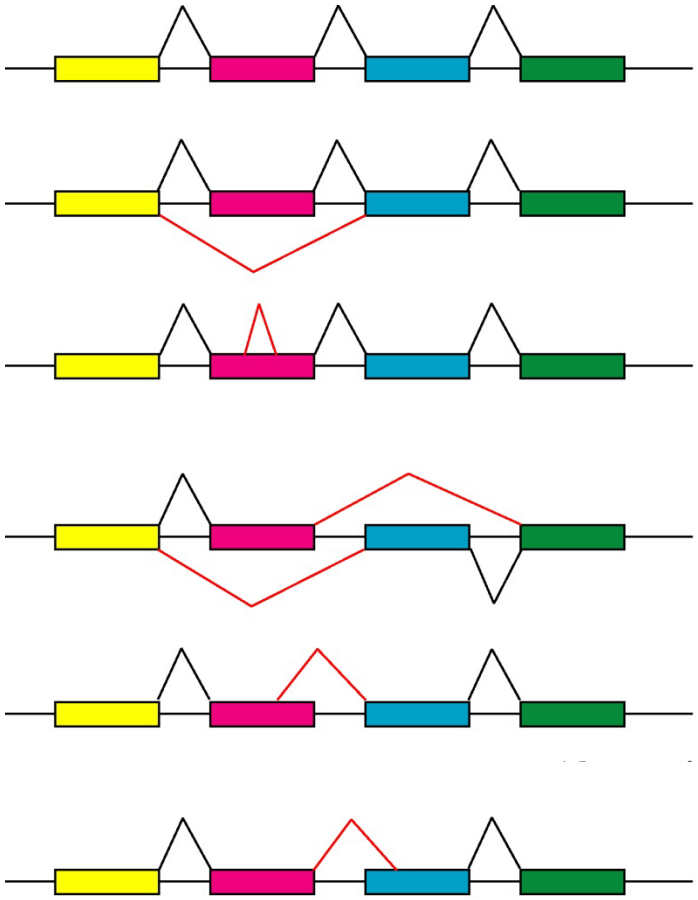
# Assessing long-distance RNA sequence connectivity



Technical Journal Club
07.07.2015
Rahel Gerosa

# Introduction

- Important driver of gene expression is the ability to produce multiple mRNA isoforms from a single gene

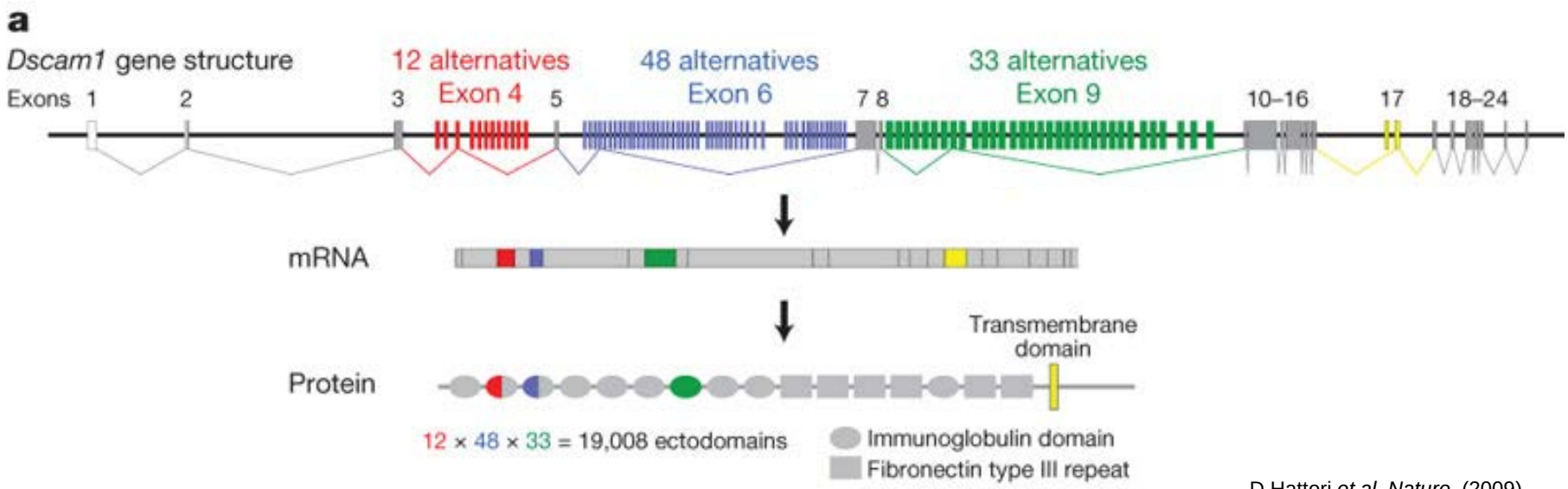
- Around 58% of Drosophila melanogaster genes and > 95% of human genes produce more than one transcript (most human genes expressing 10 or more distinct isoforms)

- Alternative promoter use, alternative splicing, and alternative polyadenylation contribute to isoform diversity

- The combination of these alternative processes increases the number of possible products exponentially → some human genes predicted to express > 100 mRNA isoforms

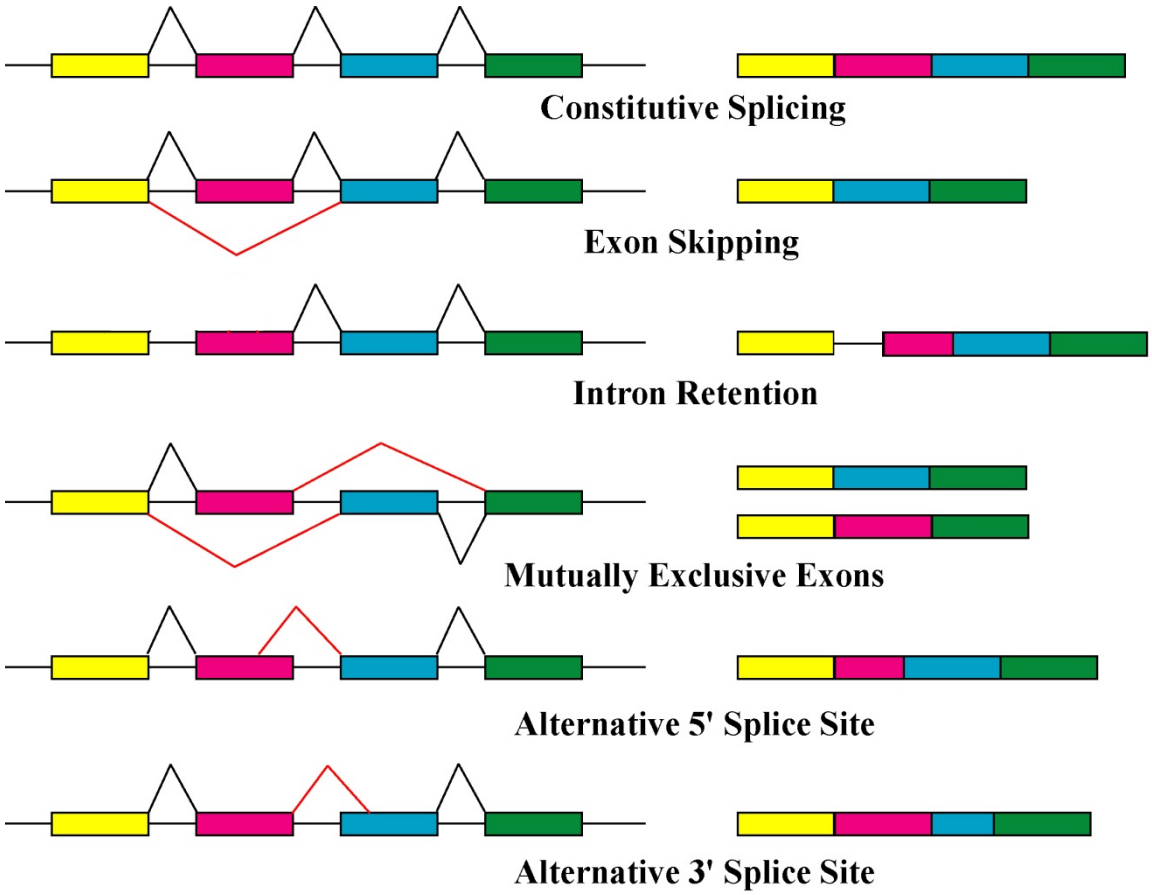
# Introduction

- Current record holder in number of isoforms is Down Syndrome Cell Adhesion Molecule (Dscam1) → 38,016 isoforms with > 7000 nt mRNAs)
- Dscam1 consists of four regions of mutually exclusive cassette exons
- These are separated by one to eight constitutive exons
- More than a quarter of human genes share this feature




D Hattori *et al. Nature*, (2009)

# Alternative Splicing




Adapted from
http://de.academic.ru/dic.nsf/dewiki/63657

# How investigate splice variants

- Understand the mechanisms underlying these complex isoform expression patterns → methods that maintain long-range exon connectivity information

- Research on isoform regulation has been hindered by transcript lengths

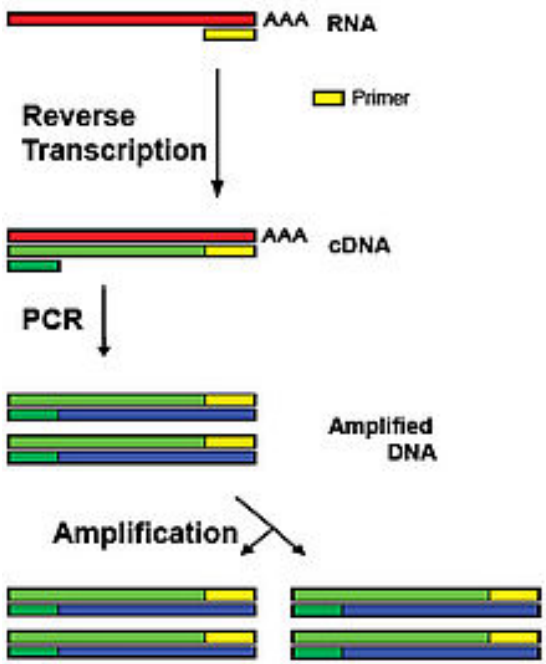
- *Dscam* transcripts span over 7 kilobases, for example, whereas short-read sequencing is limited to around 500 bases

- So far:
  - RT-PCR + Sequencing (Illumina, MiSeq)
  - Single molecule real time sequencing (SMRT)
  - CAMSeq

# Reverse transcription polymerase chain reaction (RT-PCR)

- Commonly used in molecular biology to qualitatively detect gene expression through creation of cDNA transcripts from RNA

- High-sequence similarity, combined with the long stretches of identical constitutive exons separating the distant alternative splicing regions, strongly favors template switching by RT



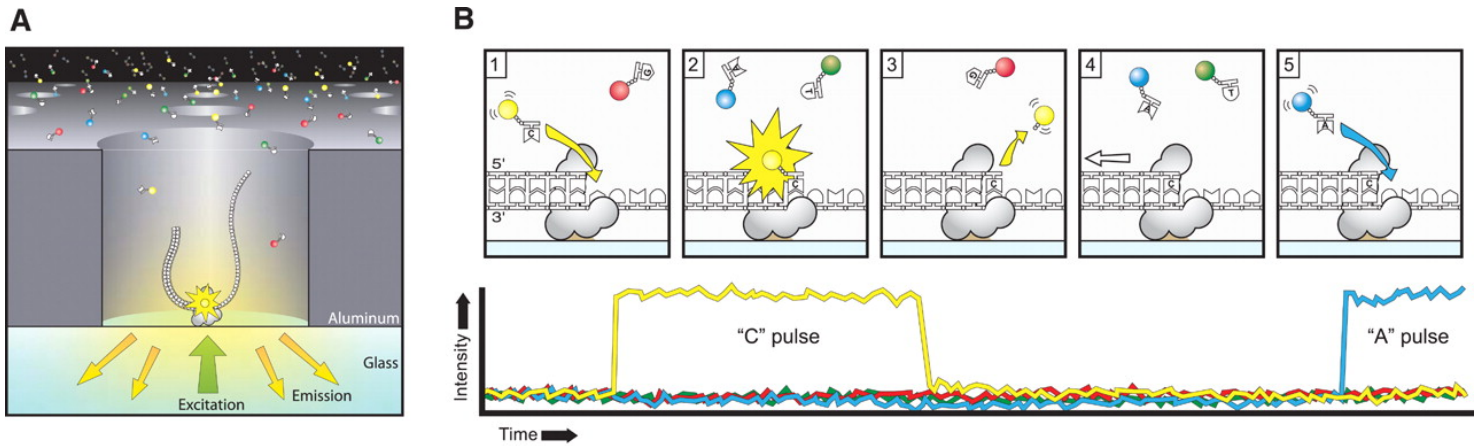

e.g. Illumina, MiSeq
Read length: 50 to 300 bp

https://en.wikipedia.org/wiki/Reverse_transcription_polymerase_chain_reaction

# Single molecule real time sequencing (SMRT)-PacBio

- A single molecule of DNA is immobilized and illuminated from below by laser light
- Enables detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase
- Single-molecule sequencing can read many kilobases of DNA (10,000 bp to 15,000 bp avg), but for RNA, it relies on reverse transcription, which is inefficient on long stretches

