Supplemental figures for

Comprehensive transcriptome analysis using synthetic long read sequencing reveals molecular co-association and conservation of distant splicing events.

Hagen Tilgner*, Fereshteh Jahanbani*, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger,
Feng Chen, Itamar Harel,
Carlos D Bustamante, Morten Rasmussen & Michael Snyder
* these authors contributed equally for this work

Nature Biotechnology: doi:10.1038/nbt.3242

1 Mapping statistics in the human brain

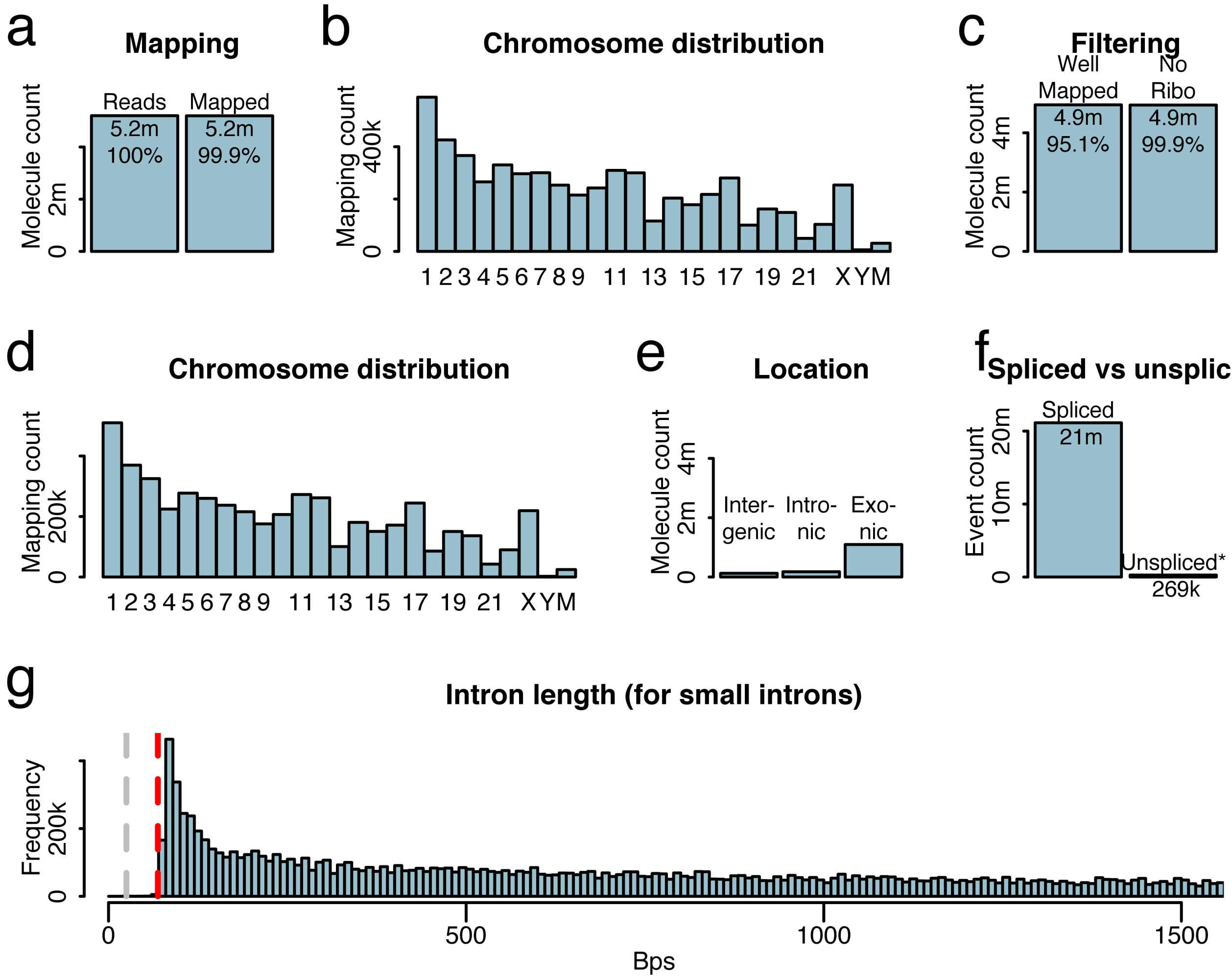

Figure S1: Mapping statistics. **(a)** Number of molecules submitted for mapping (left bar) and number (and percentage) of molecules that could be mapped to the hg19 genome using GMAP. **(b)** Chromosome distribution of molecule mappings including multiple mappings for a single molecule if GMAP produced such. **(c)** Number and percentage of molecules for which we could determine a single high-confidence-mapping (well-mapped, left bar) and those that did not overlap ribosomal RNA genes (right, percentage with respect to previous bar). **(d)** Chromosome distribution of high confidence read mappings. **(e)** Number of molecules falling entirely into intergenic, intronic and exonic regions. Note that the definition of intergenic used here is based on the Gencode-v15 annotation, which defines lncRNA genes, ribosomal RNA genes and many other kinds of short RNA genes as ”genes”. **(f)** Number of introns in high-confidence mappings (left) and number of events of incomplete splicing (intron-retention and/or partially spliced nuclear RNAs). **(g)** Intron length distribution for introns in consensus-split-molecule-mappings, showing only introns of up to 1.5kb. The grey dashed line (at 25bps) indicates that we asked GMAP to consider splits below 25bps as insertions/deletions. The red dashed line (at 70bps) indicates a cutoff under which very few annotated human introns could be found (see reference 8), suggesting a minimal intron-size of this length, under which human introns might be difficult to process for the splicing machinery. Reassuringly GMAP reported almost no introns between 25 and 70bps, indicating that intron-calls are generally high quality calls.

