Studiare la regione HLA negli esomi

E. Locatelli[1], C. Patuzzo[1], L. Moron Dalla Tor[1], L. Veschetti[1], A. Mori[1], E. Trabetti[1], D. Zipeto[1], G. Malerba[1]

[1]Sez. Biologia e Genetica, Dip. Neuroscienze, Biomedicina e Movimento, Università degli studi di Verona, Verona

elena.locatelli@univr.it

ABSTRACT

La regione HLA (6p21.3, ~4Mb) contiene un gruppo di geni altamente polimorfici coinvolti nella risposta immunitaria. L'elevata variabilità genica della regione può rendere il resequencing tramite NGS alquanto confuso, facendo credere all'allineatore che il sequenziamento contenga molti errori piuttosto che vere varianti. Inoltre ci si potrebbe chiedere se l'utilizzo del classico genoma di riferimento (GdR) sia la scelta opportuna per condurre l'allineamento e la successiva chiamata delle varianti in tale regione. Ad esempio, il GdR potrebbe essere privo di alcune regioni geniche presenti in alcuni dei molti alleli della regione HLA, rendendo impossibile l'individuazione delle varianti presenti in tali aplotipi. Con il nostro studio abbiamo cercato di capire se fosse possibile caratterizzare in maggior dettaglio il profilo genetico individuale della regione HLA a partire dai profili ottenuti dal sequenziamento dell'esoma (i.e. vcf file) tramite pipeline di analisi standard. Per semplificare le analisi abbiamo concentrato l'attenzione sul gene HLA-C, per il quale sono descritte oltre 1000 sequenze diverse (presenti nel database IPD-IMGT/HLA). Tutte le sequenze del gene HLA-C sono state allineate contro il GdR (hg19 e hg38) permettendo di individuare le varianti di ogni aplotipo e di riportare le loro coordinate rispetto al GdR. Abbiamo così costruito un database che contiene e descrive le oltre 800 varianti note del gene HLA-C contenute nei diversi aplotipi di IPD-IMGT/HLA. Lo studio sta procedendo con lo sviluppo di metodi che puntano a suggerire quali siano i diplotipi compatibili con il profilo genotipico individuale del gene HLA-C. Questo approccio potrebbe permettere di organizzare e sintetizzare l'informazione genotipica della regione HLA per condurre in maniera efficiente studi di associazione della regione con malattie comuni all'interno di studi esomici.

MATERIALI E METODI

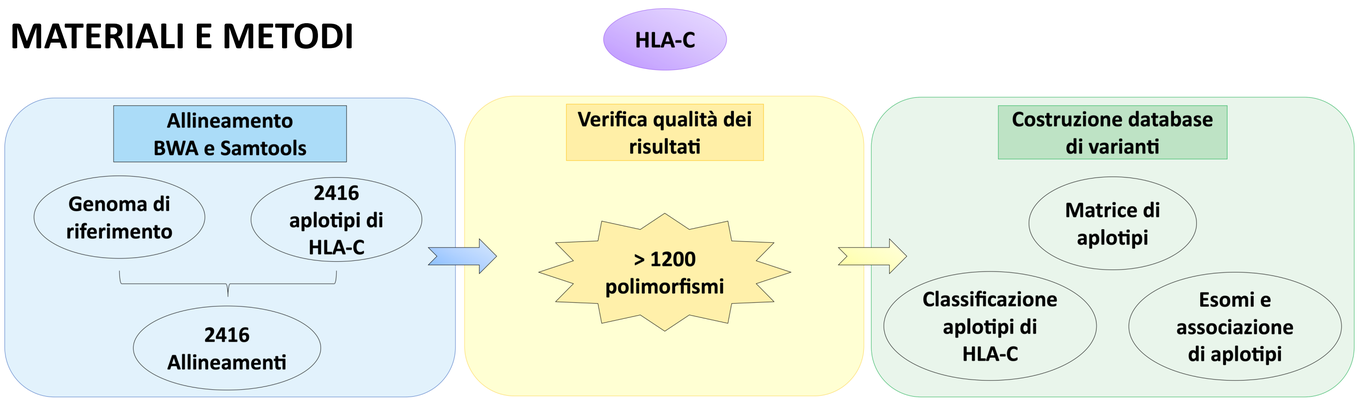

Figura a) Workflow: allineamento, controllo qualità dei risultati, costruzione database

Allineamento del genoma di riferimento e di tutti gli alleli noti di HLA-C (2416) per l'identificazione delle varianti presenti.

