K-mer Profiling Powered by Reference-assisted Assembly of NGS Data: A Highly Sensitive Protocol to Infer the Plasma Microbiome using Cell-free DNA Sequence Data

Rohita Sinha, Ellis Bixler, Michelle Altrich and Steve Kleiboeker
Viracor Eurofins, Lee's Summit, MO USA

Viracor | Eurofins Clinical Diagnostics

INTRODUCTION

Cell-free DNA (cfDNA) has emerged as an important clinical specimen to probe for pathogenic microbes, especially in organ transplant patients where the same data can be used to predict allograft rejection. Recent reports described viral, bacterial or the complete microbial diversity in plasma following cfDNA sequencing. The prevalence of certain viral families (anelloviridae) is associated with immunosuppressant dosage and the risk of antibody mediated rejection. While being informative, the cfDNA reads are inherently shorter in length (~160bp or 2x75bp) and predominated by the host DNA (~97-99%), causing challenges in their taxonomic annotation and lower specificity. Here we present a computational protocol which minimizes these challenges by merging the concept of "Reference-assisted Assembly" with K-mer profiles of NGS data, for highly sensitive and specific microbial detection.

MATERIALS AND METHODS

Viral Genome Database

A list of 129 human viral pathogenic species was obtained from the Viralzone[1] database (https://viralzone.expasy.org), and the corresponding genomes were retrieved from NCBI database using their "edirect" API. This list was further refined to select only DNA viruses (36/129).

Curation of Database Sequences

The following steps were performed: (1) Genomic sequences with key words such as 'nearly complete genomes' and 'constructed genomes' were removed, (2) We used Gramdist[2] tool to find the intra and interspecies full-genome distances (0.348 and 0.62 respectively). These numbers were used to define species which are genetically closer to each other (example Vaccinia virus, Monkeypox virus, Cowpox virus and Horsepox virus) and also the strains of viral-species which are genetically divergent (distance >0.62) from other member of their species (example Hepatitis B virus, Torque teno virus).

Targeted DNA Virus Species

The following viruses were targeted: human adenovirus, human herpesvirus 1, human herpesvirus 2, human herpesvirus 3, Epstein-Barr or human herpesvirus 4, CMV or human herpesvirus 5, human herpesvirus 6, human herpesvirus 7, human herpesvirus 8, BK polyomavirus, human parvovirus B19, JC polyomavirus, Torque teno virus, KI polyomavirus, WU polyomavirus.

Training KrakenUniq classifier

We used KrakenUniq[3] (version 0.5.8, https://github.com/fbreitwieser/krakenuniq) predict the viral origin and species of NGS reads/assemblies. KrakenUniq leverages the taxa specific K-mers and uses the total count of such unique K-mers to classifies a DNA fragment into a taxonomic bin. We trained KrakenUniq using our viral genome database; and to capture the genomic K-mer patterns at different scale, three different K-mer values (16, 21, and 31) were used. All three KrakenUniq versions were applied for the taxonomic classification of NGS reads/assemblies and the majority-wins rule was invoked to assign the viral species.

MATERIALS AND METHODS (CONTINUED)

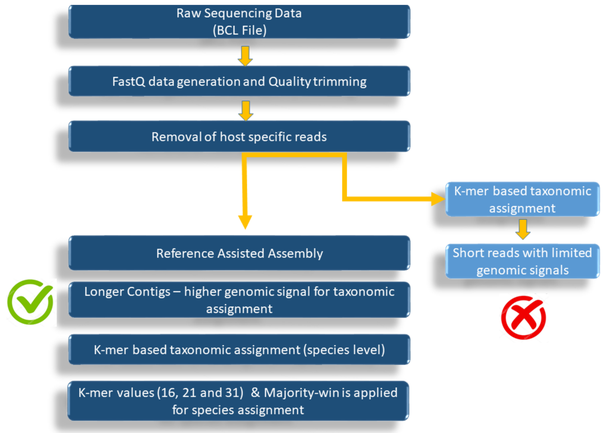

Figure 1: Flow diagram to depict the steps of our virome analysis tool. Input the pipeline is a BCL file generated using NextSeq platform. We computationally remove the reads originating from the host (human) and the non-host reads were assembled to contigs using an in-house "Reference assisted assembly" program. Assemblies were taxonomically annotated using KrakenUniq (3 versions, K-mer 16, 21, 31) and species level assignment supported by at least two versions were selected pipeline.

Simulated Test Sample Generation

We used 'wgsim'[4] tool (https://github.com/lh3/wgsim) to generate artificial fastq data using reference viral genomes and used an in-house Python script to sample reads in different proportions to generate 30 samples. Simulated samples had different degree of viral species complexities (1, 2 or 3 species) and different proportions of each species (50,100 and 500 artificial reads).

Clinical Sample Description

There were 29 clinical plasma samples obtained from a biorepository. These samples were drawn during a period of no rejection (14), acute cellular and/or humoral rejection (11), and BKV associated nephropathy (4).

