Evolutionary analysis of viral sequences in eukaryotic genomes

Sean Schneider

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:
James H. Thomas, Chair
Willie Swanson
Phil Green

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Evolutionary analysis of viral sequences in eukaryotic genomes

Sean Schneider

Chair of the supervisory committee: Professor James H. Thomas

Genome Sciences

The focus of this work is several evolutionary analyses of endogenous viral sequences in eukaryotic genomes. Endogenous viral sequences can provide key insights into the past forms and evolutionary history of viruses, as well as the responses of host organisms they infect. In this work I have examined viral sequences in a diverse assortment of eukaryotic hosts in order to study coevolution between hosts and the organisms that infect them.

This research consisted of two major lines of investigation. In the first portion of this work, I outline the hypothesis that the C2H2 zinc finger gene family in vertebrates has evolved by birth-death evolution in response to sporadic retroviral infection. The hypothesis suggests an evolutionary model in which newly duplicated zinc finger genes are retained by selection in response to retroviral infection. This hypothesis is supported by a strong association (R^2=0.67) between the number of endogenous retroviruses and the number of zinc fingers in diverse vertebrate genomes. Based on this and other evidence, the zinc finger gene family appears to act as a “genomic immune system” against retroviral infections.

The other major line of investigation in this work examines endogenous virus sequences utilized by parasitic wasps to disable hosts that they infect. These wasps package their own DNA into viral particles and inject them into the host. I found that the DNA packaged into these viral particles can be permanently transferred to the hosts that these wasps infect. I have identified 105 transferred regions in two host species: Monarch butterfly (*Danaus Plexippus*) and Silkworm (*Bombyx mori*). The last common ancestor between these species and wasps lived around 300 million years ago. Many of these regions are highly similar to one another and the sequences form 12 groups when clustered by 90% nucleotide identity. These similarities may arise from repeated integration of the same sequence or duplication after integration into the host.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

No work is done in a vacuum (particle physics aside) and this thesis is no exception. I consider myself lucky and privileged to have the love, support, and inspiration I have received from others. Without it, I have a hard time imagining how I could have made it where I did today.

I would like to first thank the Department of Genome Sciences both as an organization and as a collection of so many people who have helped me I can't list them all. The material, emotional, financial, and any other kind of support I have needed that has been given to me over the years made all of this possible in so many ways. As an organization, it made me believe that there are people out there who actually care about educating graduate students when my faith in such things was severely shaken. Thank you for the support, the education, and for believing in me. I would especially like to thank: my incoming class (especially Jarrett Egertson and Leslie Emery), Celeste Berg, Larry Ruzzo, Bob Waterston, and David Hawkins. Thank you Brian Giebel for always knowing what the heck is going on and what to do about it.

I would like to thank Jim Thomas for being my mentor, teacher, collaborator, listener, and more. You gave me the freedom to pursue projects that interest me, and were always there to listen, even if it wasn't about my project. Thank you for not only putting up with but loving conversations such as how a zombie plague is really just a much more powerful rabies strain. And thank you for talking me down from those inevitable moments of complete panic in grad school. I would like to thank current and former members of the Thomas Lab for their help and support: Ryan Emerson, Tayna Grancharova, and Rachel Bradshaw.

Thanks to the members of my thesis committee: Willie Swanson, Phil Green, John Stamatoyannopolis, and Toby Bradshaw for helpful comments, both scientific and inspirational. Your comments, suggestions, and ideas for controls greatly improved the work I present here.

I would like to additionally thank Willie Swanson for acting like a second mentor to me, both in my time here in Genome Sciences and back when I was an undergrad. I'm grateful that you took a starry-eyed undergrad you hardly knew under your wing and tried to warn him about where to apply for school, even if he didn't listen. Thank you again for helping him again when he finally made the smart choice to join Genome Sciences and for your help, kindness, and support over all these years. And thank you for surrounding yourself with great people who also helped me: Geoff Findlay, Katrina Claw, Melody Palmer, Renee George, Jan Aagaard, Jen McCreight, Steven Springer, and Joe Gasper.

Thank you to my family for everything. Mom, for your unwavering and unconditional love, support and more help than I can list. Dad, for teaching me to love science without reservation or hesitation. Mehgan, for never giving up (especially as an indomitable Donkey Kong partner during endless Super Nintendo sessions). Jesse, for always understanding where I'm coming from. Heather, for always calling things like you see them.

I want to spend a little space thanking people who inspired me without being aware of it. Thank you to Toby Bradshaw for inspiring me to study evolution for a living with your Biology 101 class. Thank you to Sid Meir for making games that inspire me to care about the past and

future of humanity and science, especially *Alpha Centauri*. Thank you to some long-dead philosophers who inspired me: Aristotle, Lucretius, Hume, Mill, Nietzsche, Russell, Wittgenstein. Thank you to Charles Darwin for the modern concept of evolution, and being a little willing to be a heretic. Thank you to Richard Dawkins for writing *The Selfish Gene*, which broadened my thinking about evolution and inspired me to love retrotransposons. Thank you to the group selectionists who railed against molecular evolution for inspiring me to prove them wrong. Thank you to everyone who pushed the Human Genome Project forward, for making this amazing field exist in my lifetime and making it possible to do what I love. Thank you to genetics for teaching me that no species is more special than another species and pushing me to think bigger than human evolution.

Finally, I would like to thank Jen for helping me in more ways than I can list, but I will attempt to list them anyway. Thank you Jen for being my rock in all those storms I weathered in grad school. Thank you for celebrating the good days with me, getting me through the tough days, and understanding me when I felt crazy. Thank you for proofreading mastery, expert typo hunting, and pushing me to present my work around the world. Thank you for being my therapist and masseuse as needed. Thank you for making me smile so many times. Thank you for believing in me. And thank you, Pixel, for letting me pet you occasionally.

Glossary

ERV – Endogenous Retrovirus. A retrovirus which has permanently integrated its genome into a host's germ line.

LTR – Long Terminal Repeats. Identical regions of sequence found at both ends of a retroviral genome that are necessary for retroviral reproduction

ENV – The retroviral "envelope" gene. Allows retroviruses to bind to and enter host cells.

ZNF – Zinc Finger

HGT – Horizontal Gene Transfer. The transfer of genetic material from one organism into the germ line of another organism

PBS –Primer Binding Site. A region in retroviral genomes which host tRNAs bind to, acting as a primer to initiate genome replication for the retrovirus.

KRAB - Kruppel-Associated-Box. A protein domain associated with zinc fingers that recruits the KAP1 complex

KAP1 - KRAB-associated protein-1 (also known as TRIM28). A histone modifying protein complex recruited by KRAB.

SETDB1 – SET domain, bifurcated 1. A methyltransferase recruited by KAP1

SCAN – Also known as the "leucine rich region", SCAN is a protein domain associated with zinc fingers that is thought to be adapted from a viral *gag* gene

IC – Independent Contrasts. A statistical method proposed by Joe Felsenstein to analyze correlations between traits of interest across an evolutionary tree

Chapter 1: Introduction

This work is divided into three chapters. The first chapter covers background topics helpful to understanding chapters two and three, each of which cover different topics in the evolution of genomes.

The second chapter covers work putting forward the hypothesis that, within metazoans, the C2H2 zinc finger family has coevolved with retroelements to act as a genomic immune system against endogenous retroviruses. This hypothesis is based on the evolutionary dynamics of the gene family, biochemical mechanics of the zinc finger proteins, and insights from other work on the organismal function of zinc finger proteins. We find evidence supporting the hypothesis in the form of a striking correlation in vertebrates between the number of endogenous retroviruses and the number of zinc finger genes in a genome assembly. It is additionally supported by extensive statistical and phylogenetic analyses. The work in this chapter was performed in extensive collaboration with Jim Thomas.

Chapter three examines horizontal gene transfer from a group of wasps which utilize endogenous virus sequences to assist in attacking the host organisms they parasitize. These wasps produce viral particles that lack viral sequences but contain fragments of the wasp genome, and inject these viral particles into the host body. By random chance, some of the viral particles entered the host germline in Silkworm (*Bombyx mori*) and Monarch Butterfly (*Danaus plexippus*) and permanently transferred wasp DNA into the host's chromosomes, which were in turn passed down to the host's offspring. These transferred sequences were found by an extensive bioinformatic search and verified by PCR amplification in diverse Silkworm samples from around the world, suggesting that these sequences are fixed throughout this species. The work presented here is the first known case of horizontal gene transfer by this mechanism.

Endogenous Retroviruses

Retroviruses are a medically and evolutionarily important group of viruses which are unique in their mode of reproduction among eukaryotic viruses. Like other viruses, the life cycle of a retrovirus requires that the virus infect a host organism and successfully enter a host cell. Unlike other viruses, retroviruses must integrate their genome into the host genome to reproduce. The basic composition of a retroviral genome is simple (Figure 1), consisting of several key protein coding regions (Stoye 2012). The group-specific antigen (*gag*) gene codes for the structural building block of the viral particle, which is polymerized into a three dimensional structure that contains and protects the viral genome. The envelope (*env*) protein binds receptors on the surface of host cells and gives the virus cell type specificities. The viral protease (*pro*) protein processes viral gene products into mature proteins, enabling production of new

viral particles. The polymerase (*pol*) gene creates several different proteins including RNAse H (an RNA cleaving enzyme found in many organisms) and two proteins that are hallmarks of the retroviruses: reverse transcriptase and integrase. Reverse transcriptase (RT) is a key enzyme that enables the retrovirus to make a DNA copy of its RNA genome, allowing it to be integrated into the host as a provirus. This process is facilitated by the integrase enzyme, which catalyzes integration of retroviral DNA into the host chromosome. These two proteins enable retroviruses to permanently incorporate themselves into the host genome.

Obligatory integration of the viral genome into the genome of the host in order to reproduce is one of the key features that distinguishes retroviruses from other viruses. When the viral DNA is integrated it forms a provirus, from which new retroviral genomes are transcribed and then packaged into new virion particles (Weiss 2006). These virion particles are released without lysing the cell, keeping the cell alive for the provirus. The provirus will continue to make new viral particles that can exit the cell to start new infections. If the provirus is integrated in a germ line cell, it can be passed down vertically to offspring of the host organism and on to their offspring and is known as an endogenous retrovirus (ERV). During host development, ERVs transcribe their genomes to make additional copies, reverse transcribe the RNA genome, and then integrate those new copies into the host genome. Because ERVs can persist for many millions of years in a host lineage, this process takes place repeatedly over many generations, resulting in ERVs with many thousands of copies in a host genome. These ERVs can abandon their free living lifestyle for many millions of years, expanding in number in the host as ‘genomic parasites’ and being passed on vertically to offspring and all future descendants.

After a retrovirus infects and expands into a new genome, a number of different ultimate outcomes can occur (Boeke and Stoye 1997). One possible outcome is that a copy of the ERV will transcribe its genome, translate the proteins, and form a virion particle that exits the host to infect a new host. In this way, endogenous retroviruses can take an exogenous form again and disperse to infect different hosts and continue the retroviral life cycle (Malik et al. 2000, Belshaw et al. 2004). Since the replication is “copy/paste”, the endogenous copy remains in the original host to continue to create more endogenous and exogenous copies. In this way, an ERV integrated in a host cell can function as a “factory” to produce more viruses as long as the cell remains alive.

Another outcome is that the retrovirus never produces exogenous copies that successfully infect other organisms, but does produce more endogenous copies(Walsh et al. 2012, Magiorkinis et al. 2012). In this scenario, the endogenous retrovirus can be an evolutionary “dead end” because it is never able to escape to infect new hosts. Nevertheless, it may continue to exist for many millions of years in the host genome, leaving sequences for researchers to investigate to learn about ancient viruses.

Occasionally, the retrovirus can be beneficial to the host in situations where the provirus is modified and adapted by the host for a new function that increases the host fitness. A classic example of this is the mammalian placenta, which requires the use of fusogenic proteins originally derived from retroviral proteins (Mi et al 2000).

Whether or not the ERV produces new virion particles that spread to new hosts, it remains in the host genome. Like all sequences in the host genome, these will be subject to random mutations and genomic rearrangements over evolutionary time, and can be removed by genetic drift. These genetic alterations can lead to particular reduced forms of ERVs (Stowe 2012). Some of the most commonly seen mutations are those that inactivate the ERV's ability to infect new cells, typically by loss of function in the *envelope* gene. These are known as retrotransposons and when this occurs, the virus can reproduce more copies in the host but is unable to spread to new hosts. Additionally, ERVs and retrotransposons will often lose their long terminal repeats (LTRs), identical regions at each end of the retroviral genome that are required for exogenous reproduction. LINEs are a group of retrotransposons that is common in vertebrates and lacks LTRs and the *envelope* gene, but have retained reverse transcriptase and thus are capable of producing more copies inside the host. SINEs are further reduced from LINEs: Having lost their own reverse transcriptase, they must "borrow" the protein from elsewhere in the cell in order to reproduce. One final outcome for ERVs is that all genetic content except the LTRs is removed, leaving "solo LTRs" in the genome that are incapable of retrotransposition. When all the different forms (retrotransposons, SINEs, LINEs, etc.) of endogenized retroviruses are included they can comprise as much as 40% of a mammalian genome (Mouse Genome Sequencing Consortium 2002).

While retroviruses excel at becoming entrenched in a host organism, a crucial step in the retrovirus life cycle is the infection of new hosts. Like most viruses, a given retroviruses typically infects members of a particular species. Occasionally, retroviruses can be horizontally transferred between distant species, allowing them to spread and adapt to new hosts. Spreading to a new species is a crucial step in the long term evolution of a virus, as it is the main defense the virus has against dangers like the extinction of the original host species or the host species developing immune resistance. These species jumps can have dramatic effects on the genome and evolution of host species. One dramatic case is the BovB retrotransposon: BovB made the jump from a python species (*Python molurus*) over into cattle, and then made copies of itself to the point where it now accounts for 25% of the cow genome, including derived SINEs (Walsh et al. 2012). Besides the increase in genomic content, there were no obvious biological effects in cattle resulting from the transfer.

Endogenous retroviruses (and derived forms like LINEs) are understood to typically not have serious negative effects on the host, instead silently, surreptitiously,

and slowly reproducing more copies over millions of years. However, this is not always the case. The retrovirus known as KoRV derives from a Gibbon Ape Leukemia Virus (Tarlington et al. 2006) which in turn derived from a mouse leukemia virus. KoRV has successfully integrated many copies into almost every individual of most wild Koala populations. Affected individuals have severe effects including chlamydiosis, neoplasia, and leukemia. In captive populations, the disease is thought to ultimately kill 80% of affected organisms (the death rates in wild populations are presently unknown). The disease is severe and widespread enough that it is thought to threaten the continued existence of the Koala species. While most ERVs are not known to seriously threaten their hosts, the existence of those like KoRV provide a compelling evolutionary reason for host species to gain adaptations to fight retroviral infections.

Endogenous Viral Elements

Viruses are specialized pathogens that complete their life cycle by invading a host cell, evading host immune response, utilizing host resources to produce more virus particles, and then releasing those particles outside of the infected cell to infect new cells and begin the cycle anew. Typically, most non-retroviral viruses do not leave any genetic information in the host cell's genome.

Occasionally, parts of viral genomes can be integrated into the genome of the host cell and persist after no virus particles remain. There are multiple ways that (non-retroviral) viral sequences can become integrated into host genomes (Figure 2), depending on the type of virus and composition of the viral genome (Katzourakis and Gifford 2010). DNA viruses can randomly integrate into host chromosomes by chance integration. Retroviruses utilize reverse transcriptase to make DNA copies of their RNA genomes. Non-retroviral RNA viruses can be bound by the reverse transcriptase originating from ERVs within the genome, which is often expressed and found sporadically throughout the cell. While reverse transcriptase preferentially binds retroviral sequences, it can bind other RNA sequences with a lower efficency and produce a DNA copy of their genome. Portions of a viral genome existing as DNA can be randomly inserted into a host chromosome by chance integration, as with DNA viruses. If the host cell survives the infection, these integrated viral sequences will be passed on to daughter cells.

If the viral sequences are integrated into a germ-line cell, the transferred viral sequence will be in gametes of the host and can be passed to the host's offspring. These sequences can then be passed into subsequent offspring and passed down vertically in a species. These stably inherited sequences are called Endogenous Viral Elements (EVEs). EVEs can be a source of novel sequences for organisms, giving the organism access to viral proteins it would not otherwise have access to. With the

explosion of eukaryotic whole genome sequence data that has become available, many EVEs have been found in eukaryotic genomes (Katzourakis and Gifford 2010).

EVEs and ERVs are of great interest to researchers, and evolutionary biologists in particular. Both retroviruses and other viruses have some of the highest mutation rates in the living world. On the one hand, this makes short term evolutionary studies of viruses easy and powerful. The downside to the high mutation rate is that viral genomes are often saturated for mutations (Emmerman and Malik 2010). This makes finding and aligning homologous sequences difficult, which in turn makes long term studies of viruses challenging. EVEs and ERVs can be a solution to this issue, upon integration into the host genome they will mutate at the host's mutation rate rather than the viral mutation rate (Gilbert and Feschotte 2010). This preserves the ancient virus sequences over time with a mutation rate many orders of magnitude lower (~10^{-3} vs ~10^{-9}, Duffy et al. 2008), allowing researchers to find preserved viral "fossils".

Endogenous viral elements and retroviruses found in extant genomes are likely a small sampling of the viral sequences integrated into hosts by viruses. As with other new mutations, most EVEs and ERVs are likely to be deleterious or neutrally evolving and potentially lost from populations by genetic drift. Many transferred viral sequences do not contain complete functional proteins and these are especially likely to be neutrally evolving. Negative selection can potentially act on transferred viral sequences, as many viruses encode for proteins toxic to their host organism and the insertion of sequences can itself be deleterious or disruptive to the host. On the other hand, endogenous viral sequences can sometimes be adaptive for the host organism (Patel et al. 2011), as the horizontal transfer provides the organism with new sequences for positive selection to act upon. Viral promoters can be transferred along with protein sequences and if they integrate near a host gene these promoters can be modified and adapted to the host. Some EVEs have been shown to be transcribed into RNA in hosts. If EVEs are transcribed in the antisense orientation, the RNA could potentially bind to single stranded viral RNAs, forming double stranded RNAs that are degraded by processes in the innate immune system such as APOBEC1 (Wedekind et al. 2003). Alternatively, if EVE sense RNA is translated into a protein, these viral proteins can activate the adaptive immune system against related viral infections. For example, a defective copy of the *gag* protein in a sheep ERV has been shown to be incorporated into newly formed viral particles, rendering them inactive (Murica et al. 2007). Organisms can also find completely novel uses for EVE sequences, a remarkable example can be found in parasitoid wasps that utilize endogenous nudivirus sequences to assist in disabling species that they parasitize (Bezier et al. 2009).

Polydnaviruses

The Hymenoptera are typically best known for their many eusocial species (ants, bees, hornets), but the vast majority of Hymenoptera species are thought to be solitary and parasitoid wasps. In the adult stage of their life cycle, female parasitoid wasps will search for an appropriate place to lay their eggs. These parasitoid wasps will typically lay their eggs inside the body of another specific insect species, usually the larval stage of a Lepidopteran insect (commonly known as caterpillars). The morphology of parasitoid insects reflects this lifestyle, females lack the ability to "sting" other organisms as seen in eusocial Hymenoptera, instead having a long needle-like ovipositor specialized at penetrating the exoskeleton of insects and injecting wasp eggs inside the victim. The eggs then develop until they reach adulthood, at which point they will all burst out of the host, killing the host in the process (Figure 3).

In reaction to a parasitoid attack, the host will launch an immune response that will attempt to destroy the growing wasp eggs and save the life of the host. If the eggs are injected into the host early in host development, the parasitoid eggs are much more likely to hatch successfully. However, if the eggs are injected later in host development the parasitoid infection may be likely to be defeated by the host immune system (Pruijssers et al. 2009). The possibility of the host fighting parasitoid infection sets up a strong evolutionary incentive for the wasp to evade the host's immune system, and conversely creates a strong incentive for hosts to defeat parasitoid infection. As a result, parasitoid wasps and their host species can evolve by a classical host-parasite "arms race". Many parasitoid wasps inject venom along with their eggs to disable the host and/or weaken the host's immune system.

Two taxa of wasps, the Braconidae and the Ichneumonidae, have stably integrated full viral genomes into their own genomes. These groups of parasitoid wasps have "domesticated" these viruses and inject them at the same time as their eggs (sometimes in addition to venom). These viruses are collectively known as "polydnaviruses" (PDVs), named after the many circles of DNA residing in virion particles. The domestication of viruses by wasps occurred independently two times in the evolution of wasps: A nudivirus was domesticated to form the Bracoviruses (used by the Braconid wasps) and an undetermined virus was domesticated to form the Ichnoviruses (used by the Ichneumonid wasps) (Bézier et al. 2009, Burke and Strand 2012). Both groups of PDVs work by disabling the host's immune system and arresting host development, allowing wasp larvae to evade the immune response and dedicating host resources toward nourishing the wasp larvae.

The genomic content and mode of reproduction in PDVs is completely different from the ancestral pre-domesticated forms of the viruses, or any other virus. The key difference is that the genetic content of PDVs is entirely wasp in origin. The genes that the wasps acquired from the viruses are used to produce virion particles but are not packaged in them (Fleming and Summers 1991). One result of this arrangement is that

PDVs are unable to reproduce after invading a host cell, as they lack the genetic code to produce more viral proteins. PDVs are produced only in the ovaries of the wasp (Stoltz et al 1976), where they are packaged with particular regions of the wasp genome specifically targeted for replication and circularization. The function of the wasp proteins coded for by the genetic content of PDVs remains largely a mystery, though it is generally thought that they function to suppress the host immune system (Hepat and Kim 2011) and arrest host development (Kwon et al. 2010).

C2H2 Zinc Fingers

As illustrated above, the genomes of organisms are constantly at risk for the dangers of viral integration, creating evolutionary pressure on host species to evolve counter-measures. Before the model presented here in chapter 2, there was no clear mechanism by which host organisms could broadly counteract endogenous viruses in their genomes. As outlined in chapter 2, the zinc finger gene family may have evolved as such a counter-measure.

The C2H2 zinc finger is a 28 amino acid protein domain that binds DNA. In genes, these zinc fingers (ZNFs) are found in large tandem arrays at the C-terminus of the protein. The tandem arrays are often accompanied by other domains known to be associated with zinc fingers on the N-terminus of the protein, such as KRAB and SCAN (Figure 4). Each finger binds a specific 3 nucleotide sequence (Figure 5), and tandem arrays of ZNFs bind larger sequences of 3 nucleotides per ZNF (Pavletich and Pabo 1991, Kim and Berg 1996). In addition to binding the 3 canonical nucleotides, each ZNF also binds an additional nucleotide from an adjacent ZNF, making binding predictions for tandem arrays of ZNFs highly complex. These tandem-ZNF genes are themselves found in large clusters on chromosomes in vertebrate genomes (Shannon et al. 2003, Emerson and Thomas 2009). This genomic organization may arise from local duplication of ZNF genes or selection for common regulation of the genomic region, however the reasons for this genomic organization are still uncertain.

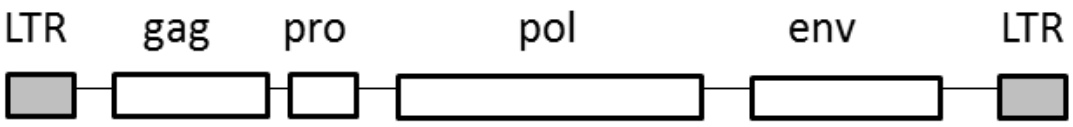
The human genome is thought to have around 470 tandem ZNF genes (Emerson and Thomas 2009). By comparison, the mouse genome has around 330 with similarly variable amounts found in other mammals. Interestingly, only 137 ZNF genes were found to be shared among placental mammals, which indicates that the gene family is evolving by birth-death evolution. This mode of evolution is common among large gene families: having new genes arise from genomic duplications and losing genes from genomic deletions. Most gene families in mammals are not as large as the ZNFs, they are rivaled only by the olfactory receptor gene family. In addition to evolving by birth-death evolution, ZNF genes are also frequently undergoing classical positive selection

in mammals (Emerson and Thomas 2009). This signature is typically seen when a genetic sequence is frequently undergoing selection for some characteristic, suggesting that there may be some constant evolutionary pressure on ZNF genes of undetermined origin.

In addition to tandem ZNF arrays, genes containing ZNFs are associated with a number of other protein domains (Figure 4). One interesting feature of ZNF genes is that the ZNFs can bind a specific DNA sequence which then brings in other protein domains to perform biochemical reactions nearby. In vertebrates, the best studied ZNF-associated domain is the KRAB domain (Birtle and Ponting 2006). The KRAB domain recruits KAP1 which nucleates a powerful histone modifying complex (Nielsen et al 1999, Ryan et al. 1999). The KAP1 complex de-acetylates histones and aids in the formation of heterochromatin. The effect of this multi-part structure is that the tandem ZNF array targets and binds a specific sequence of interest, followed by the KAP1 complex de-acetylating nearby histones, shutting down transcription in the region. This makes KRAB Zinc Finger genes (KZNFs) powerful transcriptional regulators of the genome. Another domain associated with ZNF genes is the SCAN domain, which can be present in addition to KRAB. The SCAN domain was derived from the viral capsid protein (gag) of a retrotransposon, *gypsy* (Emerson and Thomas 2011). The exact role of the SCAN domain in ZNF genes and in the organism is not clear. However, given its sequence similarity to the viral capsid protein (which self-polymerizes) it is likely that it localizes to viral proteins in some fashion.

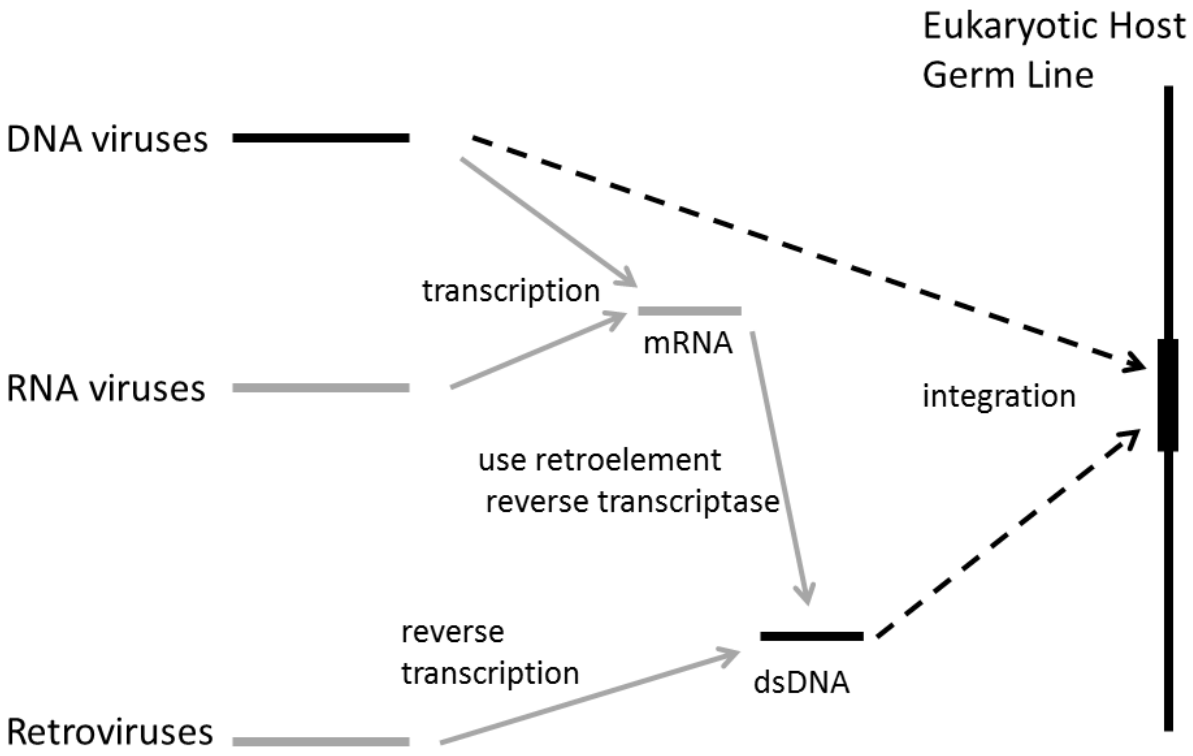
The functional role of ZNF genes in the organism has been a difficult problem for researchers to fully investigate, however some ZNF genes have been found to have organismal functions: Three genes were found to have a role in X-linked mental retardation in humans (Shoichet et al. 2003, Kleefstra et al. 2004, Lugtenberg et al. 2005), *CTCF* is an important chromatin remodeler and transcriptional insulator in eukaryotes (Filipova et al. 1996, Ishihara et al. 2006), *Zfp206* was found to induce pluripotency in embryonic stem cells, *Rsl* was found to cause sexually dimorphic gene expression in mouse livers (Krebs 2003), *OTK18* caused transcriptional repression of HIV in one study (Horiba et al 2007), and the *chato* mutation causes defects in development in mice (Garcia-Garcia 2008). The hypothesis put forth in this work is that most new zinc finger genes arise in response to retroviral infections but can later be co-opted for other functions after the retroviral infection is no longer a threat.

Figure 1.1: Typical retroviral gene structure



Identical LTRs flank both termini of the genome. Four canonical ORFs are shown, in order: group-specific antigen (*gag*), protease (*pro*), polymerase (*pol*), envelope (*env*).

Figure 1.2: Alternative paths to viral endogenization.



DNA viruses (top) can undergo random integration into chromosomes. In a germ line cell, these sequences will be passed on vertically to offspring. RNA viruses can be bound by reverse transcriptase from retroelements and reverse transcribed into a DNA copy, which can be randomly integrated by chance. Retroviruses specialize at integrating into host chromosomes, carrying reverse transcriptase and integrase proteins into the cell with the infecting virion particle to facilitate integration.

Figure1.3: Lifecycle of parasitoid wasps

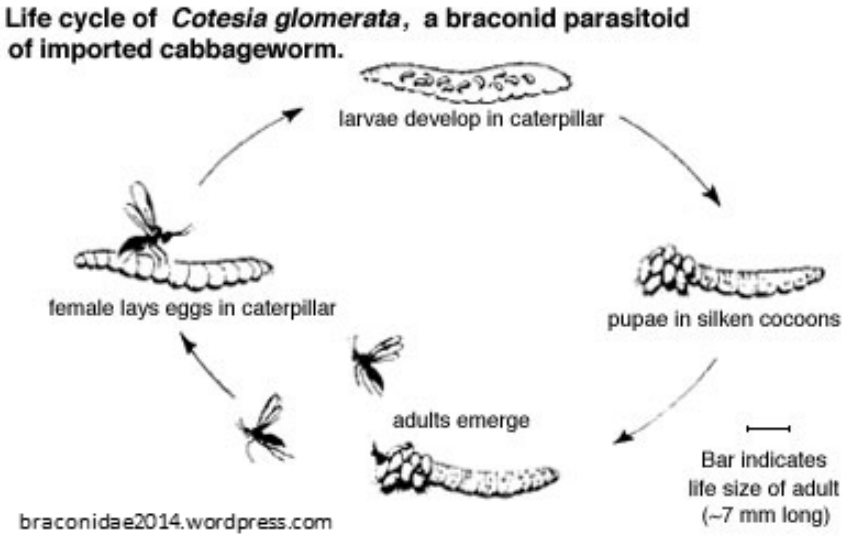
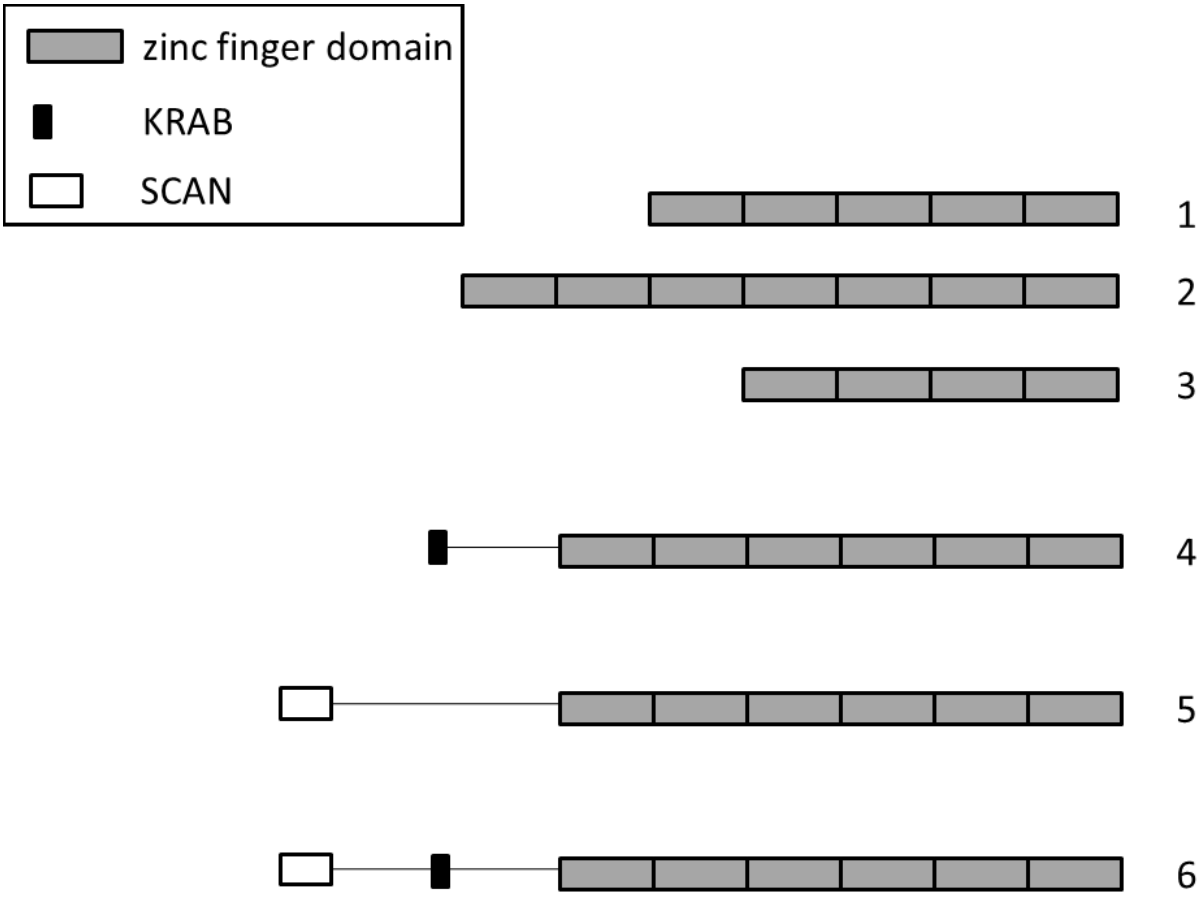


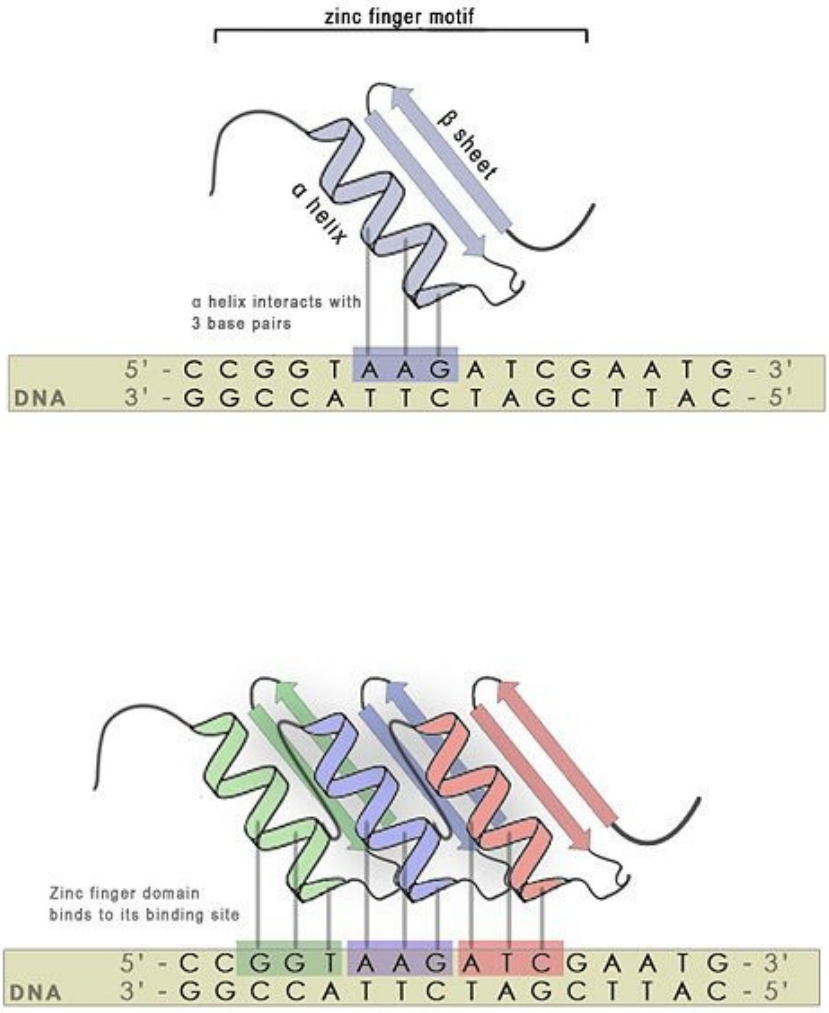
Figure from braconidae2014.wordpress.com. An adult female wasp finds a suitable host organism and lays her eggs in it (left), often accompanied with venom or polydnaviruses. If the larval wasps evade host defenses, they will grow inside the host (top). Eventually, the young wasps form pupae on the outside of the host (right). Adult wasps then emerge from the host and continue the cycle (bottom)

Figure 1.4: Gene models of C2H2 zinc fingers found in vertebrate genomes



Models 1, 2, and 3 show different ZNF genes that vary in the number of zinc finger domains contained. ZNFs can vary widely in the number of domains, and therefore in the length of the DNA sequence targeted by the protein. Model 4 has the addition of the KRAB domain, which recruits the powerful transcription repression complex, KAP1. Model 5 shows the addition of the SCAN domain, thought to potentially bind to viral proteins. Part 6 has both the SCAN and KRAB domains in the same gene.

