Determinants of Recombination in RNA viruses

by

Charles Runckel

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry

in the

GRADUATE DIVISION

[redacted]

Acknowledgements

The poliovirus recombination project was a collaboration of the DeRisi and Andino labs. I am grateful for the advice and assistance of members of the Andino lab in my first foray into virology, in particular Adam Lauring who taught me many of the basic techniques of the field. I am also indebted to Tara Christiansen for acquiring the emulsifant that made this project feasible, which was not a common laboratory reagent and had to be sourced creatively and with great patience. Finally, I am grateful to Joe and Raul for their continued support on this project over its many technical hurdles and long, stalled periods. This work was supported by the Genentech Fellowship at UCSF.

The initial honeybee virus project was performed in collaboration with Michelle Flenniken and supported by Project Apis m., a beekeeper's organization. This work was made possible by the assistance of a very committed beekeeper, Bret Adee, and other beekeepers at Adee honey farms. The expanded hymenoptera virus project was made possible by the efforts of Brian Fisher, who collected and identified a multitude of insect samples.

I could not have endured the trials and tribulations of graduate school without the unhesitating support of my friends and family. To my parents, my heartfelt thanks.

Determinants of Recombination in RNA viruses

by Charles Runckel

Abstract

Recombination is a driving force in the evolution of RNA viruses and understanding this phenomenon is critical to improving the genetic stability of live attenuated vaccines. To identify what sequence motifs are responsible for non-random patterns of mutation within similar viruses, a synthetic poliovirus was generated with densely-spaced synonymous mutations to act as markers and a recombination map was created using a novel deep sequencing technique. This map identified multiple sequence motifs that were associated with increased or decreased local recombination, which were then engineered into a new virus to successfully modulate recombination. In contrast to frequent recombination events between strains, inter-family recombination is rare. Due to the small number of known inter-family recombinants, little is know about the determinants either for the generation of such recombinants or their viability. Deep sequencing-based viral discovery techniques were employed to discover 51 new virus species, four of which represent inter-family recombinants between the superfamily Nodavirales and the family Tetraviridae. These recombination events coincided with switches between bipartite and monopartitie genome organization. No inter-family recombination events were observed in the order Picornavirales despite frequent observations of known and novel species, suggesting a predisposition towards such recombination in the Nodavirales and

Tetraviridae. These studies demonstrate new techniques to study viral recombination at all taxonomic levels, describe new motifs associated with recombination and set the stage for viral engineering to control recombination.

Table of Contents

Main Text of the Manuscript

List of Tables

List of Figures

Thesis Chapter 1 Introduction

RNA viruses are not purely asexual organisms. While the exchange of genetic material is not required for their replication, recombination is core to their ability to rapidly shift their antigen profile and cell tropism. The influenza viruses, with their many genome segments and reshuffling, are the most famous example of recombination with large shifts in immune evasion, host species and pathogenicity occurring frequently[1]. Most RNA viruses have only one or two genome segments and for them recombination achieves the same ends but occurs in a more complicated process that requires the creation of a chimeric RNA molecule. This complexity is subject to layers of bias and constraint with most determinants of recombination inferred from simpler polymerase models, the *post hoc* analysis of natural recombination events and in general suffering from small sample sets. The subtle mechanisms of recombination must be understood as the technologies of rational vaccine design[3,4] and viral pesticides mature[5].

Mechanisms of RNA virus recombination

The primary mechanism for recombination in monopartite RNA viruses, or viruses whose entire genome is encoded on a single linear strand of RNA, is thought to be the copy-choice model of homologous recombination[6]. Under this model, two non-identical viruses infect a cell simultaneously and proceed to replicate their genomes. When a polymerase stalls and dissociates for whatever reason, the nascent strand is then free to anneal to a homologous region of a different template genome, recruit any necessary replication machinery and finish genome synthesis, resulting in a chimeric linear ssRNA molecule. This model was first validated experimentally by Kierkegaard and Baltimore (1986) using conditional and selectable poliovirus mutants to show that recombination requires replication and predominantly occurs during negative strand synthesis, thus favoring copy-choice over competing models. It is unclear whether the bias towards recombination during negative-strand synthesis is simply a function of increased acceptor molecule concentration (the positive strand outnumbers the negative strand during replication by 20:1-70:1)[7,8] or if negative strand synthesis is particularly recombinogenic. The model has since been supported with observed or experimentally induced recombination of near-identical or divergent virus strains in many RNA virus species[9–14]; nucleotide homology is the major determinant of crossover site, in-line with the copy-choice model and the need to anneal the nascent strand onto an acceptor genome.

Unlike higher organisms, which have cellular machinery to facilitate recombination, viral copy-choice relies on the failure of the replicase and the natural consequences of that failure. An instance of this phenomenon can be observed *in vitro* with DNA-to-DNA and RNA-to-DNA polymerases, where incomplete extensions and re-annealing incomplete strands results in chimeric products[15,16], or artifactual recombination. Such effects are an ever-present concern in RT-PCR-based analysis of recombination and suppressing them are the major technical challenge of this project. As these simple *in vitro* systems show, copy-choice recombination is an intrinsic phenomenon of nucleic acid replication and recombination through this mechanism should be considered the default assumption for viruses unless factors are evolved to prevent or constrain replicase dissociation. While basic, this system is not necessarily uncontrolled: coronaviruses use a chaperone protein and specific RNA sequence motifs to guide controlled recombination during infection to produce sub-genomic RNAs, analogous to splicing[4,17], and the same system has been observed to facilitate recombination between coronavirus species[18].

Non-replicative end-joining has also been proposed as a mechanism for non-homologous recombination in RNA viruses. Gyml et al[19] showed that co-transfection of truncated poliovirus 5' UTR cRNA and full-length virus cRNA with a mutated non-functional 5' UTR resulted in replication competent recombinants, suggesting that RNA molecules could be cleaved and ligated in a manner amenable to viral transfection and replication. This observation was elaborated upon with transfections of truncated 5'UTR cRNA and 5'UTR-lacking viral cRNA, again resulting in recovery of replication competent virus[20]. Both experiments suggest that host factors are mediating whatever cleavage and ligation is occurring and that this proceeds in a relatively non-specific manner. It is unclear under this model how a virus could recombine and not suffer large insertions, deletions or frame-shifts as no mechanism to facilitate homologous recombination is forthcoming. Nevertheless, most experiments examining the topology, localization and tempo of recombination are technically agnostic for either model and an assumption of one or the other is not initially necessary.

Consequences of Recombination: Evolution, Muller's Ratchet and Public Health

RNA viruses evolve rapidly by virtue of high mutation rates and very large populations[21]. The stepwise accumulation of single mutations may limit the ability of viruses to make large jumps between biologically viable variations of necessary proteins, such as capsids, and prohibit the development of massively different but functional intermediates. Recombination solves both limitations by allowing viruses to exchange viable but substantially divergent proteins or protein groups and by allowing viruses to acquire host genes, presumably through viral RNA-mRNA recombination. The former is apparent in the frequent cases of recombination of capsid and non-structural genes between strains of picornavirus[14,22], norovirus[23], and other human pathogens. Such switches can alter cell receptor use and thus tropism, exemplified within the species Human Enterovirus C (HEV-C) where capsids utilize either ICAM1 or PVR (CD155) as a primary receptor and where non-structural genes shift frequently between capsids[13]. The Providence virus of moth larvae is an extreme example of this; based on phylogenetic analysis this is a recombinant of a nodavirales-like polymerase and a tetravirus-like capsid, two completely unrelated families with different protein architectures[24]. Providence virus replicates in a variety of cell lines that do not support other tetraviruses, possibly an advantage conferred by the notoriously robust nodavirales-type polymerase. The latter benefit of recombination is a potential explanation for the presence of host genes integrated into RNA virus genomes, in particular in the family dicistroviridae (a sister family of picornaviridae infecting arthropods) that utilize a “wild-card” coding region at the 5’ end of the genome preceding more classical viral components such as proteases, a polymerase and capsid genes. This region appears to have little or no homology between clades within the family and to have extremely different functions including apoptosis inhibitors[25] and RNAi suppressors[26].

Muller’s ratchet is the tendency for the acquisition of deleterious mutations to outpace their reversion or purification in asexually reproducing organisms[27]. The high error rate of virus replication would be expected to exacerbate this effect[28]. Muller’s ratchet is a core argument for the evolution of sexual reproduction, or in general for the evolutionary desirability of some mechanism for exchanging genetic material between near identical organisms. Recombination would allow deleterious mutations to be exchanged out, allowing two viruses with different deleterious mutations to produce some progeny that possessed neither. Viruses have been observed to recombine within the course of a single infection, for example in the case of an immunodeficient child immunized with attenuated poliovirus 1 (Sabin 1) who proceeded to shed virus for 3.5 years[29]. Within that time, distinct genetic lineages developed with over 3% nt divergence from the progenitor strain. Recombinants between two pairs of lineages were identified. Such recombination would be indistinguishable without such an extreme case of divergence, but presumably such recombination could occur and play a role in population fitness even in shorter-lived infections. Recombination deficient strains are necessary to test this experimentally, but have not yet been developed.

Understanding the role of genetic exchange within viruses is also essential to public health efforts to eradicate or control circulating RNA viruses and to develop attenuated vaccines. The elimination of poliovirus in most of the world has been complicated by the recombination of vaccine strains between themselves and with related non-polio strains of enterovirus, resulting in neurovirulent strains that are then free to circulate in the population[30–32]. One reversion of poliovirus 1 vaccine resulted in an outbreak of at least 34 symptomatic cases in China over 2 years[33]. During that time, five distinct recombination events were observed between circulating enterovirus strains and between lineages of the founder virus. Surveillance of the interactions between vaccines and circulating species and future efforts to develop multivalent attenuated vaccines, where multiple serotypes of live attenuated virus are administered simultaneously, will be informed by a better understanding of the causes and constraints of viral recombination.

Mechanistic determinants of recombination in RNA viruses

The distribution of crossover sites in RNA viruses are non-random and may be influenced by three main factors: protein compatibility, nucleotide homology and the recombinogenicity of nucleotide sequence elements[2,14,34–36]. These factors often overlap, making it difficult to determine which is dominant in inducing recombination and whether such crossover hotspots are evolutionarily conserved, desirable or simply tolerated. Viruses code for multiple proteins that interact with each other and with host factors. In particular, capsids consist of complex arrangements of subunits that must be structurally stable and often self-assemble, as such capsid proteins are adapted to work with each other and core capsid proteins are observed to recombine less frequently than other components of the genome, presumably because recombinants are unlikely to be viable[14,36]. This observation does not apply to surface components and other accessory features of the capsid, such as glycoproteins in the case of enveloped viruses. Non-structural genes are more adaptable, however Jiang et al[36] showed by artificial chimeras of enteroviruses that fitness of recombinants correlated with genetic distance between donor strains, suggesting that even in these genes compatibility plays a role. This is further supported by the dearth of inter-species and inter-genus recombination events observed in viral surveillance and sequencing studies compared to frequent intra-species recombination, even among viruses that share cell tropism and thus would have the opportunity to recombine[10,37]. Protein incompatibility thus influences both what strains can viably recombine and where in the genome recombination is tolerated, favoring crossovers at protein-coding boundaries and in particular between functional protein groups.

Nucleotide homology influences copy-choice recombination by limiting areas where crossovers occur, presumably by favoring the annealing of incomplete nascent strands to similar acceptor genomes. This determinant of recombination is fundamentally different than protein incompatibility in that nucleotide homology influences where recombination events tend to occur, not where crossovers are viable. Near identical strains have been observed to recombine at 100x the frequency of closely related strains with ~15-30% nucleotide divergence[6], though protein incompatibility may influence this observation. Homology would be expected to influence the rarity of recombination in capsid genes as in most virus families these are more diverse at a nucleotide level than non-structural elements, for example among the genus Enterorhinovirus, which includes poliovirus, capsid genes share ~70% nucleotide homology while non-structural genes share ~85% nucleotide homology[13,14].

Special genomic features can result in recombination hotspots by invoking either protein compatibility or nucleotide homology. Overlapping and frame-shifted genes can highly constrain the mutation of the overlap region, which must retain its coding sequence in two different frames. Such areas are highly conserved within species and prone to recombination between strains. The utilization of subgenomic RNAs during the course of infection is also recombinogenic, presumably either because subgenomic RNAs mimic incomplete genomic replication events or because the relatively high quantity of subgenomic RNA results in a more potential acceptor genomes. Noroviruses are an extreme example of both genomic features[23]: the capsid gene follows and overlaps the polymerase gene with 21 nt of shared sequence and a 1 nt frameshift. The overlap region is perfectly conserved in almost every member of the genus and moderately conserved in related genera. The capsid gene is replicated as a subgenomic RNA to produce large quantities of capsid late in infection. Five genogroups and over one hundred genotypes of norovirus are known[38], with the overwhelming majority of recombination events occurring at this combination overlap, gene boundary, sub-genomic boundary and region of perfect conservation.

RNA secondary structure elements are strongly associated with recombination, particularly in the species Human Enterovirus C (HEV-C), which possesses 23 serotypes including the three polioviruses[39]. The CRE element (cis-regulatory element) is a 62-nt hairpin whose nucleotide sequence is almost perfectly conserved between members of this species[40,41]. This and other secondary structure elements have been exhaustively associated with crossover events[14,22,34], however it is undetermined whether the mechanism of this is due to high nucleotide homology or to some influence of the secondary structure itself, for example stalling replication or inducing a slip or stutter.

Features of genome organization are generally necessary to genome replication or infection and are thus not amenable to manipulation, making studies of such recombination hotspots difficult. Non-essential sequence elements have been proposed to influence recombination, or polymerase stalling in general. High or low AU or GC content is potentially the crudest but most global element, influencing polymerase activity either through low or high base-pairing affinity or through low complexity regions causing a stutter and stall[42,43]. Low complexity regions, including homopolymers and dimer or trimer repeats, have been shown to influence polymerase stalling or slippage[44]. The frequency and genomic topology of recombination are influenced by temperature in cell culture experiments with poliovirus[45,46], suggesting a possible role for AU or GC content in recombination. These sequence elements are not only possible candidates as sequence determinants of recombination, but are also desirable in that such elements could be easily manipulated to confirm any effect and to produce recombinogenic virus strains for further studies.

The polymerase, host factors and their interactions with the RNA genome also must be considered when examining the determinants of recombination. A reduction in stalling or dissociation of a polymerase would be expected to reduce recombination; a multitude of mechanisms have been evolved to facilitate an increase in processivity both in nature (reviewed in [47]) and in engineered systems[48]. Alternately, the association of a host or viral protein could interfere with the progress of the polymerase, making potential protein binding motifs, such as poly-pyrimidine tracts or specific binding sequences, potential candidates for recombination associated sequence elements[17]. The incorporation of mismatched bases has been shown to stall polymerases, in particular an *in vitro* study of the poliovirus polymerase showed that forced mismatches result in increased turnover of the polymerase[49]. Nucleotide composition in this study influenced the rate of dissociation directly and different nucleotides were copied with different fidelities, and could thus potentially influence recombination indirectly. Determinants of recombination may not only be purely sequence based, but involve the polymerase, the interaction of viral and host proteins in the replication complex or result from interactions between the polymerase and specific sequence elements.

Recombination in Poliovirus and other Enteroviruses

Poliovirus is a species of the family Picornaviridae, positive strand RNA viruses approximately 7.4 kb in length. Polioviruses 1, 2 and 3 are antigenically distinct serotypes of the species poliovirus defined by the use of the host poliovirus receptor (PVR, aka CD155). Poliovirus is alternately described as a clade in the species Human Enterovirus C (HEV-C), which predominately uses ICAM1 (intracellular adhesion molecule 1, aka CD54) as a receptor and is paraphyletic by amino acid phylogeny if poliovirus is regarded as a separate species. Poliovirus is thus a derived clade of HEV-C with a unique capsid that is distinct by serology, cell tropism and pathogenicity, but conserved with its siblings in replication and host interaction machinery[13]. Enteroviruses are non-enveloped viruses with a protein-linked, poly-adenylated RNA genome encased in 60 copies of four capsid subunits arranged in an icosahedron. The genomic RNA consists of a positive (sense strand) single coding region whose polyprotein is subsequently cleaved into 11 proteins, including the capsid and non-structural elements such as proteases (2A and 3C) and an RNA-dependent RNA polymerase (3D). The lack of subgenomic RNAs and overlap genes makes poliovirus an attractive model for determining what sequence specific elements drive recombination without the influence of genome organization on crossover topology.

The coding region is flanked by the highly structured 5' and 3' untranslated regions (UTRs) which facilitate cap-independent translation by recruiting host factors, as well as interacting with the viral replication complex[7,50]. Host proteins are recruited by a number of nucleic acid motifs, including secondary structures and poly-pyrimidine tracts tracts[51–54]. In addition to a multitude of functional RNA secondary structures in the UTRs, at least two structures have been the identified in the coding sequence. The cis-regulatory element (CRE) is a hairpin of 60 nt with a 26-27 nt stem located in the 2C gene that is involved in the uridylation of the 3B gene (Vpg), subsequently used as a primer for genome synthesis[40]. The RNAseL element is located in 3D gene and is composed of two interacting hairpins. The structure acts as an inhibitor to host RNAse L, a component of the antiviral response[55]. This structure is thought to be specific to the HEV-C species, while the CRE element has been identified in several picornavirus genera, however its genomic position is variable from species to species[41] and it may potentially play a role in recombination incompatibility.

RNA virus recombination was first discovered in poliovirus[56,57]. Recombination studies are facilitated by the isolation of three closely related poliovirus serotypes that are capable of recombination along with serotypes of the species Human Enterovirus C (HEV-C), the identification of selectable mutants across the genome as convenient markers, the relative ease of genetic manipulation and transfection of poliovirus infectious clones, and the development of cell lines that support robust poliovirus replication. The oral polio vaccine (OPV, or Sabin strains) consists of live, attenuated strains of each of the three serotypes of poliovirus. In rare instances, vaccination results in paralysis of vaccinees by reversion to pathogenicity in the vaccine through one of three routes: mutation, recombination between vaccine strains and recombination between a vaccine strain and circulating enteroviruses.

Reversion by mutation is beyond the scope of this work, however the two categories of recombination are informative in cataloguing the constraints on recombination, especially in the near optimal scenario for recombination where three related strains of virus are administered at high titer at the same time in immunologically naïve patients. Reversion by recombination further illustrates an escape of Muller's Ratchet in that the Sabin strains were originally generated by removing selective pressure (ie body temperature, ability to infect neuronal cells) and passaging to accumulate deleterious mutations. When selective pressure is again applied (infection of the vaccinee), the ability to revert each of the numerous attenuations by mutation is limited while their scattered locations throughout the genome allow strains to recombine out attenuated mutations and produce a fit chimera.

Recombination between Sabin strains is rapid and observed in the stool of vaccinees two days after inoculation and peaking after 14 days[58]. Estimates of viable recombinants produced between serotypes are at least 1 in 10^6 of all viruses per infectious cycle; given the large numbers of viruses involved in an infection, viable recombinants are inevitable within each vaccinee. Most viable recombinants possess only one crossover site, however recombinants that progress to neurovirulence and paralysis generally have multiple crossover sites[32,58]. Strains Sabin 2 and 3 recombine more frequently than Sabin 1, though the cause is unknown. Crossover events within the capsid are rare but not unheard of; progeny of such crossovers behave serologically like one parent or the other[58,59]. Known secondary structures in the 2C (Cre) and 3D (RNAseL element) genes are strongly associated with crossover events[34], however these sites are almost perfectly conserved between strains at a nucleotide level, making the effects of homology and potentially secondary structure indistinguishable. Despite the large number of studies examining these recombinants, the actual number of distinct isolates examined in each study is almost always less than 50 and generally less than 20[2,6,34], making attempts to implicate sequence elements like AU tracts and uncharacterized secondary structures uninformative.

Phylogenies of the capsid genes VP1 and VP4/2 indicate that the polio and enterovirus serotypes of the species HEV-C are concordant, however phylogenies of the 2C and 3D non-structural genes are discordant and imply frequent recombination both after vaccination from Sabin strains and in circulating strains[13,14]. Studies of Sabin/Enterovirus recombination focus on paralysis-associated strains, which always possess capsids of poliovirus origin with no crossover events involving non-polio enteroviruses in the capsid genes. As poliovirus uses a different receptor, this is probably a functional constraint where capsid recombinants are not viable. The 5' UTR and non-structural genes are frequently recombined relative to the capsid, presumably allowing an escape from temperature sensitivity and restricted cell tropism. Jiang et al[36] used artificial recombinants of poliovirus 1 and Coxsackie A virus 20 to show that most combinations of PV1 and CAV20 genes are viable, indicating the bias in observed crossovers is likely due to the focus on neurovirulence, which leads to the isolation of poliovirus capsids and thermocompetant non-Sabin replication genes in recombinants. Growth kinetics and cell tropism of circulating Sabin/Enterovirus recombinants are similar to wild-type poliovirus[60].

Phylogenetic analysis of the 258 non-polio enterovirus and rhinovirus serotypes suggest they recombine along similar lines with capsids remaining broadly intact and non-structural regions often recombining within a species but not between species[9,14,22,61]. Two notable exceptions involve the recombination of the 5' UTR from HEV-D onto an otherwise canonical HEV-A genome[37] and a similar recombination between HEV-A and HEV-C[62]. The 2A and 2B genes, bordering the capsid genes, are the major crossover hotspots based on inter-strain recombination, however it is unclear if this is simply associated with a functional gene boundary, if this is due to nucleotide homology (as the capsids are more divergent between strains than the non-structural genes are) or due to increased recombination at this site.

Objective 1: Sequence determinants of recombination in Poliovirus

While inter-serotype recombination in Poliovirus and Enteroviruses has been extensively studied, this has mostly been through the genetic analysis of interesting isolates and lacks the resolution and depth of sampling to determine but the most obvious sequence determinants of recombination. These efforts are also unable to extricate the impact of nucleotide homology between strains and functional incompatibility between proteins, which should have no impact in intra-strain recombination, from sequence specific elements influencing recombination, which would be expected to affect intra-strain recombination.

Poliovirus strains with selectable mutants were initially the platform of choice for exploring intra-strain recombination, including the initial study supporting the copy-choice model[6] and subsequent studies[45,46] determining that the rate of intra-strain recombination is ~100x higher than inter-strain recombination in cell culture and that the rate varies based on multiplicity of infection (MOI) and temperature from 1 to 20% of progeny genomes recombinant per infectious cycle. Synonymous mutants and RFLP were used to map recombination to 500 nt windows, which suggested that the topology of recombination was not even across the genome between nearly identical strains but lacked the resolution to identify specific elements involved[45]. Strains with multiple clustered mutations to anchor strain-specific PCR primers have also been employed, allowing the rate of recombination to be determined without the any biases incurred from selection or reversion of selectable markers, with similar estimates[63,64]. This rapid assay allowed the timing of recombination to be determined as predominately occurring at 8-10 hours post infection in cell culture, as well as confirming the impact of temperature and MOI on recombination rate in the absence of selection bias. The disadvantage of PCR-based assays is that they probe only a single area, potentially biasing a global estimate with a hot or cold spot, and do not probe a sufficiently small target to associate specific sequence elements with recombination.

RNA recombination is well studied in RNA viruses, but several important questions remain unanswered:

1. Are secondary structure elements associated with recombination because they are highly conserved at a nucleotide level or because they influence recombination directly?
2. Is the capsid region of poliovirus, and other picornaviruses, truly suppressed in recombination or is this observation an artifact of protein incompatibility or nucleotide divergence between strains?
3. What sequence specific elements make different regions of nearly identical strains recombine at different rates?
4. Is it possible to engineer a strain that is non-recombinogenic, for use in attenuated vaccines? Or highly recombinogenic to study the effects of recombination on population structure and virus evolution?

Objective 2: Patterns in Inter-Family Recombination of RNA viruses

At the other end of the spectrum, recombination between RNA viruses of different families is rarely observed. Recombinant genes in these viruses generally code for a host interaction factor or a part of the envelope, where applicable. Discordant origins of the capsid and polymerase genes are extremely rare however this may reflect a bias due to undiscovered virus families: viruses frequently harbor genes of little or no conservation to other organisms or even other viruses, even other viruses in their own family, and it may simply be the case that such a virus is a recombinant and we have simply not discovered the family it recombined *from*.

Picornaviridae appears to lack major or recent rearrangements between genera in its core genes and a lack of closely related viral families coinfecting the same hosts makes it unsuitable for an investigation of inter-family recombination. Two of Picornaviridae's sister families, however, both infect the same insect hosts: the Dicistroviridae and the Iflaviridae. These viruses are generally slightly larger than picornaviruses at 8-12 kb, but share homology at a protein level in their polymerase and protease genes, as well structural characteristics such as capsid geometry[26,65]. Iflaviridae shares a general genome organization with a single coding region and mature proteins liberated from their large precursor by protease action. As in picornaviruses, the capsid precedes the polymerase gene. Dicistroviruses use two open reading frames controlled by separate internal ribosome binding sites (IRES). The non-structural genes, including the polymerase, precede the capsid genes however the gene order within each cistron is conserved with its sister families. In addition, there are several single viruses sharing strong similarities with these Picorna-like viruses, but with sufficient amino acid divergence and differences in genome organization to warrant their own family-level classification. Picorna-like viruses in an insect host are thus an attractive model for identifying inter-family recombination events.

Of all insect hosts of picorna-like viruses, honeybees are the likely model to identify a potential inter-family recombinant. Two iflaviruses and four dicistroviruses were known to infect honeybees prior to this project, viral titers were observed to be very high (10^7 to 10^{10} viral genomes per bee) even in asymptomatic bees, and bees were reported to be frequently coinfected with multiple viruses (reviewed in[66]). The recent die-offs of managed honeybees, proposed as the Colony Collapse Disorder, fortuitously provided public interest and funding for a viral discovery expedition in honeybees and later their sister species.

Deep sequencing technology is enabling the rapid discovery of new viruses based on assembly of large datasets of short sequence reads and automated annotation of those assemblies, in this case as being viral in origin. In non-human species, it is realistic to pursue an unbiased viral discovery project and identify several new species. Further, known species are regularly sequenced and recombinants between species should be easily identifiable. This project set out to determine if:

1. Interfamily recombination involving core genes occurs in Picorna-like viruses
2. If so, what patterns can be observed in the location of the break-point and characteristics of the parent viruses

References

1. Khiabanian, H., Trifonov, V. & Rabadan, R. Reassortment patterns in Swine influenza viruses. *PLoS ONE* **4**, e7366 (2009).
2. Seligman, S. J. & Gould, E. A. Live flavivirus vaccines: reasons for caution. *Lancet* **363**, 2073–2075 (2004).
3. Yount, B., Roberts, R. S., Lindesmith, L. & Baric, R. S. Rewiring the severe acute respiratory syndrome coronavirus (SARS-CoV) transcription circuit: engineering a recombination-resistant genome. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12546–12551 (2006).
4. Valles, S. M. *et al.* A picorna-like virus from the red imported fire ant, Solenopsis invicta: initial discovery, genome sequence, and characterization. *Virology* **328**, 151–157 (2004).
5. Kirkegaard, K. & Baltimore, D. The mechanism of RNA recombination in poliovirus. *Cell* **47**, 433–443 (1986).
6. Andino, R., Rieckhof, G. E. & Baltimore, D. A functional ribonucleoprotein complex forms around the 5' end of poliovirus RNA. *Cell* **63**, 369–380 (1990).
7. Novak, J. E. & Kirkegaard, K. Improved method for detecting poliovirus negative strands used to demonstrate specificity of positive-strand encapsidation and the ratio of positive to negative strands in infected cells. *J. Virol.* **65**, 3384–3387 (1991).
8. Huang, T. *et al.* Evidence of Recombination and Genetic Diversity in Human Rhinoviruses in Children with Acute Respiratory Infection. *PLoS One* **4**, (2009).

9. Tapparel, C. *et al.* New Respiratory Enterovirus and Recombinant Rhinoviruses among Circulating Picornaviruses. *Emerg Infect Dis* **15**, 719–726 (2009).

10. Chieochansin, T., Vichiwattana, P., Korkong, S., Theamboonlers, A. & Poovorawan, Y. Molecular epidemiology, genome characterization, and recombination event of human parechovirus. *Virology* **421**, 159–166 (2011).

11. Wright, C. F. *et al.* Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot-and-Mouth Disease Virus by Using Next-Generation Genome Sequencing. *J. Virol.* **85**, 2266–2275 (2011).

12. Brown, B., Oberste, M. S., Maher, K. & Pallansch, M. A. Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region. *J. Virol.* **77**, 8973–8984 (2003).

13. Simmonds, P. & Welch, J. Frequency and dynamics of recombination within different species of human enteroviruses. *J. Virol.* **80**, 483–493 (2006).

14. Luo, G. X. & Taylor, J. Template switching by reverse transcriptase during DNA synthesis. *J Virol* **64**, 4321–4328 (1990).

15. Diehl, F. *et al.* BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nat. Methods* **3**, 551–559 (2006).

16. Wu, H.-Y. & Brian, D. A. Subgenomic messenger RNA amplification in coronaviruses. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12257–12262 (2010).

17. de Haan, C. A. M., Haijema, B. J., Masters, P. S. & Rottier, P. J. M. Manipulation of the coronavirus genome using targeted RNA recombination

with interspecies chimeric coronaviruses. *Methods Mol. Biol.* **454**, 229–236 (2008).

18. Gmyl, A. P. *et al.* Nonreplicative RNA recombination in poliovirus. *J. Virol.* **73**, 8958–8965 (1999).
19. Gmyl, A. P., Korshenko, S. A., Belousov, E. V., Khitrina, E. V. & Agol, V. I. Nonreplicative homologous RNA recombination: promiscuous joining of RNA pieces? *RNA* **9**, 1221–1231 (2003).
20. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006).
21. Oberste, M. S., Peñaranda, S. & Pallansch, M. A. RNA Recombination Plays a Major Role in Genomic Change During Circulation of Coxsackie B Viruses. *J. Virol.* **78**, 2948–2955 (2004).
22. Bull, R. A., Tanaka, M. M. & White, P. A. Norovirus Recombination. *J Gen Virol* **88**, 3347–3359 (2007).
23. Walter, C. T. *et al.* Genome organization and translation products of Providence virus: insight into a unique tetravirus. *J. Gen. Virol.* **91**, 2826–2835 (2010).
24. Mari, J., Poulos, B. T., Lightner, D. V. & Bonami, J.-R. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus Cricket paralysis-like viruses. *J. Gen. Virol.* **83**, 915–926 (2002).

25. Johnson, K. N. & Christian, P. D. The novel genome organization of the insect picorna-like virus Drosophila C virus suggests this virus belongs to a previously undescribed virus family. *J. Gen. Virol.* **79 (Pt 1)**, 191–203 (1998).

26. JSTOR: The American Naturalist, Vol. 66, No. 703 (Mar. - Apr., 1932), pp. 118-138. at <http://www.jstor.org/discover/10.2307/2456922?uid=3739560&uid=2&uid=4&uid=3739256&sid=21101130522887>

27. Chao, L. Fitness of RNA virus decreased by Muller's ratchet. *Nature* **348**, 454–455 (1990).

28. Yang, C.-F. *et al.* Intratypic recombination among lineages of type 1 vaccine-derived poliovirus emerging during chronic infection of an immunodeficient patient. *J. Virol.* **79**, 12623–12634 (2005).

29. Rousset, D. *et al.* Recombinant Vaccine–Derived Poliovirus in Madagascar. *Emerg Infect Dis* **9**, 885–887 (2003).

30. Adu, F. *et al.* Isolation of recombinant type 2 vaccine-derived poliovirus (VDPV) from a Nigerian child. *Virus Research* **127**, 17–25 (2007).

31. Guillot, S. *et al.* Natural Genetic Exchanges Between Vaccine and Wild Poliovirus Strains in Humans. *J. Virol.* **74**, 8434–8443 (2000).

32. Liu, H.-M. *et al.* Serial Recombination During Circulation of Type 1 Wild-Vaccine Recombinant Polioviruses in China. *J. Virol.* **77**, 10994–11005 (2003).

33. King, A. M. Q. Preferred Sites of Recombination in Poliovirus RNA: An Analysis of 40 Intertypic Cross-Over Sequences. *Nucl. Acids Res.* **16**, 11705–11723 (1988).

34. Dedepsidis, E., Kyriakopoulou, Z., Pliaka, V. & Markoulatos, P. Correlation between recombination junctions and RNA secondary structure elements in poliovirus Sabin strains. *Virus Genes* **41**, 181–191 (2010).

35. Tolskaya, E. A. *et al.* Studies on the recombination between RNA genomes of poliovirus: The primary structure and nonrandom distribution of crossover regions in the genomes of intertypic poliovirus recombinants. *Virology* **161**, 54–61 (1987).

36. Jiang, P. *et al.* Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9457–9462 (2007).

37. Yozwiak, N. L. *et al.* Human Enterovirus 109: A Novel Interspecies Recombinant Enterovirus Isolated from a Case of Acute Pediatric Respiratory Illness in Nicaragua. *J. Virol.* **84**, 9047–9058 (2010).

38. Kroneman, A. *et al.* An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.* **51**, 121–125 (2011).

39. Human enterovirus C. at <http://www.picornaviridae.com/enterovirus/hev-c/hev-c.htm>

40. Rieder, E., Paul, A. V., Kim, D. W., van Boom, J. H. & Wimmer, E. Genetic and Biochemical Studies of Poliovirus cis-Acting Replication Element cre in Relation to VPg Uridylylation. *J Virol* **74**, 10371–10380 (2000).

41. Yang, Y., Yi, M., Evans, D. J., Simmonds, P. & Lemon, S. M. Identification of a Conserved RNA Replication Element (cre) Within the 3Dpol-Coding Sequence of Hepatoviruses. *J. Virol.* **82**, 10118–10128 (2008).

42. Schwartz, J. J. & Quake, S. R. Single molecule measurement of the 'speed limit' of DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 20294–20299 (2009).

43. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**, 335–338 (2010).

44. Wagner, L. A., Weiss, R. B., Driscoll, R., Dunn, D. S. & Gesteland, R. F. Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res.* **18**, 3529–3535 (1990).

45. Duggal, R., Cuconati, A., Gromeier, M. & Wimmer, E. Genetic recombination of poliovirus in a cell-free system. *Proc Natl Acad Sci U S A* **94**, 13786–13791 (1997).

46. Duggal, R. & Wimmer, E. Genetic Recombination of Poliovirusin Vitroandin Vivo: Temperature-Dependent Alteration of Crossover Sites. *Virology* **258**, 30–41 (1999).

47. Bruck, I. & O'Donnell, M. The ring-type polymerase sliding clamp family. *Genome Biol.* **2**, REVIEWS3001 (2001).

48. Pavlov, A. R., Pavlova, N. V., Kozyavkin, S. A. & Slesarev, A. I. Cooperation between catalytic and DNA binding domains enhances thermostability and

supports DNA synthesis at higher temperatures by thermostable DNA polymerases. *Biochemistry* **51**, 2032–2043 (2012).

49. Freistadt, M. S., Vaccaro, J. A. & Eberle, K. E. Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase. *Virol. J.* **4**, 44 (2007).

50. Trono, D., Andino, R. & Baltimore, D. An RNA sequence of hundreds of nucleotides at the 5’ end of poliovirus RNA is involved in allowing viral protein synthesis. *J. Virol.* **62**, 2291–2299 (1988).

51. Back, S. H. *et al.* Translation of polioviral mRNA is inhibited by cleavage of polypyrimidine tract-binding proteins executed by polioviral 3C(pro). *J. Virol.* **76**, 2529–2542 (2002).

52. Hellen, C. U., Pestova, T. V., Litterst, M. & Wimmer, E. The cellular polypeptide p57 (pyrimidine tract-binding protein) binds to multiple sites in the poliovirus 5’ nontranslated region. *J. Virol.* **68**, 941–950 (1994).

53. Andino, R., Rieckhof, G. E., Achacoso, P. L. & Baltimore, D. Poliovirus RNA synthesis utilizes an RNP complex formed around the 5’-end of viral RNA. *EMBO J.* **12**, 3587–3598 (1993).

54. Gamarnik, A. V. & Andino, R. Replication of poliovirus in Xenopus oocytes requires two human factors. *EMBO J.* **15**, 5988–5998 (1996).

