Poster A-12
Annotating the Eukaryotic Retrome

Authors:
Marcella A. McClure (*MSU*)
Crystal Hepp (*MSU*)
Holly Basta (*MSU*)
Tara Swope (*MSU*)
Steve Martin (*MSU*)
Anoop Winston (*MSU*)
Alex Buzak (*MSU*)

Short Abstract: The Genome Parsing Suite software identifies and annotates all Retroid agents (reverse transcriptase encoding genetic elements) in a given genome, thereby creating the data necessary to map the human and Eukaryotic Retromes. The data presented here are from the complete and ordered genomes of human, chimp, rat and mouse.

Long Abstract:

Assessing the evolutionary impact of Retroid agents (i.e., all genetic elements encoding the reverse transcriptase) on the Eukaryotic genome is only possible when each and every instance of these agents is known in a wide distribution of organisms. The Genome Parsing Suite (GPS) software1 was created to identify and annotate all Retroid agents in a given genome, thereby creating the data necessary to map the human and Eukaryotic Retromes, (i.e., all instances of Retroid agents).
The GPS is a radically different approach from Repeat Masker for identifying Retroid agents. The GPS uses RT protein sequences as WUBLAST2 queries to identify potential RTs and then proceeds to construct the Retroid Agent that encodes the RT. This method provides the user with all relevant DNA sequences (e.g., LTRs, genes etc.), all genomic positions, a measure of “goodness” of both genome and enzyme functional integrity of each potential Retroid agent, and a variety of informational tables. A GPS analysis of the human genome generates over a gigabyte that will be further analyzed to assess a variety of functional and mechanistic hypotheses regarding the variety of roles these agents play in humans.
The tracking of all RT instances is vital to understanding the co-evolution of these Retroid and human genomes and how they co-regulate each other in a variety of roles including reproduction, development and disease. Some Retroid agents provide regulatory sequences for the host cell processes, maintain telomeres, repair damaged chromosomes, and carry genetic information within and between organisms. Specifically in humans, exogenous retroviruses cause AIDS by HIV, and human T-cell leukemia by HTLV I and II. Human endogenous retroviruses (HERVs) are associated with breast cancer, testicular tumors, insulin dependent diabetes, multiple sclerosis, rheumatoid arthritis, schizophrenia, and systemic lupus erythematosus. Others Retroid agents have been demonstrated to be directly responsible for muscular dystrophy, Alport Syndrome-Diffuse Leiomyomatosis, and chronic

granulomatous disease. On the other hand, WHERV encodes syncytin, a protein expressed in the human uterus that facilitates trophoblast fusion to the syctiotrophoblast, allowing placenta formation. Other endogenous retroviruses provide this function (two different ones in mice), indicating that retroviruses have entered the mammalian genome multiple times and they have been selected for beneficial functions. Elucidating the frequency and mechanism(s) of genomic mutualism is a fundamental question in molecular evolution.

The breath of the trans-organismal study at this time includes the complete and ordered genomes of Homo sapiens (human), Pan troglodytes (chimp), Canis familiaris (dog), Rattus norvegicus (rat), Mus musculus (mouse), Danio rerio (zebrafish), Tetraodon nigroviridis (pufferfish), Anopheles gambiae (mosquito), Drosophila melanogaster (fruit fly), Apis mellifera (honey bee), Arabidopsis thaliana (mouse-ear cress), Oryza Sativa ssp. japonica cultivar (rice), Dictyostelium discoideum (slime mold) and Caenorhabditis elegans (nematode). While the genomes of other organisms have been sequenced the data are not ordered into chromosomes or processed for redundancy; when these genomes are truly complete they will be added to this ongoing analysis. The cross-correlation of gigabytes of data generated by the GPS trans-organsimal analysis necessitates further software development.

Using RT protein sequences as functional indicators of potentially active Retroid agents, we present the first detailed study of all Retroid agents in H. sapiens, P. troglodytes, R. norvegicus, and M. musculus. The GPS analysis of all WU-tBLASTn results generated from both inter- and intra-species specific RT sequence queries representing the evolutionary divergence of this gene identified 136,645 Human, 130,324 Chimp, 109,290 Rat, and 104,159 Mouse unique hits. The stage 2 GPS assessment of all unique hits found 10,557 Human, 4,105 Chimp, 8,133 Rat, and 11,882 Mouse, full-length Retroid genome. In the count of potentially active, full-length Retroid agents the read through of a stop codon or frame-shift by transitional recoding must be taken into account. While cellular translation recoding is rare, it well known to occur in Retroid agents, (e.g., mammalian LINEs, Gypsy, and Copia-like agents and HIV). In this study, therefore, Retroid genomes that are full-length and without errors, as well as those with one stop codon or frame shift are counted as potentially active In human, while there are 158 error-free Retroid genomes, and 200 potentially active ones via translational recoding. In chimp there are no potentially active Retroid agents. In rat only 2/8 potentially active Retroid agents are perfect, and in mouse 241/927 are perfect.

From these data it is noted that the higher overall genome sequence identity does not necessarily mean that two organisms will have a similar number of Retroid agents when compared to related organisms. While this is well known for insects it is rather surprising in mammals. H. sapiens and P. troglodytes are 98% identical. P. troglodytes, however, has far fewer complete Retroid agents and no potentially active ones. The same situation occurs for M. musculus and R. norvegicus, which diverged from each other 12-24 mya. The rat possesses a much lower number of complete Retroid genomes. Considering that primates and rodents diverged from each other at the same time (75 mya), this comparative genomics approach indicates that P. troglodytes and R. norvegicus may possess a mechanism for suppressing the accumulation of active Retroid agents. In the case of P. troglodytes, this could possibly account for its smaller genome size when compared to human. Retroid agents have long been considered possible agents of speciation and now the data are available to addresses this idea more directly. Retroid agents have long been considered possible agents of speciation and now the data are available to addresses this idea more directly, however, the commensual relationship between host and Retroid genomes has also played a role in the Eukaryotic landscape.

1 McClure, M.A., Richardson, H.S., Clinton, R.A., Hepp,C.M., Crowther, B.A., Donaldson, E.F.,(2005) "Automated characterization of potentially active Retroid agents in the human genome". Genomics 85 (2005) 512-523.
2 W.R. Gish, http://blast.wustl.edu, 1996-2002