Available online at www.sciencedirect.com
SCIENCE DIRECT
Methods 37 (2005) 274–279
METHODS
www.elsevier.com/locate/ymeth

Convergent functional genomics: A Bayesian candidate gene identification approach for complex disorders

B. Bertsch [a], C.A. Ogden [b], K. Sidhu [a], H. Le-Niculescu [a], R. Kuczenski [c], A.B. Niculescu [a,*]

[a] *Laboratory of Neurophenomics, Institute of Psychiatric Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[b] *Drexel University School of Medicine, Philadelphia, PA, USA*
[c] *Department of Psychiatry, University of California, San Diego, La Jolla, CA 92037, USA*

Accepted 15 March 2005

Abstract

Identifying genes involved in complex neuropsychiatric disorders through classic human genetic approaches has proven difficult. To overcome that barrier, we have developed a translational approach called Convergent Functional Genomics (CFG), which cross-matches animal model microarray gene expression data with human genetic linkage data as well as human postmortem brain data and biological role data, as a Bayesian way of cross-validating findings and reducing uncertainty. Our approach produces a short list of high probability candidate genes out of the hundreds of genes changed in microarray datasets and the hundreds of genes present in a linkage peak chromosomal area. These genes can then be prioritized, pursued, and validated in an individual fashion using: (1) human candidate gene association studies and (2) cell culture and mouse transgenic models. Further bioinformatics analysis of groups of genes identified through CFG leads to insights into pathways and mechanisms that may be involved in the pathophysiology of the illness studied. This simple but powerful approach is likely generalizable to other complex, non-neuropsychiatric disorders, for which good animal models, as well as good human genetic linkage datasets and human target tissue gene expression datasets exist.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Convergent functional genomics; Animal models; Human genetics; Gene expression; Candidate genes; Bayesian

1. Introduction

Neuropsychiatric disorders such as bipolar disorders and schizophrenia are complex, polygenic, with variable penetrance. Moreover, the imprecise phenotypical characterization inherent in current diagnostic classifications, such as DSM-IV and ICD-10, compounds the problem. As such, identifying candidate genes for them through classic genetic approaches has proven arduous. Linkage studies result in relatively broad peaks, on large chromosomal regions, with hundreds of genes in them [2,3].

The advent of microarray technology has permitted the non-hypothesis driven comprehensive profiling of gene expression changes in animal models as well as in human postmortem brains [5,9,10,14,15,18]. Moreover, in animal models in particular, specific aspects of the illnesses (endophenotypes) can be modeled by carefully chosen pharmacological treatments. However, taken by itself, the gene expression profiling approach suffers from the major caveat that it is unclear which of the gene expression changes observed are core to the pathophysiology being studied and which are epiphenomena and artefacts. This problem is particularly acute for human postmortem gene expression work [17].

To overcome the shortcomings in both the classic genetics and gene expression profiling approaches, we have developed a heuristic, translational discovery paradigm called Convergent Functional Genomics (CFG) [11,15,16]. The CFG paradigm cross-matches comprehensive animal model microarray gene expression data with human genetic linkage data, human postmortem gene expression data, and biological roles data. This Bayesian way of reducing uncertainty produces a short list of high probability candidate

* Corresponding author. Fax: +1 317 274 1365.
E-mail address: anicules@iupui.edu (A.B. Niculescu).

1046-2023/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2005.03.012

genes, pathways, and mechanisms for complex genetic disorders, such as neuropsychiatric disorders [19]. Bayesian strategies have also been used fruitfully to integrate independent datasets and lines of evidence in model organisms, such as yeast [8,22].

