Al: **Consiglio Nazionale delle Ricerche**
Ufficio Accordi e Relazioni Internazionali
P.le Aldo Moro, 7
00185 ROMA

Napoli, 03/01/2011

Programma Short term mobility 2010 - Relazione dell'attivita' scientifica svolta

Titolo del programma:	Modeling human demographic history using 1000 Genomes Project pilot data
Proponente e Fruitore:	Vincenza Colonna, ricercatore presso l'Istituto di Genetica e Biofisica "A. Buzzati-Traverso" di Napoli
Istituzione ospitante:	The Wellcome Trust Sanger Institute - Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus - Hinxton, Cambs. - United Kingdom
Periodo:	29/09/2010-26/10/2010

La conoscenza della demografia delle popolazioni umane costituisce una premessa indispensabile agli studi di genetica medica e di biologia evoluzionistica: permette di spiegare i pattern di varabilita' genetica osservati e chiarire le modalita' secondo le quail si sono realizzati alcuni processi demografici (es. migrazioni umane); permette l'identificazione dei patterns di linkage disequilibrium (ovvero l'associazione non random di due loci adiacenti su di un cromosoma), una caratteristica rilevante per l'identificazione dei geni responsabili di alcuni tratti genetici (es. malattie); e' indispensabile per la discriminazione delle regioni del genoma che hanno subito selezione naturale da quelle neutrali. Lo studio della demografia puo' essere ricondotto a due fasi: la formulazione di un range di possibili modelli e l'identificazione di quello(i) che meglio spiega(no) la variabilita' genetica osservata, attraverso il confronto con dati empirici.

Il Thousand Genomes Project (1000G) e' ad oggi il piu' grande studio della diversita' genomica umana1. La fase pilota del progetto consiste nel sequenziamento a basso coverage (2-4x) di 179 campioni da quattro popolazioni incluse nel progetto HapMap 2, il sequenziamento ad alto coverage degli esoni di circa 800 geni in 697 campioni da sette popolazioni HapMap, e dell'intero genoma di due trios (madre, padre, figlio). Lo scopo finale del 1000G e' quello di realizzare la sequenza complete del genoma di 2500 individui provenienti da 27 diverse popolazioni. I dati (pubblicamente disponibili) della fase pilota hanno permesso l'identificazione di 15 milioni di polimorfismi di singolo nucleotide (SNPs) di cui 8.5 milioni mai identificati prima, 1 milione polimorfismi di inserzione e delezione e 20,000 varianti strutturali.

Il Sanger Institute partecipa al consorzio che gestisce questo progetto assieme al National Human Genome Research Institute in Bethesda, ed al Beijing Genomics Institute in Shenzhen, China.

Il progetto parzialmente finanziato dalla borsa di studio in oggetto prevede la messa punto di sviluppo e analisi di modelli per la descrizione della demografia nelle quattro popolazioni sequenziate a basso coverage. Successivamente la metodologia sviluppata nelle popolazioni del progetto pilota sara' applicata a tutte le popolazioni del 1000G. Di seguito un elenco delle attivita' svolte durante il periodo di lavoro presso il Wellcome Trust Genome Campus, finanziato dalla borsa di studio Short-term mobility del CNR.

a) **Defizione dello schema del progetto. Ho definito in maniera dettagliata lo schema del progetto da seguire, del quale riporto a grandi linee le fasi principali:**
   1. *Elaborazione dei modelli e scelta dei parametri. Un modello e' una costruzione logica che definisce un sistema di relazioni più o meno semplificate che rappresentano uno o più fenomeni demografici nel loro svolgersi. Un modello e' descritto dai parametri che sono i valori attribuiti alle relazioni e agli eventi quantificabili (es. dimensioni delle popolazioni, tempo di separazione , etc..). I modelli elaborati per questoprogetto si basano su diversi modelli (molto simili tra loro) e parametri utilizzati in letteratura per descrivere le popolazioni umane 3-10.*
   2. *Produzione di dati simulati secondo diverse combinazioni di modelli e parametri. Con la tecnica della simuazione vengono prodotti campioni di sequenze fittizie secondo determinati modelli e parametri. La simulazione delle sequenze viene effettuata utilizzando la metodologia del Sequentially Markov Coalescent (SMC)11, 12* . Il SMC e' un algoritmo implementato di recente ed e' basato sulla teoria del coalescente13, 14. Rispetto al coalescente standard permette la simulazione di data set paragonabili per dimensioni ai data set prodotti dal 1000G, grazie alla capacita' di effettuare delle approssimazioni nei pattern di ricombinazione.
   3. *Comparazione di dati empirici e simulati. I dati simulate vengono utilizzati per calcolare delle statistiche che descrivono la variabilita' genetica dei campioni simulate second uno specifico modello. Queste statistiche vengono confrontate con quelle derivate dai dati empirici per* la scelta scelta del modello che meglio descrive i dati empirici.

b) **Produzione di nuovo software e ottimizzazione di software esistente per la simulazione e l'analisi dei dati. Il processo descritto nel paragrafo precedente viene ripetuto un numero considerevole di volte in quanto vengono testate diverse combinazioni di parametri che descrivono i diversi eventi presenti nel modello. E' quindi indispensabile automatizzare il**

processo creando un a pipeline, e valutare e ottimizzarne i tempi di calcolo. A tale scopo, per la parte inerente la fase 2 descritta al punto a) ho modificato il sotware (MaCS11) utilizzato per le simulazioni in modo da ottimizzare i tempi di simulazione, rimuovendo le parti non strettamente necessarie al progetto; ho incorporato alle simulazioni un modello di ascertainment bias per tenere conto dei falsi negativi nell'identificazione dei polimorfismi di sequenza derivati dal basso coverage. Per la parte inerente la fase 3 del precedente punto, ho scritto del software per derivare le statistiche inerenti la variabilita' genetica (siti segreganti, numero medio di differenze a coppie, site, frequency spectra, etc..) dai dati simulati per confrontarle con quelle derivate dai dati empirici. Infine ho scritto e messo a punto una pipeline che permette di automatizzare e collegare i diversi passaggi: scelta dei parametric, simulazione, calcolo delle summary statiscics, confronto con dati reali.

I tempi previsti per lo sviluppo, il completamento e la pubblicazione dei risultati di questo progetto sono di circa un anno, nel quale conptinuera' la collaborazione con l'istutuzione ospitante.