The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences

Wolfgang Helmberg*, Raymond Dunivin and Michael Feolo

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 45, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received February 20, 2004; Accepted April 8, 2004

ABSTRACT

The dbMHC resource (http://www.ncbi.nlm.nih.gov/ mhc/sbt.cgi?cmd=main) at the National Center for Biotechnology Information (NCBI) has developed an online tool for evaluating the allelic composition of sequencing-based typing (SBT) results of cDNA or genomic sequences. Whether the samples are heterozygous, haploid or a combination of the two, they can be compared with two up-to-date databases of all known alleles of several human leukocyte antigen (HLA) and killer cell immunoglobulin-like receptor (KIR) loci. The results of the submission are returned as a table of potential allele hits, along with the respective base changes and an interactive sequence viewer for close examination of the alignment.

INTRODUCTION

Direct sequencing of genomic DNA or cDNA is being used increasingly in routine diagnostic methods. For clinical procedures such as bone marrow transplants, detailed genotype information on both the recipient and the donor is required. Typically, five human leukocyte antigen (HLA) gene-coding regions of the major histocompatibility (MHC) complex are used for genotyping: HLA class I, HLA-A, HLA-B and HLA-C; and HLA class II, HLA-DRB1 and HLA-DQB1. In highly polymorphic genes such as HLAs, direct sequencing-based typing (SBT) is more efficient than traditional genetic testing methods, which are based on probe hybridization or amplification with sequence-specific primers. Sequencing strategies used in SBT differ between laboratories and can generate either heterozygous sequences, haploid sequences (after allele separation of the sample) or a combination of heterozygous and haploid sequences for each typed sample.

The dbMHC resource (http://www.ncbi.nlm.nih.gov/mhc/ sbt.cgi?cmd=main) at the National Center for Biotechnology Information (NCBI) has developed an online tool for evaluating the allelic composition of SBT samples. Whether the samples are heterozygous, haploid or a combination of the two, they can be compared with two up-to-date databases of all known alleles of several HLA and killer cell immunoglobulinlike receptor (KIR) loci. HLA-related loci are based on the IMGT/HLA database (1), which contains sequences curated by the WHO allele nomenclature committee for HLA alleles. The allele database for KIR sequences is based on entries in GenBank and the IPD database (http://www.ebi.ac.uk/ipd/kir/ index.html). The results of the submission are returned as a table of potential allele hits, along with the respective base changes and an interactive sequence viewer for close examination of the alignment.

To send feedback and suggestions for improvement, please Email the dbMHC team at: dbMHC@ncbi.nlm.nih.gov.

INPUT

Any number of sequences can be submitted; the number is indicated in the box provided (Figure 1). Sequences can be entered as simple text strings of bases or in FASTA format. All IUPAC nucleotide codes are valid; therefore, polymorphic positions of heterozygous sequences can also be represented.

By default, sequences are analyzed as heterozygote allele combinations. If a sequence is haploid, this should be indicated by using the haploid check box, thus avoiding the assumption that the submission is heterozygous. Sometimes an allele will be segregated, i.e. covering a long region that has been sequenced in two or more fragments, e.g. an amplification that spans exon 2, intron 2 and exon 3, but only exons 2 and 3 have been sequenced. In this case, it is essential to know about the haplotypic link between exons 2 and 3. This can be specified using the linked property, for either strand one or strand two (Figure 1). Once the submisson information has been added, the final step is to select the locus that has been sequenced and enter the maximum number of nucleotide mismatches allowed between the submitted sequence and the alleles in the database. A gapped interpretation can also be selected, should the submitted sequence(s) contain a longer deletion than has thus far been observed at this locus.

*To whom correspondence should be addressed. Tel: +1 301 402 2781; Fax: +1 301 480 2484; Email: helmberg@ncbi.nlm.nih.gov

