

Dating the Primigenial C4-CYP21 Duplication in Primates

Yoshihito Horiuchi,∗1 Hiroshi Kawaguchi,*1 Felipe Figueroa,* Colm O’hUigin* and Jan Klein*†

*Max-Planck-Institut für Biologie, Abteilung Immunogenetik, D-7400 Tübingen, Germany, and †Department of Microbiology and Immunology, University of Miami School of Medicine, Miami, Florida 33101

Manuscript received September 10, 1992
Accepted for publication January 29, 1993

ABSTRACT

C4 and CYP21 are two adjacent, but functionally unrelated genes residing in the middle of the mammalian major histocompatibility complex (Mhc). The C4 gene codes for the fourth component of the complement cascade, whereas the CYP21 gene specifies an enzyme (cytochrome P450c21) of the glucocorticoid and mineralocorticoid pathways. The genes occur frequently in multiple copies on a single chromosome arranged in the order C4 . . . CYP21 . . . C4 . . . CYP21. The unit of duplication (a module) is the C4-CYP21 gene pair. We sequenced the flanking regions of the C4-CYP21 modules and the intermodular regions of the chimpanzee, gorilla, and orangutan, as well as the intermodular region of an Old World monkey, the pigtail macaque. By aligning the sequences, we could identify the duplication breakpoints in these species. The breakpoint turned out to be at exactly the same position as that found previously in humans. The sequences flanking paralogous genes in the same species were found to be more similar to one another than sequences flanking orthologous genes in different species. We interpret these results as indicating that the original (primigenial) duplication occurred before the separation of apes from Old World monkeys more than 23 million years ago. The nature of the sequence at the breakpoint suggests that the duplication occurred by nonhomologous recombination. Since then, the C4-CYP21 haplotypes have been expanding and contracting by homologous crossing over which has homogenized the sequences in each species. We speculate that the reason for the concerted evolution of the primate C4-CYP21 region may be a requirement for the coevolution of certain components of the complement pathway, including the C4 component. We contrast the evolution of the C4-CYP21 region with that of other Mhc regions.

1 Permanent address: Department of Dermatology, Yokohama City University School of Medicine, Yokohama, Japan.

KROON et al. 1986), guinea pig (BITTER-SUERMANN et al. 1977), and several whale species (SPILLIAERT, PALS-DOTTIR and ARNASON 1990). Haplotypes with two C4 and two CYP21 genes are common in humans (MCLEAN et al. 1988; and others), chimpanzees (KAWAGUCHI et al. 1990; BONTROP et al. 1990; CHRISTIANSEN et al. 1991), gorillas (KAWAGUCHI and KLEIN 1992), macaques (MEVAG et al. 1983), mice (ROOS, ATKINSON and SHREFFLER 1978), rats (TOSI et al. 1985), cattle (YOSHIKA et al. 1986; CHUNG, MATTISON and MILLER 1985, 1986), pigs (KIRSZENBAUM et al. 1985), and horses (KAY et al. 1987). Haplotypes with three C4 and three CYP21 genes have been reported in humans (MCLEAN et al. 1988) and orangutans (KAWAGUCHI and KLEIN 1992). Finally, haplotypes with four C4 and four CYP21 genes occur occasionally in humans (MCLEAN et al. 1988) and orangutans (ZHANG et al. 1993). In haplotypes carrying multiple copies of C4 and CYP21 genes, the two types of gene always alternate on the linkage map (i.e., C4 . . . CYP21 . . . C4 . . . CYP21, etc.), suggesting that the basic unit is a C4-CYP21 module and that multi-modular haplotypes arise by multiplication of this unit (KAWAGUCHI, O’HUIGIN and KLEIN 1991). The human C4-CYP21 module is about 35 kilobases (kb) long and the distance between the two genes, transcription-
ally oriented in the same direction, is approximately 3 kb (CARROLL, CAMPBELL and PORTER 1985; DUNHAM et al. 1987).

