Chapter 18

Bioinformatics Assembling and Assessment of Novel Coxsackievirus B1 Genome

Jake Lin, Bryn Y. Kimura, Sami Oikarinen, and Matti Nykter

Abstract

The human microbiome project via application of metagenomic next-generation sequencing techniques has found surprising large and diverse amounts of microbial sequences across different body sites. There is a wave of investigators studying autoimmune related diseases designing from birth case and control studies to elucidate microbial associations and potential direct triggers. Sequencing analysis, considered big data as it typically includes millions of reads, is challenging but particularly demanding and complex is virome profiling due to its lack of pan-viral genomic signature. Impressively thousands of virus complete genomes have been deposited and these high-quality references are core components of virus profiling pipelines and databases. Still it is commonly known that most viral sequences do not map to known viruses. Moreover human viruses, particularly RNA groups, are notoriously heterogeneous due to high mutation rates. Here, we present the related assembling challenges and a series of bioinformatics steps that were applied in the construction of the complete consensus genome of a novel clinical isolate of Coxsackievirus B1. We further demonstrate our effort in calling mutations between prototype Coxsackievirus B1 sequence from GenBank and serial clinical isolate genome grown in cell culture.

Key words Genomics, Assembly, Bioinformatics, Enterovirus, Coxsackievirus, T1D

1 Introduction

Coxsackievirus B (CVB), with six serotypes and a member of the human enterovirus genus, have long interested researchers due to its pathogenic role in gastrointestinal inflammation and implications with cardiomyopathy. Coxsackievirus B have also been implicated in Type 1 diabetes (T1D), which is an autoimmune disorder onset typically in young children. The trigger is unknown, but clinical T1D occurs when about 90% pancreatic islet beta cell population is destroyed and thereafter regular external insulin is required to regulate blood glucose levels to avoid hyperglycaemia leading to organ damage and death. The detection of islet autoantibodies in the blood signals autoimmunity and multiple autoantibody confirmations accurately predict T1D. The window of time between autoimmunity and T1D onset can vary from months to multiple years [1]. Two of the

Andrés Moya and Vicente Pérez Brocal (eds.), *The Human Virome: Methods and Protocols*, Methods in Molecular Biology, vol. 1838, https://doi.org/10.1007/978-1-4939-8682-8_18,

autoantibodies namely insulin (IAA) and glutamic acid decarboxylase (GADA) may predict different progression models from the triggering the disease to onset of T1D. It has been reported that IAA, peaking in the first 3 years of child's life, is associated with early-age T1D onset and GADA confirmation incidence plateaus into adolescence, suggesting divergent trends and possibly multiple environmental triggers and immune interactions. Enterovirus sequences have been found in pancreas tissue [2] donated from recent T1D patients. CVB with six serotypes is the subgroup of enterovirus most cited in PubMed associated with T1D. While recent advances in metagenomic shotgun sequencing together with lowered cost of sequencing have allowed investigators to sequence virus directly from the environment, for example stool or blood samples, virus profiling and particularly genotype identification are challenging due to lack of universal viral marker, insufficient amount of preference genomes, and also large genetic distance from preference virus sequences. The genetic distance is exasperated by their innate high pace of mutation, particularly in RNA viruses including CVB. This report presents a set of bioinformatics procedures, known and novel, toward building a novel complete CVB1 consensus genome obtained from clinical virus isolate grown in pancreatic cell culture model using next-generation sequencing and multiple time points.

2 Materials

A novel wild Coxsackievirus B1 clinical isolate strain was continuously cultured in human pancreas ductal (Panc-1, abbreviated as A) and pancreas islet (1.1B4, abbreviated as B) cells to establish a persistent infection model for CVB1. We note that consensus genome assembly and variant calling are applied for Panc-1 (A) samples where quality control and contig generation were done for both cell types. The infected cells were grown in +37 °C incubator with 5% CO_2 and treated three times a week by washing and adding fresh cell culture media [3]. The cells were monitored by microscopy and culture supernatant medium was harvested once a week and stored at −80 °C for sequencing analysis.

Three time points (a prep after 1 day of culture (time point 0), and then approximately 6 (time point 1) and 12 months (time point 2) after initiation of the virus culture) of harvested and frozen cell culture mediums containing virus were selected for NGS analysis. The virus was enriched in the cell culture and therefore, the virus RNA was extracted directly for cell culture supernatant for next-generation sequencing analysis. The sample preparation was done according to the protocol presented also in this book [Kramna and Cinek, "Virome sequencing of stool samples"]:

Selected known references.

