암유전체 데이터의 다중분석을 위한 SOP

(Standard operating protocols for integrated analyses of multidimensional cancer genomics data)

목차

1. 준비사항

(1) 배경

- 본 SOP 의 대상이 되는 다중유전체 데이터는 다수의 환자군 및 샘플에서 2 개 이상의 유전분석플랫폼(genotyping platform)에서 얻어진 데이터로 본 SOP 는 데이터 간 공통되는 환자 및 샘플에 대한 데이터교차분석에 대한 내용임. 본 SOP 의 분석은 마이크로어레이 및 시퀀싱에서 얻어지는 유전데이터를 대상으로 하고 있으나, 기타 임상의료데이터로 얻어질 수 있는 임상-병리데이터 등 교차분석의 대상이 되는 데이터를 포함할 수 있음.

- 본 SOP 에서는 유전자발현 데이터 내의 각 유전자간의 발현의 교차분석을 위한 가장 대중화된 소프트웨어 중 하나인 ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks; url: wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE) [1]의 사용법을 기술함. 해당 소프트웨어의 사용 및 응용을 통해 현재 암유전체분석의 다중분석에서 일반적으로 고려되는 mRNA-microRNA 발현 간의 교차분석, mRNA-유전자카피수 간의 교차분석, mRNA 발현 및 유전자메칠화정도 간의 교차분석이 가능함.

- 본 SOP 에서는 가장 대표적인 알고리즘인 ARACNe 와 iCluster2 [2]를 기준으로 설명하고 있으나, 다중 유전체 데이터 분석을 위한 많은 다른 알고리즘이 존재하므로, 다중 유전체 데이터 분석 시 여러 다른 알고리즘를 수행하고, 각각에서 나온 결과들을 통합 비교 분석하는 것이 필요할 수 있음. 또한 연구 목적에 따라 PARADIGM [3], moCluster [4]등 다른 방법론의 사용도 고려해볼 필요가 있음.

- 본 표준 프로토콜은 데이터의 종류, 목적, 수행시기에 따라 많은 변동사항이 있을 수 있으므로, 작성된 프로토콜은 모든 연구를 대변하는 방법론이 될 수는 없음

- 본 프로토콜은 2016 년 10 월 현재를 기준으로 작성된 것으로 추후 알고리즘 버전 업데이트 등에 의해 프로토콜이 수정되어야할 수 있으며, 보다 좋은 성능을 보이는 새로운 알고리즘이 개발되는 등의 경우에는 프로토콜이 변경될 수 있음.

(2) 다중분석을 위한 데이터종류 - #1. mRNA 및 miRNA 발현 데이터

- 암유전체분석의 다중분석으로 고려될 수 있는 mRNA-microRNA 발현 간의 교차분석, mRNA-유전자카피수 간의 교차분석, mRNA 발현 및 유전자메칠화정도 간의 교차분석 등의 분석을 위해 각 데이터 세트를 준비하여야 함.

- mRNA 발현의 경우 유전자발현 microarray 및 RNAseq 기반의 전사체분석을 통해 얻어질 수 있음. 유전자발현 microarray 의 경우 현재 Affymetrix, Agilent, Illumina 등의 다양한 플랫폼이 존재하고 있으며, 그 다양성 및 분석의 복잡성으로 인해 실제 microarray image 의 low-level 분석에 관한 사항은 각 플랫폼의 생산자가 제작한 프로토콜을 따르는 것을 함.

