Published in final edited form as:

Schizophr Res. 2017 March ; 181: 83–85. doi:10.1016/j.schres.2016.09.011.

Independent Component Analysis of SNPs Reflects Polygenic Risk Scores for Schizophrenia

Jiayu Chen¹, Vince D. Calhoun^{1,2}, Godfrey D. Pearlson^{3,4}, Nora I. Perrone-Bizzozero⁵, Jessica A. Turner^{1,6}, Stefan Ehrlich^{7,8,9}, Beng-Choon Ho¹⁰, and Jingyu Liu^{1,2}

¹The Mind Research Network, Albuquerque, NM, USA

²Department of Electrical Engineering, University of New Mexico, Albuquerque, NM, USA

³Olin Neuropsychiatry Research Center, Institute of Living, Hartford, CT, USA

⁴Departments of Psychiatry and Neuroscience, Yale University, New Haven, CT, USA

⁵Departments of Neurosciences and Psychiatry, University of New Mexico School of Medicine, Albuquerque, NM, USA

⁶Psychology Department and Neuroscience Institute, Georgia State University, Atlanta, GA, USA

⁷Department of Child and Adolescent Psychiatry, Eating Disorder Services and Research Center, Technische Universit at Dresden, Faculty of Medicine, University Hospital C. G. Carus, Dresden, Germany

⁸MGH/MIT/HMS Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

⁹Department of Psychiatry, Harvard Medical School, Massachusetts General Hospital, Boston, MA, USA

¹⁰Department of Psychiatry, University of Iowa, Iowa City, IA, USA

Abstract

Schizophrenia is a psychiatric disorder with high heritability. Recent genome-wide association studies have provided a list of risk loci reliably derived from unprecedentedly large samples. However, further delineation of the diagnosis-associated susceptibility variants is needed to better characterize the genetic architecture given the disease's complex nature. In this sense, a data-driven approach might hold promise for identifying functionally related clusters of genetic variants that might not be captured by hypothesis-based models. In the current study, independent component analysis (ICA) was applied to the Psychiatric Genomics Consortium's schizophrenia-

The corresponding author: Jiayu Chen, The Mind Research Network, 1101 Yale Blvd. NE. Albuquerque, NM USA 87106-3834, Phone: (505) 504-0143; Fax: (505) 272-8002; jchen@mrn.org.

Conflict of interest: The authors declare no conflict of interest.

Contributors: JC, VDC and JL designed this study. JC analyzed the data. All authors contributed to the preparation of this manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

related single nucleotide polymorphisms (SNPs) in 104 schizophrenia patients and 142 healthy controls of European Ancestry. We found that, for 13 out of 16 extracted independent components, the associated loadings correlated highly ($r>0.5$) with the polygenic risk scores for SZ of the corresponding SNPs. These correlations were likely not inflated by the linkage disequilibrium structure (permutation $p<0.001$). In brief, we demonstrate an example of ICA analysis on SNP data yielding functionally meaningful clusters, which motivates further application of data-driven approaches as a complimentary tool for hypothesis-based methods to enrich our knowledge on the genetic basis of complex disorders.

Keywords

ICA; polygenic risk score; schizophrenia; PGC

Introduction

Genome-wide association studies (GWAS) provide evidence for a polygenic model of schizophrenia (SZ) where 23% of the variance in liability can be explained by 915,354 common single nucleotide polymorphisms (SNPs) covered in the Psychiatric Genomic Consortium (PGC) schizophrenia study (Lee et al., 2012; Purcell et al., 2009; Ripke et al., 2011). A more recent roadmap study by PGC identified 128 independent genetic associations with SZ from ~35,500 cases (Ripke et al., 2014). However, these diagnosis-associated susceptibility SNPs are likely highly heterogeneous in terms of their pathways to phenotype, considering that SZ is a complex disorder. Echoing this speculation, one recent work investigated genetic influences on schizophrenia and subcortical brain volumes and found no notable overlap at overall common variants level (Franke et al., 2016). Hence, further delineation is required to characterize the genetic architecture and better understand the pathophysiology of SZ. In this sense, a data-driven approach to dissect large sets of SNPs may hold promise for identifying functionally related variants without a prior knowledge of their biological functions.

