Array-based CGH 자료에서 유의한 유전자를 찾는 회귀분석 방법

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 신 영

Array-based CGH 자료에서 유의한 유전자를 찾는 회귀분석 방법

지도 김 동 기 교수

이 논문을 석사학위 논문으로 제출함

2004년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 신 영

김신영의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2004년 12월 일

감사의 글

의학통계학이라는 새로운 분야의 학문을 접한 후의 2년이란 시간은 참으로 빨리 흘렸습니다. 그 2년 동안 저는 학문에 있어서의 진보와 더불어 많은 인생경험을 할 수 있는 시간이었습니다. 이렇게 귀중한 시간을 저에게 허락해 주신 하나님께 감사드립니다.

의학통계학이란 길을 제시해 주시고 이끌어 주신 김동기 교수님께 진심 어린 감사의 말씀을 드립니다. 바쁘신 가운데서도 자상하면서 꼼꼼하게 통계학에 대한 원리와 연구방법을 지도해 주신 조진남 교수님과 유전통계에 대한 길을 보여주시고 학문의 목적을 정확하게 알 수 있게 해 주신 임길섭 교수님 그리고 통계학에 대해 진지하게 생각하고 고민 할 수 있게 해 주신 이학배 교수님께도 감사드립니다. 학부 때 통계학에 대한 관심을 갖게 해 주시고 기초를 쌓을 수 있게 해 주신 존경하는 김강균 교수님과 김부용 교수님께도 이 지면을 빌어 감사의 말씀을 드립니다.

인생의 모범이 되시며 항상 하나님의 뜻에 따라 살기를 실천하시며 모든 고민과 어려움 그리고 기쁨과 슬픔을 함께 할 수 있는 사랑하는 아버님과 어머님께 그 은혜를 다 보답 할 수 없지만 부모님께 자랑스런 딸로 살 수 있도록 열심히 노력하겠다는 약속과 함께 가장 큰 힘이 되어 주심에 엎드려 감사의 말씀을 드립니다. 또한 어려서부터 사랑을 너무 많이 받았는데도 제대로 감사의 말씀을 드리지 못했던 고모님들께도 감사의 말씀을 드리고 싶습니다.

새로운 길을 준비하고 있는 동생 신혜와 이제 막 사회로 진출하려는 신진이에게 그들의 길을 축복하고 어떤 일에도 나를 믿고 함께 해 주었던 동생들에게 감사의 말을 전합니다.

학문에 있어서 뿐 아니라 여러 가지 고민이나 문제들을 항상 자신의 일처럼 함께 고민하며 그 길을 제시해 주었던 기준오빠에게 진심으로 감사드립니다. 함께 지내면서 미운정 고운정이 든 성민오빠, 2년 동안 같은 방에서 여러모로 도움을

많이 준 무영오빠에게도 감사의 말씀을 드리고 싶습니다. 소리없이 도와주는 미영언니와 의학통계학과에 진학할 때 많은 도움을 주시고 또한 공부하는 모습에서나 일 하는 모습에서나 모범이 되어주신 찬미언니께도 감사의 말씀을 드립니다.

학부 선배이자 대학원 동기인 수옥 언니와 대학원 동기 은혜, 그들이 있어서 대학원 생활이 너무 즐거웠고 함께 할 수 있었던 것에 감사합니다. 앞으로의 삶 속에서도 서로 이끌어 주면서 힘이 되는 우리가 되었으면 좋겠습니다.

이 외에도 혜리씨, 소연과 민진 그리고 성은이에게 함께 이 길을 걸어감에 있어 서로에게 좋은 사람으로 남았으면 좋겠다는 말을 전하며 많이 도와줘서 고맙다는 말도 함께 하고 싶습니다. 다시 함께 일하게 된 원열 오빠께도 감사드립니다.

중학교 때부터 내 옆에서 함께 했던 친구 미숙이, 고등학교 때부터 지금까지 서로의 인생길은 다르지만 힘이 되어주고 용기를 주었던 친구 영민이와 영실이, 그리고 대학때 수학통계학부에서 함께 공부하며 우정을 나누었던 나연,지현,유미,재인 에게도 감사의 말을 전하고 싶습니다. 중국에서 고생하며 비젼을 향해 열심히 자신의 길을 가는 기도친구 준이와 기도제목을 나누며 함께 기도해 주었던 일산동안교회 청년부 사람들에게도 진심으로 감사를 표현하고 싶습니다.

내가 흔들릴 때마다 붙잡아 주고 용기를 주었던 소중한 사람 태상오빠에게 사랑과 감사를 드립니다. 바쁜 생활 가운데서도 인생에 있어서의 조언과 사랑을 아끼지 않은 오빠가 옆에 있어 든든하고 힘이 되었습니다.

랄프 왈도 에머슨은 성공을 자주 그리고 많이 웃는 것, 지성있는 사람들로부터 존경을 그리고 아이들로부터는 애정을 받는 것, 정직한 비평가들로부터 평가를 받고 거짓된 친구들의 배신을 참아내는 것, 아름다움을 감상하고 타인들이 가진 최상의 것을 발견하는 것, 세상을 조금 더 좋은 곳으로 만드는 것이라 정의하였습니다. 이제 졸업과 함께 또 하나의 시작을 하고자 합니다. 세상을 조금 더 좋은 곳으로 만들 수 있는 성공한 하나님의 사람이 되고자 노력하겠습니다.

2004년 12월

김 신 영 올림

차 례

그 림 차 례

표 차 례

국문요약

Array-based CGH 자료에서 유의한 유전자를 찾는 회귀분석 방법

Microarray 자료에서 유의한 유전자를 찾는 기존의 분석 방법을 array-based CGH 자료에 그대로 적용하는 것은 자료의 특성상 제약적인 면이 있다. 본 연구에서는 array-based CGH자료에서 통계학적으로 유의한 유전자를 찾기 위하여 가중최소 제곱법(weighted least square)에 기초한 회귀분석(regression) 방법을 제시하였다. 제시된 방법에서는 이분산성 일치적 공분산 행렬(HCCM; Heteroscedasticity Consistent Covariance Matrix)을 이용하여 분산을 추정하고 회귀계수의 유의성을 검정하였다. 단순 2단계 회귀분석, 단순 반복법 그리고 반복법의 세 가지 방법으로 유방암 자료(breast cancer data)를 분석 하였을때 유전자간, 샘플간의 이질성을 보정하는 정도가 단순 2단계회귀분석에 비해서 단순 반복법과 반복법의 경우가 더 좋았으며, 유의한 유전자의 선택에서는 단순반복법과 반복법의 경우 더 보수적인 경향을 보였다.

핵심되는 말 : Array-based CGH, 가중최소제곱법, 이분산성 일치적 공분산 행렬, 단순 2단계 회귀분석, 단순 반복법, 반복법

제 1장 서론

1. 1. 연구배경

휴먼게놈프로젝트에서 발표된 DNA서열은 인간의 DNA서열과 상당수 유전자의 위치정도를 파악한 것이며 유전자별 기능에 대해서는 알려지지 않았기 때문에 발암관련 유전자의 수적 변화를 감지하는 것은 암의 발병 원인을 연구하고 진단하며 나아가 이를 치료하기 위한 개인별 맞춤의학을 실현 시킬 수 있다는데 의의가 있다. 암의 발병 원인에 대해서 완벽하게 설명할 수는 없으나 흡연, 식이요인, 바이러스나 박테리아에 의한 감염 등의 환경적요인과 유전자의 변이에 의한 생물학적 요인을 들 수 있다. 환경적 요인에 대해서는 정확한 원인을 설명하기 어렵지만, 생물학적 요인은 다양한 유전적 변이들이 일어남으로써 암의 발달과 전이가 특정 유전자 및 염색체 부위의 수적, 구조적 변화를 수반한다는 사실이 여러 연구를 통해 증명되어 왔다. 이렇게 변이가 일어나는 유전자 및 염색체를 찾기 위한 하나의 방법으로 microarray실험이 개발되었다.

암뿐 아니라 여러 질환들을 살펴보면 특정 유전자의 과다 발현이나 과소 발현에 의하여 발생하는 경우도 있는데, 근본적으로 염색체상의 유전자 수가 늘거나 없어져서 생기는 질환이 상당수 존재한다. 기존의 microarray분석 에서는 유전자의 발현변화를 조사하고 있으나 이는 염색체상의 유전자 변이의 탐색은 불가능하다. 이미 유전자의 증폭이나 결실이 암화(tumorigenesis)를 유발한다는 많은 연구 결과들이 있으며, 여기에 비교 게놈 교잡법(이하 CGH: Comparative Genomic Hybridization)은 염색체상 특정부위의 증폭과 결실을 연구하는 방법으로 쓰이고 있다. 근래에는 기존의 CGH에 비해 높은 해상력을 가지면서 CGH에 microarray 원리를 접목하여 대용량 분석을 가능하게 한 cDNA microarray based CGH(이하 array-based CGH)방법으로 염색체상의 유전자 변이의 탐색과 발현변화를 연구하고 있다. 실험의 방법이 변하면서 실험의 특성상 분석방법도 달라지고 있는데 기

존에 microarray 자료에서 쓰이는 실험적(exploratory techniques) 분석으로는 군집분석(cluster analysis)이 가장 보편적으로 사용되고 있다(Eisen et al., 1998; Tomayo et al., 1999). 군집분석은 유전자나 샘플을 집단화 하여서 다중(multiple) array의 구조(portrait)를 알아내는 방법이나, 군집분석은 이런 형태의 연구에 민감한(sensitive) 방법은 아니다. 왜냐하면 이것은 집단간의 유사성(similarities)은 알 수 있으나 각 유전자의 차이(difference)는 알 수 없기 때문이다. 고전적 통계학적 분석방법에는(classic statistical approach) t-검정(Dudoit et al., 2000), ANOVA (Kerr et al., 2000) , 회귀분석(Thomas et al., 2001; Zhao et al., 2001), 윌콕슨 순위 합 검정등이 제안되어 사용되었다. 그러나 array-based CGH실험 자료에서는 자료 구조상 위의 방법들을 그대로 적용하는 것은 통계학적 기본가정에 위배가 되어 분석을 하는데 문제가 제기 되었다.

1. 2. 연구목적 및 방법

Microarray 자료에서 유의한 유전자를 찾는 기존의 분석방법을 array-based CGH 자료에 그대로 적용하는 것은 자료의 특성상 제약적인 면이 있다. 본 연구에서는 array-based CGH실험 자료에서 유의한 유전자를 찾는 방법으로 가중최소제곱법에 기초한 회귀분석 방법을 통하여 구체화 하고자 한다. 자료의 구조상 통계학적 기본가정에 위배가 되는 부분 즉, 모든 유전자와 샘플에서의 정규성(normality)과 등분산성(constant variance)이 성립 되지 않고, 샘플 간에 이질성(heterogeneity)이 있으며 다중 비교(multiple comparison)에서 높은 false-positive rate 때문에 유의확률(p-value)을 그대로 사용하는데 문제가 있다는 것 등을 고려하였다.

Thomas et al.(2001)과 Cheng et al.(2003)이 제시한 회귀모형을 바탕으로 가중최소제곱법을 이용하였고, 제시된 방법에서는 이분산성 일치적 공분산 행렬을 이용하여 분산을 추정하고 회귀계수의 유의성을 검정하였다.

