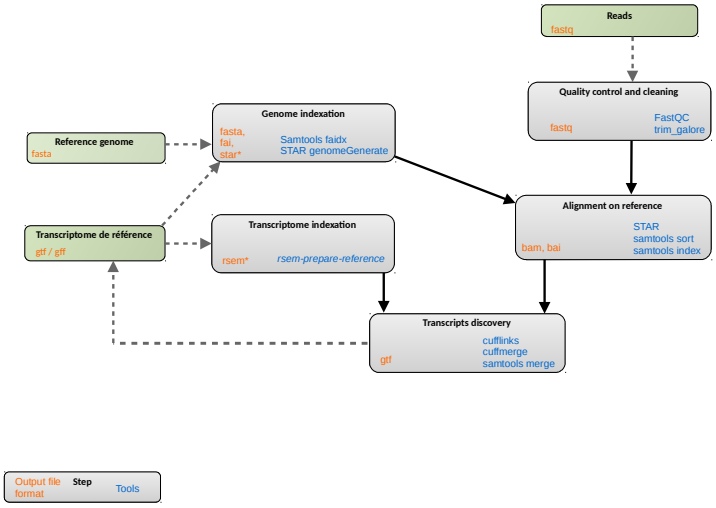
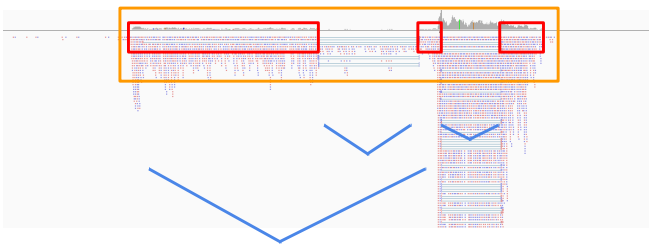
Analysis workflow



Summary -
mRNA calling & model comparison

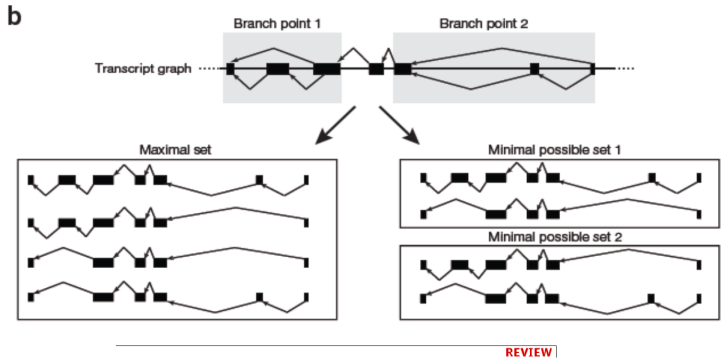
- How to reconstruct transcript ?
- Cufflinks
- Compare models (cuffcompare)
- Merge annotation (cuffmerge)
- Which strategy ?

Transcript reconstruction



Gene location
Exon location
Junctions :
- between read pair junction
- within read junction

Model building strategies



REVIEW

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

Cufflinks

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

日本語要約

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | Contributions | Corresponding author

Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621

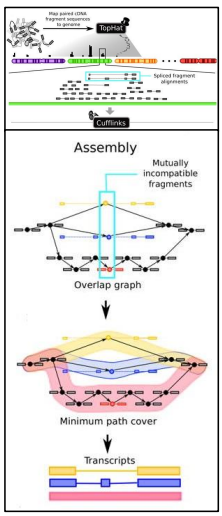
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

http://cole-trapnell-lab.github.io/cufflinks/

- **assembles transcripts**
- estimates their abundances: based on how many reads support each one
- Suite of software : cufflinks, cuffmerge, cuffcompare

Cufflinks transcript assembly

- Transcripts assembly:
  - fragments are divided into non-overlapping loci
  - each locus is assembled independently
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem):
    - nb incompatible read = mini nb of transcripts needed
    - each path = set of mutually compatible fragments overlapping each other



Cufflinks inputs and options

```
module load bioinfo/cufflinks-2.2.1
```

- Command line:

```
cufflinks [options] <aligned_reads.(sam/bam)>
```

- Some options:

```
-h/--help
-o/--output-dir
-p/--num-threads
```

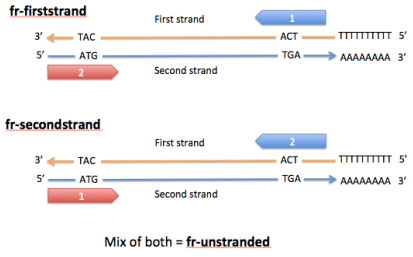
`-G/--GTF <reference_annotation.(gtf/gff)>`
estimate isoform expression, no novel transcripts

