Published on Aug 25, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-1a71-ef4d

# A capsid-based search recovers viral sequences from human brain sequencing data

We implemented a lightweight method to identify viruses in 342 human brain bulk and single-cell sequencing data sets, and identified two glioblastoma cells from a single patient that contained deltapolyomavirus sequences.

**Contributors (A-Z)**

Adair L. Borges, Feridun Mert Celebi, Rachel J. Dutton, Megan L. Hochstrasser, Kira E. Poskanzer, Taylor Reiter

*Version 1 · Mar 31, 2025*

## Purpose

Viruses have naturally evolved to infect cell types and organ compartments across the body, making them very effective delivery systems for genetic medicines. However, tissue-specific targeting is still a major challenge in this area, and some compartments are more difficult to target than others. The brain is one of the most protected organs in the human body, making drug delivery to this organ a huge challenge. We hoped to identify viruses that evolved the ability to cross the blood–brain barrier.

We consider it likely that many healthy humans have latent/asymptomatic viral infections in different organ compartments, including the brain. We sought to identify viruses capable of entering the brains of healthy people to determine molecular signatures of neurotropism. We therefore decided to implement a method capable of capturing viral genomes from sequencing data from human tissues and single cells, starting in human brain data sets.

We implemented a lightweight pipeline that first uses a marker gene-based approach to identify sequencing libraries containing viral capsids, and then uses read mapping and assembly to pull out viral sequences. Using this approach, along with several steps of aggressive filtering, we found 11 possible hits. We are most confident that we've recovered deltapolyomavirus sequences from two glioblastoma cells of a single patient. While the biological significance of this finding is unclear, we are pleased with how our approach performed.

We will not be following up on this work at this time due to the scarcity of appropriate sequencing data sets, and are sharing the details for others who are interested in similar applications. While we did not test this approach in other tissue types, we think it could be used to find viral sequences in other human sequencing data sets.

- Access all of our **code** and **data**, including our initial capsid search results, candidate viral capsid reads, capsid assemblies, read mapping results, and assemblies from mapped reads in this GitHub repository.

# We've put this effort on ice! ☒

**#StrategicMisalignment #TechnicalGap**

We do not have access to the quantity of human single-cell WGS data that we estimate we would need to continue this project, and do not plan to generate those data sets in the near future.

Learn more about the Icebox and the different reasons we ice projects.

# Background and goals

A wide variety of diseases stem from pathology or misregulation in the brain, from neurodegenerative conditions to mood disorders. Finding methods of delivery that bypass this barrier has historically been a major challenge, and the toolkit for delivering genetic cargos to the brain is still very limited. We thought there might be potential to find new viruses that cross the blood–brain barrier by checking for traces of their genomes in brain sequencing data sets.

We hypothesized that healthy humans have latent viral infections that we could detect by analyzing publicly available sequencing data sets. Indeed, a previous PCR-based survey revealed that ~12% of human brain tissue samples are positive for adeno-associated virus (AAV) DNA sequences **[1]**. A variety of approaches exist for recovering microbial sequences in human data (e.g. IDseq **[2]** and PathSeq **[3][4]**), and are often applied through the lens of pathogen identification.

We set out to recover viral genomes and define the basis of their neurotropism (specificity for the brain) by searching brain sequencing data sets. We therefore created a lightweight pipeline to mine human brain DNA and RNA sequencing data sets for viral sequences.

> Access all of our **code** and **data**, including our initial capsid search results, candidate viral capsid reads, capsid assemblies, read mapping results, and assemblies from mapped reads in this GitHub repository (DOI: 10.5281/zenodo.8253080).

# The approach

We used a lightweight computational approach to rapidly scan human brain sequencing libraries for the presence of viruses, using viral capsid proteins as marker genes. After filtering to remove contaminants and false positives, we used assembly and read mapping to recover reads of viral origin.

