CAMVS(V1.0) : CGH Analyzer and Map Viewer using S-Plus(V1.0)

Sangcheol Kim[1*], Chanhee Park[1,2], Minyoung Seo[1], Hajin Jeong[1,2], Inyoung Kim[1], Hyun Cheol Chung[1,2], Sun Young Rha[1,2]

[1] Cancer Metastasis Research Center, College of Medicine, Yonsei. Univ ersity, Seoul, Korea
[2] Brain Korea 21 Project for Medical Science, College of Medicine, Yonsei Univ ersity, Seoul, Korea

[*] E-mail: kimsc77@kribb.re.kr , current address : National Genome Information Center

Abstract

DNA 단계에서의 유전자의 증폭과 소실은 종양의 발생과 진행에 중요한 역할을 한다. 유전자의 변화를 관찰하기 위해서 Comparative Genomic Hybridization(CGH) 기술이 많이 이용되어져 왔다. 최근에는 이러한 CGH 기술을 응용하여 cDNA microarray 를 이용한 고밀도 CGH(Microarray-CGH) 기술이 보고 되고 있다. Microarray-CGH 에서 유전자별 변화 정도를 유전자의 log-비의 값의 변화 정도와 염색체 위치 정보를 이용하여 DNA 단계에서의 유전자의 변화 정도를 확인 할 수 있다. 또한 동일한 유전자의 칩을 사용하여 RNA 단계에서의 발현 양상과 직접 비교할 수 있는 장점이 있다. 현재 microarray 분석법은 많이 개발되고 실용화 되고 있으나 Microarray-CGH 분석을 위한 프로그램들은 아직 초보 단계며, 생물학자들이 사용하기 힘들고, 프로그램에 분석 자료를 적용하기 어려운 경향이 있다. 위와 같은 단점을 보완하기 위해서 개발된 CAMVS(V1.0) 프로그램은 S-plus(2000)을 기반으로 개발하였고, 복잡한 분석보다는 모든 결과들을 이미지화 할 수 있으며 파일로 결과를 쉽게 확인할 수 있도록 디자인하였다. CAMVS(V1.0)는 전체 염색체를 각 실험별로 비교 분석하는 부분, 특정 염색체를 특정 실험별로 비교 분석하는 부분과 실험간의 차이를 통계적으로 비교 분석하는 3 가지 카테고리로 구성되어 있다. 쉬운 알고리즘과 사용의 편리함, 분석결과의 다양한 그래픽, 새로운 알고리즘 추가의 용이성 등이 CAMVS(V1.0)가 가지고 있는 장점이며, Microarray-CGH를 분석하는데 아주 유용한 분석 도구이다.

Introduction

This work is supported by the Korea Science and Engineering Foundation(KOSEF) through the Cancer Metastasis Research Center(CMRC) at Yonsei University College of Medicine.

유전자의 증폭과 소실은 종양의 발생과 진행에 중요한 역할을 하여 DNA 단계에서 유전적 변이를 관찰하기 위해서는 Representational Difference Analysis(RDA), Restriction Landmark Genome Scanning(RSGS), Loss of

heterozygosity(LOS) 등의 방법(1)과 Comparative Genomic Hybridization(CGH) 기술이 많이 이용되어져 왔다(2). 최근에는 이러한 CGH 기술을 응용하여 DNA 단계에서의 유전자의 유전적 변이를 관찰하기 위해서 cDNA microarray를 이용한 고밀도 CGH(Microarray-CGH) 기술이 많이 보고 되고 있다(3). Microarray-CGH는 cDNA microarray의 실험 기법과 동일한 원리로 test sample과 reference sample간의 변화 정도를 mRNA가 아닌 DNA의 발현량으로 유전적 변이를 관찰하는 방법이다. Microarray-CGH에서는 유전자의 유전적 변이 정도를 유전자의 log-비의 값의 변화 정도와 염색체 위치 정보를 이용하여 염색체 상의 유전자들의 위치와 그에 따른 유전적 변이 정도를 결정한다.

