BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1971–1978
doi:10.1093/bioinformatics/btr313

Systems biology

Advance Access publication May 19, 2011

Positional integratomic approach in identification of genomic candidate regions for Parkinson's disease

Ales Maver and Borut Peterlin*

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

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Recent abundance of data from studies employing high-throughput technologies to reveal alterations in human disease on genomic, transcriptomic, proteomic and other levels, offer the possibility to integrate this information into a comprehensive picture of molecular events occurring in human disease. Diversity of data originating from these studies presents a methodological obstacle in the integration process, also due to difficulties in choosing the optimal unified denominator that would allow inclusion of variables from various types of studies. We present a novel approach for integration of such multi-origin data based on positions of genetic alterations occurring in human diseases. Parkinson's disease (PD) was chosen as a model for evaluation of our methodology.

Methods: Datasets from various types of studies in PD (linkage, genome-wide association, transcriptomic and proteomic studies) were obtained from online repositories or were extracted from available research papers. Subsequently, human genome assembly was subdivided into 10 kb regions, and significant signals from aforementioned studies were arranged into their corresponding regions according to their genomic position. For each region, rank product values were calculated and significance values were estimated by permuting the original dataset.

Results: Altogether, 179 regions (representing 33 contiguous genomic regions) had significant accumulation of signals when *P*-value cut-off was set at 0.0001. Identified regions with significant accumulation of signals contained 29 plausible candidate genes for PD. In conclusion, we present a novel approach for identification of candidate regions and genes for various human disorders, based on the positional integration of data across various types of omic studies.

Contact: ales.maver@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 16, 2011; revised on May 3, 2011; accepted on May 8, 2011

1 INTRODUCTION

It has been anticipated that development of highly parallel technologies would significantly improve our knowledge of several unresolved issues regarding the etiology of human disease, including details of its pathogenesis and the role of contributing risk factors. Consequently, several attempts at discerning alterations in genomic sequence, gene expression and protein profiles leading to onset of various diseases have been performed, employing high-throughput methodologies. However, results from these studies have still not clarified disease etiology and pathophysiology satisfactorily and have provided a multiplicity of results that are complex to interpret (Curtis *et al*., 2005). In addition, the inherent difficulty of these studies is frequent occurrence of false positive results due to multiple testing issues, usually small numbers of samples investigated and generally low technical and statistical reproducibility of obtained results (Khan *et al*., 1999; Kim and Park, 2004).

In order to facilitate discovery of biologically significant genetic alterations based on such data, an integrative analysis of publicly available datasets from studies that investigate global alterations in human disease on different biological layers—genome, transcriptome, proteome and phenome—is proposed. Integrative approach is based on the assumption that genes consistently associated with a disease on different biological layers in different large-scale study types are likely to reflect biologically relevant associations with disease investigated. In biological sense, alterations in genomic sequence can be reflected in changes of gene expression and consequently in alterations of proteomic profile that ultimately lead to observed phenotypes. The probability of a selected gene being associated to a disease as a consequence of a noisy, false-positive result would therefore be reduced by using integrative approach, as it is unlikely that a same false positive gene has been discovered simultaneously in different study types, if we assume that there is no bias toward selecting a certain group of false positive genes in all these studies.

Considerable difficulties may arise when combining different types of biological information (Cahan *et al*., 2007). Multiple attempts at integration of such heterogeneous data have already been reported in the literature and approaches to such prioritization of genes have been based either on candidate genes' functional similarity to a set of predefined training genes with known biological role in a disease under investigation (Aerts *et al*., 2006); or on experimental data from studies investigating specific alterations in the disease state (Rasche *et al*., 2008; Sun *et al*., 2009).

In the approach proposed by Aerts *et al*. (2006) and implemented in Endeavour software, the prioritization of candidate genes was performed by investigating their similarity to a training set of genes, based on multiple sources of data, including various annotation, expression, interaction, phenotype and other databases, followed by global ranking using order statistics. Rasche *et al*. have applied the integrative approach in the case of Type 2 diabetes mellitus, using

*To whom correspondence should be addressed.

empirical data from expression profiling studies in human tissue samples, mouse models and data from association studies in addition to information extracted from annotation and literature databases. In their approach, final prioritization was accomplished by scoring each gene, based on its fold change, empirical *P*-values, consistency across experimental replicas of the study, while also accounting for measure of entropy, signifying the extent of contribution of each study to final gene score (Rasche *et al*., 2008). Sun *et al.* integrated data from four sources: GWAS, linkage, expression and literate-based data to prioritize genes involved in schizophrenia. Here, a custom set of weighting matrices was applied to gene scores in order to control the contribution of each source of data in integration (Sun *et al*., 2009). In addition, several web-based utilities offering the possibility of custom integrative analyses are available for gene prioritization in various usage settings, i.e. GeneProspector, CANDID, ToppGene, Phenopred, SUSPECTS and several others (Tranchevent *et al*., 2009).

Previous methodologies utilizing such integration approaches were predominantly investigating which genes were overlapping in results from different types of included studies, and were thus utilizing gene-centric approach toward integration (Aerts *et al*., 2006; Middleton *et al*., 2007; Sun *et al*., 2009). Notably, a considerable fraction of meaningful genetic alterations are located outside genes, between adjacent genes, or spanning several neighboring genes: single nucleotide polymorphisms (SNPs), copy number variations (CNVs), regions defined by linkage disequilibrium studies as well as epigenetic alterations (i.e. regions with differential methylation patterns) and others. As it is often difficult to unambiguously map these genetic alterations to appropriate genes, a choice of an efficient common denominator for the integration process might be problematic. Also, difficulties due to incompatible reporting of results in publications and public repositories from various studies may present a notable problems not only when analyzing data from multiple study types, but also when using data from a single type of study, if performed on different manufacturer platforms and therefore using different probe identifiers for reporting. Although tools and relation databases enabling conversion of these identifiers exist, this process is often incomplete and results in significant loss of data.