2 Biases and reproducibility for gene expression

We calculate molecules per million (MPM) for each spliced gene by counting the number of consensus-split-mapped-molecules (CSMM - those molecules that were split-mapped and for which each split respected the splice site consensus) and divided this

Table S1: average read length values by technology and dataset

system	sample	study	mean length (bp)
PacBio-CCS	Panel of human organs	Sharon D. et al, Nature Biotechnology, 2013	999.9
PacBio-CCS	GM12878	Tilgner H. et al, PNAS, 2014	1,178
PacBio-CCS	Human brain	Li S. et al, Nature Biotechnology, 2014	1,289
SLR-RNA-Seq	Human brain	this study	1,906
SLR-RNA-Seq	Mouse brain	this study	1,849

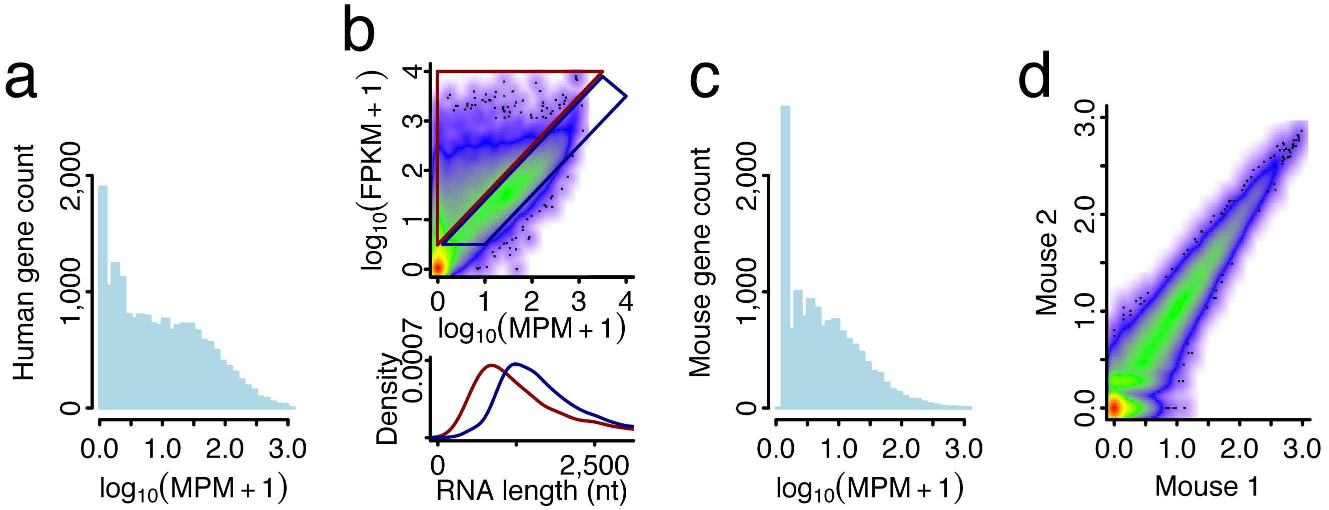

Figure S2: Histogram of human molecules per million (MPM)-values (after log_{10} transformation) for all spliced genes with one or more spliced reads **(a)**. Scatterplot of log_{10} transformed human MPM-values and log_{10} transformed human FPKM-values. The latter are calculated using STAR and Cufflinks from published short-read RNA-sequencing[31] **(top)**. Length distribution for genes identified by both approaches (blue) and for genes identified only by the short-read approach **(bottom, b)**. Histogram of mouse molecules per million (MPM)-values (after log_{10} transformation) for all spliced genes with one or more spliced reads **(c)**. Scatterplot of mouse gene MPMs (after log_{10}-transformation) between mouse 1 (x-axis) and mouse 2 (y-axis). We considered all annotated spliced genes. The Spearman correlation corresponding to the scatterplot is 0.9 ($p < 2.2e - 16$, correlation-test as implemented in test.cor in R) **(d)**.

number by the uniquely mapped molecules overall. Figure S2a shows the histogram of all human spliced genes that received at least one spliced read. To compare these expression measurements with more traditional expression FPKMs, we mapped published RNA-Seq data from human brain[31] (GSM1166113, GSM1166114, GSM1166115, GSM1166116) using STAR and and quantified the gencode 15 annotation using Cufflinks. $\underline{STAR - parameters}$ used are the following (in addition to being given an index containing the hg19-sequence and the Gencode v15 junctions):

```
--readFilesCommand zcat --outSAMunmapped Within --outFilterType BySJout --outFilterMultimapNmax 20
--alignSJoverhangMin  8 --alignSJDBoverhangMin 5 --outFilterMismatchNmax 999
--outFilterMismatchNoverLmax 0.04 --alignIntronMin 25 --alignIntronMax 1000000 --alignMatesGapMax
1000000 --runThreadN 4 --outSAMstrandField intronMotif --outStd SAM
```

