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

Wei Sun^{1,4}, Xintian You^{1,4},
Andreas Gogol-Döring^{1,4}, Haihuai He²,
Yoshiaki Kise², Madlen Sohn¹, Tao Chen¹,
Ansgar Klebes³, Dietmar Schmucker^{2,*}
and Wei Chen^{1,*}

¹Laboratory for Novel Sequencing Technology, Functional and Medical Genomics, Berlin Institute for Medical Systems Biology, Max-Delbrück-Centrum für Molekulare Medizin, Berlin, Germany, ²Laboratory of Neuronal Wiring, Vesalius Research Center, VIB, Department of Oncology, K.U. Leuven, Leuven, Belgium and ³Fakultät für Mathematik und Naturwissenschaften I, Institut für Biologie, Humboldt Universität Berlin, Berlin, Germany

The *Drosophila melanogaster* gene *Dscam* (Down syndrome cell adhesion molecule) can generate thousands of different ectodomains via mutual exclusive splicing of three large exon clusters. The isoform diversity plays a profound role in both neuronal wiring and pathogen recognition. However, the isoform expression pattern at the global level remained unexplored. Here, we developed a novel method that allows for direct quantification of the alternatively spliced exon combinations from over hundreds of millions of *Dscam* transcripts in one sequencing run. With unprecedented sequencing depth, we detected a total 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 were expressed across a broad dynamic range, with significant bias 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 nevertheless be essential for a robust self and non-self discrimination in neurons.

The EMBO Journal (2013) 32, 2029–2038. doi:10.1038/emboj.2013.144; Published online 21 June 2013

Subject Categories: RNA; neuroscience

Keywords: alternative splicing; CAMSeq; *Dscam*; ultra-deep profiling

Introduction

Alternative splicing of precursor messenger RNA (pre-mRNA) makes substantial contribution to the expansion of protein diversity (Nilsen and Graveley, 2010; Barbosa-Morais *et al*, 2012; Merkin *et al*, 2012). 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* (Ullrich *et al*, 1995), and *protocadherins* (Wu and Maniatis, 1999). The most extreme case is the *Drosophila melanogaster* homologue of Down syndrome cell adhesion molecule (*Dscam*) gene (Schmucker *et al*, 2000). The *Dscam* gene locus contains 115 exons, of which 95 are arranged into four clusters, that is, exons 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 (Zhan *et al*, 2004; Watson *et al*, 2005; 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; Watthanasurorot *et al*, 2011).

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 wild-type conditions, neighbouring 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 neighbouring neurons also express identical isoforms, then 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 clustered protocadherin 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 protocadherin-gamma receptors are important for self-avoidance in retinal cells in mice (Lefebvre *et al*, 2012) and

*Corresponding authors. D Schmucker, Laboratory of Neuronal Wiring, Vesalius Research Center, VIB, Department of Oncology, K.U. Leuven, B-3000 Leuven, Belgium. Tel.: +32 16373222; E-mail: dietmar.schmucker@vib-kuleuven.be or W Chen, Laboratory for Novel Sequencing Technology, Functional and Medical Genomics, Berlin Institute for Medical Systems Biology, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, Berlin 13125, Germany. Tel.: +49 30 94062995; Fax: +49 30 94063048; E-mail: wei.chen@mdc-berlin.de

⁴These authors contributed equally to this work

Received: 18 April 2013; accepted: 27 May 2013; published online: 21 June 2013

that due to alternative splicing and tetramer formation of protocadherins 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.

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; Zhan *et al*, 2004; Watson *et al*, 2005). 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 (Park *et al*, 2004; Graveley, 2005; Krehling and Graveley, 2005; Anastassiou *et al*, 2006; Olson *et al*, 2007; May *et al*, 2011; Yang *et al*, 2011; Wang *et al*, 2012).

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 (Neves *et al*, 2004; Chen *et al*, 2006). 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; Wilhelm *et al*, 2008; Wang *et al*, 2009). 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.

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 exons 4, 6, and 9. We analysed 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 the alternative splicing between different exon clusters was 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.

