Cutting Edge: CATERPILLER: A Large Family of Mammalian Genes Containing CARD, Pyrin, Nucleotide-Binding, and Leucine-Rich Repeat Domains

Jonathan A. Harton, Michael W. Linhoff, Jinghua Zhang and Jenny P.-Y. Ting

*J Immunol* 2002; 169:4088-4093; doi: 10.4049/jimmunol.169.8.4088

http://www.jimmunol.org/content/169/8/4088

Supplementary Material

http://www.jimmunol.org/content/suppl/2002/10/09/169.8.4088.DC1

**Why The JI?**

- **Rapid Reviews!** 30 days* from submission to initial decision
- **No Triage!** Every submission reviewed by practicing scientists
- **Speedy Publication!** 4 weeks from acceptance to publication

*average

**References**

This article cites 37 articles, 16 of which you can access for free at:

http://www.jimmunol.org/content/169/8/4088.full#ref-list-1

**Subscription**

Information about subscribing to *The Journal of Immunology* is online at:

http://jimmunol.org/subscription

**Permissions**

Submit copyright permission requests at:

http://www.aai.org/About/Publications/JI/copyright.html

**Email Alerts**

Receive free email-alerts when new articles cite this article. Sign up at:

http://jimmunol.org/alerts
Large mammalian proteins containing a nucleotide-binding domain (NBD) and C-terminal leucine-rich repeats (LRR) similar in structure to plant disease resistance proteins have been suggested as critical in innate immunity. Our interest in CIITA, a NBD/LRR protein, and recent reports linking mutations in two other NBD/LRR proteins to inflammatory disorders have prompted us to perform a search for other members. Twenty-two known and novel NBD/LRR genes are spread across eight human chromosomes, with multigene clusters occurring on 11, 16, and 19. Most of these are telomeric. Their N termini vary, but most have a pyrin domain. The genomic organization demonstrates a high degree of conservation of the NBD- and LRR-encoding exons. Except for CIITA, all the predicted NBD/LRR proteins are likely ATP-binding proteins. Some have broad tissue expression, whereas others are restricted to myeloid cells. The implications of these data on origins, expression, and function of these genes are discussed. The Journal of Immunology, 2002, 169: 4088–4093.

Jonathan A. Harton, Michael W. Linhoff, Jinghua Zhang, and Jenny P.-Y. Ting

Copyright © 2002 by The American Association of Immunologists, Inc.
(CARD, transcription enhancer, R(purine)-binding, pyrin, lots of leucine repeats) gene family. Determination of chromosomal location, genomic organization, and sequence information are also shown.

Materials and Methods

Databases and search strategies

Searches were performed using the published Celera human genome scaffold data (16), the National Center for Biotechnology Information (NCBI) "nt" database (containing GenBank, European Molecular Biology Laboratory, DNA Data Base in Japan, Protein Data Base, and completed phase 3 and 4 high-throughput genomic sequencing (HTGS) sequences), and the NCBI genome database. Initial searches were performed using the Blastp search algorithms BLASTP and TBLASTN (see supplemental data Fig. 1). Blastp identifies amino acid sequence similarities through query sequence comparison with database proteins and is more likely to find distant relationships than TBLASTN (18). BLASTN compares the query protein sequence with translations of all six reading frames of available nucleotide sequences and has the same advantages as BLASTP. We used analogous domains of the resultant sequences to identify additional sequences and/or confirm initial identities; this is known as DOUBLE-BLAST, inspired by the intermediate search sequence method of Park et al. (19, 20) and is comparable in homologue detection with Hidden Markov Methods. LRR sequences, the N-terminal pyrin domains of DEFCAP, and the CARD domains of Nod1 and Nod2 were used to perform similar searches. The N-terminal sequences of CIITA yielded no related sequences and is more likely to find distant relationships than BLASTN (18). TBLASTN compares the query sequence comparison with database proteins and is more likely to find distant relationships than BLASTN (18). TBLASTN compares the query sequence with translations of all six reading frames of available nucleotide sequences and has the same advantages as BLASTP. We used analogous domains of the resultant sequences to identify additional sequences and/or confirm initial identities; this is known as DOUBLE-BLAST, inspired by the intermediate search sequence method of Park et al. (19, 20) and is comparable in homologue detection with Hidden Markov Methods. LRR sequences, the N-terminal pyrin domains of DEFCAP, and the CARD domains of Nod1 and Nod2 were used to perform similar searches. The N-terminal sequences of CIITA yielded no related sequences obviously belonging to an NBD/LRR protein.

