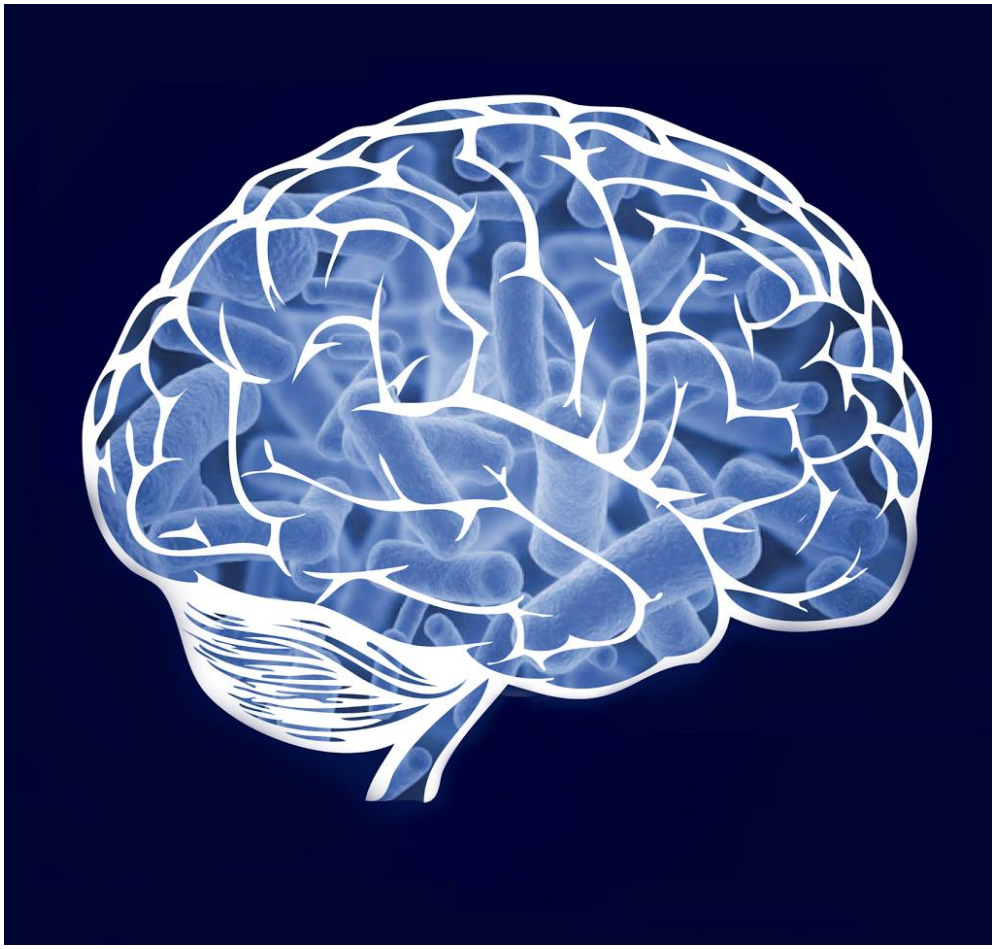
Identifying pathogens in human post mortem brain

Identifying pathogenic and commensal microbes from central nervous system sequencing studies



[35]

Iris Gorter
335213
Bio-informatica
Universitair Medisch Centrum Groningen
BSc Marissa Dubbelaar
Prof. dr. Jon Laman

Identifying pathogens in human post mortem brain

Identifying pathogenic and commensal microbes from central nervous system sequencing studies

Identifying pathogens in human post mortem brain

Identifying pathogenic and commensal microbes from central nervous system sequencing studies

Iris Gorter
335213
Bio-informatics
Institute of Life Science and Technology
MSc Fenna Feenstra
MSc Martijn Herber
BSc Marissa Dubbelaar
Prof. dr. Jon Laman
17-01-2017

Abstract

The immune system responds quickly when it is activated by an infection. The immune response activated several cells, like neutrophils, macrophages and microglia. Microglia are prime candidates of the CNS that can act like antigen presenting cells and have a phagocytic activity. During infection, microglia get activated and have increased proliferation, motility and phagocytic activity. Since microglia are APC's and phagocytic cells, they can contain parts of pathogens. When microglia are sequenced in NGS studies, DNA/RNA from pathogens can be detected.

A batch effect was observed in an early analysis of the data from Galatro and Holtman et al. This batch effect showed a separate clustering of the Brazilian and Dutch cohort. One of the hypotheses was that this batch could be explained by the exposure of different pathogens. Therefore, the aim of the project is to identify which pathogens can be found in different samples and determine if there is a difference between the two cohorts.

Another aim is to determine if there is any possibility of contamination of the samples.
Due to contamination meaning, unhygienic circumstances or the use of laboratory reagents it is possible that samples contain small amounts of pathogens.

During this project, samples of the study of Galatro and Holtman et al., were aligned to the human genome with Bowtie2. Sequences that failed to align to the human genome were identified as ‘possible pathogen sequences’. These sequences were aligned against an index of pathogens known to cause disease in brain. Approximately 97-99% of the total base pairs failed to align to the human genome are still unidentified. Next, the amount of base pairs per pathogen was used to visualize the distribution of pathogens between cohorts. A notable result is a large amount of *E. coli* in the Brazilian microglia samples.

The batch effect observed in the initial analysis was not observed during this analysis and could not be explained by the expression of the identified pathogens. There is no difference between the cohorts that can explain the batch effect.

Abbreviations

APC	Antigen presenting cells
BAM	Binary alignment map
BBB	Blood-brain barrier
BED	Browser extensible data
BP	Base pairs
CCC	Comprehensive cancer center
CNS	Central nervous system
CPM	Counts per million
DNA	Deoxyribonucleic acid
ERIBA	European Research Institute for the Biology of Aging
FACS	Fluorescence-activated cell sorting
GPTC	Groningen Proton Therapy Centre
GSVA	Gene Set Variation Analysis
MHC	Major histocompatibility complex
NGS	Next generation sequencing
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SAM	Sequence alignment map
UMCG	University Medical Center Groningen

Organization

The university medical center Groningen (UMCG) is one of the most specialized hospitals in the Netherlands. It is the biggest employer of the north in the Netherlands with approximately 13.000 employees. It contains several specialized departments, like the UMCG comprehensive cancer center (UMCG CCC) and a transplantation centre. In the end of 2017, the UMCG will open one of the first proton therapy centre in the Netherlands, the GPTC (Groningen proton therapy centre) [28]. Here, cancer patients get treated with proton therapy instead of photon therapy.
The UMCG facilitates the European Institute of Biology and Aging (ERIBA), where the focus lies on the molecular mechanisms that are responsible for aging[29].
The UMCG corporates closely with the faculty of Medical Sciences of the University of Groningen to focus on education and scientific research where approximately 4000 students start a medical course each year.

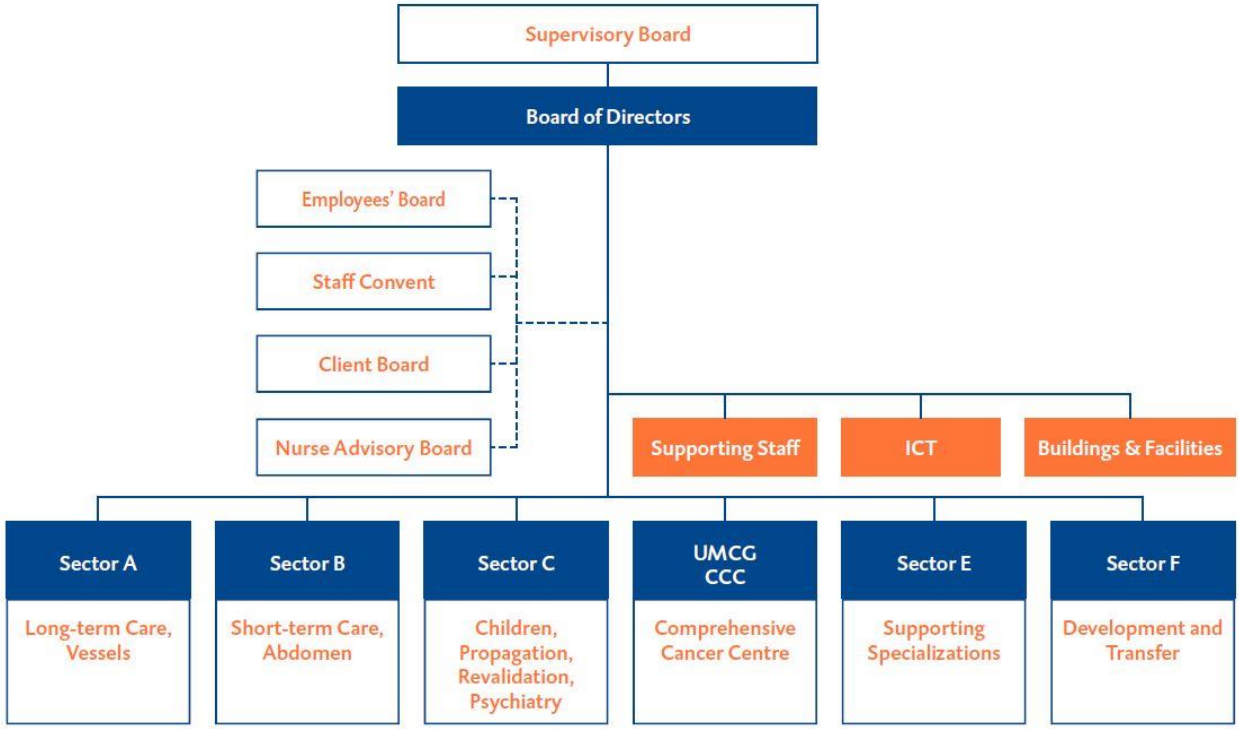


Figure 1. Organogram of the UMCG [12]

Within the UMCG there is a sector subdivision, one of them is sector F. This is where the department of Neuroscience is located, focusing on development and transferring knowledge. The main point of investigation of the Department of Neuroscience is the central nervous system (CNS). Topics of interest such as inflammatory neurodegeneration and CNS ageing.

The department is subdivided into four research groups: Anatomy, Cognitive Neuropsychiatry, Medical Physiology and the Neuroimaging Centre. The research group of Medical Physiology focuses on three research aspects:

1) Neuron-glia signaling in healthy CNS conditions.
2) The function of ion channels in neurodegenerative diseases.
3) Human motor neuron physiology.

Table of Contents

1. Introduction

The immune system might give a heavy reaction when it is activated by an infection from for example a virus. This immune response can be identified by the activation of several cells, like neutrophils, macrophages and microglia.
Every organ in the body responds to infection, the central nervous system (CNS) is no exception. Despite the series of tissue barriers that pathogens need to cross in order to access the brain and spinal cord.
An accepted hypothesis explains that the blood brain barrier (BBB) protects the CNS from pathogens like viruses and bacteria. The BBB was discovered by injecting certain dyes into the human body. The color affected the organs, but he brain remained unstained. This effect was also observed vice versa. The hypothesis describes the BBB as a barrier that prevents the escape and entry of dye from cerebral blood vessels. The BBB is composed of specialized brain microvascular endothelial cells, that let molecules as water and oxygen diffuse through this layer. A biochemical homeostasis is maintained in the CNS by the regulation of passage of molecules [1][2][3].

