Bioinformatics, 32(8), 2016, 1247–1249
doi: 10.1093/bioinformatics/btv708
Advance Access Publication Date: 15 December 2015
Applications Note

OXFORD

Genetics and population analysis

Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks)

Gilderlanio S. Araújo[1,*], Lucas Henrique C. Lima[2], Silvana Schneider[3], Thiago P. Leal[1], Ana Paula C. da Silva[2], Pedro O. S. Vaz de Melo[2], Eduardo Tarazona-Santos[1], Marília O. Scliar[1] and Maíra R. Rodrigues[1,*]

[1]Department of General Biology, [2]Department of Computer Science and [3]Department of Statistics, Federal University of Minas Gerais, Belo Horizonte, Brazil

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on 30 June 2015; revised on 5 November 2015; accepted on 29 November 2015

Abstract

Motivation: The 1000 Genomes Project (1KGP) and thousands of Genome-Wide Association Studies (GWAS) performed during the last years have generated an enormous amount of information that needs to be integrated to better understand the genetic architecture of complex diseases in different populations. This integration is important in areas such as genetics, epidemiology, anthropology, as well as admixture mapping design and GWAS-replications. Network-based approaches that explore the genetic bases of human diseases and traits have not yet incorporated information on genetic diversity among human populations.

Results: We propose Disease-ANCEstry networks (DANCE), a graph-based web tool that allows to integrate and visualize information on human complex phenotypes and their GWAS-hits, as well as their risk allele frequencies in different populations. DANCE provides an interactive way to explore the human SNP–Disease Network and its projection, a Disease–Disease Network. With these functionalities, DANCE fills a gap in our ability to handle and understand the knowledge generated by GWAS and 1KGP. We provide a number of case studies that show how DANCE can be used to explore the relationships between human complex diseases, their genetic bases and variability in different human populations.

Availability and implementation: DANCE is freely available at http://ldgh.com.br/dance/. We recommend using DANCE with Mozilla Firefox, Safari, Chrome or Internet Explorer (v9 or v10).

Contact: gilderlanio@gmail.com or maira.r.rodrigues@gmail.com

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

1 Introduction

Information on human diversity and complex human phenotypes are currently organized in independent databases. The 1000 Genomes Project (1KGP) (Auton *et al.*, 2015) database contains information on the spectrum of genetic variants in different human populations. The NHGRI-EBI GWAS Catalog (Welter *et al.*, 2014) contains information on the genetic architecture of complex diseases and traits. Despite the importance of this information, there are no tools to summarize, integrate and visualize it, which is critical for a wide spectrum of areas such as genetics, epidemiology, anthropology, to design admixture mapping or GWAS-replications and to test the genetic architecture of diseases in different populations.

Relationships between human phenotypes and their causing genetic variants have been modeled using network-based approaches,

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

which provide a comprehensive and integrated view of phenotypes interrelations (Barabási *et al.*, 2011). Despite the fact that differences in allele frequencies can account for unequal phenotype prevalences among populations, current network approaches do not integrate this information. Here, we present Disease-ANCEstry networks (DANCE), a graph-based web tool that allows to organize and visualize information on human complex phenotypes and their GWAS-hits, together with their allele frequencies in African, European and Asian populations.

2 DANCE

DANCE integrates information from two existing databases: (i) GWAS-hit SNPs reported in the NHGRI-EBI GWAS Catalog and (ii) risk-allele frequencies in Europeans, Africans and Asians from the 1KGP. From the complete list of SNPs and associated phenotypes provided by the first database, we selected 495 phenotypes and 8019 SNPs with specified risk-alleles.

Data can be visualized with two network structures/projections, an SNP–Disease network and a Disease–Disease network. The former is a bipartite network where nodes are phenotypes or GWAS-hits SNPs. Different phenotypes interconnect through shared SNPs. Population information is represented as a property of the SNP node, and includes the risk-allele frequency in a specific continental population, or a population pairwise allele frequency differentiation statistics F_{ST} (Hartl and Clark, 1998). The Disease–Disease network is the edge-weighted projection of the SNP–Disease Network, where two phenotypes are linked only if they share GWAS-hits. Weights are based on the Jaccard's Coefficient overlap index (Salton and McGill, 1986), and phenotype nodes are annotated using the Medical Subject Headings.

DANCE users can explore and visualize subnetworks of both the SNP–Disease and the Disease–Disease networks. They can interactively perform four tasks: (i) identify and visualize the set of SNPs associated with a phenotype, including their risk-allele continental frequencies and their pairwise F_{ST}; (ii) identify and visualize GWAS-hits from a particular gene and the associations with phenotypes, including their risk-allele continental frequencies or their pairwise F_{ST}; (iii) identify and visualize shared associations between two or more phenotypes by neighborhood visualization; and (iv) explore the Disease–Disease Network to identify clusters of phenotypes sharing risk-alleles.

