UNIVERSITÀ DEGLI STUDI GENOVA

Dipartimento di scienze della terra, dell'ambiente e della vita

Corso di Laurea Magistrale in:

BIOLOGIA APPLICATA E SPERIMENTALE
Curriculum BIOSANITARIO

STEREOTIPI PUBBLICI DI IMMUNOGLOBULINE: UN'ESTENSIONE DEL CONCETTO DI CLONOTIPI PUBBLICI.

Relatore:

Prof. Marco Giovine

Correlatore:

Prof. Franco Fais

Relatore esterno

Prof. Davide Bagnara

Candidato:

Luca Calabrese

Anno accademico 2021/2022

INDICE

I. Introduzione

I.a – Linfociti B ed immunoglobuline [1,2]

I linfociti B sono le cellule deputate alla produzione e all'espressione sulla membrana di anticorpi che fungono da recettore per l'antigene, il cui ingaggio è responsabile dell'avvio del processo di attivazione linfocitaria. I linfociti B fanno parte dell'immunità adattativa/acquisita in grado di rispondere in modo altamente specifico ad un'ampia varietà di antigeni grazie ad eventi di ricombinazione somatica di segmenti genici. Questo tipo di risposta immunitaria consta di componenti: uno effettore (plasmacellule) che sintetizza e secerne anticorpi ed una memoria (cellule memoria) pronta a riconoscere ri-esposizioni allo stesso antigene ed attivare una risposta più rapida, più intensa e più efficace. In aggiunta alla risposta immunitaria delle cellule B entra in gioco anche l'espansione clonale, che permette di generare un gran numero di linfociti specifici per lo stesso antigene a partire da quello naïve che, per primo, ha incontrato l'antigene.

Inizialmente i geni che codificano per i recettori antigenici sono presenti in uno stato non funzionale nella linea germinale ed è solo dopo diversi eventi di ricombinazione questi geni diventano funzionali. Infatti, durante la costruzione di un recettore vengono effettuate diverse prove che avranno successo solo grazie alla ricombinazione genica fa variare le catene pesanti e leggere delle immunoglobuline garantendo una risposta efficace contro migliaia di antigeni.

I geni delle immunoglobuline si trovano su tre cromosomi diversi il 22, il 2 ed il 14, questi tre loci contengono segmenti genici per la regione variabile e la costante, separati da regioni non codificanti. La catena pesante H vede un locus sul cromosoma 14 diviso nei segmenti V, D e J associati alle regioni codificanti per le porzioni costanti della catena pesante: Cμ per IgM, Cδ per IgD, Cγ per IgG, Cε per IgE e Cα per IgA.

La catena leggera ha due loci, il locus per λ sul cromosoma 22 con segmenti V, J e C, ed il locus per κ sul cromosoma 2 con segmenti V, J e C, entrambi senza D.

All'estremità 5' di ciascun segmento V si trova la regione Leader (L), una regione nucleotidica altamente conservata che codifica per 20-30 aminoacidi N-terminali che servono da segnale per il trasferimento della catena nel lume del reticolo.

La generazione dell'immunoglobulina prevede due tipologie di ricombinazione una combinatoria durante la quale ogni linfocita sceglie a caso un esone V tra i 100 V, un esone D tra i 20 D e un esone J tra i 6J producendo una variabilità teorica pari a 10^9-10^{11} di immunoglobuline diversa ed una giunzionale, successiva alla combinatoria, che aggiunge e/o rimuove nucleotidi casuali tra i segmenti V e D-J. Inoltre, nelle fasi di ricombinazione e maturazione entrano in gioco anche diversi enzimi RAG-1 e RAG-2 che tagliano all'altezza di sequenze segnale di ricombinazione RSS i tratti di V, D o J, inducendo la ricombinazione somatica tra due regioni e la successiva generazione di una giunzione, visto che dopo il taglio nella giunzionale entra in gioco la TdT (Transferasi desossiribonucleotidica Terminale) ed, in caso di errori, il Complesso NHEJ (Non Homologous End Joining) per riparare il DNA.

La maturazione dei linfociti B prevede diverse fasi:

***Cellula staminale**:* il DNA è ancora in forma germinale

Linfocita pro-B: in questa fase avvengono due eventi di ricombinazione che permettono l'espressione della catena pesante: giunzione D-J, seguita dal congiungimento di un segmento V al complesso DJ, con eliminazione del tratto di DNA nel mezzo.

Dividiamo questa fase in base agli eventi di ricombinazione il *Linfocita pro-B precoce*, nella quale avviene il congiungimento dei segmenti D-J e la delezione del tratto di DNA interposto ed il *Linfocita pro-B tardivo*, nella quale avviene la formazione del segmento V-D-J finale.

Le due fasi di ricombinazione vedono la messa in atto delle seguenti fasi:

*Formazione della Sinapsi***,** Rag-1 sceglie a caso due distinti segmenti genici codificanti che vengono posti in contatto grazie alla formazione di strutture ad anello e vengono mantenuti in questa posizione per i successivi eventi di taglio, processamento e ricongiunzione.

*Clivaggio e formazione hairpin***,** dopo la "scelta" dei segmenti, RAG-1 taglia il DNA mentre RAG-2 tiene vicine e ferme le estremità terminali del loop. Si otterrà una rottura a doppio filamento, ciascun "moncone" di DNA avrà un 3'-OH libero e si formerà così una forcina detta HIRPIN.

Apertura hairpin e processazione estremità appiccicose, la forcina viene aperta e le estremità appiccicose vengono tagliate dall'endonucleasi Artemis in modo casuale ed asimmetrico.

Riparo e ricombinazione giunzionale, il tratto più corto viene esteso con nucleotidi complementari inseriti da enzimi di riparo del DNA, questi aggiungono nucleotidi (P) in base alla complementarietà ed in seguito l'enzima TdT aggiunge massimo 20 nucleotidi (N) casuali tra i due segmenti, dando origine alla regione ipervariabile HCDR3, questa perdita o aggiunta di nucleotidi durante la formazione di giunzioni codificanti porterà ad un'ulteriore diversità nelle giunzioni non codificate da VDJ.

Linfocita pre-B: in questa fase il linfocita ha ricombinato i geni delle catene pesanti e le ha tradotte a proteine, nel citoplasma sono ora presenti catene pesanti μ libere.

Una piccola percentuale di catene μ va sulla membrana, in associazione a catene leggere surrogate, a formare i pre-BCR recettori che recepiscono segnali anti-apoptotici inviati dalle cellule stromali, necessari per la proliferazione e la prosecuzione della maturazione.

Quindi, vengono selezionati ed espansi i linfociti pre-B che esprimono una catena μ funzionale di membrana, è durante questa fase che avviene il primo controllo nella maturazione dei B: se il pre-BCR recepisce i segnali anti-apoptotici vi sarà , se il recettore non è funzionante la cellula va in apoptosi.

Linfocita B immaturo: la cellula perde la positività per TdT, non prolifera e possiede IgM sulla membrana. A questo stadio maturativo avviene è il secondo punto di controllo, che vede una selezione negativa basata sull'affinità verso il self: se BCR riconosce una proteina del midollo la cellula incorrerà in due destini: apoptosi o editing recettoriale grazie alla riattivazione dei geni Rag e ricombinazione delle catene leggere.

Le cellule immature sopravvissute a questa seconda selezione entrano in circolo e subiscono una maturazione iniziando ad esprimere di IgD (catene δ).

Linfocita B maturo naïve: presenta sulla membrana IgM, IgD e recettori per il complemento, ma alla fine, le IgD vengono perse e IgM rimarranno l'unico marcatore di superficie pronto a legare eventuali antigeni.

Linfocita B attivato*:* quando il linfocita lega l'antigene si attiverà ed andrà incontro all'espansione clonale, processo che darà origine a numerosi cloni linfocitari tutti muniti della stessa IgM di membrana.

Questi cloni prenderanno diversi destini: effettuare lo switch isotipico per avere immunoglobuline con diversa regione constante (IgA, IgE, IgG), maturare l'affinità per l'antigene nel centro germinativo grazie l'ipermutazione somatica a livello del gene V, differenziare plasmacellule per sintetizzare e secernere anticorpi o differenziare in cellule della memoria che rimangono in circolo con le IgM di membrana pronte a riconoscere future re-infezione del medesimo antigene.

I.b – Repertorio Immunoglobulinico [3–7]

Il repertorio immunoglobulinico è la raccolta di linfociti B con caratteristico recettore antigene-specifico di membrana. Poiché è alla base delle condizioni immunologiche dell'organismo l'identificazione dei cloni e la caratterizzazione della loro diversità sono fondamentali per capire come il sistema immunitario adattativo interviene contro malattie infettive e tumori. Le risposte immunitarie acquisite possono essere inizializzate e ottenute attraverso il riconoscimento altamente specifico dell'antigene con i recettori della superficie cellulare, essendo gli antigeni estremamente diversi i recettori in grado di riconoscere questi antigeni con elevata affinità di legame devono essere selezionati da un enorme pool di recettori ottenuti grazie a riarrangiamenti somatici casuali dei geni V, D e J durante la differenziazione dei linfociti. Di conseguenza, questo meccanismo di selezione immunologica può portare a una raccolta altamente personalizzata e singolare di linfociti in ogni individuo, questa viene chiamata repertorio immunitario primario. I repertori primari vengono poi modellati e drasticamente modificati durante le risposte guidate dall'antigene, specialmente in contesti infiammatori dovuti ad infezioni da agenti patogeni, sindromi autoimmuni e tumori, così da ottenere repertori reali funzionali ed in grado di mediare la risposta immunitaria.

Considerando l'importanza di risposte immunitarie adattative efficienti nel liberarsi delle infezioni in modo naturale o per evitare danni autoreattivi, ma anche per scopi terapeutici come la vaccinazione o la terapia cellulare, ci si rende conto dell'importanza di comprendere come i repertori linfocitari vengono selezionati durante la differenziazione.

La sfida principale nell'analisi del repertorio BCR deriva proprio dalle difficoltà che insorgono nell'interrogarne l'enorme diversità ottenuta nel processo maturativo, poiché la clonalità eterogenea nel repertorio BCR deriva dal fatto che la ricombinazione del segmento del gene V, D, J avviene a livelli di DNA indipendentemente in ogni cellula B e, inoltre, l'ipermutazione somatica e l'inserzione e la delezione di nucleotidi nelle giunzioni V-D-J aumentano enormemente la diversità giunzionale, soprattutto all'altezza della regione che

determina la specificità all'antigene, la regione determinante complementarietà 3 (HCDR3).

Tuttavia, i repertori immunitari dei recettori dell'antigene espressi sono costruiti da un sistema integrato di ricombinazione genomica ed espressione controllata e seguono modelli di sviluppo spaziale e temporale complessi, di conseguenza un'analisi efficiente del repertorio richiede sia metodi di campionamento che descrivano la diversità recettoriale ai diversi livelli, sia di mettere in atto strategie di analisi che ricostituiscano il miglior quadro della diversità immunitaria dalle informazioni parziali fornite dalla descrizione del repertorio.

Negli ultimi anni, le tecnologie in rapida crescita per il sequenziamento ad alto rendimento hanno facilitato l'avanzamento della ricerca sul repertorio, consentendo di osservarne la diversità ad una profondità senza precedenti.

La più avanzata tecnologia di throughput in genomica attualmente è il sequenziamento di nuova generazione (NGS) che ha rivoluzionato il campo di indagine dei repertori, e sfruttando questa nuova tecnica, concentrandoci sul repertorio del recettore delle cellule B (BCR) ed esaminando gli approcci che partono dall'isolamento delle cellule B fino alla costruzione di librerie di sequenziamento, è possibile osservare l'ampia diversità del repertorio stesso, ampliando gli orizzonti di comprensione della diversità invisibile e della complessità del sistema immunitario adattativo, tenendo a mente di interpretarla con cautela visto che, a causa della natura massiccia variazione delle sequenze, i dati di sequenziamento sono soggetti ad errori.

I.c – NGS [16–19]

Il sequenziamento di nuova generazione (NGS) dei repertori di immunoglobuline consente di esaminare il sistema immunitario adattativo a un livello senza precedenti. Le applicazioni includono studi sui repertori, sull'uso dei geni, sui livelli di ipermutazione somatica e sull'identificazione della variazione genetica all'interno dei loci delle Immunoglobuline attraverso metodi di inferenza. Tutte queste applicazioni richiedono librerie di partenza che consentano di generare dati di sequenze con un basso tasso di errore ed una rappresentazione ottimale del repertorio espresso. Lo sviluppo di approcci basati su NGS per l'analisi del repertorio di immunoglobuline offre nuove opportunità per studiare le risposte delle cellule B nella salute e nella malattia.

Gli approcci comunemente utilizzati per l'analisi del sequenziamento del repertorio immunitario prevedono la produzione di librerie isotipo-specifiche di cDNA di immunoglobuline, che vengono sequenziate impiegando protocolli NGS per la produzione di ampliconi che comprendono sequenze di primer parziali o sequenze complete dei segmenti genici ricombinati V, D e J delle catene pesanti o delle catene leggere.

Per produrre librerie di lunghezza e profondità sufficienti per l'analisi del repertorio, molti gruppi utilizzano attualmente protocolli Long-read Illumina, come il sistema HiSeq 2 × 250 bp o, più comunemente, il sistema MiSeq 2 × 300 bp, e due importanti tecniche di produzione di librerie, la 5′ Rapid Amplification of cDNA Ends (5′RACE) e la 5′ multiplex (5′MTPX).

Un primo passo importante dell'analisi del repertorio, necessario per una corretta assegnazione dei geni e per l'analisi dell'ipermutazione somatica, è la definizione degli alleli germinali specifici presenti nel soggetto di interesse.

L'attuale database pubblico per i geni germinali Immunoglobuline, il sistema informativo internazionale ImMunoGeneTics (IMGT), include alleli provenienti da un numero relativamente piccolo di individui e quindi copre in modo incompleto la diversità umana globale. Pertanto, è necessario disporre di protocolli per la produzione di librerie adatte all'inferenza dei geni germinali che soddisfino diversi requisiti critici:

In primo luogo, la lunghezza della sequenza della libreria deve essere sufficientemente breve, in modo da non superare i limiti tecnici della tecnologia di sequenziamento utilizzata.

In secondo luogo, le sequenze della libreria devono essere di lunghezza tale da includere l'intera sequenza V(D)J ricombinata compresa tra i primer di amplificazione.

In terzo luogo, l'amplificazione della libreria deve essere imparziale, in modo da consentire l'inclusione di tutti i geni V utilizzati nei repertori di catene pesanti e leggere e rappresentare un alto livello di diversità delle sequenze V(D)J.

Il posizionamento accurato dei primer della regione costante è il mezzo principale per ridurre al minimo la lunghezza degli ampliconi nella produzione di librerie e la localizzazione dei primer in prossimità del confine esonico prossimale è comunemente utilizzata per ridurre al minimo la lunghezza complessiva della sequenza della libreria.

Il confine 5′ della libreria sarà determinato dalla metodologia utilizzata: primer 5′MTPX posizionati nella regione leader o 5′ UTR dei rispettivi geni bersaglio, oppure una sequenza di amplificazione universale di commutazione del template aggiunta a monte della 5′UTR durante la sintesi del cDNA (5′RACE).

Molti degli attuali strumenti di analisi del repertorio sfruttano, aggiunte durante la produzione delle librerie, le UMI solitamente situate all'estremità 3′ delle librerie 5′MTPX e all'estremità 5' delle librerie 5′RACE. L'uso delle UMI consente di identificare le sequenze derivanti dalla stessa molecola di mRNA semplificando la correzione degli errori.