To solve these issues, we performed a data integration based on the genomic locations of features investigated in included studies—a position-centric integration in contrast to previously employed gene-centric approaches. We have previously reported an initial step in this direction in our study on sarcoidosis, where the co-location of genes with differential mRNA/protein expression to genomic regions linked to sarcoidosis has been inspected (Maver *et al*., 2009).

To demonstrate the feasibility of our approach, Parkinson's disease (PD) has been used as a model disease for our approach. PD is the second most common neurodegenerative disorder, characterized by progressive depletion of dopaminergic neurons within the substantia nigra, clinically manifesting as progressive symptoms of tremor, rigidity, bradykinesia and postural instability (Mayeux, 2003). It is hypothesized that idiopathic forms of PD are a result of complex interplay between several genetic and environmental factors. The role of genetic factors in its pathogenesis has been substantiated through observations of heritable forms of PD resulting from mutations in at least 12 different genes (Wider *et al*., 2010), and observation of familial aggregation of idiopathic cases of PD in epidemiological studies (Steece-Collier *et al*., 2002). Etiology of PD remains largely unclear in most patients and to date there are no reliable available diagnostic assays for early detection, diagnosis and follow-up of the disease. In the recent decade, several research groups have utilized high-throughput technologies in search of a reliable genetic marker for PD and in effect, abundance and availability of data studying alterations on genomic, transcriptomic and proteomic levels exist in recent publications and publicly available repositories.

In this study, we aim to present a novel approach to integration of data from these studies, and perform integration-based search to find genomic regions and genes related to PD ethiopathogenesis.

2 METHODS

Initially, a search for studies performed in patients with PD was performed, using Medline database (www.ncbi.nlm.nih.gov/pubmed/) with search string ('Parkinson disease'[ti] OR 'Parkinson's disease'[ti]) AND (transcriptom* OR proteom* OR 'genome-wide' OR 'linkage scan' OR microarray OR profiling). Additionally, Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/), ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Stanford Microarray Database (http://smd.stanford.edu) were searched up to discover suitable sources of data for inclusion in our initial dataset. Data from six sources was ultimately included in the final integration dataset: brain transcriptome, blood transcriptome, brain proteome, genome-wide association studies (GWAS), genome-wide linkage studies and phenomic data. For clarity, all the included significant results from these types of studies [SNPs associated with PD, regions of genomic linkage disequilibrium in PD, differentially expressed (DE) mRNA and proteins in PD] from various types of studies included were designated as signals in all the subsequent sections.

All the data from these sources were stored and all analytical steps were performed in R statistical environment version 2.10.0 (http://www.r-project.org/) using Bioconductor version 2.6 packages (http://www.bioconductor.org/), unless stated otherwise.

2.1 Transcriptomic data

Raw data on transcriptomic alterations in central nervous system (CNS) and blood samples from patients with PD was obtained from GEO repository using the GEOquery package for R (Sean and Meltzer, 2007). Transcriptomic alterations in CNS and blood samples were treated as two separate datasets to account for possible differences in transcriptional alterations observed in these two tissue samples. Three GEO datasets with accession numbers GSE8397, GSE7621 and GSE7307 performed on CNS tissue samples and one dataset from analyses performed on whole blood sample—GEO accession number GSE6613, were included. Raw datasets were examined using arrayQualityMetrics package, followed by normalization and filtering with affyPLM and genefilter packages, where necessary.

A meta-analysis of transcriptomic data was performed prior to their inclusion of transcriptomic data in the integration process. Meta-analysis calculations were performed using algorithms in the RankProd package in the R environment, which enabled us to combine datasets from the three transcriptomic studies performed on CNS samples using RPadvance function (Hong *et al.,* 2006). Genes DE in a dataset from whole blood samples were analyzed using rank product (RP) function from the same package, using the 'data from single origin' option. False positive rate values calculated by RankProd for each probe set were used in further integration steps.

2.2 Proteomic data

We have inspected seven studies investigating proteomic alterations in PD brain samples (Abdi *et al*., 2006; Basso *et al*., 2004; Choi *et al*., 2004; Jin *et al*., 2006; Sinha *et al*., 2007; 2009; Werner *et al*., 2008). From each study, proteins reaching *P*-values of ≤ 0.05 were regarded as DE on proteomic level

and their coding genes included in our study. Altogether, 199 genes coding for DE proteins were collected from proteomic studies.

2.3 Linkage data

Linkage data were obtained from a study by Foltynie *et al*. (2005), who performed a genome-wide linkage disequilibrium screen by genotyping 5546 microsatellite markers in pooled samples from 374 patients with PD and two pools of 219 and 1490 control subjects without PD. Altogether, 214 microsatellite markers were found to be significantly associated with PD, and were included in the integration.

2.4 GWAS

Data from GWAS were obtained from Open Access Database of Genome-wide Association Results project data (Johnson and O'Donnell, 2009), that included complete dataset for SNPs reaching *P*-values of ≤ 0.05 from two GWAS in PD performed by Maraganore *et al*. (2005) on Perlegen 250K platform and Fung *et al*. (2006) on Illumina Infinium 100K platform. For the integration purposes 1604 SNPs reaching *P*-values of ≤ 0.05 were included in our panel of signals.

2.5 Phenomic data

Additional layer of integration was introduced by implementation of data from Phenotype ontology project (HPOP, http://www.human-phenotype-ontology.org). This resource contains relations between genes implicated in human monogenic disorders and phenotypic traits resulting from alterations in these genes (Robinson *et al*., 2008).

