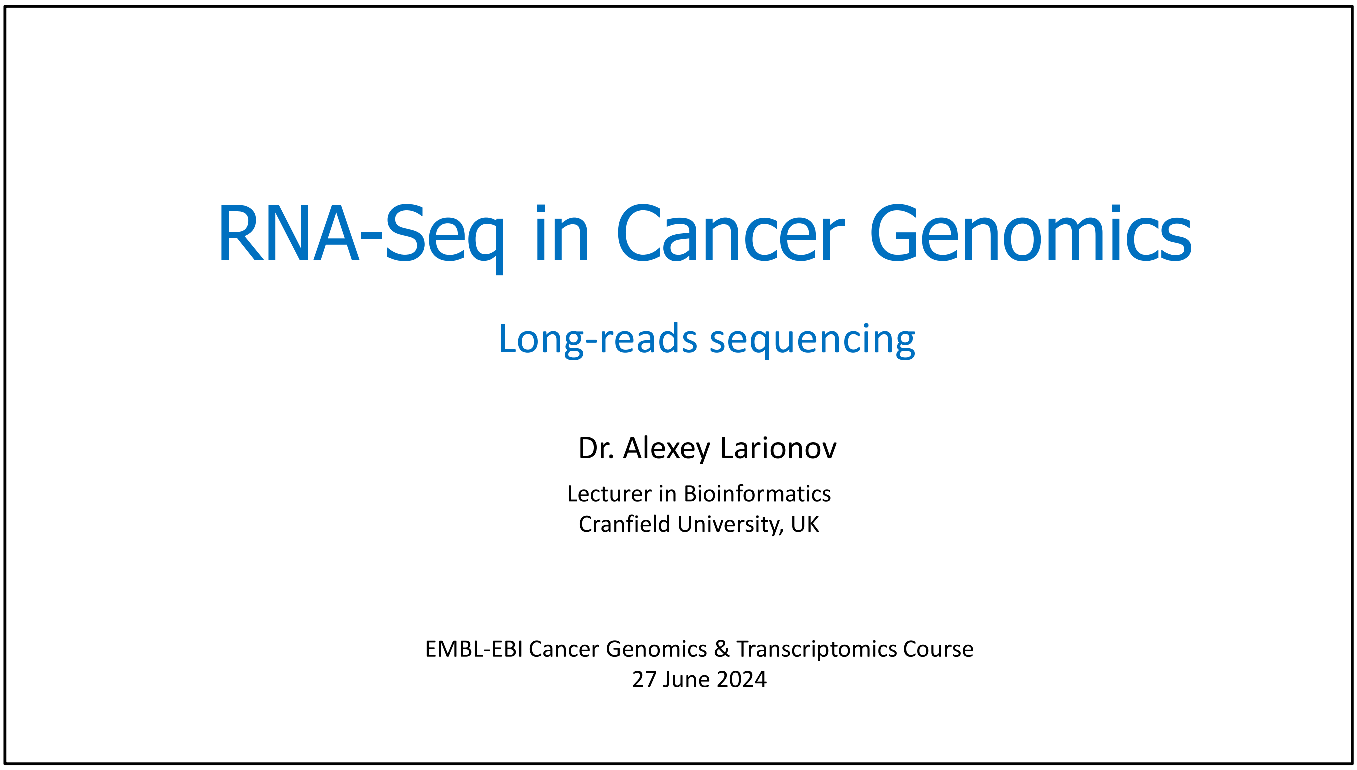


RNA-Seq in Cancer Genomics

Long-reads sequencing

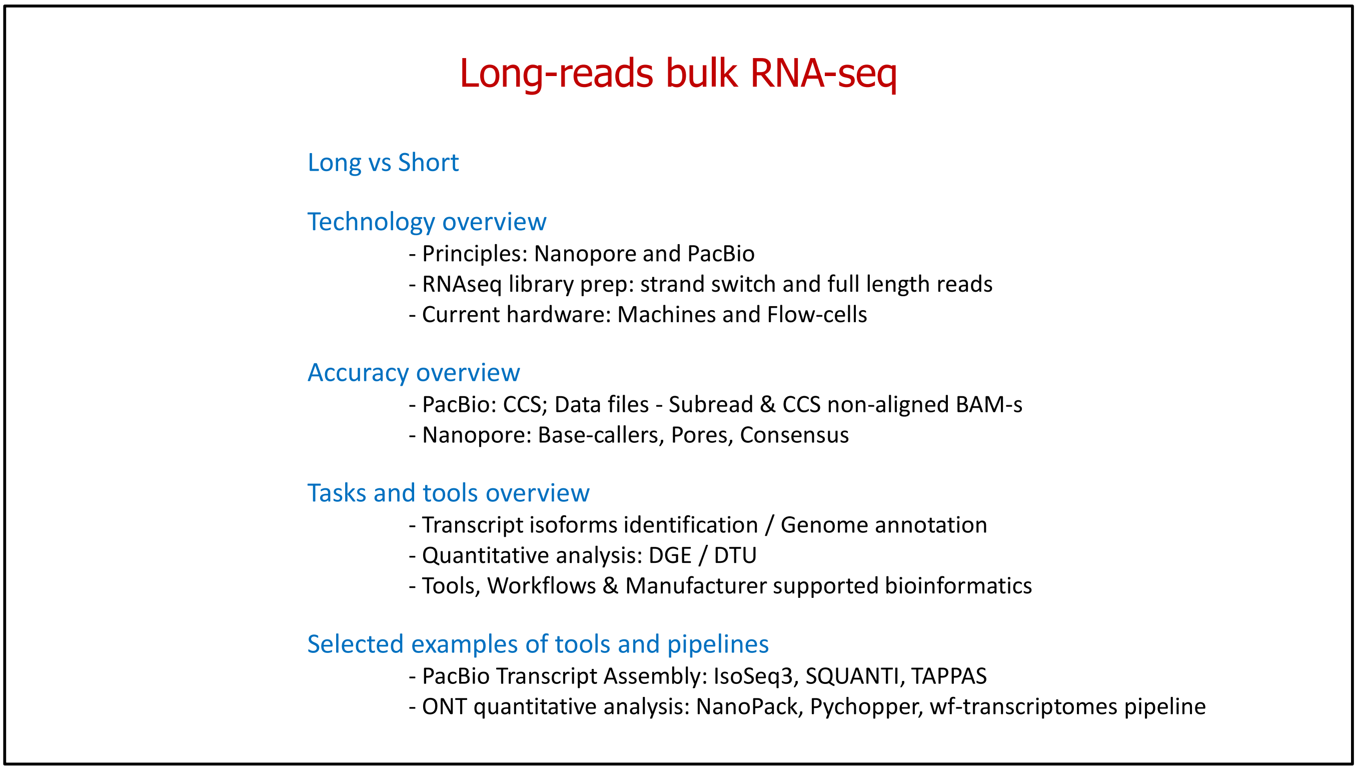
Dr. Alexey Larionov

Lecturer in Bioinformatics
Cranfield University, UK

EMBL-EBI Cancer Genomics & Transcriptomics Course
27 June 2024

Because of the broad scope of the course, and the limited time, it will be a high-level review of long-read RNA sequencing.

The participants who already have experience with long-reads RNA-seq are very welcome to contribute their comments along the lecture.



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

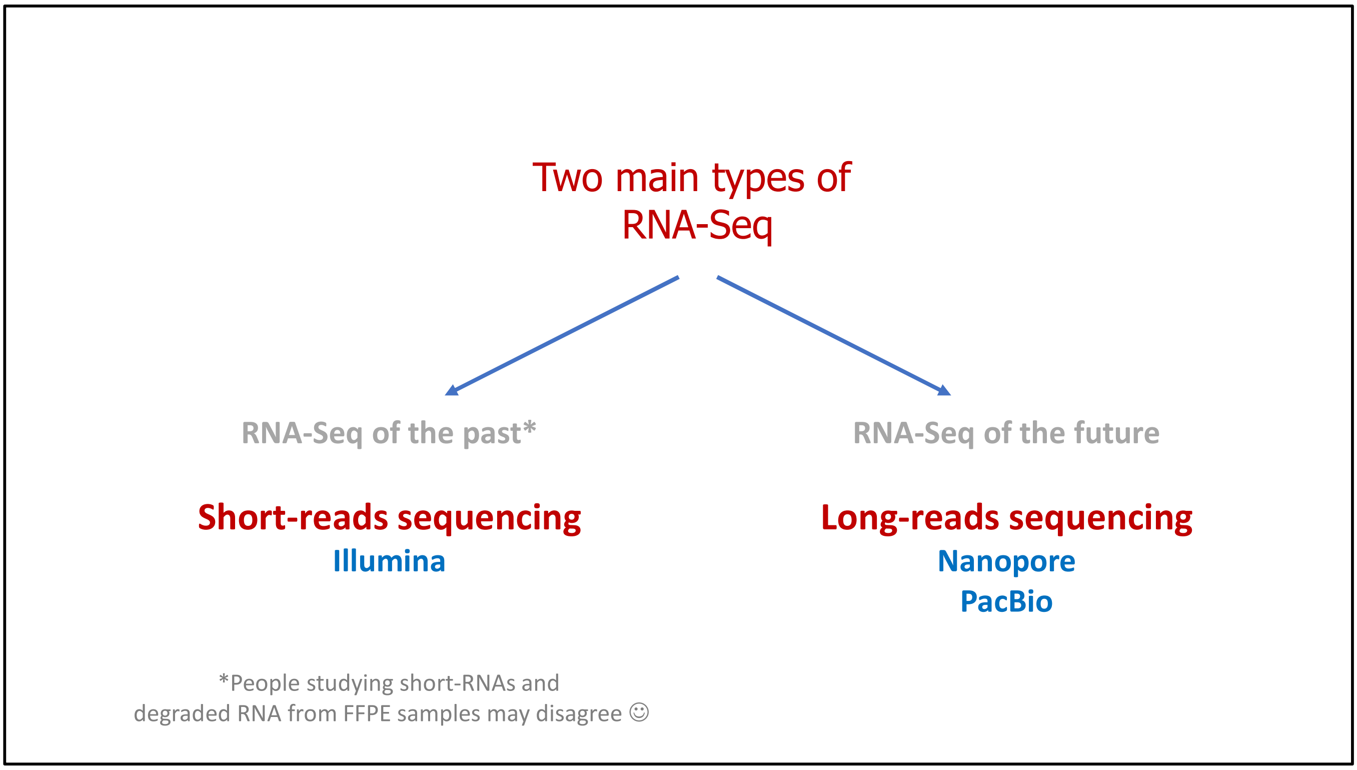
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, wf-transcriptomes pipeline

I will start with a high-level overview of the technologies, then focus on some RNA-seq specific aspects, and then will go to bioinformatics aspects discussing tools, tasks and some examples.



Two main types of RNA-Seq

RNA-Seq of the past*

Short-reads sequencing

Illumina

RNA-Seq of the future

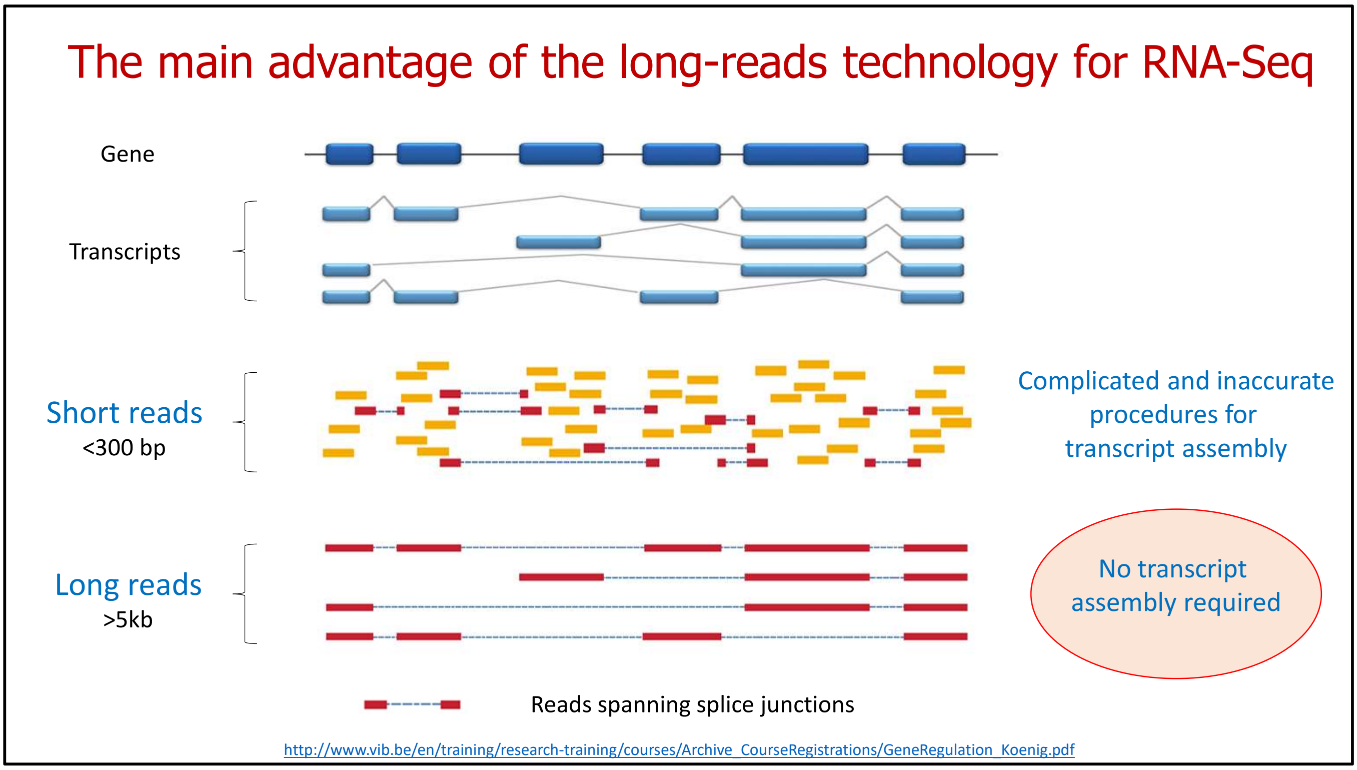
Long-reads sequencing

Nanopore

PacBio

*People studying short-RNAs and degraded RNA from FFPE samples may disagree ☺

I already mentioned today that there are two main types of RNA-sequencing: the short-read and the long-read sequencing :)



The main advantage of the long-reads technology for RNA-Seq

Gene

Transcripts

Short reads
<300 bp

Complicated and inaccurate procedures for transcript assembly

Long reads
>5kb

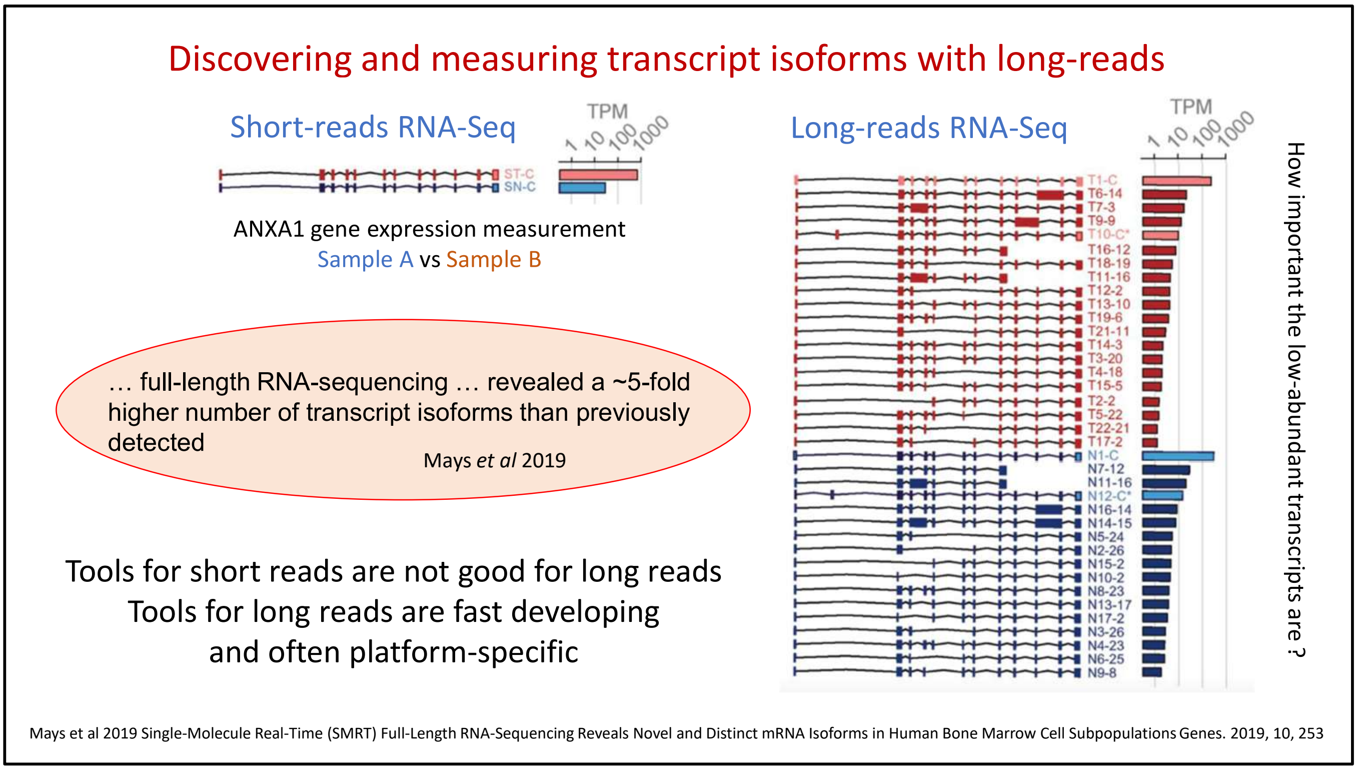
No transcript assembly required

Reads spanning splice junctions

http://www.vib.be/en/training/research-training/courses/Archive_CourseRegistrations/GeneRegulation_Koenig.pdf

The main advantage of the long-reads technology for RNA-Seq is that the long reads can span the entire transcripts, eliminating the need in the transcripts assembly (the bottleneck of the short-reads technology).

Of course, long-reads also have their limitations. Thus, later we will discus potential issues with the accuracy of some long-reads sequencing techniques. So, it is worth noting here, that the high accuracy of sequencing is not needed for RNA-seq gene/transcript expression measurement. Two or three errors per hundred nucleotides, do not complicate identifying the transcript. The length is much more important.



Discovering and measuring transcript isoforms with long-reads

Short-reads RNA-Seq

ANXA1 gene expression measurement
Sample A vs Sample B

… full-length RNA-sequencing … revealed a ~5-fold higher number of transcript isoforms than previously detected

Mays *et al* 2019

Tools for short reads are not good for long reads
Tools for long reads are fast developing
and often platform-specific

Long-reads RNA-Seq

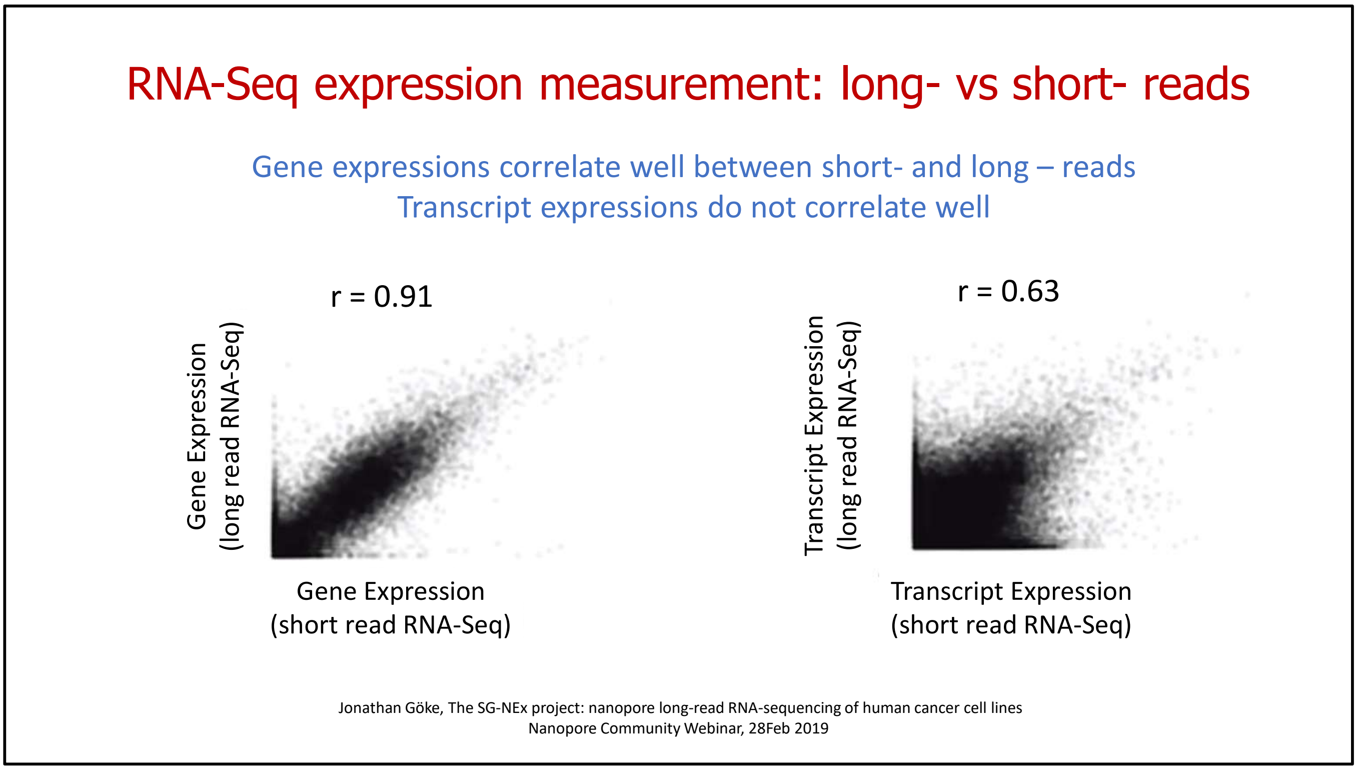
How important the low-abundant transcripts are ?

Mays et al 2019 Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct mRNA Isoforms in Human Bone Marrow Cell Subpopulations Genes. 2019, 10, 253

Because of the ability to read the entire transcripts, the long-reads RNA-seq reveals several times more transcripts than it was previously detectable by the short reads.

Interestingly, many of the new detected transcripts are low-abundant. So, we still don’t know their biologic relevance.

Another question is:
If so many transcripts were missed by the short reads, does it mean that all the previous results obtained with the short reads were wrong ?



RNA-Seq expression measurement: long- vs short- reads

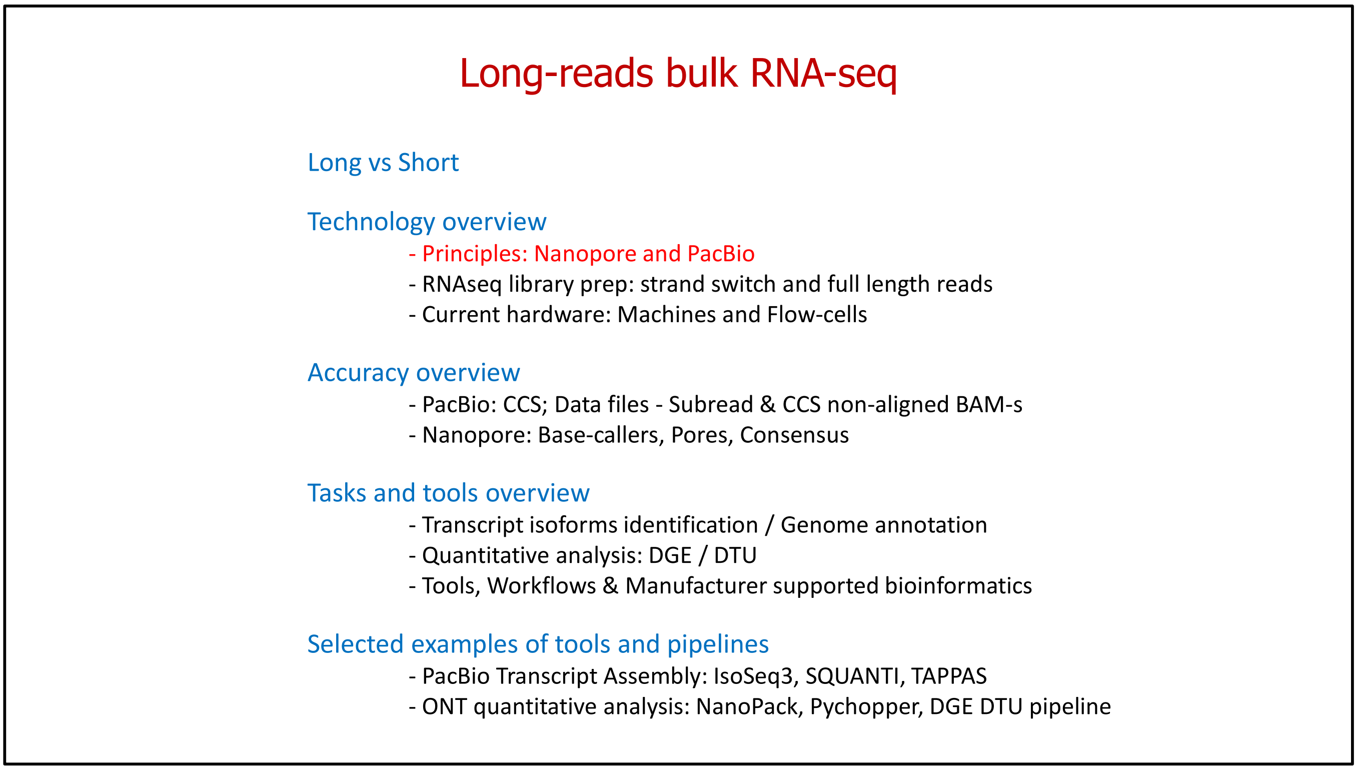
Gene expressions correlate well between short- and long – reads
Transcript expressions do not correlate well

Jonathan Göke, The SG-NEx project: nanopore long-read RNA-sequencing of human cancer cell lines
Nanopore Community Webinar, 28Feb 2019

No, it doesn’t.

On the left panel you can see that after aggregation per gene the results of short-reads RNA-sequencing correlate well with the long-reads. Of course, the right panel shows much worse correlation at the TRANSCRIPT level.

The point is that most of the gene expression results reported so far from the short reads were reported per gene, not per transcript. So, most of the previously published short-reads results are valid. Of course, the rare short-reads TRANSCRIPT-specific data might need to be validated.