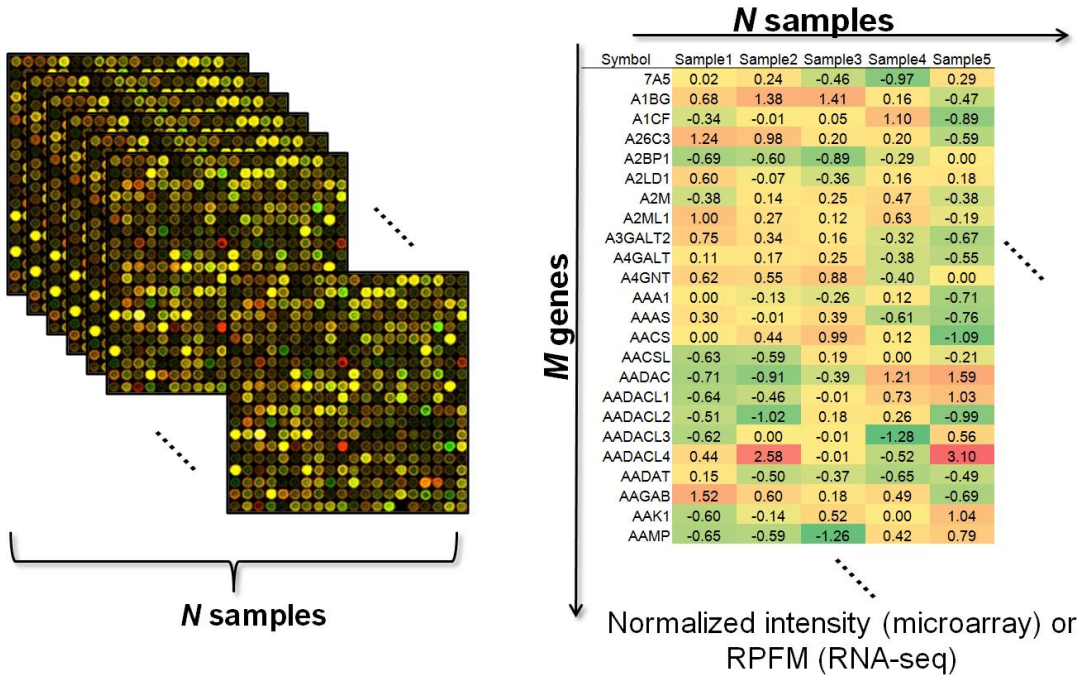


그림 1. 일반적인 mRNA expression profile 의 형태. Raw 상태의 image data 를 processing 하여 공통의 유전자군에 대응하는 normalized value 가 N 개의 샘플수에 매칭되는 2D 형태의 profile 을 일반적으로 사용함.

- mRNA 마이크로어레이의 경우 one-dye (Affymetrix, Illumina)혹은 two-dye 플랫폼(Agilent)으로 구분되며 one-dye 의 경우 개별 tumor 기반의 mRNA, two-dye 의 경우 tumor/normal mRNA 의 상대정량에 따라 각 유전자의 발현양상이 결정됨. 본 SOP 에서는 N 개의 유전자 X M 개의 샘플에 대한 유전자 발현량을 2D 매트릭스로 존재한다고 가정하고 이를 다중분석의 대상으로 함. 필요시 추가적인 quantile normalization 이 수행될 수 있음. 일반적인 2D 형태의 mRNA expression profile 은 그림 1 과 같음.

- miRNA 를 대상으로 하는 분석의 경우도 mRNA 마이크로어레이와 같은 방식으로 분석 데이터 세트를 준비함. N 개의 microRNA X M 개의 샘플에 대한 miRNA 의 상대적인 발현양을 2D matrix 로 준비함. 보통 20,000 개 이상의 mRNA 를 대상으로 하는 mRNA 분석과 달리 miRNA 의 경우, 1000 개 내외 갯수의 miRNA 를 다룸. 최근 PanCan 규모의 분석에서 제시되었듯이, miRNA 와

mRNA 의 발현량의 연관분석에서 종양에서 나타나는 다양한 유전체 변화(DNA 카피수 변화 및 메칠화)를 고려하는 선형회귀모델이 중요하다는 점이 제시된 바 있음 [5].

(3) 다중분석을 위한 데이터종류 - #2. DNA 카피수변화 프로파일

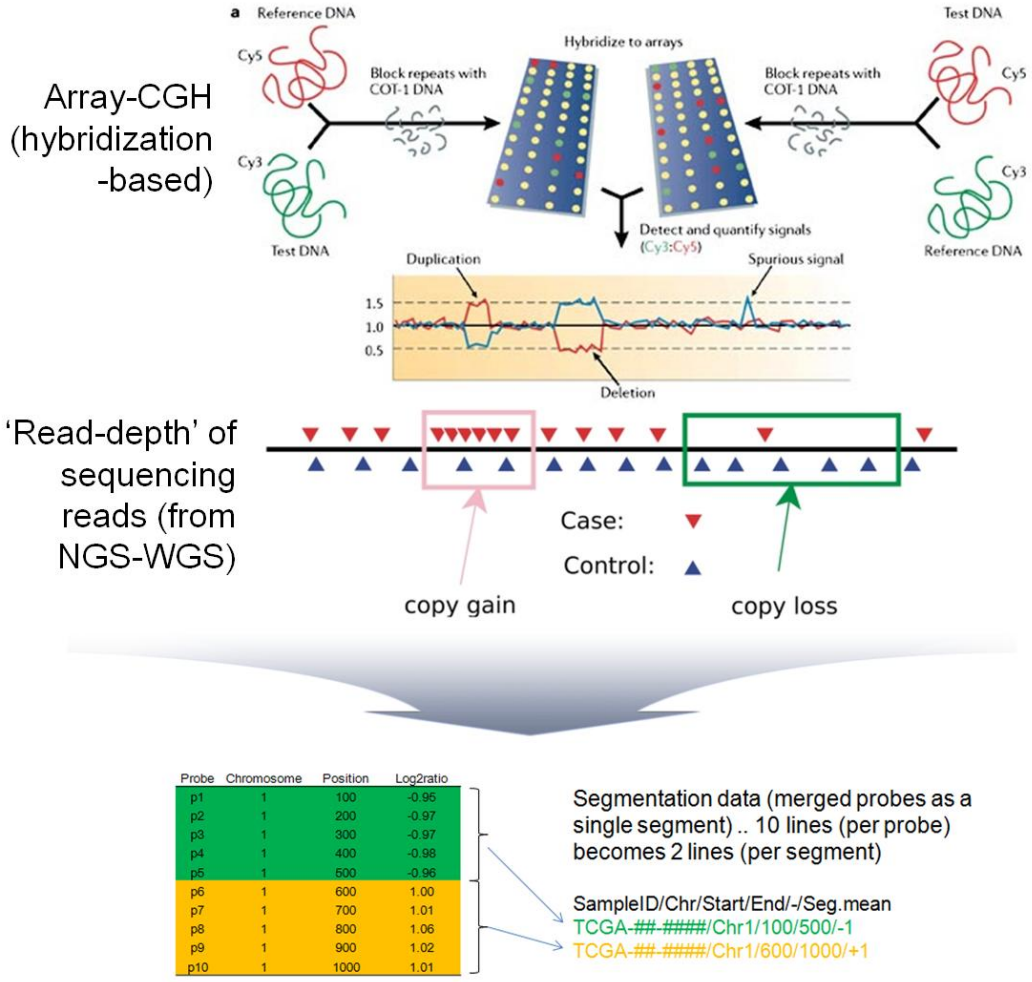


그림 2. 마이크로어레이기반의 array-CGH 및 NGS 기반의 read-depth 방식으로 추출된 카피수 프로파일은 segmentation 의 과정을 거쳐 *seg 형태로 저장되며, IGV browser 및 GISTIC 알고리즘을 통해 각각 visualization 및 driver 추출의 분석에 이용됨.

