Virus Genes 30:2, 151–156, 2005

Analysis of the Virus Population Present in Equine Faeces Indicates the Presence of Hundreds of Uncharacterized Virus Genomes★

ALAN JAMES CANN, SARAH ELIZABETH FANDRICH & SHAUN HEAPHY*
Department of Infection Immunity and Inflammation, University of Leicester, Leicester LE1 9HN, UK

Received July 20, 2004; Accepted August 9, 2004

Abstract. Virus DNA was isolated from horse faeces and cloned in a sequence-independent fashion. 268 clones were sequenced and 178140 nucleotides of sequence obtained. Statistical analysis suggests the library contains 17560 distinct clones derived from up to 233 different virus genomes. TBLASTX analysis showed that 32% of the clones had significant identity to GenBank entries. Of these 63% were viral; 20% bacterial; 7% archaeal; 6% eukarya; and 5% were related to mobile genetic elements. Fifty-two percent of the virus identities were with *Siphoviridae*; 26% unclassified phages; 17% *Myoviridae*; 4% *Podoviridae*; and one clone (2%) was a vertebrate Orthopoxvirus. Genes coding for predicted virus structural proteins, proteases, glycosidases and nucleic acid-binding proteins were common.

Key words: bacteriophage genomes, equine, gene discovery

Introduction

The composition of bacterial populations as judged by 16S ribosomal gene sequencing has been a subject of productive research for more than 20 years [1,2]. Such studies have contributed to the realization that bacterial diversity is enormous and that most bacterial species cannot be readily cultured in the laboratory. More recently the total bacterial genomic DNA from given environments has been isolated and cloned in a process sometimes called 'metagenomics'. Metagenomics allows access to a vast uncharacterized gene pool both for biotechnological purposes and to answer questions about microbial genomics and ecology [3]. At the start of this millennium, similar studies have begun to explore and identify an unexpectedly large diversity in marine eukaryotic 'picoplanktonic' cells smaller than 5 μm in diameter [4,5]. Most recently the virus (bacteriophage and eukaryotic) gene pool has started to be characterized in a sequence-independent fashion [6,7].

Virus particles at about 10^{10} ml^{-1} are 5–25 more numerous than bacterial cells in surface sea-water samples [8] and are thought to outnumber bacterial cells in many environments [9]. Current estimates put the number of viruses present in the biosphere at 10^{31} [9]. Bacteriophages in particular must play an enormous part in global ecosystems by preying on common prokaryotes, modulating population sizes, cycling key nutrients and via their role in genetic exchange and its effect on evolution [9]. They are also likely to be just as important in the normal and diseased physiology of microhabitats such as the human gut.

Virus particles are a source of considerable genomic diversity [10,11], the scope of which has yet to be realized. The diversity of virus genomes has been hidden by the inability to culture most of the host bacteria and sometimes by the lytic nature of

*Author for all correspondence:
E-mail: sh1@le.ac.uk
★The nucleotide sequence data reported in this paper have been submitted to the GenBank Genome Survey Sequences Division and have been assigned accession numbers CL536634 to CL536634 inclusive.

the viruses themselves. Even virus DNA cloning has been made difficult by DNA modifications and host-range restriction. Sequence-independent techniques have now been developed that overcome these problems, at least for viruses with a double-stranded DNA genome. They have been used to study the metagenomic virus gene pool of sea-water [6] and the human gut [7]. The latter environment has been estimated to contain 1200 virus genotypes and in excess of 600 bacterial species.

The equine gut flora is less well characterized. In one significant molecular analysis of 16S phylotypes, the bacterial flora were found to be broadly similar to that of the ruminant and human gut at the phylum level of analysis [12]. However, 89% of the phylotypes had less than 97% identity to known database entries, i.e. probably represent unknown species. Apart from a few pathogenic enteric viruses, almost nothing is known about the virus population and microbial diversity of the equine gut. To establish methodologies for virus metagenomics we chose to analyze virus sequences from the virus-rich environment that horse faeces was likely to represent. Such virus metagenomic libraries can be used to determine virus diversity, virus population dynamics, identify new genes, and even determine genomic sequences of uncultivated viruses. Given that gastro-intestinal disease is the most important cause of mortality in domesticated horses, a thorough understanding of the equine gut flora including viruses is also important for veterinary reasons [12].

