Immunogenetics (2009) 61:493–502
DOI 10.1007/s00251-009-0383-x

ORIGINAL PAPER

A clonotype nomenclature for T cell receptors

**Maryam B. Yassai · Yuri N. Naumov ·
Elena N. Naumova · Jack Gorski**

Received: 20 May 2009 / Accepted: 15 June 2009 / Published online: 1 July 2009

Abstract T cell receptor (TCR) nucleotide sequences are often generated during analyses of T cell responses to pathogens or autoantigens. The most important region of the TCR is the third complementarity-determining region (CDR3) whose nucleotide sequence is unique to each T cell clone. The CDR3 interacts with the peptide and thus is important for recognizing pathogen or autoantigen epitopes. While conventions exist for identifying the various TCR chains, there is a lack of a concise nomenclature that would identify both the amino acid translation and nucleotide sequence of the CDR3. This deficiency makes the comparison of published TCR genetic and proteomic information difficult. To enhance information sharing among different databases and to facilitate computational assessment of clonotypic T cell repertoires, we propose a clonotype nomenclature. The rules for generating a clonotype identifier are simple and easy to follow, and have a built-in error-checking system. The identifier includes the V and J region, the CDR3 length as well as its human or mouse origin. The framework of this naming system could also be expanded to the B cell receptor.

Keywords TCR · Nomenclature · CDR3 · Clonotype

Introduction

A hallmark of immunity is the intrinsic ability to recognize and eliminate foreign molecules, cells, and organisms. The adaptive immune system is comprised of B and T cells. During T and B cell development these cells express unique heterodimeric receptors that can be used in pathogen recognition. Each of these receptor chains is generated by a somatic rearrangement process that joins different segments of the TCR and BCR genes and creates a novel gene. This joining process is imprecise with insertion of non-templated nucleotides (N nucleotides) in the junction site, as well as 3′- and 5′-nucleotide deletion from the germline genes participating in the rearrangement. This region of random nucleotide insertion or deletion referred to as the third complementarity-determining region (CDR3). The resulting CDR3 have a unique nucleotide sequence that is specific to that particular B or T cell and all its progeny; hence, the clonotypic nature of the receptors. The CDR3 is the portion of these receptors that is most involved in interactions with intact soluble antigens (B cells) or intracellular processed antigens presented as immunogenic peptides loaded in MHC molecules (T cells)

The initial phase of the adaptive immune responses involves B and T cell clonal selection on the basis of the structural complementarity of antigen-specific receptors to pathogen-derived epitopes (Davis and Chien 2003; Kolar and Capra 2003). The cells recruited into the immune response execute their effector function role. After patho-

Electronic supplementary material The online version of this article (doi:10.1007/s00251-009-0383-x) contains supplementary material, which is available to authorized users.

M. B. Yassai (✉) · J. Gorski
Molecular Genetics Laboratory, Blood Research Institute,
BloodCenter of Wisconsin,
Milwaukee, WI 53226, USA
e-mail: maryam.yassai@bcw.edu

Y. N. Naumov
Department of Pathology,
University of Massachusetts Medical School,
Worcester, MA 01655, USA

E. N. Naumova
Department of Public Health and Family Medicine,
Tufts University School of Medicine,
Boston, MA 02111, USA

gen clearance, a proportion of these cells will be retained as memory. Memory provides more rapid and effective immune protection against recurring pathogen present in the environment. The collection of cells that respond to a particular pathogen is referred to as the repertoire.

T and B cells can also be implicated in responses to non-pathogenic environmental stimuli (allergies). More serious is the lack of tolerance to self that results in responses to self-antigens giving rise to autoimmune disease. In each case, a repertoire of allergen- or self-specific B or T cells is generated.

The repertoire recognizing a molecule would be a sum of the repertoires responding against all the component epitopes of the molecule. The repertoire against an organism would be the sum of all the repertoires against all the molecules from the pathogen.

Measuring an immune response at the level of the repertoire is becoming very common (Correia-Neves et al. 2001; La Gruta et al. 2008; Naumov et al. 1998; Pewe et al. 2004; Probert et al. 2007; Venturi et al. 2008). An antigen-specific response can be viewed in the context of how many T cells are recruited and the structure of their antigen receptors. The nature of the naïve and antigen-experienced cells repertoire is of interest in basic and clinical immunology, immune-pharmaceutics, and vaccine development. However, comparison of datasets from similar, or even identical, experiments from different laboratories is cumbersome due to lack of the unified clonal identification procedure where the clonotypic antigen-receptor serves a marker of clonal identity. Having a quick way to assign specific identifiers for specific receptor sequence would facilitate such comparison studies.

There are two subsets of T cells based on the exact pair of receptor chains expressed. These are either the alpha (α) and beta (β) chain pair, or the gamma (γ) and delta (δ) chain pair, identifying the αβ or γδ T cell subsets, respectively. The expression of the β and δ chain is limited to one chain in each of their respective subsets and this is referred to as allelic exclusion (Bluthmann et al. 1988; Uematsu et al. 1988). These two chains are also characterized by the use of an additional DNA segment, referred to as the diversity (D) region during the rearrangement process. The D region is flanked by N nucleotides which constitutes the NDN region of the CDR3 in these two chains.

