ALS

Amyotrophic lateral sclerosis (ALS) is an adult-onset, lethal neurodegenerative disease. In the United States the disease is also known as Lou Gehrig's disease named after a famous baseball player, who died from the disease during his career. Initially patients notice gradual onset of muscle weakness, but over time the weakness spreads throughout the body and leads to severe disability. Although onset is gradual, the disease course is relentless and about 50% of patients die within 3 years, usually due to respiratory failure. The course of the disease is highly variable and differs from patient to patient, even amongst familial ALS patients within one family. For instance, survival may range from a few months to over ten years.

Up to one third of patients with ALS survive for either more than 48 months, or die before 18 months after symptom onset. Unfortunately, our understanding about the molecular basis for ALS is limited, hampering development of effective treatments.

The disease is characterized by degeneration of motor neurons in the spinal cord, brainstem, and motor cortex. Patients exhibit upper motor neuron signs (hyperreflexia, spasticity and hypertonia) as well as lower motor neuron signs (atrophy, weakness and fasciculations). The diagnosis is made per exclusionem according to the El Escorial criteria. The incidence of ALS is approximately 2.0 per 100,000 population per year in most western countries. There are a few isolated areas in the world, such as Guam (USA), the Kii peninsula (Japan) and Irian Jaya (Indonesia), where ALS is more prevalent. Males appear to be affected more frequently than females with a ratio of 1.6 : 1.0. The average age of onset in population-based studies is in the sixties, but ranges from 18 to well over 80 years of age.

In 5-10% of patients there is a positive family history for ALS, although it is not always possible to establish the mode of inheritance in each pedigree and not all familial cases may suffer from a genuine Mendelian or monogenic disorder. About 10-20% of familial ALS patients carry autosomal dominant mutations in the SOD1 gene, although in the Netherlands this mutation is extremely rare, illustrating geographical variation in causal ALS mutations. Mutations in other genes (such as VAPB, ANG, FUS, TARDBP and FUS) have been reported, but are found at a much lower frequency and with variable penetrance, suggesting the involvement of other genes.

In contrast, in the vast majority of ALS cases there is no family history, and the disease is considered to be "sporadic", indicating a more complex picture than a single gene mutation causing ALS. Small family size and reduced penetrance all give rise to "sporadic" ALS, while in fact a clear genetic contribution is present as the recent C9orf72 discovery has revealed and a recent meta-analysis of three twin studies (ALS heritability of 0.61 (95% CI 0.38–0.78)).

ALS can be considered a complex, multifactorial disease. Identification of genetic risk factors can provide critical information about the fundamental mechanisms that lead to ALS and point to novel directions for therapy.

Genes and human complex disease

There are at least two main competing models that try to explain the genetic basis of human complex disease: rare alleles with moderate to high effect versus many variants with small effects ("infinitesimal model"). There are many arguments pro and contra each model. In short, arguments for rare alleles are that evolutionary theory predicts that disease alleles should be rare, and that many rare variants have been shown to underlie human diseases. Arguments against rare alleles stem from simulations of genome-wide-association study (GWAS) data that are not consistent with rare variant explanations, and that sibling recurrence rates of many diseases are greater than the postulated effect sizes of rare variants. The argument for common alleles is the success of GWASs that have successfully identified thousands of common variants underlying many human diseases, but arguments against are that there is still a lot of "missing heritability" in many diseases and that very few common variants have been functionally validated.

For the past 7 years, genome-wide association studies have identified thousands of common single-nucleotide polymorphisms for a wide range of diseases (including multiple sclerosis, Alzheimer's, Parkinson's, Type 2 diabetes). As a result of these studies, the field is now starting to examine which genes (or chromosomal regions) play an important role in disease mechanisms. Despite this success, GWAS have been limited in their lack of coverage of low-frequency and rare variation. Evidence accumulates that multiple rare variants (copy number variants and mutations) are related to many complex human diseases as well, including schizophrenia and autism, but they have been notoriously difficult to study.

In addition, GWAS does not provide (sufficient) information about more complex types of variation like repeat expansions or structural variants. Next-generation sequencing technologies now make it possible to interrogate the genome comprehensively and provide an unprecedented opportunity to learn about the genetic basis of complex diseases like ALS, where the contribution of rare variants and more complex types of variants (like repeat expansions) may be more relevant than that of common variation. Therefore, we have proposed a combination of GWAS with systematic (whole genome) sequencing of a large cohort of ALS patients and appropriate controls.

