THE OHIO STATE UNIVERSITY

NIH workshop (NIEHS/NHGRI, May 24-26, 2022)

https://www.niehs.nih.gov/news/events/rnaworkshop2022/index.cfm

Capturing RNA Sequence and Transcript Diversity, from Technology Innovation to Clinical Application

Kin Fai Au Lab

May 24, 27 and 31, 2022

Department of Biomedical Informatics

Ohio State University

Group discussion on 5/24/2022 (Kin Fai Au)

Kin Fai Au (KFA)'s thoughts

- Learn boundary or limitation of different seq data
- Develop methods to solve problems utilizing strengths of the data
- Hybird: LR + SR, which part/section can/should be integrated to improve performance
- Suggest to Haoran to develop platform to allow
  - individual users without computational resource; we help them do analyses; they share data with us
  - data mine -> new findings
  - construct analysis pipeline, and visualization
  - m6A; the linkage information provided by long reads needs to be mined.
  - Dingjie Wang (DW): add all tools in the pipeline
    - KFA: not necessary at this moment, only focus on popular ones based on benckmark papers
- Understand complexity and diversity of the data in nature/real samples
  - KFA: e.g., look at the frequency of a given 10-mer sequence/motif in natural genome, will lower than theorical calculation
    - Yunhao Wang (YW): benefit the generation of gold standard to reduce cost
- **Co-occurance** between sequence, structure, mods, post-transcriptional events (alternative splicing and others): innovative statistical/other method development -> novel findings

Group discussion on 5/27/2022 (Aifu Li)

Aifu Li (AL)'s thoughts

- Capture multiple different RNA modifications on the same RNA molecule
  - KFA: of >150 RNA mods, it is important to know which ones can be detected by ONT, not limited by the well-known m6A; the linkage information provided by long reads needs to be mined.
  - YW: when talking about RNA mods, the RNA species should be considered, mRNAs and non-coding RNAs.
- Generate gold standard/reference for RNA modification detection
  - KFA: challenge to generate positive controls, cost a lot.
- Identify the RNA mods with the most important biological functions
  - Bo Li (BL): agree, but need more efforts from the whole RNA community
- RNA mods on RNA virus, RNA mods vs RNA localization
  - KFA: both of them are interesting, need methodology development; more importantly, co-occurrence (association) between RNA events/mods/others

Group discussion on 5/27/2022 (Haoran Li)

Haoran Li (HL)'s thoughts (see Haoran's slide for more details, Page 9-10)

- Difficulty in computing and sharing resource
- Analysis reproducibility and standards
  - KFA: using public platform, such as GitHub, to customize analysis pipeline
- Appyter
  - KFA: high computing cost for data processing
- Gene ananotation
- Single molecules vs averaged ones

Group discussion on 5/31/2022 (Ying Zhang)

Ying Zhang (YZ)'s thoughts

- RNA mod detection solution: convert mod to mutation profile by RT (reverse transcription) enzyme
  - KFA: convert/detect multiple mods to specific mutations, on the same molecules
  - KFA: suggest to HL, computationally, de novo detection of new modification by new signal in raw data
  - YZ: new RT enzyme development
  - YW: efficiency/sensitivity, specificity (one mod to one mutation pattern, one to multiple) sing public platform, such as GitHub, to customize analysis pipeline
  - BL: like bisulfite seq for DNA
  - Hsin-Lun Hsieh: convert one mod -> another mod -> enhance detection signal
    - KFA: the same as we are doing for DNA 5hmC with Jerry's lab now
- RNA secondary structure by mod footprint
  - KFA: alignment <10% for direct RNA-seq on heavily-modified RNA molecules; mutation solution improves alignment, because more accurate base-calling; new aligner to handle mutated sequences
  - YW: need new base-calling methods plus gold standards (positive control)

Group discussion on 5/31/2022 (Bo Li)

Bo Li (BL)'s thoughts (see Bo's slide for more details, Page 11-15)