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

Tasks and tools overview

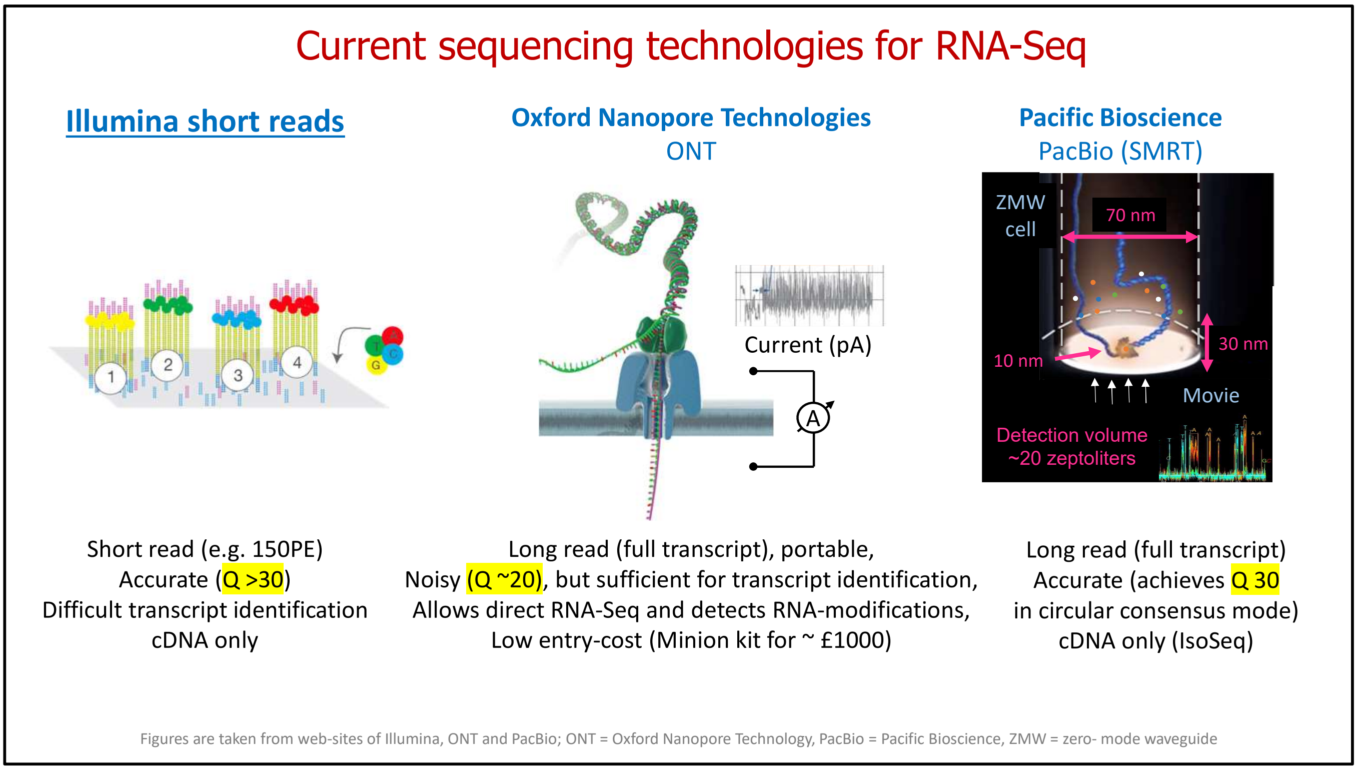
- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Now, lets review the current long reads sequencing technologies.
At the moment, there are two long-read sequencing technologies: Oxford Nanopore and Pacific Bioscience.

We will not discuss Synthetic long reads here, except for a single slide to acknowledge the existence of an alternative to the true long-reads technologies.



Current sequencing technologies for RNA-Seq

Illumina short reads

Short read (e.g. 150PE)
Accurate (Q >30)
Difficult transcript identification
cDNA only

Oxford Nanopore Technologies
ONT

Long read (full transcript), portable,
Noisy (Q ~20), but sufficient for transcript identification,
Allows direct RNA-Seq and detects RNA-modifications,
Low entry-cost (Minion kit for ~ £1000)

Pacific Bioscience
PacBio (SMRT)

Long read (full transcript)
Accurate (achieves Q 30
in circular consensus mode)
cDNA only (IsoSeq)

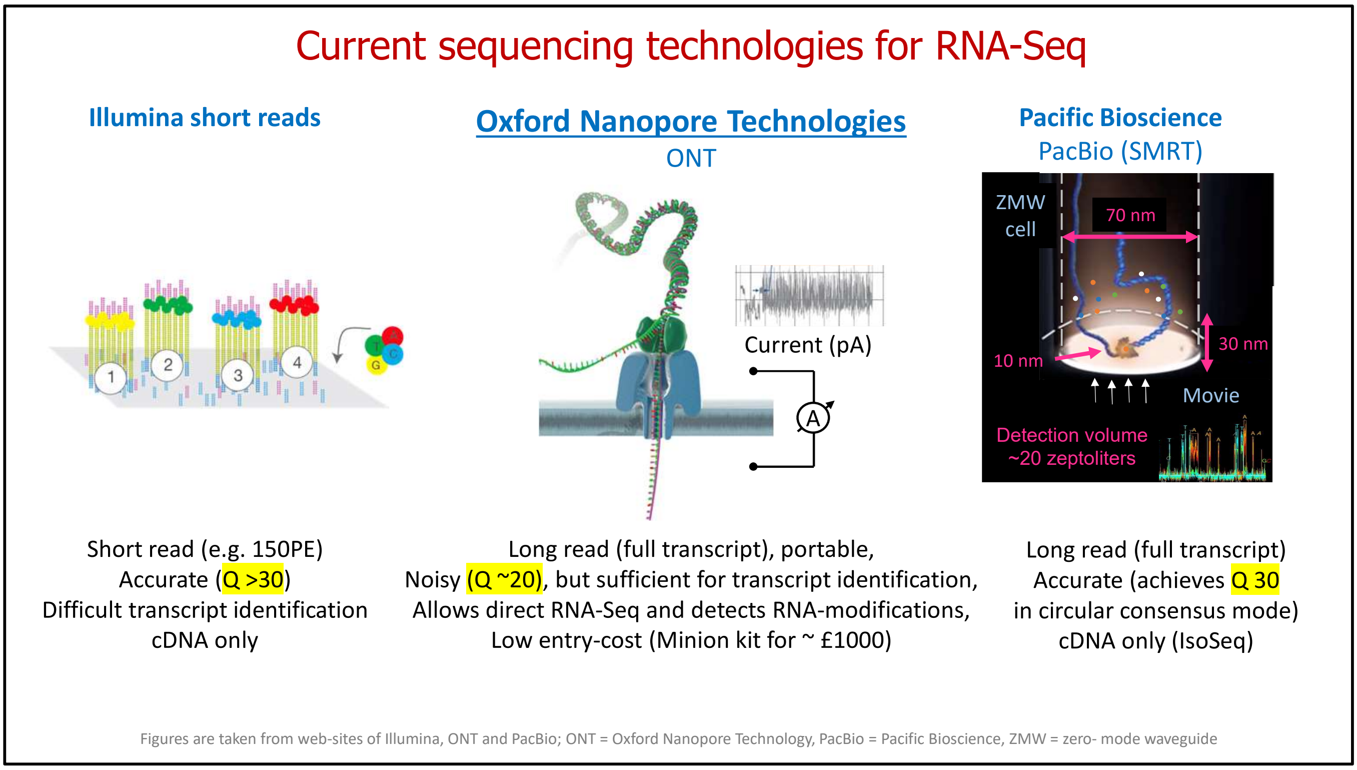
Figures are taken from web-sites of Illumina, ONT and PacBio; ONT = Oxford Nanopore Technology, PacBio = Pacific Bioscience, ZMW = zero- mode waveguide

I included the Illumina short-reads technology on this slide (on the left) to explain why it can not produce long reads.

The Illumuna sequencing is based on amplifying each initial DNA fragment to a cluster, and then adding one nucleotide per cycle, with different nucleotides coded by different colors. There are very nice videos on the Illumina web site, which explain the technology in great details.

Unfortunately, some DNAs in clusters occasionally miss a cycle, and after 2 or 3 hundred cycles the molecules in the clusters are going out of sync. So, this technology can not read longer fragments.

The advantage of the Illumina technology is that, until the clusters go out of phase, the sequencing is very accurate: less than one error in a thousand of bases.



Current sequencing technologies for RNA-Seq

Illumina short reads

Oxford Nanopore Technologies
ONT

Pacific Bioscience
PacBio (SMRT)

Short read (e.g. 150PE)
Accurate (Q >30)
Difficult transcript identification
cDNA only

Long read (full transcript), portable,
Noisy (Q ~20), but sufficient for transcript identification,
Allows direct RNA-Seq and detects RNA-modifications,
Low entry-cost (Minion kit for ~ £1000)

Long read (full transcript)
Accurate (achieves Q 30
in circular consensus mode)
cDNA only (IsoSeq)

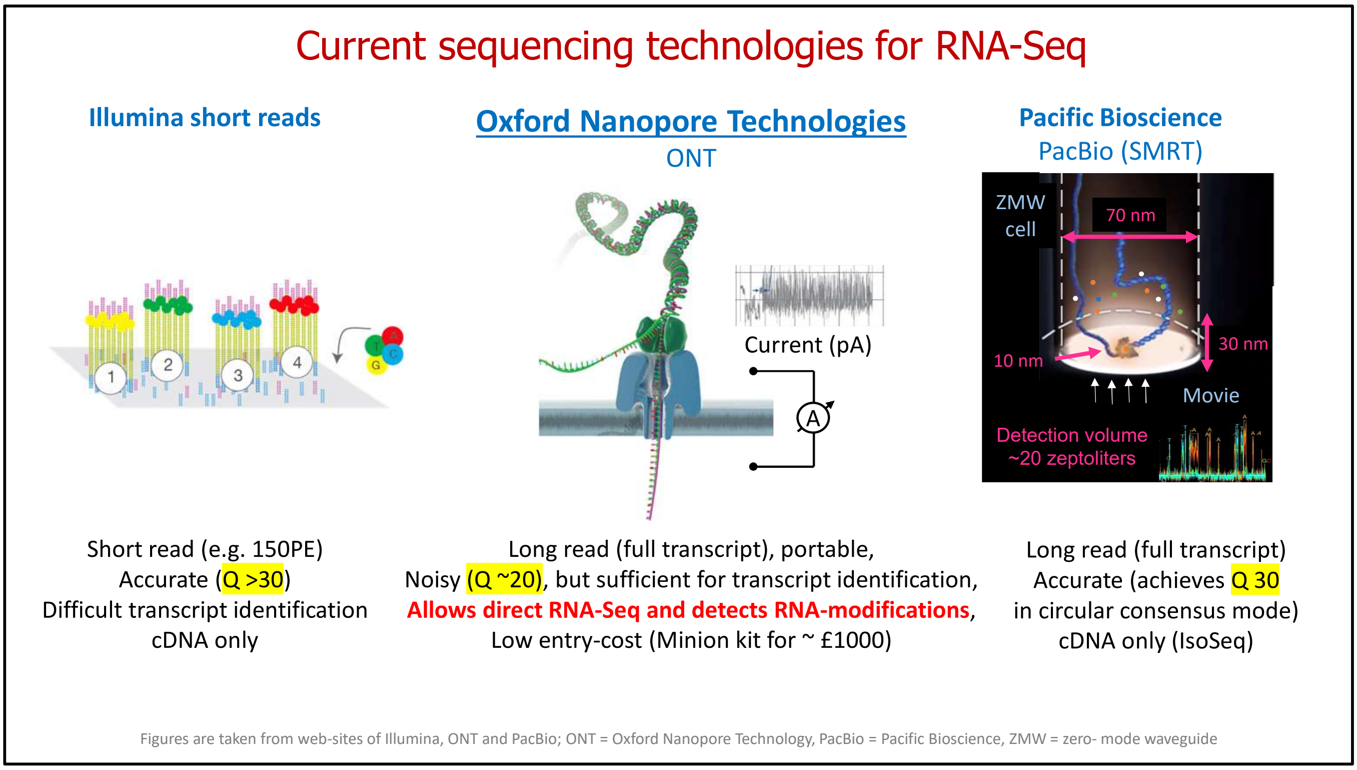
Figures are taken from web-sites of Illumina, ONT and PacBio; ONT = Oxford Nanopore Technology, PacBio = Pacific Bioscience, ZMW = zero- mode waveguide

The Nanopore technology is shown here in the middle.

It just passes DNA fragment through a pore, while measuring the ion current passing through the same pore at the same time. Because different nucleotides have different size and charge, they block the ions' passage in a different way, and so they have distinct current signatures.

The length of Nanopore sequencing is limited just by the length of the fragment, reaching hundreds of thousand or even millions of bases.

The negative side of this technology is that it is still much less accurate than the short-reads sequencing: with some errors per each hundred of nucleotides. However, for many RNAseq applications such accuracy is absolutely sufficient.



Current sequencing technologies for RNA-Seq

Illumina short reads

Oxford Nanopore Technologies

ONT

Pacific Bioscience

PacBio (SMRT)

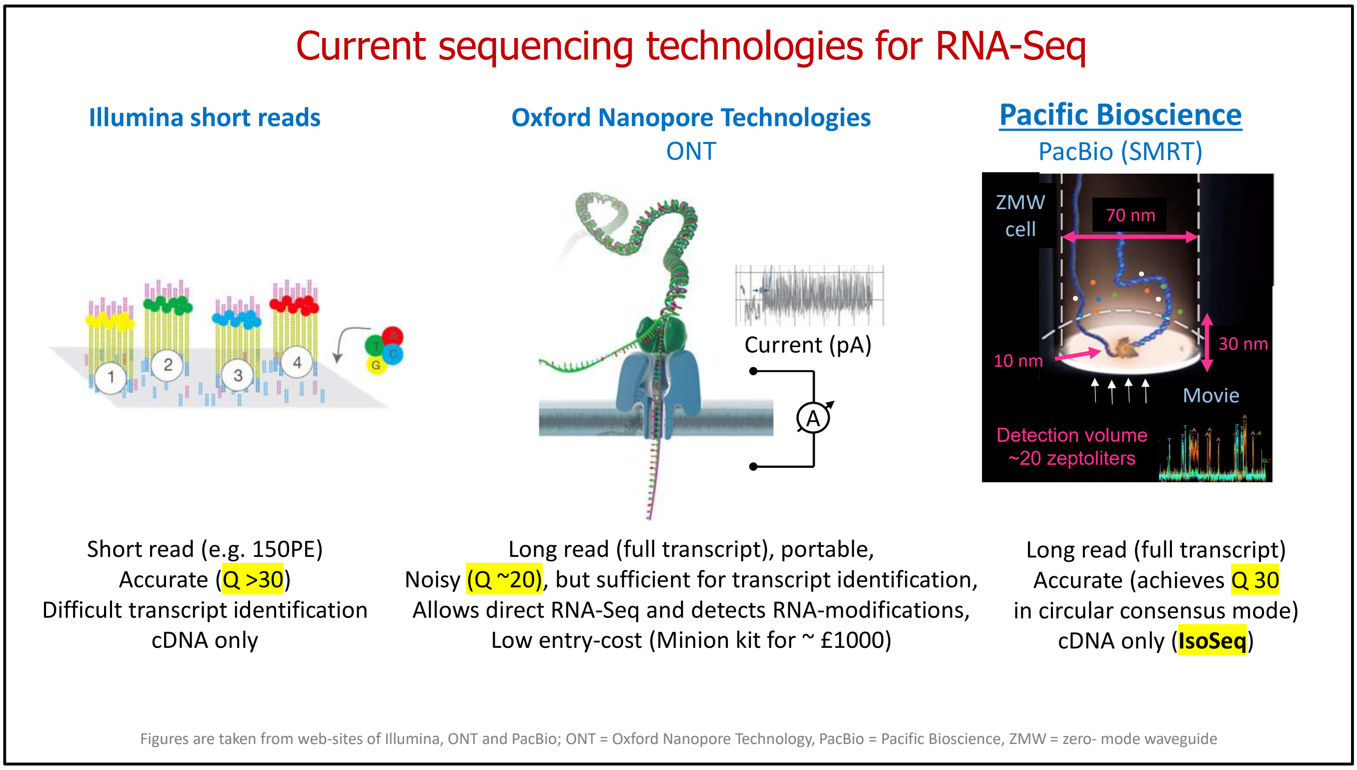
Short read (e.g. 150PE)
Accurate (Q >30)
Difficult transcript identification
cDNA only

Long read (full transcript), portable,
Noisy (Q ~20), but sufficient for transcript identification,
Allows direct RNA-Seq and detects RNA-modifications,
Low entry-cost (Minion kit for ~ £1000)

Long read (full transcript)
Accurate (achieves Q 30
in circular consensus mode)
cDNA only (IsoSeq)

Figures are taken from web-sites of Illumina, ONT and PacBio; ONT = Oxford Nanopore Technology, PacBio = Pacific Bioscience, ZMW = zero- mode waveguide

Importantly, the nanopore can sequence RNA directly, without converting it to cDNA before sequencing.



Current sequencing technologies for RNA-Seq

Illumina short reads

Oxford Nanopore Technologies
ONT

Pacific Bioscience
PacBio (SMRT)

Short read (e.g. 150PE)
Accurate (Q >30)
Difficult transcript identification
cDNA only

Long read (full transcript), portable,
Noisy (Q ~20), but sufficient for transcript identification,
Allows direct RNA-Seq and detects RNA-modifications,
Low entry-cost (Minion kit for ~ £1000)

Long read (full transcript)
Accurate (achieves Q 30
in circular consensus mode)
cDNA only (**IsoSeq**)

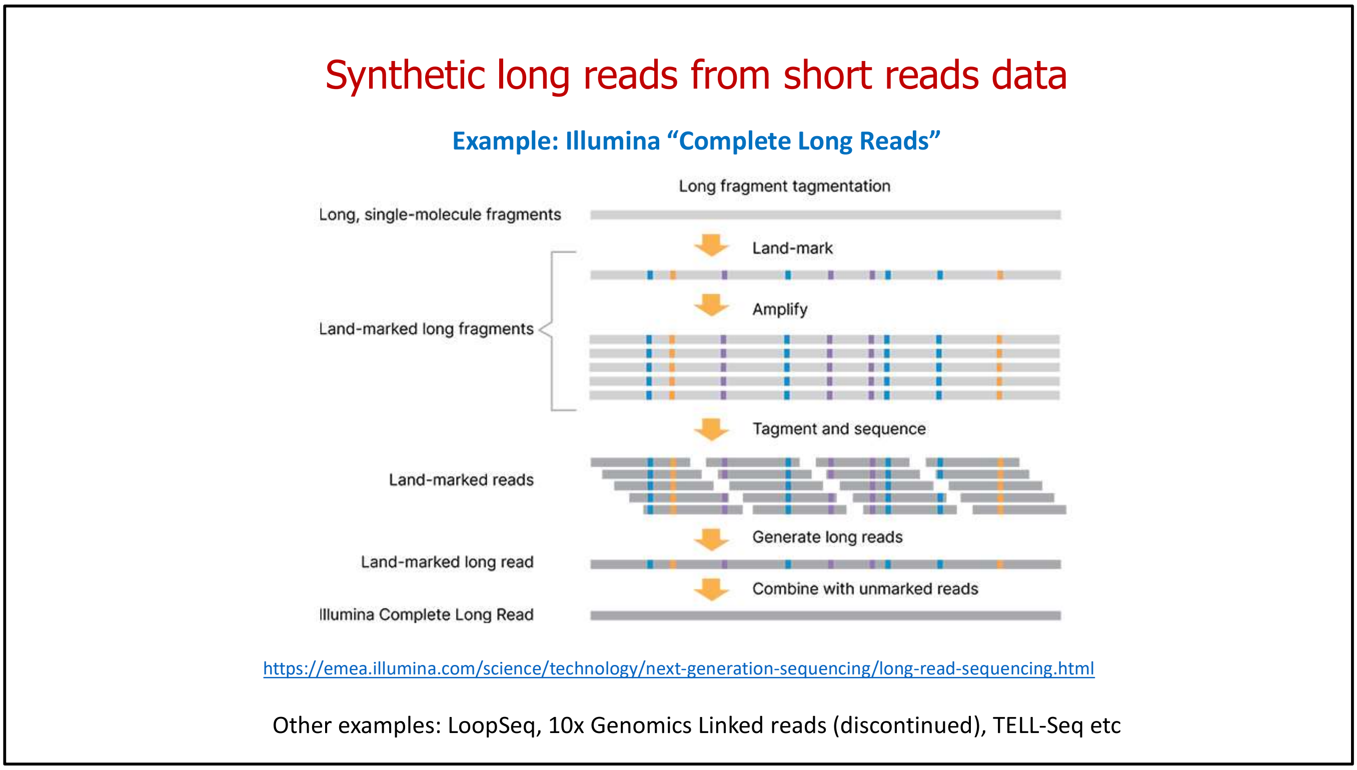
Figures are taken from web-sites of Illumina, ONT and PacBio; ONT = Oxford Nanopore Technology, PacBio = Pacific Bioscience, ZMW = zero- mode waveguide

Finally, the PacBio technology is shown on the right. PacBio is also marketed as SMRT sequencing, meaning that it reads Single Molecule in Real Time.

They place a single DNA-polymerase at the bottom of a tiny well. Then the DNA synthesis is just filmed by a tiny camera in real time. Different nucleotides are labelled by different colors. When a new nucleotide is being added to the DNA, the DNA-polymerase retains it for a certain time, which is recorded as a peak in the movie.

To make such movie possible, it's important to exclude filming of the other nucleotides, which are not being retained by DNA-polymerase, but still are present in the solution. This is made by illuminating only a very small volume around the DNA-polymerase. Because of the Brownian movement, the non-retained nucleotides quickly cross the volume, so they and are not registered in the movie. Selective illumination of such a small volume is done by some physical miracle, called Zero-Mode Waveguide (ZMW). Somehow, when the diameter of the well is small enough, the light does not go through the well, but only propagates to a certain depth. A well with about 70 nm in diameter allows to illuminate a volume of just 20 zeptoliters.

Honestly, illuminating and filming a single DNA-polymerase molecule in action sounds like SciFi to me. However, somehow it works.



Synthetic long reads from short reads data

Example: Illumina "Complete Long Reads"

Long fragment tagmentation
Long, single-molecule fragments
Land-mark
Land-marked long fragments
Amplify
Tagment and sequence
Land-marked reads
Generate long reads
Land-marked long read
Combine with unmarked reads
Illumina Complete Long Read

https://emea.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html

Other examples: LoopSeq, 10x Genomics Linked reads (discontinued), TELL-Seq etc

Finally, I would like to mention that there are alternatives to the true long-reads sequencing technologies.

For instance, Illumina provides a solution, when special "Land-marks" are placed to long DNA fragments before they are shredded and sequenced as short reads.

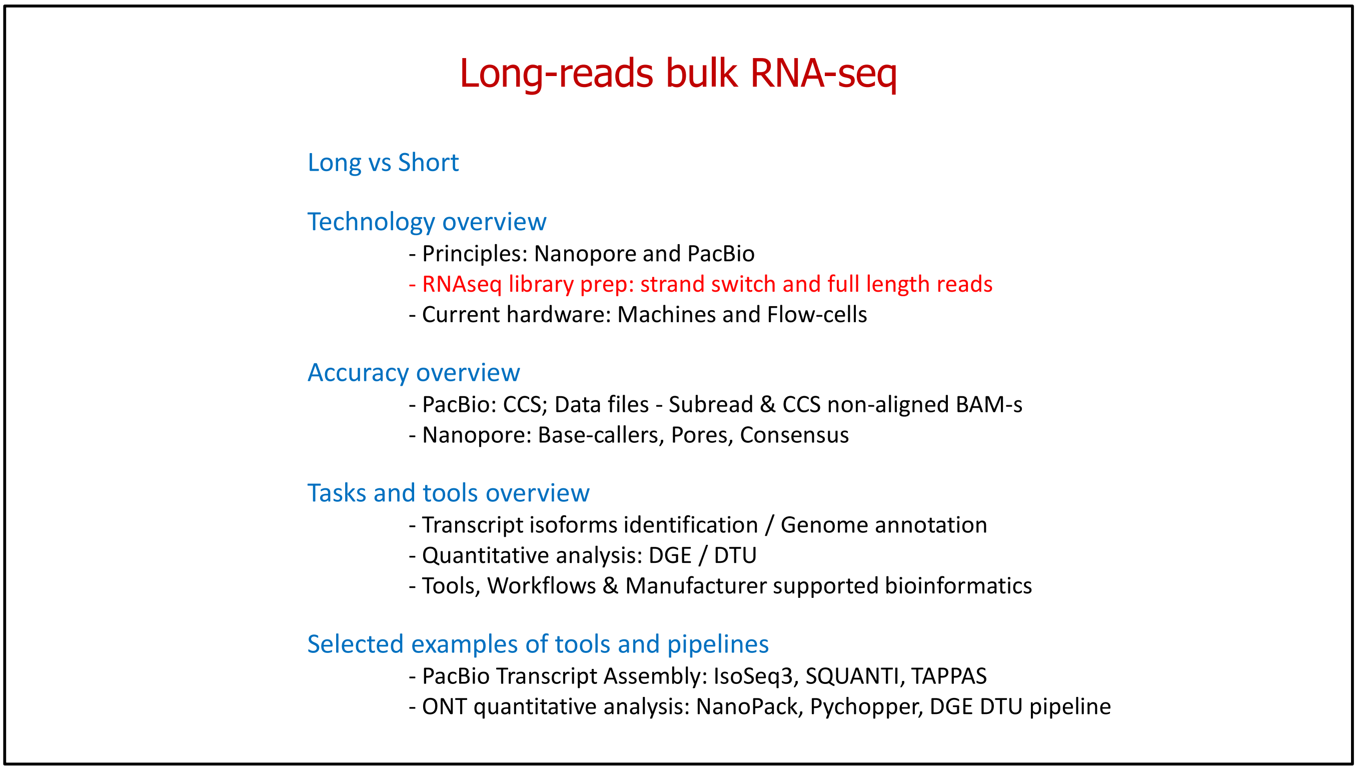
Then the initial long DNA fragments can be computationally re-assembled from the "land-marks" information.

There are other similar technologies that allow to assemble synthetic long reads from the short reads data.

This could be a cheaper alternative to the true long-reads technologies: the future will show.

