BJMG ,Supplement 15 (2012) 45-50
10.2478/v10034-012-0018-7

Proceedings of the MACPROGEN Final Conference held
at Ohrid, Republic of Macedonia, March 29-April 1 2012

INTEGRATIVE 'OMIC' APPROACH TOWARDS UNDERSTANDING THE NATURE OF HUMAN DISEASES

Peterlin B*, Maver A

***Corresponding Author:** Professor Borut Peterlin, M.D., Ph.D., Institute of Medical Genetics, Department of Gynecology and Obstetrics, University Medical Centre Ljubljana, 3, Šlajmerjeva Street, Ljubljana 1000, Slovenia; Tel./Fax: +386(0)1-540-1137; E mail: borut.peterlin@guest.arnes.si

ABSTRACT

The combination of improving technologies for molecular interrogation of global molecular alterations in human diseases along with increases in computational capacity, have enabled unprecedented insight into disease etiology, pathogenesis and have enabled new possibilities for biomarker development. A large body of data has accumulated over recent years, with a most prominent increase in information originating from genomic, transcriptomic and proteomic profiling levels. However, the complexity of the data made discovery of high-order disease mechanisms involving various biological layers, difficult, and therefore required new approaches toward integration of such data into a complete representation of molecular events occurring on cellular level.

For this reason, we developed a new mode of integration of results coming from heterogeneous origins, using rank statistics of results from each profiling level. Due to the increased use of next-generation sequencing technology, experimental information is becoming increasingly more associated to sequence information, for which reason we have decided to synthesize the heterogeneous results using the information of their genomic position. We therefore propose a novel positional integratomic approach toward studying 'omic' information in human disease.

Keywords: Data integration, Genomics, Transcriptomics, High-throughput technologies

INTRODUCTION

The development of microarray technology in the last decade and the upsurge of next-generation sequencing in the last few years has provided an abundance of data originating from various biological levels, most prominently from genomic and transcriptomic levels [1,2]. Such novel approaches have contributed greatly towards our understanding of physiological cellular processes, as well as molecular changes that occur in human disease. The high-dimensional nature of data originating from these studies has also opened an array of new theoretical and statistical challenges that had to be faced in order to attain acceptable reproducibility and consistency of scientific results [3]. In particular, a large number of hypotheses tested in a single experiment produced a substantial amount of statistical noise, causing large numbers of false-positive detections and undue omission of many true-positive results. Although statistical methods have been developed to address these issues, difficulties in in-

Clinical Institute of Medical Genetics, Department of Obstetrics and Gynecology, University Medical Centre Ljubljana, 3 Šlajmerjeva Street, Ljubljana 1000, Slovenia

creasing specificity and sensitivity of highly parallel approaches still exist, with the greatest notoriety in the field of human diseases belonging to a group of common, complex disorders.

In an attempt to alleviate these drawbacks, we developed a method that harnesses the biological relations between data originating from studies investigating human disease on various biological levels. An example of such an approach may be illustrated by the fact that genomic alterations associated with human disease, *i.e.,* multiple sclerosis (MS), are usually investigated and interpreted separately from transcriptomic alterations occurring in MS. The biological relation between these two layers may thus be utilized to favor prioritization of genes that were detected on both layers, therefore reducing noisy results and facilitating detection of true biological data. We expect that with the inclusion of increasing the number of biological layers and increasing the number of studies in the database used for integration, the comprehensiveness and biological validity of prioritized genes would increase progressively.

MATERIALS AND METHODS

The pathway towards constructing the initial database used for subsequent integration is highly dependent on the disease of interest. While some common disorders have been investigated in several 'omic' studies that investigated several biological cellular levels, the sourcing data for other diseases may be more scarce. The search for data sources should be initiated by an overview of literature published to date. When the investigator is familiar with the studies performed, the published reports and their tables in supplemental materials may be used to extract the lists of genes or other genomic features with detected significant alterations.

A crucial step in obtaining data sources of high quality is inspection of available databases that are stored in public data repositories. These tend to be highly specialized for the biological layer of investigation. For genomic data from genome-wide association studies, data may be extracted from dbGAP (http://www.ncbi.nlm.nih.gov/gap), for epigenomic, transcriptomic and methylomic data, Array Express (http://www.ebi.ac.uk/arrayexpress/) or Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/ geo/), and for next generation sequencing databases European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ ena/) and Sequence Read Archive (SRA; http://www.ncbi. nlm.nih.gov/ sra) [4-7].

