CHAPTER 1

General introduction and outline of the thesis
Chapter 1

Microglia

Brain macrophages
Just over a century ago, del Río-Hortega was the first to identify microglia and oligodendrocytes based on their morphology (del Río Hortega, 1920). In prior staining experiments, these cells did not fit the classification of neurons or astrocytes (the so-called first and second elements). Therefore, these ‘presumed adendritic’ cells were classified as the ‘third element’ (Pérez-Cerdá, Sánchez-Gómez and Matute, 2015). This classification was made based on the absence of visible cellular processes. However, a distinct visualization of these adendritic cells was achieved using a silver carbonate staining method (del Río Hortega, 1918; Sierra et al., 2016), where the cellular processes could be noticed. This led to the identification of microglia, who owe their name to their relatively small soma size. A small, round nucleus is surrounded by a thin cytoplasmic layer from which several small processes originate, which gradually become thinner and form complex projections (del Río Hortega, 1918). This microglia morphology is known as ‘ramified’, and it is currently assumed that this structure serves the tissue surveillance function of microglia (Kreutzberg, 1996; Nimmerjahn, Kirchhoff and Helmchen, 2005). Furthermore, upon loss of homeostasis, microglia become activated and gradually convert their ramified structure to adopt a more amoeboid phenotype (Fernández-Arjona et al., 2017). Initially, microglia were assumed to have the same neuroectodermic origin as other central nervous system (CNS) cell types (discussed in Tremblay et al., 2015). However, in 1972, microglia were suggested to belong to the lineage of myeloid cells and were included in the mononuclear phagocyte system (van Furth et al., 1972). This evidence was mostly based on the appearance of these cells after blood vessel development, the observation that most microglia originate from circulating monocytes during pathology, and similarity in morphology, function, and kinetics with other macrophages (van Furth et al., 1972; Ginhoux and Prinz, 2015; Prinz, Jung and Priller, 2019).

Maturation of microglia
The origin and developmental stadia of microglia differ from other peripheral macrophages. Microglia are already present at early stages of embryonic development (Del Rio-Hortega, 1939; Alliot, Godin and Pessac, 1999), and are essential during early brain development (Schlegelmilch, Henke and Peri, 2011;
Swinnen et al., 2013; Ginhoux and Prinz, 2015). Microglia development is preserved across species and starts as a subset of erythro-myeloid progenitor (EMPs) cells, which develop in the yolk sac at embryonic day 8.5 (E8.5) in mice (Ginhoux et al., 2010). Starting at E9.5-E10.5, cell maturation can be divided into three consecutive phases (Matcovitch-Natan et al., 2016): early microglia (E10.5-E14), pre-microglia, (E14.5 until a week after birth) and adult microglia (few weeks after birth) (Figure 1). Early microglia migrate from the yolk sac to the brain until the blood-brain barrier has been formed at E13.5. In the brain, they continue to replicate and proliferate to increase their cell number (Swinnen et al., 2013). The ‘neural’ environment instructs microglia to adopt their unique identity, supportive of CNS development (Butovsky et al., 2014; Matcovitch-Natan et al., 2016; Hammond et al., 2019). The progression from early- to pre-microglia is accompanied by a profound change in morphology, where amoeboid microglia adopt a ramified cell structure with one or more protrusions at E16.5 (Kierdorf et al., 2013; Swinnen et al., 2013). In addition, pre-microglia are primarily involved in functions related to nervous system development and cellular assembly (Matcovitch-Natan et al., 2016; Thion et al., 2018). At postnatal week 6, the microglia population is stabilized, and self-sustained, there is no replenishing by infiltration of peripheral monocytes. They make up approximately 10% of all cells in the brain (Lawson, Perry and Gordon, 1992), and maintain this number (Nikodemova et al., 2015) through cell proliferation and intrinsic apoptosis, which are temporally and spatially coupled (Askew et al., 2017). A slow proliferative process of microglia is observed in both mice and humans (Askew et al., 2017; Réu et al., 2017). In humans, approximately 28% of all microglia cells are estimated to be renewed each year and the lifetime of microglia is estimated to be 4.2 years (Réu et al., 2017).
Figure 1: Microglia stages and functions: Three consecutive microglia phases are indicated by different colors. During development, morphology changes from an ameboid cell, to a cell shape with protrusions, and finally to a ramified shape. Each of these changes is accompanied by diverse functions. Early microglia are mainly involved in replication and proliferation. After the formation of the blood-brain barrier, pre-microglia support neuronal development. When microglia reach adulthood, they are involved in a variety of functions that are necessary to maintain brain homeostasis.