However, microbe compounds can be carried into the CNS by phagocytic cell types such as neutrophils, macrophages and dendritic cells. Therefore, the microbe does not need to replicate in order to be present among the genetic material of mammals. Microglia are prime candidates of the CNS that might contain microbial DNA/RNA. Microglia can act as antigen presenting cells (APC) and have a phagocytic activity. Other cells like neurons, astrocytes and oligodendrocytes have phagocytic activity as well, and can be targeted for infection.
A pathogen need certain characteristics to become a CNS pathogen:

1. A mechanism to entry the CNS
2. Survival within host’s peripheral immune cells
3. Penetration of barriers

The host needs to be colonized before the pathogen can invade the bloodstream. When a pathogen successfully enters the bloodstream, it still needs to cross several barriers, like the BBB, before it can invade the CNS. There are several approaches for a pathogen to enter the CNS. The first approach is that the pathogen can adhere to host receptors that can enter the CNS. The second approach is the Trojan Horse mechanism. A pathogen can infect a leukocyte that can cross the BBB without other cells around it noticing. A third approach is transcellular penetration by pinocytosis[2][5].

Next Generation Sequencing (NGS) studies that assess expression profiles of brain cells during normal development and CNS disease increased rapidly over the past few years. Interestingly, some (but not all) of those techniques detect DNA and/or RNA species from microbes. The term microbes here broadly refers to bacteria, viruses, phages and parasites (pro- and eukaryotes) [6].

In a previous research [3], Brazilian and Dutch brain tissue was sequenced and analyzed. In the initial analysis, a batch effect was observed between within the two cohorts. One of the hypotheses is that difference could be explained by the exposure of different pathogens. There are four cell type conditions in the sample pool: microglia (Dutch and Brazilian), whole brain and brain biopsies. These pure microglia samples are obtained with the use of Fluorescence-activated cell sorting (FACS). The whole brain samples are obtained from the right parietal cortex during the course of full body autopsy. The brain tissue biopsies were obtained of the temporal lobe from epilepsy patients.

The aim of this project is to identify pathogens that can be found in different samples and to determine if there is a difference between the two cohorts. In order to identify the presence of a pathogen, non-human aligned reads will be used in a pipeline with Bowtie2.
Identified pathogens will be analyzed further with the use of two analyzing methods:

1. Quantitative expression analysis where the number of reads is normalized to the counts per million (CPM).
2. Gene expression analysis identifying viral/immunological related genes that can be used to visualize the expression of these genes among the different samples.

Another aim is to determine the possibility of contamination in the samples.
It could be possible that unhygienic circumstances or the use of laboratory reagents that contain small amounts of pathogens, infected the samples. These pathogens will be identified as contamination.

2. Theory

The nervous system contains a large collection of cells that work together. The two principal constituents are the neuroglial cells and nerve cells. Nerve cells, neurons, are functional cells. The neuroglial cell, glia, provides structural and metabolic support for neurons. The glia cells can be divided into two major classes: microglia and macroglia. When the nervous system gets infected or damaged, microglia get activated and respond quickly. Activated microglia can encourage cells to repair tissue after injury, remove debris and inactivate invading pathogens [13]. Marcroglia can be divided into four types: oligodendrocytes, Schwann cells, astrocytes and ependymal cells. Schwann cells and oligodendrocytes create the myelin sheath around axons. Schwann cells also play a role in the creation of connecting tissue sheaths around nerves and in axon regeneration. Astrocytes play a role in structural and metabolic functions, but also contribute to the blood-brain barrier. The ependymal cells play a role in the regulation of chemical flow from cavities into the brain [13]. Figure 2 gives an overview of these cells.

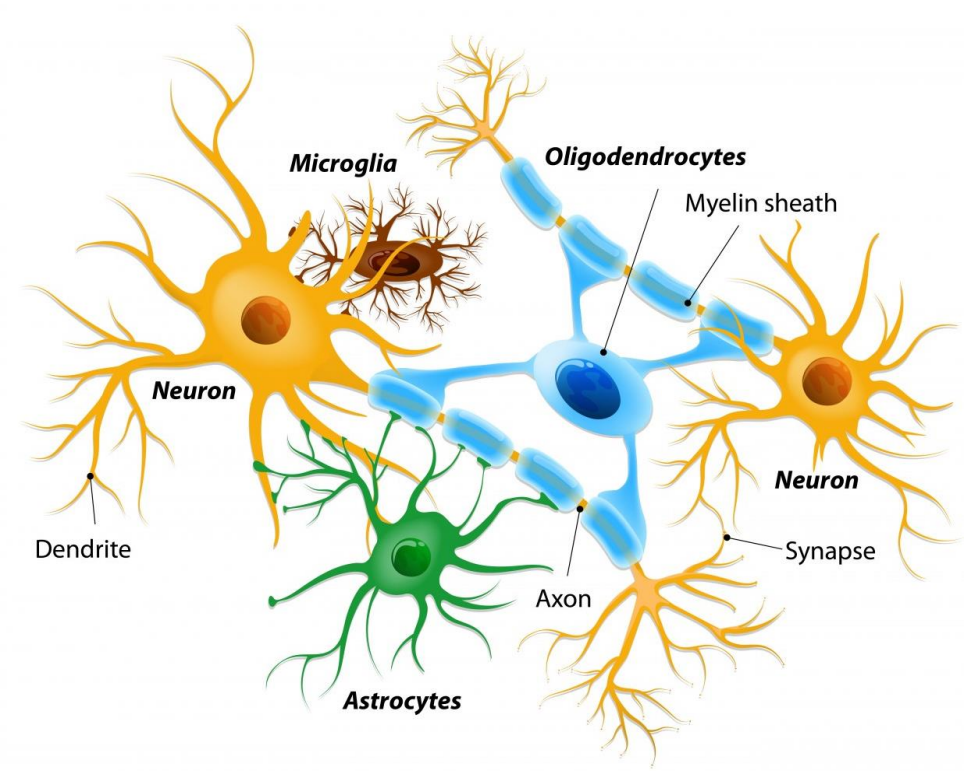


Figure 2: An overview of the neurons and glia cells in the nervous system [14]

2.1 Infection response

During infection, a lot of cells try to get the body back under control. First, macrophages attack the pathogen. Macrophages are big phagocytic cells that travel around the body. They 'eat' the microbe and digest them with enzymes and send messages to the blood vessels to excrete fluids to the 'battlefield'. Macrophages send messenger proteins to call other cells to help and fight the infection. Neutrophils patrol in the blood vessels and respond to messenger proteins. They secrete substances that kill everything that comes across, also killing healthy cells.
When macrophages and neutrophils cannot fight the infection, dendritic cells come into action. Dendritic cells present part of the pathogen on the membrane and travel to the nearest lymph node where a lot of Cytotoxic T cells and T helper cells are waiting to be activated. The dendritic cell scouts for T cells that expresses a protein of the pathogen. The T cell is activated after is has been found and starts to divide rapidly. Some T cells become memory cells, that are waiting in the lymph nodes to respond to another infection of the same pathogen. Other T cells activate B cells, resulting into B cell division. B cells start to secrete antibodies that bind to the pathogen, which makes them easy to digest for phagocytic cells [9].

Under normal conditions, microglia express low levels of major histocompatibility complex (MHC) class 1 and class 2. During infection, microglia get activated, showing an increased proliferation, motility and phagocytic activity. During this stage, they become antigen presenting cells (APC) and perform a function similar to macrophages in peripheral tissues [8].

Since microglia are APC's and phagocytic cells, they can contain parts of pathogens. When microglia are sequenced in NGS studies, DNA/RNA from pathogens can be detected.

2.2 Contamination of DNA/RNA extraction kits

An issue with techniques as polymerase chain reaction (PCR) and ribonucleic acid (RNA) isolation is the introduction of contaminating microbial deoxyribonucleic acid (DNA) and RNA. DNA and RNA contamination can occur when using molecular biology grade water, PCR reagents, DNA extraction kits and RNA extraction kits. The presence of contaminating DNA can generate misleading results, because the contaminant DNA will undergo methods like PCR, meaning amplification of pathogenic sequences [7].

For this project, contamination from kits can mean detection of pathogens that are not present in the original sample. Analysis of these pathogens can lead to a false result of the expression of that particular pathogen, and therefore generating a misleading conclusion that a pathogen can be the cause of the batch effect.

The data of this project consist of four groups: Brazilian whole brain, Dutch whole brain, Dutch microglia and Brazilian surgical samples. The microglia samples undergo a few steps before RNA isolation. First, the brain is dissociated mechanically until a homogeneous suspension is obtained. Myelin and cell debris is removed by a Percoll gradiënt configuration. The next step is to sort cells using the fluorescence-activated cell sorting (FACS) machine. RNA is extracted from the sorted microglia with the RNeasy and Allprep micro kit from Qiagen. RNA from the whole brain samples is extracted using the RNeasy lipid tissue mini kit from Qiagen [3].

A research in 2003 showed that the Qiagen kits were contaminated with DNA from Legionella [10]. In a forensic research in 2012 where two kits were compared [11], no contamination was found in the Qiagen kits anymore.

3. Materials and methods

3.1 Materials

3.1.1 GNU Bash 4.4.12 [15]

Bash was used in various scripts to execute commands multiple times for multiple files.

3.1.2 Python 3.6.3 [16]

Python was used for the creation of the pipeline. The package subprocess was used to execute a process in the command line environment and to obtain the standard output and error of the command. The package os was used to create directories. This was done to create a clear overview of the file locations. The package glob was used to get all files in a directory, to execute a command for all files in a directory. The package os.path is used to get the basename of each file. The basename was needed to get the filename without the path of the file.

3.1.3 Rstudio Desktop 1.0.153 [17]

The desktop application Rstudio was used to create R scripts and visualize plots. It also created an environment that saves variables and made installed packages available in multiple scripts without further installation.

3.1.4 R 3.4.2 (Short Summer) [18]

R was used for visualizations and statistical analysis. Several packages were used for the bar plots and analysis. R package edgeR (version 3.20.2) was used for a counts per million (CPM) normalization. Package scales (version 0.5.0) was used to scale data-frames in order to create a barplot. Package ggplot2 was used to generate a dotplot of the distribution of the amount of pathogens per cohort. Package ggplot2 (version 2.2.1) was used to create the barplots. Package heatmap3 was used to generate a heatmap of the correlation between different conditions of a sample and the presence of a pathogen.

The libraries GSVA (version 1.26.0), limma (version 3.34.4), openxlsx (version 4.0.17), biomaRt (version 2.34.1) and pheatmap (version 1.0.8) were obtained through the CRAN repository[34]**.** GSVA was used for the analysis of the gene sets in immunologic pathways, limma for the differential expression analysis and openxlsx to save a dataframe to a xlsx file. Biomart was used to convert the gene identifiers to entrez identifiers and pheatmap to generate a heatmap of the differential expression between cohorts.

