Entomopoxvirus-like long DNA sequences in human centromeric and peri-centromeric regions

Eiichi Hondo[1], Tetsuya Mizutani[2], Hiroshi Shimoda[3], Atsuo Iida[1]

[1] Laboratory of Animal Morphology, Department of Animal Sciences, Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya 464-8601, Japan

[2] Laboratory of Microbiology, Joint Faculty of Veterinary Medicine, Yamaguchi University, Yamaguchi 1677-1, Japan

[3] Center for Infectious Disease Epidemiology and Prevention Research, Tokyo University of Agriculture and Technology, Tokyo 183-8538, Japan

Correspondence: Eiichi Hondo, DVM, PhD
Email: ehondo@agr.nagoya-u.ac.jp

Abstract

Mammalian genomes harbor a range of endogenous virus-like sequences, although these sequences are typically affected by mutations and deletions, and no examples have been reported of long, conserved regions extending tens of kilobases or more. In this study, a comprehensive similarity search against the complete human T2T genome (hT2T) using high-quality viral genome data was performed. Among nearly continuous viral sequences, the longest identified was entomopoxvirus, spanning 140 kb with a similarity of over 57%. Notably, entomopoxvirus-like sequences stood out within the human genome, with a total of 2.42 Mb exhibiting similarity to human sequences. Seven entomopoxvirus species were fragmented into regions less than 37 kb each, with nearly all of these segments scattered throughout the human genome. Surprisingly, nearly all of these sequences were localized within the hsat1A regions of human centromeric and pericentromeric areas. On chromosome 3, for example, all sequences were confined within the centromere, and extensive transcription was observed across these regions. While the lack of fully sequenced genomes for other eukaryotic species limits our ability to infer the phylogenetic significance of these regions, our findings reveal remarkably long virus-like sequences within centromeric and related regions—as integral components of eukaryotic chromosomes. This discovery opens new avenues in the study of co-evolution between viruses and eukaryotes, offering intriguing insights into the genomic landscape at the core of eukaryotic biology.

Introduction

Our genomes contain various virus-like sequences known as endogenous viral elements (EVEs) [1]. These sequences often exhibit mutations and deletions compared to intact viral genomes, making it rare to find long, complete viral sequences. However, it has long been recognized that EVEs play a critical role in regulating gene expression, functioning as enhancers, and thus are significant for biological functions [2, 3]. Open chromatin regions, in particular, are considered more susceptible to the insertion of foreign viral genomes, which might enable EVEs to act as enhancers near host genes. Indeed, chimeric sequences between host genomes and HIV genomes have been observed in B cells of HIV-infected individuals, suggesting that these regions might serve as enhancer elements [4]. For these sequences to be passed to offspring, however, they must integrate through germline cells.

In mammals, EVEs have been co-opted as functional genes crucial for specific adaptations, such as PEG10 and syncytin, essential for placental development [5, 6], and SASPase, which is implicated in skin hydration and classified as a retroviral-derived EVE (R-EVE) [7]. A homolog closely related to the VP35 protein of Ebola and Marburg viruses has also been identified in *Myotis* bats, potentially linking these EVEs to the species' ability to tolerate infections by these viruses without disease symptoms [8]. R-EVEs, also known as endogenous retroviruses, constitute about 9% of our genome and are recognized as important viral-like sequences from an evolutionary perspective. For instance, endogenized viruses have been shown to influence coat color phenotypes in mice [9, 10], and R-EVEs in pigs have raised public health concerns due to their ability to release viral particles and to infect human cells [11]. Also, recent reports reveal that in certain bat species, approximately 30% of endogenous retroviruses are found in nearly intact forms, suggesting that these genome regions were acquired relatively recently [12].

In contrast, examples of extensive EVEs spanning several kilobases are known, such as the human herpesvirus 6, which exhibits highly homologous repeat sequences to those of human telomere. This suggests that certain genomic sequences in mammalian genome structure has to be discussed in the context of EVEs [13].