The resulting sam file was gzipped and sorted using $\underline{samtools}$:

```
zcat tmp.STAR.sam.gz | samtools view -bS - | samtools sort -m 28G - mapping
```

$\underline{Cufflinks - parameters}$ employed were:

```
-G Gencode.v15.file(exonsOnly) --min-intron-length 25 -p 4
```

Gene FPKMs were averaged over the Cufflinks gene FPKMs for the four datasets. Figure S2b (top) shows the scatterplot of SLR-RNA-Seq-MPMs and Cufflinks FPKMs, including one main area (blue box) of enrichment along the diagonal. A secondary area of enrichment (along the y-axis, red triangle) represents genes only recovered by the STAR-Cufflinks

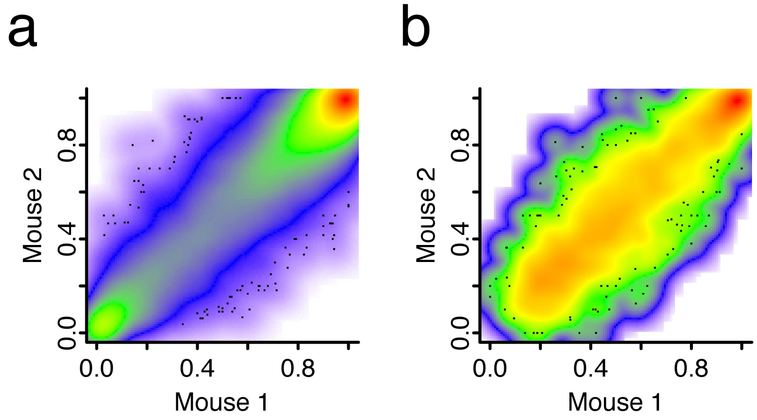

Figure S3: Scatterplot of splice site Ψ-values between mouse 1 (x-axis) and mouse 2 (y-axis). We considered all splice sites that were overlapped by 10 or more spliced reads (CSMMs) in each mouse. The Spearman correlation corresponding to the scatterplot is 0.76 ($p < 2.2e - 16$, correlation-test as implemented in test.cor in R) **(a)**. Scatterplot of transcript Percent-Isoform (Π)-values for major isoforms between mouse 1 (x-axis) and mouse 2 (y-axis). We considered all exon-intron-structures, that were classified at least once as full-length and that belonged to genes with 10 or more usable spliced reads for Π-calculation. The Spearman correlation corresponding to the scatterplot is 0.88 ($p < 2.2e - 16$, correlation-test as implemented in test.cor in R). **(b)**.

approach. The STAR-Cufflinks-only genes are significantly shorter than the genes recovered by both technologies, supposedly because SLR-RNA-Seq disfavors short molecules. In more detail, SLR-RNA-Seq enriches for long molecules in two ways. First, Illumina sequencing requires heterologous adapters on the ends of the DNA fragments for efficient sequencing. Here, homologous adapters are added to the ends of each full length molecule; therefore each molecule requires at least one Nextera tagmentation event to generate the heterologous ends needed for efficient sequencing. Because the probability of a Nexera tagmentation event is largely proportional to molecule length, the shortest molecules are the least likely to acquire at least one tagmentation event, and therefore least likely to have heterologous adapters. Second, because the final library has been size selected to only include DNA inserts longer than 350bp, shorter transcripts that do get tagmented are more likely to have fragments less than 350bp. Therefore shorter transcripts will have lower coverage, resulting in more difficulty assembling and reporting the full length molecule.
STAR-Cufflinks-only showed two- to three fold enrichments of Gencode-pseudogenes ($p < 2.2e - 16$, possibly because multi-mapping short reads may cause simultaneous detection of parent gene and pseudogene) and Gencode antisense-genes ($p < 2.2e - 16$). Figure S2c shows the histogram of mouse MPMs and figure S2d the scatter plot of mouse MPMs for mouse 1 and mouse 2, which show a Spearman correlation of 0.90.

3 Novel isoform validation in public datasets

Novel isoforms can be novel (with respect to the annotation) either because they use a novel combination of known introns or because they use a new intron. While the former case is difficult to recapitulate with short-read-sequencing, the latter case can be easily checked in other RNA-Seq datasets. As pointed out in the main text, we could validate 73.8% of our 137k novel introns in this way using long-read datasets of our own as well as short-read RNA-Seq from the ABRF[31] - and ENCODE[10] studies. We then continued to assess what this percentage would be for a random set of introns.

1. Starting with our 137,000 new introns (defined through SLR-RNA-Seq), we shifted their acceptors to the nearest upstream (downstream on the reverse strand) AG-dinucleotide and the donor to the nearest upstream (downstream on the reverse strand) GU-dinucleotide. We thus achieve a set of random introns that preserves the gene expression distribution (and roughly the intron length distribution) of the original introns. We find a "background validation rate" of 1.7% (more than 40-fold lower than the 73.8% we find for the new introns from SLR-RNA-Seq).

