Author/s:
Castro-Nallar, E;Chen, H;Gladman, S;Moore, SC;Seemann, T;Powell, IB;Hillier, A;Crandall, KA;Chandr, PS

Title:
Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage

Date:
2012-09-24

Citation:
Castro-Nallar, E., Chen, H., Gladman, S., Moore, S. C., Seemann, T., Powell, I. B., Hillier, A., Crandall, K. A. & Chandr, P. S. (2012). Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage. Genome Biology and Evolution, 4 (3), pp.382-393. https://doi.org/10.1093/gbe/evs017.

Persistent Link:
https://hdl.handle.net/11343/258971

Population Genomics and Phylogeography of an Australian Dairy Factory Derived Lytic Bacteriophage

Eduardo Castro-Nallar[1], Honglei Chen[2,5], Simon Gladman[3], Sean C. Moore[3], Torsten Seemann[4], Ian B. Powell[2], Alan Hillier[3], Keith A. Crandall[1,*], and P. Scott Chandry[3]

[1]Department of Biology, Brigham Young University

[2]Dairy Innovation Australia Limited, Werribee, Victoria, Australia

[3]CSIRO Food and Nutritional Sciences, Werribee, Victoria, Australia

[4]Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria, Australia

[5]Present address: CSIRO Livestock Industries, Australian Animal Health Laboratories, East Geelong, Victoria, Australia

*Corresponding author: E-mail: keith_crandall@byu.edu.

Accepted: 14 February 2012

Data deposition: GenBank accession numbers for the sequenced genomes are JQ740787-JQ740814. Data sets used for all the analyses are available upon request.

Abstract

In this study, we present the full genomic sequences and evolutionary analyses of a serially sampled population of 28 *Lactococcus lactis*–infecting phage belonging to the 936-like group in Australia. Genome sizes were consistent with previously available genomes ranging in length from 30.9 to 32.1 Kbp and consisted of 55–65 open reading frames. We analyzed their genetic diversity and found that regions of high diversity are correlated with high recombination rate regions (*P* value = 0.01). Phylogenetic inference showed two major clades that correlate well with known host range. Using the extended Bayesian Skyline model, we found that population size has remained mostly constant through time. Moreover, the dispersion pattern of these genomes is in agreement with human-driven dispersion as suggested by phylogeographic analysis. In addition, selection analysis found evidence of positive selection on codon positions of the Receptor Binding Protein (RBP). Likewise, positively selected sites in the RBP were located within the neck and head region in the crystal structure, both known determinants of host range. Our study demonstrates the utility of phylogenetic methods applied to whole genome data collected from populations of phage for providing insights into applied microbiology.

Key words: bacteriophage, selection, phylodynamics, pyrosequencing, population genomics.

Introduction

Bacteriophage (phage) are some of the fastest evolving and abundant entities in nature (Hendrix 2003). Phage are ubiquitous, co-occurring with bacteria in environments as varied as soils, oceans, and even in human intestines. They are also present in any industrial process that capitalizes on bacterial metabolism, for example, the biotechnological production of chemicals and food products (Hendrix 2003; Brussow et al. 2004). In order to acidify milk during the production of fermented foods, such as cheese, buttermilk, and sour cream, different strains of *Lactococcus lactis*, a Gram-positive bacteria, are used as starter organisms primarily to ferment lactose to lactic acid. In particular, the cheese industry has been troubled by phage infections that can delay or halt fermentation. Given that phage are found in raw milk and can survive pasteurization (Madera et al. 2004), strict infection control measures and careful strain selection are required to mitigate potential phage induced dairy fermentation failures. The *L. lactis* phage are members of one of the largest phage orders, *Caudovirales* and are highly diverse both genetically and morphologically. This order contains three families, Myoviridae (with long contractile tails), Siphoviridae (with long noncontractile tails), and Podoviridae (with short tails). Lactococcal phage are mainly members of the Siphoviridae family, with a few members from the Podoviridae family. The three most prevalent groups of *L. lactis* phage isolated from diary environments are c2, 936, and P335; where the first two are

virulent lytic phage and the last have been reported as virulent temperate phage.

Probable sources of phage infecting dairy fermentations include raw milk, growth supplements, starter strains possessing temperate phage integrated into their genomes, factory equipment, and workers. Lytic phage infections can cause bacterial cell lysis, with subsequent consequences on the rate of acid production in the fermentation process. In cheese factories, delays to fermentation can cause significant difficulties in a process that is based on a perishable starting material (milk) that cannot be stored in the event of delay. Phage infections can also lead to negative repercussions in flavor and texture of the final product, which can result in significant economic losses.

Previous research has deciphered some portions of the replicative cycle of 936-like phage with special attention to particle adsorption and naturally occurring phage resistance mechanisms (Boucher et al. 2000; Ledeboer et al. 2002; De Haard et al. 2005; Tremblay et al. 2006). The first interaction of a phage particle and a bacterium is mediated through the specific recognition between the phage receptor binding protein (RBP), located at the tip of the tail, and the host cell receptor distributed over the cell surface. It is known that *L. lactis* phage adsorb initially to the cell surface and likely bind to various carbohydrates containing rhamnose, glucose, or galactose (Tremblay et al. 2006). This adsorption step is reversible for c2 phage, which need a secondary interaction with the bacterial cell wall through a predicted membrane attached protein (PIP). However, 936 and P335 phage do not use a secondary receptor. A number of naturally occurring plasmid and chromosomally encoded phage resistance mechanisms have been described in *L. lactis* strains. Among these, more than 20 Abortive infection (Abi) systems have been described (Boucher et al. 2000; Chopin et al. 2005). These phage resistance mechanisms act after phage adsorption, DNA penetration, and early gene expression and generally result in death of the infected cell and a diminished number of phage progeny. In addition, a novel antiphage strategy has been developed by raising antibodies against 936 RBP in Llama (*Lama glama*) with substantial inhibitory results (De Haard et al. 2005). However, as in other pathogen–host interactions, the arms race is usually led by the pathogen and in this system, lactococcal phage have evolved mechanisms to avoid cell defenses and escape host cell resistance mutations.

Caudovirales evolution is thought to be driven by the horizontal exchange of genes between distantly related phage and hosts. This mosaicism is inferred from the observation of abrupt changes in sequence similarity (Casjens 2005). However, particular to 936-like phage, this hypothesis awaits experimental and informatic evidence.

The 936-like phage are the species of lactoccocal phage most frequently isolated from dairy fermentations (Deveau et al. 2006), have been readily isolated from whey samples during the cheese productive process around the world (de Fabrizio et al. 1991; Crutz-Le Coq et al. 2002; Madera et al. 2004; Deveau et al. 2006; Fortier et al. 2006; Suárez et al. 2008; Hejnowicz et al. 2009), and probably occur wherever *L. lactis* occurs. The evolution of the 936-like group is not well studied and while numerous complete genomes have been sequenced, most previous studies have utilized random samples within a factory or from a variety of factories. These analyses have been essentially based on sequence comparison but not in the context of a rigorous phylogenetic framework (Crutz-Le Coq et al. 2002; Fortier et al. 2006; Rousseau and Moineau 2009). In addition, nothing is known about their evolution from a population genetics perspective—their phylodynamics or dispersion over a determined geographic region (Pybus and Rambaut 2009). Moreover, open reading frames (ORFs) involved in key functions have never been analyzed for diversifying selection in a phylogenetic context.

