The discovery of viable diagnostic biomarkers for Lewy body dementia using machine learning algorithms.

A Data Management Plan created using DMPonline

Creator: Thomas Goddard

Affiliation: University of Nottingham

Template: University of Nottingham generic Data Management Plan

Project abstract:

Existing diagnostic capabilities within Lewy body dementias (LBD) are limited. Almost half of the people with LBD within the UK are misdiagnosed, most commonly as Alzheimer’s disease (AD). Accurate sub-typing of LBD is required to avoid the prescription of antipsychotic medication and for creating effective management plans. A recently developed machine learning algorithm has been shown to identify the most effective combination of diagnostic biomarkers and their optimum diagnostic cut-off values, and has displayed utility in biomarker optimisation for early-stage cancer screening. This proposal describes the application of this novel machine learning algorithm on transcriptomic data, to identify diagnostic biomarkers that may differentiate cases of LBD, from cases of AD and controls without dementia. RNA will be extracted from post-mortem occipital lobe samples and RNA-sequencing will be performed by the University College London genomic facility. Differentially expressed genes and transcripts, and alternatively spliced genes will be identified through experimentally validated methods. We will apply a novel hybrid multi-objective search and optimisation algorithm to our transcriptomic data. We will apply this protocol to our normalised RNA-sequencing read-count matrix, and identify the optimum combination of differentially expressed transcripts that can be adapted as diagnostic biomarkers to differentiate LBD from AD and controls.

ID: 83591

Start date: 01-06-2022

End date: 31-05-2023

Last modified: 10-09-2021

Copyright information:

The discovery of viable diagnostic biomarkers for Lewy body dementia using machine learning algorithms.

Data description

What data will you create?

This study will create 88 sets of fast.gz files that will contain the raw RNA-sequencing data. These files will go through quality control on the Galaxy public sever and change into fastqsanger files. The fastqsanger files will be aligned and changed to BAM files using *HISAT2*. Differentially gene expression will be analysed via *edgeR* and provide a de-list-edger.tsv output, which is a result table from statistical testing, including fold change estimates and p-values. The *edgeR* output will also include De-list-edger.bed to indicate genomic locations, and edgeR_report.pdf that includes plots where significantly differentially expressed features are highlighted. *Salmon* quantification of transcript expression will produce a quant.sf file for each sample. *DEXSEQ* analysis will produce a .tsv file to summarise the differential expression of transcripts and alternative splicing. The functional analysis of differential expression will be performed by *enrichR* and will output text files. The machine learning algorithm will detect the optimal combination of transcripts and diagnostic thresholds, and will produce a summary table. The formats and software used are validated by existing literature, and will facilitate the repetition of procedures. The formats described above will enable sharing of data in both private and public domains.

Data collection / generation

What are your methodologies for data collection / generation? How will you ensure data quality? What data standards will you use?

The protocol described in this proposal will replicate methods that have been described by our research group previously. The RNeasy plus universal mini kit (Qiagen, Hilden, Germany) will be used to extract total RNA from cryo-prepared occipital tissue. The NanoDrop ND-2000 (Thermo Fisher Scientific, Waltham, USA) and the Agilent TapeStation (Agilent Technologies, Santa Clara, USA) will be used to ensure quality and quantity. The cDNA library preparation and sequencing will utilise two µg of RNA and be completed by the UCL genomics facility. Ribosomal RNA depletion will be applied to the RNA samples and the cDNA libraries will be prepared using the TruSeq RNA sample preparation kit (Illumina, San Diego, USA) to further ensure quality. The Illumina Novaseq (Illumina, San Diego, USA) will be used to perform paired-end sequencing (56 base pairs/read) and a minimum of 40 million reads per sample will be obtained. The quality control methods described above will ensure the quality

collection of data. The methods of data collection and quality control will be document in a PhD thesis and be published in peer-reviewed journals.

Data storage and security

Where and how will data will be stored, backed-up, transferred, and secured during the active phase (short to medium term) of research?

