Pavian walkthrough with metagenomics data from brain infections

July 29, 2016

General workflow

1) Import Kraken and Centrifuge report files
2) Look at the overall statistics across the samples
3) Compare the classification across samples
    - in specific domains
    - interactive table or heatmap
4) Zoom into one sample - flow chart or sunburst diagram
5) Zoom into one pathogen in one sample
    - view the alignment on the genome

1) Import Kraken and Centrifuge report files

The first step is loading data. Pavian currently supports results from Kraken (Wood and Salzberg 2014) and Centrifuge (Kim et al. 2016). It expects files ending with '.report.csv', generated with kraken-report or centrifuge-report. Before data is loaded, Pavian shows only a limited part of the interface (see Figure 1). Click 'Browse' or 'Choose files' (depending on browser) to select files to upload to Pavian. The uploaded files will be added to a new sample set with an auto-generated name. You can change the name of the samples and the sample set using the interface that appears, once the files are loaded.

In this walk through we use data from Salzberg et al. (2016). To load this data, click the 'Load example data' button. This study sequenced brain or spinal cord biopsies from 10 patients with suspected central nervous system (CNS) infections. Upon loading the data, the links to 'Results Overview', 'Comparison' and 'Sample' become available in the sidebar, and a table describes the loaded sample set (see Figure 2).

Brain biopsies data

Salzberg et al. (2016) used sequencing to detect the presence of pathogenic microbes in brain or spinal cord biopsies from 10 patients with neurological problems indicating possible infection, but for whom conventional clinical and microbiology studies yielded negative or inconclusive results. Direct DNA and RNA sequencing of brain tissue biopsies generated 8.3 million to 29.1 million sequence reads per sample.

Every samples is from a diseased patient. There are no healthy controls, and for all but one patient (PT8) there is only one sample available. The patients had different diseases; thus it was not expected that the same bug was the cause in the patients. The samples are their own control - ubiquitously present microbes probably are sequencing or laboratory contaminants.

The FASTQ files for this study are available at http://www.ncbi.nlm.nih.gov/bioproject/PRJNA314149. The reads were classified with Kraken. The Kraken report files from these samples are available at ADDLINK.