Materials and Methods

DNA Isolation

All enzymes were purchased from Promega. DNA was isolated from fresh faecal material which had been kept in a moist atmosphere below 20°C for a maximum of 24 h before processing. Faeces were resuspended in phosphate buffered saline (PBS) at approximately 1 g per 10 ml and subjected to low speed centrifugation and sequential ultrafiltration through 0.45 and 0.2 μm pore size Acrodisc membranes (Gelman). Remaining particulate material in the filtrate was precipitated by addition of 1 M NaCl and 10% w/v PEG 6000 followed by incubation at 0–4°C for at least 1 h and centrifugation at $1400 \times g$ at 4°C for 60 min. Pellets were gently resuspended in 0.5 ml 10 mM Tris pH 7.5, 10 mM $MgCl_2$ and 100 mM NaCl. Free nucleic acids were digested with 10 U DNase I and 10 μg mL^{-1} RNase A for 30 min at 37°C. Remaining protected nucleic acids were extracted with phenol:chloroform and precipitated with ethanol.

Library Construction

DNA was resuspended in TE pH 8.0 containing 10% glycerol and randomly sheared by passage through a nebulizer (Invitrogen). Sheared DNA was end-repaired and phosphorylated with T4 DNA polymerase and T4 polynucleotide kinase respectively. Double-stranded DNA linkers were then ligated to the ends of the strands using T4-DNA ligase; Not1 CTCTTGCGGCCGCT TCTC, Not2 GAGAAGCGGCCGCAAGAG. Approximately 100 ng of this DNA was subjected to 30 cycles of amplification with pfu polymerase using the Not1 oligonucleotide as a primer. For T–A cloning, adenosine residues were added to the amplified DNA with *Taq* polymerase by incubating with 0.5 mM dATP for 30 min at 72°C. Amplified DNA was size fractionated on agarose gels and the fraction corresponding to 500–2000 bp was ligated into pGEM-T Easy (Promega). DNA was introduced into *E. coli* strains DH5α or XL-1 Blue (Stratagene) by electrotransformation. Recombinant clones were identified by blue/white screening under standard conditions.

Sequence and Bioinformatics Analysis

Plasmid miniprep DNA was prepared and the presence of cloned inserts confirmed by digestion with *Eco*RI and agarose gel electrophoresis. Nucleotide sequencing of plasmid inserts was performed by a commercial company, AGOWA, using T3 and T7 primers. All sequences were trimmed of vector sequences. To identify contigs and overlaps, all the sequences in the library were compared to each other using BLASTN. Sequences were then compared to the GenBank database in January 2004 using BLASTN and TBLASTX from NCBI. Clones with identity to virus and bacterial sequences were further analyzed using BLASTP and ORF finder, also available at NCBI. Sequences have been deposited at the GenBank Genome

Survey Sequences Division under accession numbers CL536634 to CL536634 inclusive.

Efficiencies of Phage λ and λ Genomic DNA Recovery from Faeces

λZap-Express (Stratagene) phage particles were grown and titred following the manufacturer's instructions. DNA was purified and quantified by UV spectrophotometry. In order to assess potential contamination of the clone library with free DNA, i.e. non-virus contained sequences, the following samples were analyzed: 5 g of fresh horse faeces with no additions (negative control); mixed with 2.5×10^6, 2.5×10^5 and 2.5×10^4 pfu λZap phage particles, or mixed with 2.5×10^6, 2.5×10^5 and 2.5×10^4 copies of purified λZap DNA. After extraction following the same method as for the construction of the EQ1 LASL library (above), phage DNA recovery was determined by PCR amplification of the λZap-Express genome using a primer pair based on the published vector sequence which amplifies a 522 bp fragment adjacent to the multiple cloning site (pBK-CMV residues 485–505 and 987–1006). These amplifications were performed using *Taq* polymerase for 30 cycles and made semi-quantitative by analysis of a 10-fold serial dilution of vector DNA in each experiment.

Results and Discussion

Efficiency of Virus Recovery

Three control experiments with bacteriophage λ added to fresh faecal material show that virus recovery from nuclease-treated extracts prior to phenol extraction is in the range of 10–20% (data not shown). We assume that the efficiency of bacteriophage λ recovery is typical of other viruses. The large loss is presumably due to adsorption to solid matters in the sample and retention on the filtration membranes. Electron microscopy of filtered extracts (data not shown) revealed an array of virus-like particles, but this was not further investigated.