Controllo di qualità dei risultati ottenuti: > 1200 varianti.

Costruzione di un database contenente le varianti per ogni aplotipo di HLA-C.

Analisi di classificazione di ogni sierogruppo di HLA-C.

Comparazione dei genotipi di file VCF contro il database di riferimento, utilizzando un metodo «sliding window».

RISULTATI

HLA-C	Nodo di decisione	dbSNP	Variante
C*01	hg19: 31239727	rs29029490	V. Intronica
C*02	hg19: 31238334	rs41540318	V. Intronica
C*03	hg19: 31237002	rs17885436	V. Intronica
C*04	hg19: 31239742	rs41553018	V. Intronica
C*05	hg19: 31238708 hg19: 31238984 hg19: 31239407	rs9264650 rs2308584 rs17408553	V. Intronica V. Missenso V. Missenso
C*06	hg19: 31239506 hg19: 31237605	rs1050414 rs9264607	V. Sinonima V. Intronica
C*07	hg19: 31239271-31239275	rs66565287	V. Intronica: delATCCA
C*08	hg19: 31239346 hg19: 31238507 hg19: 31238322 hg19: 31239407	rs11547346 rs41544520 rs9264639 rs17408553	V. Intronica V. Intronica V. Intronica V. Missenso
C*12	hg19: 31238995 hg19: 31238992 hg19: 31239100	rs41553316 rs41550715 rs1554181495	V. Sinonima V. Sinonima Stop Gained
C*14	hg19: 31237457	rs72558161	V. Intronica
C*15	hg19: 31237237	rs17881458	V. Intronica
C*16	hg19: 31239246	rs28367581	V. Intronica
C*17	hg19: 31237016	rs41544314	V. Intronica
C*18	hg19: 31236168 hg19: 31239506 hg19: 31237987	rs111590662 rs1050414 rs41556321	500B Downstream Variant V. Sinonima Stop Gained

Tabella a) Classificazione degli aplotipi di HLA-C. Ogni sierogruppo riporta la posizione dello SNP con cui ogni allele viene classificato.

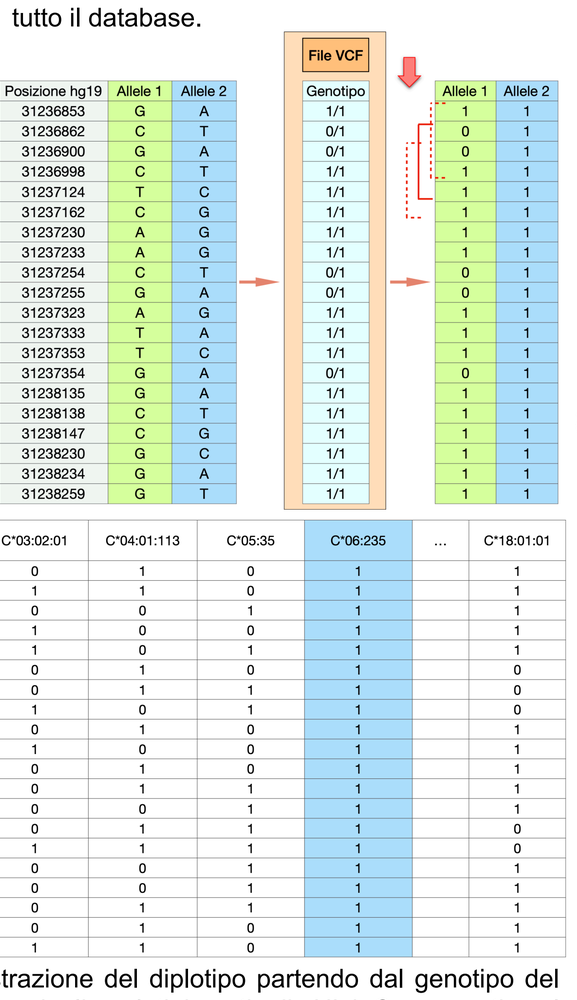

Figura b) Metodo di classificazione per i sierogruppi di HLA-C. La figura mostra la classificazione di C*01 con un errore del 1% su tutto il database.