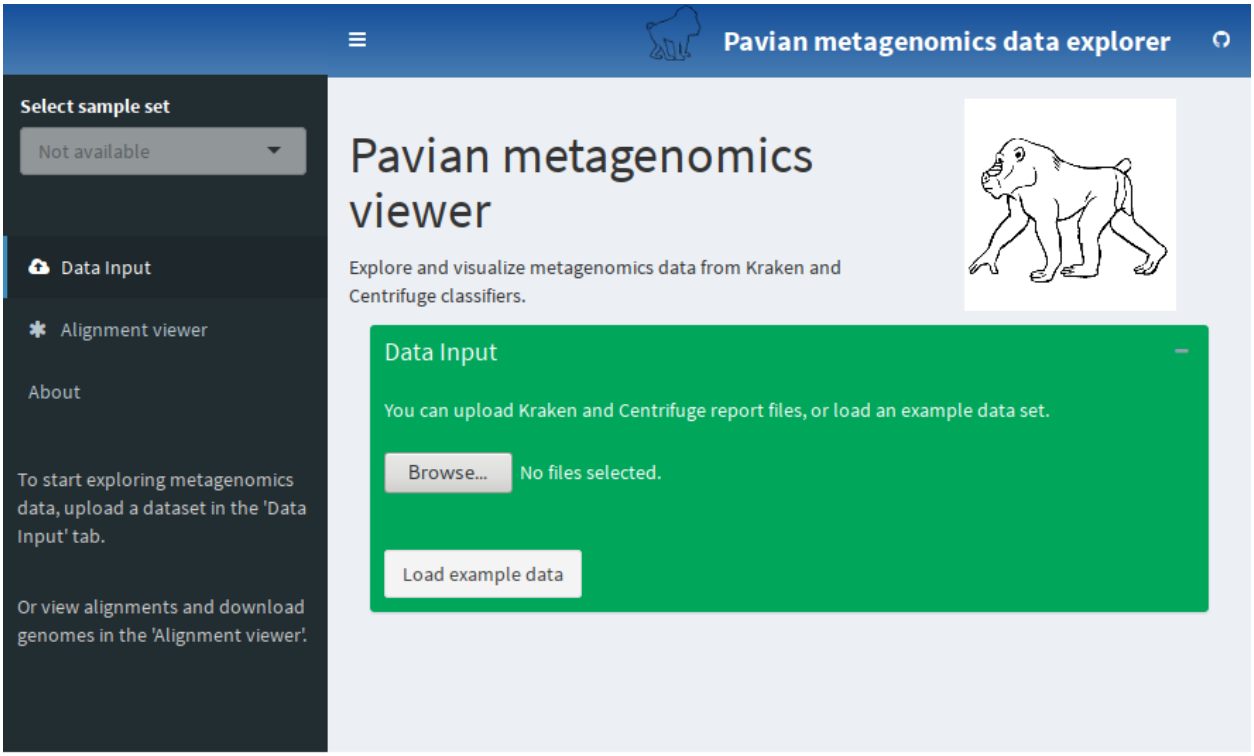

Figure 1: Pavian interface on start-up. Click 'Browse' to upload .report files, or 'Load example data' to load the brain biopsies dataset.

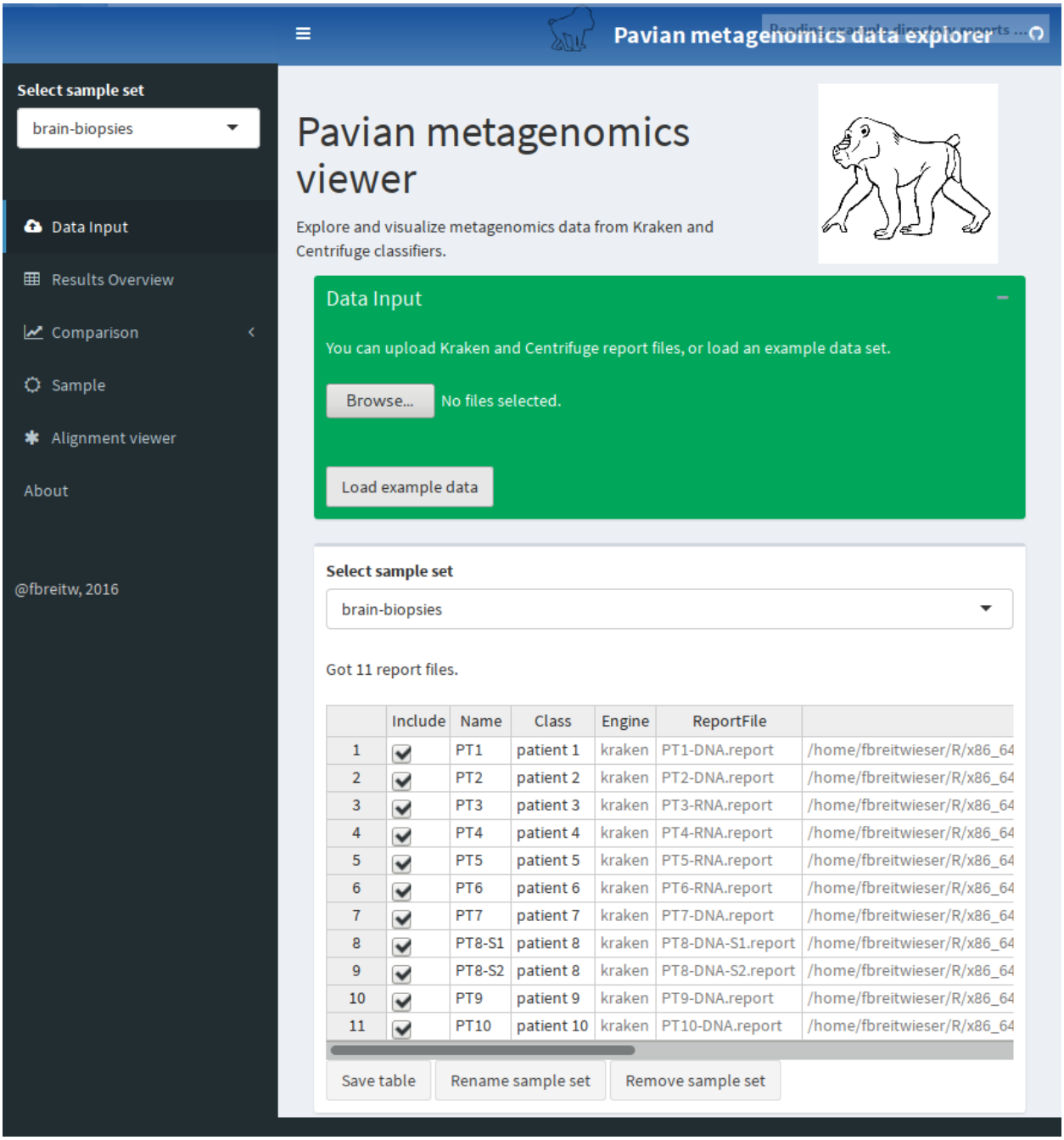

Figure 2: Pavian interface with the example data loaded. You can change the name of the sample set, or individual samples. Click 'Save table' after any change.