- DNA 카피수 변화 프로파일은 역시 시퀀싱 및 마이크로어레이 기반으로 얻어짐. 시퀀싱 기반의 카피수 변화는 보통 read-depth 기반(VarScan [6], Bic-seq [7] 등)의 tumor/normal ratio 를 genomic bin 에서 계산하며, 마이크로어레이의 경우 단일 probe 수준의 tumor/normal ratio 을 계산하게 됨. 두 플랫폼 모두 추가적인 smoothing 및 segmention 을 통해서 DNA 카피수변화프로파일의 표준 포맷인 .seg 형태를 얻게 됨. Microarray (array-CGH) 및 NGS 기반의 카피넘버 프로파일링의 개식도는 그림 2 와 같음.

- Smoothing/segmentation 에서 현재 가장 표준적으로 이용되는 알고리즘은 CBS (circular binary segmention) [8]임. 그 외, Affymetrix 사의 SNP6.0 등의 상대적으로 noisy 한 프로파일의 경우 GLAD [9]등의 알고리즘이 이용될 수 있음.

- mRNA 프로파일과 다른 점은 DNA 카피수 변화의 경우 반드시 matched normal 을 고려해야 함. 즉, tumor/normal ratio 를 구하기 위해 시퀀싱의 경우 matched normal 시퀀싱이 수행되어야 하며, one-dye array 인 Affymetrix SNP 6.0 의 경우 tumor 와 독립적으로 matched normal 의 genotyping 이 수행되어야 함. Two-dye array 인 경우는 이미 tumor/matched normal 이 고려되며 universal reference DNA 를 사용하는 것보다 해당 환자의 matched normal 을 counterpart dye 로 염색해서 사용해야 하는 것이 바람직함. 해당 환자의 matched normal 이 사용되지 않거나 two-dye chip 에서 universal reference 가 사용되는 경우 상당수의 germline CNV (copy number variation)이 같이 발굴되기 때문에 정확한 체성(somatic) CNA 프로파일을 얻기 힘들고 추후 분석에 문제를 유발할 수 있음.

- 최종적으로 얻어진 .seg 파일은 고려되는 모든 샘플에 대해 카피수 변화가 발굴된 유전체 segment 별로 log2 ratio(tumor/normal)을 알 수 있게 되며 다중분석을 위한 표준유전자세트의 유전체 coordinate 에 따라 샘플별 X 유전자별 CN 값을 매칭하게 된다. mRNA 유전자발현과 DNA 카피수 변화의

교차분석을 위해 보통 mRNA 유전자 발현량이 주어진 유전자세트(mRNA profile)를 구하고 유전자 별로 DNA 분석에 이용된 유전체버전(hg19, hg38 등)에 따른 genomic coordinate 를 (transcription start-end) 획득한 후, segment 파일의 유전자 위치정보와 교차분석하여 유전자별 해당 샘플의 카피수를 구하고 N X M 2D 매트릭스를 생성함(유전자 카피수 프로파일).

- 현재 표준 포맷으로 사용되고 있는 *.seg 형태의 경우 단일 segment 로 인지되는 유전체영역(염색체 및 시작-끝점으로 나타냄)의 평균 카피수(log ratio)를 나타내는 형식임. 이 평균 카피수는 마이크로어레이의 경우 해당 영역에 존재하는 모든 probe 의 log ratio 의 평균이며 시퀀싱 데이터는 해당 영역에 존재하는 genomic bin 혹은 해당영역의 존재하는 tumor/normal 시퀀싱 리드 수의 log 값임.

- 실제 세포주가 아닌 경우 외과적 수술로 얻어지는 암종의 경우 어느 정도의 정상세포침윤이 있고(normal contamination) 이로 인해 수%에서 수십%에 해당하는 tumor purity 를 갖게 됨. 보통 이러한 tumor purity 는 병리학자에 의한 슬라이드조직검사에 의해 대략적으로 측정되며 보통 유전체분석에 사용되는 조직은 70%이상의 tumor purity 가 확보되어야 함.