2. We first build the set of all acceptor splice sites A, originitating either from the annotation or from our aligned SLR-RNA-Seq reads. Similarly we build the set of all donor splice sites from both the annotation and our aligned reads. From these we built all possible combinations that originated from the same gene and that respected transcriptional order (donor before accecptor in transcription direction). For splice sites from reads that could not be attributed to any gene at all, we build random introns, when they were on the same chromosome and strand and no more than 100kb away from each other (but again respecting transcriptional order). We thus generated a set of 4 million random introns, of which 9.9% could be found in the RNA-Seq datasets of the previously mentioned studies (7- to 8 fold lower

than the 73.8% we find for the new introns from SLR-RNA-Seq). Note that these 4 million random introns contain almost all of the new introns that our study finds.

3. The above method produces random introns that do not match the length and gene expression distributions of the new introns found in our study. To generate random introns that match both distributions, we rounded eaach new intron's length to the next kilobase and gene's expression (read number) to the next number divisible by 10. Expression of more than 1,000 reads and length $\geq$100kb were set to these cutoff values (1,000 reads and 100kb). We then counted for each bin (illustrated in the below table) how its new introns (found in this study). We then chose randomly (out of the 4 million random introns) as many introns for each bin as indicated by the count matrix. We repeated this process 20 times. This produced 20 sets of 137k random introns whose length and expression distributions closely match the (length and expression) distributions of the 137,000 new introns our study finds. In these 20 matched random sets, the fraction of introns that could be "validated" ranged from 23.8% to 24.1% - three fold lower than the 73.8% of novel introns. Note that these random sets have significant overlap with the new introns our study finds (which are among the random introns). This may inflate the validation rate, which we deal with in point 4. This inflation can be significant, because in some ($length, expression$)-bins many of the random introns are the new introns our study finds.

	$expression = 0$	$expression = 10$	...	$expression = 1000$
$length = 0$	$c_{0,0}$	$c_{0,10}$	...	$c_{0,1000}$
$length = 1000$	$c_{1000,0}$	$c_{1000,10}$	...	$c_{1000,1000}$
$length = 2000$	$c_{2000,0}$	$c_{2000,10}$	...	$c_{2000,1000}$
...				
$length = 100000$	$c_{100000,0}$	$c_{100000,10}$	...	$c_{100000,1000}$

4. To avoid artificial inflation of the validation rate by new introns from our study (confirmed at 73.8%), we again chose 20 random sets of random introns as in the previous paragraph. This time we ensured that none of the 137k introns our study finds were chosen. The "validation" rate now ranged from 16.2% to 16.5% (four to five-fold less than the 73.8% for the new introns of this study). Both the previous and this paragraph show that the random probability of getting a validation rate of 25% or higher is less than 0.05.

4 Annotation complexity

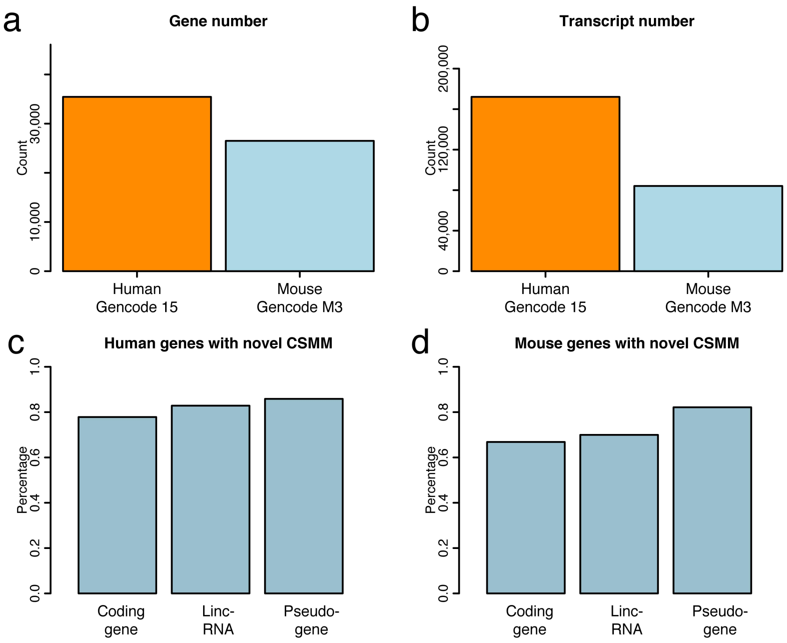

Figure S4: Annotation complexity and novel splicing isoforms. (a) Number of spliced genes in the human Gencode v15 annotation as well as the mouse Gencode M3 annotation. **(b)** Number of transcripts belonging to spliced genes in the human Gencode v15 annotation as well as the mouse Gencode M3 annotation. **(c)** Number of human spliced genes (belonging to the Gencode-defined gene categories of Coding Gene, LincRNA and Pseudogene that show at least one novel isoform. **(d)** Number of mouse spliced genes (belonging to the Gencode-defined gene categories of Coding Gene, LincRNA and Pseudogene that show at least one novel isoform.