DNA 단계에서의 유전적 변이는 크게 3단계로 나뉠 수가 있다. 유전적 변이가 없을 때는 no change, reference에 비해서 유전자의 DNA가 증가하였을 때를 gain, 그리고 DNA가 감소하였을 때를 loss라고 한다. 이러한 gain, loss가 있는 유전자 중에서 특정 기준 이상의 유전적 변이가 증가했을 때 증폭(amplification), 감소(deletion)라 하며 이렇게 결정된 유전자를 유전적 의미가 있는 유전자로 구분된다. 유전적 변이(gain, loss, amplification, deletion)를 결정하기 위한 기준은 실험실 별로 특정 기준값을 결정하여 사용하거나, 일반적으로 log-비의 값이 +/- 0.58을 기준으로 유전자의 증폭 또는 감소의 기준으로 사용하고 있다(4).

이러한 기준을 이용한 Microarray-CGH의 분석은 아직은 초보단계에 있으며 현재 개발되어진 프로그램들은 생물학자들이 사용하기 힘들고, 실제 자료를 적용하기 어려운 경향이 있다. 위와 같은 단점을 보완하기 위해서 복잡한 분석보다는 모든 결과들을 이미지와 결과 자료로 쉽게 확인할 수 있고, 생물학자들이 쉽게 사용할 수 있는 프로그램이 필요하여 CAMVS(V1.0) 프로그램을 S-plus(2000)을 기반으로 개발하였다.

Methods

CAMVS(V1.0)는 각 실험의 Microarray-CGH에서 유전자별 변화 정도를 나타내는 log-비와 염색체 위치 정보를 이용하여 DNA 단계에서의 유전자의 유전적 변이 정도를 확인할 수 있으며, 그림 1과 같이 전체 염색체를 각 실험별로 비교 분석하는 부분, 특정 염색체를 특정 실험별로 비교 분석하는 부분과 실험간의 차이를 통계적으로 비교 분석하는 3가지 카테고리로 구성되어져 있다.

분석 자료 구성

CAMVS(V1.0)를 실행하기 위해서는 각 실험의 Microarray-CGH에서 유전자의 변화 정도를 나타내는 log-비는 적절한 표준화 방법(4), 결측치 보정 방법(5)과 유전자 선별 과정 등의 사전 준비 단계를 통하여 얻는다(6). 또한 Microarray-CGH와 같은 수 많은 유전자의 자료의 생물학적 정보를 개별적으로 얻는 과정은 쉽지 않다. 이러한 과정은 SOURE(7)와 같은 database를 통하여 전체 유전자들의 여러 정보 중 염색체 정보와 염색체 상의 위치 정보를 얻을 수 있다. S-plus 2000은 다른 프로그램과 연동이 잘 되어서 DS1에는 Gene ID, 각 실험의 log-비 염색체 정보, 염색체 위치 정보 순으로 쉽

