고밀도 마이크로어레이를 이용한 위암 관련 유전자의 통계적 규명

연세대학교 대학원

의과학과

김 상 철

고밀도 마이크로어레이를 이용한 위암 관련 유전자의 통계적 규명

연세대학교 대학원

의과학과

김 상 철

고밀도 마이크로어레이를 이용한 위암 관련 유전자의 통계적 규명

지도교수 라선영

이 논문을 석사 학위논문으로 제출함

2004년 6 월 일

연세대학교 대학원

의 과 학 과

김 상 철

김상철의 석사 학위논문을 인준함

심사위원_____인

심사위원_____인

심사위원_____인

연세대학교 대학원

2004년 6 월 일

감사의 글

계절이 2번씩 바뀌서 어느덧 대학원 생활을 정리할 때가 되었습니다. 2년 동안 부족하지만 건강과 지혜로 마무리할 수 있도록 보살펴 주신 하나님께 감사드립니다. 그리고, 성호 교회 목사님을 비롯해서 성도여러분, 중고등부 학생들에게 감사드립니다. 자나 깨나 자식들 걱정하시고 하루도 빠짐없이 기도해주시고 걱정해주신 부모님께 감사드립니다. 하나뿐이 없는 오빠 비유를 맞춰준 동생에게도 감사드립니다.

생물정보학을 하기 위해서 온 저에게 따뜻한 관심과 배려와 지도해주신 정현철 선생님과 지도 교수인 라선영 선생님께 감사드립니다. 부족한 저에게 진정한 학자의 모습을 보여주신 김병수 선생님께 감사드립니다. 대학부터 지금까지 여러모로 도와주신 이선호 선생님께 감사드립니다. 누나가 없는 저에게 친 누나와 같은 존재였던 김인영 박사님께 감사드립니다. 같이 일을하면서 도움을 주신 양상화 박사님, 김미정 박사님, 박규현 선생님, 안성환 박사님, 정희철 선생님, 조성배 교수님께 감사드립니다.

이방인 같은 저를 따뜻하게 맞이하여준 실험실 식구들... 동문 식구가된 김태수 선생님(전 43기에요ㅋ), 따뜻한 눈빛으로 이야기 해주신 심웅호 선생님, 연애상담을 잘해준 면희형, 언제나 든든했고 곧 가장이 될 태문이성~, 저의 장난도 너무나 잘 받아준 하진이 형, 유찬이 아빠 우영이 형, 그의 개그를 따를자 없다 정범이 형, 푸근한 유근이형, 205번 버스 식구 영이형, 시끄럽지만 친구처럼? 편안하게 지낸 찬희형, 뻥이 심했지만 나름데로 자신의 일을 찾은 세원이형, 지금은 애 엄마가 된 가비누나, 많은 이야기를 못해봤지만 따뜻한 느낌이 드는 은송이 누나, 현정이 누나, 성시경의 열열한 팬 실험실의 왕 누나 주혜누나, 지금은 다른 곳에서 열심히 일하고있을 재휘형, 동고동락을 했어야될 친구 현석이, 마지막까지 미안함을 느끼게하는 연호, 내숭의 달인?? 정옥 삐지지마..ㅋㅋ, 그녀의 열정을 따를 수 없었다. 여러모로 고마웠던 민영이, 이거 저것 많이 챙겨준 지혜, 메신져에서 더 친해진 정연이, 새롭게 연구소를 이끌어갈 넉넉한 녀석 동혁이, 신앙생활이 부러웠던 귀연이, 내 친구의 여친이된 연주, 신촌에서 거친손?으로 반겨준 재희, 어느새 눈이

맞아 버린 동갑내기 경남, 재준. 센터 막내 우선이, 막내 혜진씨(아직 덜 친해짐). 어느 누구에게나 다 배울점이 있었습니다. 그리고 당신들과 한 시간들은 평생의 행운이었습니다.

파견? 더 많은 것을 공부하기위해 온 상경대 602호 식구들, 든든하고 자기 일을 잘했던 인욱형, 착하고 밥도 많이 먹고 열심히 사는 승희씨, 타락천사, 많은 대화를 나눴던 종철씨, 넉넉한 웃음이 좋았던 학신씨, 고생을 많이 했지만 오래 못있던 진영씨, 유진이, 까탈스런 선배 때문에 맘 고생 많이 했을 수진이, 군대가서도 건강해라 성준이. 눈 인사지만 따뜻히 대해준 상원씨, 응용 통계학과 식구들에게도 감사들 드립니다.

마주보고 대화했던 것 보다 손이 아프게 채팅했던 많은 동료들... 멋진형이라고 생각해주는 찬호씨, 삼성병원 특파원 홍희씨, 같이 졸업동기 시호씨, 모범생 한샘씨, 자신의 멋진 일을 찾을 혜진씨(화이링), 전산과 식구들에게 감사드립니다.

더 많은 것을 배우기 위해서 떠날 때 진정으로 그걸 이루도록 기도해준 연구소 식구들... 목소리 큰 허철구 선생님, 미국에 계실 첫 사수 양열렬 박사님, 월드컵 때 누구보다 들떠계셨던 고성호 박사님, 나의 10년후 모습이구 싶은 김창훈 박사님, 멋진 유머의 소유자 이병욱 선생님, 운동의 중요성을 강조해주신 이성훈 박사님, 진정한 사수 지협이 형, 자상한 누나 호진 누나, 성격 급한 경옥 누나, 멀리서 가장 많은 대화를 해주고, 장난치고, 놀아준 태형씨, 나의 체력을 알아버린 현민씨, 소녀 같은 아줌마 지혜씨, 한방식구 석표씨, 싸이 식구 진희씨, 동갑내기 찬이씨, 동갑, 동기 but 아줌마 선희씨, 몇 시간이고 수다 떨어도 지치지 않을 만큼 편했던 효수씨, 학교에서 더 친해진 희선이 형. 모든 연구소 식구들에게 감사드립니다.

졸업 논문이 완성되기까지 여러 모로 도움을 준 초,중,고 친구들..중학교 때부터 대멀이라는 별명을 가지고 있는 넉넉한 영성이, 국민학교땐 작았던 녀석이 이젠 엄청 커서 관악산에 터를 잡은 진수, 경제 상식, 영어 상식, 연애 상식이 풍부해 배울 것이 많은 영만(정신상태는 아님ㅋ). 까만 얼굴은 변함이 없는 호빈이, Go&Stop 회원 친구 녀석들에게 감사들 드립니다.

영원한 왕 형님 승무형, 친구처럼 편하지만 그래도 형님인 경진이 형, 한가지에 빠지면 미치도록 멋진 길조, 자존심이 강하고 따뜻한 경로, 나에게

매일 놀림을 받고 있지만 씩씩한 금숙, 술마시면 왕 애교 애영, 대학때나 지금이나 변함없는 승화, 나의 중매쟁이 영대, 나를 고생 많이 시킨 혁균, 유부남 상수 잘살아라.!!! 잠수중인 연도, 나의 팬 클럽.. 왕 누나 혜민누나, 얼마전 남친이 생긴 명천, 항상 푸근함과 자신이 일이 열심인? 아름, 앞으로의 중매쟁이 선영 ㅋㅋ 내가 불러 만줘도 좋아하는?? 근영, 공부를 열심히 해야할 태현, 애 아빠가 되고 연락이 뜸해진 진구씨, 중국 특파원 윤선, 소심?녀 진숙, 이름 안넣어주면 삐질 것 같은 효인, 민정, 열심히 공부하는 모습을 보여주신 이승연 교수님, 졸업하고도 반갑게 맞아주시는 차영준 교순님. 부드러운 성품의 안철환 교수님. 응용통계학과 형님들.. 왕 형님 지호형님, 농구로 친해진 희원이 형, 넉살 좋은 현석이 형. 집요한 은주, 소개는 못했지만 여러 가지 경험과 추억을 남겨준 세종대학 여러분들에게 감사드립니다.

제가 오늘날 이와 같은 논문이 완성 될 때까지 정말 많은 분들의 도움으로 된 것 같습니다. 여기에 소개되었던 분들도 감사드리지만 비록 소개는 되지 않았지만 같이 고생 많이 한 군대 전우들 또한 얼마되지 않은 아르바이트 기간에 만난 여러분들 그리고 저에게 작지만 도움을 주신 모든 여러분들에게 비록 미흡하지만 저의 논문을 바칩니다. 추후로 더욱 열심히 사회에 봉사하는 “김상철”이 되겠다는 약속을 드리겠습니다. 여러분들을 만나고 알게 된 것이 제 인생에 진정한 행복이고, 행운이 였음을 새삼 더 느낍니다.

감사합니다.

2004. 어느 늦 봄 신촌에서

김상철 올림

차 례

그림 및 표 차례 ····· iii
국문요약 ····· vi

Ⅰ. 서 론 ····· 1

Ⅱ. 재료 및 방법 ····· 4
1. 실험자료 수집 및 마이크로어레이 실험 ····· 4
가. 조직 표본 수집 ····· 4
나. Total RNA 추출 ····· 4
다. cDNA 마이크로어레이 실험 ····· 5
2. 자료의 표준화, 결측치 및 중복 유전자 처리 방법 ··· 5
가. 자료의 표준화 ····· 5
나. 결측치 처리 방법 ····· 10
다. 중복 유전자 처리 방법 ····· 11
3. 정상 조직과 암 조직 구분 유전자 검색 방법 ····· 12
가. 비모수적 t 검정과 족별 오류율 (Family-Wise Error Rate) ····· 12
나. SAM (Significance Analysis of Microarray) ···· 15
다. 경험적 베이즈 절차 (Empirical Bayesian Process) ····· 17
4. 유의적 유전자 판별 분석 ····· 20
가. 선형 판별 분석 (Linear Discrimination Analysis) ····· 20
나. 2차 판별 분석 (Quadratic Discrimination Analysis) ····· 21
다. SVM (Support Vector Machine) ····· 22

5. 유의적 유전자 기능 분류 ····· 24

Ⅲ. 결　과 ····· 26
1. 실험자료 수집 및 마이크로어레이 실험 ····· 26
2. 자료의 표준화, 결측치 및 중복 유전자 처리 방법 ··· 27
3. 정상 조직과 암조직 구분 유전자 검색 방법 ····· 31
가. 비모수적 t 검정과 족별 오류율 (Family-Wise Error Rate) ····· 31
나. SAM (Significance Analysis of Microarray) ···· 31
다. 경험적 베이즈 절차 (Empirical Bayesian Process) ·····32
라. 정상 조직과 암 조직 구분 유전자 검색 방법의 비교 ····· 33
4. 선별된 유의적 유전자들의 판별 분석 ····· 36
5. 유의적 유전자 기능 분류 ····· 37

Ⅳ. 고　찰 ····· 40

Ⅴ. 결　론 ····· 43

참고문헌 ····· 44

영문요약 ····· 48

그림 차례

그림 1. $\log_2 R$ *vs.* $\log_2 G$의 산점도 ····· 6

그림 2. M *vs.* A의 산점도 ····· 8

그림 3. 프린트 팁별 상자그림 ····· 8

그림 4. M-A 그림과 프린트 팁별 별 Loess 곡선 · 10

그림 5. SAM 결과 및 Delta value 조절 ····· 17

그림 6. 경험적 베이즈 절차 결과 (B *vs.* M) ····· 20

그림 7. 선형 판별 분석 ····· 21

그림 8. 2차 판별 분석 ····· 22

그림 9. SVM에서 자료와 커널 함수 관계 ····· 23

그림 10. DAVID 기능 분류를 위한 선택 조건 ····· 25

그림 11. 강도 의존적 Loess 적합 후 상자 그림 ·· 27

그림 12. 197자료 군집화 분석 결과 ····· 30

그림 13. 비모수적 t 검정과 족별 오류율, 경험적 베

이즈 절차, SAM 상위 3,516개 유전자 공통 집합 ····· 33
그림 14. 비모수적 t 검정과 족별 오류율, SAM, 경험적 베이즈 절차 상위 3,516개 유전자의 평균 vs. 분산 산점도 ····· 35
그림 15. 비모수적 t 검정과 족별 오류율, SAM, 경험적 베이즈 절차 상위 3,516개 유전자의 B통계량 vs. 평균 산점도 ····· 35
그림 16. 비모수적 t 검정과 족별 오류율에서 선별된 상위 유전자를 이용한 판별 분석 ····· 36
그림 17. 비모수적 t 검정과 족별 오류율에서 선별된 유전자의 기능 분류 ····· 38
그림 18. 경험적 베이즈 절차에서 선별된 유전자의 기능 분류 ····· 38
그림 19. SAM에서 선별된 유전자의 기능 분류 ···· 39

표 차례

표 1. 17K 마이크로어레이 상의 중복 스팟 수 ···· 11

표 2. 다중 검정 문제 ····· 12

표 3. 환자 자료 분포 ····· 26

표 4. 비결측치 비율과 유전자 수 ····· 29

국문요약

고밀도 마이크로어레이를 이용한 위암 관련 유전자의 통계적 규명

많은 수의 유전자를 동시에 관찰 할 수 있는 마이크로어레이의 개발은 생물학 뿐만 아니라 통계적 분석에도 새로운 도전이 되고있다. 본 논문에서는 위암 조직과 그 주위의 정상 조직에서 추출한 RNA를 사용한 마이크로어레이 실험 자료를 이용하여, 정상과 암 조직을 구별할 수 있는 유전자 군을 찾고자 하였다.

