期末報告

第二組

魏孝全	104354025	劉義瑋	104753013
江易倫	104753018	李恭儀	105753006
楊宗翰	105753024	楊子菱	105753041

Outline

- Paper 介紹
- 開發環境
- 工具準備
- 資料準備
- 實驗流程
- 實驗結果
- 結論
- 參考資料
- 小組分工

Paper介紹

Shiao, M.-S. S. et al. Expression Divergence of Chemosensory Genes between Drosophila sechellia and Its Sibling Species and Its Implications for Host Shift. Genome Biol Evol 7, 2843–58 (2015).

- Several odorant binding protein (Obp) genes and olfactory receptor (Or) genes have been suggested to be associated with the D. sechellia host shift

- Or genes and Ir genes are expressed mostly in olfactory sensory neurons on different types of sensilla in antennae.

Paper介紹

Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014).

開發團隊將這套分析流程取名為燕尾服 (Tuxedo Suite)，目的是將NGS技術定序RNA-Seq的reads透過alignment到已知基因體序列後，進行 transcript reconstruction、genes/transcripts 定量及分析不同condition間有顯著差異的genes/ transcripts，最後搭配R語言將統計結果圖像化。

- 使用tophat將reads map回genome
- 接著會透過Cufflinks做Transcript Reconstruction，以genome為參考序列進行reference-based assembly
- 透過Cuffdiff計算顯著表現的genes/ transcripts
- 使用CummeRbund呈現統計結果。

開發環境

- Ubuntu v14.04.5 LTS
- RStudio with R v3.3.2

工具準備

1. TopHat - v2.1.1
   - Fast splice junction mapper for RNA-Seq reads
   - Trapnell, C., Pachter, L. & Salzberg, S. TopHat: discovering splice junctions with RNA-Seq. Method Anal 25, 1105–1111 (2009). Citation: 5107
2. Cufflinks、Cuffmerge、Cuffdiff - v2.2.1
   - Transcriptome assembly and differential expression analysis for RNA-Seq
   - Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcrip isoform switching during cell differentiation. Nat Biotechnol 28, 511–515 (2010). Citation: 4544
3. Bowtie2 - v2.2.9
   - Tool for aligning sequencing reads to long reference sequences
   - Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). Citation: 5246
4. CummeRbund - v2.7.2
   - R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output

工具準備

5. Nextflow - v0.22
    - Data-driven computational pipelines
6. proto
    - R package for argparse
7. argparse
    - R package for parse argv
8. ggplot2
    - R package for plot

資料準備

- fastq files：raw sequence data file
  - EMBL European Bioinformatics Institute
    - GSE67587、GSE67861、GSE67862
  - DNA Data Bank of Japen
  - National Center for Biotechnology Information
    - Convert SRA file to fastq file
- fasta files：reference genome file
  - FlyBase
    - Drosophila sechellia fasta (GSE67587 & GSE67861)
    - Drosophila simulans fasta (GSE67862)
- gtf files：gene annotation file
  - FlyBase
    - Drosophila sechellia gft (GSE67587 & GSE67861)
    - Drosophila simulans gft (GSE67862)
- genome index：bowtie2-build XXXX.fasta index
  - use this coommand to build genome index

資料準備

Species	Gender	Run accession	Read pairs (bp)	Fastq size (GB)
Drosophila sechellia GSE67587	Male	SRR1952772	31828985	16
		SRR1952773	29443050	14.8
		SRR1952774	30741692	15.4
	Female	SRR1952775	36098846	24
		SRR1952776	33062046	22
		SRR1952777	34572646	24

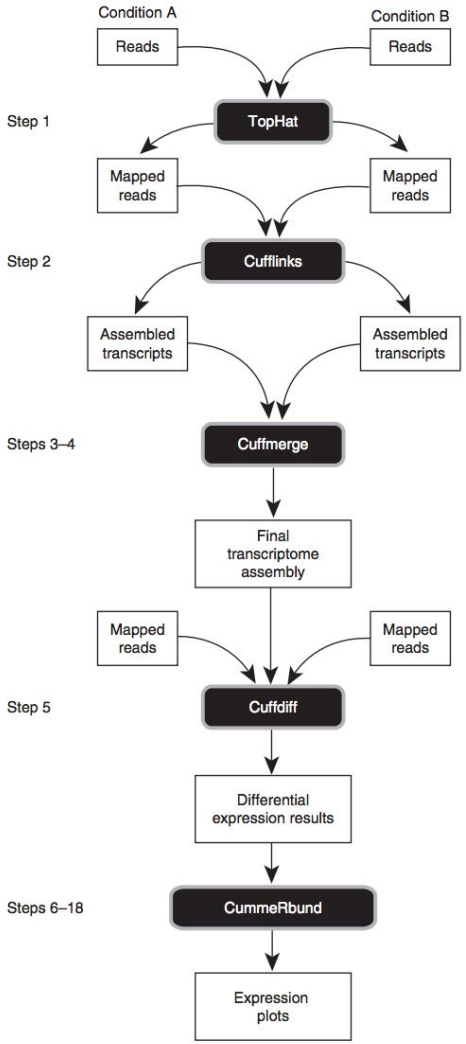
資料準備

Species	Gender	Run accession	Read pairs (bp)	Fastq size (GB)
Drosophila sechellia GSE67861	Male	SRR1973486	40328940	22
		SRR1973487	37101021	24
		SRR1973488	38741238	26
	Female	SRR1973489	29985555	19.4
		SRR1973490	28631216	18.4
		SRR1973491	29656654	19.2

