

Pavian Walkthrough

Florian P Breitwieser

2016-11-11

Contents

Introduction	1
1) Data Input - Import Kraken and Centrifuge report files	2
2) Results Overview - Look at the overall statistics across the samples	2
3) Comparison - Compare the classification across samples	3
4) Sample - Zoom into one sample with a Sankey diagram	3
5) Alignment viewer - Zoom into one pathogen in one sample	5
References	7

Introduction

Pavian (‘pathogen visualization and more’) is a interactive web tool to explore at metagenomics classification results. Pavian has been developed with this clinical metagenomics in mind. However, it should be useful to anyone who wants to analyze and visualize their metagenomics data. This document gives a walkthrough of the interface by looking at the provided example data from brain biopsies. The images have been generated using RSelenium.

Pavian metagenomics data explorer

Data Input

Pavian supports Kraken-style and MetaPhlAn report files. For Centrifuge, the script 'centrifuge-kreport' creates a report in the correct format (in the future, the standard Centrifuge report will be in the right format)! Two example datasets are available: *brain-biopsies* and *hmp-stool*. The first set is from [ten patients with suspected infections of the nervous system](#), analyzed with Kraken. The second set is sequenced stool from the [Human Microbiome Project](#), analyzed with MetaPhlAn. Note that for MetaPhlAn, the 'reads' are actually percentages.

● Upload files ● Use data on server ● Load example data

Available sample sets

● example-data/brain-biopsies
● example-data/hmp

Rename selected sample set Remove selected sample set

Got 11 report files.

	FormatOK	Include	Name	Class	Engine	ReportFile	ReportFilePath
1	true	<input checked="" type="checkbox"/>	PT1	patient 1	kraken	PT1-DNA.report	/home/fbreitwieser/R/x86_64-pc-linux-gnu-library/3.3/pavian/shinyapp/ex
2	true	<input checked="" type="checkbox"/>	PT2	patient 2	kraken	PT2-DNA.report	/home/fbreitwieser/R/x86_64-pc-linux-gnu-library/3.3/pavian/shinyapp/ex
3	true	<input checked="" type="checkbox"/>	PT3	patient 3	kraken	PT3-RNA.report	/home/fbreitwieser/R/x86_64-pc-linux-gnu-library/3.3/pavian/shinyapp/ex
4	true	<input checked="" type="checkbox"/>	PT4	patient 4	kraken	PT4-RNA.report	/home/fbreitwieser/R/x86_64-pc-linux-gnu-library/3.3/pavian/shinyapp/ex

Figure 1: Pavian interface with the example data loaded. You can change the name of the sample set, or the name that appears in the interface for individual samples. Click ‘Save table’ after any change.

1) Data Input - 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 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 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.

2) Results Overview - Look at the overall statistics across the samples

Name	Number of raw reads	Classified reads	Mammal reads	Artificial reads	Unclassified reads	Microbial reads	Bacterial reads	Viral reads	Fungal reads
PT1	12,022,284	100%	94.09%	5.894%	0.5893%	0.01671%	0.01253%	0%	0.0006108%
PT2	8,294,101	100%	99.25%	0.7503%	0.1464%	0.003985%	0.001702%	0%	0%
PT3	17,669,644	100%	96.49%	3.464%	0.3398%	0.04097%	0.03895%	0.00001704%	0.0003294%
PT4	29,101,779	100%	97.98%	2.009%	0.2182%	0.006901%	0.003767%	0%	0.00005166%
PT5	26,919,065	100%	96.96%	2.945%	0.7622%	0.0953%	0.06006%	0.03352%	0.0001872%

Figure 2: Results overview gives a quick view to the classification results.

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 2).

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) Comparison - 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 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.

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.

Name	Level	Taxonid	Overview	Mean	PT1	PT2	PT3	PT4	PT5	PT6	PT7	PT8-S1
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		262.0	656	35	612	222	494	408	130	
Bacteria	D	2		179.0	4	0	70	113	1,608	46	109	0
Propionibacterium acnes	S	1747		175.0	2	10	50	70	459	57	1,078	1
Actinomycetales	O	2037		165.0	1	1	36	4	1,712	36	20	0

Figure 3: Sample comparison provides the data from all the samples in a sample set in a concise queryable table.

4) Sample - Zoom into one sample with a Sankey 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 5).

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.