- Function difference of isoforms with small sequence difference but with different folding pattern
  - YW: structure determines function
- Standards for RNA mod detection in different cell/tissue types
  - YW: standards not just for cell types; more imperatively, standards for method development
- Find good application target, biologically significant

Group discussion on 5/31/2022 (Dingjie Wang)

Dingjie Wang (DW)'s thoughts (see Dingjie's slide for more details, Page 16-22)

- Isoform-specific function
- Quantitative analysis
- Bias in library preparation and sequencing platforms/strategies to consider for method development
  - Xiaoyu Cai: agree, need understand which bias has big effect
- Collection: tools and datasets for direct RNA-seq
- RNA input requirement, from 100 ng to 1 pg
- AI/ML for direct RNA-seq method development
- User-friendly platform and gold standard

Group discussion on 5/31/2022 (Yunhao Wang)

- Low-input RNA-seq
  - high-efficient RNA ligases to allow efficient adapter ligation
  - sequencer capture efficiency

- Standards and methods for RNA modification detection
  - synthesis, ligation ...
  - antibody, NMR, MS, direct RNA-seq

- RNA structure
  - experimental approach
  - computation: far from the knowledge of protein structure (AlphaFold)

- Interactions
  - RNA sequence, post-transcriptional events (AS, APA), modifications, structure, RBP

Capturing RNA Sequence and Transcript Diversity - From Technology Innovation to Clinical Application

Haoran Li

Department of Biomedical Informatics
Ohio State University

5/27/2022

Critical Computational Resources

- Analysis reproducibility (raised by Ali Mortazavi & Phil Bevilacqua)
  - Difficulty in sharing & comparing results for ~100 datasets
  - Raising standard for paper review
  - Central repo that documents software code used for analysis
  - Docker containers to reproduce environment
- Communication between data scientists & other researchers (raised by Ali Mortazavi & Phil Bevilacqua)
  - Jupyter notebooks to automate dataflow (https://appyters.maayanlab.cloud/#/)
- High cost for processing data (even on Cloud)
- The gene annotation to capture what they are expressed
  - How to represent results given use of different cell types;
  - Gene annotations should capture what they are expressed understand where RNA mods are occurring
- Coding culture promotion
- Difficult to get access to RNAseq data via dbGAP (Phil Bevilacqua)
  - prohibits global analysis and more study focused

Summary on RNome workshop

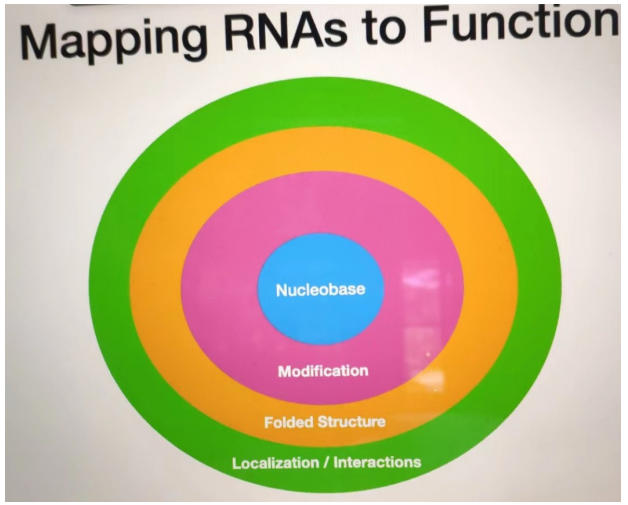
Bo Li

05/31/22

1. What is RNome?

- Sequence – Modification – Structure – Localization/Interaction - Functions

Isoforms with slight difference in sequences may have distinct functions due to the folded structure.

