ELSEVIER

Indels and Imperfect Duplication Have Driven the Evolution of Human Complement Receptor 1 (*CR1*) and *CR1*-Like From Their Precursor *CR1* Alpha: Importance of Functional Sets

C.A. McLure, J.F. Williamson, B.J. Stewart, P.J. Keating, and R.L. Dawkins

ABSTRACT: This study examines the effects of duplication and insertions-deletions (indels) by comparing human complement receptor 1 (*CR1*) and human CR1-like (*CR1L*) with syntenic genes from four other vertebrates (chimpanzee, baboon, rat, and mouse). By phylogenetic analysis, the domains of these genes can be classified into 10 distinct subfamilies (*a, b, c, d, e, f,* g^{-like}*, h, j,* and *k*), which have been largely conserved throughout vertebrate and invertebrate evolution. In spite of many complex and diverse duplications and indels, the subfamily order of domains (*a, j, e, f, b, k, d,* g^{-like}) has been maintained. The number of domain sets has increased progressively, thereby expanding the functional repertoire. *Human Immunology 66, 258–273 (2005).* © American Society for Histocompatibility and Immunogenetics, 2005. Published by Elsevier Inc.

KEYWORDS: complement control protein; *CR1; CR1-L;* evolution; duplication; insertion and deletion

ABBREVIATIONS

CCP	complement control protein
CR	complement receptor
Hosa	*Homo sapiens*
LHR	long homologous repeats
MCP	membrane cofactor protein
Mumu	*Mus musculus*
Pacy	*Papio cynocephalus*
Paha	*Papio hamadryas*
Patr	*Pan troglodytes*
Rano	*Rattus norvegicus*
RCA	regulators of complement activation
SCR	short consensus repeat
WGS	whole genomic shotgun

INTRODUCTION

Of all the mechanisms implicated in the generation of diversity, duplication and indels are now recognized to be of major importance. For example, there is a clear relationship between copy number and polymorphism when genes within the major histocompatibility complex (MHC) are compared [1]. The role of indels in generating diversity is less well known, but many examples are now known [2–5].

With the intention of establishing whether these examples are representative of the genome generally, other regions should be studied. One excellent possibility is that part of 1q32 known as the regulators of complement activation (RCA). This gene cluster contains numerous complement control proteins (CCPs) that are characterized by domains known as short consensus repeats (SCRs). Previous analyses of this cluster have revealed that duplications have been varied and complex [6, 7]. The SCRs can be regarded as the minimal units and have been duplicated individually and in combination [8]. The resulting structures have been further modified by insertion between and within SCRs.

From the Centre for Molecular Immunology and Instrumentation (C.A.M., J.F.W., B.J.S., R.L.D.), University of Western Australia, Nedlands, and C.Y. O'Connor ERADE Village (C.A.M., J.F.W., B.J.S., P.J.K., R.L.D.), Canning Vale, Western Australia.

Address reprint requests to: Dr. Roger Dawkins, Centre for Molecular Immunology and Instrumentation, University of Western Australia, PO Box 5100, Canning Vale South, Western Australia 6155; Fax: +618 9397 1559; E-mail: cmii@cyllene.uwa.edu.au.

Received November 8, 2004; revised December 27, 2004; accepted January 3, 2005.

0198-8859/05/$–see front matter
doi:10.1016/j.humimm.2005.01.004

Previous phylogenetic analysis of the protein [6] and the genomic [9] sequences has revealed that the 42 SCRs of Hosa complement receptor 1 (*CR1*) can be classified into 10 distinct subfamilies (*a, b, c, d, e, f, g*$^{-like}$*, h, j,* and *k*). The combination and order of these subfamilies relate to function [10]. For example, the *ajef* set is involved in ligand recognition and the *ch* set in membrane attachment.

Here we demonstrate that duplication and insertion have driven the evolution of the *Homo sapiens* (Hosa) *CR1* and *CR1L* genes. Comparison of these genes with syntenic genes and genomic regions of the chimpanzee (*Pan troglodytes;* Patr), Hamadryas baboon (*Papio hamadryas;* Paha), Norway rat (*Rattus norvegicus;* Rano), and house mouse (*Mus musculus;* Mumu) reveals that the order of SCRs is largely conserved but the number within each species varies greatly. Thus, Mumu *Crry* contains five SCRs and is approximately 35 kb, whereas Hosa *CR1* is over 140 kb. The genomic region in which these genes are located (between *CR2* and *CD34*) is 180 kb in Mumu and 430 kb in Hosa. These major differences can be accounted for by duplication, insertion, and deletion.