Figure 1A presents, as a case study, the SNP–Disease Network for bipolar disorder and neighboring phenotypes. We observe that bipolar disorder and four other neuropsychiatric disorders (depression, autism, schizophrenia and attention deficit hyperactivity disorder) share various risk-alleles, forming a large cluster. Besides, bipolar disorder is linked with suicide by two risk-alleles and with ulcerative colitis by one risk-allele. This large cluster of neuropsychiatric disorders also appears in the Disease–Disease Network (Supplementary Fig. S4). To identify which SNPs account for this cluster, we analyze the annotation table, available below the network view on the web interface. We find that the cluster is formed by 54 risk-alleles, from which 28 are mapped to genes and 26 are intergenic. We identify nine risk-alleles shared only by bipolar disorder and depression, six by bipolar disorder and schizophrenia and five by these three phenotypes. In addition, by looking at the risk-allele frequencies we see that five of them are fixed in Asians (frequency = 1), two of them are in TCF4 (transcription factor 4) and three in KIF5C (kinesin family member 5C). Further investigation of these genes is possible through the selection of the Gene filter. Figure 1B depicts the TCF4 gene network colored according to risk-allele frequency in Asians. We note that TCF4 SNPs links to other four

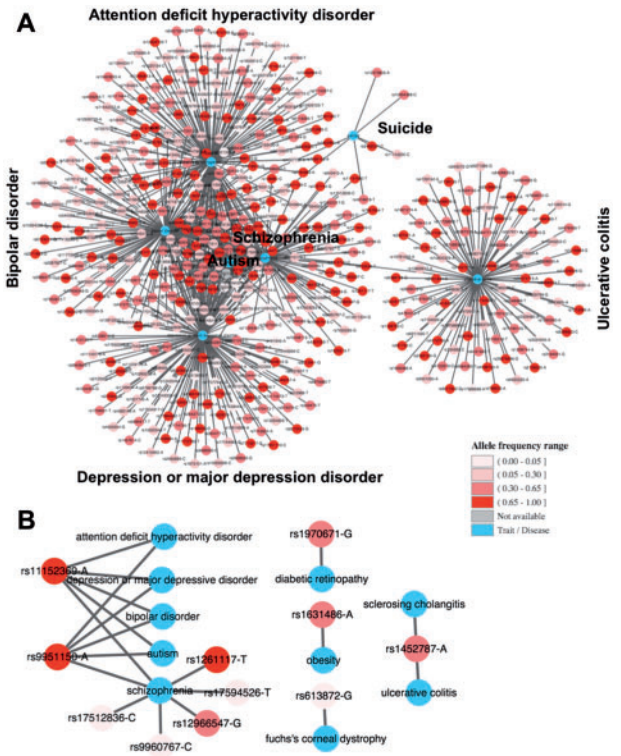

Fig. 1. SNP–Disease Network for (**A**) bipolar disorder and neighboring phenotypes and (**B**) TCF4. SNPs are colored according to risk-allele frequency in Africans (A) and Asians (B)

phenotypes besides the neuropsychiatric disorders. Three out of 11 risk-alleles of the network are non-polymorphic in Asians (one has frequency equal to 1, and the other two equal to 0), one is almost fixed (frequency = 0.997) and two are very rare (frequency $\leq$0.004). This pattern of less diversity is common for other genes and phenotypes for the Asian population, and Li *et al.* (2010) reported this low diversity of TCF4 risk-alleles. In our dataset, Asians show 602 non-polymorphic risk-alleles out of 8019; on the other hand, Europeans show 11, and Africans seven fixed alleles. Indeed, with DANCE users can assess differences in risk-allele frequencies among populations even when there is no GWAS available for the population of interest, although caution is necessary because the genetic architecture of a disease may not be entirely shared by different populations.

In the Supplementary material, we present two additional case studies. The first shows obesity-related networks, highlighting the overlap of this condition with the well-known Duffy Antigen Receptor gene, and detailing the connections of the obesity-related FTO gene (through the Gene filter). Second, we explore the Disease–Disease Network of Alzheimer's disease and find that it is linked to other five phenotypes through only one SNP risk-allele localized in TOMM40 gene, which we explore further.

All information used to construct the DANCE networks is available as tables that can be visualized, re-sorted or exported. Detailed information about DANCE implementation is in the Supplementary material. DANCE is periodically updated to incorporate the latest versions of 1KGP and GWAS-Catalog databases.

3 Conclusions

DANCE fills a gap in our ability to handle and understand the knowledge generated by the 1KGP and GWAS initiatives, being the

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

first SNP–disease network-based approach that incorporates population genetic variability.

DANCE is user-friendly, intuitive and interactive, and therefore, is also used by its authors as an educational tool at undergraduate and graduate levels. We expect it to inspire similar tools to study the genetic architecture of complex traits in model organisms such as *Drosophila* or mice. As future work, we plan to represent the population where GWAS were performed, SNPs effect size and linkage disequilibrium, and allele frequencies for other populations such as Native Americans.

Acknowledgements

Ana Lúcia Brunialti, Fabrício Benvenuto and Raquel Minardi for discussions and CAPES Agency from the Brazilian Ministry of Education for funding.

Conflict of interest: none declared.

References

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Hartl,D. and Clark,A. (1998) *Principles of Population Genetics, Vol 116*. Sinauer Association: Sunderland.

Li,T. *et al.* (2010) Common variants in major histocompatibility complex region and TCF4 gene are significantly associated with schizophrenia in Han Chinese. *Biol. Psychiatry*, **68**, 671–673.

Salton,G. and McGill,M.J. (1986) Introduction to modern information retrieval. McGraw-Hill, Inc.: New York.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, 1–6.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016