In the present study, we inferred from full genome sequence, the phylogenetic relationships within an Australian population of 936-like phage sampled serially from dairy factories over an 8-year period (1994–2001). In addition, we tested whether the population remained constant through time, and how the isolates dispersed over the geographic region from which they were sampled. Finally, we tested for evidence of diversifying selection in a set of relevant genes. Thus, the aim of the work was to provide insights into the historical relationships of this group of phage; in particular to know how they have dispersed through Australian factories, whether the population has changed in size through time, and whether there are alleles that may provide an increased fitness to the phage that carry them.

Materials and Methods

Phage Sampling and Dairy Factories

Phage samples were collected by the Cultures Division of Dairy Innovation Australia Ltd. (DIAL) as part of routine screening of factory whey samples for phage and spanned a time window from 1994 to 2001. Phage were categorized by factory of origin (factories were identified with random letters [B to I]), date of collection, and known host range. Whey samples from the Australian dairy factories serviced by the Cultures Division of DIAL are routinely screened for phage capable of infecting a large panel of *L. lactis* host strains. Whey samples containing phage are purified to single plaque and the isolated phage stored. The detection of phage resistant to *L. lactis* host defenses is used to inform starter strain selection.

The database of factory-derived phage was queried on *L. lactis* hosts, then two host range groupings were chosen to provide phage for phylogenomic analysis. All of the *L. lactis* host strains were commercial cheese starters routinely used by the industry at the time of the collection period provided by the Cultures Division of DIAL. The first host range group

was defined by strains infecting *L. lactis* strains ASCC 222 and ASCC 385. The second host range group were *L. lactis* strains ASCC 92, ASCC 818, and ASCC 962. Host range groups were selected on the basis that the phage in these groups spanned a large geographic distribution and time period.

The eight dairy factories from which phage were derived were anonymized to protect the identities of the six dairy companies and prevent giving false impressions about phage abundance. Production scale is variable between factories with all factories utilizing over 100,000 l of milk per day and some exceeding 1,000,000 l per day. The basic method of production for these samples is relevant for interpretation of the data and will be described below.

Lactococcus lactis starter cultures are all produced by the Cultures Division of DIAL as frozen cell concentrates that are then distributed to the relevant factories by direct courier. Cell concentrates are then inoculated into dedicated bulk starter culture production vessels. These vessels of approximately 20,000 l are maintained in separate sections of the factory and isolated under strictly controlled conditions of access and hygiene to prevent phage contamination. Starter cultures are then pumped from the bulk starter vessels into the cheese production vats throughout the main factory. Production cycles are approximately 3 hours in duration (time to when the next batch of starter cells will be pumped into the cheese vats). Given the scale of production, the dairy factories have surprisingly limited numbers of staff entering the cheese production area and strictly controlled access to the starter production area. Dairy factory staff is involved in greatest numbers after the risk of phage infection has ended at the final stages of the production when 20 kg bulk cheese blocks are packaged for maturation. The degree of movement between factories of staff working in the production area would be very limited and interchange between companies would be unlikely.

As stated above, the samples for this project were derived from eight dairy factories controlled by six different independent companies spread over several states in Australia. To avoid the build up of phage to a given starter culture or mix of starter cultures, many factories rotate between different collections of starter organisms, but the precise utilization of starters is commercial and kept in confidence. All starters were in use at some time during the collection period in all factories. Milk is a perishable and bulky commodity that is generally utilized by the nearest factory. The amount of milk coming from any one dairy farm going to multiple factories would be very low or nonexistent, but the precise sources for milk from these production periods is not available; so this cannot be entirely excluded. Given the retrospective nature of this analysis, no data are available on the sources of the other two ingredients for cheese production, enzyme coagulant, and salt. Similarly, records of the movements of staff, sales-persons, and suppliers at particular factories even if kept by the factories would be of commercial sensitivity and cannot be analyzed for this study.

Sequences and Alignment

Twenty-eight phage genomes were selected for complete genome sequencing by pyrosequencing (Roche Genome Sequencer FLX by Department of Primary Industry, Victoria) supplemented, on an as needed basis, by Sanger sequencing in selected regions to resolve ambiguities and low-quality positions. Phage sequences were computer annotated with the Genome Annotation Transfer Utility (Tcherepanov et al. 2006) and manual curation.

From 28 complete phage genomes, each ORF was taken and aligned individually with its orthologs. DNA sequences were visualized in Seaview 4.2.7 (Gouy et al. 2010), converted into protein and then aligned with MAFFT using L-INS-i algorithm (Katoh et al. 2005). Protein sequences were back translated into their original nucleotide sequences for further analyses. Additionally, the resulting alignment was inspected by eye to correct obvious misaligned positions. Since not all phage isolates have the same gene content or gene order, extended regions of gaps were placed when genes were absent/present in some isolates during the alignment. It is known that insertions and deletions are common in the evolution of phage; so these were interpreted as novel characters (sensu; Egan and Crandall 2008). Consequently, in further analyses, gaps were considered as a fifth character unless otherwise stated.

Genetic Diversity and Recombination

Genetic diversity (Θ) was estimated for the Australian isolates of 936-like phage using both a Watterson estimate (Watterson 1975) assuming infinite sites as implemented in DNAsp ver. 5 (Librado and Rozas 2009) and a modified Watterson estimate that relaxes the infinite sites model assumption as implemented in Pairwise, LDhat 2.5 (Hudson 2001; McVean et al. 2002). The recombination parameter, rho ($\rho = 2N_e r$), was estimated using a composite-likelihood method as implemented in Interval, also part of the LDhat 2.5 package. The correlation between Θ and ρ was tested using a nonparametric test of correlation as implemented in PASW18 (Kendall's τ test; Kendall 1938).