For this data source, an initial set of genes known to be related to PD was obtained from the Online Mendelian Inheritance in Man (OMIM) database (MIM identifier: 168600, McKusick, 2007). Afterwards, the phenotype identifiers related to these genes were obtained from HPOP's database. In the second step, we have investigated which other genes were also related to phenotypic traits of the initial gene set. This resulted in the set of genes with phenotypic consequences similar to genes in the training set. Genes with matching phenotypic traits were given score 1 for each phenotypic trait, therefore genes with more phenotypic similarities to initial PD gene set were given greater scores, enabling the prioritization according to phenotypic similarity to genes already related to PD.

2.6 Positional integration

Positional integration was performed by mapping positive signals from every type of study on a reference genome coordinate backbone. The coordinate system and length of each chromosome was defined according to NCBI reference assembly Build 36.3 (March 2008) (Fig. 1).

Genes and SNPs from proteomic, phenomic and GWAS were converted to nucleotide positions using BioMart batch mining tool (http://www.biomart.org, Smedley *et al*., 2009). Locations of probesets from transcriptomic studies and markers from linkage study were obtained from University of California Santa Cruz (UCSC) Genome browser site (http://genome.ucsc.edu). Difficulties due to incompatible builds of reference assemblies were resolved using liftOver tool from USCS Genome browser (http://hgdownload.cse.ucsc.edu/downloads.html).

Reference genome assembly was subdivided into 616 108 regions, each 10 000 bp long and with 5000 bp overlap with the neighboring region. This overlap was introduced in order to prevent losing information of signal aggregation at borders of defined regions. We have prepared an R script that automatically positioned significant signals from various types of studies into their corresponding location intervals in the genome. The result was a matrix of 616 108 lines and six columns, each column corresponding to results from transcriptomic—brain, transcriptomic—blood, proteomic, linkage, GWAS and phenomic data. For each type of study, a significant signal was noted as the specific value of $-\log 10p$ or 1 where *P*-values were not available. When multiple signals from a single type of study were located in the same region (i.e. several neighboring SNPs), the values were summarized, to increase the score of this region. When no significant signal in the interval investigated was observed, this was marked with value of 0. Values in every column were then ordered and scored from 1 (representing the most significant signal in the study) to 616 108 (representing the least significant signal in the study). In the case of two or more regions with equal score values, the lowest rank was used for further integration steps.

A rank statistics approach was utilized to combine the results originating from differing sources of data. Specifically, the RP approach, described by Breitling *et al.* (2004) was implemented and RP values were calculated according to Equation (1). In the subsequent equations, symbols used have the following significance: R is the consecutive number of each 10 kb interval, i represent each study type used in integration, $r_{i,R}$ is the rank of R-th interval in the vector of data from study i and n_i signifies number of all intervals and RP_R represents calculated rank product value for R-th interval.

$$\mathrm{RP}_R = \prod_{i=1}^{k} \left(\frac{r_{i,R}}{n_i} \right) \tag{1}$$

To estimate significance values for each region, RP values were permuted relative to regions 1000 times and expected RP estimates were compared with those from our original dataset, using Equation (2). Here the $\mathrm{rank}(\mathrm{RP}_{\mathrm{perm}})$ signifies rank of an interval in the matrix of permuted values, while the N_{perm} represents number of permutation cycles used.

$$\mathrm{RP}_{\mathrm{expected}} \approx \frac{\mathrm{rank}(\mathrm{RP}_{\mathrm{perm}})}{N_{\mathrm{perm}}} \tag{2}$$

The possibility that greater accumulation of signals occurred in region with greater density of genes was taken into consideration by permuting regions containing equal numbers of genes in separate permutation blocks. The complete set of genes required for gene density estimation was obtained from Ensembl genome assembly version 54.

Finally, false positive pre pfp (q) values were calculated as proposed in Breitling *et al*. (2004). $\mathrm{RP}_{\mathrm{expected}}$ signifies expected RP value of R-th interval after permutation, $\mathrm{rank}(R)$ signifies rank of R-th interval in the original non-permuted dataset

$$q_R = \frac{\mathrm{RP}_{\mathrm{expected}}}{\mathrm{rank}(R)} \tag{3}$$

As it is difficult to predict the importance of each single dataset in final integration step, we attempted to calibrate optimum weighting matrix that would be applied to datasets from included studies before the integration step. For this approach, we obtained three lists of training genes. First was the set of genes annotated in the KEGG database PD entry (KEGG Pathway accession number: hsa05012, Ogata *et al*., 1999), the second training set consisted of genes implicated in the monogenic forms of PD according to OMIM database (http://www.ncbi.nlm.nih.gov/omim), and the third set was obtained from the PDgene database (Lill *et al*., 2011; http://www.pdgene.org/), selecting only genes that contain at least one variant showing a nominally significant summary OR in the meta-analyses of all studies included in the PDgene database. Each type of study was given one of the three possible weights according to their importance in the integration process (1—lower importance, 2—medium importance and 3—high importance). Using three varying weights for six datasets resulted in different variations of weight factor matrices $[(p)V_3^6 = 3^6 = 729]$. For each of the weight matrices applied to original dataset significance values were calculated using a modified equation for weighted RP calculation and previously described permutation approach (Equation 4). Here, w_i signifies the weight parameter assigned to study i.

$$\mathrm{RP}_{\mathrm{R}} = \left(\prod_{i=1}^{k} \left(\frac{r_{i,R}}{n_i} \right)^{w_i} \right)^{1/\sum_{i=1}^{k} w_i} \tag{4}$$

Significance values of regions containing training genes were calculated for each weight matrix variation. Weight matrix resulting in best performance of regions containing training genes was chosen as the optimal arrangement of weights for integration approach. This optimized weight configuration has been applied to our complete dataset before final integration and the results compared with results of non-modified dataset.

