PROJECT MINE PROGRESS REPORT 2018

Introduction

Genetic studies of ALS have to date 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. We are now in a position to identify the last remaining major type of gene variation, namely rare or moderate frequency variants contributing to ALS risk. Our most recent large scale GWAS plus sequencing analyses show that the bulk of the heritability for ALS is to be expected in the rare to moderate frequency variants. These variants can only be captured exhaustively by next generation high throughput sequencing. This technology has now 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 propose to 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 will 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 will be much more confident.
2. Harmonized and detailed data collection: the combined collection of core clinical data, as defined through already existing collaborative projects in Europe (SOPHIA, Euro-MOTOR and STRENGTH) and Australia will allow 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 can be used to improve imputation of genotypes in existing and ongoing large ALS GWAS datasets, while the NGS effort is growing.
4. Expression changes can be 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 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 2017 and covered for 2018, through the Dutch ALS Foundation (Stichting ALS Netherland).
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. We cannot, therefore, solely rely on publicly available control sets such as UK10K or 1000G and ExAc. Another reason is the lower coverage these population-based datasets usually have.
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. In 2016 we launched the first version of the data browser in which researchers can go through >6,400 whole genomes from different European ancestry, and retrieve summary statistics. Herewith the overall data are made available to other researchers as well.
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

We have set a goal of 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 2018 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. This 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). (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 have 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. Herewith December 2018 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). Several potential new countries have shown their interest in joining Project MinE in the (near) future such as Argentina, India and South Africa.

Figure 1: Countries joined in Project MinE: www.projectmine.com

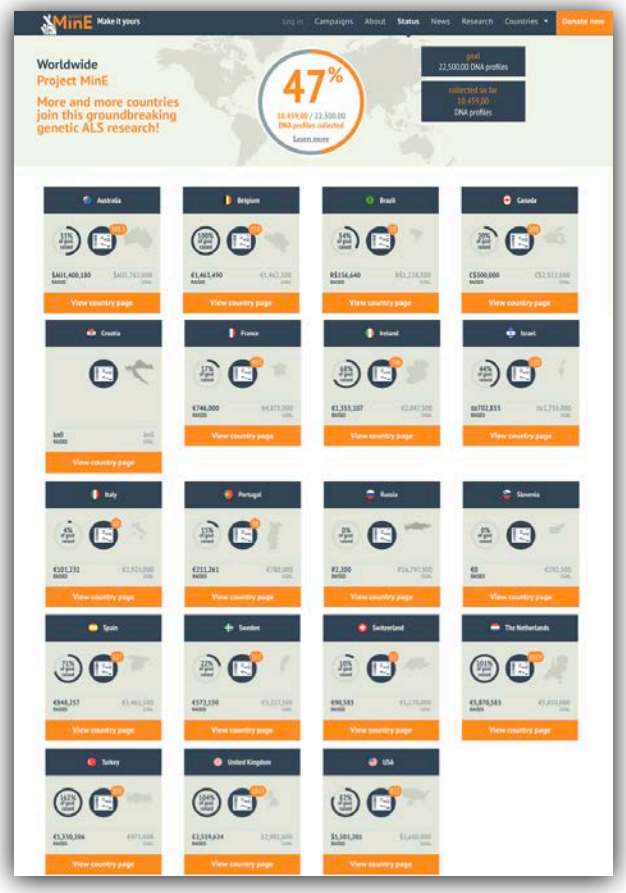

Figure 2: Status page per country December 2018: www.projectmine.com

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 new contributions were secured for Israel (60 profiles), Spain (96 profiles) and Portugal (26 profiles). Spain and the US are expected to be the next one to reach their goal in 2019-2020.

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 2018, 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 almost 9,000 WGS profiles were present as actual profiles; see Table 1 indicated as 'MinE'). These samples were sequenced through sequence provider Illumina (San Diego, USA). An additional 1370 profiles as stated under 'missing' in Table 1 are currently being processed and expected to be available in 2019. Therewith the total number of WGS 'MinE' profiles will increase towards a 10,000.

Project	Total	Cases	Controls	Missing
MinE	8934	5293	2271	1370
Collaborators	4023	3062	304	271
Total	12957	8355	2575	1641

Project	Site	Total	Cases	Controls	Missing
Australia	Australia - Brisbane	172	124	22	26
Australia	Australia - Sydney	598	593		5
MinE	Belgium	644	418	184	42
Brazil	Brazil	24			24
Canada	Canada	234	85		149
China	China	812	626	186	
MinE	France	192	0	0	192
MinE	Ireland	706	466	234	6
MinE	Israel	110	37	0	73
MinE	Italy	70	0	0	70
Malta	Malta	15	13	2	
MinE	Netherlands	3068	1956	1109	3
NYGC	NYGC	2168	1621	94	67
MinE	Portugal	84	36	17	31
MinE	Spain	520	284	135	101
MinE	Sweden	325	0	0	325
MinE	Switzerland	54	46	1	7
MinE	Turkey	803	330	110	363
MinE	UK	1801	1401	400	0
MinE	US	557	319	81	157
Total		**12957**	**8355**	**2575**	**1641**

07 / 12 / 2018 Project MinE

Table 1: Number of samples sequenced in Project Mine (MinE), from collaboration projects and under processing ('missing')- December 2018

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. Herewith the total number of genomes increased already with almost 4000 WGS profiles in 2018. This increase will continue over 2019 where even more profiles are to be expected for cases, but even more for controles (~10,000).

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 before. Researchers analyse their data on the SURFsara supercomputer. December 2018 almost 13,000 DNA profiles were available for analyses, and a next data freeze with another 8000 profiles, cleaned and ready is expected in mid 2019 to be available . The direct connection ensures that the transfer of data is safe and fast. The storage will be expanded as more data is being transferred in 2019. 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 subsidairy from PPS, Health Holland) and where data calculations will be prepared for research projects by experienced bioinformatics.

Project organization: Two Project MinE meetings were held this year; one May 20th in Oxford, UK during the ENCALS meeting and one December 6th in Glasgow, UK during the ALS/MND Conference. Topics that are addressed in project meetings are: scientific progress, progress on sample collection and analyses, progress on fund raising, project organization matters. The Consortium Agreement is final and signed by all partners. The next Project MinE meeting will be organized at the ENCALS meeting 2019 in Tours, France. 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 third Science meeting was held on May 17th at Schiphol, the Netherlands to update and share progress within the six Working groups. In addition to the Science meeting a 'hack-a-thon' was organized for those partners who actually perform complex calculations within the project for data mining. The goal was to train the project members in how to run complex calculations on the SURFsara platform and to align and uniform data processing / cleaning work flows.
The upcoming Science meetings is scheduled for May 23 and 24, at Schiphol, The Netherlands.

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). In 2018 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.
Mid 2018 the second version of the databrowser with more browsing options was released.
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.

Regarding the website we are planning to change the visuals for the overall counter, as to differentiate between the numbers of profiles on ALS cases and controls, and achieved through fundraising or through collaborations (hosting data from other WGS projects).