31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Discovering genetic polymorphism associated with gene expression levels across the whole genome.

Athos Antoniades, Ioanna Kalvari, Constantinos Pattichis, *Senior Member, IEEE*, Neil Jones, Paul M. Matthews, Enrico Domenici, Pierandrea Muglia

***Abstract*—Genetic differences have been shown to contribute to gene expression variability. A complete evaluation of the associations between a whole genome scan with 550k Single Nucleotide Polymorphisms (SNPs) and 54k detectable expression levels (probesets) was performed on 176 human peripheral blood samples. The results are presented along with visualizations that reveal cis and trans gene expression regulatory effects. The algorithmic approach followed utilized a distributed computational system. The analysis was performed using a linear regression adjusting for all relevant covariates. Permutation testing on a random subset of the top results provided an indication of the significance levels adjusted for multiple testing and the non independence of SNPs due to linkage disequilibrium. The database of the produced results can be used as a resource to assess the functional impact of genetic polymorphisms to gene expression regulation. This resource is applicable across all disease areas.**

I. Introduction

THE need to identify and understand the functionality of biological mechanisms behind genetic diseases has led to large studies where samples were genotyped using high density whole genome scans [1]. However, even when analyses of these studies were successful in identifying significant associations between a disease phenotype and genetic polymorphism, the functional impact of the polymorphism on the biological mechanisms of gene expression control remained unknown [1],[2]. Gene expression level data quantify the level of expression of genes in sampled cells. This type of data has been used to discover associations between a disease phenotype and the level of expression of a gene in a specific type of cell.

In this paper we present an analysis of a study that has both genetic polymorphism data (550k single nucleotide polymorphisms, SNPs) and gene expression data (54k probesets of mRNA expression data) across the whole genome on 176 subjects to evaluate associations between polymorphisms and gene expression. These results are disease independent and are therefore applicable to any subject domain where knowledge of the functional impact of a polymorphism to gene expression level is required.

Cis acting elements and trans acting factors can be identified from this analysis by considering the distance between SNPs and mRNA probesets in significant associations. Cis-acting members are DNA sequences in the vicinity of the structural portion of a gene expression, while trans-acting factors are typically other genes who's products bind to cis-acting sequences to regulate gene expression [14].

Similar studies as the one presented have been conducted in the past [3],[4],[5]. Those were done using different types of cells for the gene expression data, and genotyping platforms with a significantly smaller number of SNPs. However the numbers of samples in each study as well as the phenotypic traits of the subjects vary. All previous studies were able to discover many statistically significant associations between genetic polymorphisms and gene expression data.

II. Methods And Material

A. Datasets

The dataset used consisted of 176 subjects of whom 119 were diagnosed with recurrent Major Depressive Disorder (MDD) and 57 were healthy individuals. Those individuals were extracted from a larger study of 1022 Caucasians diagnosed with recurrent MDD and 1000 Caucasians age- and gender-matched non-affected controls referred to as Sample I in [1] that were genotyped with the Ilumina HumanHap550 array platform. Diagnosis of recurrent depression (at least two episodes of depression) was based on DSM-IV criteria after administration of the structured clinical interview (SCAN) [6]. The SNPs remaining after quality control for minor allele frequency threshold greater than 0.01, and missing data frequency greater than 0.1 were 511,525.

The 176 Subjects used for gene expression profiles were a subset of age and gender matched case/controls patients

Manuscript received April 23, 2009.

Athos Antoniades is with the Department of Computer Science, University of Cyprus, 75 Kallipoleos Str, P.O. Box 20537, Nicosia, 1678, Cyprus (phone: 357-22892685; fax: 357-22892701; e-mail: athos@cs.ucy.ac.cy).

Ioanna C. Kalvari, the Department of Computer Science, University of Cyprus, 75 Kallipoleos Str, P.O. Box 20537, Nicosia, 1678, Cyprus (e-mail: ioanna.kalvari@cs.ucy.ac.cy).

