RNA isoform analyses of *Drosophila Dscam* gene and *Xenopus tropicalis clustered Protocadherin* genes provide insights for neuronal self-avoidance

Dissertation

zur Erlangung des akademischen Grades des Doktors der Naturwissenschaften
(Dr. rer. nat.)

eingereicht im

Fachbereich Biologie, Chemie, Pharmazie
der
Freien Universität Berlin

vorgelegt von

Wei Sun (孙维)

aus Henan, China

Dezember 2015

Die vorliegende Arbeit wurde von Oktober 2010 bis Dezember 2015 am Max-Delbrück-Centrum für Molekulare Medizin (MDC) unter der Anleitung von Prof. Wei Chen angefertigt.

1. Gutachter: Prof. Dr. Wei Chen

2. Gutachter: Prof. Dr. Stephan Sigrist

Datum der Disputation: 11. 04. 2016

Selbststandigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Ich erkläre weiterhin, dass ich weder die vorliegende Arbeit noch deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

I hereby declare that this thesis is my own original research work and has not been submitted in any form for another degree of diploma at any university or other institute of education. Contributions from others have been clearly acknowledged in the text and references to literatures are given.

Wei SUN

2015-12-22

Preface

All the results presented here originated from collaborations with other researchers. Here I summarize my contributions and acknowledge the contributions of my collaborators in below.

Chapter 2 describes a novel sequencing-based approach for absolutely quantifying the expressions of *Drosophila Dscam* RNA isoforms, and the novel insight on neuronal self-avoidance yielded from it. The results have been published in the EMBO Journal (Sun et al., 2013). Prof. Wei Chen initiated this project. Prof. Wei Chen and I designed the novel CAMSeq technology and all other experimental procedures together. My contribution also included: 1) conducting all the experimental parts; 2) interpreting the data together with others. I would like to acknowledge contributions of Dr. Xintian You, Dr. Andreas Gogol-Döring, Dr. Haihuai He, Dr. Yoshiaki Kise, Madlen Sohn, Claudia Quedenau, Tao Chen, Mirjam Feldkamp, Claudia Langnick, Prof. Ansgar Klebes, Prof. Dietmar Schmucker, and Prof. Wei Chen.

Chapter 3 describes the annotation for *Xenopus tropicalis clustered Protocadherin* genes using the full-length 5'RACE sequencing approach. Prof. Dietmar Schmucker and Prof. Wei Chen initiated this project. Emre Etlioglu and I designed 5'RACE experiments together. My contribution also included: 1) designing all other experimental parts; 2) conducting experiments together with Claudia Quedenau; 3) analyzing and interpreting the data. I would like to acknowledge contributions of Emre Etlioglu, Claudia Quedenau, Bin Zhang, Prof. Dietmar Schmucker, and Prof. Wei Chen.

Acknowledgement

I would like to thank…

… my supervisor Prof. Wei Chen for inviting me to join his lab in Max-Delbrück-Center for Molecular Medicine (MDC). During all these years, he has been an excellent mentor and supervisor for me. He has lead me into the Systems Biology field, created the opportunities for me to work with so many cutting-edge technologies, and encouraged me to participate so many exciting research projects. I'm grateful to his supervision and mentoring, from which I would benefit for life. I also would like to thank him for proofreading my dissertation.

… my university supervisor Prof. Stephan Sigrist for giving me the opportunity to join the Free University, Berlin. His generous support has been of great help during my stay, and for all my processes in the university.

… all the members of Prof. Wei Chen's group. We have created such a joyful, inspiring and stimulating research environment for researches, which I enjoyed and benefited from for all the time. I would like to specially thank Dr. Xintian You, Dr. Andreas Gogol-Döring, and Qingsong Gao for our wonderful collaborations in different projects. I would like to thank Dr. Yuhui Hu, Dr. Yongbo Wang, Dr. Na Li, Dr. Wei Sun, Dr. Sebastian Fröhler, Tao Chen, Dr. Xi Wang, Jingyi Hou, Bin Zhang and Dr. Kun Song for teaching me many techniques during my study. I also would like to thank Madlen Sohn, Claudia Quedenau, Mirjam Feldkamp, Claudia Langnick, Anna-Maria Ströhl, and Sarah Nathalie Vitcetz for their excellent help in all my projects.

… my collaborators from other labs and institutes, especially: Prof. Dietmar Schmucker, Dr. Haihuai He, Dr. Yoshiaki Kise, Dr. Sophie A. O. Armitage, and Prof. Ansgar Klebes for Dscam project; Prof. Dietmar Schmucker, and Emre Etlioglu for cPcdh project; Prof. Claude Libert, Dr. Irina Savelyeva, Marlies Ballegeer, Dr. Jean

Jaubert and Dr. Xavier Montagutelli, for hybrid mouse project; Prof. Song Gao, Jiali Hu for FAM46B project.

… Sabrina Deter, Jennifer Stewart, and Sylvia Sibilak for their great help for all the administrative processes.

… all my friends in Germany, and the basketball crews in MDC and Berlin-Buch, especially my friend Jan Rossa for the translation of the Zusammenfassung. I also would like to thank the Chinese community in Campus, especially Dr. Yu Shi. All of you helped me a lot, and made my life in Germany wonderful.

… the China Scholarship Council (CSC) for providing me a four-year scholarship.

… Last, but above all, I want to devote my deepest thanks to my parents and my family in China. Without their love and support for all these years, it would be impossible for me to finish my PhD study. I love you.

Table of contents

Summary

The enormous isoform diversities of *Drosophila Dscam* gene and *Xenopus tropicalis (Xtro) clustered Protocadherin (cPcdh)* genes are generated from RNA alternative splicing, and play profound roles in neuronal self-avoidance.

In *Drosophila*, although appreciated as important, the Dscam isoform expression pattern at the global level still remained unexplored. Here we developed a novel method that allows for direct quantification of *Dscam* isoforms expressing patterns from over hundreds of millions of *Dscam* transcripts in one sequencing run. With such sequencing depth, we detected the expression of 18,496 isoforms, out of 19,008 theoretically possible combinations. Importantly, we demonstrated that alternative splicing between different clusters is independent. Moreover, the isoforms expressed across a broad dynamic range, with significant biases in cell/tissue and developmental stage specific patterns. Hitherto underappreciated, such bias can dramatically reduce the ability of neurons to display unique surface receptor codes. Therefore, the seemingly excessive diversity encoded in the *Dscam* locus might be essential for a robust self and non-self discrimination in neurons.

In vertebrates, cPcdh serves as counterpart as Dscam in *Drosophila*, and the function of *cPcdh* genes in neuronal self-avoidance is considered conserved across vertebrates. *Xtro* is a powerful and convenient model organism for studies of neuron development. However, the annotation of *cPcdh* genes in *Xtro* genome is still incomplete. Here by full-length 5'RACE sequencing, we refined and characterized the annotations of the *Xtro cPcdh* genes in details. In total, three *cPcdh* clusters, with at least 98 variable exons, were identified, demonstrating the genome duplication and expansion of *Xtro cPcdh* loci. Interestingly, one novel *cPcdh γ1* CE isoform we identified may serve a species-specific function for *Xtro* neuronal development. Our annotations for *Xtro cPcdh* genes provide a valuable resource for their future functional characterization.

Zusammenfassung

Die große Isoformenvielfalt des *Drosophila Dscam* Gens und der *Xenopus tropicalis (Xtro) clustered Protocadherin (cPcdh)* Gene werden durch alternatives Spleißen gebildet. Sie spielen eine wichtige Rolle in der neuronalen Selbstvermeidung (neuronal self-avoidance).

Obwohl es als wichtig erachtet wird, sind die Dscam Isoformexpressionsmuster von *Drosophila* noch nicht umfassend aufgeklärt. In diesem Projekt wurde eine neue Methode entwickelt, die es ermöglicht die über hundert Millionen *Dscam* Transkripte, die durch alternatives Spleißen gebildet werden, in einem Sequenzierungslauf zu analysieren. Mit dieser Methode konnten wir 18.496 von rechnerisch 19.008 möglichen Isoformen detektieren. Dabei konnte als wichtiges Ergebnis gezeigt werden, dass alternatives Spleißen unabhängig von den unterschiedlichen Exon-Clustern ist. Außerdem konnte nachgewiesen werden, dass die Isoformen mit einem breiten dynamischen Spektrum exprimiert werden. Zellen, Gewebe und ganze Fruchtfliegen in unterschiedlichen Entwicklungsstadien werden mit signifikanter Verzerrung in spezifischen Mustern exprimiert. Solche Verzerrung können die Möglichkeit der Neuronen einzigartige Oberflächenrezeptor-Codes zu bilden reduzieren. Deshalb ist die scheinbar große Isoformenvielfalt im *Dscam* Genort für eine stabile Selbstdiskriminierung (self/non-self discrimination) der Neuronen dennoch notwendig.

In Wirbeltieren hat cPcdh eine ähnliche Funktion wie Dscam in *Drosophila*. Diese Funktion der *cPcdh*-Gene in der neuronalen Selbstvermeidung ist konserviert in allen Wirbeltierarten. *Xtro* ist ein guter geeigneter Modelorganismus für Untersuchungen der neuronalen Entwicklung. Dennoch ist die Annotation der *cPcdh*-Gene im *Xtro*-Genom immer noch unvollständig. In diesem Projekt konnte mit Hilfe von full-length 5'RACE Sequenzierung diese Annotation vollständig und detailliert aufgeklärt werden. Es konnten drei *cPcdh* Gen-Cluster mit insgesamt mindestens 98 variablen Exons identifiziert werden, welche die genomische Duplikation und

Expansion von *Xtro cPcdh* Genorten zeigt. Besonders interessant ist die *cPcdh γ1* CE Isoform, die erstmals identifiziert werden konnte. Sie könnte eine artenspezifische Funktion für die neuronale Entwicklung von *Xtro* haben. Die gewonnenen Annotationen für die *Xtro cPcdh*-Gene bieten einen guten Ausgangspunkt für weitere funktionale Charakterisierungen dieser Gene.

Chapter 1. Introduction

1.1 Chemoaffinity Hypothesis in Neuronal Network Building

One of the most fascinating features of the neuronal system is its highly reproducible complexity and cell diversity. During development, individual neurons are highly differentiated, and well-integrated into the complex neuronal networks by correctly locating and recognizing their appropriate synaptic partners, which build the structural foundation for the enormously complex neuronal functions.

This high reproducibility of the complex pattern of synaptic connectivity must require a system that can label and guide the growth and wiring of each neuron individually during neuronal development. By analyzing the regeneration following neuronal injury, Langley (Langley, 1895) and Sperry (Sperry, 1963) proposed the similar hypothesis that there must be "some special chemical relation between each class of nerve fibre and each class of nerve cell" (Langley, 1895). This hypothesis, formalized as chemoaffinity hypothesis, proposed that each neuron "must carry some kind of individual identification tags", which can guide "each axon linking only with certain neurons to which it becomes selectively attached by specific chemical affinities" (Sperry, 1963; Zipursky and Sanes, 2010).

Since the proposal of this "chemoaffinity hypothesis", one question had been long posed: considering the total number of neurons and the total number of recognizable synaptic connections in the nervous system, is it possible that such kind of molecular tagging system is existing to match the cellular and sub-cellular diversity in neuronal system? Especially after the genome sequencing of several organisms, a dilemma seems to arise: the number of genes possessed in our genome is much less than previously predicted (only 20,000 – 25,000 protein coding genes in human genome)(International Human Genome Sequencing Consortium, 2004). How could this gene number match the requirement of multitude of the neuronal complexity (for

example, in human brain, the number of neurons is ~8.6 x 10^{10}, and the number of synapses is $10^{14} – 10^{15}$)?

1.2 Models and Molecules in Chemoaffinity Hypothesis

Many efforts have been invested on finding these "molecule tags", and in-between three general modified models for chemoaffinity guidance have been suggested (reviewed in Zipursky and Sanes, 2010).

The first model is called "gradient molecules" matching. The best example is the complementary gradients Eph kinases and their ligands, ephrins. It has been demonstrated that, in retina, the graded expressions and interactions between Eph and ephrins play critical roles in establishing the topological retinotectal map (Cheng et al., 1995; Drescher et al., 1995; McLaughlin and O'Leary, 2005).

The second model proposes that, combinatory effects of many different guidance molecules generate the neuronal individuality. These include various axonal guidance cues and their receptors, such as ephrins, semaphorins, netrins, plexins, robos, slits, and so on (reviewed in Dickson, 2002). These molecular cues interact as attractants and repellents, with the different combinations guiding the growths of specific axons to target regions via contact-mediated (short distance) and diffusible (long distance) regulatory mechanisms.

In the last model, the specificity is achieved by differential expressions of different members of multigene families of cell adhesion molecules, which possess distinct binding specificities among different family members. In this scenario, different individual neurons (or even different sub-cellular areas of neurons) express different members (or different combinations of these members) from these gene families, and therefore are stamped with different molecule identity on single cellular (or even sub-cellular) level by the members of these gene families. Several such families have been identified: the classical and type II cadherins, (Takeichi, 2007), the neurexins and neuroligins (Schreiner et al., 2014; Südhof, 2008; Treutlein et al., 2014), and the

olfactory receptors (Buck and Axel, 1991). And this hypothesis is formalized as "area code hypothesis" (Dreyer, 1998).

1.3 RNA Alternative Splicing in Chemoaffinity Hypothesis

The last model "area code hypothesis" is especially interesting, since it could match and also contribute to the theory that the molecular diversity of proteins can be substantially expanded by the increasing of the isoform diversity from RNA alternative splicing.

Alternative splicing generates multiple transcripts from the same gene by different combinations of exons, thereby increasing transcriptome plasticity and proteome diversity(Nilsen and Graveley, 2010). Recent studies using high-throughput sequencing indicate that about 25%, 60% and 90% of multi-exon genes in *C. elegans*, *Drosophila melanogaster* and humans, respectively, undergo alternative splicing(Gerstein et al., 2010; Graveley et al., 2011; Pan et al., 2008; Ramani et al., 2011; Wang et al., 2008). More recently, alternative splicing has also been proposed to be fundamentally important for the functional complexity of nervous system (Barbosa-Morais et al., 2012; Merkin et al., 2012; Raj and Blencowe, 2015).

1.4 Alternative Splicing of *Down syndrome cell adhesion molecule (Dscam)* Gene and *clustered Protocadherin (cPcdh)* Genes in Neuronal Self-discrimination

Among the genes in "area code hypothesis", two genes become very interesting candidates for molecular labeling of individual neurons: *Down syndrome cell adhesion molecule* (*Dscam*) gene in insects, and *clustered Protocadherin (cPcdh)* gene families in vertebrates (Yagi, 2012; Zipursky and Sanes, 2010). Although evolutionarily unrelated, these two genes have been demonstrated to function similarly in either insects (*Dscam*) or vertebrates (*cPcdh*). Both are critical for "neuronal self-avoidance", the mechanism that mediates neurites repelling neurites from the same neuron, but not the ones from other neurons. Such self-avoidance

mechanism guarantees the correct self/non-self discrimination during neuronal development, especially for the synaptogenesis, and thus has emerges as one critical mechanism for guiding neuronal morphology and connectivity.

These two genes mediate neuronal self-avoidance via their tremendous isoform diversities. Both could generate large numbers of isoforms by RNA alternative splicing (~38,000 *Dscam* isoforms in Drosophila; 52 *cPcdh* isoforms in human) (Schmucker et al., 2000; Wu and Maniatis, 1999). Individual neurons would express different sets of *Dscam* (in insects) or *cPcdh* (in vertebrates) isoforms in a stochastic and combinatorial manner (Esumi et al., 2005; Kaneko et al., 2006; Neves et al., 2004). Such expressing pattern would provide individual neuron a specific "*Dscam*" or "*cPcdh*" identity. During neuronal development, neurites from the same neuron would express the same Dscam or cPcdh isoforms on their surfaces. When they meet, the same protein isoforms would form homophilic binding on the surfaces, and trigger following signaling pathways, which mediates the repulsion between these neurites. On the other hand, neurites from different neurons won't trigger such repulsion since they would express different Dscam or cPcdh protein isoforms to avoid homophilic protein binding. Such mechanism guarantees the correct neuronal self/non-self discrimination, and has been demonstrated to be critical for neuronal development.

1.5 Composition of this Dissertation

One key point for such self-discrimination system is the labeling efficiency of Dscam or cPcdh isoforms for individual neurons (Forbes et al., 2011; Hattori et al., 2009; Thu et al., 2014; Yagi, 2012). In other words, how many isoforms would be enough to guarantee the neuronal individuality and how would the expression of these isoforms influence the labeling efficiency? To answer such questions, 1) precise genome annotation for all isoforms, and 2) absolute quantification of the expressing pattern for these isoforms would be necessary. This dissertation describes our works for tackling these questions for *Dscam* gene in *Drosophila* and *cPcdh* genes in *Xenopus tropicalis* by applying various RNA isoform analyses approaches. It composes two parts.

The first part (Chapter 2) is entitled "Ultra-deep profiling of alternatively spliced *Drosophila Dscam* isoforms by circularization-assisted multi-segment sequencing". The *Drosophila Dscam* gene could generate more than 19,000 different ectodomains via RNA mutual exclusive splicing of three exon clusters in its genomic locus. Such isoform diversity is critical for its function in neuronal self-discrimination. However, due to technical limitations, the expression pattern of *Dscam* isoforms, as a combination of multiple variable exons, remains unexplored at the global level. Here, we developed a novel method termed 'CAMSeq' (Circularization-Assisted Multi-Segment sequencing) that allows for direct quantification of *Drosophila Dscam* isoform expression in a high-throughput manner. Applying such system on various developmental tissues/cells, we demonstrated that 1) almost all *Dscam* isoforms indeed express, 2) the splicing choice between different exon clusters is independent from each other. Furthermore, based on our quantitative datasets, we revealed and discussed a previously often ignored impact of biased isoform expression on the labeling efficiency for neuronal individuality, and proposed two "identity-labeling" strategies that could be used for the proper wiring in complex nervous systems. These findings have a general implication, not only for the study of *Dscam* in *Drosophila*, but also for the researches on other surface receptor genes in "area code hypothesis", for example, *cPcdh* genes in vertebrates.

The second part (Chapter 3) is entitled "Full-length 5'RACE sequencing based annotation for *Xenopus tropicalis clustered Protocadherin* Genes". *Clustered Protocadherin (cPcdh)* genes are critical for neuronal development. In mouse, its isoform diversity has been demonstrated to be critical for neuronal self-discrimination. *cPcdh* gene families have been found in the genome of all sequenced vertebrates, and have been speculated to serve similar function across vertebrates, including the Western clawed frog *Xenopus tropicalis (Xtro)*, which has been a powerful model organism for developmental and neuronal biology and therefore in theory ideal for investigating *cPcdh* function in neuronal development. However, due to the relative poor quality of *Xtro* genome assembly, there is still lacking a precise genome

annotation of *Xtro cPcdh* gene families. Here, based on our genome analyses, by applying full-length sequencing of the 5'RACE products derived from *Xtro cPcdh* mRNAs, we further annotated and characterized the *Xtro cPcdh* gene loci. Our work 1) expanded the annotation of the various exon (VE) regions; 2) refined the splicing patterns in constant exon (CE) regions of *Xtro cPcdh* genes. Interestingly, we identified one novel alternative splicing event in *Xtro cPcdh γ1* cluster. The feature and expression pattern of this event suggested that it could be functionally important for *Xtro* neuronal development. In summary, our annotation and characterization for the *Xtro cPcdh* gene loci provide the foundation for the further functional investigation using *Xtro* as model to study the functional role of *cPcdh* genes.

Chapter 2. Ultra-deep profiling of alternatively spliced *Drosophila Dscam* isoforms by circularization-assisted multi-segment sequencing

Note: Results in this chapter have been published in the EMBO Journal (Sun et al., 2013). DOI: 10.1038/emboj.2013.144.

Online link: http://dx.doi.org/10.1038/emboj.2013.144

2.1 Introduction

Alternative splicing of precursor messenger RNA (pre-mRNA) makes substantial contribution to the expansion of protein diversity (Barbosa-Morais et al., 2012; Merkin et al., 2012; Nilsen and Graveley, 2010). While most genes in metazoan genomes encode only a few isoforms of mRNA, some can produce a large number of splicing isoforms (Pan et al., 2008; Wang et al., 2008), such as *CD44* (Screaton et al., 1992), *neurexin* (Südhof, 2008; Ullrich et al., 1995), and *clustered Protocadherin (cPcdh)* genes (Wu and Maniatis, 1999).

2.1.1 *Dscam* isoform diversity generated by alternative splicing is critical for neuronal development

The most extreme case is the *Drosophila melanogaster* homologue of *Down syndrome cell adhesion molecule* (*Dscam*) gene (Figure 2-1A)(Schmucker et al., 2000). The *Dscam* gene locus contains 115 exons, of which 95 are arranged into four clusters, *i.e.* exon 4, 6, 9, and 17, consisting of 12, 48, 33 and 2 variable exons, respectively. The variable exons within each cluster are spliced in a mutually exclusive manner, thereby generating potentially up to 19,008 isoforms encoding different assortments of immunoglobulin domains with differential adhesive properties (exon 4, 6 and 9 clusters) as well as two different transmembrane domains

(exon 17 cluster) (Schmucker et al., 2000). In addition, four different cytoplasmic domains could be generated by exon skipping (Yu et al., 2009). Importantly, a series of functional studies have demonstrated that a large isoform diversity is essential for its functions in both nervous and immune system (Chen et al., 2006; Dong et al., 2006, 2012; Hattori et al., 2007, 2009; Hughes et al., 2007; Matthews et al., 2007; Soba et al., 2007; Watson et al., 2005; Watthanasurorot et al., 2011; Zhan et al., 2004).