2. CFG methodology

2.1. Internal convergence

2.1.1. Agonist–antagonist animal model pharmacogenomic convergence paradigm (Fig. 1)

Signs and symptoms of psychiatric disorders can be mimicked by the use of street drugs (for example methamphetamines for bipolar disorder, PCP for schizophrenia). Gene expression changes in response to these *agonists* of the illness are of interest but may also comprise genes that have to do with neurotoxicity or other effects of the drug that might not be directly germane to the illness modeled. Gene expression changes in response to *antagonists* of the illness drugs that treat the illness (valproate for bipolar, clozapine for schizophrenia) are of interest, but again some of them may have to do with the side-effects or toxicities of the drug rather than with the therapeutic effects. We reasoned that the convergence of effects of both agonist and antagonist would identify a limited number of higher probability candidate genes. Moreover, by also using an agonist–antagonist co-treatment paradigm and identifying the genes that were not changed ("nipped in the bud") by co-treatment, we would have an additional powerful cross-validator. The convergence of changed by agonist, changed by antagonist, and nipped in the bud by co-treatment can be quite powerful and restrictive which makes it particularly useful for big gene expression datasets [19]. This approach is biased towards avoiding false positives even at the expense of having false negatives. We have termed these higher probability genes Category I candidate genes. Genes that are changed by both the agonist and the antagonist but are not nipped in the bud by co-treatment are termed Category II. Genes that are changed by either the agonist or the antagonist and nipped in the bud by co-treatment are Category III. Genes that are changed by either the agonist or the antagonist and not nipped in the bud by co-treatment are termed Category IV (Fig. 1).

The agonist/antagonist paradigm does not necessarily have to be exclusively pharmacogenomic. At least one arm of it can be genetic (selected rodent strain, knock-out). Our approach can also be applied profitably to non-neuropsychiatric complex disorders, such as asthma and hypertension [4,6].

2.1.2. Changes in multiple target tissue regions

It can be argued that if a gene is changed in multiple tissue regions of interest, it is more likely to be involved in the pathophysiology of the illness. Moreover, from a technical standpoint, the chippings from different tissue regions are independent experiments, which gives a dimension of replicability, always important and reassuring in microarray work. It is also very important to do the (animal model) biological experiments multiple times, de novo. In our experience [19], and that of others in the field [12], biological variability trumps technical variability or method used as the main factor for non-reproducibility of gene expression data.

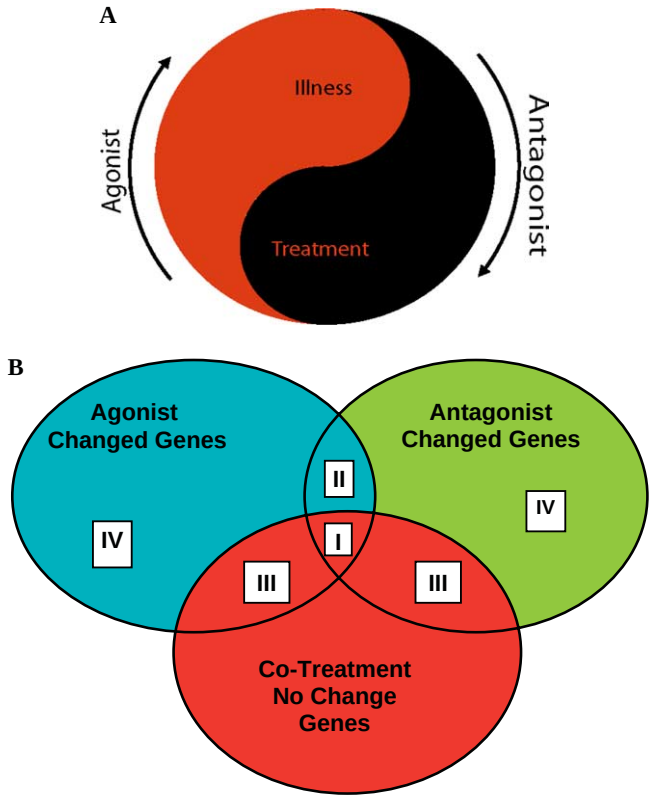

Fig. 1. Gene expression discovery engine. (A) Agonist–antagonist pharmacogenomic treatment paradigm; (B) Venn diagram categorizing genes changed by the various drug treatments and their classification into Categories I, II, III, and IV.

2.1.3. Gene expression data gene identification

A National Center for Biotechnology Information (NCBI) (Bethesda, MD: http://www.ncbi.nlm.nih.gov/) BLAST analysis of the accession number of each probe-set can be done to identify genes for which the Affymetrix database does not provide a definitive identification/name. BLAST analysis identifies the closest known gene existing in the database for the animal model species used (i.e., the highest known mouse gene at the top of the BLAST list of homologues) which then could be used to search the GeneCards database (Weizmann Institute, Rehovot, Israel: http://bioinfo.weizmann.ac.il/cards/index.shtml) to identify the human homologue. Probe-sets that do not have a known gene identified through this search remain labeled as "EST" and their accession numbers are kept as identifiers.