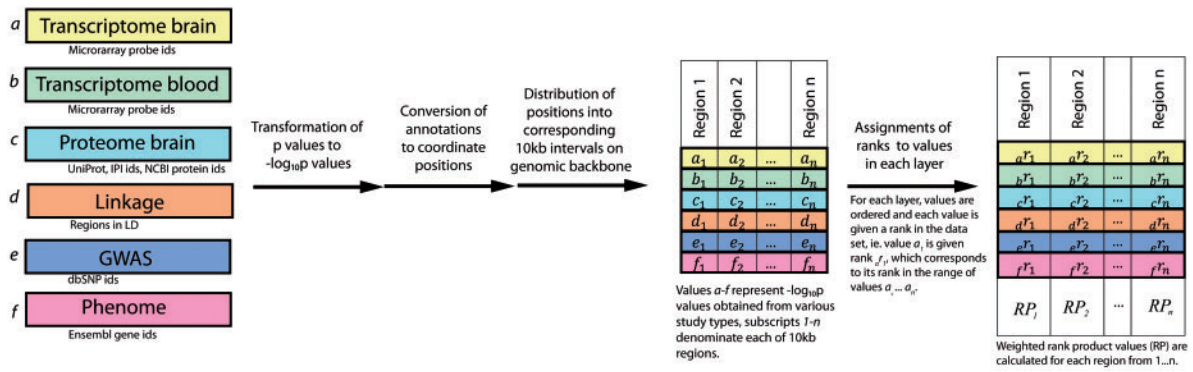

Fig. 1. Overview of steps in positional integration.

2.7 Evaluation

Evaluation was performed by searching for a direct association of genes located in top regions selected by the integration process and PD in the Medline database (www.ncbi.nlm.nih.gov/pubmed). The search was performed on articles that appeared in Medline up to July 2010 using the following search string: 'Parkinson disease AND Gene', where 'Gene' entry represented candidate genes located in the regions discovered by the integration process. In order to also include indirect associations between genes in top regions and PD, the Bitola data mining tool was utilized (http://www.mf.uni-lj.si/bitola, Hristovski *et al*., 2005). Bitola's closed discovery tool was used with the concept X set as PD (CUI:C0030567) and concept Z was chosen from genes located in the top regions. The results found by Bitola represent concepts that are related to X and Z at the same time and therefore reflect possible, indirect, relations between concepts X and Z, and were therefore used to study indirect relation of PD and genes located in the top regions.

Additionally functional profiles of genes located in the set of top region have been profiled using Gene Ontology (GO, http://www.geneontology.org, Ashburner *et al*., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/, Ogata *et al*., 1999) and Reactome (http://www.reactome.org, Joshi-Tope *et al*., 2005) pathway analysis. Hypergeometric test implemented in GOstats for R (available from http://www.bioconductor.org, Falcon and Gentleman, 2007) package has been utilized for this purpose. A complete set of genes obtained from Ensembl database (http://www.ensembl.org) was set as the gene universe. *P*-values for over-representation of functional terms were adjusted according to Benjamini–Hochberg method (Benjamini and Hochberg, 1995). Reactome Skypainter utility (http://www.reactome.org/cgi-bin/skypainter2) was used to determine overrepresentation of pathways/reactions in the final set of genes.

For comparison with other utilities available for gene prioritization, we performed analyses using Endeavour software (available at: http://homes.esat.kuleuven.be/~bioiuser/endeavour/index.php, Aerts *et al*., 2006) and ToppGene software (available at: http://toppgene.cchmc.org/, Chen *et al*., 2009). For Endeavour, a method for genome-wide scoring was selected in the Java server-based version of the software and a list of 12 PD-related genes were selected from OMIM database as a training set. All the available sources offered by Endeavour were selected for gene prioritization and other parameters were left in their default state. Similarly, for ToppGene software, a list of 12 OMIM PD-related genes were selected as a training set, complete set of Ensembl genes were selected for prediction and all available sources were included in the similarity analyses.

3 RESULTS

Data included in the integration comprised 4700 DE genes obtained by performing meta-analysis of three transcriptional studies on CNS samples, 1731 DE genes obtained by analyzing raw data from a transcriptional study performed on whole blood samples, 199 DE proteins from 7 proteomic studies, 214 regions from a linkage study in PD, 1604 SNPs from GWAS and 1253 genes with phenotypic relationship to PD.

Initially, the distribution of signals in the genome was examined. Of the complete set of 616 108 genomic regions investigated in the integration, 476 810 (77.4%) did not contain any signals from included studies, 121 011 (19.6%) contained signal originating from one study type, 16 214 (2.6%) contained signals originating from two different study types, 1969 (0.32%) contained signals from three types of studies, 103 (0.017%) regions contained signals from four study types and 1 (0.00016%) region exhibited aggregation of signals from five types of studies. Aggregation of signals from all six study types was not observed in any of the regions investigated. Of the top 10 kb regions with greatest accumulation of signals (those containing signals originating from at least four different types of studies), most were located on chromosomes 17 (18 regions), 9 (17 regions), 15 (14 regions), 18 (9 regions) and chromosomes 4 and X, each containing 7 regions. Among these, several regions were found to be adjacent to form regional blocks >10 kb, resulting from aggregation of signals in contiguous blocks of neighboring 10 kb intervals. After these pairs of adjoining regions were merged into one, there were altogether 33 contiguous genomic regions with aggregation of signals from more than four types of studies.

Before performing rank product statistics calculations, the optimal weight configuration was estimated for each type of study included (Fig. 2). When KEGG genes were used as a training set, the greatest performance was obtained when transcriptome brain, proteome and phenome layers were given higher weights and transcriptome blood and GWAS study types were given lower weights. When genes implicated in monogenic forms of PD and genes significant in meta-analyses of PD association studies were used to train best weights, optimal results were obtained with phenome layer given higher weights and other layers were given neutral or lower weights. For further analyses, optimal weight configuration based on the training set of KEGG pathway genes was used.

