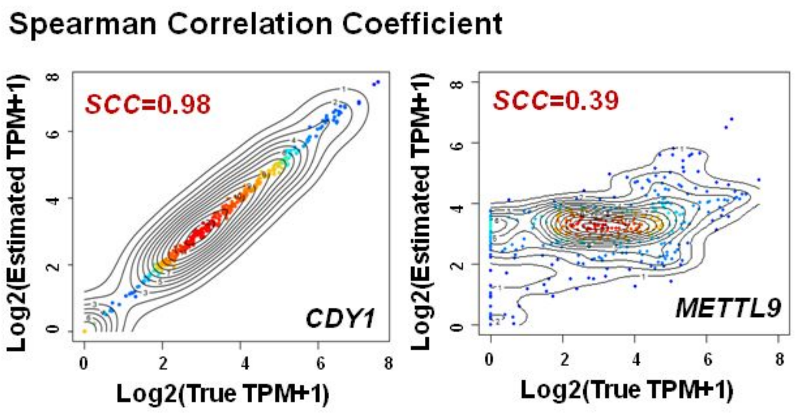
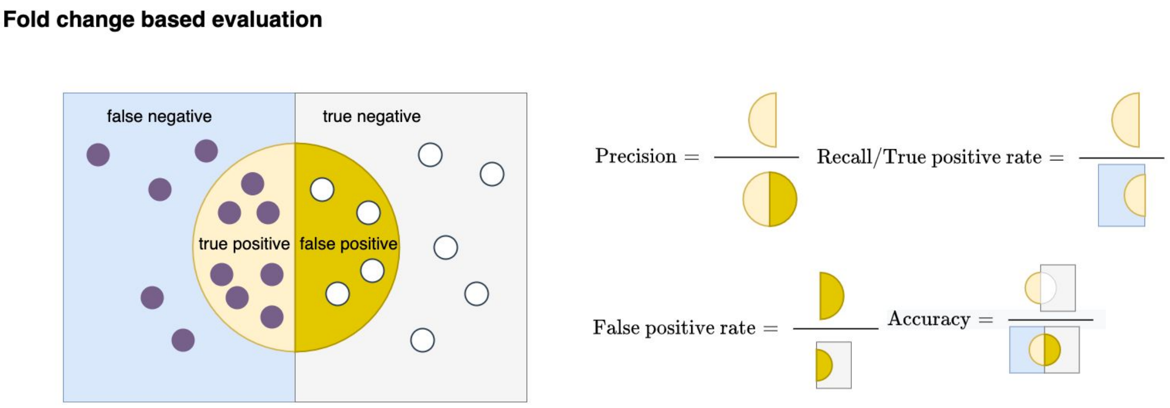
Supplementary Figures

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Pardo-Palacios et al.

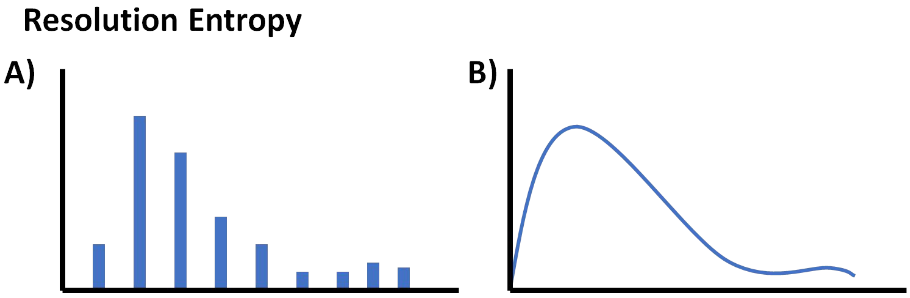


Supplementary Fig. 1: Spearman Correlation Coefficient (*SCC*) between the estimation and gold standard. The simulation study based on *SCC* reveals gene *CDY1* can be accurately quantified but not gene *METTL9*.

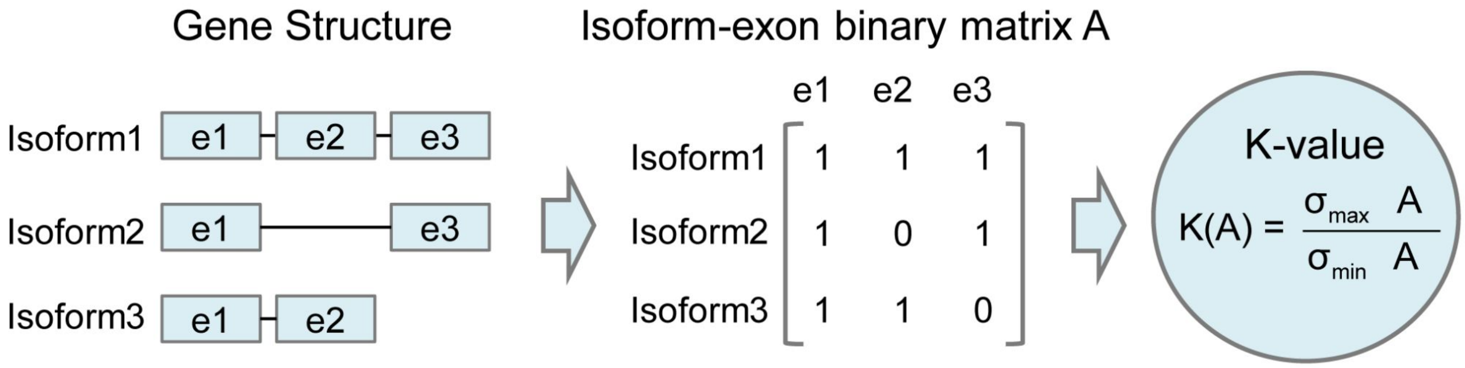


True positive: **actually** differentially expressed and quantified as **differentially expressed**
False positive: **not** differentially expressed and quantified as **differentially expressed**
True negative: **not** differentially expressed and quantified as **not** differentially expressed
False negative: **actually** differentially expressed and quantified as **not** differentially expressed