Specifically, it has been shown that dendrites that express identical Dscam isoforms on their surface repel each other (Hughes et al., 2007; Matthews et al., 2007; Soba et al., 2007). In wildtype conditions, neighboring neurons with overlapping dendritic fields express different isoforms. This limits the Dscam-Dscam interactions to sister dendrite interactions supporting self-avoidance. If the diversity of *Dscam* isoforms is decreased such that neighboring neurons also express identical isoforms, heteroneuronal repulsion occurs leading to wiring defects (Hattori et al., 2009). This illustrates that there exists critical thresholds of isoform diversity as such and also suggests that there might be additional cellular control mechanisms that ensure that different neurons express non-overlapping sets of isoforms. While receptors closely related to Dscam exist in higher vertebrates, it is surprising that they do not show a high degree of alternative splicing. It has however been proposed recently, that in vertebrate other diverse receptors and in particular the cPcdh receptors provide the functional counterpart to the Dscam isoform diversity (Schmucker and Chen, 2009; Zipursky and Sanes, 2010). This hypothesis is supported by the recent finding that cPcdh γ receptors are important for self-avoidance in retinal cells in mice (Lefebvre et al., 2012) and that due to alternative splicing and tetramer formation of cPcdhs, tens of thousands of homophilic binding specificities can be generated (Schreiner and Weiner, 2010). Overall this shows that the generation of receptor diversity by means of alternative splicing is of general importance for the process of neuronal wiring specificity, thereby emphasizing the importance of applying novel systematic and quantitative isoform expression analysis in order to dissect the underlying molecular mechanisms.

2.1.2 Current status and limitations on investigation of *Dscam* isoform expression

The alternative splicing of *Dscam* during development and in different tissues/cell types has been previously studied using customized microarrays and other PCR-based methods. These studies demonstrated both temporal and spatial regulation of splicing choices from exon 4, 6 and 9 clusters (Celotto and Graveley, 2001; Neves et al., 2004; Watson et al., 2005; Zhan et al., 2004) . The observations suggested a 'stochastic yet biased' splicing model, in which *Dscam* isoform profiles arise from a series of stochastic splicing events (Neves et al., 2004). The inclusion probability of individual variable exons is determined by the interaction between various RNA elements and specific splicing factors expressed in different cell types (Anastassiou et al., 2006; Graveley, 2005; Kreahling and Graveley, 2005; May et al., 2011; Olson et al., 2007; Park et al., 2004; Wang et al., 2012; Yang et al., 2011).

A major technical limitation in conventional profiling of *Dscam* isoforms lies in the fact that choices of the variable exons can only be investigated for each cluster separately. The frequencies of different transcript isoforms have then to be inferred based on various assumptions, for example, that alternative splicing occurs independently at different clusters. Two studies have suggested such an independent splicing mode (Chen et al., 2006; Neves et al., 2004). However, whether it indeed holds true awaits a more direct experimental examination, where the complete transcripts could ideally be quantitatively profiled. Recently, massive parallel shotgun cDNA sequencing (RNA-seq) has been used for high-throughput mRNA profiling (Cloonan et al., 2008; Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009; Wilhelm et al., 2008). But with limited read length, standard RNA-seq methods are unable to directly identify combinations of more than two *Dscam* variable exons. Moreover, given its enormous diversity, computational inference of *Dscam* isoform composition based on shotgun sequencing data would be impossible.

2.1.3 Aims of this study

In this study, we developed CAMSeq (Circularization-Assisted Multi-Segment Sequencing), a novel method that enables to quantitatively profile the expression pattern of *Dscam* isoforms consisting of exon 4, 6 and 9. We analyzed the splicing pattern of the three exon clusters at different developmental stages and in different cells/tissues. With unprecedented sequencing depth, out of 19,008 theoretically possible ones, we could detect 18,496 isoforms. They expressed across a broad dynamic range, and showed different splicing patterns at different stages as well as in different cells/tissues. Furthermore, we demonstrated that alternative splicing between different exon clusters were largely independent. Finally, our data suggest a surprisingly strong bias in isoform expression. Taken together our quantitative method and measurements enable now a thorough evaluation of how much protein diversity globally as well as per cell is essential to support a robust system distinguishing between self and non-self neurites.

2.2 Materials and Methods

2.2.1 RNA from fruit fly samples

Fruit flies from *D. melanogaster* J5 strain were raised on standard fruit fly medium at room temperature or at 25 °C. Fruit flies from embryonic, 1st larval, 2nd larval and 3rd larval stages were collected according to the time period after egg laying (embryos 13-18 h, 1st stage larvae 24-36 h, 2nd stage larvae 60-72 h and 3rd stage larvae 96-108 h). Fruit fly pupae were collected 0-48 h after puparium formation. Adult brains were dissected from 1-3 days old female after eclosion. S2 cells were maintained in Schneider's medium with 10% fetal bovine serum and 100 ng/µl of penicillin/streptomycin at room temperature. Total RNAs from fruit fly samples and S2 cells were isolated using TriZOL reagent according to manufacturer's instruction (Life Technologies).

2.2.2 *Dscam* reference RNA samples

Reverse transcription (RT) was performed on 5 µg of embryonic fruit fly total RNA with a specific primer annealed to the constitutive exon 19 (5' TGTCCTGGTGGAAGCATAG 3') using SuperScript III system with a reaction volume of 20 µl (Life Technologies). PCR was followed using 2 µl of reverse transcription product as template in 25 µl of GoTaq PCR system (Promega). The PCR primers were targeted at constitutive exons 3 and 11.

DsRef-1-F: 5'-GAGGTCCATGCCCAGGTGTACG-3'

DsRef-1-R: 5'-GTCGACATGCAGAGTGCCCTC-3'

PCR was run as following, 2 min at 95 °C, followed by 30 cycles of 30 s at 95 °C, and 2.5 min at 72 °C, and a final elongation of 10 min at 72 °C. PCR product was purified using Agencourt AMPure XP system (Beckman Coulter) and then cloned into pGEM-T Easy Vector, transformed into JM109 competent cells and plated onto LB/ampicillin/IPTG/X-gal plates according to manufacturer's instruction (Promega). Plasmids from positive colonies were purified using GeneJET plasmid DNA purification kits (Thermo Scientific) and sequences of inserted *Dscam* isoform cDNAs were confirmed using Sanger sequencing method. Plasmids from eight colonies containing different combinations of exon 4, 6 and 9 were selected. Using these eight plasmids as templates, another PCR was performed in 25 µl of Advantage 2 PCR system (Clontech) using the forward and reverse primers targeted at constitutive exons 3 and 11, with T7 promoter sequence attached at the end of forward primer.

Dsref-2-T7-F:
5'-GGATCCTAATACGACTCACTATAGGGATCCATTATCTCCCGGGACGTCC
ATGT-3'

DsRef-2-R: 5'-GTCGACATGCAGAGTGCCCTC-3'

After purification and measurement of concentrations and fragment sizes using Qubit

system (Life Technologies) and Agilent 2100 Bioanalyzer (Agilent), the eight PCR products were used as templates for i*n vitro* transcriptions with mMESSAGE mMACHINE T7 kit (Life Technologies). The resulting RNA samples were purified using Agencourt RNAClean system (Beckman Coulter) and quantified by Qubit system. The eight RNAs were then mixed together in different amounts.

2.2.3 CAMSeq

RT was performed on either 5 µg of total RNA from fly sample or 10 pg of the mixture of *Dscam* reference RNA samples with a primer annealed to the constitutive exon 11 (5' GTCGCTCTTCTTTAGATCCTTGTAC 3') using SuperScript III system with a reaction volume of 20 µl. The 1st round PCR was followed using 2 µl of RT product as template in 25 µl of Advantage 2 PCR system. The PCR primers were targeted at constitutive exons 3 and 10 with indexed barcode sequences attached at 5' ends.

CAMSeq-1-F: 5'-AGNNNNACCATTATCTCCCGGGACGTCCATGTGC-3'

CAMSeq-1-R: 5'-GTNNNNACCTTATCGGTGGGCTCGAGGATCCA-3'

("NNNN" represents barcode sequences.)

PCR was run as following, 2 min at 95 °C, followed by 22 cycles of 30 s at 95 °C, and 2.5 min at 72 °C, and a final elongation of 10 min at 72 °C. The products of 1st round PCR were purified and eluted into 10 µl of water using Agencourt AMPure XP system. After the measurement of concentrations and fragment size on Qubit system and Agilent 2100 Bioanalyzer, the purified 1st round PCR products obtained from different samples were then mixed together in equal amounts. The mixture was run on agarose gel, and DNA fragments with sizes between 1,500 bp and 2,500 bp were excised, purified and eluted into 20 µl of water using Qiaquick gel extraction kit (Qiagen). The product was then end-repaired using NEBNext End Repair Module (NEB), purified using Agencourt AMPure XP system. After measuring the concentration with Qubit system, 60 ng of the end-repaired product was used for

circularization reaction following the manufacturer's instruction (Illumina). The circularization product was purified using Agencourt AMPure XP system and quantified using Qubit system. Using 1 ng of purified circularization product as template, the 2nd round PCR was then performed in 100 μl of Phusion PCR system (Thermo Scientific). The PCR primers were targeted at constitutive exons 7 and 8 with Illumina adapters attached to the 5' ends.

CAMSeq-2-F:
5'-AATGATACGGCGACCACCGAGATCTACACTGGATACTCTGCTCGAGGATCTCTGGAAGTGC-3'

CAMSeq-2-R:
5'-CAAGCAGAAGACGGCATACGAGATCGGTCCAGCTTGTTTACGGGTTGTTCCTTCGATGA-3'

PCR was run as following, 1 min at 98 °C, followed by 15 cycles of 30 s at 98 °C, and 1.5 min at 72 °C, and a final elongation of 5 min at 72 °C. The product of 2nd round PCR was purified and eluted into 10 μl of water using Agencourt AMPure XP system. After the measurement of concentration and fragment size by Qubit system and by Agilent 2100 Bioanalyzer, the purified product was sequenced using Illumina GAIIX following manufacturer's instruction with the following modifications. On one flowcell, we performed a total of four sequencing using four specific sequencing primers targeting constitutive exon 10, exon 8, exon 5 and exon 3, respectively.

CAMSeq-barcode-seq-primer:
5'-CCTCCCAGATGGATCCTCGAGCCCACCGATAAG-3'

CAMSeq-ex9-seq-primer:
5'-GATACTCTGCTCGAGGATCTCTGGAAGTGCAAGTCA-3'

CAMSeq-ex6-seq-primer:
5'-CGATTAAGTGCCACAAAAGGACGATTGGTCATCA-3'

CAMSeq-ex4-seq-primer: 5'-CCATTATCTCCCGGGACGTCCATGTGCGAG-3'

After sequencing for each primer, the sequencing primer and the synthesized strand were washed away. By running the four sequencing for 25, 36, 36 and 36 cycles, respectively, we obtained for each DNA template molecule four sequencing reads derived from the barcode, variable exon 9, exon 6, and exon 4, respectively.

2.2.4 Pacific Bioscience (PacBio) RS sequencing of *Dscam* isoforms

The 2 kb 1st round RT-PCR product obtained from S2 cells, as described in the previous section, was directly sequenced using PacBio RS system according to the manufacturer's instruction (Pacific Biosciences).

2.2.5 Processing of CAMSeq data

Each Illumina sequencing read was split into four segments derived from barcode (1st-25thnt), exon 9 (26th-61stnt), exon 6 (62nd-97thnt) and exon 4 (98th-123rdnt), respectively. The three segment sequences corresponding to exon 4, 6 and 9 were aligned to reference *Dscam* exon sequences (http://www.ncbi.nlm.nih.gov/nucleotide/AF260530?tool=FlyBase) using bowtie2 (parameters: --very-sensitive-local -5 3). Only the reads with all the three segments that could be uniquely mapped to the respective exons were retained. The barcode segment, was used to extract the two barcode sequences derived from the 5' end of either forward or reverse primer in the 1st round PCR. The two barcode sequences were then compared with those used in the experiments. The reads containing the two barcodes with at most one mismatch from the used barcodes were retained. Those with the two barcodes derived from a same sample were used to calculate the isoform frequency, whereas those with the two barcodes derived from different samples were used to estimate the rate of forming chimeras.

2.2.6 Processing of PacBio sequencing data

Circular consensus reads obtained from PacBio sequencing were aligned to *Dscam* exons using BLAT (parameters: -tileSize=8 -stepSize=5 -oneOff=1 -minScore=20 -minIdentity=70). We retained the sequences if and only if the identity of exon 4,

exon 6 and exon 9 could all be unambiguously revealed.

2.2.7 Estimation and correction of the chimeric effect

To estimate the rate of forming chimeras, we first identify the reads derived from the inter-molecule ligation between two different molecules in the circularization step (see Processing of CAMSeq data). In these reads, the sequences of exon 4, exon 6 and one barcode *b* (forward barcode) originate from one molecule, while the sequences of exon 9 and the other barcode *b*' (reverse barcode) are from a second molecule. Assuming a second-order reaction kinetics, the rate of forming chimeras *r* is given by

$$r = F_{4,6,9,b,b'} / (F_{4,6,b} \cdot F_{9,b'}),\ b \neq b' \quad (1)$$

where $F_{4,6,9,b,b'}$ is the frequency of the chimeric product containing a distinct set of exon 4, 6, 9 as well as forward and reverse barcodes, $F_{4,6,b}$ is the frequency of reads containing the same exon 4, exon 6 and forward barcode *b*, $F_{9,b'}$ is the frequency of reads containing the same exon 9 and reverse barcode *b*'. We calculated values $r = r_{4,6,9,b,b'}$ for all exon/barcode combinations with adequate expected numbers of reads (*i.e.*, $T \cdot F_{4,6,b} \cdot F_{9,b'} \geq 100$, where T is the total number of mappable reads). Assuming that the chimera rate is independent from actual exon/barcode combination, we treated the calculated $r_{4,6,9,b,b'}$ valuesas a set of independent variables. The slope r_{avg} of a linear regression line with intercept 0 through the points $x = F_{4,6,b} \cdot F_{9,b'}andy = F_{4,6,9,b,b'}$, $b \neq b'$, was used as an average chimeric rate (Appendix Figure S2-1).

For $b = b'$ and a given chimera rate *r*, we could use Equation (1) to calculate the expected number of chimeric reads by

$$E(X_{4,6,9,b}) = r \cdot F_{4,6,b} \cdot F_{9,b} \cdot T \quad (2)$$

We then corrected the observed number of reads per isoform and barcode by subtracting the (rounded up) number of chimeras given by Formula (2) for $r = r_{avg}$.

In order to estimate the total number of expressed isoforms with high confidence, we applied Formula (2) to compute the number of potential chimeric reads using a highly

conservative estimate of chimeric rate, i.e. an upper α-quantile from the distribution of all $r_{4,6,9,b,b'}$ values, where $\alpha = 1/n$, and n = 19,008, the number of theoretically possible isoforms. We then counted for each data set, the number of different isoforms for which the number of observed reads was higher than that of chimerical reads estimated in this conservative way. For each isoform, the probability to be a false positive is at most α, thus the expected number of false positives per data set is at most $n \cdot \alpha = 1$.

2.2.8 Computation of the effective *Dscam* isoform repertoire

Given the relative frequencies f_i for all n combinations of exon 4, exon 6 and exon 9, we can compute the probability P_{11} for two identical isoforms independently sampled from the same f_i distribution as

$$P_{11} = \sum_{i=1}^{n} f_i^2 \quad (3)$$

This probability gets minimal if all isoforms express with the same probability (uniform distribution); in this case $P_{11} = 1/n$. If the splicing on the other hand is biased towards certain isoforms, it is more likely that two independently sampled isoforms are identical, so in this case P_{11} is greater than $1/n$ and the ability of the cell to create distinctive *Dscam* identities is decreased. We define the effective size n_{eff} of the *Dscam* repertoire to be the number of uniformly expressed isoforms needed to get P_{11}:

$$n_{\text{eff}} = \frac{1}{P_{11}} \quad (4)$$

The probability for a single *Dscam* transcript to have the same isoform identity as one or more of k independently expressed transcripts is

$$P_{1k} = 1 - (1 - P_{11})^k \quad (5)$$

The probability that more than h out of k Dscam transcripts independently expressed in two cells share the same isoform identity is given by a binomial distribution:

$$P_{kk} = 1 - \sum_{i=0}^{h} \binom{k}{i} \cdot P_{1k}{}^{i} \cdot (1 - P_{1k})^{(k-i)} \quad (6)$$

If we assume for example that two distinct cells are allowed to share up to 20% of their *Dscam* transcripts, we would set $h = 0.2 \cdot k$. P_{kk} can then be interpreted as the probability for two distinct cells getting the same Dscam identity by chance, see Appendix Figure S2-4A.

For a set of m cells, the probability that each cell gets a unique identity is

$$Q = (1 - P_{kk})^{\frac{m \cdot (m-1)}{2}}, \quad (7)$$

where $m \cdot (m\text{-}1) / 2$ is the number of all possible pairwise combinations of the m cells. If we set for example Q=0.95, given P_{kk}, m could be computed, see Appendix Figure S2-4B.

2.2.9 Clustering of exon 4 and 9 based on the expression patterns

We created heatmaps using the heatmap.2 function from the R package gplots to visualize the expression pattern of exon 4 alternatives (row) and exon 9 alternatives (column) . The numbers of sequencing reads were first normalized column-wise for each exon 4 alternative, and then scaled row-wise by using the parameter scale="row". The rows and the columns were hierarchical clustered by complete-linkage clustering using distance metric $d = (1\text{-}R)/2$ where R is the Pearson correlation coefficient.

2.2.10 Decomposition of *Dscam* isoform distribution datasets

If exon 4, exon 6 and exon 9 are selected independently during splicing in a homogenous cell population, the expected frequency $f_{4,6,9}$ of each isoform is given by

$$f_{4,6,9} = f_4 \cdot f_6 \cdot f_9 , \quad (8)$$

where f_4, f_6 and f_9 are the (marginal) frequencies for the exon 4, exon 6 and exon 9, respectively. If on the other hand the cell population consists of two cell types A and B with distinct splicing bias, Equation (8) may not hold. Instead we assume

$$f_{4,6,9} = \mu \cdot f_4^A \cdot f_6^A \cdot f_9^A + (1-\mu) \cdot f_4^B \cdot f_6^B \cdot f_9^B \quad (9)$$

where μ is the ratio in which the *Dscam* transcripts from *A* and *B* are mixed together. Assuming the f_6 is very similar between different cell types, Equation (9) could be simplified to

$$f_{4,9} = \mu \cdot f_4^A \cdot f_9^A + (1-\mu) \cdot f_4^B \cdot f_9^B \quad (10)$$

where $f_{4,9}$ is the expected frequency for a combination of exon 4 and exon 9.

For different fixed values μ we tried to find distributions f_4^A, f_9^A, f_4^B, f_9^B fitting to Equation (10) with minimum total log squared error. Starting with $f_4^A = f_4^B = f_4$ and $f_9^A = f_9^B = f_9$ we optimized the distributions in up to 500 rounds, where in each round we optimized for each exon 4 and exon 9 separately in random order, *i.e.* we adjusted the frequency of each variable exon such that the objective function was minimized.

2.3 Results

2.3.1 Development of CAMSeq, a novel method that enables the quantitative profiling of *Dscam* ectodomain isoforms

The genomic structure and splicing model of *Dscam* is shown in Figure 2-1A. In this study, we focused on the combinations of variable exon 4, 6 and 9, since alternative splicing in these three clusters could generate theoretically up to 19,008 different ectodomains, which contain the actual domains of the Dscam protein determining its recognition specificity. Since Illumina paired-end sequencing could only yield up to 150 nt sequences from both 5' and 3' ends of cDNA fragments shorter than 1 kb, it cannot be used to directly determine for each isoform the precise combination of the three variable exons. Therefore, we developed a new method termed 'CAMSeq', the novelty of which consists of two major components: 1) circularization followed by another PCR reduces the size of cDNA fragments to be sequenced; 2) multi-segment

sequencing yields multiple exon sequences from a same cDNA molecule. The scheme of CAMSeq is illustrated in Figure 2-1B. In brief, first, using RT-PCR with the barcode-indexed primers targeting constitutive exon 3 and exon 10, *Dscam* mRNA was reverse-transcribed and amplified. PCR products derived from different samples and labeled with different barcodes were pooled together. After circularization of the pooled 2 kb RT-PCR product and another round of PCR with the primers targeting constitutive exon 7 and exon 8, the amplification product of approximately 1 kb in length was then sequenced. As shown in Figure 2-1B, we modified the standard Illumina sequencing procedure and obtained from every template DNA molecule four sequencing reads (quadruple-reads) derived from exon 4, 6, 9 and barcode, respectively (Methods). Thereby, we could identify the exon usages simultaneously in the three clusters and unambiguously reveal the identity of the expressed isoforms.

To evaluate the accuracy of our method, we first generated a set of eight *Dscam* mRNAs with known concentrations by *in vitro* transcribing cloned *Dscam* cDNAs containing different combinations of exon 4, 6 and 9. We then prepared two reference samples by mixing these eight RNAs in different amounts spanning five orders of magnitude and applied CAMSeq on these two samples (Methods). As shown in Figure 2-1C, a straight linear relationship spanning the full dynamic range was observed between the RNA amount and the number of sequencing reads derived from each RNA, demonstrating that our method provides an accurate estimation of the relative abundance of different isoforms. Furthermore, to examine the reproducibility of our method when applied to biological samples, we analyzed twice the same RNA extracted from S2 cells, a cell line derived from *Drosophila* embryonic hemocytes. As shown in Figure 2-1D, the isoform profiles from the two replicates were highly correlated (R^2=0.993). Finally, to assess a potential systematic bias caused by cDNA circularization, the second round of PCR as well as Illumina sequencing procedure, we also measured the isoform abundance in S2 cells by directly sequencing the 2 kb RT-PCR products using PacBio RS system (Methods). A total of 63,109 PacBio reads could be used to reveal the identities of exon 4, 6 and 9 for 3,725 *Dscam* isoforms and

the isoform abundances estimated using the two approaches showed a high correlation (R^2=0.978; Figure 2-1E).