In this study, complete viral genome data files registered in NCBI/RefSeq and GenBank (specifically, the virushostdb data file) were utilized to examine their

homology to human T2T full genome (hT2T) [14]. We report newly identified regions exhibiting long viral sequence homology, especially in centromeric and pericentromeric regions, which were not detected in previous human genome builds prior to hT2T.

Results and Discussion

Using the latest hT2T human genome build, we performed homology analysis with the viralhostdb's complete viral data file (as of August 2024) via the fasta36 software (e-value threshold set at 1e-50). Among sequences with over 57% homology and lengths exceeding 5 kb, the total unique viral sequences amounted to 2.71 Mb, representing 0.087% of the hT2T genome. Notably, sequences homologous to viruses in the Entomopoxvirinae subfamily were identified in centromeres on chromosome 2, 3, 6, 10, 13, 20, 21, 22 and X (Fig. 1). Of all viruses, entomopoxvirus exhibited the greatest representation, with sequences totaling 2.42 Mb, accounting for 89.0% of all viral homologous regions. These homologous regions were primarily localized near centromeres on the short arms of chromosomes 13, 21, and 22, mixed with sequences homologous to other viruses in low gene density regions shown on heat maps in Fig. 1, with the exception of chromosome 3. Homologous regions on the terminal long arm of chromosome 20 and near the centromere on the short arm of the X chromosome corresponded to 5 kb and 8 kb regions homologous to Grapevine fleck virus and Clostridium phage, respectively, spanning intronic regions of the RTEL1 and CCNB3 genes. Other homologous regions on chromosomes 2, 6, 10, 22, and X were primarily composed of stringent simple repeat regions. Detailed locations of these homologous regions within the human hT2T genome are presented in Supplemental Table 1.

The EVE sequence in the centromere of chromosome 3 spans totally 748 kb, representing 16.25% of the centromeric region (Fig. 2a). EVEs on chromosome 13, chromosome 21, and chromosome 22 span totally 1.43 Mb (Fig. 2b), 291 kb (Fig. 2c), and 141 kb (Fig. 2d), respectively. These sequences predominantly consist of the satellite DNA hsat1A. No centromeric satellite sequences were found outside of these regions, and hsat1A regions accounted for 96.6% of the total 2.71 Mb viral-homologous regions found in this study. These regions are unique to the hT2T build, with no liftover to GRCh37 or GRCh38 possible and sparse annotation. Analysis of Simple Repeats (table: hub_170992_simpleRepeat) downloaded from the UCSC Genome Browser

indicated lower repeat density in other regions compared to chromosome 3's centromere. An examination of SGDP copy number revealed significant inter-individual variation: 11–17 on chromosome 3, 29–233 on chromosome 13, 27–245 on chromosome 21, and 31–271 on chromosome 22.

While entomopoxvirus-like sequences were identified in the complete human genome, their relatively low homology (a minimum 57%) and the functional significance to core eukaryotic cell division suggest the possibility that ancestral entomopoxviruses may have infected more primitive eukaryotes, leading to eventual endogenization. To investigate, we retrieved homologous sequences from the 249-39967 bp region of NC_023426 alphaentomopoxvirus Anomala cuprea, which shares 58% homology with a continuous 37,759 bp region in the hT2T chromosome 3 centromere (Fig. 2a; Alphaentomopoxvirus_U box on human Hsat1A_3_2), across 341 eukaryotic species from yeast to mammals (Ensembl release-113), aligning them using fasta36 (e-value 1e-50). Chromosome 3 was selected due to its minimal variation in SGDP copy number among human individuals. For comparison, homologous sequences to the 249-39967 bp region in hT2T chromosome 3 were also retrieved across 341 species and aligned with MAFFT software. However, alignments revealed no significant conservation across species for either viral or human homologous regions (data not shown).

