

Amino Acid Patterns Within Short Consensus Repeats Define Conserved Duplicons Shared by Genes of the RCA Complex

Craig A. McLure, Roger L. Dawkins, Joseph F. Williamson, Richard A. Davies,* Jemma Berry, Natalie Longman-Jacobsen, Rebecca Laird, Silvana Gaudieri†

Centre for Molecular Immunology and Instrumentation, University of Western Australia, Nedlands, Western Australia 6907

Received: 8 September 2003 / Accepted: 28 January 2004

Abstract. Complement control proteins (CCPs) contain repeated protein domains, short consensus repeats (SCRs), which must be relevant to diverse functions such as complement activation, coagulation, viral binding, fetal implantation, and self–non-self recognition. Although SCRs share some discontinuous and imperfect motifs, there are many variable positions and indels making classification in subfamilies extremely difficult. Using domain-by-domain phylogenetic analysis, we have found that most domains can be classified into only 11 subfamilies, designated *a, b, c, d, e, f, g, h, i, j, or k* and identified by critical residues. Each particular CCP is characterized by the order of representatives of the subfamilies. Human complement receptor 1 (*CRI*) has *aje**fbkd*** repeated four times and followed by *ch*. The classification crosses CCPs and indicates that a particular CCP is a function of the mix of SCRs. The *aje* set is a feature of several CCPs including human *CRI* and *DAF* and murine *Crry* and appears to be associated with the success or failure of implantation *inter alia*. This approach facilitates genomic analysis of available sequences and suggests a framework for the evolution of CCPs. Units of duplication range from

single SCRs, to septamers such as *efbk**daj***, to extensive segments such as *MCP-CRIL*. Imperfections of duplication with subsequent deletion have contributed to diversification.

Key words: Complement control proteins — Duplication — Phylogeny — Genomics — Peptide motifs

Introduction

The complement control protein (CCP) family has been conserved from invertebrates to provide required regulation of complement and other cascades and to enable self–nonself discrimination. The CCPs have evolved in parallel with the complement cascade and can be regarded as the major agents of regulation. During vertebrate evolution the family has expanded such that there are numerous copies clustered throughout the vertebrate and indeed human genome. Apparently functions have diverged and now include viral receptors, regulation of adaptive immunity, and maintenance of fetomaternal tolerance. (Bell 2000; Xu et al. 2000).

The CCP family is defined by the presence of reiterated protein domains known as short consensus repeats (SCRs). An SCR typically comprises 56–70 amino acids, including four cysteines and a tryptophan which are thought to be essential in holding the domain in its rigid triple-loop structure (Schwarzenbacher et al. 1999) which permits the formation of complex proteins

Manuscript number 0014 of the Centre for Molecular Immunology and Instrumentation of the University of Western Australia

Present addresses:

*Department of Biochemistry, University of Western Australia, Nedlands, 6907, Western Australia

†Department of Clinical Immunology, Royal Perth Hospital, Perth, 6000, Western Australia

Correspondence to: Craig A. McLure; email: cmii@cyllene.uwa.edu.au

and multiple binding affinities. Each CCP has a specific number of SCR units at the genomic level, e.g., human (*Homo sapiens*) *CR1* has 30 and *Homo sapiens* decay accelerating factor (*DAF*) has four SCRs. Numerous SCR-containing proteins are encoded within the genomic region known as the regulators of complement activation (RCA) complex on 1q32.

It remains to be determined how such domains could be responsible for such diverse functions, how each CCP achieves a remarkable degree of ligand specificity, and how self–nonself discrimination is mediated. Indeed, it is still uncertain how it will be possible to explain the similarities as well as the differences in function when component SCRs are compared.

Structure–function studies have exploited several different approaches but can be difficult to interpret. For example, the X-ray structure of two SCRs of *CR2* in complex with C3d highlights some of the problems including similarities in sequence of SCRs and therefore cross-reactivity of antibodies and inhibitory peptides (Szakonyi et al. 2001).

Valuable insights have been provided through the use of *in situ* mutagenesis. For example, the Atkinson group (Krych et al. 1998; Krych-Goldberg et al. 1999) has shown that some residues are important in the binding of *C4b* to *CR1*. Interestingly, some of the effects of mutagenesis are unexpected in that the critical mutation is remote from the putative binding site. Furthermore, a major conclusion is that combinations of SCRs are more important than single domains (Kirkitadze et al. 1999), suggesting that the evolutionary unit is a specific set of SCRs. There is a need for alternative approaches to define the sequence characteristics of these sets. Here we illustrate the value of phylogenetic analysis.

From an evolutionary perspective, the CCP family is a very useful model. Undoubtedly, duplication has been the driver of expansion and diversification but the extent of the region duplicated must have ranged from a single domain or less through multiple contiguous domains as fixed sets to genomic segments containing large genes. All of these processes are suggested from a comparison of cDNA sequences and apparent to various degrees from genomic analysis of the data available for the RCA complex (Heine-Suner et al. 1997; Rodriguez de Cordoba et al. 1999; <http://www.ncbi.nlm.nih.gov>).

With the dual aims of elucidating structure–function relationships and defining mechanisms of genomic duplication, we have analyzed genes within the RCA complex and compared the results with those obtained after earlier analyses of duplication within the major histocompatibility complex (MHC) (Dawkins et al. 1999).

We show here that SCRs can be classified into a limited number of groups and the arrangement of these groups relates to function. By analogy, each

SCR is a letter, the order of the letters defines sets (or words), and the number and content of the words create CCPs (or sentences).

We also show that, as in the MHC, it is the imperfection of the duplication processes that contributes to diversity.

Methods

Primate and Rodent RCA Sequences

The relationships between individual SCRs were investigated using the following amino acid and nucleotide sequences: *Homo sapiens CR1* (P17927 and Y00816), *Homo sapiens CR1L* (positions 1278188–1355188 on gi:22041066), *Homo sapiens* complement receptor 2 (*CR2*) (PL0009), *Homo sapiens DAF* (P08174), *Homo sapiens* membrane cofactor protein (*MCP*) (P15529), *Homo sapiens MCPL* (positions 1263188–1273188 on gi:22041066), *Homo sapiens C2* and factor B (*Bf*) (AF019413), chimpanzee (*Pan troglodytes*) *CR1* (AAA51438), *Pan troglodytes DAF* (AAF73177), baboon (*Papio hamadryas*) *CR1* (AAA62170), *Papio hamadryas DAF* (AAF73178), *Papio hamadryas MCP* (AAB66815), house mouse (*Mus musculus*) *CR1* (AAA78271), *Mus musculus* complement regulatory protein (*Crry*) (AAA37468), *Mus musculus DAF* (AAB00092), *Mus musculus MCP* (BAA31859), *Mus musculus CR2* (A43526), and Norway rat (*Rattus norvegicus*) *Crry* (AAA91821).

Strategy

Domain-by-Domain Analysis

Initially, we identified SCRs contained within RCA proteins and other CCPs using Pfam (<http://pfam.wustl.edu/>) and database annotation. Alternative definitions were also used, but in essence the defining motif was simply C..C..W..C (Hourcade et al. 1989; Reid et al. 1986).

Next, all domains were extracted and annotated before being added to a SCR database. The individual sequences could be analyzed without regard to their presence within a particular CCP. The intention was to compare and classify SCRs independently without bias.

Third, various phylogenetic analyses were used to establish relationships between individual SCR sequences. Alignments were created (ClustalW v1.8) and adjustments were evaluated depending upon the purpose of the alignment. For example, the clusters were found to differ by indels as well as residues and it was sometimes necessary to realign on the basis of the C..C..W..C motif. Given the divergence between and within duplicons of CCPs, evolutionary distances were estimated using 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 based on previous studies of SCRs. (Krushkal et al. 2000). Phylogenetic trees were constructed by neighbor-joining and parsimony methods. Each tree underwent 1000 bootstrap replications and was not rooted (Molecular Evolutionary Genetics Analysis version 2.1; Pennsylvania State University 1993).