Independent component analysis (ICA) is a type of blind source separation approach which finds extensive applications in the neuroimaging fields (Calhoun et al., 2009). However, its application to genotype data has not been as popular (Pearlson et al., 2015), presumably due to the difficulty of interpretation, given the limited knowledge on gene-gene interactions and intergenic SNPs. In this work, we leverage the information derived from a large schizophrenia GWAS, and show, as an example, that ICA analysis yields valid clustering where the identified covariation patterns highly correlate with the polygenic risk scores for SZ (PRS-SZ) of those clusters of SNPs.

Materials and Methods

A total of 246 subjects of European Ancestry, including 142 healthy controls and 104 individuals with SZ, were aggregated from two cohorts that were not used in the PGC study, i.e., the Mind Clinical Imaging Consortium (MCIC) study and a Center of Biomedical Research Excellence (COBRE) study. The institutional review board at each site approved the study and all participants provided written informed consents. Table S1 lists the

demographic information. Recruitment and clinical screening information were described in the previous study (Chen et al., 2013). The MCIC and COBRE data were genotyped using the Illumina chips and went through the standard imputation, quality control and population structure correction procedures as detailed in the supplemental information. Discrete numbers were assigned to the categorical genotypes: 0 for no minor allele, 1 for one minor allele, and 2 for two minor alleles. Finally, 5,468 SZ-related SNPs were mapped from the 108 PGC reported SZ-related regions (Ripke et al., 2014) and entered into the ICA analysis.

A detailed description of the method is provided in the Supplemental information. Briefly, ICA decomposes the input data X into a linear combination of sources (or components) such that $X=AS$. S and A denote the component and loading matrix, respectively. Each row of S represents a component while each column of A represents a loading vector associated with the component. For interpretation, the component is generally normalized (z-score) and the top contributing SNPs are selected with $|z\text{-score}|>2$. While ICA decomposition essentially identifies a set of variables covarying with an embedded pattern as reflected in the loading vector, this covariation pattern is largely contributed to by the top SNPs. Thus the top SNPs of each component can be considered as a cluster where the cluster members covary with the loading vector. In the current work, given the 246 samples and 5,468 SZ-related SNPs, we extracted 16 independent components (ICs) based on the criterion of minimal description length (Rissanen, 1978) using Infomax ICA (Amari, 1998; Bell and Sejnowski, 1995). Top SNPs were then selected with $|z\text{-score}|>2$. For each component, we computed the loadings accounted by the top contributing SNPs (denoted as A_{top}). Then correlations between A_{top} and A were computed to assess the contributions of top SNPs to the loading patterns accounted by all the input SNPs. The PRS-SZ was also computed for the top SNPs of each component. It was a linear weighted sum of the top SNPs' genotype data where weights were natural logarithms of the odds ratios of the same SNPs released by the PGC GWAS of SZ (De Jager et al., 2009; Ripke et al., 2014). Correlations were then computed between ICA loadings accounted by the top SNPs and their PRS-SZs. Furthermore, we investigated with a permutation test whether the observed A_{top} -PRS $_{top}$ correlations might be inflated by the linkage disequilibrium (LD) structure.

Results

Table 1 lists the number of top SNPs, A - A_{top} correlation, as well as A_{top} -PRS $_{top}$ correlation for each IC. The top SNPs of each component are summarized in Table S2. Table S3 details a representative component (IC1) with SNP positions and hosting genes listed. As shown in Table 1, all 16 components showed very high correlations ($r>0.9$) between the loadings accounted by all SNPs (A) and those accounted by the top SNPs (A_{top}), confirming that the ICA components reflected the covariation patterns among the top SNPs. For 13 out of the 16 components, A_{top} showed high correlations ($|r|>0.5$, $p<5.76\times10^{-17}$) with the corresponding PRS-SZs, indicating that the identified covariation patterns among top SNPs reflected their cumulative risk for SZ identified previously in much large samples. For most of the components, the top SNPs were distributed across multiple LD blocks, as shown in Figure S1, suggesting that ICA captures covariation including but not limited to LD structure. In addition, a significant $p\text{-value}<0.001$ was observed from the permutation test, confirming that the observed A_{top} -PRS $_{top}$ correlations were not likely fully attributable to the LD

structure. A speculation is that, a cluster of top SNPs covarying with the identified pattern might be reflecting their participation in the same genetic network where functional adaptation and selective pressure may play a role in maintaining the covariation (Jones et al., 2014; Rohlf et al., 2010). A DAVID functional annotation analysis (Huang et al., 2009a, b) on IC1 revealed a number of enriched pathways, including synapse, mitotic cell cycle, chromatin modification, as summarized in Table S4. The annotation results suggest some possibility of interactions.

