Trinity: Transcriptome Assembly for Genetic and Functional Analysis of Cancer [U24]

PI: Aviv Regev
Co-Investigator: Brian Haas

Co-PI: Tom Doak

ITCR meeting, May 31 2017
The Cancer Transcriptome

A window into the (expressed) genetic and epigenetic state of a tumor
The Cancer Transcriptome

A window into the (expressed) genetic and epigenetic state of a tumor

![Diagram showing genetic differences between normal and cancer cells.](image)
The Cancer Transcriptome

A window into the (expressed) genetic and epigenetic state of a tumor

![Diagram showing differences between normal and cancerous transcriptomes, highlighting SNPs, splice isoforms, and intron retention changes.](image)
The Cancer Transcriptome

A window into the (expressed) genetic and epigenetic state of a tumor
The Cancer Transcriptome

A window into the (expressed) genetic and epigenetic state of a tumor

+ the associated microbiome, virome...
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

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

Spliced alignment of RNA-Seq to genome

RNA-Seq reads
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

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Transcript reconstruction from RNA-Seq spliced alignments

Genome
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Transcript reconstruction from RNA-Seq spliced alignments

Align to genome
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Transcript reconstruction from RNA-Seq spliced alignments

Genome

Not mapping due to genome restructuring or foreign origin.
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

Not mapping due to genome restructuring or foreign origin.
Contemporary strategies for transcript analysis from RNA-Seq

RNA-Seq reads

Spliced alignment of RNA-Seq to genome

De novo transcript assembly

Transcript reconstruction from RNA-Seq spliced alignments

[diagram showing RNA-Seq reads, spliced alignment, de novo transcript assembly, and transcript reconstruction]

Not mapping due to genome restructuring or foreign origin.
The Ever-Growing Trinity User Community

- ~2k unique users per month
- >4k literature citations (~20% cancer community)
- Open Source software development contributions from the Trinity community.

[GitHub](http://trinityrnaseq.github.io)
User support and training:
• Google group and Twitter feed for community interaction and support.
• Extensive documentation, user guides, tutorials and protocols

• Demo and training videos
• On-site training workshops
Cancer Transcriptome Analysis Toolkit (CTAT)

Goal: to assist cancer researchers in applying RNA-Seq to genetic and functional analyses of cancer
Trinity Cancer Transcriptome Analysis Toolkit

Cancer RNA-Seq

Galaxy

+ Genome Alignments for Reads & Transcripts
Trinity Cancer Transcriptome Analysis Toolkit

- Cancer RNA-Seq
- Mutations
- Fusion Transcripts
- Transcript Expression
- LincRNAs
- Alternative Splicing
- Viruses & Microbes
- Single Cell Tumor Heterogeneity

Galaxy

Genome Alignments for Reads & Transcripts
Trinity Cancer Transcriptome Analysis Toolkit

Cancer RNA-Seq

Mutations

Fusion Transcripts

Transcript Expression

LincRNAs

Alternative Splicing

Viruses & Microbes

Single Cell Tumor Heterogeneity

Genome Alignments for Reads & Transcripts

Interactive Visualizations and Summary Reports
Trinity CTAT Available Through Galaxy via NCGAS at Indiana University

- Transcriptome Assembly
- Mutation Detection
- Fusion Detection

Simply Google: Trinity Galaxy or visit: https://galaxy.ncgas-trinity.indiana.edu/
Galaxy Integration of CTAT Inspectors

Interactive Visualizations and Summary Reports
Trinity Cancer Transcriptome Analysis Toolkit

- Cancer RNA-Seq
- Mutations
- Fusion Transcripts
- Transcript Expression
- LincRNAs
- Genome Alignments for Reads & Transcripts
- Viruses & Microbes
- Alternative Splicing
- Single Cell Tumor Heterogeneity
- Interactive Visualizations and Summary Reports
Mutation Detection Using RNA-Seq

- Cancer RNA-Seq
- Galaxy
- Trinity
- Live with Galaxy
- Mutations
- Genome Alignments for Reads & Transcripts
- Single Cell Tumor Heterogeneity
- Viruses & Microbes
- Alternative Splicing
- Fusion Transcripts
- Transcript Expression
- LincRNAs

Interactive Visualizations and Summary Reports
Trinity CTAT Cancer Mutation Identification Module

Custom visualizations & reports are made possible by collaboration among multiple ITCR groups.