資料準備

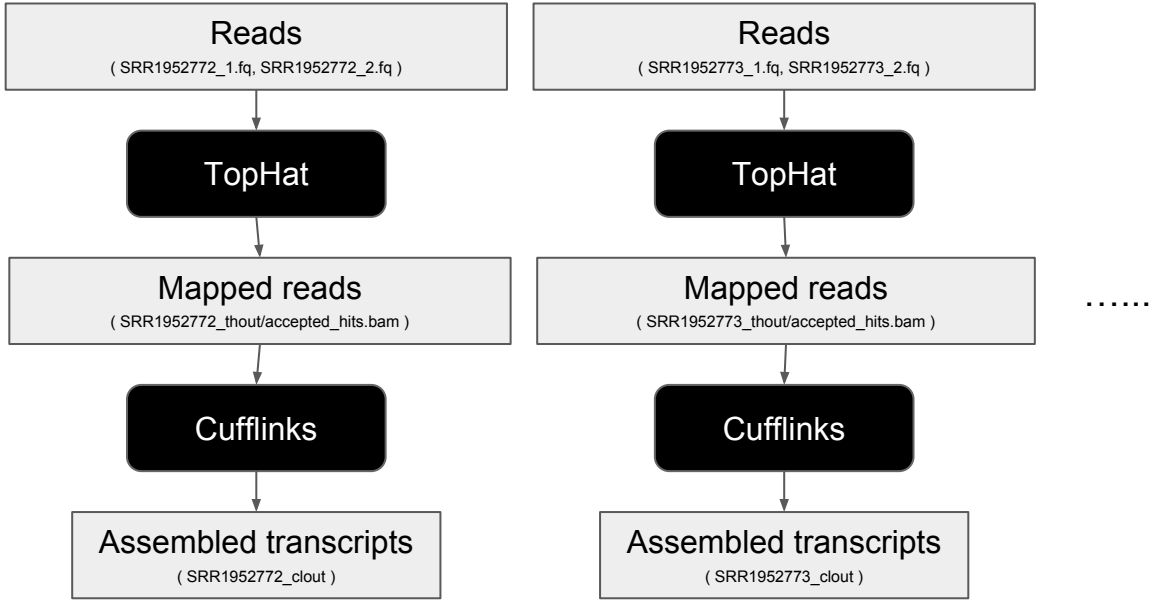
Species	Gender	Run accession	Read pairs (bp)	Fastq size (GB)
Drosophila simulans GSE67862	Male	SRR1973492	28225417	14.2
		SRR1973493	26008321	13
		SRR1973494	27138172	14.6
	Female	SRR1973495	38340684	19.2
		SRR1973496	35245620	17.6
		SRR1973497	36830863	18.4