2.2. External convergence

2.2.1. Genetic linkage convergence

To designate convergence for a particular gene, the gene has to map within 10 centiMorgans (cM) of a microsatellite

marker for which linkage evidence to the (neuropsychiatric) disorder of interest has been reported in at least one published study [15]. The University of Southampton's sequence-based integrated map of the human genome (The Genetic Epidemiological Group, Human Genetics Division, School of Medicine, University of Southampton: http://cedar.genetics.soton.ac.uk/public_html/) is used to obtain cM locations for both genes and markers. The sex-averaged cM value is calculated and used to determine convergence to a particular marker. For markers that are not present in the Southampton database, the Marshfield database (Center for Medical Genetics, Marshfield, WI: http://research.marshfieldclinic.org/genetics/) is used to evaluate linkage convergence. Further information on specific gene function and biology can be obtained from the Johns Hopkins University database, Online Mendelian Inheritance of Man (http://www.ncbi.nlm.nih.gov/omim/).

The 10 cM distance was chosen initially because the length of linkage peaks for neuropsychiatric disorders is on average 20 cM [15]. As with other parameters in our approach, it can be varied to make the approach more restrictive (to avoid false positives) or less restrictive (to avoid false negatives).

2.2.2. Biological and postmortem convergence

Information about our candidate genes is obtained using GeneCards, as well as database searches using PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) and various combinations of keywords—gene name, tissue studied (example: brain), disease (example: bipolar, schizophrenia), and postmortem. Postmortem convergence is deemed to occur for a gene if there are published reports of human postmortem data showing changes in expression of that gene or protein product in brains from patients with that particular disorder. Genes are deemed to have biological convergence if their known biological function is relevant to the pathophysiology of the disorder studied, in human or animal models. Biological association with pharmacological agents used clinically to treat the disorder are particularly interesting. The search can be extended to closely related disorders, as there is often clinical and biological overlap. This would increase sensitivity but decrease specificity. More recently, lymphocyte gene expression profiles from patients have been published. They can constitute another line of evidence for convergence, and an interesting line of work for identifying peripheral biomarkers for neuropsychiatric disorders, where biopsies are not a practical option [13,21,23].

2.3. Candidate genes, pathways, and mechanisms

2.3.1. Filtering of the data and empirical scoring (Fig. 2)

Each line of evidence has parameters inside (e.g. p value of change for gene expression data), that can be set at more stringent or less stringent thresholds, depending if one desires more specificity or more sensitivity (Fig. 2).

Arguments can be made for different ways of weighing the impact and importance of the multiple internal and external lines of evidence. If one desires more sensitivity, then the internal lines of evidence provided by the gene

