Poster P21.027.B

Assessment of read depth requirements for gene and isoform discovery: a comparative study of long-read and short-read RNA sequencing data in human heart and brain

Nina Gonzaludo*[1], Jocelyne Bruand[1], Amy Klegarth[1], Jason Underwood[1], Elizabeth Tseng[1], Birth Defects Research Laboratory[2], Kimberly A. Aldinger[3]

1. PacBio, Menlo Park, CA, USA, 2. University of Washington, Seattle, WA, USA, 3. Seattle Children's Research Institute, Seattle, WA, USA

Long-read sequencing enables full-length isoform characterization

Advancements in long-read sequencing technology have revolutionized transcriptomics research, allowing for full-length, comprehensive capture of transcript isoforms without the need for cDNA fragmentation and computational assembly methods, as is required for short-read RNA-seq (100-200 bp). With long-read RNA sequencing, researchers can improve gene and isoform discovery and annotation compared to traditional short-read RNA-seq.

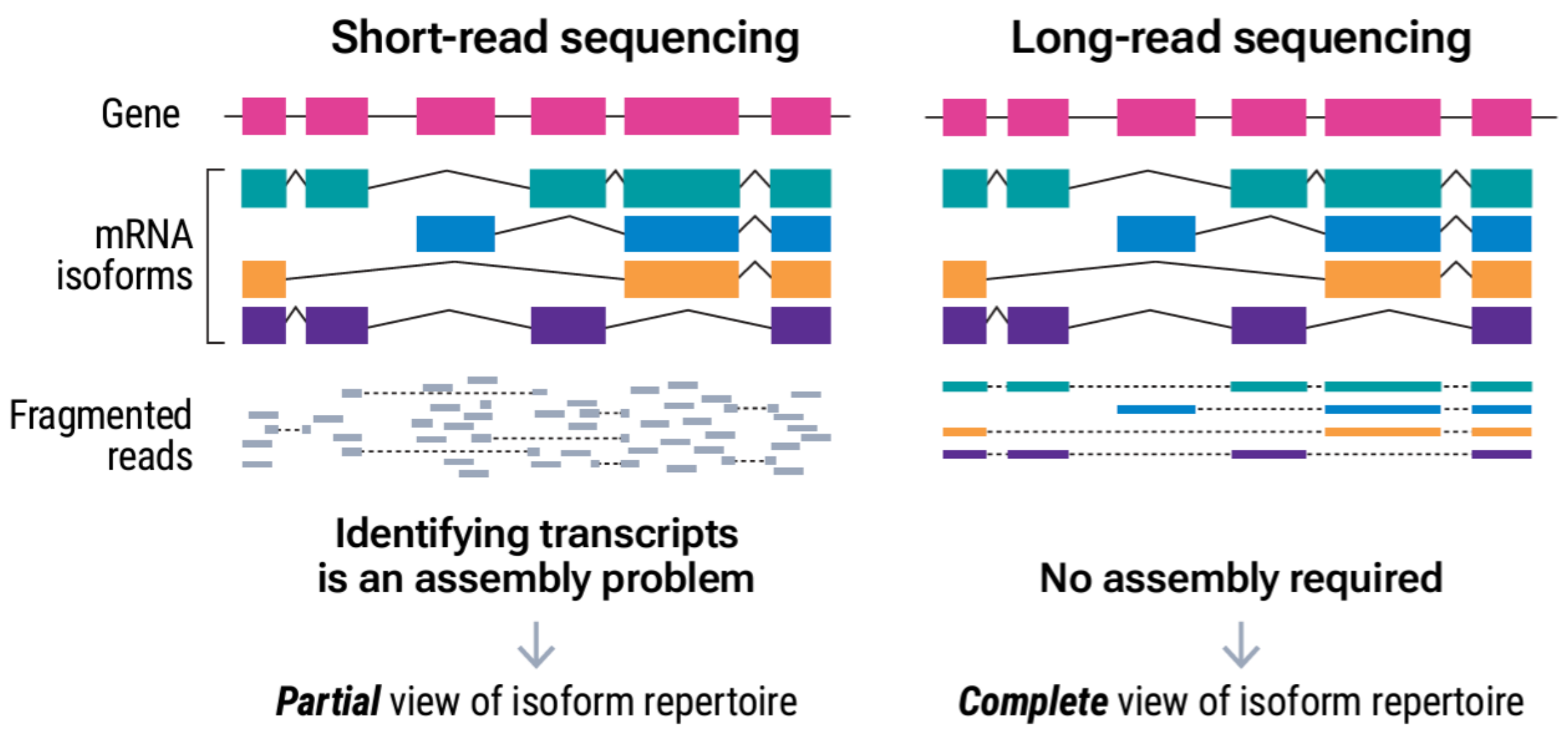
Short-read sequencing
Long-read sequencing
Gene
mRNA isoforms
Fragmented reads
Identifying transcripts is an assembly problem
No assembly required
Partial view of isoform repertoire
Complete view of isoform repertoire

Figure 1. Long-read RNA sequencing with the PacBio Iso-Seq method does not require transcript assembly, enabling full-length cDNA sequencing and providing a complete, unambiguous view of the transcriptome.