55. Han, J.-Q. *et al.* A Phylogenetically Conserved RNA Structure in the Poliovirus Open Reading Frame Inhibits the Antiviral Endoribonuclease RNase L. *J. Virol.* **81**, 5561–5572 (2007).

56. Hirst, G. K. Genetic Recombination with Newcastle Disease Virus, Polioviruses, and Influenza. *Cold Spring Harbor Symposia on Quantitative Biology* **27**, 303–309 (1962).

57. LEDINKO, N. Genetic recombination with poliovirus type 1. Studies of crosses between a normal horse serum-resistant mutant and several guanidine-resistant mutants of the same strain. *Virology* **20**, 107–119 (1963).

58. Cuervo, N. S. *et al.* Genomic Features of Intertypic Recombinant Sabin Poliovirus Strains Excreted by Primary Vaccinees. *J. Virol.* **75**, 5740–5751 (2001).

59. Zhang, Y. *et al.* Type 2 vaccine-derived poliovirus from patients with acute flaccid paralysis in china: current immunization strategy effectively prevented its sustained transmission. *J. Infect. Dis.* **202**, 1780–1788 (2010).

60. Riquet, F. B. *et al.* Impact of exogenous sequences on the characteristics of an epidemic type 2 recombinant vaccine-derived poliovirus. *J. Virol.* **82**, 8927–8932 (2008).

61. McIntyre, C. L., McWilliam Leitch, E. C., Savolainen-Kopra, C., Hovi, T. & Simmonds, P. Analysis of Genetic Diversity and Sites of Recombination in Human Rhinovirus Species C. *J. Virol.* **84**, 10297–10310 (2010).

62. Smura, T. *et al.* Enterovirus surveillance reveals proposed new serotypes and provides new insight into enterovirus 5’-untranslated region evolution. *J. Gen. Virol.* **88**, 2520–2526 (2007).

63. Jarvis, T. C. & Kirkegaard, K. Poliovirus RNA recombination: mechanistic studies in the absence of selection. *EMBO J* **11**, 3135–3145 (1992).

64. Tang, R. S., Barton, D. J., Flanegan, J. B. & Kirkegaard, K. Poliovirus RNA recombination in cell-free extracts. *RNA* **3**, 624–633 (1997).

65. Leat, N., Ball, B., Govan, V. & Davison, S. Analysis of the complete genome sequence of black queen-cell virus, a picorna-like virus of honey bees. *J. Gen. Virol.* **81**, 2111–2119 (2000).

66. Runckel, C. *et al.* Temporal Analysis of the Honey Bee Microbiome Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia. *PLoS ONE* **6**, e20656 (2011).

Chapter 2 Preface

This work represents my primary project during graduate school and focuses on a mechanistic understanding of the determinants of recombination in the absence of viability restrictions. The patterns of recombination observed would be expected to apply broadly at the level of intra-strain and inter-strain recombination. The recombination-deficient strain proposed in this work is now under design and construction.

This manuscript is in submission to PLoS Pathogens under the Creative Commons License.

Identification and manipulation of the molecular determinants influencing poliovirus recombination

Charles Runckel[1,2], Oscar Westesson[3], Raul Andino[4], Joseph L. DeRisi[1,2]

1 Howard Hughes Medical Institute, Bethesda, Maryland, United State of America, **2** Departments of Medicine, Biochemistry and Biophysics, and Microbiology, University of California San Francisco, San Francisco, California, United States of America, **3** The UC Berkeley - UCSF Joint Graduate Group in Bioengineering, **4** Department of Microbiology and Immunology, University of California San Francisco, San Francisco, California, United States of America

Abstract

The control and prevention of communicable disease is directly impacted by the genetic mutability of the underlying etiological agents. In the case of RNA viruses, genetic recombination may impact public health by facilitating the generation of new viral strains with altered phenotypes and by compromising the genetic stability of live attenuated vaccines. The landscape of homologous recombination within a given RNA viral genome is thought to be influenced by several factors, however a complete understanding of the genetic determinants of recombination are lacking. Here, we utilize gene synthesis and deep sequencing to create a detailed recombination map of the poliovirus 1 coding region. We identified over 57 thousand breakpoints throughout the genome and we show the majority of breakpoints to be concentrated in a small number of specific "hotspots," including those associated with known or predicted RNA secondary structures. Nucleotide base composition was also found to be associated with recombination frequency, suggesting that recombination is modulated across the genome by predictable and alterable motifs. We tested the predictive utility of the nucleotide base composition association by generating an artificial hotspot in the poliovirus genome. Our results imply that modification of these motifs could be extended to whole genome re-designs for the development of recombination-deficient, genetically stable live vaccine strains.

Author Summary

Viral recombination is critical to understanding the evolution of viral groups and impacts vaccine design, but is poorly understood. In the poliovirus vaccine, recombination is one potential mode of failure where vaccine strains recombine to produce a pathogenic product. We combine gene synthesis and deep sequencing to generate a high-resolution recombination map of poliovirus, both as a model RNA virus and a continuing threat that has yet to be eradicated. This map shows that recombination is concentrated into hotspots and suggests that predictable and alterable motifs in the RNA sequence are associated with recombination frequency. We demonstrate the utility of these observations by re-designing a poliovirus strain to recombine more frequently than normal, facilitating future studies on the role of viral recombination during infection. This result suggests that a large-scale redesign of the entire poliovirus genome to dampen recombination may be feasible, with implications for producing safer and more stable live vaccines.

Introduction

Recombination in RNA viruses is a source of genetic diversity and rapid evolutionary change and may result in the emergence of new strains by facilitating shifts in cell tropism, antigen profile and pathogenicity. The mechanism of RNA virus recombination can proceed through re-assortment of genome segments, as is the case for the Influenza A virus, or through the generation of chimeric viral genomes during replication for non-segmented viruses. This recombination is frequent in the wild with different recombinant genotypes rising to dominance and declining over a timescale of only a few years[1]. Sequencing of large numbers of viral isolates has revealed instances of intra-species recombination in many human-infecting RNA viruses with major public health implications, including norovirus[2], astrovirus[3], flavivirus[4] and at least eight species of picornavirus[5–10]. Rare inter-species recombinants, such as the enteroviruses HEV90[11] and HEV109[12], have also been described.

Viral recombination not only impacts public health by the evolution of new viral strains, but may also undermine live-attenuated vaccines by producing a fully pathogenic strain derived from the attenuated strains. The oral poliovirus vaccine (OPV) is the most famous example, where three attenuated serotypes of poliovirus are typically administered simultaneously. One week after inoculation, over a third of Sabin-2 and Sabin-3 viruses shed are recombinant[13]. In the worst case, recipients can develop vaccine-associated paralytic poliomyelitis, potentially through a recombined strain. Vaccine derived polioviruses (VDPVs) may also recombine with other circulating strains of enterovirus to create pathogenic chimeras[14]. Such events have caused outbreaks in numerous locations[15–18] and remain an ever-present consideration for newly designed live attenuated vaccines, such as the recently proposed tetravalent Dengue virus vaccine[19]. For engineered vaccine strains, a greater understanding of the underlying molecular determinants influencing recombination in RNA viruses has the potential to mitigate unwanted outcomes.

Besides its global health importance, poliovirus has also long served as a model RNA virus and in particular as a model system for the study of recombination. Viral recombination was first demonstrated in poliovirus[20] and there have been extensive studies in cell culture examining the timing and topology of recombination between different serotypes and between nearly identical construct strains[21–25]. Recombination among poliovirus strains in the wild have been readily observed and provide further opportunity for post hoc genetic analysis[17]. Together, cell culture and phylogenetic studies have indicated that recombination is not randomly distributed through the genome[25,26]. A model for the mechanism of poliovirus recombination was proposed by Kirkegaard and Baltimore (1986). Briefly, the “template-switch” model consists of premature termination of replication and association of the nascent strand with a different template genome, followed by a resumption of replication yielding a chimeric daughter genome. Consistent with this template-switch model, nucleotide homology between viral species may be a major determinant of recombination frequency[21].

Protein incompatibility has also been suggested to constrain the generation of viable recombinants. For example, recombination between the genes encoding the interlocking capsid proteins has rarely been observed[5,8,27]. However, a lower frequency of recombination in the genes encoding structural proteins may also be the result of differing levels of nucleotide homology, since capsid genes tend to possess greater sequence diversity than the non-structural genes[8].

The effects of RNA secondary structure add yet another confounding element to the analysis. Enterovirus genomes possess well-documented RNA secondary structures that have been associated with recombination breakpoints[28,29], however it is difficult to disentangle the relative contributions of nucleotide identity and the secondary structure itself with respect to recombination, especially since the sequences of these structures are highly conserved[30,31].

In efforts to overcome these issues, previous cell-culture studies have employed nearly identical strains with selectable markers, restriction-enzyme specific mutations[24,25], or unique PCR-primer annealing sites[22,23] to detect recombination events over parts of the poliovirus genome at an effective resolution of ~500-1000 nt. It has been estimated from these studies that the frequency of recombinant progeny arising from a single passage of two co-cultured strains is roughly 1-20%[21–25] and some studies have indicated that the relative recombination frequency varies in different regions of the genome, with the structural genes having a lower frequency than the non-structural genes[24,25].

In order to obtain a higher resolution map and to elucidate the sequence-specific determinants underlying poliovirus recombination, we have developed an approach utilizing a synthetic poliovirus genome engineered to contain 368 specific markers. By ultra deep sequencing, we examined the resulting viral population produced by co-infection of cells with wild-type and synthetic poliovirus genomes. The resulting high-resolution map of recombination frequencies allowed us to uncover key genomic features that both enhance or repress recombination. Based on these results, we then reengineered a portion of the genome to increase the frequency of recombination. These results identify RNA features influencing recombination and demonstrate that they may be altered with predictable outcomes. These results also suggest possible routes to attenuating recombination frequencies in synthetic vaccine strains.

Results

Construct Strain Design and Validation

Gene synthesis is inherently free from the limitations of traditional site directed mutagenesis and cloning procedures and thus enables any number of genetic modifications. Using gene synthesis, we have designed and synthesized a poliovirus genome engineered explicitly for the purpose of measuring enterovirus recombination. In total, we specified 368 synonymous marker mutations, spaced every 18nt, spanning the poliovirus 1 coding region (Fig 1A) with the intent of using Illumina deep sequencing technology to detect recombinants between wild-type and mutant poliovirus. This synthetic genome was chemically synthesized (Blue Heron, Inc.) and then tested for viability by transfection.

The initial full-length synthetic mutant virus construct was not viable when transfected into HeLaS3 cells. Therefore, we chose to arbitrarily divide the parental synthetic construct into two subconstructs (C1 and C2, Fig 1a and M&M). These constructs were viable and achieved CPE in a similar time as wild-type, within two passages post transfection. To check for additional mutations or reversion of our engineered markers, the genomes of both constructs were recovered and re-sequenced after three passages. No reversions were detected, however three additional mutations were revealed, all in the capsid region (G1872U, U2134C and A2663G) of C1 (black triangles, Fig 1a). No mutations were observed in C2. One-step growth curves of the constructs reveal robust amplification for the C2 strain compared to wild-type, while the C1 strain was consistently slower than wild-type by 5-10-fold at 4, 6 and 8 hours post infection. However, C1 ultimately produced a similar number of competent virions (~10^9 pfu/mL) as wild-type by 10 hours and beyond (fig S1a). Plaque size and morphology were similar between strains (fig S1b) and direct competition assays, where viruses were co-inoculated at equal titer and allowed to compete, showed equivalent representation after one passage. In contrast, by passage 4, the wild-type had completely out-competed both synthetic construct strains (fig S1c). These results indicate that there was a mild loss of fitness incurred by the synonymous mutations in the synthetic construct strains, yet they were viable and competitive with the wild-type strain for at least one infectious cycle. These results are consistent with previous observations[32,33].

Recombination Mapping by Deep Sequencing

Monolayers of HeLaS3 cells were coinfected with wild-type virus and each of the synthetic construct strains at a multiplicity of infection (MOI) of 10 PFUs/cell. Viral RNA was harvested after 24 hours. Illumina-compatible libraries were generated from the RNA using a standard protocol intended for RNA-seq meta-transcriptome applications[34]. A “no coinfection” control was conducted in parallel, wherein cells were infected with each virus in separate cultures, harvested at 24 hr and pooled prior to library generation. The no-coinfection control libraries provide a measure of the false-positive rate, since these samples were cultured separately and thus do not contain any recombinant virus. However, we observed high rates (1.5 breakpoints per genome) of recombinant sequences in the dataset, presumably caused by template switching during reverse transcription and/or PCR[35,36]. To circumvent the occurrence of false-recombination during library preparation, we employed a serial oil/water-emulsion droplet technique to effectively create single molecule reaction vessels for all subsequent enzymatic operations[37]. Each step of the process, beginning with reverse transcription and proceeding through fragment amplification and Illumina adaptor PCR, was conducted within separate emulsions as diagrammed in Fig 1C. After optimization of the library preparation, biological replicates of the coinfection experiment and matching no-coinfection controls were prepared and sequenced using an Illumina HiSeq2000.

Data Analysis

The error rate of Illumina sequencing and the error rate of enzymatic amplification present challenges for the interpretation of recombination mapping data. With previously reported recombination frequencies of 1-20% per genome per infectious cycle and 366 marker pairs, the mean ratio of recombinant to non-recombinant marker pairs is expected to be 1 in 10^4 to 1 in 10^3. Using published enzyme error rates, the highest fidelity commercially available enzymes possess a theoretical error rate of 1:40,000[37,38]. Illumina sequencing has published error rates per base of 0.1-1%[39]. To surmount both of these confounding sources of error, we only designated a read as evidence of a recombination breakpoint if, and only if, the candidate breakpoint was supported by a minimum of two markers on each side (Fig 1B). This requirement effectively squares the overall error rate at a cost of approximately 50% of the data set.

After quality filtering by removing reads of with any ambiguous basecalls (Ns) and trimming 10nt off of the error-prone 3' end of each read, 75 and 66 million reads (each now 90nt long) were obtained for the biological replicates, yielding a total of 110.8 and 99.0 million marker pairs mapped, disallowing any mismatches or ambiguities in alignment (Table 1). Marker pairs within 40nt of amplicon primer binding sites were also removed, in addition to those modified for RFLP analysis (see M&M). In total, 82% of marker pairs passed all quality thresholds and were used for this analysis. The signal-to-noise ratio of the coinfection to the no-coinfection control, defined as the sum of recombination frequencies observed at each marker pair in the experimental dataset divided by the no-coinfection control, ranged from 23.1:1 to 29.5:1, and averaged 26.6:1. While the biological replicates were highly correlated (R^2=0.72, Figure S2a), there was no similarity to the no-coinfection control (R^2=0.10) as expected. We observed a 2-fold variation in per marker pair recombination frequency between replicates, however the rank order of marker pairs was highly similar (Spearman ρ=0.91, Figure S2b) thus permitting identification of associations despite small differences in magnitude.

Overall Topology of Recombination

Over 57,000 individual recombination breakpoints were observed in this mapping experiment. The overall distribution of recombination breakpoints was highly consolidated with 47% of the total breakpoints observed in only 10% of the marker pairs with a mean recombination frequency of 0.14% versus 0.024% in the lower 90% of marker pairs. Breakpoint occurrences were observed between all but two of the marker pairs, with no significant difference between capsid and non-structural genes when considering mean or median recombination frequencies averaged over those regions (0.031% vs 0.042% mean crossovers per 17-nt, p>0.1). Gene boundaries have been proposed as recombination hotspots[40], however no association was observed examining either the precise site of gene boundaries, or those sites and their adjacent marker pairs. The total recombination frequencies measured were 10% and 12% for the biological replicates respectively. These frequencies are within and favoring the upper bound of previous estimates[21–23,41].

RNA Secondary Structure

RNA secondary structure has previously been identified as an enhancer of recombination and our results strongly support this association[29]. The largest peak of recombination frequency coincides with the RNAseL element ($p<10^6$), an RNA secondary structure located in the 3C gene and associated with host nuclease inhibition[31]. Recombination over this element was 3.5 times higher than the rest of the genome and included the largest recombination hotspot observed (0.44% recombinant). The CRE element, the only other well characterized RNA structure in the coding region[42], was not modified in our synthetic constructs due to concerns over viability of the mutant. We examined predicted secondary structure over the entire genome using unafold[43] and a sliding 52-nt window corresponding to each marker pair and the adjacent marker pairs. Windows with a predicted folding energy of less than -8 kcal/mol associated significantly ($p=0.0005$) with reduced recombination frequency (Fig 3), with structured regions exhibiting a higher rate of recombination.

Sequence Composition

We examined associations between sequence composition and recombination frequency. We found that GC content bias was associated with recombination: high GC marker pairs (>55% over a 17nt window) were associated with a 1.3x increase in recombination frequency ($p=0.027$) and low GC content (<40%) was associated with a 2.1x decrease ($p=0.00017$) (Fig 4a). Tracts of AU or GC nucleotides were also associated with reduced and increased recombination, respectively (Fig 4b,c). The magnitude of the effect increased with the length of the tract, from an increase of 1.3-fold for GC tetramers to 1.9-fold for hexamers (p=0.00004). AU tracts showed an inverse effect, with AU tetramers associated with 2.6-fold reduction in recombination and 3.9-fold for hexamers ($p<10^6$). Effects were observed for GC content even after accounting for AU and GC tracts, and vice versa, suggesting that the two may have independent effects or that the effects are related and involve a more complicated relationship than can be determined with this data set. The “no-coinfection” control data was subjected to the same analysis, however none of the models achieved statistical significance. We also applied these analyses to the dataset shifted one marker pair up or downstream to identify effects that may not manifest themselves locally; no significant associations were observed.

Other motifs were also examined with regard to recombination frequency. Homopolymer tracts and all dinucleotide pairs were compared with no significant associations except for AU and GC tracts; and their associated homopolymers lacked sufficient occurrences to achieve significance. No association was observed between the overall complexity of the sequence between marker pairs, as measured by LZW compression score[44,45]. To investigate whether more complex or cryptic sequence motifs were associated with recombination frequency, we employed fReduce[46] and BioProspector[47]. As these software packages are intended to identify short sequence motifs associated with transcription factor binding sites, we substituted recombination frequency as faux expression data and inter-marker regions as promoter sequences. These analyses yielded no significant predictions, however one caveat is that rare motifs or highly degenerate motifs would be unlikely to be detected in this analysis due to the small size of the genome.

An engineered hotspot

The aggregated analyses revealed both secondary structure and AU/GC content as being significantly associated with bias in poliovirus recombination. To further validate and understand the relationship between GC- and AU-rich regions and recombination, we redesigned and synthesized a portion of the poliovirus capsid region with 40 synonymous mutations over a 332-nt region with the intention of creating or extending GC tetramers or disrupting AU tracts whenever possible. For this region, the number and length of GC tracts was increased (Table 2, Fig 5a) while AU tracts 4nt or longer were eliminated. The GC content of the region was increased by 12% which resulted in a 26% increase in the overall predicted folding energy (-108.3 kcal/mol vs. -136.7 kcal/mol) (Unafold, M&M). This GC-rich construct was cloned into a wild-type poliovirus infectious clone. Synonymous mutations flanking the GC-rich region were added to both the test region construct and the wild type construct.

Coinfection experiments were performed and Illumina-compatible libraries generated for each virus pair as for the mapping experiment. This assay consisted of a single amplicon, requiring only a one-step RT-PCR in emulsion. Six coinfections each were performed with marked and unmarked wild-type virus, with the GC-rich construct and with no-coinfection controls of wild-type virus. The GC-rich construct was found to increase the rate of recombination by 7.4x over the 332-nt region (fig 5b). This result supports our finding that the presence of GC-rich regions positively influences the rate of poliovirus recombination at those regions.

Discussion

By combining synthetic poliovirus genome constructs with the large read depth conferred by Illumina sequencing, we describe a recombination map covering 82% of the poliovirus 1 coding region with over 57 thousand recombinants identified. A whole genome recombination rate of 0.10 to 0.12 crossovers per genome per infectious cycle was observed for biological replicates. This rate is within the previously published estimates of 1-20% for near identical strains in cell culture[21,22,25]. It is important to note that our recombination estimate differs in form from most previous experiments by examining the RNA of all virions produced rather than examining viable isolates.

This mapping technique is amenable to any virus for which there is an infectious clone and suitable cell line for transfection and coinfection, and could subsequently be applied to animal infections. Notably, this strategy is also possible in poorly studied viruses as no pair of selectable mutations need be identified and characterized prior to construct design. Poliovirus was used here as a well-understood model, but was also advantageous due to robust growth in cell culture. While our synthetic virus had an identical protein coding sequence to the wild-type, there are presumably undiscovered RNA secondary structure elements in the poliovirus genome that were disrupted by the markers. Three mutations in the C1 strain arose, however none of these coincided with markers and thus cannot be considered direct revertants. Whether these mutations represent compensatory changes to currently unknown secondary structure elements or rose to prominence in the population for other reasons is unknown.

The sample preparation requirements of ultra-high throughput sequencing are prone to artifactual recombination by template switching during library production. Previous studies using RT-PCR to characterize recombination frequency may have avoided this issue by using extremely low starting concentrations of template. Library preparation techniques require quantities of template orders of magnitude greater than that required for RT-PCR, necessitating the development of the emulsion-based library generation protocol described here. We note that our emulsion generation method (bead milling) produces variable vesicle sizes that require generous template dilutions, and it is likely that this could be improved by utilizing microfluidic droplet makers[48]. Alternatively, Ozsolak et al[49] have sequenced RNA molecules directly without reverse transcription, which could provide a more direct means of assaying recombination with a similar viral construct design.

Phylogenetic studies rarely observe enterovirus recombinants with crossovers in the capsid region. This observation could be the result of protein incompatibility affecting viability, low nucleotide homology preventing recombination from occurring at all, or some sequence-based factor dampening recombination. Our results do not support a significant difference in recombination rate between the capsid and the non-structural region, even including the large hotspot at the RNAseL element.

The extremes of GC content, and in particular long tracts of only AU or GC nucleotides, are also associated with bias in recombination frequency. In the simplest interpretation, incomplete RNAs terminating in GC-rich sequences could be expected to anneal to a new template genome more robustly than AU-rich sequences as a straightforward matter of thermodynamics and in line with the established copy-choice mechanism (treated in King 1988[26]). This interpretation suggests that in poliovirus, thermodynamic factors influence annealing of the nascent strand to the recipient genome to a greater extent than the initial dissociation of the donor genome. In the converse scenario, GC-rich regions would instead be less prone to fraying or dissociation from the original template and be associated with reduced recombination. The inverse symmetry of GC and AU effects further favors a simple thermodynamic model. An alternate and not exclusive model would consider RNA secondary structure to be the mechanism for recombination modulation, with GC and AU content influencing recombination indirectly by altering secondary structure stability. Our results supports earlier associations of the RNAseL element with recombination and further suggest that local secondary structure, as predicted *in silico*, also globally influences recombination rate. We also note that a recently described RNA secondary structure (Burril et al, personal communication) also corresponds to a recombination hotspot in the 3D region. These conclusions suggest that it is plausible that a global redesign of the poliovirus genome could be implemented with the intent of reducing recombination potential by disrupting secondary structure elements and modulating nucleotide use.

The frequency of AU and GC tracts is associated with the genomic GC content in Picornavirus species. Poliovirus represents a moderate case with a GC content of 46%. Other Enterovirus species, the genus Cardiovirus and most newly described or proposed genera have a similar GC content and AU/GC tract frequency (Figure 6). The genera Parechovirus, Hepatovirus and the Rhinovirus species all possess higher than average AU content, while the genera Apthovirus and Kobuvirus are GC rich. Based on the AU and GC tract associations described, we would predict that recombination rates within the GC-rich clades would be be greater than poliovirus (eg. Aichivirus, FMDV), and that the AT-rich clades (parechoviruses, hepatoviruses, rhinoviruses) would have less recombination potential than poliovirus. A major caveat of this prediction is that other factors, such as replication kinetics, the formation of replication rosettes, and differences in the viral polymerase could potentially confound such a simple relationship. No comparable recombination studies *in vitro* using nearly identical strains have been performed in these other picornaviruses, however phylogenetic studies on Human Rhinovirus species have suggested a low incidence of inter-serotype recombinants[6] compared to the closely related but more GC-balanced Enterovirus species. Limited studies of the very GC-rich Foot-and-Mouth Disease Virus have suggested that recombination between strains is frequent[5]. While circumstantial, these observations are consistent with our predictions.

The GC/AU and secondary structure motifs are straightforward to identify and can be engineered, with caveats. We modified a test region representing 4.5% of the genome to create or extend GC-rich tracts with synonymous mutations and eliminate AU tracts. The net effect of this modification was an increase in GC content (by 12%) and an increase in predicted folding energy (by 26%). This redesign underscores the difficulty of modifying coding sequence while leaving other, possibly vital, sequence factors in place. GC-content in virus sequences may be a form of adaptation to the host[50,51] and it is possible that making GC-content changes across an entire genome will render a virus non-viable or adjust its growth parameters, such as cell tropism and permissive temperature. CpG and UpA elements in RNA are underrepresented in mammalian RNA viruses[52,53] and have been associated with immune stimulation[54] and endonuclease susceptibility[55,56]. Notably, Burns et al (2009) re-engineered Poliovirus 2 to increase GC content by 15% while maintaining CpG and UpA frequency without compromising viability in cell culture, however when only 9% of the genome was saturated with UpA and CpG elements the virus was rendered almost nonviable[33].

Lessons from poliovirus vaccines clearly teach the need for a better understanding of recombination potential and the factors that influence it. Ultimately, knowledge and manipulation of these factors may assist in the development and validation of recombination deficient attenuated vaccine strains.

Methods

Virus Design and Manipulations

Six different staggers are possible when synonymously recoding a sequence every 18-nt. A python script generated all possible staggers of the pAL-WT[57] plasmid containing a modified poliovirus 1 genome with the variant placing the fewest possible mutations on tryptophans or methionines, which cannot be synonymously mutated, selected for further redesign. A poliovirus codon table was used to mutate optimal codons to the second most optimal codon, and mutate all other codons to the optimal codon. When methionines or tryptophans were encountered, the marker was shifted one codon 5' or 3'. Every ~500 nt, sites of synonymous hyper-divergence were engineered with at least 5 mismatches within 9 consecutive nucleotides to act as specific primer sites for PCR- or qPCR-based low-resolution recombination assays. In addition, 22 single synonymous mutations were made to create unique restrictions sites in the infectious clone plasmid to facilitate future modification and RFLP assays. The design was submitted to Blue Heron (OriGene) for chemical synthesis. The construct infectious clone plasmid and pAL-WT were subsequently digested with BglII and ApaI (NEB), reciprocal fragments ligated and chemically transformed into Transformax cells (Epicentre) with a 30C overnight incubation step followed by subsequent bacterial culture at 37C (GenBank accessions JX286703-4).

Infectious clone plasmid DNA was linearized with MluI (NEB) prior to T7 *in vitro* transcription. 10 ug of RNA was electroporated in a 4 mm cuvette (300V, 1000 uF, 24 Ohms) into 5 x 10^6 HeLaS3 cells as a standard reaction; up to 50 ug of RNA was attempted for the construct cRNA (adapted from [57]). Cells were maintained in 50% DMEM/50% F12 media, 10% newborn calf serum and 2 mM glutamine; immediately after transfection cells were maintained in 10% Fetal Bovine Serum instead of NCS. Virus stocks were harvested after cytopathic effect (CPE) was observed by 3 rounds of freeze/thaw at -80C and 37C. Viruses were passaged at high Multiplicity of Infection (MOI) with a 1:20 dilution of harvested media into fresh media and cells.

Plaque assays were performed on ~10^6 HeLaS3 cells in 6-well plates by washing cells with PBS, inoculation of 10-fold dilutions of virus in media, incubation for 60 minutes at 37C, an additional wash with PBS and overlay with 1% agarose and 50% DMEM/50% F12 with 1%NCS and 2 mM glutamine. One-step growth curves were performed in similar fashion with a 0.1 MOI virus inoculum and overlay in 10% NCS and media instead of an agarose formulation. Cultures were frozen at 2-hour intervals and harvested as above prior to plaque assay to determine viral load.

Coinfections were inoculated on 4x10^6 cells with two virus stocks at an MOI of 10 each, washed after 1 hour and incubated for 24 hours in 10% NCS media prior to harvest and freeze-thaw. Viral RNA was extracted by Trizol (Invitrogen)/chloroform followed by isopropanol precipitation. Virus stocks for the competition assay were passaged at 0.1 MOI for an additional four passages. Competition assay RNA was amplified by non-strain specific primers, cloned by

Topo-TA (Invitrogen) and colonies PCR amplified with strain specific primers to determine strain frequency.

Emulsion Library Construction

Emulsion conditions were adapted from [37], emulsions were created by overlaying 600 uL of 2% EM90 (Degussa) and 0.05% Triton X-100 in light mineral oil (Sigma) with 200 uL of aqueous reaction mix on ice in 2 mL round-bottom tubes with 5 mm zinc-plated steel ball bearings. Solutions were shaken in a TissueLyzer II at 15 Hz for 10 sec and 17 Hz for 10 sec. Reactions were prepared in parallel to achieve a template occupancy ratio of 1:10,000. 100 uL aliquots of emulsion were then transferred to 0.2 mL PCR tubes with a wide-bore pipette for thermocycling. For extraction, 100 uL of diethyl ether and 1 uL of 1% Cresol red (as an aqueous phase indicator dye) was added to each reaction and transferred to a 1.7 mL tube. PCR tubes were washed with an additional 100 uL of diethyl ether, which was also added to the recovery tube. Emulsions were broken by vortexing at maximum speed (3000 rpm) for 30 seconds and centrifugation at 13.2 k rpm for 1 minute followed by removal of the oil phase. This wash and breaking was repeated once with diethyl ether, once with ethyl acetate and then twice with diethyl ether. The aqueous phase was dried in a speed-vac centrifuge for 10 minutes and column purified (Zymo).

Reverse transcription and PCR reaction mixes were adapted to function under emulsion conditions: Bovine Serum Albumin (NEB) was added to a final concentration of 5% to serve as a bulking agent at the oil interface, detergentcontaining reaction buffers were avoided and enzymes were added to 5% final reaction volume. All thermocyler incubation times were extended to at least 1 minute to facilitate heat transfer. Reverse transcriptions were performed with SuperScript II (Invitrogen) with manufacturer's buffers and PCR reactions performed with Phusion (NEB) with detergent-less High Fidelity buffer. Reverse transcription was performed separately with three specific primers and each reaction was then amplified by PCR with the appropriate specific primer pair. Large PCR products were size-selected on a LabChip XT with the DNA 2k beta chip and quantitated by BioAnalyzer. Products were then subjected to transposase-based library preparation by Nextera (Epicentre) followed by emulsion PCR with Phusion. The product of this reaction was also size selected for 400-500 nt products using the LabChip DNA 750 chip, quantitated by qPCR (Kapa) and applied directly to sequencing on an Illumina HiSeq 2000 with 100 nt paired end reads.

Data analysis

Deep sequencing data was filtered for quality: all sequences with more than 1 N were removed and sequences without a perfect match of at least 55 nt to either wild-type or construct strains were discarded. Reads were trimmed from 100 nt to 90 nt due to error rates of over 1% per base in the terminal region. Custom scripts were used to generate all possible recombinant and non-recombinant wild-type and construct sequences spanning four markers (55 nt) and count perfect matches in the dataset. We identified an additional source of artifactual recombination that occurs during library preparation: both the RT and PCR steps utilize specific primer sites and at locations immediately 3’ of the primer sites (see PCR amplicons in figure 2b) extremely high levels of apparent recombination were observed in both the no-infection control and experimental datasets. These false-recombinants presumably arose due to abortive initiation. We removed sites 40 nt 3’ of the primer sites from all subsequent analyses (3% of marker pairs). The ends of the PCR amplicons exhibited low read coverage and were also removed from this analysis (2% of marker pairs). Furthermore, the short region spanning the region of overlap between the two synthetic constructs was not covered by an amplicon in this analysis (6%). A total of 22 of 366 marker pairs were designed to either create or destroy a restriction site, providing target sites for RFLP assays of recombination. These marker pairs (6%) were also excluded from analysis.

Secondary structure predictions of the poliovirus genome were determined by unafold[43] analysis of overlapping four-marker tiles (52 nt without the flanking markers). Other analysis platforms are discussed specifically in the text. The following models were considered for their presence between each marker pair: presence of a homopolymer of 4 nt or longer (4 models), presence of a dinucleotide tract of 4 nt or longer (6 models), or presence of a gene boundary (2 models). Non-binary models were considered by binning continuous scores into three similar size bins and attempting to associate the upper or lower bin vs the rest of the dataset (2 models each): GC content, LZW score, and unafold folding energy (over a 52 nt tile). In addition, two additional models were considered from the top output of the BioProspector and fReduce analysis packages for a total of twenty models; a multiple testing correction was applied to all association tests to compensate for this. Association tests were performed as Student's t-tests using the OpenEpi statistical calculator (www.openepi.com). Biological replicates were considered as discrete data points in this analysis, for a total of 580 marker pair data points.

Artificial Hotspot Experiment

A 400 nt DNA molecule was synthesized by IDT, added to a larger poliovirus PCR amplicon by fusion PCR and cloned into the prib(+)XpAlong [58] plasmid at restriction sites AatII and NheI. Triplet marker sites were added by modified primers amplifying construct or wild-type DNA, followed by similar fusion PCR and cloning steps. Viruses were generated and propagated from the infectious clones as above. The coinfection experiment was performed identically, however the library generation was executed in a single emulsion step using SuperScriptIII/Platinum Taq one-step RT-PCR mix (Invitrogen) and specific primers, otherwise as above. Amplicons were sequenced on a HiSeq 2000 diluted to a ratio of <1:10 with an unrelated insect RNA library to dampen decoupling effects; the poliovirus reads were prepared with unique DNA indices and were separated after sequencing. A lane that experienced severe over-clustering, which exacerbates the decoupling effect, was discarded from analysis.

Figure Legends

Figure 1. Experimental overview. A. Synonymous mutations were made in a synthetic poliovirus 1 genome every 18 nt. Construct and wild-type plasmid DNA was exchanged to create two partially-tagged strains, C1 and C2. Mutations observed in recovered populations are indicated by arrows in C1. Recombination (or a lack of recombination) is determined by Illumina sequencing, with recombination rate calculated as the ratio of discordant to concordant marker pairs at any given location. B. Wild-type and construct viruses co-infect a HeLa monolayer at high MOI. RNA is extracted after the infectious cycle is complete and reverse transcribed in an oil droplet emulsion. The emulsion is broken and the cDNA is amplified to ~2.6 kb PCR amplicons in another emulsion. This emulsion is broken and recovered large PCR amplicons are fragmented and adapters ligated to the sheared ends by transposase. Illumina compatible fragments are again amplified by PCR in emulsion prior to extraction, quantitation and Illumina sequencing.

Figure 2. Recombination map. A. Sequencing coverage depth per marker pair. B. Frequency of discordant vs concordant marker pairs across the genome. Positions of high recombination (over 0.001 discordant) in red. Not assayed areas marked in blue (see Methods). C. Individually infected virus strains were pooled after RNA extraction to determine false recombination from the library preparation steps, displayed at the same scale as section B.

Figure 3. Secondary structure is associated with recombination frequency. Marker pairs are binned based on the unafold calculated RNA folding energy of the marker pair and the flanking pairs (52-nt fragments). Biological replicates are shown in black and grey. Statistical associations are determined by Student's t-test after multiple testing correction. Increased folding is associated with recombination.

Figure 4. GC content affects recombination frequency. A. Marker pairs are binned by the GC content of the intervening 17 nt. B. Marker pairs are binned by the presence of tracts of consecutive A or U nucleotides of varying lengths. Bins are non-exclusive (ie marker pairs in the AU 6-mer bin are also included in the 4-mer bin). C. As B, binned by the presence of G or C tracts.

Figure 5. Creating a recombination hotspot. A. Wild-type poliovirus, with AU and GC tracts highlighted, was tagged with triple synonymous markers separated by 332 nt. Another construct was made by synonymously mutating bases to create or extend GC tracts and destroy AU tracts. An identical triple marker pair was installed on a derivative strain. B. Marked and unmarked viruses were co-infected and marker discordance calculated. The GC-rich construct, the wild-type strains and wild-type strains without co-infection as a control were assayed (n=6 each).