Results

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 1A. In this study, we focused on the combinations of variable exons 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 as '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 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 labelled 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 ~1 kb in length was then sequenced. As shown in Figure 1B, we modified the standard Illumina sequencing procedure and obtained from every template DNA molecule four sequencing reads (quadruple-reads) derived from exons 4, 6, 9, and barcode, respectively (Materials and 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 exons 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 (Materials and methods). As shown in Figure 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 analysed twice the same RNA extracted from S2 cells, a cell line derived from *Drosophila* embryonic hemocytes. As shown in Figure 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 (Materials and methods). A total of 63 109

Figure 1 CAMSeq, a novel massive parallel sequencing-based method for quantitative profiling *Dscam* isoforms. **(A)** Alternative splicing of *Drosophila 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 Immunoglobulin (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 ~2 kb RT-PCR product and another round of PCR with the primers targeting constitutive exon 7 and exon 8, the amplification product of ~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 exons 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 $F_{4,6,9,b,b'}$, whereas X axis is the product between the frequency of reads containing the same exons 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, $F_{4,6,b} \times F_{9,b'}$. Assuming a second-order reaction kinetics, the mean chimerical rate could be represented by the slope of regression line, i.e., 0.0097 (Materials and methods).

PacBio reads could be used to reveal the identities of exons 4, 6, and 9 for 3725 *Dscam* isoforms and the isoform abundances estimated using the two approaches showed a high correlation ($R^2 = 0.978$; Figure 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 labelled with different barcodes joined together (Materials and methods). As shown in Figure 1F, the mean chimerical rate was $\sim 1\%$.

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

We used CAMSeq to analyse *Dscam* isoform expression at different developmental stages (embryo, 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 1B; Materials and 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 exons 4, 6, and 9 as well as the barcode representing a specific sample (Supplementary Table S1). 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 clusters 4 and 9 showed moderate to dramatic changes, whereas the differences in exon 6 clusters were relatively modest (Figure 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 (Materials and methods; Supplementary Table S1). The number was quite close to 18 612, the maximum number of potential isoforms if excluding the pseudo-exon 6.11, indicating that all the remaining *Dscam* isoforms were expressed.

In each sample, the relative abundance of different isoforms spanned at least four orders of magnitude (Figure 2B; Supplementary Table S2). The most abundant 10 and 100 isoforms derived 0.7–2.0%, and 5.3–12.2% of all reads from one sample (Supplementary Table S1). Importantly, our comparison between different samples showed that S2 cells express a significantly more restricted repertoire of *Dscam* isoforms in which only 7317 isoforms were detected with the most abundant 100 isoforms accounting for 25.6% of all reads (Supplementary Table S1). 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 2A). Given the observation that the splicing choice of exon 9 was most variable between different cell types (Figure 2A), this corroborates our hypothesis that the larger repertoire observed in the other samples was due to the much higher cell-type diversity.

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 exons 4, 6, and 9. Comparison between the observed and expected isoform frequencies in different samples showed mixed results

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

Figure 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 exons 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–D) In a similar way, we determined whether the splicing choices were independent between exons 4 and 6 (B), exons 6 and 9 (C), and exons 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 exons 4 and 9, the yellow and blue groups. (G) The splicing choice within each group was largely independent between exons 4 and 9. X axis depicted the observed isoform frequencies from the whole brain data set, whereas in Y axis, the expected isoform frequencies were the sum of the expected frequencies of the two groups.

(Figure 3A; Supplementary Figure S1A). 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 3A; Supplementary Figure S1A).

Given that the splicing choice of exons 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 analysed splicing choices between exons 4 and 6, exons 6 and 9, as well as exons 4 and 9, separately. Indeed, whereas the splicing appeared to be independent between exons 4 and 6, as well as between exons 6 and 9 in all samples (Figure 3B and C; Supplementary Figure S1B and C), the splicing between exons 4 and 9 showed different patterns between S2 and other samples (Figure 3D; Supplementary Figure S1D). 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 3E and Supplementary Figure S2A, 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 data set being largely independent between the two clusters (Figure 3F and 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 (Supplementary Figures S1E, F and S2B).