Figure 1.5: Model of zinc finger binding to DNA



From http://2010.igem.org/Team:Slovenia/PROJECT/proof/domain. Each zinc finger binds a specific 3 nucleotide sequence (top). Tandem arrays of zinc fingers can bind longer sequences (bottom), varying with the number of fingers in the protein.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U.S.A. 101, 4894–4899.

Bézier A, Annaheim M, Herbinière J, Wetterwald C, Gyapay G (2009) Polydnaviruses of braconid wasps derive from an ancestral nudivirus. Science 323(5916):926-30.

Birtle Z, Ponting CP. 2006. Meisetz and the birth of the KRAB motif. Bioinformatics 22: 2841-2945.

Boeke JD, Stoye JP: Retrotransposons, endogenous retro-viruses, and the evolution of retroelements. In Retroviruses. Edited by Coffin JM, Hughes SH, Varmus HE. Cold Spring Harbor, NY: Cold Spring Harbor Press; 1997: 343-435.

Burke GR, Strand MR (2012) Deep sequencing identifies viral and wasp genes with potential roles in replication of Microplitis demolitor Bracovirus. J Virol. 86(6):3293-306.

Duffy, S., Shackelton, L.A., and Holmes, E.C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet.9, 267–276.

Emerman M, Malik HS (2010). Paleovirology--modern consequences of ancient viruses. PLoS Biol. Feb 9;8(2):e1000301. doi: 10.1371/journal.pbio.1000301.

Emerson RO, Thomas JH. 2009. Adaptive evolution in zinc finger transcription factors. PLoS Genet. 5: e1000325.

Emerson RO and Thomas JH. Gypsy and the Birth of the SCAN Domain. J Virol. Nov 2011; 85(22): 12043–12052.

Fleming JG, Summers MD (1991) Polydnavirus DNA is integrated in the DNA of its parasitoid wasp host. Proc Natl Acad Sci U S A. 88(21):9770-4.

Filipova GN, Fagerlie S, Klenkova EM, Myers C, Dehner Y, et al. (1996) An Exceptionally Conserved Transcriptional Repressor, CTCF, Employs Different Combinations of Zinc Fingers To Bind Diverged Promoter Sequences of Avian and Mammalian c-myc Oncogenes. Mol Cell Biol 16(6): 2802–2813.

Garcia-Garcia MJ, Shibata M, Anderson KV (2008) Chato, a KRAB zinc-finger protein, regulates convergent extension in the mouse embryo. Development 135: 3053–3062.

Gilbert, C., and Feschotte, C. (2010). Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol.8, e1000495

Horiba M, Martinez LD, Buescher JL, Sato S, Limoges J, et al. (2007) OTK18, a zinc-finger protein, regulates human immunodeficiency virus type 1 long terminal repeat through two distinct regulatory regions. J Gen Virol 88: 236–241.

Ishihara K, Oshimura M, Nakao M (2006) CTCF-Dependent Chromatin Insulator Is Linked to Epigenetic Remodeling. Mol Cell 23: 733–742.

Kim CA, Berg JM. 1996. A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. Nat. Struct. Biol. 3: 940-945.

Kleefstra T, Yntema HG, Oudakker AR, Banning MJ, Kalscheuer VM, et al. (2004) Zinc finger 81 (ZNF81) mutations associated with X-linked mental retardation. J Med Genet 41: 394–399.

Krebs CJ, Larkins LK, Price R, Tullis KM, Miller RD, et al. (2003) Regulator of sex-limitation (Rsl) encodes a pair of KRAB zinc-finger genes that control sexually dimorphic liver gene expression. Genes Dev 17: 2664–2674.

Lugtenberg D, Yntema HG, Banning MJG, Oudakker AR, Firth HV, et al. (2005) ZNF674: A New Krüppel-Associated Box–Containing Zinc-Finger Gene Involved in Nonsyndromic X-Linked Mental Retardation. Am J Hum Genet 78(2): 265–278.

Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. (2012) Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A. May 8;109(19):7385-90.

Malik HS, Henikoff S, Eickbush TH. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 2000 Sep;10(9):1307-18.

Mi S et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000 Feb 17;403(6771):785-9.

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520–562.

Murcia PR, Arnaud F, and Palmarini M (2007). The transdominant endog-enous retrovirus enJS56A1 associates with and blocks intracellular trafficking of Jaagsiekte sheep retrovirus Gag. J. Virol.81, 1762–1772.

Nielsen AL, Ortiz JA, You J, Oulad-Abdelghani M, Khechumian R, Gansmuller A, Chambon P, Losson R. 1999. Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. EMBO J. 18: 6385-6395.

Patel MR, Emerman M, Malik HS (2011). Paleovirology - ghosts and gifts of viruses past. Curr Opin Virol. Oct;1(4):304-9. doi: 10.1016/j.coviro.2011.06.007.

Pavletich NP, Pabo CO. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252: 809-817.

Pruijssers AJ, Falabella P, Eum JH, Pennacchio F, Brown MR, Strand MR (2009) Infection by a symbiotic polydnavirus induces wasting and inhibits metamorphosis of the moth Pseudoplusia includens. Journal of Experimental Biology. 212: 2998–3006.

Ryan RF, Schultz DC, Ayyanathan K, Singh PB, Friedman JR, Fredericks WJ, Rauscher FJ 3rd. 1999. KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Krüppel-associated box-zinc finger proteins in heterochromatin-mediated gene silencing. Mol. Cell. Biol. 19: 4366-4378.

Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res. 13: 1097-1110.

Shoichet SA, Hoffmann K, Menzel C, Trautmann U, Moser B, et al. (2003) Mutations in the ZNF41 gene are associated with cognitive deficits: identification of a new candidate for X-linked mental retardation. Am J Hum Genet 73: 1341–1354.

Stoltz DB, Vinson SB, MacKinnon EA (1976) Baculovirus-like particles in the reproductive tracts of female parasitoid wasps. Can J Microbiol. 22(7):1013-23.

Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012 May 8;10(6):395-406.

Tarlington RE, Meers J, Young PR: Retroviral invasion of the koala genome.Nature2006, 442:79-81

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA. 110:1012–1016.
Wang ZX, Kueh JLL, Teh CHL, Rossbach M, Lim L, et al. (2007) Zfp206 Is A Transcription Factor That Controls Pluripotency of Embryonic Stem Cells. Stem Cells 25: 2173–2182.

Wedekind JE, Dance GS, Sowden MP, Smith HC (2003) Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends Genet. Apr;19(4):207-16.

Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006 Oct 3;3:67.

Chapter 2: Coevolution of retroelements and tandem zinc finger genes

Introduction

Vertebrate genomes contain large and highly variable numbers of tandem C2H2 zinc finger (tandem ZNF) transcription factor genes. Outside of mammals almost nothing is known about the function of these genes. Mammalian tandem ZNF genes are dominated by those with a KRAB domain, which expanded from a KRAB – tandem ZNF fusion gene near the root of tetrapod vertebrates (Birtle and Ponting, 2006). Most KRAB zinc finger (KZNF) proteins consist of an N-terminal KRAB domain followed by multiple tandem ZNF domains (Bellefroid et al., 1991; Huntley et al., 2006). Some KZNF proteins have an additional SCAN protein-interaction domain N-terminal to their KRAB domain (Edelstein and Collins, 2005). The KRAB domain represses transcription by binding TRIM28 (also called KAP1), which is part of a large protein complex that modifies histones to promote closed chromatin (e.g. Nielsen et al., 1999; Ryan et al., 1999; Lechner et al., 2000; Schultz et al., 2002; Sripathy et al. 2006). The tandem ZNF domains confer DNA-binding specificity in a modular manner, with a turn-helix segment of each ZNF domain binding to 3 nucleotides in target DNA sites (Pavletich and Pabo, 1991; Kim and Berg, 1996).

Tandem ZNF genes have been gained by an ongoing process of lineage-specific duplication and divergence (e.g. Shannon et al., 2003; Emerson and Thomas, 2009; Nowick et al., 2010). Throughout vertebrates, tandem ZNF gene expansions are characterized by strong positive selection that has changed the DNA binding specificity and number of ZNF domains in a gene while retaining a conserved KRAB domain (Schmidt and Durrett, 2004; Emerson and Thomas, 2009). Though most tandem ZNF genes outside of mammals lack a KRAB domain, the structure and evolution of their zinc fingers is strikingly similar to that of KZNF genes in mammals (Emerson and Thomas, 2009). These evolutionary patterns combined with remarkably little functional information have given rise to a set of long-standing puzzles. What are the organismal functions of tandem ZNF genes? Why are there so many genes and why do the gene number and their repertoire of DNA binding sites change so quickly?

Recent work suggests a plausible functional explanation for KRAB tandem ZNF gene evolution. First, a series of papers showed that restriction of murine leukemia virus (MLV) in mouse embryonic stem cells results from transcriptional repression by the mouse-specific KZNF gene *ZNFp809* acting via the TRIM28 complex (Wolf and Goff, 2007; Wolf and Goff, 2009). *ZNFp809* binds integrated MLV DNA at its primer-binding site (PBS), which MLV requires to prime reverse transcription via a complementary host tRNA (Wolf and Goff, 2009). Shortly thereafter, two papers showed that deletion of TRIM28 (or the TRIM28 effector SETDB1) in mouse embryonic cells causes massive transcriptional derepression of several mouse endogenous retroviruses (ERVs) (Rowe

et al., 2010; Matsui et al., 2010). Since TRIM28 and SETDB1 are thought to be shared effectors of all KZNF proteins, this result suggests that a suite of KZNF genes repress transcription of diverse ERVs (Jacobs et al. 2014. For a review, see Rowe and Trono 2011).

These findings have the potential to explain the large number of evolutionarily volatile KZNF genes based on a host-pathogen interaction. ERVs are the genomic footprints of historical retroviral infections that resulted in viral insertions in the germ line (reviewed in Blikstad et al., 2008; de Parseval and Heidmann, 2005). When an ERV first appears in a genome, it typically appears as a burst of multiple elements, due to either transposition (Belshaw et al., 2004, 2005) recurrent viral insertions. This process results in a genomic signature that reflects a sampling of the history of retroviral infections, or at least the subset of infections that successfully established a germ line copy. Based on the diversity and ages of ERVs present today, both the mouse and human lineages have episodically suffered a large number of infections by diverse retroviruses over the past 80 million years (e.g. de Parseval and Heidmann 2005; Stocking and Kozak, 2008). Less extensive analysis suggests that a similar process occurs throughout mammals (e.g. Mouse Genome Sequencing Consortium, 2002; Mikkelsen et al., 2007).

Presumably both new retroviral infections and divergence of endogenous long terminal repeat (LTR) retroelements will drive selection for a host response. One possible host response is the generation of new transcriptional repressors that evolve to target the DNA of the new retrovirus or retrotransposon. The simple modular biochemistry of KZNF transcriptional repression makes KZNF genes particularly suitable for such a role. The existing TRIM28 complex should require only recruitment by a new DNA binding specificity to result in a repressed chromatin state. This repressed state can spread many kilobases from the DNA binding site (Groner et al., 2010), so in principle binding anywhere in a retroelement could provide effective repression.

In this section we present data that supports the hypothesis that most vertebrate tandem ZNF genes evolved to repress retroviral or LTR retroelement activity.

Results

LTR retroelement correlation with ZNF domains and genes: ERVs and LTR retrotransposons differ primarily in the presence or absence of an envelope (ENV) coding sequence. These two types of retroelements interconvert by gain or loss of ENV sequences. Because the ENV coding sequence is extremely diverse and rapidly evolving, distinguishing these two groups is difficult without detailed analysis. We will refer to both types of elements as LTR retroelements, and we did not attempt to

distinguish them in our analysis. Our initial goal was to test for correlation between LTR retroelement load and tandem ZNF coding potential across a wide range of vertebrate genomes. Existing genome annotations are uneven, so we implemented genome searches to detect both LTR retroelements and ZNF domains in a manner independent of annotation status and phylogeny. In the case of retroelements, this was made possible by the fact that LTR retroelements include ancient protein coding domains that distinguish them from all other known genomic features (e.g. reviewed in Gogvadze and Buzdin, 2009). In the case of ZNF domains, this was made possible by the fact that the C2H2 ZNF domain has the same length and sequence profile throughout animals (Emerson and Thomas, 2009).

We found a striking correlation across vertebrates between the number of LTR retroelements and the number of ZNF domains. This correlation holds within mammals and outside of mammals, and when all 26 taxa are combined (Figure 2.1and Figure 2.2). To account for the fact that shared phylogenetic history probably accounts for some of this correlation, we computed corrected correlations and P-values (Figure 2.1and Table 2.1) using the method of Independent Contrasts (IC, Felsenstein, 1985; Garland, Bennett, and Rezende, 2005). The IC-corrected correlations remained strong and highly significant, and were robust to widely different score thresholds for counting LTR retroelements and ZNF domains, and to counting the total number ZNF domains or the number of putative tandem ZNF genes (Table 2.1 and Table 2.2). Given existing evidence that KRAB ZNF genes can repress retroviral and LTR retrotransposon transcription, the most obvious inference is that the historical LTR retroelement content of each genome has driven the number of ZNF domains, but we considered other possible explanations.

First, it seemed possible that each genome has a characteristic rate of segmental duplication or duplicate retention and that this rate drives both LTR retroelement and ZNF domain content. We examined this possibility by testing for IC-corrected correlations of non-LTR (LINE-like) retroelements with ZNF domains and of LTR retroelements with other large dynamic protein domain families (olfactory receptor and immunoglobulin C1 and V domains). None of these correlations were statistically significant (Table 2.3). Second, it seemed possible that unknown constraints on genome size influence the potential for LTR retroelement and ZNF domain content in each genome. We tested this possibility by normalizing LTR retroelement content to genome size and testing the normalized correlation to ZNF domains. The correlation remained highly significant (Table 2.3). Given that LTR retroelements are a significant contributor to genome size, it is unsurprising that genome size itself positively correlates with LTR retroelement and ZNF domain content, though these correlations were weak and statistically non-significant after IC correction (Table 2.3).

Other features of these data can be explained by a model in which LTR retroelements drive tandem ZNF gene evolution. First, testing various score cutoffs for counting ZNF domains showed that the correlation to LTR retroelements is strongest very near the score that best distinguishes human ZNF domains in known genes from those in pseudogenes (Figure 2.3 and Table 2.1). This result suggests that the correlation is stronger for functional ZNF genes than for pseudogenes. Second, LTR retroelement correlation to total ZNF domain number was slightly stronger than to the number of putative tandem ZNF genes (Table 2.1). This result suggests that the total DNA binding potential of ZNF genes is more important than the number of genes. Finally, the most prominent correlation outlier in mammals is mouse, which has fewer ZNF domains than predicted by its LTR retroelement content (Figure 2.1). The mouse genome is known to have several groups of recently and currently active ERVs (Stocking and Kozak, 2008), suggesting the possibility that the host ZNF response is lagging behind a recent burst of LTR retroelement activity. Alternative explanations of this mouse result are considered in the Discussion section.

If the major function of ZNF genes is to transcriptionally repress LTR retroelements, then the sequence diversity of LTR retroelements should be an important factor in driving ZNF number. We estimated the relative sequence diversity of LTR retroelements in each species by extracting their reverse transcriptase coding regions and measuring their total protein tree length. Unsurprisingly, we found that retroelement diversity correlates strongly with retroelement number so it was difficult to distinguish the influence of copy number and copy diversity. As expected given this result, retroelement diversity also strongly correlated with ZNF number, though not quite as strongly as did retroelement copy number (Table 2.2).

The 26 genome assemblies analyzed above are all based on >5-fold sequence read coverage, but they vary in read coverage and in the degree of assembly finishing (Figure 2.2). This variation could affect the apparent retroelement and ZNF gene content differentially, for example by collapsing recent ZNF gene duplicates into apparent single genes. Such variation in assembly quality is difficult to detect and control for, but we made one simple test by repeating the analysis only on the 16 published genomes (higher than average read coverage and finishing effort). Using the LTR retroelement and ZNF domain cutoffs that gave the best correlation across all species, the IC-corrected correlation for published genomes was higher than for all genomes (R-squared 0.71 vs. R-squared 0.67) and the correlation remained highly significant despite the smaller data set (P-value $7.2*10^{-5}$). This result suggests that improved genome assemblies will most likely improve the observed correlations.

In addition to LTR retroelements containing part or all of their internal sequences, vertebrate genomes contain large numbers of solo LTR sequences that arise by recombination between flanking LTRs (Copeland, Hutchison and Jenkins, 1983). It is

very difficult to obtain unbiased counts of solo LTRs because they have no generally shared sequence features. In addition, solo LTRs are more abundant for older retroelements, because they have had more time to recombine since their original insertion. Nevertheless, we assessed correlation between ZNF sequences and total annotated LTR retroelement sequence content (including solo LTRs) for the 16 published vertebrate genomes, since they have the best annotated general repeat content. The IC-corrected correlation was significant (R-squared 0.36, P =0.018), though weaker than for elements with internal sequence. This lower correlation could result from uneven annotation of solo LTR sequences, the expected skew toward older retroelements that are less reflective of recent selective pressure on ZNF genes, or other unknown factors.

Though mammalian tandem ZNF genes are dominated by those encoding a KRAB domain, this domain association is less common in other tetrapods and is absent in fish (Table 2.2; Looman et al., 2002; Birtle and Ponting, 2006; Thomas and Emerson, 2009). The fact that ZNF domain content correlates strongly with LTR retroelement content throughout all of these groups suggests that the function of non-KRAB ZNF genes in other tetrapods and in fish is related to the function of KRAB ZNF genes in mammals. This inference is also supported by similarities in sequence evolution of tandem ZNF genes in each of these groups (see below and Emerson and Thomas, 2009). We speculate that other domains in these taxa play a role analogous to the KRAB domain in mammals, or that tandem ZNF proteins bound to DNA can directly repress transcription.

Recent LTR retroelement activity: The data above reflect an historical aggregate of LTR retroelement activity and tandem ZNF gene duplications. To test whether these characters are temporally correlated, we estimated the age of LTR retroelement insertions based on divergence between the two long terminal repeats of each retroelement (Johnson and Coffin, 1999) and the age of ZNF gene duplicates based on synonymous site divergence (d_S). Though mutation rates surely vary among the species, this variation should not affect relative divergence rates of ZNF genes and LTR retroelements within a species. Using 5% and 10% divergence cutoffs, we found significant IC-corrected correlations between recent LTR retroelement activity and recent tandem ZNF gene duplications across the combined taxa (Table 2.4). The highest correlation was between sequence diversity among recently active LTR retroelements and the number of ZNF domains in recent tandem ZNF gene duplicates (Figure 2.4), but comparisons of the numbers of LTR retroelements and ZNF gene duplicates were also highly significant (Table 2.4). Mouse, opossum, and lizard show evidence of especially high recent LTR retroelement activity, and all three species have correspondingly high numbers of recent tandem ZNF gene duplicates (Figure 2.4,

compare the steepness of the curves near the origin). In addition, all three species show possible evidence of an earlier period of relatively quiescent LTR retroelement activity associated with fewer tandem ZNF gene duplicates, as evidenced by the plateaus on each curve. Alternatively, these plateaus could result from a higher rate of deletion removing older LTR retroelements. Except for a very recent drop in LTR retroelement activity, the patterns on the human lineage suggest relatively slow and constant rates of LTR retrotransposition and ZNF gene duplication. Other genomes with relatively low recent LTR retroelement activity (e.g. horse, elephant, and medaka) have curves broadly similar to that of human (not shown).

Primate LTR retroelements and tandem ZNF gene duplicates: Among existing vertebrate genome sequences, primates provide the densest phylogeny and the best annotation of LTR retroelements. Using RepeatMasker annotations and tandem ZNF gene annotations in humans as a starting point, we investigated in detail the appearance of new LTR retroelements and tandem ZNF genes on the primate lineage (Supplemental Methods). We could divide the primate lineage into 6 distinct branches based on available sequences: a basal primate branch (before the divergence of basal primates), a Simian branch (before the divergence of New World from Old World monkeys), a Catarrhine branch (before the divergence of Old World monkeys from apes), a Hominoid/Hominid branch (before the divergence of orangutan from human; this branch is bisected by gibbons, which currently lack a whole genome assembly), a Hominina branch (before the divergence of chimpanzee from human), and a human specific branch. Using a combination of insertion site analysis and sequence trees of retroelement internal sequences, we defined the branch on which each of 48 primate-specific LTR retroelement families first appeared (Figure 2.5). Another 4 retroelement families were imperfectly resolved, appearing just before or just after the divergence of New World monkeys.

Starting with annotated human tandem ZNF genes, we used a combination of genome sequence searches, maximum-likelihood trees, and synteny to determine the branch on which each new tandem ZNF gene duplicate appeared. Among ZNF gene duplicates, we distinguished between those that diverged by at least 5% in amino acid sequence in an attempt to distinguish between selected duplicates and possibly neutral copy number variation. Since events on each branch of the phylogeny are statistically independent, we analyzed correlations without using IC. Correlation between appearance of new LTR retroelement families and new tandem ZNF genes was remarkably strong and statistically significant. The correlation was highest when ZNF genes with low divergence were excluded and the ambiguous retroelement families were split equally on the two possible branches (Figure 2.5). These results are consistent with our global analysis of vertebrate correlations, suggesting that many or

most tandem ZNF genes in primates arose in response to the appearance of new families of ERVs. The most prominent deviation from perfect correlation is on the Hominoid/Hominid branch, where no new LTR retroelement families but 14 new tandem ZNF genes appeared. The immediately preceding Catarrhine branch was subject to a particularly intense burst of new LTR retroelements (Figure 2.5); we speculate that some of the 14 Hominoid/Hominid ZNF genes arose in response to this slightly earlier burst. In contrast to the human-specific branch, new LTR retroelement families have arisen on the chimpanzee- and macaque-specific branches (Polavarapu et al., 2006; Jern et al., 2006; Rhesus Macaque Genome Sequencing and Analysis Consortium et al., 2008), but the number of families is small and we lack statistical power to test for tandem ZNF gene response.

Predictions for duplicate gene divergence: The hypothesis that most tandem ZNF genes function to repress LTR retroelements predicts certain patterns of molecular evolution driven by the epidemiology of retroelements. First, each host genome should acquire distinct expansions of tandem ZNF genes in response to lineage-specific retroelement challenges. Second, the duplicate genes that comprise these expansions should be subject to positive selection to modify their DNA binding specificity as they adapt to new retroelements. These two predictions have already been confirmed for several of the genomes we analyzed here, including human, mouse, cow, frog, fugu, and zebrafish (Emerson and Thomas, 2009). We extended these analyses to several additional species, namely rat, horse, elephant, opossum, lizard, tetraodon, medaka, and lamprey. In all cases, we found large species-specific clades of tandem ZNF genes with overwhelming evidence of positive selection affecting predominantly nucleotide contact residues of ZNF domains (Figure 2.6).

A final prediction of our hypothesis is that when a new divergent duplicate tandem ZNF gene pair arises, often one of the duplicates will retain the ancestral DNA binding specificity while the other duplicate acquires a new or modified DNA binding specificity that targets a new retroelement. After optimizing its new DNA binding function, the divergent duplicate should be subject to purifying selection. These patterns are expected in cases in which repression of an ancestrally targeted retroelement remains selectively significant, so that only one copy of a duplicate gene pair is free to alter specificity to protect against a new challenge. These patterns are predicted by some other hypotheses for ZNF gene evolution, including when the ancestral tandem ZNF gene has been exapted for host transcriptional regulation(see Discussion). As described in the next sections, we explored these predictions among human duplicate KZNF genes since they are best annotated and there exist multiple closely-related primate genomes that help resolve the time of duplication and ancestral gene identity.

Tracing the origins of human KZNF genes: We identified 34 cases in which two human KZNF genes were clearly closest relatives (see Materials and Methods). Each pair is assumed to have arisen by gene duplication from an ancestral gene at some time during tetrapod evolution. In order to trace the evolutionary history of each pair of genes, we used the two human proteins to find all closely related sequences in a set of increasingly divergent mammalian genomes. Because of the patterns of conservation and divergence detailed below, these searches were remarkably effective in unambiguously tracing the ancestry of each gene.

Considering one duplicate pair, one common result was as follows. In one or more of the most closely related species, one clear copy of each gene was found, indicating that both genes were present in the last common ancestor of human and those species. On deeper branches in the tree, each species had only one gene closely related to the two human genes, suggesting that their last common ancestral species had one copy of the gene, which later duplicated and diverged on the human lineage. The other common result was that clear copies of both genes were found back to some point on the phylogenetic tree, but deeper in the tree no specific ancestral genes were found (more accurately, many possible ancestral genes were found but it was unclear which of them was the true ancestor). Below we consider the latter case first, in which the precise origin of two closely-related human genes is unclear, but the pattern of divergence of the two copies from each other can nevertheless be analyzed.

Divergence of gene pairs of uncertain origin: We could unambiguously identify and analyze 19 human duplicate gene pairs with two copies in a number of species but no clear specific ancestor. Table 2.5 summarizes key features of these duplicates. The phylogenetic depth of the traceable ancestry of the two genes varied from early in the primate lineage to early in the placental mammalian lineage. It may be presumed that the two genes arose by one or more rounds of duplication and divergence from some specific ancestral gene, but the identity and sequence of the ancestral gene is indeterminate. For example, clear copies of both *ZNF273* and *ZNF680* were identified from all five primate species but no species outside of primates.

Three patterns were apparent in most or all cases:

1) Orthologs of each gene were subject to purifying selection across the entire set of DNA binding domains: both the number of ZNF domains and the amino acid sequence of each ZNF domain are highly conserved. Averaged across 280 ZNF domains from 22 genes randomly selected from these duplicates, the nucleotide and phosphate contact residues are among the slowest evolving (Figure 2.7). The most plausible explanation for this pattern is that each orthologous ZNF domain is subject to purifying selection to retain its DNA-binding specificity.

2) A second pattern is evident when comparing two duplicate genes to each other: amino acid changes are more abundant in major nucleotide contact residues than elsewhere (18 of 19 duplicate pairs). When the divergence between the duplicates was relatively low, this difference did not reach statistical significance, but in 12 of 19 duplicates changes were enriched in major nucleotide contact residues with $P < 0.01$ (Fisher's exact test). Summed over all 222 testable ZNF domains from all 19 duplicate pairs, changes between paralogs that are conserved among orthologs occurred in 250 of 666 major nucleotide contact residues (37.5%) but only 489 of 3,552 other residues (13.8%), a highly significant enrichment. This pattern is summarized graphically in Figure 2.7.

3) A third pattern is that entire ZNF domains were often lost or gained in one duplicate gene relative to the other, consistent with previous observations (Looman et al., 2002; Huntley et al., 2006). When such a difference was observed it was strongly conserved among orthologs of each of the two genes, suggesting that these domain arrangements are also subject to purifying selection. Some such events involved insertion or deletion of ZNF domains and others involved point mutations that disrupt the canonical finger structure (Table 2.5). These results suggest that finger gain and loss contribute to changes in DNA binding specificity between duplicate genes.

Duplicate divergence is asymmetric: In the 19 cases discussed above, the absence of an identified ancestral gene in outgroup species precluded analysis of the symmetry of divergence following duplication. In the other 15 duplicate cases, the ancestral gene state could be identified based on the pattern of gene number and gene type in various species (Table 2.6). For example, copies of both human *ZNF557* and *ZNF558* were clearly identified in chimpanzee and orangutan, but only one related gene was found in macaque, marmoset, cow, dog, horse, and rodents, suggesting that a single ancestral gene duplicated on the branch leading to great apes. By comparison of the two duplicates with the single gene from outgroup species, we could address whether divergence occurred in one or both duplicate copies. As shown in Table 2.6, the results usually indicated highly asymmetric divergence of the duplicate genes. Amino acid changes following duplication were strongly biased toward nucleotide contact residues: in total, conserved changes occurred in 155 of 504 major nucleotide contact residues (30.8%) but in only 250 of 2,688 other residues (9.3%). After an initial period of divergence, the divergent copy became subject to purifying selection, since its copies in all descendant species are very similar in amino acid sequence. These patterns suggest that one duplicate retains the ancestral DNA binding specificity, whereas the other duplicate acquires a new or modified DNA binding specificity.

Positive selection following duplication: If new duplicate KZNF genes are subject to selection to acquire new DNA binding specificities, codon-based methods for analyzing selection might be able to detect branch-specific positive selection. We used the

branch-site maximum-likelihood models implemented in codeml to test this possibility. For 17 of the 34 duplicates, highly significant evidence ($P < 0.01$) was obtained for positive selection on the branch connecting the two duplicate copies (Table 2.6). In many cases, the number of available sequences and their total tree length are well below the optimum for detection of positive selection by this method (Anisimova, Bielawski, and Yang, 2001), so it is possible that divergence of all of the duplicates involved positive selection but reached statistical significance only in the strongest cases. As expected, the specific residues implicated in positive selection are strongly enriched in the major nucleotide contact residues. These results directly support the idea that initial duplicate divergence is driven by selection to acquire new DNA binding specificity.

Stable genes: Though duplication is common in the KZNF family, some genes are old and highly conserved. Marsupial mammals diverged from placental mammals about 180 Mya (Kumar and Hedges, 1998; Mikkelsen et al., 2007). Using systematic genome searches, we identified 20 human KZNF genes that have clear orthologs in the marsupial opossum genome and are present in all or nearly all placental mammalian genomes (*ZKSCAN1*, *ZNF3*, *ZNF18*, *ZNF192*, *ZNF202*, *ZNF205*, *ZNF212*, *ZNF213*, *ZNF263*, *ZNF282*, *ZNF398*, *ZNF436*, *ZNF446*, *ZNF496*, *ZNF641*, *ZNF746*, *ZNF764*, *ZNF777*, *ZNF783*, and *ZNF786*). Each gene was present in single copy in opossum and throughout placental mammals and their ZNF domains were invariant in number and highly conserved. These results indicate that some KZNF genes adopted stable functional roles early in mammalian evolution and that they have subsequently retained the same DNA binding specificity. Explanations that reconcile this finding with the retroelement hypothesis are given in Discussion.

Discussion

Based on the striking correlations between LTR retroelement content and C2H2 ZNF domain content throughout vertebrates and over time, we propose that most tandem ZNF genes originate to repress transcription of LTR retrotransposons or retroviruses. The linear-regression lines for the raw correlations pass close to 0 (the origin) LTR retroelements and ZNF domains (Figure 2.1), suggesting that most or all tandem ZNF genes are involved. Consistent with this hypothesis, recent publications show that a mouse KZNF gene represses murine leukemia retrovirus (Wolf and Goff, 2009) and that an unknown suite of KZNF genes probably repress a wide variety of IAP and MusD LTR retroelements (Rowe et al., 2010; Matsui et al., 2010). The vast majority of tandem ZNF genes have no experimentally-determined organismal function, a situation fully compatible with retroelement repression because this function should be difficult to ascertain. Nevertheless, a handful of tandem ZNF genes are implicated in

other processes, including sex-limited gene expression, imprinting, and mouse embryonic development (Krebs et al., 2005; Li et al., 2008; Mackay et al., 2008; García-García, Shibata, and Anderson, 2008). Similarly, some KZNF genes arose early in mammalian evolution and have been retained throughout Therian mammals with nearly invariant DNA binding domains. These genes are unlikely to have current day retroelement-related functions since there is no evidence for such widely shared retroelements. One plausible explanation is simply that some tandem ZNF genes evolved directly to fulfill other host functions and that they were never involved in retroelement repression. Alternatively, it is well-established that transcriptional promoters and enhancers present in retroelements are sometimes exapted for host transcription, following chance integration in an appropriate location to confer useful transcriptional regulation on a host gene (reviewed in Cohen, Lock, and Mager, 2009). A few studies provide indirect evidence that host exaptation of retroelement regulatory sequences may be extremely common (Lowe, Bejerano, and Haussler, 2007; Conley, Piriyapongsa, and Jordan, 2008). In addition, in at least two cases a retroviral gene itself appears to have been adopted for a host function (Best et al., 1996; Dupressoir et al., 2009). Thus it is possible that tandem ZNF genes that now function as host transcription factors initially evolved to repress LTR retroelements and were later retained on the basis of their regulatory role for a host gene.

In mammals, transposition competence of new ERV families is usually relatively transient, decaying over a period of several million years; after transposition specific ERV sequences evolve neutrally and eventually lose protein coding function and transcription competence (e.g. de Parseval and Heidmann, 2005; Stocking and Kozak, 2008). During this transition, selection to retain specifically protective ZNF genes will attenuate. Unless they are exapted for a distinct host function, most such ZNF genes should eventually become pseudogenes or be deleted from the genome. This predicted pattern has not been analyzed in detail, but the general expectation of abundant ZNF pseudogenes is clearly met in the human genome and probably in other genomes (Figure 2.3).