# Single molecule real time sequencing (SMRT)-PacBio



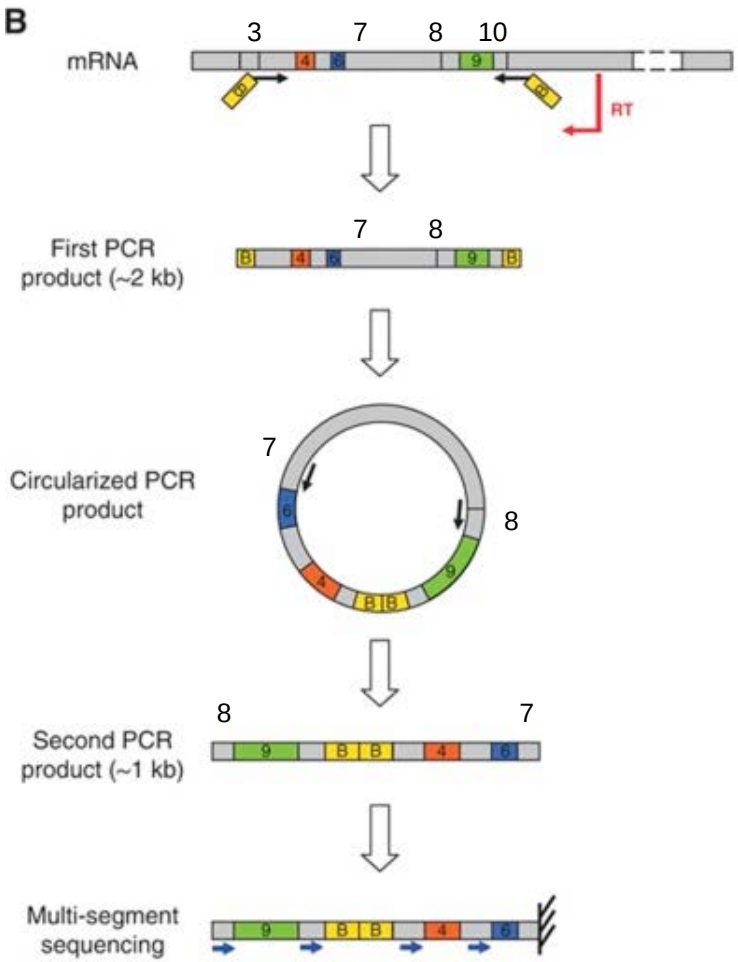

Eid et al., Science, 2009

(1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an increase of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse

# Circularization-Assisted Multi-Segment Sequencing (CAMSeq)

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

- Four specific sequencing primers targeting constitutive exon 3, 5, 8, and exon 10
- Obtained from every template DNA molecule four sequencing reads derived from exons 4, 6, 9, and barcode, respectively




Wei Sun et al., EMBO Journal, 2013

# Problems with RT-PCR, PacBio and CAMSeq

- Sequences are too long → not possible to sequence with current high-throughput technologies
- Full-length transcripts must be inferred by piecing together multiple short overlapping reads
- Many exon variants arose from exon-duplication events, so their sequences are highly similar
- Single molecule methods capable of longer reads (e.g., Pacific Biosciences) have limited read depths, making it difficult to fully analyze transcripts expressed over many orders of magnitude
- This high-sequence similarity, combined with the long stretches of identical constitutive exons separating the distant alternative splicing regions, strongly favors template switching by RT

- New technology is needed → SeqZip

eLIFE
elifesciences.org

April 2015

# Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation

Christian K Roy[1,2], Sara Olson[3], Brenton R Graveley[3], Phillip D Zamore[1,2]*, Melissa J Moore[1,2]*

Christian K. Roy

Melissa J Moore

interested in post-transcriptional gene regulation in eukaryotes via mechanisms involving RNA and RNA-protein (RNP) complexes

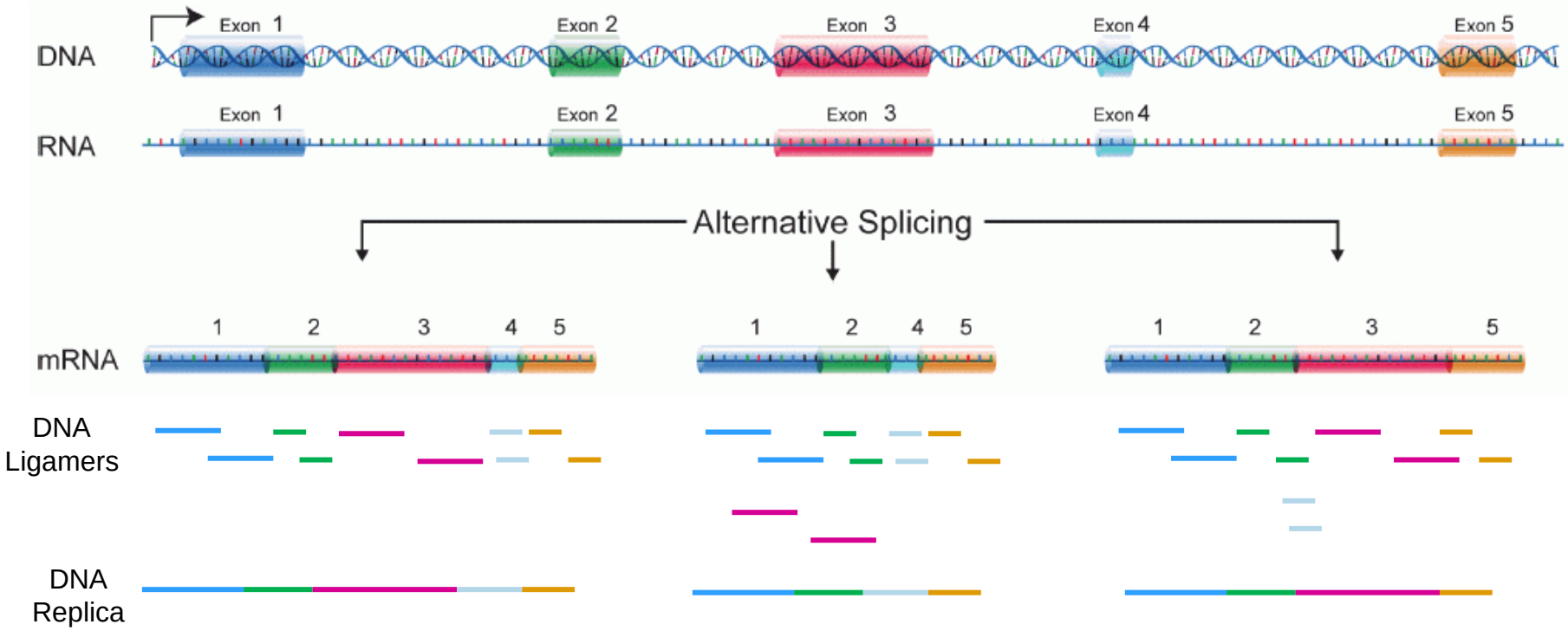
# Advantages of SeqZip

- To better probe exon connections in distant parts of the same molecule
- Has no RT step
- Eliminates long intervening regions of common sequence
- Unique exon-specific barcodes introduced during the ligation step further discourage template switching during subsequent amplification

# Principles of SeqZip

A reverse transcription-free method to assess sequence connectivity

- Uses sets of DNA oligonucleotides termed ‘ligamers’ (~ 40–60 nt) that are hybridized to the 5 ′ and 3 ′ ends of a single alternatively spliced exon or the beginning and end of a large block of constitutively included exons, looping out the sequences in between (loops can be hundreds to thousands of nucleotides long)




Adapted from https://en.wikipedia.org/wiki/Alternative_splicing

# Principles of SeqZip

A reverse transcription-free method to assess sequence connectivity

- Uses sets of DNA oligonucleotides termed ‘ligamers’ (~ 40–60 nt) that are hybridized to the 5′ and 3′ ends of a single alternatively spliced exon or the beginning and end of a large block of constitutively included exons, looping out the sequences in between (loops can be hundreds to thousands of nucleotides long)

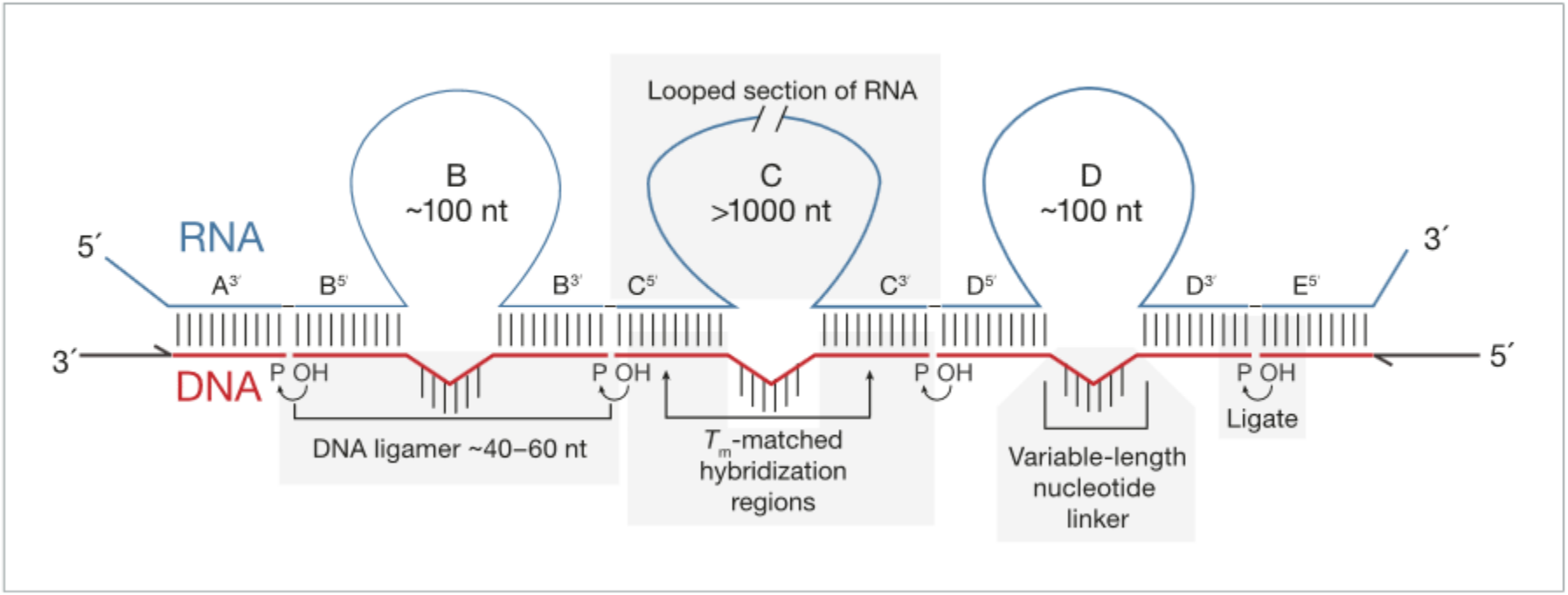
- Juxtaposed ligamers hybridized to single RNA molecules are then joined by enzymatic ligation with T4 RNA ligase 2 (Rnl2) to make a DNA replica of the spliced RNA

- The resultant DNA ligation products
  - capture the intramolecular connectivity among exons of interest
  - compress the sequence space necessary to identify those exons

- Exon connectivity is subsequently decoded by assessing the sizes or sequences of the ligation products