Figure 6. AU- and GC-tract frequency in Picornavirus species. Type strains of picornavirus species were analyzed for the presence of AU or GC 4-mers.

Figure S1. Fitness characterization of construct strains. A. One-step growth curves. Virus strains were applied to HeLa monolayers, washed and time-point samples frozen every two hours (x-axis) in triplicate (error bars). Samples were thawed and titered by plaque assay (y-axis). B. Plaques formed by construct strains were not visually different from the wild-type. C. Competition assay. Viruses were co-infected at equal titer, harvested and passaged into fresh cells four times. Viral RNA was extracted, amplified by strain-conserved primers, cloned and transformed into bacteria, and the relative quantity of each strain determined by strain specific colony PCR.

Figure S2. Comparison of biological replicates. HeLa monolayers were co-infected in parallel and proceeded through all steps of library preparation and sequencing separately. A. The recombination frequency at each marker pair is presented as a separate data point. B. Rank ordered list of marker pairs and corresponding recombination frequency.

Table 1 Mapping statistics

	Replicate 1	Replicate 2
Reads mapped	74,891,647	65,738,764
Marker pairs mapped	110,815,512	98,986,069
Genome equivalents mapped (mapped pairs / 290 pairs)	382,122	341,331
Recombination events observed	31,410	26,336
Wild-type : Construct Reads	0.78	0.82
Recombination rate (sum of observed per marker pair fraction recombinant)	0.117	0.101
Control recombination rate	0.00396	0.00437
Signal-to-noise ratio	29.5 : 1	23.1 : 1

Table 2 Features of the GC-rich construct

Feature	Wild-type	Construct
GC content	53.6%	65.7%
GC 4+ mers	11	19
GC 5+ mers	5	16
GC 6+ mers	2	7
GC 7+ mers	1	6
AU 4+ mers	5	0
AU 5+ mers	4	0
AU 6+ mers	2	0
AU 7+ mers	2	0
mfold energy (kcal/mol)	-108.3	-136.7
CpG elements	30	64
UpA elements	38	12

Figure 1

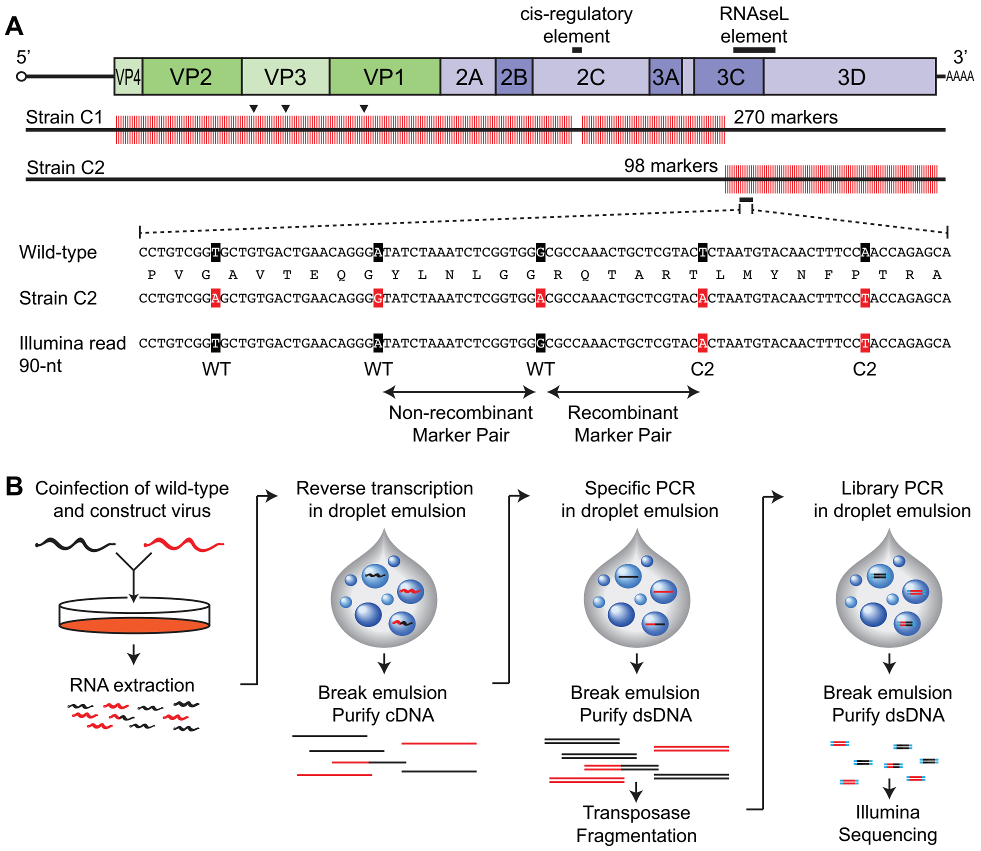
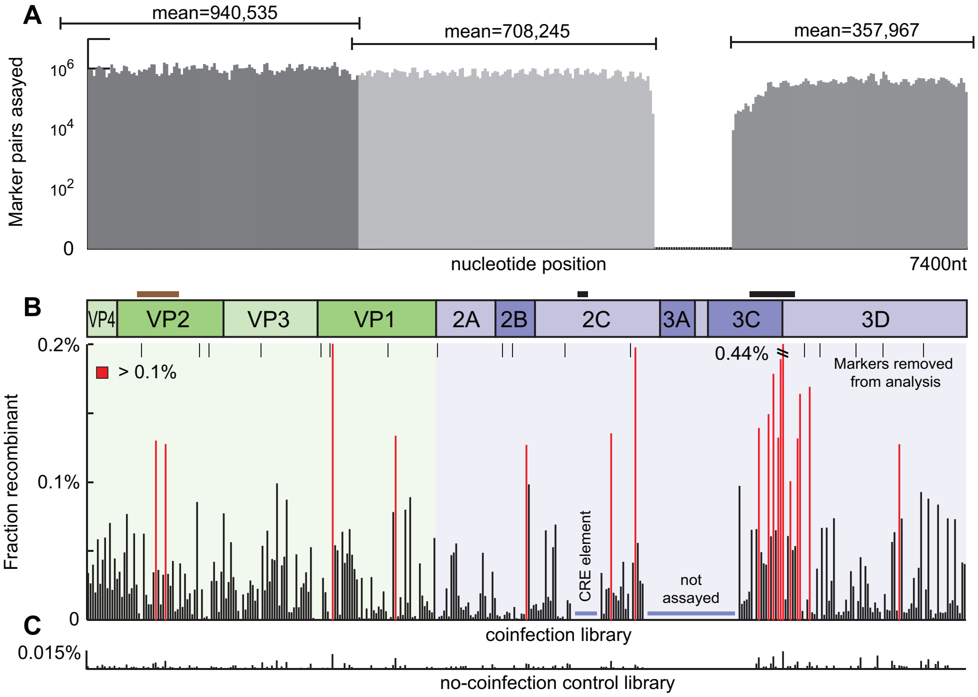

A
cis-regulatory element
RNAseL element
5'
3'
AAAA
VP4
VP2
VP3
VP1
2A
2B
2C
3A
3C
3D
Strain C1
270 markers
Strain C2
98 markers
Wild-type
CCTGTCGGTGCTGTGACTGAACAGGGATATCTAAATCTCGGTGGGCGCCAAACTGCTCGTACTCTAATGTACAACTTTCCAACCAGAGCA
P V G A V T E Q G Y L N L G G R Q T A R T L M Y N F P T R A
Strain C2
CCTGTCGGAGCTGTGACTGAACAGGGGTATCTAAATCTCGGTGGACGCCAAACTGCTCGTACACTAATGTACAACTTTCCTACCAGAGCA
Illumina read 90-nt
CCTGTCGGTGCTGTGACTGAACAGGGATATCTAAATCTCGGTGGGCGCCAAACTGCTCGTACACTAATGTACAACTTTCCTACCAGAGCA
WT
WT
WT
C2
C2
Non-recombinant Marker Pair
Recombinant Marker Pair
B
Coinfection of wild-type and construct virus
RNA extraction
Reverse transcription in droplet emulsion
Break emulsion
Purify cDNA
Specific PCR in droplet emulsion
Break emulsion
Purify dsDNA
Transposase Fragmentation
Library PCR in droplet emulsion
Break emulsion
Purify dsDNA
Illumina Sequencing

Figure 2



A
mean=940,535
mean=708,245
mean=357,967
Marker pairs asayed
10^6
10^4
10^2
0
nucleotide position
7400nt
B
VP4
VP2
VP3
VP1
2A
2B
2C
3A
3C
3D
0.2%
0.1%
0
Fraction recombinant
🟥 > 0.1%
0.44%
Markers removed from analysis
CRE element
not assayed
coinfection library
C
0.015%
no-coinfection control library

Figure 3

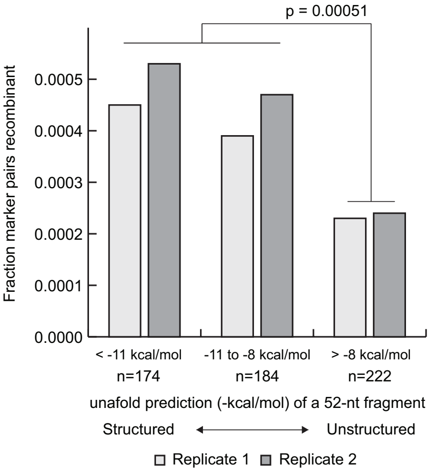

p = 0.00051
Fraction marker pairs recombinant
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
< -11 kcal/mol
n=174
-11 to -8 kcal/mol
n=184
> -8 kcal/mol
n=222
unafold prediction (-kcal/mol) of a 52-nt fragment
Structured
Unstructured
Replicate 1
Replicate 2

Figure 4

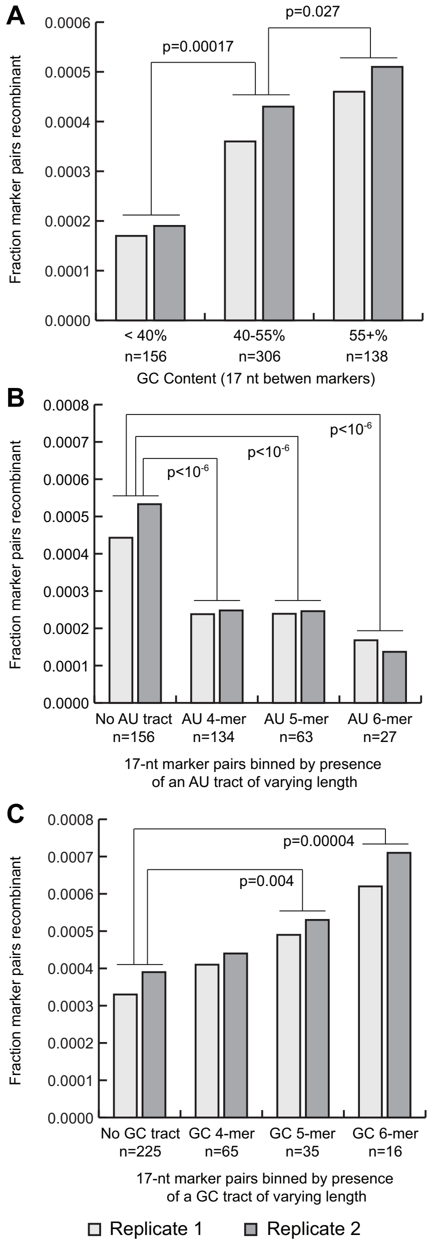

A
Fraction marker pairs recombinant
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
p=0.027
p=0.00017
< 40%
n=156
40-55%
n=306
55+%
n=138
GC Content (17 nt betwen markers)
B
Fraction marker pairs recombinant
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
$p<10^{-6}$
$p<10^{-6}$
$p<10^{-6}$
No AU tract
n=156
AU 4-mer
n=134
AU 5-mer
n=63
AU 6-mer
n=27
17-nt marker pairs binned by presence
of an AU tract of varying length
C
Fraction marker pairs recombinant
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
p=0.00004
p=0.004
No GC tract
n=225
GC 4-mer
n=65
GC 5-mer
n=35
GC 6-mer
n=16
17-nt marker pairs binned by presence
of a GC tract of varying length
□ Replicate 1
■ Replicate 2

Figure 5

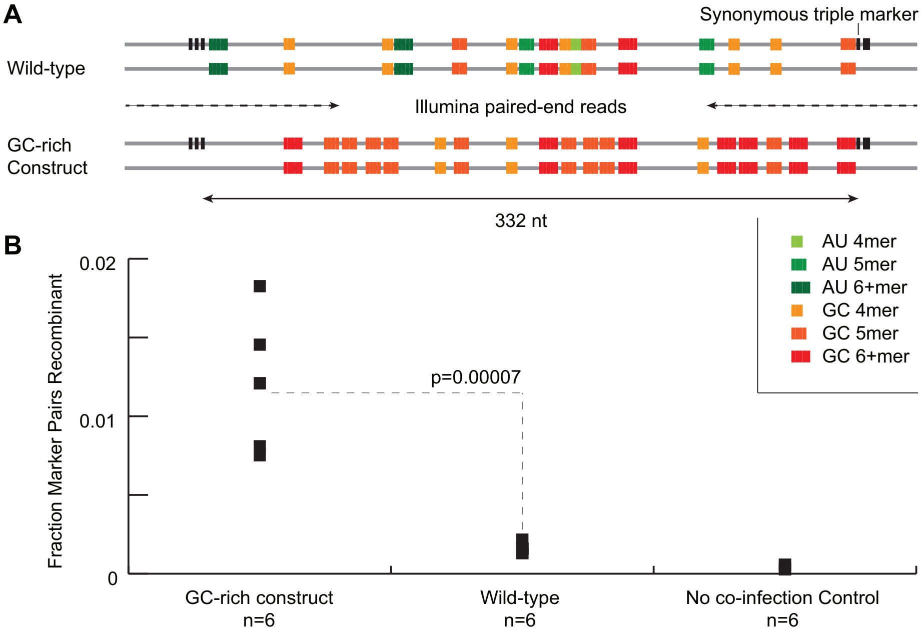
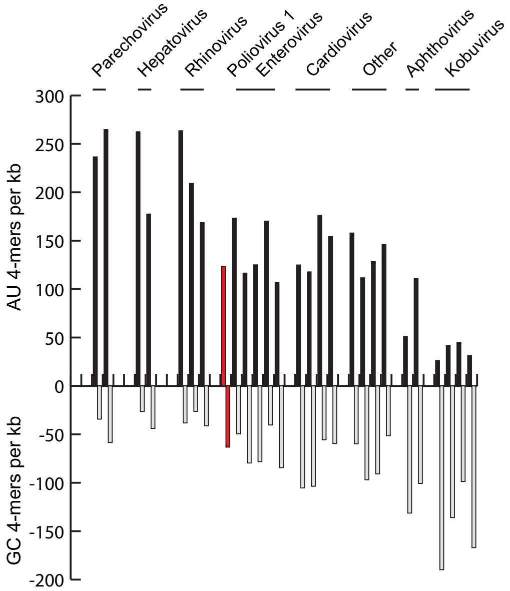
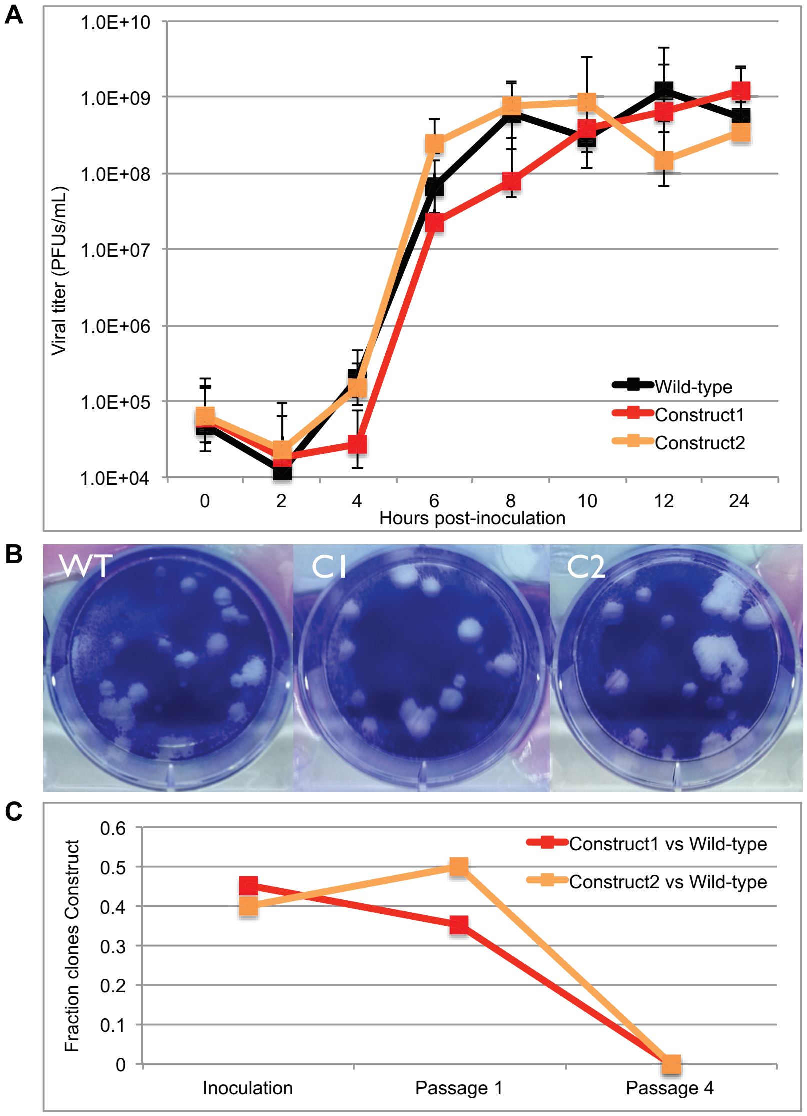

A
Synonymous triple marker
Wild-type
Illumina paired-end reads
GC-rich
Construct
332 nt
AU 4mer
AU 5mer
AU 6+mer
GC 4mer
GC 5mer
GC 6+mer
B
Fraction Marker Pairs Recombinant
0.02
0.01
0
p=0.00007
GC-rich construct
n=6
Wild-type
n=6
No co-infection Control
n=6

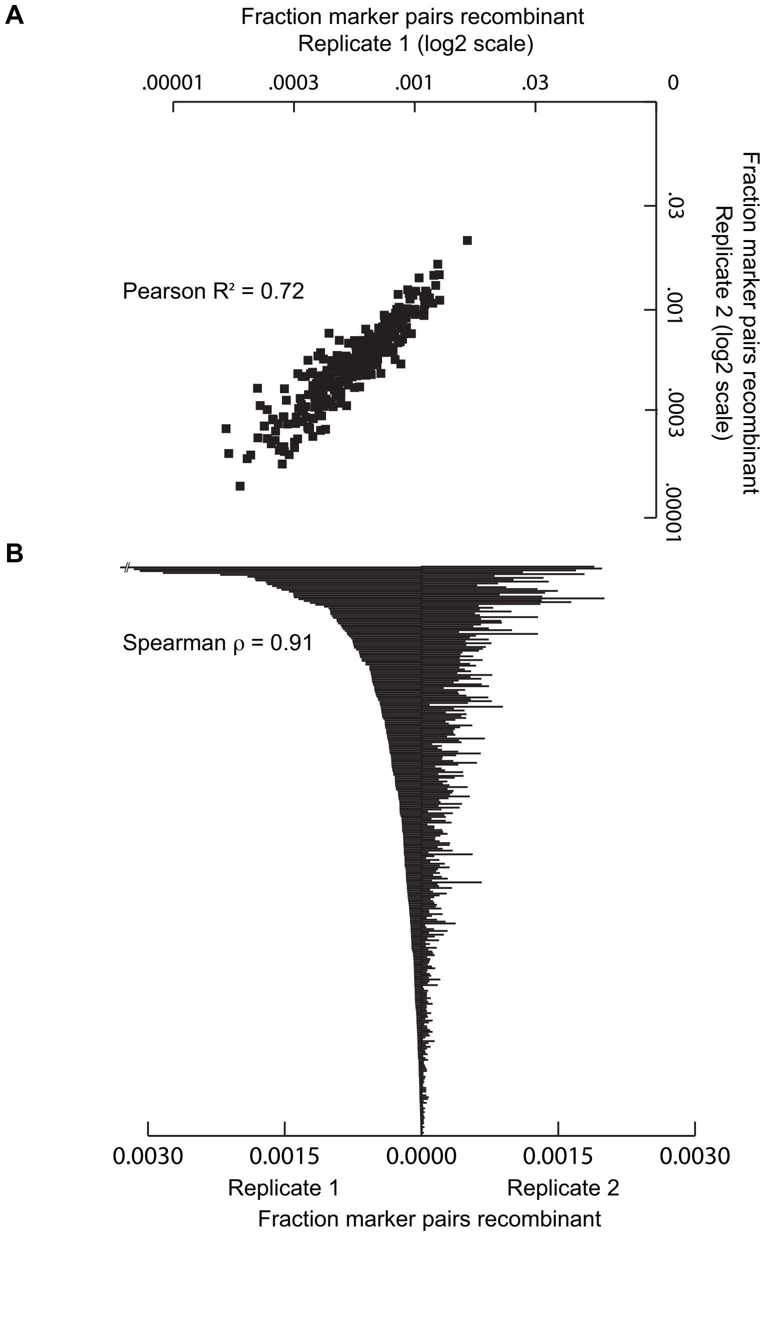
Figure 6


Parechovirus
Hepatovirus
Rhinovirus
Poliovirus 1
Enterovirus
Cardiovirus
Other
Aphthovirus
Kobuvirus
AU 4-mers per kb
GC 4-mers per kb
300
250
200
150
100
50
0
-50
-100
-150
-200

Supplemental Figure 1



Supplemental Figure 2



1. McWilliam Leitch EC, Bendig J, Cabrerizo M, Cardosa J, Hyypiä T, et al. (2009) Transmission networks and population turnover of echovirus 30. J Virol 83: 2109–2118. doi:10.1128/JVI.02109-08.

2. Bull RA, Tanaka MM, White PA (2007) Norovirus Recombination. J Gen Virol 88: 3347–3359. doi:10.1099/vir.0.83321-0.

3. Wolfaardt M, Kiulia NM, Mwenda JM, Taylor MB (2011) Evidence of a Recombinant Wild-Type Human Astrovirus Strain from a Kenyan Child with Gastroenteritis. J Clin Microbiol 49: 728–731. doi:10.1128/JCM.01093-10.

4. Twiddy SS, Holmes EC (2003) The Extent of Homologous Recombination in Members of the Genus Flavivirus. J Gen Virol 84: 429–440. doi:10.1099/vir.0.18660-0.

5. Heath L, Van Der Walt E, Varsani A, Martin DP (2006) Recombination Patterns in Aphthoviruses Mirror Those Found in Other Picornaviruses. J Virol 80: 11827–11832. doi:10.1128/JVI.01100-06.

6. McIntyre CL, McWilliam Leitch EC, Savolainen-Kopra C, Hovi T, Simmonds P (2010) Analysis of Genetic Diversity and Sites of Recombination in Human Rhinovirus Species C. J Virol 84: 10297–10310. doi:10.1128/JVI.00962-10.

7. Huang T, Wang W, Bessaud M, Ren P, Sheng J, et al. (2009) Evidence of Recombination and Genetic Diversity in Human Rhinoviruses in Children with Acute Respiratory Infection. PLoS One 4. doi:10.1371/journal.pone.0006355.

8. Simmonds P, Welch J (2006) Frequency and dynamics of recombination within different species of human enteroviruses. J Virol 80: 483–493. doi:10.1128/JVI.80.1.483-493.2006.

9. Oberste MS, Peñaranda S, Pallansch MA (2004) RNA Recombination Plays a Major Role in Genomic Change During Circulation of Coxsackie B Viruses. J Virol 78: 2948–2955. doi:10.1128/JVI.78.6.2948-2955.2004.

10. Chieochansin T, Vichiwattana P, Korkong S, Theamboonlers A, Poovorawan Y (2011) Molecular epidemiology, genome characterization, and recombination event of human parechovirus. Virology 421: 159–166. doi:10.1016/j.virol.2011.09.021.

11. Smura T, Blomqvist S, Paananen A, Vuorinen T, Sobotová Z, et al. (2007) Enterovirus surveillance reveals proposed new serotypes and provides new insight into enterovirus 5'-untranslated region evolution. J Gen Virol 88: 2520–2526. doi:10.1099/vir.0.82866-0.

12. Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, et al. (2010) Human Enterovirus 109: A Novel Interspecies Recombinant Enterovirus Isolated from a Case of Acute Pediatric Respiratory Illness in Nicaragua. J Virol 84: 9047–9058. doi:10.1128/JVI.00698-10.

13. Cuervo NS, Guillot S, Romanenkova N, Combiescu M, Aubert-Combiescu A, et al. (2001) Genomic Features of Intertypic Recombinant Sabin Poliovirus

Strains Excreted by Primary Vaccinees. J Virol 75: 5740–5751. doi:10.1128/JVI.75.13.5740-5751.2001.

14. Guillot S, Caro V, Cuervo N, Korotkova E, Combiescu M, et al. (2000) Natural Genetic Exchanges Between Vaccine and Wild Poliovirus Strains in Humans. J Virol 74: 8434–8443. doi:10.1128/JVI.74.18.8434-8443.2000.

15. Adu F, Iber J, Bukbuk D, Gumede N, Yang S-J, et al. (2007) Isolation of recombinant type 2 vaccine-derived poliovirus (VDPV) from a Nigerian child. Virus Research 127: 17–25. doi:10.1016/j.virusres.2007.03.009.

16. Rousset D, Rakoto-Andrianarivelo M, Razafindratsimandresy R, Randriamanalina B, Guillot S, et al. (2003) Recombinant Vaccine–Derived Poliovirus in Madagascar. Emerg Infect Dis 9: 885–887. doi:10.3201/eid0907.020692.

17. Yang C-F, Naguib T, Yang S-J, Nasr E, Jorba J, et al. (2003) Circulation of Endemic Type 2 Vaccine-Derived Poliovirus in Egypt from 1983 to 1993. J Virol 77: 8366–8377. doi:10.1128/JVI.77.15.8366-8377.2003.

18. Liu H-M, Zheng D-P, Zhang L-B, Oberste MS, Kew OM, et al. (2003) Serial Recombination During Circulation of Type 1 Wild-Vaccine Recombinant Polioviruses in China. J Virol 77: 10994–11005. doi:10.1128/JVI.77.20.10994-11005.2003.

19. Seligman SJ, Gould EA (2004) Live flavivirus vaccines: reasons for caution. Lancet 363: 2073–2075. doi:10.1016/S0140-6736(04)16459-3.

20. Hirst GK (1962) Genetic Recombination with Newcastle Disease Virus, Polioviruses, and Influenza. Cold Spring Harbor Symposia on Quantitative Biology 27: 303–309. doi:10.1101/SQB.1962.027.001.028.

21. Kirkegaard K, Baltimore D (1986) The mechanism of RNA recombination in poliovirus. Cell 47: 433–443.

22. Jarvis TC, Kirkegaard K (1992) Poliovirus RNA recombination: mechanistic studies in the absence of selection. EMBO J 11: 3135–3145.

23. Tang RS, Barton DJ, Flanegan JB, Kirkegaard K (1997) Poliovirus RNA recombination in cell-free extracts. RNA 3: 624–633.

24. Duggal R, Cuconati A, Gromeier M, Wimmer E (1997) Genetic Recombination of Poliovirus in a Cell-Free System. PNAS 94: 13786–13791.

25. Duggal R, Wimmer E (1999) Genetic Recombination of Poliovirusin Vitroandin Vivo: Temperature-Dependent Alteration of Crossover Sites. Virology 258: 30–41. doi:10.1006/viro.1999.9703.

26. King AMQ (1988) Preferred Sites of Recombination in Poliovirus RNA: An Analysis of 40 Intertypic Cross-Over Sequences. Nucl Acids Res 16: 11705–11723. doi:10.1093/nar/16.24.11705.

27. Oberste MS, Maher K, Pallansch MA (2004) Evidence for Frequent Recombination within Species Human Enterovirus B Based on Complete

Genomic Sequences of All Thirty-Seven Serotypes. J Virol 78: 855–867. doi:10.1128/JVI.78.2.855-867.2004.

28. Dedepsidis E, Kyriakopoulou Z, Pliaka V, Markoulatos P (2010) Correlation between recombination junctions and RNA secondary structure elements in poliovirus Sabin strains. Virus Genes 41: 181–191. doi:10.1007/s11262-010-0512-5.

29. Tolskaya EA, Romanova LI, Blinov VM, Viktorova EG, Sinyakov AN, et al. (1987) Studies on the recombination between RNA genomes of poliovirus: The primary structure and nonrandom distribution of crossover regions in the genomes of intertypic poliovirus recombinants. Virology 161: 54–61. doi:10.1016/0042-6822(87)90170-X.

30. Yang Y, Yi M, Evans DJ, Simmonds P, Lemon SM (2008) Identification of a Conserved RNA Replication Element (cre) Within the 3Dpol-Coding Sequence of Hepatoviruses. J Virol 82: 10118–10128. doi:10.1128/JVI.00787-08.

31. Han J-Q, Townsend HL, Jha BK, Paranjape JM, Silverman RH, et al. (2007) A Phylogenetically Conserved RNA Structure in the Poliovirus Open Reading Frame Inhibits the Antiviral Endoribonuclease RNase L. J Virol 81: 5561–5572. doi:10.1128/JVI.01857-06.

32. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E (2006) Reduction of the Rate of Poliovirus Protein Synthesis through Large-Scale

Codon Deoptimization Causes Attenuation of Viral Virulence by Lowering Specific Infectivity. J Virol 80: 9687–9696. doi:10.1128/JVI.00738-06.

33. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, et al. (2009) Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. J Virol 83: 9957–9969. doi:10.1128/JVI.00508-09.

34. Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, et al. (2011) Temporal Analysis of the Honey Bee Microbiome Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia. PLoS ONE 6: e20656. doi:10.1371/journal.pone.0020656.

35. Luo GX, Taylor J (1990) Template switching by reverse transcriptase during DNA synthesis. J Virol 64: 4321–4328.

36. Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res 23: 2049–2057.

37. Diehl F, Li M, He Y, Kinzler KW, Vogelstein B, et al. (2006) BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. Nat Methods 3: 551–559. doi:10.1038/nmeth898.

38. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, et al. (2008) Deciphering Human Immunodeficiency Virus Type 1 Transmission

and Early Envelope Diversification by Single-Genome Amplification and Sequencing. J Virol 82: 3952–3970. doi:10.1128/JVI.02660-07.

39. Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, et al. (2011) Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot-and-Mouth Disease Virus by Using Next-Generation Genome Sequencing. J Virol 85: 2266–2275. doi:10.1128/JVI.01396-10.

40. Simon-Loriere E, Martin DP, Weeks KM, Negroni M (2010) RNA structures facilitate recombination-mediated gene swapping in HIV-1. J Virol 84: 12675–12682. doi:10.1128/JVI.01302-10.

41. Duggal R, Cuconati A, Gromeier M, Wimmer E (1997) Genetic recombination of poliovirus in a cell-free system. Proc Natl Acad Sci U S A 94: 13786–13791.

42. Rieder E, Paul AV, Kim DW, van Boom JH, Wimmer E (2000) Genetic and Biochemical Studies of Poliovirus cis-Acting Replication Element cre in Relation to VPg Uridylylation. J Virol 74: 10371–10380.

43. Markham NR, Zuker M (2008) UNAFold. In: Keith JM, editor. Bioinformatics. Totowa, NJ: Humana Press, Vol. 453. pp. 3–31. Available:http://www.springerprotocols.com/Full/doi/10.1007/978-1-60327-429-6_1?encCode=RklCOjFfNi05MjQtNzIzMDYtMS04Nzk=&tokenString=HBFxsNLwzBNE9XqKEvi7KQ==. Accessed 15 May 2012.

44. Gusev VD, Kulichkov VA, Chupakhina OM (1993) The Lempel-Ziv complexity and local structure analysis of genomes. BioSystems 30: 183–200.

45. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 4: R9.

46. Wu RZ, Chaivorapol C, Zheng J, Li H, Liang S (2007) fREDUCE: detection of degenerate regulatory elements using correlation with expression. BMC Bioinformatics 8: 399. doi:10.1186/1471-2105-8-399.

47. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput: 127–138.

48. Zeng Y, Novak R, Shuga J, Smith MT, Mathies RA (2010) High-Performance Single Cell Genetic Analysis Using Microfluidic Emulsion Generator Arrays. Anal Chem 82: 3183–3190. doi:10.1021/ac902683t.

49. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, et al. (2009) Direct RNA sequencing. Nature 461: 814–818. doi:10.1038/nature08390.

50. Rabadan R, Levine AJ, Robins H (2006) Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J Virol 80: 11887–11891. doi:10.1128/JVI.01414-06.

51. Dunham EJ, Dugan VG, Kaser EK, Perkins SE, Brown IH, et al. (2009) Different Evolutionary Trajectories of European Avian-Like and Classical Swine H1N1 Influenza A Viruses. J Virol 83: 5485–5494. doi:10.1128/JVI.02565-08.

52. Karlin, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? Journal of Virology 68: 2889–2897.

53. Rima BK, McFerran NV (1997) Dinucleotide and Stop Codon Frequencies in Single-Stranded RNA Viruses. J Gen Virol 78: 2859–2870.

54. Sugiyama T, Gursel M, Takeshita F, Coban C, Conover J, et al. (2005) CpG RNA: Identification of Novel Single-Stranded RNA That Stimulates Human CD14+CD11c+ Monocytes. J Immunol 174: 2273–2279.

55. Duan J, Antezana M (2003) Mammalian Mutation Pressure, Synonymous Codon Choice, and mRNA Degradation. Journal of Molecular Evolution 57: 694–701. doi:10.1007/s00239-003-2519-1.

56. Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B (1989) Evolution of the Genome and the Genetic Code: Selection at the Dinucleotide Level by Methylation and Polyribonucleotide Cleavage. PNAS 86: 192–196.

57. Lauring AS, Andino R (2011) Exploring the Fitness Landscape of an RNA Virus by Using a Universal Barcode Microarray. J Virol 85: 3780–3791. doi:10.1128/JVI.02217-10.

58. Herold J, Andino R (2000) Poliovirus Requires a Precise 5′ End for Efficient Positive-Strand RNA Synthesis. J Virol 74: 6394–6400. doi:10.1128/JVI.74.14.6394-6400.2000.

Acknowledgements

This work and JDR are supported by the Howard Hughes Medical Institution. CR was supported by the Genentech Fellowship at UCSF. RA is supported by grants from the NIAID (R01 AI36178 and AI40085).

Data submissions

Synthetic poliovirus constructs were submitted to GenBank (see materials and methods).

Chapter 3 Preface

This work represents early steps in the development of virus discovery techniques using deep sequencing, and the Lake Sinai viruses described are the first novel species-level viruses discovered using Illumina short reads and no assisting platform, such as microarray or PCR screen. While the discovery of novel picorna-like viruses and particularly an inter-family recombinant was the general intent of this project, the fortuitous discovery of the Lake Sinai viruses, which represent a recombination between the families of Tetraviridae and Paranodaviridae, suggest the Nodavirus-like superfamily as a potential better target for discovering inter-family recombinants. The techniques and targets initially described here are expanded in a mature demonstration of the technology in Chapter 4.

This manuscript was published in PLoS One in 2011 under the Creative Commons License.

Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, *Nosema*, and *Crithidia*.