During the preparation of the sequencing libraries, in the circularization step, in addition to self-circularization, chimeras could also form from intermolecular ligation events where two or more DNA molecules are joined together. Although occurring at a much lower frequency compared to self-circularization, these chimeras could nevertheless lead to an overestimation of the number of isoforms that were detected. To rule out the 'chimera' effect and obtain an accurate number of detected isoforms, we estimated the rate of forming chimeras by counting the number of apparent chimeras in which Dscam cDNAs derived from different samples and were labeled with different barcodes joined together (Methods). As shown in Figure 2-1F, the mean chimerical rate was approximately 1%.

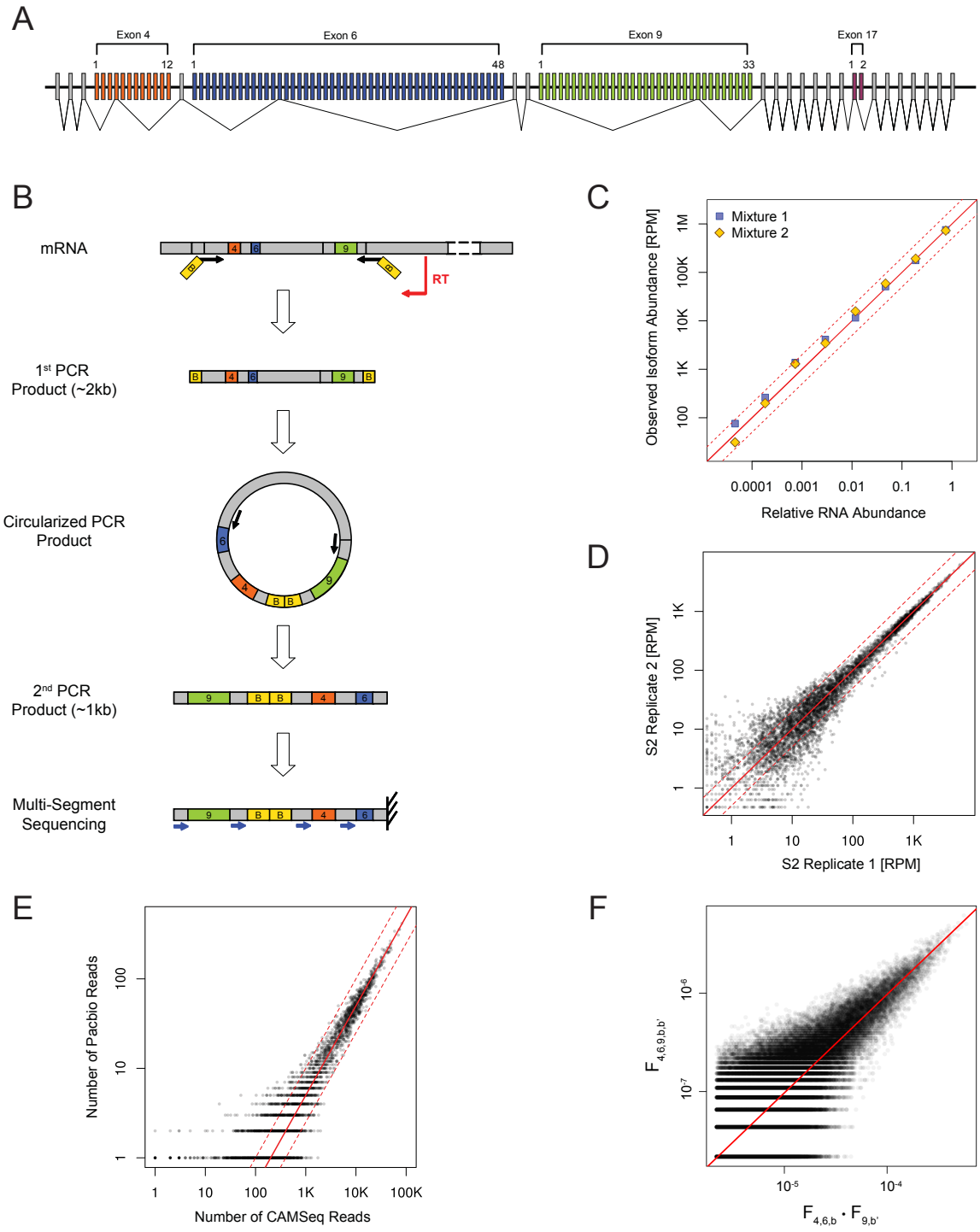


Figure 2-1. CAMSeq, a novel massive parallel sequencing based method for quantitative profiling Dscam isoforms. (**A**) Alternative splicing of *Drosophila Melanogaster Dscam* gene. Constitutive exons are depicted in grey whereas alternative exons from exon clusters 4, 6, 9 and 17 are depicted in orange, blue, green and purple, respectively. The exon 4, 6 and 9 alternatives code for the variable extracellular Immunoglobin (Ig) domains. The two exon 17 alternatives code for the

single transmembrane domain. (**B**) Outline of CAMSeq. In brief, first, using RT-PCR with the barcode-indexed primers targeting constitutive exon 3 and exon 10, *Dscam* mRNA was reverse-transcribed and amplified (Barcodes are depicted in yellow and marked with "B"). After circularization of the approximately 2 kb RT-PCR product and another round of PCR with the primers targeting constitutive exon 7 and exon 8, the amplification product of approximately 1 kb in length was then sequenced. Here, using a modified sequencing procedure with four specific sequencing primers targeting constitutive exon 3, exon 5, exon 8 and exon 10 respectively (blue arrows), we obtained from every template DNA molecule four sequencing reads derived from exon 4, 6, 9 and barcode, respectively. (**C**) Analysis of sequencing data obtained from two controlled mixtures of eight *in vitro* synthesized *Dscam* mRNAs. The relative RNA abundance (X-axis) was plotted against the normalized number of derived sequencing reads (RPM, reads per million total reads) (Y-axis). (**D**) Comparison of sequencing data obtained from two replicate experiments on the same total RNA extracted from S2 cells showed a high correlation (R^2=0.993). (**E**) Comparison of sequencing data obtained from CAMSeq to that from PacBio sequencing showed a high correlation (R^2=0.978). (**F**) Estimation of chimeric rate. To estimate the rate of forming chimeras, we first count the number of chimeric reads derived from different samples with different barcodes joined together. Here Y-axis represent the frequency of such chimeric reads F4,6,9,b,b', whereas X-axis is the product between the frequency of reads containing the same exon 4 and 6 as well as the same forward barcode, and that of reads containing the same exon 9 as well as the same reverse barcode, F4,6,b ·F9,b'. Assuming a second-order reaction kinetics, the mean chimerical rate could be represented by the slope of regression line, i.e. approximately 1% (Methods).

2.3.2 Detection of *Dscam* isoforms expressed at different developmental stages and in different cells/tissues

We used CAMSeq to analyze *Dscam* isoform expression at different developmental

stages (embryos, first instar larvae (L1), second instar larvae (L2), and third instar larvae (L3), and pupae) and in adult brain. Here, *Dscam* cDNAs from each sample were amplified with the primers containing distinct barcode sequences at both 5' and 3' ends (Figure 2-1B; Methods). The PCR products from different samples were then pooled in equal amounts and the mixture went through the remaining steps as described above. For each sample, we obtained between 5.71 and 15.22 million quadruple-reads that could be used to unambiguously identify the usage of exon 4, 6 and 9 as well as the barcode representing a specific sample (Appendix Table S2-1). In all samples, we could detect the presence of all variable exons from exon 4, 6 and 9 clusters, except exon 6.11. During development, the exon usages in cluster 4 and 9 showed moderate to dramatic changes, whereas the differences in exon 6 clusters were relatively modest (Figure 2-2A).

After subtracting all potential chimerical reads, we detected with high confidence between 13,216 and 16,886 isoforms in each sample, and 18,496 isoforms in at least one sample (Methods; Appendix Table S2-1). The number was quite close to 18,612, the maximum number of potential isoforms if excluding the pseudo-exon 6.11, indicating all the remaining *Dscam* isoforms expressed.

In each sample, the relative abundance of different isoforms spanned at least four orders of magnitudes (Figure 2-2B). The most abundant 10 and 100 isoforms derived 0.7% to 2.0%, and 5.3% to 12.2% of all reads from one sample (Appendix Table S2-1). Importantly, our comparison between different samples showed that S2 cells express a significantly more restricted repertoire of *Dscam* isoforms in which only 7,317 isoforms were detected with the most abundant 100 isoforms accounting for 25.6% of all reads (Appendix Table S2-1). Such striking difference between S2 and all the other samples might be explained by the fact that S2 cells are a homogeneous cell population whereas other samples consist of different types of *Dscam* expressing cells with the splicing preferences towards different sets of variable exons. As a result, at a similar sequencing depth, we could detect much fewer isoforms in S2 cells. Interestingly, when we compared dynamic ranges of different exon usages in the three

clusters between S2 and other samples, it turned out that exon 9 cluster expressed a relatively limited set of exons in S2 compared to other samples (Figure 2-2A). Given the observation that the splicing choice of exon 9 was most variable between different cell types (Figure 2-2A), this corroborates our hypothesis that the larger repertoire observed in the other samples was due to the much higher cell-type diversity.

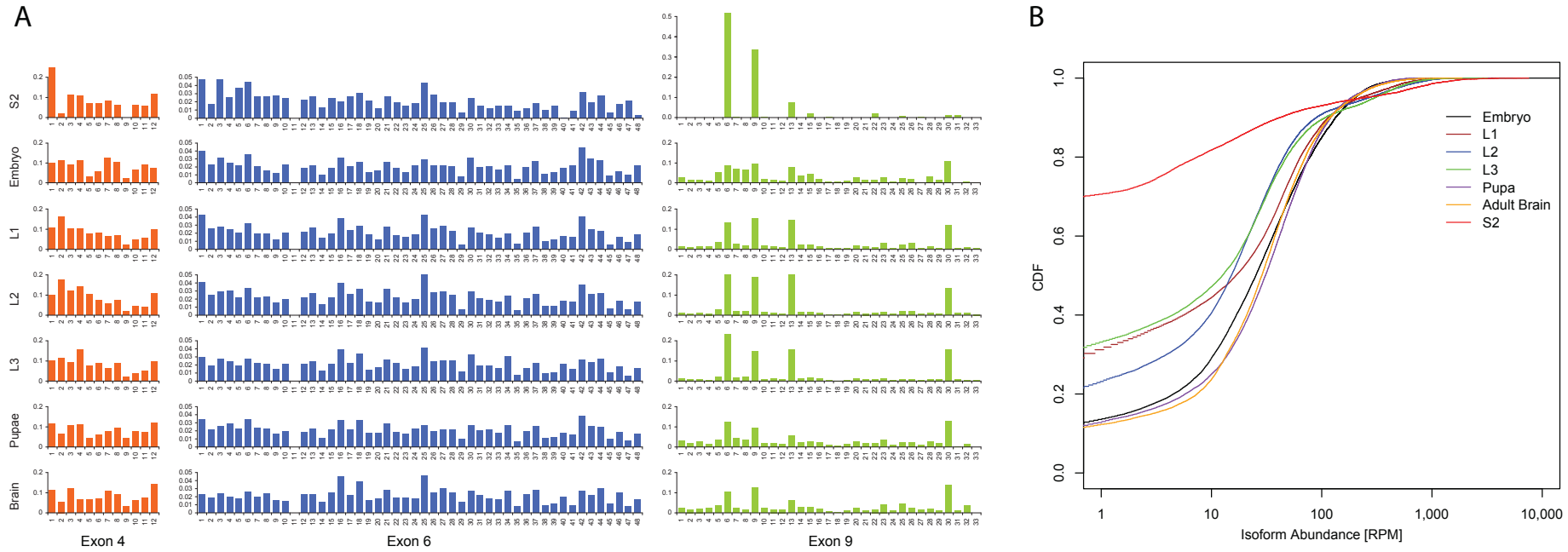


Figure 2-2. The relative expression of *Dscam* variable exons and *Dscam* isoforms during development and in S2 cells. (**A**) The relative expression of variable exon 4, 6 and 9 in different samples. (**B**) Cumulative distribution function of abundances (RPM) of *Dscam* isoforms in different samples.

2.3.3 Independent splicing choice between the different exon clusters

To address whether the splicing choices at different exon clusters are independent or not, we first estimated the relative abundance of all variable exons, and then assuming an independent splicing model, calculated the expected relative frequencies of different isoforms by simply multiplying the frequencies of their respective variable exon 4, 6 and 9. Comparison between the observed and expected isoform frequencies in different samples showed mixed results (Figure 2-3A; Appendix Figure S2-1A). Whereas a straight linear relationship was observed in S2 cells, demonstrating unambiguously the independent splicing choice among different exon clusters, other samples showed only weak to modest correlations between the observed and expected frequencies (Figure 2-3A; Appendix Figure S2-1A).

Given that the splicing choice of exon 4 and 9, especially the latter, was quite variable between different cell types, we hypothesized that the different observation between S2 and other samples was due to the fact that other samples consisted of different cell types expressing distinct sets of exon 4s and 9s. To corroborate this hypothesis, we further analyzed splicing choices between exon 4 and 6, exon 6 and 9, as well as exon 4 and 9, separately. Indeed, whereas the splicing appeared to be independent between exon 4 and 6, as well as between exon 6 and 9 in all samples (Figure 2-3B, C; Appendix Figure S2-1B, C), the splicing between exon 4 and 9 showed different patterns between S2 and other samples (Figure 2-3D and Appendix Figure S2-1D). Notably, we could cluster the variable exon 4s and 9s based on their expression patterns in adult brain and other samples. As shown in Figure 2-3E and Appendix Figure S2-2A, exon 9s could be clearly divided into two groups, one containing only five exons and the other consisting of the remaining 27. Given the differential usages of variable exon 4s within the two groups, we could *in silico* decompose the whole adult brain data into two sets with different usages of exon 4s and 9s, and the splicing choices within each dataset being largely independent between the two clusters (Figure 2-3F, G). In a similar way, other samples could also be decomposed into two or three groups expressing distinct sets of exon 4s and 9s, and all with independent

splicing choices among different exon clusters (Appendix Figure S2-2B and Appendix Figure S2-1E, F).

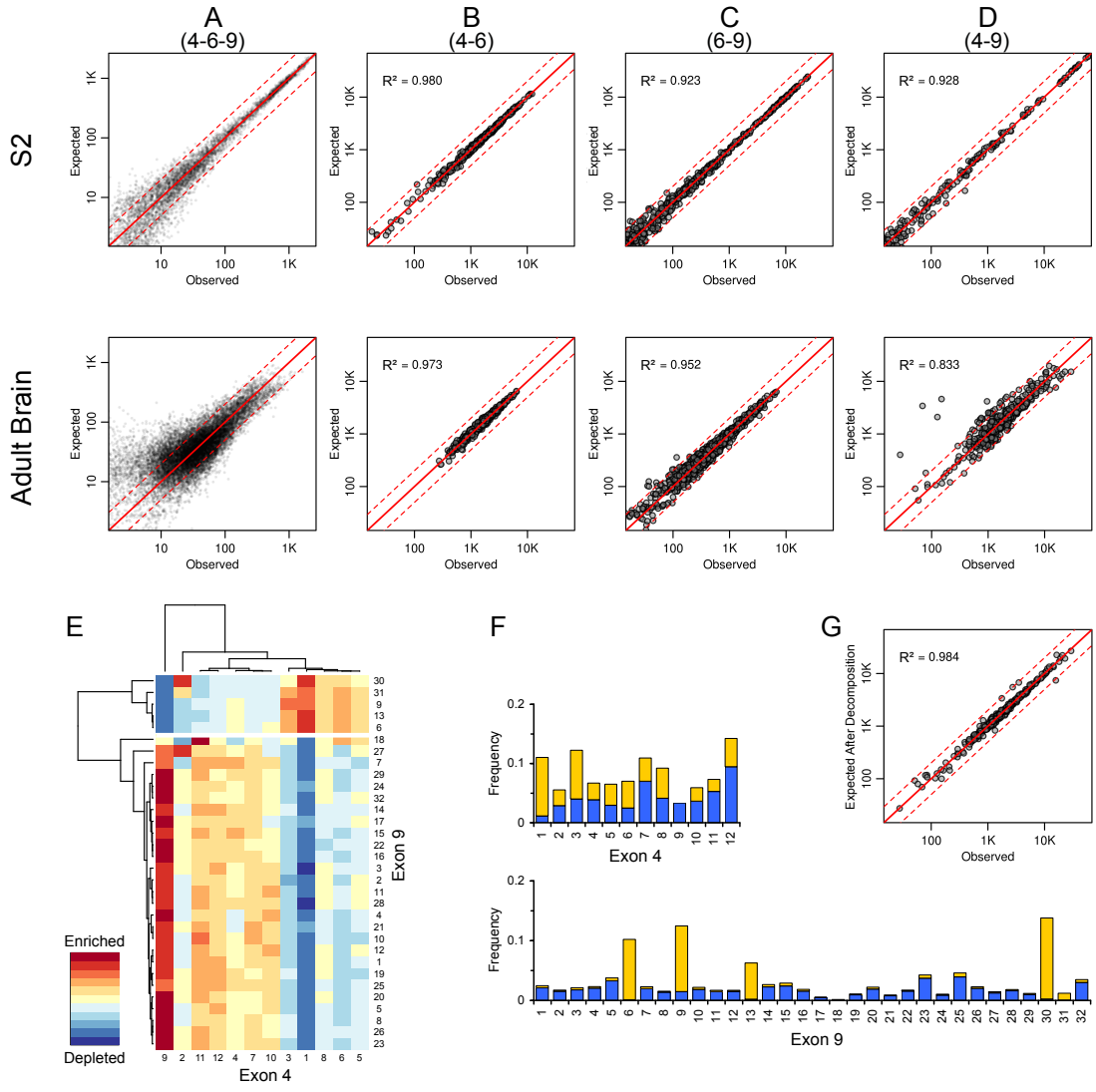


Figure 2-3. Independent splicing choice between the different variable exon clusters. (**A**) Observed isoform frequencies were depicted in X-axis. Expected frequencies were calculated by multiplying the frequencies of their respective variable exon 4, 6 and 9, and depicted in Y-axis. To determine whether the splicing between the three clusters was independently controlled, the two frequencies were compared. (**B**), (**C**), (**D**): In a similar way, we determined whether the splicing choices were independent between exon 4 and 6 (**B**), exon 6 and 9 (**C**), exon 4 and 9 (**D**),

respectively. (**E**). In adult brain sample, the variable exon 4s and 9s were clustered based on their expression patterns, the exon 9s could be clearly divided into two groups, one containing only five exons and the other consisting of the remaining 27. (**F**) Given the differential usages of variable exon 4 within the two groups, the whole brain data were *in silico* decomposed into two sets with different usages of exon 4 and 9, the yellow and blue groups. (**G**) The splicing choice within each group was largely independent between exon 4 and 9. X-axis depicted the observed isoform frequencies from the whole brain dataset, whereas in Y-axis, the expected isoform frequencies were the sum of the expected frequencies of the two groups.

2.4 Discussion

2.4.1 Novelty and advantages of CAMSeq

We developed CAMSeq, a new massive parallel sequencing based approach for quantitatively profiling *Dscam* isoform expression. All previous global analyses of the alternative splicing of *Dscam* using microarrays measured the relative abundance of variable exons from different clusters separately. In contrast, our new method allows identifying the expression of *Dscam* isoforms directly by determining for each isoform the precise combination of exon 4, 6 and 9. Furthermore, our sequencing approach provided an accurate quantitative measurement, demonstrated by several control experiments. Finally, the sequencing depth achieved in this study enabled us to detect almost all the possible isoforms except those containing pseudo-exon 6.11. This is consistent with the previous findings (Celotto and Graveley, 2001; Neves et al., 2004) and the observation that the amino acid sequence of exon 6.11 lacks critical residues essential for proper Immunogloblin (Ig) domain folding (Dietmar Schmucker, personal communications). Notably, we could detect a very minor fraction of isoforms skipping either of exon 4, 6 or 9, consistent with previous observations (Kreahling and Graveley, 2005). Taken together, with the unprecedented sequencing depth, we achieved an ultra-high sensitivity of detecting lowly expressed isoforms

without detection of any false positive sequences.

2.4.2 Independent splicing choices between the different exon clusters

We demonstrated that alternative splicing between different exon clusters is independent in a uniform cell population (S2 cells). In a previous study, using a genetic approach, Chen *et al* generated two fly lines in which different parts of exon 4 clusters were deleted. Subsequent expression analysis of the splicing pattern in exon 6 and 9 clusters revealed no significant difference between the larval central nervous system (CNS) of the control and that of the two mutant strains, implicating that splicing of exon 6 and 9 are independent from that of exon 4 (Chen et al., 2006). In another study, using S2 cells, Neves *et al* analyzed the relative abundance of variable exon 4s and 6s in the isoforms containing two different exon 9s. They did not find specific exon 4 and 6 alternatives associated with either of the two exon 9s and therefore suggested splicing choices of the three clusters were independent (Neves et al., 2004). Our quantitative data are consistent with these findings, and, for the first time, provided direct and comprehensive experimental evidence for the independent splicing regulation of the three exon clusters in a distinct cell type.

However, in more complex samples, we observed some potential splicing dependence, especially between exon 4 and 9. We attributed such observation to the cellular heterogeneity of these samples. While the splicing is independent within a distinct cell type, different types of cells with differential usages of exon 4s and 9s, combined together, could give the misleading impression of dependence, as demonstrated by our *in silico* data decomposition (Figure 2-3 and Appendix Figure S2-1, S2-2).

