Copenhagen University

Bioinformatics MSc Project

The Human Virome In The 1000 Genomes Project

Adham Mandour

supervised by

Dr. Anders Albrechtsen

August 2018

Introduction

Over the past several years interest and active research in viral metagenomics has been steadily increasing, and not without due cause; in any given environmental habitat, including the human microbiome, at least 1031 viral particles exit at any point in time. Humans in particular carry a 10-100 fold viral particles to human cells, including the gastrointestinal tract where extensive research on the bacterial microbiome has already been under way for more than a decade [64]. This may in fact be less innocuous that it sounds; firstly due to immune systems' response to viral infections, and its role in regulating transkingdom interactions within a microbiome through immunomodulation. Viral immunomodulation has been implicated as a significant contributing factor in chronic autoimmune diseases (lupus, rheumatoid arthritis, multiple sclerosis and asthma and several more). It has also been shown to confer beneficial effects in the form of protection against bacterial infections in the gastrointestinal tract (as shown in mice) and a heightened antiviral immunity and superior response such as the case in young human adults carrying herpesvirus[9]. On an individual level, there are also widely disparate variations, even among immediate relatives [32] [42] .Furthermore, the heavily underestimated diversity of viruses, the number of novel viruses being continuously discovered within the human body, as well as the shortcomings of fully sequenced viral genomes with "Hallmark" universal genes remain open issues in profiling the human virome, and a challenge in taxonomy in general [48]. Efforts have been made to probe DNA viral diversity within humans however, and a recent study found 3,761 novel DNA viral sequences circulating as cell-free circular DNA in 1,351 blood samples [22].

The aforementioned not only highlights the significance and prevalence of the human virome within the body, but also the difficulty in fully characterizing what would be a typical virome of a healthy individual. Towards that end, a recent metagenomics study (JCVI) with a focus on homology-dependent methods rather than novel discovery, explored the unmapped sequences of 8,420 deep sequenced individuals [53] in an attempt to profile a normal blood phenotype. Sequences were obtained from whole genome of sequencing of blood, and the study found 94 viruses (19 human DNA viruses with the rest being concluded as reagent contaminations) in 42% of all -disease free- individuals. The researchers also found significant contributions of age, sex and ancestry in viral infections [34].

Aim:Our goal is to apply similar methods of the latter study on the freely available 1000 Genomes[4] to capture a profile of the human virome. We futher aim to conduct Genome-Wide Analysis (GWA) on the most prevalent viruses found to test for associated loci with viral abundances. There are several marked differences in contrast with the methodology of JCVI study with this study; firstly, samples from the 1000 Genomes Project are obtained from cell cultures (lymphoblastoid cell lines) and are of low-coverage (4x – 8x).Furthermore, because the samples are publicly available, health related records including age and gender of the indivudal was not recorded. Individuals sequenced in the 1000 Genomes Project self-reported as "healthy" however. These factors combined would lead us to expect significant differences between the results of this study and the previous one, in particular lower viral abundances, but we hope to nonetheless ascertain the validty of the method within a low-computational setting. To add another layer of complexity and to exclude atleast one

source of variance, the same methods were applied on low-coverage (11x) whole-exome sequence data of 2,000 Danish individuals.

Methods

Data Selection And Processing

Study Samples And Sequencing

All samples were obtained from the 1000 Genomes repository where low coverage bam files were available (4x–8x). In particular, 4 pairs of similar populations were included to represent a wide range of ethnic diversity, and pairs were included specifically to increase the overall number of samples: GBR, British, CHB, Han–Chinese, CLM, Colombians, IBS, Spanish, JPT, Japanese, LWK, Luhya, Peruvian, PEL, and Yoruba, YRI. Different instiutions handled sequencing of the low coverage genomes, but it is assumed that the platform used was Illumina HiSeq 2000/2500 generating 100bp reads.

Population key	GBR	CHB	CLM	IBS	JPT	LWK	PEL	YRI
Number of Samples	93	105	96	109	104	101	88	111

In addition, low-coverage whole-exome sequencing data of 2,000 Danish individuals that was part of two larger studies [1] [28] was also obtained and included in this study.

Extraction of Unmapped Reads

Following the methodology of the JCVI study, both unmapped read pairs and partially unmapped reads were used in this study. Paired unmapped reads and singletons were extracted from each bam file using Samtools [27] . String Graph Assemlber [49] was used to identify and filter low complexity and low sequencing quality reads with parameters dust–threshold = 2.5 and quality–filter = 50

Building Viral Databases

RefSeq & Mixed Viral Databases

Two databases were contructed. The first database was construted with the NCBI RefSeq viral reference genomes [41].For the purpose of identifying and removing potential false positives, a more comprehensive database was constructed with the initial RefSeq viral genomes along with the UCSC genomes of human, chimp, mouse, chicken, and fruit fly, and representative genomes of Fungi(14/2) Bacteria(1,543/1636) and Archaea(140/389) (where the first number is the number of genomes found on the NCBI database as opposed to the ones included in the JCVI study). It also included NCBI nt

vectors (117,660/274,565), (758/778) plasmids and for the explicit purpose of detecting commercial vectors or plasmids, NCBI's non-redundant UniVec database was also added.

The large differences in the number of vectors is attributed to the ambiguity of searching for "nt vectors" as mentioned verbatim in the JCVI paper, as opposed to searching the ncbi database with the literal parameter "expression/cloning vectors" as used in this study.

Viral Read Identification And Estimation

Identification of Viral Reads

Unmapped reads were searched against the NCBI RefSeq viral reference genomes [41] using blastn [3] and filtered with *e*-value $\leq$ 1e–10. Reads with unique hits to phages were removed from further analysis as they are non-relevant to the interest of this study, and to reduce computational demand. Candidate reads based on the initial blast run were then collected and searched against the comprehensive database with blastn (for the 93 GBR samples only), and bwa-mem for all populations. For blastn, reads were filtered with *e*-value $\leq$ 1e-20 and a minimum score threshold matching the following formula:

$$(Matches - 70) - (Gaps \times 2.5 + Mismatches \times 2) \geq 0$$

A single BWA-mem run was matching the scoring matrix used in blastn with parameters –k 28 –B 2 –A 1 –O 0 –E 2.5.

Lastly, reads with higher-scoring hits to non-viral organisms than to viral organisms were flagged as false positives, and discarded.

Estimating Viral Abundances

Viral abundance can be interpreted as viral genomes per human cell/viral copies per 100,000 human cells using the following equation:

$$virus\ abundance = \frac{2 \times \frac{number\ of\ reads\ mapped\ to\ viral\ genome}{viral\ genome\ size}}{\frac{number\ of\ reads\ mapped\ to\ human\ genome}{human\ genome\ size}}$$

Viral Integration

After filtering for false positives, reads were collected and searched against a seperate database (Viral RefSeq Genomes + HG38 Human Genome) with bwa-mem on default settings to detect integration events. An integration event for a read is determined if one mate of a pair was assigned to a viral genome, and the other mate to the human genome, or if a single read was assigned as a split read on the human genome.

Genome-Wide Analysis

Genotype Data & SNP Association Tests

Genotype data with ¿ 80 million variants was obtained from the 1000 genomes project variant call file and filtered for low complexity regions, SNPs falling within CNV regions, MAF¿0.05 and genotype frequency¿0.01 and subsetted for the individuals included in this study using plink[10] with the command –geno 0.01 –maf 0.05 –chr 1-22. Out of the original markers, 6,911,733 remained after quality control. Genome-Wide association was conducted for EBV and Human adenovirus C. The former was log-transformed, and approximated to a normal distribution through quantile transormation and the latter was recoded into case-control depending on whether or not the virus was detected in an indivudal. Finally, PCA was performed on the genotype data, and a linear/logistic model association test was run on each viruse with covariates from the top 10 principle components to adjust for population stratification.

Results

Finding Viral Sequences in Unmapped Reads

RefSeq Viral Database Runs

Through the initial blastn run on the RefSeq database, only 0.02% of unmapped reads were assigned to viruses in all populations (all samples carried a median of 221,863,216 unmapped reads), with a median of 7,111 viral reads. A substanitally lower but expected figure from previously reported median of 400,000:

Population key	GBR	CHB	CLM	IBS	JPT	LWK	PEL	YRI
Viral Unmapped Reads(%)	5.3e-06	1.9e-06	3.9e-06	3.8e-06	10e-06	3.5e-06	9.4e-06	9.3e-06

The viral Refseq search yeilded hits to 106 viruses represented in 575 strain of mostly phages (S:Table 4). In particular, the Human Endogenous Retrovirus K113 was the most prevalent (n=688), followed by Human gammaherpesvirus 4 (EBV, HHV4) (n=661). The bateriophage phiX174 was also highly prevalent as it was in the previous study (n=367). The prevalence of EBV is much higher than previously reported as it is used in transforming B-cell lines into immortalized lymbphoblastoid cell lines(LCLs) used in the 1000 genomes project [**Neitzel1986**].

Mixed Viral Database Runs

It was initially planned that a second round of blast on a comprehensive database was to be performed to filter out potential false positives with *e*-value $\leq$ 1e-20. This proved too computationally demanding however and BWA Maximal Extent Matching algorithm (BWA-mem) [27] was

used instead. The JCVI study filtered reads from the second blast run with bit score $\geq$ 190. This presented two issues (1) translating blastn's bitscore to a reasonable approximation to be applied on reads of lengths of 100bp used in this study as opposed to 150bps, and (2) approximating bwa-mem runs to match blastn as closely as possible, and subsequently filtering reads outside of bwa's native alignment score metric, but in concordance with blastn's scoring scheme.

For testing, candidate reads from the 93 GBR samples were searched against the comprehensive database using both blastn and bwa-mem to ascertain the validity of using bwa-mem. For blastn, candidate reads for each sample were split into 1200 roughly equal parts and blastn was run in parallel batches of 64 utilizing all available cores, for a total of approximately 111,600 blastn runs for all samples. This was a necessary step to reduce memory usage, and as blastn time is linear to query size, it did not increase overall runtime

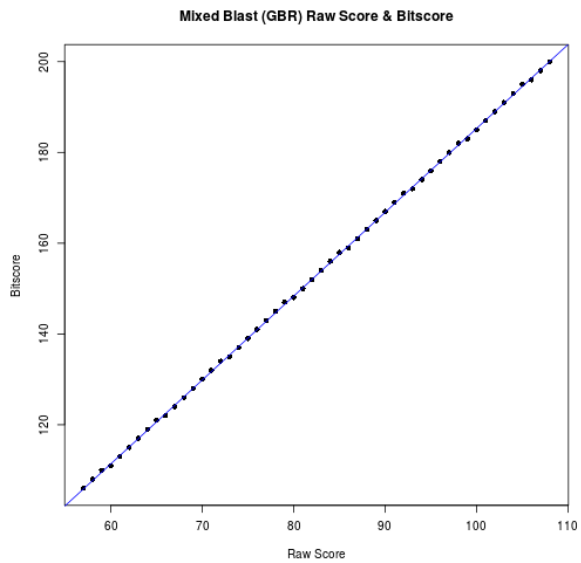

Figure 1: The plot describes the linear relationship between raw scores and bitscores. The raw scores and bitscore coefficient (1.835) is determined by the equation in BLAST: $bitscore = \frac{S\lambda - ln(k)}{ln(2)}$ where parameters K and λ depend on the scoring matrix

The linear relationship between raw score and bitscore was determined to be 1.835 ((Figure 1) which translates to a raw score of 103.5 for a bitscore of 190. The implies that the original intent of the JCVI study was to include reads that have a lower bound of 103.5 positive matches for all hits, or roughly 69% positive matches with no gaps or mismatches between a query of 150bps and a possible match, which is further equivalent to a minimum of 70 positive matches when applied to a query of size 100bps.

BWA-mem results for the test run on GBR (and all subsequent populations) were then filtered based on the following relation to guarantee a minimum raw score of 70 for all hits:

$$(Matches - 70) - (Gaps \times 2.5 + Mismatches \times 2) \geq 0$$

Viral abundances were then estimated from the results of both BWA-mem and blastn to draw a

rough comparison as shown in figure (Figure 2)

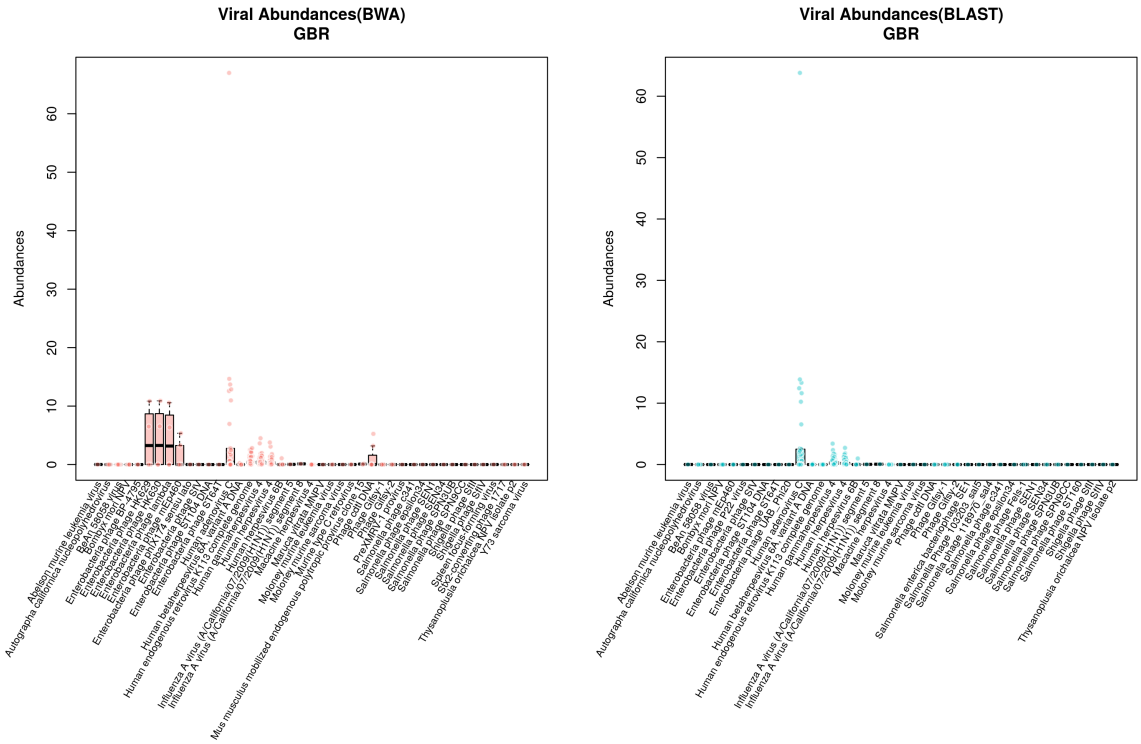

Figure 2: The figure shows estimated viral abundances side by side for the blastn derived results and bwa. blastn assigned more reads to phages than BWA, but all other human and low abundance non-human viruses within the focus of this study are nearly identical in terms of read counts and viral abundances.