Microglia support a healthy brain environment

Microglia are the innate immune cells in the brain and important to maintain a healthy CNS environment (Figure 2). During their ‘surveillance’ state, microglia are stationary while the processes scan the surrounding environment for potential threats (Davalos et al., 2005; Nimmerjahn, Kirchhoff and Helmchen, 2005). Microglia express a set of genes termed ‘the sensome’ which contains surface receptors that detect signals from their environment (Hanisch, 2002; Hickman et al., 2013; Rodríguez-Iglesias, Sierra and Valero, 2019). When substantial disturbance of the homeostatic environment is observed, microglia become ‘reactive’ (Sierra et al., 2016; Prinz, Jung and Priller, 2019). Their morphology changes to an amoeboid shape, processes are retracted and the cell body size expands. Additionally, there is a difference in gene expression, and they start migrating to the affected location (Streit, Walter and Pennell, 1999; Hanisch, 2002). In this state, microglia phagocytose (Sierra et al., 2010), present antigens, secrete chemokines to attract more immune cells and release anti- or pro-inflammatory cytokines (Hanisch, 2002; Sierra et al., 2016; Wright-Jin and Gutmann, 2019). The interplay between surveilling and reactive microglia enables the maintenance and protection of other brain cells.
Figure 2: Microglia - homeostasis support: Microglia can perform a variety of functions to support CNS homeostasis. Starting at the top, several functions can be identified including: (A) release of cytokines to induce a pro- or anti-inflammatory response, (B) phagocytosis, (C) synaptic regulation, (D) antigen presentation, (E) transformation of precursor cells, (F) moderation in the neuronal network, and (G) scanning the nearby parenchyma.

**Microglia assisting other CNS cells**

Microglia are suggested to be involved in the regulation of myelination, and maturation and survival of oligodendrocytes, a process that includes repair and restoration of homeostasis (Nicholas, Wing and Compston, 2001; Clemente et al., 2013; Shigemoto-Mogami et al., 2014; Frost and Schafer, 2016; Wright-Jin and Gutmann, 2019). Besides, microglia play a role in the differentiation of neural progenitor cells into astrocytes (Nakanishi et al., 2007; Frost and Schafer, 2016) and neurons (Hou et al., 2020). Microglial interference in neuronal networks has been thoroughly investigated. Several activities include support of neural progenitor cells and neurons in survival, proliferation, and maturation (Butovsky and Weiner, 2018; Prinz, Jung and Priller, 2019; Tay, Carrier and Tremblay, 2019). This is done by regulation of the neural cell number through phagocytosis of apoptotic newborn cells (Sierra et al., 2010, 2013; Cunningham, Martinez-Cerdeno
and Noctor, 2013). After the formation of this neuronal network, microglia continue their support by regulating this network (Wright-Jin and Gutmann, 2019). Without this support, neurons are not likely to integrate into functional neuronal networks (Huang and Reichardt, 2001; Ueno et al., 2013). Furthermore, microglia regulate synaptic homeostasis (Ji et al., 2013; Parkhurst et al., 2013; Wake et al., 2013) by removing non-functional synapses and remodeling of the synaptic circuits (Kettenmann et al., 2011; Paolicelli et al., 2011).
Next generation sequencing

**Human genome project**
In 2001, the first two human genome sequence drafts were presented (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001), by the human genome project (HGP) and Celera Genomics, using different first-generation sequencing approaches. The HGP used a physical map approach which was based on Sanger sequencing, while Celera applied a whole-genome shotgun approach. The combined effort was presented as the “finishing” human genome sequence (International Human Genome Sequencing Consortium, 2004) and its reveal had a major impact on medicine and biology. The results of the human genome project and follow-up studies have shown that the human deoxyribonucleic acid (DNA) is more than 99% similar among individuals, and that diseases are often due to minor changes in the nucleotide sequence. These early findings have initiated massive efforts to investigate individual genomes to identify genetic functions, and deficits underlying disease. Research on biological pathways, gene networks, and molecular systems is still ongoing. In addition, the elucidation of the genome has led to further exploration of mutations, modifications, and expression changes that occur on various molecular levels (Collins et al., 2003; Giani et al., 2020).