2.4.3 Influence of *Dscam* isoform expressing patterns on neuronal self-avoidance

Dscam diversity is essential for neurite self-avoidance and plays a profound role in wiring the fruit fly brain. Using an elegant genetic approach, the Zipursky lab demonstrated that thousands of isoforms are essential to provide neurons with a robust mechanism to distinguish between self and non-self during self-avoidance (Hattori et

al., 2009). Moreover, they used mathematical modeling to support the hypothesis that the full molecular diversity encoded by the *Dscam* gene locus is almost five times larger than what may be considered as necessary. However, in such a model, all the potential isoforms were randomly sampled with equal probability. Apparently, such assumption of uniform isoform expression is an oversimplification. Starting with a realistic *in vivo* data set, we performed a similar modeling study using our actual quantitative datasets. First, we estimated the number of different isoforms that could be obtained by randomly sampling certain numbers of *Dscam* mRNA copies. As expected, this number is dependent on the biased choice of the variable exons. Due to the biased exon usage, the number of distinct isoforms that could be present in a certain number of neurons is much lower than that under the assumption that all isoforms expressed with equal probability (Figure 2-4A). For example, *Drosophila* mushroom body (MB) comprises some 2,500 neurons. If each individual MB neuron expresses 20 *Dscam* mRNA copies(Zhan et al., 2004) and all possible isoforms express with equal probability, about 17,300 different isoforms would be present in one MB. In contrast, if the isoforms express based on the pattern we measured from the adult brain, only 12,300 different isoforms would be present in one MB (Figure 2-4A; Methods). With such reduced repertoire, obviously the number of neurons with unique Dscam identity also decreases (Figure 2-4B; Methods). For instance, if up to 20% of Dscam isoforms were allowed to share between two neurons, under the assumption of uniform isoform expression, 68,500 neurons could be distinguished from each other. But with the more realistic size of adult brain *Dscam* repertoire evaluated by our quantitative CAMSeq analysis, only 3,200 neurons could be uniquely labeled (Figure 2-4B; see Appendix Figure S2-3B for the conditions in which up to 0% or 10% of isoforms were allowed to share between two neurons). The same labeling capacity could also be coded by about 5,500 uniformly expressed isoforms. To facilitate the comparison of labeling capacities between cell types with different splicing biases, we suggest to define the effective size of a certain *Dscam* repertoire as the number of uniformly expressed isoforms that could label the same number of neurons with unique identity (Methods).

Obviously the fruit fly nervous system consists of many different cell types expressing different *Dscam* splicing repertoires. As suggested by our decomposition analysis of the brain dataset (the yellow group in Figure 2-3F), it is very likely that in some distinct types of neurons, the *Dscam* isoform repertoire might be similarly small as observed in S2 cells. Due to the usage of a rather limited set of variable exons, the number of different *Dscam* isoforms present in a certain number of cells would be quite small and only dozens of cells could be labeled with unique Dscam identities when any pair of neurons were allowed to share 20% of their expressed *Dscam* isoforms (Figure 2-4B). Indeed, based on these calculations we would suggest that the effective size of *Dscam* repertoire of S2 cells is only around 800. On the other hand, different types of neurons would manifest the preferences towards different sets of exons, thereby lowering the probability to share too many of the same isoforms. The low effective size of isoform repertoire is then counteracted by a cell-type specific splicing bias. Therefore, in spite of a smaller effective size of isoform repertoire within a distinct type of neurons, the interconnecting neurons, consisting of different cell types, can still easily discriminate self from non-self.

In general terms we would like to speculate that in any complex nervous system two "identity-labeling" strategies could be used for the proper wiring in a large group of interconnecting neurons. That is, they can either be a homogeneous cell populations with low bias in exon usage and thus expressing randomly from a relatively large surface receptor repertoire, or consist of different cell groups with each distinctly controlling the expression of a limited but selective sets of receptor isoforms. Notably, with the second strategy, surface receptor isoforms could be used not only to distinguish self and non-self (“individual identity”), but also potentially to differentiate between different groups of cells (“group identity”). In this scenario, the neurites from different types of neurons are allowed to connect with a higher probability than those from the same cell type. Although it is an intriguing model for understanding the neuronal wiring specificity, it needs the further experimental evidences to validate and improve, especially considering one recent published study

demonstrating the preferential formation of chemical synapses between sister neurons from the same precursor cell (Yu et al., 2012).

Genetic studies have been instrumental in understanding why the enormous *Dscam* molecular diversity is required in neuronal wiring. In these studies, connectivity phenotypes in different nervous systems were assessed in the strains with different sets of variable exons deleted (Chen et al., 2006; Hattori et al., 2009). Often, the effects of different deletions on the *Dscam* repertoire were implicitly assumed to be solely dependent on the number of deleted exons. Such assumption would hold true if all the variable exons express with equal probability. However, due to splicing bias, the effect will also depend on identities of the deleted exons. Importantly, we observed that the effect would be unequal in different cell types with distinct splicing patterns (Appendix Table S2-2). Counter-intuitively, there might be some extreme scenarios in which the effective *Dscam* repertoire could even increase when the exons with predominant splicing bias are removed (Appendix Table S2-2). In addition, the number of neurons that could be labeled with unique Dscam identity will become sensitive to the *Dscam* expression level when the repertoire gets sufficiently small and there is an optimal range of *Dscam* mRNA copies per cell that maximize the total number of labeled neurons (see S2 sample in Figure 2-4B). Normal wiring pattern could break down if the expression of *Dscam* fluctuates out of such a range. Therefore, due to all these complications, the results in the genetic studies need to be complemented by quantitative expression data in order to better interpret the influence of molecular diversity on neuronal wiring specificity.

Taken together, due to the biased usage of variable exons, which could be surprisingly strong in some distinct cell types, the accessible *Dscam* isoform repertoire is more restricted than previously appreciated. Moreover, the splicing of *Dscam* is determined by the interaction between its various RNA elements and specific splicing factors. Therefore, it seems clear that, dependent on the expression levels of the splicing factors and other interacting RNAs, the abundances of *Dscam*-accessible splicing complexes could fluctuate and thus lead to uncertainties in *Dscam* splicing outputs.

To accommodate these limitations, during evolution, which, as Francois Jacob put, is a tinkering process, *Drosophila Dscam* gene locus might have adapted to this limitation by way of expanding exon number to encode an extremely high isoform diversity (Jacob, 1977). Such diversity, although seemingly beyond the necessity, is nevertheless essential to assure the neurons with a robust discrimination system to distinguish between self and non-self.

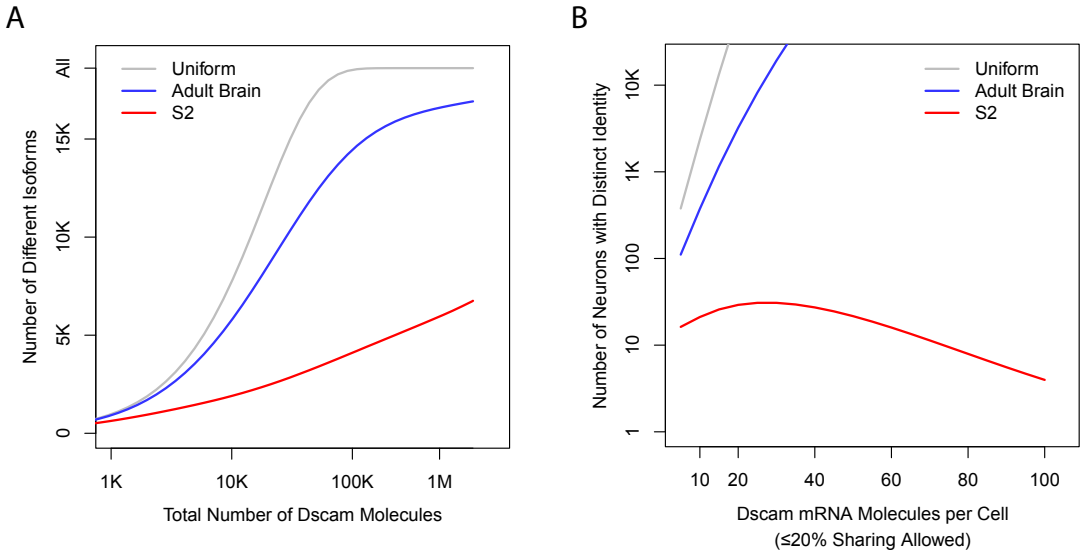


Figure 2-4. Monte Carlo simulation of *Dscam* repertoire and the number of neurons that could be labeled with unique *Dscam* identity. (**A**) The number of different isoforms (Y-axis) could be obtained by randomly sampling different numbers of *Dscam* mRNA molecules (X-axis) based on the distribution of *Dscam* isoform abundances in adult brain, S2 cells or a hypothetical uniform distribution (Methods). (**B**) The number of neurons that obtain unique identities at more than 95% likelihood (Y-axis) when each neuron expresses different numbers of *Dscam* mRNA molecules (X-axis), if allowing 20% of isoforms shared between any pair of neurons, calculated based on the distribution of *Dscam* isoform abundances in adult brain, S2 cells or a hypothetical uniform distribution (Methods). See Appendix Figure S4B for the condition in which up to 0% or 10% of isoforms are allowed to share between any pair of neurons.

Chapter 3. Full-length 5'RACE transcript sequencing-based annotation for *Xenopus tropicalis clustered Protocadherin* genes

3.1 Introduction

Neuronal self-avoidance is not unique in invertebrates. It is also observed in vertebrate nervous system. Although in the mouse genome there exists *DSCAM* orthologs *DSCAM1* and *DSCAML1*, they lack the molecular diversity as presented in their insect ortholog (Schmucker and Chen, 2009; Zipursky and Sanes, 2010). Thus, in mammals, the *DSCAM* gene couldn't provide the molecule coding ability for self-recognition. Instead, there must exist other genes performing such self-avoidance function in vertebrates as *Dscam* gene in insects. Indeed, in vertebrates, the most promising candidates are the *clustered Protocadherin (cPcdh)* genes (Chen and Maniatis, 2013; Chen et al., 2006; Hirayama and Yagi, 2013; Yagi, 2012; Zipursky and Sanes, 2010).

3.1.1 *cPcdh* genes as the functional counterpart of *Drosophila Dscam* gene for neuronal self-avoidance

In mouse, 58 *cPcdh* proteins are encoded in three tandem gene clusters (named *cPcdh* α cluster, β cluster, and γ cluster), encoding 14, 22, and 22 members, respectively (Figure 3-1) (Wu and Maniatis, 1999). The *cPcdh* α and γ RNA transcripts consists one large variable first exon and three small 'constitutive' subsequent exons: 1) the first large exons encoded the entire ectodomain, the transmembrane domain and a short cytoplasmic part. Due to alternative usage of first exon, the first exons are different for different *cPcdh* members, thus named as variable exons (VEs). 2) The three subsequent small exons encode the remaining cytoplasmic part, are shared among either different cPcdh α members or cPcdh γ members, and thus named as

constant exons (CEs). Comparing to *cPcdh α* and *γ* members, each *cPcdh β* RNA transcript contain only one VE, thus only have relatively shorter cytoplasmic parts and share no constant region among all cPcdh β members.

The *cPcdh* genes are considered as the functional counterpart of insect *Dscam* gene for vertebrate neuronal self-avoidance for several reasons. First, *cPcdh* genes are predominantly expressed in the neuronal systems, and almost all isoforms express in a scattered manner across the whole brain regions (Esumi et al., 2005; Kaneko et al., 2006; Noguchi et al., 2009; Yokota et al., 2011). In single neuron level, each neuron expresses multiple isoforms, and all isoforms are chosen and expressed in a stochastic and combinatorial manner (Esumi et al., 2005; Kaneko et al., 2006). Second, functional studies have demonstrated their functional importance in neuronal wiring (Chen and Maniatis, 2013). The loss of cPcdh α in mice leads to axon projection defects (Hasegawa et al., 2008, 2012; Katori et al., 2009). cPcdh γ knock-out in mice will lead to neonatal death with neurological defects, including neuron death and reduction in the synapse formation(Chen et al., 2012; Garrett and Weiner, 2009; Garrett et al., 2012; Lefebvre et al., 2008; Prasad and Weiner, 2011; Prasad et al., 2008; Wang et al., 2002; Weiner et al., 2005). Third, cPcdh proteins could form heteromultimeric protein oligomers from different cPcdh members (Rubinstein et al., 2015; Schreiner and Weiner, 2010; Thu et al., 2014). The heterotetramers and other possible multimers formed by cPcdh members are the homophilic binding unit that introduces cell-cell adhesions and interactions. In theory, such homophilic binding via the protein oligomers could dramatically increase the molecule tagging diversity provided from cPcdh members. Finally, one recent study observed that the conditional deletion of the mouse *cPcdh γ* genes could lead to the defect in dendritic self-avoidance, providing the critical functional connection between *Drosophila Dscam* gene and vertebrate *cPcdh* genes for neuronal self-avoidance (Lefebvre et al., 2012).

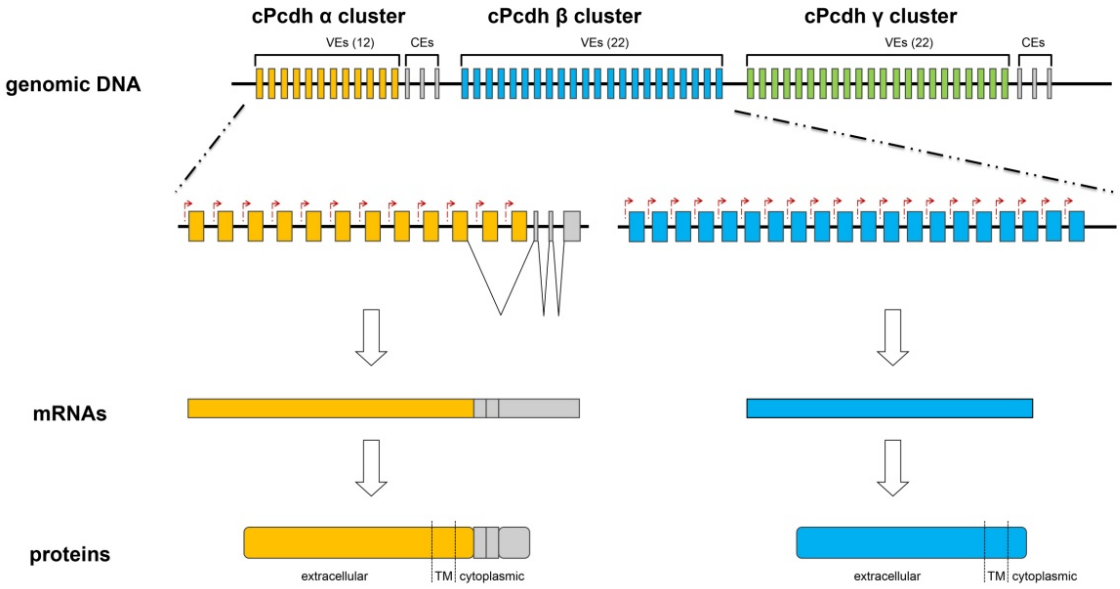


Figure 3-1. Genomic, mRNA and protein structures of the mouse *cPcdh* genes. Mouse *cPcdh* genes are organized into three gene clusters (*α*, *β* and *γ*) on their genome loci. *cPcdh α*, *β* and *γ* gene clusters could generate 12, 22, and 22 different isoforms, respectively, via alternative usage of first exon. These alternative first exons are named as variable exons (VEs), shown as yellow (*α*), blue (*β*), and green (*γ*). Through alternative first exon usage, *cPcdh α* or *γ* genes could generate different RNA isoforms by combining different VE with three following constant exons (CEs) (shown as grey) presenting at the 3' ends of all isoforms of *cPcdh α* or *γ* genes. On the other hand, RNA transcripts of *cPcdh β* genes only contain only one VE, without any CE. The cPcdh VEs encode the whole extracellular part, the transmembrane (TM) part, and a short part of cytoplasmic part, while the three cPcdh CEs encode only the C-terminal of the cytoplasmic part. Comparing with *cPcdh α* and *γ* genes, the RNA structure difference results into a shorter cytoplasmic part for the proteins encoded by *cPcdh β* genes.

3.1.2 Evolution of *cPcdh* genes

Evolutionarily, *cPcdh* genes are speculated as a vertebrate innovation, since they could be identified in all vertebrates investigated, but not in the vast majority of

invertebrate species (Jiang et al., 2009; Kohmura et al., 1998; Noonan et al., 2004a, 2004b; Sugino et al., 2004; Tada et al., 2004; Wu et al., 2001; Yu et al., 2007; Zou et al., 2007). More importantly, considering the general structural similarities of *cPcdh* genes among different species, it has been speculated that cPcdh proteins may serve conserved neuronal self-avoidance functions in all vertebrate species. Such property raises the possibility to use other vertebrate model organisms to investigate the functional mechanism of *cPcdh* genes in neuronal self-avoidance.

3.1.3 *Xenopus tropicalis (Xtro)* as potential model for functional study of *cPcdh* genes

The Western clawed frog, *Xenopus tropicalis (Xtro)*, is a powerful model organism, especially for developmental and neuronal studies. Its large brood size, embryonic transparence and *ex utero* development make it a suitable and valuable model for investigating the function of *cPcdh* genes in neuronal development (Harland and Grainger, 2011). However, the relatively poor assembly and annotation of *Xtro* genome becomes the obstacle for such study (Hellsten et al., 2010). So far there's no complete and appropriate annotation for *cPcdh* genes in *Xtro* genome, especially regarding their expressing and splicing patterns. As demonstrated in our study on *Drosophila Dscam* gene, the complete annotation for all isoforms of *cPcdh* genes would be the essential resource for studying the function of cPcdh in *Xtro* system.

3.1.4 Aims of this study

In this study, based on our latest genome analysis of *Xtro cPcdh* gene loci annotation from genome analyses (Etlioglu et al., 2016), we further expanded and refined the annotation of *Xtro cPcdh* genes by applying long-read full-length sequencing of the 5'RACE (Rapid Amplification of cDNA Ends) products derived from *Xtro cPcdh* mRNAs. For the three *cPcdh* gene clusters existed in *Xtro* genome (*α* cluster, *γ1* cluster and *γ2* cluster), in total we identified at least 98 VEs expressed in the diploid outbred *Xtro*, demonstrating the expansion of *Xtro* VEs compared to their mammalian ortholog (14 for *Xtro cPcdh α* cluster, 47 for *Xtro cPcdh γ1* cluster, and 37 for *Xtro cPcdh γ2* cluster). We also characterized the splicing patterns in CE regions.

Interestingly, we identified one novel alternative splicing event occurred in CE region of *cPcdh γ1* cluster, which could generate a novel cPcdh γ1 protein isoform with a shorter cytoplasmic part, which may serve as the evolutionary compensation for the lacking of *cPcdh β* cluster in *Xtro*. Overall, our precise annotation for *Xtro* cPcdh genes would provide a solid foundation for the further functional investigations.

3.2 Materials and Methods

3.2.1 RNA sample preparation

Brains and spinal cords (3 brains and 3 spinal cords were pooled) from *Xtro* in stages 50 and 60 were dissected out and snap frozen in liquid nitrogen. Snap frozen tissues were then lysed in TriZOL by syringing through a 21-gauge needle. Following RNA extraction was processed following manufacturer's protocol (Life Technologies).

3.2.2 full-length 5'RACE sequencing for *Xtro cPcdh* genes

1 μg of *Xtro* totalRNA sample was used per 5'RACE reaction. The 5'RACE reactions were performed with SMARTer® RACE 5'/3' Kit using manufacturer's protocol with the following modifications (TaKaRa Clontech). The RT was performed using mixture of three following gene-specific RT primers.

γ1-RT: 5'-TCGTTCTCATTTTCCAGTTTCTTTCC-3'

γ2-RT: 5'-TACTGTACCATAAGAACTAGAGGCAG-3'

α-RT: 5'- ACATTTGACAGAATAAAGCTTTAAGAC-3'

The PCR was performed with three following gene-specific PCR primers in three separate PCR reactions, with Tm at 54°C and 30 cycles of amplification.

γ1-PCR: 5'-TACNNNNNNACTGCCCTGTTGGTGTCAGCCAATC-3'

γ2-PCR: 5'-TACNNNNNNACCAATTCGCTTGGGGAATTCTTCTGGGG-3'

α-PCR: 5'-TACNNNNNNACGGAAGGTGCATCAACAGTAGGAAGAA-3'

("NNNNNN" representing various barcode sequences for indexing different samples during sequencing; three indexes were used: "ATCACG"; "TTAGGC"; "ACAGTG")

After 5'RACE reaction, the products were purified and eluted into 10 µl of water each using Agencourt AMPure XP system. The purified 5'RACE products were prepared into sequencing libraries using DNA Template Prep Kit 2.0 (3Kb - 10Kb), then sequenced on PacBio RS SMRT sequencing platform with DNA/Polymerase Binding Kit P6, DNA Sequencing Reagent 4.0, and MagBead Standard Seq v2.

3.2.3 PacBio sequencing reads processing and alignment

ROI reads obtained from PacBio sequencing were aligned to *Xtro cPcdh* VEs and CEs annotations from our previous genome analyses (Etlioglu et al., 2016) using BLAST with default setting.

3.2.4 VE annotations using PacBio sequencing reads

PacBio IsoSeq v1 analysis (PacBio) was performed for PacBio long-read sequencing results with the following parameters beyond default: estimated cDNA size: 2~3 kb; minimum quiver accuracy: 0.99. The resulting polished high-quality isoforms were retained and aligned to *Xtro cPcdh* VE annotations based on our previous genome analyses using BLAST with default setting (Etlioglu et al., 2016). The isoforms were manually grouped according to the numbers of mismatches from the alignment (see Appendix Table S3-3 for the manual grouping). Grouped isoforms were then aligned and clustered using ClustalO tool with default setting and parameter "--outfmt=vie". The vie output files were then aligned and parsed to obtain the consensus sequences individually. These consensus sequences are then denoted as "rna-VE", VE annotated from full-length sequencing. These rna-VE were also aligned to 1) *Xtro cPcdh* VEs annotations based our previous genome analysis, and 2) *Xtro* genome sequence of scaffold3 of xenTro 7 (www.xenbase.org) using BLAST with default setting for various purpose.

3.2.5 CE annotations using PacBio sequencing reads

The polished high-quality isoforms from 3.2.4 section were also aligned to *Xtro cPcdh* CEs annotations based on our previous genome analyses using BLAST with default setting (Etlioglu et al., 2016). The isoforms with alignment containing large number of mismatches (>10) or gaps (>10) were retained and further characterized using ClustalO tool and parsed to obtain the consensus sequences as described in 3.2.4 section. The Percentage of Splice-In (PSI) value of the novel γ1 CE isoform was calculated as the ratio between the ROI reads aligned to the novel γ1 CE isoform and the ROI reads aligned to all γ1 CE isoforms.