The state of ERVs and tandem ZNF genes in the mouse is of particular interest because further experimental tests of our hypothesis are most feasible there. Our data show that the mouse reference genome assembly has an unusually high number and diversity of ERVs relative to tandem ZNF genes. One possible explanation is that tandem ZNF gene response in mouse lags behind the recent high ERV activity that is known in mouse (Stocking and Kozak, 2008), but alternative explanations are possible. First, mouse has a higher rate of genomic deletion than human (Mouse Genome Sequencing Consortium et al., 2002), which should remove older ERVs more quickly, potentially freeing ZNF genes involved in their defense for directional positive selection to protect against new ERV challenges. This possibility may be testable by a focused

analysis of the patterns of duplication and positive selection among mouse ZNF genes. Second, a strong recent evolutionary ZNF response to high ERV activity is expected to result in an abundance of unfixed ZNF gene duplicates, causing heterogeneity in the number of ZNF genes in mouse populations. This possibility predicts that sequences from other wild *Mus musculus* isolates will vary in ZNF gene content, with some isolates having more or fewer ZNF genes than the reference genome. Finally, though the mouse genome assembly is one of the highest quality we analyzed, it is possible that a recent burst of ZNF gene duplication would be obscured by assembly collapse of multiple similar paralogs, resulting in an underestimate of ZNF gene number.

In mammals, the large majority of LTR retroelements are clearly ERVs as indicated by the presence of a viral ENV gene or close relatedness to a known retrovirus. Outside of mammals, this relationship is less clear. Fish appear to have relatively few ERVs and a large burden of LTR retrotransposons with no clear retroviral connection (Basta, Buzak, and McClure, 2007; Basta et al., 2009). In chicken and finch, most retroelements are classified by RepeatFinder as ERVs but detailed analysis is lacking, and in lizard no analysis of LTR retroelements is available. Retroviruses have repeatedly evolved from vertically transmitted retrotransposons by acquisition of an ENV gene (e.g. Doolittle and Feng, 1992; Laten, Majumdar, and Gaucher, 1998; Malik, Henikoff, and Eickbush, 2000). Conversely, integrated retroviruses can readily convert to vertically transmitted transposons by loss of the envelope gene (Ribet et al., 2008). It is possible that tandem ZNF genes repress exogenous retroviruses, endogenous retroviruses, and LTR retrotransposons, but the balance of activities for these groups remains unclear and may vary in different species.

Implications for retroviral repression in mammals: Judging from patterns of endogenized retroviral sequences, mammals have been subject to an ongoing barrage of retroviral infections of diverse types (Blikstad et al., 2008; de Parseval and Heidmann, 2005). New retroviral infections in a particular species can arise by a shift or expansion of host range by a retrovirus that infects another species (e.g. Benveniste and Todaro, 1974; Martin et al., 1999; Chen et al., 1996; Gao et al., 1999). The consequence for the new host is the occasional appearance of an unpredictable new retroviral challenge. If a retrovirus successfully integrates in the germ cell lineage, it may also result in the spread of a new deleterious ERV in the host genome. It is well established that mammals combat retroviral infection in multiple ways, including attacking viral RNA with APOBEC cytidine deaminases and ZAP, interfering with viral capsid with Fv1 and TRIM5alpha, and preventing viral particle release with Tetherin (reviewed in Wolf and Goff, 2008). The pervasiveness of retroviruses in vertebrates and the multiple layers of viral restriction by the host support the idea that there should also be strong selection on the host to repress retroviral transcription. The size, diversity,

and rapid evolution of the tandem ZNF gene family suggests that these genes fill this role.

The sequence divergence patterns of new duplicate genes suggests the following model for the contribution of KZNF genes to host response to a new retroviral infection in mammals. Starting from either a new duplicate gene or a pre-existing copy number polymorphism, a KZNF gene with significant, even minor, off-target binding to a new retroviral sequence is driven to fixation and starts to evolve improved target recognition by changes in amino acid sequence and changes in ZNF number. This pattern of duplicate evolution corresponds in many ways to that proposed for bacterial genes (Bergthorsson, Andersson, and Roth, 2007). The initial off-target binding to a new retrovirus may arise purely by chance or may result from sequence relatedness of the new retrovirus to a previously encountered retrovirus for which the host has already evolved a cognate KZNF gene. If the previously encountered retrovirus (or its endogenized copies) remain selectively significant for the host, there will be pressure for one copy of the KZNF gene to retain its ancestral DNA binding specificity and for adaptation to the new retrovirus to act on the other copy. If the previously encountered retrovirus is no longer selectively significant for the host, targeting a new retrovirus could be achieved by directional selection on an ancestral KZNF gene without gene duplication, though we didn't observe any clear instances of this pattern.

Other possible evolutionary drivers: A number of other potential drivers of tandem ZNF gene duplication and divergence have been suggested and probably apply in specific cases, listed below. Based on the expansion and diversification of KZNF gene sequence and expression patterns on the primate lineage, it has been suggested that these genes underlie the evolution of novel primate traits including an enlarged brain (Hamilton et al., 2003; Nowick et al., 2009; Nowick et al., 2010). Based on expansion of genes in a cluster of KZNF genes in mouse that includes two genes that modify sex-limited expression of other genes, it has been suggested that KZNF genes play a role in speciation via modification of sex-specific traits (Krebs et al., 2005). One KZNF gene with an ortholog in mouse (*ZNFp57*) and human (*ZNFP57*) has been shown to be required for genomic imprinting at several loci (Li et al., 2008; Mackay et al., 2008). Since imprinting involves maternal-zygotic conflict (e.g. Smith, Garfield, and Ward, 2006), this process has the potential to drive KZNF duplication and diversification. Finally, the *PRDM9* tandem ZNF gene is strongly implicated in specification of recombination hotspots (Myers et al., 2010; Baudat et al., 2010; Parvanov, Petkov, and Paigen, 2010). Though recombination hotspots evolve rapidly, the domain structure and evolution of *PRDM9* are clearly different from all other tandem ZNF genes (Oliver et al., 2009; Thomas et al. 2009) and it has not been subject to the expansion seen in the genes described here. None of these explanations alone suffice to explain the general

correlations between genomic LTR retroelement content and tandem ZNF coding potential. In contrast, the established potential for host exaptation of retroviral regulatory elements provides a plausible mechanism by which tandem ZNF genes initially selected for retroelement repression could over time adopt a variety of other host functions.

Materials and Methods

Species key and genome assemblies: Mammals: hsap = human = *Homo sapiens* (hg18), ptro = chimpanzee = *Pan troglodytes* (panTro2), ppyg = orangutan = *Pongo pygmaeus abelii* (ponAbe2), mmul = macaque = *Macaca mulatta* (rheMac2), pham = baboon = *Papio hamadryas* (papHam1), cjac = marmoset = *Callithrix jacchus* (calJac3), btau = cow = *Bos taurus* (bosTau4), cfam = dog = *Canis familiaris* (canFam2), ecab = horse = *Equus caballus* (equCab2), mmus = mouse = *Mus musculus* (mm9), rnor = rat = *Rattus norvegicus* (rn4), cpor = guinea pig = *Cavia porcellus* (cavPor3), lafr = elephant = *Loxodonta africana* (loxAfr3), ocun = rabbit = *Oryctolagus cuniculus* (oryCun2), mdom = opossum = *Monodelphis domestica* (monDom5), oana = platypus = *Ornithorynchus anatinus* (ornAna1). Non-mammalian tetrapods: acar = lizard = *Anolis carolinensis* (anoCar1), tgut = finch = *Taeniopygia guttata* (taeGut1), xtro = frog = *Xenopus tropicalis* (xenTro2), ggal = chicken = *Gallus gallus* (galGal3). Teleost fish: drer = zebrafish = *Danio rerio* (danRer6), olat = medaka = *Oryzias latipes* (oryLat2), gacu = stickleback = *Gasterosteus aculeatus* (gasAcu1), trub = fugu = *Takifugu rubripes* (fr2), tnig = tetraodon = *Tetraodon nigroviridus* (tetNig2). Jawless fish: pmar = lamprey = *Pteromyzon marinus* (petMar1). All genome assemblies were obtained from the UCSC Genome Browser Gateway (http://genome.ucsc.edu). For analysis of primate LTR retroelement and tandem ZNF gene origins, additional low coverage genome assemblies were obtained from ENSEMBL (htpp://www.ensembl.org).

Retroelement counts: RepeatMasker data were unavailable for several genomes of interest and misleading for others, apparently because some genomes contain abundant retroelement sequences that do not yet appear in the RepBase sequences used as queries by RepeatMasker (Smit, Hubley, and Green, 2010). To make counts of retroelements in a manner independent of repeat annotation status and species phylogeny, we used the fact that LTR retroelements are distinguished from all other known sequences by the appearance of a characteristic pattern of conserved coding elements, namely protease, reverse transcriptase, RNaseH, and integrase domains (many retroelements also encode gag and env proteins, but these are poorly conserved across the broad phylogenetic space we wished to analyze).

To detect matches to the characteristic LTR retroelement protease, reverse transcriptase, RNaseH, and integrase (most elements) or tyrosine recombinase (*DIRS*

superfamily elements) domains we used the Pfam database profiles PF00077 (retroviral aspartyl protease type 1), PF08284 (retroviral aspartyl protease type 2), PF00078 (reverse transcriptase), PF00075 (RNaseH), PF00665 (Integrase catalytic), and cd00799 (Cre recombinase, closely related to *DIRS* tyrosine recombinase). Rpsblast was run with all the profiles on each target genome with a very permissive E-value to allow for fragmentary domain matches expected from older neutrally-evolving retroelements. Matches with a heuristically-chosen blast score of 22 or higher were retained and were sorted in genome order. A custom local dynamic programming algorithm was applied to extract matches within 3 kb of each side of reverse transcriptase (RT) matches and in one of the expected LTR retroelement orders. The algorithm constrained the aligned matches as follows: [0 or more PROT (PF00077 or PF08284)] – [1 or more RT] – [0 or more RNH] – [0 or more INT].

ZNF domain searches: We performed a search for zinc finger domains on selected genomes using the program rpsblast with the –p F option (6-frame translation of DNA query). The search profile consisted of a 28 amino acid weight matrix profile of the ZNF domain (including the 7 amino acid linker region upstream of the 21 amino acid ZNF core). This profile was generated from the set of functional human tandem ZNF proteins using the psiblast program as directed in the NCBI blast documentation. Subsequent analysis showed that this profile is nearly identical to profiles derived from tandem ZNF proteins from other species (examples shown in Figure 2.8). Genome searches were carried out in two forms: one search of the entire genome assembly and a second search of all open reading frame (ORF) segments of 100 codons or longer. From ORF searches, we counted: 1) the number of ORFs with one or more ZNF match above some score cutoff, and 2) the total number of ZNF matches above some score cutoff. Counts from all searches with various score cutoffs are reported in Table 2.2. To be sure that the ZNF domain matches reflect bona fide tandem ZNF genes, we determined the ZNF domain profile and the number of tandem ZNF domains for each genome with large numbers of ZNF genes (examples shown in Figure 2.8). To determine the ZNF domain score distribution expected for genes and pseudogenes, all human ZNF domain matches were divided into those in known RefSeq genes (plus a few probable genes not yet appearing in RefSeq; Huntley et al., 2006) and those outside genes (which are likely to belong to pseudogenes and gene fragments). We made a density histogram of the rpsblast scores for each group (representing how well each hit matches the ZNF profile) and superimposed the histograms (Figure 2.3). The pseudogene scores presumably reflect a distribution of times of neutral evolution since pseudogenization. The crossover point where the hit density for genes first exceeds the hit density for pseudogenes occurs at about rpsblast score 57. This crossover point is close to the peak correlations of LTR retroelements and ZNF domains (Table 2.2).

Phylogenetic correction by independent contrasts: Comparing two characters in a scatter plot assumes statistical independence, an assumption that is violated when related species are used as data points (Felsenstein, 1985; Garland et al, 2005). This can create spurious correlations across broad phylogenies (Whitney and Garland, 2010). To account for such phylogenetic concerns we tested our data using Felsenstein's independent contrasts method implemented in the PDAP package (Midford et al, 2005) of Mesquite (Maddison and Maddison, 2010). Specifically, we used the positivized x vs. y contrasts (mode 9 in PDAP) to measure the correlation between respective ZNF metrics and ERV metrics. The tree used in the analysis was based on best estimates for species divergence times derived largely from TimeTree (http://www.timetree.org/; Hedges and Kumar, 2006). Pearson correlations were forced to go through the origin.

Genomic positions of ZNF genes and LTR retroelements: It was formally possible that a high local genome-position correlation of ZNF genes and LTR retroelements might explain some of the correlation reported in Table 2.1, Table 2.2, and Figure 2.1. This possibility is remote, because ZNF domains are strongly clustered and occupy only a tiny fraction of each genome (e.g. for 100 Kb bins, 0.9% and 1.4% of the human and mouse genomes respectively have a significant rpsblast match to the Pfam ZNF domain profile). Nevertheless, we systematically tested correlations in genome positions among ZNF domains and LTR retroelements (Table 2.7). For most mammalian genomes we found a statistically significant but very weak correlation (mean R-squared 0.003). For non-mammalian genomes correlations were even weaker and most were only marginally significant. We conclude that colocalization in the genome cannot explain more than a tiny part of the general correlations.

Identification of closest human gene pairs and their orthologs in other species: Pairs of human duplicate KZNF genes were identified as reciprocal best blastp matches or as neighbors on a pairwise distance tree among all human KZNF genes, with further tests to eliminate unclear gene pairs. Orthologs of these human tandem ZNF genes were identified as follows.

The protein encoded by the ZNF-containing exon from human was used as a tblastn query to the entire set of available Eutherian genome assemblies. DNA corresponding to the two best tblastn matches from each species was extracted. Potential human outgroup paralogs were identified by using the entire set of candidate orthologous proteins as queries in a blastp search of all human ZNF proteins. The single best human match to each candidate ortholog and the 20 best matches overall were extracted, redundant matches were removed, and these proteins were appended to the candidate ortholog set. This final set of proteins was aligned using Dialign-tx (Subramanian et al. 2008) and a maximum-likelihood protein tree was made using phyml (Guindon and Gascuel 2003).

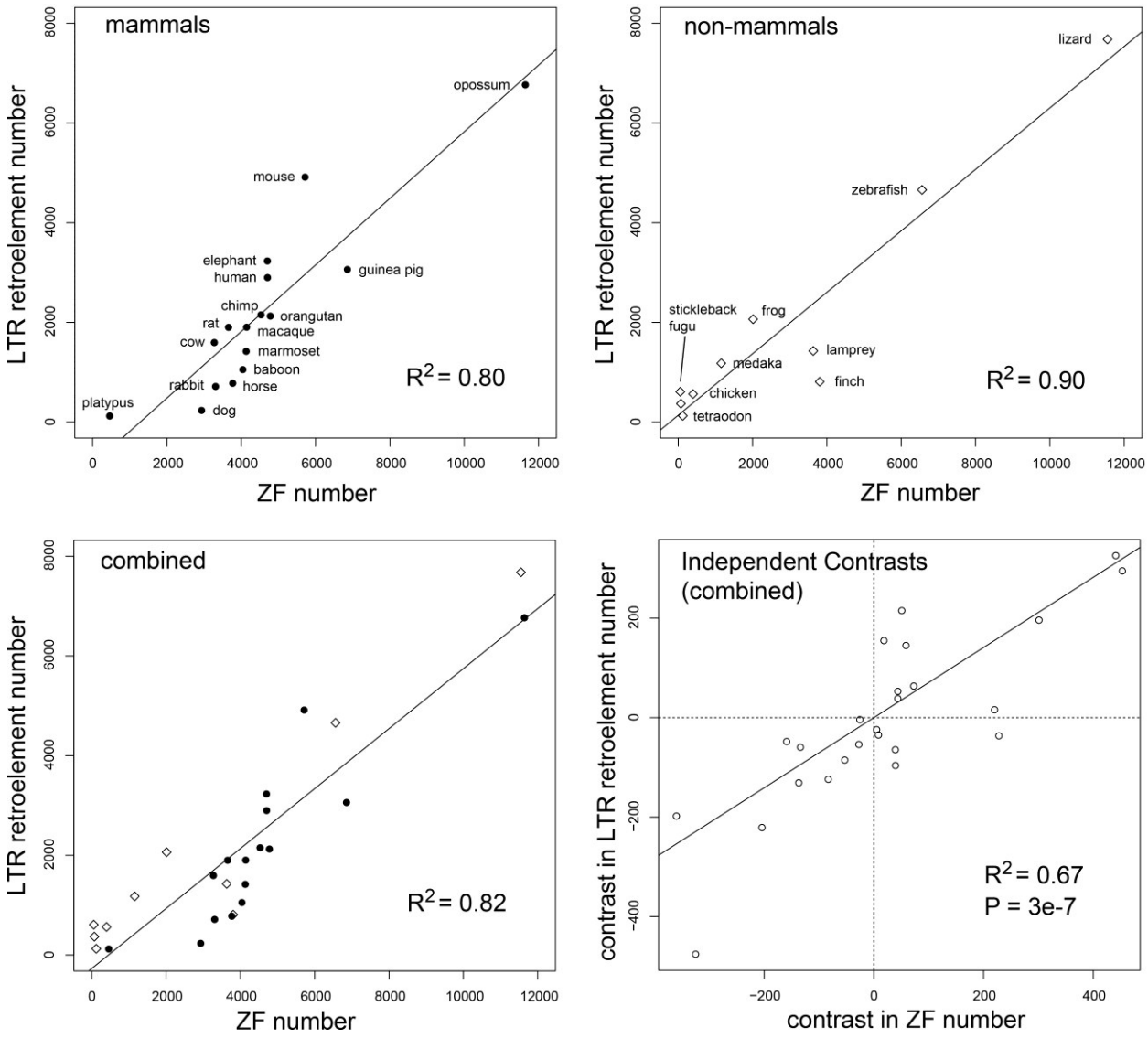
Rate of change in orthologous fingers: We measured the relative rates of change of specific amino acid residues in orthologous KZNF domains as follows. 11 KZNF gene pairs were arbitrarily selected from among the duplicate pairs with an indeterminate ancestral gene. For each pair, all available orthologs were gathered from the five primates, cow, dog, and horse. The 22 genes encoded a total of 280 ZNF domains. Each orthologous finger protein set was "aligned" (these are gap-free alignments since all fingers conformed to the standard 28 amino acid finger domain) and PhyML 3.0 was used to estimate rates of change at each site (JTT matrix, 20 rate categories, gamma parameter 1.0; Guindon and Gascuel, 2003; Guindon et al., 2009). The "lk" output file from PhyML gives the likelihood that each site belongs in each of the 20 estimated rate categories (Anisimova and Gascuel, 2006). The peak likelihood rate value was extracted for each position in each ZNF domain. These rates were averaged across all 280 orthologous finger groups. Note that this method does not measure the absolute divergence of the sequences, which varied from gene to gene depending on available orthologs. Since all 28 positions were present in all aligned finger sets, this method does produce an average estimate of the relative rates of change at each ZNF site, as plotted in Figure 2.7.

Asymmetry calculation: For duplicate KZNF genes with an identified ancestral gene, asymmetry of divergence between the two duplicates was computed as follows. For each aligned site, the ancestral state was inferred when all copies of the ancestral gene (the single copy gene present in early branching species) encoded the same amino acid. At such sites, when all copies of each of the two duplicate genes encoded the same amino acid and at least one gene diverged from the ancestral state (i.e. the site changed and was conserved among orthologs) the site was counted as informative. When duplicate copy 1 (the copy overall most similar to the ancestral state) was divergent the site received a score of -1; when duplicate copy 2 was divergent the site received a score of +1; when both were divergent the site received a score of 0. When averaged across all informative sites, the expected score is 0 if divergence is perfectly symmetric and the expected score is 1 if divergence is perfectly asymmetric.

Tests for positive selection: For species-specific expansion analysis, clades of species-specific tandem ZNF exons were collected and analyzed by site models 7, 8, and 8A implemented in codeml PAML 3.15 (Zhang, Nielsen, and Yang, 2005; Yang, Wong, and Nielsen, 2005). Additional details are described in Emerson and Thomas, 2009. Strong evidence of positive selection was detected in 30 of 35 clades. To determine the types of protein sites subject to positive selection, ZNF sites with Bayes-Empirical-Bayes P-values of 0.98 or higher were counted summing over all the clades. Counts for each of the 28 classes of ZNF sites are shown in Figure 2.6. For each human duplicate gene pair, branch-site models implemented in codeml were applied to test for branch-specific positive selection (Zhang, Nielsen, and Yang, 2005; Yang,

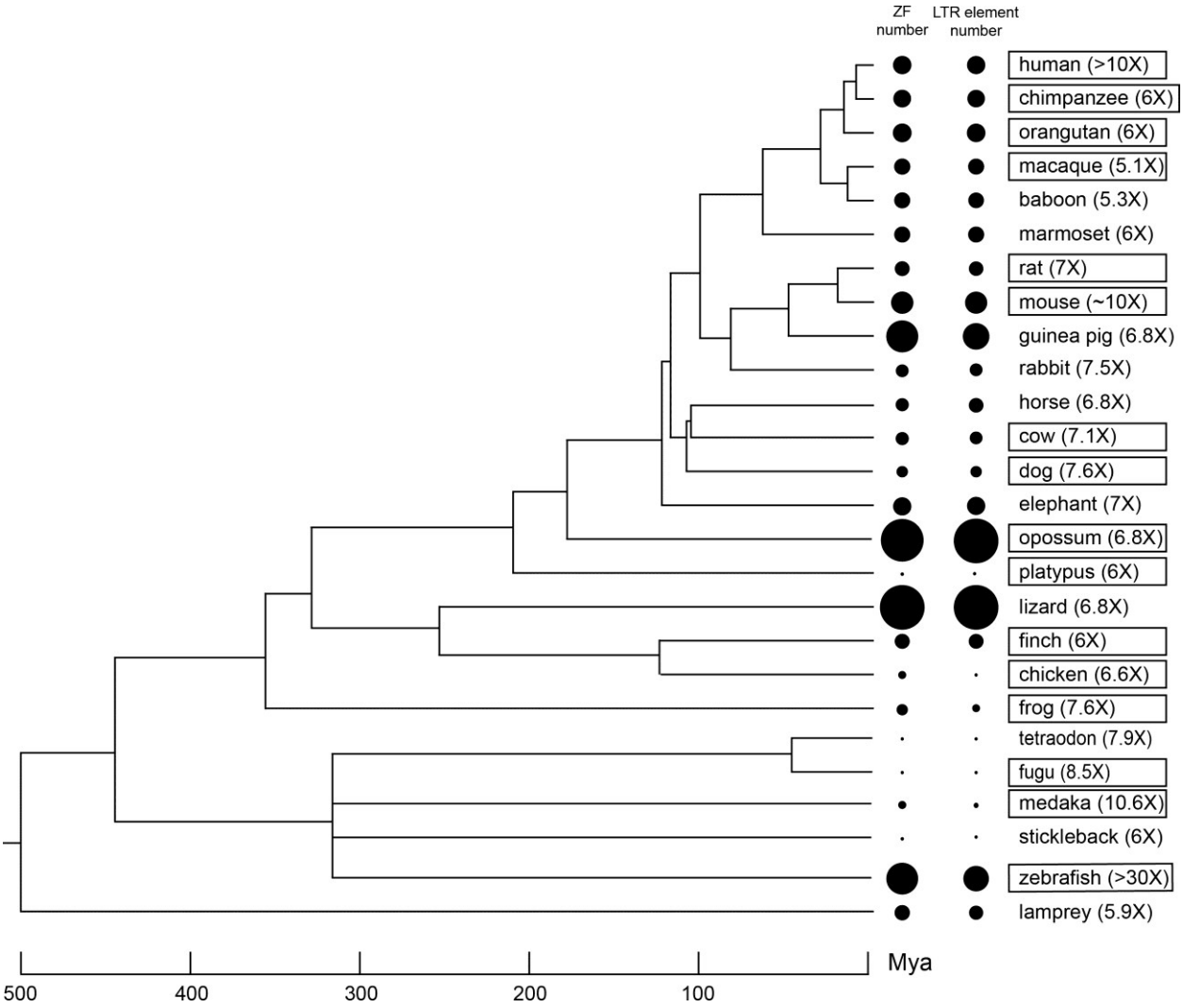
Wong, and Nielsen, 2005). For the control model A1, the foreground d_N/d_S was constrained to be 1.0 on all branches of the tree, whereas for selection model A, the foreground d_N/d_S was allowed to differ on the branch joining the duplicate copies. Statistical analysis was based on a chi-square test of twice the difference in log likelihoods between the two models with one degree of freedom. Specific codeml output details and statistics are shown in Table 2.8.

Figure 2.1: Correlation of genomic LTR retroelements and ZF domain content



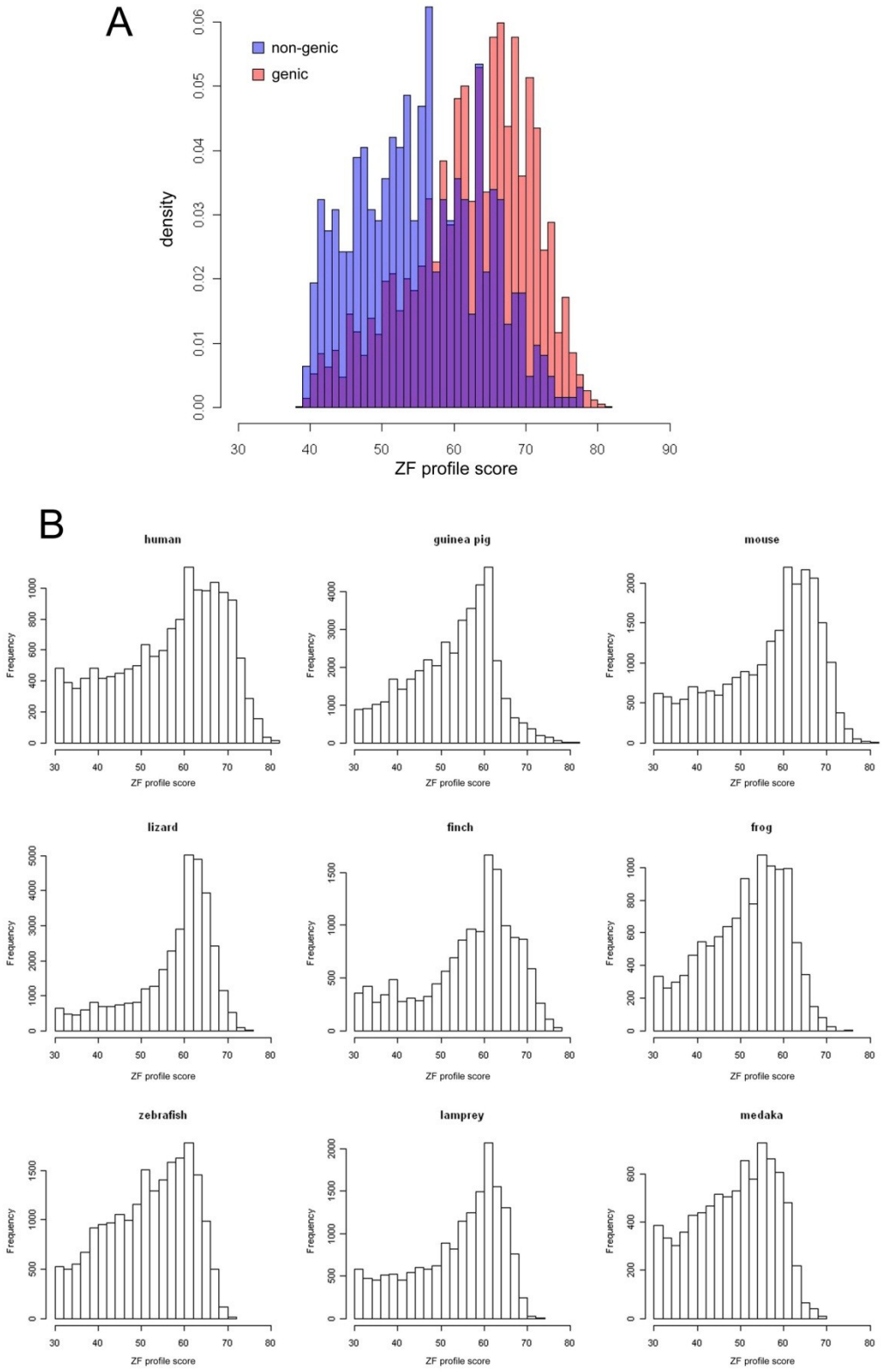
Three panels show the number of detected LTR retroelements plotted against the number of detected ZF domains in different vertebrate groups or all groups combined. The lines show the linear least-squares best fit with its squared correlation coefficient R^2. The fourth panel (lower right) shows the combined data after correction for phylogenetic relatedness by the method of Independent Contrasts. The line is the linear least-squares best fit forced to go through the origin and its associated R^2 and P value. The summed score cutoff for LTR retroelements was 50. The ZF number was determined from all genomic open reading frames with 4 or more ZF domain matches with a minimum average score of 55. These counting criteria gave the maximum correlation for combined data, but a wide variety of other counting criteria also gave highly significant correlations (Table 2.2).

Figure 2.2: Phylogenetic tree and the number of LTR retroelements and ZF domains detected in vertebrate genomes



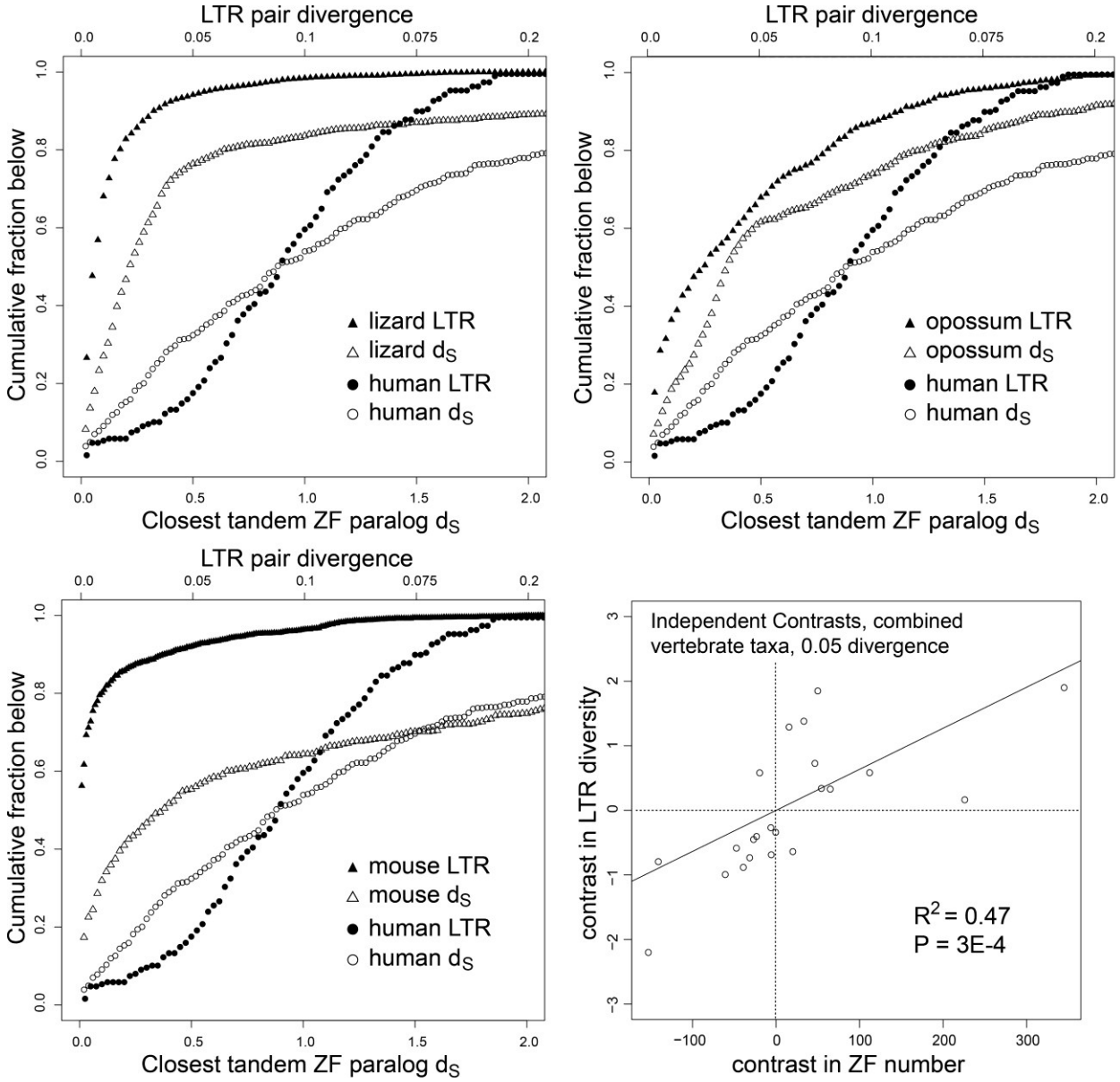
The phylogenetic tree was drawn by hand using estimates of branch points according to TimeTree data summaries (Hedges et al., 2006). The tree was also used for Independent Contrast calculations. At each tree tip the number of detected tandem ZF domains and LTR retroelements is proportional to the diameter of the circle. Following each species identifier is a summary of sequencing read coverage, derived from summaries given on the UCSC genome browser and the relevant genome sequencing center reports. Published genomes used in some analyses are boxed.

Figure 2.3: Histograms of ZF domain profile matches



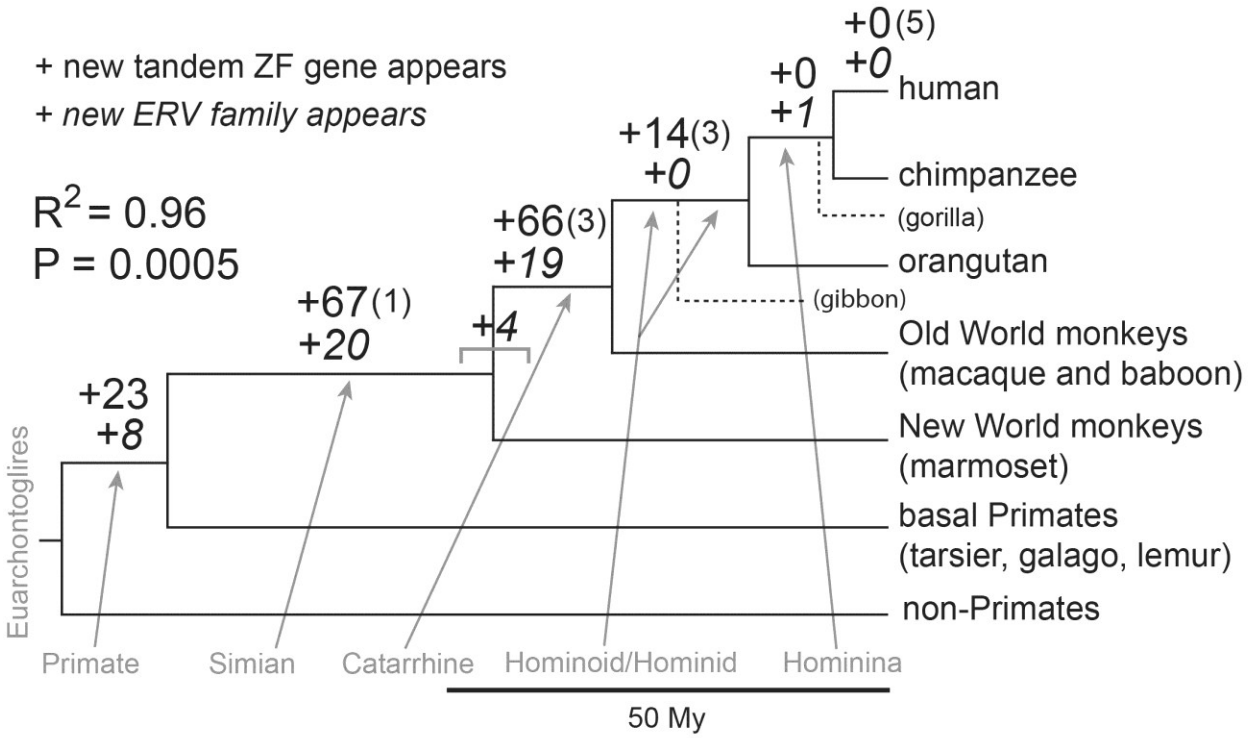
Panel A shows the density of profile scores of human ZF domains in coding genes (red) and those in non-coding regions (blue). Most of the non-genic matches are associated with non-coding duplicates of ZF genes (pseudogenes). Panel B shows the frequencies of profile scores of ZF domains for representative genomes. The leftward tail in each histogram presumably represents (mostly) pseudogene matches.

