

Network approach to Genome-Wide Association studies

Daniel Remondini^{1,2*}, Mirko Francesconi³, Francesco Lescai^{4,5}, Gastone Castellani^{1,3}

¹INFN, Alma Mater Studiorum Università di Bologna

²Dipartimento di Fisica, Alma Mater Studiorum Università di Bologna

³DIMORFIPA, Alma Mater Studiorum Università di Bologna

⁴CIG - Centro Interdipartimentale "L.Galvani", Alma Mater Studiorum Università di Bologna

⁵ITB – Istituto di Tecnologie Biomediche, CNR Milano

*address correspondence to Daniel Remondini: daniel.remondini@unibo.it

Single nucleotide polymorphisms (SNP) are variations of single base along the DNA sequence generally characterized by a bi-allelic feature, i.e. two possible variants exist.

During the past six years these common variations became progressively important given the possibility to easily implement the SNP genotyping by dense arrays and robotic instruments.

By identifying most of the approximately 10 million SNPs estimated to occur commonly in the human genome, the International HapMap Project is identifying the basis for a large fraction of the genetic diversity in the human species.

The International HapMap Project is a multi-country effort to identify and catalogue genetic similarities and differences in human beings. Using the information in the HapMap, researchers will be able to find genes that affect health, disease, and individual responses to medications and environmental factors. The Project is a collaboration among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States.

The goal of the International HapMap Project is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared.

However, testing all of the 10 million common SNPs in a person's chromosomes would be extremely expensive. The development of the HapMap will enable geneticists to take advantage of how SNPs and other genetic variants are organized on chromosomes. Genetic variants that are near each other tend to be inherited together. These regions of linked variants are known as haplotypes.

For geneticists, SNPs act as markers to locate genes in DNA sequences, and an increasingly dense catalogue of such variation is elucidating the complex architecture of the human genome. Moreover an increasing convergence emerges on the existence of a wired interplay between multiple factors with a small effect size, underlying complex traits.