Phylogenetic Inference

A phylogeny for phage isolates was estimated using Bayesian inference as implemented in BEAST 1.5.3 (Drummond and Rambaut 2007). This method was chosen because of its ability to account for uncertainty in phylogenetic estimates, the variety of clock models and tree priors available, and the relative speed of the computation. The best-fit model of evolution was selected under Akaike information Criterion (AIC) (Akaike 1974) and parameter values using model averaging as implemented in jModelTest 1.0.1

(Posada 2008). The best-fit model was used for subsequent analyses; nevertheless, base frequencies and rates were estimated in BEAST. The Molecular Clock hypothesis was tested on the data against an uncorrelated Lognormal Relaxed Molecular Clock by Bayes Factors as implemented in BEAST 1.5.3 (Drummond et al. 2006). The molecular clock was calibrated with sampling dates and a uniform prior distribution for the mean rate of substitution ranging from 10^{-3} to 10^{-8} substitutions per site per year (s/s/y) (Holmes 2009). Bayesian posterior probabilities were determined by running four chains (independent runs) of 100 million steps each. Parameters were sampled every 10,000 steps. A burn-in of 10% was used to discard parameters initially sampled. Autocorrelation was assessed by checking the Effective Sample Size statistic (>200). Convergence and mixing of the Markov chains were assessed using Tracer 1.5 (http://beast.bio.ed.ac.uk/Main_Page). Different runs tend to converge at the same parameter values in the stationary phase. Mixing was evaluated qualitatively by looking at oscillations in the parameter space explored by the trace of each chain. Trees were summarized in TreeAnnotator (http://beast.bio.ed.ac.uk/Main_Page) by targeting the Maximum Clade Credibility tree and support values represent posterior probability estimates.

Past Population Dynamics

Past population dynamics of the Australian phage, all the putative protein-coding genes were inferred in BEAST 1.5.4 using the Extended Bayesian Skyline Plot (eBSP) model as a tree prior and an uncorrelated relaxed molecular clock (Drummond et al. 2005). The number of grouped intervals was co-estimated with the relative genetic diversity through time. Bayes factors were used to test different tree priors representing different demographic scenarios: expansion, exponential and constant growth, and the eBSP.

Phylogeographic Analysis

Phylogeographic analyses were performed by ancestral reconstruction of discrete states in a Bayesian statistical framework (Lemey et al. 2009). The most parsimonious description of the diffusion process was identified by a discrete analysis (see http://beast.bio.ed.ac.uk/Tutorials). The inferences were summarized and geographical ancestral states were color-coded onto the tree topology.

Selection Analysis

Adaptive selection was assessed by estimating the ratio of nonsynonymous to synonymous nucleotide substitution rates (dN/dS) in a site-by-site basis as implemented in www.datamonkey.org (Delport et al. 2010) using Fixed-Effects Likelihood (FEL), Random-Effects Likelihood (REL), and Single Likelihood Ancestor Counting (SLAC). Positively selected sites were mapped onto the crystal structure of the RBP (PDB accession 2BSD; Spinelli et al. 2005) using Cn3D 4.1 software available at ftp://ftp.ncbi.nih.gov/cn3d/Cn3D-4.1.msi.

Results

Genome Organization, Genetic Diversity, and Recombination

The genome length of Lactococcal phage belonging to the 936-like species has been reported to range from 28 to 32 Kbp and to consist of 50-64 ORF (Deveau et al. 2006; Rousseau and Moineau 2009). Similarly, the new phage genomes analyzed for this study ranged in length from 30.9 Kbp to 32.1 Kbp (31.5 on average) and consisted of 55 to 65 ORFs (61 on average) that, to a great extent, were conserved in gene order and sequence homology (fig. 1). Homologues of 41 ORFs were present in all 28 phage studied from a total pool of 74 ORFs. The remaining 33 ORFs not present in all phage but detected in one or more phage were primarily located in the early-transcribed region.

For phylogenetic analysis, all the ORFs from all the phage (not just orthologs common to all phage) were assembled into a coding genome that, when aligned, had a length of 34,086 bases (no missing data) (fig. 2). The genetic diversity across the coding genome was measured using a sliding window approach that looked at genetic diversity (Θ/site) and recombination rate (ρ/kb) (fig. 2). Three main regions of relatively high genetic diversity (Θ per site>0.10) were identified. The leftmost peak is located in the late-expressed region and mainly spans the neck portal protein. The second peak is located close to the late/early gene expression junction and includes some genes whose presence is variable across the phage. This area of increased diversity includes part of the tape measure protein, the holin, the lysin, and some early genes of unknown function with the largest peak over the receptor binding protein. The rightmost peak spans the start of the early region and all of the middle expressed genes. This region encodes the DNA polymerase subunit, the Holliday junction endonuclease, and several genes of unknown function. Θ was estimated at 0.03009 using a likelihood method that relaxes the infinite sites assumption (LDhat) and at 0.04497 using the infinite sites Watterson estimate (DNAsp5). The recombination rate averaged across all segregating sites was estimated as $\rho = 0.0015$, while the recombination rate for each site varied considerably and correlated significantly with regions of high genetic diversity ($R = 0.665$ and two-tailed P value = 0.01, Kendall's ρ Test of correlation).

Phylogenomic Analysis

The best-fit model of evolution for the phage genome alignment, chosen of 88 implemented models, was TVM+I+G (transversion model; Variable base frequencies, variable transversions, transitions equal). Because one of the purposes of this study was to infer historical relationships with

