Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 85 (2005) 512–523

GENOMICS

www.elsevier.com/locate/ygeno

Automated characterization of potentially active retroid agents in the human genome

Marcella A. McClure[a,*], Hugh S. Richardson[a], Rochelle A. Clinton[a], Crystal M. Hepp[a], Brad A. Crowther[a], Eric F. Donaldson[b]

[a]*Department of Microbiology and the Center for Computational Biology, Montana State University at Bozeman, 109 Lewis Hall, Bozeman, MT 59717, USA*

[b]*University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA*

Received 10 June 2004; accepted 29 December 2004

Abstract

Retroid agents are genomes that encode the reverse transcriptase (RT) and replicate by way of an RNA intermediate. Some retroid agents are implicated in disease via insertional mutagenesis, while others have been found to encode proteins essential to primate reproduction or provide regulatory sequences for host cell processes. The Genome Parsing Suite (GPS), a generic multistep automated process, was developed to characterize all RT-like sequences in the human genome database and to annotate the gene complement of the retroid agents that encode these sequences. In this report the GPS analyzes all significant WU-tBLASTn hits returned for 30 representative RT queries. A total of 128,779 unique RT signals were identified, and 7594 of these were retrieved by RTs not previously reported in the human genome. We have identified 9652 full-length long interspersed nuclear elements (LINEs). Only 159 LINEs are without stop codons or frameshifts.

Keywords: Endogenous retroviruses; Retrotransposons; Retroposons; Retroid; Repetitive elements; Genome Parsing Suite software

Retroid agents are RT-encoding genomes that replicate by reverse transcription of an RNA intermediate. Although once considered to be “junk” DNA, it is abundantly clear, as first suggested by Barbara McClintock, that these transposable elements and viruses are important in the evolution and development of eukaryotes. The full impact these agents have on human evolution, development, and disease can be known only when all such agents are identified, mapped, and evaluated as to probable function and historical relationship. Some retroid agents are implicated in disease while others have been found to encode proteins essential to primate reproduction [1–3], provide regulatory sequences for host cell processes [4–6], maintain telomeres [7], repair damaged chromosomes [8], and exchange genetic information among and between organisms [9].

The LINEs are responsible for various diseases, indirectly as mobilizers of noncoding repetitive elements [10] and directly as insertional mutagens. For example, the X-linked disorders Duchenne and Fukuyama-congenital type muscular dystrophies [11,12], Alport syndrome–diffuse leiomyomatosis [13], and chronic granulomatous disease [14] are caused by LINE insertion. In humans only two exogenous retroviruses are known to cause diseases: acquired immunodeficiency by HIV and human T cell leukemia by HTLV. Endogenous retroviruses are associated with a variety of diseases such as rheumatoid arthritis [15], systemic lupus erythematosus [16], and schizophrenia [17].

The encoding of the RT function defines the gene common among all the retroid agents that are coevolving within eukaryotes. The RT is one of two functional

Abbreviations: LINE, long interspersed nuclear element; HERV, human endogenous retrovirus; RT, reverse transcriptase; GPS, Genome Parsing Suite.

[*] Corresponding author. Fax: +1 406 994 7373.
E-mail address: mars@parvati.msu.montana.edu (M.A. McClure).

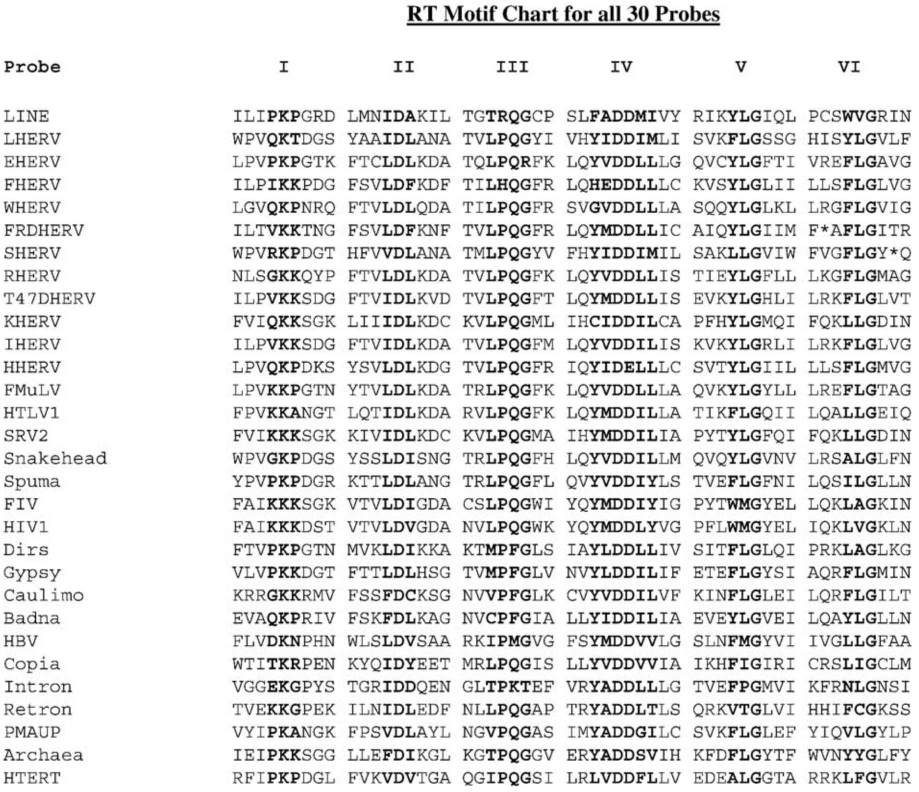
0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2004.12.006