Recent studies indicate, however, that in reality the module contains three genes, the third one being transcribed from the DNA strand complementary to that from which the C4 and CYP21 genes are transcribed (MOREL et al. 1989). The third gene codes for a protein related to the extracellular matrix protein tenasin (MATSUMOTO et al. 1992; GITELMAN, BRISTOW and MILLER 1992) and is either referred to as MHC-F3 because of its location in the Mhc region and the presence of fibronectin type III domains (MATSUMOTO et al. 1992), or simply as "X" (MOREL et al. 1989). The last exon coding for the 3'-untranslated (3'UT) region of the CYP21 gene overlaps with the last exon of the MHC-F3(X) gene, but the two genes are oriented in opposite directions. In haplotypes with multiple copies of the CYP21 gene, all but one of the MHC-F3(X) copies are truncated (GITELMAN, BRISTOW and MILLER 1992) and most likely pseudogenes.

GITELMAN, BRISTOW and MILLER (1992) sequenced the DNA in the region between the two modules of a bimodular human haplotype and found that the sequence in the 3' part of the region matches that found upstream from the first module, whereas the sequence in the 5' part aligns with that found downstream from the second module. The two parts of the intermodular sequences overlap by four nucleotides in the center of the region. This result suggests that the two modules arose by duplication from a single module and that this event occurred by nonhomologous recombination.

In the present study we attempt to answer the following questions: When did the duplication occur? Did nonhomologous recombination occur repeatedly in different primate species and different haplotypes, or did it occur only once, and if so, by what mechanism did the subsequent variation in module number arise? To this end, we sequenced the relevant segments of the C4-CYP21 region in the chimpanzee, gorilla, orangutan and pigtail macaque.

MATERIALS AND METHODS

Source of cosmid clones and DNA: The relevant DNA segments were obtained from cosmid clones. Three cosmid libraries were used. The first library was constructed using DNA isolated from the Epstein-Barr virus (EBV)-transformed B cell line, Hugo, established from the common chimpanzee (Pan troglodytes) at the TNO Institute of Applied Radiobiology and Immunology, Rijswijk, The Netherlands (MAYER et al. 1988). The second library was based on DNA of a western lowland gorilla (Gorilla gorilla) isolated from the fibroblast line Sylvia, which was established from a skin biopsy sample by KIRBY D. SMITH, The Johns Hopkins University School of Medicine, Baltimore, Maryland. The DNA for the third cosmid library was isolated from the cell line CP81, which was established from monocyteic leukemia cells of a 13-year-old female orangutan at the Los Angeles Zoological Garden (RASHEED et al. 1977). The pigtail macaque (Macaca nemestrina) DNA was derived from the EBV-transformed B cell line 86081 kindly provided to us by LAKSHMI GAUR, HLA Laboratory, Puget Sound Blood Center, Seattle, Washington. Genomic DNA was isolated from the indicated cell lines according to the method described by MANIATIS, FRITSCHE AND SAMBROOK (1982) and the libraries were constructed and screened according to STEINMETZ et al. (1985); for a full description of the libraries, see KAWAGUCHI et al. (1990) and KAWAGUCHI and KLEIN (1992).

Analysis of cosmid clones: DNA isolated from cosmid clones following the protocol given in MANIATIS, FRITSCHE and SAMBROOK (1982) was digested with restriction endonucleases, and the resulting fragments were separated by agarose gel electrophoresis and transferred to nitrocellulose membranes (Sartorius, Göttingen, Germany). The relevant fragments were subcloned into pHuBluescript II SK+ phagemid vector (Stratagene, Heidelberg, Germany) according to the standard method (DAVIS, DIBNER and BATTYE 1986).

Sequencing: Subclones were ligated to the pHuBluescript II SK+ phagemid vectors and the recombinant clones were picked up by the colony hybridization method (DAVIS, DIBNER and BATTYE 1986). Double-stranded DNA was prepared by the plasmid mini-boiling method (HOLMES and QUGLEY 1981). Five micrograms of DNA were denatured in 0.2 M NaOH, 0.2 mM EDTA for 30 min at 57° and sequenced by the dideoxy chain-termination method (SANGER, NICKLEN and COULSON 1977), using the Sequenase version 2.0 kit (U.S. Biochemicals, Cleveland, Ohio). Fragments were sequenced on both strands two to three times to eliminate or resolve sequencing errors and ambiguities.