The CDR3 of each of the two receptor chains defines the clonal specificity. For αβ T cells the CDR3 is in most contact with the peptide bound to the MHC (Rudolph et al. 2006). For this reason, CDR3 sequences have been the main focus for sequencing studies. In the past three decades, TCR clone sequences have been presented in publications in many different forms. Some, using an alias as an identifier and present a whole nucleotide sequence of a clone by identifying the V, D, and J segments (Elliott et al. 1988). In some publications, the information about the V and the J usage and the amino acids of the V/NDN/J junction sequences (Kent et al. 2005) are given, while in other publications, both nucleotides and amino acid sequences of all different segments that have been recombined to make up the CDR3 region of the TCR clones are given (Maslanka et al. 1996; Naumov et al. 1998; Shin et al. 2005). However, a full sequence could be quite bulky. Often, for simplicity, each sequence is assigned its own alias that could be a number or a combination of letters and numbers to ease the tracking of information (Cameron et al. 2002; Chien et al. 1987; Correia-Neves et al. 2001; Davis and Bjorkman 1988; Elliott et al. 1988; Kalams et al. 1994; La Gruta et al. 2008; Lehner et al. 1995; McHeyzer-Williams and Davis 1995; Naumov et al. 1998, 2006; Pewe et al. 2004; Venturi et al. 2008). With the arrival of new ultra-high throughput or massively parallel sequencing techniques these data sets are bound to grow larger. Without a proper standardization, the general compilation of such information across published and documented data sources is problematic. Thus, there is a need for a nomenclature which allows to properly enumerating the TCR chains and tracing them to the T cell clones.

The primary purpose of this naming system is to have a unique identifier for the CDR3 of each TCR chain, so that information about the T cell clones in publications, databases, and other forms of communication can be unambiguously associated with the correct T cell clone. The proposed nomenclature is intended to provide the immunology community an easy route to share genetic information about clonal and clonotypic T cell receptors.

Materials and methods

T cell clonotypes

To properly document and enumerate TCR CDR3, we have developed a working definition of a clonotype and a nomenclature that reflects the sequence information of the CDR3 of that particular receptor:

1. A TCR clonotype is a unique nucleotide sequence that arises during the gene rearrangement process for that receptor. The combination of nucleotide sequences for the surface expressed receptor pair would define the T cell clonotype.
2. Clonotyping is a process to identify the unique nucleotide CDR3 sequences of a TCR chain. This generally involves PCR amplification of the cDNA using V-region-specific primers and either constant region (C) specific or J-region-specific primer pairs, followed by nucleotide sequencing of the amplicon.

3. Clonotype nomenclature is the system for assigning identifiers and tracing records of clonotype identification.

The clonotype nomenclature

The clonotype nomenclature refers to a system of names that are fully controlled through explicit and rigid syntactic rules. We have also defined a list of desired features for a clonotype identifier that allows computational assignment. To minimize the identifier length and to maintain clarity a mix of letters and digits is used. There is a firm restriction on the use of capitalization and character formatting in a formal name. For the clonotype nomenclature, lowercase letters are reserved for amino acid sequences in the V and J regions and uppercase letters are reserved for amino acid sequences in the region between the V and J. Thus, the uppercase corresponds to amino acids encoded by the N or NDN regions. To make names fully computable, we are avoiding the use of subscripts, superscripts, accents, and word separators; Greek symbols are replaced by uppercase Roman letters; the period ('.') is used as a symbol separator.

The clonotype-naming process

The name contains information on amino acid sequence originating from V, J, and NDN regions. The name consists of five segments: (1) CDR3 amino acid identifier, (2) CDR3 nucleotide sequence identifier, (3) variable (V) segment identifier, (4) joining (J) segment identifier, and (5) CDR3 length identifier. The name can be constructed and deconstructed in the same manner. Access to a standard genetic code table and the germline configuration of the V and J regions identified in the name is all that is needed to reconstruct the actual nucleotide sequence of the clonotype.

The rules for clonotype naming are as follows:

1. CDR3 amino acid identifier

This segment uses the one-letter code and always starts and ends with a lowercase letter. The starting lowercase letter represents the last amino acid from the V segment which is completely (all three nucleotides) encoded from the V region. The final lowercase letter represents the first amino acid entirely encoded by the J region. Uppercase letters represent amino acids that are encoded fully or in part by the NDN region.

2. Nucleotide sequence identifier (ID)

A series of digit numbers (ID) with a leading period for a symbol separator is reserved for a nucleotide identifier. Each digit in this number reflects the specific codon for each uppercase amino acid in the name. These numbers are not limited and appear in the same order as their amino acid counterparts. The identifier assignment is based on the standard codon table (Table 1). The codons for each amino acid are numbered sequentially from top to bottom and then across and down for the six codon amino acids. The three termination codons are assigned "O" and numbered 1 for Ochre (TAA), 2 for Amber (TAG), and 3 for Opal (TGA). The letter "O" is chosen because two of the three terminators start with this letter and no amino acid is associated with this letter.

3. TCR V region identifier

The V gene family (also referred to as group) is identified by an uppercase Roman letter followed by a

Table 1 Genetic codes and their assigned ID numbers

	T			C			A			G		
T	TTT	F	1	TCT	S	1	TAT	Y	1	TGT	C	1
	TTC	F	2	TCC	S	2	TAC	Y	2	TGC	C	2
	TTA	L	1	TCA	S	3	TAA	Ochre=O	1	TGA	Opal=O	3
	TTG	L	2	TCG	S	4	TAG	Amber=O	2	TGG	W	1
C	CTT	L	3	CCT	P	1	CAT	H	1	CGT	R	1
	CTC	L	4	CCC	P	2	CAC	H	2	CGC	R	2
	CTA	L	5	CCA	P	3	CAA	Q	1	CGA	R	3
	CTG	L	6	CCG	P	4	CAG	Q	2	CGG	R	4
A	ATT	I	1	ACT	T	1	AAT	N	1	AGT	S	5
	ATC	I	2	ACC	T	2	AAC	N	2	AGC	S	6
	ATA	I	3	ACA	T	3	AAA	K	1	AGA	R	5
	ATG	M	1	ACG	T	4	AAG	K	2	AGG	R	6
G	GTT	V	1	GCT	A	1	GAT	D	1	GGT	G	1
	GTC	V	2	GCC	A	2	GAC	D	2	GGC	G	2
	GTA	V	3	GCA	A	3	GAA	E	1	GGA	G	3
	GTG	V	4	GCG	A	4	GAG	E	2	GGG	G	4