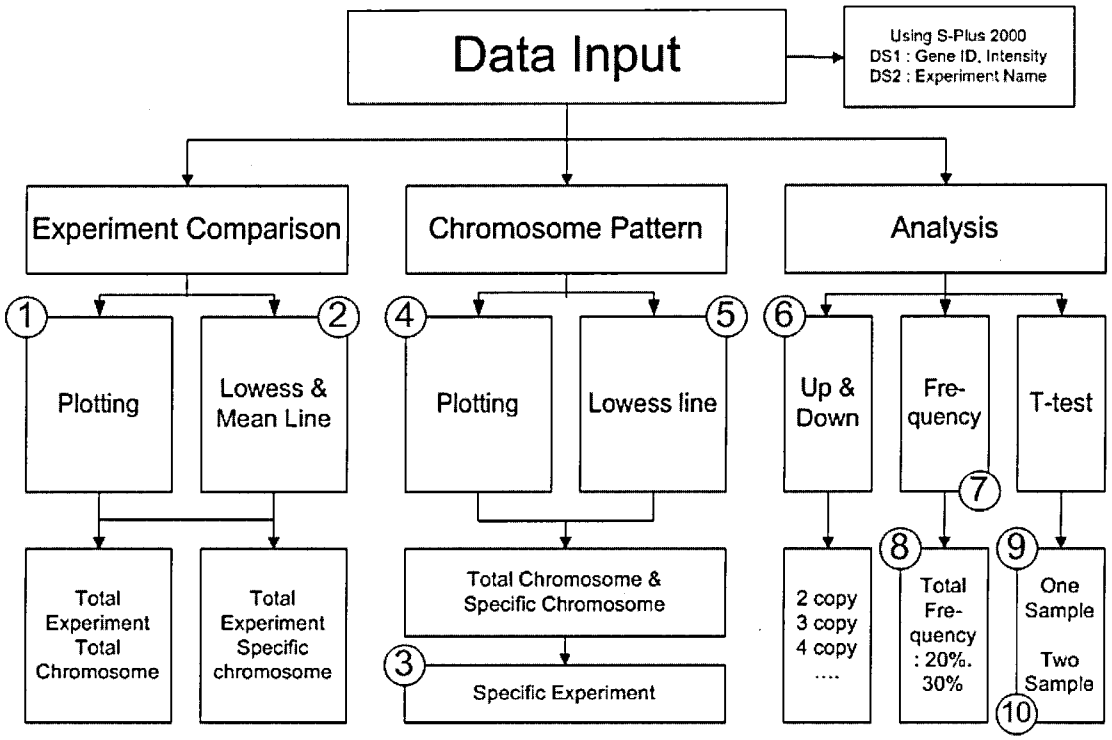
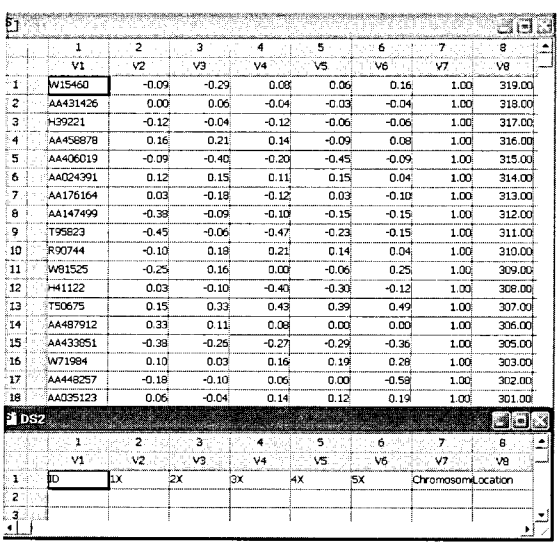


그림1. CAMVS(V1.0)의 구성 모식도

게 upload하여 입력 가능하다. DS2에는 DS1에서 각 행의 실험 정보 및 염색체 정보를 입력을 한다(그림2).



	1	2	3	4	5	6	7	8
	V1	V2	V3	V4	V5	V6	V7	V8
1	W15460	-0.09	-0.29	0.08	0.06	0.16	1.00	319.00
2	AA431426	0.00	0.06	-0.04	-0.03	-0.04	1.00	318.00
3	H39221	-0.12	-0.04	-0.12	-0.06	-0.06	1.00	317.00
4	AA458878	0.16	0.21	0.14	-0.09	0.08	1.00	316.00
5	AA406019	-0.09	-0.40	-0.20	-0.45	-0.09	1.00	315.00
6	AA024391	0.12	0.15	0.11	0.15	0.04	1.00	314.00
7	AA176164	0.03	-0.18	-0.12	0.03	-0.10	1.00	313.00
8	AA147499	-0.38	-0.09	-0.10	-0.15	-0.15	1.00	312.00
9	T95823	-0.45	-0.06	-0.47	-0.23	-0.15	1.00	311.00
10	R90744	-0.10	0.18	0.21	0.14	0.04	1.00	310.00
11	W81525	-0.25	0.16	0.00	-0.06	0.25	1.00	309.00
12	H41122	0.03	-0.10	-0.40	-0.30	-0.12	1.00	308.00
13	T50675	0.15	0.33	0.43	0.39	0.49	1.00	307.00
14	AA487912	0.33	0.11	0.08	0.00	0.00	1.00	306.00
15	AA433851	-0.38	-0.25	-0.27	-0.29	-0.36	1.00	305.00
16	W71984	0.10	0.03	0.16	0.19	0.28	1.00	303.00
17	AA448257	-0.18	-0.10	0.06	0.00	-0.58	1.00	302.00
18	AA035123	0.06	-0.04	0.14	0.12	0.19	1.00	301.00