Seven entomopoxvirus species were fragmented into regions less than 37 kb each, with nearly all of these segments scattered throughout the human genome. (Supplemental Table 3; [15]). Given that entomopoxviruses are linear double-stranded DNA viruses that replicate within the host's cytoplasm, the mechanism by which these sequences were integrated into the host genome requires discussion. Traditional molecular mechanisms for integration are inadequate to explain this. A plausible hypothesis is that entomopoxvirus sequences might have been integrated during co-infection with retroviruses or by utilizing enzymes by endogenous retroviruses, facilitating chromosomal integration during cell division. Given the low homology (57–70%) and the unlikely scenario of entomopoxvirus infection to primates, such a scenario may have occurred in ancient times, though definitive evidence remains elusive.

The discovery of moderately homologous yet extensive (hundreds of kb) continuous viral-homologous regions in the human genome, other than human herpesvirus-related telomeric associations, is unprecedented. These sequences,

undetected in GRCh37 or GRCh38, were identified only with hT2T, and remarkably, 96.6% of the viral-homologous 2.71 Mb was comprised of hsat1A. Located in α-satellite non-coding regions of the genome forming centromeres and pericentromeres [16], α-satellite sequences, while often assumed to be identical across centromeres, actually show substantial complexity and chromosomal specificity in variations and polymorphisms. Higher-order repeat units (HOR) and Hsat constituting α-satellites feature consecutive repeat sequences but exhibit diversity in monomer types, mutations, and arrays. For example, HOR homology across non-homologous human chromosomes ranges from 50–70% [17] and varies significantly within the same species.

Over the past 20 years, extensive efforts have been devoted to genome sequencing and assembly, yet an estimated 5-10% of the human genome has remained as unexplored "black box" regions [18]. Most of these regions are found within or around the centromeres of chromosomes, containing highly repetitive DNA sequences that complicate assembly from short DNA sequencing reads. However, the recent completion of a telomere-to-telomere human genome assembly (hT2T-CHM13) has allowed a comprehensive characterization of the centromeric and pericentromeric repeats, which make up 6.2% (189.9 Mb) of the genome [14]. Despite this achievement, it is still not possible to conduct molecular phylogenetic analyses on α-satellite regions in non-human species. High-order repeats (HORs), which form the core of the centromere and are essential for kinetochore function and cell division in eukaryotes [16], are believed to contain sequences widely conserved across eukaryotes. Nonetheless, they remain unanalyzable with current methods. Similarly, Hsat, another class of repeats, plays a crucial role in chromosomal stability, heterochromatin formation, and epigenetic control essential for life [17], but no cross-species analyses are feasible at this time. In fact, a search for homologous sequences across 341 species in this study yielded no regions with significant homology across species, highlighting the need for complete genome assemblies in additional species.

Since the release of the hT2T genome build, functional studies on hsat1A in humans have begun to progress, with transcripts from hsat1A regions identified, and expression differences between normal and tumor cells reported [19]. In this study, we used long-read RNA-seq data from humans and mapped it to the hT2T genome using Minimap2 to examine transcriptional products in regions identified with entomopox-like sequences on chromosome 3, 13, 21, and 22. Transcripts were analyzed for read species (bottom panel in “Transcripts” on Fig. 2) and read count (top panel) in chromosomes 3, 13, and 21, with limited read counts observed in the individual

analyzed (European ancestory). A relatively large number of reads were detected primarily in the Hsat3 region on chromosome 3, on the other hand, the limited transcripts identified on viral-like sequences ("Regions" in Fig. 2) include long non-coding RNAs spanning several hundred kilobases. Even at low expression levels, a few copies of these transcripts might be sufficient to regulate centromere function on these chromosomes, suggesting they could be functionally relevant molecules.

While the exact functions of these transcripts remain unknown, impaired centromere function could lead to diseases involving chromosomal missegregation [20] (Barra and Fachinetti, Nat Commun 2018). Furthermore, this study found that SGDP copy numbers for all centromere-pericentromere regions containing entomopox-like sequences varied between individuals, implying inter-individual variability in hsat1A regions [21]. This variability suggests that hsat1A may be an essential region, influencing not only disease susceptibility but potentially even evolutionary processes.

