Genes Encoding Three New Members of the Leukocyte Antigen 6 Superfamily and a Novel Member of Ig Superfamily, Together with Genes Encoding the Regulatory Nuclear Chloride Ion Channel Protein (hRNCC) and a $N^\omega$-$N^\omega$-Dimethylarginine Dimethylaminohydrolase Homologue, Are Found in a 30-kb Segment of the MHC Class III Region

Gloria Ribas, Matt Neville, Joanne L. Wixon, Jianhua Cheng and R. Duncan Campbell

*J Immunol* 1999; 163:278-287; 
http://www.jimmunol.org/content/163/1/278
Genes Encoding Three New Members of the Leukocyte Antigen 6 Superfamily and a Novel Member of Ig Superfamily, Together with Genes Encoding the Regulatory Nuclear Chloride Ion Channel Protein (hRNCC) and a $N^\omega$-$N^\omega$-Dimethylarginine Dimethylaminohydrolase Homologue, Are Found in a 30-kb Segment of the MHC Class III Region\textsuperscript{1,2}

Gloria Ribas,\textsuperscript{3*} Matt Neville,* Joanne L. Wixon,\textsuperscript{4*} Jianhua Cheng,* and R. Duncan Campbell\textsuperscript{5*}\textsuperscript{†}

Many of the genes in the class III region of the human MHC encode proteins involved in the immune and inflammatory responses. We have sequenced a 30-kb segment of the MHC class III region lying between the heat shock protein 70 and TNF genes as part of a program aimed at identifying genes that could be involved in autoimmune disease susceptibility. The sequence analysis has revealed the localization of seven genes, whose precise position and order is cen-G7-G6-G6A-G6B-G6C-G6D-G6E-tel, five of which are fully encoded in the sequence, allowing their genomic structures to be defined. Three of them (G6C, G6D, and G6E) encode putative proteins that belong to the Ly-6 superfamily, known to be GPI-anchored proteins attached to the cell surface. Members of the family are specifically expressed and are important in leukocyte maturation. A fourth gene, G6B, encodes a novel member of the Ig superfamily containing a single Ig V-like domain and a cytoplasmic tail with several signal transduction features. The G6 gene encodes a regulatory nuclear chloride ion channel protein, while the G6A gene encodes a putative homologue of the enzyme $N^\omega$-$N^\omega$-dimethylarginine dimethylaminohydrolase, which is thought to be involved in regulating nitric oxide synthesis. In addition, three microsatellite markers, 9N-I, 82-2, and D6S273 are contained within the sequence, the last two of which have been reported to be strongly associated with the autoimmune disease ankylosing spondylitis. The Journal of Immunology, 1999, 163: 278–287.

L

eukocyte Ag 6 (Ly-6)\textsuperscript{6} Ags are a group of leukocyte Ags first identified in the mouse that consist of 70–80 aa containing 10 Cys residues (1, 2). One distinctive feature of these Ags is that they are attached to the cell surface by a GPI anchor. The site of attachment is an Asn residue, based on known attachment sites of similar GPI-anchored proteins (1, 3). The GPI anchor is directly involved in signal transduction (4, 5). Members of the Ly-6 family are differentially expressed in several hematopoietic lineages, especially in T lymphocytes, and appear to function in signal transduction and cell activation (1, 3, 6–8).

Chromosome mapping studies using somatic cell hybrids, recombiant inbred mouse lines, and in situ hybridization have revealed that all members of the mouse Ly-6 multigene family occupy a single genetic locus on chromosome 15 that is closely linked to the sis and myc proto-oncogenes and to loci that mediate susceptibility to radiation-induced lymphoid malignancy (9–11). The physical map spans 630 kb and contains at least 18 distinct Ly-6-related sequences (12). Six expressed Ly-6 mouse genes are known (Ly-6A/E,-B,-C, and -G; TSA-1/Sca-2; and ThB) and a seventh, Ly-6F, has been identified at the molecular level (3, 7, 13–18). The sequences of the leader peptide and membrane-anchoring portions show high similarities, but the sequences of the middle segments apart from the invariant cysteine residues are much more variable. The middle segments probably play an important role in the functional diversity that may distinguish the Ly-6 gene products. To date, three human members of this family, E48 (homologue of mouse ThB), TSA-1/Sca-2 (homologue of mouse TSA-1), and GML, have been mapped in chromosome 8 (8q24-pter region) (19–21), which is syntenic to mouse chromosome 15. There are other Ly-6-related molecules with diverse functions, such as the three domain urokinase plasminogen activator receptor.

\textsuperscript{1} This work was supported by a Formacion Personal Investigador postdoctoral fellowship from the Spanish Government (to G.R.) and a Medical Research Council studentship (to J.L.W.).

