Mining genetic, transcriptomic and imaging data in Parkinson's disease

Guglielmo Cerri
Department of Computer Science
University of Verona
Verona, Italy, IT
guglielmo.cerri@univr.it

Manuel Tognon
Department of Computer Science
University of Verona
Verona, Italy, IT
manuel.tognon@univr.it

Andre Altmann
Centre for Medical Image Computing (CMIC)
University of London
London, UK
a.altmann@ucl.ac.uk

Rosalba Giugno
Department of Computer Science
University of Verona
Verona, IT
rosalba.giugno@univr.it

***Abstract*— Parkinson's Disease (PD) is one of the most common and diffused neurodegenerative diseases, with a prevalence of 0.3% in the general population of industrialized countries and ~1% in subjects over the age of 60 [1]. PD is associated with both genetic and neuroimaging factors. Despite the factors causing PD are still not completely clear, during the last decade there has been many significant advances in early clinical diagnosis, via brain imaging, and in our understanding of the genetics of Parkinson's disease. *Imaging genetics* is an emerging longitudinal research field bridging genetic insights into the biology of complex diseases with quantitative neuroimaging phenotypes. *Imaging genetics* primarily focuses on identifying and characterizing how genes and genetic variation influence neuroanatomical and neurophysiological traits exploiting brain images. Until now have been proposed several *imaging genetics* methods to improve our knowledge on different complex neurodegenerative disease, such as Alzheimer's disease, but only few studies focused on PD. Using PD as a case study, here we exploit the advantages of existing methods to analyze heterogeneous datasets. To do so, we designed and propose a multi-view *imaging genetics* workflow interpolating genotyping, transcriptomic data, and functional and morphological brain images. We show how to process and interpret data to retrieve several potential genetic variants, which could constitute potential genetic biomarkers of PD onset and progression. The method consists of three steps. Each phase explores different aspects and uses different tools to handle the different data type. We will move from the classical quality control phase, which corrects potential problems in the dataset, to the more technical and core part of our method, consisting of principal component analysis (PCA), Genome Wide Association Study (GWAS), and results merge, validation and interpretation. Each step carries out the analysis, by using different state-of-the-art tools and scripts written in various programming language like R, Python and Bash.**

Index Terms*—*GWAS, neurodegenerative disease, Parkinson's disease, SNPs, imaging genetics, PLINK, Python, R

EXTENDED ABSTRACT

I. INTRODUCTION

Parkinson's disease (PD) is a brain disorder that leads to shaking, stiffness and difficulties with walking, balance, and coordination. Affected people may also show mental and behavioral changes, such as sleep problems, depression, memory loss or fatigue. PD is an age-related disease, with an increased prevalence in populations of subjects over the age of 60. About 5 to 10% of PD patients have an "early-onset" variant and it is often, but not always, inherited. PD is characterized by the loss of groups of neurons involved in the control of voluntary movements. Here we present a novel i*maging genetics* workflow on Parkinson's Disease, focused on discovering new potential candidate genetic biomarkers for Parkinson's disease onset, through a longitudinal study involving genotyping, transcriptomic, functional (Dopamine Transporter Scan) and morphological (Magnetic Resonance Imaging) imaging data [2—4].

The proposed tutorial has the aim to encourage and stimulate the attendees on the biomedical research with the advantage of integration of heterogenous data. In the last decade the use of images together with genetics data has become widespread among the bioinformatics researchers. This allowed the detailed investigation of several complex diseases, significantly improving our knowledge about their origins and causes. While in recent years many imaging genetics analyses have been developed and successfully applied to characterize brain functioning and complex neurodegenerative diseases such as Alzheimer's disease, to our knowledge, no standard imaging genetics workflow has been proposed for PD. The novelty of our workflow can be summarized as follows:

- We propose a domain free and easy-to-use workflow, integrating heterogenous data, such as genotyping, transcriptomic, and neuroimaging data.