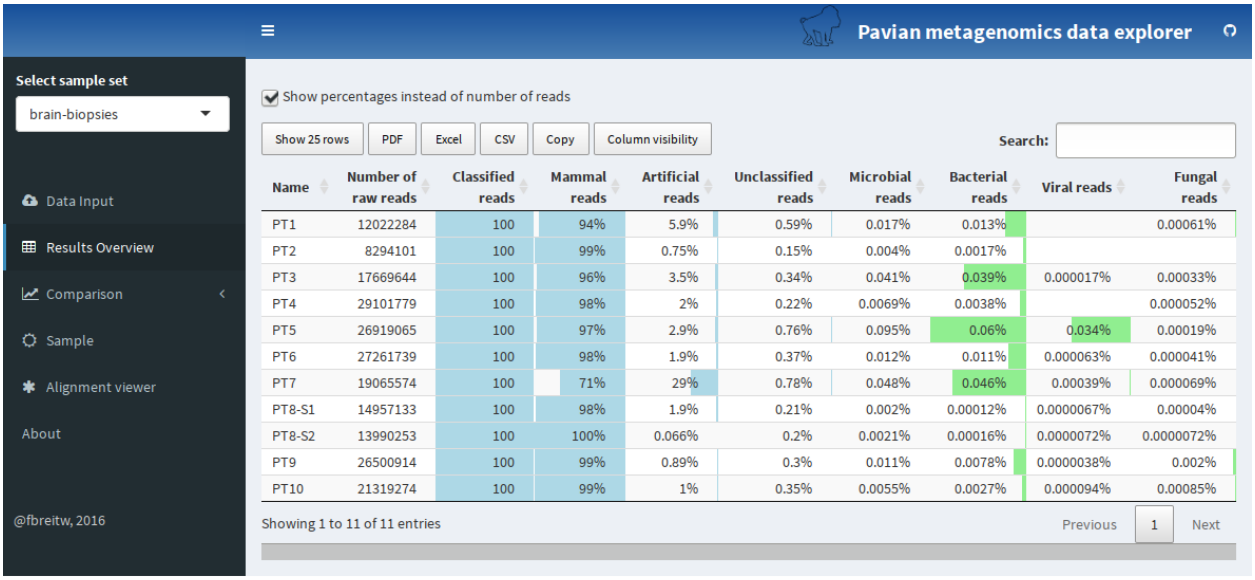

Figure 3: Results overview gives a quick view to the classification results.

2) Look at the overall statistics across the samples

Select 'Results Overview' in the sidebar to load the report files. After the files are loaded, the number of reads in the different samples, as well as the overall classification in different categories are displayed in a table (see Figure 3).

The samples of the brain biopsies dataset have between 8.3 and 29 million reads, most of them classified as 'mammal' (a 'mammal' is usually the host in these studies). There is a varying number of reads classified as artificial - usually below 5%, but two outliers in sample PT1 and PT7 with 5.9 and 29%, respectively. The number of reads classified as microbial is below .1% in all samples.

3) Compare the classification across samples

Click 'Comparison' and 'All data' to delve into the data (see Figure 4). The sample comparison view juxtaposes the identification results from the samples in a query-able table with taxa as rows and samples as columns. The table provides a visual guide to the values with inline bars, and the column taxid links to NCBI. You can decide to show percentages or z-scores, only show results at a certain taxon level, and filter uninteresting taxa. By default, the taxa 'Homo sapiens', 'synthetic construct' and 'unclassified' are filtered.

From this view, you can also select to see the data in the table in a heatmap representation (Figure 5). Note that the data displayed always correspond to the table - to show more rows in the heatmap, first select to show more rows in the table. Furthermore, you can look at how the samples are correlated with each other in 'Samples Clustering'.

In the brain biopsies dataset, we immediately see an outlier in sample PT5, which is the only one with a substantial number of JC polyomavirus reads. Note the table is rather wide with ten samples, and the fourth column provides an overview of the values in all columns. You can restrict the view to bacteria, viruses, or eukaryotic microbes by selecting the appropriate link in the sidebar.