ID numbers are assigned to the codons for each amino acids in the codon table by numbering sequentially from top to bottom and across giving "1" for the first sequence that coded any amino acid and "2" for the second sequence that coded the same amino acid and so on. The termination sequences are assigned the letter "O". The letter "O" is used because none of the one letter amino acid codes are represented as "O". ID number of "1" is assigned to the first termination code TAA (Ochre, O1), and 2 for the TAG (Amber, O2), and 3 for the TGA (Opal, O3) specific subfamily (also referred to as subgroup) identifier. In order to sort the clonotypes based on the V gene usage we assign a fixed number of characters for the V gene subfamilies. The names of the human V gene subfamilies are as originally described by Hood and colleagues (Rowen et al. 1996). Each V gene is assigned a subfamily number (two digits) followed by S and another number to define the subfamily member. In the case of TCR AV and TCR BV genes, some subfamilies have more than one member. The members are identified by S1, S2, S3, … for human and −1, −2, −3, … for mouse. The names of the mouse V genes are based on the ImMunoGeneTics (IMGT) database (Giudicelli et al. 2005). For mouse distal V alpha genes that are repeats of the proximal ones, we omitted the "–" in the name to keep the total characters to five, similar to that for human V genes. The breakdown of the assigned characters is shown in Table 2.

The identification of a subfamily member from a TCR sequence focused on the CDR3 depends on the specificity of the V region primer and the sequence homology of the subfamily members in the DNA segment 3′ of the V primer. If the primer is specific enough to distinguish a specific subfamily member, the clonotype name will have the specific subfamily member's name. If the primer pairs to the region that all subfamily members have identical sequences, then the DNA sequence 3′ of the primer will determine the TCRV name. If all subfamily members have identical sequence for this region, the subfamily member's name will be SX for human, and −X for mouse. If some family members can be defined but others cannot, the indistinguishable subfamilies are referred to using Y and Z. The possible members that comprise Y and Z should be further explained.

Some AV genes can rearrange to either alpha J genes (resulting in a TCR alpha chain) or delta J genes (resulting in a TCR delta chain). These are called ADV genes. Based on their location in the AV locus region, we simplify the nomenclature by using the alpha gene name. The J region identifier then specifies to which constant region the VA is linked. Shown in Table 3 are the human and mouse alpha/delta genes and our corresponding nomenclature. There are two genes that do not follow this rule; the human delta V1 gene which is located between the AV23 and AV24 genes only rearranges to the delta J genes and yet has not been found rearranging to alpha J genes, and mouse AV15-2/DV6-2 and AV15D-2/DV6-2 genes are similar and yet have not been found rearranging to the alpha J genes. In our naming system, the delta V name will be used for these genes; D1 for human and D6-2 for mouse.

Allelic forms of V regions exists (http://imgt.cines.fr/textes/IMGTrepertoire/Proteins/#B). Currently, the clonotype nomenclature does not account for these. They could be identified by enlarging the V region identifier by one or two characters. The need for this level of characterization is unclear at this time so the identifier has been kept shorter for sake of usability.

4. TCR J region identifier

The J gene identifier appears after the V gene identifier. The J gene family is expressed by Roman letters as defined for the V gene identifier above. Human (Rowen et al. 1996) and mouse (Giudicelli et al. 2005) J genes are named as described. For sake of brevity and to facilitate sorting, the "S" for designation of subfamily members in human and the "–" for designation of the subfamily members for mouse is dropped, resulting in a two-digit number. The detail of J character assignment is shown in Table 4. It should be noted that there are five subfamily in human gamma J family; 1, 2, P, P1, and P2, that two of the subfamilies have been identified by assigned numbers (gamma J1 and gamma J2), one has a been identified by assigned letter (gamma JP) and two have been identified by a letter and a number (gamma JP1 and gamma JP2). In order to have the same characters for all human gamma J genes, we are assigning a number to the ones that do not have a number identifier as follows; GJP = GJ3, GJP1 = GJ4, and GJP2 = GJ5.

There are a number of alleles of J regions that have been reported (Lefranc and Lefranc 2001 & http://imgt.cines.fr/textes/IMGTrepertoire/Proteins/#B). Currently, the nomen-

Table 2 Breakdown of assigned characters for human and mouse V genes in clonotype identifier

TCR	TCR chain assigned ID	Human V gene characters (No.) letter (No.)	Human V gene overall characters	Mouse V gene characters (No.) "− "or letter (No.)	Mouse V gene overall characters
TCR α	A	(2 digit) S (1 digit)	5	(2 digit) "–" or "D" (1 digit)	5
TCR β	B	(2 digit) S (1 digit)	5	(2 digit) "–" (1 digit)	5
TCR γ	G	(2 digit)	3	(1 digit)	2
TCR δ	D	(1 digit)	2	(1 digit) "–" (1 digit)	4

