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

ITCR meeting, June 2016
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 comparison between normal and cancerous states with SNPs and mutations](image)
The Cancer Transcriptome

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

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

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

- SNPs
- Splice Isoforms
- Intron Retention
- Gene Fusions

Normal

Cancer

Junction
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
Contemporary strategies for transcript analysis from RNA-Seq

Spliced alignment of RNA-Seq to genome

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

De novo transcript assembly

Spliced alignment of RNA-Seq to genome

RNA-Seq reads

Transcript reconstruction from RNA-Seq spliced alignments

Genome
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
Contemporary strategies for transcript analysis from RNA-Seq

De novo transcript assembly

Spliced alignment of RNA-Seq to genome

Transcript reconstruction from RNA-Seq spliced alignments

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.
The Ever-Growing Trinity User Community

- ~1.5k unique users per month
- >3k 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

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 → Trinity + 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
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 Cancer Transcriptome Analysis Toolkit

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

Interactive Visualizations and Summary Reports
Mutation Detection Using RNA-Seq

Cancer RNA-Seq → Galaxy → Trinity → Genome Alignments for Reads & Transcripts

Mutations → Fusion Transcripts → Transcript Expression → LincRNAs

Viruses & Microbes → Single Cell Tumor Heterogeneity → Interactive Visualizations and Summary Reports
Custom visualization is a product of multiple labs efforts.
Mutation Analysis and Visualization from Within Galaxy

- Table of Predicted Variants with scores, attributes and rankings.
- Individual mutation report, including genome evidence view and annotations.
- Mupit 3D protein structure view
  (ITCR - Rachel Karchin and Mike Ryan)

** will demo **
Fusion Transcript Detection

1. Cancer RNA-Seq
2. Genome Alignments for Reads & Transcripts
3. Mutations
4. Fusion Transcripts
5. Transcript Expression
6. LincRNAs
7. Viruses & Microbes
8. Alternative Splicing
9. Single Cell Tumor Heterogeneity

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

**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

---

**Paired-end Illumina RNA-Seq**

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

**Chr-A**

**Chr-B**

**Junction read**

**Spanning frag**

**/1 /1**

**/2 /2**
Align reads to the genome, Identify discordant pairs and junction/split reads.

De novo RNA-Seq assembly

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

Align transcripts to genome, Identify Fusion Transcripts

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

Paired-end Illumina RNA-Seq

STAR-alignments

Just discordant or unmapped reads

DISCASM

Paired-end Illumina RNA-Seq

Compare to:

Prada
SoapFuse
TophatFusion
ChimeraScan
Defuse
FusionCatcher
Ericscript
FusionHunter
Maps splice
Jaffa

* In collaboration with Alex Dobin, developer of STAR

* In collaboration with Tom Wu, developer of GMAP

START alignsments to the genome, identifying discordant pairs and junction/split reads. The output can be used for de novo RNA-Seq assembly or for fusion transcript discovery.

GMAP-fusion aligns transcripts to the genome, identifying fusion transcripts.

Competition tools include:

Prada
SoapFuse
TophatFusion
ChimeraScan
Defuse
Ericscript
FusionHunter
Maps splice
Jaffa
Cancer-associated Viruses and Microbiome

Evaluation of Fusion-Finding Accuracy
(using 75 Cancer Cell Lines, with TP = min 3 tools agree)

All Fusion Prediction Accuracies

Fusion predictions ranked according to min evidence support.

De novo Assembly-based Fusion Prediction Accuracy

Cancer-associated Viruses and Microbiome
STAR-Fusion and DISCASM/Trinity Improve on both Speed and Accuracy of Fusion Detection

Using 30M PE reads, 5 samples ea.
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
LincRNA Identification

Cancer RNA-Seq

Genome Alignments for Reads & Transcripts

Mutations

Fusion Transcripts

Transcript Expression

LincRNAs

Viruses & Microbes

Single Cell Tumor Heterogeneity

Alternative Splicing

Interactive Visualizations and Summary Reports
Considers conserved ORFs and dN/dS

Example:

Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (MALAT1), non-coding RNA

Freely available, open source: https://slncky.github.io/
SLNCKY - LncRNA Analysis and Visualization from Within Galaxy

Table of Predicted LncRNAs with their orthologs

Individual LncRNA report, including alignment view and evolutionary metrics.

** will demo **
SLNCKY-based Re-discovery of PCAT1: Prostate Cancer Associated Transcript 1

Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature Biotechnology, 2011

John R Prensner\textsuperscript{1,8}, Matthew K Iyer\textsuperscript{1,8}, O Alejandro Balbin\textsuperscript{1}, Saravana M Dhanasekaran\textsuperscript{1,2}, Qi Cao\textsuperscript{1}, J Chad Brenner\textsuperscript{1}, Bharathi Laxman\textsuperscript{3}, Irfan A Asangani\textsuperscript{1}, Catherine S Grasso\textsuperscript{1}, Hal D Kominsky\textsuperscript{1}, Xuhong Cao\textsuperscript{1}, Xiaojun Jing\textsuperscript{1}, Xiaojue Wang\textsuperscript{1}, Javed Siddiqui\textsuperscript{1}, John T Wei\textsuperscript{1}, Daniel Robinson\textsuperscript{1}, Hari K Iyer\textsuperscript{3}, Nallasivam Palanisamy\textsuperscript{1,2,6}, Christopher A Maher\textsuperscript{1,2} & Arul M Chinnaiyan\textsuperscript{1,2,4,6,7}
Single Cell Tumor Heterogeneity

Cancer RNA-Seq

Galaxy → Trinity → Genome Alignments for Reads & Transcripts

Mutations

Fusion Transcripts

Transcript Expression

LincRNAs

Viruses & Microbes

Alternative Splicing

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

Chromosome gain/loss
Differential Expression of surface receptors
TP53 Mutations

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

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq
Tirosh, Izaar, ... , Regev, Garraway; Science 2016
Trinity CTAT InferCNV: Utility to identify large-scale CNV from single cell RNA-Seq

https://github.com/broadinstitute/inferCNV

CTAT InferCNV by Tim Tickle and Itay Tirosh
Goal: cancer transcriptome toolkit accessible to any cancer researcher

Starting point: RNA-Seq data (fastq files)
Access Trinity CTAT via Galaxy

The National Center for Genome Analysis Support hosts the public web interface for running Trinity jobs.

- Backed by three devoted nodes running on the Karst system with 512GB memory each

We’re steadily growing since the official launch in January, 2015

Total Galaxy Users per Month

![Graph showing the increase in Galaxy users per month from December 2014 to May 2016.](image)
Scaling from individual samples
Scaling from individual samples to many samples
Firecloud

Scalable Cancer Computing Solution

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

Also enables integration of Trinity CTAT into:

Clinical Research Sequencing Platform (CRSP)
Got Cancer RNA-Seq? Run Trinity!

Trinity

Mutation detection
Expression
Single cell tumor heterogeneity
Fusion transcripts
Splicing
Viruses

Lots more to come!!!

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

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

Alex Dobin

Cathy Wu
Jing Sun
Peggy Hsu
Jintaek Kim
Sachet Shukla
Dan Landau

Nathalie Pochet
Nik Obholzer

Jill Mesirov
James Robinson

Tom Wu

Bill Barnett
Thomas Doak
Carrie Ganote
Robert Henschel
Ben Fulton