Syst Synth Biol (2015) 9:S57–S62
DOI 10.1007/s11693-015-9185-7

COMMENTARY

On genome annotation of Brucellaphage Gadvasu (BpG): discovery of ORFans for integrated systems biology approaches

Deepti Chachra[1] · Pushpinder Kaur[1,2] · Prasad Siddavatam[3] · Prashanth Suravajhala[3,4,5] · Hari Mohan Saxena[1]

Received: 4 August 2015 / Revised: 12 November 2015 / Accepted: 16 November 2015 / Published online: 21 November 2015

Abstract Brucellaphage Gadvasu (BpG) is a lytic phage infecting *Brucella* spp. Brucellaphages contain dsDNA as genetic material and are short-tailed particles with host-specificity. Here, we report the challenges on annotation in the complete genome sequence of BpG when compared with that of a recent broad host-range brucellaphage Pr, an original reference genome. The extracted DNA was subjected to genome sequencing with Illumina technology and assembled using SSAKE/Velvet. A significant number of genes were found to be similar between the phages with sequence analysis revealing conserved open reading frames that correspond to 33 gene ontology classifiers, transcriptional terminators and a few putative transcriptional promoters. The analyses revealed that the genome constitutes 1269 contigs and 275 genes encoding 260 proteins. The sequence comparison from the reference data indicated that the genome shares an approximately 70 % nucleotide similarity and differs mainly in the region encoding proteins. We bring this commentary providing an overview of how this exemplar genome can allow us to understand these known unknown regions in brucellaphages.

Keywords Brucellaphage · Whole genome sequence · De novo assembly · Next generation sequencing · Genome mapping · Annotation

Introduction

Brucella is a gram-negative coccobacilli bacterium infecting mammalian hosts especially cattle and buffaloes. In the recent-past, human brucellosis has become a serious public health concern worldwide due to its zoonotic potential. *Brucella* can be infected by lytic phages known as brucellaphages (Farlow et al. 2014). With the recent genome sequencing strains Pr and Tb (Flores et al. 2012), there has been a significant interest in understanding the phage genomes. Brucellaphages are short-tailed particles containing dsDNA as genetic material (Flores et al. 2012) and are quite similar in physicochemical properties, spatio-temporal morphology and pathogenesis (Zhu et al. 2009). Previously, Jablonski designated a few variants of brucellaphages as wild type, tu (turbid), r (rapid lysis), rm (rapid lysis, minute) (Jablonski 1962). A comparative whole genome analysis study of six diagnostic brucellaphages revealed the presence of several candidate host range determinants (Farlow et al. 2014). Additionally they are known to yield genetically distinct phages and demonstrated recombination among brucellaphages (Tevdoradze et al. 2015). In this work, we have assembled the genome of a new lytic brucellaphage named as Gadvasu

Electronic supplementary material The online version of this article (doi:10.1007/s11693-015-9185-7) contains supplementary material, which is available to authorized users.

✉ Prashanth Suravajhala
prash@bioclues.org

1 Department of Veterinary Microbiology, College of Veterinary Science, GADVASU, Ludhiana, Punjab 141004, India

2 Food and Agricultural Products Center, Oklahoma State University, Stillwater, OK 74078-6055, USA

3 Bioclues.org, Kukatpally, Hyderabad 500 072, India

4 Bioinformatics Organization, 28 Pope St, Hudson, MA 01749, USA

5 Present Address: Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark

(abbreviated after Guru Angad Dev Veterinary and Animal Sciences University ~GADVASU, where the work was carried out). Sequence analysis of a lytic phage could throw some light on the diversity at the level of the lytic machinery, viz. the endolysins, holins etc. (Krupovic et al. 2011). This could aid in selection and effective therapeutic application of phages in infections with antibiotic resistant bacteria. A repertoire of genetically diverse lytic phages facilitate repeated use of phages for therapy in the same host without attracting neutralizing immune response against them. Generally, comparative genomic studies on phages help in understanding the phage-pathogen interactions and the strategies evolved by the bacteria to evade the phage mediated destruction in the body (Brüssow et al. 2004). Interpreting such differences observed among cohorts could shed some light on the evolution of brucellaphage. In addition, it would be an interesting example of host-pathogen co-evolution in understanding *Brucella* organisms from different geographical regions.