I.d – Clonotipi pubblici di immunoglobuline [8–10]

Il genoma umano contiene approssimativamente 20mila geni codificanti proteine, ma la grandezza della collezione di recettori antigenici data dalla ricombinazione casuale dei segmenti genetici e da aggiunte giunzionali casuali è sconosciuta.
In linea di principio, gli esseri umani possono dare una risposta anticorpale a qualsiasi molecola non self grazie ad un ampio repertorio di anticorpi naïve, la cui diversità è ampliata dall'ipermutazione somatica in seguito all'esposizione all'antigene. Si stima che la diversità del repertorio di anticorpale umano naïve sia di almeno 10^{12} anticorpi unici, ma poiché il numero di cellule B del sangue periferico in un essere umano adulto sano è dell'ordine di 5×10^{9}, la popolazione di cellule B circolanti campiona solo una piccola frazione di questa diversità.
Le analisi su vasta scala dei repertori di anticorpi umani sono proibitivamente difficili a causa sia delle enormi dimensioni, poiché le informazioni codificate da tutti i geni riorganizzati supera le dimensioni del genoma umano di oltre quattro ordini di grandezza, sia dalla localizzazione della maggior della popolazione di linfociti B, che localizzata in organi o tessuti è inaccessibili al campionamento, circoscrivendo lo studio genetico del repertorio anticorpale al set di linfociti B circolanti. Però, grazie alle nuove tecnologie ad alto rendimento, questo limitato set di dati si può osservare con una profondità tale da rivelare si repertori in gran parte unici per ogni individuo, ma anche una sottopopolazione di clonotipi anticorpali condivisi. Infatti, non è stato stabilito se gli individui possiedano repertori unici (privati) o repertori condivisi (pubblici), però l'osservazione di clonotipi anticorpali condivisi, ovvero un insieme di sequenze che utilizzano gli stessi geni V/J e codificano per una sequenza di amminoacidi HCDR3 identica, nei repertori di cellule B e l'identificazione delle sequenze di questi potrebbe consentire una migliore comprensione del ruolo dei repertori dei linfociti B, in quanto la loro esistenza potrebbe essere il risultato di un meccanismo mediante il quale questi clonotipi condivisi vengano selezionati o arricchiti dopo la ricombinazione in base alle proprietà biochimiche delle regioni HCDR3.

II. Metodi

II.a – pRESTO [20,21]

pRESTO è un insieme di tool informatici che fornisce una struttura integrata per gestire tutte le fasi di elaborazione delle sequenze prima dell'assegnazione del segmento germinale, che può essere gestito da altri software, come IMGT/HighV-QUEST, progettato per gestire reads single-end o paired-end, testato su dati provenienti da diverse piattaforme di sequenziamento e comprende un'ampia gamma di funzioni progettate per soddisfare le esigenze di vari protocolli Rep-Seq.
Il pacchetto software pRESTO è fornito come una serie di utility a riga di comando, tutte implementate come moduli di Python applicabili indipendentemente dalla piattaforma di sequenziamento usata, in quanto ogni strumento accetta sequenze sotto forma di file FASTA o FASTQ (con schema di punteggio Phred). Gli strumenti più dispendiosi dal punto di vista computazionale sono parallelizzati, consentendo agli utenti di sfruttare i sistemi specificando il numero di sottoprocessi da eseguire.
Inoltre, per soddisfare le esigenze particolari dei progetti Rep-Seq, pRESTO utilizza uno schema di annotazione che etichetta le singole reads estendendo le descrizioni della sequenza. Queste funzioni di annotazione di pRESTO consentono agli utenti di ordinare e suddividere le sequenze nelle corse multiplex, semplificando il flusso di lavoro e riducendo la possibilità di errore umano.
Ad esempio, in una singola corsa multiplex, l'isotipo del recettore è spesso determinato dalla particolare sequenza del primer della regione costante, il sistema di annotazione di pRESTO associa queste informazioni a ogni read, invece di richiedere un complesso sistema di file separati per ogni serie di annotazioni, semplificando così l'analisi comparativa.
pRESTO fornisce anche:

- Diversi metodi per manipolare queste annotazioni, consentendo di personalizzare le pipeline integrando filtri di sequenza testuali o aritmetici nel flusso di lavoro.

- Strumenti completi di controllo della qualità per filtrare le reads in base alle proprietà della sequenza, come i punteggi di qualità Phred, l'etichettatura valida del codice a barre, l'identità del primer e l'abbondanza di reads duplicate.
- Strumenti per misurare la diversità e i profili di errore di insiemi di reads annotate; tali informazioni possono essere utilizzate per stimare i tassi di errore di sequenziamento e rimuovere dall'analisi gruppi di reads UID altamente variabili, consentendo di migliorare l'accuratezza del sequenziamento e della quantificazione dei dati di repertorio.
- Operazioni speciali su misura per le tecnologie di barcoding UID, tra cui strumenti per l'allineamento multiplo di gruppi di reads UID e la generazione di sequenze di consenso da gruppi di reads UID.
- Supporto nell'assemblaggio de novo di reads paired-end sovrapposte
- Possibilità agli utenti di propagare le annotazioni tra i record paired-end, il che è necessario per i protocolli in cui il codice a barre del campione o l'UID si trovano solo su una read della coppia.

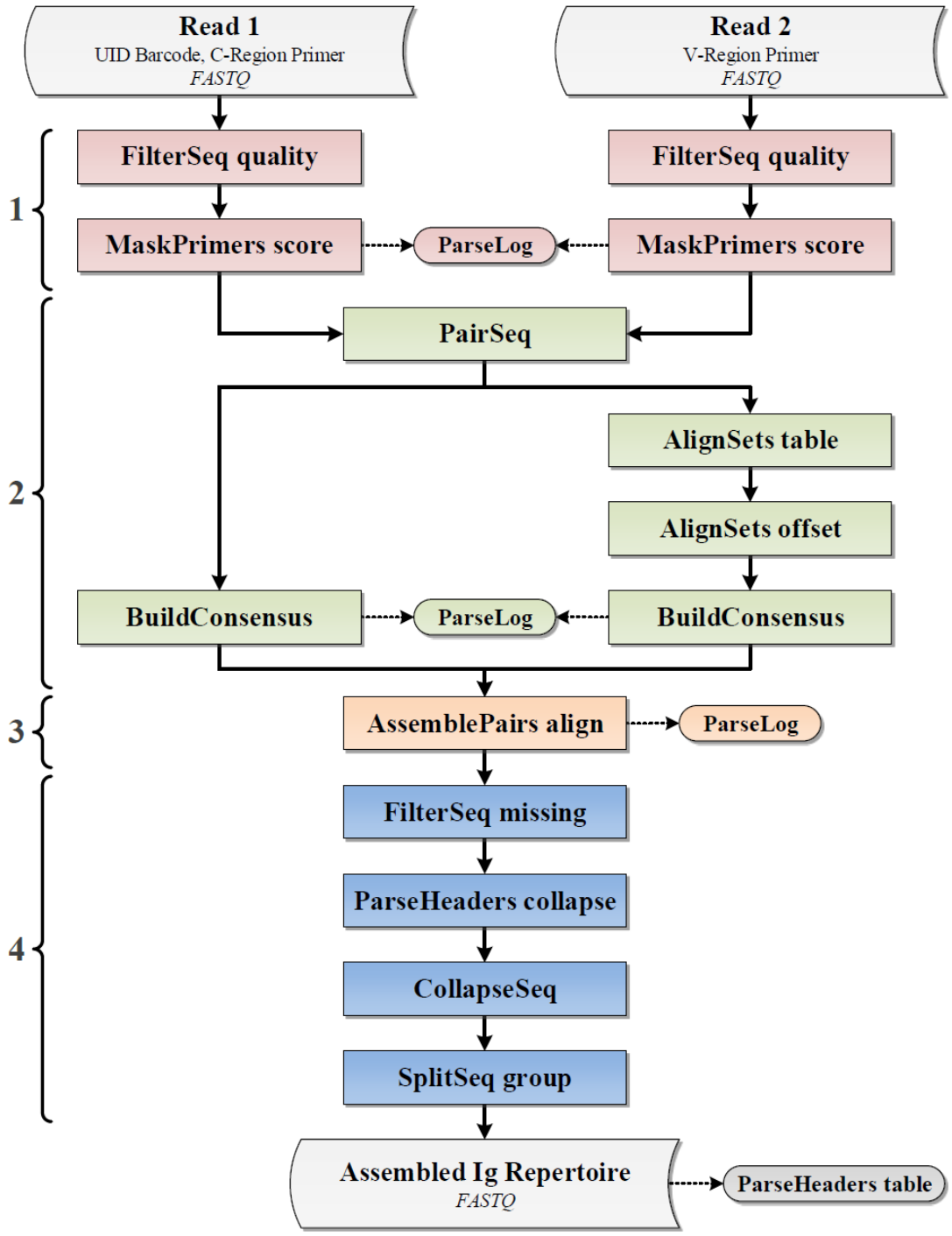
Ottenuto l'output in formato FastaQ del sequenziamento dei campioni effettuato con Illumina MiSeq inizia il primo passaggio di processamento dei dati, scarichiamo librerie di Python per questo scopo contenenti comandi e sotto-comandi da usare per poter elaborare al meglio le sequenze.

Usando questi comandi e relativi sottocomandi processiamo due sequenze FastaQ di esempio per ottenere un'unica sequenza, partendo da:

- Due file FastaQ delle sequenze senso ed antisenso del nostro campione:
  1. R1.fastq
  2. R2.fastq
- La sequenza del Primer Forward:
  3LsdcDNA.fasta
- La sequenza del Primer Reverse:
  IgHC_primers_n2m.fasta
- La sequenza della Regione costante per identificare gli isotipi all'interno del campione:
  IgHC_primers_subisotipe_n2m.fasta

Prima di guardare i comandi ed i sotto comandi specifici utilizzati è utile dare uno sguardo al Flowchart dei vari procedimenti, così da avere una visione intera di quello che avviene.

Partendo da due Read (o sequenze) si effettuano quattro macro passaggi: controlli di qualità ed annotazione delle sequenze, generazione delle sequenze UID di consenso, assemblamento delle porzioni terminali appaiate delle sequenze ed infine filtraggio per ottenere un repertorio ad alta fedeltà.



Ora osserveremo più precisamente i comandi ed i rispettivi sotto-comandi di questi macro passaggi utilizzati.

Con il comando “Split Seq” e relativo sottocomando “Samplepair” estrapoliamo 10.000 sequenze dal nostro file intero:

SplitSeq.py samplepair -1 Intera_R1.fastq -2 Intera_R2.fastq -n 10000

Con il comando “Type” rinominiamo i due file delle sequenze estrapolate:

Type Intera_R1_sample1-n10000.fastq > R1.fastq

Type Intera_ R2_sample1-n10000.fastq > R2.fastq

Con il comando “Filter Seq” e sottocomando “Quality” identifichiamo e rimuoviamo delle reads di bassa qualità, in questo caso le reads con punteggi medi di qualità inferiori a 20:

FilterSeq.py quality -s R1.fastq -q 20 --nproc 4

FilterSeq.py quality -s R2.fastq -q 20 --nproc 4

Con il comando “Mask Primers” e sottocomando “Align” identifichiamo e tagghiamo i primer PCR di entrambe le reads ed annotiamo ciascuna sequenza di reads 1 con il barrcode UID che precede il primer:

MaskPrimers.py align -s R1_quality-pass.fastq --log MP1_Align.out

-p 3LsdcDNA.fasta --maxlen 30 --mode cut --barcode --nproc 4

MaskPrimers.py align -s R2_quality-pass.fastq --log MP2_Align.out

-p IgHC_primers_n2m.fasta --maxlen 30 --mode tag --nproc 4

Con il comando “Parse Log” generiamo una tabella di nome sequenza (ID), identità primer (PRIMER), barrcode UID (BARRCODE) e tasso di errore corrispondenza primer (ERRORE) dai file processati con Mask Primer:

ParseLog.py -l MP1_Align.out MP2_Align.out -f ID PRIMER BARCODE ERROR

Con il comando “Pais Seq” l'annotazione BARCODE identificata da MaskPrimers viene copiata dalla reads 1 alla rispettiva read 2, inoltre vengono rimosse le reads non accoppiate e quelle accoppiate vengono ordinate:

PairSeq.py -1 R1_quality-pass_primers-pass.fastq

-2 R2_quality-pass_primers-pass.fastq --coord illumina --1f BARCODE

Con il comando “Buil Consensus” generiamo una singola sequenza di consenso per ogni barrcode UID. Con --maxdiv vengono rimossi i gruppi di reads UID con

statistiche ad alta diversità, con --prcons per la read 2 rimuoviamo le singole sequenze che non condividono un'annotazione di primer comune con la maggior parte del set, rimuoviamo interi gruppi di reads che hanno assegnazioni di primer ambigue e costruiamo un'assegnazione di primer di consenso per ogni UID:

BuildConsensus.py -s R1_quality-pass_primers-pass_pair-pass.fastq --bf BARCODE --pf PRIMER --maxdiv 0.1 --log bc1.out --nproc 4

BuildConsensus.py -s R2_quality-pass_primers-pass_pair-pass.fastq --bf BARCODE --pf PRIMER --prcons 0.7 --maxdiv 0.1 --log bc2.out --nproc 4

Con il comando “Parse Log” generiamo una tabella di UID (BARRCODE), conteggi di reads (CONSCOUNT), reads di primer di isotipi di consenso 1 (PRCONS) e statistiche sulla diversità (DIVERSITÀ) dai file processati con Build Consenus:

ParseLog.py -l BC1.out BC2.out -f BARCODE CONSCOUNT PRCONS DIVERSITY

Con il comando “Pair Seq” l'annotazione BARCODE identificata da MaskPrimers viene copiata dalla reads 1 alla rispettiva read 2, inoltre vengono rimosse le reads non accoppiate e quelle accoppiate vengono ordinate:

PairSeq.py -1 R1_quality-pass_primers-pass_pair-pass_consensus-pass.fastq -2 R2_quality-pass_primers-pass_pair-pass_consensus-pass.fastq --coord presto

--1f BARCODE

Con il comando “Assemble Pair” e relativo sottocomando “Align” ciascuna sequenza di consenso UID paired-end viene assemblata in una sequenza Ig a lunghezza intera, durante questa fase l'annotazione dell'isotipo di consenso (PRCONS) dalla reads 1 e il numero di reads utilizzate per definire la sequenza di consenso (CONSCOUNT) per entrambe le reads vengono propagate nelle annotazioni della sequenza Ig a lunghezza intera:

AssemblePairs.py align -1 R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass.fastq -2 R2_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass.fastq --coord presto --rc tail --1f CONSCOUNT --2f CONSCOUNT PRCONS --log AP.out --nproc 4

Con il comando “Parse Log” generiamo una tabella con lunghezza di sovrapposizione (OVERLAP), tassi di errore (ERROR) e valori p (PVAL) di ciascuna operazione di assemblaggio:

ParseLog.py -l AP.out -f ID OVERLAP ERROR PVAL

Con il comando “Filter Seq” e sottocomando “Missing” le sequenze con un'alta percentuale di nucleotidi con valori N = 10 della regione non primer vengono rimosse:

FilterSeq.py missing -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass.fastq -n 10 --inner --nproc 4

Con il comando “Parse Header” e sottocomando “Collapse” l'annotazione che specifica il numero di reads non elaborate utilizzate per costruire ciascuna sequenza viene aggiornata:

ParseHeaders.py collapse -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass.fastq -f CONSCOUNT --act min

Con il comando “Collapse Seq” le sequenze nucleotidiche duplicate vengono rimosse basandosi sul requisito che le sequenze duplicate condividano lo stesso primer isotipico:

CollapseSeq.py -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass_reheader.fastq -n 10 --uf PRCONS --cf CONSCOUNT --act sum --inner

Con il comando “Split Seq” e sottocomando “Group” dalle sequenze univoche vengono filtrate quelle con almeno 2 sequenze che contribuiscono:

SplitSeq.py group -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass_reheader_collapse-unique.fastq -f CONSCOUNT --num 2

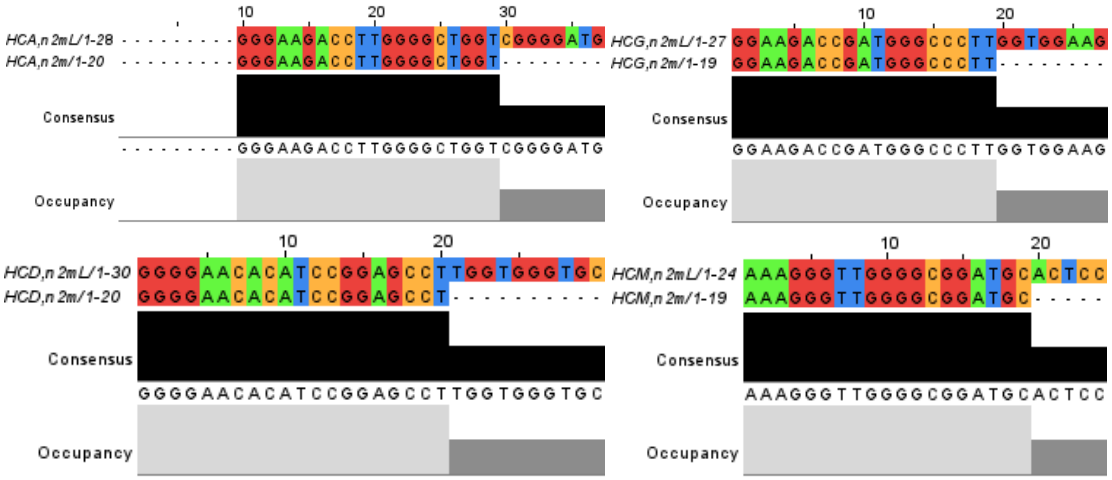
Con il comando “Parse Header” e sottocomando “Table” le annotazioni del repertorio vengono convertite in una tabella:

ParseHeaders.py table -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass_reheader.fastq -f ID PRCONS CONSCOUNT DUPCOUNT

Con il comando “Mask Primers” e relativo sottocomando “Align” identifichiamo e tagghiamo la regione costante della nostra sequenza, così da avere informazioni sul tipo di isotipi nel campione, usiamo --revpr poiché la sequenza usata per l’identificazione della regione costante va “girata” in quanto la reads 2 che contiene le informazioni sull’isotipo è stata “ribaltata”.

