Pavian walkthrough with metagenomics data from brain infections

July 28, 2016

General workflow

1) Import Kraken and Centrifuge report files (a sample set)
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

Import Kraken and Centrifuge report files

The first step is loading data. Pavian expects files ending with '.report.csv', generated with kraken-report or centrifuge-report, resp. Click 'Choose files' 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 walkthrough 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 1).

Brain biopsies data

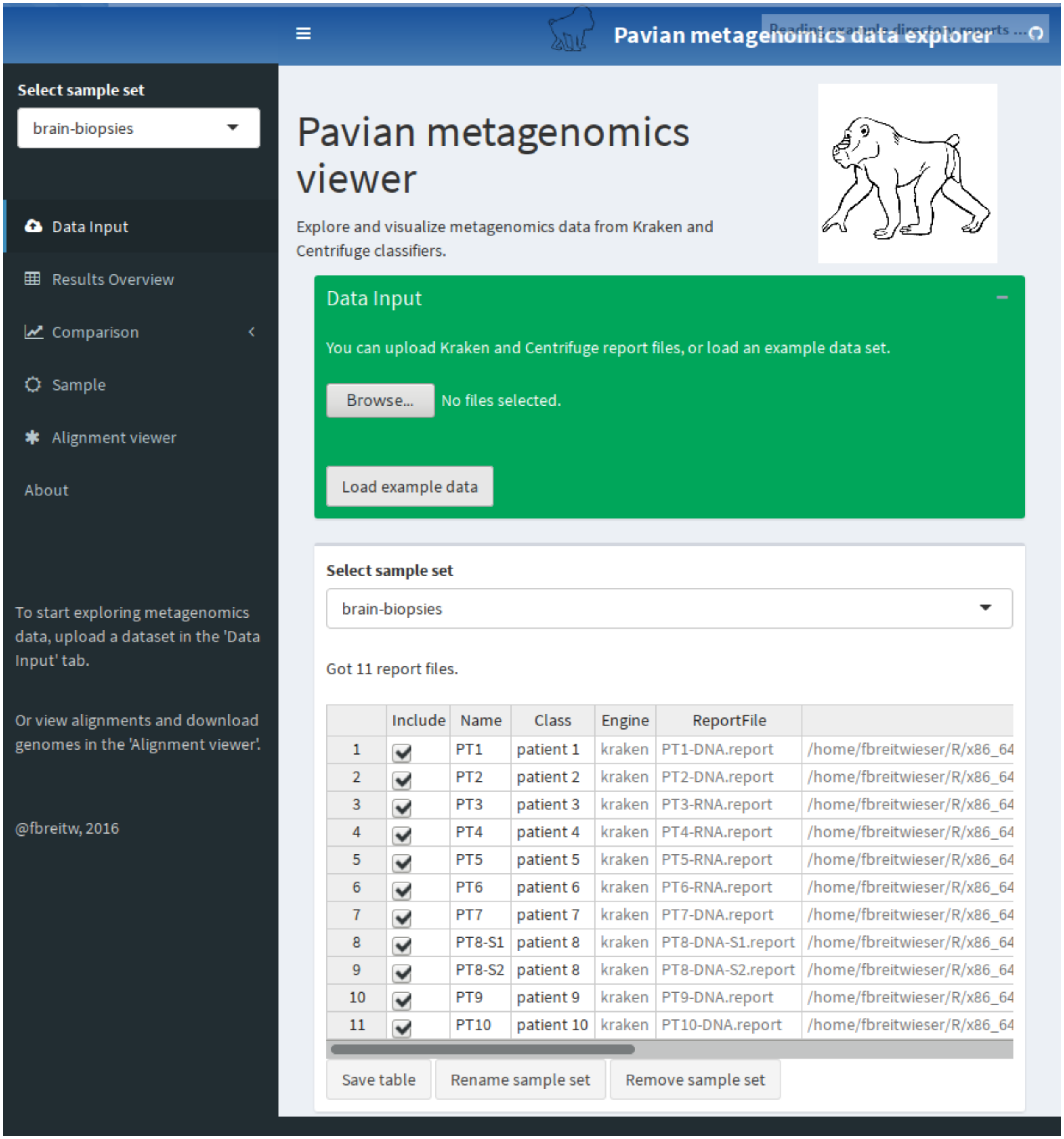
Salzberg et al. (2016) used sequencing to detect the presence of pathogenic microbes in brain or spinal cord biopsies from 10 patients with neurologic problems indicating possible infection, but for whom conventional clinical and microbiology studies yielded negative or inconclusive results. Direct DNA 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 were disregarded as sequencing or laboratory contaminants. Most of the reads in each sample are human, or from common contaminants - bacteria that always appear but are unlikely to cause the infection, such as P. acnes.