Discussion

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 to identify the expression of *Dscam* isoforms directly by determining for each isoform the precise combination of exons 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 Immunoglobulin (Ig) domain folding (Dietmar Schmucker, personal communications). Notably, we could detect a very minor fraction of isoforms skipping either of exons 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.

We demonstrated that the 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* (2006) 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 exons 6 and 9 is independent from that of exon 4. In another study, using S2 cells, Neves *et al* analysed 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 that 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 exons 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 3; Supplementary Figures S1 and S2).

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 laboratory 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 modelling 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 modelling study using our actual quantitative data sets. 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 4A). For example, *Drosophila* mushroom body (MB) comprises some 2500 neurons. If each individual MB neuron expresses 20 *Dscam* mRNA copies (Zhan *et al*, 2004) and all possible isoforms express with equal probability, then 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, then only 12 300 different isoforms would be present in one MB (Figure 4A; Materials and methods). With such reduced repertoire, obviously the number of neurons with unique *Dscam* identity also decreases (Figure 4B; Materials and 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 3200 neurons could be uniquely labelled (Figure 4B; see Supplementary Figure S3B for the conditions in which up to 0 or 10% of isoforms were allowed to share between two neurons). The same labelling capacity could also be coded by about 5500 uniformly expressed isoforms. To facilitate the comparison of labelling 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 (Materials and 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 data set (the yellow group in Figure 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

Figure 4 Monte Carlo simulation of *Dscam* repertoire and the number of neurons that could be labelled 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 (Materials and methods). (B) The number of neurons that obtain unique identities at >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 (Materials and methods). See Supplementary Figure S3B for the condition in which up to 0 or 10% of isoforms are allowed to share between any pair of neurons.

number of cells would be quite small and only dozens of cells could be labelled with unique *Dscam* identities when any pair of neurons were allowed to share 20% of their expressed *Dscam* isoforms (Figure 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-labelling’ strategies could be used for the proper wiring in a large group of interconnecting neurons. That is, they can either be a homogeneous cell population 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 set 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.

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 will be unequal in different cell types

with distinct splicing patterns (Supplementary Table S3). 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 (Supplementary Table S3). In addition, the number of neurons that could be labelled 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 labelled neurons (see S2 sample in Figure 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.

Materials and methods

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, first larval, second larval, and third larval stages were collected according to the time period after egg laying (embryos 13–18 h, first stage larvae 24–36 h, second stage larvae 60–72 h, and

third stage larvae (96–108 h). Fruit fly pupae were collected 0–48 h after puparium formation. Adult brains were dissected from 1- to 3-day-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).

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'-TGTCTGCTGGTGAAGCATAG-3') using SuperScript III system with a reaction volume of 20 μl (Life Technologies). PCR was followed using 2 μl of RT 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'-GAGGTCCATGCCAGGTGTA CG-3'; DsRef-1-R: 5'-GTCGACATGCAGAGTGCCTC-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 exons 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'-GGATCCTAATACGACTCACTATAGGGATCCATTATCTCCCGGACG TCCATGT-3'; DsRef-2-R: 5'-GTCGACATGCAGAGTGCCTC-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 *in 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.

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'-GTCGCTCTCTTATGATCC TTGTAC-3') using SuperScript III system with a reaction volume of 20 μl. The first 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'-AGNNNNACCAT TATCTCCGGGACGTCCATGTC-3'; CAMSeq-1-R: 5'-GTNNNNAC CTTATCGGTGGGCTCGAGGATCCA-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 first 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 first 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 1500 and 2500 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 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 second 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'-AATGATACGGCGACCAC CGAGATCTACACTGGATACTCTGCTCGAGGATCTCTGGAAGTGC-3';

CAMSeq-2-R: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCCAGCTT GTTTACGGGTGTTCTCTTCGATGA-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 second 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'-CCTCCAGATGGATCCTCGAGCC CACCGATAAG-3'; CAMSeq-ex9-seq-primer: 5'-GATACTCTGCTCGA GGATCTCTGGAAGTGCAAGTCA-3'; CAMSeq-ex6-seq-primer: 5'-CG ATTAAGTCCACAAAAGGACGATTGGTCATCA-3'; CAMSeq-ex4-seq-primer: 5'-CCATTATCTCCCGGACGTCCATGTGCGAG-3'). After each sequencing, 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.