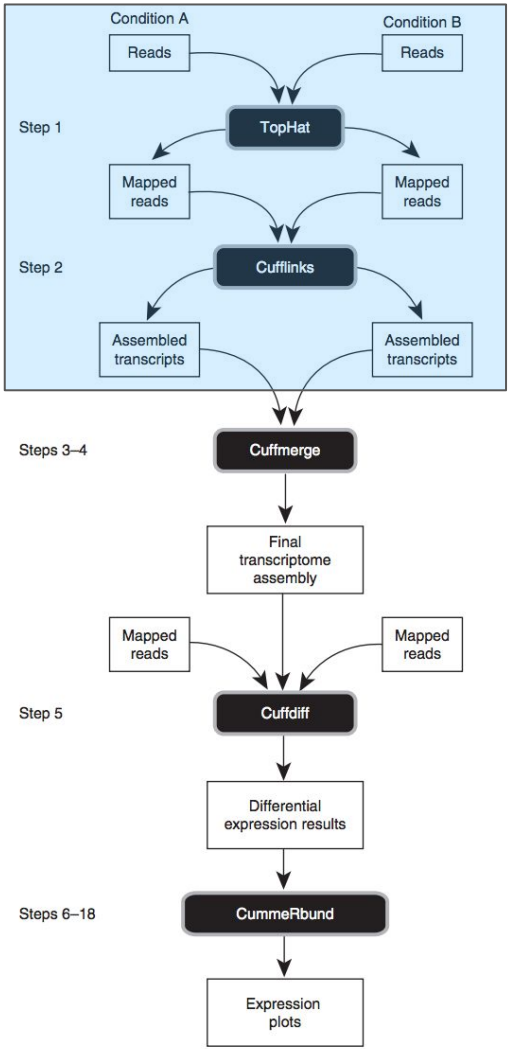
實驗流程



Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

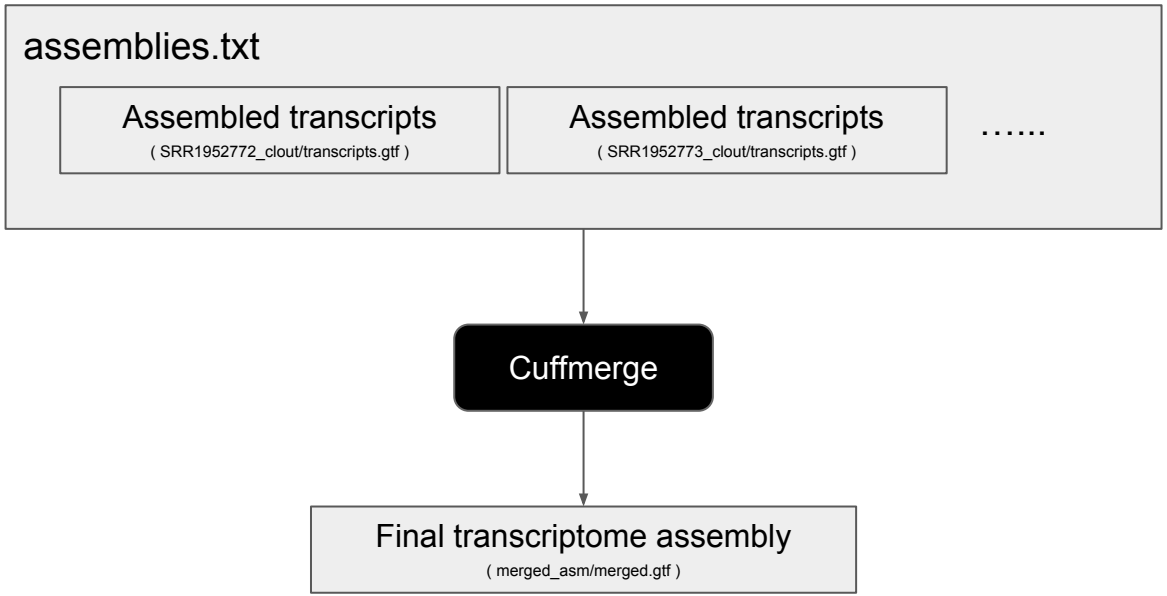
實驗流程

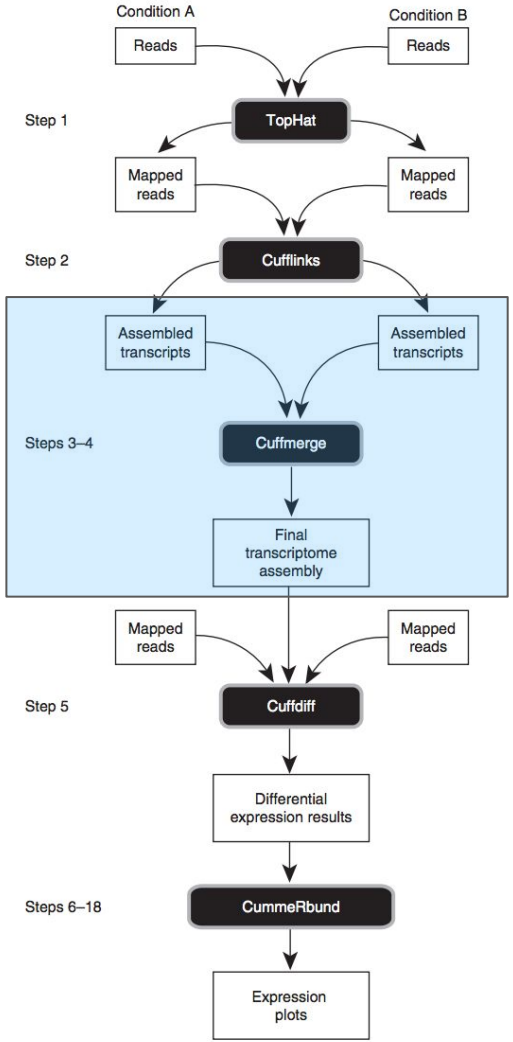




Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

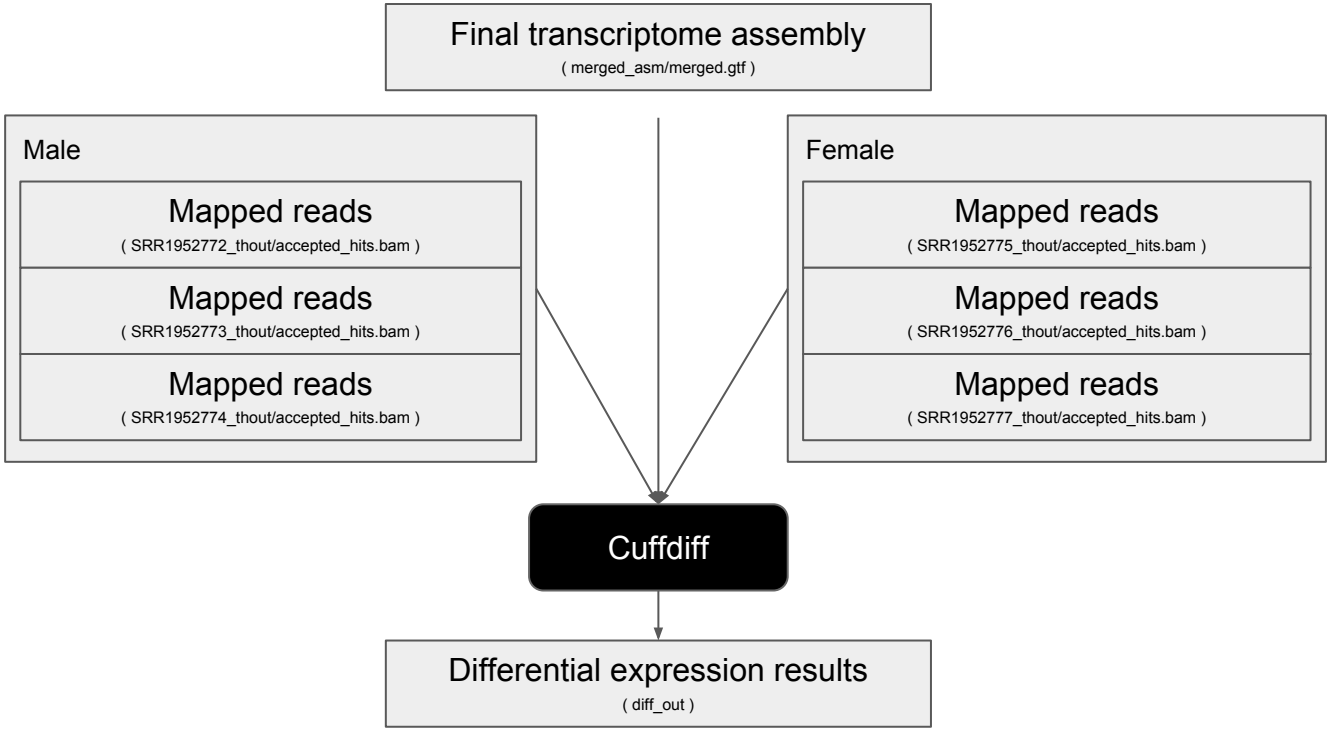
實驗流程

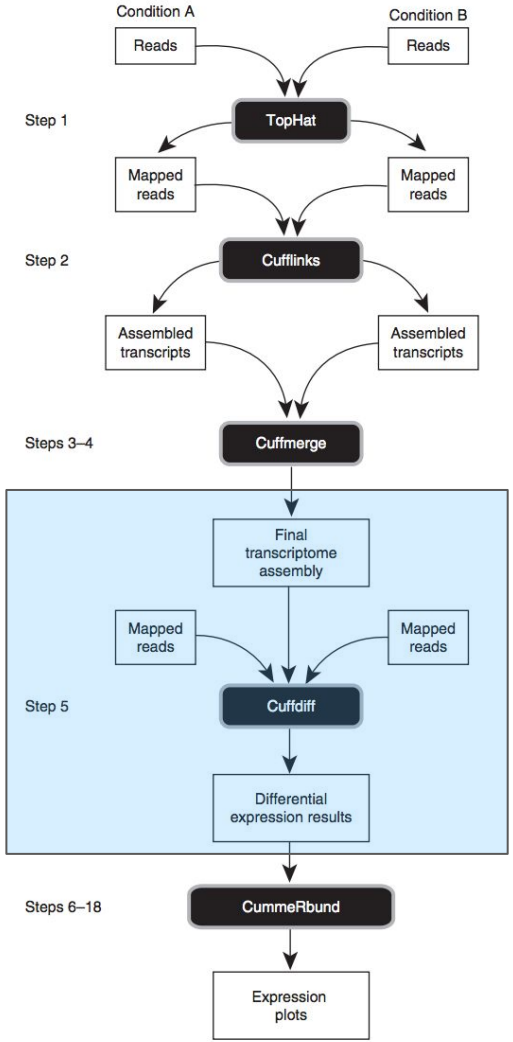




Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

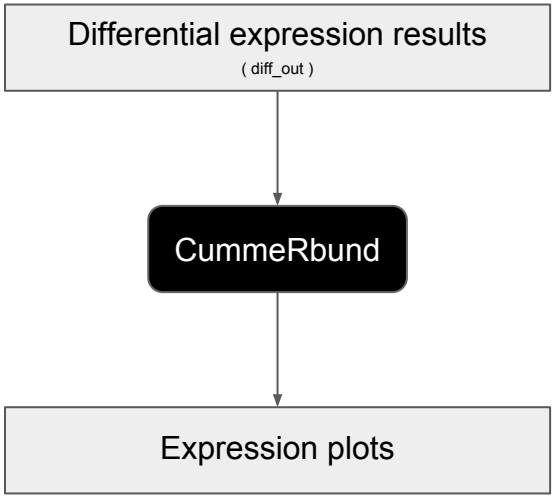
實驗流程

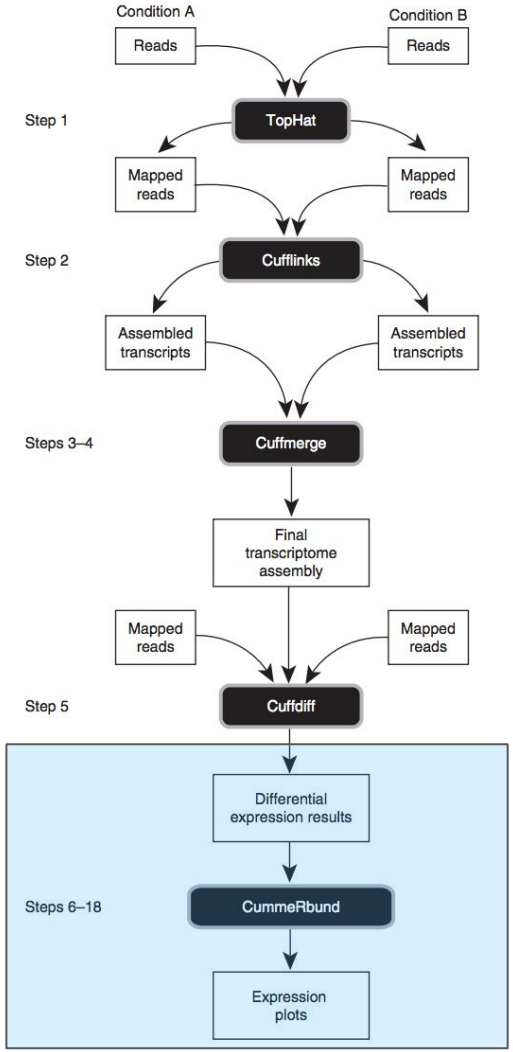




Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

實驗流程





Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

實驗流程

Align the RNA-seq reads to the genome ● TIMING ~6 h

1| Map the reads for each sample to the reference genome:

```
$ tophat -p 8 -G genes.gtf -o C1_R1_thout genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout genome C2_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout genome C2_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout genome C2_R3_1.fq C1_R3_2.fq
```

Assemble expressed genes and transcripts ● TIMING ~6 h

2| Assemble transcripts for each sample:

```
$ cufflinks -p 8 -o C1_R1_clout C1_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R2_clout C1_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R3_clout C1_R3_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R1_clout C2_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R2_clout C2_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R3_clout C2_R3_thout/accepted_hits.bam