# Principles of SeqZip

A reverse transcription-free method to assess sequence connectivity



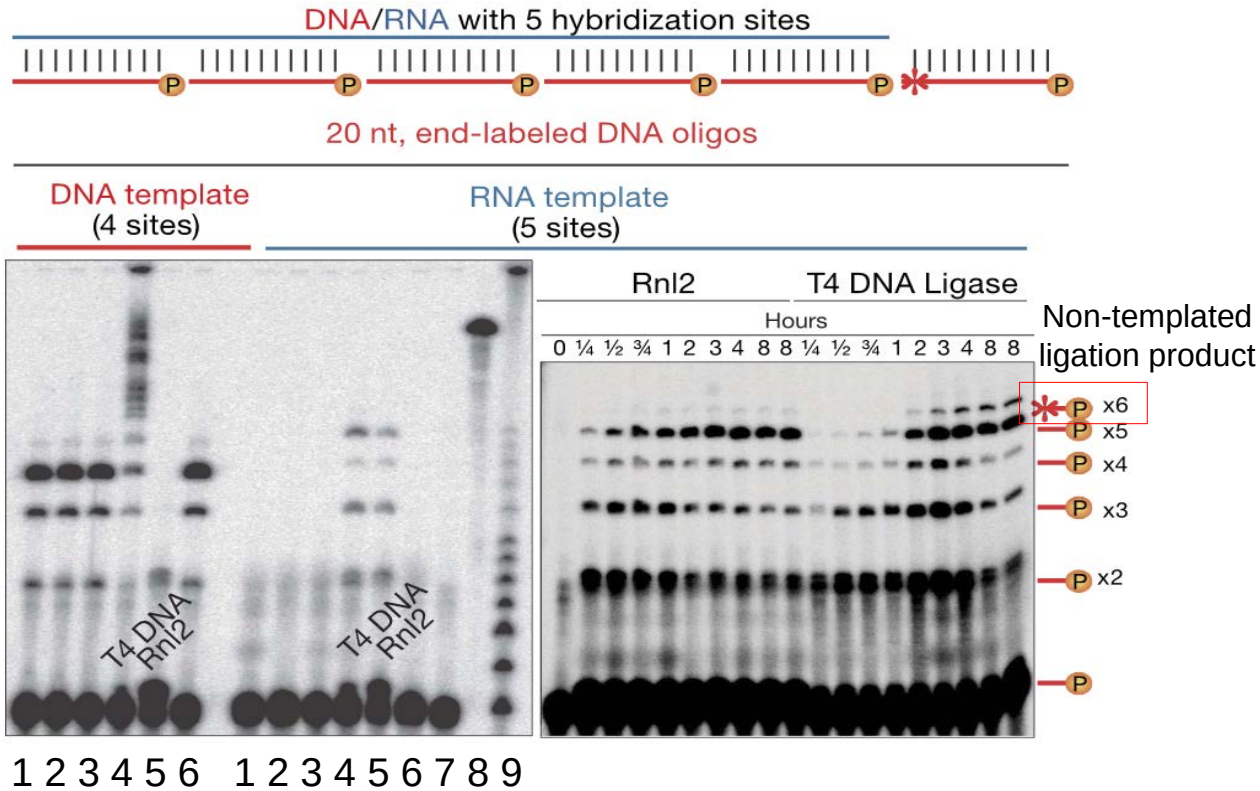

- Requires efficient ligation of multiple DNA oligos (legamers 40-60 nt) hybridized to an RNA template with little or no non-templated ligation (looping out the sequences in between)

# RNA-templated DNA-to-DNA ligation

- Only T4 DNA ligase joined DNA fragments templated by RNA
- It also catalyzes non-templated DNA ligation which would reduce SeqZip fidelity
- Tested several other commercially available enzymes to ligate four or five 5′ 32 P-radiolabeled 20-nt DNA oligos hybridized to adjacent positions on either DNA or RNA
- Only T4 DNA ligase and RNA ligase 2 (Rnl2) joined the DNA oligos when hybridized to the RNA template
- Rnl2 could not ligate DNA oligos hybridized to the DNA template, eliminating the possibility of contaminating genomic DNA confounding SeqZip

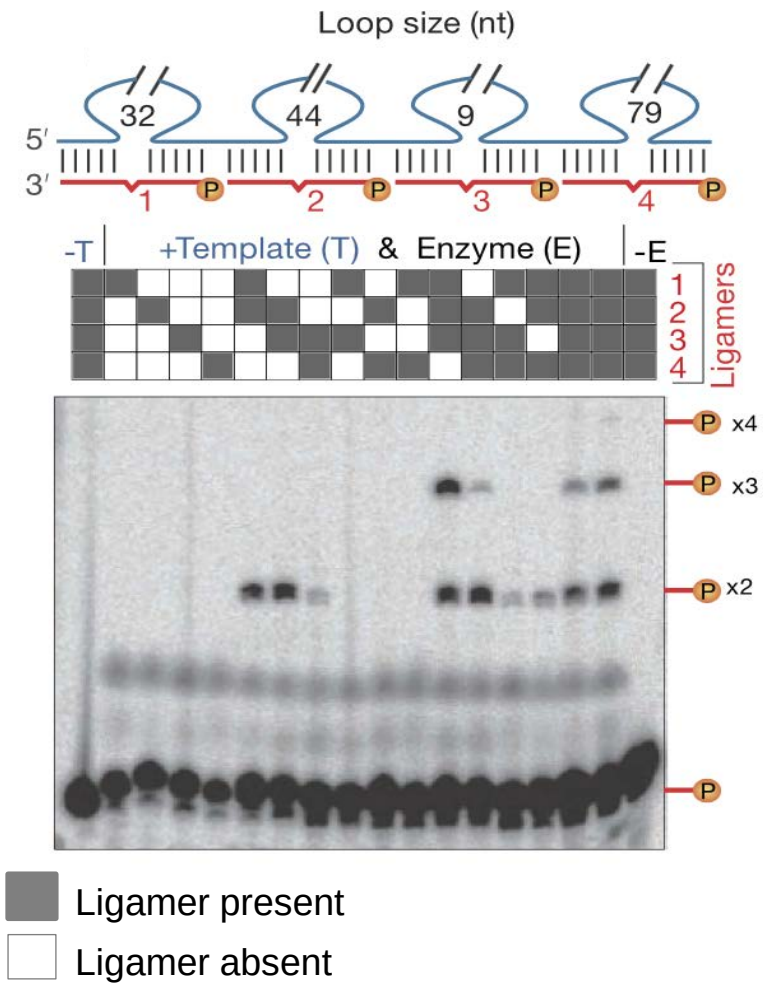
1. Tth DNA ligase
2. Tsc DNA ligase
3. Thermostable DNA ligase
4. T4 DNA ligase
5. T4 Rnl2
6. E. coli DNA ligase

7. 32P-oligos only
8. 32P-labeled RNA
9. 32P-labeled DNA ladder



# RNA-templated DNA-to-DNA ligation

- Test the ability of Rnl2 to ligate ligamers designed to loop out various lengths of a template RNA
- Rnl2 can join multiple 32 P-labeled ligamers each looping out sections of the template but only when they are adjacently hybridized



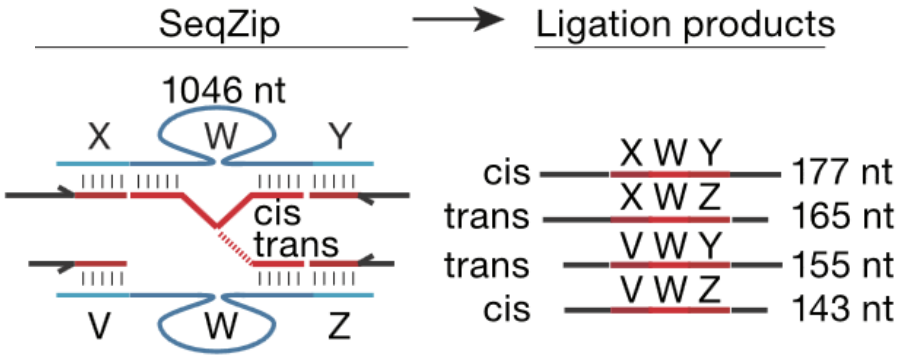

307 nt transcript → 244 nt condensed to 104 nt DNA sequence

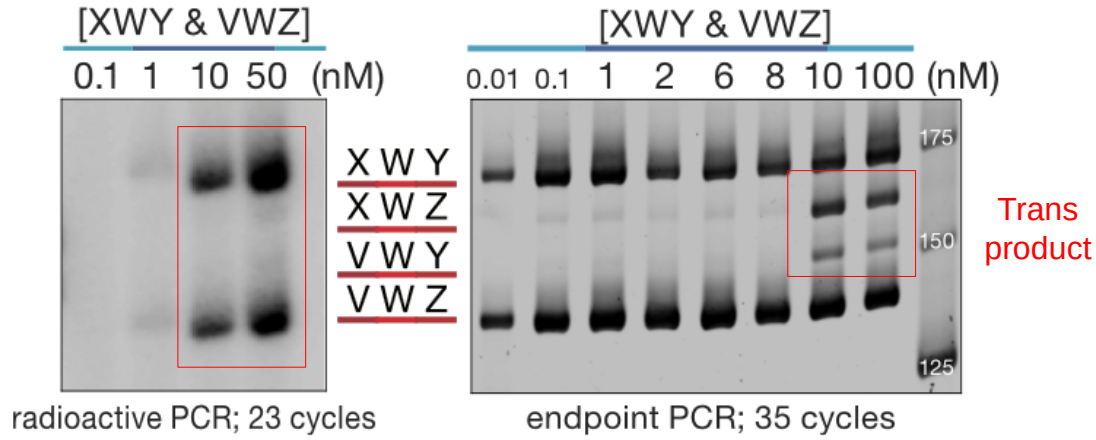
# Cis- and trans-transcript hybridization

- A ligamer designed to loop out the sequences in between widely spaced regions of complementarity has the potential to bridge two RNA molecules
- Such intermolecular (trans) hybridization would interfere with measurement of intramolecular (cis) RNA connectivity, producing artifacts similar to template switching in RT-based methods

- Test the frequency of such trans events → mixed two RNAs, each comprising a common 1106 nt sequence
- Ligamer set in which a single internal ligamer (W) looped out 1046 nt of the shared internal sequence

- To disfavor trans hybridization, the general conditions for SeqZip use cellular RNA concentrations at which mRNAs are present at <1 nM

