

Dec 06, 2024

Mouse CGRm39 tdTomato Trace Alignment

DOI

<https://dx.doi.org/10.17504/protocols.io.kxygxw3bzv8j/v1>

Michael Morley¹, Joseph.Planer¹

¹Upenn



Michael Morley

Upenn

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.kxygxw3bzv8j/v1>

Protocol Citation: Michael Morley, Joseph.Planer 2024. Mouse CGRm39 tdTomato Trace Alignment. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.kxygxw3bzv8j/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: December 06, 2024

Last Modified: December 06, 2024

Protocol Integer ID: 114508

Keywords: scRNA-seq, lineage tracing, mouse cgrm39 tdtomato trace alignment, mouse cgrm39 tdtomato trace alignment this pipeline, mouse genome reference crcm39, gencode m27 gene annotation, cell lineage, lineage tracing, transcripts mapping, expression of transcripts mapping, sequencing depth, tdtomato, transcript, gene, cell

Funders Acknowledgements:

Endothelial cell signaling in regeneration of the lung

Grant ID: R01HL162683

Abstract

This pipeline is modification of standard STARsolo pipeline to enable cell lineage tracing. Here we start with Mouse genome reference CRCm39 and Gencode M27 gene annotation. For lineage tracing, cells were annotated based on expression of transcripts mapping to either the 871bp region overlapping the 3x SV40 polyA or the 1kb region overlapping the bGH polyA following the tdTomato coding sequence. These regions which are labeled site A/B were chosen because of the higher sequencing depth. A simple call procedure can be used to annotate a cell as 'Traced' if >50% of transcripts from Site B, the recombined allele.

Attachments



features.gtf

0B



tdTomato-no-ROSA26.f...

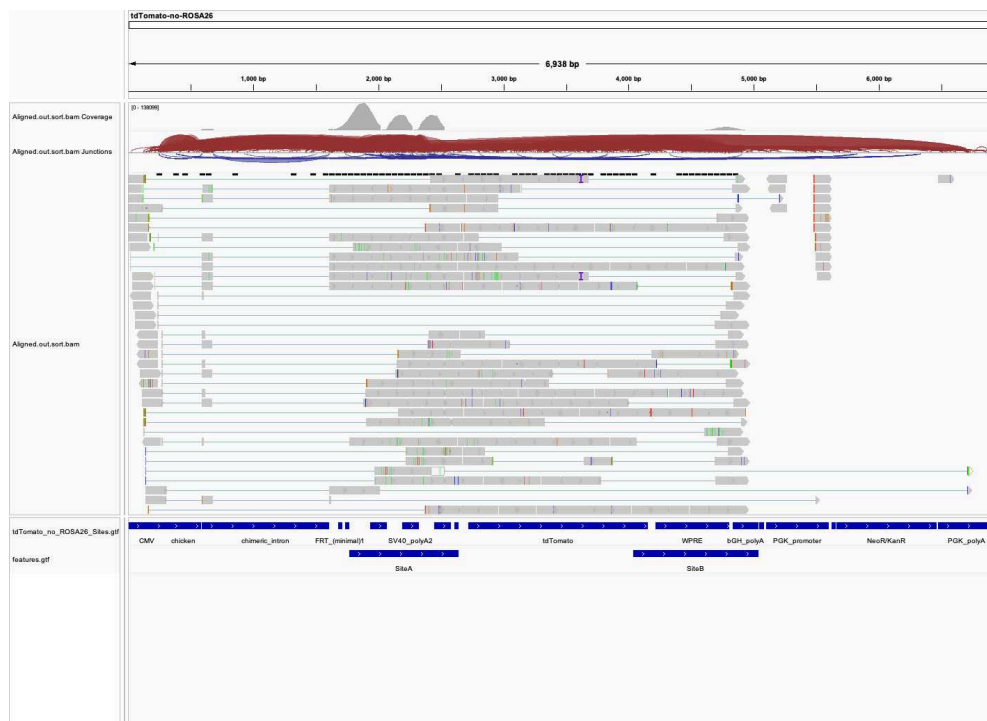
6KB

Create Custom reference for STAR

- 1 We need to add the tdTomato construct sequence to the GRCm39 reference genome. The fasta file is available in the files section

```
cat tdTomato-no-ROSA26.fasta GRCm39.genome.fa >
GRCm39.genome.tdTomato.fa
```

- 2 We need to add features for this sequence. We create 2 features SiteA 871bp region overlapping the 3x SV40 polyA and Site B 1kb region overlapping the bGH polyA following the tdTomato coding sequence. these sites can be appended to the Gene Annotation GTF



```
cat gencode.vM27.annotation.gtf features.gtf >
gencode.vM27.annotation_Tdtomato.gtf
```

3 Create a STAR reference

```
STAR --runThreadN 20 \  
--runMode genomeGenerate \  
--genomeDir mm39_tdTomato_STAR \  
--genomeFastaFiles GRCm39.genome.tdTomato.fa \  
--sjdbGTFfile gencode.vM27.annotation_Tdtomato.gtf \  
--sjdbOverhang 100
```

Alignment with STARsolo (version 2.9.7a)

- 4 The STARsolo algorithm is used as a fast, memory-efficient alternative to 10x Genomics Cell Ranger.

STARsolo directly aligns raw sequencing reads from scRNA-seq data to the human reference genome.

Like Cell Ranger, STARsolo performs:

Barcode and UMI (Unique Molecular Identifier) extraction: Identifies cell barcodes and UMIs from the raw sequencing reads.

Read alignment: Reads are aligned to the genome using STAR.

Transcript quantification: Assigns reads to genes based on the alignment to compute gene expression per cell.

Gene-cell matrix generation: Produces a matrix similar to Cell Ranger's output, containing gene expression counts across cells.

Achieving Cell Ranger v3.0 Output:

To match the output of Cell Ranger v3.0, specific parameters are set within the STARsolo command, particularly:

--soloType CB_UMI_Simple: Defines the type of input for the barcodes and UMIs.

--soloCBwhitelist: Uses the 10x Genomics whitelist to ensure correct identification of valid barcodes.

--soloUMIdedup: UMI deduplication is set to handle potential PCR artifacts by counting each transcript only once per UMI.



--soloFeatures: This is set to generate gene expression matrices, capturing standard features like gene and transcript annotations.

--soloCellFilter EmptyDrops_CR

Post-Processing:

Once the alignment and quantification are completed, the resulting gene-cell matrix is further processed using SoupX to remove ambient RNAs

Determine Traced Cells

5

By adding in the Site A/B features these will be counted and present in the count matrix. SiteA would have counts from the un-recombined allele and SiteB from the recombined allele. A simple ratio of SiteB/SiteA+SiteB can be used to define a cutoff to determine a "traced".