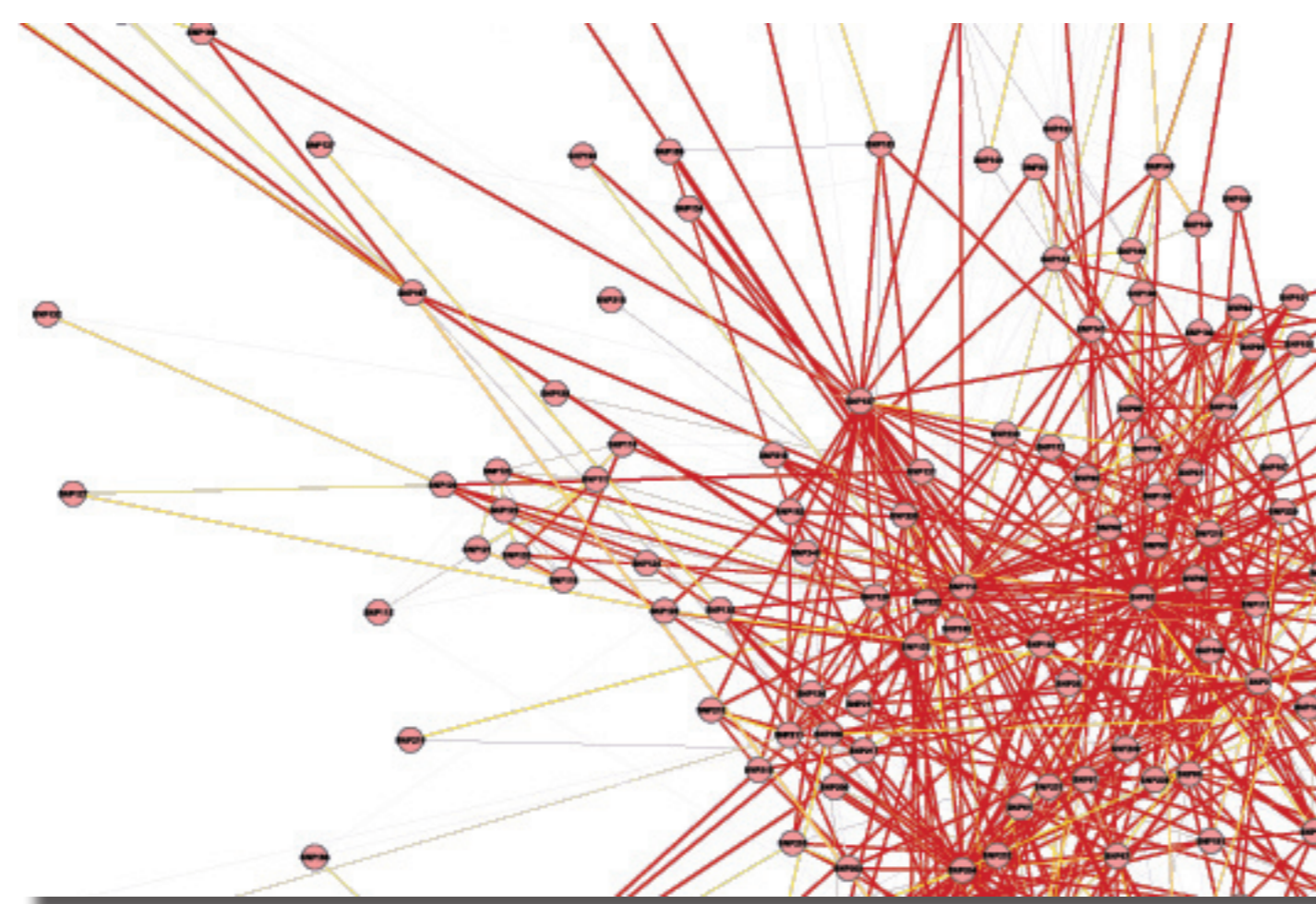
High throughput genetic analysis is progressively gathering laboratories in large consortia, capable to exploit the potential of new technologies and very large sample sets in catching the genetic determinants of complex diseases[1,2,6,11,28,29].

This resource and technological scale is driving genetic studies design from a family-based inheritance (co-segregation) studies (i.e. linkage studies), to population-based association studies.

However, despite well established statistical methods in family genetic studies, as far as large and genome-wide association (GWA) studies are concerned, a consistent debate is still ongoing on the best approaches to overcome the major issues inherent to such study designs[3-5,9,10,13,16-20,22].

The most widely used statistical tests are single point statistics (chi-square, or Cochran-Armitage test) along the genome; these tests can be integrated with haplotype (or multi-marker) analysis once the linkage disequilibrium (LD) structure is drawn and thus haplotype blocks have been identified.

All these tests can be performed under



different assumptions and with slightly different approaches, and multivariate analyses are generally performed.

Two main obstacles can be envisaged looking at the literature and according to many groups' experiences:

- 1) the false positive rates, and consequently the efficacy of the corrections adopted;
- 2) The capability to identify low-penetrance variations across the human genome.

As for false positives, many different approaches have been proposed and, provided the sample collection to be large enough, a two-stage design has been shown to be very effective in detecting key leads in the genome, often replicated in other populations. It's not the purpose of this paper to address this area[3,24].

As for the identification of low-penetrance polymorphisms is concerned, the area is of a major consideration when disentangling the picture of any complex trait. Indeed, it's quite realistic for complex phenotypes to be determined by a combination of many different polymorphic loci each of them accounting for a minor part of the total variance, hence very difficult to be detected when a genome-wide genotyping is performed and when GWA significance rates are applied[26].