- The workflow addresses the complexity of integrating real multisource data when a limited number of data are available by proposing three step-based metho. The first step integrates genotyping and imaging features considering each feature individually, the second phase summarizes imaging features in a single measure, while the last step focuses on linking potential functional effects caused by the genetic biomarkers found during the two previous phases.

- We propose a validation of the method on genetic and imaging data related to PD, showing our new results.

II. AIMS AND OBJECTIVES

The attendees will acquire an experience on how to conduct a complete *imaging genetics* workflow, in a specific case study like PD. After the tutorial session, the attendees will be able to conduct themselves an *imaging genetics* pipeline, which could also be applied to study other neurological diseases. The tutorial will introduce the participants to the biological background, especially with the notion of DNA, RNA, Single nucleotide polymorphism (SNP) and Genome-Wide Association Study (GWAS). The participants will have the opportunity to get familiar with PLINK [5], a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyzes in a computationally efficient manner. PLINK provides a large range of functionalities designed to analyze genotyping data, such as data management, summary statistics, quality control, population stratification detection, or variant-trait associations. Moreover, the audience will learn some notions about the widely used

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX
IEEE

Python and R programming languages, which are extensively used throughout the presented workflow. During the tutorial, will be used different Python and R packages such as *Pandas*, to efficiently manage genotyping data, LIMMA [6], to prioritize candidate genes, variantPartition [7], to handle repeated measures for the RNA-seq data, and *ClusterProfiler* [8], to search the biological pathways that contains the significative genes founded and *WikiPathways* [9, 10] as annotation databases. Finally, they will acquire useful skills regarding tools often used in this field like TATES [11] and GRACE [12]. TATES, which is used in the *Individual View* phase, evaluates the SNP-imaging phenotype associations by assigning to each genomic variant a comprehensive *P*-value. GRACE, which is utilized during the *Integrated View* step, iteratively computes a semiparametric model, explaining the disease progression over time, and a single measure, describing different quantitative neuroimaging features. During the tutorial will be also tackled the fundamental step of results interpretation (*Functional Interpretation*). In fact, it is crucial to link the candidate SNPs biomarkers to the influenced genes and pathways. To do so, we will focus on searching SNP genomic locations on dbSNP [13], and on retrieving the protein-protein interaction network (PPI) with STRINGdb [14], for those genes affected by candidate biomarkers. Moreover, we will introduce how to annotate genetic variants which are not located within genes, by using GTEx [15] data portal.

The tutorial code is wrapped in different Jupyter notebooks (formerly IPython Notebooks), that is a web-based and system-independent interactive computational environment for easy analysis reproducibility. Therefore, the attendees will obtain the experience to use alone the Jupyter Notebook, a very useful open-source framework that allows users to create and share documents containing live code, equations, visualizations and narrative text.

III. MUVIG: MULTI VIEW BASED IMAGING GENETICS WORKFLW ON PARKINSON DISEASE

The tutorial is based on an imaging genetics workflow *MUVIG*: *Multi view based imaging genetics analysis on Parkinson disease*, presented in [16]. The workflow covers completely an imaging genetics analysis, from data preprocessing to results interpretation, integrating first genetic and imaging data, and then transcriptomic data.

The data used for this tutorial were obtained from the Parkinson's Progression Marker Initiative (PPMI) data portal [17]. Currently, PPMI is the most complete and comprehensive collection of PD related data. The dataset that will be used throughout the tutorial consists of genetic and neuroimaging dat. Genetic data consist in a set of genetic variants, more specifically insertions and deletions (indels) and Single Nucleotide Polymorphisms (SNPs), and RNA counts (transcriptomic data). The neuroimaging data consist of DaTSCAN and MRI images, which have been shown to be effective biomarkers for PD onset and progression by several studies [2—4, 18].

