

# LONG READ RNA SEQUENCING

ALISANDRA DENTON

HHU Duesseldorf

24.08.2022

## WHY LONG-READS?

For *quantifying gene expression* the minimum length that can be uniquely mapped will be the most efficient use of money.

### Speaker notes

It's the total read counts that are important for differential expression analyses. Once reads are long enough to map uniquely, there's little else to gain there. So in what sort of cases \_do\_ we want long reads?

## WHY LONG-READS?

For *quantifying gene expression* the minimum length that can be uniquely mapped will be the most efficient use of money.

So what else?

### Speaker notes

It's the total read counts that are important for differential expression analyses. Once reads are long enough to map uniquely, there's little else to gain there. So in what sort of cases \_do\_ we want long reads?

# WHAT ABOUT ISOFORMS?



## Speaker notes

Short reads can be used for definition/quantification of some isoforms.

# WHAT ABOUT ISOFORMS?



Both isoforms supported



Speaker notes

Short reads can be used for definition/quantification of some isoforms.

# WHAT ABOUT ISOFORMS?



Both isoforms supported



## Short reads can detect simple cases of alternative splicing

Speaker notes

Short reads can be used for definition/quantification of some isoforms.

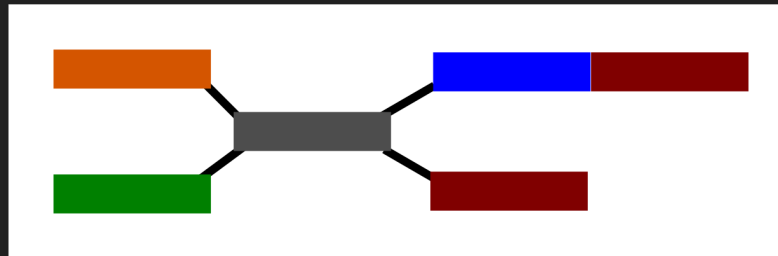
# WHAT ABOUT ISOFORMS?



## Speaker notes

Data not even theoretically there to resolve 2x options beyond read length. The "Frayed rope" is in fact a classic problem when trying to resolve graphs, whether we're talking exons or kmers.

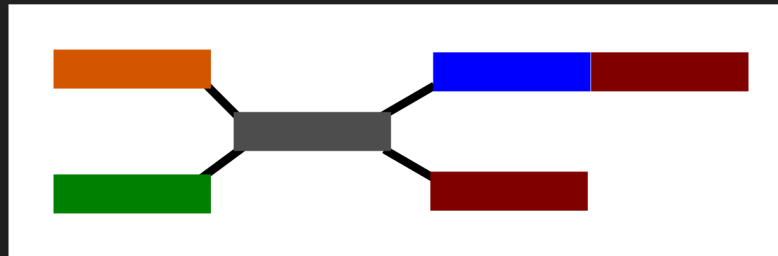
# WHAT ABOUT ISOFORMS?



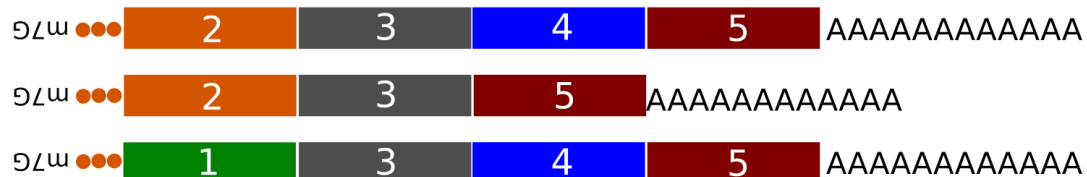
## Speaker notes

Data not even theoretically there to resolve 2x options beyond read length. The "Frayed rope" is in fact a classic problem when trying to resolve graphs, whether we're talking exons or kmers.

# WHAT ABOUT ISOFORMS?



2-4 isoforms possible, none is solidly supported



## Speaker notes

Data not even theoretically there to resolve 2x options beyond read length. The "Frayed rope" is in fact a classic problem when trying to resolve graphs, whether we're talking exons or kmers.

# WHAT ABOUT ISOFORMS?



Speaker notes

Thus long reads help in defining and quantifying isoforms / real gene models.

# WHAT ABOUT ISOFORMS?



Long reads help for complicated cases of alternative splicing!

Speaker notes

Thus long reads help in defining and quantifying isoforms / real gene models.

# WHAT ABOUT ISOFORMS?



Long reads help for complicated cases of alternative splicing!

- Gene structure annotation

Speaker notes

Thus long reads help in defining and quantifying isoforms / real gene models.

# WHAT ABOUT ISOFORMS?



Long reads help for complicated cases of alternative splicing!

- Gene structure annotation
- Detecting complex isoforms

Speaker notes

Thus long reads help in defining and quantifying isoforms / real gene models.

# WHAT ABOUT ISOFORMS?



Long reads help for complicated cases of alternative splicing!

- Gene structure annotation
- Detecting complex isoforms
- Quantifying complex isoforms

Speaker notes

Thus long reads help in defining and quantifying isoforms / real gene models.

# TO REALLY UNDERSTAND THE DATA, YOU MUST UNDERSTAND THE PROTOCOL

## Long read RNAseq options

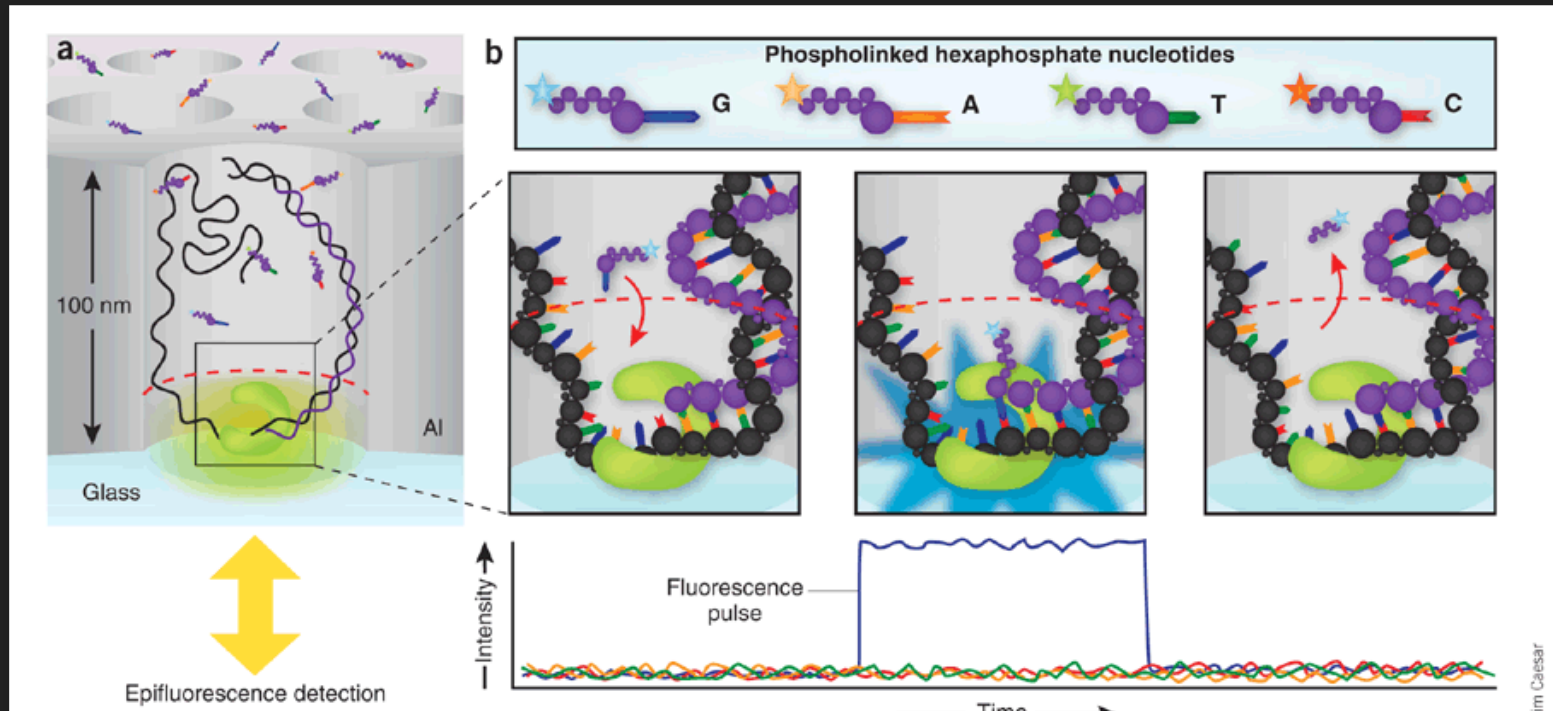
- Iso-Seq: Pacific Biosciences
- direct RNAseq: Oxford Nanopore Technologies

### Speaker notes

Recall from the Illumina data how 3' bias could indicate low-quality RNA, just one example how you can't really understand the data without understanding how it is made. Let's take a detailed look at our long read options. Qualifier: please keep in mind I've never actually done any of these protocols, so have reading expertise only!