Neil Jones was with the Discovery Technology Group, Molecular Discovery Research, Glaxo Smith Kline, (e-mail: neiljon12@googlemail.com)

Paul M. Matthews is with the Glaxo Smith Kline Clinical Imaging Center, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 ONN (e-mail: paul.m.matthews@gsk.com).

Enrico Domenici is with Translational Medicine, Mood and Anxiety Disorders, Neurosciences Centre of Excellence in Drug Discovery, GlaxoSmithKline R&D Medicine Research Centre, Via Fleming 4, Verona 37138 Italy (e-mail: enrico.h.domenici@gsk.com).

Pierandrea Muglia is with Discovery Medicine, Neurosciences Centre of Excellence in Drug Discovery, GlaxoSmithKline R&D Medicine Research Centre, Via Fleming 4, Verona 37138 Italy (e-mail: pierandrea.2.muglia@gsk.com).

generated after excluding subjects with comorbidities for major medical conditions, heavy smokers, subjects sampled after a meal or subjects with more than 0.1% missing genotypes. Whole blood mRNA expression data were generated on these subjects by using Affymetrix HU133 plus v.2 GeneChips™ in two batches, each batch analyzing half the samples.

Samples were randomized prior to processing using the NuGEN Ovation RNA Amplification System 2 (NuGEN, San Carlos, USA). The resulting fragmented and labeled material was hybridized to the human U133plus2 chip for sixteen hours. The array was washed and scanned according to Affymetrix protocols. Genechip data quality was assessed using report files generated in GCOS (GeneChip® Operating System) and checked against in house criteria for probe and hybridization quality. In addition, gene array data quality was assessed for homogeneity of quality control metrics by Principal Component Analysis (PCA) using Simca by Umetrics.

The global analysis of gene expression was initially processed by normalizing probeset intensity data using Rosetta Resolver for the visual assessment of key trends by gene expression. Further analysis was undertaken using the normalized probeset intensities by performing a General Linear Model (GLM) based analysis using SAS v9.1 and modeling either for triad or gender and disease. Post hoc testing was then performed to allow comparison of gene expression levels between the control versus depressed patient groups as well as identification of gene expression changes which showed differences between males and females.

B. Methodology

When choosing a statistical methodology for performing the tests the type of data as well as the need to adjust for covariates needs to be taken into account. In this project, linear regression was used as it allowed for quantitative traits and the adjustment of multiple covariates [7]. The covariates used were the disease status of the subjects, the batch number and their age at the time the blood samples used in the mRNA expression data were taken.

The multiple testing problem in this project needed to be addressed as the number of tests was high (10^{13}) [8]. Bonferoni correction was used to address this issue [9]. This approach utilizes a simple heuristic to correct the p-values for the number of tests performed. It tends to work well for independent tests [9]. However, SNPs are not completely independent as some of them may be in Linkage Disequilibrium, a non random association of alleles at two or more loci [10].

The optimal way with regard to the quality of the results to address this problem was through permutation testing, a computationally intensive approach generating the significance levels empirically [8],[11]. One of the many merits of this approach is that it preserves the correlation structure between SNPs, therefore it is not negatively affected by Linkage Disequilibrium between SNPs. However, permutation testing is so computationally intensive that it was prohibitive to perform on all 10^{13} tests.

In order to get a reasonable estimation of the deviation between the Bonferoni adjusted p-values for multiple testing and the permutation derived p-values for the top results, an analysis was conducted on a subset of 18344 of the top results with a p-value smaller than 10^{-5}.

Due to the high number of tests that needed to be performed using linear regression, the computational capacity needed to complete the analysis within a reasonable time frame was very high. Therefore, a computer grid composed of 200 processing units was used and a distributed computing algorithm was implemented to enable the use of all processing cores in parallel to generate the needed results.