domains of the RNA-dependent DNA polymerase. The RT provides the polymerase function and the ribonuclease H (RH) domain removes the RNA template for second-strand DNA synthesis. Margaret Dayhoff introduced the concept of discrete “islands” of highly conserved amino acids that are maintained over long evolutionary time spans that we refer to as an ordered series of motifs (OSM) [18]. The OSM of the RT [18–20] contain the key amino acids that fold to form the active site of the enzyme [21]. These six highly conserved motifs provide easy identification of potential RT function (Fig. 1). While all retroid agents encode the RT gene, many share a more extensive gene complement [18] (Fig. 2). In keeping with the historically accurate [22,23] and officially recognized nomenclature [24], seven major classes of retroid agents are easily discernible from phylogenetic analyses [25]. These classes include retroviruses, pararetroviruses, retrotransposons, retroposons, retrointrons, retroplasmids, and retrons. The cellular telomere elongation reverse transcriptase (TERT) is the only copy of the RT gene found in the human genome that is not encoded by a retroid agent.

Reports on the number and distribution of retroid agents in the human genome estimate that they make-up approximately 17% of the total DNA [26–28]. Given the evolutionary and developmental importance of these agents we have developed new software, the Genome Parsing Suite (GPS), to identify and characterize RT signals in any genome database and to annotate the retroid agents that encode these potential RTs. The GPS approach is quite different in concept from RepeatMasker [29], a program designed to identify and mask out retroid agents in the human genome with consensus DNA for repetitive elements. The GPS utilizes protein rather than nucleotide sequences to screen for the presence of retroid agents, thereby providing a deeper query into a genome. Once nucleotide substitution reaches mutational saturation, DNA sequences are no longer useful for homology searches, while the corresponding protein sequences easily retain enough signal to identify distant potential homologues. The prototype GPS provides information about the retroid agent, including genes present, condition of genes, agent boundaries, location of the agent in the genome, etc.

To date only three types of retroid agents have been reported in the human genome: the retroposons, LINEs; the retroviruses, human endogenous retroviruses (HERVs), human T cell leukemia (HTLV) [30], murine leukemia [31], and mouse mammary tumor [32,33]; and the TERT. While most HERVs are thought to be transcriptionally silenced, some have been shown to encode gene products important to human reproduction [3]. LINEs, on the other hand, are predicted to number more than 500,000 copies [26] in the human genome, with 5000–6000 of those predicted to be full-length genomes [34,35].

Although developed to analyze the evolutionary footprints of retroid agents in a given genome the GPS can be