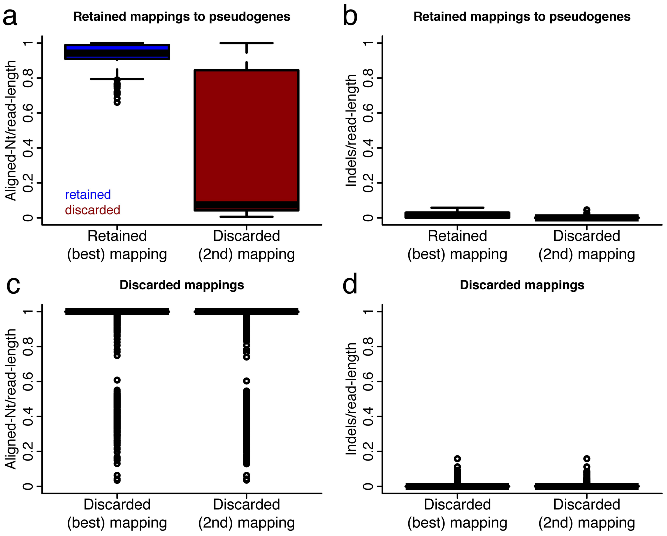

Figure S5: Quality of mapping to pseudogenes (a) Fraction of a read's nucleotides that is aligned to a genomic nucleotide for primary retained mappings for pseudogenes and for secondary (discarded) mappings. **(b)** Number of deletions and insertions per nucleotide of read sequence for mappings to pseudogenes and for secondary (discarded) mappings. **(c)** Same measure as in (a) for reads discarded, because primary and secondary mapping were not deemed sufficiently different. **(d)** Same measure as in (b) for reads discarded, because primary and secondary mapping were not deemed sufficiently different.

5 Uniqueness of mappings to pseudogenes

Mapping reads to pseudogenes can present considerable challenges when the pseudogene has only few nucleotide changes that distinguish it from its parent gene. In this work we only considered mappings to pseudogenes (and to any other location in the genome), if they were clearly better than any secondary mapping. Figure S5 shows that for the retained mappings to pseudogenes, there are easily observable differences in the quality of the primary mapping and any secondary mapping. Mappings that were discarded because of the existence of a secondary mapping did not show such easily observable differences.

6 Statistical considerations

In this manuscript, we performed 1,536 parallel (that is barcoded) sequencing experiments (sequenced in 4 lanes). 1,536 times we select roughly 1,000 molecules randomly out of a pool of billions of cDNA-molecules. These 1,000 molecules are then amplified in a separate well, fragmented and barcoded, so that later on "Synthetic Long Reads" (SLRs) or "contigs" can be assembled from these 1,000 molecules only - that is without interference of other molecules. Although we have taken further precautions, in theory we cannot rule out that two non-identical isoforms from the same gene end up in the same well. Hypothetically this situation (a **collision**) could lead to the assembly of a hybrid false positive SLR. We denote with **n** the overall number of transcripts for a fixed gene and with **N** the number of non-identical transcripts for this gene in a well. We are interested in the probability of a collision, conditioned to that the gene has been detected in the well - which we denote by

$$P(N \geq 2|N \geq 1)$$

We can easily determine the probability that for the gene in question at least one of its transcripts is present in the well, by determining the fraction of wells, in which the gene was detected. This experimentally determined probability can be written as

$$P(N \geq 1)$$

However, we cannot proceed this way for $P(N \geq 2|N \geq 1)$, as doing so would assume that contig-assembly can detect collisions - which we precisely want to test. Also, we cannot assume to know n - the number of overall existing transcripts for a gene. We will however derive an upper bound ($\frac{P(N\geq1)+(1-P(N\geq1))*ln(1-P(N\geq1))}{P(N\geq1)}$) on $P(N \geq 2|N \geq 1)$, that is calculated knowing only $P(N \geq 1)$. We proceed by

1. Motivating why we assume $4 * 10^{10}$ as the number of sscDNA molecules in 100ng of sscDNA
2. Giving the probability for a transcript to be among 1,000 randomly chosen molecules (out of $4 * 10^{10}$)
3. Showing that for most transcript-pairs such probabilities (although technically dependent) are very close to independence.

4. Based on the assumption of independence (see above), deriving an upper bound on $P(N \geq 2|N \geq 1)$.
5. Comparing the (above) upper bound (for each gene) to the fraction of molecules (for this gene), that are novel with respect to the annotation, and showing that for most genes, we have more novel isoforms than could be explained by putative collisions (and resulting hybrid and false positive SLRs)

6.1 Number of ssDNA molecules in 100ng

Assuming an average RNA-molecule length of 2kb, we estimate the number of molecules in 100ng of sscDNA as

$$\frac{10^{-7}g}{2000nt * 303\frac{g}{mol*nt}} * 6.022 * 10^{23}\frac{molecules}{mol} \quad = \quad 9.9 * 10^{10} molecules$$

To be on the safe side (with errors in our quantifications), we will lower this number by 60% to $4 * 10^{10}$.