Con questo passaggio andiamo a ricercare la sequenza dell’isotipo sul primer reverse sulla regione costante usato per amplificare R2:

MaskPrimers.py align -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass_reheader_collapse-unique_atleast-2.fastq -p IgHC_primers_subisotipe_n2m.fasta --maxerror 0.3 --mode cut --maxlen 30 --pf C_region --log MP2_IGHC.out --nproc 4 --revpr4



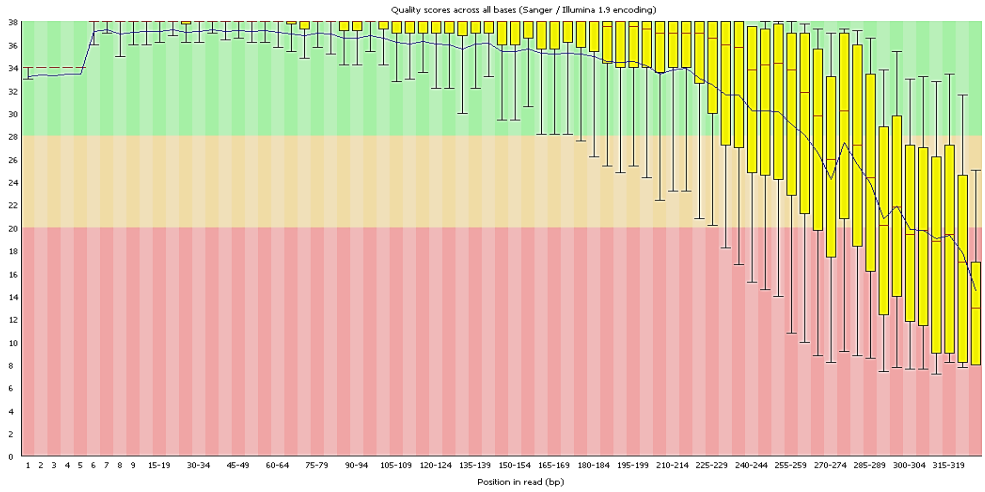
Con il comando “Parse Header” e relativo sottocomando “Table” le annotazioni del repertorio finale vengono quindi convertite in una tabella:

ParseHeaders.py table -s R1_quality-pass_primers-pass_pair-pass_consensus-pass_pair-pass_assemble-pass_missing-pass_reheader_collapse-unique_atleast-2_primers-pass.fastq -f ID PRCONS CONSCOUNT DUPCOUNT C_region

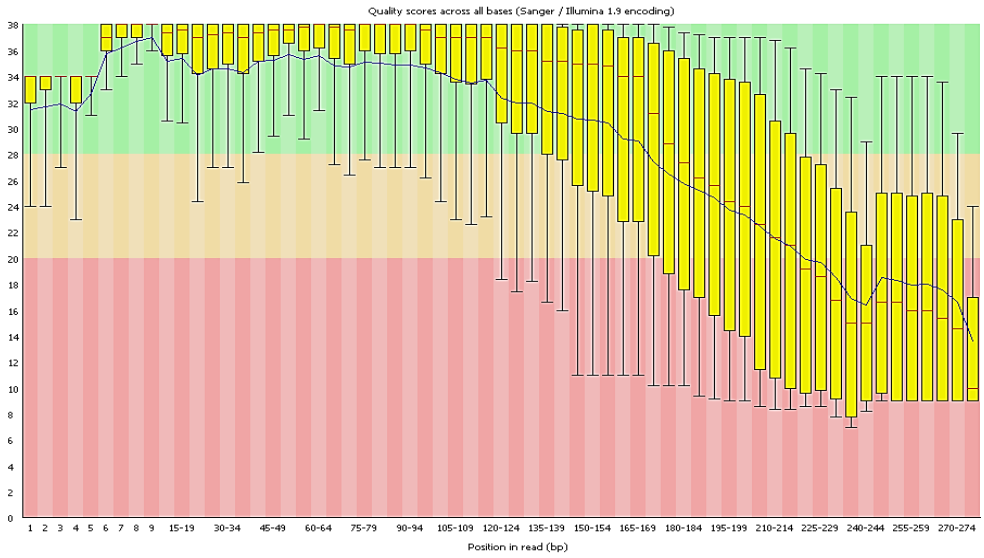
In tutti questi passaggi abbiamo la generazione di nuovi file FastaQ, queste sequenze ogni volta diverse da quelle di prima possono essere analizzate tramite FastaQC, software che permette di analizzare graficamente ciò che succede da un passaggio all’altro.

Inizialmente quello che possiamo osservare per le due diverse sequenze è questo:

Intera_R1

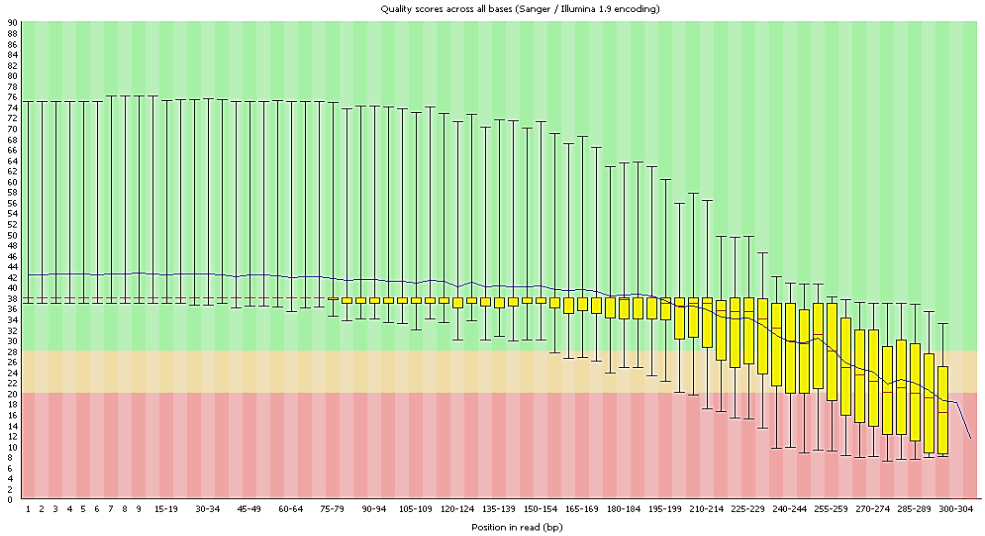


Intera_R2

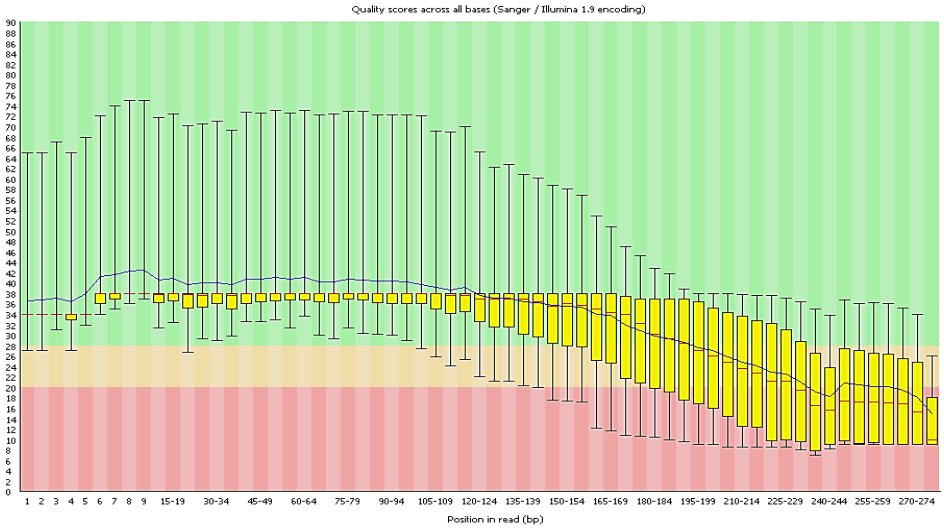


Dopo l'ultimo PairSeq per ogni sequenza quello che possiamo osservare è:

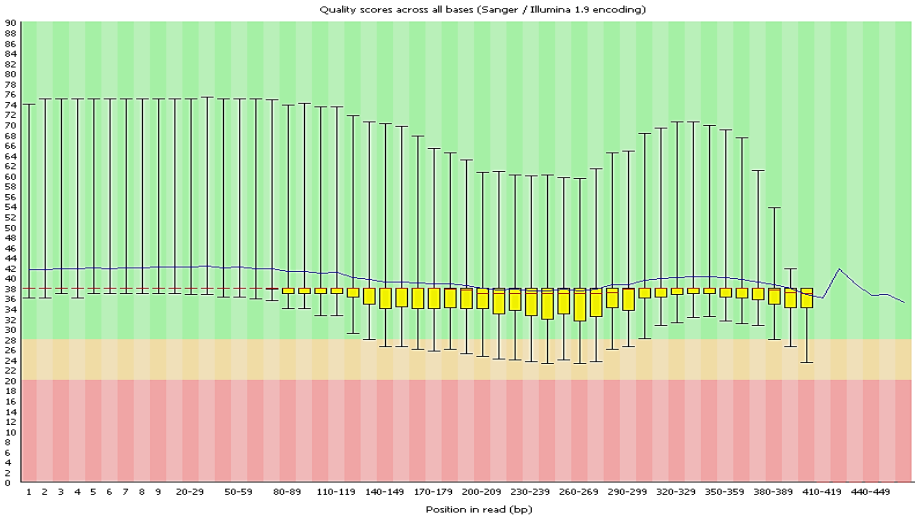
PairSeq_R1:



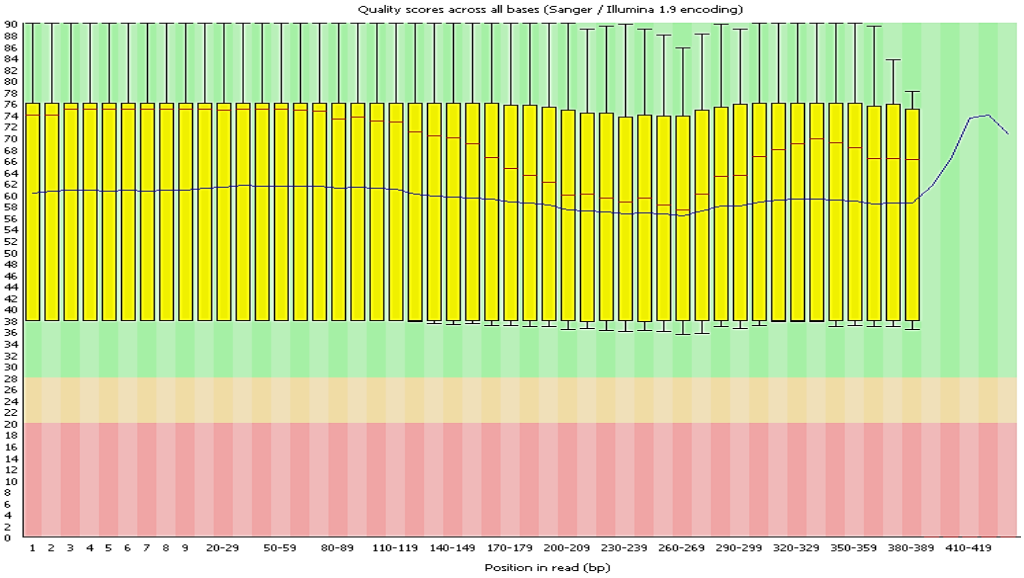
PairSeq_R2:



Dopo Assemble Pairs align quello che osserveremo è:



Alla fine di tutto, dopo l'ultimo Mask Primers align, dall'appaiamento delle due sequenze iniziali R1 e R2 avremo un'unica sequenza, con risultato:



II.b – Change-O [22]

La caratterizzazione su larga scala dei repertori di immunoglobuline delle cellule B, l'elevata diversità germinale e somatica del repertorio immunoglobulinico rappresenta una sfida per un'analisi biologicamente significativa, che richiede metodi computazionali specializzati e Change-O fornisce strumenti per analisi avanzate di dati di sequenziamento del repertorio immunoglobulinico su larga scala. Change-O comprende strumenti per determinare l'insieme completo degli alleli della regione variabile delle immunoglobuline di un individuo, per suddividere le sequenze di in popolazioni clonali, creare alberi di discendenza, dedurre modelli di targeting dell'ipermutazione somatica, misurare la diversità del repertorio, quantificare la pressione di selezione e calcolare le proprietà chimiche della sequenza, utilizzano un formato di dati comune, che consente di integrare perfettamente più analisi in un unico flusso di lavoro. L'estrazione di misure di interesse biologico e clinico dal repertorio germinale annotato è un processo lungo e soggetto a errori. Qui entra in gioco Change-O, un toolkit che copre una serie di compiti di analisi complessi per i dati di sequenziamento del repertorio immunoglobulinico. La suite Change-O è composta da quattro pacchetti software: una raccolta di pipeline di Python (changeo-ctl) e tre pacchetti R separati.

I dati vengono passati alle utility di Change-O sotto forma di file di testo delimitati da tabelle, ciascuna utility identifica i dati di input rilevanti sulla base di nomi di colonne standardizzati e aggiunge nuove colonne al file con le informazioni di output da portare alla fase di analisi successiva. Ad esempio, Change-O fornisce strumenti per importare i dati dallo strumento IMGT/HighV-QUEST, utilizzato di frequente, e una serie di utility per eseguire operazioni di base sul database, come l'ordinamento, il filtraggio e la modifica delle annotazioni.

