Project[1] Number: 803984

Project Acronym: GIDE

Project title: Molecular diversification of inhibitory neurons during development

DATA MANAGEMENT PLAN

[1] The term ‘project’ used in this template equates to an ‘action’ in certain other Horizon 2020 documentation

1. Data Summary

This project uses single-cell RNA sequencing (scRNAseq) to study the developmental diversification of inhibitory neurons. In scRNAseq experiments, RNA sequence information from individual cells is obtained with optimized next-generation sequencing (NGS) technologies. Statistical analyses of the sequencing data will be performed using R and Python programming languages. Raw data files are obtained in the following formats:

1) BCL file format: The NextSeq, HiSeq, and NovaSeq Sequencing Systems that are used in this proposal generate raw data files in binary base call (BCL) format. This sequencing file format requires conversion to FASTQ format for use with data analysis tools used in this proposal.

2) FASTQ file format: This is a text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems, and can be used as input for a wide variety of secondary data analysis solutions.

The datasets that will be generated in this proposal will be used together with publicly available datasets provided e.g. by the international public repository “Gene Expression Omnibus” (GEO). The origin of the existing data are previous publications including our own.

In this project about 4TB of NGS raw data files will be generated. They will be useful to other basic research institutes

2. FAIR data

2. 1. Making data findable, including provisions for metadata

After publication, data used in this project will be made discoverable with metadata, identifiable and locatable by means of a standard identification mechanism. We will use the naming convention established by the GEO. The GEO deposit procedure enables and encourage submitters to supply MIAME and MINSEQE compliant data. Furthermore, it provides a keyword search to optimize possibilities for re-use. For GEO submission, additional information is supplied by completing all fields of a metadata template spreadsheet. In this context, metadata refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. A clear version number will be provided.

In addition to making raw data files publicly available we will deposit scripts to reproduce our analyses on Github or Gitlab.

2.2. Making data openly accessible

All data used in this proposal will be made available after publication in peer reviewed journals. Raw data files and processed data files will be made available via the GEO. GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. R and python scripts will be made available via Github or Gitlab repositories. GEO creates links to experiment family downloads in various formats and supplementary files are provided at the foot of each GEO Series record. These files are compressed using gzip (.gz or .tgz extension). To unzip and read these files, a utility such as WinZip or 7-Zip can be used. No documentation about the software to access the data is needed. There is no need for a data access committee. The GEO offers well described conditions for access.

2.3. Making data interoperable

The data used in the project is interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries. We will use standard vocabularies for all data types present in our data set, to allow inter-disciplinary interoperability.

2.4. Increase data re-use (through clarifying licences)

Data submitted to GEO will be freely available. We will try to make our research data available as soon as possible. This usually is the moment when our research articles get published. The data used in the project will be useable by third parties after the end of the project. There is no intended limit on the duration how long the data remain re-usable. Data quality assurance processes will be described in the research articles.

3. Allocation of resources

GEO and GitHub are free. Gitlab is provided by the Max-Planck-Gesellschaft, so there are no additional costs for this project. Usually, one of the authors of the research article will be responsible for data management in this project. Long term archiving of all research data will be realised by the Max-Planck-Gesellschaft

4. Data security

We follow the data safety and protection standards of the Max-Planck-Gesellschaft.

5. Ethical aspects

There are no ethical or legal issues that can have an impact on data sharing. No personal data will be published in this project

6. Other issues

We do not make use of other national/funder/sectorial/departmental procedures for data management.