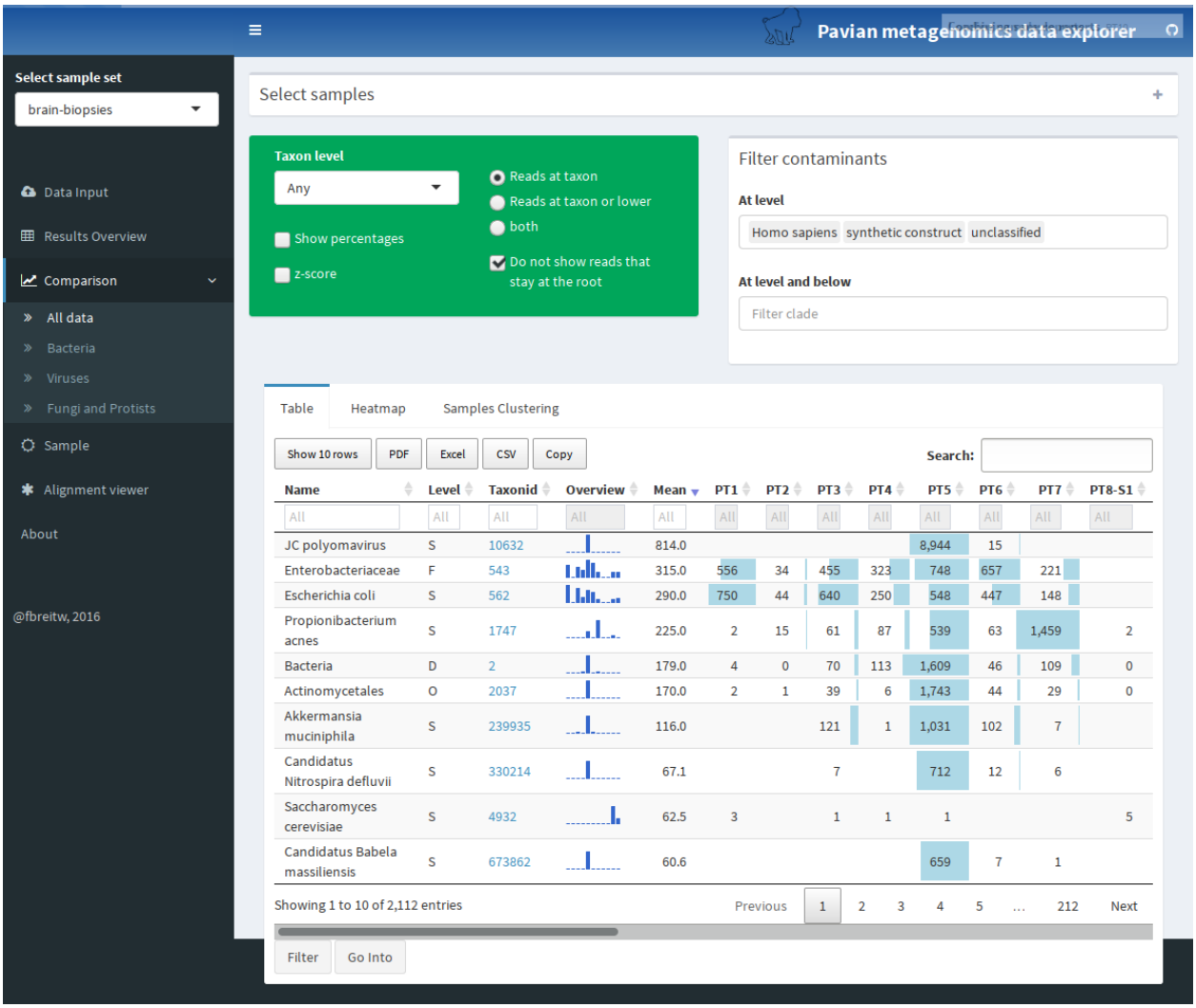

Figure 4: Sample comparision provides the data from all the samples in a sample set in a concise queryable table.

