PROJECT MINE PROGRESS REPORT 2021

Introduction

Genetic studies of Amytrophic Lateral Sclerosis (ALS) have comprised four main types: candidate gene sequencing studies, family based linkage studies, genome-wide association studies (GWAS), and studies of copy number variation. These study designs have allowed the identification of rare gene variation contributing to familial risk and to common gene variation contributing to apparently sporadic ALS risk. The last remaining major type of gene variation, namely rare or moderate frequency variants contributing to ALS risk are to be identified. Large scale GWAS plus sequencing analyses show that the bulk of the heritability for ALS is in the rare to moderate frequency variants. These variants can only be captured exhaustively by next generation high throughput sequencing. This technology has matured to the extent that it is feasible financially and practically, with the remaining hurdle being interpretation of findings. The problem of interpretation arises because each individual harbors many rare variants that would be predicted to cause harmful effects, but without apparent hurt, suggesting that there are evolutionary buffers preventing deleterious gene variants from always causing harm. This means that the only way to determine if rare variants found in a gene implicate that gene in disease causation is to compare the frequency of rare variants between very large numbers of people with ALS and normal controls, including control sequences in public databases.

We therefore sequence the ALS samples available to many of us in several countries/biobanks using next generation sequencing as part of a multinational collaboration under the banner of Project MinE. By sharing data with similar projects from across Europe, Australasia and the US, we have the ability to identify new ALS genes with a high level of confidence, leading to increased understanding of the mechanism of ALS and a greater probability of developing diagnostic tests and effective therapies.

Project MinE is unique in several aspects:

1. Size: many population based sequencing projects use low coverage exome (WES) or whole genome sequencing (WGS). Coverage in Project MinE is effectively 45x, compared to 4-12X in population-based projects, including UK10K and GoNL. This means that individual genotyping is much more confident.
2. Harmonized and detailed data collection: the combined collection of core clinical data, as defined through collaborative projects in Europe (SOPHIA, Euro-MOTOR and STRENGTH) and Australia allows for further detailed analyses of genes that determine age at onset, progression through ALS stages and survival in ALS.
3. Improve ongoing GWAS efforts: sequences are used to improve imputation of genotypes in existing and ongoing large ALS GWAS datasets, while the NGS effort is growing.
4. Expression changes are mapped to intergenic or genic sequences using RNA seq or expression arrays with WGS, which is a clear advantage as this is not possible with WES.
5. WGS provides better and more complete coverage of the exome than exome sequencing (especially in “difficult” regions, i.e. GC rich or including repeats)

6. Data storage and processing is centralized but flexible: a setup is available to Project MinE at the SURF supercomputer. This means that all raw data are directly delivered "through the wire" at this supercomputer. Therefore, there is no need to keep track of many hard drives for data delivery. Partners of Project MinE have default access to their own data, and data can be shared after a formal data access procedure. Also, SURF allows for supercomputing using the data directly, i.e. without the need to download the data and perform calculations on local high performance compute solutions. These data storage and calculation hours were funded over the past years, through the Dutch ALS Foundation (Stichting ALS Netherland) and co-funding by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. For 2021 this was funded by the Dutch ALS Foundation (Stichting ALS Netherland)
7. Combined WGS data generation with methylation (until 2019): of every sample that was submitted to Illumina, we get WGS, plus 450K methylation and 2.5M OmniExpress GWAS chips. This allows for state-of-the art analyses on gene-environment interactions (alcohol, smoking, occupational), and sub clustering of patient groups based on methylation profiles. As of 2020 the main focus for generating new data is on WGS only, as the current methylation data set is quite extensive for research purposes. In 2021 the first new data set of WGS was generated through sequenced by Hartwig Medical Foundation.
8. Proper controls: a requirement for Project MinE participation is to submit cases and locally/ancestrally matched controls. This is to ensure that no population stratification or false positives are found, which is especially crucial with rare genetic variation. Another reason is the lower coverage these population-based datasets usually have. However, more and high quality control data are increasingly available. Therefore, we seek for collaboration where we can use appropriate control data to expand the data base and therewith the analytical power.
9. Availability of data to other consortia: anonymized Project MinE data (from the Netherlands) is part of the International Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/home). This project allows every researcher who has GWAS data to impute up their dataset to an unparalleled low level of minor allele frequencies, to help find new disease genes. This way, Project MinE helps facilitate the discovery of disease genes outside of ALS/MND. Since 2016 the overall data are made available to other researchers as well through the data browser in which researchers can go through >6,400 whole genomes from different European ancestry, and retrieve summary statistics.
10. A combined good price for data generation: due to the formation of a consortium with a "franchise" construction, we are able to negotiate favorable pricing for genomics data generation, while individual PI's keep total control of their data.