I componenti più costosi dal punto di vista computazionale dispongono di un supporto multiprocessing integrato includendo anche una documentazione dettagliata e una registrazione opzionale degli errori. Change-O, insieme a pRESTO, fornisce i componenti chiave di un ecosistema analitico che consente un'analisi sofisticata dei dataset di sequenziamento del repertorio immunoglobulinico ad alto rendimento. Usando Change-O e sottomettendogli i

file da IMGT andiamo a generare i nostri database delle sequenze con relative annotazioni, per farlo usiamo il comando:

Makedb.py imgt -i "File da IMGT" -s "File Seq FASTA" -r "File Germline"

Dove: -i è il file ottenuto da IMGT, -s è il file FASTA della sequenza non processata da IMGT e -r è il file FASTA della germline

II.c – Rstudio

RStudio è un software potente e semplice di interagire con la programmazione R, considerato come un ambiente di sviluppo integrato che fornisce una soluzione unica per tutti i calcoli statistici e grafici. RStudio è una versione più avanzata di R, dotata di una finestra a più riquadri che consente di accedere a tutti gli elementi principali in un'unica schermata, come sorgente, console, ambiente e cronologia, file, foto e grafici. Impiegando questo ambiente di lavoro come prima cosa grazie ad una serie di pipeline manipoliamo i database, così da avere metadati ed annotazioni accessibili ed utili. Osserviamo una pipeline di esempio impiegate per manipolare i file dei sequenziamenti che andranno a costituire il database per la mia indagine sulle Public Sequence.

Importiamo i dati del sequenziamento "Ex.tvs", creando "dbEX" database di partenza:

```
dbEX <- readChangeoDb("Ex.tsv")
```

Selezioniamo solo le sequenze con caratteristiche di qualità che a noi interessano:

```
dbEX <- dbEX %>%
     filter(consensus_count >= 2, productive == TRUE,
            stop_codon == FALSE, vj_in_frame == TRUE)
```

Dalla colonna identificativa, unica per ogni sequenza, estrapoliamo le informazioni su campione (Sample), Sottotipo (Subset) e Tessuto (Tissue), generando tre nuove colonne.

```
dbEX <-  dbEX %>%
     separate(sequence_id, c("Sample", "Subset", "Tissue"),
              sep = "_", remove = FALSE, extra = "drop")
```

Andando a rimuovere elementi dalla colonna v_call otterremo informazioni sul gene V:

```
dbEX$v_call <- gsub("Homsap ", "", dbEX$v_call)
dbEX$v_gene <- sapply(strsplit(dbEX$v_call, "[*]"), "[", 1)
dbEX$v_gene <- gsub("IGHV", "", dbEX$v_gene)
```

Andando a rimuovere elementi dalla colonna d_call otterremo informazioni sul gene D:

```
dbEX$d_call <- gsub("Homsap ", "", dbEX$d_call)
dbEX$d_gene <- sapply(strsplit(dbEX$d_call, "[*]"), "[", 1)
dbEX$d_gene <- gsub("IGHD", "", dbEX$d_gene)
```

Andando a rimuovere elementi dalla colonna j_call otterremo informazioni sul gene J:

```
dbEX$j_call <- gsub("Homsap ", "", dbEX$j_call)
dbEX$j_gene <- sapply(strsplit(dbEX$j_call, "[*]"), "[", 1)
dbEX$j_gene <- gsub("IGHJ", "", dbEX$j_gene)
```

Andando a rimuovere elementi dalla colonna cregion otterremo informazioni sull'isotipo:

```
dbEX$cregion <- sapply(strsplit(dbEX$cregion, "-"), "[", 1)
```

Usando le informazioni contenute nella colonna del gene V generiamo una nuova variabile, per avere informazioni sulla famiglia di geni V:

```
dbEX$v_family <- sapply(strsplit(dbEX$v_gene, "-"), "[", 1)
dbEX$v_family <- gsub("IGHV", "", dbEX$v_family)
```

Usando la colonna con la lunghezza della giunzione nucleotidica (junction_length)e dividendo la lunghezza per tre, numero di nucleotidi per codone amminoacidico, generiamo una nuova stinga con il numero di amminoacidi nella regione HCDR3 (N_aa_junction):

```
dbEX <- dbEX %>%
     mutate(N_aa_junction = junction_length/3)
```

Infine, andiamo a selezionare le variabili utili alla nostra analisi bioinformatica:

```
dbEX <- dbEX %>%
        select(sequence_id, Sample, Subset, Tissue,
               junction_length, N_aa_junction,
               cregion, v_family, v_gene, d_gene, j_gene,
               sequence, cdr3, junction_aa)
```

Dove sequence_id è il codice identificativo unico per ogni sequenza, Sample è la sigla identificativa del campione, Subset è il sottotipo del Linfocita B, Tissue è il tessuto di provenienza del campione, junction_length è il numero di nucleotidi della regione HCDR3, N_aa_junction è il numero di amminoacidi della regione HCDR3, cregion è l'isotipo, v_family è la famiglia del gene V espresso nella sequenza, v_gene è il gene V espresso nella sequenza, d_gene è il gene D espresso nella sequenza, j_gene è il gene J espresso nella sequenza, sequence è la sequenza nucleotidica per intero, HCDR3 è la sequenza nucleotidica della regione HCDR3 e junction_aa è la sequenza amminoacidica della regione HCDR3.

Facendo "correre" queste poche righe di codice si ottiene un database di partenza con tutte le informazioni utili all'analisi bioinformatica:

	sequence_id	Sample	Subset	Tissue	junction_length	N_aa_junction	cregion	v_family	v_gene	d_gene	j_gene	sequence	cdr3	junction_aa
1	B10_DN_PB	B10	DN	PB	63	21	IgA	1	1-18	4-11	6	CAGGTTCAGCTGGTGC...	GCGAGATGGGCCAGAC...	CARWARLTVTTGRY...
2	B10_DN_PB	B10	DN	PB	36	12	IgM	3	3-66	1-26	4	GAGGTGCAGCTGGTG...	GCGAGATTGAGTGGGA...	CARLSGSGRDYW
3	B10_DN_PB	B10	DN	PB	84	28	IgA	3	3-43	3-9	6	GAGGTGCAGCTGGTG...	GCAACAGAAACTCGGC...	CATETRPPLRNILSG...
4	B10_DN_PB	B10	DN	PB	54	18	IgA	3	3-64D	6-6	4	GAGGTGCAGCTGGTG...	GTGAAAGATCGATATAG...	CVKDRYSIYSLGIDL...
5	B10_DN_PB	B10	DN	PB	84	28	IgA	3	3-43	3-9	6	GAAGTGCAGCTGGTG...	GCAACAGAAACTCGGC...	CATETRPPLRNILSG...
6	B10_DN_PB	B10	DN	PB	45	15	IgG	1	1-3	1-26	6	CAGGTCCAGCTTGTGC...	GCGAGAAGCAACCACT...	CARSNHYYYYGMD...
7	B10_DN_PB	B10	DN	PB	72	24	IgA	1	1-18	2-15	5	CAGGTTCAGCTGGTGC...	GCGAGATTTGGTAGGG...	CARFGRDVYDCSG...
8	B10_DN_PB	B10	DN	PB	60	20	IgM	1	1-2	3-10	5	CAGGTGCAGCTGGTG...	GCGAGAGAGAAGGGTT...	CAREKGYYGSGTIP...
9	B10_DN_PB	B10	DN	PB	42	14	IgG	2	2-5	3-3	4	CAGATCACCTTGAGGG...	GCACTCTCCTCCGTTTG...	CALSSVWSGIMGYW
10	B10_DN_PB	B10	DN	PB	60	20	IgM	5	5-51	2-21	3	GAGGTGCAGCTGGTG...	GCGAGACTAGGGGCTT...	CARLGAFVVVTAIP...
11	B10_DN_PB	B10	DN	PB	33	11	IgG	3	3-21	3-16	4	GAGGTGCAGCTGGTG...	ACGAATACGGGTCAGCT...	CTNTGQLSDYW
12	B10_DN_PB	B10	DN	PB	66	22	IgA	1	1-18	6-19	6	CAGGTTCAGCTGGTGC...	GCGAGATGGTCGCGAAT...	CARWSRIRAVAGTS...
13	B10_DN_PB	B10	DN	PB	60	20	IgG	4	4-34	6-13	5	CAGGTGCAGCTTCAGC...	GCGAGAGACGGTAGCG...	CARDGSDTWFAIRP...
14	B10_DN_PB	B10	DN	PB	63	21	IgG	1	1-2	2-15	5	CAGGTGCAGCTGGTG...	GCGAGAGGTTGTAGTG...	CARGCSGGRCYPPL...
15	B10_DN_PB	B10	DN	PB	54	18	IgA	3	3-48	5-24	6	GCGGTGCAGCTGCTG...	GCGNGAGTGGGAGATG...	CAXVGDGHTSFYF...
16	B10_DN_PB	B10	DN	PB	51	17	IgM	1	1-18	4-23	2	CAGGTTCAGCTGGTGC...	GCGAGAGCGACCCGGG...	CARATRGGNSYWV...
17	B10_DN_PB	B10	DN	PB	51	17	IgG	4	4-39	3-3	4	CAGCTGCAGCTGCAG...	GCGAGACAACGGGGAC...	CARQRGRFLEWLLT...
18	B10_DN_PB	B10	DN	PB	54	18	IgA	1	1-46	3-16	5	CAGGTGCGGCTGGTTC...	GCGAGACAGTTCTGCTC...	CARQFCSETACHF...
19	B10_DN_PB	B10	DN	PB	84	28	IgA	3	3-43	3-9	6	GAAGTGCAGCTGGTG...	GCAACAGAAACTCGGC...	CATETRPPLRNILSG...
20	B10_DN_PB	B10	DN	PB	45	15	IgG	3	3-23	5-24	1	GAGGTGCAGCTGTTG...	GCGAAAATGCCCACATT...	CAKMPTFSMVLLLG...

Usando queste pipeline su tutte le sequenze dei 57 campioni otterremo, in seguito alla loro unione, un database composto 14 colonne ed un numero di righe nell'ordine di 10^6 che ovviamente verrà manipolato e ridotto fino a contenere solo informazioni utili.

Sfruttando il potere di calcolo di R ed un ampio numero di librerie quali *dplyr*, *rlist*, *purrr*, *data.table*, *tidyr*, *tidyverse*, *biostrings*, *stringdist*, *rowr*, *alakazam*, *shazam*, *ggplot2*, *ggseqlogo* e *scales*, e tutti i loro comandi e sotto-comandi è stato possibile ottenere dati utili ed indagare tutta la ricchezza del repertorio immunoglobulinico grazie a grafici, focalizzando l'attenzione sulle Sequenza Pubbliche. Sequenze che, come ricordiamo, si possono distinguere secondo più definizioni: Sequenze con uguali amminoacidi al HCDR3, Sequenze con uguali amminoacidi al HCDR3 e stesso gene V e Sequenze con uguali amminoacidi al HCDR3, stesso gene V e stesso gene J.

Prima di andare a fare indagini mirate bisogna ottenere i dati necessari, in quanto le sequenze pubbliche sono una piccola percentuale del repertorio e, quindi, prima dobbiamo isolarle dalle altre.

Per fare ciò ci siamo basati sulla definizione "base" di sequenza pubblica, ovvero: catene immunoglobuliniche che condividono la stessa sequenza nella regione variabile del HCDR3:

```
FL_Tot <-  list.files(pattern = "-db.tsv",
                      full.names = FALSE)
US_TOT_Vec <- foreach(n = FL_Tot, .combine = c) %do% {
   fread(n, select = "junction_aa") %>%
      distinct() %>%
      flatten() %>%
      unlist()}
```

Come punto di partenza abbiamo generato un vettore contenente solo le sequenze amminoacidiche della giunzione (*junction_aa*) dei 60 campioni. Da questo vettore tramite "*duplicated*" e "*%in%*" ricerchiamo ed estraiamo solo le sequenze duplicate che ci serviranno:

```
Duples_TOT <- US_TOT_Vec[duplicated(US_TOT_Vec)|
                         duplicated(US_TOT_Vec,
                                    fromLast = T) == 1]
Dup_T <- US_TOT_Vec[US_TOT_Vec%in%Duples_TOT]
```

Grazie a "count" contiamo quante giunzioni duplicate abbiamo trovato e, per avere un dato consultabile anche in seguito, generiamo un dataframe ordinato per sapere quante volte la sequenza duplicata compare nel vettore.

```
Dup_T_c <- plyr:: count(Dup_T)
  colnames(Dup_T_c) <- c("Junction_aa", "Count")
    Dup <- data.frame(Dup_T_c)
```

A questo punto generiamo un database con tutte le sequenze dei 60 campioni con tutte le 14 variabili di partenza:

```
FL_TOT <-  list.files(pattern = "-db.tsv",
                      full.names = FALSE)
```

```
US_TOT_DF <- foreach(n = FL_TOT, .combine = smartbind) %do%
{fread(n, select = c("sequence_id", "Sample", "Tissue",
                     "junction_aa", "junction_length",
                      "N_aa_junction", "cdr3",
                       "v_family", "v_gene", "j_gene"))%>%
 distinct(junction_aa, .keep_all = T)}
```

Ora sfruttando il vettore con solo i HCDR3 duplicati andiamo a generare un database con le sequenze duplicate:

```
Duples_TOT_DF <- US_TOT_DF %>%
   mutate(dup = junction_aa%in%Dup$Junction_aa) %>%
   filter(dup == "TRUE")
```

Inoltre, usando sempre lo stesso vettore generiamo un database di controllo senza le sequenze duplicate che contiene le sequenze uniche:

```
DF_none_Duples <- US_TOT_DF %>%
 mutate(dup = US_TOT_DF$junction_aa%in%Dup$Junction_aa) %>%
   filter(dup == "FALSE")
```

Però al database generale delle sequenze duplicate mancano informazioni necessarie per andare avanti con l'analisi.

Innanzitutto, definiamo il Clan di appartenenza di ogni sequenza, basandoci sulla colonna del Gene V:

```
Duples_TOT_DF <- Duples_TOT_DF %>%
 mutate(Clan_VH =
        case_when(v_family == 'IGHV1'|v_family == 'IGHV5'|
                    v_family == 'IGHV7' ~ 'Clan_1',
                  v_family == 'IGHV2'|v_family == 'IGHV4'|
                    v_family == 'IGHV6' ~ 'Clan_2',
                  v_family == 'IGHV3' ~ 'Clan_3'))
```

Facciamo la stessa cosa per il database di controllo con le sequenze uniche:

```
DF_none_Duples <- DF_none_Duples %>%
   mutate(Clan_VH =
      case_when(v_family == 'IGHV1'|v_family == 'IGHV5'|
                    v_family == 'IGHV7' ~ 'Clan_1',
                v_family == 'IGHV2'|v_family == 'IGHV4'|
                    v_family == 'IGHV6' ~ 'Clan_2',
                v_family == 'IGHV3' ~ 'Clan_3'))
```

Per riuscire ad identificare e discriminare le sequenze pubbliche tra di loro, ovvero riconoscere quali hanno la stessa sequenza amminoacidica al HCDR3, andiamo ad aggiungere una colonna al database che contiene un tag identificativo che ci permette proprio di fare ciò.

Con "*match*", andiamo a riconoscere le sequenze amminoacidiche uguali in *junction_aa* ed aggiungere un numero identificativo unico per ogni sequenza nella colonna *Pub_CDR3*, che poi verrà modificato per essere più comprensibile:

```
Dupples_TOT_DF <- Dupples_TOT_DF %>%
   transform(Pub_CDR3 = match(junction_aa,
                              unique(junction_aa)))
Dupples_TOT_DF$Pub_CDR3 <- lapply(Dupples_TOT_DF$Pub_CDR3,
                                  function(x)paste("Jn", x,
                                                   sep="."))
Dupples_TOT_DF$Pub_CDR3<-as.character(Dupples_TOT_DF$Pub_CDR3
)
```

Ora nel database c'è una colonna che ci permetterà di riconoscere le sequenze pubbliche, permettendo di muovere i primi passi verso un'analisi più approfondita del repertorio pubblico.