본 연구에서 제시하고자 하는 방법은 기존의 분석 방법에 비해 보다 효율적으로 이용할 수 있고 통계학적 가정에 충실하게 적합 시켰다는 측면에서 그 의미를 갖고 있다고 할 수 있다. 본 논문에서 다루게 될 내용들은 다음과 같다. 먼저 microarray 와 CGH 그리고 array-based CGH실험에 관한 설명과 개발된 분석 방법을 간략히 소개하고 이어서 사용된 통계학적 이론과 개념을 언급하고, 본 연구의 핵심인 구체적인 방법론을 논의 한다. 아울러 제시된 분석 방법의 평가를 위해 실재 자료를 이용하여 분석한 세 가지 방법의 결과를 비교, 해석하고 토론한다.

제 2장 cDNA microarray실험과 cDNA microarray CGH

2. 1. Microarray

2. 1. 1. cDNA microarray의 소개

분자생물학적 실험기술의 발달과 공학(robotics) 기술의 발달로 등장 하게 된 DNA microarray 혹은 마이크로칩(microchip)은 유전자 형별 분석이나 EST(Expressed Sequence Tag)의 발현정도를 동시에 관찰할 수 있는 기술이다. 유전자 증폭 기술이 발달하면서 PCR(Polymerase Chain Reaction)이라는 유전자 증폭 기술을 이용하여 유전자를 복제 할 수 있게 되었다(Bej et al.,1991). 이런 유전자들을 유리 혹은 나일론 판 위에 행렬의 형태로 찍어놓은 것을 microarray 라고 한다. cDNA microarray 경우 DNA 클론들을 PCR에 의해 증폭하고 정제 한 후 슬라이드 위에 일정크기로 찍어서(spotting) 칩(chip)을 제작한다. 칩은 수천개의 DNA를 작은 공간에 고밀도로 붙여 놓은 것으로 이것을 통하여 전체 유전체(genome)의 발현양상을 탐색 할 수 있게 되었고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰이 가능하게 되었다. 유전자가 반응을 나타내는 것을 유전자의 발현(expression)이라고 하는데, microarray 실험은 각 처리에 따른 유전자의 발현도의 차이를 측정하는 실험이다. 그 실험과정을 간단히 살펴보면 다음과 같다.

2. 1. 2. cDNA microarray의 실험방법

실험군과 대조군에서 mRNA를 추출하여 RT-PCR(Reverse Transcription Polymerase Chain Reacion) 방법을 통해 cDNA로 역전사(Reverse Transcription)

과정을 통하여 다른 색깔의 형광 시료로 염색하여 빨강색(Cy5) 과 녹색(Cy3)을 띤 cDNA를 합성한다. 염색된 cDNA를 혼합하여 제작해놓은 칩에 결합(hybridization)을 시킨다. 여러 번 세척을 하여 반응하지 않은 cDNA를 제거한 후에 레이저 형광 스케너(laser fluorecence scanner)를 통하여 각 유전자의 형광정도를 읽는데 이는 유전자의 발현정도를 알려주는 것으로 Genepix, Imagene과 같은 이미지처리 소프트웨어를 이용하여 수치화 하게 된다. (그림 1)

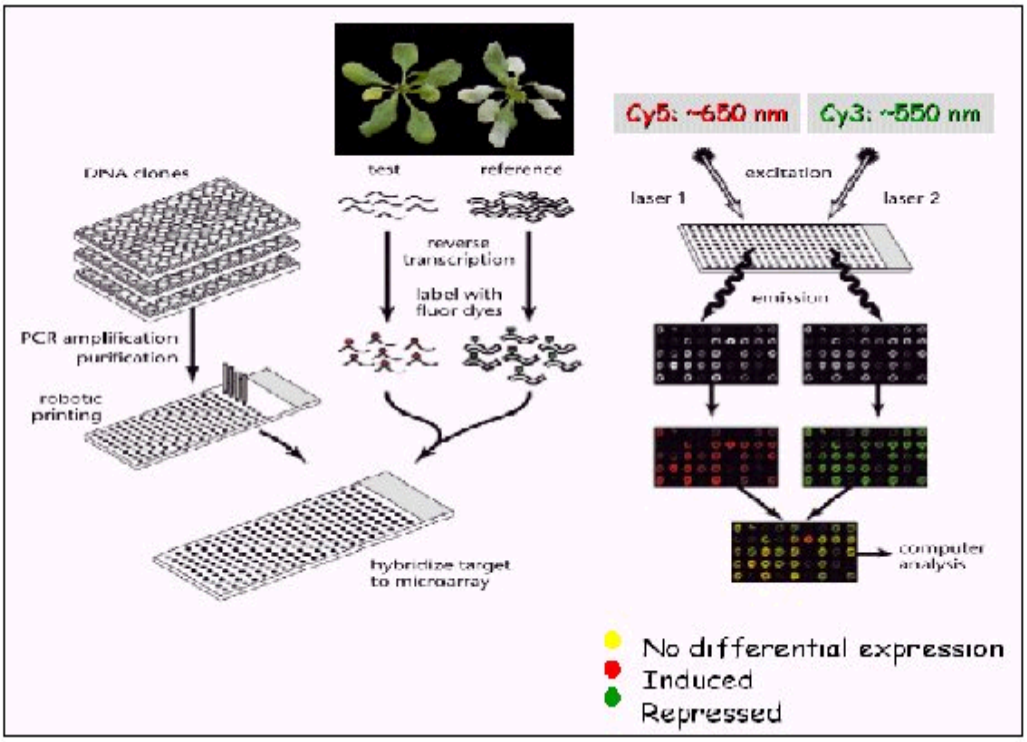


그림 1. cDNA microarray 실험과정

2. 1. 3. cDNA 칩

DNA 칩의 개념은 1995년 stanford university의 Pat Brown에 의해서 약 2-3천개의 유전자를 1㎠ 넓이 안에 집약시킨 형태로 처음 개발되었다. 초기엔 유전자의 발현변화 측정을 목적으로 cDNA 칩을 만들었지만 이후에 돌연변이를 검색 할 수 있는 oligonucleotide 칩도 개발되었다. DNA 칩은 길이 500bp이상의 유전자를 붙여 대량의 유전자 발현 확인을 목적으로 하는 cDNA 칩과 20개 내외의 염기들을 붙여 돌연변이 진단등 주로 DNA분석을 목적으로 하는 oligonucleotide 칩으로 분류되어 사용되고 있다.

2. 1. 4. cDNA microarray 자료분석

Microarray 실험은 수천 개의 유전자의 발현을 동시에 관찰할 수 있고 또한 유전자들 상호간의 관계를 알아낼 수 있는 방법이다. 이 방법은 각 샘플 마다 실험군과 대조군에 대해 염료처리를 하고 두 처리 간에 발현도가 유의한 차이를 보이는 유전자를 찾아내는 것이 주 목적이다. 초기에는 유의한 유전자를 찾는 분석방법으로 t-test ,윌콕슨 순위 합 검정(wilcoxon rank sum test, Snedecor and Cochran, 1980)등을 사용하여 분석하였다. 현재는 SAM, SMA, ANOVA 모형과 함께 베이지안 기법을 이용한 분석 방법들이 사용 되고 있다. 분석방법이 다양한 이유는 microarray 실험이 많은 변동 요인(error)를 가지고 있고 또한 다양한 실험 설계 방법이 존재하기 때문이다. microarray 자료 분석에 사용되는 또 하나의 방법은 이렇게 선택된 유전자들을 군집분석 등을 통하여 자료의 차원을 줄이는 방법이다. 위계적 군집화(hierarchical clustering), K 평균 군집화(K-means clustering), 의사결정나무(decision tree) 와 같은 분석방법 뿐 아니라 인공 신경망(artificial neural network), SVM(Support Vector Machine) 등과 같은 데이터 마이닝 기법을 이용한 분석도 이루어지고 있다.

2. 2. 비교 게놈 교잡법

2. 2. 1. CGH의 소개

암 연구에 있어 CGH는 고형 종양(solid tumor)에 생긴 염색체의 결실이나 증폭의 유무를 전 염색체를 대상으로 해석하는 방법이다.

CGH는 고형종양에서 전 염색체를 대상으로 DNA의 카피수 변화(copy number change)를 해석하는 방법으로 위력이 있고 변화를 보이는 염색체영역을 좁히기 위한 스크리닝법으로 획기적인 분자 유전학적 방법이다. 하지만 인간의 분열 중기 염색체(meta phase)표본을 이용하기 때문에, 일반적으로 10-20 Mb이상에 걸친 변화가 없으면 CGH로는 검출이 어려운 검출감도나 해상도에 한계가 있다.

2. 2. 2. CGH의 실험방법

CGH의 실험과정은 종양 DNA를 FITC로 준거(reference) DNA를 로다민으로 표지한후 인간 정상 분열 중기 염색체표본에 동시에 결합시킨다. 이것을 CCD카메라를 사용하여 교잡(hybridization)한 후의 형광화상을 취해 유사칼라 (FITC,녹색; 로다민, 빨강색)를 붙여서 표시하는 것과 동시에, 전용 소프트웨어를 사용하여 염색체상의 각각의 형광강도의 비율을 계산하고 프로파일로 나타낸다. DNA카피수의 증가를 획득(gain), 감소를 소실(loss)라고 한다.

2. 3. Array-based CGH

2. 3. 1. Array-based CGH의 소개

CGH는 인간의 분열 중기 염색체로 유도된 염색체를 현미경 상에서 판단해야 하기 때문에 염색체 상의 미세한 부분의 변화를 알 수 없어서 실제 해상력은 20 Mb 이상이며 조작과 결과 분석의 어려움으로 인하여 많은 시간과 노력을 요구한다. 최근에 개발된 array-based CGH는 유리 판(glass slide) 위에 분열중기로 유도된 염색체 대신에 염색체 상에서 정확한 위치를 알고 있는 DNA fragment(BAC, PAC, P1)를 심어 놓은 것으로 DNA microarray 법을 종래의 metaphase CGH법에 응용한 것이 array-based CGH법이다. 이것은 그림2에 제시한 것처럼, 슬라이드 상에 어떤 염색체 영역을 커버 할 수 있는 Bacterial Artificial Chromosome(BAC)클론, 혹은 염색체상에 위치가 그려진(mapping) cDNA를 array 상으로 배치한 실험방법이다. 이 방법은 metaphase CGH법으로 염색체 영역을 좁힌 후, 그 영역에 상당하는 cosmid혹은 BAC클론을 array화 한 DNA칩을 제작하고 게놈의 변화를 해석하는 것이다.