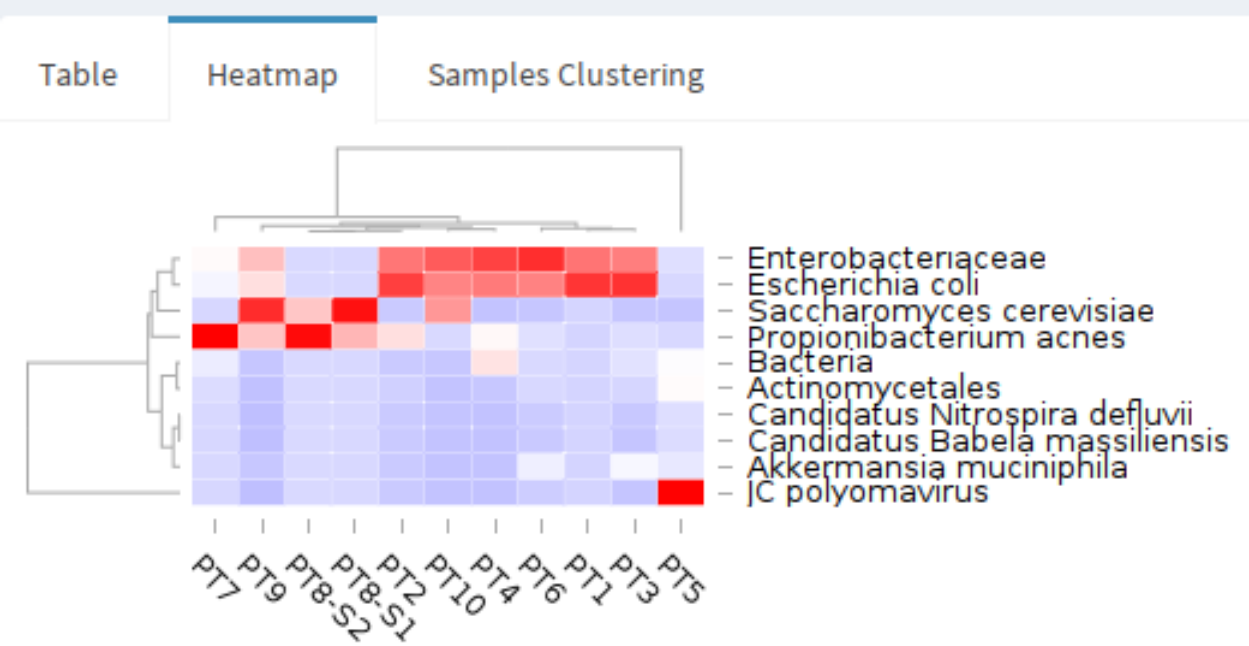

Figure 5: In the heatmap view, the data can be visualized in this intuitive way, clustering samples and taxa by abundances.

4) Zoom into one sample - flow chart or sunburst diagram

As JC Polyomavirus is a prime suspect in sample PT5, let's look further into the sample. Select 'Sample' in the sidebar, and then 'PT5' (Figure 6).

The sample comparison view allows a user to juxtapose the identification results from multiple samples (Figure 1C). The main view is an interactive table with taxa as rows and samples as columns. As the number of samples grows, this table can get very wide; thus to provide overview of the abundances in wide tables, the third column contains an inline barchart representation of the counts for a given species (row) across all samples. By default, read counts at all taxonomy levels are shown, but it is possible to only show specific taxonomical levels. The table can be queried and filtered.

The same results may also be inspected with an interactive heatmap, shown in Figure 1D. The samples and microbes may be clustered to group together samples with similar microbial profiles. Clicking on a row or column focuses in on one sample and microbe.

5) Zoom into one pathogen in one sample

JC polyomavirus has a high read count in sample PT5. However, do these reads cover the genome, or are they localized in a limited stretch of the genome? The coverage of the genome can provide a strong indication whether an assignment is spurious or not. A high read count for a particular species does not always mean that the microbe is present, as the reads may map to contaminated regions of a database genome or common sequences.

Click 'Alignment viewer' to get to Pavian's two functionalities to help in this in-depth investigation. First, it provides a convenient interface to the NCBI RefSeq assemblies (Kitts et al. 2016) in 'Download genomes for alignment'. There, go to 'viral' genome assemblies, and select 'Get assembly info'. Note that every time this is selected, all the associated assembly summary information from NCBI is downloaded. After a short