De novo assembly revealed that larger contigs were found to be associated with host-cell mediation

The brucellaphage was isolated from sewage (effluent waste water) from a dairy farm located in tropical, trans-gangetic central plain region (latitude: 30.9, 75.85) (Chachra et al. 2012). The extracted DNA (as per the method of Sambrook et al. 1987) was subjected to high-throughput Illumina genome sequencing by Beijing Genomics Institute (BGI). The raw data obtained was subjected to different steps (see Fig. 1 for details).The second part of the annotation was manual which included *de novo* assembly to join the short, single-end Illumina® reads into contigs which form longer blocks of the sequences (Warren et al. 2007; Zerbino 2010). The sequences were trimmed and SSAKE assembly files were obtained which included multi-line fasta file contigs. The contigs were merged using the Staden DNA sequence analysis package and further subjected to pregap algorithm to remove duplicates (Staden 2001). The larger contigs were compared against the reference genome and were further blasted using blast2seq to correct the potential bad joins in the contigs.

Approximately 1050 Mbp of nt sequence reads were assembled from the restriction fragments of the nucleotide sequences. The BpG assembly was mapped against the reference genome reads attributing to absence of functional elements in the phage. The larger contigs obtained using Velvet resulted 1269 contigs with a saturated mean redundancy. We observed that the reference genomes share an approximate 70 % nucleotide identity with the BpG assembly with 660 and 260 contigs mapping to the

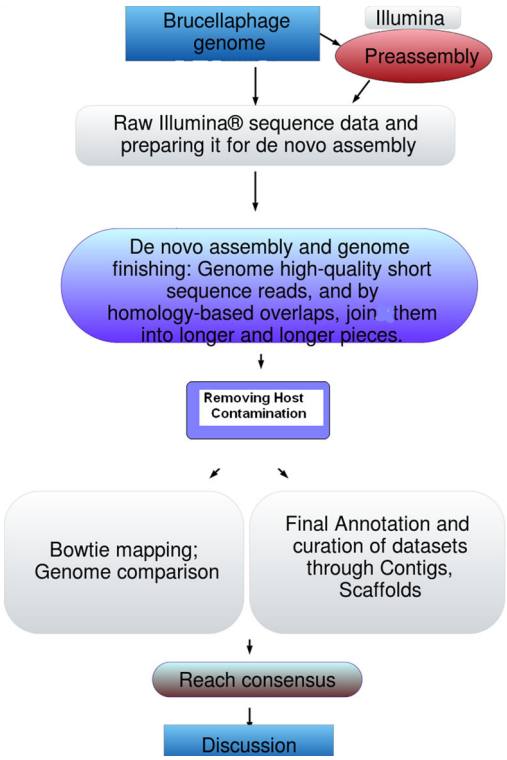

Fig. 1 Annotation strategy used to describe Brucellaphage Gadvasu. The draft sequences were subjected to pre-assembly processing. A raw FastQ data file of size 1050 Mbp containing thousands of sequence reads were obtained for the genome. The raw data was subjected to different steps. An open source script was run to convert the FastQ output into a standard format using Sanger-encoding base qualities (see *Maq software package*) to remove the duplicated identifier lines in each read, making the file size more manageable. Adapter contaminants, if any, were ligated onto fragments of sheared genomic DNA and then size-selected on a gel. If the size selection included fragments that are too small, the resulting library includes short inserts where the sequencing reads have both genomic DNA and adaptor sequence from the far side. The FastX clipper tool truncated the recognizable Illumina adapter sequences (see *FastX toolkit*). All the sequencing reads were further checked against the host genome using Bowtie's to eliminate the host genome interference (Langmead et al. 2009). Before indexing the host genome, the identifiers of sequence reads matching the host genome were extracted and mononucleotide reads were filtered (see *Fastq clipper*). As the *de novo* assembly can be improved by removing poor-quality base calls from the end of short-sequence reads produced by Illumina®, we used an adapted version of the quality-trimming script, which is packaged with the SSAKE3 *de novo* assembly software (Warren et al. 2007). Additional data analysis was performed utilizing the genome assembly and gene ontology (GO) tools. The GO categorization of genes identified were a part of functions responsible for infected macrophages and other GO biological processes (see supplementary tables)