Methods

All viral genome data were downloaded from https://www.genome.jp/virushostdb/ on August 6, 2024, and the full-length human genome hT2T-CHM13v2.0 was obtained from NCBI. A homology search was performed using the entire viral dataset as the query and hT2T as the reference, using fasta36 with parameters -b 1 -d 1 -E 1e-50 -m BB. Preliminary experiments established a homology threshold of 57% identity and 5 kb continuity to ensure complete capture of entomopoxvirus data, which guided human sequence collection (fasta36_data_collection.py). All python codes described here are stored in GitHub repository (https://github.com/ehondo2020/Takoyaki). Using the output files, human chromosomal locations of viral sequences were extracted (human_genomic_position_extract.py), merged to remove duplicates, and the largest range was determined (chromosomal_range_determination.py). The resulting human genomic position data was converted into BED files, and suppl_table2.txt was created with make_single_range_1.py and _2.py, while make_single_range_3.py was used to finalize the list of human genome positions corresponding to each virus (suppl_table1.txt) for visualization with RIdeogram

(https://github.com/TickingClock1992/RIdeogram) and Gviz (https://github.com/ivanek/Gviz).

CenSat and simple_repeats data from the UCSC Genome Browser, along with SGDP copy number for relevant regions, were downloaded for Gviz visualization. Simple_repeats counts were collected using the options, a bin size of 5,000 and window size of 500. RNA-seq data were obtained from the Whole Genome Sequencing Project and mapped to hT2T-CHM13v2.0 using Minimap2. Chromosome 3 centromeric regions in the human hT2T genome (37,759 bp) were compared for homology with other eukaryotic species. Genome data for 341 species were downloaded from Ensembl's FTP site (release-113), and a batch homology search was performed using fasta36 with parameters -b 1 -d 1 -E 1e-50 -m BB.

References

[1] Deep mining reveals the diversity of endogenous viral elements in vertebrate genomes. Jose Gabriel Nino Barreat & Aris Katzourakis. Nat Microbiol 2024 Nov;9(11):3013-3024. doi: 10.1038/s41564-024-01825-4.

[2] Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Edward B Chuong, Nels C Elde, Cédric Feschotte. Science 2016 Mar 4;351(6277):1083-7. doi: 10.1126/science.aad5497.

[3] Regulation of endogenous retroviruses in murine embryonic stem cells and early embryos. Xinyi Lu. J Mol Cell Biol 2024 Jan 17;15(8):mjad052. doi: 10.1093/jmcb/mjad052.

[4] Early antiretroviral therapy in neonates with HIV-1 infection restricts viral reservoir size and induces a distinct innate immune profile. Pilar Garcia-Broncano et al. Sci Transl Med 2019 Nov 27;11(520):eaax7350. doi: 10.1126/scitranslmed.aax7350.

[5] Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Ryuichi Ono et al. Nat Genet 2006 Jan;38(1):101-6. doi: 10.1038/ng1699.

[6] Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. S Mi, X Lee, X Li, G M Veldman, H Finnerty, L Racie, E LaVallie, X Y Tang, P Edouard, S Howes, J C Keith Jr, J M McCoy. Nature 2000 Feb 17;403(6771):785-9. doi: 10.1038/35001608.

[7] SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. Takeshi Matsui et al. EMBO Mol Med 2011 Jun;3(6):320-33. doi: 10.1002/emmm.201100140.

[8] Conservation of Structure and Immune Antagonist Functions of Filoviral VP35 Homologs Present in Microbat Genomes. Megan R Edwards et al. Cell Rep 2018 Jul 24;24(4):861-872.e6. doi: 10.1016/j.celrep.2018.06.045.

[9] Dilute (d) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. N A Jenkins, N G Copeland, B A Taylor, B K Lee. Nature 1981 Oct 1;293(5831):370-4. doi: 10.1038/293370a0.

[10] A role for the rare endogenous retrovirus β4 in development of Japanese fancy mice. Akira Tanave & Tsuyoshi Koide. Commun Biol 2020 Feb 4;3(1):53. doi: 10.1038/s42003-020-0781-z.

[11] Monitoring for PERV Following Xenotransplantation. Joachim Denner. Transpl Int 2024 Oct 7:37:13491. doi: 10.3389/ti.2024.13491.

