

Short Communication

doi: 10.1111/j.1469-1809.2011.00659.x

A Report of the First Biennial Meeting on Capita Selecta in Complex Disease Analysis (CSCDA2010), Leuven, Belgium, August 25–27, 2010

Kristel Van Steen^{1,2*} and Kristel Slegers^{3,4}¹*Systems and Modeling Unit, Montefiore Institute, University of Liege, Grande Traverse 10, 4000 Liège, Belgium*²*Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium*³*Neurodegenerative Brain Diseases Group, Department of Molecular Genetics, VIB, Antwerpen, Belgium*⁴*Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, Antwerpen, Belgium*

Summary

There is a need for interdisciplinary assessments and interpretations of -omics underpinnings of human complex diseases. However, often investigators from different, yet overlapping, disciplines experience difficulties in understanding the other discipline's language and there is a clear need for establishing a platform that nourishes interdisciplinary team processes and allows tearing down the professional's tower of Babel. To accommodate these needs, the biennial mini-conference Capita Selecta in Complex Disease Analysis was instigated. Abstracts are freely available online [<http://www.aimontefiore.org/cscda2010/>].

Keywords: Complex disease analysis, gene-gene and gene-environment interactions, integrated -omics analysis

Introduction

The first meeting of "Capita Selecta in Complex Disease Analysis" brought together 73 international researchers from a variety of disciplines, including systems biology, biostatistics, biomedical sciences, medical and molecular genetics, bioinformatics, engineering, and computer science. The geographical spread of the participants covered Europe and the United States. Content wise, this year's meeting had two focal points: detecting and investigating epistatic effects and integrating -omics data sets in the postgenomic era. It was motivated by the current belief of a growing number of investigators that advanced methods in statistical genetics should make better use of the vast amount of information that can be retrieved from proteomics, transcriptomics, epigenomics, and metabolomics, to name but a few.

The conference talks of CSCDA2010 were preceded by several short courses given by invited lecturers (*David Evans, University of Bristol, UK and Christophe Lambert, Golden Helix, USA*) on basic concepts in statistical genetics of genome-wide data sets. These educational sessions allowed PhD stu-

dents and investigators with different backgrounds to acquire the necessary basis for the subsequent mini-conference. The mini-conference itself comprised a series of high-level seminars, alternating invited talks of keynote speakers and selected contributions, with an emphasis on interaction between participants and invited speakers. In particular, morning sessions were devoted to keynote speakers and afternoon sessions were moderated by an expert researcher, stimulating discussions between case-study presenters and the audience. This rather unusual format worked particularly well, and bridged the gap between newcomers in the field, and well-established researchers. It also enhanced brainstorming about relevant topics related to human genetics data analysis and translation to personalized medicine, in a friendly and rather informal atmosphere. In what follows, we present a few highlights of the scientific part of the meeting.

A World of Interactions

The session on "interactions" was opened by a keynote lecture by *Jason Moore (Dartmouth Medical School, USA)*. It has become clear that only a small percentage of the total genetic heritability of a trait can be explained by the loci identified to-date. This may be attributed to the fact that in reality epistatic effects can be found in addition to gene-environment

*Corresponding author: Kristel Van Steen, Systems and Modeling Unit, Montefiore Institute, University of Liege, Grande Traverse 10, 4000 Liège, Belgium. Tel: 324 366 2692; Fax: 324 366 2989; E-mail: kristel.vansteen@ulg.ac.be

interactions and there is the potential for multiple associations with small effect sizes, as well as non-SNP polymorphisms or epigenetic effects. Actually, Moore went some steps further and argued that interactions among loci or between genes and/or environmental factors may make a rather substantial contribution to variation in complex traits, such as disease susceptibility. Furthermore, he stated that the scientific community currently largely underdelivers on using genetics to improve health care by not fully exploiting the existing complexity.

The term epistasis has several distinct, yet related meanings (Phillips, 2008), and Moore pointed out that a cautionary attitude should be adopted when translating apparent “evidence” of statistical epistasis into genetic epistasis, or vice versa. In addition, mechanistic interactions will always be hard to prove, because markers themselves are not causal. Therefore, unraveling the genetic architecture of complex diseases is not an easy task (Van Steen, 2011). The complexity of the problem offers a number of challenges for several interdisciplinary fields, including bioinformatics and statistics, Moore continued. One of these challenges pertains to the breakdown of standard parametric approaches in the presence of sparse data. Another challenge is the follow-up on findings and to investigate the importance of these findings for public health.

Structural alterations (e.g., insertions and deletions) result in increased single-nucleotide changes in their neighborhood. This relationship holds well across different time scales and therefore appears to be a time-invariant principle of genome evolution. A better understanding of this relationship, for instance by comparing different genomes, will therefore not only have its impact on GWAs, but also on gene therapy and disease genomics. To illustrate this, *Madan Babu (MRC Laboratory of Molecular Biology, Cambridge, United Kingdom)* investigated in more detail whether mutations of different sizes are independent of each other or, rather, linked. At the species level, it turns out that genes with structural alterations in nearby regions, show increased single nucleotide changes. In human genomes, the single nucleotide substitution rates are higher near sites of structural alteration and decrease with increasing distances. Increased insights are needed toward how this established linkage affects the functional elements in the human genome. Understanding the role of such a relationship and its implications for uncovering the network of genetic interactions appears to be an important direction of future research.