The workflow is composed by three main parts: *Individual view, Integrated view* and *Functional interpretation*. The *Individual view* focuses on finding candidate SNP biomarkers integrating genetic and imaging data, by considering each imaging trait individually and combining the respective results. The *Integrated view* searches for genetic variants-phenotype associations, by computing a single phenotypic measure to summarize the imaging features. Here the idea is, instead of considering each imaging feature individually, grouping them by imaging modality. Finally, the *Functional interpretation* works as validation, focusing on linking the results obtained during the two previous phases with their potential functional consequences. During this part are employed transcriptomic and imaging data simultaneously to search for genes, whose expression pattern reflects the changes observed in brain images. Therefore, differentially expressed genes are linked to the cellular pathways in which they participate.

The source code of this workflow is freely distributed to the communities and is available in our github page at: https://github.com/InfOmics/MUVIG. The code is divided for each mentioned section inside the source (*src*) folder.

IV. TUTORS

The tutorial is proposed through a session of 3 hours. The tutors are Prof. Rosalba Giugno, Associate Professor in Computer Science at Department of Computer Science, University of Verona; Guglielmo Cerri, research fellow at Department of Computer Sciene, University of Verona; Manuel Tognon, PhD student of Computer science working at department of Computer Science, University of Verona;Dr. Andre Altmann, MRC Senior Research Fellow, head of COMputational Biology in Imaging and geNEtics (COMBINE) (part of UCL's Centre for Medical Image Computing) and proleptic Lecturer, University College London (UCL).

REFERENCES

[1] L. M. L. De Lau and M. M. B. Breteler. Epidemiology of Parkinson's disease. *The Lancet Neurology* 5.6 (2006): 525-535.

[2] G. Pagano, F. Niccolini, M. Politis. Imaging in Parkinson's disease. *Clinical Medicine* 16.4 (2016): 371.

[3] I. Gayed, U. Joseph, M. Fanous, D. Wan, *et al.*.The impact of DaTscan in the diagnosis of Parkinson disease. *Clinical nuclear medicine* 40.5 (2015): 390—393.

[4] P. Tuite. Magnetic resonance imaging as a potential biomarker for Parkinson's disease. *Translational research* 175 (2016): 4—16.

[5] S. Purcell, B. Neale, K. Todd-Brown, *et al.*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* 81.3 (2007): 559-575.

[6] M.E. Ritchie, B. Phipson, D.I. Wu, *et al.*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43.7 (2015): e47-e47.

[7] G. E. Hoffman and P. Rousso. dream: Powerful differential expression analysis for repeated measures designs. *Bioinformatics* (2020).

[8] G. Yu, L. Wang, Y. Han, *et al.*. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* 16.5 (2012): 284-287.

[9] A. R. Pico, T. Kelder, M. P. Van Iersel, *et al.*. WikiPathways: pathway editing for the people. *PLoS Biol* 6.7 (2008): e184.

[10] M. Martens, A. Ammar, A. Riutta, *et al.*. WikiPathways: connecting communities. *Nucleic Acids Research* 49.D1 (2021): D613-D621.

[11] S. Van der Sluis, D. Posthuma, C. V. Dolan. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* 9.1 (2013): e1003235.

[12] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, *et al.*. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia* 10 (2014): S400-S410.

[13] S. T. Sherry, M. H. Ward, M. Kholodov, *et al.*. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29.1 (2001): 308-311.

[14] D. Szklarczyk, A. L. Gable, D. Lyon, *et al.*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* 47.D1 (2019): D607-D613.

[15] J. Lonsdale, J. Thomas M. Salvatore, *et al.*. The genotype-tissue expression (GTEx) project. *Nature genetics* 45.6 (2013): 580-585.

[16] G. Cerri, M. Tognon, N. Oxtoby, *et al.*. Multi view based imaging genetics analysis on Parkinson disease. *Briefings in Bioinformatics* (submitted) 2021.

[17] K. Marek, D. Jennings, S. Lasch, *et al.*. The Parkinson's Progression Markers Initiative (PPMI). *Progress in neurobiology* 95.4 (2011): 629—635.

[18] N. Pyatigoraskaya, B. Magnin, M. Mongin, *et al.*. Comparative study of MRI biomarkers in the substantia nigra to discriminate idiopathic Parkinson disease. *American Journal of Neuroradiology* 39.8 (2018): 1460—1467.