DS2

	1	2	3	4	5	6	7	8
	V1	V2	V3	V4	V5	V6	V7	V8
1	ID	1X	2X	3X	4X	5X	ChromosomeLocation	
2								
3								

그림 2. CAMVS(V1.0) 입력 자료 형태

사용자 입력 변수 설정

CAMVS(V1.0)에서 제공하는 방법을 사용자가 이용하기 위해서는 사용자 입력 부분의 변수를 설정해야 한다. 입력 부분은 크게 프로그램 전체에서 고정적으로 사용되는 변수와 특정 방법에 필요한 변수로 구성되어 있으며 각 입력 변수들은 디폴트로 17K microarray 를 이용할 때 가장 유용한 결과를 얻을 수 있도록 설정되었다. 만약 사용자가 이용하는 칩의 유전자의 수가 달라지면 변수를 조정함으로써 결과를 얻을 수 있도록 구성되었다. 그림 3 은 디폴트로 설정된 변수들과 각각의 변수들을 설명하고 있다.

```
con.class<-10                          # method : 1 , 1-2 ,1-3 comparison of experiment number : default 10
All.span<-0.025                        # total chromosome lowess function span : default 0.025
chro.span<-0.2                         # each chromosome lowess function span  : default 0.2
P.Fold.change<-  0.64                  # Fold.chagne  0.58 positive if method :0 process omit
M.Fold.change<- -0.64                  # Fold.chagne  -0.58 negative if method :0 process omit
scale.1<-0.1                           # method : 1-1 plot scale adjust
ADD.lines <- c( 0, 0 )                 # method : 3 ADD lines
cut.point <- c(log(1.5,2),log(2,2),log(3,2),log(4,2))  # method : 6 Fold.change Up down Frequency
cut.proportion<-0.2                    # method : 7 total data great than proportion of %
cut.count<-c(1,10,13,15,18,21,24)      # method : 8 Frequency of each data great than of threshold
One.Class  <- c(2:31)                  # method : 9 One sample T-Test class
Two.Class.1 <- c(2:16)                 # method : 10 Two sample T-Test class
Two.Class.2 <- c(17:31)                # method : 10 Two sample T-Test class
```

그림 3. CAMVS(V1.0) 사용자 입력

실험별 분석

실험간의 비교 분석은 그림 1 에서와 같이 2 가지로 구분된다. CAMVS(V1.0)의 방법 1 은 전체 실험을 비교 분석하는 방법으로 그림 4 의 A 에서처럼 각 실험간의 비교를 전체 염색체에서 유전자들의 발현량의 정도와 중심 경향은 lowess 곡선으로 표현하였다. 또한, 각 유전자들이 사용자가 설정한 증폭, 감소 기준값을 벗어났을 경우에는 양쪽에 - 부호로 쉽게 확인할 수 있다.

CAMVS(V1.0)의 방법 2 는 전체 염색체 중심 경향을 lowess 곡선으로 표현하여 한 장의 그림에서 각 실험에서 전체 염색체의 변화 양상을 파악할 수 있고, 전체 실험을 유전자별로 평균을 계산하여 lowess 곡선을 통하여 전체 유전자의 중심 경향을 그림 4 의 B 처럼 확인할 수 있다.

CAMVS(V1.0)에서는 방법 1, 2 로 실험간의 비교 분석뿐만 아니라, 전체 실험의 변화 양상을 파악할 수 있다. 특이적으로 변화하는 염색체나, 사용자가 관심 있어하는 특정 염색체를 프로그램을 수행하는 단계에서 선택하면 그림 4 의 결과처럼 유전자의 발현량의 정도, lowess 곡선을 통한 중심경향, 기준값(증폭, 감소)을 벗어났는지 등을 확인할 수 있다.