CVB1 Prototype (GenBank accession no NC_001472.1).

CVB1 wild type (GenBank accession no AY186745.1).

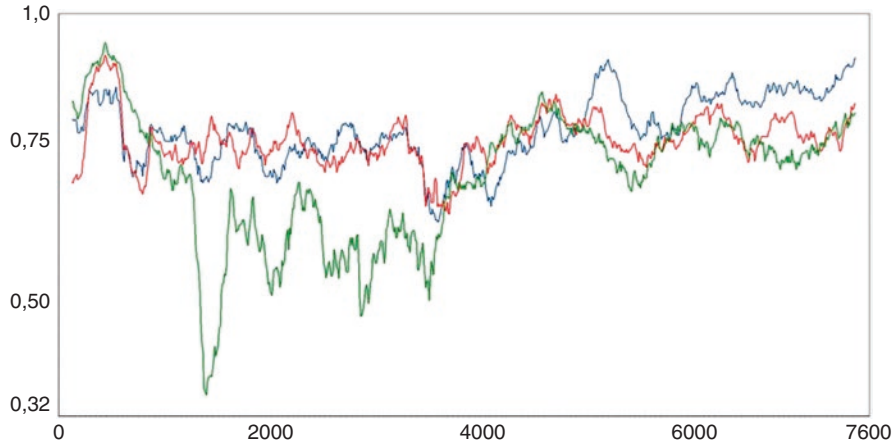

Fig. 1 Reference similarity plots. Set against CVB1 prototype, CVB1 clinical isolate in red and CVB1 wild type in blue are compared. CVB5 in green confirms the large genetic diversity in structural capsid regions (~800–3000) between serotypes

Table 1
Bioinformatics dependencies (tested in Linux environment) and applied tools

Name	Purpose
Bedtools [4]	Tool for genome arithmetic
BWA [5]	Short read alignment
FastQC [6]	Read quality report util
IGV [7]	Aligned track visualization
KmerGenie [8]	Sequence utility
Samtools [9]	Sequence utility
Taxonomer [10]	Virome profiler, no assembly, direct mapping
Velvet [11]	Short read assembler
Vipie [12]	Multi-sample virome profiler; de novo assembly

CVB1 Clinical isolate (Hyöty lab, University of Tampere, unpublished).

CVB5 Prototype (GenBank accession no MF973166.1).

Figure 1 demonstrates the substantial sequence differences between CVB1 prototype and CVB1 wild type, exposing the need for an updated and specific clinical isolate. Variant analysis between virus sequences analyzed from different time points in cell culture also mandates relevant consensus reference sequences (Table 1).

Custom code requires Python (Version 3.4+) and pysam [4, 13].

Detailed directions and README file are accessible at: https://sourceforge.net/projects/contig-mutation-caller

3 Methods

3.1 Virome Profiling and Assessment

Preliminary quality control (QC) and virus profiling were done in Vipie [12], a web-based virome NGS pipeline. Resultant virome profiles from CVB1 prototype samples (Fig. 2a) show that a clear majority, over 99% of viral reads mapped to its homologous sequence in the GenBank (accession NC_001472.1), while for clinical virus isolate (Fig. 2b) grown in pancreatic ductal and islet cells, the mapped profile accessions are highly diverse. The profile includes more than 20 different matched EV accessions with proportions