Power of Project MinE

Our goal is analyzing DNA profiles of at least 15,000 ALS patients and 7,500 locally/ancestry matched controls. Achieved power is of course dependent on aggregated allele frequency differences in specific genes between cases and controls. For example, to 'rediscover' SOD1 with sufficient certainty ('statistical significance') 2,200 ALS genomes and 1,100 controls are needed, for FUS and TARDBP mutations 6,000 ALS genomes and 3,000 controls are needed, and for gene 'X' with 0.5% allele frequencies in ALS cases while being nearly absent in controls, the whole set of 22,500 samples are needed.

Status of Project MinE – accomplished in 2021 and future perspectives

In general, 2021 was again a year in which the world was struck by a pandemic. This had a massive impact on day-to-day live and also on ALS research. Project MinE paused in some areas, like no live meetings, no fundraising events, no lab activities, etc. Fortunately, other activities were continued by seeking solutions through digital communication and remote-(data)access.

Despite the delaying impact of COVI-19 pandemic the following was achieved:

- 586 new DNA samples were collected and shipped to sequencing partner Hartwig Medical Foundation (these samples were provided by Spain, Belgium, Turkey, France and Israel)
- Funds were generated for new samples by FUNDELA (Spain), ARSLA (France) and ALS Liga (Belgium).
- An increase of the total number of DNA profiles funded (+14 France, +29 Spain) to 10,993 end of Dec 2021 (49% of total aim of Project MinE).
- Upload of 50,000 DNA profiles from UKBiobank to SURF integrated into the Project MinE.
- Therefore, harmonization process is set-up, in which data from external DNA data banks is cleaned and standardized for comparison with data from Project MinE.
- Development and use of cloud-based calculation tool to extract information at minimal costs from US data set (>30,000 profiles) to enrich Project MinE with new DNA profiles from ALS patients and controles.
- Processing 6 data requests; the data set of Project MinE is accessible for researchers through the Project MinE data browser (Data Browser | Project MinE) whereas more extensive data access (WGS datafreeze 2, holding 9,600 WGS profiles), can be requested through the Project MinE data request procedure. Access to this data set is organized through formal routing of requesting and granting, which after granting is formalized through a Data Sharing Agreement to comply with GDPR regulations. In addition, a renewed and updated publication policy was released to promote broad, transparent and responsible data sharing (FAIR principles).

- Posts summary GWAS data of Project MinE manuscript of W. van Rheenen, et al. Nature Genetics (2021) was made available as summary statistics, shortly after publication. https://www.projectmine.com/research/download-data/
- three new publications in peer-reviewed publications were delivered in which Project MinE data were used. Publications | Project MinE or search on PubMed – Project MinE Consortium
- Our two yearly Project MinE General Assembly meetings at ENCALS and MNDA symposium were cancelled due to COVID-19 pandemic.
- Even our 6th Project MinE Scientific meeting is due to COVID-19 pandemic, postponed to January of 2022, where scientists involved in Project MinE meet digital and working groups will give an update on their progress (phenotype data, gene burden testing, epigenetic data, infrastructure, structural variation and repeated elements, and non-coding variation).