Commenti e riassunto della review di Science

Introduzione

Le analisi filogenetiche sull'evoluzione del cancro nascono con lo scopo di capirne la progressione, di cercare di prevederla ed infine di controllarla.

Le mutazioni del cancro differiscono rispetto a quelle nelle specie per 4 diversi fattori:

- i tipi di mutazioni (errori)
- la loro frequenza
- l'intensità della selezione (se non la sua inesistenza)
- la grande eterogeneità dei sottocoloni

Fenotipi di ipermutabilità:

- instabilità cromosomica dovuta al malfunzionamento della proteina p53
- instabilità microsatellite (DNA Mismatch repair anormale, produzione di sequenze GC/GA ripetute dovute alla mancata correzione degli errori)
- grande quantità di mutazioni puntuali (es: disfunzione proteine APOBEC, responsabili della risposta antivirale, in particolare contro i retrovirus)
- CNVs (dove vi sono cambiamenti con scala e locazione diversi a seconda del meccanismo che le ha indotte, come BFB breakage-fusion-bridge)
- Katageis, dove diverse SNVs accadono in una regione cromosomica ristretta (a volte causati dalla sovraespressione della proteina APOBEC3B)
- Chromothripsis, dove un cromosoma si distrugge e viene riarrangiato in modo apparentemente casuale (si crede che fenomeni come questo possano essere delle alternative alla teoria evolutiva del cancro)
- chromoplexy, dove del dna passa da un cromosoma all'altro a seguito di una catena di riarrangiamenti dovuti a diversi eventi BFB

> BFB: evento in cui i telomeri di un cromosoma si rompono, alla replicazione il cromosoma genererà due fratelli anch'essi privi di telomeri che, proprio per questo si fondono alle estremità. Al momento della replicazione, i centromeri dei due fratelli sono tirati in direzione opposta, questo causa una rottura in una posizione causle, così le due cellule figlie non avranno il cromosoma corretto. Il

> processo poi si ripete ad ogni duplicazione dato che neppure i due cromosomi suddivisi hanno telomeri.

Il modo in cui le mutazioni di tipo SNV si accumulano differisce da paziente a paziente ed inoltre in funzione del **trattamento**. Infatti questi possono causare rotture a doppio filamento influendo loro stesse nel panorama delle mutazioni, oppure, viceversa, possono sopprimere l'ipermutabilità.

E' stato spesso assunto che la selezione e la differenziazione delle cellule cancerose sia guidata da un processo evolutivo dinamico che le porta ad adattarsi all'ambiente. Altri studi hanno però mostrato come queste considerazioni giochino un ruolo secondario nell'evoluzione tumorale, in contrasto con l'evoluzione delle specie. Gli elevati livelli di **eterogeneità nelle sub-popolazioni colonali** fanno infatti pensare a una condizione di "equilibrio puntuale" piuttosto che alla sopravvivenza del più forte, anche se alcuni tumori appaiono proprio privi di selezione in mancanza di un trattamento (mentre, in caso di trattamento, c'è accordo sul fatto che il progresso sia evolutivo).

L'alta eterogneità pre-trattamento gioca un ruolo di rilievo nella soppravvivenza del tumore, garantendo che questo riesca a sopravvivere favorendo la propagazione delle sue sub-popolazioni adatte alla nuova condizione (comprovato dal fatto che spesso la progressione alla forma metastatica e la resistenza ai farmaci deriva proprio dalle popolazioni poco frequenti nelle fasi iniziali).

L'analisi filogenetica dei tumori

Il riconoscimento del fatto che l'evoluzione abbia un ruolo in quella dei tumori ha portato alla formazione di un'estensione della filogenetica verso l'analisi del cancro, giungendo alla creazione di quello che si può definire un nuovo campo. Dopo i primi passi pionieristici (Tsao, idea, Desper, applicazione) lo scopo di questa materia è rimasto quello di riuscire a ricostruire l'evoluzione tumorale sfruttando l'informazione data dalle variazioni geniche, spesso prorio mediante alberi filogenetici.

Questi metodi si differenziano per diverse caratteristiche:

- **Design dello studio**: influenza il tipo di dato utilizzato per l'analisi, possono essere svolti su più tumori di diversi pazienti, bulk su un singolo paziente o single-cell su un singolo paziente.
- **Tipologia dei dati analizzati**: marker tumorali, come CGH a larga scala (per analizzare CNVs) o FISH (fluorecense in situ hybridization), oppure metodi basati sul Next Generation Sequencing per l'estrazione di SNVs, CNVs, espressione genica, metilazione o altri marker istonici (modifiche alla struttura cromosomica).
- **Modello matematico**: il modello utilizzato incide sull'applicabilità e l'adesione del metodo filogenetico ai dati, per esempio supportando o meno certi tipi di mutazioni come le SVs (structural variants, come le CNVs) o assumendo determinati processi evolutvi.

