Latest Developments in the Trinity Cancer Transcriptome Analysis Toolkit

PI: Aviv Regev
Co-Investigator: Brian Haas

ITCR meeting, Utah. May 2019
A Plethora of Biological Sequence Analyses Enabled by Transcriptomics

Figure 2: Transcriptome profiling for genetic causes and functional phenotypic readouts.
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

(typically ~75 to 150 base length reads)

Two paradigms for transcriptome analysis
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

Genome
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

Transcript reconstruction from RNA-Seq spliced alignments
Contemporary strategies for transcript analysis from RNA-Seq

- Spliced alignment of RNA-Seq to genome
- De novo transcript assembly
- Transcript reconstruction from RNA-Seq spliced alignments
Contemporary strategies for transcript analysis from RNA-Seq

- Spliced alignment of RNA-Seq to genome
- De novo transcript assembly
- Align to genome
- Transcript reconstruction from RNA-Seq spliced alignments
- RNA-Seq reads
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Align to genome

Transcript reconstruction from RNA-Seq spliced alignments

Genome

Tumor viruses & microbes?
Contemporary strategies for transcript analysis from RNA-Seq

- Spliced alignment of RNA-Seq to genome
- De novo transcript assembly
- Mutation Detection
Contemporary strategies for transcript analysis from RNA-Seq

- **De novo transcript assembly**
- **Spliced alignment of RNA-Seq to genome**
- **Align to genome**
- **Fusion Detection**

![Diagram showing RNA-Seq reads, spliced alignment, de novo transcript assembly, align to genome, and fusion detection.](image)
Trinity is a highly popular and effect software for *de novo* transcriptome assembly and analysis.

Nature Biotechnology, 2011
~8k citations

Nature Protocols, 2013
~3k citations

Traffic from last 2 weeks on GitHub:
• > 3k unique visitors
• 51 unique cloners

**Trinity** was just recently shown to be top ranked among 10 different assemblers when assessed across diverse transcriptomes and accuracy metrics. Hozer and Marz, Gigascience, May 2019. [https://doi.org/10.1093/gigascience/giz039](https://doi.org/10.1093/gigascience/giz039)
Cancer Transcriptome Analysis Toolkit

Transcript Reconstruction

Mutations + CTAT-Mutations

Fusion Transcripts

Cancer RNA-Seq

Single Cell Tumor Heterogeneity

Viruses & Microbes + Kraken + Centrifuge

Alternative Splicing

Transcript Expression

LncRNAs slinky

Interactive Visualizations and Summary Reports
CTAT Software Logistics

Data Resources

CTAT Genome Library
- GENCODE genome & annotations
- Fusion Library
- COSMIC mutations
- RNA-editing sites
- Metadata for application support.

Software Availability

- GitHub
- Docker
- Singularity
- Bioconda
- Galaxy
- FireCloud
Trinity CTAT Integrates ITCR Collaborations
CTAT Mutations Pipeline – integrates CRAVAT and IGV-reports
CTAT – Fusion Inspector Reports

https://github.com/FusionInspector/FusionInspector/wiki

(all self-contained dynamic html IGV reports generated as outputs)
Benchmarking the CTAT Mutation Pipeline Using the Genome In a Bottle DATA

Precision - Recall for Variant Detection
Truth Set Requires Min 10 RNA-seq Read Coverage

90% Precision & 90% Recall, requiring min. 10 RNA-seq coverage and filtered for RNA-editing sites.
Evaluating Fusion Prediction Accuracy

Evaluated 23 different methods on 56 cancer cell lines.
Evaluations

Also, Top Performer in a DREAM (independent) Competition

DREAM competition organized in part by Josh Stuart and Kyle Ellrott
Generating Cancer Transcriptome Resources

FireCloud
Scalable Cancer Computing Solution for the NCI Cloud

• Leveraged Docker and WDL workflows

• Enabled rapid processing of all of TCGA and GTEx RNA-seq data through STAR-Fusion

Thank you NCI Cloud Credits Program!

~20k samples processed, August-Sept. 2018
Total cost: ~$25k, so ~ $1.25 / sample

Findings supplement our CTAT Fusion Annotation Library
Single Cell Resolution of Tumor Heterogeneity via RNA-Seq

Trinity CTAT Integrates ITCR Collaborations
InferCNV Plug-in to Next Generation Clustered Heatmaps

Detecting the driving fusion in **synovial sarcoma** single cell RNA-Seq

The SS18-SSX fusion is detected in
- 56% of the malignant cells.
- 2% of the non-malignant cells.

Fusion predictions are accurately discriminating between the SS18-SSX1 and SS18-SSX2 fusions.

**Goal:**
Enable CTAT Fusion, Mutation, and Copy Number Profiles Navigable at Single Cell Resolution

Livnat Jerby, Cyril Neftel, Nicolo Riggi, Aviv Regev, Mario Suva
Got Cancer RNA-Seq? Use Trinity CTAT!

Transcript Reconstruction
Mutation detection
Expression
Fusion transcripts
Splicing
Viruses
Single cell tumor heterogeneity

More to come!!!
Latest Developments in the Trinity Cancer Transcriptome Analysis Toolkit

U24: Trinity: Transcriptome assembly for genetic and functional analysis of cancer

Christophoros H. Georgiou1, Brian J. Haem, Maylah D. Fang1, Maxwell Brown, Anna Rogers1

Royal Institute of Technology, Stockholm, Sweden.

Abstract: The Trinity Cancer Transcriptome Analysis Toolkit (TCA) provides researchers with a suite of tools for exploring cancer biology via RNA-seq. The diverse CTAT portfolio largely build single-cell transcriptomes to examine functional and genetic characteristics of cancer at high resolution.

CTAT Mutation Pipeline

Benchmarking against multiple pipelines is shown is Travis Lane (Kleidoscope) and Trimmomatic to CTAT with Ben Astrof (TrimFRAGS) with this data set with a VAF of 30%. In this experiment, we applied the quality filtering criteria to the data set from the second experiment.

Here, we show the performance comparison of the different pipelines on a real-world dataset. These results highlight the importance of rigorous quality control in single-cell RNA-seq experiments.

CTAT Fusion Detection Pipeline

While the de novo assembler-based fusion detection methods are unable to detect the full-length transcripts, the alternative splicing-based methods provide more complete fusion genome sequence evidence and can reconstruct full-length transcript, such as fusion genes.

CTAT Single Cell Copy Number Alteration Detection

Chromosomal structural alterations, including rearrangements and amphipaths, are frequently associated with various diseases such as cancer. Their detection can help guide patient treatment and improve accuracy. The pipeline was designed to identify such changes in single cells from RNA-seq data. We applied the pipeline to a real-world cancer dataset and successfully detected several copy number alterations.

Using NGS and RNA-seq data, we detected multiple cases of chromosome rearrangements in single cells, demonstrating the potential of the CTAT pipeline for detecting copy number alterations in single cells.
Acknowledgements

Team Trinity CTAT

Brian Haas
Aviv Regev
Christophe Georgescu
Vrushali Fangal
Maxwell Brown

CSH
Alex Dobin

Mario Suva
Livnat Jerby

Steven Salzberg

Rachel Karchin
Mike Ryan

Jill Mesirov
James Robinson

John Weinstein
Bradley Bloom
Chris Wakefield

BROAD INSTITUTE

UC San Diego Health Sciences

Informatics Technology for Cancer Research