Figure 2.4: ZF gene duplicate and LTR divergence time courses.



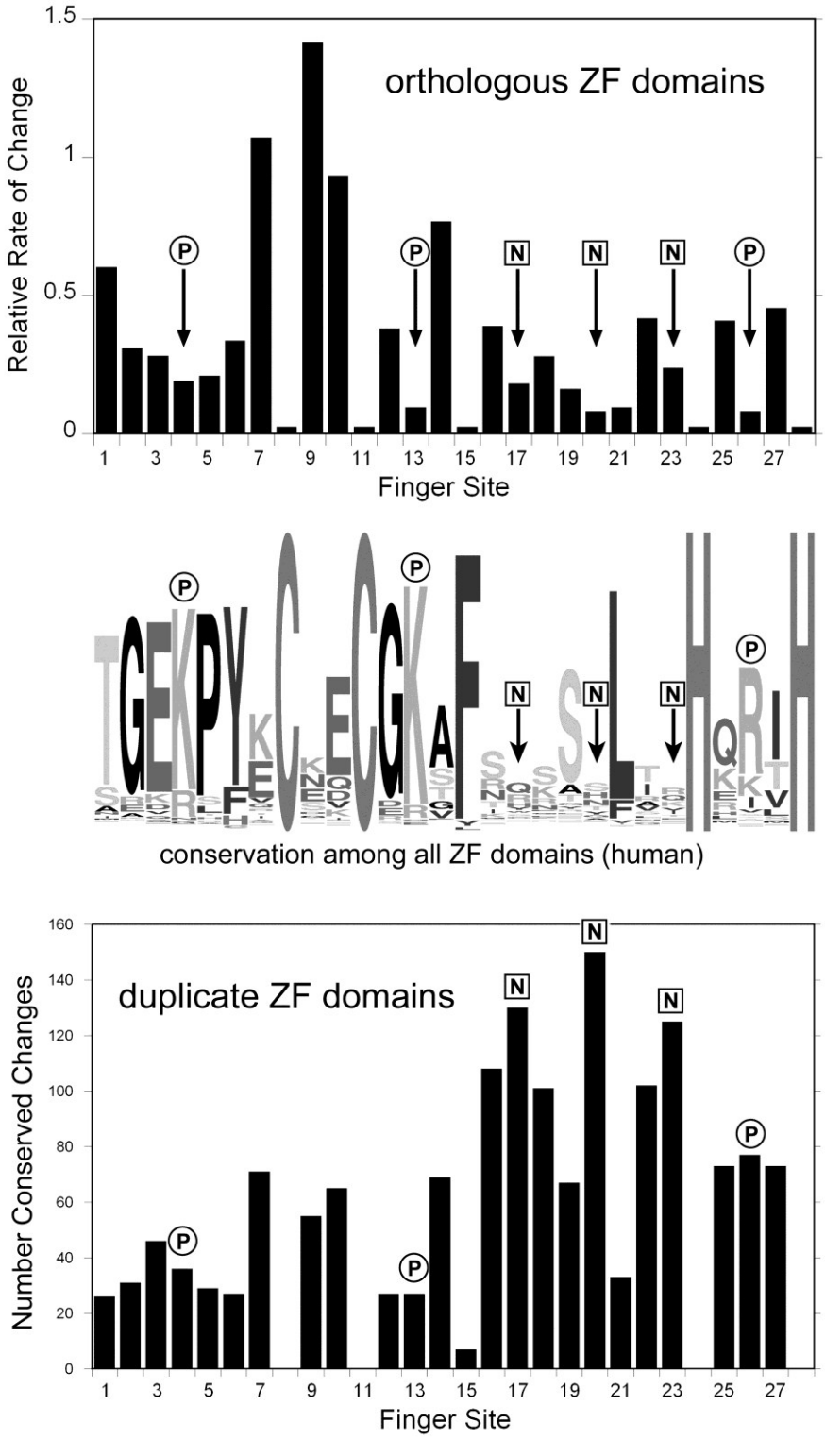
Three panels show cumulative histograms of LTR nucleotide divergence and closest ZF paralog d_S for the indicated species. The axes have been scaled to best display the full curve for both data sets. The human data are included in all three panels for comparison. The fourth panel (lower right) shows statistical analysis at or below one divergence point (0.05 LTR divergence/0.05 paralog d_s) for all species combined after correction by Independent Contrasts (see Table 2.4). The line is the linear least-squares best fit forced to go through the origin and its associated R^2 and P value.

Figure 2.5: Primate phylogeny with the appearance of new endogenous retroviral families and new tandem ZF genes



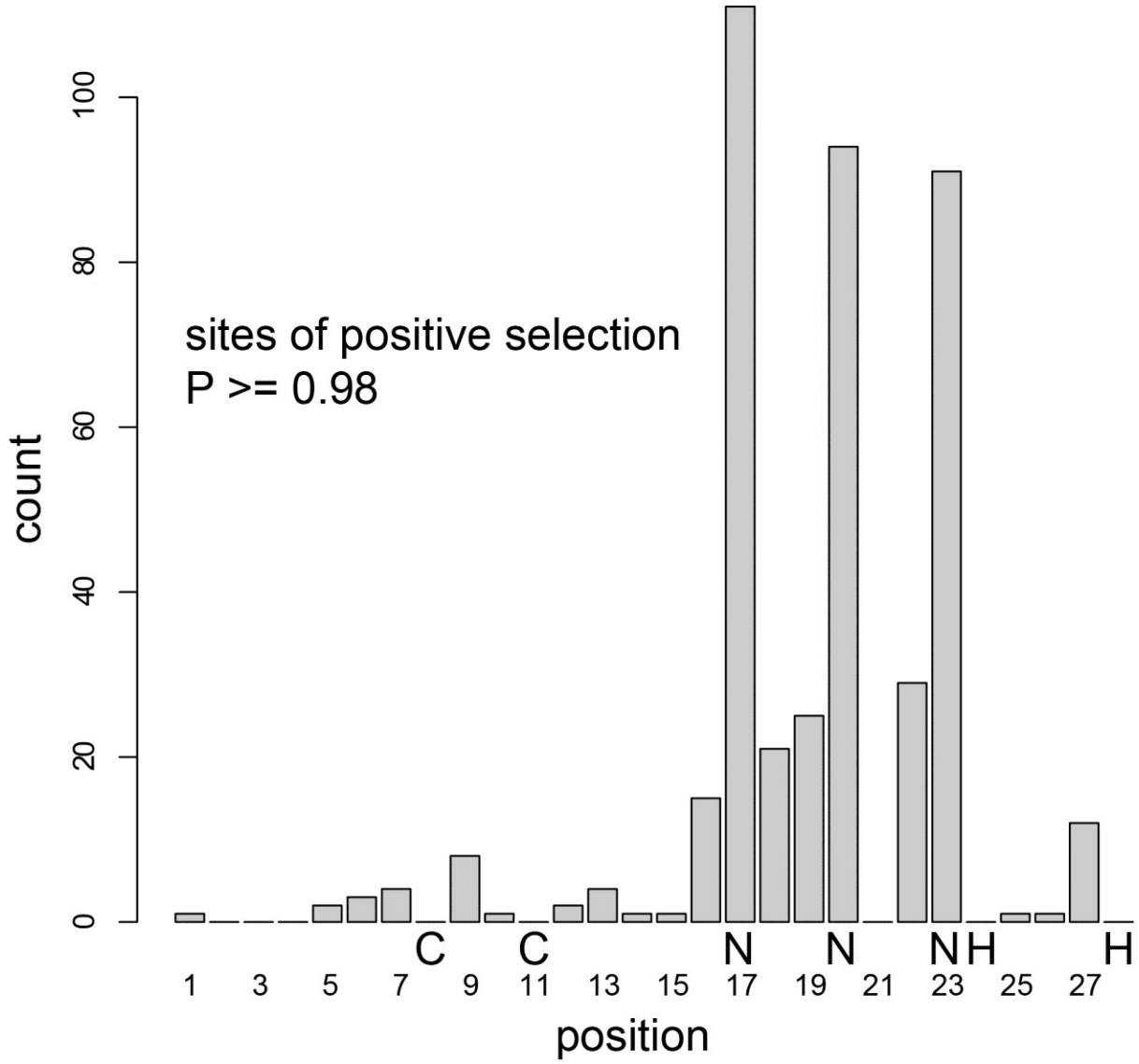
Data were derived by tracking the first appearance of human ERV families and tandem ZF genes. On each branch leading to the human, the top number indicates the number of tandem ZF gene duplicates (with additional duplicates that diverged by less than 5% in amino acid sequence in parentheses) and the bottom number indicates the number of new ERV families. Four families of ERVs could not be confidently assigned to a specific branch and are shown straddling the Simian/Catarrhine branch point. The gorilla genome is low coverage and was not systematically analyzed, but the single ERV (*HERV-Fc1*) that appears on the branch leading to human and chimpanzee is clearly present in gorilla (not shown).

Figure 2.6 Changes in orthologous zinc fingers and duplicate zinc fingers compared to diversity among all fingers



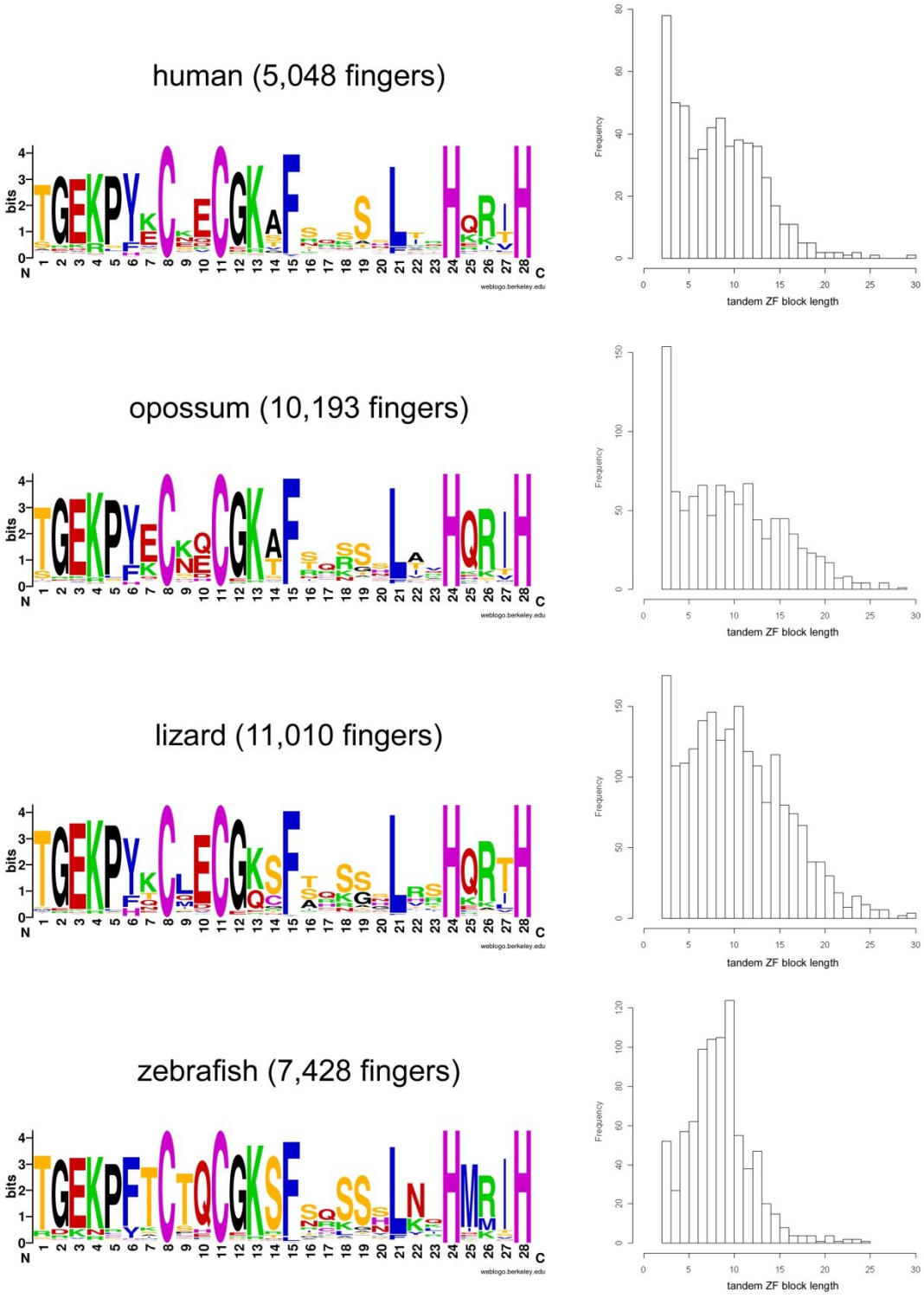
The top panel shows the averaged relative rates of divergence in 280 orthologous ZF domains from 22 randomly chosen KZNF genes. The center panel shows the diversity among all ZF domains from human KZNF genes as a logo plot (http://weblogo.berkeley.edu). The lower panel shows the number of conserved changes observed among the 390 testable zinc fingers in all 34 duplicate gene pairs analyzed. Circles labeled P indicate residues that make major phosphate contacts and squares labeled N indicate residues that make the major nucleotide contacts.

Figure 2.7: Sites of positive selection in ZF domains among all species-specific expansions



Bar plot of the number of sites at each position in the ZF domain with strong evidence of positive selection (codeml Bayes-Empirical-Bayes P-value 0.98 or higher). The X-axis is the position within the ZF domain. The canonical C2H2 residues that define the zinc finger are shown below the bars, together with the positions of major nucleotide contacts (labeled “N”).

Figure 2.8: ZF domain matches are dominated by bona fide tandem ZF regions



Each row shows a logo plot made from all the canonically spaced ZF domains (Cx2Cx12Hx3H) found in each genome with their upstream 7 amino acids added, and a

histogram of the length of blocks of strictly tandem ZF domains among the same ZF domains. The data for the graphs were derived by extracting all genomic ORFs with 4 or more ZF domain matches with an average profile score of 50 or more, and retaining those that contain 4 or more ZF domains with canonical spacing (the vast majority for each genome). The zebrafish logo plot has a few differences from the others, but these are not shared in other fish (data not shown). Zebrafish also has shorter tandem ZF domain blocks and this pattern is shared with other fish (data not shown).

Table 2.1: Statistical tests for correlation of LTR retroelement counts and ZF domain counts.

		ORFs with 4 or more ZF domains									
		ZF minScore 40 [3]		ZF minScore 45		ZF minScore 50		ZF minScore 55		ZF minScore 60	
LTR cutoff [1]	ZF count type [2]	R squared [4]	P value [4]	R squared	P value	R squared	P value	R squared	P value	R squared	P value
minScore 80	ORF	0.300	$3.8*10^{-03}$	0.318	$2.7*10^{-03}$	0.363	$1.1*10^{-03}$	0.436	$2.4*10^{-04}$	0.575	$7.2*10^{-06}$
	ZF	0.517	$3.5*10^{-05}$	0.548	$1.6*10^{-05}$	0.606	$2.8*10^{-06}$	0.671 [5]	$3.1*10^{-07}$	0.668	$3.4*10^{-07}$
minScore 150	ORF	0.243	$1.0*10^{-02}$	0.261	$7.7*10^{-03}$	0.303	$3.6*10^{-03}$	0.375	$8.8*10^{-04}$	0.526	$2.7*10^{-05}$
	ZF	0.455	$1.6*10^{-04}$	0.486	$7.6*10^{-05}$	0.546	$1.6*10^{-05}$	0.617	$2.0*10^{-06}$	0.638	$9.8*10^{-07}$
minScore 200	ORF	0.222	$1.5*10^{-02}$	0.238	$1.1*10^{-02}$	0.281	$5.4*10^{-03}$	0.352	$1.4*10^{-03}$	0.513	$3.9*10^{-05}$
	ZF	0.424	$3.1*10^{-04}$	0.455	$1.6*10^{-04}$	0.516	$3.6*10^{-05}$	0.590	$4.6*10^{-06}$	0.622	$1.7*10^{-06}$

[1] Score cutoff for counting an LTR retroelement match in each genome.

[2] ZF counts were made either using the number of ORFs containing 4 or more ZF domains (ORF) or by counting the total number of ZF domains in ORFs with 4 or more ZF domains (ZF).

[3] The minimum rpsblast score required for each ZF domain match, as used for the two types of ZF counts. For example if minScore is 40, then ORF count is the number of ORFs with 4 or more ZF domains at or above score 40, and ZF count is the number of ZF domains in all ORFs at or above score 40.

[4] R-squared and P values were computed by the method of Independent Contrasts (Midford et al, 2005; Maddison and Maddison 2010).

[5] Figure 1 is a graph of the data for these cutoff values (the peak for the values in this table).

Table 2.2: Data and statistical analysis of vertebrate LTR retroelement and ZF domain content

Statistical Summaries:

		ORFs with 4 or more ZF domains									
		ZF4 minScore 40		ZF4 minScore 45		ZF4 minScore 50		ZF4 minScore 55		ZF4 minScore 60	
LTR retroelement cutoff	ZF count type	R squared	P value	R squared	P value	R squared	P value	R squared	P value	R squared	P value
minScore 80	ORF count	0.300	3.8E-03	0.318	2.7E-03	0.363	1.1E-03	0.436	2.4E-04	0.575	7.2E-06
	ZF count	0.517	3.5E-05	0.548	1.6E-05	0.606	2.8E-06	0.671	3.1E-07	0.668	3.4E-07
minScore 150	ORF count	0.243	1.0E-02	0.261	7.7E-03	0.303	3.6E-03	0.375	8.8E-04	0.526	2.7E-05
	ZF count	0.455	1.6E-04	0.486	7.6E-05	0.546	1.6E-05	0.617	2.0E-06	0.638	9.8E-07
minScore 200	ORF count	0.222	1.5E-02	0.238	1.1E-02	0.281	5.4E-03	0.352	1.4E-03	0.513	3.9E-05
	ZF count	0.424	3.1E-04	0.455	1.6E-04	0.516	3.6E-05	0.590	4.6E-06	0.622	1.7E-06

		whole genome ZF domains									
		ZF minScore 40		ZF minScore 45		ZF minScore 50		ZF minScore 55		ZF minScore 60	
LTR retroelement cutoff	ZF count type	R squared	P value	R squared	P value	R squared	P value	R squared	P value	R squared	P value
minScore 80	ZF count	0.493	6.4E-05	0.532	2.4E-05	0.597	3.7E-06	0.661	4.5E-07	0.682	2.0E-07
minScore 150	ZF count	0.431	2.7E-04	0.470	1.1E-04	0.537	2.1E-05	0.607	2.7E-06	0.645	7.8E-07
minScore 200	ZF count	0.401	5.2E-04	0.437	2.4E-04	0.503	5.0E-05	0.574	7.5E-06	0.618	2.0E-06

		ORFs with 4 or more ZF domains									
		ZF4 minScore 40		ZF4 minScore 45		ZF4 minScore 50		ZF4 minScore 55		ZF4 minScore 60	
	ZF count type	R squared	P value	R squared	P value	R squared	P value	R squared	P value	R squared	P value
LTR diversity (tree length)	ORF count	0.184	2.9E-02	0.199	2.2E-02	0.244	1.0E-02	0.327	2.3E-03	0.559	1.1E-05
	ZF count	0.376	8.6E-04	0.416	3.7E-04	0.494	6.2E-05	0.607	2.8E-06	0.750	1.1E-08

Other correlations

		R squared	P value
LTR minScore 80	LINE-like	0.183	0.05
LTR minScore 80	mammal Olf genes	0.015	0.67
LTR minScore 80	mammal Olf pseudogenes	0.003	0.84
LTR minScore 80	IG C1 domain count	0.089	0.14
LTR minScore 80	IG V domain count	0.027	0.42
LTR minScore 80	genome assembly size	0.048	2.9E-01
ZF count minScore 55	LINE-like	0.063	0.26
ZF count minScore 55	genome assembly size	0.385	1.1E-01
ZF count minScore 55	LTR minScore 80 normalized to genome size	0.581	6.0E-06

Published genomes only, total annotated LTR content:

species	ZF domains minScore 55	RepMask annotated length (LTR + internal sequence)	R squared	P value
opossum	12807	345438106	0.3603	0.0180
mouse	6380	289890855		
orangutan	4895	277238083		
human	5340	359398096		
macaque	4670	221455701		
chimpanzee	5045	247609569		
rat	4139	236203878		
cow	3869	100311384		
dog	3214	101481791		
platypus	582	5796037		
zebrafish	7988	76472359		
medaka	1804	264009		
fugu	237	2750721		
finch	4437	42382077		
frog	2604	5812504		
chicken	543	14450218		

All R squared and P values were computed using Independent Contrasts with the correlation constrained to passing through the origin and a two-tailed P-value.

Data:

species	database version	ACGT genome assembly size	RepMask LINE length	RepMask LINE count	RepMask LTR length	RepMask LTR count	LTR 80-150	LTR 150-200	LTR 200-250	LTR 250up	LTR 80up	LTR 150up	LTR 200up	LTR branch length (blast)	LTR branch length (ML)
Monodelphis domestica (opossum)	monDom5	3501643220	1056459444	2574678	345438106	649672	1192	1280	1261	3034	6767	5575	4295	1121.7	ND
Cavia porcellus (guinea pig)	cavPor3	2663352932	392836933	891142	120922391	448360	717	453	279	1612	3061	2344	1891	293.5	ND
Mus musculus (mouse)	mm9	2617627118	538302066	965731	289890855	851246	541	458	421	3494	4914	4373	3915	578.8	ND
Pongo pygmaeus (orangutan)	ponAbe2	3091708170	687049053	1607106	277238083	762392	498	592	412	625	2127	1629	1037	260.6	ND
Homo sapiens (human)	hg18	2844730767	850110775	2041771	359398096	972481	456	876	525	1040	2897	2441	1565	316.3	ND
Loxodonta africana (elephant)	loxAfr3	3118548474	903361630	1869808	209661419	684373	1746	998	268	219	3231	1485	487	522.9	ND
Macaca mulatta (macaque)	rheMac2	2870775727	553795251	1295793	221455701	597632	362	422	376	742	1902	1540	1118	268.8	ND
Pan troglodytes (chimpanzee)	panTro2	2886534993	614641339	1435834	247609569	660972	490	554	376	731	2151	1661	1107	269.4	ND
Papio hamadryas (baboon)	papHam1	2761826660	NA	NA	NA	NA	424	286	157	185	1052	628	342	193.6	153.9
Callithrix jacchus (marmoset)	calJac3	2929124957	598784758	1408441	207435719	625191	425	307	218	469	1419	994	687	245.1	189.9
Equus caballus (horse)	equCab2	2454407628	543179729	1272532	144955862	446560	325	178	135	140	778	453	275	135.9	109.3
Rattus norvegicus (rat)	rn4	2477053777	583884382	899092	236203878	741614	354	266	230	1051	1901	1547	1281	229.3	ND
Bos taurus (cow)	bosTau4	2731796898	626246310	1655587	100311384	444065	367	310	266	651	1594	1227	917	164.9	130.3
Canis familiaris (dog)	canFam2	2384979816	482138327	1261122	101481791	339360	91	40	31	72	234	143	103	45.5	38.8
Oryctolagus cuniculus (rabbit)	oryCun2	2604006039	437510567	1099885	107548734	398493	133	106	83	392	714	581	475	100.3	83.2
Ornithorynchus anatinus (platypus)	ornAna1	1842235522	398091329	2089269	5796037	16004	26	29	26	40	121	95	66	20.0	15.7
Danio rerio (zebrafish)	zv8/danRer6	1506876822	40076235	123842	76472359	156251	1047	959	1468	1185	4659	3612	2653	175.1	ND
Pteromyzon marinus (lamprey)	petMar1	831696438	NA	NA	NA	NA	522	548	306	52	1428	906	358	177.0	ND
Oryzias latipes (medaka)	oryLat2	700386597	404989	1061	264009	607	274	363	320	222	1179	905	542	123.6	138.2
Gasterosteus aculeatus (stickleback)	gasAcu1	446612119	3446483	7988	1536218	1999	129	125	132	223	609	480	355	43.0	55.8
Takifugu rubripes (fugu)	JGI4.0/fr2	351195798	10118902	26523	2750721	6822	88	114	113	54	369	281	167	42.8	42.7
Tetraodon nigroviridus (tetraodon)	tetNig1	312394419	NA	NA	NA	NA	28	30	21	47	126	98	68	16.7	16.6
Anolis carolinensis (lizard)	anoCar1	1741478929	NA	NA	NA	NA	548	1046	1824	4260	7678	7130	6084	611.9	ND
Taeniopygia guttata (finch)	taeGut1	1222848000	39456769	144164	42382077	93014	107	123	122	457	809	702	579	63.4	59.1
Xenopus tropicalis (frog)	JGI4	1356860127	43407992	165576	5812504	10905	587	772	383	322	2064	1477	705	116.5	110.5
Gallus gallus (chicken)	galGal3	1042566360	65599880	202700	14450218	36318	197	128	108	131	564	367	239	54.5	49.1

species	LINE 60-100	LINE 100-150	LINE 150-200	LINE 200up	LINE all	IG C1 domain	IG V domain	OR functional (Hayden)	OR pseudo (Hayden)
Monodelphis domestica (opossum)	18840	32188	2689	440	54157	106	506	1273	110
Cavia porcellus (guinea pig)	10426	11669	768	44	22907	111	601	561	1120
Mus musculus (mouse)	9910	22418	844	179	33351	125	534	1190	74
Pongo pygmaeus (orangutan)	12772	14430	780	329	28311	84	340	511	420
Homo sapiens (human)	12269	15382	618	105	28374	124	388	581	252
Loxodonta africana (elephant)	66718	57135	4798	846	129497	176	405	1988	2376
Macaca mulatta (macaque)	10198	10354	527	70	21149	97	338	464	269
Pan troglodytes (chimpanzee)	12692	13868	794	289	27643	96	300	550	315
Papio hamadryas (baboon)	10733	2875	74	3	13685	115	288	NA	NA
Callithrix jacchus (marmoset)	12842	11956	753	184	25735	150	306	NA	NA
Equus caballus (horse)	9406	8619	604	48	18677	163	399	1459	852
Rattus norvegicus (rat)	12542	26676	2068	1706	42992	106	674	1420	293
Bos taurus (cow)	52059	71348	8276	3277	134960	119	540	1583	798
Canis familiaris (dog)	6991	7151	886	74	15102	102	397	883	161
Oryctolagus cuniculus (rabbit)	7760	9184	995	43	17982	91	435	604	373
Ornithorynchus anatinus (platypus)	7097	2271	87	26	9481	75	307	433	313
Danio rerio (zebrafish)	2470	1824	58	26	4378	112	923	NA	NA
Pteromyzon marinus (lamprey)	3411	1752	109	25	5297	11	145	NA	NA
Oryzias latipes (medaka)	3749	1431	197	52	5429	131	611	NA	NA
Gasterosteus aculeatus (stickleback)	946	452	32	17	1447	104	739	NA	NA
Takifugu rubripes (fugu)	1084	617	49	15	1765	64	442	NA	NA
Tetraodon nigroviridus (tetraodon)	393	154	12	1	560	49	315	NA	NA
Anolis carolinensis (lizard)	14075	13248	950	312	28585	78	284	NA	NA
Taeniopygia guttata (finch)	909	128	6	0	1043	39	231	NA	NA
Xenopus tropicalis (frog)	5867	3773	323	139	10102	239	604	NA	NA
Gallus gallus (chicken)	4342	1710	157	32	6241	37	298	NA	NA

	minScore 40		minScore 45		minScore 50		minScore 55		minScore 60		total genomic ZF domains					
species	ORFs	total fingers	ORFs	total fingers	ORFs	total fingers	ORFs	total fingers	ORFs	total fingers	minScore 40	minScore 45	minScore 50	minScore 55	minScore 60	genomic KRAB domains
Monodelphis domestica (opossum)	1127	13629	1082	13199	1020	12563	967	11643	904	10211	15828	15044	14032	12807	11143	934
Cavia porcellus (guinea pig)	1360	12705	1231	11350	1054	9391	852	6854	449	2916	15861	13667	11006	8156	4202	807
Mus musculus (mouse)	677	7626	634	7135	599	6555	555	5715	480	4313	9204	8335	7460	6380	4823	659
Pongo pygmaeus (orangutan)	663	6391	630	6036	590	5502	536	4780	444	3671	7091	6481	5745	4895	3835	576
Homo sapiens (human)	589	6199	563	5882	521	5382	480	4706	411	3700	7622	6990	6210	5340	4227	536
Loxodonta africana (elephant)	590	6130	561	5785	531	5349	488	4702	409	3607	7487	6866	6145	5278	4077	510
Macaca mulatta (macaque)	519	5451	495	5165	467	4742	425	4144	363	3276	6642	6096	5394	4670	3719	420
Pan troglodytes (chimpanzee)	596	5998	567	5679	528	5207	478	4530	412	3586	7435	6738	5962	5045	4012	535
Papio hamadryas (baboon)	572	5441	539	5125	496	4642	447	4043	377	3145	6849	6243	5492	4721	3735	480
Callithrix jacchus (marmoset)	524	5403	495	5108	458	4689	421	4133	362	3254	7009	6389	5647	4891	3883	482
Equus caballus (horse)	486	5091	465	4823	428	4385	388	3770	336	2922	5988	5524	4918	4202	3270	445
Rattus norvegicus (rat)	498	5222	467	4867	429	4354	388	3658	328	2594	6377	5800	5024	4139	2988	472
Bos taurus (cow)	476	4595	443	4259	403	3812	357	3275	285	2457	5785	5224	4540	3869	2986	375
Canis familiaris (dog)	369	3826	352	3621	330	3328	304	2932	262	2342	4562	4154	3694	3214	2595	306
Oryctolagus cuniculus (rabbit)	460	4413	436	4185	389	3749	344	3307	301	2716	5686	5173	4557	3948	3226	394
Ornithorynchus anatinus (platypus)	95	775	81	681	74	590	62	458	41	304	1210	996	775	582	405	58
Danio rerio (zebrafish)	1554	13209	1459	11448	1278	9364	960	6561	527	3145	14677	12812	10629	7988	4512	0
Pteromyzon marinus (lamprey)	716	5349	674	4995	610	4497	511	3633	329	2033	6357	5945	5393	4493	2876	0
Oryzias latipes (medaka)	479	3622	429	3072	366	2344	216	1158	19	95	4442	3758	2876	1804	563	0
Gasterosteus aculeatus (stickleback)	118	773	83	512	48	285	13	55	0	0	1238	886	523	225	60	0
Takifugu rubripes (fugu)	86	551	60	364	38	224	14	72	0	0	1112	781	463	237	69	0
Tetraodon nigroviridus (tetraodon)	72	453	53	333	33	198	22	122	10	41	975	690	405	232	95	0
Anolis carolinensis (lizard)	1174	14527	1143	13959	1103	13146	1034	11551	875	8186	16037	15275	14213	12501	9093	340
Taeniopygia guttata (finch)	736	4690	725	4586	705	4385	655	3806	442	2234	5505	5276	4969	4437	3287	5
Xenopus tropicalis (frog)	513	4738	476	4105	430	3283	323	2015	88	472	5515	4783	3836	2604	1144	31
Gallus gallus (chicken)	89	674	82	616	78	563	62	398	25	140	1142	934	747	543	300	72

Table 2.3: Control correlations

		R squared [1]	P value [1]
LINE-like [2]	ZF4 count minScore 55 [3]	0.063	0.26
LTR minScore 80 [4]	mammal Olf genes [5]	0.015	0.67
LTR minScore 80	mammal Olf pseudogenes [5]	0.003	0.84
LTR minScore 80	IG C1 domain count [6]	0.089	0.14
LTR minScore 80	IG V domain count [6]	0.027	0.42
ZF4 count minScore 55	LTR minScore 80 normalized to genome size [7]	0.581	$6.0*10^{-06}$
LTR minScore 80	genome assembly size [8]	0.048	0.29
ZF4 count minScore 55	genome assembly size	0.385	0.11
LTR minScore 80	LINE-like	0.183	0.05

[1] R-squared and P values were computed on the transformation by the method of Independent Contrasts.

[2] LINE-like elements counted for each genome (Supplemental Methods).

[3] ZF4 counts for ORFs with 4 or more ZF domains as described for Table 1.

[4] LTR retroelements counted for each genome with a minimum score of 80 (Supplemental Methods).

[5] Mammalian olfactory gene and pseudogene counts were taken from Hayden et al. 2010. Data were available for all the mammals except baboon and marmoset.

[6] IG C1 (immunoglobulin constant domain type 1) and IG V (immunoglobulin variable domain) domain counts were made from genomic searches with profiles PF07654 (C1-set) and PF07686 (V-set) with minimum rpsblast scores of 40 and 35 respectively. Score cutoffs were chosen to reflect the approximate number of each domain in the human genome.

[7] LTR retroelement counts were divided by the genome assembly size before computing the correlation statistics.

[8] Genome assembly size was computed by counting the number of A, C, G, and T residues directly from the genome assemblies used for all analyses.