- 최근, 시퀀싱 및 마이크로어레이에서 얻어진 전장유전체규모의 DNA 카피수 프로파일을 이용해서 tumor purity 및 tumor ploidy 를 예측하고 이를 이용하여 log2 수준의 DNA 카피수를 absolute 수준(CN = 1, 2, 3, ...)으로 변환하는 알고리즘이 제시되었고(ABSOLUTE, https://www.broadinstitute.org/cancer/cga/absolute) [10], 이를 이용할 경우 기존 GISTIC [11] 등의 알고리즘으로 간접적으로 예측하던 absolute CN 을 유전자별로 매칭할 수 있음. 단, ABSOLUTE 를 사용하기 위해서는 어느 정도의 CNA 가 존재하여야 하며 CNA 가 비교적 없는 혈액기원의 종양이나 MSI(+)종양의 경우 ABSOLUTE 를 사용할 수 없음. 또한 DNA 카피수 변화가

아닌 유전자발현프로파일을 이용하여 간접적으로 tumor purity 를 계산하는 알고리즘(ESTIMATE) [12]가 제시된 바 있음.

(4) 다중분석을 위한 데이터종류 - #3. DNA 메칠화 프로파일

- DNA, 특히 프로모터 등의 유전자의 조절부위에 위치한 CpG dinucleotide 의 메칠화 양상 또한 마이크로어레이 및 차세대시퀀싱으로 프로파일이 가능하며, 암유전체의 epigenetic 변화를 프로파일링하는 데 있어 활발히 이용되고 있음. Epigenetics 변화의 다른 측면인 histone 변화의 경우, 다양한 종류의 histone modification 및 에피게놈 변화를 측정할 수 있는 Chip-seq 이 연구 단계에 있으나, 암유전체분석에서는 아직 활발히 이용되고 있지 않음.

- CpG decay 에 의해 CpG dinucleotide 는 expected value 에 비해 observed value 가 낮으며(observed/expected = 1/5) 유전자조절부위와 밀접한 연관을 갖는 CpG island 등에서 주로 관찰됨. 대략, 4 천만개에 이르는 CpG dinucleotide 중 대표성을 갖는 일부의 CpG dicleotide 를 probe 화 한 DNA methylation microarray 가 대표적으로 이용되고 있으나 최근 전장유전체규모의 methylation 분석(whole-genome bisulfite sequencing)이나 methylation domain 을 ChIP (chromatin immunoprecipiation)으로 분석하는 기법 및 CpG capture sequencing 등의 차세대시퀀싱 기반의 메칠화 분석기법이 이용되고 있음.

- 특정 CpG dicleotide 의 메칠화는 유전자발현양상과는 다르게 해당 CpG dinucleotide 의 메칠화 정도(0-100%)를 beta value 로 환산하여 사용하게 되며 이로 인해 수십-수천배의 fold change 를 가지는 유전자발현량과는 달라 제한된 변화폭을 가짐. 보통 유전자발현과의 상호분석을 위해 해당유전자의

mRNA 발현량과 해당유전자의 프로모터영역/CpG island 의 메칠화정도의 연관도를 계산하게 됨.

2. 다중유전체데이터 연관분석 - ARACNe

(1) ARACNe 소개

- ARACNe (algorithm for the reconstruction of accurate cellular networks)는 B cell co-expression network 을 추출하는 수단으로 처음 보고되어(Nat Genet 4:382, 2005) [13] 악성신경교종의 핵심전사인자를 찾는 등 (Nature 463:328, 2010) [14]의 다양한 목적으로 이용되고 있음.

- ARACNe 의 경우 MI (mutual information)을 distance measure 로 이용하며 전사조절인자(transcription factor)와 target 간의 직접/간접(direct/indirect)관계를 유추할 수 있는 DPI (data processing inequality)를 도입함. MI 의 경우 non-linear setting 에서 statistical correlation 을 결정할 수 있는 추정치로서 통상적으로 이용되는 연관도(Pearson or Spearman correlation)에 비해 장점을 가지고 있음.

- 통상적으로 100 개(샘플/환자) 이상의 유전자발현프로파일을 이용하는 것이 좋으며 안정적인 MI 를 계산할 수 있는 최소한의 숫자로 정함.

(2) ARACNE 설치 및 실행

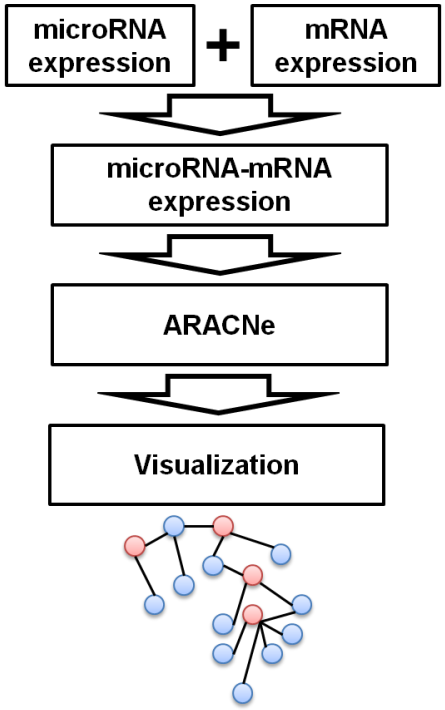


그림 3. ARACNe 를 이용한 microRNA, mRNA 통합 분석 개요.

- 그림 3 은 ARACNe 를 이용한 microRNA 와 mRNA 통합 분석방법에 대한 개요를 보임.

- ARANCe 를 설치하기 위해 권장되는 기본 하드웨어 및 소프트웨어 조합은 다음과 같음.