**Genomics**
The genomics field contains several research domains including DNA sequencing, epigenomics, and transcriptomics. Encoded genetic information is the central component of these research fields with, as central element DNA, a double-stranded helical structure of nucleic acids (Watson and Crick, 1953). DNA consists of fixed nucleotide pairs, where adenine forms an interaction with thymine and cytosine bonds with guanine (Chargaff, 1950). These nucleotide combinations could be further analyzed to investigate aspects as genetic variation, characterization of gene (and protein) functions, and heredity (known as genetics) (Griffiths et al., 2000). One approach to analyze these aspects is through genome-wide association studies (GWAS) that relate genetic variants (e.g. single-nucleotide variants, copy number variants) to biological traits, including disease (Tam et al., 2019). In the case of disease-related research, the outcome of this analysis generally results in the identification of potential risk alleles of genes as important factors in disease pathology (The ENCODE...
Project Consortium, 2012; Kundaje et al., 2015). Although these studies have been proven to be effective, it is not possible to explain biological processes, including disease pathologies with only GWAS results. Generally, cells only use a part of the complete DNA template to differentiate into a specific cell type to perform cell-specific functions (Goldberg, Allis and Bernstein, 2007). Specific use and accessibility of DNA is regulated by changes in the epigenomic landscape, which explains the response of the epigenome to the environment for each cell (Waddington, 2014). Alterations in the microenvironment cause epigenetic changes, which in turn are involved in the regulation of the transcriptional process of DNA to ribonucleic acid (RNA). The complete collection of RNA molecules is called the transcriptome. Single-stranded RNA molecules act as a courier and mediate the translation to 3D protein structures, which can perform specific tasks in the cell. **Figure 3** provides a visual summary of mechanisms involved in genomics, epigenomics, and transcriptomics.

**Figure 3: Three aspects of genomics analyses**: Genomics entails investigation of the complete genome, whereas each level can be investigated using various approaches. (A) Generally, genetic analyses include GWAS studies that study genetic variants of the DNA. (B) Further exploration could reveal DNA regions that are accessible for transcription, and might explain cellular differences, these analyses are performed on the epigenome. (C) After transcription, a collection of RNA molecules could be quantified to specify the gene expression during various biological conditions.
**Epigenomics**

Although the same genetic information is stored in all cell types of the human body, every distinct cell displays a specific gene expression pattern (Kundaje et al., 2015). Furthermore, cell-specific gene expression profiles can be altered, for example in response to environmental changes. A major proportion of the chromosomes has a compressed structure that warrants the protection of chromosomal ends as well as chromosomal separation during mitosis (Kouzarides, 2007). The compact structure of DNA precludes gene transcription (Rivera and Ren, 2013). Before transcription can occur, an open chromatin structure is necessary. This process can be investigated at an epigenomic level, where an assembly of chemical factors regulate DNA accessibility, thereby regulating gene expression (DeAngelis, Farrington and Tollefsbol, 2008; Reddy, 2017). Negatively charged chromosomal DNA is wrapped and ordered using positively charged protein units called histones (Figure 4A, center panel) (Kornberg and Lorch, 1992; Maeshima et al., 2014). A histone H3-H4 tetramer and two histone H2A-H2B dimers form the core of a nucleosome, that can coil 147 DNA base pairs (Rivera and Ren, 2013) in approximately 1.65 turns (Luger et al., 1997). These base pairs are locked by histone unit H1 to enable a higher-order structure in chromatin (Allan et al., 1981; Fyodorov et al., 2018). The chromatin accessibility of the nucleosome can be altered using posttranslational modifications (PTMs) (Kouzarides, 2007; Winter and Amit, 2014).

Upon PTMs, including acetylation, methylation, and phosphorylation, chemical groups are added at various positions of the histone N-terminal tails using the corresponding enzymatic families. Each PTM involves or recruits unique chromatin-modifying enzymes, so-called readers, writers, and erasers that fulfill a unique purpose in this process. Readers are protein modules that can recognize a specific amino acid residue modification (Marmorstein and Zhou, 2014; Treviño, Wang and Walker, 2015; Hyun et al., 2017).

Writer and eraser proteins are responsible for the reversible modification, a process that consists of adding, or removing a chemical group on a marked nucleotide residue, respectively (Bowman and Poirier, 2015). The majority of DNA is part of a compact chromatin structure, which is less accessible for transcriptional processes (Allahverdi et al., 2011; Bannister and Kouzarides, 2011). A change of the chromatin structure can occur through various mechanisms, one example is the binding of a chemical residue on one of the N-terminal histone tails of the core nucleosome.
Figure 4: Epigenetic alterations: (A) A closed core nucleosome that makes interaction with non-histone proteins challenging. (B) Open chromatin which is realized by posttranslational modifications. For clarity, three modification types are visualized; acetylation, methylation, and phosphorylation. Two examples of chromatin remodeling are, (C) repositioning, where the location of the core nucleosome shifts, and (D) replacement, this type of remodeling enables the replacements of canonical histones by another variant.
This chemical change could convert the compact chromatin composition (Figure 4B) (du Preez and Patterton, 2013; Erler et al., 2014) to an open chromatin structure (Fu et al., 2017) which could stimulate or refrain various gene activity effects, depending on the modification type, location and degree (Grewal and Rice, 2004; DeAngelis, Farrington and Tollefsbol, 2008).