111명의 환자 중에서 86명의 환자는 정상과 암 조직의 쌍 자료, 13명에선 정상 조직 자료, 12명의 환자에선 암 조직 자료가 얻어져 총 197개의 자료를 이용하였다. 이중 86 쌍 자료를 비모수적 t검정과 족별 오류율 (Family Wise Error Rate), SAM (Significance Analysis of Microarray), 경험적 베이즈 절차 (Empirical Bayesian Process)를 이용하여 정상 조직과 암 조직을 구분할 수 있는 유의한 유전자를 선별하였다. 비모수적 t검정과 족별 오류율에서 조절 p-값이 0.0001보다 작은 경우에 3,516개의 유전자가 선별되었고, SAM의 경우엔 FDR (False Discovery Rate)가 0.00378%일 때 선별된 유전자는 5,657, 경험적 베이즈 절차에서 B값이 0 이상일 때 선별된 유전자는 6,661개이

었다. 비모수적 t검정과 족별 오류율에서 선정된 유전자를 순서를 고려하여, 13명의 정상 조직과 12명의 암 조직으로 구성된 시험군에 적용하여 선정된 유전자들의 암과 정상을 구분할 수 있는 타당성을 선형 판별 분석 (Linear Discrimination Analysis), 2차 판별 분석 (Quadratic Discrimination Analysis) 및 SVM (Support Vector Machine)을 이용하여 평가하였다. 판별 분석 결과 상위 2개의 유전자로 정확성이 0.96~1임을 확인하였다. 각 방법에서 선별된 유전자들의 기능은 생리학적 과정 (physiological process), 세포 과정 (cellular process), 발달 과정 (developmental process) 순으로 관련되어 있음을 생물정보학 데이터 베이스를 통하여 관찰하였다 (DAVID : Database for Annotation Visualization and Integrated Discovery, http://apps1.niaid.nih.gov/david/).

위암에서 정상 조직과 암 조직을 구분할 수 있는 유전자는 많이 존재하며, 그 중 몇 개의 유전자로도 새로운 자료를 판별 분석 할 수 있음을 통계학적 기법을 이용하여 확인하였다. 이와 같이 선별된 유전자 군들은 위암의 발생과 진행을 이해하는 자료일 뿐만 아니라 실제 임상에 사용할 수 있는 생물학적 표지자의 후보물질이 될 수 있는 중요한 정보를 제공할 수 있다고 생각한다.

핵심되는말 : 위암, 마이크로어레이, 통계학적 방법, 유전자 선별, 판별 분석, 데이터 베이스

고밀도 마이크로어레이를 이용한 위암 관련 유전자의 통계적 규명

<지도교수 **라 선 영**>

연세대학교 대학원 의과학과

김 상 철

Ⅰ. 서 론

사람의 사망 원인 중에서 암으로 사망하는 비율이 점점 증가하는 추세이다. 암 중에서 위암은 1980년대 초까지만 전 세계적으로 가장 많이 발생하는 암 중 하나였으나 꾸준히 감소하여 현재 선진국에서는 발병률이 낮아지는 추세이다[1]. 그러나, 위암은 우리나라에서 높은 발생률과 사망률을 가지는 암이며, 또한 성별에 관계없이 빈번하게 발생하는 암이다. 이러한 위암의 치료 방법으로는 외과적 치료, 화학적 치료가 병행하고 있으며 조기 진단과 치료방법의 발전으로 조기 위암에 대해서는 생존율이 증가 되고 있지만, 진행된 위암의 치료는 아직 미미한 수준에 이르고 있다[2]. 그러므로 적절한 치료와 정확한 예측을 위한 새로운 분자 생물학적 표지자의 발견과 이들을

특이적으로 억제하는 치료제 개발이 중요한 과제로 주목되고 있다.

암에서의 특이적인 유전자 발현은 매우 중요한 의미를 가지며 이들은 환자의 예후와 생존에 깊이 관련되어 있다. 이들 유전자의 발현은 유기적으로 연관되어 있기 때문에, 암의 병태생리를 이해하기 위해서 몇 개의 유전자 단위가 아니라 유전자군을 총체적으로 관찰하는 것이 필수적임이 제시되고 있다. 따라서 기존의 연구방법은 다수의 유전자 발현을 관찰하기에는 한계가 있기에 마이크로어레이(microarray)를 통한 관찰이 바람직하다.

마이크로어레이는 기존의 방법보다 많은 수의 유전자를 빠르고 효율적으로 밝혀낼 수 있는 장점을 가지고 있어 다양한 방면의 연구에 도입되고 있다[3]. 암의 연구에 있어서 마이크로어레이를 이용하여 다양한 암 조직과 정상 조직의 유전자 발현정도를 비교하여 암의 진행정도, 전이의 유무, 환자 예후의 예측, 조기 암과 악성 암의 차이에 대하여 규명한 결과가 보고되고 있다[4,5,6,7,8,9,10].

현재 마이크로어레이를 이용한 다양한 실험이 진행되며 많은 자료들이 공개되었고, 이러한 자료들을 기초로 통계학적인 분석이 이루어지고 있다. 그러나, 마이크로어레이는 자료의 특성상 표본(specimen) 수에 비해 많은 유전자들을 가지고 있는 문제점과 복잡한 실험과정에서 많은 변이를 유발할 수 있는 요인을 가지고 있어서, 마이크로어레이 자료의 특성에 맞는 통계학적인 이해와 분석법이

필요하다. 지난 몇 년간 몇 가지 표준적 분석 절차가 제시되었고[11,12,13], 암 연구에서 정상 조직과 암 조직을 구별하는 유전자 선별에도 다양한 통계적 방법들이 제시되었다[11,12,13,14,15]. 또한 다양한 판별 분석 방법들을 이용하여 정상과 암 조직을 구분할 수 있도록 선별된 유전자 군의 타당성을 평가하고자 하였다[16,17].

위암에서도 이러한 마이크로어레이를 이용한 실험 결과들이 일부 보고 되고 있다. 그러나 대개는 기본적인 통계적 분석 방법들이 사용하여, 정상 조직과 암 조직을 구별하는 유전자 선별에 사용된 통계적 방법들은 fold change, t-검정, 분산분석 등 전통적인 통계 분석이 이루어져왔다[18,19].

본 논문에서는 마이크어레이 실험에서 표준화 단계부터 적합한 표준화 방법을 선택하고, 위암 조직과 정상 조직을 구별할 수 있는 유전자를 다양한 통계적 방법을 이용하여 선별하고 각 방법이 어느 정도의 유사성을 가지고 있는지 확인하였다. 이렇게 선별된 유전자들은 여러 가지 분류기법으로 타당성을 평가하였다. 또한, 선별된 유전자를 생물정보학 데이터 베이스를 이용하여 어떤 기능을 하는지 기능별로 분류를 하였다.

Ⅱ. 재료 및 방법

1. 실험자료 수집 및 마이크로어레이 실험

실험자료 수집 및 마이크로어레이 실험은 연세 대학교 의과대학 세브란스 병원 및 암전이 연구센터의 표준 프로토콜로 수행했다.

가. 조직 표본 수집

연세 대학교 의과대학 세브란스 병원에서 위암으로 판정된 환자의 위암 조직과 주위 정상 조직을 각각 수술 당시에 수집하여 즉석 냉동하여 실험까지 -70℃ 냉동고에 보관했다.

나. Total RNA 추출

(1) Trizol (Invitrogen, USA)을 사용하여 암 조직 및 정상 조직에서 total RNA를 추출했다.

(2) Total RNA의 질을 높이기 위해서 RNeasy kit (Qiagen, Germany)를 사용하여 RNA를 정제했다.

(3) 정제된 RNA를 실험에 적합한 농도 (3ug/ul이상)로 만들기

위해서 Microcon YM (Millipore, USA)을 사용하여 RNA를 농축했다.

(4) 모든 RNA의 질과 양은 Gene Spec Ⅲ (Hitachi, Japan)과 Gel Documentation-Photo System (Vilber Lourmant, France)를 사용하여 측정했다.

다. cDNA 마이크로어레이 실험

Test sample과 reference RNA의 total RNA 40ug을 사용하여 test sample은 Cy5-dUTP로, reference sample은 Cy3-dUTP로 labeling했다. Reference RNA는 11개 장기를 대표하는 암 조직 주에서 추출한 RNA pool인 Yonsei Reference RNA (Cancer Metastasis Research Center)를 사용하였다. 각각 Cy3-dUTP와 Cy5-dUTP로 labeling된 cDNA를 정제하고 혼합한 후, 17,000개의 인간 유전자가 점적되어 있는 마이크로어레이 (GT-CMRC, KOREA)를 사용하여 보합을 시행하였다. 보합이 끝난 마이크로어레이는 washing하여 GenePix 4000B scanner로 스캔하고, 영상 데이터를 얻었다.

2. 자료의 표준화, 결측치 및 중복 유전자 처리 방법

가. 자료의 표준화

마이크로어레이의 각 유전자의 발현량은 결합 시킨 형광색에 따라서 녹색은 G로 표기되며 Cy3가 labeling되어 532 *nm* 의 파장으로 스캔하여 얻고, 빨간색은 R로 표기되며 Cy5가 labeling되어 635 *nm* 의 파장으로 스캔하여 얻는다. 분석에 필요한 두 값을 구하기 위해서 스팟에 발현 강도인 635, 532 *nm* 각각의 파장으로 얻은 자료의 foreground 중앙값에서 스팟 주변의 발현 강도인 background의 중앙값을 뺀 값을 실제 발현량 R, G 값으로 사용한다. 일반적으로 R, G 각각에는 로그를 취한다. 유전자의 밀집 정도를 색의 밝기 정도로 표현함으로써 자료가 어느 부분에 집중적으로 밀집되어 있는지의 해석을 용이하게 하였다[20] (그림 1).