```

3| Create a file called assemblies.txt that lists the assembly file for each sample. The file should contain the following lines:

```
./C1_R1_clout/transcripts.gtf
./C2_R2_clout/transcripts.gtf
./C1_R2_clout/transcripts.gtf
./C2_R1_clout/transcripts.gtf
./C1_R3_clout/transcripts.gtf
./C2_R3_clout/transcripts.gtf
```

4| Run Cuffmerge on all your assemblies to create a single merged transcriptome annotation:

```
cuffmerge -g genes.gtf -s genome.fa -p 8 assemblies.txt
```

Identify differentially expressed genes and transcripts ● TIMING ~6 h

5| Run Cuffdiff by using the merged transcriptome assembly along with the BAM files from TopHat for each replicate:

```
$ cuffdiff -o diff_out -b genome.fa -p 8 -L C1,C2 -u merged_asm/merged.gtf \
./C1_R1_thout/accepted_hits.bam,./C1_R2_thout/accepted_hits.bam,./C1_R3_thout/
accepted_hits.bam \
./C2_R1_thout/accepted_hits.bam,./C2_R3_thout/accepted_hits.bam,./C2_R2_thout/
accepted_hits.bam
```

Explore differential analysis results with CummeRbund ● TIMING variable

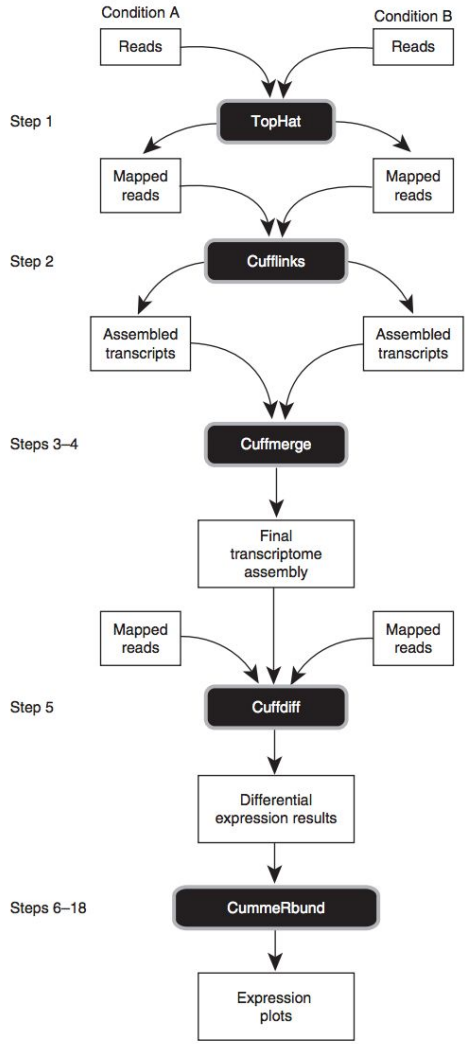
6| Open a new plotting script file in the editor of your choice, or use the R interactive shell:

7| Load the CummeRbund package into the R environment:

```
> library(cummeRbund)
```

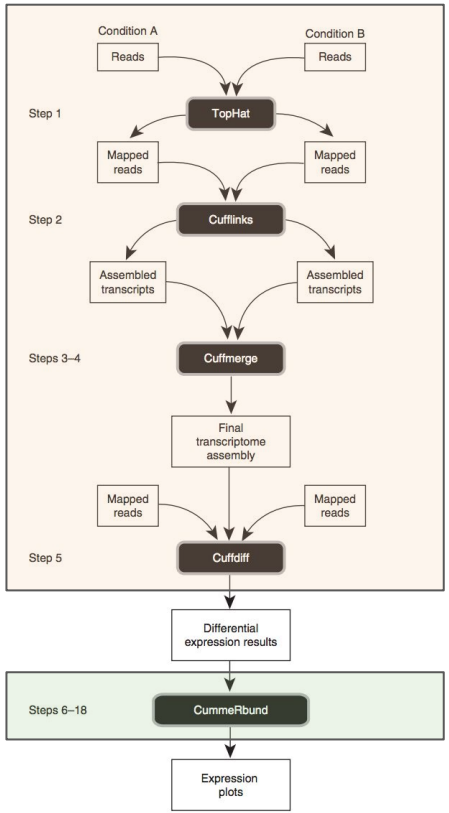
8| Create a CummeRbund database from the Cuffdiff output:

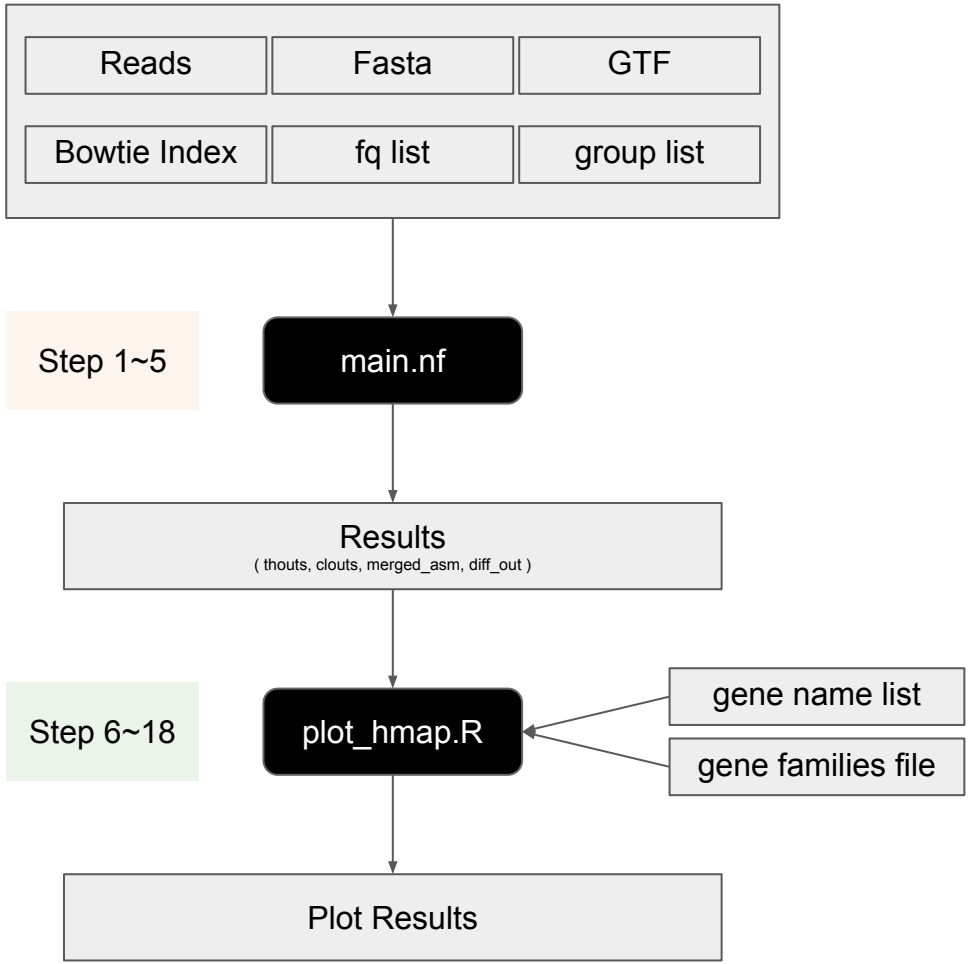
```
> cuff_data <- readCufflinks('diff_out')
```



Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014)