\textsuperscript{2} Address correspondence and reprint requests to Dr. R. Duncan Campbell, HGMP Resource Centre, Hinxton, Cambridge, United Kingdom.

\textsuperscript{3} Current address: HGMP Resource Centre, Hinxton, Cambridge, U.K. CB10 1SB.

\textsuperscript{4} Current address: Department of Biomolecular Sciences, The Mill Building, University of Manchester Institute of Science and Technology, P.O. Box 88, Manchester, U.K. M60 1QD.

\textsuperscript{5} Address correspondence and reprint requests to Dr. R. Duncan Campbell, HGMP Resource Centre, Hinxton, Cambridge, U.K. CB10 1SB. E-mail address: rcampbel@hgmp.mrc.ac.uk

\textsuperscript{*} Abbreviations used in this paper: Ly-6, leukocyte Ag 6; Tsa-1/Sca-2, thymic shared Ag-1/stem cell Ag-2; ThB, thymocyte B Ag; E48, monoclonal antibody E48; GML, GPI-anchored molecule-like protein; RA, rheumatoid arthritis; AS, ankylosing spondylitis; DDAll, enzyme $N^\omega$,$N^\omega$-dimethylarginine dimethylaminohydrolase; uPAR, urokinase plasminogen activator receptor; SH2–3, Src homology 2–3; ADMA, asymmetric dimethyl-L-arginine; EST, expressed sequence tag.

\textsuperscript{†} Medical Research Council Immunocchemistry Unit, Department of Biochemistry, Oxford University, Oxford, United Kingdom; and HGMP Resource Centre, Hinxton, Cambridge, United Kingdom.

Received for publication February 2, 1999. Accepted for publication April 14, 1999.
The Journal of Immunology

Table I. Transcript-specific RT-PCR primers

<table>
<thead>
<tr>
<th>Genes</th>
<th>Primers</th>
<th>TM</th>
<th>G + C (%)</th>
<th>Position*</th>
<th>Genomic Length (kb)</th>
<th>cDNA Length² (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6A</td>
<td>F 5’-TGTCTGGGAGTATAACTGAGG 3’</td>
<td>60.2</td>
<td>55</td>
<td>12397</td>
<td>0.75</td>
<td>356</td>
</tr>
<tr>
<td></td>
<td>R 5’-GACTGCGCACTGTGGAGAC 3’</td>
<td>60.5</td>
<td>63</td>
<td>13153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G6B</td>
<td>F 5’-CTCCTTACTGCGCTTCTG 3’</td>
<td>60.2</td>
<td>58</td>
<td>16615</td>
<td>1.10</td>
<td>575</td>
</tr>
<tr>
<td></td>
<td>R 5’-GAGGAGCTGCTCCCTTCACTCC 3’</td>
<td>60.0</td>
<td>57</td>
<td>17735</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G6C</td>
<td>F 5’-GCTGTACCTGCTGCAACG 3’</td>
<td>60.1</td>
<td>60</td>
<td>21221</td>
<td>1.01</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>R 5’-GTTCAGGTTCGTTGCTGTC 3’</td>
<td>58.0</td>
<td>60</td>
<td>22235</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G6D</td>
<td>F 5’-TGGCAAGGAGCACTGGTAC 3’</td>
<td>60.0</td>
<td>58</td>
<td>23629</td>
<td>2.2</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>R 5’-AGCAATTCCTGCTAAAGGCC 3’</td>
<td>58.0</td>
<td>60</td>
<td>25873</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G6E</td>
<td>F 5’-TCACGTGCCAGACACTCTG 3’</td>
<td>62.4</td>
<td>60</td>
<td>27608</td>
<td>1.1</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>R 5’-CCCTGACATCTTCTCTG 3’</td>
<td>61.9</td>
<td>61</td>
<td>28720</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actin</td>
<td>F 5’-CCTCCAGCATCTCCTCTGC 3’</td>
<td>60.0</td>
<td>72</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 5’-TGTTGTGAAGCCTGTAAGG 3’</td>
<td>60.0</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Positions of primers (5’) within the genomic sequence generated in this study.

² PCR conditions: 1 mM MgCl₂, 60°C annealing time, 30 cycles.

activator receptor (uPAR) or CD87, and CD59 (complement protein-blocking assembly of the membrane attack complex) that exhibit a low degree of sequence identity and are encoded by genes that are not in close linkage with any of the Ly-6 family.

In addition to the three new members of the Ly-6 superfamily, the cosmid was found to contain a gene (G6B) encoding a putative molecule of nuclear channel protein. The G6B gene encodes a regulatory molecule of a nuclear channel protein, a new member of the Ig superfamily, and two other genes. The G6B gene encodes the microsatellite marker D6S273 and 82-2. In a genome-wide screen for susceptibility loci in AS, the data obtained by Brown et al. (45) clearly indicated that these two

markers, which achieved the highest LOD scores of all markers tested, were strongly associated with susceptibility to the disease.