Charles Runckel[1]*, Michelle L. Flenniken[2*], Juan C. Engel[3], J. Graham Ruby[1], Donald Ganem[1], Raul Andino[2] and Joseph L. DeRisi[1,#]

[1]Howard Hughes Medical Institute, Departments of Medicine, Biochemistry and Biophysics, and Microbiology, University of California, San Francisco, CA 94143;
[2]Department of Microbiology and Immunology, University of California, San Francisco;
[3]Sandler Center for Drug Discovery and Department of Pathology, University of California, San Francisco. *These authors contributed equally.
# Corresponding author: joe@derisilab.ucsf.edu

Abstract

Honey bees (*Apis mellifera*) play a critical role in global food production as pollinators of numerous crops. Recently, honey bee populations in the United States, Canada, and Europe have suffered an unexplained increase in annual losses due to a phenomenon known as Colony Collapse Disorder (CCD). Epidemiological analysis of CCD is confounded by a relative dearth of bee pathogen field studies. To identify what constitutes an abnormal pathophysiological condition in a honey bee colony, it is critical to have characterized the spectrum of exogenous infectious agents in healthy hives over time. We conducted a prospective study of a large scale migratory bee keeping operation using high-frequency sampling paired with comprehensive molecular detection methods, including a custom microarray, qPCR, and ultra deep sequencing. We established seasonal incidence and abundance of known viruses, *Nosema sp*. [1], *Crithidia mellificae*, and bacteria. Ultra deep sequence analysis further identified four novel RNA viruses, two of which were the most abundant observed components of the honey bee microbiome. Our results demonstrate episodic viral incidence and distinct pathogen patterns between summer and winter time-points. Peak infection of common honey bee viruses and *Nosema* occurred in the summer, whereas levels of the trypanosomatid *Crithidia mellificae* and Lake Sinai virus 2, a novel virus, peaked in January.

Keywords: *Apis mellifera*, honey bee pathogens, pan-arthropod pathogen microarray, *Crithidia mellificae*, black queen cell virus, sacbrood virus, acute bee paralysis virus, Lake Sinai virus, aphid lethal paralysis virus strain Brookings, Big Sioux River virus, *Nosema ceranae*, *Nosema apis,* phorid, *Apocephalus borealis*, *Spiroplasma apis, Spiroplasma melliferum*

Author Summary

Agricultural crops, accounting for approximately one-third of the human diet, are pollinated by honey bees. Unfortunately, U.S. honey bee populations have suffered increased annual losses since 2006. These losses are partially attributed to colony collapse disorder (CCD), an unexplained phenomenon that is associated with increased pathogen incidence. Numerous viruses, microbes, and mites threaten honey bee colony health. In order to understand honey bee diseases which result in the death of a colony, we must first characterize the viruses and microbes associated with healthy colonies. Here we document the dynamics of honey bee pathogens from 20 commercially managed, migratory honey bee colonies over a 10-month time period. In order to comprehensively assess the pathogens in each sample, we developed a custom microarray capable of simultaneous detection of over 200 potential insect pathogens. Using this and other molecular biology techniques including quantitative PCR and ultra deep sequencing, we demonstrated episodic viral incidence, documented peak incidences of viruses and microbes and further characterized a trypansomal parasite. By thoroughly sequencing the nucleic acid in select samples, we discovered four new viruses, two of which were the most abundant honey bee viruses monitored in this 431-sample study. This result illustrates the power of this technique for viral discovery and broadens the spectrum of potential disease causing agents in honey bees. Interestingly, neither pathogen incidence nor abundance increased after long distance transport inherent to migratory beekeeping operations. This work provides a basis for future epidemiologic studies aimed at determining the causes of CCD.

Introduction

Western honey bee (*Apis mellifera*) are highly social insects that live in colonies of ~30,000 individuals [2,3]. Honey bees are essential pollinators of agriculturally important crops including apples, almonds, alfalfa, and citrus. Current agricultural practices, such as large-scale monocultures, demand a seasonal abundance of honey bees in geographic locations incapable of maintaining sufficient pollinator populations year-round. Migratory beekeeping operations fulfill this need. For example, each February in the Central Valley of California 1.3 million honey bee colonies (~ 50% of the U.S. population) are required for almond pollination [4,5,6]. Pollination of this and other U.S. crops is valued at ~$15 billion annually [6].

There are numerous threats facing honey bee populations and the recent losses of honey bee colonies in the United States, Canada, and Europe is alarming. In the U.S., annual honey bee colony losses increased from 17-20% to 32% during the winter of 2006/07 with some operations losing 90% of their hives [7]. Average annual losses have remained high, averaging 32.6% from 2007-2010 [7,8,9]. One factor contributing to increased losses is Colony Collapse Disorder (CCD), an unexplained loss of honey bee colonies fitting a defined set of criteria [10,11]. While factors such as pesticide exposure, transportation stress, genetic diversity, and nutrition affect colony health, the most significant CCD-associated variable characterized to date is increased pathogen incidence [11]. Although greater pathogen incidence correlates with CCD, the cause is unknown in part due to insufficient knowledge of the pathogenic and commensal organisms associated with honey bees [11,12].

Parasitic threats to honey bee colonies include viruses, *Nosema*, bacteria, and *Crithidia*. The majority of honey bee infecting viruses are positive-sense single-stranded RNA viruses of the *Picornavirales* order. They included acute bee paralysis virus (ABPV) [13], black queen cell virus (BQCV) [14], Israeli acute bee paralysis virus (IAPV) [15], Kashmir bee virus (KBV) [16], deformed wing virus (DWV) [17], sacbrood virus (SBV) [18], and chronic bee paralysis virus (CBPV) [19] (reviewed in Chen and Siede, 2007 [20]). Several DNA viruses that infect honey bees have also been described [20]. Viral infections in bees can remain asymptomatic, or cause deformities, paralysis and/or death [20,21]. Symptoms associated with specific viruses include wing deformities (DWV), hairless, dark, shiny bees (CBPV), swollen yellow larva and/or dark-brown larva carcasses in the cells of worker-bees (SBV) or queen-bees (BQCV), however accurate diagnosis requires molecular biology techniques as asymptomatic bees frequently test positive for one or more viruses [20,22]. In addition to viral infections, honey bees are also readily parasitized by the microsporidia *Nosema* [1,20]. Historically U.S. honey bees were predominantly infected by *Nosema apis*, but recently *Nosema ceranae* infections dominate [1,23]. The effects of *Nosema* infection on individual bee and colony health are unclear [1,24]. Some reports suggest infections decrease longevity and may lead to collapse [25,26,27], but since *Nosema* is widespread and often detected in healthy colonies its role in colony health requires further investigation [11,24,28]. Another fungal pathogen *Ascophaera apis,* the causative agent of Chalkbrood disease, kills infected larvae, but does not typically cause colony loss [29,30]. Bacterial pathogens of honey bees include *Paenibacillus larvae* and *Melissococcus pluton,* the causative agents of American and European Foulbrood disease [31,32,33,34]. In addition to microbial infections, mite infestation (*Ararapis woodi, Tropilaelaps sp.,* and *Varroa destructor)* also weakens and kills honey bee colonies [35,36]. Introduction of *V. destructor* mites, which feed on the hemolymph of developing honey bees and transmit viruses (DWV, KBV, IAPV), in the late 1980s was devastating to the U.S. honey bee population [37,38,39]. Notably, the restricted genetic diversity of the U.S. honey bee population may make it particularly susceptible to catastrophic and episodic losses [40,41].

To gain a more complete understanding of the spectrum of infectious agents and potential threats found in commercially managed migratory honey bee colonies, we conducted a 10-month prospective investigation. Our broad-scale analysis incorporated a suite of molecular tools (custom microarray, polymerase chain reaction (PCR), quantitative PCR (qPCR) and deep sequencing) enabling rapid detection of the presence (or absence) of all previously identified honey bee pathogens as well as facilitating the detection of novel pathogens. This study provides a comprehensive temporal characterization of honey bee pathogens and offers a baseline for understanding current and emerging threats to this critical component of U.S. agriculture.

Results

Following devastating losses suffered by U.S. commercial beekeeping operations in 2006-2007, we initiated a prospective study monitoring a typically managed, large-scale (>70,000 hives), migratory commercial beekeeping operation over 10-months. Honey bees from 20 colonies were consistently sampled beginning with the introduction of a new queen in April 2009 (Mississippi (MS), through transport to summer foraging grounds in South Dakota (SD), and transfer to California (CA) for almond pollination (Figure 1). During our study, these colonies were exposed to antimicrobial treatments, transportation stress, different pollen and nectar sources, and three distinct geographic locations: MS, SD, and CA, (U.S.A.).

A molecular analysis pipeline consisting of custom microarray, polymerase chain reaction (PCR), quantitative PCR (qPCR) and ultra deep sequencing was employed to characterize the honey bee microbial flora. Pathogen screening was performed using the “Arthropod Pathogen Microarray” built on the same design principles used for human pathogen microarray screening [42,43]. The array’s design couples highly-conserved nucleic acid targets with hybridization-based detection to identify previously uncharacterized organisms [44,45,46,47,48,49]. Specifically, the APM was designed to detect virtually all known microbial parasites of insects. Endpoint PCR provided sensitive detection while qPCR documented abundance of select pathogens. Ultra deep sequencing facilitated the discovery of novel and highly divergent microbes. Together the results from our monitoring study provide insight regarding the incidence of virus and microbe infections in honey bee colonies.

Arthropod Pathogen Microarray design and validation

The Arthropod Pathogen Microarray (APM) is a custom DNA microarray capable of detecting over 200 arthropod associated viruses, microbes, and metazoans. This DNA microarray includes oligonucleotides representing every arthropod-infecting virus with published nucleic acid sequence in the International Committee on Taxonomy of Viruses database as of November 2008 [50,51]. Design principles used for APM oligonucleotides (70-mers) were based on previous pan-viral microarrays using ArrayOligoSelector (AOS)[52]. In addition, non-viral pathogens including, *Nosema* (microsporidia), *Crithidia* (trypanosomatid), *Varroa* (mite), *Tropilaelaps* (mite) and *Acarapis* (tracheal mite) as well as *Paenibacillus larvae* and *Melissococcus pluton* bacterial species [51] were represented on the microarray (Table 1). This new diagnostic tool is composed of 1536 oligonucleotides, including viral, non-viral and positive control targets (Table 1). Array analysis is performed computationally using e-predict [52,53]. The sensitivity of the APM was estimated to be 1.9 x10^5 viral genome copies (1 pg *Drosophila* C virus *in vitro* transcribed genomic RNA) in an *A. mellifera* RNA (1 μg) background (see Materials and Methods). Array specificity was confirmed by performing pathogen-specific PCRs in conjunction with nucleic acid sequencing. Test samples included honey bees from managed and feral colonies, *Vespula* sp. (yellow jackets), and *Bombus* sp. (bumble bees)

(Supplemental Table 1). A sample from a collapsed colony in Montana tested positive for the highest number of viruses (BQCV, DWV, KBV, IAPV) and documented the array's ability to simultaneously detect multiple pathogens. Analysis of symptomatic honey bees, such as hairless, shiny bees and bees with deformed wings, confirmed the presence CBPV and DWV, respectively [54,55]. Likewise, analysis of *Varroa destructor* RNA validated the array's ability to detect mites and their associated viruses (DWV). Interestingly, pathogens normally associated with honey bees, DWV and ABPV, were also detected in a yellow jacket sample (*Vespula sp.)* obtained near a hive entrance from which the honey bees also tested positive for ABPV and DWV. We utilized the APM to detect several pathogens (BQCV, DWV, SBV and *Nosema*) in CCD-affected colony samples from an Oklahoma based migratory beekeeping operation (Feb. 2009). In total we detected and sequence confirmed ten previously characterized honey bee pathogens using the array including: CBPV, IAPV, DWV, ABPV, BQCV, SBV, KBV, *Nosema apis*, *N. ceranae* and *Varroa destructor*.

Temporal monitoring of 20 migratory honey bee colonies

Honey bee samples were collected during their travels from Mississippi through South Dakota to California resulting in a prospectively collected 10-month time-course of 431 data points, each consisting of 50-100 bees isolated separately from both the entrance (older foragers) and brood comb (younger house bees). Hives #10, #14 and #19 were lost in December due to queen death or infertility.

We analyzed all the entrance samples (5 bees per colony each time-point) using the APM.

Nosema

There was an abundance of *Nosema* infections in our monitor colonies throughout the entire time-course. APM monitoring revealed that approximately half of the colonies in April and May were *Nosema* positive (Figure 2A). Notably, nearly every colony was infected during a surge in August and September. In order to determine which *Nosema* species was responsible for infections, each hive was analyzed at a single time-point per month by species specific PCR. In April and May, *N.apis* was predominant whereas in June, July, and October through December, *N. ceranae* was exclusively detected (Figure 2B). During the highest incidence of *Nosema* (August – September), 75% of all colonies were infected with *Nosema ceranae* and less than 25% with *Nosema apis*, most of which were co-infected with *N. ceranae*. Quantitative-PCR data from pooled monthly samples confirmed that *Nosema ceranae* was prevalent throughout the time-course and peaked in August (Figure 2C). While seasonal variation may play a role, an anti-fungal (Fumagillian) was used to abrogate *Nosema* infection [56] and may be responsible for the observed decrease in *Nosema* abundance from November to January (Figure 2).

Viruses

The APM readily detected common honey bee viruses in samples collected throughout the time-course. In total, we report 69 virus incidences in 63 of 431 total samples (Figure 3). Overall virus incidence was sporadic, which we attribute to either rapid clearance (< 2 weeks) or mild infection in predominantly healthy monitor colonies. The majority of infections occurred during July, August, and September when the monitor colonies were in South Dakota. The most prevalent virus infections observed during our 10-month study were SBV, BQCV and ABPV; however the frequencies of specific viruses were insufficient for statistical tests. Other viruses including DWV, IAPV, and KBV were infrequently detected in the latter half of our time-course. A total of six double virus infections were detected, frequently involving ABPV or SBV. There were only three cases in which the same virus (BQCV) was detected in consecutive time points from a particular monitor colony (Figure 3A Hives #4, #6, and #20). Typically a single virus was detected in multiple colonies at a given time-point and these infections did not persist. For example, there were waves of SBV infection in April and January and of BQCV in July and early August (Figure 3A). qPCR analysis of pooled monthly samples confirmed and extended APM findings. BQCV, SBV and ABPV levels peaked in mid-summer to early fall at $6.6\text{x}10^{9}$- $8\text{x}10^{10}$ genome copies per bee (Figure 4), consistent with previously characterized levels of these viruses [55,57,58].

Ultra deep sequencing, discovery of novel viruses

A summer South Dakota time-point (August 5, 2009) was selected for deep sequencing due to high *Nosema* load and the presence of several common honey bee viruses, including ABPV, BQCV and SBV. All expected microbes (*Nosema ceranae, Crithidia mellificae)* and viruses were detected [ABPV (39,352 reads aligned by BlastN e-value < $1x10^{-7}$), BQCV (2,868 reads) and SBV (4,414 reads)]. In addition, we detected *Spiroplasma* sequences (70,407 reads) consistent with the presence of both *Spiroplasma apis* and *S. melliferum* (66 reads and 44 reads aligning to the RNA PolB gene of each, respectively).

Four distinct novel viruses were discovered via deep sequencing. Paired-end sequencing reads (2 x 63 nt) of unknown origin were screened by tBlastx [59] against all known insect viruses present in Genbank [60]. Screening hits with an e-value greater than $1x10^{-3}$ were used to target de novo contig assembly using the complete data set. Short contigs were screened by tBlastx against the non-redundant nucleotide database (NR) at an e-value threshold of $1x10^{-5}$. Hits to viral sequence, but not host sequences, were further assembled (see materials and methods). In each case, PCR primers were initially designed to bridge or confirm assembled contigs by Sanger sequencing. Confirmed contigs were extended with the PRICE assembler package (see Materials and Methods). In total, sequences from four novel viruses were recovered and Sanger validated. These include two members of *Dicistroviridae*, and two RNA viruses distantly related to *Nodaviridae*.

Aphid Lethal Paralysis virus strain Brookings

Investigation of contigs aligning to the Aphid Lethal Paralysis Virus genome, in the family *Dicistroviridae,* recovered a 4,125 nt contig (Genbank Q871932)) spanning the RNA-dependent RNA Polymerase (RdRp) gene, the internal ribosome entry site (IRES) structure and the capsid coding region. The recovered sequence aligned with 83% nucleotide and 89% amino acid identity to the canonical ALPV genome over the RdRp gene. The two viruses shared 97% nucleotide homology along 171 nt of the IRES. The high sequence similarity between this new isolate and canonical ALPV makes it unclear whether this is a novel species or a new strain of ALPV. Regardless, ALPV has not previously been reported in association with honey bees. We propose the designation ALPV strain Brookings (after the SD county from which the virus was isolated). Specific PCR primers were designed for the Brookings strain and utilized to analyze additional time-course samples, resulting in detections on thirty distinct occasions, including in Mississippi, South Dakota and California. Incidence peaked in May, when 7 out of 20 hives were infected, whereas maximum abundance occurred in August albeit at a relatively low level, 4.42x10^4 copies per 100 ng of RNA sample (approximately 2.21x10^7 copies per bee), as compared to previously characterized honey bee viruses (Figures 3 and 4). Frequent detection of ALPV strain Brookings throughout the time-course from multiple geographic locations suggests that this virus is not simply a “passenger” obtained from forage (nectar and pollen) shared with other insects. However, further investigation is required to determine whether ALPV strain Brookings is a honey bee pathogen.

Big Sioux River virus

A second novel dicistrovirus, designated Big Sioux River Virus (BSRV) after its place of discovery, is most similar to the *Rhopalosiphum padi Virus* (RhPV). Four contigs of size 1473, 861, 1164 and 1311 nt (Genbank JF423195-8) derived from the non-structural region, the IRES, and the capsid gene. BSRV shares low amino acid identity with RhPV; only 78% in the non-structural region and 69% in the capsid gene. This level of amino acid divergence is consistent with the taxonomic rank of a new species (Supplemental Figure 2). Twenty-eight incidences of BSRV were detected from 197 time-course samples by specific PCR with most individual colony detections occurring in samples collected from April to July 2009 in Mississippi and South Dakota. Incidence was low from October onwards (Figure 3B). Peak abundance was $7.64x10^3$ copies per 100 ng of RNA sample (approximately $3.8x10^6$ copies per bee) and occurred in August (Figure 4). Of note, BSRV associated significantly with *Nosema apis* infections (p=0.003, OR 6.0) and also with ALPV-Brookings (p=0.014, OR=4.5).

Lake Sinai Virus strain 1 and 2

Three contigs had significant alignment to chronic bee paralysis virus (CBPV) and members of the family *Nodaviridae*. Both the individual reads and our initial contigs were further assembled and extended using the complete data set (see Materials and Methods). Two separate contig sequences (5.5 kb each) were generated by de novo assembly. Both contigs were confirmed by specific PCR and Sanger sequencing. The first contig represents a novel RNA virus that we designate Lake Sinai virus (LSV1) (HQ871931), after Lake Sinai in Brookings County, South Dakota. The second contig also represented a related, yet divergent (71% nt identity), RNA virus which we designated Lake Sinai virus 2 (LSV2) (HQ888865). The 5' end of LSV1 was determined by RACE (rapid amplification of cDNA ends). The 5' end of the LSV2 assembly was within 57 nt of the LSV1 RACE results [19,55]. The 3' ends of both viruses were refractory to traditional RACE methods and attempts at 5' RACE on the negative strand were also unsuccessful.

Both LSV genomes display similarities to the RNA1 molecule of chronic bee paralysis virus (CBPV) with predicted open reading frames (ORFs) of similar size and arrangement with the notable exception that LSV1 and 2 ORFs are contained on a single RNA rather than in the bipartite configuration of CBPV [19,55] (Figure 5B). LSV1 and 2 possess the Orf1 gene, which is of unknown function, with predicted products (of 847 and 846 aa) previously unique to CBPV (853 aa). The Orf1 genes of LSV1 and CBPV share minimal (18%) amino acid identity. All three viruses encode an RdRp that partially overlaps and exists in a frame shift with respect to Orf1[19]. Both LSVs possess a triple stop codon within 10 residues of the end of the Orf1 gene whereas CBPV has two adjacent stop codons. The RdRp genes are considerably more conserved with 80% identity between the two LSV strains and 25% amino acid identity between them and CBPV. Both LSV RdRp genes have the DxSRFD and SG amino acid motifs in the NTP binding pocket (residues 375-380 and 436-437 in LSV1) conserved between the families *Nodaviridae*,*Tombusviridae* and CBPV. An amino acid phylogeny of the *Nodavirales* superfamily RdRp places the LSV strains on the same branch as CBPV, and separated from the larger *Nodavirus* and *Tombusvirus* families (Figure 5A).

As previously noted, the capsid protein of LSV1 and 2 is encoded on the same RNA as Orf1 and the RdRp unlike that of CBPV, which possesses a bipartite genome (Figure 5B). The capsids of LSV1 and 2 have significant profile similarity to the capsid gene of *Nudaurelia capensis beta-tetravirus* by HHpred [61] (e-value $1.0x10^{-26}$) and they exhibit weak direct protein alignment by Blastx (e-value $1.0x10^{-04}$). Similarity to *tetravirus* capsid genes consistently outranked similarity to CBPV or *nodavirus* capsids by these methods. *Tetraviruses* are not close relatives of the *Nodavirales* superfamily, although *Betatetraviruses* have a similar monopartite genome organization to LSV (Figure 5B). LSV1 and 2 share 70% amino acid identity over the capsid. The LSV1 capsid overlaps the RdRp gene in the +1 reading frame for 125 nt before ending in a pair of stop codons (separated by two residues). The LSV2 capsid is in frame with the RdRp and separated by 18 nt without a redundant stop codon.

Seven of twenty hives sampled on August 5, 2009 were positive for LSV1 and an additional five hives in the time-course, from July (SD) and January/February (CA) were found to be positive for LSV1, all with greater than 95% nucleotide identity. LSV2 was more prevalent and was detected by PCR in 30 of 197 time-course samples from all three geographic regions. LSV2 incidence surged in April, July and January during which over a third of all 20 monitor hives were infected. Strain specific qPCR demonstrated high abundance ($\geq 2x10^6$ copies per 100 ng RNA) of both LSV strains in our monitor colonies throughout the majority of the time-course (Figure 4). LSV1 copy number peaked in July, at $1.39x10^8$ copies per 100 ng of RNA sample (approximately $7.0x10^{10}$ copies per bee). Notably, LSV2 was the most abundant virus detected in this study ($\sim 10^{11}$ copies per bee). Copy number peaked in both April and January, at $7.22x10^8$ copies per 100 ng of RNA sample (approximately $3.61x10^{11}$ copies per bee) and $1.42x10^9$ copies per 100 ng of RNA sample (approximately $7.1x10^{11}$ copies per bee), respectively. Positive sense RNA viruses, like LSV 1 and 2, utilize a negative strand template to produce viral genome copies, therefore detection of the negative-strand intermediate is indicative of an actively replicating infectious virus [37,62,63]. We utilized negative-strand specific RT-PCR to detect the replicative forms of both LSV1 and LSV2 (Supplemental Figure 4). cDNA synthesis reactions were performed using tagged negative strand-specific LSV1 and 2 primers followed by exonulcease I digestion of excess unincorporated RT-primers [63] (Materials and Methods and Supplemental Table 2). PCR amplification using a tag-specific forward primer and LSV-specific reverse primers confirmed the presence of the replicative forms of both LSV1 and LSV2 in the July RNA sample (Supplemental Figure S4). Together, this data and the abundance of LSV1 and 2, compared to other significant honey bee viruses, suggests that LSV1 and LSV2 are novel honey bee viruses that play significant roles in colony health.

Crithidia mellificae

The broad scope of our microarray platform enabled identification of an unexpected microbe, *Crithidia mellificae*, in our time-course samples (Figure 6). Given that *Crithidia bombi* is a bumble bee pathogen and trypanosomatids were previously described in honey bees [10,64,65], 5 unique oligonucleotides each from *Crithidia oncopelti* and *C. fasciculata* rRNA sequences were included on the microarray. Oligonucleotides from these two distantly related organisms were predicted to hybridize to all other *Crithidia* species with published sequence [51]. Three oligonucleotides and their reverse complements derived from *Crithidia oncopelti* were repeatedly detected in samples throughout the time-course. Pilot Sanger sequencing of randomly amplified genomic DNA from a honey bee intestinal sample yielded a 121 base-pair (bp) stretch of the kinetoplast minicircle with 74% homology to the *Crithidia fasciculata* kinetoplast (BlastN e-value = 3.5 x 10^{-8}). Specific PCR retrieved 593 nt of the GAPDH gene to confirm phylogenetic placement.

We sought to further characterize this parasite by microscopy, PCR, culturing and DNA sequencing. Honey bee intestines were dissected in a sterile environment from which *Crithidia mellificae* was cultured. Light microscopy of these parasites enabled visualization of the flagella and motility (Figure 6; Supplemental Movies S5 and S6). Fixed sample imaging facilitated DAPI visualization of the kinetoplast DNA, as well as nuclear DNA (Figure 6). Previous studies describing trypanosomatids in honey bees lacked DNA-sequencing data with the exception of Cox-Foster *et al.* (2007) who published a 715-nt sequence of 18S ribosomal RNA that was too conserved between trypanosomatids for precise taxonomic assignment [10]. Together, the features observed by microscopy (flagella and kinetoplast) and phylogenetic analysis unambiguously identify this species taxonomically. We have deposited the GADPH sequence (JF423199) for future molecular identification, and genomic sequencing of *C. mellificae* is underway.

In order to specifically monitor *Crithidia mellificae,* additional oligonucleotides complementary to the *C. mellificae* rRNA and kinetoplast sequence were designed and included on the APM beginning in October 2009. These additional oligonucleotides enabled robust *C. mellificae* detection in later time-course samples, 33% of which tested positive (Figure 6). In addition, we screened samples throughout the time-course (April 2009 – Jan. 2010) by PCR and qPCR specific to the *C. mellificae* rRNA gene. *C. mellificae* infection was detected by PCR at every time-point and in turn from every geographic location sampled in our study (MS, SD, and CA). Likewise, *C. mellificae* was readily detected in pooled monthly RNA samples by qPCR throughout the year (Figure 6C). In contrast to BQCV, SBV, ABPV and *Nosema ceranae*, which exhibited peak levels in late summer and early fall, peak trypanosomatid levels occurred in January 2010. Despite this, *C. mellificae* infections statistically associated with *N. ceranae* infections (Chi Square p=0.004, OR=3.1). *C. mellificae* was also detected in numerous hobbyist and study hives in the San Francisco Bay Area (CA), as well as samples from a CCD-affected apiary in Oklahoma, indicating wide geographic distribution (Supplemental Table 1).

Spiroplasma melliferum and *S. apis*

Spiroplasma, a close relative of the genus *Mycoplasma*, are bacterial parasites that have been implicated as pathogens of insects, vertebrates and plants. Strains of *spiroplasma* similar to flower-associated parasites were identified as a pathogen of honey bees in France, *Spiroplasma apis* [66], and the United States, *Spiroplasma melliferum* [67]. Pilot Sanger sequencing of a pooled honey bee sample (August 2009) identified an rRNA-derived sequence from a *Spiroplasma*. Pan-*spiroplasma* and pan-*mycoplasma* PCRs targeting the 16S rRNA gene detected sporadic infections over most of the time-points and a surge of 9 infections in August and 6 infections in September. Sequence data indicates that these isolates have high homology to previously identified *spiroplasma* isolates (>98% nucleotide identity). *Spiroplasma* infections had strong associations with *N. ceranae* (Chi Square p=0.015, OR=7.2) and *C. mellificae* (p=0.000076, OR=16.3), however this may be an artifact of the short surge of *Spiroplasma* coinciding with a period of high *Nosema* load.

Phorid fly *(Apocephalus borealis)*

Apocephalus borealis, phorid flies, have previously been associated with bumble bee parasitism [68] and have recently been described as a parasite of honey bees in the San Francisco Bay Area [69]. *Phoridae* family members (*e.g. Pseudacteon* sp.) are well-characterized parasites of ants and other insects. These flies lay eggs inside the insect hosts, which are in turn consumed by the larvae during development. Although, *A. borealis* parasitism of honey bees is uncommon, we analyzed our time-course samples for the presence of phorid rRNA by PCR. Pooled monthly samples were weakly positive for *Apocephalus borealis* in December and January (Supplemental Figure S4). We sequenced PCR amplicons from two individual (October 2009 Hive $^{\#}$7 and $^{\#}$10) and one pooled-monthly (December 2009) samples and determined that the phorid rRNA sequences from our time-course shared 99% similarity to honey bee-parasitizing phorids captured in San Francisco. This is the first report of phorid flies in honey bee samples outside of California and thus expands their known geographic range (SD,CA), although the range *A. borealis* as a bumblebee pathogen extends across North America [70].

Discussion

The importance of honey bees to global agriculture and the emergence of CCD calls for increased longitudinal monitoring of infectious processes within honey bee colonies. The data presented herein represent the finest resolution time-course of honey bee associated microbes to date. We demonstrate the utility of an arthropod pathogen microarray (APM) for simultaneous detection of numerous pathogens and the power of ultra deep sequencing for viral discovery. Several previous studies examined honey bee samples from diseased or CCD-affected and healthy colonies [10,11,21,71,72], but few have temporally monitored multiple pathogens [58,73,74]. Although these studies differed in sampling strategy, geography, colony management (*e.g.* migratory commercial versus stationary hobbyist, chemically treated versus organic), and pathogen monitoring technology (*e.g.* serology, PCR, spore counts, microarray) they provide a framework for our surveillance of previously characterized honey bee pathogens.

*Nosem*a infection was prevalent in our 20 monitor colonies. *N. ceranae* was the predominant species. *N. apis* was detected in individual colony samples in April (Mississippi) and May (South Dakota), but was undetectable in pooled monthly samples, indicating relatively low levels. *N. ceranae* abundance peaked in early-spring and late-summer. Lower *N. ceranae* levels from November to January likely reflects antifungal (Fumagillan) treatments applied in the fall, but may also represent natural seasonal variation [56]. In comparison, another U.S.-based

(Mississippi, Arkansas) study, which calculated *Nosema* levels using qPCR of genomic DNA calibrated to spore counts, also reported overall dominance by *N. ceranae*, but higher *Nosema* levels in November 2008 as compared to March 2009 [75]. *Nosema* spore count data from non-CCD and CCD-affected colonies in California and Florida was not significantly different and approximately 50% of the colonies assayed were infected [11]. Data from European studies indicate varying prevalence of *N. apis* and *N. ceranae* [23,27,76,77]. For example, a retrospective analysis of honey bee samples from Spain, Switzerland, France and Germany indicated peak levels of *Nosema* (presumably *N. apis*) in early spring and mid-winter from 1999 to 2002, whereas from 2003 to 2005 *Nosema* incidences remained relatively high throughout the year, a result the authors attribute to increased prominence of *N. ceranae* associated with recent increased bee losses [27]. In contrast, a recent (2005-2009) time-course study in Germany demonstrated greater *Nosema* incidence in the spring, detected *N. apis* more frequently than *N. ceranae*, and found no correlation between colony loss and *Nosema* infection [76]. Variable *Nosema species* prevalence and abundance at both the apiary and individual colony level indicate that standardized, molecular biology-based monitoring of large sample cohorts is required in order to understand the dynamics of *Nosema* infection, which are likely influenced by multiple factors including host genetic variation, climate, exposure levels, and treatment regimes [75,78]. Recently, higher levels of *Nosema bombi* were detected in North American bumble bee species experiencing population decline [79]. Although, like CCD, the causes of bumble bee decline are complex and not fully characterized, this report underscores the importance of further characterizing the epidemiology and pathogenicity of *Nosema*.

We monitored the incidence of all known honey bee viruses, discovered 4 new honey bee associated viruses, and quantified the relative abundance of select viruses in time-course samples. Overall, virus infections in our monitor colonies were quickly cleared and no chronic infections of previously characterized honey bee viruses were observed. Our data suggest that healthy colonies are undergoing constant cycles of viral infection and clearance. The most prevalent, previously characterized viruses in our study were BQCV, ABPV and SBV. The peak incidence of BQCV (25%) occurred in July, whereas ABPV (6.3%) and SBV (12.5%) peaked in August. Summer peak virus incidence was also reported in a PCR based honey bee virus (BQCV, ABPV, and SBV) survey of 36 geographically distributed apiaries in France (BQCV, ABPV, DWV, SBV, CBPV, KBV) [74], a qPCR time-course study of 15 colonies in England (BQCV and ABPV) [58], and an unpublished East-coast U.S. based survey (BQCV) [20]. Another virus, invertebrate iridescent virus-6, claimed to be associated with CCD and prevalent (75%) in healthy colonies but not supported in subsequent analysis [80,81], was never detected by the APM (n=431), end-point PCR (n=197), or in any of the 20 samples that were deep sequenced [82].

Seasonality of specific pathogens in our time-course study representing 2,155 individual bees from 431 samples varied, although many including BQCV, APBV, SBV, *Nosema*, exhibited reduced June and peak August levels. Peak incidences of these organisms in the spring and late summer are likely attributable to increased brood rearing [20,74,83] and foraging during these seasons [84]. Increased brood rearing during the summer, results in a greater number of bees capable of transmitting pathogens to other members of the colony living in very close proximity [20]. Honey bee viruses are transmitted vertically via infected queens and horizontally via the oral-fecal route or through the exoskeleton [20,22]. Foraging activity also increases pathogen exposure [84] and may also stress the bees so that inapparent infections reach detectable levels. Although other sources of stress, such as transportation and poor nutrition, are hypothesized to increase pathogen levels [11], these factors were minimal during the summer when the monitor colonies were stably situated in South Dakota foraging on diverse pollen and nectar sources, including alfalfa (*Medicago sativa L.*), sweet clover (*Melilotus spp.)* and a variety of other flowering plants in June with increasing availability of corn (*Zea mays ssp.*) and soybean (*Glycine max)* pollen later in the summer. Notably, these colonies were part of a typically managed commercial beekeeping operation and therefore received nutritional supplements, protein paddies and sugar syrup throughout the year (Materials and Methods). Adequate monitor colony nutrition may have played an important role in the rapid virus clearance observed in our study. Although further experimental validation is needed, recent work examining the effects of nutrition on DWV titer in caged-bee studies demonstrated that viral titer was reduced by pollen and protein supplementation [85]. In addition, anti-mite and antimicrobial treatments in the spring and late-fall may have accounted for the lower pathogen levels at those times of year and in turn for the relatively high levels during the summer (Materials and Methods). We did not observe either increased incidence or abundance of any of the microbes and viruses monitored in our study after long distance transport.

Although several monitor colonies were lost (n=3; one unfertile (drone laying) queen, two queen-less colonies) and many (n=8) had fewer than 6 frames of bees in February 2010, none exhibited CCD characteristics and none of the numerous viruses and microbes we surveyed correlated with the weak colonies. Interestingly, our sample cohort had very few incidences of IAPV and DWV. IAPV, a virus that has received much attention due to its correlation with CCD-affected samples in an early study [10], although not in a subsequent expanded study [11], was detected in our monitor hives in December. The colonies in our study cleared or reduced IAPV infection to levels below detection within one week, indicative of a mild infection (Figure 3). IAPV infection has been shown to cause paralysis and death in mini-colony and cage studies [15,86], although its role in CCD is unclear [11,87,88]. Likewise, DWV incidence in our time-course samples was very low (0.7%) and presumably cleared rapidly. In contrast a French time-course documented increased DWV incidence throughout the year (spring 56%, summer 66%, autumn 85%) [74] and two U.S. studies also report high DWV incidence [20,72]. Our results are not indicative of poor DWV detection by the array or our sampling strategy, since DWV was detected in both entrance and interior samples from other colonies. In addition, DWV-specific PCR of pooled monthly time-course samples was negative (Supplemental Figure S3). Therefore negligible DWV in our monitor colonies may be attributed to low exposure and/or good colony health. A thorough one-year investigation of virus (ABPV, BQCV, DWV) and *V. destructor* in England found a correlation between DWV copy number and over-winter colony loss [58]. Lack of DWV in our monitor colonies is consistent with low *Varroa destructor* incidence, since mites are known to transmit DWV [37,89,90]. Low incidence of both DWV and *V. destructor* in our study may be partially attributed to our analysis of entrance samples, which consist of actively foraging and/or guarding adult bees. Since *Varroa* mites parasitize larva they are more readily detected in larva and young bee samples as well as hive bottom boards. More significantly, monitor colonies received miticide treatments in order to reduce *V. destructor* burden.