實驗流程





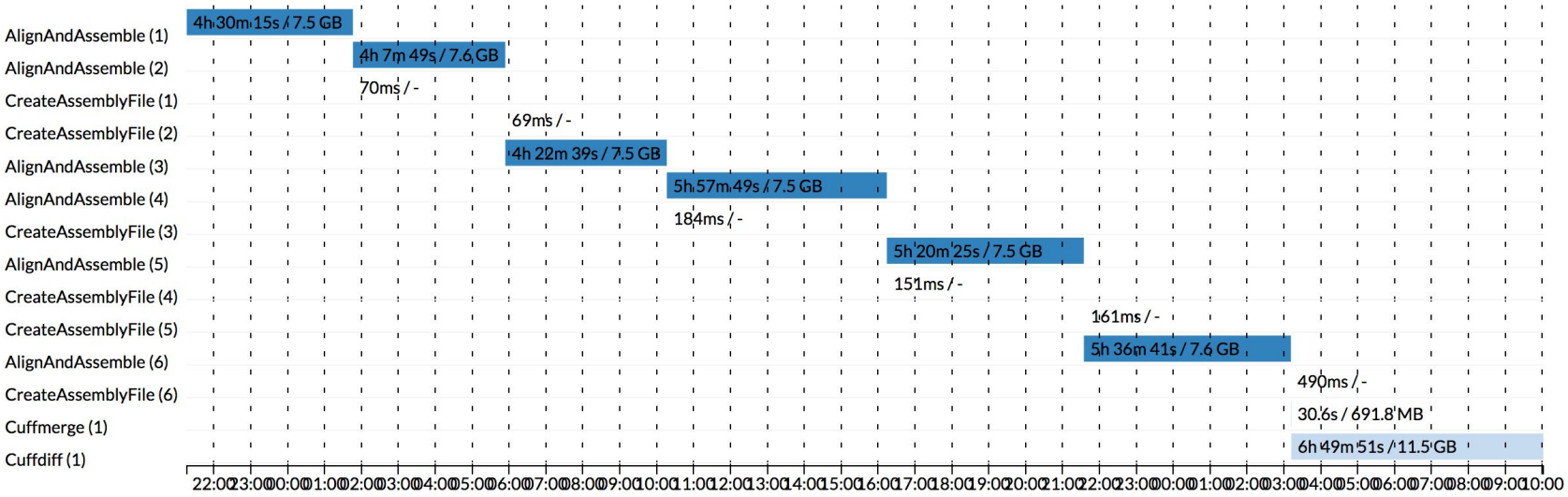
[GitHub BioFinalProject]

實驗流程

GSE67587

Processes execution timeline

Launch time: 08 Jan 2017 21:15
Elapsed time: 1d 12h 46m 3s



Created with Nextflow -- http://nextflow.io

Hardware:

- RAM: 20GB
- CPU: Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz
- Thread: 8 threads

實驗結果

Species	Gender	Run accession	Read pairs (bp)	Mapped Percentage (Left/Right/Pairs)			Multiple Align Percentage (Left/Right/Pairs)			Tophat Execution Times
Drosophila sechellia GSE67587	Male	SRR1952772	31828985	75.9%	74.6%	72.4%	31.2%	30.8%	29.7%	03:04:51
		SRR1952773	29443050	79.2%	78.2%	76.2%	32.3%	31.9%	30.9%	02:45:26
		SRR1952774	30741692	79.6%	78.7%	76.7%	32.5%	32.2%	31.2%	02:53:09
	Female	SRR1952775	36098846	75.8%	74.5%	72.5%	33.5%	33.0%	31.9%	03:42:29
		SRR1952776	33062046	79.3%	78.3%	76.4%	34.7%	34.3%	33.3%	03:13:30
		SRR1952777	34572646	79.6%	78.7%	76.9%	34.9%	34.6%	33.6%	03:28:17