Materials and Methods

Sequencing strategy

The EL3 cosmid clones were sequenced using an M13 shotgun strategy (46) with fluorescent dye primer and dye terminator sequencing chemistries (Amersham, Arlington Heights, IL). Cosmid DNA was sonicated, and fragments from 1–1.5 kb were selected for cloning into M13 mp18. Recombinant M13 mp18 phage DNA was purified from culture supernatants using a Vistra DNA Labstation and was cycle sequenced using Thermo-Seq enzymes and a 96-well format on a Hybaid Omnigene thermocycler (Middlesex, U.K.; 95°C for 5 min followed by 20 cycles of 95°C for 30 s and 60°C for 30 s) in the presence of the fluorescent dye-labeled M13 universal primer (5’-TGACCGGCAGCAAATG-3’). The sequencing reactions were run on an Applied Biosystems 377 automated DNA sequence generator (Foster City, CA), and sequence data were analyzed with the Applied Biosystems 377-dedicated software. Individual sequence traces were processed and reassembled using the programs PREGAP and GAP v4.0/B4 from the Staden suite of software (Medical Research Council Laboratory of Molecular Biology, Cambridge, UK).

Ambiguities within the sequence were resolved, and the sequences across areas of single orientation read were confirmed with dye terminator sequencing chemistries, while gaps between contigs were closed by either sequencing the reverse strand of long clones (>800 nt) that extended into the gap or by sequencing of PCR products covering the gaps.

Transcript profiling

The expression of transcripts was investigated in the cell lines Jurkat 6 and Molt 4 (T cell), Raji (B cell), HeLa (epithelial), HT1080 (epithelial), and SW620 (adenocarcinoma) with RT-PCR using total RNA and the Promega RT system, according to the manufacturer’s protocol (gene-specific primers are shown in Table I). PCR primers were designed to give products containing more than one exon so that amplification products arising from genomic DNA contamination were easily discernible. The first-round cDNA synthesis was performed in a final volume of 20 μl with 1 μg of total RNA; 5 μl of this reaction mix was used per 25 μl of PCR reaction using the transcript-specific primers and amplification conditions listed in Table I. Each transcript-specific RT-PCR reaction was performed in at least triplicate to allow for any variation between reactions. Control amplification reactions with primers derived from β-actin were conducted for each first-round cDNA synthesis reaction. The identities of PCR products were confirmed by direct dye terminator sequencing.

Sequence analysis

The Wisconsin Package version 9-UNIX (GCG), maintained at the University of Oxford Molecular Biology Data Center, was used for the majority of the sequence analysis and database interrogation. The DNA sequence generated was screened against the EMBL, SwissProt, PDB, EMBL-EST, and TIGR-EST (search@hcd.tigr.org) databases to position...
known genes and identify possible new coding regions. Repetitive ele-
ments were identified with the aid of the Repeat Masker server (A. F. A. Smir and P. Green, RepeatMasker at http://ftp.genome.washington.edu/ RH/RepeatMasker.html), and potential coding regions were defined using the NIX exon prediction program (http://www.hgmp.mrc.ac.uk/registered/ Webapp/nix/) at the HGMP Resource Center, Hinxton (Cambridge, U.K.). Predictions of protein secondary structure, solvent accessibility, and trans-
membrane regions were conducted using the Jpred-consensus secondary
structure prediction server (http://circinus.ebi.ac.uk:8081/) or PredictPro-
tein program (phd@EMBL-Heidelberg.de). The GCG program SIG-
CLEAVE and the SMART (Simple Modular Architecture Research Tool)
server (http://coot.embl-heidelberg.de/SMART) were used to identify lead-
peptides. Sequence motifs and protein domains were identified using a
combination of the GCG program MOTIF, the Prosite Profilescan server
(http://lrec3.unil.ch/software/PFSCAN-form.html), and the SMART
server. Multiple alignments of amino acid sequences were performed using
Clustalx software, making use of protein structure information from se-
quences within the PDB database to aid in the alignment wherever possi-
ble. Alignments were hand-edited using the GCG9 SeqLab multiple align-
ment editor.