nucleotide and protein sequences respectively that are associated with 605 protein families (see Fig. 2: Supplementary Table 1a and Table 1b). The average GC content of the BpG genome was found to be 31.53 %. Based on the sequence similarities of putative gene product to proteins in the databases, we noticed that the BpG has some unique

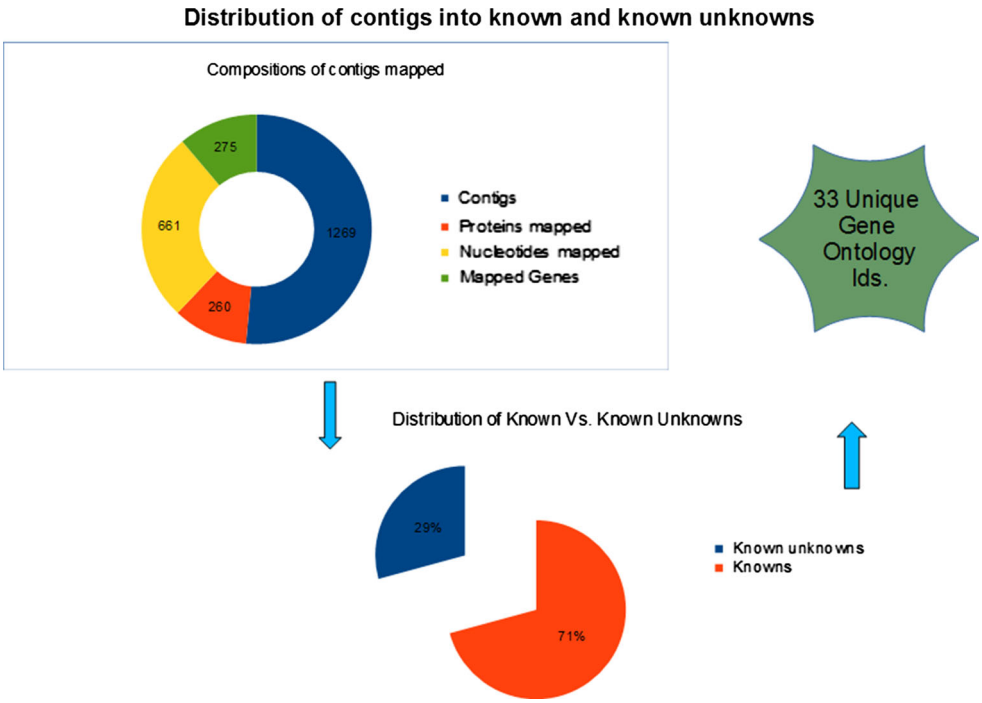

Fig. 2 Composition of contigs, nucleotide, protein and mapped genes along with *outline* of genome organization and characteristics of BpG. The figure also shows the number of unique GO identifiers mapped from the distributed genes

scaffolds (Supplementary Table 1a and Table 1b) that the reference genome did not have. Although the restriction fragments and single nucleotide polymorphisms (SNP) were predicted from the nucleotide sequences, no further annotation was done as we consider they could be in agreement with those sequences detected experimentally. Our analyses suggest that the contigs matching the genes map to 33 unique gene ontology (GO) ids (Supplementary Table 2a) and their genes are conserved across different taxa (Fig. 3; Supplementary Table 2b). While we argue that additional studies could possibly aid in understanding the host-phage relationships better, our analyses show that the BpG's changes whence compared to its peers' specific to the nucleotide/scaffold conservation could be linked to the overall distribution of base percentage, quality reads and data filtering.