- **Differenze negli algoritmi applicati**: diversi studi sono basati su metodologie che venivano già impiegate per la filogenetica delle specie (metodi di massima parsimonia, minima evoluzione, neighbour joining, UPGMA (Unweighted Pair Group Method with Arithmetic Mean, massima verosimiglianza, inferenza probabilistica bayesiana) a volte combinate. Di rado gli algoritmi hanno implementazioni che tengano conto della pecularità nell'evoluzione del tumore.

Gli **alberi evolutivi** sono centrali in questo genere di studi; utilizzati fin dai primi esperimenti che hanno provato ad applicare questi sistemi per trovare le mutazioni guida nel cancro o di ordinarle temporalmente, fino a quelli che ora mettono in discussione se l'evoluzione del tumore possa essere effettivamente lineare, come nel caso delle specie, o se invece abbia un comportamento ramificato. Il tutto spesso con lo scopo di predire le traiettorie evolutive del cancro.

> l'autore fa notare che il fatto che si stiano sviluppando molti tool con output diversi e a volte contrastanti è preoccupante. Le discrepanze sono causate dalle varie differenze nei fattori sopraelencati ed in particolare a seconda della tipologia di dati analizzati. Il fatto che non ci sia selezione su un tipo di marker non esclude che invece non ci possa essere in un altro.

Variazioni sulla filogenesi dei tumori

Vi sono sostanzialmente tre metodologie di studio:

- **Cross sectional**: usa dati da **diversi tumori** per costruire i percorsi evolutivi comuni.
- **Regional bulk**: costruisce degli alberi per il **singolo paziente** valutando le variazioni fra le **diverse regioni** tumorali.
- **single cell**: per il **singolo paziente** e una **singola regione** per valutare le modifiche fra le singole cellule.

Cross-sectional

> Il primo esempio è dovuto a Fearon e Vogelstein, con un modello evolutivo lineare e realizzazione manuale, il primo esempio con metodi filogenetici deriva da Desper (albero oncogenetico, archi mutazioni con probabilità)

I primi modelli erano di tipo **character-based** (i più informativi e complessi algoritmicamente) ossia sfruttanti un insieme discreto di marker tumorali basandosi sulla **massima parsimonia** (i più semplici della loro classe) (Desper) che ha il difetto di assumere una certa rarità fra le mutazioni, non troppo adatta al contesto dei tumori. Seguono poi i metodi probabilistici basati sulla **massima verosimiglianza** o sul **campionamento bayesiano**, più adatti allo scopo ma più complessi. Approcci più recenti cercano di sfruttare al meglio le tecniche moderne di sequenziamento e molti di questi sono basati sul metodo

MCMC Markov Chain Monte Carlo, che permette di analizzare un numero maggiore di ricostruzioni filogenetiche possibili considerando diversi parametri ma ad un costo computazionale ancora più elevato.

Un'alternativa sono i metodi **distance-based** che permettono l'analisi su dataset molto ampi (vengono stimate delle distanze fra i samples e cercati i più vicini) sacrificando precisione (minima evoluzione).

> Questi metodoi sembrano far dipendere molto il loro risultato dal modello matematico selezionato, e potrebbero non essere affidabili in caso di elevata eterogeneità nel tumore

Regional Bulk

Passo in avanti rispetto ai cross-sectional in termini di specificità, prendono dati da regioni diverse e ne costruiscono un albero evolutivo. Hanno applicazioni da prima dell'avvento dell'NGS e ve ne sono di basati su diversi metodi che sfruttano sia soluzioni applicate per l'analisi filogenetica nelle specie che euristiche custom.

> **Deconvoluzione dei cloni dalla sequenza bulk**: ossia la capacità di estrapolare le sub-popolazioni clonali dai dati completi.

Single cell

La metodologia più specifica, vengono analizzate cellule individuali estratte da un singolo paziente. Le prime metodologie applicate per quest'analisi sono state effettuate con sistemi precedenti al sequenziamento come FISH o microsatellite markers, anora utilizzati per la loro capacità di analizzare un numero decisamente superiore di cellule. Molti dei sistemi software in questo ambito sfruttano metodologie create per lo studio filogenetico su specie senza sfruttare algoritmi o modelli espliciti.