# ISO-SEQ: PACBIO SEQUENCING

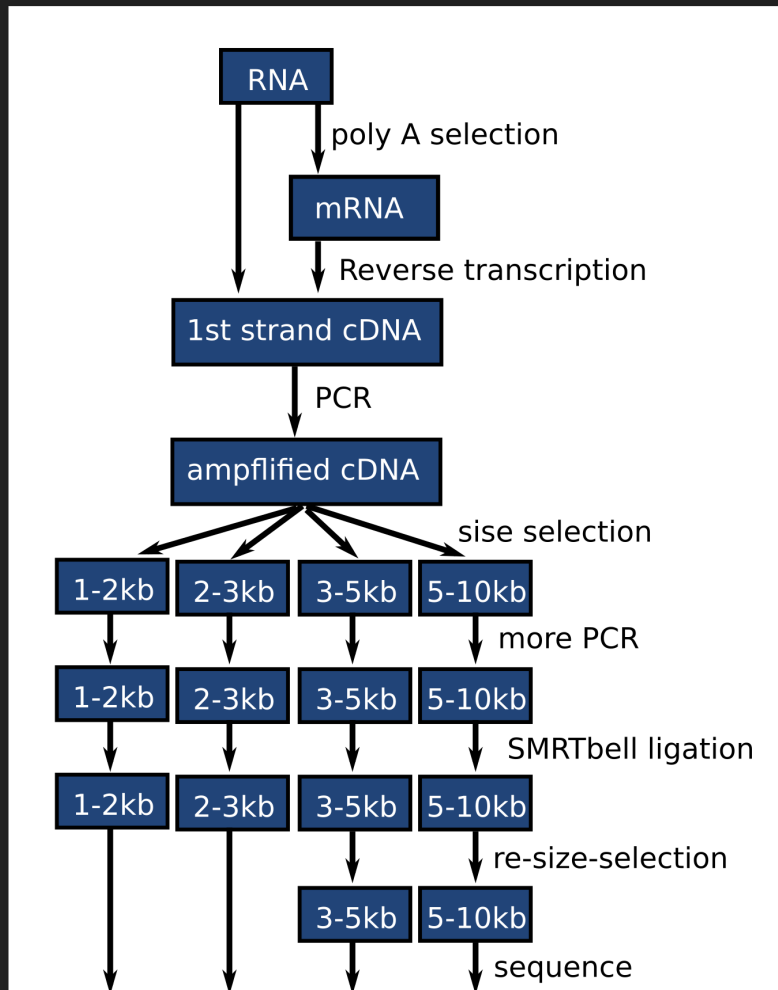
"Single Molecule Real Time sequencing (SMRT)"



## Speaker notes

You've seen the basics of PacBio sequencing before, many wells, sequencing by synthesis, incorporation of fluorescent nucleotides can be detected by laser / camera combo; but we aren't mashing raw tissue on to the flow cell, how do we get here?

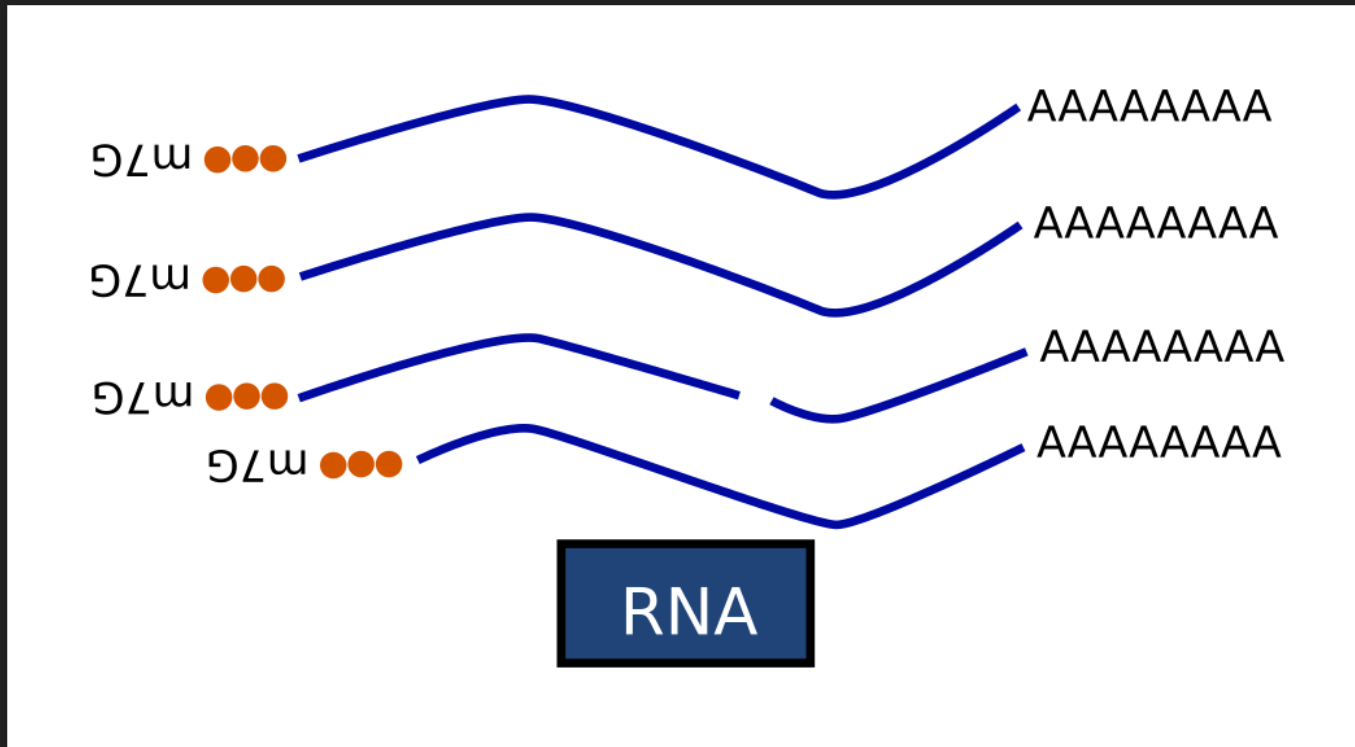
# ISO-SEQ: LIBRARY PREPARATION



## Speaker notes

There are a few important things to notice already from the overview: PCR, independent processing of size selected libraries.