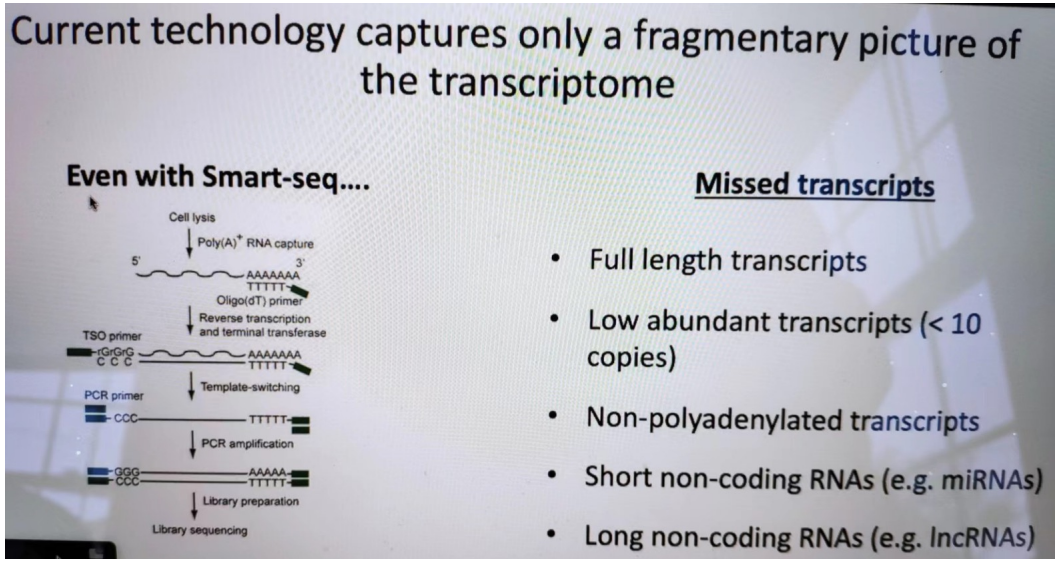


Yunsun Nam, UT Southwestern

1. What is RNome?

- Different types of RNA and diverse functions

Desire technology to capture the full picture.



Anna Pyle, Yale

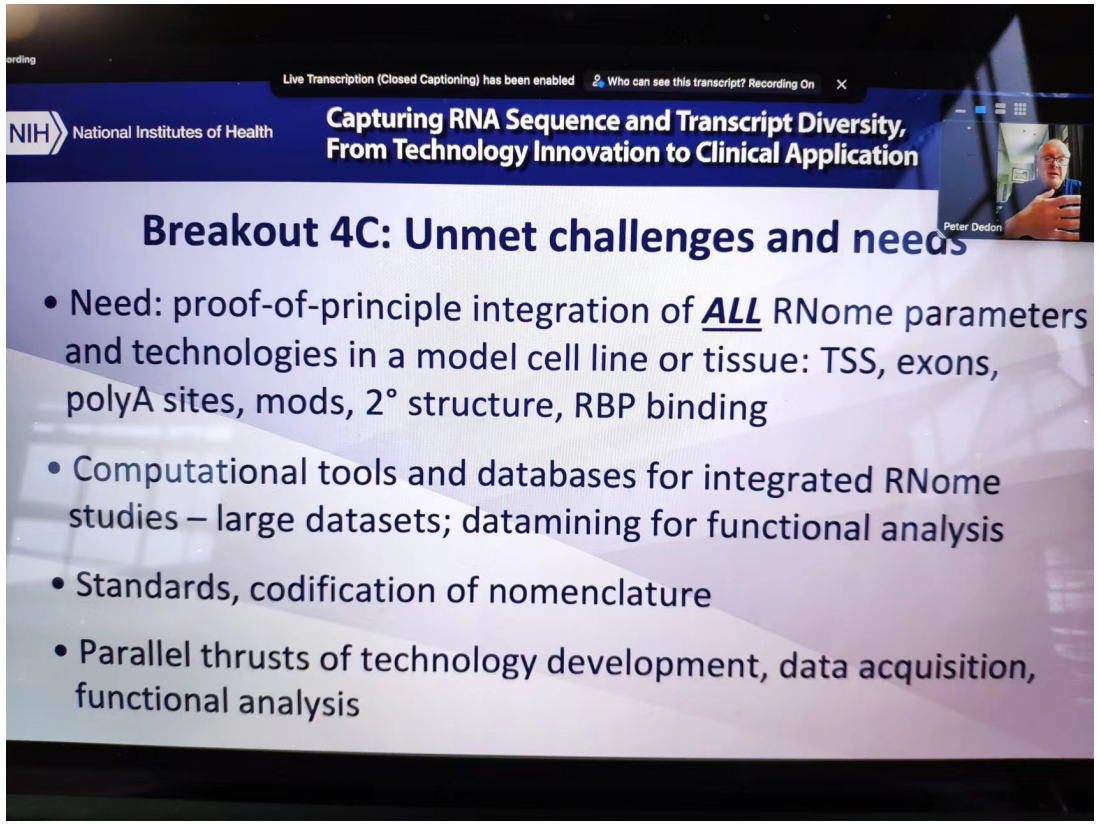
2. Opportunities and challenges in RNome