[12] Signature of viral fossils: a comparative genomics approach to understand the diversity of endogenous retroviruses in bats. Vinita Lamba et al. Virus Res 2024 Oct 25:350:199484. doi: 10.1016/j.virusres.2024.199484.

[13] Endogenization and excision of human herpesvirus 6 in human genomes. Xiaoxi Liu et al. PLoS Genet 2020 Aug 10;16(8):e1008915. doi: 10.1371/journal.pgen.1008915.

[14] The complete sequence of a human genome. Sergey Nurk et al. Science 2022 376, 44–53.

[15] The Genome of Melanoplus sanguinipes Entomopoxvirus. C L Afonso 1, E R Tulman, Z Lu, E Oma, G F Kutish, D L Rock. J Virol 1999 Jan;73(1):533-52. doi: 10.1128/JVI.73.1.533-552.1999.

[16] Complete genomic and epigenetic maps of human centromeres. Nicolas Altemose et al. Science 2022 376, eabl4178.

[17] Alpha satellite DNA biology: finding function in the recesses of the genome. Shannon M McNulty & Beth A Sullivan. Chromosome Res 2018 Sep;26(3):115-138. doi: 10.1007/s10577-018-9582-3.

[18] An assessment of the sequence gaps: unfinished business in a finished human genome. Evan E Eichler, Royden A Clark, Xinwei She. Nat Rev Genet 2004 May;5(5):345-54. doi: 10.1038/nrg1322.

[19] Human Satellite 1A analysis provides evidence of pericentromeric transcription. Mariana Lopes et al. BMC Biol 2023 Feb 8;21(1):28. doi: 10.1186/s12915-023-01521-5.

[20] The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. V Barra & D Fachinetti. Nat Commun 2018 Oct 18;9(1):4340. doi: 10.1038/s41467-018-06545-y.

[21] The Structure of Simple Satellite Variation in the Human Genome and Its Correlation With Centromere Ancestry. Iskander Said, Daniel A Barbash, Andrew G Clark. Genome Biol Evol 2024 Aug 5;16(8):evae153. doi: 10.1093/gbe/evae153.

Acknowledgements

This work was partially supported by Grants-in-Aid for Scientific Research from the JSPS (24K21913 to E.H.). Computations were performed on the NIG supercomputer at ROIS National Institute of Genetics, Japan.