For a detailed comparison, the full tabular results on GBR are given in the supplement, including prevalence, median read count and depth coverage for each virus found (blastn: S:Table 3 and BWA: (S:Table 2))

The Virome In The 1000 Genomes

Human Viruses

After eliminating potential false positives, the number of viruses found in all populations (excluding phages) fell to 34 (S:Table 1 and S:Figure 1), of which 17 human viruses were identified((Table 1) and (Figure 3)). Coverage plots were generated for viral hits with relatively high depth coverage or a high number of reads to confirm the presence of full genomic copies of said viruses(Figure 4). The human viruses assigned included 5 strains of Influenza A which are more likely segments of a whole of either strain or the same reads assigned to more than one strain as they are found concurrently in only two individuals from the same population (LWK). As expected, EBV was found in 96% of individuals (n=767) but also Human gammaherpesvirus 4 type 2 (HHV) at exactly the

same level of prevalence. This is likely due to reads with multiple hits to different viruses as EBV types I and II are largely indistinguishable except for genes EBNA-2A and EBNA-3A/B/C which comprise less than 5% of the genome and which furthermore have high relative similarity between them [37] (99% total shared nucleotide identiy verified independatly with blast multiple sequence alignment of the RefSeq genomes). This is also further supported by the high prevalence of Maccacine herpesvirus 4 which is also closely related to EBV[33]. Additionally, both coverage plots for EBV and HHV show the characteristic 12kb deletion exclusive to the B95-8 EBV strain used in cell immortilization(Figure 4)[24]. HHV6A and HHV6B were assigned to 16 samples each, but at further inspection, both viruses were assigned to the same individuals again possibly indicating ambiguity in unique assignment of reads to either type (again with 95% shared nucleotide identiy verified independantly). There is the possiblity of coninfection however, as both viruses share 9598 reads in common out of 16544 and 22464 total reads for HHV6A and HHV6B respectively. Human endogenous retrovirus K113 was found in relative high abundance at 13% (n=102) and Human Adenovirus C at 20% (n=156) unlike the previous study which assigned Adenovirus C at less than 0.0.1% of the sample population. The coverage plot for Adenovirus C demonstrates lack of coverage across the genome with several sharp peaks. Randomly picked reads were searched against the ncbi nucleotide database (online) and were assigned equal scores and perfect identity matches to the RefSeq human adenovirus C, as well as various expression/cloning vectors.

Table 1: Included are the 17 human viral hits including several strains of Influenza A

BWA on GBR			Number of Reads		Viral Abundances		Depth of Coverage	
Virus	Number of Individuals	Percentage of Individuals	Median	Maximum	Median	Maximum	Minimum	Maximum
Human adenovirus C	156	19.57	0	232368	0.03	101.664	0.002	138.825
Human betaherpesvirus 6A, variant A DNA	16	2.01	0	17162	0	1.091	0	4.597
Human gammaherpesvirus 4	767	96.24	1539	3679906	0.29	17.854	0.001	83.452
Human herpesvirus 2 strain HG52	3	0.38	0	5	0	0.001	0.001	0.002
Human herpesvirus 4	766	96.11	1318	3071683	0.25	13.533	0	67.076
Human herpesvirus 6B	16	2.01	0	22682	0	1.09	0	6.026
Human herpesvirus 7	7	0.88	0	55	0	0.006	0.001	0.017
Human papillomavirus	5	0.63	0	7	0.01	0.015	0.01	0.027
Human parvovirus B19	2	0.25	0	9	0.02	0.023	0.154	0.154
Human T-lymphotropic virus 1	4	0.5	0	252	0.08	1.057	0.197	2.389
Influenza A virus (A/California/07/2009(H1N1))	4	0.5	0	20	0.01	0.022	0.011	0.126
Influenza A virus (A/Guangdong/1/1996(H5N1))	2	0.25	0	4	0.01	0.009	0.047	0.047
Influenza A virus (A/Hong Kong/1073/99(H9N2))	2	0.25	0	2	0.01	0.011	0.053	0.053
Influenza A virus (A/Korea/426/1968(H2N2))	2	0.25	0	1	0	0.001	0.007	0.007
Influenza A virus (A/Puerto Rico/8/1934(H1N1))	2	0.25	0	3	0	0.004	0.02	0.02
Merkel cell polyomavirus isolate R17b	2	0.25	0	2	0.01	0.008	0.028	0.028
Torque teno virus	4	0.5	0	5	0.02	0.018	0.035	0.111

All other viruses were found in under 1% of all individuals, including some that were reported in the previous study below the same threshold of prevalence. These include Influenza A (2 individuals), Parovirus B19 (2 individuals), Papiliomvarius and a Herpesvirus 2 strain. Lastly, T-lymphotropic virus 1 was found in four individuals. A perfectly matched read (no clipping, no gaps and no mismatches) was searched against the ncbi database with blastn (online) and did indeed correspond to the virus by having the maximum score in both cases, pointing at the possiblity of Leukemia [69]. Another oncogenic virus that was found was Merkel cell polyomavirus (two individuals, two reads) normally found in apprximately 80% of merkel cell carcinoma patients. [55]

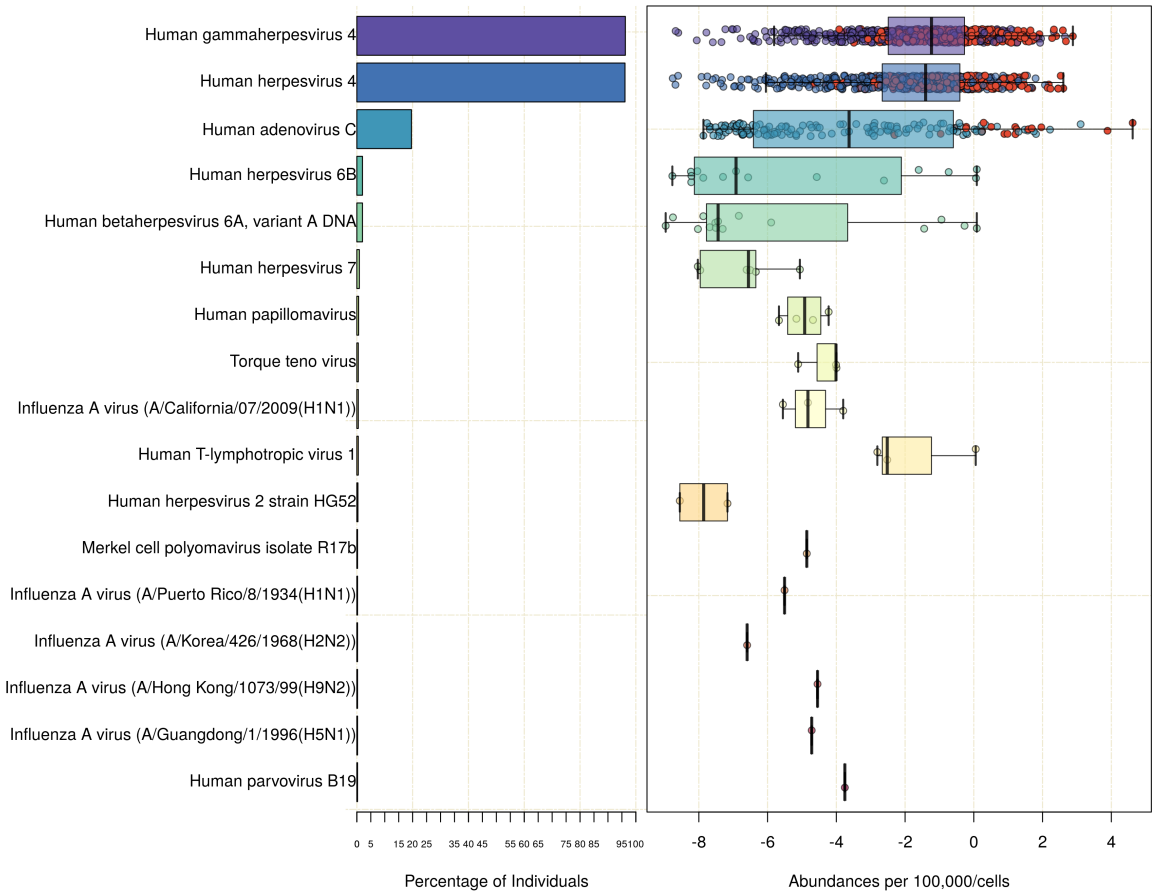

Figure 3: Prevalence And Viral Load In All 797 Samples.The figure shows prevalence of the 17 assigned human viruses along with their corresponding log-transformed viral load/abundance (as genome copies per 100,000/human cells. Red dots indicate evidence of integration.

Curiously, the first and third most prevalent viruses in the JCVI study (Human herpesvirus 7(HH7) and Torque teno virus at 20% and 9% repsectively) were almost entirely absent in this study.

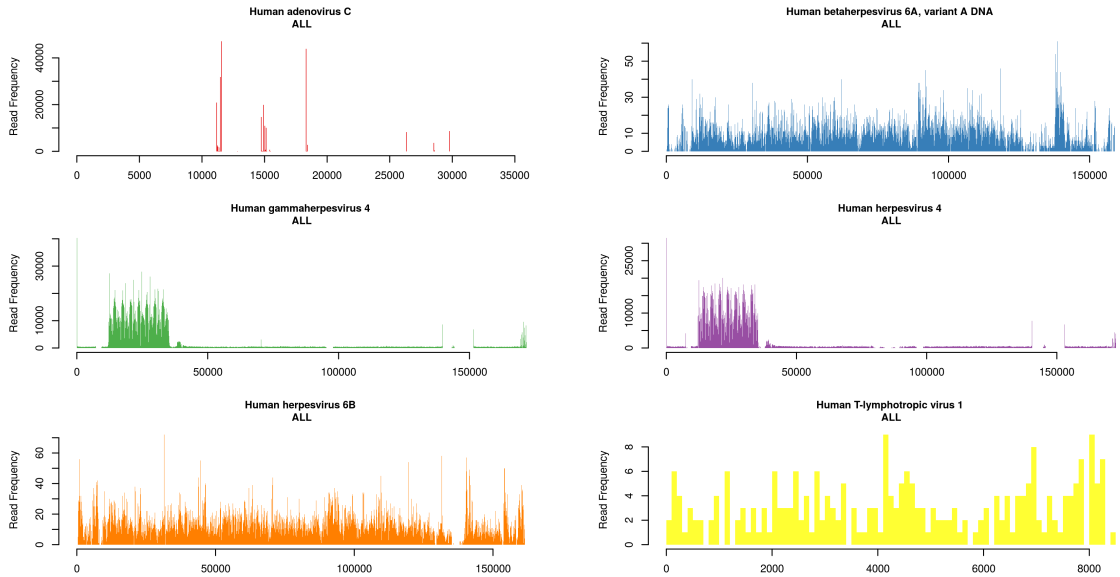

Figure 4: Genome coverage.The y-axis can be interpreted as the depth of coverage/reads per window. The figure shows well defined coverages for most of the selected viruses with the exception of adenovirus, and HH4/EBV which both show excessive accumelation of reads around the EBN-2/LP genes

Whole-Exome Danish Population

The same methods were applied on 2,000 low coverage (approximately 11x using a NimbleGen 2.1M HD array (34.1 Mb, 21,810 genes)) whole-exome sequencing data from a Danish population. The median number of unmapped reads for all samples was 1,930,828, of which only 2786 reads aligned to viruses.

The prevalences and abundances were significantly lower (less than 3%) (Figure 5) with several completely absent viruses. The coverage plots (Figure 6) atleast show concordance with the coverages for EBV/HH4 with the same over–representation around the EBN-2 genes, a few corresponding peaks in Adenovirus C, and a somewhat full genomic coverage for HBH-6A/B.

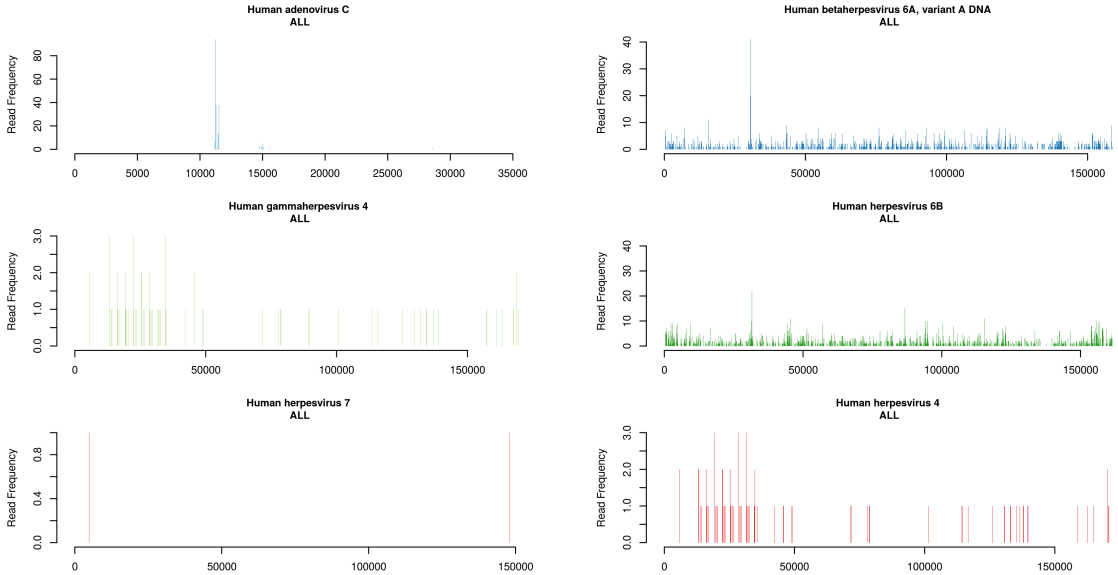

Figure 5: Genome coverage For The Danish Population.The y-axis can be interpreted as the depth of coverage/reads per window. The figure shows poor coverages for most of the selected viruses with the exception of HBH–6A/b, but the same excessive accumelation of reads around the EBN-2/LP genes for EBV/HH4

Viral Integration

Collectively, viral integration events were found in 376 samples out of 774 samples in two viruses only (excluding HH4).

Percentage of Integrated Hits Per Population								
Virus	GBR	CHB	CLM	IBS	JPT	LWK	PEL	YRI
EBV	0.04	0.14	0.15	0.13	0.06	0.09	0.085	0.05
Human Adenovirus C	9.94e-04	0.0	3.73e-04	0.0	4.5e-04	1.49e-04	6.75e-05	8.26e-04

The table shows that despite the prevalence of integration events among samples, the proportion of integrated reads per overall samples was quite low.

Other Viruses

Other non-human viruses found were found in agreement with the previous study including Moloney murine leukemia and archaeal viruses such as Geobacilus virus E3 which were assumed to be enviromental/cell contaminants, or the presence of which as part of the human microbiome is still under dispute.