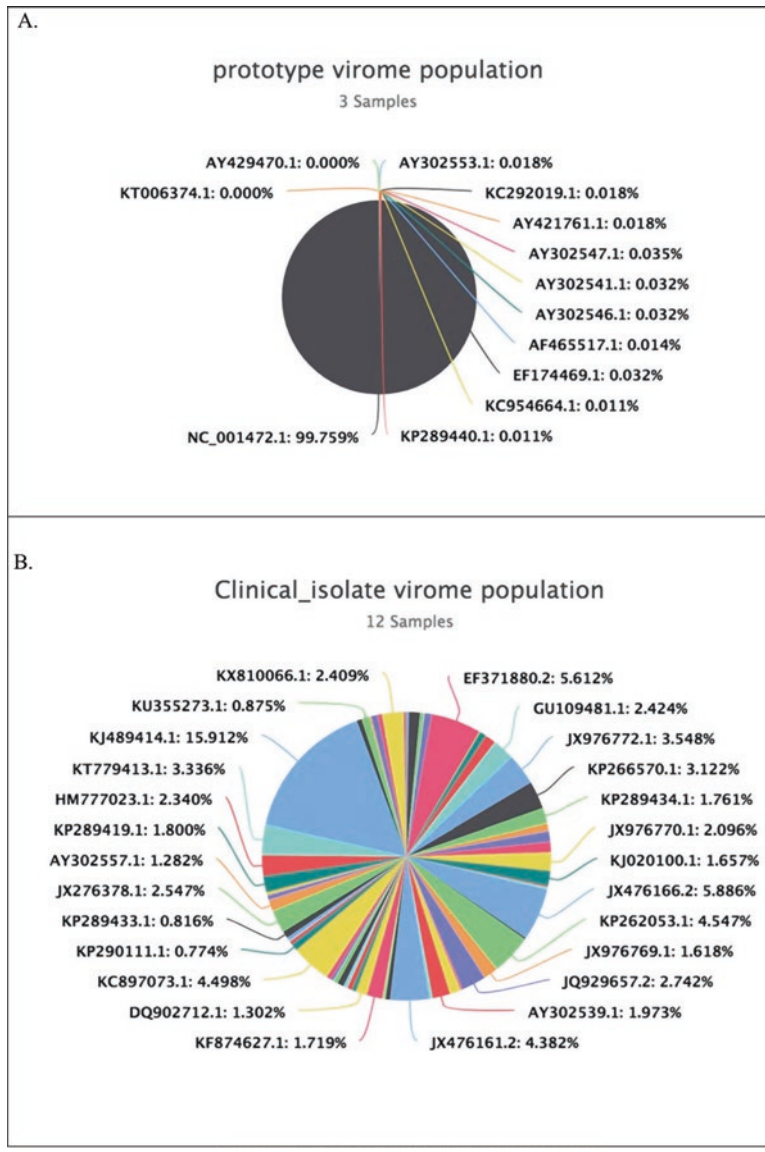

Fig. 2 Vipie virome population profiling—known CVB1 prototype strain (**a**) and wild CVB1 clinical isolate (**b**)

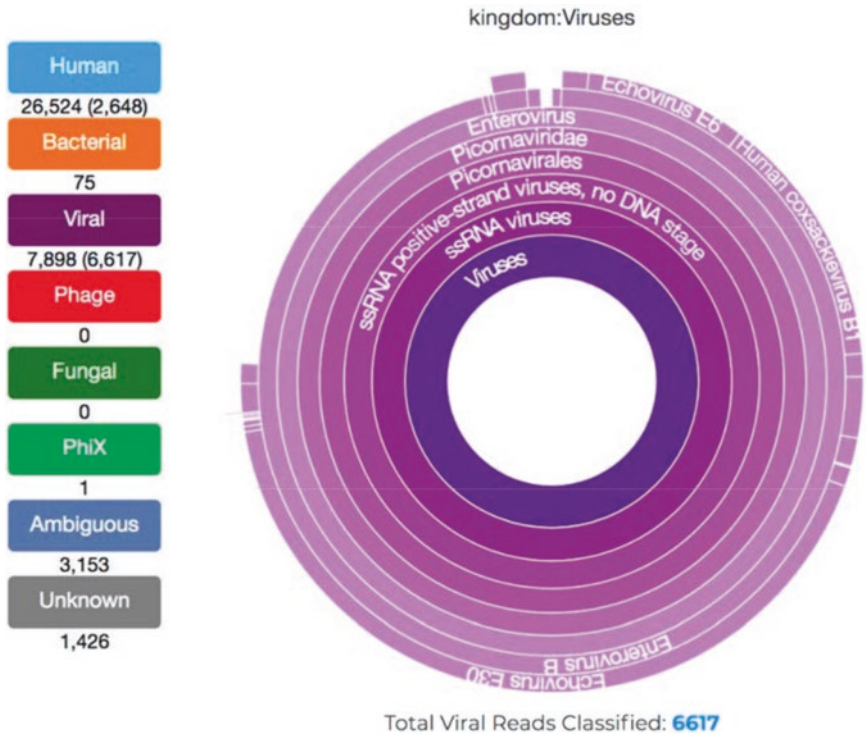

Fig. 3 Virome profile validation of clinical isolate of CVB1-Taxonomer pipeline operates on one sample at a time and confirms CVB1 but also other EV accessions

between 1% and 15%. The heterogeneous and divergent profiles are expected in the analysis of clinical virus isolate and confirmed the need for enhanced custom methods and also motivate the need for novel genome for accurate and sensible mutation assessments.