PacBio sequencing of *Dscam* isoforms

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

Processing of CAMSeq data

Each Illumina sequencing read was split into four segments derived from barcode (1st–25th nt), exon 9 (26th–61st nt), exon 6 (62nd–97th nt), and exon 4 (98th–123rd nt), respectively. The three segment sequences corresponding to exons 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 first 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.

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.

Estimation and correction of the chimeric effect

To estimate the rate of forming chimeras, we first identify the reads derived from the intermolecular 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

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

where $F_{4,6,9,b,b'}$ is the frequency of the chimeric product containing a distinct set of exons 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., $T \times F_{4,6,b} \times F_{9,b'} \geq 100$, 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 $r_{4,6,9,b,b'}$ values as a set of independent variables. The slope r_{avg} of a linear regression line with intercept 0 through the points $x = F_{4,6,b} \times F_{9,b'}$ and $y = F_{4,6,9,b,b'}$, $b \neq b'$ was used as an average chimeric rate (Figure 1F).

For $b = b'$ and a given chimera rate r , we could calculate the expected number of chimeric reads by

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

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 $r = r_{\text{avg}}$.

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, that is, 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 $n \times \alpha = 1$.

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, then 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:

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

If we assume for example that two distinct cells are allowed to share up to 20% of their *Dscam* transcripts, then we would set $h = 0.2 \times k$. P_{kk} can then be interpreted as the probability for two distinct cells getting the same *Dscam* identity by chance, see Supplementary Figure S3A.

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

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

where $m \times (m-1)/2$ is the number of all possible pairwise combinations of the m cells. If we set for example $Q = 0.95$, given P_{kk} , then m could be computed, see Supplementary Figure S3B.

References

- Anastassiou D, Liu H, Varadan V (2006) Variable window binding for mutually exclusive alternative splicing. *Genome Biol* 7: R2
Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodenic V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ (2012)

Clustering of exons 4 and 9 based on the expression patterns

We created heatmaps using the heatmap.2 function from the R package plots 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 $d = (1 - R)/2$ where R is the Pearson correlation coefficient.

Decomposition of *Dscam* isoform distribution data sets

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

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

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

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

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

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

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.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (<http://www.embojournal.org>).

Acknowledgements

We thank Mirjam Feldkamp, Claudia Langnick, and Claudia Quedenau from Wei Chen's group for their excellent technical assistance. We also thank Oliver Goldenberg from Illumina and Christoph Koenig from Pacific Biosciences for their supports. We thank Christine Kocks from MDC for helpful discussion. As part of the Berlin Institute for Medical Systems Biology at the MDC, the research group of Wei Chen is funded by the Federal Ministry for Education and Research (BMBF) and the Senate of Berlin, Berlin, Germany (BIMSB 0315362A, 0315362C). WS is supported by the Chinese Scholarship Council (CSC). HH and DS were funded by VIB and FWO. YK was supported by an HFSP fellowship.

Author contributions: WC conceived and designed the project. WS performed the experiments with the help of MS and TC. XY and AGD analysed the data. HH, YK, AK, and DS provided all fruit fly materials. WS, XY and AGD contributed to part of the manuscript. DS and WC wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