Peptide Searches

Using the amino acid alignment shown in Fig. 1, selected amino acid motifs from the group consensus sequences were searched against the databases SwissProt, Prosite, and GenBank/DBJ/EMBL using BLAST (<http://www.ncbi.nlm.nih.gov/Blast>) and Motif 2 (<http://motif.genome.ad.jp/MOTIF2.html>).

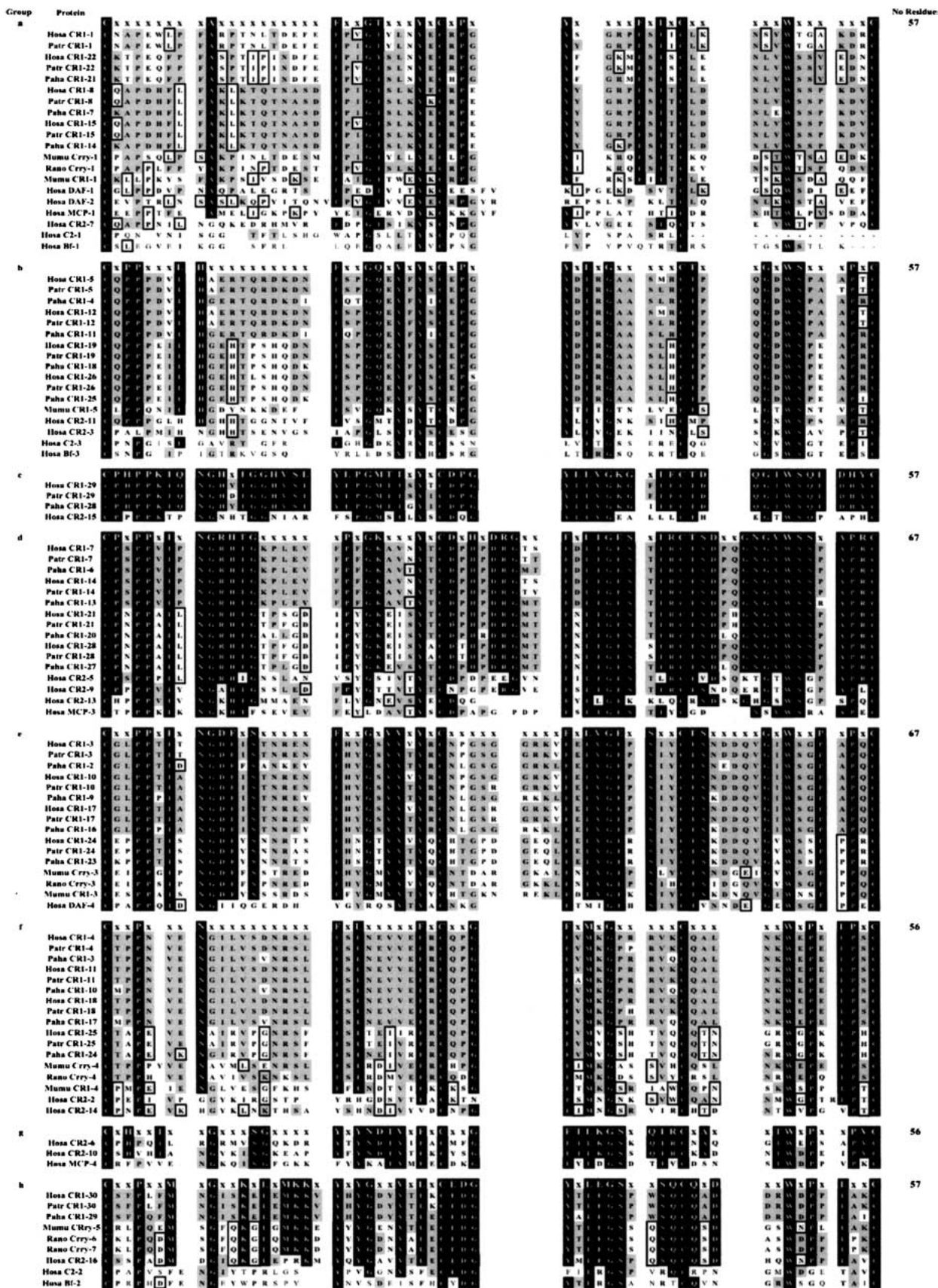


Fig. 1. Legend appears on 146.

long branch lengths and require further analysis, the intention was to develop a series of hypotheses to describe the likelihood of common origins.

So as to define further clusters we then pooled the SCRs of *CR1* and *DAF*; the SCRs fell within the *a*, *j*, and *e* clusters such that *CR1* and *DAF* can be written $(ajefbkd)_4$, xx and $aaaje$, respectively, indicating a common set of *aje* (Table 1).

When SCRs of *CR2* are included, clusters *c*, *g*, *h*, and *i* appear, giving a total of 11 subfamilies. Interestingly, the two unclassified (x) and terminal (29 and 30) SCRs of *CR1* cluster with corresponding SCRs of *CR2*, i.e., *c* contains *CR1* SCR 29 and *CR2* SCR 15 and *h* contains *CR1* SCR 30 and *CR2* SCR 16; *CR1* and *CR2* share a set consisting of *ch*. Thus, as illustrated in Table 1, *Homo sapiens CR1* can be written $(ajefbkd)_4ch$. *CR2* is $jfbkdgaidgbidfch$, which can be rewritten $jf(a|bi|kdf|g)_3ch$, where *a|b*, *i|k*, and *f|g* are introduced to recognize alternatives yielding three tetramers with *a* or *b*, *i* or *k*, *d*, *f* or *g* preceded by *jf* and followed, as in the case of *CR1*, by *ch*. Alternatively *CR2* could be $j(f|ga|bi|kd)_3fch$, where the tetramer is similar in content although different in order.

As shown in Figs. 1 and 2, this analysis suggests that SCR subfamilies have diverged to greater or lesser degrees since the original sets were duplicated. For example, the original tetramer of *CR2* may have contained the progenitors of *a + b*, *i + j + k*, *d + e*, and *f + g*, which is now present in *CR1* as *ajef* and in *CR2* as *a|bi|kdf|g*. Within *CR2*, we deduce that the *d* subfamily has been conserved relative to the other more divergent components of the tetramer. Interestingly, the terminal *ch* set has been retained by both *CR1* and *CR2*.

SCR Subfamilies and Their Sets Are Shared by Different RCA Proteins

With rare exceptions, all SCRs in our database could be classified into these 11 subfamilies (*a* to *k*) or their putative progenitors and all CCPs examined could be described according to the schema used for *CR1* and *CR2* (Table 1). For example, *MCP* is *aj/kdx*, suggesting a common origin with *CR1* and *CR2*, divergence or shuffling yet relative retention of the proximal *aj* set. *CRIL* is particularly interesting. When the available genomic sequence is examined (Fig. 3), we identify extra SCRs and a structure of $a(jefbkd)_2d$, indicating a close but imperfect relationship with *CR1*, viz. $(ajefbkd)_4ch$. Other sequences may well indicate further heterogeneity, but on the basis of the current data, it appears that the internal *jefbk* set was duplicated as a pentameric segment or that deletions occurred after duplication of the *ajefbkd* septamer. Apparently, divergence is ongoing