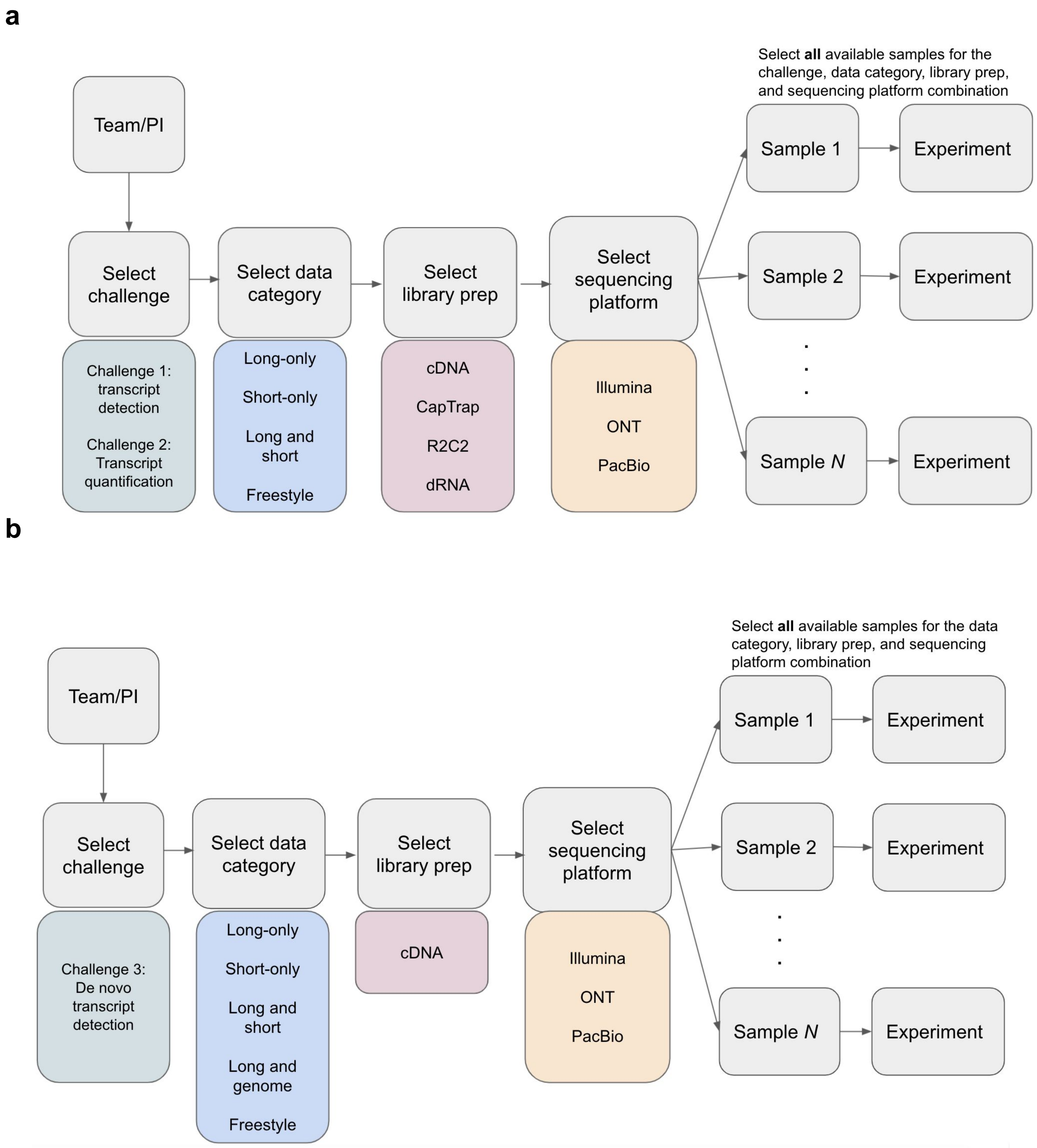
Supplementary Fig. 2: Log-fold-change-based evaluation metrics. This figure illustrates how ROC statistics such as precision, recall and accuracy are calculated. It measures the performance on detecting real biological changes.