Visto che la definizione più profonda di sequenza pubblica prende in considerazione anche i geni V, andiamo a generare una colonna simile alla precedente, che conterrà le informazioni sul gene. Per fare ciò ci basterà andare a manipolare la colonna già esistente, ma accorciando il nome ci consentirà una visualizzazione più veloce e comprensibile:

```
Duples_TOT_DF <- Duples_TOT_DF %>%
  mutate(ID_GeneV = gsub("IGHV", "", Duples_TOT_DF$v_gene))
```

Per aver dati ancora più accessibili, facili da leggere, manipolare ed interpretare inseriamo la colonna Tag_Cl_Fm_Gn identificativa, con le informazioni per la sequenza HCDR3 pubblica, il clan del gene V, la famiglia ed il gene V:

```
Duples_TOT_DF <- Duples_TOT_DF %>%
   unite(Tag_Cl_Fm_Gn, c("Pub_CDR3", "Clan_VH", "v_gene"),
         sep = "_", remove = F) %>%
   relocate(Tag_Cl_Fm_Gn, .after = ID_GeneV)
```

Facciamo la stessa cosa anche per una seconda colonna *Tag_Jn_Gn* che però conterrà solo le informazioni sulla sequenza HCDR3 pubblica ed il gene V:

```
Duples_TOT_DF <- Duples_TOT_DF %>%
   unite(Tag_Jn_Gn, c("Pub_CDR3", "ID_GeneV"),
         sep = "_", remove = F) %>%
   relocate(Tag_Jn_Gn, .after = Tag_Cl_Fm_Gn)
```

Usando le nuove variabili generate generiamo due sotto-database contenenti informazioni più stringenti, che ci permetteranno di fare le analisi.

Il primo è un sotto-database con sequenze pubbliche solo per stesso HCDR3

```
US_DUP_Jn <- Duples_TOT_DF %>%
   dplyr:: group_by(Pub_CDR3) %>%
   dplyr:: summarise(Q = n()) %>%
   dplyr:: filter(Q > "1")
PB_Jn <- Duples_TOT_DF %>%
   mutate(Q = Pub_CDR3%in%US_DUP_Jn$Pub_CDR3) %>%
   filter(Q == "TRUE")
```

Il secondo è un sotto-database con sequenze per stesso HCDR3 e gene V:

```
NN_DUP_JG <- Duples_TOT_DF %>%
   dplyr:: group_by(ID_GeneV, Pub_CDR3, Tag_Jn_Gn) %>%
   dplyr:: summarise(Q = n()) %>% dplyr:: filter(Q > "1")
```

```
PB_Jn_vG <- Duples_TOT_DF %>%
   mutate(TF = Tag_Jn_Gn%in%NN_DUP_JG$Tag_Jn_Gn) %>%
   filter(TF == "TRUE")
```

Infine, come per ogni studio valido grazie a "*slice_sample*" generiamo due database di controllo andando a sortare casualmente sequenze del database di controllo con solo le sequenze uniche:

```
Controll <- DF_none_Duples %>%
   dplyr:: slice_sample(n = 22e+04) %>%
   mutate(ID_GeneV = gsub("IGHV", "", Controll$v_gene))

Controll_1 <- DF_none_Duples %>%
   dplyr::slice_sample(n = 866e+02) %>%
   mutate(ID_GeneV = gsub("IGHV", "", Controll_1$v_gene))
```

Facendo correre queste pipeline otterremo i dati che ci permetteranno di effettuare la nostra analisi bioinformatica ed andare ad indagare le caratteristiche di questo repertorio di sequenze pubbliche e verificare se la loro esistenza sia dovuta al caso o ad un qualche meccanismo di selezione messo in atto durante il processo maturativo delle cellule B.

II.d Analisi statistiche

In modo da dare una base statistica alle analisi effettuate sul repertorio di clonotipi pubblici sono stati applicati l'indice di Pearson, per osservare il grado di correlazione tra i dati del repertorio di clonotipi pubblici ed il repertorio di controllo, ed il test del chi-quadro per osservare grazie al p-value se le differenze siano: non significative (p-value > 0,05), significative (p-value < 0,05), molto significative (p-value < 0.01), estremamente significative (p-value < 0.001).

III. Risultati

Identificazione delle sequenze pubbliche

Abbiamo analizzato il repertorio di catene pesanti di immunoglobuline di 56 donatori, con una media di 125.700 sequenze per donatore (da un minimo di 5000 a un massimo di 442.800 sequenze) (Tabella 1). Abbiamo quindi individuato clonotipi pubblici, definiti come sequenze trovate in almeno 2 donatori che condividano il gene *IGHV* e la sequenza amminoacidica del *HCDR3*, con una media di ~1400 sequenze per donatore (da 130 a 7900 sequenze per donatore).

Campioni	Sequenze iniziali	Clonotipi	Campioni	Sequenze iniziali	Clonotipi
S1	268556	1297	S29	137693	1227
S2	47673	293	S30	164052	1317
S3	44544	308	S31	162869	1071
S4	143630	2493	S32	138759	1580
S5	125952	907	S33	227544	2248
S6	117599	1007	S34	193643	2364
S7	77523	543	S35	265521	1870
S8	21471	1739	S36	222344	2016
S9	121114	779	S37	155229	914
S10	132656	1122	S38	84092	642
S11	132978	1118	S39	57030	429
S12	148235	1077	S40	52718	362
S13	101574	2208	S41	5501	130
S14	97447	2378	S42	138303	668
S15	68090	636	S43	173966	848
S16	54403	309	S44	40315	276
S17	47906	271	S45	54105	540
S18	69300	530	S46	42513	675
S19	94221	1098	S47	60476	899
S20	95447	1027	S48	250641	2255
S21	79722	963	S49	177509	7918
S22	48139	304	S50	60636	6512
S23	205330	1359	S51	19069	608
S24	303275	2288	S52	70374	766
S25	259834	3266	S53	68752	657
S26	280910	2695	S54	65526	940

La percentuale di clonotipi pubblici nel repertorio di ogni donatore (Figura 1.a) mostra che la maggior parte è compresa tra lo 0,5 ed il 3% con una mediana dello 0,8%.

Si può osservare una relazione lineare (indice di Pearson = 0,741, p < 0,001) tra la dimensione del repertorio campionato e quello pubblico di ogni donatore (Figura 1.b), indicando che almeno parte della varianza osservata sia dovuta alle differenze di campionamento.

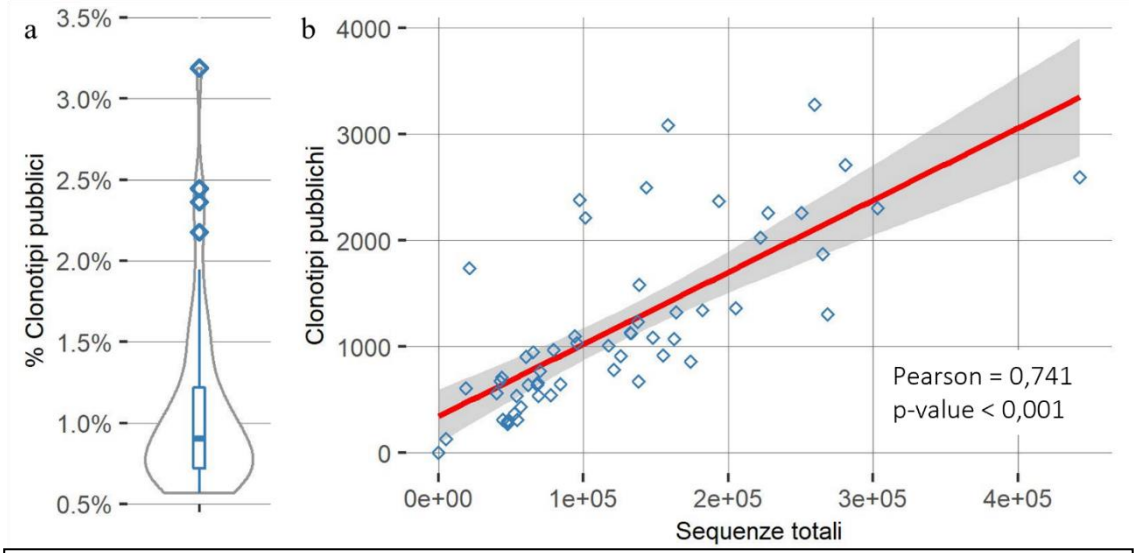


Figura 1
(a) Frequenza di clonotipi pubblici per campione
(b) Correlazione lineare (linea rossa) e p-value tra database totale e database di clonotipi pubblici

Lunghezze HCDR3 nei clonotipi pubblici

La distribuzione delle lunghezze amminoacidiche degli *HCDR3* dei clonotipi pubblici (Figura 2.a), mostra un andamento normale con la mediana, in corrispondenza di 14 aminoacidi, affiancata da percentuali non indifferenti di lunghezze amminoacidiche comprese tra 11 e 16 aminoacidi.

La distribuzione delle lunghezze del repertorio pubblico a confronto con un database di controllo (Figura 2.a), ottenuto campionando casualmente il repertorio completo dei donatori, denota come vi sia una differenza considerevole nella distribuzione delle lunghezze, in quanto nel repertorio pubblico si osserva un arricchimento di *HCDR3* corti rispetto al repertorio di controllo (mediane rispettivamente di 14 amminoacidi e 18 amminoacidi).

La maggior parte dei clonotipi pubblici è condiviso solamente tra 2 campioni (Figura 2.b), ma analizzando il 95 percentile dei clonotipi più condivisi si denota come vi siano 50 clonotipi condivisi in più di 10 soggetti fino ad una condivisione

massima in ben 21 soggetti e la che maggior parte delle sequenze più condivise è racchiusa in un range di lunghezze *HCR3* tra i 7 ed i 15 amminoacidi.

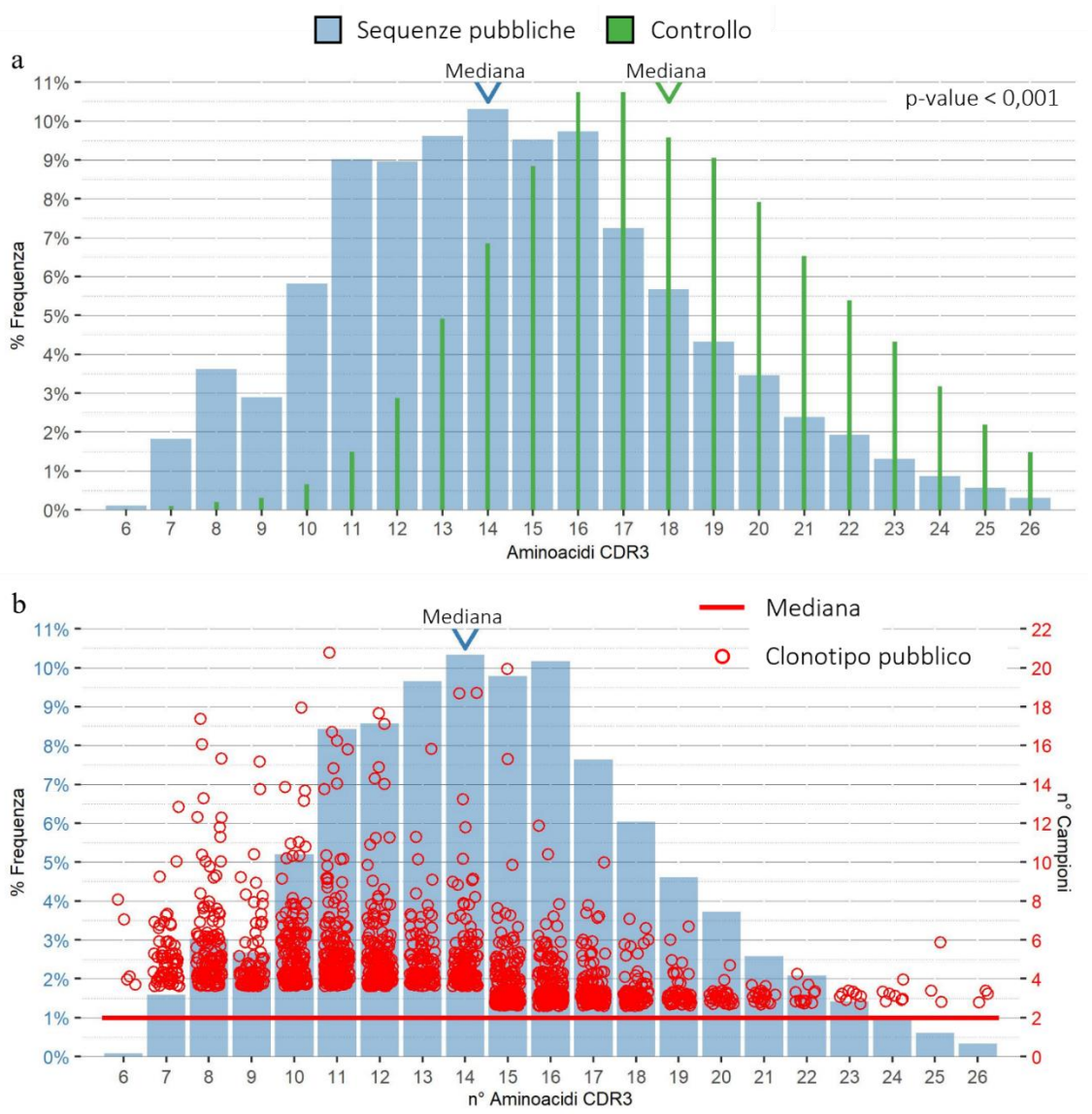


Figura 2
(a) Distribuzione lunghezze amminoacidiche dei CDR3 pubblici (blu) e controllo (verdi)
(b) Distribuzione lunghezze amminoacidiche dei CDR3 pubblici (blu) e condivisione dei clonotipi (rosso), con mediana di condivisione (linea orizzontale) e dati del 95 per centile

Stato mutazionale dei geni IGHV nei clonotipi pubblici

La frequenza di mutazioni del repertorio pubblico messa a confronto con la frequenza di mutazione del repertorio di controllo (Figura 3.a) mostra come i clonotipi pubblici siano arricchiti per sequenze mutate rispetto al repertorio di controllo.

Guardando le mutazioni delle singole lunghezze *HCDR3* del repertorio pubblico (Figura 3.b) si osserva un trend dove le sequenze più lunghe siano più mutate, al contrario delle più corte che invece arricchite di germline.