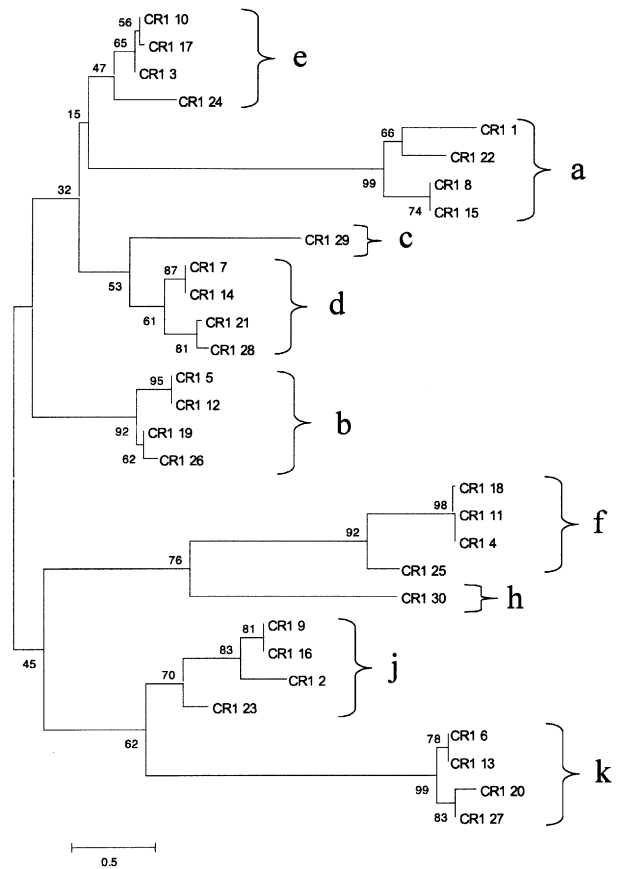


Fig. 2. Amino acid phylogeny of individual domains from Hosa *CR1* reveals seven distinct clusters, each containing four SCRs. *CR1* domains are numbered 1 to 30 (5' to 3'). There are seven distinct clusters "abdefjk" containing every eighth domain, e.g., *a* contains 1, 8, 15, and 22. By contrast, *c* and *h* contain a single member (29 and 30, respectively). The data indicate that *CR1* comprises four septamers, each containing in order *ajefbkd*. The neighbor-joining tree is unrooted and the bar at the bottom of the tree reflects the distance as calculated by the gamma distance model ($\alpha = 0.93$).

since, for example, the introns between *j1* and *j2* differ in length and, especially, the content of insertions indicated by gaps (Figs. 4A and 4B). We therefore asked whether there were other ways of defining the subfamilies.

Substitutions and Indels Define SCR Subfamilies

Once the major clusters were identified by phylogenetic analysis, it became obvious that some residues are characteristic of, or even restricted to, a single cluster. Many of these residues had previously been regarded as alternatives within all SCRs but we prefer the possibility that there are sites which define specific subfamilies. Further evidence for this interpretation is shown in Table 2, where it can be seen that substitutions correlate with coding indels and therefore the number of residues, i.e., the length of the SCR.

Table 1. Conservation of the complement regulatory SCR set *aje* across vertebrates

Species ^a	Protein	No. of SCRs	Internal SCR structure ^b
Hosa	<i>CR1</i>	30	(a, j, e, f, b, k, d) ₄ , c, h
Hosa ^a	<i>CRIL</i>	12	a, (j, e, f, b, k) ₂ , d
Patr	<i>CR1</i>	30	(a, j, e, f, b, k, d) ₄ , c, h
Paha	<i>CR1</i>	29	j, e, f, b, k, d, (a, j, e, f, b, k, d) ₃ , c, h
Mumu	<i>MCRI</i>	21	(a, j, e, f, b, k), j, f, b, k, d, g, a, i, d, g, i, d, f, c, h
Hosa	<i>MCP</i>	4	a, j/k, d, g
Hosa	<i>MCPL</i>	3	a, j/k, d
Paha	<i>MCP</i>	4	a, j/k, d, g
Mumu	<i>MCP</i>	4	a, j/k, d, g
Mumu	<i>Crry</i>	5	a, j, e, f, h
Rano	<i>Crry</i>	7	a, j, e, f, k, h, h
Hosa	<i>DAF</i>	4	a, a, j, e
Patr	<i>DAF</i>	4	a, a, j, e
Paha	<i>DAF</i>	4	a, a, j, e
Mumu	<i>DAF</i>	4	a, a, j, e
Hosa	<i>CR2</i>	16	j, f, b, k, d, g, a, i, d, g, b, i, d, f, c, h
Mumu	<i>CR2</i>	15	j, f, b, k, d, g, a, i, d, g, i, d, f, c, h

^a Hosa, *Homo sapiens*; Mumu, *Mus musculus*; Rano, *Rattus norvegicus*; Patr, *Pan troglodytes*; Paha, *Papio hamadryas*.

^b Each SCR has been classified, using phylogenetic analysis, relative to the *CR1* SCR nomenclature described in Fig. 2. In addition, SCR sequences were compared to amino acid patterns corresponding to each *CR1* SCR cluster.

^c *CRIL* genomic sequence from NCBI database, positions 1103188 to 1453188 on contig NT_021877.12. Analysis indicates that this internal region contains additional *CRIL* SCRs that had not been described in the *CRIL* mRNA (A34924) (*ajejbkj*).

Conservation of SCR Subfamilies and Their Sets in Other Vertebrates

The validity of the phylogenetic approach was evaluated by examining non-human primates and other vertebrates. If the clusters are meaningful, they should be apparent throughout at least recent evolution. In terms of the order and position of subfamilies, *DAF* is the same in primates and *Mus musculus*. *Pan troglodytes* and *Papio hamadryas* *CR1* are identical to *Homo sapiens* *CR1* with the exception that the first *Papio hamadryas* SCR *a* has been deleted at the protein level (Table 1).

More interesting is the comparison with the mouse. *Mus musculus* *CR1* has been contrasted with *Homo sapiens* *CR1* (Holers et al. 1992; Kalli and Fearon 1994; Molina et al. 1994; Prechl and Erdei 2000), but here we show that in spite of all the differences, the sequence designated *MCRI* contains *ajejbk* and is therefore identical to the first six SCRs of *Homo sapiens* *CR1* and *CRIL*. The remaining sequence of *MCRI* is similar to *CR2*, as expected from earlier observations, suggesting that *MCRI* is in fact *CR1* and *CR2* or, in view of the present findings, a composite of *CRIL* and *CR2*. A comparison of *Homo sapiens* and *Mus musculus* *CR2* reveals the expected three tetramers (*a/b, k/i, d, f/g*) followed by *ch*, with the exception that the third tetramer in *Mus musculus* has lost the first SCR. Interestingly, as with *Homo sapiens*, *d* is less divergent than the other SCRs.

The differences between primate and rodent *CR1* suggest that the true orthologue for *Homo sapiens* *CR1* and *CRIL* may be *Crry* rather than *Mus musculus* *CR1*. Using our classification, the sequence designated *Mus musculus* *Crry* is *ajejfh*, raising the interpretation that four external domains (*ajej*) may subservise at least some *CR1* or *CRIL* functions. The *Rattus norvegicus* *Crry* sequence is designated *ajejkh* and therefore shares the *ajej*, although the terminus may be different.

Human CR1 is Closer to *Mus musculus* Crry than *Mus musculus* CR1

So as to investigate the evolutionary relationships between *Homo sapiens* *CR1* and *Mus musculus* *CR1* and *Crry*, we included all component SCRs in the same phylogenetic analysis as shown in Fig. 5. Clearly *Crry* *ajej* are very similar to the *ajej*, respectively, of the first, second, and third septamers of human *CR1*. Murine *CR1* is more remote.