Polymerase chain reaction (PCR): Genomic DNA was amplified in the GeneAmp PCR system 9600 (Perkin-Elmer, Überlingen, Germany). Two hundred nanograms of DNA were initially denatured by heating to 94° followed by a 30-cycle profile of 20 sec at 94°, 20 sec annealing at 50°, and 30 sec extension at 72° (using the GeneAmp PCR reagent kit (Perkin-Elmer). The primers used for amplification were 5'-GACTCCTTGATGGATGTTGA-3' (Tu336) and 5'-AAGGACAGCCTGGCGCCCT-3' (Tu337), which were specific for sequences located 105 base pairs (bp) upstream and 116 bp downstream of the human recombination breakpoint in the intermodular region, respectively. The amplified products were isolated by electrophoresis on low melting point agarose gel, cloned in Bluescript II SK+ vector, and sequenced.

Construction of dendrograms: The neighbor-joining genetic distance method of SAITOU and NEI (1987) was used for evaluating evolutionary relationships among nucleotide sequences. Genetic distances were calculated by the two-parameter method (KIMURA 1980).

RESULTS

Contig maps of the C4-CYP21 region in the chimpanzee (Figure 1), gorilla (Figure 2) and orangutan (Figure 3) were constructed previously by KAWAGUCHI et al. (1990) and KAWAGUCHI and KLEIN (1992). They revealed the existence of two C4-CYP21 modules each in the chimpanzee and gorilla, and three modules in the orangutan. To identify the breakpoints of the duplication that produced these multimodular chromosomes, we isolated cosmid fragments located upstream from the most 5' C4 gene (region A in Figures
Dating a Primate Duplication

333

C-621

I FIGURE 1.—Contig map of the chimpanzee C4-CYP21 region. Segments at the top indicate the extent of the relevant cosmids (numbered; an asterisk specifies an allelic cosmid). The boxes in the middle portion indicate the genes, arrows the transcriptional orientation of the genes, and the circled letters the sequenced regions. The lower part of the figure gives the restriction map and the blowups of the relevant restriction fragments (B, BamHI; E, EcoRI; H, HindIII; K, KpnI; S, SacI). The rectangles at the lowest part indicate the fragments sequenced (sequencing strategy is shown by arrows and homology by different shading).

1–3), between the modules (region B in Figures 1 and 2, as well as regions B' and B' in Figure 3), and downstream from the most 3' CYP21 gene (region C in Figures 1–3). The chimpanzee (Figure 1) 2.4-kb A-region fragment was derived from the cosm id clone C-62 after digestion with EcoRI and BamHI. Further digestion with SacI produced a 0.95-kb fragment which was subcloned and sequenced. Two allelic chimpanzee B-region fragments were derived from cosmids C-65 and C-04 (2.0-kb HindIII-BamHI fragments which were then used to produce 0.6-kb SacI fragments for sequencing). The chimpanze e C-region fragment was derived from cosm id clone C-81 (a 1.6-kb KpnI fragment produced, upon digestion with HindIII, a 1-kb fragment, which was sequenced). The gorilla (Figure 2) A-region fragment was obtained from cosm id G-1 upon digestion with EcoRI and BamHI (2.3 kb) and subsequent trimming with SacI (0.95 kb). The B-region fragments were derived from cosmids G-21 and G-2 (mapping to identical regions in homologous chromosomes) after digestion, first with EcoRI (14.5 kb) and then with SacI (0.6 kb). The gorilla C-region fragment was produced by digestion of the cosm id clone G-18, first with EcoRI (8.0 kb) and then with SacI (2.4 kb). The orangutan (Figure 3) A-region fragment came from the cosm id clone O-15, digested first with EcoRI/BamHI (2.4 kb) and then with SacI (0.95 kb). The orangutan B- and B'-region fragments were derived from clones O-17 and O-4, respectively, by digestion with KpnI (6.0 kb) and then with SacI (0.6 kb). The orangutan C-region fragment was produced from clone O-6 by digestion with EcoRI (14.5 kb), followed by digestion with SacI (2.5 kb). All these fragments were subcloned and sequenced. In the case of the pigtail macaque, the DNA derived from the intermodular region was amplified by PCR using primers based on the human sequence. The amplification produced the expected 263-bp band, which was then sequenced.