Additionally, chromatin composition can be remodeled by either adjusting the location of the core nucleosome using multiprotein complexes (Saha, Wittmeyer and Cairns, 2006; Teif and Rippe, 2009) or by replacing canonical histones with other histone variants that have a subtle change in their amino acid composition (Figure 4C) (Henikoff and Ahmad, 2005; Szenker, Ray-Gallet and Almouzni, 2011; Giaimo et al., 2019). These conversions are initiated by chromatin remodelers that are driven through the hydrolyzation of adenosine triphosphate (ATP) using large, multiprotein complexes (Saha, Wittmeyer and Cairns, 2006; Teif and Rippe, 2009; Ryan and Owen-Hughes, 2011).

In summary, initiation of chromatin alterations is a reaction to nucleosome characteristics, that can provide the necessary biochemical responses that are linked to specialized biological functions (Narlikar, Sundaramoorthy and Owen-Hughes, 2013), including DNA repair, sex chromosome inactivation, developmental regulation, and transcription (Henikoff, 2008; Talbert and Henikoff, 2010).
Figure 5: Chromatin sequencing methods: An overview of the differences among the discussed methodologies. (A) ChIP-seq is divided into a chromatin immunoprecipitation step that uses an antibody that is specific for a protein or histone modification. This step is followed by purification and amplification of the sequences, and is then sequenced (Raha, Hong and Snyder, 2010). (B) ATAC-seq can access and quantify the open chromatin using Tn5 transposase, this enzyme can fragment and ligate the open regions. After this step, the fragments are amplified and sequenced (Buenrostro et al., 2015). (C) Hi-C detects chromatin interactions using formaldehyde to crosslink the protein-DNA complexes. These crosslinked fragments are ligated, amplified and paired-end sequenced (Belton et al., 2012). (D) PLAC-seq is a method that combines the procedures of Hi-C and ChIP-seq. Crosslinking of proteins and DNA result in a sample fragment that is followed by a chromatin immunoprecipitation step that identifies the interaction of a particular protein (Fang et al., 2016). This figure adapted from the aforementioned papers, and the review of (Furey, 2012).
Although many aspects of the epigenetic landscape are known, new approaches are being developed to understand the effect of different PTMs, ‘open’ chromatin regions, interaction of distinct biological, and biochemical factors. The introduction of massive parallel sequencing, or high throughput sequencing led to the development of various methods that could quantify chromatin accessibility, and interactions. Examples of the sequencing methods that analyze chromatin accessibility include chromatin immunoprecipitation (ChIP) (Barski et al., 2007; Johnson et al., 2007) (Figure 5A), and assay for mapping transposase accessible chromatin (ATAC) through sequencing (Buenrostro et al., 2013) (Figure 5B). The ChIP-seq method includes a treatment with formaldehyde to crosslink proteins to DNA (Solomon, Larsen and Varshavsky, 1988), whereas ATAC-seq uses a hyperactive Tn5 transposase to cut open chromatin regions, and ligate sequencing adapters (Adey et al., 2010). Ultimately, purified fragments can be sequenced, and quantified to identify binding sites of DNA-associated proteins as transcription factors, polymerases, and epigenetic chromatin modifications (The ENCODE Project Consortium, 2012; Wang et al., 2012).

Sequencing methods that provide an overview of chromatin interactions include Hi-C sequencing (Belton et al., 2012) (Figure 5C), and proximity ligation-assisted ChIP-Seq (PLAC-seq) (Fang et al., 2016) (Figure 5D). Both approaches provide a spatial map that explains the interaction among genomic regions, that enable further exploration of the interaction between promoters and distal regulatory elements. In general, Hi-C sequencing is a commonly used approach to investigate extensive chromatin reorganization, and interactions that are formed in the genome to create a 3D model (Dixon et al., 2012; Rao et al., 2014). Whereas, PLAC-seq can provide a map of long-range chromatin interactions that reveal promoter-enhancer connections (Fang et al., 2016). Although both procedures focus on different aspects, they both contribute to a better understanding of the genome architecture, a feature that could identify novel targets that could be addressed in GWAS studies (Mishra and Hawkins, 2017). Altogether, the epigenome provides an overview that describes the epigenetic regulation, which is fundamental to understand the mechanisms that are shaping the corresponding gene expression.
Transcription factors RNA polymerase II Enhancer sites Preinitiation complex TFIIH and TFIIE bind to finalize the preinitiation complex. (B) Influences on the transcription process can be made by enhancer elements that are generally found at a longer distance from the promoter region. (C) RNA polymerase II transcribes DNA to a pre-mRNA molecule. Adding a 5’ cap and a 3’ poly-A tail to the sequence and removal of the intronic reads transform the pre-mRNA into a stable mRNA molecule. This figure is adapted from (Poss, Ebmeier and Taatjes, 2013), (Gupta et al., 2016), and (Haberle and Stark, 2018).