SCR Subfamilies and Function

Previous studies have suggested that the first three and particularly the second and third SCRs are involved in binding to *C3* and *C4*. From Table 1, it can be deduced that the *aje* set encodes the region of relevance. Other sets of subfamilies clearly have a

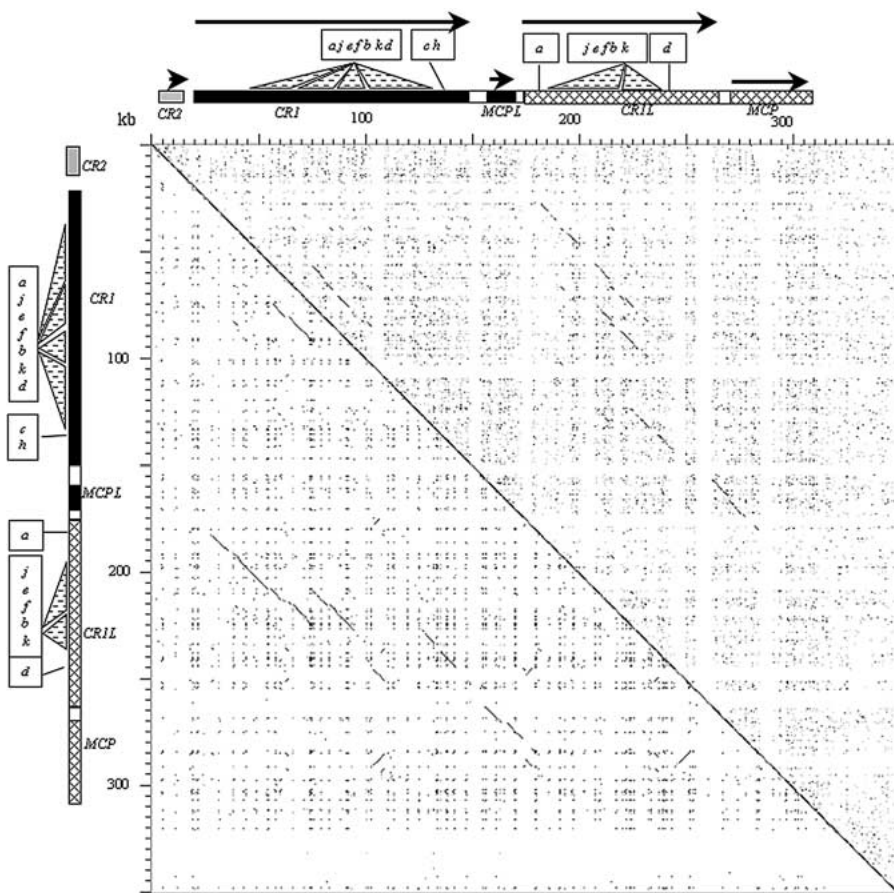


Fig. 3. Multiple duplications and divergence following segmental duplication of *CR1* and *MCP* precursors. Dot-plot analysis of 350-kb genomic region against itself. Genes *CR1*, *MCP*, *CR1L*, and *MCP1* and part of *CR2* are shown. The dot plot beneath the diagonal line is of the unmasked genomic sequence, while the dot plot above the line uses sequence that was masked using Repeat Masker (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>). Since segmental duplication of the *CR1* precursor, it is apparent that divergence has occurred by at least three separate processes: first, the basic unit has replicated such that *CR1* is now $a(jefbkd)_3ch$, whereas *CR1L* is now $(ajejbk)_2d$; second, individual SCRs have been added or lost, e.g., *c* and *h*; third, there have been insertions/deletions to explain the expansion of the region between *j(i)* and *j(ii)* of *CR1* septamer 4 (see Fig. 4a). Arrows indicate the direction of the coding sequence.

different significance. For example, the *ch* set relates to the terminus.

Peptide Patterns Marking Functional Subfamilies

If, as implied above, the classification of SCRs relates to function, it should be possible to determine which amino acid residues are specific and critical to the function of subfamilies, at least within orthologous proteins in related species. From Fig. 1 and Table 2 it can be seen that there C..C..W..C are defining in terms of SCRs generally but that most of the other patterns are actually characteristic of specific subfamilies at least within primates. For example, *a* contains an A which is specific, with some minor exceptions, within the *b* subfamily, possibly indicating the proximity of *a* and *b*. The proximal I/L/V marker, used in earlier studies (see patterns at the base of Fig. 1), is an I in *d*, *e*, and *c* but M in *h* and P in *j*. It can also be seen that the D/E marker is in *a* as a T, in *b* as Q, in *c* as M, in *d* as K, in *j* as S, in *k* as A. These residues are preceded by a G which is not present in *f*, *g*, or *i*. The more terminal proline is found in *b*, *d*, *e*, *f*, *g*, *j*, and *k* but not in *a* and not necessarily in *h*, *c*, or *i*, etc. With minor exceptions most of the so-called motifs can be

used to define subfamilies of the SCRs of *CR1* and *Cry* (Table 2).

Indels and Extensive Continuous Sequences

A striking feature of these analyses is that the different groups of SCRs have specific insertions and deletions (indels) and therefore specific lengths (Table 2). Furthermore, as a consequence, there are many more or less specific continuous sequences which might not have been expected from previous analyses.

Evolution of SCRs and CCPs

From the analyses of protein sequence, it is apparent that RCA proteins have evolved through a series of processes involving duplication and divergence of SCR domains. It is also clear that individual SCRs and particularly sets of SCRs have become specialized in terms of position, proximity, and undoubtedly function. Thus, for example, *aje* and *ch* occur on different proteins but in the same relative position. Duplications, deletions, and other processes of divergence have contributed to this process of specialization or selection.

