# cDNA cloning and characterization of the protein encoded by RD, a gene located in the class III region of the human major histocompatibility complex

Jianhua CHENG, Kevin J. MACON and John E. VOLANAKIS\*

Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, U.S.A.

The RD gene, initially defined in the mouse, has been mapped between the Bf and C4A genes in the human major histocompatibility complex class III region. Using the mouse cDNA as a probe, we isolated and sequenced human RD cDNA clones. The composite nucleotide sequence consisted of 1301 nucleotides, excluding a poly(A) tail at the 3' end. It contained a single open reading frame encoding a polypeptide of 380 amino acid residues with a calculated molecular mass of 42274 Da. The

## INTRODUCTION

In humans, the class III region of the major histocompatibility complex (MHC) spans about 1100 kb of DNA and is located between the class I region at the telomeric end and the class II region at the centromeric end. The entire class III region has been cloned in overlapping cosmid and YAC clones [1-5] and the resulting gene map includes 36 loci. Among these are the genes encoding the complement proteins C4A, C4B, C2 and factor B, the cytochrome P-450 enzyme 21-hydroxylase, the major heatshock protein HSP70, tumour necrosis factors  $\alpha$  and  $\beta$ , and valyl-tRNA synthetase. The structure and function of the products of the remaining genes are currently unknown. However, the extremely high gene density observed in this region of chromosome 6 and the apparent association of certain human diseases with MHC class III haplotypes [6,7] has stimulated speculation about possible functional relationships among the proteins encoded by these genes [6].

The RD gene was initially identified in the mouse and mapped in the H-2S subregion of the MHC between the Bf and C4-Slp genes [8]. Nucleotide sequencing of cDNA clones predicted a 42 kDa polypeptide characterized by an unusual region consisting entirely of a basic (mostly Arg)/acidic (mostly Asp) dipeptide repeated tandemly 26 times. Codon usage indicated that this Arg-Asp (R-D) periodicity could not be explained by recent duplication events, suggesting that strong selective pressures are responsible for the conservation of the RD motif [8]. Using mouse RD cDNA as a probe, the human RD gene was also mapped in the class III MHC region and it was shown to be expressed in a variety of tissues [8]. Partial cDNA clones for human RD have been isolated and characterized [9]. In addition, genomic human RD clones also have been isolated and sequenced [10]. Alignment of the human genomic nucleotide sequence with the corresponding mouse cDNA sequence indicated that the

most striking structural feature of the deduced amino acid sequence is a region consisting entirely of 24 tandem repeats of an Arg-Asp (or Glu) dipeptide. The human RD cDNA was expressed in *Escherichia coli* as a fusion protein with glutathione S-transferase and used to produce antisera in rabbits. Western blot analysis and immunoprecipitation of lysates of biosynthetically labelled HeLa cells indicated that RD is a 44 kDa nuclear protein.

human RD gene comprises 11 exons and spans approx. 6 kb of DNA. Several discrepancies exist between the two reported human RD nucleotide sequences [9,10].

In this paper we present the complete nucleotide sequence of the human RD cDNA, which was derived from two overlapping clones. In addition, we have raised antisera against recombinant RD polypeptide fragments and utilized them to identify and characterize the RD protein in HeLa cells as a 44 kDa nuclear constituent.

#### **MATERIALS AND METHODS**

#### Isolation and nucleotide sequencing of RD cDNA clones

Human RD cDNA clones were isolated from a T cell cDNA library in  $\lambda$ gt11 (HL1016b; ClonTech, Palo Alto, CA, U.S.A.). The library was screened with the insert of the mouse RD cDNA clone WL623 [8], provided by Dr. M. Lévi-Strauss, Institut Pasteur, Paris, France. Several clones were isolated and two, HRD4.1 and HRD1.1, were used to derive the nucleotide sequence (see Figure 1). HRD4.1 extending from nt 200 to the poly(A) tail was sequenced in its entirety, while HRD1.1 was used to derive the nucleotide sequence of the 5' region. Both strands were sequenced by the chain-termination method, using M13 mp18 and mp19 as cloning vectors and modified bacteriophage T7 polymerase [11] (Sequenase; United States Biochemical, Cleveland, OH, U.S.A.). To resolve compression bands, dITP/ddITP were substituted for dGTP/ddGTP in the sequencing reactions [12]. Nucleotide sequence data were analysed by using the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group (Madison, WI, U.S.A.) and the MacVector Sequence Analysis Software (International Biotechnologies, Inc., New Haven, CT, U.S.A.).

Abbreviations used: MHC, major histocompatibility complex; GST, glutathione S-transferase; IPTG, isopropyl β-D-thiogalactoside; RT, room temperature; ORF, open reading frame; UTR, untranslated region; snRNP, small nuclear ribonucleoprotein.
\* To whom correspondence and reprint requests should be addressed.

The nucleotide sequence data reported are available from GenBank under accession no. L03411.

#### PCR

Genomic DNA was isolated from the white blood cells of unrelated individuals as described [13]. Two oligonucleotide primers, 5'-AGTGATCGACTTCGAGAACT-3' and 5'-TCAGTTAAGGTCACCTTGAC-3', derived from the 5' ends of exon 7 and intron 7 of the RD gene respectively [10], were used to amplify a 365 bp region of the gene. Amplification was performed by using  $0.5 \mu g$  of genomic DNA and 2.5 units of DNA Taq polymerase (Amplitaq; Perkin-Elmer Cetus, Norwalk, CT, U.S.A.). The PCR was carried out for 30 cycles in a 100  $\mu$ l reaction volume using a TempCycler (Coy Laboratory Products, Ann Arbor, MI, U.S.A.). Each cycle consisted of 1 min of denaturation at 94 °C, 45 s of annealing at 55 °C, and 1 min of extension at 72 °C, except for the last extension step which lasted 3 min. The PCR product was digested with EcoO109I and the overhangs were filled in. The resulting 271 bp fragment (nt 549-819) (Figure 1), containing the entire RD dipeptide repeat-encoding region, was cloned into the SmaI site of M13 mp18. From each amplified genomic DNA three independent M13 clones were isolated and sequenced.

#### Expression and purification of RD peptides

Two fragments from the coding region of the RD cDNA were expressed in E. coli as fusion polypeptides with glutathione Stransferase (GST). The first fragment (nt 210-1310; Figure 1), encoding all but the 41 N-terminal amino acid residues of the RD polypeptide, consisted of the 1.1 kb insert of the cDNA clone HRD4.1 and was obtained by *Eco*RI digestion of the  $\lambda$ gt11 phage. The second fragment, a 372 bp long DNA fragment (nt 210-581, Figure 1) encoding amino acid residues 42-165, was obtained by EcoRI/HindIII double digestion of the HRD4.1 phage clone, followed by filling in the restriction enzymes overhangs. Both cDNA fragments were isolated by electrophoresis in low-temperature-melting agarose and were subcloned into the expression vector pGEX-2T (Pharmacia LKB Biotechnology, Piscataway, NJ, U.S.A.) in-frame with the GST gene, using the EcoRI and SmaI cloning sites for the first and second fragments, respectively. The orientation of the RD cDNA insert with respect to the GST gene was ascertained by using agarose gel electrophoresis of appropriate restriction digests. Competent E. coli cells, strain NM554, were transformed by the pGEX-2T-RD recombinants. Expression transformants were selected by inducing fusion protein expression with isopropyl  $\beta$ -D-thiogalactoside (IPTG) and analysing the cell lysates by SDS/PAGE [14]. Large-scale purification of fusion proteins was carried out from lysates of IPTG-induced cultures of transformants by using affinity chromatography on gluthathione-agarose as described [15]. For most experiments described here, RD polypeptides were cleaved from the GST carrier protein by digestion of the matrix-bound fusion polypeptide with thrombin as described [16].

#### Anti-RD sera

Rabbit anti-RD sera were raised by immunizing with the polypeptide RD-(42-380). For some experiments, antibodies against the RD-(42-380) or the RD-(42-165) polypeptide were purified from these antisera by affinity chromatography. The corresponding gel matrices were prepared by coupling approx. 3 mg of either fusion polypeptide to 5 ml of tresyl-activated agarose (Schleicher and Schuell, Keene, NH, U.S.A.), according to the manufacturer's instructions.

## Cell culture

HeLa cells were grown in monolayer cultures in RPMI 1640 (Flow Laboratories, McLean, VA, U.S.A.) supplemented with 10% heat-inactivated fetal bovine serum, 2 mM glutamine, 100 units/ml penicillin and 10  $\mu$ g/ml streptomycin at 37 °C in 5% CO<sub>2</sub>.

## Western blot analysis

Total cellular extracts were prepared by lysing HeLa cells in 60 mM Tris/HCl, 150 mM NaCl buffer, pH 7.4, containing 0.25% Nonidet P-40 and proteinase inhibitors (2 mM phenylmethanesulphonyl fluoride and 1  $\mu$ g/ml each of leupeptin, aprotinin, pepstatin and chymostatin). Nuclei were isolated from HeLa cells and nuclear extracts prepared as described by Dignam et al. [17], except that all buffers contained the proteinase inhibitor cocktail described above. Extract proteins were fractionated by SDS/PAGE on 10% acrylamide gels and blotted on to nitrocellulose membranes electrophoretically as described [18]. Membrane strips were incubated at room temperature (RT) for 1 h in PBS, pH 7.4, containing 5% non-fat dried milk and 0.1 % Tween-20 to block protein binding sites. Antibody-reactive protein bands were then detected by incubating the membrane strips at RT for 2 h in the same solution, containing a 1:500 dilution of pre-immunization or anti-RD rabbit serum, or 10  $\mu$ g/ml affinity-purified antibodies. After washing, bound antibody was visualized by using a 1:1000 dilution of alkaline phosphatase-conjugated goat anti-rabbit IgG (Southern Biotechnology Associates, Birmingham, AL, U.S.A.) followed by alkaline phosphatase substrate (Sigma Chemical Co., St. Louis, MO, U.S.A.).