After selection of KEGG pathway-based weighting matrix, significance values for each interval were estimated by calculation of weighted RP values and performing 1000 permutations. Genome-wide distribution of obtained significance values is presented in

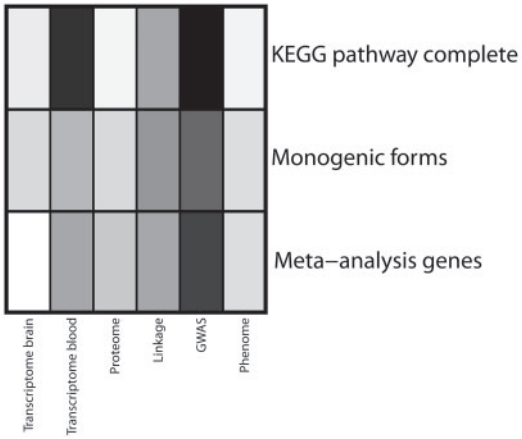

Fig. 2. Estimation of optimal weight matrices using three different training gene sets. The heatplot represents which layers were weighted more (lighter) and which were weighted less (darker) in order to maximize the final rank values for regions containing genes in the three training sets.

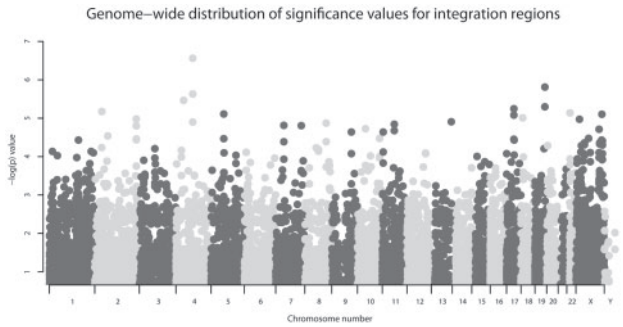

Fig. 3. Genome-wide plot displaying the distribution of calculated *P*-values across the genome. *X*-axis represents locations of the region on genomic backbone and *Y*-axis represents −log10p estimates of *P*-values obtained by permutation.

Figure 3. Altogether, 179 (0.029%) regions attained significance values ≤ 0.0001, comprising 33 discrete regional blocks after neighboring intervals were merged. Altogether, these regions contained 29 Ensembl genes. Peaks of highest significance were observed UCHL1 and SNCA gene regions on chromosome 4, GFAP gene region on chromosome 17, APOE gene region on chromosome 19, in addition to novel regions on chromosomes 7, 11 and 20. When applying a less restrictive cutoff of *P*-values ≤ 0.05, altogether 23 647 (3.84%) regions reached significance. This accounted for 2748 contiguous regions, containing 2183 genes. Calculation of pfp (q) values revealed 10 regions attaining pfp values < 0.05, containing 14 genes.

Among the regions with best significance scores, several contained genes with already established role in PD through association studies or genes that have been implicated in genetic forms of PD, such as SNCA and UCHL1. For demonstration, the detailed view of supporting evidence for inclusion of SNCA gene region is presented in Figure 4a.

On the other hand, several novel genetic candidates were revealed after inspecting the list of top regions. For presentation, YWHAE gene was selected, being located in one of the top regions, and supported by evidence from at least four originating omic studies. The supporting inclusion evidence for YWHAE gene region is presented in Figure 4b. Further exploration of data and inspection of interpositions of results from source studies may be performed on our project's web site, which is available at http://integratomics.dyndns-remote.com/index.php.

3.1 Evaluation

Search for direct associations of genes in regions with highest significance values ($P \leq 0.0001$) with PD has uncovered that of 29 genes located in these regions, 15 genes were previously directly associated with PD (51.7%), of which 7 genes were co-occuring with PD in at least 10 Medline entries (MAPT, BAX, APOE, GFAP, SNCA, PRNP and UCHL1).

Search for the indirect associations using the Bitola software has revealed that several remaining novel genes were related to PD indirectly: 11 genes through their role in neurodegeneration processes, 3 genes through their role in regulation of neuronal function, 1 gene through involvement in regulation of apoptosis in neuronal cells and 4 genes through their association with various other disorders of CNS.

A selection of overrepresented GO terms associated with the set of genes located in top integratomic regions are presented in Figure 5 (presented are terms with hypergeometric test *P*-values < 0.01). The largest proportion of genes from integration was associated with GO terms related to specific neuronal terms (most prominently, axonal function), metabolism of small molecules, lipid/sterol metabolism and apoptosis. It has also been inspected whether genes in regions with evidence originating from greater number of different study types would result in enrichment of terms related to PD. It can be noted in Figure 5 that with progressive inclusion of more study types, a greater proportion of genes in the resulting set were related to GO annotations generally associated with PD: mitochondrial function, lipid/cholesterol metabolism, metabolism of small molecules and neural development.

The Reactome enrichment analysis has revealed that resulting set of genes are involved in pathways of Membrane trafficking ($P = 0.011$), Lipid digestion, mobilization and transport (=0.0024) and Synaptic transmission ($P = 0.044$).

Calculations in Endeavour software using 12 training genes from OMIM database have shown 111 new genes to have prioritization *P*-values < 0.0001. We compared the list of top 1000 genes provided by Endeavour and list of top 1000 genes in highest ranked regions obtained by our integrative approach. Overall, both lists of top-ranked genes shared 176 (17.6%) genes (hypergeometric test *P*-value $< 1 \times 10^{-16}$). Analysis with ToppGene gene prioritization suite has ranked 199 genes with similarity *P*-values < 0.0001. Even higher concordance between top 1000 list of genes from ToppGene and our approach was observed (overall 225 genes matching in both lists, hypergeometric test *P*-value for enrichment $< 1 \times 10^{-41}$). The coherence of genes selected by Endeavour and ToppGene and genes located in regions prioritized by our approach may be observed in Supplementary Figure 1.