I will not discuss Synthetic Long Reads in this lecture later. You may see a comparison between PacBio and Synthetic Long Reads following this link:

https://www.pacb.com/blog/the-hifi-difference-true-long-reads-vs-synthetic-long-reads/



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

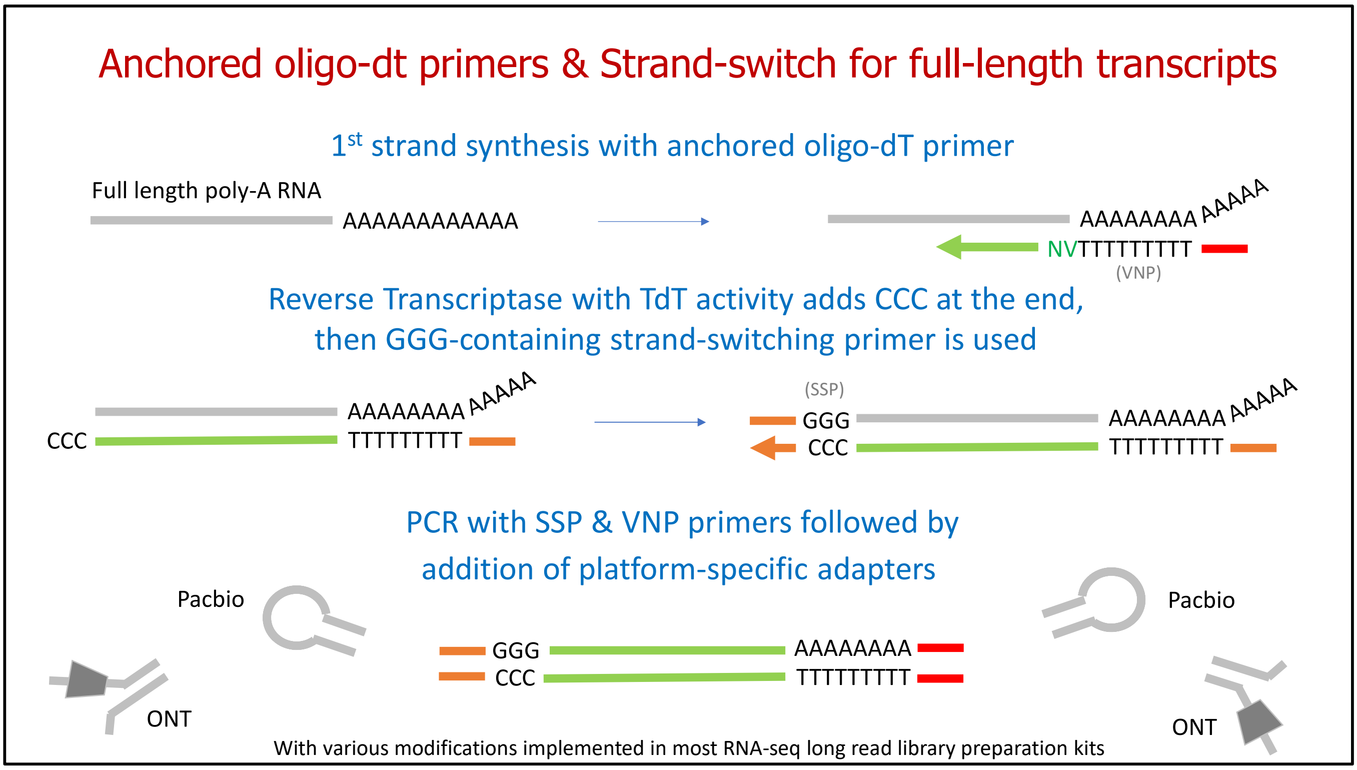
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

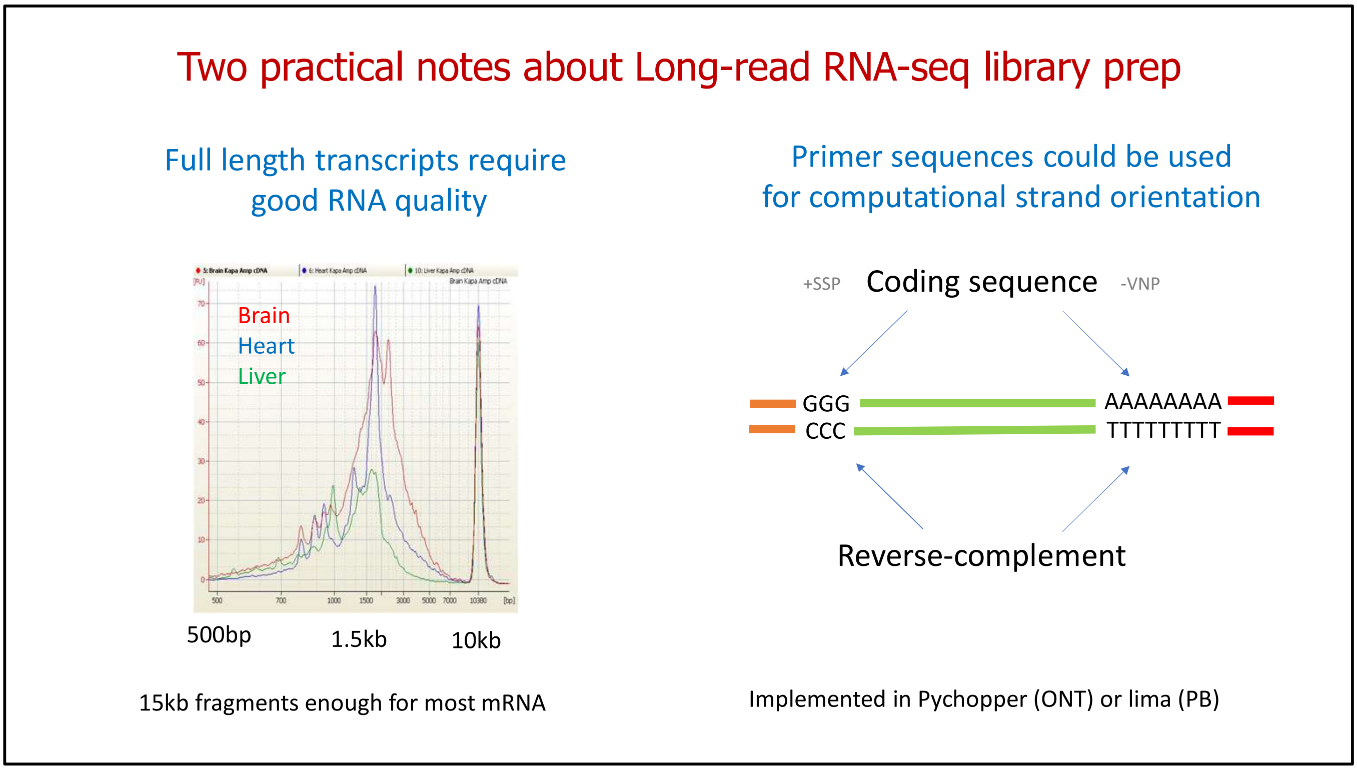
- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Now let’s talk about RNA-seq library preparation for the true long-reads sequencing technologies: ONT and PB.


Anchored oligo-dt primers & Strand-switch for full-length transcripts
1st strand synthesis with anchored oligo-dT primer
Full length poly-A RNA
AAAAAAAAAAAA
AAAAAAAA AAAAA
NVTTTTTTTTT
(VNP)
Reverse Transcriptase with TdT activity adds CCC at the end,
then GGG-containing strand-switching primer is used
CCC
AAAAAAAA AAAAA
TTTTTTTTT
(SSP)
GGG
CCC
AAAAAAAA AAAAA
TTTTTTTTT
PCR with SSP & VNP primers followed by
addition of platform-specific adapters
Pacbio
ONT
GGG
CCC
AAAAAAAA
TTTTTTTTT
Pacbio
ONT
With various modifications implemented in most RNA-seq long read library preparation kits

This is a *simplified* scheme explaining the widely used technique for making cDNA libraries for full-length transcripts.

SSP: Strand Switching Primer (sometime called template-switching)
VNP: V = A or G or C ; N = A or G or C or T ; P= Primer
TdT : Terminal deoxynucleotidyl Transferase



Two practical notes about Long-read RNA-seq library prep

Full length transcripts require good RNA quality

15kb fragments enough for most mRNA

Primer sequences could be used for computational strand orientation

+SSP Coding sequence -VNP

GGG AAAAAAAA

CCC TTTTTTTTT

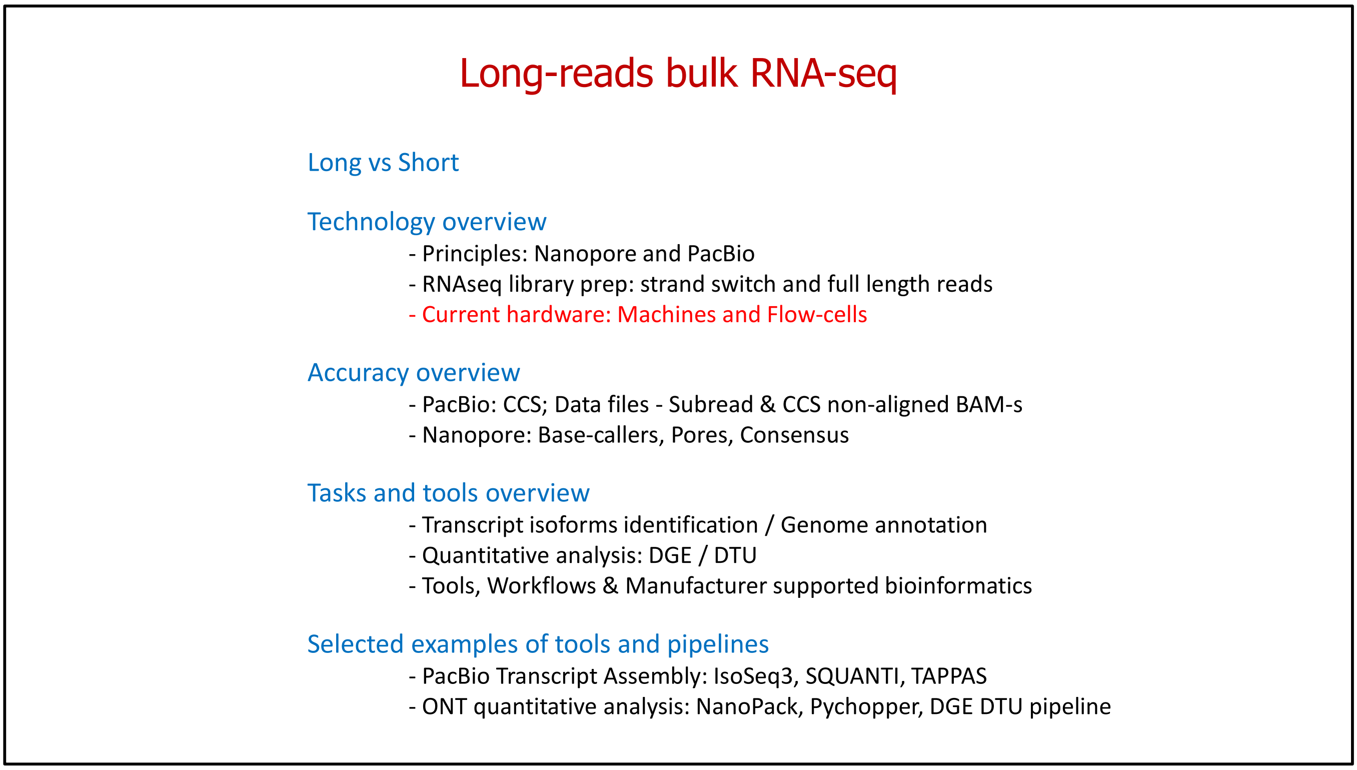
Reverse-complement

Implemented in Pychopper (ONT) or lima (PB)

Of course, to make the full-length cDNA libraries, the full-length RNA molecules should be present in the extract in the first place.

If the non-degraded RNA was used, then 15kb read length should be enough to sequence most of human transcripts in any tissue.

An additional advantage of using the Strand-switching technique described in the previous slide is that it allows computational detection of the strand orientation. So, all the long-reads libraries are stranded: the coding sequence is flanked by ***direct*** SSP and ***reverse-complemented*** oligo-dT primers



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

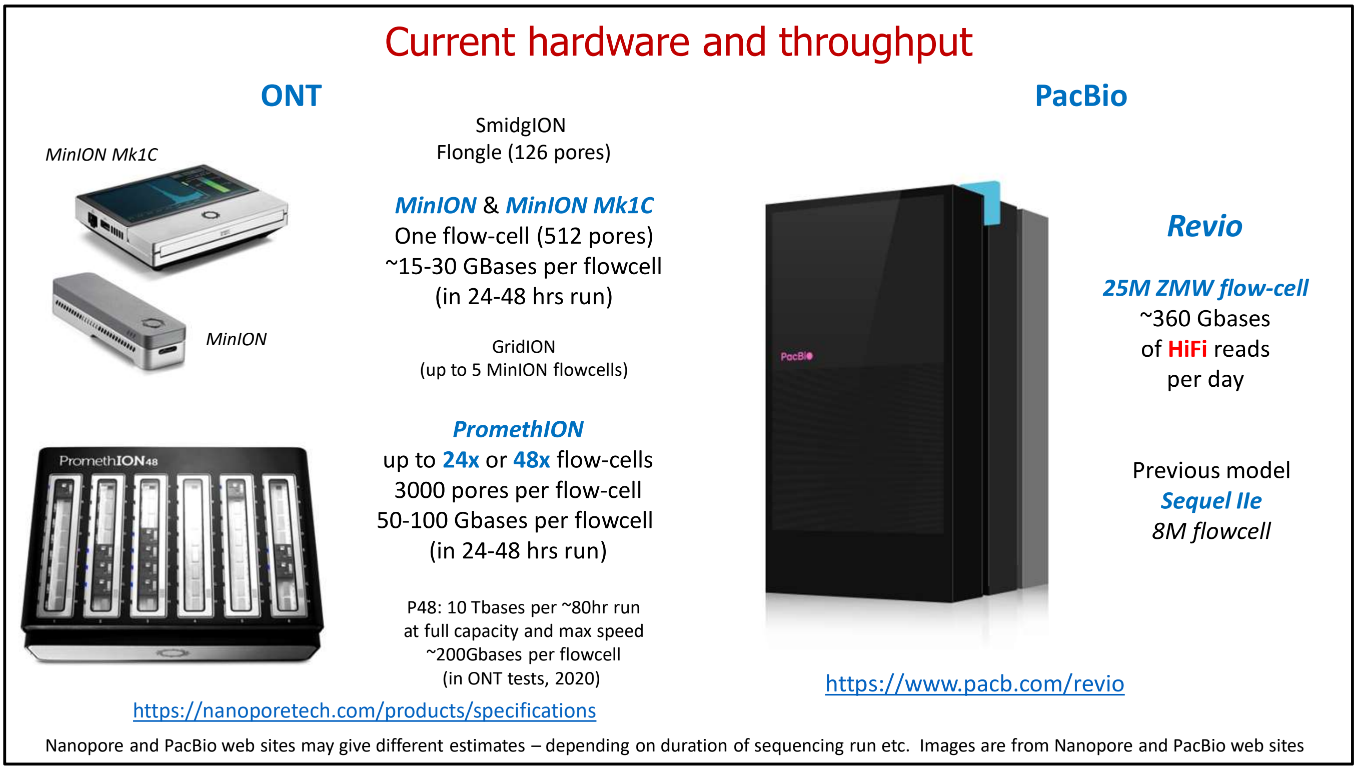
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Then the prepared libraries go to the sequencing machines.



Current hardware and throughput

ONT

MinION Mk1C

MinION

SmidgION
Flongle (126 pores)

MinION & ***MinION Mk1C***
One flow-cell (512 pores)
~15-30 GBases per flowcell
(in 24-48 hrs run)

GridION
(up to 5 MinION flowcells)

PromethION
up to **24x** or **48x** flow-cells
3000 pores per flow-cell
50-100 Gbases per flowcell
(in 24-48 hrs run)

P48: 10 Tbases per ~80hr run
at full capacity and max speed
~200Gbases per flowcell
(in ONT tests, 2020)

https://nanoporetech.com/products/specifications

PacBio

Revio

25M ZMW flow-cell
~360 Gbases
of **HiFi** reads
per day

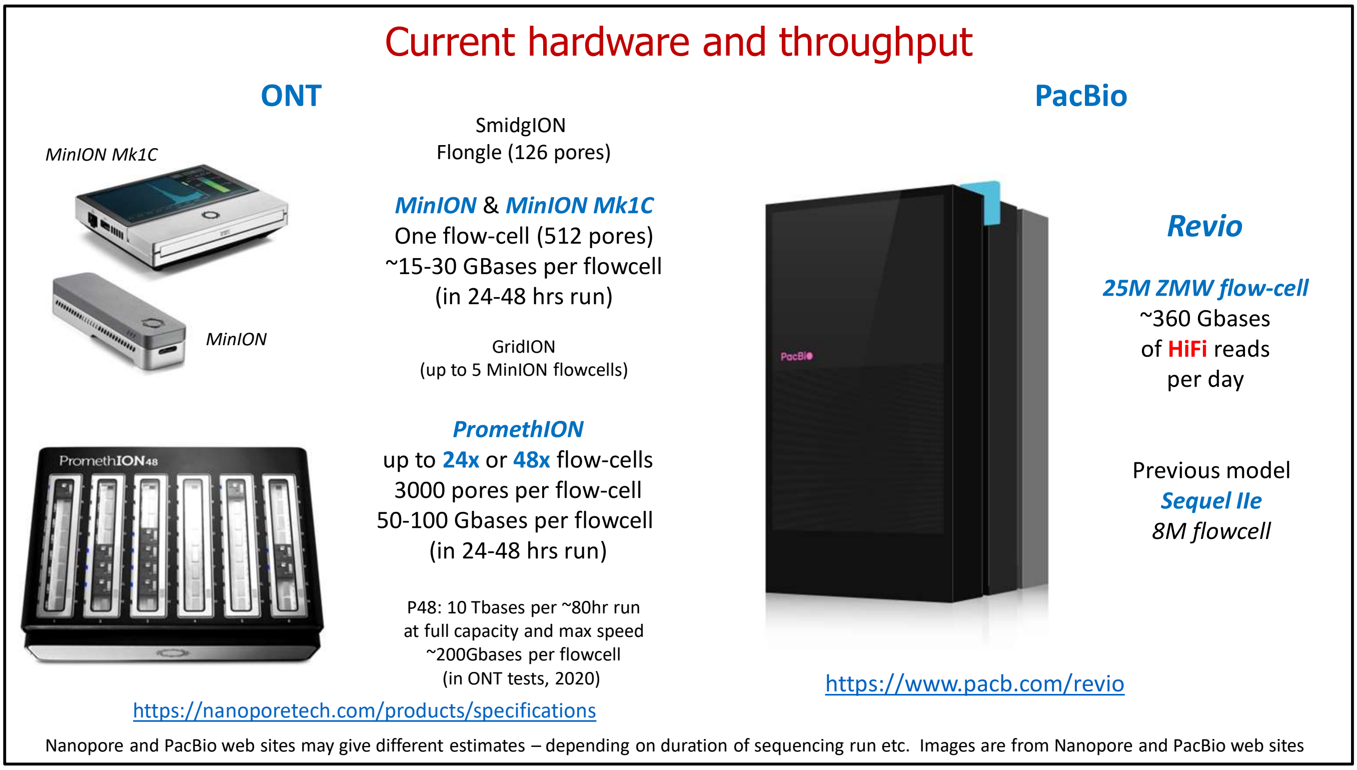
Previous model
Sequel IIe
8M flowcell

https://www.pacb.com/revio

Nanopore and PacBio web sites may give different estimates – depending on duration of sequencing run etc. Images are from Nanopore and PacBio web sites

There are many models in Nanopore.

For a long time, the entry-level ONT model was **Minion**. It’s a size of 10x2x3 cm, required connection to a laptop by USB cabel.
MinION Mk1C uses the same flowcell as Minion but doesn’t need a laptop and includes on-board basecalling.
The highest end of Nanopore is **PromethION**, it has much larger flowcells and up to 48 of them.
1x PromethION flow-cell may allow 50-75x human genome may reach >30 consensus accuracy (caveats discussed later).
The theoretical throughput of 48 flowcell Promethion run at full speed and max capacity is mind-blowing.
There are intermediate and smaller models.



Current hardware and throughput

ONT

MinION Mk1C

MinION

SmidgION
Flongle (126 pores)

MinION & ***MinION Mk1C***
One flow-cell (512 pores)
~15-30 GBases per flowcell
(in 24-48 hrs run)

GridION
(up to 5 MinION flowcells)

PromethION
up to **24x** or **48x** flow-cells
3000 pores per flow-cell
50-100 Gbases per flowcell
(in 24-48 hrs run)

P48: 10 Tbases per ~80hr run
at full capacity and max speed
~200Gbases per flowcell
(in ONT tests, 2020)

https://nanoporetech.com/products/specifications

PacBio

Revio

25M ZMW flow-cell
~360 Gbases
of **HiFi** reads
per day

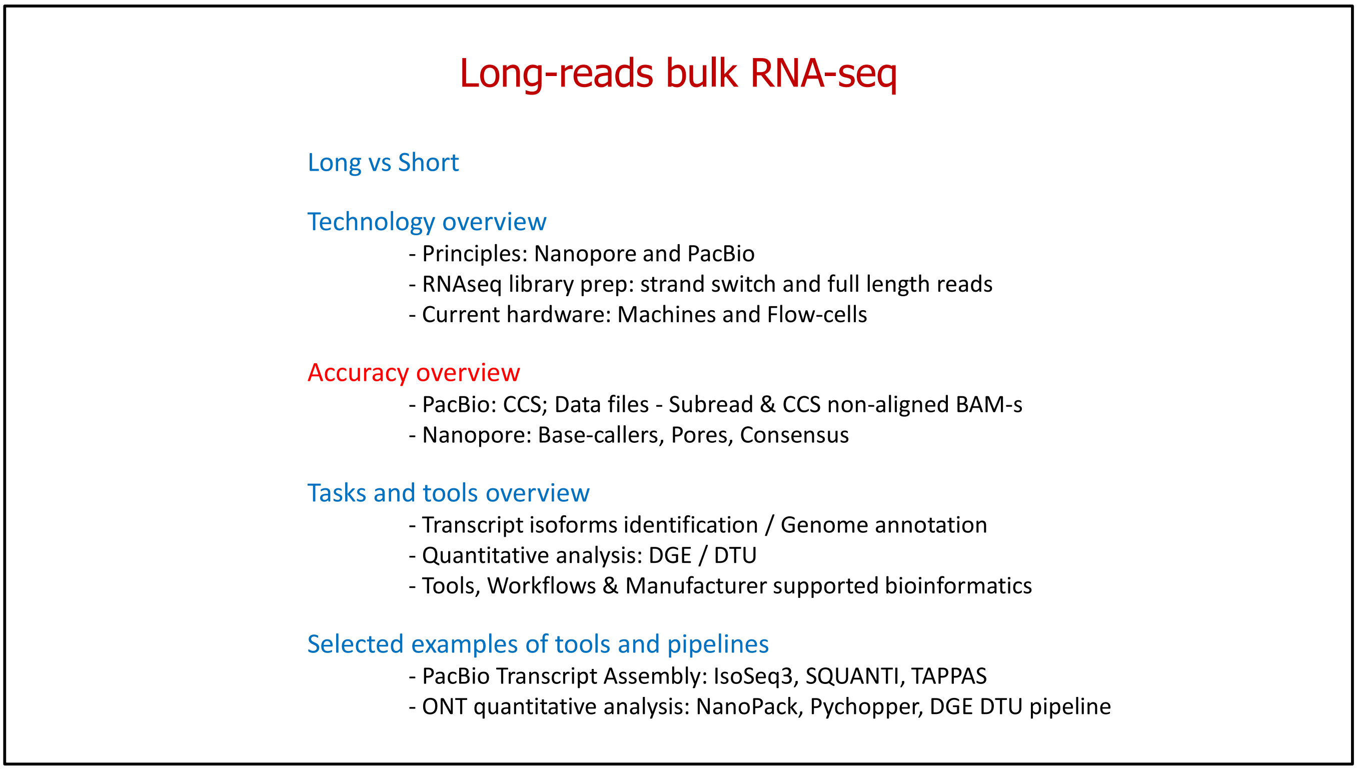
Previous model
Sequel IIe
8M flowcell

https://www.pacb.com/revio

Nanopore and PacBio web sites may give different estimates – depending on duration of sequencing run etc. Images are from Nanopore and PacBio web sites

The recently launched (in 2023) "latest and greatest" model of PacBio is **Revio**.

It was a large step forward against the previous models.
Importantly PacBio can produce HiFi reads – their accuracy could be even higher than in Illumina (discussed later).
Although PacBio is still more expensive than ONT per base, **Revio** already allows to sequence human genome for USD1000 (with long accurate phased reads). Combined with **Kinnex** library preparation (a new technique discussed later) **Revio** may make PacBio RNA-seq even more affordable.



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

Tasks and tools overview

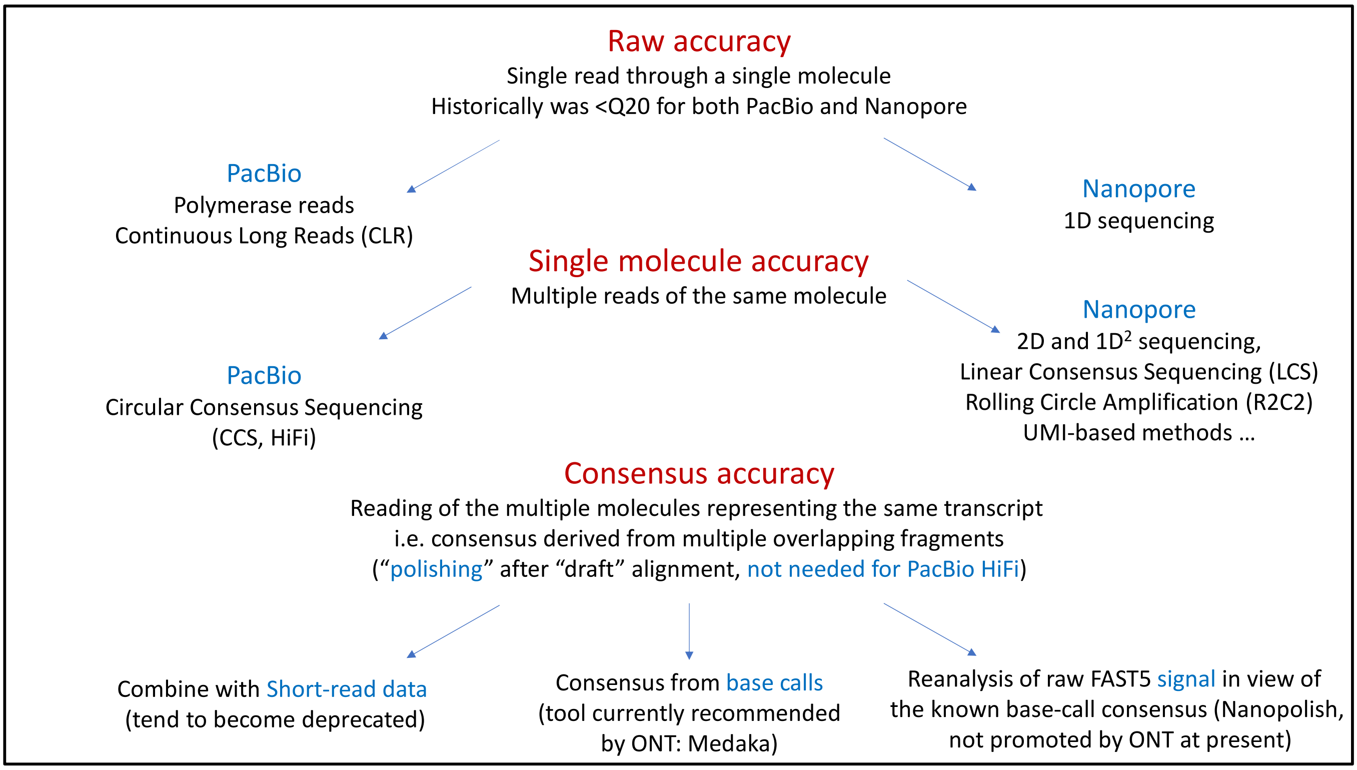
- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

I have already mentioned previously “consensus accuracy” or “HiFi reads”.

These terms relate to the ways how long reads sequencing technologies improve the accuracy of their reads.