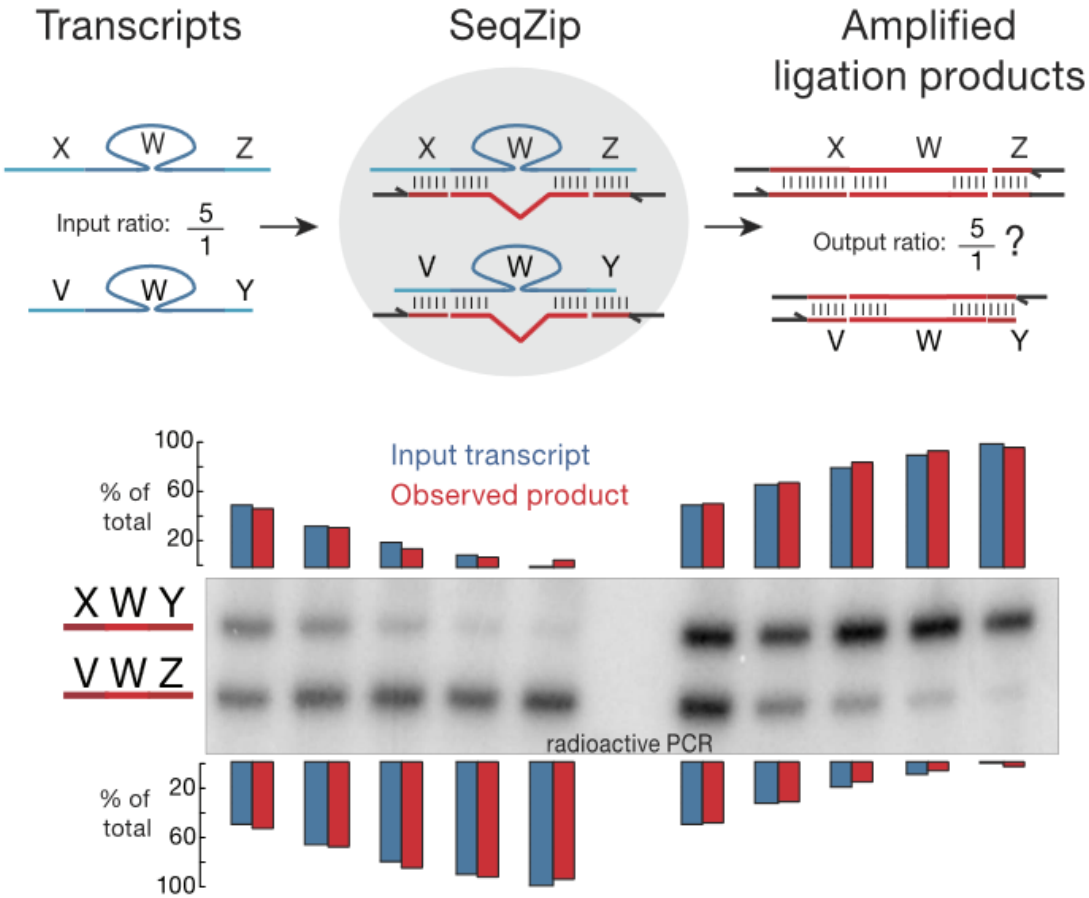



Cis hybridization products predominated

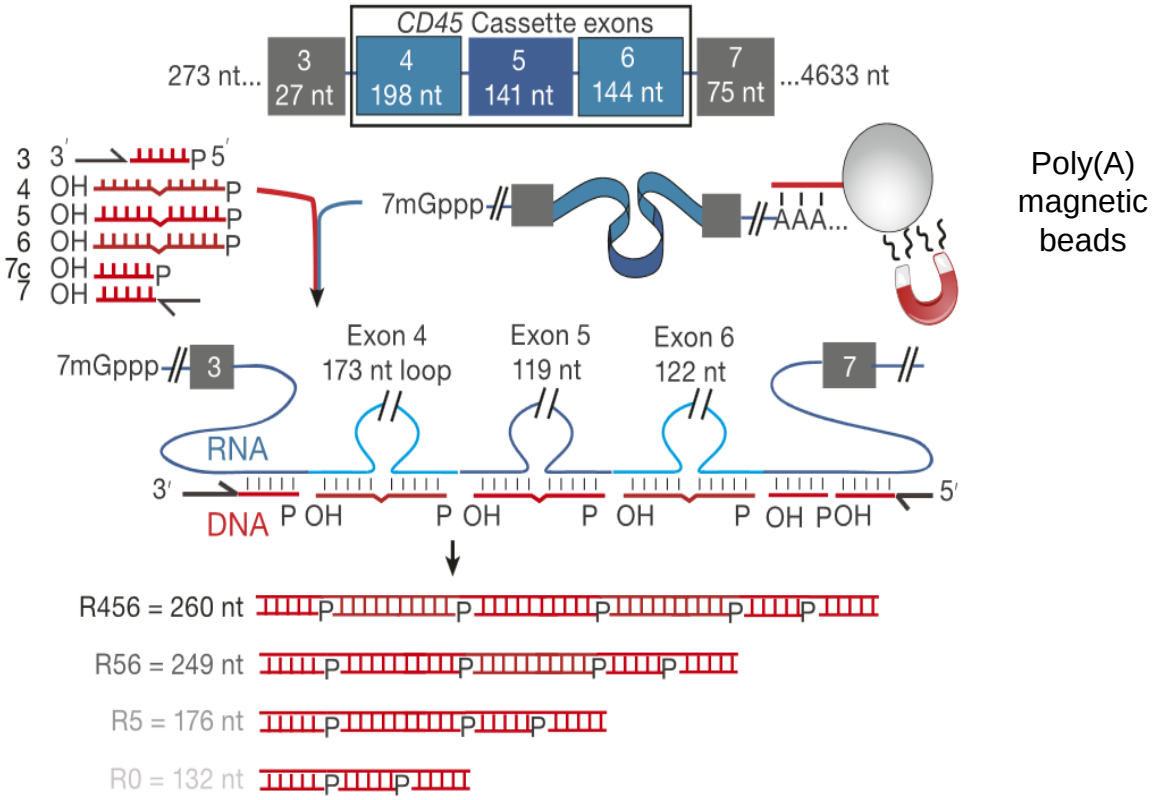
# Accurately report on relative input RNA concentrations

- To be useful as a quantitative method, SeqZip should accurately report on input RNA abundances
- Mixed two target RNAs at ratios varying from 100:1 to 1:100
- Radioactive PCR revealed that their respective SeqZip product ratios paralleled these input ratios over the entire series



# Measure endogenous mRNA isoform expression

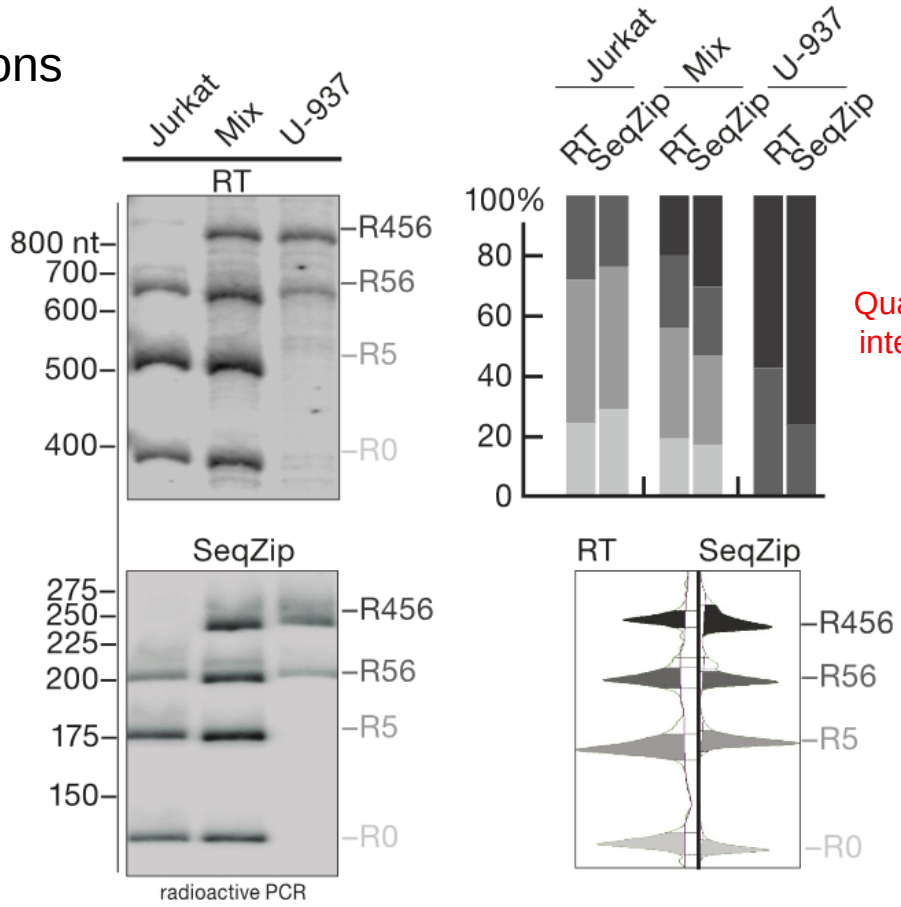
- Assess alternative exon inclusion in endogenous human CD45 (PTPRC) mRNA
- Various combinations of exons 4, 5 and 6
  - R456 = isoform containing exons 4, 5 and 6
  - R56 = isoform containing exons 5 and 6
  - R5 = isoform containing exons 5
  - R0 = isoform containing neither exons 4, 5 or 6



# Measure endogenous mRNA isoform expression

- Jurkat cells (resembling naive, primary T cells) predominantly express isoforms R56, R5, or R0
- U-937 cells (resembling activated T cells) predominantly express the R56 and the R456 isoform

- The three adjacent cassette exons occupy only 585 nt, making this region amenable to analysis by both reverse transcription and SeqZip
- Both methods reported the expected isoform abundances




Quantified band intensities from gels

Mirrored lane profiles from the mix lanes

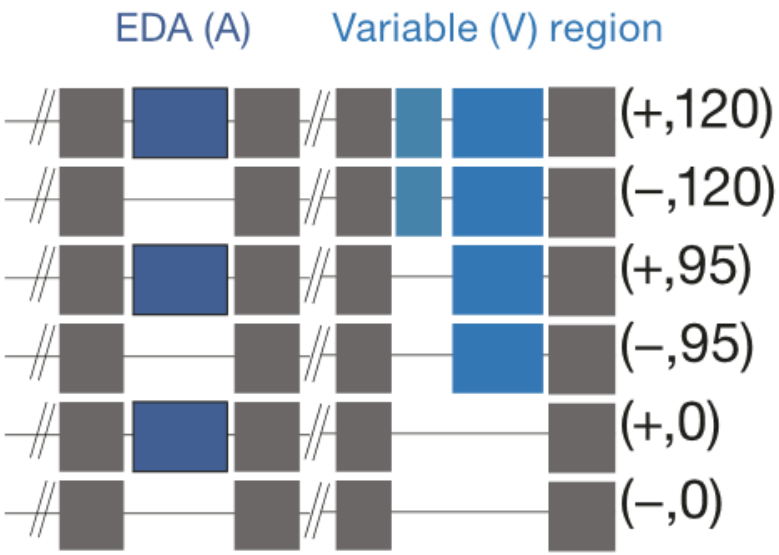
# Fibronectin

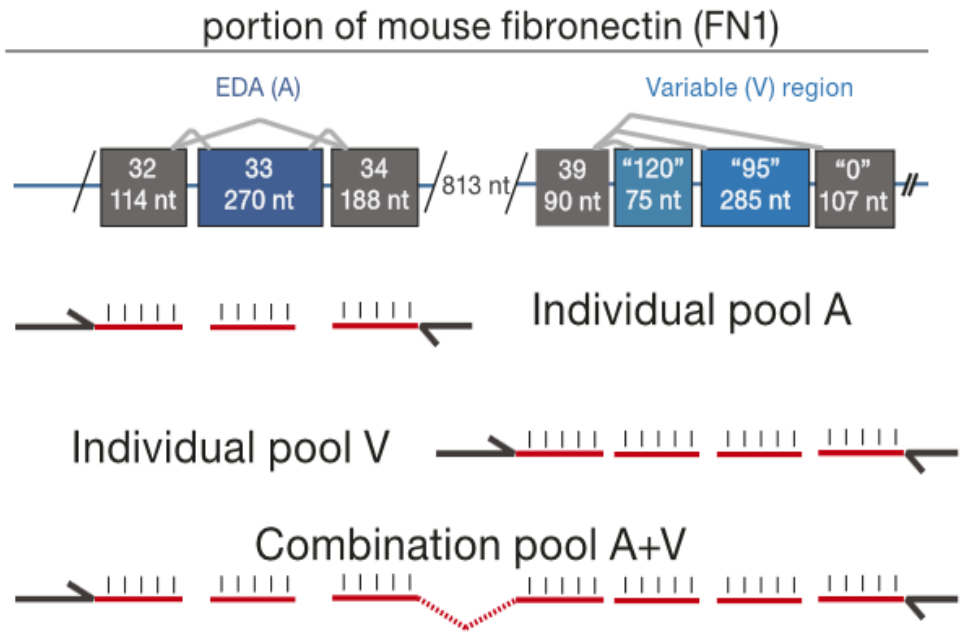
- For a more complex splicing pattern → fibronectin
- Glycoprotein of the extracellular matrix that binds to membrane-spanning receptor proteins called integrins
- Binds extracellular matrix components such as collagen, fibrin, and heparan sulfate proteoglycans
- The original suggestion that an upstream splicing decision can affect a downstream splicing decision came from analysis of the EDA and V regions of fibronectin

# Fibronectin

Three well-characterized regions of alternative splicing:

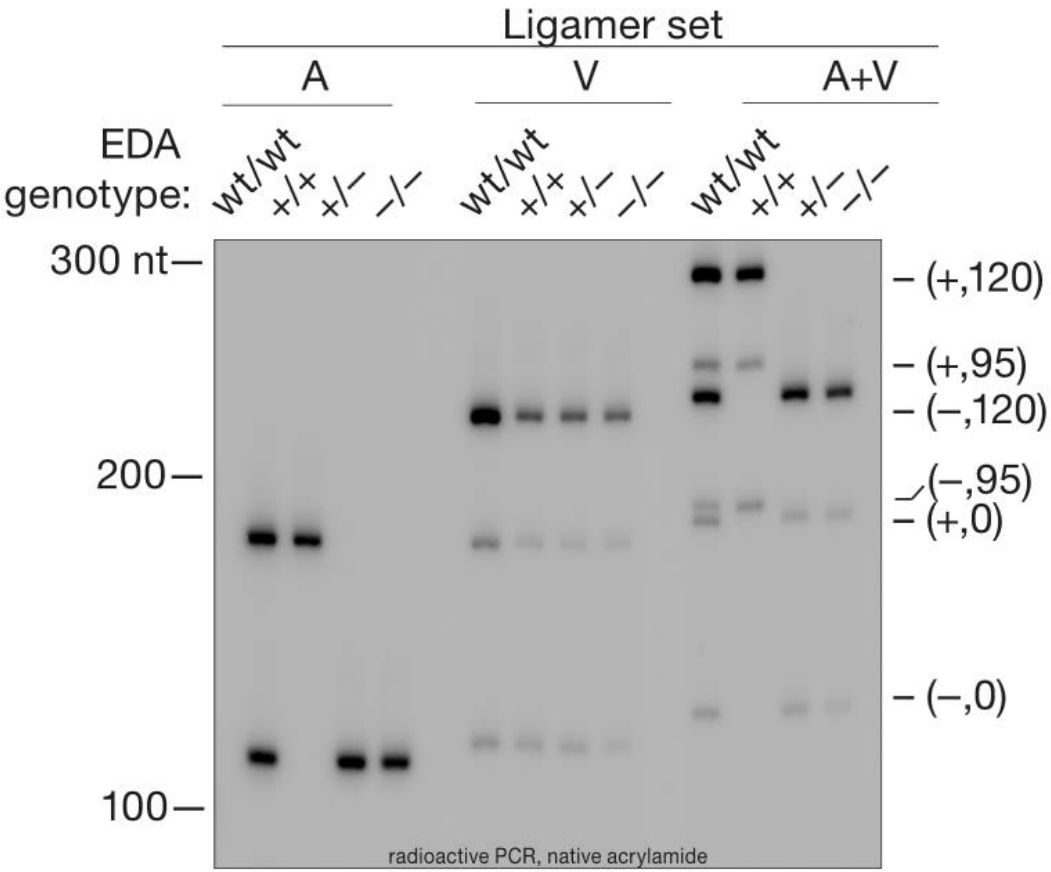
1. the EDB exon included in embryos and adult brain but not other adult tissues
2. the EDA exon variably <u>included or excluded</u> across multiple developmental and adult tissue types
3. the variable (V) region in which use of three alternative 3′ splice sites leads to inclusion of 120, 95, or 0 additional amino acids in the FN1 protein



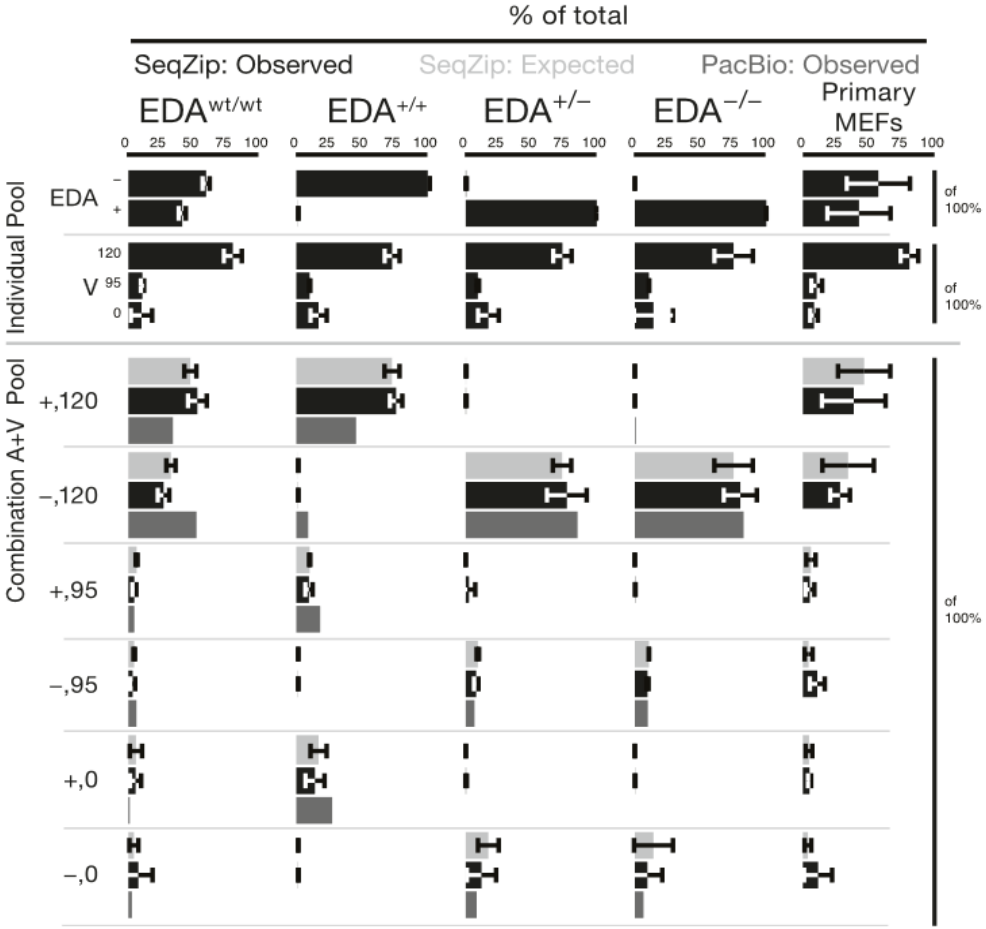
# Fn1 isoform abundance

- The effects of EDA inclusion or exclusion on V region splicing were previously tested by creating mice via homologous recombination with intronic splicing enhancers modified to favor either constitutive inclusion (+/+) or exclusion (−/−) of the EDA exon
- That study also analyzed mice heterozygous for the modified locus (+/−) and the wild-type parental strain (wt/wt)
- Immortalized mouse embryonic fibroblasts generated from all four mouse lines and performed SeqZip analysis



# Fn1 isoform abundance

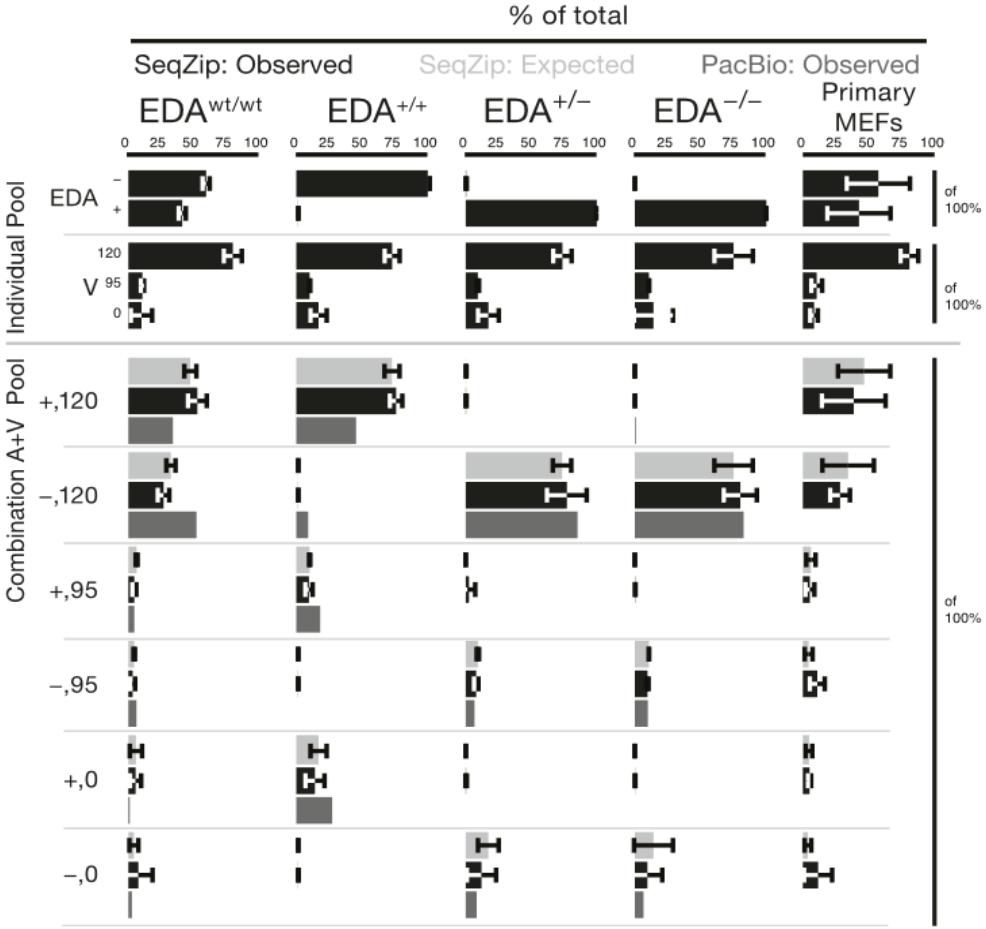
- EDA and V isoform ratios determined from PCR band intensities of the A and V pool ligation products (SeqZip: Observed) were used to calculate expected EDA:V isoform abundances, assuming no interdependence between the two regions
- Also generated cDNAs and sequenced them on a Pacific Biosciences RSII instrument (PacBio:Observed), a single molecule platform with sufficient read length to maintain connectivity between the EDA and V regions
- In both the SeqZip and PacBio data sets, constitutive EDA inclusion or exclusion was as expected in the +/+ and −/− cells, respectively
- Unexpectedly, however, we could not detect any EDA inclusion in the +/− cells despite confirming the presence of both alleles in gDNA



# Fn1 isoform abundance

- EDA and V isoform ratios determined from PCR band intensities of the A and V pool ligation products (SeqZip: Observed) were used to calculate expected EDA:V isoform abundances, assuming no interdependence between the two regions
- Also generated cDNAs and sequenced them on a Pacific Biosciences RSII instrument (PacBio:Observed), a single molecule platform with sufficient read length to maintain connectivity between the EDA and V regions
- In both the SeqZip and PacBio data sets, constitutive EDA inclusion or exclusion was as expected in the +/+ and −/− cells, respectively
- Unexpectedly, however, we could not detect any EDA inclusion in the +/− cells despite confirming the presence of both alleles in gDNA

Results support the view that the EDA and V regions of mouse Fn1 are spliced autonomously

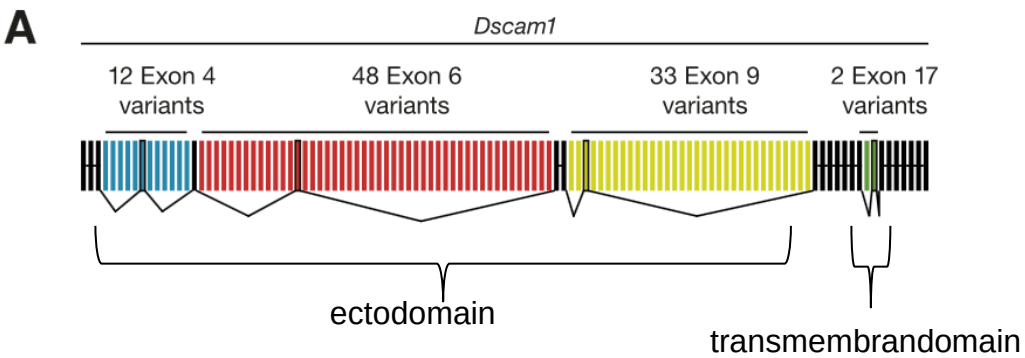


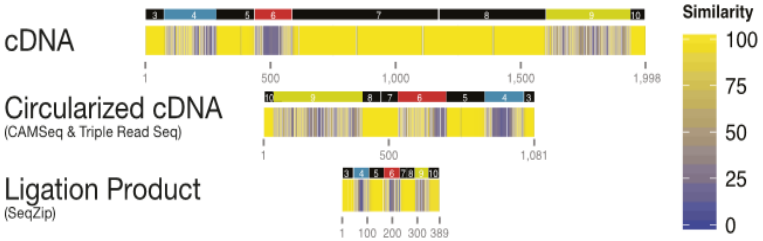
# Dscam1

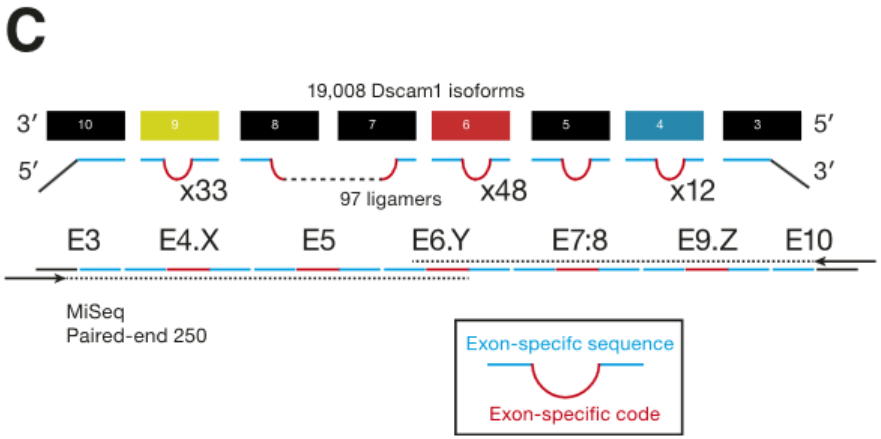
- Dscam1 produces with 38’016 possible isoforms the greatest known isoform diversity of any single gene
- Important in neuronal wiring and pathogen recognition
- Isoforms in the fruit fly, each encoding a unique protein that will help axons to chart a course in the developing brain
- Complete characterization of Dscam1 isoform diversity presents an extreme technical challenge
- The four regions of mutually exclusive cassette exons span > 4300 nt in full-length mRNAs

