

Factor Models for QTL Studies

Oliver Stegle,^{1,*} Leopold Parts,² Anitha Kannan,³ Richard Durbin,² and John Winn³

¹University of Cambridge, UK

²Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

³Microsoft Research, Cambridge, UK

(Dated: Feb 22, 2008)

The recent availability of large scale data sets profiling single nucleotide polymorphisms (SNPs) and gene expression across different human populations, has directed much attention towards discovering patterns of genetic variation and their association with gene regulation. Two aspects of the nature of expression profiles make the identification and interpretation of such associations difficult. Firstly, we expect that a variety of environmental, developmental and other factors influence gene expression which can obscure such associations. Secondly, the regulatory network linking genes makes it difficult to pinpoint causal relationships between SNPs and regulatory elements.

We address the first issue by proposing FA-eQTL, a factor-model that explicitly takes non-genetic variability into account, and thereby can significantly improve the power of an expression Quantitative Trait Loci (eQTL) study [2]. We discuss a variational Bayesian implementation of this model (Fig. 1), and point out rapid approximations that are applicable in certain situations. Applying our model to simulated and real world data we can demonstrate a significant improvement in performance. On data from the HapMap project [3], we find more than three times as many significant associations than a standard eQTL method.

To address co-expression of genes, we further extended FA-eQTL by jointly reducing the dimensionality of the expression profile and modelling non-genetic factors. We discuss results applying this enhanced QTL-model to biological data, including human [3] as well as datasets from yeast [1].

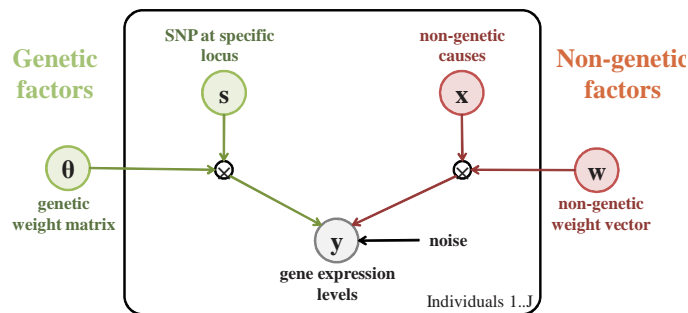


FIG. 1: The Bayesian network for our model that includes both genetic (green) and non-genetic factors (red) when explaining gene expression levels. The rectangle indicates that contained variables are duplicated for each individual.

-
- [1] R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102:1572–7, Feb. 2005. PMID: 15659551.
 - [2] O. Stegle, A. Kannan, R. Durbin, and J. Winn. Accounting for Non-Genetic Factors Improves the Power of eQTL Studies. to appear *RECOMB*, 2008.
 - [3] B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavare, P. Deloukas, and E. T. Dermizakis. Population genomics of human gene expression. *Nat Genet*, 39:1217–1224, Oct. 2007.

*Electronic address: os252@cam.ac.uk