EBioMedicine 28 (2018) 168–179

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Ancestral Variations of the *PCDHG* Gene Cluster Predispose to Dyslexia in a Multiplex Family

Teesta Naskar [a,e], Mohammed Faruq [b], Priyajit Banerjee [c], Massarat Khan [d], Rashi Midha [a,1], Renu Kumari [b,f], Subhashree Devasenapathy [a], Bharat Prajapati [a], Sanghamitra Sengupta [e], Deepti Jain [c], Mitali Mukerji [b], Nandini Chatterjee Singh [a,*], Subrata Sinha [a,g,*]

[a] Cellular and Molecular Neuroscience Division, National Brain Research Centre, Manesar, Gurgaon, Haryana 122051, India
[b] Genomics and Molecular Medicine, CSIR-Institute of Genomics and Integrative Biology, New Delhi 110007, India
[c] Regional Centre for Biotechnology, NCR Biotech Science Cluster, Faridabad, Haryana 121001, India
[d] Maharashtra Dyslexia Association, Mumbai, Maharashtra 400088, India
[e] Department of Biochemistry, University of Calcutta, Kolkata, West Bengal 700019, India
[f] Academy of Scientific and Innovative Research, CSIR-Institute of Genomics & Integrative Biology (AcSIR-IGIB), New Delhi 110020, India
[g] Department of Biochemistry, All India Institute of Medical Sciences, New Delhi 110029, India

ARTICLE INFO

Article history:
Received 6 August 2017
Received in revised form 22 December 2017
Accepted 28 December 2017
Available online 9 January 2018

Keywords:
Dyslexia
Protocadherin gamma
Ancestral variations
Neanderthal genome
Dominant inheritance
Trans-homophilic interaction
Neuronal connection

ABSTRACT

Dyslexia is a heritable neurodevelopmental disorder characterized by difficulties in reading and writing. In this study, we describe the identification of a set of 17 polymorphisms located across 1.9 Mb region on chromosome 5q31.3, encompassing genes of the *PCDHG* cluster, *TAF7*, *PCDH1* and *ARHGAP26*, dominantly inherited with dyslexia in a multi-incident family. Strikingly, the non-risk form of seven variations of the *PCDHG* cluster, are preponderant in the human lineage, while risk alleles are ancestral and conserved across Neanderthals to non-human primates. Four of these seven ancestral variations (c.460A > C [p.Ile154Leu], c.541G > A [p.Ala181Thr], c.2036G > C [p.Arg679Pro] and c.2059A > G [p.Lys687Glu]) result in amino acid alterations. p.Ile154Leu and p.Ala181Thr are present at EC2: EC3 interacting interface of γA3-PCDH and γA4-PCDH respectively might affect trans-homophilic interaction and hence neuronal connectivity. p.Arg679Pro and p.Lys687Glu are present within the linker region connecting trans-membrane to extracellular domain. Sequence analysis indicated the importance of p.Ile154, p.Arg679 and p.Lys687 in maintaining class specificity. Thus the observed association of *PCDHG* genes encoding neural adhesion proteins reinforces the hypothesis of aberrant neuronal connectivity in the pathophysiology of dyslexia. Additionally, the striking conservation of the identified variants indicates a role of *PCDHG* in the evolution of highly specialized cognitive skills critical to reading.

1. Introduction

Reading is a specific, advanced cognitive activity of humans. However individuals with dyslexia experience varying degrees of difficulty in performing this skill despite adequate intelligence or education and absence of neurological illness or sensory deficits (Peterson and Pennington, 2015). Worldwide epidemiological data suggests that the prevalence of dyslexia is approximately 5–12% (Shaywitz et al., 2007), while in India it is reported to be 9–11% (Mogasale et al., 2012). Dyslexia is known to have a strong neurodevelopmental origin, as a result of aberrations in neuronal migration and connectivity, as elucidated in a number of studies involving postmortem brains and neuroimaging (Paulesu et al., 1996; Skeide et al., 2015). Studies on postmortem brains of individuals with dyslexia have shown specific histological anomalies including ectopias and heterotopias resulting from abnormal neuronal migration (Galaburda and Kemper, 1979; Galaburda et al., 1985). Functional magnetic resonance imaging (f-MRI) and positron emission tomography (PET) scan studies have reported the neural correlates of acquired skills like reading and writing (Paulesu et al., 2014). These studies attributed dyslexia to poor or delayed neuronal maturation and disrupted functional connectivity of neurons.