Table 2.4: Correlations between recent LTR retroelements and recent ZF duplicates

90% LTR identity, 0.1 dS ZF ORF				
	LTR num	LTR diversity	ZF ORF count	ZF domain count
acar	4996	52.3	301	3456
btau	420	29.8	70	741
cfam	40	4.8	13	106
cjac	184	16.6	20	171
cpor	1384	32.4	501	4457
drer	1741	73.7	823	7948
ecab	68	6.6	17	119
gacu	254	17	13	91
ggal	115	8.2	57	421
hsap	887	52.5	44	468
lafr	58	10.7	33	297
mdom	1228	57.4	180	1854
mmul	366	30.4	21	213
mmus	2958	52.2	184	2026
oana	23	2.8	1	6
ocun	245	16.9	55	521
olat	517	48.8	96	668
pmar	211	19.5	364	2612
rnor	665	22	73	791
tgut	206	14.6	416	2629
tnig	52	6	4	24
trub	89	10.6	8	47
xtro	600	38.3	87	753

90% LTR, 0.1 dS				
	LTR num vs ZF ORF	LTR num vs ZF domain	LTR div vs ZF ORF	LTR div vs ZF domain
p	0.04740	0.00982	0.02649	0.00683
R squared	0.17438	0.27741	0.21337	0.29990

95% LTR identity, 0.05 dS ZF ORF				
	LTR num	LTR diversity	ZF ORF count	ZF domain count
acar	4642	48.1	181	1975
btau	192	10.7	65	661
cfam	19	2.5	12	101

cjac	73	7.1	12	114
cpor	1176	21.2	359	3109
drer	1520	57.8	740	7186
ecab	31	3.3	9	60
gacu	237	15.3	10	74
ggal	80	5.9	54	401
hsap	391	15.2	26	275
lafr	10	1.8	23	205
mdom	865	31.4	107	1048
mmul	170	11.6	8	60
mmus	2562	26.6	133	1446
oana	8	1.1	1	6
ocun	172	11.2	35	294
olat	394	37.4	65	404
pmar	78	8.8	265	1818
rnor	476	12.7	44	476
tgut	141	11	223	1407
tnig	38	4.3	4	24
trub	58	7.7	5	29
xtro	483	31.5	58	493

95% LTR, 0.5 dS				
	LTR num vs ZF ORF	LTR num vs ZF domain	LTR div vs ZF ORF	LTR div vs ZF domain
p	0.04821	0.00434	0.00171	0.00033
R squared	0.17324	0.24719	0.38063	0.46651

P and R squared values were computed using Independent Contrasts.

Table 2.5: Summary of duplicate pairs with an indeterminate ancestral gene.

phylogenetic depth [1]	duplication depth [2]	human gene 1	human gene 2	informative fingers [3]	nt contact changes [4]	ntc adjacent changes [5]	other changes [6]	P-val nt contact vs. other [7]	fingers indel [8]	fingers defective [9]	P-val branch-specific pos selection [10]
Primate specific	> cjac	*ZNF273*	*ZNF680*	10	14	15	29	0.001	0	2	0.0005
Primate specific	> cjac	*ZNF100*	*ZNF430*	10	7	1	11	0.01	0	0	1
Primate specific	> cjac	*ZNF836*	*ZNF841*	15	15	16	29	0.001	2	2	< 0.0001
Eutheria	deep	*ZNF570*	*ZNF583*	11	4	19	17	0.75	1	0	0.5398
Eutheria	deep	*ZNF383*	*ZNF829*	8	3	8	7	0.2	1	1	0.9287
Eutheria	deep	*ZFP30*	*ZFP82*	13	4	10	27	NA	0	0	0.024
Eutheria	deep	*ZNF264*	*ZNF805*	13	5	4	10	0.07	0	0	0.0049
Eutheria	deep	*ZNF226*	*ZNF234*	16	12	13	10	< 0.0001	0	1	< 0.0001
Eutheria	deep	*ZFP112*	*ZNF45*	13	28	20	45	< 0.0001	0	0	0.005
Eutheria	deep	*ZNF568*	*ZNF569*	15	14	16	26	0.001	2	0	1
Boreoeutheria	deep	*ZNF354A*	*ZNF354B*	13	2	5	1	0.07	0	0	0.116
Eutheria	deep	*ZNF619*	*ZNF621*	7	2	8	7	0.43	3	0	0.8065
Primate specific	> cjac	*ZNF564*	*ZNF136*	13	25	41	59	< 0.0001	0	1	0.2415
Primate specific	> cjac	*ZNF124*	*ZNF670*	6	14	14	27	0.0001	0	0	0.0144
Eutheria	deep	*ZNF382*	*ZNF567*	9	15	14	35	0.002	3	2	0.0022
Eutheria	deep	*ZNF41*	*ZNF484*	13	12	13	31	0.02	2	0	0.0821
Eutheria	deep	*ZNF81*	*ZNF175*	12	20	23	28	< 0.0001	1	0	0.0831
Primate specific	> mmul	*ZNF675*	*ZNF681*	10	17	5	20	< 0.0001	0	1	0.0019
Primate specific	> mmul?	*ZNF528*	*LLNL759*	15	37	30	70	< 0.0001	0	0	< 0.0001
			total sites:	222	666	1110	3552				
			total changes:		250	275	489	< 0.0001	15	10	
			change frequency:		0.375	0.248	0.138				

[1] the oldest branch identified with an ortholog for either gene of the duplicate pair.

[2] the phylogenetic branch on which the duplication occurred. > cjac = before marmoset, > mmul = before macaque, deep = before Boreoeutherian split).

[3] the number of ZF domains shared between the duplicate copies.

[4] changes that occurred at one of the three major nucleotide contact sites. For the three "changes" columns, changes in fingers are defined as amino acid residues that are invariant among all orthologous copies of a gene and different between the two duplicates.

[5] changes that occurred at a site immediately adjacent to a major nucleotide contact site (there are five such sites because one is an invariant zinc-coordinating H residue).

[6] changes that occurred at one of the remaining 16 sites (16 = 28 - 3 - 5 - 4 invariant zinc-coordinating residues).

[7] result of a one-sided Fisher's exact test for whether the changed nt contact sites are more frequent than changed other sites, not corrected for multiple testing (NA = not applicable, because they are less frequent).

[8] the number of ZF domains involved in 28 amino acid indel changes between the duplicates (some indels involve more than one adjacent ZF domain).

[9] the number of ZF domains in which one duplicate copy has lost one or more zinc-coordinating residue.

[10] the P-value for positive selection from the branch-site model of codeml, with the branch joining the duplicates labeled (see Materials and Methods), not corrected for multiple testing.

Table 2.6: Summary of duplicate pairs with an inferred ancestral gene.

phylogenetic depth [1]	duplication depth	human gene 1 [2]	human gene 2 [2]	informative fingers	nt contact changes	ntc adjacent changes	other changes	P-val nt contact vs. other	asymmetry score [3]	fingers indel	fingers lost	P-val branch-specific pos selection
Primate specific	> ppyg	*ZNF431*	*ZNF714*	12	13	5	15	< 0.0001	0.91	0	0	< 0.0001
Primate specific	> ppyg	*ZNF679*	*ZNF716*	7	3	7	7	0.19	0.41	3	0	0.0377
Primate specific	> mmul	*ZNF160*	*ZNF665*	18	15	12	32	0.002	1	2	0	0.018
Primate specific	> ppyg	*ZNF468*	*ZNF28*	10	3	4	8	0.24	0.93	7	0	0.0059
Primate specific	> ppyg	*ZNF611* [4]	*ZNF600*	14	8	7	5	0.0001	0.47	3	3	< 0.0001
Primate specific	> ppyg	*ZNF799* [4]	*ZNF443*	13	3	5	7	0.2	0.87	1	0	0.5463
Eutheria	> mmul	*ZNF773*	*ZNF419*	9	5	5	1	0.0004	1	2	0	0.0104
Eutheria	> mmul	*ZNF33B* [4]	*ZNF33A*	16	7	4	4	0.0003	1	0	0	0.0018
Eutheria	> cjac	*ZNF585A*	*ZNF585B*	21	5	2	2	0.001	0.56	0	0	0.4948
Primate specific	> ppyg	*ZNF736* [4]	*ZNF727*	7	10	10	17	0.002	0.89	2	0	0.0013
Eutheria	> mmul	*ZNF558*	*ZNF557*	9	15	9	29	0.05	0.93	1	0	0.0296
Theria	> cjac	*ZNF764*	*ZNF747*	4	3	3	3	0.05	0.44	3	0	0.0024
Eutheria	> mmul?	*ZNF133*	*ZNF343*	10	23	17	41	< 0.0001	0.95	2	1	0.4348
Primate specific	> mmul?	*ZNF17*	*ZNF749*	9	22	24	38	< 0.0001	0.76	1	3	< 0.0001
Euarchontoglires	> cjac	*ZIK1*	*ZNF416*	9	20	22	41	< 0.0001	0.98	0	2	< 0.0001
			totals:	168	155	136	250			27	9	
			sites:		504	840	2688					
			frequencies:		0.308	0.162	0.093					

[1] Most column headers are as defined in Table 3. Euarchontoglires is the clade that includes rodents, lagomorphs, and primates.

[2] The inferred ancestral gene state is listed in the left column and the divergent duplicate gene in the right column.

[3] The asymmetry score can vary from 0 to 1 and is a measure of the extent to which amino acid changes occurred in the divergent duplicate relative to the conserved duplicate (see Materials and Methods). A score of 1 means that changes occurred exclusively in the divergent duplicate, and a score of 0 means that changes were equally distributed between the duplicates. Only changed sites in which all the orthologs had the same amino acid were counted (conserved changes).

[4] These duplicate gene pairs are also described in Nowick et al. 2010.

Table 2.7: Correlations of genomic LTR retroelements and ZF domains

genome	comparison basis	bin size [1]	R-squared	P-value (ANOVA)
acar	our data [2]	100,000	3.07E-04	1.87E-02
acar	our data	1,000,000	4.90E-04	3.39E-01
btau	our data	100,000	2.67E-03	2.20E-16
cfam	our data	100,000	7.61E-04	1.13E-05
cjac	our data	100,000	1.05E-03	5.12E-08
cpor	our data	100,000	3.90E-03	2.20E-16
drer	our data	100,000	3.52E-04	2.27E-02
drer	our data	1,000,000	1.68E-03	1.32E-01
ecab	our data	100,000	5.22E-03	2.20E-16
gacu	our data	100,000	1.42E-03	1.03E-02
ggal	our data	100,000	2.16E-03	1.07E-06
hsap	RepeatMasker [3]	100,000	3.20E-03	2.20E-16
hsap	RepeatMasker	1,000,000	6.50E-03	2.20E-16
hsap	our data	100,000	4.15E-03	2.20E-16
hsap	our data	1,000,000	3.64E-02	2.20E-16
lafr	our data	100,000	2.41E-03	2.20E-16
mdom	our data	100,000	1.73E-04	1.39E-02
mdom	our data	1,000,000	1.39E-03	2.73E-02
mmul	our data	100,000	5.53E-03	2.20E-16
mmus	RepeatMasker	100,000	1.40E-02	2.20E-16
mmus	RepeatMasker	1,000,000	5.70E-02	2.20E-16
mmus	our data	100,000	4.40E-03	2.20E-16
mmus	our data	1,000,000	3.04E-02	2.20E-16
oana	our data	100,000	1.74E-04	1.28E-01
ocun	our data	100,000	2.00E-03	1.24E-13
olat	our data	100,000	2.58E-04	1.43E-01
pham	our data	10,000	2.87E-06	3.94E-01
pmar	our data	10,000	8.84E-07	7.45E-01
ppyg	our data	100,000	2.75E-03	2.20E-16
ptro	our data	100,000	3.76E-03	2.20E-16
rnor	our data	100,000	3.93E-03	2.20E-16
tgut	our data	100,000	2.58E-02	2.20E-16
tnig	our data	100,000	1.24E-04	5.04E-01
trub	our data	100,000	2.36E-04	3.31E-01

xtro	our data	100,000	5.94E-06	7.71E-01
		mean (100 Kb bin, our data):	2.99E-03	
		eutherian mammal mean (100 Kb bin, our data):	3.27E-03	

1 Contigs less than half the length of the bin size were excluded from the counts to reduce over-representing zero match classes in short contigs. Contigs for pham (baboon) and pmar (lamprey) are small, so a smaller bin size was used.

2 Counts in each genomic bin are from data summarized in Table S1. Specifically, the LTR retroelement counts are from all elements (minScore 80), and the ZF counts are the numbers of ZF domains in ORFs with four or more fingers (minScore 55). These are the same data shown in main text Figure 1.

3 LTR retroelement counts in each genomic bin are the number of individual LTR class segments reported in the current RepeatMasker table from the UCSC genome site. The ZF domain counts are from total genomic ZF domains (minScore 55).

Table2.8: Summary of codeml branch-site results

human gene 1	human gene 2	foreground omega (model A)	number total sites	number selected sites	delta ML	P-val pos selection
ZNF836	ZNF841	23.1	890	12	26.58	2.53E-07
ZIK1	ZNF416	107.1	583	39	22.56	2.04E-06
ZNF226	ZNF234	12.4	729	11	17.67	2.63E-05
ZNF17	ZNF749	9.8	743	7	14.87	0.0001
ZNF528	LLNL759	18.1	545	15	14.06	0.0002
ZNF600	ZNF611	220.9	728	20	12.11	0.0005
ZNF431	ZNF714	11.5	508	4	9.17	0.0025
ZNF273	ZNF680	3.6	460	5	6.02	0.0005
ZNF736	ZNF727	7.3	434	5	5.14	0.0013
ZNF33B	ZNF33A	119.0	730	19	4.90	0.0018
ZNF675	ZNF681	4.9	573	0	4.80	0.0019
ZNF382	ZNF567	2.2	575	23	4.68	0.0022
ZNF764	ZNF747	442.8	305	13	4.62	0.0024
ZNF264	ZNF805	122.3	546	21	3.96	0.0049
ZFP112	ZNF45	2.0	420	36	3.94	0.0050
ZNF468	ZNF28	18.4	681	0	3.78	0.0059
ZNF773	ZNF419	184.7	447	11	3.28	0.0104
ZNF124	ZNF670	2.5	335	0	2.99	0.0144
ZNF160	ZNF665	2.5	731	0	2.80	0.0180
ZFP30	ZFP82	2.4	457	10	2.55	0.0240
ZNF558	ZNF557	2.5	288	11	2.37	0.0296
ZNF679	ZNF716	2.8	415	0	2.16	0.0377
ZNF41	ZNF484	2.4	799	NA	1.51	NS
ZNF81	ZNF175	1.6	614	NA	1.50	NS
ZNF354A	ZNF354B	3.1	528	NA	1.24	NS
ZNF136	ZNF564	1.3	490	NA	0.69	NS
ZNF133	ZNF343	1.2	604	NA	0.31	NS
ZNF585A	ZNF585B	2.1	672	NA	0.23	NS
ZNF570	ZNF583	1.3	493	NA	0.19	NS
ZNF799	ZNF443	1.8	608	NA	0.18	NS
ZNF619	ZNF621	1.1	482	NA	0.03	NS
ZNF383	ZNF829	1.1	399	NA	0.00	NS
ZNF100	ZNF430	1.0	463	NA	0.00	NS
ZNF568	ZNF569	1.0	608	NA	0.00	NS

For each duplicate gene pair, key output values from the branch-site model of codeml are listed, using selection model A and control model A1 (see Materials and Methods and PAML 3.15 documentation for details). Foreground omega is the additional foreground d_N/d_S class assigned by codeml run with model A. Number sites - the total number of sites in the alignment. Number selected sites - the number of sites with evidence for positive selection with P >= 0.95 (Bayes-Empirical-Bayes output). Delta ML - the difference in log-likelihoods for model A and model A1. P-val pos selection – the P-value for positive selection, computed by a chi-square test from twice delta ML with one degree of freedom. Values are not corrected for multiple testing. In cases where the P-value was not significant (P > 0.05 or NS, not significant), the number of selected sites is labeled NA (not applicable).

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18: 1585-1592.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55: 539-552.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. Ann. Rev. Genomics Hum. Genet. 7: 149-173.

Basta HA, Buzak AJ, McClure MA. 2007. Identification of novel retroid agents in Danio rerio, Oryzias latipes, Gasterosteus aculeatus and Tetraodon nigroviridis. Evol. Bioinform. Online 3: 179-195.

Basta HA, Cleveland SB, Clinton RA, Dimitrov AG, McClure MA. 2009. Evolution of teleost fish retroviruses: characterization of new retroviruses with cellular genes. J. Virol. 83: 10152-10162.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836-840.

Bellefroid EJ, Poncelet DA, Lecocq PJ, Revelant O, Martial JA. 1991. The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. Proc. Natl. Acad. Sci. USA 88: 3608-3612.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. USA 101: 4894-4899.

Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol Biol Evol. 22: 814-817.

Benveniste RE, Todaro GJ. 1974. Evolution of C-type viral genes: inheritance of exogenously acquired viral genes. Nature 252: 456-459.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. USA 104:17004-17009.

Best S, Le Tissier P, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. Nature 382:826-829.

Birtle Z, Ponting CP. 2006. Meisetz and the birth of the KRAB motif. Bioinformatics 22: 2841-2945.

Blikstad V, Benachenhou F, Sperber GO, Blomberg J. 2008. Evolution of human endogenous retroviral sequences: a conceptual account. Cell. Mol. Life Sci. 65: 3348-3365.

Chen Z, Telfer P, Gettie A, Reed P, Zhang L, Ho D, Marx PA. 1996. Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. J. Virol. 70: 3617-3627.

Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448: 105-114.

Conley AB, Piriyapongsa J, Jordan IK. 2008. Retroviral promoters in the human genome. Bioinformatics 15: 1563-1567.

Copeland NG, Hutchison KW, Jenkins NA. 1983. Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. Cell 33: 379-387.

de Parseval N, Heidmann T. 2005. Human endogenous retroviruses: from infectious elements to human genes. Cytogenet. Genome Res. 110: 318-332.

Doolittle RF, Feng DF. 1992. Tracing the origin of retroviruses. Curr Top Microbiol Immunol. 176: 195-211.

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T. 2009. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc. Natl. Acad. Sci. USA 106: 12127-12132.

Edelstein LC, Collins T. 2005. The SCAN domain family of zinc finger transcription factors. Gene 359: 1-17.

Emerson RO, Thomas JH. 2009. Adaptive evolution in zinc finger transcription factors. PLoS Genet. 5: e1000325.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125: 1-15.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397: 436-441.

García-García MJ, Shibata M, Anderson KV. 2008. Chato, a KRAB zinc-finger protein, regulates convergent extension in the mouse embryo. Development 135: 3053-3062.

Garland T Jr, Bennett AF, Rezende EL. 2005. Phylogenetic approaches in comparative physiology. J. Exp. Biol. 208: 3015-3035.

Gogvadze E, Buzdin A. 2009. Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. 66: 3727-3742.

Goodwin TJ, Poulter RT. 2002. A group of deuterostome Ty3/ gypsy-like retrotransposons with Ty1/ copia-like pol-domain orders. Mol. Genet. Genomics 267: 481-491.

Groner AC, Meylan S, Ciuffi A, Zangger N, Ambrosini G, Dénervaud N, Bucher P, Trono D. 2010. KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. PLoS Genet. 6: e1000869.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537: 113-137.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696-704.

Hamilton AT, Huntley S, Kim J, Branscomb E, Stubbs L. 2003. Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. Cold Spring Harb. Symp. Quant. Biol. 68: 131-140.

Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. Jan;20(1):1-9.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972.

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 16: 669-677.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature. 2014 Sep 28. doi: 10.1038/nature13760.

Jern P, Sperber GO, Blomberg J. 2006. Divergent patterns of recent retroviral integrations in the human and chimpanzee genomes: probable transmissions between other primates and chimpanzees. J Virol. 80: 1367-1375.

Johnson WE, Coffin JM. 1999. Constructing primate phylogenies from ancient retrovirus sequences. Proc. Natl. Acad. Sci. USA 96: 10254-10260.

Kim CA, Berg JM. 1996. A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. Nat. Struct. Biol. 3: 940-945.

Krebs CJ, Larkins LK, Khan SM, Robins DM. 2005. Expansion and diversification of KRAB zinc-finger genes within a cluster including Regulator of sex-limitation 1 and 2. Genomics 85: 752-761.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392: 917-920.

Laten HM, Majumdar A, Gaucher EA. 1998. SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci U S A 95: 6897-6902.

Lechner MS, Begg GE, Speicher DW, Rauscher FJ 3rd. 2000. Molecular determinants for targeting heterochromatin protein 1-mediated gene silencing: direct chromoshadow domain-KAP-1 corepressor interaction is essential. Mol. Cell. Biol. 20: 6449-6465.

Li X, Ito M, Zhou F, Youngson N, Zuo X, Leder P, Ferguson-Smith AC. 2008. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. Dev. Cell 15:547-557.

Looman C, Abrink M, Mark C, Hellman L. 2002. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. Mol Biol Evol. 19: 2118-2130.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc. Natl. Acad. Sci. USA 104: 8005-8010.

Mackay DJ, Callaway JL, Marks SM, White HE, Acerini CL, et al. (19 authors). 2008. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. Nat. Genet. 40: 949-951.

Maddison WP, Maddison DR. 2010. Mesquite: A modular system for evolutionary analysis. Version 1.1. http://mesquiteproject.org.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10:1307-1318.

Martin J, Herniou E, Cook J, O'Neill RW, Tristem M. 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J. Virol. 73: 2442-2449.

Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC, Shinkai Y. 2010. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464: 927-931.

Midford PE, Garland T Jr, Maddison WP. 2005. PDAP Package of Mesquite. Version 1.07.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (63 authors). 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167-177.

Mouse Genome Sequencing Consortium, et al. (222 authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876-879.

Nielsen AL, Ortiz JA, You J, Oulad-Abdelghani M, Khechumian R, Gansmuller A, Chambon P, Losson R. 1999. Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. EMBO J. 18: 6385-6395.

Nowick K, Gernat T, Almaas E, Stubbs L. 2009. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. Proc. Natl. Acad. Sci. USA 106: 22358-22363.

Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggests selection for novel function in primate-specific KRAB-ZNF genes. Mol. Biol. Evol. 2010 Jun 23. Epub ahead of print.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 5: e1000753.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. Science 327: 835.

Pavletich NP, Pabo CO. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252: 809-817.

Polavarapu N, Bowen NJ, McDonald JF. 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol. 7: R51.

Rhesus Macaque Genome Sequencing and Analysis Consortium, et al. (176 authors). 2008. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.

Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T. 2008. An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res. 18: 597-609.

Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard PV, Layard-Liesching H, Verp S, Marquis J, Spitz F, Constam DB, Trono D. 2010. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463: 237-240.

Rowe HM, Trono D. Dynamic control of endogenous retroviruses during development. Virology. 2011 Jan 18. Epub ahead of print.

Ryan RF, Schultz DC, Ayyanathan K, Singh PB, Friedman JR, Fredericks WJ, Rauscher FJ 3rd. 1999. KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Krüppel-associated box-zinc finger proteins in heterochromatin-mediated gene silencing. Mol. Cell. Biol. 19: 4366-4378.

Schmidt D, Durrett R. 2004. Adaptive evolution drives the diversification of zinc-finger binding domains. Mol. Biol. Evol. 21: 2326-2339.

Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ, 3rd. 2002. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev. 16: 919-932.

Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res. 13: 1097-1110.

Smit, AFA, Hubley, R, Green, P. 2010. RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.

Smith FM, Garfield AS, Ward A. 2006. Regulation of growth and metabolism by imprinted genes. Cytogenet. Genome Res. 113: 279-291.

Sripathy SP, Stevens J, Schultz DC. 2006. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. Mol. Cell. Biol. 26: 8623-8638.

Stocking C, Kozak CA. 2008. Murine endogenous retroviruses. Cell. Mol. Life Sci. 65: 3383-3398.

Subramanian A, Kaufmann M, Morgenstern B. 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol. 3: 6.

Thomas JH, Emerson RO, Shendure J. 2009. Extraordinary molecular evolution in the PRDM9 fertility gene. PLoS One 30: e8505.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

Whitney KD, Garland T Jr. 2010. Did Genetic Drift Drive Increases in Genome Complexity? PLoS Genet. 6: e1001080.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973-982

Wolf D, Goff SP. 2007. TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell 131: 46-57.

Wolf D, Goff SP. 2008. Host restriction factors blocking retroviral replication. Ann Rev. Genet. 42: 143-163.

Wolf D, Goff SP. 2009. Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature. 458:1201-1204.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22: 1107-1118.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22: 2472-2479.

Chapter 3: Accidental genetic engineers: Horizontal sequence transfer from parasitoid wasps to their Lepidopteran hosts

Introduction

The transfer of genetic information between different species, horizontal gene transfer (HGT), is an unexpected and transformational discovery resulting from the expansion of whole genome sequence data. Examples of HGT violate the traditional idea that shared sequences among organisms come from a shared ancestral species. HGT can be a source of novel sequences that organisms would otherwise not have access to (Boto 2014) such as a fungal carotenoid gene that gave pea aphids a bright orange coloration (Moran and Jarvick 2010) or antifreeze genes transferred between very distant fish species (Graham 2008). Not all horizontal transfers are clearly adaptive. There are a number of cases where transposons have been transferred between parasite species and their host species (Gilbert et al. 2010, Thomas et al 2008, Gilbert et al 2014, Kuraku et al. 2012, Walsh et al. 2013).

Nevertheless, HGT is rarely reported among multicellular eukaryotes, presumably due to a combination of the poor efficiency of foreign sequences entering the germline and the difficulties of detecting HGT events. Viruses, however, can enter cells and facilitate HGT. Many fragments of virus genomes have been found in eukaryotes (Katzourakis and Gifford 2010), and these endogenous viral elements provide a rich source of information on the evolution and history of viruses. Here we present horizontally transferred sequences (HTS) from polydnaviruses(PDVs) which were created by a process similar to the process by which endogenous viral elements are formed. A key difference between EVEs and PDVs is that instead of transferring viral sequences, PDVs transfer genetic material from specific portions of the wasp genome.

Polydnaviruses are an unusual group of virus-like particles named after their dsDNA genomes that exist in many circular fragments (Stoltz 1984). PDVs are utilized alongside or instead of venom to aid in parasitoid reproduction by weakening the secondary host's immune

system so that the parasitoid eggs can develop. Unlike typical viruses, PDVs are incapable of self-replication and instead are endosymbiotic viruses of two clades of parasitic wasps, Braconids (bracoviruses) and Ichneumonids (ichnoviruses). These two wasp clades have stably integrated viral sequences into their genomes and inherit them vertically (Fleming and Summers 1986, Fleming and Summers 1991, Stoltz 1990, Belle et al. 2002). Viral particle production occurs exclusively in the ovaries of these parasitoid wasps (Theilman and Summers 1986, Stoltz et al 1976).

Specific portions of the wasp's genome are packaged (Flemming and Summers 1991) into the virus-like particles (see Figure 3.1), but genes encoding viral functions are not included, leaving PDV particles without the molecular machinery to replicate in the secondary host. An important evolutionary consequence of the mechanism of PDV production is that the propagation of PDVs depends on survival of the wasp eggs growing in the secondary host. As a result, suppression of secondary host immunity is crucial to the survival of the wasp/virus mutualism. Genes transferred by PDV particles encode a number of mechanisms to suppress secondary host immunity, such as controlling gene expression with modified histones (Hepat and Kim 2011), interfering with the haemocyte cytoskeleton (Labropoulou et al. 2008), and inhibiting phagocytosis (Pruijssers and Strand 2007). In addition to suppressing the host immune system, PDV genes can also function to shift resources from secondary host larval development to growing wasp larvae (Kwon et al. 2010). A number of cystatins in *Cotesia* bracoviruses are thought to suppress the development and immune response of the secondary host (Espagne 2005). These cystatin genes have been under recent positive selection, consistent with classical host/parasite evolution (Serbielle 2008).

Despite the many mechanisms PDVs employ to attack the host, these parasitoid attacks vary widely in lethality depending on their timing. For a horizontally transferred sequence to be seen in the population, a secondary host must survive a PDV attack and reproduce. In one study, infection by a bracovirus shortly after the start of the fifth instar (the final stage of

development before metamorphosis in the species) resulted in a 100% lethality rate, while infection 60 hours post-ecdysis (the shedding of the old cuticle) had 0% lethality and generally had little effect on the larvae (Pruijssers et al 2009). These experiments illustrate the possibility of secondary host survival with a foreign sequence after PDV infection.

PDVs have now been sequenced from several wasps, (Webb et al. 2006, Desjardins et al. 2008) including many related braconid wasps in the *Cotesia* family (Espagne et al. 2004, Choi et al. 2009, Chen et al. 2011). Analysis of *Cotesia* relatives reveals that many of the genes packaged into PDVs are part of large gene families. These genes are eukaryotic in origin and some of these gene families only appear to be found in wasps (Chen et al. 2008). In addition to positive selection acting on coding sequences, these gene families also appear to be subject to birth-death evolution typical of large gene families under strong or quickly changing selection (Friedman and Hughes 2006).

Bracoviruses and ichnoviruses arose independently from different viruses early in the evolution of Braconid and Ichneumonid wasps (Volkoff et al. 2010). Braconid wasps originated around 70-100 MYA (Thézé et al. 2011, Whitfield 1997, Whitfield 2002) when a nudivirus stably integrated into the germ line of an ancestral braconid wasp (Bézier et al. 2009, Burke and Strand 2012) The origin of ichnoviruses is less clear. Reflecting their different origins, the two groups have different viral gene content and structure. Ichnoviruses package a suite of dsDNA circles in each virion particle, while bracoviruses only include one (of many) circles per virion particle (Albrecht et al. 1994). This means different bracovirus particles from the same wasp have different contents (see Figure 3.1) and that it is possible for a wasp to inject more of one dsDNA circle than another (Beck et al. 2007).

The fact that PDVs routinely transfer wasp DNA to the secondary host has prompted researchers to investigate how long this DNA persists and whether it can permanently integrate into the secondary host (Gundersen-Rindal and Dougherty 2000, Doucet et al. 2007, Beck et al. 2011). Bracovirus DNA circles from the wasp *Cotesia congregata* were shown to persist for at

least 6 days throughout the living body of the secondary host, *Manduca sexta* (Le et al. 2003). Additionally, Lepidopteran cell cultures infected with PDVs show stable integration into chromosomes and expression of both bracoviruses (Gundersen-Rindal and Lynn 2003) and ichnoviruses (Volkoff et al. 2001). Interestingly, both studies found biased integration favoring some PDV segments over others. What has been absent until now is a demonstration that this type of integration can occur outside of laboratory cell cultures, in the germ line of living organisms.

Results

The initial screen of this study was an extensive bioinformatic search for PDV sequences in eukaryotic genomes. The search was performed using tblastx, searching 386 PDV sequences against 165 plant and animal whole genome sequences (see methods). PDV sequences contain an unknown mix of coding sequence and intergenic regions (as do the genome assemblies), tblastx was utilized because it can sensitivity locate matching protein coding regions in a background of non-coding regions.

In our initial screen there were 18,797 hits in total before any filters were applied. Within these, 99.44% of our tblastx matches were found in invertebrates even though the invertebrate genomes were only 65/165 (39.4%) of the genome assemblies queried (and an even smaller fraction of the sequence space searched). In the remaining 0.56% of tblastx matches there were 101 vertebrate matches and 11 plant matches. The invertebrate matches had an average tblastx score of 116.4 (e-values 1E-06 to 1E-200) while the non-invertebrates had an average match score of 58.3 (e-values 3E-07 to 5E-024). As a result of being poor matches, 88.4% of the non-invertebrate matches were removed by the initial length (67 amino acids) and tblastx score (minimum 70) thresholds alone, with the rest being removed by later filters. In contrast, 31.3% of invertebrate hits were filtered out by those thresholds.

After applying a series of quality filters (see methods), our bioinformatic search revealed many candidate PDV-derived horizontally transferred sequences (HTS) in two Lepidopteran

species (see Table 3.1). This is consistent with the natural host range of these wasps, which is heavily enriched for targeting Lepidoptera larvae. These sequence similarities violate the known insect phylogeny and suggest HGT.

To test whether these matches could be explained by unknown insect genes giving a false signal resembling HGT, several analyses were performed. The candidate HTS were used as queries in a search against a combined database with the same 165 eukaryotic genomes, the 387 PDV sequences, and NCBI's non-redundant viral sequence list(see methods). In all cases the candidates matched more closely to PDVs. The same search was performed by web-BLAST against NCBI's non-redundant database and all the candidates matched best to PDVs.

In an additional analysis, each of the remaining HTS candidates were searched against the set of 165 eukaryotic genomes with the goal of finding other sequences related to the HTS. In every case, no new matches were found. This indicates that the only sequences with significant sequence similarity to the HTS candidates in the two Lepidopterans are wasp PDV sequences. Thus, our reported 105 HTS are each only found in two places: the Lepidopteran genome assembly (either *Bombyx mori* or *Danaus plexippus*) and the PDV assembly from a braconid wasp.

This distant distribution is quite unusual. Sequences with a high relatedness between Lepidoptera and Hymenoptera typically come from highly conserved genes with a broad distribution among insects. In an additional analysis, we identified 2,631genes from the wasp *Nasonia vitripennis* with homology to genes in *Bombyx mori*. 300 of these were selected at random to perform a phylogenetic analysis examining how frequently these genes could be found in other insect species. We found that these random *Nasonia* genes are present in 96-99% of insect assemblies (Figure 3.2). This is dramatically different from the candidate HTS, which are only found in a single species (plus the donor wasp). We also measured the degree of nucleotide conservation for each gene between the wasp PDV copy and the copy in the other insect. The tblastx match regions between *Bombyx* and *Nasonia* for these 2,631 genes had

median length 537 nucleotides (range 183-2745) and median 71.3% nucleotide identity (range 58.5-87.3), a much lower degree of conservation than in our HTS candidates (median 86.95%). The 105 HTS candidates were found to have significantly different percent identities than the 2,631 conserved *Nasonia* genes found in *Bombyx* using the Mann-Whitney u-test ($p<2.3E-09$). All 300 genes had a broader phylogenetic distribution than any of the PDV matches. The combination of the candidate HTS only being found in one very distant species and having a higher nucleotide identity than conserved genes is indicative of these candidate sequences being horizontally transferred. None of the HTS found in *Bombyx mori* had any homology to HTS in *Danaus plexippus* and vice versa, indicating separate integration events.

The HTS were sorted into homology groups, with a member being added to a group if it had 90% similarity over 90% of its length to another member of the group (HTS are listed with groups in Table 3.1). Some of these HTS exist as a single copy while others appear to be part of such groups (Figure 3.3). Homology groups containing more than one sequence were aligned by DIALIGN (Morgenstern 2004) and placed onto trees using PHYML (Guindon et al. 2010). Those that appear together in a homology group are the result of an unknown mixture of two factors: repeated integration of the same PDV locus (circle) into the secondary host and subsequent duplication of the sequences in the secondary host. Many repeated integrations of the same sequence are expected from previous cell culture studies (Gundersen-Rindal and Lynn 2003, Volkoff et al. 2001). It is possible that there are mis-assemblies in the genome sequences that affect the apparent copy number of these sequences. This could increase or decrease the copy number by splitting alleles into paralogs or vice versa. For these reasons it is difficult to know exactly how many copies exist in the destination genomes, though it must be at least one copy to show up in the assembly. Because the destination genome assemblies do not have chromosome information there is no information regarding whether the transferred sequences preferentially integrate into certain portions of the destination genome.

Three PDV queries (NC_006658.1, HQ009558.1, EF067323.1) were found to contain significant matches to transposable elements (a gypsy, helitron, and Jockey respectively). Hits matching these sequences were discarded for clarity and focus in this manuscript. However, it is worth noting that if the central hypothesis of this paper is correct (that PDVs will tend to transfer genomic sequences from parasitoid wasps to their host species) it would predict that transposable elements like these would frequently make the transfer from wasp to host along with their surrounding genomic DNA. If a transposable element is targeted for packaging into a PDV, it will be among a small portion of the genome that is heavily enriched, copied in large numbers, and transferred to the host species making PDVs a strong vector for spreading of transposable elements. Once transferred to the host species, these could expand in number and spread throughout the new genome.