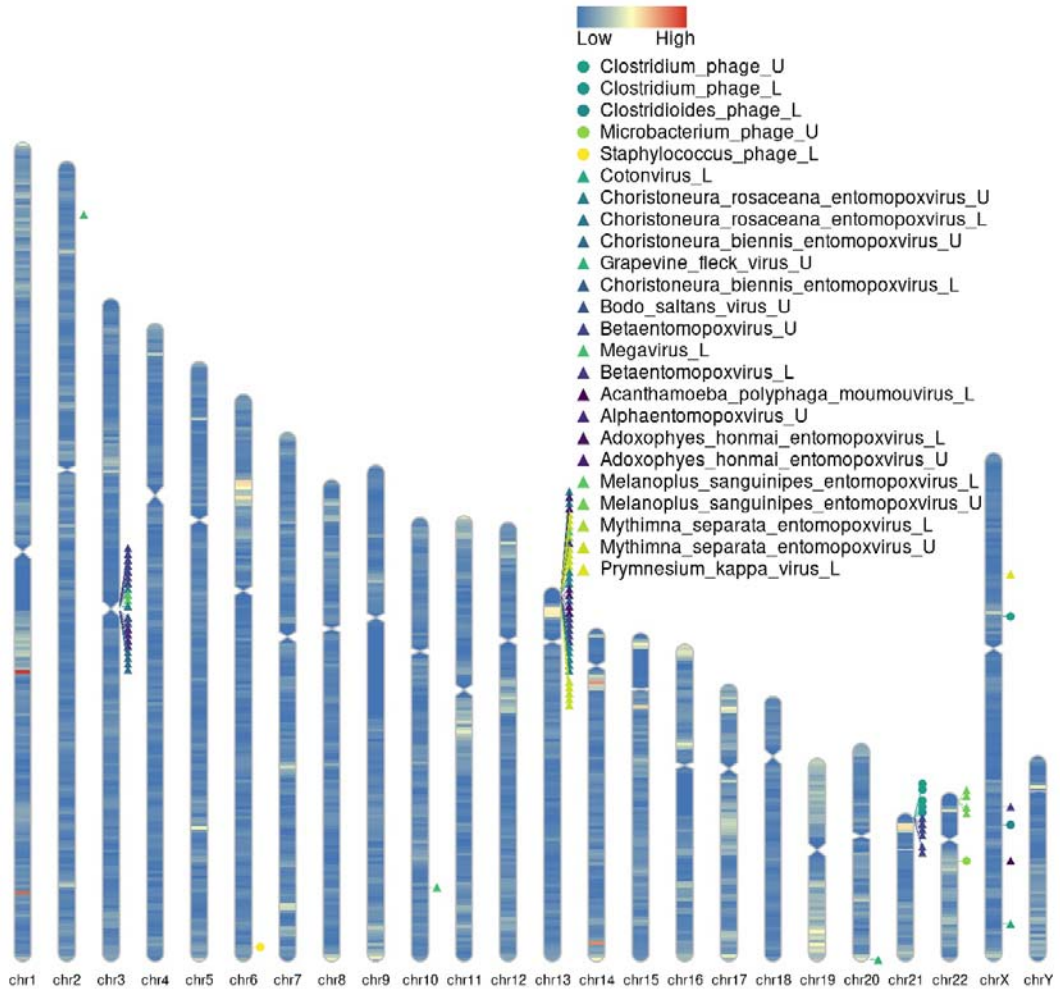

Fig.1 Locations of virus-like sequences and homologous sequences on the human T2T genome. All chromosomes were built using hT2T files (sequence data and .gff annotation file). Each chromosome is shaded to reflect gene density. Clusters of homologous sequences are observed within the centromere of chromosome 3, and on chromosomes 13, 21, and 22. Most scattered sequences across the genome are tandem simple repeat arrays; however, on chromosome 20, a region with moderate gene density near the end of the long arm contains a sequence homologous to Grapevine fleck virus, forming part of the intronic region of the RTEL1 gene. Additionally, on the X chromosome, a sequence in the proximal short arm near the centromere shows high homology to a Clostridium phage and is located within an intron of the CCNB gene. The suffixes "_L" and "_U" at the end of figure labels indicate sequences with homology to multiple viruses (L) or to a single virus (U). For sequences labeled "_L,"

the longest homologous sequence name is provided, and a list of additional viruses is available in Supplemental Table 2.