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 sampels are available at ADDLINK.

Look at the overall statistics across the samples

Select 'Results Overview' in the sidebar to load the report files. This will show a vire with the number of reads in the different samples, as well as the overall classification in different categories (see Figure 2).



Pavian metagenomics viewer

Explore and visualize metagenomics data from Kraken and Centrifuge classifiers.

You can upload Kraken and Centrifuge report files, or load an example data set.

To start exploring metagenomics data, upload a dataset in the 'Data Input' tab.

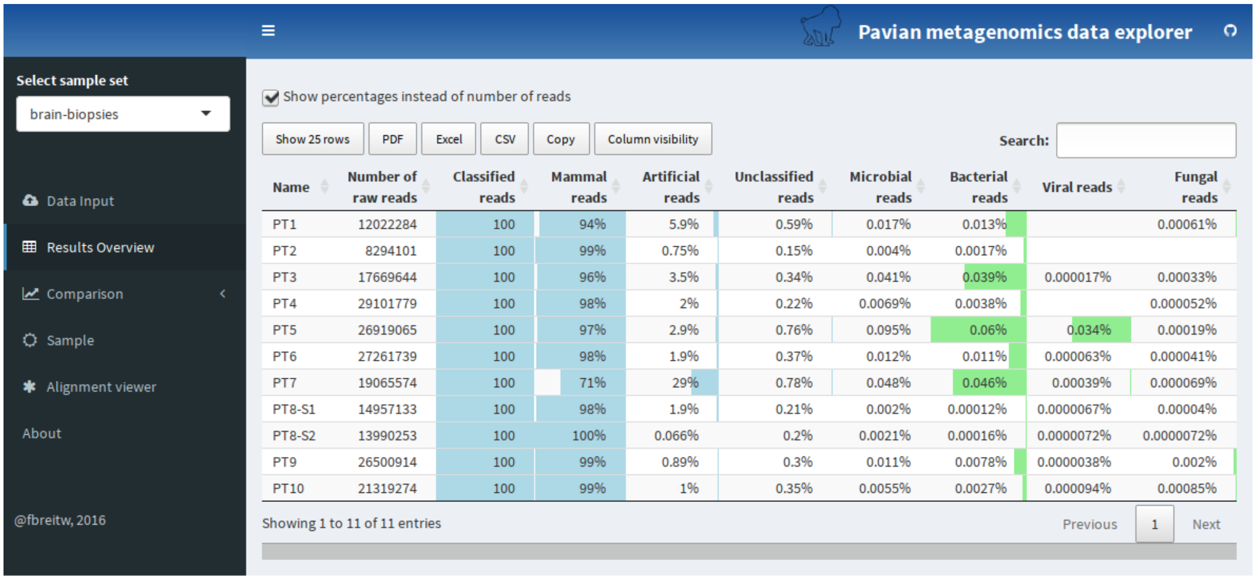
Or view alignments and download genomes in the 'Alignment viewer'.

@fbreitw, 2016

Got 11 report files.