Transcript standard in certain cell line or tissue

Modification standard

Tech: Nanopore direct RNA sequencing

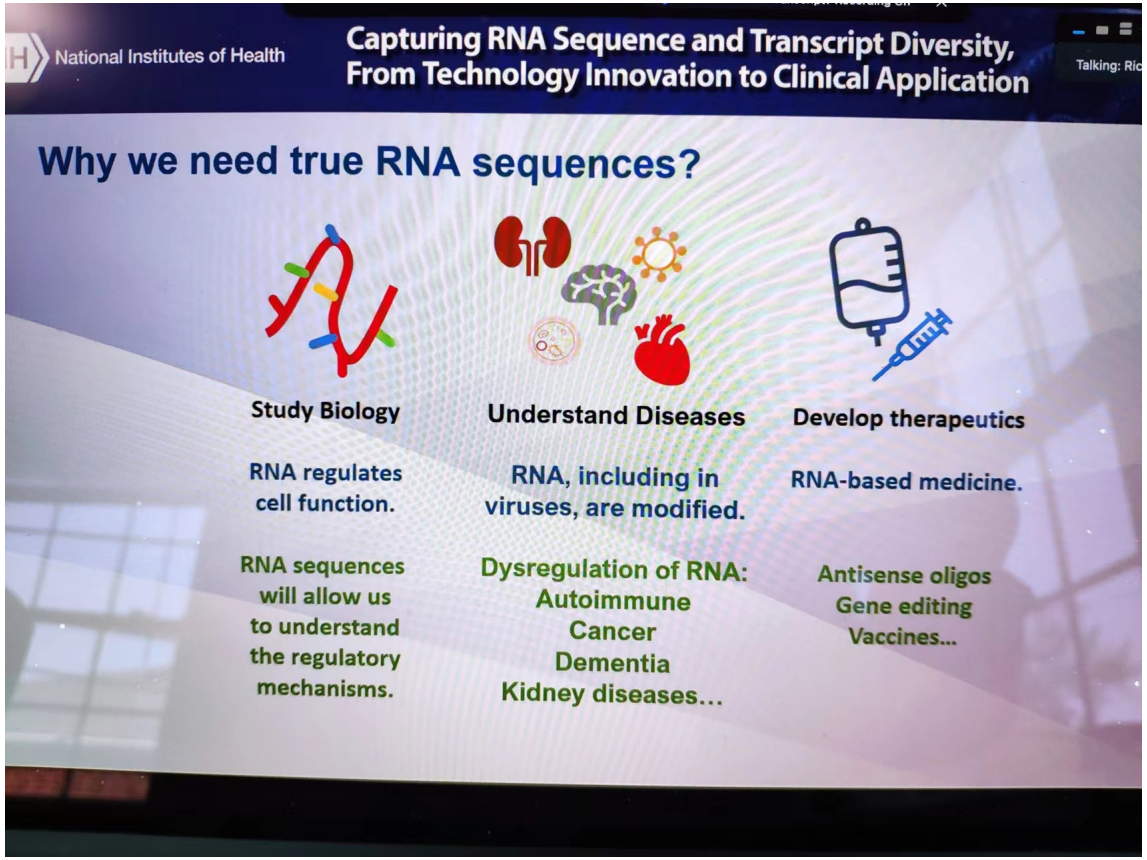
RNA mods



3. Application and translation

Understanding biological functions

Human health



Capturing RNA Sequence and Transcript Diversity - From Technology Innovation to Clinical Application