cDNA selection

Five nanograms of cosmid F9N DNA (overlapping the centromeric part
of EL3), obtained by the alkaline lysis method, was biotinylated with 16-
dUTP (20 μM) and radioactively labeled with [α-32P]dCTP (0.06 μM)
using the nick translation kit from Promega (Madison, WI) (47). A fetal
brain cDNA library (gift from M. Lovett, Department of Biochemistry,
University of Texas Southwestern Medical Center, TX) was supplied li-
gated to a linker, to facilitate amplification using primer 3'-CTCGAGAAT
TCTGGATCCTC. The reaction was heated to 94°C for 10 min and then
subjected to 30 cycles of 45 s at 94°C, 45 s at 58°C, and then 2 min at 72°C
before a final stage at 72°C for 10 min. The PCR products were ethanol
precipitated and resuspended in 7.5 μl of distilled water. Two micrograms
of the selected cDNA was mixed with 2 μg of COT-1 DNA and 10 ng of
BamHI-digested Lorist 4 vector in a total volume of 10 μl. This mixture
was boiled for 5 min, 10 μl of warmed 2.5% cDNA hybridization solution
(50% deionized formamide, 1 M NaCl, 50 mM Tris-HCl (pH 7.4), 0.2%
BSA, 0.2% Ficoll 400, 0.2% polyvinylpyrrolidone, 0.1% sodium pyro-
phosphate, 10% dextran sulfate, and 0.5% SDS) was added, and the cDNA
sample was preannealed at 65°C for 4 h. The preannealed starting cDNAs
were mixed with denatured template and 2× cDNA hybridization solution
in a final volume of 20 μl. This mixture was then incubated at 65°C for
50 h. After recovery of the template and elution of the selected cDNAs, the
selected cDNAs were preannealed and hybridized to the remaining tem-
plate in the same way as the starting cDNA. cDNAs were recovered from
the hybridization mixes using streptavidin M-280 beads (Dynal, Great
Neck, NY). Ten-microliter beads were washed in 10 μl of binding buffer
(supplied by the manufacturer) before mixing with the hybridization mix
for 15 min. The beads were then washed twice with 1× SSC/0.1% (w/v)
SDS and three times for 15 min in 0.1× SSC/0.1% (w/v) SDS at 65°C to
remove nonspecifically bound cDNAs. The selected cDNAs were eluted
with 1 M Tris-HCl (pH 7.5).

The secondary selected cDNA products were amplified with linker
primers with adaptors containing dump bases to allow UdG cloning using
the Cloneamp kit (Life Technologies, Grand Island, NY). PCR products
(50–100 ng) were mixed with 25 ng of prepared pAMP 18 vector in the
presence of 1 U of uracil DNA glycosylase in UdG annealing buffer. The
ligation products were transformed into competent TG1 cells plated out
onto L plates (with ampicillin) at low density, and the cDNA inserts were
characterized by either dye primer or dye terminator sequencing using the
M13–21 forward primer.

Results

The complete genomic sequence of the cosmid cosEL3 was ob-
tained from a total of 700 templates. This sequence can be found in
the EMBL database under accession number AJ012008/ HSA012008. Analysis of the sequence was performed by database
searching using BLAST and NIX (@hgmp.mrc.ac.uk). A sche-
matic representation of the exons identified, the repeat elements, and
the microsatellites present in the cosmid is shown in Fig. 1. The
location of exon coordinates and other sequences, such as
polyadenylation signals, can be found in the EMBL database,
submission HSA012008. The derived amino acid sequences of
the five proteins encoded by genes located in the cosmid are
shown in Fig. 2.

The Ly-6-like genes

The G6C gene. Six different human and three mouse ESTs from
the EST database aligned over the G6C genomic region. Only one
of six human EST entries (H03135) matched exactly the G6C
gene, while the others are only partial transcripts containing the
last exon (accession nos. H03945, W56634, R27318, W56597, and
R25237). One of three mouse EST entries (W12301) is almost a
complete transcript, while the other two contain only exons 1 and
2 (accession nos. AA93044 and AA500454). The G6C gene spans
about 3 kb and contains three exons. RT-PCR analysis using

FIGURE 1. Location of cosEL3 in the MHC class III
region. ~150 kb centromeric of the TNF locus. Open
boxes indicate the boundaries of the genes. Differently
filled boxes indicate the exons of the genes found in
cosEL3. Black boxes represent repeat elements. Horiz-
ontal arrows show the orientation of the genes, and
vertical double arrows show the locations of the three
microsatellites detected in cosEL3.
primer sequences located in exons 2 and 3 has revealed expression of G6C only in the T cell lines Molt 4 and Jurkat 6 (Fig. 3). Sequencing of the PCR product has allowed the exon boundaries to be confirmed.

The G6D gene. Only one mouse cDNA (accession no. AA794551) was found in the EST databases matching the internal exons of G6D. The gene spans about 3.5 kb and contains only three exons in reverse orientation to those of G6C. The microsatellite marker D6S273 lies in between the second and third exons very close to a highly repetitive region of about 1.5 kb in length (see Fig. 1). This microsatellite has recently been found to be strongly associated with susceptibility to AS (45). RT-PCR analysis using primer sequences located in exons 2 and 3 has revealed expression of G6D transcript is expressed in all the cell lines analyzed (Fig. 3). Sequencing of the PCR product has allowed confirmation of the exon boundaries.