#### Biosynthetic labelling and immunoprecipitation

HeLa cells were grown in 10 cm diam. Petri dishes to about 75 % confluency. The culture media were removed and the monolayer was washed with warm (37 °C) PBS, pH 7.4. Methionine-free media (4 ml) were added and after 1 h of incubation they were supplemented with 0.25 mCi of [35S]methionine (DuPont-NEN Research Products, Boston, MA, U.S.A.). The cells were incubated at 37 °C in a CO, incubator for 18 h. The media were then removed, and the cells were washed three times with PBS, pH 7.4, and detached with a rubber policeman. Nuclear extracts were prepared as described [17] and incubated with  $100 \,\mu l/ml$ Protein A-Sepharose 6B (Sigma) at 4 °C for 2 h. After centrifugation, the supernatant was made 10% glycerol and 0.5% Triton X-100 and incubated with 100 µl/ml Protein-A-Sepharose-bound anti-RD IgG. The antibody-coated beads were prepared by incubating Protein A-Sepharose with affinitypurified anti-RD-(42–165) IgG (10  $\mu$ g of IgG/100  $\mu$ l of beads) at 4 °C overnight and washing three times with PBS, pH 7.4, containing 0.1 % Triton X-100. Control beads were prepared by incubating Protein A-Sepharose with normal rabbit IgG. The nuclear extract/beads mixture was rotated at 4 °C for 3 h. The beads were then collected by centrifugation and washed four times with PBS, pH 7.4, containing 0.1% Triton X-100, and once with 50 mM Tris/HCl, pH 6.8. Bound proteins were then extracted into SDS/PAGE sample buffer (62.5 mM Tris, pH 6.8, 5% 2-mercaptoethanol, 10% glycerol, 2% SDS, 0.005%Bromophenol Blue) by heating at 90 °C for 10 min, and subjected to SDS/PAGE on 10% acrylamide gels. After electrophoresis, the gels were soaked in 500 mM sodium salicylate for 30 min, dried by vacuum, and subjected to autoradiography at RT.

## RESULTS

The nucleotide sequence of human RD cDNA, obtained by sequencing both strands of the two overlapping clones, HRD4.1 and HRD1.1, is shown in Figure 1. It consists of 1301 nt, which include a single 1143-nt-long open reading frame (ORF), a 86-ntlong 5'-untranslated region (UTR) and a 72-nt-long 3'-UTR, which is followed by a 17-nt-long poly(A) tail. A potential polyadenylation signal, AGUAAA, which deviates from the canonical AAUAAA signal but has been shown to be functional