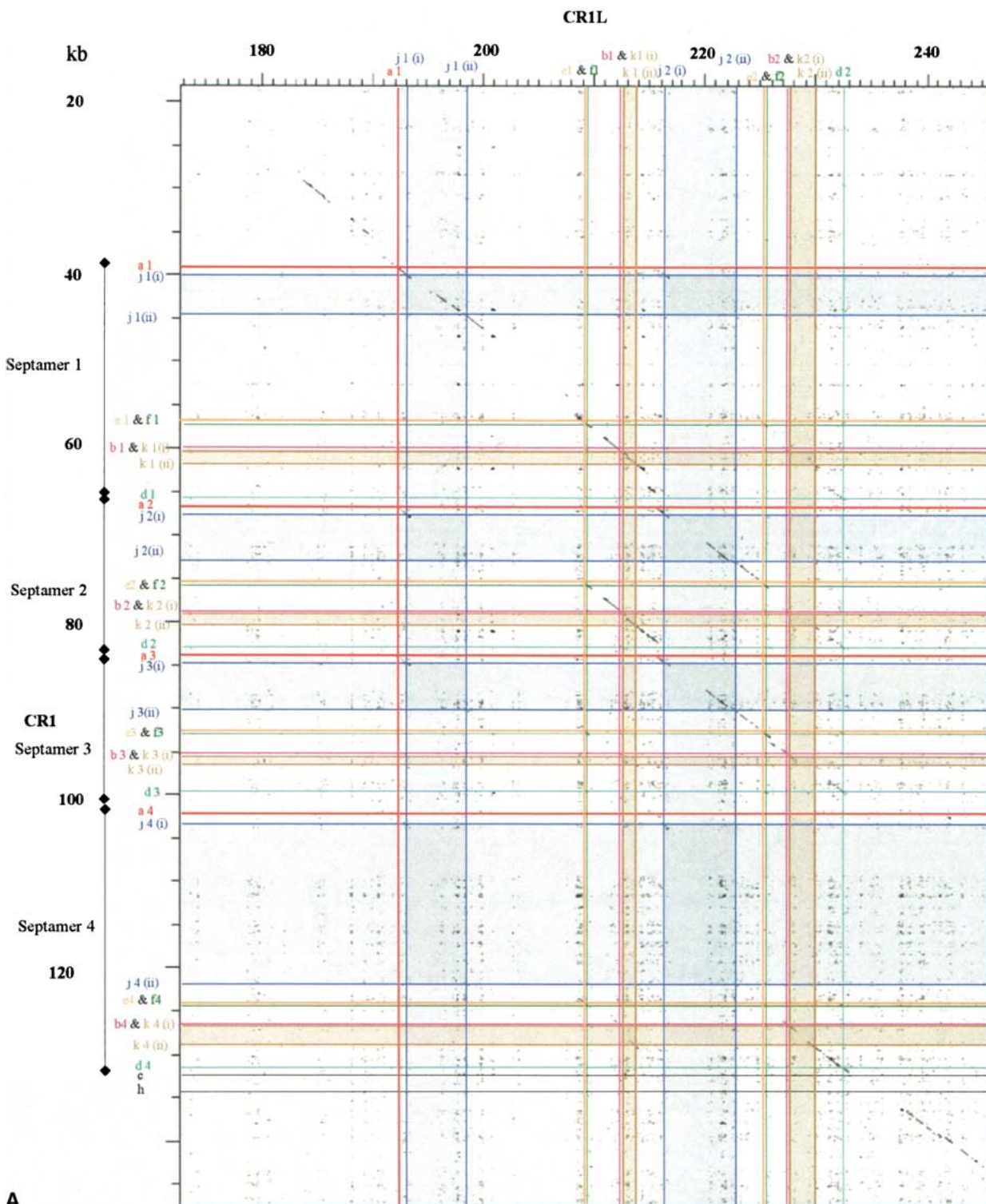


Fig. 4A. Analysis of masked genomic sequence of *CRIL* against *CR1*. Sequences were masked using Repeat Masker (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>). Colored lines running through the plot indicate positions of individual SCRs such that *a* is red, *j* is dark blue, *e* is orange, *f* is green, *b* is pink, *k* is brown, *d* is light blue, and *c* and *h* are black. Filled regions show noncoding sequence between split exons (*j* and *k*). For the purpose of discussion we assume that the *CRIL* was the progenitor of *CR1*

which arose through further piecemeal duplication of components within *CRIL*. For example, the region of *CRIL* commencing at *e1* has duplicated en bloc and there has been further diversification, especially in noncoding regions between *j2(i)* and *j2(ii)*. There are other possible units of duplication commencing at *e2*. Note also that in this sequence there are 12 SCRs within *CRIL*. The deletion of the segment including *d1*, *a2* is indicated by a vertical gap between *k1(ii)* and *j2(i)*.

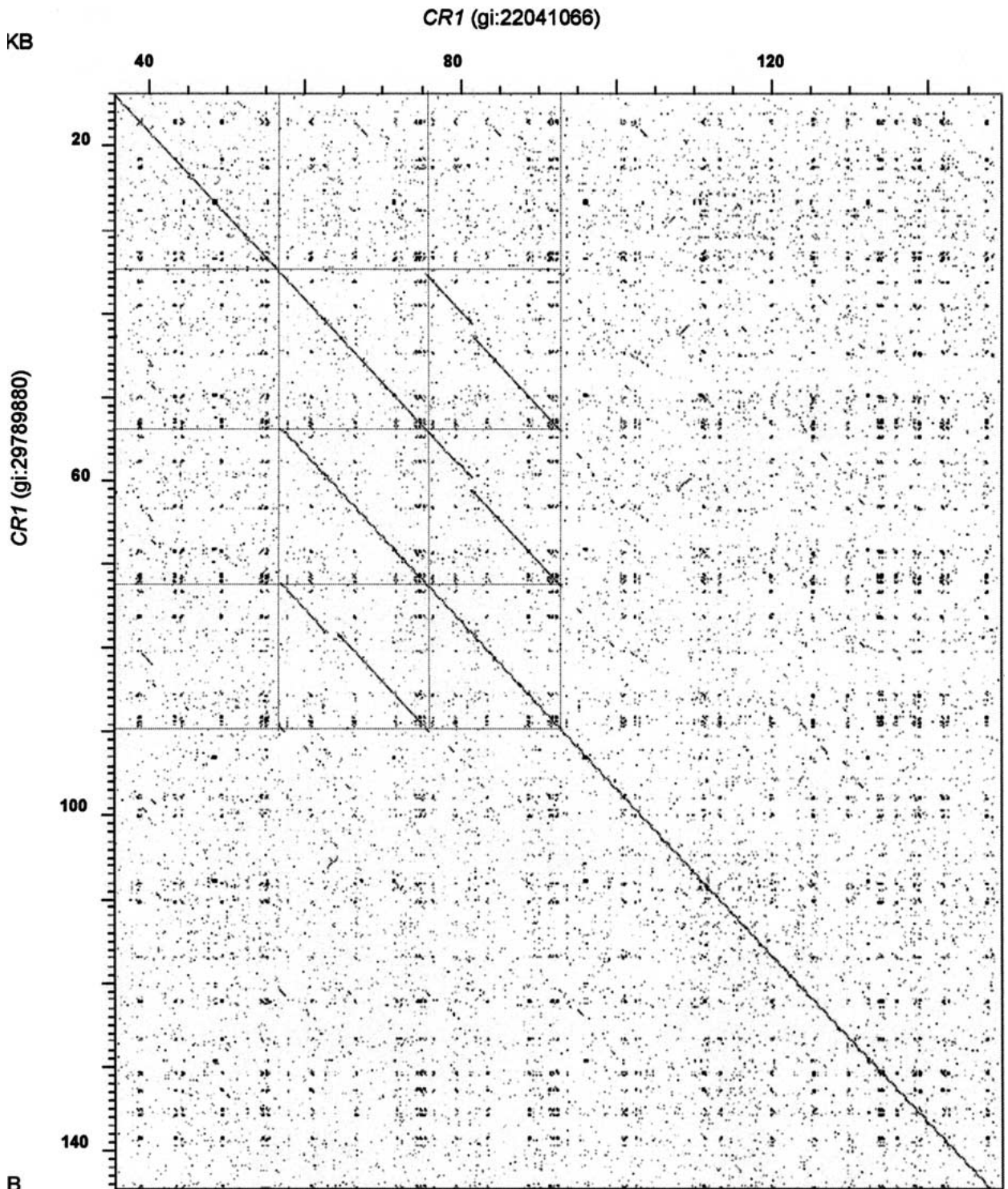


Fig. 4B. Dot-plot analysis of revised genomic sequence reveals an additional duplcon. Revised genomic contig (gi:29789880) containing *CR1* shows evidence of an additional duplicated unit when compared against previously analyzed contig (gi:22041066). The same unit (*e1* to *j2[i]*) can be identified in Fig. 4A.

Units of Duplication and Deletion

Having demonstrated that there are specific sub-families of SCRs, it becomes clear that there must have been many diverse units of duplication and deletion. For example, *a* alone has been duplicated in *DAF*. The set *ajefbkd* has been duplicated as a septamer in the

evolution of *CR1*. This septamer may have arisen by insertion and addition of *d* to *ajefbk*, which is found in *CR1L* and *MCRI*, or by a more complex process of shuffling. Based on similarities between *CR1* and *CR2*, we expect that there was a much earlier tetramer corresponding to *ajef* and a trimer corresponding to *bkd* in *CR1* and it appears that these may have been key ele-

ments in the formation of many CCPs. Note also that *aje* and *bkd* may have arisen from an ancestral form of a trimer designated $a + b, j + k, d + e$. That is, progenitors of $a + b$, of $j + k$, and of $d + e$ may have given rise to *aje* and also *bkd*.

Examination of genomic sequence must be qualified by the difficulty in assembling duplicated sequences but appears to indicate extreme complexity. Large units have also been duplicated and/or deleted; shuffling of segments, whole genes, or their fragments has been superimposed on the addition or loss of smaller units. Thus the primordial *CR1* and *MCP* sequences were duplicated as a segment of at least 100 kb which has now extended through further internal duplication and insertion.

It is apparent that duplication may have been imperfect, at least to the extent that some SCRs have diverged substantially and independently while maintaining their position within a duplicon. For example as shown in Fig. 2, in *Homo sapiens CR1*, *a* of septamer 1 (*a1*) has diverged from *a2* and *a3* (which are identical), whereas *f1*, *f2*, and *f3* are conserved. In terms of divergence and conservation within *CR1*, there are similarities between *a* and *j*, between *e* and *f*, and among *b*, *d*, and *k*. It is interesting that these groupings are similar to the sets identified above, suggesting that there may be constraints imposed by function and selection.