- Chen BE, Kondo M, Garnier A, Watson FL, Püettmann-Holgado R, Lamar DR, Schmucker D (2006) The molecular diversity of Dscam is functionally required for neuronal wiring specificity in *Drosophila*. *Cell* **125**: 607–620
- Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *NatMethods* **5**: 613–619
- Dong Y, Cirimotich CM, Pike A, Chandra R, Dimopoulos G (2012) Anopheles NF- κ B-regulated splicing factors direct pathogen-specific repertoires of the hypervariable pattern recognition receptor AgDscam. *Cell Host Microbe* **12**: 521–530
- Dong Y, Taylor HE, Dimopoulos G (2006) AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol* **4**: e229
- Graveley BR (2005) Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**: 65–73
- Hattori D, Chen Y, Matthews BJ, Salwinski L, Sabatti C, Grueber WB, Zipursky SL (2009) Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms. *Nature* **461**: 644–648
- Hattori D, Demir E, Kim HW, Viragh E, Zipursky SL, Dickson BJ (2007) Dscam diversity is essential for neuronal wiring and self-recognition. *Nature* **449**: 223–227
- Hughes ME, Bortnick R, Tsubouchi A, Bäumer P, Kondo M, Uemura T, Schmucker D (2007) Homophilic Dscam interactions control complex dendrite morphogenesis. *Neuron* **54**: 417–427
- Jacob F (1977) Evolution and tinkering. *Science* **196**: 1161–1166
- Kreahling J, 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
- Lefebvre JL, Kostadinov D, Chen WV, Maniatis T, Sanes JR (2012) Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. *Nature* **488**: 517–521
- Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**: 523–536
- Matthews BJ, Kim ME, Flanagan JJ, Hattori D, Clemens JC, Zipursky SL, Grueber WB (2007) Dendrite self-avoidance is controlled by Dscam. *Cell* **129**: 593–604
- May GE, Olson S, McManus CJ, Graveley BR (2011) Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. *RNA* **17**: 222–229
- Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**: 1593–1599
- Mortazavi A, Williams BA, McCue K, Schaeffer L, 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, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349
- Neves G, Zucker J, Daly M, Chess A (2004) Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet* **36**: 240–246
- Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463
- Olson S, Blanchette M, Park J, Savva Y, Yeo GW, Yeakley JM, Rio DC, Graveley BR (2007) A regulator of Dscam mutually exclusive splicing fidelity. *Nat Struct Mol Biol* **14**: 1134–1140
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–1415
- Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR (2004) Identification of alternative splicing regulators by RNA interference in *Drosophila*. *Proc Natl Acad Sci USA* **101**: 15974–15979
- Schmucker D, Chen B (2009) Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. *Genes Dev* **23**: 147–156
- Schmucker D, Clemens JC, Shu H, Worby Ca, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684
- Schreiner D, Weiner JA (2010) Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. *Proc Natl Acad Sci USA* **107**: 14893–14898
- Screaton GR, Bell MV, Jackson DG, Cornelis FB, Gerth U, Bell JI (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
- Soba P, Zhu S, Emoto K, Younger S, Yang S-J, Yu H-H, Lee T, Jan LY, Jan Y-N (2007) *Drosophila* sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. *Neuron* **54**: 403–416
- Ullrich B, Ushkaryov YA, Südhof TC (1995) Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* **14**: 497–507
- Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476
- Wang X, Li G, Yang Y, Wang W, Zhang W, Pan H, Zhang P, Yue Y, Lin H, Liu B, Bi J, Shi F, Mao J, Meng Y, Zhan L, Jin Y (2012) An RNA architectural locus control region involved in Dscam mutually exclusive splicing. *Nat Commun* **3**: 1255
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63
- Watson FL, Püttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, Schmucker D (2005) Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* **309**: 1874–1878
- Wathanasurot A, Jiravanichpaisal P, Liu H, Söderhäll I, Söderhäll K (2011) Bacteria-induced Dscam isoforms of the Crustacean, *Pacifastacus leniusculus*. *PLoS Pathog* **7**: e1002062
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239–1243
- Wu Q, Maniatis T (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**: 779–790
- Yang Y, Zhan L, Zhang W, Sun F, Wang W, Tian N, Bi J, Wang H, Shi D, Jiang Y, Zhang Y, Jin Y (2011) RNA secondary structure in mutually exclusive splicing. *Nat Struct Mol Biol* **18**: 159–168
- Yu H-H, Yang JS, Wang J, Huang Y, Lee T (2009) Endodomain diversity in the *Drosophila* Dscam and its roles in neuronal morphogenesis. *J Neurosci* **29**: 1904–1914
- Zhan X-L, Clemens JC, Neves G, Hattori D, Flanagan JJ, Hummel T, Vasconcelos ML, Chess A, Zipursky SL (2004) Analysis of Dscam diversity in regulating axon guidance in *Drosophila* mushroom bodies. *Neuron* **43**: 673–686
- Zipursky SL, Sanes JR (2010) Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. *Cell* **143**: 343–353