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

In Log Out Alignments Primers/Probes Tryping kits SET Graphic Views diskHCHornes Sequencing Based Typing Number of sequences 3 Change Clear Help Ouery Haploid Sequence (copy&paste or select a filename) Linked Not linked Ouery 1 P PA*01012, Exon2 # Browse C C @ Clear Ouery 3 P PExon3, displaid # Browse C C @ Clear Mismatches: 1 Locus HLAA ■ Haploid C Gaped C Interpret Alignment Download Interpretation Query 1 Query 3 db Version 2 04 Mismatches Mismatches A*010101 - A*0102 - 4 - 4 A*010102 - A*0102 - 4 - 4 A*010102 - A*0103 - 3633 A=>AG A*010102 - A*0103 - 5599 C<>C A*0102 - A*0103 - 5599 C<>C A*0102 - A*0108 - 4 Cuery 1 Sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCCCCCCCCCCCCATGAGGTATTTCTTCACATCCGTOTCCCGGCCCGG	S NCBI	dbMHC Sea	uence Inter	pretation	Longod in an: Ci	unat			
Sequencing Based Typing Number of sequences: 3 Change Clear Help Ouery Haploid Sequence (copy&paste or select a filename) Linked Not linked Ouery 1 PA*010101, Exon2 Browse C C C Clear Ouery 2 PA*01010, Exon2 Browse C C C Clear Ouery 3 PExon3, diploid Browse C C Clear Mismatches: Locus: HLA-A Haploid Gapped Interpret Alignment Download Interpretation Query 1 Query 2 Query 3 Goversin 2.04 Mismatches Mismatches Mismatches Alignment Download Interpretation Query 1 Query 2 Query 3 Goversin 2.04 Mismatches Mismatches Alignment Download Interpretation Query 1 Sense [backto top] - - - - Ar010102 Ar010101, Exon2 - - - - - Query 1 sense [backto top] - - - - - FASTA header >A'010101, Exon2 Co	og In Log Out	Alignments Primers	Probes Typing	Kits SBT	Graphic View	w dbMHC Ho	me		
Query Haploid Sequence (copy&paste or select a filename) Linked Not linked Query 1 Ø A*010101, Exon2 Browse C @ @ Clear Query 2 Ø A*0102, Exon3, diploid Browse C @ @ Clear Query 3 F >Exon3, diploid Browse C @ @ Clear Mismatches: I Locus HLAA Haploid G Gapped G Interpret Alignment Download Interpretation Query 1 Query 2 Query 3 db/Version 2.04 Mismatches Mismatches A'01001. A'0102 - - - - - - - A'0102. A'0103 - 363: A<>A.G - - - - - A'0102. A'0108 -		Sequencing	Based Typi	ng Number	of sequences:	3 Change	Clear	Help	
Query Haploid Sequence (copy&paste or select a filename) Linked Not linked Query 1 Pat010101, Exon2 Browse C C C Clear Query 2 Pat01021, Exon2 Browse C C C Clear Query 3 PExon3, diploid Browse C C C Clear Mismatches: I Locus HLAA Haploid Gapped Interpret Alignment Download Interpretation Query 1 Query 2 Query 3 Aviolo2.Aviolo2 - - - Aviolo2.Aviolo2 -									
Query 1 IF >A*010101, Exon2 Image: Constraint of the second of th		Query Haploid S	Sequence (copy	&paste or sele	ect a filename)		Linked N	Not linked	1
Query 2 ▼ >A*0102, Exon2 Image: Construction of the second		Query 1 💌 👂	A*010101, Exor	n2 🕂		Browse	cc	e	Clear
Query 3 □ ▷Exon3, diploid □ Browse C C Clear Mismatches: □ Locus: HLAA ■ Haploid □ Gapped □ Interpret Alignment Download Interpretation Query 1 Query 2 Query 3 □ Alignment Download Interpretation Query 1 Query 2 Query 3 □ Anolog Anolog Anolog □ <		Query 2 💌 🖻	A*0102, Exon2	÷		Browse	cc	c	Clear
Mismatches: 1 Locus: HLA-A ■ Haploid □ Gapped □ Interpret Alignment Download Interpretation Query 1 Query 2 Query 3 db/version 2.04 Mismatches Mismatches Mismatches A*010101.**0102 - A*0102.**0104N - A*0102.**0108 - A*0102.**0108 - A*0102.**0108 - A*0102.**0108 - A*0102.**0108 - A*0102.**0108 - A*0102.**0109 171: C<>A - Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GercecArtCatTacGerGerCeGeGeCeGeGeGeGeGeGeGeGeGeGeGeGeGe		Query 3 🗖 🖡	Exon3, diploid	a 🔅		Browse	cc	e	Clear
Interpretation Query 1 Query 2 Query 3 dbVersion 2.04 Mismatches Mismatches Mismatches A*0102.A*0102 - - - A*0102.A*0102 - 498: C<>T,C A*0102.A*0103 - 363.A<>A,G A*0102.A*0109 - - A*0102.A*0109. - - Guery 1 sense [back to top] - FASTA header >A*010101. Exon2 - Exon2: 1.270 GettectAttectCattectAttectGetectCattecttectCattecttecttecttecttecttecttecttecttectte		Mismatches: 1	Locus: HLA-A	▼ Ha	aploid 🗖 Gapp	ped 🗆 Interp	oret Al	ignment	Download