# Analysis of Dscam1 isoforms

- Alternative splicing of four blocks of mutually exclusive cassette exons (exons 4, 6, 9, and 17) can potentially produce 38,016 possible mRNA isoforms
- Long regions of sequence homology promote template switching during RT-PCR, this can generate novel isoforms not originally present in the biological sample
- SeqZip can maintain the connectivity information, compresses sequence length and increases sequence heterogeneity → greatly decreasing the potential for template switching

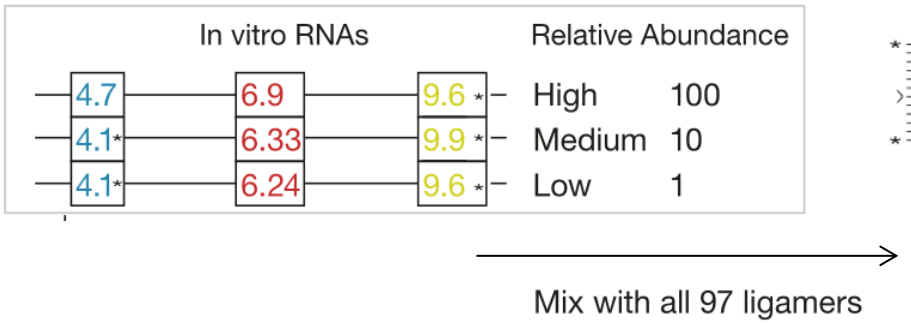


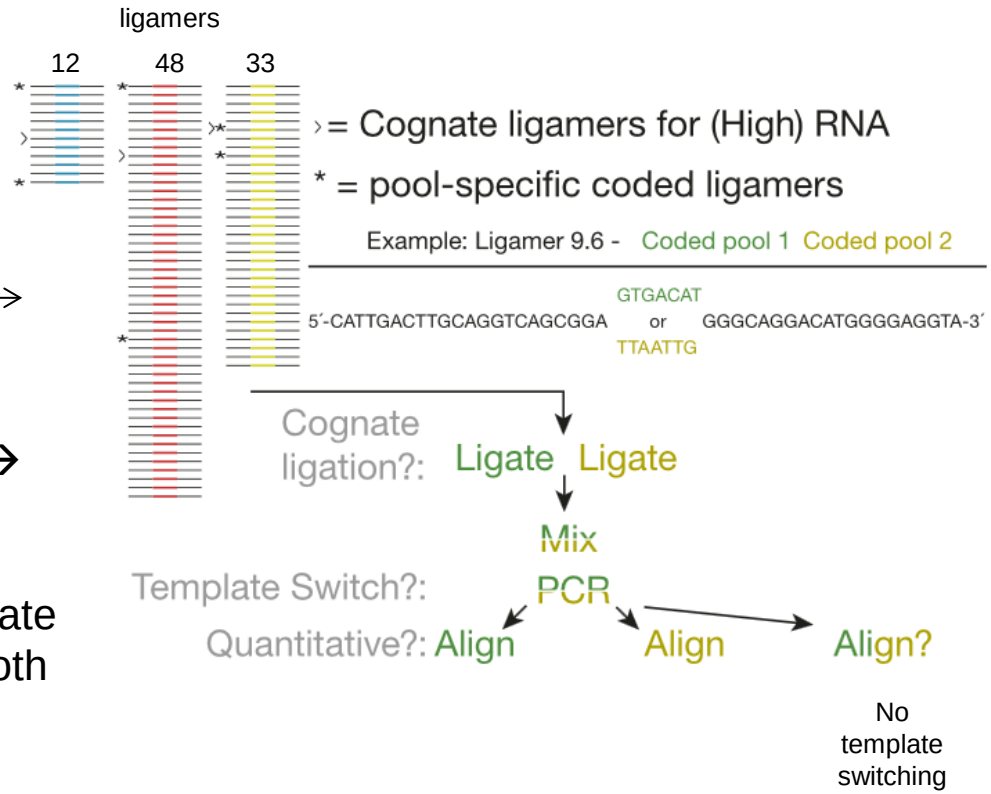



# Analysis of Dscam1 isoforms

- <u>SeqZip should greatly reduce template switching</u>
- Mixed together three different in vitro-transcribed Dscam1 isoforms in the presence of total RNA from a mouse hepatoma cell line
- This mixture was then divided into two separate ligation reactions, each containing a complete 97 ligamer pool that differed only in the 7 nt ligamer barcode assigned to two exons in each cluster

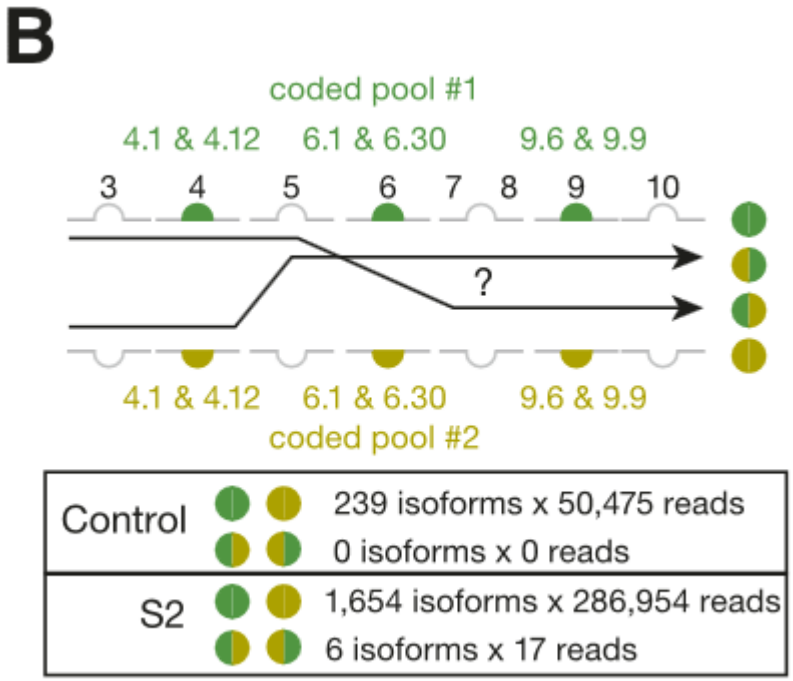



- Differentially coded samples were mixed → PCR → MiSeq
- 50,475 reads obtained in these control reactions → none were indicative of template switching (no ligation product contained both pool 1 and pool 2 barcodes)

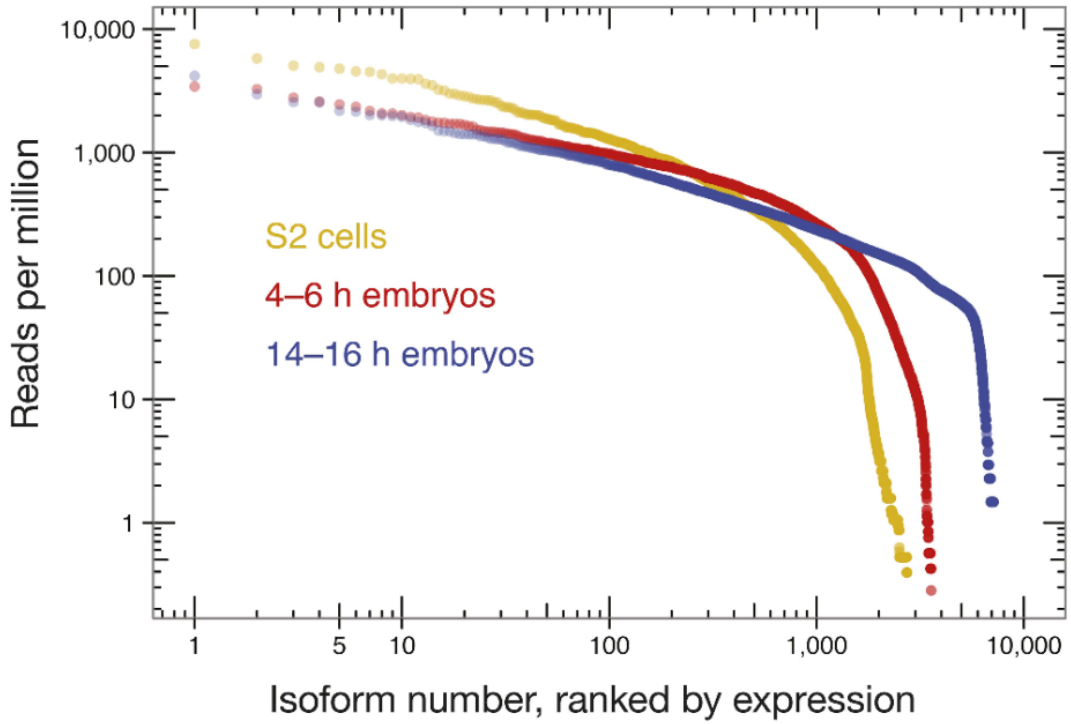
# Analysis of Dscam1 isoforms

- Same differential coding approach was applied to Drosophila S2 cell poly(A)-selected RNA
- Just 17 of 111,242 reads (0.015%) corresponded to template-switched isoforms
- SeqZip design greatly diminishes template switching (RT based system: 34-55%)



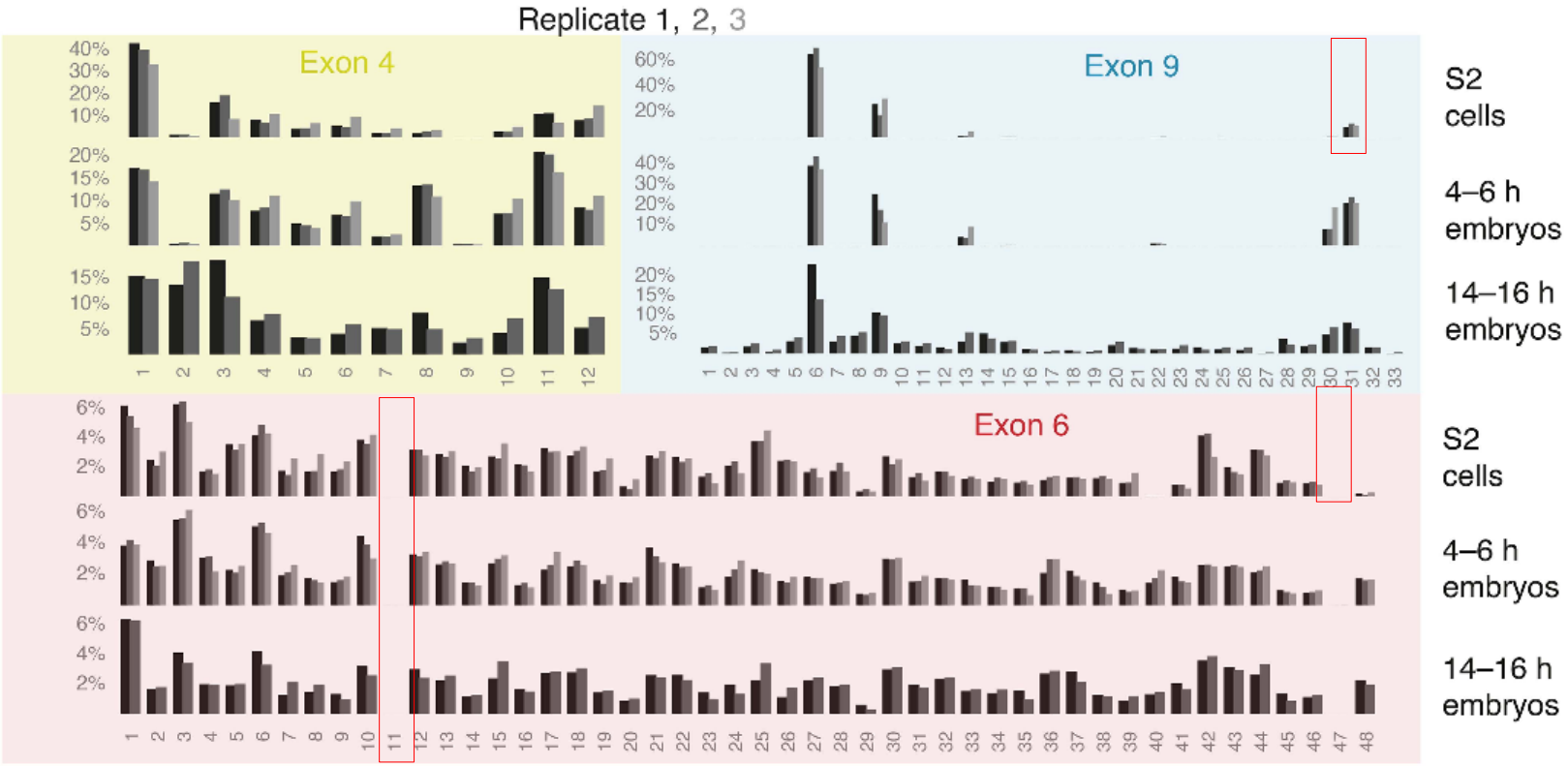
# Analysis of Dscam1 isoforms

- Used SeqZip to measure Dscam1 isoform identity and abundance in S2 cells, as well as 4–6 hr and 14–16 hr Drosophila embryos
- Reduced the median size of mRNA sequences analyzed from 1734 nt to 356 nt
- Products could be fully sequenced using paired-end reads in a single Illumina or MiSeq run
- Detected 8’397 of 18’612 possible isoforms
- Isoform abundances were highly correlated between both technical and biological replicates