Vipie profiled results were confirmed using virome profiling pipeline Taxonomer. Though Taxonomer restricts submissions to one sample (two paired end sequences files), the accurate and easy to use tool produces attractive and intuitive results. Clinical CVB1 isolate sequence from time point 0 was analyzed using Taxonomer, the results reported CVB1 along with two Echovirus strains (Fig. 3). The amount of human, bacterial, and unknown viral reads identified are consistent with Vipie aligned distribution report. Taxonomer tool is ideal for fast metagenomics profiling as it maps the inputs reads directly onto known references but as the tool skips de novo assembly, it is not suitable for consensus construction as contigs are not produced.

3.2 QC and De Novo Assembly

Analysis of changes occurring in the CVB1 virus genome over time during infection of pancreas model cells was carried out to investigate the possible role of CVB1 in Type 1 diabetes. To compare changes between time points a consensus sequence of the genome for each time point was created from fastq files generated through

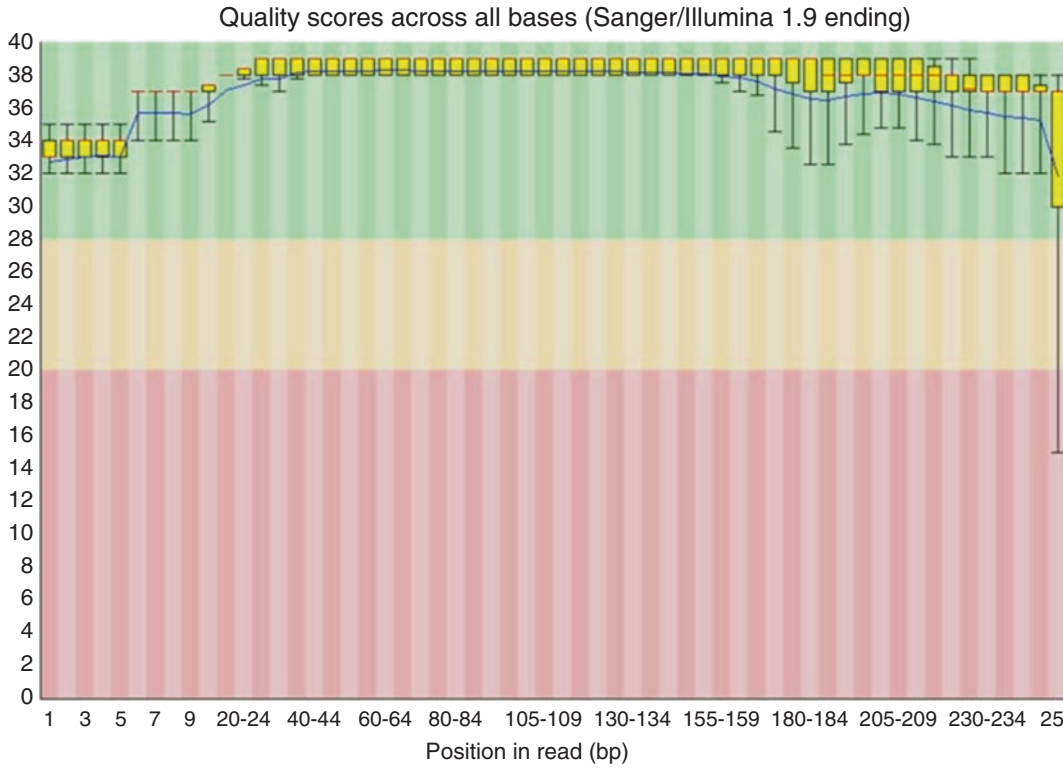

Fig. 4 A per base quality score plot from FastQC. The x-axis indicates insert positions along the reads and the y-axis represents possible quality score ranges. A box-and-whisker plot of quality scores is calculated for each position or range of positions across all reads. The median quality score stays well above 28, the cutoff for "very good quality scores" until the very end of the insert length

NGS. This section will discuss how the clinical isolate samples were processed from quality control of the fastq files to reporting changes across time points. The fastq files from clinical sourced CVB1 were examined in FastQC and determined to have acceptable quality so no trimming was done on the reads (Fig. 4). To assess quality of the sequenced reads, the phred quality scores were examined for each position in the reads. The phred score is −10 log10 Probability {nucleotide was correctly identified}. Therefore, a phred score of 20 means there is a 99% probability the nucleotide was correctly identified during sequencing.

