# Doing right now

add alngroup to each seq name

concatenate all PNMA seqs and outgroup

make tree to make sure I have groups right

summarize PNMA results (genes/pseudogenes)

get PNMA model 0 dN/dS

# 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

# Alignment notes

turns out that PNMA5 and PNMA3 DO show evidence for programmed frameshift. I am not looking at the whole gene in my alignments!!!

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009743/>

**PNMA1** keeping both Dasypus\_novemcinctus seqs – only 95% identical

**PNMA3** ignoring one isoform of PNMA3, as it has an unconserved last exon: PNMA3\_Homo\_sapiens\_NM\_001282535

**PNMA5**: Mouse has a longer ORF than human in 3’ direction – it may be the full-length protein: not sure if it still frameshifting. It has lost the stop codon that human and rat and most (but not all) other mammals have. Some patches within the extension DO show conservation.

**PNMA6E/F**: keeping both horse seqs (97% identical), both Chinese hamster seqs (97% identical), both armadillo seqs (96% identical). Seems to be one of those situations where two genes maintain similarity by gene conversion?

**PNMA8A** – removed Chinese hamster seq that appears to be two half-genes fused inappropriately LOC100762167\_Cricetulus\_griseus\_XM\_035461250

last exon messed up

**PEG10** – removed a second dodgy rat copy from Peg10\_Rattus\_norvegicus\_XM\_008762737 – has various Ns. processed pseudogene??

**RTL1** – mouse has weird in-frame expansions

**RTL3** – removed a copy of extra Chinese hamster copies, probably alleles/splices, one appears to be a pseudogene

**RTL8A** – removed a human isoform that had no matches to the small last exon: RTL8A\_Homo\_sapiens\_NM\_001134321

Kept 2 seqs from these species, as the pairs are different enough: armadillo, rat

# Results notes

**CCDC8** – have good alignment. In placental mammals, but not marsupials.

**ZCCs 12/18**

12 and 18 diverged from each other after the Afrotheria+Xenarthra – versus other placental mammals split. But rodents have two versions of the 12-like gene, and 0 versions of the 18-like gene. Armadillo (Xenarthra) also has two genes, but they’re basal to the 12/18 split. Could be a gene-conversion-like situation.

outgroup is not helpful for the tree

# Model 0 dN/dS results

ARC dN/dS = 0.0322 (whole gene, 10 placental mammals, dS tree len 5.4)

Arc1: 0.0926 (whole gene, 8 Drosophila species, dS tree length is ~3.7)

Arc2: 0.0383 (whole gene, 8 Drosophila species, dS tree length is ~3.7)

CCDC8 (PNMA family): 0.14491 (whole gene minus poorly aligned chunks, 14 placental mammals, dS tree len 3.3)

ZCCHC12 (PNMA family): 0.252 (whole gene, 10 placental mammals not including rodents/Afrotheria/Xenarthra, dS tree len 1.3)

ZCCHC18 (PNMA family): 0.5351 (whole gene, 10 placental mammals not including rodents/Afrotheria/Xenarthra, dS tree len 1.1)

MOAP1 (PNMA family): 0.29261 (whole gene, 9 placental mammals not including pseudogenes, dS tree len 1.37)

PNMA1 (PNMA family): 0.06165 (whole gene, 14 placental mammals, dS tree len 1.8361)

PNMA2 (PNMA family): 0.12210 (whole gene, 15 placental mammals, dS tree len 3.43)

PNMA3 (PNMA family): 0.20187 (whole gene, 10 placental mammals, dS tree len 2.28)

# Papers

## dArc1 in Dmel: Ashley, Cell, 2018

various functional studies on dArc1

intergenic region between dArc1 and dArc2 is polymorphic present/absent, and that polymorphism seems to affect darc1 protein and mRNA levels

## ARC in mammals: Pastuzyn, Cell, 2018

Mammalian Arc mediates intercellular mRNA transfer in neurons, mediating synaptic plasticity

ARC and dArc1/2 are both independently derived from Ty3/gypsy element:

Rat ARC and drosophila Arc1 form capsid structures

Rat ARC capsids bind and encapsulate RNA

Rat ARC capsid assembly requires RNA

Arc Protein and Arc mRNA Are Released by Neurons in Extracellular Vesicles

“Highly conserved, unique orthologs of the murine Arc genes were identified throughout the tetrapods (mammals, birds, reptiles, amphibians), but were conspicuously absent from all fish lineages and other deuterostomes examined (94 species). The closest relatives of Arc in the coelacanth, zebrafish, and carp genomes were encoded by prototypical Ty3/gypsy retrotransposons, with indications of recent transposition activity. Similarly, orthologs and paralogs of Drosophila Arc (darc1, darc2) were identified in all schizophoran (true) flies represented in the database but were not detected in any other dipteran (e.g., mosquitoes) or protostome species (286 species; Figure S1B). The closest retrotransposon relatives of the fly Arc genes were found in the genomes of the silkworm and Argentine ant. Interestingly, while Arc appears to be a single-copy gene in all tetrapods examined, the gene has experienced multiple rounds of duplication during schizophoran evolution (Figure S1B). Phylogenetically, tetrapod Arc genes cluster with Ty3/gypsy retrotransposons from fish, whereas the fly Arc homologs group with a separate lineage of Ty3/gypsy retrotransposons from insects (Figure 1A). These results indicate that the tetrapod and fly Arc genes originated independently from distinct lineages of Ty3/gypsy retrotransposons, as conjectured previously (Abrusa´ n et al., 2013), but still share significant homology in the retroviral Gag domain.”

Fig 1A says that mammalian ARC is a tetrapod gene (shared with lizard Anolis carolensis) (and alignment in Figure S1A)

Figure S1B shows tree of Dipteran Arcs – the two Drosophila Arcs arose by duplication of each other in the early Dipteran ancestor. Some other dipterans have experienced subsequent duplications so have >2 Arcs.

# Make sure I remember

possibility of programmed frameshift

# using UCSC alignments

## parsing maf format

did this, but I did not like the results. maf alignments don’t do very well with gappy regions

## using liftOver

did this. I liked the results better than the parsed mafs, but I know I’m getting some non-orthologous regions aligning when there’s a duplication.

# 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

# Will’s message of Dec 18 2020

he made two alignments: PNMA and MART family, each with closest capsid of known structure (dArc for PNMA and Ty3 for MART)