

Global inference of human phenotype-genotype associations by integrating interactome, phenome and diseasome

Xuebing Wu¹, Michael Q. Zhang^{1,2}, Shao Li^{1,*}

1. MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/
Department of Automation, Tsinghua University, Beijing, China
2. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

*E-mail: shaoli@mail.tsinghua.edu.cn

Uncovering genotypes underlying specific phenotype, especially human disease, is one of the principal goals for genetics research and is of vital importance for biomedicine. We introduce CIPHER (Correlating Interactome and PHEnome to pRedict disease genes), a systems biology approach for phenotype-genotype association inference and disease gene discovery. The core of CIPHER is a mathematical model dissecting the modularity of human genetic diseases, that similar diseases are caused by functionally related genes [1]. CIPHER explains phenotype similarity using genotype distance in molecular network.

By integrating protein interaction network, phenotype similarity network and human disease network [2], CIPHER systematically quantifies the relevance between genes and diseases to prioritize candidate genes. It has the following advantages: **i**) it achieves very high power in detecting known disease genes, **ii**) it scales well to genome-wide scan of disease genes, **iii**) it is applicable to nearly all the phenotypes, as far as their clinical features are sufficiently characterized, and finally **vi**) it parallels one classical model in transcriptional binding motif discovery from expression data, and is readily to implement the sophisticated strategies capable of identifying motif interactions to explore gene interactions in genetic diseases.

Benchmark test on artificial linkage loci shows that CIPHER outperforms a method [3] using similar data sources (Figure 1) and successfully ranks known disease genes as top 1 in 709 of 1444 linkage intervals, yielding a 53.3 fold enrichment (FE) over random selection, significantly outperforms existing methods (with FE<24). In genome-wide scan of disease genes, CIPHER correctly selects known disease genes out of 8919 in ~10% of the cases, achieving a fold enrichment of ~900, again significantly outperforms existing methods (with FE<70). In a case study of breast cancer, CIPHER ranks most of the known susceptibility genes within top 5% of the genome, and also assigns ranks to novel susceptibility genes MAP3K1 [4], GGA1 [5] and IKBKE [6] within top 0.4%, 1% and 8% of the ranked human genome respectively. Tentatively, we further explore gene-gene interactions in breast cancer using an excellent approach [7] for inferring transcriptional factors interactions, and discover a sub-network centered around BRCA1 and BRCA2, the only two major risk factors identified for breast cancer.

We use CIPHER to infer genome-wide molecular basis for 1126 disease phenotypes, systematically examining the the genetic background of a wide spectrum of phenotypes. The resulted disease relevance profile is a rich resource for disease gene discovery, which can be used to guide the selection of disease candidates for gene mapping. It also provides comprehensive information on genetic overlaps of various complex phenotypes, which can be immediately exploited in the design of genetic mapping approaches that involve joint linkage or association analysis of multiple seemingly disparate phenotypes. Finally, clustering of the

profile (Figure 2) presents a global picture of the modularity of genetic diseases, suggesting the existence of disease module, which comprises a set of functionally related genes and a set of genetically overlapped diseases, in which the gene set is highly relevant to the disease set. Our disease relevance profile establishes a 'multi-diseases multi-genes' view of genetic diseases, and will facilitate the study of systems biomedicine.

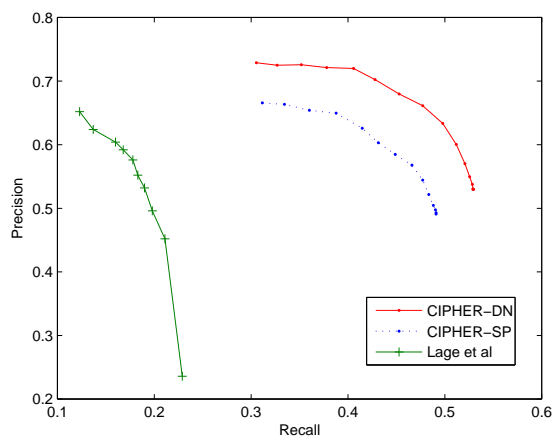


Figure 1: Performance compared with the method of Lage et al. [3]. Precision is plotted against recall, which shows that CIPHER's precision for high-scoring candidates can approach 73% and 67%, better than that of Lage et al.'s method, while maintaining a high recall. CIPHER-DN and CIPHER-SP are two implementations using different neighborhood systems: direct neighbor (DN) and shortest path (SP), respectively.