	Include	Name	Class	Engine	ReportFile	
1	☑	PT1	patient 1	kraken	PT1-DNA.report	/home/fbreitwieser/R/x86_64
2	☑	PT2	patient 2	kraken	PT2-DNA.report	/home/fbreitwieser/R/x86_64
3	☑	PT3	patient 3	kraken	PT3-RNA.report	/home/fbreitwieser/R/x86_64
4	☑	PT4	patient 4	kraken	PT4-RNA.report	/home/fbreitwieser/R/x86_64
5	☑	PT5	patient 5	kraken	PT5-RNA.report	/home/fbreitwieser/R/x86_64
6	☑	PT6	patient 6	kraken	PT6-RNA.report	/home/fbreitwieser/R/x86_64
7	☑	PT7	patient 7	kraken	PT7-DNA.report	/home/fbreitwieser/R/x86_64
8	☑	PT8-S1	patient 8	kraken	PT8-DNA-S1.report	/home/fbreitwieser/R/x86_64
9	☑	PT8-S2	patient 8	kraken	PT8-DNA-S2.report	/home/fbreitwieser/R/x86_64
10	☑	PT9	patient 9	kraken	PT9-DNA.report	/home/fbreitwieser/R/x86_64
11	☑	PT10	patient 10	kraken	PT10-DNA.report	/home/fbreitwieser/R/x86_64

Figure 1: Pavian interface with the sample files loaded.



Name	Number of raw reads	Classified reads	Mammal reads	Artificial reads	Unclassified reads	Microbial reads	Bacterial reads	Viral reads	Fungal reads
PT1	12022284	100	94%	5.9%	0.59%	0.017%	0.013%		0.00061%
PT2	8294101	100	99%	0.75%	0.15%	0.004%	0.0017%		
PT3	17669644	100	96%	3.5%	0.34%	0.041%	0.039%	0.000017%	0.00033%
PT4	29101779	100	98%	2%	0.22%	0.0069%	0.0038%		0.000052%
PT5	26919065	100	97%	2.9%	0.76%	0.095%	0.06%	0.034%	0.00019%
PT6	27261739	100	98%	1.9%	0.37%	0.012%	0.011%	0.000063%	0.000041%
PT7	19065574	100	71%	29%	0.78%	0.048%	0.046%	0.00039%	0.000069%
PT8-S1	14957133	100	98%	1.9%	0.21%	0.002%	0.00012%	0.0000067%	0.00004%
PT8-S2	13990253	100	100%	0.066%	0.2%	0.0021%	0.00016%	0.0000072%	0.0000072%
PT9	26500914	100	99%	0.89%	0.3%	0.011%	0.0078%	0.0000038%	0.002%
PT10	21319274	100	99%	1%	0.35%	0.0055%	0.0027%	0.000094%	0.00085%

Showing 1 to 11 of 11 entries

Figure 2: Results overview.