Column one represents the TCR chain. Column two represents the human V characters and their placeholders. For example, the V04S1 is displayed as 04S1 that has two digits (04) and the letter (S), and the 1 digit (1). Column three represents the overall human V gene characters, for example the 5 for B04S1. Columns four and five represent the same for mouse V genes in the clonotype identifier

Table 3 Human and mouse α/δ gene assignment

Human alpha/delta v genes (by Rowen et al.)	Human assigned V name (proposed nomenclature)	Mouse alpha/delta V genes (by IMGT)	Mouse assigned V name (proposed nomenclature)
hADV14S1	A14S1	AV4-4/DV10	A4-4
hADV23S1	A23S1	AV6-7/DV9	A6-7
hADV29S1	A29S1	AV13-4/DV7	A13-4
hADV36S1	A36S1	AV14D-3/DV8	A14D3
hADV38S2	A38S2	AV15-1/DV6-1	A15-1
		AV16D-1/DV11	A16D1
		AV21-1/DV12	A21-1

Column one represents the Human α/δ V gene names based on (Rowen et al. 1996). Column two represents the human α/δ V gene names assigned by the present nomenclature. Column three represents the mouse α/δ V gene names based on IMGT nomenclature. Column four represents the mouse α/δ V gene names assigned by the present nomenclature.

clature does not take these into account. If needed, the J identifier could be extended by one character to include an allele identifier.

5. CDR3 length identifier

The length of the clonotype is determined by the number of amino acids between the C-terminal-conserved cysteine (C) of the V region, and phenylalanine (F) of the J region which is part of the FG×GT conserved motif in all J regions. The C and the F are not counted in the length. The number representing the length is preceded by a letter L that serves as a symbol separator. In order to sort the clonotypes based on their length, we assigned three characters for the length, the first being the letter "L", followed by the two digits specifying the length.

Results and discussion

Generating a TCR clonotype identifier The use of the nomenclature is demonstrated for a TCR β-chain clonotype from our studies of CD8 T cells from HLA-A2.1 individuals responding against the influenza A matrix protein M1-derived peptide, $M1_{58–66}$ (Fig. 1). The nucleotide sequence is shown in the center with the amino acid translation underneath. The nucleotides corresponding to the NDN region are bold and underlined. The last amino acid which is completely encoded from the V gene (agt) is Serine, so the clonotype name will start with lowercase letter "s". Amino acids in the NDN region (IRSS) are presented as uppercase letters. It should be noted that I is partially encoded by the V gene and the second S is partially encoded by the J gene. The first amino acid which is completely encoded from the J gene is Tyrosine (tac) and is denoted as a lowercase "y". A period is used as a symbol separator between the amino acid identifier and the nucleotide identifier. Each uppercase letter in the NDN region (IRSS) gets a nucleotide identifier number based on the codon table (Table 1). An Ile that is encoded by ATT gets number 1, Arg that is encoded by CGG gets number 4, Ser that is encoded by AGT gets number 5, and Ser that is encoded by AGC gets number 6. In this case, the four-digit numbers "1456" identifies the nucleotide sequence of NDN region (atT, CGG, AGT, AGc). The origin of the V gene is expressed by B for beta, followed by the subfamily19 and subfamily member S1. The "S" in the V region identifier indicates that the clonotype is derived from human T cell. The origin of the J gene is shown by B for beta and subfamily 2 and the subfamily member 7. The length of the clonotype is the number of amino acids between the

Table 4 Breakdown of assigned characters for human and mouse J genes in clonotype identifier

TCR	TCR Chain assigned ID	Human J Gene Characters (No.)	Human J Gene Overall Characters	Mouse J Gene Characters (No.)	Mouse J Gene Overall Characters
TCR α	A	(2 digit)	3	(2 digit)	3
TCR β	B	(2 digit)	3	(2 digit)	3
TCR γ	G	(1 digit)	2	(1 digit)	2
TCR δ	D	(1 digit)	2	(1 digit)	2

Column one represents the TCR chain. Column two represents the human J characters which is number and their place holders that are 2. For example, J2S1 is displayed as 21 (two digits). Column 3 represents the overall human J gene characters for example B21. Columns 4 and 5 represent the same for mouse J genes in the clonotype identifier

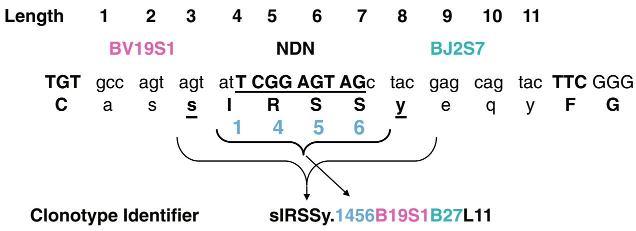
Length 1 2 3 4 5 6 7 8 9 10 11
BV19S1
NDN
BJ2S7
TGT gcc agt agt atT CGG AGT AGc tac gag cag tac TTC GGG
C a s s I R S S y e q y F G
1 4 5 6
Clonotype Identifier
sIRSSy.1456B19S1B27L11

Fig. 1 An example of TCR β-chain clonotype identifier. The BV and the BJ regions are fully identified. The single-letter -code amino acid translation is shown *below the nucleotide sequence*. The *bold uppercase letters* represent the conserved amino acids from the V (*C*) and from the J (*FG*). The amino acids that are not completely encoded by the germline, which are predominantly encoded by the NDN, are also in *uppercase* (IRSS). *Below the NDN-encoded amino acids* is the codon ID for each of them as assigned from the Table 1. The *bold underlined lowercase letters* represent the last amino acid that is completely encoded by the V gene (*s*) and the first amino acid that is completely encoded by the J region (*y*). The clonotype identifier takes the uppercase NDN amino acids and flanks them with the lowercase V and J encoded amino acids. This is followed by the codon ID for the uppercase NDN sequence. The V and J chains are next identified. Finally, the length of the CDR3 is determined by counting the number of amino acids between the uppercase C and uppercase FG. This count is shown in the *top line*

cysteine (C) of the V region and phenylalanine (F) of the J region which is part of FG×GT, which in this case is 11. Thus, the length identifier is the letter "L" for length followed by number 11.