References

- [1] Oti, M. *et al.* The modular nature of genetic diseases. *Clin. Genet.* **71**, 1-11 (2007)
- [2] Goh, K. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685-8690 (2007)
- [3] Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309-316 (2007)
- [4] Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1093 (2007)
- [5] Sjöblom, T. *et al.* The Consensus Coding Sequences of human breast and colorectal cancers. *Science* **314**, 268-274 (2006)
- [6] Boehm, J.S. *et al.* Integrative Genomic approaches identify IKBKE as a breast cancer oncogene. *Cell* **129**, 1065-1079 (2007)
- [7] Das, D. *et al.* Interacting models of cooperative gene regulation. *Proc. Natl. Acad. Sci. USA* **101**, 16234-16239 (2004)

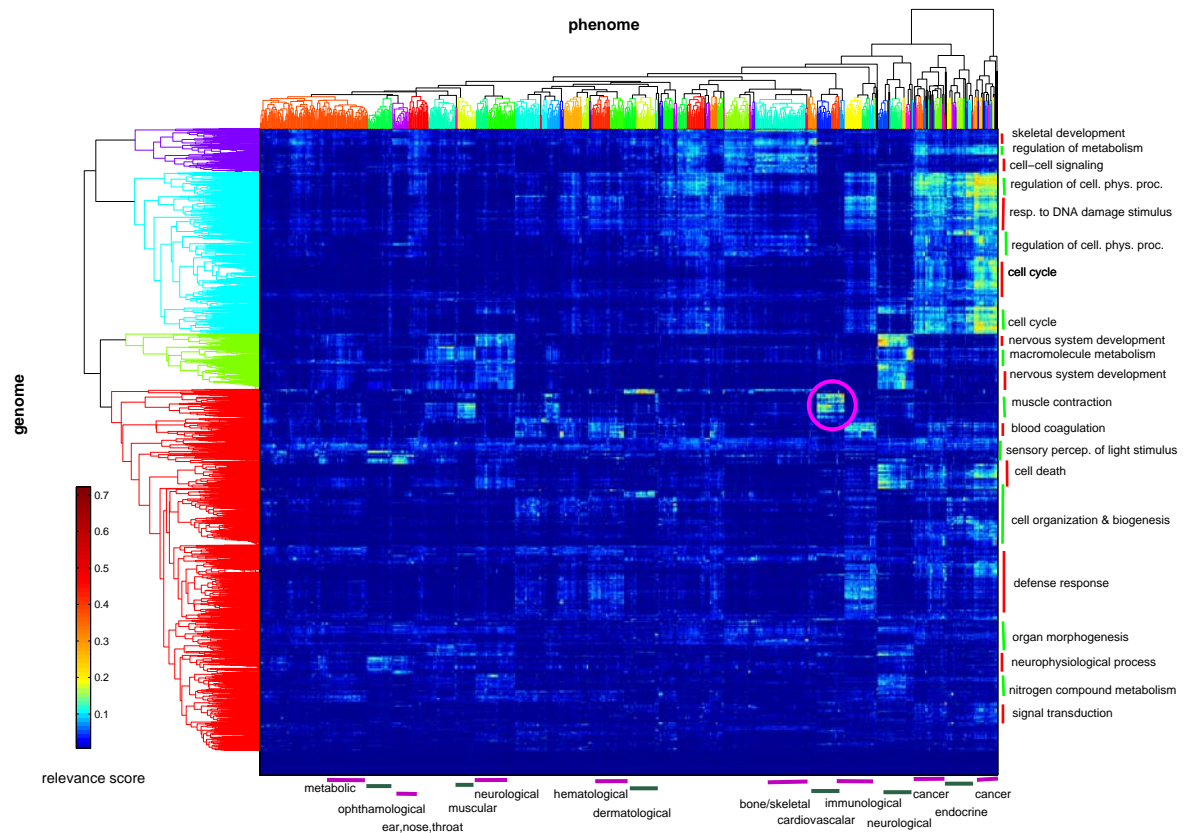


Figure 2: Hierarchical clustering of disease relevance profiles, depicting ~ 9 million relevance scores quantifying all associations between 8919 genes and 1126 diseases. The color of each cell represents the relevance score R_{pg} of a disease (column) and a gene (row), where red/blue indicates high/low relevance score (upper left). Phenotype clusters are manually inspected and annotated with enriched disease categories (bottom), and gene clusters are annotated with enriched biological process terms of Gene Ontology (right). Highly scored blocks are defined as disease modules, each comprising a set of functionally related genes implicated in a set of genetically overlapped diseases. For example, the pink circled region indicates a module comprised of a gene set of muscle contraction involving in a set of cardiovascular diseases.