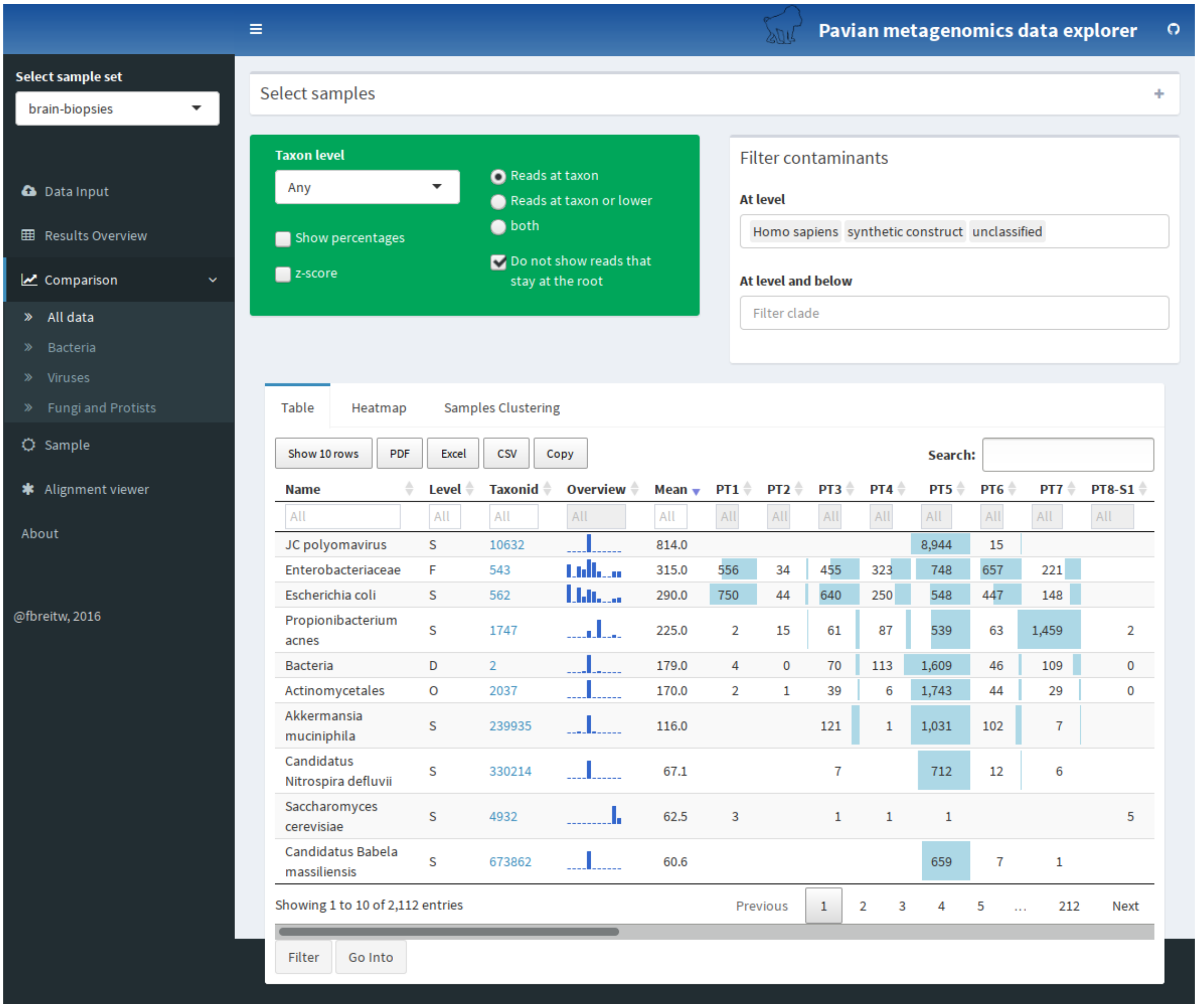
The samples of the brain biopsies dataset have between 8.3 and 29 million reads, most of them classified as 'mammal' (a 'mammal' is usually the host in these studies). There is a varying number of reads classifie as artificial - usually below 5%, but two outliers in sample PT1 and PT7 with 5.9 and 29%,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

Compare the classification across samples

Click 'Comparison' and 'All data' to delve into the data (see Figure 3). The sample comparison view juxtaposes the identification results from the samples in a queryable 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. 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'.

For the brain biospies dataset, the table is rather wide with ten samples. The fourth column therefore provides an overview of the values in the columns. By default, the rows are sorted by to the mean over all samples. In this sample set, we immediately see an outlier in sample PT5, which is the only one with a substantial number of JC polyomavirus reads.