Relevance to CCPs Encoded Within the MHC

C2 and *Bf* contain C..C..W..C but are encoded within the MHC rather than the RCA complex. We therefore asked whether subfamilies of SCR can be identified. By phylogenetic analysis (Martinez et al. 2001), it can be seen that the *h* subfamily is present within *C2*-SCR2 and *Bf*-SCR2, suggesting that the subfamily existed before the separation of the MHC and RCA clusters. In both genes, SCR3 has some features of *a* and *b*, as might be expected if *a* and *b* had a common origin, viz $a + b$. In both genes, SCR1 is truncated but possibly related to *a*.

A remarkable feature of these analyses is the similarity between *C2* and *Bf* (see Fig. 1). Thus although the two forms are quite different in length and intron structure, they have retained almost identical SCRs. The differences between *C2* and *Bf* may relate to differences in their functional roles in the complement cascades but the residual similarities suggest that there is local conservation of sequence.

Nucleotide Analysis

Phylogenetic analysis of the nucleotide sequence (synonymous sites) of *CR1* revealed a very similar

result to the amino acid sequence as shown in Fig. 2. Not only were the clusters the same but so too were the different patterns of divergence. This finding emphasizes the apparent paradox: although it appears that *CR1* has been formed by successive duplications of a septamer, some of the components of each septamer appear to have a different evolutionary history.

To search for indicators of selection, we used a window analysis to plot the synonymous (dS) and nonsynonymous (dN) changes throughout the length of the *CR1* SCRs. The differences within a subfamily (see Fig. 2) can be predominantly synonymous (e.g., *a1* [*CR1*-1]- versus *a2* [*CR1*-8] and *a3* [*CR1*-15]) or nonsynonymous (e.g., *j2* [*CR1*-9] and *j3* [*CR1*-16] versus *j4* [*CR1*-23]) (Fig. 6), suggesting that there has been some divergence of function since the time of duplication of the septamers. In some comparisons (e.g., *j* and *f*), the dN/dS ratio reverses as the window progresses along the sequence (Fig. 6). The *d* and *b* comparisons reveal limited synonymous and nonsynonymous changes. These findings emphasize the complexity of duplication and suggest that there have been further indels and shuffling of the SCRs following duplication of the septamers.

Discussion

Units of Duplication

This work arose from our interest in comparing the MHC and the RCA complex, especially from the point of view of duplication. Within the MHC, the duplication process generally involves segments of heterogeneous genes, fragments, and indels and there is emerging evidence to suggest that the RCA complex may contain similar segments containing fragments of *CR1* and *MCP* (Hourcade et al. 2000). However, in contrast to the MHC, the dominant process within the RCA complex appears to have involved sequential duplication of one or more relatively homogeneous domains (SCRs). So as to define the latter process, we classified SCR sequences excised from several different CCPs. Remarkably, most SCRs examined could be grouped into only 11 subfamilies, designated *a* to *k*. As shown above, *Homo sapiens CR1* can be written as (*ajefbkd*)₄ *ch*, and *DAF* as *aaaje*. From such examples, it can be concluded that the subfamilies have been conserved but also shuffled to create specific sets which have themselves been conserved. Duplication has operated at multiple levels of complexity including individual SCRs, specific sets of SCRs such as *aje* and even *ajefbkd*, whole genes, and even segments with more than one gene.

After examining CCPs of different species, we conclude that this process of duplication can be im-

Table 2. Characteristics amino acid patterns of nine SCR groups

SCR group	SCR Pattern ^b	Number of residues ^c	Contributing primate
a	Cxxxxxxxx-AxxxxxxxxxxxxFxxGTTxxxYxCxPx-----Yxxxx-FxIxCxx-----xxxWxxxxxxxxxC	57	<i>Crry</i> , primate and rodent <i>CRI</i>
b	CxPPxxxL-Hxxxxxxxxxxxxx-FxxGQxVxYxCxPx-----YxLxGxxxxxxxxCTx-----xGxWSxxxPxG	57	<i>Crry</i> , primate and rodent <i>CRI</i>
c	CPHPPKIq-NGHxIGGHVSL-YLPGMTIxYxCDPq-----YLLVKGxIFCTD---QGIWSQLDHYC	57	primate <i>CRI</i>
d	CPxPPxIx-NGRHTGxxxxxxxx-xPxGKxxxYxCDxHPDRGxxFLIGBSxIRCTSDxxGNGVWSSxAPRC	67	primate <i>CRI</i>
e	CxxPPxIx-NGDFxSxxxxxxxx-FxxGxVxYxCxxxxxxxxxxxxFxlVGExSxxCTSxxxxxxxxGxWxxPxPxG	67	<i>Crry</i> , primate and rodent <i>CRI</i>
f	CxxPxxxx-Nxxxxxxxxxxxxx-FxLxxxxxxxxFxCxxG-----FxmGxxxxxxxxCxxx-----xxWxPxLPxG	56	<i>Crry</i> , primate and rodent <i>CRI</i>
h	CxxPxxM--xGxxKxLxMKKxYxYxVxLxCEDG-----YxLEGSxxSQQxD-----xxWxPxLxxC	57	Primate <i>CRI</i> and <i>Crry</i>
j	CxxPxxPx-xGxVHxxxxxxxx--xGSxxxYxCxxG-----xRLIGxxSxxCxxxxxxxx-xxWxxxPxG	58	<i>Crry</i> , primate and rodent <i>CRI</i>
k	CxxxxGxLxxGxVxxPxxLq---LGAKVxFVCxxG-----xxLKGSxSxCVLxGxxx-xxWNxSVPVC	59	Rat <i>Crry</i> . and primate <i>CRI</i>
a	CxxPPxIx-NGxIxxxxxxxxx-F--GDxIxYxCxxGxxxxxxxxFxxxGxxxLxCxx-----xAxWxxxxPxG	61	
	L A L Y E L F Y L L G		
	V S V V V V V		

^a The consensus SCR sequence derived by Hourcade et al. (1989).

^b The proteins used to define the groups are *Mus musculus* and *Rattus norvegicus Crry*, *Homo sapiens*, *Pan troglodytes*, *Papio hamadryas CRI*, and the “a, j, e, f, b, k” portion of *Mus musculus CRI*. The residues essential for defining any of the groups were assigned only when all group members had a single residue at a specific position. Positions where multiple residues were present are designated with an x. An indel at a position is indicated by a —.

^c The number of residues in each group refers to the total number of amino acids only.

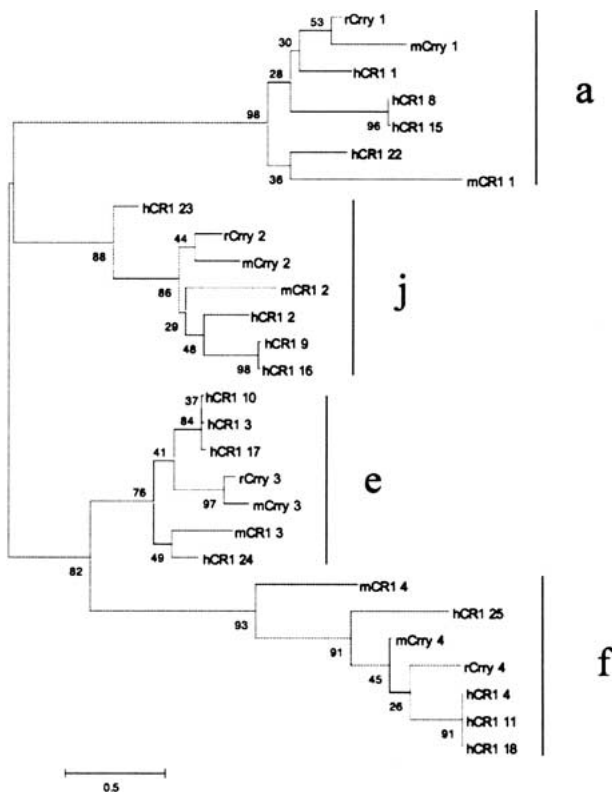


Fig. 5. Rodent *Crry* is phylogenetically closer to Hosa (*Homo sapiens*) *CRI* than to Mumu (*Mus musculus*) *CRI*. In *a*, *e*, and *f*, the *Crry* domains are more similar to the corresponding SCRs of the first, second, and third septamers of Hosa *CRI* (*hCRI*) than Mumu *CRI* (*mCRI*). Domains from Mumu *CRI* and the fourth septamer of Hosa *CRI* are more divergent. The neighbor-joining tree is unrooted and the bar at the bottom of the tree reflects the distance as calculated by the gamma distance model ($\alpha = 0.93$).