Virus DNA Recovery

This procedure was designed to isolate double stranded DNA virus genomes. Single-stranded DNA virus genomes are unlikely to be recovered and RNA virus genomes will not be recovered. Using the protocol described we were routinely able to isolate approximately 100 ng of DNase-protected DNA per gram of fresh horse faecal material, used to make library Eq1. Similar amounts of DNA have been recovered from chicken faeces and a lake sediment sample, which have also been used to make libraries, the procedure therefore seems robust. Two experimental observations confirm that the isolated DNA is derived from virus particles, substantially free of DNA derived from the host or associated bacteria. First, filtered faecal extracts contained high levels of a temperature-sensitive (75°C for 10 min) endogenous nuclease activity sufficient to degrade 20 μg ml^{-1} of λ DNA within 30 min at 37°C (data not shown). The combination of endogenous nuclease and added DNase I makes it likely that only virus particle-protected DNA, or DNA associated with other colloidal material is isolated by our procedure. The endogenous nuclease activity alone can degrade at least 20-fold more DNA than we recover. Second, we used PCR analysis to assess potential contamination of virus encapsidated DNA with free DNA using our DNA extraction procedure on 5 g horse faeces spiked with either λZap-Express phage particles or λZap-Express DNA. Primers amplifying a 522 bp fragment of adjacent to the multiple cloning site of λZap-Express were used, see materials and methods for details. After extraction, no amplicon was detected from faeces alone. We could detect the 522 bp amplicon when 2.5×10^4 pfu of λZap-Express particles were added to horse faeces. We were unable to detect 5×10^6 genome equivalents of free λZap-Express DNA added to faeces. As our limit of sensitivity in this assay was between 1000 and 10,000 genome equivalents, this suggests that any free DNA is indeed degraded during our isolation procedures. The 200-fold or greater sensitivity in detection of λ encapsidated DNA compared to free phage DNA suggests that less than 0.5% of the clones in the Equation (1) LASL library would arise from contamination of the library with free DNA.

Taken together, this evidence suggests the vast majority of clones in our three libraries should

represent phage genomes, and that bacterial-related sequences in the library (see below) represent transduced bacterial genes.

Population Analysis of Equation (1) Clones

A total of 268 clones were analyzed. The average sequence-read length was 664 nucleotides, ranging from 113 to 826 nucleotides and the total number of nucleotides sequenced was 178140. Of the 268 clones, 263 contained unique sequences, one sequence appeared twice and another sequence occurred three times. This gives a Shannon Index, a measure of species diversity and evenness, of 5.6 [13]. This value is higher than is often reported for microbial communities but lower than that being reported from virus communities, ranging from 6.4 in human faeces to 9 in marine sediments [6,7]. Chao1 is a commonly used non-parametric estimation of the number of classes in a population [14]. It is particularly useful in studies such as this where the data sets are skewed toward the low-abundance classes. In the following equation where S_{Chao1} is the total number of different clones in a population, S_{obs} is the number of observed clones, n_1 is the number of clones observed once, n_2 is the number of clones observed twice.

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{n_1^2}{2n_2} \quad (1)$$

In the Eq1 library $S_{\text{obs}} = 268$. S_{Chao1} therefore $= 268 + (263^2/2 \times 2) = 17560$ i.e. the Eq1 library contains 17560 distinct clones with an average size of 664 bp. Assuming that the average virus genome size is 50 kb, typical of tailed bacteriophages that constitute the majority of virus clones in this library (see below) then (50000/664) or 75.3 clones comprise a complete virus genome and the Equation(1) library contains (17560/75.3) or 233 different virus genomes. The value of S_{obs} is high enough to make this calculation dependent on the assumptions that almost all of the clones are virus in origin and of the average virus genome size. A similar study of the uncultured virus population of human faeces estimated the number of virus species using Chao1 analysis to be broadly similar to this value i.e. approximately 147 [7]. However the same data set was estimated by Monte Carlo analysis to contain up to 1200 species. If we assumed only 20% of the clones are virus in origin, an absolutely minimal estimate based on our sequence analysis described below, then the estimated number of virus genomes in the Equation (1) library would drop to 47.