# Data set selection

We initially considered using four types of publicly available sequencing data sets: Bulk RNA sequencing, bulk DNA sequencing, single-cell RNA sequencing, and single-cell DNA sequencing. We decided to survey these different types of data because we had no idea which type would be most amenable to viral sequence mining. In bulk tissue sequencing, researchers survey many cells at once, increasing the odds that a viral genome may be present. In single-cell sequencing data sets, they cast a much smaller net, but there may be a more favorable ratio of viral to host nucleic acids. We were also unsure whether DNA or RNA sequencing would yield more viral hits. DNA sequencing will catch double-stranded DNA (dsDNA) viruses, but may not capture single-stranded DNA (ssDNA) viruses and will definitely miss RNA viruses. RNA sequencing will catch RNA viruses, but can only capture DNA viruses if they are transcriptionally active. As we dug further into these data types, we quickly realized that most public single-cell RNA sequencing experiments use 10x technology, which only sequences a small 3′ or 5′ fragment of the transcript. We can leverage 10x single-cell RNA sequencing data sets for microbial detection (for example, using SIMBA **[5]**), but we specifically wanted our pipeline to be able to reconstruct viral genomes from sequencing data. This would be impossible with the short fragments that this type of approach returns. We chose to move forward with only paired-end bulk RNA/DNA sequencing and paired-end single-cell DNA sequencing.

We curated a list of 561 publicly available sequencing runs (spanning 81 individuals) from the NCBI Sequence Read Archive (SRA). While our goal was to identify viral genomes from latent infections in healthy individuals, we also included data from glioblastoma and COVID-19 patients. More of these patient data sets were available and they could still contain intriguing hits, they're just more likely to include viruses that can only infect people with a compromised blood–brain barrier and/or immune system. As we began to work with the data, we found that samples with file sizes larger than three gigabytes were difficult to work with in our computing setup, due to increased hard drive space and RAM requirements. We also found that the deeply sequenced samples appeared to have higher rates of background contamination (presumably from sample handling, kit contamination, and index hopping), as measured by running `sourmash gather` (version 4.7.0) **[6]** on the samples with our SeqQC contamination database **[7]**. Higher contamination rates strongly limit our ability to detect true positives and are a major red flag for us. We chose to move

forward with 319 “small” samples (< 3 GB) and 23 large samples that we had already downloaded and processed. This left us with a total of 342 sequencing runs.

## Computational rationale

In designing our approach, we first aimed to implement a fully organism-agnostic, database-free pipeline that would be able to discover any non-human read in sequencing data. We thought this would cast the broadest net to discover new or unusual viral agents, and would even capture non-viral entities as well. Our plan was to use read mapping against the human reference genome followed by more targeted cleanup steps inspired by approaches such as Read Origin Protocol **[8]**, PathSeq **[3] [4]**, and IDseq **[2]**. We anticipated that we would be able to remove > 98% of reads, leaving us with a small set of putatively microbial reads to deal with. We then planned to *de novo* assemble these leftover reads and then classify the resulting contigs. However, when we tried implementing this approach on a subset of the data, we ran into several snags. First, it took a long time to process even a few samples. Second, for the data sets that we did process this way, very few of the leftover reads actually assembled into contigs. This created more challenges — accurately annotating short reads is challenging, and it was difficult to assess shared sequence content within and between samples. In rethinking our pipeline, we decided to use a marker gene-based approach inspired by recent petabase-scale discovery efforts for novel RNA viruses **[9]**. This type of method is highly scalable while being more robust to the magnitude of evolutionary divergence common in viruses.

View the workflow code for our initial subtractive mapping approach.

## Capsid-based search

We used DIAMOND BLASTx (version 2.0.8) **[10]** to align the translated read set (including all human reads) against a database of viral capsid amino acid sequences that we downloaded from the Virus Orthologous groups (VOG) database. Most viruses have some type of capsid, making this a reliable indicator gene for the presence of a virus. We found that a 90% ID cutoff reliably returned non-human/viral sequences, as determined by BLASTn against the NCBI nt ( `-remote` ). We tried lower cutoffs (67% and