Transcriptomics

As a result of chromatin remodeling and modification, open DNA loci are created that can bind non-histone proteins (e.g. RNA polymerase II and regulatory proteins) that initiate a cell type specific gene transcription (DeAngelis, Farrington and Tollefsbol, 2008; Zentner and Henikoff, 2013; Jambhekar, Dhall and Shi, 2019). Assembly of general transcription factors (TFs) such as the TATA-binding protein (TBP), TFIIA, TFIIB, TFIIF, TFIIE, and TFIIH forms a complex that can bind RNA polymerase II (pol II) at an open promoter region (Figure 6A) (Sainsbury, Bernecky and Cramer, 2015; Petrenko et al., 2019). The binding of TBP is the first step, it creates a recognition site for the other general TFs to bind. Next, TFIIA and TFIIB bind to TBP to create a docking point for the pol II/TFIIF complex. Finally, TFIIE and TFIIH bind downstream of the complex to protect the sequence preceding pol II from cleavage, and forming the preinitiation complex (Figure 6A) (Buratowski et al., 1989; Conaway and Conaway, 1993; Horn, Kugel and Goodrich, 2016). Furthermore, hydrolysis of ATP is necessary for transcription to occur (Serizawa, 1997). After the formation and activation of the preinitiation complex,
the nucleotides on DNA are one-on-one transcribed to a pre-mRNA molecule (Figure 6C). This molecule consists of intronic and exonic parts, also known as the non-coding intragenic and coding expressing regions, respectively. The next step includes several adaptations as adding a cap at the 5’ termini (Furuichi et al., 1975), removal of the intronic sequences, and the ligation of a poly(A) tail at the 3’ end of the sequence (Lim and Canellakis, 1970). Only then a stable messenger RNA (mRNA) molecule is generated that after cytoplasmic transport serves as a template for protein synthesis (Jacob and Monod, 1961; Furuichi, LaFiandra and Shatkin, 1977; Bernstein, Peltz and Ross, 1989).

Besides the open chromatin regions, and binding of general transcription factors, gene expression can be altered by regulatory elements that can be located at a long distance from the promoter of the respective gene (Miele and Dekker, 2008). Initial demonstration of an enhancing effect dates back to 1981, where a 72 base pair repeat sequence motif of the simian virus 40 genome was able to increase the expression of β-globulin (Banerji, Rusconi and Schaffner, 1981). Enhancers can bind TF, and stimulate the promoter site using activators (Banerji, Rusconi and Schaffner, 1981; Serfling, Jasin and Schaffner, 1985). This process can create a structure where the distant enhancer is folded near the promoter region, creating a chromatin loop, where it can promote the recruitment of pol II (Figure 6B) (Carter et al., 2002).

By analyzing the diversity and number of these transcripts, it is possible to create a transcriptome profile that represents the current cellular state (Wang, Gerstein and Snyder, 2009). Initially, mRNA was converted into expressed sequencing tags (ESTs), a 200-500 base pair long sequence of an expressed gene. These ESTs were used to generate a high-resolution genome map that consisted of various genes and their chromosomal location (Adams et al., 1991). Unfortunately, comparisons of different studies based on relative abundant transcripts were limited due to the differences in mRNA expression from individual genes (Weinstock et al., 1994). This issue was resolved after the introduction of technologies as microarray, where fluorescent probes could be quantified to measure a difference in gene expression between two mRNA samples (Schena et al., 1995). The development of high throughput sequencing led to the creation of the bulk RNA-Seq technique (Figure 7A), a method that can be used to map and quantify transcripts in a tissue or cell population (Wang, Gerstein and Snyder, 2009). This provided the opportunity to explore the transcriptome to determine gene expression changes, or to identify novel genes in different tissues/cell types during different diseases and progressive stages.
Chapter 1

(Mele et al., 2015; Casamassimi et al., 2017). Recently, novel transcriptomic methods have been developed, as spatial transcriptomics (ST) (Ståhl et al., 2016) (Figure 7C), and single-cell RNA sequencing (scRNA-Seq) (Tang et al., 2009) (Figure 7B). In ST analyses, the origin of transcriptomic expression can be assigned to the position of a tissue slice, using positional molecular barcodes that provide a region-specific gene expression profile (Maniatis et al., 2019). ScRNA-Seq allows the quantification of the transcripts per cell, revealing a cellular clusters within the total population (Zeisel et al., 2015; Tasic et al., 2016). Differences between cellular clusters can reveal novel cellular markers (Wagner, Regev and Yosef, 2016). The abovementioned transcriptomic analyses can be used together to describe gene expression changes under different conditions, in different regions, and cells (Zeisel et al., 2018; Moncada et al., 2020).

