Characterization of Retroid Agents in the Human Genome: An Automated Approach

Marcella A. McClure, Rochelle A. Clinton, Hugh S. Richardson, Vijay A. Raghavan, Crystal M. Hepp, Brad A. Crowther, Angela K. Olsen, Eric F. Donaldson[1] and Aaron R. Juntunen.

Keywords: Retroid agents, retroviruses, retrotransposons, Genome Parsing Suite, human genome

1 Introduction.

Retroid agents are genomes that replicate by reverse transcription of an RNA intermediate. Although once considered to be "junk" DNA, some Retroid agents are implicated in disease via insertional mutagenesis while others have been found to encode proteins essential to mammalian reproduction or to provide regulatory sequences for host cell processes. We have developed new software, the Genome Parsing Suite (**GPS**), to identify and characterize reverse transcriptase (RT) signals in the human genome database (HGD), and to annotate the Retroid Agents that encode them. The **GPS** approach is quite different in concept from **RepeatMasker** [1], a program designed to identify and mask out Retroid Agents in the human genome with consensus DNA for repetitive elements. Recent work conducted on the December, 2001 freeze of the human genome identified 90 L1 LINEs with intact open reading frames (ORF) [2]. The prototype **GPS** provides precise information about the Retroid Agent, including genes present, condition of genes, agent boundaries, location of the agent in the genome etc. The **GPS** utilizes protein rather than nucleotide sequences to screen for the presence of Retroid Agents, thereby providing a deeper query into a genome. Using the **GPS** to analyze 257,278 RT hits retrieved by **BLAST** from the July 2003 freeze of the HGD a total of 95,537 unique RT signals have been identified, and 111 signals are not unambiguously classified to date. As expected, a human LINE sequence pulls out 93% of the unique RT hits. This is less than the reported 500,000 LINEs found by **RepeatMasker**. The estimated number of LINEs in the HGD using **RepeatMasker** does not account for two possibilities regarding small sequence length hits; 1) some may be random, and 2) small hits that are close in proximity actually belonging to the same highly divergent LINE genome. In addition the **RepeatMasker** results include small remnants of untranslated regions (UTRs), while the **GPS** screens for potentially functional Retroid Agents. We have identified 156 LINEs that contain the coding capacity to be potentially functional. All chromosomes except 19 and 21 have full length LINEs without stop codons or frame shifts (table 1). The **GPS** is designed to not only identify highly conserved Retroid agents, but also very distant ancestors. A complete analysis of Retroid information content per genome will also allow the correlation of the position of these agents with higher order genomic feature, e.g., regions of increased gene expression or silencing within CpG islands. The development of this research tool provides the ability to quickly evaluate all Retroid information of a given genome and generate hypotheses regarding the nature of the Retroid Agent landscape in genomes from all three domains of life. By populating the **GPS** with protein sequences representing all known RT genes not only is a complete analysis of the major Retroid Agents present in the any genome feasible, but low frequency RT genes and in some cases “cryptic” retroviral genomes may be discovered.

2 Software and files.

The **GPS** can be populated with any set of phylogenetically distributed protein sequences and the corresponding ordered-series-of-motifs (OSM) representing functional or structurally important amino acids to search, annotate and assess probable function of new members of a protein family in any organismal genome database. Tests were performed to determine which of three external search methods provide the greatest amount of raw data to analyze. **BLAST** [3] with the PAM70 matrix, retrieved more significant RT hits than **WU-BLAST** [4] using the same matrix, or **RepeatMasker** driven by **Cross_Match** [1] using our in-house RT gene libraries. Twenty-one representative RT sequences were used to generate the data presented in table 1. The GPS is designed to initially evulate the RT retrieved hits and then search the surrounding genomic sequences for other genes encoded by Retroid agents. Raw **BLAST** hits are evaluated for the presence of the RT OSM. Unique hits are determined by comparison of: 1) hits in the same reading frame by multiple probes to the same location, and 2) compound hits at the same location from multiple reading frames. In stage one, a hit with a blast score that is at least 10% greater than all others is determined to be the unique hit. All hits meeting this 10% criterion are reevaluated in the next procedure. In stage two, if more than one probe retrieves a hit to the same location, the unique hit is

All at: Montana State University-Bozeman, Dept. of Microbiology and the Center for Computational Biology, 109 Lewis Hall, Bozeman MT 59717. Email: mars@parvati,msu.montana.edu

[1] University of North Carolina, Chapel Hill, NC 27514. Email: eric_donaldson@med.unc.edu

determined by a score of at least 50% greater than all others. Ambiguity arises when the range of scores are all within 50% of one another. The unique hit will then be determined from among these ambiguities at the Retroid agent gene component analysis stage of the GPS. This procedure removes redundancy of retrieved hits due to cross coverage by multiple queries and solves the problem of multiple small hits retrieved by a given query which represents a distantly related RT gene. Failure to account for these small hits results in each hit being scored as a unique hit, thereby overestimating the number of potential RT genes and Retroid genomes within a host genome. Based on the OSM scores for potential RTs, corresponding Retroid agent genome sequences are extracted from the organismal database. The **GPS** then analyzes these potential Retroid agents for the expected genes given the RT classification based on our gene component libraries. In addition, each potential Retroid agent sequence will be evaluated by all component libraries to screen for recombination events. Future extensions to the **GPS** will include the ability to populate the system with nucleotide sequences using our in-house Retroid nucleotide libraries and a graphical display of all results.

3 Figures and tables.

Chr	Unique	Unclass	Full	Perfect
1	6722	4	127	17
2	7793	4	167	12
3	6858	5	133	11
4	7136	14	154	10
5	6361	7	139	12
6	5793	8	122	10
7	4944	3	86	9
8	4817	2	109	7
9	3842	3	71	4
10	4029	2	74	6
11	4486	9	95	9
12	3890	6	83	6
13	3167	2	42	3
14	2850	3	51	3
15	2520	5	35	5
16	1624	2	26	7
17	1458	1	19	1
18	2448	1	46	5
19	1222	6	18	0
20	1374	3	26	3
21	986	1	8	0
22	745	3	7	2
x	9091	4	187	13
y	1381	13	23	1
Totals	95537	111	1848	156

Table 1. Chr is chromosome number. Unique indicates all unique RT signals, Unclass includes all RTs that could not be unambiguously classified by query RTs. Full indicates full length LINEs and Perfect indicates LINEs that are full length and contain no frame-shifts or stop codons.

4 References.

[1] Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html

[2] Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V. & Kazazian, H. H., Jr. (2003)*Proc Natl Acad Sci U S A* 100, 5280-5.

[3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997)*Nucleic Acids Res* 25, 3389-402.

[4] Gish, W. 1996-2003 http://blast.wustl.edu