80%), but these frequently returned human reads. We also required that both read pairs hit a viral capsid to reduce spurious hits. We excluded all hits related to bacteriophage capsids, as these likely come from exogenous contamination. Given that we do not expect to find bacteria in the brains of healthy, glioblastoma, or COVID-19 patients, it is unlikely that the bacteria-infecting phages would be present either. We also discarded a common hit to a repetitive region of a Mimivirus capsid (VOG ID: 2487768.AYV81705.1), as well as an obvious case where an entire NCBI BioProject was likely contaminated with human mastadenovirus sequences (VOG ID: 10515.AP_000172.1). After encountering this putatively contaminated BioProject (PRJNA527986), we made a filtering rule where we required that a capsid hit be represented at the family level in more than one BioProject to avoid recovering potential laboratory contaminants. If we were to repeat this, we wouldn't implement this filtering rule since we are concerned it could lead to the removal of legitimate hits — rather, we would just remove the single problematic BioProject. After filtering, we assembled the reads that hit the capsid sequences to make mini assemblies using MEGAHIT (version 1.2.8) **[11]**, and used BLASTn to search these mini-assemblies as well as any unassembled reads against the NCBI Nucleotide database to determine their origin. We then discarded all capsid hits that had any read matches to human sequences or cloning vectors.

> View our BLAST workflow code and the code we used to build the BLAST database.

## Viral genome reconstruction

We took the runs that seemed to have legitimate virus signal in them after all of our filtering and built them into an assembly graph using BCALM (version 2.2.3) **[12]** to try and recover the whole viral genome. In most cases, we were not able to recover viral contigs, despite having an indication from our capsid-based search that there were potentially reads of viral origin present. In a further effort to reconstruct these genomes, we used spacegraphcats (version 2.1.2) **[13]** to query with the capsid reads to try and bait out viral reads located nearby in the assembly graph. This approach failed for the samples on which we tried it, for one of two reasons. For some samples, there were no additional connected reads in the assembly graph neighborhood, indicating that sequencing was too shallow to capture the whole viral genome. For the

other samples, there were many additional connected reads. However, when we visualized the assembly graph neighborhood with Bandage **[14]**, and when we BLASTed the capsid reads back to these graphs, the assembly graphs were too tangled to be helpful in genome recovery. We therefore did not pursue this approach further.

> View the workflow code for our assembly graph-based approach to reconstruct viral genomes.

## Read mapping

Using BLAST results of putative viral reads and mini-assemblies against the NCBI Nucleotide database, we identified the viral genomes most likely to be representative of the viruses in the brain samples (Table 2). We also pulled the genomes of the viruses from which the initial capsid hit originated (Table 2). We then used read mapping of the capsid-positive sequencing runs against those genomes as another way to try and bait out all of the reads of potential viral origin. We used Bowtie 2 (version 2.5.1) **[15]** for nucleotide read mapping with parameters `--no-unal --very-sensitive-local`. We also tried read mapping with the protein read mapper PALADIN **[16]** and DIAMOND BLASTx **[10]**, but found that both of these methods returned too many false positives, so we moved forward with nucleotide mapping alone.

> View the workflow code for our read mapping strategy for recovering viral sequences.

## Serratus Explorer NT search query

We used the serratus.io web server to compare the results of the Serratus NT search to our capsid-based search. The Serratus NT search is a pre-computed database of the results of read mapping the full SRA against the vertebrate viral pangenome **[9]**, and represents a complementary approach to the marker gene-based search that we used. We searched the database using the SRA run as our query and retrieved hits

that had an alignment ID of > 75% and a non-zero score. This filtering is not stringent, and likely would not be appropriate for initial virus discovery. However, our goal here was to compare this read mapping approach to the marker gene-based approach. A protein-based search is inherently more evolutionarily flexible, so we loosened the read mapping parameters accordingly.

> Access all of our **code** and **data**, including our initial capsid search results, candidate viral capsid reads, capsid assemblies, read mapping results, and assemblies from mapped reads in this GitHub repository.

# The results

## Initial hits

We searched 342 human brain sequencing data sets for viral capsids using DIAMOND BLASTx searches of translated reads against a capsid database from VOGDB. With a 90% ID cutoff and the requirement that both read pairs hit a capsid protein, we found 49 capsid sequences (representing unique viral lineages) in the human brain data. After removing phage hits and hits that were only present in a single NCBI BioProject at the family level, we were left with eight viral lineages. Next, we assembled the reads that hit those viruses, BLASTing both the mini-assemblies and the unassembled reads to discard all hits that potentially derived from human sequence or cloning vectors.