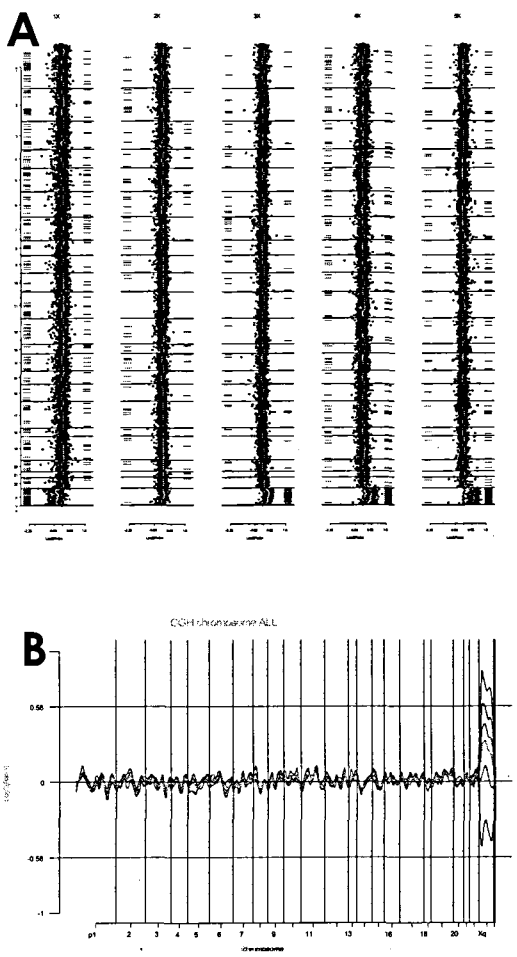


그림 4. 실험별 분석 결과

염색체의 경향 분석

CAMVS(V1.0)에서 염색체의 경향 분석은 그림 1 에서와 같이 2 부분으로 구성된다. 방법 3 은 각 실험별로 CGH-Microarray 의 경향을 파악하기 위한 방법이다. 실험 별로 전체 염색체가 유전자별로 발현하는 정도와 중심 경향을 lowess 곡선을 통하여 확인하고, 증폭 기준값을 벗어났을 경우 +(빨강), 감소 기준값을 벗어났을 경우는 - (녹색) 로 확인할 수 있다(그림 5 A).

방법 4 는 실험별로 염색체 별로 나눠서 그림 4 의 B 의 결과처럼 발현량의

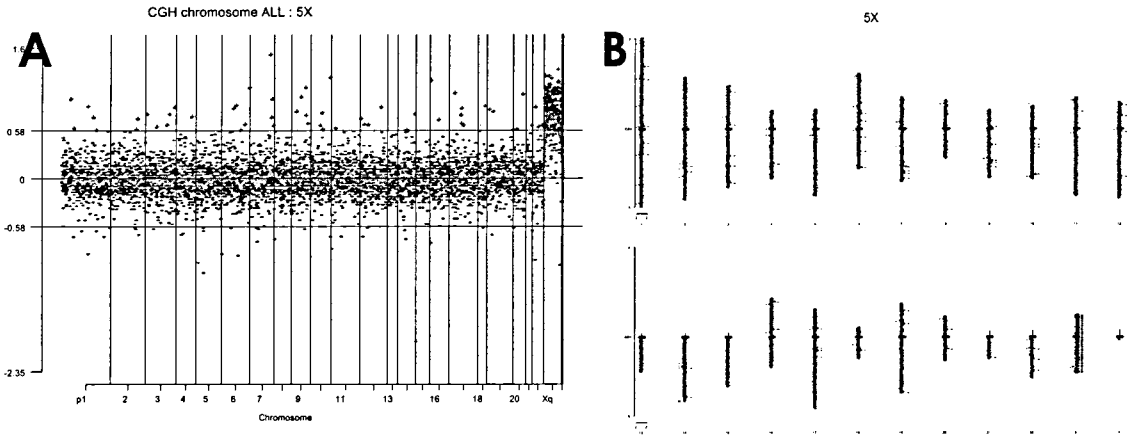


그림 5. 염색체의 경향 분석

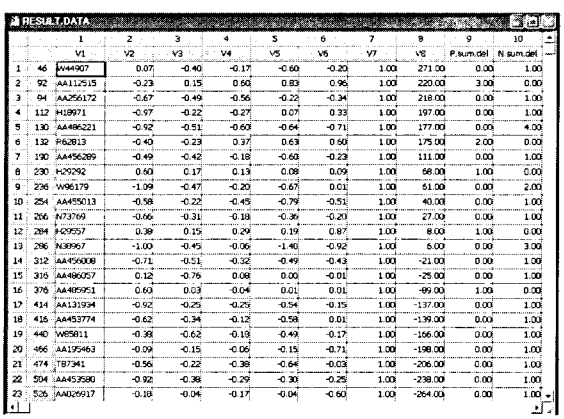
정도, 중심경향 기준값 벗어났을 경우엔 gain 은 빨간색의 - 부호, loss 의 경우엔 녹색의 - 부호로 확인 가능하다.