However, determining the optimal read depth for robust annotation and isoform discovery remains a challenge. This study investigated the read depth requirement for gene and isoform discovery using PacBio's full-length Kinnex kits for RNA sequencing, when compared to short-read RNA-seq using Illumina.

Full-length RNA sequencing using PacBio Kinnex

Kinnex full-length RNA kits combine full-length isoform sequencing with a concatenation method of combining smaller amplicons into larger fragment libraries for throughput increase. This enables full-length isoform discovery and capture of abundance information.

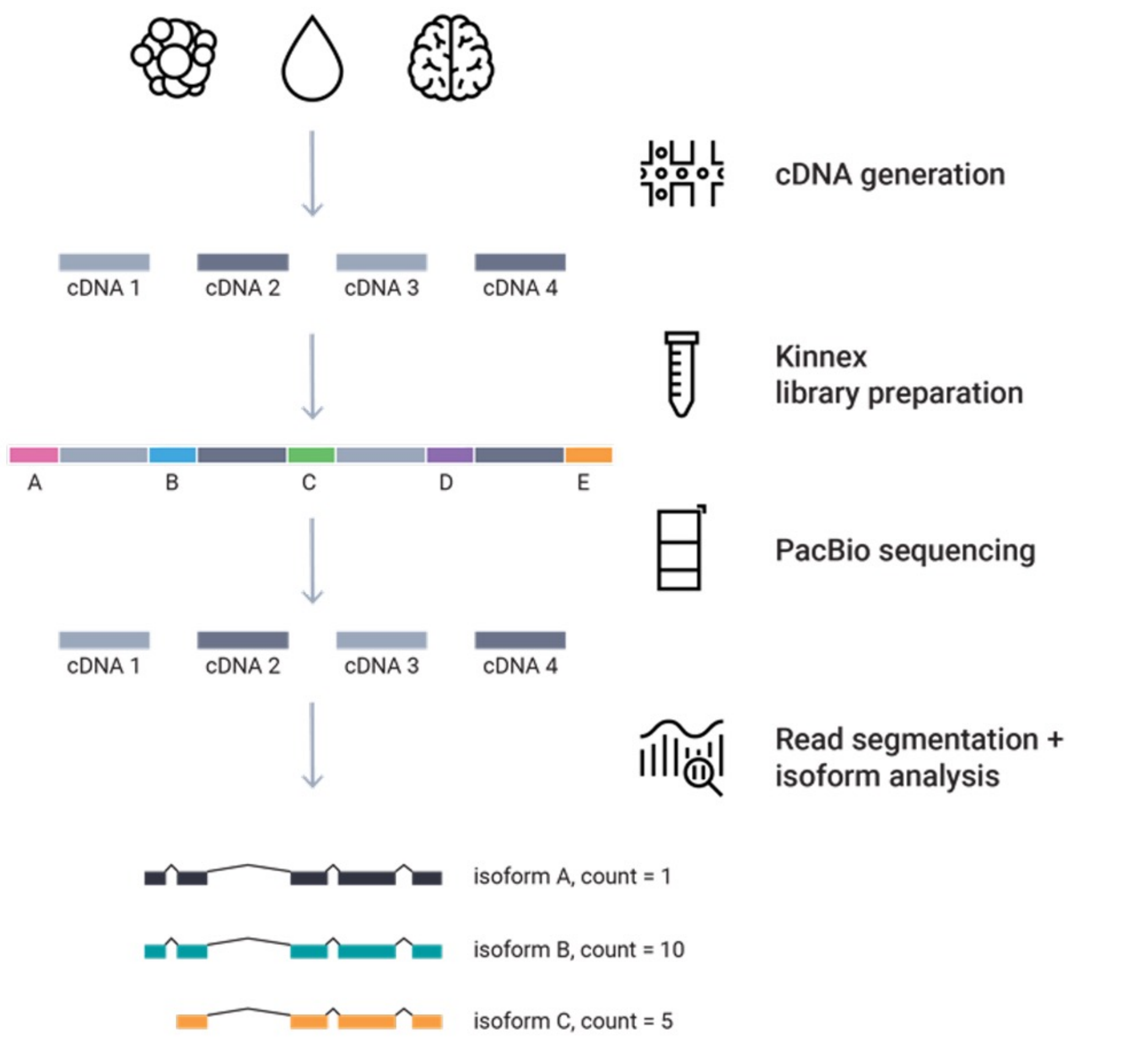

Figure 2. Kinnex full-length RNA library and analysis workflow.