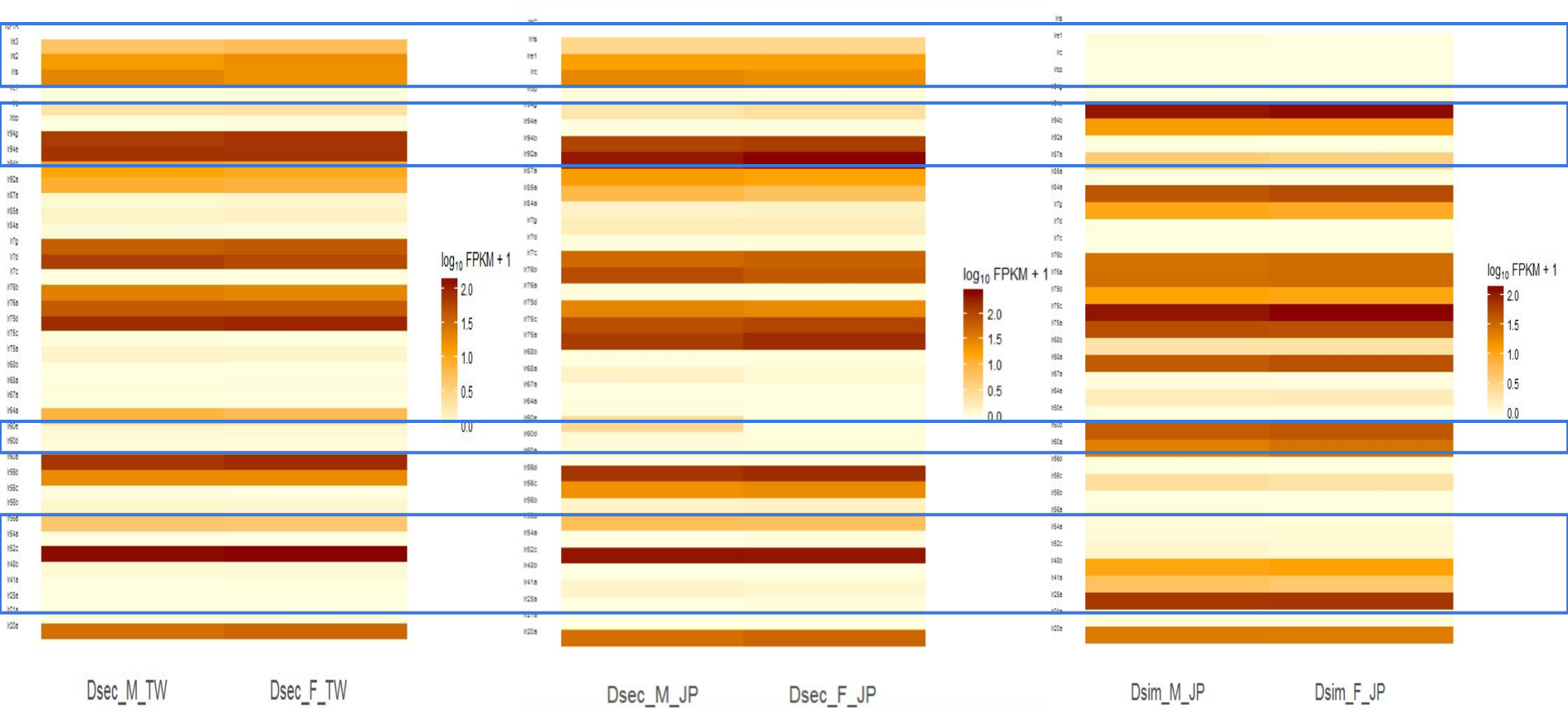
實驗結果

Species	Gender	Run accession	Read pairs (bp)	Mapped Percentage (Left/Right/Pairs)			Multiple Align Percentage (Left/Right/Pairs)			Tophat Execution Times
Drosophila sechellia GSE67861	Male	SRR1973486	40328940	43.2%	37.5%	30.0%	2.7%	2.3%	2.0%	03:13:49
		SRR1973487	37101021	44.5%	39.5%	32.1%	2.8%	2.5%	2.1%	03:12:02
		SRR1973488	38741238	77.6%	76.4%	74.0%	8.1%	7.9%	6.9%	03:21:53
	Female	SRR1973489	29985555	43.9%	37.8%	29.6%	2.8%	2.4%	2.0%	02:25:34
		SRR1973490	28631216	45.2%	39.8%	31.6%	2.9%	2.5%	2.1%	02:33:10
		SRR1973491	29656654	44.9%	40.1%	31.7%	2.8%	2.5%	2.1%	02:38:52

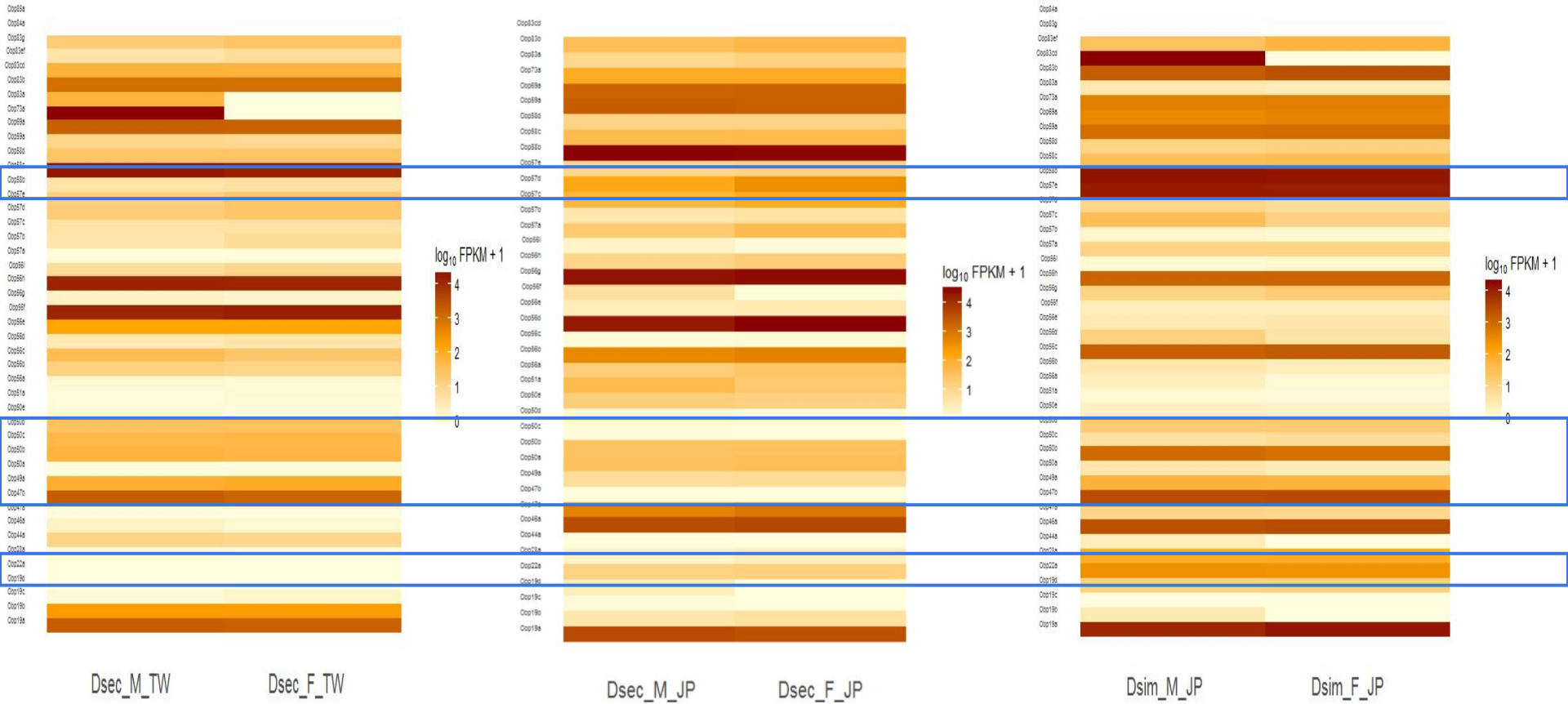
實驗結果

Species	Gender	Run accession	Read pairs (bp)	Mapped Percentage (Left/Right/Pairs)			Multiple Align Percentage (Left/Right/Pairs)			Tophat Execution Times
Drosophila simulans GSE67862	Male	SRR1973492	28225417	61.6%	52.2%	44.4%	0.3%	0.3%	0.2%	03:57:20
		SRR1973493	26008321	63.2%	54.4%	46.9%	0.3%	0.3%	0.2%	03:59:25
		SRR1973494	27138172	62.9%	55.2%	47.4%	0.3%	0.3%	0.2%	03:57:41
	Female	SRR1973495	38340684	64.4%	54.2%	45.7%	0.3%	0.3%	0.2%	05:23:34
		SRR1973496	35245620	66.2%	56.6%	48.5%	0.3%	0.3%	0.2%	04:48:21
		SRR1973497	36830863	65.9%	57.5%	48.9%	0.3%	0.3%	0.2%	05:05:57