What is currently known about the genetic basis of ALS?

Mutations in 12 genes have been found to cause familial ALS, the most common being the repeat expansion in C9orf72, then SOD1, followed by FUS and TARDBP. All genes mutated in familial ALS have also been found mutated in patients diagnosed with sporadic ALS and, besides a lower mean age of onset, no clinical difference exists between the two groups.

Besides mutations in known familial ALS genes, few robustly replicated genetic associations have been found yet in sporadic ALS.

Both genetic models as described before appear applicable to ALS. The C9orf72 discovery clearly shows that a rare allele with moderate to high penetrance, can lead to a sporadic form of ALS. Interestingly, the latest and greatest GWAS in ALS, showed a genome-wide significant association (5% difference in allele frequency) in a locus that harboured this rare allele. The lesson learned was that not a common variant in that locus contributed to the cause of ALS as in the infinitesimal model, but that in stead, the identified variant was a proxy for a rare, but moderate penetrant allele (C9orf72 repeat expansion) that was present in ~5% of sporadic ALS patients.

Three emerging phenomena have arisen based on recent genetic studies in ALS:

1. Genome significant signals in ALS GWASs are not necessarily driven by common variants that are over-represented in ALS as a genetic "risk-factor", but instead are a proxy for a rare, moderately penetrant genetic variant (C9orf72).
2. Intronic, non-coding genetic variation is causal to ALS for a fair amount of all ALS patients (C9orf72).

3. Genetic variation, either coding or non-coding, including tandem repeats and rare mutations, in genes that were known to cause other neurodegenerative diseases than ALS, e.g. SCA (ATXN2), HSP (NIPA1), Parkinson's disease (ANG), are shared with ALS (genetic pleiotropy).

Points two and three are crucial when prioritizing the search for the causal variants in future GWAS discoveries.

What is the optimal strategy to move forward in ALS genetics?

Whole exome/genome sequencing in non-familial ALS

This is the ultimate goal in analysing the ALS genome, however, the current costs for one genome at a reasonable high coverage are still high to reach the required power for this analysis. This approach would pose several computational, analytical and data storage challenges, but these can be overcome.

A combination of GWAS and sequencing

The notion that rare variation, not necessarily tagged by common SNP markers, will be relevant to ALS, and that whole genome sequencing of thousands of samples will be prohibitively expensive, motivates our proposal to use comprehensive sequencing combined with massively expanding GWAS in ALS to interrogate the genome for rare variants, either as a direct cause for ALS or as a proxy for a haplotype harbouring a moderately penetrant genetic variant like C9orf72. These forms of genetic variation have never been systematically studied in ALS. Since non-coding and intronic variation appears relevant, we prefer whole genome sequencing instead of whole exome sequencing.

It is our goal to ultimately whole genome sequence all included ALS samples and a selection of controls.

Progress

We have selected 1000 samples to be whole genome sequenced by Illumina. These samples were prepared in our laboratory and checked for appropriate concentration and degradation to ensure successful genotyping. In 2013, we sent out these 1000 samples and we thusfar have received raw data from Illumina of 942 of these samples. The turn-around time of sequencing has been excellent by Illumina.

The primary objective of these genomes is to use these for imputation of genome-wide association data, collected separately. We do not expect to have sufficient power with these 1000 genomes alone to perform independent analyses on these data. We will need more genome/exome data for that purpose.

In order to be able to use these data for imputation, we had to deal with several challenges: We have to deal with 100TB of data that needs to be securely stored, but more importantly, we need access to readily available High-Performance-Computing (HPC) facilities to be able to work on these data.

The first step is to recalibrate all raw data, since we know that standard quality scores by Illumina are inaccurate. We have delevoped an in-house pipeline to parallelise storage and recalibration of these data. To work on the data further, we needed further investments in hardware for storage and computing, as announced in the original grant proposal of project MinE.

The next step will be to derive haplotypes from recalibrated data and to use these for massive imputation of GWAS data, also on the HPC facility.
To conclude, we have received almost all whole genome sequenced data of the DNA samples selected and sent out and we have succesfully managed to get our pipeline up and running for storage and calculations.