Equation (1) Clone Sequence Analysis

BLASTN analysis was generally uninformative since most of the nucleotide sequences obtained are highly novel. Four clones (1.5%) had an E value < 0.0001. Two of these (accession numbers CL536728 & CL536820) showed 100% identity over a short length of 27 or 28 nucleotides, to a mouse and a phage sequence respectively. Two showed lower identity to 104 or 131 nucleotides. The first came from a read of 672 nucleotides (accession no CL536711) and showed 86% identity over the 104 nucleotides ($E = 4 \times 10^{-7}$) to a genomic sequence from *Agrobacterium tumefaciens*. The second most significant hit from a short sequence read of 132 nucleotides (accession no CL536891, 96% ID over 131 residues $E = 2 \times 10^{-57}$) was to an 18S rDNA sequence from *Trichoderma viride*. This is a widely distributed soil fungus. Using TBLASTX analysis to classify the sequences, see Fig. 1a, 68% (181 clones) did not find a significant homology in Genbank (i.e. $E > 0.001$). The figure given by Breitbart et al. [7], was 59%, possibly reflecting the fact that human bacteria and viruses are better characterized than equine ones. Of the 87 significant hits, 20% or 55 clones were virus, including one prophage hit assigned with reference to Casjens [15], 6% or 17 clones were bacterial, 2% or 6 clones were archaeal, less than 2% or 5 clones were eukarya, less than 2% or 4 clones, were related to mobile genetic elements and plasmids, see Fig. 1b. This is different to the results presented by Breitbart et al. [7] where "bacterial" hits were twice as common as virus hits and archaeal and mobile elements were also more common. Since we have demonstrated that our library is not contaminated with free DNA sequences, many of the "bacterial" hits are likely to be genuinely phage related, although the database annotations suggested that only two of these hits were phage related. Many of the "bacterial" hits likely represent either transduced bacterial genes or cryptic prophage sequences [11,16].

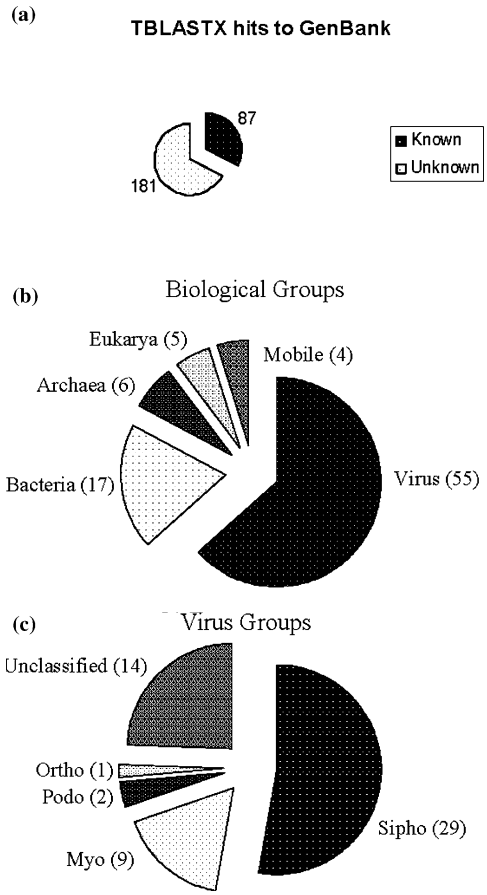

Fig. 1. Overview of Eq1 clone library sequence classifications based on TBLASTX sequence similarities in GenBank. (a) Numbers of sequences with a significant match, i.e. $E < 0.001$. (b) Classification of identifiable hits into biological groups, number of clones is indicated in brackets. (c) Classification of the virus hits into virus groups, *Siphoviridae, Myoviridae, Podoviridae, Orthopoxviridae* and unclassifieds. The number of clones is indicated in brackets.

Of the 55 significant virus homologies, 29/55 (52%) were to *Siphoviridae* (λ-like phages), 14/55 (26%) were to unclassified viruses, 9/55 (17%) were to *Myoviridae* (T4-like phages), 2/55 (4%) were to *Podoviridae* (T7-like phages) and 1/55 (2%) was to Orthopoxvirus (vertebrate poxviruses) see Fig. 1c.