This neurodevelopmental disorder also has a strong genetic component which could be heterogeneous in nature (Mascheretti et al., 2017). The heritability of dyslexia, as well as its constituent sub-phenotypes, has been shown by a number of twin and family-based studies (Fisher et al., 2002; Schumacher et al., 2007). Genome-wide studies using complex pedigrees have reiterated the heritability of dyslexia sub-phenotypes and mapped several genomic loci designated as *DYX1-DYX9* includes candidate genes like *ROBO1* (Hannula-Jouppi et al.,

* Corresponding authors.
E-mail address: subrata.sinha@nbrc.ac.in (S. Sinha).
[1] Present address: National Institute of Mental Health and Neuro Science, Bangalore, Karnataka 560029, India.

https://doi.org/10.1016/j.ebiom.2017.12.031

2005), *KIAA0319* (Paracchini et al., 2006), *DCDC2* (Meng et al., 2005), *DYX1C1* (Tapia-Paez et al., 2008), mostly implicated in neurite outgrowth, neural connectivity, migration and development. In addition to the genes of these *DYX* loci, many other genes such as *FOXP2*, *CNTNAP2* (Peter et al., 2011), *SLC2A3* (Roeske et al., 2011), *GRIN2B* (Mascheretti et al., 2015), *CEP63* (Einarsdottir et al., 2015) and *PCDH11X* (Veerappa et al., 2013) have been shown to be associated with dyslexia.

The studies so far indicate that dyslexia is likely to be a collection of many different endophenotypes resulting in multiple molecular and cellular pathologies. However, the basic molecular underpinnings of this disability are still elusive. Therefore, for a better understanding of the pathophysiology of reading (dis)ability, an effort to identify novel susceptible genes to the disorder, we have investigated the genetic basis of dyslexia inheritance, applying whole exome sequencing and genome wide SNP array, in a three generational family from a highly endogamous group from Western India.

We identified 17 variations present at or adjacent to the protocadherin gamma (*PCDHG)* gene cluster that co-segregated with dominantly inherited dyslexia in the family being studied. The clustered protocadherins play important roles in several steps of neural morphogenesis and connectivity. Remarkable features of the γ-PCDH proteins, including their extensive molecular diversity, enriched synaptic localization, isoform specific homophilic adhesion, cell specific expression pattern, dendritic expression and spine morphogenesis suggest their indispensable role in the development and maintenance of neural circuits and their functional maturity and connectivity (Chen and Maniatis, 2013; Schreiner and Weiner, 2010; Kostadinov and Sanes, 2015).

Clustered α and β-*PCDH* as well several non-clustered *PCDH* have been reported to be associated with many neurodevelopmental disorders like autism spectrum disorder (ASD) (Anitha et al., 2013), schizophrenia (Jiang et al., 2017), epilepsy (Cooper et al., 2016), intellectual disability. In an Indian family based study, a genome wide scan identified copy number variations of *PCDH11X*, a non-cluster protocadherin as being associated with dyslexia (Veerappa et al., 2013).

In the present study, identification of multiple variations co-segregated with dyslexia like a single haplotype block provides mechanistic insights into the disease pathophysiology. In addition, the presence of the variations on extracellular domains of γPCDHs along with the importance of p.Ile154Leu and p.Ala181Thr in trans-homophilic interactions strengthen the hypothesis of aberrant neuronal connectivity in the pathophysiology of dyslexia and could guide to generate physiologically relevant cellular and animal models. Interestingly, the striking evolutionary conservation of seven of these dyslexia associated variants, including four non-synonymous amino acid changes (c.460A > C [p.Ile154Leu], c.541G > A [p.Ala181Thr], c.2036G > C [p.Arg679Pro] and c.2059A > G [p.Lys687Glu]), indicates their evolutionary significance in the development of cognitive substrates underlying the unique human ability to read.