ATC	GTTO	STG	AGCO	:CGC	ГАТ: <b>1</b>	CAG	CGG	CCA	GCG	CGG	GCG	CGG	CCG	GAG	GAC	CGI	GGG	GCC	cc	60
CGG	TTGO	CG	cccc	CTC	GGG	AGC	CAC	CAT	GTI	GGT	GAT	ACC	ccc	CGG	ACT	GAG	CGA	GGA	AG	120
								М	L	v	I	Р	P	G	L	S	E	E	E	12
									-	•	-	-	Ĩ	-	Ξ.		-5-			
AGG	AGGC	TC	IGC J	GM	GAA	ATT	CAA	CAA	GCT	CAA	GAA		GĂA		GGC	ATT	GCT	GGC	TC	180
8		τ.	007	ĸ	x	P	N	ĸ	Т.	ĸ	ĸ	ĸ	ĸ	ĸ	2	τ.	τ.	200	τ.	32
				·	·									. 2.	.2.	1		~	5	52
				-		~~~~			~ ~ ~	~~~	200	-	-	~~ ~	200	- - -	, 	3.70	36	240
IGA	NGAP	196C /	~~~~	1140	- Au					~~~	~~~~	100	191		nce		- MC 1	~ ~ ~	-703	240
ĸ	ĸ	9	5	5	5	5	T	1	2	0	6	6	v	~	R	2	4	3	5	32
AGC	AGCC	TG:	ICAT	GGA	CAC	AGC	CAC	AGC	AAC	AGA	GCA	GGC		GCA	GCI	GGT	GAA	GTC	AG	300
Q	P	v	M	D	T	A	т	•	т	R	Q	•	ĸ	Q	L	v	ĸ	s	Ģ	12
								_												
GAG	CCAI	CAC	STGC	CAT	CAA	GGC	TGA	GAC	CAA	GAA	СТС	AGG	CTT	CAA	GÇG	TTC	TCG	AAC	cc	360
A	I	s	A	I	ĸ	λ	Е	T	ĸ	N	s	G	F	ĸ	R	s	R	T	L	92
					Ŧ															
TTG	AGGG	GA	AGTI	.777	GGA	CCC	CGA	GAA	GGG	ACC	AGT	CCC	CAC	TTT	CCA	GCC	GTT	CCA	GA	420
E	G	ĸ	L	ĸ	D	P	E	ĸ	G	P	v	P	T	F	Q	P	F	Q	R	112
										Ť										
GGA	GCAT	AT	TGC	TGA	TGA	TGA	CCT	GCA	AGA	GTC	ATC	CAG	ACG	TCC	CCA	GAG	GAA	ATC	TC	480
s	I	s	A	D	D	D	L	Q	E	S	s	R	R	P	Q	R	ĸ	s	L	132
			1																	
TGT	ATGA	GAG	CT1	TGT	GTC	TTC	TAG	TGA	TCG	ACT	TCG	AGA	ACT	AGG	ACC	AGA	TGG	AGA	λG	540
Y	Е	s	F	v	s	s	s	D	R	L	R	Е	L	G	P	D	G	E	Е	152
-																				
AGG	CAGA	LCGC	sece	AGG	GGC	TGG	TGA	TGG	TCC	ccc	TCG	AAG	стт	TGA	CTG	GGG	CTA	TGA	AG	600
2	R	G	P	G	λ	G	D	G	P	P	R	S	F	D	W	G	Y	E	E	172
	-	-	-	-		-	-	-	-	-	•••	-	-	-		-	-	-	-	
2200	-		3760		CTC	CTC	ACC	CTC	~~~	TCC	ccc		cca	CAG	000	GGA	cce	CAG	cc	660
20	- -	6		201	e			 e	P			с.	200	°	P	D.			ŭ	192
ĸ	3	9	~		3	3	~		£	•			ñ		n		~			1.72
		C 3 1	CCC	CON	CAC	ACA	CCC	ACA	TCC	CCA	cro	CCA	TCC	ACA	ccc			ACA	C R	720
ATG	AGAG	GA/	ACCG	GGA	CAG	AGA	CCG	AGA	TCG	GGA	6CG	GGA	TCG	AGA	2CCG	GGA	TCG	AGA	CA	720
ATG/ E	R	igaj N	R	GGA D	CAG R	AGA D	CCG R	AGA D	TCG R	GGA E	GCG R	GGA D	TCG R	AGA D	R	D	R	D	R	720 212
E	R	IGA/ N	R	D	CAG R	D	R	AGA D	TCG R	GGA E	GCG R		TCG R	AGA D	R		R			720 212
GAGI	AGAG R ACAG	IGA/ N	ACCG R AGCG	GGA	CAG R CAG	GGA	CCG R TCG	AGA	TCG R TCG	GGA E GGA	GCG R TCG	GGA D AGA	TCG R TCG	AGA D AGA	CCG R CCG	GGA D GGA	R ACG	AGA D GGA	СА  СА	720 212 780
GAGI D	R R ACAG R	IGA N AGJ E	ACCG R AGCG R	GGA D GGA D	CAG R CAG R	AGA D GGA D	CCG R TCG R	AGA D GGA D	TCG R TCG R	GGA E GGA D	GCG R TCG R	GGA D AGA D	TCG R TCG R	AGA D AGA D	CCG R CCG R	GGA D GGA E	R ACG R	AGA D GGA D	CA R CA R	720 212 780 232
GAGI D	R R ACAG R	IGA/ N AG/ E		GGA D GGA D	CAG R CAG R	GGA	CCG R TCG R	AGA D GGA D	TCG R TCG R	GGA E GGA D	GCG R TCG R	GGA D AGA D	TCG R TCG R	AGA D AGA D	CCG R CCG R	GGA D GGA E	R ACG R	AGA D GGA D		720 212 780 232
GAGI GAGI GGGGI	AGAG R ACAG R ATCG	IGA/ N AGJ E IGGJ		GGA D GGA	CAG R CAG R TCG	AGA D GGA D	CCG R TCG R	AGA D GGA D AGA	TCG R TCG R	GGA E GGA D	GCG R TCG R GGG	GGA D AGA D TCC	TCG R TCG R TTT	AGA D AGA D CCG	CCG R CCG R J CAG	GGA GGA E GTC	ACG R GGA	AGA D GGA D TTC	CA R CA R AT	720 212 780 232 840
GAGI GAGI GGGI D	AGAG R ACAG R ATCG R	IGA/ N IAGJ E IGGJ E		IGGA D IGGA D IGGA	CAG R CAG R TCG R	AGA D GGA D AGA D	CCG R TCG R CCG R	AGA D GGA D AGA D	TCG R TCG R CCG R	GGA E GGA D AGA E	GCG R TCG R GGG G	GGA D AGA D TCC P	TCG R TCG R TTT F	AGA D AGA D CCG R	CCG R CCG R L CAG R	GGA GGA E GTC S	ACG R GGA D	AGA D GGA D TTC S	CA R CA R AT F	720 212 780 232 840 252
GAGI GAGI GGGI D	AGAG R ACAG R ATCG R	IGA/ N AGJ E IGGJ E		GGA D GGA D GGA	CAG R CAG R TCG R	AGA D GGA D AGA D	CCG R TCG R CCG R	AGA D GGA D AGA D	TCG R TCG R CCG R	GGA E GGA D AGA E	GCG R TCG R GGG G	GGA D AGA D TCC P	TCG R TCG R TTT F	AGA D AGA D CCG R	CCG R CCG R J CAG R	GGA GGA E GTC S	ACG R GGA D	AGA D GGA D TTC S	CA R CA R AT F	720 212 780 232 840 252
GAGI GAGI D GGGI D TCCC	AGAG R ACAG R ATCG R	IGA/ N IAG/ E IGG/ E		IGGA D IGGA D IGGA	CAG R CAG R TCG R CCC	AGA D GGA D AGA D TAG	CCG R TCG R CCG R GAA	AGA D GGA D AGA D	TCG R TCG R CCG R GAA	GGA GGA D AGA E TAC	GCG R TCG. R GGG G	GGA D AGA D TCC P	TCG R TCG R TTT F TGT	AGA D AGA D CCG R ATA	CCG R CCG R CAG R TGG	GGA GGA GTC S AGA	ACG R GGA D AGA	AGA D GGA D TTC S CAT	CA R CA R AT F	720 212 780 232 840 252 900
GAGI GAGI GGGI D TCCC P	AGAG R ACAG R ATCG R TGA	IGA N IAGJ E IGGJ E ACC R		GGA D GGA D GGA D AGC A	CAG R CAG R TCG R CCC P	AGA D GGA D AGA D TAG R	CCG R TCG R CCG R GAA	AGA D GGA D AGA D AGG G	TCG R TCG R CCG R GAA	GGA GGA D AGA E TAC	GCG R TCG R GGG G TCT L	GGA D AGA D TCC P CTA Y	TCG R TCG R TTT F TGT V	AGA D AGA D CCG R ATA Y	CCG R CCG R J CAG R TGG	GGA GGA GTC S AGA E	ACG R GGA D AGA	AGA D GGA D TTC S CAT	CA CA CA R AT F GA	720 212 780 232 840 252 900 272
GAGI GGGI GGGI TCCC P	AGAG R ACAG R ATCG R TGA	IGAJ N AGJ E IGGJ E ACG R	AGCG R AGCG R AGCG R GCG R	GGA D GGA D GGA D AGC A	CAG R CAG R TCG R CCC P	AGA D AGA D TAGA R	CCG R TCG R CCG R GAA K	AGA D GGA D AGA D AGG G	TCG R TCG R CCG R GAA	GGA GGA D AGA E TAC' T	GCG R TCG R GGG G G TCT L	GGA D AGA D TCC P CTA Y	TCG R TCG R TTT F TGT V	AGA D AGA D CCG R ATA Y	CCG R CCG R CAG R TGG G	GGA GGA GTC S AGA E	ACG R GGA D AGA	AGA D GGA D TTC S CAT	CA R CA R AT F GA T	720 212 780 232 840 252 900 272
GAGI GGGI D GGGI TCCC P CACC	AGAG R ACAG R ATCG R CTGA E CCAC	IGAJ N IAGJ E IGGJ E IACO R	AGCG R AGCG R AGCG R GCG R	GGA D GGA D GGA D AGC A CCG	CAG R CAG R TCG R CCC P TGG	AGA D GGA D AGA D TAG R GGC	CCG R TCG R CCG R GAA X CTT	AGA D GGA D AGA D AGG G CTC	TCG R TCG R CCG R GAA N	GGA GGA D AGA E TAC T	GCG R TCG R GGG G TCT L TGG	GGA D AGA D TCC P CTA Y AAA	TCG R TCG R TTT F TTT F TGT V CAT	AGA D AGA D CCG R ATA Y CAT	CCG R CCG R CAG R TGG TGA	GGA GGA GGA GTC S AGA E	ACG R GGA D AGA D CTC	AGA D GGA D TTC S CAT M CAT	CA R CA R AT F GA T GG	720 212 780 232 840 252 900 272 960
GAGI GGGI GGGI TCCC P CACC P	AGAG R ACAG R ATCG R TGA E CCAC	IGAJ N IAGJ IGGJ IGGJ R IACO R CC1 L		GGA D GGA D GGA D AGC A CCG R	CAG R CAG R TCG R CCC P TGG G	AGA D GGA D AGA D TAG R GGC A	CCG R TCG R CCG R GAA X CTT F	AGA D GGA D AGA D AGG G CTC S	TCG R TCG R CCG R GAA N TCC P	GGA GGA D AGA E TAC T	GCG R TCG. R GGG G G TCT L IGG; G	GGA D AGA D TCCC P CTA Y AAAA N	TCG R TCG R TTT F TGT V CAT	AGA D AGA D CCCG R ATA Y CAT I	CCG R CCG R CAG R TGG G TGA	GGA GGA GTC S AGA E CCT L	ACG R GGA D AGA D CTC S	AGA D GGA D TTC S CAT M CAT	CA CA CA CA CA CA CA CA CA CA CA CA CA C	720 212 780 232 840 252 900 272 960 292
GAGI GGGI D TCCC P CACC P	AGAG R ACAG R ATCG R CTGA E CAC T	IGA N IAGJ E IGGJ E IACO R CCT L	ACCG R AGCG R AGCG R GGCG R TTCT L	GGA D GGA D GGA D GGA D A GGA A CCG R ↓	CAG R CAG R TCG R CCCC P TGG G	AGA D GGA D AGA D TAG R GGC A	CCG R TCG R CCG R GAA X CTT F	AGA D GGA D AGA D AGG G CTC' S	TCG R TCG R CCG R GAA N TCC <sup>-</sup> P	GGA E GGA D AGA E TAC T T T T T	GCG R TCG. R GGG G G G TCT L IGG. G	GGA D AGA D TCC P CTA Y AAAA N	TCG R TCG R TTT F TGT. V CAT	AGA D AGA D CCCG R ATA Y CAT I	CCG R CCG R CAG R TGG G TGA	GGA GGA B GTC S AGA E CCT L	ACG R ACG R GGA D AGA D CTCC S	AGA D GGA D TTC S CAT M CAT	CA R CA R CA R CA R GA T GG D	720 212 780 232 840 252 900 272 960 292