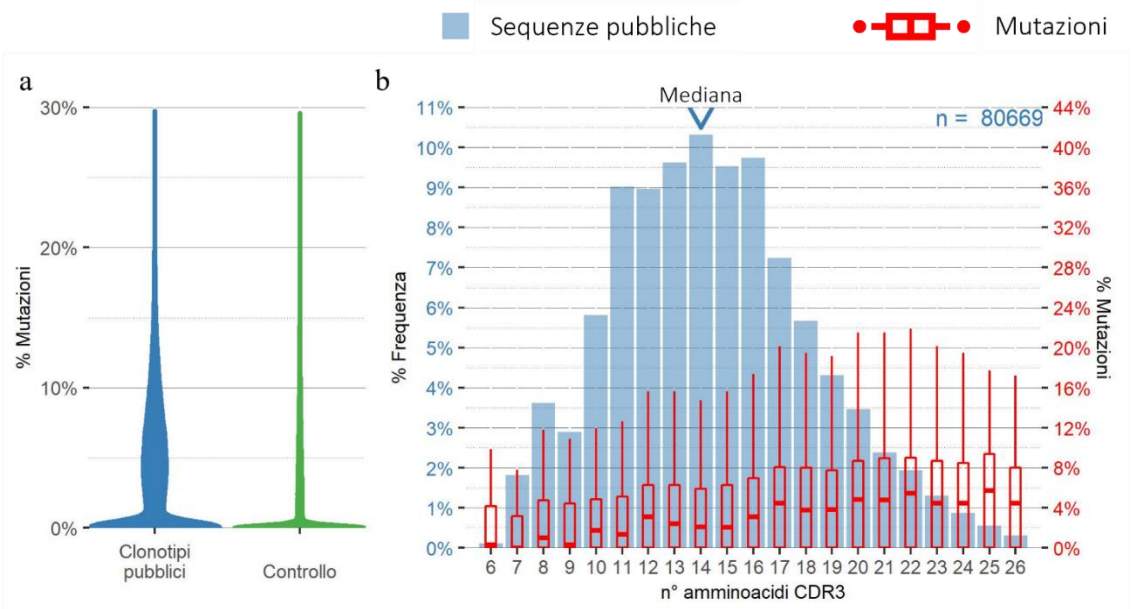


Figura 3
(a) Frequenza di mutazioni nel database pubblico (blu) e nel controllo (verde)
(b) Distribuzione lunghezze amminoacidiche dei CDR3 pubblici (blu) e frequenza di mutazioni per singola lunghezza (rosso)

Repertorio dei geni IGHV nei clonotipi pubblici

Confrontando l'uso dei geni *IGHV* del repertorio pubblico con il controllo (Figura 4.a) è possibile apprezzare la differenza di espressione, più o meno marcata, tra i due repertori; con geni come il *IGHV* 3-23 o il *IGHV*3-7 significativamente più espressi nel repertorio controllo ($p < 0{,}001$ e $p < 0{,}01$) o come l'1-69 più espresso nel repertorio pubblico ($p < 0{,}001$).

Osservando il rapporto tra l'espressione di diversi geni *IGHV* tra i clonotipi pubblici e il repertorio di controllo (Figura 4.b), si nota come appena una decina di geni siano maggiormente espressi nei clonotipi pubblici arrivando, in alcuni casi, ad essere più del doppio di quelli nel repertorio di controllo.

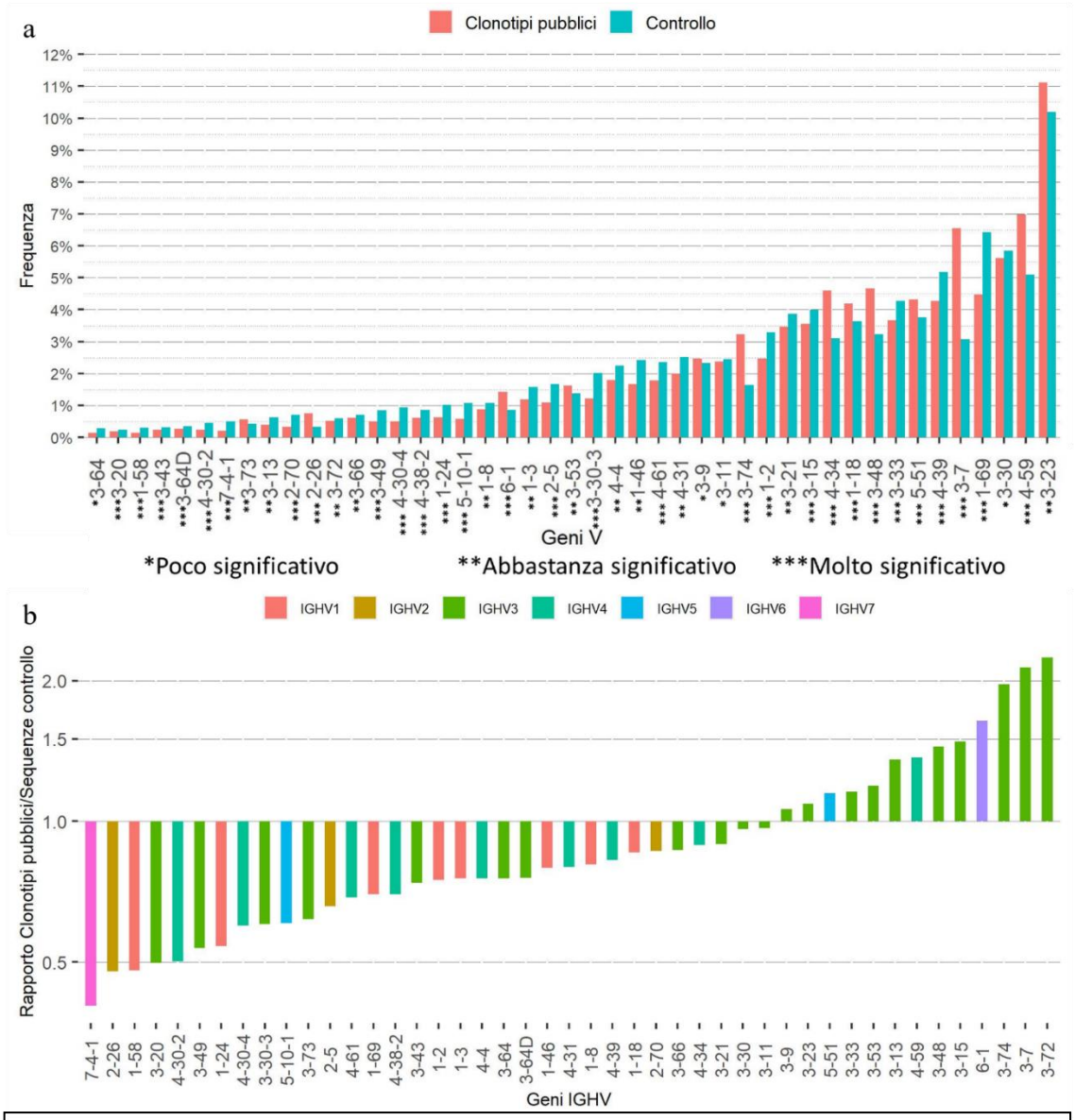


Figura 4
(a) Frequenza di espressioni IGHV nel database pubblico (rosa) e nel controllo (azzurro)
(b) Rapporto dell'espressione IGHV tra clonotipi pubblici e controllo

Analisi dei clonotipi pubblici per singoli IGHV

La Figura 5 mostra quattro esempi rappresentativi, segregati per singoli geni *IGHV* (*IGHV1-8*, *IGHV3-48*, *IGHV3-74* e *IGHV5-51*), di ciò che si è già descritto nella figura 2 (distribuzione delle lunghezze amminoacidiche degli *HCDR3* e grado di condivisione dei clonotipi pubblici). Si nota come le lunghezze degli *HCDR3* dei clonotipi pubblici non presentino la stessa distribuzione osservata per i dati aggregati (Figura 2.b): simmetria della distribuzione e mediana appaiono gene specifico, inoltre alcuni casi (geni *IGHV1-8* e *IGHV3-48*) presentano un picco assente nel controllo.

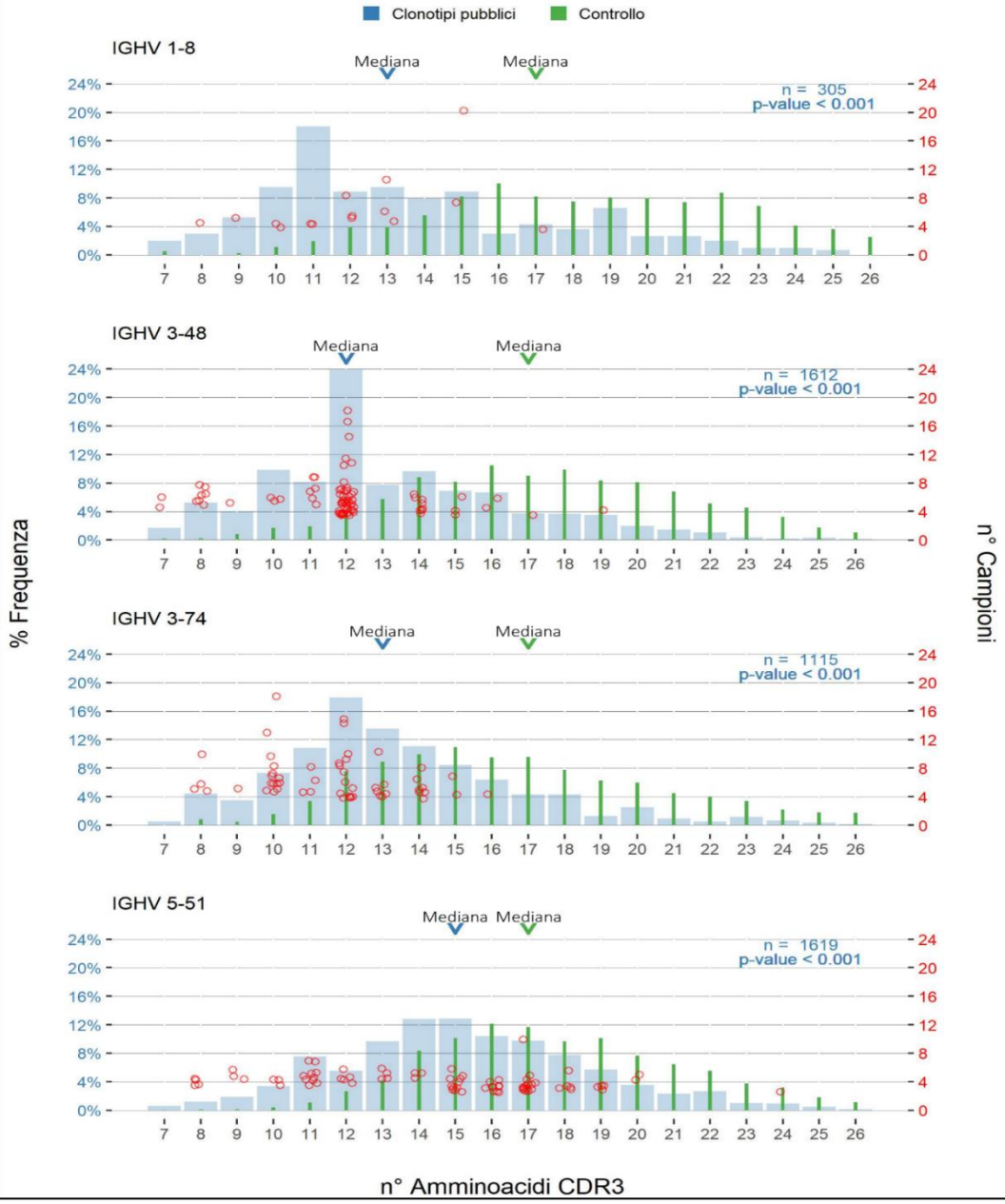


Figura 5
Distribuzione delle lunghezze amminoacidiche del CDR3 pubblici (blu) e di controllo (verde) e, sovrapposta, condivisione dei clonotipi pubblici (rosso)

<u>Composizione amminoacidica dei clonotipi pubblici per singoli geni IGHV</u>

Abbiamo generato i LOGOs della sequenza amminoacidica degli *HCDR3* delle lunghezze contenenti i clonotipi più condivisi dei geni analizzati (*IGHV1-8*, *IGHV3-48*, *IGHV3-74* e *IGHV5-51*) e confrontati con i controlli. Si può osservare che per i geni *IGHV1-8*, *IGHV3-48*, *IGHV3-74* dai LOGOs dei clonotipi pubblici emergono motivi amminoacidici predominanti, caratteristica non presente nelle sequenze controllo non pubbliche con lo stesso gene *IGHV* e lunghezza dell'*HCDR3* (Figura 6.a, 6.b e 6.c).

Diversamente per il gene *IGHV5-51* non è evidente nessun motivo ricorrente (Figura 6.d).

Focalizzando l'attenzione sulle sequenze *HCDR3* dei geni *IGHV1-8*, *IGHV3-48* e *IGHV3-74* (Figura 6.a, 6.b e 6.c) si nota la presenza di amminoacidi come Prolina in posizione 5, Arginina in posizione 6, Aspargina in posizione 7 e Triptofano in posizione 8 per il gene *1-8* Triptofano in posizione 7, Alanina in posizione 9 e Fenilalanina in posizione 9 per il gene *3-48* (Tabella 6.b) e Serina in posizione 4, Acido Aspartico in posizione 5, Triptofano in posizione 6 e Fenilalanina in posizione 7 per il gene *3-74*, non presenti nel controllo.

Isolando la sequenza più rappresentata dei clonotipi pubblici con gene *IGHV1-8*, *IGHV3-48* e *IGHV3-74* e la sequenza del clonotipo più condiviso con gene *IGHV5-51* ed osservando in quanti campioni fosse condivisa si può contare come: la sequenza d'interesse per il gene *IGHV1-8* è condivisa da undici soggetti, la sequenza d'interesse per il gene *IGHV3-48* è condivisa da diciassette soggetti, la sequenza d'interesse per il gene *IGHV3-74* è condivisa da diciotto soggetti e la sequenza isolata per il gene *IGHV5-51* è condivisa da dieci soggetti.

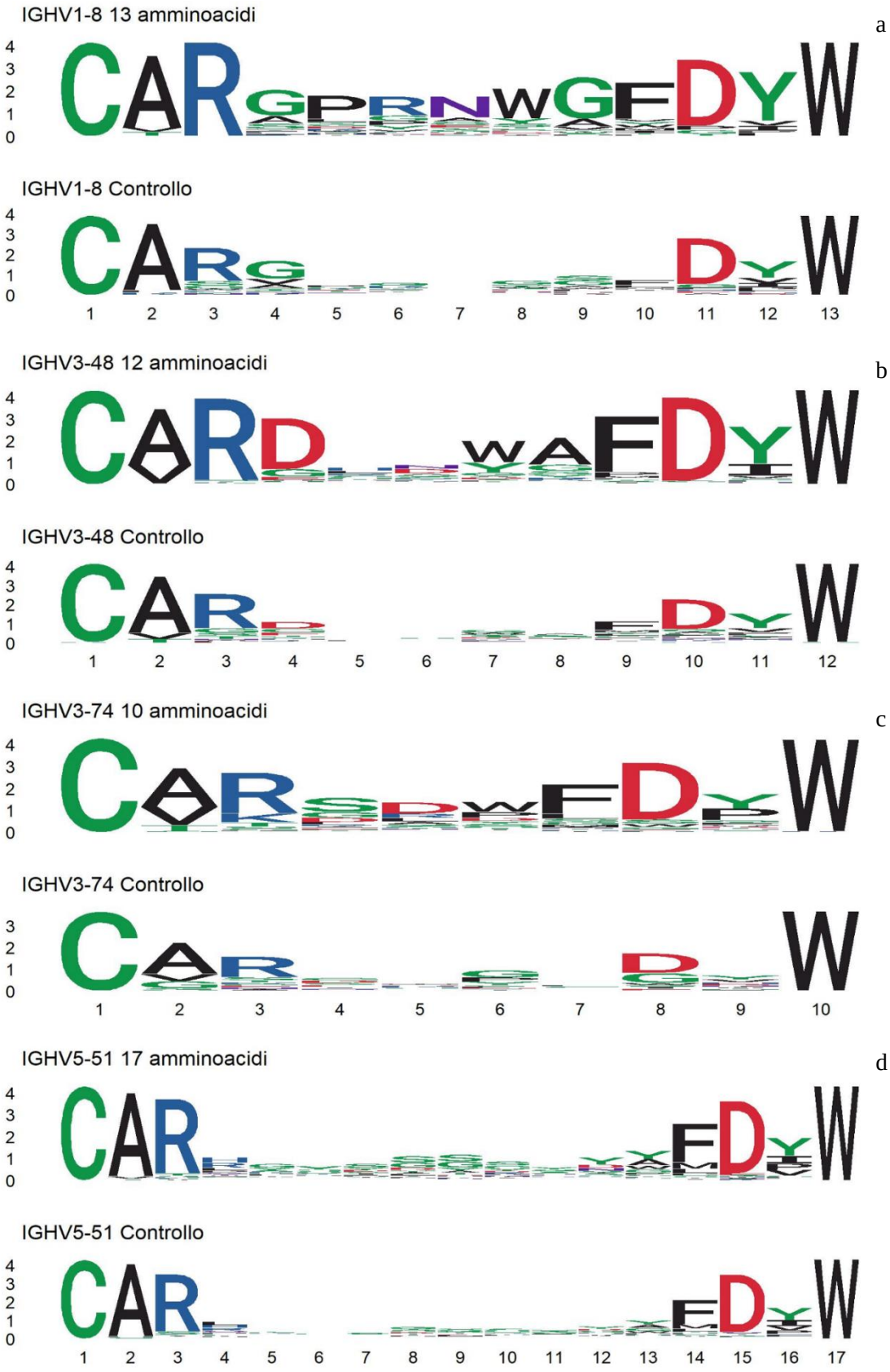


Figura 6
(a) Composizione amminoacidica dei CDR3 clonotipo IGHV1-8 e controllo
(b) Composizione amminoacidica dei CDR3 clonotipo IGHV3-48 e controllo
(c) Composizione amminoacidica dei CDR3 clonotipo IGHV3-74 e controllo
(d) Composizione amminoacidica dei CDR3 clonotipo IGHV5-51 e controllo

Analisi delle sequenze nucleotidiche e amminoacidiche delle sequenze consensus

Abbiamo analizzato le sostituzioni amminoacidiche, del gene *IGHV* rispetto alla sequenza germinale, e le differenze nucleotidiche della regione *HCDR3* tra le sequenze appartenenti allo stesso clonotipo pubblico (Figura 7). Possiamo avere innanzitutto informazione su dove si va a collocare la nostra sequenza d'interesse e, in seguito, in che regioni sono presenti le mutazioni alla sequenza amminoacidica della regione variabile e nella sequenza nucleotidica del *HCDR3*.

