Searching the Murine Genome for Endogenous Retroviral Genes

Kristin Henry and Ty Davis

Abstract

When a retrovirus becomes established in germ-line DNA, it becomes endogenous. At least eight percent of the human genome has been found to be of retroviral origin. Endogenous retroviral sequences have been found in many species. In our study, we search the mouse (murine) subset of the EST genome database for protein sequences from known murine retroviruses. We use BLAST search results with Phylogenetic Trees and Self-Organizing Maps to discover patterns and familial relationships between the sequences.

Introduction

HIV may be the most famous retrovirus, but it's not alone. The complete genomes of at least 60 different retroviruses are currently listed at The National Center for Biotechnology Information (NCBI).

What makes a retrovirus different from other viruses, is that it inserts it's code into the nuclear DNA of the infected cell. They also infect germ cells, and can be passed down through the generations along with the genes for eye color and all the details that make up our bodies. Researchers have found that at least eight percent of the human genome consists of retroviral content. These endogenous retroviral sequences may have a role in how our dna functions.

In our project, we used BLAST and Biopython libraries, to search for retroviral sequences within the mouse genome. We then looked at these sequences, and clustered them based on their similarities with Phylogenetic Tree and basic Self-Organizing Map clustering.

Methods

Searching the Mouse Genome for MERVs
- started with genomes of murine (mouse) **retroviruses**
- looked for conserved regions of three retroviral proteins (env, gag, pol)
- used conserved regions as probes "search terms" to search for matches in mouse genome, with BLAST

Searching All Genomes:
- used conserved region of mouse retroviral pol gene as probe
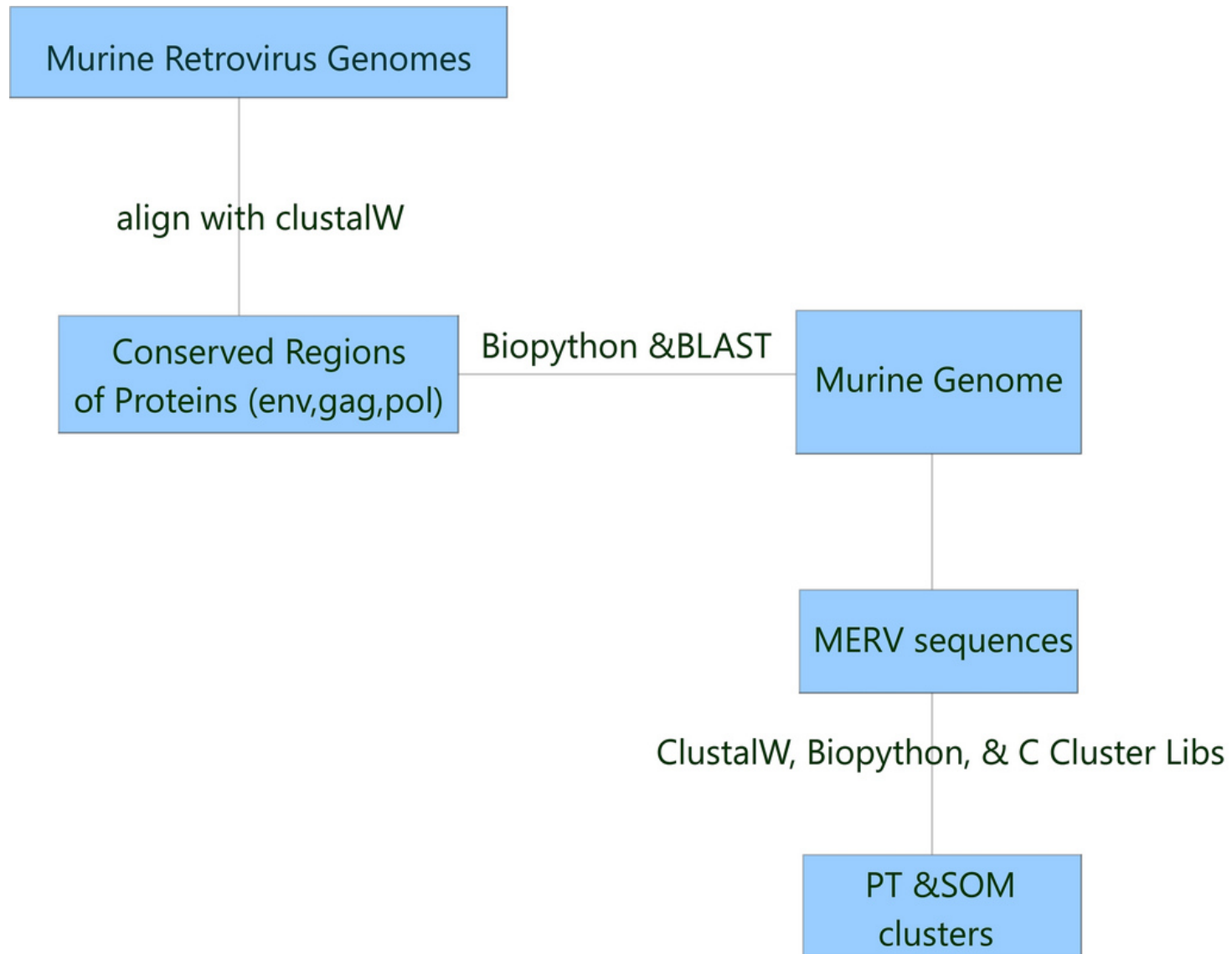- PSI-BLAST entire NCBI genetic database with this probe