To gain insight into the function of open reading frames (ORFs) in candidate HTS, we searched the translated protein sequence of the HTS against the PfamA and PfamB databases (http://pfam.sanger.ac.uk/search) using a highly permissive e-value (1e-03). Two sequences (PDV94 and PDV96) were found to contain the BEN domain, which is known to be found in PDVs and is associated with growth arrest and transcriptional regulation. PDV94 also matches to the protein family Pfam-B_3333 (function unknown). The function of ORFs in other candidate HTS is unknown.

Based on previous work, we expect that PDV loci transferred to secondary hosts are likely subject to recurrent positive selection before being transferred (Desjardins et al. 2008, Bézier et al. 2008). To test for positive selection in coding regions, the HTS were aligned by codon to the matching PDV sequence and then analyzed for synonymous and non-synonymous changes by maximum likelihood using PAML (Yang 2007). The results indicate that many of the sequences have a dn/ds > 1, suggesting that they have undergone recent positive selection (Table 3.1). A dn/ds <1 generally indicates that a sequence is conserved, though it does not rule out the possibility of positive selection on a small number of residues in the protein. It is possible

that the positive selection we observed occurred prior to transfer (Desjardins et al. 2008, Bézier et al. 2008), however it is also possible that some of these sequences were co-opted and selected upon after transfer to the secondary host as well.

To rule out the possibility of an apparent HTS being the result of genome assembly error, the presence of HTS were tested by PCR directly from Bombyx genomic DNA with DNA from other insects used as a control (Figure 3.4). Several representative HTS (PDV32, PDV99, PDV100, and PDV101) from *Bombyx mori* were arbitrarily selected: including both a representative large sequence that appears a single time and a representative smaller sequence that appears many times in the assembly. The HTS successfully amplified from all *Bombyx* strains tested and did not amplify in control species, confirming that these sequences are not artifacts of the *Bombyx* sequencing project.

Discussion

While previous work has shown a strong trend for HGT between parasitic species and host species, (Gilbert et al. 2010, Thomas et al 2008, Gilbert et al 2014, Kuraku et al. 2012, Walsh et al. 2013) the findings presented here are the first known instances of HGT between a *parasitoid* species and host species. Most parasites typically do not kill the host they feed upon, making them a passive vector for horizontal transfer if there happens to be a relevant infection by a vector species (e.g. retrovirus). Unlike parasites, successful parasitoids ultimately kill their host, meaning the secondary host must also defeat the parasitoid infection and survive to reproduce for HGT to occur. Parasitoid/PDV attacks vary widely in lethality. Attacks earlier in development can have a 100% lethality rate while attacks as little as 60 hours later can have 0% lethality and few effects (Pruijssers et al. 2009). Based on this, it seems likely that the ancestral infections discussed in this manuscript occurred in larvae that were at a later stage of development. This window of time in which PDV infections can occur but have little effect on the host provides a natural avenue for HGT.

Those parasitoids that utilize PDVs likely increase the chance of horizontal transfer by many orders of magnitude by enriching a small portion of their genome and injecting it directly into the host in a specialized viral vector. Combined with the variable lethality of PDV infections, this gives a plausible range of natural circumstances in which horizontal transfer can occur. Previous work has shown this potential for PDVs to facilitate HGT in somatic insect cell lines, but never in germ line cells or living organisms in natural populations. Our findings represent the first documented cases of HGT using a PDV vector in a live organism.

Though the mechanisms by which PDVs integrate into live organisms are poorly understood, this HGT raises the possibility of experimentally manipulating the system. These PDVs could be used as gene delivery vectors for the many plant (Dangerfield and Austin 1998, Infante et al. 1995) and animal species these viruses naturally affect, and possibly others. The targeting and efficiency of gene transfer using PDVs could be improved, making PDVs a potentially useful vector that will naturally self-terminate due to a lack of reproductive capability. If PDVs could be produced from wasp ovary cell cultures, this targeted gene transfer could be achieved using artificial injections without rearing wasps or infecting hosts with wasp larvae.

One interesting feature of these PDV producing wasps with respect to HGT is that, unlike many other instances of HGT from one multicellular eukaryote to another, there is no independently reproducing third species required to act as an intermediary. In many other instances of eukaryote HGT, this role is played by an organism (e.g. retrovirus) which is free living and then randomly acquires and carries the sequence from the donor species for a time, often replicating it as a part of its genome, until infecting the recipient species and transferring the sequence to the recipient germ line. HGT facilitated by PDVs is quite different. In this case everything needed to complete the HGT is contained within the wasp/host system. On the donor side, the sequences targeted for packaging in PDVs is non-random: a small subset of the wasp genome is replicated many times and packaged into the PDV vector. On the recipient side, the

host is a species specifically targeted by the wasp (and not an undirected infection as in the case of other vectors).

These differences predict several effects we expect to see when HGT is facilitated by PDVs. One is that there should be an increased number of horizontal transfers going from wasp to host. Another prediction is that the transferred regions should be regions targeted for PDV packaging, rather than random wasp sequences. Because PDVs cannot replicate independently, another prediction is that these transferred sequences would not be expected to be found in species that are not targeted by these wasps. In the case of other vectors, the same virus that infects the recipient's germ line could continue reproducing and infect many other organisms (and species), producing a different, more dispersed pattern of HGT than one expects with PDVs.

This unusual genetic phenomenon creates an evolutionary paradigm in which a host species can acquire genetic information from a parasitoid species that attacks it. Previous work has discussed the acquisition of sequences from viral pathogens (Aswad and Katzourakis 2012, Bertsch et al. 2009, Flegel 2009, Koonin 2010) however discussion of the acquisition of sequences from eukaryotic parasitoids has been absent up to this point. As with mutations or other horizontally transferred sequences, these sequences have the potential to be adaptive, maladaptive, or neutral for the host (probably more often neutral/maladaptive). Maladaptive sequences are typically removed over time by selection, so sequences found in extant organisms are likely depleted for maladaptive sequences. Occasionally, transferred sequences could be co-opted and adapted by the host in a number of ways. The chance of a sequence being co-opted would be expected to increase with the number and diversity of transferred sequences. The ways that the sequences could be adapted to the new host include controlling the host's gene regulation, altering its physiology, or using those sequences as a direct countermeasure against wasp parasitization to fight future PDV infections. For example, the secondary host could use the transferred sequences to produce anti-sense RNA to mark

transcribed PDV mRNAs for degradation by normal processes that affect double stranded RNA. Another possibility is that the secondary host could co-opt protein interacting domains taken from the wasp to bind and block proteins produced by the virus. The secondary host could also find completely novel uses for the sequences, such as regulating its growth or immune system. Since it is estimated that there are hundreds of thousands of PDV producing wasp species (and their corresponding host species, ranging from insects to plants), our findings are likely a first look at a large and unusual phenomenon. This sort of horizontal transfer is likely to be discovered more frequently as more eukaryote species and PDVs are sequenced.

Tracking PDV HTS in secondary hosts can be a useful tool for evolutionary studies. It should be possible to deduce parasitoid/host relationships, potentially even if there is no known parasitoid affecting the secondary host species. This is the case with the data presented here, which predict a parasitoid which affected *Bombyx mori* (or an ancestor species) though none is currently known. Given the moderate sequence divergence observed between the HTS found in *Bombyx/Danaus* and the donor PDVs, this suggests that the transfers were from some ancestors of *Cotesia* wasps to some ancestors of *Bombyx/Danaus* and may not involve any presently extant species. Evolutionary analysis of PDV HTS is applicable not just to present parasitoids but also parasitoids from the past, opening up the exciting possibility of tracking relationships that involve extinct species.

Methods

BLAST searches

The initial screen of this study was an extensive tblastx search of 387 PDV sequences against 165 eukaryote genomes including 75 vertebrate species, 25 plant species, and 65 invertebrate species. Among the invertebrates, there are 48 insect species including 15 Hymenoptera, 5 Lepidoptera, and 24 Diptera. A complete list of PDV queries and eukaryotic whole genome assemblies can be found in Table 3.2 and Table 3.3.

PDV sequences contain an unknown mix of coding sequence, regulatory sequence, introns, intergenic regions, and repeat content. When a PDV is naively searched against whole genome sequences, any of these features could match to the whole genome sequence. As a result, matches could appear in species that are not known to be parasitized by PDV wasps (e.g. mammals). These could be caused by simple repeat sequences or by proteins common to the larger group (e.g. metazoans). Tblastx was used because it is highly sensitive at detecting evolutionarily divergent proteins with different amino acids but conserved function (e.g. substituting a non-polar amino acid for a different non-polar amino acid) using the BLOSUM62 matrix. Sequences from this screen were kept if they had all the following: an e-value below 1.0e-10(significant after multiple correction testing), a BLAST score greater than 70, and a match length of at least 67 amino acids. Remaining hits found on the same contig nearby one another (<5kb) were merged into a single hit.

Using these ORF containing protein matches as initial regions, we expanded our investigation to regions surrounding these open reading frames to get a more complete understanding of how much sequence was transferred. The protein match from the PDV sequence plus 5kb of flanking DNA in both directions were searched against the destination genomes for a strong match in DNA sequence. Portions of these ORF/extension regions that match to the destination genome could indicate a large stretch of DNA transferred by PDV to the destination genome. Sequences in this search were kept if they had the following: a match length of at least 200 nucleotides and a percent identity of at least 80%. The remaining sequences were our initial pool of potentially HTS.

Removing repeats

Sequences matching a PDV sequence found to contain a transposon (NC_006658.1, HQ009558.1, EF067323.1) were removed. While we expect transposons to be transferred from PDVs to destination genomes just as a coding sequence would, they were removed for clarity and focus in this manuscript. Transposons were found by searching the 387 PDV sequences

with RepeatMasker 3.3.0 (default settings). PDV sequences with significant matches to transposons were removed from the study.

Remaining candidate sequences were filtered for simple repeats. The DNA sequence of each candidate was searched by RepeatMasker 3.3.0 for significant matches with the following settings: -noint, –poly, default settings elsewhere. Any candidate with a significant match to simple repeats was removed.

Homology Analysis

The set of candidate sequences in the destination genomes were then analyzed and placed into homology groups. All the candidate sequences were searched against themselves using blastn with a max e-value of 1.0e-10. Candidate sequences were placed in a group together when a sequence has a 90% nucleotide identity over 90% of its length to any member in the group. Homology groups (and their candidates) containing more than one destination genome were removed from the study. Though such sequences may be horizontally transferred, we removed the sequences on the possibility that they could be vertically transmitted. There is no biological mechanism precluding protein sequences shared by metazoans or eukaryotes being included in a PDV. If such a shared sequence is included in a PDV, our screen for protein matches would find them and report a dispersed pattern of hits across many destination genomes. Because of this concern, homology groups showing that pattern were removed from the study.

In addition to searching for homology groups within our candidate transferred regions (above), we also searched for homology between our candidate sequences and the whole genome sequences of a large set of organisms. Specifically, we searched the candidate HTS against a database including our set of 165 eukaryotic genomes, the original 387 PDV sequences, and NCBI's non-redundant virus list (containing "conventional" viruses). In each search the strongest match of all the sequence space searched (whether viral, PDV, or eukaryote) was recorded. A sequence that matches more strongly to any eukaryotic sequence

than to a PDV or virus sequence (self-hits excluded) could indicate that the sequence is an insect gene shared between those species by a distant common ancestor and not actually a PDV HTS. If the sequences in the destination genomes are truly horizontally transferred regions, they should match more strongly (>20 blast score gap) to PDV sequences than to any other sequence. Searches were also performed by web-BLAST against NCBI's non-redundant database.

Testing conserved insect regions

We started with the coding regions of known Bombyx mori mRNAs (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bombyx_mori/RNA/) and used each as a tblastx query to the parasitic wasp *Nasonia vittripennis* genome assembly. *Bombyx* coding genes that had a *Nasonia* match score with the threshold used in the PDV searches (BLAST score 70 or higher) were extracted and a single splice form was retained for each such gene. This produced a set of 2,631 *Bombyx* genes that include a highly conserved segment present in the *Nasonia* genome. We selected 300 random genes from these 2,631 and used the same tblastx test to determine the representation of such highly conserved gene segments across a representative set of 26 insect genome assemblies. Results of the representation of these genes among tested insect genomes are shown in Figure 3.2, mapped onto a phylogenetic tree of the species tested.

The phylogenetic tree in Figure 3.2 was constructed as follows. A suite of 1,400 *Apis mellifera* coding exons that are present in single copy widely in insects was obtained, starting with all the coding exons annotated by Ensembl Amel4.0, and using tblastn to eliminate exons that are absent or present in more than one copy in the *Apis mellifera* (Amel 4.5 assembly), *Tribolium casteneum* (Tcas3.0 assembly), *Drosophila melanogaster* (dm3 assembly), or *Ladonia fulva* (Lful_1.0 assembly) genomes. This suite of exons was used as query in tblastn searches of all the species shown in Figure 3.2 plus the Arachnid *Ixodes scapularis* (IxscaW3 assembly) for rooting. The single best tblastn match for each genome was extracted and those

that were reciprocal best blast matches to the correct *Apis mellifera* coding exon were retained for alignment and tree building. For each exon, muscle3.8.31 (default parameters) was used to generate a multiple protein alignment, and positions in the alignment with more than 10% gaps were removed (Edgar 2004). All the exon alignments were then concatenated to generate a large protein multiple alignment (average of 59,188 amino acids per species). This alignment was used to construct a maximum-likelihood phylogenetic tree with phyml (LG model, 6 rate classes, SPR moves). The tree obtained agrees well with all recent analyses of these species (Simon et al. 2012, Peters et al. 2014, Johnson et al. 2013), though no single published tree contains this exact set of species.

PCR

Four HTS were tested by PCR amplification. Primers were designed such that the reverse primer was placed in a region alignable between the wasp PDV sequence (see Fig. 3) and the Lepidopteran sequence (i.e. having strong homology between the sequences). The forward primer was placed in a region that was unalignable between the wasp and Lepidopteran sequences (regions lacking homology). Thus the PCR product spanned the junction between the transferred regions and non-transferred regions.

The PCR product was run on a 1.5% agarose gel with a 100bp ladder and stained with ethidium bromide. The 20ul PCR reaction contained: 5ul genomic DNA at 30ng/ul, 5ul New England Biolabs 5X taq master mix, 1ul forward primer, 1ul reverse primer, 8ul H20.

Different strains of *Bombyx mori* were tested for the presence of these hits (all strains courtesy of Marian Goldsmith, University of Rhode Island). **p50**: Inbred Chinese-originating strain used in sequencing by Japanese part of *B. mori* sequencing project. Originated from lab of Toru Shimada, University of Tokyo. **106**: Chinese strain. **Nistari**: a multivoltine (multiple generations per year; no diapause) strain originally from India then brought to an INRA lab in Lyon. **108**: Chinese strain **401**: Chinese strain reported to have BT-resistance. **418**: Chinese strain. **214**: Japanese strain. **555**: European strain.

Selected primers are drawn on the relevant sequences in Figure 3.3 and detailed below. Gels shown in Figure 3.4.

Figures 3.4A and 3.4B:

PDV32. Forward primer 5'-TTTCACCATCGTCTCGTCCC-3' (nscaf2876: 1094875- 1094894).

Reverse primer 5'- AGGCAGCTGGTTGTGAACAG-3' (nscaf2876: 1095911- 1095892).

Expected product size: 1037.

PDV101. Forward primer 5'- AGCAACGTGAGAACTCTACGAA-3' (nscaf3026: 4678131-4678110). Reverse primer 5'- AGGCAGCTGGTTGTGAACAG-3' (nscaf3026: 4677042-4677061). Expected product size: 1090.

Figure 3.4C:

PDV101 was run with two different forward primers against a single reverse primer. Reverse primer 5'-TGCGCTTTGTCTCGGATCTT-3' (nscaf3026: 4676695-4676676).

Forward primer 1 5'- AGTGATGCGGAACCAGTGAG-3' (nscaf3026: 4676010-4676029)

Expected product size = 686

Forward primer 2 5'-TGCTGCAATTGACTAAACCGC-3' (nscaf3026: 4675647-4675667.)

Expected product size: 1049.

Figure 3.4D:

PDV32. Reverse primer 5'-CAAAACGTGCCAGAGCCAAA-3' (nscaf2876: 1095967-1095948).

Forward primer 5'-TGGCAGACGAGCTCACAAAT-3' (nscaf2876: 1095021-1095040.)

Expected product size: 947.

Figure 3.4E:

PDV100. Forward primer 5'-TAGACACATCAGCGCAACCA-3' (nscaf2953: 1176106- 1176125)

Reverse primer 5'-AGCCAGAGTACCCGTTTTCG-3' (nscaf2953: 1176947-1176928)

Expected product size: 842

Figure 3.4F:

PDV100. Forward primer 5'-ACCAGACGAGCTTGTTGTGA-3' (nscaf2953: 1175996-1176015)

Reverse primer 5'-TACAGTTCCGGGAGTACGGA-3' (nscaf2953: 1176885- 1176866)

Expected product size: 890

Figure 3.4G:

PDV99. Forward primer 5'-TTCGACTCACGAGGAGCCTA-3' (nscaf2734: 12767- 12786)

Reverse primer 5'-AGTGGGACGAAAGTTGCCAG-3' (nscaf2734: 13515- 13496)

Expected product size: 749

Figure 3.4H:

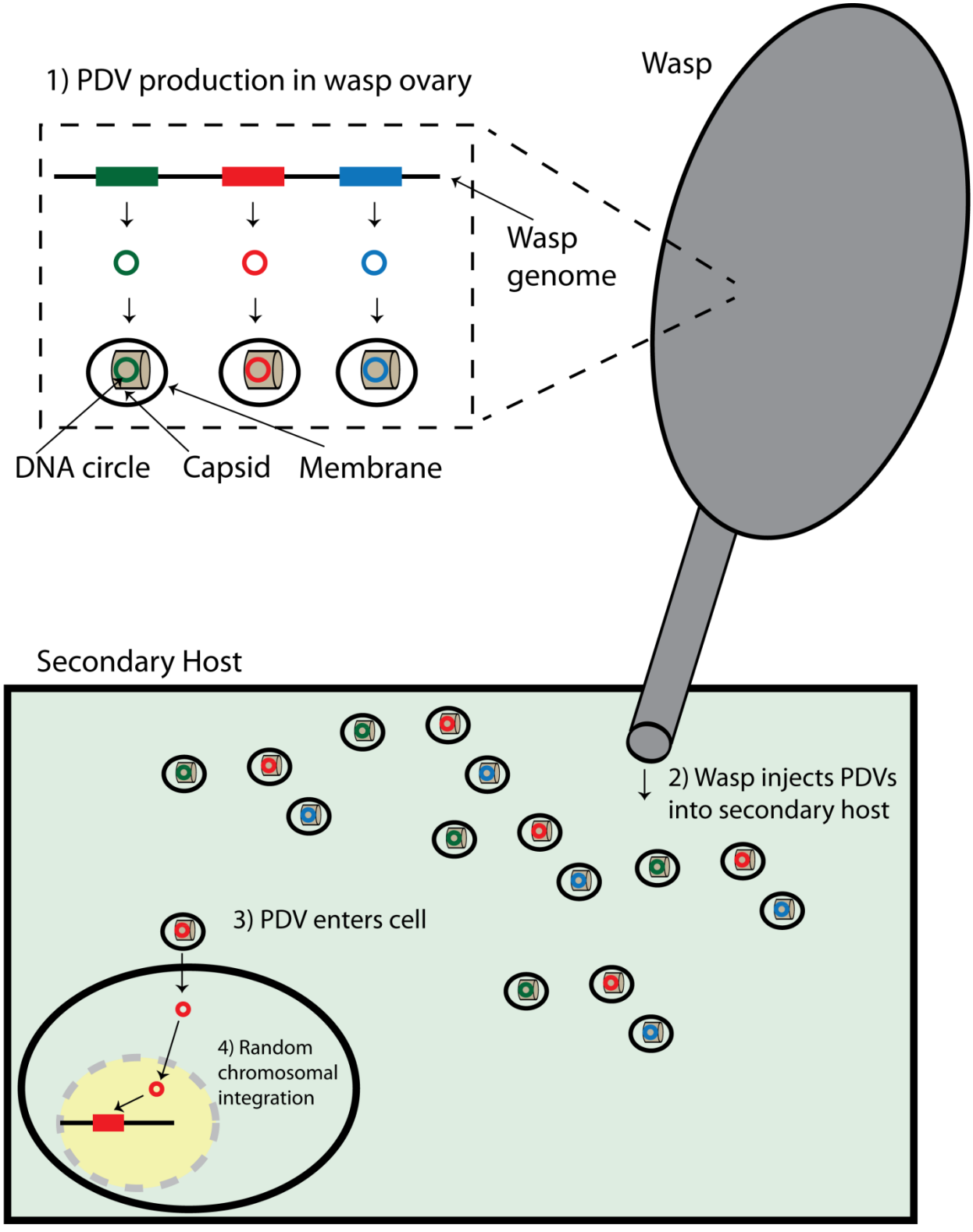
PDV99. Forward primer 5'-GGGCGTTAACAATGCCAAGG-3' (nscaf2734: 12589-12608)

Reverse primer 5'-CGAAAGTTGCCAGTTTCCGC-3' (nscaf2734: 13508-13489)

Expected product size: 920

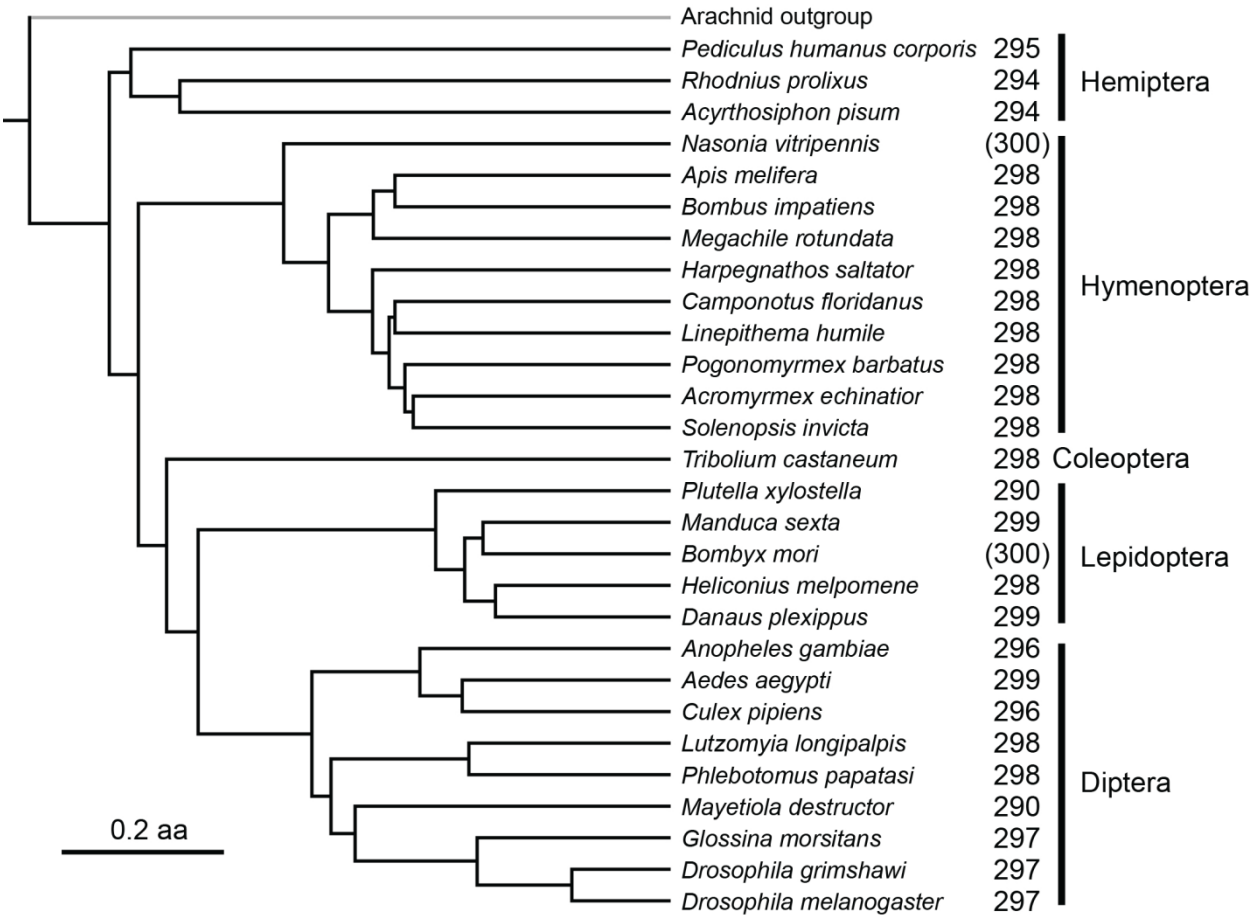
Genomic DNA from other insect species was also used as a negative control. Wild type *Drosophila melanogaster* was obtained from Celeste Berg. *Apis melifera* and *Chlosyne lacinia* were obtained from Charles Laird.

Figure 3.1: Bracovirus production and integration



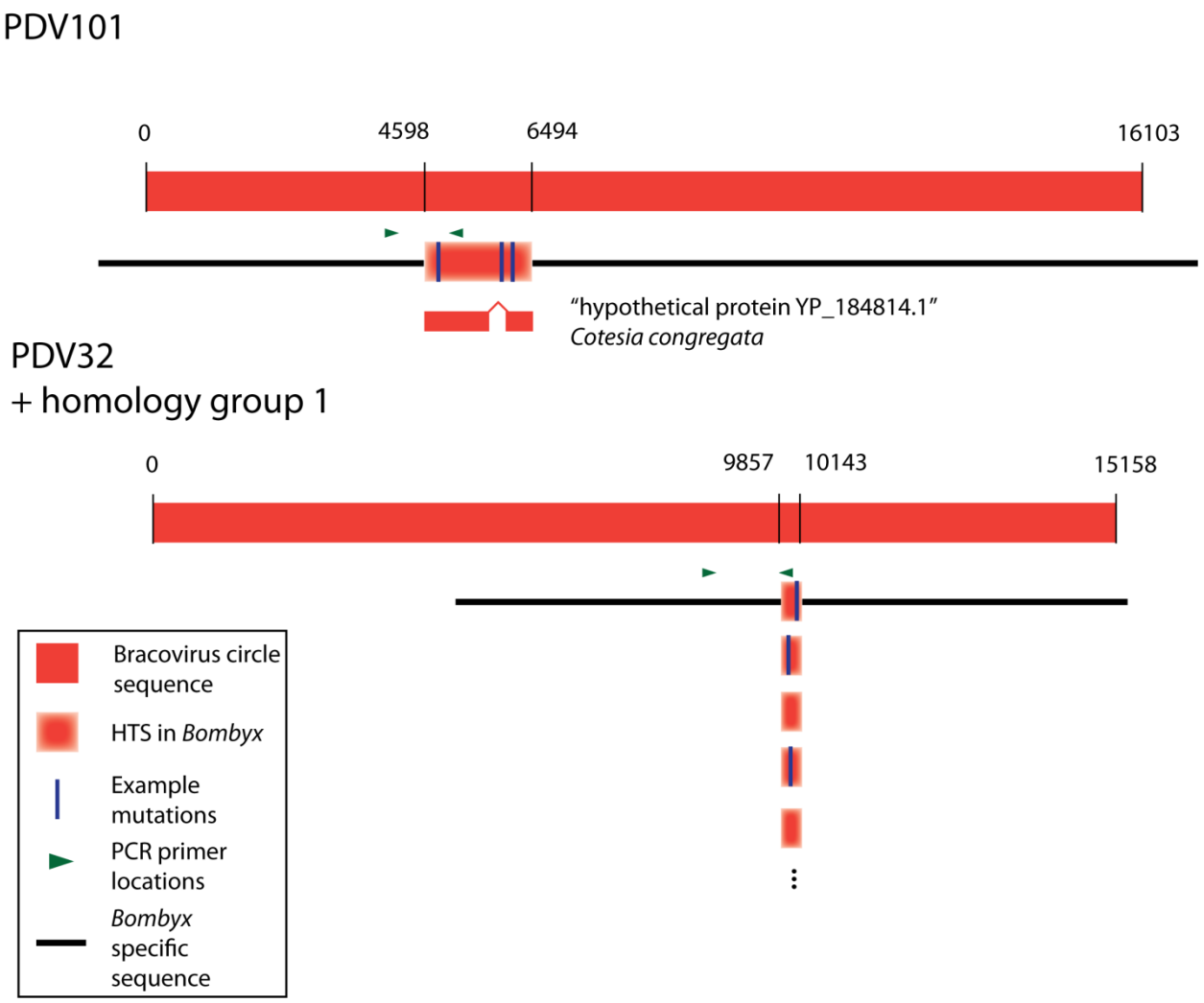
In part 1 specific portions of the wasp genome are targeted, replicated, and circularized into bracovirus DNA circles. These circles are produced in the wasp ovary then packaged in a viral capsid and lipid membrane. In Part 2 the bracoviruses are injected into the secondary host. In part 3 the bracovirus enters the secondary host cell, shedding the capsid. This typically occurs in haemocytes, with the proteins produced by the transferred DNA destroying or disabling the haemocyte's ability to defend the secondary host against wasp larvae. Occasionally, bracovirus DNA will integrate into the secondary host nuclear DNA as shown in part 4. If steps 3 and 4 occur in a germ line cell the bracovirus DNA may be passed on to the secondary host's offspring. Subsequent genetic drift or selection can result in the HTS becoming fixed in the population.

Figure 3.2: Representation of highly conserved sequences in insect genomes



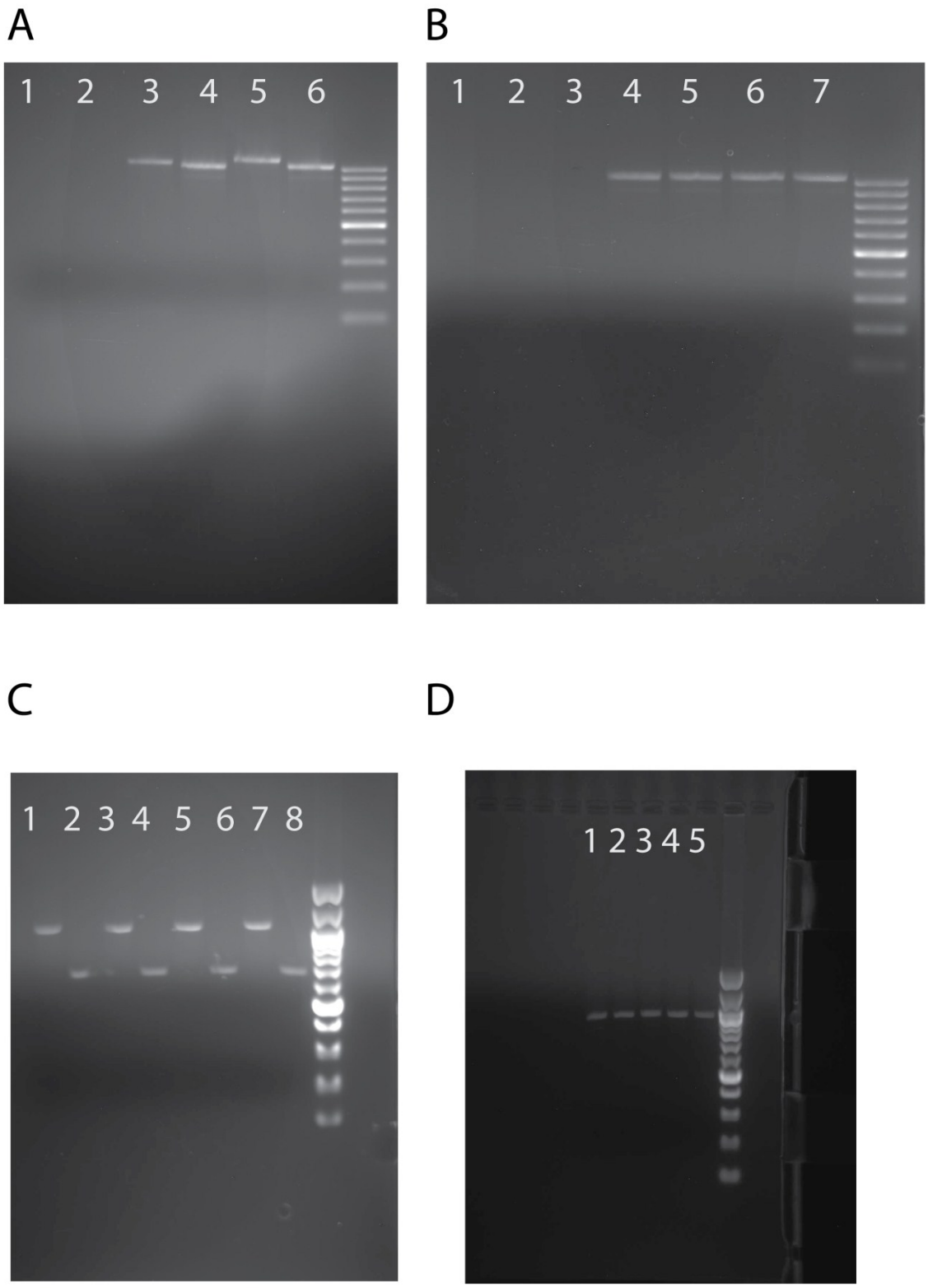
Phylogenetic tree generated from sequence data using PHYML. Coding regions from *Nasonia vittripennis* were used as tblastx queries against the *Bombyx mori* genome assembly to identify highly conserved regions. 300 conserved regions were selected at random and searched for in a broad range of insect genome assemblies using tblastx. In each insect species, 290-299 of these highly conserved regions were found (a rate of 96-99%). Missing sequences are an unknown combination of genomic deletions and incomplete assemblies. With a divergence time of ~300 MYA between *Bombyx* and *Nasonia*, there has been sufficient time for genomic deletions to occur. Branch lengths are based on amino acid divergence between species.

Figure 3.3: Example HTS in Silkworm(*Bombyx mori*)

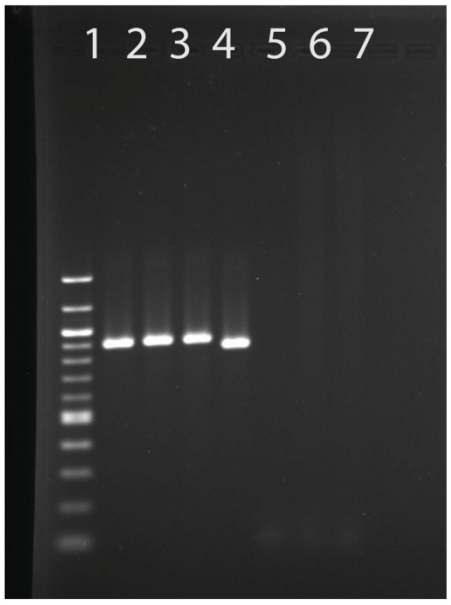


Two examples of HTS found in *Bombyx* (the same two HTS targeted for PCR in figure 4) aligned to bracovirus sequences from *Cotesia congregata*. The top portion shows a HTS (PDV101) aligned to *Cotesia congregata* bracovirus circle 12 (NC_006644.1). Predicted protein in the donor sequence shown (YP_184814.1). The bottom portion show a HTS (PDV32) aligned to *Cotesia congregata* bracovirus circle 5 (NC_006637.1). PDV32 is part of homology group 1, which is shown aligned to PDV32. In both examples the HTS have undergone point mutations and insertion/deletions that are too numerous to accurately represent here so example mutations have been drawn. Primers used in the PCR reaction shown in figure 4 are shown with green triangles.

