Good [morning/afternoon], we're presenting a study by Uapinyoying et al. (2020) that demonstrates how **long-read sequencing** can uncover complex transcript diversity in some of the largest and most repetitive genes in the mammalian genome.
RNA-seq has revolutionized transcriptomics, but short-read sequencing like Illumina struggles with ultra-long, highly repetitive transcripts — especially in muscle structural genes like titin (106 kb), nebulin (22 kb), and Nrap (5.5 kb).
These genes have hundreds of exons, complex alternative splicing, and tissue-specific isoforms that are often missed or misassembled with short reads.
This study addresses this limitation using PacBio Iso-Seq, a long-read RNA-seq platform, paired with new bioinformatic pipelines to resolve full-length isoforms and unannotated exons of these large muscle genes. Results were confirmed by PCR and Sanger sequencing.

- exCOVator: Identifies differential exon usage and new exons - find unannotated exons.

- exPhaser is used to quantify and annotate splicing patterns within larger transcript structures. It takes as input cassette exons identified by exCoVator, which are involved in defining isoformse

The team used mouse tissues: fast-twitch (EDL), slow-twitch (soleus), and cardiac muscle.

They isolated 5–10 kb mRNA and applied PacBio SMRT sequencing with size-selected Iso-Seq libraries. Even with long reads (~10 kb), titin and nebulin are longer than read length, so full-transcript coverage isn't always possible.

Their key findings:

Nrap (~5.5kb):

- Found differential splicing of exon 12, previously thought to be exclusive to cardiac tissue.

- Nrap-c (lacking exon 12) was assumed to be cardiac-specific, but is actually **expressed in all three muscle types**, disproving the old classification

- RT-PCR and Sanger sequencing validated this unexpected result.

Nebulin (~22 kb):

- Too long for full coverage, but **internal oligo(dT) priming** enabled partial read-through
- Found novel exons and muscle-type-specific splicing between exons 137–152.
- Exon 138 was included in soleus (slow) and excluded in EDL (fast).
- Also identified an unannotated exon (u-002), conserved across species.
- Phased 14 exons in the Z-disk/super-repeat region to quantify isoforms — nearly all known transcripts in RefSeq/Gencode were incorrect or missing.

Titin (~106 kb):

- The largest known human protein, too long for any current read
- They showed **exon 191**, thought to be constitutive, is actually a **cardiac-specific cassette exon**, spliced out in 64% of heart transcripts
- This was confirmed by RT-PCR and supported by human GTEx and DCM datasets.
- Other cardiac-specific or alternative exons were also mapped and phased.

This study shows that with the right computational pipelines, **long-read sequencing can quantify differential exon usage and isoform expression**, even in genes >100 kb long.

It corrects long-standing assumptions in gene annotations, reveals new exons, and connects splicing differences to muscle specialization.

- **285 differentially used exons** and **14 novel exons** discovered
- **Transcript-level phasing and quantification** of large genes
- **Correction of current annotations** – most observed isoforms were *not* in RefSeq or GENCODE
- The method even detected **rare isoforms** and **novel transcripts**

And most importantly:

From a technical standpoint, this is one of the **first practical demonstrations** that PacBio long-read sequencing can be used for **semi-quantitative** analysis — not just transcript discovery. This *approach opens the door to differential expression analysis of difficult transcripts* — those that have long been considered beyond the reach of traditional RNA-seq methods.

This has clear applications in **muscle biology, genome annotation, and clinical diagnostics** to manage cardiac disorders in human