After all the sources have been investigated, a collected database of features [genes, mRNAs, microRNAs (miRNAs), CpG islands, proteins and others] with significant alterations in chosen disease should be prepared for each included study. We also advise collection of information, such as significance values and fold change values, on which prioritization of features for each biological layer will be performed in the later steps. If the latter information is not available, all the significant alterations in a given study will have the same importance in integration. In the following section, significant results from various study types will be collectively referred to as "signals" for reasons of clarity.

Data Integration. Before data can be integrated, they have to be reduced to a universal common denominator. Due to increasing heterogeneity of genetic information, tying biological information to gene-level annotation is becoming increasingly more difficult. Genomic variation and methylation patterns are two examples of information that is prohibitively difficult to associate with genes in any straightforward manner, as such alterations occur in genes, between genes or spread across several genes. For this reason, we opted for an integration based on the genomic position of features originating from various data sources. This required the signals from all databases to be converted to their genomic positions and projected on the genome assembly backbone. This step then allows for complete omission of difficult annotation conversion steps, required before final integration can be performed, greatly simplifying the synthesis of heterogeneous data.

After signals are positioned on the genomic backbone, the complete assembly is divided into bins of equal size. For each study, a score is given to each of the bins, depending on the score of alterations residing in that segment of the genome. After this step, the scores of all bins are prioritized and their rank scores calculated. The integration step is attained when the non parametric rank product for each of the bins is calculated, and on the basis of rank scores of bins originating from each data source, as we have previously described [8]. The lower final rank product signifies that higher ranks

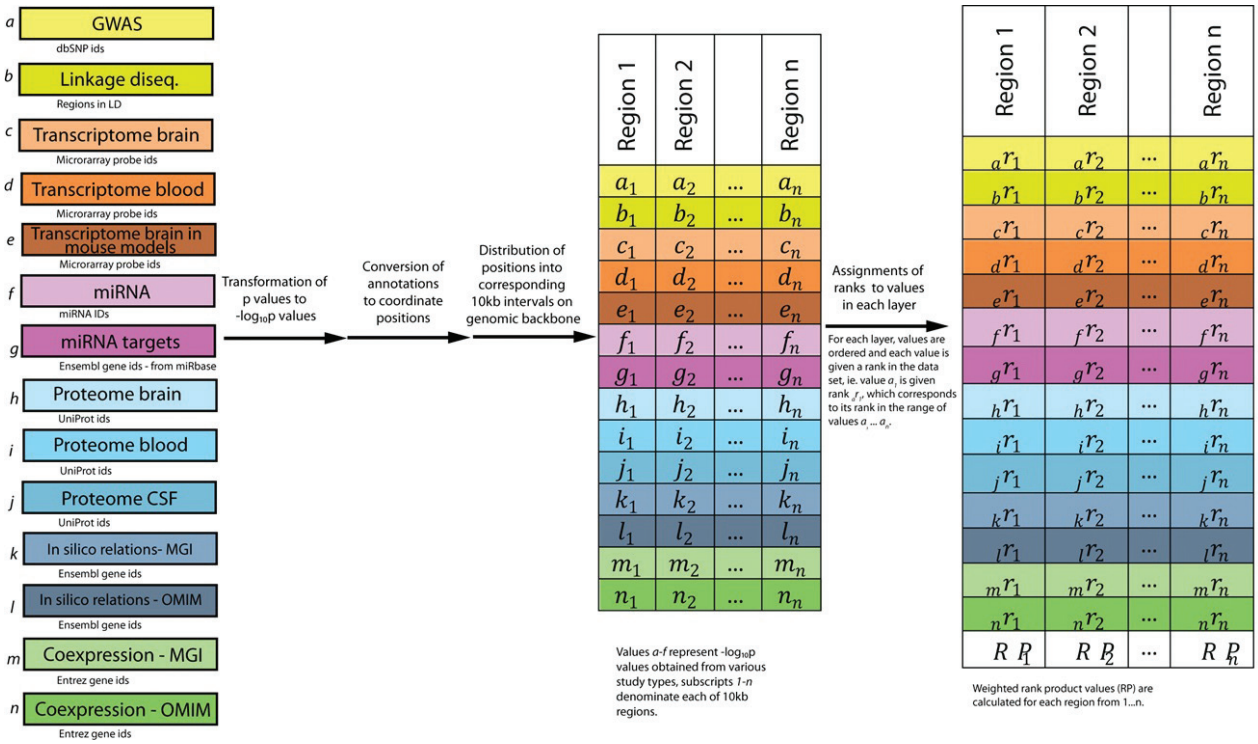

Figure 1. Process of integration of numerous heterogeneous data sources. First, data on significant alterations on a certain biological layer is obtained from selected studies (data from various layers is coded by letters a-n and differing colors). These alterations or signals are then positioned into genomic bins of fixed size and bin-scores for each of the bins is estimated. For each of the layers in a-n, bins are then prioritized on the basis of this score and the rank of each bin is separated. The final integration step is then performed by calculating rank products for each of the genomic bins, based on their rank in each of data sources.

were attained by bins on several separate biological layers [9]. Therefore, these bins represent genomic regions where accumulation of signals is detected on various biological levels, and thus represent regions of interest for further investigation. Ultimately, a permutational test may be employed to determine the significance of signal accumulation in each bin [8]. The detailed overview of the process may be observed in Figure 1.

