Transcriptome analysis using RNA-Seq

유전체 빅데이터 분석 예비 전문가 과정
실무 프로젝트 리포트

9기 강 영 광

2021.08.17 ~ 19

실무 프로젝트 분석 방법

제공된 샘플 정보를 토대로 분석 방법을 정리하세요.

레퍼런스 정보

- 종 : Arabidopsis thaliana
- 데이터베이스 : Ensembl (release 44)
- Gene model
  - No. All genes : 888,095 (제공된 GTF 기준)
  - No. Masked genes : 18,747 (제공된 GTF 기준)

RNA-Seq 샘플 정보

Tissue	Treatment	Time point	Sample Name	No. Raw Reads
Shoot	Control	6h	Shoot-Control-6h	16,270,534
Shoot	Control	12h	Shoot-Control-12h	14,775,484
Shoot	Control	24h	Shoot-Control-24h	14,887,260
Shoot	PA01	6h	Shoot-PA01-6h	14,491,286
Shoot	PA01	12h	Shoot-PA01-12h	15,526,964
Shoot	PA01	24h	Shoot-PA01-24h	14,427,486

분석방법

*RNA seq 6개 sample(fastq), genome sequence(fasta), gene model(gtf, mask.gtf) 을 가지고 분석을 시작한다.

00.Raw reads quality check : FastQC를 이용해서 한다. read들의 quality 상태를 확인한다.
1)Adapter trimming : cutadapt (TurSeq mRNA Library Prep Kit)
2)Quality trimming : trimmomatic (주문요청서 조건)
3)Clean reads quality check : FastQC를 한다.

01.TopHat2 – bowtie2를 활용하여 Mapping을 한다.

02.<Cufflinks package - Quant mode>
-Cuffquant (주문요청서 조건)
-Cuffnorm (주문요청서 조건)
-Cuffdiff (주문요청서 조건) 을 실행한다.

03.Cuffnom data table를 통해 Expressed gene selection을 하고,
R studio(vis_exp)를 통해 Scatter plot, Box plot, MDS plot으로 visualization을 한다.

04.Cuffdiff data를 통해 DEG selection을 수행한다.

05.HTSeq-count로 read matrix를 만들고 TCC를 수행한다.
TCC data table를 통해 DEG selection을 하고,
R studio(vis_deg)를 통해 Volcano plot, MA plot으로 visualization을 한다.

06.Cuffdiff DEG selection data와 TCC DEG selection data의 Common DEG를 Venn diagram으로visualization을 한다.

07.Common DEG list를 DAVID를 활용한 Gene Set Enrichment Analysis를 통해 경향성을 파악한다.

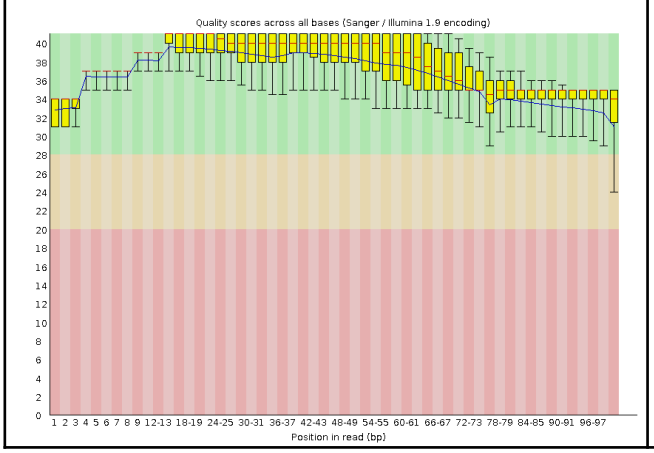
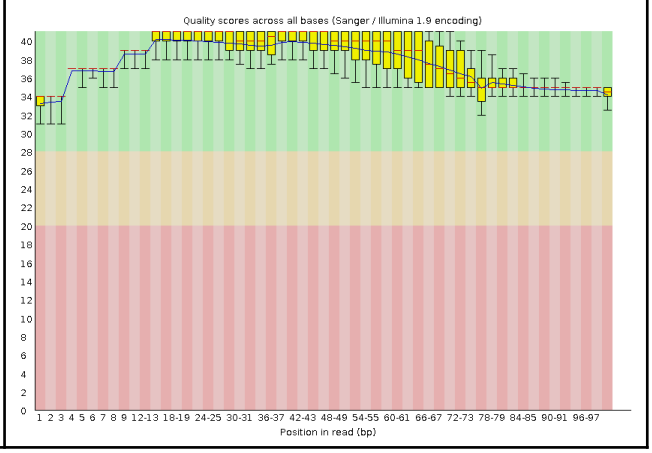
RNASeq Sample Statistics