Assembly of putative novel genes and construction of genomic maps

Pyrin and LRR sequences identified within contigs containing NBDs were examined for location and orientation to determine the likelihood of residing in the same operon as an identified NBD. Pyrin and LRR domains were considered contiguous with an NBD if they fell upstream and downstream of the NBD, respectively, in the same orientation. CARD domains occur both upstream (Nod1/2) and downstream (DEFCAP) of the NBD (21), but none of the novel sequences contained CARD domains. As sequence data became available for more than a single domain, a putative genomic organization was generated by comparing the cDNA sequence with the genome sequence.

Cell lines, preparation of RNA, and RT-PCR

HeLa, MCF7, Jurkat, RAJ1, and RAMOS cell lines were cultured in either DMEM (high glucose) or RPMI 1640 with 10% FCS, l-glutamine, and penicillin/streptomycin. Peripheral blood leukocytes were obtained as buffy coats from the American Red Cross (Durham, NC). Total RNA was prepared using the SV Total RNA Isolation Kit (Promega, Madison, WI). Total RNA was reversely transcribed to cDNA using Moloney murine leukemia virus reverse transcriptase and amplified in an MJ Thermocycler (MJ Research, Cambridge, MA) in a separate reaction with primers specific for each target sequence. Amplification products were electrophoresed on 0.8% agarose and visualized with ethidium bromide.

Results and Discussion

Identification of novel CIITA-related sequences

BLAST searches of the published Celera and NCBI genomic databases using the NBD and LRR of CIITA, Nod1, Nod2, DEFCAP, and resultant target sequences as queries revealed 22 potential genes and pseudogenes, including the presently known genes, unified by the presence of an NBD and downstream LRRs (Table I). New genes were assigned a name based on chromosome number and order of discovery (e.g., 19.1, first found on chromosome 19). Nod1, Nod2, and DEFCAP contain CARD domains that may be involved in recruiting caspases (12, 13, 21). DEFCAP also has an N-terminal pyrin domain with homology to the familial Mediterranean fever protein (7). BLAST searches were also performed for the CARD domains of Nod1/2, the pyrin domain of DEFCAP, and resulting target sequences. CARD domain homologies were not found for any of the novel sequences. The majority of the putative genes have upstream pyrin domains, but the upstream N-terminal sequences of several remain unknown.

Conservation of intron-exon organization

We determined exon/intron sizes and positions for the known and some predicted NBD/LRR proteins by the location of the sequence corresponding to the mRNA/cDNA assuming intactness of the contig (Fig. 1A). The genomic organization is complex and remarkably similar for all the sequences examined, with large NBD exons (~1500 nt) and LRR exons of ~76 nt, 174 nt, or both depending of the gene. CARD and pyrin domains are ~300 nt long.

CATERPILLER domains

Table I highlights the distinct domains of each sequence. Nod1, Nod2, and CIITA have N-terminal CARD or CARD-like domains. Thirty have N-terminal pyrin domains. CIITA is unique in having an N-terminal acidic trans activation domain. Five of these sequences do not have CARD, pyrin, or CIITA-like activation domains upstream of their NBDs. The diversity of these N-terminal sequences suggests multiple functional modes.

The predicted nucleotide specificity based on motifs found in the CATERPILLER genes is shown in Table I. This is compared with another family, containing plant and animal proteins, grouped on the basis of a NTPase domain and C-terminal repeats of either the LRR or WD40 type, called the NACHT family, which includes NAIP, CIITA, HET-E, and TP1 (22). Remarkably, the majority are predicted to be ATP-binding proteins, with the exception of CIITA, which binds GTP, and HET-E. A GTP-binding protein-like magnesium coordination (G3) motif (DXGX) occurs in a number of the other sequences, but excepting the more distantly related ApaI, it follows the more typical kinase 2 site found in ATP-binding proteins.

We aligned the NBDs of these predicted proteins, each ~500 aa long, and observed 12 groupings of conserved residues (motifs) (Fig. 1B). The full protein alignment of the NBD domains is shown in supplemental data Fig. 2. Although the seven NACHT motifs are present, the larger number of compared sequences permits a refined definition of the NACHT domain that excludes WD40 repeat-containing members, thus distinguishing a CATERPILLER NBD from the broader NACHT family. These motif definitions also suggest a divergence between the majority of the NBDs that we describe and those like NAIP. Functionally important motifs likely include motif I, which contains the Walker A sequence found in most nucleotide-binding proteins (23), and motifs III and V that overlap or are adjacent to leucine-charged domain motifs (24). These motifs are important for CIITA function (8). Motif III contains the kinase 2 motif which coordinates magnesium ions in ATP-binding proteins (23).

