Genotypic analysis of gene expression in the dissection of the aetiology of complex neurological diseases

Daniah Trabzuni Mphil Msc, Department of Molecular Neuroscience, Institute of Neurology, Queen Square, London, WC1N 3BG, England.

Dr.Mina Ryten MBBS PhD MRCP, Department of Molecular Neuroscience, Institute of Neurology, Queen Square, London, WC1N 3BG, England.

Dr Robert Walker PHD, Brain Bank Manager (MRC Sudden Death Project), University of Edinburgh, Neuropathology Unit, Wilkie Building, Teviot Place, Edinburgh, EH8 9AG

Colin Smith MD FRCPath, Senior Lecturer in Pathology, Neuropathology, University of Edinburgh, Wilkie Building, Teviot Place, Edinburgh , EH8 9AG.

Professor John Hardy PhD MD (Hon) FMedSci, Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, Queen Square, London, WC1N 3BG, England

Abstract

The overall aim of this project is to translate newly discovered genetic risk traits for complex neurological conditions into a deeper understanding of pathogenesis. Until recently there seemed little hope of developing a genetic understanding of common diseases of the central nervous system (CNS). However, whole genome association studies of human disease are revolutionising our understanding of the aetiology of complex diseases, such as Parkinson's and Alzheimer's disease. These studies have demonstrated what has long been suspected, that "normal" variability contributes to the risk of common neurological diseases. While some of the risk loci identified have been assigned to coding changes in genes, the majority have not, and many have not even mapped to recognisable genes. Thus, knowing genetic risk variants for common diseases has not provided an automatic understanding of pathogenesis or obvious therapies. In order to address this problem we will study the heritability of gene expression within the human brain. The basis of our approach is the hypothesis that heritable differences in transcriptional regulation, which are present and measurable in control populations, are important drivers of pathology in the human CNS. If common heritable differences in transcriptional regulation can drive pathology in the human CNS, then we would expect to find strong associations between the risk SNPs identified in GWAs for neurological diseases and specific mRNA expression phenotypes of functional significance in control human brain. We intend to use post-mortem control human brain tissue to collect samples from well-defined brain regions known to be particularly affected in the most common neurological (e.g. the substantia nigra). Since risk-associated SNPs will be present in the control as well as the case population, using control brain tissue we can study downstream affects on gene expression without the complications of neuronal death, glial response and symptomatic treatments. Using microarray technology, we will produce large quantities of high quality, genome-wide paired SNP and exon-specific expression data. Data analysis will be

focused on identifying downstream gene expression changes associated with individual SNPs known to increase the risk of developing a neurodegenerative disease. Thus, we hope to bridge the gap between genetic risk and pathophysiology. In this way, we may be able to provide new therapeutic strategies for the early and effective treatment of diseases affecting CNS.

Keywords: Neurological diseases, Genome wide association studies (GWAS), Expression quantitative trait loci (eQTLs), Splicing quantitative trait loci (sQTL), Single neucleotide polymorphism (SNPs).

Background

Over the past decade substantial progress has been made in the identification of the genetic causes of many monogenic human neurodegenerative diseases, such as Huntington disease. However, standard genetic methods have been less successful in defining and characterizing gene(s) in polygenic diseases such as obesity, type 2 diabetes, cardiovascular disease, Parkinson's and Alzheimer's disease (Risch and Merikangas 1996). Genome wide association studies (GWAS) have succeeded where previous approaches have failed. In the last three years hundreds of single nucleotide polymorphisms (SNPs) or copy number variants (CNVs) have been identified, which are significantly associated with common human diseases, including complex neurological diseases (Hardy and Singleton 2000). For example, CNV is associated with susceptibility to Parkinson's and Alzheimer's disease (Lupski 2007). In addition, one of the most important observations from the GWAS is that the majority of the risk SNPs identified in polygenic disorders are located in the non-coding regions of the genome. Thus, for some diseases we may be able to identify risk SNPs, but still have no understanding of the mechanism by which the risk loci increase the risk of disease.

This has led some researchers to suggest that genetic variation amongst individuals may result in changes in transcriptional regulation and that this may be an important mechanism for disease.Expression of mRNA, both quantity and quality, can beconsidered a measurable and heritable quantitative trait. SNPs (and the genetic loci in linkage disequilibrium) have been shown to act as expression quantitative trait loci (eQTLs) in many tissues, including human brain (Myers, Gibbs et al. 2007; Heinzen, Ge et al. 2008). More recently, it has been shown that processing of precursor mRNA within human brain can be influenced by specific SNPs and these SNPs have been termed splicing quantitative trait loci (sQTLs) (Heinzen, Ge et al. 2008). Most importantly, inherited risk loci for neurological diseases can be associated with gene expression changes in control brain tissue and these changes are consistent with known pathophysiology. For example, inheritance of the MAPT H1c haplotype results in increased expression of the 4 repeat tau isoform in control human brain tissue and this isoform has been implicated in progressive supranuclear palsy, a sporadic tauopathy (Pittman, Myers et al. 2005; Myers, Pittman et al. 2007). As a consequence, the exploration of inherited quantitative variations in human tissues has become one of the major priorities for medical genetics.