Deep sequencing analysis revealed the presence of four novel viruses (ALPV-Brookings, BSRV, LSV1 and LSV2), illustrating the power of this technique for honey bee virus discovery. The Lake Sinai viruses are extremely divergent from known insect viruses in both amino acid identity and genome organization. They are most closely related to CBPV, a known pathogen of honey bees [55]. Since the presence of viral nucleic acid does not necessarily indicate infection, as pollen pellets of infected and non-infected workers are known to harbor honey bee viruses [84], we confirmed the presence of the replicative forms of LSV1 and 2 in time-course samples. The magnitude of LSV throughout the time-course also suggests that these are bona fide honey bee viruses. LSV2 was the most abundant virus in our study. It is intriguing that peak virus copy number per bee occurred in April (~ 3.6×10^{11}) and January (~ 7.1×10^{11}) since colonies typically collapse during the winter months. In contrast, LSV1 copy number peaked in July, similarly to the previously described honey bee viruses monitored in our study. Frequent detections of both ALPV-Brookings and BSRV (~15% incidence in the time-course) by PCR screen in different geographic regions argues against simple carryover from other insects during foraging, but does not rule out potential re-infection from stored pollen (bee bread) [84]. Research to determine the potential pathogenicity of these four new viruses in honey bees is underway.

Crithidia mellificae was readily detected throughout the time-course. In contrast to most other prevalent microbes and viruses, relative *Crithidia* levels peaked in the winter (January 2010). The effects of *C. mellificae* on the honey bee host remain relatively uncharacterized compared to those of *C. bombi* on bumble bee, which include reduced worker fitness and colony survival [64,92]. To date, there are only a few reports of *C. mellificae* infection of honey bees in the literature including early work describing the first isolation and culture of this organism in 1967 from Australian honey bees [65]. This work tested the effect of feeding *C. mellificae* to honey bees and demonstrated similar mortality rates in infected and uninfected bees [65]. More recently, similar trypanosomatid prevalence and loads were reported in CCD-affected colonies and healthy controls [10,11]. Although current data suggest that *C. mellificae* does not dramatically affect colony health additional pathogenesis research in honey bees is warranted considering the detrimental effects of *C. bombi* on bumble bee colonies.

The importance of honey bees in agriculture and the emergence of CCD underscores the need to monitor honey bee associated viruses and microbes in healthy colonies over time. The confinement of *Spiroplasma* infection to a two-month window demonstrates the value of time-course sampling as opposed to single-collection screens. The development of high throughput platforms, such as the APM, will facilitate monitoring of exogenous agents in order to better understand their effect on honey bee health and survival. Our discovery and genomic characterization of four new viruses will facilitate future monitoring. Temporal characterization of these and the other microbes described herein offers a more complete view of the possible microbe-microbe and microbe-environment interactions. Further studies examining any subtle or combinatorial effects of these novel microbes are warranted. Increased analysis of prospectively collected samples is essential to address the hypothesis that either one or more viruses and/or microbes cause CCD. To our knowledge, this is the first U.S. honey bee pathogen monitoring study to report both comprehensive pathogen incidence and relative abundance of specific pathogens over time. Results from our molecular analysis pipeline (APM, PCR, qPCR, ultra deep sequencing) provide a basis for future epidemiologic studies aimed at determining the causes of CCD.

Materials and Methods

Collaborating commercial beekeeping operation information

Twenty monitor hives were established in April 2009 by a large-scale (>72,000 hives), migratory commercial beekeeping operation (Mississippi, California, and South Dakota, U.S.A.) that experienced CCD-losses in 2007/08. Standard beekeeping management practices for an operation of this size were employed. Treatment regimes throughout the year were as follows: (1) anti-mite treatment April 2009, just prior re-queening – amitraz; (2) antibacterial treatment May 2009 - oxytetracycline hydrochloride (OTC) (Terramycin™); (3) anti-fungal (*Nosema sp.)* treatment August 25, September 12, and October 13, 2009 - fumagillan; (4) antibacterial treatment late August, early September, 2009 - tylosin tartrate; (5) anti-mite treatment September 12, 2009, after harvesting honey; (6) anti-mite treatment – early November and early December 2009 - essential oils from lemon grass and spearmint (Honey-B-Healthy™). Honey bees colonies were periodically supplemented with sugar syrup and protein supplement. In April (1 gallon) and October (2 gallons) bees were fed 50% (weight/volume) sucrose; in November all colonies received 3 gallons of a 1:1 mixture of high fructose corn syrup-55 (HFCS-55, 55% fructose, 42% glucose) and sucrose syrup. Additional sugar syrup was given to colonies based on colony weight (< 80 lbs - 3 gallons, 80-90 lbs - 2 gallons., 90-100 lbs – none). This operation experienced an average 18% colony loss from November 2009 to February 2010. Colonies with younger queens (≤ 2 years old) experienced 11% loss, whereas colonies with older queens experience 21% loss.

Honey Bee sampling and storage

Samples (~ 50-100 bees) were collected into 50 mL Falcon tubes using a modified hand-held vacuum cleaner from both the entrance and interior of the hive and immediately put on dry ice for overnight shipment to our laboratory. Samples were stored at -80ºC until RNA extraction; excess bees were archived for long-term -80ºC storage. Time-course samples were collected monthly from April 15 (week 1) through July 14 (week 14), 2009 and weekly samples were attempted thereafter, however due to inclement weather or shipping logistics the samples for weeks 15, 28-30, 32, and 39-41 were not collected. A total of 864 samples were obtained and 431 exterior samples were analyzed.

Honey bee sample preparation

We determined that analysis of five honey bees per sample was sufficient for our colony monitoring project. Arthropod pathogen microarray (APM) analysis of test samples revealed that combined analysis of 5 bees reproducibly detected most, if not all, of the pathogens detected from 10 or 15 independently analyzed bees from the same sample. In addition, we confirmed the consistency of APM results by performing multiple analyses of a single RNA sample. Based on our test results and practical sample handling considerations, we reasoned that repeated analysis of 5 bees from each colony over-time (115 bees per colony) was sufficient for this study.

Honey bee samples, 5 bees per colony each time-point, were homogenized in 1 mL 50% TRIzol Reagent (Sigma) and 50% phosphate buffered saline (PBS, UCSF Cell Culture) solution in a 2 mL micro-centrifuge tube containing one sterile zinc-coated steel ball bearing (5 mm) using a TissueLyzer II (Retsch), for 4 minutes at 30 Hz. RNA was isolated according to TRIzol Reagent (Invitrogen) manufacturer's instructions. In brief, TRIzol reagent honey bee homogenate was combined with 0.1 ml chloroform and mixed by vortexing for 5 seconds, samples were incubated at room temperature for 5 minutes, prior to centrifugation for 10 minutes at 13,200 x g in a table top centrifuge. Next, 700 ⌈L of the aqueous phase was transferred to a new microfuge tube containing 490 ⌈L isopropanol. Following mixing, the samples were incubated at -20 ºC for 20 minutes and then either centrifuged (13,200 x g for 15 min) or further purified utilizing Zymo-III RNA columns according to manufacture's instructions (Zymo). RNA was extracted from five bees collected from the colony entrance for each of the time-course samples.

Arthropod Pathogen Microarray design and synthesis

Design principles used for APM oligonucleotides (70 nt) were based on previous pan-viral microarrays using ArrayOligoSelector (AOS) [52]. Briefly, array oligonucleotides were selected for uniqueness against an insect nucleic acid background, for ~50% GC content to maintain high complexity, and for cross-reactivity of highly-conserved nucleic acid features with evolutionarily related targets (<-50 kcal/mol predicted binding energy). Arthropod pathogen oligonucleotides (GEO GPL11490) were synthesized by Invitrogen, suspended at 40 pmol/ ⌠L in 3X SSC and 0.4 pmol/ ⌠L control oligo and printed on poly-L-lysine slides (Thermo) with silicon pins as previously described [93]. Each oligonucleotide and its reverse complement were printed twice for redundancy. Arrays were allowed to air-dry and stored and room temperature. Prior to use, oligonucelotides were cross-linked to slides via UV exposure (600 mJ), washed with 3X SSC / 0.2% SDS and blocked using a methylpyrrolidone solution (335 mL 1-methyl-2-pyrrolidinone, 5.5 grams succinic anhydride, 15 mL 1M sodium borate).

Sample Preparation for Arthropod Pathogen Microarray

(Reverse Transcription, CyDye Labeling, Hybridization, Scanning)

For each sample, 5 ⌠L (~ 15 ⌠g nucleic acid) of extracted material was randomly primed and amplified as previously described [42,43]. Briefly, an adapter-linked random nonamer (5‘GTTTCCCACTGGAGGATANNNNNNNNN) was used to prime the reverse transcription reaction using SuperScript II (Invitrogen). The same oligo is used for two rounds of second-strand synthesis with Sequenase (USB) in order to produce adapter-flanked sequences from both RNA and DNA starting material. One-quarter of the random priming reaction is used in a 50 ⌠LTaq PCR reaction for 25 cycles with a single primer

(5’GTTTCCCACTGGAGGATA). One-tenth of the amplified material was further amplified for 10-20 cycles with a Cy3-linked primer (5‘Cy3 - GTTTCCCACTGGAGGATA). Samples were purified with the Zymo DNA Clean and Concentrator (Zymo) and resuspended in a buffer of 3X SSC, 50 mM HEPES and 0.5% SDS, and t hybridized on the APM overnight at 65ºC. Arrays were washed and scanned with an Axon 4000A scanner. Samples were analyzed manually and scored as positive for a pathogen if at least three unique oligonucleotides hybridized with at least five times background intensity. Arrays were further analyzed by a second unbiased method using the E-Predict algorithm [52,53], wherein all virus genomes were computationally hybridized to the array oligos and array results are compared to expected binding profiles. The top 5 unique oligos were removed and the algorithm reiterated twice in order to improve detection of low titer target(s) during a co-infection. Known honey bee pathogens were called positive if they exceeded a similarity score of 0.001 and were the highest ranked call in any iteration. In the event of a disagreement between the two analysis methods, a specific PCR reaction was performed, using material from the first PCR step, to resolve the call.

Assessment of Arthropod Pathogen Microarray sensitivity

In order to estimate the sensitivity of the arthropod pathogen microarray (APM) two positive control samples were prepared in the presence and absence of pathogen-free honey bee RNA. A full-length (9,264 nucleotide) *Drosophila* C virus (DCV) clone was *in vitro* transcribed, serially diluted into honey bee RNA, reverse-transcribed, amplified, dye-labeled and hybridized to the APM as described above. Detection of at least 3 of the 8 unique DCV oligonucleotides and their reverse complements resulted in an estimated DCV detection level of 1.9 x10^5 genome copies (1 pg DCV genomic RNA) in an *A. mellifera* RNA (1 μg) background. Similarly, detection of a BQCV genome segment (452 nt), corresponding to one array oligo and its reverse complement, diluted into either pathogen-free honey bee RNA (0.5 μg) or water indicated detection limits of 1.2 x10^5 genome segment copies (30 fg BQCV RNA segment) and 1.2 x10^4 genome segment copies (3 fg BQCV RNA segment) respectively.

PCR Screen

Reaction conditions for polymerase chain reaction (PCR) amplifications of select samples were performed under the following conditions: 5 ⌠L of 1:10 dilution of RdB DNA and 10 pmol of each forward and reverse primers were amplified with Taq polymerase with the following cycling conditions: 95ºC for 5 min; 95ºC for 30s, 50-60ºC for 30s, 72ºC for 1 min, 35 cycles; final elongation 72ºC for 7 min., hold at 4ºC. Select samples were Sanger sequenced directly from ExoI and SAP treated PCR product or from colony PCR of TOPO cloned (Invitrogen) gel-extracted bands. Bands produced by PCR assays for known honey bee pathogens were sequenced until each molecular weight product was unambiguously associated with either a true positive or non-target amplification of the honey bee genome or microbiome. All PCR results for the four novel viruses were confirmed by Sanger sequencing.

Quantitative PCR (qPCR)

qPCR was performed on pooled samples from each month. Equivalent amounts of RNA (10 ⌈g) from each hive sample (monitor hives 1-20) were pooled according to the month in which they were collected (April 2009 to January 2010). Pooled RNA was further purified using Qiagen RNAeasy columns, including on column DNase Treatment (Qiagen). cDNA synthesis reactions were performed with SuperScriptIII (Invitrogen) according to manufacturer's instructions. In brief, RNA from each pooled sample (5 ⌈g), random hexamer (1.25 ⌈g) and dNTPs (0.5 mM each) were combined in a 50 ⌈L reaction volume, incubated at 65°C (5 min), cooled on ice (1 min) and subsequently combined with 50 ⌈L of 2x First-Strand Buffer containing SSIII (1000 U), DTT (5 mM), and RNaseOUT (200 U). Reverse transcription reactions were incubated for 12 hours at 42°C followed by inactivation of the reaction (70°C, 15 min). qPCR was performed in triplicate wells using 2 ⌈L of cDNA as template in 20 ⌈l reactions composed of HotStartTaq 2X Mastermix (Denville), 1X SYBR Green (Invitrogen), $MgCl_2$ (3 mM), and forward and reverse primers (600 nM each) (Supplemental Table 2) on a LightCycler480 (Roche). The qPCR thermo-profile consisted of a single pre-incubation 95ºC (10 min), 35 cycles of 95ºC (30 s), 60ºC (30 sec), and 72ºC (30 s). No RT control reactions using pooled RNA as the template for qPCR were performed in triplicate on each plate. Target qPCR amplicons were cloned into pGEM-T (Promega) or TOPO CR 2.1 (Invitrogen) vectors and sequence verified. Plasmid standards, containing from 10^9 to 10^2 copies per reaction, were used as qPCR templates to assess primer efficiency and generate the pathogen-specific standard curves used to quantify the viral genome or rRNA copy number. The linear standard equations generated by plotting the crossing point (Cp) versus the log$_{10}$ of the initial plasmid copy number for each primer set were as follows: BQCV Cp = -5.67x +59.44, R^2 = 0.975; SBV Cp = -5.34x +56.33, R^2 = 0.976; ABPV Cp = -4.03x +43.7, R^2 = 0.995; LSV1 Cp = -4.21x +46.56, R^2 = 0.993; LSV2 Cp = -3.66x +40.76, R^2 = 0.998; ALP-Br Cp = -2.91x +34.76, R^2 = 0.980; BSRV Cp = -3.28x +36.93, R^2 = 0.999; *Nosema ceranae* Cp = -7.03x +69.43, R^2 = 0.975; *Crithidia* rRNA Cp = -3.13x +36.44, R^2 = 0.994 (LightCycler 480 Software, Abs Quant/2nd Derivative Max, high sensitivity mode, Roche). The detection limits of each qPCR primer set were as follows: *Crithidia* and ALP-Br - 10^2 copies, LSV2 and BSRV -10^3 copies, BQCV, SBV, ABPV, LSV1 and *Nosema* -10^4 copies. Specific qPCR amplicons had Cp values of < 30. Pathogen copy number data were reported per RT-qPCR reaction (Figure 4). Values obtained from the no RT control reactions, all below the detection limit of the assays, were subtracted from the total pathogen copy number for each month. An estimate of the number of viral genomes per bee can be obtained by multiplying the reported qPCR copy number values by 500. This estimate is based on the following: typical RNA yield was approximately 50 ⌠g per bee, each qPCR reaction was performed on cDNA generated from 100 ng RNA, therefore each well represents 1/500th of an individual bee. We choose to represent the raw data, since each monthly-pooled sample was composed of variable bee numbers due to differential sampling frequency each month. In addition, qPCR with a host primer set, *Apis m.* Rpl8, was performed using 1 ⌠L cDNA template on each qPCR plate to ensure consistency and cDNA quality. qPCR products were analyzed by melting point analysis and 2% agarose gel electrophoresis (Supplemental Figure S1).

Negative strand-specific RT-PCR

LSV strain 1 and 2 positive samples were analyzed for the presence of negative-strand RNA, which is indicative of virus replication, using strand-specific RT-PCR [37,62,63]. RNA from select samples (*e.g.* pooled July sample) was further purified using Qiagen RNAeasy columns, including on column DNase Treatment (Qiagen). cDNA synthesis reactions were performed with SuperScriptIII (Invitrogen) according to manufacturer's instructions using negative strand-specific LSV1 and 2 primers tagged with an additional 21 nt of sequence (5'-GGCCGTCATGGTGGCGAATAA) at their 5' end [63]; the tag sequence shares no homology with LSV nor to the honey bee genome (primer sequences listed in Supplemental Table 2). In brief, RNA from each sample (1 ⌠g), tagged-negative strand specific LSV primer (10 pmole) or random hexamers (50 ng) and dNTPs (0.5 mM each) were combined in a 10 ⌠L reaction volume, incubated at 65°C (5 min), cooled on ice (1 min) and subsequently combined with 10 ⌠L of 2x First-Strand Buffer containing SSIII (200 U), DTT (5 mM), and RNaseOUT (40 U). Reverse transcription reactions were incubated for 1 hour at 50°C followed by inactivation of the reaction (70°C, 15 min). Unincorporated primers present in the RT reactions were digested with exonuclease I (Fermentas), 0.1 Units per reaction which corresponds to a 10-fold excess of enzyme relative to the initial primer concentration, at 37°C for 30 min followed by heat inactivation at 80°C for 15 minutes. PCR was performed using 2 ⌠L of exonuclease I treated cDNA template in 25 ⌠l reactions containing 10 pmol each of a tag-specific forward primer (TAGS) and an LSV-specific reverse primer using the following cycling conditions: 95ºC for 5 min; 95ºC for 30s, 58ºC for 30s, 72ºC for 30s, 35 cycles; final elongation 72ºC for 4 min., hold at 4ºC. In addition to amplification and detection of the LSV replicative form using tagged-negative strand primed cDNA template and TAGS forward and LSVU-R-1744 PCR primers, negative and positive controls were performed (Supplemental Figure S4 – labeled (1)). Negative controls included utilizing unprimed RT reaction as a template for PCR amplification using TAGS forward and LSVU-R1744 primers (labeled (2)), LSV tagged negative-strand primed cDNA template in PCR reaction in which only the LSVU-R1744 primer was added in order to ensure that all of the unincorporated RT primer was digested with exonuclease I and thus not involved in priming the PCR reaction (labeled (5)), and no template PCR using LSV qPCR primer sets (labeled (6)). Positive controls included using random hexamer primed cDNA as template for PCR amplification using LSV1 or LSV2 -specific forward primer and LSVU-R-1744 (labeled (3)) and random hexamer primed cDNA amplified using LSV-specific qPCR primer sets (labeled (6)). PCR products were analyzed using agarose (2%) gel electrophoresis (Supplemental Figure S4).

Crithidia mellificae strain SF - Microscopy, Culturing and DNA Purification

Honey bees were collected from a San Francisco, CA (U.S.A.) colony previously identified to be *Crithidia* positive by microarray and PCR testing. Honey bees were immobilized by chilling at 4ºC for 20 minutes, briefly washed in 70% ethanol, and decapitated prior to dissection. The SF strain was isolated from honey bee intestines dissected in a sterile environment, minced and placed in a T25 flask and cultured in BHT medium composed of Brain Heart Infusion (BHI) 28.8 g/L (DIFCO), tryptose 4.5 g/L (DIFCO), glucose 5.0 g/L, Na_2HPO_4 0.5 g/L, KCl 0.3 g/L, hemin 1.0 mg/L, fetal bovine serum (heat inactivated) 2% v/v, pH 6.5, and containing penicillin G sodium (10^6 units/L) and streptomycin sulfate (292 mg/L) at 27ºC [94]. Free active *Crithidias* were observed 24 hours post inoculation. Parasites were maintained by subculture passage every 4 days; stable liquid nitrogen stocks were archived. Light microscopy of live parasites was performed using a Leica DM6000 microscope equipped with Hamamatsu C4742-95 camera and Volocity Software (PerkinElmer). Imaging fixed parasites (4% paraformaldehyde, 20 min) facilitated visualization of DAPI (4',6-diamidino-2-phenylindole) stained nuclear and kinetoplast DNA. Images of fixed *Crithidia mellificae* were obtained using both the Leica DM6000 microscope and a Zeiss LSM 510-M microscope equipped with both a 63x objective numerical aperture 1.4, and a 100x objective numerical aperture 1.4.

For DNA purification, *Crithidia mellificae* (~10^6 trypanosomes/mL culture medium) were pelleted by centrifugation (800xg for 6 min) and washed with PBS prior to DNA extraction. DNA was extracted using the DNeasy Genomic DNA

Extraction Kit (Qiagen) as per the manufacturers instructions. Bees from Crithidia positive hives were homogenized by TissueLyser as above and DNA extracted using the DNeasy kit for the initial PCR screens, after suspension in either PBS or 1X Micrococcal Nuclease Buffer (NEB).

Ultra Deep Sequencing Library Preparation

Total nucleic acid from all twenty monitor hives at time-point 17 (August 5, 2009) was pooled (approximately 3 μg per hive). One quarter was treated with RNase A/T1 (Fermentas) and genomic DNA was isolated using a DNeasy column (Qiagen). 50 ng of genomic DNA was prepared for deep sequencing by Nextera recombinase (Epicentre) per the manufacturer's instructions. The remaining nucleic acid was treated with Turbo DNase (Ambion) and column purified (Zymo) before being split into thirds. One third was enriched for mRNAs with dT-linked Dynabeads (Invitrogen). RNA from this fraction and from a second unenriched fraction were primed for RT and second-strand synthesis with an adapter linked oligo as above using oligo SolCommonN (5'CGCTCTTCCGATCTNNNNNN). The third fraction of RNA was primed with an anchored oligo dT and subjected to two rounds of second strand synthesis with SolCommonN. Half of the initial material was amplified with primer SolCommon (5'CGCTCTTCCGATCT) with KlenTaq (Sigma) at an annealing temperature of 37ºC for 20 cycles. Reactions were cleaned by Zymo column, analyzed by NanoDrop spectrophotometer and 50 ng was used in a four-primer PCR reaction. In a 50 ⌠L KlenTaq reaction, 10 pmol each of primers 5Sol1 (5'AATGATACGGCGACCACCGA) and 5Sol1

(5’CAAGCAGAAGACGGCATACG) and 0.5 pmol of Sol1 (5’AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) and Sol2 (5’CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT) were incubated for 2 cycles annealing at 37ºC and 10 cycles at 55ºC. Products were run on an 8% native acrylamide TBE gel (Invitrogen) and a 300-350 nt smear was cut out and electro-eluted. The product was further amplified at an annealing temperature of 55ºC with primers 5Sol1 and 5Sol2 for 5-10 cycles until at least 30 ng of material was produced, as determined by NanoDrop. Libraries were sequenced on an Illumina Genome Analyzer II with a V3 cluster generation kit and V5 sequencing reagent as per the manufacturer’s instructions, producing paired-end 65 nt reads.

Solexa Data Analysis and Virus Genome Recovery

Six pools of sequence data were downloaded from Genbank: *Nosema ceranae* (draft genome), *Spiroplasma* (*S. citri* draft genome and all sequences longer that 500 nt), DNA viruses of arthropods (all complete genomes), all small RNA viruses of arthropods except *dicistroviridae* and *iflavirus* (complete genomes), all members of *dicistroviridae* and iflavirus except those infecting honey-bees (complete genomes), and all known honey bee RNA viruses (complete genomes). Each pool was converted into a Blast library and queried against the entire Solexa dataset by BlastN and tBlastx. Hits with an e-value greater than 1x10^{-3} were extracted along with their paired end, regardless of similarity. Each pool was assembled using the Geneious sequence analysis package [95]. Contigs greater than 250 nt were queried again against the dataset by tBlastx with an e-value threshold of $1x10^{-5}$. Any positive hits were then queried against the NR database with the same parameters to eliminate spurious hits.

Contigs that appeared divergent or that were derived from non-honey bee associated viruses were extended using the entire read dataset using a paired-end contig extension algorithm ("PRICE" Graham Ruby, manuscript under preparation). The extended contigs were then independently confirmed by PCR recovery and Sanger sequencing. Individual paired-end reads that were discordant with the recovered contigs were used to further nucleate new contigs via contig extension. Primer3 [96] was used to design primers bridging adjacent contigs, as determined by mapping onto known virus genomes. Individual viruses or other microbes were queried with a BlastN threshold e-value of $1x10^{-7}$ (W7) to determine read counts.

Statistical Analysis

Associations were calculated treating each hive sample at each time-point as a distinct event. P-values (Chi-square values) and odds ratios listed were calculated by the OpenEpi statistical package v2.3 (http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm). Only seven microbes with incidences in the study set of at least 10% (20 incidences in 197 samples) were examined for association, resulting in 28 discrete association tests and the corresponding Bonferroni multiple testing correction. Microbes occurring infrequently were not used in association tests and so did not contribute to multiple testing correction.

Figure Legends

Figure 1. Temporal monitoring of the honey bee microbiome from 20 monitor colonies within a large-scale migratory U.S. beekeeping operation using a custom arthropod pathogen microarray, PCR, quantitative PCR, and ultra deep sequencing. The colonies were established with new queens in Mississippi (MS) in April 2009, moved to South Dakota (SD) in May 2009, and finally to California (CA) in November 2009; monitoring concluded in January 2010.

Figure 2. *Nosema* detection and quantification in time-course samples from 20 honey bee colonies. (A) Arthropod pathogen microarray detection of *Nosema sp.* in each colony (5 bees per sample) throughout the 10-month time-course. Colonies were managed using standard commercial beekeeping practices and treatments, which are listed below panel A and further described in Materials and Methods. (B) *Nosema ceranae* and *Nosema apis* incidence assessed by species-specific end-point PCR from a single time-point (n=20) each month; the positive sample percentages in each pie-chart are indicated in red. (C) Relative abundance of *Nosema ceranae* throughout the time-course assessed by qPCR of pooled monthly RNA samples; quantification of rRNA copy number based on a standard curve as described in materials and methods.

Figure 3. Detection of viruses and microbes in time-course samples from 20 honey bee colonies. (A) Arthropod pathogen microarray detection of viruses:

sacbrood virus (SBV), black queen cell virus (BQCV), acute bee paralysis virus (ABPV), Israeli acute bee paralysis virus (IAPV), Kashmir bee virus (KBV), deformed virus (DWV) in each colony (5 bees per sample). (B) Incidence of select parasites assessed by end-point PCR from a single time-point each month (each chart n=20, except January n=17); the positive sample percentages in each pie-chart are indicated in red.

Figure 4. Relative abundance of select viruses assessed by RT-qPCR of pooled monthly time-course samples. Viral genome copy numbers per 100 ng RNA were calculated based on standard curves [(black queen cell virus (BQCV), sacbrood virus (SBV), acute bee paralysis virus (ABPV), Lake Sinai virus strain 1 (LSV1), Lake Sinai virus strain 2 (LSV2), aphid lethal paralysis virus strain Brookings (ALP-Br), and Big Sioux River virus (BSRV)]; multiplying reported values by 500 provides a copy number per bee estimate, as further described in Materials and Methods. LSV2, a novel virus, reached the highest copy number observed in this study in January 2010 (1.42x10^9 copies per 100 ng of RNA sample; approximately 7.1x10^{11} copies per bee); note the x-axis on each graph was independently scaled.

Figure 5. Phylogenetic placement and genome organization of Lake Sinai viruses.

(A) RdRp amino acid phylogeny of the *Nodavirales* superfamily. Lake Sinai virus strain 1 (LSV1; HQ871931), Lake Sinai virus strain 2 (LSV2: HQ888865), chronic bee paralysis virus (CBPV; NC010711), boolarra virus (BoV; NC004142), *Nodamura* virus (NoV; NC002690), barfin flounder nodavirus BF93Hok (BFV; NC011063), grapevine Algerian latent virus (GALV; NC011535), melon necrotic spot virus (MNSV; NC001504), pothos latent virus (PoLV; NC000939) and carrot red leaf virus (CtRLV; NC006265). Protein sequences were aligned by ClustalW and a tree generated by the Neighbor-Joining method with 100 replicates [95] (B) Genome organization of the Lake Sinai viruses and similar RNA viruses.

Figure 6. *Crithidia mellificae,* SF strain detection and quantification. (A) Light and fluorescent microscope images illustrate key features of this trypanosomatid parasite including DAPI stained kinetoplast DNA (yellow arrow) and nuclear DNA (white arrow), as well as the flagellar pocket (bottom panel, red arrow); scale bar = 5 μm. (B) Arthropod pathogen microarray detection of *Crithidia mellificae* in each colony (5 bees per sample) from October 2009 to January 2010. (C) Relative abundance of *Crithidia mellificae* throughout the time-course as assessed by RT-qPCR of pooled monthly time-course samples; quantification of rRNA copy number based on a standard curve as described in Materials and Methods.

Supplemental Figure Legends

Supplemental Table 1. Arthropod pathogen microarray results from test samples.

Supplemental Table 2. Primers used in this study, * denotes primer sets used for PCR screening results in Figure 3B, ** denotes qPCR primer sets used to obtain the results in Figure 4 and Supplemental Figure S3.

Supplemental Figure S1. Gel electrophoresis of RT-qPCR products from pooled-monthly samples. qPCR products were amplified using the primer sets listed in Supplemental Table 2: *Nosema ceranae* 249 bp, *Crithidia mellificae* 153 bp, black queen cell virus (BQCV) 141 bp, sacbrood virus (SBV) 103 bp, acute bee paralysis virus (ABPV) 177 bp, Lake Sinai Virus strain 1 (LSV1) 174 bp, Lake Sinai Virus strain 2 (LSV2) 225 bp, Aphid Lethal Paralysis Virus Strain Brookings (ALP-Br) 192 bp, and Big Sioux River virus (BSRV) 281 bp. Molecular weight ladder (L), April 2009 (A), May (M), June (J6), July (J7), August (A), September (S), October (O), November (N), December (D), January 2010 (J1); RNA no RT control (--), plasmid standard copy number from 10^X.

Supplemental Figure S2. Dicistrovirus Phylogeny.

Dicistrovirus IRES elements were aligned by ClustalW and a Neighbor-Joining tree generated by the Geneious Tree Builder (100 replicates). IAPV – Israel acute paralysis virus (NC009025), KBV – Kashmir bee virus (NC004807), ABPV – acute bee paralysis virus (NC002548), SINV1 – *Solenopsis invicta* virus 1 (NC006559), TSV – *Taura* syndrome virus (NC003005), ALPV – acute lethal paralysis virus (NC004365), ALPV strain Brookings (Q871932), RhPV –

Rhopalosiphum padi virus (NC001874), BSRV – Big Sioux River virus (JF423195-8), CrPV – cricket paralysis virus (NC003924), DCV – *Drosophila* C virus (NC001834), TV – *Triatoma* virus (NC003783), HPV – *Himetobi P* virus (NC003782), PSV – *Plautia Stali* intestine virus (NC003779), HCV – *Homalodisca coagulata* virus (NC008029), and BQCV – black queen cell virus (NC003784); red text – common honey bee viruses; blue text – novel viruses.

Supplemental Figure S3. RT-PCR results from pooled-monthly samples. (A) *Nosema apis* (268 bp), (B) deformed wing virus (DWV; 194 bp), (C) *Apocephalus borealis* (phorid fly; 500 bp), (D) *Apis mellifera* ribosomal protein L8 (Rpl8; 100 bp). Molecular weight ladder (L), April 2009 (A), May (M), June (J6), July (J7), August (A), September (S), October (O), November (N), December (D), January 2010 (J1); RNA only no RT control (--), water (H_2O), and positive control (+).

Supplemental Figure S4. Detection of the replicative form of LSV1 and LSV2 by negative strand-specific RT-PCR. The pooled July RNA sample was analyzed for the presence of LSV negative-strand RNA, which is indicative of virus replication, using strand-specific RT-PCR as described in Materials and Methods; RT-PCR products from reactions were analyzed by agarose (2%) gel electrophoresis.

Supplemental Figure S5 and S6. *Crithidia mellificae*, strain SF movies.

Light microscopy of live parasites was performed using a Leica DM6000 microscope (100x objective) equipped with Hamamatsu C4742-95 camera and Volocity Software (PerkinElmer).

Acknowledgments

We would like to thank Bret Adee and the staff at Adee Honey Farms, Inc. and Steve Thomas, San Francisco Beekeepers Association, for assistance with sampling; Christi Heintz, Executive Director Project *Apis m*., for facilitating this study; and members of the DeRisi and Andino laboratories for reviewing this manuscript. This work was supported by a research grant from Project *Apis m.* (www.projectapism.org) awarded to JDR JDR is supported by the Howard Hughes Medical Institute. C.R. is supported by at the Genentech Graduate Student Fellowship. M.L.F. is supported by an A.P. Giannini Foundation Medical Research Fellowship and research funds from UC-Davis's Häagen Dazs Postdoctoral Fellowship in Honey Bee Biology.

Author Contributions

Conceived and designed experiments: JDR., M.L.F. and C.R.

Performed the experiments, analyzed data: C.R., M.L.F., GR, and JDR

Wrote the paper: M.L.F., C.R., JDR., and R.A.