Modello	descrizione
Massima parsimonia	Il più semplice, assume che avvenga il numero minimo di mutazioni
Minima evoluzione	Analogo del precedente per i metodi distance based, minimizza la quantità di evoluzione
Probabilistico	metodologia adatta anche a scenari complessi, dati inesatti e campionamento su parametri evolutivi ignoti. Si dividono in di massima verosimiglianza, che ricerca l'albero più adatto ai dati, e Bayesiani, utilizzati per identificare lo spazio degli alberi più plausibile con i parametri, più consistenti con dati e modello
Minimi quadrati pesati	Modello distance-based che crea l'albero più attendibile approssimando un insieme di distanze fra taxa (input) con l'approssimazione dei minimi quadrati

Algoritmo	desrizione
Combinatoriale	Classe di metodi adatta ad analisi character-based che prevede di trovare l'ottimo su un determinato insieme di possibili modelli; sono i più efficienti ma adatti a soli modelli semplici
Heuristic search	Sistemi che cercano di trovare un albero approssimato che si avvicini a quello ottimo, si adatta a modelli ignoti.
Neighbour Joining	metodo veloce che va a raffinare i sottoalberi approssimando un albero di minima evoluzione, può genereare output non consistenti dal punto di vista temporale
UPGMA	Unweighted Pair group with aritmetic mean: metodo per la costruzione dell'albero ricostruendo sottoalberi gerarchicamente e riunendoli assieme, è veloce ma richiede l'assuzione che le mutazioni avvengano con un ritmo costante
Markov chain Monte Carlo (MCMC)	Algoritmi adatti a diversi modelli probabilistici che permettono l'esplorazione dei range dei parametri e dei loro intervalli di confidenza. Generalmente è troppo costoso computazionalmente per permetterne l'utilizzo

Un esempio di applicazione pratica

Per capire meglio come può essere effettuata un'analisi corretta per un proprio studio con i metodi di inferenza filogenetica l'autore propone una simulazione realistica dove evidenzia le molteplici possibilità di errore.

- **DOMANDA**: quali sono la sequenza e le tempistiche con cui le CNVs si visualizzano nell'evoluzione del cancro al seno?
- **DATI**: sequenziamento del DNA su tutto il genoma con copertura di 50x da 200 cellule tumorali con genoma normale di riferimento.

Consistenza fra modello e dati

Sfruttando un programma già pronto basato sul neighbour joining è possibile ottenere un albero filogenetico già direttamente, molto simile a quelli degli studi a cui è stato applicato.

PROBLEMA: Questo però può essere dovuto al fatto che l'approccio impiegato non è adatto all'analisi delle mutazioni di tipo CNV, mentre è nato per quelle SNV, dato che assume che le mutazioni si accumulino **una per volta** con una **frequenza costante** e un cambiamento di molte basi sarebbe mal interpretato.

SOLUZIONE: Per riuscire a trattare le CNV si deve utilizzare un modello compatibile, come ad esempio quello probabilistico bayesiano che si adatta anche a scenari evolutivi complessi.

Compatibilità fra algoritmo e modello

PROBLEMA: cambiando il modello è necessario cambiare anche l'algoritmo che opera sui dati.

SOLUZIONE: In questo contesto si può scegliere di usare metodi che sfruttino l'MCMC Sampling, che sono lo standard per riuscire ad approssimare modelli probabilistici complessi su cui non si ha una teoria precisa.

Compatibilità fra modello e dati

PROBLEMA: Dato che ogni algoritmo porta con sè delle limitazioni specifiche bisogna accertarsi che queste siano rispettate dai dati. Nel caso di esempio si ha che l'algoritmo MCMC ha la caratteristica di avere una complessità esponenziale secondo il numero di cellule, dove 200 non è una quantità accettabile.

SOLUZIONE: Queste metodologie sono normalmente adatte solo a decine di specie, dunque una scelta possibile è quella di adattare i dati a questo algoritmo, e di passare da un'analisi single cell ad una bulk, ove sono sequenzate 10 diverse regioni provenienti da 20 guppi di cellule tumorali differenti, adattando così dati e algoritmo.

Compatibilità fra metodo e domande

PROBLEMA: Con la soluzione imposta al passo precedente non è possibile avere la risoluzione necessaria per poter rispondere alla domanda originale, dato che nelle CNV accade spesso che i tumori mostrino difetti di replicazione che portano all'accumulazione di queste mutazioni in sub-cloni che sono rari negli stadi iniziali della malattia. (troppe poche regioni, ove si possono verificare multiple CNV)

APPROCCI POSSIBILI: a questo punto bisogna considerare di rianalizzare a fondo il problema. Una prima idea può essere quella di prendere i dati dello studio ed analizzarli con un modello di massima parsimonia, più semplice e veloce rispetto a quello probabilistico, una seconda quella di usare algoritmi più esotici ed infine una terza quella di sfruttare diverse tipologie di marker (FISH ad esempio).

In alcune circostanze potrebbe non essere possibile trovare un accordo fra le componenti dello studio considerato, a quel punto è necessario riflettere sul fatto che non esistono metodi già pronti per molte importanti domande e che potrebbe essere necessario svilupparne uno (il che richede competenze notevoli in biologia computazionale).