사용자가 관심 있어하는 특정 염색체를 방법 3 을 수행하는 단계에서 선택하면 그림 5 의 A 결과를 보여주며 기준값을 벗어났을 경우엔 유전자 ID 를 확인 가능함으로 사용자가 이해하기 용이하도록 되어있다.

분석

CAMVS(V1.0)에서는 분석 부분은 3 부분으로 구성되어있다(그림1). 방법 6은 기준값(1.5, 2, 2.5, … Fold change)을 여러 가지를 설정하여 각 실험별로 기준값을 벗어나는 유전자의 빈도수를 계산하여 준다. 이 방법은 실험별로 자료의 형태가 얼마나 유사한지를 확인 가능하게 하여 사용자가 자료 별 특징을 파악하는데 유용하게 이용할 수 있다.

Microarray-CGH 실험에서는 각 유전자가 전체 실험에서 증폭, 감소가 되는 빈도의 비율이 분석 시에 중요한 부분이다. 방법 7에서는 사용자가 전체에서 적합한 비율을 선택하면 그 비율을 넘는 유전자만을 선택하여 Microarray-CGH 분석하는데 용이하게 하였다. 결과는 그림 6과 같이 DS1에 입력한 자료에서 기준값(증폭, 감소)을 벗어난 빈도수를 S-Plus 결과 자료에서 확인할 수 있다.


RESULT.DATA

	1	2	3	4	5	6	7	8	9	10
	V1	V2	V3	V4	V5	V6	V7	V8	P.sum.del	N.sum.del
1	46	W44907	0.07	-0.40	-0.17	-0.60	-0.20	1.00	271.00	0.00	1.00
2	92	AA112515	-0.23	0.15	0.60	0.83	0.96	1.00	220.00	3.00	0.00
3	94	AA256172	-0.67	-0.49	-0.56	-0.22	-0.34	1.00	218.00	0.00	1.00
4	112	H18971	-0.97	-0.22	-0.27	0.07	0.33	1.00	197.00	0.00	1.00
5	130	AA486221	-0.92	-0.51	-0.60	-0.64	-0.71	1.00	177.00	0.00	4.00
6	132	R62813	-0.40	-0.23	0.37	0.63	0.60	1.00	175.00	2.00	0.00
7	190	AA456289	-0.49	-0.42	-0.18	-0.60	-0.23	1.00	111.00	0.00	1.00
8	230	H29292	0.60	0.17	0.13	0.08	0.09	1.00	68.00	1.00	0.00
9	236	W96179	-1.09	-0.47	-0.20	-0.67	0.01	1.00	61.00	0.00	2.00
10	254	AA455013	-0.58	-0.22	-0.45	-0.79	-0.51	1.00	40.00	0.00	1.00
11	266	N73769	-0.66	-0.31	-0.18	-0.36	-0.20	1.00	27.00	0.00	1.00
12	284	H29557	0.39	0.15	0.29	0.19	0.87	1.00	8.00	1.00	0.00
13	296	N38967	-1.00	-0.45	-0.06	-1.40	-0.92	1.00	6.00	0.00	3.00
14	312	AA456008	-0.71	-0.51	-0.32	-0.49	-0.43	1.00	-21.00	0.00	1.00
15	316	AA486057	0.12	-0.76	0.08	0.00	-0.01	1.00	-25.00	0.00	1.00
16	376	AA485951	0.60	0.03	-0.04	0.01	0.01	1.00	-99.00	1.00	0.00
17	414	AA131934	-0.92	-0.25	-0.25	-0.54	-0.15	1.00	-137.00	0.00	1.00
18	416	AA453774	-0.62	-0.34	-0.12	-0.58	0.01	1.00	-139.00	0.00	1.00
19	440	W85811	-0.38	-0.62	-0.19	-0.49	-0.17	1.00	-166.00	0.00	1.00
20	466	AA195463	-0.09	-0.15	-0.06	-0.15	-0.71	1.00	-198.00	0.00	1.00
21	474	T87341	-0.56	-0.22	-0.38	-0.64	-0.03	1.00	-206.00	0.00	1.00
22	504	AA453580	-0.92	-0.38	-0.29	-0.30	-0.25	1.00	-238.00	0.00	1.00
23	526	AA026917	-0.18	-0.04	-0.17	-0.04	-0.60	1.00	-264.00	0.00	1.00