Classificazione dei sierogruppi di HLA-C

La tabella «a» mostra la posizione degli SNP secondo cui ogni sierogruppo viene classificato.

La figura «b» mostra un esempio di classificazione con bassa probabilità di errore per C*01 su tutto il database.

Identificazione dei diplotipi dal database

Metodo di estrazione del diplotipo maggiormente compatibile tra tutti gli aplotipi di HLA-C presenti nel database con il genotipo del file VCF (tabelle b e c) .

Il risultato viene visualizzato grazie ad una heatmap, che evidenzia la compatibilità tra il genotipo del VCF e i diplotipi di HLA-C (figura c).

Tabella b)

Posizione hg19	C*01:02:01:01	C*02:02:02:15	C*03:02:01	C*04:01:113	C*05:35	C*06:235	…	C*18:01:01
31236853	1	1	0	1	0	1		1
31236862	1	0	1	1	0	1		1
31236900	1	0	0	0	1	1		1
31236998	1	1	1	0	0	1		1
31237124	0	1	1	0	1	1		1
31237162	1	1	0	1	0	1		0
31237230	1	1	0	1	1	1		0
31237233	1	1	1	0	1	1		0
31237254	0	0	0	1	0	1		1
31237255	0	0	1	0	0	1		1
31237323	1	1	0	1	0	1		1
31237333	1	1	0	1	1	1		1
31237353	0	1	0	0	1	1		1
31237354	0	0	0	1	1	1		0
31238135	0	1	1	1	1	1		0
31238138	0	1	0	0	1	1		1
31238147	0	1	0	0	1	1		1
31238230	0	1	0	1	1	1		1
31238234	0	1	0	1	0	1		1
31238259	0	1	1	1	0	1		1

Tabella c) Metodo di estrazione del diplotipo partendo dal genotipo del file VCF. Identificazione degli aplotipi noti di HLA-C presenti nel database. In questo esempio l'aplotipo di HLA-C evidenziato in verde è compatibile con l'allele 1 del VCF; l'aplotipo azzurro con l'allele 2.

