Azione 3

Verifica di congruenza dei dati e delle informazioni

Pipeline QC di dati pedigree, di dati quali-quantitativi e di dati genomici

Le statistiche e i controlli effettuati sui dati genomici, qualitativi e del pedigree verranno ottimizzati durante l'intero periodo del progetto I-BEEF2, al fine di avere una procedura ad-hoc per ciascuna tipologia di dato.

Nello specifico, il dato genomico viene scaricato dal server di proprietà di LGS e posizionato in una directory specifica del server ANACLI. La cartella zippata preparata dal laboratorio di analisi contiene diversi file quali l'elenco delle matricole a cui corrisponde un ID univoco dato dal laboratorio, la mappa genomica del chip usato, un report qualitativo del DNA e una cartella contenente un file *ped* e uno *map*.

Il primo passaggio è quello di scompattare i file nella directory specifica e verificare che non ci siano delle cartelle duplicate.

Tramite il numero di matricola si risale alla razza dell'individuo facendo controlli incrociati tra le anagrafiche di ciascuna razza e i vari elenchi che ANACLI invia al laboratorio, così da diminuire ogni possibilità di errore dovuto ad esempio alla non corretta trascrizione del numero di matricola. In caso in cui la matricola non venga rintracciata in nessuno degli archivi sopra descritti, viene collocata automaticamente in un file di testo che verrà successivamente visionato per effettuare i dovuti controlli.

Vengono individuate e trascritte anche altre informazioni come il sesso e il numero di generazione di appartenenza.

Terminati questi controlli iniziali, e valutata la provenienza dei soggetti, si creano i file nel formato *bed* tramite l'utilizzo del software PLINK. Questi file contengono tutti gli individui e i marker genomici di quella determinata cartella inviata da LGS così da poter risalire, in qualsiasi momento, alla data e al nome di quell'archivio.

Successivamente vengono individuati ed elencati i soggetti che hanno una qualità dei dati non soddisfacente, cioè inferiore ad una certa soglia di completezza, settata al 90-95%. Questo passaggio rappresenta il primo controllo qualità che verrà poi ripetuto ed aggiustato per ciascuna analisi, rendendolo più o meno stringente.

Per quanto riguarda i dati pedigree, le anagrafiche di ciascuna razza vengono scaricate periodicamente e un controllo qualità viene effettuato con una pipeline costruita *ad hoc* per ciascuna razza. Gli errori più comuni sono:

- la matricola del soggetto che risulta uguale al padre o alla madre;
- animali di sesso maschile che appaiono nel campo di madre, e animali di sesso femminile che appaiono nel campo di padre;
- Data di nascita errata, posteriore alla data di nascita della progenie.

Le correzioni degli errori sopra descritti sono effettuate tramite delle pipeline che permettono di costruire il pedigree, segnalando gli eventuali errori. Il software usato a questo scopo è il software R, più precisamente il pacchetto “optisel”, che permette non solo di costruire un pedigree corretto da utilizzare per le analisi successive, ma anche stimare il coefficiente di consanguineità, la completezza delle anagrafiche e il numero di generazioni.