After filtering, we were left with a total of 11 samples from three different BioProjects that we predicted to contain virus capsid marker genes. Based on the capsid hits, we predicted seven out of the 11 viruses to be polyomaviruses (alpha and delta). There were two hits to gammapapillomavirus, one hit to mastadenovirus, and one hit to parvovirus (Table 1). We confirmed the deltapolyomavirus hits by BLASTing the capsid reads, but the alphapolyomavirus hits did not match anything, or in one case, the capsid reads matched many different viruses (Table 1). We saw similar patterns for the other hits - some were supported by the BLAST results and others were not. We also queried Serratus Explorer, a precomputed database that has results from read mapping the entire SRA against a vertebrate viral pangenome **[9]**. The Serratus read mapping data did not find any viruses in the whole-genome sequencing (WGS) runs

that we flagged as virus-positive using our capsid-based search, but did have more overlap with our analysis of RNA sequencing libraries (Table 1).

| Run | BioProject | Experiment |
|---|---|---|
| **SRR1778915** | PRJNA273155 | WGS of WGA DNA from single nuclei from brain tumor patient B **[17]** |
| **SRR1779200** | PRJNA273155 | WGS of WGA DNA from single nuclei from brain tumor patient B **[17]** |
| **SRR8750801** | PRJNA527986 | Bulk RNA-seq, Inferior temporal, Healthy adult human ha **[18]** |
| **SRR8750456** | PRJNA527986 | Bulk RNA-seq, Substantia nigra, Healthy adult human hb **[18]** |
| **SRR8750473** | PRJNA527986 | Bulk RNA-seq, Supramarginal, Healthy adult human hc **[18]** |
| **SRR8750734** | PRJNA527986 | Bulk RNA-seq, Dorsolateral prefrontal, Healthy adult human hb **[18]** |
| **SRR14788345** | PRJNA736951 | WGS of bulk brain, Cerebellum, Healthy adult human 8305 **[19]** |
| **SRR14862871** | PRJNA736951 | WGS of bulk brain, Left temporal, Healthy adult human 8305 **[19]** |
| **SRR14862884** | PRJNA736951 | WGS of bulk brain, Cerebellum, Healthy adult human 8305 **[19]** |
| **SRR14999724** | PRJNA736951 | WGS of bulk brain, Right prefrontal, Healthy adult human 8307 **[19]** |
| **SRR14862871** | PRJNA736951 | WGS of bulk brain, Left temporal, Healthy adult human 8305 **[19]** |

Table 1

**Sequencing runs with viral hits**.

For each sequencing run, we have listed the BioProject that it comes from, a description of the experiment, analysis of that run on the Serratus Explorer database, BLAST results of the reads/mini-assemblies searched against the nr database, and the capsid marker gene hit from our initial search.

## Attempts to recover additional viral sequences

To recover all the viral sequences in the 11 sequencing runs (instead of just the reads mapping to capsid), we tried assembling the full read sets. We also used read mapping against a handful of viral genomes (Table 2) — either top BLAST hits or some of our original capsid hits — to bait out viral reads using an assembly-free approach. This worked well for the two data sets (SRR1778915 and SRR8750456) that hit a deltapolyomavirus capsid in our initial scan (Table 1). These read sets are single-cell whole genomes from two different glioblastoma cells from the same patient, from whom a total of 98 glioblastoma cells were sequenced. We were able to recover genomic information from both data sets using read mapping against a representative deltapolyomavirus genome (STL polyomavirus isolate HB201). Specifically, SRR1778915 had 170 reads that mapped to the deltapolyomavirus genome, covering 4551 bp of the 4775 bp genome. SRR8750456 had 32 reads that mapped, covering 1689 bp of the genome. We were also able to successfully assemble the higher-coverage data sets (SRR1778915) into viral genome fragmented across seven contigs.

We were not able to recover viral contigs from the other nine data sets in which we initially detected virus, despite using spacegraphcats **[13]** to try and pull viral regions directly from the assembly graph. Our read mapping approach didn't help to clarify things either. For seven out of the nine remaining samples, read mapping resulted in coverage of less than 100 bp of the viral genomes, which we do not consider to be legitimate. Two samples (both bulk RNA sequencing) had >100 bp of the query viral genomes, but were still extremely low-signal. Specifically, SRR8750473 had four reads that mapped across 194 bp of the human adenovirus type 7 genome. SRR8750456 had 12 reads that mapped across 189 bp of the human papillomavirus type 4 genome, and two reads that mapped across 130 bp of the human adenovirus type 7 genome.