3.1.5 SAMtools 1.6 [19]

SAM tools was used for sam and bam manipulation. The function view was used to convert a sequence alignment map (SAM) file to a binary alignment map (BAM) file and to remove all reads

that had flag 4 (unaligned) The function sort was used to sort the BAM file, followed by the function merge that merged two BAM files from different runs together.

3.1.6 Bowtie2 2.3.2 [20]
Bowtie2 is a tool that aligns sequences to a reference sequence. This was used to align the human genome to the samples and align against the sequences of pathogens. It uses an index of the reference sequence to keep the memory footprint small. This makes Bowtie2 a fast and memory-efficient alignment tool. Bowtie2 was chosen because it is known to perform well when aligning to a large mammalian genome. It is also chosen because it is well documented, fast and easy to use.

3.1.7 BBMap 37.66 [21]
BBMap is a short read aligner that contains bioinformatic tools such as pileup. The pileup tool was used to get the coverage information out of a BAM file.

3.1.8 Perl 5.26.0 [22]
Perl is a programming language that was used to retrieve genomes of multiple species in a genus. It connects to the Eutils server from NCBI[27] and uses a search query to get all genomes in fasta format.

3.1.9 Bedtools 2.26.0 [23]
Bedtools contains a lot of different tools for various genomics analysis tasks. The function bamToFastq was used to convert BAM files to fastq. This was done because the samples were available in BAM format, but Bowtie2 does not accept BAM as a valid input file. The -fq1 and -fq2 parameters were given to create paired end fastq files. Bedtools was also used for the function bamtobed. This function converts a BAM file to a browser extensible data (BED) file. A BED file contains a clear overview of rows and columns with all the information of a BAM file.

3.1.10 Git 2.14.1 [24]
The version control tool Git was used to use an online version control with the online GitHub service.

3.1.11 Picard 2.14 [31]
Picard is a software that contains various tools to manipulate SAM and BAM files. SortSam was used to sort the BAM file using the queryname of the BAM file, AddOrReplaceReadGroups to replace all read groups with a single read group and FixMateInformation to check if the mate pairs accord with each other and to fix them if necessary.
CreateSequenceDictionary was used to create a sequence dictionary from the human genome. The sequence dictionary is required to reorder the BAM file. The last step was to reorder reads in the BAM file using the reference dictionary with ReorderSam.

3.1.12 Htseq 0.6.1 [32-33]
The function Htseq-count from the software Htseq was used to generate counts of the overlap of reads with genes.

3.2 Methods

3.2.1 Filter out human DNA

The first step was converting the BAM data to paired-end fastq data using the function bamToFastq from bedtools. A bash script was used to call the bedtools command for every BAM file. An example of the command:

bamToFastq -i inputfile.bam -fq output_r1.fastq -fq2 output_r2.fastq

The '-i' parameter represents the input file in BAM format. The '-fq' and '-fq2' represent the paired end fastq files.

The next step was to build the bowtie2 index for the human genome. *Homo sapiens* (assembly GRCh38.p11) [25] was used for the index using the '*bowtie2-build*' command. After building the index, the paired-end samples were aligned against the index using the following example command:

Bowtie2 -x human_index -1 input_r1.fastq -2 input_r2.fastq --un-conc output.fastq -S sam_output.sam --no-unal --no-hd --no-sq -p 32

The '--un-conc' parameter wrote the unaligned reads to a new fastq file. This step was necessary to make the aligning against the sequences of pathogens quicker. The chance of finding a pathogen that has a similar sequence in the human genome is smaller with this approach. The '--no-unal' '--no-hd' '--no-sq' parameters removes some rapport functions of the SAM files to make the SAM files less big, because these are not used. The parameter '-p' sets the amount of threads higher to make the alignment quicker. For the aligning step, only two mismatches were allowed.

3.2.2 Finding pathogens

Sequences that were not aligned against the human genome were used to align to the pathogenic genomes and scaffolds. The list of pathogenic genomes and scaffolds was formulated with pathogens that are known to cause disease in the brain of that are noticed by microglia. Part of the pathogens were downloaded using a perl script with the Eutils function from NCBI[27] and a part from the NCBI Genome chapter[26]. The pathogenic genomes were added into one multifasta file using the shell concatenate command:

*cat *.fasta > pathogen_genomes.fasta*

An index was built with the '*bowtie2-build*' command. The next step was to align the pathogen index to the left over samples:

Bowtie2 -x pathogen_index -1 input_r1.fastq -2 input_r2.fastq -S sam_output.sam -p 32

This command was again executed for every set of fastq files with a bash script. This time no new fastq file was created. Also the parameters to make the SAM less big were not used, because the lines it removes are needed for the processing. Again the threads were specified to make the process faster. Only two mismatches are allowed with the default alignment settings.

The result of Bowtie2 was a SAM file. To process this, the SAM file was converted to BAM using the SAMtools '*view'* function:

samtools view -b -S -o output.bam input.sam

The '-b' parameter represents the creation of a BAM file. The '-o' parameter was used for the output file.
After converting the SAM file to BAM, the not aligned sequences were filtered. When a read failed to align, the read flag was set to 4. With the SAMtools '*view'* function, the reads are filtered:

samtools view -b -F 4 inputfile.bam > output.bam

The next step was to convert the BAM files to a BED file using Bedtools. A BED file starts with a accession number, base pair number start and base pair number end.

bedtools bamtobed -i input

Obtaining the scientific names based on the accession numbers was done with a python script that used a dictionary with the accession number and the corresponding pathogen name. To create a clear overview of the found pathogens in each sample, a pathogen was only reported once. This means that if multiple scaffolds were found of a pathogen, it was only reported once.

3.2.3 Visualisation of distribution

A bar plot was created to give an overview of the found pathogens.. The BAM file that was created with the Bowtie2 run was used to perform a pile up using BBMap:

pileup.sh in=input.bam out=out.txt

In the resulting file were the amounts of aligned base pairs (bp) reported. When a certain accession number has 0.0000% of aligned bp, it was excluded of the file using a python script. A python script was used to convert all accession numbers to pathogen names and count all found bp for each pathogen. The next step was to identify the number of unidentified bp. The total bp was determined using SAMtools '*view':*

samtools view input.bam -c -o output.txt

The amount of unidentified was calculated as followed:

Total base pairs - total amount of aligned base pairs = unidentified base pairs

The file that contained aligned bp per pathogen and the unidentified bp was saved in a tab separated file. This file was loaded in R as a data frame. For each data frame, a CPM normalization was performed using the edgeR library. The plots were created using ggplot from the library ggplot2.

To visualize the influence of pathogens that are labeled as contamination, a dot plot was created. For every cohort, two data frames were created. A data frame with the total amount of bp that had been identified as a pathogen and a dataframe with the total amount of bp excluding the bp that had been identified as contamination pathogens. The plot was created using ggplot2.

3.2.4 Correlation analysis

To research the possibility that some characteristics have an influence on the found bp, a heatmap was generated with all the characteristics of the donors and the amount of found pathogens. The amount of bp per pathogen per sample was normalized using the CPM function of edgeR. A heatmap was generated using heatmap3 with the amount of bp per pathogen in each sample and with the characteristics origin, tissue, sequence machine, gender, age and PMD.

3.2.5 Gene Set Variation Analysis

To identify different immunologic genes that may correlate to a certain presence of a pathogen, a differential expression analysis was performed on the human data. The first step was to create the counts of genes per sample. This has been done with bowtie2, SAMtools, Picard and Htseq-count. At first, the samples are aligned to the human genome with bowtie2:

bowtie2 -x humanindex -1 sample_r1.fastq -2 sample_r2.fastq -S output.sam

Next, a lot of preprocessing steps were done that were essential before running the Htseq-count function:

samtools view -Sbo output.bam input.sam

java -jar picard.jar SortSam I = input.bam O = output.bam SO = queryname

java -jar picard.jar AddOrReplaceReadGroups INPUT = input.bam OUTPUT = output.bam LB = samplename PU =samplename SM = samplename PL= illumina CREATE_INDEX = true

java -jar picard.jar FixMateInformation INPUT = input.bam

java -jar picard.jar CreateSequenceDictionary R = genomeFasta O = genomeDictionary.dict

java -jar picard ReorderSam I = input.bam O = output.bam ALLOW_INCOMPLETE_DICT_CONCORDANCE = true R = genomeFasta

java -jar picard.jar MarkDuplicates INPUT = input.bam OUTPUT = output.bam CREATE_INDEX = true METRICS_FILE = sample.metrics.log

The counts of overlap of reads with genes were determined with the count function from Htseq:

Htseq-count -f bam -r name -i gene input.bam -s no genomeGTF > sample_counts.txt

Followed by adding all the generated counts per sample in one file. This file was loaded into R together with a target file that contains the sample name, origin of sample, age of donor, post-mortem delay and cause of death. The gene identifier in the counts dataframe needed to be converted to entrez identifiers. From the *Homo Sapiens* ensembl database, the ensembl gene id, entrez gene, external gene id and the wikigene description was retrieved. The gene identifier was changed to the entrez gene id.
This information was necessary in order to perform the gene set variation analysis (GSVA). The immunological pathways were downloaded from the WEHI Bioinformatics Research Page [30]. A heatmap was generated with the function pheatmap from the library pheatmap, showing the most abundance immunological functions with the found genes between two cohorts.

4. Results

48 donors from two different cohorts are used for this research. Donors from the Netherlands are obtained the Netherlands brain bank. Brazilian samples were received from the human brain bank of the pathology department from the University of São Paulo medical school. 15 out of 16 whole brain donors also have microglia sequenced. In total, 66 samples are used. An overview of the donors and samples can be observed in Appendix 1 (page 33).

4.1 Identified pathogens

There are four cohorts compared to determine the different occurrence of pathogens. The four cohorts are: Dutch microglia, Brazilian microglia, Brazilian whole brain and Brazilian biopsy samples. In total, 54 pathogens are identified. 12 pathogens that are not present in all of the cohorts can be observed in Table 1. The remaining 42 pathogens are present in all cohorts. 49 pathogens are identified in the Dutch microglia cohort, 52 in the Brazilian microglia cohort, 50 in the Brazilian whole brain cohort and 44 in the Brazilian biopsy cohort.

Pathogen	**NL_MG**	**BR_MG**	**BR_WB**	**Biopsy**
Anaerococcus vaginalis				✓
Bacillus phage	✓	✓	✓	
BeAn 58058	✓	✓	✓	
Borrelia burgdorferi	✓	✓	✓	
Enterobacteria phage	✓	✓	✓	
Finegoldia magna				✓
Human endogenous	✓	✓	✓	
Human immunodeficiency		✓	✓	
Phytophthora infestans	✓	✓	✓	
Rabies virus		✓		
Saccharomyces cerevisiae	✓	✓	✓	
Tick-borne encephalitis		✓		

Table 1: Overview of pathogens that occur differently between the four cohorts. Each check mark means the presence of a pathogen in one or more samples of a cohort.

An overview of all identified pathogens is available in Appendix 1 (page 39).