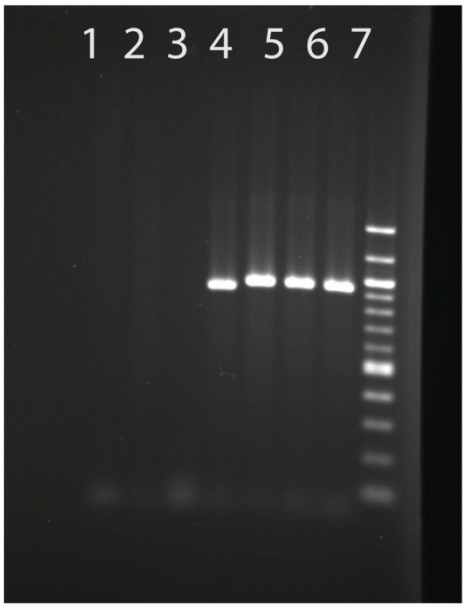
Figure 3.4: PCR amplification of HTS in Silkworm(*Bombyx mori*)



E

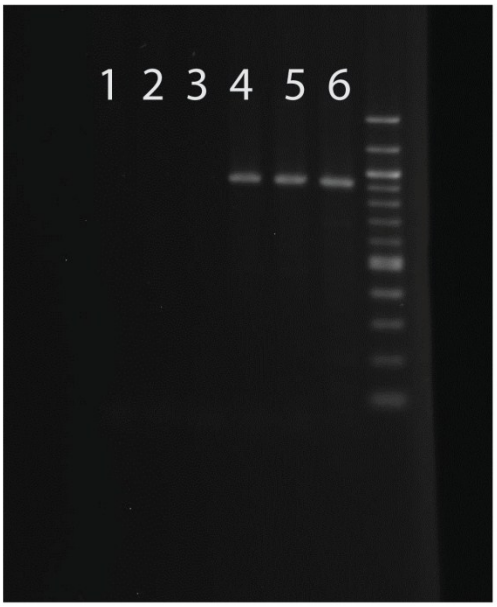


F



G

H



Gels displaying PCR amplification of HTS in *Bombyx mori*. Gels were ethidium bromide stained and run with a 100bp ladder (brighter bands at 500bp and 1000bp). Primers targeting PDV32 and PDV101 are shown in fig 3. Figure 4A shows results for PDV32 and PDV101. Figure 4B shows results for PDV32. Figures 4C and 4D show alternative PCR primers amplifying PDV32 and PDV101. Figures 4E and 4F show result for primers targeting PDV100. Figures 4G and 4H show result for primers targeting PDV99 Note that in parts E and F the Bombyx strain 106 appears to have a deletion polymorphism yielding a smaller fragment than other strains.

A) PCR results showing amplification of a product spanning the junction of the transferred wasp DNA and native *Bombyx* sequence. Lanes 2, 4, and 6 have primers targeting PDV32 (expected band size: 1037) and lanes 1, 3, and 5 have primers targeting PDV101(expected band size 1090). Lanes 5 and 6 use *Bombyx mori* strain 418(Chinese), lanes 3 and 4 *Bombyx mori* strain 214(Japanese), lanes 1 and 2 wild type *Drosophila melanogaster* as a negative control. **B)** We tested the same primer set above targeting PDV32 against a diverse panel of insect genomic DNA (all lanes tested with the same primers). DNA used in the reactions was as follows. Lane 1: *Chlosyne lacinia* (Lepidoptera). Lane 2: *Apis mellifera* (honeybee, Hymenoptera). Lane 3: wild type *Drosophila melanogaster* (Diptera). Lane 4: *Bombyx mori* strain 555 (European). Lane 5: *Bombyx mori* strain Nistari (Indian). Lane 6: *Bombyx mori* strain 418 (Chinese). Lane 7: *Bombyx mori* strain 214 (Japanese). **C)** PCR results for reaction targeting PDV 101 with alternative primers. Lanes alternate between the two different forward primers for the reaction (expected product sizes of 1049 and 686). Tested four strains (in order) with each pair of primer sets: 418 (Chinese), 214(Japanese), Nistari (Indian multivoltene), 555(European). **D)** PCR results for alternative primers targeting PDV32 (expected product size of 947). Five strains were tested: 418 (Chinese), 214(Japanese), 401(Chinese BT-resistant), Nistari (Indian multivoltene), 555(European). **E)** PCR results for primers targeting PDV100. Lane1: B. mori 214(Japanese). Lane2: B. mori 401(Chinese). Lane 3: B. mori 108(Chinese). Lane 4: B. mori 106(Chinese). Lane 5: Drosophila Melanogaster negative control. Lane 6:

Apis Melifera negative control. Lane 7: Chlosynne lacinia (butterfly) negative control. **F)** PCR results for primers targeting PDV100. Lane 1: Chlosynne lacinia (butterfly) negative control. Lane 2: Apis Melifera (honeybee) negative control. Lane 3: Drosophila Melanogaster negative control. Lane 4: B. mori 106(Chinese). Lane 5: B. mori 108(Chinese). Lane 6: B. mori 401(Chinese). Lane7: B. mori 214(Japanese). **G)** PCR results for primers targeting PDV99. Lane 1: B. mori 214(Japanese). Lane 2: B. mori 401(Chinese). Lane 3: B. mori 108(Chinese). Lane 4: Apis Melifera (honeybee) negative control. Lane 5: Drosophila Melanogaster negative control. Lane 6: Tenebrio molitor(mealworm) negative control. **H)** Lane 1: Tenebrio molitor(mealworm) negative control. Lane 2: Drosophila Melanogaster negative control. Lane 3: Apis Melifera (honeybee) negative control. Lane 4: B. mori 108(Chinese). Lane 5: B. mori 401(Chinese). Lane 6: B. mori 214(Japanese).

Table 3.1: Information on PDV HTS

PDV#	Group	Donor wasp species	matching viral sequence(query)	Recipient species	Contig(database)	%id	sequence length	query start	query stop	database start	database stop	e-value	score	dnds	Pfam	PCR
1	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2964	88.78	349	9855	10152	2829230	2829578	2.00E-76	285	0.8336	-	-
2	1	Cotesia congregata	NC_006637.1	Bombyx mori	nscaf3034	86.67	349	5647	5984	4872439	4872787	3.00E-72	272	1.054	-	-
3	1	Cotesia congregata	NC_006637.1	Bombyx mori	nscaf3099	86.2	351	5648	5995	1793619	1793969	5.00E-71	268	1.5639	-	-
4	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2827	87.84	376	9856	10146	1348326	1348701	2.00E-70	266	1.3718	-	-
5	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3026	87.84	343	9856	10146	4299130	4298788	2.00E-70	266	3.4834	-	-
6	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold14798	86.89	341	9854	10176	151	491	2.00E-70	266	1.1736	-	-
7	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold1935	86.89	336	9854	10176	551	216	2.00E-70	266	1.1235	-	-
8	1	Cotesia congregata	NC_006637.1	Bombyx mori	scaffold23953	86.97	341	5646	5970	599	259	8.00E-70	264	0.7043	-	-
9	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold27348	86.85	340	9855	10176	496	157	8.00E-70	264	3.3829	-	-
10	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2839	86.81	339	9856	10176	103175	102837	3.00E-69	262	2.2367	-	-
11	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3031	86.81	339	9856	10176	4974594	4974932	3.00E-69	262	1.5076	-	-
12	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold13551	86.81	339	9856	10176	697	359	3.00E-69	262	3.0969	-	-
13	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold15214	86.81	339	9856	10176	105	443	3.00E-69	262	0.8643	-	-
14	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold1767	86.81	339	9856	10176	3074	3412	3.00E-69	262	1.3091	-	-
15	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold22468	86.81	339	9856	10176	614	276	3.00E-69	262	1.1926	-	-
16	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold2548	87.58	311	9854	10146	316	6	3.00E-69	262	1.2627	-	-
17	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold26569	86.81	339	9856	10176	62	400	3.00E-69	262	0.3233	-	-
18	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold31722	86.81	339	9856	10176	513	175	3.00E-69	262	1.8467	-	-
19	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold5751	86.81	341	9856	10176	348	8	3.00E-69	262	0.523	-	-
20	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2818	87.54	299	9855	10146	1627974	1628272	1.00E-68	260	2.6921	-	-
21	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3034	87.5	346	9856	10146	864952	864607	5.00E-68	258	1.3069	-	-
22	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold30325	87.5	339	9856	10146	554	216	5.00E-68	258	0.5153	-	-
23	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2330	86.54	340	9855	10176	4328068	4327729	2.00E-67	256	0.6716	-	-
24	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2767	86.54	347	9855	10176	4060686	4061032	2.00E-67	256	1.947	-	-
25	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2902	86.5	339	9856	10176	1373714	1374052	8.00E-67	254	1.3858	-	-
26	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2964	86.5	339	9856	10176	1791019	1791357	8.00E-67	254	1.3964	-	-
27	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3027	86.5	341	9856	10176	1985313	1985653	8.00E-67	254	2.9499	-	-
28	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3099	86.5	339	9856	10176	1638115	1638453	8.00E-67	254	0.971	-	-
29	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold29692	86.5	341	9856	10176	424	84	8.00E-67	254	1.0683	-	-
30	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold41206	86.5	339	9856	10176	34	372	8.00E-67	254	0.8739	-	-
31	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2795	87.16	321	9856	10146	1669602	1669282	1.00E-65	250	3.6816	-	-
32	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2876	87.33	559	9857	10143	1095376	1095934	2.00E-65	250	1.0603	-	yes
33	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3032	87.16	299	9856	10146	1509080	1508782	1.00E-65	250	1.8447	-	-
34	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2767	86.24	340	9855	10176	1961830	1961491	5.00E-65	248	1.2487	-	-
35	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2931	87.25	352	9856	10146	371407	371056	5.00E-65	248	3.3829	-	-
36	1	Cotesia congregata	NC_006637.1	Bombyx mori	nscaf3013	86.36	357	5646	5970	462656	462300	5.00E-65	248	3.2836	-	-
37	1	Cotesia congregata	NC_006648.1	Bombyx mori	scaffold1807	86.15	370	9857	10176	542	173	8.00E-64	244	3.3829	-	-
38	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2398	87.59	282	9856	10117	104187	103906	4.00E-62	238	3.3829	-	-
39	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf2811	87.04	344	9856	10152	125368	125711	5.00E-62	238	1.794	-	-
40	1	Cotesia congregata	NC_006654.1	Bombyx mori	nscaf2556	89.04	243	749	526	155814	156056	5.00E-61	234	2.7673	-	-
41	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf1898	85.58	339	9856	10176	2938281	2937943	1.00E-59	230	3.3829	-	-
42	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3031	88.66	254	9857	10090	4635218	4635471	3.00E-59	228	2.7586	-	-
43	1	Cotesia congregata	NC_006654.1	Bombyx mori	scaffold2598	88.5	241	747	526	774	534	2.00E-57	222	2.4064	-	-
44	1	Cotesia congregata	NC_006654.1	Bombyx mori	nscaf2813	88.89	207	747	545	113948	114154	1.00E-55	216	0.0672	-	-
45	1	Cotesia congregata	NC_006640.1	Bombyx mori	nscaf3075	95.65	290	119	253	681633	681922	2.00E-54	212	3.3829	-	-
46	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3034	85.34	309	9851	10146	1406043	1405735	9.00E-54	210	3.306	-	-
47	1	Cotesia congregata	NC_006637.1	Bombyx mori	scaffold6369	85	271	5713	5970	385	115	7.00E-39	161	3.4834	-	-
48	1	Cotesia congregata	NC_006648.1	Bombyx mori	nscaf3015	91.51	243	9857	9959	1534168	1533926	3.00E-31	135	2.3231	-	-
49	1	Cotesia congregata	NC_006637.1	Bombyx mori	nscaf2888	96.83	236	5714	5776	1685050	1685285	2.00E-23	109	3.3829	-	-
50	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300013	88.8	390	8604	8364	473382	473771	2.00E-67	256	0.967	-	-
51	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300179	88.52	385	8605	8362	232886	232502	9.00E-67	254	3.3829	-	-
52	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300247	88.61	376	8369	8605	109246	109621	5.00E-65	248	3.508	-	-
53	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300545	85.8	851	13361	13040	27049	27899	1.00E-61	238	1.3504	-	-
54	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300710	85.8	1159	13361	13040	4425	3267	2.00E-61	238	0.7679	-	-
55	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300033	88.61	679	8605	8369	60141	60819	6.00E-60	232	2.3508	-	-
56	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300024	82.92	436	12977	13400	986654	986219	1.00E-59	230	1.2444	-	-
57	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300019	85.63	414	13040	13361	40017	40430	2.00E-58	226	0.2809	-	-
58	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300097	84.37	429	12977	13302	180823	181251	2.00E-58	226	0.7242	-	-
59	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300121	88.57	382	8364	8573	266213	266594	5.00E-56	218	0.2191	-	-
60	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300234	87.22	270	8582	8362	11869	12138	1.00E-55	216	0.1749	-	-
61	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300223	88.68	697	8582	8372	766921	767617	1.00E-54	214	0.1961	-	-
62	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300013	88.13	267	8582	8364	480370	480636	2.00E-54	212	0.3327	-	-
63	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300056	85.29	275	13136	13402	590211	589937	8.00E-54	210	0.4903	-	-
64	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300101	84.89	285	12971	13236	448376	448092	1.00E-52	206	0.4806	-	-
65	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300135	89.06	382	8561	8370	256214	255833	2.00E-52	206	0.4798	-	-
66	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300352	88.44	207	13236	13040	110880	110674	1.00E-51	202	0.0103	-	-
67	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300140	87.11	241	8605	8381	304568	304328	7.00E-51	200	0.7245	-	-
68	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF301259	88.29	313	8573	8369	2641	2953	9.00E-51	200	0.7488	-	-
69	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300205	85.66	278	13131	13385	239143	239420	1.00E-49	196	0.6184	-	-
70	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300119	87.8	209	13344	13141	347057	347265	1.00E-48	192	0.7192	-	-
71	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300135	88.12	224	13106	13305	429838	429615	2.00E-48	192	0.6922	-	-
72	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300168	86.35	297	13142	13388	182194	182490	2.00E-48	192	0.4731	-	-
73	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300118	86.55	252	13110	13344	488329	488078	3.00E-47	188	0.4604	-	-
74	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300442	87.69	282	8570	8376	72560	72841	3.00E-47	188	0.3398	-	-
75	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300004	84.73	262	12971	13222	1070284	1070545	1.00E-46	186	0.3117	-	-
76	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300353	88.36	265	13137	13323	70745	70481	1.00E-46	186	0.4727	-	-
77	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300006	87.44	291	13040	13236	99518	99808	5.00E-46	184	1.1323	-	-
78	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300065	86.55	262	13142	13361	146914	147175	1.00E-43	176	1.8185	-	-
79	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300012	93.16	261	13106	13222	1264705	1264965	3.00E-41	168	0.4706	-	-
80	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300021	89.81	233	13141	13297	1452534	1452766	2.00E-41	168	0.7265	-	-
81	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300360	93.16	252	13106	13222	21665	21414	3.00E-41	168	0.6006	-	-
82	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300259	87.43	254	13144	13323	271099	271352	1.00E-40	167	0.4284	-	-
83	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300174	88.55	227	13040	13203	444412	444638	4.00E-40	165	1.4229	-	-
84	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300228	84.29	268	13144	13401	177424	177157	2.00E-39	163	0.431	-	-
85	2	Cotesia vestalis	HQ009532.1	Danaus plexippus	DPSCF300001	85.78	226	8605	8404	2148566	2148341	5.00E-39	161	46.654	-	-
86	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300127	84.38	290	13326	13075	210642	210931	3.00E-38	159	1.5257	-	-
87	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300109	89.51	241	13061	13203	88497	88257	9.00E-38	157	0.7161	-	-
88	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300909	85.65	208	13140	13344	3284	3491	8.00E-38	157	0.678	-	-
89	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300020	90.68	263	13105	13222	372924	373186	1.00E-34	147	0.857	-	-
90	2	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300203	83.26	264	13143	13361	174272	174009	9.00E-26	117	86.2626	-	-
91	3	Cotesia vestalis	HQ009550.1	Bombyx mori	nscaf2853	85.71	237	27596	27815	4835780	4836016	3.00E-47	188	0.9501	-	-
92	3	Cotesia vestalis	HQ009550.1	Bombyx mori	nscaf3022	85.12	224	27605	27815	624379	624602	6.00E-42	170	0.0095	-	-
93	3	Cotesia vestalis	HQ009550.1	Bombyx mori	nscaf2930	82.3	212	27605	27810	5618511	5618300	4.00E-24	111	1.3172	-	-
94	4	Cotesia congregata	NC_006654.1	Danaus plexippus	DPSCF300640	88.46	6570	747	2945	6168	12737	1.00E-200	2337	1.5364	BEN, Pfam-B_3333	-
95	4	Cotesia congregata	NC_006654.1	Danaus plexippus	DPSCF300080	89.55	4091	1407	2945	500672	504762	1.00E-200	1770	0.303	-	-
96	4	Cotesia congregata	NC_006660.1	Danaus plexippus	DPSCF300216	90.33	2188	3833	4752	118567	116380	1.00E-200	1112	1.0305	BEN	-
97	5	Cotesia congregata	NC_006660.1	Danaus plexippus	DPSCF300320	91.15	409	9502	9919	22086	21678	7.00E-154	543	0.7362	-	-
98	5	Cotesia congregata	NC_006660.1	Danaus plexippus	DPSCF300367	91.67	339	9502	9849	24521	24859	3.00E-131	468	0.3233	-	-
99	6	Cotesia vestalis	HQ009550.1	Bombyx mori	nscaf2734	85.94	207	27606	27733	13474	13680	4.00E-24	111	0.3989	-	yes
100	7	Cotesia congregata	NC_006645.1	Bombyx mori	nscaf2953	87.9	415	2730	2976	1176763	1177177	2.00E-64	246	0.4589	-	yes
101	8	Cotesia congregata	NC_006644.1	Bombyx mori	nscaf3026	87.68	1918	4598	6494	4677185	4675268	1.00E-200	1861	0.4301	-	yes
102	9	Cotesia congregata	NC_006633.1	Danaus plexippus	DPSCF300035	92.39	891	13801	13513	46422	47312	6.00E-110	398	0.6675	-	-
103	10	Cotesia vestalis	HQ009530.1	Danaus plexippus	DPSCF300063	82.38	245	12710	12901	1108985	1108741	3.00E-22	105	0.4964	-	-
104	11	Cotesia vestalis	HQ009545.1	Danaus plexippus	DPSCF300088	82.52	338	6218	6543	1248817	1248480	2.00E-48	192	0.4756	-	-
105	12	Cotesia congregata	NC_006639.1	Danaus plexippus	DPSCF300184	87.72	1547	5608	6731	130535	128989	1.00E-200	1134	0.0696	-	-

The first column “PDV#” is an arbitrary numbering scheme for the HTS. “Group” sorts the found PDV sequences into related groups. “Donor wasp species” gives the wasp species that the matching PDV derives from. “matching viral sequence” is the viral sequence a match was found for. “Recipient species” is the species the transferred sequence was found in. “Contig” shows the contig of a sequence. “%id” shows the percentage of matching nucleotides reported by BLAST. “sequence length” is the length of the horizontally transferred sequence. “query start” and “query stop” are the start and stop of the match on the query sequence (virus). “database start” and “database stop” are the start and stop of the match on the database (eukaryote genome). “evalue” is the probability of the match by chance. “score” is the BLAST alignment score. “dn/ds” is the ratio of synonymous to non-synonymous changes between the query virus and the HTS found. . “Pfam” shows protein family matches for the sequence. “PCR” shows if the sequence was tested by PCR.

Table 3.2: Polydnavirus sequences used as search queries

Polydnavirus Accession #	Wasp Species	Length	Total length
NC_007995.1	Campoletis sonorensis ichnovirus	8864	
NC_008003.1	Campoletis sonorensis ichnovirus	15510	
NC_008007.1	Campoletis sonorensis ichnovirus	15812	
NC_007988.1	Campoletis sonorensis ichnovirus	7760	
NC_007990.1	Campoletis sonorensis ichnovirus	8171	
NC_007992.1	Campoletis sonorensis ichnovirus	8600	
NC_008004.1	Campoletis sonorensis ichnovirus	7887	
NC_008008.1	Campoletis sonorensis ichnovirus	7604	
NC_007993.1	Campoletis sonorensis ichnovirus	9155	
NC_007998.1	Campoletis sonorensis ichnovirus	11285	
NC_007994.1	Campoletis sonorensis ichnovirus	9391	
NC_007996.1	Campoletis sonorensis ichnovirus	10223	
NC_007987.1	Campoletis sonorensis ichnovirus	6567	
NC_007985.1	Campoletis sonorensis ichnovirus	12098	
NC_008000.1	Campoletis sonorensis ichnovirus	14531	
NC_007999.1	Campoletis sonorensis ichnovirus	11841	
NC_008005.1	Campoletis sonorensis ichnovirus	6283	
NC_007997.1	Campoletis sonorensis ichnovirus	6138	
NC_008001.1	Campoletis sonorensis ichnovirus	14825	
NC_007989.1	Campoletis sonorensis ichnovirus	8284	
NC_007991.1	Campoletis sonorensis ichnovirus	8315	
NC_008002.1	Campoletis sonorensis ichnovirus	19557	
NC_008006.1	Campoletis sonorensis ichnovirus	10757	
NC_007986.1	Campoletis sonorensis ichnovirus	7276	
AF411011.1	Campoletis_sonorensis_ichnovirus_segment_A	6283	
AY029395.1	Campoletis_sonorensis_ichnovirus_segment_O1	11285	
AY029396.1	Campoletis_sonorensis_ichnovirus_segment_P	11841	
AY029397.1	Campoletis_sonorensis_ichnovirus_segment_T	14531	
AY029398.1	Campoletis_sonorensis_ichnovirus_segment_U	14825	305499
Z31378.2	Chelonus_inanitus_bracovirus_segment_V16.8	17789	17789
NC_006659.1	Cotesia congregata bracovirus	41573	
NC_006649.1	Cotesia congregata bracovirus	32108	
NC_006636.1	Cotesia congregata bracovirus	15876	
NC_006657.1	Cotesia congregata bracovirus	19820	
NC_006660.1	Cotesia congregata bracovirus	20197	
NC_006647.1	Cotesia congregata bracovirus	8785	
NC_006639.1	Cotesia congregata bracovirus	24748	
NC_006641.1	Cotesia congregata bracovirus	15959	
NC_006655.1	Cotesia congregata bracovirus	15279	
NC_006637.1	Cotesia congregata bracovirus	14489	
NC_006654.1	Cotesia congregata bracovirus	13597	
NC_006643.1	Cotesia congregata bracovirus	12903	
NC_006653.1	Cotesia congregata bracovirus	26062	
NC_006651.1	Cotesia congregata bracovirus	19161	
NC_006652.1	Cotesia congregata bracovirus	4981	

NC_006640.1	Cotesia congregata bracovirus	5032	
NC_006645.1	Cotesia congregata bracovirus	21388	
NC_006646.1	Cotesia congregata bracovirus	31972	
NC_006648.1	Cotesia congregata bracovirus	15158	
NC_006661.1	Cotesia congregata bracovirus	11186	
NC_006638.1	Cotesia congregata bracovirus	15230	
NC_006635.1	Cotesia congregata bracovirus	29874	
NC_006633.1	Cotesia congregata bracovirus	27346	
NC_006662.1	Cotesia congregata bracovirus	17477	
NC_006650.1	Cotesia congregata bracovirus	18768	
NC_006658.1	Cotesia congregata bracovirus	30655	
NC_006642.1	Cotesia congregata bracovirus	14286	
NC_006656.1	Cotesia congregata bracovirus	12682	
NC_006634.1	Cotesia congregata bracovirus	14975	
NC_006644.1	Cotesia congregata bracovirus	16103	567670
EF067319.1	Cotesia_plutellae_bracovirus_segment_S10	14184	
DQ075360.1	Cotesia_plutellae_bracovirus_segment_S11	11982	
EF067320.1	Cotesia_plutellae_bracovirus_segment_S14	6349	
EF067321.1	Cotesia_plutellae_bracovirus_segment_S16	7482	
DQ075354.1	Cotesia_plutellae_bracovirus_segment_S2	13848	
EF067322.1	Cotesia_plutellae_bracovirus_segment_S21	12172	
EF067323.1	Cotesia_plutellae_bracovirus_segment_S22	18065	
EF067324.1	Cotesia_plutellae_bracovirus_segment_S27	13909	
AY651829.1	Cotesia_plutellae_bracovirus_segment_S28	21258	
DQ075355.1	Cotesia_plutellae_bracovirus_segment_S3	11017	
AY651828.1	Cotesia_plutellae_bracovirus_segment_S30	23535	
AY651830.1	Cotesia_plutellae_bracovirus_segment_S33	23697	
EF067325.1	Cotesia_plutellae_bracovirus_segment_S35	11724	
EF067326.1	Cotesia_plutellae_bracovirus_segment_S36	18893	
EF067327.1	Cotesia_plutellae_bracovirus_segment_S37	14945	
EF067328.1	Cotesia_plutellae_bracovirus_segment_S38	14096	
DQ075356.1	Cotesia_plutellae_bracovirus_segment_S4	14943	
EF067329.1	Cotesia_plutellae_bracovirus_segment_S41	17528	
EF067330.1	Cotesia_plutellae_bracovirus_segment_S48	15432	
DQ075357.1	Cotesia_plutellae_bracovirus_segment_S5	14572	
EF067331.1	Cotesia_plutellae_bracovirus_segment_S50	13251	
EF067332.1	Cotesia_plutellae_bracovirus_segment_S51	17543	
DQ075358.1	Cotesia_plutellae_bracovirus_segment_S8	10763	
DQ075359.1	Cotesia_plutellae_bracovirus_segment_S9	10741	351929
HQ009524.1	Cotesia_vestalis_bracovirus_segment_c1	14213	
HQ009533.1	Cotesia_vestalis_bracovirus_segment_c10	17683	
HQ009534.1	Cotesia_vestalis_bracovirus_segment_c11	23267	
HQ009535.1	Cotesia_vestalis_bracovirus_segment_c12	21425	
HQ009536.1	Cotesia_vestalis_bracovirus_segment_c13	22954	
HQ009537.1	Cotesia_vestalis_bracovirus_segment_c14	13743	
HQ009538.1	Cotesia_vestalis_bracovirus_segment_c15	10941	
HQ009539.1	Cotesia_vestalis_bracovirus_segment_c16	14847	
HQ009540.1	Cotesia_vestalis_bracovirus_segment_c17	10780	
HQ009541.1	Cotesia_vestalis_bracovirus_segment_c18	12173	
HQ009542.1	Cotesia_vestalis_bracovirus_segment_c19	6353	

HQ009525.1	Cotesia_vestalis_bracovirus_segment_c2	7139	
HQ009543.1	Cotesia_vestalis_bracovirus_segment_c20	39213	
HQ009544.1	Cotesia_vestalis_bracovirus_segment_c21	20333	
HQ009545.1	Cotesia_vestalis_bracovirus_segment_c22	8662	
HQ009546.1	Cotesia_vestalis_bracovirus_segment_c23	14264	
HQ009547.1	Cotesia_vestalis_bracovirus_segment_c24	36834	
HQ009548.1	Cotesia_vestalis_bracovirus_segment_c25	8112	
HQ009549.1	Cotesia_vestalis_bracovirus_segment_c26	16127	
HQ009550.1	Cotesia_vestalis_bracovirus_segment_c27	27881	
HQ009551.1	Cotesia_vestalis_bracovirus_segment_c28	23103	
HQ009552.1	Cotesia_vestalis_bracovirus_segment_c29	26482	
HQ009526.1	Cotesia_vestalis_bracovirus_segment_c3	17848	
HQ009553.1	Cotesia_vestalis_bracovirus_segment_c30	4512	
HQ009554.1	Cotesia_vestalis_bracovirus_segment_c31	6877	
HQ009555.1	Cotesia_vestalis_bracovirus_segment_c32	3811	
HQ009556.1	Cotesia_vestalis_bracovirus_segment_c33	2683	
HQ009557.1	Cotesia_vestalis_bracovirus_segment_c34	15157	
HQ009558.1	Cotesia_vestalis_bracovirus_segment_c35	5667	
HQ009527.1	Cotesia_vestalis_bracovirus_segment_c4	14038	
HQ009528.1	Cotesia_vestalis_bracovirus_segment_c5	11406	
HQ009529.1	Cotesia_vestalis_bracovirus_segment_c6	15209	
HQ009530.1	Cotesia_vestalis_bracovirus_segment_c7	17462	
HQ009531.1	Cotesia_vestalis_bracovirus_segment_c8	15829	
HQ009532.1	Cotesia_vestalis_bracovirus_segment_c9	13187	540215
NC_008862.1	Glypta fumiferanae ichnovirus	2298	
NC_008866.1	Glypta fumiferanae ichnovirus	2358	
NC_008899.1	Glypta fumiferanae ichnovirus	2879	
NC_008845.1	Glypta fumiferanae ichnovirus	1963	
NC_008882.1	Glypta fumiferanae ichnovirus	2625	
NC_008864.1	Glypta fumiferanae ichnovirus	2341	
NC_008887.1	Glypta fumiferanae ichnovirus	2723	
NC_008907.1	Glypta fumiferanae ichnovirus	3058	
NC_008898.1	Glypta fumiferanae ichnovirus	2875	
NC_008918.1	Glypta fumiferanae ichnovirus	3357	
NC_008939.1	Glypta fumiferanae ichnovirus	3008	
NC_008851.1	Glypta fumiferanae ichnovirus	2080	
NC_008876.1	Glypta fumiferanae ichnovirus	2548	
NC_008852.1	Glypta fumiferanae ichnovirus	2094	
NC_008872.1	Glypta fumiferanae ichnovirus	2454	
NC_008853.1	Glypta fumiferanae ichnovirus	2135	
NC_008911.1	Glypta fumiferanae ichnovirus	3130	
NC_008840.1	Glypta fumiferanae ichnovirus	1859	
NC_008879.1	Glypta fumiferanae ichnovirus	2590	
NC_008884.1	Glypta fumiferanae ichnovirus	2636	
NC_008871.1	Glypta fumiferanae ichnovirus	2452	
NC_008931.1	Glypta fumiferanae ichnovirus	4886	
NC_008892.1	Glypta fumiferanae ichnovirus	2785	
NC_008877.1	Glypta fumiferanae ichnovirus	2572	
NC_008869.1	Glypta fumiferanae ichnovirus	2408	
NC_008873.1	Glypta fumiferanae ichnovirus	2462	

NC_008927.1	Glypta fumiferanae ichnovirus	4066
NC_008937.1	Glypta fumiferanae ichnovirus	2759
NC_008846.1	Glypta fumiferanae ichnovirus	1963
NC_008922.1	Glypta fumiferanae ichnovirus	3519
NC_008926.1	Glypta fumiferanae ichnovirus	4059
NC_008843.1	Glypta fumiferanae ichnovirus	1945
NC_008895.1	Glypta fumiferanae ichnovirus	2854
NC_008888.1	Glypta fumiferanae ichnovirus	2723
NC_008868.1	Glypta fumiferanae ichnovirus	2395
NC_008839.1	Glypta fumiferanae ichnovirus	1750
NC_008909.1	Glypta fumiferanae ichnovirus	3111
NC_008919.1	Glypta fumiferanae ichnovirus	3391
NC_008875.1	Glypta fumiferanae ichnovirus	2540
NC_008920.1	Glypta fumiferanae ichnovirus	3403
NC_008916.1	Glypta fumiferanae ichnovirus	3331
NC_008890.1	Glypta fumiferanae ichnovirus	2768
NC_008863.1	Glypta fumiferanae ichnovirus	2318
NC_008891.1	Glypta fumiferanae ichnovirus	2775
NC_008883.1	Glypta fumiferanae ichnovirus	2631
NC_008932.1	Glypta fumiferanae ichnovirus	5156
NC_008908.1	Glypta fumiferanae ichnovirus	3083
NC_008930.1	Glypta fumiferanae ichnovirus	4829
NC_008921.1	Glypta fumiferanae ichnovirus	3498
NC_008933.1	Glypta fumiferanae ichnovirus	2060
NC_008837.1	Glypta fumiferanae ichnovirus	1533
NC_008900.1	Glypta fumiferanae ichnovirus	2924
NC_008861.1	Glypta fumiferanae ichnovirus	2295
NC_008860.1	Glypta fumiferanae ichnovirus	2285
NC_008936.1	Glypta fumiferanae ichnovirus	2588
NC_008901.1	Glypta fumiferanae ichnovirus	2930
NC_008906.1	Glypta fumiferanae ichnovirus	3036
NC_008844.1	Glypta fumiferanae ichnovirus	1948
NC_008905.1	Glypta fumiferanae ichnovirus	2997
NC_008894.1	Glypta fumiferanae ichnovirus	2819
NC_008925.1	Glypta fumiferanae ichnovirus	3821
NC_008885.1	Glypta fumiferanae ichnovirus	2641
NC_008941.1	Glypta fumiferanae ichnovirus	4459
NC_008880.1	Glypta fumiferanae ichnovirus	2598
NC_008874.1	Glypta fumiferanae ichnovirus	2509
NC_008914.1	Glypta fumiferanae ichnovirus	3243
NC_008870.1	Glypta fumiferanae ichnovirus	2415
NC_008849.1	Glypta fumiferanae ichnovirus	2074
NC_008902.1	Glypta fumiferanae ichnovirus	2933
NC_008923.1	Glypta fumiferanae ichnovirus	3743
NC_008857.1	Glypta fumiferanae ichnovirus	2199
NC_008915.1	Glypta fumiferanae ichnovirus	3307
NC_008924.1	Glypta fumiferanae ichnovirus	3761
NC_008904.1	Glypta fumiferanae ichnovirus	2944
NC_008934.1	Glypta fumiferanae ichnovirus	2309
NC_008928.1	Glypta fumiferanae ichnovirus	4340