In these cases, we could not validate our initial viral hits using read mapping or assembly, but they're also not clearly attributable to laboratory contamination. It is possible that a virus is truly present, but at an extremely low level, hindering our ability to accurately identify and classify it. However, these are extremely weak signals and it is difficult to say anything conclusive.

| Viral genome name | Viral genome accession | Inclusion criteria |
|---|---|---|
| Pbunalikevirus phiFenriz | GCA_002597305.1 | BLAST match (phage, included to capture background contamination and/or spurious read mapping) |
| Severe acute respiratory syndrome coronavirus 2 | GCF_009858895.2 | BLAST match (Included to capture background contamination and/or spurious read mapping) |
| STL polyomavirus, ViralProj186434 | GCF_000904055.1 | Match to VOG capsid (deltapolyoma virus) |
| Human papillomavirus 4, ViralProj15492 | GCF_000864845.1 | BLAST match |
| Alphapolyomavirus cardiodermae, ViralProj185188 | GCF_000903895.1 | Match to VOG capsid (alphapolyoma virus) |
| Human parvovirus B19, ViralProj14090 | GCF_000839645.1 | BLAST match |
| Human adenovirus 7, ViralProj15114 | GCF_000859485.1 | BLAST match |
| Human papillomavirus isolate HPV-msk_013 | MH777161.1 | BLAST match |

Table 2

**Representative viral genomes that we used for read mapping**.

We defined representative viral genomes for each putative hit (Table 1), and then used these for read mapping to try and recover additional viral sequences. We used BLAST of capsid reads to identify the best hit reference genome (BLAST match), and included the reference genome from which the initial capsid match originated (Match to VOG capsid).

## Intriguing polyomavirus hits

Out of all the possible viral hits that we found, we consider the polyomavirus-related hits to be most interesting from both a technical and a biological perspective. Polyomaviruses are small double-stranded, non-enveloped viruses with a 5 kb genome that infect humans, other mammals, and birds **[20]**.

On the technical side, we unambiguously detected deltapolyomavirus reads in the glioblastoma single-cell sequencing data set, and assembled one partial deltapolyomavirus genome. It is also intriguing that we detected alphapolyomavirus capsids in several other data sets, though we were never able to successfully assemble out a viral sequence to work with, and most of the capsid-encoding reads did not have any BLAST matches.

From a biological standpoint, polyomaviruses match several of the criteria we set out for ourselves at the beginning of this study. Polyomaviruses can establish asymptomatic infection in healthy individuals **[20]**, and at least one member of the polyomavirus family can transit the human blood–brain barrier to infect cells in the brain **[21][22][23]**.

The particular deltapolyomavirus that we found is very closely related to STL polyomavirus, which was initially discovered in fecal samples from children **[24]**. Seroprevalence studies show that STL polyomaviruses are highly prevalent in human populations **[25]**, causing asymptomatic (and potentially latent) infection. However, the significance of finding deltapolyomavirus sequences in two out of 98 glioblastoma cells from a single patient is difficult to assess.

From a therapeutic delivery standpoint, we were hoping to have enough hits to hint at the molecular basis of neurotropism rather than identifying a single virus. While that was not an outcome of the current project, with further validation our results could point toward exploring deltapolyomaviruses for delivery to brain, similar to previous work on betapolyomaviruses **[26]**.

While our biological findings would require much more data to properly contextualize, we consider our technical findings to be the most immediately useful takeaway from our study. At the outset, we were unsure what data types would be most amenable to viral sequence discovery. In our small study, we only analyzed a few single-cell WGS data sets and many more bulk data sets. However, we found the single-cell WGS data

sets to be much easier to work with computationally and ended up yielding our most promising results. As an added bonus, the multiple displacement amplification (MDA) whole-genome amplification step used in single-cell WGS is strongly biased toward capture of circular ssDNA and dsDNA viral genomes **[27]**. Though we will not be moving forward with this work in the near term, we consider our most useful finding from this study to be a much stronger understanding of how to implement a search for viral sequences in human sequencing data.

# Recommendations

Below, we share our learnings for others who want to use a similar approach. You will notice that our recommended approach does not include all the filtering steps that we used, but those can be added back in at the user's discretion depending on end goals.