The Brazillian biopsy samples are used as a control group. The biopsy samples are collected in a surgical room, which is thought to be more hygienic in comparison with a dissection room. This creates the hypothesis that pathogens that are not present in the Brazilian biopsy samples, but are present in other cohorts, are possible environmental contamination pathogens.

4.2 Basepair overview

The amount of bp that are identified as pathogens differs between cohorts, but also between samples within a cohort. It is possible that the amount of identified bp contains possible contamination pathogens. A dotplot is created in order to visualize the mean amount of bp of pathogens per sample.

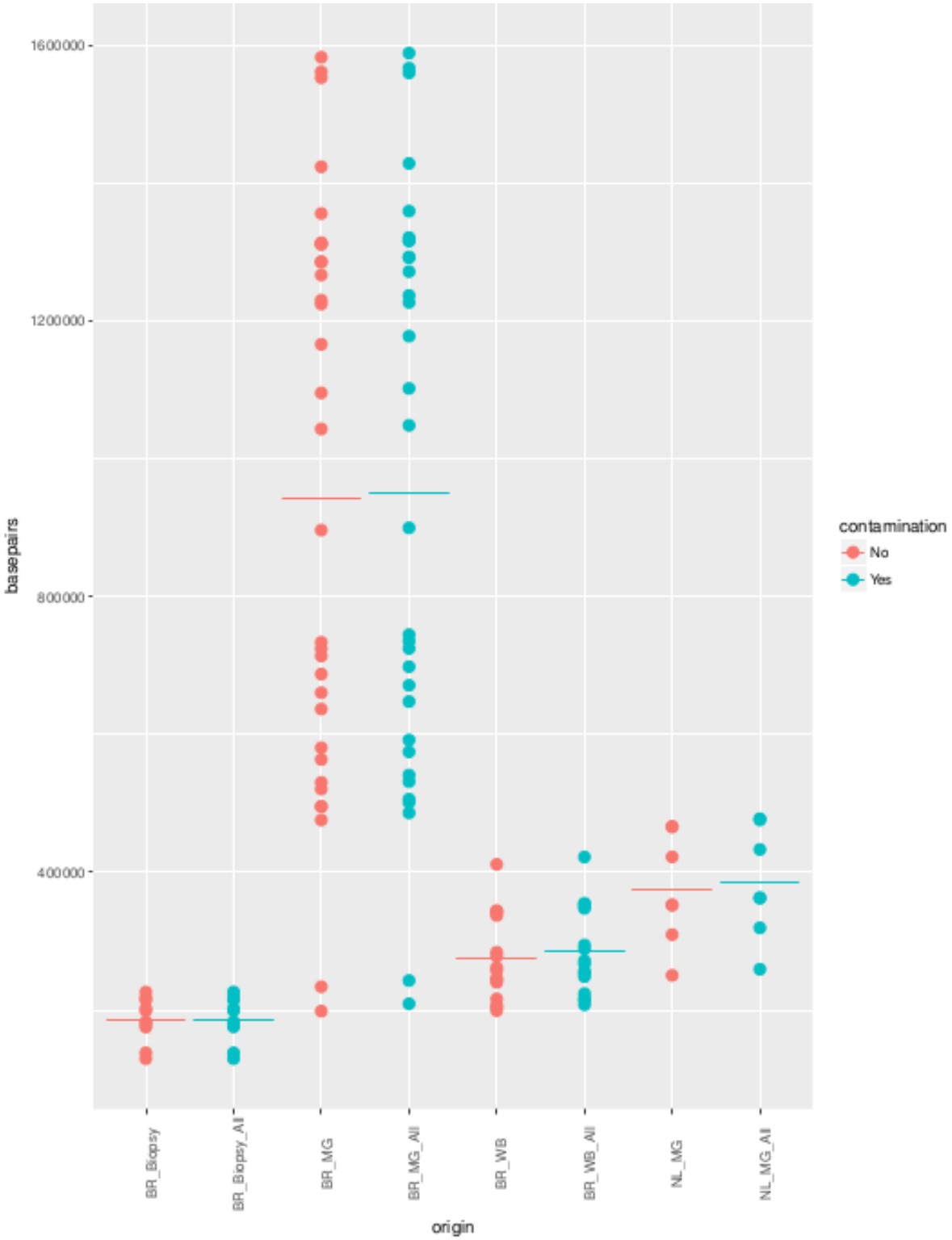


Figure 3: Overview of the total identified bp per sample with and without possible contamination pathogens. The total amount of identified bp varies a lot in the Brazilian microglia cohort. The amount of bp that were labeled as contamination is very low.

Figure 3 visualizes the amount of bp per sample that has been identified as a pathogen. The dots that are without contamination have bp removed that have been identified as a contamination pathogen. The total amount of bp that has been identified in the Brazilian microglia cohort varies a lot between samples. It is possible that some samples have a high amount of a certain pathogen present, which increases the total amount of identified bp.

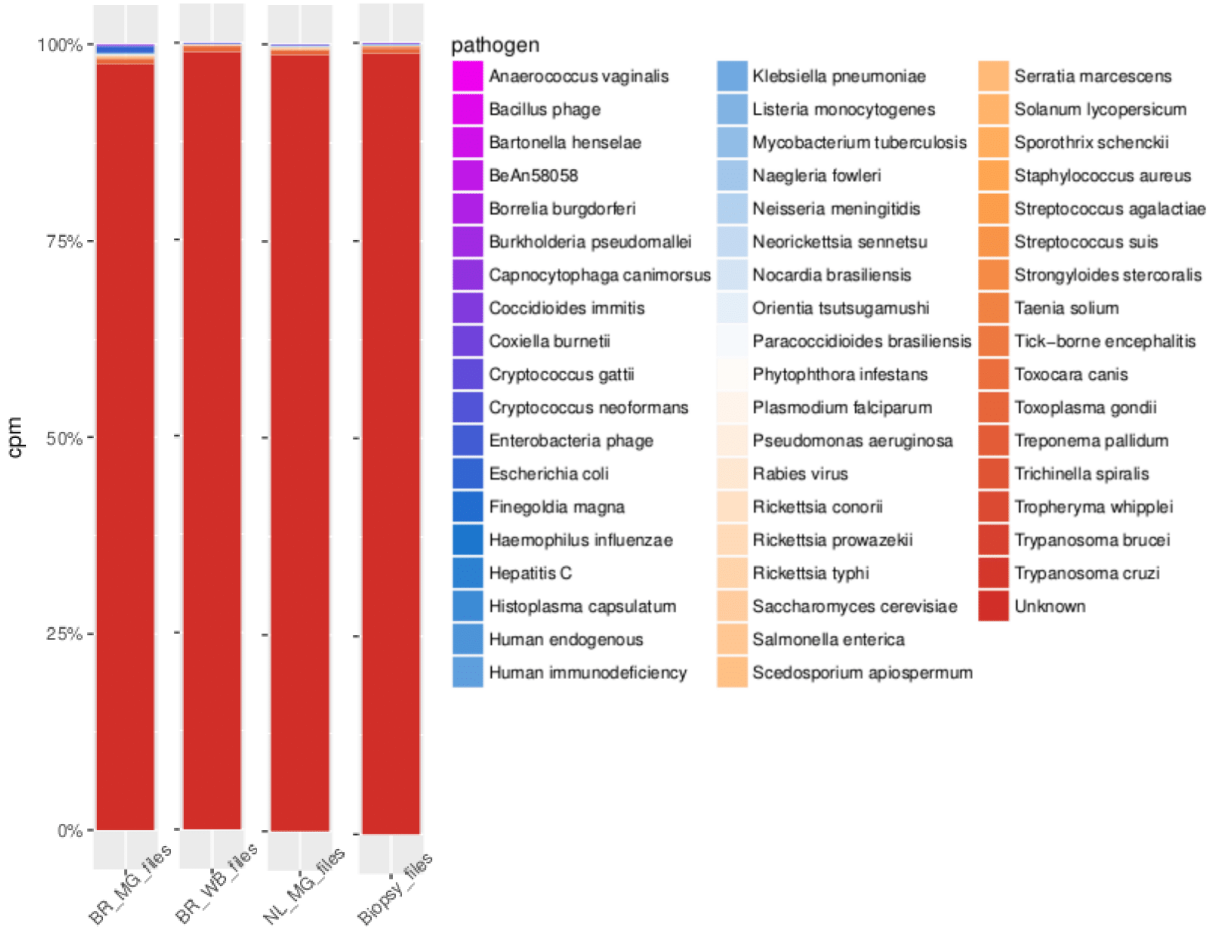


Figure 4: Overview of the normalized mean (CPM) pathogen bp distribution with unidentified bp. The legend shows all identified pathogens and unknown. 97-99% of all bp are unknown, which are not identified as a pathogen.

The barplot in figure 4 visualizes the normalized amount of bp that have been identified as bp from pathogens. Each bar represents the mean of the normalized bp that failed to align to the human genome per cohort. Overall, about 99% is ‘Unknown’. This means that these bp are not identified as pathogens. In the Brazilian microglia cohort, the amount of ‘Unknown’ is approximately 97%, so it has more bp identified in comparison with the other cohorts. This observation is also seen in figure 3.

Unidentified bp are removed from the barplot to visualize difference of pathogen presence per cohort.