Furthermore, from the conserved genes, we found that there are many Gag-Pol polyproteins which are known to regulate their own translation. It is believed that they are targeted via a multipartite membrane-binding signal and further play a role in nuclear localization of the phage genome during infection (Johnson et al. 2014; see Supplementary Table 2a). The key regulation of Gag-Pol Polyprotein motifs might aid the phage to phosphorylate the molecules at the time of virus maturation possibly helping the phage in host specificity. Our comparative bioinformatics' analysis from Pfam studies make known that there are several genes mapped that correspond to the 33 gene ontology functions. The YP_00 prefixed accessions correspond to the regions of protein families that linked to the known unknown regions (Supplementary Table 3). Among them are a few fiber proteins and ORFs linked to the domains of unknown function (DUF). These genes are associated with membrane-bound lytic murein transglycosylases which form an inherent part of phage activity. The analysis shows that a third of the genes seem to be either conserved ORFs or mapped to polymerases which are additionally needed for the phage to incorporate host mediation and basic metabolism. Correlation between these genes using systems biology approaches could allow us to understand functional overview of brucellaphages genomes. Proteins those are denoted as "structural or domains of unknown function" can further be identified using mass spectrometry and crosschecked whether or not "ORFans" show any similarity to known sequences in existing databases (Flores et al. 2012). Furthermore, it would be interesting to see if there is enrichment of ORFs present on phage hypothetical proteins (Supplementary Table 3). On a further note, when the set of ORFs were enriched with the online mendelian inheritance in man (OMIM), we found that these are associated with chemokine receptors which are well known to play a role specific to fundamental processes of homeostasis and metabolism of the cell (http://omim.org/entry/604833). Chemokines are a group of small molecules regulating and trafficking various types of leukocytes in the organism. These genes are associated with a poxvirus named Molluscum contagiosum virus (MCV) that initiates (Homey et al. 2002) and so we argue that these genes could serve as reliable candidates linked to chemokines and their targets. So far, very less