Nilanjana Chatterjee (National Cancer Institute, USA) explained about the needs to understand interactions, in general, and interactions of genes with environmental factors in particular. Although this brings about an extra component of difficulty in the analysis (namely, the reliable measuring of environmental agents), understanding these types of interactions is necessary in order to improve the power for gene

discovery. As for the case of gene–gene interaction detection, gene–environment interactions may be removable or non-removable, depending on whether or not they are present irrespective of the scale on which the effect modification is measured. As Chatterjee indicated, first, the target of the study needs to be identified: whether interest is in a particular environmental agent and a handful of genes, or the more challenging large-scale screening of interactions among several environmental factors and the genome. Second, the appropriate study design (for instance, case-only or not) and corresponding statistical test need to be selected. Third, unbiased estimates of interaction odds ratios need to be computed.

The afternoon case-study sessions nicely showcased some subtopics that are currently of interest in complex disease analysis. *Marylyn Ritchie (Vanderbilt University Medical School, USA)* acted as a moderator for this session and organized the discussions around the following topics, based on the input during the young-investigator session: (1) Does trait heterogeneity correlate with genetic heterogeneity? What about changing phenotypes? How can we identify clinically important subphenotypes? (2) With full genome sequences arising on the horizon, which genetic variants should we use for better explaining genetic heritability of traits? Are the estimates of genetic heritability derived in the early days of genomics reliable? (3) Although pure interactions have been statistically replicated, often no sufficient foundations have been given for a biological validation. So what does replication mean in this context? (4) Which analysis method to use when having different types of phenotypic data and different types of both rare and common genetic variants available? What are the potentials for performing integrated analyses, hence merging information from different data sources?

Adopting the Systems Biology View to Human Complex Disease Analysis

Ivo Gut (Centro Nacional de Análisis Genómico, Barcelona, Spain) opened the session on “–omics data integration.” He argues that by using additional layers of biological information, researchers are better able to pinpoint causative variants and genes than by using standard genome-wide association techniques using arrays of SNPs and/or CNVs, despite the fact that these arrays have been useful in identifying candidate genomic regions and hence in providing handles on functional networks. With the evolution of the latest second-generation nucleic acid sequencing technology, one way of looking beyond the popular GWAs-related tagging SNPs, which often reside outside coding regions or known regulatory elements, is to extend genetic association methods, or to develop integrated methods for –omics analysis, using in-depth sequencing data of candidate regions of interest, or data derived from

genome-wide sequencing efforts. Apart from improved power to identify causal variants, evidence is growing that these sequencing efforts may even uncover the rarer variants with higher effects and with more direct functional consequences for common diseases. With respect to using sequencing data as reference data in analyses, Gut launches a cautionary note, in that often it is forgotten that the integrated use of reference sequences will bias the experiment toward the reference, away from the actual situation.

In line with adopting a systems biology point of view at the human genetics level, *Peter van der Spek (Erasmus MC, Rotterdam, the Netherlands)* illustrated the importance of considering genomic information as a whole, and integrating different data sources, such as genomics, proteomics, cytogenetics, and imaging data to identify genetic associations with, for instance, brain development. He showed how relating molecular data with imaging data paves the way for improved (image-guided) diagnosis and intervention.

Peter Holmans (Cardiff University School of Medicine, United Kingdom) presented an overview of current issues involved in performing pathways analyses. One explanation for the small effect sizes observed in complex traits is genetic heterogeneity, defined as the presence of disease susceptibility alleles in different genes in affected individuals. If the susceptibility genes act together in a biological pathway, then a joint analysis of the set of genes in that pathway may boost up the power to detect a genetic association.

Several questions were raised during the keynote lecture. (1) How should pathways be defined? Which pathway database resources should be used and how sensitive are the results to different database usages? (2) Should prior information regarding pathways that are known to be “involved” in the disease pathways be used in a GWA, or should information about “all pathways” be considered? (3) What is the best way to map SNPs to genes? (4) How to score pathways and to formally assess statistical significance, while adequately correcting for multiple testing? (5) Should replication be performed at the gene level or at the pathways level? Should we rather envisage biological replication and/or target replication in different diseases?