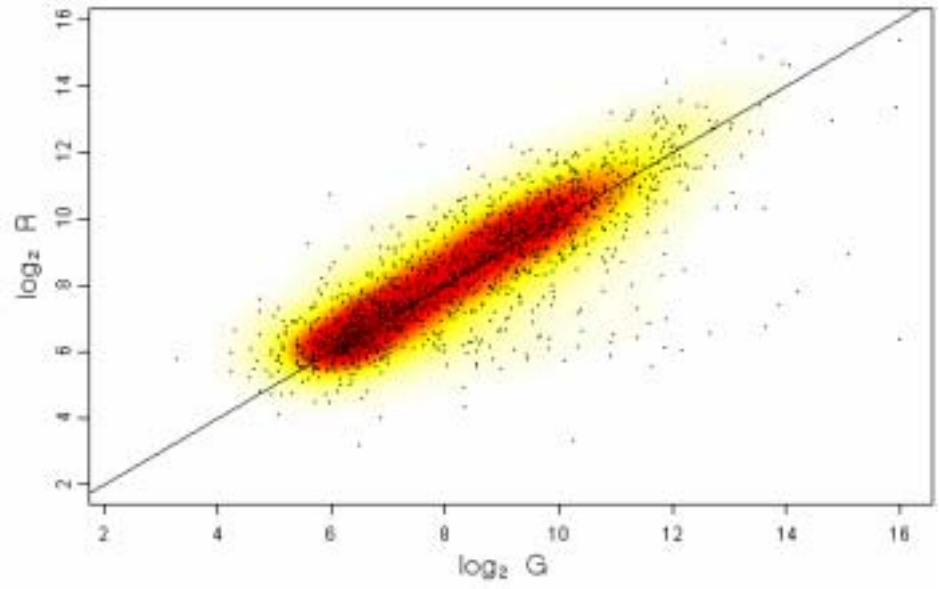


그림 1. $\log_2 R$ *vs.* $\log_2 G$ 의 산점도

$\log_2 R$ *vs.* $\log_2 G$ 그림을 시계 반대 방향으로 45도 틀어 놓은 그림으로 M-A 그림을 그릴 수 있다[11]. 단, 여기서 $M = \log_2(\frac{R}{G})$ 이고 $A = 0.5 * \log_2(RG)$ 로 정의된다. M은 로그로 변환한 상대 강도이고 A는 두 염료 강도의 평균을 나타내주므로 스팟의 강도를 나타내 준다. 즉, 그림 1의 자료를 대상으로 M-A 그림을 그리면 그림 2와 같다.

M-A 그림은 스팟의 이상점 식별이나 상대 강도가 강도 의존적으로 변하는 행태를 M-A 그림이 더 잘 나타내주고 있기 때문에 $\log_2 R$ *vs.* $\log_2 G$ 그림보다는 M-A 그림을 더 선호한다. 또한, 표준화를 구현하는데도 M-A 그림이 유용하다.

그림 2의 M-A 그림은 M값을 0을 중심으로 대칭형태를 보인다. 그렇지만, 혹 발생할 수도 있는 프린트 팁에 의한 공간 효과를 보여주지 못함으로 프린트 팁별 효과를 볼 수 있는 상자 그림을 통하여 프린트 팁별 효과를 판단하였다(그림 3).

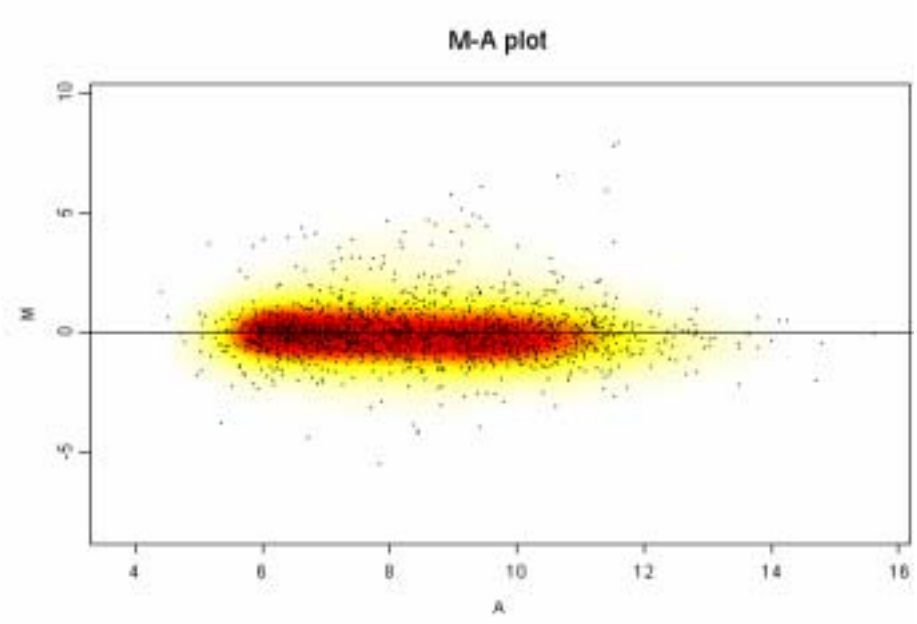


그림 2. M *vs.* A의 산점도

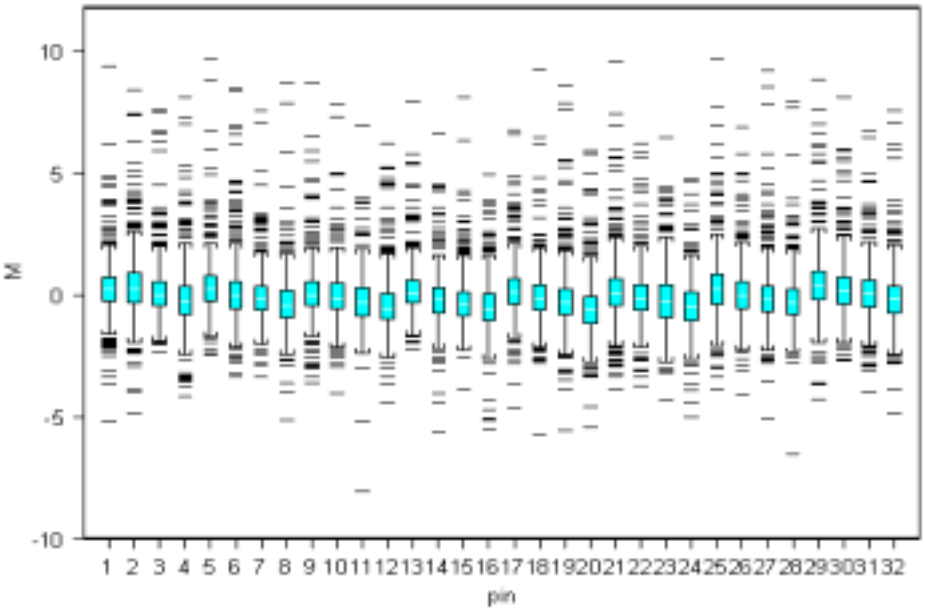


그림 3. 프린트 팁별 상자그림

그림 3은 각 프린트 팁별로 발현량의 범위 및 중심 경향을 보여준다. 마이크로어레이 실험에서 4의 배수 번째 프린트 팁들이 전반적으로 낮은 발현량을 보이는 주변 효과 (edge effect)를 보이는 것을 확인할 수 있다. 또한 강도 의존적 Loess 적합 방법을 각 프린트 팁 별로 적용하는 방법을 사용하였을 때 그림 4와 같이 프린트 팁별 평균값이 차이가 나고, 양 끝의 곡선의 형태의 차이를 보일 경우 프린트 팁에 의한 공간 효과와 강도 의존적 추세를 동시에 보정하여 주는 방법인 강도 의존적 Loess 적합 방법을 사용해야 한다. 식은 다음과 같다.

$$M' = \log_2(\frac{R}{G}) - c_i(A) = \log_2(\frac{R}{k_i(A) \cdot G}) \qquad \text{[식 1]}$$

식 1에서 $c_i(A)$는 i번째 프린트 팁에서 Loess 적합 값이다[21] .

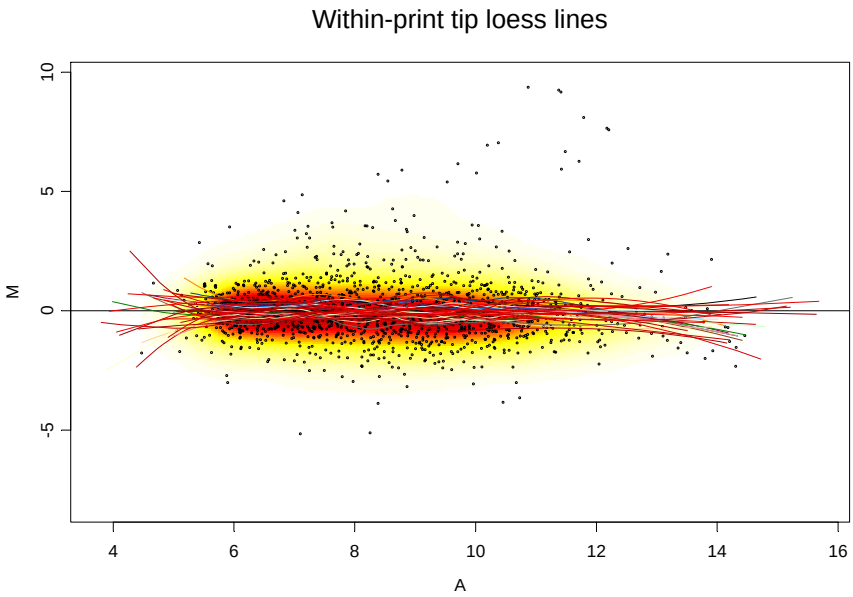


그림 4. M-A 그림과 프린트 팁별 Loess 곡선

나. 결측치 처리 방법

마이크로어레이 실험을 수행하다 보면 여러 기준치에 적합하지 못하여 결측치로 되는 스팟들이 존재한다. 수십개의 마이크로어레이 실험 후 어떤 유전자에서 불과 몇 개의 결측치로 인하여 유전자 전체 자료가 제거되는 것은 자료의 손실이 많다고 판단하여 적절한 방법으로 결측치를 대치하는 방법을 이용한다. 이러한 결측치를 대치하는 방법에는 크게 3가지가 있다. 첫째로 각 유전자의 결측치를 비결측치의 평균값으로 대치하는 방법이 있다. 둘째로 SVD (Singular Value Decomposition) 알고리즘을 이용한 방법으로, 이 방법은 선형

결합 근사로 가상 직교 발현 경향을 이용하여 값을 얻는다. 마지막으로 KNN (K-Nearest Neighbor) 방법으로 K를 고정시키고, 결측된 유전자와 결측되지 않은 유전자의 거리를 계산하여 가장 가까운 K개의 유전자들의 평균으로 결측치를 추정하는 방법이다. 본 논문에서는 두 집단으로 나눠진 자료형태에서 가장 좋은 결과를 보이는 KNN 방법을 사용하였다[22].

다. 중복 유전자 처리 방법

표 1은 17K 마이크로어레이에서 스팟들의 중복 정도를 나타낸 것이다. Blank를 제외한 나머지는 중복 스팟은 유전자이고, 이것은 실험의 안전성을 확인 할 수 있고, 스케닝을 수행할 때 기준을 잡을 수 있다. 표 1에서의 중복 스팟 수가 3, 4개의 유전자는 house keeping 유전자로 어느 상황에서도 발현량이 동일하며, 실제 분석에서는 1개의 유전자로 통합을 해야 된다. 일반적으로 중복된 유전자는 평균을 취하여 사용을 한다.

중복 스팟 수	1	2	3	4	589(Blank)
개수	14389	1318	10	5	1

표 1. 17K 마이크로어레이 상의 중복 스팟 수

3. 정상 조직과 암 조직 구분 유전자 검색 방법

가. 비모수적 t검정과 족별 오류율 (Family-Wise Error Rate)

고전적으로 어떤 유전자가 기준값을 벗어났는지를 판단하는 방법으로 t-검정을 사용하게 되는데, 마이크로어레이와 같은 수천 수만개의 유전자를 검정할 때 다중 검정이 문제가 된다. 예를 들어 10,000개의 유전자를 유의수준 0.05로 하여 검정을 하였을 때 10,000*0.05=500 적어도 500개의 유전자가 유의하다 판단될 수 있다. 이러한 다중 검정 문제를 보안하기 위해서 여러 가지 방법이 제시되었다.