1. 분석 조건에 맞게 Preprocessing 과 Reference mapping 을 진행하고 결과를 아래 테이블에 정리하세요

Sample name	Raw reads	Clean reads	Clean Reads %	Mapped reads	Mapped reads %
Shoot-Control-6h	16,270,534	13,806,244	84.85	13,664,174	83.98
Shoot-Control-12h	14,775,484	12,632,712	85.50	12,491,556	84.54
Shoot-Control-24h	14,887,260	12,666,664	85.08	12,538,400	84.22
Shoot-PA01-6h	14,491,286	12,312,722	84.97	12,181,360	84.06
Shoot-PA01-12h	15,526,964	13,317,106	85.77	13,178,521	84.88
Shoot-PA01-24h	14,427,486	12,290,412	85.19	12,154,444	84.25

- Raw reads : count all reads (R1+R2) [Hints : fastqc Total Sequences, cutadapt log]
- Clean reads : count all reads (R1+R2) after adapter trimming and quality trimming [Hints : fastqc, trimmomatic log]
- Clean reads % : (Clean reads/Raw reads) * 100, 소수점 2자리
- Mapped reads : count all mapped reads (Left reads + Right reads) [Hints : Tophat align_summary.txt]
- Mapped reads % : (Mapped reads/Raw reads) * 100, 소수점 2자리

2. Shoot-Control-6h 샘플에서 Preprocessing 전과 후의 fastqc (Per base sequence quality) 를 비교하세요

Before preprocessing	After preprocessing
	

3. 위 결과들을 토대로 Preprocessing 의 관점에서 데이터를 해석하세요.

- Fastqc(Per base sequence quality)는 각 서열에서 base 당 quality 값을 boxplot으로 보여준다. Preprocessing과정을 통해 adapter trimming(cutadapt)과 Quality trimming(Trimmomatic) 전후의 결과를 비교해보면, trimming 이후에 y축의 Q(Pred score)값을 높여 quality가 좋아진 것을 확인 할 수 있다. 그러므로, Quality가 더 높은 clean reads을 얻게 되었음을 확인할 수 있다.

Expression analysis

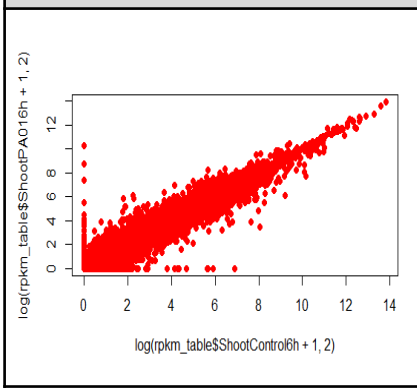
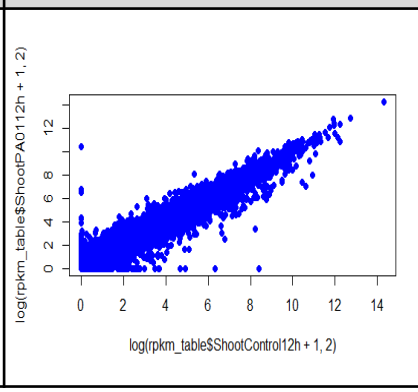
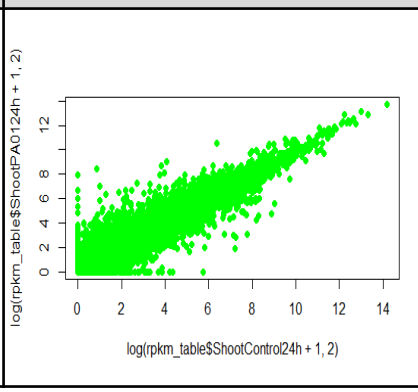
4. Cuffnorm 결과를 토대로 6 샘플의 RPKM 테이블을 제작하고, 각 조건별 expressed gene 의 수를 확인하고 어떤 의미를 나타내는지 설명하세요.