It should be noted that the example shown here does not account for allelic differences in the V or the J genes. An example of the same clonotype identifier with the J allele information would be sIRSSy.1456B19S1B271L11, with the first two digits of the J identifier specifying the subfamily member (2S7) and the final digit specifying the allele.

V region nomenclature For TCRAV and TCRBV, some BV subfamilies have more than one member. The identification of the subfamily members depends on two factors. The first is the specificity of the V region primer that is used for amplifying the particular V subfamily member. Primers could be designed that are specific for only one subfamily member. If the primer is specific enough to anneal only to one of the V subfamily members, then the clonotype identifier will use the subfamily member's name such as S1, S2, S3 (for human), and −1, −2, −3 (for mouse). The second factor is the sequence homology between the subfamily members in the region 3′ of the V primer up to the conserved cysteine, the nucleotide differences downstream the conserved cysteine is not considered due to possibility of excision during the rearrangement process. For some choices of primer, there may be sufficient differences in the region between the primer and the conserved cysteine that the particular subfamily member can be identified. If this is the case, then the name of subfamily member is used. In other cases, the sequence between the V primer and the conserved cysteine is associated with multiple sequences. We reserve the letter X for use if the primer does not allow any distinction of subfamily members. Y and Z can be used to designate subsets of possible subfamily members and these must be defined. These designations will be specific for the primers used and once defined can be used over and over. An example of V gene identification is shown in Supplementary Table 1.

Identifying other chains Additional examples of using the nomenclature for human α-TCR, β-TCR, γ-TCR, and δ-TCR are shown in Table 5. Since the δ-chain could be the result of either Vδ- or Vα-chain genes rearranging to Jδ, we show an example of the naming of both such possibilities (examples 4 and 5).

Table 5 Examples of human TCR α, β, γ, and δ clonotype identifiers

	TCR V Gene	NDN	TCR J Gene	Clonotype Identifier
α A.A. coding	**TGT** gct tat agg **a** C a y **r**	**CG** T 4	agt gga ggt agc aac tat aaa ctg aca **TTT GGA** **s** g g s n y k l t F G	**rTs.4A38S2A53L13**
β A.A. coding	**TGC** agc **gt** C **s** V 4	**G AAG GAC GGG GC** K D G A 2 2 4 1	**t** ggg tac acc **TTC GGT** **g** y t F G	**sVKDGAg.42241B29S1B12L09**
γ A.A. coding	**TGT** gcc ttg tgg gag gtg C a l w e **v**		caa gag ttg ggc aaa aaa atc aag gta **TTT GGT** **q** e l g k k i k v F G	**vq.G09G3L14**
δ A.A. coding	**TGT** gac acc C d **t**	**TGG GGG AG** W G S 1 4 6	**c** tcc tgg gac acc cga cag atg ttt **TTC GGA** **s** w d t r q m f F G	**tWGSs.146D2D3L13**
δ A.A coding	**TGT** gac atg aga **g** C d m **r**	**CG GTG T** A V 4 4	**ac** acc gat aaa ctc atc **TTT GGA** Y **t** d k l i F G 2	**rAVYt.442A14S1D1L08**

The top row of each block is the nucleotide sequence of the TCR chain, middle row is the amino acid translation, and bottom row is the assigned number for the amino acids that are not completely encoded by the germline gene. The uppercase underlined letters (shown in bold) represent the conserved amino acids from the V region (C) and the J region (F), and the amino acids that are not completely encoded from the germline gene. The lowercase underlined letters (shown in bold) represent the last amino acid that is completely encoded by V gene and the first amino acid that is completely encoded by J gene

Decoding TCR clonotype identifier By decoding the name, the nucleotide sequence of the TCR chain can be derived in a reverse manner as that used for the encoding. Using the first example shown in Table 5 “rTs.4A38S2A53L13”, the “A38S2” and “AJ53” shows that the clonotype origin is human and the sequences of the alpha V38S2 and alpha J53 genes are needed for the decoding (Fig. 2). The entire length between the “C” and “FG” is 13 amino acids long. So the genomic sequence of the TCR AV38S2 is obtained and the positions of the amino acids lined up with a length ruler starting with the position immediately after the conserved cysteine. The TCR AJ53 is then placed so that the last amino acid before the conserved FG lines up with the end of length ruler. The two lowercase letters “r” and “s” in the name identify the last V and the first J position encoded by the germline. This leaves one position to be filled by the N nucleotides and this is the threonine represented in the clonotype name as “T”. The codon table shows that the codon 4 for T is ACG. The T can be encoded entirely by N nucleotides (ACG) or the initial nucleotide, a, could be V-germline-encoded and the rest of the sequence is N derived. While Fig. 2 uses the second possibility (aCG), the sequence is decoded regardless.

D regions

Our nomenclature does not define the D region of the clonotype. The D regions could be defined after decoding by a homology search. Because of the truncation of D regions, they are often difficult to unambiguously assign. Defining the D region usage would be left to the individual investigator.

