler capire in cosa differiscono nell'espressione genica.
Dato che sto facendo una analisi statistica ho bisogno di replicati (almeno 3) e tanto più è eterogeneo il campione di interesse tanto maggiore deve essere il numero di replicati.
Immaginiamo di voler caratterizzare l'effetto di un gene X in termini di trascritto: introduciamo il gene nella cellula in modo da over-esprimere il gene e andare a vedere cosa succede nella cellula facciamo lo stesso in una cellula di controllo. Coltiviamo le cellule su 3 terreni di coltura e da ognuna si prepara la library di NGS e vediamo un numero elevati di geni differenziali rispetto ai controlli che però non sono confermati di campion primari (approccio sbagliato).

L'approccio corretto: da una linea generiamo 3 linee indipendenti in cui transfetto il gene che in inserirà in regioni diverse del genoma portando a patterns diversi e in questo modo riduco il rischio di falsi positivi.

Esperimenti di trascrittomica:
Scegliere il tipo di esperimento di trascrittomica usare. Ne abbiamo due:
- uno in cui purifichiamo l'mRNA con coda di poli-A. Il vantaggio è che ci focalizziamo solo su mRNA che codifica su proteine. Lo svantaggio è che perdiamo le info su tutti gli altri RNA
- Si considera tutto l'mRNA. Il vantaggio è che non ci limitiamo solo all'mRNA il problema è che la maggior parte di RNA è formato da rRNA, per cui si fa una deplezione di rRNA.

Bisogna capire quanti cluster usare:
- minimo 20 milioni
- Max 100 milioni

Passando da 20 a 100 aumentano i costi e il vantaggio è che possiamo leggere meglio geni espressi a basso livello e si rilevano eventi di splicing differenziale

Bisogna scegliere se fare sequenziamento single read o paired read.

Come analizzo i dati
Si usa la principal component analysis: si prende un dato che è quello della trascrittomica in 20000 dimensioni e schiacciarlo in uno spazio a due dimensioni.
Immaginiamo di aver ottenuto due dati relativi a solo due geni (A su X e B su Y). PCA crea un asse che massimizza la varianza dei campioni cioè cerca di far in modo che la proiezione dei campioni lungo l'asse sia la più diffusa possibile, se si proiettano invece su un asse che è perpendicolare a quello precedente si ha uno spread minore. La PCA ruota questo asse di poco sempre di più per misurare la varianza proiettata sull'asse e registra la varianza ad ogni configurazione. La configurazione che aumenta al massimo la varianza viene scelto come asse numero 1 detto anche principal component 1 e proietta i campioni su quest'asse. Questo viene fatto per scartare i geni che hanno un'espressione simile in tutti i campioni.

Log2 fold change: Differenza di espressione tra casi e controlli in scala logaritmica

17

RNA-Seq: geni differenziali (DEG)