figure 1. Diagram of the project's data flow

Results

Conserved sequences from murine retroviral genes:

Env

MSTLSKSKDKIDPPLILFLAAASSPHQVYNITWEVTNETASPLWPL
DLCLHWDEPCNTANRHGFYVCPGPSCGGPSYCASWGCETTGRAYWK
PSSSWDITVDNNTSNAQCKNCNPLAIQFTNAGKQVTWHWGLRLYSG

GAG

MGQTVTTPLSLTLEHWDVQRANQVEVKKRRWVTFCSAEWPTFGVGW
PQDGTFNDIIQVKKVFSPGPHGHPDQVPYIVTWEAIAEPPPWVKPF
VPKLLPPSAPLPPPPQPPPQSALYPALTPPIKPKPKPQVLPDSGGP
LIDLLTEDPPPYEGPPSSDGDGDEAAPTPPSPMVSRLRGKREPPAS
QAFPLGQFSSSDLYNWKNNNPSFSEDPGKLTALIESVLTHQPTWDD
CQQLLGTLLTGEEKQRVLLEARKAVRGEDGRPTQLPNEIEAFPLER
PDWDYTTRGRNHLVYRQLLLAGLQNAGRSPTNLAKVKGITQGPNES
PSAFLERLKEAYRRYTPYDPEDPGQETNVSMSFIWQSAPDIGRKLE
RLEDLKSKTLGDLVREAEKIFNKRETPEEREERIKRETEEKEERRR
AEDEQKEKERDRRRQREMSKLLATVVSGQRQDRQGGERRRPQLDKD
QCAYCKEKGHWAKDCPKKPRGPRGPRPQASLLTLDD

POL

TLDDQGGQGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVLTQNPGP
LSDKSAWVQGATGGKRYRWTTDRRVHLATGKVTHSFLHVPDCPYPL
LGRDLLTKLKAQIHFEGSGAQVVGPMGQPLQVLTLNIEDEYRLHET
SKPDVPLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPLKATSTPVS
IKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKPGT
NDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLK
DAFFCLRLHPTSQSLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTL
FDEALHRDLADFRIQHPDLILLQYVDDLLLAATSELDCQQGTRALL
QTLGDLGYRASAKKAQICQKQVKYLGYLLKEGQRWLTEARKETVMG
QPTPKTPRQLREFLGTAGFCRLWIPGFAEMAAPLYPLTKTGTLFKW
GPDQQKAYQEIKQALLTAPALGLPDLTKPFELFVDEKQGYAKGVLT
QKLGPWRRPVAYLSKKLDPVAAGWPPCLRMVAAIAVLTKDAGKLTM
GQPLVILAPHAVEALVKQPPDRWLSNARMTHYQALLLDTDRVQFGP
IVTLNPATLLPLPEEGLQHDCLDILAEAHGTRPDLTDQPLPDADHT
WYTDGSSFLQEGQRKAGAAVTTETEVIWAKALPAGTSAQRAELIAL
TQALKMAEGKKLNVYTDSRYAFATAHIHGEIYRRRGLLTSEGKEIK
NKDEILALLKALFLPKRLSIIHCPGHQKGNRAEARGNRMADQAARE
VATRETPETSTLLIENSAPYTREHFHYTVTDIKDLTKLGATYDSAK
KCWVYQGKPVMPDQFTFELLDFLHQLTHLSFSKTKALLERSYSPYY
MLNRDRTLKDITETCKACAQVNASKSAVKQGTRVRGHRPGTHWEID
FTEVKPGLYGYKYLLVFVDTFSGWVEAFPTKKETAKVVTKKLLEEI
FPRFGMPQVLGTDNGPAFVSKVSQTVADLLGVDWKLHCAYRPQSSG
QVERMNRTIKETLTKLTLATGSRDWVLLLPLALYRARNTPGPHGLT
PYEILYGAPPPLVNFPDPDMAKVTHNPSLQAHLQALYLVQHEVWRP
LAAAYQEQLDRPVVPHPFRVGDTVWVRRHQTKNLEPRWKGPYTVLL
TTPTALKVDGIAAWIHAAHVKAADTRIEPPSESTWRVQRSQNPLKI
RLTRGTS