그림6. CAMVS(V1.0) 증폭, 감소 빈도수 확인

방법 8은 방법 7에서 얻은 결과를 확장하여 다양한 기준을 사용자가 해석하기 쉽게 그림으로 표현한 것이다. 또한 사용자가 여러 개의 빈도수를 선택하여 특정 유전자나 염색체 상에서 gain, loss, 증폭(amplification), 감소(deletion)를

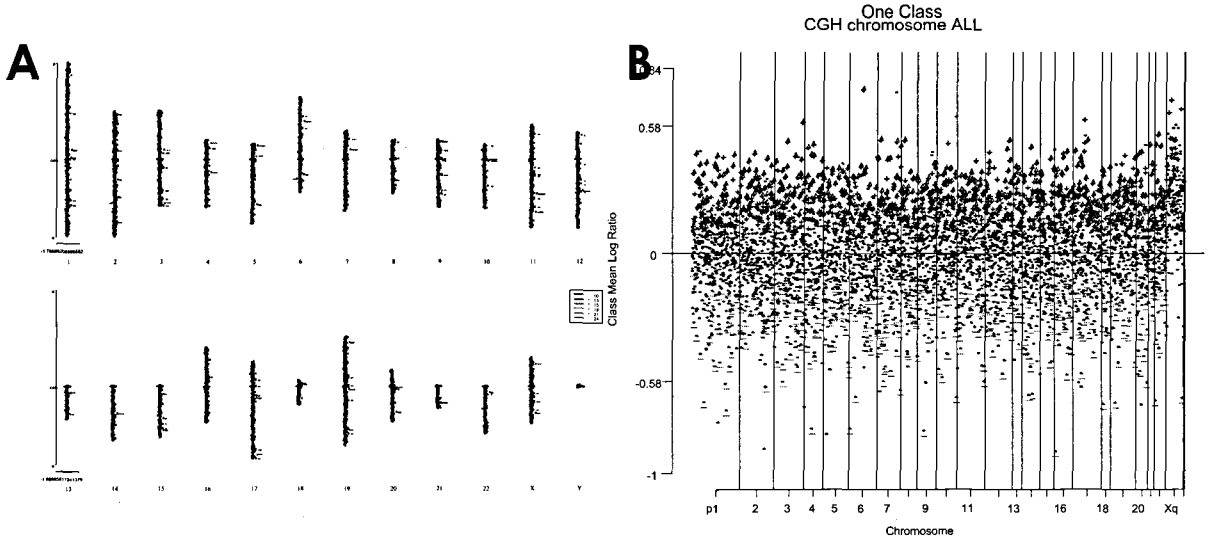


그림 7. CAMVS(V1.0) 빈도수, *t-test* 분석 결과

확인할 수 있다. 결과는 그림 7의 A 그림으로 확인할 수 있듯이 사용자 선택한 빈도수에 따라서 다른 색과 다른 길이로 표현함으로써 특정 염색체의 어느 부분이 증폭, 감소되는지를 그림으로 해석하기 용의하게 하였다. 또한, 각주를 추가함으로써 한 장의 그림 에서 비교 분석이 가능하도록 하였다.

CAMVS(V1.0)에서 방법 9, 10은 통계분석 방법인 *t-test* 를 one sample, two sample 일 경우를 적용시켜 각 유전자마다 *t-test* 결과를 통하여 p-value 가 0.05, 0.01보다 작은 경우에 그림 7의 B 결과처럼 다른 기호와 색깔로 표시하였고, 각각의 결과는 따로 파일에 저장되어 추후에 생물학적으로 해석할 때 이용하기 쉽게 하였다.