Dingjie Wang, PhD

Department of Biomedical Informatics
The Ohio State University

05/31/2022

➢ Transcript diversity and specific function prediction

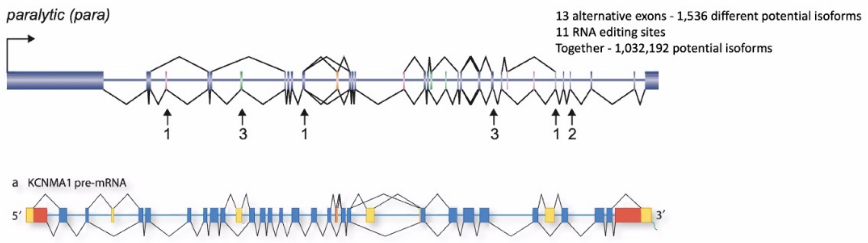
◆ **Anna Marie Pyle**, Yale university

Challenges from the biology side

- Most RNAs involved in active gene expression are **very long** (>1kb)
- RNA is **fragile** and often low abundance
- Functionally **relevant information** within RNA is **not captured** by its encoded sequence

Functionally relevant information within RNA is not captured by its encoded sequence

2. Expressed mRNAs and lncRNAs are alternatively spliced into complex isoforms, each with a different sequence.
a. Isoforms differ by cell and tissue-type
b. Isoforms depend on stage of development, infection, stress or environment



Nilsen & Graveley (2010) *Nature*, **463**:457-63

Adapted from Brent Gravely

◆ **Samie Jaffrey**, Weill Cornell Medicine

Once we have a good sense of transcript diversity, another major question is which ones are biologically important, and which ones are just stochastic missplicing errors that have no biological consequence.

◆ **Kin Fai Au, OSU**

And the next question is which isoforms have isoform-specific functions

◆ **Angela Brooks, UCSC**

Agreed with Kin Fai! We have very little knowledge of isoforms-specific functions

➢ Confound analysis of expression and splicing from RNA-seq data

◆ Chris Burge, MIT

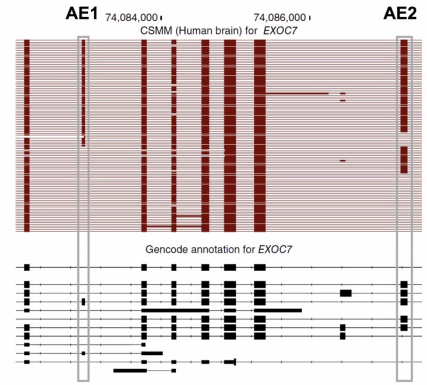
Outline

- **Issues that confound analysis of expression and splicing from RNA-seq data**
  - **short read: GC bias and "shrinkage"**
  - **long read/mod data: association between features in a transcript**
  - **single-cell: spurious binary splicing**
  - **newer technologies: all of the above**
- **Issues that confound interpretation and translation:**
  - **classification and nomenclature of exons/isoforms**

Two alternative exons in *EXOC7* human brain SLR-RNA-seq

- **AE1 and AE2 appear positively associated, i.e. transcripts with AE1 mostly contain AE2 ($P < 2x10^{-19}$)**

Tilgner et al. Nature Biotech. 2015



Mixtures of cell types cause problems

What if EXOC7 is expressed mainly in two cell types:

cell type A: AE1 and AE2 are **lowly** and **independently** included (mostly 0,0 mRNAs)
cell type B: AE1 and AE2 are **highly** and **independently** included (mostly 1,1 mRNAs)