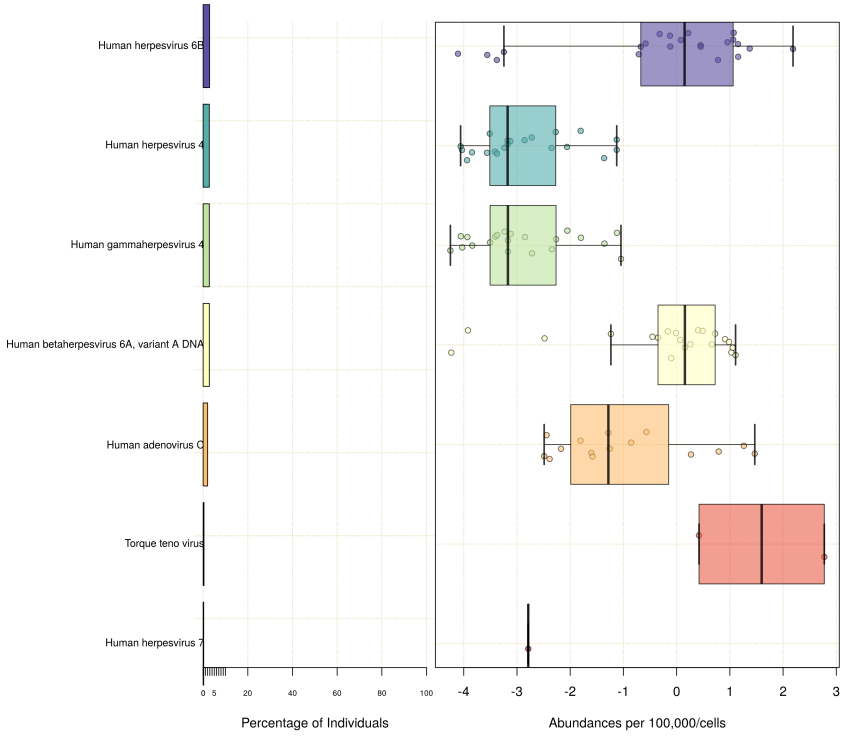

Figure 6: Prevalence And Viral Load In All 797 Samples.The figure shows prevalence of the 7 assigned human viruses along with their corresponding log-transformed viral load/abundance (as genome copies per 100,000/human cells). All viral prevalences are below 3%

Viral Load Distributions

In order to investigate geographical or ethnic differences in relation to viral loads for Adenovirus C and EBV, a Kruskal-Wallis test was performed and was found to be significant in both viruses (Adeno;p–value=0.0012, EBV;p–value=1.597e-09) indicating that atleast one population sample for either virus was drawn from a different distribution compared to the rest. A post-hoc Dunn test was performed with bonferroni correction on viral abundances (as shown in (Figure 7) with the results presented in Table 2.

Table 2: Pair-wise Dunn Test: An asterik implies signifcance post-bonferonni correction

Adenovirus C								EBV							
Adeno	CHB	CLM	GBR	IBS	JPT	LWK	PEL	EBV	CHB	CLM	GBR	IBS	JPT	LWK	PEL.
CLM	0.5554	–	–	–	–	–	–	CLM	1	–	–	–	–	–	–
GBR	0.1137	1	–	–	–	–	–	GBR	1	1	–	–	–	–	–
IBS	0.0487	1	1	–	–	–	–	IBS	0*	0*	0*	–	–	–	–
JPT	1	0.7033	0.1605	0.0670	–	–	–	JPT	1	1	1	0*	–	–	–
LWK	1	1	1	0.6642	1	–	–	LWK	0.3763	1	1	0.0005*	1	–	–
PEL	0.2505	0.0045*	0.0009*	0.0004*	0.1928	0.0319	–	PEL	1	1	1	0.0002*	1	1	–
YRI	1	1	1	0.6491	1	1	1	YRI	1	1	1	0*	1	1	1

In the case of Adenovirus C, there appeared to be no significant differences between the samples with the exception of the Peruvian sample which was significant when compared to the European

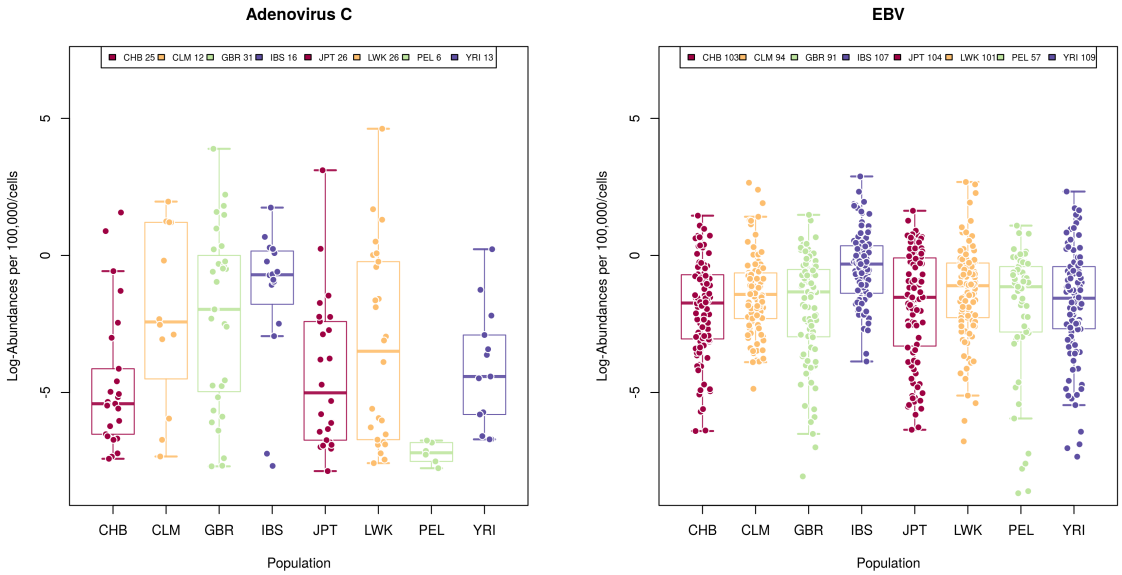

Figure 7: Viral Load Distributions Per Population For Adenovirus C and EBV: It may be noteworthy to mention however the in the case of Adenovirus C, the Peruvian samples appears to have a lower median and lower and dispersion in relation to all other samples. It also had the lowest number of samples. In the case of EBV, the Spanish sample had the highest median of all other groups.

samples and the Columbian sample. Conversely, there were significant pair-wise differences between the Spanish sample and all other samples in the case of EBV.

It becomes apparent that there is no continental/larger ethnic specific differences in terms of viral load for either virus. Furthermore, given the nature of the Kruskal-Wallis/Dunn test combined with the fact that the distributions of the sample were not ascertained or assumed to be identical, it is difficult to assert with confidence that the statistical differences signify either signficance differences in the medians of each group, or due to the sample groups having significantly different distribtuions. In other words, the p–values may either signify differences in the medians of the groups if the assumption of identical distributions holds, or more generally, that the samples of the groups were drawn from significanlty different distributions. A short commenatary on the medians and spread of the data is included on the figure caption (Figure 5).

GWAS

Principle Component Analysis & QQplots

Logistic and Linear assocication tests were performed on adenovirus C and EBV repsectively, while controling for population stratification using the top 10 principle components as covariates. The purpose of these association tests was to investiage a possible link between viral load and general population allelic variation at known snps.The PCA plot (Figure 8) of the top 2 eigenvectors shows clear seperation between non-mixed populations, and a sweeping range for admixed popu-

lation of Peru and Columbia. QQplots are also presented here for the validation of controling for population stratification using PCA (Figure 9)).

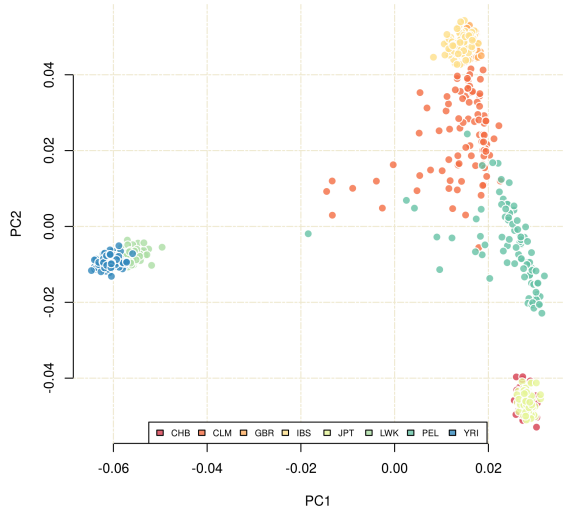

Figure 8: PCA for the 8 population: African, European and Asian populations nearly stack together, while the Peruvian/Columbian(Dark green/Orange) samples skew between European or Asian populations

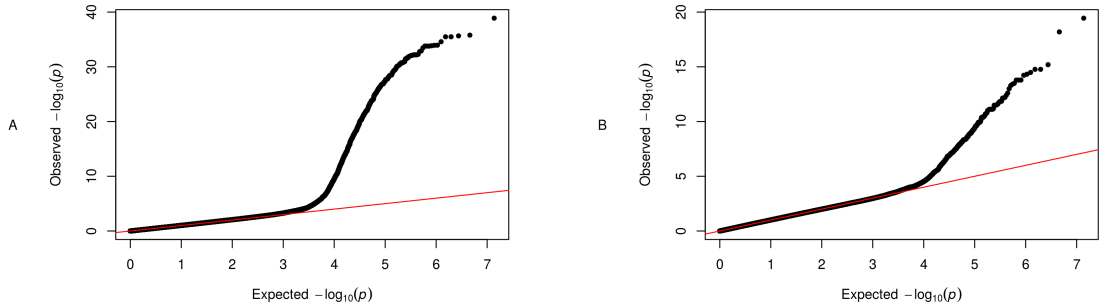

Figure 9: QQplots: The genomic inflation factors were found to be 1.03684 and 1.02208 for (A) EBV and (B) Adenovirus C respectively, indicating adequate correction for population stratification

Manhatten plots were generated for both association tests with log transformed p-values and a cutoff of log(5e–8)(Figure 10)). The number of snps showing positive association was found to be 923 for EBV, and 172 for adenovirus.

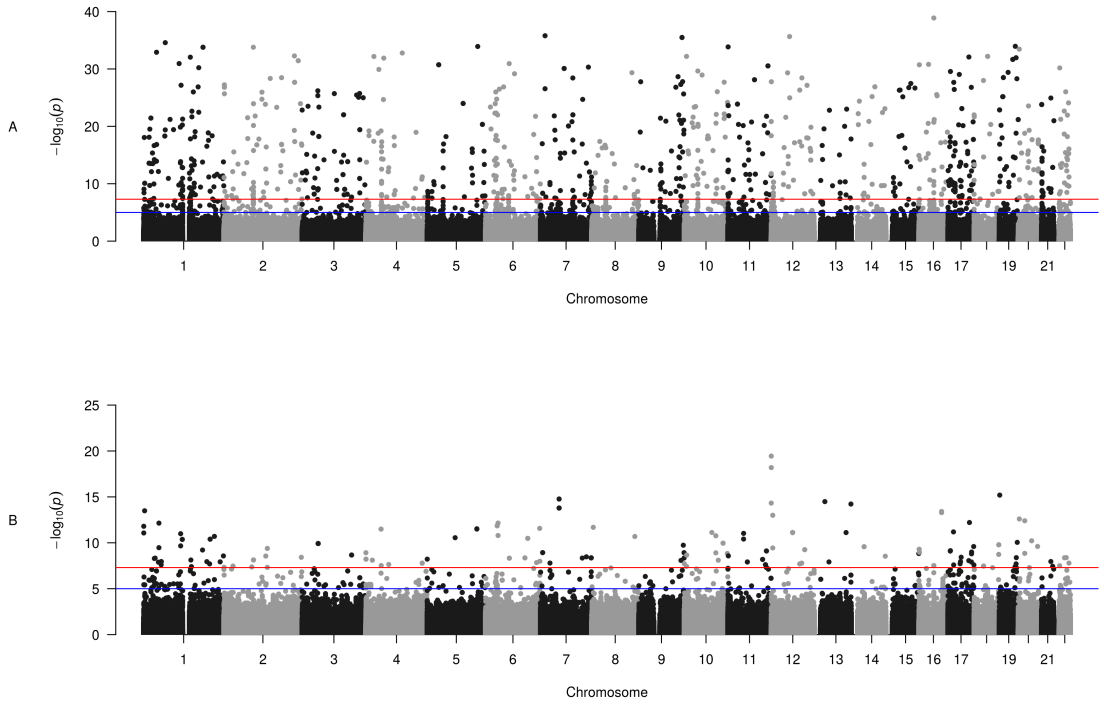

Figure 10: Manhatten plots: Log-transformed p–values were plotted by chromsome position in a so-called "Manhatten" plot with a bonferonni corrected genome-wide wide (red) threshold of significance at log(5e–08)

There appeared to be no discernable pattern of linkage-disequilibrium (LD) between the positively associated SNPs. Furthermore, in either EBV or Adenovirus C, none of the positively associated SNPs were found within the the human leukocyte antigen (HLA) region, excluding the possibility of direct association of any of the SNPs to a possible role in regulating the human immune system. Given the high number of SNPs with no clear pattern of LD, and no clear in regulating the immune system, it is assumed that most of the results are false positives likely stemming from the low sample size used in this study.

Integration Site Distribution

To further investiage the possible causes of false positives, one hypothesis is that false positives may have occured in the form of genome-wide associated SNPs in LD with the viral integration sites. The distribution of integration sites across chromosomes and distances of genome-wide associated SNPs to integration sites then had to be quantified. For EBV and Adenovirus C, the frequency of integration (Figure 11) appears roughly linear to the lengths of the chromosomes across the whole genome.

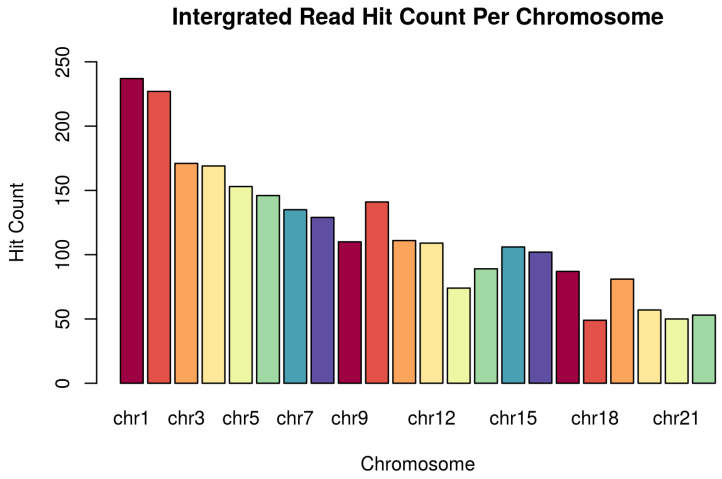

Figure 11: Frequency of Integrated Sites: The plot shows the frequency of integrated hits for EBV/Adenovirus across chromosomes. The number of hits is roughly linear to chromosome length

In the case of EBV, the distribtion of integration sites per chromosome (Figure 12) may appear random at first glance and not specific to any one particular region in the genome, but there is evidence that there may be preferences for CNVs or SVs.[65].

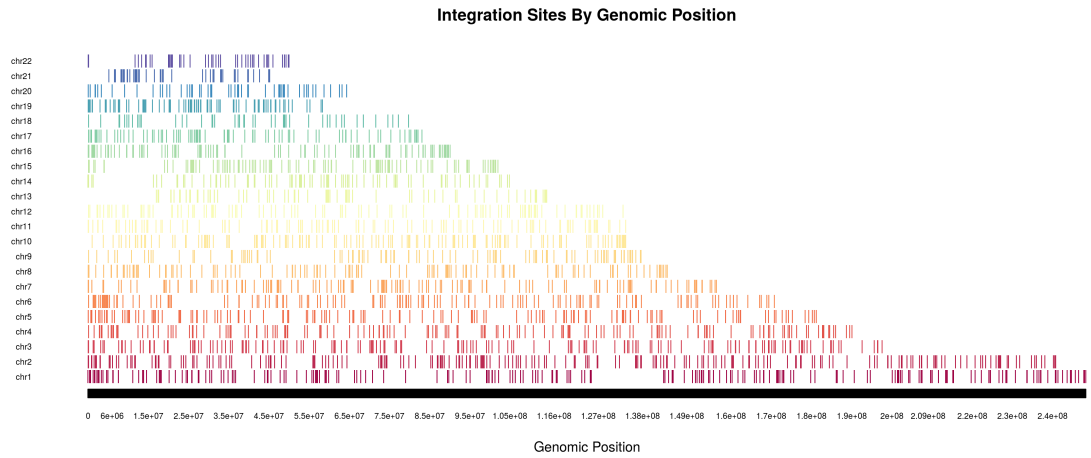

Figure 12: Distribution of Integrated Sites: The genomic positions for each chromosome is on the x-axis, and individual chromosomes are represented on the y-axis. The plot indicates where integrated sites were found with a vertical dash.