Raw accuracy
Single read through a single molecule
Historically was <Q20 for both PacBio and Nanopore

PacBio
Polymerase reads
Continuous Long Reads (CLR)

Nanopore
1D sequencing

Single molecule accuracy
Multiple reads of the same molecule

PacBio
Circular Consensus Sequencing
(CCS, HiFi)

Nanopore
2D and 1D^2 sequencing,
Linear Consensus Sequencing (LCS)
Rolling Circle Amplification (R2C2)
UMI-based methods …

Consensus accuracy
Reading of the multiple molecules representing the same transcript
i.e. consensus derived from multiple overlapping fragments
(“polishing” after “draft” alignment, not needed for PacBio HiFi)

Combine with Short-read data
(tend to become deprecated)

Consensus from base calls
(tool currently recommended
by ONT: Medaka)

Reanalysis of raw FAST5 signal in view of
the known base-call consensus (Nanopolish,
not promoted by ONT at present)

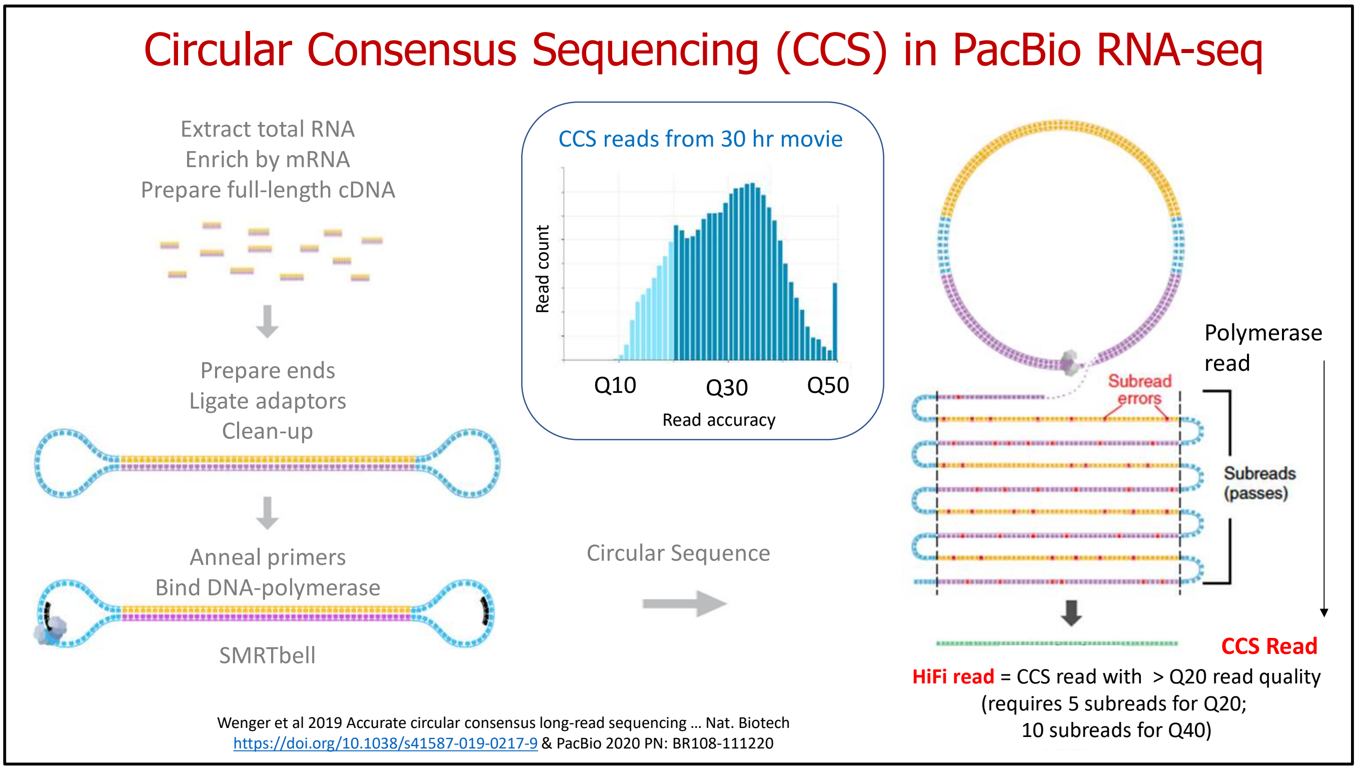
Typical ***raw*** Illumina base Quality is Q30-40 (which is less than 1 error per 1000 bases). The ***raw*** long-reads sequencing is significantly less accurate.

This slide contains many new terms. I will explain some of them now, and some later. For PacBio the ***raw*** data may be called Polymerase reads (or Continuous Long Reads, CLR). For Nanopore the ***raw*** data may be called 1D sequencing.

To improve accuracy the long-read technologies have tried to read the same sequence many times.

This solved the problem for PacBio: when a DNA fragment is circularized and red many times (as will be shown in the next slide), the Circular Consensus Reads (CCS) accuracy easily exceeds the Illumina raw base quality (if the same sequence is red for 5 and more times).

Nanopore also tried reading the same molecule twice (called 2D and 1D2). Unfortunately, because of the non-random distribution of Nanopore errors, this was less successful than in PacBio. So, Nanopore still uses consensus accuracy: reading multiple fragments from the same gene to improve the accuracy. With sufficient depth (and price) Nanopore claims that it may match Illumina accuracy (as will be shown later). However, it’s still a work in progress (in 2024).



Circular Consensus Sequencing (CCS) in PacBio RNA-seq

Extract total RNA
Enrich by mRNA
Prepare full-length cDNA

Prepare ends
Ligate adaptors
Clean-up

Anneal primers
Bind DNA-polymerase

SMRTbell

CCS reads from 30 hr movie

Circular Sequence

CCS Read

HiFi read = CCS read with > Q20 read quality
(requires 5 subreads for Q20;
10 subreads for Q40)

Wenger et al 2019 Accurate circular consensus long-read sequencing … Nat. Biotech
https://doi.org/10.1038/s41587-019-0217-9 & PacBio 2020 PN: BR108-111220

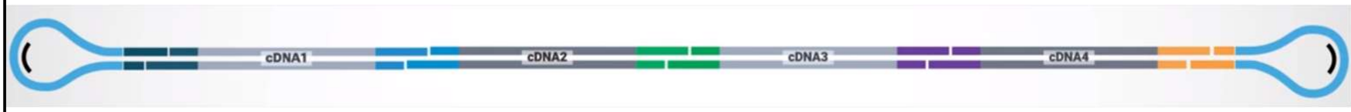
This slide explains the PacBio Circular Consensus Reads.

Because PacBio ***raw*** data (so called Polymerase Reads or Continuous Long Reads) have low accuracy (<20), PacBio came up with a nice trick to improve the accuracy. During the library preparation they make the fragment circular. Then, during the sequencing the same fragment is red again and again many times. Because the errors are random, the consensus sequence after multiple reads becomes as accurate as the Illumina short reads (or even better accuracy). The length of the circular consensus reads in PabBio may easily achieve 10-15 kilobases, which is enough for most of human full-length RNAs.

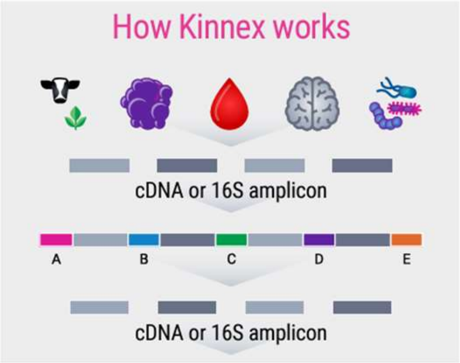
Importantly, unlike to Nanopore, PacBio can not sequence RNA directly: it has to be converted to cDNA before sequencing.

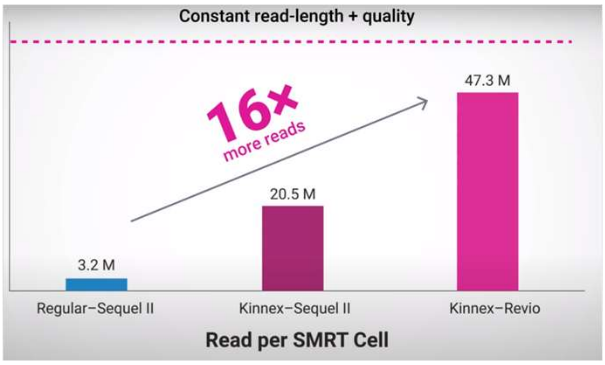
Kinnex technology in PacBio RNA-seq

A new library prep method than can significantly increase yield and decrease price of PB RNA-seq



Kinnex (formerly MAS-seq) concatenates several cDNAs into one array before HiFi sequencing





https://www.pacb.com/wp-content/uploads/Kinnex-brochure.pdf

https://www.youtube.com/watch?v=NICUp8C6rms

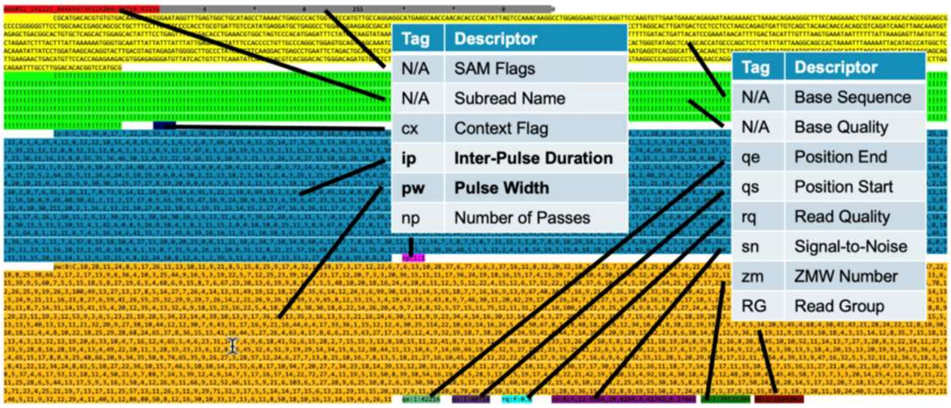
Kinnex (formerly MAS-seq) method concatenates several cDNAs into one array before taking it into CCS. This is still very new. However, combined with the high output of the Revio sequencer, this may pave the way for more affordable PacBio RNA-seq.

PacBio CCS file formats : BAM-s everywhere

- PacBio machine produces raw data as a BAM of **unaligned subreads** (with all base qual=0 & all read qual=0.8) *
- The CCS workflow produces a BAM with **unaligned consensus reads** with meaningful base & read qualities
- Alignment (mapping) programs will produce **aligned BAM** files that *retain PacBio tags*

* the latest versions of sequencers (Sequel IIe) may output the consensus BAM-s to reduce data size

Additional tags in PacBio BAM-s



Tag	Descriptor
N/A	SAM Flags
N/A	Subread Name
cx	Context Flag
ip	**Inter-Pulse Duration**
pw	**Pulse Width**
np	Number of Passes

Tag	Descriptor
N/A	Base Sequence
N/A	Base Quality
qe	Position End
qs	Position Start
rq	Read Quality
sn	Signal-to-Noise
zm	ZMW Number
RG	Read Group

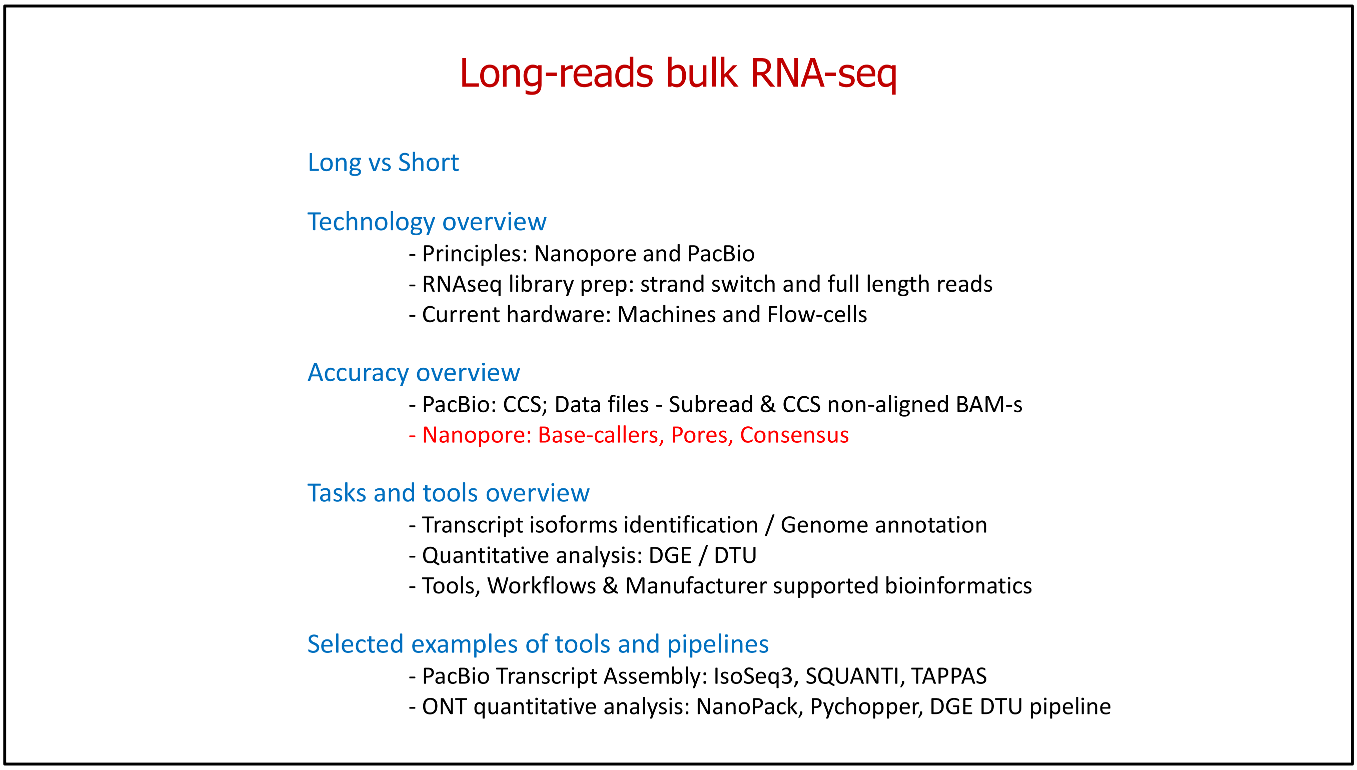
PacBio Webinar: PacBio Data Deep Dive: A Closer Look at HiFi Sequencing, 24 March 2021
https://pacbiofileformats.readthedocs.io/en/12.0/index.html

A short side-tracking about PacBio file formats.

In addition to the base sequence and base quality PacBio needs to record the kinetic information (pulse width, inter-pulse duration) and some other information about their movies. So, they decided not to use FASTQ file format. Instead, they initially decided to use their own version of BAM files.

Also, PacBio uses a special sort of XML file format. You may see more details here: https://pacbiofileformats.readthedocs.io/en/12.0/DataSet.html

In the recent machines, PacBio implements CCS workflow on sequencer, by default outputting only CCS reads, and discarding the sub-reads information to reduce the data size.



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

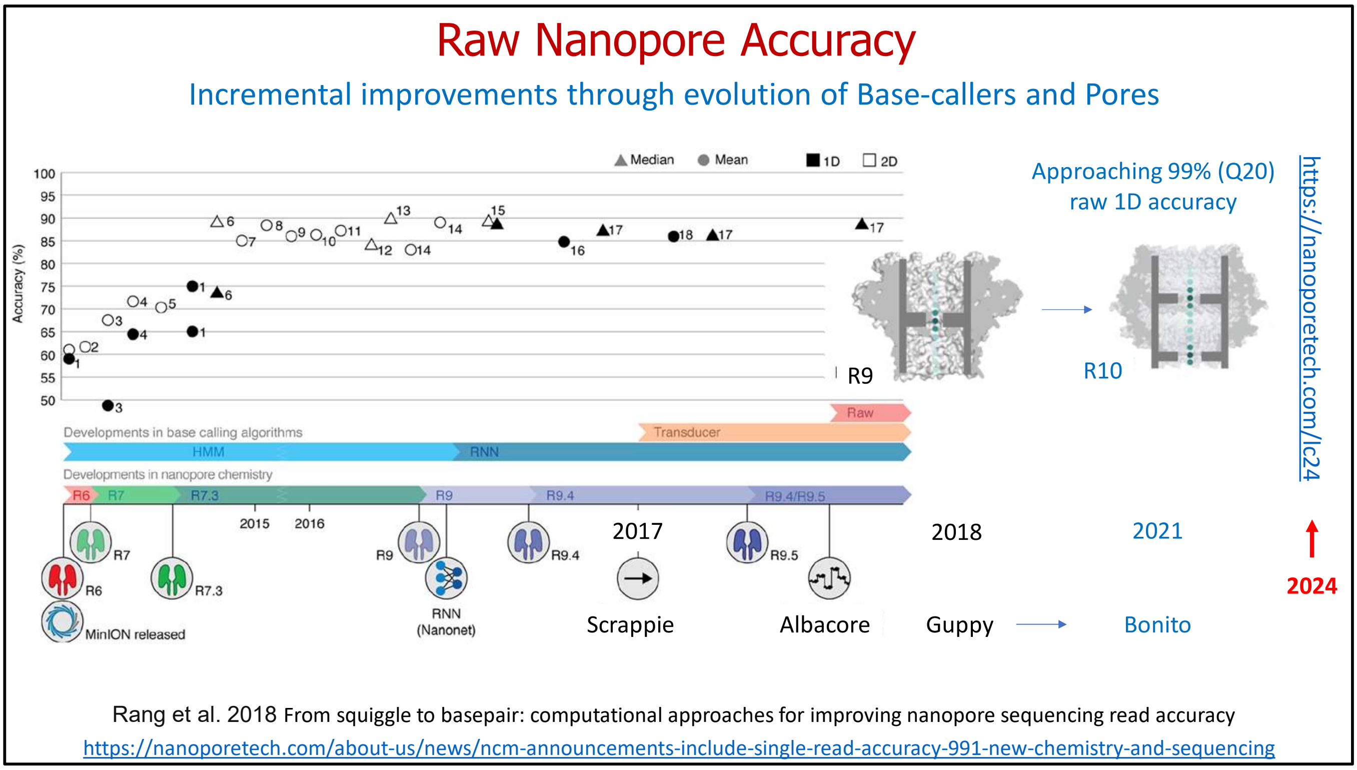
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Coming back to the accuracy of long reads sequencing.



Raw Nanopore Accuracy

Incremental improvements through evolution of Base-callers and Pores

Approaching 99% (Q20) raw 1D accuracy

R9

R10

2017

2018

2021

2024

Scrappie

Albacore

Guppy → Bonito

https://nanoporetech.com/lc24

Rang et al. 2018 From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy

https://nanoporetech.com/about-us/news/ncm-announcements-include-single-read-accuracy-991-new-chemistry-and-sequencing

Once upon a time (about ten years ago) the accuracy of Nanopore reads was awful: 50% errors.

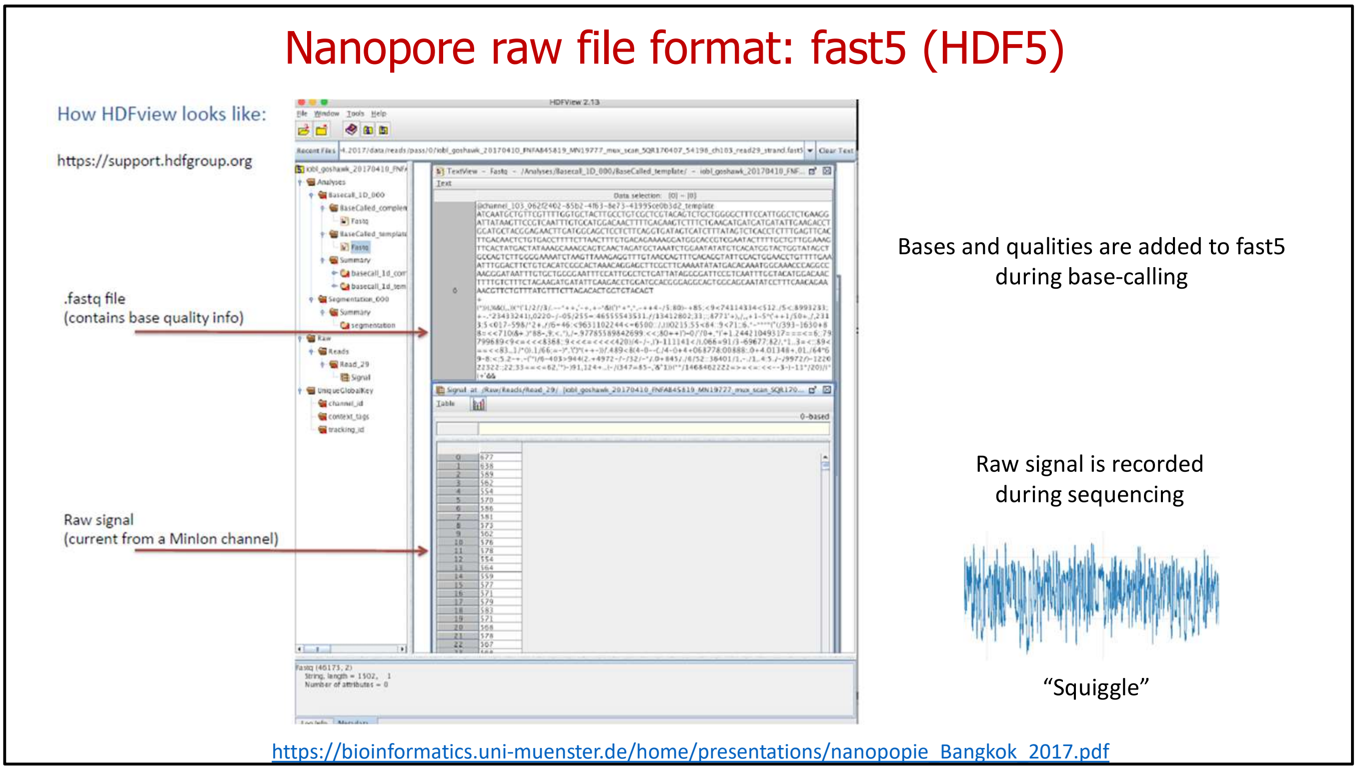
However, it was dramatically improved since that by developing the new pores, and most importantly, by the new algorithms for calling bases from squiggles (it’s a term to describe the ionic current fluctuations recorded by the pore).

The breakthrough in the basecalling algorithms development happened when Nanopore decided to use machine-learning for this.

Practically, this means that the training sets and even some hyperparameters (such as depth of the network etc) may vary even between different models of the same basecaller.   For instance, model trained on human DNA may not be perfect for bacteria or plants …

Of course, the best available models were trained on the human data, which is handy for our Cancer research context ☺

A couple of years ago, Nanopore announced that the latest at the time version of basecaller **Bonito** with data from **R10 pores** approached ***99% raw accuracy***, which is a remarkable progress comparing with initial 50% just 10 year ago.



Nanopore raw file format: fast5 (HDF5)

How HDFview looks like:

https://support.hdfgroup.org

.fastq file
(contains base quality info)

Raw signal
(current from a MinIon channel)

```
4.2017/data/reads/pass/0/iobl_goshawk_20170410_FNFA845819_MN19777_mux_scan_SQR170407_54198_ch103_read29_strand.fast5
```

```
Data selection: [0] ~ [0]
0  @channel_103_062f2402-85b2-4f63-8e73-41995ce0b3d2_template
   ATCAATGCTGTTCGTTTTGGTGCTACTTGCCTGTCGCTCGTACAGTCTGCTGGGGCTTTCCATTGGCTCTGAAGG
   ATTATAAGTTCCGTCAATTTGTGCATGGACAACTTTTGACAAGTCTTTCTGAACATGATCATGATATTGAAGACCT
   GCATGCTACGGAGAACTTGATGGCAGCTCCTCTTCAGGTGATAGTCATCTTTATAGTCTCACCTCTTTGAGTTCAC
   TTCACAACTCTGTGACCTTTTCTTAACTTTGTGACAGAAAAGCATGGCACCGTCGAATACTTTTGCTGTTGGAAAC
   TTCACTATGACTATAAAGCAAAGCAGTCAACTAGATGCTAAATCTGGAATATATGTCACATGGTACTGGTATAGCT
   GCCAGTCTTGGGGAAAATCTAAGTTAAAGAGGTTTGTAACCACTTTCACAGGTATTCCACTGGAACCTGTTTTGAA
   ATTTGGACTTCTGTCACATCGGCACTAAACAGGAGCTTCGCTTCAAAATATATGACACAAATGGCAAACCCAGGCC
   AACGGATAATTTGTGCTGGGGAATTTCCATTGGCTCTGATTATAGGCGATTCCCTCAATTTGCTACATGGACAAC
   TTTTGTCTTTCTAGAAGATGATATTGAAGACCTGGATGCACGGGAGGCACTGGCAGCAATATCCTTTCAACAGAA
   AACGTTCTGTTTATGTTTCTTAGACACTCGTCTACACT
   +
   (")9),%&(),,)((")(1/2//3/.--"++,'-+,+-"&(()"+",",-++4-/5;80)-+85;<9<74114334<512./5<;8993231;
   +-."23433241),0220-/-05/255=46555543531.//13412802;33;;;4771'+),/,,+1-5"(++1/50+,/,231
   3;5<017-598/"2+,//6=46:<9631102244<=6500::/,))0215;55<84::9<71::6,"-""""("(/393-1630+8
   8=<<710(&+,)"88-,9;<,"),/-.97785589842699.<<;80++)')-0/'/0+,"/'+1.24421049317===<=6;;79
   799689<9<==<<<8368;;9<<<<=<<<<420)/4-/-,))-111141</),066=91/3-69677;82/,"1..3=<;;89<
   ==<<83..1/"0).1/66;=-)",'))"(++-))/.489<8(4-0--(./4-0+4+068778.00888::0+4.01348+,01,/64"6
   9-8:<;5.2-+,-("")/6-403>944(2.+4972-/-/32/-"/.0+845/./6/52::36401/1,-./1..4:5./-/9972/)-1220
   22322:;22:33+==<=62,")-)91,124+..(-/(347=85-,'&"1)(""/1468462222=>=<=:<<--3-)-11"/20)/(
   (+'&&
```

0	677
1	638
2	589
3	562
4	554
5	570
6	586
7	581
8	573
9	562
10	576
11	578
12	554
13	564
14	559
15	577
16	571
17	579
18	583
19	571
20	568
21	578
22	507

```
Fastq (46173, 2)
  String, length = 1502,  1
  Number of attributes = 0
```

Bases and qualities are added to fast5 during base-calling

Raw signal is recorded during sequencing

"Squiggle"

https://bioinformatics.uni-muenster.de/home/presentations/nanopopie_Bangkok_2017.pdf

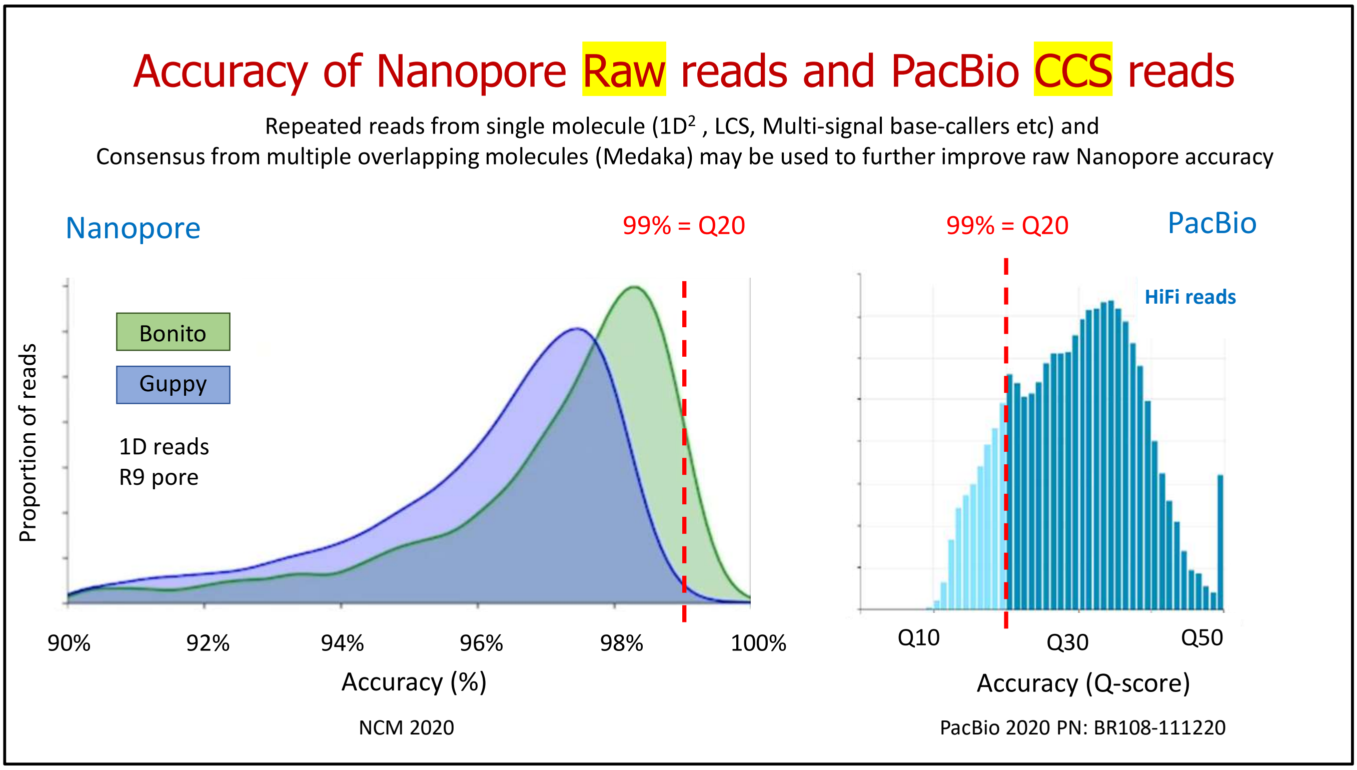
Again: a side-tracking about file formats.