GAGI GGGI D TCCC P CACC P	AGAG R ACAG R ACAG R T CTGA C T CACC	IGA/ N IAG/ E IGG/ E IACO R I CCTI L	ACCG R AGCG R AGCG R GGCG R TTCT L GAAA	GGA D GGA D GGA D GGA D A GGA A CCG R L CCG	CAG R CAG R TCG R CCCC P TGG G	AGA D GGA D AGA D TAG R GGC A CTT	CCG R TCG R CCG R GAA X CTT F CGT	AGA D GGA D AGA D AGG G CTC S CAC	TCG R TCG R CCG R GAA N TCC <sup>C</sup> P CTA	GGA E GGA D AGA TAC T TTT F TGA	GCG R TCG R GGG G G G G TCT L TGG G	GGA D AGA D TCC P CTA Y AAAA N GAT	TCG R TCG R TTT F TGT V CAT I GGA	AGA D AGA D CCCG R ATA Y CAT I GTC.	CCG R CCG R CAG R TGG G TGA D	GGA GGA E GTC S AGA E CCT L AGA	ACG R ACG R GGA D AGA D CTCC S	AGA D GGA D TTC S CAT M CAT M GGC	CA R CA R CA R CA R CA R CA CA CA CA CA CA CA CA CA CA CA CA CA	720 212 780 232 840 252 900 272 960 292 1020
GGGJ GGGJ TCCC P CACC P	AGAG R ACAG R ATCG R CTGA E CCAC T CACC P	IGAJ N IAGJ E IGGJ E IACO R I CAO	ACCG R AGCG R AGCG R GGCG R TTCT L SAAA	GGA D GGA D GGA D GGA D CCG R CCG R C CCG C	CAG R CAG R TCG R CCCC P TGG G TGCC	AGA D GGA D AGA D TAG R GGC A CTT	CCG R TCG R CCG R GAA X CTT F CGT	AGA D GGA D AGA D AGG G CTC' S CACC	TCG R TCG R CCG R GAA N TCC P CTA	GGA E GGA D AGA TAC' T TTT' F TGA	GCG R TCG R GGG G G G TCT L TGG G K	GGA D AGA D TCCC P CTA Y AAAA N GAT	TCG R TCG R TTT F TGT V CAT I GGA	AGA D AGA D CCCG R ATA Y CAT I GTC. S	CCG R CCG R CAG R TGG G TGA D AGC A	GGA D GGA E GTC S AGA E CCT L AGA D	ACG R ACG R GGA D AGA D CTCC S TCA	AGA D GGA D TTC S CAT M CAT M GGC A	CA R CA R CA R CA R CA R CA R CA R CA R	720 212 780 232 840 252 900 272 960 292 1020 312
ATG E GAGJ D TCCCC P CACC P ACCC P	AGAG R ACAG R ATCG R CTGA E CCAC T CACC P	IGAJ N IAGJ IGGJ E IGGJ R IGGJ R I I I I I I I I I I I I I	ACCG R AGCG R AGCG R GGCG R TTCT L SAAA	GGA D GGA D GGA D GGA D CCG R C CCG R C C C C	CAG R CAG R TCG R CCCC P TGG G TGC A	AGA D GGA D AGA D TAG R GGC A CTT F	CCG R TCG R CCG R GAA X CTT F CGT	AGA D GGA D AGA D AGA G CTC S CAC	TCG R TCG R CCG R GAA N TCC' P CTA	GGA E GGA D AGA E TAC' T TTT' F TGA	GCG R TCG. R GGG G G TCT L TGG. G K	GGA D AGA D TCCC P CTA Y AAAA N GAT	TCG R TCG R TTT F TGT V CAT I GGA	AGA D AGA D CCGG R ATA Y CAT I GTC S	CCG R CCG R CAG R TGG G TGA D AGC A	GGA D GGA E GTC S AGA E CCT L AGA D	ACG R GGA D AGA D CTC S TCA	AGA D GGA D TTC S CAT M GGCC A	CAR CAR CAR AT F GAT GG D CG V	720 212 780 232 840 252 900 272 960 292 1020 312
ATGA	AGAG R ACAG R ATCG R CTGA CTGA CTGA CCACC P CACC P	GAJ N AGJ E GGJ E GGJ R C C C C C C C C C C C C C C C C C C	ACCG R AGCG R AGCG R GGCG R TTCT L SAAA	GGA GGA D GGA D GGA D AGC A CCG C C C GGG	CAG R CAG R TCG R CCC P TGG G TGC A	AGA D GGA AGA D TAG R GGCC A CTT F		AGA D GGA D AGA D AGG G CTC S CAC T GGA	TCG R TCG R CCG R GAA N TCC' P CTA	GGA E GGA D AGA E TAC T T TTT F F TGA E	GCG R TCG. R GGG G G TCT L TGG. G K K	GGA D AGA D TCCC P CTA Y AAAA N GAT M GGCT	TCG R TCG R TTT F TGT V CAT I GGA E CAA	AGA D AGA D CCGG R ATA Y CAT I GTC S AGT	CCG R CCG R CAG R TGG G TGA D AGC A	GGA D GGA E GTC S AGA E CCT L AGA D CAT	ACG R GGA D AGA D CTC S TCA Q AGC	AGA D GGA D TTC S CAT M CAT M GGC A	CAR CAR CAR AT F GA T GG D CG V AA	720 212 780 232 840 252 900 272 960 292 1020 312 1080
ATGA	AGAG R ACAG R ACCAG R CTGA CTGA CCACC P CCACC P CTGA E	$\begin{array}{c} GAJ\\ N\\ \hline \\ GGJ\\ \hline \\ GGJ\\ \hline \\ GGJ\\ \hline \\ CAG\\ \hline \\ CC1\\ \hline $	ACCG R AGCG R AGCG R GGCG R TTCT L GAAA	GGA GGA GGA GGA A GGA A CCCG R CCCG C G G		AGA D GGA A AGA D TAG R GGC A CTT P CCA	CCG R TCG R CCG R GAA X CTT F CGT V GGT V	AGA D GGA D AGA D AGG G CTC' S CAC' T GGA	TCG R TCG R CCG R GAA N TCC P CTA STC S	GGA <u>B</u> GGA <u>D</u> AGA <u>C</u> TAC T TTT F TGA <u>V</u> V	GCG R TCG. R GGGG G G TCT L IGG; G K K NCA	GGA D AGA D TCC P CTA Y AAAA N GAT M GGCT	TCG R TCG R TCG R TCG R TCG T TCG T TCG R TTTT F TCG R TCG R TTTT F TCG R TTTT F TCG R TTTT F TCG R TCG R TTTT F TCG R TCG R TTTT F TCG R TCG R TTTT F TCG R TCG R TCG R TCG R TCG R TCG R TCG R TTTT F TCG R TCG R TCG R TCG R TCG R TTTT F TCG R TCG R TCG R TCG R TCG R TTTT F TCG R TCG TCG TCG TCG TCG TCG TCG TCG TCG TCG	AGA D AGA D CCCG R ATA Y CAT I GTC S AGT <sup>o</sup> V	CCG R CCG R CAG R TGG G TGA D AGC A CAA	GGA D GGA E GTC S AGA E CCT L AGA D CAT. I	ACG R GGA D AGA D CTCC S TCA Q AGCC	AGA D GGA D TTC S CAT M GGC A CCG. R	CAR CAR CAR CAR CAR CAR CAR CAR CAR CAR	720 212 780 232 840 252 900 272 960 292 1020 312 1080 332
ATGA E GAGJ D GGGJ D TCCCC P CACCC P CACCC P ACCCC P TTGC	AGAGAG R ACAG R ATCG R CTGA C CACC T CACC P CTGA E	GAJ N AGJ E GGJ R ACC R L CAC R CCT L CAC R L CAC I L CAC I L CAC I L CAC I L CAC I L	ACCG R AGCG R AGCG R GGCG R TTCT L GAAA	$\begin{array}{c} GGA\\ D\\ GGA\\ D\\ GGA\\ CGG\\ C\\ C\\ C\\ G\\ G\\ G\\ C\\ C\\ G\\ G\\$	CAG R CAG R TCG R CCCC P TGG G TGC A CCCC	AGA D GGA D AGA D TAG R GGC A CTT F CCA	CCG R TCG R CCG R GAA X CTT F CGT V V	AGA D GGA D AGA D AGG G CTC' S CACC T GGA E	TCG R TCG R CCG R GAA N TCC P CTA S TCC S	GGA E GGA D AGA TAC T TTTT F TGA E TTGA V	GCG R TCG. R GGGG G G TCT L IGG; G K K Q	GGA D AGA D TCC P CTA Y AAAA N GAT M GGCT L	TCG R TCG R TCG R TCG R TTT F TGT V CAT I GGA E CAA K	AGA D AGA D CCCG R ATA Y CAT I GTC S AGT <sup>o</sup> V	CCG R CCG R CAG R TGG G TGA D AGC A N	GGA D GGA GTC S AGA E CCT L AGA D CAT. I	ACG R GGA D AGA D CTCC S TCA Q AGCC A	AGA D GGA D TTC S CAT M GGC A CCG. R	CA R CA R CA R CA R CA R CA R CA R CA R	720 212 780 232 840 252 900 272 960 292 1020 312 1080 332
ATGA	AGAGAG R ACAG R ATCG R CTGA C CACC T CACC P CTGA E	$\begin{array}{c} GAJ\\ N\\ \end{array}$	ACCG R AGCG R AGCG R GGCG R TTCT L SAAA N SGCA	GGA GGA GGA AGC A CCG C GGA GGA	CAG R CAG R TCG R TCG G TGG G TGG C CCC P TGG G T TGG C T	AGA D GGA D AGA D TAG R GGC A CTT F CCA Q	CCG R TCG R CCG R GAA X CTT F CGT V CGT V	AGA D GGA D AGA D AGG G CTC' S CAC T GGA E	TCG R TCG R CCG R GAA N TCC' P CTA STC' S	GGA GGA D GGA D AGA TAC T TTAC T TTAC T TTAC T TTAC T TTAC T TTAC T TTAC T TTAC T TTAC T TTAC T T TTAC T T T T T T T T T T T T T	GCG R TCG R GGG G G TCT L TGG G K K K NCA	GGA D AGA D TCC P CTA Y AAAA N GAT M GGT L	TCG R TCG R TTT F TGT F TGT I GGA E CAA K	AGA D AGA D CCCG R ATA Y CAT I GTC. S AGT <sup>C</sup> V	CCG R CCG R CAG R TGG G TGA D AGC A N	GGA GGA GGC GTC S AGA CCT L AGA D CAT. I ↓ CCCC	ACGA B GGA D AGA D CTCC S TCA Q AGC A	AGA D GGA D TTC S CAT M GGCC A CCA R CCA	CAR CAR ATF GAT GGD CGV AAK GA	720 212 780 232 840 252 900 272 960 272 960 292 1020 312 1080 332