3.2.6 Re-analysis of the published RNA-seq data for various *Xtro* developmental stages

Published RNA sequencing data were downloaded from Short Read Archive (SRA, SRP012375) and were used to quantify the expression and splicing pattern of *cPcdh* genes during Xenopus development (Tan et al., 2013). With Tophat2, the RNA-seq data were aligned to the reference genome (xenTro3, www.xenbase.org), as well as the transcriptome (Ensembl version 82) with the novel *γ1* CE isoform added. The splicing pattern regarding the novel *γ1* CE isoform was measure by the PSI value derived from the ratio of exon junction reads using this splicing site (splicing in) to the sum of junction reads using either canonical 5' splicing site (splicing out) or novel splicing site. We applied TPM (Tags/unique mapped reads Per Million mapped tags/unique mapped reads) value, which assesses the *Xtro cPcdh* gene expression. By normalizing the exon junction reads linked last two constitutive exon of *cPcdh* to total exon junction reads.

3.3 Results

Even in the latest *Xtro* genome build (xenTro7), the *cPcdh* gene loci are not completely annotated. To identify the genomic regions of *cPcdh* genes in *Xtro* genome, we firstly performed TBLASTN searching for mouse *cPcdh α*, *β* and *γ* genes in xenTro7 (Etlioglu et al., 2016). Two unlinked *cPcdh* genomic loci were identified

in *Xtro* genome, both flanked by non-*cPcdh* genes (Figure 3-2). In the first locus, one *cPcdh α* gene cluster (named as *cPcdh α* cluster) and one *cPcdh γ* cluster (named as *cPcdh γ1* cluster) were identified. In the second locus, one *cPcdh γ* cluster (named as *cPcdh γ2* cluster) was identified. Interestingly, there's no *cPcdh β* cluster identified in *Xtro* genome. By further genomic analyses, these *Xtro cPcdh* genes are annotated in more details. First, all three *Xtro cPcdh* genes shared the classical structures as their mammalian orthologs: in their RNA transcripts, a first alternative VE is splicing together with CEs, which are identical for all RNA isoforms from the same cluster. Second, similar to their mammalian orthologs, three CEs could be identified for all the three *Xtro cPcdh* clusters. Third, different numbers of VEs were identified for different clusters. 14, 46 and 36 VEs are identified for *cPcdh α*, *γ1*, and *γ2* clusters, respectively (Figure 3-2).

This genome analysis is valuable for annotating the *Xtro cPcdh* genes. However, such approach has several disadvantages: 1) it could not guarantee that all annotated exons are authentic exons that do express (*i.e.* false positives); 2) it could not guarantee that all authentic exons have been annotated (*i.e.* false negatives). To partially overcome these disadvantages, and further annotate and refine the *Xtro cPcdh* gene structures in a more comprehensive way, approaches based on RNA transcript profiling would be suitable. As their mammalian orthologs, the diversities of *Xtro cPcdh* genes are mainly generated from alternative usage of first exon (alternative VEs) at the 5' end of the transcripts, linked to the three constant exons (CEs) at the 3' end of the transcripts. Such structures make 5'RACE combined with full-length sequencing analysis a suitable approach to further annotate the gene structures, and characterize the expression and splicing patterns for *cPcdh* RNA transcripts. Recently, the rapid development on the long-read sequencing technologies have enabled full-length sequencing for long DNA molecules in high-throughput manner. In this study, we applied PacBio long-read full-length sequencing analysis for the 5'RACE products to annotate the *Xtro* Pcdh α, γ1, and γ2 genes.

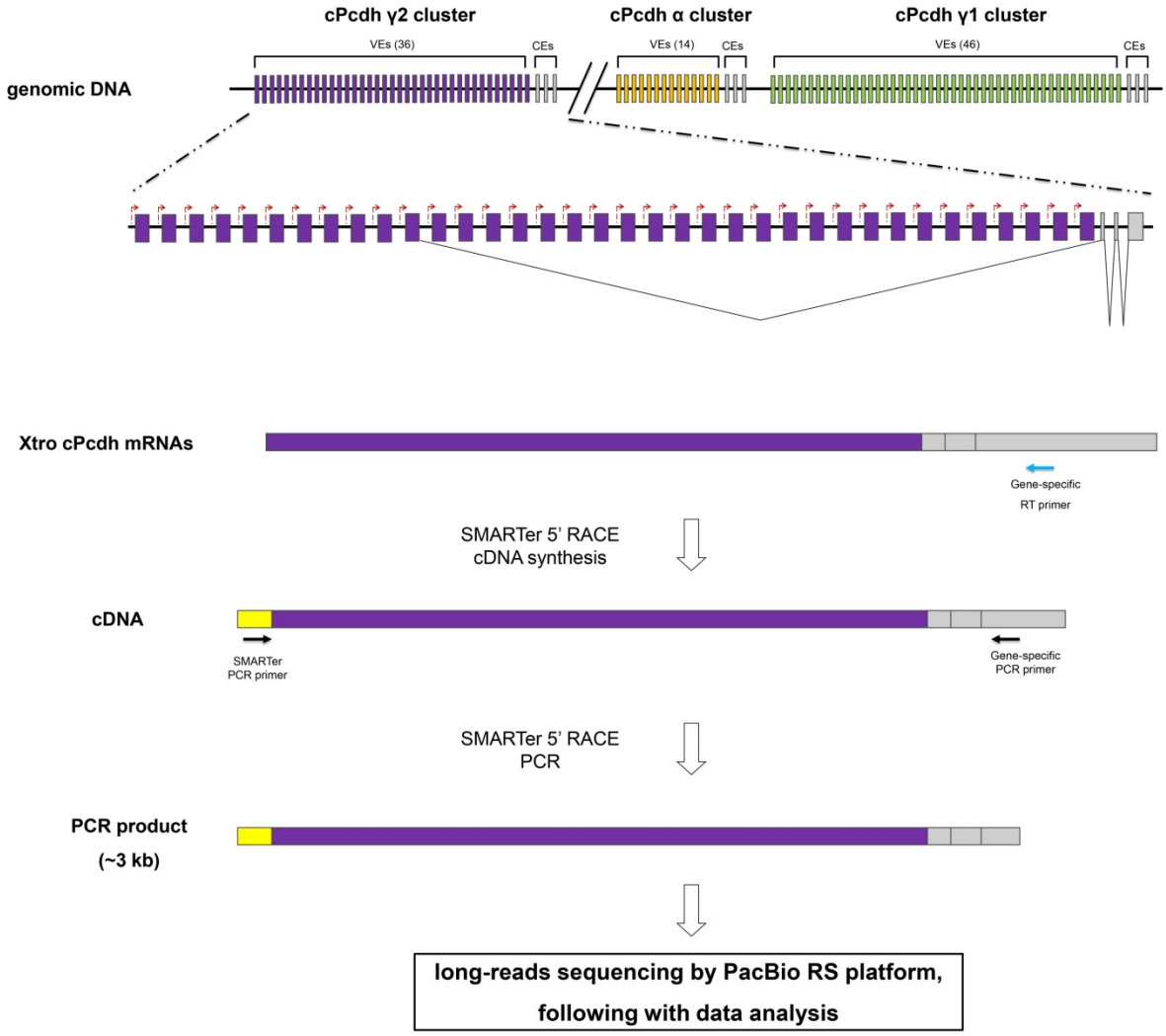


Figure 3-2. Experiment design of full-length 5'RACE sequencing for *Xtro cPcdh* genes. Based on our previous genome analyses of *Xtro cPcdh* genes, three *cPcdh* gene clusters, *α*, *γ1* and *γ2*, are identified in *Xtro* genome (Etlioglu et al., 2016). RNAs from all three clusters are spliced as one VE linked with three following CEs. Based on this structure feature, gene-specific SMARTer 5'RACE RT primers and PCR primers are designed on the CE3 of *Xtro cPcdh* mRNAs. After SMARTer 5'RACE, the products (~3 kb) are preceded with PacBio long-read sequencing.

3.3.1 Establishment of full-length 5'RACE sequencing for *Xtro cPcdh* genes

The experimental design for our strategy is demonstrated in Figure 3-2. In general, for

Xtro Pcdh α, γ1, and γ2 genes, all RNA regions from the 5' end to the middle part of constant exon 3 (CE3) could be amplified by the 5'RACE using gene-specific primers targeting CE3. Such 5'RACE were performed for total RNA samples from biological triplicates of mixed *Xtro* neuronal tissues (brains and spinal cords) in stage 50 (metamorphosis) and stage 60 (froglets). These 5'RACE products were then sequenced with PacBio long-read instrument.

In total, we obtained 45,108, 36,105, and 49,932 putative full-length 5'RACE reads for *Xtro* Pcdh α, γ1, and γ2 genes, respectively. These full-length reads contain both gene–specific RACE primer sequences and the 5' specific RACE oligo sequences (Appendix Table S3-1). To check the specificities of these 5'RACE products, we first mapped these reads onto the CE regions of these three clusters (methods). After mapping, >99% of reads contains the CE regions of corresponding *cPcdh* gene, indicating the high specificities of our 5' RACE experiments (Appendix Table S3-1).

3.3.2 Annotation for *Xtro cPcdh* VEs

First, we investigated the false positive rate of the VE annotations from genome analyses, *i.e.* how many of these VEs do express. After mapping these full-length 5'RACE reads to current VE annotations, it showed that all annotated VE do express (Appendix Table S3-2). However, we noticed that a portion of full-length 5'RACE reads either could not be mapped to current VE annotations, or mapped with very large numbers of mismatches, indicating there still exist novel VEs beyond current annotations.

The highly gene-specific full-length 5'RACE reads made it possible using them to build precise RNA transcript annotations for these three *Xtro cPcdh* genes. Considering the gene structures of *Xtro cPcdh* clusters, we first aim to further annotate and refine the VEs. Considering the features of PacBio long-read sequencing (long read length, but relative high sequencing error rate (~11%)), in order to annotate the VEs using these full-length 5'RACE reads, we developed a customized analysis pipeline by integrating various sequence clustering tools with current *Xtro cPcdh*

annotation (methods, Appendix Table S3-3). After applying this pipeline, in total, we first identified 24, 99 and 66 VEs for *Xtro cPcdh α*, *γ1*, and *γ2* clusters, respectively. Hereafter we will denote these VEs as rna-VEs. The rna-VE numbers are much more than the VE numbers identified by genome analyses (14, 46, and 36 VEs for *Xtro cPcdh α*, *γ1*, and *γ2* clusters). Since our RNA samples were generated from outbred frog strains, and these rna-VEs were annotated according to their sequence differences, we suspected that many rna-VEs might be the over-annotations due to the allelic variations from parental differences. To investigate such possibility, we compared the sequence differences between the rna-VEs and the VEs annotated from genome analyses. Indeed, most of the rna-VEs (24 of 24 *α* rna-VEs; 94 of 99 *γ1* rna-VEs; 66 of 66 *γ2* rna-VEs) could be aligned on VEs annotated from genome analyses. Among these rna-VEs that could be aligned, most of them have very similar sequences comparing with VEs annotated from genome analyses (Figure 3-3A). All these indicate that many rna-VEs were actually over-annotated due to the allelic variations from outbred frog strains. However, interestingly, there are still several rna-VEs either not mappable to current VE annotations (five rna-VEs) or mapped with large numbers of mismatches (10 rna-VEs containing >100 mismatches comparing with current VE annotations), indicating there still exist novel VEs beyond annotations from genome analyses. After aligning these rna-VE sequences to the current *Xtro* genome, indeed, we identified and refined two more novel *cPcdh* VEs beyond current annotations in the *Xtro* genome (one in γ1 cluster, and one in γ2 cluster), together with our previous VE annotation from genome analyses we could annotate at least 98 VEs in *Xtro* genome (14 for α cluster, 47 for γ1 cluster, 37 for γ2 cluster) (Figure 3-3B, Appendix Table S3-4). However, even after the genome alignment, we have noted that there are still 9 rna-VEs containing considerable too large sequence variations comparing with current VE annotations and genome reference (mismatch number > 100, Figure 3-3A). Thus we couldn't rule out the possibility that these rna-VEs are indeed the authentic VEs in the genome, due to the relatively poor quality of current *Xtro* genome reference.

Moreover, there are several interesting observations from full-length 5'RACE sequencing data. When analyzing the transcription start sites (TSSs) of *Xtro cPcdh* genes using full-length 5'RACE sequencing data, we noted TSSs of *cPcdh α-14* VE spread in a wide genomic range (~ 2 kb), while the TSSs of all other *cPcdh* VEs distribute in a narrow region (within 200bp) (Figure 3-3C), indicating that *cPcdh α-14* VE may use a totally different transcriptional regulatory mechanism comparing with that of other VEs'. We observed a small fraction of reads (~1.2%) supporting the RNA structure of *cPcdh α* VE splicing together with *cPcdh γ1* CEs. Considering the genomic structure of *Xtro cPcdh* genes, it indicates the long-distance splicing for the extremely long precursor mRNAs transcribing from TSS of *cPcdh α* VEs till the end of *cPcdh γ1* CE genomic regions. And this phenomenon could also be observed in the mammalian *cPcdh* clusters (Wu and Maniatis, 1999).

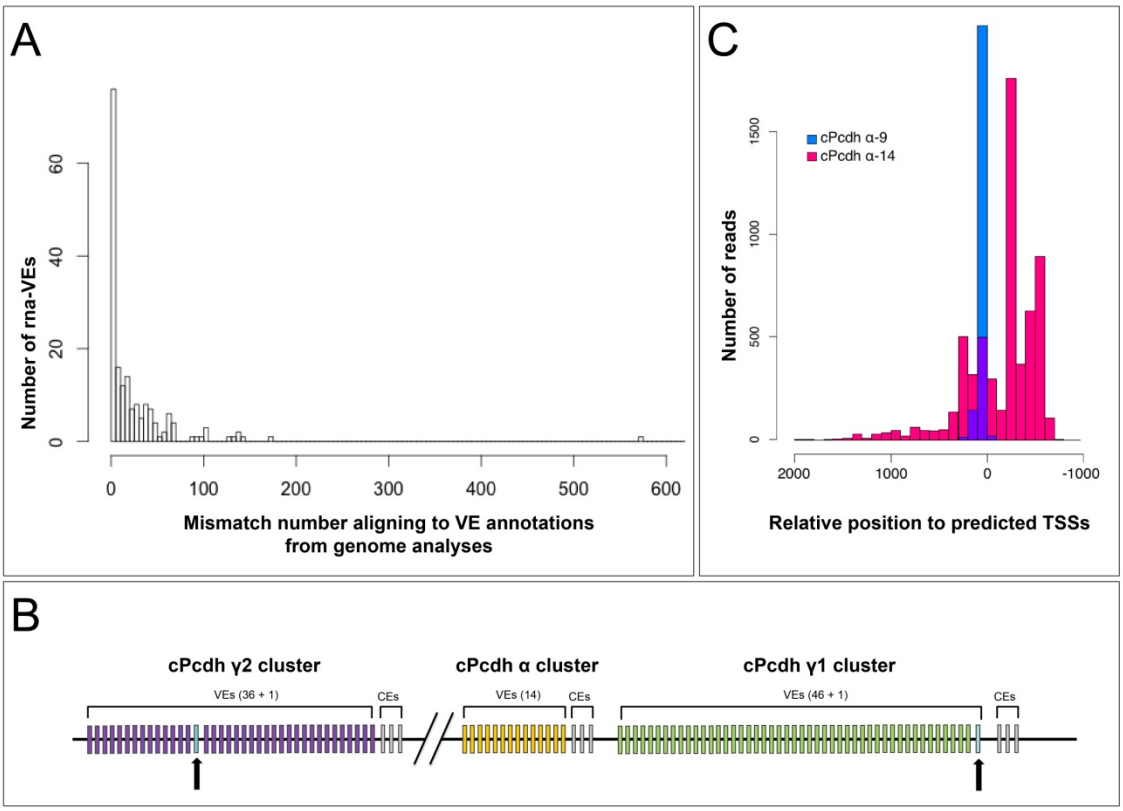


Figure 3-3. *Xtro cPcdh* VE annotation using full-length 5'RACE sequencing. (**A**) This histogram demonstrates the similarity between rna-VEs and VE annotations from genome analyses. X-axis shows the mismatch numbers for rna-VEs when aligning to

the VE annotations from genome analyses. Y-axis shows the numbers of rna-VEs with different mismatch numbers. (**B**) The improvement of the annotation of *Xtro cPcdh* genome loci from full-length 5'RACE sequencing. The arrows indicate the novel VEs annotated from full-length 5'RACE sequencing. (**C**) The broad spreading of TSSs for *Xtro cPcdh α14* transcripts. X-axis represents the relative positions of the 5' ends of full-length reads to the predicted TSS positions from genome analyses (bp). Y-axis represents the numbers of reads aligned to corresponding VEs. Blue and Red bars represent the results of *cPcdh α-9* and *cPcdh α-14* isoforms, respectively. TSSs from all other *cPcdh* isoforms demonstrated the similar distribution as those from *cPcdh α-9* (data not shown).

3.3.3 Annotation for *Xtro cPcdh* CEs

Our customized analysis pipeline not only allows annotating the VE regions, but also enables to evaluate and further refine our annotations for the CEs of *Xtro cPcdh* genes. We first investigated the quality of the CE annotations from genome analyses. All the CEs and splicing structures annotated from genome analyses could also be identified in our analysis from full-length 5'RACE sequencing, demonstrating the high quality of our annotation of *Xtro cPcdh* genes based on genome analyses (Appendix Table 3-1). Interestingly, we identified one novel splicing event in the CE region of *Xtro cPcdh γ1* RNA transcripts (Figure 3-4, Appendix Table S3-4). In this splicing event, compared to the canonical *γ1* CE annotation, we identified one novel alternative 5' splicing site in the intron between *γ1* CE2 and CE3, resulting in a novel CE isoform with 16nt extension at the 3' part of *γ1* CE2 (Figure 3-4A). This 16nt insertion located in the canonical *γ1* CE coding region and it would introduce a premature stop codon, thus resulting into a novel cPcdh γ1 protein isoform with shortened cytoplasmic part (80 amino acids shorter) (Figure 3-4A). After comparing the sequences at the intronic regions close to 3' ends of *cPcdh γ* CE2s between *Xtro*, mouse and human, we find this novel splicing site is not conserved in vertebrates (Figure 3-4B). Therefore we

suspect that there won't exist similar alternative splicing for mouse and human *cPcdh γ* CEs (Figure 3-4B). Indeed, re-analyzing public RNA-seq data confirmed our hypothesis on human and mouse *cPcdh γ* genes (data not shown). Considering no *cPcdh β* cluster existing in *Xtro* genome (mammalian *cPcdh β* genes also coding cPcdh proteins with short cytoplasmic part) (Wu and Maniatis, 1999), proteins from this novel *cPcdh γ1* CE isoform may serve as the *Xtro* functional counterpart for mammalian cPcdh β proteins (Figure 3-4A).

We analyzed the expression pattern of this novel *cPcdh γ1* CE isoform. The Percentage of Splicing-In (PSI) values of this novel splicing event in stage 50 and 60 were 22.8% and 32.5%, respectively (Figure 3-4C). Such high PSI values indicated this novel *cPcdh γ1* CE isoform are indeed expressed at a reasonable high level.

To further investigate the expression patterns of this novel *cPcdh γ1* CE isoform and the expression patterns of all three *Xtro* cPcdh genes during development, we reanalyzed the published *Xtro* transcriptome profiling data for 23 distinct developmental stages (2-cell stage to stage 44-45) using Illumina short-read high-throughput RNA sequencing (Tan et al., 2013). As demonstrated in Figure 3-5, all the three *cPcdh* gene clusters increase expression after stage 13-14, which is the end of gastrulation and the beginning of the neurula stage, consistent with the function of cPcdh proteins in neuronal development. Interestingly, *cPcdh α* genes demonstrated another peak of expression wave at the early developmental stage (stage 8 to stage 13-14), indicating cPcdh α may have additional function in the early *Xtro* development (Figure 3-5). More importantly, we could also identify the novel *cPcdh γ1* CE isoform from this public RNA-seq dataset, and observe its inclusion (PSI value: 6% to 15%) since the beginning of *cPcdh γ1* cluster expression (Figure 3-4D).