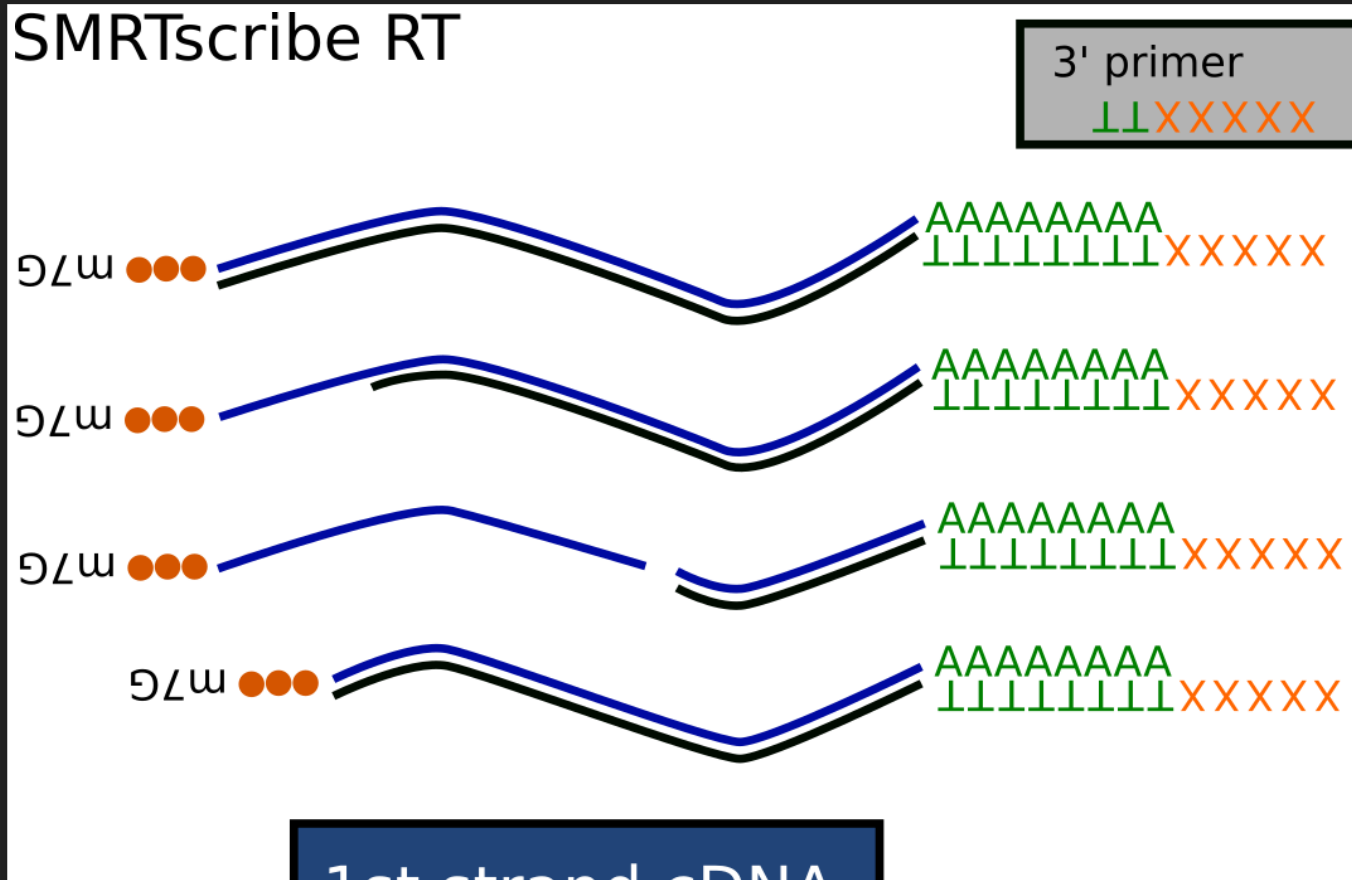
# ISO-SEQ: LIBRARY PREPARATION



Speaker notes

But let's follow some example mRNAs as they go through the library prep protocol.

# ISO-SEQ: LIBRARY PREPARATION

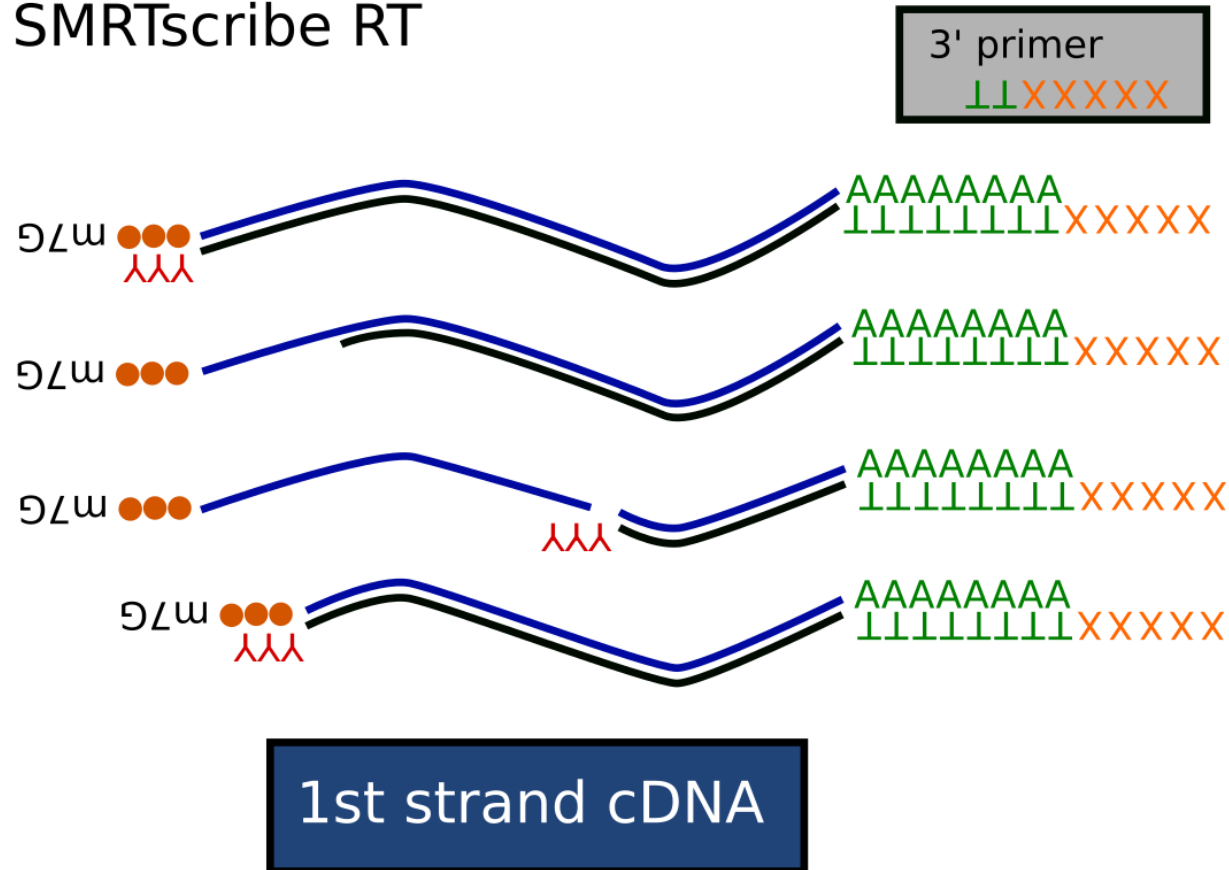


## Speaker notes

Full reverse transcription, incomplete reverse transcription, full RT to point where mRNA was broken, and finally full RT on shorter isoform. Throughout, [X, Y, Z] will be used to indicate specific / controlled sequences and as appropriate, their reverse complement (e.g. primer sequences)

# ISO-SEQ: LIBRARY PREPARATION

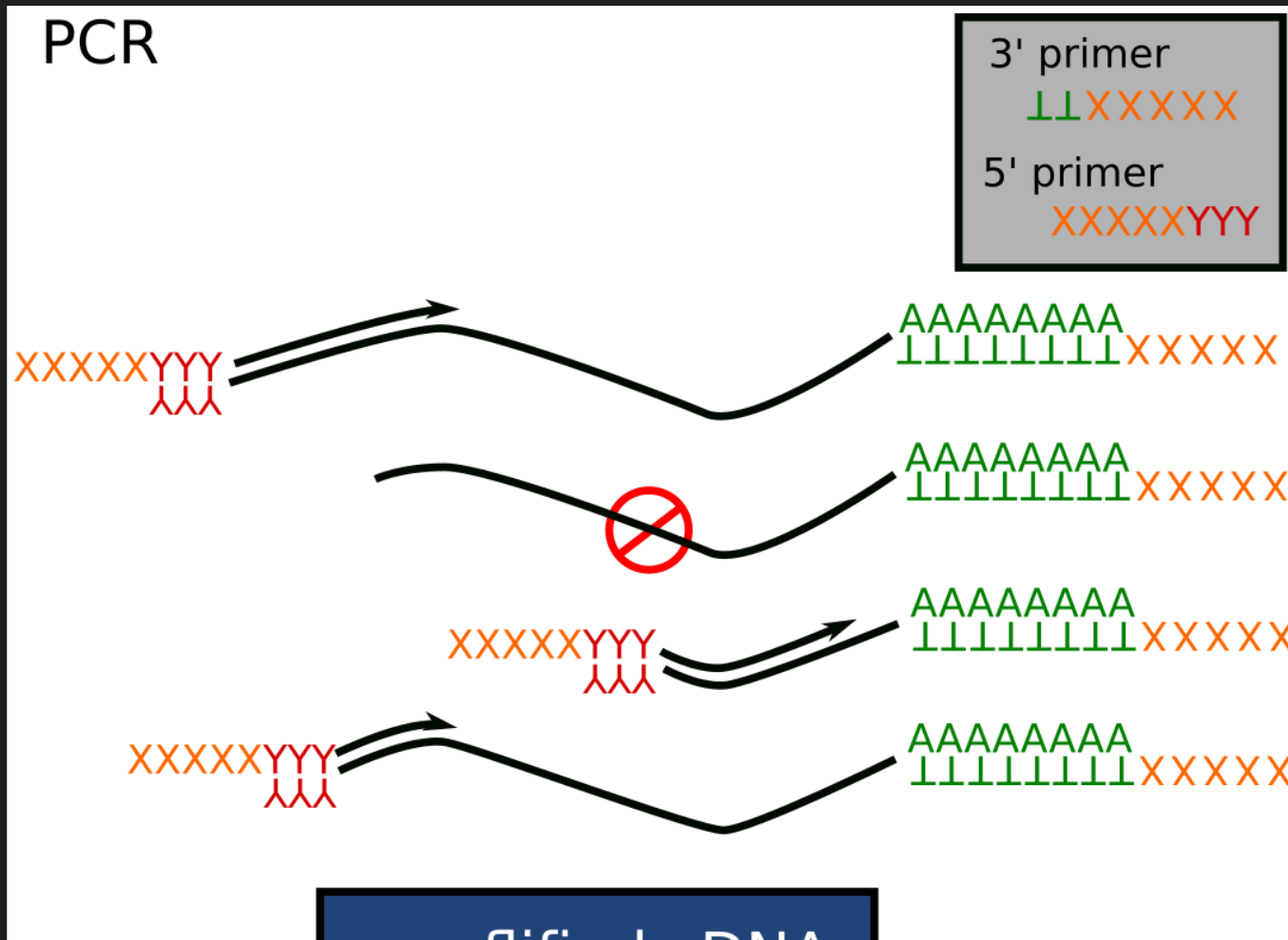
SMRTscribe RT



## Speaker notes

SMRTscribe can tag double stranded ends to help determine whether transcripts are full length. However, detecting a double stranded end is not the same as detecting the 5' cap!

