CORTEX XXX (2014) 1–10

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach

Q11 Kristin K. Nicodemus [a,*], Brita Elvevåg [b,c], Peter W. Foltz [d,e], Mark Rosenstein [d], Catherine Diaz-Asper [f] and Daniel R. Weinberger [f,g,h]

[a] Neuropsychiatric Genetics Group, Department of Psychiatry, Trinity Centre for Health Sciences, Trinity College Dublin, St James Hospital, Dublin, Ireland
[b] Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Norway
[c] Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway
[d] Pearson Knowledge Technologies, Boulder, CO, USA
[e] Department of Psychology, University of Colorado, Institute for Cognitive Science, Boulder, CO, USA
[f] Clinical Brain Disorders Branch, National Institute of Mental Health/NIH, Bethesda, MD, USA
[g] Lieber Institute for Brain Development, Baltimore, MD, USA
[h] Departments of Psychiatry, Neurology, Neuroscience and The Institute of Genomic Medicine, Johns Hopkins School of Medicine, Baltimore, Q1 MD, USA

ARTICLE INFO

Article history:
Received 15 July 2013
Reviewed 01 October 2013
Revised 14 November 2013
Accepted 11 December 2013
Published online xxx

Keywords:
Cognition
Verbal learning and recall
Gene
Latent semantic analysis
Schizophrenia

ABSTRACT

Background: Category fluency is a widely used task that relies on multiple neurocognitive processes and is a sensitive assay of cortical dysfunction, including in schizophrenia. The test requires naming of as many words belonging to a certain category (e.g., animals) as possible within a short period of time. The core metrics are the overall number of words produced and the number of errors, namely non-members generated for a target category. We combine a computational linguistic approach with a candidate gene approach to examine the genetic architecture of this traditional fluency measure.

Methods: In addition to the standard metric of overall word count, we applied a computational approach to semantics, Latent Semantic Analysis (LSA), to analyse the clustering pattern of the categories generated, as it likely reflects the search in memory for meanings. Also, since fluency performance probably also recruits verbal learning and recall processes, we included two standard measures of this cognitive process: the Wechsler Memory Scale and California Verbal Learning Test (CVLT). To explore the genetic architecture of traditional and LSA-derived fluency measures we employed a candidate gene approach focused on SNPs with known function that were available from a recent genome-wide association study (GWAS) of schizophrenia. The selected candidate genes were associated with language and speech, verbal learning and recall processes, and processing speed. A total of 39 coding SNPs were included for analysis in 665 subjects.

Results and discussion: Given the modest sample size, the results should be regarded as exploratory and preliminary. Nevertheless, the data clearly illustrate how extracting the

* *Corresponding author*. Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity Centre For Health Sciences, Trinity College Dublin, St James's Hospital, Dublin 8, Ireland.
E-mail address: nicodemk@tcd.ie (K.K. Nicodemus).
0010-9452/$ – see front matter © 2014 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.cortex.2013.12.004

meaning from participants' responses, by analysing the actual content of words, generates useful and neurocognitively viable metrics. We discuss three replicated SNPs in the genes ZNF804A, DISC1 and KIAA0319, as well as the potential for computational analyses of linguistic and textual data in other genomics tasks.

1. Introduction

1.1. *Category fluency and schizophrenia: the role of the intermediate phenotype*

A complex combination of susceptibility genes and environmental factors is assumed to contribute to the overall clinical presentation of psychiatric disorders. Applying a reductionist approach to the diverse presenting phenomenology is not only daunting, but likely overlooks much of the associated deficits in the case of schizophrenia (but see Morar et al., 2011) where cognitive deficits are quite central to the neurodevelopmental course of the illness (Elvevåg & Weinberger, 2001). With such complex medical disorders one way to reduce the complexity of genetic effects is the 'intermediate phenotype' approach where it is argued that the putative risk genes should show greater effects at the intermediate level. Applied to psychiatry, this research strategy argues for bridging the gap between the emergent psychosis and the effects of genes on cells that directly modulate neurocognition (Goldberg & Weinberger, 2004; Meyer-Lindenberg & Weinberger, 2006; Tan, Callicott, & Weinberger, 2008). Such a research framework is appealing (but see Flint and Munafo (2007) for a different opinion), as the resulting intermediate phenotypes (e.g., working memory, episodic memory, semantic memory) are more amenable to systematic neurobiological research than the transient phenomenology (Elvevåg & Weinberger, 2009). Crucially, in psychiatric disorders it is at this intermediate phenotype level that genetic associations often show both stronger penetrance (Tan et al., 2008) and inheritance (Snitz, MacDonald, & Carter, 2006) than at the level of clinical diagnosis. Consequently, several major challenges emerge, namely the unavoidable required refinements to the intermediate phenotype and the management of the huge amount of data resulting from investigations of intermediate phenotypes.

Given the increasing importance of genome-wide association studies (GWAS) in neuropsychiatric research, it is increasingly apparent that intermediate phenotypes are potentially the means with which genomic discoveries will be made, but also may be limiting factors. Indeed, this new approach is magnitudes more complex than any enterprise embarked on hitherto in psychiatric genetics and arguably requires sophisticated phenotypes in order to unravel the complexities and thus eventually the pathologies within neural functional systems. Bilder and colleagues argue that cognitive ontologies need to be developed and refined to not only enable greater consistence and collaboration in research, but also to facilitate connections between intermediate phenotypes and genes (Bilder et al., 2009).

One crucial part of this puzzle is a modern cognitive neuroscientific re-operationalization of common psychometric concepts and terms. Here we focus on one of the most widely used neuropsychological tests – the category fluency task – to illustrate the current limitations of the 'verbal descriptions' of the underlying cognitive constructs and the issues that emerge when trying to explore the genetic architecture of the associated constructs. Specifically, the recall process likely involves a search for meanings as reflected in the 'clustering' of words in the output. Many approaches have been employed to examine the structure of the clustering, but are often problematic given the subjective judgements of cluster boundaries or have turned out to be simply unreliable (Voorspoels et al., in press). We have previously adopted LSA as an objective and reliable methodology to chart the flow of meaning in words and discourse (Elvevåg, Foltz, Weinberger, & Goldberg, 2007), and briefly describe this technique below. Our current motivation is that the 'content' of words has rarely been considered a useful candidate in investigations concerning genomics. This absence may be partially due to the notoriously subjective and labour intense efforts required in quantifying the content of words. However, advances in computational linguistics provide a viable framework within which the meanings of words can be rigorously investigated.