Despite this issue being of a key importance, most of the papers reporting GWA studies applied single point statistics, multi-marker analysis and haplotypes analyses, performed LD mapping, adopted different false-positive rate corrections[7,8,12,21,25]. Few of them actually included interaction analysis and other similar approaches capable to grasp the effect of interactions and across-genome combinations, rather than the major effect of single markers or (despite more importantly) the major contribution of a specific haplotype in a locus[14,15,23].

Recent approaches have been proposed, that take into account the experience of gene expression analysis, as a similar problem has been experienced in microarrays data analysis. These approaches try to exploit the so called "a priori biological knowledge", by taking into account the biological role of genes that can be grouped in larger structures such as pathways

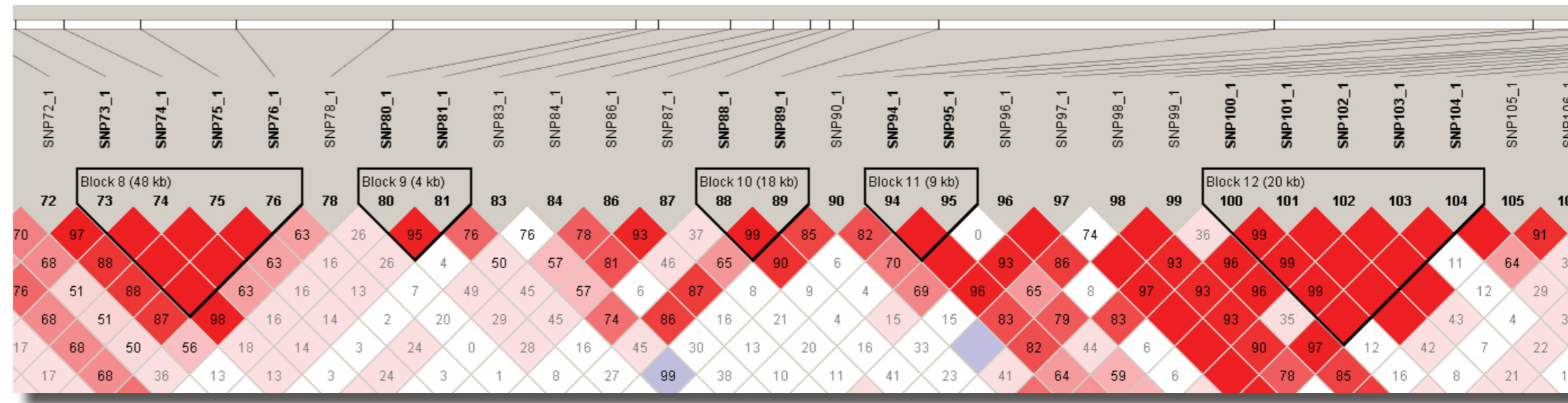
and ontologies[27].

Appropriate methods are in development, as single nucleotide polymorphisms (SNPs) genetic analysis imply slightly different datasets, as well as values which are not independent each other within specific sub-groups, due to physical reasons, i.e. the existence of linkage disequilibrium (LD) and LD blocks across the genome.

This poster proposes a multiscale method for network reconstruction and analysis starting from single SNP relationships,

with a bottom-up approach progressively recovering higher level structures: LD blocks, gene loci and biological pathways.

The network reconstruction method is based on the calculation of snps correlation matrices (or other distances used to estimate LD like D'): one



of the current issues is how to interpret intermediate values of (D').

The key idea here is to relax the threshold level of D' trying to grasp long distances correlations in the network that likely reflect complex genetic interactions and which are the

most interesting targets for the analysis. The resulting networks are highly connected. The use of such networks can improve the efficiency in identifying multi-marker relationships, but it can pave the way to the implementation of innovative approaches never applied to genetic epidemiology before.

Within a network reconstruction, as an example, clusters will represent linkage disequilibrium blocks, thus providing a more unifying method for the identification of these physical correlations in the genome. At the same time edges between these LD clusters will provide a more immediate tool for highlighting long-range interactions where the appropriate threshold is set.