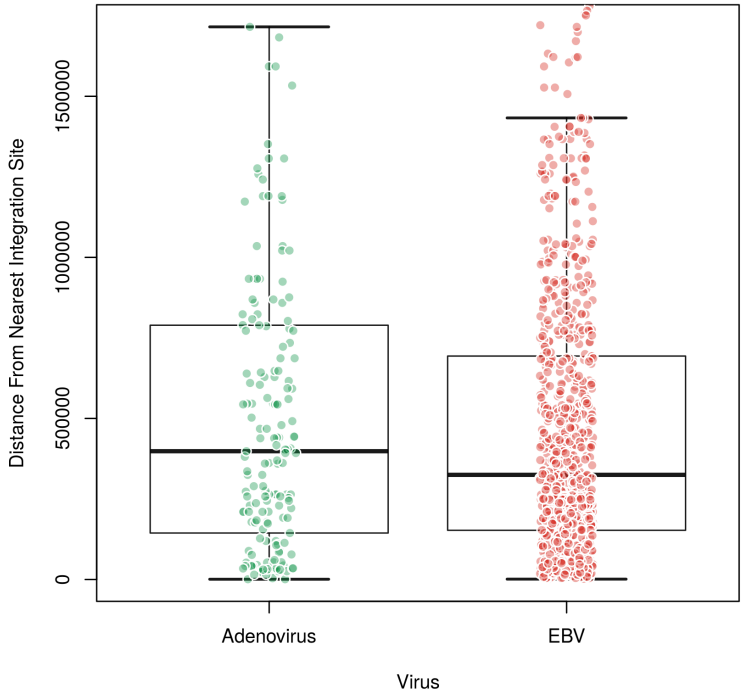

Figure 13: Boxplot of Distances Between SNPs and Integration Sites: Each point corresponds to a genome-wide positively associated SNP and its absolute distance to the nearest integration site

Distances between genome-wide associated SNPs and the nearest integration was also resolved(Figure 13). The plots carry a median distance of 398229 bps for EBV, and 324871 for adenovirus C. A previous study has shown that –over a sete of diverse ethnic populations– the proportion SNP pairs showing no evidence of recombination (D'1.0) vary nonlinearly by distance. For a distance of less than 5kb, the proportions were 76, 63 and 54% in Ashkenazi, Caucasian and African American samples respectively. At a distance of 50kb, the proportions vary from 35 to 30% [47], so there does lie the possibility that some of the genome-wide SNPs are in high enough LD with the integrated sites to be positively associated with viral load.

Discussion

The results of this study have highlighted several issues with the methods followed in the JCVI study when applied to low-coverage immortalized cell lines, and with the JCVI study itself. Firstly, there is a clear discrepancy between previous findings in terms of viral prevalence in both this study and the previous one in what could be claimed to be a "profile" of the healthy human virome. For instance, despite the high seroprevalence of Merkel Cell Polyomavirus and Human Polyomaviruses and its subtypes (ranging from 40% to 86%) in a healthy population[36], in this study and the JCVI study, its prevalence was determined at 0.25% and 0.63% respectively, and 0.55% and 0.19%. The same wide discrepancy is found in the case of Human betaherpesvirus 6 which has a seroprevalence of over 95% in developed countries, and ranges between 70 and 90% in ethnically diverse groups [9], while a prevalence of below 5% is reported in both studies. Without reconclining such wide differences in prevalence, generalizing the findings of the methods applied herein and elsewhere to the general population should be taken with caution. A number of factors could be implicated in this issue, one of which is the use of blastn as a heurestic method for assigning reads, another is the RefSeq viral database itself. The following two examples serve to illustrate these issues, but it is by no means exclusive to these select examples alone.

The EBV B95-8 strain is used in B-cell immortilization because of its nature as a non-integrating episomal chromosome. It carries a characteristic 12kb deletion not found in wild strains or other EBV types starting precisely from positions 139,724 to 151,554 [46]. The coverage for EBV in this study present several peculiarities; near matching coverages between HH4 and EBV including the 12kb deletion, small accumelation of reads roughly in the middle of the 12kb region, extreme accumelation of reads around the EBN2 gene region (critical for transformed B-cell growth and proliferation [20]), and evidence of viral integration concentrated around the same area.

Almost matching coverage between the two types could be explained by the ambiguity of assigning reads with multiple hits to viral genomes, an issue which was not addressed in the previous study despite the appearance of an uncharacteristic gap in the coverage plots of HSV1 and HHV7 around roughly the same area. The small accumelation of reads within the 12kb deletion (found in both studies) could be attributed to the presence of wild strain EBV, as well as the accumelation integrated reads around EBN2. To confound matters more, Macacine herpesvirus 4 was found in 70% of all samples, which presents a more serious illustration of the issue of ambigious hits assigned by blastn. It is worth mentioning that a study based on unmapped reads from roughly 1000 individuals from the 1000 genomes project found no evidence of co-infection between EBV and HH4, and was able to confirm the presence of wild-type EBV through the presence of uniquely mapped reads within the 12kb deletion region [46]. They also confirmed the presence of EBV in approximately 90% of the samples.

On the other hand, Human adenovirus C presents a different kind of problem when dealing with the non-specificity of blasn and the RefSeq database. There are 57 types of Human adenovirus spread across 7 species implicated in a wide range of diseases with a complementary range of symptoms and severity [43]; the most prevalent of which is respiratory infections. Human adenovirus

type 3 alone has 248 confirmed strains which have been shown to be highly heterogenuous and rapidly mutating [14]. The RefSeq viral database contains viral genome on the level of types or species only, and not on a strain level. With this background in mind, we can proceed forward to clearly highlight the issue. Firstly, adenovirus was found in only 9% of the previous study, so the focus will be on the results presented here. Despite the high read counts and relatively high prevalence, the coverage plot for adenovirus C failed to demonstrate full genome coverage. This could purely be attributed to vector contaminants, as adenovirus is frequently used in a wide array of expression and cloning vectors. But it could also to a failure to correctly assign true viral reads to some highly heterogenuous strain of adenovirus, while only retaining reads of conserved regions present in the Adenovirus C RefSeq genome.

In adressing these concerns, it is assumed that the researchers at JCVI attempted to assemble viral contigs for that purpose. They managed to assemble contigs for 8 viruses in 137 samples only, as opposed to 19 viruses in 3,342 with read mapping. Given the requirement of high coverage necessary for viral contig assembly, it would have been impossible to attempt contig assembly using low coverage data.

In short, blastn's inherent method for homology search is an advantage for searching against highly divergant and rapdily mutating viruses, but it can lead to situations such as the ones listed above. Ultimately, metagenomic approach may be the better route to take. A metagenomic tool for assigning reads based on 31-kmer exact matches was shown to have higher sensitivity and precision than blastn on the condition that the database was built from GenBank genomes, and not RefSeq [62]. However, the authors of the tool recommended against the use of GenBank draft genomes due to the presence of contaminants. However, a new method was devised (as recent as June 25th, 2018) for removing contaminants and low-complexity regions from draft genomes stored on GenBank[29]. In theory, applying the method of effectively purging draft genomes of comtaiminants combined with a metagenomic read assignment could not only lead to higher sensitivy/precision without sacrificing blastn's unique advantage, but could also unlock more available viral genomes, types and strains stored on GenBank.

Alternatively, future researchers could instead restrict their analysis to viral hits on the species level. However, given the wide ranging pathology of virus types/strains, and the utlity of this type of reasearch in ascertaining blood transfusion safety, the outcome of such a study would be severly handicappned and underutilized. As a positive example from this study, the coverage plots for Human herpesvirus 6A and 6B (Figure 4) could be taken as convincing evidence of the presence of either or both strains. Addtional empirial testing such as a digital droplet PCR assay could then be performed to properly validate and quanity the presence of either strain [25].

Finally, one would not expect to find viral sequences in whole exon sequence data. The presence of viruses in unmapped reads could lead us to atleast three interpretation within the real of conjecture; the viruses are integrated within exon regions, they are homologous to regions in the human genome and not actual full viruses, or they are simply lab contaminents. In terms of homologous sequences, 100 random reads were selected and searched against the USCS chimp and human (HG38) genomes using BLAT and were found to have aligned sequence lengths of around 20, which might

exclude the possiblity of homology. In retrospect, a more thorough determination into human-viral homology might have been more fruitful as a quality control step.

The reads could also be viruses that were not properly washed away during array hybridization. What the findings do confirm with some certainty is that the over-representation of reads aligned to the EBN-2 genes in EBV or HH4 is not an issue exclusive to the use of whole-genome sequences from LCLs or to a particular ethnic group. Furthermore, the issue of ambigious read assignment persists in either case.

We also performed genome-wide association on adenovirus C and EBV to find associated loci to viral load. The sheer number of associations with no clear patterns of LD, combined with the absence of association of SNPs falling within the HLA region, was enough to indicate a high prevalence of false positives.In terms of sample sizes, it was previously estimated that the minimum required sample size for detecting association with 80% statistical power varied as a function of certain effect sizes [57] (beta; as measured in standard deviation units) and MAFs. For beta of 1, and MAF of 0.1, one would need a thousand individuals. For a beta of 0.1, double the order of magnitude . By these constraints, this study is underpowered. Furthermore, given the ubiquity of viral integration events for adenovirus C and EBV across the human genome, it is an open possiblity that the contribution of integrated reads to viral abundances positively correlates with markers in LD to nearby integration sites, increasing the chance of false positives. One possible solution is to determine LD between nearby markers to each integration site and exclude them.

In conclusion, rather than reaching concrete findings, the results of this study have instead provided the opportunity to highlight critical issues with the methods followed in both studies. From the outlook of species level comparisons, there appeared to be viral infections of varying degrees in most samples. Humans herpesvirus 6 provided the strongest evidence out of all other viruses, but there is also an obvious lack of agreement in reported population-level prevalences in previous findings with both this study and the JCVI study. There is a great wealth of publically available data of unmapped reads that is yet to be mined and exploited to its full usefuleness. In order to do so, issue of contamination and ambigious read assignment need to be accounted for. The need for precision is even more greatly emphasized given the recent evidence of possible links between viral infections and chronic diseases such as diabetes [2]. Furthermore, to guarantee the success of a genome-wide association study, viral integration also needs to be corrected for beforehand. In short, capturing a profile of the human virome from unmapped reads is within reach and there lay many a great possiblities on where it could lead to.

Supplementary Data

Full BWA Results

Below are the figures and tables for the full BWA results on all 797 samples included in this study

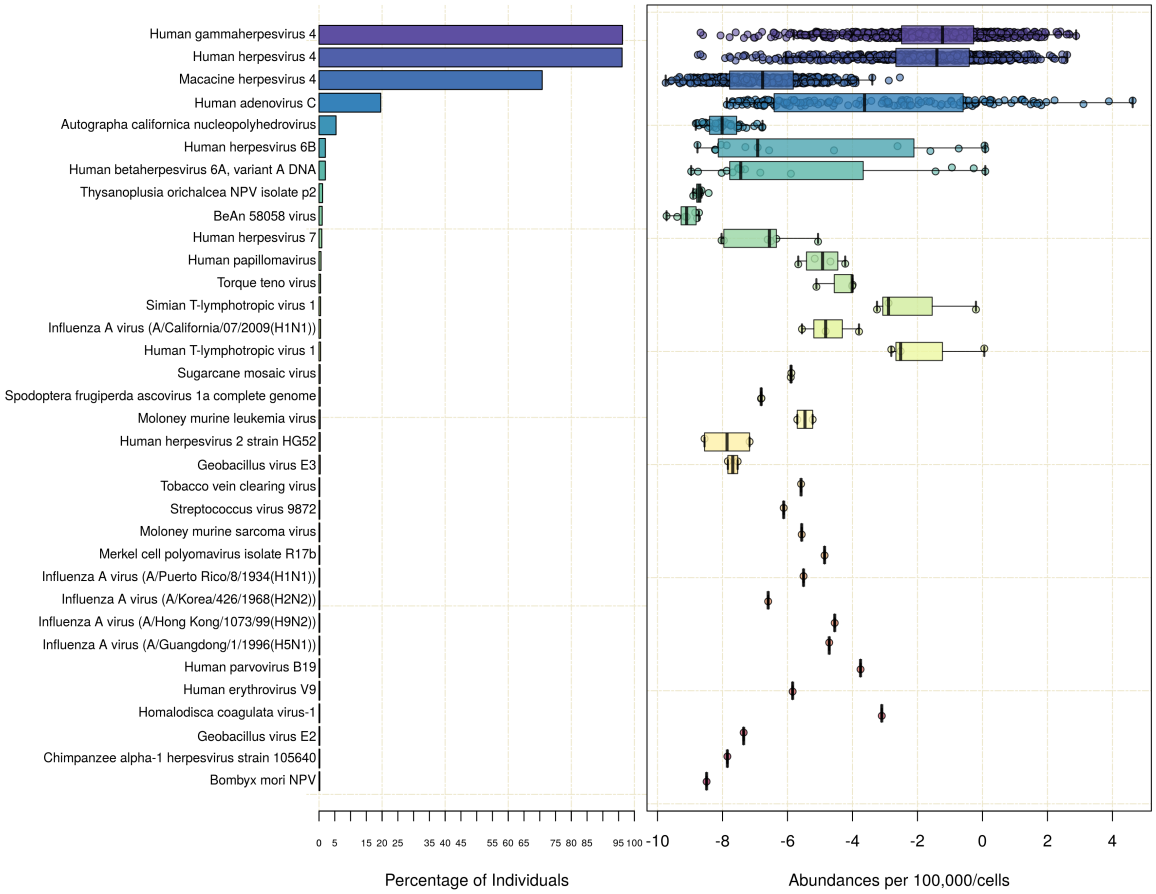

Figure 14: The figure shows estimated viral abundances side by side for the blastn derived results and bwa. The abundances are nearly identical in both cases.

