PROJECT MINE PROGRESS REPORT 2016

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. Nevertheless, if desired, this is possible. For the next 12 months, this is already funded 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. This year we launched a 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 2016 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).

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 is organized now in a similar way to Europe. Principal Investigators (PIs) there have adopted the core clinical data definitions from Europe and follow the structure of Project MinE. There are continuing talks with Swedish and French PIs to further contribute to the project. Spring 2015 Brazil joined as new partner. In 2016 Canada joined in, and herewith, end of December 2016 a total of 17 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, Slovenia, India, and Russia.

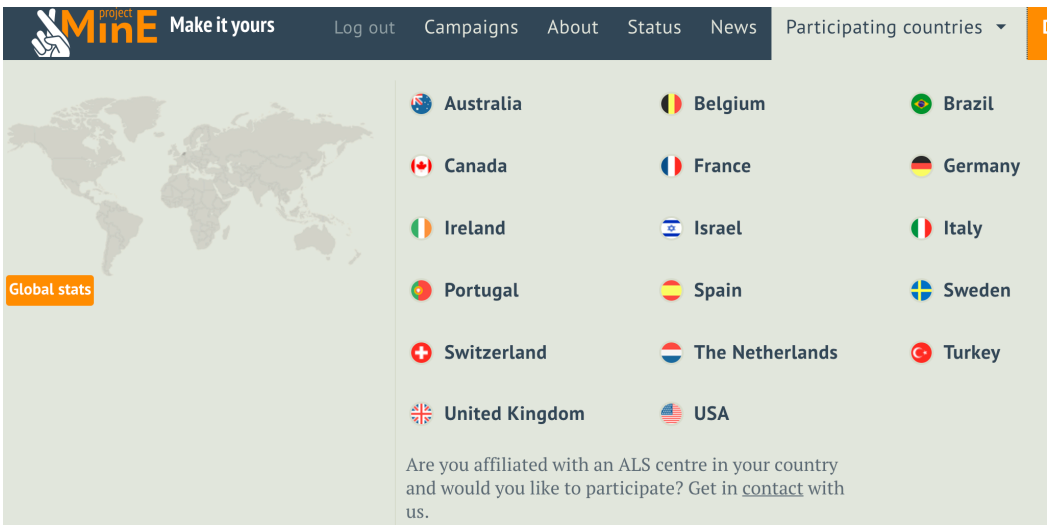

Figure 1: Countries joined in Project MinE: www.projectmine.com

Figure 2: Status page per country Dec 2016: 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, Fundela in Spain, and governments (the Netherlands, Belgium, Sweden (pending)). 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. The UK is expected to be the next one to reach their goal in 2017-2018.

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. During the ACS in September 2016 it was officially announced for the UK to have a 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', '232Km in 2016', 'Mis retos por la ELA') and other very successful (sports) events such as Project MinE Gala Event of Prize4Life in Feb, Tour de ALS, Ledro4Life, RainBowRun and the Good Run. Six sport events are already scheduled for their next edition in 2017 (ACS, London City Swim, Tour de ALS, Ledro4Life, RainBowRun, The Good Run). 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 almost all countries.

Number of samples sequenced: By the end of 2016, Project MinE will have assembled an impressive number of WGS profiles with appreciable power in a relatively short time period (almost 7,000 WGS profiles; see Table 1 below). Samples were sequenced through sequence provider Illumina (San Diego, USA). More batches will be sequenced through 2017 as soon as funds are available; for a batch of 1500 samples funds are secured or pending in UK, Ireland, Spain, Turkey, Israel, Sweden, Switzerland, and in USA, France and Italy, respectively). During the first quarter of 2016, when we combine Project MinE data with the ongoing Biogen/Colombia University Medical Center project, we will already be able to analyse 8,000 ALS DNA profiles and 8,226 controls (16,226 samples in total).

Total goal per country				sequenced 2013-2016			%
	Cases	Controls	Total	Cases	Controls	Total	of goal
Australia	1.000	500	1.500			35	2%
Canada	*666*	*333*	*1.000*				
Belgium	500	250	750	384	191	575	77%
Brazil	100	50	150	10	5	15	10%
France	2.000	500	2.500				
Germany	2.000	1.000	3.000				
Ireland	750	375	1.125	272	136	408	36%
Israel	150	75	225				
Italy	1.000	500	1.500				
Nederland 1+2+3	2.000	1.000	3.000	2.000	1.000	3.000	100%
Portugal	200	200	400				
Spain	500	250	750	244	116	360	48%
Sweden	1.000	650	1.650				
Switzerland	280	300	580				
Turkey	500	500	1.000	150	74	224	22%
UK 1+2+3+4	1.334	666	2.000	1.102	365	1.467	73%
USA (1)	700	700	1.400	417	158	575	41%
USA (2)	1.300	0	1.300	1.300		1.300	
total	**15.980**	**7.849**	**23.830**	**5.879**	**2.045**	**7.959**	

Table 1: Number of samples sequenced Project Mine – Dec 2016.

In 2017 the Project MinE consortium will also host data from other WGS projects worldwide (e.g. Answer ALS, CREATE and data at the Broadinstitute (US)). Already sequenced data from ALS patients and controls will be imported to the Project at SURFsara to become available for the Project MinE researchers for analyses. Herewith the total number of genomes will increase substantially (1500-2000 genomes are to be expected).

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 2016 almost 8,000 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 in 2017. 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 where data calculations will be prepared for research projects by experienced bioinformatics.

Project organization: Two Project MinE meetings were held this year; one May 19th st in Milan, Italy during the ENCALS meeting and one December 7th in Dublin, Ireland 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 almost all partners. The next Project MinE meeting will be organized at the ENCALS meeting 2017 in Ljubljana, Slovenia. Before this, we will have a first Project MinE Science meeting in March 27th at Schiphol, The Netherlands. This is to set-up and coordinate the research efforts of those partners who now have substantial data sets within Project MinE. The aim is to form and structure working groups around five major topics for ALS genetic research. Each working group will define their aims, tasks and deliverables and will report to the General Assembly of the consortium every six months. The working groups will be announced at the Project MinE website (research page).

Besides the meetings communications between (the project coordinator and) partners are going through platform Basecamp. The intention for 2017 is to expand the current governance structure with a supervisory board, and 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 organisations through advise or promotion through the Project MinE website and social media.