perfect or at least modified by deletion and insertion. As shown in Table 1, variants such as *ajefbk*, *ajef* and *jefbk* occur, although possibly only as additional copies. Functional diversification and species differences may be direct consequences of imperfect duplication.

Overlap Between CCPs

Historically, distinct CCPs have been identified and contrasted with those already known. Preferred ligands have been defined and differences in expression have been described. The number of SCRs in each CCP has been used as a means of classification and equivalents in different species have been sought often with confusing and controversial conclusions.

The present study has revealed far more structural overlap than could have been expected from earlier observations and from the existing nomenclature. Because single domains from multiple CCPs were examined, unexpected relationships were revealed. These can be considered under three headings: within a gene of a species, between genes of a species, and between species.

Within-Gene Relationships

For more than a decade, it has been appreciated that the amino acid sequence of *CRI* contains four long homologous repeats (LHR) or septamers followed by two different carboxy-terminal SCRs. So-called allotypes with three to six septamers have been described. Functional studies by the Atkinson (Uhrinova et al. 2003) and Barlow (Kirkitadze and Barlow 2001) groups have revealed similarities between septamers and the importance of particular sets of SCRs. For example (in *CRI*), functional Site 1, which binds C4b and has decay accelerating activity (DAA), resides in SCR 1–3, 8–10, and 15–17 of *CRI* and all three sets are *aje* using our nomenclature. Interestingly, within *DAF*, it has been shown that DAA resides in SCR 2–4 (*aje*) and that SCR1 (*a* preceding *aje*) is redundant. Thus, our approach has confirmed, extended, and simplified earlier observations (Krushkal et al. 2000). Furthermore, phylogenetic analysis alone could have predicted these functional relationships while suggesting a useful nomenclature.

Although there are insufficient data at present, it can be predicted that there will be considerable diversity when the same set is compared within a gene (isotypic) or between subjects (polymorphism).

Between-Gene Relationships

As shown in Table 1, there are many similarities between different CCPs within a species, and as indicated above, these correlate with shared functional sites, e.g., DAA is encoded within *aje*. Even more remarkable is the fact that another transmembrane CCP, *CR2*, shares the carboxy-terminal *ch*. The comparison of *CRI* and *CR2* suggests that an early and functional set is that represented by *fbkd* in *CRI*. These three sets can be regarded as the putative building blocks of CCPs and their functions as relevant to the evolution of innate and acquired immunity.

The degree of overlap between CCPs has been underestimated in the past. Attempts to define each CCP as a distinct protein must be reviewed. We expect considerable although not complete functional and immunological overlap between *CRI*, *CRIL*, and *DAF* inter alia. On the other hand, we also expect that there will be polymorphism.

Between-Species Relationships

It is often said that CCPs are very different between species but the present analysis yields a different conclusion. Within primates, subfamily classification of SCRs reveals striking similarities and the order is consistent allowing for occasional dele-

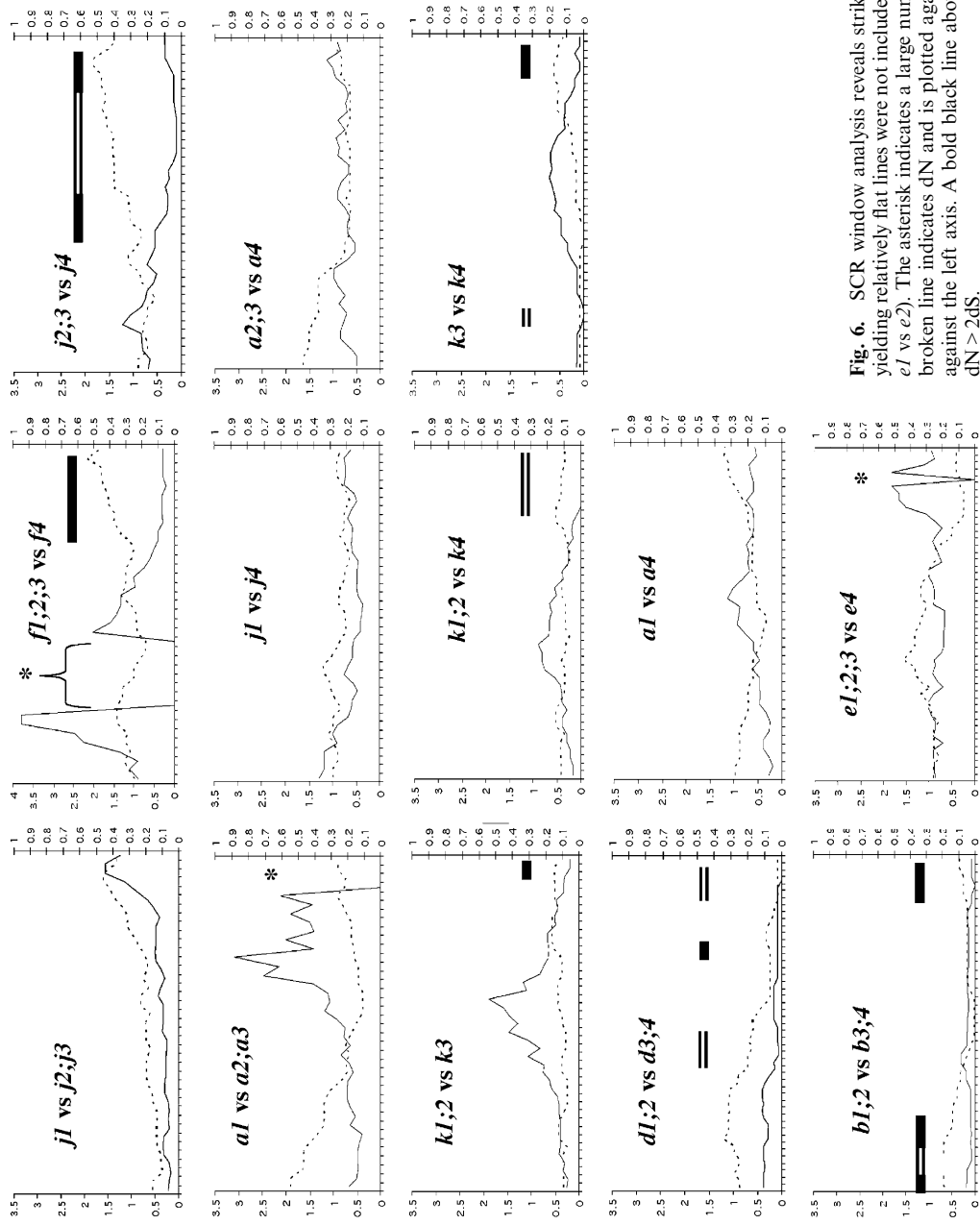


Fig. 6. SCR window analysis reveals striking differences in dS and dN patterns. Comparisons yielding relatively flat lines were not included and correlated with short genetic distances in Fig. 2 (e.g., *e1* vs *e2*). The asterisk indicates a large number based on the WINA program (Endo et al. 1996). The broken line indicates dN and is plotted against the right axis, and the black line indicates dS plotted against the left axis. A bold black line above the graph indicates dN > dS; a double line indicates dN > 2dS.

tion as in the case of the first *a* in *Papio hamadryas CR1*.