gi|254939518|ref|NP_001157094. DPSYFVFLYLTKDSRFLTLNIMNKTTSEVWLIDGLSPWDPPVLIQKRIHGMLYYV
gi|254939520|ref|NP_666096.3|p DPSYFVFLYLTKDSRFLTLNIMNKTTSEVWLIDGLSPWDPPVLIQKRIHGMLYYV

gi|62000680|ref|NP_001013845.1 PSCGVPSCCRPVCCQTTCRPSCGVSSCCRPVCCQTTCRPSCGVSSCCRPVCCQT
gi|148839318|ref|NP_083163.3|s LSTLSITGDGLDSSNLKITLEANSQSVSVQGIRFTPGPPGPKDLCFTWRE
gi|254281307|ref|NP_766624.3|p PVLVVIGAGSSNSYGVYTIRDQENRTSLELCWFQKEPALYATVHGYFLV
gi|345199290|ref|NP_001230832. YECKQCGKAFAVSYTLQIHNRTHTGEKPYECKQCGKAFAQSSNLRIHKQTHTGER
gi|147904569|ref|NP_082180.2|c GLDEGEPAGSWDLTVEGRAAHSQCAYSNP
gi|33239256|ref|NP_666833.1|ol TLARMYLKPSSSYDLGQDRQVA

gi|77812697|ref|NP_035782.3|ti PEIKLSWYKGTEKLEPSNKYEISIDGDRHTLRVKNCQP
gi|67763822|ref|NP_659208.2|he KKHGFKLGPTPKTLGSSLENAWGSGRAGPSYSRPA
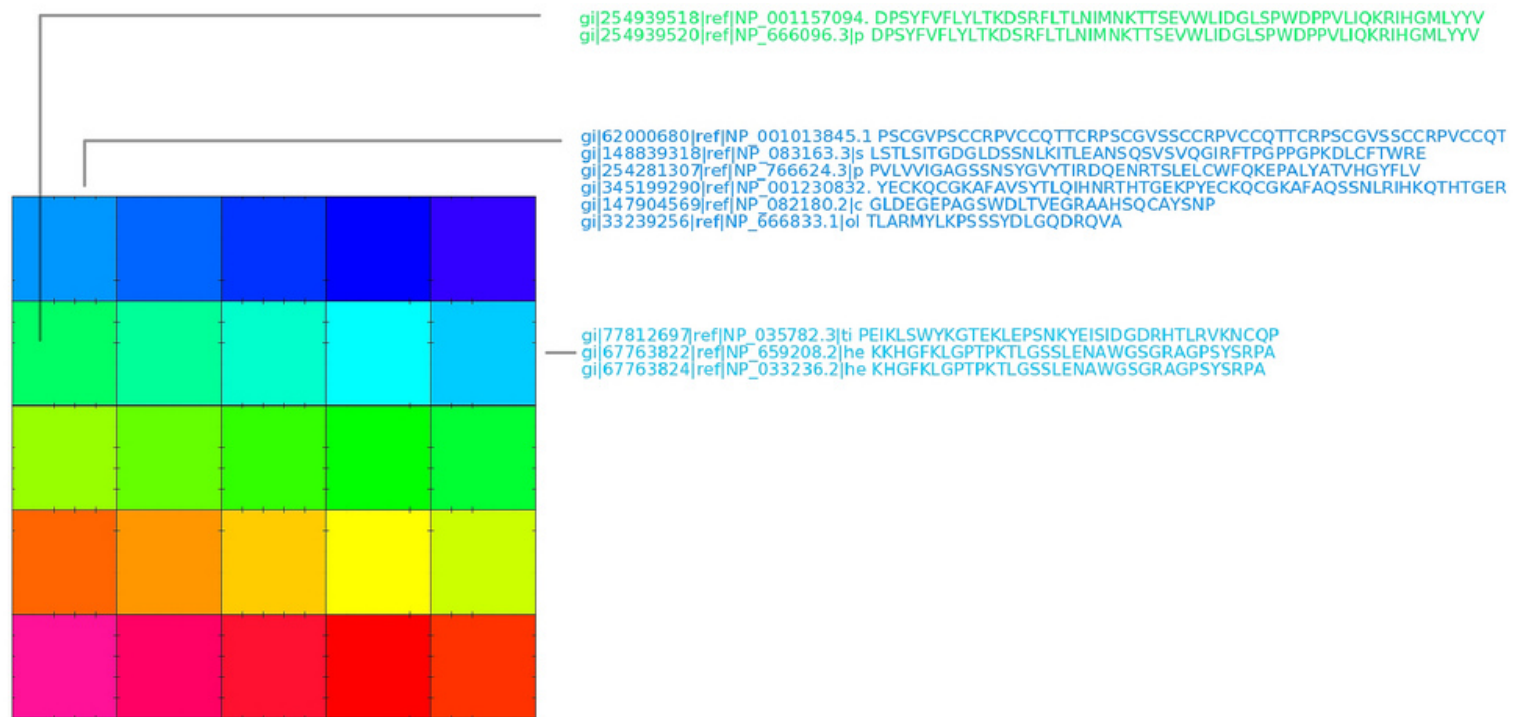gi|67763824|ref|NP_033236.2|he KHGFKLGPTPKTLGSSLENAWGSGRAGPSYSRPA

