Updates on RNA-seq data analysis pipeline for plants

Lin Wang

Research Associate

Brutnell Lab, BTI, Cornell

iPlant Collaborative — Empowering a new plant biology

BTI Boyce Thompson Institute for Plant Research

Cornell University Founded A.D. 1865

Presentation Outline

- Overview of the current state of RNA-seq data analysis pipelines
- Overview of our custom pipeline and the comparison to the most popular “Tuxedo” package
- Future development of our pipeline

The current status of RNA-seq data analysis pipeline

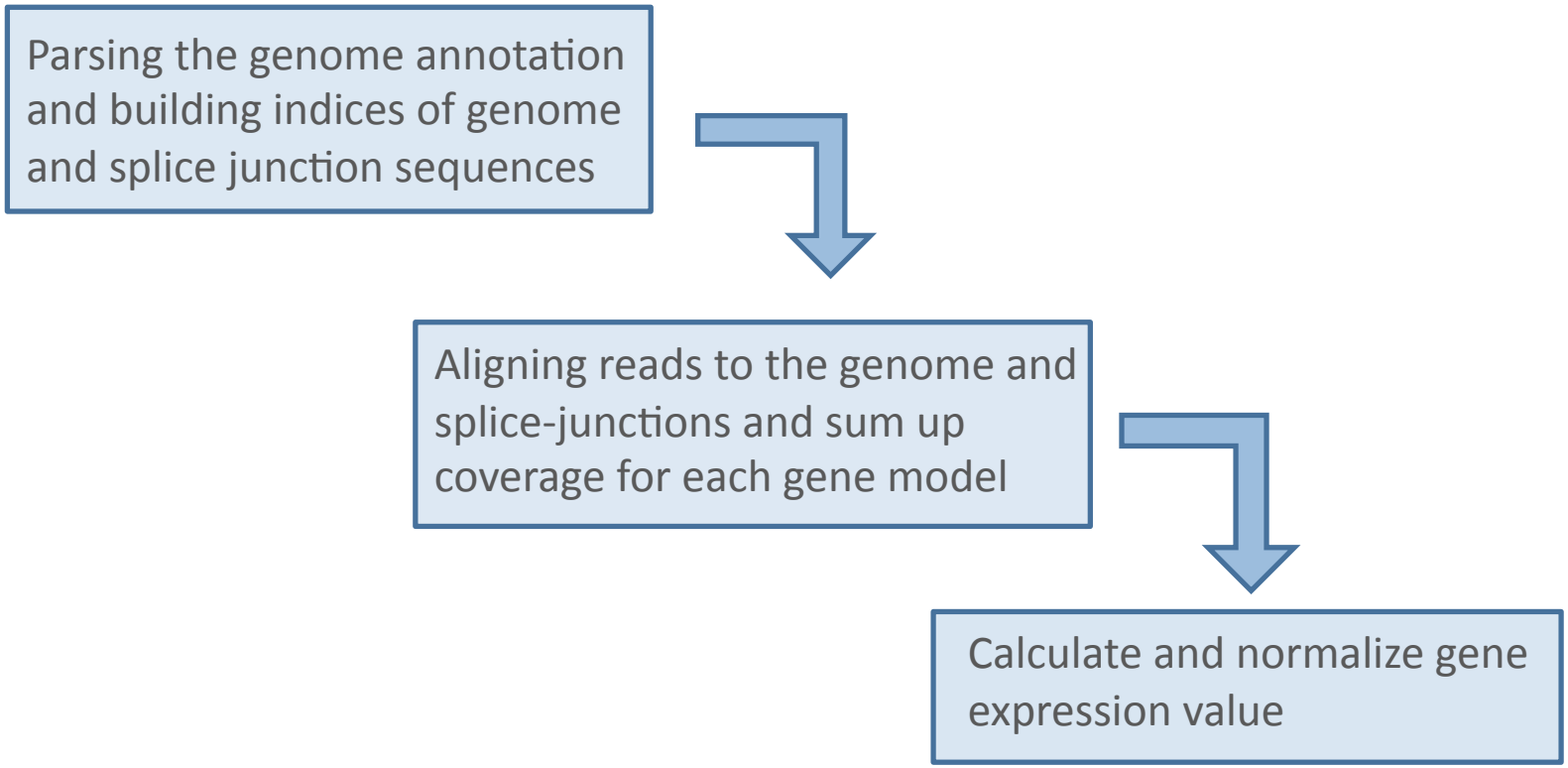
- No commonly accepted standard
  - Depending on the need of analysis, no single stand-alone package can do it all
  - Method for normalization is still a point of debates
  - The RNA-seq technology itself is constantly evolving
- The “tuxedo” package is likely the most widely used and acknowledged
  - Bowtie, Tophat, Cufflinks (Cufflinks published in Nature method)
  - Still buggy, especially Cufflinks with its latest version (more on this later)