An important component of virome sequencing analysis involves estimation of potential bacterial and human mapped reads. Figure 5 shows that there are partial reads mapping to human and bacterial 16S ribosome marker, a majority of sample reads are classified as dark viral matter. The high number of dark viral proportions pinpoint the lack of sufficient references and also greater genetic distance between clinical strains and references due to higher evolutionary mutation rates [14].

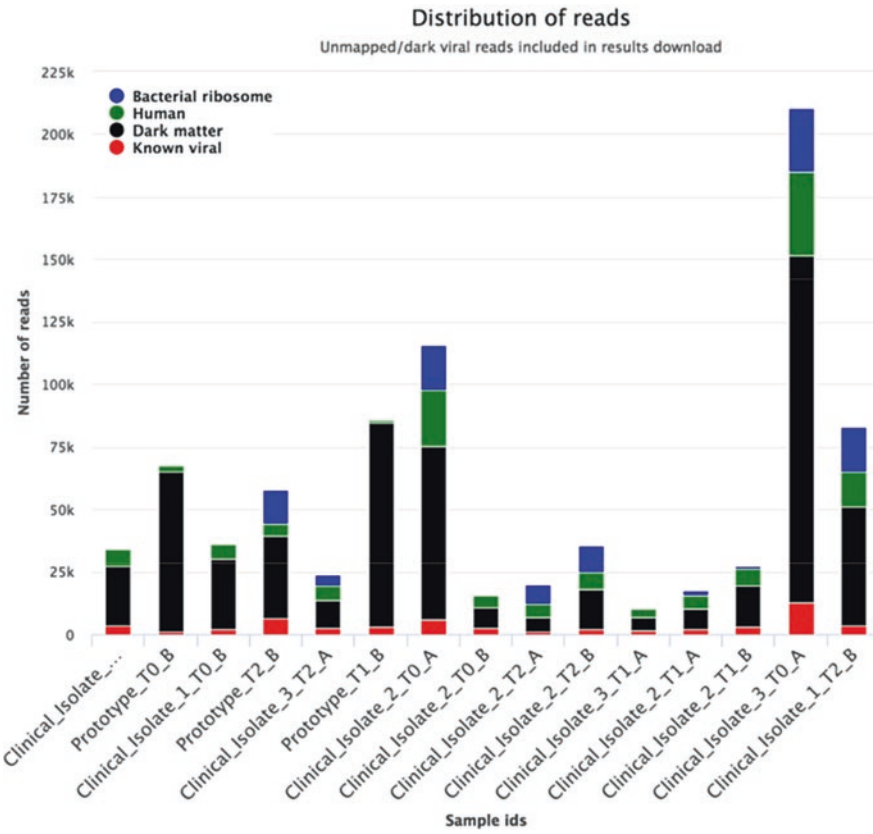

Fig. 5 Mapped distribution report produced from Vipie reports partial virome reads mapping to bacterial and human references. Because of lack of updated genome references and high mutation rates, it can be seen that a majority of reads, in black, are classified as unknown viral reads (Dark matter)

Sanger sequencing was done on the clinical isolate particles when they first arrived at the research facility and a nearly complete genome sequence was produced. This sequence was used as the reference to help assemble a consensus sequence from for the first time point. The steps to assemble a consensus sequence involved first creating contigs (a sequence created from overlapping reads) from the fastq reads, aligning the contigs to a reference, and using the nucleotide sequences from the contigs to form the consensus. Vipie uses Velvet by default, a de Bruijn graph-based [15] de novo assembler, to assemble contigs from short-read fastq files produced by NGS sequencing. The program KmerGenie was used to estimate the optimal k-mer size and average coverage cut-off parameters used during de novo assembly. In addition, the minimum contig length and average read size for each sample from the fragment analyzer were given as parameters to the assembler. Velvet assembly consists of two steps. Operating on quality filtered and interlaced read files, the first step creates a "hash" set of k-mer sized length nodes, storing all (k-1) overlapped nucleotides. The second step attempts to walk the nodes

