Allegato tecnico all'Accordo di Ricerca Collaborativa tra Cineca e Network for Italian Genomes (nel seguito NIG) per la creazione di un repository centralizzato per i dati di varianti genomiche umane prodotti da sequenziamento massivo parallelo

Obiettivi del progetto

Obiettivo principale del progetto è strutturare e utilizzare l'infrastruttura abilitante del Cineca per consentire l'archiviazione e l'analisi dei dati prodotti dal sequenziamento del genoma da parte dei partner NIG e la creazione di un repository che consenta di ricavare dati di varianti genomiche umane con particolare riferimento alla popolazione italiana.

Il networking, che esiste e si rafforzerà ancora di più tra i partner del progetto, rappresenta una risorsa importante per la diagnosi e la ricerca, sia per la ricerca di base sia per la ricerca applicata alla salute umana. Molti esempi testimoniano come l'accesso a un numero relativamente alto di dati genomici, relativi alla popolazione specifica, abbia svolto un ruolo fondamentale per l'identificazione di geni di malattie genetiche, geni di suscettibilità e lo sviluppo di nuovi trattamenti.

Lo scopo del repository centralizzato è proprio quello di coordinare l'archiviazione dei dati in una piattaforma unica per migliorare l'accesso e l'analisi dei dati genomici italiani da parte della comunità biomedica. L'idea complessiva è di creare un ambiente che assicuri la tutela della privacy e la riservatezza dei dati per i soggetti italiani coinvolti, interessati o no da malattia, e dei loro familiari.

Indicativamente, la realizzazione del repository permetterà ai ricercatori di NIG di raggiungere i seguenti obiettivi:

- Identificazione delle varianti comuni (> 1%) nella popolazione italiana diversamente presenti in banche dati europee ed extraeuropee, inclusa valutazione della loro diffusione regionale e locale. Eventuale annotazione come fattore di rischio.
- Identificazione di varianti rare (< 1%) e diffusione loco-regionale. Annotazione del possibile significato patogenetico.
- Imputazione dei possibili aplotipi nazionali associati a varianti comuni e rare. Co-occorrenza (sintenia) nel blocco aplotipico con altre varianti già annotate per significato funzionale.
- Costruzione di un Genoma di Riferimento Italiano (ItaGenomeRefSeq) e relativo costante aggiornamento.
- Diffusione dei risultati mediante pubblicazione sul sito NIGweb con accesso condizionato a consorziati, utilizzatori ecc., secondo le policy condivise.

Vincoli ed assunti

Sarà responsabilità di Cineca adottare adeguate misure di sicurezza per garantire la riservatezza e la protezione dei dati, implementando protocolli per l'accesso ai dati solo ai membri della comunità scientifica indicati da NIG, secondo i livelli di volta in volta scelti da NIG per l'accesso al singolo esperimento (e.g. dato pubblico; dato riservato con accesso da lista di ricercatori identificati; dato pubblico da data definita)

Sarà responsabilità di NIG che i suddetti livelli di accesso siano rispettosi della legge nazionale sulla privacy, raccogliendo, ove richiesto, il consenso informato scritto per l'archiviazione dei dati.

Di seguito le policy di sicurezza che saranno implementate:

Autenticazione e autorizzazione

- La consultazione dei dati genetici trattati con strumenti elettronici è consentita previa adozione di sistemi di autenticazione basati sull'uso combinato di informazioni note agli incaricati e di dispositivi, anche biometrici, in loro possesso.
- Sarà prevista una procedura di autenticazione per l'identificazione univoca degli utenti tramite opportune credenziali di autenticazione
- L'accesso ai dati sarà consentito ai soli utenti in possesso di credenziali valide
- Le credenziali non utilizzate da almeno 180 giorni saranno disattivate
- I codici identificativi già utilizzati non saranno riassegnati ad altri utenti
- Le password dovranno avere una lunghezza minima di 8 caratteri
- L'utente sarà tenuto a modificare la password iniziale al primo accesso
- Le password avranno una durata di 90 giorni
- Gli utenti saranno chiamati a modificare la propria password dopo il primo accesso e successivamente ogni 90 giorni
- Le password non saranno conservate localmente in chiaro (ma sarà salvato un loro hash)
- Le password saranno trasmesse in rete su canali sicuri e cifrati (https)
- Saranno previsti meccanismi di controllo della robustezza delle password
- Saranno previsti meccanismi di separazione dei privilegi in funzione del profilo utente adottando il principio del minimo privilegio
- L'accesso alle risorse di sistema sarà limitato ai soli account privilegiati
- Ciascuna Università (o ente aderente al NIG) potrà accedere solo ai risultati statistici elaborati dall'applicativo e/o a informazioni aggregate; pertanto non saranno conoscibili i dati riferiti a singole unità statistiche/interessati (a meno del caso in cui i dati dell'interessato siano trattati dal titolare degli stessi che potrà, naturalmente, trattarli singolarmente).

