

Analisi bioinformatica di Esoma Intero in trio

L'analisi dell'esoma intero applicata ai trio (genitori e figlio/figlia affetto) è un approccio molto potente al fine di identificare mutazioni 'de novo' potenzialmente patogeniche nei probandi in esame. Sequenziando il paziente e i suoi genitori, le varianti possono essere filtrate in base alla coerenza o meno calcolata sulla base delle leggi dell'eredità Mendeliana e priorizzate in base a parametri tecnici come la qualità di chiamata delle varianti, la frequenza nella popolazione e la rilevanza del gene per la patologia studiata.

Il numero globale iniziale di varianti identificabili in un esperimento di esoma intero in trio (padre, madre, figlio/figlia affetto) è di circa 150.000 (nella maggior parte dei casi varianti di singolo nucleotide). Le varianti di potenziale rilevanza patologica alla fine di questa procedura sono tuttavia dell'ordine delle decine, semplificando quindi molto il lavoro di validazione. L'analisi di esomi 'in trio' parte dall'analisi dell'esoma e procede con una procedura analitica sviluppata 'ad hoc' per l'identificazione delle putative varianti patogeniche.

Analisi dell'esoma

L'analisi bioinformatica dei risultati degli esperimenti di Esoma applicata ai trio si compone di due fasi consequenziali: mappaggio sul genoma di riferimento ed analisi delle variazioni con le relative statistiche.

Mappaggio sul genoma di riferimento

L'analisi inizia con il mappaggio delle sequenze prodotte dal sequenziatore sul genoma di riferimento nella versione più recente. Il mappaggio viene eseguito con software standard e parametri ottimizzati per la tecnologia Ion Torrent, in grado di gestire correttamente le peculiarità di questa piattaforma di sequenziamento (ad esempio per quanto riguarda le piccole inserzioni/delezioni).

Le statistiche riassuntive sui risultati globali del mappaggio e dell'arricchimento dell'esoma sono messe a disposizione del cliente nel report finale. Vengono forniti gli allineamenti nei formati standard binario ".bam" (con i relativi indici .bai) al fine di consentire la visualizzazione diretta degli allineamenti delle read sul genoma di riferimento con il programma Integrated Genome Viewer (files .bam ed indici .bai: <http://www.broadinstitute.org/igv/>).

Analisi delle variazioni e statistiche relative

La seconda parte dell'analisi, eseguita sempre tramite procedure standard e parametri ottimizzati per la tecnologia Ion Torrent, consiste nell'identificazione analitica ('call') dei polimorfismi di sequenza e delle piccole inserzioni e delezioni, sempre rispetto al genoma di riferimento. Le varianti sono successivamente annotate mediante il confronto con database noti come dbSNP, Cosmic e Clinvar.

Le variazioni identificate in questo passaggio, sia SNP che INDEL, vengono riportate per ogni campione in un file nel formato standard VCF, che vengono forniti al cliente e possono essere analizzati autonomamente con diversi strumenti analitici, anche basati su web. Vengono inoltre restituiti al cliente files in formato Excel contenenti l'annotazione delle varianti per tutti i campioni considerati. Queste annotazioni comprendono la profondità del sequenziamento (coverage totale di sequenza) nell'intorno della posizione in esame; la classificazione in omo- o eterozigote; il gene coinvolto e la sua localizzazione nel genoma; le conte delle read comprendenti il nucleotide corrispondente alla variante e quelle corrispondenti alla referenza genomica; i relativi valori di qualità; la frequenza relativa dell'allele variante; l'effetto predetto della variante sulla sequenza proteica.

L'analisi di inserzioni e delezioni (INDEL) entro i 20 nucleotidi rappresenta una risorsa addizionale molto interessante e poco esplorata di identificazione di variazioni potenzialmente causative. Le tabelle dei risultati di queste variazioni riportano anche le call degli alleli, la sequenza nel contesto della inserzione o delezione e la valutazione di omozigotà/eterozigotà/emizigotà delle varianti identificate.

Analisi del trio ed identificazione delle putative varianti patogeniche

I punti principali che possono guidare nella prioritizzazione delle varianti, e quindi nella loro riduzione numerica, nell'analisi di esomi in trio sono:

- ⇒ La struttura della famiglia (il pedigree viene sempre considerato nell'analisi).
- ⇒ Il criterio di selezionare principalmente varianti (nell'affetto) che abbiano un effetto funzionale sulla proteina codificata dal gene bersaglio della mutazione. Evidenze come mutazioni nel promotore e nei segnali di splicing non sono priorizzate in prima istanza.



⇒ L'utilizzo intensivo di criteri di controllo della qualità del dato come la qualità di chiamata e la copertura di sequenza per base.

L'esoma in trio ha il grande vantaggio di poter ricostruire facilmente la fase delle varianti utilizzando le informazioni del pedigree. I tre tipi di mutazioni che sono interessanti come potenzialmente patogeniche nel probando (quindi nel figlio o figlia affetti dei soggetti in esame) sono:

- ⇒ Varianti De Novo;
- ⇒ Varianti Eterozigoti Composte;
- ⇒ Varianti Omozigoti Recessive.