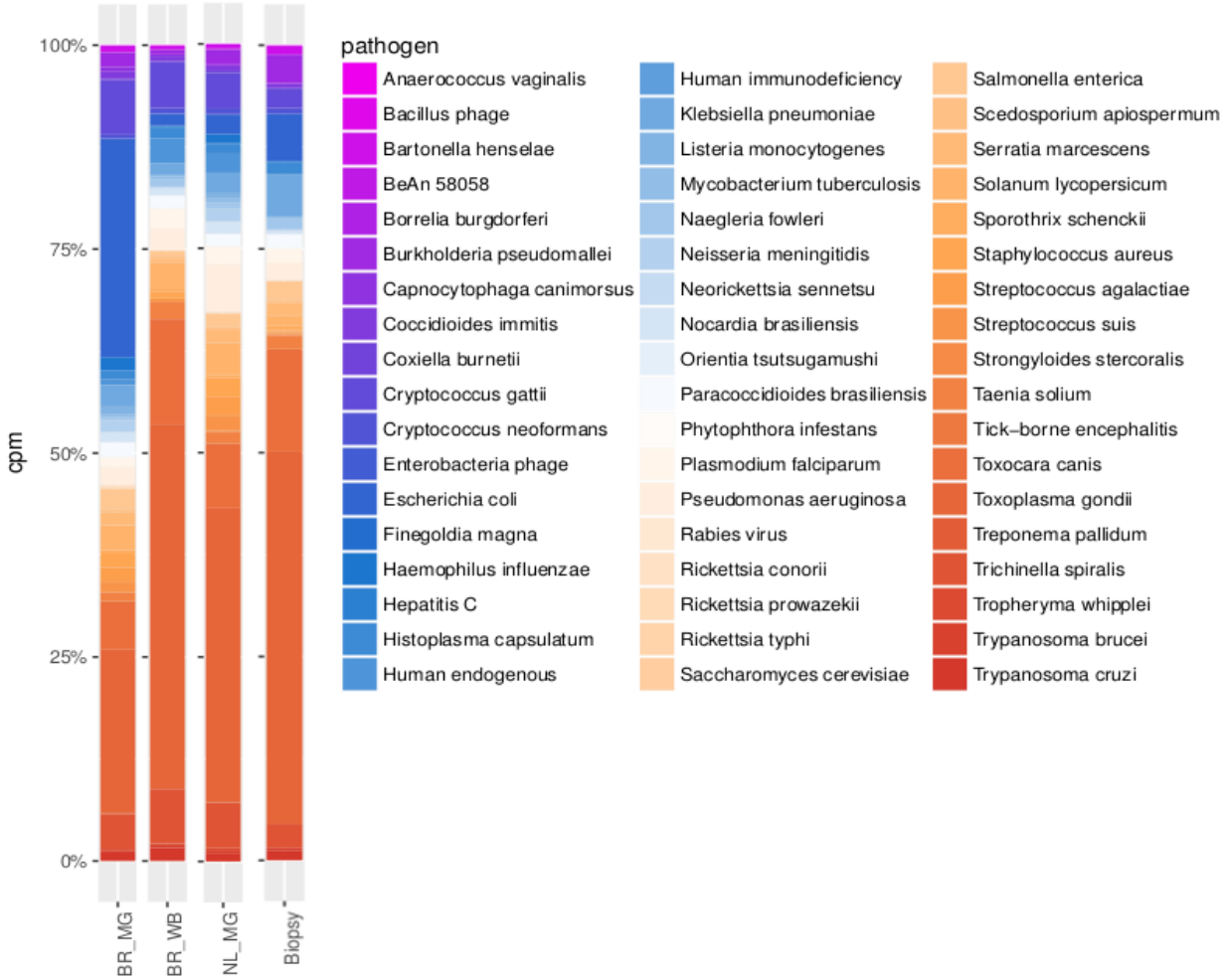


Figure 5: The distribution of pathogens without unidentified bp. Determined with mean of the normalized (CPM) bp per cohort. The legend contains all 54 identified pathogens. The Brazil microglia cohort shows an higher amount of Escherichia Coli.

Figure 5 contains an overview of pathogens per cohort. When removing unidentified bp it becomes visible that *Toxoplasma Gondii* is greatly present in all the cohorts. In the Brazilian microglia, *Escherichia Coli* is greatly present. The Brazilian whole brain samples have a visible similarity with the Brazilian biopsy samples. The Dutch samples have similarities with the Brazilian whole brain samples and with the Brazilian biopsy samples. A difference is observed in the upper part of the bar, that represents the first 25 pathogens of the legend. Also, *Toxocara Canis* is present in a smaller amount. The observation of a smaller amount of *T. Canis* is also observed in the Brazilian microglia samples. In Brazilian microglia samples it is visible that *T. Gondii* is less present. The exact numbers of each pathogen per cohort is available in Appendix 2 (page 36).

In the legend of figure 4 and figure 5 it is visible that there are traces of the *Human Immunodeficiency virus.* This might suggest that some donors were infected with a HIV infection. However, this virus is hardly present (Appendix 2). This might suggest that the observation of HIV in the samples is an computational artefact.

4.3 Correlation analysis

It was hypothesized that age and PMD might show a pattern in pathogen presence.
To depict age or PMD related effects, a heatmap was generated containing all known characteristics from the donors: age, gender, PMD, sequencing machine, tissue and origin.

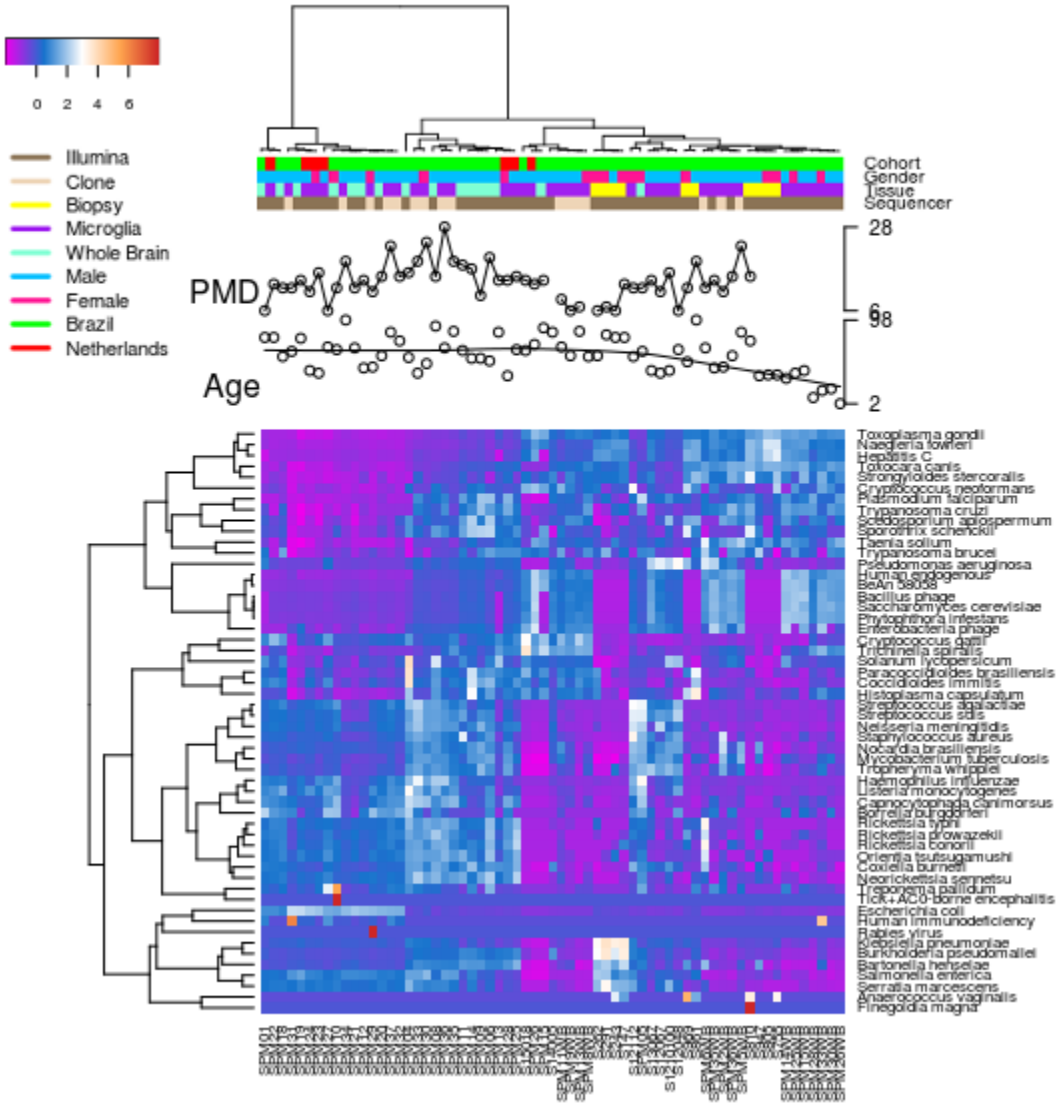


Figure 6: Heatmap of the amount of pathogens per sample with the characteristics of the samples. Larger image is available in Appendix 5 (page 45).

In figure 6 it is visible that there is some small correlation with age. The age of the donors shows a declining trend. It also shows no correlation with cohort and gender. The published results of

Galatro and Holtman et al. [3] shown that there was no correlation found between the different cohorts. In figure 6 a comparable result is observable. There is also no correlation between the present pathogens and sequencing machine.

4.4 Gene Set Variation Analysis

The gene set variation analysis is performed with three comparisons: Brazil biopsy vs Brazil whole brain, Brazil microglia vs Dutch microglia and Brazil whole brain vs Brazil microglia. The results of the last two comparisons are available in Appendix 4 (page 43).

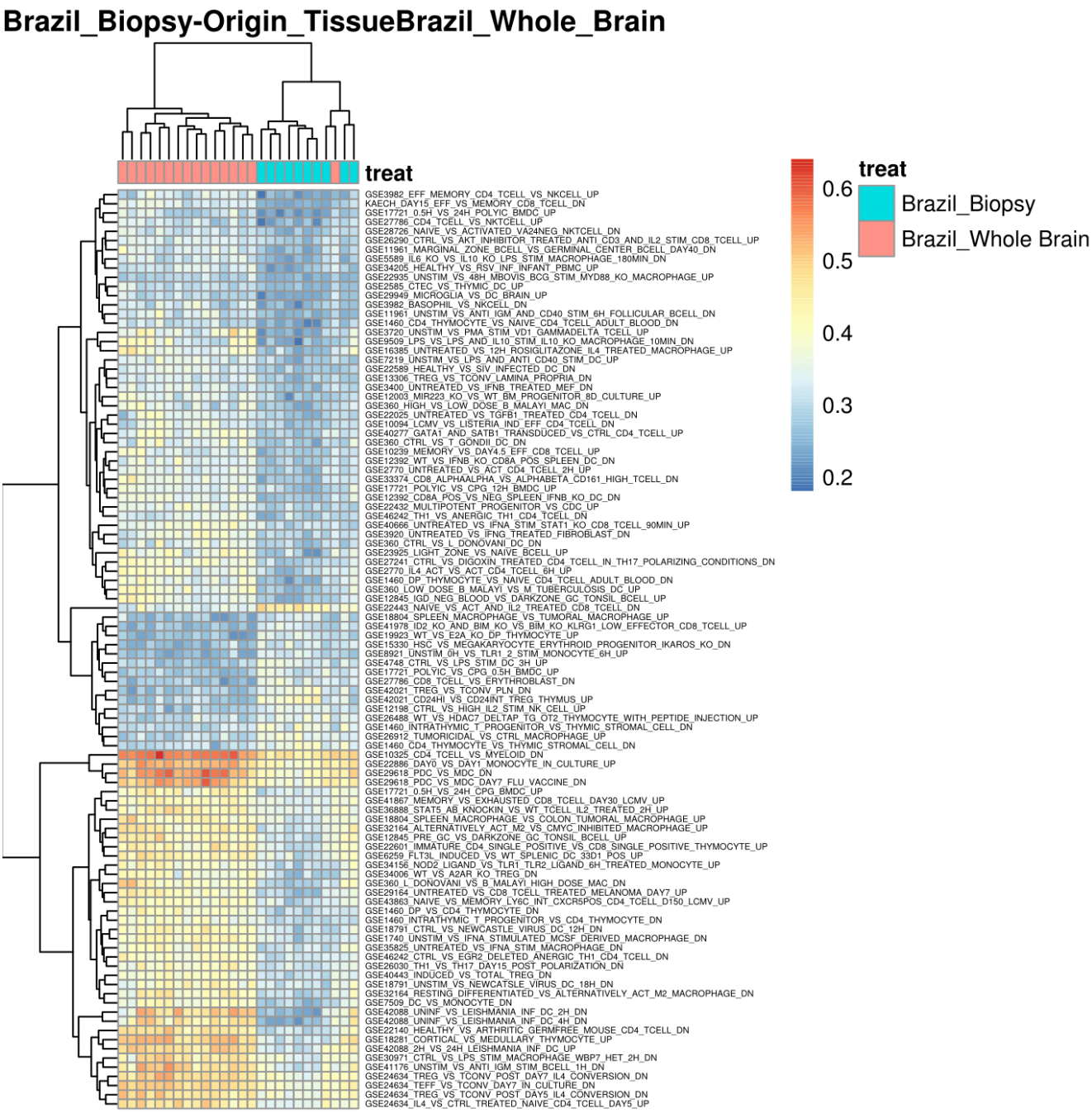


Figure 7: Heatmap of immunologic pathways of Brazil biopsy samples against Brazil whole brain samples. Upregulation In a CD4 Tcell pathway in Brazilian whole brain samples can indicate an active infection at time of death.