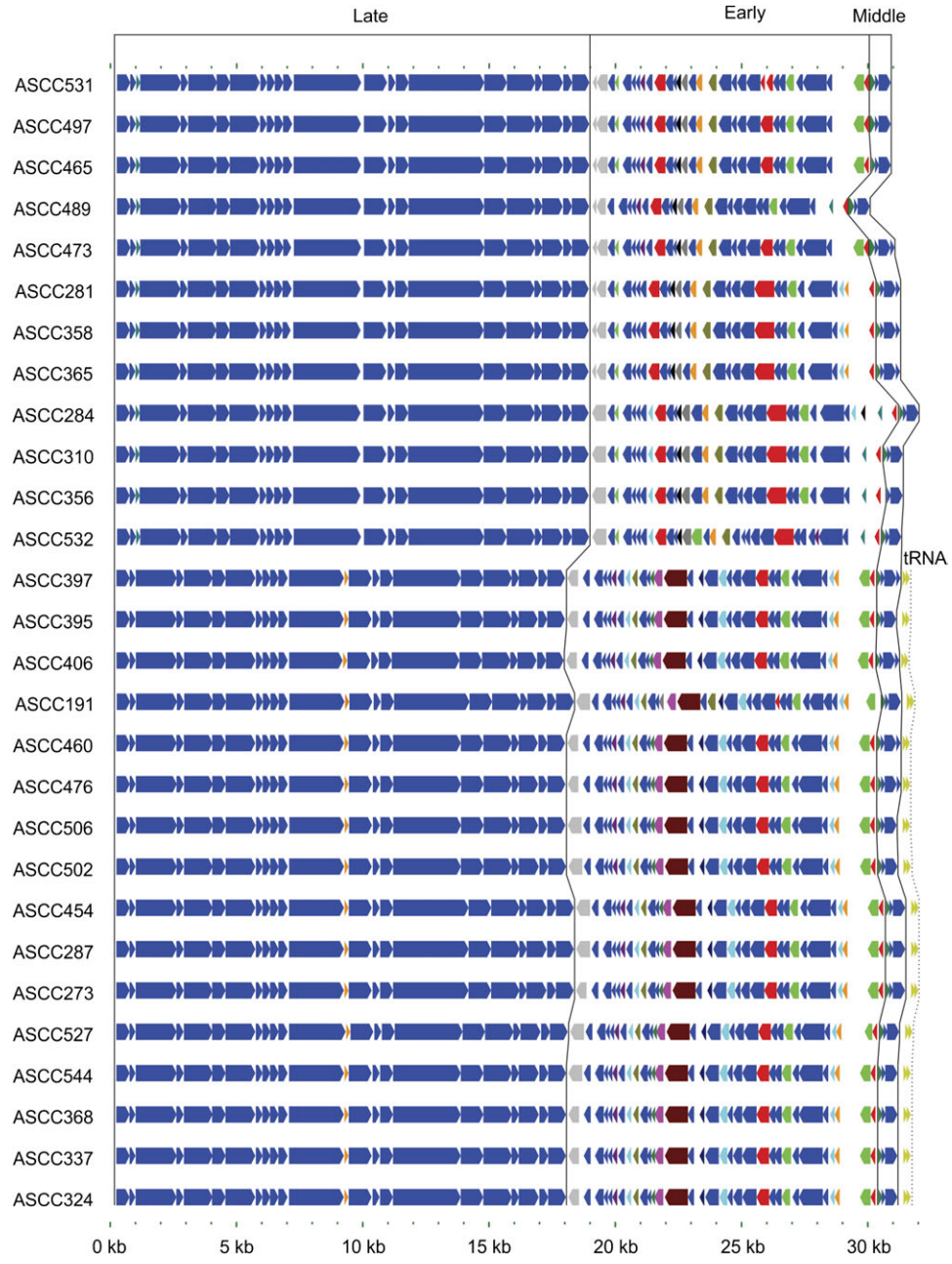

Fig. 1.—Schematic genome alignment of Australian cheese factory derived 936-like phage used for this study. Phage genomes are represented as colored arrows indicative of the direction of transcription of protein and tRNA encoding genes. Those orthologs conserved in all phage are colored dark blue while those present in a subset of phage are in a variety of colors. The gray ortholog present in all genomes is an intact gene in some genomes and an apparent pseudogene in others. Black lines indicate the boundaries between those regions encoded during the early, middle, and late phases of transcriptions. The tRNA genes encoded by some phage are also indicated by olive green arrows.

respect to divergence times, a molecular clock was tested using Bayes factors. The relaxed uncorrelated lognormal model was then used to infer a phylogeny (fig. 3). The tree prior used was chosen based on Bayes Factors (table 1). Four demographic models were tested and the one with highest support, eBSP, was chosen for subsequent analysis. The phylogenetic inference showed strong support for all the clades and a spatio-temporal relationship between the isolates. However, clades and subclades did not map cleanly with geographic location. For example, most factory E isolates were closely related to each other (subclade within lower large clade in fig. 3), while the others were scattered throughout the tree, suggesting potential transmission routes (see below). In this particular case, factory E isolates

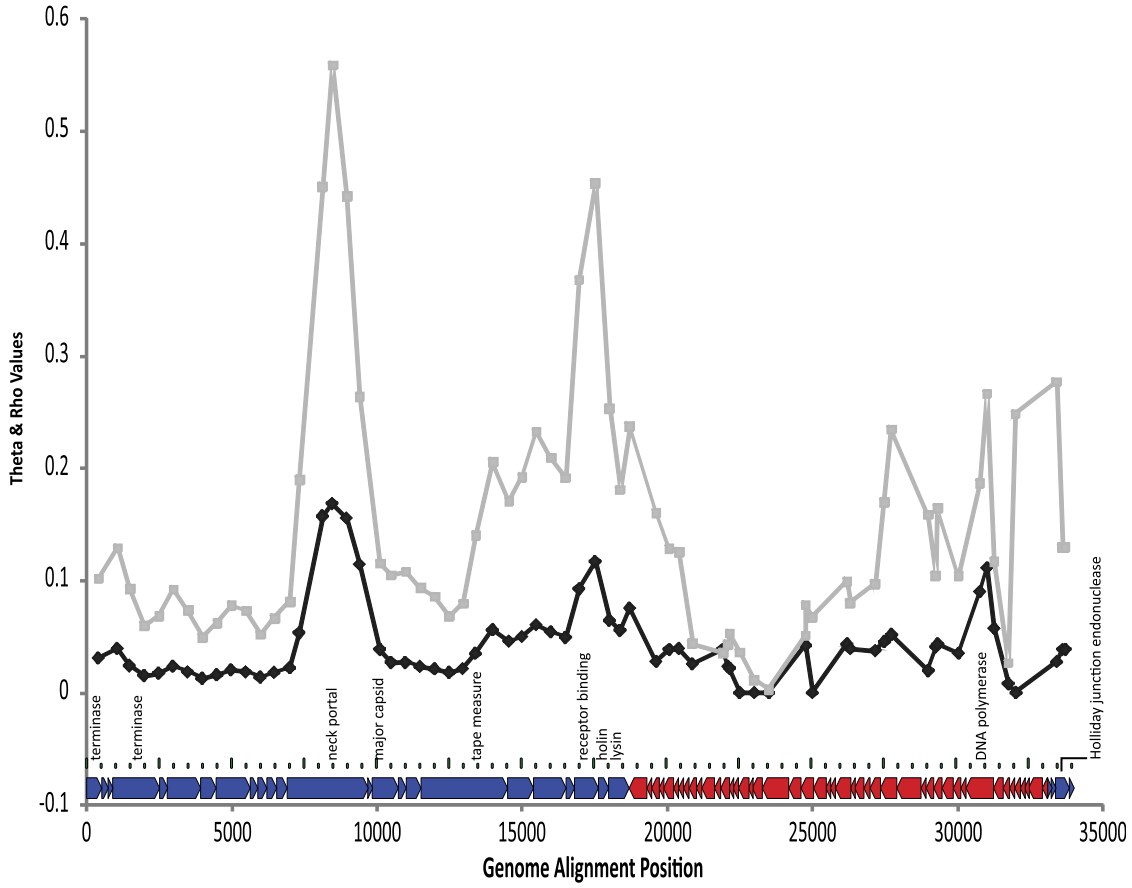

Fig. 2.—Plot of recombination rate and genetic diversity for aligned phage genomes. Genetic diversity plotted as Θ per site (black line) and recombination rate ρ per kb (gray line) as a function of time as determined in dnaSP 5.0 and LDhat 2.5, respectively. A schematic representation of all the orthologs present in the concatenated alignment used for phylogenetic analysis annotated with selected genes to provide a point of reference is below the graph. Orthologs colored red are early transcripts while those in blue represent late and middle transcripts (left and right, respectively).

did not form a monophyletic group since some members are related to phage isolated from other locations. The same result applies to isolates from other disparate locations. The structure of the tree matched perfectly with the known host specificity, i.e., the upper clade contains all isolates capable of infecting *L. lactis* ASCC host strains 222 and 385. Likewise, the lower clade contains all isolates capable of infecting *L. lactis* ASCC host strains 92, 818 and 962. Branch colors represent the localities (factories) where the phage was present and where the ancestral state was inferred. Additionally, we obtained age estimates for the upper and lower clades. Likely values for the upper clade age range from 14 to 128 years old (95% credible interval, CI) with a mean of 60 years. In turn, the mean age for the lower clade was estimated at 110 years with a 95% CI from 27 to 236 years; much older than the upper clade, but overlapping in their respective credible intervals.