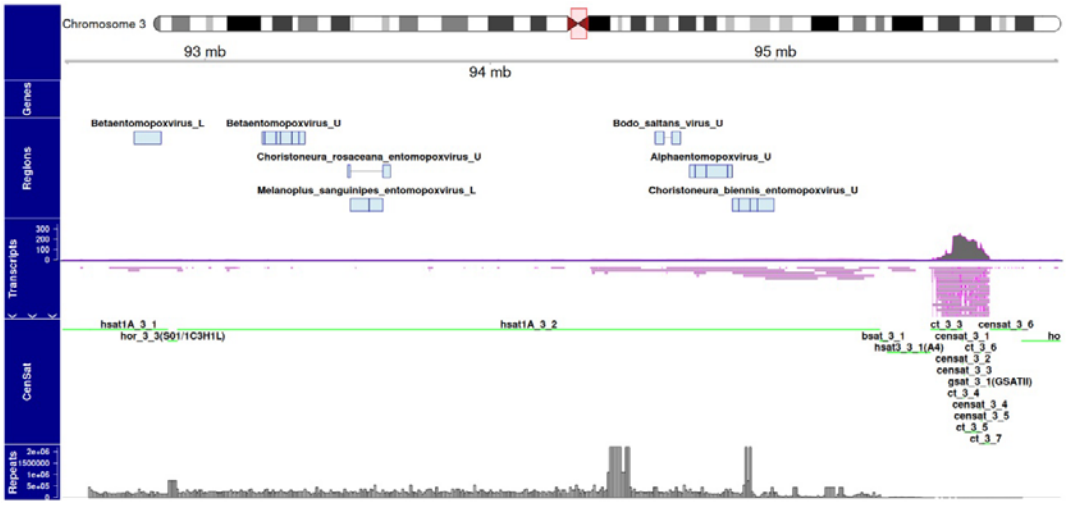

Fig2a Centromeric region with high similarity to viruses within chromosome 3. This region is unique to hT2T, cannot currently be lifted over from GRCh37 or 38, and shows no distribution of genes. On the left, the “Regions” panel displays bars indicating viral regions with high similarity. The mapping results of long-read RNA-seq data from European human samples used in this study, aligned with minimap2, are shown on the left under “Transcripts.” Although transcription from viral-related regions is observed (bottom of “Transcripts”), transcription levels are low (top of “Transcripts”). Long non-coding RNAs of over 10 kb have been identified in the virus-like sequences. In contrast, numerous transcripts have been observed primarily from the hsat3 region on the right side of the figure. “Repeats” shows the distribution of simple repeats; this region contains numerous intermittent short repeats, displayed here with a bin size of 5,000 and window size of 500.

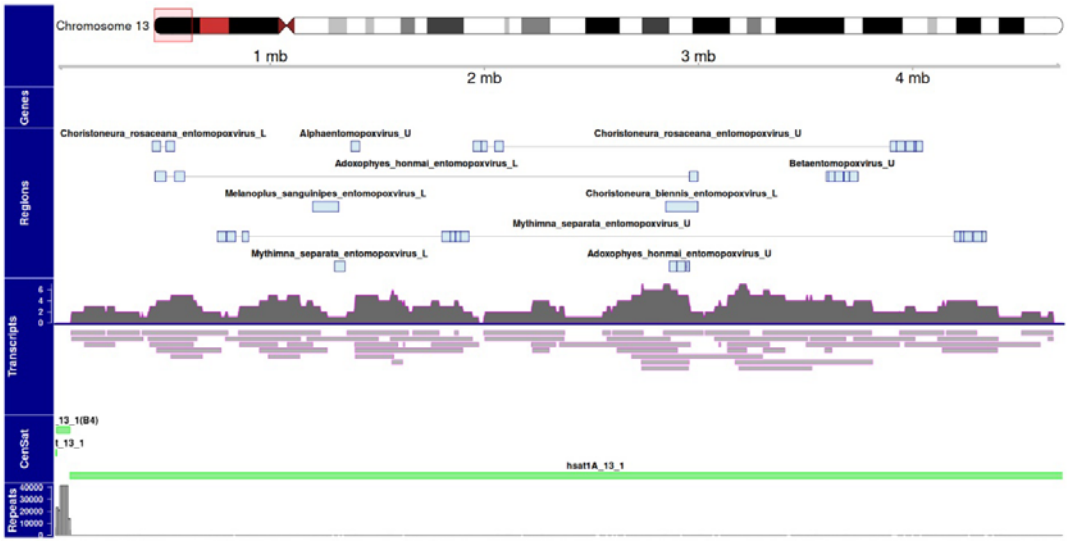

Fig2b Virus-like sequences distributed in the short arm of chromosome 13. Non-coding RNAs are transcribed across this entire region, with transcripts observed from all parts of the virus-like sequences. Simple repeats are scarcely detected. Shown with a bin size of 5,000 and window size of 500.