4 DISCUSSION

In our study, we have performed a positional integratomic search for candidate genomic regions and candidate genes, based on

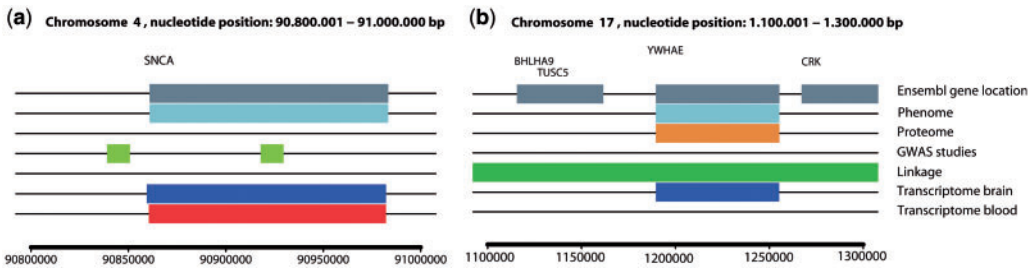

Fig. 4. (**a**) Evidence substantiating placement of known SNCA region on chromosome 4 among the top regions. Plot represents detailed view of the interposition of signals originating from several types of studies included. (**b**) Evidence substantiating placement of novel YWHAE region on chromosome 17 among the top regions. Plot represents detailed view of the interposition of signals originating from several types of studies included.

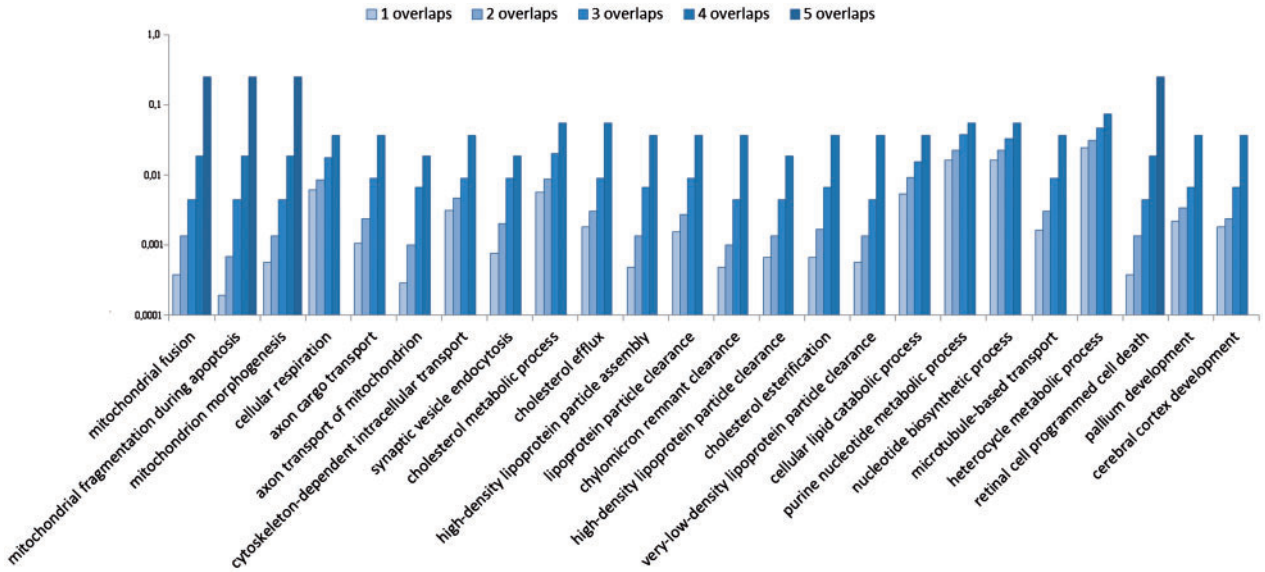

Fig. 5. Profile of GO term enrichment of genes located in the top regions. Increasing proportion of genes related to processes of mitochondrial function, cholesterol and lipid metabolism and nervous system function/development is apparent as genes in regions with greater number of overlaps are included in enrichment analyses.

the integration of data-originating studies that investigated global alterations in PD on different biological layers. From large-scale studies in PD, 179 regions were selected by demonstrating significantly increased aggregation of signals across at least four different biological layers of data from large-scale studies in PD.

Recent technological advances have enabled identification of molecular alterations occurring in human diseases on a global scale. However, these approaches are intrinsically burdened by the high dimensionality of data leading to multiple testing issues, usually small numbers of samples investigated and generally low technical and statistical reproducibility of obtained results (Khan *et al*., 1999; Kim and Park, 2004). We therefore hypothesized that an integrated approach, involving several layers of biological information interconnected by flow of information defined by central dogma of molecular biology could provide additional insight into events occurring in complex diseases. It is not uncommon that intermediate and minor alterations found in various large-scale studies are disregarded in further studies as they are interpreted as non-specific noise in the data. However, a proportion of these alterations could nevertheless represent biologically relevant information. Consistent observation of such alterations in various types of large-scale studies may suggest a biologically important signal, although they may only reach intermediate or minor significance levels in separate studies. Therefore, this additional level of complexity may provide a more complete insight into disease pathogenesis and assist in searching for optimal disease biomarkers for PD and other complex disorders.

Integration of data from various sources has previously been employed to select for genes with consistently highest significance scores from various different studies (Aerts *et al*., 2006; Sun *et al*., 2009). Previous integration studies have utilized gene-centric approaches, combining information from various types of studies using gene names or accessions as a common denominator. However, several issues may arise when using gene-centric approach to integration of multi-sourced data.