Phylodynamics

The phage dataset was tested against four demographic models including the extended eBSP. The eBSP model had the highest support and was chosen for subsequent analysis. To estimate the historical diversity of this group of phage, an eBSP analysis was performed (fig. 4). Mean and median values for relative genetic diversity (*y* axis) together with credibility intervals were plotted through time (*x* axis, time 0 represents the earliest sampling, i.e., 2001). We inferred a rather constant genetic diversity and population size, with slight variation in the credible intervals. The genetic diversity and population sizes are in part dependent on the number of hosts available for infection. Since the production processes in the factories are standardized and the time window is small (8 years for a DNA virus), it is possible that the analysis represents the limited/constant capacity of the factories during such years or that the temporal signal was not present in the sequences.

Phylogeography

The phylogeographic pattern of dispersion shows multiple sources for several isolates (fig. 5). The diagram depicts a matrix of factories reflecting the relative distances

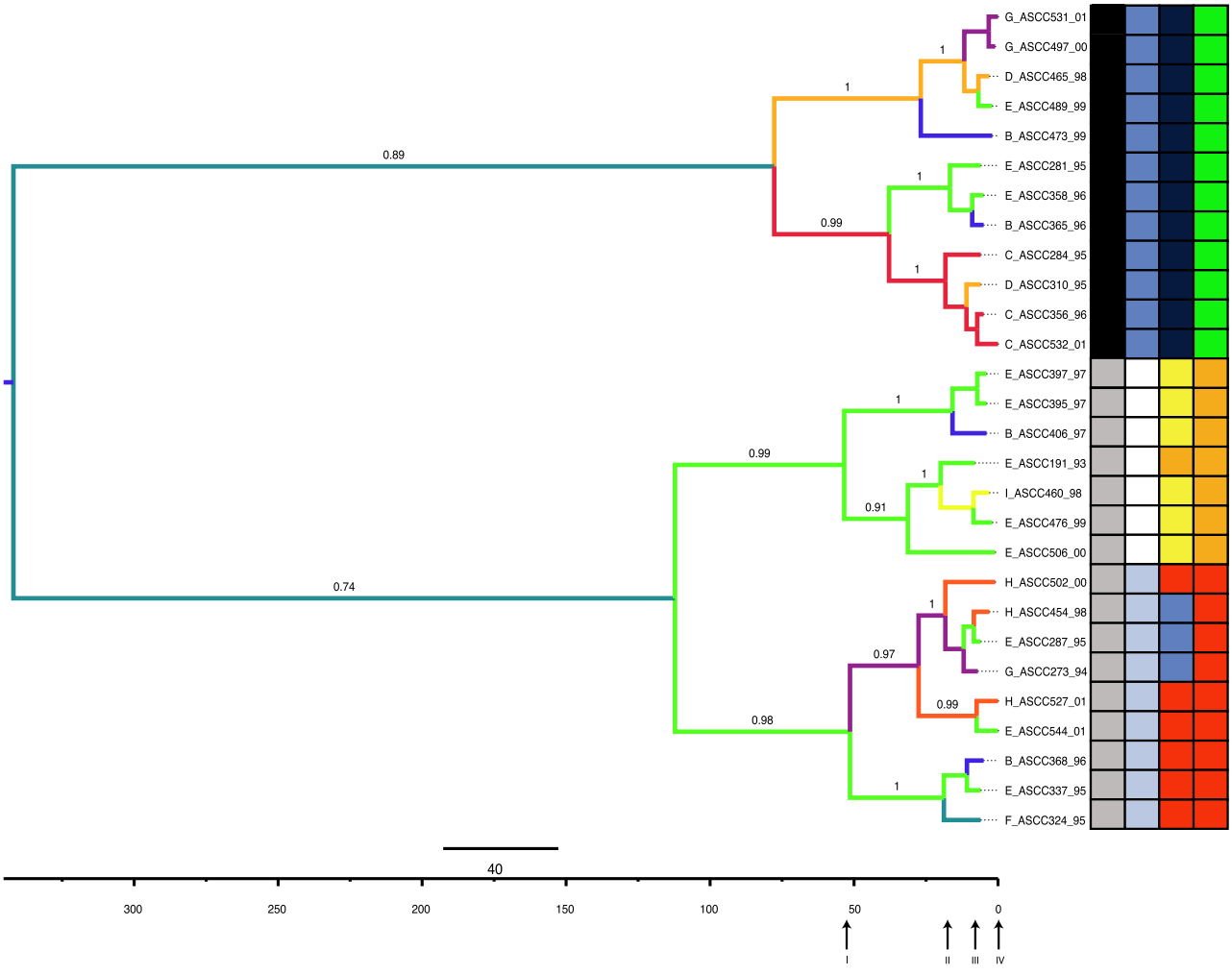

Fig. 3.—Maximum clade credibility phylogeny of the twenty-eight 936-like phage. Branch values represent posterior probability support. Branches are colored according to the ancestral state of geographic location. Taxa names indicate factory key_phage_year of isolation. The tree is midpoint rooted. On the grid: first column indicates host tropism, black: 222/385 and gray: 92/818/962 *Lactococcus lactis* host strains. Second, third, and fourth columns are codons under positive selection 155, 165, and 167, respectively. Blue: methionine; white: gap; light blue: leucine; dark blue: phenylalanine; yellow: tyrosine; orange: serine; red: threonine; green: alanine. Arrows below time line relate dispersion analysis described in figure 5.

between factories, but not linked to any geographic features. The pattern suggests a human-driven dispersion throughout the region rather than natural means because distant locations are connected in different directions instead of, for example, by means of natural elements like wind directions, bird migrations, etc. For instance, ancestral lineages in factories C, D, and E were evolved from an ancestor present in factory F. Likewise, the factory F ancestral lineage gave rise to lineages present in factories G and I (fig. 3).

