Identification of protein subnetworks implicated in Autism Spectrum Disorders (ASD)

C. Correia[1,2], Diekmann Y[1], Oliveira G[3], Pereira-Leal JB[1], A.M. Vicente[1,2] and the Autism Genome Project Consortium
1) Instituto Gulbenkian de Ciência, Portugal; 2) Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal; 3) Hospital Pediátrico de Coimbra, Portugal

INTRODUCTION

In contrast to the many rare variants that have been catalogued in families with ASDs, genome-wide association studies (GWAS) have thus far met limited success in the identification of common risk variants[1]. This suggests that ASD, like most complex diseases, may result from the accumulation and interaction of many genetic variants with small individual risk, which cannot be detected in current GWAS in a single SNP analysis framework[2].Therefore, alternative analytic approaches to SNP or haplotype-based methods must be developed to increase the power of GWAS, shifting the focus from individual markers to the study of the cumulative effect of multiple genes acting on the same biological process.
In the last decade the exceptional growth of molecular interaction data enabled the development of network-based approaches, in which there are no predefined gene sets, in contrast to pathway-based approaches[3]. The integration of interaction data with high throughput expression data has been used successfully through the identification of biologically meaningful subnetworks with a high prediction performance[4], but the application of this strategy to other types of high throughput data such as GWAS, is still very limited[5].

OBJECTIVE: To indentify biological networks associated with ASD risk, and with predictive value for autism diagnosis, we have applied a network-based approach to the Autism Genome Project (AGP) consortium GWAS.

MATERIALS AND METHODS

- 2258 Probands and both parents (trios) were collected, as part of the Autism Genome Project (AGP) Consortium genome analysis project, as previously reported[6], and genotyped in the Illumina 1M SNP and the 1M duo SNP arrays. A total of 732 781 SNPs, passing the QC details described elsewhere[6], were tested for association using the Transmissions Disequilibrium Test (TDT) implemented in PLINK v1.07.
- A large human protein-protein interaction (PPI) network, consisting of 12372 proteins and 58365 interactions, was built compacting data from six public PPI databases: BIND, BioGRID, HPRD, IntAct, MINT, MPIDB and MPPI. SNPs were assigned to genes according the Refseq gene coordinates (NCBI NGRch37) +/-10 kb upstream and downstream, and mapping to the PPI was performed converting Entrez gene IDs to Uniprot IDs (release 2010_04).
- Nearest neighbor shortest path length, percentage of direct interactions and size of largest connected component between proteins within autism sets were calculated using python module Network X. Statistical significance of the calculated properties was assessed using 1000 samples of random proteins from the human PPI network of size equal to the autism sets. Functional enrichment analysis was performed using DAVID.
- A sample of 943 ASDs families (4,444 subjects) from the Autism Genetic Resource Exchange genotyped using the Illumina HumanHap550 BeadChip[7] was used for replication purposes. Only SNPs in common between this dataset and the AGP dataset and passing the same QC criteria were tested (425 280 SNPs).

RESULTS

No SNPs reaching genome-wide significance and very few with $P<1x10^{-5}$ were identified in single SNP analysis (TDT) (Figure 1)

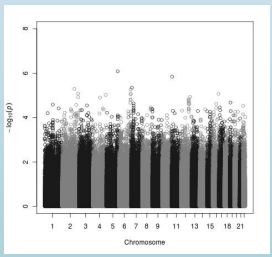


Figure 1- Manhattan Plot showing the –log10 (*P*-values) of SNPs passing the quality control criteria from the TDT analysis of the AGP sample.

Hypothesis: There are variants with small effect, confined to a limited number of biological pathways, which are not detected by TDT

1. **Define the *P*-value threshold** for which we can infer biology from the data. Using sets of proteins encoded by genes where SNPs at different *P* thresholds map, we determined the *P*-value for which the size **of largest connected component (LCC)** (group of interconnected proteins in the network) in our data **is significantly larger than in random samples** (Table 1; Figure 2).

RESULT 1: Threshold selected is TDT *P*-value<0.01

Table 1. Number of SNPs within genes (according to the definition used) selected at different *P*-value thresholds (0.1 to $1x10^{-8}$) and the corresponding number of genes, proteins and nodes in the global Protein-Protein Interaction Network (PPI)

Log_{10} *P*-value	-1	-1,5	-2	-2,5	-3	-3,5	-4	-4,5	-5	-5,5	<-5.5
# SNPs	44157	14692	4955	1675	543	175	69	25	9	1	0
# Genes	9502	4729	2059	860	323	115	45	17	7	1	0
# Proteins	8802	4434	1958	812	308	111	43	15	7	1	0
# Proteins in PPI	5048	2592	1138	462	181	61	26	12	6	1	0

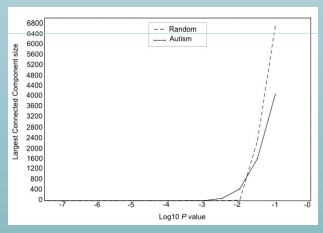
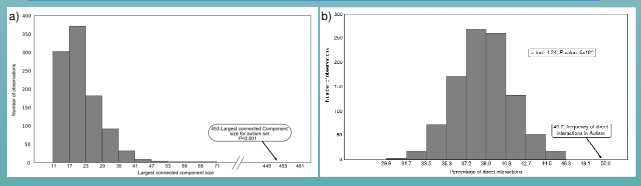


Figure 2- Largest connected component size determined for sets of proteins derived from SNPs selected at different TDT *P*-value thresholds in the AGP sample (solid line) and in random samples (dashed line).

