The 1st International Workshop on Algorithms, Tools and new Fronties
on the use of Networks in Biology and Clinical Science
(BioNet-2017)

A new approach to disentangle genetic and epigenetic components on disease comorbidities: studying correlation between genotypic and phenotypic disease networks

G. Tradigo[a,c], R. Vacca[c], T. Manini[c], V. Bird[c], T. Gerke[c], P. Veltri[b], M. Prosperi[c,*]

[a]University of Calabria, ponte Bucci, Arcavacata di Rende 87036, Italy
[b]Universitá di Catanzaro, viale Europa, Catanzaro 88100, Italy
[c]University of Florida, 2004 Mowry Rd., Gainesville FL 32610-0231, USA

Abstract

Disease comorbidity is a result of complex epigenetic interplay. A disease is rarely a consequence of an abnormality in a single gene; complex pathways to disease patterns emerge from gene-gene interactions and gene-environment interactions. Understanding these mechanisms of disease and comorbidity development, breaking down them into clusters and disentangling the epigenetic — actionable — components, is of utter importance from a public health perspective. With the increase in the average life expectancy, healthy aging becomes a primary objective, from both an individual (i.e. quality of life) and a societal (i.e. healthcare costs) standpoint. Many studies have analyzed disease networks based on common altered genes, on protein-protein interactions, or on shared disease comorbidites, i.e. phenotypic disease networks. In this work we aim at studying the relations between genotypic and phenotypic disease networks, using a large statewide cohort of individuals (100, 000+ from California, USA) with linked clinical and genotypic information, the Genetic Epidemiology Research on Adult Health and Aging (GERA). By comparing their phenotypic and genotypic networks, we try to disentangle the epigenetic component of disease comorbidity.

Keywords: GWAS, comorbidity network, genetic analysis

1. Introduction

Disease comorbidity is a result of complex epigenetic interplay. A disease is rarely a consequence of an abnormality in a single gene; complex pathways to disease patterns emerge from gene-gene interactions and gene-environment interactions. Understanding these mechanisms of disease and comorbidity development, breaking down them into clusters and disentangling the epigenetic — actionable — components, is of utter importance from a public health

* Corresponding author. Tel.: +1-352-273-5860; fax: +1-352-273-5365.
  E-mail address: m.prosperi@phhp.ufl.edu

perspective. With the increase in the average life expectancy, healthy aging becomes a primary objective, from both an individual (i.e. quality of life) and a societal (i.e. healthcare costs) standpoint.

Genome-wide association studies (GWAS) are now very common due to a decrease in sequencing cost and increase in throughput, and large data bases are available publicly linking single nucleotide polymorphisms (SNPs) to diseases, e.g. the GWAS Catalog (https://www.ebi.ac.uk/gwas/). Recent studies have exploited these data bases to create high resolution SNP networks [4,5]. However, there is a lack of studies that analyze SNPs (single nucleotide polymorphisms) and disease comorbidities using the same data, linked at the individual level; this is the objective of our investigation, here presented.

In this work we aim at studying the relations between genotypic and phenotypic disease networks, using a large statewide cohort of 110,266 individuals from California, USA, with linked clinical and genotypic information, the Genetic Epidemiology Research on Adult Health and Aging (GERA). DNA data has been extracted from saliva samples and stored in a text file repository. To disentangle the epigenetic component of disease comorbidity, we build two types of comorbidity networks (i.e. genotypic and phenotypic) for each ethnic group in the database and for the whole patient population. The network construction workflow will be described in Section 2. We then report networks' structural features by using standard network metrics, showing the most significant ones.

2. Methods

Study design and data sets.. The data source for this study is the Genetic Epidemiology Research on Adult Health and Aging (GERA)[6], a public resource funded by the National Institutes of Health (NIH). GERA is a subset of the Kaiser Permanente's Research Program on Genes, Environment, and Health (RPGEH). RPGEH links together comprehensive electronic medical records, data on relevant behavioral and environmental factors, and biobank data (genetic information from saliva and blood) from 500,000 consenting health plan members enrolled among the six million-member Kaiser Permanente Medical Care Plan of Northern California and Southern California. Data from over 100,000 participants from various ethnic groups, with ages from 27 to 97 years (average age: 63), are freely available in GERA, with associated genotyping information, demographics, health-related behaviors, and grouped health conditions on the basis of the International Classification of Diseases v.9 (ICD-9) ontology, from an average of 23.5 years of electronic medical records. High-density genotyping was conducted using custom designed Affymetrix Axiom arrays [7,8].

The final goal, on which we are still working, is to analyze inter- and intra-ethnic group network differences and to analyze and validate the extracted knowledge with experts in the clinical domain. Instead of measuring molecular markers, we aim to categorize patients by measuring co-factors, such as existing health conditions and prescription drug use. The data analysis methods needed to achieve this aim are complex, and may have limited previous research efforts like this. Accurate prognostic classification at diagnosis remains an urgent and unmet challenge, due to confounding by screening practices and comorbid conditions in an aging population. Our endotyping effort may reveal unique clinical profiles that can help guide prognosis and treatment decisions.

Ethics Statement.. This study has been performed in accordance with the Declaration of Helsinki. The research protocol has been approved by University of Floridas Institutional Review Board. The GERA data request has been approved on April 22, 2016, and is deposited on the GERA website under Dr. Travis Gerke's name.

Genome-wide Association Study.. For the analisys of the GERA dbGap database files we used the PLINK tool [1,2], version 1.90b3.42 64-bits. In PLINK the whole GWAS pipeline is implemented in a single tool, allowing to effectively search the most significant genotypic information (i.e. SNPs) explaining the differences in the phenotypic feature set (e.g. age, gender, BMI), also called covariates. From the GERA dbGap database we obtained genotypic data, extracted at University of California San Francisco using custom designed Affymetrix Axiom arrays, for four ethnic groups: (*i*) *AFR* showing genetic similarity with African-Americans, (*ii*) *EUR*, defined, in the GERA Genotypic Data description, as Non-Hispanic White, (*iii*) *EAS* containing patients with East Asian genetic traits and (*iv*) *LAT*, with DNA belonging to the Latinos. It is worth noting that these groups were made up directly from genetic evidence and not from self-declared race memberships by individuals [7]. We also built an integrated dataset, called *ALL*, by combining those groups and we stored the race information as a feature in the phenotype covariates file. In fact, together with the genotypic data, we obtained from the GERA dbGap database a set of phenotypic files containing demographic and