# ISO-SEQ: LIBRARY PREPARATION



## Speaker notes

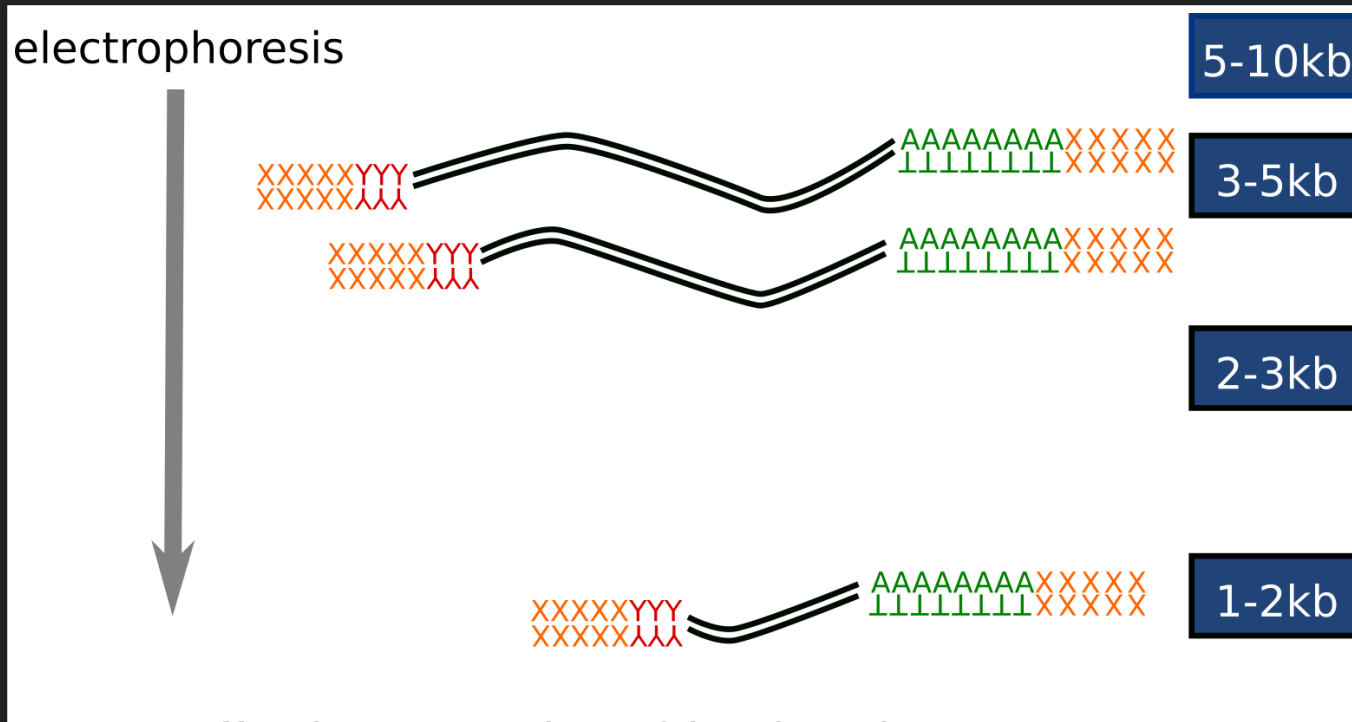
Generally speaking, only sequences with the scribe tag will be further amplified during the following PCR. But note as well that the full long and short isoforms are now indistinguishable from the degradation product.

PCR

3' primer  
 5' primer

amplified cDNA

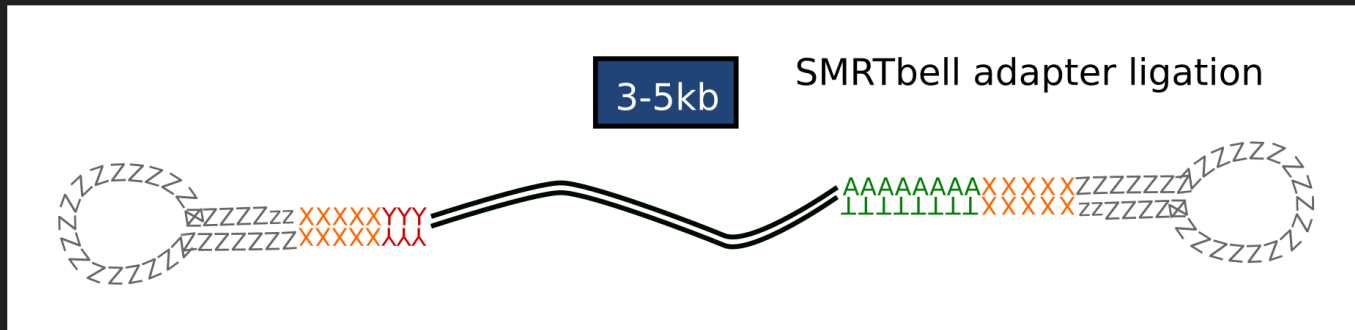
# ISO-SEQ: LIBRARY PREPARATION



## Speaker notes

If a library has drastically different sizes and PCR is performed the shortest sequences will be amplified most efficiently and dominate the sequencing results. Same holds for Illumina, but the size range here is much broader. Thus, we need size selection. You'll want a Blue Pippen or a Sage ELF, as they have many fold higher recovery than extraction from a gel.