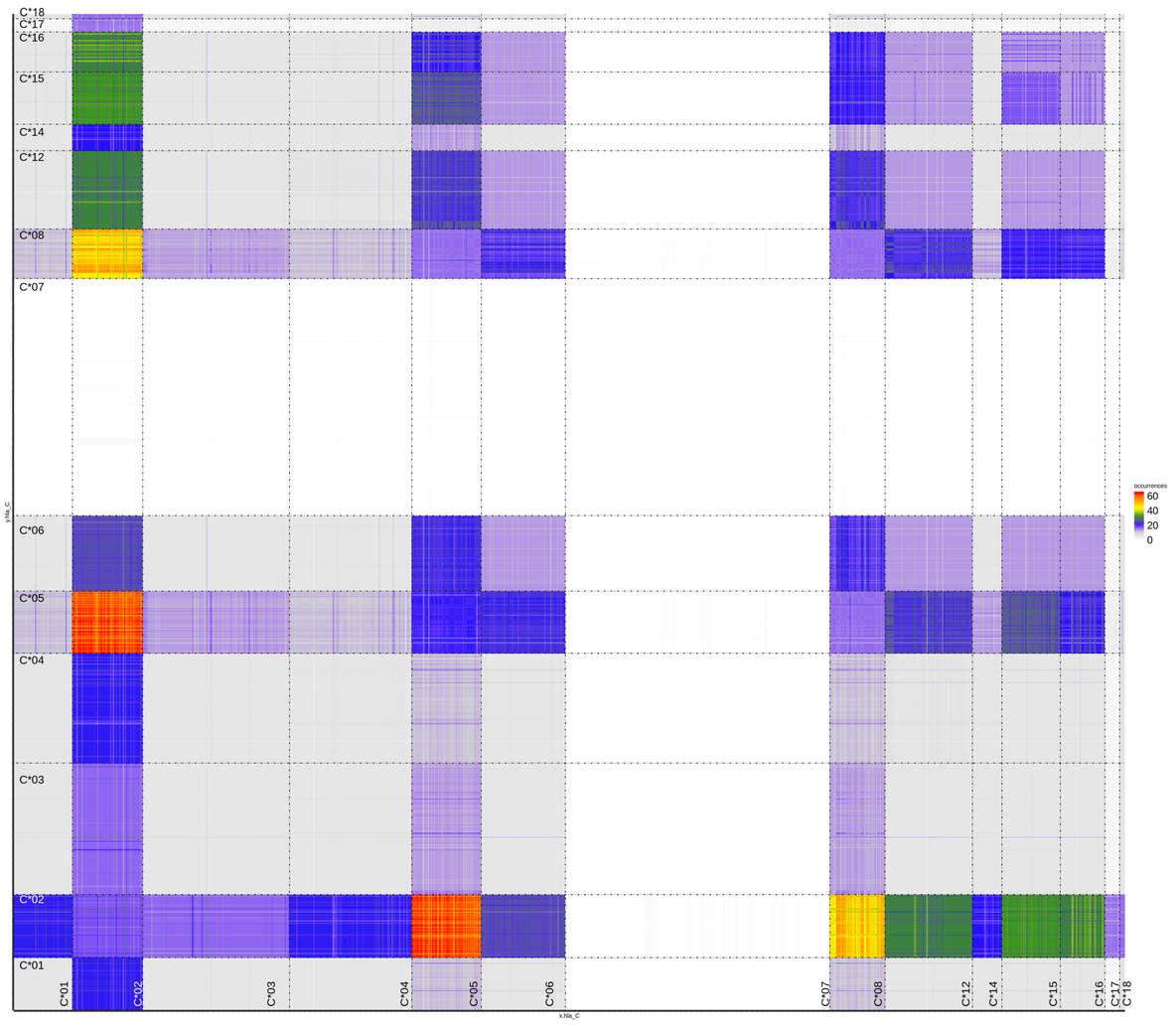

Figura c) Heatmap per la visualizzazione dei diplotipi più probabili (il numero maggiore di hit) per un dato genotipo di un file VCF. In questo esempio i due riquadri rossi mostrano come soluzione migliore il diplotipo C*02-C*05. Sull'asse x e y troviamo tutti i 2416 aplotipi di HLA-C.

CONCLUSIONI

Abbiamo messo a punto un sistema rapido e veloce che punta a stimare gli alleli del gene HLA-C partendo dai file VCF di studi standard dell'esoma e da un database da noi allestito contenente tutti gli alleli noti di HLA-C. Il metodo appare robusto a livello dei gruppi di HLA-C (C*01, C*02, …), inoltre notiamo che non tutti gli aplotipi giustificano i dati contenuti nei VCF, suggerendo la presenza di alleli non ancora caratterizzati. Queste informazioni saranno utilizzate per futuri studi di caso-controllo. L'analisi sarà estesa agli altri geni di HLA (partendo da -A e -B).

References

[1] IPD-IMGT/HLA: https://www.ebi.ac.uk/ipd/imgt/hla/

[2] Illumina HLA sequencing: https://www.illumina.com/clinical/hla-sequencing.html

[3] Seán Turner, et al. [Sequence-Based Typing Provides a New Look at HLA-C Diversity], The Journal of Immunology, 1998, 161: 1406–1413.

[4] Erlich et al. [Next-generation sequencing for HLA typing of class I loci], BMC Genomics 2011 12:42.

[5] Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, et al. (2017) [Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles], PLoS Genet 13(6): e1006862.