  - ARACNE (http://amdec-bioinfo.cu-genome.org/html/caWorkBench/upload/aracne.zip): ARACNE source code can be downloaded from http://amdec-bioinfo.cu-genome.org/html/caWorkBench/upload/aracne_source.zip

  - JDK 1.5 (http://java.sun.com/j2se/1.5.0/download.jsp) 이상

  - Computer operating systems: Windows, GNU Linux or Mac OS X (version 10.4 or higher, on a PPC architecture)

- ARACNe 는 platform-independent java 기반의 jar 실행파일이 제공되며 기본옵션의 세팅만으로 실행가능함. input 파일은 ARANCe 의 input 형태의 mRNA expression profile 로 sample (column)및 probeID(row)에 대응하는 발현량을 tab-delimted 형태의 텍스트로 입력됨. ProbeID 는 해당 마이크로어레이의 probe ID 를 직접적으로 사용할 수 있으며 non-redundance gene ID 로 대체할 수 있음. ARACNe 가 input 으로 받는 파일의 형태는 다음과 같음

Col header 1	Col header2	Sample name 1	Sample name 2	…
Description				
Description				
ProbeID 1	Probe annot 1	4.5	9.8	5.6
ProbeID 2	Probe annot 2	3.6	0.5	2.8
…	…	…	…	…

- ARACNe 의 input 파일은 다수의 샘플에서 측정된 단일 expression profile 로 일부유전자(regulator)를 설정할 경우 해당 regulator 에 대해 co-expression network 을 추론함.

- miRNA-mRNA 간의 다중분석을 예로 들면, 추출된 miRNA 및 mRNA 에 만들어진 profile 을 merging 해서 만든 하나의 matrix 형태의 파일을 입력 파일로 사용. Methylation-mRNA 의 다중분석을 위해 동일한 방법(profile 을 merging 후, regulator-target 의 설정)으로 수행할 수 있음.

-

- ARACNe 의 실행 command 및 option 은 다음과 같음

```
java -jar ARACNE-java.jar [OPTIONS]
ARACNE options:
-i <file> Input gene expression profile dataset
-o <file> Output file name (optional) [*]
-j <file> Existing adjacency matrix (.adj) file
-a <fixed_bandwidth|variable_bandwidth|adaptive_partitioning>
Algorithm (fixed_bandwidth | variable_bandwidth |
```

```
adaptive_partitioning), default: adaptive_partitioning
-k <kernel width> Kernel width (accurate method only),
default: determined by program
-b <# bins> No. of bins (fast method only), default: 6
-t <threshold> MI threshold, default: 0
-p P-value for MI threshold (e.g. 1e-7), default:
```

- 즉, ARACNe 의 실제 실행 예는 다음과 같음

```
java -jar ARACNE2.jar -i <input_file> -a
adaptive_partitioning -p 1e-7
```

- MI value 에 대한 cutoff 를 직접적으로 정할 수 있으나 보통 MI threshold 에 대응하는 P value 를 정할 수 있음. 보통 1e-7(default)를 P value cutoff 로 설정하고 이를 만족하는 pair 만 report 할 수 있으며 output file size 를 최소화하기 위해 P value cutoff 를 설정하는 것을 권장함.

- miRNA 를 regulator 로 설정하고 ARACNe 를 수행하는 경우에는 - l 옵션을 이용하여 입력 데이터 매트릭스 파일에서 miRNA 에 해당하는 probeID 들의 리스트를 파일로 입력함. miRNA 뿐 아니라 TF 들도 함께 regulator 로 사용하기 위해서는 TF 의 probeID 들도 함께 입력함.

```
java -jar ARACNE2.jar -i <input_file> -a
adaptive_partitioning  -p 1e-7 -l <tf_list>
```

- P value (혹은 MI value) cutoff 를 설정하고 ARANCe 를 수행할 경우 output 을 다음과 같음. 보통 transcriptional regulator(TF) 혹은 merge 된 regulator(microRNA 등)을 설정하고 이 regulator 각각의 subnetwork (significance cutoff 를 만족하는 correlation/MI 를 보이는 target set)을 추출하는 방식으로 진행됨.

```
> Parameter name1 Parameter value 1
```

> … …				
ProbeID 1	ProbeId 2	0.08	ProbeId 5	0.15
ProbeID 2	ProbeId 1	0.08	ProbeId 3	0.22
…	…	…	…	…

- 각 행(row)가 특정 TF 의 subnetwork 에 관한 information 으로 상기 예에서 첫 줄은 probeID 1 유전자(혹은 TF)에 대해 미리 설정된 P value 를 만족하는 correlation 을 보이는 유전자(예에서는 ProbeID 2 및 ProbeID 5)와 해당하는 MI value 를 표기함. MI 에 대한 유의도(예, 1e-7)를 미리 설정한 경우, 각각의 row 에 해당 유의도를 만족하는 유전자만 표기되나 설정하지 않은 경우, 모든 유전자가 MI 값과 함께 표기됨.

- 이 결과 데이터를 이용하면 microRNA 를 입력 TF 리스트로 넣은 경우에는 유전자 발현양 프로파일 데이터를 이용한 microRNA 에 의한 mRNA 조절을 예측할 수 있음.

- ARANCe 의 경우 indirect association (A-B, B-C 간 연관에 의해 A-C 간 연관을 보이는 것)을 filter 하기 위한 DPI(data processing inequality)옵션을 설정할 수 있음. DPI 의 초기설정값은 1(100%)로 계산된 모든 edge 가 선택되나 DPI 값을 0(no tolerance)에서 0.15(15% tolerance)로 조정함으로써 false positive 를 줄일 수 있음.

