GWAS next generation: identifying mechanisms of action in association studies.

María Rodríguez Martínez, Paola Nicoletti, Gonzalo López, Mukesh Bansal, Yishai Shimoni, Andrea Califano.

Columbia University, 10033 NY, USA.

Genome wide association studies (GWAS) have emerged as a powerful tool for the identification of genetic variants that are associatedwith complex phenotypes and disease.Despite the many newly discovered associations, thevariants identified by these studies typically explain only a small fraction of the heritable component of disease risk[1]. Furthermore,few genetic variants are found within coding regions of genes, and theelucidation of the molecular mechanism by which these loci influencethe phenotype remains challenging. Most loci mapto inter-genic regions of unknown function and, whilesome of them can be connected to nearby genes by linkage disequilibrium, a sizable fraction lie in genomic regions with no clear connection to known disease biology.

The genetic component of complex phenotypes can also arise from a large number of small effect loci[2]. In this case, the heritability would not be due to a single common or rare variant, but rather to combinations of common variants, each one contributing a small additive effect. These combinations can be epistatic interactions among common alleles, or multiple genetic variations that interact through different layers of genomic regulation. Complex phenotypes therefore would have a much more complex genetic architecture due to the joint action of very many loci of small effect [3]. Identifying interactions between multiple loci requires the application of statistical and computational methods that detect patterns of epistasis across the genome. This involves performing genome-wide searches of high order combinations of SNPs or SNPs and genes, and requires testing a large number of hypothesis with often limited sample sizes, leading to a reduced statistical power. Furthermore the computational search becomes unmanageable for more than a few hundredSNPs.

In this work we explore an innovative approach to identify the molecular mechanisms of genetic variants previously associated to disease. We have implemented gVITaMIN (Genetic Variability IdenTifies Missing INteractions), an algorithm that searches forfunctional genetic associations following a two-step approach. First, gVITaMIN searches for direct associations between a locus and gene expression levels. However a SNP can be functionally important for a phenotype without displaying any association with gene expression, therefore in a second step, gVITaMIN searches

for associations between a locus and changes in gene activity. Concretely, we analyze whether aputative locusinfluences the regulatory activity of a transcription factor (TF) over a large set of its target genes (TG). This influence is measured as a difference in the correlation between the TF and its TGs conditioned upon the presence of the variant.

In recent years, a plethora of epigenetic modificationsin the human genome have been characterized and shown to play diverse roles in gene regulation, cellular differentiation and the onset of disease. In particular, regulatory elements such as transcriptional enhancers and silencers, or chromatin marks such as promoters and enhancers, have been shown to play a crucial role in the establishment and maintenance of specific gene regulatory programs. These elements can be perturbed by genetic variants. For instance, mutations in regulatory elements can disrupt or enhance the binding of transcription factors and alter gene expression; polymorphisms that overlap with chromatin marks can prevent regulation through methylation or acetylation and hinder transcription factor binding, etc. In order to integrate this level of genomic regulation, we use to the ENCyclopedia Of DNA Elements (ENCODE) [4, 5] to identify loci that map to characterized functional elements of the human genome. These loci are scored according to their proximity to the genomic element, and linked to the gene or genetic program associated to the functional mark.

Finally, cumulative associations in a particular pathway are likely to pinpoint specific regulatory programs associated with a disease.We therefore search for functional variants associated to a phenotype that cluster on biological units, such as genes or pathways.We combinethe TFs and TGs predicted to be associated to the loci with the genes and genetic programs identified through genomic mapping, and search for enriched pathways on these genes usingGene Set Enrichment Analysis (GSEA) [6]. These pathways are likely to be phenotypically relevant to the physiopathology of the disease and provide insights into the molecular mechanisms underlying them.

We have applied gVITaMIN to the study of genetic variants associated to breast cancer susceptibility. We will present preliminary data identifying an intriguing association between BARD1, a genethat forms a heterodimer with BRCA1, and it is essential for the stability of BRCA1. Several variants at this locus have been reported to be associated with high risk neuroblastoma and colon cancer, suggesting a role in disease that is active across different cancers phenotypes.

BIBLIOGRAPHY

1. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease*. N Engl J Med, 2010. **363**(2): p. 166-76.
2. Valdar, W., et al., *Genome-wide genetic association of complex traits in heterogeneous stock mice*. Nat Genet, 2006. **38**(8): p. 879-87.
3. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease*. Nat Rev Genet, 2010. **11**(6): p. 446-50.
4. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types*. Nature, 2011. **473**(7345): p. 43-9.
5. Ernst, J. and M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome*. Nat Biotechnol, 2010. **28**(8): p. 817-25.
6. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.