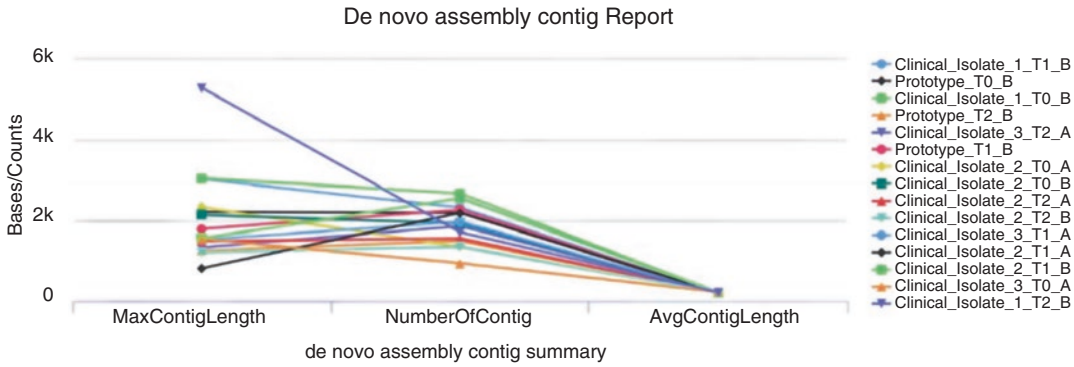

Fig. 6 Contig report. Contig outputs, with maximum size ranges around 2000–5500 and average length of 300 are shown. An estimated 2000 contigs are produced and based on Vipie QC read report (not shown). Contig sizes do not directly correlate with raw number of sequence reads

by linking the unique nodes, the path of the hash nodes represents the contigs. Explained more in notes, essentially k-mer size selection represents a balance between specificity and precision as larger k-mer sizes increase specificity due to increased overlaps and at the same time lower sensitivity as more reads are discarded. Generated CVB1 contig report from Vipie is shown in Fig. 6 where each sample produced roughly 1200–2500 contigs and a sample maximum length of more than 5000. Contig results are also stored as separate fasta files for each sample and accessible directly from Vipie secured results where they can be downloaded.

3.3 Consensus Construction

The contigs were aligned to the reference strain using BWA producing alignment files in the sam file format. The amount of the reference covered by each set of contigs was recorded. In addition, contigs with mapping quality below MAPQ score of 20 were removed. Consensus sequences were created for all virus strains collected at the three different time points of the experiment. The time point 0 sequence was used as a reference for time point 1 and time point 1 for time point 2. It is likely that this approach will work and be sustainable for studies with denser and higher number of serial samples. Time point $(i + 1)$ consensus sequences were made using the sequences from the contigs in the order that they were aligned to the reference at i. Vipie by default produces contigs using Velvet, the contigs are scaffolds, implying nonoverlapping. For each referential position covered by a contig, a nucleotide was appended to the consensus sequence. In scenarios when differences between the contig and the reference were found, the nucleotides from the contigs were appended to the targeted consensus sequence. The result was a sequence identical to the reference except for substitutions and indels found in the relevant time point contigs. The consensus contig coverages and assembly parameters generated from

Table 2
Contig coverage of reference

Strain time point	Kmer/coverage	Contig genome coverage relative to latest consensus
Clinical isolate 1 t0	21/2	96.4%
Clinical isolate 1 t1	25/2	91.7%
Clinical isolate 1 t2	31/2	80.2%

KmerGenie for all three time points are shown in Table 2 and the steps below detail the described approach:

3.3.1 De Novo Assembly Parameters

Make a kgen.in **file for paired-end reads that contains the names of the forward and reverse** fastq files:

```
vim kgen_8.in
```

KmerGenie to determine the correct hash length and cutoff:

```
/kmergenie-1.7044/kmergenie kgen_2.in
```

3.3.2 Serial Based Consensus Comparison and Construction

Step 1: Create bwa index from consensus sequence from previous point or reference sequence.

```
bwa index--p Clinical_Isolateref Clinical_
Isolateref.fa
```

Step 2: Align the contigs from the current point to the consensus sequence from previous time point or reference. Result is an alignment file describing where contigs aligned along the previous time point sequence.

```
bwa mem--t 3--L 6,6 /path/ref_Seq/
Clinical_Isolateref Timepoint_[i]_contigs.
fa > Timepoint_[i]_alignedto_Clinical_
Isolateref.sam
```

Step 3: Extract contigs with MAPQ > = 20 (adjustable via parameter).

```
/usr/bin/python2.6 /path/extract_sam_ac-
cession_simp.py Timepoint_[i]_aligned-
to_Clinical_Isolateref.sam Timepoint_[i]_
alignedto_Clinical_Isolateref_extracted.
sam 20
```

Visualization–- IGV

Step 4: Sort contigs according to start position.

```
igvtools sort Timepoint_[i]_aligned-
to_Clinical_Isolateref_extracted.
sam Timepoint_[i]_alignedto_Clinical_
Isolateref_extracted_sorted.sam
```

Step 5: Create an index of the contig alignment sam file.