1.2. *Latent semantic analysis: building a semantic space*

Latent Semantic Analysis (LSA) is a statistical approach to the acquisition and representation of meaning, which allows similarities among the elements of a language (e.g., words, sentences, or passages) to be computed based on word co-occurrence patterns in large corpora of naturally produced discourse. LSA is a computational model of meaning that closely mimics human understanding of the contextual use of language, which has been widely used for information retrieval, machine understanding of text, and applications such as automated essay scoring (for an overview, see Landauer, Kintsch, McNamara, & Dennis, 2007). Unlike standard keyword-based methods, LSA can detect subtle aspects of semantic content. LSA has been widely used for cognitive modelling of learning and memory processes as well as for computing coherence in language and thought processes. The reduced dimension semantic representation from LSA allows comparison by computing the semantic similarity between individual terms or groups of terms (see Supplementary Methods for further details and an example).

In the case of the category fluency task, the total number of words produced has been shown to be an important metric and poor performance (i.e., production of substantially fewer words than expected based on demographically based

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

Table 1 – Candidate genes, SNPs and SNP functions.

Gene	SNP	Function
Language and speech ($N = 9$)		
ATP2C2	rs2303853	Gly411Ser
DCDC2	rs2274305	Ser221Gly
DYX1C1	rs600753	Glu191Gly
KIAA0319	rs807534	Tyr424/968/1004/1013Cys
KIAA0319	rs807541	Ala309/853/889/898Ala
KIAA0319	rs4576240	Thr97/133/142/Pro
NAGPA	rs887854	Asn495Asn
ZNF804A	rs1366842	Thr707Lys
ZNF804A	rs12477430	His747Arg
Episodic memory ($N = 14$)		
ADCY8	rs12545028	Arg523Arg
BDNF	rs6265	Val66/74/81/95/148Met
CAMK2G	rs2675671	Lys49Lys
CLSTN2	rs17348572	Ile331Thr
CLSTN2	rs7632885	Val366Ile
CLSTN2	rs10804675	Val847Val
COMT	rs4680	Val158Met
GRIN2B	rs3026160	Cys838Cys
HTR2A	rs6314	His368/452Tyr
HTR2A	rs6313	Ser34Ser
WWC1 (KIBRA)	rs17551608	Arg250Cys
WWC1 (KIBRA)	rs3822659	Ser735Ala
WWC1 (KIBRA)	rs3733980	Val801Val
WWC1 (KIBRA)	rs3203960	Leu1005/1011/Leu
Verbal fluency ($N = 5$)		
CACNA1C	rs1544514	Ala174Ala
DISC1	rs2492367	Ile119/469/501Ile
DISC1	rs6675281	Leu485/607/639Phe, Thr572Thr
DISC1	rs12133766	Leu499/621/653/Leu, 3′ UTR
SLC6A3 (DAT)	rs6350	Asn38Asn
Processing speed ($N = 11$)		
ATRNL1	rs10885721	Thr1203Thr
C20orf196	rs1699233	Gly107Gly
CRTC3	rs8033595	Ser72Asn
DIP2C	rs3740304	Tyr1551Tyr
DIP2C	rs2288681	Ala1274Ala
NFKBIL1	rs2230365	Ser103/126/Ser
PDE1C	rs3213709	Gly610/670/Gly
PDE1C	rs2302450	Ala594/654Ala
PDE1C	rs1860790	Asn591/651Asn
PKNOX1	rs234781	Thr114Thr
SPATA7	rs3179969	Val42/74Met

normative data) has been associated with a variety of clinical disorders, including schizophrenia (Bokat & Goldberg, 2003; Lezak, 1995). A possible common mechanism associated with less than optimal performance on this simple task relates to speed of performance, but there are many other components, namely language, speech, verbal learning and recall, semantic organization (Schwartz, Baldo, Graves, & Brugger, 2003), and fluency in general.

2. Methods

To explore the genetic architecture of traditional measures (e.g., number of valid words generated) and LSA-derived measures of verbal fluency (e.g., average vector length; measures described in Section 2.1) we adopted a candidate gene approach and focused on SNPs with known function that were available from genome-wide association SNP chips. Note that although for some SNPs the function is known based on the sequence of the DNA (e.g., whether there is an amino acid change), for most SNPs the result of this function on cognition is not known. Candidate genes associated with language and speech were selected: ATP2C2 (1 SNP), DCDC2 (1), DYX1C1 (1), KIAA0319 (3), NAGPA (1) (Graham & Fisher, 2013) and ZNF804A (2) (Becker et al., 2012), verbal learning and recall (as a subset of the concept of episodic memory): ADCY8 (1), BDNF (1), CAMK2G (1), CLSTN2 (3), COMT (1), GRIN2B (1), HTR2A (2), WWC1 (3) (Papassotiropoulos & de Quervain, 2011), verbal fluency: CACNA1C (1) (Krug et al., 2010), DISC1 (3) (Palo et al., 2007), and SLC6A3 (1) (Pauli et al., 2012) and processing speed: ATRNL1 (1), C20orf196 (1), CRTC3 (1), DIP2C (2), NFKBIL1 (1), PDE1C (3), PKNOX1 (1), SPATA7 (1) (Luciano et al., 2011) (Table 1). Previous evidence for association with cognition and evidence for potentially deleterious effects of non-synonymous SNPs on their resulting protein product identified using the "Sorting Tolerant from Intolerant" (SIFT) algorithm (Kumar, Henikoff, & Ng, 2009) are given in Supplementary Table 1, which provides *a priori* hypotheses on the direction of association for many SNPs selected.

A total of 39 coding SNPs were included for analysis. Note that the groupings of genes by association with phenotypes is not complete; indeed, genes investigated may have pleiotropic effects across phenotypes but were grouped by associations reported in the literature. For comparative purposes, we included two standard measures of verbal learning and recall: the logical memory test from the Wechsler Memory Scale-Revised (WMS-R; Wechsler, 1987) and the California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987).

2.1. *LSA candidate gene association study*

2.1.1. *Participants*

Healthy control participants ($N = 307$), patients with schizophrenia ($N = 194$) and their unaffected siblings ($N = 164$) were included, all of whom gave informed consent to participate in the Clinical Brain Disorders Branch National Institute of Mental Health 'Sibling Study' protocol, which is an U.S. investigation of neurobiological abnormalities related to genetic risk for schizophrenia (Weinberger DR, PI). All participants were screened by two board-certified psychiatrists using semi-structured psychiatric interviews, third-party informants, toxicology screening, and cognitive testing exclusions as previously described (Huffaker et al., 2009). All patients met DSM-IV criteria for schizophrenia or schizoaffective disorder, depressed type, all siblings were free from schizophrenia spectrum disorders, and all controls were free from DSM-IV lifetime psychiatric illness or substance abuse. All participants self-identified as Caucasian.