To record raw data ("squiggle") along with the bases, their qualities and some other data, Nanopore initially used a sort of XML data format, called fast5.

There are viewers for such files. This slide illustrates the structure of the file and some data contained in different tags.

However, the fast5 files could be really big (many hundreds of gigabytes), and processing fast5 files may be slow. A newer version of the Nanopore data files is called POD5.

You may find more information about the Nanopore data formats here: https://community.nanoporetech.com/technical_documents/data-analysis/v/datd_5000_v1_revr_22aug2016/file-formats



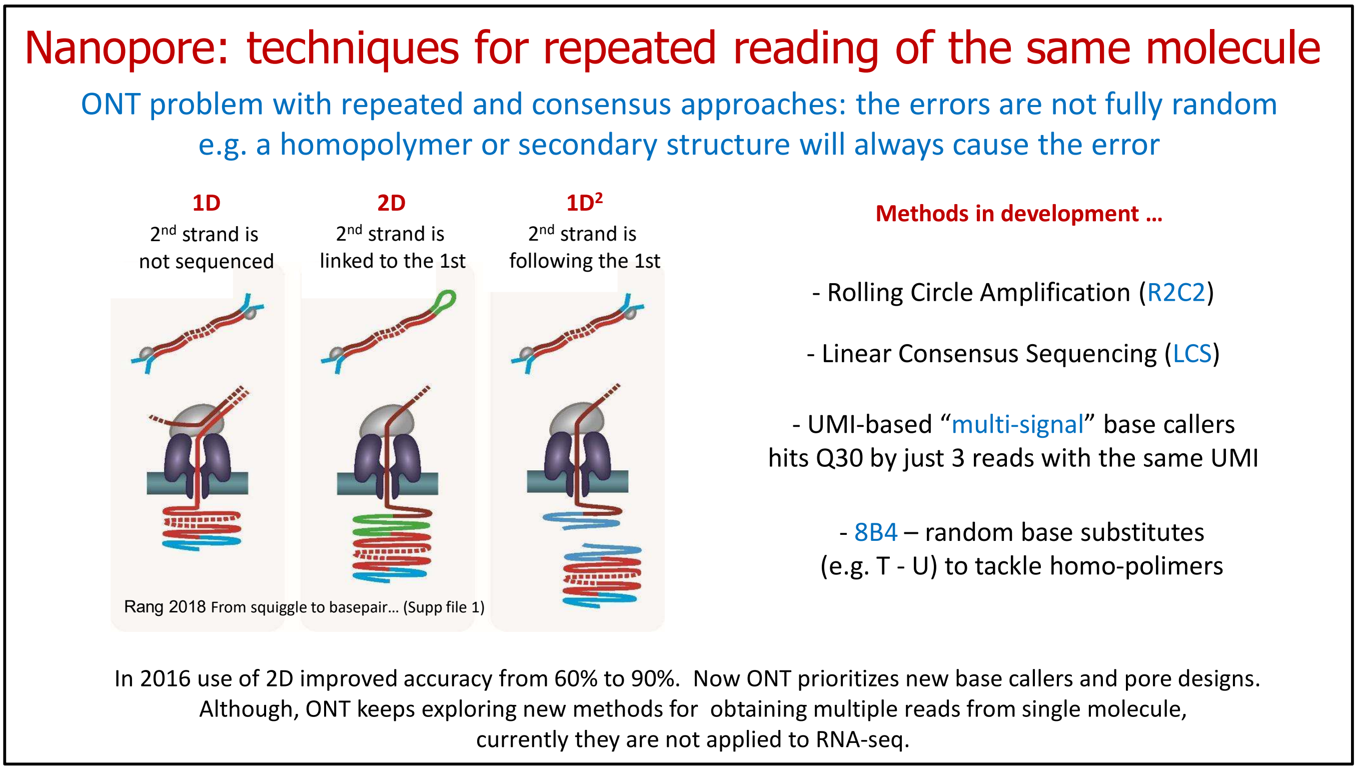
Accuracy of Nanopore Raw reads and PacBio CCS reads

Repeated reads from single molecule (1D^2 , LCS, Multi-signal base-callers etc) and
Consensus from multiple overlapping molecules (Medaka) may be used to further improve raw Nanopore accuracy

NCM 2020

PacBio 2020 PN: BR108-111220

However, 99% accuracy is still just Q20. So, the accuracy of ***raw*** ONT data is still much lower than ***HiFi*** PabBio.



Nanopore: techniques for repeated reading of the same molecule

ONT problem with repeated and consensus approaches: the errors are not fully random
e.g. a homopolymer or secondary structure will always cause the error

1D
2nd strand is not sequenced

2D
2nd strand is linked to the 1st

$1D^2$
2nd strand is following the 1st

Rang 2018 From squiggle to basepair… (Supp file 1)

Methods in development …

- Rolling Circle Amplification (R2C2)

- Linear Consensus Sequencing (LCS)

- UMI-based “multi-signal” base callers
hits Q30 by just 3 reads with the same UMI

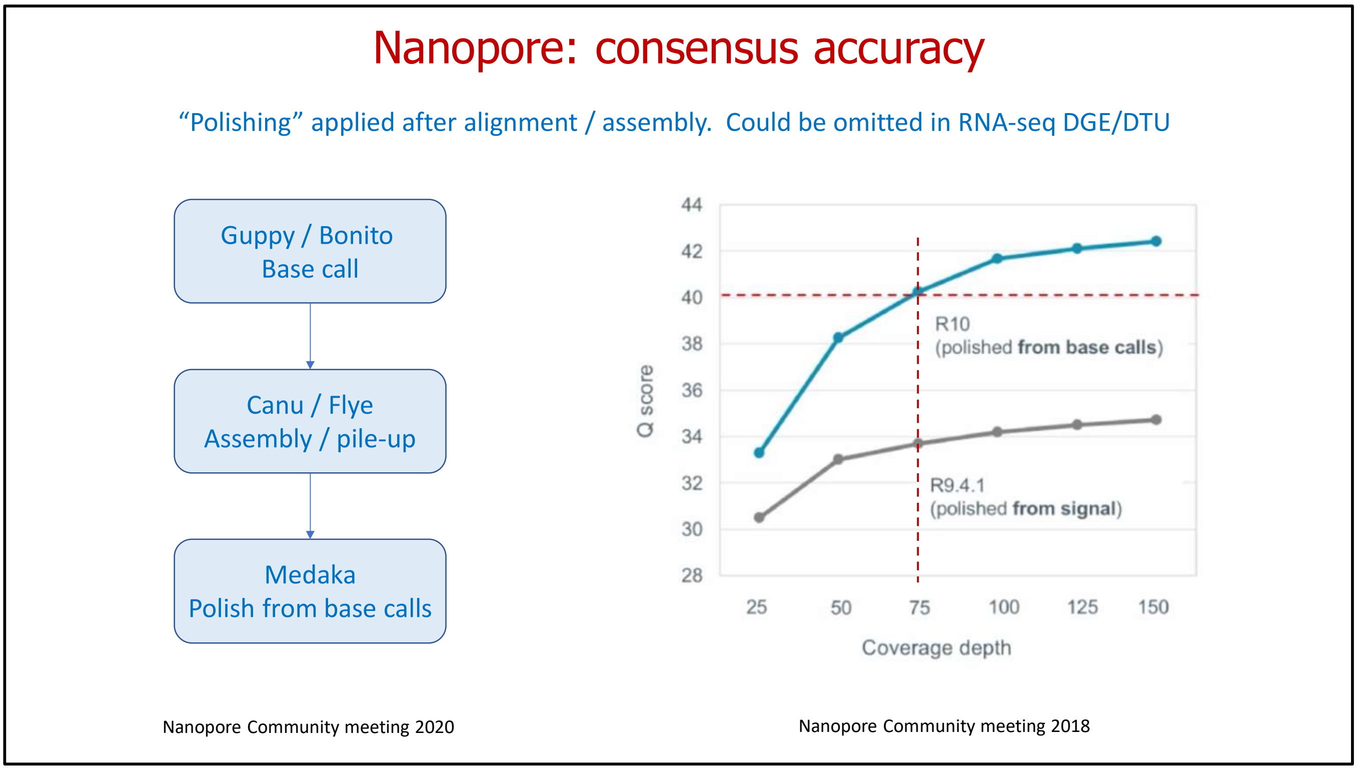
- 8B4 – random base substitutes
(e.g. T - U) to tackle homo-polimers

In 2016 use of 2D improved accuracy from 60% to 90%. Now ONT prioritizes new base callers and pore designs.
Although, ONT keeps exploring new methods for obtaining multiple reads from single molecule,
currently they are not applied to RNA-seq.

Nanopore also explored with the multiple reads from single molecule, but this direction looks abandoned at the moment.

- First, it red the same molecule twice by adding a hairpin (so called 2D libraries);
- Then it tried to do the same without adding a hairpin (1D2): because the 2nd strand is still in vicinity of the pore, in ~75% cases it may follow the first strand even w/o hairpin.

These methods gave a modest improvement, but nowhere near the CCS reads in Pacbio.
There are some other experimental approached in development.



Nanopore: consensus accuracy

"Polishing" applied after alignment / assembly. Could be omitted in RNA-seq DGE/DTU

Nanopore Community meeting 2020

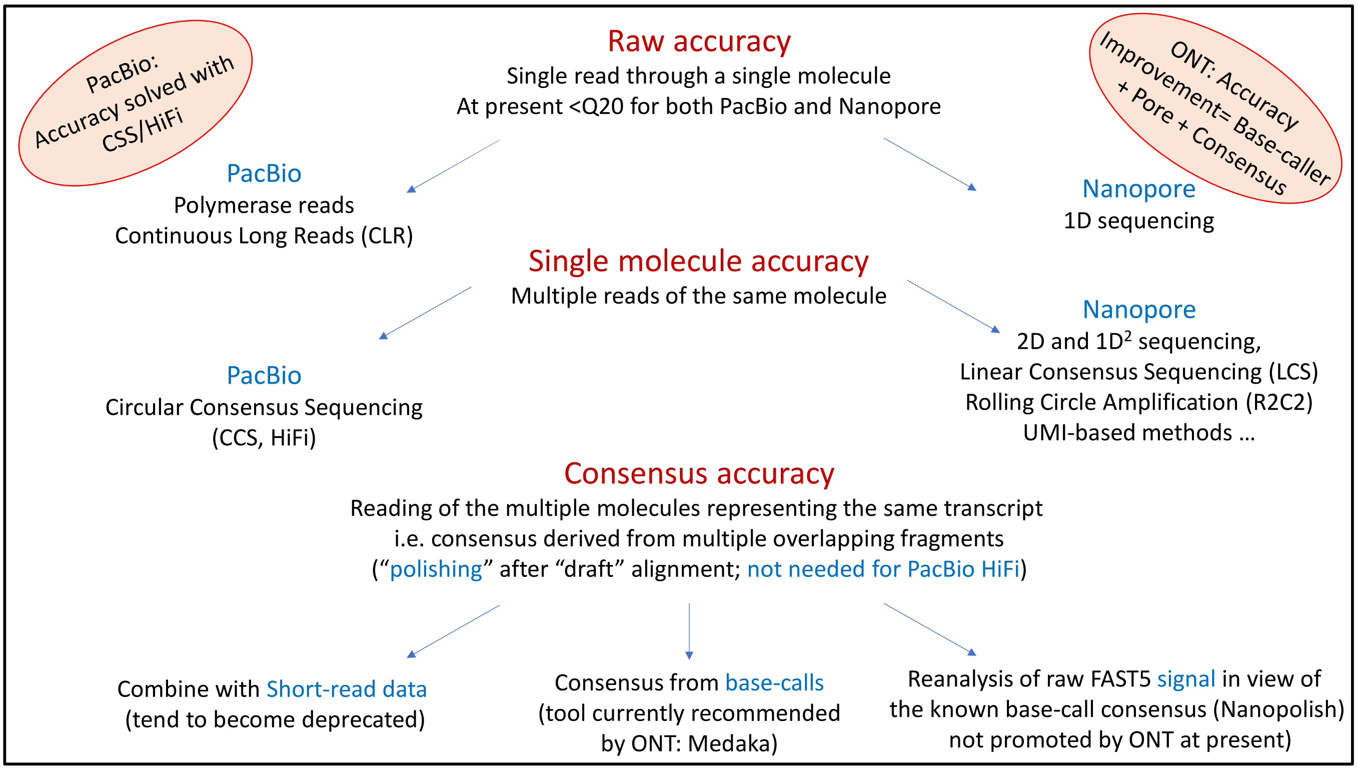
Nanopore Community meeting 2018

The issue complicating Nanopore basecalling is that the errors are not random. If some sequence produces a squiggle that is hard to decipher at the first pass, it is also hard at the second pass. Most errors are in homopolymers (repeats of the same nucleotide), which is not the common pattern in RNA (except for the poly-A tail).

Nanopore works on it, both on technical and marketing sides.

Their current marketing clams are that with sufficient depth of sequencing the Nanopore ***consensus*** accuracy may become the best of the long read sequencing method.

This is yet to happen … However, with sufficient depth (and associated cost) they truly can rich a very good accuracy.



PacBio: Accuracy solved with CSS/HiFi

Raw accuracy

Single read through a single molecule

At present <Q20 for both PacBio and Nanopore

ONT: Accuracy Improvement= Base-caller + Pore + Consensus

PacBio

Polymerase reads

Continuous Long Reads (CLR)

Nanopore

1D sequencing

Single molecule accuracy

Multiple reads of the same molecule

PacBio

Circular Consensus Sequencing (CCS, HiFi)

Nanopore

2D and 1D^2 sequencing,

Linear Consensus Sequencing (LCS)

Rolling Circle Amplification (R2C2)

UMI-based methods …

Consensus accuracy

Reading of the multiple molecules representing the same transcript

i.e. consensus derived from multiple overlapping fragments

(“polishing” after “draft” alignment; not needed for PacBio HiFi)

Combine with Short-read data (tend to become deprecated)

Consensus from base-calls (tool currently recommended by ONT: Medaka)

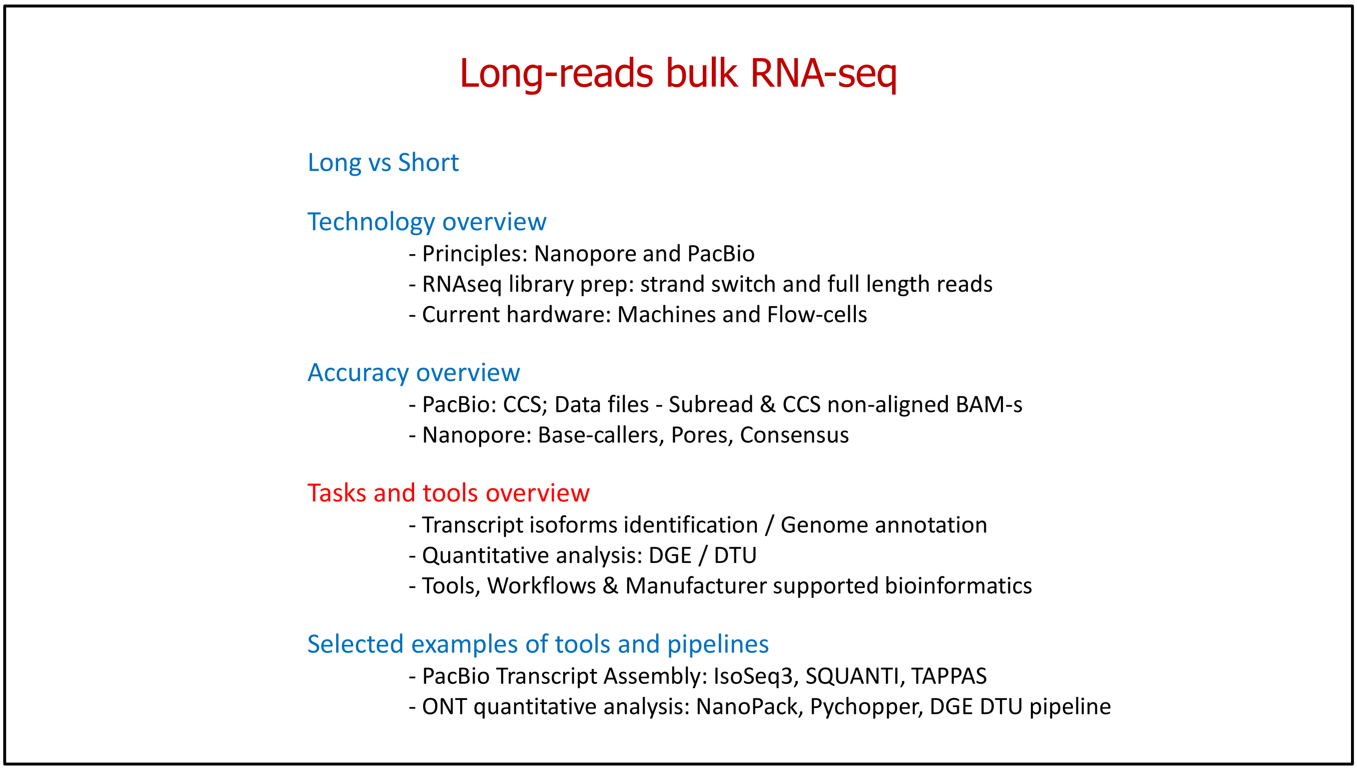
Reanalysis of raw FAST5 signal in view of the known base-call consensus (Nanopolish) not promoted by ONT at present)

To summarize, The ***raw*** reads accuracy is low for both PacBio and Nanopore.

PacBio solved it by Circular Consensus Sequencing (***CCS***), easily reaching Q40 and above for reads of 15kb.

Nanopore achieved quality Q15-20 by new ***basecallers and pores design***. This accuracy is enough for most RNAseq applications.

Because Nanopore is still cheaper, which allows to get higher depth for the same price, at the moment it is considered that Nanopore could be better for transcripts quantification and differential expression studies, while PacBio could be better for accurate transcript isoforms discovery (Pardo-Palacios et al 2023, https://www.biorxiv.org/content/10.1101/2023.07.25.550582v1)



Long-reads bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

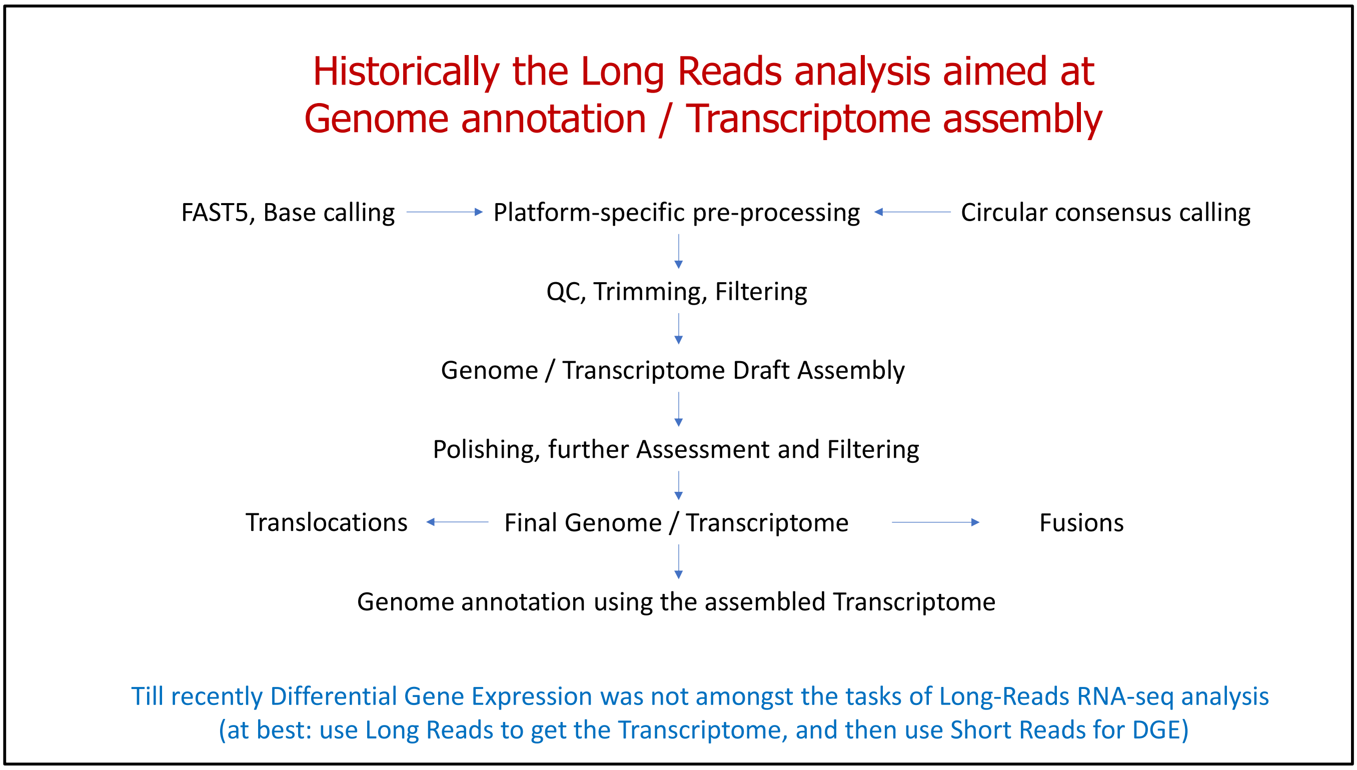
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

This brings us to the next point: the tasks currently performed by long-reads RNAseq



Historically the Long Reads analysis aimed at Genome annotation / Transcriptome assembly

FAST5, Base calling
Platform-specific pre-processing
Circular consensus calling
QC, Trimming, Filtering
Genome / Transcriptome Draft Assembly
Polishing, further Assessment and Filtering
Translocations
Final Genome / Transcriptome
Fusions
Genome annotation using the assembled Transcriptome

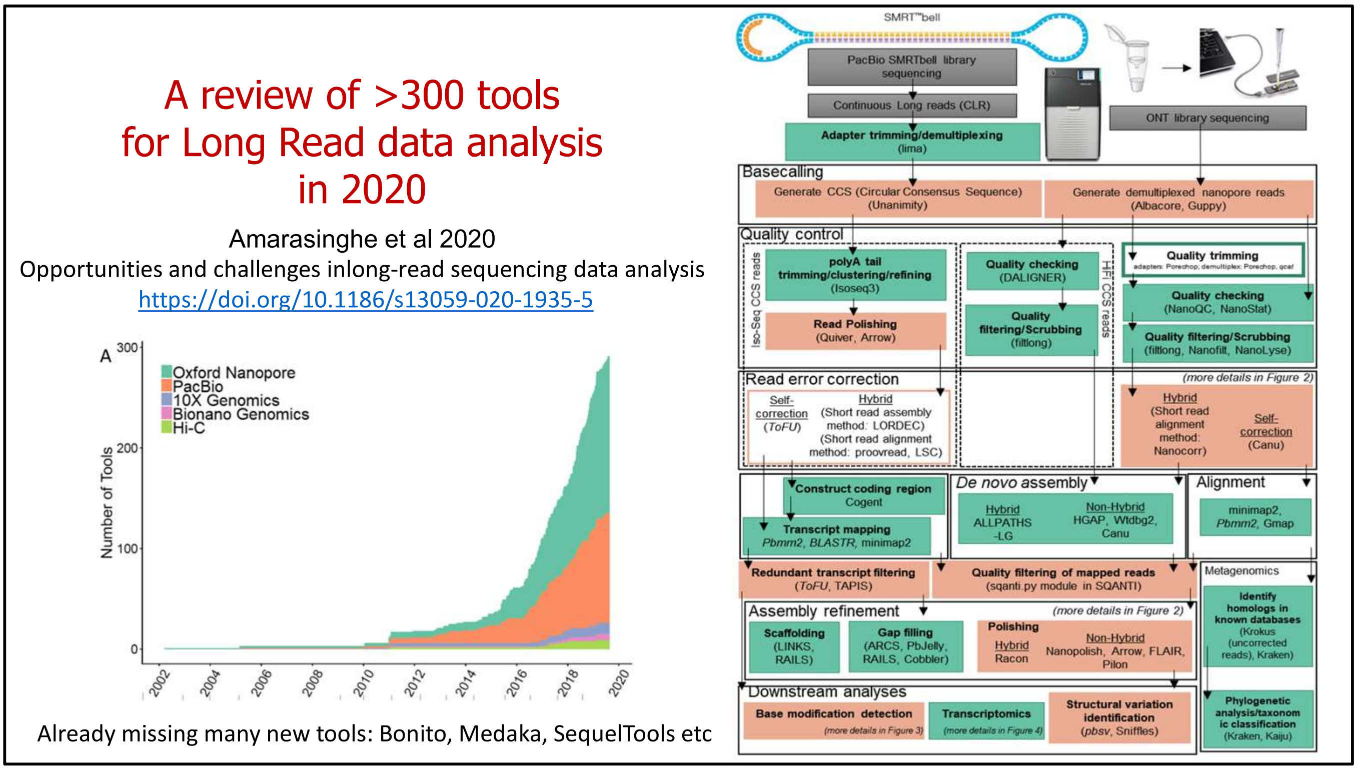
Till recently Differential Gene Expression was not amongst the tasks of Long-Reads RNA-seq analysis (at best: use Long Reads to get the Transcriptome, and then use Short Reads for DGE)

Till very recently the long reads RNAseq was too expensive to get sufficient depth for reliable transcripts ***quantification***.
So, most of the tools, pipelines and papers published about long reads sequencing were focused on transcripts ***identification***.
Identification of transcripts could be used for

- Genome annotation
- Transcriptome assembly (in absence of a reference genome) or
- Fusion detection

Such tasks as "Genome annotation" or "Transcriptome assembly" in absence of a reference genome may sound alien for human cancer researcher because human cancer research enjoys the best annotated reference genome and transcriptome available (i.e. human genome and transcriptome :)

However, after adoption of Nanopore Prometheon (and maybe PacBio Revio+Kinnex in the near future) the long-reads technologies are producing sufficient and affordable depth of sequencing for the quantitative analysis too.