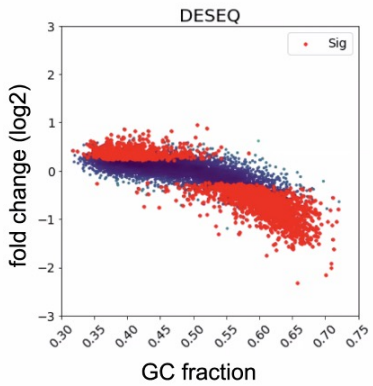
This type of heterogeneity in the data can yield **false positive associations** (related to Simpson's paradox)

With some algebra, can show that analysis of a **mixture** of two cell types with independent splicing of two exons but different inclusion levels tends to produce **spurious positive associations** (but not spurious negative associations)

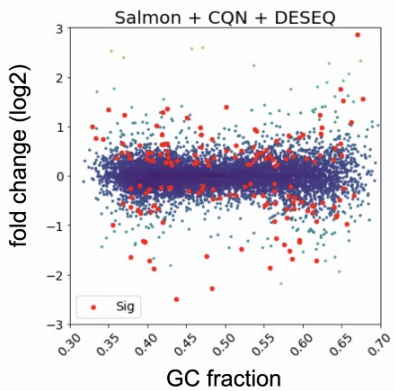
In practice, the constituent cell types in a sample and their splicing are often not fully known

Detecting and Correcting GC-bias in RNA-seq Data

Very strong relationship between gene %GC and log fold change suggests presence of bias



Running data through Salmon + CQN corrects for GC bias and gives more plausible set of significant genes



GC bias, likely related to PCR steps in library prep, is relatively common in RNA-seq datasets and can yield spurious conclusions if not corrected

Michael McGurk

See also Van Nostrand et al. Nature 2020

Conclusions, Part 1

- **All sequencing technologies – even established ones – involve biases in library prep or sequencing**, and these biases may differ between samples or batches

Direct RNA sequencing mostly avoids library prep, but there will still be biases in the capture of different RNA molecules, and biases in the sequencing itself

With any new technology, it is essential to learn the nature of these biases and develop robust statistical procedures to correct them to minimize their impacts on conclusions

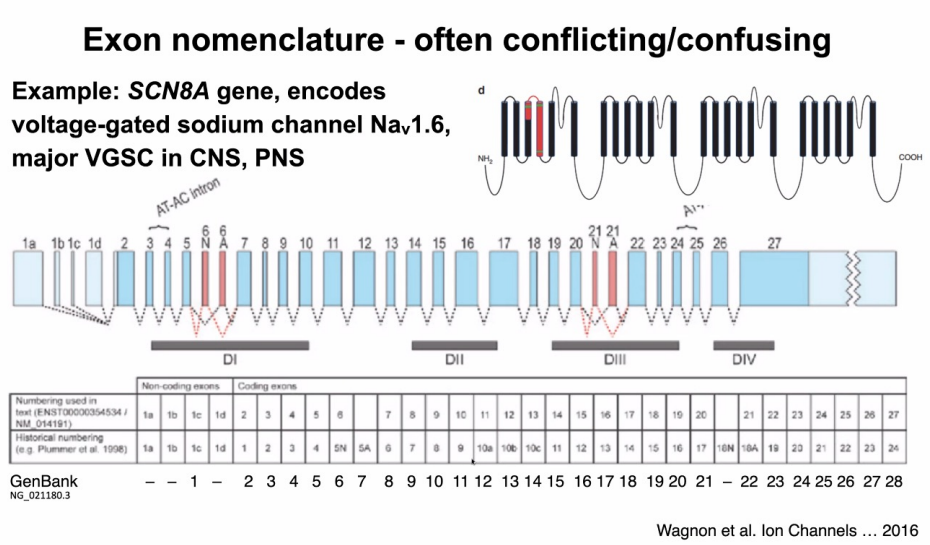
- All **analyses of tissues/cell mixtures are inherently biased** for detection of positive associations between mRNA features, unless the cell type composition is accounted for