Table 3: Included are 34 viral hits including several strains of Influenza A

BWA on GBR			Number of Reads		Viral Abundances		Depth of Coverage	
Virus	Number of Individuals	Percentage of Individuals	Median	Maximum	Median–index5	Maximum	Minimum	Maximum
Human gammaherpesvirus 4	767	96.24	1539	3679906	0.29	17.854	0.001	83.452
Human herpesvirus 4	766	96.11	1318	3071683	0.25	13.533	0	67.076
Macacine herpesvirus 4	564	70.77	3	9167	0	0.079	0	0.267
Human adenovirus C	156	19.57	0	232368	0.03	101.664	0.002	138.825
Human endogenous retrovirus K113 complete genome	102	12.8	0	147	0	0.018	0.008	0.058
Autographa californica nucleopolyhedrovirus	43	5.4	0	90	0	0.001	0.001	0.005
Human betaherpesvirus 6A, variant A DNA	16	2.01	0	17162	0	1.091	0	4.597
Human herpesvirus 6B	16	2.01	0	22682	0	1.09	0	6.026
Thysanoplusia orichalcea NPV isolate p2	9	1.13	0	8	0	0	0.001	0.001
BeAn 58058 virus	8	1	0	7	0	0	0.001	0.001
Human herpesvirus 7	7	0.88	0	55	0	0.006	0.001	0.017
Human papillomavirus	5	0.63	0	7	0.01	0.015	0.01	0.027
Human T-lymphotropic virus 1	4	0.5	0	252	0.08	1.057	0.197	2.389
Influenza A virus (A/California/07/2009(H1N1))	4	0.5	0	20	0.01	0.022	0.011	0.126
Simian T-lymphotropic virus 1	4	0.5	0	202	0.06	0.818	0.125	1.816
Torque teno virus	4	0.5	0	5	0.02	0.018	0.035	0.111
Geobacillus virus E3	3	0.38	0	4	0	0.001	0.001	0.001
Human herpesvirus 2 strain HG52	3	0.38	0	5	0	0.001	0.001	0.002
Moloney murine leukemia virus	3	0.38	0	4	0	0.005	0.018	0.018
Spodoptera frugiperda ascovirus 1a complete genome	3	0.38	0	15	0	0.001	0.004	0.005
Sugarcane mosaic virus	3	0.38	0	2	0	0.003	0.008	0.008
Bombyx mori NPV	2	0.25	0	1	0	0	0.001	0.001
Chimpanzee alpha-1 herpesvirus strain 105640	2	0.25	0	2	0	0	0.001	0.001
Geobacillus virus E2	2	0.25	0	1	0	0.001	0.002	0.002
Homalodisca coagulata virus-1	2	0.25	0	12	0.05	0.045	0.128	0.128
Human erythrovirus V9	2	0.25	0	1	0	0.003	0.019	0.019
Human parvovirus B19	2	0.25	0	9	0.02	0.023	0.154	0.154
Influenza A virus (A/Guangdong/1/1996(H5N1))	2	0.25	0	4	0.01	0.009	0.047	0.047
Influenza A virus (A/Hong Kong/1073/99(H9N2))	2	0.25	0	2	0.01	0.011	0.053	0.053
Influenza A virus (A/Korea/426/1968(H2N2))	2	0.25	0	1	0	0.001	0.007	0.007
Influenza A virus (A/Puerto Rico/8/1934(H1N1))	2	0.25	0	3	0	0.004	0.02	0.02
Merkel cell polyomavirus isolate R17b	2	0.25	0	2	0.01	0.008	0.028	0.028
Moloney murine sarcoma virus	2	0.25	0	1	0	0.004	0.013	0.013
RD114 retrovirus	2	0.25	0	1	0	0.003	0.01	0.01
Squirrel monkey retrovirus - HLB	2	0.25	0	2	0.01	0.005	0.017	0.017
Streptococcus virus 9872	2	0.25	0	2	0	0.002	0.006	0.006
Tobacco vein clearing virus	2	0.25	0	1	0	0.004	0.011	0.011

BLASTN/BWA–mem Approximation

Below are the figures and tables for the comparative blastn/bwa runs on all 106 GBR samples included in this study.

Table 4: Results for one run on all GBR samples with blastn

BWA on GBR			Number of Reads		Viral Abundances		Depth of Coverage	
Virus	Number of Individuals	Percentage of Individuals	Median	Maximum	Minimum	Maximum	Minimum	Maximum
Human gammaherpesvirus 4	92	98.92	1777	280065	0.26	4.396	0.001	11.829
Human herpesvirus 4	92	98.92	1352	238477	0.23	3.767	0	10.727
Macacine herpesvirus 4	61	65.59	2	646	0	0.019	0	0.041
Human adenovirus C	32	34.41	0	74327	0.14	48.909	0.002	101.88
Human endogenous retrovirus K113 complete genome	16	17.2	0	28	0	0.01	0.008	0.032
Human betaherpesvirus 6A, variant A DNA	7	7.53	0	1733	0	0.236	0	0.805
Human herpesvirus 6B	6	6.45	0	8298	0	1.065	0	3.679
Autographa californica nucleopolyhedrovirus	5	5.38	0	11	0	0.001	0.001	0.003
Phage cdtI DNA	5	5.38	0	15	0	0.005	0.002	0.02
Enterobacteria phage mEp460	3	3.23	0	3	0	0.001	0.002	0.005
Influenza A virus (A/California/07/2009(H1N1))	3	3.23	0	4	0.01	0.008	0.011	0.015
Thysanoplusia orichalcea NPV isolate p2	3	3.23	0	2	0	0	0.001	0.001
BeAn 58058 virus	2	2.15	0	1	0	0	0.001	0.001
Enterobacteria phage HK629	2	2.15	0	1	0	0.001	0.002	0.002
Enterobacteria phage HK630	2	2.15	0	1	0	0.001	0.002	0.002
Enterobacteria phage SfV	2	2.15	0	1	0	0.001	0.003	0.003
Enterobacteria phage ST104 DNA	2	2.15	0	1	0	0.001	0.003	0.003
Enterobacteria phage ST64T	2	2.15	0	1	0	0.001	0.003	0.003
Phage Gifsy-1	2	2.15	0	1	0	0	0.002	0.002
Phage Gifsy-2	2	2.15	0	2	0	0.001	0.005	0.005
Salmonella phage c341	2	2.15	0	1	0	0.001	0.003	0.003
Salmonella phage epsilon34	2	2.15	0	1	0	0.001	0.003	0.003
Shigella phage SfII	2	2.15	0	1	0	0.001	0.003	0.003

Table 5: Results for one run on all GBR samples with BWA

BWA on GBR			Number of Reads		Viral Abundances		Depth of Coverage	
Virus	Number of Individuals	Percentage of Individuals	Median	Maximum	Median–index5	Maximum	Minimum	Maximum
Human gammaherpesvirus 4	93	100	1448	231943	0.213	3.409	0.001	10.982
Human herpesvirus 4	93	100	950	186648	0.143	2.72	0	10.054
Macacine herpesvirus 4	53	56.99	1	1130	0.001	0.027	0	0.125
Human adenovirus C	35	37.63	0	144881	0.204	63.801	0.004	130.802
Human endogenous retrovirus K113 complete genome	11	11.83	0	17	0.004	0.006	0.008	0.024
Human betaherpesvirus 6A, variant A DNA	7	7.53	0	1062	0.001	0.143	0	0.49
Autographa californica nucleopolyhedrovirus	6	6.45	0	26	0.001	0.003	0.001	0.007
Bombyx mori NPV	6	6.45	0	13	0	0.002	0.001	0.005
Human herpesvirus 6B	6	6.45	0	7719	0.002	0.99	0	3.422
Phage Gifsy–1	6	6.45	0	98	0.002	0.044	0.002	0.198
Phage cdtI DNA	5	5.38	0	21	0.001	0.009	0.002	0.038
Enterobacteria phage mEp460	4	4.3	0	12	0.002	0.004	0.005	0.019
Phage Gifsy–2	4	4.3	0	51	0.002	0.025	0.002	0.11
Influenza A virus (A/California/07/2009(H1N1))	3	3.23	0	3	0.004	0.004	0.006	0.015
Salmonella phage Fels–1	3	3.23	0	4	0.001	0.002	0.002	0.007
Salmonella phage SEN34	3	3.23	0	5	0.001	0.002	0.002	0.01
Salmonella phage SPN3UB	3	3.23	0	9	0.002	0.004	0.002	0.018
Abelson murine leukemia virus	2	2.15	0	3	0.014	0.014	0.039	0.039
BeAn 58058 virus	2	2.15	0	1	0	0	0.001	0.001
Enterobacteria phage P22 virus	2	2.15	0	1	0.001	0.001	0.003	0.003
Enterobacteria phage SfV	2	2.15	0	10	0.006	0.006	0.029	0.029
Enterobacteria phage ST104 DNA	2	2.15	0	2	0.001	0.001	0.005	0.005
Enterobacteria phage ST64T	2	2.15	0	2	0.001	0.001	0.005	0.005
Enterobacteria phage UAB_Phi20	2	2.15	0	1	0.001	0.001	0.003	0.003
Maruca vitrata MNPV	2	2.15	0	1	0	0	0.001	0.001
Moloney murine leukemia virus	2	2.15	0	3	0.01	0.01	0.027	0.027
Moloney murine sarcoma virus	2	2.15	0	1	0.005	0.005	0.013	0.013
Salmonella enterica bacteriophage SE1	2	2.15	0	2	0.001	0.001	0.005	0.005
Salmonella Phage 103203_sal5	2	2.15	0	1	0.001	0.001	0.003	0.003
Salmonella phage 118970_sal4	2	2.15	0	1	0.001	0.001	0.002	0.002
Salmonella phage c341	2	2.15	0	1	0.001	0.001	0.003	0.003
Salmonella phage epsilon34	2	2.15	0	1	0.001	0.001	0.003	0.003
Salmonella phage SEN1	2	2.15	0	1	0.002	0.002	0.003	0.003
Salmonella phage SPN9CC	2	2.15	0	1	0.001	0.001	0.003	0.003
Salmonella phage ST160	2	2.15	0	2	0.001	0.001	0.005	0.005
Shigella phage SfII	2	2.15	0	8	0.005	0.005	0.02	0.02
Shigella phage SfIV	2	2.15	0	10	0.006	0.006	0.026	0.026
Thysanoplusia orichalcea NPV isolate p2	2	2.15	0	1	0	0	0.001	0.001

Full Putative Viruses

Table 6: All putative viruses found in 797 individuals included in this study searched against the RefSeq viral genomes using blastn

Blast on All Samples With Viral Database			Number of Reads	
Virus	Number of Individuals	Percentage of Individuals	Median Reads per Individual	Maximum Reads per Individual
Human endogenous retrovirus K113 complete genome	693	86.62	249	382240
Human gammaherpesvirus 4	665	83.12	1424	3745843
Human herpesvirus 4	665	83.12	1200	3227899
BeAn 58058 virus	611	76.38	123	216639
Macacine herpesvirus 4	567	70.88	19	59749
Enterobacteria phage phiX174 sensu lato	369	46.12	0	24550669
Enterobacteria phage WA13	197	24.62	0	619883
Enterobacteria phage lambda	196	24.5	0	81221
Enterobacteria phage HK630	175	21.88	0	56403
Enterobacteria phage ID2 Moscow/ID/2001	162	20.25	0	309494
Enterobacteria phage alpha3	160	20	0	283302
Enterobacteria phage ID18	160	20	0	340760
Enterobacteria phage HK629	159	19.88	0	44814
Bacteriophage phiK complete genome	155	19.38	0	408205
Enterobacteria phage St-1	154	19.25	0	414655
Human adenovirus C	149	18.62	0	465037
Enterobacteria phage G4	120	15	0	173846
Enterobacteria phage T4	104	13	0	1560
Shigella phage SHFML-11	104	13	0	1319
Enterobacteria phage RB14	103	12.88	0	1195
Enterobacteria phage RB27	103	12.88	0	1209
Enterobacteria phage RB3	103	12.88	0	1366
Escherichia phage e11/2	103	12.88	0	1313
Escherichia phage ECML-134	103	12.88	0	1538
Escherichia phage UFV-AREG1	103	12.88	0	1353
Escherichia phage wV7	103	12.88	0	1150
Shigella phage SHFML-26	103	12.88	0	1215
Bacteriophage RB32	102	12.75	0	1415
Enterobacteria phage AR1 DNA	102	12.75	0	1100
Enterobacteria phage RB51	102	12.75	0	1123
Enterobacteria phage RB68	102	12.75	0	1123
Enterobacteria phage vB_EcoM_ACG-C40	102	12.75	0	1321
Escherichia phage slur14	102	12.75	0	1352
Shigella phage pSs-1	102	12.75	0	1187
Shigella phage SHBML-50-1	102	12.75	0	1110
Yersinia phage phiD1 complete genome	102	12.75	0	1314
Yersinia phage PST	102	12.75	0	1263
Enterobacteria phage imeo9	101	12.62	0	1293
Escherichia phage HY03	101	12.62	0	1119
Escherichia phage sluro2	101	12.62	0	1208
Escherichia phage sluro7	101	12.62	0	1325
Escherichia phage vB_EcoM-UFV13	101	12.62	0	1214
Shigella phage Shfl2	99	12.38	0	1177
Escherichia phage HY01	98	12.25	0	1141
Enterobacteria phage M13	97	12.12	0	432
Stx2-converting phage 1717	87	10.88	0	6848
Enterobacteria phage BP-4795	86	10.75	0	6870
Enterobacteria phage P1	86	10.75	0	265
Enterobacteria phage HK106	83	10.38	0	18088
Enterobacteria phage fd strain 478	82	10.25	0	332
Enterobacteria phage SfI	82	10.25	0	24828
Enterobacteria phage Sf6	81	10.12	0	203
Shigella phage SfIV	80	10	0	277
Enterobacteria phage HK97	79	9.88	0	18022
Phage cdtI DNA	77	9.62	0	22638
Salmonella phage SJ46	75	9.38	0	162
Stx2 converting phage II DNA	74	9.25	0	234
Stx2 converting phage vB_EcoP_24B	67	8.38	0	304
Enterobacteria phage YYZ-2008	65	8.12	0	6746
Propionibacterium phage PHL150M00	65	8.12	0	203
Propionibacterium phage P104A	63	7.88	0	188
Propionibacterium phage PHL141N00	63	7.88	0	194
Enterobacterial phage mEp234	62	7.75	0	17454
Propionibacterium phage PAC1	62	7.75	0	200

Propionibacterium phage PHL071N05	62	7.75	0	175
Propionibacterium phage PHL092M00	62	7.75	0	184
Propionibacterium phage PHL132N00	62	7.75	0	193
Shigella phage SfII	62	7.75	0	160
Propionibacterium phage P1.1	61	7.62	0	215
Propionibacterium phage P100D	61	7.62	0	192
Propionibacterium phage PA1-14	61	7.62	0	183
Propionibacterium phage PHL111M01	61	7.62	0	182
Enterobacteria phage HK140	60	7.5	0	11420
Enterobacteria phage HK542	60	7.5	0	35845
Enterobacteria phage HK544	60	7.5	0	35723
Propionibacterium phage Enoki	60	7.5	0	184
Propionibacterium phage P14.4	60	7.5	0	178
Propionibacterium phage PHL113M01	60	7.5	0	188
Propionibacterium phage Wizz0	60	7.5	0	181
Enterobacteria phage 933W	59	7.38	0	221
Enterobacteria phage mEp460	59	7.38	0	22155
Enterobacteria phage VT2phi_272	59	7.38	0	162
Propionibacterium phage Moyashi	59	7.38	0	166
Propionibacterium phage PHL055N00	59	7.38	0	200
Propionibacterium phage PHL082M00	59	7.38	0	186
Propionibacterium phage Attacne	58	7.25	0	173
Propionibacterium phage P9.1	58	7.25	0	186
Propionibacterium phage PHL060L00	58	7.25	0	176
Propionibacterium phage PHL085N00	58	7.25	0	151
Propionibacterium phage PHL095N00	58	7.25	0	190
Propionibacterium phage Procrass1	58	7.25	0	191
Propionibacterium phage Solid	58	7.25	0	185
Propionibacterium phage PA6	57	7.12	0	174
Propionibacterium phage PHL030N00	57	7.12	0	164
Propionibacterium phage PHL067M10	57	7.12	0	179
Propionibacterium phage PHL116M00	57	7.12	0	159
Propionibacterium phage PHL179M00	57	7.12	0	194
Propionibacterium phage SKKY	57	7.12	0	159
Bacteriophage HK022	56	7	0	11418
Propionibacterium phage P105	56	7	0	184
Propionibacterium phage PHL025M00	56	7	0	188
Autographa californica nucleopolyhedrovirus	55	6.88	0	206
Enterobacteria phage HK446	55	6.88	0	11394
Propionibacterium phage MrAK	55	6.88	0	177
Propionibacterium phage PHL009M11	55	6.88	0	170
Propionibacterium phage PHL171M01	55	6.88	0	136
Enterobacteria phage Sf101	54	6.75	0	11392
Propionibacterium phage Kubed	54	6.75	0	142
Propionibacterium phage PAD20 endogenous virus	54	6.75	0	189
Propionibacterium phage PAS50 endogenous virus	54	6.75	0	181
Propionibacterium phage PHL112N00	54	6.75	0	185
Propionibacterium phage P101A	53	6.62	0	189
Salmonella phage SEN22	53	6.62	0	229
Enterobacteria phage IME10	52	6.5	0	11420
Propionibacterium phage P100_1	52	6.5	0	171
Propionibacterium phage PHL070N00	52	6.5	0	145
Propionibacterium phage Pirate	52	6.5	0	171
Escherichia phage TL-2011b	50	6.25	0	152
Propionibacterium phage ATCC29399B_C	50	6.25	0	162
Propionibacterium phage PHL152M00	50	6.25	0	121
Propionibacterium phage Stormborn	50	6.25	0	159
Enterobacteria phage HK633	49	6.12	0	11363
Propionibacterium phage Ouroboros	49	6.12	0	118
Enterobacteria phage mEp235	48	6	0	11449
Escherichia phage PA2	48	6	0	126
Propionibacterium phage PHL041M10	48	6	0	153
Propionibacterium phage PHL114L00	48	6	0	115
Stx1 converting phage DNA	48	6	0	163
Enterobacterial phage mEp213	46	5.75	0	1782
Propionibacterium phage BruceLethal	46	5.75	0	113
Propionibacterium phage Lauchelly	46	5.75	0	129
Enterobacteria phage HK620	44	5.5	0	11338
Enterobacteria phage mEp043 c-1	44	5.5	0	1761
Propionibacterium phage ATCC29399B_T	44	5.5	0	100
Propionibacterium phage Pacnes 2012-15	43	5.38	0	112
Propionibacterium phage PHL010M04	43	5.38	0	101
Propionibacterium phage PHL301M00	43	5.38	0	127
Propionibacterium phage QueenBey	43	5.38	0	118
Enterobacteria phage mEpX2	42	5.25	0	11401
Enterobacteria phage VT2-Sakai	42	5.25	0	147
Bombyx mori NPV	39	4.88	0	105
Enterobacteria phage Min27	39	4.88	0	173
Propionibacterium phage P100_A	39	4.88	0	112
Salmonella phage c341	39	4.88	0	1769