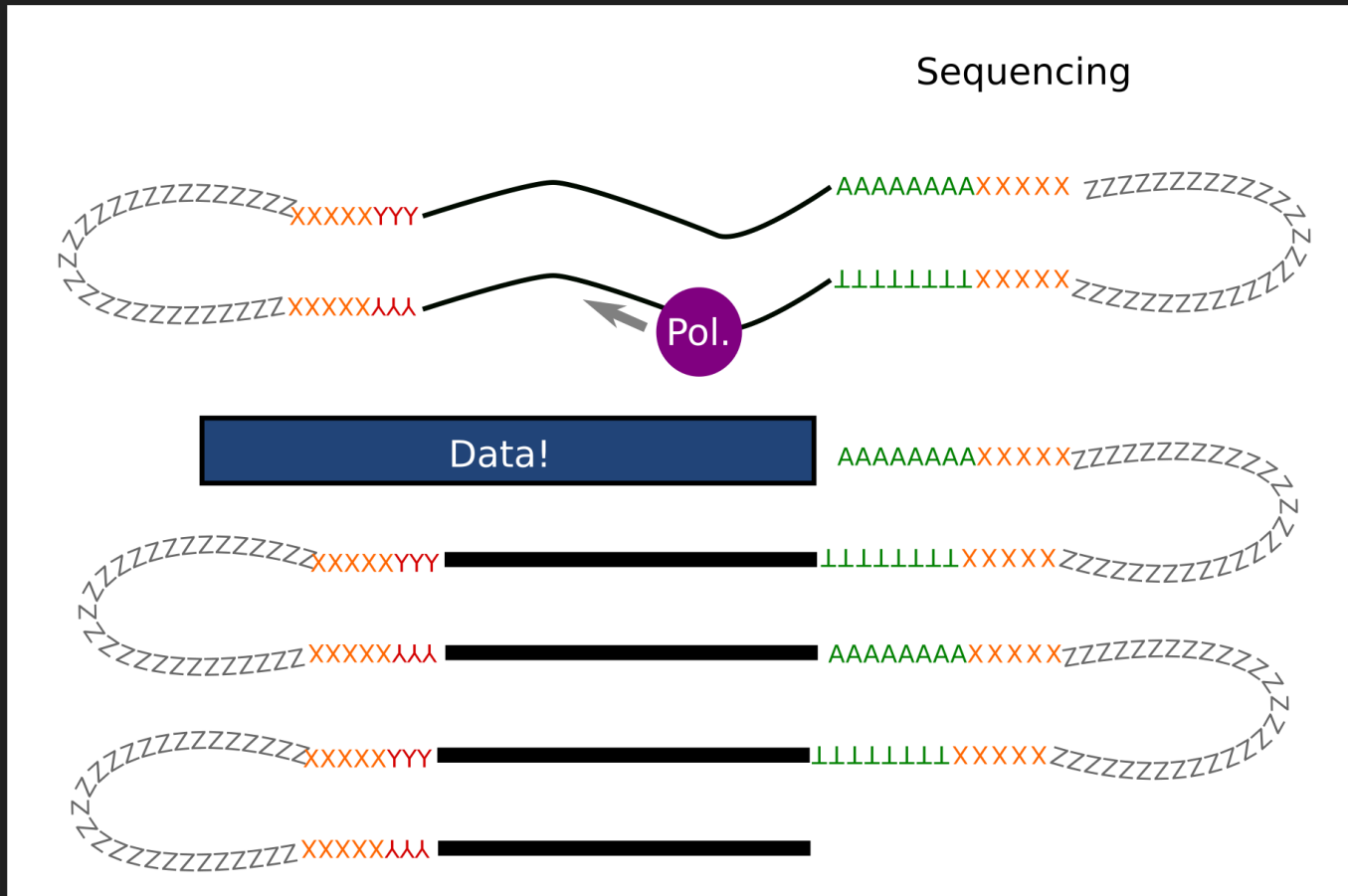
# ISO-SEQ: LIBRARY PREPARATION



Speaker notes

Finally the last hairpin sequencing adapters "SMRTbells" are ligated.

# ISO-SEQ: LIBRARY PREPARATION



## Speaker notes

These hair pin adapters allow sequencing to proceed in a circle, so that each sequence in the library produces multiple passes / subreads.

# ISO-SEQ: LIBRARY PREPARATION

## Alternatives / variation:

### Speaker notes

5' end selection can stop the sequencing of degradation products, but it may well be at the loss of long or readily degradable transcripts. Looking 1:1 at Illumina and PacBio RNAseq from the same study, shows much higher 3' bias on the PacBio (as all RT starts from poly-A). If you have a gene of interest that you think has complex splicing and you'd like to have it better annotated and you don't need the rest, then please take gene specific primers. It will be cheaper, easier to analyze and you'll have much better depth for your target. If you are trying to do genome-wide annotation, then highly expressed genes of e.g. photosynthesis genes will be sequenced very very deeply before transcription factors are detected, normalization can help with this by reducing the dynamic range in a library. Keep in mind that most of the above is inappropriate or should be done with care for `_quantification_`

# ISO-SEQ: LIBRARY PREPARATION

## Alternatives / variation:

- 5' end selection

### Speaker notes

5' end selection can stop the sequencing of degradation products, but it may well be at the loss of long or readily degradable transcripts. Looking 1:1 at Illumina and PacBio RNAseq from the same study, shows much higher 3' bias on the PacBio (as all RT starts from poly-A). If you have a gene of interest that you think has complex splicing and you'd like to have it better annotated and you don't need the rest, then please take gene specific primers. It will be cheaper, easier to analyze and you'll have much better depth for your target. If you are trying to do genome-wide annotation, then highly expressed genes of e.g. photosynthesis genes will be sequenced very very deeply before transcription factors are detected, normalization can help with this by reducing the dynamic range in a library. Keep in mind that most of the above is inappropriate or should be done with care for `_quantification_`

# ISO-SEQ: LIBRARY PREPARATION

## Alternatives / variation:

- 5' end selection
- Deep sequencing with target primers

### Speaker notes

5' end selection can stop the sequencing of degradation products, but it may well be at the loss of long or readily degradable transcripts. Looking 1:1 at Illumina and PacBio RNAseq from the same study, shows much higher 3' bias on the PacBio (as all RT starts from poly-A). If you have a gene of interest that you think has complex splicing and you'd like to have it better annotated and you don't need the rest, then please take gene specific primers. It will be cheaper, easier to analyze and you'll have much better depth for your target. If you are trying to do genome-wide annotation, then highly expressed genes of e.g. photosynthesis genes will be sequenced very very deeply before transcription factors are detected, normalization can help with this by reducing the dynamic range in a library. Keep in mind that most of the above is inappropriate or should be done with care for `_quantification_`

# ISO-SEQ: LIBRARY PREPARATION

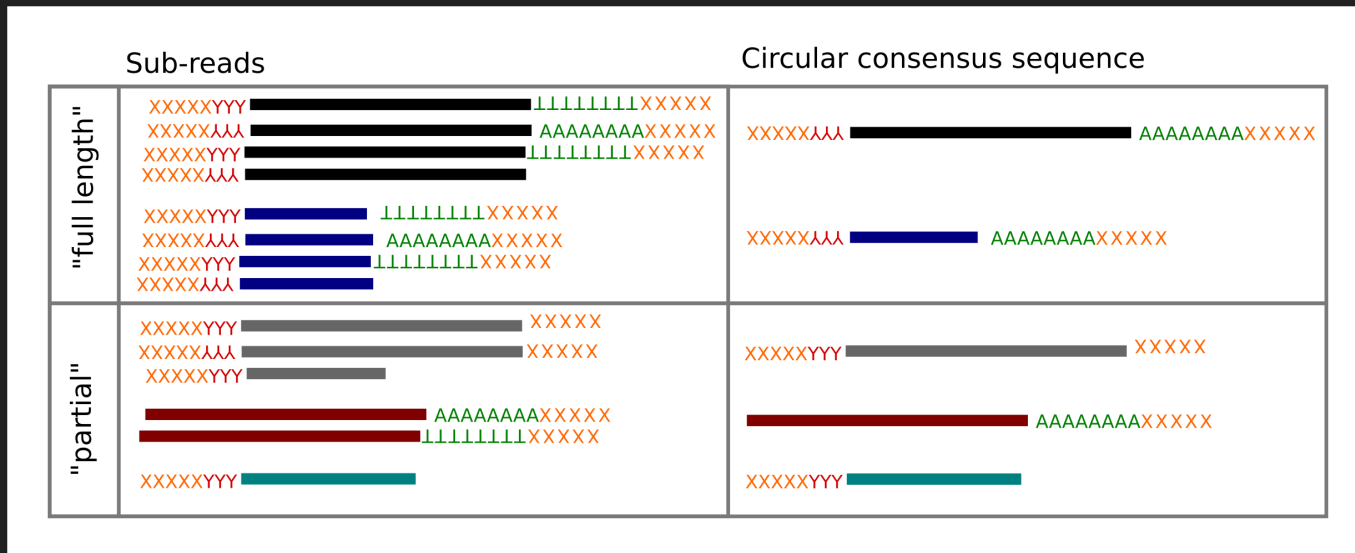
## Alternatives / variation:

- 5' end selection
- Deep sequencing with target primers

### Speaker notes

5' end selection can stop the sequencing of degradation products, but it may well be at the loss of long or readily degradable transcripts. Looking 1:1 at Illumina and PacBio RNAseq from the same study, shows much higher 3' bias on the PacBio (as all RT starts from poly-A). If you have a gene of interest that you think has complex splicing and you'd like to have it better annotated and you don't need the rest, then please take gene specific primers. It will be cheaper, easier to analyze and you'll have much better depth for your target. If you are trying to do genome-wide annotation, then highly expressed genes of e.g. photosynthesis genes will be sequenced very very deeply before transcription factors are detected, normalization can help with this by reducing the dynamic range in a library. Keep in mind that most of the above is inappropriate or should be done with care for `_quantification_`

# ISO-SEQ: DATA PROCESSING



## Speaker notes

Knowing the protocol, we can better understand what's necessary in data processing. We get subreads, and can convert to CCS. CCS ~= "real reads". There's also reads w/(o) adapters poly-A that are handled specifically during processing.