	RPKM >= 0.3	RPKM >= 1	RPKM >= 10	RPKM >= 100
expressed gene	20,893	18,984	11,725	1,837

- expressed gene : 적어도 한 샘플에서 RPKM 발현량 기준을 만족하는 유전자의 수

- RPKM은 상대 값을 이용해 발현 양을 표시하는 normalization method 중 하나이다. 내가 확인한 유전자의 길이를 정규화 했을 때, 전체 reads 된 것 중 해당 유전자에만 mapping된 reads를 말한다. 또한 RPKM은 실제 유전자의 발현 유무를 판단할 수 있는 기준으로 활용되며, RPKM값 0.3이상을 갖는 유전자들은 in vivo에서 실제로 발현한다고 판단 할 수 있다.
  그러므로, 한 유전자에서 6 샘플 RPKM 값 중 max(최대값)을 찾고, max값이 RPKM 기준값(0.3,1,10,100)과 비교할 때, expressed gene의 수만큼 실제로 유전자가 발현한다는 의미를 갖는다.

5. 같은 시간대에서 Control 샘플과 Case 샘플의 발현량을 Scatter plot 으로 비교하고 어떤 의미가 있는지 설명하세요.

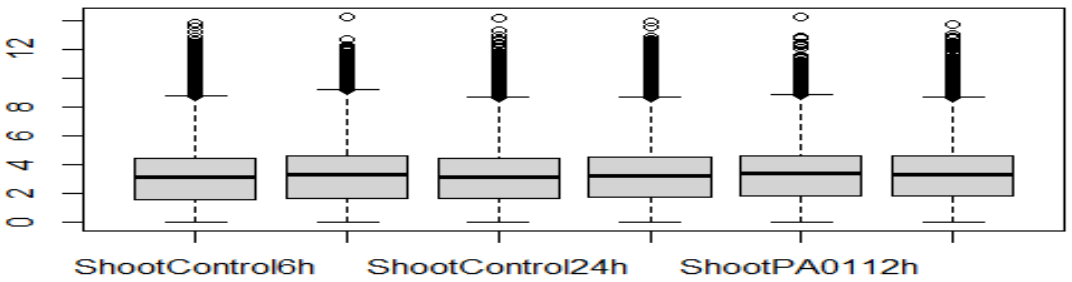
6 시간	12 시간	24 시간
		

- 6h,12h,24h Scatter plot의 spot들 분포의 넓이가 넓어지는 것을 볼 때, 6h 보다 12h후의 두 sample(control, case)간의 발현 분포의 차이가 커진다라는 의미를 갖는다. 그래서 PA01을 처리했을 때, Arabidopsis Shoot sample에서 발현의 변화가 있음을 의미한다.