2.1.2. *Measures*

We used a category fluency task where a participant generated words in response to the cue 'animal' for 1 min. We transcribed the words produced during this task and these written records were used in computing measures of

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

coherence from the actual sequence of words (animals)[1] generated usingLSA (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). In the present work, a semantic space was derived from the commonly used TASA (Touchstone Applied Science Associates, Inc.) corpus, consisting of 44,486 documents by 98,646 unique terms and represented in 300 dimensions (see http://lsa.colorado.edu). Each term or group of terms is represented as vectors in this semantic space and the cosines between vectors for terms are used to measure the degree of semantic similarity between the terms. Typically a cosine close to 0 indicates no similarity, while a cosine close to 1 indicates high semantic similarity. These cosines closely match human ratings of similarity of meaning and have been empirically demonstrated in many contexts (see Landauer & Dumais, 1997; Landauer et al., 2007).

2.1.3. *Coherence measures*

To the extent that category fluency is a proxy of some aspect of language (likely the retrieval aspect) our theory-driven coherence measure assays search for meanings/associations during recall. *A priori*, we selected eight measures (listed below) based on theoretical motivations indicating that they would likely be sensitive assays to subtle differences in fluency tasks. Fluency (both letter and category) has been shown to be heritable (Aukes et al., 2008; Bratko, 1996; Sakakibara et al., 2013; Vandenberg, 1962). LSA may generate more (qualitative) information about fluency performance than traditional scoring methods alone, with the benefit of some LSA measures being mostly independent of differences in overall word production. Although we focused on the 'animal' category, the overall counts in 'animals' mirrored findings for all the categories combined.

2.1.3.1. The total number of valid words. This measure is a simple count of the number of valid words generated within 1 min. Only words related to the cue 'animal' were counted as valid terms.

2.1.3.2. *N* word sequence coherence ($N = 1$, 2 and 3). This is an index of word coherence averaged over moving window sets of size *N* sequential words. A moving window technique with window-size *N* starts with the first *N* words and compares the semantic similarity of the first word to the other $N - 1$ words in the window. The window then moves forward one word and the comparison is repeated with the next set of *N* words. This approach enables the exploration of the influences of previously generated words on determining the current utterance. Specifically, a moving window may be thought of as an assay with biological validity, as it is metric of a type of working memory measure in which the previously uttered word impacts the subsequent ones. Thus even though the actual category fluency task taps into some aspect of verbal learning and recall, the actual process of sequentially listing animals requires some type of working memory.

2.1.3.3. Vector length. This measure indexes unusualness, with low frequency words having a higher 'information value' (thus leading to a higher vector length). The vector length provides a measure of the semantic information value, with more complex terms typically having greater vector length. There are two ways this measure can be computed: (i) average vector length, and (ii) overall vector length. Average vector length is the average of all the vector lengths for the words, while overall vector length treats all the words as a context and computes the vector length of that context. The context vector length is computed by summing the scaled vectors for all the words in the context and taking the vector length of the resulting summed vector. Since additional words typically add "information value", overall vector length is influenced by word count whereas average vector length is independent of word count. Indeed, overall vector length may be thought of as a metric of the quality of the chunk of information retrieved. Furthermore, this measure has recently been shown to be related to disconnected speech and functional behaviour in a large sample of elderly patients with schizophrenia (Holshausen, Harvey, Elvevåg, Foltz, & Bowie, 2013).

2.1.3.4. Cosine to 'animal'. This is an index of how close/similar (in semantic space) the animal words generated are to the cue 'animals' as measured by the average of the cosines between the generated animal terms to the term 'animal'. Put differently, this is a measure of semantic proximity and cohesiveness.

2.1.3.5. Average cosine between all terms. This measure is the average of all $n^*(n-1)/2$ pairwise cosine similarities between the *n* valid words and describes the overall cohesion, with higher values indicating increased cohesion.

We used standard unpaired *t*-tests with unequal variance to test for differences in these measures between (1) controls and probands (schizophrenia patients) and controls and siblings and (2) males and females within each group (e.g., controls, probands and siblings). Controls generated significantly more valid words and thus had a longer resulting overall vector length (p-values $< .00625$, passing Bonferroni correction for the number of phenotypes tested) and average vector length and average cosine (p-values $< .01$) than probands (Table 2).

No significant differences were found between siblings and controls on any measure. We did not test for differences between probands and siblings as they are related thus violating the assumption of statistical independence. Sex-specific analyses are given in Supplementary Table 2. The average vector length was significantly longer in female controls than male controls (p-value $< .05$) and in female siblings versus male siblings (p-value $< .01$). Additionally, in siblings, males produced significantly more valid words than females (p-value $< .05$), and proximity to the cue 'animal' was significantly higher in female siblings than male siblings (p-value $< .05$). No significant differences were observed between male and female probands on any of these measures.

We related our novel measures to the number of valid words via correlation coefficients (Table 3). Even though all *p*-

[1] We examined one category only because there is considerable blurring of semantic boundaries between the other two categories, namely fruits and vegetables (e.g., an avocado and tomato are examples of fruits, but they are often generated as exemplars of the vegetable category) and consequently the semantic search process can be expected to be somewhat more complex.

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

Table 2 – Means and standard deviations for number of valid words and LSA-derived phenotypes by group.

Phenotype	Controls (N = 471)	Probands (N = 194)	Siblings (N = 164)
Number of valid words	21.97 (4.97)	16.13 (5.18)**	21.13 (4.62)
Average vector length	.062 (.013)	.067 (.019)*	.063 (.012)
Overall vector length	.51 (.086)	.44 (.11)**	.50 (.079)
Cosine to 'Animal'	.14 (.030)	.14 (.042)	.15 (.029)
Average cosine	.096 (.026)	.11 (.037)*	.099 (.026)
Cohesion (window-size = 1)	.18 (.050)	.18 (.062)	.19 (.053)
Cohesion (window-size = 2)	.12 (.046)	.12 (.053)	.12 (.049)
Cohesion (window-size = 3)	.10 (.041)	.10 (.056)	.10 (.041)

*t-test comparing controls with either probands or siblings, *p*-value <.01.
**t-test comparing controls with either probands or siblings, *p*-value <.00625 (passing Bonferroni critical value for 8 tests).

values for the correlations were <.05 (uncorrected) for controls and siblings (with the exception of cohesion with a window-size of 2 in siblings), we can see from the strength of correlation that the novel metrics are associated with the number of valid words but not completely dependent on it.