First, difficulties may arise due to inconsistencies in gene annotation used for reporting of results in various types of large-scale studies. Results of these studies are often provided using differing

annotations for reporting significant outcomes, i.e. using microarray probe identifiers, gene symbols, gene names and other annotations. Conversion of these annotations to a common gene identifier is often imperfect. This issue has been well recognized previously in meta-analyses of microarray expression data (Cahan *et al*., 2007). In our preliminary assessment of input data, we have found that 306 (7.5%) of Affymetrix probe identifiers showing significant gene expression differences in brain of patients with PD could not be converted to HGNC (Human Gene Nomenclature Committee) gene annotations using the BioMart utility (http://www.biomart.org/). Even greater difficulties were encountered in conversion of identifiers for significantly associated with PD to HGNC gene names, where only 50.6% of SNP identifiers could be converted to HGNC gene names. It is expected that with inclusion of greater number of studies differing by design, an increasing proportion of meaningful information is lost in the integration process due to these annotation disparities.

Second, it is well established that genetic changes located outside gene's coding region may contribute to disease susceptibility. Various forms of genetic changes are located in adjacent gene regulatory regions several kilobases upstream or downstream and influence gene expression and/or function (Kleinjan and van Heyningen, 2005). These long range effects have been shown extend to 1 Mb in either direction from the gene and may reside in regions occupied by other genes (Kimura-Yoshida *et al*., 2004; Kleinjan *et al*., 2001). Such long-range interactions would be overlooked in gene-centric integrations approaches, i.e. a SNP located farther upstream from a gene would not be associated with a DE gene, and thus this interaction would be omitted from further analyses. As the role of these interactions may be an important layer of complexity in multifactorial diseases, this is a significant issue in the gene-centric integration approaches (Steidl *et al*., 2007).

To address these and other issues occurring in integrative approaches, we based the integration of data from various types of studies on positions of genetic changes associated with PD. Difficulties due to poor conversion of annotations were resolved by converting annotations to their positions on genome coordinates. Where direct conversion was not available, we have used BLAST services to find corresponding genomic positions of feature, based on sequences of probes by microarray manufacturers. Also, this approach takes into account interactions between nearby genetic changes and is not limited by types of genetic changes to be included in the integration process, being flexible enough to allow inclusion of anticipated data from studies investigating epigenetic modifications, CNV, microRNA alterations in human disease. Since it would be statistically problematic to directly merge significance values from studies reporting alterations in PD on different biological layers, especially due to significant differences in methodological and statistical designs, we have used a prioritization approach to select best regions that contained most significant values on the greatest number of biological layers included. To evaluate genes in regions singled-out by integration process, we investigated existent literature for direct and indirect associations of these genes with PD. We have found that integration process revealed genes from which a notable proportion was previously studied in relation with PD (51.7%). Of these several are previously well-investigated PD genes, found to have a role in monogenic forms of PD, such as UCHL1 and SNCA.

On the other hand, several genes located in top regions have not previously been investigated in PD. A large fraction of remaining genes could be indirectly associated with PD with further data-mining approaches. As an example, YWHAE gene, presented in the results section, has been indirectly associated with PD in only one study, according to the Medline database (using search string 'Parkinson's disease AND YWHAE'), while multiple lines of evidence from large-scale studies suggests its role in PD (altered mRNA expression in brain, altered protein levels in samples from PD, location in region linked to PD in genomic scans and phenotypic compatibility with PD). According to data from GO, it is involved in the processes of neuronal migration, CNS development, intracellular signaling and regulation of apoptosis, which fits our current view of PD pathogenesis (Gandhi and Wood, 2005). In addition, it has recently been found to potentially interact with Parkin protein, an ubiquitin-protein ligase (E3), mutations of which cause juvenile onset—autosomal recessive PD (Davison *et al*., 2009).

Other genes located in the top regions have functional profiles compatible with PD pathogenetic processes, and belong to GO categories of nerve–nerve synaptic transmission, mitochondrion transport along microtubule, CNS development, amyloid precursor protein metabolic process, myelination among others and are therefore feasible genes to include in the further studies of PD pathogenesis.

It is necessary to point out that possible difficulties are inherent to position-centric integration approach. Choice of sizes of regions used for integration is not straightforward and by choosing regions too small, may result in missing important long-range interactions, while choosing to large regions may result in high amount of false positive genes. We chose to study regions 10 kb in size, as this reduced a number of genes observed to an average of 0.5 gene per each region; therefore, easing the process of supervised selection of the genes where this aggregation occurs in the selected region. On the other hand, possibility of missing interactions occurring at lengths farther than 10 kb exists in this case.

As noted in Section 1, several types of general purpose software are available that enable performing integrative gene prioritization and calculate ranking of candidate genes based on similarities to a predefined training set of genes instead of using disease-specific experimental results. For comparison, we investigated coherence of output from available tools with the results of our integrative analysis. Notably, we performed analyses using Endeavour and ToppGene software, selecting OMIM genes related to monogenic forms of PD as an initial training set. While a notable proportion of genes obtained by these gene prioritization approaches calculation were compatible with our list of top genes (list of top 1000 genes obtained by our approach and that obtained by Endeavour or ToppGene software shared 17.6 and 22.5% of genes, respectively), there are also differences between our and Endeavour/ToppGene ranked lists (Supplementary Fig. 1). This divergence is primarily a consequence of inclusion of different sources of data—our results were based on disease-specific experimental data and Endeavour/ToppGene prioritization is based on annotational and functional similarities to the training set of genes. Also, Endeavour and similar approaches base prioritization on a smaller set of genes with known relation to a disease under investigation, potentially conferring some limitation to discovery of entirely new causative genes and disease mechanisms. It is our view that results from both approaches should be viewed as complementary utilities in integrative gene prioritization.