A review of >300 tools for Long Read data analysis in 2020

Amarasinghe et al 2020

Opportunities and challenges inlong-read sequencing data analysis

https://doi.org/10.1186/s13059-020-1935-5

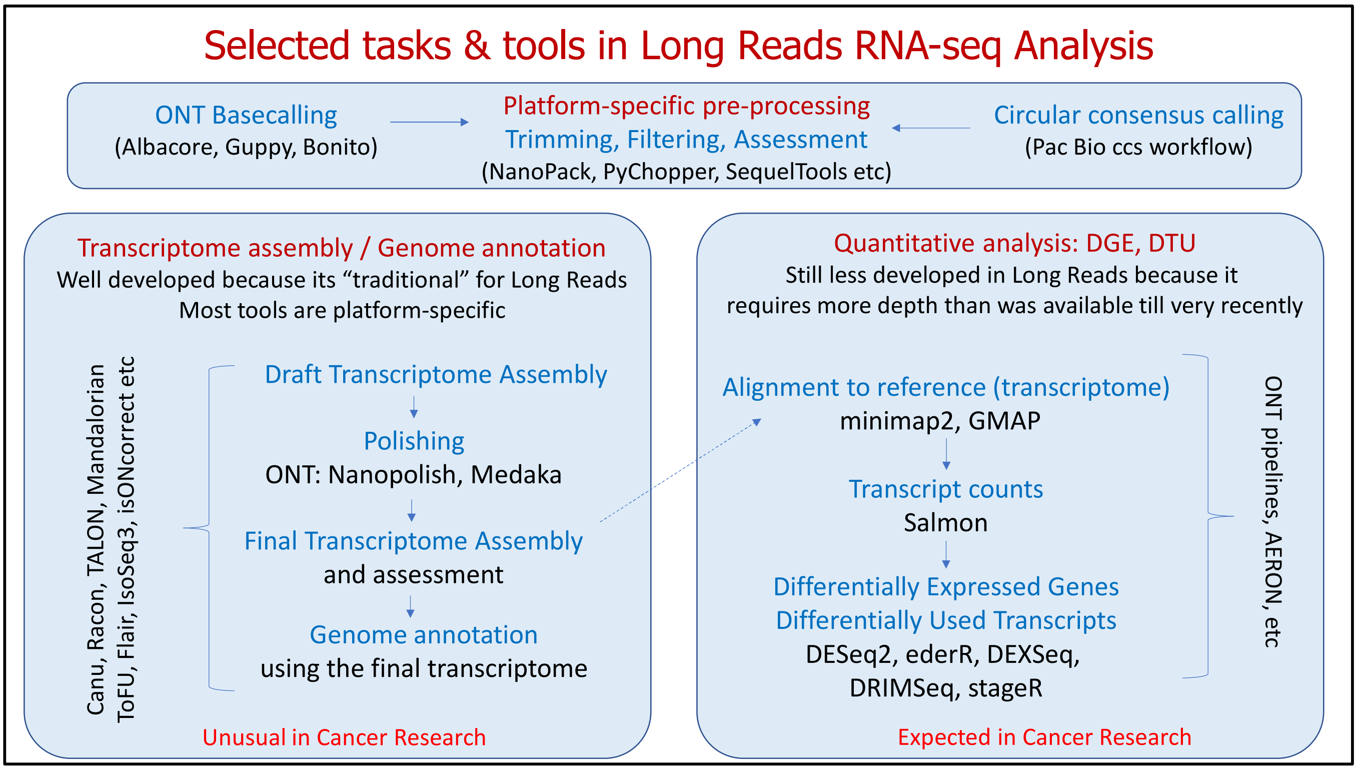
Already missing many new tools: Bonito, Medaka, SequelTools etc

A
300
200
100
0
Number of Tools
🟩 Oxford Nanopore
🟧 PacBio
🟦 10X Genomics
🟪 Bionano Genomics
🟩 Hi-C
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
SMRT™bell
PacBio SMRTbell library sequencing
Continuous Long reads (CLR)
Adapter trimming/demultiplexing (lima)
ONT library sequencing
Basecalling
Generate CCS (Circular Consensus Sequence) (Unanimity)
Generate demultiplexed nanopore reads (Albacore, Guppy)
Quality control
Iso-Seq CCS reads
polyA tail trimming/clustering/refining (Isoseq3)
Read Polishing (Quiver, Arrow)
Quality checking (DALIGNER)
Quality filtering/Scrubbing (filtlong)
HiFi CCS reads
Quality trimming adapters: Porechop; demultiplex: Porechop, qcat
Quality checking (NanoQC, NanoStat)
Quality filtering/Scrubbing (filtlong, Nanofilt, NanoLyse)
Read error correction
Self-correction (ToFU)
Hybrid (Short read assembly method: LORDEC) (Short read alignment method: proovread, LSC)
(more details in Figure 2)
Hybrid (Short read alignment method: Nanocorr)
Self-correction (Canu)
Construct coding region Cogent
Transcript mapping Pbmm2, BLASTR, minimap2
De novo assembly
Hybrid ALLPATHS -LG
Non-Hybrid HGAP, Wtdbg2, Canu
Alignment
minimap2, Pbmm2, Gmap
Redundant transcript filtering (ToFU, TAPIS)
Quality filtering of mapped reads (sqanti.py module in SQANTI)
Metagenomics
Identify homologs in known databases (Krokus (uncorrected reads), Kraken)
Assembly refinement
(more details in Figure 2)
Scaffolding (LINKS, RAILS)
Gap filling (ARCS, PbJelly, RAILS, Cobbler)
Polishing
Hybrid Racon
Non-Hybrid Nanopolish, Arrow, FLAIR, Pilon
Downstream analyses
Base modification detection (more details in Figure 3)
Transcriptomics (more details in Figure 4)
Structural variation identification (pbsv, Sniffles)
Phylogenetic analysis/taxonomic classification (Kraken, Kaiju)

It looks like the tools for the long-reads data analysis are proliferating even faster than the tools for short-reads analysis :)

Only this review mentions hundreds of tools …

It's now a separate branch of bioinformatics – not the tools development but the tools comparison :))



Selected tasks & tools in Long Reads RNA-seq Analysis

ONT Basecalling
(Albacore, Guppy, Bonito)

Platform-specific pre-processing
Trimming, Filtering, Assessment
(NanoPack, PyChopper, SequelTools etc)

Circular consensus calling
(Pac Bio ccs workflow)

Transcriptome assembly / Genome annotation
Well developed because its “traditional” for Long Reads
Most tools are platform-specific

Canu, Racon, TALON, Mandalorian
ToFU, Flair, IsoSeq3, isONcorrect etc

Draft Transcriptome Assembly

Polishing
ONT: Nanopolish, Medaka

Final Transcriptome Assembly
and assessment

Genome annotation
using the final transcriptome

Unusual in Cancer Research

Quantitative analysis: DGE, DTU
Still less developed in Long Reads because it
requires more depth than was available till very recently

Alignment to reference (transcriptome)
minimap2, GMAP

Transcript counts
Salmon

Differentially Expressed Genes
Differentially Used Transcripts
DESeq2, ederR, DEXSeq,
DRIMSeq, stageR

ONT pipelines, AERON, etc

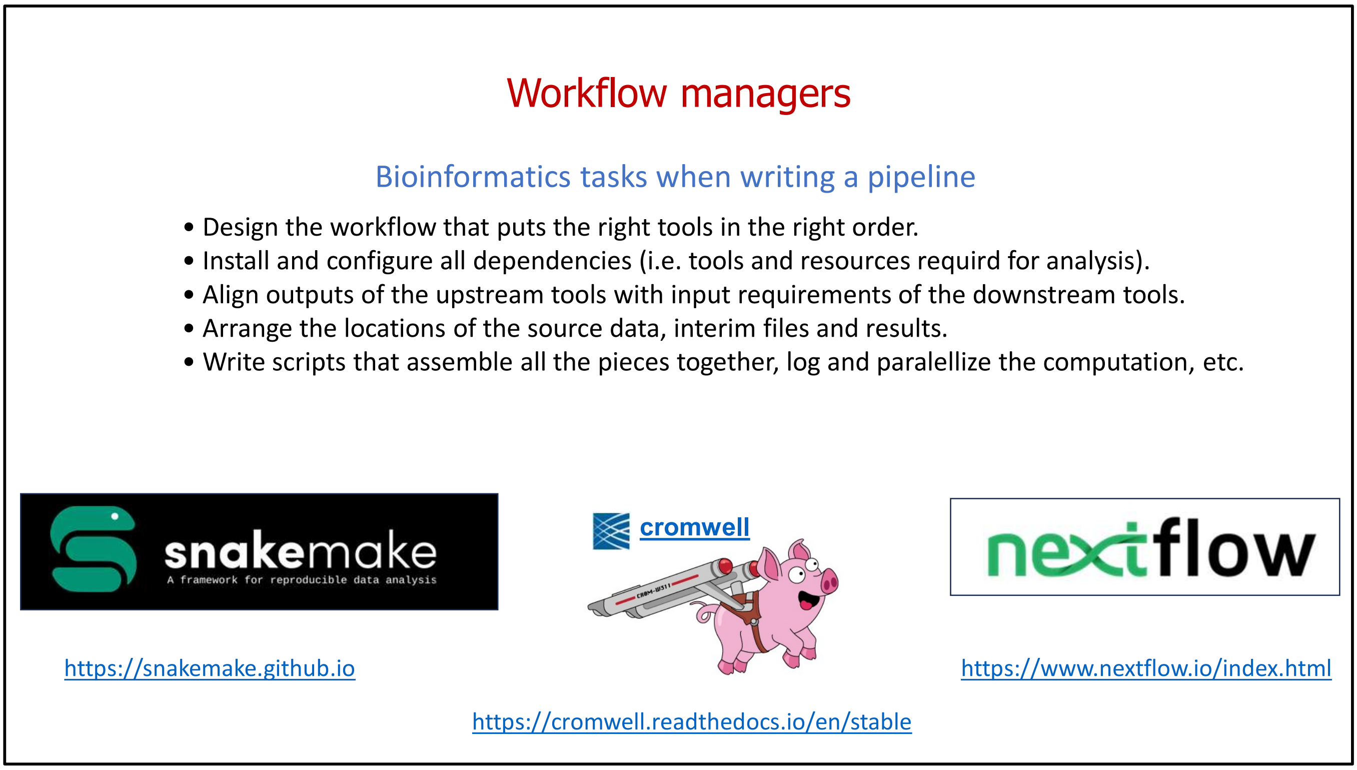
Expected in Cancer Research

To simplify the chart shown in the previous slide, for the purpose of this short introductory lecture,
here I split long reads RNA-seq tools to 3 large categories:

- Pre-processing
- Transcriptome assembly/Genome annotation
- Quantitative analysis

I do not claim that mentioned tools are better than many others …
The specific tools mentioned here are just to illustrate what was used in the literature, some of them are already deprecated.

And, of course, there are many RNAseq tasks and tools that are not even mentioned here, such as Nanopore Direct RNA-seq, RNA-modifications, poly-A tail length etc



Workflow managers

Bioinformatics tasks when writing a pipeline

- Design the workflow that puts the right tools in the right order.
- Install and configure all dependencies (i.e. tools and resources requird for analysis).
- Align outputs of the upstream tools with input requirements of the downstream tools.
- Arrange the locations of the source data, interim files and results.
- Write scripts that assemble all the pieces together, log and paralellize the computation, etc.

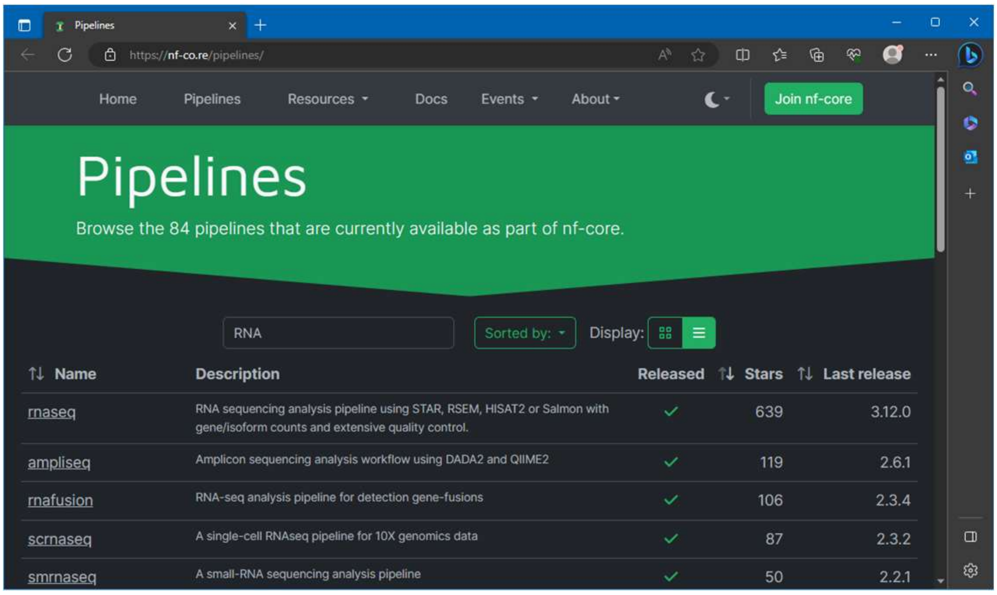
https://snakemake.github.io

https://www.nextflow.io/index.html

https://cromwell.readthedocs.io/en/stable

Even a simple bioinformatics task often requires multiple steps and tools.
It takes lots of effort and expertise to write a pipeline connecting different tools.
Workflow managers help to connect multiple tools into pipelines and facilitate reproducible research.

NF-core: Nextflow pipelines repository



Pipelines

Browse the 84 pipelines that are currently available as part of nf-core.

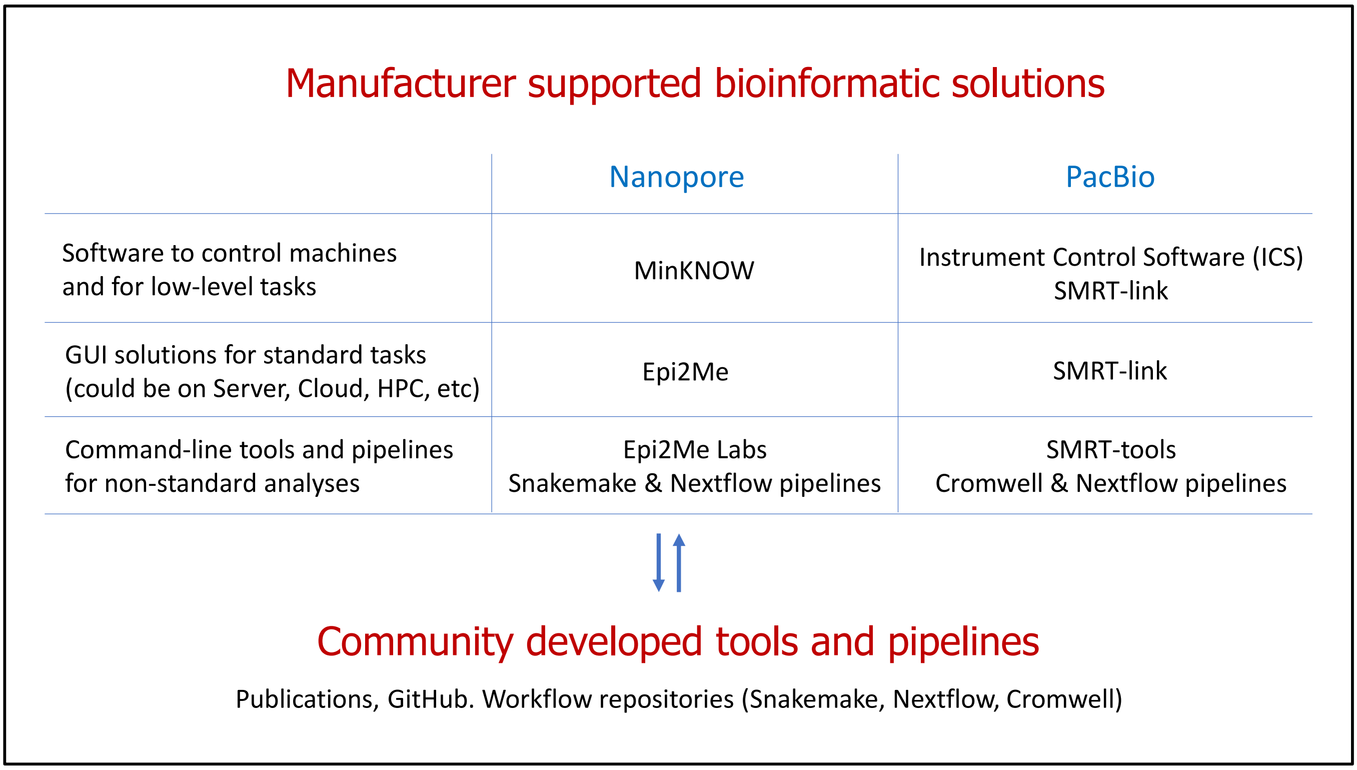
RNA

Name	Description	Released	Stars	Last release
rnaseq	RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.	✓	639	3.12.0
ampliseq	Amplicon sequencing analysis workflow using DADA2 and QIIME2	✓	119	2.6.1
rnafusion	RNA-seq analysis pipeline for detection gene-fusions	✓	106	2.3.4
scrnaseq	A single-cell RNAseq pipeline for 10X genomics data	✓	87	2.3.2
smrnaseq	A small-RNA sequencing analysis pipeline	✓	50	2.2.1

Snakemake workflow catalogue: many hundreds of workflows

https://snakemake.github.io/snakemake-workflow-catalog

Writing and publishing pipelines is currently a strong trend in bioinformatics community.



Manufacturer supported bioinformatic solutions

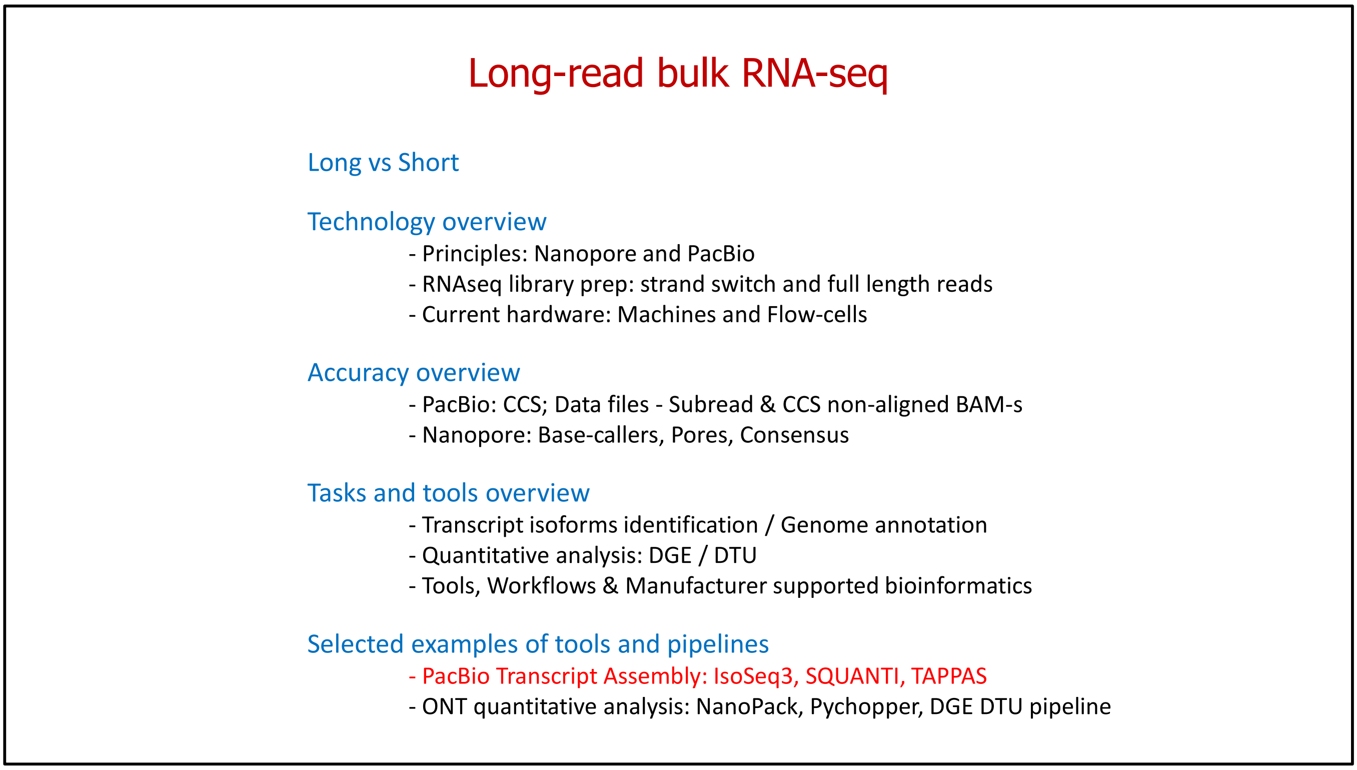
	Nanopore	PacBio
Software to control machines and for low-level tasks	MinKNOW	Instrument Control Software (ICS) SMRT-link
GUI solutions for standard tasks (could be on Server, Cloud, HPC, etc)	Epi2Me	SMRT-link
Command-line tools and pipelines for non-standard analyses	Epi2Me Labs Snakemake & Nextflow pipelines	SMRT-tools Cromwell & Nextflow pipelines

Community developed tools and pipelines

Publications, GitHub. Workflow repositories (Snakemake, Nextflow, Cromwell)

Both Nanopore and PacBio provide extensive Bioinformatics support.
With some pinch of salt, their manufacturer supplied bioinformatics is summarized in such table.

Along with the separate tools, both Nanopore and PacBio develop and publish the entire pipelines, using different workflow managers.
We will try one in our practical session …



Long-read bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

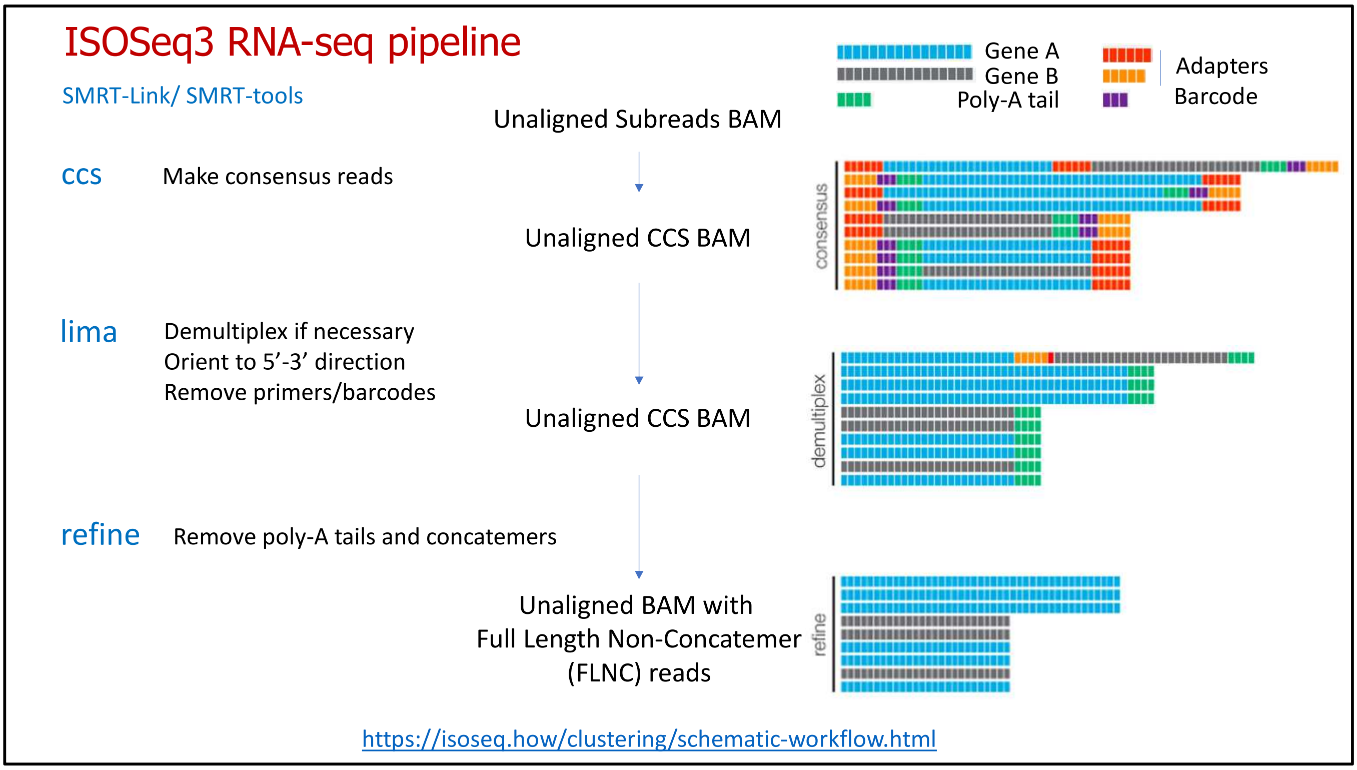
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Finally, I will briefly discuss some examples of the long-reads RNA-seq tools and pipelines.



ISOSeq3 RNA-seq pipeline

SMRT-Link/ SMRT-tools

Unaligned Subreads BAM

ccs Make consensus reads

Unaligned CCS BAM

lima Demultiplex if necessary
Orient to 5’-3’ direction
Remove primers/barcodes

Unaligned CCS BAM

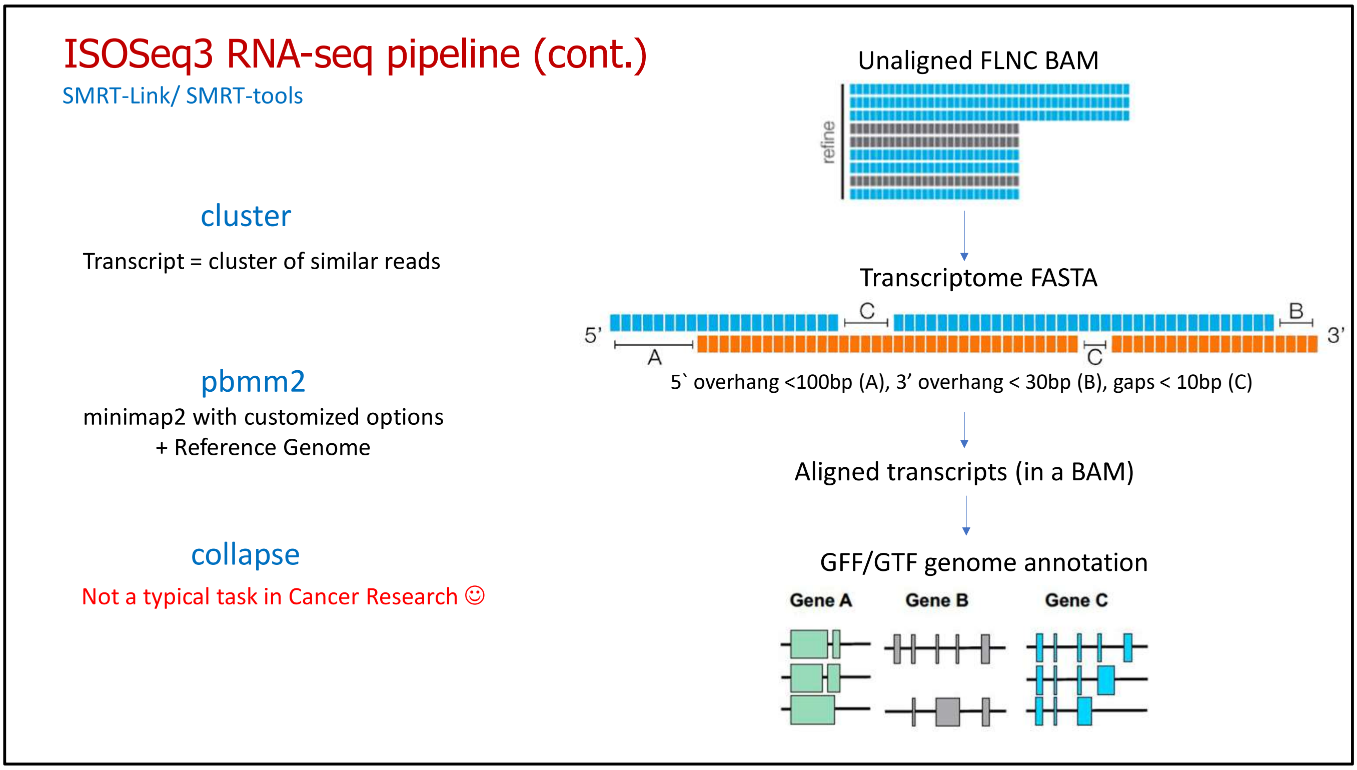
refine Remove poly-A tails and concatemers

Unaligned BAM with
Full Length Non-Concatemer
(FLNC) reads

https://isoseq.how/clustering/schematic-workflow.html

I will start with IsoSeq (=Isoforms Sequencing) pipeline provided by PacBio.
It was designed for transcripts isoforms identification.