- Total RNA as input (300 ng, RIN ≥ 7)
- Generates barcoded cDNA (up to 12-plex) using Iso-Seq express 2.0 kit
- Create Kinnex libraries by concatenating 8 cDNA into an array
- Sequence on PacBio Sequel II, IIe, or Revio systems
- SMRT Link outputs isoform read count information
- Achieve 15 million reads on Sequel II and IIe systems or 40 million reads on Revio system

Conflict of interest disclosure: all PacBio authors listed are shareholders and employees of PacBio. Dr. Aldinger has nothing to disclose.

Long & short-read RNA sequencing on human heart and brain samples

Long-read RNA-Seq was performed using the PacBio Kinnex full-length RNA kit and sequenced with one Revio SMRT Cell per each of the 8 samples (Table 1) on a Revio system. Short-read RNA-seq was performed on heart samples using Illumina TruSeq Stranded mRNA kits, with sequencing on NovaSeq6000.

Sample	Type	Age (postconceptional day)	Age (postconceptional week)
Heart 1 (Trisomy 21)	Bulk	98	14
Heart 2 (Trisomy 21)	Bulk	137	20
Heart 3 (Control)	Bulk	137	20
Heart 4 (Control)	Bulk	96	14
Cerebellum 1	Purkinje cell layer (PCL)		15
Cerebellum 2	External granule cell layer (EGL)		15
Cerebellum 3	Bulk		15
Cerebellum 4	Bulk		14

Table 1. RNA was extracted from developing organs for isoform discovery. Short-read RNA-sequencing was performed on heart.

	Heart 1* (T21)	Heart 2* (T21)	Heart 3* (Control)	Heart 4* (Control)	Cerebellum 1*	Cerebellum 2*	Cerebellum 3	Cerebellum 4*
HiFi Reads (millions)	6.35	6.06	6.56	6.34	6.32	7.17	6.05	7.49
Transcripts (S-reads, millions)	35.3	31.9	33.3	36.8	46.3	51.5	38.6	43.9
Mean transcript length (bp)	2,363	2,367	2,255	2,287	2,075	2,073	2,426	2,007
% reads w/ full arrays	61.82%	58.33%	56.16%	65.63%	85.83%	84.13%	71.81%	64.53%
Mean array size (concat factor)	5.56	5.26	5.09	5.8	7.33	7.19	6.38	5.86

Table 2. HiFi sequencing results, 1 sample per Revio SMRT Cell. *Sample HiFi data publicly available at https://downloads.pacbcloud.com/public/dataset/Kinnex-full-length-RNA/

Read depth sensitivity analysis for gene and isoform discovery

To simulate lower read depths and isoform discovery sensitivity, long reads were down-sampled. Samples of similar types or specimens were combined for simpler comparison. For these samples, unique known gene and isoform counts peaked at roughly 25k genes and 60-80k isoforms. Down-sampling results suggest that 80% of known genes and isoforms may be detectable at 10-20M reads per sample.

