Transcriptome Annotation Construction Software (TACoS): leveraging long and short read attributes for a fully source aware transcriptome annotation

Richard Kuo
Motivation

- Functional Annotation of Animal Genomes (FAANG) Consortium

- Public annotations (Ensembl, NCBI, etc.)
  - Mostly just model organisms
  - Require funding to annotate
  - Take time to annotate
  - Annotation decisions are not as transparent and not customizable
  - Slower to adapt to new technologies
  - Require public data

- RNAseq annotations
  - For non-model organisms
  - Highly dependent on analysis pipelines used
  - Likely to be of lower quality
Iso-Seq Motivation

- Collapsing redundant transcripts
- Merging annotations
- Preparing for use by FAANG
Iso-Seq Collapsing

TSSC – Transcription Start Site Collapse
ECC – Exon Cascade Collapse

But what about 5’ cap selected libraries?
• **Threshold for:**
  - Transcript start
  - Transcript end
  - Exon start
  - Exon End

• **Grouping:**
  - How much wobble walking should be allowed?

• **Current implementation:**
  - 10 bp difference allowed
  - Unlimited wobble walking
Wobble: Transcript Start/End

Transcript Start Wobble

Transcript End Wobble
Wobble: Exon Start/End

Exon Start Wobble

Exon End Wobble

Number of Exons
0 50000 100000 150000

Wobble
0 3 6 9 12 16 20 24 28 32 36 40 44 48 52

Wobble
0 6 13 21 29 37 45 53 61 69 77 85 93 102 112 122 132
TACoS Collapsing

- Control over transcript collapsing
  - Define threshold for transcription start/end and exon boundaries

- Manages 5’ cap selected and non cap selected sequencing data
  - Allows for Iso-Seq Tofu Collapse style of collapsing

- Provides source information for all predicted events
  - Support for each final model
  - Support for each transcript feature (TSS/TTS, splice junctions)

- Flags uncertainties
  - Poly A truncation
  - Genomic issues (assembly gap)*
  - RT Switch*

*future feature
Additional Features

• Variant Calling
  – SNP and Indels
  – Can be used for phasing

• Wide vs narrow transcription start/end sites

• Wobbly splice junctions

• Can be used from FLNC stage or post-ICE stage
  – Adjust coverage and identity settings

FLNC – Full length non-chimeric reads
ICE – Iterative Clustering for Error Correction

https://tacobueno-live-iglusjax.stackpathdns.com/assets/food/tacos/Taco_Crispy_Beef_990x725.jpg
TACoS Annotation Merging

• Manages short read data merge with Iso-Seq
  – Uses transcript start and end sites from Iso-Seq, verifies splice junctions with RNAseq
  – CAGEseq for transcription start sites*

• Merge public annotation with Iso-Seq*
  – Attach public ID’s and flag differences

• Provides source information for all predicted events

• Flags uncertainties
  – Genomic issues*
  – Multi-mapping issues*

• Collapse models for RNAseq count based software*
  – Kallisto, Salmon etc.

*future feature
Similar Tools

• EVM
  – Weighted inputs
  – Does not weight features

• Augustus
  – Does not yet incorporate Iso-Seq

• IDP
  – Requires Iso-Seq and RNAseq from same sample
  – Incorporates validation pre-assembly

• Anyone know of any other tools?
TACoS Summary

- Primarily based on Iso-Seq data
- Manages Iso-Seq from after error correction to transcriptome annotation
- Integrates other data
  - Sequencing Data: RNAseq, CAGEseq, CHIPseq*
  - Database Data: Public annotations*, protein databases*
- Biotype classification*
ICE-Cycle

- Tool for splitting FLNC sequences by size for running ICE
- Define overlapping windows
- Collect split groupings
- Re-run ICE with split groupings
- Merge all ICE runs
- Not pretty...

github.com/rkizen/ice_cycle

FLNC – Full length non-chimeric reads
ICE – Iterative Clustering for Error Correction
Acknowledgement

Professor Dave Burt
Professor Alan Archibald
Katarzyna Miedzinska
Bob Paton
Lel Eory

Steve Picton
Elizabeth Tseng

Karim Gharbi
Marian Thomson