Table 1: Oligonucleotide targets for the Arthropod Pathogen Microarray

Dicistrovirus	**Total:264**
acute bee paralysis virus	38
black queen cell virus	42
Israel acute paralysis virus	26
Kashmir bee virus	42
other Dicistroviruses	116
Iflavirus	**Total:128**
deformed wing virus	22
honey bee slow paralysis virus	24
sacbrood virus	22
other Iflaviruses	60
Other Virus Families	**Total: 794**
Ascovirus	80
Baculovirus	138
Birnavirus	12
Cypovirus	98
Densovirus	110
Idnoreovirus	10
Iridovirus	46
Luteovirus	10
Nimavirus	20
Nodavirus	68
Okavirus	10
Poxvirus	74

Rhabdovirus	10
Tetravirus	30
Totivirus	10
Unassigned Virus Families	**Total: 88**
chronic bee paralysis virus	26
Solenopsis Invicta virus II	26
Acyrthospihon Pisum virus	12
Nora virus	12
kelp fly virus	12
Bacteria	**Total: 70**
Achromobacter	14
Paenibacillus	22
Melissococcus	10
Enterococcus	12
Wolbachia	6
Brevibaccilus	6
Fungi/Protists	**Total: 102**
Crithidia	20
Nosema	20
Ascophaera	10
Aspergillus	20
Metarhizium	20
Hirsutella	12
Mites	**Total: 80**
Varroa	32

Tropilaelaps	16
Acarapis	2
Nematodes	30
Positive Controls	**Total: 32**

Supplemental Table 1. Arthropod Pathogen Microarray test samples

Sample	**Arthropod Pathogen Microarray Results**
Overwinter Collapse (2007; MT)	BQCV, DWV, IAPV, KBV
***Apis mellifera* (CA)**	SBV
Varroa destructor	*Varroa*, DWV
Apis mellifera (drone; CA)	BQCV, DWV
Apis mellifera (Hive 53; CA - House Bees)	BQCV, KBV
Apis mellifera (Hive 53; CA - Foragers)	BQCV
Apis mellifera (Hive 42; CA - House Bees)	BQCV
Apis mellifera (Hive 42; CA - Foragers)	BQCV
***Bombus sp.* (CA)**	*Crithidia, Varroa*

Apis mellifera	
(Hive UCD; CA - Hairless Bee)	CBPV
Apis mellifera	
(Hive UCD; CA - Healthy Bee)	*Crithidia*
***Vespula* sp.**	
(Hive SM, CA, Feb. 2009)	ABPV, DWV, *Crithidia*
Apis mellifera	ABPV, DWV, BQCV,
(Hive SM, CA, Feb. 2009)	*Nosema*
Varroa destructor	
(Hive SM, CA Feb. 2009)	*Varroa*, DWV
Bee Bread (Healthy Hive Fall 2009)	positive controls
Bee Bread	
(CCD-affected Hive 1,Winter 2008/9**)**	DWV
Bee Bread	
(CCD-affected Hive 2,Winter 2008/9)	DWV, VDV-1
Overwinter Collapse (2010; OK)	DWV, SBV, *Nosema*
Overwinter Collapse (2010; OK)	BQCV, *Nosema*

Figure 1

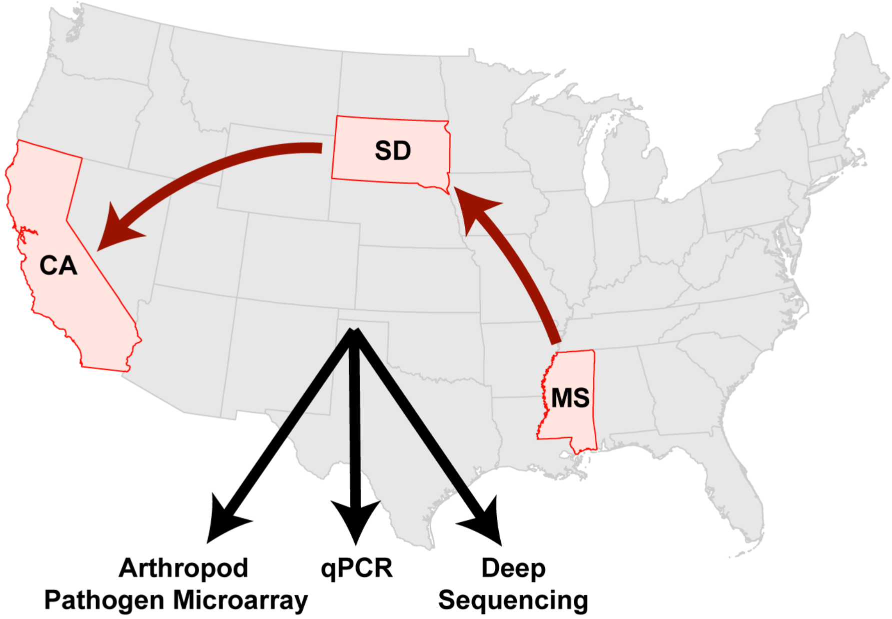
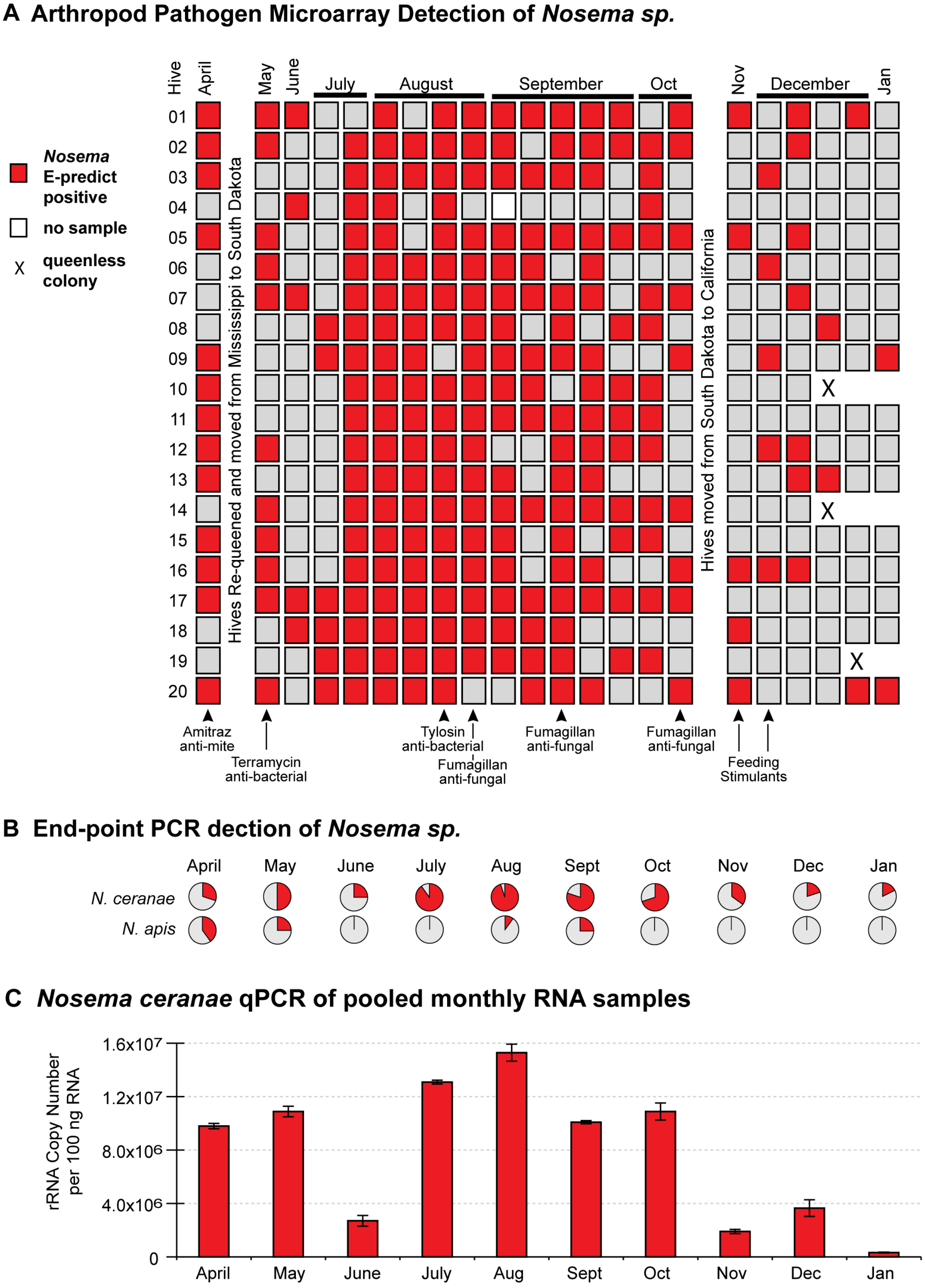

SD
CA
MS
Arthropod
Pathogen Microarray
qPCR
Deep
Sequencing

Figure 2

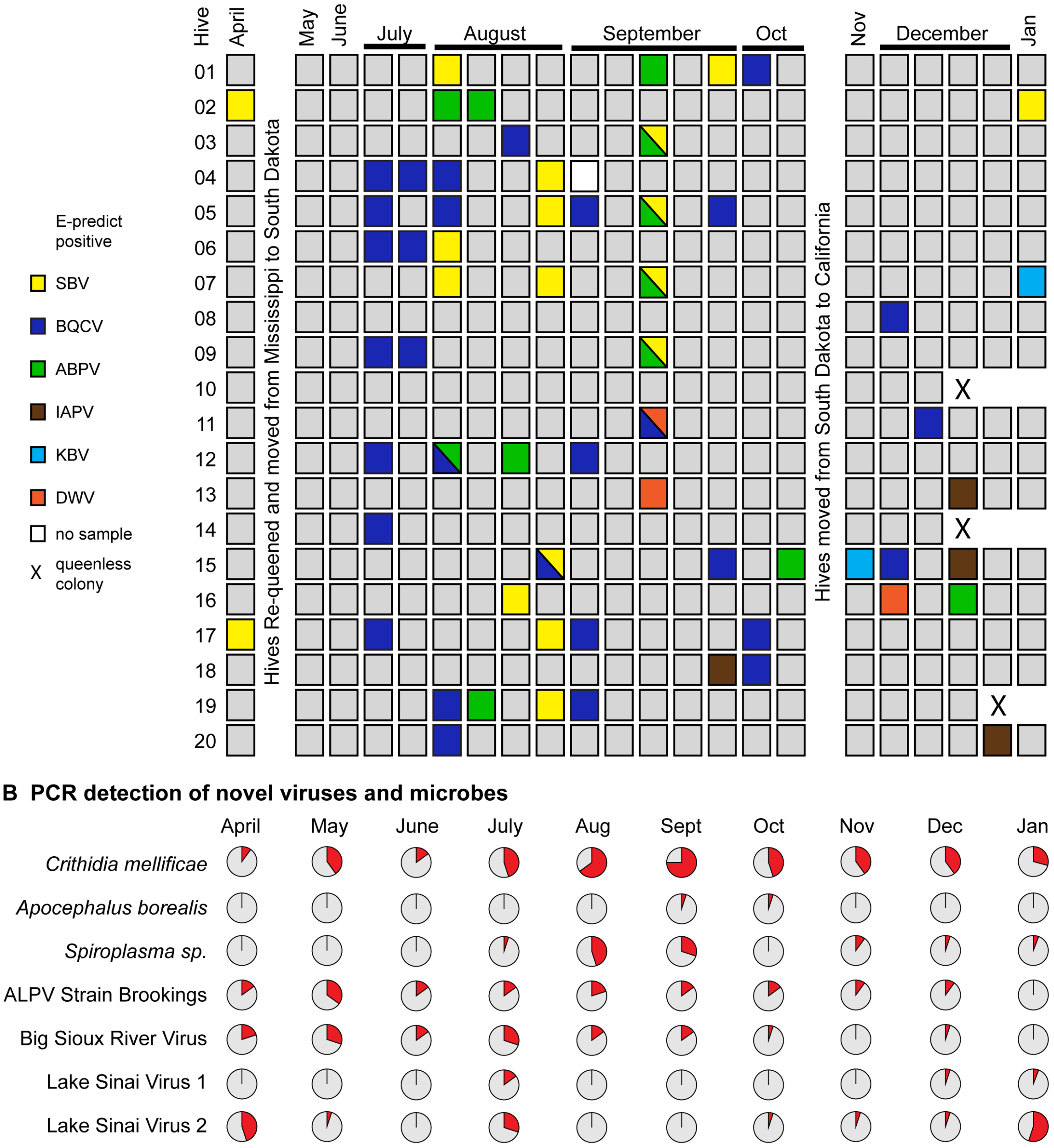


A Arthropod Pathogen Microarray Detection of *Nosema sp.*

B End-point PCR dection of *Nosema sp.*

C *Nosema ceranae* qPCR of pooled monthly RNA samples

Figure 3



A Arthropod Pathogen Microarray detection of common honey bee viruses

B PCR detection of novel viruses and microbes

Figure 4

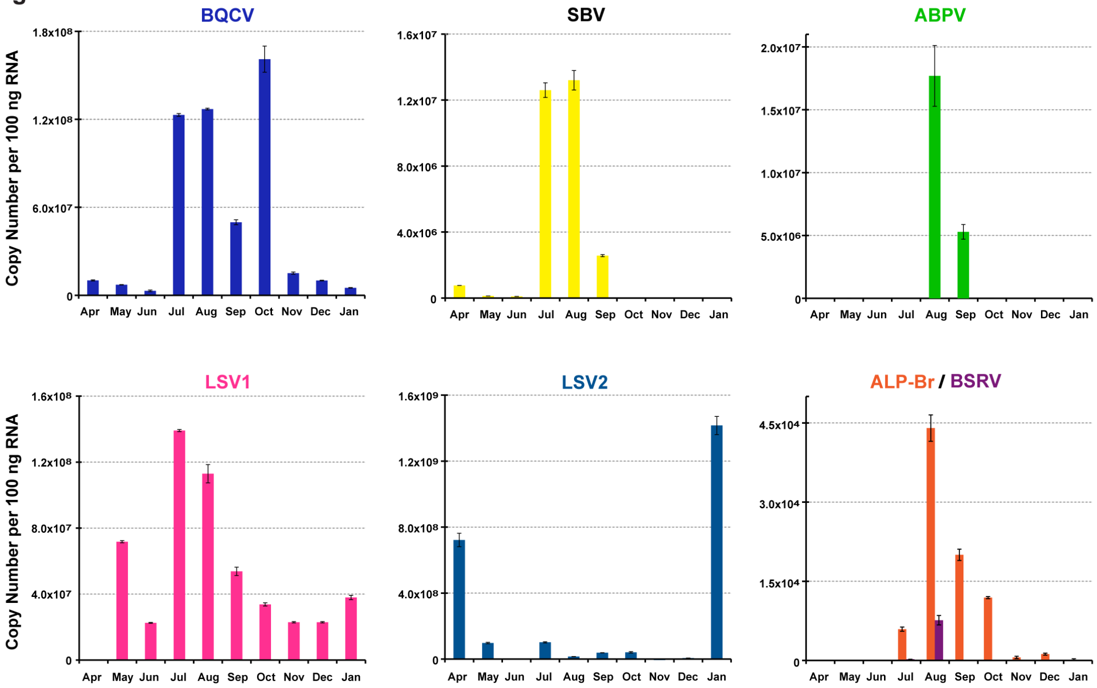

BQCV
SBV
ABPV
LSV1
LSV2
ALP-Br / BSRV
Copy Number per 100 ng RNA
Copy Number per 100 ng RNA
1.8x10^{8}
1.2x10^{8}
6.0x10^{7}
1.6x10^{7}
1.2x10^{7}
8.0x10^{6}
4.0x10^{6}
2.0x10^{7}
1.5x10^{7}
1.0x10^{7}
5.0x10^{6}
1.6x10^{8}
1.2x10^{8}
8.0x10^{7}
4.0x10^{7}
1.6x10^{9}
1.2x10^{9}
8.0x10^{8}
4.0x10^{8}
4.5x10^{4}
3.0x10^{4}
1.5x10^{4}
0
Apr May Jun Jul Aug Sep Oct Nov Dec Jan

Figure 5

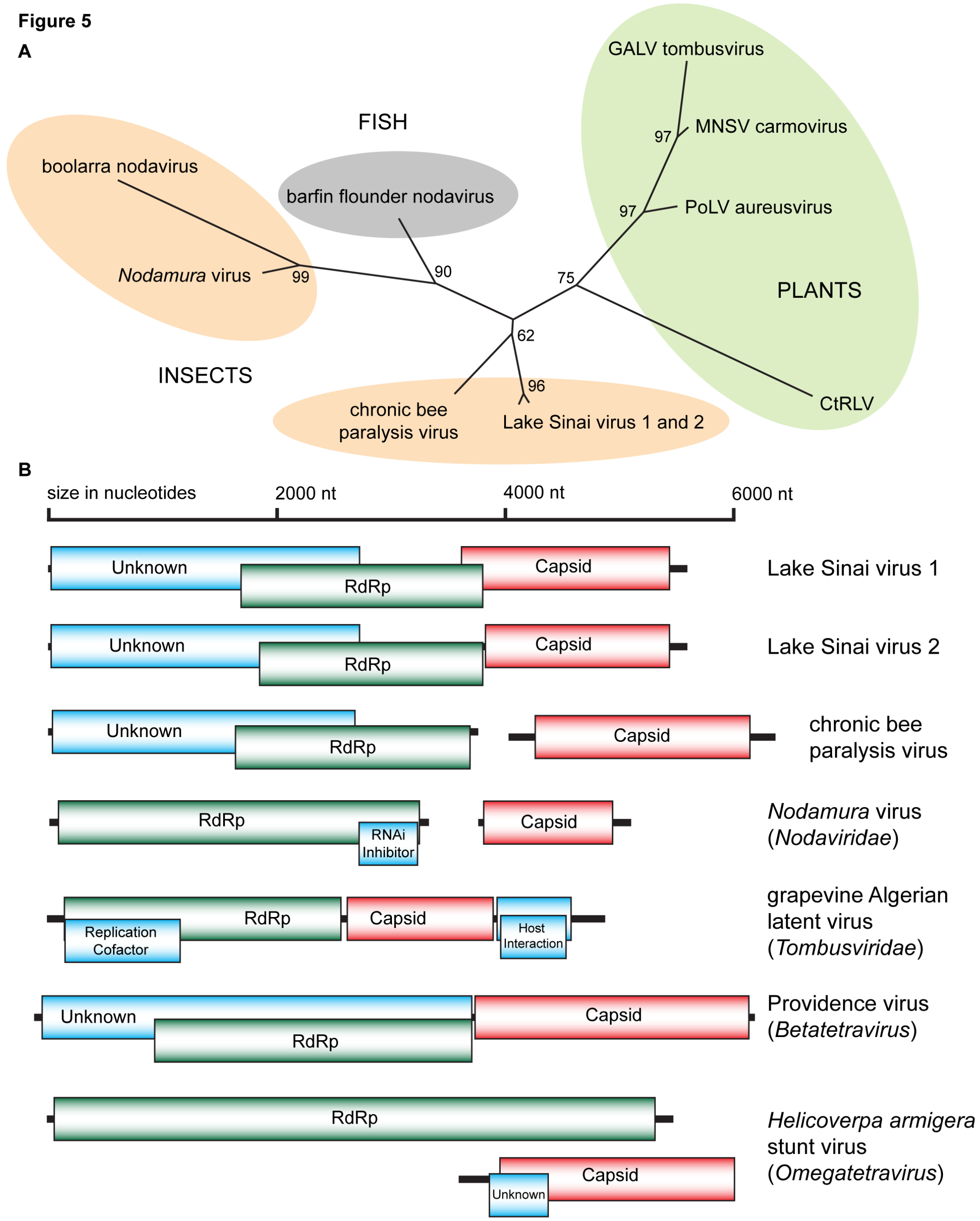
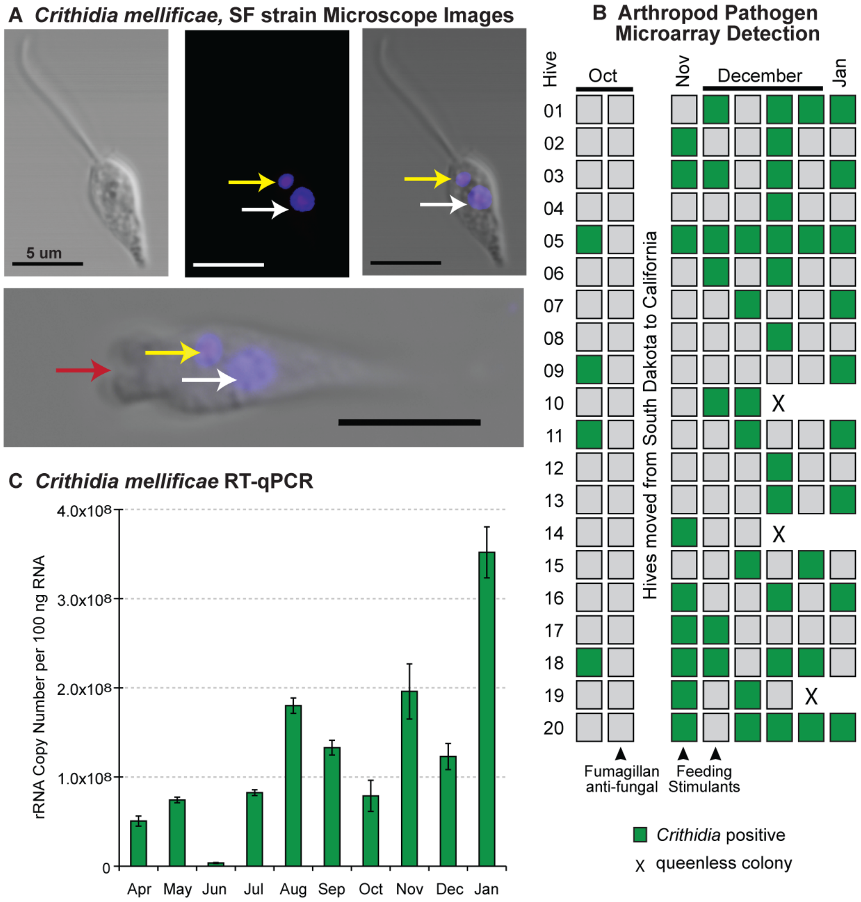

A
FISH
boolarra nodavirus
barfin flounder nodavirus
Nodamura virus
99
90
75
62
96
97
97
GALV tombusvirus
MNSV carmovirus
PoLV aureusvirus
PLANTS
INSECTS
chronic bee paralysis virus
Lake Sinai virus 1 and 2
CtRLV
B
size in nucleotides
2000 nt
4000 nt
6000 nt
Unknown
RdRp
Capsid
Lake Sinai virus 1
Unknown
RdRp
Capsid
Lake Sinai virus 2
Unknown
RdRp
Capsid
chronic bee paralysis virus
RdRp
RNAi Inhibitor
Capsid
Nodamura virus (Nodaviridae)
RdRp
Replication Cofactor
Capsid
Host Interaction
grapevine Algerian latent virus (Tombusviridae)
Unknown
RdRp
Capsid
Providence virus (Betatetravirus)
RdRp
Capsid
Unknown
Helicoverpa armigera stunt virus (Omegatetravirus)

Figure 6



A *Crithidia mellificae,* SF strain Microscope Images

B Arthropod Pathogen Microarray Detection

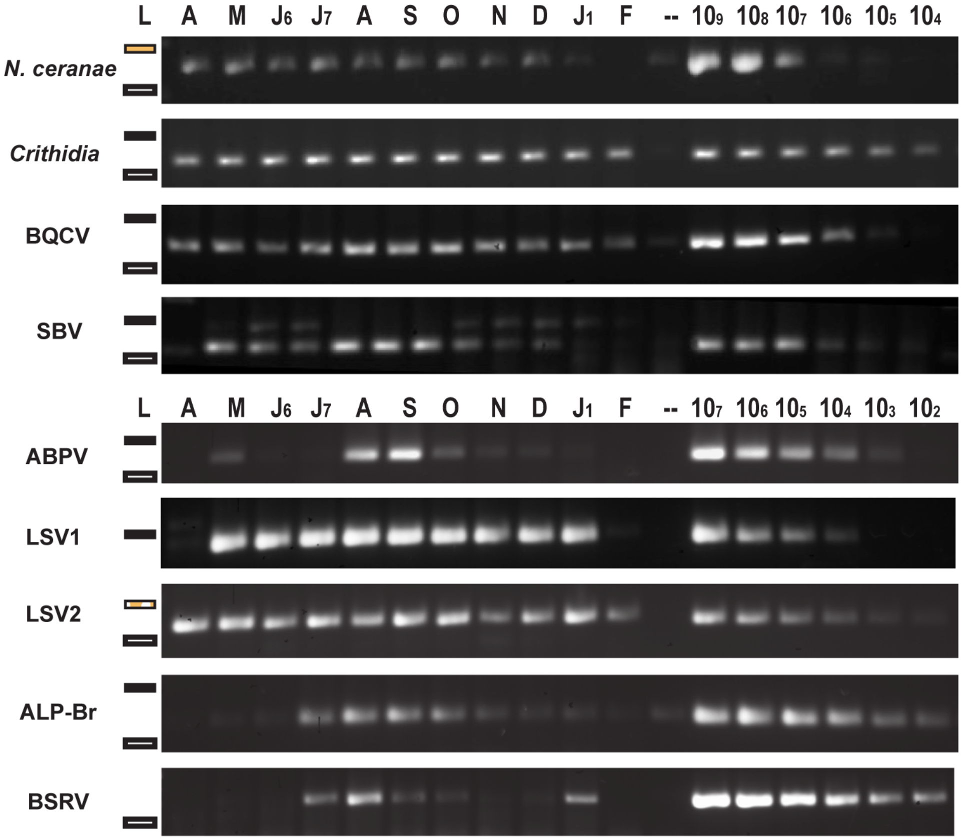
Hive	Oct		Nov	December				Jan
01	⬜	⬜	⬜	🟩	⬜	🟩	🟩	🟩
02	⬜	⬜	🟩	⬜	⬜	🟩	⬜	⬜
03	⬜	⬜	🟩	🟩	⬜	🟩	⬜	🟩
04	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜
05	🟩	⬜	🟩	🟩	🟩	🟩	🟩	🟩
06	⬜	⬜	⬜	🟩	⬜	🟩	⬜	⬜
07	⬜	⬜	⬜	⬜	🟩	⬜	⬜	🟩
08	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜
09	🟩	⬜	⬜	⬜	⬜	⬜	⬜	🟩
10	⬜	⬜	⬜	🟩	🟩	X		
11	🟩	⬜	⬜	⬜	🟩	⬜	⬜	🟩
12	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜
13	⬜	⬜	⬜	⬜	⬜	🟩	⬜	🟩
14	⬜	⬜	🟩	⬜	⬜	X		
15	⬜	⬜	⬜	⬜	🟩	⬜	🟩	⬜
16	⬜	⬜	🟩	⬜	⬜	🟩	⬜	🟩
17	⬜	⬜	🟩	🟩	⬜	⬜	⬜	⬜
18	🟩	⬜	🟩	🟩	⬜	🟩	🟩	⬜
19	⬜	⬜	🟩	⬜	🟩	⬜	X	
20	⬜	⬜	🟩	⬜	🟩	🟩	🟩	🟩

🟩 *Crithidia* positive

X queenless colony

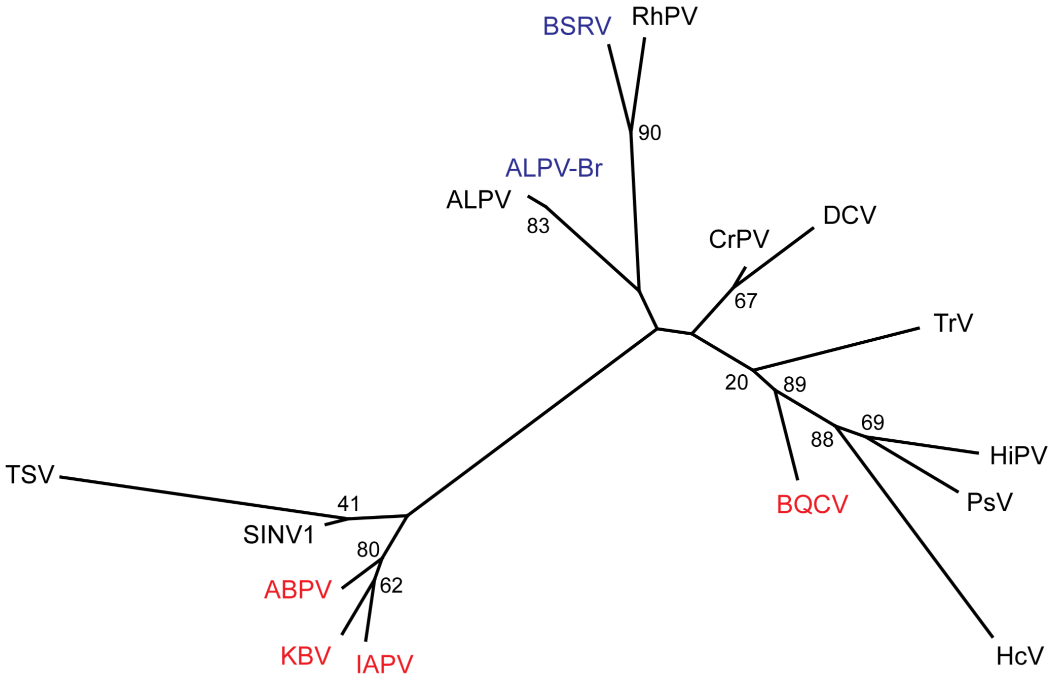
C *Crithidia mellificae* RT-qPCR

Supplemental Figure S1 Agarose gel electrophoresis of qPCR products from pooled-monthly honey bee RNA samples

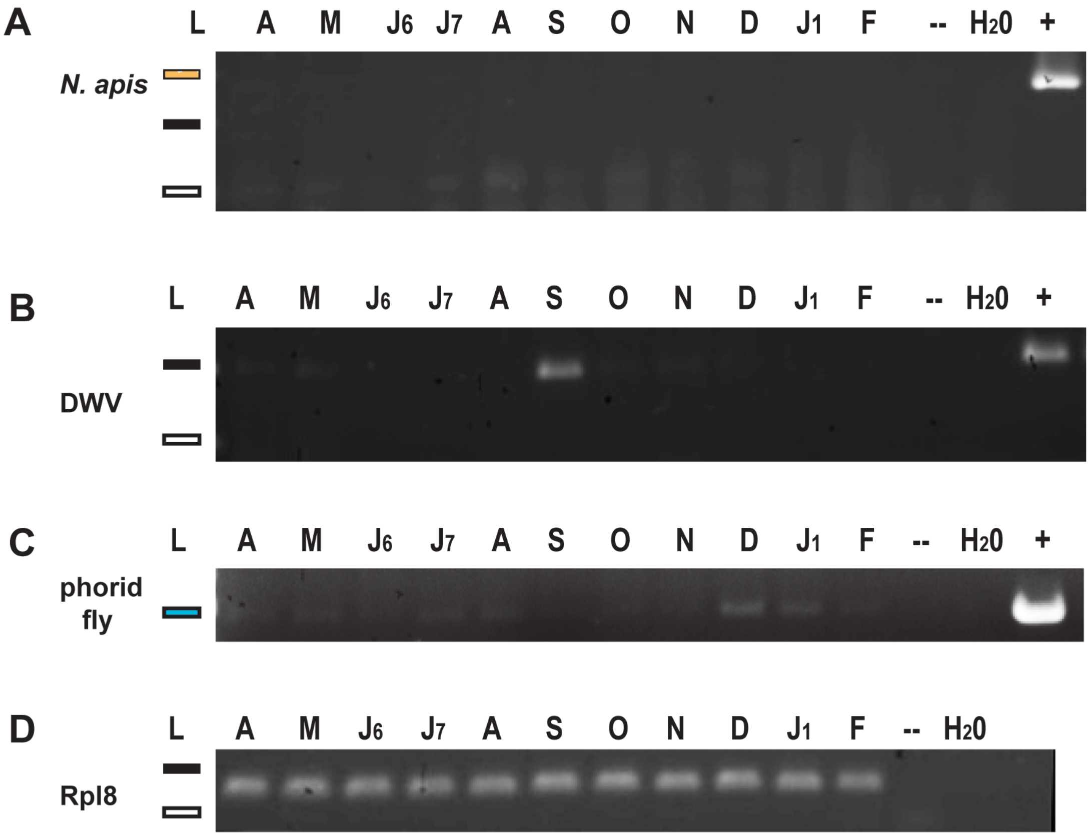

L A M J_6 J_7 A S O N D J_1 F -- 10_9 10_8 10_7 10_6 10_5 10_4
N. ceranae
Crithidia
BQCV
SBV
L A M J_6 J_7 A S O N D J_1 F -- 10_7 10_6 10_5 10_4 10_3 10_2
ABPV
LSV1
LSV2
ALP-Br
BSRV

Molecular weight ladder (L), 🟧 = 300 bp, ▬ = 200 bp, ▭ = 100 bp.
April 2009 (A), May (M), June (J6), July (J7), August (A), September (S),
October (O), November (N), December (D), January 2010 (J1);
RNA only no RT control (--), plasmid standard copy number from 10^#.

Supplemental Figure S2 Phylogeny of the Dicistroviridae, derived from a nucleotide alignment of the internal IRES

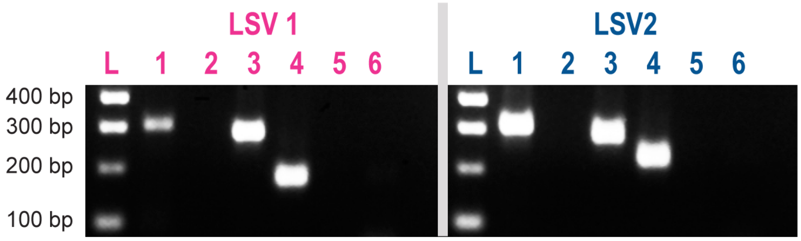


Supplemental Figure S3
Agarose gel electrophoresis of RT-PCR products from pooled-monthly samples


A
L A M J6 J7 A S O N D J1 F -- H20 +
N. apis
B
L A M J6 J7 A S O N D J1 F -- H20 +
DWV
C
L A M J6 J7 A S O N D J1 F -- H20 +
phorid fly
D
L A M J6 J7 A S O N D J1 F -- H20
Rpl8

Molecular weight ladder (L), 🟦 = 500 bp, 🟧 = 300 bp, ▬ = 200 bp, ▭ = 100 bp

Supplemental Figure S4
Detection of the replicative form of LSV1 and LSV2 by negative strand-specific RT-PCR



(L) Molecular weight ladder,
(1) LSV negative strand amplification using tagged-negative strand primed cDNA template and TAGS forward and LSVU-R-1744 PCR primers,
(2) Negative control - unprimed RT reaction amplified using TAGS forward and LSVU-R1744 PCR primers,
(3) Positive contol - random hexamer primed cDNA amplified using LSV1 or LSV2 -specific forward primer and LSVUR-1744 reverse primer,
(4) Postivie control - random hexamer primed cDNA amplified using LSV-specific qPCR primer sets,
(5) Negative control - LSV tagged negative-strand primed cDNA template in PCR reaction in which only the LSV-U-R1744 primer was added,
(6) Negative control - no template PCR using LSV qPCR primer sets

REFERENCES

1. Cornman RS, Chen YP, Schatz MC, Street C, Zhao Y, et al. (2009) Genomic analyses of the microsporidian Nosema ceranae, an emergent pathogen of honey bees. Plos Pathog 5: e1000466.
2. Pennisi E (2006) Genetics - Honey bee genome illuminates insect evolution and social behavior. Science 314: 578-579.
3. Winston ML (1987) The biology of the Honey Bee: First Harvard Press.
4. Johnson R (2010) Honey Bee Colony Collapse Disorder.
5. USDA-ARS Questions and Answers: Colony Collapse Disorder
6. Morse RA, Calderone NW (2000) The Value of Honey Bees as Pollinators of U.S. Crops in 2000. Gleanings in Bee Culture Supple 1-15.
7. Vanengelsdorp D, Underwood R, Caron D, Hayes J (2007) An estimate of managed colony losses in the winter of 2006-2007: A report commissioned by the apiary inspectors of America. Am Bee J 147: 599-603.
8. van Engelsdorp D, Hayes J, Caron D, Pettis J (2010) Preliminary Results: Honey Bee Colonies Losses in the U.S., winter 2009-2010. Apiary Inspectors of America (AIA) and USDA-ARS Beltsville Honey Bee Lab.
9. Vanengelsdorp D, Hayes J, Underwood RM, Pettis J (2008) A Survey of Honey Bee Colony Losses in the US, Fall 2007 to Spring 2008. PLoS ONE 3: e4071.

10. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, et al. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318: 283-287.

11. Vanengelsdorp D, Evans JD, Saegerman C, Mullin C, Haubruge E, et al. (2009) Colony collapse disorder: a descriptive study. PLoS ONE 4: e6481.

12. Martinson VG, Danforth BN, Minckley RL, Rueppell O, Tingek S, et al. (2011) A simple and distinctive microbiota associated with honey bees and bumble bees. Mol Ecol 20: 619-628.

13. Govan VA, Leat N, Allsopp M, Davison S (2000) Analysis of the complete genome sequence of acute bee paralysis virus shows that it belongs to the novel group of insect-infecting RNA viruses. Virology 277: 457-463.

14. Leat N, Ball B, Govan V, Davison S (2000) Analysis of the complete genome sequence of black queen-cell virus, a picorna-like virus of honey bees. J Gen Virol 81: 2111-2119.

15. Maori E, Lavi S, Mozes-Koch R, Gantman Y, Peretz Y, et al. (2007) Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination. J Gen Virol 88: 3428-3438.

16. de Miranda JR, Drebot M, Tyler S, Shen M, Cameron CE, et al. (2004) Complete nucleotide sequence of Kashmir bee virus and comparison with acute bee paralysis virus. J Gen Virol 85: 2263-2270.

17. Lanzi G, de Miranda JR, Boniotti MB, Cameron CE, Lavazza A, et al. (2006) Molecular and biological characterization of deformed wing virus of honeybees (Apis mellifera L.). J Virol 80: 4998-5009.

18. Ghosh RC, Ball BV, Willcocks MM, Carter MJ (1999) The nucleotide sequence of sacbrood virus of the honey bee: an insect picorna-like virus. J Gen Virol 80 (Pt 6): 1541-1549.

19. Olivier V, Blanchard P, Chaouch S, Lallemand P, Schurr F, et al. (2008) Molecular characterisation and phylogenetic analysis of Chronic bee paralysis virus, a honey bee virus. Virus Res 132: 59-68.

20. Chen YP, Siede R (2007) Honey bee viruses. Adv Virus Res 70: 33-80.

21. Baker A, Schroeder D (2008) Occurrence and genetic analysis of picorna-like viruses infecting worker bees of Apis mellifera L. populations in Devon, South West England. J Invertebr Pathol 98: 239-242.

22. Chen YP, Pettis JS, Collins A, Feldlaufer MF (2006) Prevalence and transmission of honeybee viruses. Appl Environ Microbiol 72: 606-611.

23. Klee J, Besana AM, Genersch E, Gisder S, Nanetti A, et al. (2007) Widespread dispersal of the microsporidian Nosema ceranae, an emergent pathogen of the western honey bee, Apis mellifera. J Invertebr Pathol 96: 1-10.