When primates are compared with rodents, some differences are apparent but these can be explained simply by single indels or by different degrees of duplication and alternative shuffling of conserved sets of SCRs. Nomenclature and annotation may give false impressions as to true and false orthologues. For example, so-called *MCRI* is a fusion of *Homo sapiens CR1/CR1L* and *Homo sapiens CR2* (with a single deletion). The human equivalent of *Mus musculus Crry* has been sought because of its likely importance in successful pregnancy. Indeed, although the nomenclature is necessarily historic and arbitrary, there could be a case for revision. In the meanwhile, we argue that the human equivalent of the CCP which protects the murine fetus is a mixture of *CR1* and all these human CCP which contain at least *aje* or more of the *ajefbkd CR1* septamer and therefore the conformational unit implied by the collocation of up to seven SCRs on the trophoblast surface. It will be important to consider these possibilities (i.e., *aje* irrespective of the CCP) when defining the immunopathology and designing antibodies or peptides which might block or initiate the process of abortion.

Evolution of SCRs

While the data are still limited and the process obviously very complex, we propose a framework for the evolution of SCRs and CCPs. Early SCRs were beneficial because of their inherent binding affinities, especially when occurring together. Any high-affinity binding has the potential to contribute to innate immunity by way of localization and immobilization of invading organisms. The same binding may also play a role in self–nonself discrimination in that there could be selection for receptors which are unoccupied and therefore free to bind foreign (rather than host) ligands. Selection would have favored closely linked (duplicated) gene clusters since increased copy number would result in increased concentrations of the receptor while facilitating the assembly of the multi-component products and the conservation of specific sets.

As individual SCR domains diverged, functional diversity increased especially when sets were imperfectly duplicated. Some sets, however, were conserved particularly when crucial to the regulation of coevolving cascades such as the complement system. Duplication of genes such as the *C3+4+5* precursors favored the imperfect but segmental duplication of specific sets of SCRs. Imperfections included insertion and deletion of SCRs as well as errors in linear sequence. Some coding polymorphism arose but was not especially favored because it was nec-

essary to retain specificity for particular ligands. Noncoding insertions may have been important in relation to achieving differential but precise regulation of the expression of CCPs. During primate evolution these processes intensified due to the recruitment of human endogenous retroviruses (HERVs) which facilitated further divergence (Gaudieri et al. 2000).

Acknowledgments. The authors gratefully acknowledge the support of the Immunogenetics Research Foundation, the National Health and Medical Research Council of Australia, the Australian Research Council, and Equitech Pty Ltd. S.G. is supported by a Howard Florey Fellowship; J.B. and C.M. are supported by APA scholarships.

References

- Bell E (2000) Murine embryonic survival depends on regulation of complement. *Immunol Today* 21:109
- Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J (1999) Genomics of the major histocompatibility complex: Haplotypes, duplication, retroviruses and disease. *Immunol Rev* 167:275–304
- Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13:685–690
- Gaudieri S, Dawkins RL, Habar K, Kulski JK, Gojobori T (2000) Nucleotide diversity within the human major histocompatibility complex: Function of hitch-hiking effect, duplications, indels and recombination. In: Kasahara M (ed) *The major histocompatibility complex: Evolution, structure, and function*. Springer Verlag, Tokyo, pp 186–200
- Heine-Suner D, Diaz-Guillen M, deVillena F, Robledo M, Benitez J, Rodriguez de Cordoba S (1997) A high-resolution map of the regulator of the complement activation gene cluster on Iq32 that integrates new genes and markers. *Immunogenetics* 45: 422–427
- Holers V, Kinoshita T, Molina H (1992) The evolution of mouse and human complement C3-binding proteins: Divergence of form but conservation of function. *Immunol Today* 13:231–236
- Hourcade D, Holers VM, Atkinson JP (1989) The regulators of complement activation (RCA) gene cluster. *Adv Immunol* 45:381–416
- Hourcade D, Liszewski M, Krych-Goldberg M, Atkinson J (2000) Functional domains, structural variations and pathogen interactions of MCP, DAF and CR1. *Immunopharmacology* 49:103–116
- Kalli K, Fearon D (1994) Binding of C3b and C4b by the CR1-like site in murine CR1. *J Immunol* 152:2899–2903
- Kirkitadze M, Barlow P (2001) Structure and flexibility of the multiple domain proteins that regulate complement activation. *Immunol Rev* 180:146–161
- Kirkitadze M, Dryden D, Kelly S, Price N, Wang X, Krych M, Atkinson J, Barlow P (1999) Co-operativity between moduls within a C3b-binding site of complement receptor type 1. *FEBS Lett* 459:133–138
- Krushkal J, Bat O, Gigli I (2000) Evolutionary relationships among proteins encoded by the regulator of complement activation gene cluster. *Mol Biol Evol* 17:1718–1730
- Krych M, Hauhart R, Atkinson J (1998) Structure-function analysis of the active sites of complement receptor type 1*. *J Biol Chem* 273:8623–8629

- Krych-Goldberg M, Hauhart R, Subramanian V, Yurcisin B, Crimmins D, Hourcade D, Atkinson J (1999) Decay accelerating activity of complement receptor type 1 (CD35). *J Biol Chem* 274:31160–31168
- Martinez OP, Longman-Jacobsen N, Davies R, Chung EK, Yang Y, Gaudieri S, Dawkins RL, Yu CY (2001) Genetics of human complement component C4 and evolution the central MHC. *Frontiers Biosci* 6:D904–D913
- Molina H, Kinoshita T, Webster C, Holers V (1994) Analysis of C3b/C3d binding sites and factor I cofactor regions within mouse complement receptors 1 and 2. *J Immunol* 153:789–795
- Prechl J, Erdei A (2000) Immunomodulatory functions of murine CR1/2. *Immunopharmacology* 49:117–124
- Reid K, Bentley D, Campbell R, Chung L, Sim R, Kristensen T, Tack B (1986) Complement system proteins which interact with C3b or C4b. A superfamily of structurally related proteins. *Immunol Today* 7:230–234
- Rodriguez de Cordoba S, Diaz-Guillen MA, Heine-Suner D (1999) An integrated map of the human regulator of complement activation (RCA) gene cluster on Iq32. *Mol Immunol* 36:803–808
- Schwarzenbacher R, Zeth K, Diederichs K, Gries A, Kostner G, Laggner P, Prassl R (1999) Crystal structure of human β 2-glycoprotein I: Implications for phospholipid binding and the antiphospholipid syndrome. *EMBO J* 18:6228–6239
- Sonnhammer E, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: SC1–10
- Szakonyi G, Guthridge J, Li D, Young K, Holers V, Chen X (2001) Structure of complement receptor 2 in complex with its C3d ligand. *Science* 292:1725–1728
- Uhrinova S, Lin F, Ball G, Bromek K, Uhrin D, Medof M, Barlow P (2003) Solution structure of functionally active fragment of decay accelerating factor. *Proc Natl Acad Sci USA* 100:4718–23
- Xu C, Mao D, Holers M, Palanca B, Cheng A, Molina H (2000) A critical role for murine complement regulator Crry in fetomaternal tolerance. *Science* 287:498–501

Web sites

- <http://pfam.wustl.edu>, PFAM protein Family Database home page
- <http://www.ncbi.nlm.nih.gov>, NCBI GenBank
- <http://www.ncbi.nlm.nih.gov/Blast>, NCBI Blast Web page
- <http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>, NCBI Blast 2 sequences
- <http://motif.genome.ad.jp/MOTIF2.html>, MOTIF home page
- <http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>, Repeat Masker