6.2 Probability of transcript-occurrence among 1,000 random molecules

Let's consider a gene g with n transcripts $t_1, t_2, ..., t_n$ with associated molecule counts $c_1, c_2, ..., c_n$ among the total $4 * 10^{10}$ molecules. Denoting with T_i the number of molecules corresponding to transcript t_i that are among the 1,000 randomly chosen molecules (from the total of $4 * 10^{10}$), we can further define $p_i = P(T_i \geq 1)$. This probability can be written as $p_i := P(T_i >= 1) = 1 - \frac{\binom{c_i}{0} * \binom{4*10^{10} - c_i}{1000}}{\binom{4*10^{10}}{1000}} = 1 - \frac{\binom{4*10^{10} - c_i}{1000}}{\binom{4*10^{10}}{1000}}$

6.3 Quasi-independence of p_i and p_j

Two of these transcript probabilities (p_1 and p_2) are strictly speaking not independent. The following considerations show however that they are very close to independence as long as the two associated molecules c_1 and c_2 are relatively small. We will now consider two transcripts t_1 and t_2 and the associated probabilities of NOT being among the 1,000 randomly chosen molecules: $P(T_i = 0) = 1 - p_i = \frac{\binom{4*10^{10} - c_i}{1000}}{\binom{4*10^{10}}{1000}}$

In the strict sense, the two probabilities ($P(T_1 = 0)$ and $P(T_2 = 0)$) are independent if and only if $P(T_2 = 0|T_1 = 0) = P(T_2 = 0)$ or in other words $\frac{P(T_2=0|T_1=0)}{P(T_2=0)} = 1$. In our case, we have

$$P(T_2 = 0|T_1 = 0) \quad = \quad \frac{P(T_2 = 0 \cap T_1 = 0)}{P(T_1 = 0)} = \frac{\binom{c_1+c_2}{0} * \binom{4*10^{10} - c_1 - c_2}{1000}}{\binom{4*10^{10}}{1000}} * \frac{\binom{4*10^{10}}{1000}}{\binom{4*10^{10} - c_1}{1000}} = \frac{\binom{4*10^{10} - c_1 - c_2}{1000}}{\binom{4*10^{10} - c_1}{1000}} \tag{1}$$

and therefore using this last equation 1, we have

$$\begin{aligned} \frac{P(T_2 = 0|T_1 = 0)}{P(T_2 = 0)} &= \frac{\binom{4*10^{10} - c_1 - c_2}{1000}}{\binom{4*10^{10} - c_1}{1000}} * \frac{\binom{4*10^{10}}{1000}}{\binom{4*10^{10} - c_2}{1000}} = \frac{\prod_{i=0}^{999} \frac{4*10^{10} - (c_1+c_2) - i}{i+1}}{\prod_{i=0}^{999} \frac{4*10^{10} - c_1 - i}{i+1}} * \frac{\prod_{i=0}^{999} \frac{4*10^{10} - i}{i+1}}{\prod_{i=0}^{999} \frac{4*10^{10} - c_2 - i}{i+1}} \\ &= \\ &= \prod_{i=0}^{999} \frac{4 * 10^{10} - (c_1 + c_2) - i}{4 * 10^{10} - c_1 - i} * \frac{4 * 10^{10} - i}{4 * 10^{10} - c_2 - i} \end{aligned} \tag{2}$$

To illustrate the "closeness to independence", we explored all values $(c_1, c_2) \in \{4 * 10^0, 4 * 10^1, 4 * 10^2, 4 * 10^3, 4 * 10^4, 4 * 10^5, 4 * 10^6\}^2$ (with $4 * 10^6$ equating to one in 10,000 molecules of the total $4 * 10^{10}$), to calculate formula 2. Figure S5a shows that for all these (c_1, c_2), $\frac{P(T_2=0|T_1=0)}{P(T_2=0)}$ is almost equal to 1. In fact, we find $\frac{P(T_2=0|T_1=0)}{P(T_2=0)} \in [0.99999, 1]$. When including values of $4 * 10^7$ and $4 * 10^8$, $\frac{P(T_2=0|T_1=0)}{P(T_2=0)}$ starts to deviate from 1 (Figure S5b). Thus for very highly expressed genes (such as GAPDH for example), assuming independence is not a good idea. Note, that there can be at most 1,000 transcripts with $4 * 10^7$ or more molecules (because the total number of molecules is $4 * 10^{10}$). In fact there will be much fewer, because each lowly expressed transcript and very highly expressed transcript lowers the total number of molecules that can be distributed onto transcripts with counts $\geq 4 * 10^7$. Given that we expect at the very least 200,000 transcripts in a complex tissue such as brain, we conclude that for the vast majority of transcripts, independence is a reasonable assumption.

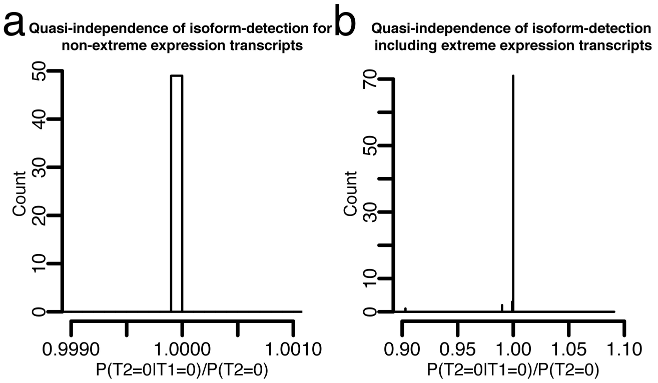

Figure S6: Exploration of quasi-independence. **(a)** Using hypothetical transcripts with counts of molecules of up to 4,000,000. **(b)** Using counts of molecules of up to 400,000,000.