In all four primate species, part of the B (B')-region sequence was found to align with part of the A-region sequence (we refer to the former segment as B-A), while an adjacent part of the B (B')-region sequence aligned with part of the C-region sequence [this seg-
ment is referred to as B-C; see Figures 4 and 5 in which the human sequences derived from Gittelman, Bristow and Miller (1992) are also given. At the site of transition from B-C to B-A, four nucleotides (CAAG) were found that occurred also at the 5' end of homology in the A region and at the 3' end of homology in the C region. The only deviations were a deletion of the C in the gorilla A-region sequence, a G → A substitution in the orangutan A-region sequence, and an A → T substitution in the chimpanzee C-region. The CAAG tetranucleotide in the B region is the site of change in homology: the segment 5' of this site is homologous to the C region, whereas the 3' segment is homologous to the A region. This homology switch indicates that the breakpoint that resulted in the primigenial duplication of the module lies at this tetranucleotide site. The tetranucleotide overlap between the A and C regions suggests that the duplication occurred by nonhomologous recombination (see DISCUSSION). The presence of the same B-A to B-C segment transition in humans, the great apes, and Old World monkeys (here represented by the pigtail macaque) indicates that the primigenial duplication occurred more than 23 million years (my) ago, the estimated time of separation of the ape and Old World monkey lines (Martin 1990). In the orangutan, which has three C4-CYP21 modules and hence two intermodular regions, the same B-A to B-C and B'-A to B'-C transition indicates that the third module most likely arose by unequal crossing over from misaligned multimodular chromosomes. Had the third module arisen by an independent nonhomologous recombination, one would expect the breakpoint to have occurred by chance at a position different to that between the first and the second module.

Theoretically, it should be possible to estimate how long ago the primigenial duplication occurred from the divergence of the A- from the B-A- and of the C- from the B-C-region sequences. This calculation assumes, however, that the A and B-A, as well as the C and B-C regions, evolved independently. To test whether this assumption is fulfilled, we constructed dendrograms based on the sequences of the three regions (Figures 6 and 7). Had the various regions evolved independently, one would expect the A-region sequences of the different species to cluster together and the different B-A sequences to form another cluster; similarly, the C-region sequences of the different species should form a cluster separately from the B-C region sequences. Figures 6 and 7, however, reveal a different picture: The sequences do not cluster according to regions, but according to species. Thus there is one branch with the chimpanzee A and B-A sequences, another branch with the gorilla A and B-A sequences, and another branch still with the orangutan A and B-A sequences (Figure 6). A similar picture emerges in the dendrogram of the B and B-C sequences (Figure 7). (Because of their shortness, the human and pigtail macaque sequences were not included in these comparisons.) We note also that the orangutan B-C and B'-C sequences are very similar to each other: They differ by a single nucleotide substitution only and share several substitutions and deletions absent in all the other sequences. They must, therefore, have diverged very recently.

Taken together, these data indicate that the A and B-A, as well as C and B-C regions of the C4-CYP21 modules are not evolving independently and hence that their sequences cannot be used for estimating the time the primigenial duplication occurred. The observation that different regions within the same species are more similar to each other than the same regions of different species suggests that concerted evolution (Dover et al. 1982) has been taking place and has led to intraspecific homogenization of sequences. This interpretation is consistent with the conclusion reached in an earlier publication describing the C4 and CYP21 genes themselves: Comparison of exonic and intronic human, chimpanzee, gorilla, and orang-
The human sequence is from GITELMAN, BRISTOW and MILLER (1992). B*-A is a sequence allelic to B-A; B'-A is a sequence paralogous to B-A.