In figure 7 there is a clear difference between the Brazil biopsy samples and the Brazil whole brain samples. This may be due the presence of different pathogens. A CD4 Tcell pathway is upregulated in all Brazillian whole brain samples, except one. This can indicate an active infection in the days before death. Also a pathway of plasmacytoid dendritic cells and monocytes is upregulated in these samples, which also supports the hypothesis of an active infection.

5. Conclusion and discussion

The aim of the project was to confirm or discard the hypothesis that the batch effect found in a previous research from Galatro and Holtman et al. [3] is caused by the exposure of different pathogens.

The aim of the project was to confirm of discard the batch effect found in the initial analysis from Galatro and Holtman et al. [3] by identifying the presence of pathogens.
It was suspected that pathogens are present and detectable in the human brain and microglia cells. A difference in presence of pathogens between Brazilian samples and Dutch samples can explain a small batch effect.

To confirm this hypothesis, pathogens have been identified in all samples and were visualized with the use of a barplot to observe a difference between cohorts. The Brazilian biopsy samples were used as a control group to identify pathogens that could be present due contamination from the autopsy room. A dotplot was used to observe if the pathogens that have been labeled as contamination were highly present and could influence the outcome. To observe a correlation between different characteristics of the samples and the presence of pathogens, a heatmap was used.

In total, 54 different pathogens are identified. 12 of these pathogens are not present in all cohorts. There is little difference between the amount of presence of pathogens between cohorts, which cannot explain the batch effect.

The removal of pathogens that are labeled as possible contamination pathogens was performed to see if there is a difference between cohorts that would not have been visible with those pathogens. The difference between the number of identified bp with and without the pathogens that are labeled as possible pathogens in the dotplot of figure 3 is little. This can mean that the pathogens that are labeled as contamination can also be computational artefacts.

Approximately 99% of the bp failed to align to the human genome and are not identified as a pathogen. It is possible that these bp are part of pathogens that have not been identified. This makes it also possible that there is a difference between cohorts, but that the pathogens that differ have not been identified yet.

It is also possible that the Bowtie2 aligner could have an influence on the amount of found pathogens. All samples were aligned with the default alignment settings, which only allows 2 mismatches per sample. This creates the possibility that some pathogens are present, but not found because they have more mismatches.

Another possibility is that sequences that failed to align to the human genome are in fact human sequences, but are not recognized by the human genome. This makes it possible that the amount of unknown is truly lower.

The Brazil microglia samples have approximately 97% of the bp unidentified. This is a difference between the other cohorts, since they have approximately 99% unidentified. In figure 5 it is observable that all Brazil microglia samples have *E. coli* largely present, which can be the cause of this difference.

In the legend of figure 4 and figure 5 it is also visible that HIV is present, which can indicate a HIV infection in some donors. However, in Appendix 2 it is concluded that this is only in very small amounts present, which can indicate an computational artefact.

The heatmap with all characteristics of the samples shows little correlation of age. PMD, sequencer, cohort and tissue do not seem to have influence on the presence of pathogens.

The final conclusion is that the batch effect cannot be explained with the observations. There is no significant difference of the occurrence of pathogens between cohorts and samples.

Contamination analysis

The most recent research of the Qiagen DNA and RNA extraction kits did not show contamination. This makes it not likely that the kits used in the initial research were contaminated with pathogens. All samples were extracted by one group of laboratory technicians, so it is not likely that contamination occurred due different laboratory techniques.

Recommendations

In this research, a cutoff for bp was not used. A recommendation for further analysis is to determine a cutoff for bp. In Appendix 2 it is noticeable that a few pathogens are identified with a low number of bp. This makes it possible that these pathogens are not really present but are reported false positive. Therefore they can be considered as computational artefacts.
When the pathogens with an occurrence lower than the determined cutoff, they can be removed. This can create a different distribution between cohorts.

Another recommendation is to do a statistical analysis of the presence of pathogens. This can give an indication if a presence of a pathogen is significant and if the bp that are removed because of contamination were significantly present or not.

A recommendation to identify more pathogens is to extend the list of genomes, for example the viral database of NCBI.

To avoid kit contamination, a recommendation is to sequence a control kit and DNA free water to proof that the kit used is indeed contamination free.

In a future research, a recommendation is to do a immunohistochemistry on brain slices with pathogen antibodies to proof the presence of pathogen traces in the brain.

References

1. Drevets (2016). Immunology of Central Nervous System Pathogens. In Encyclopedia of Immunobiology, volume 4. Elsevier Ltd, 173-183
2. Sorge and Doran et al. 2012 in Future Microbiology
3. Galatro and Holtman et al. 2017 in Nature Neuroscience
4. Endothelial Cells (Microvascular) *http://www.promocell.com/products/human-primary-cells/endothelial-cells-microvascular/1* Consulted on 05-10-2017
5. Zuchero (2016). Discovery of Novel Blood-Brain Barrier targets to Enhance Brain Uptake of Therapeutic Antibodies, In Neuron, volume 89, issue 1, 70–82
6. Frey and Bishop-Lilly (2015). Next-Generation Sequencing for Pathogen Detection and Identification, In Methods in Microbiology, volume 42, 525-554
7. Salter et al. 2014 in BMC Biology
8. Yang and Han et al in Journal of Clinical Neuroscience Volume 17, Issue 1, January 2010, Pages 6-10
9. David D. Chaplin. Overview of the Immune Response in The Journal of allergy and clinical immunology Volume 125, Issue 2, Supplement 2, 2010, S3-S23
10. Evans and Murdoch et al. 2003 in Journal of Clinical Microbiology
11. Phillips K, McCallum N, Welch L. A comparison of methods for forensic DNA extraction: Chelex-100 and the QIAGEN DNA Investigator Kit (manual and automated) in Forensic Science International: Genetics, Volume 6, Issue 2, 2012, 282-285
12. UMCG Organogram. Consulted on 21-11-2017 https://www.umcg.nl/SiteCollectionImages/UMCG/Over_het_UMCG/Organisatie/organogram-eng-large.JPG
13. John H Martin, Organization of the Central Nervous System in Neuroanatomy Text and Atlas, fourth edition, McGraw-Hill, 2012, 3-9
14. Figure neurons and glia cells. Consulted on 21-11-2017 https://cerebralpalsynewstoday.com/2016/05/04/microglial-cells-interaction-with-dendrimer-in-cerebral-palsy-may-lead-to-novel-therapeutics/
15. Bash. Consulted on 23-11-2017. https://www.gnu.org/software/bash/
16. Python. Consulted on 23-11-2017 https://www.python.org/downloads/release/python-363/
17. Rstudio desktop. Consulted on 23-11-2017 https://www.rstudio.com/products/rstudio/
18. R. Consulted on 23-11-2017 https://www.r-project.org/
19. SAMtools. Consulted on 23-11-2017 https://sourceforge.net/projects/samtools/files/samtools/1.6/
20. Bowtie2. Consulted on 23-11-2017 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
21. BBMap/pileup. Consulted on 23-11-2017 https://github.com/BioInfoTools/BBMap/blob/master/sh/pileup.sh
22. Perl. Consulted on 23-11-2017 http://www.perl.org/
23. Bedtools. Consulted on 23-11-2017 http://bedtools.readthedocs.io/en/latest/
24. Git. Consulted on 23-11-2017 https://git-scm.com/

25. NCBI Homo Sapiens genome. Consulted on 20-09-2017 https://www.ncbi.nlm.nih.gov/genome/?term=homo%20sapiens
26. NCBI Genome Entrez search interface. Consulted on 20-09-2017 https://www.ncbi.nlm.nih.gov/genome/
27. Entrez Programming Utilities Help Manual. Consulted on 20-09-2017 https://www.ncbi.nlm.nih.gov/books/NBK25501/
28. UMCG Proton centre. Consulted on 07-12-2017 https://www.umcgroningenptc.nl/en/umc-groningen-protonen-therapie-centrum
29. ERIBA UMCG. Consulted on 07-12-2017 http://eriba.umcg.nl/about/
30. Human immunologic pathways data. Consulted on 18-12-2017. http://bioinf.wehi.edu.au/software/MSigDB/human_c7_v5p2.rdata
31. Picard. Consulted on 18-12-2017 https://github.com/broadinstitute/picard/releases
32. Htseq. Consulted on 18-12-2017 https://htseq.readthedocs.io/en/release_0.9.1/
33. Anders S, Pyl PT, Wolfgang H. HTSeq – a Python framework to work with high-throughput sequencing data in Bioinformatics, Volume 31, Issue 2, 2015, 166-169
34. CRAN Repository. Consulted on 07-01-2017 https://cloud.r-project.org/
35. Front page image. Consulted on 05-12-2017 https://www.pasteur.fr/en/education/programs-and-courses/e-learning-mooc/mooc-microbes-brain

Appendix 1

Donor	Microglia/ whole brain	Age	Gender	Origin	Post mortem Delay
spm09	Microglia	78	M	Brazil	6
spm08	Microglia	78	M	Brazil	13
spm06	Microglia	56	M	Brazil	12
spm05	Microglia	62	M	Brazil	12
spm04	Microglia	77	F	Brazil	14
spm36	Microglia	40	M	Brazil	11
spm35	Microglia	37	M	Brazil	16
spm33	Microglia	67	M	Brazil	6
spm30	Microglia	64	M	Brazil	12
spm28	Microglia	98	F	Brazil	19
spm26	Microglia	66	M	Brazil	12
spm25	Microglia	43	M	Brazil	14
spm15	Microglia	44	M	Brazil	11
spm14	Microglia	57	F	Brazil	15
spm13	Microglia	84	M	Brazil	23
spm11	Microglia	74	M	Brazil	15
spm37	Microglia	55	M	Brazil	16
spm34	Microglia	37	M	Brazil	19
spm32	Microglia	48	M	Brazil	24
spm31	Microglia	91	F	Brazil	15
spm29	Microglia	66	M	Brazil	28
spm27	Microglia	85	M	Brazil	19
spm24	Microglia	63	M	Brazil	18
spm23	Microglia	54	M	Brazil	17
spm22	Microglia	54	M	Brazil	10
spm21	Microglia	51	M	Brazil	20

spm20	Microglia	84	M	Brazil	14
spm19	Microglia	34	M	Brazil	14
spm18	Microglia	64	M	Brazil	15
spm12	Microglia	63	M	Brazil	14
spm10	Microglia	70	F	Brazil	13
spm01	Microglia	89	F	Brazil	14
S15018	Microglia	84	F	Netherlands	NA
S14005	Microglia	67	M	Netherlands	9
S13067	Microglia	57	M	Netherlands	6
S12112	Microglia	85	M	Netherlands	7
S12110100	Microglia	56	F	Netherlands	NA
S12067	Microglia	57	M	Netherlands	6
S12048	Microglia	81	M	Netherlands	7
spm9	Whole Brain	78	M	Brazil	6
spm8	Whole Brain	78	M	Brazil	13
spm6	Whole Brain	56	M	Brazil	12
spm5	Whole Brain	62	M	Brazil	12
spm36	Whole Brain	40	M	Brazil	14
spm35	Whole Brain	37	M	Brazil	11
spm33	Whole Brain	40	M	Brazil	16
spm30	Whole Brain	64	M	Brazil	6
spm2	Whole Brain	50	M	Brazil	12
spm28	Whole Brain	98	F	Brazil	19
spm26	Whole Brain	66	M	Brazil	12
spm25	Whole Brain	43	M	Brazil	14
spm15	Whole Brain	44	M	Brazil	11
spm14	Whole Brain	57	F	Brazil	15
spm13	Whole Brain	84	M	Brazil	23
spm11	Whole Brain	74	M	Brazil	15
S969	Biopsy	34	F	Brazil	NA
S947	Biopsy	35	M	Brazil	NA
S861	Biopsy	35	M	Brazil	NA