The G6E gene. No EST sequence matching G6E was found in EST database searches. Although the gene appears to be incomplete, the first two exons encode the putative signal peptide, and the first 24 aa of the Ly-6 domain containing the first four cysteines. A stop codon was found at the end of the second predicted exon, which might indicate that G6E is a pseudogene. In support of this, G6E was found not to be expressed in any of the cell lines tested, whether activated or not by IFN-γ and PMA (Fig. 3), even though two different primer pairs were used.

The G6C/D/E proteins

The G6C and G6D genes encode polypeptides of 125 and 133 aa, respectively. G6E is an incomplete gene and encodes a partial polypeptide sequence of only 51 aa. The G6C and G6D polypeptides each contain a leader peptide, a Ly-6-like domain, a transmembrane region (type 1a), and sequences that indicate they could be anchored to the cell membrane via GPI anchors. Cleavage of the putative signal peptide and the hydrophobic transmembrane region after the Asn residue next to the last Cys residue would yield mature proteins of about 79 – 85 aa in length. Alignment with all protein members of the Ly-6 superfamily is shown in Fig. 4 together with the predicted secondary structural elements (indicated by arrows (β-sheets), cylinder (α-helix), and lines (loops)). The percent similarity and identity between the different members of the Ly-6 family are shown in Table II. G6D lacks two main cysteines obstructing two different disulfide bonds. This is rare in the family because the 10 cysteines are vital to maintain the structure. However, one other member of this family is found to have two cysteines missing (uPAR1, see alignment in Fig. 4). Phylogenetic analysis of the Ly-6 alignment is shown in Fig. 5, which reveals that both G6C and G6D belong to the Ly-6 superfamily, although they are differently related to mouse Ly-6 proteins.

The G6B gene

The G6B gene, which spans about 2 kb, has been predicted by several exon prediction programs in reverse orientation to the G6A
and G6C genes and contains six exons tightly packed between two sets of repetitive elements (Fig. 1). A cDNA selection experiment using fetal brain cDNA as template identified one clone (3/1B7), which contained the end of exon 2, exons 3 and 4, and the beginning of exon 5. One EST entry (accession no. AA699838) was also found containing only the last exon with the predicted polyadenylation signal. We were unable to determine any mRNA expression when performing RT-PCR analysis on a number of cell lines, both unactivated and activated.

The predicted G6B protein has 241 aa and contains several interesting features corresponding to a putative signal transduction receptor. The translated sequence immediately downstream from the initiator methionine codon shows a high proportion of hydrophobic amino acid residues, characteristic of a posttranslationally cleaved leader sequence from position 1–18. Exon 2 encodes an external domain with high similarity to Ig variable domains and contains four cysteines, two of them in good consensus sequences to interact in a disulfide bond, while the other two might interact with another molecule and induce dimerization, a feature common in Ig receptors. G6B also contains two key residues that form a salt bridge, Arg85 in strand D with Asp102 in strand F. Alignment with some Ig V domains is shown in Fig. 6A (50). The percent similarity and identity between protein sequences shown in the alignment are given in Table II. G6B also appears to have two putative Asn glycosylation sites at amino acid positions 77 (NQTN) and 88 (NTTC).

Another hydrophobic region predicted as a transmembrane region (PSORT II, Nakai, PHD) is positioned in the translated sequence from aa 143 to 165. Residues 168–175 correspond to a putative signal transduction domain, containing a proline-rich sequence motif with a consensus sequence X-X-Pro-Pro-X-Pro-X-X for a SH3 binding domain (reviewed by Cohen et al. (51) and Yu et al. (52)) (Fig. 6B), and two putative phosphorylated tyrosine motifs (53). One of them seems to have the consensus sequence for an SH2 binding domain (Y-hydrophobic-X-hydrophobic: YADL).

Table II. Similarity (S) and identity (I) between members of the Ly-6 superfamily and G6C/D, and G6B and Ig V domains