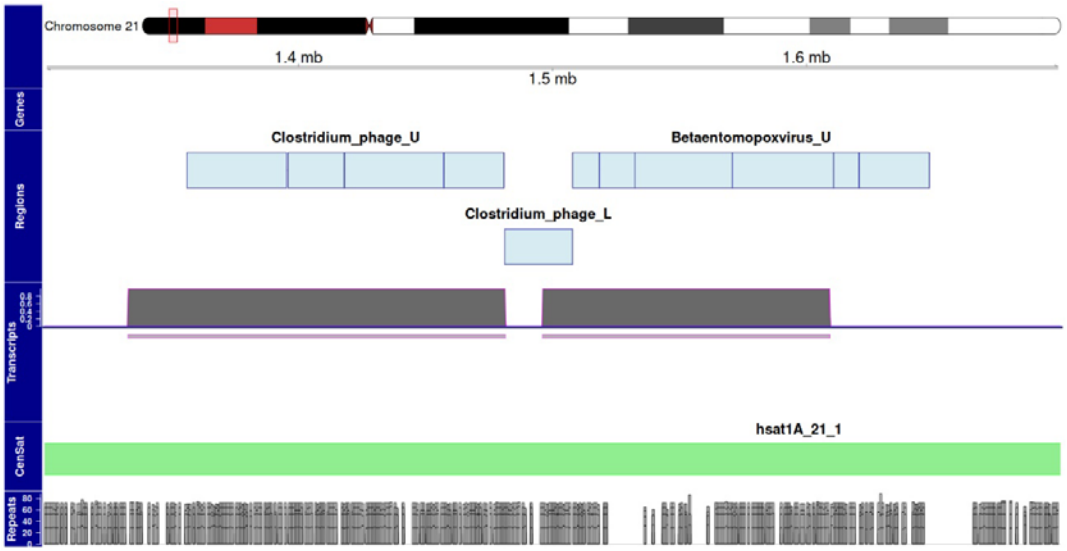

Fig2c Virus-like sequences distributed in the short arm of chromosome 21. This region includes one unique sequence each for entomopoxvirus and Clostridium phage (unlike "_L," these sequences do not overlap with other viral sequences). The small

Clostridium phage box contains sequences overlapping with other viruses. Intriguingly, the mapping of transcripts reveals non-coding RNA sequences over 100 kb in length, which closely match the positions of the two virus-unique sequences. Simple repeats are also sparse in this region. Displayed with a bin size of 5,000 and window size of 500.

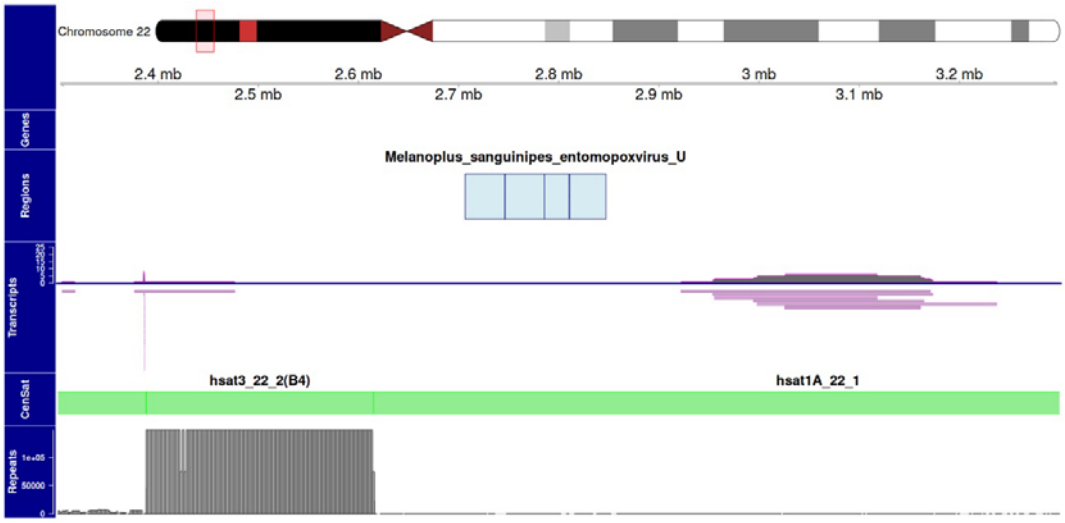

Fig2d Virus-like sequences distributed in the short arm of chromosome 21. Only a unique sequence for entomopoxvirus is identified in this region; however, no transcripts are detected, and no particularly distinctive features are observed.