- RNA-Seq Fastqs
- GATK Best Practices
- Variant Annotation
- Variant Filtration
- Visualization

- CRAVAT
- MuPIT
- IGV.js

- Indiana University
- Galaxy
- Mesirov & Robinson
- Karchin & Ryan
Mutation Inspector Report
Exploring a Single Variant

Powered by IGV.js

(ITCR collaborators – Jill Mesirov and James Robinson)
MuPIT Provides Additional 3D Context for Mutation

Visualize variants within protein structure.

(ITCR collaborators - Rachel Karchin and Mike Ryan)
Fusion Transcript Detection

- Cancer RNA-Seq
- Genome Alignments for Reads & Transcripts
- Mutations
- Fusion Transcripts
- Transcript Expression
- LincRNAs
- Viruses & Microbes
- Alternative Splicing
- Single Cell Tumor Heterogeneity

Interactive Visualizations and Summary Reports
Top-down Approaches to Fusion Transcript Discovery

**Paired-end Illumina RNA-Seq**

**STAR-Fusion**

* In collaboration with Alex Dobin, developer of STAR

Align *reads* to the genome, Identify discordant pairs and junction/split reads.

**De novo RNA-Seq assembly**

**Trinity or Oases (MK)**

**GMAP-fusion**

* In collaboration with Tom Wu, developer of GMAP

Align *transcripts* to genome, Identify Fusion Transcripts
Top-down Approaches to Fusion Transcript Discovery

Paired-end Illumina RNA-Seq

STAR-Fusion
* In collaboration with Alex Dobin, developer of STAR

Align reads to the genome, Identify discordant pairs and junction/split reads.

De novo RNA-Seq assembly

Trinity or Oases (MK)

GMAP-fusion
* In collaboration with Tom Wu, developer of GMAP

DISCASM

STAR-alignments

Just discordant or unmapped reads

Align transcripts to genome, Identify Fusion Transcripts

Chr-A

Chr-B

/1

Junction read

/1

Spanning frag

/2

/2

Chr-A

Chr-B
New Results

**STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq**

Brian Haas, Alexander Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas Doak, Natalie Pochet, Jing Sun, Catherine Wu, Thomas Gingeras, Aviv Regev

doi: https://doi.org/10.1101/120295

This article is a preprint and has not been peer-reviewed [what does this mean?].
Benchmarking Fusion-finding Tools

- **Simulated data**
  - 5 replicates
  - 2500 Simulated fusions
  - 30M PE sim RNA-Seq data

- **Genuine data**
  - 65 Cancer Cell Lines

Precision – Recall Curves

Precision = TP / (TP + FP)
Recall = TP / (TP + FN)

Accuracy = area under the curve (AUC)
Benchmarking Fusion-finding Tools
(results shown for simulated data)

Fusion Prediction Accuracy
(AUC value distribution across 5 replicates)

Fusion Prediction Sensitivity vs. Expression

<table>
<thead>
<tr>
<th>STAR–Fusion</th>
<th>nFuse</th>
<th>InFusion</th>
<th>ChimPipe</th>
<th>JAFFA–Direct</th>
<th>ChimeraScan</th>
<th>TopHat–Fusion</th>
<th>deFuse</th>
<th>MapSplice</th>
<th>FusionCatcher</th>
<th>JAFFA–Hybrid</th>
<th>EricScript</th>
<th>PRADA</th>
<th>SOAP–fuse</th>
<th>JAFFA–Assem</th>
<th>FusionHunter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AUC values

Expression (log₂[TPM]) bin

Percent of fusions
STAR-Fusion is accurate and **FAST**
New Results

**STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq**

Brian Haas, Alexander Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas Doak, Natalie Pochet, Jing Sun, Catherine Wu, Thomas Gingeras, Aviv Regev

doi: https://doi.org/10.1101/120295

---

**Reproducible data analysis**
- All fusion prediction results from all programs included
- Single command to reanalyze data, generate all figures and tables
New Results

**STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq**

Brian Haas, Alexander Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas Doak, Natalie Pochet, Jing Sun, Catherine Wu, Thomas Gingeras, Aviv Regev

doi: https://doi.org/10.1101/120295
Bottom-up Fusion ‘In silico Validation’ Using FusionInspector

Add to whole genome. Align reads, score and assess.