Interestingly, probands showed no association between the number of valid words generated and cohesion, as measured by the cosine to 'animal', average cosine or average vector length measures (Table 3), whereas in controls and siblings the greater number of words generated the less cohesive the set of words were to each other and the fewer unique words were used. Correlations among all phenotypic measures showed significant correlation among most LSA-derived measures, but only the number of valid words was significantly correlated with WMS logical memory 1 and 2, and WMS measures were significantly correlated with CVLT measures but not with the LSA-derived measures (Supplementary Table 3), indicating the LSA-derived measures may measure different aspects of cognition.

2.1.4. *Genotyping and quality control*

DNA was extracted from peripheral blood leukocytes using standard methods. Genotyping was performed in waves using the Illumina HumanHap 550K/610/660K Quad, the Illumina HumanOmni 2.5M Quad and the Illumina HumanOmni 2.5M-8 chips. The overlap between these 5 genotyping platforms was used for the present study. Quality control (QC) was performed after genotype-calling from the intensity plots using the Illumina GenomeStudio (Illumina, version 2010.1). SNPs with minor allele frequencies < 1%, high missing rates (>5%) and deviation from Hardy Weinberg expectation (*p*-value < .0001) were removed. Individuals were removed from analysis if their genotyping rate was below 97%. Sex-checks, sample duplications and cryptic relatedness were examined by identity-by-state analysis of autosomal chromosomes. If the IBS sharing coefficient for unrelated individuals was >.10 one individual was randomly selected to be retained for analysis and the other related individual was removed from further analysis. After QC, 665 individuals remained with valid genotype and phenotype values. After merging all SNP platforms and taking the consensus across all platforms, 278,675 SNPs remained for analysis, of which as detailed in the Introduction we selected 39 for the present study (see Table 1 for details).

Analyses were conducted in controls and any SNP showing association with verbal fluency at the uncorrected *p*-value .05 level was tested separately in probands and their healthy siblings as replication samples. Due to the exploratory nature of our study and the small sample sizes, we used uncorrected *p*-values and relied on replication in independent samples of siblings or probands. Sex-stratified analyses were conducted in controls (142 males and 165 females) with the same replication strategy applied to probands (159 males and 35 females) and siblings (66 males and 98 females). All analyses were performed in PLINK v. 1.07 (Purcell et al., 2007) or using the R Statistical Computing Environment (R Development Core Team, 2011). The genomic inflation factor (GC lambda) in controls was estimated from a full GWAS. Across all phenotype measures the largest genomic inflation factor was 1.02, suggesting very little evidence for population stratification in our study. Experiment-wise empirical *p*-values were computed by repeating the entire experiment 1000 times using 1000 null replicates where each phenotype was permuted within the group at hand. To determine the empirical *p*-value for replication between control and siblings or probands, we counted the number of times, per 1000 replicates, the *p*-value was <.05 for controls and then for the replication group for the same SNP and same phenotype with the same direction of association with the exception of SNPs in ZNF804A, which have consistently shown less impaired cognition in cases who carry the allele that increases risk for schizophrenia (see Discussion).

Table 3 – Correlations of LSA-derived phenotypes with number of valid words.

Phenotype	Controls		Siblings		Probands	
	r^2	*p*-value	r^2	*p*-value	r^2	*p*-value
Cosine to animal	−.12	**.0093**	−.19	**.0012**	−.038	.53
Average cosine	−.20	**$2.22e^{-05}$**	−.26	**$8.50e^{-06}$**	−.086	.16
Average vector length	−.33	**$1.35e^{-12}$**	−.38	**$2.06e^{-11}$**	.026	.67
Overall vector length	.56	**$p < 2.2e^{-16}$**	.53	**$p < 2.2e^{-16}$**	.62	**$p < 2.2e^{-16}$**
Cohesion (window-size = 1)	.24	**$3.70e^{-07}$**	.18	**.0028**	.32	**$1.28e^{-07}$**
Cohesion (window-size = 2)	.13	**.0062**	.075	.20	.25	**$4.0e^{-05}$**
Cohesion (window-size = 3)	.10	**.034**	.16	**.0063**	.29	**$2.39e^{-06}$**

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

Table 4 – Candidate gene associations with LSA-derived variables.

