Epistasis analysis reveals associations between gene variants and bipolar disorder

In complex phenotypes (e.g., psychiatric diseases) single locus tests, commonly performed with Genome-Wide Association Studies, have proven to be limited in discovering strong gene associations. A growing body of evidence suggests that epistatic non-linear effects may be responsible for complex phenotypes arising from the interaction of different biological factors. A major issue in epistasis analysis is the computational burden due to the huge number of statistical tests to be performed when considering all the potential genotype combinations. In this work, we developed a computational efficient pipeline to investigate the presence of epistasis at a genome-wide scale in bipolar disorder, which is a typical example of complex phenotype with a relevant but unexplained genetic background. By running our pipeline we were able to identify 13 epistasis interactions between variants located in genes potentially involved in biological processes associated with the analyzed phenotype.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3242v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

Epistasis analysis reveals association between gene variants and bipolar disorder

Carlo Maj[1*], Elena Milanesi[1], Massimo Gennarelli[1], Luciano Milanesi[2], Ivan Merelli[2]

[1] Genetics Unit, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
[2] Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, Italy.

*cmaj@fatebenefratelli.eu

Introduction

Single locus tests, commonly performed with Genome-Wide Association Studies, have proven to be limited in discovering strong gene associations for complex phenotypes (e.g., psychiatric diseases) [1]. With this approach significant associations can be hardly found and usually the identified associations are not replicated in different cohorts. This could be due to the fact that genetic architecture of complex diseases involves different genes. As a matter of fact, the role of genetic factors in complex phenotype differs from the traditional Mendelian phenotype in which a single variant presents a high penetrance rate (i.e., a given variant causes the phenotype).

To outpace this problem, the study of how genetic variants interact to define a specific phenotype, called epistasis, is gaining progressive interest. In biological terms we refer to epistasis when the effect of a gene depends also on another gene. In fact, a big number of variants can play minor additive roles generating a shared polygenic effect [2], while epistatic nonlinear interactions may be responsible for specific biological phenotypes [3].

In statistical terms, we refer to epistasis when the effect of a combined set of variants differs from the linear combination of the marginal effect [4]. The computation of the epistasis of two (or more) loci is, therefore, an indication that the phenotype is affected by the genotype combination rather the single variants, suggesting an underlying potential biological association [5].

A major issue in epistasis analysis is the computational burden due to the huge number of statistical tests to be performed when considering all the potential genotype combinations [6]. For this reason, the majority of the studies evaluating epistatic effects have been done considering specific target genes, while only a few studies performed a genome-wide analysis.

In this work, we developed a computational efficient pipeline to investigate the presence of epistasis at a genome-wide scale in bipolar disorder which is a typical example of complex phenotype with a relevant, but unexplained genetic background. In fact, despite it is well established that bipolar disorder is characterized by a pivotal genetic component, only few variants have been found to be associated with this disease [7]. Polygenic cumulative effects and epistasis can overcome the limit of the traditional single locus association by investigating the role of multifactorial and combined effects of different genetic variants.

Methods

Our epistasis analysis has been performed considering Whole Exome data from dbGaP study (phs000021.v2.p1) between 1166 bipolar disorder patients and 6181 controls. The Psychiatric Genomic Consortium (PGC) study found only few significant peaks overcoming genome-wide significance, although a huge number of nominally associated variants have been identified [8].

In order to biologically interpret the results, we selected for the epistasis evaluation only the variants nominally associated. Noteworthy, most of the data from PGC derives from GWAS data which include a large number of intergenic/intronic variants that are not present in the exome dataset, which is instead limited to exons and flanking regions. Linkage Disequilibrium (LD) filtering have been performed to filter variants in linkage, which are inherited together. Therefore, in statistical terms, they carry the same information and can be filtered out using a single variant as a proxy for the other variants in linkage. To perform LD pruning we considered a threshold of 2 in the Variance Inflator Factor (VIF) using a window size of 50 SNPs. After LD and PGC nominal association filtering, we analyzed the epistasis between 1628 variants spread among the whole genome.

In this work, we implemented a pipeline for the epistasis analysis that relies on PLINK 1.9 [9] to model and test SNP interactions, since its regression-based approach though not the most

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3242v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

accurate allows to perform a comprehensive exploration of the input space at least for binary interactions.

In particular, we first screened the interactions using the *FastEpistasis module*, which uses an efficient parallel algorithm to test pair-wise interactions. When using a single core computer, this screening approach is 15 times faster than PLINK [10]. Moreover, it is implemented using a multi-thread approach in order to exploit all the CPU-cores of a server, and it can be further distributed among different servers. FastEpistasis performs an imprecise, but fast scan for epistasis based on inspection of 3x3 joint genotype count tables. The idea is to convert the three genotype categories into two allele categories by considering allele products:

	YY	**Yy**	**yy**
XX	*a*	*b*	*c*
Xx	*d*	*e*	*f*
xx	*g*	*h*	*i*

⇒

	Y	**y**
X	*4a+2b+2d+e*	*4c+2b+2f+e*
x	*4g+2h+2d+e*	*4i+2h+2f+e*

and to compute the odd ratios between loci X and Y.
The test for epistasis is based on the Z-score between the odd ratio of cases and controls:

$$Z = \frac{\log\left(ORcases\right) - \log\left(ORcontrols\right)}{sqrt\left(SE\left(ORcases\right) + SE\left(ORcontrols\right)\right)}$$

which follows a standard normal distribution.