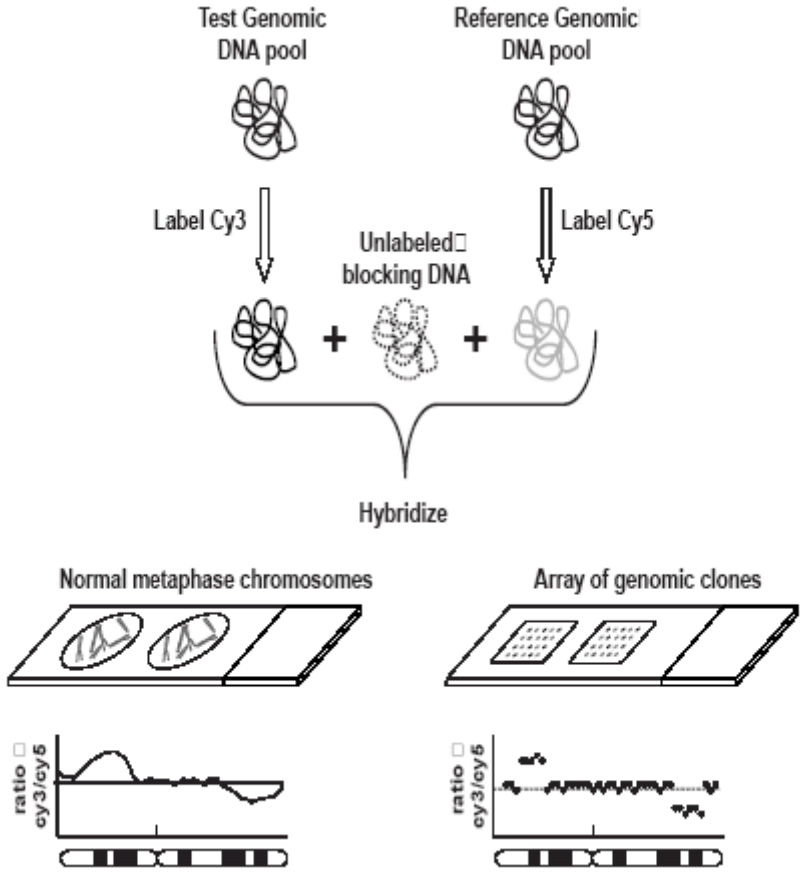


그림 2. Array-based CGH법의 원리 (Fridlyand et al., 2004)

2. 3. 2. Array-based CGH의 실험방법

실험방법은 종양 DNA를 빨간색(Cy5)으로 준거 정상 DNA를 녹색(Cy3)으로 표시한 후 슬라이드글라스에 찍은 BAC혹은 cDNA(microarray)를 동시에 교잡시킨다. 암의 염색체 상에서 증폭이 있는 영역은, 그 배열과 일치한 DNA 영역에서 빨간색 형광이 강해지고 암으로 결실 하고 있는 영역에 일치한 DNA 영역에서는 정상 DNA에서의 녹색형광이 강해진다. 변화가 없는 영역에 상당하는 DNA 영역에서는 빨간색과 녹색이 합성되기 때문에 노랑색으로 검출된다. CCD카메라 혹은

형광 스캐너를 이용하여 결합한 후의 형광화상을 취해, 유사칼라 (Cy3, 녹색; Cy5, 빨강색)을 붙여서 표시하는 것과 동시에, 각각의 형광강도의 비율(Red/Green)을 계산하고, 프로파일로 나타낸다. 여기서 중요한 것은 찍을(spot) 할 DNA가 염색체상에서 위치가 정해진 것이어야 한다는 것이다. (그림3) 이처럼 array-based CGH는 종양 게놈중의 DNA 카피수의 차이, 염색체 증폭이나 결실을 두 종류의 형광강도의 변화 로 검출하고 직접 유전자에 까지 접근하려고 하는 새로운 게놈 변화의 해석법이라 할 수 있다.

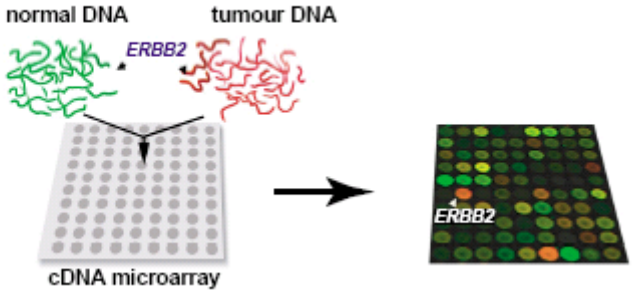


그림 3. cDNA microarray CGH의 실험 (Pollack et al.,1999)

2. 3. 3. cDNA 칩 과 BAC 칩 사용의 차이점

Microarray에서 사용되는 cDNA 칩은 비교적 수집이 쉬운 cDNA를 칩위에 올려놓은 것으로 많은 종류의 유전자를 대상으로 진단할 수 있는 장점이 있다. 그러나 이러한 많은 종류의 유전자에 의한 발현이 그 종류만큼 다양하여 발현량에 따라 정상세포와 암세포를 나누는 기준을 세우기가 매우 어려운 단점이 있다. 이에 비해 array-based CGH에서 사용되는 BAC 칩의 경우 암 관련 유전자의 발현량이 아닌 수적, 구조적 변화를 감지하는 방식이기 때문에 정상세포와 암세포와의 구별이 훨씬 명확해 질 수 있다. BAC 칩위에 올라가는 BAC DNA는 복잡하고 어려운 4단계(BAC library 제작 단계, End-sequencing 단계, Bioinformatics 단계, BAC Clone mapping by FISH 단계)를 거쳐서 명확하게 확인된 BAC clone을 추출하며 BAC 칩은 cDNA 칩과 달리 명확하게 확인이 된 BAC DNA를 이

용하기 때문에 보다 정확한 결과를 제공해 줄 수 있다.

cDNA 칩과 BAC 칩의 차이점을 아래 표1에 나타내었다.

표 1. cDNA 칩과 BAC 칩의 차이점

구분	cDNA chip	BAC chip
DNA sourse	cDNA	BAC (Bacterial Artificial Chromosome
실험목적	유전자의 발현량 측정	염색체 또는 유전자의 수적/구조적 변화 측정
Clone의 선별	Gene이나 STS의 무작위적 선별	발암관련 유전자(Oncogene &Tumor suppressor gene)을 포함하는 BAC clone
기술적 기반	Northern blot, Dot blot	FISH (Fluorescent in Situ Hybridization) CGH (Comparative Genomic Hybridization)

2. 3. 4. Array-based CGH 자료 분석

Array-based CGH 분석은 간단하게는 실험-준거군의 log ratio의 값의 2 표준편차를 기준으로 염색체의 증폭과 결실을 정의하는 것으로부터 시작하여, 보편적인 기준을 1.15이상을 증폭, 0.85 이하를 결실로(Lundsteen et al., 1995; Schleger, 2000) 제시하였다. 후에 분계점 방법(threshold method; Pollack et al., 2002), 회귀모형을 바탕으로 한 비모수적 방법 (Cheng et al., 2003)등을 이용한 모델을 제시했다. 다른 방법으로는 Hidden Markov Model (Fridlyand., 2004), Break point model(Jong et al., 2003)등의 모델이 제시되고 있다. 개발된 프로그램으로는 Matlab tool인 CGH-Plotter(Autio et al., 2003)과 CLAC 알고리즘을 이용한 CGH-Miner(Wang et al., 2004)가 염색체의 증폭과 결실을 찾는 데 쓰이고 있다.

제 3장 통계학적 배경

3. 1. 회귀분석

3. 1. 1. 행렬을 이용한 다중 회귀모형

다중회귀모형에서는 하나의 반응변수에 대해 여러 개의 설명변수가 주어져 있다. 설명변수를 $X_1, X_2, X_3, \cdots, X_p$ 라 한다면 자료는 아래와 같이 표현 될 수 있다.

표 2. 자료의 구조

자료의 순서	변수값					
	Y	X_1	X_2	X_3	……	X_p
1	y_1	x_{11}	x_{12}	x_{13}	……	x_{1p}
2	y_2	x_{21}	x_{22}	x_{23}	……	x_{2p}
⋮	⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	x_{n3}	……	x_{np}

이 표현방법에 의하면 x_{ij} 는 j번째 설명변수에서 i번째 자료의 값을 나타낸다. 다중회귀모형을 $y = (y_1, y_2, \cdots, y_n)^T$, $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$, $\beta = \beta_1, \beta_2, \cdots, \beta_p)^T$ 의 벡터와 다음과 같은 행렬로 표시된 설명변수 자료를 이용하면

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$y = \mathrm{X}\beta + \epsilon \qquad , \ \epsilon \ \sim N(0, \sigma^2 I) \tag{3-1}$$

과 같이 표현된다.

식 (3-1)의 분산-공분산(variance-covariance) 행렬은 ϵ_i가 각각 독립이므로 대각원소의 분산값 σ^2을 제외한 모든 공분산 값은 0이 되어 $\sigma^2 I$와 같은 상수행렬로 표시된다.

3. 1. 2. 최소제곱법

다중회귀모형에서의 최소제곱법은 다음의 함수를 최소화 시켜주는 β의 추정량 $\hat{\beta}$ 을 구하는 방법이다.

$$Q(\beta) = \sum (y_i - x_i^T \beta)^2 = (y - \mathrm{X}\beta)^T (y - \mathrm{X}\beta) \tag{3-2}$$

식 (3-2)를 $\hat{\beta}$에 대해 편미분하여 정리하면 다음과 같은 정규방정식 (normal equation)을 얻는다.

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\beta = \mathrm{X}^{\mathrm{T}} y$$

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} y$$

$\hat{\beta}$ 을 이용하여 얻는 통계량은 다음과 같다.

1) y의 추정값 $\hat{y}$

$$\hat{y} = \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}y$$

2) 오차의 추정값 e

$$\begin{aligned} e = y - \hat{y} &= y - \mathrm{X}\hat{\beta} \\ &= y - \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}y \end{aligned}$$

3) 잔차제곱합 SSE

$$\begin{aligned} SSE = e^{T}e &= (y - \mathrm{X}\hat{\beta})^{T}(y - \mathrm{X}\hat{\beta}) \\ &= y^{T}y - \hat{\beta}^{T}\mathrm{X}^{\mathrm{T}}\mathrm{y} \end{aligned}$$

4) σ^2의 추정량 $\hat{\sigma^2}$ (p는 모수의 수)

$$\hat{\sigma}^2 = \frac{SSE}{n - p} = MSE$$

5) $\hat{\beta}$의 분산 공분산 행렬

$$Var(\hat{\beta}) = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$$

(3-3)

$\sigma^2(X^TX)^{-1}$의 대각 원소의 값은 해당되는 계수의 분산이며 행렬의 (i, j)번째

요소는 i번째 회귀계수와 j번째 회귀계수의 공분산이다.

σ^2을 알지 못하는 경우 MSE로 대체 될 수 있다

6) 가설검정

귀무가설 $H_0;\ \beta_k = 0$ 에 대한 통계량은

$$t^* = \frac{\hat{\beta}_k - \beta_k (= 0)}{s.e(\hat{\beta}_k)} \tag{3-4}$$

이다.

여기서 $\hat{\beta}_k$의 표준오차는 $MSE(X^T X)^{-1}$의 대각원소의 제곱근을 의미한다.

3. 1. 3. 가중최소제곱법

최소제곱법은 등분산 가정, 즉 오차의 분산은 일정하다고 가정하는 경우에 사용되는 방법이다. 만약 이러한 가정이 만족되지 못하는 경우에는 최소제곱법에 의한 추정량은 가우스-마코프의 정리에 의해 최량선형불편추정량(Best Linear Unbiased Estimator ; BLUE)이 아니다. 자료의 성격상 이러한 등분산가정은 만족이 되지 않는 경우가 흔히 발생한다. 이런 경우에 일반화 다중 회귀모형(generalized multiple regression model)을 기초로 한 가중최소제곱법을 적용 시킬 수 있다.

일반화 다중 회귀모형은

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ \cdots\ + \beta_{ip} + \epsilon_i$$

과 같이 나타낼 수 있다.

여기서 $\beta_0, \beta_1, \cdots, \beta_p$ 는 모수이고, $X_{i1}, X_{i2}, \cdots, X_{ip}$ 는 알려진 상수 이며 , 오차는 $\epsilon_i \sim iid\ N(0, \sigma_i^2)$, $i = 1, 2, \cdots, n$ 이다.

일반화 다중모형에서 오차의 분산-공분산 행렬은 식 (3-1) 보다 더 복잡한 형태이다.

$$\sigma_{\epsilon}^2 = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

가중최소제곱법은 아래 식 (3-5)의 Q를 최소화하면서 회귀계수를 추정하는 방법이다.

$$Q(\beta) = \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2 \qquad (3\text{-}5)$$

여기서 σ_j^2의 역수를 가중치(weight)로 정의하며,

$$w_i = \frac{1}{\sigma_i^2}$$

$w_i = 1$ 일때 최소제곱추정법이 된다.