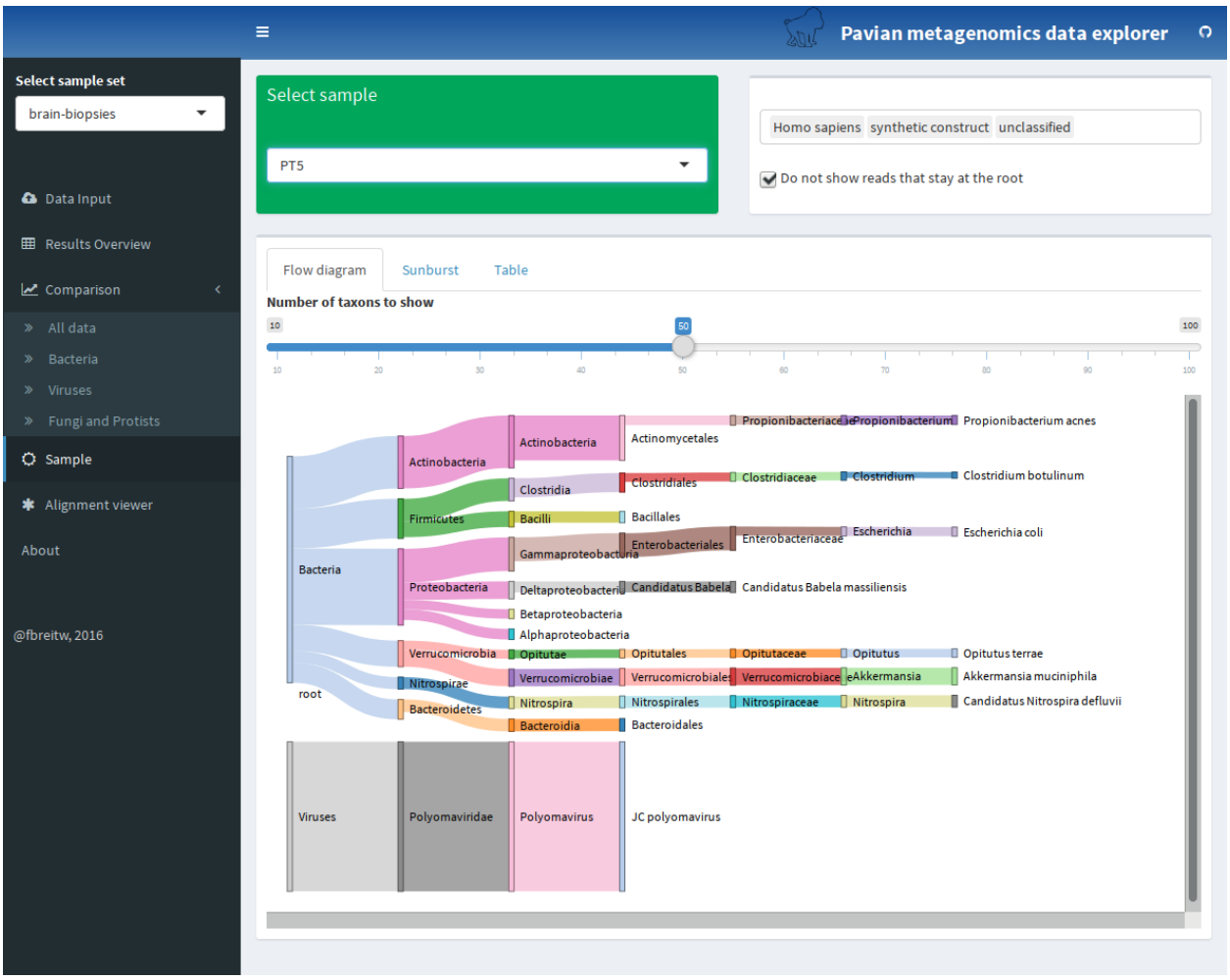

Figure 6: Sankey diagram of sample PT5 shows that JC Polyomavirus dominates the sample, and no other microbiota have been detected in a significant amount.

while a table appears with all viral RefSeq genome. Search for 'JC Pol' to find the reference genome for this species (see Figure 7). Note that the reference genome is the strain Mad1 originally uploaded in 1993. Once a row in the table is selected, a link to the genomic sequence (*_genomic.fna.gz) is displayed below, as well as Linux commands to copy and paste for downloading and building an index based on the genome. Note that the user has to then manually align the sample sequences and build a BAM file and BAM index.