	# not rejected	# rejected	
# true null hypothese	U	V	m_0
# non-true null hypotheses	T	S	m_1
	W	R	m

표 2. 다중 검정 문제

표 2는 m개의 유전자가 귀무가설 $M_0 = \{i : H_i = 0\}$ 과 대립가설 $M_1 = \{i : H_i = 1\}$ 을 만족할 경우에 $m_0 = |M_0|$, $m_1 = |M_1|$ (| |는

집합의 원소의 개수의 절대값)라 하면, V를 1종 오류 (false positives)라 하고, T를 2종 오류 (false negatives)라 한다. 다중 검정 상황에서 1종 오류에 관한 방법은 PCER (Per-comparison error rate), FWER (Family-wise error rate), FDR (False discovery rate), pFDR (Positive false discrovey rate)등이 있다. PCER방법은 1종 오류의 개수의 기대값을 전체 가설의 개수로 나눈 값으로, 표 2를 이용하면 $PCER = E(V)/m$ 이다. FWER은 1종 오류가 적어도 1번 이상 일어날 확률로, 표 2를 이용하면 $FWER = \Pr(V > 0)$ 이다. FDR은 기각역 중에서 1종 오류의 비율의 기대값이며, 표 2를 이용하면 $FDR = E(\frac{V}{R})$ 이다. 그렇지만 R이 0일때 FDR은 정의 되어지지 않으므로 R=0 인 경우는 $\frac{V}{R}$ 을 0으로 놓아서 다음과 같이 정의하였다.

$$FDR = E(\frac{V}{R} 1_{\{R>0\}}) = E(\frac{V}{R} \mid R > 0) \Pr(R > 0)$$

pFDR은 기각역이 0보다 클 때 1종 오류의 수를 기각역의 수로 나눈 기대값으로, 표 2를 이용하면 $pFDR = E(\frac{V}{R} \mid R > 0)$ 이다. 이러한 1종 오류에 관한 방법들의 비율을 비교하면 $PCER \le FDR \le pFDR \le FWER$ 순이 된다[23].

Westfall & Young[24] 이 제시한 비모수적 t검정과 족별 오류율 방법은 자료에 의한 p-value의 순서와 반복 순열을 수행하면서 가장 큰 유의수준을 이용하는 stepdown방법이고, maxT 조절 p-value는

검정 통계량의 순서가 $|t_{s_1}| \geq |t_{s_2}| \geq |t_{s_3}| \geq \cdots \geq |t_{s_{m-1}}| \geq |t_{s_m}|$ 일때 $\tilde{p}_{r_i} = \max_{k=1,\cdots,i}\{\Pr(\max_{l=k,\cdots,m} |T_{s_l}| \geq |t_{s_k}| \,\|\, H_M)\}$로 정의된다. 여기서 $H_M = \cap_{i=1}^{m}\{H_i = 0\}$ 이다. 알고리즘은 Ge 등의 논문에 다음과 같이 나와있다[23].

Permutation algorithm for step-down maxT adjusted p-values - based on Westfall & Young(1993) Algorithm.

For the original data, order the observed test statistics such that $|t_{s_1}| \geq |t_{s_2}| \geq \cdots |t_{s_m}|$. For the bth permutation, b=1, ⋯ ,B:

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \cdots, t_{m,b}$ for each hypothesis.
3. Next, compute $u_{i,b} = \max_{l=i,\ldots,m} |t_{sl,b}|$, the successive maxima of test statistics by

$$u_{m,b} = |t_{sm,b}|$$

$$u_{i,b} = \max(u_{i+1,b}, |t_{si,b}|) \quad \text{for } i = m-1, \cdots, 1.$$

The above steps are repeated B times and the adjusted p-value are estimated by

$$\tilde{p}^*_{s_i} = \frac{\#\{b : u_{i,b} \geq |t_{s_i}|\}}{B} \quad \text{for } i = 1, \cdots, m$$

with the monotonicity constraints enforced by setting

$$\tilde{p}^*_{s_i} \leftarrow \tilde{p}^*_{s_i} \;,\; \tilde{p}^*_{s_i} \leftarrow \max(\tilde{p}^*_{s_{i-1}}, \tilde{p}^*_{s_i}) \quad \text{for } i = 2, \cdots, m.$$

이 방법의 장점은 다중 검정을 하는데 일어나는 1종 오류를 유의 수준에서 통제할 수 있고, Bootstrap이나 순열 반복과 같은 방법을 통하여 각 원소들의 귀무가설을 추정 할 수 있다. 이와 같은 장점 때문에 마이크로어레이에서 각 유전자별 귀무가설을 추정하고, 다중 검정의 1종 오류를 통제할 수 있다. 그러나, 각 원소의 분포가 단봉 형태가 아니거나 원소들의 분포가 동일하지 않을 경우에는 이 방법을 사용할 때 정확성이 떨어진다[23]. R-언어에서 비모수적 t검정과 족별 오류율 방법를 Package로 제공한다(http://www.bioconductor.org).

나. SAM (Significance Analysis of Microarray)

SAM은 Tusher 등이[11] 이 제시한 방법으로 http://www-stat.stanford.edu/~tibs/SAM/ 웹사이트에서 다운로드 받을 수 있다. SAM방법은 이 표본, 일 표본, 대응 자료, 절단 자료 등 여러 가지 방법을 적용할 수 있다. 위에서 제시한 여러 방법의 귀무가설을

검정하는 순열 검정방법은 동일하나, 계산되는 통계량은 차이가 있다. 다음은 SAM에서 제시하는 이 표본에 대한 검정 통계량이다.

$$d_i = \frac{r_i}{s_i + s_0}; i = 1,2,\cdots,p \quad \text{[식 2]}$$

$$r_i = \bar{x}_{i2} - \bar{x}_{i1}, \quad \bar{x}_{i1} = \sum_{j \in C_1} \frac{x_{ij}}{n_1}, \quad \bar{x}_{i2} = \sum_{j \in C_2} \frac{x_{ij}}{n_2}$$

$$s_i = [(\frac{1}{n_1} + \frac{1}{n_2})\{\sum_{j \in C_1}(x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C_2}(x_{ij} - \bar{x}_{i2})^2\}/(n_1 + n_2 - 2)]^{1/2}$$

식 2는 t-검정과 유사하지만, s_0 라는 fudge factor를 분모에 추가하여 사용한다. 이유는 작은 M값을 갖는 유전자가 작은 분산을 가지면 귀무가설을 기각하게 되는 것을 보안하고, 분산의 영향을 줄이기 위하여 s_0 를 추가한다. 또한, 순열 회수를 100에서 5,000번까기 조절하면서, 예측치를 계산한다. 이때 유의한 유전자를 선택할 때 기준을 잡는 것은 FDR (False discovery rate)이다. FDR은 반복 수행 결과 잘못되었다 불리는 유전자의 중앙값을 유의하다 불리는 유전자 수로 나눈 값이다. 그림 5의 결과는 SAM을 수행한 결과이다. List Delta Table를 선택하여 Delta 값에 따른 FDR과 유전자수를 확인하여, 적정한 Delta값을 선택하면 가운데 점선들이 움직이면서

유의한 유전자들을 빨간점, 녹색점으로 나타낸다. 그리고 List Significant Genes을 선택하면 유의한 유전자의 정보를 얻는다[11] .

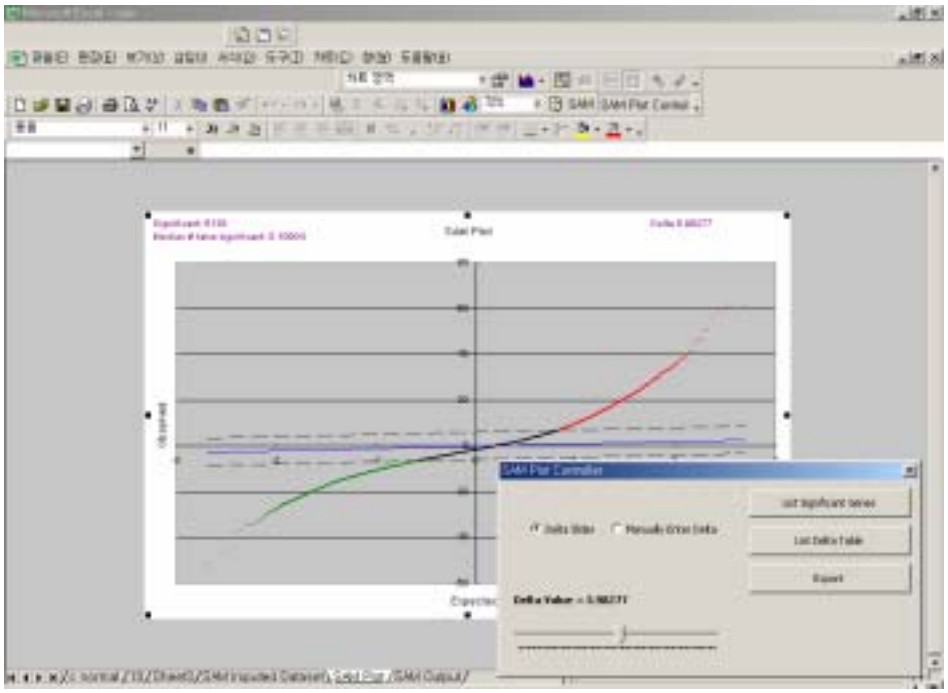


그림 5. SAM 결과 및 Delta value 조절

다. 경험적 베이즈 절차 (Empirical Bayesian Process)

베이지안 방법으로 로그 사후 승산비를 이용한 경험적 베이즈 절차는 다음과 같다. 자료에서 각 유전자들을 $M_{ij} \sim N(\mu_j, \sigma_j^2)$ 라하고, 여기서 i는 유전자, j는 표본을 나타낸다. 지수자 변수(I)를

사용하여 g번째 유전자에 대하여 발현의 차이가 있으면, I=1이라 정의하고, 아니면 0이라 정의한다.

$$I_g = \begin{cases} 0 & if \quad \mu_g = 0 \\ 1 & if \quad \mu_g \neq 0 \end{cases} \qquad [식\ 3]$$

로그 사후 승산비를 이용한 B-통계량은 다음과 같다.

$$B_g = \log \frac{\Pr(I_{g=1} \mid (M_{ij}))}{\Pr(I_{g=0} \mid (M_{ij}))} \qquad [식\ 4]$$

유전자들의 독립성을 가정하고 베이즈 이론을 적용하면 식 4는 다음과 같이 표현 할 수 있다.

$$\begin{aligned} B_g &= \log \frac{p}{1-p} \frac{\Pr((M_{ij}) \mid I_{g=1})}{\Pr((M_{ij}) \mid I_{g=0})} \\ &= \log \frac{p}{1-p} \frac{\Pr((M_g) \mid I_{g=1})}{\Pr((M_g) \mid I_{g=0})} \frac{\prod_{i \neq g} \Pr(M_i \mid I_g = 1)}{\prod_{i \neq g} \Pr(M_i \mid I_g = 0)} \qquad [식\ 5] \\ &= \log \frac{p}{1-p} \frac{\Pr((M_g) \mid I_{g=1})}{\Pr((M_g) \mid I_{g=0})} \end{aligned}$$

식 5에 여러 사전 정보들을 추정하여 마지막으로 얻을 수 있는

식은 다음과 같다.

$$B_g = \log \frac{p}{1-p} \frac{1}{\sqrt{(1+nc)}} [\frac{a + s_g^2 + M_{.g}^2}{a + s_g^2 + \frac{M_{.g}^2}{1+nc}}]^{\nu + \frac{n}{2}} \quad [식\ 6]$$

또한, 슬라이드 내에 반복된 유전자가 m개 존재할 때는 슬라이드 내의 분산과 슬라이드간의 분산을 이용하여 나타내면 다음과 같다.

$$B_g = \log \frac{p}{1-p} \frac{1}{\sqrt{(1+mnc)}} [\frac{a + \frac{1}{mn}(SSB_g + kSSW_g) + M_{g..}^2}{a + \frac{1}{mn}(SSB_g + kSSW_g) + \frac{M_{g..}^2}{1+mnc}}]^{\nu + \frac{mn}{2}}$$

[식 7]

식 7의 SSB_g, SSW_g 는 각각 슬라이드간 제곱합과 슬라이드내 제곱합을 나타낸다. 경험적 베이즈 절차는 R 언어 (www.r-project.org)에서 Package로 제공되고, 사용자의 입력 자료 형태만 일치하게 하면 프로그램화하여 이용할 수 있다. 그림 6은 경험적 베이즈 절차의 결과를 보여준다. Y축은 B값을 나타내고 x축은 M값의 평균을 나타내며 B값이 큰 경우가 더 유의적인 유전자라 판단할 수 있다[12] .