Result / Discussion

유전자의 증폭과 소실은 종양의 발생과 진행을 DNA 단계에서 유전적 변이를 관찰할 수 있는 cDNA microarray를 이용한 고밀도 CGH (Microarray-CGH) 분석법은 많이 개발되고 있다. 그러나, Microarray-CGH분석을 위한 프로그램들은 부족하고, 실험자가 사용하기 힘들고, 실제 자료를 적용하기 어려워서 해석하기 쉽고, 사용하기 쉬운 프로그램이 필요하여 CAMVS(V1.0) 가 개발하였다.

각 실험의 Microarray-CGH에서 각 유전자의 log-비와 염색체 위치 정보를 이용하여 DNA 단계에서의 유전자의 변화를 CAMVS(V1.0)를 이용하여 전체 염색체를 각 실험별로 비교 분석, 특정 염색체를 특정 실험별로 비교 분석과 실험간의 차이를 통계적으로 비교 분석하는 3가지의 카테고리를 분석할 수 있다.

Microarray-CGH의 실험이 활발히 진행됨에 따라 사용자의 요구가 다양해지고, 다양한 통계적 방법들이 새로이 제시되고 있다(8,9,10). 이러한 현실에 부합하기 위해서 CAMVS(V1.0) 프로그램은 새로운 알고리즘을 추가되고, 다양한 요구에 부합하도록 발전해야 될 것이다. 또한, S-plus와 유사한 코드를 가지는 R 언어로 개발이 된다면,

사용자가 S-plus이용에 대한 부담 없이 사용이 가능하고 다른 개발자들에 의해서 추가 개발된다면 DNA 단계에서의 종양의 발생과 진행을 이해하는데 유용할 것 이다.

Acknowledgements

이 프로그램이 완성되기까지 아낌없는 지도와 격려를 해주신 연세대학교 통계학과 김병수 교수님과 세종대학교 응용수학과 이선호 교수님께 감사 드립니다.

References

[1] Gray,J. and Collines,C. Genome changes and gene expression in human solid tumors. *Carcinogenesis*, 21, 2000, 443-452

[2] Pollack,J., Perou,C., Alizadeh,A., Eisen,M., Pergamenschikov,A., Williams,C., Jeffrey,S., Botstein,D. and Brown,P. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23, 1999, 41-46

[3] Seo,MY. Rha,SY. Yang,SH. Kim,SC. Lee,GY. Park,CH. Yang,WI. Ahn,JB. Park,BW. Chung,HC. The pattern of gene copy number changes in bilateral breast cancer surveyed by cDNA microarray-based comparative genomic hybridization. *Int J Mol Med.*, 13(1), 2004, Jan, 17-24.

[4] Yang,YH. Dodit,S. Luu,P. Lin,DM. Peng,V. Ngai,J. et al. Normalization for cDNA microarray data : a robust composite method adderessing single and multiple slide systematic variation. *Nucleic Acides Res.*, 30, 2002, e15

[5] Troyanskaya,O. Cantor,M. Sherlack,G. Brown,P. Hastie,T. Tibshirani,R. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17, 2001, 520-525

[6] Kim,BS. Kim,IY. Lee,SH. Kim,SC. Rha,SY. Chung,HC. Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, 2004, Accepted

[7] Diehn,M. Sherlock,G. Binkley,G. Jin,H. Matese,JC. Hernandez-Boussard,T. Rees,CA. Cherry,JM. Botstein,D. Brown,PO. Alizadeh, AA. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, 31(1), 2003, Jan 1, 219-23.

[8] Autio,R., Hautaniemi,S., Kauraniemi,P., Harja,OY., Astola,J. Wolf,M. and Kallioniemi,A. CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics,* 19, 2003, 1714-1715

[9] Wang,P. Kim,Y. Pollack,J. Narasimhan,B. Tibshirani,R. A method for calling gains and losses in array CGH data *Technical report*, 2004, 1-28

[10] Myers,CL. Dunham,MJ. Kung,SY. Troyanskaya,OG. Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics,* 2004, Accepted