In order to reduce the size of the produced data only statistically significant results prior to adjustment for multiple testing using a threshold of $p < 0.01$ were recorded. However, we expected that the number of results passing that threshold would still be too high to be effectively visualized. To address this, a database application was implemented that was tailored to the needs of this project. It is capable of quickly producing subsets of results by querying probesets or SNPs' genomic regions of interest, by specifying the number of the most significant results to include in the subset, or by querying on the ontology or any annotation information of any gene, mRNA probeset or SNP in the dataset [11],[12].

C. Query and Visualization of Results

All results included detailed annotation information for the genetic regions of both the mRNA expression data and the genetic polymorphisms as well as information derived from gene ontology [13]. The graphs presented in the paper were created using the Spotfire DXP software package. Any other software platform that enables query and visualization of large spreadsheet type data could be used; however Spotfire DXP has a high capacity in terms of handling large data sizes such as the ones generated in this project.

III. Experimental Results

The analyses using linear regression of all SNPs against all probesets took 12 days and 12 hours while the permutation analyses on the subset of top results took 5 days and 8 hours on the 200 processing unit system. These times were estimated based on the amount of time allocated to this project by the distributed computing resource's job scheduler. The actual run times were slightly longer as each analysis took 1 extra day due to other programs running for short periods of time on the same resource with higher priority than this project.

Out of the 4,981,737 associations that passed a threshold of p-value < 10^{-5}, 18344 were randomly selected for permutation testing. This was the maximum number of tests that could be performed in the allocated time on the machines. The number of permutations was set to 10^{10}.

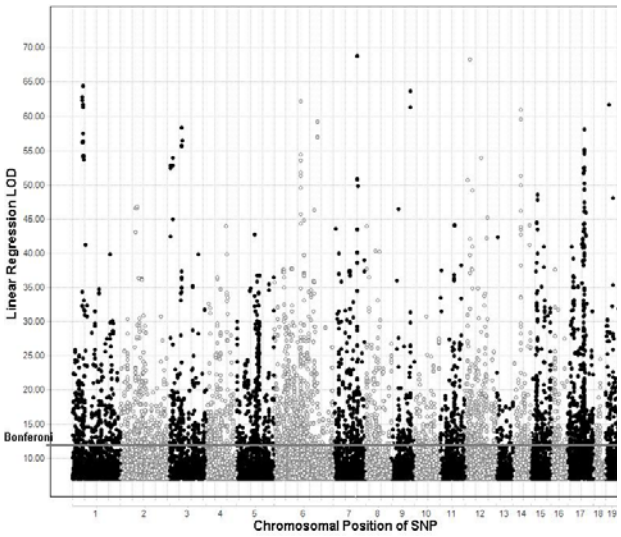

Fig. 1. Distribution of top results in the genome. The top results are highly significant and distributed across the entire genome.

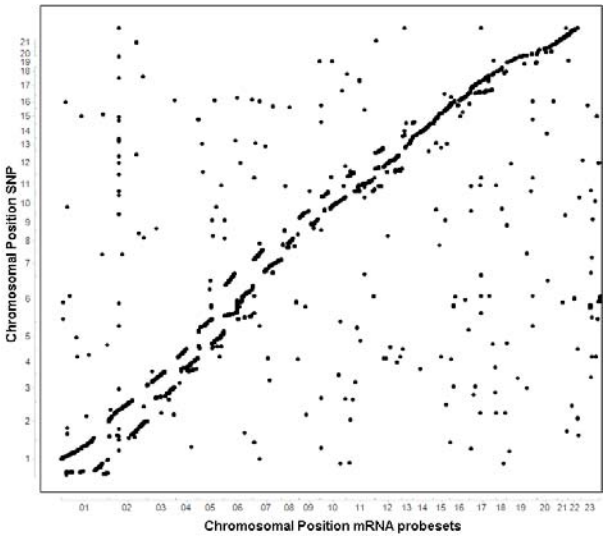

Fig. 2. Chromosomal Position of mRNA vs SNP clearly there are more results along the Y=X diagonal. These are driven by cis acting elements.