6.4 Number of different transcripts of a gene among 1,000 random molecules

For a gene with n transcripts, we denote with N the number of transcripts, for which at least one molecule is among the 1,000 molecules. Assuming independence of $p_1, p_2, .., p_n$ (see previous paragraph) we write

$$\begin{aligned}
P(N=0) &= P(T_i = 0 \cap T_2 = 0 \cap .. \cap T_n = 0) = \prod_{i=1}^{n}(1-p_i) \\
P(N \geq 1) &= 1 - P(N=0) = 1 - \prod_{i=1}^{n}(1-p_i) \\
P(N=1) &= \sum_{i=1}^{n} p_i \prod_{j \neq i}^{n}(1-p_j)
\end{aligned}$$

The collision probability (that is for a detected gene the probability that two or more non-identical isoforms are chosen when choosing randomly 1,000 molecules) can now be written as

$$P(N \geq 2 | N \geq 1) = \frac{P(N \geq 2 \cap N \geq 1)}{P(N \geq 1)} = \frac{P(N \geq 2)}{P(N \geq 1)} = \frac{P(N \geq 1) - P(N=1)}{P(N \geq 1)} \tag{3}$$

All the above formulas depend heavily on n, a number that we don't know (and that we cannot take from known annotations, because of the large number of novel transcripts that we find).

Lemma 6.1 $\forall n$: $P(N \geq 2 | N \geq 1) \leq \frac{P(N \geq 1) + (1 - P(N \geq 1)) * ln(1 - P(N \geq 1))}{P(N \geq 1)}$
This lemma holds true, irrespective of n and only depends on $P(N \geq 1)$ - a value that we can easily observe (as the fraction of wells, in which the gene was detected). The below lines arise from basic mathematics - we add it here to support our work, not because we believe to be the first authors to ever do this calculation.

Proof Because of equation 3, the proposition becomes

$$\begin{aligned}
\frac{\cancel{P(N \geq 1)} - P(N=1)}{\cancel{P(N \geq 1)}} &\leq \frac{\cancel{P(N \geq 1)} + (1 - P(N \geq 1)) * ln(1 - P(N \geq 1))}{\cancel{P(N \geq 1)}} \\
\iff -\sum_{i=1}^{n} p_i \prod_{j \neq i}^{n}(1-p_j) &\leq \prod_{i=1}^{n}(1-p_i) * ln(\prod_{i=1}^{n}(1-p_i)) \\
\iff -\sum_{i=1}^{n} \frac{p_i}{1-p_i} &\leq ln(\prod_{i=1}^{n}(1-p_i)) = \sum_{i=1}^{n} ln(1-p_i) \\
\iff \sum_{i=1}^{n}(ln(1-p_i) + \frac{p_i}{1-p_i}) &\geq 0 \qquad (4)
\end{aligned}$$

Inequality 4 is true because $ln(1-p) + \frac{p}{1-p} \geq 0$. With $q = 1-p$, we write $f(q) = ln(q) + \frac{1-q}{q} = ln(q) - 1 + \frac{1}{q}$, which is positive for any $q \in (0,1]$, as the following shows: Let's consider q=1: Here we have $f(1) = ln(1) - 1 + \frac{1}{1} = 0 - 1 + 1 = 0$. Let's consider f': $f'(q) = \frac{1}{q} - \frac{1}{q^2}$ Since $\forall q \in (0,1) : q^2 < q \Rightarrow f'(q) < 0 \Rightarrow f$ is monotonously decreasing on $(0,1)$ and therefore ≥ 0 on the same interval.

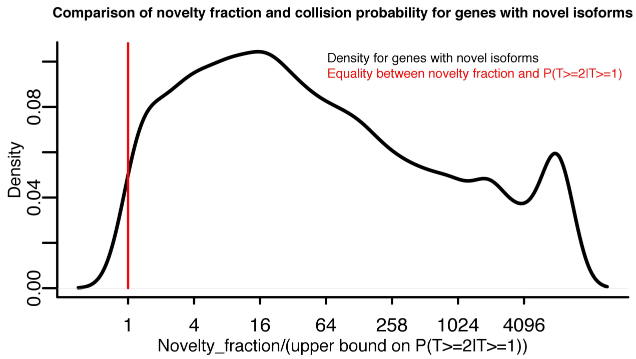

Figure S7: Gene-wise comparison of the fraction of novel CSMM (among all CSMM for this gene) and the upper bound on the collision probability calculated earlier.

6.5 Alterntative proof using Poisson-probabilties

The above lemma can also be derived, assuming that the number of a specific transcript of a gene presented in a given well follows a Poisson distribution with mean parameter λ, which is generally less than 0.1. In brief, we can state: $-P(N = 1) = -\lambda e^{-\lambda} = ln(e^{-\lambda}) * e^{-\lambda} = ln(1 - P(N \geq 1)) * (1 - P(N \geq 1))$. With this we can write $P(N \geq 2 | N \geq 1) = \frac{P(N\geq2)}{P(N\geq1)} = \frac{P(N\geq1)-P(N=1)}{P(N\geq1)} = \frac{P(N\geq1)+ln(1-P(N\geq1))*(1-P(N\geq1))}{P(N\geq1)}$

6.6 Comparison of novel isoforms and collision probability

Here we consider the possibility that novel isoforms are merely a consequence of having two non-identical isoforms from the same gene within the same well (which could possibly lead to a hybrid assembly of a false positive new isoform). To this end, we calculated for each gene the collision probability as well as the novelty rate (that is the fraction of molecules for this gene that describes a new isoform).