- Target single-cell WGS data, as those gave us the most viral information while also being small and quick to process.
- Curate the viral capsid DB to make it as big and specific as possible.
- Integrate with the Amazon S3 SRA mirror, which should reduce sample download times, which was one of the longest steps.
- BLASTx the raw reads against the capsid database. Filter the results to reads that have > 90% identity and reads where both pairs map to a viral capsid.
- Assemble the reads and BLAST the product and any unassembled reads against NCBI nt to check if you recovered viral sequences. Remove all hits that have matches to the human or to laboratory vectors.
- For the BLAST results that pass this filter, pull down the viral genome that the results matched and map the library against this genome using Bowtie 2 with appropriate parameters. Be sure to include some controls here (non-integrating RNA virus for DNA samples, phage for samples with low bacterial burden, etc.) for viruses you don't expect to be present. Consider including the human genome as well. Also, filter BAMs to include mapped reads only at the write step to save on hard drive space.
- Filter to only properly mapped reads and use SAMtools `stats`, `flagstat`, `depth`, and `coverage` to summarize alignments.

- Filter to genome mappings that have a coverage of ~100 or 150 base pairs at a minimum. The more coverage the better, as this increases the workable information you will have. For the deltapolyomaviruses we found, we had coverage over 1,000 bp, which in one case corresponded to successful genome assembly. Having this level of genomic recovery is critical for assessing viral overlap between samples, opens up implementation of phylogenetic approaches, and can facilitate SNP comparison across samples.
- Use depth results to see where the reads mapped against the genome, which in cases of low coverage can help determine if the virus is legitimately present (reads spread across the genome) or if a small region of the genome is recruiting a lot of reads (less legitimate).
- Use SAMtools `stats` results to determine the "error" rate in mapping, which is useful for estimating variants without having to do a proper variant-calling workflow.
- Last, assemble any promising samples with substantial genomic coverage of a virus of interest from the full read set to recover viral contigs.

> Access all of our **code** and **data**, including our initial capsid search results, candidate viral capsid reads, capsid assemblies, read mapping results, and assemblies from mapped reads in this GitHub repository.

# Key takeaways

We implemented a quick, lightweight pipeline for scanning human RNA and DNA sequencing data sets for viral sequences, using capsids as viral marker genes. We were able to recover deltapolyomavirus sequences from two glioblastoma cells from a single patient, though we are unsure as to the significance of this finding. Overall, we found that WGS single-cell sequencing data sets were the easiest to work with for this type of viral discovery effort, suggesting this approach may be useful for discovery of ssDNA or dsDNA viruses in other contexts.

# Next steps

Based on our hit rates, we estimated that we would need single-cell DNA sequencing data from at least 100 cells from > 50 healthy individuals to assess if the viral signal we saw from deltapolyomaviruses was repeatable and present in non-glioblastoma states. Glioblastomas are thought to arise from astrocytes, so if we were to continue this work, we would start by doing targeted sequencing of astrocytes from healthy donors to see if we can detect more deltapolyomavirus genomes. To identify unique molecular adaptations underlying viral neurotropism, we would want an even larger sequencing cohort sampled across multiple cell and tissue types. This would be a substantial and expensive undertaking.

Alternatively, it may be possible the single-cell WGS data gave us better viral coverage because the multiple displacement amplification (MDA) step in single-cell WGS is over-amplifying the circular deltapolyomavirus genomes, driving them to relatively high copy number and letting us detect them. In this case, it may be possible to enrich for circular viral DNA sequences in bulk samples using this amplification approach, sacrificing the single-cell resolution but gaining throughput. One could also leverage related enrichment approaches for extrachromosomal circular DNA like mobilome-seq **[28]** to push the ratio of viral to host DNA higher. At the end of the day, we are icing this project due to strategic misalignment and technical limitations. These types of sequencing surveys are not efforts that Arcadia is going to invest in right now, and to our knowledge, these types and quantities of data are not currently available in publicly accessible databases.