Gene class	Phenotype	Gene	SNP	Chromosome	Beta	*p*-value	Group
Verbal fluency	Overall vector length	DISC1	rs6675281	1	.022	.033	All controls
Verbal fluency	Average vector length	SLC6A3	rs6350	5	4.0E-03	.029	All controls
Verbal fluency	**Average vector length**	**DISC1**	**rs12133766**	**1**	**−.0093**	**.018**	**Male controls**
Verbal fluency[a]	*Average vector length*	*DISC1*	*rs12133766*	*1*	*−.0085*	*.049*	*Male probands*
Verbal fluency	Cohesion (WS = 1)	DISC1	rs12133766	1	−.029	.028	Male controls
Episodic memory	Average vector length	WWC1	rs17551608	5	−.0028	.045	All controls
Episodic memory	Number of valid words	WWC1	rs17551608	5	1.38	.011	All controls
Episodic memory	Number of valid words	WWC1	rs17551608	5	1.73	.0062	Female controls
Episodic memory	Cohesion (WS = 2)	WWC1	rs3733980	5	−.013	.043	Male controls
Episodic memory	Number of valid words	KIAA0319	rs807534	6	−1.45	.032	All controls
Episodic memory	**Number of valid words**	**KIAA0319**	**rs807534**	**6**	**−1.74**	**.032**	**Female controls**
Episodic memory	*Number of valid words*	*KIAA0319*	*rs807534*	*6*	*−1.67*	*.041*	*Female siblings*
Episodic memory	Number of valid words	KIAA0319	rs807541	6	−1.35	.028	Female controls
Episodic memory	Overall vector length	KIAA0319	rs4576240	6	.041	.046	Male controls
Episodic memory	Overall vector length	CLSTN2	rs10804675	3	.031	.019	Male controls
Episodic memory	Overall vector length	CLSTN2	rs10804675	3	−.028	.0026	Female controls
Episodic memory	Average vector length	CLSTN2	rs10804675	3	−.0033	.015	Female controls
Episodic memory	Cohesion (WS = 1)	CLSTN2	rs10804675	3	−.015	.010	Female controls
Episodic memory	Cohesion (WS = 2)	BDNF	rs6265	11	−.014	.023	Male controls
Episodic memory	**Number of valid words**	**ZNF804A**	**rs1366842**	**2**	**1.31**	**.042**	**Male controls**
Episodic memory	*Number of valid words*	*ZNF804A*	*rs1366842*	*2*	*−1.12*	*.033*	*Male probands*
Episodic memory	Average cosine	ZNF804A	rs1366842	2	−.0064	.042	Female controls
Episodic memory	Cohesion (WS = 3)	HTR2A	rs6314	13	−.017	.033	Female controls
Episodic memory	Cohesion (WS = 1)	DCDC2	rs2274305	6	−.013	.035	Male controls
Language	Cohesion (WS = 2)	DYX1C1	rs600753	15	.010	.0062	All controls
Language	Cohesion (WS = 2)	DYX1C1	rs600753	15	.011	.040	Male controls
Language	Cosine to 'Animal'	DYX1C1	rs600753	15	.0051	.048 (.055)	All controls
Language	Cosine to 'Animal'	DYX1C1	rs600753	15	−.0082	.022	Female controls
Language	Cohesion (WS = 1)	ATP2C2	rs2303853	16	.033	.032	Male controls
Processing speed	Average cosine	ATRNL1	rs10885721	10	−.0053	.013	All controls
Processing speed	Average cosine	ATRNL1	rs10885721	10	.0065	.038	Female controls
Processing speed	Average cosine	PKNOX1	rs234781	21	.010	.0063	All controls
Processing speed	Average cosine	PKNOX1	rs234781	21	.011	.039	Female controls
Processing speed	Average vector length	PDE1C	rs3213709	7	.0027	.032	All controls
Processing speed	Cohesion (WS = 1)	PDE1C	rs3213709	7	.010	.025	All controls
Processing speed	Cohesion (WS = 1)	PDE1C	rs2302450	7	.011	.017	All controls
Processing speed	Cohesion (WS = 1)	PDE1C	rs2302450	7	.020	.0024	Male controls
Processing speed	Cohesion (WS = 1)	SPATA7	rs3179969	14	.011	.010	All controls
Processing speed	Cohesion (WS = 1)	SPATA7	rs3179969	14	.018	.0050	Male controls

[a] Significant replication sample results are given in italics.

2.2. *Candidate gene results for LSA metrics*

2.2.1. *Candidate genes associated with LSA measures*

Two genes selected to be linked with verbal fluency, DISC1 (2 SNPs) and SLC6A3 (1 SNP), were associated in all controls and male controls and probands with measures of overall and average vector length and cohesion with a window-size of 1 (Table 4 lists all association tests with a *p*-value < .05, uncorrected).

Intriguingly, DISC1 rs12133766, which is a synonymous change (Leu499/621/653), was negatively associated with average vector length in male controls (β = −.0093, *p*-value = .018) and this negative association was replicated in male probands (β = −.0085, *p*-value = .049), with an experiment-wise empirical *p*-value of .051. This result indicates that males carrying a copy of the minor allele at rs12133766 used less complex terms in response to the cue 'animal' than those who did not carry a copy of the minor allele, regardless of the number of words generated. No association was observed for this SNP in combined male and female controls or in female controls individually (all *p*-values > .05). Verbal learning and recall-associated genes such as WWC1 (2 SNPs), KIAA0319 (3 SNPs), CLSTN2 (1 SNP), BDNF (1 SNP), ZNF804A (1 SNP), HTR2A (1 SNP) and DCDC2 (1 SNP) showed association with the number of valid words generated, cohesion of words generated across all 3 window-sizes, overall and average vector length and average cosine (Table 4). SNP rs807534, a missense SNP (Tyr424/968/1004/1013Cys) in the gene KIAA0319, was negatively associated with the number of valid words generated in female controls, with individuals carrying the minor allele generating 1.74 fewer words on average (β = −1.74, *p*-value = .032). The negative association with the minor allele at this SNP was able to be replicated in female siblings, with individuals carrying the minor allele generating 1.67 fewer words (β = −1.67, *p*-value = .041), with an experiment-wise *p*-value of .049. The SNP rs1366842 in ZNF804A is a missense SNP (Thr707Lys) and was positively associated with the number of valid words

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

generated in male controls, with males carrying the minor allele generating 1.31 additional words than those carrying major alleles ($\beta = 1.31$, p-value = .042) but *negatively* associated with the number of valid words generated in male probands, thus that carriers of the minor allele generated, on average, 1.12 fewer words than those carrying major alleles ($\beta = -1.12$, p-value = .033), experiment-wise empirical p-value of .045. Two genes previously associated with language and speech, DYX1C1 (1 SNP) and ATP2C2 (1 SNP), showed association with cohesion and cosine to the word 'animal' (p-values ranged from .0062 to .048), although these associations were not replicated in probands or siblings. Of the genes reported to show association with processing speed, ATRNL1 (1 SNP), PKNOX1 (1 SNP), PDE1C (2 SNPs) and SPATA7 (1 SNP) showed significant uncorrected association with cohesion, average cosine and average vector length (p-values ranged from .0024 to .039). However, none of these associations were replicated in probands or siblings.

Across the genes grouped by phenotype (e.g., "Gene Class" in Table 4), the number of associations in controls (including overall and sex-stratified analyses) with a p-value < .05 in the group of genes previously associated with verbal learning and recall was 17, which was not significantly larger than expected by chance (empirical p-value > .05). However, within the group of genes previously associated with verbal learning and recall, the phenotype most frequently associated was the number of valid words (6/17, 35%), which was significantly more frequently associated than expected by chance within all replicates with 17 or greater associations (empirical p-value = .0080). The number of associations within the group of genes previously associated with processing speed was not greater than expected by chance (empirical p-value > .05). However, SNPs within this group were associated primarily with cohesion with a window-size of 1 (5/10, 50%; empirical p-value = .0040) and average cosine (4/10, 40%; empirical p-value = .028). Genes previously associated with language and speech were associated most frequently with cohesion with a window-size of 2 (2/5, 40%) and cosine to 'animal' (2/5, 40%), but neither the clustering with these phenotypes nor the overall number of associations was significant (empirical p-values all > .05). Similarly, genes previously associated with verbal fluency in controls were more likely to be associated with average vector length (2/4, 50%), although the number of associations with phenotypes and the frequency of association with average vector length did not exceed that expected by chance (all empirical p-values > .05).