행렬 W 는 가중치 w_i를 포함한 대각행렬(diagonal matrix)이다.

$$\mathrm{W} = \begin{bmatrix} w_i & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

식 (3-5)를 $\hat{\beta}$에 대해 편미분하여 정리하면 다음과 같은 정규방정식(normal equation)을 얻는다.

$$\mathrm{X}^{\mathrm{T}}\mathrm{WX}\beta = \mathrm{X}^{\mathrm{T}}\mathrm{Wy}$$

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{WX})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Wy}$$

σ_i^2을 안다고 가정 했을 때 분산-공분산 행렬은 아래와 같다.

$$\sigma_{\beta}^2 = (\mathrm{X}^{\mathrm{T}}\mathrm{WX})^{-1}$$

3. 1. 4. 모형진단

모형진단은 모형이나 가정에 어떤 문제점이 있나를 알아보는 것으로 회귀 모형에 세운 가정이 실제 문제에서 적당한지를 알아보는 것이다. 여기서 기본적으로 사용되는 통계값은 잔차(residual)이다. 잔차의 형태로부터 추정모형이 관측된 자료를 얼마나 정확하게 적합하고 있는지, 즉 모형의 가정이 바람직한지의 여부를 알 수 있다. 회귀모형에 기본적인 가정은 다음과 같다.

첫째, 오차항(ϵ_i)의 등분산성이다. 어떠한 $X = x$ 값에 대해서도 Y의 분산은 같다는 가정이다.

둘째는 독립성(independency)이다. 만약 오차항들이 서로 독립이라면 잔차 들은 무작위로 흩어져 있을 것이고, 그렇지 않다면 오차항들 사이에 상관관계가 있다고 할 수 있다. 오차항의 독립성을 평가하는 한 측도로 더빈-왓슨(Durbin -Watson) 통계량이 있다. DW통계량은 잔차들의 상관계수를 측정하게 되며 아래와 같이

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

로 정의한다.

셋째는 오차의 정규성이다. 어떠한 $X = x$ 값에 대해서도 Y의 분포는 정규분포를 따른다는 가정이다. 정규분포는 대부분의 오차에 대해 가정 보편적이고 타당성 있는 분포이다. 회귀분석의 추정과 검정에 쓰이는 t-분포는 정규분포에 약간 위배되더라도 덜 민감한 특성을 갖고 있기 때문에, 정규성 가정이 만족되지 않았을 때의 심각성은 위의 두 경우보다는 덜하다. 오차항의 정규성은 잔차의 정규 확률도(normal probability plot ;Q-Q plot)로 검토 할 수 있다. 정규 확률도는 잔차의 정규분포 하에서의 기댓값을 가로축으로 하고 실제 관찰된 잔차를 세로축으로 하여 그린 그림이다. 이 그림이 45도의 기울기를 가진 직선에 가까우면 가까울수록 오차항이 정규분포를 따른다고 볼 수 있다.

3. 2. 이분산성 일치적 표준 오차

3. 2. 1. 이분산성 일치적 공분산 행렬의 소개

회귀모형에서 통계학적 가정(등분산성, 정규성, 독립성)들이 만족되는 경우에 최소제곱법은 가장 효과적인 최량선형불편추정량을 추정 할 수 있다. 종종 오차의 분산에 이분산성(heteroscedasticity)이 관찰치들 사이에 나타나는데 이런 경우에 최소제곱추정량은 비편향 추정값은(unbiased) 되지만 비효율적(inefficient)이 된다.

이에 각각의 관찰치 간 오차의 표준편차(standard deviation)의 역수를 가중치로 둔 이분산성을 고려한 통계학적 방법이 제시되고 있다.(Green, 2000; Carroll and Ruppert, 1988) 이분산성의 형태(form)가 알려진 경우에는 앞에서 제시한 가중최소제곱법을 이용 할 수 있다. 그러나 보통의 경우 이분산의 형태를 알 수 없는데 이런 경우 단순가중을 두는 방법은 실용적(impractical)이지 못하다.

이분산성은 부정확한 함수의 형태에 의해 야기되는데, 이것은 종속변수의 분산안정화변환(variance stabilizing transformation)방법 (Weisberg, 1980)이나 종속변

수, 독립변수 모두의 변환으로 (Carroll and Ruppert, 1988) 정확하게 할 수 있다.

이분산성 일치적 공분산 행렬을 기초로 한 검정(test)은 이분산성이 알려지지 않은 형태 일때 회귀계수의 공분산 행렬을 일치추정량으로 제시한다.

HCCM의 발달과정은 초기에 Eicker(1963, 1967)와 Huber(1967), White (1980)가 HC0로 알려진 HCCM 형태를 소개하였고 후에 Mackinon and White(1985)가 HC0의 제안점등을 들어 이를 보안한 HC1, HC2, HC3를 제안하였다. 여기서 그들은 작은 표본에서는 HC3를 사용하는 것을 권장하였고 Davidson and Mackinnon (1993)는 HC0 보다는 HC2 나 HC3를 사용할 것을 권장하였다.

3. 2. 2. 다중회귀모형에서의 이분산성 일치적 공분산 행렬

식 (3-1)의 모형에서 $E(\epsilon) = 0$, 그리고 $E(\epsilon\epsilon^T) = \Phi$로 정의하면 최소제곱량 $\hat{\beta}$의 분산은

$$var(\hat{\beta}) = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\Phi(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$$

(3-6)

이다.

오차가 등분산(homoscedastic)일때 에는 $\Phi = \sigma^2\mathrm{I}$ 이므로 식 (3-3)에서와 같이 정의 될 수 있으나 이분산 일때에는 식 (3-3)=식 (3-6) 가 될 수 없다.
만약 오차가 이분산이고 Φ가 알려진 경우에는 식 (3-6)을 이용하여 이분산성을 보정 할 수 있다. 그러나 이분산성의 형태가 알려지지 않은 경우에는 HCCM이 사용되어야 한다. 위에서 말한 4가지 형태에 대한 정리는 다음과 같다.

1) HC0

HC0는 White, Eicker, Huber 추정량과 같이 HCCM에서 가장 보편적으로 사용되는 방법이다. HCCM은 ϕ_{ii}를 추정하기 위해 e_i^2 를 사용한다. 이것을 수식으로 나타내면

$$\hat{\phi}_{ii} = \frac{(e_i^2 - 0)^2}{1} = e_i^2 \ , \quad \hat{\Phi} = diag[e_i^2]$$

$$\begin{aligned} HC0 &= (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\hat{\Phi}\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \\ &= (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{diag}[\mathrm{e_i}^2]\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \end{aligned} \tag{3-7}$$

으로 나타 낼 수 있다.

2) HC1

Hinkley(1977)는 HC0를 자유도로 보정한 HC1을 제안했는데 이는 HC0의 작은 표본 성질(property)을 증가시킨 것이다. HC1은

$$\begin{aligned} HC1 &= \frac{N}{N-K}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{diag}[\mathrm{e_i}^2]\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \\ &= \frac{N}{N-K}HC0 \end{aligned} \tag{3-8}$$

과 같이 나타 낼 수 있다.

3) HC2

식 (3-7)에서는 잔차 e 를 기초로 한 $\hat{\Phi}$을 구한다. 그런데, 비록 오차가 이분산이라도 잔차는 그렇지 않을 수 있다. 이에 $h_{ii} = x_i(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}x_i^T$ 라고 정의할 때

$$var(e_i) = \sigma^2(1-h_{ii}) \neq \sigma^2 \ , \ \frac{1}{N} \leq h_{ii} \leq 1$$

가 된다.

HC2는 e_i^2이 σ^2에 편향(biased)된 추정치라 해도 $\frac{e_i^2}{1-h_{ii}}$는 덜 편향되는 것을 Mackinnon과 White(1985)가 Horn과 Duncan(1975)의 이론을 바탕으로 제안한 것으로

$$HC2 = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} diag\left[\frac{e_i^2}{1-h_{ii}}\right]\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \qquad (3\text{-}9)$$

로 정의된다.

4) HC3

HC3는 Efron(1982)의 jack-knife 추정량과 비슷하게 표현되는데,

$$HC3 = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} diag\left[\frac{e_i^2}{(1-h_{ii})^2}\right]\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \qquad (3\text{-}10)$$

로 나타낼 수 있다.

여기서 e_i^2을 $(1-h_{ii})^2$으로 나누는 것은 분산이 큰 관찰치의 “과대 영향(over-influence)”을 수정한(adjust) 효과를 준다.

3. 3. 다중 검정

3. 3. 1. 개요

다중검정 문제는 많은 가설(hypothesis) 검정이 같은 자료 안에서 이루어 질 때 발생한다.

다중검정은 얻어진 자료의 가치나, 그 자료의 새로운 면을 발견하기위한 대안적인 통계학적 방법들이기에 의미가 크나, false significance를 높이는 단점이 있다.

예를 들어 10개의 가설을 5%의 유의수준(significance level)로 검정한다면, 이 검정들로 얻어진 유의확률(p-value)들의 분포(distribution)는 균등분포(uniform)이고 각각 독립적(independent)라고 가정할 것이다. 귀무가설(null hypothesis)은 0.05하에서 유의성을 검정하게 되는데, 적어도 10개의 검정중에 하나 이상의 검정의 유의수준의 확률이 0.401로 낮아지게 되며, 만약 20개의 가설이라면 0.642로 확률이 올라간다. 이런 변화들이 다중 검정의 단점을 보여준다.

다중검정은 가설검정의 집합(a family of hypothesis tests)에서 보정된 유의확률에 의해 문제에 접근하는 방법이다. 보정된 유의확률이란 전체 가설의 집합을 고려하면서 주어진 가설을 기각(reject)할 가장 작은 유의수준이다. 결정방법(decision rule)은 보정된 유의확률이 α보다 작을때 귀무가설을 기각한다. 거의 모든 경우에 이 방법은 FWER(FamilyWise Error Rate)을 α수준, 혹은 그 보다 낮게 조정한다.

3. 3. 2. 유의확률을 보정하는 방법들

다중검정을 하는 방법은 여러 가지가 있는데 간단히 소개하면 아래의 표와 같이 정리 할 수 있다.

표 3. 유의확률을 보정하는 방법들

Method	type	control
Bonferroni	single-step adjusted p-value	FWER
Holm(1979)	step-down adjusted p-value	
Hochberg(1988)	step-up adjusted p-value	
SidakSS	single-step adjusted p-value	
SidakSD	step-down adjusted p-value	
Benjamini and Hochberg(1995)	step-up adjusted p-value	FDR
Benjamini and Yekutieli(2001)	step-up adjusted p-value	

Bonferroni 와 Sidak 보정방법은 raw p-value의 간단한 함수들이다. 이 두 방법은 너무 보수적인(conservative)것으로 알려져 있다. Step-down 방법은 Hochberg의 Step-up 방법과 함께 보수적 인 것 들을 약간 제외한 방법이다. 위에서 제시한 방법 이외에도 bootstrap 방법이나 permutation을 이용한 방법도 있다. (Westfall and Young, 1989,1993)

제 4장 Array-based CGH 자료 분석에서 가중 최소 제곱법을 이용한 회귀분석

4. 1. 회귀모형

본 논문에서 제시할 모형은 Thomas et al.(2001)가 제시한 회귀모형과 Cheng et al.(2003)이 제시한 ARO모형을 적합, 변형 시켜서 array-based CGH 자료에 적용해 본 것이다.

