Identification of closely linked SNP markers using information theory

D.V. Raje*#, Venkatesh, K*, S. Bundela* and M.S. Rekha Rao*
*Ocimum Biosolutions Ltd., Road # 1, Banjara Hills, Hyderabad – 500 034 (INDIA).
Email: dhananjay.v@ocimumbio.com

1.0 INTRODUCTION

Chronic Fatigue Syndrome (CFS) is a disease that has no definite diagnostic clinical symptoms or laboratory abnormalities and hence identification of such illness is often complex. The definition of CFS as a unique disease is not obvious as it may represent a common response to a collection of other illnesses. Hence, establishing well defined measures to diagnose CFS has been a challenge in medical science. There are attempts to refer different sources of information like clinical assessment, microarray expression, proteomics data and SNPs and to correlate them to improve upon the diagnostics.

In the present study, our focus is on SNP data and the interest is to identify pairs of SNPs that are closely linked and specific to the disease. The grouping of subjects based on intake classification was referred for analysis. The mutual information content was used as a measure to identify distinguishing SNP pairs across the groups. Also, the clinical data was analyzed using logistic regression to obtain statistically significant factors associated with the disease. Finally, results of both the analyses were compared. The computations were performed in *R* programming language.

2.0 DATA ANALYSIS

The clinical data (2006) and the SNP data (2007), publicly available at CAMDA website, were considered for analysis. The screening resulted into 164 subjects that had both clinical and SNP information. The intake classification (IC) from the clinical data was relied upon and used for grouping subjects into three distinct categories viz., Non-fatigued (54), ISF (55) and CFS (55). The responses were quantified for suitability of analysis. The SNP data consisted of 166 SNPs spanning 39 genes representing different genotypes. The three genotypes corresponding to each SNP marker were coded numerically as 0, 1 and 2; where codes 0 (allele 1) and 1 (allele 2) indicates homozygous condition and 2 indicates heterozygous condition (both alleles). This coding strategy was adopted for all the SNPs. The missing / incomplete data points for SNPs were replaced with their respective median values. The interest here was to identify closely linked SNPs, which could be relevant in disease mapping. The data was analyzed for each group using mutual information content to determine SNP pairs that share maximum information in the group. Mutual information measures the mutual dependence between the two variables. Intuitively, it measures the information shared by the two variables *i.e.*, how much knowing one of the variables reduces uncertainty about the other; and is given by

$$MI(X,Y) = -\sum_{x} p(x)\log_2(p(x)) - \sum_{y} p(y)\log_2(p(y)) + \sum_{x,y} p(x,y)\log_2(p(x,y)) \quad \ldots(1)$$

Here X and Y represent two SNPs being compared. The first two terms provides Shannon's entropy, while the third term provides joint probability of occurrence of values of X and Y. There are nine distinct combinations of the three genotypes. The *MI* values could be obtained for all the possible pairs of SNPs resulting into an information matrix. The maximum of *MI* across each row could be

determined and ranked from highest to lowest. The top ranked SNP pairs could be selected, as they share much of the information between them. Such analysis could be carried out for each of the study group. The unique pairs for each group could be identified and referred to as SNP markers. For the present data set, a matrix of *MI* with pairwise information content for all 166 SNPs was obtained. The pairs with *MI* greater than 1.0 were selected for analysis and treated as closely linked pairs. The paired SNPs showed that they have a preferred choice of genotypes, i.e. if we know about the genotype in one SNP, we can predict the genotype in the other. If such occurrence is specific to a group, then the pair could be identified as marker for the group. Some of the pairs satisfying the above criterion are shown in Figure 1. It is evident from the figure that few SNP pairs are unique to the groups, while some are shared by one or more groups. The pairs DRD2(rs7486613)-HTR7(rs11756841), TPH2(rs1247728)-ACE(rs8872233), BDNF(rs11592758)-HTR5A(rs7744285) show close relationships in ISF & CFS patients and not in the non-fatigued group. In other words, there is a preferred choice of genotypes across these pairs in the two diseased groups, which is absent in the non-fatigued group. Simultaneously, the clinical factors that are most relevant to the disease were investigated. The factors from CDC symptoms inventory were considered and treated as independent variables, while IC was treated as dependent variable. The responses on these factors were numerically coded and analyzed using logistic regression to extract statistically significant symptoms across the groups. The factors like depression (OR = 31; $p < 0.05$) and post-exertion fatigue (OR = 34.994; $p < 0.01$) indicated significant influence on CFS. Similarly, depression (OR = 15.34; $p < 0.05$) and post-exertion fatigue (OR = 221.62; $p < 0.01$) showed significant impact on ISF. These symptoms were studied for any associations with the identified SNP marker pairs. DRD2, HTR1A (found in CFS), TPH2 were associated with depression; IL-1B was associated with muscle pain (found in CFS); BDNF was associated with depression and impaired memory; ACE, HTR7, MAOA(found in IFS), HTR2C were associated with sleep, depression and fatigue; NRC31 was associated with sleep and fatigue. In short, more SNP's related to depression were found in the diseased group.

We believe that the use of information theory on a much larger data set would ascertain these SNPs as markers and would yield a definite genotypic profile of the selected markers in the ISF & CFS patients.

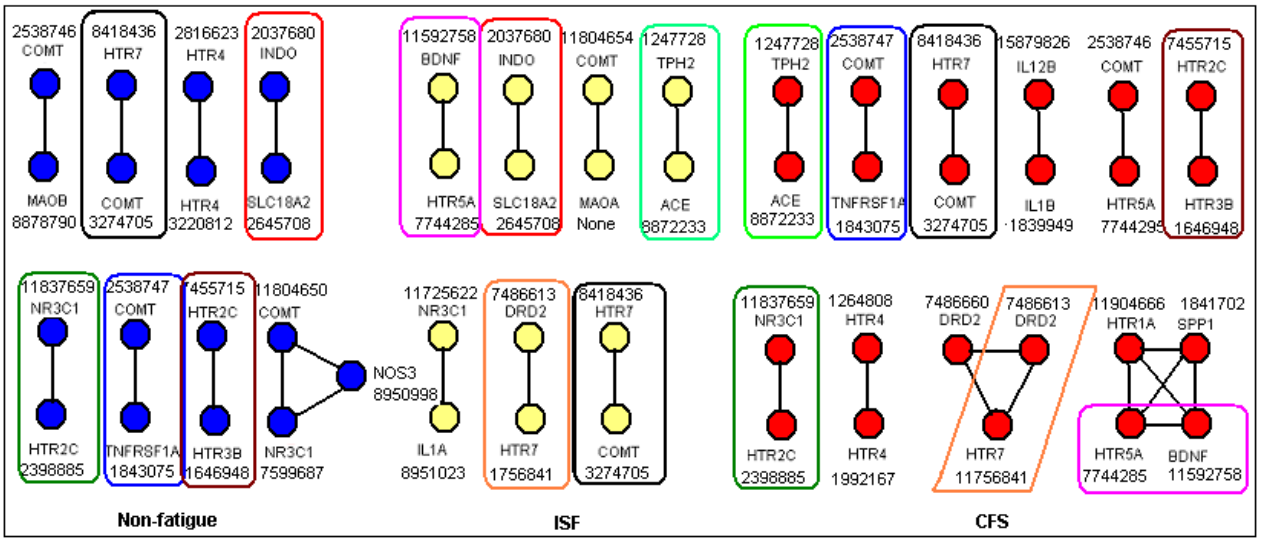

Figure 1: Closely linked SNP markers in three different study groups.