- 약 200 여개의 샘플과 10,000 여개의 유전자로 구성된 인풋 파일을 이용하여 ARANCe 를 수행할 때, 9G 이상의 메모리를 사용할 정도로 많은 메모리와 계산량을 필요로 하므로 개인용 데스크탑 보다는 서버용 컴퓨터에서 수행하기를 권장함.

(3) ARACNe 의 범용성 및 R 구현성

- ARANCe 는 특정 mRNA 발현 프로파일에서 MI 기반의 일반적인 co-expression network 을 생성하는 데 이용할 수 있으나 일반적으로 해당 프로파일에서 'regulator' subset(일반적으로 transcription factor)를 설정하고 이러한 regulator 에 대응하는 subnetwork 을 추출하는 데 이용하고 있음.

- mRNA expression 과 같이 miRNA expression 을 merging 하고(동일한 샘플이 확보될 경우) 이를 regulator 로 설정하고 miRNA 별로 potential target 을 추출할 수 있음. 이러한 다중분석은 현재 miRNA + mRNA 발현프로파일로 시도되고 있으나 다른 종류의 조합(DNA 카피수 혹은 DNA 메칠화 + mRNA 발현 프로파일)에도 응용할 수 있음.

- ARACNe 가 추론하는 network 은 통상적으로 이용하는 연관계수(Pearson/Spearman)을 계산함으로써 구현될 수 있음. 예를 들어, 샘플수준으로 매칭된 mRNA expression 및 miRNA expression 프로파일이 있을 때 pairwise 연관계수는 R 에서 cor(t(mRNA_exp_matrix), t(miRNA_exp_matrix), method="pearson or "spearman")으로 계산이 가능하며, R 의 cor.test 함수를 이용하여 특정 유의도를 만족하는 pair/edge 를 선택할 수 있음.

(4) ARANCe 결과의 visualization

- ARANCe 결과의 network visualization 을 위해서 CytoScape [15]등의 tool 을 이용할 수 있음. CytoScape 의 경우 network 의 인자(node-edge)를 text file 로 인지하여 network visualization 및 해당 network 의 다양한 property 를 계산할 수 있는 범용적인 툴로서 상기 ARACNe 결과를 단순 변환을 통해 CytoScape input 파일로 전환하여 이용함.

- Cytoscape 의 input file 로서 node 와 node 간의 관계로 individual 라인으로 변환 후, text 형태로 가져올 수 있음. Cytoscape v2 이상의 경우, Import->Network->File 로 파일을 지정후, "source"및 "target"컬럼을 지정함으로써(v3.3 이상의 경우, GUI 형식) network visualization 이 가능함.

- Network 로딩 후, 'organic'등의 다양한 layout 을 통한 visualization 이 가능하며, Tools->NetworkAnalyzer->Network analysis 를 통해 connectivity 등의 다양한 network property 를 계산할 수 있음.

- 그림 4 는 cytoscape 를 이용한 subnetwork 분석의 예를 보임.