NC_008886.1	Glypta fumiferanae ichnovirus	2646	
NC_008881.1	Glypta fumiferanae ichnovirus	2615	
NC_008858.1	Glypta fumiferanae ichnovirus	2235	
NC_008865.1	Glypta fumiferanae ichnovirus	2358	
NC_008889.1	Glypta fumiferanae ichnovirus	2738	
NC_008913.1	Glypta fumiferanae ichnovirus	3157	
NC_008896.1	Glypta fumiferanae ichnovirus	2856	
NC_008935.1	Glypta fumiferanae ichnovirus	2502	
NC_008867.1	Glypta fumiferanae ichnovirus	2380	
NC_008917.1	Glypta fumiferanae ichnovirus	3343	
NC_008938.1	Glypta fumiferanae ichnovirus	2840	
NC_008855.1	Glypta fumiferanae ichnovirus	2179	
NC_008893.1	Glypta fumiferanae ichnovirus	2808	
NC_008912.1	Glypta fumiferanae ichnovirus	3141	
NC_008859.1	Glypta fumiferanae ichnovirus	2262	
NC_008910.1	Glypta fumiferanae ichnovirus	3117	
NC_008850.1	Glypta fumiferanae ichnovirus	2079	
NC_008842.1	Glypta fumiferanae ichnovirus	1940	
NC_008848.1	Glypta fumiferanae ichnovirus	2066	
NC_008841.1	Glypta fumiferanae ichnovirus	1916	
NC_008847.1	Glypta fumiferanae ichnovirus	2005	
NC_008903.1	Glypta fumiferanae ichnovirus	2934	
NC_008854.1	Glypta fumiferanae ichnovirus	2177	
NC_008838.1	Glypta fumiferanae ichnovirus	1657	
NC_008897.1	Glypta fumiferanae ichnovirus	2871	
NC_008940.1	Glypta fumiferanae ichnovirus	3140	
NC_008878.1	Glypta fumiferanae ichnovirus	2579	
NC_008856.1	Glypta fumiferanae ichnovirus	2195	
NC_008929.1	Glypta fumiferanae ichnovirus	4707	291597
EU001259.1	Glyptapanteles_flavicoxis_bracovirus_segment_1	38757	
EU001267.1	Glyptapanteles_flavicoxis_bracovirus_segment_10	50693	
EU001268.1	Glyptapanteles_flavicoxis_bracovirus_segment_11	29631	
EU001269.1	Glyptapanteles_flavicoxis_bracovirus_segment_12	8074	
EU001270.1	Glyptapanteles_flavicoxis_bracovirus_segment_13	31710	
EU001271.1	Glyptapanteles_flavicoxis_bracovirus_segment_14	28120	
EU001272.1	Glyptapanteles_flavicoxis_bracovirus_segment_15	20527	
EU001273.1	Glyptapanteles_flavicoxis_bracovirus_segment_17	13595	
EU001274.1	Glyptapanteles_flavicoxis_bracovirus_segment_18	18404	
EU001275.1	Glyptapanteles_flavicoxis_bracovirus_segment_19	11120	
EU001260.1	Glyptapanteles_flavicoxis_bracovirus_segment_2	35842	
EU001276.1	Glyptapanteles_flavicoxis_bracovirus_segment_20	9701	
EU001277.1	Glyptapanteles_flavicoxis_bracovirus_segment_21	8137	
EU001278.1	Glyptapanteles_flavicoxis_bracovirus_segment_22	16845	
EU001279.1	Glyptapanteles_flavicoxis_bracovirus_segment_23	16987	
EU001280.1	Glyptapanteles_flavicoxis_bracovirus_segment_24	13182	
EU001281.1	Glyptapanteles_flavicoxis_bracovirus_segment_26	28551	
EU001282.1	Glyptapanteles_flavicoxis_bracovirus_segment_27	11118	
EU001283.1	Glyptapanteles_flavicoxis_bracovirus_segment_28	17937	
EU001284.1	Glyptapanteles_flavicoxis_bracovirus_segment_29	12859	
EU001261.1	Glyptapanteles_flavicoxis_bracovirus_segment_3	16057	

EU001285.1	Glyptapanteles_flavicoxis_bracovirus_segment_31	9488	
EU001262.1	Glyptapanteles_flavicoxis_bracovirus_segment_4	13357	
EU001263.1	Glyptapanteles_flavicoxis_bracovirus_segment_5	41240	
EU001264.1	Glyptapanteles_flavicoxis_bracovirus_segment_6	23555	
EU001265.1	Glyptapanteles_flavicoxis_bracovirus_segment_7	13134	
EU001266.1	Glyptapanteles_flavicoxis_bracovirus_segment_9	3759	542380
EF051505.1	Glyptapanteles_indiensis_bracovirus_segment_1	21829	
EU001243.1	Glyptapanteles_indiensis_bracovirus_segment_10	36820	
EU001244.1	Glyptapanteles_indiensis_bracovirus_segment_11	25596	
EU001245.1	Glyptapanteles_indiensis_bracovirus_segment_13	26307	
EU001246.1	Glyptapanteles_indiensis_bracovirus_segment_14	25602	
EU001247.1	Glyptapanteles_indiensis_bracovirus_segment_15	18205	
EF051506.1	Glyptapanteles_indiensis_bracovirus_segment_2	23513	
EU001250.1	Glyptapanteles_indiensis_bracovirus_segment_20	9686	
EU001251.1	Glyptapanteles_indiensis_bracovirus_segment_22	18864	
EU001252.1	Glyptapanteles_indiensis_bracovirus_segment_23	18522	
EU001253.1	Glyptapanteles_indiensis_bracovirus_segment_24	37261	
EU001254.1	Glyptapanteles_indiensis_bracovirus_segment_25	18594	
EU001255.1	Glyptapanteles_indiensis_bracovirus_segment_26	24831	
EU001256.1	Glyptapanteles_indiensis_bracovirus_segment_28	15161	
EF051507.1	Glyptapanteles_indiensis_bracovirus_segment_3	16355	
EU001258.1	Glyptapanteles_indiensis_bracovirus_segment_30	12911	
EF051508.1	Glyptapanteles_indiensis_bracovirus_segment_4	13591	
EF051509.1	Glyptapanteles_indiensis_bracovirus_segment_5	25971	
EF051510.1	Glyptapanteles_indiensis_bracovirus_segment_6	20380	
EF051511.1	Glyptapanteles_indiensis_bracovirus_segment_7	22407	
EF051512.1	Glyptapanteles_indiensis_bracovirus_segment_8	10057	
AY871265.1	Glyptapanteles_indiensis_bracovirus_segment_F	18583	461046
NC_008969.1	Hyposoter fugitivus ichnovirus	3957	
NC_008986.1	Hyposoter fugitivus ichnovirus	5047	
NC_008993.1	Hyposoter fugitivus ichnovirus	5530	
NC_008955.1	Hyposoter fugitivus ichnovirus	3548	
NC_008956.1	Hyposoter fugitivus ichnovirus	3153	
NC_008991.1	Hyposoter fugitivus ichnovirus	5328	
NC_008970.1	Hyposoter fugitivus ichnovirus	4006	
NC_008976.1	Hyposoter fugitivus ichnovirus	4188	
NC_008977.1	Hyposoter fugitivus ichnovirus	4225	
NC_008988.1	Hyposoter fugitivus ichnovirus	5070	
NC_008981.1	Hyposoter fugitivus ichnovirus	4495	
NC_008947.1	Hyposoter fugitivus ichnovirus	2755	
NC_008990.1	Hyposoter fugitivus ichnovirus	5298	
NC_008971.1	Hyposoter fugitivus ichnovirus	4126	
NC_008972.1	Hyposoter fugitivus ichnovirus	4144	

NC_008995.1	Hyposoter fugitivus ichnovirus	5848	
NC_008984.1	Hyposoter fugitivus ichnovirus	4729	
NC_008985.1	Hyposoter fugitivus ichnovirus	4914	
NC_008968.1	Hyposoter fugitivus ichnovirus	3868	
NC_008992.1	Hyposoter fugitivus ichnovirus	5353	
NC_008989.1	Hyposoter fugitivus ichnovirus	5108	
NC_008951.1	Hyposoter fugitivus ichnovirus	3590	
NC_008996.1	Hyposoter fugitivus ichnovirus	5404	
NC_008961.1	Hyposoter fugitivus ichnovirus	3293	
NC_008982.1	Hyposoter fugitivus ichnovirus	4562	
NC_008953.1	Hyposoter fugitivus ichnovirus	3935	
NC_008987.1	Hyposoter fugitivus ichnovirus	5050	
NC_008966.1	Hyposoter fugitivus ichnovirus	3638	
NC_008950.1	Hyposoter fugitivus ichnovirus	5254	
NC_008964.1	Hyposoter fugitivus ichnovirus	3573	
NC_009001.1	Hyposoter fugitivus ichnovirus	4461	
NC_008983.1	Hyposoter fugitivus ichnovirus	4720	
NC_008959.1	Hyposoter fugitivus ichnovirus	3213	
NC_009002.1	Hyposoter fugitivus ichnovirus	4808	
NC_009000.1	Hyposoter fugitivus ichnovirus	3831	
NC_008973.1	Hyposoter fugitivus ichnovirus	4180	
NC_008978.1	Hyposoter fugitivus ichnovirus	4247	
NC_008965.1	Hyposoter fugitivus ichnovirus	3591	
NC_008954.1	Hyposoter fugitivus ichnovirus	4351	
NC_008997.1	Hyposoter fugitivus ichnovirus	6654	
NC_008979.1	Hyposoter fugitivus ichnovirus	4264	
NC_008949.1	Hyposoter fugitivus ichnovirus	3510	
NC_008962.1	Hyposoter fugitivus ichnovirus	3385	
NC_008994.1	Hyposoter fugitivus ichnovirus	5579	
NC_008960.1	Hyposoter fugitivus ichnovirus	3225	
NC_008957.1	Hyposoter fugitivus ichnovirus	2851	
NC_008963.1	Hyposoter fugitivus ichnovirus	3573	
NC_008980.1	Hyposoter fugitivus ichnovirus	4367	
NC_009003.1	Hyposoter fugitivus ichnovirus	6145	
NC_008998.1	Hyposoter fugitivus ichnovirus	8851	
NC_008946.1	Hyposoter fugitivus ichnovirus	4692	
NC_008948.1	Hyposoter fugitivus ichnovirus	4442	
NC_008958.1	Hyposoter fugitivus ichnovirus	2857	
NC_008999.1	Hyposoter fugitivus ichnovirus	3402	
NC_008952.1	Hyposoter fugitivus ichnovirus	4063	
NC_008967.1	Hyposoter fugitivus ichnovirus	3841	246092
AF464931.1	Hyposoter_didymator_ichnovirus_segment_SH-A	2570	
AF364057.1	Hyposoter_didymator_ichnovirus_segment_SH-B	3536	

AF464930.1	Hyposoter_didymator_ichnovirus_segment_SH-C	3951	
AF364055.1	Hyposoter_didymator_ichnovirus_segment_SH-E	4644	
AF156933.2	Hyposoter_didymator_ichnovirus_segment_SH-F	4895	
AF479654.1	Hyposoter_didymator_ichnovirus_segment_SH-G	5633	25229
AY556383.1	Hyposoter_fugitivus_ichnovirus_segment_A1	2755	
AY935249.1	Hyposoter_fugitivus_ichnovirus_segment_B1	3153	
AY570798.1	Hyposoter_fugitivus_ichnovirus_segment_B11	3590	
AY577428.1	Hyposoter_fugitivus_ichnovirus_segment_B17	3935	
AY563518.1	Hyposoter_fugitivus_ichnovirus_segment_B7	3510	
AY597814.1	Hyposoter_fugitivus_ichnovirus_segment_B8	3548	
AY577429.1	Hyposoter_fugitivus_ichnovirus_segment_C10	4351	
AY556384.1	Hyposoter_fugitivus_ichnovirus_segment_C12	4442	
AY547319.1	Hyposoter_fugitivus_ichnovirus_segment_C16	4692	
AY570799.1	Hyposoter_fugitivus_ichnovirus_segment_C2	4063	
AY563519.1	Hyposoter_fugitivus_ichnovirus_segment_D5	5254	43293
NC_007031.1	Microplitis demolitor bracovirus	7228	
NC_007034.1	Microplitis demolitor bracovirus	10790	
NC_007029.1	Microplitis demolitor bracovirus	6307	
NC_007038.1	Microplitis demolitor bracovirus	15218	
NC_007030.2	Microplitis demolitor bracovirus	4576	
NC_007037.1	Microplitis demolitor bracovirus	15058	
NC_007033.1	Microplitis demolitor bracovirus	7823	
NC_007032.1	Microplitis demolitor bracovirus	9604	
NC_007036.1	Microplitis demolitor bracovirus	13704	
NC_007040.1	Microplitis demolitor bracovirus	15096	
NC_007035.1	Microplitis demolitor bracovirus	11238	
NC_007039.1	Microplitis demolitor bracovirus	17355	
NC_007028.1	Microplitis demolitor bracovirus	3611	
NC_007041.1	Microplitis demolitor bracovirus	13279	
NC_007044.1	Microplitis demolitor bracovirus	34334	185221
AB291138.1	Tranosema_rostrale_ichnovirus_segment_A1	4121	
AB291139.1	Tranosema_rostrale_ichnovirus_segment_A2	4508	
AB291140.1	Tranosema_rostrale_ichnovirus_segment_B1	4870	
AB291141.1	Tranosema_rostrale_ichnovirus_segment_B2	4957	
AB291164.2	Tranosema_rostrale_ichnovirus_segment_B3	5290	
AF529168.2	Tranosema_rostrale_ichnovirus_segment_B4	5410	
AY940454.1	Tranosema_rostrale_ichnovirus_segment_C1	5590	
AB291142.1	Tranosema_rostrale_ichnovirus_segment_C2	5592	
AB291143.1	Tranosema_rostrale_ichnovirus_segment_C3	5609	
AB291144.1	Tranosema_rostrale_ichnovirus_segment_C4	5740	
AB291145.1	Tranosema_rostrale_ichnovirus_segment_C5	5865	
AB291146.1	Tranosema_rostrale_ichnovirus_segment_C6	5950	
AB291149.2	Tranosema_rostrale_ichnovirus_segment_D1	7405	

AB291150.1	Tranosema_rostrale_ichnovirus_segment_D2	6839	
AB291153.1	Tranosema_rostrale_ichnovirus_segment_D5	7012	
AB291154.1	Tranosema_rostrale_ichnovirus_segment_D6	7049	
AB291155.1	Tranosema_rostrale_ichnovirus_segment_D7	7055	
AB291156.1	Tranosema_rostrale_ichnovirus_segment_E1	7408	
AF421353.1	Tranosema_rostrale_ichnovirus_segment_F1	7990	
AB291158.1	Tranosema_rostrale_ichnovirus_segment_F3	8551	
AB291160.2	Tranosema_rostrale_ichnovirus_segment_G2-1	4757	
KC176798.1	Tranosema_rostrale_ichnovirus_segment_G2-2	4903	
AB291161.2	Tranosema_rostrale_ichnovirus_segment_G3-1	4286	
KC176799.1	Tranosema_rostrale_ichnovirus_segment_G3-2	4550	
AB291163.1	Tranosema_rostrale_ichnovirus_segment_G5	10139	151446

“Polydnavirus queries” lists all the PDV sequences used in the search by accession number. “Wasp species” gives the species of wasp the PDV was sequenced from. “Length” is the base pair length of the PDV sequence. “Total” gives the total amount of base pairs sequenced for that species of wasp.

Table 3.3: Eukaryote whole genome sequences used as search databases

Abbreviation	Eukaryote Genome	vertebrate	plant	invertebrate
aaeg	Aedes_aegypti[insect_fly]_AAGE			yes
acal	Aplysia_californica[mollusc]_Aplcal2.0			yes
acar	Anolis_carolinensis[lizard]_anoCar1	yes		
acech	Acromyrmex_echinatior[insect_ant]_AEVX			yes
aceph	Atta_cephalotes[ant]_ADTU			yes
acoe	Aquilegia_coerulea[columbine]		yes	
agam	Anopheles_gambiae[insect_fly]_anoGam1			yes
agamM	Anopheles_gambiae_M[insect_fly]_ABKP			yes
agamS	Anopheles_gambiae_S[insect_fly]_ABKQ			yes
alyr	Arabidopsis_lyrata		yes	
amel	Ailuropoda_melanoleuca[panda]_ailMel1	yes		
andar	Anopheles_darlingi[insect_fly]_ADMH			yes
apflo	Apis_florea[insect_bee]_AEKZ			yes
apisu	Acyrthosiphon_pisum[insect_aphid]_Acyr20071212			yes
apmel	Apis_melifera[insect_bee]_apiMel3			yes
aque	Amphimedon_queenslandica[porifera]_ACUQ1			yes
astep	Anopheles_stephensi[fly]_November_2011			yes
atha	Arabadopsis_thalinia		yes	
bdis	Brachypodium_distachyon[grass]		yes	
bflo	Branchiostoma_floridae[lancelet]_JGI2			yes
bmor	Bombyx_mori[moth]_v2.0			yes
boimp	Bombus_impatiens[insect_bee]_AEQM			yes
boter	Bombus_terrestris[insect_bee]_AELG			yes
btau	Bos_taurus[cow]_bosTau6	yes		
camil	Callorhinchus_milii[shark]_AAVX	yes		
capit	Capitella_spI[polychaete]_Capit1			yes
ccle	Citrus_clementina[clementine]		yes	
cfam	Canis_familiaris[dog]_canFam2	yes		
cflo	Camponotus_floridanus[insect_ant]_AEAB1			yes
cgri	Cricetulus_griseus[hamster]_CriGri_1.0	yes		
chir	Capra_hircus[goat]_CHIR_1.0	yes		
chof	Choloepus_hoffmanni[sloth]_choHof1	yes		
cint	Ciona_intestinalis[ciona]_JGI2			yes
cjac	Callithrix_jacchus[marmoset]_calJac3	yes		
cpap	Carica_papaya[papaya]		yes	
cpip	Culex_pipiens[insect_fly]_cpip3			yes
cpor	Cavia_porcellus[guinea_pig]_cavPor3	yes		
cqui	Culex_quinquefasciatus[insect_fly]_AAWU			yes
crei	Chlamydomonas_reinhardtii[green_algae]		yes	
csat	Cucumis_Sativus[cucumber]		yes	
csin	Citrus_sinensis[sweet_orange]		yes	
dalbo	Drosophila_albomicans_DroAlb_1.0			yes
dana	Drosophila_ananassae_dana_r1.3			yes
danple	Danaus_plexippus[butterfly]_v3			yes
dere	Drosophila_erecta_dere_r1.3			yes

dgri	Drosophila_grimshawi_dgri_r1.3			yes
dmad	Daubentonia_madagascarensis[aye-aye_lemur]_DauMad_1.0	yes		
dmel	Drosophila_melanogaster_dm3			yes
dmoj	Drosophila_mojavensis_dmoj_r1.3			yes
dnov	Dasypus_novemcinctus[armadillo]_dasNov2	yes		
dord	Dipodomys_ordii[kangaroo_rat]_dipOrd1	yes		
dpers	Drosophila_persimilis_AAIZ			yes
dpse	Drosphila_pseudoobscura_dpse_r2.13			yes
dpul	Daphnia_pulex[crustacean]_Dappu1			yes
drer	Danio_rerio[zebrafish]_danRer6	yes		
dsech	Drosophila_sechellia_AAKO			yes
dsimu	Drosophila_simulans_AAGH			yes
dvir	Drosphila_virilis_dvir_r1.2			yes
dwil	Drosphila_willistoni_dwil_r1.3			yes
ecab	Equus_cabalus[horse]_equCab2B	yes		
eeur	Erinaceus_europaeus[hedgehog]_EriEur2.0	yes		
egra	Eucalyptus_grandis[eucalyptus]		yes	
etel	Echinops_telfairi[tenrec]_echTel1	yes		
fcat	Felis_catus[cat]_FelCat6.2	yes		
gacu	Gasterosteus_aculeatus[stickleback]_gasAcu1	yes		
ggal	Gallus_gallus[chicken]_galGal3	yes		
ggor	Gorilla_gorilla[gorilla]_gorGor3	yes		
glomor	Glossina_morsitans[insect_tsetse_fly]_tetsev1			yes
gmax	Glycine_max[soybean]		yes	
gmor	Gadus_morhua[cod_fish]_GadMor2010	yes		
hburt	Haplochromis_burtoni[cichlid_fish]_AstBur1.0	yes		
hgla	Heterocephalus_glaber[naked_mole_rat]_AFSB01	yes		
hmag	Hydra_magnipapillata[cnidarian]_ABRM.1			yes
hmelp	Heliconius_melpomene[butterfly]_ASM31383v2			yes
hrob	Helobdella_robusta[annelid]_Helro1			yes
hsal	Harpegnathos_saltator[insect_ant]_AEAC1			yes
hsap	Homo_sapiens_hg19	yes		
ixsca	Ixodes_scapularis[arachnid_tick]_IxscaW1			yes
lacha	Latimeria_chalumnae[coelacanth_fish]_LatCha1	yes		
lafr	Loxodonta_africana[elephant]_loxAfr3	yes		
lerin	Leucoraja_erinacea[skate]_LER1	yes		
lgig	Lottia_gigantea[mollusc]_Lotgi1			yes
lhumi	Linepithema_humile[ant]_Lhum_UMD_V04			yes
locu	Lepisosteus_oculatus[gar_fish]_LepOcu1	yes		
lutlon	Lutzomyia_longipalpis[insect_sand_fly]_Llon0.1			yes
mdest	Mayetiola_destructor[insect_fly]_Mdes0.5			yes
mdom	Monodelphis_domestica_monDom5	yes		
mesc	Manihot_esculenta[cassava]		yes	
meug	Macropus_eugenii[wallaby]_macEug1	yes		
mfas	Macaca_fascicularis[macaque]_CAEC01	yes		
mgal	Meleagris_gallopavo[turkey]_Turkey_2.01	yes		
mgut	Mimulus_guttatus[monkeyflower]		yes	
mhap	Meloidogyne_hapla[nematode]_ABLG1			yes

minc	Meloidogyne_incognita[nematode]_CABB1			yes
mluc	Myotis_lucifugus[microbat]_myoLuc2	yes		
mmul	Macaca_mulatta[macaque]_rheMac2	yes		
mmur	Microcebus_murinus[lemur]_micMur1	yes		
mmus	Mus_musculus[mouse]_mm9	yes		
mput	Mustela_putorius[ferret]_AGTQ	yes		
mrotu	Megachile_rotundata[bee]_MROT_1.0			yes
msex	Manduca_sexta[moth]_Msex_1.0			yes
mtru	Medicago_truncatula[legume]		yes	
mundu	Melopsittacus_undulatus[budgie_bird]_6.3	yes		
mzeb	Maylandia_zebra[cichlid_fish]_MetZeb1.0	yes		
nbric	Neolamprologus_brichardi[cichlid_fish]_NeoBri1.0	yes		
ngir	Nasonia_giraulti[insect_wasp]_ADAO			yes
nleu	Nomascus_leucogenys[gibbbon]_nomLeu1	yes		
nlon	Nasonia_longicornis[insect_wasp]_ADAP			yes
nvec	Nematostella_vectensis[cnidarian]_Nemve1			yes
nvit	Nasonia_vitripennis[insect_wasp]_Nvit2.0			yes
oana	Ornithorynchus_anatinus[platypus]_ornAna1	yes		
ocun	Oryctolagus_cuniculus[rabbit]_oryCun2	yes		
ogar	Otolemur_garnettii[galago]_otoGar2	yes		
olat	Oryzias_latipes[medaka_fish]_oryLat2	yes		
onil	Oreochromis_niloticus[tilapia]_AERX	yes		
opri	Ochotona_princeps[pika]_ochPri2	yes		
oros	Odobenus_rosmarus[walrus]_Oros_1.0	yes		
osat	Oryza_sativa[rice]		yes	
ovari	Ovis_aries[sheep]_Oar_v3.1	yes		
pcap	Procavia_capensis[hyrax]_proCap1	yes		
pham	Papio_hamadryas[baboon]_papHam1	yes		
phlpap	Phlebotomus_papatasi[sand_fly]_Ppap2.0			yes
phuco	Pediculus_humanus_corporis[insect_louse]_AAZO1			yes
pmar	Pteromyzon_marinus[lamprey]_petMar1	yes		
pmol	Python_molurus[snake]_AEQU	yes		
pnye	Pundamilia_nyererei[cichlid_fish]_PunNye1.0	yes		
pogbar	Pogonomyrmex_barbatus[insect_ant]_ADIH			yes
ppac	Pristionchus_pacificus[nematode]_ABKE1			yes
ppat	Physcomitrella_patens[moss]		yes	
pper	Prunus_persica[peach]		yes	
ppyg	Pongo_pygmaeus[orang]_ponAbe2	yes		
ptri	Populus_trichocarpa[poplar]		yes	
ptro	Pan_troglodytes[chimp]_panTro3	yes		
pvam	Pteropus_vampyrus[bat]_pteVam1	yes		
pxyl	Plutella_xylostella[moth]_DBM_FJ_V1.1			yes
rcom	Ricinus_Communis[castorbean]		yes	
rnor	Rattus_norvegicus[rat]_rn4	yes		
rpro	Rhodnius_prolixus[insect_bug]_ACPB			yes
sara	Sorex_araneus[shrew]_SorAra2.0	yes		
sbic	Sorghum_bicolor[sorghum]		yes	
sbol	Saimiri_boliviensis[squirrel_monkey]_AGCE	yes		

shar	Sarcophilus_harrisii[Tasmanian_devil]_DEVIL7	yes		
sita	Setaria_italica[millet]		yes	
skow	Saccoglossus_kowalevskii[acorn_worm]_Skow1.1			yes
smoe	Selaginella_moellendorffii[spikemoss]		yes	
solinv	Solenopsis_invicta[ant]_AEAQ			yes
spur	Strongylocentrotus_purpuratus[echinoderm]_strPru2			yes
ssal	Salmo_salar[salmon]_ASM23337v1	yes		
sscr	Sus_scrofa[pig]_susScr9	yes		
stri	Spermophilus_tridecemlineatus[squirrel]_speTri1	yes		
tadh	Trichoplax_adherens[placozoa]_Triad1			yes
tbel	Tupaia_belangeri[tupaia]_tupBel1	yes		
tcas	Tribolium_castaneum[insect_beetle]_Tcas3.0			yes
tgut	Taeniopygia_guttata_taeGut1	yes		
tman	Trichechus_manatus[manatee]_AHIN	yes		
tnig	Tetraodon_nigroviridus[tetraodon]_tetNig2	yes		
trub	Takifugu_rubripes[fugu]_fr2	yes		
tsyr	Tarsius_syrichta[tarsier]_tarSyr1	yes		
ttru	Tursiops_truncatus[dolphin]_Ttru1.4	yes		
vcar	Volvox_carteri[green_algae]		yes	
vpac	Vicugna_pacos[vicugna]_vicPac1	yes		
vvin	Vitis_vinifera[wine_grape]		yes	
xmac	Xiphophorus_maculatus[platy_fish]_4.4.2	yes		
xtro	Xenopus_tropicalis[frog]_xenTro2	yes		
zmay	Zea_mays[corn]		yes	
	Total	75	25	65

“Abbreviation” gives an abbreviation for the eukaryote database used. “Eukaryote Genome” lists the eukaryote genome databases used in full.

Albrecht U, Wyler T, Pfister-Wilhelm R, Gruber A, Stettler P (1994). Polydnavirus of the parasitic wasp Chelonus inanitus (Braconidae): characterization, genome organization and time point of replication. J Gen Virol. 75 (Pt 12):3353-63.
Aswad A, Katzourakis A. (2012) Paleovirology and virally derived immunity. Trends Ecol. Evol. 27:627–636.
Beck MH, Inman RB, Strand MR (2007) Microplitis demolitor bracovirus genome segments vary in abundance and are individually packaged in virions. Virology. 359(1):179-89.
Beck MH, Zhang S, Bitra K, Burke GR, Strand MR (2011) The encapsidated genome of Microplitis demolitor bracovirus integrates into the host Pseudoplusia includens. J Virol. 85(22):11685-96.
Belle E, Beckage NE, Rousselet J, Poirié M, Lemeunier F, Drezen JM (2002) Visualization of polydnavirus sequences in a parasitoid wasp chromosome. J Virol. 76(11):5793-6.
Bertsch C, Beuve M, Dolja VV, Wirth M, Pelsy F, et al. (2009) Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. Biol. Direct 4:21.
Bézier A, Annaheim M, Herbinière J, Wetterwald C, Gyapay G (2009) Polydnaviruses of braconid wasps derive from an ancestral nudivirus. Science 323(5916):926-30.
Bézier A, Herbinière J, Serbielle C, Lesobre J, Wincker P, et al. (2008) Bracovirus gene products are highly divergent from insect proteins. Arch Insect Biochem Physiol. 67(4):172-87.
Boto L (2014) Horizontal gene transfer in the acquisition of novel traits by metazoans. Proc Biol Sci. 281(1777):20132450
Burke GR, Strand MR (2012) Deep sequencing identifies viral and wasp genes with potential roles in replication of Microplitis demolitor Bracovirus. J Virol. 86(6):3293-306.
Chen YF, Gao F, Ye XQ, Wei SJ, Shi M et al. (2011) Deep sequencing of Cotesia vestalis bracovirus reveals the complexity of a polydnavirus genome. Virology 414(1):42-50.
Choi JY, Kwon SJ, Roh JY, Yang TJ, Yoon SH (2009) Sequence and gene organization of 24 circles from the Cotesia plutellae bracovirus genome. Arch Virol. 154(8):1313-27.
Dangerfield PC, Austin AD (1998) Biology of Mesostoa kerri (Insecta: Hymenoptera: Braconidae: Mesostoinae), an Endemic Australian Wasp that Causes Stem Galls on Banksia marginata . Australian Journal of Botany 46: 559–569.
Desjardins CA, Gundersen-Rindal DE, Hostetler JB, Tallon LJ, Fadrosh DW (2008) Comparative genomics of mutualistic viruses of Glyptapanteles parasitic wasps. Genome Biol. 9(12):R183.
Doucet D, Levasseur A, Béliveau C, Lapointe R, Stoltz D, Cusson M (2007) In vitro integration of an ichnovirus genome segment into the genomic DNA of lepidopteran cells.J Gen Virol. 88(Pt 1):105-13.
Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797
Espagne E, Douris V, Lalmanach G, Provost B, Cattolico L et al. (2005) A virus essential for insect host-parasite interactions encodes cystatins. J Virol. 79(15):9765-76.
Espagne E, Dupuy C, Huguet E, Cattolico L, Provost B (2004) Genome sequence of a polydnavirus: insights into symbiotic virus evolution. Science. 306(5694):286-9.

Flegel TW (2009) Hypothesis for heritable, anti-viral immunity in crustaceans and insects. Biol. Direct. 4:32.
Fleming JA, Summers MD (1986).Campoletis sonorensis Endoparasitic Wasps Contain Forms of C. sonorensis Virus DNA Suggestive of Integrated and Extrachromosomal Polydnavirus DNAs. J Virol. 57(2):552-62.
Fleming JG, Summers MD (1991) Polydnavirus DNA is integrated in the DNA of its parasitoid wasp host. Proc Natl Acad Sci U S A. 88(21):9770-4.
Friedman R, Hughes AL. (2006) Pattern of gene duplication in the Cotesia congregata Bracovirus. Infect Genet Evol. 6(4):315-22.
Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, et al. (2014) Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat Commun. 5:3348.
Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1350.
Graham LA, Lougheed SC, Ewart KV, Davies PL (2008). Lateral transfer of a lectin-like antifreeze protein gene in fishes. PLoS ONE. DOI: 10.1371/journal.pone.0002616
Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O (2010) "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." Systematic Biology. 59(3):307-21
Gundersen-Rindal D, Dougherty EM (2000) Evidence for integration of Glyptapanteles indiensis polydnavirus DNA into the chromosome of Lymantria dispar in vitro. Virus Res. 66(1):27-37.
Gundersen-Rindal DE, Lynn DE (2003) Polydnavirus integration in lepidopteran host cells in vitro. J Insect Physiol. 49(5):453-62
Hepat R, Kim Y (2011) Transient expression of a viral histone H4 inhibits expression of cellular and humoral immune-associated genes in Tribolium castaneum. Biochem Biophys Res Commun. 415(2):279-83.
Infante F., Hanson P. and Wharton R. A. (1995) Phytophagy in the genusMonitoriella(Hymenoptera: Braconidae) with description of new species. Annals of the Entomological Society of America 88: 406–415.
Johnson BR, Borowiec ML, Chiu JC, Lee EK, Atallah J, Ward PS (2013) Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. Curr Biol. 23(20):2058-62.
Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes (2010) PLoS Genet. 6(11):e1001191.
Koonin EV (2010) Taming of the shrewd: novel eukaryotic genes from RNA viruses. BMC Biol. 8:2.
Kuraku S, Qiu H, Meyer A (2012) Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys. Genome Biol Evol. 4(9):929-36.
Kwon B, Song S, Choi JY, Je YH, Kim Y (2010) Transient expression of specific Cotesia plutellae bracoviral segments induces prolonged larval development of the diamondback moth, Plutella xylostella. J Insect Physiol. 56(6):650-8.

Labropoulou V, Douris V, Stefanou D, Magrioti C, Swevers L, Iatrou K (2008) Endoparasitoid wasp bracovirus-mediated inhibition of hemolin function and lepidopteran host immunosuppression. Cell Microbiol. (10):2118-28.
Le NT, Asgari S, Amaya K, Tan FF, Beckage NE (2003) Persistence and expression of Cotesia congregata polydnavirus in host larvae of the tobacco hornworm, Manduca sexta. J Insect Physiol. 49(5):533-43.
Moran NA, Jarvik T (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328(5978):624-7.
Morgenstern B (2004) DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ. Nucleic Acids Research 32. W33-W36.
Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, Ziesmann J, et al. (2014) The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol. Biol. 14: 52
Pruijssers AJ, Falabella P, Eum JH, Pennacchio F, Brown MR, Strand MR (2009) Infection by a symbiotic polydnavirus induces wasting and inhibits metamorphosis of the moth Pseudoplusia includens. Journal of Experimental Biology. 212: 2998–3006.
Pruijssers AJ, Strand MR (2007) PTP-H2 and PTP-H3 from Microplitis demolitor Bracovirus localize to focal adhesions and are antiphagocytic in insect immune cells. J Virol. 81(3):1209-19
Serbielle C, Chowdhury S, Pichon S, Dupas S, Lesobre J (2008) Viral cystatin evolution and three-dimensional structure modelling: a case of directional selection acting on a viral protein involved in a host-parasitoid interaction. BMC Biol. 6:38.
Simon S, Narechania A, Desalle R, Hadrys H (2012) Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. Genome Biol Evol. 4(12):1295-309.
Stoltz DB (1990).Evidence for chromosomal transmission of polydnavirus DNA.J Gen Virol. 71 (Pt 5):1051-6.
Stoltz DB, Krell P, Summers MD, Vinson SB (1984) Polydnaviridae - a proposed family of insect viruses with segmented, double-stranded, circular DNA genomes. Intervirology. 21(1):1-4.
Stoltz DB, Vinson SB, MacKinnon EA (1976) Baculovirus-like particles in the reproductive tracts of female parasitoid wasps. Can J Microbiol. 22(7):1013-23.
Theilmann DA, Summers MD (1986) Molecular analysis of Campoletis sonorensis virus DNA in the lepidopteran host Heliothis virescens. J Gen Virol. 67 (Pt 9):1961-9.
Thézé J, Bézier A, Periquet G, Drezen JM, Herniou EA (2011) Paleozoic origin of insect large dsDNA viruses. Proc Natl Acad Sci U S A. 108(38):15931-5.
Thomas J, Schaack S, Pritham, EJ (2010) Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol. Evol. 2: 656–664.
Volkoff AN, Jouan V, Urbach S, Samain S, Bergoin M et al. (2010) Analysis of virion structural components reveals vestiges of the ancestral ichnovirus genome. PLoS Pathog. 6(5):e1000923.
Volkoff AN, Rocher J, Cérutti P, Ohresser MC, d'Aubenton-Carafa Y (2001) Persistent expression of a newly characterized Hyposoter didymator polydnavirus gene in long-term infected lepidopteran cell lines. J Gen Virol. 82(Pt 4):963-9.

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA. 110:1012–1016.
Webb BA, Strand MR, Dickey SE, Beck MH, Hilgarth RS (2006)Polydnavirus genomes reflect their dual roles as mutualists and pathogens. Virology. 347(1):160-74.
Whitfield, JB (1997) Molecular and morphological data suggest a single origin of the polydnaviruses among braconid wasps. 84(11): 502-507.
Whitfield JB (2002) Estimating the age of the polydnavirus/braconid wasp symbiosis. Proc Natl Acad Sci U S A. 99(11):7508-13.
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586-91.