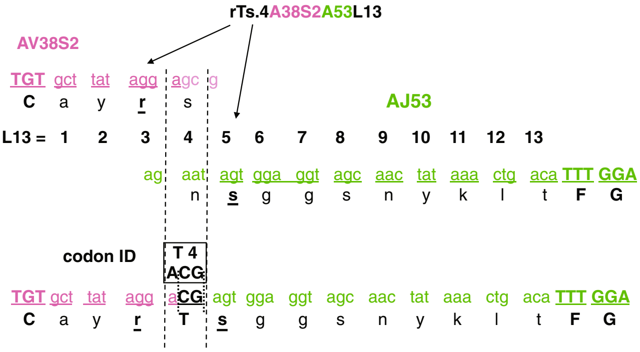
rTs.4A38S2A53L13
AV38S2
TGT gct tat agg agc g
C a y r s
AJ53
L13 = 1 2 3 4 5 6 7 8 9 10 11 12 13
ag aat agt gga ggt agc aac tat aaa ctg aca TTT GGA
n s g g s n y k l t F G
codon ID
T 4
ACG
TGT gct tat agg aCG agt gga ggt agc aac tat aaa ctg aca TTT GGA
C a y r T s g g s n y k l t F G

Fig. 2 Deriving the nucleotide sequence of the CDR3 by decoding the clonotype TCR β-chain identifier. The genomic sequence of the TCRV gene (AV38S2) is obtained and the positions of the amino acids lined up with a length ruler starting with the position immediately after the conserved cysteine. The TCRJ gene (AJ53) is then placed so that the last amino acid before the conserved FG lines up with the end of length ruler. The two lowercase letters “r” and “s” in the name identify the last V and first J position encoded by germline. This leaves one position to be filled by the N nucleotides and this is the threonine represented as “T” in the clonotype name. The codon table shows that codon 4 for T is ACG, and the only way that the T can be encoded is that the initial nucleotide, A, is from the V germline sequence and the rest of the sequence is N derived

Properties of the naming system

The nomenclature described here has a number of important properties:

1. The nomenclature is exhaustive: it ensures that each clonotypes has an identifier.
2. The nomenclature is compact: an identifier is relatively short.
3. The nomenclature is an open system in that the identifiers are not restricted. Therefore, it allows expansion or addition of identifiers.
4. The nomenclature is compartmental allowing manipulation of identifiers. Manipulations could be sort, select, arrange, etc. Single or combinations of identifiers can be manipulated.

Advantages of the naming system

By having these characteristics, the nomenclature has several general advantages. By combining all five elements of the CDR3 region, this system permits any clonotype to be defined. Our nomenclature is more compact than either a nucleotide- or amino-acid-based naming system. It is two-thirds shorter than the CDR3 nucleotide sequences, while still describing the nucleotide sequence. The CDR3 amino acid sequence is pared to the NDN contribution only. It distinguishes clonotypes that use different encoding for the same CDR3 amino acid sequence. For example, we have found 207 different clonotypes that use the same BV19S1 and the same BJ2S7, and have the exactly the same CDR3 amino acid sequence (CASSIRSSYEQYF). Even if the amino acid identifier of the name is the same without nucleotide identifier, it is impossible to distinguish between them. We show some examples of this in Table 6 from our analysis of the HLA-A2-restricted response to influenza $M1_{58–66}$. This shows the power of the nomenclature for defining population studies that deal with a large number of similar clonotypes.

By being compartmental, the proposed nomenclature can enumerate all possible names. Each compartment is an identifier. While it is unlikely that new J or V regions will be uncovered in mice or man, these could be easily absorbed into the name. The compartmentalization allows the level of identification of the V region to reflect in the name. If the identification of polymorphic variants of either V or J regions becomes important, the size of the compartment for these regions could be expanded to

Table 6 Examples of different human clonotype sequences coding identical amino acids in the CDR3β

CDR3 Amino Acid Sequence	TCR V Gene	NDN	TCR J Gene	Clonotype Identifier
CASSIRSSYEQYF	**TGT** gcc agt agt ata **C** a s s **i**	**AGA AG** **R S** **5 6**	**C** tcc tac gag cag tac **TTC GGG** **s** y e q y **F G**	**iRSs.56B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt agt at **C** a s **s I** **1**	**T CGC AG** **R S** **2 6**	**C** tcc tac gag cag tac **TTC GGG** **s** y e q y **F G**	**sIRSs.126B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt agt ata **C** a s s **i**	**AGG AGC AGT** **R S S** **6 6 5**	tac gag cag tac **TTC GGG** **y** e q y **F G**	**iRSSy.665B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt agt at **C** a s **s I** **1**	**T CGG AGC TCT** **R S S** **4 6 1**	tac gag cag tac **TTC GGG** **y** e q y **F G**	**sIRSSy.1461B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt ag **C** a **s S** **6**	**C ATC AGA AG** **I R S** **4 6 1**	**C** tcc tac gag cag tac **TTC GGG** **s** y e q y **F G**	**sSIRSs.6256B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt agt at **C** a s **s I** **1**	**T AGA TCC AGC TAT** **R S S Y** **5 2 6 1**	gag cag tac **TTC GGG** **e** q y **F G**	**sIRSSYe.15261B19S1B27L11**
CASSIRSSYEQYF	**TGT** gcc agt ag **C** a **s S** **6**	**C ATA AGG TCC AG** **I R S S** **3 6 2 6**	**C** tac gag cag tac **TTC GGG** **y** e q y **F G**	**sSIRSSy.63626B19S1B27L11**