In order to assess whether the replicated associations observed in DISC1 with average vector length and KIAA0319 and ZNF804A with number of valid words generated captured association with variation in these novel measures that was not accounted for by traditional measures of verbal learning and recall, we tested for association between these 3 SNPs and traditional measures of verbal learning and recall: the CVLT 15 correct, the CVLT long delay free recall correct, and the WMS-R logical (verbal learning and recall) memory 1 (immediate recall) and 2 (30 min delayed recall), plus the difference between logical memory 1 and 2. No significant association was found between DISC1 rs12133766 and these measures in male controls (p-value range = .11–.56) or in male probands (p-value range = .22–.98), nor between KIAA0319 rs807534 in female controls (p-value range = .31–.84) or female siblings (p-value range = .25–.86), nor between ZNF804A rs1366842 in male controls (p-value range = .080–.73) or in male probands (p-value range = .30–.81).

3. Discussion

We have shown that genes previously associated with verbal fluency (DISC1) and verbal learning and recall (ZNF804A and KIAA0319) were associated and replicated using traditional measures of category fluency (e.g., the number of valid words generated to the word 'animal') and also to a novel LSA-derived measure of average vector length, which is a measure of the quality of information retrieved. We further found that the genes associated with verbal learning and recall were significantly more frequently associated with the number of valid words generated than expected by chance alone. Genes previously associated with processing speed were significantly more frequently associated with cohesion of words generated and with average cosine. Further, the 3 SNPs in DISC1, ZNF804A and KIAA0319 were not significantly associated with traditional measures of verbal learning and recall using logical memory from the Wechsler Memory Scale or the CVLT, suggesting the novel computational linguistics metrics may measure components of category fluency that differ from more traditionally-used measures and may provide a more nuanced phenotype than the crude count of number of words produced. Although associations observed in female controls were replicated in female siblings and associations observed in male controls were replicated in male probands, the inability of our study to replicate associations in female probands ($N = 35$) and male siblings ($N = 66$) is likely due to the small sample sizes in these groups. The effect sizes observed were modest, similar to those found in other cognitive SNP association studies (e.g., Luciano et al., 2011). As is common in genome-wide association studies, our focus was on independent replication of our associations observed in controls.

3.1. *ZNF804A*

The GWAS-identified ZNF804A SNP rs1344706 (O'Donovan et al., 2008) has shown replicated increased risk for schizophrenia in several studies including a recent meta-analysis that reported a genome-wide significant level of association (Williams et al., 2011). Further studies have linked ZNF804A to neurocognitive measures including verbal working memory and episodic working memory (e.g., Walters et al., 2010). However, within schizophrenia cases, the schizophrenia-associated risk allele has shown relatively preserved cognition (Becker et al., 2012; Chen et al., 2012; Van Den Bossche et al., 2012; Walters et al., 2010), whereas in healthy controls the same schizophrenia-associated risk allele appears to impart poorer performance (Balog, Kiss, & Keri, 2011; Esslinger et al., 2011; Lencz et al., 2010; Voineskos et al., 2011). The SNP in our study, rs1366842, is approximately 10 kb from rs1344706 and is in modest linkage disequilibrium with this SNP (HapMap, CEU and TSI populations, $r^2 = .40$). However, given incomplete linkage disequilibrium, (the risk-associated A allele at rs1344706 occurs with the minor allele A of rs1366842

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

less frequently (approximately 20% in HapMap CEU and TSI populations) than the major allele at rs1366842 (approximately 38% of chromosomes)), the effect of rs2366842 may represent an independent effect on cognition in probands and controls. Consistent with previous reports of preserved cognitive function in probands but impaired cognitive function in controls according to the risk allele of the genome-wide significant SNP rs1344706, we also observed differential association with rs1366842 and the number of valid words generated to the cue 'animal', with controls carrying the minor allele at this SNP generating significantly more words and with cases generating significantly fewer words than those not carrying the minor allele, respectively.

3.2. *DISC1*

DISC1 is a schizophrenia and major depression candidate gene that was discovered in a large family, where a balanced translocation between chromosomes 1 and 11 segregated with psychiatric illness (Millar et al., 2000). As a scaffold protein, DISC1 has been shown to interact with multiple genes that are associated with neuronal migration and neurodevelopment (Porteous, Millar, Brandon, & Sawa, 2011). A recent study reported association between a rare-variation burden score within DISC1 and the Moray House Test, a test of verbal reasoning and general cognition, in a large population-based cohort from Scotland (Thomson et al., 2013). In bipolar disorder families from Finland, several SNPs and haplotypes in DISC1 were found to be associated with bipolar disorder and with various neurocognitive traits, including general intellectual functioning, verbal ability, verbal working memory and verbal learning, along with category fluency (Palo et al., 2007). Interestingly, the association with category fluency observed in this study (Palo et al., 2007) was with the SNP rs821616, which is located 190,453 bp from our DISC1 SNP (rs12133766) and thus our findings are likely to represent an independent signal within DISC1. Variation in DISC1 has also been associated with reductions in gray matter volume in the hippocampus and poorer performance on the logical memory 2 subtest from the Wechsler Memory Scale in healthy subjects (Callicott et al., 2005) and also with gray matter thickness, memory and cognitive processing (Carless et al., 2011).

3.3. *KIAA0319*

KIAA0319 encodes a protein considered important for neuronal adhesion/attachment and has been linked to a range of reading-related traits and dyslexia (Venkatesh, Siddaiah, Padakannaya, & Ramachandra, 2013; for a review see Graham & Fisher, 2013) which is a highly heritable condition in which genetic factors likely contribute up to 75% of the variance in the phenotype (DeFries, Fulker, & LaBuda, 1987). Importantly, a recent meta-analysis that integrated case–control and transmission/disequilibrium test studies supported the role of this gene in the risk of dyslexia (Zou et al., 2012). Recent work with rats (who have almost identical speech discrimination thresholds to humans) whose expression of the homologue of the human gene KIAA0319 was reduced (by *in utero* RNAi transfection of *Kiaa0319*) displayed much variability in neural excitability, neural discriminability and latency to speech sounds, findings which offer a putative mechanism for how this so-called dyslexia gene may impair phoneme processing namely by altering auditory cortical responses (Centanni et al., 2013). Interestingly, this gene has also been associated with a reduction in the asymmetric activation of the superior temporal sulcus in human participants during an fMRI reading study, although the sample size was small for studying complex genetic traits (Pinel et al., 2012).