FIGURE 4.—Nucleotide sequence of the A and B-A regions of C4-CYP21 in different primates: Hsow, human; Patr, chimpanzee; Gogo, gorilla; Popy, orangutan; Mane, pigtail macaque. Simple majority consensus appears at the top. Dashes (-) indicate identity with the consensus sequence, dots (.) undetermined sequence, asterisks (*) insertions/deletions, and N, sequence ambiguity. The critical tetranucleotide is boxed.

The human sequence is from GITELMAN, BRISTOW and MILLER (1992). B*-A is a sequence allelic to B-A; B'-A is a sequence paralogous to B-A.

The answer to the questions that stimulated the present study is that the primigemial duplication probably occurred only once in the Catarrhini line which includes Old World monkeys, apes and humans. The duplication occurred before the separation of the Old World monkey and ape lines more than 23 my ago. Attempts to push this date of occurrence even further back in time have thus far failed for technical reasons: With the set of primers used in this study we have not been able to amplify by PCR the relevant segments of the C4-CYP21 regions in the New World monkeys and prosimians (our unpublished data). Presumably, these regions have diverged too much from the human

**Discussion**

Utman sequences revealed larger genetic distances between orthologous genes of different species than between paralogous genes of the same species (KAWAGUCHI, O’HURIN and KLEIN 1992; KAWAGUCHI et al. 1992). It appears, therefore, that the entire C4-CYP21 chromosomal region is evolving in a concerted fashion.

**Figure 4**

-Nucleotide sequence of the A and B-A regions of C4-CYP21 in different primates: *Hsow*, human; *Patr*, chimpanzee; *Gogo*, gorilla; *Popy*, orangutan; *Mane*, pigtail macaque. Simple majority consensus appears at the top. Dashes (-) indicate identity with the consensus sequence, dots (.) undetermined sequence, asterisks (*) insertions/deletions, and N, sequence ambiguity. The critical tetranucleotide is boxed.

The human sequence is from GITELMAN, BRISTOW and MILLER (1992). B*-A is a sequence allelic to B-A; B'-A is a sequence paralogous to B-A.
region for the primers to work. It may be necessary to construct genomic libraries from representatives of these two primate groups and sequence the relevant clones to find out how far back the primigenial duplication can actually be traced.

An alternative explanation positing repeated, independent duplications is unlikely. We assume that the duplication occurred by nonhomologous recombination. Although the mechanism of nonhomologous recombination is not known (Meuth 1989; Roth and Wilson 1986), it is believed to require extensive homology between the exchange partners and to occur at random sites in the genome. Because of the latter condition, the probability of nonhomologous recombination occurring at least four times at exactly the same site is extremely low. One could argue, however, that the duplications were in fact produced by homologous recombination between repetitive elements at the exchange site. However, no repetitive elements could be identified at this site. Furthermore, even if the CAAG tetranucleotide and its flanks were a recombinational hotspot, one would not expect individual, independent recombinations to produce identical breakpoints. In all the current models of homologous recombination, the initial nicking of the DNA is presumed to occur randomly in the aligned region and the subsequent steps of the recombinational process can be expected to introduce further
variation in the resulting products so that even if the exchange occurs in the same region several times, the breakpoints do not fall in exactly the same site.

The existence of different primate haplotypes with varying numbers of C4-CYP21 modules indicates that subsequent to the rare primigenial duplication, frequent secondary rounds of duplication have been occurring repeatedly. Others have documented that the secondary duplications occur by homologous but occurring repeatedly. Others have documented that because of a selection pressure on the C4 genes, which the CYP21 genes develop congenital adrenal hyperplasia (CAH), which is a lethal condition (White, New and Dupont 1986). Thus homogenization of CYP21 genes has fatal consequences and it is unlikely that this process is favored by natural selection. So, the concerted evolution of the C4-CYP21 region in the Catarrhini (and very likely also in other mammals), must be occurring because of a selection pressure on the C4 genes, with which the CYP21 genes accidentally became locked into a permanent partnership by the primigenial duplication.