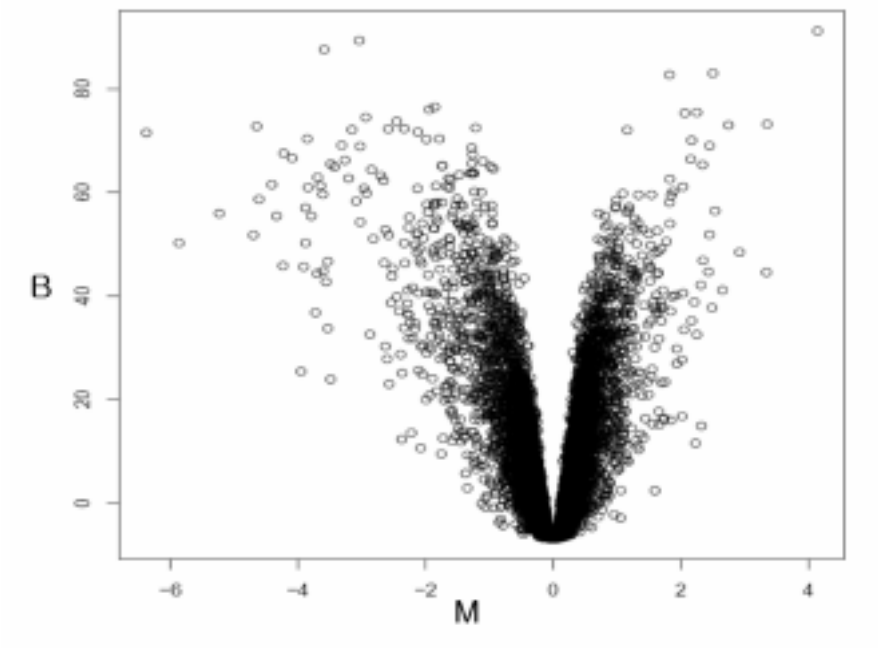


그림 6. 경험적 베이즈 절차 결과 (B *vs.* M)

4. 유의적 유전자 판별 분석

가. 선형 판별 분석 (Linear Discrimination Analysis)

선별된 유전자를 이용하여 새로운 자료에서 정상과 암 조직을 얼마나 정확하게 분류를 할 수 있는지의 타당성 검정 방법으로 선형

판별 분석 방법이 있다. 선형 판별 분석 방법은 그림 7와 같이 두 집단을 선을 이용하여 분류하는 방법이다.

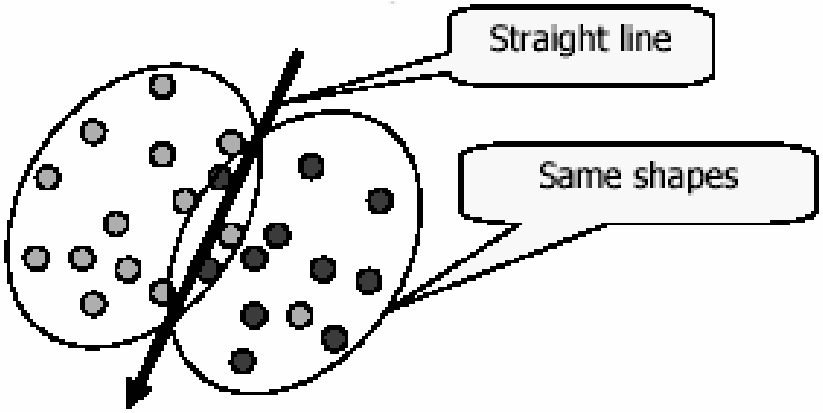


그림 7. 선형 판별 분석

K개의 집단이 있고, 각 집단의 평균과 공분산을 μ_k 와 Σ_k 로 표기하면, 선형 판별 분석 방법은 집단들의 공분산 행렬이 같다, $\Sigma_k = \Sigma$, 는 가정 하에서 식 8이 유도되어졌고 이 식을 이용하여 판별 분석한다[17] .

$$\ell(x) = \arg\min_k (x-\mu_k)\Sigma^{-1}(x-\mu_k)^T \quad \text{[식 8]}$$

선형 판별 분석 방법은 R-언어를 이용하여 프로그램화 하였다.

나. 2차 판별 분석 (Quadratic Discrimination Analysis)

2차 판별 분석 방법은 그림 8과 같이 두 집단을 2차 곡선을 이용하여 분류한다.

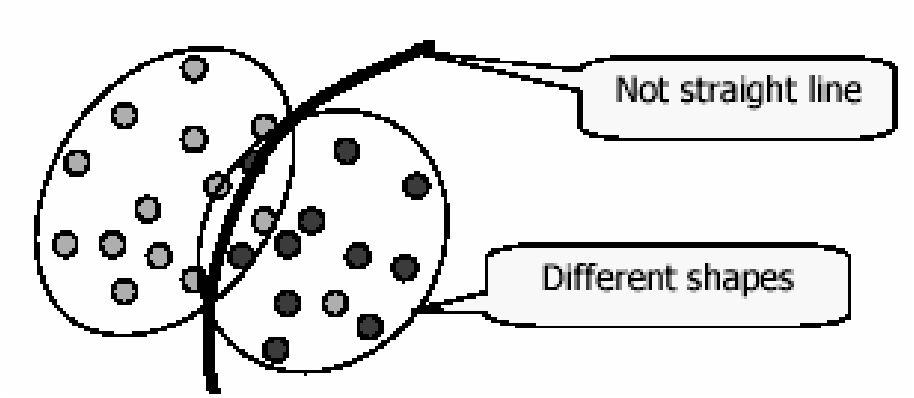


그림 8. 2차 판별 분석

2차 판별 분석 방법은 집단별 공분산 행렬이 서로 다르다고 가정하에서 식 9가 유도되어졌고 이 식을 이용하여 판별 분석한다[17].

$$\ell(x) = \arg\min_k \{(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^T + \log|\Sigma_k|\} \quad \text{[식 9]}$$

2차 판별 분석 방법은 R-언어를 이용하여 프로그램화 하였다.

다. SVM (Support Vector Machine)

SVM은 최근 기계 학습 분야에서 매우 각광 받는 방법으로, 2개의 그룹을 나누는 분류 방법이다. SVM은 그림 9과 같이 두개의 집단을

분리할 수 있는 최적의 초 평면 (hyperplane)을 찾아준다. 여기서 최적의 초 평면이란 두 집단에 속한 가장 가까운 자료와 거리가 최대가 되도록 하는 초 평면을 뜻한다. 자료들이 선형 분리가 불가능한 공간에 분포하고 있는 경우 커널함수를 사용하여 자료들을 선형 분리 가능한 고차원 공간으로 사상 시킨 후, 최적의 초 평면을 찾아주게 된다. 커널함수는 고차원 공간의 스칼라 곱에 대응된다고 정의되며, 커널함수의 종류엔 선형 커널함수, 다항 커널함수 등이 있다. SVM방법에 대한 프로그램은 http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html에서 제공하고 있으며, R-언어에서 Package로 제공한다[16].

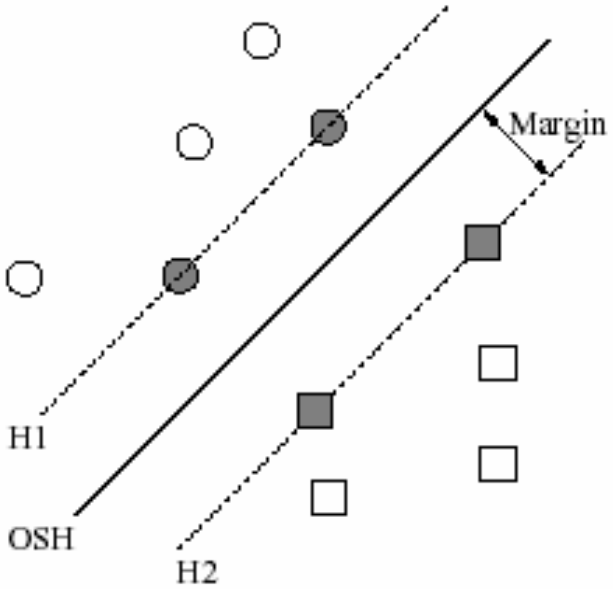
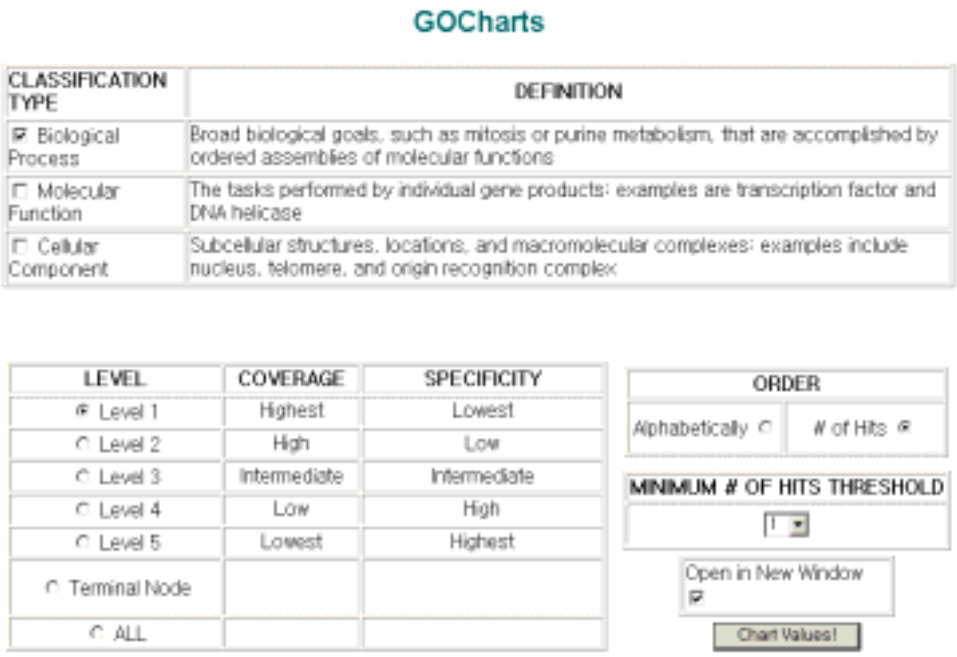


그림 9. SVM에서 자료와 커널 함수의 관계

5. 유의적 유전자 기능 분류

통계적 방법으로 선별된 유의한 유전자의 기능을 찾기 위해서 각 유전자마다 기존의 데이터 베이스를 이용하는 것은 시간적으로 많은 노력이 필요하며, 사용자의 주관에 따른 분류가 되기 때문에 오류가 생길 수 있다. 그렇지만 생물정보학을 이용하여 만들어진 데이터 베이스를 이용한다면 이러한 시간적, 주관적인 문제를 해결할 수 있다. 본 논문에서는 DAVID (Database for Annotation Visualization and Integrated Discovery; http://apps1.niaid.nih.gov/david/)를 이용하여 유전자들의 기능을 분류하였다. DAVID에 접속하여 선별된 유전자 리스트를 전송을 하면 그림 10과 같은 여러 조건을 선택할 수 있는 창이 나타난다. 여기서 사용자가 자신에게 적합한 분류를 선택하여 수행을 하면 분류에 따른 결과가 나타난다.