However, for 120 of the permuted tests the number of permutations was not adequate to accurately estimate a p-value since the empirically derived signal was stronger than 10^{-9}. The LOD ($-\log_{10}$ p-value) scores from the remaining 18224 empirically derived p-values were on average just -1.73e-1 off from the linear regression derived LOD scores with a standard deviation for the difference of 2.89e-1. All but one of the empirically tested associations were statistically significant.

Associations with an unadjusted p-value less than e-7 were plotted and are shown in fig. 1. The y axis represents the LOD score derived from the linear regression analyses, and the X axis represents the chromosomal location of the SNP in the association with the labels representing the chromosomes. The statistical significance threshold (p-value 0.01) after Bonferoni correction for multiple testing is p-value = 3.62e-13, LOD=12.44, a line identified as Bonferoni is ploted to help identify the statistically significant results after Bonferoni adjustment. As can be seen from fig. 1, the number of associations that pass the threshold of statistical significance 0.01 using Bonferoni correction is very high (6104) and they span the entire genome.

In fig. 2, the results with a p-value less than 10^{-10} were selected. The chromosomal location of the SNP and the center of the probeset are on the Y and X axis respectively. Probesets that map to more than one location were not included. In this graph there seem to be many associations across the Y = X diagonal. These associations are between SNPs and mRNA sequences that are very close together or overlapping. In biological terms, these are called cis-acting elements. The associations between polymorphisms and probesets derived from mRNA sequences on different chromosomes are considered to be trans-acting. These may be transcription factors for the mRNA in the association. However, for associations where both the SNP and the mRNA probeset are on the same chromosome but they are not very close together (over 100k base pairs for example) deriving a cis or trans acting status is questionable, since the level of LD between the SNP and mRNA sequence should be taken into account.

IV. Discussion

The number of results that passed levels of statistical significance after adjusting for multiple testing is high compared to analyses of similar genetic data for disease status traits, especially considering the small sample size of this analysis [1],[2],[14]. This can be attributed to the nature of the hypothesis tested here. The disease status of a subject may be caused by a multitude of factors. Genetic predisposition is one of them, but the number of genetic loci associated with the disease is unknown in the majority of diseases. Moreover, environmental factors typically play an important role and they are very hard to identify, or adjust for [14]. When looking for disease associated genes in similar studies it is common to find a few or no statistically significant results [1],[2],[14]. By focusing on gene mRNA expression levels the uncertainty is reduced, and the search for associated factors becomes more focused. Previous studies conducted using similar approaches confirm the high number of statistically significant associations across the whole genome providing more evidence for the explanation given above [2],[3].

By considering the distance between the genetic location of the polymorphism and the mRNA sequence in a significant association it is possible to estimate if trans acting factors or cis acting elements are responsible for the association. Specifically, if the mRNA and the genetic polymorphism are on different chromosomes then the SNP in the association is associated with a trans acting factor of the probeset mRNA sequence. If they are on the same chromosome and in high linkage disequilibrium they are

considered to be cis acting elements of the mRNA probeset. However, if they are on the same chromosome, not in LD and are far apart (over 100k base pairs) they cannot be conclusively identified as cis elements or trans acting factors[14].

An evaluation of the performance of Bonferoni correction for multiple testing as compared to permutation testing was also attempted on a subset of the top results. The evaluation revealed that the empirically derived p-values were for most of the results very close to the unadjusted values derived with linear regression. This implies that the multiple testing problem did not affect the most significant associations as much as expected based on Bonferoni correction. This is another indication that the top significance levels observed are driven by true effects rather than randomness associated with the multiple testing problem. Note however that the number of permuted associations represented just 0.37% of the total number of associations with p-values over 10^{-5}.