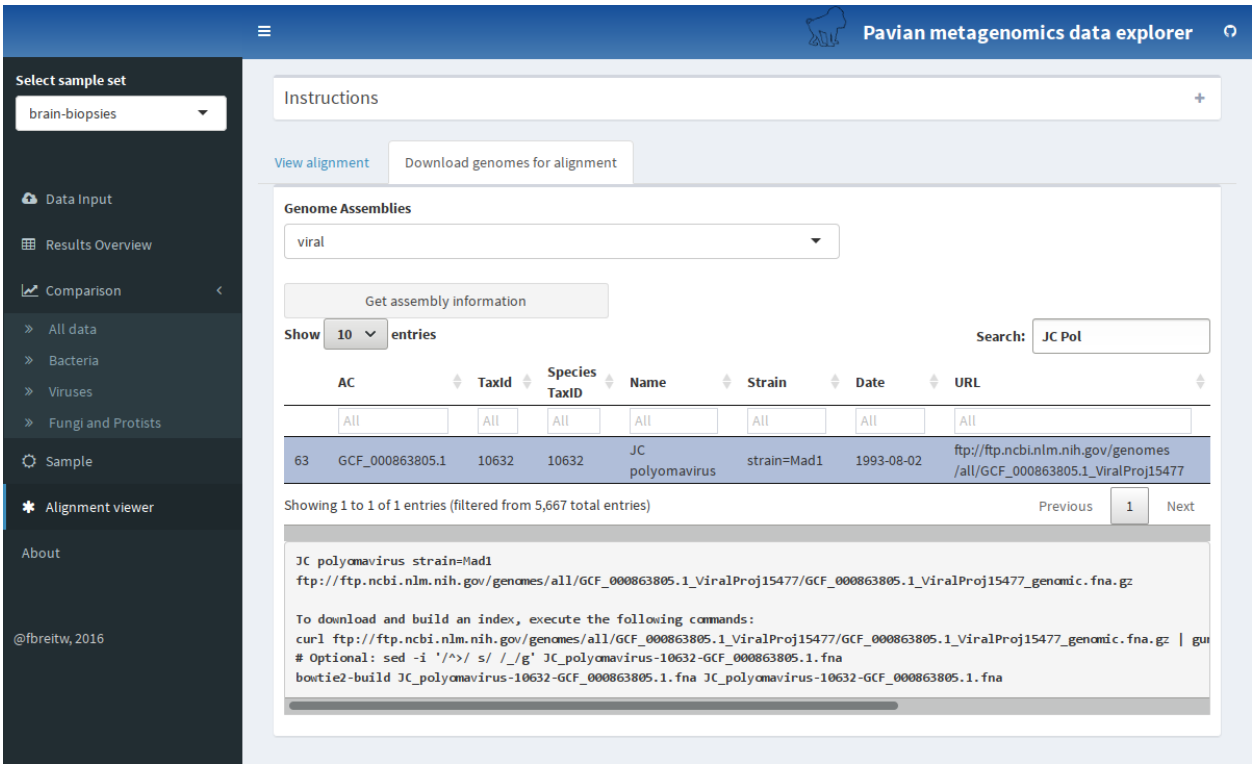

Figure 7: Pavian provides an intuitive interface to RefSeq genome assembly data

With a BAM file and BAM index (BAI) available, we can use the genome viewer. Click 'View alignment'. Normally, the two files have to be uploaded by clicking the 'Browse' button. Pavian includes the alignment of the sequences of PT5 to the reference genome of JC polyomavirus, and it is loaded by clicking 'Show alignment pileup' when no files are uploaded. Upon loading the files, the coverage over the genome is displayed (Figure 8). You can select a region to zoom into. For this genome we see that the reads align pretty randomly across the genome, but certain small region are not covered. As the reference genome is a strain that was isolated in 1993, it is very likely that this patient does have a different strain. However, it provides confidence that it is the same species.