dbVersion 2.04 Mismatches Mismatches Mismatches A*010101 - A*0102 - - - A*01012 - A*0102 - - 498: C<>T,C A*0102 - A*0103 - - 498: C<>T,C A*0102 - A*0103 - - 363: A<>A,G A*0102 - A*0108 - - 559: C<>C,T A*0102 - A*0109 171: C<>A - Query 1 sense [back to top] FASTA header FASTA header >A*010101, Exon2 - Exon2: 1 - 270 GCTCCACTCATGAGGTATTTCTTCACATCCGTGTCCCGGCCGG		Interpretation	Querv 1	Query 2	Query 3				
A*010101 - A*0102		dbVersion 2.04	Mismatches	Mismatches	Mismatches				
A*0102 - A*0104N - - - A*0102 - A*0102 - - 498: C<>T,C A*0102 - A*0103 - - 363: A<>A,G A*0102 - A*0108 - - 559: C<>C,T A*0102 - A*0109 171: C<>A - - Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGTGTCCCGGCCGG		A*010101 - A*010	02 -	-	-				
A*010102 - A*0102 - - 498: C<>T,C A*0102 - A*0103 - - 363: A<>A,G A*0102 - A*0108 - - 559: C<>C,T A*0102 - A*0109 171: C<>A - - A*0102 - A*0109 171: C<>A - - Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGTGTCCCGGCCGG		A*0102 - A*0104	- V	-	-				
A*0102 - A*0103 - - 363: A<>A,G A*0102 - A*0108 - - 559: C<>C,T A*0102 - A*0109 171: C<>A - - Query 1 sense [back to top] - - FASTA header >A*010101, Exon2 - - Exon2 : 1 - 270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGTGTCCCGGGCCGGGCCGGGCGGG		A*010102 - A*010	02 -	•	498: C<>T_C				
A*0102-A*0108 559: C<>C,T A*0102 - A*0109 171: C<>A Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCTCCACTCCATGAGGTATTTCTTCACATCCGTGGCCGGGCCGGCGGCGGGGGGGG		A*0102 - A*0103	-	-	363: A<>A,G				
A*0102 - A*0109 171: C<>A - Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGTGTCCCGGCCGG		A*0102 - A*0108	-	•	559: C<>C,T				
Query 1 sense [back to top] FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCC >>> Exon2 Multiple Intr/Exon • 313 • 392 CC >>> Exon2] Exon2 Exon2] Exon2 Exon2] Exon2 Stop Guery 1 Sense [back to top] File File HLA-A • Intr/Exon • 313 • 392 CC >>> Exon2] Exon2] Exon2] Exon2] Exon2] Exon2] Exon2] Exon3 340 343 A/Intr. 320 S20 330 S40 343 350 S40 343 550 S40 343 55		A*0102 - A*0109	171: C<>A	-	-				
FASTA header >A*010101, Exon2 Exon2: 1 - 270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGGGCCGGCC		Query 1	sense [backt	o top]					
Exon2: 1-270 GCTCCCACTCCATGAGGTATTTCTTCACATCCGTGTCCCGGGCCGGCC		FASTA header	>A*010101, Ex	con2			_		
Download/3D LocusLink Show Region		Exon2: 1-270	GCTCCCACTCCAT GGGGAGCCCCGCT	GAGGTATTTCTT TCATCGCCGTGG	CACATCCGTGTCC GCTACGTGGACGA	CGGCCCGGCCGC CACGCAGTTCGT			
eles HLA-A Intr/Exon 313 392 CDNA Codon FASTA Reference: Reference Image: Code of the code o	Download/3D	LocusLink Show	Region		otein 🔽 Misma	tches IT Code		Help	
CC> >> Exon2 Exon2] [Exon3 Exon3 A/Intr. 320 330 340 343 350 360 370 380 390 400 10101 CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACG GTTCTCA CACCATCCAG AT AATGTATG GCTGCGACGAC GGGGGGCCTCCC cy 1:	Alleles HLA-A	▼ Intr/Exon ▼	313 _ 392		odon FASTA	A Reference:	Referen	ce 💌	
A/Intr. 320 330 340 343 350 360 370 380 390 400 10101 CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACG GTTCTCA CACCATCCAG AT ATGTATG GCTGCGACGT GGGGCCGGAC GGGCGCTTCC ry 1:	<pre><<c> >> Exon2</c></pre>		Exon2	[Exon3					Exon3
10101 CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACG GTTCTCA CACCATCCAG ATLATGTATG GCTGCGACGT GGGGCCGGAC GGGCGCTTCC cy 1:	NA/Intr.	320 330	340 34	3 350	360	370	380	390	9 400
cy 2:	010101 CCCTGO	CGCGG CTACTACAAC	CAGAGCGAGG AC	G GTTCTCA CA	CCATCCAG ATAAT	TGTATG GCTGC	GACGT GO	GGCCGGA	C GGGCGCTTCC
ry 3: 102 103 	ery 1:			- ******* **	******* ***	****** *****	***** **	******	* ********
102 103 103 	ery 3: *****	***** *********	*******	*					
103	0102								
103	0103				G				
	0103				G				
and, "_", identical to reference " ", deletion "t", not segmenced	gend. "-". ide	antical to reference	" ". deletie	. ###+ not a	aguangad				