5 CONCLUSIONS

A novel integrative approach to selection of candidate regions and genes involved in human disease is presented. In contrast to previous methodologies employed for integration, this approach is based on genomic positions of alterations in human disease (position-centric), circumventing several issues met in previous gene-centric integrative studies.

Here, this approach was used for discovery of putative regions and genes related to PD, using a comprehensive and unique set of data sources (GWAS, linkage studies, expression studies and proteomic studies). Of note is an additional source of phenomic information, obtained by inclusion of data from HPOP and utilizing the information on phenotypic changes in monogenic disorders in searching of genes for complex disorders.

In the near future, we expect the availability of data on PD to grow further. This new information will be straightforwardly added to our current dataset and further improve comprehensiveness of the search for putative genomic regions and genes related to PD. While we have demonstrated the use of this approach in PD, large availability of data reported for other common diseases enables application of our strategy to other complex diseases or genetic diseases with undetermined genetic causes.

The data from analysis in our study are accessible for further exploration and inspection of content of ranked regions on the web address: http://integratomics.dyndns-remote.com/index.php. As further sources of experimental data are added to the initial set of studies, the prioritization of regions will be updated accordingly and the results accessible on this site.

Funding: The study was supported by the Slovenian Research Agency grant No. J3-2377.

Conflict of Interest: none declared.

REFERENCES

Abdi,F. *et al.* (2006) Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *J. Alzheimers Dis.*, **9**, 293–348.

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Basso,M. *et al.* (2004) Proteome analysis of human substantia nigra in Parkinson's disease. *Proteomics*, **4**, 3943–3952.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Cahan,P. *et al.* (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, **401**, 12–18.

Chen,J. *et al.* (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.

Choi,J. *et al.* (2004) Oxidative modifications and down-regulation of ubiquitin carboxyl-terminal hydrolase L1 associated with idiopathic Parkinson's and Alzheimer's diseases. *J. Biol. Chem.*, **279**, 13256–13264.

Curtis,R.K. *et al.* (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Davison,E.J. *et al.* (2009) Proteomic analysis of increased Parkin expression and its interactants provides evidence for a role in modulation of mitochondrial function. *Proteomics*, **9**, 4284–4297.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Foltynie,T. *et al.* (2005) A genome wide linkage disequilibrium screen in Parkinson's disease. *J. Neurol.*, **252**, 597–602.

Fung,H.C. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.

Gandhi,S. and Wood,N.W. (2005) Molecular pathogenesis of Parkinson's disease. *Hum. Mol. Genet.*, **14**, 2749–2755.

Hong,F. *et al.* (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825–2827.

Hristovski,D. *et al.* (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, **74**, 289–298.

Jin,J. *et al.* (2006) Proteomic identification of a stress protein, mortalin/mthsp70/GRP75: relevance to Parkinson disease. *Mol. Cell Proteomics*, **5**, 1193–1204.

Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.

Joshi-Tope,G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.

Khan,J. *et al.* (1999) DNA microarray technology: the anticipated impact on the study of human disease. *Biochim. Biophys. Acta*, **1423**, M17–M28.

Kim, R.D. and Park, P.J. (2004) Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol.*, **5**, R70.

Kimura-Yoshida,C. *et al.* (2004) Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, **131**, 57–71.

Kleinjan,D.A. *et al.* (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.*, **10**, 2049–2059.

Kleinjan,D.A. and van Heyningen,V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.*, **76**, 8–32.

Lill,C.M. *et al.* (2011) The PDGene Database. Alzheimer Research Forum. Available at: http://www.pdgene.org/ (last accessed date April 27, 2011).

Maraganore,D.M. *et al.* (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.

Maver,A. *et al.* (2009) Search for sarcoidosis candidate genes by integration of data from genomic, transcriptomic and proteomic studies. *Med. Sci. Monit.*, **15**, SR22–SR28.

Mayeux,R. (2003) Epidemiology of neurodegeneration. *Annu. Rev. Neurosci.*, **26**, 81–104.

McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Middleton,F.A. *et al.* (2007) Integrating genetic, functional genomic, and bioinformatics data in a systems biology approach to complex diseases: application to schizophrenia. *Methods Mol. Biol.*, **401**, 337–364.

Ogata,H. *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.

Rasche,A. *et al.* (2008) Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics*, **9**, 310.

Robinson,P.N. *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Sean,D. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

Sinha,A. *et al.* (2007) Blood proteome profiling in case controls and Parkinson's disease patients in Indian population. *Clin. Chim. Acta*, **380**, 232–234.

Sinha,A. *et al.* (2009) Identification of differentially displayed proteins in cerebrospinal fluid of Parkinson's disease patients: a proteomic approach. *Clin. Chim. Acta*, **400**, 14–20.

Smedley,D. *et al.* (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Steece-Collier,K. *et al.* (2002) Etiology of Parkinson's disease: genetics and environment revisited. *Proc. Natl Acad. Sci. USA*, **99**, 13972–13974.

Steidl,U. *et al.* (2007) A distal single nucleotide polymorphism alters long–range regulation of the PU.1 gene in acute myeloid leukemia. *J. Clin. Invest.*, **117**, 2611–2620.

Sun,J. *et al.* (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics*, **25**, 2595–6602.

Tranchevent,L.C. *et al.* (2010) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, **12**, 22–32.

Werner,C.J. *et al.* (2008) Proteome analysis of human substantia nigra in Parkinson's disease. *Proteome Sci.*, **6**, 8.

Wider,C. *et al.* (2010) Genetics of Parkinson disease and essential tremor. *Curr. Opin. Neurol.*, **23**, 388–393.