Expression analysis

6. 적어도 한 샘플에서 RPKM 이 0.3 이상인 유전자들을 대상으로 Boxplot 을 그리고 어떤 의미가 있는지 설명하시오

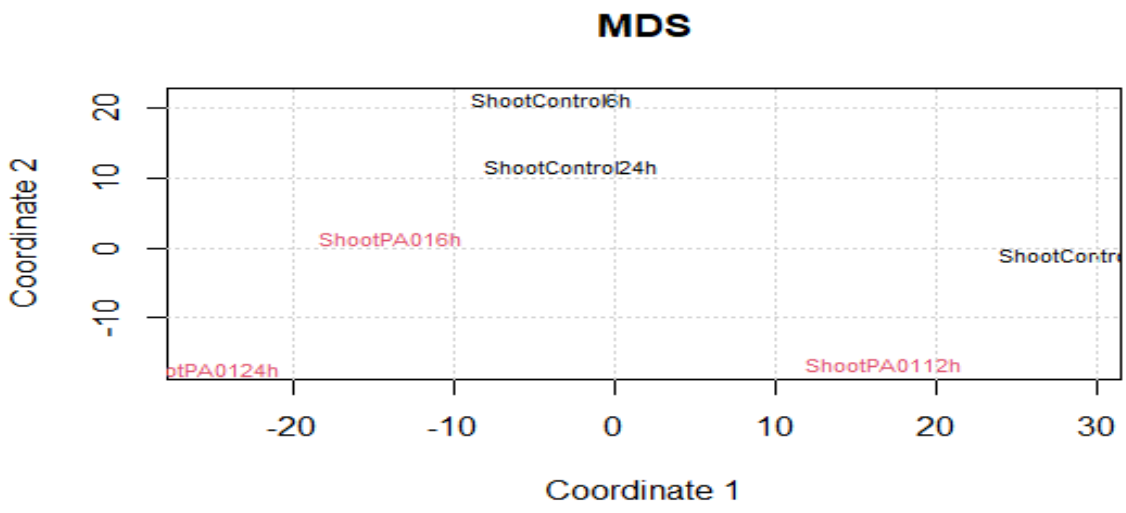
Box plot



- 6 sample들의 boxplot을 볼때, box가 균등한 것을 보아 6 sample data들의 normalization이 잘 되었음을 의미한다. 만약 normalization 잘 안되었을 경우, 개중의 한 box가 튀어 올라와 있다면 다른 normalization 방법(TPM)을 써야하는지 고려를 해봐야한다. Tuxedo protocol 자체가 내 data에 맞지 않는 분석 방법일 수 있다.

7. 적어도 한 샘플에서 RPKM 이 0.3 이상인 유전자들을 대상으로 MDS 를 그리고 어떤 의미가 있는지 설명하시오.

MDS



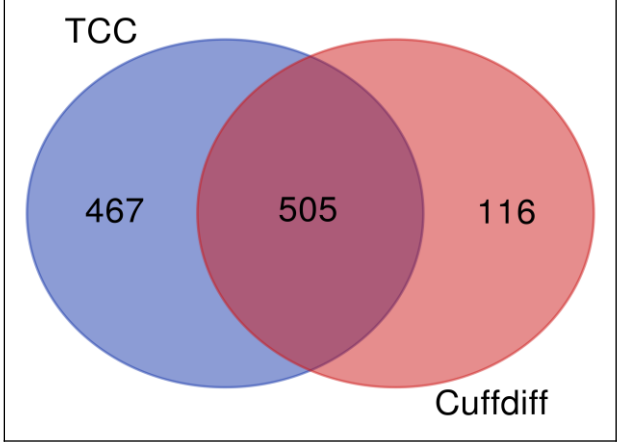
- clustering 분석의 한 종류로 Control 간의 거리는 가까워야 하고, Control과 Case의 발현 차이는 클 것이므로 거리가 멀 것이라고 예상할 수 있다. 그래서 sample간의 발현 값을 가지고 거리를 계산하고 그 거리를 시각화하여 sample간 얼마나 유사성을 갖는지 확인할 수 있다. 그러므로, 위의 결과는 Control 과 Case의 sample간의 거리가 있으므로, 분석이 잘 되었고 분석 발현 양을 잘 얻었다는 것을 확인할 수 있다.

DEG analysis

8. 분석 조건에 맞게 CUFFDIFF와 TCC 를 사용한뒤, 각 분석 도구 간의 DEG 양상을 venn diagram 으로 확인하세요. (참고 : http://bioinformatics.psb.ugent.be/webtools/Venn)