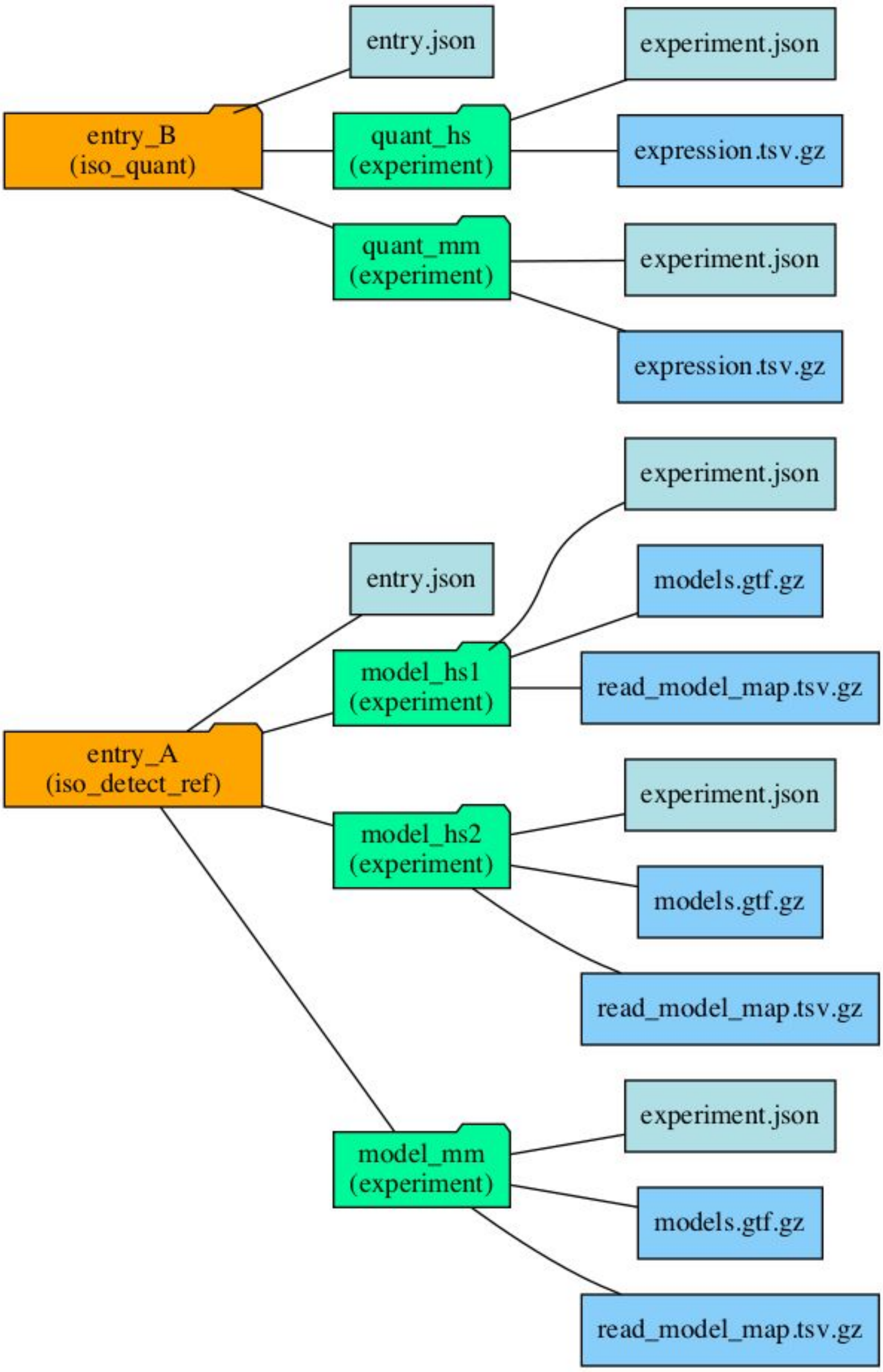
Supplementary Fig 3. Resolution Entropy. (A) The software output only a few certain discrete values has lower resolution entropy as it cannot capture the continuous and subtle difference of gene expressions. (B) The software with continuous output values has higher resolution entropy.



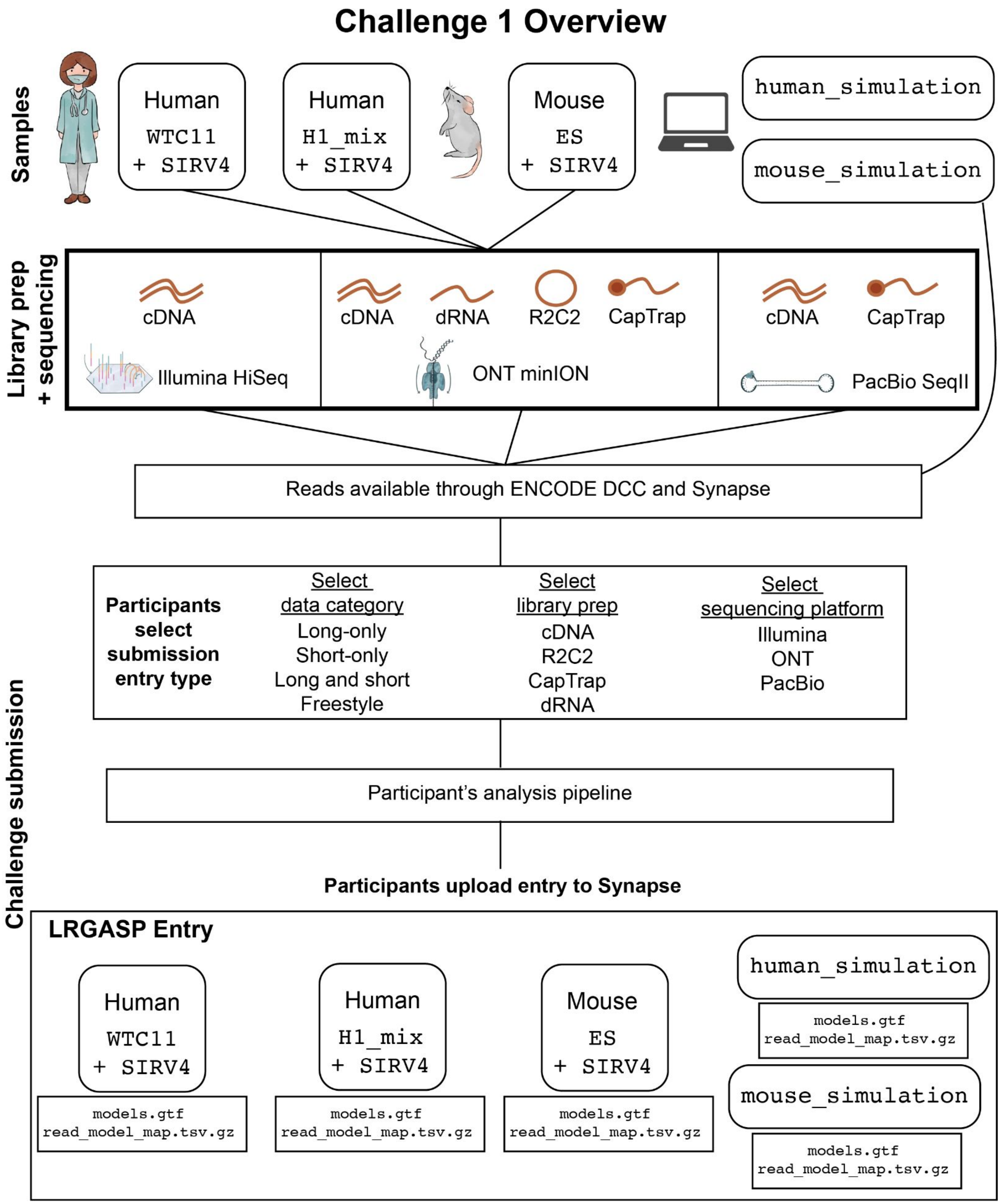
Supplementary Fig. 4: Description of K-value. A measure of the complexity of exon-isoform structures for each gene.