Salmonella phage epsilon34	39	4.88	0	1769
Enterobacteria phage mEpX1	38	4.75	0	11336
Propionibacterium phage PHL199M00	38	4.75	0	113
Stx2-converting phage 86	36	4.5	0	149
Enterobacteria phage P22 virus	32	4	0	1705
Enterobacteria phage T7	32	4	0	65
Enterobacteria phage UAB_Phi20	32	4	0	1705
Yersinia pestis phage phiA1122	32	4	0	59
Enterobacteria phage 13a	31	3.88	0	62
Enterobacteria phage If1	31	3.88	0	58
Escherichia phage CICC 80001	31	3.88	0	59
Salmonella phage vB_SemP_Emek	30	3.75	0	1770
Staphylococcus phage SPbeta-like	27	3.38	0	417
Enterobacteria phage phi80	25	3.12	0	40
Shamonda virus N and NSs genes, segment S, genomic RNA, isolate Ib An 55550	24	3	0	30
Simian adenovirus 49	23	2.88	0	91
Enterobacteria phage ES18	21	2.62	0	34
Escherichia phage 64795_ec1	21	2.62	0	34
Salmonella phage SPN9CC	21	2.62	0	40
Shigella phage 75/02 Stx	21	2.62	0	49
Enterobacteria phage SfV	20	2.5	0	57
Phage Gifsy-1	20	2.5	0	247
Salmonella phage 118970_sal4	20	2.5	0	37
Salmonella phage vB_SosS_Oslo	20	2.5	0	49
Enterobacteria phage mEp237	19	2.38	0	29
Enterobacteria phage ST104 DNA	17	2.12	0	43
Escherichia phage HK75	17	2.12	0	37
Human adenovirus 54	17	2.12	0	13008
Human adenovirus D	17	2.12	0	12315
Salmonella Phage 103203_sal5	17	2.12	0	32
Salmonella phage 118970_sal3	17	2.12	0	395
Escherichia phage TL-2011c	16	2	0	35
Shigella phage POCJ13	16	2	0	37
Stenotrophomonas phage IME15	16	2	0	18
Chimpanzee adenovirus Y25	15	1.88	0	5466
Clostridium phage phiCD505	15	1.88	0	216
Clostridium phage phiMMP02	15	1.88	0	269
Enterobacteria phage vB_EcoM_VR5	15	1.88	0	30
Human adenovirus E	15	1.88	0	5446
Clostridium phage CDMH1 complete genome	14	1.75	0	367
Clostridium phage phi CD119	14	1.75	0	154
Clostridium phage phiC2	14	1.75	0	525
Clostridium phage phiCD211	14	1.75	0	121
Clostridium phage phiCD27	14	1.75	0	332
Clostridium phage phiCDHM19	14	1.75	0	271
Clostridium phage phiMMP03	14	1.75	0	521
Human betaherpesvirus 6A, variant A DNA	14	1.75	0	21783
Simian adenovirus 16 strain C-8	14	1.75	0	152
Thysanoplusia orichalcea NPV isolate p2	14	1.75	0	15
Clostridium phage phiMMP01	13	1.62	0	439
Enterobacteria phage HK225	13	1.62	0	29
Escherichia phage vB_EcoM_JS09	13	1.62	0	14
Escherichia phage vB_EcoM_PhAPEC2	13	1.62	0	14
Human herpesvirus 6B	13	1.62	0	25082
Shigella phage Shf125875	13	1.62	0	14
Shigella phage SHSML-52-1	13	1.62	0	14
Salmonella enterica bacteriophage SE1	12	1.5	0	45
Salmonella phage STML-198	12	1.5	0	12
Salmonella phage STP4-a	12	1.5	0	12
Abelson murine leukemia virus	11	1.38	0	17
Bacillus phage SPBc2	11	1.38	0	1211
Clostridium phage phiCD111	11	1.38	0	45
Clostridium phage phiCD146	11	1.38	0	38
Clostridium phage phiCD38-2	11	1.38	0	37
Clostridium phage phiCD506	11	1.38	0	50
Clostridium phage phiCDHM11 complete genome	11	1.38	0	88
Clostridium phage phiCDHM13 complete genome	11	1.38	0	88
Clostridium phage phiMMP04	11	1.38	0	80
Escherichia phage Av-05	11	1.38	0	13
Lactococcus phage BM13	11	1.38	0	25
Maruca vitrata MNPV	11	1.38	0	12
Phage Gifsy-2	11	1.38	0	96
Salmonella phage ST160	11	1.38	0	42
Salmonella phage ST64B	11	1.38	0	223
Cynomolgus adenovirus 1 isolate UK/UK-1/2004	10	1.25	0	18
Enterobacteria phage ST64T	10	1.25	0	40
Mus musculus mobilized endogenous polytropic provirus clone 15	10	1.25	0	111
Shigella phage Ss-VASD	10	1.25	0	29
Staphylococcus phage phiBU01	10	1.25	0	16

Staphylococcus prophage phiN315	10	1.25	0	24
Clostridium phage phiCD6356	9	1.12	0	21
Lactococcus Phage ASCC191	9	1.12	0	14
Lactococcus phage bIL170	9	1.12	0	16
Lactococcus phage SL4	9	1.12	0	15
Ralstonia phage RSS30 DNA	9	1.12	0	12
Salmonella phage RE-2010	9	1.12	0	18
Simian adenovirus 8 strain P-5	9	1.12	0	10
Staphylococcus phage 23MRA	9	1.12	0	22
Staphylococcus phage StB20-like	9	1.12	0	14
Bacteriophage phi-105 DNA	8	1	0	12
Clostridium phage phiCD481-1	8	1	0	33
Enterobacteria phage HX01	8	1	0	7
Enterobacteria phage RB69	8	1	0	7
Escherichia phage P13374	8	1	0	29
Escherichia phage phi191	8	1	0	28
Human adenovirus B2	8	1	0	75
Lactococcus phage 340	8	1	0	13
Lactococcus phage P680	8	1	0	14
Murine type C retrovirus	8	1	0	40
Staphylococcus phage phiNM3	8	1	0	18
Staphylococcus phage StauST398-4	8	1	0	19
Bacteriophage r1t integrase, repressor protein (rro), dUTPase, holin and lysin genes	7	0.88	0	14
Enterobacteria phage Fels-2	7	0.88	0	11
Human adenovirus B1	7	0.88	0	60
Human herpesvirus 7	7	0.88	0	55
Lactococcus phage 50101	7	0.88	0	11
Lactococcus phage jm2	7	0.88	0	11
Lactococcus phage phi7	7	0.88	0	11
Lactococcus phage phiLC3	7	0.88	0	12
Lactococcus phage Tuc2009	7	0.88	0	10
Lactococcus phage ul36	7	0.88	0	10
Lactococcus prophage bIL286	7	0.88	0	9
Lactococcus prophage bIL309	7	0.88	0	9
Moloney murine leukemia virus	7	0.88	0	26
Pestivirus giraffe-1 H138 complete genome	7	0.88	0	13
Pseudomonas phage JBD25	7	0.88	0	55
Salmonella phage S16	7	0.88	0	6
Salmonella phage vB_SnwM_CGG4-1	7	0.88	0	6
Simian adenovirus 1	7	0.88	0	26
Staphylococcus phage 77	7	0.88	0	12
Uncultured crAssphage	7	0.88	0	15
Bacillus prophage phBC6A52	6	0.75	0	6
Lactococcus phage CB14	6	0.75	0	12
Lactococcus phage jj50	6	0.75	0	7
Lactococcus phage sk1	6	0.75	0	7
Pseudomonas phage Pf1	6	0.75	0	13
Ralstonia phage RSY1 DNA	6	0.75	0	9
Salmonella phage Fels-1	6	0.75	0	33
Salmonella phage SEN34	6	0.75	0	9
Salmonella phage SPN3UB	6	0.75	0	14
Staphylococcus phage 3A	6	0.75	0	11
Staphylococcus phage P954	6	0.75	0	8
Staphylococcus phage phi5967PVL DNA	6	0.75	0	9
Staphylococcus phage phiPVL108	6	0.75	0	6
Staphylococcus phage phiSa119	6	0.75	0	9
Staphylococcus phage PVL proviral DNA	6	0.75	0	8
Staphylococcus phage SMSAP5	6	0.75	0	9
Staphylococcus phage StB20	6	0.75	0	10
Staphylococcus phage vB_SepiS-phiIPLA5	6	0.75	0	10
Staphylococcus prophage phi 12	6	0.75	0	13
Staphylococcus prophage phiPV83	6	0.75	0	6
Streptococcus phage Abc2	6	0.75	0	9
Streptococcus prophage EJ-1	6	0.75	0	10
Y73 sarcoma virus	6	0.75	0	5
Bovine herpesvirus 4 long unique region	5	0.62	0	6
Burkholderia phage KS14	5	0.62	0	5
Burkholderia phage phiE125	5	0.62	0	6
Enterobacteria phage fiAA91-ss	5	0.62	0	12
Escherichia phage APCEc01	5	0.62	0	4
Escherichia phage HK639	5	0.62	0	5
Escherichia phage pro483	5	0.62	0	9
Lactococcus phage 28201	5	0.62	0	5
Lactococcus phage 63301	5	0.62	0	7
Lactococcus phage CB13	5	0.62	0	9
Lactococcus phage P008	5	0.62	0	5
Lactococcus phage TP901-1	5	0.62	0	6
Lactococcus prophage bIL285	5	0.62	0	5
Moloney murine sarcoma virus	5	0.62	0	16

PreXMRV-1 provirus	5	0.62	0	31
Pseudomonas phage JBD18	5	0.62	0	23
Pseudomonas phage MD8	5	0.62	0	5
Ralstonia phage 1 NP-2014	5	0.62	0	6
Staphylococcus phage 11	5	0.62	0	16
Staphylococcus phage 55	5	0.62	0	35
Staphylococcus phage B166	5	0.62	0	21
Staphylococcus phage CNPH82	5	0.62	0	24
Staphylococcus phage phi2958PVL	5	0.62	0	9
Staphylococcus phage phiPVL-CN125	5	0.62	0	6
Staphylococcus phage phiSauS-IPLA35	5	0.62	0	7
Staphylococcus phage StauST398-2	5	0.62	0	10
Staphylococcus phage vB_SauS_phi2	5	0.62	0	6
Staphylococcus phage YMC/09/04/R1988	5	0.62	0	8
Staphylococcus prophage phi 13	5	0.62	0	7
Streptococcus phage 2972	5	0.62	0	11
Streptococcus phage ALQ13.2	5	0.62	0	11
Streptococcus phage O1205	5	0.62	0	6
Streptococcus phage P9	5	0.62	0	9
Streptococcus phage PH10	5	0.62	0	7
Streptococcus phage Sfi19	5	0.62	0	7
Streptococcus phage Sfi21	5	0.62	0	6
Streptococcus phage TP-778L complete genome	5	0.62	0	6
Streptococcus prophage 315.6	5	0.62	0	11
Streptococcus thermophilus bacteriophage Sfi11	5	0.62	0	6
Bovine viral diarrhea virus 1	4	0.5	0	5
Choristoneura fumiferana granulovirus	4	0.5	0	51
Enterobacteria phage P2	4	0.5	0	12
Enterobacteria phage P88	4	0.5	0	6
Escherichia phage pro147	4	0.5	0	6
Gordonia phage Blueberry	4	0.5	0	25
Gordonia phage Cucurbita	4	0.5	0	25
Human T-lymphotropic virus 1	4	0.5	0	272
Klebsiella phage vB_KpnM_KB57	4	0.5	0	8
Lactococcus phage 4268	4	0.5	0	4
Lactococcus phage 712	4	0.5	0	6
Lactococcus phage 98201	4	0.5	0	4
Lactococcus phage bIBB29	4	0.5	0	5
Lactococcus phage jm3	4	0.5	0	6
Murine osteosarcoma virus	4	0.5	0	13
Pseudomonas phage F10	4	0.5	0	5
Pseudomonas phage JBD44	4	0.5	0	5
Pseudomonas phage JBD88a	4	0.5	0	5
Pseudomonas phage JBD93	4	0.5	0	6
Pseudomonas phage MP22	4	0.5	0	5
Pseudomonas phage MP29	4	0.5	0	5
Pseudomonas phage MP48	4	0.5	0	6
Pseudomonas phage PA1/KOR/2010	4	0.5	0	5
Pseudomonas phage phi2	4	0.5	0	5
Pseudomonas phage phi3	4	0.5	0	11
Pseudomonas phage vB_PaeP_Tr60_Ab31	4	0.5	0	5
Salmonella phage FSL SP-004	4	0.5	0	63
Salmonella phage SEN1	4	0.5	0	51
Salmonella phage SEN4	4	0.5	0	5
Salmonella phage SEN5	4	0.5	0	5
Simian T-lymphotropic virus 1	4	0.5	0	268
Spleen focus-forming virus	4	0.5	0	43
Staphylococcus phage 47	4	0.5	0	7
Staphylococcus phage 71	4	0.5	0	19
Staphylococcus phage 88	4	0.5	0	7
Staphylococcus phage CNPx	4	0.5	0	37
Staphylococcus phage JS01	4	0.5	0	5
Staphylococcus phage PH15	4	0.5	0	37
Staphylococcus phage phi7401PVL DNA	4	0.5	0	6
Staphylococcus phage phiETA	4	0.5	0	20
Staphylococcus phage phiETA2	4	0.5	0	20
Staphylococcus phage SA12	4	0.5	0	17
Staphylococcus phage SA13	4	0.5	0	14
Staphylococcus phage SA97	4	0.5	0	32
Staphylococcus phage SAP-26	4	0.5	0	25
Streptococcus phage 858	4	0.5	0	8
Streptococcus phage DT1	4	0.5	0	7
Streptococcus phage phiARI0923	4	0.5	0	4
Streptococcus prophage 315.4	4	0.5	0	4
Torque teno virus 10	4	0.5	0	3
Yersinia phage L-413C	4	0.5	0	12
Bacillus phage PfEFR-5	3	0.38	0	6
Bacteriophage D3	3	0.38	0	2
Bacteriophage WPhi	3	0.38	0	9
Dinoroseobacter phage DFL12phi1	3	0.38	0	4