In contrast to CYP21, both C4 loci on a bimodular chromosome are functional and individuals with complete C4 deficiency are extremely rare [less than 20 individuals of this type have been identified in the human population thus far; see Hauptmann et al. (1986)]. The complete absence of functional C4 genes leads to severe systemic lupus erythematosus (SLE), an inflammatory autoimmune disorder which, because of trapping of antigen-antibody complexes in capillaries, may affect multiple organs. The condition, in contrast to CAH, need not be fatal. Partial C4 deficiency, on the other hand, is common in the human population, because some 40% of individuals have fewer than four functioning C4 genes (Morgan and Walport 1991). The clinical course of the partial C4 deficiency, however, may vary from no detectable symptoms at all to mild episodic disorders and increased susceptibility to SLE. The C4 loci therefore tolerate homogenization much better than the CYP21 loci, but this fact in itself does not explain why their homogenization occurs in the first place. We propose that the reason lies in the multifunctionality of the C4 molecule. During its activation and inactivation, the C4 molecule must interact with at least eight other molecules—the antigen, the antibody, complement factors C3, C2, C1, as well as the C4 and C4bp complement receptors (Porter 1985). The molecules specified by the two human C4 loci, C4A and C4B, differ in two of the eight interactions (those with the antigen and the antibody) and are identical in the remaining six. The C4A molecules have, in comparison to the C4B molecules, a higher propensity to form
amide bonds with amino groups, whereas the C4B molecules form preferentially ester bonds with hydroxyl groups. It has therefore been suggested that C4A is primarily involved in promoting the physiological disposal of immune complexes, whereas the hemolytically more active C4B molecule participates preferentially in clearance of microorganisms (LAW, DODDS and PORTER 1984). The functional difference between the C4A and C4B molecules resides in the C4d fragment, specifically in the amino acid position 1106 (LAW, DODDS and PORTER 1984; KAWAGUCHI et al. 1992); the six other interaction sites are localized in other parts of the C4 molecule. It seems, therefore, that C4 is under pressure to retain the C4A-C4B difference in the C4d fragment and identity in the rest of the molecule (divergence of the two C4 genes outside the C4d region could lead to a failure in one or several of the six interactions). Indeed, sequence comparisons of the two C4 genes in bimodular chromosomes of different primates indicate retention of the C4A-C4B difference in the different species and striking homogenization in the rest of the C4 genes (KAWAGUCHI et al. 1992), as well as in the flanking regions (this communication). The reason for the concerted evolution of the C4 genes could thus be the need to maintain identity of six interaction sites in the C4A and C4B molecules, while allowing functional differentiation of the molecules at two other sites.

The C4-CYP21 region is part of the Mhc in all mammals thus far tested but its mode of evolution is not characteristic of the entire complex (KLEIN et al. 1993). The DP and DQ regions, which code for two different families of class II molecules, have remained remarkably stable since the separation of the ancestral Catarrhini and Platyrrhini at least, more than 37 my ago (GRAHOVAC et al. 1993; GAUR et al. 1992). On the other hand, the DR region, which codes for another family of class II proteins, displays a pronounced haplotype polymorphism passed on from species to species (KLEIN et al. 1991). The different primates DR haplotypes vary in length as well as in the number and composition of their genes. This region is thus marked by instability which seems to be favored by natural selection. Hence there seems to be different selection pressures influencing different regions of the Mhc, with the result that each region evolves in its own way.

We thank RYSZARD LORENZ for technical assistance, ANICA MILOSEV for the preparation of the graphics, and LYNNNE YAKES for editorial help.

LITERATURE CITED


KAWAGUCHI, H., Z. ZALESKA-RUTCZYNSKA, F. FIGUEROA, C. O'HUDIN and J. KLEIN, 1992 C4 genes of the chimpanzee,


Communicating editor: W-H. Li