RESULTS AND DISCUSSION

Results originating from the positional integratomic approach represent a prioritized list of genomic regions, where regions containing the greatest accumulation of heterogeneous biological alterations in an investigated disease rank highest and are characterized by lowest permutation test *p* values. As the integrative approach is performed for regions (bins) across the whole genome, the resulting genome-wide distribution of results from integration of data in human disease may be inspected. Genome-wide distribution of integration results for MS as an example of a complex autoimmune human disorder is represented in Figure 2. Here, the greatest accumulation of signals is observed on chromosome 6, specifically in the well-known human leukocyte antigen (HLA) region, suggesting that data from heterogeneous biological sources of 'omic' data indicate the role of this region in MS. Moreover, other regions have also attained high integration scores, suggesting importance of non-HLA regions in MS. Specifically, a region containing an interleukin-7 receptor gene (IL7R) attained very high integrative scores, not only on the basis of detections from genome-wide association studies, but also on the basis of evidence from expression profiling studies in blood and brain tissues. Additionally, the same region has been ranked high due to information obtained from various bioinformatic sources of data, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and co-expression information [10,11]. Such a heterogeneous body of evidence offers information of great relevance to true biological disease alterations and thus provides plausible candidate selection for further studies.

Figure 2. The genome-wide distribution of significance values, based on the permutation test of integration scores. Each region or genomic bin is represented by a dot whose height represent significances in the $-\log_{10}P$ form, with regions characterized by high accumulation of heterogeneous data attaining higher $-\log_{10}P$ values. The HLA region on chromosome 6 attained the highest score in these analyses with *p* values below $1 \bullet 10^{-9}$. Notably, non-HLA regions score high as well, offering a landscape of new genomic regions for further down-stream investigations

The positional approach offers great flexibility and control over parameters on which the final prioritization of genomic regions is based. Based on scientific questions, a researcher may be more interested in a contribution of only selected biological layers to the final integration score. For this reason, we have implemented means to allow custom weighting of different sources of data. For example, if one is interested in the relation between genomic variation and differential methylation, one may attribute those two sources greater weights and regions where signals from GWAS (genome-wide association studies), and global methylation studies aggregate will be obtained. Additional levels of control may also be obtained by customizing the size of genomic bins, allowing for detection of interactions that spread across larger genomic regions.

There has been great interest in deciphering the genetic factors with medium-to-low effect sizes as the explanation for the phenomenon of missing heritability in MS and other complex disorders [12,13]. Here, an integrative approach may assist in promoting detection of the genomic variant with its actual role in such complex disorder, and distinguishing them from spurious noise originating from statistical noise generated in genome-wide association studies. As large-scale studies, which attempt to detect low-effect susceptibility factors in human disease, have to be performed on large sample sizes, requiring great resources and effort [14], this approach may be a mode of comprehensive evidence-based selection of molecular determinants to investigate in such downstream validation studies.

With continuing development of high-throughput technologies, it is expected that the amount of the resulting data in large databases will continue to rise. For this reason, novel approaches for interpretation and understanding will also have to be prepared to face these challenges. As it is difficult for a single researcher or research group to have a comprehensive overview over such a vast information landscape, new means of presentation and access to these results will have to be envisaged. A position-based, integrative approach not only represents the means to quick insight into heterogeneous evidence from several large-scale studies, but is also a basis