2. Using a TDT *P*<0.01, what are the properties of the network derived by these sets of proteins? What does this indicate? (Figure 3)

RESULT 2: 453 proteins are interconnected in the LCC (*P*<0.001) and 50% of proteins directly interacting (*P*=$5x10^{-5}$)-» indicates that ASD- associated proteins are involved in a small number of interconnected biological processes and preferentially directly interact with each other

Figure 3 – a) Histogram of the sizes of the largest connected component in 1000 random sets of interacting proteins. The observed size for autism set is shown by an arrow. An Empirical *P*-value was calculated. b) Histogram of the frequencies of direct interactions in 1000 random set of proteins. The relative frequency observed for autism set is shown by an arrow. Result of the z-test is indicated.



3. What are the pathways, cellular compartments and expression tissues enriched in the largest component of interacting autism genes? (Table2)

RESULT 3: This group of interconnected proteins (LCC) is mainly enriched in pathways related to cell adhesion, localized in the synaptic compartment and expressed in the brain.

4. Can we replicate our results in an independent GWAS dataset ? (Table2)

RESULT 4: 219 proteins are interconnected in the LCC (*P*<0.001) and 39.4%% of proteins directly interacting (*P*=0.008) from Wang *et al* dataset.
2 of the top 5 pathways enriched in the AGP LCC, namely pathways related to cancer and focal adhesion, also ranked high in the Wang dataset. The most enriched cellular components and tissue expression are very similar between the two datasets.

	AGP dataset					Wang et al dataset				
Term	Count	%	*P*-Value	Benjamini	Rank	Count	%	*P*-Value	Benjamini	Rank
Canonical Pathway										
hsa04520:Adherens junction	16	3,5	5.53E-7	6.85E-5	1	6	2,7	0,019	0,068	28
hsa05200:Pathways in cancer	34	7,5	2.11E-6	1.31E-4	2	20	9,1	2.39E-5	3.46E-4	6
hsa04510:Focal adhesion	24	5,3	1.07E-5	4.41E-4	3	17	7,8	2.10E-6	3.66E-5	5
hsa04730:Long-term depression	12	2,6	1.31E-4	0,004	4	6	2,7	0,012	0,056	21
hsa04514:Cell adhesion molecules (CAMs)	16	3,5	4.21E-4	0,010	5	-	-	-	-	-
GO cellular compartment										
GO:0044459~plasma membrane part	126	27,8	5.99E-16	2.44E-13	1	64	29,2	1.23E-8	4.94E-7	7
GO:0042995~cell projection	60	13,2	2.27E-14	5.00E-12	2	33	15,1	3.93E-9	3.69E-7	3
GO:0043005~neuron projection	38	8,4	2.01E-12	2.95E-10	3	27	12,3	2.87E-12	8.09E-10	1
GO:0005886~plasma membrane	169	37,3	5.00E-12	5.49E-10	4	90	41,1	3.57E-8	1.12E-6	9
GO:0045202~synapse	37	8,2	2.70E-11	2.37E-9	5	26	11,9	4.14E-11	5.84E-9	2
Tissue Expression										
Brain	262	57,8	1.97E-11	4.44E-9	1	123	56,2	1.33E-5	0,002	1
Fetal brain	47	10,4	2.71E-8	3.04E-6	2	22	10,0	2.87E-4	0,024	2
Platelet	33	7,3	5.13E-6	3.84E-4	3	16	7,3	0,002	0,063	5
Epithelium	97	21,4	1.33E-5	7.45E-4	4	49	22,4	5.64E-4	0,032	3
Skeletal muscle	32	7,1	2.65E-5	0,001	5	17	7,8	9.81E-4	0,041	4

Table 2. Canonical Pathways, GO cellular compartment and tissue expression enrichment analysis of the proteins included in largest connected components calculated for the AGP and the Wang *et al* datasets.

CONCLUSIONS

- Using the AGP GWAS we showed that network analysis is an effective strategy to uncover, from a GWAS, low effect sizes *loci* that can not be detected using single SNP analysis. Using a network property, called largest connected component we have shown that there are many relevant susceptibility *loci* with *P*<0.01 that are being overlooked and should be further explored.
- Our analysis have shown that autism-associated proteins are functionally related, preferentially directly interact with each other and are involved in a small number of interconnected biological processes, previously associated with autism. While cancer related pathways may result from a bias toward highly studied pathways, it is also known that many processes implicated in cancer, such as cellular proliferation or apoptosis are also important processes for nervous system development. These observations were replicated in a smaller GWAS publicly available[7].
- Identification of small subnetworks implicated in autism and predictive for diagnosis are underway.

REFERENCES:1- Betancur C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting (2011) *Brain Res.* 1380: 42-77 ; 2- Witte JS. Genome-wide association studies and beyond (2010).Annu Rev Public Health 31:9-20; 3- Barabasi et al. Network Medicine: a network based approach to human disease (2011) Nat Rev Genet 12(1):56-68; 4- Chuang et al. Network-based classification of breast cancer metastasis (2007) Mol Syst Biol 3:140; 5- Baranzini et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis (2009). Hum Mol Genet 18(11):2078-90; 6- Anney et al. A genome-wide scan for common alleles affecting risk for autism (2010) Hum Mol Genet 19(20):4072-82; 7-Wang et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders (2009) Nature 459 (7246):528-33

ACKNOWLEDGMENTS: We thank the Autism Genome Project and the Funding agencies, Autism Speaks, Medical Research Council, Health Research Board (Ireland), Genome Canada, and Hilibrand Foundation. C. Correia was supported by a fellowship from the Fundação para a Ciência e Tecnologia (SFRH/BPD/64281/2009)