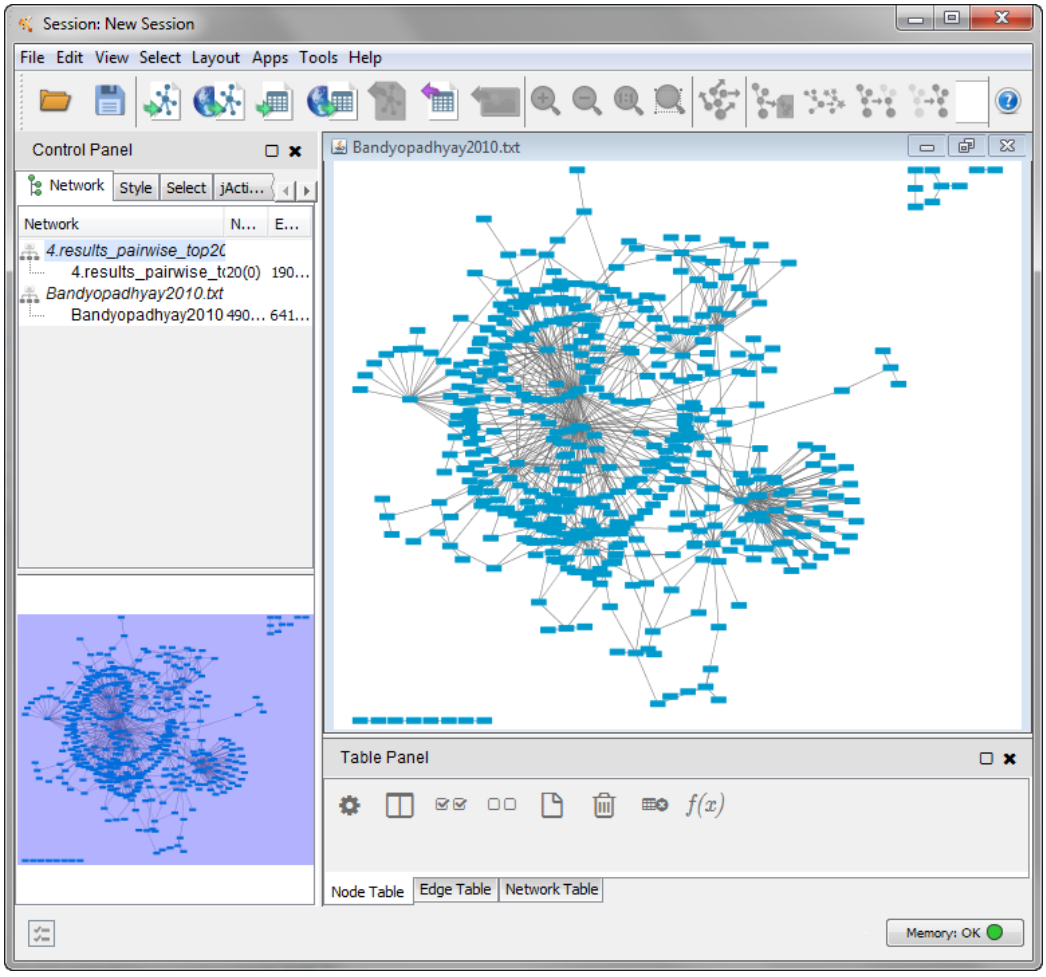


그림 4. Cytoscape(V3.2)의 snapshot. BioGrid 에서 제공하는 yeast-two-hybrid 를 통한 gene-gene network 의 예로 'Organic' layout 이 선택됨.

3. 다중유전체데이터 연관분석 - iCluster

(1) 전분석 - 유전자클러스터링

- 유전자발현기반의 클러스터링은 mRNA 발현프로파일의 기본적인 QC 후의 데이터 quality 및 임상연관성을 보기 위해 수행됨. Phenotype 의 명확한 차이를 보이는 군(예, 환자 및 정상군) 간에 유의한 유전발현차이가 있다면 클러스터링 등의 방법에 의해 그 차이를 확인할 수 있으며 이는 phenotype 의 차이를 설명할 수 있는 유전발현 차이가 이미 데이터에 내재되어 있음을 시사함.

- 유전자 발현 클러스터링은 해당 소프트웨어(Cluster/TreeView) 혹은 범용소프트웨어인 R 로 수행함. 보통 2 만개가 넘은 유전자를 대상으로 하고 있기 때문에 유의한 유전자서브셋을 추출하여 분석을 진행하게 됨. 일반적으로 샘플간의 유전발현다양성을 나타내는 지표인 MAD(median absolute deviation)이나 표준편차가 높은 일부(보통, 100-1000 개)의 유전자를 선택하여 클러스터링을 수행하게 됨.

- 계통적(hierarchical) 클러스터링은 유전자 혹은 샘플간의 거리(distance)를 측정하는 방법 및 계산된 거리를 어떻게 linkage 할 것인가에 따라 다양한 조합의 시도가 가능함. 보통 Pearson correlation(1-PCC)을 distance measure 로 일반적으로 사용하나 이외, euclidean distance, (non-parametric) Spearman correlation 등을 사용할 수 있음. Linkage 방법의 경우 거리매트릭스(distance matrix)에서 최단 거리의 쌍을 선택후 해당 쌍의 유전자 2 개의 거리를 평균(average), 최소(single), 최대(complete)선택에 따른 linkage 옵션을 조정할 수 있음.

- 계통클러스터링 이외에, K-means 클러스터링 및 PCA(principle component analysis)를 수행할 수 있음. K-means 클러스터링의 경우 이미 정해진 갯수(k 값)에 대해 해당 데이터세트를 구분하는 것으로 정해진 반복수(iteration) 동안 랜덤하게 나뉘어진 클러스터군을 거리 및 유사도에 의해 재배열하는 과정을 반복하여 최종적으로 k 군을 얻게됨. PCA 의 경우, 해당 변수(보통 유전자수)의 변이도를 가장 잘 대변하는 주성분(PCA)을 순서대로 분리하여, PCA1-PCA2 두개의 변수(혹은 PCA1-3 의 3 차원)의 2 차원구분(보통 scatter plot)을 통해 데이터를 구분하게 됨.

(2) iCluster 소개

- iCluster 는 다중 데이터로 부터 클러스터링을 수행할 수 있는 수단으로 보고되어, 2009 년 Bioinformatics 저널에 처음 보고된 후 [16], 2012 년 GBM 에서 molecular subtype 을 구분하기 위한 방법으로 iCluster2 가 plos one 저널에 발표되었음 [2]. 본 SOP 에서는 iCluster2 를 기준으로 실행 방법을 설명함.

- iCluster 는 copy number 데이터, 유전자 발현 데이터, DNA methylation 데이터들을 동시에 입력으로 받아서, 은닉변수모델(latent variable model)을 이용하여 클러스터링을 수행함. iCluster2 에서는 이를 변형한 variance-weighted shrinkage 방법에 기반하여 클러스터링을 수행함.

(3) iCluster 설치 및 실행

- iCluster 는 R 을 이용하여 수행할 수 있음. R 에서 iCluster 패키지를 설치하기 위해서는 다음과 같은 명령어를 수행함.

```
> install.packages("iCluster")
```

- 입력 데이터는 다중 유전체 데이터가 하나의 list 구조에 들어있는 형태임. 즉, copy number 데이터, 유전자 발현 데이터, methylation 데이터 세 가지 데이터가 각각, cn, exp, methyl 이라는 이름의 matrix 로 존재한다고 한다면 다음과 같은 명령어를 이용하여 list 구조 형태의 입력 데이터로 변형할 수 있음. 각각의 유전체 데이터 matrix(cn, exp, methyl)에서 공통적으ㅗㄹ 행(row)은 샘플, 열(column)은 유전자를 의미함.

```
> data <- list(cn, exp, methyl)
```

- iCluster2 실행을 위한 명령어 및 옵션은 다음과 같다.