It includes preprocessing that is required for virtually any PacBio RNAseq task.
Tools used in this pipeline are available separately from their pages on PacBio’s GitHub, as well they are included into SMRT-Link software or SMRT-tools toolset available from PacBio.



ISOSeq3 RNA-seq pipeline (cont.)

SMRT-Link/ SMRT-tools

cluster

Transcript = cluster of similar reads

pbmm2

minimap2 with customized options
+ Reference Genome

collapse

Not a typical task in Cancer Research ☺

Unaligned FLNC BAM

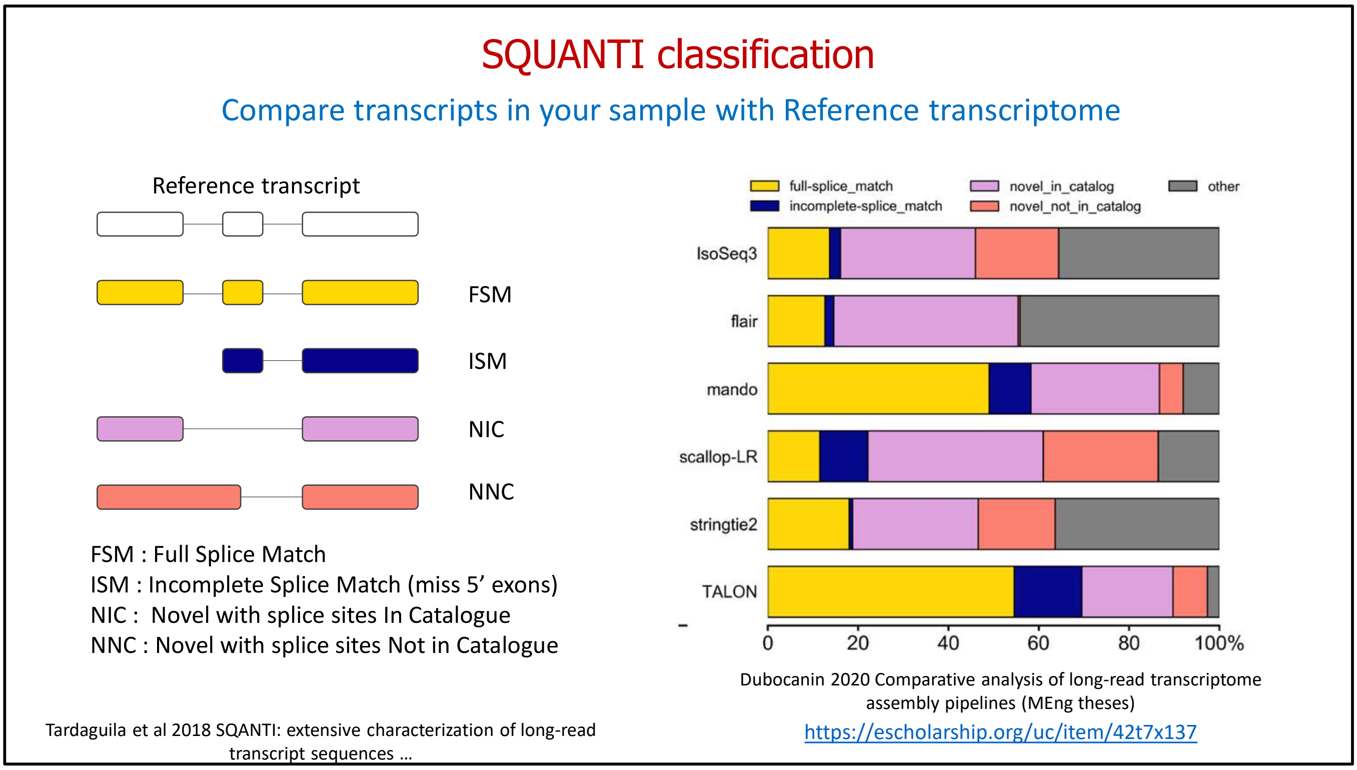
Transcriptome FASTA

5` overhang <100bp (A), 3' overhang < 30bp (B), gaps < 10bp (C)

Aligned transcripts (in a BAM)

GFF/GTF genome annotation

You can see that after alignment IsoSeq pipeline does not count specific transcripts (as we would expect for short-reads analysis).
Instead, it just "collapses" similar transcripts and generates genome annotation.
Not a typical task in Cancer Research.
However, the initial steps of IsoSeq pipeline still would be needed for the Iso-Forms quantification too.



SQUANTI classification

Compare transcripts in your sample with Reference transcriptome

FSM : Full Splice Match
ISM : Incomplete Splice Match (miss 5’ exons)
NIC : Novel with splice sites In Catalogue
NNC : Novel with splice sites Not in Catalogue

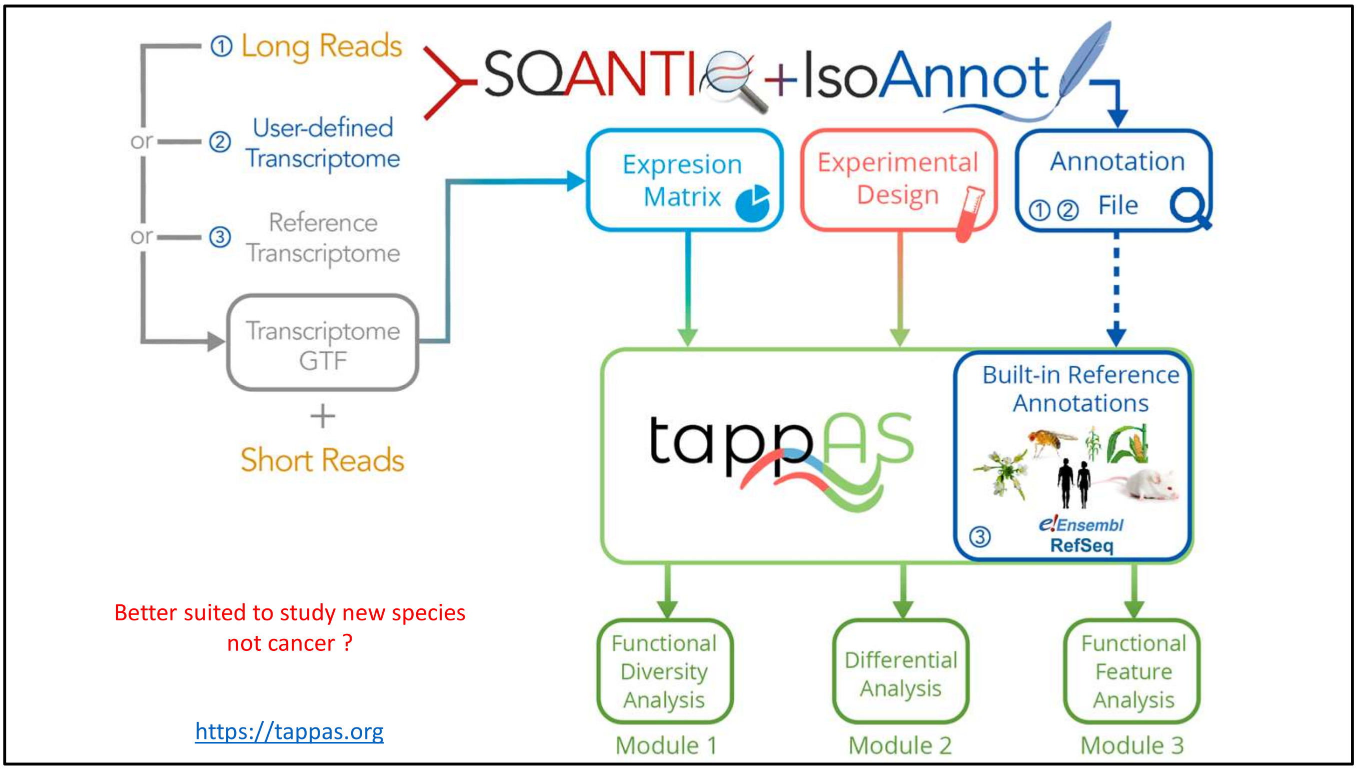
Tardaguila et al 2018 SQANTI: extensive characterization of long-read transcript sequences …

Dubocanin 2020 Comparative analysis of long-read transcriptome assembly pipelines (MEng theses)
https://escholarship.org/uc/item/42t7x137

The down-stream step after the transcript isoforms discovery is to classify these isoforms.

Squanti is a very popular tool downstream of the transcriptome assembly.
It compares the detected transcripts with the previously available reference transcriptome.

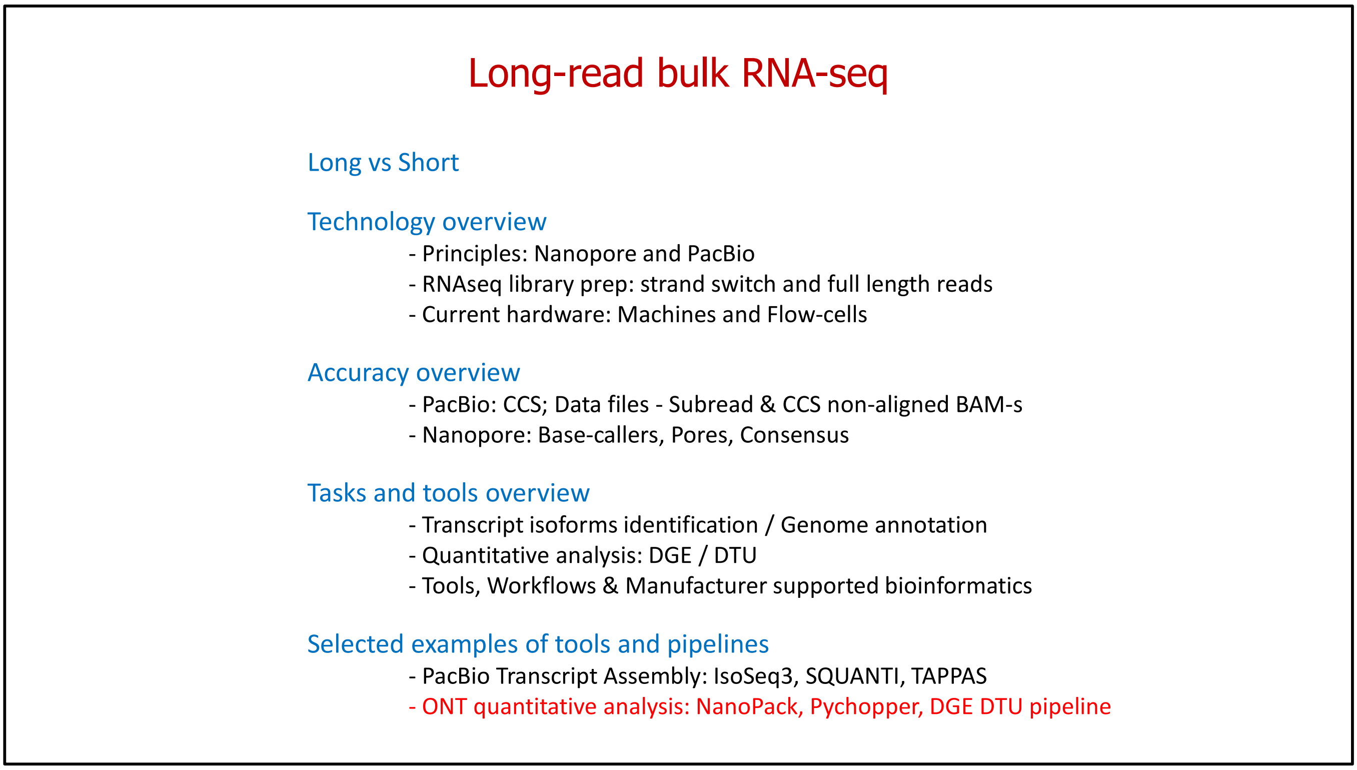
Surprisingly to me, despite the long reads supposedly spanning entire transcripts, there could be quite a disparity between results of different transcriptome assembly pipelines.
(shown on the right)



Better suited to study new species
not cancer ?

https://tappas.org

Squanti is a part of a wider ecosystem of tools for the transcriptome annotation, called ***TAPPAS.***
Apparently, it also includes a differential expression functionality.
However, initially the ***TAPPAS*** quantification module relied on the additional short-reads data.



Long-read bulk RNA-seq

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

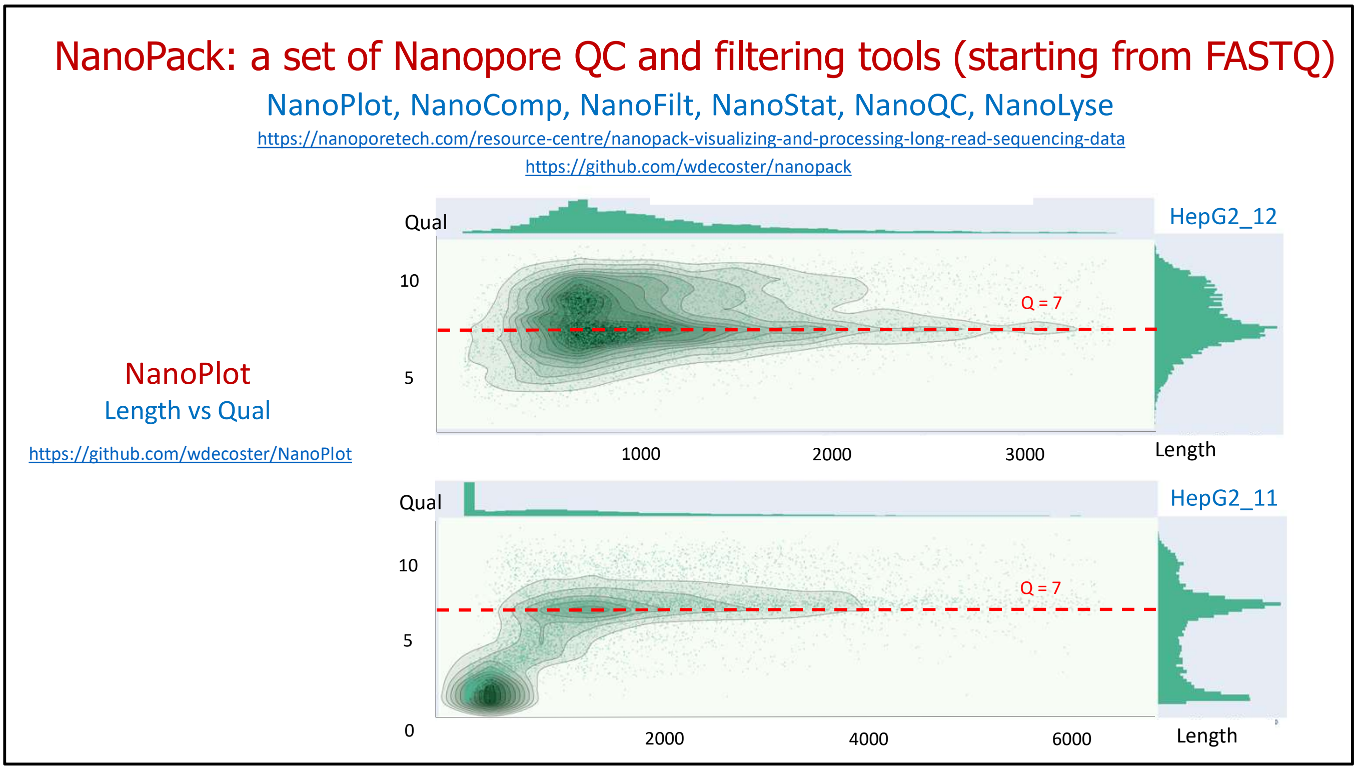
Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

The last part of the lecture will illustrate some Oxford Nanopore tools that we will use during the practical session.



NanoPack: a set of Nanopore QC and filtering tools (starting from FASTQ)

NanoPlot, NanoComp, NanoFilt, NanoStat, NanoQC, NanoLyse

https://nanoporetech.com/resource-centre/nanopack-visualizing-and-processing-long-read-sequencing-data

https://github.com/wdecoster/nanopack

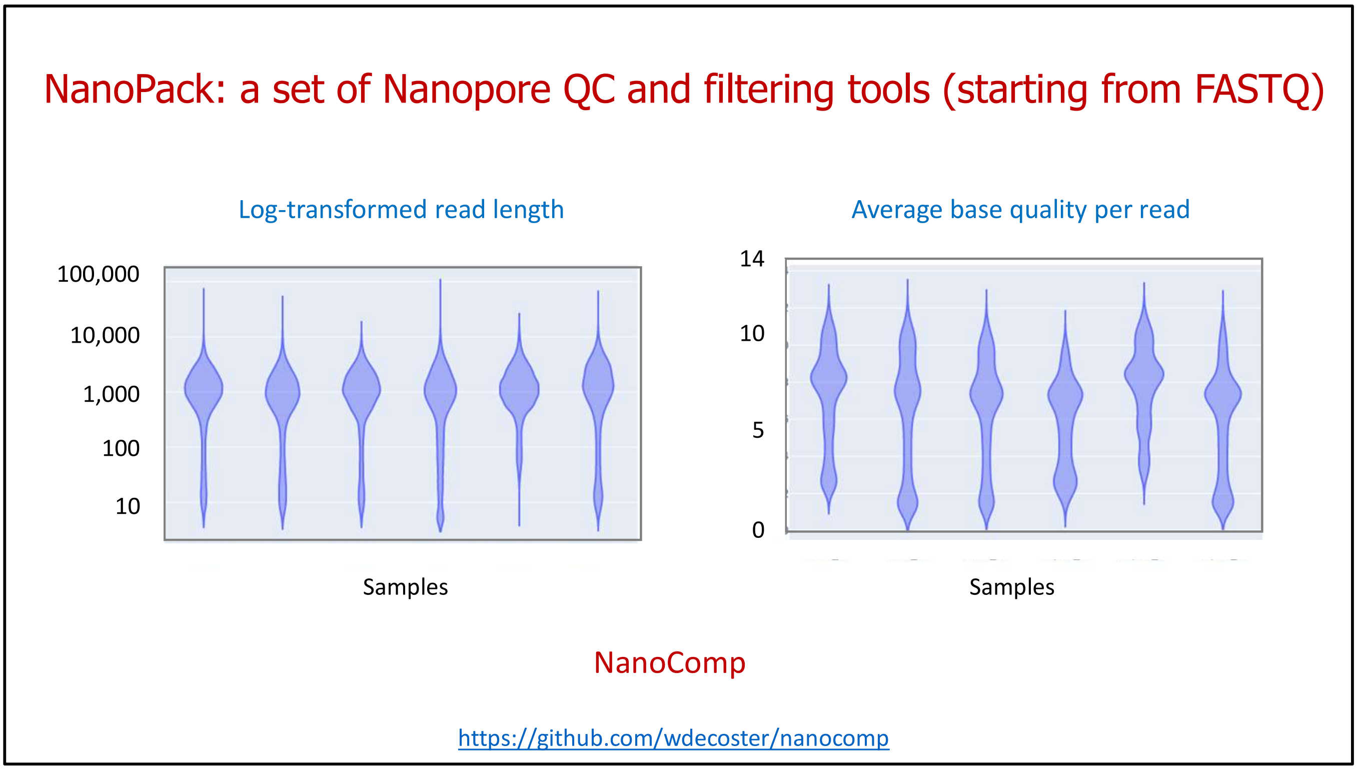
NanoPlot

Length vs Qual

https://github.com/wdecoster/NanoPlot

NanoPack - Simple and straight-to-the-point toolset for QC and filtering Nanopore data (using FASTQ files).

NanoPlot shows a simple and very informative plot of reads Length vs reads Quality.

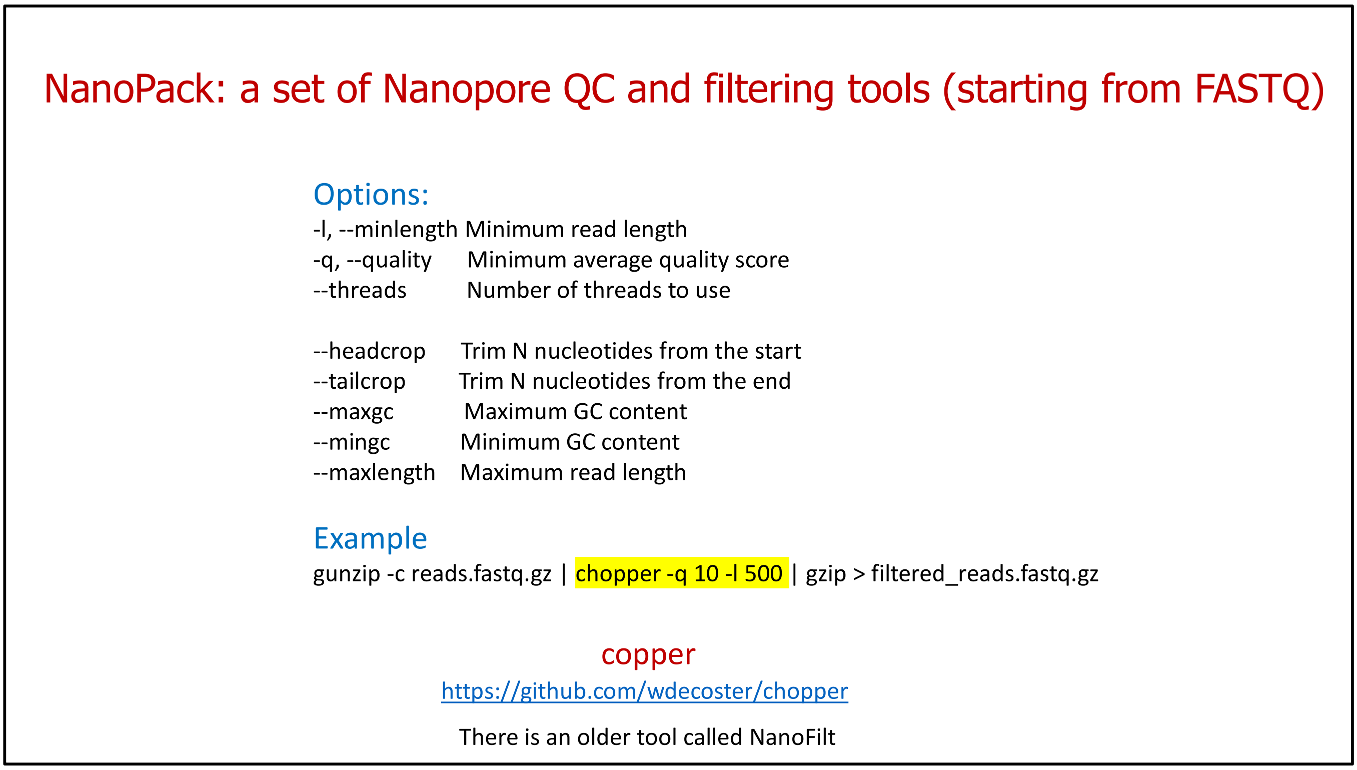


NanoPack: a set of Nanopore QC and filtering tools (starting from FASTQ)

NanoComp

https://github.com/wdecoster/nanocomp

NanoComp allows to compare quality metrics between multiple samples.
Actual plots could be interactive (showing additional information when mouse hoovers over the plot).



NanoPack: a set of Nanopore QC and filtering tools (starting from FASTQ)

Options:

-l, --minlength	Minimum read length
-q, --quality	Minimum average quality score
--threads	Number of threads to use
--headcrop	Trim N nucleotides from the start
--tailcrop	Trim N nucleotides from the end
--maxgc	Maximum GC content
--mingc	Minimum GC content
--maxlength	Maximum read length

Example

```
gunzip -c reads.fastq.gz | chopper -q 10 -l 500 | gzip > filtered_reads.fastq.gz
```

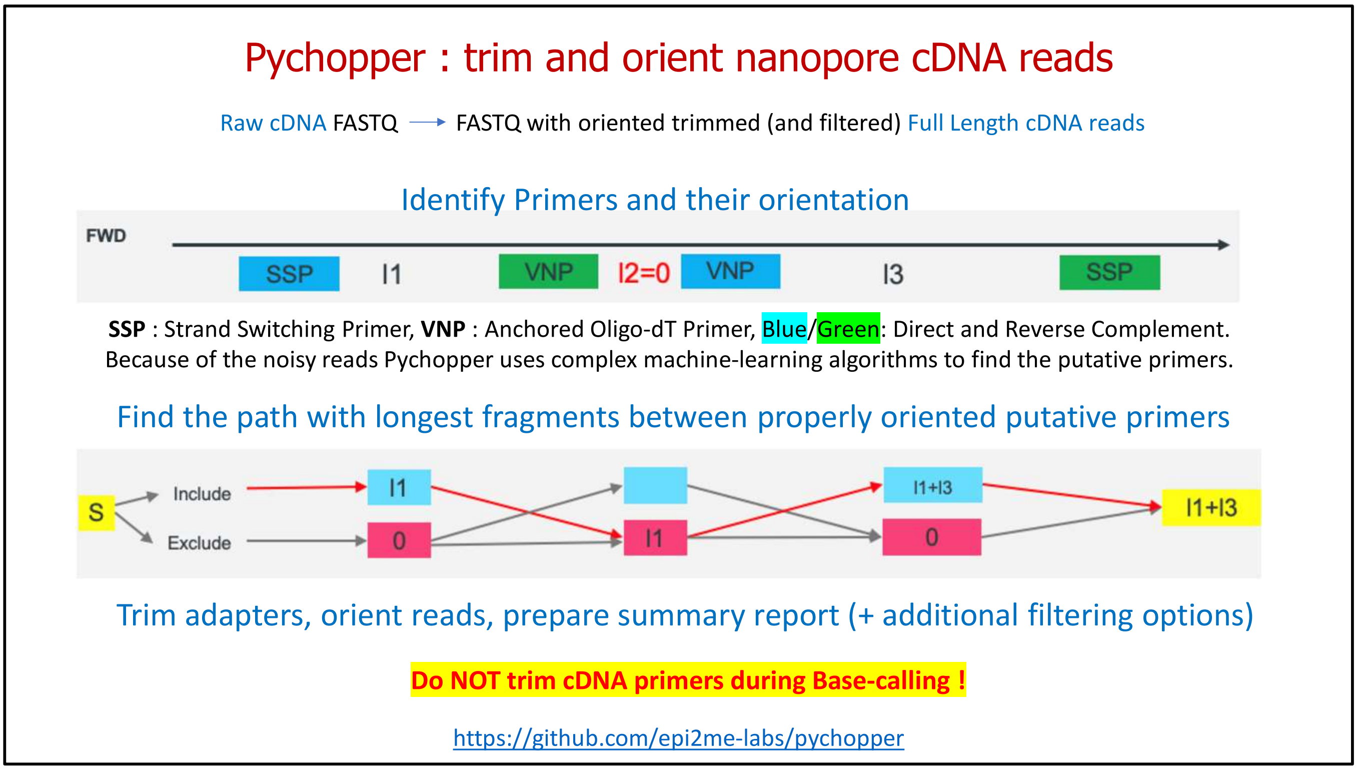
copper

https://github.com/wdecoster/chopper

There is an older tool called NanoFilt

After evaluating the quality of the reads, we need to remove bad reads.
A common practice for Nanopore data includes removal of reads by mean qual > 7 (compare to minimal Q20 in PacBio HiFi :)
Copper can filter reads basing on many different parameters too.