GOCharts

CLASSIFICATION TYPE	DEFINITION
☑ Biological Process	Broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions
☐ Molecular Function	The tasks performed by individual gene products; examples are transcription factor and DNA helicase
☐ Cellular Component	Subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex

LEVEL	COVERAGE	SPECIFICITY
◉ Level 1	Highest	Lowest
○ Level 2	High	Low
○ Level 3	Intermediate	Intermediate
○ Level 4	Low	High
○ Level 5	Lowest	Highest
○ Terminal Node		
○ ALL		

ORDER	
Alphabetically ○	# of Hits ◉

MINIMUM # OF HITS THRESHOLD

1

Open in New Window ☑

Chart Values!

그림 10. DAVID 기능 분류를 위한 선택 조건

Ⅲ. 결과

1. 실험자료 수집 및 마이크로어레이 실험

연세 대학교 의과대학 세브란스 병원에서 위암으로 판정된 111명의 환자의 위암 조직과 주위 정상 조직을 이용하여 연세의대 암전이 연구센터의 프로토콜에 의거하여 실험을 하였다.

Hybridization		#. number
(Reference, Normal tissue)	(Reference, Tumor tissue)	
Yes	Yes	86
Yes	Missing	13
Missing	Yes	12

표 3.환자 자료 분포

실험 결과로 표 3과 같이 111명의 환자의 정상 조직과 암 조직을 이용하여 마이크로어레이 실험을 하였다. 86명의 환자는 정상 조직과 암 조직의 total RNA의 양이 충분하여 모두 실험하였다. 반면 25명의 환자에서는 13명은 암 조직의, 12명은 정상 조직의 total RNA의 양이 부족하거나 조직 RNA가 분해되어 실험을 시행하지 못하여 쌍이 아닌

자료를 얻었다. 최종적으로 197개의 마이크로어레이 실험 자료를 얻었다.

2. 자료의 표준화, 결측치 및 중복 유전자 처리 방법

그림 2, 3, 4를 통하여 표준화의 필요성을 확인했다. 가장 접합한 표준화 방법으로 강도 의존적 Loess 적합 방법이라 판단하여 197개의 마이크로에레이에서 수행을 하였다. 표준화를 수행한 결과는 그림 11와 같다. 그림 3와 같이 기존의 프린트 팁별 효과를 제거하여 모든 프린트 팁에서 동일성을 유지하도록 하였다[21].

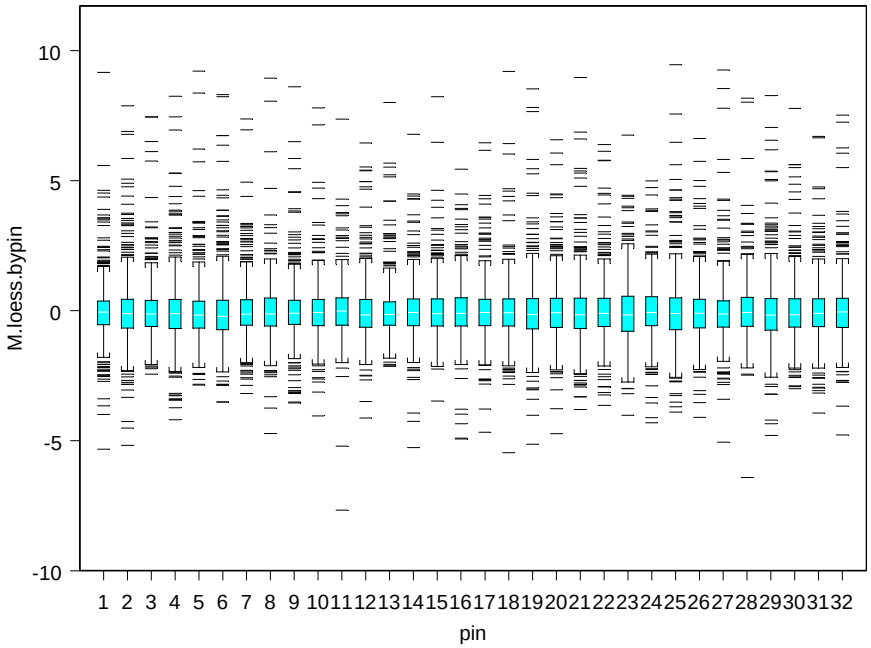


그림 11. 강도 의존적 Loess 적합 후 상자 그림

표 4의 결과를 이용하여 비결측치 비율을 감소하면서 스팟의 개수 증가 정도를 계산하였을 때 80%부터 안정적으로 증가함을 확인하고, 비결측치 비율을 80%로 사용하였다. 이렇게 결정된 결측치를 SAM에서 제공하는 KNN (K-Nearest Neighbor) 방법으로 결측치를 보정하였다. 이때 사용된 K는 10으로 하였다[22].

결측치 보정을 한 후 얻어진 자료는 15,274*197 형태의 자료를 얻었고, 실제 분석을 위하여 중복된 유전자는 평균값으로 치환을 하여 사용하였으므로, 최종적으로 얻어진 자료는 14,124개의 유전자와 197실험 자료를 얻었다. 그림 12은 197개의 자료를 계층적 군집화 분석을 통하여 어떤 경향을 보이는지 확인한 결과이다. 크게 3집단으로 나눠짐이 관찰되며, 1개의 정상 집단과 2개의 암 집단으로 구분된다. 이를 통해서 정상 집단과 암 집단은 전체적인 유전자 발현 차이가 있음과 또한 암 집단 내에도 차이가 존재한다는 것을 확인했다[14].

비결측치비율(%)	스팟 수	비결측치비율(%)	스팟 수
100	9960	50	15975
95	14142	45	16082
90	14776	40	16178
85	15069	35	16245
80	15274	30	16315
75	15433	25	16384
70	15554	20	16443
65	15671	15	16518
60	15770	10	16599
55	15846	5	16696
		0	17664

표 4. 비결측치 비율과 유전자 수

.

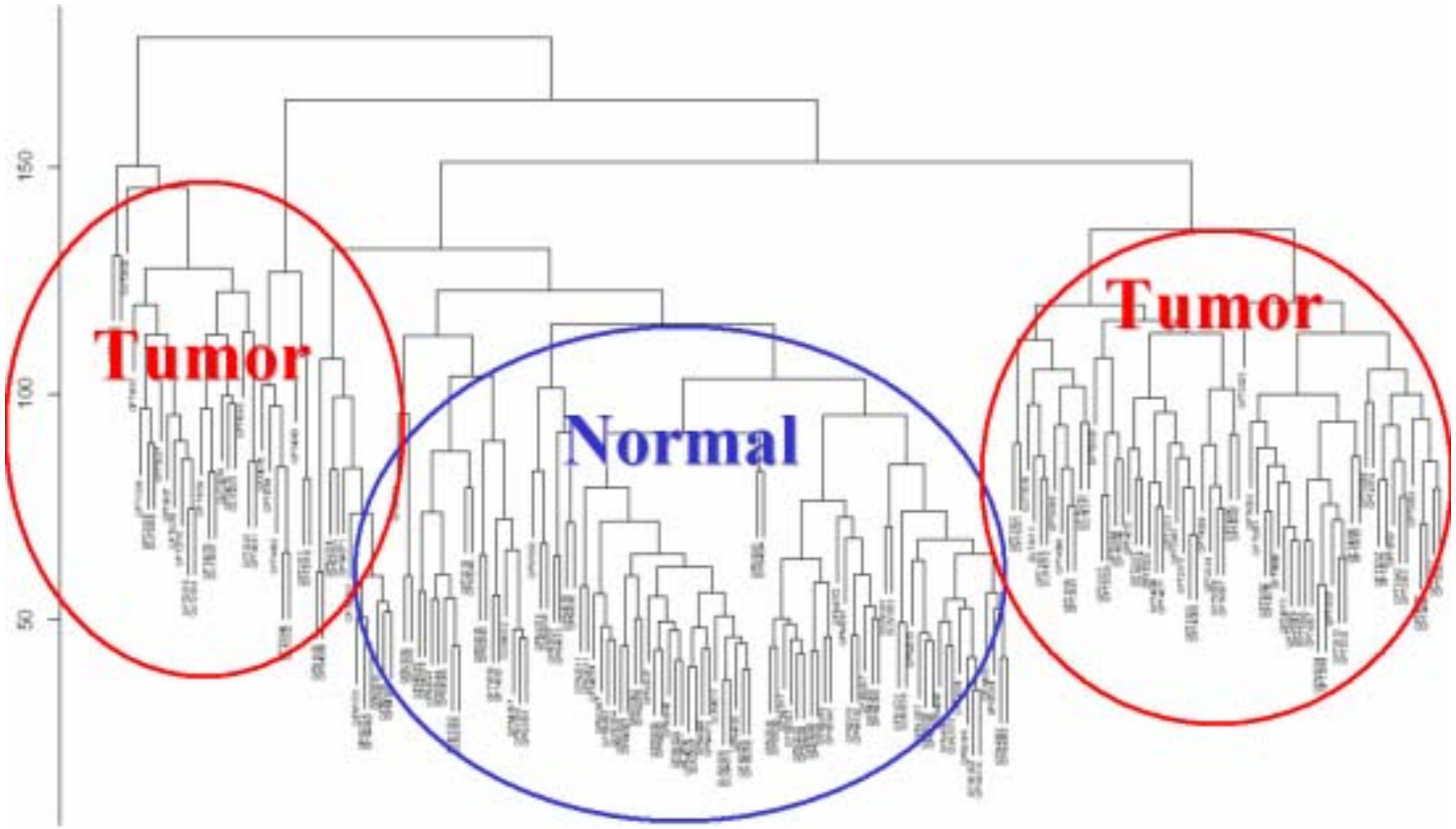


그림 12. 197자료 군집화 분석 결과

3. 정상 조직과 암 조직 구분 유전자 검색 방법

86명의 쌍자료를 대상으로 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차 방법을 이용하여, 정상 조직과 암 조직 구분 유전자를 검색하였다.

가. 비모수적 t검정과 족별 오류율 (Family-Wise Error Rate)

R-언어에서 Package로 제공되는 함수를 이용하여 자료의 특성에 적합하게 프로그래화하여 86명의 쌍 자료를 이용하여 비모수적 t검정과 족별 오류율 방법을 수행하였다. 순열조합을 10,000번 시행한 결과, 조절 p-value가 0.0001보다 작은 유전자가 3,516개이었다. 보수적인 방법인 비모수적 t검정과 족별 오류율 방법을 이용했을 시, 정상과 암 조직을 구분하는 유전자가 작은 p-value에서도 많이 선별되므로 정상 조직에서 암 조직으로 변화가 되면 유전자 변화가 많이 일어남을 관찰하였다.

나. SAM (Significance Analysis of Microarray)

SAM은 엑셀로 개발되어서 사용자가 쉽게 사용할 수 있도록

되어있고, 최근 생물학자들이 가장 많이 사용하는 방법 중에 하나이다. SAM에는 여러 가지 방법들이 있는데 본 논문에서 수행한 방법은 paired data 방법을 사용하였고, 순열조합은 5,000번을 시행하였다. 결과는 그림 5과 같이 그림과 수치(FDR)를 이용하여 사용자가 적정 수준을 선택할 수 있었다. SAM을 수행한 결과 FDR (False Discovery Rate)가 가장 작은 0.00378%일 때 선별된 유전자는 5,657개이다. SAM방법을 이용했을 시 비모수적 t검정과 족별 오류율 방법과 유사하게 정상과 암 조직을 구분하는 유전자가 많이 선별됨을 관찰하였다.

다. 경험적 베이즈 절차 (Empirical Bayesian Process)

경험적 베이즈 절차는 R-언어에서 Package로 제공되는 함수를 이용하여 자료의 특성에 적합하게 프로그래화하였다. 프로그램을 통하여 여러 과정을 수행한 후 그림 6과 같은 결과와 수치화 자료를 얻었다. B-통계량이 0 이상일 때 선별된 유전자는 6,661개이었다. 경험적 베이즈 절차 방법을 이용했을 시 비모수적 t검정과 족별 오류율 방법, SAM방법과 유사하게 정상과 암 조직을 구분하는 유전자가 많이 선별됨을 관찰하였다.