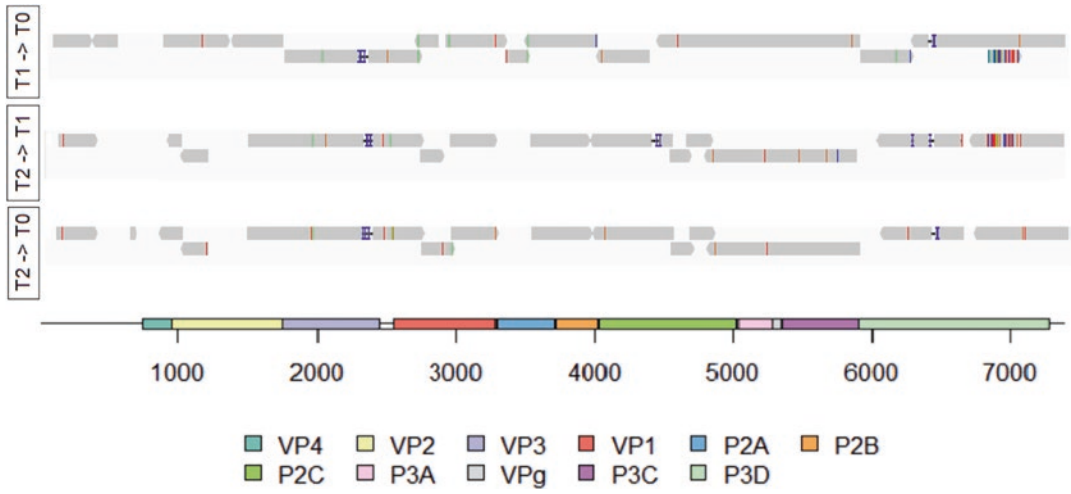

Fig. 7 Contigs from time point 0 are shown aligned to the clinical isolate reference genome in the top track. Time point 1 contigs are aligned to the time point 0 consensus sequence in the middle track and time point 2 contigs are aligned to the time point 1 consensus sequence in the bottom track

```
igvtools index Timepoint_[i]_alignedto_
Clinical_Isolateref_extracted_sorted.sam
```

Step 6: Determine differences between timepoints and construct a consensus sequence for the current time point.

```
python /mutation_caller_V4.py Timepoint_
[i]_alignedto_Clinical_Isolateref_ex-
tracted_sorted.sam /path/ref_Seq/
Clinical_Isolateref.fa /path/Contigs/clini-
cal_isolate/mut_seqs_clinical_isolate
```

Step 7: Repeat with next time point(Fig. 7).

3.4 Variant Calling

Comparisons were made between the prototype sequence and different time points of clinical isolate. A script was written in python that tabulates single nucleotide polymorphisms (SNPs) and indels (insertions and deletions). The script made use of pysam, a module that creates easy access to nucleotide read information from an aligned (.sam) file. During the comparison between the reference and time point 0, each contig from the time point 0 sample was examined for indels from nucleotide differences compared to the window region of the reference to which it aligned. These differences, relatively consistent between time points, are reported in Table 3.

It has been reported that CVB3 variants and deletions within 5′ terminal [16–18] impact virulence and are also of great interest pertaining to this project. The variant results thus far have not matched, and probably should not since this study is using novel clinical isolate CVB1, a different serotype but warranting considerably more analysis. The fact also is that the variants reported are based on majority consensus but minor allele fractions, while less

Table 3
The total number of SNPs and indels found in clinical isolate serial samples grown in PANC-1 cells

	Clinical Isolate 1T0	Clinical Isolate 1T1	Clinical Isolate 1T2
SNPs	15	24	18
Deletions	1	0	3
Insertions	0	1	2

study, have important biological roles including associated cancer odds ratios [19]. The tools and steps introduced are largely applicable for minor allele study and benefits from the constructed consensus. These steps will also benefit from additional analysis to include comprehensive translation and classification of variant functional and potential important untranslated region (UTR) impacts.