However, the mapping of genetic factors that underlie quantitative traits requires a large number of tissue samples and obtaining these samples can be challenging (Stranger, Forrest et al. 2005). A number of studies have studied the genetics of gene expression and their association to clinical traits, but due to issues around tissue access most of these studies have used blood or lymphoblastoid cell lines as a sample resource

(Emilsson, Thorleifsson et al. 2008). This study will use post-mortem control human brain samples to focus specifically on neurological disorders.

In summary, it is important to expand our understanding of how risk SNPs in non-coding regions have consequences on gene expression and splicing and how this leads to the progression of human disease.

Aims:

The first aim will examine and investigate the mechanism of the bi-directional relationship between the heritable expression quantitative traits (eQTL), splicing quantitative traits (sQTL) and genome SNPs loci, which are located in the non-coding regions, in 10 different human control brain regions. We hypothesise **that these variations have a fundamental role and significant association to increase the risk of the neurological disorders**. The results of this study will hopefully present a comprehensive understanding of the progression and the etiology of the human complex diseases in more depth; thus providing a new wide applicable treatment era in the future.

The second aim hopes to establish a vital and reliable dataset of eQTL and sQTL that are significantly paired with the genomic SNPs loci data in human control brains. This dataset will serve as one of the major and required platforms in the genome-wide association area for human control brains, and will be a very rich resource for further research into biological and genetic disorders such as, Parkinson's, Alzheimer diseases, obesity, type 2 diabetes and cardiovascular diseases, which are considered as sporadic cases and those that do not follow a particular Mendelian inherence pattern.

Furthermore, having the advantages of applying and analyzing high throughput technologies in the field of genetics such as the; GeneChip Human Exon and high density BeadChip Arrays, will improve the researcher's skills in collecting, analyzing, presenting, and interpreting data using advanced and modified statistical and bioinformatics models. These skills and experiences will be transferable to be applied on different populations that have been estimated to have a high risk rate to develop genetic disorders and especially polygenic disorders. For example, the population of Saudi Arabia has an isolated genetic pool and consanguinity rate of 60% due to first cousin marriages within the same tribe. This will lead not only to a higher susceptibility to develop monogenic diseases that follow mendelian patterns of inheritance, but also variety of polygenic diseases of which their genetic architecture in not fully understood such as, type 2 diabetes, deafness, cardiovascular disease, obesity, Parkinson's disease and many more.

Therefore this population would be ideal for this type of study considering different types of tissues for example, adipose tissues, kidney biopsies and blood samples, as these types of tissues are accessible and easy to collect compared with brain tissues.

Material and Methods

Sample collection:

150-200 frozen human control brain tissues were collected prospectively from the MRC Sudden Death Brain and Tissue Bank in Edinburgh, at a rate of 4 brains per week. Samples that were considered in this study were within the criteria of sudden death, post-mortem within 3 days and no significant abnormality in the brain tissue. It is well known that the brain is presenting a high diversity of cell types and hence, ten different regions will be dissected and collected from each control brain. These regions are; Cerebellum, thalamus, basil ganglia, substantial nigra, hippocampus, medulla, frontal cortex, parietal cortex, white matter and spinal cord.

RNA and DNA isolation and quality check:

Genomic DNA will be isolated from the cortex of each brain tissue by using QIAamp DNA Mini Kit from Qiagen. Nanodrop will be used to measure the concentration of the isolated DNA.

Total RNA, including microRNA (miRNA), was isolated from each of the brain region using TRIZOL reagent according to manufacturer's protocol (Chomczynski and Sacchi 1987). The quality of each RNA sample was checked by: a) measuring the RNA concentration and 260/280 ratio using Nanodrop and b) examine each RNA sample on Agilent 2100 Electrophoresis Bioanalyzer chip to evaluate the RNA integrity and degradation level in each sample. RNA integrity number (RIN) is the main indication for the quality of the RNA sample. Samples with a high yield of RNA, 260/280 ratio >1.8 and RIN number >5 will be considered in this study.

RNA and DNA microarray analysis:

DNA genotyping will be performed using an Illumine Bead platform. Each of the DNA samples will be applied and processed using Illumina Human 1M BeadChip SNP array, while RNA expression analysis will be performed using Affymetrix Chip station with Affymetrix GeneChip Human Exon array for each RNA sample. Two sets of raw data will be generated, genomic and transcriptomic data.