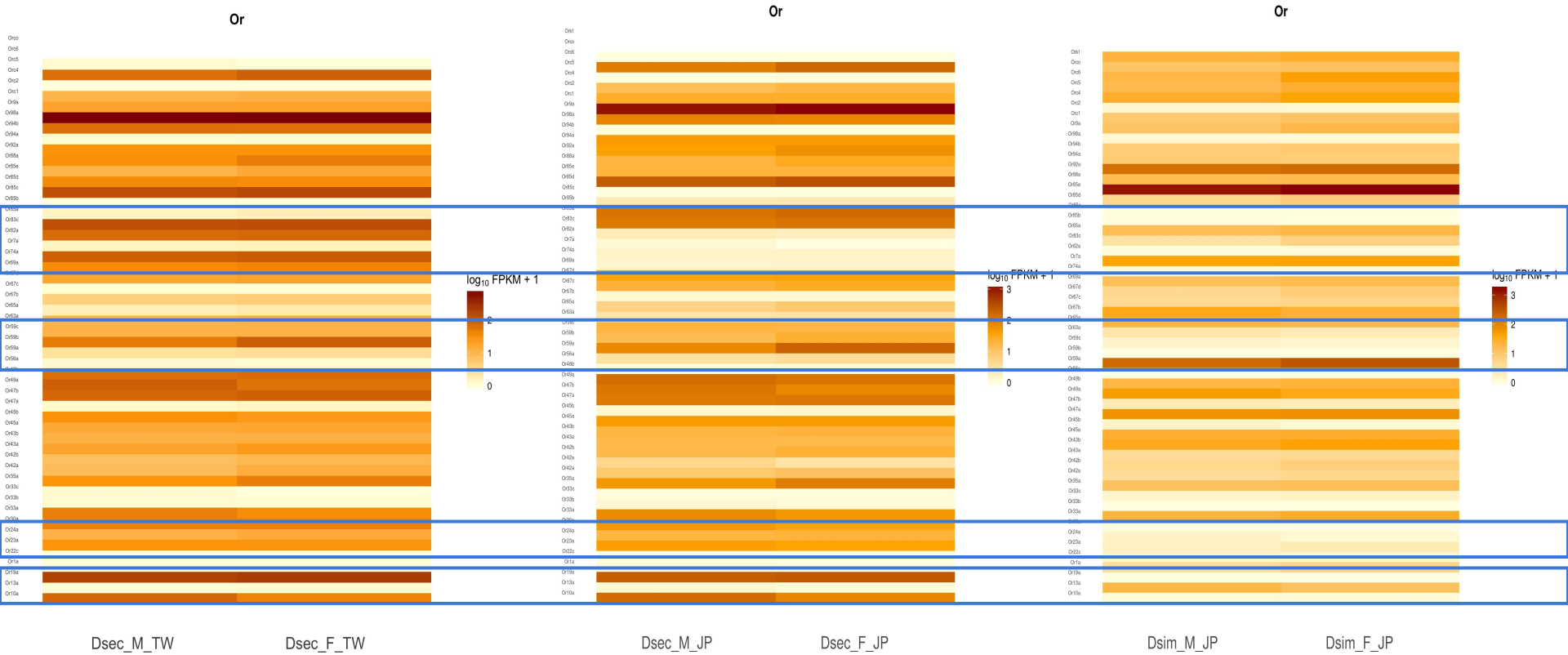
實驗結果 - Ir系列基因


log10 FPKM + 1
2.0
1.5
1.0
0.5
0.0
Dsec_M_TW
Dsec_F_TW
Dsec_M_JP
Dsec_F_JP
Dsim_M_JP
Dsim_F_JP

実験結果 - Obp系列基因



実験結果 - Or系列基因


Or
Or
Or
log10 FPKM + 1
log10 FPKM + 1
log10 FPKM + 1
Dsec_M_TW
Dsec_F_TW
Dsec_M_JP
Dsec_F_JP
Dsim_M_JP
Dsim_F_JP

結論

- 遇到的問題
  a. **fastq 下載速度太慢**：此篇實驗需要的序列數目有18條，每條的檔案大小約10G上下，網速不夠
     - 撰寫scriptvu與分散下載，大家分工將檔案載下來
  b. **fastq 載不下來**：對EBI網站的存取次數過多，站方似乎阻擋下載fastq的要求。
     - 下載SRA，用sra toolkit v2.8.1轉fastq
  c. **流程中的指令**：需要執行不同的程式，且時間長短不一
     - 使用NextFlow產生通用的Script，能依序執行不同的指令
  d. **cuffmerge**：無法在虛擬機執行
     - 將檔案放到不跟host分享的資料夾內
  e. **執行緒意外結束**：tophat可能會沒跑完
     - 檢查align_summary的input 跟 fastq wc -l 的大小/4 要相同
     - wc -l 是文件的統計行數，然後fastq的格式是:
       >
       AAAAAATTTTT.... (sequencing)
       >
       f1ds3f6wr4g..... (quality)
       所以記錄一個read會用到4行，wc -l 除4會是read數量。

參考資料

1. Shiao, M.-S. S. et al. Expression Divergence of Chemosensory Genes between Drosophila sechellia and Its Sibling Species and Its Implications for Host Shift. Genome Biol Evol 7, 2843–58 (2015).
2. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2014).
3. Nature Protocol教學實作:重建轉錄體(Transcript Reconstruction)分析於已知基因體序列(Genome-Based)物種的RNA-Seq概念及分析實作
4. 論文閱讀:Differential gene and transcript expression analysis of RNA-seq experiments with Tophat and Cufflinks
5. Install cummeRbund
6. Install cummeRbund Error
7. Update bioconductor

小組分工

姓名	學號	貢獻
魏孝全	104354025	Drosophila simulans (GSE67862)
劉義瑋	104753013	Nextflow Script、R Script、資料下載、執行、校對、報告
江易倫	104753018	無
李恭儀	105753006	Drosophila sechelia(GSE67861)資料下載、執行、校對、產生heat map、報告
楊宗翰	105753024	Drosophila simulans(GSE67862)資料下載、執行、校對、cummeRbund視覺化程式修改、heatmap、報告
楊子菱	105753041	Drosophila sechelia (GSE67861)資料下載、執行、校對、產生heat map、報告