<table>
<thead>
<tr>
<th></th>
<th>G6C</th>
<th>G6D</th>
<th>G6B</th>
</tr>
</thead>
<tbody>
<tr>
<td>S (%)</td>
<td>I (%)</td>
<td>S (%)</td>
<td>I (%)</td>
</tr>
<tr>
<td>G6D</td>
<td>24.36</td>
<td>20.51</td>
<td></td>
</tr>
<tr>
<td>Mouse Ly-6-a</td>
<td>29.73</td>
<td>27.03</td>
<td>25.97</td>
</tr>
<tr>
<td>Mouse Ly-6-e</td>
<td>30.56</td>
<td>25.00</td>
<td>26.67</td>
</tr>
<tr>
<td>Mouse Ly-6-g</td>
<td>24.70</td>
<td>21.90</td>
<td>26.67</td>
</tr>
<tr>
<td>Mouse Ly-6-f</td>
<td>28.38</td>
<td>22.97</td>
<td>23.38</td>
</tr>
<tr>
<td>Human GML</td>
<td>34.62</td>
<td>24.36</td>
<td>34.65</td>
</tr>
<tr>
<td>Human Tsa-1/Sca-2</td>
<td>24.66</td>
<td>21.92</td>
<td>23.68</td>
</tr>
<tr>
<td>Human E48</td>
<td>28.57</td>
<td>25.71</td>
<td>25.00</td>
</tr>
<tr>
<td>Human CD59</td>
<td>34.33</td>
<td>26.87</td>
<td>30.44</td>
</tr>
<tr>
<td>Human uPAR-1</td>
<td>32.39</td>
<td>26.76</td>
<td>25.00</td>
</tr>
<tr>
<td>Human Ig-V (ph1672)</td>
<td>41.5</td>
<td>29.79</td>
<td></td>
</tr>
<tr>
<td>Human Ig-V (U77554)</td>
<td>37.87</td>
<td>25.24</td>
<td></td>
</tr>
<tr>
<td>C. plumbeus (U50610)</td>
<td>37.3</td>
<td>27.66</td>
<td></td>
</tr>
<tr>
<td>Mouse Ig-V (L21019)</td>
<td>40.5</td>
<td>26.19</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 4. Alignment of the extracellular region of G6C (1–79) and G6D (1–82), generated by the Clustalx program, with the corresponding region of the related sequences from the Ly-6 superfamily members are shown. Residues that are common or similar between the different sequences are shown boxed (asterisk indicates that all residues are identical; Prettyplot program). The three Ly-6 modules in uPAR are labeled 1–3 starting from the outermost, N-terminal domain. Positions of β-strands (A–E) are indicated by arrows, and the single α helix is indicated by a cylinder above the sequence. Disulfide bonds between cysteines are shown by lines. The sequences of the following proteins are from SWISSPROT (mouse Ly-6 genes (P05533, P09568, P35460, P35461), human CD59 (P13987), rat CD59 (P27274), GML (Q99445), human uPAR (Q30405), mouse uPAR (P35456), rat uPAR (P49616), Cobrotoxin, Naja siamensis (P01391)) and EMBL (E48 (X82693), ThB (X63782), human Tsa-1/Sca-2 (U56145), mouse Tsa-1/Sca-2 (U04268), chicken Tsa-1/Sca-2 (L34554), rat Ly-6 (M30692, M30689, M30690), SP-10 from Palfree et al. (68), and Squid Sgp2 from Williams et al. (79)).
The G6A gene

The G6A gene contains eight exons that span about 3 kb of DNA (Fig. 1). The 5'-untranslated region of the mRNA is encoded in the first two exons and part of the third (from positions 12,052 to 12,257). When either the first or second exon is used, the splicing of the third exon changes, and it starts at position 12194. The most common ESTs (accession no. of ESTs starting at exon 3: AA298342, AA134375; accession no. of ESTs starting at exon 2: H87910, AA133714; accession no. of ESTs starting at exon 1: AA298342, AA134375, and R11949). The longest open reading frame from all splice variants starts at position 12,258 in exon 3 and codes for a 285-aa polypeptide of 31.4 kDa.

Database searches revealed that the putative human polypeptide shared 96.7% identity with the derived amino acid sequence encoded by the mouse cDNA clone (7u) isolated from a malignant melanoma cDNA library (54), and 48.75% identity (70.46% similarity) with the rat kidney enzyme DDAH (43) (EC 3.5.3.18; Fig. 7). The active center of the enzyme is probably between residues 119 and 166, which are highly conserved (see Fig. 7). Three small motifs common to the prokaryote enzyme family that catabolize L-arginine to L-citruline (55) are found in the DDAH family as well (highlighted in Fig. 7).

The results obtained in the RT-PCR experiments showed that G6A was expressed widely (Fig. 3). However, Spanjaard et al. (54) only detected expression of the mouse mRNA in large intestine, with low levels in lung and brain. In rat, expression was detected in kidney, pancreas, liver, brain, lung, and heart (56), and the protein was clearly shown to exist in peritoneal neutrophils and macrophages (56). Recently, the DDAH has been characterized in bovine brain tissue, and the protein is detected in total brain and kidney, at low levels in heart, liver, and lung, but not in blood, muscle, spleen, or intestine extracts (57). The PCR product obtained from the RT-PCR experiment has been sequenced, and the exon boundaries checked.