Le varianti De Novo (neomutazioni) sono varianti non condivise con nessuno dei due genitori. Di seguito le caratteristiche tipiche di questo tipo di mutazioni:

- ⇒ Corrispondono ad un c.d. "Errore Mendeliano", cioè non rispettano la dinamica di base delle variazioni monogeniche, non essendo ereditate
- ⇒ Possono essere originate da errori di replicazione del DNA, lesioni genetiche spontanee, elementi genetici trasponibili etc
- ⇒ Intervengono in fasi precoci dello sviluppo
- ⇒ Possono essere uniche per ogni paziente
- ⇒ Sono varianti rare, con MAF < 1% rispetto ai dati relativi agli esomi del Progetto di Sequenziamento '1000 Genomes' (<http://www.1000genomes.org/>);
- ⇒ Alterano la sequenza aminoacidica della proteina;
- ⇒ Il probando (figlio/figlia) è generalmente **eterozigote**;
- ⇒ Le misure di qualità che si possono utilizzare per la selezione/prioritizzazione sono la profondità allelica per l'allele di riferimento ed alternativo (numero alleli WT/numero alleli ALT); la profondità di sequenziamento per quel particolare sito polimorfico; alti valori di qualità di chiamata.

La seconda classe di polimorfismi considerati in questa analisi è quella di eterozigote composto del probando. I filtri da utilizzare in questo caso sono:

- ⇒ I genitori sono eterozigoti per polimorfismi differenti **nello stesso gene**;
- ⇒ Il probando (figlio/figlia) è eterozigote per (almeno) due SNP **nello stesso gene**, una da ognuno dei due genitori;
- ⇒ Il polimorfismo altera la sequenza aminoacidica del prodotto genico;
- ⇒ Sono varianti rare, con MAF < 1% rispetto ai dati del Progetto di Sequenziamento Esomi del progetto "1000 Genomes" contenuti in dbSNP;
- ⇒ Si utilizzano criteri di filtraggio sulla profondità di chiamata del genotipo, sul punteggio di qualità, sull'indice di eterozigosità dell'allele alternativo.

La terza classe di polimorfismi considerati in questa analisi sono varianti recessive rare in omozigosi. I criteri per i filtri di questa condizione sono:

- ⇒ I genitori devono essere eterozigoti per lo stesso polimorfismo;
- ⇒ Il probando deve essere omozigote;
- ⇒ Sono varianti rare, con MAF < 1% rispetto ai dati del Progetto di Sequenziamento Esomi dell'1000 Genome Project;
- ⇒ Il polimorfismo altera la sequenza aminoacidica del prodotto genico;
- ⇒ Si utilizzano criteri di filtraggio sulla profondità di chiamata del genotipo, sul punteggio di qualità, sull'indice di eterozigosità dell'allele alternativo.

Nella **Figura 1** vengono riassunte le caratteristiche rilevanti di queste tre categorie di mutazioni



Figura 1 – categorie di mutazioni identificabili con un’analisi in trio

Le varianti prodotte dalla prima parte della procedura di analisi Full Exome vengono quindi sottoposte ad una serie di comparazioni e filtri specifici, tramite procedure standard per la tecnologia Ion Torrent, per l’identificazione delle mutazioni de novo; delle varianti in eterozigosi composta; delle varianti recessive in omozigosi secondo lo schema concettuale illustrato. I risultati vengono riassunti in tavole Excel, e vengono fornite anche le liste globali di tutte le variazioni identificate.

I risultati della procedura completa vengono riportati evidenziando le varianti chiamate nelle tre categorie: De Novo; Eterozigote composto; Omozigote recessivo. Le varianti saranno quindi annotate con le seguenti informazioni:

- Tipo di variazione (SNV/Indel);
- Categoria (De Novo; Eterozigote Composto; Omozigote Recessivo);
- L’allele variante e le read assegnate all’allele variante ed all’allele di riferimento;
- L’ID ed il nome del gene;
- Annotazioni relative ai principali database, quali dbSNP, COSMIC, Clinvar
- Le conseguenze sul prodotto genico della variante: posizione nel cDNA, nella CDS, nella proteina; gli aminoacidi di riferimento e varianti con i relativi codoni;
- Le predizioni dell’effetto della variante sulla funzionalità della proteina prodotte da SIFT e Polyphen;
- Lo status del probando, della madre e del padre per la variazione in esame (Omozigote; Eterozigote; Corrispondente al Genoma di Riferimento);
- Prioritizzazione delle varianti in base alla rilevanza per la patologia, anche tramite liste di geni candidati fornite dal cliente.

Informazioni per gli ordini

Prodotto	N. Catalogo
Analisi Bioinformatica III: DNA (esoma intero in trio)	DNA-BF03