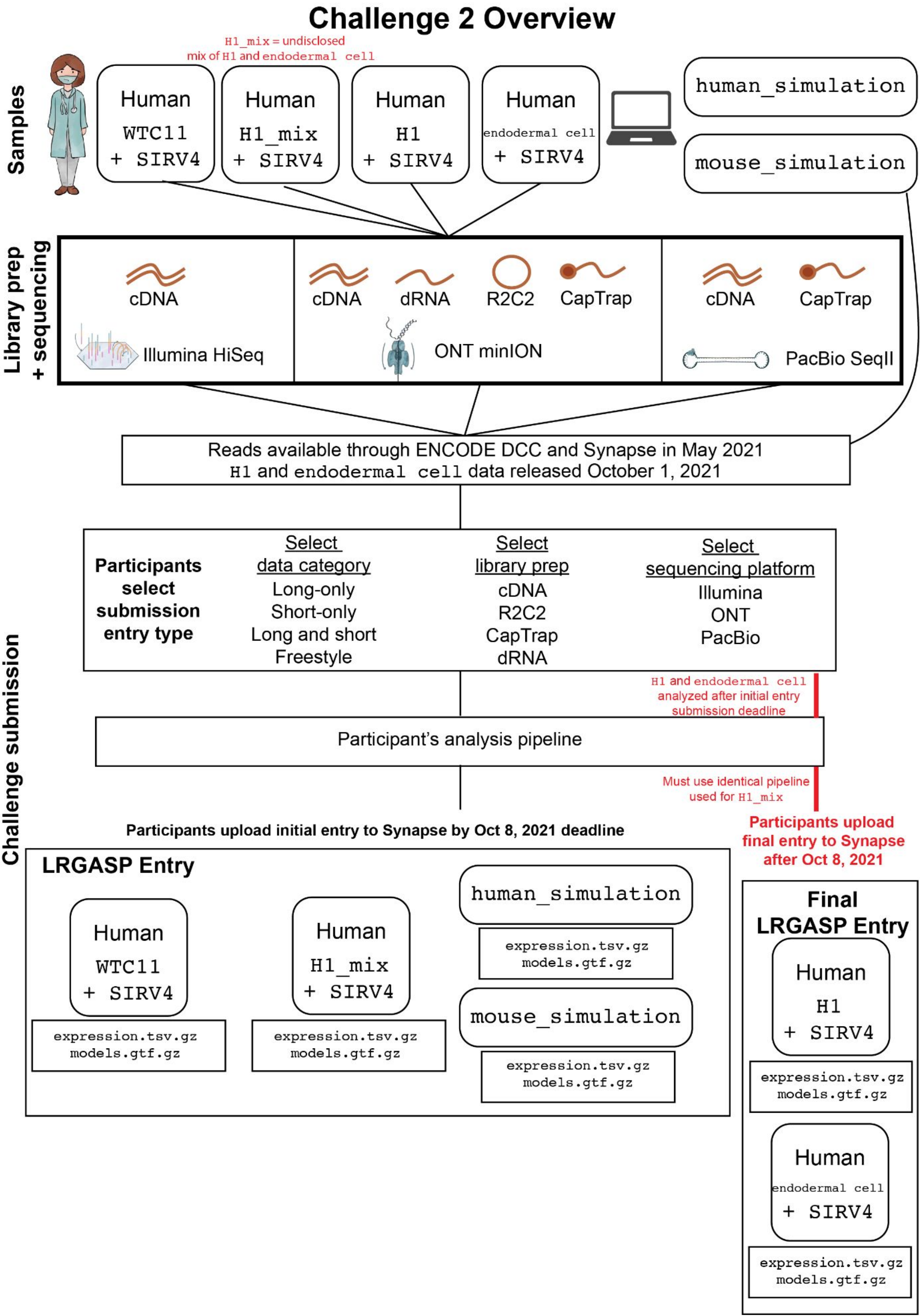
Supplementary Fig. 5: Challenge submission. a, Overview of submissions to Challenges 1 and 2. Each entry will be derived from a specific data category, library prep, and sequencing platform combination. All available samples for the selected combination must be included in an entry **b**, Overview of submissions for Challenge 3