For large datasets, it is reasonable to start with this command (using liberal p-value thresholds) to identify candidate pairs for further investigation, and then follow up with a more rigorous and computationally expensive analysis on those pairs as the one based on regression methods.

Indeed the second step of our pipeline is represented by the classic PLINK epistasis analysis, performed using a linear/logistic regression (according to the phenotype) to fit the interaction model:

$$Y = \beta_0 + \beta_1 g_{x_1} + \beta_2 g_{x_2} + \beta_3 g_{x_1} g_{x_2}$$

for each inspected variant pair (x_1, x_2), where g_{x1} and g_{x2} are allele counts and the β_3 is the coefficient to test for significance. Since the test is based on the interaction coefficient β_3 the *epistasis module* measures how much the association between the two inspected variants and the analyzed phenotype differs from a pure linear/logistic additive model. Considering that the number of interactions analyzed in this step is a fraction of the original combinatory number of SNP pairs, we were able to correct results for multiple testing by computing a q-value.

The last step of the pipeline was the PLINK *twolocus function*, which computes tables of joint genotype counts and frequencies between the two specified variants. This analysis is computed for interactions that pass the q-value filter in the second step, in order to highlight the distribution of the SNPs in both cases and controls. Indeed, for case/control data, two similar sets of tables are reported, which stratify the two-locus genotype counts and frequencies by cases and controls.

We performed our analysis using a twenty-core virtual server relying on the Intel Xeon Sandy Bridge technology, harboring 64GB of memory. Using this server, the screening step of 946650 variants in 7347 samples took 5760 minutes. The exact computation of the epistasis effect among couples of filtered polymorphisms took 363 minutes. The final step, involving the computation of joint genotypes tables for 6.69 E6 pairwise interactions took 2387 minutes.

Results

Our epistasis analysis detected thirteen epistatic interactions within the analyzed variants (see Table 1). Interestingly, none of them is genome-wide associated at single locus analysis, while only one (i.e., rs1060570) is only nominally associated. In other words, such epistasis interactions involve variants that would be filtered out by using standard single-locus GWAS association. Their effects on the phenotype arise from a nonlinear combination of genotype distribution rather than from an additive cumulative effect. Some of these associations concern genes which could be biologically related to neuropsychiatric disorders. Of particular interest seems to be the GABRA4-

AFG3L3 association, where the first gene codifies for a receptor involved in synaptic transmission and AFG3L3 is an enzyme involved in axonal development. Another interesting interaction is between TAOK2 and a microRNA (MIE499B) which is predicted to regulate its gene expression.

On the other hand, none of the 16 single locus genome-wide significant associations has been found in any significant epistasis interaction, suggesting that their effect is exerted at an individual level or it is purely addictive. Following the statistical evaluation of epistasis values, it is required to interpret the potential biological meaning of the identified interactions. PLINK offers the *twolocus function* which displays the distribution of the combined genotypes between case and controls.
By comparing the distributions, it is possible to assess which is the potential genotype combination of risk/protection for the analyzed phenotype (see Figure 1).

Table 1: Identified significant epistatic interactions

Chr1	Snp1	Chr2	Snp2	Gene1	Gene2
14	rs1060570	19	rs2302224	DPF3	PTPRS
10	rs11191741	16	rs12443685	SH3PXD2A	ABCC11
4	rs16869654	20	rs2236523	SLIT2	ADRM1
7	rs1799370	10	rs10740579	TNS3	PCDH15
6	rs1892172	22	rs737945	RSPO3	ASCC2
12	rs2044846	19	rs254259	USP15	NDUFA3
4	rs2055943	18	rs11080572	GABRA4	AFG3L2
11	rs2509010	12	rs7961392	MMP27	RFX4
3	rs3736156	18	rs1788799	TKT	NPC1
16	rs3814883	20	rs2425009	TAOK2	MIR499B
7	rs728275	9	rs7856971	OR6B1	ASTN2
3	rs9628	13	rs7335339	DCP1A	ATP8A2
3	rs9822460	15	rs3803406	OR5K4	ALPK3