We will use UoN-provided storage for our working data. UoN licenses Microsoft OneDrive, an ISO 27001 information security management compliant service that allows secure and controlled sharing of data amongst the research team. University of Nottingham OneDrive encrypts data both in transit and at rest and is approved against the University's Handling Restricted Data Policy. The service provides continual failover support. This service provides up to 5TB free-at-point-of-use, and as we do not anticipating generating more than 5TB we we will not require any additional costs for use of this service. Data will be backed-up on UoN-provided external hard-drives that are encrypted and password protected.

Data management, documentation, and curation

What are your principles, systems, and major standards for data management and creation? What metadata and documentation will you keep?

This study will not collect any personal data that can be used to identify individuals from the biological samples. This proposal will utilise samples from the Brains for Dementia Research (BDR) network of brain banks, which have ethical approval for UK-based studies (London City and East NRES committee 08/H0704/128+5). We will use UoN-provided Microsoft OneDrive storage for our working data. There will be a folder for each stage of data collection and quality control. The files for individual samples will be named according to the brain region, study group, and the randomly allocated sample number (e.g. OCC_LBD_1). All additional files from data processing will follow this format, and if files are combined to produce a single analysis file, the format will represent method used, brain region and study group (e.g. DEXSEQ_OCC_LBD).

Ethics & Privacy

Are there any ethical or privacy related issues associated with your data?

This study will not collect any personal data that can be used to identify individuals from

the biological samples. This proposal will utilise samples from the Brains for Dementia Research (BDR) network of brain banks, which have ethical approval for UK-based studies (London City and East NRES committee 08/H0704/128+5).

Data preservation

How will you ensure the long term storage and preservation of data?

All research data created by the project will be deposited in the UoN research data archive, https://rdmc.nottingham.ac.uk/ The UoN data archive is underpinned by commercial digital storage which is audited on a twice-yearly basis for compliance with the ISO 27001 standard. UoN will retain and preserve research data in line with Race against Dementia's requirements, but data will be retained for longer periods of time where it is of continual value to users.

Data sharing and access

How will the data generated be shared and published?

Our dataset does not contain any personal or commercially sensitive information and thus will be shared via the University of Nottingham data archive under a CC-BY license. There will be no need to update the data past the project period. All published outputs will contain a Data Availability Statement including the datacite DOI which directs to the relevant data set. Data will be released at the same time as any published outputs which are underpinned by the data or by 1 year from the end of the project at the latest.

Roles & responsibilities

Who will be responsible for managing data, data security, data quality, and data security both during the award and post-award?

The research supervisor, Dr Anto Rajamani, is the owner of data and has responsibility for all data within this study, in line with University of Nottingham policy. The primary investigator, Thomas Goddard, will act as a data steward, who will manage data on a day-to-day basis. The primary investigator will be responsible for collecting, processing and analysis of data, including transcription. Access to data will also be given to research supervisors (Dr Keeley Brookes, Prof Kevin Morgan and Dr Grazziela Figueredo), who have a responsibility to ensure data security and quality. The University of Nottingham has a responsibility to ensure data stored using Office365 remains secure.

Relevant policies

What are the relevant institutional, departmental or study policies on data sharing and data security?

We will ensure that our research aligns with the requirements of the University's Research Data Management Policy, Information Security Policy, Code of Research Conduct and Research Ethics. We will abide by the Secure Data Handling Policy and Data Protection Policy. All third party commercial data or new data that may be suitable for commercial exploitation will be protected by the University's Intellectual Property policy.

IPR

Who will own the copyright and IPR of any data that you will collect or create? Will you create a licence(s) for its use and reuse? If you are planning to use existing data as part of your research, do any copyright or other restrictions determine its use?

The intellectual property of data will remain with the University of Nottingham.

Budgeting

What are the costs or funding required for capturing, processing, storing, and archiving your data?

This project is being funded by the Race Against Dementia seed grant. The procedures by which data will be captured will amount to £18,213.59. Data processing will utilise the Digital Research Service at UoN and has been budgeted at £1,780.

Further Help

Would you like your plan to be reviewed by specialists in Libraries?

Saving this plan after checking the "Yes" box will immediately notify Libraries DMP review service, please only do this when you are ready for review.

- Yes