If one cares about the source of the association, need to:
- work with pure cell populations
- separate cell types as in ScISOr-seq, or
- develop new statistical approaches for dealing with unknown mixtures

◆**What features affect quantitative analysis?**
- Technology biases
- Gene features
- Single molecule and single cell

➢ Confound analysis of classification and nomenclature of exons/isoforms

◆ Chris Burge, MIT

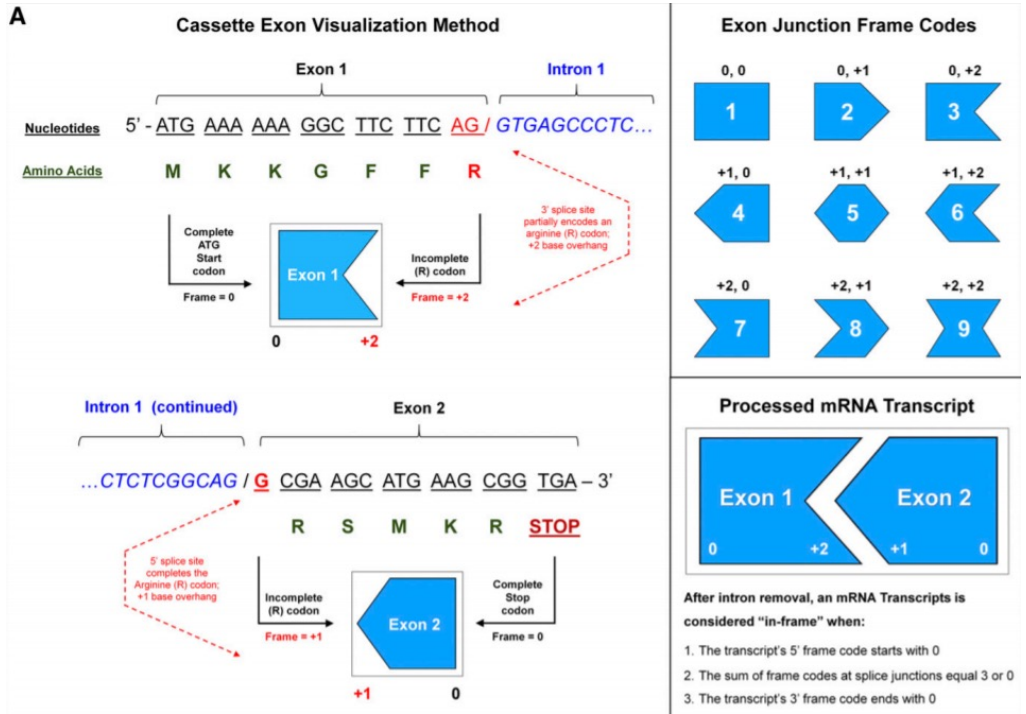


◆ Kin Fai Au, OSU

I would suggest to re-define "exon/intron" by defining "unsplicable elements: that is, using all splice sites to split the gene body into smaller bins (than the conventional exons). Each such bin are the basic unit to define gene isoform structure. For example, the alternative 5' splicing event can be defined as the alternative combination of two such bins. This generalized definition may help to standardize the nomenclature

◆ Jo Yeakley, BioSpyder Technologies

Andrew Annalora (Oregon State University) has a variant graphical representation of exons (Drug Metabolism and Disposition April 2020, 48 (4) 272-287; DOI: https://doi.org/10.1124/dmd.119.089102) where instead of rectangles for exons, he bumps out the 5' or 3' edge to indicate coding frame, so it's easy to see where a cassette exon maintains coding frame of the rest of the transcript. This seems like an accessible way to find the low hanging fruit for splice variants that are likely to be biologically important.



Conclusions, Part 2

Controlling for mRNA capture efficiency, sequencing depth/complexity is essential in analyses of splicing, particularly from single cells

The same genomic regions are often processed as first or internal exons in different isoforms

Development of a consistent nomenclature for exons, isoforms, and modifications would facilitate research and impact clinical diagnosis/treatment

➢ What tools or datasets are necessary for analysis of direct RNA sequencing data?