라. 정상 조직과 암 조직 구분 유전자 검색 방법의 비교

그림 13는 위에서 제시한 3가지 방법들에서 유의한 유전자가 가장 작은 개수가 선별된 비모수적 t검정과 족별 오류율을 기준으로 상위 3,516개의 유전자를 선택하여 3가지 방법이 얼마나 유사성을 보이는지 확인한 그림이다. 3,296개의 유전자는 3가지 방법 모두에서 공통적으로 나왔고, 비모수적 t검정과 족별 오류율과 경험적 베이즈 절차 방법은 3,472개가 공통 유전자로 두 방법이 상당이 유사한 결과를 보인다. SAM의 경우엔 비모수적 t검정과 족별 오류율, 경험적 베이즈 절차와 대부분 공통 유전자를 가지지만 178개의 유전자가 차이가 있었다.

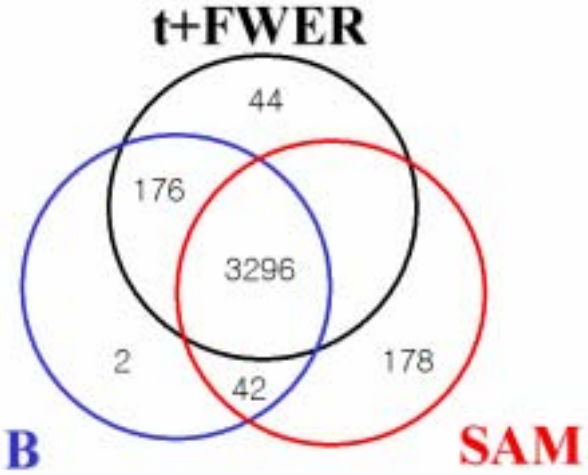


그림 13. 비모수적 t검정과 족별 오류율, 경험적 베이즈 절차, SAM 방법의 상위 3,516개 유전자 공통 집합

그림 14, 15은 쌍 86환자의 암 자료 발현량에서 정상 자료의 발현량을 뺀 평균값을 이용하여 3가지 방법에서 상위 3,516개의 유전자들에 대한 그림이다. 그림 14는 한 환자의 암 자료에서 정상 자료를 뺀 값의 평균 (Mean)과 분산 (Var)에 대한 그림으로 일반적으로 3가지 방법 모두 평균은 크고 분산은 작은 경우에 유의한 유전자로 선별이 되어진다. 그러나, 평균이 크다고 해도 분산이 클 경우에는 유의한 유전자로 선택이 안된다. 비모수적 t검정과 족별 오류율, 경험적 베이즈 절차의 경우에는 SAM보다 작은 평균을 가져도 분산이 작을 경우에 선별되어지는 유전자가 존재한다. 반면에 SAM의 경우엔 분산이 크더라도 평균이 크면 선별된다. 즉, 비모수적 t검정과 족별 오류율, 경험적 베이즈 절차에서 선별된 유전자가 SAM에서 선별된 유전자 보다 자료가 안정적이라 할 수 있다. 그림 15의 경우에는 평균값의 절대값이 커짐에 따라 B값도 커지는 경향을 확인했다. 또한 3가지 방법 모두에서 선별된 유전자는 B값이 크며, SAM의 방법에서 선별된 유전자는 상대적으로 작은 B값에서도 선별되어짐을 관찰하였다.

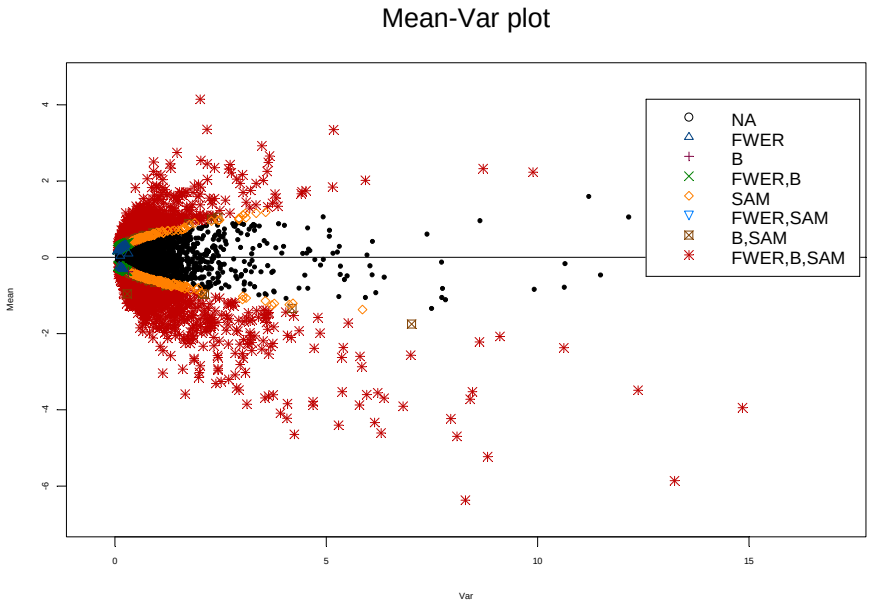


그림 14. 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차 방법의 상위 3,516개 유전자의 평균 vs. 분산 산점도

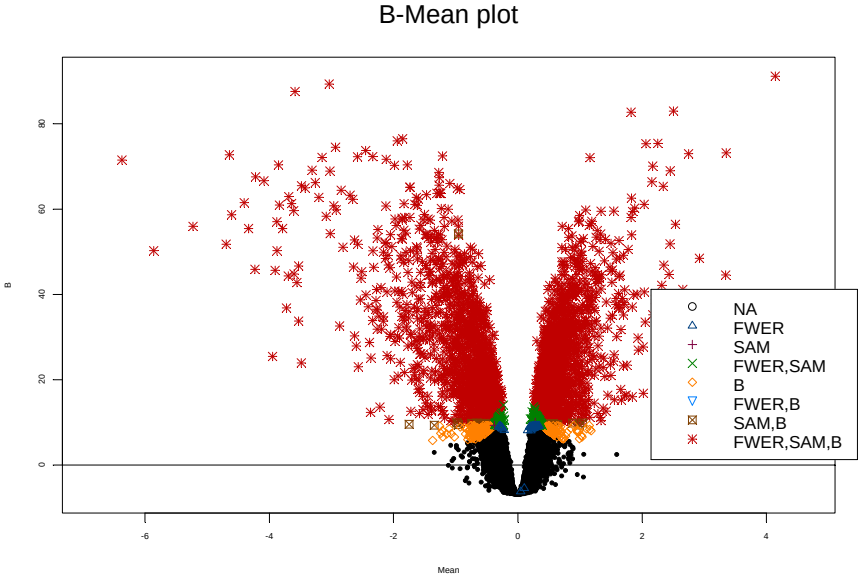


그림 15. 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차 방법의 상위 3,516개 유전자의 B통계량 vs. 평균 산점도

4. 선별된 유의적 유전자들의 판별 분석

비모수적 t검정과 족별 오류율 방법으로 선별된 유전자를 쌍 자료가 아닌 정상 13개, 암 12개의 시험군 자료를 이용하여 선형 판별 분석, 2차 판별 분석, SVML, SVMP 방법을 시행하였다.

그림 16은 86개의 쌍 자료에서 비모수적 t검정과 족별 오류율에서 가장 큰 t-값 순으로 유전자를 1개씩 증가 시키면서 25개(정상13, 암12)의 자료를 이용하여 판별 분석한 결과이다. 4가지 분류 모두 최소 2~9개의 유전자로 정상과 암을 구분할 수 있다는

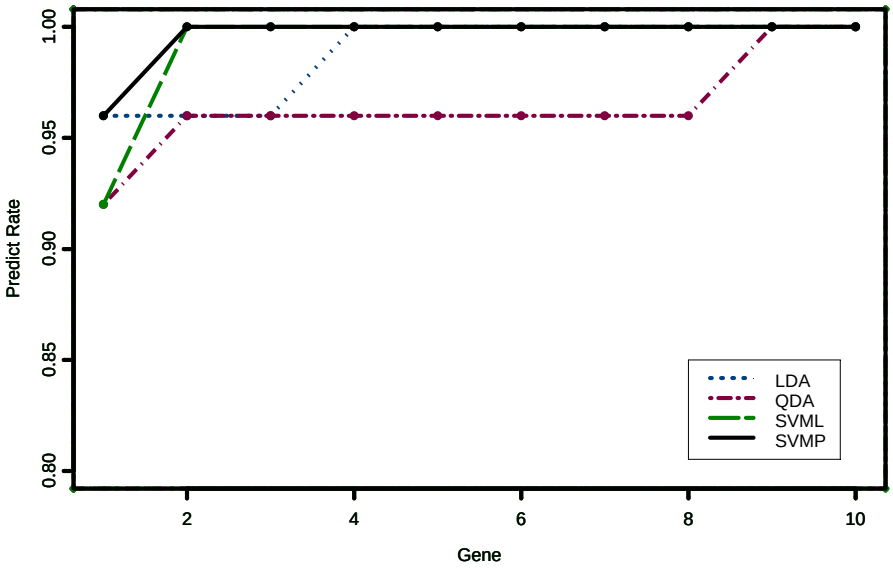


그림 16. 비모수적 t검정과 족별 오류율에서 선별된 상위 유전자를 이용한 판별 분석 결과

결과를 확인하여 위암에서는 암과 정상을 구분을 위해서 필요한 유전자의 수는 10개 보다 작을 수 있다는 것을 관찰했다. 4가지 방법 중에서 비모수적 t검정과 족별 오류율 방법으로 선별된 유전자를 이용하여 SVML 방법, SVMP 방법이 2개의 유전자로도 정상과 암을 구분할 수 있는 결과로 가장 좋은 결과를 얻었다. 선형 판별 분석은 4개의 유전자로 구분이 가능하다는 결과를 얻었고, 2차 판별 분선 방법은 4가지 방법 중 가장 많은 유전자를 사용하여야 정상과 암을 구분할 수 있는 결과를 보임을 확인하였다.

5. 유의적 유전자 기능 분류

3가지 방법에서 유의하게 선별된 유전자를 Genebank와 같은 데이터 베이스를 통하여 각 유전자의 기능을 찾는 것은 불가능하므로 생물정보학을 이용한 데이터 베이스인 DAVID를 이용하여 유전자들의 기능을 분류하였다. 그림 10에서 Classification Type에서는 Biological Process를 선택을 하고, 적용범위는 가장 높게, 특수성은 가장 낮게, 분류의 최소 개수의 유전자를 1개까지 적용하였다.

그림 17은 비모수적 t검정과 족별 오류율에서 선별된 유전자를 DAVID로 수행한 결과이다. 결과에서 나오는 유전자 수가 입력한 유전자 수보다 큰 이유는 한 유전자가 2개 이상의 기능을 가진

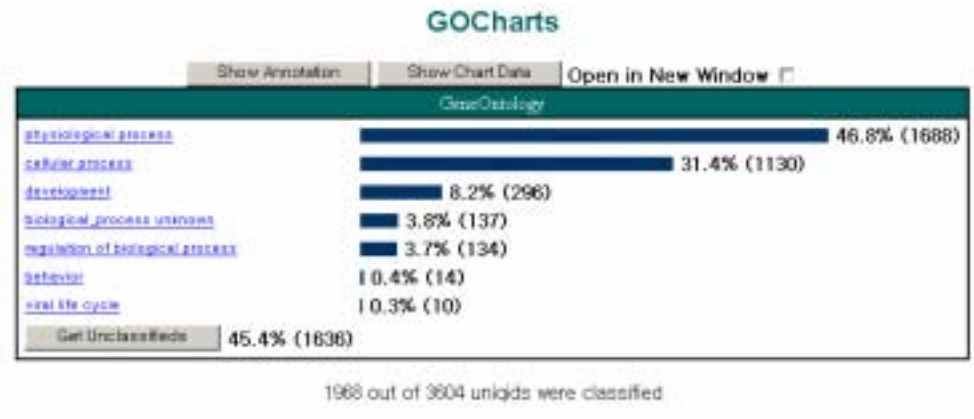


그림 17. 비모수적 t검정과 족별 오류율에서 선별된 유전자의 기능 분류

유전자는 각각을 따로 분류를 하였기 때문이다. 분류 결과 생리학적 과정 (physiological process)가 46.8%로 가장 많고, 세포과정 (cellular process)은 31.4%로, 발달 (development) 8.2% 순으로 구분되었다.