We propose a model for the evolution of human *CR1* and human *CR1L*.

MATERIALS AND METHODS

Rodent *Crry* and Primate *CR1* and *CR1L* Sequence Analysis

The following amino acid and nucleotide sequences were analyzed and compared in this study: a Hosa genomic sequence containing *CR1, MCPL, CR1L,* and *MCP* at 1q32 was taken from the NCBI database (http://www.ncbi.nlm.nih.gov/) (position 1124945–1449694 on contig NT_021877.16 [gi:37539616]; accession nos. AL691452.10, AL137789.11, AL365178.10, and AL035209.1), Patr genomic sequence containing *CR1, MCPL, CR1L,* and *MCP* was taken from Ensembl database (http://www.ensembl.org/Pan_troglodytes/) (position 187628675–187978675 on chromosome 1; contigs: AADA01119442, AADA01119441, AADA01219300, AADA01119440, AADA01201432, AADA01291272, AADA01361130, AADA01361131, AADA01361132, AADA01311109, AADA01218853, AADA01251680, AADA01167980, AADA01185290, AADA01222324, AADA01117941, AADA01189110, AADA01100515, AADA01249515, AADA01157797, AADA01291331, AADA01127022, AADA01098546, AADA01005029, AADA01181018, AADA01214894, AADA01117928) (missing Patr *CR1L* genomic sequence was sourced from the NCBI Whole Genomic Shotgun sequence database [http://www.ncbi.nlm.nih.gov/]; accession numbers gi:39217870, gi:39217871, gi:39217872, and gi:38938087), Mumu genomic sequence containing *Crry* and *MCP* on chromosome 1 was taken from the NCBI database (http://www.ncbi.nlm.nih.gov/; position 3163341–3279341 on contig NT_039190.2, strain C57BL/6J), Rano genomic sequence containing *Crry* and *MCP* on chromosome 13 was taken from the NCBI database (http://www.ncbi.nlm.nih.gov/) (position 2136644–2259333 on contig NW_047404.1), Patr *CR1L* amino acid (AAA50460.1), Paha *CR1* amino acid sequence (AAA62170) and yellow baboon (*Papio cynocephalus;* Pacy) *CR1L* amino acid sequence (AAA99004).

SCR Alignments and Subfamilies

SCRs within genomic sequence were identified by Pfam (http://www.sanger.ac.uk/software/Pfam/dnasearch.shtml). SCR content of amino acid sequences was identified by the defining motif C..C..C..W..C [11, 12].

All SCRs identified were extracted and included in a database as an individual sequence labeled by the species and the gene from which it came and numbered according to its position within the gene (*i.e.,* the first SCR of *CR1* in Hamadryas baboon is labeled Paha_*CR1*_1). The partial, incomplete, or degenerate SCRs identified have also been included in this database and assigned a reference label (*i.e.,* Patr_*CR1*_10) but were removed and not included in any alignments or phylogenetic studies. The *g*$^{-like}$ SCRs have been labeled as in [9], so as not to confuse the existing numbering, for example Hosa_*CR1*_SCR7–8 lies between Hosa_*CR1*_SCR7 and Hosa_*CR1*_SCR8.

The sequences were aligned by ClustalW (http://www.es.embnet.org/cgi-bin/clustalw.cgi) and adjustments made on the basis of the conservation of the conserved motif C..C..C..W..C and previous SCR alignments of subfamilies, described in [6]. Phylogenetic and molecular evolutionary analyses were performed by MEGA version 2.1 [13]. Evolutionary distances were calculated by the gamma distance model, which accounts for multiple amino acid substitutions and variation of substitution rate among sites. A gamma shape parameter of 0.93 was used on the basis of previous studies of SCRs [14]. Phylogenetic trees were constructed by the neighbor-joining method.

Subfamilies established through phylogenetic analysis were assigned colors, whereby *a* = red, *j* = orange, *e* = yellow, *f* = green, *b* = dark blue, *k* = purple, *d* = mauve, *g* = gray, *c* = light blue and *h* = aqua. These are consistent throughout all figures with color. The SCR "Rano_*Crry*_5" was used as a reference outlier for phylogenetic analysis of the relationship within individual subfamilies.

Organization and Use of Subfamilies to Define Functional Sets of *CR1*-Related Genes

By using the phylogenetic analysis to first define subfamilies (Figure 1a(ii)) and second to define associations