Enterobacteria phage phiP27	3	0.38	0	5
Enterobacteria phage PsP3	3	0.38	0	54
Enterobacterial phage mEp390	3	0.38	0	2
Friend murine leukemia virus FB29 complete genome	3	0.38	0	5
Geobacillus phage GBSV1	3	0.38	0	3
Geobacillus virus E3	3	0.38	0	4
Human herpesvirus 1 strain 17	3	0.38	0	2
Human herpesvirus 2 strain HG52	3	0.38	0	5
Influenza A virus (A/California/07/2009(H1N1)) segment 5	3	0.38	0	4
Influenza A virus (A/California/07/2009(H1N1)) segment 8	3	0.38	0	3
Mouse mammary tumor virus	3	0.38	0	3
Pseudomonas phage B3	3	0.38	0	11
Pseudomonas phage JBD24	3	0.38	0	4
Pseudomonas phage JBD30	3	0.38	0	3
Pseudomonas phage JBD5	3	0.38	0	4
Pseudomonas phage MP42	3	0.38	0	3
Pseudomonas phage NP1	3	0.38	0	5
Pseudomonas phage PaBG	3	0.38	0	2
Pseudomonas phage PAJU2	3	0.38	0	4
Pseudomonas phage phi297	3	0.38	0	3
Pseudomonas phage vB_PaeS_PM105	3	0.38	0	6
Pseudomonas phage vB_PaeS_PMG1	3	0.38	0	2
Ralstonia phage phiRSA1	3	0.38	0	2
Salmonella phage SSU5	3	0.38	0	3
Staphylococcus phage 3MRA	3	0.38	0	3
Staphylococcus phage 42e	3	0.38	0	4
Staphylococcus phage B236	3	0.38	0	18
Staphylococcus phage phiETA3	3	0.38	0	17
Staphylococcus phage phiIPLA-C1C	3	0.38	0	11
Staphylococcus phage StauST398-1	3	0.38	0	6
Staphylococcus phage StauST398-5	3	0.38	0	17
Staphylococcus phage vB_SepiS-phiIPLA7	3	0.38	0	8
Stenotrophomonas phage S1	3	0.38	0	7
Streptococcus phage 7201	3	0.38	0	4
Streptococcus phage IC1	3	0.38	0	3
Streptococcus phage PH15	3	0.38	0	4
Streptococcus phage SM1	3	0.38	0	3
Streptococcus phage TP-J34 complete genome	3	0.38	0	2
Streptococcus prophage 315.1	3	0.38	0	5
Streptococcus prophage 315.2	3	0.38	0	3
Streptococcus virus 9872	3	0.38	0	4
Torque teno virus 1	3	0.38	0	2
Acinetobacter phage IME_AB3	2	0.25	0	1
Acinetobacter phage YMC/09/02/B1251_ABA_BP	2	0.25	0	2
Aeromonas phage PX29	2	0.25	0	2
Aeromonas phage vB_AsaM-56	2	0.25	0	2
Bacillus phage 250	2	0.25	0	2
Bacteriophage 186	2	0.25	0	11
Brucella phage BiPBO1	2	0.25	0	1
Burkholderia phage KL3	2	0.25	0	1
Burkholderia phage KS10	2	0.25	0	1
Caulobacter phage CcrRogue	2	0.25	0	1
Chimpanzee alpha-1 herpesvirus strain 105640	2	0.25	0	2
Cronobacter phage ENT39118	2	0.25	0	1
Cronobacter phage ENT47670	2	0.25	0	1
Encephalomyocarditis virus	2	0.25	0	1
Enterobacteria phage DT57C	2	0.25	0	1
Enterobacteria phage T1	2	0.25	0	2
Enterobacteria phage T5	2	0.25	0	1
Enterococcus phage phiFL4A	2	0.25	0	1
Escherichia phage ADB-2	2	0.25	0	1
Escherichia phage Pollock	2	0.25	0	1
Flavobacterium phage Fpv2	2	0.25	0	1
Flavobacterium phage Fpv20	2	0.25	0	1
Flavobacterium phage Fpv3	2	0.25	0	1
Geobacillus virus E2	2	0.25	0	1
Haemophilus phage SuMu	2	0.25	0	1
Human erythrovirus V9	2	0.25	0	1
Human papillomavirus type 10 genomic DNA	2	0.25	0	2
Human papillomavirus type 134	2	0.25	0	1
Human papillomavirus type 92	2	0.25	0	2
Human papillomavirus type 96	2	0.25	0	2
Human parvovirus B19	2	0.25	0	9
Influenza A virus (A/California/07/2009(H1N1)) segment 1 polymerase PB2 (PB2) gene	2	0.25	0	2
Influenza A virus (A/California/07/2009(H1N1)) segment 2 polymerase PB1 (PB1) gene	2	0.25	0	3
Influenza A virus (A/California/07/2009(H1N1)) segment 4 hemagglutinin (HA) gene	2	0.25	0	4
Influenza A virus (A/California/07/2009(H1N1)) segment 6 neuraminidase (NA) gene	2	0.25	0	2

Influenza A virus (A/California/07/2009(H1N1)) segment 7 matrix protein 2 (M2) and matrix protein 1 (M1) genes	2	0.25	0	4
Influenza A virus (A/goose/Guangdong/1/1996(H5N1)) polymerase (PA) and PA-X protein (PA-X) genes	2	0.25	0	1
Influenza A virus (A/goose/Guangdong/1/1996(H5N1)) segment 7	2	0.25	0	6
Influenza A virus (A/Goose/Guangdong/1/96(H5N1)) neuraminidase (NA) gene	2	0.25	0	2
Influenza A virus (A/Goose/Guangdong/1/96(H5N1)) nucleocapsid protein (NP) gene	2	0.25	0	1
Influenza A virus (A/Hong Kong/1073/99(H9N2)) segment 5	2	0.25	0	1
Influenza A virus (A/Hong Kong/1073/99(H9N2)) segment 7	2	0.25	0	3
Influenza A virus (A/Hong Kong/1073/99(H9N2)) segment 8	2	0.25	0	2
Influenza A virus (A/Korea/426/1968(H2N2)) segment 2	2	0.25	0	2
Influenza A virus (A/Korea/426/1968(H2N2)) segment 7	2	0.25	0	3
Influenza A virus (A/New York/392/2004(H3N2)) segment 7	2	0.25	0	1
Influenza A virus (A/New York/392/2004(H3N2)) segment 8	2	0.25	0	1
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 4	2	0.25	0	1
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 5	2	0.25	0	1
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 6	2	0.25	0	1
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 7	2	0.25	0	2
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 8	2	0.25	0	2
Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 7 matrix protein 2 (M2) and matrix protein 1 (M1) genes	2	0.25	0	1
Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 8 nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes	2	0.25	0	1
Influenza A virus pb1 gene for polymerase Pb1, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	2	0.25	0	1
Klebsiella phage KP34	2	0.25	0	1
Klebsiella phage NTUH-K2044-K1-1 DNA	2	0.25	0	1
Klebsiella phage vB_KpnP_KpV41	2	0.25	0	1
Klebsiella phage vB_KpnP_KpV475	2	0.25	0	1
Lactobacillus phage CL1	2	0.25	0	1
Lactobacillus phage CL2	2	0.25	0	1
Lactobacillus phage iLp1308	2	0.25	0	1
Lactococcus phage bIL67	2	0.25	0	2
Lactococcus phage BK5-T	2	0.25	0	1
Lactococcus phage D4412	2	0.25	0	1
Lactococcus phage M6165	2	0.25	0	1
Lactococcus phage PLgT-1	2	0.25	0	1
Lactococcus prophage bIL310	2	0.25	0	2
Merkel cell polyomavirus isolate R17b	2	0.25	0	2
Microbacterium phage Min1	2	0.25	0	1
Mycobacterium phage Che9c	2	0.25	0	2
Mycobacterium phage Sbash	2	0.25	0	2
Potato virus X	2	0.25	0	2
Pseudomonas phage D3112	2	0.25	0	2
Pseudomonas phage DMS3	2	0.25	0	2
Pseudomonas phage EL	2	0.25	0	2
Pseudomonas phage H70	2	0.25	0	1
Pseudomonas phage JBD69	2	0.25	0	2
Pseudomonas phage JD024	2	0.25	0	2
Pseudomonas phage LPB1, complate genome	2	0.25	0	1
Pseudomonas phage MP38	2	0.25	0	2
Pseudomonas phage phiCTX	2	0.25	0	1
Pseudomonas phage PS-1 DNA	2	0.25	0	1
Pseudomonas phage vB_PaeM_PS24	2	0.25	0	1
Pseudomonas phage vB_PaeS_PAO1_Ab30	2	0.25	0	3
Pseudomonas phage YMC11/02/R656	2	0.25	0	1
Pseudomonas phage YMC11/07/P54_PAE_BP	2	0.25	0	2
Ralstonia phage p12J	2	0.25	0	1
Ralstonia phage RS603 DNA	2	0.25	0	1
Ralstonia phage RSM1	2	0.25	0	1
Ralstonia phage RSM3	2	0.25	0	1
RD114 retrovirus	2	0.25	0	2
Rous sarcoma virus	2	0.25	0	1
Salmonella phage BP12A	2	0.25	0	1
Salmonella phage Shivani	2	0.25	0	1
Salmonella phage Vi06 complete sequence	2	0.25	0	3
Shigella phage pSf-2	2	0.25	0	2
Shigella phage Shfl1	2	0.25	0	1
Shigella phage SHSML-45	2	0.25	0	1
Simian adenovirus 18	2	0.25	0	1
Simian adenovirus 3	2	0.25	0	1
Simian T-cell lymphotropic virus 6 strain Cmo8699AB	2	0.25	0	1
Squirrel monkey retrovirus - HLB	2	0.25	0	2
Staphylococcus aureus phage phiNM2	2	0.25	0	2
Staphylococcus aureus phage phiNM4	2	0.25	0	1
Staphylococcus phage 29	2	0.25	0	1
Staphylococcus phage 52A	2	0.25	0	2
Staphylococcus phage 53	2	0.25	0	2
Staphylococcus phage 69	2	0.25	0	2

Staphylococcus phage 6ec	2	0.25	0	1
Staphylococcus phage 80alpha	2	0.25	0	2
Staphylococcus phage 85	2	0.25	0	2
Staphylococcus phage 92	2	0.25	0	1
Staphylococcus phage IME-SA4	2	0.25	0	2
Staphylococcus phage phiJB	2	0.25	0	1
Staphylococcus phage phiMR25	2	0.25	0	4
Staphylococcus phage phiNM1	2	0.25	0	3
Staphylococcus phage phiSauS-IPLA88	2	0.25	0	3
Staphylococcus phage phiSLT	2	0.25	0	1
Staphylococcus phage PT1028	2	0.25	0	1
Staphylococcus phage SA1	2	0.25	0	4
Staphylococcus phage SA11	2	0.25	0	2
Staphylococcus phage StB12	2	0.25	0	2
Staphylococcus phage StB27	2	0.25	0	1
Staphylococcus phage TEM123	2	0.25	0	4
Staphylococcus phage vB_SepS_SEP9	2	0.25	0	2
Staphylococcus phage X2	2	0.25	0	1
Stenotrophomonas phage vB_SmaS-DLP_2	2	0.25	0	1
Streptococcus phage A25	2	0.25	0	1
Streptococcus phage K13	2	0.25	0	2
Streptococcus phage phi3396	2	0.25	0	1
Streptococcus phage phiARI0131-1	2	0.25	0	1
Streptococcus phage phiARI0131-2	2	0.25	0	2
Streptococcus phage phiARI0460-1	2	0.25	0	2
Streptococcus phage phiARI0462	2	0.25	0	1
Streptococcus phage phiBHN167 complete genome	2	0.25	0	1
Streptococcus phage SpSL1	2	0.25	0	1
Streptococcus phage YMC-2011	2	0.25	0	1
Streptococcus prophage 315.5	2	0.25	0	1
Streptococcus virus 9871	2	0.25	0	1
Streptococcus virus 9874	2	0.25	0	1
Thermus phage phiYS40	2	0.25	0	2
Thermus phage TMA	2	0.25	0	1
Thermus thermophilus phage IN93	2	0.25	0	1
Tobacco vein clearing virus	2	0.25	0	1
Torque teno virus 12	2	0.25	0	1
Torque teno virus 16	2	0.25	0	1
Torque teno virus 19	2	0.25	0	1
Torque teno virus 6	2	0.25	0	1
Torque teno virus 8	2	0.25	0	1

References

[1] A. Albrechtsen et al. "Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes". In: *Diabetologia* 56.2 (2013), pp. 298–310. ISSN: 0012186X. DOI: 10.1007/s00125-012-2756-1.

[2] Emrah Altindis et al. "Viral insulin-like peptides activate human insulin and IGF–1 receptor signaling: A paradigm shift for host–microbe interactions". In: *Proc. Natl. Acad. Sci.* (2018), p. 201721117. ISSN: 0027–8424. DOI: 10.1073/pnas.1721117115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1721117115.

[3] SF F Altschul et al. *Basic local alignment search tool*. 1990. DOI: 10.1016/S0022-2836(05)80360-2. URL: http://www.cmu.edu/bio/education/courses/03510/LectureNotes/Altschul1990.pdf%7B%5C%%7D5Cnhttp://www.sciencedirect.com/science/article/pii/S0022283605803602%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/2231712.

[4] Adam Auton et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74. ISSN: 14764687. DOI: 10.1038/nature15393. arXiv: 15334406.

[5] David J. Balding. "A tutorial on statistical methods for population association studies". In: *Nat. Rev. Genet.* 7.10 (2006), pp. 781–791. ISSN: 14710056. DOI: 10.1038/nrg1916. arXiv: 92-5-105182-8.

[6] J. Rodney Brister et al. "NCBI viral Genomes resource". In: *Nucleic Acids Res.* 43.D1 (2015), pp. D571–D577. ISSN: 13624962. DOI: 10.1093/nar/gku1207.

[7] William S. Bush and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies". In: *PLoS Comput. Biol.* 8.12 (2012). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002822. arXiv: NIHMS150003.

[8] William S. Bush and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies". In: *PLoS Comput. Biol.* 8.12 (2012). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002822. arXiv: NIHMS150003.

[9] Ken Cadwell. "The Virome in Host Health and Disease". In: *Immunity* 42.5 (2015), pp. 805–813. ISSN: 10974180. DOI: 10.1016/j.immuni.2015.05.003. arXiv: 15334406. URL: http://dx.doi.org/10.1016/j.immuni.2015.05.003.

[10] Christopher C. Chang et al. "Second-generation PLINK: Rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1 (2015), pp. 1–16. ISSN: 2047217X. DOI: 10.1186/s13742-015-0047-8. arXiv: 1410.4803.