회귀 모형은

$$Y_{jk} = \delta_k + \lambda_k(a_j + b_j x_k + \epsilon_{jk}) \qquad (4\text{-}1)$$

로 정의 한다.

여기서 $Y_k = \begin{bmatrix} Y_{1k} \\ Y_{2k} \\ \vdots \\ Y_{Jk} \end{bmatrix}$ 는 array를 나타내며 Y_{jk}는 k번째 샘플의 j번째 유전자를 표현한다. $(j = 1, 2, \cdots, J;\ k = 1, 2, \cdots\ K)$

$x_k = \begin{matrix} 0 \ , \text{if} \ \ normal\ sample \\ 1 \ , \text{if} \ \ tumor\ \ sample \end{matrix}$ 는 k번째 샘플에 관여하는 공변량,

(a_j, b_j) ; 특정유전자 회귀계수(gene-specific regression coefficients),

δ_k ; 특정 샘플 가법 이질성 요인(sample-specific additive heterogeneity factor),

λ_k ; 특정 샘플 승법 이질성 요인(sample-specific multiplicative heterogeneity factor),

ϵ_{jk} ; 확률변수(random variable) 이다.

x_k가 이분형(binary)이기 때문에 a_j는 정상 샘플$(x_k = 0)$에서의 j번째 유전자의 발현량(expression level)의 평균을 측정 할 수 있고 ,b_j는 두 샘플 그룹 사이의 j번째 유전자의 발현량의 평균의 차이를 측정 할 수 있다. 이질성 요인인 δ_k 와 λ_k는 다량의 mRNA 샘플간의 변동(variation) 정도를 나타낸다. 이질성 요인을 추정하기위해 3.1.3 에서 소개한 가중최소제곱법을 이용하여 모수를 추정하기로 하였다. 앞으로 제시할 3가지의 방법으로 array-based CGH 자료를 분석할 때 그 방법에 따라 유의한 유전자를 찾는데 어떤 차이점이 있으며 얼마나 잘 발견하는지에 대해 알아보고자 한다.

4. 1. 1. 단순 2단계 회귀분석

단순 2단계 회귀분석 (이하 M1)은 두 번의 회귀분석 모형이다. 먼저 샘플간의 이질성 요인의 보정을 위해서 모형을

<step 1>

$$Y_{jk} = \delta_k + \lambda_k(\bar{y}_j + \epsilon_{jk})$$

여기서 $\bar{y}_j$; j 번째 유전자들의 샘플 간 평균

로 가정한다.

가중최소제곱법방법으로 추정된 이질성 요인인 $\hat{\delta}_k$ 와 $\hat{\lambda}_k$를 가지고 관찰된 원래의 자료를 $\frac{(Y_{jk} - \hat{\delta}_k)}{\hat{\lambda}_k}$ 로 보정한다. 여기서 가중치(weight)은 표본분산 (sample variance)를 사용하기로 하였다. (Carroll and Ruppert, 1988)

<step 2>

모형을 다시 설명하면, 보정된 모형은 아래와 같다.

$$\frac{(Y_{jk}-\hat{\delta}_k)}{\hat{\lambda}_k} = a_j + b_j x_k + \epsilon_{jk}$$

가중최소제곱방법으로 추정된 $\hat{a}_j$, $\hat{b}_j$ 에서 $\hat{b}_j$의 분산을 추정함에 있어서 3.2에서 소개한 HCCM 방법을 이용하여 분산을 추정하였다. 이는 이분산성이 알려지지 않은 형태 일 때 회귀계수의 공분산 행렬을 일치추정량으로 제시하는 방법으로 오차의 이분산성을 고려한 통계량을 세울 수 있다.

4. 1. 2. 단순 반복법

단순 반복 방법 (이하 M2)은 다음과 같다. 우선, 특정 샘플 가법 이질성 요인 (δ_k=0) , 특정 샘플 승법 이질성 요인 (λ_k=1)로 초기값(initial value)을 주면 모형은 다음과 같이 표현된다.

<step1>

$$Y_{jk} = a_j + b_j + \epsilon_{jk}$$

최소제곱법을 이용하여 $\hat{a}_j^{(1)}$과 $\hat{b}_j^{(1)}$ 을 추정한다. 여기서, $MSE^{(1)} = \hat{\sigma}_j^{2(1)}$ 을 가정한다.

<step 1> 에서 추정된 $\hat{a}_j^{(1)}$와 $\hat{b}_j^{(1)}$를 식(4-1)에 넣으면 다음과 같이 표현된다.

$$Y_{jk} = \delta_k + \lambda_k(a_j^{(1)} + b_j^{(1)} x_k + \epsilon_{jk})$$

여기서 가중치는 <step 1>에서 구한 $MSE^{(1)} = \hat{\sigma}_j^{2(1)}$ 를 이용하여 가중 최소제곱법으로 $\hat{\delta}_k^{(1)}$ 과 $\hat{\lambda}_k^{(1)}$을 추정한다.

<step2>

모형을 다시 설명하면, 보정된 모형은 아래와 같다.

$$\frac{(Y_{jk} - \hat{\delta}_k^{(1)})}{\hat{\lambda}_k^{(1)}} = a_j + b_j x_k + \epsilon_{jk}$$

가중최소제곱방법으로 $\hat{a}_j$, $\hat{b}_j$ 를 추정하고 HCCM 방법으로 분산을 추정하여 통계량을 계산한다.

4. 1. 3. 반복법

세 번째로 제시하는 반복법(이하 M3)은 두 번째 방법의 확장된 방법이다. 두 번째 소개한 방법이 반복법의 단순한 형태라면 이번에 소개하는 방법은 수렴(converge) 될 때 까지 반복하는 방법이다.

두 번째 방법에서와 마찬가지로 , 특정 샘플 가법 이질성 요인 (δ_k=0) , 특정 샘플 승법 이질성 요인(λ_k=1)로 초기값을 주면 모형은 다음과 같이 표현된다.

<step1>

$$Y_{jk} = a_j + b_j + \epsilon_{jk}$$

최소제곱법을 이용하여 $\hat{a}_j^{(1)}$과 $\hat{b}_j^{(1)}$ 을 추정하고 $MSE^{(1)} = \hat{\sigma}_j^{2(1)}$ 을 가정한다. <step 1> 에서 추정된 $\hat{a}_j^{(1)}$와 $\hat{b}_j^{(1)}$를 식 (4-1) 에 넣으면 다음과 같이 표현된다.

$$Y_{jk} = \delta_k + \lambda_k(a_j^{(1)} + b_j^{(1)}x_k + \epsilon_{jk}) \quad (4\text{-}2)$$

여기서 가중치는 <step 1>에서 구한 $MSE^{(1)} = \hat{\sigma}_j^{2(1)}$ 를 이용하여 가중 최소 제곱법으로 $\hat{\delta}_k^{(1)}$ 과 $\hat{\lambda}_k^{(1)}$을 추정한다.

<step2>

모형을 다시 설명하면, 보정된 모형은 아래와 같다.

$$\frac{(Y_{jk} - \hat{\delta}_k^{(1)})}{\hat{\lambda}_k^{(1)}} = a_j + b_j x_k + \epsilon_{jk} \quad (4\text{-}3)$$

가중 최소 제곱방법으로 $\hat{a}_j^{(2)}$, $\hat{b}_j^{(2)}$ 를 추정하고 $MSE^{(2)} = \hat{\sigma}_j^{2(2)}$ 를 얻는다. 식(4-2)로 돌아가서 $\hat{a}_j^{(1)}$ 대신에 $\hat{a}_j^{(2)}$, $\hat{b}_j^{(1)}$대신에 $\hat{b}_j^{(2)}$을 넣고 계산을 반복하여 n번 시행한 후 $\hat{\delta}_k^{(n)}$, $\hat{\lambda}_k^{(n)}$ 이 수렴될때 최종 $\hat{a}_j^{(n)}$,$\hat{b}_j^{(n)}$으로 통계량을 세운다.

4. 2. 회귀계수에 대한 검정

식 (3-4) 에서처럼 귀무가설 $H_0;\ \beta_j = 0$ 에 대한 통계량은 다음과 같이 구할수 있다.

$$t_1^* = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{HC0}}$$

$$t_2^* = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{HC1}}$$

$$t_3^* = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{HC2}}$$

$$t_4^* = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{HC3}}$$

여기서 $\hat{\beta}_j$의 표준오차는 $MSE(X^TX)^{-1}$의 대각원소의 제곱근을 의미한다.

3. 2. 2 에서 논의 한 것을 다시 살펴보면, 이분산성의 형태가 알려지지 않은 경우에는 $MSE(X^TX)^{-1}$ 대신에 HCCM이 사용되어야 한다. 따라서 HC0, HC1, HC2, HC3 등 4가지 방법으로 표준오차를 추정하여 통계량을 구하여 비수정적 유의확률을 구하게 된다.

4. 3. 다중 검정

4. 2 에서 제시한 통계량으로 얻어진 유의확률은 다중 검정 문제를 가지고 있다. (이는 3. 3 에서 자세한 내용은 논의 되었다.) 이를 해결하기 Bonferroni, Holm, Hochberg, Sidak, Benjamini and Hochberg, Benjamini and Yekutieli이 제시한 방법으로 수정된 유의확률을 구하여 유의한 유전자를 찾는다.

제 5장 자료 분석

5. 1. R package

R 은 통계 전산(statistical computing)과 그래픽(graphics)을 위한 언어 이다. GNU(Gnu is not Unix) project로서 S 언어 와 비슷하며, AT&T이자 지금은 Lucent Technologies인 Bell 연구실의 Ross Ihaka 과 Robert Gentleman 에 의해 개발되었다. R은 다양한 통계적 분석- 선형(linear) 혹은 비선형 모델(nonlinear modeling), 고전적 통계학적 검정법 , 시계열(time-series) 분석, 군집분석등-과 그래픽적인 기술등의 발전 가능하다.

R 은 Open Source route를 따르며 http://cran.r-project.org 에서 다운받을 수 있는 Free Software이다. 또한 별도의 라이브러리를 사용할 수 있어 개발도 가능하고 자료도 공유할 수 있고 S로 쓰여진 많은 프로그램을 R에서 사용할 수 있는 장점이 있다. 2004년 11월에 나온 release R.2.0.1이 현재 최신 버전이며, 본 논문에서는 다중검정을 할 수 있는 (multtest package사용) release R.2.0.0을 이용하여 분석하였다.

5. 2. 실제 자료 분석

5. 2. 1. 유방암 자료

Stanford Genomics Breast Cancer Consortium에 공개되어있는 유방암 자료를 이용하여 분석해 보았다. 이 자료[1]는 또한 CGH-Miner 프로그램에 예제 자료로 사용되고 있다. 이 자료는 2개의 정상 준거 arrays 와 20개의 암(tumor) arrays 샘플에서의 array-based CGH분석을 통해 얻어진 log ratio 값이다. 실험에 대한 설명은 2장에서 언급하였다.

자료의 구성을 살펴보면 아래와 같다.