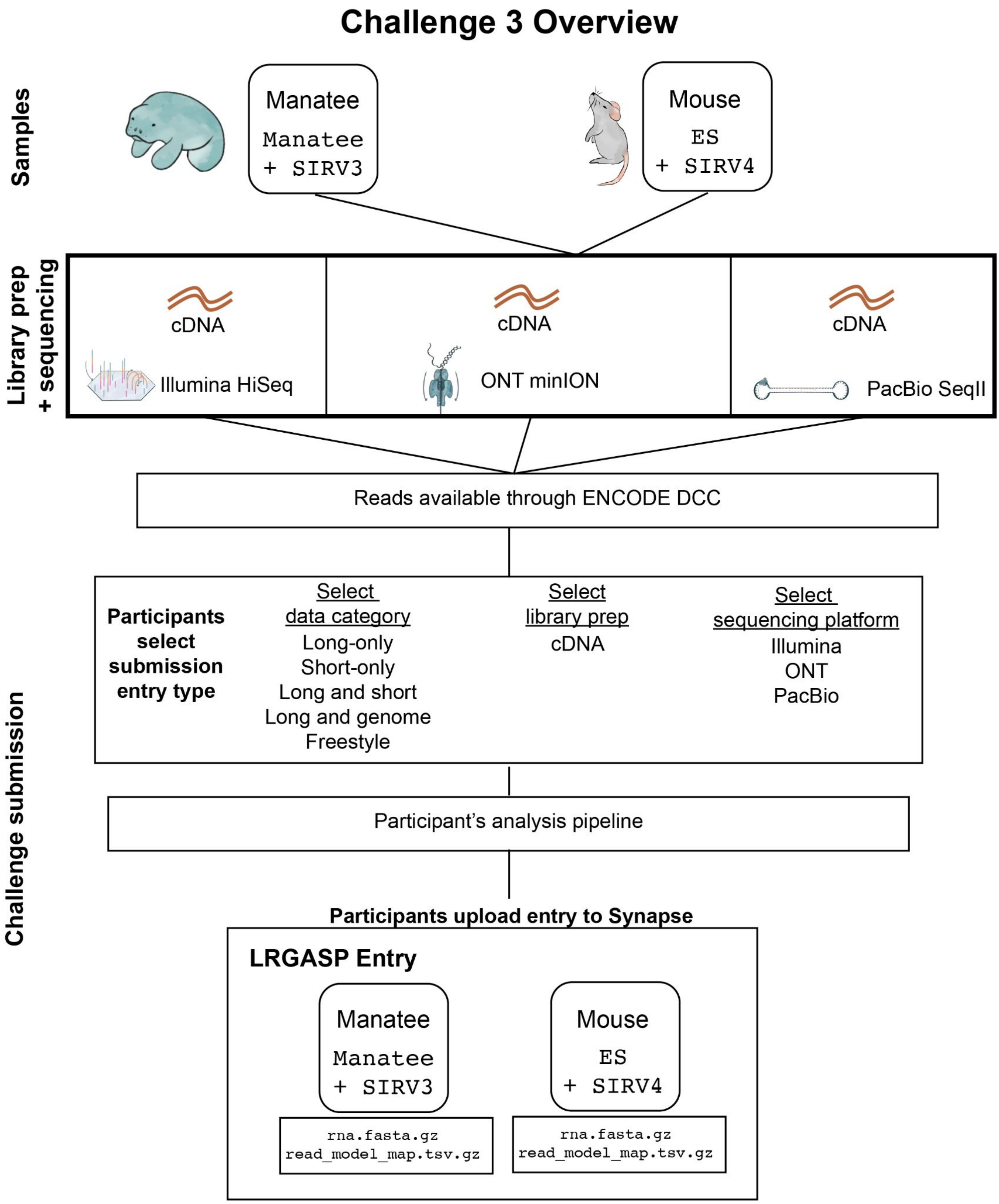
Supplementary Fig. 6. Schematic of directory structure and files that would be included in each entry



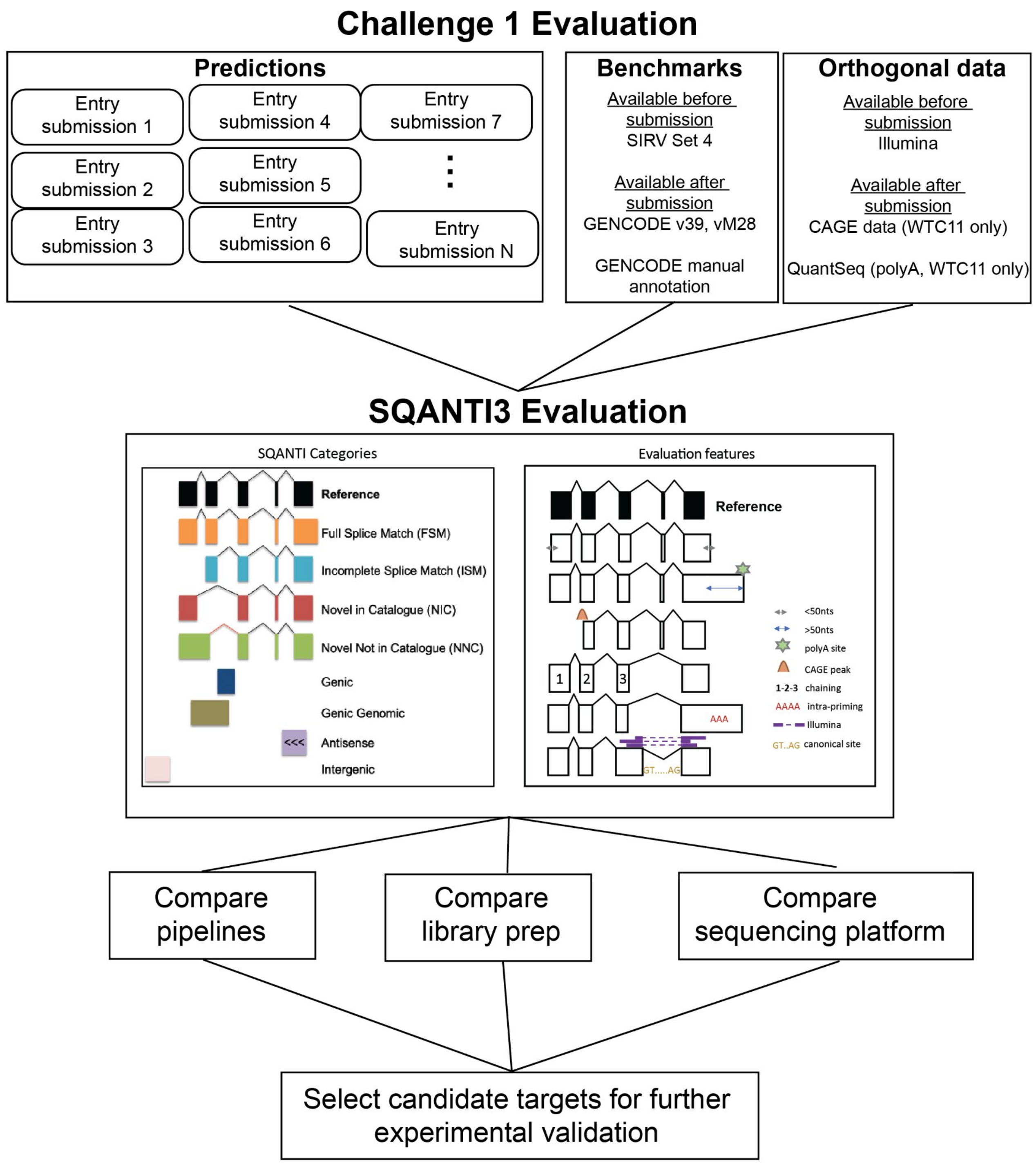
Supplementary Fig. 7: Flow diagram of Challenge 1: Transcript isoform detection with a high-quality genome. Samples, library prep methods, and sequencing platforms used in the challenge are indicated at the top. Participants select which data category, library prep, and sequencing platform to analyze, run their pipelines to generate transcript predictions, and submit an entry which includes predictions for all samples. The entries include a .gtf file of the transcript models and a .tsv file that assigns reads that supported each transcript model.