Discussion

In summary, we demonstrate that ICA decomposition of SZ-related SNPs yielded meaningful clusters. The ICA components were dominated by top SNPs whose covariation patterns (reflected in loadings) correlated highly with their polygenic risk scores for SZ. And the observed correlations between ICA loadings and PRS-SZs were significantly higher than those observed from randomly weighted genotype data with the same data structure, justifying the ICA decomposition. Given that the PRS-SZs were calculated using odds ratios derived from a GWAS including ~35,500 cases, it is noteworthy that ICA is able to identify patterns highly correlated with the PRS-SZs in just 246 subjects. Meanwhile, although this observation increases the confidence that the identified components are likely functionally meaningful, how these SNPs are functionally interactive awaits further investigation. Overall, this type of blind source separation analysis holds promise for identifying subsets of functionally related genetic variants that may not be recognized by hypothesis-based models, which might help characterize the genetic architecture of the susceptibility SNPs identified from GWAS and contribute to delineating the pathophysiological pathways underlying the complex disorder.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was funded by the National Institutes of Health grants P20GM103472, R01EB005846, 1R01EB006841 and 1R01MH094524-01A1, as well as an NSF EPSCoR grant #1539067.

Role of funding: The funding sources had no involvement in the study design, data analysis and interpretation of data, and in the writing of the manuscript.

References

- Amari S. Natural Gradient Works Efficiently in Learning. *Neural Comput.* 1998; 10:251–276.
- Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 1995; 7(6):1129–1159. [PubMed: 7584893]
- Calhoun VD, Liu J, Adalı T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage.* 2009; 45(1, Supplement 1):S163–S172. [PubMed: 19059344]
- Chen J, Calhoun VD, Pearlson GD, Perrone-Bizzozero N, Sui J, Turner JA, Bustillo JR, Ehrlich S, Sponheim SR, Canive JM, Ho BC, Liu J. Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *Neuroimage.* 2013; 83C:384–396.

- De Jager PL, Chibnik LB, Cui J, Reischl J, Lehr S, Simon KC, Aubin C, Bauer D, Heubach JF, Sandbrink R, Tyblova M, Lelkova P, Havrdova E, Pohl C, Horakova D, Ascherio A, Hafler DA, Karlson EW. Benefit, Beyond, Ltf, Studies, C. Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol*. 2009; 8(6):1111–1119. [PubMed: 19879194]
- Franke B, Stein JL, Ripke S, Anttila V, Hibar DP, van Hulzen KJ, Arias-Vasquez A, Smoller JW, Nichols TE, Neale MC, McIntosh AM, Lee P, McMahon FJ, Meyer-Lindenberg A, Mattheisen M, Andreassen OA, Gruber O, Sachdev PS, Roiz-Santianez R, Saykin AJ, Ehrlich S, Mather KA, Turner JA, Schwarz E, Thalamuthu A, Yao Y, Ho YY, Martin NG, Wright MJ, Schizophrenia Working Group of the Psychiatric Genomics, C., Psychosis Endophenotypes International, C., Wellcome Trust Case Control, C. Enigma C, O'Donovan MC, Thompson PM, Neale BM, Medland SE, Sullivan PF. Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. *Nat Neurosci*. 2016; 19(6):420–431. [PubMed: 26854805]
- Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009a; 37(6):1–13. [PubMed: 19033363]
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009b; 4(6):44–57. [PubMed: 19131956]
- Jones AG, Burger R, Arnold SJ. Epistasis and natural selection shape the mutational architecture of complex traits. *Nat Commun*. 2014; 5
- Lee SH, DeCandia TR, Ripke S, Yang J, Schizophrenia Psychiatric Genome-Wide Association Study, C., International Schizophrenia, C., Molecular Genetics of Schizophrenia, C. Sullivan PF, Goddard ME, Keller MC, Visscher PM, Wray NR. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet*. 2012; 44(6):247–250. [PubMed: 22344220]
- Pearlson GD, Liu J, Calhoun VD. An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in genetics*. 2015; 6:276. [PubMed: 26442095]
- Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009; 460(6):748–752. [PubMed: 19571811]
- Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang HL, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau RA, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai GQ, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Chong SA, Cloninger CR, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodriguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard MV, Hougaard DM, Ikeda M, Joa I, Julia A, Kahn RS, Kalaydjieva L, Karachanak-Yankova S, Karjalainen J, Kavanagh D, Keller MC, Kennedy JL, Khrunin A, Kim Y, Klovins J, Knowles JA, Konte B, Kucinskas V, Kucinskiene ZA, Kuzelova-Ptackova H, Kahler AK, Laurent C, Keong JLC, Lee SH, Legge SE, Lerer B, Li MX, Li T, Liang KY, Lieberman J, Limborska S, Loughland CM, Lubinski J, Lonnqvist J, Macek M, Magnusson PKE, Maher BS, Maier W, Mallet J, Marsal S, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melegh B, Melle I, Meshulam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Mors O, Murphy KC, Murray RM, Myin-Germeys I, Muller-Myhsok B, Nelis M, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nikitina-Zake L, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, O'Neill FA, Oh SY, Olincy A, Olsen L, Van Os J, Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietilainen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quedsted D, Rasmussen HB, Reichenberg A, Reimers

MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Schall U, Schubert CR, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi JX, Sigurdsson E, Silagadze T, Silverman JM, Sim K, Slominsky P, Smoller JW, So HC, Spencer CCA, Stahl EA, Stefansson H, Steinberg S, Stogmann E, Straub RE, Strengman E, Strohmaier J, Stroup TS, Subramaniam M, Suvisaari J, Svrakic DM, Szatkiewicz JP, Soderman E, Thirumalai S, Toncheva D, Tosato S, Veijola J, Waddington J, Walsh D, Wang D, Wang Q, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wong EHM, Wormley BK, Xi HS, Zai CC, Zheng XB, Zimprich F, Wray NR, Stefansson K, Visscher PM, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Borglum AD, Cichon S, Darvasi A, Domenici E, Ehrenreich H, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jonsson EG, Kendler KS, Kirov G, Knight J, Lencz T, Levinson DF, Li QQS, Liu JJ, Malhotra AK, McCarroll SA, McQuillin A, Moran JL, Mortensen PB, Mowry BJ, Nothen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sham PC, Sklar P, St Clair D, Weinberger DR, Wendland JR, Werge T, Daly MJ, Sullivan PF, O'Donovan MC. Consortium, P.G., Conso, P.E.I., Consor, W.T.C.-C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014; 511(7510):421–+. [PubMed: 25056061]

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DH, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jonsson EG, Bitter I, Pietilainen OP, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Borglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Haan L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthøj B, Godard S, Hamshire M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jurgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krasucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang KY, Lichtenstein P, Lieberman JA, Linszen DH, Lonnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattingdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentoft M, Norton N, Nothen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Orntoft TF, Owen MJ, Pantelis C, Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quedsted D, Quinn EM, Rasmussen HB, Rethelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CC, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Terenius L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*. 2011; 43(6):969–976. [PubMed: 21926974]

Rissanen J. Modeling by Shortest Data Description. *Automatica*. 1978; 14(6):465–471.

Rohlf s RV, Swanson WJ, Weir BS. Detecting Coevolution through Allelic Association between Physically Unlinked Loci. *Am J Hum Genet*. 2010; 86(6):674–685. [PubMed: 20381007]

Table 1
Correlations between the top SNPs' ICA loadings (A_{top}) and polygenic risk scores for schizophrenia (PRS_{top})

IC index	Number of Top SNPs	A- A_{top} correlation (r-value)	A_{top} - PRS_{top} correlation (r-value)
1	273	0.9179	-0.5784
2	231	0.9245	-0.7578
3	143	0.9311	0.7733
4	116	0.9139	-0.6993
5	184	0.9206	0.1900
6	163	0.9227	-0.7092
7	146	0.9307	-0.5833
8	199	0.9228	-0.4623
9	260	0.9281	0.4999
10	130	0.9255	-0.8827
11	131	0.9307	0.6596
12	278	0.9367	-0.5136
13	122	0.9448	0.9700
14	148	0.9377	-0.8777
15	166	0.9490	-0.9280
16	80	0.9797	0.9934