Table 1

Bayes Factors Hypothesis Testing on Demographic Tree Priors for Australian 936 Phages

	Model				
Model	**Constant Population**	**Expanding Population**	**Exponential Growth**	**Extended Bayesian Skyline Plot**	**Likelihood**
Constant population	—	15.07	16.794	−31.877	−92,650.146
Expanding population	−15.07	—	1.725	−46.947	−92,684.846
Exponential growth	−16.794	−1.725	—	−48.671	−92,688.817
Extended Bayesian Skyline Plot	31.877	46.947	48.671	—	−92,576.747

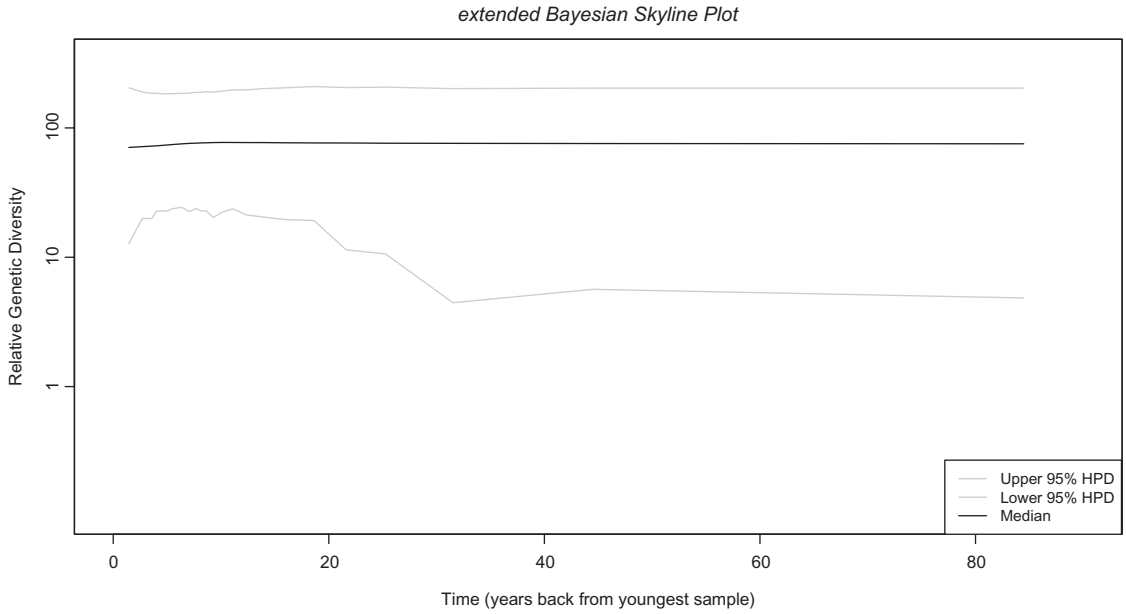

Fig. 4.—Plot of genetic diversity over time. Extended Bayesian Skyline Plot describing population changes as a function of time. The *y* axis represents relative genetic diversity and *x* axis represents years backward in time from the most contemporaneous sample, that is, 2001. Mean and median are shown with 95% credibility intervals.

Selection Analysis

Five protein-coding genes were chosen for selection analysis: RBP gene (host specificity), tape measure protein (tail length determinant), major capsid protein (head structure), lysin (bacterial cell wall degradation), and neck portal (capsid structure). Sites in which the d*N*/d*S* rates ratio was statistically greater than one were considered to be under positive selection (Sharp 1997). The REL method found more sites than the SLAC and FEL methods (table 2). Since the crystal structure for the RBP protein has been resolved (Spinelli et al. 2005), the selected sites were mapped onto the structure to visualize the positioning of these in a 3D context (fig. 6). The functional polyprotein is a homotrimer (fig. 6*B*) in which each monomer is intertwined with each other. All positively selected sites were located at the head and neck structures where host range specificity is known to reside. Accordingly, these sites were mapped to the tips of the phylogeny along with host specificity (fig. 3 grid).

Discussion

Genome Organization, Genetic Diversity, and Recombination

It has been commonly assumed that phage evolve in a modular fashion, by shuffling conserved cassettes of genome regions due to the observed similarity among these regions when comparing distant phage (Casjens 2005). Here, we looked for recombination rates and genetic diversity over all segregating sites (fig. 2). Low recombination rates were inferred over the first ~8 kb, which represents late expressed genes. This observation is somewhat expected since this region contains packaging and morphogenesis genes; therefore, changes here are likely to be purified due to structural and functional constraints, that is, a change here is more likely to yield a protein product unable to form viable viral particles. In addition, over the central region of the genome alignment, a spike centered on the receptor binding protein ~17.5 kb and another region between 27.7 and 33.4 kb exhibit high recombination rates. The former spike is also part of the late expressed genes; whereas, the latter region lies within early/middle expressed genes. However, it seems that genetic diversity along the genome is more the result of nucleotide substitutions rather than recombination as evidenced by the ratio $\Theta/\rho = 20.06$. Needless to say, further experimental analyses will shed more light on the relative role of recombination in the evolution of Lactoccocal phage. For now, the correlation of higher genetic diversity with higher recombination rate (fig. 2) is suggestive of an important role for recombination in generating novel combinations of alleles in the phage populations and the discrete regions of high recombination rate suggest that modularity could be a common feature in this group.

Phylogenomic Analysis

In order to infer historical relationships between isolates, a relaxed uncorrelated molecular clock was used and calibrated with sampling dates (Drummond et al. 2002, 2005; Drummond and Rambaut 2007). The phylogeny inferred here shows two major clades that share a common ancestor. Interestingly, each clade contains all the isolates

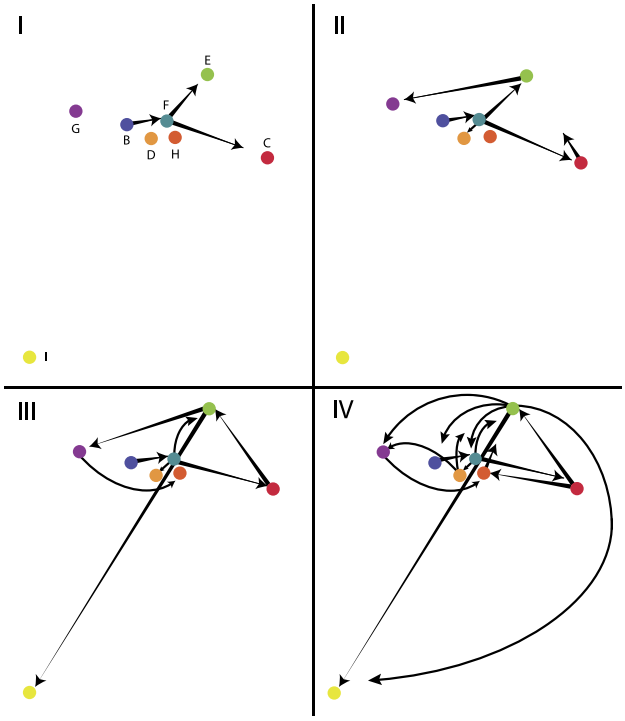

Fig. 5.—Dispersion pattern of Australian 936-like phages. Phylogeographic analysis for time-points I–IV (root to tips; linked to positions designated in fig. 3) showing the dispersion routes of the ancestral lineages that explain the current sampling geographic distribution. Letters indicate phage from eight different geographic locations, color-coded as in figure. I–IV are four "time slices" arbitrarily chosen to describe the dispersion pattern. Although no scale is provided, the overall diagram spans hundreds of kilometers.

able to infect one of the two main groups of *L. lactis* host strains used in this study (figure 3). Ancestral states were color mapped in the topology showing a geographical structure that can be further used to infer phylogeographic patterns (see below). For the lower clade, it is clear that ancestral lineages were present somewhere near factory E and then were dispersed and independently evolved to give rise to all the diversity sampled that is capable of infecting *L. lactis* ASCC host strains 92, 818, and 962 in different sampled localities.