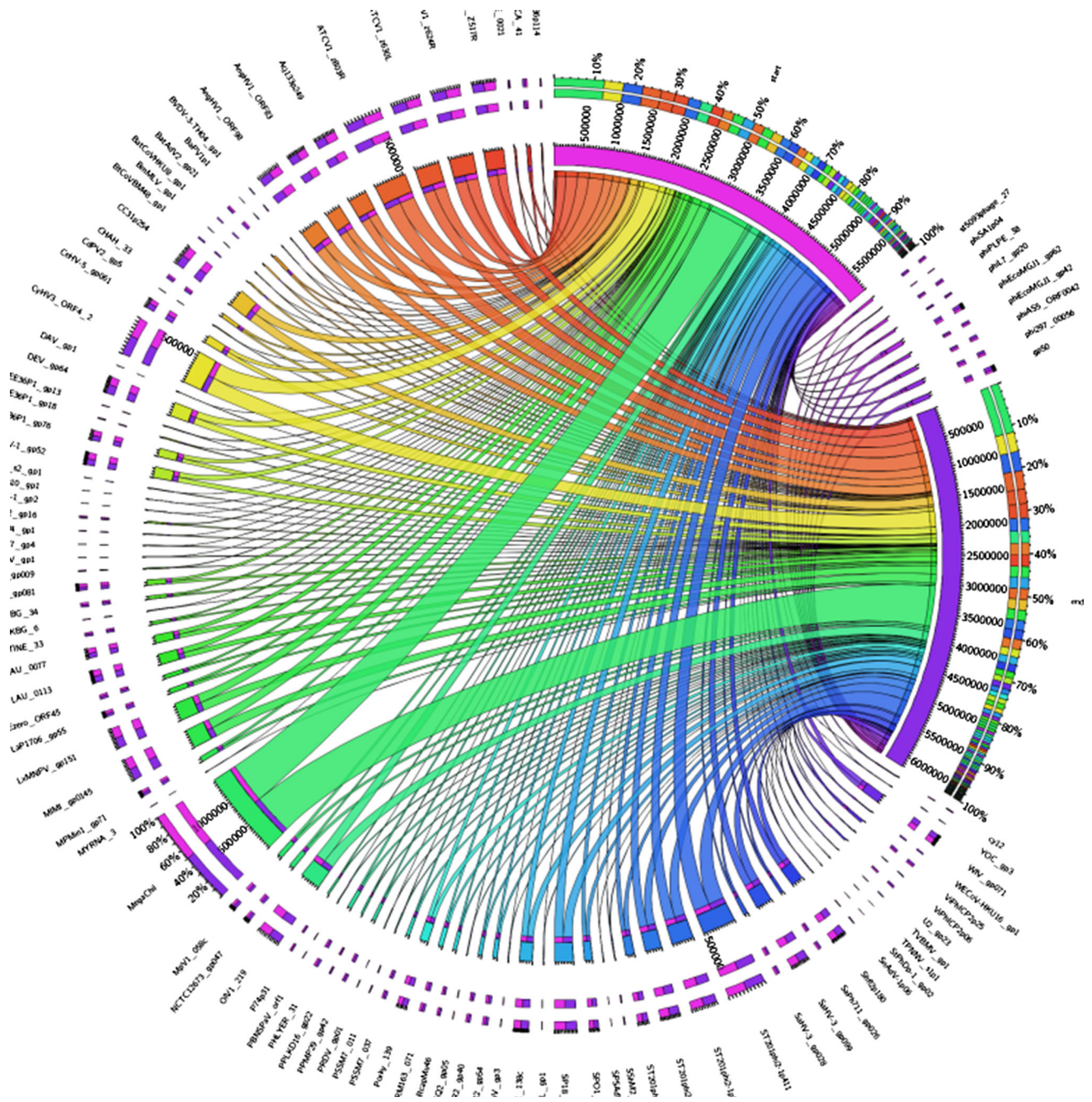

Fig. 3 A visual map of genes specific to their base positions enlisted in supplementary table 2b. The gene names with the *row ribbons* are placed first with the segment colors interpreted by the segment count (number of bases). (Color figure online)

knowledge is inferred from chemokine receptors in brucellaphage.

Virulence associated pathways

From the candidate ORFans, we asked if there are any congruent pathways linked to ontology terms specific to viruses. To check this, we used GenBank to map the proteins associated with GO Biosystems and found that the protein clusters are associated with 55 pathways, 6 structural complexes and 15 functional datasets. Furthermore, with the annotation, we sought to identify if any pathogenic or virulent pathways exist. Among these associations, we found putative DNA polymerases (marked in red in Gene Ontology Biosystems: see Fig. 4) which are catalytic subunits of non-segmented (herpes) dsDNA viruses (Davison et al. 2006). These are known to be involved with wide