# ISO-SEQ: DATA PROCESSING

## Starting from the CCS

Circular consensus sequence

XXXXXXXXXX ██████████ AAAAAAAAAA XXXXXX  
XXXXXXXXXX ██████████ AAAAAAAAAA XXXXXX

Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

## Starting from the CCS

Circular consensus sequence

XXXXXXXXXX ██████████ AAAAAAAAAA XXXXXX  
XXXXXXXXXX ██████████ AAAAAAAAAA XXXXXX

- Trim adapters and demultiplex

Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

# Starting from the CCS

## Circular consensus sequence

XXXXXXLL ██████████ AAAAAAAAAAXXXXXX  
XXXXXXLL ██████████ AAAAAAAAAAXXXXXX

- Trim adapters and demultiplex
- Orient with poly-A and trim

## Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

## Starting from the CCS

Circular consensus sequence

XXXXXX^^^ XXXXXXXXXXXXXXXXXXXX AAAAAAAAAAXXXXXX  
XXXXXX^^^ XXXXXXXXXXXX AAAAAAAAAAXXXXXX

- Trim adapters and demultiplex
- Orient with poly-A and trim
- Cluster into "full length" draft transcripts

Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

## Starting from the CCS

Circular consensus sequence

XXXXXXXXXX ██████████ AAAAAA XXXXXX  
XXXXXXXXXX ██████████ AAAAAA XXXXXX

- Trim adapters and demultiplex
- Orient with poly-A and trim
- Cluster into "full length" draft transcripts
- Polish with all data

Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

## Starting from the CCS

Circular consensus sequence

XXXXXXXXXX ██████████ AAAAAA XXXXXX  
XXXXXXXXXX ██████████ AAAAAA XXXXXX

- Trim adapters and demultiplex
- Orient with poly-A and trim
- Cluster into "full length" draft transcripts
- Polish with all data

Speaker notes

A common analysis pipeline (similar to that which we will do today) might look like this...

# ISO-SEQ: DATA PROCESSING

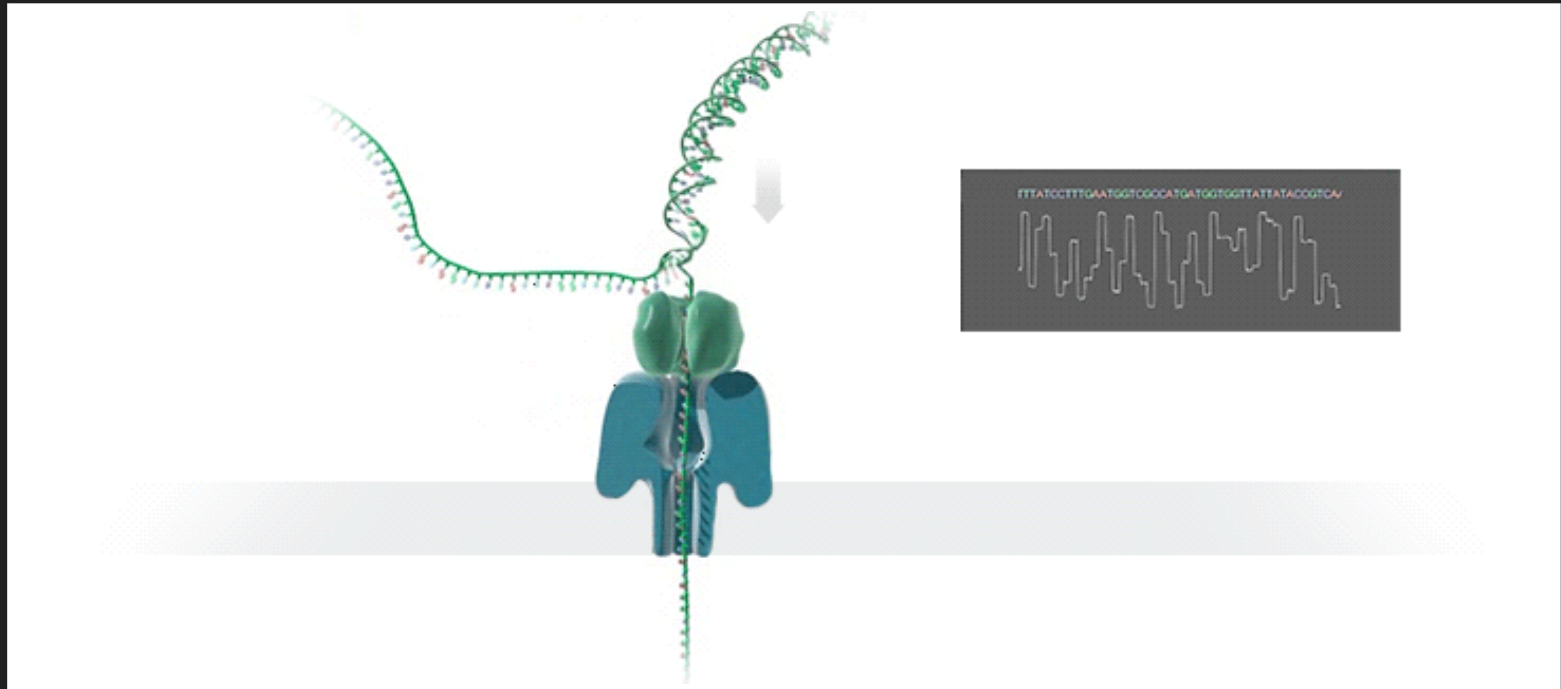


## Speaker notes

The per bp costs are for the subreads. The CCS represents the (mostly) unique pieces of evidence, and does have decent quality. The 'high quality' transcript is generally the final result you're after.

# DIRECT RNASEQ: NANOPORE

Tracing disruption in current as RNA/DNA sequence passes through a Nanopore

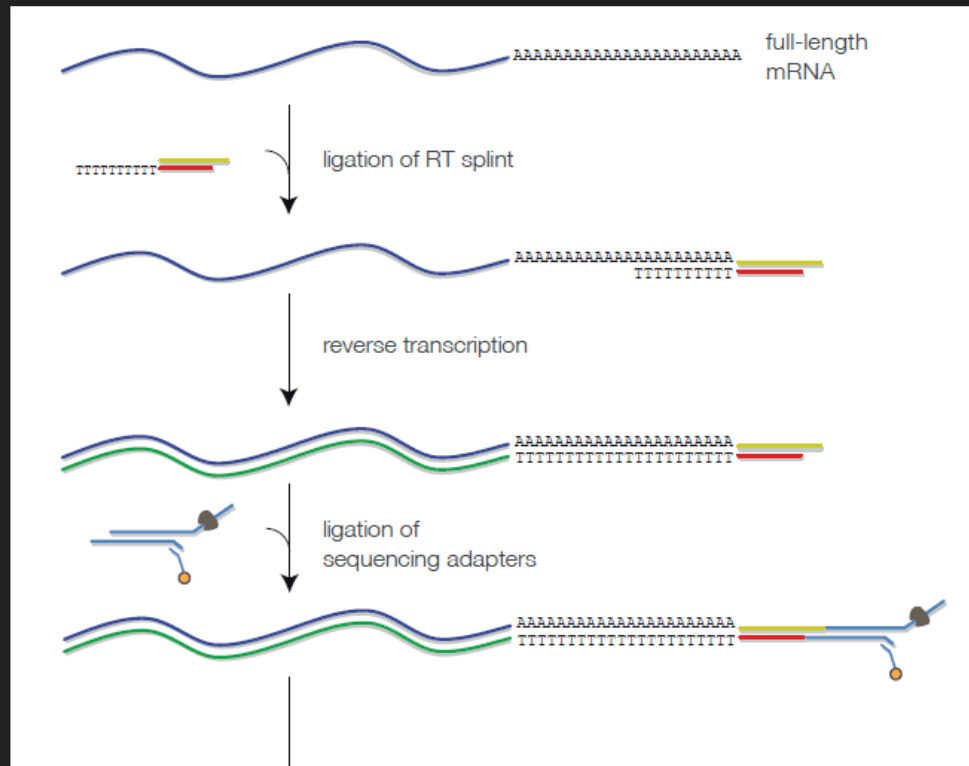


## Speaker notes

You've probably seen the basics where bases passing through a nanopore cause a change in the current.

# DIRECT RNASEQ: LIBRARY PREPARATION

One option: sequence mRNA and 1st strand cDNA



## Speaker notes

Perhaps unsurprisingly, the protocol is comparatively simple (and maybe also not final). Ligate adapters to poly-A (3' selection), RT, and ligate sequencing adapters (1D squared, so mRNA and 1st strand can be sequenced). On mRNA itself is where modification could in theory be detected.

# DIRECT RNASEQ: LIBRARY PREPARATION

## Variations

### Speaker notes

It's unclear to me if one can just go try direct RNAseq and expect any results yet, (2 papers total), But what does seem to be working a bit better is creating cDNA libraries; which increases yield, makes it more doable, and comes with all the caveats of these methods (PCR bias, no more modification detection). Major obvious diff (advantage?) compared to PacBio is the reduced PCR and reduced duplicity during sequencing.

# DIRECT RNASEQ: LIBRARY PREPARATION

## Variations

- sequence cDNA (2x yield)

### Speaker notes

It's unclear to me if one can just go try direct RNAseq and expect any results yet, (2 papers total), But what does seem to be working a bit better is creating cDNA libraries; which increases yield, makes it more doable, and comes with all the caveats of these methods (PCR bias, no more modification detection). Major obvious diff (advantage?) compared to PacBio is the reduced PCR and reduced duplicity during sequencing.