Although credible intervals of timing estimates for both clades were partially overlapping, the mean values are 50 years apart. This suggests that this lineage has had more time to diversify and disperse than the upper clade. This is evidenced by the dispersion pattern inferred (fig. 5) in which isolates with 92, 818, and 962 host strain tropism exhibited a wider range. In contrast, the upper (younger) clade has a narrower geographic range (fig. 5). It is worth mentioning that although the tips of the tree are associated with this host range, changes in host strains could have occurred in the past in response to phage detection, for which ancestors of the sampled phage might have been associated with other *L. lactis* strains.

Phylodynamics

Demographic analysis revealed a constant population size/genetic diversity through time. If we assume that effective population size in viruses, looking back in time, reflects the number of infections that occurred, we could say that this result (fig. 4) might represent the stable number of outbreaks through time (for the time period analyzed). Further information could shed light on this issue as it has been done in other studies where census sizes or recorded demographic/outbreak information has been correlated with past population dynamics inferences (Pérez-Losada et al. 2005; Bennett et al. 2010; Tazi et al. 2010). Interestingly, Bayes Factor analysis did not show the strongest support for the constant demographic model (table 1). Instead, the eBSP model showed the strongest support. This could be explained because this model, in a piecewise fashion, tries

Table 2
Positive Selected Sites (d*N*/d*S* > 1) on Australian 936 Phages Genes

	Selection Detection Method			
Gene (Overall d*N*/d*S*)	**SLAC***	**FEL***	**REL****	**Consensus (At Least in Two Methods)**
RBP (0.44962)	165, 167	155, 167	45, 57, 137, 139, 155, 167, 173, 223, 229, 233, 259, 261	155, 167
Major capsid protein (0.194507)	Not found	Not found	Not found	—
Endolysin (0.207828)	Not found	Not found	Not found	—
Tape measure protein (0.21852)	Not found	Not found	42, 158, 241, 319, 320, 390, 527, 530, 584, 678, 729, 738, 762, 770, 802, 818, 847, 849, 854	—
Neck portal protein (0.472364)	Not found	Not found	Not found	—

Note.—*P value < 0.1; **Bayes factor > 50. All analyses performed online at www.datamonkey.org.

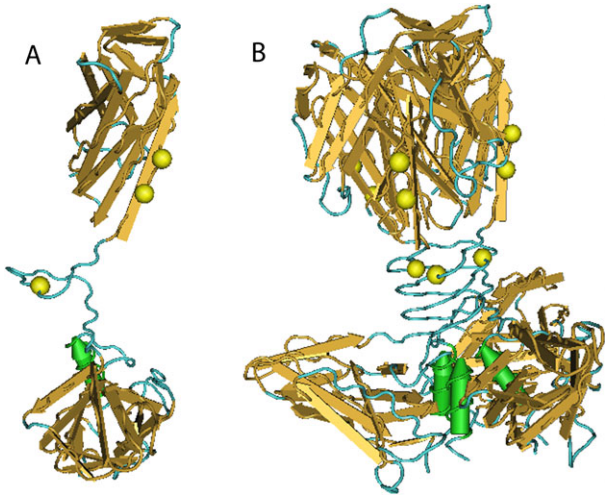

Fig. 6.—Molecular mapping of positive selected sites detected in RBP. (*A*) Monomer structure with three sites mapped on it; (*B*) Homotrimer (active structure) with sites mapped in each chain. Molecular structure retrieved from NCBI structure database, PDB code 2BSD.

to accommodate population dynamics giving an overall best fit of the model to the data (Heled and Drummond 2008).

Phylogeography

The dispersion pattern (fig. 5) suggests artificial dispersal because of the multiple routes and long distances the lineages have taken to reach the present distribution. If the dispersal of lineages over the geographic region was explained by some sort of natural carrier or whatever natural means, we would expect geographically close factories to have genetically closely related phage lineages, for example, factories H, D, and B, as two neighbor factories would share phage having a most recent common ancestor. However, this seems not to be the case. As described in the cheese production outline above, the complex nature of the dairy manufacturing environment, the unrecorded movements of sales persons, technicians, and consumable suppliers combined with the lack of historical production records, this retrospective study does not lend itself to a comprehensive epidemiological analysis. The specific mechanism of transfer between factories would be better addressed by a well-designed prospective study to provide additional evidence for the artificial movement hypothesis. Such a study would be challenging since it would require a substantial level of factory testing for phage detection, substantive recording of personnel movements (both factory personnel and external personnel), and exhaustive recording of incoming consumables.

Natural Selection

The final step of this study was to look for diversifying selection in some biologically important genes (table 2 and fig. 6). We used three methods to achieve some degree of confidence in the potentially selected sites. It has been shown that with large datasets (>50 taxa) all methods tend to converge to the same answer (Kosakovsky Pond and Frost 2005). However, when using small data sets (this case, 28 taxa) a consensus-based inference is advisable. One of the methods, the REL method, assumes that each rate is represented by a simple distribution and is commonly recognized as a liberal method. On the other hand, the SLAC method provides a "quick and dirty" alternative. In turn, the FEL method was used to infer the more conservative estimate of positive selected sites. Not surprisingly, REL found many more sites under selection than FEL and SLAC when examining RBP and tape measure genes. Regarding major capsid protein, endolysin, and neck portal genes, it was not possible to find any sites under positive selection with any of the three methods. This is likely due to the fact that little variation was found in those genes, the presence of identical sequences that reduced the data set even more, and the inherent conservativeness of the methods. Similarly, no positively selected sites were found in the tape measure protein gene sequence when inspected under SLAC and FEL methods. However, positively selected sites were found in RBP, the protein that determines host specificity. Fortunately, the crystal structure of an ortholog of this protein has been resolved (p2 phage, 936-like; Spinelli et al. 2005), so it was possible to map the selected sites onto the structure (fig. 6). The active form of the protein is a homotrimer whose main domains are: the shoulders, a β-sandwich attached to the phage tail; the neck, an interlaced β-prism; and the head, the receptor recognition domain composed of seven-stranded β-barrel (from bottom up in fig. 6*B*). We mapped positively selected sites found by all three methods and these were located in the neck and one toward the head. The neck is a rigid structure, homologous to viruses infecting host of different kingdoms (adenovirus, reovirus, other phage as well). It is thought that this structure along with the head have coevolved with their host and thus are responsible of host range (Spinelli et al. 2005). In fact, across 936-like phage, the shoulder structure is highly conserved (~90% amino acid identity), whereas neck and head show greater variability (down to 15%, fig. 2*E* in Spinelli et al. 2005). Interestingly, character states in codon 155 were related to monophyletic groups and to a large degree related to host specificity, with the upper clade having a methionine and the lower clade with either a gap or a leucine. Likewise, codons 165 and 167 also followed this pattern. It is known that RBP is the sole determinant of host specificity (Dupont et al. 2004). Thus, the fact that these sites were found to be under positive selection and located in relevant domains could indicate that they are somehow involved in the phage–cell interaction directly or indirectly by serving as a scaffold for a proper conformation or making contacts with the cell ligand. The use of this observation as a predictor of host specificity awaits

experimental confirmation. In addition, we also note that isolates from the same factories are related to either one bacterial host or the other, suggesting a low mutational barrier for the physiological adaptation between bacterial strains.