The presence of LRR sequences downstream of the NBD was required for inclusion as a CATERPILLER family member. The LRR sequences following NBDs have two exon arrangements, a singlet (~74 nt) containing one motif iteration or a duplex (~180 nt) containing two (Table I (column 8), Fig. 1A, and supplemental data Fig. 3). The sole absolute requirement for inclusion as an LRR is conservation of the hydrophobic residues "leucines" comprising the motif. BLAST searches for LRRs may miss some sequences due to a greater likelihood of less similarity between non-LRR-motif residues. Thus, without actual cDNA clones, it is impossible to be highly confident that all of the LRR exons downstream of the...
NBD have been identified for each putative gene. Given this caveat, it appears that all of the genes on chromosome 19 have doublet LRR exons whereas those on chromosome 16 have singlets. DEFCAP and the potential pseudogene 12 have both singlet and doublet exons.

**Phylogenetic analysis of the NBD and evolutionary issues**

An analysis using protein alignment and tree generation software (Data Analysis in Molecular Biology and Evolution) (25) was performed to examine the potential phylogenetic relationship of the predicted NBD protein sequences (Fig. 2A). Apaf1 and RPM1 (Table I) were included because their NBD regions are similar to those of this family. Except for 11.3, the newly identified NBD sequences are more closely related to one another than Apaf1 (Fig. 2A), suggesting that NBD/WD40 repeat proteins are more distantly related. Interestingly, the NBD of RPM1, an NBD/LRR R protein of *Arabidopsis*, is most closely related to Apaf1. The novel NBD most closely related to RPM1 is 11.3 which has an NBD exon interrupted by an intron. Consistent with divergent evolution, the NBDs of the known and putative proteins with upstream CARD domains are more closely related to each other than to those NBDs with upstream pyrin domains which form their own grouping phylogenetically. Further analysis of NBD/LRR-type plant R proteins and other eukaryotic NBD/LRR proteins will help resolve issues of divergent vs convergent evolution.

The assignment of the CATERPILLER genes to chromosomal positions is shown in Fig. 2B. Most are found in clusters on chromosomes 11, 16, and 19. Three occur at 11p15, three more between 16p12 and 16p13, and nine at 19q13. Proximities of the six sequences on a single contig at 19q13.4 strongly suggest that gene duplication has occurred for these sequences. All except four of these sequences are near the telomere, suggesting that those found singly may have their origins in chromosomal recombination. Among those not at the telomeric end of chromosomes, one (X) is likely a pseudogene. In *Saccharomyces*, fermentation gene alleles are thought to have been generated by the duplication of genes close to the telomeric end and subsequent genomic dispersion by recombination (26). Comparative genomics studies will best address these questions.

The presence of multiple individual exons containing one or two LRRs implies that exon shuffling may occur and that natural selection may favor the maintenance or elimination of a given LRR sequence or pair while simultaneously preserving other aspects of the gene in question (see supplemental data Figs. 3 and 4A). The specificity of plant R proteins is principally dependent on the LRR, and these are targets for diversifying selection (15). In Flax, a 6-aa difference in the LRR of P vs P2 depends on the LRR, and these are targets for diversifying selection (27). The LRRs of RPS2 contain a small stretch important for cooperation with host factors determining *Arabidopsis* resistance to *Pseudomonas*
Unequal recombination, gene conversion, and accumulated mutations likely generate novel specificities for the NBD/LRR class of R proteins.

**Evidence for expression of the CATERPILLER genes**

In light of these data, the NBD/LRR protein family is larger than currently known. Significant information is available on the expression patterns of the known genes and this reflects their biologic role. CIITA has three different isoforms arising from three different promoters. Nod1 has a wide tissue distribution (12), whereas Nod2 and CIAS1 are restricted to monocytes, consistent with inflammatory roles (4, 13). To begin to examine the expression of the other sequences, we have used the NCBI database to search for expressed sequence tags encoding at least part of the sequence (see Table II). UniGene sequence entries exist for CIAS1, Nod1, Nod2, DEFCAP, Nalp2, and 16.1. Fourteen of the genes are represented in GenBank human expressed sequence tag (EST) database. The gene we identify as 19.3 has been previously described as a partial cDNA encoding a 344-aa protein (RNO2) composed of LRRs and is expressed in bone marrow, peripheral blood leukocytes, and nitric oxide-treated HL-60 cells (29). No EST entry was found for 11.2, 12, 19.1, 19.2, 19.5, 19.8, or X. We have also conducted a

---

**FIGURE 1.** Motifs and genomic organization. A. Genomic organization for known and some predicted members of the CATERPILLER family are shown to scale. Black boxes represent exons. Unusually large introns are interrupted and their size indicated below in kilobase pairs. Exons with ambiguous positions are shown as gray boxes. The large 3’ exons of Nod1 and Nod2 are 3’ untranslated sequences. B. NBDs were aligned using Clustal with minor manual adjustments. Twelve motifs defining the CATERPILLER NBD are shown. Capital letters indicate residues (single letter code) that have a frequency of >50% or are invariant. Lower case letters indicate residues with frequency <50% but with a predominant characteristic (a = acidic, b = basic, h = hydrophobic, p = serine/threonine, r = aromatic). *, Those residues used to define the NACHT family. Superscripts 1, 2, and 3 indicate NACHT motifs V, VI, and VIII, respectively.