# DIRECT RNASEQ: LIBRARY PREPARATION

## Variations

- sequence cDNA (2x yield)
- amplify and sequence cDNA (5x yield)

### Speaker notes

It's unclear to me if one can just go try direct RNAseq and expect any results yet, (2 papers total), But what does seem to be working a bit better is creating cDNA libraries; which increases yield, makes it more doable, and comes with all the caveats of these methods (PCR bias, no more modification detection). Major obvious diff (advantage?) compared to PacBio is the reduced PCR and reduced duplicity during sequencing.

# DIRECT RNASEQ: LIBRARY PREPARATION

## Variations

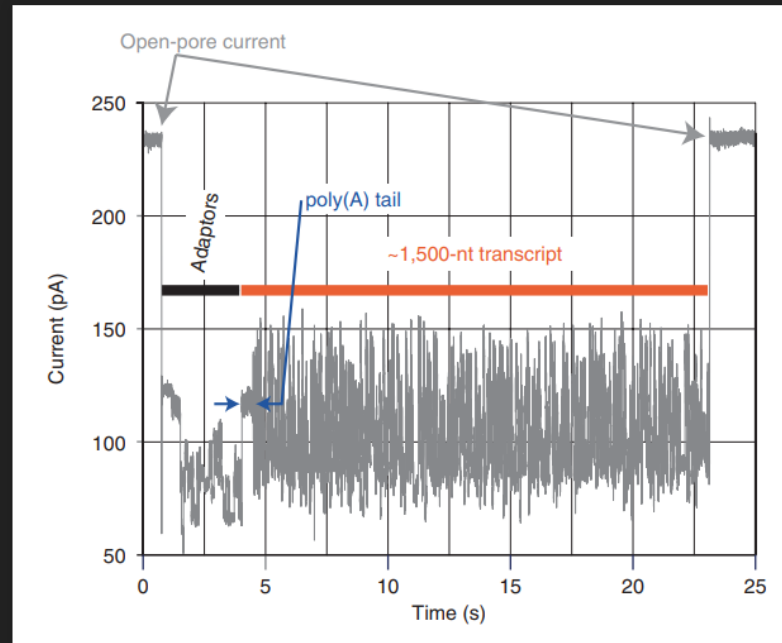
- sequence cDNA (2x yield)
- amplify and sequence cDNA (5x yield)
- targeted adapters

### Speaker notes

It's unclear to me if one can just go try direct RNAseq and expect any results yet, (2 papers total), But what does seem to be working a bit better is creating cDNA libraries; which increases yield, makes it more doable, and comes with all the caveats of these methods (PCR bias, no more modification detection). Major obvious diff (advantage?) compared to PacBio is the reduced PCR and reduced duplicity during sequencing.

# DIRECT RNASEQ: DATA PROCESSING

Raw data is a trace of the current over time

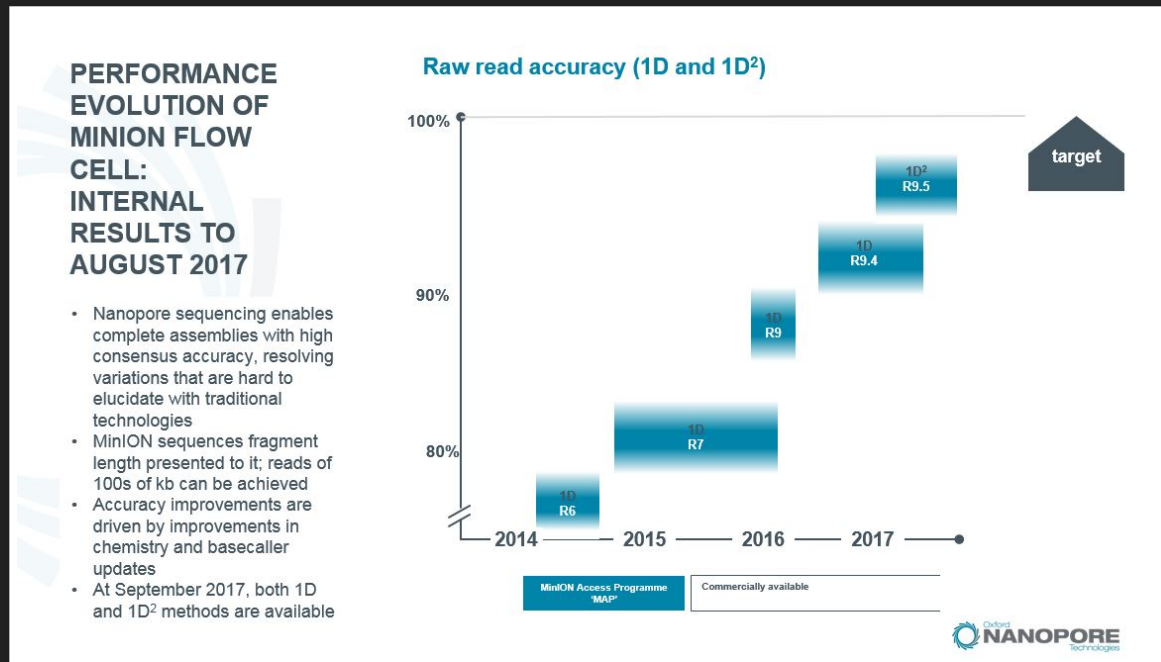


source: Geralde et al., 2017

Base calling can be performed with ONT's Recurrent Neural Network base-caller: Albacore

# DIRECT RNASEQ: DATA PROCESSING

Advertised quality looks better these days

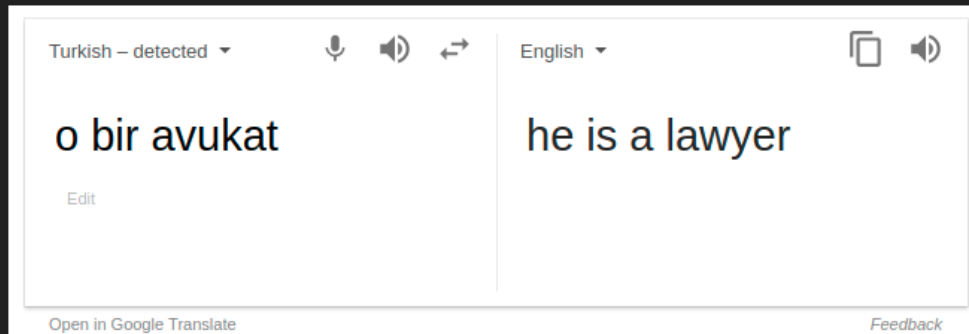
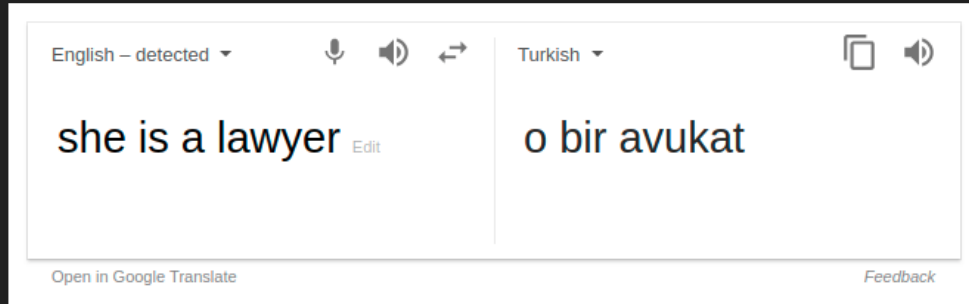


## Speaker notes

But from what I know of (at least) plant studies with nanopore, the theoretical values that were trained on amplified ("easy") DNA have not been realized. For instance, we got 1D values closer to 80% accuracy for the wild tomato genome.