Conclusion

In this study, we showed that recombination rate is concentrated within a few regions over the genome with these increased recombination rates correlated with increased regions of genetic diversity providing evidence for the modular nature of phage genome structure. In addition, our whole genome phylogenetic analysis of a population of 936-like phage shows isolates cluster together according to their host tropism. Our phylogeographic analysis suggests the mechanism of dispersion is most likely human-associated movement and agrees with timing estimates. Furthermore, positively selected sites in the receptor-binding protein, the sole determinant of host range (Dupont et al. 2004), lies on the domains that interact with the host receptor and correlate well with host specificity. A comprehensive assessment of 936-like phage evolutionary features will necessitate extensive sequencing efforts along with the recording of isolate characteristics. Nevertheless, our results clearly show the dominant role of host specificity in the evolutionary dynamics of these phage and demonstrate the utility of evolutionary approaches to key questions in applied microbiology.

Acknowledgments

E.C.N. would like to thank Justin Bagley for providing useful comments. E.C.N. was funded by Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), Gobierno de Chile – Becas Chile, and by a Brigham Young University (BYU) Graduate Mentoring Award 2011–12.

Literature Cited

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automat Contr. 19:716–723.

Bennett SN, et al. 2010. Epidemic dynamics revealed in dengue evolution. Mol Biol Evol. 27:811–818.

Boucher I, Emond E, Dion E, Montpetit D, Moineau S. 2000. Microbiological and molecular impacts of Abik on the lytic cycle of Lactococcus lactis phages of the 936 and P335 species. Microbiology 146:445–453.

Brussow H, Canchaya C, Hardt W. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev. 68:560–602.

Casjens SR. 2005. Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microbiol. 8:451–458.

Chopin M, Chopin A, Bidnenko E. 2005. Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 8:473–479.

Crutz-Le Coq A, Cesselin B, Commissaire J, Anba J. 2002. Sequence analysis of the lactococcal bacteriophage biL170: insights into structural proteins and Hnh endonucleases in dairy phages. Microbiology 148:985–1001.

de Fabrizio SV, Ledford RA, Shieh YSC, Brown J, Parada JL. 1991. Comparison of lactococcal bacteriophage isolated in the United States and Argentina. Int J Food Microbiol. 13:285–293.

De Haard HJW, et al. 2005. Llama antibodies against a lactococcal protein located at the tip of the phage tail prevent phage infection. J Bacteriol. 187:4531–4541.

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457.

Deveau H, Labrie SJ, Chopin M, Moineau S. 2006. Biodiversity and classification of lactococcal phages. Appl Environ Microbiol. 72:4338–4346.

Drummond AJ, Ho S, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Nicholls G, Rodrigo A, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Drummond AJ, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Drummond AJ, Rambaut A, Shapiro B, Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 22:1185–1192.

Dupont K, Vogensen FK, Neve H, Bresciani J, Josephsen J. 2004. Identification of the receptor-binding protein in 936-species lactococcal bacteriophages. Appl Environ Microbiol. 70:5818–5824.

Egan AN, Crandall KA. 2008. Incorporating gaps as phylogenetic characters across eight DNA regions: ramifications for North American Psoraleeae (Leguminosae). Mol Phylogenet Evol. 46:532–546.

Fortier L, Bransi A, Moineau S. 2006. Genome sequence and global gene expression of Q54, a new phage species linking the 936 and C2 phage species of Lactococcus lactis. J Bacteriol. 188:6101–6114.

Gouy M, Guindon S, Gascuel O. 2010. Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27:221–224.

Hejnowicz MS, Golebiewski M, Bardowski J. 2009. Analysis of the complete genome sequence of the lactococcal bacteriophage biBB29. Int J Food Microbiol. 131:52–61.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. BMC Evol Biol. 8:289.

Hendrix RW. 2003. Bacteriophage genomics. Curr Opin Microbiol. 6:506–511.

Holmes EC. 2009. The evolutionary genetics of emerging viruses. Annu Rev Ecol Evol Syst. 40:353–372.

Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics 159:1805–1817.

Katoh K, Kuma K, Toh H, Miyata T. 2005. Mafft version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kendall MG. 1938. A new measure of rank correlation. Biometrika 30:81–93.

Kosakovsky Pond SL, Frost SDW. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol. 22:478–485.

Ledeboer AM, et al. 2002. Preventing phage lysis of Lactococcus lactis in cheese production using a neutralizing heavy-chain antibody fragment from llama. J Dairy Sci. 85:1376–1382.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol. 5:e1000520.

Librado P, Rozas J. 2009. Dnasp V5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Madera C, Monjardin C, Suarez JE. 2004. Milk contamination and resistance to processing conditions determine the fate of Lactococcus lactis bacteriophages in dairies. Appl Environ Microbiol. 70:7365–7371.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241.

Pérez-Losada M, Viscidi RP, Demma JC, Zenilman J, Crandall KA. 2005. Population genetics of Neisseria gonorrhoeae in a high-prevalence community using a hypervariable outer membrane porB and 13 slowly evolving housekeeping genes. Mol Biol Evol. 22:1887–1902.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol. 25:1253–1256.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 10:540–550.

Rousseau GM, Moineau S. 2009. Evolution of Lactococcus lactis phages within a cheese factory. Appl Environ Microbiol. 75:5336–5344.

Sharp PM. 1997. In search of molecular Darwinism. Nature 385:111–112.

Spinelli S, et al. 2005. Lactococcal bacteriophage P2 receptor-binding protein structure suggests a common ancestor gene with bacterial and mammalian viruses. Nat Struct Mol Biol. 13:85–89.

Suárez V, Moineau S, Reinheimer J, Quiberoni A. 2008. Argentinean Lactococcus lactis bacteriophages: genetic characterization and adsorption studies. J Appl Microbiol. 104:371–379.

Tazi L, et al. 2010. Population dynamics of Neisseria gonorrhoeae in Shanghai, China: a comparative study. BMC Infect Dis. 10:13.

Tcherepanov V, Ehlers A, Upton C. 2006. Genome annotation transfer utility (GATU): rapid annotation of viral genomes using a closely related reference genome. BMC Genomics. 7:150.

Tremblay DM, et al. 2006. Receptor-binding protein of Lactococcus lactis phages: identification and characterization of the saccharide receptor-binding site. J Bacteriol. 188:2400–2410.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.

Associate editor: Martin Embley