24. Fries I (2010) Nosema ceranae in European honey bees (Apis mellifera). J Invertebr Pathol 103 Suppl 1: S73-79.

25. Higes M, García-Palencia P, Martín-Hernández R, Meana A (2007) Experimental infection of Apis mellifera honeybees with Nosema ceranae (Microsporidia). J Invertebr Pathol 94: 211-217.

26. Higes M, Martín-Hernández R, Botías C, Bailón EG, González-Porto AV, et al. (2008) How natural infection by Nosema ceranae causes honeybee colony collapse. Environ Microbiol 10: 2659-2669.

27. Martín-Hernández R, Meana A, Prieto L, Salvador AM, Garrido-Bailón E, et al. (2007) Outcome of colonization of Apis mellifera by Nosema ceranae. Appl Environ Microbiol 73: 6331-6338.

28. Chen Y, Evans JD, Smith IB, Pettis JS (2008) Nosema ceranae is a long-present and wide-spread microsporidian infection of the European honey bee (Apis mellifera) in the United States. J Invertebr Pathol 97: 186-188.

29. Aronstein KA, Murray KD (2010) Chalkbrood disease in honey bees. J Invertebr Pathol 103: S20-S29.

30. Qin X, Evans JD, Aronstein KA, Murray KD, Weinstock GM (2006) Genome sequences of the honey bee pathogens Paenibacillus larvae and Ascosphaera apis. Insect Mol Biol 15: 715-718.

31. Forsgren E (2010) European foulbrood in honey bees. J Invertebr Pathol 103 Suppl 1: S5-9.

32. Genersch E (2010) American Foulbrood in honeybees and its causative agent, Paenibacillus larvae. J Invertebr Pathol 103: S10-S19.

33. Forsgren E, Lundhagen AC, Imdorf A, Fries I (2005) Distribution of Melissococcus plutonius in honeybee colonies with and without symptoms of European foulbrood. Microb Ecol 50: 369-374.

34. de Graaf DC, Alippi AM, Brown M, Evans JD, Feldlaufer M, et al. (2006) Diagnosis of American foulbrood in honey bees: a synthesis and proposed analytical protocols. Lett Appl Microbiol 43: 583-590.

35. Rosenkranz P, Aumeier P, Ziegelmann B (2010) Biology and control of Varroa destructor. J Invertebr Pathol 103: S96-S119.

36. Sammataro D, Gerson U, Needham G (2000) Parasitic mites of honey bees: life history, implications, and impact. Annu Rev Entomol 45: 519-548.

37. Boncristiani HF, Di Prisco G, Pettis JS, Hamilton M, Chen YP (2009) Molecular approaches to the analysis of deformed wing virus replication and pathogenesis in the honey bee, Apis mellifera. Virol J 6: 221.

38. Chen YP, Pettis JS, Evans JD, Kramer M, Feldlaufer MF (2004) Transmission of Kashmir bee virus by the ectoparasitic mite Varroa destructor. Apidologie 35: 441-448.

39. Shen MQ, Cui LW, Ostiguy N, Cox-Foster D (2005) Intricate transmission routes and interactions between picorna-like viruses (Kashmir bee virus and sacbrood virus) with the honeybee host and the parasitic varroa mite. Journal of General Virology 86: 2281-2289.

40. Delaney DA, Meixner MD, Schiff NM, Sheppard WS (2009) Genetic Characterization of Commercial Honey Bee (Hymenoptera: Apidae)

Populations in the United States by Using Mitochondrial and Microsatellite Markers. Ann Entomol Soc Am 102: 666-673.

41. Tarpy D, Seeley T (2006) Lower disease infections in honeybee (Apis mellifera) colonies headed by polyandrous vs monandrous queens. Naturwissenschaften 93: 195-199.

42. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. (2002) Microarray-based detection and genotyping of viral pathogens. P Natl Acad Sci Usa 99: 15687-15692.

43. Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, et al. (2003) Viral discovery and sequence recovery using DNA microarrays. Plos Biol 1: E2.

44. Chiu CY, Rouskin S, Koshy A, Urisman A, Fischer K, et al. (2006) Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. Clin Infect Dis 43: e71-76.

45. Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, et al. (2006) Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. Plos Pathog 2: e25.

46. Chiu CY, Alizadeh AA, Rouskin S, Merker JD, Yeh E, et al. (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus microarray. J Clin Microbiol 45: 2340-2343.

47. Kistler A, Avila PC, Rouskin S, Wang D, Ward T, et al. (2007) Pan-viral screening of respiratory tract infections in adults with and without asthma

reveals unexpected human coronavirus and human rhinovirus diversity. J Infect Dis 196: 817-825.

48. Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, et al. (2008) Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections. P Natl Acad Sci Usa 105: 14124-14129.

49. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, et al. (2008) Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. Virol J 5: 88.

50. Buchen-Osmond C (2003) The universal virus database ICTVdB. Computing in Science & Engineering 5: 16-25.

51. Medicine USNLo National Center for Biotechnology Information (NCBI). 8600 Rockville Pike, Bethesda MD, 20894 USA: U.S. National Library of Medicine.

52. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 4: R9.

53. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. Genome Biol 6: R78.

54. de Miranda JR, Genersch E (2010) Deformed wing virus. J Invertebr Pathol 103: S48-S61.

55. Ribiere M, Olivier V, Blanchard P (2010) Chronic bee paralysis: A disease and a virus like no other? J Invertebr Pathol 103: S120-S131.

56. Williams GR, Sampson MA, Shutler D, Rogers REL (2008) Does fumagillin control the recently detected invasive parasite Nosema ceranae in western honey bees (Apis mellifera)? J Invertebr Pathol 99: 342-344.

57. Gauthier L, Tentcheva D, Tournaire M, Dainat B, Cousserans F, et al. (2007) Viral load estimation in asymptomatic honey bee colonies using the quantitative RT-PCR technique. Apidologie 38: 426-U427.

58. Highfield AC, El Nagar A, Mackinder LCM, Noël LM-LJ, Hall MJ, et al. (2009) Deformed wing virus implicated in overwintering honeybee colony losses. Appl Environ Microbiol 75: 7212-7220.

59. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

60. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36: D25-30.

61. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244-248.

62. Craggs JK, Ball JK, Thomson BJ, Irving WL, Grabowska AM (2001) Development of a strand-specific RT-PCR based assay to detect the replicative form of hepatitis C virus RNA. J Virol Methods 94: 111-120.

63. Plaskon NE, Adelman ZN, Myles KM (2009) Accurate strand-specific quantification of viral RNA. PLoS ONE 4: e7468.

64. Brown M, Schmid-Hempel R, Schmid-Hempel P (2003) Strong context-dependent virulence in a host-parasite system: reconciling genetic evidence with theory. J Anim Ecol 72: 994-1002.

65. Langridge D, McGhee R (1967) Crithidia Mellificae n. sp. an Acidophilic Trypanosomatid of the Honey Bee Apis Mellifera. J Protozool 14: 485-&.

66. Mouches C, Bove J, Tully J, Rose D, McCoy R, et al. (1983) Spriroplasma-Apis, a new species from the honeybee Apis mellifera. Ann Inst Pasteur Mic A134: 383-397.

67. Clark T, Whitcomb R, Tully J, Mouches C, Saillard C, et al. (1985) Spiromplasma melliferum, a new species from the honeybee (Apis mellifera). Int J Syst Bacteriol 35: 296-308.

68. Brown BV (1993) Taxonomy and Preliminary Phylogeny of the Parasitic Genus Apocephalus, Subgenus Mesophora (Diptera, Phoridae). Systematic Entomology 18: 191-230.

69. Core A, Runckel C, Ivers J, Quock C, DeNault S, et al. (2011) Phorid-parasitized Honeybee Take Flight in the Dead of Night. San Francisco, CA: San Francisco State University.

70. Otterstatter MC, Whidden TL, Owen RE (2002) Contrasting frequencies of parasitism and host mortality among phorid and conopid parasitoids of bumble-bees. Ecological Entomology 27: 229-237.

71. Johnson RM, Evans JD, Robinson GE, Berenbaum MR (2009) Changes in transcript abundance relating to colony collapse disorder in honey bees (Apis mellifera). P Natl Acad Sci Usa 106: 14790-14795.

72. Welch A, Drummond F, Tewari S, Averill A, Burand JP (2009) Presence and prevalence of viruses in local and migratory honeybees (Apis mellifera) in Massachusetts. Appl Environ Microbiol 75: 7862-7865.

73. Genersch E, von der Ohe W, Kaatz H, Schroeder A, Otten C, et al. (2010) The German bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies. Apidologie 41: 332-352.

74. Tentcheva D, Gauthier L, Zappulla N, Dainat B, Cousserans F, et al. (2004) Prevalence and seasonal variations of six bee viruses in Apis mellifera L. and Varroa destructor mite populations in France. Appl Environ Microbiol 70: 7185-7191.

75. Bourgeois AL, Rinderer TE, Beaman LD, Danka RG (2010) Genetic detection and quantification of Nosema apis and N. ceranae in the honey bee. J Invertebr Pathol 103: 53-58.

76. Gisder S, Hedtke K, Möckel N, Frielitz M-C, Linde A, et al. (2010) Five-year cohort study of Nosema spp. in Germany: does climate shape virulence and assertiveness of Nosema ceranae? Appl Environ Microbiol 76: 3032-3038.

77. Paxton RJ, Klee J, Korpela S, Fries I (2007) Nosema ceranae has infected Apis mellifera in Europe since at least 1998 and may be more virulent than Nosema apis. Apidologie 38: 558-565.

78. Hamiduzzaman MM, Guzman-Novoa E, Goodwin PH (2010) A multiplex PCR assay to diagnose and quantify Nosema infections in honey bees (Apis mellifera). J Invertebr Pathol.

79. Cameron SA, Lozier JD, Strange JP, Koch JB, Cordes N, et al. (2011) Patterns of widespread decline in North American bumble bees. P Natl Acad Sci Usa.

80. Foster LJ (2011) Interpretation of data underlying the link between CCD and an invertebrate iridescent virus. Mol Cell Proteomics.

81. Knudsen G, Chalkley RJ (2011) Commentary on Bromenshenk et al: The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis. submitted PLoS ONE.

82. Bromenshenk JJ, Henderson CB, Wick CH, Stanford MF, Zulich AW, et al. (2010) Iridovirus and Microsporidian Linked to Honey Bee Colony Decline. Plos One 5: e13181.

83. Bailey L, Ball BV, Perry JN (1981) The Prevalence of Viruses of Honey Bees in Britain. Annals of Applied Biology 97: 109-118.

84. Singh R, Levitt AL, Rajotte EG, Holmes EC, Ostiguy N, et al. (2010) RNA Viruses in Hymenopteran Pollinators: Evidence of Inter-Taxa Virus Transmission via Pollen and Potential Impact on Non-Apis Hymenopteran Species. PLoS ONE 5: e14357.

85. DeGrandi-Hoffman G, Chen Y, Huang E, Huang MH (2010) The effect of diet on protein concentration, hypopharyngeal gland development and virus

load in worker honey bees (Apis mellifera L.). J Insect Physiol 56: 1184-1191.

86. Maori E, Paldi N, Shafir S, Kalev H, Tsur E, et al. (2009) IAPV, a bee-affecting virus associated with Colony Collapse Disorder can be silenced by dsRNA ingestion. Insect Mol Biol 18: 55-60.

87. Chen Y, Evans JD (2007) Historical presence of Israeli acute paralysis virus in the United States. Am Bee J 147: 1027-1028.

88. Palacios G, Hui J, Quan PL, Kalkstein A, Honkavuori KS, et al. (2008) Genetic analysis of Israel acute paralysis virus: Distinct clusters are circulating in the United States. J Virol 82: 6209-6217.

89. Chen YP, Higgins JA, Feldlaufer MF (2005) Quantitative real-time reverse transcription-PCR analysis of deformed wing virus infection in the honeybee (Apis mellifera L.). Applied and Environmental Microbiology 71: 436-441.

90. Shen MQ, Yang XL, Cox-Foster D, Cui LW (2005) The role of varroa mites in infections of Kashmir bee virus (KBV) and deformed wing virus (DWV) in honey bees. Virology 342: 141-149.

91. Berényi O, Bakonyi T, Derakhshifar I, Köglberger H, Nowotny N (2006) Occurrence of six honeybee viruses in diseased Austrian apiaries. Appl Environ Microbiol 72: 2414-2420.

92. Yourth CP, Brown MJF, Schmid-Hempel P (2008) Effects of natal and novel Crithidia bombi (Trypanosomatidae) infections on Bombus terrestris hosts. Insect Soc 55: 86-90.

93. Eisen MB, Brown PO (1999) DNA arrays for analysis of gene expression. Meth Enzymol 303: 179-205.

94. Engel JC, Parodi AJ (1985) Trypanosoma cruzi cells undergo an alteration in protein N-glycosylation upon differentiation. J Biol Chem 260: 10105-10110.

95. Drummond AJ, Ashton B, Buxton S CM, Cooper A, Heled J, et al. (2010) Geneious v5.1, Available from http://www.geneious.com.

96. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, et al. (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35: W71-74.

Chapter 4 Preface

This work represents the mature development of the viral discovery techniques described in Chapter 3, and applies them in a high-throughput and parallel way. At the outset of this project, the assumption is made that not only will some novel viruses be discovered, but that enough novel species will be observed to make statements regarding the distribution of host and pathogen, as well as strong odds of identifying at least some inter-family recombinants. This is a departure from previous virus discovery efforts, but just as the scope of projects shifted with new technology from single genes to gene pathways and families, so too will virome projects shift to communities and viral families. With the novel recombinants identified in this project, the previous work and other studies in the literature, many phylogenetically distinct inter-family recombination events can now be described and patterns in the location and conditions of recombination can be proposed.

High-throughput viral discovery in the clade Hymenoptera reveals inter-family recombination events and novel virus families

Runckel C[1], Fisher BL[2], DeRisi JL[1]

[1] Howard Hughes Medical Institute, Departments of Medicine, Biochemistry and Biophysics, and Microbiology, University of California, San Francisco, CA 94143;
[2] California Academy of Sciences, San Francisco, CA 94143

Abstract

Hymenopterans are a diverse group of social and non-social insects known to be frequently infected with RNA viruses. These viruses have overwhelmingly derived from the Picornavirales, suggesting a unique susceptibility either of this clade or of social organisms in general to viruses of this order. We employ deep sequencing techniques to discover 43 new virus species in 20 different families of hymenopterans, including social and non-social species. Several inter-family recombinants within the Nodavirus-like superfamily and between that superfamily and the Tetraviridae are observed. Picorna-like viruses account for one-third of new virus species observed and social hymenopterans were associated ($p=0.009$) with picornavirus-like species compared to non-social hymenopterans. Despite the large number of novel picornavirus-like species observed, we did not identify any inter-family recombinants, in contrast to the Nodavirales. We also identify novel negative-strand RNA viruses, ssDNA viruses and dsRNA viruses in hymenopterans for the first time, as well as a new family of positive-strand RNA viruses, the Lasiviridae.

Introduction

Hymenoptera is a broad order of insects including wasps, bees, ants and sawflies. The clade is notable for the frequent utilization of haplo-diploid sex determination and the evolution of eusociality on multiple occasions. Hymenopterans exhibit a number of colonial lifestyles, including a single fertile queen (monogyny), many queens in a colony (polygyny), and supercolonies (large contiguous colonies stretching for hundreds of miles)[1,2]. These lifestyles result in dense populations of closely related individuals including sisters, half-sisters or cousins, in a manner rarely observed in the animal kingdom. Such arrangements could be conducive to disease transmission and even pandemics, such as the worldwide spread from east Asia of mites, viruses and microsporidians infecting honeybees[3,4].

Numerous studies of pollinators and agricultural pests have led to the identification of viral and non-viral pathogens, and surveys have indicated that bees and ants are frequently infected with high loads of virus[5,6]. Most hymenoptera viruses are close relatives of the family picornaviridae, including the families dicistroviridae and iflaviridae, and are thus positive strand, poly-adenylated RNA viruses. In contrast, few negative- or double-strand RNA viruses or DNA viruses are known as pathogens of hymenopterans. This may be due to a biological predisposition to ssRNA virus infection, or more likely to methodological bias in discovery. Early research into bee viruses relied on isolation and propagation by microinjection of bee homogenates[7,8] and was biased to detect viruses causing paralysis or deformity, favoring fast replicating lytic viruses. In the genomic era, most new hymenopteran viruses have been identified serendipitously from transcriptome sequencing projects[9–11]. The first generation of sequencing projects relied on poly-A tail selection of sample RNAs, again favoring the discovery of poly-adenylated viruses. Both research models bias towards the discovery of picorna-like viruses. Virus families with broad host ranges covering most insect families and the related subphylum Crustacea have not been identified in hymenopterans, including nodaviruses and densoviruses. We set out to characterize the viral flora of hymenopterans from social and non-social species using both unbiased deep sequencing techniques and techniques intended to bias for previously under-represented virus families. Our intention was to determine if picorna-like viruses are indeed more common than other virus-types in Hymenoptera and to observe any patterns in the viral community in social versus non-social insects of this order.

Results

Sample Set

Samples for the pilot study were collected opportunistically and were all from social species due to their ease of collection and identification. Whole samples or extracted nucleic acid were stored in several different methods to assess sequencing quality prior to a larger survey collection involving social and non-social species. Samples included “Yellow-jackets” (*Vespula sp*., from Oregon and Idaho), Asiatic and Dwarf honeybees (*Apis ceranae* and *A. florea*, from Thailand and Vietnam), and five ant species from California (see table 1). Once an efficient sample processing protocol was established, social and non-social samples were collected live and extracted the same day in northern California. Care was taken to sample as many different and distantly related families of the order Hymenoptera as possible; in total 20 different families are represented in this survey.

Discovery Pipeline

We developed a modified virus discovery pipeline based on two principles: first, to enrich a sample for viral nucleic acid with multiple molecular biology techniques as opposed to standard deep sequencing methodology with a single unbiased library, and second to detect and assemble complete viral genomes by targeted and iterative search and assembly from short deep sequencing reads. Current methods of sequence-based virus discovery generally rely on a single library preparation technique due to the difficulty or cost of library generation[12,13].

Ideally, ribosome subtraction techniques are employed but frequently virus discovery efforts are serendipitous, or otherwise are the result of mining transcriptome data acquired for other purposes; these libraries are overwhelming generated by poly-A selection of mRNA[14] and as such are biased against numerous virus families which do not have poly-adenylated genomes or intermediates.

We employed four different enrichment technqiues: microccocal nuclease treatment for encapsidated nucleic acid, dT priming of RNA to select for poly-adenylated RNAs, 5'-monophosphate exoribonuclease treatment to remove rRNAs while leaving non-poly-A RNAs and reverse transcription of RNA without an added primer. This last technique represents a specific viral enrichment, due to the fact that some virus clades use self-priming secondary or repeat structures and thus efficiently reverse transcribe themselves. Secondary structures are also observed to cleave by hydrolysis at the 3' end of internal hairpins, leaving a suitable priming site and selecting for highly-structured RNAs, which favors viral enrichment. We note that none of these techniques is universal for all viruses and some will efficiently suppress certain viral clades; taken together this four-pronged enrichment was however extremely effective in enriching viral sequences, to over half of total reads in some datasets.

Conventional viral analysis of deep sequencing data relies on undirected assembly of short reads and subsequent detection of viral sequences by BLAST or other alignment algorithms. Unbiased assembly is especially difficult in transcriptome datasets due to variations in coverage level by transcript[14,15]. We employ an iterative search and assembly technique to detect individual short reads of interest and grow them by targeted assembly, followed by additional search and assembly steps. Significance thresholds are initially set low and raised with each step, allowing assembled sequences with <20% amino acid divergence to be reliably detected. tBLASTx[16] and HMMER[17] provided different methods of identifying divergent sequences. A conventional analysis pipeline was run in parallel using the Trinity assembler for unguided assembly followed by tBLASTx and HMMER; the iterative pipeline identified all the novel viruses the conventional pipeline detected, while the conventional pipeline detected only 79% of the novel viral species identified by the iterative pipeline.

Nodavirus-like Superfamily

The family nodaviridae consists of the broad and diverse genus Alphanodavirus, infecting insects and crustaceans, and the highly conserved Betanodaviruses, which infect fish (Figure 1). Nodaviruses are simple bipartite viruses with polymerase and capsid genes on different RNA segments and additional small orfs, in some cases encoding RNAi inhibitors[18], overlapping the core genes. Distantly related to nodaviruses are the diverse plant viruses of the Tombusviridae, differing in organization with monopartite genome organization and employing sub-genomic RNAs to produce capsid specific transcripts[19]. Recently, the number and diversity of nodaviruses and related, unassigned viruses have expanded due to deep sequencing-based viral discoveries. These include the tetnoviruses[20], distantly related to nodaviruses by amino acid similarity but possessing a monopartite genome organization, and nodaviruses of nematodes[21], which appear to preserve the standard bipartite genome organization.

The Chronic Bee Paralysis virus (CBPV) was a single, unassigned virus with a unique gene, Orf1, of unknown function and a bipartite genome with polymerase and capsid orthologous to nodaviruses, though with sufficient divergence not to be included in the family[22]. We previously described a new clade of honeybee-infecting viruses closely related to CBPV, the Lake Sinai Viruses (LSV1, 2 and 3) that shared the Orf1 gene and a similar polymerase, however employed a monopartite genome organization with a capsid gene related to those of tetraviruses, not CBPV or nodaviruses[5]. We now observe highly conserved LSV2 sequences (>95% nucleotide homology to South Dakota isolates) in *Apis ceranae* samples from Thailand, expanding the geographic and host range of this virus. We propose to name this clade Paranodavirus and now describe a new member virus observed in the invasive Argentine ant, *Linepithema humile* (Figure 1). This new virus, the Argentine ant paranodavirus (AAPV), exhibits similar genome organization to CBPV and amino acid similarity of 33, 42 and 39% in the Orf1, polymerase and capsid genes, respectively (Figure 1).

We also describe a highly divergent nodavirus observed in *Vespula vulgaris*, the Vespula vulgaris nodavirus 1 (VVNV1). This virus is positioned at the base of the nodaviridae and tombusviridae families by phylogenetic comparison of the RdRP gene. The capsid is similarly highly divergent, with low level blastp homology (e-12 to e-14) to the capsids of nematode, mosquito and fish nodaviruses. Two large orfs preceding the polymerase have no alignment to any known protein by blastp or HHpred. Notably, this virus breaks with the nodavirus convention of a bipartite genome organization and arranges the polymerase and capsid on the same strand with a slight overlap. This and other recent discoveries indicate the genome organization of this super-family is highly elastic: phylogenies based on amino acid similarity of the polymerase gene indicate at least 4-6 switches between mono- and bipartite genome organization, with the range due to non-robust support for the exact placement of the tetnoviruses in the nodavirales tree. Whether the ancestral virus is mono- or bipartite cannot be inferred with the phylogenetic tree as currently known. Notably, all observed incidences of monopartite organization involve the capsid gene positioned 3’ of the polymerase gene, often with either a short (<20 nt) intervening space or a short or medium sized (<200 nt) overlap and frame-shift with the polymerase gene.

The Noda-like superfamily is not only prone to genome reorganization, but also to recombination between members and with the family Tetraviridae. This recombination is displayed as a two-part map based on amino acid similarity as determined by an all-versus-all BLASTp of viral capsid or polymerase genes (Figure 2). The polymerase gene is used to position each virus (the nodes) on a two-dimensional space using an edge-weighted force-directed layout, with the degree of similarity influencing the spacing between viruses. A second layer of information in the form of interconnecting lines (edges) between virus nodes is displayed, with the color and presence of each line representing similarity of the capsid gene. This capsid overlay does not influence the position of each node. For example, the SmVA virus[23] (lower right), infecting fungi, possesses a polymerase gene similar to Nodaviridae but a capsid similar to Tombusviridae, thus the virus node is positioned near Nodaviridae but the capsid-similiarity representing lines tie it to the Tombusviridae. Similarly, Providence Virus[24] (center), classified as a tetravirus, possesses a tetra-like capsid gene but the polymerase has no similarity to other tetravirus polymerases and instead is highly related to Tombusvirus polymerases. The Lake Sinai Viruses, as mentioned, also have polymerase and capsid genes of discordant origins (left). In addition to the apparently common combinations of a Noda-like polymerase and Tetra-like capsid, we also describe a potential recombinant with a reverse arrangement, the Braconidae Tetra-Noda-like virus (BTNV). The different clades of origin for the various genes suggest at least four independent inter-family recombination events, three of which involve the Tetraviridae.

In addition to these unusual recombinant noda-like viruses, we also identified four apparently canonical nodaviruses with 30-55% amino acid identity to previous described alphanodaviruses.

Picorna-like viruses

We recovered a picorna-like virus with an unusual genome organization and divergent predicted proteins that place this virus outside of previously described picorna-like virus families. Isolated from the ant *Aphaenogaster occidentalis*, the A. occidentalis Picorna-like virus (AoPLV) consists of a 9,621 nt recovered assembly encoding two large orfs. Unlike the family dicistroviridae, which employs a similar dicistronic organization with replication-related genes preceeding a capsid, AoPLV is arranged inversely with the capsid 5’ of the replication genes. This arrangement is in turn similar to the family Iflaviridae, but that family employs a single cistron. At least two stop codons are present in each frame between the two cistrons, suggesting this is a bona fide genomic reorganization and not a sequencing or assembly artifact. Sanger sequencing of the internal IRES also confirmed the assembled sequence.

Basel picornaviruses with distinct genome organizations have also previously been described from dogs, carp and fire ants. The Solenopsis invicta virus 2[10], originally reported from imported red fire ants, appears to employ a separate frame-shifted orfs for each capsid gene instead of a single orf whose product is subsequently cleaved by a viral protease, as is the case in Picorna-, Dicistro- and Iflaviruses. We describe a similar virus (AAPLV) isolated from the Argentine ant, *Linepithema humile*, with an identical genome organization. Sequence divergence between the two viruses supports them as distinct species by Picorna-like virus standards, with 51% amino acid similarity in the polymerase gene and 25-40% similarity in the four putative capsid genes.

Eight novel dicistroviruses and six novel iflaviruses with canonical genome organizations were also identified, predominantly in the social insect samples. The amino acid similarity boundaries for distinct species and genera are not well specified for either family, fortunately all viruses identified were either very similar to known viruses (>95% amino acid in the polymerase gene) or very different (23-68%). This level of divergence is consistent with at least distinct species, using similarities between ICTV recognized species as a benchmark. Only isolates with at least 3kb of sequence (approximately one third of the expected genome size) and 1kb of sequence in the polymerase gene were included in this tabulation; additional short assemblies of probable picorna-like virus origin were also recovered but without large assemblies it is uncertain if they are truly novel and where they should be placed phylogenetically. Picorna-like virus families are presented in an all-versus-all BLASTp analysis similar to that used for the Noda-like virus superfamily; no clear inter-family recombination events are identified using the same criteria.

Posaviruses are proposed viruses of nematodes distantly related to picorna-like viruses and astroviruses; this family was identified from transcriptome sequencing of the parasitic roundworm *Ascaris suum* and from virome sequencing of pig stool[25], the worm's host. We report a novel Posavirus (AAPoV) observed in the Argentine ant and in a separate Anthrophora bee specimen; the two isolates share >90% nucleotide similarity. This virus exhibits similar genome organization to previously described posaviruses with a large replicase orf followed by a frame-shifted capsid gene, and at 15-25% identity in the polymerase gene is typical in terms of divergence between other members of the posavirus family.

We also recovered an RNA virus from the Diapriidae survey specimen. This virus was similar to the previously described Rosy apply aphid virus and A. pisum virus in both sequence identity (22-28%) and genome organization, and represents a third member of this family.

Lasivirus

The Laem-Singh virus was initially associated with retinopathy in crustaceans and more broadly with the Monodon Slow Growth Syndrome, a multisymptom malady of farmed shrimp[26]. Only 689 nt of the RNA-dependent RNA polymerase (RdRP) gene has been recovered, however RT-PCR and in-situ hybridization has shown this virus to infect at least five species of crustaceans in multiple tissues. The virus initially appeared phylogenetically related to plant viruses of the family Luteoviridae, though with sufficient divergence to potentially warrant placement in its own family. We identify five related viruses in this clade, which we propose to name Lasivirus (LAem-SIngh virus) after the founding species. Contigs of 2,736 to 3,357 nt were assembled from *Vespula vulgaris* (VVLV) and from the Proctotrupidae (PLV), Eurotpmidae (ELV), Torymidae (TLV), and Braconidae (BLV) survey samples (Figure 4). This virus was strongly enriched by the self-priming technique suggesting a hairpin or terminal repeat end and represented 0.6% to 68.4% percent of SP enriched libraries.

Lasiviruses are predicted to contain two large ORFs: a putative serine protease followed by a predicted RdRP separated either by a short intergenic region or overlapping with a frame-shift. The related Sobemoviridae, Luteoviridae and the Mushroom bacilliform viruses share a related polymerase gene and appear roughly equally distant in terms of amino acid identity at 20-25% in this highly conserved gene (Figure 4a). The Lasivirus serine protease is highly divergent and does not align significantly to anything in the NR database by BlastP, however HHpred identified strong serine protease motifs in all five (e-4 to e-33) consistent with the putative serine proteases encoded in the related virus families. Deep sequencing assemblies on all five lasiviruses terminated within 180 nt of the RdRP stop codon, however all three related virus clades possess a capsid located 3’ of the RdRP gene in a monopartite genome arrangement; this could indicate some sequencing- or assembly-resistant sequence unique to lasiviruses or that lasiviruses employ two or more genome segments. Given the low conservation of capsid sequences in these virus families, a novel capsid gene could evade detection. Further, the SP enrichment appears highly variable between genome segments, so high coverage of the polymerase gene does not guarantee equally high coverage of the capsid.

Within lasiviruses, BLV, PLV and VLV form a sub-clade characterized by high amino acid conservation of 50-60% in the RdRP gene and 35-40% in the protease gene, as compared to 25-30% and 10-20% conservation, respectively, between other lasiviruses (Figure 4b and c). Interestingly, the original Laem-Singh virus appears more similar to ELV at 40% amino acid similarity compared to other Laem-Singh viruses. Unlike in dicistroviruses, nodaviruses and baculoviruses, a crustacean virus in this case does not appear to exist on a related but distant clade compared to the insect viruses.

Non-(+)-ssRNA viruses

Baltimore class IV viruses, with positive sense single-strand RNA genomes, make up the majority of known and novel viruses observed in this survey, however we detected several novel negative- and double-strand RNA viruses as well. Two rhabdoviruses were recovered with over 12kb assemblies each, including the complete polymerase gene. The first, identified in the Halictidiae survey sample, appears to include divergent polymerase and nucleocapsid genes most similar to the genera Cytorhabdovirus and Nucleorhabdovirus, both genera infecting plants, however at sufficient amino acid divergence (24% and 9%, respectively) to potentially suggest a novel clade. The second Rhabovirus was isolated from the "Pavement ant" Tetramorium caespitum. This virus shows little similarity to the Halictidiae Rhabdovirus (15% amino acid similarity in the L gene) and is most similar to a proposed new clade of mosquito-borne rhabdoviruses including the Nyamanini virus, albeit also at low similarity (28% aa). Interestingly, the strongest alignment against the non-redundant (NR) database is to ancient, integrated rhabdoviral sequences in the *Aedes aegypti* genome[27], with approximately one-third of the polymerase gene aligning at 41% identity. Several lines of evidence support TcRV as a viral sequence as opposed to a similar integration: TcRV is enriched in micrococcal nuclease treated and filtered samples, suggesting encapsidation of the viral RNA; TcRV encodes a canonical and full-length L gene product including an N-terminal polymerase domain and a C-terminal RNA-capping domain, unlike truncated *A. aegypti* integrations which span smaller regions; and TcRV was only detected in one of two T. caespitum samples analyzed.

Several segmented RNA viruses were identified, however due to the high sequence divergence observed not all expected genome segments were recovered. The polymerase segment of a Bunyavirus (AABV) was recovered from the Argentine ant, *Linepthema humile*, and is most similar to a proposed new genus currently consisting of mosquito-borne bunyaviruses, including Gouleako virus. We also recovered portions of three segments of an orthomyxovirus from the Proctotrupidae survey specimen; the PA and PB2 polymerase and Hemagglutinin genes all support a relationship to Quaranfil virus, a recently described virus of mosquitoes, at a low similarity of 19-31%. Finally, we detected the polymerase segment of a reovirus present in the Braconidae survey specimen; this virus had low (34%) identity to the Colarado tick fever virus, a Coltivirus utilizing human and tick hosts.

DNA viruses have also been described in hymenopterans, though these have primarily been from dsDNA genome families. We describe two densoviruses from the Argentine Ant (AADeV) and *Vespula* sample (VDeV). These possessed relatively high similarity to the previously described *P. citri* and *B. germanicus* densoviruses and similar genome organization.

Braconidae

Of the non-social species, the highest number of viruses detected in a single sample was from the Braconidae. Braconids and some other parasitic wasps inject viruses or virus-like particles into hosts alongside their larvae; this is thought to act as a decoy for the host immune system[28]. We identify six new viruses in the Braconid sample, including the previously mentioned lasivirus (BLV), reovirus (BRV) and proposed tetra-nodavirus recombinant (BTNV), as well as three tetraviruses (BTV1-3). The tetraviruses displayed canonical genome organization and high similarity in the polymerase gene to previously described tetraviruses of Drosophila species (~50% amino acid identity).

Discussion

The viral flora of both social and non-social hymenopterans was examined in an unbiased and several different biased methods. We analyzed more social insects than non-social in terms of total number of samples, total number of individuals and total mass of sample because social insects are easy to acquire and increased sampling and sampling quantity facilitates viral discovery. The absolute number of viruses described can thus not be used to argue whether one group or another is more prone to supporting a virus infection; the relative distribution of virus types is amenable to comparison, however, as the samples were enriched and sequenced by similar methods. We find that social insects are more likely to be associated with viruses of the order Picornavirales (Fisher exact p=0.009, comparing virus species observed in this study, both novel and previously described). This suggests that the order hymenoptera is not itself intrinsically prone to frequent or diverse picorna-like virus infections, but rather the social species of hymenoptera are and a bias in previous investigations towards bees and ants skewed viral discoveries towards that order. These observations cannot distinguish between a difference in the diversity of viruses between social and non-social species or a difference in the frequency of infection and viral load, which would influence viral observations, however the viral survey methodology described here would be applicable to a study controlling for sample number, mass and geographic distribution that would be able to distinguish those two models.

While positive single-stranded RNA viruses from several families appear to be common in hymenopterans, we observed a diverse set of negative ssRNA, dsRNA and ssDNA viruses. Of particular note, we observed six different viruses in a single sample of braconid wasps. Interestingly, braconids are known to have integrated a large DNA virus in the ancient past (>300 million years ago)[28], the capsid of which is expressed and injected into its caterpillar hosts along with the wasp's larvae to act as an immune decoy. Other wasps have been described with non-integrated DNA viruses infecting the reproductive organs with similar effect, and a Dicistrovirus was recently described infecting wasp ovary cells[29], however it is not yet known if this RNA virus serves a similar function. Whether these six RNA viruses also serve as decoys and what their distribution is among wasps is unknown, however further studies in different species of braconid and other parasitic wasps could reveal new viruses that are potentially harnessed by the host. Three of the viruses are of the tetravirus family, previously known to infect butterflies and moths during the larval stage, as well as flies; this supports their potential role as caterpillar pathogens used as immune decoys and suggests a possible route for the wasp to have acquired the virus in the first place.

While 72% of the novel virus species we have described fall within known families based on sequence-based phylogenetics and genome organization, 17% are of sufficient divergence for phylogenetic placement outside currently defined families and 11% are only the second or third member of their clade to be described. New viruses fell into four different novel family-level clades by amino acid divergence and genome organization, suggesting that the total number of virus families is not yet known. Further, this demonstrates the ability of deep sequencing to identify new family-level groups. If all the members of a superfamily (ie Noda-like or Picorna-like) are removed from the sequence database used for comparison, all four novel clades are still identified suggesting that the technique remains robust at the superfamily level, however as no new superfamily-level viruses were identified this is only a *post hoc* argument.

