

# Prioritisation of Positional Disease Gene Candidates

Frances S. Turner<sup>1</sup>, Colin A. M. Semple<sup>1</sup>

**Keywords:** disease gene prediction, complex trait, co-expression, GO terms.

## Introduction

Although much work has been done in the linkage mapping of many common disorders with a genetic component such as diabetes, asthma, obesity and schizophrenia, few of the genes involved have been identified. These studies have often mapped disease susceptibility to regions of tens of megabases, potentially containing many hundreds of genes [4]. A method of prioritising such positional candidates for further study is therefore required. Typically this is done manually, by looking for genes whose characteristics (function/expression etc) fit with what is already known about the disease. Computational approaches have been developed to automate this process and produce a smaller list of candidate genes [1][5]. However, by definition such approaches would not be successful in identifying novel disease genes that do not fit with prior knowledge of the disease or where little is known about the underlying aetiology of the disease.

Some studies aiming to identify novel genes predisposing to polygenic/complex diseases assume that genes contributing to the same/similar disease phenotype may be in the same pathway and/or have functional similarities [1][2][7]. Genes of potential interest may be selected from a large list of positional candidates according to similarities such as shared functional annotation, co-expression, or evidence of protein interaction with genes known to contribute to that disease. However, looking for potential disease related pathways from long lists of positional candidate genes is likely to lead to numerous false positives due to the large number of non-disease genes compared to the number of disease genes. There has been no attempt to systematically combine all the functional annotation available to search for genes in similar pathways with mapping data in a statistically rigorous method.

## Results

We previously developed POCUS (Prioritisation Of Candidate genes Using Statistics) [6] an approach that does not rely on assumptions about the underlying aetiology of the disease, only that some of the different genes responsible may share an identifiable common feature. POCUS was based upon a search for functional annotation, including Gene Ontology (GO) terms and InterPro domains, shared between genes within different susceptibility regions for the same disease. Each gene within a susceptibility region is scored according to the features it shares with genes in other regions. The scores reflect the probability of the observed similarity being seen by chance so the false positive rate is controlled. Here we present numerous refinements to the algorithm that increase its speed and further reduce false positives, we also consider a wider range of annotation including multi-platform co-expression measures. We also show how a user may include prior knowledge of a disease, in the form of genes known to affect the phenotype under consideration, in POCUS to improve the success of the method.

We demonstrate that POCUS can successfully identify genes from common pathways using curated human pathway data from the Reactome database [3]. We also test the technique on 29 polygenic diseases for which at least three of the genes responsible have been identified and are present in the Online Mendelian Inheritance in Man (OMIM) database [2].

---

<sup>1</sup> MRC Human Genetics Unit, Crewe Road, Edinburgh, UK, E-mail: fturner@hgu.mrc.ac.uk

## Conclusions

We show that for some diseases it is possible to identify known disease genes from a long list of positional candidates by taking the top scoring genes identified by POCUS. This method was developed for human disease susceptibility regions but could be applied to model organisms where mapping data is often of higher resolution.

## References

- [1] Freudenberg, J. and Propping, P. 2002. A Similarity-Based Method for Genome-Wide Prediction of Disease-Relevant Human Genes. *Bioinformatics* 18:S110-5.
- [2] Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V.A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 30:52-55.
- [3] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33:D428-32.
- [4] McCarthy, M.I., Smedley, D. and Hide, W. New methods for finding disease-susceptibility genes: impact and potential. 2003. *Genome Biol.* 4:119.
- [5] Perez-Iratxeta, C., Bork, P. and Andrade, M.A. 2002. Association of Genes to Genetically Inherited Diseases Using Data Mining. *Nat Genet* 31:316-319.
- [6] Turner F.S., Clutterbuck D.R. and Semple, C.A. 2003. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 4:R75.
- [7] Yonan, A.L., Palmer, A.A., Smith, K.C., Feldman, I., Lee, H.K., Yonan, J.M., Fischer, S.G., Pavlidis, P. and Gilliam, T.C. 2003. Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. *Genes Brain Behav* 2:303-320.