

## PROJECT MINE PROGRESS REPORT 2019

---

### Introduction

Genetic studies of ALS have comprised four main types: candidate gene sequencing studies, family based linkage studies, genome-wide association studies, 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)

## PROJECT MINE PROGRESS REPORT 2019

---

6. Data storage and processing is centralized but flexible: a setup is available to Project MinE at the SURFsara 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, SURFsara 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 2018 and covered for 2019, through the Dutch ALS Foundation (Stichting ALS Nederland) and co-funding by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.
7. Combined WGS data generation with methylation: of every sample that is 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.
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

## PROJECT MINE PROGRESS REPORT 2019

---

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 2019 and future perspectives

**'New' ALS genes:** Project MinE data already contributed to the discovery of TUBA4A and TBK1. In 2016 two novel ALS genes, C21orf2 and NEK1, and 3 novel GWAS loci were discovered. The results were published in established peer reviewed research journal (Nature Genetics). In 2017 Project MinE scientists discovered a shared genetic origin for ALS/MND and schizophrenia. Knowledge of the shared biological pathways between these diseases will help to develop new treatments for ALS that are based on stabilizing disrupted brain networks. Results were published in Nature Communications. The year 2018 brought NIPA1 repeat expansions as proven risk factor for ALS (published in Neurobiol Aging), the identification of KIF5A as novel ALS gene (published in Neuron), and that CHCHD10 variants are not related to pure ALS as was suggested in earlier findings based on fewer data (Ann Neurol). This year exome sequencing also including Project MinE data implicated a novel gene, DNAJC7, encoding a heat shock protein that, when disrupted leads to protein aggregation, a pathological hallmark of neurodegeneration (Nature Neuroscience). (see all publications <https://www.projectmine.com/research/publications/>)

**International partners involved:** Project MinE includes the UK, the Netherlands, the USA, Ireland and Belgium. These countries are the “frontrunners” of the project. Italy, Spain, Turkey, Portugal and Israel are following and are committed to reach their target. Australia and Canada are organized now in a similar way to Europe, and transferred their first data sets to Project MinE. Principal Investigators (PIs) there have adopted the core clinical data definitions from Europe and follow the structure of Project MinE. In 2017 Sweden, Switzerland, France and Brazil were able to contribute to the project by securing samples and funds. In 2018 Russia and Slovenia started on collecting samples and funds, Croatia is preparing for participation. Mid 2018 Malta joined and contributed with already sequenced profiles to the project. In 2019 no new countries were joining, therewith end of this year a total of 19 (and one as tentative) countries were shown on the Project MinE website to be connected to the project (see Figure 1 and 2). Potential new countries have shown their interest in joining Project MinE in the (near) future such as Argentina, India and South Africa.

## PROJECT MINE PROGRESS REPORT 2019



Figure 1: Countries joined in Project MinE: [www.projectmine.com](http://www.projectmine.com)



Figure 2: Status page per country December 2019: [www.projectmine.com](http://www.projectmine.com)

## PROJECT MINE PROGRESS REPORT 2019

---

**Funds and events:** Funding is provided by the specific local ALS foundations (MNDA in the UK, Stichting ALS the Netherlands, ALS Liga in Belgium, AriSLA in Italy, ALSA in the US, MND Australia (MNDRIA), Prize4Life in Israel, Irish ALS in Ireland, FUNDELA in Spain, Kirac Foundation in Turkey, APELA in Portugal, ARSLA in France, Swiss ALS Foundation and the ALS Association Switzerland in Switzerland, ALS Canada in Canada, Muscular Dystrophy Association of Slovenia in Slovenia, and governments (the Netherlands, Belgium, Sweden ). Whereas Belgium reached their goal (all funds were raised for sequencing 750 samples) in 2015, The Netherlands touched the finish line of collecting funding for sequencing a total of 3,000 samples in March 2016. In 2017 Turkey and the UK reached their goals of 500 and 1800 profiles, respectively. In 2018 contributions were secured for Israel (60 profiles), Spain (96 profiles) and Portugal (26 profiles), followed by contributions for France (18 profiles) and Sweden (87 profiles) in 2019

The majority of the funding in the Netherlands and the USA is based on donations through the City Swims in Amsterdam (ACS) and New York (NACS, first edition) respectively. Other countries, such as Belgium, UK, Spain are in contact with the board of the City Swims to start up their own swim locally. The UK had their first swim in London September 2017. Special contributions were made by several foundations (MNDA, FUNDELA, Suna And Kiraç Foundation), by anonymous donors, through personal campaigns ('Als het licht uitgaat', 'Km solidarios Serra de Tramuntana', 'de Peramides a Veleros', 'El gran reto de Jorge Abarca', 'Mi penultimo reto por la ELA' 'Mary Bucles por un mundo sin ELA') and other very successful (sports) events such as the Good Run in Ireland. August has been announced to be internationally the month of the IceBucketChallenge. Various new local initiatives were welcomed and supported by the local ALS foundations and contributed to the increase in funds over this year in a lot of the countries.

**Number of samples sequenced:** By the end of 2019, Project MinE will have assembled an impressive number of WGS profiles with appreciable power in a relatively short time period. End December a total number of 10,398 WGS profiles were present as actual profiles (see Table 1). These samples were sequenced through sequence provider Illumina (San Diego, USA). The blood epigenetic data of these samples was completed in 2019 as well. Illumina had made errors in processing a subset of these arrays in the years prior, as it turned out while completing this dataset in 2019. Illumina offered to redo those samples and these reruns have finished satisfactorily (2,000 samples). Primary methylation analyses will be finished in 2020.