**FIGURE 2.** Phylogenetic tree for NBDs and chromosomal locations of CATERPILLER genes. A. Deduced amino acid sequences from NBD exons were compared with one another using alignment and tree generation software in the Data Analysis in Molecular Biology and Evolution software package (25). *, A predicted gene with unknown N-terminal sequences. B. The chromosomal location of each known or predicted sequence. For chromosomal locations with multiple sequences, the name order does not correspond to the ordering on the chromosome.
preliminary survey of the expression of these new genes, summarized in Table II, and have detected message for every nonpseudogene except 19.1 and 19.2. Nearly all of the family members are expressed in hemopoietic cells and are likely restricted in that lineage, whereas CIAS1 in familial cold urticaria and Nod2 in immune function. CIITA directly controls MHC II gene expression.

### Immunologic significance

Of the known genes, CIITA, CIAS1, and Nod2 are clearly linked to immune function. CIITA directly controls MHC II gene expression, whereas CIAS1 in familial cold urticaria and Nod2 in Crohn’s disease are likely regulating in a subdivision of CATERPILLER proteins: group I, CARD-containing (e.g., Nod1; group II, pyrin-containing (e.g., DEFCAP); group III, trans activation domain (e.g., CIITA); and unknown (e.g., 16.1) (see Table I). However, these grouping may be oversimplified. For example, multiple cell type-specific forms of CIITA are known. The dendritic cell form has a CARD-like N terminus followed by the activation domain, although no caspase recruitment activity has been described (36). It is of interest that Nod2 and cryopyrin are also expressed as multiple transcripts (4, 13). Whether these different transcripts code for proteins of somewhat different function is clearly of interest. Additionally, self-association has also been demonstrated for CIITA and Nod1, whereas heterodimerization of CIAS1 with apoptotic protein ASC may involve the pyrin domain of CIAS1 (5, 10, 12). Self- and heteroassociation might amplify and generate diversity necessary to mediate appropriate responses.

Genes coding proteins structurally related to CIITA, Nod2, and others in having an NBD, multiple C-terminal LRRs, and few different N-terminal domains abound in the human genome. The sequences and genomic organization of these genes suggest a high degree of relatedness, a common origin, and a potential link to the basic immune response genes of plants. Studies on CIITA, CIAS1, DEFCAP, Nod1, and Nod2 reveal some interesting parallels with the plant proteins and strongly suggest that this family of proteins will likely influence mammalian immune responses.

### Related issues

The number of mammalian NBD/LRR sequences we were able to identify is significantly smaller than that occurring in some plants (35). The mammalian family may be larger than we describe as NAIP and Ipaf (CARD12), despite having NBDs and LRRs, were not detected using our parameters (except when using 16.2), likely due to the absence of some of the CATERPILLER motifs in their NBDs. Limited BLAST searches of translated nucleotide sequences from *Drosophila* and *Caenorhabditis elegans*, genomic databases failed to identify any NBD/LRR genes. A similar search of the *Danio rerio* (zebrafish) database did yield likely NBD/LRR sequences, and the mouse genome has at least as many genes in this family as did humans (J. A. Harton, unpublished observation).

The preponderance of NBD/LRR proteins in plants is due to reliance on individual effector molecules for recognizing pathogen-specific products. Higher order eukaryotes have developed a highly complex adaptive immune system driving a staggering array of protein-specific immune responses with a limited number of genes.

N-terminal variation in the known and predicted genes suggests a subdivision of CATERPILLER proteins: group I, CARD-containing (e.g., Nod1); group II, pyrin-containing (e.g., DEFCAP); group III, trans activation domain (e.g., CIITA); and unknown (e.g., 16.1) (see Table I). However, these grouping may be oversimplified. For example, multiple cell type-specific forms of CIITA are known. The dendritic cell form has a CARD-like N terminus followed by the activation domain, although no caspase recruitment activity has been described (36). It is of interest that Nod2 and cryopyrin are also expressed as multiple transcripts (4, 13). Whether these different transcripts code for proteins of somewhat different function is clearly of interest. Additionally, self-association has also been demonstrated for CIITA and Nod1, whereas heterodimerization of CIAS1 with apoptotic protein ASC may involve the pyrin domain of CIAS1 (5, 10, 12). Self- and heteroassociation might amplify and generate diversity necessary to mediate appropriate responses.

### Note added in proof

During the review of this manuscript a report describing the initial characterization of Pypaf7, which we refer to as 19.3/Monarch-1, was published. 2002, *J. Biol. Chem.* 277:29874.

---

References