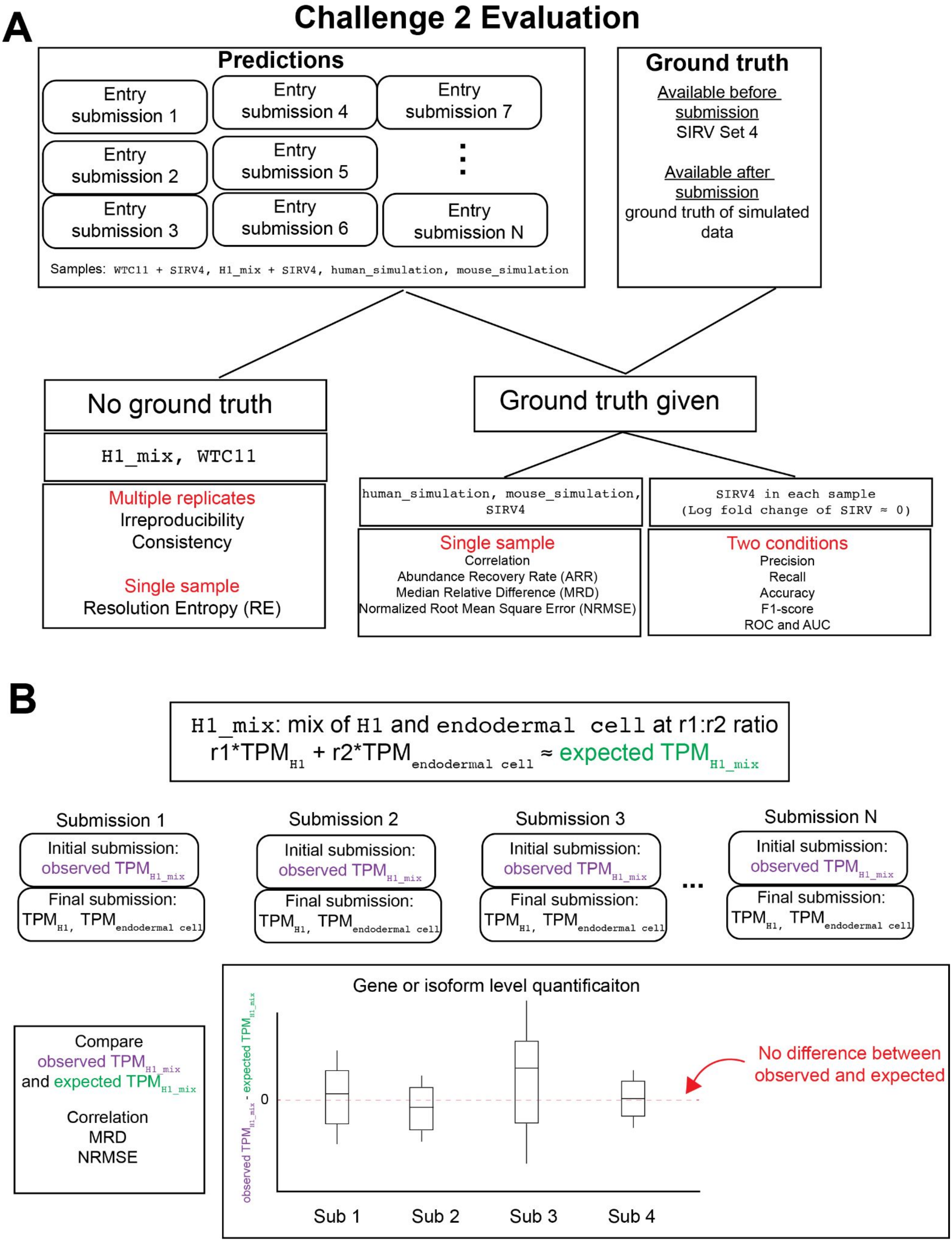
Supplementary Fig. 8. Flow diagram of Challenge 2: Transcript isoform quantification. Samples, library prep methods, and sequencing platforms used in the challenge are indicated at the top. Participants select which data category, library prep, and sequencing platform to analyze, run their pipelines to generate transcript predictions, and submit an entry which includes predictions for all samples. The entries include a .gtf file of the transcript models that are quantified and a .tsv file of the expression quantification. The H1 and endodermal cell samples were released after the initial submission deadline and participants were required to submit the quantification after the deadline.