3.4. *Application to cognitive neuroscience*

Automated text analysis technologies hold much promise of integrating fragmented information spread across multiple fields of expertise into a complete picture exposing the interrelated roles of various genes, proteins, and chemical reactions in cells and organisms. Such methods could be applied to delineating gene clusters that share a similar biological function and establishing connections between genes and disease. Already there are demonstrations of how statistical text mining methods can rapidly obtain functional information about genes, (protein-to-protein interactions, gene function annotation, and measures of gene-to-gene similarity (Raychaudhuri, 2006)), and establish links to specific disease states (e.g., Raychaudhuri et al., 2009; Semeiks, Grate, & Mian, 2005; Xuan, Wang, Watson, & Meng, 2007). We have illustrated some of the potential of automated text analysis with a widely used neuropsychological task, namely category fluency. However, such a framework for analysing the content of words is equally applicable to other assessment tasks, for example, the output generated in verbal learning and recall tests, or even directly from natural speech. Here, we have shown that genes previously associated with verbal fluency and verbal learning and recall showed a sex-dependent association with the number of valid words generated and LSA-derived average vector length that was able to be replicated, despite small sample sizes. Our study provides an intriguing first attempt to delineate the underlying genomic architecture of category fluency using phenotypes obtained via computational linguistics approaches. Ideally, our preliminary findings will provide the impetus for larger studies such as genome-wide association studies of similar phenotypes, which would require substantially larger sample sizes. As these modelling approaches are agnostic to language, combination of data across centres and languages within a consortium-level effort would provide the statistical power to better understand how common variation contributes to these novel phenotypes.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland and the Marie-Curie Action COFUND under Grant Number 11/SIRG/B2183 to Dr. Nicodemus. Dr. Elvevåg was supported by the Northern Norwegian Regional Health Authority (Helse Nord RHF). All calculations were performed on the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing. This cluster was funded through grants from Science Foundation Ireland.

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2013.12.004

REFERENCES

Aukes, M. F., Alizadeh, B. Z., Sitskoorn, M. M., Selten, J.-P., Sinke, R. J., Kemner, C., et al. (2008). Finding suitable phenotypes for genetic studies of schizophrenia: heritability and segregation analysis. *Biological Psychiatry, 15*, 128–136.

Balog, Z., Kiss, I., & Keri, S. (2011). ZNF804A may be associated with executive control of attention. *Genes Brain and Behaviour, 10*, 223–227.

Becker, J., Czamara, D., Hoffmann, P., Landerl, K., Blomert, L., Brandeis, D., et al. (2012). Evidence for the involvement of ZNF804A in cognitive processes of relevance to reading and spelling. *Translational Psychiatry, 2*, e136.

Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., et al. (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cognitive Neuropsychiatry, 14*, 419–450.

Bokat, C. E., & Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophrenia Research, 64*, 73–78.

Bratko, D. (1996). Twin study of verbal and spatial abilities. *Personality and Individual Differences, 21*, 621–624.

Callicott, J. H., Straub, R. E., Pezawas, L., Egan, M. F., Mattay, V. S., Hariri, A. R., et al. (2005). Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia. *PNAS (USA), 102*, 8627–8632.

Carless, M. A., Glahn, D. C., Johnson, M. P., Curran, J. E., Bozaoglu, K., Dyer, T. D., et al. (2011). Impact of *DISC1* variation on neuroanatomical neurocognitive phenotypes. *Molecular Psychiatry, 16*, 1096–1104.

Centanni, T. M., Booker, A. B., Sloan, A. M., Chen, F., Maher, B. J., Caraway, R. S., et al. (2013). Knockdown of the dyslexia-associated gene Kiaa0319 impairs temporal responses to speech stimuli in rat primary auditory cortex. Advance Access published February 8, 2013 *Cerebral Cortex*. http://dx.doi.org/10.1093/cercor/bht028.

Chen, M., Xu, Z., Zhai, J., Bao, X., Zhang, Q., Gu, H., et al. (2012). Evidence of IQ-modulated association between *ZNF804A* gene polymorphism and cognitive function in schizophrenia patients. *Neuropsychopharmacology, 36*, 1572–1578.

DeFries, J. C., Fulker, D. W., & LaBuda, M. C. (1987). Evidence for a genetic aetiology in reading disability in twins. *Nature, 329*, 537–539.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California verbal learning test*. San Antonio, TX: Psychological Corporation.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*, 304–316.

Elvevåg, B., & Weinberger, D. R. (2001). Neuropsychology in context of the neurodevelopmental model of schizophrenia. In C. A. Nelson, & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 577–595). Cambridge, Mass: MIT Press.

Elvevåg, B., & Weinberger, D. R. (2009). Introduction: genes, cognition and neuropsychiatry. *Cognitive Neuropsychiatry, 14*, 261–276.

Esslinger, C., Kirsch, P., Haddad, L., Mier, D., Sauer, C., Erk, S., et al. (2011). Cognitive state and connectivity effects of the genome-wide psychosis variant in ZNF804A. *NeuroImage, 54*, 2514–2523.

Flint, J., & Munafo, M. R. (2007). The endophenotype concept in psychiatric genetics. *Psychological Medicine, 37*, 163–180.

Goldberg, T. E., & Weinberger, D. R. (2004). Genes and the parsing of cognitive processes. *Trends in Cognitive Sciences, 8*, 325–335.

Graham, S. A., & Fisher, S. E. (2013). Decoding the genetics of speech and language. *Current Opinion in Neurobiology, 23*, 43–51.

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2013). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.02.006 (Epub ahead of print]).

Huffaker, S. J., Chen, J., Nicodemus, K. K., Sambataro, F., Yang, F., Mattay, V., et al. (2009). A primate-specific, brain isoform of KCNH2 affects cortical physiology, cognition, neuronal repolarization and risk of schizophrenia. *Nature Medicine, 15*, 509–518.

Krug, A., Nieratschker, V., Markov, V., Krach, S., Jansen, A., Zerres, K., et al. (2010). Effect of CACNA1C rs1006737 on neural correlates of verbal fluency in healthy individuals. *NeuroImage, 49*, 1831–1836.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*, 1073–1081.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Kintsch, W., McNamara, D. S., & Dennis, S. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lencz, T., Szeszko, P. R., DeRosse, P., Burdick, K. E., Bromet, E. J., Bilder, R. M., et al. (2010). A schizophrenia risk gene, *ZNF804A*, influences neuroanatomical and neurocognitive phenotypes. *Neuropsychopharmacology, 35*, 2284–2291.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Luciano, M., Hansell, N. K., Lahti, J., Davies, G., Medland, S. E., Räikkönen, K., et al. (2011). Whole genome association scan for genetic polymorphisms influencing information processing speed. *Biological Psychology, 86*, 193–202.

