

## **Progetto di Ricerca**

### **Valutazione e sviluppo di metodi per l'identificazione di varianti genetiche "Copy Number Variants" (CNV) mediante tecnologie di sequenziamento di nuova generazione**

#### **Introduzione**

Il DNA dell'uomo è caratterizzato dalla presenza di alterazioni genomiche che possono variare in dimensione dalla singola base nucleotidica fino a grandi eventi che coinvolgono interi cromosomi.

Fino a pochi anni fa si pensava che la maggiore parte della variabilità genetica coinvolgesse singole basi del genoma (Single Nucleotide Polymorphism SNP), mentre alcuni studi pubblicati tra il 2006 e il 2007 hanno dimostrato che la più grande fonte di variabilità genomica risiede nelle variazioni del numero di copie (Copy Number Variants, CNV). Le CNV sono definite come regioni genomiche che coinvolgono segmenti di DNA maggiori di 50 basi e che si possono trovare in un numero variabile di copie. Le CNV giocano un ruolo fondamentale nella genetica delle patologie complesse e negli ultimi anni sono stati pubblicati molti articoli scientifici che dimostrano il loro coinvolgimento in patologie neuropsichiatriche e anche nella suscettibilità all'HIV, al morbo di Alzheimer e al morbo di Parkinson.

L'avvento delle tecnologie di sequenziamento di seconda generazione (Sequenziamento ad alta produttività, High throughput sequencing, HTS) ha aperto nuove opportunità per l'identificazione delle CNV, consentendo di raggiungere risoluzioni fino a pochi anni fa impensabili.

Uno degli approcci più promettenti per l'identificazione di CNV consiste nel contare il numero di sequenze (Read Count, RC) che vengono allineate a particolari regioni del genoma. L'approccio RC è stato utilizzato per identificare CNV in campioni di tumore e per studiare la variabilità genomica di popolazioni.

#### **Scopo**

Sebbene alcuni metodi statistici siano già stati presentati in letteratura, rimane ancora molto da fare per capire approfonditamente la capacità dei metodi RC di identificare le CNV e predire il numero esatto di copie di DNA. Lo scopo di questo studio si articola in tre punti fondamentali:

1. studiare le proprietà statistiche e le fonti di variazione sistematica dei dati RC.
2. Implementare e valutare metodi statistico-matematici per l'identificazione automatizzata di CNV da dati RC.
3. Sviluppare un software che includa i metodi sopracitati che consenta al ricercatore biomedico e al clinico di analizzare facilmente i dati di RC e generare i risultati attraverso un'interfaccia "user-friendly".

## **Risultati attesi**

Da queste analisi ci attendiamo di definire i limiti di risoluzione dei metodi RC nell'identificare regioni genomiche coinvolte in CNV e di comprendere la loro capacità di predire il numero esatto di copie di quelle regioni. Inoltre, tale progetto consentirà lo sviluppo di un software fruibile on-line da rendere disponibile per le attività di ricerca e diagnostica del nostro Dipartimento e dei ricercatori esterni ad esso a livello nazionale ed internazionale.

## **Impatto clinico**

La scoperta della presenza di CNV nel genoma umano ha avuto un grande impatto per molte aree della ricerca biomedica. Alterazioni genomiche possono condurre a un'ampia gamma di manifestazioni cliniche che includono ritardi mentali e malformazioni congenite. Inoltre, alterazioni nel numero di copie di DNA sono state associate all'autismo, al morbo di Parkinson e a vari tipi di tumore.

I risultati ottenuti con questo progetto consentiranno al ricercatore biomedico di pianificare esperimenti di sequenziamento del genoma per identificare CNV con la risoluzione desiderata. Inoltre, lo sviluppo di un software "user-friendly" permetterà di sfruttare i dati prodotti dai sequenziatori di nuova generazione anche nel laboratorio di diagnostica.

Alkan C., et al. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 2011;12:363-376.

Campbell P.J., et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* 2008;40:722-729.