2. Materials and Methods

2.1. Participants

An extended family KA25 with familial dyslexia were identified for this study (Fig. 1). Twenty members out of twenty six, were included in this study. The study was approved by the Institutional Human Ethics Committee of National Brain Research Centre, Manesar, India and signed informed consent was obtained from all the participants in accordance with the Declaration of Helsinki. In case of children, the signed informed consent was obtained from their parents. Members of the family were in the age range of 4–78 years; while tests for reading were administered to the 7–70 years age group. Except for two participants (IV-8 and IV-9), all others had a graduate degree and had received at least 15 years of academic education in English. They all reported English as their preferred and proficient language and hence languages and reading assessments were carried out in English, using Dyslexia Assessments for Languages of India (DALI) (see Web Resource). DALI is a standardized and validated battery of assessments developed by the National Brain Research Centre, India and is available in four Indian languages namely Hindi, Marathi, Kannada and English. Non-verbal intelligence quotient was assessed using Standard Progressive Matrices (Raven, 2000). Participants were interviewed individually to ascertain reading history, difficulties during schooling and performances in their remedial classes for those who had undergone the remedial program.

2.2. Whole Exome Sequencing

Whole exome sequencing was performed for individuals II-1 II-9, III-1, III-2, III-3, III-4, III-8, III-9, IV-1, IV-2, IV-3, IV-6 and IV-7 on Illumina Hi-seq 2000. For each sample, 2 μg of non-degraded high molecular weight genomic DNA was used by following manufacturer's protocol.

Bioinformatics analyses and quality check of sequence reads were done through genome reassembly pipeline of NGS toolkit (Patel and Jain, 2012). After ensuring quality, raw sequence reads of two end-sequenced read files were mapped to human indexed reference genome file (Grch37/hg19) by using Bowtie2 (Langmead and Salzberg, 2012). Variant calling from (Sequence Alignment Map) SAM aligned file was performed by using SAMTools (Li et al., 2009) following conversion of SAM file to BAM (Binary Alignment/Map format) file using 'samtools view' parameter, sorting of BAM file using 'samtools sort' parameter, indexing of BAM file using 'samtools index' parameter, generating variation BCF (binary) file using 'samtoolsmpileup' and converting BCF (binary) file to VCF (text) variation file using bcftools. To recalibrate the base quality score as well the local alignment around insertions and deletions was done by GATK31 method. After passing the data QC (80% coverage, >25× depth), total 156,294 variations were found to be shared among all thirteen individuals.

2.3. Variant Prioritization

The variants were prioritized on the basis of dominant inheritance pattern of the disorder in this family. Therefore all 156,294 variations from whole exome sequencing, were filtered out for the risk alleles that were either heterozygous or homozygous in affected individuals against homozygous non-risk genotypes in unaffected individuals. As per the dominant model, it was assumed that one copy of the risk allele was sufficient to develop the disorder.

2.4. Genome Wide SNP Scan

Each DNA sample from the family KA25 was genotyped for 719665 SNP marker using Illumina Human Omni express12v1-1. 1 μg DNA per sample was used for the fragmentation process followed by PCR enrichment for SNP. Initial genomic data scan was performed by using iScan (Illumina). Variants were annotated by ANNOVAR (Wang et al., 2010) and after generating the base call files all the individual files were merged and processed in Genome Studio 1.7 where a final file was generated and analyzed.

2.5. Multiple Sequence Alignment

To investigate conservation of each identified variations within the Neanderthal genome (Prüfer et al., 2017) and primate groups, we have performed the multiple sequence alignment of the flanking sequences of each SNPs using NCBI blast tool, clustal-w and t-coffee (Notredame et al., 2000). Results were generated using ESPript (Gouet et al., 1999). Primate sequences were collected from both UCSC and NCBI database. Neanderthal genome sequences were obtained from the 'Ancient Genome Browser' (see Web Resource) of the Department of Evolutionary Genetics, Max Planck Institute of Evolutionary Anthropology, Leipzig, Germany and also from the UCSC Genome Browser.