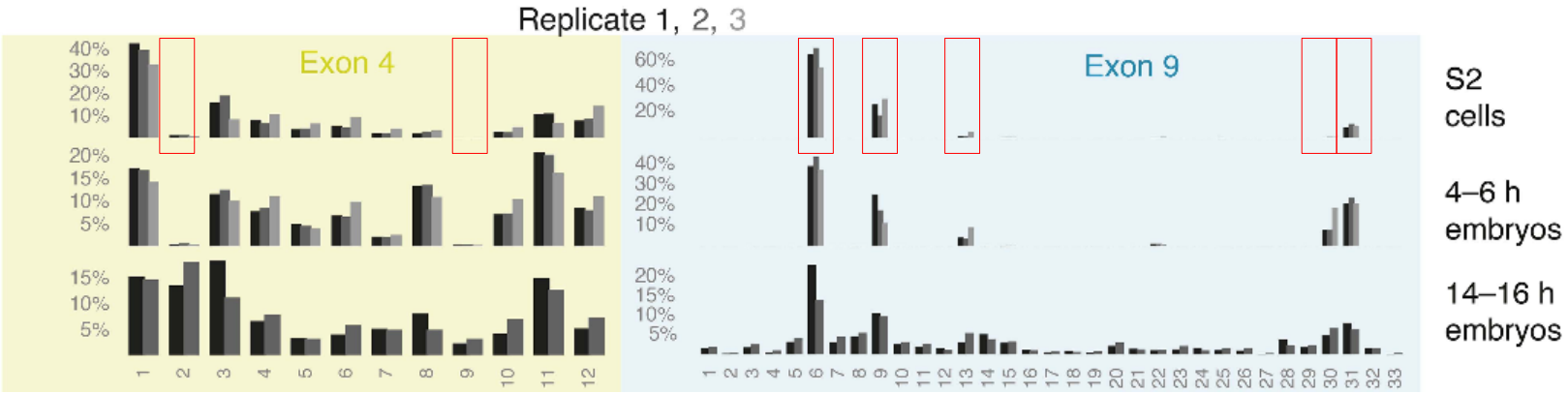
# Analysis of Dscam1 isoforms

- Of the 97 possible exons represented in the ligamer set, all were detected <u>except exon 6.11</u>, which is generally thought to be an unused pseudo-exon (additional evidence for the specificity by SeqZip)
- Patterns of individual exon use in S2 cells were directly comparable between the SeqZip and CAMSeq data sets (exceptions: 6.47 and 9.31)
- Isoform diversity increases with tissue complexity
- As previously shown, cluster 4 and 9 exon usage patterns change during development, whereas the cluster 6 pattern remains more static



# Analysis of Dscam1 isoforms

- In S2 cells, Dscam1 mRNAs incorporate very little of exon 4 cassettes 2 and 9 and use almost exclusively exon 9 cassettes 6, 9, 13, 30, and 31
- This pattern is the characteristic of hemocytes and consistent with the macrophage-like nature of S2 cells
- Whereas 4–6 hr embryos are similar to S2 cells in exon clusters 4 and 9, 14–16 hr embryos show increased exon diversity, particularly in cluster 9

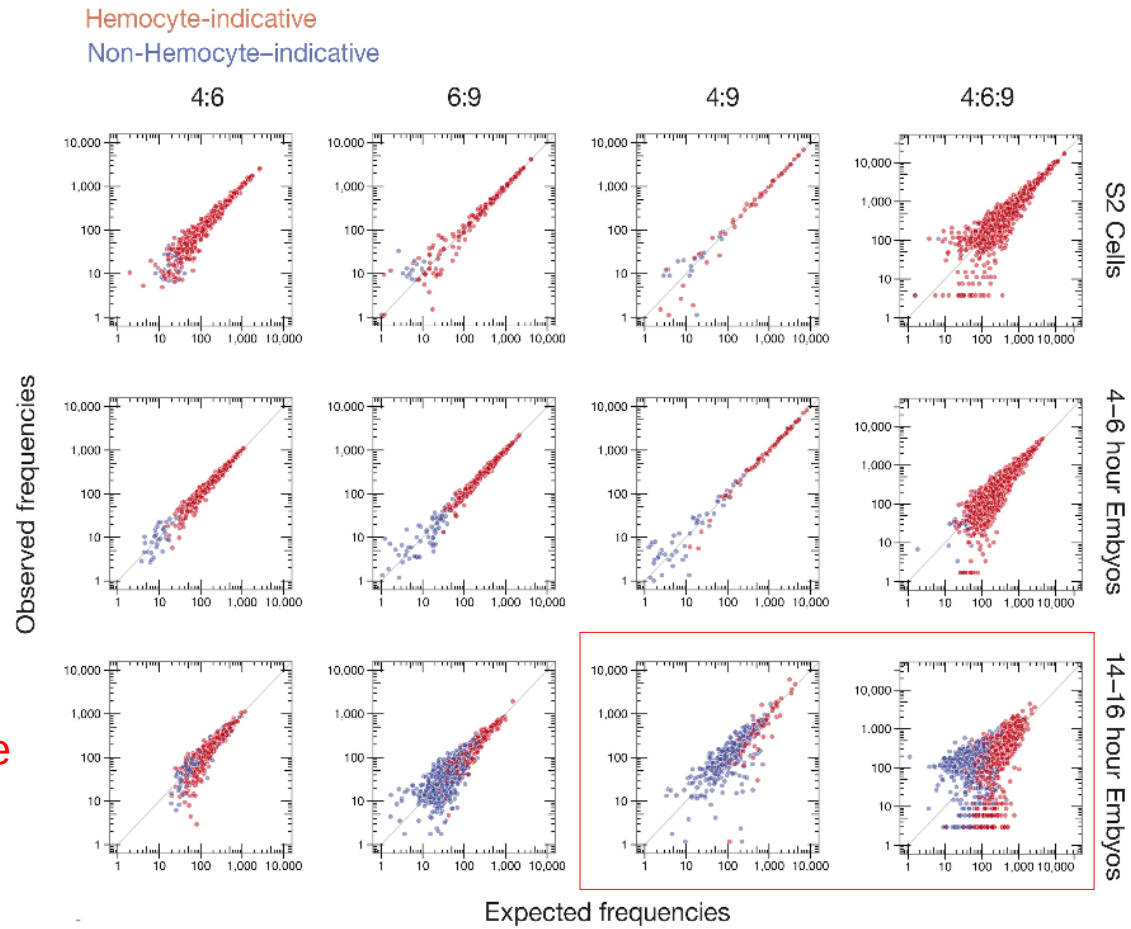


# Analysis of Dscam1 isoforms

- Dscam1 isoforms associated with hemocytes (lacking exon 4.2 and 4.9 or containing exon 9 cassettes 6, 9, 13, 30, or 31) are the most abundant in all three samples, but other isoforms emerge as development proceeds
- Examine the <u>possibility of coordinated splicing</u> → calculated expected pairwise and three-way exon combination frequencies for every transcript isoform observed in each sample → <u>assumtion: no coordination</u>

Comparison of expected and observed frequencies → no statistically significant differences between expectation and observation for S2 cells and 4–6 hr embryos

did fall outside of the expected range




Most are Non-hemocytes

# Analysis of Dscam1 isoforms

Because the majority of 4:6:9 combination frequencies (99.3%) were consistent with the null hypothesis of no coordination, the data agree with previous studies that individual cassettes in Dscam clusters 4, 6, and 9 are chosen independently, with exon choice in one cluster having no detectable effect on subsequent exon choice in another cluster

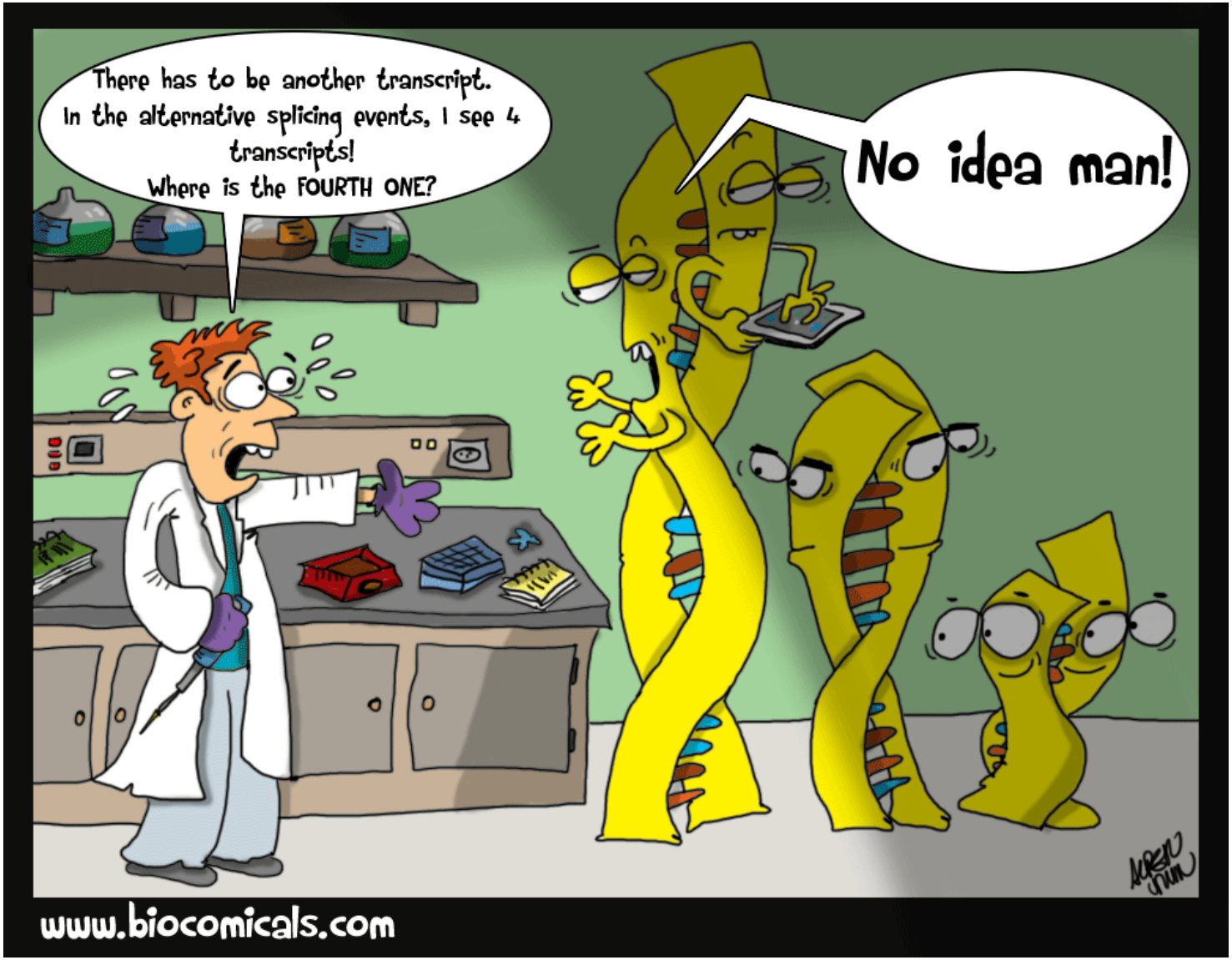
# Advantages and Disadvantages of SeqZip