Open Reading Frame Analysis of the Equation (1) Library

Eighty-two clones identified by TBLASTX analysis as having identity to either a bacterial, virus or mobile entry were further examined using ORF finder from NCBI. Read lengths varied from 132 to 845 nucleotides, some clones were read completely. ORF finder identified both complete and incomplete ORFs, some of the latter potentially coding for proteins greater than 263 amino acids long. BLASTP analysis identified proteins likely to be involved in phage DNA replication, morphogenesis and cell lysis as shown in Table 1. ORFs with no predictable function were the most common. Enzymes probably involved in bacterial cell wall hydrolysis were the second most commonly identified class, followed by structural proteins. Many of the proteins were involved in nucleic acid interactions and nucleotide metabolism. Similar proteins were found by Breitbart et al. [7].

This work has given us a glimpse into the virus population complexity of the horse gut. We need to sequence many more clones to establish more contigs which will allow us to build up a database of virus ORFs and even virus genomes. Future work could be extended to include both single-stranded virus DNA genomes and RNA virus genomes. We could also look at the changing population dynamics of both virus and bacteria in the horse gut under varying conditions and attempt to isolate infectious virus corresponding

Table 1. Categories of probable virus ORF function as indicated by BLASTP analysis

Protein classification	Number of matches
Unknowns	27
Glycosidases various	8
Capsid/structural	5
ATPases	5
Proteases	4
Endonucleases	4
Helicases	3
Transposases	3
Terminases	3
Tape measure proteins	3
Portal proteins	
DNA methylases	2
DNA polymerases	2
Oxidoreductases	2
Nucleotide transferases	2
Dehydrogenase	1
Integrase	1
SH3 domain protein	1
dUTPase	1
Transcriptional regulator	1
Chemotaxis factor	1

to some of these genomes. We have also constructed a virus library from chicken faeces and, although our sequence analysis is much less complete, many of the observations made above in terms of virus diversity may be similar. Finally, in this work we have found that the sequence homology of our clones to GenBank entries is very low. We have no idea what approximately 68% of the sequence information codes for. Presumably it does something- is it related to known functions or unknown functions?

Acknowledgements

We would like to thank Raymond and Zamire Dalgleish for assistance with various aspects of the work presented here.

References

1. Stahl D.A., Lane D.J., Olsen G.J., and Pace N.R., Science *224*, 409–411, 1984.
2. DeLong F.E. and Pace N.R., Syst Biol *50*, 470–478, 2001.
3. Rodriguez-Valera F., FEMS Microbiol Lett *231*, 153–158, 2004.
4. Lopez-Garcia P.F., Rodriguez-Valera F., Pedros-Alio C., and Moreira D., Nature *409*, 603–607, 2001.
5. Moon-van der Staay S.Y., De Wachter R., and Vaulot D., Nature *409*, 607–610, 2001.
6. Breitbart M., Salamon P., Andresen B., Mahaffy J.M., Segall A.M., Mead D., Azam F., and Rohwer F., Proc Natl Acad Sci USA *99*, 14250–14255, 2002.
7. Breitbart M., Hewson I., Felts B., Mahaffy J.M., Nulton J., Salamon P., and Rohwer F.J., Bacteriol *185*, 6220–6223, 2003.
8. Fuhrman J.A., Nature *399*, 541–548, 1999.
9. Wommack K.E. and Colwell R.R., Microbiol Mol Biol Rev *64*, 69–114, 2000.
10. Hendrix, R.W., Curr Opin Microbiol *6*, 506–511, 2003.
11. Pedulla M.L., Ford M.E., Houtz J.M., Karthikeyan T., Wadsworth C., Lewis J.A., Jacobs-Sera D., Falbo J., Gross J., Pannunzio N.R., Brucker W., Kumar V., Kandasamy J., Keenan L., Bardarov S., Kriakov J., Lawrence J.G., Jacobs W.R. Jr., Hendrix R.W., and Hatfull G.F., Cell *113*, 171–182, 2003.
12. Kristian D., Stewart C.S., Flint H.J., and Shirazi-Beechey S.P., FEMS Microbiol Ecol *38*, 141–151, 2001.
13. Shannon C.E., MD Comput *14*, 306–317, 1997.
14. Chao A., Scandinavian J Stat. *11*, 783–791, 1984.
15. Casjens S., Mol Microbiol *49*, 277–300, 2003.
16. Boyd E.F., Davis B.M., and Hochhut B., Trends Microbiol *9*, 137–144, 2001.