표 4. Array-based CGH 자료의 구성

염색체	유전자	Normal sample		Tumor sample		
		X_1	X_2	X_3	……	X_k
chromosome 1	1	r_{11}	r_{12}	r_{13}	……	r_{1k}
	2	r_{21}	r_{22}	r_{23}	……	r_{2k}
	⋮	⋮	⋮	⋮		⋮
	j_1	$r_{j_1 1}$	$r_{j_1 2}$	$r_{j_1 3}$	……	$r_{j_1 k}$
⋮	⋮	⋮	⋮	⋮		⋮
chromosome 23	1	r_{11}	r_{12}	r_{13}	……	r_{1k}
	2	r_{21}	r_{22}	r_{23}	……	r_{2k}
	⋮	⋮	⋮	⋮		⋮
	j_{23}	$r_{j_{23} 1}$	$r_{j_{23} 2}$	$r_{j_{23} 3}$	……	$r_{j_{23} k}$

여기서 정상 샘플은 정상 대 정상 (normal vs normal) 교잡을 한 것이고, 암 샘플은 암 대 정상(tumor vs normal) 교잡을 한 array 이다. 염색체는 1번부터 23번

1) Pollack et al.(2002) http://genome-www.stanford.edu/breast_cancer

까지 조사되었고 각각의 염색체 내에서 유전자는 염색체 내의 거리로 순서화 되어 있다.

5. 2. 2. 자료 기술

Stanford Genomics Breast Cancer Consortium 제시된 자료는 23개 염색체에서 6,691개의 유전자이다. 여기서는 결측값(missing value)을 제외한 4,600개의 유전자만을 이용하여 분석 하였다. 이는 원래 자료의 68.75% 만을 사용한다는 단점이 있지만 결측값을 추정하는 방법을 이용하면 그 방법에 따라서 결과에 영향을 줄 수 있으므로 이 논문에서 제시하는 방법간의 비교를 위해 결측값을 제외시킨 자료를 이용하였다. 각 염색체 별 유전자의 빈도수는 다음과 같다.

표 5. 염색체별 유전자의 빈도수

염색체	유전자 개수	percent(%)	염색체	유전자 개수	percent(%)
1	450	9.78	13	92	2.00
2	332	7.22	14	176	3.83
3	265	5.76	15	120	2.61
4	187	4.07	16	178	3.87
5	245	5.33	17	260	5.65
6	265	5.76	18	76	1.65
7	226	4.91	19	220	4.78
8	167	3.63	20	123	2.67
9	177	3.85	21	56	1.22
10	202	4.39	22	103	2.24
11	266	5.78	23	167	3.63
12	247	5.37	**합계**	4600	100

5. 3. 방법간의 보정 정도 비교

염색체 6번에서의 각 방법 별 Y_{jk} 의 값을 보정하기 전에 값과 각 방법을 사용한 후 보정된 값으로 그려본 상자그림(boxplot) 이다. 보정하기 전보다 보정된 이후에 각 샘플 간의 분산이 줄어든 것 을 알 수 있다.

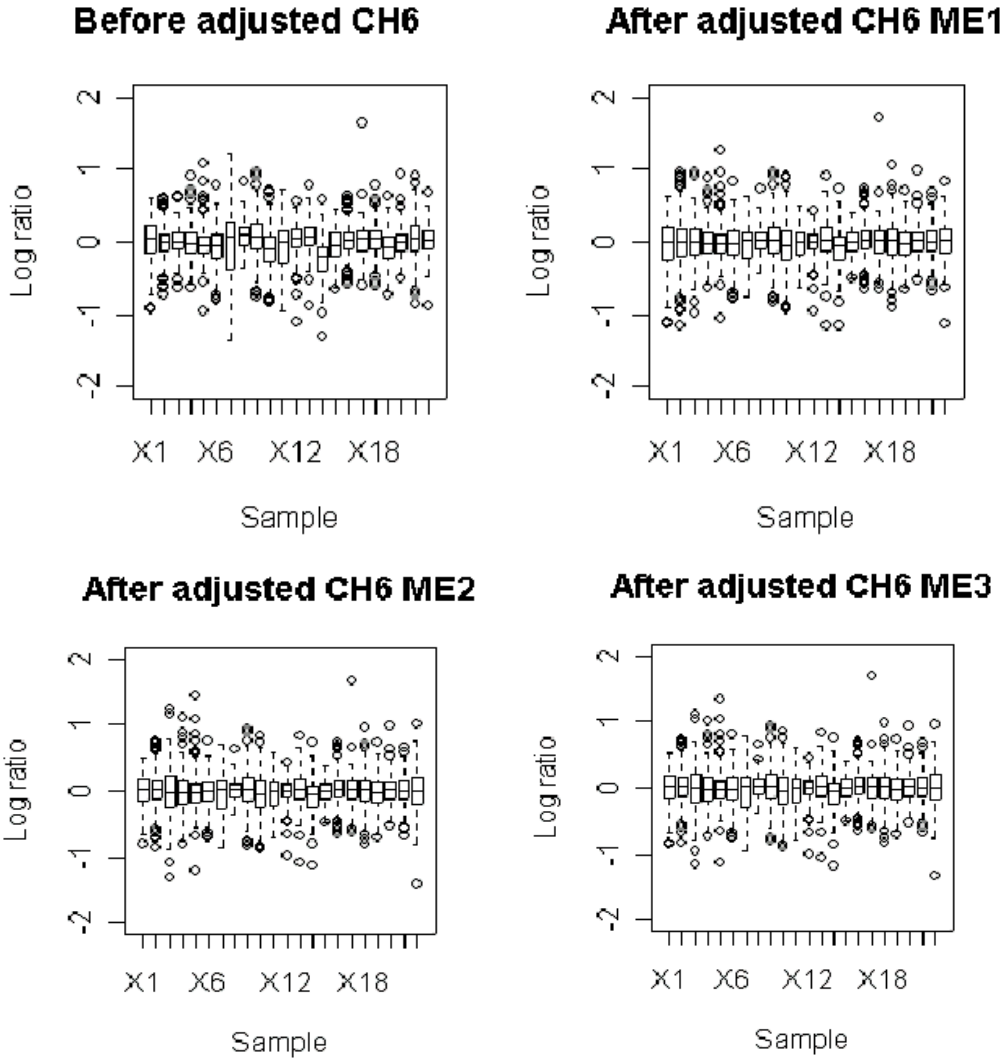


그림 4. 염색체 6번에서의 상자 그림

5. 4. 3가지 방법에 따른 δ_k 와 λ_k

표 6. 3가지 방법에 따른 $\hat{\delta}_k$ 과 $\hat{\lambda}_k$

k	$\hat{\delta}_k$			$\hat{\lambda}_k$		
	M1	M2	M3	M1	M2	M3
1	0.031303	-0.007634	-0.006748	0.551251	1.277665	1.227062
2	0.020500	0.007634	0.007140	0.453965	0.722335	0.750508
3	0.009469	0.007984	0.004466	0.494001	0.407041	0.495300
4	-0.008947	0.006653	0.003486	0.673205	0.528453	0.607919
5	0.040967	0.055990	0.054378	0.487653	0.441506	0.481959
6	-0.023792	-0.013035	-0.008978	1.172179	1.149262	1.047477
7	-0.028807	-0.016398	-0.012639	1.449713	1.301960	1.207657
8	0.051617	0.049458	0.053989	1.129443	1.135457	1.021786
9	-0.094747	-0.091842	-0.084952	1.779053	1.719629	1.546756
10	-0.048914	-0.043213	-0.041552	0.961089	0.913696	0.872035
11	0.075454	0.079755	0.085277	1.678807	1.557346	1.418786
12	0.111183	0.096683	0.099750	1.004501	0.974213	0.897268
13	-0.065965	-0.054113	-0.045937	1.871594	1.824691	1.619552
14	0.045551	0.031164	0.033469	0.762041	0.763138	0.705320
15	-0.036280	-0.028972	-0.022272	1.531878	1.513624	1.345525
16	-0.016432	-0.010038	-0.006740	1.135313	1.107185	1.024451
17	-0.003162	-0.008702	-0.009190	0.713222	0.696331	0.708583
18	0.052996	0.036342	0.039674	0.804907	0.843624	0.760038
19	-0.010990	-0.002453	0.001930	1.773332	1.643434	1.533462
20	-0.036384	-0.024580	-0.027147	0.442913	0.405251	0.469676
21	-0.079081	-0.077912	-0.080935	0.555162	0.552120	0.627964
22	0.014460	0.007227	0.003965	0.574776	0.522040	0.603861

표 6 는 8번 염색체를 각 각의 방법으로 분석 하였을때, 추정된 특정 샘플 가법 이질성 요인과 승법 요인을 제시하고 있다. 결과를 보면 전체적으로 M 1에 비하여 M2와 M3는 특정 샘플 가법 요인인 δ_k 는 작게 추정되고, 특정샘플 승법 요인

인 λ_k는 더 크게 추정됨을 볼 수 있었다. 이는 특정 샘플의 승법 요인이 더 크게 추정되는 M2와 M3에서 더 λ_k 값에 민감하게 반응할 것 이라는 예측을 할 수 있다.

5. 5. 잔차 그림

모형검정에 있어 3. 1. 4에서 설명한 통계학적 가정들의 타당성을 알 수 있는 가장 보편적인 방법은 표준화 잔차(studentized residual)를 y 축으로 하고 $\hat{y}$을 x 축으로 하는 잔차 그림을 그려보는 것이다. 표준화 잔차는 e_i를 그것의 표준오차 값으로 나눈 것으로 $h_{ii} = x_i(X^TX)^{-1}x_i^T$라고 정의했을때, 아래와같다.

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

가정된 모형이 타당하다면 각각의 표준화 잔차 r_i는 평균 0, 분산 1을 갖는다. 오차는 정규분포를 따른다는 가정이 타당하다면 r_i의 약 95%는 -2에서 +2 사이에 놓이고, r_i의 99.7%는 -3에서 +3 사이에 놓일 것이다.

Arrray-based CGH 자료 분석에서는 오차의 분산은 특별히 표집(sampling)과 유전자의 교차 교배(cross-hybridization)에 관계가 깊다고 가정하고, 이런 확률변수의 분포는 미지의 분포를 하며 정규분포와 같은 어떤 분포도 따를수 있다는 즉, 분포에 대한 가정을 하지 않는 확률변수라고 생각한다. (Thomas et al., 2001)

본 논문에서의 방법은 오차의 이분산성을 가정하고 그를 해결 할 수 있는 방법을 제시한 것이다. M1을 최소제곱법으로 구한 모델의 오차의 잔차 그림을 보면 아래 그림 5와 같다. 그림 5는 표준화 잔차로 그린 그림이다.