The G6B gene

The gene G6 was originally defined using a 2.4-kb HindIII genomic fragment lying adjacent to a CpG island, which hybridized to mRNA species of 1.5 and 1.4 kb in the cell lines U937, U937 stimulated by PMA, HepG2, and Molt4, and mRNA species of 1.5 and 1.45 kb in the cell line Raji (31). A fragment containing the 5'-untranslated sequence in exon 2 was only detected in one of the mRNA species (58). Two mRNA species have also been described by Valenzuela et al. (42) (1.0 and 1.2 kb) that are differentially regulated by PMA, retinoic acid, IFN-γ, and IL-2. They also detected a difference in the 5'-untranslated region between the two species. The gene spans 6 kb and contains seven exons (Fig. 1). The microsatellite marker 82-2 which has been found to be strongly associated with susceptibility to AS (45), is located between the second and third exons. The longest open reading frame of G6 starts in exon 2 and codes for a polypeptide of 241 aa with a predicted Mr of 26.9 kDa. The amino acid sequence is identical with the polypeptide described by Valenzuela et al. (42), which they claimed could be a putative nuclear chloride ion channel protein.

Two PEST regions have been found in G6 (wwwserver at JMB Jena). These regions are rich in proline (P), glutamic acid (E), serine (S), and threonine (T) and begin and end with relatively charged residues, but most commonly lysine, arginine, and histidine. Those regions are characteristic of proteins that are rapidly degraded within eukaryotic cells (59). PEST sequences have been reported in a number of proteins that are located in the nucleus, including c-Fos, c-Myc, p53, Bmi1, and Upil (59–61).

Interestingly there are two basic sequences (KRR and KKYR) at positions 49 and 192 in the protein that satisfy the criteria for nuclear localization signals (62). Support of this, there are four putative casein kinase II sites near these nuclear localization signals at positions 44 (TTUD), 174 (TLAD), 198 (TIPE), and 222 (TCPD) (63). Casein kinase II sites have been located close to putative casein kinase II sites near these nuclear localization signals in many confirmed nuclear localization signal sequences found in other proteins, and phosphorylation at these sites has been implicated in the rate of nuclear transport. Other post-translation modifications that could occur in G6 are shown in Fig. 2.

No N-terminal signal peptide is found in the G6 protein sequence (PsortII and Nakai server). None of the programs used (TM prediction, PHD) predicted a good transmembrane region in the protein, suggesting a globular three-dimensional structure for G6.

Discussion

Ly-6-like genes

The genes G6C and G6D, and possibly also G6E, encode small Cys-rich proteins of 79–82 aa that appear to be novel members of the Ly-6 family. They are probably anchored in the cell membrane by a C-terminal GPI moiety, a post-translational modification in common with each member of the Ly-6 family described to date.
The finding of Ly-6-like proteins in humans is important because of the independent regulation of each member of this family, generating distinct patterns of expression during hemopoiesis and immune responses. This information suggests that they provide distinct functions affecting various stages of leukocyte development.

Only a few human genes with this homology have been described to date, the E48 homologue (19) to the mouse ThB gene (64), the Tsa-1/Sca-2 homologue (20) (or 98604 (65) or RIG-E (66)) to the mouse Tsa-1/Sca-2 gene (6), and the GML (21). The E48 (involved in cell-cell adhesion), Tsa-1/Sca-2 (recently, physically and functionally associated with the TCR via CD3 (67)), and GML (the expression of which is induced by p53) genes have been localized in human chromosome 8 (19–21), which is the syntenic region of mouse chromosome 15, where the mouse Ly-6 gene cluster has been localized (12). No Ly-6 domain has been found in combination with domains of any other superfamily; this may be because the exon structures known for this superfamily are not suited to exon shuffling. The SP-10 (sperm acrosomal protein) is another human member of the Ly-6 superfamily that is encoded by a gene located in chromosome 11q23-q24 (68). However, it contains some unique features that distinguish it from other members of the superfamily, including G6C and G6D. For example, more than half of the original protein sequence is unrelated to Ly-6, being encoded by a completely different exon not present in any other Ly-6 member, and it lacks the GPI anchor, which is a distinctive feature of the Ly-6 Ags.

The G6B gene

G6B is an Ig-like receptor with different signal transduction motifs found in the putative cytoplasmic domain of the molecule. The Ig superfamily is a large group of related proteins that function, ..
mainly in the immune system, in cell-cell recognition or in the structural organization and regulation of muscle (49). The various members of the superfamily are built of homologous domains of approximately 100 residues and with a structure formed by two β-sheets packed face to face. A disulfide bond is almost always formed by cysteines localized in strands B and F, which are one of the most conserved in the Ig domain. Some of the members have more than two cysteines, and it is believed that this contributes to dimerization. Individual members of the superfamily can differ in the number and size of strands in the two β-sheets and in the size and conformation of the links between the strands. Similarities in sequences and structures were found, however, and these allow the different members to be grouped into what Williams and Barclay (49) call sets. V set and V-related domains have about 65–75 aa residues between the conserved disulfide bond, and there are four β-strands in each β-sheet plus a short β-strand segment across the top of the domain. G6B is more related to the V set, as it contains 71 aa in between the two cysteines most likely to form the disulfide bond.