Breakout 5A: Infrastructure and Bioinformatics

Q1: What tools or datasets are necessary for analysis of direct RNA sequencing data?

- Angela involved in large community effort in tool devel for long-reads. Surprised on how diff tools report isoforms. A lot of diversity in reporting. Benchmarks - some more accurate than another.
- DIrect RNA Seq, Detect RNA secondary structure,
  - do not have stand pipeline/software to analyze raw data, lack of standard, computational storage needs are vast
  - What kind of information actually needs to be saved? What needs to be extracted from raw data? Utilize viral RNA seq data
  - We dont know all RNA mod or how they will affect current studies, but we need to keep raw data so that we can go back
  - More efficient tools to process the data,
- More Organized way to access fast 5 files/ data that is published, creating a tool that does this
- IS there a way to keep/mask the mods that are known bases/clearly identified, just look for the "funny" ones, that way to ficuse on developing software on ide
  - Challenge with a high error rate with direct RNA
  - Public data can be used to help decrease error rate
  - Repository of nanopore reads,
  - What kind of data set is necessary? Dont need all the combinations,
- s are also needed for MassSpec-based direct sequencing of RNA and modifications
-
-

Motivations for Direct RNA Nanopore Sequencing

- Analyze isoforms directly (no assembly)
- Poly-A tail length assessment
- Accurate quantification in PCR-free system
- Accurate analysis of gene fusion events
- Analyze different RNA types (e.g. rRNA, tRNA, etc.)
- RNA modifications – a new frontier
- No reliance on reverse transcriptase errors (~1e-4)

Slide: Courtesy of Miten Jain

Challenges

RNA modifications

- Understanding raw nanopore traces and relationship to modifications
- Need gold standards!
- Need to preserve RNA molecules and develop ways to increase coverage of full-length RNAs
- Need to reduce input requirements (100 ng —> 1 pg)
- Need new technologies (one technology can't do it all!)

➢ AI/ML machine learning approaches in dRNA sequencing

Q2: How can AI/ML machine learning approaches be used to support direct RNA sequencing? (summary)

- What is the incentive for doing more direct RNA seq?
  - One of the most critical problem to solve
- We need better or established ground truth sets
- Generate large data sets are needed
- We can also make use of existing short-read or existing dRNA data together for AI/ML approaches
- Challenges sharing data
  - Standard format which is compressible

➢User-friendly platform and gold standard

Technology Development

Technology integration - baseline information everything that you can measure about RNA (single cell, spatial, subcellular location and time for cell lines as well as for single cells within tissues) that can be used to understand disease and altered health states

A database (searchable) is needed to store and maintain information. Standardization and communication of the nomenclature for mods. N

In-depth understanding of a model RNA including functional consequences

Need proof-of-principle that's applicable to many people (viruses) also a representative mRNA (e.g., beta actin)

Breakout 4A: Technologies on the horizon

- Need for standards
  - Way to send to labs to validate
  - ABRF project
  - See how labs detect
- Funding for technology consortium, just to get fundamental platforms in place
- How can you map modifications in complex mixtures, purification methods that make more amenable to MS/sequencing
  - Huge need for purification methodology to separate into pools of lower complexity
  - Current methods need work
  - Limit: deconvoluting complexity of sample
- Use reference standards to classify?
  - No NIST equivalent
  - 2 that EPA are using
- Most lacking in current generation of cross-linkers?
  - Ways to push forward?
  - 4SU most common
  - Cross link to protein then separate in MS once collided with gas? MS gas-based cleavable cross-linker for protein-protein, need for nucleic acid-protein
- Bioinformatics tools, for RNA seq and analysis
  - Platform not as common, not many user-friendly tools for analysis of MS data
- Separation of RNA, complicated system to study, bioinformatics tools could be useful here
- Mass spec isn't very user-friendly, drawback of tool
- Big investment in MS and tools needed, but other tech platforms should also be explored
  - Tip enhanced