S810	Biopsy	31	M	Brazil	NA
S805	Biopsy	37	F	Brazil	NA
S755	Biopsy	40	F	Brazil	NA
S291	Biopsy	9	F	Brazil	NA
S262	Biopsy	17	F	Brazil	NA
S243	Biopsy	19	M	Brazil	NA
S147	Biopsy	2	F	Brazil	NA

Table 1: overview of characteristics of samples

Appendix 2

Pathogen	NL_MG	BR_MG	BR_WB	Biopsy
Anaerococcus vaginalis	0	0	0	103
Bacillus phage	222	224	216	0
Bartonella henselae	1993	7396	1162	2017
BeAn 58058	348	356	342	0
Borrelia burgdorferi	108	538	56	0
Burkholderia pseudomallei	7064	16668	1447	6638
Capnocytophaga canimorsus	736	5742	623	104
Coccidioides immitis	2105	7725	1729	973
Coxiella burnetii	1027	2427	328	151
Cryptococcus gattii	16817	63076	16126	4491
Cryptococcus neoformans	2632	4505	1865	1208
Enterobacteria phage	247	185	177	0
Escherichia coli	9114	254503	4110	10883
Finegoldia magna	0	0	0	14
Haemophilus influenzae	4303	15061	430	187
Hepatitis C	335	317	299	299
Histoplasma capsulatum	4616	10756	3720	2429
Human endogenous	8916	6216	8623	0

Human immunodeficiency	0	85	13	0
Klebsiella pneumoniae	9619	24850	4040	9817
Listeria monocytogenes	2617	9692	601	101
Mycobacterium tuberculosis	1659	3250	952	271
Naegleria fowleri	2767	2946	2477	2339
Neisseria meningitidis	6220	13383	616	432
Neorickettsia sennetsu	194	1283	126	68
Nocardia brasiliensis	5471	10602	2383	737
Orientia tsutsugamushi	173	1443	119	96
Paracoccidioides brasiliensis	5852	16845	4632	3135
Phytophthora infestans	163	146	155	0
Plasmodium falciparum	8345	11301	6853	3441
Pseudomonas aeruginosa	22336	20657	7336	3880
Rabies virus	0	1	0	0
Rickettsia conorii	244	1854	200	108
Rickettsia prowazekii	224	1522	135	108
Rickettsia typhi	263	1733	145	94
Saccharomyces cerevisiae	207	208	202	0
Salmonella enterica	6036	22731	1696	4218

Scedosporium apiospermum	1811	3693	1479	782
Serratia marcescens	5870	15556	844	2961
Solanum lycopersicum	15082	29152	8896	2316
Sporothrix schenckii	1360	3873	1161	940
Staphylococcus aureus	9130	16129	2295	351
Streptococcus agalactiae	8473	16401	590	225
Streptococcus suis	7244	12118	532	327
Strongyloides stercoralis	405	476	440	328
Taenia solium	5605	10272	5774	3131
Tick-borne encephalitis	0	2	0	0
Toxocara canis	30464	56457	36782	23503
Toxoplasma gondii	138328	190525	127779	85272
Treponema pallidum	76	1239	154	29
Trichinella spiralis	21344	39965	19051	5260
Tropheryma whipplei	1562	2917	706	236
Trypanosoma brucei	860	2015	852	522
Trypanosoma cruzi	3926	9497	4582	2278

Table 2: Overview of the mean basepairs per pathogen per cohort

Appendix 3

A list of all identified pathogens.

Anaerococcus vaginalis

Bacillus phage

Bartonella henselae

BeAn 58058

Borrelia burgdorferi

Burkholderia pseudomallei

Capnocytophaga canimorsus

Coccidioides immitis

Coxiella burnetii

Cryptococcus gattii

Cryptococcus neoformans

Enterobacteria phage

Escherichia coli

Finegoldia magna

Haemophilus influenzae

Hepatitis C

Histoplasma capsulatum

Human endogenous

Human immunodeficiency

Klebsiella pneumoniae

Listeria monocytogenes

Mycobacterium tuberculosis

Naegleria fowleri

Neisseria meningitidis

Neorickettsia sennetsu

Nocardia brasiliensis

Orientia tsutsugamushi

Paracoccidioides brasiliensis

Phytophthora infestans

Plasmodium falciparum

Pseudomonas aeruginosa

Rabies virus

Rickettsia conorii

Rickettsia prowazekii

Rickettsia typhi

Saccharomyces cerevisiae

Salmonella enterica

Scedosporium apiospermum

Serratia marcescens

Solanum lycopersicum

Sporothrix schenckii

Staphylococcus aureus

Streptococcus agalactiae

Streptococcus suis

Strongyloides stercoralis

Taenia solium

Tick-borne encephalitis

Toxocara canis

Toxoplasma gondii

Treponema pallidum

Trichinella spiralis

Tropheryma whipplei

Trypanosoma brucei

Trypanosoma cruzi

Appendix 4

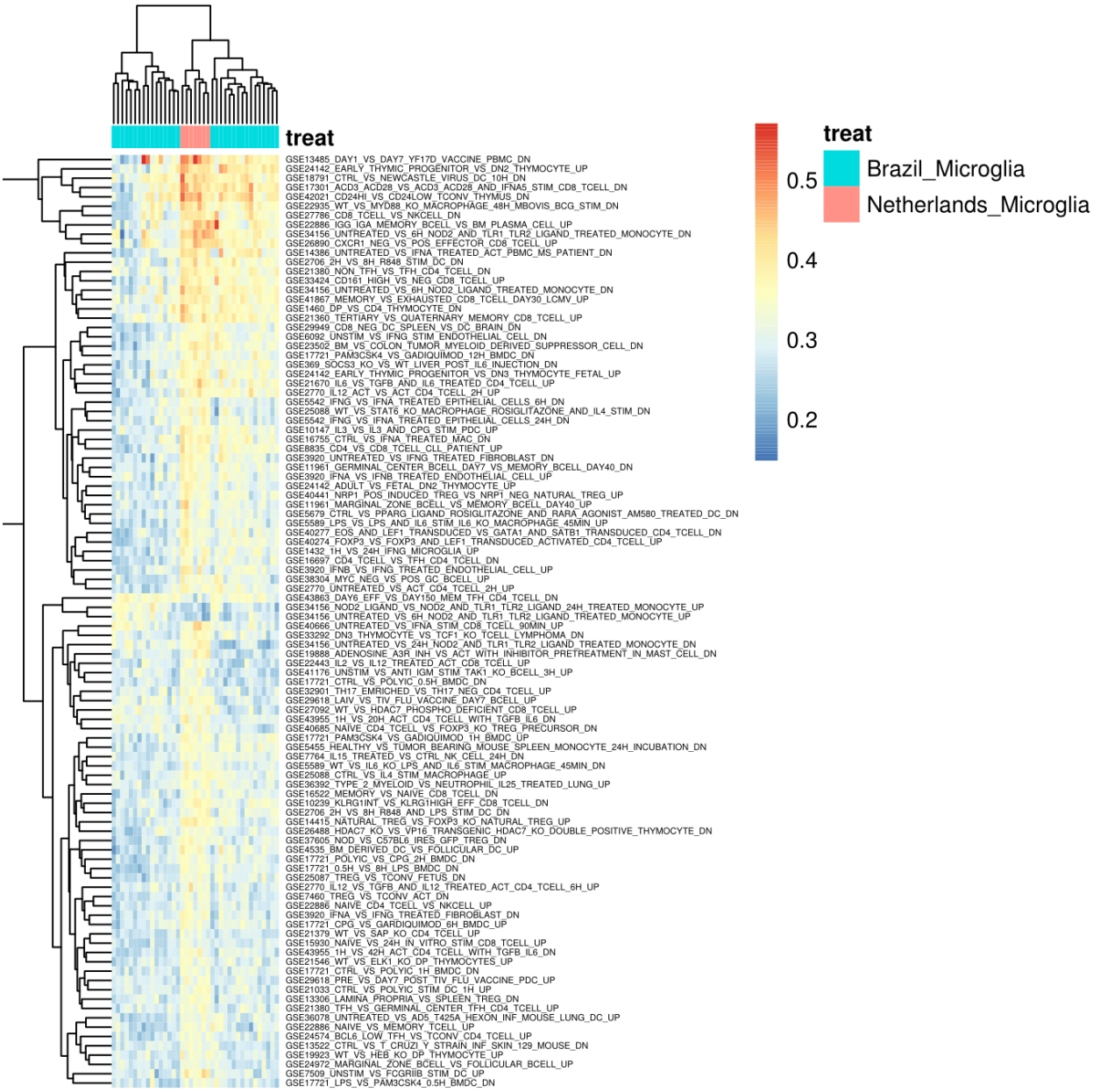


Figure 1: Heatmap of immunologic pathways comparing Brazil Microglia with the Dutch Microglia