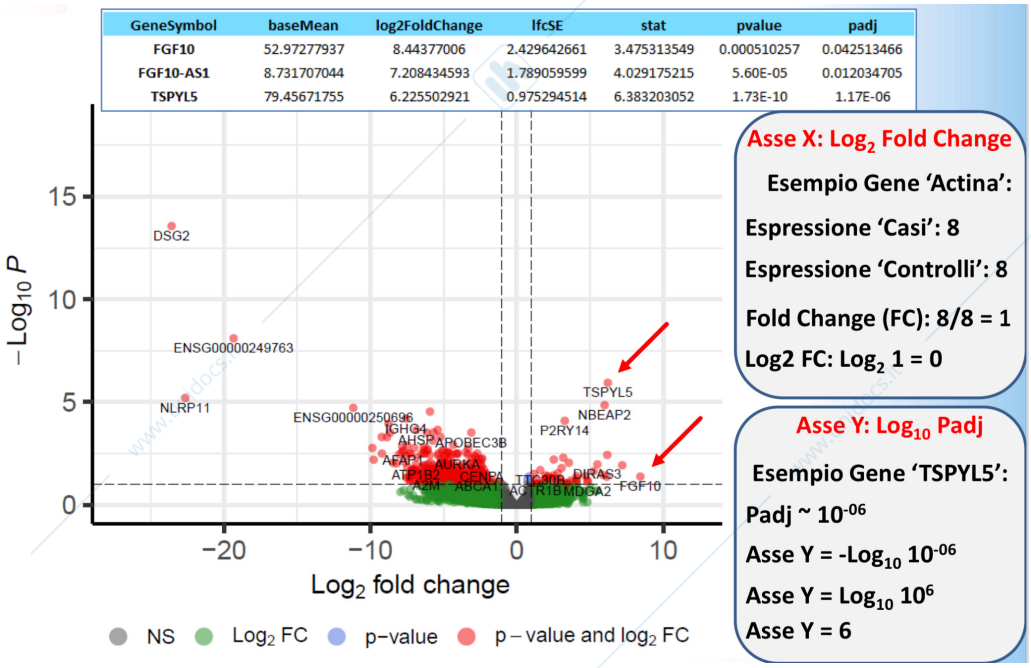
GeneSymbol	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FGF10	52.97277937	8.44377006	2.429642661	3.475313549	0.000510257	0.042513466
FGF10-AS1	8.731707044	7.208434593	1.789059599	4.029175215	5.60E-05	0.012034705
TSPYL5	79.45671755	6.225502921	0.975294514	6.383203052	1.73E-10	1.17E-06
ENPP5	31.12797975	6.184700999	1.399752449	4.418424847	9.94E-06	0.003778425
TTC23L	4.881465952	6.15782943	1.775928161	3.467386555	0.000525546	0.043386464
NBEAP2	20.77280788	6.003944207	1.016613396	5.905828342	3.51E-09	1.42E-05
STOML3	18.7183532	5.892410813	1.649463996	3.572318539	0.000353835	0.035429008
DIRAS3	30.87502966	5.507981106	1.354428957	4.066644527	4.77E-05	0.01094141
PTGFR	147.1980966	5.247621313	1.403373827	3.739289711	0.000184541	0.023926444
HNRNPKP1	12.08199275	4.868480604	1.493621866	3.259513479	0.001116035	0.067583591

Gene Symbol: gene su cui è stata applicata l'analisi statistica

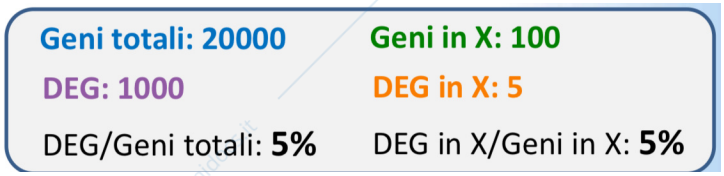
Log_2 FoldChange: rapporto tra l'espressione media di un gene nei casi (numeratore) e nei controlli (denominatore), espresso in scala logaritmica in base 2

padj: p-value 'aggiustato' per test statistici multipli. Visto che nell'analisi DEG applichiamo migliaia di test statistici (uno per gene), il solo p-value standard genererebbe troppi falsi positivi (FP). Applichiamo quindi correzioni ad hoc (e.g. Benjamini-Hochberg FDR o Bonferroni) per controllare i FP

Ottengo un volcano plot:
Sullo 0 abbiamo geni che non hanno differenza significativa, se un gene ha un'espressione 256 volte maggiore nei casi rispetto ai controlli il log2 sarà pari a 8 se invece è minore il log sarà -8.
Il gene essendo in basso vuol dire che è statisticamente significativo ma il p adj non è particolarmente basso quindi si può dire che FGF10 è un gene che è fortemente differenziale in termini di log 2 FC ma in termini di p value no. Quindi è maggiormente espresso nei casi rispetto ai controlli ma con una deviazione standard molto alta -> misura non precisa. I geni che mi interessano di più sono quelli in alto a destra e in alto a sinistra (fortemente differenziali e hanno una statistica differenziale solida).



Analisi di pathway: si parte da una lista di geni differenziali per arrivare a pathway differenziali per capire cosa sta succedendo nelle cellule di interesse.
A partire da tutti i geni del genoma individuo quelli differenziali. Poi considero i geni di un pathway X e vado a vedere tra tutti i geni differenziali quanti sono presenti nel pathway X.



-> non ho arricchimento