Data analysis of the raw data:

The raw data will pass through different stages and quality control steps to be available to interpret. The Genomic data set will be processed using BeadStudio II software to generate SNPs calls. Then further analysis and quality control steps will be applied using the PLINK program to check samples for different factors such as, age, sex, ethnic background, Minor allele frequency value (MAF) and the p-value. By the end of the analysis the genomic variants, SNPs and copy number variant (CNV), will be detected and identified.

The raw transcriptomic data is more challenging to analyze. The quantity of the generated data will be extremely large. Some algorithms such as, PLIER (Probe Logarithmic Intensity), which is provided in Affymetrix power tool (APT) software, can be used to evaluate the arrays and do all the required quality control

steps. The the data will then be passed through a number of summarization and normalization steps as described (Shah and Pallas 2009). Later, the identification of the expression variants; splicing variants and expression variants will take a place.

After identification of both sets of variants, the association test will be applied using a linear regression model as one of the association models that can be used. This will be followed by a multiple test correction step using well known methods such as, Bonferroni correction, false discovery rate and permutation to generate a null distribution of p-value. The assigned significance threshold of p value in this study is equal to 0.05. This statistical association is an indication for a bi-directional relationship of how one or multi SNPs loci can affect one or multi expression variants or vice versa.

At this stage this particular dataset is ready to be used in various biological applications. Other available and resourceful databases such as, A Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/26525384) can be used in parallel with this dataset to enrich and maximize the use of the information available. For example, some risk SNPs have been published as being associated with Alzheimer's disease and Schizophrenia, but their overall effect of regulating expression or splicing variants to cause the disease phenotype is not conclusive. Therefore, comparing these associated risk SNPs to the generated eQTL or sQTL will help to preferences some genes than others in the genes candidate list. In addition, if any significant splicing variant has been identified and associated with certain disease its function can be investigated further by RNA sequence.

Moreover, there are many different amenable ways to use this dataset for biological and genetic studies to present a clearer understanding of the mediated biological steps between risk SNP identification and disease pathophysiology. Eventually it is hope that this will lead to a new wide applicable treatment era in the future.

Time frame

The estimated time frame for the project is two and a half years. The first year will be mainly occupied by sample collection, DNA and RNA isolation and quality checks. Following this, will begin the generation of the data from the microarray analysis and the processing of quality control steps. The most challenging and time consuming part of this study will be the final analysis for the generated data.

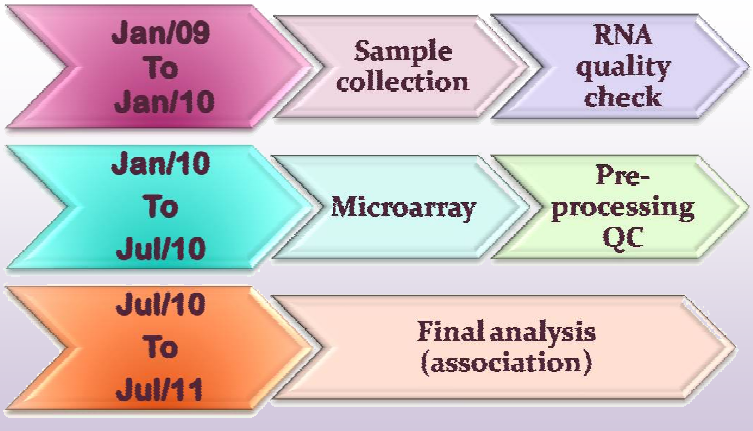

Figure :
Time frame for the proposed project

References

Chomczynski, P. and N. Sacchi (1987). "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction." Anal Biochem **162**(1): 156-9.

Emilsson, V., G. Thorleifsson, et al. (2008). "Genetics of gene expression and its effect on disease." Nature **452**(7186): 423-8.

Hardy, J. and A. Singleton (2000). "The future of genetic analysis of neurological disorders." Neurobiol Dis **7**(2): 65-9.

Heinzen, E. L., D. Ge, et al. (2008). "Tissue-specific genetic control of splicing: implications for the study of complex traits." PLoS Biol **6**(12): e1.

Lupski, J. R. (2007). "Structural variation in the human genome." N Engl J Med **356**(11): 1169-71.

Myers, A. J., J. R. Gibbs, et al. (2007). "A survey of genetic human cortical gene expression." Nat Genet **39**(12): 1494-9.

Myers, A. J., A. M. Pittman, et al. (2007). "The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts." Neurobiol Dis **25**(3): 561-70.

Pittman, A. M., A. J. Myers, et al. (2005). "Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration." J Med Genet **42**(11): 837-46.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Shah, S. H. and J. A. Pallas (2009). "Identifying differential exon splicing using linear models and correlation coefficients." BMC Bioinformatics **10**: 26.

Stranger, B. E., M. S. Forrest, et al. (2005). "Genome-wide associations of gene expression variation in humans." PLoS Genet **1**(6): e78.