Data Release Date: July 9, 2021, **Dataset Version:** ng00105.v1

Release Information:
The first release (July 9, 2021) includes 1000Genomes Imputed GWAS and RNAseq data. There are also phenotype files (for both sets of data), genotype APOE values, and eQTL/sQTL summary statistics.

Microglia Genomic Atlas – MiGA:
The Microglia Genomic Atlas (MiGA) is a genetic and transcriptomic resource comprised of 255 primary human microglia samples isolated *ex vivo* from four different brain regions of 100 human subjects with neurodegenerative, neurological, or neuropsychiatric disorders, as well as unaffected controls. We performed systematic analyses to investigate sources of microglial heterogeneity, including brain region, age, and sex. We further performed expression and splicing QTL analyses in each region and performed a meta-analysis across the four regions to increase our discovery power. We then performed colocalization and used fine-mapping and microglia-specific epigenomic data to prioritize genes and variants that influence neurological disease susceptibility through gene expression and splicing in microglia. With this approach, we have built the most comprehensive resource to date of cis genetic effects on the microglial transcriptome and propose underlying molecular mechanisms of potentially causal functional variants in several brain disorders.

Dataset Description:
Human post-mortem brain samples were obtained from the Netherlands Brain Bank (NBB) and the Neuropathology Brain Bank and Research CoRE at Mount Sinai Hospital. The permission to collect human brain material was obtained from the Ethical Committee of the VU University Medical Center, Amsterdam, The Netherlands, and the Mount Sinai Institutional Review Board. For the Netherlands Brain bank, informed consent for autopsy, the use of brain tissue and accompanied clinical information for research purposes was obtained per donor ante-mortem.

DNA Genotyping
Samples were genotyped using the Illumina Infinium Global Screening Array (GSA). Genotype imputation was performed for those 90 donors through the Michigan Imputation Server v1.4.1 (Minimac 4) using the 1000 Genomes (Phase 3) v5 (GRCh37) European panel and Eagle v2.4 phasing in quality control and imputation mode with rsq filter set to 0.3. Following imputation, variants were lifted over to the GRCh38 reference to match the RNA-seq data using Picard liftoverVCF and the “b37ToHg38.over.chain.gz” liftover chain file.

RNA extraction and sequencing
RNA was isolated using RNeasy Mini kit (Qiagen) adding the DNase I optional step or as described in detail before (Melief J, et al., 2016). Library preparation was performed at Genewiz using the Ultra-low input system which uses Poly-A selection. SMART-Seq v4 Ultra Low Input RNA Kit was used for library construction using 100 ng of RNA. The libraries were

sequenced as 150 bp on fragments with an average read depth of 29 million (ranging from 14-82M) read pairs on the Illumina HiSeq 2500.

RNA-seq data processing

RNA-seq data was processed using the RAPiD pipeline (Wang YC, et al., 2015). RAPiD aligns samples to the hg38 genome build using STAR (Dobin A, et al., 2013) using the GENCODE v30 transcriptome reference and calculates quality control metrics using Picard. RNA-seq quality control was performed applying three filters to remove samples: 1) samples with less than 10M reads aligned from STAR; 2) samples with more than 20% of the reads aligned to ribosomal regions; 3) samples with less than 10% of the reads mapping to coding regions; 4) samples from brain regions with fewer than 20 donors. Estimated transcript abundance was obtained using RSEM (Li B and Dewey CN, 2011) and transcripts were summed to the gene level with tximport (Love MI, et al., 2017). Genes with more than 1 read count per million (CPM) in 30% of the samples were kept for downstream analysis. Gene level read counts were normalized as transcripts per million mapped reads (TPM) to adjust for sequencing library size differences.

Quantitative Trait Loci mapping

To perform expression QTL (eQTL) mapping, we followed the latest pipeline created by the GTEX consortium (Aguet et al. 2019). We completed a separate normalization and filtering method to previous analyses. Gene expression matrices were created from the RSEM output using tximport (Love, Soneson, and Robinson 2017). Matrices were then converted to GCT format, TMM normalized, filtered for lowly expressed genes, removing any gene with less than 0.1 TPM in 20% of samples and at least 6 counts in 20% of samples. Each gene was then inverse-normal transformed across samples. After filtering, we tested a total of 18,430 genes. Then, PEER (Stegle et al. 2012) factors were calculated to estimate hidden confounders within our expression data. We created a combined covariate matrix that included the PEER factors and the first 4 genotyping ancestry MDS values as input to the analysis. We tested numbers of PEER factors from 0 to 20 and found that between 5 and 10 factors produced the largest number of eGenes in each region.

To test for cis-eQTLs, linear regression was performed using the tensorQTL (Taylor-Weiner et al. 2019) cis_nominal mode for each SNP-gene pair using a 1 megabase window within the transcription start site (TSS) of a gene. To test for association between gene expression and the top variant in cis we used tensorQTL cis permutation pass per gene with 1000 permutations. To identify eGenes, we performed q-value correction of the permutation P-values for the top association per gene (Storey 2003) at a threshold of 0.05.

We performed splicing quantitative trait loci (sQTL) analysis using the splice junction read counts generated by regtools (Feng et al. 2018). Junctions were clustered using Leafcutter (Li et al. 2018), specifying for each junction in a cluster a maximum length of 100kb. Following the GTEx pipeline, introns without read counts in at least 50% of samples or with fewer than 10 read counts in at least 10% of samples were removed. Introns with insufficient variability across samples were removed. Filtered counts were then quantile normalized using prepare_phenotype_table.py from Leafcutter, merged, and converted to BED format, using the coordinates from the middle of the intron cluster. We created a combined covariate matrix that included the PEER factors and the first 4 genotyping ancestry MDS values as input to the analysis. We mapped sQTLs with between 0 and 20 PEER factors as covariates in our

QTL model and determined 5 to be optimal in MFG, STG and THA. 0 PEER factors were used for SVZ.
To test for cis sQTLs, linear regression was performed using the tensorQTL nominal pass for each SNP-junction pair using a 100kb window from the center of each intron cluster. Although junctions were initially grouped together into clusters, we tested each SNP-junction pair separately, which is the standard approach (Li et al. 2018; Aguet et al. 2019). To test for association between intronic ratio and the top variant in cis we used tensorQTL permutation pass, grouping junctions by their cluster using --grp option. To identify significant clusters, we performed q-value correction using a threshold of 0.05.

File Manifest: https://st1.niagads.org/portal/download-public/NG00105.v1/fm

Subject Consents:
Sequenced subjects in this dataset belong to the following consent levels as indicated by the submitting study IRBs:

Consent Level*	# Subjects
GRU-IRB-PUB	108
Total	108

*Consent level definitions can be found on the Data Use Limitations page.

Dataset Accession Numbers Available in ng00108.v1:

Type	Description	Accession
Dataset	MiGA – Microglia Genomic Atlas	NG00105
Study	MiGA – Microglia Genomic Atlas	sa000018
Sampleset	MiGA – Microglia Genomic Atlas	snd10022
Fileset	MiGA – Microglia Genomic Atlas – GWAS Data	fsa000008
Fileset	MiGA – Microglia Genomic Atlas – QTL Summary Statistics	fsa000009
Fileset	MiGA – Microglia Genomic Atlas – RNASeq Data	fsa000010

Related Publications:
KATIA DE PAIVA LOPES*, GIJSJE SNIJDERS*, JACK HUMPHREY* et al. “Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies”. bioRxiv, 2020.
Link: https://www.biorxiv.org/content/10.1101/2020.10.27.356113v1