그림 18. 경험적 베이즈 절차에서 선별된 유전자의 기능 분류

그림 19. SAM에서 선별된 유전자의 기능 분류

그림 18, 19는 경험적 베이즈 절차와 SAM에서 선별된 유전자를 DAVID를 이용하여 기능을 분류한 결과로 비모수적 t검정과 족별 오류율의 분류 결과와 유사하였다. 경험적 베이즈 절차의 결과와 비모수적 t검정과 족별 오류율의 결과에서는 기능이 알려지지 않은 유전자가 약 45.2~4%로 유사하게 나왔지만, SAM의 결과에서는 44.2%로 기능이 알려진 유전자가 많이 존재함을 관찰하였다.

Ⅳ. 고찰

위암에서 마이크로어레이를 이용한 실험 결과들이 일부 보고 되고 있으며, 이러한 보고들은 아주 기본적인 통계학적 방법들을 사용하였다. 데이터의 표준화시 house keeping 유전자를 이용한 표준화, 전체 데이터의 중앙값을 0으로 보정하는 기본적 표준화가 사용되었다. 정상 조직과 암의 조직을 구별하는 유전자 선별에 사용된 통계적 방법들은 fold change, t-검정, 분산분석 등 전통적인 통계 분석이 이루어져왔다. 하지만 이제까지 제시된 많은 방법들 중 어느 방법이 최선의 방법인지 제시가 불가능한 현실이다.

본 논문에서는 마이크로어레이 자료의 특성에 맞는 통계학적인 이해와 분석법의 필요성을 고려하여 기존의 다른 논문에서 사용되었던 전체적인 표준화 방법 뿐만 아니라 표준화 단계부터 어떤 표준화를 사용하는 것이 모든 자료의 오차를 줄이고, 동일한 실험 조건을 유지하는지를, M-A plot과 상자 그림, 프린트 팁별 Loess 곡선 등을 이용하여 확인하였다. Yang 등[21] 이 제시한 모든 방법을 사용하여서 그 중 본 실험 결과를 가장 좋게 표준화 할 수 있는 방법을 선택하였다.

결측지 보정 과정에서는, 결측치를 증가하면서 유전자 증가폭이 수치적으로 안정적인 값을 선택하여 정보 손실을 줄이는 가장 적합한

값을 추정하고, 두 집단으로 나눠졌을 때 다른 보정 방법보다 더 좋은 결과를 보이는 결측치 보정 방법 (KNN)을 수행하였다. 중복된 유전자는 평균치로 치환을 하여 분석에 용이하게 하였다.

위암조직과 정상 조직을 구별할 수 있는 유전자를 최근에 여러 통계학자들이 제시하고 있는 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차 방법을 이용하여 선별하였다. 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차 3가지 방법의 결과 위암에서 정상과 암 조직을 구분할 수 있는 후보 유전자가 많다는 것을 확인하였다. 비모수적 t검정과 족별 오류율, SAM, 경험적 베이즈 절차의 3가지 방법에서 선별된 상위 3,516개의 유전자 중 대부분의 유전자 3,296개가 동일하게 선별되었다. 그렇지만, 그림 14에 비모수적 t검정과 족별 오류율, 경험적 베이즈 절차는 평균이 작고, 분산이 작은 경우에 선별되었지만, SAM의 경우 다른 방법 보다 평균이 크고, 분산이 클 경우에 선별되어진다. 즉 SAM방법이 분산에 영향을 많이 받는다는 것을 확인하였다. 3가지 방법에서 공통적으로 선별된 유전자가 다른 유전자들보다 우선적으로 추후 확인 연구의 대상이 될 수 있는 유전자라고 할 수 있을 것이다.

비모수적 t검정과 족별 오류율로 선별된 유전자를 선형 판별 분석, 2차 판별 분석, SVML, SVMP 방법을 이용하여 판별 분석을 한 결과 몇 개의 유전자로 위암과 정상을 정확히 분류됨을 관찰하였다. 이것으로

위암에서 단지 몇 개의 유전자를 이용하여 정상과 암을 구분할 수 있음을 확인하였다.

또한, 다양한 통계적 방법으로 선별된 많은 유전자를 생물정보학 데이터 베이스 (DAVID)를 이용하여 유전자를 기능별로 구분을 하였다. 선별된 유전자들은 선별 방법과 상관없이 생리학적 과정 (physiological process)이 가장 많고, 세포 과정 (cellular process), 발달 (development) 순으로 구분되었다.

본 논문에서는 유전자 각각을 이용하여 선별하였지만, 추후엔 유전자간의 연관성을 고려한 방법을 적용함이 필요하고, 생물학적 표지자 개발을 위한 분석 방법을 통하여 진단 칩 개발에 이용할 수 있는 유전자 선별 방법을 사용할 수 있다. 또한, 다양한 분류 방법을 이용하여 좀더 정확한 검정도 필요하다. 이렇게 통계학적으로 검정된 유전자들을 생물정보학 데이터 베이스를 이용하여 기능을 확인할 수 있었으나, 자세한 생물학적 기능을 검정 하는 것이 필요하다. 기능 검정뿐만 아니라 유전자들의 경로 분석을 통한 연관성 검정도 추후의 과제 중 하나이다. 즉, 본 논문에서 다룬 통계학적인 분석은 생물의 기능 및 경로를 연구하는 분야인 생물정보학의 시작점이라 할 수 있다.

Ⅴ. 결론

본 논문은 위암에서 정상과 암을 구분할 수 있는 후보 유전자군을 찾고자 하였다. 다양한 통계적 방법으로 후보 유전자를 찾았고, 여러 가지 판별분석 방법으로 확인하였다. 또한, 후보 유전자들의 생물학적 기능을 확인하기 위하여 생물정보학 데이터 베이스를 이용하여 분류를 하였다. 그러나, 이러한 것은 단지 통계적 방법으로 선별되었기에 생물학적으로 자세한 검정이 필요하다.

Ⅵ. 참고 문헌

1. Howson CP, Hiyama T, Wynder EL. The decline in gastric cancer: epidemiology of an unplanned triumph. Epidemiol Rev 1986;8:1-27.
2. Roukos DH. Current status and future perspectives in gastric cancer management. Cancer Treatment Rev 2000;26:243-255.
3. Yoshitaka H, Hirokazu T, Shuichi T, Naoka M, Ja-Mun C, Masashi F, et al. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res 2002;62:233-240.
4. Alex B, Hong L, Jia L, Paul W, Stephen L, Andrew JH, et al. Distinctive Patterns of gene expression in premalignant mucasa and gastric cancer. Cancer Res 2003;63:2569-2577.
5. Sakakura C, Hagiwara A, Nakanishi M, Shimomura K, Takagi T, Yasuoka R, et al. Differential gene expression profilies of gastric cancer cells established from primary tumour and malignant ascites. Brit J Cancer 2002;87:1153-1161.
6. Hiroshi I, Ayumi M, Koshi M, Hiroaki U, Masaki M. Prognostic socre of gastric cancer determined by cDNA microarray. Clin Cancer Rec 2002;8:3475-3479.

7. Jianhua W, Shishu C. Screening and identification of gastric adenocarcinoma metastasis-related gene using cDNA microarray coupled to FDD-PCR. J Cancer Res Clin 2002;128:547-553.

8. Marjan MW, Ernst JK, Cindy P, Antoine MS, Ivar S, Daniel P. et al. Genemic profiling of gastric cancer predicts lymph node status and survival. Oncogene 2003;22:1870-1879.

9. Suguru H, Yoichi F, Meihua L, Seiji S, Tatsushi K, Takeshi W, et al. Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23,040 genes. Cancer Res 2002;62:7012-7017.

10. Liu LX, Liu ZH, Jiang HC, Qu X, Zhang WH, Wu LF, et al. Profiling of differentially expressed genes in human gastric carcinoma by cDNA expression array. World J Gastroentero 2002;8(4):580-585.

11. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 2001;98:5116-5121.

12. Lonnstedt I, Speed T. Replicated microarray data. Stat Sinica 2002;31-46.

13. Dudoit S, Yang YH, Callow MW, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sinica 2002;12:111-139.

14. Eisen MP, Spellman PT, Brown PO, Botstein D. Clustering analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 1998;95:14863-14864.

15. Tamayo P, Slonim D, Mesirov J, Xhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with selp-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci 1999;96:2907-2912.

16. Terrence SF, Nello C, NigelD, David WB, Michel S, David H. Support vector machine classification and validation of cancer tissue samples using Microarray expression data, Bioinformatics 2000;16:906-914.

17. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 2002;97:77-87.

18. Lee S, Baek M, Yang H, Bang YJ, Kim WH, Ha JH, et al.

Identification of genes differentially expressed between gastric cancers and normal gastric mucosa with cDNA microarrays. Cancer Lett 2002;184:197-206.

19. Suganuma K, Kubota T, Saikawa Y, Abe S, Otani Y, Furukawa T, et al. Possible chemoresistance-related genes for gastric cancer detected by cDNA microarray. Cancer Sci 2003;94:355-359.

20. Eilers PHC, Goeman JJ. Enhancing scatterplots with smoothed densities. Bioinformatics 2004;20:623-628.

21. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normlization for cDNA microarray data:a robust composite method adderssing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.

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

23. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Technical report 2003;1-41.

24. Westfall PH, Young SS. Resampling-based multiple testing : Examples and methods for p-value adjustment. John Wiley & Sons 1993;116-117.

Abstract

Statistical Identification and Validation of Gastric Cancer-Related Genes Using High-Density cDNA Microarray

Sang Cheol Kim

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Sun Young Rha)

The development of microarray that simultaneously monitors thousands of genes has new challenge on statistics as well as biology. Using microarray data with RNA extracted from gastric cancer and normal tissue around it, this study aims to search for the genes discriminating between normal and cancer tissue using various statistical methods.

The study used total of 197 data among 111 patients - paired data of normal and cancer tissue from 86, normal from 13, and cancer from 12 patients, respectively. Using 86 paired data as a

training set, significant genes discriminating normal tissue from cancer tissue were selected in a manner of Family-Wise Error Rate (FWER), SAM (Significance Analysis of Microarray) and Empirical Bayesian Process. Number of significance genes were 3,516 in the FWER when adjust p-value is less than 0.0001, 5,657 in the SAM when FDR (False Discovery Rate) is 0.00378%, and 6,661 in the empirical Bayesian process when B-statistic is more than 0, respectively. To validate the predictive accuracy of the selected genes from FWER, LDA (Linear Discrimination Analysis), QDA (Quadratic Discrimination Analysis) and SVM (Support Vector Machine) were used in a separate test set of 25 unpaired samples. As a result top 2 genes showed the prediction rate of 0.96~1. The selected genes from each method were categorized into physiological, cellular, and the developmental process according to the Database for Annotation Visualization and Integrated Discovery (DAVID, http://apps1.niaid.nih.gov/ david/).

Significant number of genes to discriminate a normal and cancer tissue of the gastric cancer were selected with various statistical methods. These selected gene set might serve as an important database for understanding the carcinogenesis and progression of

gastric cancer, and the potential biomarker candidates for the clinical application.

Key Words : gastric cancer, microarray, statistical method, gene selection, discriminant analysis, database