4 Notes

1. CVB1 was used in this study because of possible association to T1D. It also represents a good model for novel virus assembly. CVB5 was selected as an example strain to demonstrate diverse EV sub-serotype genetic distances.
2. Pancreatic ductal and islet cell lines were selected as they are known to be able to establish a persistent infection model for CVB1 and highly relevant for T1D.
3. Virome sample profiling was validated with VirusTAP [20]. The tool is web based and includes de novo assembly and outputs contigs. As VirusTap removes all human and bacterial reads prior to assembly, it is not fit for consensus.
4. In the scenario within consensus construction where particular samples exhibit magnitudes more contamination or dark matter reads relative to other samples, the contigs generated likely will in effect possess similar proportions. Remapping directly to virus genomes of interest with a low MAPQ score cutoff and then using the aligned reads for de novo assembly can assist with building more desirable contigs.
5. KmerGenie [8] is recommended for optimizing k-mer size, "k," and k-mer coverage cutoff for Velvet and all de Bruijn graph solutions. K-mer size must be odd and smaller than read insert length. The program estimates the k-mer size that maximizes the number of unique k-mers found in the reads in a fastq file or set of paired fastq files produced from NGS sequencing. In a de Bruijn graph, these k-mers can be linked by finding common prefixes and suffixes thus forming contiguous regions of the original genome. Shorter k-mers produce

more connections between reads increasing the proportion of correctly connected reads (high sensitivity) but also increasing the proportion of incorrectly connected reads (low specificity). In contrast, longer k-mers produce fewer connections decreasing the proportion of correctly connected reads but increasing the proportion of correctly unconnected reads. The k-mer coverage is the number of times the k-mer was observed in the reads. A k-mer coverage cutoff determines the number of times a k-mer must be seen in the reads before it is used in the assembly.

References

1. Pociot F, Lernmark Å (2016) Genetic risk factors for type 1 diabetes. Lancet 387(10035):2331–2339
2. Krogvold L, Edwin B, Buanes T et al (2015) Detection of a low-grade enteroviral infection in the islets of langerhans of living patients newly diagnosed with type 1 diabetes. Diabetes 64(5):1682–1687
3. Sane F, Caloone D, Gmyr V et al (2013) Coxsackievirus B4 can infect human pancreas ductal cells and persist in ductal-like cell cultures which results in inhibition of Pdx1 expression and disturbed formation of islet-like cell aggregates. Cell Mol Life Sci 70(21):4169–4180
4. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842 https://doi.org/10.1093/bioinformatics/btq033
5. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760
6. Babraham Bioinformatics, FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 02 2018
7. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
8. Chikhi R, Medvedev P (2013) Informed and automated k-Mer size selection for genome assembly, HiTSeq
9. Li H, Handsaker B, Wysoker A et al (2009) 1000 genomes project, the sequence alignment/map format and SAM tools. Bioinformatics 25(16):2078–2079
10. Flygare S, Simmon K, Miller C et al (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol 17(1):111
11. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829
12. Lin J, Kramna L, Autio R et al (2017) Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. BMC Genomics 18(1):378
13. pysam. https://github.com/pysam-developers/pysam. Accessed 02 2018
14. de Bruijn NG (1946) A combinatorial problem. Koninklijke Nederlandse Akademie v Wetenschappen 49:758–764
15. Attar N (2016) Viral evolution: more of the world's a phage. Nat Rev Microbiol 14:269
16. Massilamany C, Gangaplara A, Basavalingappa RH et al (2015) Mutations in the 5' NTR and the non-structural protein 3A of the Coxsackievirus B3 selectively attenuate myocarditogenicity. PLoS One 10(6):e0131052
17. Chapman NM, Kim KS, Drescher KM et al (2008) 5′ terminal deletions in the genome of a coxsackievirus B2 strain occurred naturally in human heart. Virology 375(2):480–491
18. Rinehart JE, Gómez RM, Roos RP (1997) Molecular determinants for virulence in coxsackievirus B1 infection. J Virol 71(5):3986–3991
19. Pomerantz MM, Freedman ML (2011) The genetics of cancer risk. Cancer J 17(6):416–422
20. Yamashita A, Sekizuka T, Kuroda M (2016) VirusTAP: viral genome-targeted assembly pipeline. Front Microbiol 7:32