`-g/--GTF-guide <reference_annotation.(gtf/gff)>`
use reference transcript annotation to guide assembly

```
--max-bundle-length [3,500,000]
--max-bundle-frags [500,000]
--library-type
```

library prep used for input reads

Cufflinks library types



Cufflinks outputs

- **`transcripts.gtf`**
  contains assembled isoforms (coordinates and abundances)
- **`genes.fpkm_tracking`**
  contains the genes FPKM
- **`isoforms.fpkm_tracking`**
  contains the isoforms FPKM
- **`skipped.gtf`**
  contains skipped loci (too many fragments)

Cufflinks GTF description

`transcripts.gtf` (coordinates and abundances):

- contains assembled isoforms
- can be visualized with a genome viewer
- attributes: ids, FPKM, confidence interval, read coverage & support
  - score: most abundant isoform = 1000
    minor isoforms = minor FPKM/major FPKM
  - cov: estimate for depth across the transcript

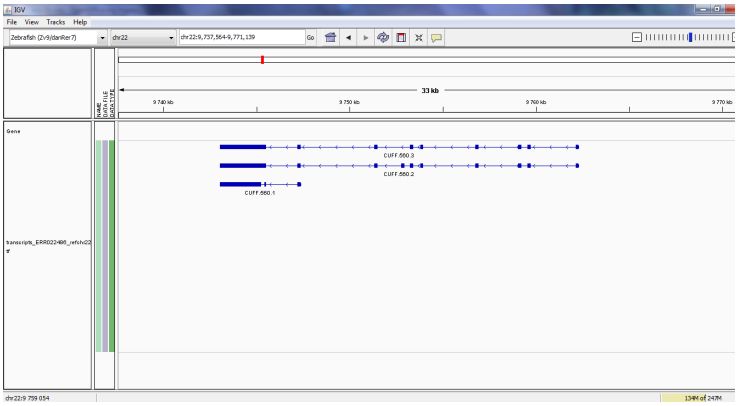
```
1   Cufflinks   transcript  459812  460830  1       -   .
1   Cufflinks   exon        459812  460830  1       -   .
1   Cufflinks   transcript  463572  478996  1000    -   .
1   Cufflinks   exon        463572  463746  1000    -   .
1   Cufflinks   exon        466228  466405  1000    -   .
```

```
gene_id "ENSDTAG00000013841"; transcript_id "ENSDTAT00000018387"; FPKM "0.0000000000"; frac "0.000000";
gene_id "ENSDTAG00000013841"; transcript_id "ENSDTAT00000018387"; exon_number "1"; FPKM "0.0000000000"; frac "0.000000";
gene_id "CUFF.2"; transcript_id "ENSDTAT00000015319"; FPKM "25.4745974237"; frac "1.000000";
gene_id "CUFF.2"; transcript_id "ENSDTAT00000015319"; exon_number "1"; FPKM "25.4745974237"; frac "1.000000";
gene_id "CUFF.2"; transcript_id "ENSDTAT00000015319"; exon_number "2"; FPKM "25.4745974237"; frac "1.000000";
```

```
conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000"; full_read_support "no";
conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";
conf_lo "21.387219"; conf_hi "29.561976"; cov "422.904985"; full_read_support "yes";
conf_lo "21.387219"; conf_hi "29.561976"; cov "422.904985";
conf_lo "21.387219"; conf_hi "29.561976"; cov "422.904985";
```


Cufflinks GTF description

`transcripts.gtf` (coordinates and abundances):
visualization in IGV



Cufflinks / Cuffcompare

Compare assemblies between conditions:

- compare your assembled transcripts to a reference annotation
- track Cufflinks transcripts across multiple experiments

Command:

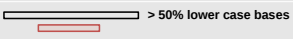
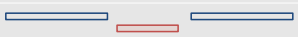
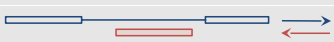
```
cuffcompare [-r <reference.gtf>] [-o <outprefix>] <input1.gtf>
…
```

Outputs:

- `<outprefix>.stats` - overall summary statistics
- `<outprefix>.combined.gtf` - "union" of all transfrags
- `<cuff_in>.refmap` - transfrags matching to reference transcript
- **`<cuff_in>.tmap`** - best reference transcript for each transfrag
- `<outprefix>.tracking` - tracking transfrags across samples

Cuffcompare

Class code de cuffcompare

=	complete match	
c	contained	
j	novel isoform	
e	single exon	
i	within intron	
o	exonic overlap	
p	polymerase run-on	
r	repeat	
u	unknown, intergenic	
x	exonic overlap on the opposite strand	
s	intronic overlap on the opposite strand	

13

http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#transfrag-class-codes

Cufflinks / Cuffmerge

Merge together several assemblies:

- merge novel isoforms and known isoforms
- filters a number of transfrags that are probably artifacts
- build a new gene model describing all conditions

Command:

```
cuffmerge [options] -o <assembly_GTF_list>
```

Options:

- `-o/--output-dir`
- `-g/--ref-gtf`
- `-s/--ref-sequence`
- `--min-isoform-fraction`
  discard isoforms with abundance below this [0.05]
- `-p/--num-threads`

14

Cufflinks / Cuffmerge

`merged.gtf` (coordinates and legacy):

- contains merged input assemblies
- can be visualized with a genome viewer
- attributes: ids, name, oId, nearest_ref, class_code, tss_id, p_id

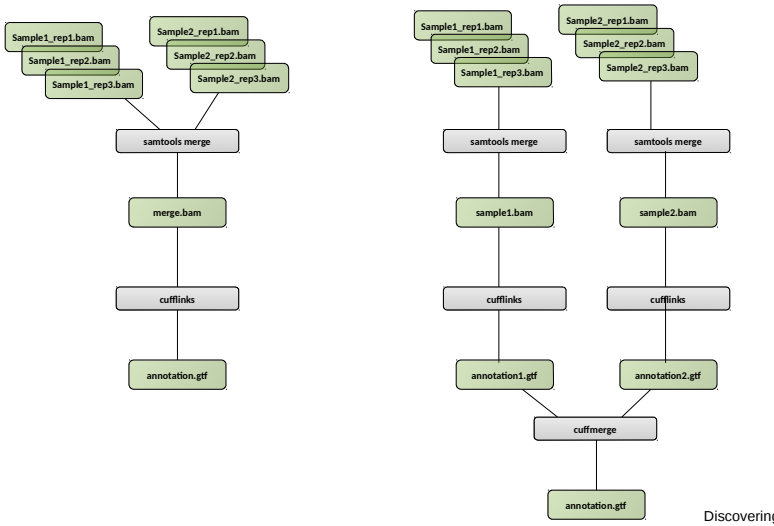
```
1  Cufflinks  exon  34627   35558   .  +  .
1  Cufflinks  exon  242394  242646  .  +  .
1  Cufflinks  exon  275623  275681  .  +  .
1  Cufflinks  exon  242402  242646  .  +  .
1  Cufflinks  exon  254559  254693  .  +  .
1  Cufflinks  exon  247340  249673  .  +  .
1  Cufflinks  exon  351546  351874  .  +  .
1  Cufflinks  exon  355064  355237  .  +  .
1  Cufflinks  exon  357793  357952  .  +  .
1  Cufflinks  exon  361144  362915  .  +  .
```

```
gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; gene_name "ENSBTAG00000006858";
gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "2"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "1";
gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "2";
gene_id "XLOC_000003"; transcript_id "TCONS_00000004"; exon_number "1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "1"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "2"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "3"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "4"; gene_name "RCAN1";
```

```
oId "ENSBTAT00000009004"; nearest_ref "ENSBTAT00000009004"; class_code "="; tss_id "TSS1";
oId "CUFF.1.1"; nearest_ref "ENSBTAT00000007283"; class_code "x"; tss_id "TSS2";
oId "CUFF.1.1"; nearest_ref "ENSBTAT00000007283"; class_code "x"; tss_id "TSS2";
oId "CUFF.1.2"; class_code "u"; tss_id "TSS2";
oId "CUFF.1.2"; class_code "u"; tss_id "TSS2";
oId "CUFF.2.1"; class_code "u"; tss_id "TSS3";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
```

15

Which strategy ?



Hands-on: transcripts assembly

Using cufflinks:

Exercise 6:
- reconstruct known and novel transcripts
- compare annotations