	DEG	UP	DOWN
TCC	972	595	377
CUFFDIFF	621	417	204

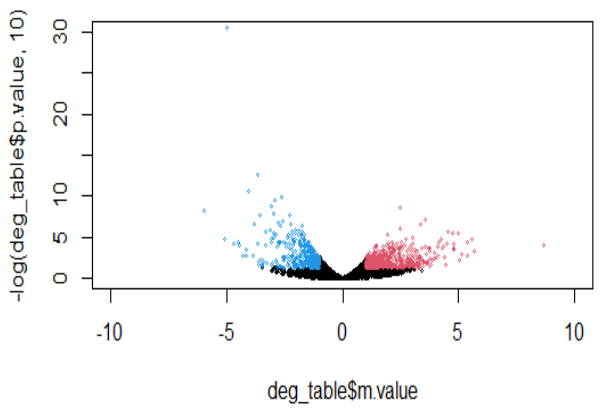
Venn diagram



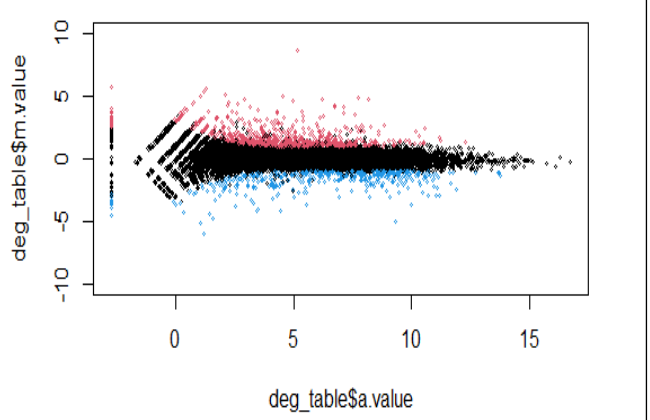
- TCC_DEG 972개/ Cuffdiff_DEG 621개 공통 DEG 505개를 venn diagram으로 확인할 수 있다.
- 차후 common DEG를 활용하여 GESA 분석을 통해 경향성을 분석한다.

9. TCC 분석 결과를 토대로 Volcano plot 과 MA plot 을 그리고 DEG 분포에 대해 설명하세요

Volcano plot



MA plot



- UP regulator DEG(빨강), DOWN regulator DEG (파랑), House keeping gene(검정)
- Volcano plot은 X축이 두 group간 발현의 차이를 나타내는 것이고, Y축은 –log(p-value)값으로 유의하면 유의할수록, y값이 커진다. 그러므로, Volcano plot에서 상단에 있는 Up regulated된 DEG, p-value가 유의한 plot 상단에 찍힌 DEG유전자들은 DEG 발현 값의 차이도 크고, 유의한 DEG들이 가장 먼저 찾아볼 DEG들이 된다.
- MA plot은 X축 두 group간의 발현의 평균(A.value), Y축 두 group간의 발현의 차이(P.value)이다. Y축의 0을 기준으로 양수일 때, control 대비 case에서 up regulation된 DEG이고, 0을 기준으로 음수일 때, control 대비 case에서 down regulation된 DEG들이 분포하고 있음을 의미한다.

Gene-Set Enrichment Analysis

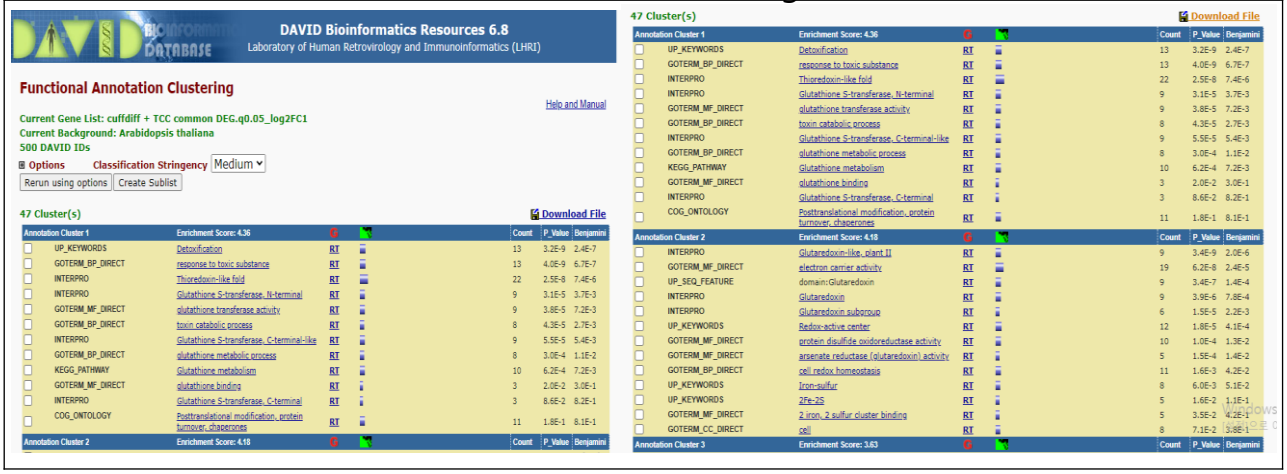
10. CUFFDIFF 와 TCC 에서 확인된 common DEG 를 대상으로 DAVID 를 이용한 Gene set Enrichment Analysis 수행하고 결과를 분석 디자인에 맞게 해석하세요.