ATGA	AGAG R ACAG R ATCG R CTGA CTGA C CACC T CACC P CACC P CTGA C CACC P	GAJ N AGJ E GGJ E ACC R L CAG L CAG L CAT	ACCG R AGCG R AGCG R GGCG R TCT L SAAA N CCAA	GGA D GGA D GGA A CCG C G GGA	CAG R CAG R TCG R CCC P TGG G CCC P TGG G CCC P TGG G CCC C P TGG G CCC C P	AGA D GGA D AGA D TAG R GGCC A CCA Q CGC	CCG R TCG R CCG R GAA X CTT F CGT V V TAC	AGA D GGA D AGA C C C C C C C C C C C C C C C C C C	TCG R TCG R CCG R GAA N TCC P CTA S TCC S CTA	GGA GGA GGA D AGA TAC' T TTTT' F TGA C S GTC' S	GCG R TCG R GGG G G G TCT L TGG G K K Q TGT V	GGA D AGA D TCCC P CTA Y AAAA N GAT M GGTT L CTG	TCG R TCG R TTT F TGT F TGT V CAT I GGA E CAA K GGGG	AGA D AGA D CCCG R ATA Y CAT I GTC. S AGT <sup>C</sup> V CTCC S	CCG R CCG R CAG R TGG G TGA D AGC A CAA N CAA	GGA GGA GGA GTC S AGA E CCT L AGA D CAT I ↓ CGC A	TCG R ACG R GGA D AGA D CTCC S TCA S TCA A G C TCA S TCA V	AGA D GGGA D TTC S CAT M GGCC A CCA R CCA	CAR CAR ATF GAT GGD CGV AAK GAN	720 212 780 232 840 252 900 272 960 272 960 292 1020 312 1080 332 1140 352
ATC: E GAGJ D TCCCC P CACCC P CACCC P TTGC A CACCC P ACCCCC P ACCCCC ACCCC P ACCCCC ACCCC P ACCCCC ACCCCC P ACCCCCC ACCCCC ACCCCC ACCCCC ACCCCCC ACCCCCC ACCCCCC ACCCCCCC ACCCCCCCC	AGAG R ACAG R ATCG R CTGA CTGA CTGA C CACC P CACC P CTGA C CACC P	SAG N SAGJ E GGJ E ACC R CAT L CAT L CAT	ACCG R AGCG R AGCG R AGCG R TCT L SAAA N N CAA	GGAA D GGAA D GGAA CCGG CCGG CCGG CCGG C	CAG R CAG R TCG R CCC P TGG G CCC P TGG G CCC P TGG C CAG R R T CAG R R TCG R TCG TCG R TCG TCG TCG TCG TCG TCG TCG TCG TCG TCG	AGA D GGA D AGA D TAG R GGCC A CCTT C CA	CCG R TCG R CCG R GAA X CTT F CGT F V GGT V TAC	AGA D GGA D AGA G G CTC S CACC T GGA E G	TCG R TCG R CCG R GAA N TCC P CTA S TCC S CTA	GGA GGA D GGA D AGA T TAC' T T T T T G T S S	GCG R TCG. R GGGG G G TCT L IGG G K K K NCA K V	GGA D AGA D TCCC P CTA Y AAAA N GAT L CTG	TCG R TCG R TTT F TGT F TGT V CAT I GGA K GGGG G	AGA D AGA D CCG R ATA Y CAT I GTC S AGT V CTC S	CCG R CCG R CAG CAG CAG CAG CAA D AGC A CAA N CAA	GGA GGA GGC S AGA CCT L AGA D CCT L AGA D CCT A CCT CCT	TCG R ACG R GGA D AGA D CTC S TCA Q Q AGC A TGT V	AGA D GGGA D TTC S CAT M GGCC A CCAT R CCA R CCA	CARCAR AT F GAT GGD CGV AAK GAN	720 212 780 232 840 252 900 272 960 292 1020 312 1080 332 1140 352
ATGA	AGAG R ACAG R ATCG R CTGA C TGA C TGA C TGA C TGA C CACC P C AGCC P	SGAJ SAGJ E SGGJ E SGGJ E ACC R CAT L CAT M TAN	ACCG R AGCG R AGCG R GGCG R TCT L SAAA N CAA N CAA N CAA A CAA A CAA A CAA A CAA A CAA A CAC C C R A CCG R C C C C C C C C C C C C C C C C C	GGA D GGA D GGA D GGA C C G G G G G G G G G G G C	CAG R CAG R TCG R TCG CCC P TGG G CCC P TGG G CCC C P TGG G CCC C P	AGA D GGA D AGA D TAG R GGC A CTT F CCA Q CGC A		AGA D GGA D AGA D AGA CTC'S CACC T GGA E GGA CTCGG	TCG R TCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG CCG	GGA GGA GGA D AGA TAC' T TTGA E TGT V GTC' S CAC'	GCG R TCG. R GGGG G G TCT <sup>1</sup> L TGG. G MAAM K NCAM Q TGTC V	GGA D AGA D TCCC P CTA Y AAAA N GAT L CTG CTG	TCG R TCG R TTT F TGT F TGT V CAT I GGGA K GGGG G G	AGA D AGA D CCG R ATA Y CAT I GTC S CTC S	CCG R CCG R CAG R CAG G G TGA D CAA N CCT L	GGA GGA GGC S AGA CCT L AGA D CCT L AGA D CCT L AGA D CCT L AGA D CCT L AGA D CCT L AGA D CCT L AGA D CCT C C CCT C CCT C CCT C CCT C CCT C CCT C CCT C C C CCT C CCT C C C C C C C C C C C C C	TCG R ACG R GGA D AGA D CTC S TCA Q AGC A TGT V	AGA D GGA D TTC S CAT M CAT M GGC A CCG. R CCA Q	CAR CAR AT F GAT GGD CGV AAK GAN CT	720 212 780 232 840 252 900 272 960 292 1020 312 1080 332 1140 352
ATGA	AGAG R ACAG R ATCG R CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	GAJ SAGJ EGGJ EGGJ EGGJ R CAG R CAG CAG CAG CAG CAG CAG CAG CAG	ACCG R AGCG R AGCG R GCG R TCT L AAAA C AAAA C C AAAA C C AAAA C C C A C	GGA D GGA D GGA D GGA A CCG C C G G G D TTG	CAG R CAG R TCG R CCC P TGG G CCC P TGG G TGC TGC A CCA	AGA D GGA D AGA D TAG R GGC A CTT F CCA CCC A CCC CCC		AGA D GGA D AGA D AGA CTC'S CAC' T GGA E T GGA CTC'S CAC' T GGA	TCG R TCG R TCG R CCG R GAA N TCC' P CTA STC' S CAA K SAG	GGA GGA GGA AGA TAC' T T T T T T G T G T G T G T G T C' S GA T T T T T T T T T T T T T	GCG R TCG. R GGG G G TCT L TGGG G K K Q TGT V V CCAC	GGA D AGA D TCC P CTA Y AAAA N GAT U CTG W GAT	TCG R TCG R TTT F TGT F TGT T CAT I GGA E CAA K GGG G G TGT	AGA D AGA D CCG R ATA Y CAT I GTC S CTC S CTA	CCG R CCG R CAG R CAG R CAG R CAG CAA CCA N CCT L CAG	GGA GGA GGC S AGA E CCT L AGA D CAT. I ↓ CGC A TGA	TCG R ACG R GGA D AGA D CTCC S TCA Q AGC A TCA V TGA	AGA D GGA D TTC S CAT M CAT M GGC A CCA Q CCT V	CAR AT F GAT GGD GGV AAK GAN CTV	720 212 780 232 900 252 900 272 960 292 1020 312 1080 332 1140 352
ATGA	AGAG R ACAG R ATCG R CTGA CCAC T CACC P CACC P CACC P CACC P CACC P	GAJ SAGJ EGGJ EGGJ EGGJ R CAG R CAG CAG CAG CAG CAG CAG CAG CAG	AGCG R AGCG R AGCG R GGCG R TCT L AAAA C CAAA C C C C C C C C C C C C C	GGA D GGA D GGA D GGA A CCGG C C G G G A CCGG C C C G G C C C C	CAG R CAG R TCG R TCG G CCC P TGG G G CCC P TGG G CCC P TGG G CCC P TGG G CCC P TGG G CCC P TGG G CAG CAG R R TCG R R TCG R R TCG R R TCG R R TCG R R TCG R R TCG R CAG CAG R TCG C CCC R TCG C CCC R TCG C CCC R TCG C CCC R TCG C CCC R TCG C CCC R TCG C CCC R TCG C CCC C C CCC R T CCC C C C C CCC R T CCCC C C C	AGA D GGA D AGA D TAG R GGC A CCA Q CGC A CCG R	CCG R TCG R CCG R GAA X CTT F CGT V TAC T CGA	AGA D GGA D AGA D AGG G CTC'S CTC'S CACC T GGA E T GGA K	TCG R TCG R GAA N TCC P TA STC S S CAA K SAG	GGA GGA D AGA TAC T TTTT F TGA GTC S GACC T	GCG R TCG. R GGG G G TCT L TGG G K K Q TGT V CCAC	GGA D AGA D TCC P CTA Y AAAA N GAT L CTG W GAT	TCG R TCG R TTT F TGT F TGT V CAT I GGA E CAA K GGG K G G IGT V	AGA D AGA D CCG R ATA Y CAT I GTC S CTC S CTA Y	CCG R CCG R CAG R TGG G TGA D AGC A N CAA N CCT L CAG	GGA GGA GGCA GTC S AGA CCT L AGA D CGC A TGA	TCG R ACG R GGA D AGA D CTCC S TCA Q AGC A TGT V TGA D	AGA AGA D GGA D TTC S CAT M GGC A CAT M GGC A CCA R CCA R CCA R CCA C C C C C C C	CAR CAR AT F GAT GGD GGV NAK GAN CTY	720 212 780 232 840 252 900 272 960 272 960 292 1020 312 1080 332 1140 352
ATGA	AGAG R ACAG R ATCG R TGA TGA CTGA C TGA C TGA C CACC P CACC P C CCCC P	$\begin{array}{c} \operatorname{GAJ} \\ \operatorname{GGJ} \\ \operatorname{GGJ} \\ \operatorname{GGJ} \\ \operatorname{GGJ} \\ \operatorname{GGJ} \\ \operatorname{GGJ} \\ \operatorname{CAG} \\ CAG$	AGCG R AGCG R AGCG R GCG R TCT L AAAA C AGCG R C C A C C A C C C R A GCG R C C C R A GCG R C C C R A C C C R A C C C R A C C C R A C C C C	GGA D GGA D GGA D GGA A CCGG C G G G C C G G G C C C C C	CAG R CAG R TCG R CCCC P TGG G CCCC P TGG G CCCC TGG C TGC TGC TGC TGC C A C CCCC P	AGA D GGA D AGA D TAG R GGC A CCA Q CGC R CCG R	CCG R TCG R CCG R GAA X CTT F CCG R GAA X CTT F V TAC T CGGA	AGA D GGA D AGA D AGG G CTC'S CACC T GGA E T GGA K	TCG R TCG R TCG R GAA N TCC' P TA STC' S TCC' S TCC' R SAG R	GGA GGA D AGA E TACC T TTTT F TGA S GTC S GACC T	GCG R TCG. R GGGG G G TCT L TGG. G K K CAC Q Q	GGA D AGA D TCCC P CTA Y AAAA N GAT' K GAT' L CTGC W GAT' I	TCG R TCG R TTT F TGT F TGT CAT I GGA E CAA E CAA K GGGG G G G G CAA	AGA D AGA D CCCG R ATA Y CAT I GTC S CTC S CTC S CTA Y	CCG R CCG R CCG R CAG R TGG CAG R TGA CAA N CCT L CAG S	GGA GGA GGC S GTC S AGA CCT L AGA D CCT I L CCC I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I L CCT I C C C C	TCG R ACGA C GGA D AGA D CTCC S TCA Q Q AGCC A TGTC V TGA C	AGA AGA D GGA D TTC S CAT M GGCC A CAT M GGCC A CCA R CCA R CCA R CCA C CAT	CAR ATF GAT GGD GV AAK GAN CY	720 212 780 232 840 252 960 272 960 292 1020 312 1080 332 1140 352