Figure 7: Transcriptome sequencing methods: (A) Bulk RNA-Seq requires whole tissue (or sorted cells) that are used to investigate the gene expression of a whole population. (B) Single-cell RNA-Seq requires a unique molecular identified and/or barcode to distinguish the gene expression per cell. (C) Spatial transcriptomics can quantify the gene expression and visualize this based on the position on the tissue section. These sequences consist of a cleavage site, a T7 amplification and sequencing handle, a spatial barcode, and a unique molecular identifier besides the mRNA sequence. This figure is adapted from (Tang et al., 2009; Wang, Gerstein and Snyder, 2009; Ståhl et al., 2016)
Microglia transcriptome

Mouse core profile
By combining knowledge of chromatin alterations, and gene expression to define the current cellular state, it has been possible to outline the microglia gene expression profile (Eggen, Boddeke and Kooistra, 2017; Sousa, Biber and Michelucci, 2017). In macrophages, the TF PU.1 acts as a myeloid lineage master regulator that can modulate chromatin marks to influence the epigenome (Nerlov and Graf, 1998; Ostuni and Natoli, 2011; Pham et al., 2013; Holtman, Skola and Glass, 2017; Yeh and Ikezu, 2019). It binds to a purine-rich sequence, located near the promoter of target genes, to coordinate the attachment of other TFs and cofactors. This process initiates the regulation of gene expression (Smith et al., 2013). However, gene expression in tissue-specific macrophages are diverse and suggests that PU.1 cooperates with diverse TFs (Sousa, Biber and Michelucci, 2017). In-depth analysis in mice led to the identification of an epigenetic landscape, and associated gene expression patterns which were shown to be distinct among tissue-specific macrophages (Gosselin et al., 2014; Lavin et al., 2014). In addition, it revealed that microglia belong to the myeloid family. Although murine microglia show expression of Pu.1 and MafB, which is in line with other macrophages, they also express unique key factors that distinguish them from other macrophages. Examples are transcription factors Sall1/3, Cx3cr1, Irf8, and binding motifs of the SMAD and MEF2 family (Kierdorf et al., 2013; Gosselin et al., 2014; Lavin et al., 2014; Matcovitch-Natan et al., 2016). SMAD motifs are known to communicate with cytokines of the TGF-β superfamily (Macias, Martin-Malpartida and Massagué, 2015; Itoh et al., 2019), which contributes to a microglia-specific gene expression pattern (Butovsky et al., 2014). TGF-β deficient mice were shown to have a phosphorylated TGF-β-activated kinase (TAK1), a regulator of cell death (Mihaly, Ninomiya-Tsuji and Morioka, 2014), and a release of proinflammatory cytokines (Zölter et al., 2018). Binding motifs of the MEF2 family regulate several homeostatic genes (Yeh and Ikezu, 2019), that encode for anti-inflammatory genes as Il10 and reduce pro-inflammatory cytokines as Tnfα (Yang et al., 2015). Altogether, functionalities of the unique identified epigenome and transcriptome of microglial cells include the signatures as maintenance of CNS homeostasis (Buttgereit et al., 2016; Mass et al., 2016; Yeh and Ikezu, 2019), microglia-neuron interaction and neuronal maintenance (Reshef et al., 2017; Zölter et al., 2018).

Analysis of developmental stages in murine microglia led to the
observation that microglia maturation occurs in a gradual manner (Matcovitch-Natan et al., 2016). Epigenetic profiling revealed that the enhancer F13a1 is a unique marker for cells in the yolk sac. Identification of yolk sac-specific genes illustrated activity in the defense response, proliferation, and cell cycle. Early microglia continue proliferation, and cell cycle functions, similar to the cells in the yolk sac. Once microglia are accustomed to their microenvironment, they initiate a cascade of processes that transform early microglia into a pre-microglia stage. This pre-microglia stage can be recognized by the high expression of genes involved in regulation of cytokine secretion, neuronal migration, and development processes. Canonical TFs Sall1 and MafB are open during this stage and maintain accessibility during adulthood. Interestingly, early microglia and pre-microglia were shown to have a similar chromatin landscape, and were not subjected to a distinct chromatin remodeling. This suggests that there is no necessity for chromatin remodeling in large alterations of the transcriptomic profile. Furthermore, activation of unique adulthood enhancers (e.g. Irf8), were accompanied by a gene expression profile that consists of canonical microglia genes (Matcovitch-Natan et al., 2016).