figure 2. SOM clustering diagram of retroviral *env* genes found in mouse genome

figure 3. Phylogenetic Tree clustering diagram of retroviral *env* genes found in mouse genome
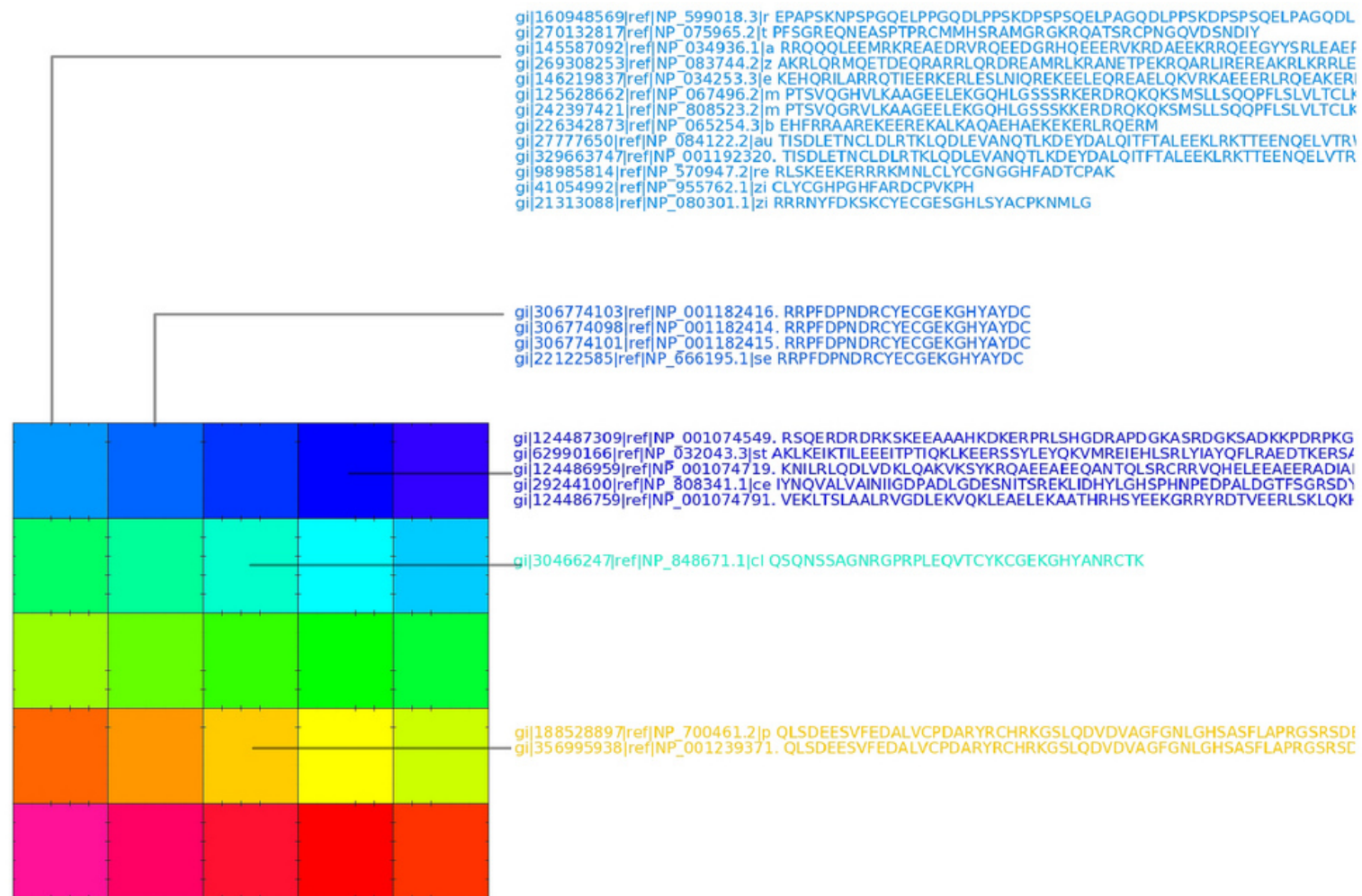
gi|160948569|ref|NP_599018.3|r EPAPSKNPSPGQELPPGQDLPPSKDPSPSQELPAGQDLPPSKDPSPSQELPAGQDL
gi|270132817|ref|NP_075965.2|t PFSGREQNEASPTPRCMMHSRAMGRGKRQATSRCPNGQVDSNDIY
gi|145587092|ref|NP_034936.1|a RRQQQLEEMRKREAEDRVRQEEDGRHQEEERVKRDAEEKRRQEEGYYSRLEAEF
gi|269308253|ref|NP_083744.2|z AKRLQRMQETDEQRARRLQRDREAMRLKRANETPEKRQARLIREREAKRLKRRLE
gi|146219837|ref|NP_034253.3|e KEHQRILARRQTIEERKERLESLNIQREKEELEQREAELQKVRKAEEERLRQEAKERI
gi|125628662|ref|NP_067496.2|m PTSVQGHVLKAAGEELEKGQHLGSSSRKERDRQKQKSMSLLSQQPFLSLVLTCLK
gi|242397421|ref|NP_808523.2|m PTSVQGRVLKAAGEELEKGQHLGSSSKKERDRQKQKSMSLLSQQPFLSLVLTCLK
gi|226342873|ref|NP_065254.3|b EHFRRAAREKEEREKALKAQAEHAEKEKERLRQERM
gi|27777650|ref|NP_084122.2|au TISDLETNCLDLRTKLQDLEVANQTLKDEYDALQITFTALEEKLRKTTEENQELVTRV
gi|329663747|ref|NP_001192320. TISDLETNCLDLRTKLQDLEVANQTLKDEYDALQITFTALEEKLRKTTEENQELVTR
gi|98985814|ref|NP_570947.2|re RLSKEEKERRRKMNLCLYCGNGGHFADTCPAK
gi|41054992|ref|NP_955762.1|zi CLYCGHPGHFARDCPVKPH
gi|21313088|ref|NP_080301.1|zi RRRNYFDKSKCYECGESGHLSYACPKNMLG

gi|306774103|ref|NP_001182416. RRPFDPNDRCYECGEKGHYAYDC
gi|306774098|ref|NP_001182414. RRPFDPNDRCYECGEKGHYAYDC
gi|306774101|ref|NP_001182415. RRPFDPNDRCYECGEKGHYAYDC
gi|22122585|ref|NP_666195.1|se RRPFDPNDRCYECGEKGHYAYDC

gi|124487309|ref|NP_001074549. RSQERDRDRKSKEEAAAHKDKERPRLSHGDRAPDGKASRDGKSADKKPDRPKG
gi|62990166|ref|NP_032043.3|st AKLKEIKTILEEEITPTIQKLKEERSSYLEYQKVMREIEHLSRLYIAYQFLRAEDTKERSA
gi|124486959|ref|NP_001074719. KNILRLQDLVDKLQAKVKSYKRQAEEAEEQANTQLSRCRRVQHELEEAEERADIAI
gi|29244100|ref|NP_808341.1|ce IYNQVALVAINIIGDPADLGDESNITSREKLIDHYLGHSPHNPEDPALDGTFSGRSDY
gi|124486759|ref|NP_001074791. VEKLTSLAALRVGDLEKVQKLEAELEKAATHRHSYEEKGRRYRDTVEERLSKLQKF