Harpaz and Chothia (50) defined a new set called I defined by 20 key amino acid positions that form the characteristic folds of the set. Ig molecules that follow set I seem to be involved in the cell adhesion and cell surface recognition processes. Some of the previous V and C2 sets fall into this new set I. Alignment of G6B with some other variable domains in a manner similar to that used by Harpaz and Chothia highlighted the conservation of 18 of the 20 key sites in G6B. This alignment is based on a secondary structure prediction that defined the approximate position of all the β-strands. Other features that indicate an Ig domain are that the majority are encoded within one single exon, which is always phase 1. Both features are found in the predicted variable domain of G6B, providing more evidence for its inclusion in the superfamily. Two residues important for the formation of a salt bridge are also conserved, the arginine residue in strand D with aspartic acid in strand F. Of the four cysteines found in the predicted sequence of G6B, the two in strands B and F most likely form a disulfide bond. One of the other cysteines is localized after the Ig-like domain and before the helical transmembrane region, which could indicate the possible existence of G6B as a dimer, either disulfide linked to itself or to another polypeptide.

Interestingly, the BG genes within the chicken MHC complex have been found to encode molecules composed of a single extracellular domain that resembles an Ig V-type domain, a transmembrane region, and a 217-aa cytoplasmic domain (69). We have compared G6B with the BG molecules and have found that they share 20% identity (26% similarity). Moreover, the intracellular segment of sequence is 3 times larger in the BG molecules compared with G6B, and they share no similarity whatsoever. Thus, although G6B and the BG molecules are members of the Ig superfamily, the lack of sequence similarity indicates that they are unrelated to each other.

One of the internal motifs corresponds to a short proline-rich sequence of ~10 aa that could bind to a SH3 (Src homology 3) domain (70) in a left-handed PPII helix conformation (a 3-fold symmetry helix, with proline residues usually in one face). PPII helices in globular proteins are also ideal mediators of protein-protein interactions since most of these motifs are located on the surface of their proteins (71).

G6B also contains two putative phosphorylated tyrosine residues with one of them at least in a consensus sequence for a SH2-binding motif. Thus, G6B contains all the possible features to be a signal transduction receptor.

The G6A gene

The G6A protein is a homologue of the enzyme DDAH that is thought to regulate the production of NO by metabolizing Nω-mono-methyl-L-arginine and asymmetric dimethyl-L-arginine to L-citruline, both of them analogues of L-arginine and direct inhibitors of constitutive and inducible NO synthases (44, 72). Although G6A is probably not the human DDAH enzyme itself, it is likely that it could still function in a similar way as DDAH, but have different specificity. Taking into account that Nω-mono-methyl-L-arginine has a therapeutic effect in different inflammatory processes (73–75) and NO production has been found to contribute to the development of insulin-dependent diabetes mellitus (76, 77),

FIGURE 6. Alignment of the G6A amino acid sequence with mouse AA833494 and rat D86041, which corresponds to the enzyme DDAH. Three small motifs in common with the prokaryotic deiminase enzyme family are highlighted with triangles.
examination of the exact function of G6A might add some valuable information to these processes.

Among all arginine-using or -producing enzymes only the di-methyldargininase and arginine deiminase enzyme families contain three motifs in common, which are thought to be involved in functionality or in maintaining their structure (55) (Fig. 7). Both families catalyze the hydrolysis of l-arginine into l-citruline, but the arginine deiminase enzyme is found exclusively in prokaryotes. Interestingly, arginine deiminase is known to be involved in strongly inhibiting cell growth, suppressing IL-2 production and IL-2R expression, and inducing apoptotic cell death in T lymphoblasts (78).

The G6 gene

The genomic localization of NCC27 is in the central part of the MHC class III region and corresponds exactly with the gene described in this paper as G6. NCC27 has been demonstrated to be in the nuclear membrane (42). Although the authors describe the protein as a chloride ion channel itself, they do not reject the possibility of NCC27 acting as a regulatory subunit of a multipeptide chloride ion complex. We think that this second hypothesis might be more appropriate due to the presence of several PEST sequences (seven in bovine p64 and two in G6 or NCC27) that could act as signals for the rapid degradation of the protein, and the clear lack of a transmembrane region. Under physiological conditions, nuclear localization must be subject to complicated regulatory mechanisms, since the presence of certain proteins in the nucleus is required only at very specific moments in the cell cycle or only in response to short-lived stimuli.

It is apparent that many of the genes located in the MHC class III region are good candidate genes for susceptibility to diseases such as insulin-dependent diabetes mellitus or AS due to the possibility of their involvement in the immune and/or inflammatory mechanisms, since the presence of certain proteins in the nucleus is required only at very specific moments in the cell cycle or only in response to short-lived stimuli.

References