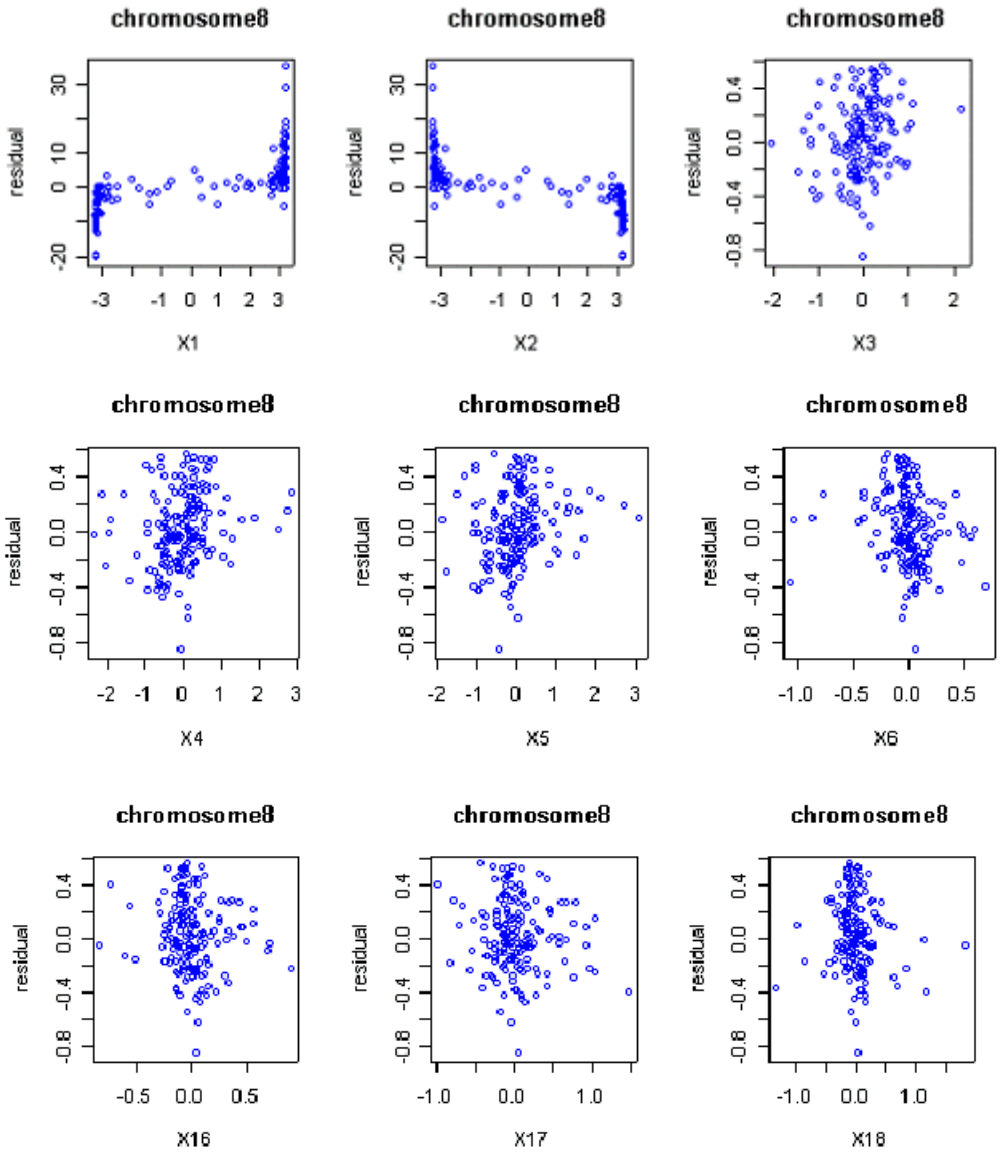
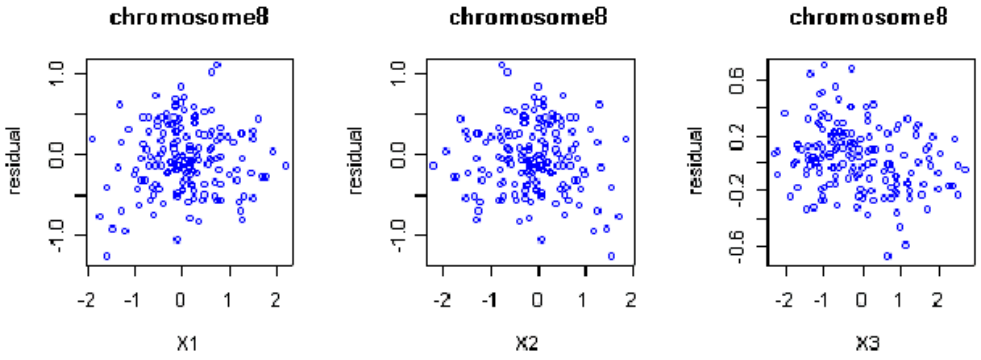


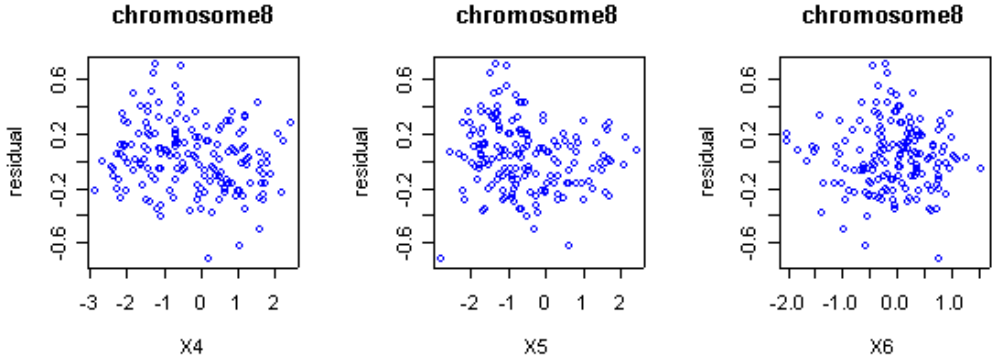
그림 5. 최소제곱법으로 추정한 모형의 표준화 잔차 그림

그림5에서 보면 잔차가 경향을 띠는 모습을 볼 수 있다. 즉 이분산성이라는 것을 알 수 있고 최소제곱법으로 모델을 설명 하는데 것에 문제점이 있음을 알 수 있다. 이 논문에서 제시하고 있는 방법들로 잔차그림을 그려보면, 다음과 같다.

a) M 1 으로 샘플 1, 2, 3 의 표준화 잔차 그림



b) M 2로 샘플 4, 5, 6 의 표준화 잔차 그림



c) M 3로 샘플 16, 17, 18 의 표준화 잔차 그림

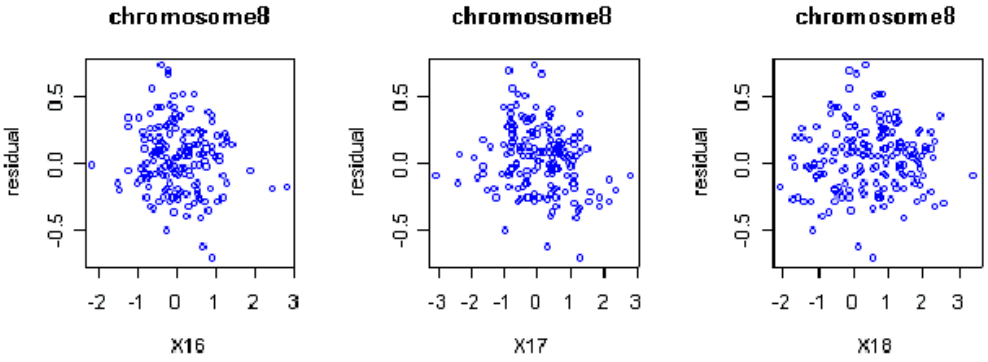


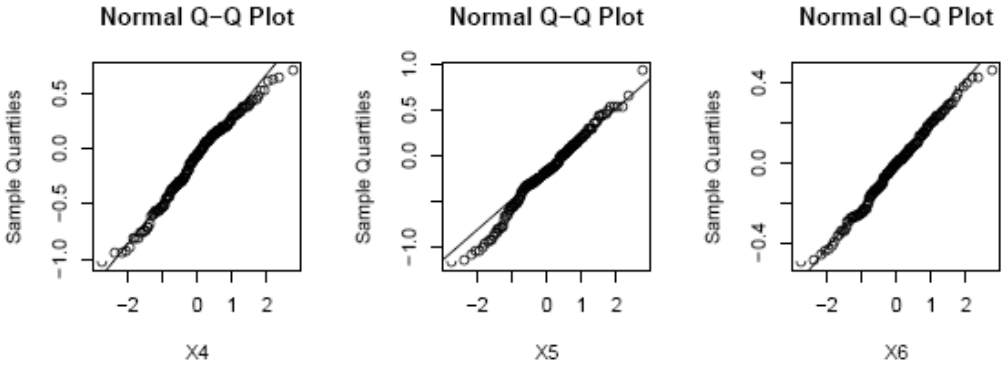
그림 6. 각 방법 별 표준화 잔차 그림

그림 6에 따르면 각각의 방법에서 오차의 분산이 등분산 가정을 따른다는 것을 알 수 있다. 여기서는 염색체 8번을 각각 방법에 따라서 샘플 별로 표준화 잔차 그림을 그린 것이다. 이것으로 각 방법 별로 모델이 적합하다는 것을 알 수 있다. 즉, 최소제곱법으로 회귀분석을 하는 것보다, 가중최소제곱법으로 적합 시킨 모형이 더 효과적이라는 것을 알 수 있다.

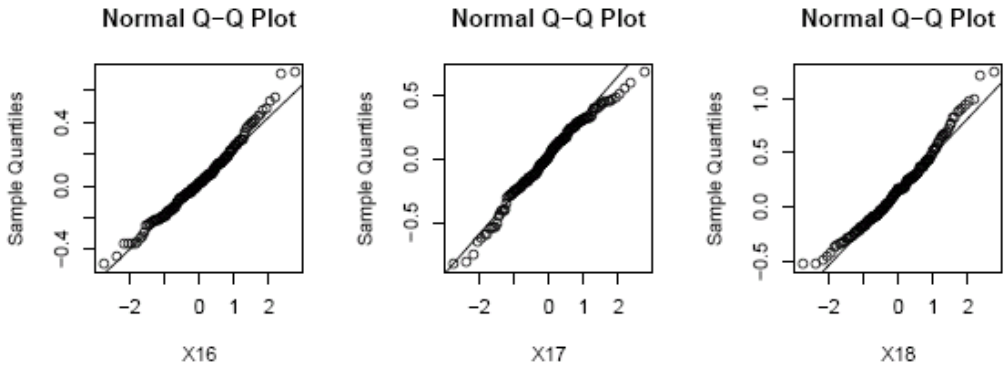
5. 6. 정규 확률 그림

각 방법을 적용한 결과 오차가 정규성을 따르는지에 대해서 정규 확률 그림으로 알아보았다. 아래 그림7을 보면 전체적으로 정규성을 따른다고 볼 수 있다.

a) M 1로 샘플 4, 5, 6 의 정규확률그림



b) M 2로 샘플 16, 17, 18 의 정규확률그림



c) M 3로 샘플 19, 20, 21 의 정규확률그림

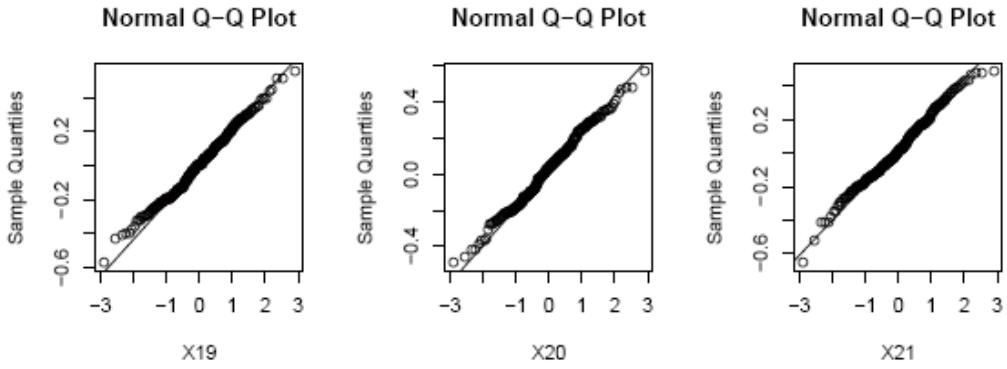
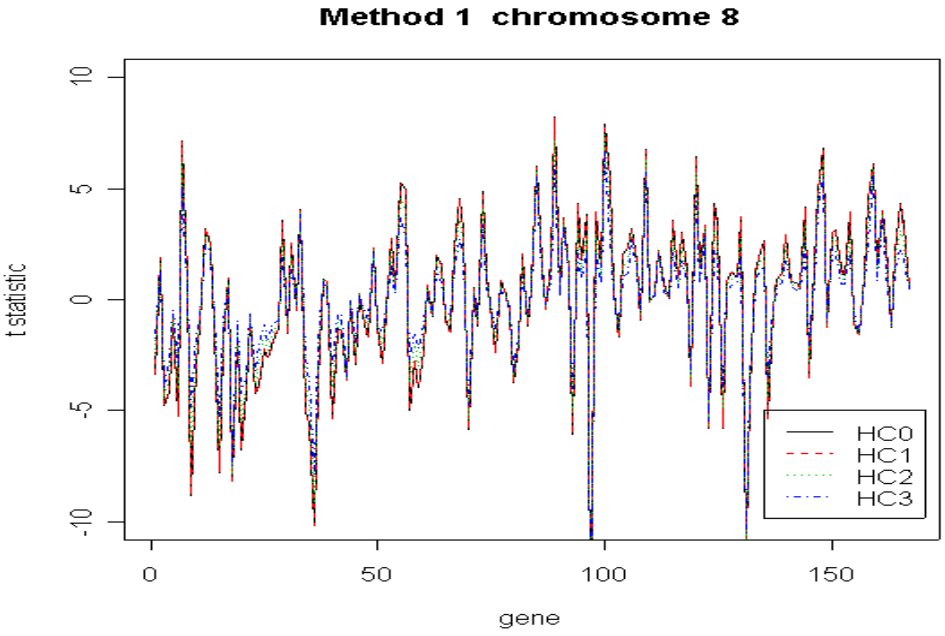