gi|30466247|ref|NP_848671.1|cl QSQNSSAGNRGPRPLEQVTCYKCGEKGHYANRCTK

gi|188528897|ref|NP_700461.2|p QLSDEESVFEDALVCPDARYRCHRKGSLQDVDVAGFGNLGHSASFLAPRGSRSDI
gi|356995938|ref|NP_001239371. QLSDEESVFEDALVCPDARYRCHRKGSLQDVDVAGFGNLGHSASFLAPRGSRSD

figure 4. SOM clustering diagram of retroviral *gag* genes found in mouse

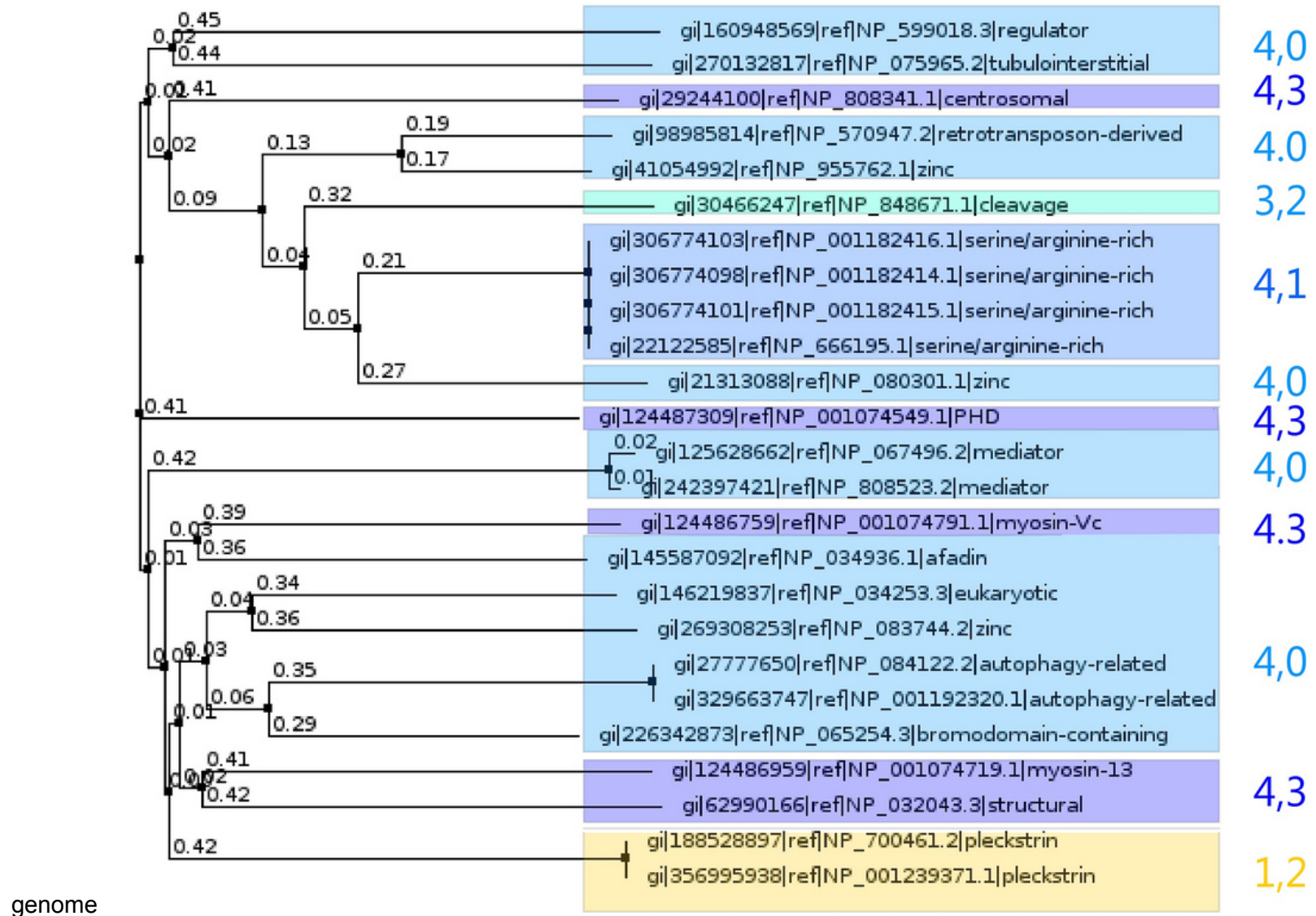figure 5. Phylogenetic Tree clustering diagram of retroviral *gag* genes found in mouse genome
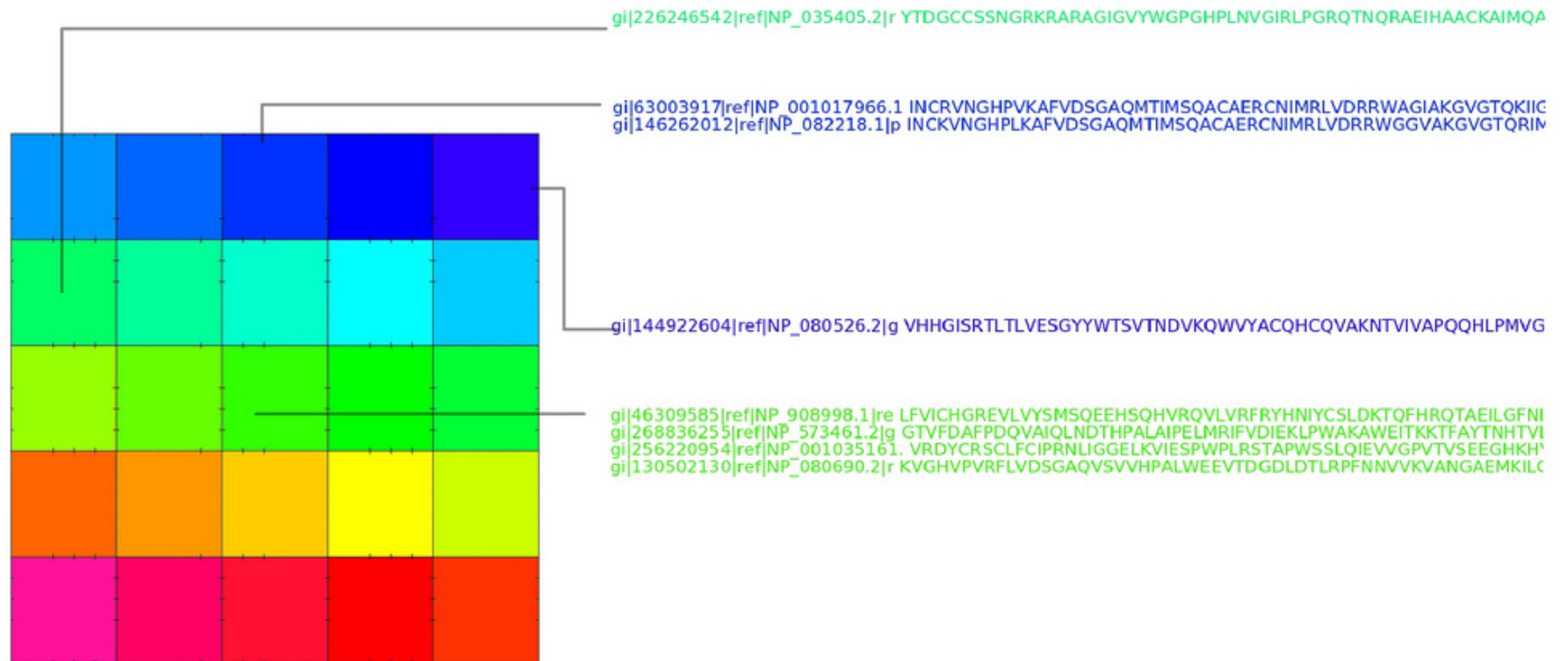
gi|226246542|ref|NP_035405.2|r YTDGCCSSNGRKRARAGIGVYWGPGHPLNVGIRLPGRQTNQRAEIHAACKAIMQA