Taxon level

Any ▾

● Reads at taxon
○ Reads at taxon or lower
○ both

□ Show percentages
□ z-score

☑ Do not show reads that stay at the root

Filter contaminants

At level: Homo sapiens, synthetic construct, unclassified

At level and below: Filter clade

Name	Level	Taxonid	Overview	Mean	PT1	PT2	PT3	PT4	PT5	PT6	PT7	PT8-S1
All	All	All	All	All	All	All	All	All	All	All	All	All
JC polyomavirus	S	10632		814.0					8,944	15		
Enterobacteriaceae	F	543		315.0	556	34	455	323	748	657	221	
Escherichia coli	S	562		290.0	750	44	640	250	548	447	148	
Propionibacterium acnes	S	1747		225.0	2	15	61	87	539	63	1,459	2
Bacteria	D	2		179.0	4	0	70	113	1,609	46	109	0
Actinomycetales	O	2037		170.0	2	1	39	6	1,743	44	29	0
Akkermansia muciniphila	S	239935		116.0			121	1	1,031	102	7	
Candidatus Nitrospira defluvii	S	330214		67.1			7		712	12	6	
Saccharomyces cerevisiae	S	4932		62.5	3		1	1	1			5
Candidatus Babela massiliensis	S	673862		60.6					659	7	1	

Showing 1 to 10 of 2,112 entries

Figure 3: Sample comparision.

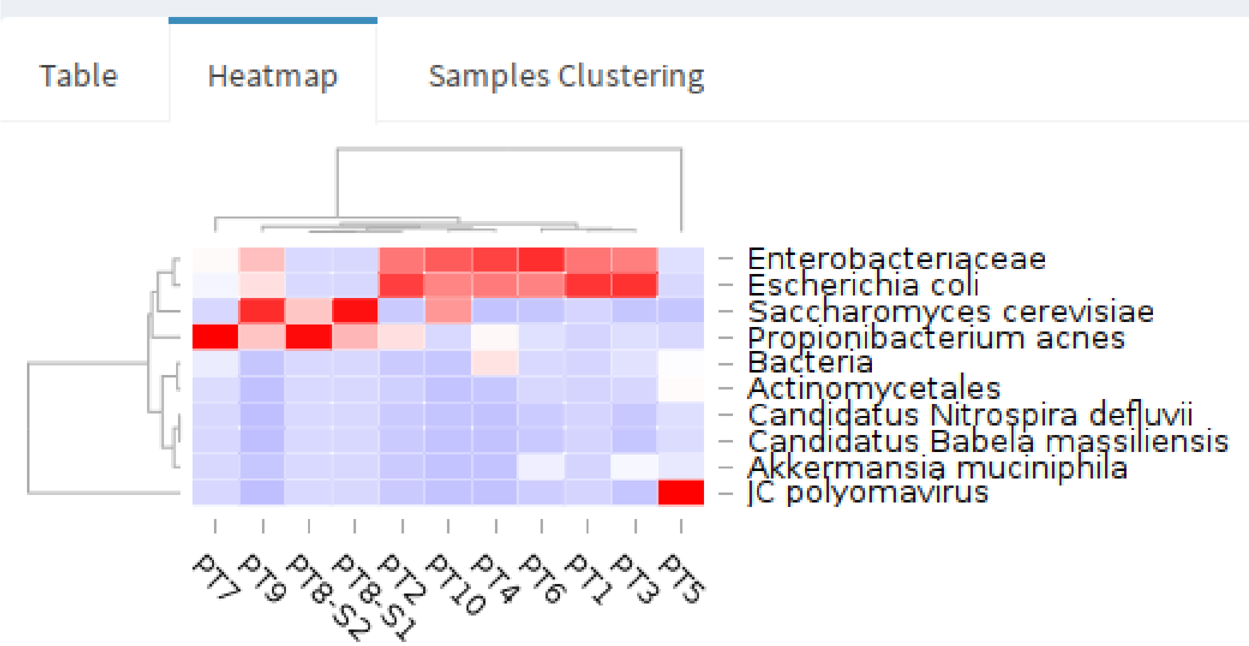


Figure 4: Sample comparision.

Zoom into one sample - flow chart or sunburst diagram

Zoom into one pathogen in one sample

```
- view the alignment on the genome
```

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.