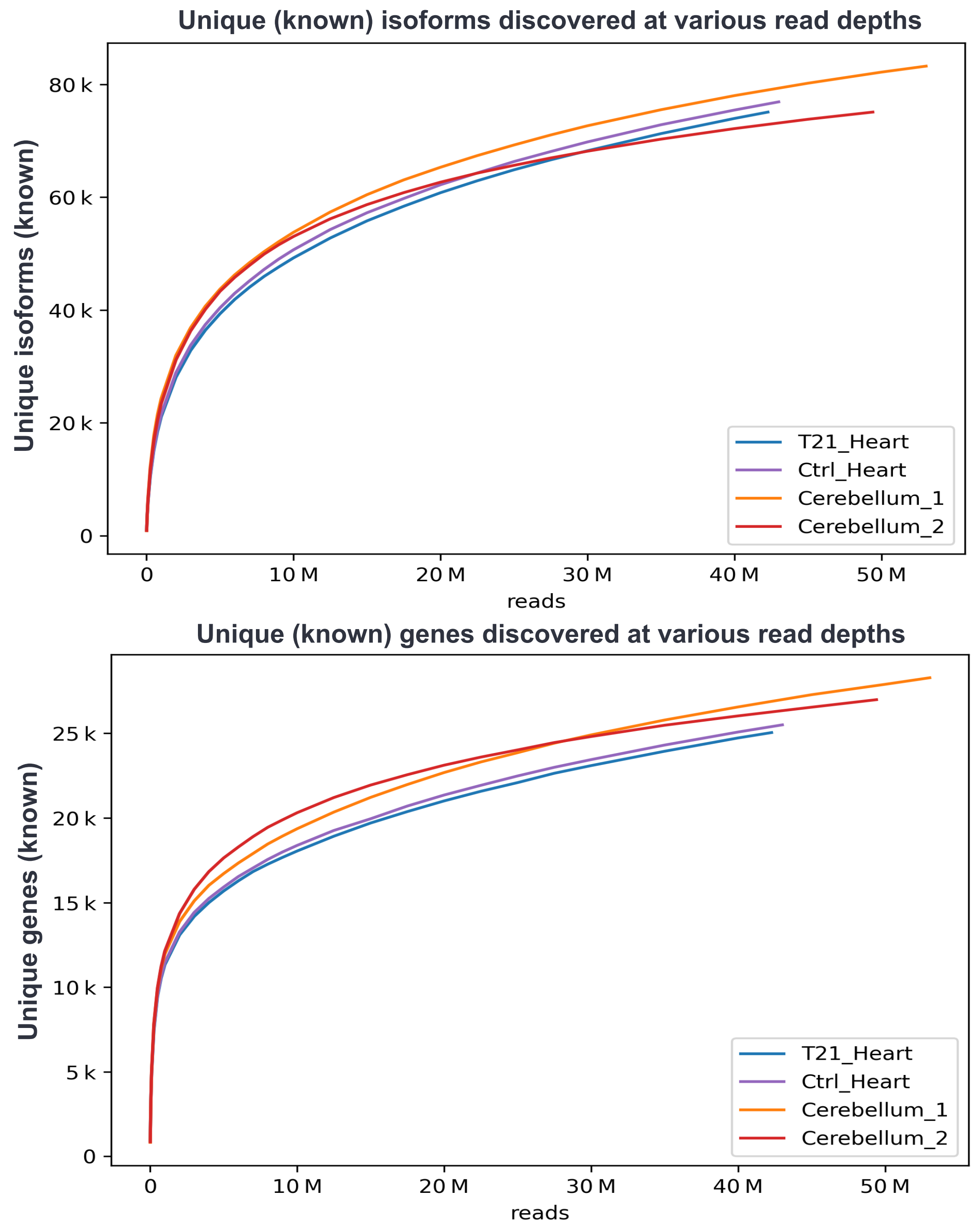

Figure 3. Saturation curves were computed for observations of > 1 HiFi read per isoform. Heart samples 1 and 2 were combined as "T21_Heart", and 3 and 4 as "Ctrl_Heart". Cerebellum samples 1 and 2 from the same specimen were combined as "Cerebellum_1", and 3 and 4 as "Cerebellum_2". Unique known isoforms were quantified at down-sampled thresholds to simulate lower long-read sequencing read depths.

Fewer new isoforms discovered after 20M reads

To estimate reads needed to saturate all known genes and isoforms, we compute the slope for each interval of down-sampled reads, then identified the point at which the slope falls below 0.1%. That is, we identified the point at which each additional 1000 reads results in less than 1 new gene or isoform discovered.