참고 : https://david.ncifcrf.gov/summary.jsp

- Input genes : Common DEG

- **Annotation Clustering)**

Annotation clustering

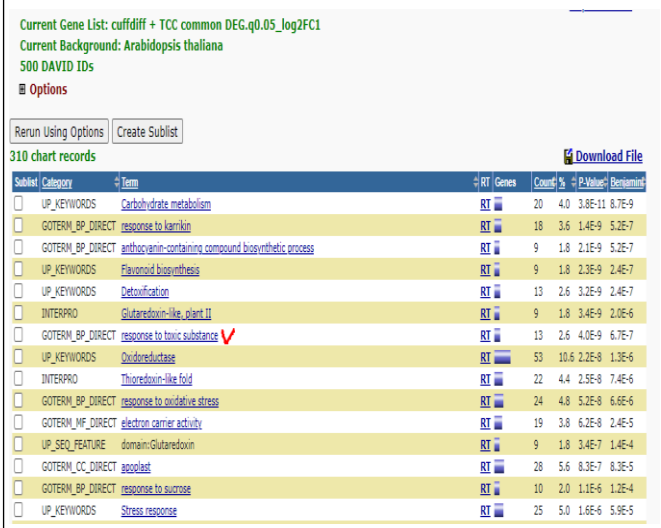
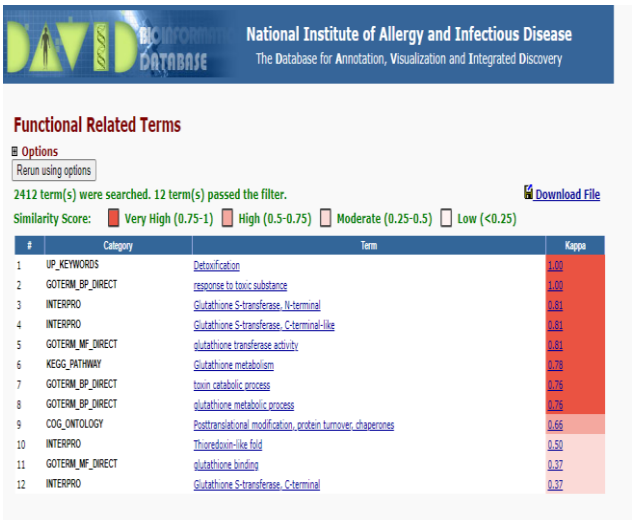


- 설명: commen DEG에서 얻어진 gene-set마다 Enrichment Score 값과 통계량(P-value, Benjamin)을 확인할수 있고, 유사한 term들 끼리 clustering 되어있다.
- common DEG 500개 gene list를 놓고 DAVID를 통한 GSEA분석을 했을 때, 유의한 term들의 cluster가 47개 확인된다.
- 그 중에 Enrichment Score값이 큰 것들이 하나의 cluste1을 보거나, Biological process 에 속하는 cluster에 있는 term들의 유의한 생물학적인 경향성을 파악하는게 좋다.

=> 8 페이지 annotation chart!

Gene-Set Enrichment Analysis

- **Annotation Chart)**

response to toxic substance	Related Terms
	

- 설명: GOTERM_BP_DIRECT의 response to toxic substance라는 Gene-set 안에 13개 gene, P-value, Benjamin(사후통계량)을 확인할 수 있다.
- common DEG 500개 gene list를 놓고 DAVID를 통한 GSEA분석을 했을 때, term들의 chart가 310개 확인된다.
- 우측 사진) response to toxic substance(gene-set) 관련된 terms들이 나온다. 또한, 내가 넣어준 common DEG들과 관련된 terms들을 알 수 있다.