ATGA E GAGJ D GGGJ D TCCCC P TCCC P CACCC P ACCCC P ACACC S ACAC	AGAGA R ACAG R ATCG R CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	$\begin{array}{c} \mathbf{G}\mathbf{A}\mathbf{J}\\ \mathbf{S}\mathbf{G}\mathbf{G}\mathbf{J}\\ \mathbf{B}\\ \mathbf{G}\mathbf{G}\mathbf{J}\\ \mathbf{C}\mathbf{A}\mathbf{G}\mathbf{G}\\ \mathbf{C}\mathbf{C}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{C}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{C}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{I}\\ \mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}\mathbf{G}$	AGCG R AGCG R AGCG R AGCG R TCT L GGCG R TCT L SAAA S CAA S CAA S C C C C C C C C C C	GGA GGA GGA GGA GGA CCG CGG GGA TTG C TTG C TTG TTG	CAG R CAG R TCG R CCC P G G CCC P G G CCC C P G G CCC C P G G CCC C P G G CCC C P G G CCC C P G G CCC C P G G C C C C	AGA D GGA D AGA D TAG R GGC A CCA Q CCA CCGC R CCA CCGC R CCA	CCG R TCG R CCG R GAA K CTT F CCG R GAA K CTT F CCG T T CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C CCG R C C C C	AGA D GGA D AGG G G C T GGA C T GGA C T G G A C T C C C C C C C C C C C C C C C C C	TCG R TCG R CCG R GAA N TCC' P CTA S GTC' S CAA K SAGG	GGA GGA GGA C GGA T T T T T T T T T T T T T	GCG R TCG R GGG G G TCT L TGG G K K K K V CCA Q AGAO	GGA D AGA D TCCC P CTA Y AAAA N GAT' I GAT' I GAT' I SCT(	TCG R TCG R TTT F TGT F TGT V CAT I GGGA CAA K GGGG G G G G G G G G G G G G G G G	AGA D AGA D CCG R ATA Y CAT I CAT I CAT I CAT I CAT I CTC S CTC S CTC I S CTC I S C CTC I S C CCG R AGA	CCG R CCG R CAG R TGG G TGA D CAG CAA N CCT L CAG S CTT	GGA GGA GGC S AGA C C C C C C C C C C C C C C C C C C	ACG R GGA D AGA D CTCC S TCA Q AGCC A TCA TGA C TCA C TCA C C TCA C C C C C C C C C	AGA AGA D GGA D TTC S CAT M CAT M GGC A CAT CAT CAT CAT CAT CAT CAT	CAR ATF GAT GGD GV NAK GAN CTY AT	720 212 780 232 840 252 900 272 960 292 1020 312 1080 352 1140 352 1200 372
ATGA	AGAG R ACAG R ATCG R CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	GAJ N AGJ E GGJ E ACC R C C C C C C C C C C C C C C C C C	ACCG R AGCG R AGCG R AGCG R TTCT L AGCG R TTCT L AGCG C T L AGCG C T L AGCG R C C A C C C A C C C C C C C C C C C C	GGA D GGA D GGA CCG C C GGA D CCG C GGA D TTG C TGT V	CAG R CAG R TCG R CCC P TGG G TGC T TGC T TGC A CCA H GAC T D	AGA D GGA D AGA D TAG R GGC A CCA Q CCA CCG R CCA CCG R G G	CCG R TCG R CCG R GAA K CTT F CCG R GAA K CTT F CGT T TAC T CGA	AGA D GGA D AGG G G C T GGA C T GGA C T GGA K C TAA C TAA C	TCG R TCG R CCG R GAA N TCC <sup>·</sup> P TA STC <sup>·</sup> S TCC <sup>·</sup> P TA STC <sup>·</sup> S TA S R S GGG	GGA GGA GGA C GGA T T T T T T T T T T T T T	GCG R TCG R GGG G G TCT L TGG G G K K Q CCAC Q A GACAC	GGA D AGA D TCC P CTA Y AAAA N GGT K GGT CTG C TG C TG C TG C TG C TG	TCG R TCG R TTT F TGT T TGT T TGT CAT I GGA CAA K GGGG G G G G G G G G G G G G G G G	AGA D AGA D CCGG R ATA Y CATA S CATA S CATA V CTCC S CTA Y TTCC	CCG R CCG R CAG R TGG G TGA D AGC A N CAA N CAA N CAA N CCT L CAG S CTT	GGA GGA GGA GTC S AGA CAT. I ↓ CGC A CAT. I ↓ CGC A GTGA	TCG R ACG R GGA D AGA D CTCC S TCA Q AGC V TGA V TGA D CCTC	AGA AGA D GGA CAT S CAT M GGCC A CCG R CCA Q CCGT V CAT	CAR AT F GAT GGD GV NAK GAN CTY AT	720 212 780 232 840 252 900 272 960 272 960 272 1020 312 1080 332 1140 352 1200 372
ATGA E GAGJ GGGJ GGGJ TCCCC P CACCC P CACCC P CACCC P TTGC A A CAC S ACAJ K	ACAG R ACAG R ATCG R CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	GAJ N AGJ B GGJ ACC R CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CAG CC1 CC1 CAG CC1 CC1 CC1 CC1 CC1 CC1 CC1 CC1 CC1 CC	AGCG R AGCG R AGCG R GCG R TCT L SAAA SGCG R TCT L SAAA SGCG R TCT L SAAA SGCG R TCT L SAAA SGCG R TCT L SGCG R S C S C S C S C S R S C S S C S S C S R S S C S S S S	GGA GGA GGA GGA GGA A GGA A CCR CCC G GGA D TCC TCC V	CAG R CAG R TCG R CCC P TGG G TGC TGC A CAG R TCG CAG R TCG R TCG R TCG R TCG R TCG R TCG CAG R TCG TCG TCG TCG TCG TCG TCG TCG TCG TCG	AGA D GGA D AGA D AGA D TAG R GGC A CCA Q CGC A CCG R CCA CCA CCA CCA CCA CCA CCA CCA CCA C	CCG R TCG R GAA K CTT F GGA V TAC T GGA D T F	AGA D GGA D AGA G G G C TC S C TC S C TC S C TC G G A C T C S C TC S C TC S C TC S C TC S C TC S C TC S C A C TC S C A C T C S C A C C S C A C C S C A C C S C C S C C S C C S C C S C C S C C S S C S S C S S C S S C S	TCG R TCG R CCG R GAA N TCC <sup>°</sup> P CTA STC <sup>°</sup> STC <sup>°</sup>	GGA GGA GGA GGA C GGA C T T T GA C S GACC T C S GACC T C S C C C C C C C C C C C C C	GCG R TCG. R GGG G G G G G G TCT L IGGG G G TCT L IGGG C C C C C C C C C C C C C C C C C	GGA D AGA D TCCC P CTA Y AAAA N GAT L CTG M GAT I I SCT	TCG R TCG R TTT F TGT F TGT F TGT V CAT I GGA CAA CAA CAA CAA CAA CAA CAA CAA CAA	AGA D AGA D CCG R ATA Y CAT I CAT I S GTC S AGT V CTC S CTA Y I I TCO	CCG R CCG R CAG R TGG G TGA TGG CAG N CAA N CCT L CAG S CTT	GGA GGA GGA GGC S AGA E CCT L AGA D CAT I CGC A D CAT I CGC A CAT I CGC A CAT I CGC A CCT CT C CT C C C C C C C C C C C C C	TGR R GGA D AGA D CTC S TCA Q AGC S TGT V TGA D CTC C CTC	AGAA AGAA D GGAA TTCC S CAT M GGCCA R CCAC R CCAC Q CCAT	CAR AT F AT GGD GV AAK GAN CY AT	720 212 780 232 840 252 900 272 960 292 1020 312 1080 332 1140 352 1200 372
ATGA E GAGJ <u>D</u> GGGJ D TCCCC P CACCC P CACCC P CACCC P TTGC A CACA S ACAA K GCCCC	AGAG R ACAG R ATCG R CTGA CTGA CTGA CTGA CTGA CTGA CCAC C CCAC C CCAC C CCAC C CCAC C C CCAC C C CCAC C C C C C C C C C C C C C C C C C C C	GAJ N AGJ E GGJ E GGJ E ACC R CAT B CAT M K AAA N TGC	AGCG R AGCG R AGCG R GCG R TCT L GCG R TCT L GCG C C L GGG G G CCT L CCT L CCAN	GGA D GGA D GGA CCG C GGA D CCG C GGA D TTG C TGT V TCT	CAG R CAG R TCG R CCC P TGG G TGC A CAG R TCG R CCC P TGG G TGC A CAG R TCG R CAG R TCG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R CAG R C CAG R C CAG R C CAG R C CAG R C CAG R C CAG R C CAG R C CAG R C C C C C C C C C C C C C C C C C C	AGA D GGA D AGA D TAG R GGC A CCA CCA CCA CCA CCA CCA CCA CCA C		AGA D GGA D AGG G G CTC'S CACC T GGA CTC'S CACC T GGA K CTAC K CTAC K	TCG R TCG R GAA N TCG R GAA N TCC P CTA S TCG R S S G GAA N TCG R S S G G S S S G G S S S S G G S S S S	GGA GGA GGA GGA C GGA C T T T T T T T T T T T T T	GCG R TCG. R GGG G G G G G G TCT L IGGG G G K NAA K NCA Q Q CCA Q Q AGAQ	GGA D AGA D TCCC P CTA Y AAAA N GAT CTG CTG SGAT I GGAT I GGAT	TCG R TCG R TTT F TGT F TGT V CAT I GGA CAA CAA CAA CAA CAA CAA CAA CAA CAA	AGA D AGA D CCG R ATA Y CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CAT I CCG R AGA A A CA A C R A CA A C R A CA A C R A CA A C R A CA A C CCG R A CA A CA	CCG R CCG R CAG G TGA G TGA CAG A CAA N CCT L CAG S CTT	GGA GGA GGC S AGA CCT L AGA D CAT. I CGC A TGA D GTGA Ta.	TCG R ACGA D AGA D CTCC S TCA Q AGC TCA TCA TCA C TCA C C C C C C C C C C	AGA AGA D GGA TTC S CAT M GGC A CAT M GGC A CCA Q CGT V CAT	CAR AT F GAT GGD GGV NAK GAN CTY AT	720 212 780 232 840 252 900 272 960 292 1020 312 1080 352 1140 352 1200 372 1260 380 1301