---

# References


1 Gao G, Vandenberghe LH, Alvira MR, Lu Y, Calcedo R, Zhou X, Wilson JM. (2004). Clades of Adeno-Associated Viruses Are Widely Disseminated in Human Tissues. https://doi.org/10.1128/jvi.78.12.6381-6388.2004

2 Kalantar KL, Carvalho T, de Bourcy CFA, Dimitrov B, Dingle G, Egger R, Han J, Holmes OB, Juan Y-F, King R, Kislyuk A, Lin MF, Mariano M, Morse T, Reynoso LV,

Cruz DR, Sheu J, Tang J, Wang J, Zhang MA, Zhong E, Ahyong V, Lay S, Chea S, Bohl JA, Manning JE, Tato CM, DeRisi JL. (2020). IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. https://doi.org/10.1093/gigascience/giaa111

3 Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, Meyerson M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. https://doi.org/10.1038/nbt.1868

4 Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, Meyerson M. (2018). GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. https://doi.org/10.1093/bioinformatics/bty501

5 Mahmoudabadi G, Crasta S, Quake SR. (2022). Single Cell Transcriptomics Reveals the Hidden Microbiomes of Human Tissues. https://doi.org/10.1101/2022.10.11.511790

6 Irber L, Brooks PT, Reiter T, Pierce-Ward NT, Hera MR, Koslicki D, Brown CT. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. https://doi.org/10.1101/2022.01.11.475838

7 Chou S, Reiter T. (2024). Speeding up the quality control of raw sequencing data using seqqc, a Nextflow-based solution. https://doi.org/10.57844/ARCADIA-CXN6-CH62

8 Mangul S, Yang HT, Strauli N, Gruhl F, Porath HT, Hsieh K, Chen L, Daley T, Christenson S, Wesolowska-Andersen A, Spreafico R, Rios C, Eng C, Smith AD, Hernandez RD, Ophoff RA, Santana JR, Levanon EY, Woodruff PG, Burchard E, Seibold MA, Shifman S, Eskin E, Zaitlen N. (2018). ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. https://doi.org/10.1186/s13059-018-1403-7

9 Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Peña M, Korobeynikov A, Chikhi R, Babaian A. (2022). Petabase-scale sequence alignment catalyses viral discovery. https://doi.org/10.1038/s41586-021-04332-2

10 Buchfink B, Xie C, Huson DH. (2014). Fast and sensitive protein alignment using DIAMOND. https://doi.org/10.1038/nmeth.3176

11 Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. https://doi.org/10.1093/bioinformatics/btv033

12 Chikhi R, Limasset A, Medvedev P. (2016). Compacting de Bruijn graphs from sequencing data quickly and in low memory. https://doi.org/10.1093/bioinformatics/btw279

13 Brown CT, Moritz D, O'Brien MP, Reidl F, Reiter T, Sullivan BD. (2020). Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. https://doi.org/10.1186/s13059-020-02066-4

14 Wick RR, Schultz MB, Zobel J, Holt KE. (2015). Bandage: interactive visualization of *de novo* genome assemblies. https://doi.org/10.1093/bioinformatics/btv383

15 Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. https://doi.org/10.1038/nmeth.1923

16 Westbrook A, Ramsdell J, Schuelke T, Normington L, Bergeron RD, Thomas WK, MacManes MD. (2017). PALADIN: protein alignment for functional profiling whole metagenome shotgun data. https://doi.org/10.1093/bioinformatics/btx021

17 Francis JM, Zhang C-Z, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, Homer H, Haidar S, Blumenstiel B, Pedamallu CS, Ligon AH, Love JC, Meyerson M, Ligon KL. (2014). *EGFR* Variant Heterogeneity in Glioblastoma Resolved through Single-Nucleus Sequencing. https://doi.org/10.1158/2159-8290.cd-13-0879

18 Khrameeva E, Kurochkin I, Han D, Guijarro P, Kanton S, Santel M, Qian Z, Rong S, Mazin P, Sabirov M, Bulat M, Efimova O, Tkachev A, Guo S, Sherwood CC, Camp JG, Pääbo S, Treutlein B, Khaitovich P. (2020). Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. https://doi.org/10.1101/gr.256958.119