Meyer-Lindenberg, A., & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience, 7*, 818–827.

Millar, J. K., Wilson-Annan, J. C., Anderson, S., Christie, S., Taylor, M. S., Semple, C. A., et al. (2000). Disruption of two novel genes by translocation co-segregating with schizophrenia. *Human Molecular Genetics, 22*, 1415–1423.

Morar, B., Dragović, M., Waters, F. A. V., Chandler, D., Kalaydjieva, L., & Jablensky, A. (2011). Neuregulin 3 (NRG3) as a susceptibility gene in a schizophrenia subtype with florid delusions and relatively spared cognition. *Molecular Psychiatry, 16*, 860–866.

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., et al. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics, 40*, 1053–1055.

Palo, O. M., Antila, M., Silander, K., Hennah, W., Kilpinen, H., Soronen, P., et al. (2007). Association of distinct allelic haplotypes of DISC1 with psychotic and bipolar spectrum disorders and with underlying cognitive impairments. *Human Molecular Genetics, 16*, 2517–2528.

Papassotiropoulos, A., & de Quervain, D. J. F. (2011). Genetics of human episodic memory: dealing with complexity. *Trends in Cognitive Sciences, 15*, 381–387.

Pauli, A., Prata, D. P., Mechelli, A., Picchioni, M., Fu, C. H., Chaddock, C. A., et al. (2012). Interaction between effects of

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004

genes coding for dopamine and glutamate transmission on striatal and parahippocampal function. Mar 22 [Epub ahead of print] *Human Brain Mapping*. http://dx.doi.org/10.1002/hbm22061.

Pinel, P., Fauchereau, F., Moreno, A., Barbot, A., Lathrop, M., Zelenika, D., et al. (2012). Genetic variants of FOXP2 and KIAA0319/TTRAP/THEM2 locus are associated with altered brain activation in distinct language-related regions. *Journal of Neuroscience, 32*, 817–825.

Porteous, D. J., Millar, J. K., Brandon, N. J., & Sawa, A. (2011). DISC1 at 10: connecting psychiatric genetics and neuroscience. *Trends in Molecular Medicine, 17*, 699–706.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics, 81*, 559–575.

Raychaudhuri, S. (2006). *Computational text analysis for functional genomics and bioinformatics*. Oxford: Oxford University Press.

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C. Y., , International Schizophrenia Consortium, Purcell, S. M., et al. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genetics, 5*, e1000534.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sakakibara, E., Takizawa, R., Nishimura, Y., Kawasaki, S., Satomura, Y., Kinoshita, A., et al. (2013). Genetic influences on prefrontal activation during a verbal fluency task in adults: a twin study based on multichannel near-infrared spectroscopy. *NeuroImage*. http://dx.doi.org/10.1016/j.neuroimage.2013.03.052 (Epub ahead of print).

Schwartz, S., Baldo, J., Graves, R. E., & Brugger, P. (2003). Pervasive influence of semantics in letter and category fluency: a multidimensional approach. *Brain and Language, 87*, 400–411.

Semeiks, J. R., Grate, L. R., & Mian, I. S. (2005). Text-based analysis of genes, proteins, aging, and cancer. *Mechanisms of Ageing and Development, 126*, 193–208.

Snitz, B. E., MacDonald, A. W., & Carter, C. S. (2006). Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. *Schizophrenia Bulletin, 32*, 179–194.

Tan, H. Y., Callicott, J. H., & Weinberger, D. R. (2008). Intermediate phenotypes in schizophrenia genetics redux: is it a no brainer? *Molecular Psychiatry, 13*, 233–238.

Thomson, P. A., Parla, J. S., McRae, A. F., Kramer, M., Ramakrishnan, K., Yao, J., et al. (2013). 708 common and 2010 rare DISC1 locus variants identified in 1542 subjects: analysis for association with psychiatric disorder and cognitive traits. Advance online publication, 4 June 2013 *Molecular Psychiatry*. http://dx.doi.org/10.1038/mp.2013.68.

Vandenberg, S. G. (1962). The hereditary abilities study: hereditary components in a psychological test battery. *American Journal of Human Genetics, 14*, 220–223.

Van Den Bossche, M. J., Docx, L., Morrens, M., Cammaerts, S., Strazisar, M., Bervoets, C., et al. (2012). Less cognitive and neurological deficits in schizophrenia patients carrying risk variant in ZNF804A. *Neuropsychopharmacology, 35*, 2284–2291.

Venkatesh, S. K., Siddaiah, A., Padakannaya, P., & Ramachandra, N. B. (16 May 2013). Analysis of genetic variants of dyslexia candidate genes KIAA0319 and DCDC2 in Indian population. Advance online publication *Journal of Human Genetics*. http://dx.doi.org/10.1038/jhg.2013.46.

Voineskos, A. N., Lerch, J. P., Felsky, D., Tiwari, A., Rajji, T. K., Miranda, D., et al. (2011). The ZNF804A gene: characterization of a novel neuronal risk mechanism for the major psychoses. *Neuropsychopharmacology, 36*, 1871–1878.

Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D. R., & Elvevåg, B. (2013). Deriving semantic structure from category fluency: clustering techniques and their pitfalls. *Cortex* (in press).

Walters, J. T. R., Corvin, A., Owen, M. J., Williams, H., Dragovic, M., Quinn, E. M., et al. (2010). Psychosis susceptibility gene *ZNF804A* and cognitive performance in schizophrenia. *Archives of General Psychiatry, 67*, 692–700.

Wechsler, D. (1987). *Wechsler memory scale—Revised*. San Antonio, TX: Psychological Corporation.

Williams, H. J., Norton, N., Dwyer, S., Moskvina, V., Nikolov, I., Carroll, L., et al. (2011). Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Molecular Psychiatry, 16*, 429–441.

Xuan, W., Wang, P., Watson, S. J., & Meng, F. (2007). Medline search engine for finding genetic markers with biological significance. *Bioinformatics, 23*, 2477–2484.

Zou, L., Chen, W., Shao, S., Sun, Z., Zhong, R., Shi, J., et al. (2012). Genetic variant in KIAA0319, but not in DYX1C1, is associated with risk of dyslexia: an integrated meta-analysis. *American Journal of Medical Genetics Part B, 159B*, 970–976.

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004