RESULTS

Table 1. Results for simulated samples.

Species Used in 1 Virus Mix	*Contig Length Range*	*Prediction Accuracy*
Human herpesvirus 1, Human herpesvirus 2, KI polyomavirus, BK polyomavirus, Human cytomegalovirus, WU polyomavirus	4542 to 11995 bp	100%
Species Used in 2 Virus Mix	***Contig Length Range***	***Prediction Accuracy***
Epstein-Barr virus, Human adenovirus 4, Human herpesvirus 2, Human herpesvirus 6, BK polyomavirus, Human herpesvirus 7, Human cytomegalovirus, KI polyomavirus, JC polyomavirus, Torque teno virus, WU polyomavirus	4357 to 44679 bp	100%
Species Used in 3 Virus Mix	***Length Range***	***Prediction Accuracy***
KI polyomavirus, Human cytomegalovirus, Human parvovirus B19, Epstein-Barr virus, WU polyomavirus, Human herpesvirus 1, Human herpesvirus 8, BK polyomavirus, Human herpesvirus 7, Human adenovirus, Human herpesvirus 6, JC polyomavirus, Human herpesvirus 2	3633 to 45633 bp	100%

RESULTS (CONTINUED)

- Clinical samples were comprised of 15 rejection and 13 non-rejection cases.
- Viral DNA fragments were detected in 22/29 samples: 11/15 rejection and 10/13 non-rejection.
- ViraOme predictions for HHV7, CMV, BKV, JCV, EBV, and HSV1/2 were further validated using our qPCR assays.

Table 2. Clinical Samples: Computational Predictions & Validation by qPCR.

Sample ID	*Rejection Status*	*ViraOme Prediction (qPCR test)*	*qPCR Confirmation*	*Viral Count*	*Contig Lengths*
1	Rejected	EBV	Not-tested		147 bp
2	Rejected	BKV, (EBV)	EBV	145 IU/ml	1328, 407 bp
3	Rejected	BKV, TTV	Not-tested		527, 452 bp
4	Rejected	(ADV),TTV	Not-detected	N/A	145, 248 bp
5	Rejected	(BKV, ADV), TTV	BKV	870 c/ml	419, 192, 144 bp
6	Rejected	TTV, (ADV)	Not-detected	N/A	2435,80 bp
7	Rejected	TTV	Not-tested		2452 bp
8	Rejected	(BKV, EBV), JCV	BKV, EBV	110,100 c/ml, 9IU/ml	5022, 138, 3998 bp
9	Rejected	(BKV), TTV	BKV	838,500 c/ml	4717, 354 bp
10	Rejected	(BKV, JCV), ADV, TTV	BKV, JCV	2.1 x 10^6 c/ml, 239,400 c/ml	5009, 3422, 145, 253 bp
11	Rejected	(BKV), HSV1/2	BKV	1 x 10^7 c/ml	5104, 148, 108 bp
12	Non-rejected	(HHV7)	HHV7	6 c/ml	146 bp
13	Non-rejected	TTV	Not-tested		523 bp
14	Non-rejected	(CMV), TTV	CMV	300 IU/ml	398, 96 bp
15	Non-rejected	EBV	Not-tested		379 bp
16	Non-rejected	(ADV, CMV), TTV, HHV7	Not-detected	N/A	229, 148, 147, 144 bp
17	Non-rejected	(BKV, JCV), TTV	BKV, JCV	448,300 c/ml, 5500 c/ml	4963, 364, 574 bp
18	Non-rejected	(BKV, JCV), ADV, EBV	BKV, JCV	42,400 c/ml, 400 IU/ml	4461, 877, 146, 102 bp
19	Non-rejected	TTV	Not-tested		1821 bp
20	Non-rejected	TTV, (ADV)	Not-detected	N/A	610, 102 bp
21	Non-rejected	(BKV, JCV)	BKV, JCV	451,800 c/ml, 105,600 c/ml	5097, 4901 bp

CONCLUSIONS

The ViraOme tool, designed to detect pathogenic viruses using cfDNA data performed well both on simulated and clinical samples with a majority of results confirmed by qPCR. Our results emphasize how computational predictions can complement clinical diagnostic approaches. It's also worth noticing that we failed to confirm the *in silico* prediction of adenovirus in all the samples, which may either be attributed to the possibilities of a low titer (hence, short contig lengths) or adenovirus being integrated to the host genome, leading to non-amplification of PCR targets.

References

1- ViralZone: a knowledge resource to understand virus diversity, Nucleic Acids Res. Jan 2011.
2- A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences, *BMC Bioinformatics,* 2010
3- KrakenUniq: confident and fast metagenomics classification using unique k-mer counts, Genome Biology, 2018
4- wgsim – Read simulator for next generation sequencing Github Repository: http://github.com/lh3/wgsim

MM 1195 REV0 1020