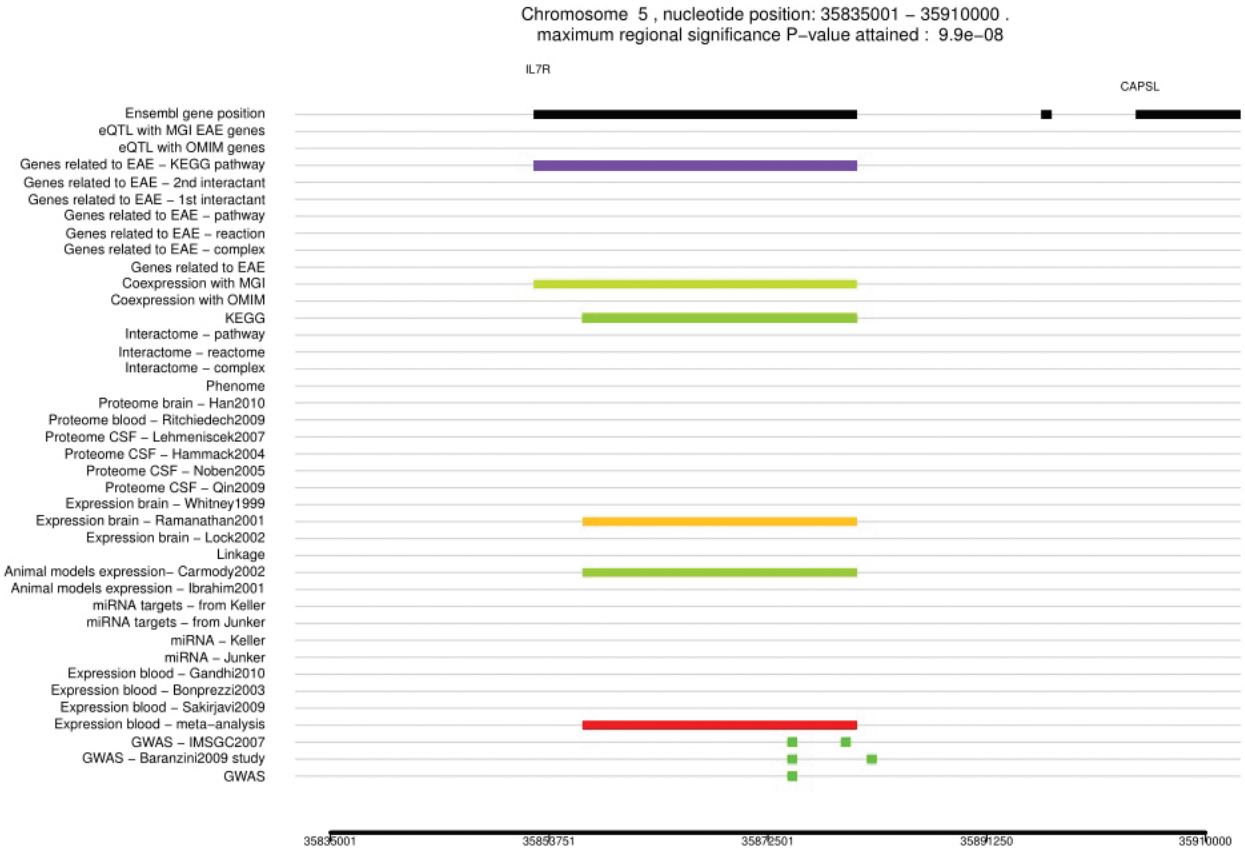

Figure 3. The supporting evidence substantiating the high score of the IL7R region on chromosome 5. Although the genetic variants in the IL7R gene have been detected in genome-wide association scans, there is also a substantial body of data supporting its relevance in MS. This support originates from whole-genome expression profiling studies in blood (red color) and brain (yellow color) as well as in expression profiling of brain samples from experimental autoimmune encephalitis animal models (dark green color). Additional in-silico relations, such as KEGG pathway relations and co-expression data support its relevance in MS.

toward the preparation of an interactive genome browser-like solutions for fast and easy access to this body of information.

REFERENCES

1. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006; 7(1): 55-65.
2. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008; 5(1): 16-18.
3. Shi L, Reid LH, Jones WD, *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24(9): 1151-1161.
4. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011; 39(Database issue): D19-D21.
5. Leinonen R, Akhtar R, Birney E, *et al.* The European nucleotide archive. Nucleic Acids Res. 2011; 39 (Database issue): D28-D31.
6. Parkinson H, Kapushesky M, Shojatalab M, *et al.* ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007; 35(Database issue): D747-D750.
7. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 2006; 411: 352-369.
8. Maver A, Peterlin B. Positional integratomic approach in identification of genomic candidate regions for Parkinson's disease. Bioinformatics. 2011; 27(14): 1971-1978.
9. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinform Comput Biol. 2005; 3(5): 1171-1189.
10. Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol. 2012; 802: 19-39.

11. Obayashi T, Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res. 2011; 39(Database issue): D1016-D1022.
12. Sawcer S, Hellenthal G, Pirinen M, *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476(7359?): 214-219.
13. Manolio TA, Collins FS, Cox NJ, *et al.* Finding the missing heritability of complex diseases. Nature. 2009; 461(7265): 747-753.
14. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005; 37(4): 413-417.