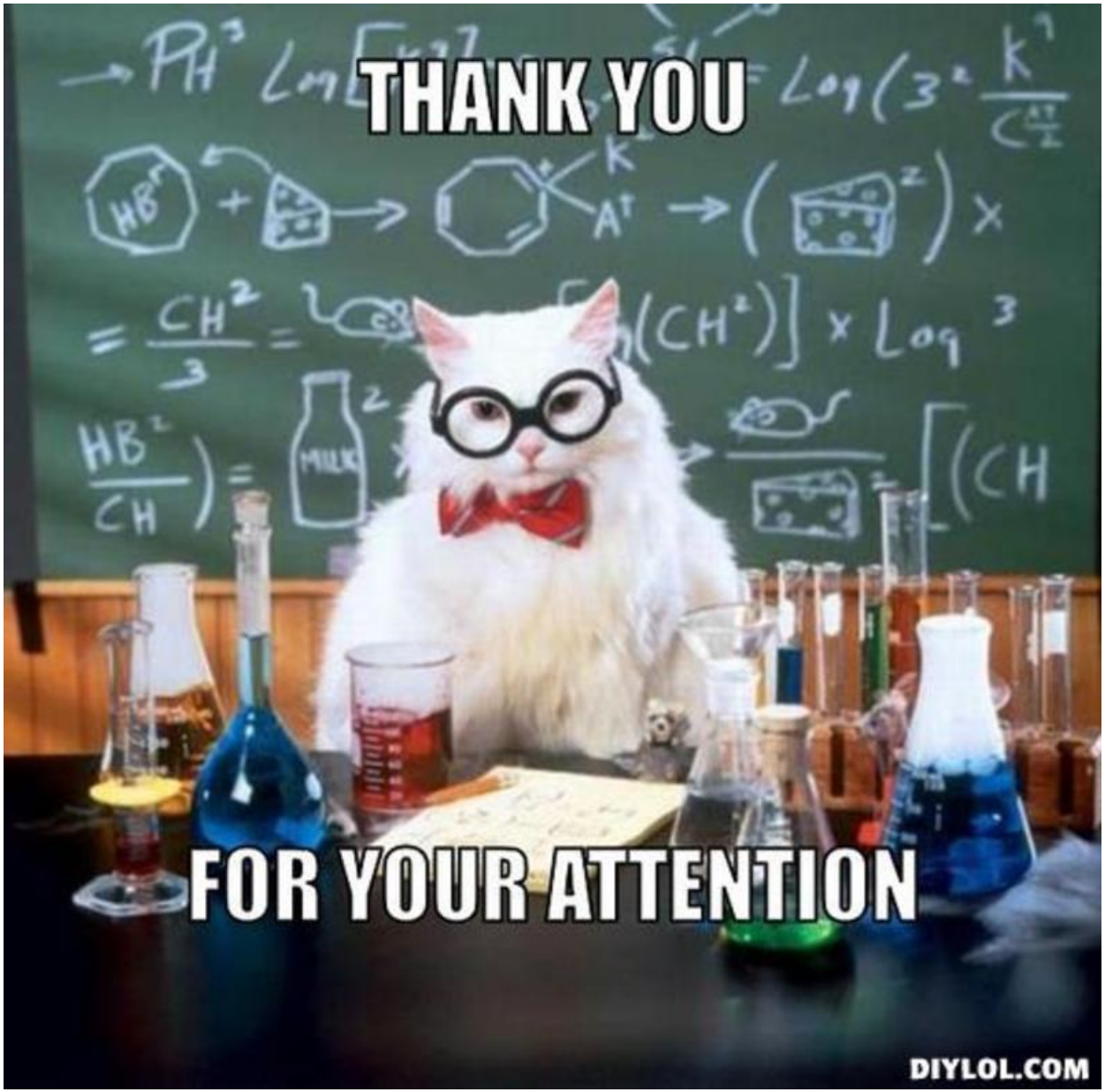
<u>Advantage</u>

- No RT step
- Eliminates long intervening regions of common sequence
- Discourage template switching during subsequent amplification

<u>Disadvantage</u>

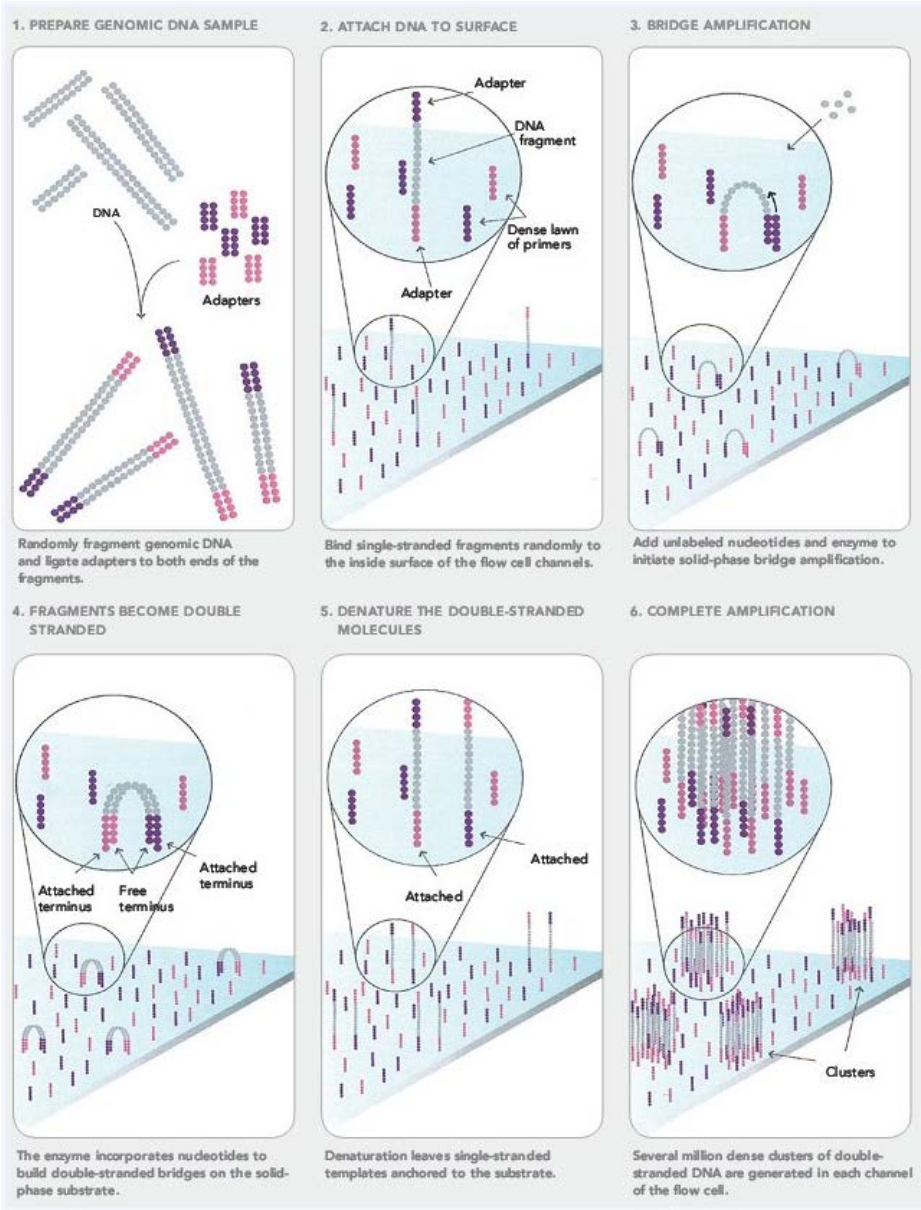
- Many mammalian transcripts have more varied alternative splicing
- Alternative 5′ and 3′ splice site usage and intron retention
- Ligamer design against these types of alternative splicing quickly becomes difficult

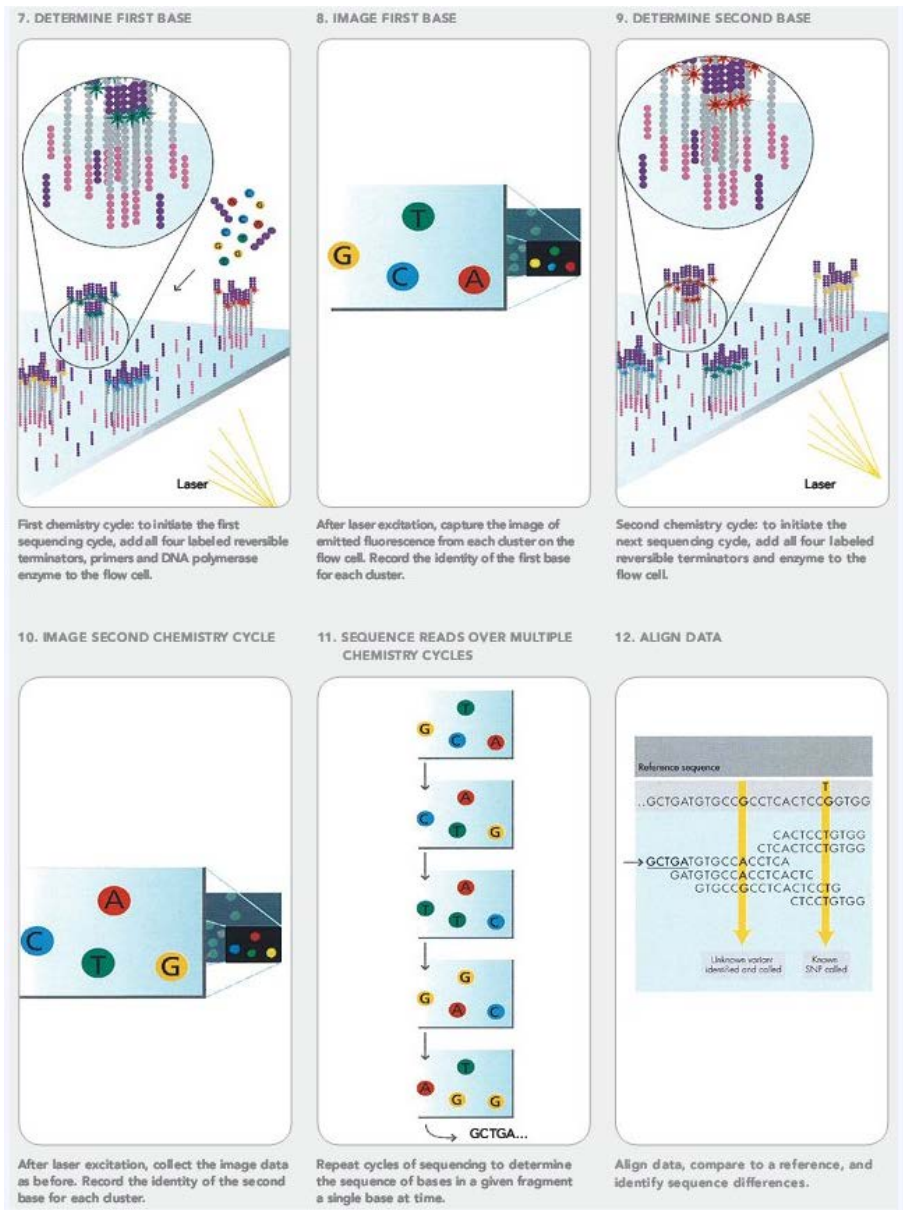

There has to be another transcript. In the alternative splicing events, I see 4 transcripts! Where is the FOURTH ONE?
No idea man!
www.biocomicals.com




http://beta.diylol.com/memes/2031-chemistry-cat/posts/566875-thank-you-for-your-attention

# Illumina


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Attached terminus
Free terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.


7. DETERMINE FIRST BASE
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
8. IMAGE FIRST BASE
T
G
C
A
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
9. DETERMINE SECOND BASE
Laser
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
10. IMAGE SECOND CHEMISTRY CYCLE
A
C
T
G
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
GCTGA...
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
12. ALIGN DATA
Reference sequence
...GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
Align data, compare to a reference, and identify sequence differences.