The top row of each block is the nucleotide sequence encoding CDR3, middle row is the amino acid translation, and bottom row is the assigned number for the amino acids that are not completely encoded by the germline gene. The uppercase underlined letters (shown in bold) represent the conserved amino acids from the V region (C) to the J region (F), and the amino acids that are not completely encoded from the germline gene. The underlined lowercase letters (shown in bold) represent the last amino acid that is completely encoded by the V gene, and the first amino acid that is completely encoded by the J gene

facilitate the addition. If the system were to be used for naming of BCR, an identifier for the heavy chain constant region could be added after the J identifier. The structure allows these identifiers to be fully computable and the character assignment of gene identifier makes it easy to sort based on the V gene, J gene, and the length. It also supports a built-in error checking for the digits in ID and number of amino acids in the NDN region, which is a one to one relation for all functional TCR clones. For example, if there are four amino acids in the NDN region, there would be four-digit numbers in the ID part of the name and the errors are easily found.

Table 7 Identifiers of the HLA-A2.1:M1$_{58–66}$-specific clones/clonotypes found in multiple studies

Clonotype identifier	Moss et al. 1991 (Clone ID)	Lehner et al. 1995 (Clone ID)	Naumov et al. 1998 (Clonotype ID)	Naumov et al. 2006 (Clonotype ID)
sIRSs.146B19S1B27L11		DDD8	132	19.27
iRSs.66B19S1B27L11	1a8		71	16.27
sIRSs.226B19S1B27L11	B1b	MODG5		
iRSt.62B19S1B22L11	B1c			
sIRs.24B19S1B23L11	B1d			
sMRSs.166B19S1B27L11		JNJ1		43.27
iRSSy.626B19S1B27L11		NMH8		
sIRSAy.2662B19S1B27L11		KEF9	94	28.27
sTRs.23B19S1B23L11		HLE19		13.23
sMRSs.163B19S1B27L11		MODG4		
sMRs.16B19S1B23L11		JN5K2		

Different HLA-A2.1 individuals were recruited in the studies reported as Naumov et al. (1998) and (Naumov et al. (2006). The numbers of individuals sharing identical influenza M1$_{58–66}$-specific clonotypes are shown in the first column. The clonotype identifiers are shown in second column. The identifiers of the M1$_{58–66}$-specific CD8 T cell clones reported by Moss et al. (1991) and Lehner et al. (1995) and clonotypes reported by Naumov et al. (1998) and Naumov et al. (2006) are shown in the corresponding columns, respectively

Comparing TCR clonotypes To the extent that clonotypes are public (1), they can be identified in many laboratories. Thus, a fixed nomenclature will avoid difficulties associated with local identifiers. For example, the M1 response in HLA-A2 individuals has been studied by many groups. We show that some of the published clonotypes identified by Moss et al. in 1991 and by Lehner et al. in 1995, and us (Naumov et al. 1998, 2006) were observed in more than one study (Table 7). This shows the power of a common robust naming system in comparing the results of related studies that have been published independently.

Alternative codon numbering systems We also examined an alternative approach for codon numbering. We used the same codon numbering table, as described above, but then generated a list of all the possible ways for encoding of a particular NDN sequence. The observed sequence is then defined by its index position on the list. The IRSS amino acid sequence could be used as an example: when the clonotype identifier encodes I1 R1, S1, S1, ID number would be 1 instead of 1111. When the clonotype identifier encodes I1 R1, S1, S2, ID number would be 2 instead of 1112. Since IRSS has 648 possible encoding combinations (3×6×6×4), the identifiers would be shorter using one to three characters. However, the disadvantage of using this approach is that it is less direct and requires a computer program for optimal implementation. This takes away the ability for an individual to manually identify or decode a particular sequence.

Summary We have implemented a rational nomenclature system that makes the TCR sequences easier to read and compare. The nomenclature rules are simple and easy to implement. Having the codon table available, it would be easy to name any TCR clonotype or clone without developing customized naming software. Nevertheless, the rules are simple enough to be encoded in computer programs. It has a built-in error-checking system which is the one to one correspondence between each digit in ID and each uppercase amino acid in the clonotype identifier. The benefits of our consistent nomenclature would accrue exponentially as the number of TCR under the study increases.

Implementing this nomenclature would facilitate deployment of clonotype databases run by individual laboratories for specific immune responses and immune diseases. The clonotype names within these databases would be reliable, error-free, and allow easy cross-referencing and comparison of T cell repertoires by different laboratories. The clonotypes could be cataloged in a single database and annotated as to their occurrence and associations with particular responses. Such a catalog is only possible by providing an easy-to-use nomenclature. T cell clones can be unambiguously identified by naming both chains. The same identification could be provided for single-cell PCR data where both chain sequences are available. The framework of this naming system could also be implemented for B cell receptors if a system was added to account for somatic hypermutation. Such a convention would open up the possibility of creating a BCR catalog which would be a useful tool for investigators working on BCR repertoires.

Acknowledgments We thank Dr. Marie-Paule Lefrance for clarification of the mouse TCR nomenclature. We also thank Dr. Andrea Ferrante for helpful discussions. This work was funded by the National Institutes of Health Grant U19 AI062627.