# DIRECT RNASEQ: DATA PROCESSING

## Training data bias

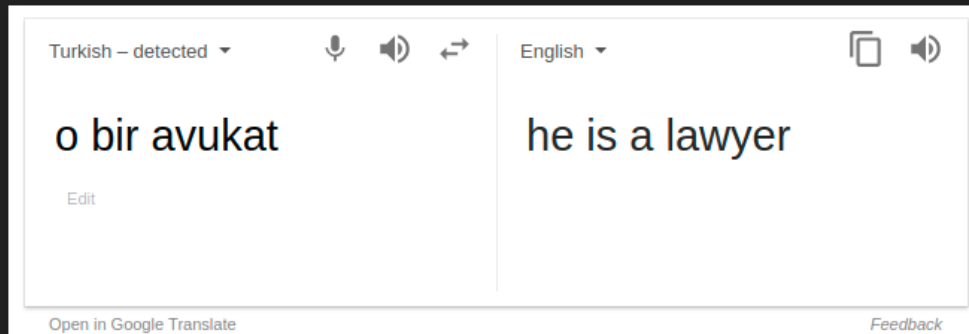
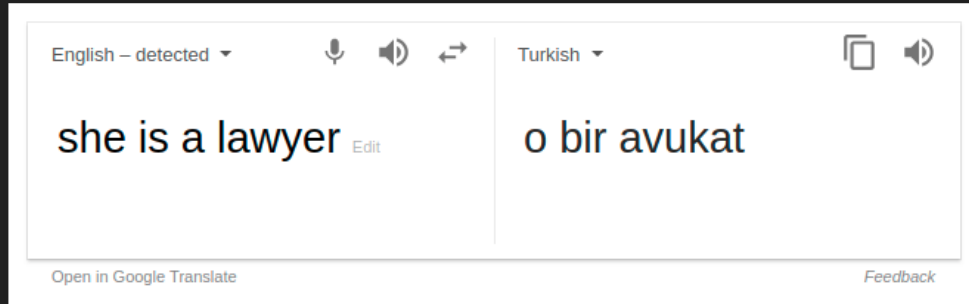


### Speaker notes

Throughout machine learning any bias in the training data is transferred into bias in the final model. E.g. Turkish doesn't genderize a simple term like, 'they are a doctor', so the model has to choose when translating it back, and it chooses that which is most common in the training data. The same idea is how the published accuracy by ONT is not reflected in a dataset very different from what ONT is training on.

# DIRECT RNASEQ: DATA PROCESSING

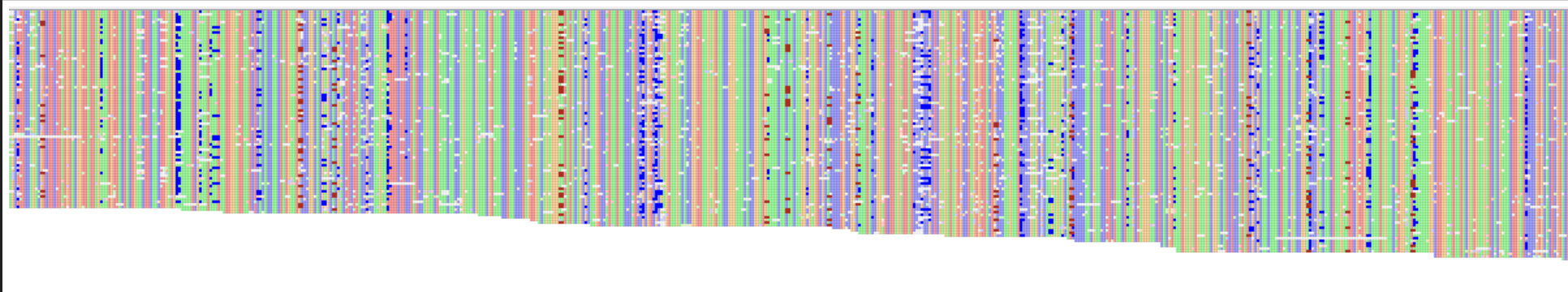
## Training data bias



### Speaker notes

Throughout machine learning any bias in the training data is transferred into bias in the final model. E.g. Turkish doesn't genderize a simple term like, 'they are a doctor', so the model has to choose when translating it back, and it chooses that which is most common in the training data. The same idea is how the published accuracy by ONT is not reflected in a dataset very different from what ONT is training on.

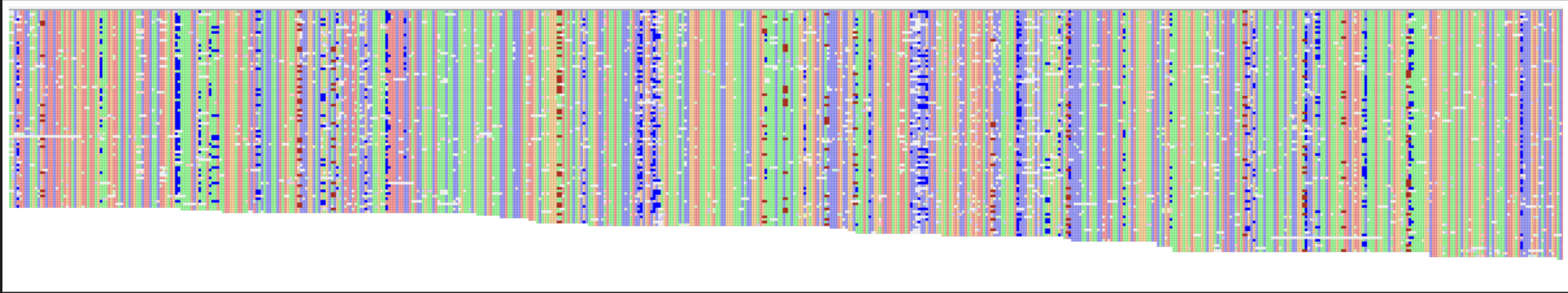
# DIRECT RNASEQ: DATA PROCESSING



## Speaker notes

Contrast with mapped subreads from Iso-seq. InDels show up in vertical lines. Since the error is non-random, self-polishing is not a real option, and data must be complemented by something without a biased error model. e.g. polished with Illumina data, or simply mapped to a trustable genome.

# DIRECT RNASEQ: DATA PROCESSING

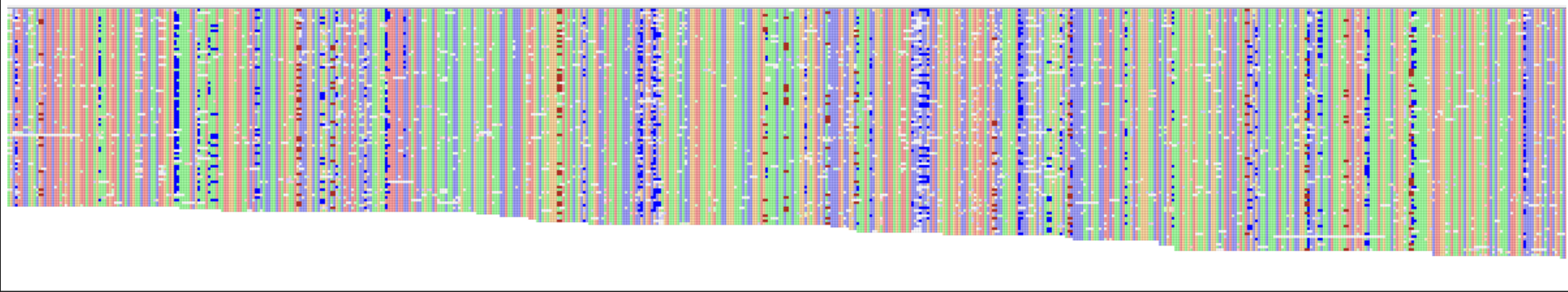


Error is not just high, but non-random

## Speaker notes

Contrast with mapped subreads from Iso-seq. InDels show up in vertical lines. Since the error is non-random, self-polishing is not a real option, and data must be complemented by something without a biased error model. e.g. polished with Illumina data, or simply mapped to a trustable genome.

# DIRECT RNASEQ: DATA PROCESSING



Error is not just high, but non-random

ONT cannot self-polish to acceptable accuracy for  
*most de novo purposes*

## Speaker notes

Contrast with mapped subreads from Iso-seq. InDels show up in vertical lines. Since the error is non-random, self-polishing is not a real option, and data must be complemented by something without a biased error model. e.g. polished with Illumina data, or simply mapped to a trustable genome.

# TO TRY ISO-SEQ OR DIRECT RNASEQ?

Further description

Sequencing Platform Info

Google-doc of costs

# TO TRY ISO-SEQ OR DIRECT RNASEQ?

Both have read lengths >> transcript lengths.

## Iso-Seq:

- lower, random, error.
- 10x subreads / "real" read.
- Has semi-established analysis pipelines.

## direct RNAseq:

- higher biased error.
- Potentially no PCR necessary.
- Potentially detect base modifications.

### Speaker notes

Trade error vs yield. Iso-Seq being more established is of course a huge advantage for beginners or where costs are manageable (e.g. targeted / well funded). But genomes can clean up the error, so Nanopore is looking really good for gene annotation if one is ready to proceed without being able to check seq answers for everything yet.

# TO TRY ISO-SEQ OR DIRECT RNASEQ?

Both have read lengths >> transcript lengths.

## Iso-Seq:

- lower, random, error.
- 10x subreads / "real" read.
- Has semi-established analysis pipelines.

## direct RNAseq:

- higher biased error.
- Potentially no PCR necessary.
- Potentially detect base modifications.

### Speaker notes

Trade error vs yield. Iso-Seq being more established is of course a huge advantage for beginners or where costs are manageable (e.g. targeted / well funded). But genomes can clean up the error, so Nanopore is looking really good for gene annotation if one is ready to proceed without being able to check seq answers for everything yet.

**THANK YOU FOR YOUR ATTENTION!**