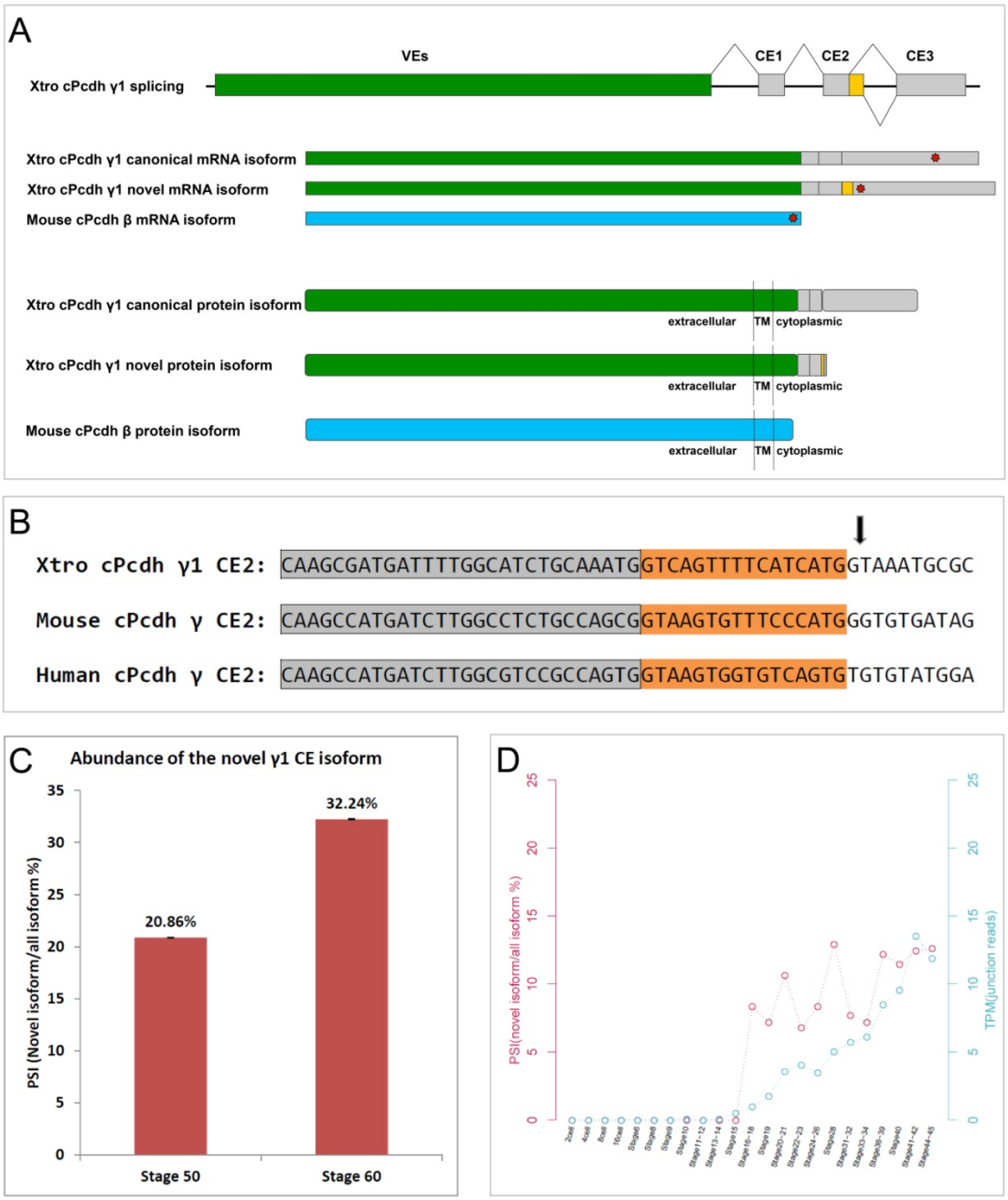


Figure 3-4. A novel *Xtro cPcdh γ1* CE isoform identified from full-length 5'RACE sequencing. (**A**) A novel *Xtro cPcdh γ1* isoform could be generated by RNA alternative splicing at the CE2, which results into a 16-nt CE2 3' extension (yellow part). It will introduce the frame-shift and a pre-mature stop codon (red stars), which will produce *Xtro* cPcdh γ1 proteins with shorter cytoplasmic part. Blue bars demonstrate the mouse *cPcdh β* mRNAs and proteins. (**B**) The sequence comparisons among the corresponding regions in *Xtro*, mouse, and human indicate this novel *Xtro cPcdh γ1* alternative splicing event is *Xtro* species-specific. The arrow indicates the

splicing site is specific for *Xtro*. (**C**) The expressing abundances of the novel *Xtro cPcdh γ1* isoform. Percentage of Splice-In (PSI) values were calculated from our full-length 5'RACE sequencing for *Xtro* brain samples at stage 50 and stage 60. (**D**) The expressing dynamics of Xtro *cPcdh γ1* genes and the novel *γ1* CE isoform during *Xtro* development based on re-analysis of published RNA-seq data. Blue dots indicate the expression level of *Xtro cPcdh γ1* genes at different developmental stages. Red dots indicate the PSI values for the novel *Xtro cPcdh γ1* CE isoform at different developmental stages.

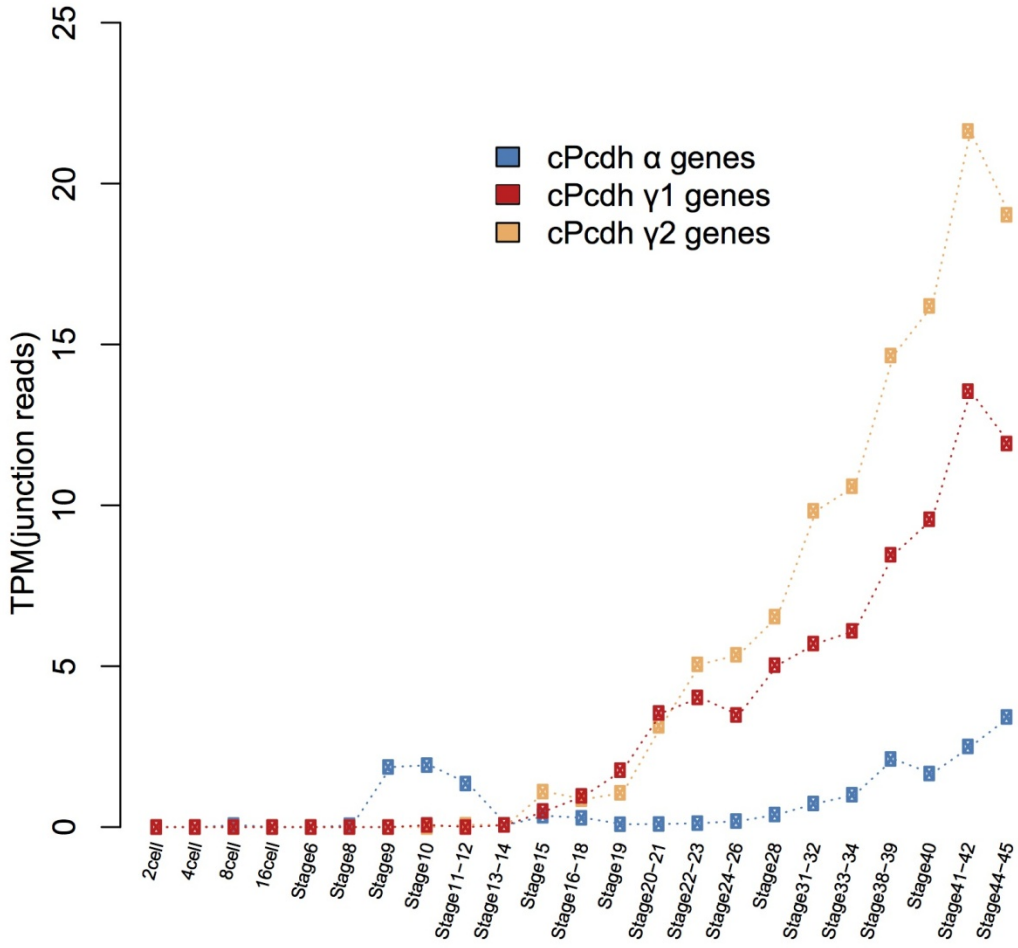


Figure 3-5. The expressing dynamics of *Xtro cPcdh* genes during development. Blue, red, and orange dots indicate the expression levels of *Xtro cPcdh α*, *γ1*, and *γ2* genes at different developmental stages, respectively.

3.4 Discussion

3.4.1 Novelty and Advantages of full-length 5'RACE sequencing-based annotation

High-throughput sequencing technologies have been revolutionary for genomic and transcriptomic studies (Wang et al., 2009). One disadvantage of the 2nd generation high-throughput sequencing technologies is the short read length (typically 2x100 bp with pair-end sequencing), which couldn't directly span the full-length RNA transcripts, thus making it difficult to precisely assemble and characterize the landscape of the RNA isoforms due to the diversity and complexity generated from alternative splicing (Steijger et al., 2013). However, recent development of the long-read high-throughput sequencing technologies offered the opportunity to solve this problem. Their long reads could be used to directly produce full-length RNA transcript sequences without assembly, thus providing the possibility for precise transcript annotation (Branton et al., 2008; Eid et al., 2009). Such full-length RNA sequencing has been successfully applied on the annotations and characterizations for both whole transcriptome (Au et al., 2013; Sharon et al., 2013; Tilgner et al., 2014) and isoform diversity for complex genes, such as mammalian *neurexins* (Schreiner et al., 2014; Treutlein et al., 2014) and *Dorsophila Dscam* genes (Armitage et al., 2014; Bolisetty et al., 2015; Sun et al., 2013).

Genome analyses such as conservation analysis have been a powerful tool for annotating conserved genes in a new species. Yet, it has several disadvantages. First, it is only suitable for annotating the conserved regions between the previous annotated species and the new one. Novel evolutions of the target gene in the new species won't be able to be characterized by the genome analyses. Thus it will introduce potential false negative into the annotation. Second, it could only provide the putative annotation for the target gene, *i.e.* it could not guarantee that the annotated RNA isoforms indeed express. Thus it will introduce potential false positive into the annotation. These problems would be more severe for those genes with rapid

evolution and complex gene structure, such as *cPcdh* genes. Full-length RNA sequencing has become a powerful tool to identify and characterize complex splicing isoforms for the genes, such as *neurexin* genes (Schreiner et al., 2014; Treutlein et al., 2014) and would also be a suitable complement for the annotation based on genome analyses.

Here, we applied 5'RACE together with PacBio long-read full-length sequencing to further characterize and refine our previous annotation of the *Xtro cPcdh* genes loci generated from genomic analyses. By applying our customized full-length 5'RACE sequencing analysis, we successfully evaluated the false positive and false negative rates of our previous *Xtro cPcdh* annotation. From our data, we observed the expressions of all isoforms and exons annotated from genome analyses, indicating the 'zero' false positive rate of the annotation from genome analyses. However, we observed several novel exons and splicing events not presenting in the annotations from genome analyses, indicating certain false negative in the genome analyzing annotation. Together with several other observations, such as the long distance spliced isoforms between *Xtro cPcdh α* VEs and *cPcdh γ* CEs, as well as the spreading TSS distribution of *Xtro cPcdh α-14* isoforms, it demonstrated that the full-length RNA (5'RACE) sequencing is indeed a suitable and powerful approach for gene isoform annotation.

3.4.2 Genome duplication and expansion of *Xtro cPcdh* gene loci

In our genome annotation of *Xtro cPcdh* gene loci, we first identified two unlinked cPcdh cluster loci in *Xtro* genome, one containing one *α* cluster and one *γ* cluster (*γ1*), the other containing one *γ* cluster (*γ2*) (Etlioglu et al., 2016). But there's no *β* cluster in *Xtro* genome. In comparison, all mammalian *cPcdh* genes resident in one locus, consist of one *α*, *β*, and *γ* clusters, indicating the genome duplication of the *Xtro cPcdh* loci. Actually, this is not the unique phenomenon for *Xtro*. The genome duplications of the *cPcdh* loci have also been observed in zebrafish and fugu genomes (Noonan et al., 2004b; Tada et al., 2004; Yu et al., 2007), indicating that *cPcdh* loci

are rapidly evolving during vertebrate evolution.

Moreover, we identified at least 98 *cPcdh* VEs in *Xtro* genome (14 for *α* cluster, 47 for *γ1* cluster, 37 for *γ2* cluster) (Figure 3-3). This is much more than the VE numbers identified in mammalian genomes (52 VEs in human genome; 58 VEs in mouse genome; 59 VEs in rat genome)(Wu et al., 2001; Zou et al., 2007), indicating not only the duplication, but also the expansion of *cPcdh* loci in *Xtro* genome. Interestingly, analyses for the recently sequenced octopus (*Octopus bimaculoides*) genome not only identified *Pcdh* genes, but also demonstrated that its *Pcdh* loci exhibiting extensive duplications and expansions (168 *Pcdh* genes were identified in octopus genome), indicating that *cPcdh* may not be just an innovation for vertebrates, but also for other *Metazoan subphyla* (Albertin et al., 2015). These evidences further confirmed the rapid evolution of the *cPcdh* genes.

3.4.3 Potential functional importance of the novel *cPcdh γ1* CE isoform

In mammals, the structural of *cPcdh β* gene is special, compared with *cPcdh α* and *γ* gene. Mature *cPcdh β* RNA transcripts don't contain CEs in the 3' end, thus producing cPcdh proteins with shorter cytoplasmic part (Figure 3-4A). Such special structure of *cPcdh β* genes leaded to the speculation that cPcdh proteins with short cytoplasmic part may play certain unique functional roles in neuronal development. However, in *Xtro* genome, the absence of the *cPcdh β* cluster questioned the functional importance of cPcdh proteins with short cytoplasmic part. Interestingly, in the full-length 5'RACE sequencing based annotation for *Xtro cPcdh* genes, we identified one novel splicing event in the CE regions of *cPcdh γ1* clusters. This event will result in a novel *cPcdh γ1* CE isoform with a premature stop codon, coding for the cPcdh γ1 proteins possessing a shortened cytoplasmic part (Figure 3-4A). And the RNA transcripts of such short cytoplasmic cPcdh γ1 proteins do express in reasonable high levels during the *Xtro* neuronal development. Previously, it has been proposed that *Pcdh* may have been evolved in different species in a lineage-specific fashion (Albertin et al., 2015; Yu et al., 2008). Thus based on our observations here, we

speculated that this novel *cPcdh γ1* CE isoform might potentially provide the function for *Xtro* similar as *cPcdh β* genes for mammals, representing one possible convergent evolution for generating cPcdh proteins with short cytoplasmic part.

3.4.4 Summary and perspective

In summary, here based on our latest *Xtro cPcdh* gene loci annotation from genome analyses, we further annotated and refined the *Xtro cPcdh* gene loci in details by applying long-read full-length sequencing of the 5'RACE products from *Xtro cPcdh* mRNAs. Our detailed annotation would provide a solid foundation for the further functional investigations of cPcdh genes in *Xtro* frog as a model organism.

There are also several open questions derived from this study. First, although predicted as constitutive, whether the part of *cPcdh* RNA transcripts downstream of our primer targeting regions on CE3s contain any alternative regions still needs to be experimentally examined. To do so, full-length 3'RACE sequencing for *cPcdh* RNA transcripts would be suitable. Second, for VE annotations, regarding the large number of rna-VEs we identified, more efforts should be devoted to precisely define they are either the authentic VEs presenting in the *Xtro* genome, or truly the over-annotations due to allelic variations. Third, as demonstrated in the second part of this dissertation, for genes functioning in neuronal self-avoidance, the absolute quantification for the expression of different isoforms would be essential for understanding their functional importance. However, due to the possible different template-switching efficiencies for different *cPcdh* RNA isoforms in the RACE reactions, full-length 5'RACE sequencing could not provide the quantitative information for *cPcdh* RNA isoform expression (data not shown). Thus, it would be necessary to develop novel approaches that enable absolutely quantifying the expression of *Xtro cPcdh* isoforms, as CAMSeq for *Drosophila Dscam* isoforms expression. Finally, the novel *cPcdh γ1* CE isoform we identified in this study provides an interesting scenario regarding the potential convergent evolution of Pcdh protein isoforms with short cytoplasmic part. To reveal its role in neuronal development, more functional and mechanistic investigation

would be necessary in the future.

Chapter 4. Discussion

4.1 Evolution of Neuronal Self-avoidance

The studies in *Drosophila Dscam* gene and vertebrate *cPcdh* genes have provided valuable insights into the molecular mechanisms of neuronal self-avoidance. Furthermore, these studies have also much deepened our understanding regarding the evolution of neuronal self-avoidance.

Although *Dscam* orthologs are present throughout *deuterostomes*, the extensive isoform diversities of *Dscam* could only be observed in insects and crustaceans. Thus, the function of Dscam in neuronal self-avoidance seems to be a *Pancrustacea*-specific evolutionary innovation generated from the diversification of *Dscam* isoforms (Armitage et al., 2012). Mammalian *Dscam* ortholog genes have no extensive diversity. Recent studies have revealed that the *cPcdh* genes play critical roles in mammalian neuronal self-avoidance, indicating the mammalian seem to have adapted and evolved an independent machinery to achieve neuronal self-avoidance similar as *Dscam* gene for insects. Considering the emerging and similarities of *clustered Protocadherin (cPcdh)* genes in vertebrates, *cPcdh* was speculated as a vertebrate innovation from duplication and expansion of non-clustered *Pcdh* ancestors. Recently, analysis of the annelids, molluscs, squid, and octopus genomes has revealed that there also exist other types of *cPcdh* genes in their genomes, which are evolved and expanded from distinct non-clustered *Pcdh* ancestors (Albertin et al., 2015). Together with previous study on *Pcdh* genes in elephant shark, it seems that *Pcdh* genes have been the choice for regulating neuronal self-avoidance in many species (Yu et al., 2008). The *Pcdh* genes in these species have undergone separate, lineage-specific expansions to fulfill the molecular diversity required for their neuronal self-non-self discriminations. Interestingly, *Pcdh* orthologs could not be identified in *Drosophila* genome. The lost of *Pcdh* ancestor gene may be the reason for *Drosophila* to choose *Dscam* as neuronal self-avoidance gene. Combining all these evidence together, we

speculate that, the mechanisms of neuronal self-non-self discrimination may be evolutionary innovations appeared after the divergences of the *bilateria*, and may have co-evolved with the neuronal systems in lineage-specific manners.

4.2 Importance of RNA Isoforms Profiling in Neuronal Studies

Our work has illustrated the importance of absolute quantification of *Drosophila Dscam* RNA isoform expression on understanding the molecular mechanism of neuronal self-avoidance. The same principle should also apply for the study of *cPcdh* genes in vertebrate neuronal self-avoidance. We have also demonstrated the unique advantages of full-length 5'RACE sequencing on annotating and investigating the splicing patterns of *Xtro cPcdh* genes. In general, our researches emphasized the power and importance of RNA isoform analyses in deciphering the mechanisms of neuronal self-avoidance.

From broader perspective, identifying and quantifying the isoform diversities generated from RNA alternative splicing is not only important for investigating the neuronal self-avoidance, but also critical for studying other processes involved in neuronal development and function. Evolutionary analysis for RNA alternative splicing across vertebrate species have demonstrated that, among all tissues and organs, the neuronal tissues not only possess the most abundant alternative splicing, but also possess the most conserved alternative splicing (Barbosa-Morais et al., 2012; Merkin et al., 2012). Such feature suggests that the alternative splicing in neuronal tissues may play rather important roles for neuronal system (Raj and Blencowe, 2015). One example is the *neurexin* genes. The alternative splicing of *neurexin* genes has been demonstrated to be important for the target selection of neuronal wiring (Südhof, 2008). Neurexin proteins are presynaptic cell-adhesion molecules. Through RNA alternative splicing, *neurexin* genes could potentially generate thousands of different isoforms(Schreiner et al., 2014; Treutlein et al., 2014). Different neurexin isoforms with different exons included or excluded will produce protein isoforms exhibiting different binding affinities for different cell surface receptors, thus differ in organizing

different postsynaptic compartment and sharping the postsynaptic connections. Alternative splicing occurred in many other genes, such as alternative splicing in *Disable-1* (Yano et al., 2010) and *Unc13b* (Quesnel-Vallières et al., 2015), would also regulate the neuronal development from many perspectives. Considering the complicated alternative splicing events and consequent protein functional diversities in neuronal system, it has been proposed that the complicated and coordinated expression of different isoforms from various proteins may be of general importance for the correct formation of the neuronal connectivity map and the proper functioning of the neuronal system. Following this hypothesis, two tasks would be essential in the future: 1) discovering all the potential RNA isoforms generated from alternative splicing; 2) accurately quantifying the expressions of different isoforms in the temporal-spatial specific manner, even in the single-cell or subcellular resolution. We could envision that, such information would be of great help for precisely dissecting the processes of the neuronal development, and for understanding the mechanisms of the neuronal wiring and functioning.

Bibliography

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., and Rokhsar, D.S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature *524*, 220–224.

Anastassiou, D., Liu, H., and Varadan, V. (2006). Variable window binding for mutually exclusive alternative splicing. Genome Biol. *7*, R2.

Armitage, S. a O., Freiburg, R.Y., Kurtz, J., and Bravo, I.G. (2012). The evolution of Dscam genes across the arthropods. BMC Evol. Biol. *12*, 53.

Armitage, S. a O., Sun, W., You, X., Kurtz, J., Schmucker, D., and Chen, W. (2014). Quantitative profiling of Drosophila melanogaster Dscam1 isoforms reveals no changes in splicing after bacterial exposure. PLoS One *9*, e108660.

Au, K.F., Sebastiano, V., Afshar, P.T., Durruthy, J.D., Lee, L., Williams, B. a, van Bakel, H., Schadt, E.E., Reijo-Pera, R. a, Underwood, J.G., et al. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. U. S. A. *110*, E4821–E4830.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science (80-.). *338*, 1587–1593.

Bolisetty, M.T., Rajadinakaran, G., and Graveley, B.R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol. *16*, 204.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S. a, Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2008). The potential and

challenges of nanopore sequencing. Nat. Biotechnol. *26*, 1146–1153.

Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell *65*, 175–187.

Celotto, A.M., and Graveley, B.R. (2001). Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. Genetics *159*, 599–608.

Chen, W. V, and Maniatis, T. (2013). Clustered protocadherins. Development *140*, 3297–3302.

Chen, B.E., Kondo, M., Garnier, A., Watson, F.L., Püettmann-Holgado, R., Lamar, D.R., and Schmucker, D. (2006). The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila. Cell *125*, 607–620.

Chen, W. V, Alvarez, F.J., Lefebvre, J.L., Friedman, B., Nwakeze, C., Geiman, E., Smith, C., Thu, C.A., Tapia, J.C., Tasic, B., et al. (2012). Functional significance of isoform diversification in the protocadherin gamma gene cluster. Neuron *75*, 402–409.

Cheng, H.J., Nakamoto, M., Bergemann, a D., and Flanagan, J.G. (1995). Complementary gradients in expression and binding of ELF-1 and Mek4 in development of the topographic retinotectal projection map. Cell *82*, 371–381.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*, 613–619.

Dickson, B.J. (2002). Molecular mechanisms of axon guidance. Science (80-.). *298*, 1959–1964.

Dong, Y., Taylor, H.E., and Dimopoulos, G. (2006). AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. PLoS Biol. *4*, e229.

