# Should we look for

presence and conservation overall, and of these features:

* protease cleavage sites that would release CA as an individual protein?
* conservation within mammals of the bit that resembles capsid
* if we want this thing to recruit RNAs, do we require NC homology too?

analyze PNMA family evolution?

rates in orthodb?

paml primates?

model 0 dN/dS

Will is looking at expression using the Human Protein Atlas: he doesn’t care where they are expressed, as long as there is evidence for expression.

# Analysis I want to do

look at phyloP scores across genes (use Bioc?) – just ORF? or include surround?

use bioc orthodb?

synteny: is Genomicus parseable? (perhaps in combo with Bioc Ensembl data)

in-frame alignments –

get 12-field bed, each transcript separately

get multiz mafs

maf2fasta\_JY.bioperl

# Make sure I remember

possibility of programmed frameshift

# parsing UCSC 100way alignments

## maf format

(not readable by Geneious)

http://genome.ucsc.edu/FAQ/FAQformat.html#format5

s = sequence lines

i = information about context of sequence lines immediately preceding

e = information about empty parts of the alignment block

Kent\_tools has various maf tools, but no obvious fasta conversion

module load LAST

maf-convert --help

can convert to axt, blast, blasttab, chain, gff, html, psl, sam, tab

# PNMA family

Ensembl supertree contains two gene trees:

ENSGT01000000214513 (ZCCHC12, ZCCHC18 and some likely ERVs from amphibians, some ERVs/genes/not sure from fish) and

ENSGT01010000222436 (MOAP, CCDC8, and PNMAs 1,2,3,5,6A,6F,6E,8A,8B,8C)

# Retrovirus gag (mostly from wikipedia)

## HIV gag:

Group-specific antigen, or gag, is the polyprotein that contains the core structural proteins of an Ortervirus (except Caulimoviridae). It was named as such because scientists used to believe it was antigenic. Now it is known that it makes up the inner shell, not the envelope exposed outside. It makes up all the structural units of viral conformation and provides supportive framework for mature virion.

~500a.a.

programmed frameshift

All orthoretroviral gag proteins are processed by the protease (PR or pro) into MA (matrix), CA (capsid), NC (nucleocapsid) parts, and sometimes more. If Gag fails to cleave into its subunits, virion fails to mature and remains uninfective.

MA (matrix): is responsible for targeting Gag polyprotein to the plasma membrane via interaction with PI(4,5)P2 through its highly basic region (HBR). HIV MA also makes contacts with the HIV trans-membrane glycoprotein gp41 in the assembled virus and, indeed, may have a critical role in recruiting Env glycoproteins to viral budding sites.

CA The p24 capsid protein (CA) is a 24 kDa protein fused to the C-terminus of MA in the unprocessed HIV Gag polyprotein. After viral maturation, CA forms the viral capsid. CA has two generally recognized domains, the C-terminal domain (CTD) and the N-terminal domain (NTD). The CA CTD and NTD have distinct roles during HIV budding and capsid structure.

NC The HIV nucleocapsid protein (NC) is a 7 kDa zinc finger protein in the Gag polyprotein and which, after viral maturation, forms the viral nucleocapsid. NC recruits full-length viral genomic RNA to nascent virions.

## gag in other retroviruses:

The gag gene of Spumaretrovirinae (e.g. P14349) and Metaviridae (e.g. Q86TG7) only have a recognizable nucleocapsid part. It also lacks a myristoylation sequence.[7]

The Spumaretroviral (SV) gag is related to orthoretroviral gag, as structural work has shown that part of the N-terminal domain shares functional and structural homology with the typical capsid protein.[8] The SV gag is not processed like the orthoretrovieral gag; only a tiny 3kDa cut at the C-terminal is requried, and other cleavage sites are generally inefficient.[9]

The Metaviral (MV, Ty3/gypsy) gag, too, is known to have a structurally homologous capsid protein. Each capsid is assembled from 540 proteins. Unlike orthoretroviral CA proteins, it does not require dramatic maturation.[10] The animal Activity-regulated cytoskeleton-associated protein (ARC) gene is repurposed from the metaviral gag.[11] This gene is responsible for transporting mRNA among neural cells, a key part of neuroplasticity. It has independently arose in Tetrapoda and Drosophila.[12]

Caulimoviridae members rarely get a gag assignment to its capsid-containing ORF, but the CP-PRO-POL layout does show analogy with the canonical gag-pol setup. Whether the parts stick together into a polyprotein depends on the genus.

## readthrough site

Gag/pol translational readthrough site (or Retroviral readthrough element) is a cis-regulatory element found in retroviruses.[1] The readthrough site facilitates the mechanism of translation readthrough of the stop codon at the gag-pol junction producing the gag and pol fusion protein in certain retroviruses. Retroviruses whose gag and pol genes are in the same reading frame often depend upon approximately 5% read-through of the gag UAG termination codon to form the gag-pol polyprotein. This readthrough is usually dependent on a pseudoknot located eight nucleotides downstream of the stop codon (UAG). Sequence conservation is found in the second pseudoknot loop.

# Viral phylogeny

Ortervirales is an order that contains all accepted species of single-stranded RNA viruses that replicate through a DNA intermediate (Group VI) and all accepted species of double-stranded DNA viruses (except Hepadnaviridae) that replicate through an RNA intermediate (Group VII). [1][2] The name is derived from the reverse of retro.

There are five families in this order:

Family Belpaoviridae — LTR retrotransposon, Bel/Pao family

Family Metaviridae — LTR retrotransposon, Ty3/gypsy family (incl sushi-ichi)

Family Pseudoviridae — LTR retrotransposon, Ty1/copia family

Family Retroviridae — Retroviruses, e.g. HIV

Family Caulimoviridae — dsDNA-RT virus infecting plants