The results database of this project can be used to provide statistically significant evidence of genetic polymorphisms to mRNA expression levels across any disease area. Since the MDD status, gender and age of the subjects were used as covariates, the results are independent of these factors. As an example, consider a statistically significant association found in this experiment between SNP A and probeset of mRNA sequence of gene B that are on different chromosomes. Assume now that in an unrelated research project SNP A was found to be associated with a disease. Having knowledge of the association between SNP A and gene B researchers would have an indication to study whether changes in expression of gene B could be the true causative factor for the disease predisposition, and that the genetic polymorphism associated with SNP A was simply regulating the expression levels of gene B. In another scenario expression levels of gene B could be found to be associated with a disease. In this case, the product of the gene SNP A is associated with could be used to help regulate the expression of gene B in diseased subjects.

V. Conclusion

A database has been created with statistically significant associations between genetic polymorphisms and mRNA expression across the whole genome covering 550k SNPs and 54k mRNA probesets. These results can be applied to any disease area where knowledge of cis or trans effects between mRNA and SNPs that are involved in the statistically significant results of the analysis is required. The distance between the mRNA probeset and the SNP in the association can help determine if it is a trans effect or a cis acting element. This can be used to identify transcription factors and their binding sites for genes that can be used as targets to help regulate gene expression in diseased subjects.

Acknowledgment

The authors wish to thank Clyde Francks and Federica Tozzi for useful discussions, Israel Gloger and Robert Alexander for their enthusiastic support, and the Florian Holsboer with the staff at the Munich Max-Planck Institute of Psychiatry.

References

[1] P. Muglia, et al., "Genome-Wide association study of recurrent major depressive disorder in two European case-control cohorts", *Nature Publishing, Molecular Psychiatry*, 23 December 2008; doi:10.1038/mp.2008.131.

[2] J. F. Thompson, et al., "Comprehensive Whole Genome and Candidate Gene Analysis for Response to Statin Therapy in the Treating to New Targets (TNT) cohort." *Cardiovascular Genetics,* 12 February 2009, doi: 10.1161/CIRCGENETICS.108.818062.

[3] A. L. Dixon, L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, I Gut, M Farrall, G.M. Lathrop, G. R. Abecasis, W. Cookson. "A whole-genome association study of global gene expression". *Nature Publishing, Nature Genetetics* 16 September 2007; 39, 1202-1207 | doi:10.1038/ng2109.

[4] Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. "Allelic variation in human gene expression," Science, 2002, vol. 297, pp. 1143.

[5] E. E. Shadt, et. al., "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, 2003, vol. 422, pp. 297-302.

[6] American Psychiatry Association (Corporate Author), "Diagnostic and Statistical Manual of Mental Disorders DSM-IV, 4th edition, June 2000, ISBN: 978-0890420256.

[7] E. Vittinghoff, et al., "Regresion Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models", 8 June 2008, Springer, ASIN:BOOOQECINM

[8] J. D. Storey, O. Tibshirani, "Statistical significance for genome-wide studies" *PNAS,* 2003, 100, 9440–9445.

[9] Y. Benjamini, Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal Royal Statistical Society*. 1995, Ser. B, vol. 57, pp. 289-300.

[10] G. R. Abecasis, W. O. Cookson, L. R. Cardon, "Selection Strategies for disequilibrium mapping of quantitative traits in nuclear families," *The American Journal of Human Genetics*., 1999, vol. 65, pp. A245.

[11] P. I. Good, "Permutation, Parametric, and Boostrap Tests of Hypotheses", 1 December 2004, Springer, ISBN: 979-0387202792.

[12] M. Ashburner, M. et. al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Publishing, Nature Genetics*, vol. 25, pp. 25-29, 2000.

[13] M. A. Harris et al., "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*., 2004, vol. 32, pp. D258-D261

[14] T. Bishop, P. Sham, "Analyses of Multifactorial Disease", Academic Press, 15 December 2000, ISBN:978-0121016104.