Figure S7 shows that for the majority of genes, the fraction of novel CSMM exceeds strongly the upper bound of the collision probability. 84% of all genes have at least twice as many novel CSMM as predicted by the upper bound on the collision probability.

A further line of thought indicating that the collisions (that is two non-identical isoforms from the same gene within the same well) do not cause an important number of false positive novel isoforms is the following: For the 3,042 genes without a single novel CSMM, one would expect 2920 false positive CSMM - none of which is observed.

7 Validation of dMAPs through PacBio-sequencing of PCR products

We verify our observations of distant AS events using an orthogonal sequencing approach, where we amplify specific genes and sequence them on the PacBio platform. We designed primers for two of the human genes that exhibit dMAPs, BIN1 (Forward, 5-CCAGAAGCTGGGGAAGGCAG, Reverse, 5-CAGCACCACATCACCAGCCT) and CAPN7 (Forward, 5-ACCCTCAAAACAAGGATGGTGA, reverse, 5-TCCAGGTAGCAAAACCCACACC). Primers where placed in flanking constitutively expressed exons and targeted length of PCR products were >1kb in order to avoid preferential sequencing of the shorter isoforms. Total RNA from human brain was converted to sscDNA using an oligo-dT primer following manufacturers suggested protocol (LifeTech, CA, Invitrogen, CatNo: 18080-051). The sscDNA was amplified with the gene specific primers using a Kapa HotStart Readymix (Kapa Biosystems, KK2601), cycling conditions were: 98C 30 sec, 35 cycles of 98C 10 sec, 58C 30 sec, 72C 90sec, followed by a final extention of 2 min at 72C. The amplified product was converted to a PacBio SMRT-library following manufacturers guidelines for 2kb products without any shearing of the PCR product (Pacific Biosystems, CA, PN 001-143-835-06). Libraries were sequencing on a PacBio RSII, using C2/P4 chemistry, according to manufacturers guidelines.

The six cases (BIN1, CAPN7, ABCD4, EXOC7, MAP4 and NRCAM) were chosen after sequencing four lanes of SLRSeq (the first four of the final eleven lanes we use in the paper). They are all positively associated - that is they have a significant one-tailed p-value, with the alternative hypothesis being apositive association between the two exons. We considered an event validated if the one-tailed test (in the same direction) based on the PCR-result was also significant. The below table

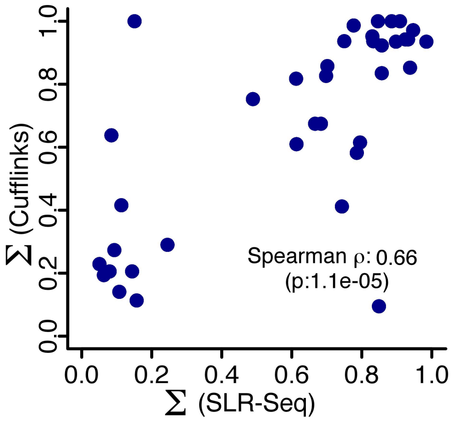

Figure S8: Using STAR and Cufflinks to predict Σ values of dMAPs and dMEPs. Advantages of this approach include (a) genome-guided alignment of short-reads removes the need to continously assembly an entire fragment: a few not covered bases do not necessarily break up the gene structure (whereas they do break up the SLR) and (b) easy incorporation into existing pipelines.
The disadvantages of this approach include (a) requirement for both exons to be annotated and (b) Cufflinks is unaware that overlapping short-reads from one well are very likely to represent fragments of PCR-products from one molecule.

gives the p-values from our genome wide scan (after sequencing 4 lanes and before correction for multiple testing) and the p-value from the PCR result.

Table S2: Validation of six dMAPs

Gene	Genome-wide p-val	PCR p-val
BIN1	1.335e-09	< 2.2e-16
CAPN7	1.043e-12	< 2.2e-16
ABCD4	3.035e-05	3.027e-07
EXOC7	1.622e-09	< 2.2e-16
MAP4	4.742e-06	3.986e-15
NRCAM	4.906e-07	1.413e-09

8 Comparison of SLR-RNA-Seq and short-read-analysis for dMAPs and dMEPs

In order to explore if short read mappers (e.g. STAR) and transcript analysis tools (Cufflinks) would draw similar conclusions, we mapped the first four lanes of SLR-RNA-Seq short-reads (that is before assembly into SLRs) to the hg19 genome using STAR. We then employed Cufflinks to quantify the transcripts of the Gencode v15 annotation and added up the FPKMs for all transcripts corresponding to the "both-exons-included" isoform, the "exon1-included-exon2-excluded" isoform, the "exon1-excluded-exon2-included" isoform and the "both-exons-excluded"-isoform separately. Based on these four values we calculated a Cufflinks-Σ in strict analogy to the SLR-Seq-Σ. $STAR - parameters$, Samtools and Cufflinks parameters were identical to the ones employed above in section 2 . Figure S8 shows that overall the STAR-Cufflinks approach performed reasonably well in predicting the SLR-RNA-Seq Σ values.