gi|63003917|ref|NP_001017966.1 INCRVNGHPVKAFVDSGAQMTIMSQACAERCNIMRLVDRRWAGIAKGVGTQKIIG
gi|146262012|ref|NP_082218.1|p INCKVNGHPLKAFVDSGAQMTIMSQACAERCNIMRLVDRRWGGVAKGVGTQRIN

gi|144922604|ref|NP_080526.2|g VHHGISRTLTLVESGYYWTSVTNDVKQWVYACQHCQVAKNTVIVAPQQHLPMVG

gi|46309585|ref|NP_908998.1|re LFVICHGREVLVYSMSQEEHSQHVRQVLVRFRYHNIYCSLDKTQFHRQTAEILGFNI
gi|268836255|ref|NP_573461.2|g GTVFDAFPDQVAIQLNDTHPALAIPELMRIFVDIEKLPWAKAWEITKKTFAYTNHTVI
gi|256220954|ref|NP_001035161. VRDYCRSCLFCIPRNLIGGELKVIESPWPLRSTAPWSSLQIEVVGPVTVSEEGHKH\
gi|130502130|ref|NP_080690.2|r KVGHVPVRFLVDSGAQVSVVHPALWEEVTDGDLDTLRPFNNVVKVANGAEMKILC

figure 6. SOM clustering diagram of retroviral *pol* genes found in mouse genome

figure 7. Phylogenetic Tree clustering diagram of retroviral *pol* genes found in mouse genome