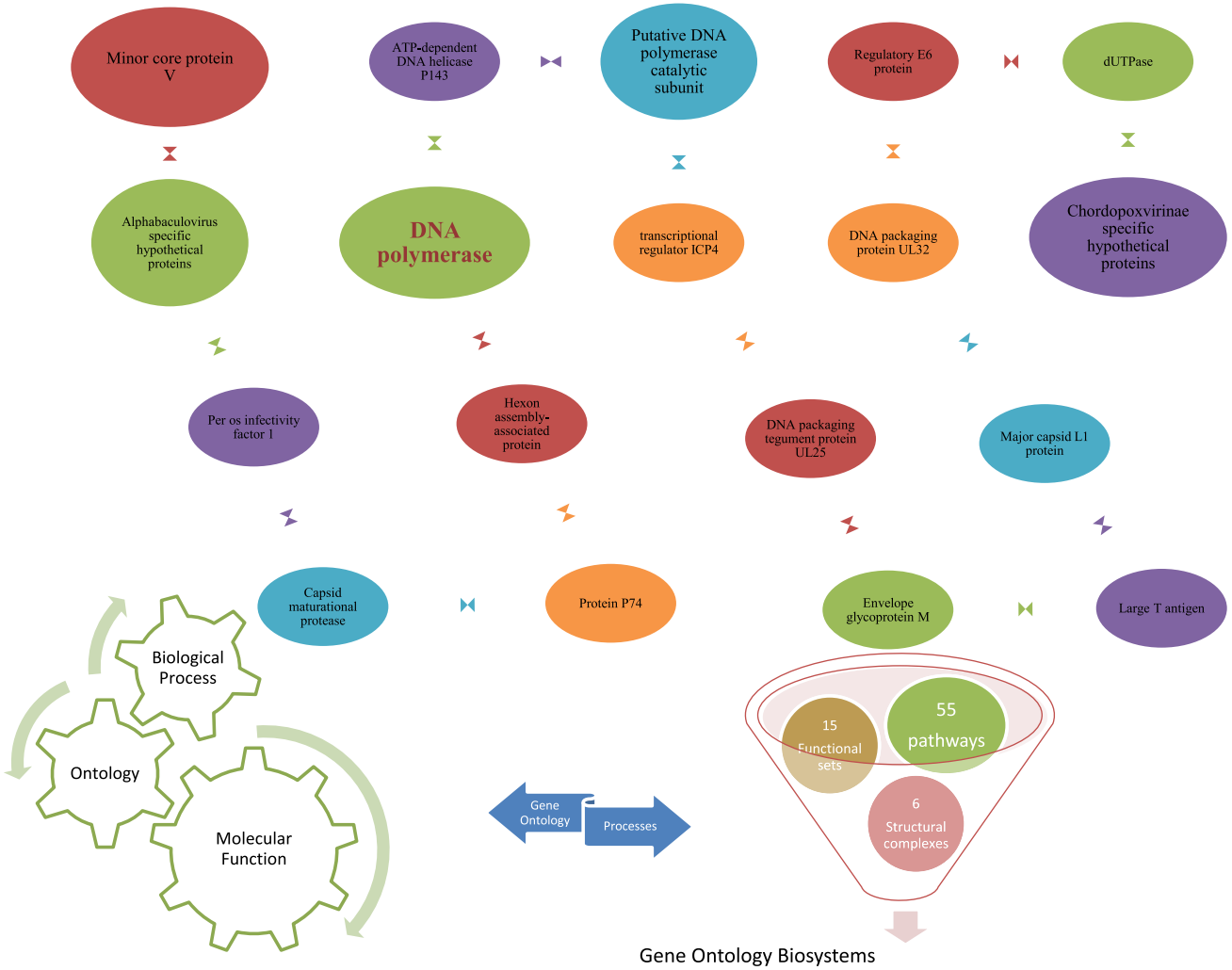

Fig. 4 Protein clusters associated with Gene Ontology Biosystems. Among the 18 protein clusters, we believe DNA polymerases (*highlighted in red*) play a significant role in interpreting the virulence factors. (Color figure online)

range of protein elements, viz: tegument, envelope, capsid containing capsomers and genomic DNA and thus, we believe that this protein cluster might be involved with virulent pathways. To improve the correctness of this pathway mapping, we checked from the review of literature that a large screen of such virulent pathways are associated with (Zn) metal binding and envelope/transporter associated pathways. Interestingly, from the ORFs, we also observed that a putative Zn-binding protein, viz. IcHV1_ORF78 (see supplementary table 2b) forms an integral part of this protein cluster. It would be interesting to see if any regulatory sequences for phages are helpful in understanding the virulence spectrum. It may also be noted that Tevdoradze et al. have also showed how essential the lytic and lysogenic cycles are propagated on certain alternate vaccine strains which forms a key basis for understanding these virulent pathways. Their mode of behavior and functional linkages in relation to the mutations of each virulence-associated gene(s) could be key choice to list the candidate proteins.