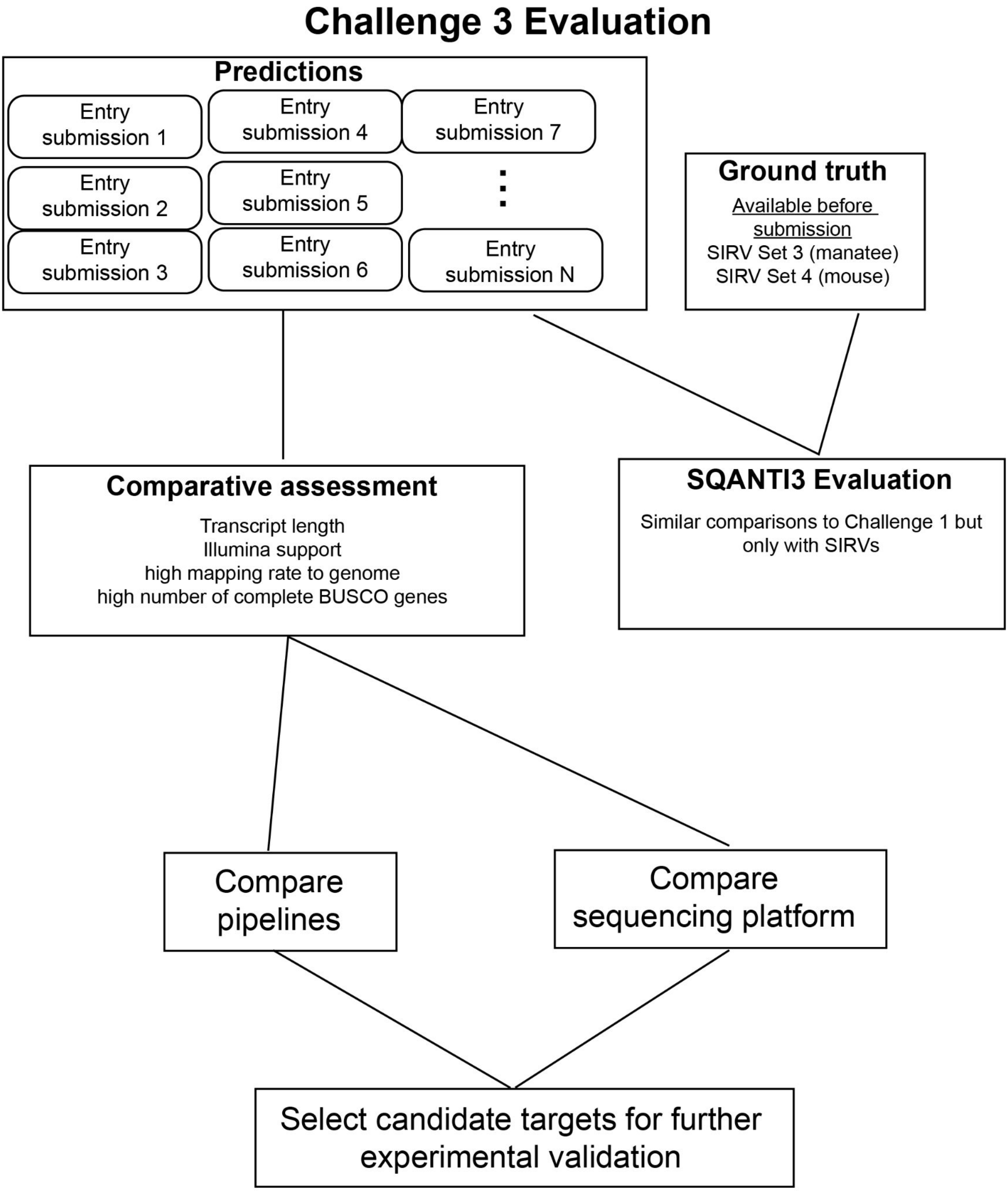
Supplementary Fig. 9. Flow diagram of Challenge 3. Samples, library prep methods, and sequencing platforms used in the challenge are indicated at the top. Participants select which data category and sequencing platform to analyze, run their pipelines to generate transcript predictions, and submit an entry which includes predictions for all samples. The entries include a .fasta file of the transcript models and a.tsv file that assigns reads that supported each transcript model.



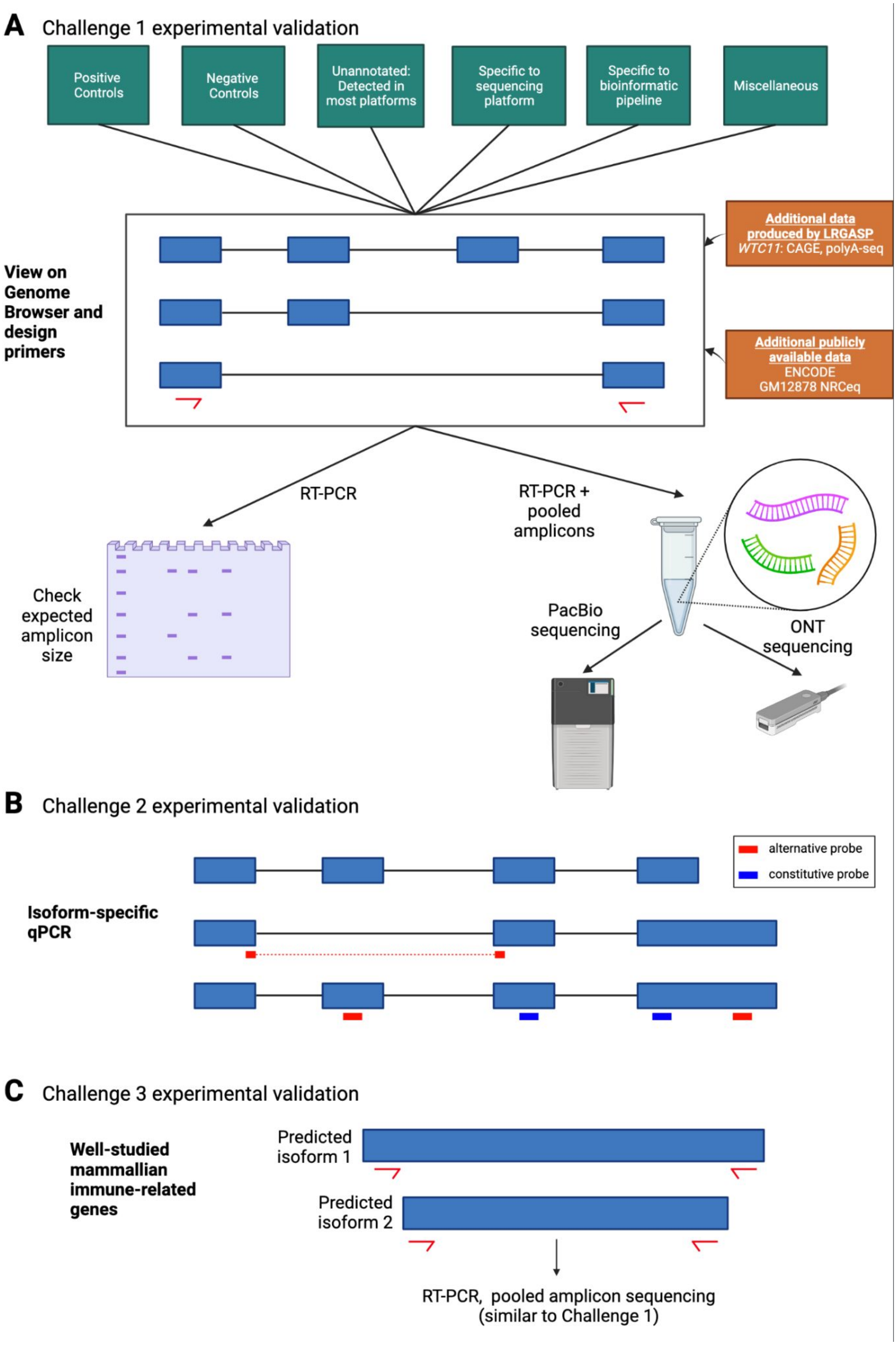
Supplementary Fig. 10. Flow diagram of the evaluation for Challenge 1. Benchmarks and additional orthogonal data that will be used for the evaluation are indicated. For example, CAGE and QuantSeq data from WTC11 cells were generated and made available only after participant submissions; therefore, they represent “hidden” data. These will be used to define 5’ transcript starts and 3’ ends.