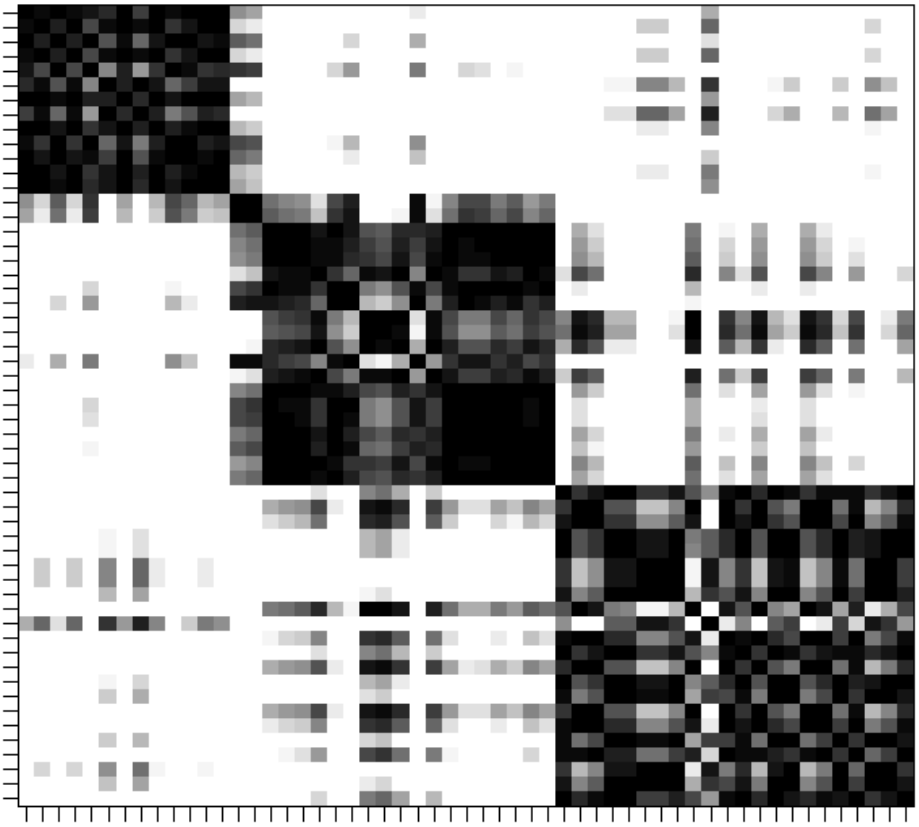
```
> iCluster2(datasets, k, lambda=NULL, scale=T,
scalar=F, max.iter=10, verbose=T)


- datasets: A list containing data matrices. For each
data matrix, the rows represent samples,
and the columns represent genomic features.
- k: Number of classes for the samples.
- lambda: Penalty term for the coefficient matrix of
the iCluster model.
- scalar: Logical value. If true, a degenerate
version assuming scalar covariance matrix
is used.
- max.iter: maximum iteration for the EM algorithm
- scale: Logical value. If true, data matrix is
column centered
- verbose: Logical value. If true, print message.
```

- 클러스터 3 개를 만드는 경우에 대한 실제 실행 예시는 다음과 같다.

```
> fit=iCluster2(datasets=data, k=3,
lambda=list(0.44,0.33,0.28))
```

- 클러스터링 결과를 확인하기 위해서는 plotiCluster 함수를 이용한다. 각 샘플들이 어떻게 클러스터링 되었는지 symmetric matrix 의 heatmap 형태로 그림이 그려진다. 그림에서 각 row 와 column 은 각 샘플을 의미한다.

```
> plotiCluster(fit=fit, label=rownames(data[[1]]))
```



4. 참고문헌

[1] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 7(Suppl 1):S7, 2006.

[2] Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using iCluster. PLoS One. 7(4):e35236, 2012.

[3] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, Bioinformatics 26(12):i237-245, 2010.

[4] Meng C, Helm D, Frejno M1, Kuster B., moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J Proteome Res. 15(3):755-765, 2016.

[5] Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C., Analysis of microRNA-target interactions across diverse cancer types. Nat Struct Mol Biol. 20(11):1325-1332, 2013.

[6] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. 22(3):568-576, 2012.

[7] Xi R, Lee S, Xia Y, Kim TM, Park PJ., Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. 44(13):6274-6286, 2016.

[8] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 5(4):557-572, 2004

[9] Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E., Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics, 20(18):3413-3422, 2004.

[10] Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 30(5):413-421, 2012.

[11] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol., 12(4):R41, 2011.

[12] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RG., Inferring tumour purity and stromal and immune cell admixture from expression data., Nat Commun. 4:2612, 2013.

[13] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 37(4):382-390, 2005.

[14] Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 463(7279):318-325, 2010.

[15] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498-2504, 2003.

[16] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 25(22):2906-2912, 2009.