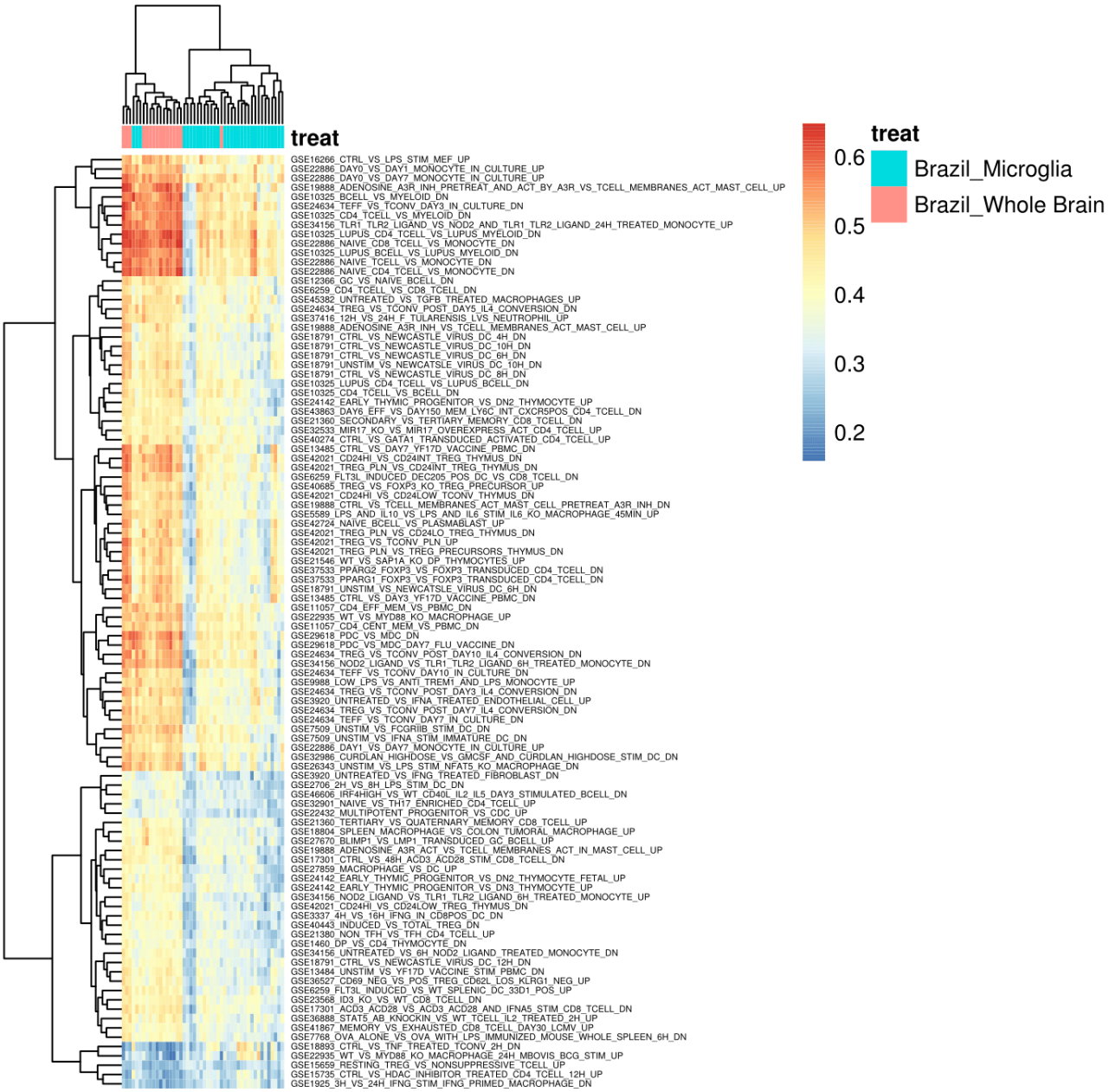
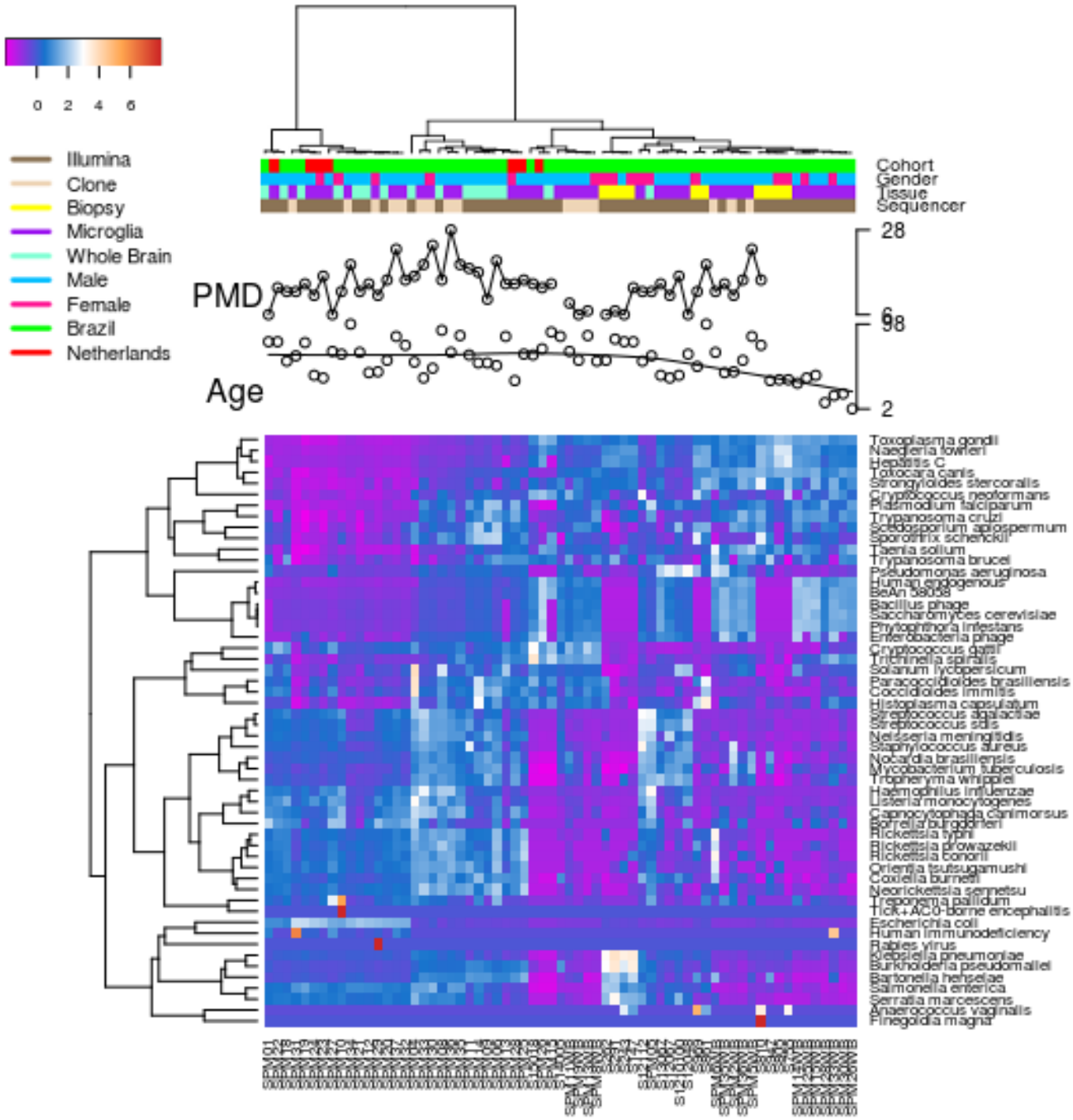


Figure 2: Heatmap of immunologic pathways of Brazil microglia vs Brazil whole brain

Appendix 5


0 2 4 6
Illumina
Clone
Biopsy
Microglia
Whole Brain
Male
Female
Brazil
Netherlands
Cohort
Gender
Tissue
Sequencer
PMD
Age
28
6
98
2
Toxoplasma gondii
Naegleria fowleri
Hepatitis C
Toxocara canis
Strongyloides stercoralis
Cryptococcus neoformans
Plasmodium falciparum
Trypanosoma cruzi
Scedosporium apiospermum
Sporothrix schenckii
Taenia solium
Trypanosoma brucei
Pseudomonas aeruginosa
Human endogenous
BeAn 58058
Bacillus phage
Saccharomyces cerevisiae
Phytophthora infestans
Enterobacteria phage
Cryptococcus gattii
Trichinella spiralis
Solanum lycopersicum
Paracoccidioides brasiliensis
Coccidioides immitis
Histoplasma capsulatum
Streptococcus agalactiae
Streptococcus sobs
Neisseria meningitidis
Staphylococcus aureus
Nocardia brasiliensis
Mycobacterium tuberculosis
Tropheryma whipplei
Haemophilus influenzae
Listeria monocytogenes
Capnocytophaga canimorsus
Borrelia burgdorferi
Rickettsia typhi
Rickettsia prowazekii
Rickettsia bonarii
Orientia tsutsugamushi
Coxiella burnetii
Neorickettsia sennetsu
Treponema pallidum
Tick-AC0-borne encephalitis
Escherichia coli
Human immunodeficiency
Rabies virus
Klebsiella pneumoniae
Burkholderia pseudomallei
Bartonella henselae
Salmonella enterica
Serratia marcescens
Anaerococcus vaginalis
Finegoldia magna

Samenvatting

Het immuunsysteem reageert snel wanneer het geactiveerd wordt door een infectie. De immuunrespons activeert vele cellen, zoals neutrofielen, macrofagen en microglia. Microglia zijn de hoofd kandidaten van het centrale zenuwstelsel die zich kunnen gedragen als antigeen presenterende cellen en kunnen een fagocyterende functie aannemen. Tijdens een infectie worden microglia geactiveerd. Ze gaan snel delen, meer bewegen en gaan fagocyteren. Omdat microglia antigeen presenterende cellen en fagocyterende cellen zijn, kunnen ze delen van pathogenen opnemen. Wanneer microglia gesequenced worden, kan er DNA/RNA van pathogenen gedetecteerd worden.

In een onderzoek van Galatro and Holtman et al., is er een batch effect gevonden tussen het Braziliaanse cohort en het Nederlandse cohort. De hypothese was dat dit veroorzaakt werd door de blootstelling aan verschillende pathogenen. Het doel van dit onderzoek is om pathogenen te identificeren in specimens en te bepalen of er een verschil is tussen de cohorts. Om dit doel te behalen is er een Bowtie2 pipeline gemaakt die potentiële pathogenen identificeert in de specimens. De geïdentificeerde pathogenen worden geanalyseerd met een kwantitatieve expressie analyse en een genexpressie analyse.
Een ander doel is om te bepalen of er sprake is van contaminatie in de specimens. Contaminatie kan veroorzaakt worden door onhygiënische omstandigheden of gebruik van laboratorische DNA/RNA kits.

De specimens zijn tegen het humane genoom aligned. De sequenties die niet als onderdeel van het humane genoom herkend werden, zijn aligned tegen pathogenen die bekend staan om infecties te veroorzaken in het brein en die opgemerkt worden door microglia. Ongeveer 97-99% van de totale baseparen die niet alignen tegen het humane genoom zijn reeds nog ongeïdentificeerd. De baseparen per pathogeen zijn gebruikt om de verdeling tussen de cohorts te visualiseren door middel van R. Een significant resultaat is de grote hoeveelheid *E. coli* in de Braziliaanse microglia specimens. Een visualisatie van de correlatie van de eigenschappen laat een kleine impact van leeftijd zien op de gevonden pathogenen.

Voor zover kan het batch effect niet verklaard worden met de pathogenen die tot nu toe gevonden zijn. Er is geen significant verschil tussen de cohorts die het batch effect kunnen verklaren. Het is mogelijk dat er nog pathogenen niet geïdentificeerd zijn in de specimens die het batch effect mogelijk wel kunnen verklaren.