Overview of our custom RNA-seq pipeline

- Coding language: Perl and R
- Aligner used: BWA (though, though any aligner can be swapped in with some modification to the code)
- Splice-junction alignment is done separately by building a splice-junction database (similar to Tophat)
- Evenly weighted distribution of “multi-reads”

Overview of our custom RNA-seq pipeline

Our pipeline can be modularized in a three-step process:



The disadvantages of our pipeline vs. the "Tuxedo" package

- Speed (most "tuxedo" packages are compiled with C)
- Acceptance (All "Tuxedo" packages are published)
- More functionalities (Novel splice junction and transcription unit detection)
- Optional bias correction (caused by random priming and reverse-transcriptase)

So…why are we not switching?

The advantages of our pipeline vs. the "Tuxedo" package

- Complete control over the pipeline (we know best how our libraries are constructed and we know the biology – or plant biology if that matters)

- Better correlation with qRT-PCR (tested with 41 genes)

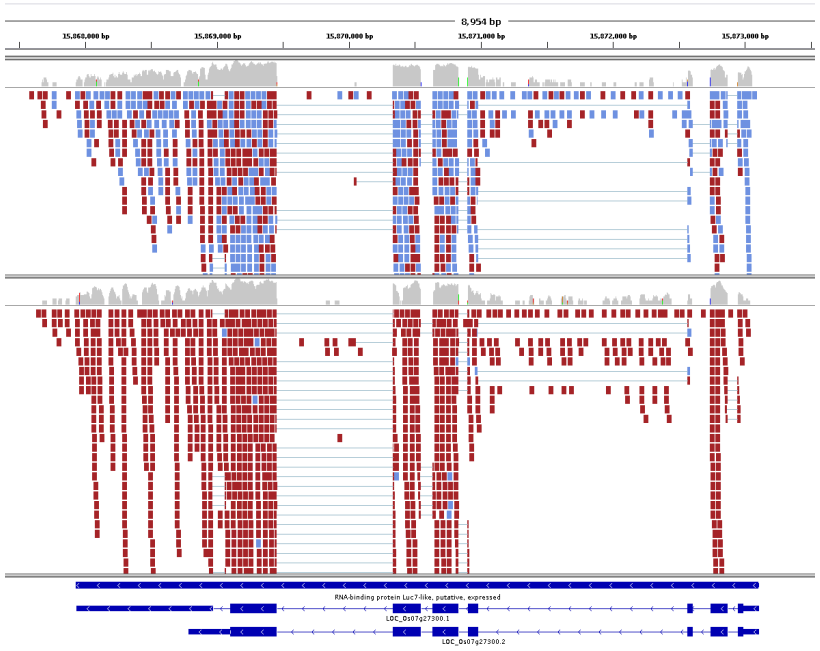
	base sec	-1 sec	+4 sec	tip sec
Cufflinks	0.978804	0.820685	0.77274	0.789175
Our pipeline	0.978954	0.858298	0.850948	0.855966

Simple Pearson correlation

The advantages of our pipeline vs. the "Tuxedo" package

- Cufflinks (v. 1.0.1) is very problematic when used to calculate RPKM of strand-specific data

Non-Strand-specific (NSS)



Strand-specific (SS)

RPKM calculated with Cufflinks:

NSS: 132.79

SS: 32.38

RPKM calculated with our pipeline:

NSS: 137.60

SS: 125.82

The advantages of our pipeline vs. the "Tuxedo" package

- Easily adjustable for different ways of normalization -How gene expression value are calculated/estimated

  – Most commonly used method: Normalize to total reads mapped (RPKM/FPKM: Reads/fragments Per Kilobase of exon model per Million mapped reads)

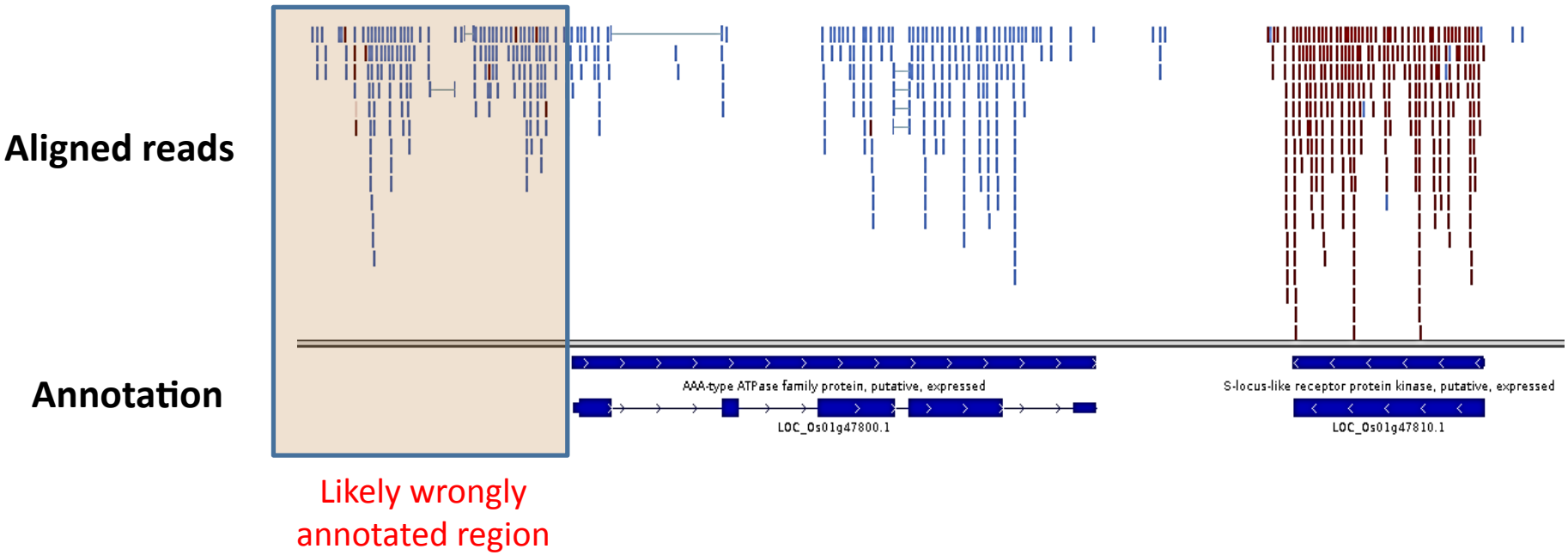
  – Alternative normalization methods:
    - Normalizing to the 3rd quintile of total reads have been shown to outperform normalizing to total reads (This is particularly true for plant transcriptomics analysis)
    - Normalizing to spiking RNA (more close related to our protocol)
    - Trimmed Means of M (TMM) *(Robinson and Oshlack, Genome Biology 2010, 11:R25)*

We are continuing to refine our pipeline – works in progress

- **Adding on the missing functionalities**
  - Counting reads from all possible splice junctions
  - Bias correction (integrate the published R package)
  - Incorporate a gene-centric view of calculating gene expression
  - Filtering step for raw reads off the sequencing machine
  - Incorporate index splitting and trimming (Fastx toolsets)
  - Filtering low complexity reads (defined as a single nucleotide consist over 95% of total read length)