An interesting discussion provided a few answers. For instance, (1) pathways can be defined in a variety of ways, including via systems biology approaches. In addition, pathway-based analysis of GWA data without prior selection of a relevant pathway may point toward biologically important genes that would be left undetected by GWAs targeting common variants. (2) Jointly considering multiple contributing factors in the same pathway might complement traditional single SNP/gene approaches and can provide additional insights in interpreting GWA data on complex diseases. (3) Caution should be taken in mapping SNPs to genes, so as to interpret GWA findings, especially in the light of the existence

of so-called “synthetic associations” to SNPs that are quite distant from the true causative variants. (4) Univariate pathway approaches include testing for enrichment of gene groups in gene lists, or comparing distributions of scores within each gene group to random selections. Ideally, the significance of pathway scores is assessed on the basis of biological connotations. Alternatively, multivariate pathway approaches such as Principal Components Analysis, Singular Value Decomposition, or Multidimensional Scaling are considered. (5) Finally, similar to GWAs (Igl et al., 2009) a distinction should be made between “replication” (both original and confirmation sample are drawn from the same population; systematic differences are reduced to a minimum) and “validation” (the confirmation sample originates from a population, which is different than that from which the original sample was drawn) of pathway analysis results.

The heterogeneity of the topics covered by the “-omics integration” afternoon case-study session, reflected the wide span of interests this topic involves. *Lude Franke (UMC Groningen, the Netherlands)* acted as a moderator for the afternoon session and summarized the young-investigators’ contributions on data integration as follows: (1) Most integrated analyses involved bio-data integration via statistical models or bioinformatics tools, or via a variety of visualization techniques. (2) Some investigators dwelled upon possible ways to integrate (combine) different study designs in a single analysis. (3) With a wealth of information available, there is always the concern of having overfitted the data; proper measures should be taken to account for this. (4) When data become too enormous for easy handling, data reduction methods may be used.

One popular data reduction method is the principal components method. Franke extended the discussions by elaborating on the usefulness of deriving principal components from gene expression data. He emphasized the biological relevance of these components, while still explaining a sufficient percentage of the total variation in the data, and argued in favor of them to be used while relating gene expression to genetic variations.

The aforementioned type of analysis linking expression data to SNPs only involves integrating two data sources. Yet the complex nature of SNP genotype effects on gene expression, the concerns about the choice of relevant tissues, the collection of large data sets, the statistical analysis, and the computational burden of the analysis give a flavor of the difficulties future researchers in this area are bound to encounter.

It is all in the Genes

Inflammatory bowel disease (IBD), with Crohn’s disease (CD) and ulcerative colitis as major phenotypes, involves the interplay of environmental risk factors with immunological

changes that trigger onset of disease in a genetically susceptible host. The disease carries a very heterogeneous presentation. Séverine Vermeire (University Hospital Leuven, Belgium) gave an excellent historical overview about the success story of IBD genetic analysis, whether via genome-wide linkage studies or association studies. Despite the fact that over 100 loci have been identified, the most understood gene for IBD, namely *CARD15*, only explains about 20% of the genetic predisposition to CD. Moreover, to date, it is unclear how many susceptibility genes underlie IBD and how these variants may interact with each other and/or with environmental agents. Joining forces and brainstorming about more efficient ways of integrating the various sources of biological and nonbiological information will be most helpful in elevating the research on IBD and other complex human diseases to the next level.

In Conclusion

Much of the discussion during CSCDA2010 involved subjects that are still developing. In particular data integration should not only involve combining data using statistical approaches and data fusion with biological knowledge using a variety of bioinformatics and computational tools, but should also involve a thorough investigation of the interactions between the different components of biological systems; these themes linked the two research days of CSCDA2010.

In addition, there is a clear need to provide a roadmap in order to find your way in the wealth of currently available bioinformatics tools. Supplementary efforts are required to set up and carry out well-designed experiments to biologically validate model-based findings. The integration of diverse experimental sources and the adoption of a systems view, using interdisciplinary tools and personnel, are needed in order to further unravel the causes of complex diseases and to generate improved diagnostics and therapies.

The second meeting of Capita Selecta in Complex Disease Analysis (CSCDA2012) will be held in Liège (Belgium) from May 30 through June 1, 2012. For more details, we refer the reader to <http://www.aimontefiore.org/>.

Acknowledgements

We thank all active participants of CSCDA2010, who kindly provided their presentation slides, hereby facilitating the drafting of this meeting report.

Furthermore, K Van Steen acknowledges the research opportunities offered by the Belgian Network BioMAGNet (Bioinformatics and Modelling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Program (Phase VI/4), initiated by the Belgian State, Science Policy Office. This work is also supported in part by the IST Program of the European Community, under the PASCAL2 Network of Excellence (Pattern Analysis, Statistical Modelling and Computational Learning), IST-2007-216886. K Slegers acknowledges the Fund for Scientific Research-Flanders (FWO-V).

References

- Igl, B. -W., König, I. R. & Ziegler, A. (2009) What do we mean by 'replication' and 'validation' in genome-wide association studies? *Hum Hered* **67**, 66–68.
- Phillips, P. C. (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, **9**, 855–867.
- Van Steen, K. (2011) Travelling the world of gene-gene interactions. *Brief Bioinform* 2011 Mar 26. [Epub ahead of print] PMID: 21441561.

Received: 2 Jan 2011

Accepted: 31 Mar 2011