Dong, Y., Cirimotich, C.M., Pike, A., Chandra, R., and Dimopoulos, G. (2012). Anopheles NF-κB-regulated splicing factors direct pathogen-specific repertoires of the hypervariable pattern recognition receptor AgDscam. Cell Host Microbe *12*, 521–530.

Drescher, U., Kremoser, C., Handwerker, C., Löschinger, J., Noda, M., and Bonhoeffer, F. (1995). In vitro guidance of retinal ganglion cell axons by RAGS, a 25 kDa tectal protein related to ligands for Eph receptor tyrosine kinases. Cell *82*, 359–370.

Dreyer, W.J. (1998). The area code hypothesis revisited: olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos. Proc. Natl. Acad. Sci. USA *95*, 9072–9077.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science (80-.). *323*, 133–138.

Esumi, S., Kakazu, N., Taguchi, Y., Hirayama, T., Sasaki, A., Hirabayashi, T., Koide, T., Kitsukawa, T., Hamada, S., and Yagi, T. (2005). Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. Nat. Genet. *37*, 171–176.

Etlioglu, H.E., Sun, W., Chen, W., and Schmucker, D. (2016). Identification, characterization and comparative analysis of the Xenopus tropicalis clustered Protocadherin loci. Manuscr. Prep.

Forbes, E.M., Hunt, J.J., and Goodhill, G.J. (2011). The Combinatorics of Neurite Self-Avoidance. Neural Comput. *23*, 2746–2769.

Garrett, A.M., and Weiner, J.A. (2009). Control of CNS synapse development by γ-protocadherin-mediated astrocyte-neuron contact. J. Neurosci. *29*, 11723–11731.

Garrett, A.M., Schreiner, D., Lobas, M. a., and Weiner, J. a. (2012). γ-Protocadherins

Control Cortical Dendrite Arborization by Regulating the Activity of a FAK/PKC/MARCKS Signaling Pathway. Neuron *74*.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science (80-.). *330*, 1775–1787.

Graveley, B.R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell *123*, 65–73.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

Harland, R.M., and Grainger, R.M. (2011). Xenopus research: metamorphosed by genetics and genomics. Trends Genet. *27*, 507–515.

Hasegawa, S., Hamada, S., Kumode, Y., Esumi, S., Katori, S., Fukuda, E., Uchiyama, Y., Hirabayashi, T., Mombaerts, P., and Yagi, T. (2008). The protocadherin-alpha family is involved in axonal coalescence of olfactory sensory neurons into glomeruli of the olfactory bulb in mouse. Mol Cell Neurosci *38*, 66–79.

Hasegawa, S., Hirabayashi, T., Kondo, T., Inoue, K., Esumi, S., Okayama, A., Hamada, S., and Yagi, T. (2012). Constitutively expressed Protocadherin-α regulates the coalescence and elimination of homotypic olfactory axons through its cytoplasmic region. Front. Mol. Neurosci. *5*, 97.

Hattori, D., Demir, E., Kim, H.W., Viragh, E., Zipursky, S.L., and Dickson, B.J. (2007). Dscam diversity is essential for neuronal wiring and self-recognition. Nature *449*, 223–227.

Hattori, D., Chen, Y., Matthews, B.J., Salwinski, L., Sabatti, C., Grueber, W.B., and Zipursky, S.L. (2009). Robust discrimination between self and non-self neurites

requires thousands of Dscam1 isoforms. Nature *461*, 644–648.

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N.H., Shu, S., Taher, L., et al. (2010). The genome of the Western clawed frog Xenopus tropicalis. Science (80-.). *328*, 633–636.

Hirayama, T., and Yagi, T. (2013). Clustered protocadherins and neuronal diversity.

Hughes, M.E., Bortnick, R., Tsubouchi, A., Bäumer, P., Kondo, M., Uemura, T., and Schmucker, D. (2007). Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron *54*, 417–427.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Jiang, X.-J., Li, S., Ravi, V., Venkatesh, B., and Yu, W.-P. (2009). Identification and comparative analysis of the protocadherin cluster in a reptile, the green anole lizard. PLoS One *4*, e7614.

Kaneko, R., Kato, H., Kawamura, Y., Esumi, S., Hirayama, T., Hirabayashi, T., and Yagi, T. (2006). Allelic gene regulation of Pcdh-α and Pcdh-γ clusters involving both monoallelic and biallelic expression in single Purkinje cells. J. Biol. Chem. *281*, 30551–30560.

Katori, S., Hamada, S., Noguchi, Y., Fukuda, E., Yamamoto, T., Yamamoto, H., Hasegawa, S., and Yagi, T. (2009). Protocadherin-alpha family is required for serotonergic projections to appropriately innervate target brain areas. J. Neurosci. *29*, 9137–9147.

Kohmura, N., Senzaki, K., Hamada, S., Kai, N., Yasuda, R., Watanabe, M., Ishii, H., Yasuda, M., Mishina, M., and Yagi, T. (1998). Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. Neuron *20*, 1137–1151.

Kreahling, J., and Graveley, B. (2005). The iStem, a long-range RNA secondary structure element required for efficient exon inclusion in the Drosophila Dscam pre-mRNA. Mol. Cell. Biol. *25*, 10251–10260.

Langley, J.N. (1895). Note on Regeneration of Prae-Ganglionic Fibres of the Sympathetic. J. Physiol. *18*, 280–284.

Lefebvre, J.L., Zhang, Y., Meister, M., Wang, X., and Sanes, J.R. (2008). gamma-Protocadherins regulate neuronal survival but are dispensable for circuit formation in retina. Development *135*, 4141–4151.

Lefebvre, J.L., Kostadinov, D., Chen, W. V, Maniatis, T., and Sanes, J.R. (2012). Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature *488*, 517–521.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523–536.

Matthews, B.J., Kim, M.E., Flanagan, J.J., Hattori, D., Clemens, J.C., Zipursky, S.L., and Grueber, W.B. (2007). Dendrite self-avoidance is controlled by Dscam. Cell *129*, 593–604.

May, G.E., Olson, S., McManus, C.J., and Graveley, B.R. (2011). Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. RNA *17*, 222–229.

McLaughlin, T., and O'Leary, D.D.M. (2005). Molecular gradients and development of retinotopic maps. Annu. Rev. Neurosci. *28*, 327–355.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science (80-.). *338*, 1593–1599.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*,

621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Neves, G., Zucker, J., Daly, M., and Chess, A. (2004). Stochastic yet biased expression of multiple Dscam splice variants by individual cells. Nat. Genet. *36*, 240–246.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Noguchi, Y., Hirabayashi, T., Katori, S., Kawamura, Y., Sanbo, M., Hirabayashi, M., Kiyonari, H., Uchimura, A., and Yagi, T. (2009). Total expression and dual gene-regulatory mechanisms maintained in deletions and duplications of the Pcdha cluster. J. Biol. Chem. *284*, 32002–32014.

Noonan, J.P., Grimwood, J., Danke, J., Schmutz, J., Dickson, M., Amemiya, C.T., and Myers, R.M. (2004a). Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. Genome Res. *14*, 2397–2405.

Noonan, J.P., Grimwood, J., Schmutz, J., Dickson, M., and Myers, R.M. (2004b). Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. *14*, 354–366.

Olson, S., Blanchette, M., Park, J., Savva, Y., Yeo, G.W., Yeakley, J.M., Rio, D.C., and Graveley, B.R. (2007). A regulator of Dscam mutually exclusive splicing fidelity. Nat. Struct. Mol. Biol. *14*, 1134–1140.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Park, J.W., Parisky, K., Celotto, A.M., Reenan, R.A., and Graveley, B.R. (2004).

Identification of alternative splicing regulators by RNA interference in Drosophila. Proc. Natl. Acad. Sci. USA *101*, 15974–15979.

Prasad, T., and Weiner, J. a. (2011). Direct and Indirect Regulation of Spinal Cord Ia Afferent Terminal Formation by the γ-Protocadherins. Front. Mol. Neurosci. *4*, 1–12.

Prasad, T., Wang, X., Gray, P. a, and Weiner, J. a (2008). A differential developmental pattern of spinal interneuron apoptosis during synaptogenesis: insights from genetic analyses of the protocadherin-gamma gene cluster. Development *135*, 4153–4164.

Quesnel-Vallières, M., Irimia, M., Cordes, S.P., and Blencowe, B.J. (2015). Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. Genes Dev. *29*, 746–759.

Raj, B., and Blencowe, B.J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron *87*, 14–27.

Ramani, A.K., Calarco, J. a., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M., et al. (2011). Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Genome Res. *21*, 342–348.

Rubinstein, R., Thu, C.A., Goodman, K.M., Wolcott, H.N., Bahna, F., Mannepalli, S., Ahlsen, G., Chevee, M., Halim, A., Clausen, H., et al. (2015). Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions. Cell *163*, 629–642.

Schmucker, D., and Chen, B. (2009). Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. Genes Dev. *23*, 147–156.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C. a, Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671–684.

Schreiner, D., and Weiner, J.A. (2010). Combinatorial homophilic interaction

between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proc. Natl. Acad. Sci. USA *107*, 14893–14898.

Schreiner, D., Nguyen, T.-M., Russo, G., Heber, S., Patrignani, A., Ahrné, E., and Scheiffele, P. (2014). Targeted Combinatorial Alternative Splicing Generates Brain Region-Specific Repertoires of Neurexins. Neuron *84*, 386–398.

Screaton, G.R., Bell, M. V, Jackson, D.G., Cornelis, F.B., Gerth, U., and Bell, J.I. (1992). Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. Proc. Natl. Acad. Sci. USA *89*, 12160–12164.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014.

Soba, P., Zhu, S., Emoto, K., Younger, S., Yang, S.-J., Yu, H.-H., Lee, T., Jan, L.Y., and Jan, Y.-N. (2007). Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. Neuron *54*, 403–416.

Sperry, R.W. (1963). Chemoaffinity in the orderly growth of nerve fiber patterns and connections. Proc. Natl. Acad. Sci. USA 703–710.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., The RGASP Consortium, Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P., Akerman, M., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods *10*, 1177–1184.

Südhof, T.C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. Nature *455*, 903–911.

Sugino, H., Yanase, H., Hamada, S., Kurokawa, K., Asakawa, S., Shimizu, N., and Yagi, T. (2004). Distinct genomic sequence of the CNR/Pcdhalpha genes in chicken. Biochem. Biophys. Res. Commun. *316*, 437–445.

Sun, W., You, X., Gogol-Döring, A., He, H., Kise, Y., Sohn, M., Chen, T., Klebes, A.,

Schmucker, D., and Chen, W. (2013). Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing. EMBO J. *32*, 2029–2038.

Tada, M.N., Senzaki, K., Tai, Y., Morishita, H., Tanaka, Y.Z., Murata, Y., Ishii, Y., Asakawa, S., Shimizu, N., Sugino, H., et al. (2004). Genomic organization and transcripts of the zebrafish Protocadherin genes. Gene *340*, 197–211.

Takeichi, M. (2007). The cadherin superfamily in neuronal connections and interactions. Nat. Rev. Neurosci. *8*, 11–20.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res. *23*, 201–216.

Thu, C.A., Chen, W. V, Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between α, β, and γ protocadherins. Cell *158*, 1045–1059.

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M.P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc. Natl. Acad. Sci. U. S. A. *1640*, 10–12.

Treutlein, B., Gokce, O., Quake, S.R., Südhof, T.C., and Sudhof, T.C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proc. Natl. Acad. Sci. USA *111*, E1291–E1299.

Ullrich, B., Ushkaryov, Y.A., and Südhof, T.C. (1995). Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. Neuron *14*, 497–507.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore,

S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, X., Weiner, J. a., Levi, S., Craig, A.M., Bradley, A., and Sanes, J.R. (2002). Gamma protocadherins are required for survival of spinal interneurons. Neuron *36*, 843–854.

Wang, X., Li, G., Yang, Y., Wang, W., Zhang, W., Pan, H., Zhang, P., Yue, Y., Lin, H., Liu, B., et al. (2012). An RNA architectural locus control region involved in Dscam mutually exclusive splicing. Nat. Commun. *3*, 1255.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Watson, F.L., Püttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., Rebel, V.I., and Schmucker, D. (2005). Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science *309*, 1874–1878.

Watthanasurorot, A., Jiravanichpaisal, P., Liu, H., Söderhäll, I., and Söderhäll, K. (2011). Bacteria-Induced Dscam Isoforms of the Crustacean, Pacifastacus leniusculus. PLoS Pathog. *7*, e1002062.

Weiner, J.A., Wang, X., Tapia, J.C., and Sanes, J.R. (2005). Gamma protocadherins are required for synaptic development in the spinal cord. Proc. Natl. Acad. Sci. USA *102*, 8–14.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239–1243.

Wu, Q., and Maniatis, T. (1999). A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell *97*, 779–790.

Wu, Q., Zhang, T., Cheng, J.F., Kim, Y., Grimwood, J., Schmutz, J., Dickson, M., Noonan, J.P., Zhang, M.Q., Myers, R.M., et al. (2001). Comparative DNA sequence

analysis of mouse and human protocadherin gene clusters. Genome Res. *11*, 389–404.

Yagi, T. (2012). Molecular codes for neuronal individuality and cell assembly in the brain. Front. Mol. Neurosci. *5*, 45.

Yang, Y., Zhan, L., Zhang, W., Sun, F., Wang, W., Tian, N., Bi, J., Wang, H., Shi, D., Jiang, Y., et al. (2011). RNA secondary structure in mutually exclusive splicing. Nat. Struct. Mol. Biol. *18*, 159–168.

Yano, M., Hayakawa-Yano, Y., Mele, A., and Darnell, R.B. (2010). Nova2 regulates neuronal migration through an RNA switch in disabled-1 signaling. Neuron *66*, 848–858.

Yokota, S., Hirayama, T., Hirano, K., Kaneko, R., Toyoda, S., Kawamura, Y., Hirabayashi, M., Hirabayashi, T., and Yagi, T. (2011). Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster. J. Biol. Chem. *286*, 31885–31895.

Yu, H.-H., Yang, J.S., Wang, J., Huang, Y., and Lee, T. (2009). Endodomain diversity in the Drosophila Dscam and its roles in neuronal morphogenesis. J. Neurosci. *29*, 1904–1914.

Yu, W.-P., Yew, K., Rajasegaran, V., and Venkatesh, B. (2007). Sequencing and comparative analysis of fugu protocadherin clusters reveal diversity of protocadherin genes among teleosts. BMC Evol. Biol. *7*, 49.

Yu, W.-P., Rajasegaran, V., Yew, K., Loh, W.-L., Tay, B.-H., Amemiya, C.T., Brenner, S., and Venkatesh, B. (2008). Elephant shark sequence reveals unique insights into the evolutionary history of vertebrate genes: A comparative analysis of the protocadherin cluster. Proc. Natl. Acad. Sci. USA *105*, 3819–3824.

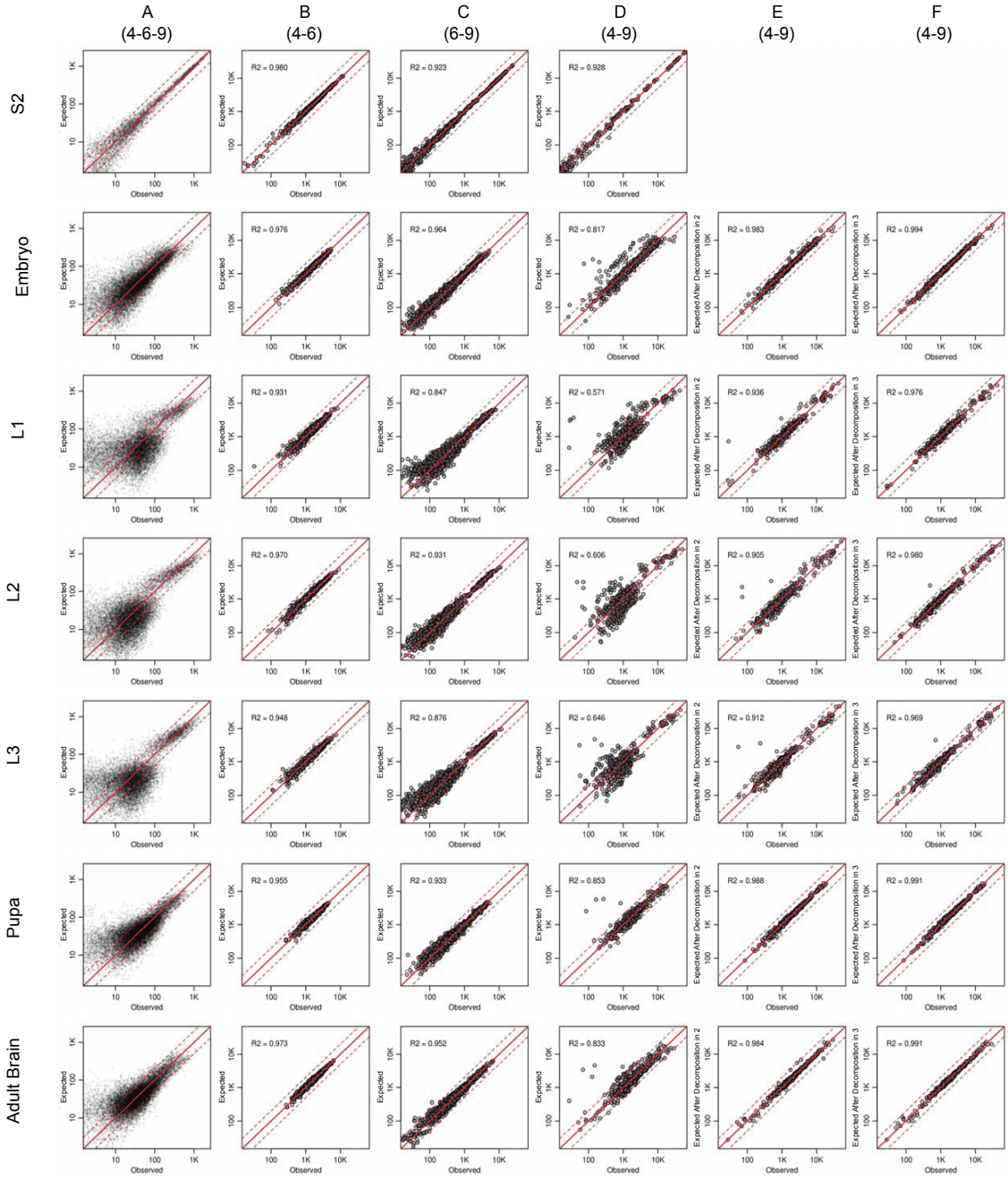
Yu, Y.-C., He, S., Chen, S., Fu, Y., Brown, K.N., Yao, X.-H., Ma, J., Gao, K.P., Sosinsky, G.E., Huang, K., et al. (2012). Preferential electrical coupling regulates neocortical lineage-dependent microcircuit assembly. Nature *486*, 113–117.

Zhan, X.-L., Clemens, J.C., Neves, G., Hattori, D., Flanagan, J.J., Hummel, T., Vasconcelos, M.L., Chess, A., and Zipursky, S.L. (2004). Analysis of Dscam diversity in regulating axon guidance in Drosophila mushroom bodies. Neuron *43*, 673–686.

Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell *143*, 343–353.

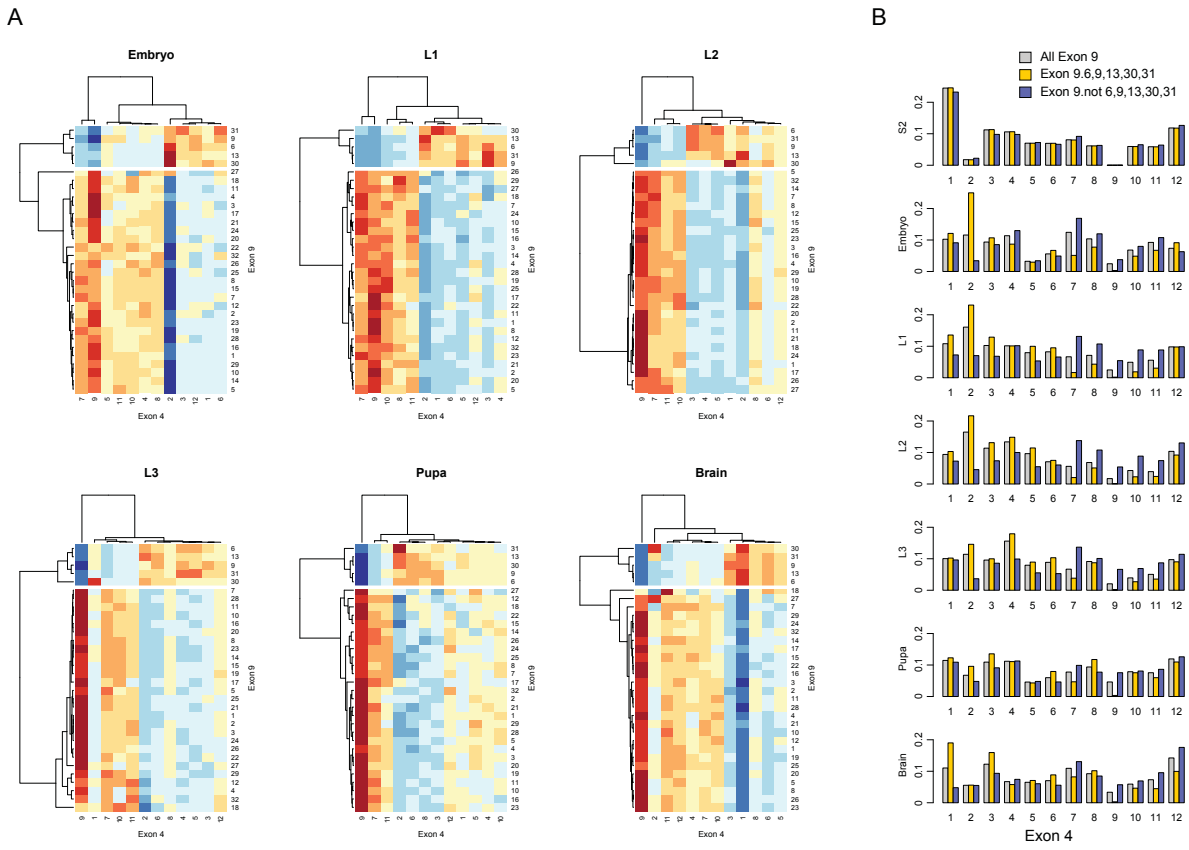
Zou, C., Huang, W., Ying, G., and Wu, Q. (2007). Sequence analysis and expression mapping of the rat clustered protocadherin gene repertoires. Neuroscience *144*, 579–603.