References

Bluthmann H, Kisielow P, Uematsu Y, Malissen M, Krimpenfort P, Berns A, von Boehmer H, Steinmetz M (1988) T-cell-specific deletion of T-cell receptor transgenes allows functional rearrangement of endogenous alpha- and beta-genes. Nature 334:156–159. doi:10.1038/334156a0

Cameron TO, Cohen GB, Islam SA, Stern LJ (2002) Examination of the highly diverse CD4(+) T-cell repertoire directed against an influenza peptide: a step towards TCR proteomics. Immunogenetics 54:611–620. doi:10.1007/s00251-002-0508-y

Chien YH, Iwashima M, Wettstein DA, Kaplan KB, Elliott JF, Born W, Davis MM (1987) T-cell receptor delta gene rearrangements in early thymocytes. Nature 330:722–727. doi:10.1038/330722a0

Correia-Neves M, Waltzinger C, Mathis D, Benoist C (2001) The shaping of the T cell repertoire. Immunity 14:21–32. doi:10.1016/S1074-7613(01)00086-3

Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. Nature 334:395–402. doi:10.1038/334395a0

Davis MM, Chien YH (2003) T cell antigen receptors. In: Paul WE (ed) Fundamental immunology, 5th edn. Lippincott Williams & Wilkins, Philadelphia, pp 227–258

Elliott JF, Rock EP, Patten PA, Davis MM, Chien YH (1988) The adult T-cell receptor delta-chain is diverse and distinct from that of fetal thymocytes. Nature 331:627–631. doi:10.1038/331627a0

Giudicelli V, Chaume D, Lefranc MP (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33:D256–D261. doi:10.1093/nar/gki010

Kalams SA, Johnson RP, Trocha AK, Dynan MJ, Ngo HS, D'Aquila RT, Kurnick JT, Walker BD (1994) Longitudinal analysis of T cell receptor (TCR) gene usage by human immunodeficiency virus 1 envelope-specific cytotoxic T lymphocyte clones reveals a limited TCR repertoire. J Exp Med 179:1261–1271. doi:10.1084/jem.179.4.1261

Kent SC, Chen Y, Bregoli L, Clemmings SM, Kenyon NS, Ricordi C, Hering BJ, Hafler DA (2005) Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope. Nature 435:224–228. doi:10.1038/nature03625

Kolar GR, Capra JD (2003) Immunoglobulins: structure and function. In: Paul WE (ed) Fundamental immunology, 5th edn. Lippincott Williams & Wilkins, Philadelphia, pp 47–68

La Gruta NL, Thomas PG, Webb AI, Dunstone MA, Cukalac T, Doherty PC, Purcell AW, Rossjohn J, Turner SJ (2008) Epitope-specific TCRbeta repertoire diversity imparts no functional advantage on the CD8+ T cell response to cognate viral peptides. Proc Natl Acad Sci USA 105:2034–2039. doi:10.1073/pnas.0711682102

Lefranc MP, Lefranc G (2001) The T cell receptor facts book. Academic, London

Lehner PJ, Wang EC, Moss PA, Williams S, Platt K, Friedman SM, Bell JI, Borysiewicz LK (1995) Human HLA-A0201-restricted cytotoxic T lymphocyte recognition of influenza A is dominated by T cells bearing the V beta 17 gene segment. J Exp Med 181:79–91. doi:10.1084/jem.181.1.79

Maslanka K, Yassai MB, Gorski J (1996) Molecular identification of T cells that respond in a primary bulk culture to a peptide derived from a platelet glycoprotein implicated in neonatal alloimmune thrombocytopenia. J Clin Invest 98:1802–1808. doi:10.1172/JCI118980

McHeyzer-Williams MG, Davis MM (1995) Antigen-specific development of primary and memory T cells in vivo. Science 268:106–111. doi:10.1126/science.7535476

Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, Bodmer HC, McMichael AJ, Bell JI (1991) Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. Proc Natl Acad Sci USA 88:8987–8990. doi:10.1073/pnas.88.20.8987

Naumov YN, Hogan KT, Naumova EN, Pagel JT, Gorski J (1998) A class I MHC-restricted recall response to a viral peptide is highly polyclonal despite stringent CDR3 selection: implications for establishing memory T cell repertoires in "real-world" conditions. J Immunol 160:2842–2852

Naumov YN, Naumova EN, Clute SC, Watkin LB, Kota K, Gorski J, Selin LK (2006) Complex T cell memory repertoires participate in recall responses at extremes of antigenic load. J Immunol 177:2006–2014

Pewe LL, Netland JM, Heard SB, Perlman S (2004) Very diverse CD8 T cell clonotypic responses after virus infections. J Immunol 172:3151–3156

Probert CS, Saubermann LJ, Balk S, Blumberg RS (2007) Repertoire of the alpha beta T-cell receptor in the intestine. Immunol Rev 215:215–225. doi:10.1111/j.1600-065X.2006.00480.x

Rowen L, Koop BF, Hood L (1996) The complete 685-kilobase DNA sequence of the human β T cell receptor locus. Science 272:1755–1762. doi:10.1126/science.272.5269.1755

Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 24:419–466. doi:10.1146/annurev.immunol.23.021704.115658

Shin S, El-Diwany R, Schaffert S, Adams EJ, Garcia KC, Pereira P, Chien YH (2005) Antigen recognition determinants of gamma-delta T cell receptors. Science 308:252–255. doi:10.1126/science.1106480

Uematsu Y, Ryser S, Dembic Z, Borgulya P, Krimpenfort P, Berns A, von Boehmer H, Steinmetz M (1988) In transgenic mice the introduced functional T cell receptor beta gene prevents expression of endogenous beta genes. Cell 52:831–841. doi:10.1016/0092-8674(88)90425-4

Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? Nat Rev Immunol 8:231–238. doi:10.1038/nri2260