Figure 1. The interface of the SBT resource. The page is divided into two frames. The top frame contains the submission form and displays the interpretation results; the bottom frame contains the interactive alignment viewer. In this example are three query sequences (Query 1–3), all testing the HLA-A locus from one sample. 'Query 1' and 'Query 2' are haploid, whereas 'Query 3' is heterozygous. One nucleotide mismatch is allowed between the query and matched sequences. Two allele combinations match perfectly; four allele combinations show one mismatch each. The blockwise display of the query sequences starts with 'Query 1'. Nucleotides 1–270 of 'Query 1' are part of exon 2. The bottom frame shows the three query sequences aligned to the reference sequence A^* 010101. The mismatch for the allele combination A^* 0102–A0103 at position 363 (A<>A,G) is highlighted. The nucleotide 'A' in the heterozygous 'Query 3' does not match with the combination A,G of A^* 0102–A0103.

Algorithm

Queries to the SBT Tool are processed in several steps. The first involves a BLAST-based (2,3) search on a custom-made BLAST database that contains the allele sequences of each locus. Each sequence is split into blocks of untranslated regions (UTRs), exons and introns. In the second step, the block ends are fine-tuned to ensure a complete alignment. This is important when the query sequences have as-yetundescribed polymorphisms occurring around block boundaries. An alignment consisting of the best matches between blocks from the allele database and blocks of query sequences is created against a reference sequence. In the next step, all alleles that have nucleotide mismatches exceeding the userdefined cutoff are discarded. Insertions/deletions are counted as one mismatch. In submissions that contain several sequences, subsequent queries are run against only those sequences that were not discarded because of mismatches. If the default heterozygous analysis is requested, all of the remaining allele matches are combined at the end of the process to display a heterozygous allele.

OUTPUT

Allele combinations that match the submitted sequences within the user-defined cutoff are listed in a table. Nucleotide mismatches, insertions and deletions between the submitted sequences and any derived allele combination are listed according to their positions within the alignment and the nature of the mismatch. Each sequence submitted is split up according to its aligned UTR/exon/intron components. The blocks are listed according to their boundary positions within each sequence.

Heterozygous sequences that contain a deletion in one allele, which generates a base shift with almost continuous polymorphic positions after the deletion, can be correctly diagnosed, although the user should check to ensure that the electropherogram has been read correctly.

Interactive alignment viewer

Specifying the locus that query sequence(s) will be BLASTed against is part of the submisison process (Figure 1). As soon as the user does this, the interactive alignment viewer is loaded into the bottom frame of the Web page. It displays the reference sequence with the aligned query sequences. To this can be added, as the user wishes, the matched alleles listed in the table of allele combinations and/or as many other alleles from the specified locus as desired (Figure 1). The alignment can be navigated by scrolling forward and backward, by selecting a distinct exon/intron/UTR from the block display or by specifying a position within the alignment. The nucleotide mismatches listed in the results table link directly to the position of the mismatch in the alignment viewer.

The various function buttons on the alignment viewer can be used to manipulate the alignment in the followng ways: translate the nucleotide sequence into protein, change the reference sequence, highlight single nucleotide polymorphism (SNP) positions, display sequences in FASTA format and download complete or partial alignments of sequences displayed.

DOWNLOAD OPTIONS AND FURTHER ANALYSIS

The results of the SBT tool—the tabulated allele matches, the sequence blocks used and/or the alignment—can be downloaded as a text file.

If there is a protein structure available for the locus selected, it can be viewed using the NCBI Cn3D viewer (4,5). All haploid sequence submissions are translated into proteins. When these sequences represent new alleles with mutations that lead to non-synonymous amino acid changes, those positions can be highlighted on the protein structure in the Cn3D viewer, and the file can be downloaded. This feature also allows the display of any amino acid differences between any subset of alleles.

ALLELE DATABASES

Currently, the databases are restricted to classical HLA and KIR genes, but we plan to expand the list of loci in the future. Allele databases populated from non-curated external sources (such as the KIR database) are updated periodically by an automated GenBank search. Sequences that match any existing allele with a maximum of 30 nt differences are included; some sequences may be entered by hand. Such sequences link back to the source GenBank record via the Accession number. (HLA loci, which already undergo high-quality manual curation, are not updated using GenBank sequences.)

ACKNOWLEDGEMENTS

The authors would like to thank Jo McEntyre and Belinda Beck for their altruistic support preparing this article.

REFERENCES

- Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P. and Marsh, S.G. (2003) IMGT/HLA and IMGT/ MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.*, 31, 311–314.
- 2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.
- 4. Hogue, C.W. (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.*, **22**, 314–316.
- Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A. and Bryant, S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, 25, 300–302.