Session information

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Arch Linux
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
```

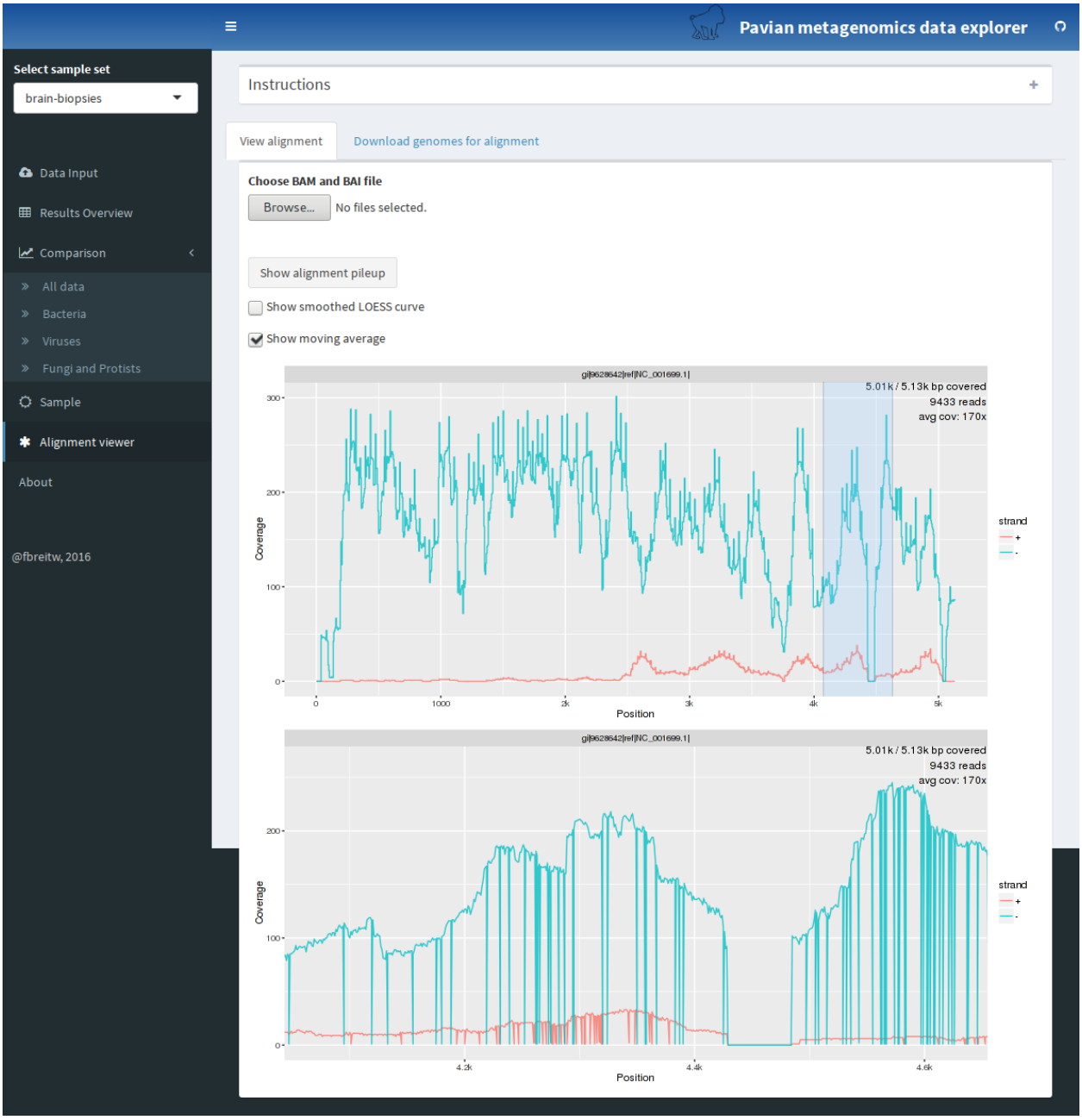

Figure 8: The alignment of the reads of sample PT5 to JC polyomavirus reference genome can be explored with the alignment viewer. There is a high coverage of most regions of the genome.

```
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] pavian_0.1.6
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.6          knitr_1.13           magrittr_1.5
##  [4] munsell_0.4.3        xtable_1.8-2         colorspace_1.2-6
##  [7] R6_2.1.2             stringr_1.0.0        plyr_1.8.4
## [10] tools_3.3.1          shinydashboard_0.5.1 grid_3.3.1
## [13] gtable_0.2.0         htmltools_0.3.5      yaml_2.1.13
## [16] digest_0.6.9         shiny_0.13.2         ggplot2_2.1.0
## [19] formatR_1.4          htmlwidgets_0.6.2    mime_0.4
## [22] evaluate_0.9         rmarkdown_1.0.9001   stringi_1.1.1
## [25] scales_0.4.0         rhandsontable_0.3.2  jsonlite_1.0
## [28] httpuv_1.3.3
```

References

Kim, Daehwan, Li Song, Florian P Breitwieser, and Steven L Salzberg. 2016. "Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences." *BioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/054965.

Kitts, Paul A., Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G. Smith, et al. 2016. "Assembly: A Resource for Assembled Genomes at NCBI." *Nucleic Acids Res* 44 (D1). National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.: D73–D80. doi:10.1093/nar/gkv1226.

Salzberg, Steven L., Florian P. Breitwieser, Anupama Kumar, Haiping Hao, Peter Burger, Fausto J. Rodriguez, Michael Lim, et al. 2016. "Next-Generation Sequencing in Neuropathologic Diagnosis of Infections of the Nervous System." *Neurol Neuroimmunol Neuroinflamm* 3 (4). ter Science,; Biostatistics (S.L.S.), Johns Hopkins University, Baltimore, MD.: e251. doi:10.1212/NXI.0000000000000251.

Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biol* 15 (3): R46. doi:10.1186/gb-2014-15-3-r46.