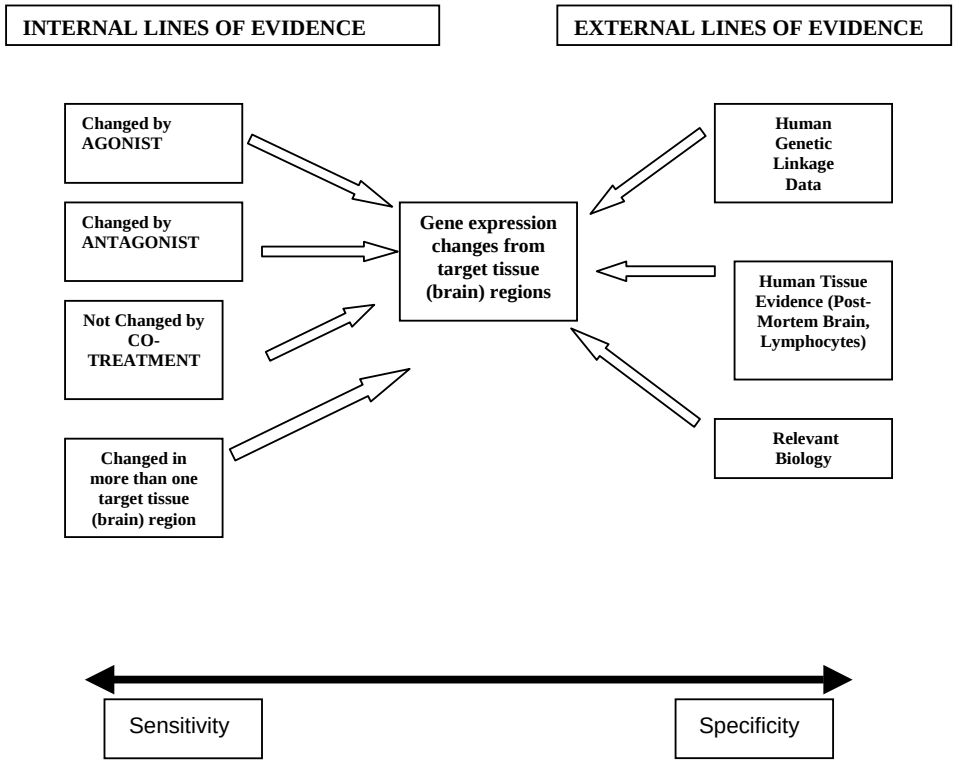

Fig. 2. Convergence. Multiple converging independent internal and external lines of evidence for Bayesian cross-validation of findings.

expression datasets should be weighed more heavily. If one desires more specificity, then the external lines of evidence provided by the human genetic linkage data, postmortem data or biological roles data should be weighed more heavily. We have used equal weighing in the work that we have published so far [19], assigning an empirical score of 1 to each independent internal and external line of evidence. With this simple approach, a pyramid of probability (Fig. 3) can be built for the genes in the dataset, with the highest probability candidate genes at the top.

It is very clear that any particular line of evidence in our approach can have caveats and uncertainties. The power of the approach derives from the Bayesian integration of multiple independent lines of evidence. As in a network with multiple nodes, even if one node becomes questionable or non-functional, the network overall has resilience and retains viability for its designed purpose.

2.3.2. Gene ontology analysis

The NetAffx Gene Ontology Mining Tool (Affymetrix, Santa Clara, CA: http://www.affymetrix.com/index.affx) can be used to categorize the genes in the different datasets into functional categories, using the Biological Process ontology branch. A simple hierarchy (gold, silver, and bronze) can be used to classify the different gene ontology (GO) categories, based on the number of Category I, II, III, and IV genes that they have (Fig. 4).

2.3.3. Pathways and mechanisms

The top candidate genes can be organized into networks of inter-relationships and pathophysiological mechanisms