# Figure 1 Nucleotide and predicted amino acid sequence of human RD cDNA

Nucleotide and amino acid residue numbers are on the right. Arrows mark the first nucleotides of exons 2–11. A leucine zipper motif is underlined by a broken line and the RD-repeat region by a solid line. Boxes show the putative RNA-binding motif [9], the potential N-glycosylation site, and the putative poly(A) signal, in that order. These nucleotide sequence data are available from GenBank under accession number L03411. *in vitro* [19], is located 18 nt upstream of the poly(A) tail attachment site. The ORF, ending with a UAG termination codon, encodes a 380-amino-acid-residue polypeptide with a calculated molecular mass of 43274 Da. The putative AUG initiation codon lies within a sequence GCCACCAUGU, very similar to the consensus sequence motif of eukaryotic translation initiation sites, except for the presence of a pyrimidine at the 4+ position [20].

Comparison of the RD cDNA sequence reported here to that published previously [9] reveals two major differences. First, the present sequence is 216 nt longer at the 5' end, while that reported by Surowy et al. [9] is 141 nt longer at the 3' end. The difference at the 5' end results in a longer ORF, encoding 55 additional amino acid residues. The deduced amino acid sequence of the extended N-terminal segment is very similar to that of the mouse RD [8]. In addition, as mentioned above, the AUG codon initiating the longer ORF lies in a sequence context that is highly suitable for translation initiation. Thus it appears likely that the cDNA clones isolated previously [9] were not full-length. The length difference at the 3' end is perhaps more interesting, since both sets of cDNA clones include poly(A) tails. It seems possible that the observed difference results from the differential utilization of alternative poly(A) signals. This view is supported by the similarity between the extended 3'-UTR sequence reported by Surowy et al. [9] and that reported previously for the mouse RD [8]. However, the extended 3'-UTR of the human RD does not contain an additional typical poly(A) signal. Another difference between the two cDNA sequences is found in the codon for amino acid residue 174 (Figure 1), which is reported here as AGT for Ser as compared with the previously assigned ATG for Met [9]. It is not clear whether this difference is due to sequencing error or to genetic polymorphism. However, the nucleotide sequence of a genomic clone for RD [10] also has ATG (Ser) at the corresponding position.

Alignment of the cDNA sequence reported here with that of the previously characterized genomic RD clone [10] confirms the exon-intron organization of the RD gene that was deduced previously from comparison with the nucleotide sequence of the mouse cDNA (Figure 1). The present data also allow the definition of the boundaries of intron 1 of the RD gene, which was not previously possible. Inspection of the two nucleotide sequences indicates that intron 1 follows nt 78 of the cDNA sequence (Figure 1). The resulting first intron boundaries, ...CGGgtaagg... and ...cccctgactagG... (where exonic nucleotides are capitalized), are in good agreement with the consensus sequences for intron splicing [21]. Thus exon 1 encodes all but the last 8 nt of the 5'-UTR, which are encoded by exon 2 of the RD gene. Comparison of the ORF reported here with that proposed by Speiser and White [10] on the basis of the genomic sequence indicates that two codons of the latter sequence, GAC and GAT, both encoding Asp residues and corresponding to positions immediately following amino acid residues 116 and 265 (Figure 1), are not present in our cDNA clone. We have no explanation for this discrepancy. However, these codons are also absent from the partial cDNA sequence published by Surowy et al. [9]. Six additional nucleotide differences exist between the present sequence and that reported previously by Speiser and White [10]. In each case our sequence is in agreement with that published by Surowy et al. [9].

Inspection of the nucleotide sequence encoding the dipeptide repeat region indicates that it consists of four 36-nt-long imperfect tandomly repeated sequences. The presence of these nucleotide repeats, which display 81-89% identity with their consensus sequence, raises the possibility of unequal crossing-over events leading to length polymorphisms of the RD gene [22]. To



Figure 2 SDS/PAGE analysis of purified recombinant RD peptides

Lane M, size markers; lane 1, GST-RD-(42-380) fusion peptide; lane 2, isolated RD-(42-380) peptide, lane 3, GST-RD-(42-165) fusion peptide. The size of the markers in kDa is shown on the left.



Figure 3 Western blot analysis of RD protein of HeLa cells

Total cellular (a) or nuclear (b) extracts of HeLa cells were subjected to SDS/PAGE and then blotted to nitrocellulose filters. Strips from each blot were incubated with control serum from rabbits before immunization (lanes 1), anti-RD serum (lanes 2), or affinity-purified anti-RD(42–165) antibodies (lanes 3). Bound antibodies were visualized by using alkaline phosphatase-conjugated goat anti-rabbit IgG, followed by alkaline phosphatase substrate. Positions and size in kDa of markers is shown. The 44 kDa polypeptide is indicated by arrows.

investigate this possibility we amplified by PCR and sequenced this region from the genomic DNA of 12 unrelated individuals. At least 16 distinct MHC haplotypes were represented among these 12 individuals, including two haplotypes which were found to be associated with IgA deficiency and common variable immunodeficiency [7]. One of these haplotypes, characterized by deletion of the C4A gene, has also been found to be associated with systemic lupus erythematosus, insulin-dependent diabetes mellitus and coeliac disease [6]. None of the PCR fragments



Figure 4 Immunoprecipitation analysis of RD protein

RD was precipitated from nuclear extracts prepared from <sup>35</sup>S-methionine-labelled HeLa cells by using affinity-purified anti-(RD42–165) antibodies (lane A). Lane B contains a mock immunoprecipitate obtained by using IgG purified from normal rabbit serum. Immunoprecipitates were subjected to SDS/PAGE followed by autoradiography. The positions and size in kDa of size markers is shown on the left.

sequenced showed any length differences from the cDNA sequence shown in Figure 1.