Appendix

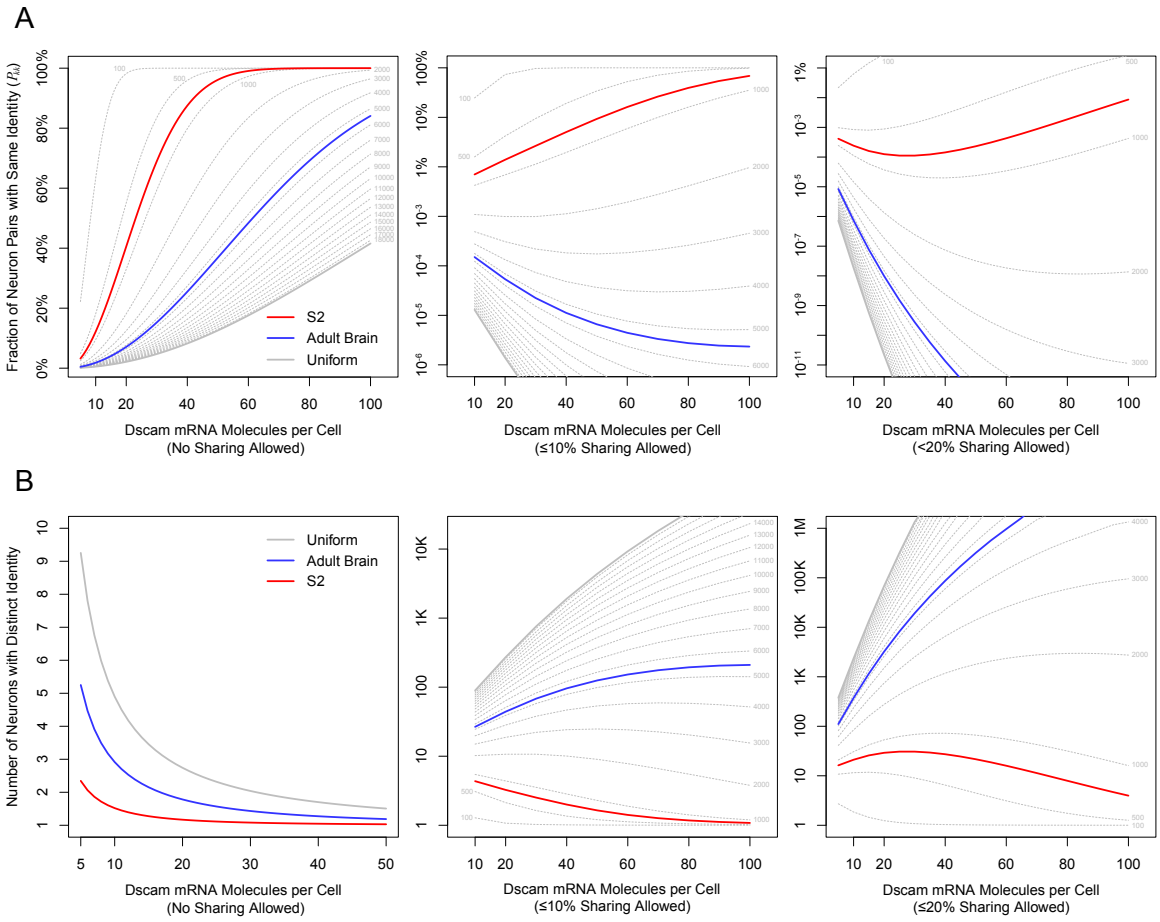


Appendix Figure S2-1. Independent splicing choice between the different variable exon clusters. (**A**). Observed isoform frequency was depicted in X-axis. The expected frequency was calculated by multiplying the frequencies of their respective variable exon 4, 6 and 9 and depicted in Y-axis. To determine whether the splicing between the three clusters was independently controlled, the two frequencies were

compared. The solid red line is the diagonal and the parallel dash red lines represent ranges of one-fold change. (**B**), (**C**), (**D**): In a similar way, we determined whether the splicing choice was independent between exon 4 and 6 (**B**), exon 6 and 9 (**C**), exon 4 and 9 (**D**), respectively. (**E**), (**F**): The whole dataset from each sample could be *in silico* decomposed into two (**E**) or three (**F**) groups. The splicing choice within each group was largely independent between exon 4 and 9. X-axis depicted the observed isoform frequency from the whole dataset, whereas in Y-axis, the expected frequency was the sum of the expected frequencies of the two (**E**) or three (**F**) groups.



Appendix Figure S2-2. (**A**). The variable exon 4s and 9s were clustered based on their expression patterns, the exon 9s could be clearly divided into two groups, one containing only five exons and the other consisting of the remaining 27. (**B**). Given the differential usage of variable exon 4s within the two groups, we could *in silico* decompose the whole data into two sets with different usage of exon 4s and 9s.



Appendix Figure S2-3. (**A**). The fraction of neuron pairs with the same *Dscam* identity (P_{kk}, Y-axis; Methods) when each neuron expresses different number of *Dscam* mRNA molecules (X-axis), if up to 0% (left), 10% (middle) or 20% (right) of isoform are allowed to share between any pair of neurons, calculated based on the distribution of *Dscam* isoform abundances in adult brain, S2 cells, a uniform distribution or in hypothetical samples with different effective size of *Dscam* repertoire (dash lines) (Methods). (**B**). The number of neurons that obtain unique identities at more than 95% likelihood (Y-axis) when each neuron expresses different numbers of *Dscam* mRNA molecules (X-axis), if allowing up to 0% (left), 10% (middle) or 20% (right) of isoforms shared between any pair of neurons, calculated based on the distribution of *Dscam* isoform abundances in adult brain, S2 cells, a hypothetical uniform distribution or in hypothetical samples with different effective

size of *Dscam* repertoire (dash lines) (Methods).

Appendix Table S2-1. Summary of *Dscam* isoform profile in the different samples

Sample	No. Quadruple Reads (million)	No. Detected Isoforms	% Reads from the most abundant 10 (100) isoforms
S2 cell	12.67	7,317	4.8 (25.6)
Embryo	12.89	16,862	1.1 (7.0)
L1	5.71	14,145	1.8 (10.3)
L2	14.35	15,118	2.0 (12.2)
L3	13.57	13,216	1.7 (10.8)
Pupa	11.90	16,876	0.67 (5.3)
Adult brain	15.22	16,886	1.1 (7.0)
Total	86.32	18,496	

Appendix Table S2-2. The effect of different genetic deletions on the effective size of *Dscam* repertoire (excluding peusdo-exon 6.11)

Note: 1. delta 4.1-4.3, delta 4.4-4.12, aggregate 9, aggregate 6,9 strains were described in (Hattori et al., 2009); delta 4.2-4.6 and delta 4.4-4.8 strains were described in (Chen et al., 2006); **2.** We assumed here that the biased usages of the remaining exons are unchanged after deletion. Since this assumption might not hold, as demonstrated by Chen et.al, the number should not be interpreted as the approximation of the real situation. Instead, we used the table to demonstrate the different effects of the deletions of the same number, but of different exons; and the unequal effect of the same deletion in different samples.

Strain	Uniform	Embryo	L1	L2	L3	Pupa	Brain	S2
WT	18,612	5,489	3,119	2,301	2,586	6,503	5,442	769
delta 4.1-4.3	13,959	5,303	3,206	2,079	1,966	5,235	5,376	726
delta 4.4-4.12	4,653	1,047	697	566	666	1,468	929	185
aggregate 9	564	429	409	385	422	463	432	300

aggregate 6,9	12	11	11	10	11	11	11	8
delta 4.2-4.6	10,857	4,165	2,462	2,122	2,006	4,191	3,337	411
delta 4.4-4.8	10,857	2,649	1,537	1,176	1,462	3,805	2,802	382

Appendix Table S3-1. Sequencing summary of the full-length 5'RACE sequencing for *Xtro cPcdh* genes

Note: "replicates 1-3": the full-length 5'RACE read numbers for replicates of different *cPcdh* clusters; "Full-length reads": the sum of the full-length 5'RACE read numbers for replicates of different *cPcdh* clusters in different stages; "mapped to CEs": the sum of the full-length 5'RACE read numbers that could be aligned on the CEs in different stages; "total full-length reads" and "total mapped to CEs": the sums of corresponding read numbers from stage 50 and 60.

Samples		α	γ1	γ2
stage 50	replicate_1	9,751	6,451	8,441
	replicate_2	6,040	6,631	6,789
	replicate_3	6,162	5,868	8,667
	Full-length reads	21,953	18,950	23,897
	mappable to CEs	21,748	18,829	23,831
stage 60	replicate_1	10,677	5,898	9,556
	replicate_2	4,932	5,016	8,362
	replicate_3	7,546	6,241	8,117
	total	23,155	17,155	26,035
	mappable to CEs	23,020	17,080	25,980
total full-length reads		45,108	36,105	49,932
all mapped to CEs		44,768	35,909	49,811

Appendix Table S3-2. Full-length 5'RACE reads mapping statistics on VEs annotated from genome analyses and full-length 5'RACE analyses

Note: "Nr. reads" indicates the numbers of full-length 5'RACE reads mapped to the corresponding VEs.

VEs	Nr. reads	VEs	Nr. reads	VEs	Nr. reads	VEs	Nr. reads
α-1	2722	γ1-12	505	γ1-37	574	γ2-14_novel	980
α-2	940	γ1-13	447	γ1-38	947	γ2-15	591
α-3	960	γ1-14	695	γ1-39	1308	γ2-16	1174
α-4	3820	γ1-15	1215	γ1-40	455	γ2-17	1239
α-5	5127	γ1-16	862	γ1-41	267	γ2-18	1371
α-6	1153	γ1-17	787	γ1-42	258	γ2-19	393
α-7	1030	γ1-18	531	γ1-43	139	γ2-20	619
α-8	2292	γ1-19	198	γ1-44	263	γ2-21	405
α-9	2593	γ1-20	1115	γ1-45	501	γ2-22	531
α-10	4236	γ1-21	637	γ1-46	207	γ2-23	1295
α-11	6438	γ1-22	608	γ1-46_novel	708	γ2-24	953
α-12	1147	γ1-23	1042	γ2-1	996	γ2-25	1130
α-13	165	γ1-24	1490	γ2-2	212	γ2-26	463
α-14	6348	γ1-25	417	γ2-3	666	γ2-27	849
γ1-1	886	γ1-26	646	γ2-4	437	γ2-28	988
γ1-2	157	γ1-27	466	γ2-5	628	γ2-29	272
γ1-3	1087	γ1-28	210	γ2-6	261	γ2-30	1343
γ1-4	175	γ1-29	1020	γ2-7	894	γ2-31	1119
γ1-5	132	γ1-30	321	γ2-8	307	γ2-32	1358
γ1-6	446	γ1-31	485	γ2-9	508	γ2-33	1989
γ1-7	207	γ1-32	584	γ2-10	785	γ2-34	1565
γ1-8	514	γ1-33	1264	γ2-11	414	γ2-35	2624
γ1-9	586	γ1-34	556	γ2-12	1526	γ2-36	2332
γ1-10	290	γ1-35	771	γ2-13	1546		
γ1-11	424	γ1-36	1116	γ2-14	1946		

Appendix Table S3-3. Criteria of manual grouping of IsoSeq high-quality isoforms for the further clustering in rna-VE annotation building

Note: 1. For the high-quality isoforms from IsoSeq analysis, only the ones with length of 2.7 ~ 4 kb were retained for further analysis. The retained IsoSeq isoforms were aligned to VE annotations from genome analyses. These IsoSeq isoforms were first grouped according to their aligned lengths on VE annotations from genome analyses. Only the ones with aligned lengths >= 2 kb were further grouped according to the

modalities of distributions of the numbers of mismatches in the alignments. Others with with aligned lengths < 2 kb were grouped according to the modalities of distributions of their aligned lengths; **2.** "VE": VE annotated from genome analyses; "rna-VE": number of rna-VEs grouped according to mismatch numbers; "cutoffs": the mismatch numbers used in the cutoffs.

VE	rna-VE	cutoffs	VE	rna-VE	cutoffs	VE	rna-VE	cutoffs
α-1	1	-	γ1-19	3	10; 35	γ2-5	1	-
α-2	3	12; 30	γ1-20	2	6	γ2-6	3	10; 20
α-3	1	-	γ1-21	3	25; 40	γ2-7	3	20; 80
α-4	1	-	γ1-22	2	6	γ2-8	2	20
α-5	2	8	γ1-23	3	15; 25	γ2-9	3	4; 20
α-6	1	-	γ1-24	3	20; 45	γ2-10	2	25
α-7	1	-	γ1-25	3	4; 20	γ2-11	2	10
α-8	1	-	γ1-26	2	20	γ2-12	2	5
α-9	1	-	γ1-27	2	20	γ2-13	1	-
α-10	2	25	γ1-28	2	20	γ2-14	1	-
α-11	2	16	γ1-29	3	20; 32	γ2-15	1	-
α-12	2	75	γ1-30	2	40	γ2-16	2	300
α-13	2	16	γ1-31	2	40	γ2-17	1	-
α-14	1	-	γ1-32	2	40	γ2-18	1	-
γ1-1	1	-	γ1-33	2	40	γ2-19	1	-
γ1-2	1	-	γ1-34	2	40	γ2-20	2	10
γ1-3	2	10	γ1-35	1	-	γ2-21	2	20
γ1-4	1	-	γ1-36	3	75; 110	γ2-22	1	-
γ1-5	1	-	γ1-37	1	-	γ2-23	2	20
γ1-6	2	20	γ1-38	1	-	γ2-24	2	7
γ1-7	2	5	γ1-39	1	-	γ2-25	2	10
γ1-8	2	50	γ1-40	3	20; 35	γ2-26	1	-
γ1-9	2	40	γ1-41	2	40	γ2-27	2	20
γ1-10	2	10	γ1-42	2	10	γ2-28	1	-
γ1-11	3	5; 20	γ1-43	2	12	γ2-29	3	20; 60
γ1-12	2	20	γ1-44	2	20	γ2-30	3	5; 20
γ1-13	2	10	γ1-45	2	20	γ2-31	2	40
γ1-14	2	10	γ1-46	2	20	γ2-32	2	20
γ1-15	2	20	γ2-1	1	-	γ2-33	3	50; 120
γ1-16	2	20	γ2-2	1	-	γ2-34	2	20
γ1-17	2	6	γ2-3	2	8	γ2-35	2	60
γ1-18	2	15	γ2-4	1	-	γ2-36	2	8

Appendix Table S3-4. Genome coordinates of annotated VEs and CEs in *Xtro* genome

Note: 1. The coordinates are assigned into *Xtro* genome build xenTro7 scaffold_3 (www.xenbase.org). 2. "Novel": novel exons annotated in full-length 5'RACE sequencing.

Exons	5' end	3' end	Exons	5' end	3' end
α-1	51187032	51184560	γ1-38	50784449	50781894
α-2	51181672	51179172	γ1-39	50777031	50774466
α-3	51175457	51173018	γ1-40	50772116	50769555
α-4	51170581	51168131	γ1-41	50767639	50765052
α-5	51164432	51161987	γ1-42	50761386	50758826
α-6	51156016	51153595	γ1-43	50755061	50752510
α-7	51135553	51133079	γ1-44	50749507	50746980
α-8	51127540	51125099	γ1-45	50744143	50741603
α-9	51119718	51117265	γ1-46	50738906	50736336
α-10	51113091	51110621	γ1-46_novel	50733976	50731454
α-11	51104827	51102380	γ1-CE1	50727636	50727575
α-12	51098136	51095659	γ1-CE2	50723973	50723884
α-13	51088610	51086178	γ1-CE2_novel	50723973	50723868
α-14	51066765	51063862	γ1-CE3	50721331	-
α-CE1	51044416	51044356	γ2-1	56318477	56315963
α-CE2	51042597	51042508	γ2-2	56313475	56310951
α-CE3	51024104	-	γ2-3	56306842	56304301
γ1-1	50991259	50988675	γ2-4	56299725	56297214
γ1-2	50985977	50983443	γ2-5	56292943	56290405
γ1-3	50976418	50973874	γ2-6	56286106	56283561
γ1-4	50971831	50969212	γ2-7	56280320	56277766
γ1-5	50966456	50963933	γ2-8	56274149	56271594
γ1-6	50961157	50958633	γ2-9	56265546	56263011
γ1-7	50956279	50953761	γ2-10	56256801	56254273
γ1-8	50951056	50948543	γ2-11	56248976	56246441
γ1-9	50944790	50942273	γ2-12	56242744	56240210
γ1-10	50938074	50935563	γ2-13	56233406	56230845
γ1-11	50932720	50930382	γ2-14	56224782	56222254
γ1-12	50928160	50925533	γ2-14_novel	56217995	56215403
γ1-13	50922977	50920464	γ2-15	56211737	56209164
γ1-14	50918272	50915643	γ2-16	56207024	56204417
γ1-15	50909565	50906974	γ2-17	56201417	56198818
γ1-16	50902914	50900396	γ2-18	56192906	56190293
γ1-17	50898589	50896047	γ2-19	56183586	56180986

γ1-18	50892956	50890385	γ2-20	56177962	56175330
γ1-19	50888297	50885691	γ2-21	56172758	56170215
γ1-20	50883348	50880894	γ2-22	56166939	56164409
γ1-21	50877284	50874826	γ2-23	56160265	56157739
γ1-22	50871988	50869443	γ2-24	56152478	56149955
γ1-23	50864534	50861973	γ2-25	56146750	56144142
γ1-24	50859436	50856934	γ2-26	56138314	56135705
γ1-25	50853659	50851128	γ2-27	56127450	56124844
γ1-26	50848457	50845859	γ2-28	56119305	56116721
γ1-27	50843728	50841203	γ2-29	56112737	56110131
γ1-28	50838678	50836105	γ2-30	56106282	56103673
γ1-29	50833563	50831003	γ2-31	56099397	56096782
γ1-30	50828040	50825473	γ2-32	56091708	56089098
γ1-31	50822892	50820364	γ2-33	56085258	56082634
γ1-32	50818155	50815568	γ2-34	56077996	56075382
γ1-33	50812425	50809860	γ2-35	56069868	56067259
γ1-34	50807670	50805100	γ2-36	56062197	56059570
γ1-35	50801981	50799417	γ2-CE1	56044817	56044752
γ1-36	50797299	50794730	γ2-CE2	56036496	56036407
γ1-37	50790175	50787613	γ2-CE3	56033796	-

Publications

As First Author or Co-first Author (*Co-first Authors)

1. Gao Q*, **Sun W***, Ballegeer M, Libert C, Chen W. *Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing.* **Molecular Systems Biology**. 2015, 11: 816

2. Armitage SAO*, **Sun W***, You X*, Kurtz J, Schmucker D, Chen W. *Quantitative Profiling of Drosophila melanogaster Dscam1 Isoforms Reveals No Changes in Splicing after Bacterial Exposure*. **PLoS ONE**. 2014, 9(10): e108660.

3. **Sun W***, You X*, Gogol-Döring A*, He H, Kise Y, Sohn M, Chen T, Klebes A, Schmucker D, Chen W. *Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing.* **The EMBO Journal**. 2013, 32(14): 2029-38.

Co-authorship

4. Bertelsen B, Nazaryan-Petersen L, **Sun W**, Mehrjouy MM, Xie G, Chen W, Hjermind LE, Taschner PE, Tümer Z. *A germline chromothripsis event stably segregating in 11 individuals through three generations.* Genetics in Medicine. 2015 Aug 27. doi: 10.1038/gim.2015.112. [Epub ahead of print]

5. Hou J, Wang X, McShane E, Zauber H, **Sun W**, Selbach M, Chen W. *Extensive allele-specific translational regulation in hybrid mice.* Molecular Systems Biology. 2015, 11: 825.

6. You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C, Wang X, Hou J, Liu H, **Sun W**, Sambandan S, Chen T, Schuman EM, Chen W. *Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity.* Nature Neuroscience. 2015, 18(4): 603-10.

7. Bertelsen B, Melchior L, Jensen LR, Groth C, Nazaryan L, Debes NM, Skov L, Xie G, **Sun W**, Brøndum-Nielsen K, Kuss AW, Chen W, Tümer Z. *A t(3;9)(q25.1;q34.3) translocation leading to OLFM1 fusion transcripts in Gilles de la Tourette syndrome, OCD and ADHD.* Psychiatry Research. 2015, 225(3): 268-75.

8. Szentiks CA, Tsangaras K, Abendroth B, Scheuch M, Stenglein MD, Wohlsein P, Heeger F, Höveler R, Chen W, **Sun W**, Damiani A, NikolinV,Gruber AD, Grobbel M, Kalthoff D, Höper D, Czirják GÁ, DeRisi J, Mazzoni CJ, Schüle A, Aue A, East ML, Hofer H, Beer M, Osterrieder K, Greenwood AD. *Polar bear encephalitis: Establishment of a comprehensive next-generation pathogen analysis pipeline for captive and free-living wildlife.* Journal of Comparative Pathology, 2014, 150(4): 474-88.

9. Gao Q, **Sun W**, You X, Fröhler S, Chen W. *A systematic evaluation of hybridization-based mouse exome captures system.* BMC Genomics 2013, 14: 492.

10. Mayer J, Tsangaras K, Heeger F, Avila-Arcos M, Stenglein MD, Chen W, **Sun W**, Mazzoni CJ, Osterrieder N, Greenwood AD. *A novel endogenous betaretrovirus group characterized from polar bears (Ursusmaritimus) and giant pandas (Ailuropodamelanoleuca).* Virology. 2013, 433(1): 1-10.