RT Motif Chart for all 30 Probes

Probe	I	II	III	IV	V	VI
LINE	ILI**PKP**GRD	LMN**IDA**KIL	TG**TRQG**CP	SL**FADDMI**VY	RIK**YLG**IQL	PCS**WVG**RIN
LHERV	WPV**QKT**DGS	YAA**IDL**ANA	TV**LPQG**YI	VH**YIDDIM**LI	SVK**FLG**SSG	HIS**YLG**VLF
EHERV	LPV**PKP**GTK	FTC**LDL**KDA	TQ**LPQR**FK	LQ**YVDDLL**LG	QVC**YLG**FTI	VRE**FLG**AVG
FHERV	ILP**IKK**PDG	FSV**LDF**KDF	TI**LHQG**FR	LQ**HEDDLL**LC	KVS**YLG**LII	LLS**FLG**LVG
WHERV	LGV**QKP**NRQ	FTV**LDL**QDA	TI**LPQG**FR	SV**GVDDLL**LA	SQQ**YLG**LKL	LRG**FLG**VIG
FRDHERV	ILT**VKK**TNG	FSV**LDF**KNF	TV**LPQG**FR	LQ**YMDDLL**IC	AIQ**YLG**IIM	F*A**FLG**ITR
SHERV	WPV**RKP**DGT	HFV**VDL**ANA	TM**LPQG**YV	FH**YIDDIM**IL	SAK**LLG**VIW	FVG**FLG**Y*Q
RHERV	NLS**GKK**QYP	FTV**LDL**KDA	TV**LPQG**FK	LQ**YVDDLL**IS	TIE**YLG**FLL	LKG**FLG**MAG
T47DHERV	ILP**VKK**SDG	FTV**IDL**KVD	TV**LPQG**FT	LQ**YMDDLL**IS	EVK**YLG**HLI	LRK**FLG**LVT
KHERV	FVI**QKK**SGK	LII**IDL**KDC	KV**LPQG**ML	IH**CIDDIL**CA	PFH**YLG**MQI	FQK**LLG**DIN
IHERV	ILP**VKK**SDG	FTV**IDL**KDA	TV**LPQG**FM	LQ**YVDDIL**IS	KVK**YLG**RLI	LRK**FLG**LVG
HHERV	LPV**QKP**DKS	YSV**LDL**KDG	TV**LPQG**FR	IQ**YIDELL**LC	SVT**YLG**IIL	LLS**FLG**MVG
FMuLV	LPV**KKP**GTN	YTV**LDL**KDA	TR**LPQG**FK	LQ**YVDDLL**LA	QVK**YLG**YLL	LRE**FLG**TAG
HTLV1	FPV**KKA**NGT	LQT**IDL**KDA	RV**LPQG**FK	LQ**YMDDIL**LA	TIK**FLG**QII	LQA**LLG**EIQ
SRV2	FVI**KKK**SGK	KIV**IDL**KDC	KV**LPQG**MA	IH**YMDDIL**IA	PYT**YLG**FQI	FQK**LLG**DIN
Snakehead	WPV**GKP**DGS	YSS**LDI**SNG	TR**LPQG**FH	LQ**YVDDIL**LM	QVQ**YLG**VNV	LRS**ALG**LFN
Spuma	YPV**PKP**DGR	KTT**LDL**ANG	TR**LPQG**FL	QV**YVDDIY**LS	TVE**FLG**FNI	LQS**ILG**LLN
FIV	FAI**KKK**SGK	VTV**LDI**GDA	CS**LPQG**WI	YQ**YMDDIY**IG	PYT**WMG**YEL	LQK**LAG**KIN
HIV1	FAI**KKK**DST	VTV**LDV**GDA	NV**LPQG**WK	YQ**YMDDLY**VG	PFL**WMG**YEL	IQK**LVG**KLN
Dirs	FTV**PKP**GTN	MVK**LDI**KKA	KT**MPFG**LS	IA**YLDDLL**IV	SIT**FLG**LQI	PRK**LAG**LKG
Gypsy	VLV**PKK**DGT	FTT**LDL**HSG	TV**MPFG**LV	NV**YLDDIL**IF	ETE**FLG**YSI	AQR**FLG**MIN
Caulimo	KRR**GKK**RMV	FSS**FDC**KSG	NV**VPFG**LK	CV**YVDDIL**VF	KIN**FLG**LEI	LQR**FLG**ILT
Badna	EVA**QKP**RIV	FSK**FDL**KAG	NV**CPFG**IA	LL**YIDDIL**IA	EVE**YLG**VEI	LQA**YLG**LLN
HBV	FLV**DKN**PHN	WLS**LDV**SAA	RK**IPMG**VG	FS**YMDDVV**LG	SLN**FMG**YVI	IVG**LLG**FAA
Copia	WTI**TKR**PEN	KYQ**IDY**EET	MR**LPQG**IS	LL**YVDDVV**IA	IKH**FIG**IRI	CRS**LIG**CLM
Intron	VGG**EKG**PYS	TGR**IDD**QEN	GL**TPKT**EF	VR**YADDLL**LG	TVE**FPG**MVI	KFR**NLG**NSI
Retron	TVE**KKG**PEK	ILN**IDL**EDF	NL**LPQG**AP	TR**YADDLT**LS	QRK**VTG**LVI	HHI**FCG**KSS
PMAUP	VYI**PKA**NGK	FPS**VDL**AYL	NG**VPQG**AS	IM**YADDGI**LC	SVK**FLG**LEF	YIQ**VLG**YLP
Archaea	IEI**PKK**SGG	LLE**FDI**KGL	KG**TPQG**GV	ER**YADDSV**IH	KFD**FLG**YTF	WVN**YYG**LFY
HTERT	RFI**PKP**DGL	FVK**VDV**TGA	QG**IPQG**SI	LR**LVDDFL**LV	EDE**ALG**GTA	RRK**LFG**VLR

Fig. 1. RT motif chart for all 30 probes. The six motifs found in all RT sequences comprising the OSM of the queries. The highly conserved residues are in bold. These are the residues counted in the OSM score as described under Methods.