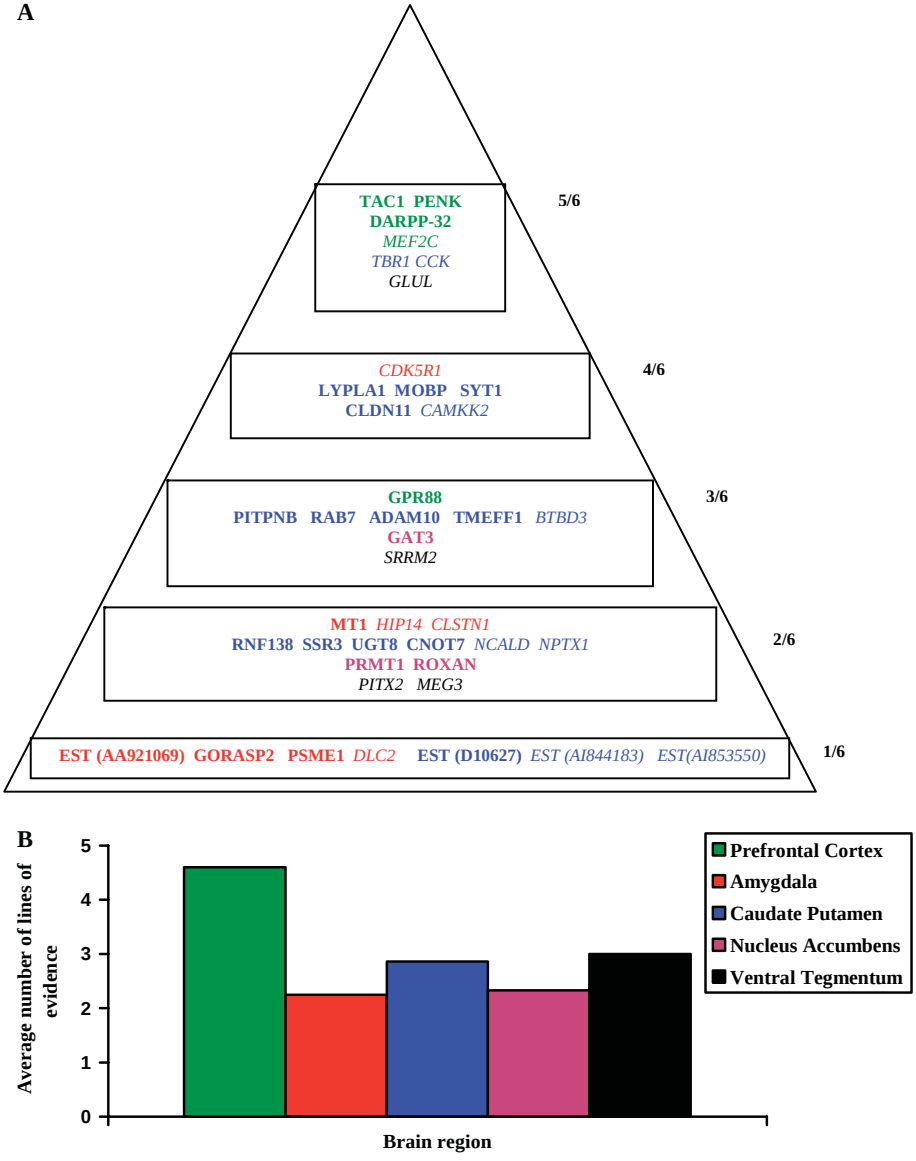

Fig. 3. Pyramid of top candidate genes for bipol disorder. (A) Probability pyramid generated by the tabulation of independent converging lines of evidence. Plain text—increased by methamphetamine; italics—decreased by methamphetamine; bold—(B) Comparison of different target brain regions in terms of average number of lines of evidence per candidate gene. (from Ogden et al. [19]).

		CATEGORIES					
		I	II	III METH	III VPA	IV METH	IV VPA
GO ANALYSIS – BIOLOGICAL PROCESSES –		NUMBER OF GENES					
1.	Cell communication	2	4	9	12	18	36
2.	Cell growth and/or maintenance	1	5	10	20	25	56
3.	Metabolism	2	9	12	19	35	72
4.	Morphogenesis	1	4	2	4	6	14
5.	Response to stress	1		2	1	5	1
6.	Response to external stimuli	1		2	2	7	2
7.	Reproductive behavior	1				1	
8.	Behavioral fear response	1					
9.	Cell motility		1	2	3	1	3
10.	Homeostasis		1	1			3
11.	Reproduction		1		1		2
12.	Pattern specification		1			1	1
13.	Embryonic development		1				
14.	Cell differentiation			2	2		3
15.	Learning and/or memory				1	1	2
16.	Cell death				1	1	2
17.	Death				1	1	2
18.	Circulation			4			1
19.	Rhythmic behavior				1		1
20.	Genetic transfer				1		
21.	Bone remodelling					1	1
22.	Hemostasis					1	
23.	Regulation of gene expression epigenetic					1	
24.	Membrane fusion						1
25.	Secretion						2

Fig. 4. Gene ontology analysis. Biological processes obtained from Gene ontology analysis of our complete dataset Categories I, II, III, and IV. Meth, methamphetamine; VPA, valproate. (from Ogden et al. [19]).

by exhaustive manual searches of the literature (PubMed). While this can be precise and exhaustive, it is impractical, especially for larger datasets, and commercial packages such as Ingenuity (Mountain View, CA: http://www.ingenuity.com) and others perform similar functions. With continuous improvements, it is hoped that they can be as reliable as the manual searches done by an expert, with the added power of computing and the useful graphical display interfaces of signaling pathways and networks. One informative feature of Ingenuity that we have used [19] is identifying which of the candidate genes in your dataset are the targets of existing drugs. For clinicians, that may provide an empirical starting point for rational polypharmacy (combination therapy).

3. Making sense of data

Genes that change together may work together in Co-Acting Gene Expression (CAGE) groups [11,15]. Moreover, these genes may be in epistasis with one another and thus provide a way of revisiting human genetic linkage datasets with empirically derived, testable hypotheses for epistatic interactions. Linkage peaks that were weak by themselves may become stronger when tested in conjunction with other peaks. Conversely, some of the epistatic interactions may be suppressive of each other. We have termed this approach of using gene expression data to unlock the secrets of epistasis EpiExpress.

4. Pursuing leads

Candidate genes can be further validated by studying the phenotype of transgenic mice in which the gene of interest is ablated (knock-out, siRNA) [20], or overexpressed. More definitive proof consists in demonstrating association of polymorphism in the gene with the illness in human genetic studies [1,7]. The final nail in the coffin should be evidence that those polymorphisms have functional significance.