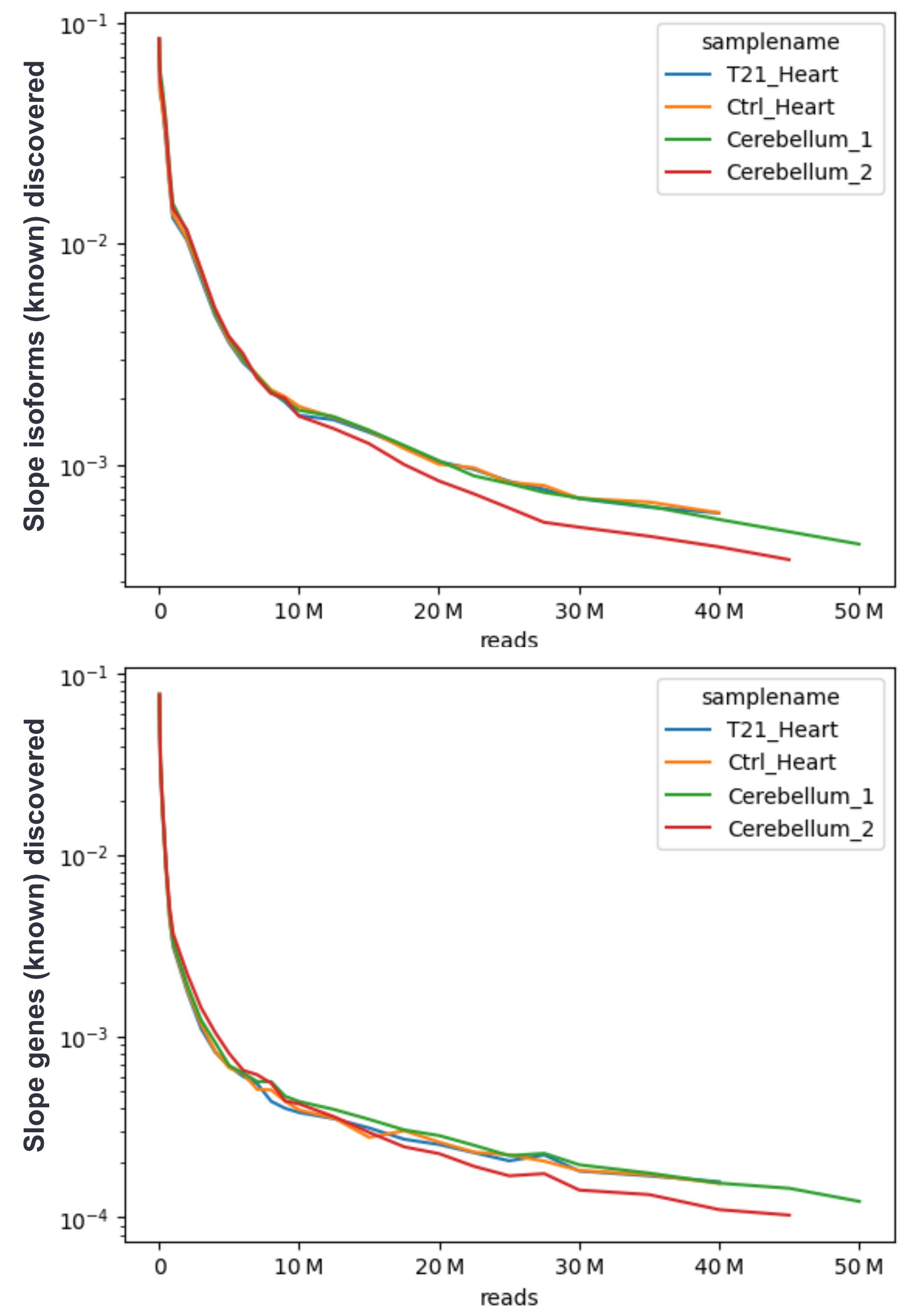

Figure 4. Isoform and gene discovery slopes by read depth.

Based on these samples, we reached a slope of <0.1% new isoform per read discovered at around 20M reads across all samples. For genes, a slope of <0.1% new genes discovered occurred at around 4M reads for heart samples and 5M reads for brain samples.

Kinnex discovers more genes than short reads at lower depth

In general, long-read RNA-sequencing identified more genes overall compared to short-read RNA-seq.

	PacBio Kinnex			Illumina short-read RNA-seq		
Sample	Transcripts (S-reads, millions)	Known Genes	Novel Genes	Total Reads (QC passed, millions)	Unique gene IDs (FPKM > 1)	Unique gene IDs (TPM>1)
Heart 1 (T21)	35.3	21,429	4,348	123.6	14,387	14,871
Heart 2 (T21)	31.9	22,011	4,682	81.5	15,005	15,802
Heart 3 (Control)	33.4	22,152	5,051	91.4	14,934	15,600
Heart 4 (Control)	36.8	22,306	5,153	81.6	14,869	15,546

Table 3. Unique genes identified by both platforms for heart samples, based on GENCODE 39 annotation.

Conclusions

- The PacBio full-length Kinnex RNA kit provides complete transcript coverage for isoform and gene discovery in tissues of interest, enabling understanding of biology and disease.
- With Kinnex, fewer long reads reads are needed for gene discovery compared to short-read RNA seq.
- The majority of known genes and isoforms can be discovered using full-length Kinnex kits at 10-20M reads per sample, suggesting multiplexing may be a cost-effective yet comprehensive option.

For more information, visit pacb.com/Kinnex