## PROJECT MINE PROGRESS REPORT 2019



Site 	N 
Australia-Sydney	597
Belgium	774
Brazil	68
Canada	210
France	376
Ireland	701
Israel	110
Italy	70
Malta	36
Netherlands	2963
Portugal	82
Spain	570
Sweden	362
Switzerland	54
Turkey	802
UK	2066
US	557
<b>s</b>	<b>10398</b>

Table 1: Number of samples sequenced in Project Mine (MinE December 2019)

In addition to the profiles as contributed through funding and samples collection in Project MinE the consortium also hosts data from other WGS projects worldwide (e.g. Answer ALS, CREATE, TOPMed and data at the Broadinstitute (US)). Already sequenced data from ALS patients and controls is imported to the Project at SURFsara to become available for the Project MinE researchers for analyses. A start was made combining external WGS data from patients with ALS and external control cohorts. Major challenge there was to realign and recall all MinE data, since most external datasets were aligned and called using GATK (Regier, et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. (*Nature Communications*, 9(1), 1–8), and not Illumina's standard pipeline (Isaac's). This is ongoing now using the supercompute infrastructure at SURFsara. Our goal is to realistically arrive in 2020 at a WGS dataset, harmonized, including at least 12,000 ALS samples and 30,000 controls. Limits are not compute power, expertise or samples, but further funding for sequencing. Nevertheless, this will be a unique dataset leading to many new important breakthroughs for ALS.

**Data storage:** After whole genome sequencing, Illumina sends the data to the supercomputer of SURFsara in the Netherlands, where the data is stored securely. SURFsara, a non-profit agency available for research, guarantees safe and fast storage of all petabytes of data for Project MinE. This is a crucial part of Project MinE, as it needs more capacity for storage and analyses than any project

## PROJECT MINE PROGRESS REPORT 2019

---

before. Researchers analyse their data on the SURFsara supercomputer. December 2019 almost 18,000 DNA profiles were available for analyses. The direct connection ensures that the transfer of data is safe and fast. The storage will be expanded as more data is being transferred over the years. These will come from newly sequenced samples, and from samples already sequenced within other smaller WGS projects. These consortia are willing to share that data to Project MinE. The calculation capacity on SURFsara is covered through the Project Beyond MinE and subsidary from PPS, Health Holland) and where data calculations will be prepared for research projects by experienced bioinformatics.

The data is made available for external researchers through the databrowser for exploring certain details of the genome through aggregated data. For more extensive research in the data set of Project MinE we have established a fixed set (datafreeze 2) containing 9,600 WGS profiles. Access to this data set is organized through formal routing of requesting and granting, which after granting is formalised through a Data Sharing Agreement to comply with GDPR regulations.

**Project organization:** One Project MinE meeting was held this year; one May 15<sup>th</sup> in Tours, France during the ENCALS Topics that are addressed in project meetings are: scientific progress, progress on sample collection and analyses, progress on fund raising, project organization matters.

The next Project MinE meeting will be organized at the ENCALS meeting 2020 in Edinburgh, UK. Besides the general Project MinE meetings there are Project MinE Science meetings. Here we set-up and coordinated the research efforts of those partners who have substantial data sets within Project MinE. We formed and structured working groups around six major topics for ALS genetic research. Each working group defined their aims, tasks and deliverables and reports to the General Assembly of the consortium every six months. The working groups are announced at the Project MinE website (research page). The fourth Science meeting was held on May 23<sup>th</sup> at Schiphol, the Netherlands to update and share progress within the six Working groups.

Since 2017 the Project MinE website is expanded with a (renewed) research page to show and share the scientific output generated from the Project MinE data. This page displays the scientific publications of the project, the Working groups with their actions and goals, the data browser and how to request for access to the data of Project MinE (data sharing). Also in 2019 we have received several requests for data sharing. Few of these requests were directed towards the use of the databrowser, whereas the others were granted access by the consortium members.

Besides the meetings communications between (the project coordinator and) partners are going through platform Basecamp, we structured a data access procedure as to facilitate the initiation of data sharing research projects between two or more partners which use the Project MinE data from those countries, and to stimulate partners to raise funds by providing researchers with up to date information on Project MinE for grant submissions and by supporting local fundraising organizations through advise or promotion through the Project MinE website and social media.

## PROJECT MINE PROGRESS REPORT 2019

---

### New sequencing facility:

Illumina in San Diego discontinued their Fast Track Service in 2019. This meant, we had to look for a new sequencing provider. We have explored many options and evaluated all providers using a fixed set of parameters (including pricing, turn around time, flexibility regarding failed samples, method of data delivery, etc.). We have chosen the Hartwig Foundation, based on this exercise (<https://www.hartwigmedicalfoundation.nl/en/>). There was one “disadvantage”: they had never used PCR-free library preparation, as they are used to receiving very small amounts of tissue/ DNA. PCR-free library preparation is crucial to Project MinE as all previous 9,600 samples have been processed in that way, it leads to better (i.e. more even) coverage genome-wide, and it is crucial to be able to detect repeat expansions such as in *C9orf72*.

Hartwig was willing to test a number of samples from us, and we have evaluated these. The good news is that the data quality is on par, if not better, so next batches will be sent to Hartwig in 2020.