References

[1] T.B. Barrett, R.L. Hauger, J.L. Kennedy, A.D. Sadovnick, R.A. Remick, P.E. Keck, S.L. McElroy, M. Alexander, S.H. Shaw, J.R. Kelsoe, Mol. Psychiatry 8 (5) (2003) 546–557.
[2] W. Berrettini, Neuromol. Med. 5 (1) (2004) 109–117.
[3] N. Craddock, M.C. O'Donovan, M.J. Owen, J. Med. Genet. 42 (3) (2005) 193–204.
[4] D.J. Erle, Y.H. Yang, Genome Biol. 4 (11) (2003) 232.
[5] S.J. Evans, P.V. Choudary, C.R. Neal, J.Z. Li, M.P. Vawter, H. Tomita, J.F. Lopez, R.C. Thompson, F. Meng, J.D. Stead, D.M.

Walsh, R.M. Myers, W.E. Bunney, S.J. Watson, E.G. Jones, H. Akil, Proc. Natl. Acad. Sci. USA 101 (43) (2004) 15506–15511.
[6] R.S. Friese, P. Mahboubi, N.R. Mahapatra, S.K. Mahata, N.J. Schork, G.W. Schmid-Schonbein, D.T. O'Connor, Am. J. Hypertens. 18 (5 Pt 1) (2005) 633–652.
[7] E. Hattori, C. Liu, J.A. Badner, T.I. Bonner, S.L. Christian, M. Maheshwari, S.D. Detera-Wadleigh, R.A. Gibbs, E.S. Gershon, Am. J. Hum. Genet. 72 (5) (2003) 1131–1140.
[8] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, Science 302 (5644) (2003) 449–453.
[9] S. Kaiser, L.K. Nisenbaum, Physiol. Genomics 16 (1) (2003) 1–7.
[10] S. Kaiser, L.A. Foltz, C.A. George, S.C. Kirkwood, K.G. Bemis, X. Lin, L.M. Gelbert, L.K. Nisenbaum, Neurobiol. Dis. 16 (1) (2004) 220–235.
[11] J.R. Kelsoe, A.B. Niculescu 3rd, CNS Spectr. 7 (3) (2002) 215–216, 223–226.
[12] J.E. Larkin, B.C. Frank, H. Gavras, R. Sultana, J. Quackenbush, Nat. Methods 2 (5) (2005) 337–344.
[13] F.A. Middleton, C.N. Pato, K.L. Gentile, L. McGann, A.M. Brown, M. Trauzzi, H. Diab, C.P. Morley, H. Medeiros, A. Macedo, M.H. Azevedo, M.T. Pato, Gene expression analysis of peripheral blood leukocytes from discordant sib-pairs with schizophrenia and bipolar disorder reveals points of convergence between genetic and functional genomic approaches, Am. J. Med. Genet., B Neuropsychiatr. Genet. May 12 (2005) [Epub ahead of print].
[14] K. Mirnics, F.A. Middleton, G.D. Stanwood, D.A. Lewis, P. Levitt, Mol. Psychiatry 6 (3) (2001) 293–301.
[15] A.B. Niculescu 3rd, D.S. Segal, et al., Physiol. Genomics 4 (1) (2000) 83–91.
[16] A.B. Niculescu 3rd, J.R. Kelsoe, Am. J. Psychiatry 158 (10) (2001) 1587.
[17] A.B. Niculescu, Genome Biol. 6 (4) (2005) 215.
[18] L.K. Nisenbaum, Genes Brain Behav. 1 (1) (2002) 27–34.
[19] C.A. Ogden, M.E. Rich, et al., Mol. Psychiatry 9 (11) (2004) 1007–1029.
[20] S.V. Rakhilin, P.A. Olson, A. Nishi, N.N. Starkova, A.A. Fienberg, A.C. Nairn, D.J. Surmeier, P. Greengard, Science 306 (5696) (2004) 698–701.
[21] R.H. Segman, N. Shefi, T. Goltser-Dubner, N. Friedman, N. Kaminski, A.Y. Shalev, Mol. Psychiatry 10 (5) (2005) 500–513, 425.
[22] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman, D. Botstein, Proc. Natl. Acad. Sci. USA 100 (14) (2003) 8348–8355.
[23] M.T. Tsuang, N. Nossova, et al., Am. J. Med. Genet. B Neuropsychiatr. Genet. 133 (1) (2005) 1–5.