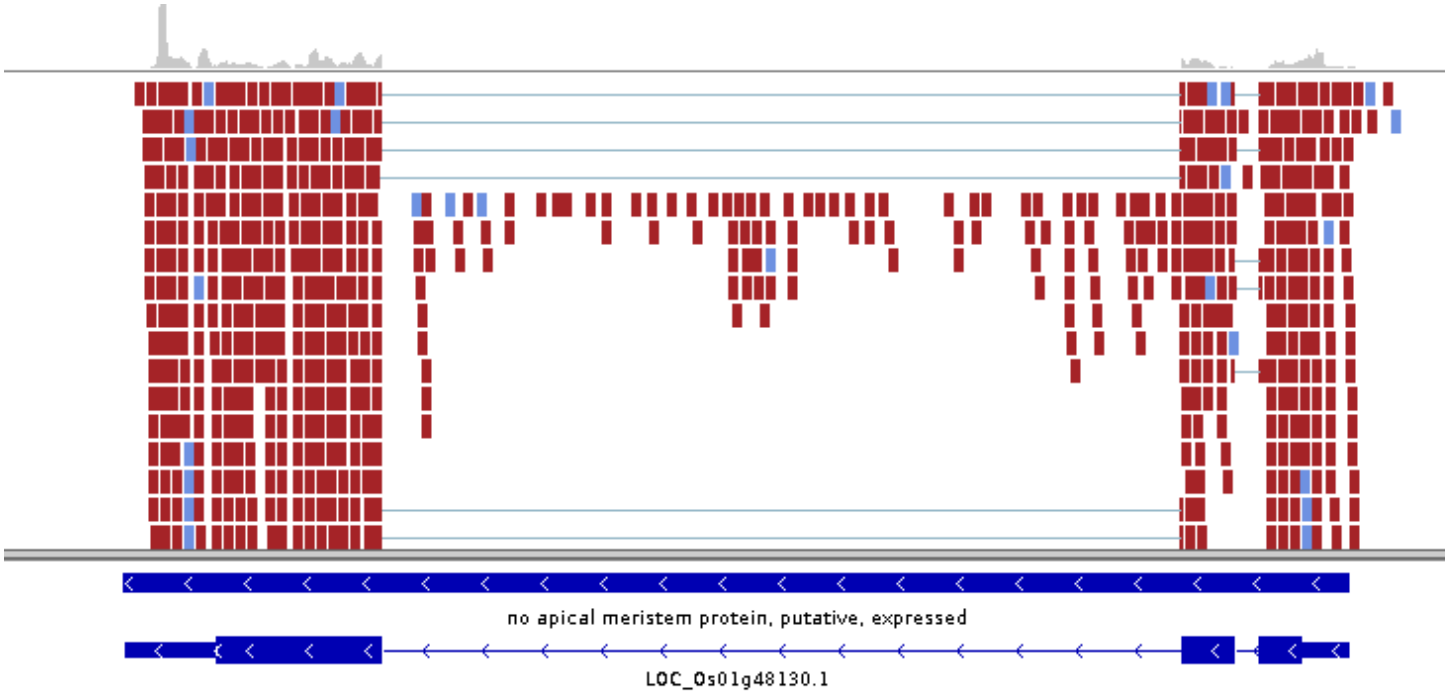
We are continuing to refine our pipeline – works in progress

- Add an annotation-refinement module in front of everything



We are continuing to refine our pipeline – works in progress

- Add an intron-alignment/alternative-splicing calling module (in collaboration with statistician Peng Liu from Iowa State University)



Last but not least…

- We are finally prepared to publish our wet-lab RNA-seq method together with the analysis pipeline
  - Strand-specific RNA-seq library construction method that is 10x cheaper than current cheapest Illumina kit
  - Multiplexing adaptors that are compatible for HiSeq2000
  - Spiking RNA as a tool for alternative normalization
  - Our custom RNA-seq analysis pipeline

Acknowledgements

- Cornell, computational biology service unit
  - Qi Sun
  - Lalit Ponnala
- Department of Statistics, University of Iowa
  - Peng Liu
  - Yaqing Si
- iPlant collaborative