* STAR enhancements to support FusionInspector

Make mini-fusion contigs

All fusion predictions
FusionInspector Fusion View

Powered by IGV.js

ITCR collaborators - Jim Robinson and Jill Mesirov
Driving Cancer Project: Search for Gene Fusions in Chronic Lymphocytic Leukemia (CLL)

- A common adult leukemia in Europe and North America
- Tremendous clinical heterogeneity
- Incurable by conventional chemotherapy
- Molecular understanding largely unknown

* Work done in collaboration with Cathy Wu, Dana Farber Cancer Center & BI.
Defining Recurrent Fusion Transcripts in Chronic Lymphocytic Leukemia

270 CLL tumor samples, 18 normal samples

<table>
<thead>
<tr>
<th>Filter GTEx</th>
<th>Total fusions</th>
<th>Unique fusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>8322</td>
<td>4898</td>
<td></td>
</tr>
<tr>
<td>6587 (79%)</td>
<td>4644 (95%)</td>
<td></td>
</tr>
<tr>
<td>5893 (71%)</td>
<td>4558 (94%)</td>
<td></td>
</tr>
<tr>
<td>599 (7%)</td>
<td>320 (7%)</td>
<td></td>
</tr>
<tr>
<td>341 (4%)</td>
<td>62 (1%)</td>
<td></td>
</tr>
</tbody>
</table>

Expression Filter
0.1 FFPM

Define recurrent
Top 25/62 Recurrent Fusions in CLL

(*) Known oncogene
Validating Fusion Predictions via RT-PCR

SuperScript 2, random primer

ThermoScript, random primer

ThermoScript, origo-dT primer

* Work by Jintaek Kim

V mark means validated by Sanger-seq.
Targeted Fusion Transcript Validated by Sanger Sequencing

Alternatively Spliced Fusion Transcript Validated

* Work by Jintaek Kim, Dana Farber
Single Cell Tumor Heterogeneity

Cancer RNA-Seq

Genome Alignments for Reads & Transcripts

Mutations

Viruses & Microbes

Fusion Transcripts

Alternative Splicing

Transcript Expression

LincRNAs

Interactive Visualizations and Summary Reports
Centrifuge +/- Trinity Applied to HPV-Driven HNSCC

* Collaboration with Steven Salzberg, JHU. Centrifuge: Kim et al. Genome Research, 2016

**Longer Sequences and Taxonomic Resolution Using Trinity**

* Trinity reconstructed HPV contigs. (259 base to 1.3 kb length)

* RNA-Seq data from: Gene expression analysis of TIL rich HPV-driven head and neck tumors reveals a distinct B-cell signature when compared to HPV independent tumors. Wood et al. Oncotarget, 2016
Single Cell Tumor Heterogeneity

Cancer RNA-Seq → Galaxy → Trinity

Genome Alignments for Reads & Transcripts

Mutations
Fusion Transcripts
Transcript Expression
LincRNAs
Alternative Splicing
Viruses & Microbes

Interactive Visualizations and Summary Reports
Single Cell Resolution of Tumor Heterogeneity via RNA-Seq

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

Tirosh, Izaar, ...., Regev, Garraway; Science 2016

Large-scale Copy Number Variation Inferred from Single Cell RNA-Seq Data

Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma

Patel, Tirosh, ..., Regev, Bernstein; Science 2014
Trinity CTAT InferCNV: Utility to identify large-scale CNV from single cell RNA-Seq

https://github.com/broadinstitute/inferCNV

First official software release: May 30, 2017

* Example from oligodendroglioma
Firecloud

Scalable Cancer Computing Solution for the NCI Cloud

- Integration of Trinity CTAT into Docker and WDL workflows
- Process TCGA data
- Shareable workflows and data resources

Integration of Trinity CTAT into:

Clinical Research Sequencing Platform (CRSP)

Pilot study in pediatric oncology underway
Got Cancer RNA-Seq? Run Trinity!

https://galaxy.ncgas-trinity.indiana.edu/

Lots more to come!!!
Acknowledgements

Aviv Regev
Brian Haas
Timothy Tickle
Asma Bankapur
Christophe Georgescu
Ami-levy Moonshine

UC San Diego
Health Sciences
Jill Mesirov
James Robinson

BROAD INSTITUTE

CSH
Alex Dobin

BRIGHAM AND WOMEN’S HOSPITAL
Nathalie Pochet

JOHNS HOPKINS
BIOMEDICAL ENGINEERING
Steven Salzberg

ITCR
Informatics Technology for Cancer Research

DANA-FARBER CANCER INSTITUTE
Cathy Wu
Jing Sun
Peggy Hsu
Sachet Shukla

Thomas Doak
Carrie Ganote
Robert Henschel
Cicada Brokaw