Standard Operation Procedure: Selecting controls in Project MinE

Background:
The careful and adequate selection of controls should be considered as important as including correctly diagnosed patients in genetic association studies. The use of inappropriate controls can lead to spurious result, either false negative or false positive. Therefore, the selection of the right controls is essential to the success of this project and should be done with great care.

Project MinE is a large international collaboration in which many centers from all over the world are participating. Each of the contributing centers has collected DNA samples from cases and controls over time and it is therefore possible to perform a study of this magnitude. Indeed, the size of the study is the key to success.
Ideally, all cases and controls in the study would have been collected according to predefined criteria. ALS cases are diagnosed according to the El Escorial criteria and therefore cases have ascertained in a standardized manner. For controls this is however not the case. Centers have collected controls using different strategies. Considering we need very large numbers of controls and we need controls from each population in the study, we will therefore accept that controls are derived from different sources as they long as certain criteria are fulfilled.

Criteria:
- Matched for age (+/- 5 years from age of patient).
- Matched for gender.
- Matched for nationality or ethnicity (as relevant).
- No medical history of neuromuscular or neurodegenerative disease.
- No known family history for neuromuscular or neurodegenerative disease.

Accepted sources of controls:
- Population based (preferred)
- Spouse controls
- Blood bank controls
- Controls from other studies

DNA requirements:
20 ng/µl and a minimum of 100 µl for standard NGS sequencing.

Rationale for criteria:
Age matching:
The rationale for age matching is derived from population genetics. Nowadays travel is far simpler and easier than in the past. Therefore, younger individuals are more likely to move around for employment or other reasons compared to elderly individuals, who are more likely to have spent their entire life in one place. So elderly individuals are more likely to have alleles specific to a certain geographic region, whereas younger individuals are likely to be more mixed. In other words, age matching prevents over- or underrepresentation of alleles specific to certain geographic regions.

Age matching:
The use of so-called hyper-normal controls has also been suggested. Hyper-normal controls are individuals who are free of disease at high age (for instance >90 years of age). The advantage of using these samples is that one can be relatively certain that they will not

develop ALS. Therefore ALS risk variants will be underrepresented in this hyper-normal control group, which would potentially facilitate the discovery of ALS risk factors. There are however also disadvantages to this approach. Firstly, individuals who live to be healthy at high age are likely to have protective factors against a numbers of diseases, such as cardiovascular disease and cancer. In other words, when using these controls in an association study there is chance that the hits will be associated with longevity rather than ALS. Secondly and a more practical concern is that collecting large numbers of hyper-normal controls (>7,500) does not seem realistic. Having a subset of hyper-normal controls potentially introduces a form of bias. Therefore, hyper-normal controls will be not allowed in this study.

Matched for nationality or ethnicity (as relevant):
Ideally we would like to have a homogeneous study population composed out of individuals from similar descent. Considering the majority of all samples will be from Europe and the US, we are looking for Caucasian individuals from European descent. In previous studies we used self reported ethnicity / descent and required that all 4 grandparents were born in The Netherlands. In a small country with relatively little ethnic diversity matching for nationality is appropriate. In other populations with greater ethnic diversity matching for nationality will not suffice. For instance, as pointed out at the Project MinE kick-off meeting in Amsterdam, it does not make sense to have Sephardic controls for Ashkenazi cases for the Israeli samples. Additionally, including only a small number of individuals from a specific ethnic minority is not useful. This will result in a number of rare alleles likely to be unique to that specific ethnic group and without a substantial group of controls for comparison, we will have no way of determining whether these alleles are disease associated or population specific.

No medical history of neuromuscular or neurodegenerative disease:
Many of the genes that have been implicated ALS have also been demonstrated to be associated with other neurodegenerative / neuromuscular disorders. For instance, *C9orf72* has been implicated in frontotemporal dementia, Alzheimer's disease, schizophrenia, *TARDBP* in FTD and PD, *VCP*, *hnRNPA1* and *hnRNPA2B1* in Paget's disease of the bone, FTD and IBM, *FUS* in essential tremor, etc. Therefore controls with medical history of any neurodegenerative or neuromuscular disease could potential introduce ALS risk alleles into the control population.

What about disease controls?
Since we now know that ALS genes can have marked pleiotropy, ranging from psychosis (*C9orf72*), to Parkinson's disease (*ANG*), and even Huntington's (*C9orf72*), and also knowing that ALS patients in general have a beneficial cardiovascular history, the choice for any disease controls is undesirable. There is a high risk of biased results, either positive or negative, without knowing the potential pleiotropic nature of the genes of interest.

No known family history for neuromuscular or neurodegenerative disease:
Many familial ALS (and neurodegenerative) genes are incompletely penetrant. Therefore individuals that have relatives with neuromuscular or neurodegenerative disease could potentially also be carriers of risk alleles for these disorders. Therefore, these individuals are not the ideal controls. We realize that extensive, in-depth family histories may not be available for all controls. However, when it is known that a certain subject has a first or second degree with a neuromuscular disease, they should not be included in the study.

Accepted sources of controls:
Controls from different sources will be accepted in this study as long they fulfill the criteria above. The following types of controls will be accepted: population based controls, spouse controls, blood bank controls and controls from other studies. Population based controls are preferred above any other source. Spouse controls and blood bank controls will also be accepted. Controls gathered for other studies may also be used, especially if they have already been genotyped. Participating members to Project MinE are strongly encouraged to actively look and engage scientific partners that have access to or are involved in large scale whole genome sequencing projects.

PAN questionnaire:
It is highly desirable that all controls also fill in a standardized questionnaire on life style and environmental exposures in order to facilitate gene –environmental studies. Participating centers in the EuroMOTOR study are already using this questionnaire. It has been translated into English and Italian. We are working on translations in other languages: German, French, Spanish and Portuguese. Translations of the questionnaire will be made available through the ENCALS website. Data can be stored in an online database (Progeny). Accounts and access can be set through the UMC Utrecht.