Per *1-8_SI* (Figura 7.a) è possibile notare come vi siano alterazioni all'altezza del *HCDR1* e del *FR3* anche se più rare, mentre la sequenza d'interesse va ad allinearsi con la regione *HCDR3* e non presenta mutazioni amminoacidiche, cosa che non si può dire della corrispettiva sequenza nucleotidica, che presenta divergenze sia a livello della regione di inserzione dei nucleotidi P ed N sia del gene J, tenendo anche conto delle due regioni all'altezza delle giunzione con la regione D che presentano nucleotidi diversi.

Per *3-48_SI* (Figura 7.b) è possibile notare come vi siano alterazioni all'altezza del *HCDR1* e del *HCDR2* osservando come quest'ultima abbia un'elevata variabilità, mentre la sequenza d'interesse va ad allinearsi con la regione *HCDR3* e non presenta mutazioni amminoacidiche, mentre la corrispettiva sequenza nucleotidica presenta divergenze sia a livello della regione di inserzione dei nucleotidi N, del gene D e del gene J, tenendo anche conto della regioni all'altezza delle giunzione tra il primo sito d'inserzione di nucleotidi N e la regione D che presentano amminoacidi tutti diversi, di cui quali non possiamo dire nulla in quanto sono quegli nucleotidi aggiunti casualmente durante il triggering.

Per *3-74_SI* (Figura 7.c) è possibile notare come vi siano alterazioni all'altezza del *HCDR1* e del *HCDR2* osservandone l'elevata variabilità, mentre la sequenza d'interesse va ad allinearsi con la regione *HCDR3* e non presenta mutazioni, mentre la corrispettiva sequenza nucleotidica che non presenta la regione D a causa degli eventi di ricombinazione, poi presenta divergenze sia a livello del gene V, della regione di inserzione dei nucleotidi N, e del gene J, al contrario delle sequenze precedenti questa oltre a non presentare le regione D non sembra avere quella regione di nucleotidi inseriti casualmente durante il triggering.

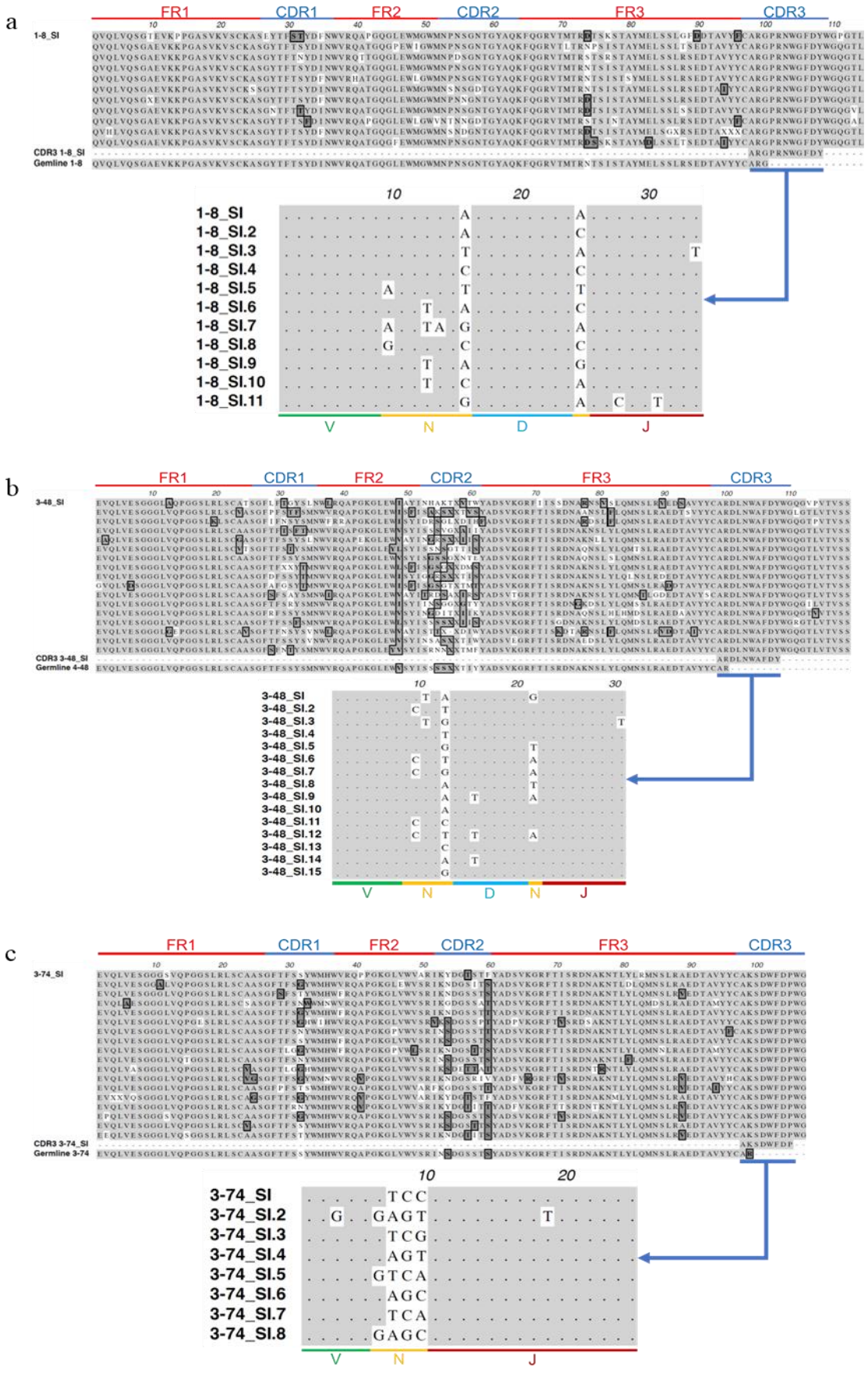


Figura 7
(e) Composizione amminoacidica dei CDR3 clonotipo IGHV1-8 e controllo
(a) Composizione amminoacidica dei CDR3 clonotipo IGHV3-48 e controllo
(b) Composizione amminoacidica dei CDR3 clonotipo IGHV3-74 e controllo

Calcolo similarità dei clonotipi pubblici con la sequenza consensus

Considerando che nei LOGOs ogni clonotipo ha peso equivalente indipendentemente dal suo grado di condivisione, la presenza di un clonotipo pubblico identico o molto simile alla sequenza consenso (Figura 8) non spiega la presenza di evidenti pattern amminoacidici. Abbiamo dunque preso la sequenza del clonotipo identica o quella a più alta similarità al logos e calcolato la similarità amminoacidica con le altre sequenze di uguale lunghezza e IGHV (Figura 9).

Figura 8
Sequenze CDR3 dei clonotipi più condivisi

Emerge che alcuni gruppi (e.s. IGHV1-8, IGHV3-48 e IGHV3-74) presentano un andamento bimodale e delle similarità, dove uno dei due cluster di clonotipi sarà altamente simile al consensus. Diversamente, per altri gruppi esemplificati dfal gene IGHV5-51, si nota un andamento unimodale dovuto all'assenza rilevante di sequenze simili alla sequenza consensus. Dei gruppi con distribuzione bimodale delle similarità, possiamo osservare quanto i due cluster siano dissimili a livello amminoacidico ed il grado di conservazione della sequenza consensus nel cluster delle sequenze simili (Figura 10).

Notando come per IGHV1-8 (Figura 10.a) il motivo del consensus sia evidente solo per il cluster delle simili, per IGHV3-48 (Figura 10.b) il motivo sia mantenuto nelle cluster delle sequenze simili, ma con l'alanina in posizione 8

presente anche nel cluster delle sequenza delle dissimili e per IGHV3-74 (Figura 10.c) la distribuzione degli amminoacidi non è differente dalle precedenti con la sequenza consensus riscontrabile nel gruppo delle simili con la fenilalanina in posizione 7 anche nelle dissimili.

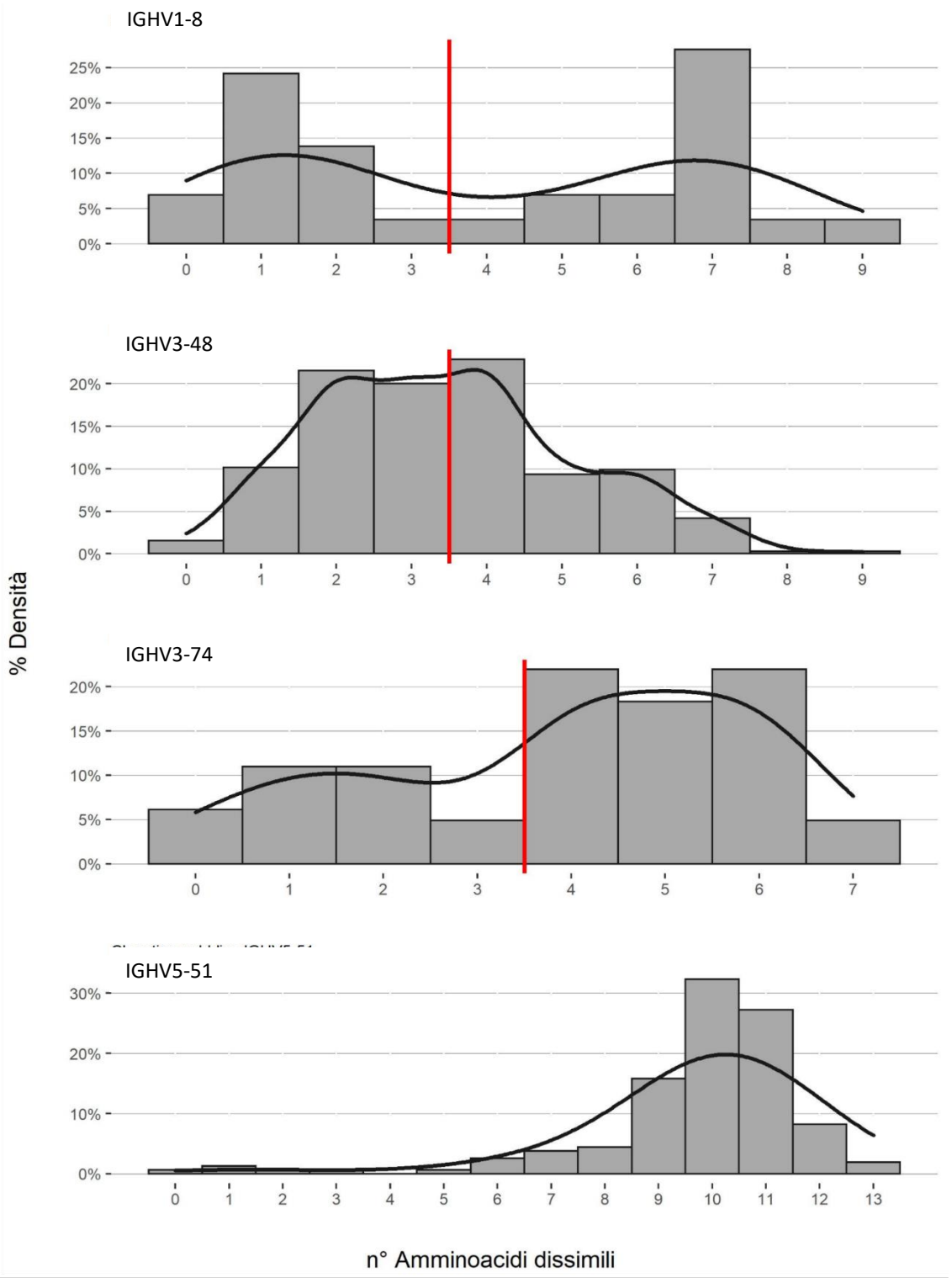


Figura 9
Similarità amminoacidica con le altre sequenze di uguale lunghezza e IGHV, cut-off per definire sequenze simili (linea rossa)

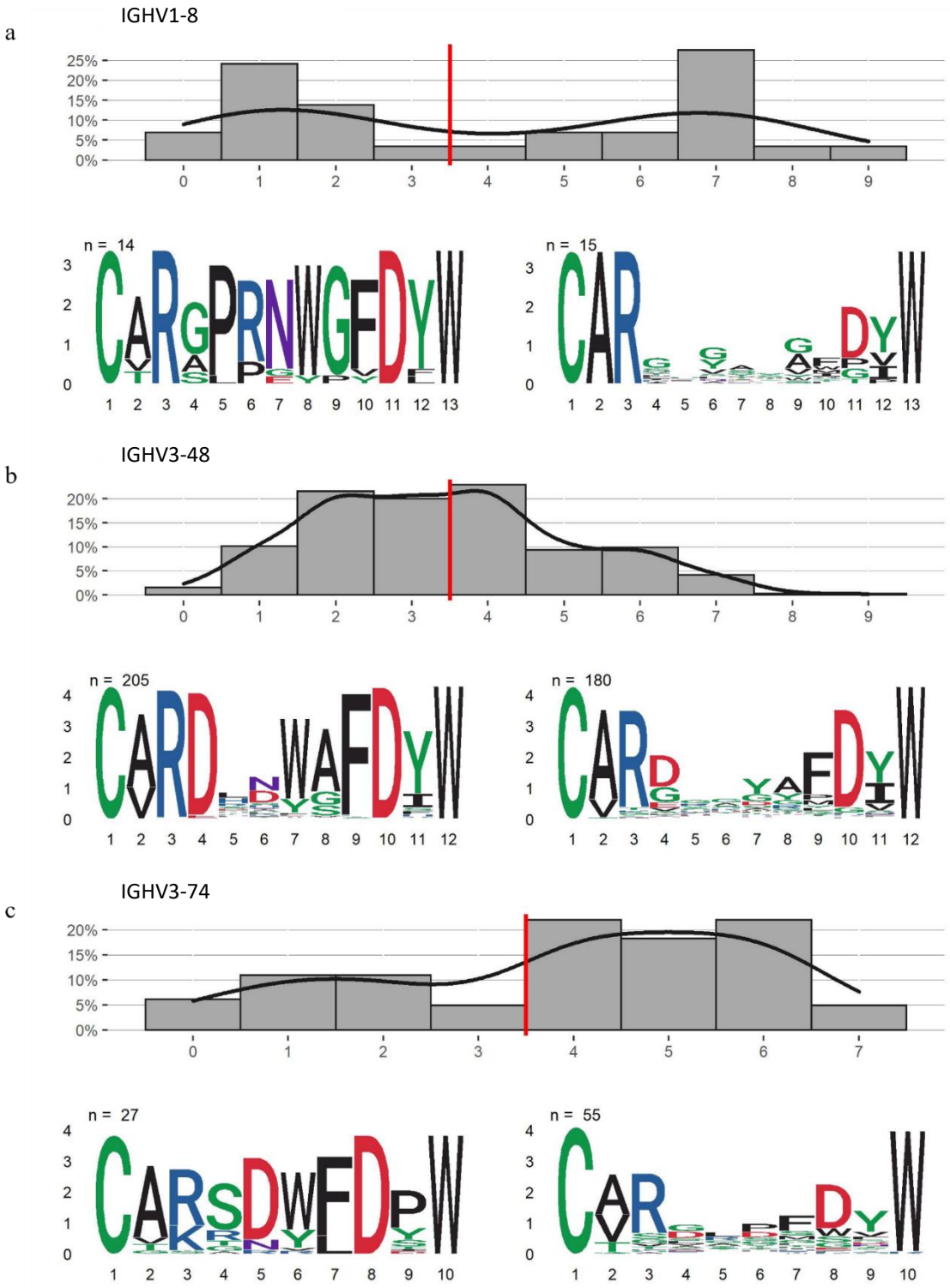


Figura 10
(a) Grado di conservazione della sequenza consensus nel cluster delle sequenze simili con IGHV1-8
(b) Grado di conservazione della sequenza consensus nel cluster delle sequenze simili con IGHV3-48
(c) Grado di conservazione della sequenza consensus nel cluster delle sequenze simili con IGHV3-74