The use of this network approach will also allow us to map on this SNP representation the different scales of biological information available: cluster groups will describe genes and inter-genetic regions, and connections between groups of cluster-groups will integrate the information on biological pathways.

Different variables can then be integrated into the description of the network's edges and nodes, and include phenotypic variables, case/control status, other association single-point measures etc. thus improving the efficiency of the analysis.

Several network description measures can be highly helpful in this scenario in disentangling the interactions at these different scales underlying complex traits.

A critical problem in applying such a methodology is that network reconstruction algorithms are computationally very expensive and require parallel computing to be solved. The dimensionality of the matrices can vary from 50k x 50k in the case of a genome wide low density genotyping up to 1M x 1M in case of top density genome-wide studies. The application of algorithms for the analysis and comparison of networks of such high dimensionality is challenging. Moreover the embedding of other covariates such as the pathological state with respect to physiological one the a priori biological knowledge in terms of genes and genes interaction, pathways and network of pathways is even more challenging in terms of computation, data storage and modelling.

References:

- [1] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661-78.
- [2] Butcher LM, Plomin R. The Nature of Nurture: A Genomewide Association Scan for Family Chaos. Behav Genet 2008.
- [3] Clarke GM, et al. Fine mapping versus replication in whole-genome association studies. Am J Hum Genet 2007;81(5):995-1005.
- [4] Curtis D. Allelic association studies of genome wide association data can reveal errors in marker position assignments. BMC Genet 2007;8:30.
- [5] Dong C, et al. Gene-centric characteristics of genome-wide association studies. PLoS ONE 2007;2(12):e1262.
- [6] Florez JC, et al. A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. Diabetes 2007;56(12):3063-74.
- [7] Hakonarson H, et al. Polychronakos C. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. Diabetes 2008;57(4):1143-6.
- [8] Hinney A, et al. Genome Wide Association (GWA) Study for Early Onset Extreme Obesity Supports the Role of Fat Mass and Obesity Associated Gene (FTO) Variants. PLoS ONE 2007;2(12):e1361.
- [9] Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. Hum Hered 2007;64(4):203-13.
- [10] Ioannidis JP, et al. Heterogeneity in meta-analyses of genome-wide association investigations. PLoS ONE 2007;2(9):e841.
- [11] Ionita-Laza I, et al. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. Am J Hum Genet 2007;81(3):607-14.
- [12] Kayser M, et al. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am J Hum Genet 2008;82(2):411-23.
- [13] Kingsmore SF, et al. Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov 2008;7(3):221-30.
- [14] Kooperberg C, Leblanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol 2008;32(3):255-63.
- [15] Kooperberg C, et al. Sequence analysis using logic regression. Genet Epidemiol 2001;21 Suppl 1:S626-31.
- [16] Li C, Li M, et al. Evaluating cost efficiency of SNP chips in genome-wide association studies. Genet Epidemiol 2008.
- [17] Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet 2008.
- [18] Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol 2008;32(3):215-26.
- [19] Macgregor S, et al. Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. Nucleic Acids Res 2008;36(6):e35.
- [20] Pearson TA, Manolio TA. How to interpret a genome-wide association study. Jama 2008;299(11):1335-44.
- [21] Raelson JV, et al. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A 2007;104(37):14747-52.
- [22] Rao DC. An overview of the genetic dissection of complex traits. Adv Genet 2008;60:3-34.
- [23] Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. Biostatistics 2008;9(1):187-98.
- [24] Skol AD, et al. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 2007;31(7):776-88.
- [25] Todd JA, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39(7):857-64.
- [26] Tomlinson I, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39(8):984-8.
- [27] Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 2007;81(6).
- [28] Wilk JB, et al. Framingham Heart Study genome-wide association: results for pulmonary function measures. BMC Med Genet 2007;8 Suppl 1:S8.
- [29] Zeggini E, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008.