19 Breuss MW, Yang X, Schlachetzki JCM, Antaki D, Lana AJ, Xu X, Chung C, Chai G, Stanley V, Song Q, Newmeyer TF, Nguyen A, O'Brien S, Hoeksema MA, Cao B, Nott A, McEvoy-Venneri J, Pasillas MP, Barton ST, Copeland BR, Nahas S, Van Der Kraan L, Ding Y, Gleeson JG, Breuss MW, Yang X, Antaki D, Chung C, Averbuj D, Courchesne E, Ball LL, Roy S, Weinberger D, Jaffe A, Paquola A, Erwin J, Shin J, McConnell M, Straub R, Narurkar R, Mathern G, Walsh CA, Lee A, Huang AY, D'Gama A, Dias C, Maury E, Ganz J, Lodato M, Miller M, Li P, Rodin R, Borges-Monroy R, Hill R, Bizzotto S, Khoshkhoo S, Kim S, Zhou Z, Park PJ, Barton A, Galor A, Chu C, Bohrson C, Gulhan D, Lim Elaine, Lim Euncheon, Melloni G, Cortes I, Lee J, Luquette J, Yang L, Sherman M, Coulter M, Kwon M, Lee Semin, Lee Soo, Viswanadham V, Dou Y, Chess AJ, Jones A, Rosenbluh C, Akbarian S, Langmead B, Thorpe J, Cho S, Abyzov A, Bae T, Jang Y, Wang Y, Molitor C, Peters M, Gage FH, Wang M, Reed P, Linker S, Urban A, Zhou B, Pattni R, Zhu X, Amero AS, Juan D, Povolotskaya I, Lobon I, Moruno MS, Perez RG, Marques-Bonet T, Soriano E, Moran JV, Sun C, Flasch DA, Frisbie TJ, Kopera HC, Kidd JM, Moldovan JB, Kwan KY, Mills RE, Emery SB, Zhou W, Zhao X, Ratan A, Vaccarino FM, Cherskov A,

Jourdon A, Fasching L, Sestan N, Pochareddy S, Scuder S, Glass CK, Gleeson JG. (2022). Somatic mosaicism reveals clonal distributions of neocortical development. https://doi.org/10.1038/s41586-022-04602-7

20 Moens U, Calvignac-Spencer S, Lauber C, Ramqvist T, Feltkamp MCW, Daugherty MD, Verschoor EJ, Ehlers B. (2017). ICTV Virus Taxonomy Profile: Polyomaviridae. https://doi.org/10.1099/jgv.0.000839

21 Major EO, Amemiya K, Tornatore CS, Houff SA, Berger JR. (1992). Pathogenesis and molecular biology of progressive multifocal leukoencephalopathy, the JC virus-induced demyelinating disease of the human brain. https://doi.org/10.1128/cmr.5.1.49

22 Caldarelli-Stefano R, Vago L, Omodeo-Zorini E, Mediati M, Losciale L, Nebuloni M, Costanzi G, Ferrante P. (1999). Detection and typing of JC virus in autopsy brains and extraneural organs of AIDS patients and non-immunocompromised individuals. https://doi.org/10.3109/13550289909021994

23 Tan CS, Ellis LC, Wüthrich C, Ngo L, Broge TA Jr, Saint-Aubyn J, Miller JS, Koralnik IJ. (2010). JC Virus Latency in the Brain and Extraneural Organs of Patients with and without Progressive Multifocal Leukoencephalopathy. https://doi.org/10.1128/jvi.00609-10

24 Lim ES, Reyes A, Antonio M, Saha D, Ikumapayi UN, Adeyemi M, Stine OC, Skelton R, Brennan DC, Mkakosya RS, Manary MJ, Gordon JI, Wang D. (2013). Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing. https://doi.org/10.1016/j.virol.2012.12.005

25 Lim ES, Meinerz NM, Primi B, Wang D, Garcea RL. (2014). Common Exposure to STL Polyomavirus During Childhood. https://doi.org/10.3201/eid2009.140561

26 Ye D, Zimmermann T, Demina V, Sotnikov S, Ried CL, Rahn H, Stapf M, Untucht C, Rohe M, Terstappen GC, Wicke K, Mezler M, Manninga H, Meyer AH. (2021). Trafficking of JC virus-like particles across the blood–brain barrier. https://doi.org/10.1039/d0na00879f

27 Kim K-H, Bae J-W. (2011). Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. https://doi.org/10.1128/aem.00289-11

28 Lanciano S, Carpentier M-C, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, Ghesquière A, Panaud O, Mirouze M. (2017). Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. https://doi.org/10.1371/journal.pgen.1006630