Stato mutazionale del gene IGHV nei cluster di clonotipi pubblici

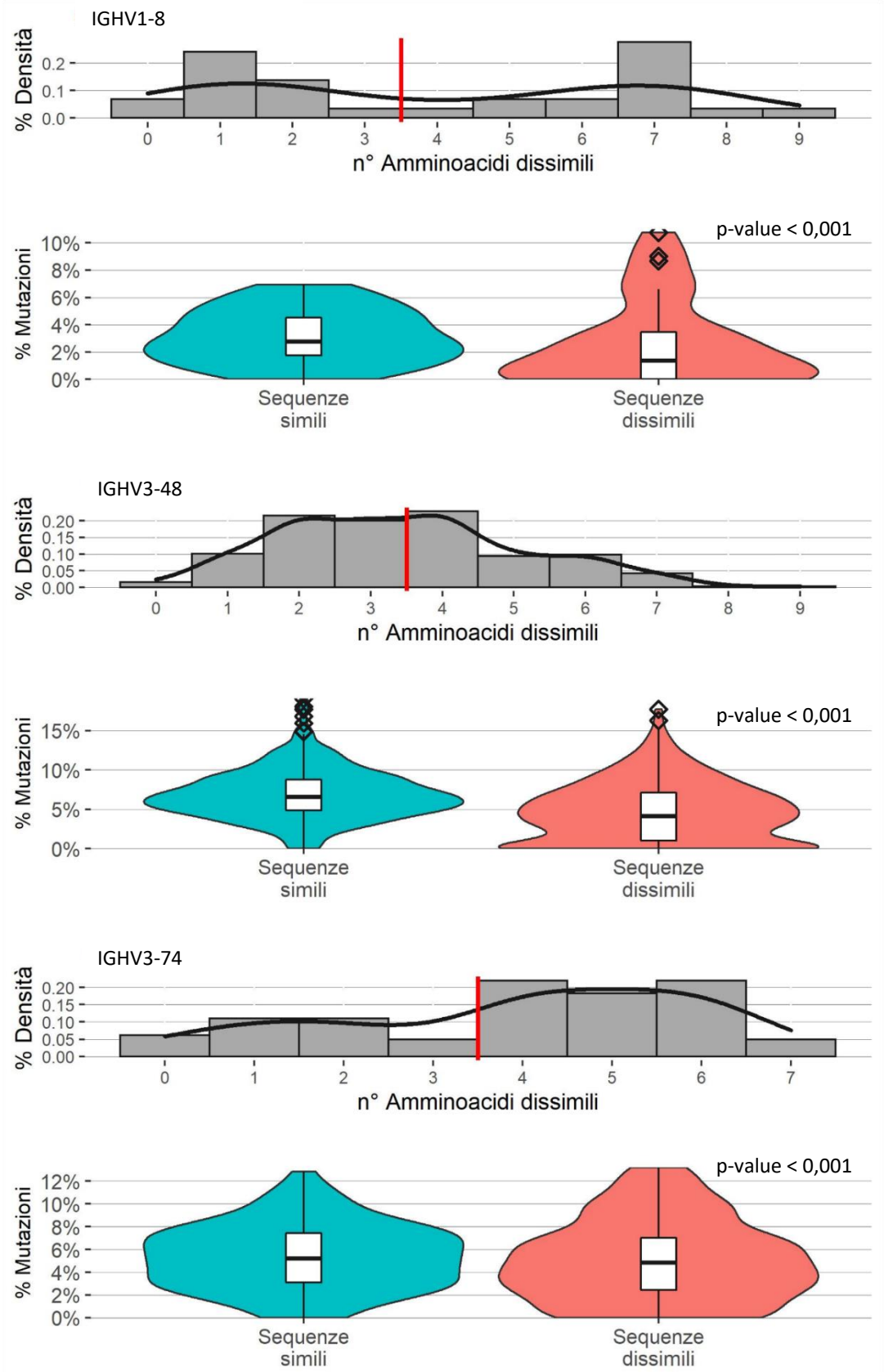


Figura 11
(a) Stato mutazionale delle sequenze simili e dissimili con IGHV1-8
(b) Stato mutazionale delle sequenze simili e dissimili con IGHV3-48
(c) Stato mutazionale delle sequenze simili e dissimili con IGHV3-74

Avendo definito due cluster di clonotipi pubblici, abbiamo analizzato le percentuali di mutazioni nei due gruppi (Figura 11).

Possiamo osservare che le sequenze simili a IGHV1-8 (Figura 11.a) siano per lo più mutate rispetto a quelle del secondo cluster delle dissimili, arricchite per sequenze non mutate.

Mentre per quanto riguarda IGHV3-48 entrambi i cluster presentano un alto livello di clonotipi mutati ma con delle sostanziali differenze: il cluster delle simili ha la base stretta viste le poche sequenze non mutate, mentre il cluster delle dissimili presenta un arricchimento dei clonotipi non mutati.

Per IGHV3-74 (Figura 11.c) notiamo per entrambi i cluster una mediana della frequenza di mutazioni sul 5% ma con due basi diverse, il cluster delle simili vede una base molto più piccola, al contrario del cluster delle sequenze dissimili.

Le differenze tra le frequenze di mutazioni dei due cluster di sequenze sono rimarcate dal valore del p-value, che essendo inferiore a 0,001 ci indica una differenza statisticamente significativa.

<u>Presenza degli HCDR3 simili ai clonotipi pubblici all'interno del repertorio IGHV</u>

Infine abbiamo esplorato se le sequenze amminoacidiche degli HCDR3 dei clonotipi dei cluster silimi al consensus fossero presenti anche associate a geni IGHV diversi da quelloi del clonotipo pubblico (Figura 12).

Questa analisi mostra che gli HCDR3 caratterizzanti i tre cluster simili al consensus, associati ai geni IGHV1-8, IGHV3-48 e IGHV3-74, sono presenti soltanto con il gene d'origine, quindi monogenici. Contrariamente, sono presenti HCDR3 di clonotipi pubblici i quali si possono osservare associati a diversi geni IGHV (Figura 12), possiamo quindi parlare di HCDR3 poligenici.

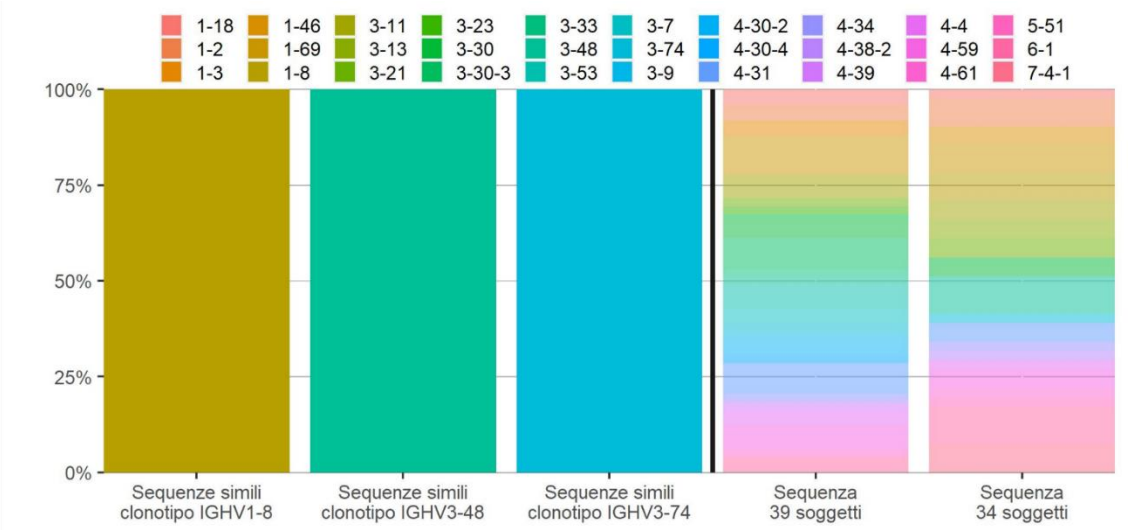


Figura 12
Repertori IGHV di CDR3 delle sequenze simili (a sinistra della verticale nera) e delle sequenze dissimili (a destra della verticale nera) nel database totale delle sequenze

IV. Conclusioni

Sequenze immunoglobuliniche con uguali amminoacidi alla regione HCDR3, con espressione dello stesso gene IGHV e ritrovate in più soggetti diversi vengono identificate come clonotipi pubblici di immunoglobuline.

Diversi studi[8,9] hanno identificato e caratterizzato principalmente quantitativamente la presenza di clonotipi pubblici condivisi tra pochi soggetti (da 3 a 10), ma fattori limitanti, quali il ristretto gruppo di individui o il basso numero di sequenze, ha permesso una caratterizzazione poco approfondita.

Con questa analisi, disponendo di un database di partenza di oltre 7 milioni di sequenze derivanti da 56 repertori immunoglobulinici di altrettanti individui, abbiamo potuto effettuare un'analisi qualitativa più approfondita.

Abbiamo potuto ricapitolare osservazioni già effettuate da lavori precedenti quali la frequenza di clonotipi pubblici all'interno di un repertorio, la predilezione di HCDR3 corti rispetto ad un repertorio da sangue periferico.

Con questa analisi, disponendo di un database di partenza di oltre 7 milioni di sequenze derivanti da 56 repertori immunoglobulinici di altrettanti individui, abbiamo potuto effettuare un'analisi qualitativa più approfondita.

Abbiamo potuto ricapitolare osservazioni già effettuate da lavori precedenti quali la frequenza di clonotipi pubblici all'interno di un repertorio, la predilezione di HCDR3 corti rispetto ad un repertorio da sangue periferico.

Nel nostro studio vi sono evidenze che suggeriscono che almeno una parte dei clonotipi pubblici derivi da un processo di selezione visto: il diverso utilizzo di geni IGHV rispetto ai controlli, gli HCDR3 corti, l'accumulo di sequenze mutate. Inoltre, sia la distribuzione delle lunghezze HCDR3 che il livello di condivisione dei clonotipi pubblici risulta essere gene IGHV specifica ed il differente grado di espressione dei geni IGHV tra clonotipi pubblici e sequenze di controllo potrebbe indicarne un qualche ruolo nella selezione.

A queste osservazioni sui clonotipi pubblici, si aggiunge in parallelo l'individuazione di quantità notevoli di linfociti B che esprimono lo stesso gene IGHV ma presentano sequenze amminoacidiche con alto livello di similarità del HCDR3 e bias per lo stato mutazionale del gene IGHV.

Abbiamo definito questi gruppi di clonotipi pubblici come stereotipi pubblici. In seguito ad un'analisi più approfondita si può notare come gli HCDR3 appartenenti agli stereotipi pubblici, siano presenti nel resto del repertorio associati esclusivamente al gene IGHV caratterizzante. Contrariamente, gli HCDR3 di altri clonotipi pubblici, si possono trovare associati a differenti geni IGHV.
La ricorrente selezione e mantenimento di anticorpi con specifiche caratteristiche, quali gli stereotipi pubblici, potrebbero essere il risultato di diversi processi:

1. Eventi stocastici, sequenze statisticamente probabili da riarrangiare
2. Mantenuti per selezione:
    a. Infezioni comuni e ricorrenti (e.s. Raffreddore)
    b. Anticorpi con funzioni a cavallo con l'immunità innata assimilabili agli anticorpi naturali

Queste analisi espandono il concetto di clonotipo pubblico e mettono le basi per ulteriori studi che potrebbero portare alla definizione ed alla identificazione di possibili meccanismi di selezione che prediligono sequenze con caratteristiche chimico-fisico-strutturali particolari e specifiche, permettendo di completare ed estendere le conoscenze sui processi maturativi e di sintesi delle immunoglobuline.

V. Bibliografia

1. Backhaus, O. Generation of Antibody Diversity. in *Antibody Engineering* (InTech, 2018). doi:10.5772/intechopen.72818.
2. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology* **125**, (2010).
3. Kim, D. & Park, D. Deep sequencing of B cell receptor repertoire. *BMB Reports* vol. 52 540–547 Preprint at https://doi.org/10.5483/BMBRep.2019.52.9.192 (2019).
4. Lin, S. G. *et al.* Highly sensitive and unbiased approach for elucidating antibody repertoires. *Proc Natl Acad Sci U S A* **113**, 7846–7851 (2016).
5. Six, A. *et al.* The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Frontiers in Immunology* **4**, (2013).
6. Hershberg, U. & Luning Prak, E. T. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 370 Preprint at https://doi.org/10.1098/rstb.2014.0239 (2015).
7. Dekosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology* **31**, 166–169 (2013).
8. Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).
9. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
10. Imkeller, K. & Wardemann, H. Assessing human B cell repertoire diversity and convergence. *Immunological Reviews* vol. 284 51–66 Preprint at https://doi.org/10.1111/imr.12670 (2018).
11. R.Ramesh, A. M. S. K. P. *Polimerase chain reaction*. (1992).
12. Zhu, H. *et al.* PCR past, present and future. *BioTechniques* vol. 69 317–325 Preprint at https://doi.org/10.2144/BTN-2020-0057 (2020).

13. Lorenz, T. C. Polymerase chain reaction: Basic protocol plus troubleshooting and optimization strategies. *Journal of Visualized Experiments* (2012) doi:10.3791/3998.
14. Garibyan, L. & Avashia, N. Polymerase chain reaction. *Journal of Investigative Dermatology* **133**, 1–4 (2013).
15. Solanki, G. *POLYMERASE CHAIN REACTION*. *International Journal of Pharmacological Research www.ssjournals.com IJPR* vol. 2 www.ssjournals.com (2012).
16. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition* **98**, 236–238 (2013).
17. Bernat, N. V. *et al.* High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Frontiers in Immunology* **10**, (2019).
18. McCombie, W. R., McPherson, J. D. & Mardis, E. R. Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine* **9**, (2019).
19. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **5**, (2010).
20. Heiden, J. A. vander, Yaari, G. & Kleinstein, S. H. *pRESTO Example Workflow: Illumina MiSeq 2x250bp B-cell receptor repertoire with UID barcoding 1 Overview of Experimental Data*. (2014).
21. Vander Heiden, J. A. *et al.* PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
22. Gupta, N. T. *et al.* Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).