- Mouse retroviral proteins found in genomes of several species:
  - porcine (pig), mus musculus (house mouse), sus scrofa (wild boar), sus scrofa domesticus (domestic pig), baboon, Feline, Gibbon, mus dunni (pygmy mouse), and even the Human retroviruses.
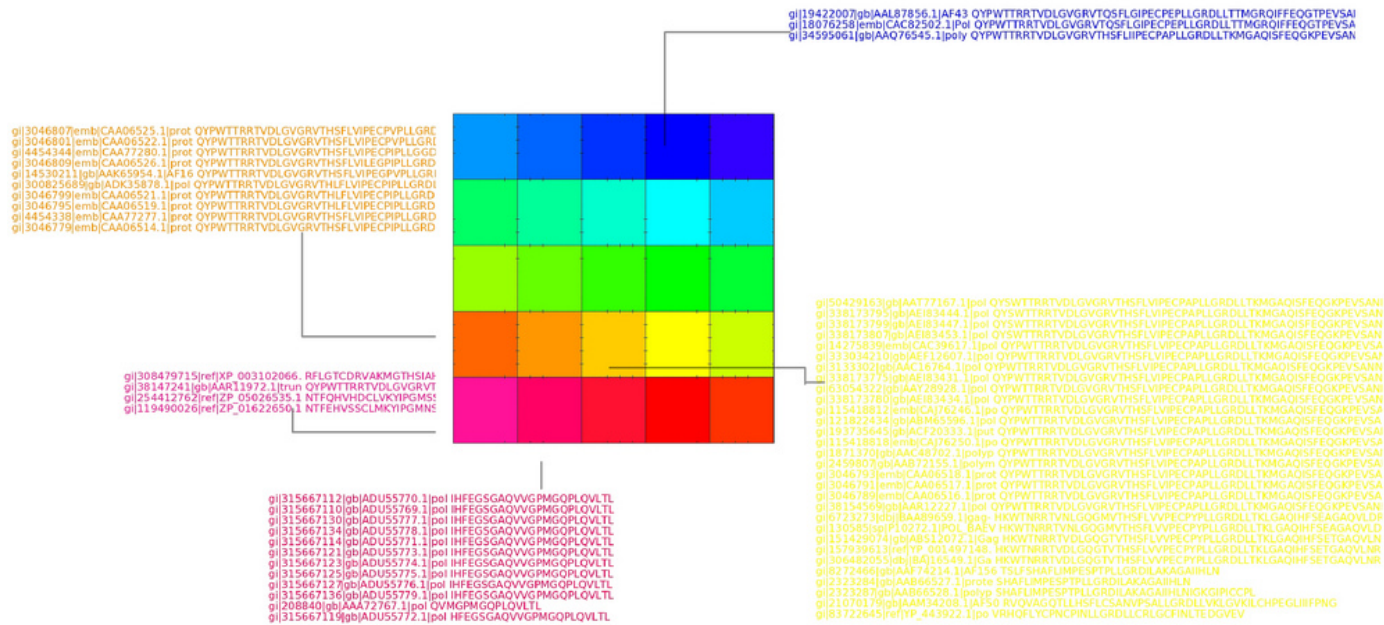

figure 8. SOM clustering diagram of retroviral *pol* genes found genomes across species

Conclusion

We found:

- retroviral sequences in the mouse genome
- mouse retrovirial sequences in other species (pig, cow, human, and others)
- SOM clustering was in accordance with Phylogenetic Tree clustering

References:

Self-Organizing Map-Based Discovery and Visualization of Human Endogenous Retroviral Sequence Groups
by Merja Oja, and Goran O. Sperber, Jonas Blomberg, Samuel Kaski
International Journal of Neural Systems, Vol 15, No 3 (2005) 163-179
([abstract](#))

Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins
by Nathalie de Parseval, Vladimir Lazar, Jean-Francois Casella, Laurence Benit, and Thierry Heidmann
Journal of Virology. 2003 Oct; Vol 77(No 19):10414-22.
([abstract](#))

Retroviral Genomes
http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=11632