[11] L De Bolle, L Naesens, and E De Clercq. "Update on human herpesvirus 6 biology, clinical features, and therapy". In: *Clin.Microbiol.Rev.* 18.0893-8512 (Print) (2005), pp. 217–245. ISSN: 0893-8512. DOI: 10.1128/CMR.18.1.217. URL: c:%7B%5C%%7D5CKarsten%7B%5C%%7D5CPDFs%7B%5C%%7D5CVirologie-PDFs%7B%5C%%7D5CViro-2005%7B%5C%%7D5CDe%20Bolle%20et%20al.-Update%20on%20human%20herpesvirus%206%20biology%20clinical%20features%20and%20therapy.pdf.

[12] Keyue Ding and Iftikhar J. Kullo. "Genome-wide association studies for atherosclerotic vascular disease and its risk factors". In: *Circ. Cardiovasc. Genet.* 2.1 (2009), pp. 63–72. ISSN: 1942325X. DOI: 10.1161/CIRCGENETICS.108.816751.

[13] Amy Y. Hafez and Micah A. Luftig. "Characterization of the EBV-induced persistent DNA damage response". In: *Viruses* 9.12 (2017). ISSN: 19994915. DOI: 10.3390/v9120366.

[14] Ezazul Haque et al. "Worldwide increased prevalence of human adenovirus type 3 (HAdV3) respiratory infections is well correlated with heterogeneous hypervariable regions (HVRs) of hexon". In: *PLoS One* 13.3 (2018), pp. 1–12. ISSN: 19326203. DOI: 10.1371/journal.pone.0194516.

[15] Mikko Heino. "Quantitative Traits". In: *Stock Identif. Methods* (2014), pp. 59–76. ISSN: 14710064. DOI: 10.1016/B978-0-12-397003-9.00004-7. URL: http://linkinghub.elsevier.com/retrieve/pii/B9780123970039000047.

[16] E. Hellebrand et al. "Epstein-Barr virus vector-mediated gene transfer into human B cells: Potential for antitumor vaccination". In: *Gene Ther.* 13.2 (2006), pp. 150–162. ISSN: 09697128. DOI: 10.1038/sj.gt.3302602.

[17] H Hill. *Fixing Teacher Professional Development*. 2009. DOI: 10.1177/0964663912467814. arXiv: arXiv:1011.1669v3.

[18] Joel N. Hirschhorn and Mark J. Daly. "Genome-wide association studies for common diseases and complex traits". In: *Nat. Rev. Genet.* 6.2 (2005), pp. 95–108. ISSN: 14710056. DOI: 10.1038/nrg1521.

[19] Joel N. Hirschhorn and Mark J. Daly. "Genome-wide association studies for common diseases and complex traits". In: *Nat. Rev. Genet.* 6.2 (2005), pp. 95–108. ISSN: 14710056. DOI: 10.1038/nrg1521.

[20] Myung-Soo Kang and Elliott Kieff. "Epstein-Barr virus latent genes". In: *Exp. Mol. Med.* 47.1 (2015), e131. ISSN: 2092-6413. DOI: 10.1038/emm.2014.84. URL: http://www.nature.com/doifinder/10.1038/emm.2014.84.

[21] Mary F. Kearney et al. "Multiple sources of contamination in samples from patients reported to have XMRV infection". In: *PLoS One* 7.2 (2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0030889.

[22] Mark Kowarsky et al. "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA". In: *Proc. Natl. Acad. Sci.* 114.36 (2017), pp. 9623–9628. ISSN: 0027-8424. DOI: 10.1073/pnas.1707009114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1707009114.

[23] Chitose Kurihara et al. "An easy method for preparation of Cre-loxP regulated fluorescent adenoviral expression vectors and its application for direct reprogramming into hepatocytes". In: *Biotechnol. Reports* 12 (2016), pp. 26–32. ISSN: 2215017X. DOI: 10.1016/j.btre.2016.10.003. URL: http://dx.doi.org/10.1016/j.btre.2016.10.003.

[24] Wonkeun Lee et al. "An Epstein-Barr Virus Isolated from a Lymphoblastoid Cell Line Has a 16-Kilobase-Pair Deletion Which Includes gp350 and the Epstein-Barr Virus Nuclear Antigen 3A An Epstein-Barr Virus Isolated from a Lymphoblastoid Cell Line Has a 16-Kilobase-Pair Deletio". In: *J. Virol.* 75.18 (2001), pp. 8556–8568. DOI: 10.1128/JVI.75.18.8556.

[25] Emily C. Leibovitch et al. "Coinfection of human herpesviruses 6A (HHV-6A) and HHV-6B as demonstrated by novel digital droplet PCR assay". In: *PLoS One* 9.3 (2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0092328.

[26] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp324. arXiv: 1303.3997.

[27] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp352. arXiv: 1006.1266v2.

[28] Kirk E. Lohmueller et al. "Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes". In: *Am. J. Hum. Genet.* 93.6 (2013), pp. 1072–1086. ISSN: 00029297. DOI: 10.1016/j.ajhg.2013.11.005.

[29] Jennifer Lu and Steven L. Salzberg. "Removing contaminants from databases of draft genomes". In: *PLOS Comput. Biol.* 14.6 (2018), e1006277. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006277. URL: http://dx.plos.org/10.1371/journal.pcbi.1006277.

[30] Rajendra Mandage et al. "Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples". In: *PLoS One* 12.6 (2017), pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0179446.

[31] Mark I McCarthy et al. "Genome-wide association studies for complex traits: Consensus, uncertainty and challenges". In: *Nat. Rev. Genet.* 9.5 (2008), pp. 356–369. ISSN: 14710056. DOI: 10.1038/nrg2344. arXiv: NIHMS150003.

[32] Samuel Minot et al. "The human gut virome : Inter-individual variation and dynamic response to diet The human gut virome : Inter-individual variation and dynamic response to diet". In: (2011), pp. 1616–1625. DOI: 10.1101/gr.122705.111.

[33] Walter N. Moss and Joan A. Steitz. "Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA". In: *BMC Genomics* 14.1 (2013). ISSN: 14712164. DOI: 10.1186/1471-2164-14-543.

[34] Ahmed Moustafa et al. "The blood DNA virome in 8,000 humans". In: (2017), pp. 1–20.

[35] Heidemarie Neitzel. "A routine method for the establishment of permanent growing lymphoblastoid cell lines". In: *Hum. Genet.* 73 (1986), pp. 320–326.

[36] Jérôme T J Nicol et al. "Age-specific seroprevalences of merkel cell polyomavirus, human polyomaviruses 6, 7, and 9, and trichodysplasia spinulosaassociated polyomavirus". In: *Clin. Vaccine Immunol.* 20.3 (2013), pp. 363–368. ISSN: 15566811. DOI: 10.1128/CVI.00438-12.

[37] Anne L. Palser et al. "Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection". In: *J. Virol.* 89.10 (2015), pp. 5222–5237. ISSN: 0022-538X. DOI: 10.1128/JVI.03614-14. URL: http://jvi.asm.org/lookup/doi/10.1128/JVI.03614-14.

[38] Thomas a Pearson and Teri a Manolio. "How to interpret a genome-wide association study." In: *Jama* 299.11 (2008), pp. 1335–1344. ISSN: 1538-3598. DOI: 10.1001/jama.299.11.1335. URL: http://www.ncbi.nlm.nih.gov/pubmed/18349094.

[39] A L Price et al. "New approaches to population stratification in genome-wide association studies". In: *Nat. Rev. Genet.* 11.7 (2010), pp. 459–463. ISSN: 1471-0056. DOI: 10.1038/nrg2813. arXiv: NIHMS150003.

[40] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nat. Genet.* 38.8 (2006), pp. 904–909. ISSN: 10614036. DOI: 10.1038/ng1847.

[41] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic Acids Res.* 35.SUPPL. 1 (2007), pp. 61–65. ISSN: 03051048. DOI: 10.1093/nar/gkl842. arXiv: NIHMS150003.

[42] Alejandro Reyes et al. "Viruses in the faecal microbiota of monozygotic twins and their mothers". In: *Nature* 466.7304 (2010), pp. 334–338. ISSN: 00280836. DOI: 10.1038/nature09199.

[43] Christopher M. Robinson et al. "Molecular evolution of human adenoviruses". In: *Sci. Rep.* 3 (2013), pp. 1–7. ISSN: 20452322. DOI: 10.1038/srep01812.

[44] Noah A Rosenberg et al. "Genome-wide association studies in diverse populations". In: *Nat. Rev. Genet.* 11.5 (2010), pp. 356–366. ISSN: 14710056. DOI: 10.1038/nrg2760.

[45] Simon Roux et al. "Viral dark matter and virus-host interactions resolved from publicly available microbial genomes". In: *Elife* 4.JULY2015 (2015), pp. 1–20. ISSN: 2050084X. DOI: 10.7554/eLife.08490.001. arXiv: arXiv:1011.1669v3.

[46] Gabriel Santpere et al. "Genome-wide analysis of wildtype epstein-barr virus genomes derived from healthy individuals of the 1000 genomes project". In: *Genome Biol. Evol.* 6.4 (2014), pp. 846–860. ISSN: 17596653. DOI: 10.1093/gbe/evu054.

[47] Sagiv Shifman et al. "Linkage disequilibrium patterns of the human genome across populations". In: *Hum. Mol. Genet.* 12.7 (2003), pp. 771–776. ISSN: 09646906. DOI: 10.1093/hmg/ddg088.

[48] Peter Simmonds et al. "Consensus statement: Virus taxonomy in the age of metagenomics". In: *Nat. Rev. Microbiol.* 15.3 (2017), pp. 161–168. ISSN: 17401534. DOI: 10.1038/nrmicro.2016.177.

[49] JT Simpson and Richard Durbin. "Efficient de novo assembly of large genomes using compressed data structures". In: *Genome Res.* (2012), pp. 549–556. DOI: 10.1101/gr.126953.111.Freely. URL: http://genome.cshlp.org/content/22/3/549.short.

[50] Robert Sladek et al. "A genome-wide association study identifies novel risk loci for type 2 diabetes". In: *Nature* 445.7130 (2007), pp. 881–885. ISSN: 00280836. DOI: 10.1038/nature05616. arXiv: t8jd4qr3m [13960].

[51] Barbara E. Stranger, Eli A. Stahl, and Towfique Raj. "Progress and promise of genome-wide association studies for human complex trait genetics". In: *Genetics* 187.2 (2011), pp. 367–383. ISSN: 00166731. DOI: 10.1534/genetics.110.120907.

[52] Amalio Telenti et al. "Deep Sequencing of 10 , 000 Human Genomes". In: (2016), pp. 1–29.

[53] Amalio Telenti et al. "Deep sequencing of 10,000 human genomes". In: *Proc. Natl. Acad. Sci.* 113.42 (2016), pp. 11901–11906. ISSN: 0027-8424. DOI: 10.1073/pnas.1613365113. arXiv: 061663. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1613365113.

[54] Chao Tian et al. "Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections". In: *Nat. Commun.* 8.1 (2017). ISSN: 20411723. DOI: 10.1038/s41467-017-00257-5. arXiv: s41467-017-00257-5 [DOI: 10.1038]. URL: http://dx.doi.org/10.1038/s41467-017-00257-5.

[55] Yanis L. Tolstov et al. "Human Merkel cell polyomavirus infection II. MCV is a common human infection that can be detected by conformational capsid epitope immunoassays". In: *Int. J. Cancer* 125.6 (2009), pp. 1250–1256. ISSN: 00207136. DOI: 10.1002/ijc.24509. arXiv: NIHMS150003.

[56] Kristien Verdonck et al. "Human T-lymphotropic virus 1 : recent knowledge about an". In: *Lancet Infect.Dis.* 7.4 (2007), pp. 266–281.

[57] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *Am. J. Hum. Genet.* 101.1 (2017), pp. 5–22. ISSN: 15376605. DOI: 10.1016/j.ajhg.2017.06.005. URL: http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

[58] William Y S Wang et al. "Genomewide association studies: Theoretical and practical concerns". In: *Nat. Rev. Genet.* 6.2 (2005), pp. 109–118. ISSN: 14710056. DOI: 10.1038/nrg1522. arXiv: NIHMS150003.

[59] Christian Widmer et al. "Further improvements to linear mixed models for genome-wide association studies". In: *Sci. Rep.* 4 (2014). ISSN: 20452322. DOI: 10.1038/srep06874.

[60] John S. Witte. "Genome-Wide Association Studies and Beyond". In: *Annu. Rev. Public Health* 31.1 (2010), pp. 9–20. ISSN: 0163-7525. DOI: 10.1146/annurev.publhealth.012809.103723. arXiv: NIHMS150003. URL: http://www.annualreviews.org/doi/10.1146/annurev.publhealth.012809.103723.

[61] William S M Wold and Karoly Toth. "Adenovirus vectors for gene therapy, vaccination and cancer gene therapy." In: *Curr. Gene Ther.* 13.6 (2013), pp. 421–33. ISSN: 1875-5631. DOI: 10.2174/1566523213666131125095046. arXiv: 15334406. URL: http://www.ncbi.nlm.nih.gov/pubmed/24279313%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4507798.

[62] Derrick E. Wood and Steven L. Salzberg. "Kraken Ultrafast metagenomic sequence classification using exact alignments". In: *Genome Biol.* 15.3 (2014). ISSN: 1474760X. DOI: 10.1186/gb-2014-15-3-r46. arXiv: /www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164{\&}tool=pmcentrez{\&}rendertype=abstract. [Figures, S., 2010. Supplementary information. Nature, 1(c), pp.1--7. Available at: http:].

[63] Qingfa Wu et al. "Identification of Viruses and Viroids by Next-Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms". In: *Annu. Rev. Phytopathol.* 53.1 (2015), pp. 425–444. ISSN: 0066-4286. DOI: 10.1146/annurev-phyto-080614-120030. URL: http://www.annualreviews.org/doi/10.1146/annurev-phyto-080614-120030.

[64] Q. Wu et al. "Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm". In: *Proc. Natl. Acad. Sci.* 109.10 (2012), pp. 3938–3943. ISSN: 0027-8424. DOI: 10.1073/pnas.1117815109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1117815109.

[65] Kai Xiao et al. "Genome-wide analysis of Epstein-Barr Virus (EBV) integration and strain in C666-1 and Raji cells". In: *J. Cancer* 7.2 (2016), 214–224.pdf. ISSN: 18379664. DOI: 10.7150/jca.13150.

[66] Jian Yang et al. "Advantages and pitfalls in the application of mixed-model association methods". In: *Nat. Genet.* 46.2 (2014), pp. 100–106. ISSN: 15461718. DOI: 10.1038/ng.2876. arXiv: NIHMS150003.

[67] Rokusuke Yoshikawa et al. "Characterization of RD-114 virus isolated from a commercial canine vaccine manufactured using CRFK cells". In: *J. Clin. Microbiol.* 48.9 (2010), pp. 3366–3369. ISSN: 00951137. DOI: 10.1128/JCM.00992-10.

[68] Lawrence S. Young and Alan B. Rickinson. "Epstein-Barr virus: 40 Years on". In: *Nat. Rev. Cancer* 4.10 (2004), pp. 757–768. ISSN: 1474175X. DOI: 10.1038/nrc1452.

[69] Lan-lan Zhang et al. "Human T-cell lymphotropic virus type 1 and its oncogenesis". In: *Acta Pharmacol. Sin.* 38.8 (2017), pp. 1093–1103. ISSN: 1671-4083. DOI: 10.1038/aps.2017.17. URL: http://www.nature.com/doifinder/10.1038/aps.2017.17.

[70] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nat. Genet.* 44.7 (2012), pp. 821–824. ISSN: 10614036. DOI: 10.1038/ng.2310. arXiv: arXiv:1305.4366v2.