Interfamily recombination among RNA viruses is rare. Capsid and polymerase phylogenies often display different phylogenetic origins at a strain or species level, but at the family level they are almost always conserved. Despite describing 16 new picorna-like virus species and making multiple detections of four known picorna-like viruses, we did not observe any unequivocal inter-family recombination events. In contrast, we observed three inter-family recombination events between the family Tetraviridae and the Noda-like superfamily in addition to two previously described events. Why inter-family and inter-superfamily recombination events are so frequent in these clades but not others is, for now, a matter of speculation.

All of the recombination-prone families are relatively simple in genome organization with single capsid genes and few accessory factors, potentially simplifying the process of meshing two different protein sets. The frequent genome reorganization between mono- and bipartite organization could suggest an easy route to joining two unrelated viral genome segments. Tetraviruses and some monopartite plant noda-like viruses employ a subgenomic RNA encoding the capsid gene which is expressed at a high copy at some stages of the infectious cycle and such subgenomic RNAs have been strongly associated with recombination in other clades, most notably in Caliciviridae. A high copy number tetravirus capsid subgenomic RNA being joined to a noda-like virus polymerase would be a reasonable and likely scenario for the initial generation of the recombinant, followed by the robust noda-like polymerase ensuring viability. The location of genome packaging signals and the function of some accessory factors in still unknown and could also play roles in inter-family recombination. The discovery of more recombinant species will shed light on the viral characteristics that lead to inter-family recombination.

Methods

Sampling

Pilot samples were acquired from a variety of geographic locations and stored by several means to compare nucleic acid quality after storage and amenability to viral enrichment techniques. Storage platforms included Whatman FTA cards (General Electric), dry ice, no preservation (natural desiccation) and live acquisition. Specimens were homogenized by mortar and pestle or by TissueLyzer II (Roche) with a sterile 5 mm zinc-plated steel ball-bearing for 2 min at 30 hz in ice cold PBS. Portions of select samples were passed through 0.45 and 0.22 um filters and digested with 1000 gel units/mL micrococcal nuclease (NEB) at 37C for 30 min prior to the addition of 100 uL 0.5M EDTA and nucleic acid extraction. Hymenopteran samples used in the survey portion of the study were collected live, killed on dry ice and immediately homogenized in ice cold PBS by TissueLyzer. One-tenth of each sample was pooled for micrococcal nuclease treatment as above and the remainder was extracted immediately.

Molecular Biology

Nucleic acid extraction and library preparation for deep sequencing were performed as previously described[5]. Briefly, RNA and DNA were isolated by Trizol(Life Technologies)/choloroform extraction and isopropanol precipitation by centrifugation and cleanup by Zymo RNA Clean and Concentrator (Zymo Research). Samples were split and one portion digested with Terminator

Exonuclease (Epicentre Biotechnologies) to eliminate host and microbial rRNA. Micrococcal nuclease-treated and Terminator-treated were reverse transcribed with SuperScript III (Invitrogen) at 42C for 1 hour with an adapter-linked random nonamer (3Sol13_N, GCTCTTCCGATCTNNNNNNNNN, 40 pmol) and denaturation for 4 minutes at 94C, while undigested nucleic acid samples were also reverse transcribed without any primer to enrich for RNAs with hairpin tail structures or other self-priming motifs. Second-strand synthesis was performed by the addition of Sequenase and 1x sequenase buffer (USB) followed by incubation at 37C for 8 minutes and denaturation at 94C for 4 minutes. Poly-adenylated RNA was enriched by performing a similar reverse transcription reaction with oligo-dU-dT (5'-TTTTUTTTTUTTTTUTTTTU-3', IDT) followed by digestion of the oligo with UDG/APE1 (NEB) at 37C for 30 minutes to eliminate random priming on the homopolymeric oligo, observed to otherwise compromise ~5% of final library amplicons. Oligo 3Sol13_N was added and second strand synthesis was performed twice; all libraries now consisted of cDNA flanked by the shared 13 nt 3' end of the Illumina TruSeq adapters and were now treated identically from this stage onwards.

Five microliters of adapter-linked cDNA was amplified by one-primer PCR with 100 pmol oligo 3Sol13 (GCTCTTCCGATCT) with KlenTaq (Sigma) as per manufacturer's instructions for 15 cycles with an annealing temperature of 37C. Reactions were cleaned by Zymo DNA Clean and Concentrator (Zymo) and resuspended in water prior to quantification of product yield by Nanodrop (Thermo). Products that did not produce at least 5 ng/uL in 20 uL were subjected to further amplification as required. Ten nanograms of input material was used in a palindrome suppression PCR with Klentaq as above. This PCR adds the full-length deep sequencing adapters and selects for amplicons with different adapters by optimizing primer concentrations to suppress amplicons with same adapter via palindrome suppression. Reactions were then pooled and size-selected on a LabChip XT (Caliper) with the DNA500 and 750 chips, and amplified for a further 5 cycles with primers 5Sol1_20 (AATGATACGGCGACCACCGA) and 5Sol2_21 (CAAGCAGAAGACGGCATACGA). Libraries were sequenced on a HiSeq 2000 with v1.5 paired-end cluster generation kits and a 100-nt read length.

Data processing and analysis

Reads with more than 5 ambigious bases (Ns) were removed from the dataset and the remaining reads subjected to two parallel analyses. In the first, reads were assembled with the ABySS[31] and Trinity[15] assemblers and open-reading frames (ORFs) larger than 300 nt were translated and queried against Genbank's non-redundant database (NR) by blastp. Contigs matching annotated viruses were extended by the PRICE assembler for 30 cycles and queried again against NR. In the second pipeline, raw reads are queried against a hand-curated database of insect viruses by Blastx[16] or by HMMER[17]. Low significance hits are extended by PRICE and the resulting large contigs again queried by Blastx or HMMER. In both cases, contigs were deemed viral in origin if they contained at least one gene that aligned more similarly to known viral genes than to any other genes in the database and if the Blast or HMMER expectation value (e-value) was 10^{-10} or lower. Known retrotransposons, endogenous retroviruses and misannotated viral sequences were manually removed.

Phylogenies and Clusters

Phylogenies were generated by protein alignment by ClustalW[32] and dendrograms were generated by the Neighbor-Joining method using the Jukes-Cantor model for genetic distance using the Geneious sequence package[33]. Cluster figures were generated by all-against-all Blastp followed by layout in Cytoscape[34], using the Edge-weighted Force-directed layout option and using the Blastp bit-score as the edge-weight value. Layouts were performed in two steps: an initial step using the entire dataset of several virus families followed by subsequent remapping of family-level groupings within each dataset.

Figure Legends

Figure 1 RdRP amino acid Phylogeny of the Nodavirus-like Superfamily. Amino acid sequences of the RdRP gene for members of the Nodavirus-like superfamily, including the families Nodaviridae and Tombusviridae and the proposed family Paranodaviridae, were aligned by ClustalW and presented in a Neighbor-Joining phylogeny. Families are demarcated by dashed lines and novel species are designated with circles. Genome organization of selected species are overlaid on the phylogeny.

Figure 2 Blast-Cluster Analysis of the Nodavirus-like Superfamily and the Tetraviridae. All polymerase genes in the selected families were aligned against each other by Blastp and arranged in two dimensions using an Edge-weighted layout on a logarithmic scale by Blastp bitscore. Polymerase edges were then masked and capsid alignments overlaid with similarity displayed in red or blue. Novel viruses described in this study are displayed as red nodes. Recombination events are displayed as nodes grouping to one area by polymerase similarity but with connections to another area by capsid similarity.

Figure 3 Blast-Cluster Analysis of the Picornavirales. As figure 2, for families of the order Picornavirales.

Figure 4 RdRP amino acid Phylogeny of the proposed Lasiviridae. As figure 1, for the families Lasiviridae, Sobemoviridae and Luteoviridae.

Table 1 Viruses observed in Social Insects

Table 2 Viruses observed in Non-social Insects

Table 1 Viruses observed in Social Insects

Sample	Virus Designation	Assembly	Similarity (% protein identity)	Clade
Formicidae (Linepithema humile)	Argentine ant Paranodavirus (AAPNV) (Accession here)	1,964 3,633	Capsid: 39% to CBPV Polymerase: 42% to CBPV	Paranodavirus (proposed)
	Argentine ant Dicistrovirus 1 (AADV1)			Dicistroviridae (Picornavirales)
	Argentine ant Dicistrovirus 2 (AADV2)			Dicistroviridae (Picornavirales)
	Argentine ant Iflavirus (AAIV)			Iflaviridae (Picornavirales)
	Argentine ant Picorna-like Virus (AAPLV)	11,347	Capsid: 25-40% (4 ORFs) to SINV2 Polymerase: 51% to SINV2	Picornavirales
	Argentine ant Posavirus (AAPoV) (Posavirus)	8,487	Capsid: 35% to A. suum posavirus Polymerase: 32% to A. suum posavirus	Posavirus
	Argentine ant Bunyavirus (AABV)	5,303	Polymerase: 28% to Gouleako virus	Bunyaviridae
	Argentine ant Densovirus (AADeV)	5,144	Capsid: 48% to P. citri densovirus Polymerase: 40% to P. citri densovirus	
Formicidae (Tetramorium caespitum)	T. caespitum Dicistrovirus (TcDV)			Dicistroviridae (Picornavirales)
	T. caespitum Rhabdovirus (TcRV)	12,095	Nucleocapsid (N): No match Polymerase (L): 28% to Nyamanini Virus	Rhaboviridae (Mononegavirales)
Formicidae (Monomorium ergatogyna)	n/a			
Formicidae (Componotus vicinus)	n/a			
Formicidae (Aphaenogaster	A. occidentalis Picorna-like virus (AoPV)	9,621	Capsid: 20-30% to the Iflaviridae, Dicistroviridae and Picornaviridae	Picornavirales

occidentalis)			Polymerase: 28% to the Iflaviridae, Dicistroviridae and Sequiviridae	
	A. occidentalis Iflavirus (AoIV)			Iflaviridae (Picornavirales)
Formicidae (Formica ca01)	n/a			
Apidae (Apis ceranae)	Deformed Wing Virus (DWV)*	n/a	Capsid: DWV >95% Polymerase: DWV >95%	Iflaviridae (Picornavirales)
	Lake Sinai Virus 2 (LSV2)*	n/a	Capsid: LSV2 >95% Polymerase: LSV2 >95%	Paranodavirus (proposed)
Apidae (Apis florea)	Deformed Wing Virus (DWV)*	n/a	Capsid: DWV >95% Polymerase: DWV >95%	Iflaviridae (Picornavirales)
Apidae (Bombus sp)	Bombus Iflavirus 1 (BIV1)			Iflaviridae (Picornavirales)
	Bombus Iflavirus 2 (BIV2)			Iflaviridae (Picornavirales)
	Acute Bee Paralysis Virus (ABPV)*	n/a	Capsid: ABPV >95% Polymerase: ABPV >95%	Dicistroviridae (Picornavirales)
Vespidae (Vespula sp)	Vespula Nodavirus 1 (VNV1)	3134	Polymerase: 58% to Nodamura virus	Nodaviridae
	Vespula Nodavirus 2 (VNV2)	1285	Polymerase: 55% to Pieris rapae virus	Nodaviridae
	Vespula Noda-like Virus (VNLV)	3168	Capsid: 24% to Tombusviridae and Lake Sinai Viruses Polymerase: 30% to Wuhan, Beta- and Nematode nodaviruses	Nodaviridae-like
	Vespula Densovirus (VDeV)	1397 1126	Capsid: 58% to B. germanica densovirus Polymerase: 65% to B. germanica densovirus	Parvoviridae
	Vespula Lasivirus (VLV)	2971	See Fig 4	Lasivirus (proposed)
	Vespula Iflavirus (VIV)			Iflaviridae (Picornavirales)
	Vespula Dicistrovirus 1 (VDV1)			Dicistroviridae (Picornavirales)
	Vespula Dicistrovirus 2 (VDV1)			Dicistroviridae

				(Picornavirales)
	Vespula Dicistrovirus 3 (VDV1)			Dicistroviridae (Picornavirales)
	Vespula Dicistrovirus 4 (VDV1)			Dicistroviridae (Picornavirales)
	Vespula Dicistrovirus 5 (VDV1)			Dicistroviridae (Picornavirales)
	Acute Bee Paralysis Virus (ABPV)*	n/a	Capsid: ABPV >95% Polymerase: ABPV >95%	Dicistroviridae (Picornavirales)
	Black Queen Cell Virus (BQCV)*	n/a	Capsid: BQCV >95% Polymerase: BQCV >95%	Dicistroviridae (Picornavirales)
	Deformed Wing Virus (DWV)*	n/a	Capsid: DWV >95% Polymerase: DWV >95%	Iflaviridae (Picornavirales)

Table 2 Viruses observed in Non-social Insects

Sample	Virus Designation	Assembly	Similarity (% protein identity)	Clade
Apidae *(Xylocopa tabaniformis orpifex)*	Argentine ant Posavirus (AAPoV2) (Posavirus)		See Argentine ant entry	Posavirus
Braconidae	Braconidae Lasivirus (BLV)	2937	See Fig 4	Lasivirus (proposed)
	Braconidae Tetravirus virus 1 (BTV1)		Polymerase: 51% to D. melanogaster tetravirus SW2	Tetraviridae
	Braconidae Tetravirus virus 2 (BTV2)		Polymerase: 47% to D. melanogaster tetravirus SW2	Tetraviridae
	Braconidae Tetravirus virus 3 (BTV3)			Tetraviridae
	Braconidae Tetra- Noda-like virus (BTNV)		Capsid: 24% to Wuhan and Betanodaviruses Polymerase: 32% to D. melanogaster tetravirus SW2	Nodaviridae-like Tetraviridae-like (putative recombinant)
	Braconidae Reovirus (BRV)	3284 1236	Polymerase: 34% to Colorado tick fever virus	Reoviridae
Chalcididae	n/a			
Cynipidae	n/a			
Diapriidae	Diapriidae RNA Virus (DRV)	10,047	Capsid: 22% to Rosy Apple Aphid Virus Polymerase: 28% to RAAV	Rosy Apple Aphid virus-like
Eupelimidae	n/a			
Eurytomidae	Eurytomidae Lasivirus (ELV)	2736	See Fig 4	Lasivirus (proposed)
Figitidae	n/a			
Ichneumonidae	n/a			
Megaspilidae	n/a			
Pompilidae	n/a			
Proctotrupidae	Proctroptrupidae Lasivirus (PLV)	2758	See Fig 4	Lasivirus (proposed)
	Proctroptrupidae Iflavirus (PLV)			Iflaviridae

				(Picornavirales)
	Proctroptrupidae Orthomyxovirus (POV)	1432 2549 660	Polymerase PA: 25% to Quaranfil virus Polymerase PB2: 19% to Quaranfil Hemagglutinin: 31% to Quaranfil	Orthomyxoviridae
Pteromalidae	Bombus Iflavirus 2 (BIV2)		See Bombus entry	Iflaviridae (Picornavirales)
Scelionidae	Seclionidae Nodavirus (SNV)	2831	Polymerase: 39% to Bat Guano associated nodavirus	Nodaviridae
Sphecidae	n/a			
Tenthredinidae	n/a			
Torymidae	Torymidae Lasivirus (TLV)	2257	See Fig 4	Lasivirus (proposed)
	Torymidae Tetravirus virus (TTV)			Tetraviridae
Halictidiae *(Lasiogloccum tegulariforme)*	Halictidiae Rhabdovirus A	12,769	Nucleocapsid (N): 9% to Nucleorhabdovirus Polymerase (L): 24% to Cytorhabdovirus	Rhabdoviridae

Figure 1

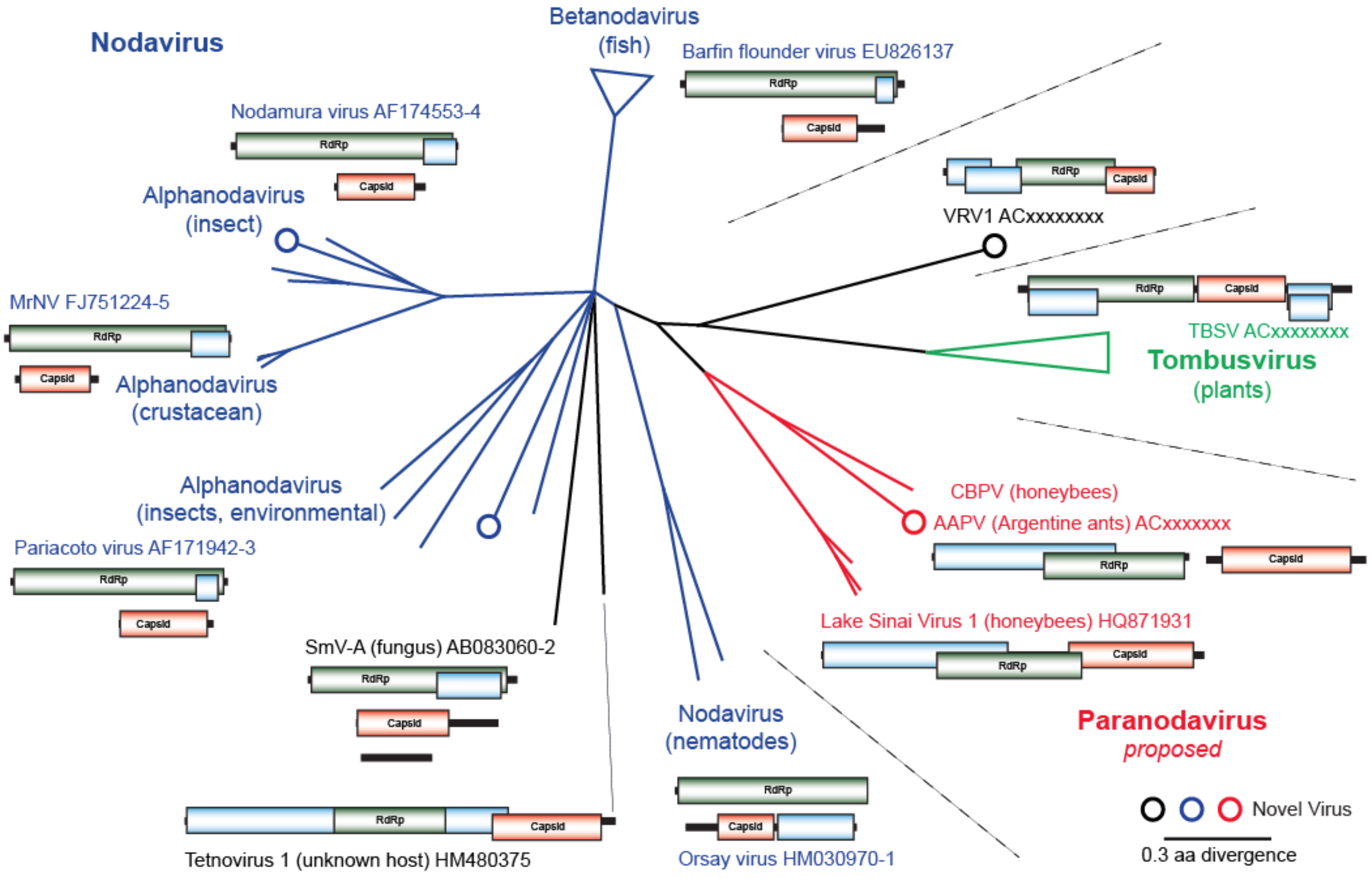

Nodavirus
Betanodavirus
(fish)
Barfin flounder virus EU826137
RdRp
Capsid
Nodamura virus AF174553-4
RdRp
Capsid
Alphanodavirus
(insect)
VRV1 ACxxxxxxxx
RdRp
Capsid
MrNV FJ751224-5
RdRp
Capsid
Alphanodavirus
(crustacean)
TBSV ACxxxxxxxxx
RdRp
Capsid
Tombusvirus
(plants)
Alphanodavirus
(insects, environmental)
Pariacoto virus AF171942-3
RdRp
Capsid
CBPV (honeybees)
AAPV (Argentine ants) ACxxxxxxx
RdRp
Capsid
Lake Sinai Virus 1 (honeybees) HQ871931
RdRp
Capsid
SmV-A (fungus) AB083060-2
RdRp
Capsid
Nodavirus
(nematodes)
Paranodavirus
proposed
RdRp
Capsid
Orsay virus HM030970-1
Tetnovirus 1 (unknown host) HM480375
RdRp
Capsid
Novel Virus
0.3 aa divergence

Figure 2

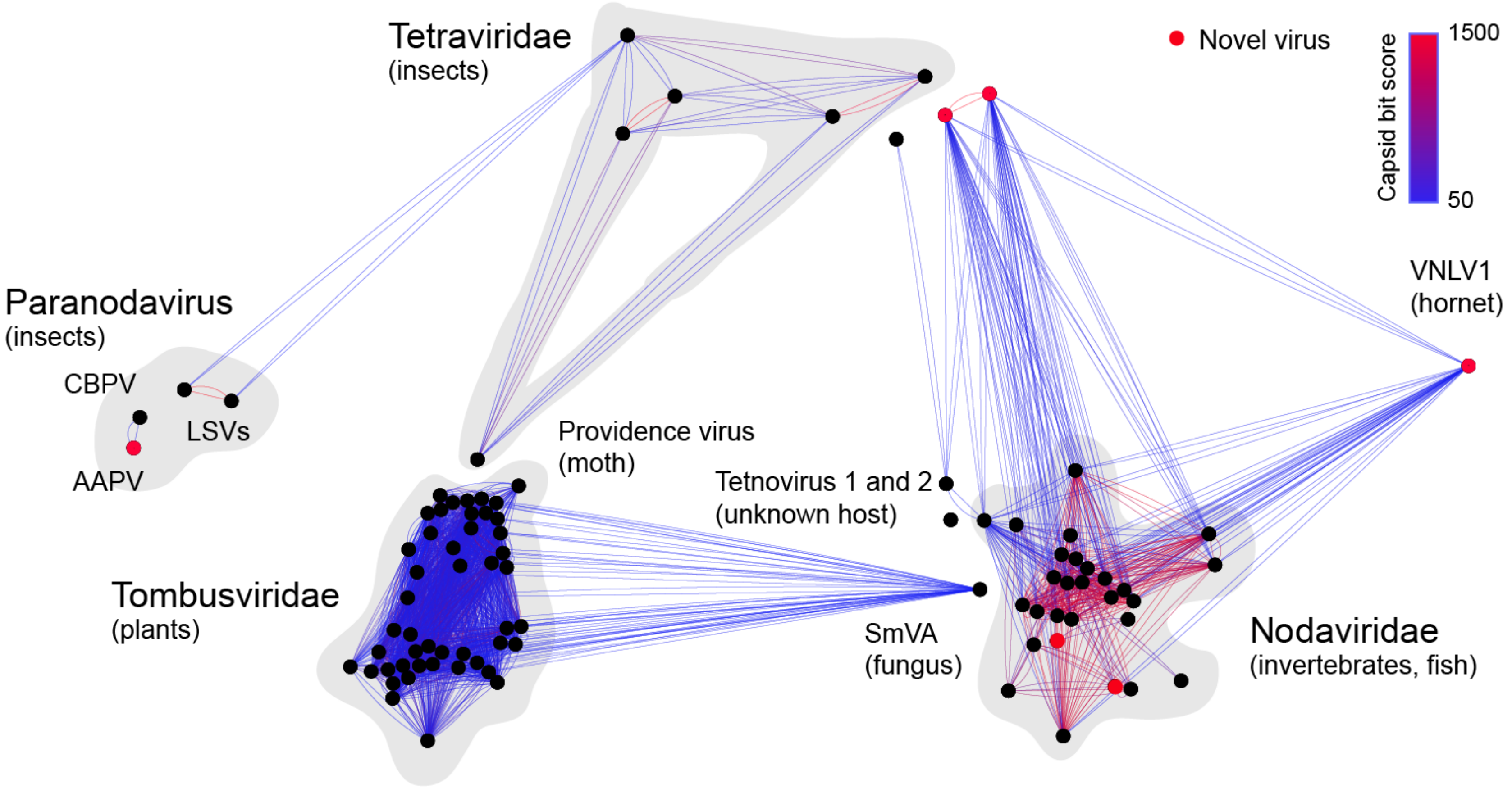

Tetraviridae
(insects)
Novel virus
Capsid bit score
1500
50
VNLV1
(hornet)
Paranodavirus
(insects)
CBPV
LSVs
AAPV
Providence virus
(moth)
Tetnovirus 1 and 2
(unknown host)
Tombusviridae
(plants)
SmVA
(fungus)
Nodaviridae
(invertebrates, fish)

Figure 3

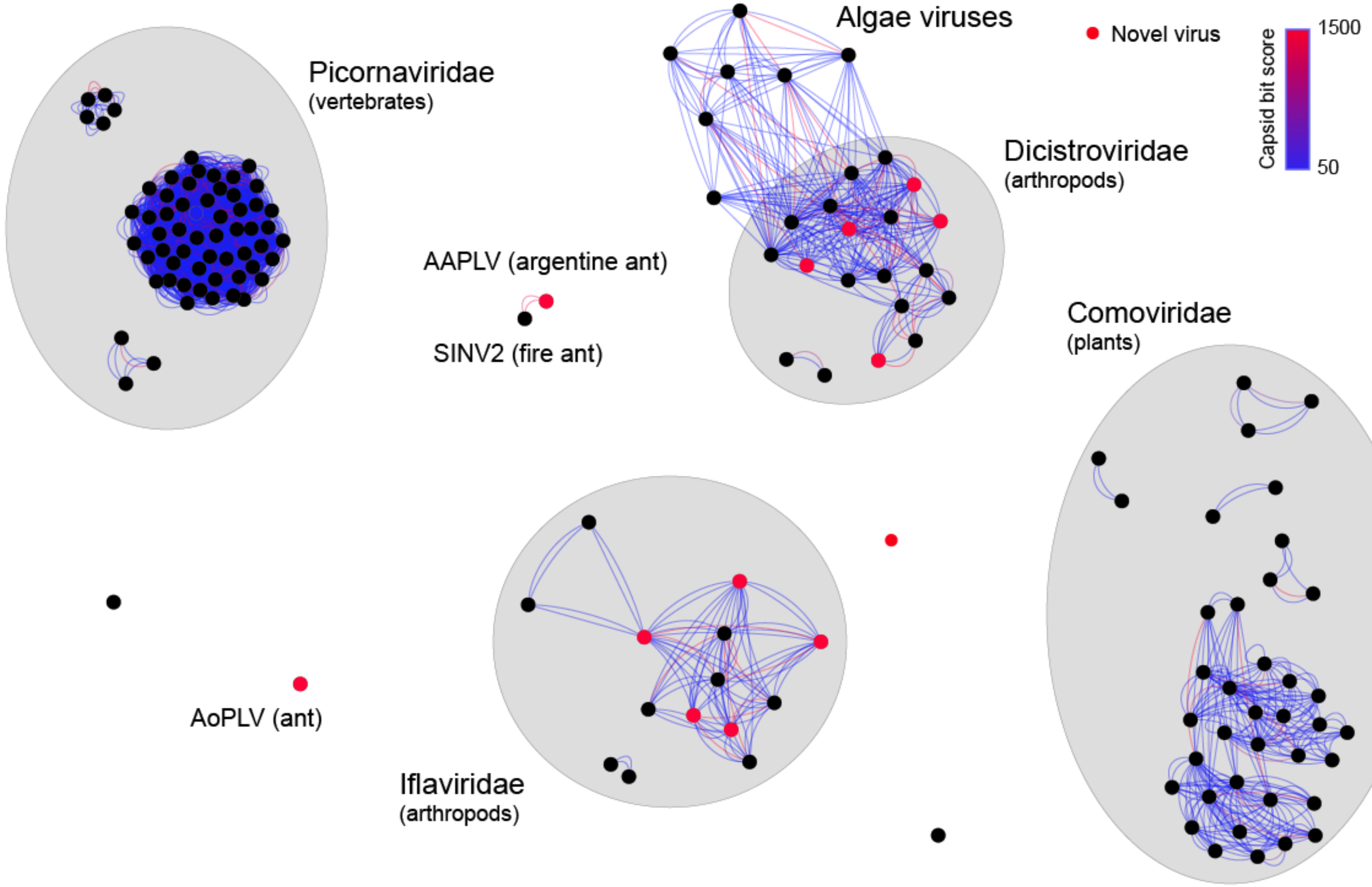
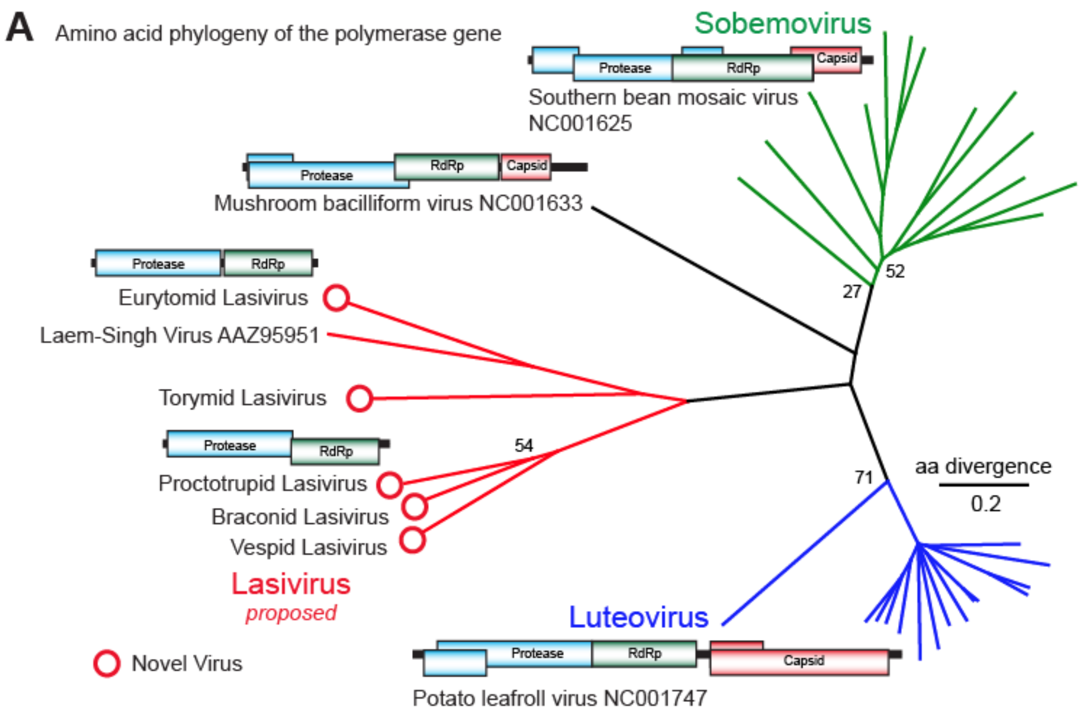

Picornaviridae
(vertebrates)
AAPLV (argentine ant)
SINV2 (fire ant)
Algae viruses
Dicistroviridae
(arthropods)
Novel virus
Capsid bit score
1500
50
Comoviridae
(plants)
AoPLV (ant)
Iflaviridae
(arthropods)

Figure 4

A Amino acid phylogeny of the polymerase gene



B Pairwise amino acid identity (%), Protease gene

	ELV	TLV	BLV	PLV
TLV	14			
BLV	15	12		
PLV	19	12	39	
VLV	18	12	39	38

C Pairwise amino acid identity (%), Polymerase gene

	LS*	ELV	TLV	BLV	PLV
ELV	40				
TLV	29	27			
BLV	25	28	29		
PLV	30	28	29	58	
VLV	27	27	30	50	53

References

1. Hughes, W. O. H., Oldroyd, B. P., Beekman, M. & Ratnieks, F. L. W. Ancestral monogamy shows kin selection is key to the evolution of eusociality. *Science* **320**, 1213–1216 (2008).
2. Giraud, T., Pedersen, J. S. & Keller, L. Evolution of supercolonies: the Argentine ants of southern Europe. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6075–6079 (2002).
3. Chen, Y., Evans, J. D., Smith, I. B. & Pettis, J. S. Nosema ceranae is a long-present and wide-spread microsporidian infection of the European honey bee (Apis mellifera) in the United States. *J. Invertebr. Pathol.* **97**, 186–188 (2008).
4. Vanengelsdorp, D. & Meixner, M. D. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *J. Invertebr. Pathol.* **103 Suppl 1**, S80–95 (2010).
5. Runckel, C. *et al.* Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. *PLoS ONE* **6**, e20656 (2011).
6. Celle, O. *et al.* Detection of Chronic bee paralysis virus (CBPV) genome and its replicative RNA form in various hosts and possible ways of spread. *Virus Res.* **133**, 280–284 (2008).
7. Bailey, L. The multiplication of sacbrood virus in the adult honeybee. *Virology* **36**, 312–313 (1968).

8. Bailey, L. & Woods, R. D. Three previously undescribed viruses from the honey bee. *J. Gen. Virol.* **25**, 175–186 (1974).

9. Valles, S. M., Strong, C. A. & Hashimoto, Y. A new positive-strand RNA virus with unique genome characteristics from the red imported fire ant, Solenopsis invicta. *Virology* **365**, 457–463 (2007).

10. Hashimoto, Y. & Valles, S. M. Infection characteristics of Solenopsis invicta virus 2 in the red imported fire ant, Solenopsis invicta. *J. Invertebr. Pathol.* **99**, 136–140 (2008).

11. Oliveira, D. C. S. G. *et al.* Data mining cDNAs reveals three new single stranded RNA viruses in Nasonia (Hymenoptera: Pteromalidae). *Insect Mol. Biol.* **19 Suppl 1**, 99–107 (2010).

12. Greninger, A. L. *et al.* The complete genome of klassevirus - a novel picornavirus in pediatric stool. *Virol. J.* **6**, 82 (2009).

13. Yozwiak, N. L. *et al.* Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* **6**, e1485 (2012).

14. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).

15. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

17. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).

18. Qi, N. *et al.* RNA binding by a novel helical fold of b2 protein from wuhan nodavirus mediates the suppression of RNA interference and promotes b2 dimerization. *J. Virol.* **85**, 9543–9554 (2011).

19. Wang, S., Mortazavi, L. & White, K. A. Higher-order RNA structural requirements and small-molecule induction of tombusvirus subgenomic mRNA transcription. *J. Virol.* **82**, 3864–3871 (2008).

20. Kapoor, A., Simmonds, P., Lipkin, W. I., Zaidi, S. & Delwart, E. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. *J. Virol.* **84**, 10322–10328 (2010).

21. Félix, M.-A. *et al.* Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses. *PLoS Biol.* **9**, e1000586 (2011).

22. Olivier, V. *et al.* Molecular characterisation and phylogenetic analysis of Chronic bee paralysis virus, a honey bee virus. *Virus Res.* **132**, 59–68 (2008).

23. Yokoi, T., Yamashita, S. & Hibi, T. The nucleotide sequence and genome organization of Sclerophthora macrospora virus A. *Virology* **311**, 394–399 (2003).

24. Walter, C. T. *et al.* Genome organization and translation products of Providence virus: insight into a unique tetravirus. *J. Gen. Virol.* **91**, 2826–2835 (2010).

25. Shan, T. *et al.* The fecal virome of pigs on a high-density farm. *J. Virol.* **85**, 11697–11708 (2011).

26. Sritunyalucksana, K., Apisawetakan, S., Boon-Nat, A., Withyachumnarnkul, B. & Flegel, T. W. A new RNA virus found in black tiger shrimp Penaeus monodon from Thailand. *Virus Res.* **118**, 31–38 (2006).

27. Fort, P. *et al.* Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol. Biol. Evol.* **29**, 381–390 (2012).

28. Thézé, J., Bézier, A., Periquet, G., Drezen, J.-M. & Herniou, E. A. Paleozoic origin of insect large dsDNA viruses. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15931–15935 (2011).

29. Zhu, J.-Y., Ye, G.-Y., Fang, Q., Wu, M.-L. & Hu, C. A pathogenic picorna-like virus from the endoparasitoid wasp, Pteromalus puparum: initial discovery and partial genomic characterization. *Virus Res.* **138**, 144–149 (2008).

30. Hahn, C. S., Lustig, S., Strauss, E. G. & Strauss, J. H. Western equine encephalitis virus is a recombinant virus. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5997–6001 (1988).

31. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).

32. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

33. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

34. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Aug. 14, 2012

Author Signature

Date