Validazione dei dati

- Saranno definiti i punti di ingresso ed di uscita dell'applicazione che richiederanno specifici controlli di validazione dei dati
- Saranno previsti meccanismi di validazione dell'output tramite l'applicazione di opportune codifiche
- Saranno considerati e contrastati attacchi di tipo SQL Injection e Cross Site Scripting

Gestione delle sessioni utente

- Le sessioni utente saranno soggette ad un periodo di time-out oltre il quale sarà richiesta l'autenticazione
- Il contenuto degli identificatori di sessione saranno cifrati
- Gli identificatori di sessione saranno trasmessi su canali sicuri e cifrati (https)
- Saranno previsti meccanismi di logout per forzare la chiusura della sessione

Registrazione delle operazioni

- Saranno individuati gli eventi chiave per la sicurezza che dovranno essere soggetti a registrazione tramite file di log
- I log saranno conservati in un'area non accessibile agli utenti per garantirne l'inalterabilità (e.g. syslogd)
- I file di log saranno soggetti a procedure di backup periodico
- I file di log saranno soggetti a procedure di rotazione, cancellando i log più vecchi

- La frequenza di rotazione ed il periodo di mantenimento dei log saranno variabili configurabili dagli amministratori del sistema
- L'accesso ai file di log sarà consentito unicamente agli account privilegiati
- Saranno valutate procedure di verifica del funzionamento dei log

Crittografia e disponibilità dei dati

- Il sistema realizzato da Cineca dovrà essere realizzato in modo da trattare i dati genetici con tecniche di cifratura o mediante l'utilizzazione di codici identificativi o di altre soluzioni che li rendano temporaneamente inintelligibili anche a chi è autorizzato ad accedervi e permettano di identificare gli interessati solo in caso di necessità.Le password non saranno conservate localmente in chiaro (ma sarà salvato un loro hash)
- Le credenziali di accesso saranno trasmesse in rete su canali sicuri e cifrati (https)
- Sono previsti protocolli di comunicazione sicuri che garantiscano, previa verifica, l'identità digitale del server che eroga il servizio e della postazione client da cui si effettua l'accesso ai dati, ricorrendo a certificati digitali emessi in conformità alla legge da un'autorità di certificazione.
- Saranno predisposte procedure di backup dei metadati e dei profili utente con frequenza giornaliera ed un periodo di retention di 30 giorni (configurabile)
- Saranno predisposte procedure di archiviazione tramite sincronizzazione per i file raw più vecchi di 7 giorni (configurabile)
- Saranno previsti dei meccanismi di ripristino dei dati in caso di danneggiamento
- Gli ambienti di produzione e sviluppo opereranno su insiemi di dati disgiunti

Requisiti organizzativi e gestionali

- Tutti i meccanismi di sicurezza adottati saranno documentati
- La conformità alle misure di sicurezza saranno attestate dalle ISO 9001:2008 e ISO 27001
- Saranno previsti adeguati meccanismi di controllo degli accessi
- La connessione a strumenti di gestione e configurazione avverrà su canali sicuri e cifrati (https)

Trattamento dei dati personali e sensibili

- Saranno previsti meccanismi di protezione dei dati contro minacce di intrusione e dell'azione di programmi malevoli
- Saranno previsti aggiornamenti periodici di tutti i componenti software per prevenire vulnerabilità e correggerne i difetti
- Saranno previsti meccanismi di backup con frequenza giornaliera per profili utente e metadati e meccanismi di archiviazione per file raw con frequenza settimanale
- Saranno previsti meccanismi di ripristino dei dati in caso di danneggiamento

Analisi dei dati genomici

Cineca si impegna ad adottare una procedura accelerata per la valutazione dei progetti di ricerca bioinformatica presentati da ricercatori NIG che avessero come obiettivo l'analisi dei dati genomici inclusi ma non limitati a Whole Exome Sequencing, Target Re-sequencing, Whole Genome Sequencing, funzionali alla generazione di metadati utili al popolamento del repository.

A seguito di valutazione tecnica positiva dei suddetti progetti, Cineca si impegna altresì ad erogare adeguate risorse di calcolo per la realizzazione delle analisi, sia mediante accesso diretto al cluster Pico che tramite pipeline web per l'analisi di dati NGS, già implementate da Cineca (http://www.hpc.cineca.it/content/hpc-next-generation-sequencing)

Tempistica

(MX = X mesi a partire dalla firma del presente documento)

Trasferimento dati grezzi (FASTQ) su risorse CINECA (M3)

- preparazione ambiente
- modalità e meccanismo di trasferimento e tempistiche di upload
- deposito dati su file-system
- sviluppo di pipeline concordata per la sottomissione di dati di WES
- sviluppo di pipeline concordata per la sottomissione di dati di WGS
- sviluppo di pipeline concordata per la sottomissione di dati di pannelli di geni
- ulteriore armonizzazione dati provenienti da piattaforme diverse

Questa fase richiederà un lavoro di armonizzazione dei dati che sarà effettuato da CINECA con il supporto di almeno 2 rappresentanti del NIG.

Progettazione repository per gestione dati post-analisi e varianti valide (M7)

- struttura database
- meta-dati
- accesso utenti
- condivisione dati
- privacy/sicurezza
- interfacce

Implementazione prima release repository (database metadati) (M9)

Test e verifica delle funzionalità prima release (M10)

Implementazione seconda release repository (M12)