**Human core microglia profile**

The major part of the transcriptome and epigenome in human and mouse microglia is well conserved (Galatro, Holtman, et al., 2017; Gosselin et al., 2017). Analysis of both species, illustrated that factors including PU.1, IRF, MEF2, SMAD, and MAF are similarly enriched. Additionally, TFs that are known to be involved in functional microglia roles as SALL1, STAT3, and RELA, were identified in both mouse and human (Gosselin et al., 2017). Although many aspects of the microglia epigenome and transcriptome are equivalent between mice and humans, some characteristics are unique for humans. First, various microglia transcriptions factors (e.g. SALL3 and SMAD1) were found to be less enriched in humans when compared to mice (Gosselin et al., 2017). Furthermore, some gene expression differences were observed (Gosselin et al., 2017; Geirsdottir et al., 2019). Second, only a fraction of disease-associated genes found in humans was expressed in rodents (Geirsdottir et al., 2019). That reveals essential differences in the microglia profile between humans and other species. Third, single-cell sequencing of human microglia revealed an organization of several microglia subtypes, regardless of sex. This observation differs from other mammals, where one microglia type was detected during homeostasis (Geirsdottir et al., 2019).
Big data era

Scientific data increase
In the past years, many scientific contributions led to further understanding of the complete human genome. The advancements in sequencing technology resulted in a decrease in costs, which enabled the generation of thousands deep sequenced human genomes that were used to report novel single nucleotide variants, and to highlight the importance to increase the quality of human genome sequencing (Telenti et al., 2016). Furthermore, it is expected that, in Europe one million sequenced human genomes will be generated in 2022 (Saunders et al., 2019). These initiatives provide great opportunities and novel exploratory analyses. Although the increase of computational science has major opportunities, it also has drawbacks concerning data storage and distribution (Marx, 2013), aspects that contribute to the trustworthiness of the research output (Elsevier, 2019).

The open science data principle
Various difficulties arise with the increase of big data. (Meta)data becomes more precise and can be obtained in a variety of formats. Several challenges include data processing and management, especially regarding privacy, security, and ethical characteristics. Moreover, the increase of big data leads to challenges in various technical aspects as capturing, integration, transformation, and analysis (Sivarajah et al., 2017; Papageorgiou et al., 2018; Navarro et al., 2019). Nowadays, reusage of data is not possible for many published datasets, which is often due to incomplete metadata, or unsystematical data storage. Ideally, we need to create and sustain an environment where (meta)data from different repositories can be shared among researchers (Spector-Bagdady et al., 2019). For this purpose, guidelines regarding data management are necessary to create a standardized workflow for storing (novel) data, and to promote data reusage for future analyses and technologies.

The FAIR concept provides four key principles, data needs to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Making data FAIR can be done using FAIRification, a workflow which is divided into three steps. The first step is pre-FAIRification, and includes the recognition of a FAIRification objective and the analysis of (meta)data. The next step, FAIRification, requires a definition of a conceptual model from the (meta)data. This concept is then used to describe the available items in the (meta)data. Furthermore, this information
needs to be connected to the (meta)data to integrate it with (future) applications. All of this information needs to be stored while providing access to humans and machines. The last procedure, post-FAIRification, evaluates the previous steps, to determine if the (meta)data is FAIR (Jacobsen et al., 2020). Currently, the FAIR principles are a central component in the concept for the Dutch Personal Health Train, where search algorithms will provide access to various data sources (e.g. biobanks, hospitals, and data repositories). A uniform defined and maintained infrastructure enables further interaction of these components, which results in (re)usage of patient (meta)data (Beyan et al., 2020). Another solution is data storage using databases that support the FAIR principles, of which MOLGENIS (Swertz et al., 2004) is an example. This platform consists of a versatile database structure that can be adjusted to one's scientific data, using the FAIR principles. Furthermore, it consists of many functionalities as storage, sharing, and visualization of data in a secured, high-performance application, which could be further modified using external coding scripts. Altogether, these platforms provide an opportunity for the research community and act as an example of how data should be shared in the future.