Expression of the two RD cDNA fragments, nt 210–1301 and 210–581 (Figure 1), in *E. coli* by using the pGEX-2T vector resulted in the synthesis of GST–RD fusion peptides of the predicted size. As shown in Figure 2, the purified GST–RD (42–380) fusion protein (lane 1) and the thrombin-cleaved and isolated RD-(42–380) peptide (lane 2) had apparent molecular masses of 66 and 36 kDa respectively, as compared with amino-acid-sequence-based estimates of 66 and 38.72 kDa respectively. The GST–RD-(42–165) fusion protein (lane 3) had an apparent molecular mass of 42 kDa, as compared with the estimated 40.8 kDa. Purified RD-(42–380) peptide was used to immunize rabbits. The antisera produced and affinity-purified antibodies thereof were utilized in the identification and characterization of human RD protein in HeLa cells.

Figure 3 shows Western blots of total cellular (Figure 3a) and nuclear (Figure 3b) extracts of HeLa cells developed with either unfractionated anti-RD-(42-380) serum (lanes 2) or with affinitypurified anti-RD-(42-165) antibodies (lanes 3). Lanes 1 show control blots reacted with serum obtained from the same rabbits before immunization. As shown, the anti-RD-(42-380) serum reacted with three major species of 73, 44 and 33 kDa. At least three additional, less prominent bands were present in the total cellular but not the nuclear extract. In contrast, the affinitypurified anti-RD-(42-165) antibodies reacted only with the 44 kDa species from either type of extract. It thus appears that the 44 kDa polypeptide, which is present in nuclei, represents the RD protein of HeLa cells. The identity of the additional bands reacting with the whole antiserum is not known. However, it seems possible that they represent cross-reactive proteins, since a number of nuclear proteins, including U1 small nuclear ribonuceloprotein (snRNP) 70 kDa, have basic-acidic dipeptide repeat regions [23,24], similar to although shorter than that present in RD protein. These charged regions are thought to be immunogenic [24] and thus likely to cross-react with RD protein.

This interpretation is consistent with the failure of antibodies against the RD-(42-165) peptide to recognize any of the additional protein bands, since this region of RD protein does not include the RD dipeptide repeats. If antibodies against the acidic-basic dipeptide repeat region are indeed responsible for the additional bands seen in the Western blot, the intensity of these bands may be a function of the relative abundance of the corresponding proteins in the HeLa cells extracts.

To further characterize the RD protein, we performed biosynthetic labelling experiments, followed by immunoprecipitation from nuclear extracts, SDS/PAGE and autoradiography. Affinity-purified anti-RD-(42–165) antibodies were used in these experiments, since they were shown to be specific for RD protein in the Western blotting analysis. However, as shown in Figure 4, immunoprecipitates obtained with these antibodies contained not only a 44 kDa species, which presumably represents RD protein, but also at least two additional polypeptides of about 84 and 42 kDa. Although the identity of these polypeptides is unknown, it seems possible that they constitute components of RD-containing macromolecular complexes.

## DISCUSSION

The RD gene, named for an unusual region of its putative protein product characterized by alternating positively and negatively charged amino acid residues, mainly the repeating dipeptide Arg (R)-Asp (D), was initially identified as an MHC class III gene in mice. Here we report the full-length nucleotide sequence of the human RD cDNA, derived from overlapping clones. The reported sequence reconciles discrepancies between previously reported human RD sequences [9,10] and also allows for the description of the boundaries of the first intron of the gene (Figure 1).

The single ORF of the RD cDNA sequence encodes a 380amino-acid-residue-long polypeptide with a calculated molecular mass of 43274 Da. Evidence for the expression of an RD protein of similar size was obtained for HeLa cells by Western blotting and immunoprecipitation experiments. Western blots, developed with antibodies against a region of the putative RD polypeptide not containing the repeating charged dipeptide and not exhibiting similarity with other proteins, identified a 44 kDa protein, present in total cellular and in nuclear extracts of HeLa cells (Figure 3). The nuclear localization of the RD protein was confirmed by indirect immunofluorescence (results not shown). The amino acid sequence similarity between the RD protein and the U1 snRNP 70 kDa protein and the presence of an snRNP-like domain in the RD protein [9] provide suggestive evidence for an association with snRNP particles or other macromolecular structures associated with spliceosomes [25,26]. The immunoprecipitation experiments indicating that the RD protein is most likely complexed with at least two other nuclear proteins is consistent with this hypothesis.

As mentioned previously [8], the most notable feature of the amino acid sequence of the RD protein is a 48-residue-long region starting at position 196 (Figure 1) and consisting exclusively of the tandemly repeated dipeptide of Arg and a negatively charged amino acid residue. Nineteen of the 24 negatively charged residues are Asp and the remaining five are Glu. As noted, [8,10], both the relatively poor conservation of codon usage and the lower similarity between mouse and human nucleotide sequences for this region than for the remaining cDNA indicate selective pressures for the conservation of the charge periodicity. This conclusion is supported by our failure to detect size polymorphism of this region of the RD gene, despite the presence of imperfect tandem nucleotide repeats. Admittedly the size of our sample was small; however, several MHC haplotypes were represented and two of them, which are associated with IgA deficiency and common variable immunodeficiency [7], are relatively common among Caucasians [27]. Furthermore, a similar search among 107 individuals indicated that only 3.3% of the chromosomes carried genes with 22 or 23 repeats [28]. These considerations suggest an important functional role for the RD region of the protein.

The expert secretarial assistance of Mrs. Paula Kiley is acknowledged. This work was supported in part by U.S. Public Health Service Grants AR03555 and Al21067. J.E.V. holds the Anna-Lois Waters Chair of Medicine in Rheumatology.

### REFERENCES

- 1 Sargent, C. A., Dunham, I. and Campbell, R. D. (1989) EMBO J. 8, 2305-2312
- 2 Spies, T., Bresnahan, M. and Strominger, J. L. (1989) Proc. Natl. Acad. Sci. U.S.A. 86, 8955–8958
- 3 Spies, T., Blanck, G., Bresnahan, M., Sands, J. and Strominger, J. L. (1989) Science 243, 214–217
- 4 Kendall, E., Sargent, C. A. and Campbell, R. D. (1990) Nucleic Acids Res. 18, 7251–7257
- 5 Ragoussis, J., Monaco, A., Mockridge, I., Kendall, E., Campbell, R. D. and Trowsdale, J. (1991) Proc. Natl. Acad. Sci. U.S.A. 88, 3753–3757
- 6 French, M. A. H. and Dawkins, R. L. (1990) Immunol. Today 11, 271-274
- 7 Volanakis, J. E., Zhu, Z. B., Schaffer, F. M., Macon, K. J., Palermos, J., Barger, B. O., Go, R., Campbell, R. D., Schroeder, H. W., Jr. and Cooper, M. D. (1992) J. Clin. Invest. 89, 1914–1922
- 8 Lévi-Strauss, M., Carroll, M. C., Steinmetz, M. and Meo, T. (1988) Science 240, 201–204
- 9 Surowy, C. S., Hoganson, G., Gosink, J., Strunk, K. and Spritz, R. A. (1990) Gene 90, 299–302
- 10 Speiser, P. W. and White, P. C. (1989) DNA 8, 745-751
- 11 Tabor, S. and Richardson, C. C. (1987) Proc. Natl. Acad. Sci. U.S.A. 84, 4767-4771
- 12 Mills, D. R. and Kramer, F. R. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 2232-2235
- 13 Sykes, B. C. (1983) Lancet II, 787-788
- 14 Laemmli, U. K. (1970) Nature (London) 227, 680–685
- 15 Smith, D. B. and Johnson, K. S. (1988) Gene 67, 31-40
- 16 Gearing, D. P., Nicola, N. A., Metcalf, D., Foote, S., Wilson, T. A., Gough, N. M. and Williams, R. L. (1989) Bio/Technology 7, 1157–1161
- 17 Dignam, J. D., Lebovitz, R. M. and Roeder, R. G. (1983) Nucleic Acids Res. 11, 1475–1489
- 18 Towbin, H., Staehelin, T. and Gordon, J. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 4350–4354
- 19 Sheets, M. D., Ogg, S. C. and Wickens, M. P. (1990) Nucleic Acids Res. 18, 5799–5805
- 20 Kozak, M. (1991) J. Biol. Chem. 266, 19867-19870
- 21 Mount, S. M. (1982) Nucleic Acids Res. 10, 459-472
- 22 Jeffreys, A. J., Royle, N. J., Wilson, V. and Wong, Z. (1988) Nature (London) 332, 278–281
- 23 Douvas, A. and Sobelman, S. (1991) Proc. Natl. Acad. Sci. U.S.A. 88, 6328-6332
- 24 Brendel, V., Dohlman, J., Blaisdell, B. E. and Karlin, S. (1991) Proc. Natl. Acad. Sci. U.S.A. 88, 1536–1540
- 25 Reed, R. (1990) Proc. Natl. Acad. Sci. U.S.A. 87, 8031-8035
- 26 Fu, X.-D. and Maniatis, T. (1990) Nature (London) 343, 437-441
- 27 Baur, M. P., Neugebauer, M. and Albert, E. D. (1984) in Histocompatibility Testing 1984 (Albert, E. D., Baur, M. P. and Mayr, W., eds), pp. 756–760, Springer-Verlag, Berlin and Heidelberg
- 28 White, P. C., Vitek, J., Lahita, R. G. and Speiser, P. W. (1992) Hum. Genet. 89, 243–244

Received 24 December 1992/16 April 1993; accepted 27 April 1993