Important: NanoPack's ***copper*** is not the same as ***Pychopper*** discussed on the next slide!



Pychopper : trim and orient nanopore cDNA reads

Raw cDNA FASTQ → FASTQ with oriented trimmed (and filtered) Full Length cDNA reads

Identify Primers and their orientation

SSP : Strand Switching Primer, **VNP** : Anchored Oligo-dT Primer, Blue/Green: Direct and Reverse Complement.
Because of the noisy reads Pychopper uses complex machine-learning algorithms to find the putative primers.

Find the path with longest fragments between properly oriented putative primers

Trim adapters, orient reads, prepare summary report (+ additional filtering options)

Do NOT trim cDNA primers during Base-calling !

https://github.com/epi2me-labs/pychopper

Pychopper trims, orients and filters RNAseq reads.

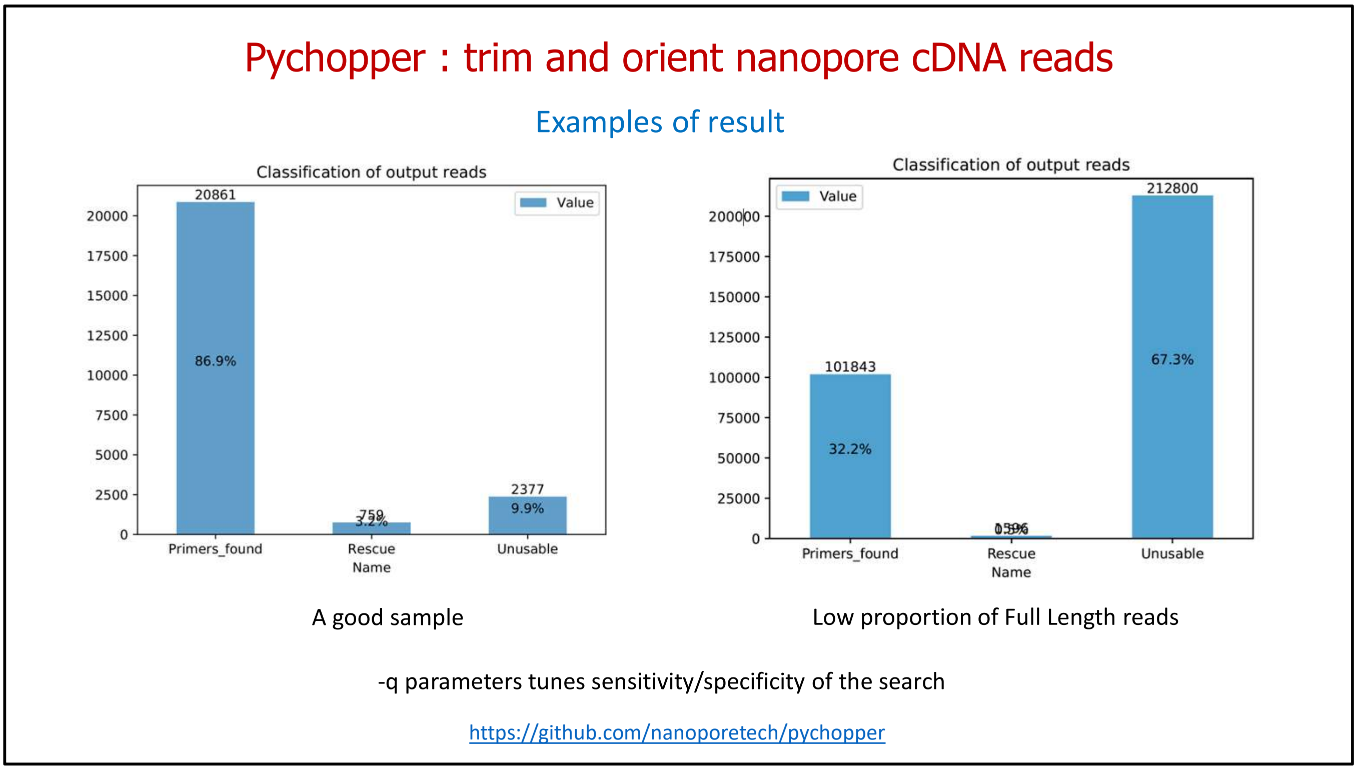
If you remember, initially raw RNAseq long reads should include strand switching primer (***SSP***) on one side and anchored oligo-dT primer (***VNP***) on the other.

The strand and orientation can be detected by the position and sequences (***direct*** or ***reverse-complement***) of these primers.

Because of the low quality of raw Nanopore reads (especially in old Nanopore data) ***Pychopper*** uses complex approximation algorithms to detect the primer sequences.

These are diagrams from the ***Pychopper*** web page, that supposedly should explain some of the employed algorithms.

To be honest, I struggle to understand these diagrams, but you are welcome to try it yourself (the link to web page is provided :)



Pychopper : trim and orient nanopore cDNA reads

Examples of result

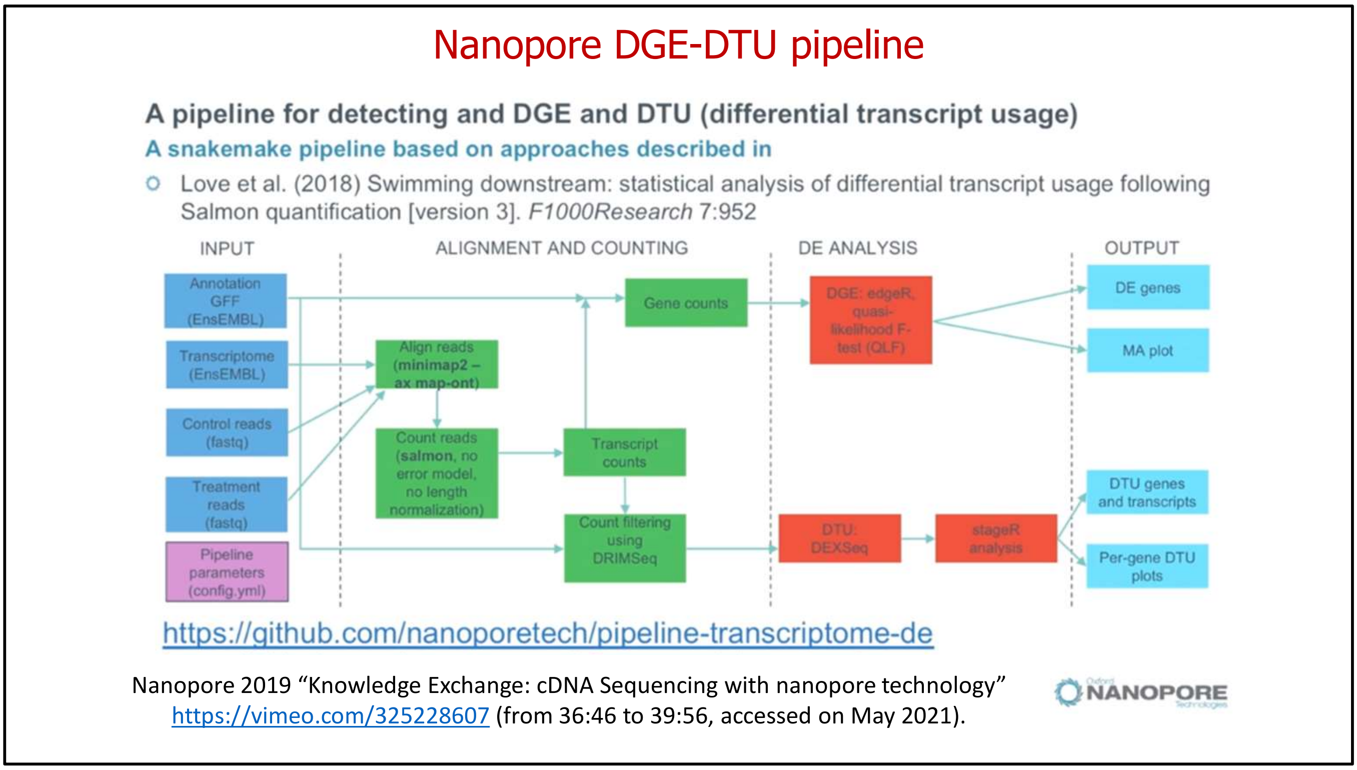
A good sample

Low proportion of Full Length reads

-q parameters tunes sensitivity/specificity of the search

https://github.com/nanoporetech/pychopper

The good thing about ***Pychopper*** is that it provides informative plots about the preprocessing results.
Here you can see a good and a bad example.
The proportion of rejected reads may be tuned by –q parameter when running Pychopper (in the recent versions of Pychopper it is tuned using a sub-set of reads).



Nanopore DGE-DTU pipeline

A pipeline for detecting and DGE and DTU (differential transcript usage)

A snakemake pipeline based on approaches described in

- Love et al. (2018) Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification [version 3]. *F1000Research* 7:952

INPUT
ALIGNMENT AND COUNTING
DE ANALYSIS
OUTPUT
Annotation GFF (EnsEMBL)
Transcriptome (EnsEMBL)
Control reads (fastq)
Treatment reads (fastq)
Pipeline parameters (config.yml)
Align reads (minimap2 – ax map-ont)
Count reads (salmon, no error model, no length normalization)
Gene counts
Transcript counts
Count filtering using DRIMSeq
DGE: edgeR, quasi-likelihood F-test (QLF)
DTU: DEXSeq
stageR analysis
DE genes
MA plot
DTU genes and transcripts
Per-gene DTU plots

https://github.com/nanoporetech/pipeline-transcriptome-de

Nanopore 2019 “Knowledge Exchange: cDNA Sequencing with nanopore technology” https://vimeo.com/325228607 (from 36:46 to 39:56, accessed on May 2021).

This slide illustrates the ***DGE-DTU pipeline*** that was recommended by Nanopore for quantification of transcript isoforms some years ago.

The pipeline started with properly oriented full-length reads prepared by Pychopper.
Then the pipeline
- Mapped reads to the known transcriptome with minimap2
- Passed mapped reads to Salmon for count
- Used these counts for DGE (Differential Gene Expression, edgeR) and
- Differential Transcript Usage analysis (DTU, DEXSeq & stageR)

All the required scripts were provided in the pipeline GitHub:
https://github.com/nanoporetech/pipeline-transcriptome-de
The sequential steps included in the pipeline were orchestrated using the ***Snakemake*** workflow manager.

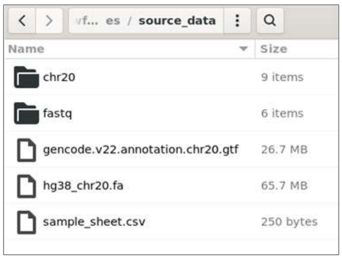
The currently recommended version of this Nanopore pipeline is called ***wf-transcriptomes***:
https://github.com/epi2me-labs/wf-transcriptomes
It implements the same steps, but uses the ***Nextflow*** workflow manager.
We will run the ***wf-transcriptomes*** pipeline during our practical session.

Nanopore wf-transcriptomes pipeline

https://github.com/epi2me-labs/wf-transcriptomes

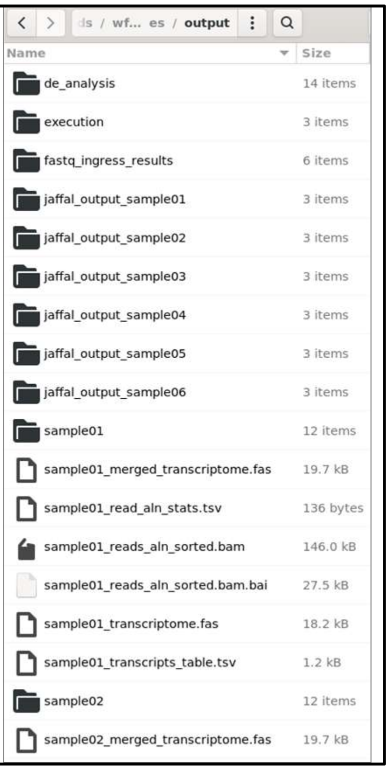
Input Folder → Output folder

1) Install pipeline and tools
2) Copy data and resources in the source folder
3) Run the pipeline



Name	Size
chr20	9 items
fastq	6 items
gencode.v22.annotation.chr20.gtf	26.7 MB
hg38_chr20.fa	65.7 MB
sample_sheet.csv	250 bytes

```
nextflow run epi2me-labs/wf-transcriptomes \
--fastq  source_data/fastq \
--de_analysis \
--ref_genome source_data/hg38_chr20.fa \
--transcriptome-source reference-guided \
--ref_annotation
  source_data/gencode.v22.annotation.chr20.gtf \
--direct_rna \
--minimap2_index_opts '-k 15' \
--sample_sheet source_data/sample_sheet.csv \
--jaffal_refBase source_data/chr20/ \
--jaffal_genome hg38_chr20 \
--jaffal_annotation genCode22 \
-profile standard
```



Name	Size
de_analysis	14 items
execution	3 items
fastq_ingress_results	6 items
jaffal_output_sample01	3 items
jaffal_output_sample02	3 items
jaffal_output_sample03	3 items
jaffal_output_sample04	3 items
jaffal_output_sample05	3 items
jaffal_output_sample06	3 items
sample01	12 items
sample01_merged_transcriptome.fas	19.7 kB
sample01_read_aln_stats.tsv	136 bytes
sample01_reads_aln_sorted.bam	146.0 kB
sample01_reads_aln_sorted.bam.bai	27.5 kB
sample01_transcriptome.fas	18.2 kB
sample01_transcripts_table.tsv	1.2 kB
sample02	12 items
sample02_merged_transcriptome.fas	19.7 kB

In practice, you first need to install the pipeline and tools following the authors' recommendations.
This step may require some IT knowledge, so this was already done on your VM.

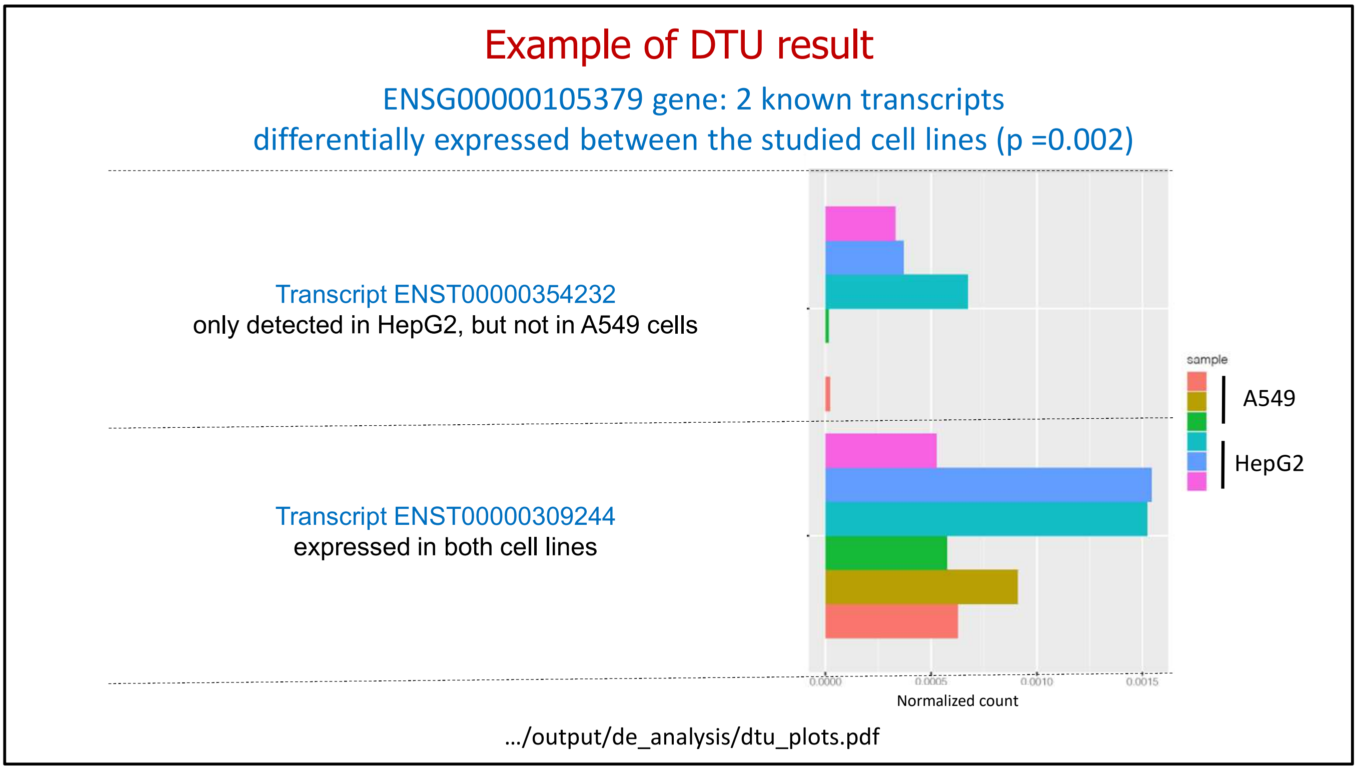
Then, all what you need to is:

- Add your source data (FASTQ files, reference genome, etc)
- Start the pipeline

After the pipeline completes, the results will appear as many files and sub-folders in the output folder.

Nanopore aims to develop user-friendly GUI tool to install and run its pipelines: EPI2ME Desktop.
At the moment, Nanopore plans works on connecting EPI2ME Desktop to a cloud, because user's machines may not have enough resources to run the analysis.

A copy of EPI2ME Desktop was installed into your VM.
However, it is not yet intuitive enough to use.
So, you will run the pipeline using the provided script.



Example of DTU result

ENSG00000105379 gene: 2 known transcripts
differentially expressed between the studied cell lines (p =0.002)

Transcript ENST00000354232
only detected in HepG2, but not in A549 cells

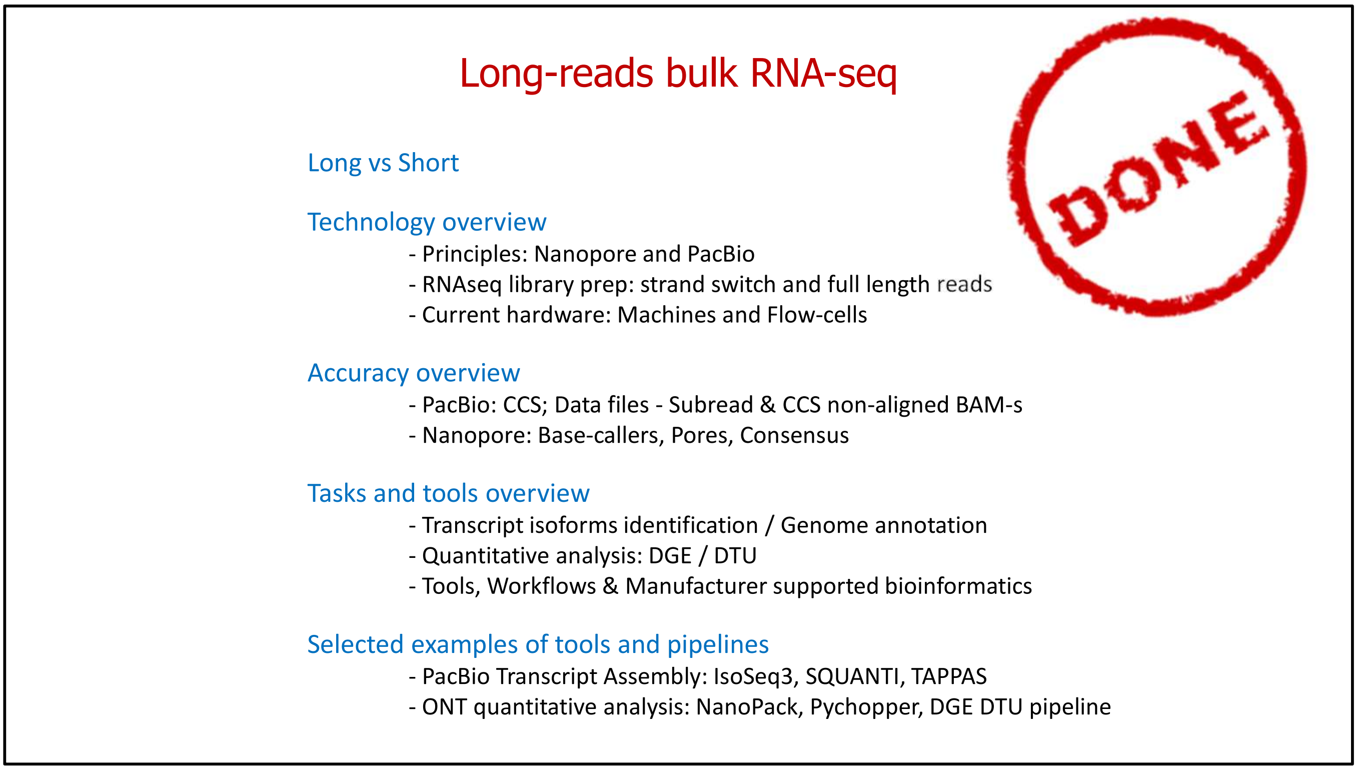
Transcript ENST00000309244
expressed in both cell lines

…/output/de_analysis/dtu_plots.pdf

This is an example of DTU analysis produced by the pipeline.
In wf-transcriptomes output folder you may find such plots
…/output/de_analysis/dtu_plots.pdf file.

This example compares two cell lines: A549 and HepG2.
You can see that a specific gene (ETFB = ENSG00000105379) has two alternative transcripts.
One of them (ENST00000309244) is strongly expressed in both cell lines.
While the other (ENST00000354232) is only expressed in HepG2.

The output folder contains much more results, including the outputs of every involved tool, such as FASTQs trimmed by Pychopper (if cDNA RNAseq was analyzed), BAM files produced by minimap2 etc. You will be able to explore the content of the ***wf-transcriptomes*** output folder during the practical session.



Long-reads bulk RNA-seq

DONE

Long vs Short

Technology overview

- Principles: Nanopore and PacBio
- RNAseq library prep: strand switch and full length reads
- Current hardware: Machines and Flow-cells

Accuracy overview

- PacBio: CCS; Data files - Subread & CCS non-aligned BAM-s
- Nanopore: Base-callers, Pores, Consensus

Tasks and tools overview

- Transcript isoforms identification / Genome annotation
- Quantitative analysis: DGE / DTU
- Tools, Workflows & Manufacturer supported bioinformatics

Selected examples of tools and pipelines

- PacBio Transcript Assembly: IsoSeq3, SQUANTI, TAPPAS
- ONT quantitative analysis: NanoPack, Pychopper, DGE DTU pipeline

Selected references

Pardo-Palacios F. *et al.* **2023**: Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. bioRxiv https://www.biorxiv.org/content/10.1101/2023.07.25.550582v1

Kovaka S. *et al.* **2023:** Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. Nat Methods https://doi.org/10.1038/s41592-022-01716-8

Foord C. *et al.* **2023:** The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing Nat Methods https://doi.org/10.1038/s41592-022-01715-9

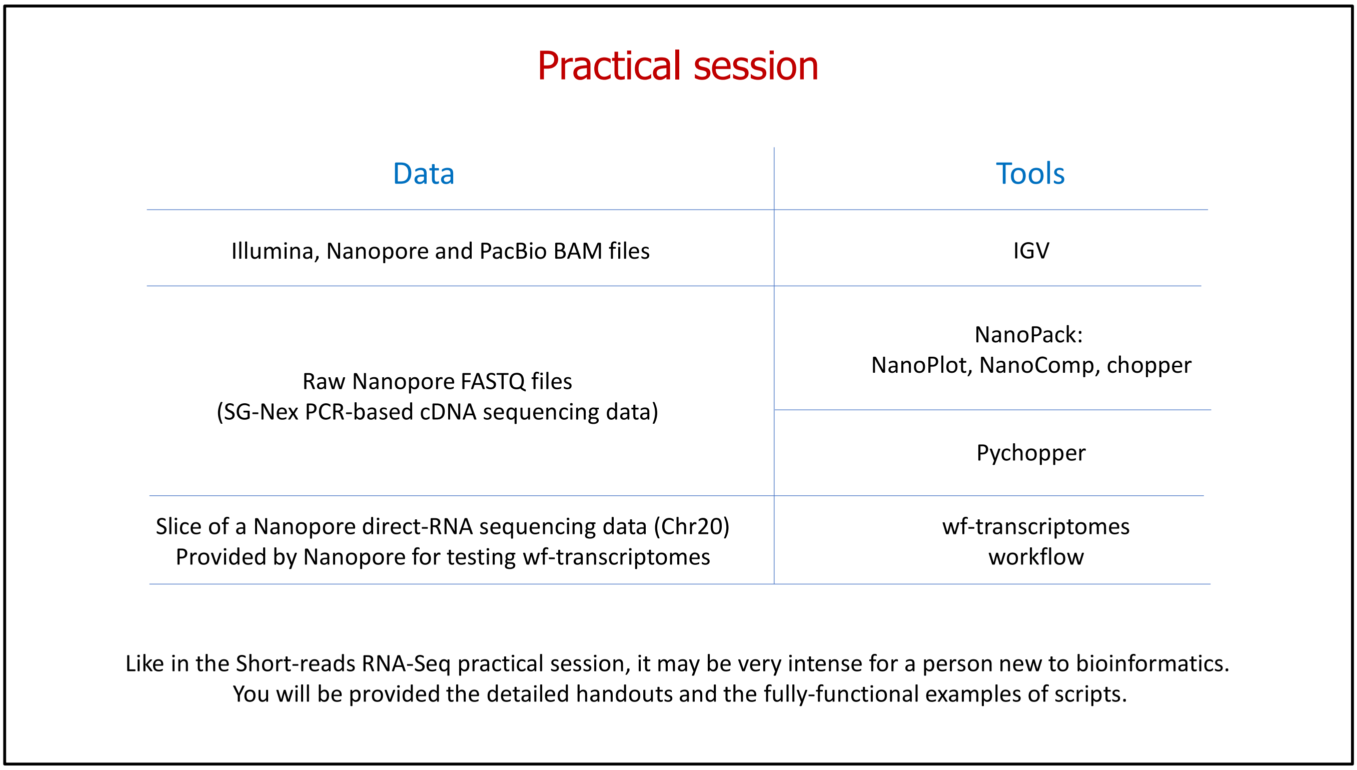
Logsdon G. *et al.* **2020:** Long-read human genome sequencing and its applications. *Nat Rev Genet*. https://doi.org/10.1038/s41576-020-0236-x

Amarasinghe S. *et al.* **2020:** Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* https://doi.org/10.1186/s13059-020-1935-5

Stark R. *et al.* **2019:** RNA sequencing: the teenage years. *Nat Rev Genet.* https://doi.org/10.1038/s41576-019-01

Wenger A. *et al.* **2019:** Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* https://doi.org/10.1038/s41587-019-0217-9

Love M. et al **2018:** Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification *F1000Research,* https://doi.org/10.12688/f1000research.15398.3



Practical session

Data	Tools
Illumina, Nanopore and PacBio BAM files	IGV
Raw Nanopore FASTQ files (SG-Nex PCR-based cDNA sequencing data)	NanoPack: NanoPlot, NanoComp, chopper
	Pychopper
Slice of a Nanopore direct-RNA sequencing data (Chr20) Provided by Nanopore for testing wf-transcriptomes	wf-transcriptomes workflow

Like in the Short-reads RNA-Seq practical session, it may be very intense for a person new to bioinformatics.
You will be provided the detailed handouts and the fully-functional examples of scripts.

Samples are selected by size !
Don’t over-interpret data quality and biology :)