Conclusions

We have compared Brucellaphage Gadvasu to that of a recently sequenced brucellaphage Pr. There are genes with varying lengths that form an integral part of some unique genomic repertoire (about 30 %) that do not map to the reference. Functional analysis revealed that there are many known unknowns in the form of a conserved set of genes yielding putative proteins. Our exploration of the new genes that are associated with the lytic function of the phage and investigation of any duplication of genes leading to enhanced lytic capability of the phage has yielded the known unknowns. Whether or not the genes are distributed across the genome can be found out from the genes

associated with host-virility in them. The Gag-Pol polyproteins could be a set of candidate genes remarkable to study host-organism interactions, DNA repair and number of other scaffolds associated with evolutionary pathogenesis. The rare broad (genus-specific) lytic activity of this phage in contrast to the usual species-specific lytic activity of most phages possibly indicates that there is an evolutionary pressure on the phage in the face of re-emergence of Brucellosis. While we have attempted to understand the edifice of these genes along with the conserved regions between phages, these could be a source of studies for detecting evolutionary pattern. The research on genome organization of brucellaphages and comparative genomics of recently sequenced brucellaphages has just begun to yield insights into the conservation of nucleotide sequences, modes of interactions between host and pathogens and their relationships. However, scientific significance of such genome sequences depends on understanding and identifying functional elements of the genomes. In view of the less information available about brucellaphages, there is a need to annotate and curate the genome sequence with genes and ORFs, further providing a complete repertoire of metadata or wiki-enabled collection of data.

Availability of supporting data

Genomic sequences of the phage were deposited with accessions as follows: Bioproject PRJNA202394: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA202394.

Acknowledgments The authors are thankful to the Director of Research, GADVASU, Ludhiana for providing the funds under RKVY Scheme, Government of India. The authors thank Sivaramaiah Nallapeta for useful suggestions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Brüssow H, Canchaya C, Hardt W (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68(3):560–602

Chachra D, Kaur H, Chandra M, Saxena HM (2012) Isolation, electron microscopy and physicochemical characterization of a brucellaphage against brucella abortus vaccine strain S19. Internet J Microbiol 10(2):1–7

Davison AJ, Cunningham C, Sauerbier W, McKinnell RG (2006) Genome sequences of two frog herpesviruses. J Gen Virol 87(Pt 12):3509–3514

Farlow J, Filippov AA, Sergueev KV, Hang J et al (2014) Comparative whole genome analysis of six diagnostic brucellaphages. Gene 541(2):115–122

Fastq Clipper. http://genomics-pubs.princeton.edu/prv/…/fastq-filter_mono_nucleotide.pl

FastX toolkit. http://hannonlab.cshl.edu/fastx_toolkit/commandline.html

Flores V, López-Merino A, Mendoza-Hernandez G, Guarneros G (2012) Comparative genomic analysis of two brucellaphages of distant origins. Genomics 99(4):233–240

Homey B, Alenius H, Muller A, Soto H, Bowman EP, Yuan W, McEvoy L, Lauerma AI, Assmann T, Bunemann E, Lehto M, Wolff H, Yen D, Marxhausen H, To W, Sedgwick J, Ruzicka T, Lehmann P, Zlotnik A (2002) CCL27-CCR10 interactions regulate T cell-mediated skin inflammation. Nat Med 8:157–165

Jablonski L (1962) Variability of Brucella phages. Nature 193:703–704

Johnson SF, Collins JT, D'Souza VM, Telesnitsky A (2014) Determinants of Moloney murine leukemia virus Gag-Pol and genomic RNA proportions. J Virol 88(13):7267–7275

Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 75(4):610–635

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25

Maq software package. http://www.maq.sourceforge.net/fq_all2std.pl

Sambrook J, Fritsch EF, Maniatis T (1987) Molecular cloning: laboratory manual, vol 1. Cold Spring Harbour Laboratory Press.

Staden R, Judge DP, Bonfield JK (2001) Sequence assembly and finishing methods bioinformatics. In: Baxevanis AD, Ouellette BFF (eds) A practical guide to the analysis of genes and proteins, 2nd edn. Wiley, New York

Tevdoradze E, Farlow J, Kotorashvili A, Skhirtladze N, Antadze I, Gunia S, Balarjishvili N, Kvachadze L, Kutateladze M (2015) Whole genome sequence comparison of ten diagnostic brucellaphages propagated on two Brucella abortus hosts. Virol J 12:66

Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinform 11(Unit 11.5):1–13

Zhu CZ, Xiong HY, Han J, Cui BY et al (2009) Molecular characterization of tb, a new approach for an ancient brucellaphage. Int J Mol Sci 10(7):2999–3011