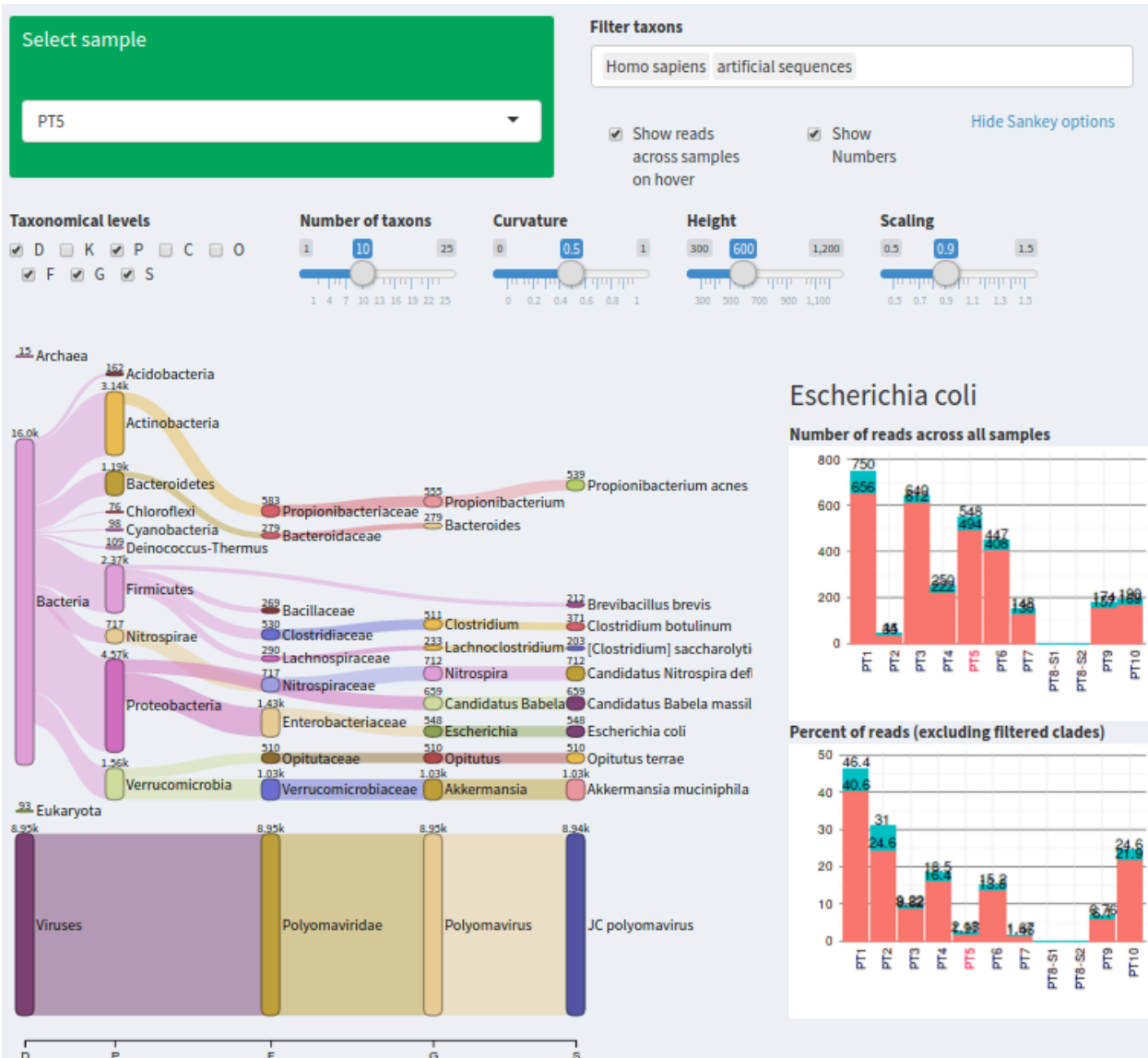


Figure 4: Sankey diagram of sample PT5 shows that JC Polyomavirus dominates the sample, and no other microbiota have been detected in a significant amount. When hovering nodes, the rea distribution for this taxon in other samples is displayed on the right. Here, *Escherichia coli* is selected.

5) Alignment viewer - 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 while a table appears with all viral RefSeq genome. Search for ‘JC Pol’ to find the reference genome for this species (see Figure 6). 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.

Get assembly information

Show 10 entries Search: JC

AC	Taxid	Species TaxID	Name	Strain	Date	URL	
5920	GCF_000863805.1	10632	1891763	JC polyomavirus	strain=Mad1	1993-08-02	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GC

Showing 1 to 1 of 1 entries (filtered from 6,019 total entries) Previous 1 Next

JC polyomavirus strain=Mad1

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/863/805/GCF_000863805.1_ViralProj15477/GCF_000863805.1_ViralProj15477_genomic.fna.gz

To download and build an index, execute the following commands:

```
curl
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/863/805/GCF_000863805.1_ViralProj15477/GCF_000863805.1_ViralProj15477_genomic.fna.gz
| gunzip -c > JC_polyomavirus-10632-GCF_000863805.1.fna # Optional: sed -i '/^>/ s/ /_g' JC_polyomavirus-10632-
GCF_000863805.1.fna bowtie2-build JC_polyomavirus-10632-GCF_000863805.1.fna JC_polyomavirus-10632-GCF_000863805.1.fna
```

Figure 5: 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 7). 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. Still, the coverage of this genome provides confidence that it is the same species.



Figure 6: 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.

References

- Kim, Daehwan, Li Song, Florian P Breitwieser, and Steven L Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research*. 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.