**Databases**

Many researchers benefited from the creation of the first scientific molecular database, the “Atlas of protein sequence and structure”, published in 1965. This encyclopedia contained information regarding the evolution of protein sequences in different species, the methodology of protein sequence comparisons, suggestions for protein notations including the one-letter amino acid code, and information regarding the amino acids (Dayhoff et al., 1965; ‘Margaret Oakley Dayhoff 1925–1983’, 1984). After 15 years of maintenance, this initiative led to the development of the first online nucleic sequence database (Dayhoff et al., 1981), one year before the creation of GenBank (Bilofsky et al., 1986). This action would later lead to the development of sequence search databases as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990). Computational research has expanded enormously in the past 5 decades, making it very difficult to create a database that consists of all major findings in one research field, similar to the protein atlas of Dayhoff. Nowadays, raw sequencing data is stored in the sequencing read archives (SRA) and European nucleotide archives (ENA). This information, together with processed data, can be found on archiving databases as the gene expression omnibus (Edgar, Domrachev and Lash, 2002; Barrett et al., 2012) and array express (Athar et al., 2019) which provide a detailed overview.
of the many datasets that are available in research fields. However, the accuracy and throughput of the curation on (meta)data or new datasets, and samples can and needs to be improved. Implementation of the FAIR guiding principles on these archiving databases, would lead to a reliable open-access environment allowing further exploration of published data (Wang, Lachmann and Ma’ayan, 2019).

**Reusage of sequencing data**

For bioinformatic analysis, SRA files need to be downloaded from one of the archiving databases. Often these files are formatted as a FASTQ file, that stores the biological sequences, and base quality scores (Cock et al., 2010). After downloading, these files can be processed using a variety of functions (**Figure 8**) before performing statistical analyses (Robinson, McCarthy and Smyth, 2010; Love, Huber and Anders, 2014; Ritchie et al., 2015). Each of these steps require bioinformatics knowledge to use the appropriate tools, parameters, and operating system.

Altogether, the integration of published and novel data is time-consuming and can be complicated for researcher that are unfamiliar with the analysis methods. Due to the increasing demand for better data transmission, a large number of online web pages have emerged that provide access to the data of a single publication. In general, these web pages have a search option that allows a quantitative exploration of the available genes under various conditions (Olah et al., 2018; Hammond et al., 2019; Li et al., 2019; Van Hove et al., 2019). The convenience of these web pages lies in the interception of various time-consuming processes as quality checking, downloading, aligning, and analysis of the samples. Since these steps were performed already, other researchers can obtain the gene expression information more easily. However, the development of applications that provide a structured overview of publicly available data provides a better solution (Zhang et al., 2014; Holtman et al., 2015; Mancarci et al., 2017). Concluding, the influence of the aforementioned observations might lead to analyses that can combine different omics data and development of multi-omics applications that can process, and visualize results of novel analyses (Regev et al., 2017; Conesa and Beck, 2019).
Figure 8: From raw reads to count matrix: Raw fastq files undergo various steps, that can be divided into preparation of the reads, mapping and counting, adapted from (Batut et al., 2018; Doyle, Phipson and Dashnow, 2020). Initially, the raw data in fastq files need to be quality checked (Andrews, 2015), trimmed (Hannon Lab, no date; Krueger, 2012), and checked again to proceed to the next step, mapping. The second step includes the alignment of the reads to the reference genome (Langmead et al., 2009; Dobin et al., 2013; Kim, Langmead and Salzberg, 2015; Bray et al., 2016). The outcome needs to be quality checked to determine if the alignment outcome is correct, optionally there might be several quality checking steps that depend on the sequencing technology that has been applied in a specific dataset. The last step, counting, ensures that the read from the alignment are annotated to genes, and are quantified into a count matrix (Liao, Smyth and Shi, 2014; Anders, Pyl and Huber, 2015).
Thesis outline

Transcriptomic analyses are a major contributor in determining the gene expression pattern of tissues and cells like microglia. Basic aspects, regarding microglia transcriptomics, have been addressed in the introduction above, a more elaborate description is provided in chapter 2. Furthermore, in chapter 2, the functionality of microglia is addressed and described as a continuous spectrum, as opposed to prior microglia phenotype classification. Although many questions remain unresolved when the transcriptome of human microglia is compared to that of other species.

The human microglia transcriptome profile is obtained from cells that were isolated from post mortem (Galatro, Holtman, et al., 2017) or surgically (Gosselin et al., 2017) obtained tissue. In chapter 3 the effect of post mortem delay on microglia gene expression in mice was investigated. Analysis of various PMD time points led to the observation of 50 differentially expressed genes that showed a subtle change over time and was present in both mice and humans.

In chapter 4 the microglia RNA expression profile was derived from a study on two macaque cohorts. In total, 666 genes were characterized as the macaque microglia gene expression profile that was used to identify overlapping and non-overlapping genes in zebrafish, mice, and humans.

In chapter 5 the setup and outline of the glia database BRAIN-SAT is described. Seminal published papers of the glia research field, that contain RNA-Seq data, have been collected and processed to data tables. This data is accessible on an online accessible platform with interactive features, that can be used by scientists in the research field, to verify the expression of target genes or to identify novel targets for further analysis.

Finally, chapter 6 provides a summary and discussion of results from chapters 3-5. Furthermore, suggestions for further development are provided.