A
Challenge 2 Evaluation
Predictions
Entry submission 1
Entry submission 2
Entry submission 3
Entry submission 4
Entry submission 5
Entry submission 6
Entry submission 7
⋮
Entry submission N
Samples: WTC11 + SIRV4, H1_mix + SIRV4, human_simulation, mouse_simulation
Ground truth
Available before submission
SIRV Set 4
Available after submission
ground truth of simulated data
No ground truth
H1_mix, WTC11
Multiple replicates
Irreproducibility
Consistency
Single sample
Resolution Entropy (RE)
Ground truth given
human_simulation, mouse_simulation, SIRV4
Single sample
Correlation
Abundance Recovery Rate (ARR)
Median Relative Difference (MRD)
Normalized Root Mean Square Error (NRMSE)
SIRV4 in each sample (Log fold change of SIRV ≈ 0)
Two conditions
Precision
Recall
Accuracy
F1-score
ROC and AUC
B
H1_mix: mix of H1 and endodermal cell at r1:r2 ratio
r1*TPM$_{H1}$ + r2*TPM$_{endodermal\ cell}$ ≈ expected TPM$_{H1\_mix}$
Submission 1
Submission 2
Submission 3
Submission N
Initial submission: observed TPM$_{H1\_mix}$
Final submission: TPM$_{H1}$, TPM$_{endodermal\ cell}$
...
Compare observed TPM$_{H1\_mix}$ and expected TPM$_{H1\_mix}$
Correlation
MRD
NRMSE
Gene or isoform level quantificaiton
observed TPM$_{H1\_mix}$ - expected TPM$_{H1\_mix}$
0
Sub 1
Sub 2
Sub 3
Sub 4
No difference between observed and expected

Supplementary Fig. 11. Flow diagram of the evaluation for Challenge 2. (A) Evaluation of Challenge 2 can be separated into metrics when a ground truth is known or a ground truth is unknown. (**B**) Example analyses to evaluate transcript expression using the cell mixing experiment. A sample, H1_mix, was initially provided for quantification which was a mix of H1 cells and endodermal cells at an undisclosed ratio. After the initial submission, the individual H1 and endodermal cell samples were released and participants submitted quantifications for each.



Supplementary Fig. 12. Flow diagram of the evaluation for Challenge 3. Only SIRVs are available for ground truth information. The evaluation will be based on a comparative assessment of the predictions followed by targeting specific candidates for further validation.


A
Challenge 1 experimental validation
Positive Controls
Negative Controls
Unannotated: Detected in most platforms
Specific to sequencing platform
Specific to bioinformatic pipeline
Miscellaneous
View on Genome Browser and design primers
Additional data produced by LRGASP
WTC11: CAGE, polyA-seq
Additional publicly available data
ENCODE
GM12878 NRCeq
RT-PCR
RT-PCR + pooled amplicons
Check expected amplicon size
PacBio sequencing
ONT sequencing
B
Challenge 2 experimental validation
alternative probe
constitutive probe
Isoform-specific qPCR
C
Challenge 3 experimental validation
Well-studied mammallian immune-related genes
Predicted isoform 1
Predicted isoform 2
RT-PCR, pooled amplicon sequencing (similar to Challenge 1)

Supplementary Fig. 13: Experimental validation approaches for the LRGASP challenges. (A) Multiple categories of types of transcript will be selected for validation (shown in green boxes). These loci will be viewed in the UCSC Genome Browser along with additional datasets to aid in the manual design of primers. Amplicons will be analyzed by fragment size and also pooled to perform long-read sequencing with PacBio and ONT **(B)** A select number of genes will be selected for transcript isoform-specific qPCR. A combination of probes detecting constitutive and alternative regions will be used. (C) RT-PCR validation will be performed similar to Challenge 1, except transcript will be selected from well-studied mammalian immune-related genes.