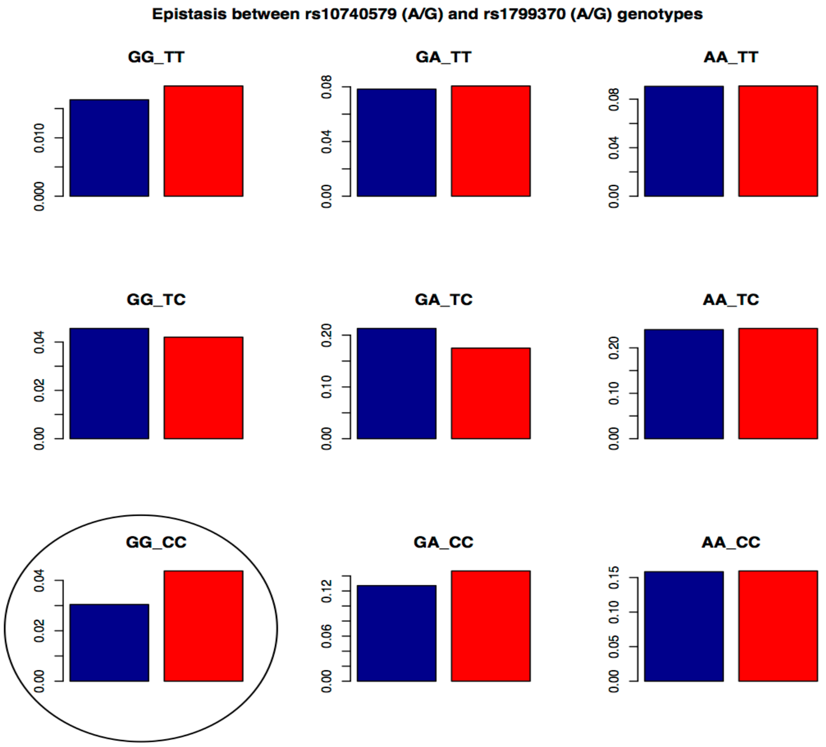

Figure 1 Example of combined genotype distributions. rs10740579 (column) and rs1799370 (row) have similar frequencies between cases and controls (0.31 vs 0.30 for the former, 0.42 vs 0.43 for the latter). However, specific genotype combinations, as the double homozigous GG CC (circled) show different frequencies between cases (in red) and controls (in blue). Interestingly, both the involved

genes (rs10740579 is within an intron of PCDH15 while rs1799370 is within an intron of TNS3) are implicated in the same biological process (i.e., cell adhesion).

Conclusions

In this work, we presented a whole genome epistasis evaluation in bipolar disorder identifying a number of potentially biologically relevant epistatic interactions. These data confirm the genetic background complexity of this disorder where both polygenic additive effects and the establishment of gene-gene non-linear interactions, play a synergic role. Please note that we filtered nominally associated variants and we considered only pairwise combinations among SNPs. Therefore, part of the missing hereditability could be due also by new emerging associations possibly involving high order interactions.

This result suggests the importance of including also epistasis evaluation in standard genetic association analysis. In fact, the combinatorial effect of genetic variations may explain potential association which would be otherwise excluded from standard single-locus analysis. Indeed, it is expected that for phenotype with a strong non-Mendelian genetic is the effect of multiple variants to be somehow related to the phenotype (possibly also in interactions with the environment). Epistasis analysis has the capability to identify specific associations at genetic level which can allow researchers to better characterize the biological mechanisms underlying a given phenotype.

Bibliography

[1] A. L. Collins and P. F. Sullivan, “Genome-wide association studies in psychiatry: what have we learned?,” *Br. J. Psychiatry*, vol. 202, no. 1, pp. 1–4, Jan. 2013.

[2] N. Chatterjee, J. Shi, and M. García-Closas, “Developing and evaluating polygenic risk prediction models for stratified disease prevention.,” *Nat. Rev. Genet.*, vol. 17, no. 7, pp. 392–406, 2016.

[3] T. F. Mackay and J. H. Moore, “Why epistasis is important for tackling complex human disease genetics.,” *Genome Med.*, vol. 6, no. 6, p. 124, 2014.

[4] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, “A survey about methods dedicated to epistasis detection.,” *Front. Genet.*, vol. 6, p. 285, 2015.

[5] A. Le Rouzic, “Estimating directional epistasis.,” *Front. Genet.*, vol. 5, p. 198, 2014.

[6] W.-H. Wei, G. Hemani, and C. S. Haley, “Detecting epistasis in human complex traits.,” *Nat. Rev. Genet.*, vol. 15, no. 11, pp. 722–33, Nov. 2014.

[7] M. A. Escamilla and J. M. Zavala, “Genetics of bipolar disorder.,” *Dialogues Clin. Neurosci.*, vol. 10, no. 2, pp. 141–52, 2008.

[8] Psychiatric GWAS Consortium Bipolar Disorder Working Group, “Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4.,” *Nat. Genet.*, vol. 43, no. 10, pp. 977–83, Sep. 2011.

[9] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, “PLINK: a tool set for whole-genome association and population-based linkage analyses.,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–75, Sep. 2007.

[10] T. Schüpbach, I. Xenarios, S. Bergmann, and K. Kapur, “FastEpistasis: a high performance computing solution for quantitative trait epistasis.,” *Bioinformatics*, vol. 26, no. 11, pp. 1468–9, Jun. 2010.