그림 7 . 각 방법별 정규 확률 그림

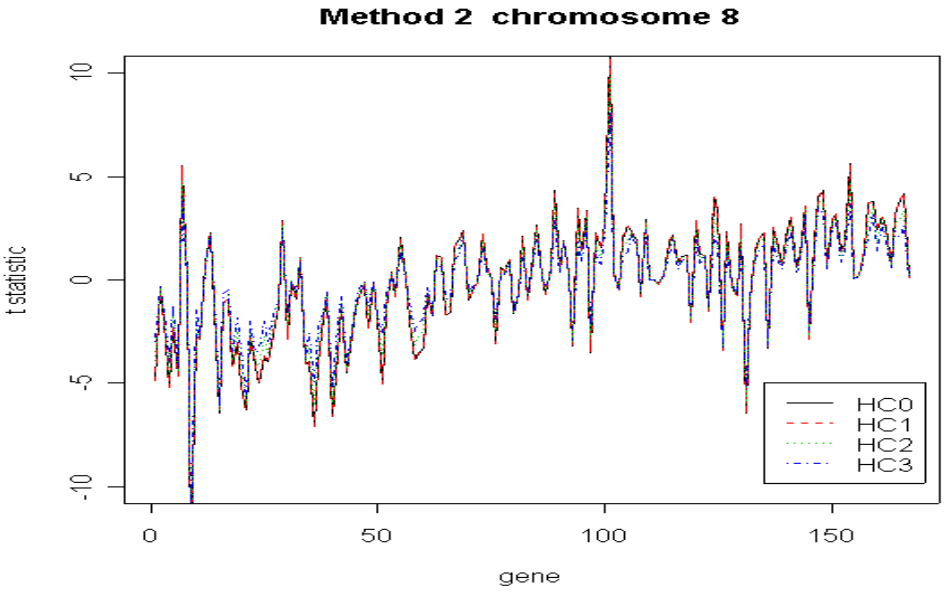
5. 7. t-통계량의 변화

4. 2에서 세운 통계량이 각 방법 에 따라 , 취하는 분산행렬에 따라서 어떤 변화가 있는지를 알아보기 위해서 x축은 유전자 이고 y축은 t^* 가 되는 그래프를 그려보았다.

a) M 1에서 t-통계량



b). M 2에서 t-통계량



c) M 3에서 t-통계량

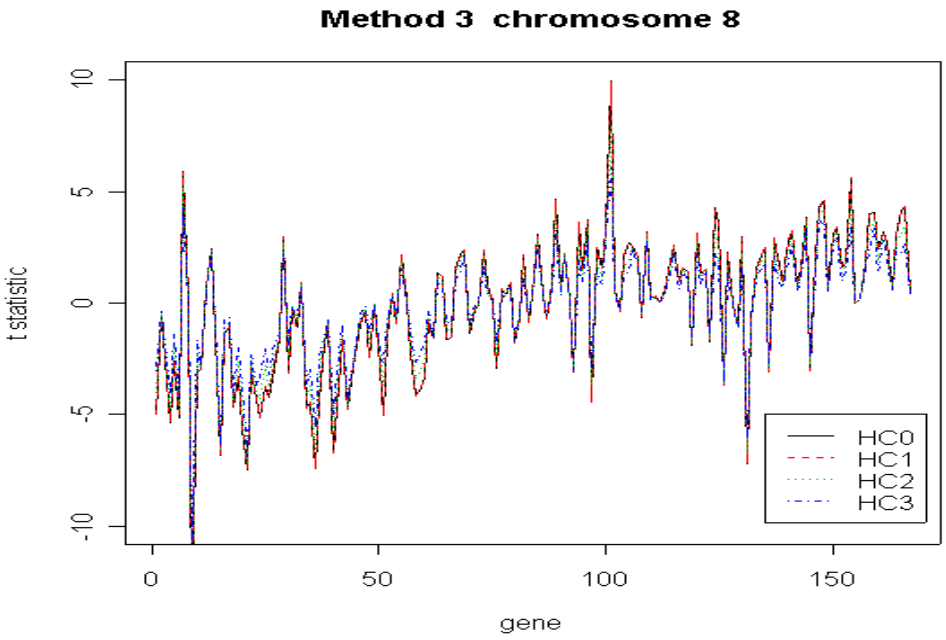


그림 8 . 각 방법별 t 통계량의 분포

각 방법 마다 4가지의 분산추정 방법으로 계산하였기 때문에 그림 8에서 볼 수 있듯이 각 방법 마다 4가지의 그래프가 그려진다. 이를 살펴보면, 각 각의 방법간에 패턴은 비슷하나, HC1(빨간색 선)의 경우는 분산을 작게 추정해서 상대적으로 t값이 큰 부분에서는 다른 것들 보다 더 크게 되고, HC3(파란색선)의 경우는 분산이 큰 관찰치의 “과대 영향”을 수정하는 효과를 갖고 있기 때문에 다른 것들에 비해서 큰 값에서는 작아지고 작은 값에서는 조금 커지는 모습을 보인다. 이는 후에 다중비교를 한 결과를 보면서 결과가 어떻게 달라지는 지를 살펴보기로 한다.

5. 8. 다중 검정 결과

다중검정은 위에서 구한 통계량으로 보정된 유의확률, Bonferroni, Holm, Sidak, Bonferroni and Hochberg, Benjamini and Yekutieli가 제시한 방법으로 각각의 보정된 유의확률을 구해준다.

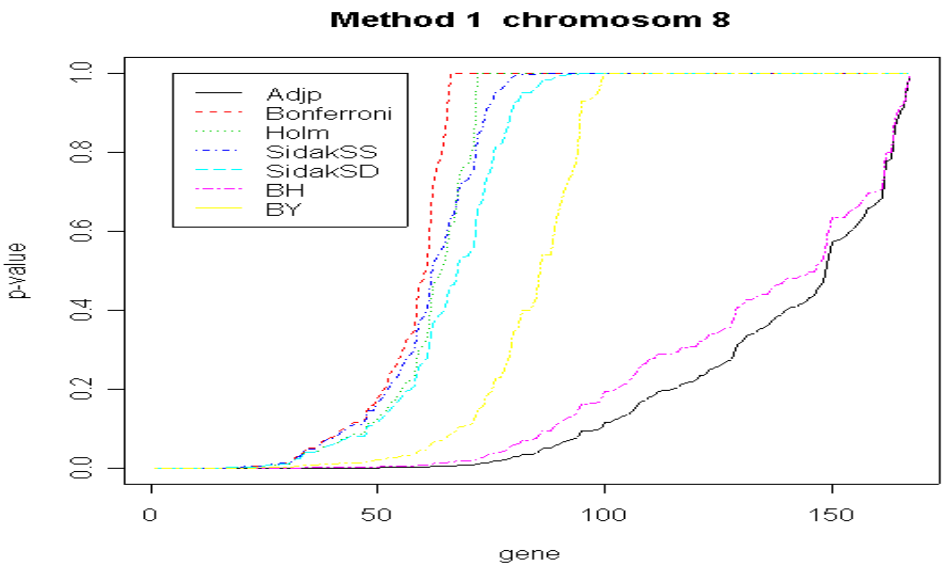
여기서는 R 2.0.0의 multtest package를 사용하여 계산하였으며, 본 논문이 제시하고 있는 M1, M2, M3 에 따라서 또한 다중검정방법들에 따라 그 결과가 어떻게 다른지에 대해서 유의확률 그림을 통해 알아보았다.

5. 8. 1. M1, M2, M3 별 유의 확률 그림

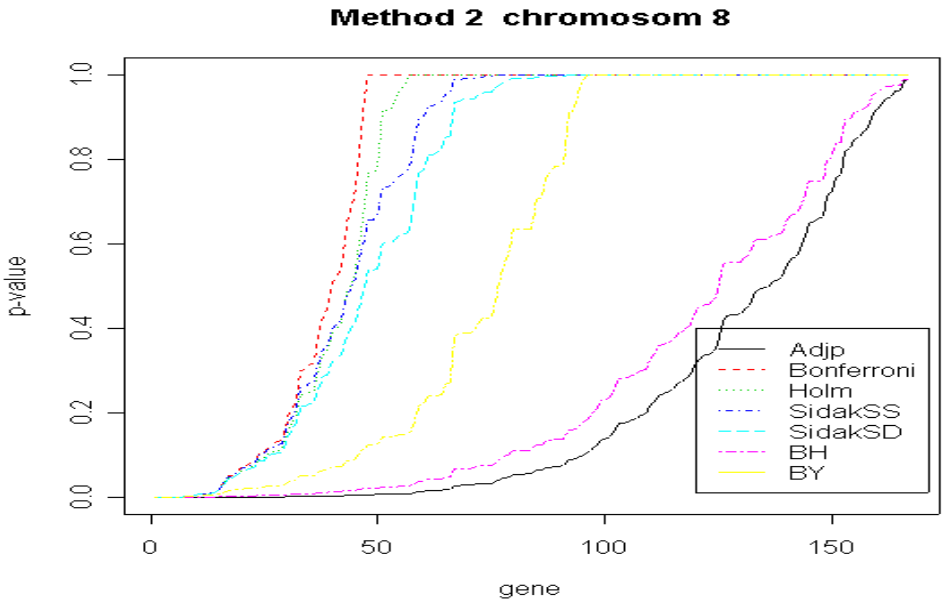
각 방법 별 유의 확률 그림을 그려보면 유의확률을 통제(control) 하는 과정이 FWER인지 아니면 FDR인지에 따라 차이가 있음을 알 수 있다.

유의확률 그림의 비교를 통하여 각 각의 방법이 얼마나 보수적인지, 혹은 조금 덜 보수적인 선택을 하는지를 알 수 있으며, 그 선택은 선택된 유전자의 수나 명단을 보고 결정하는 것이 좋다.

a) M 1에서의 다중검정별 유의한 유전자 선택의 차이



b) M 2에서의 다중검정별 유의한 유전자 선택의 차이



c) M 3에서의 다중검정별 유의한 유전자 선택의 차이

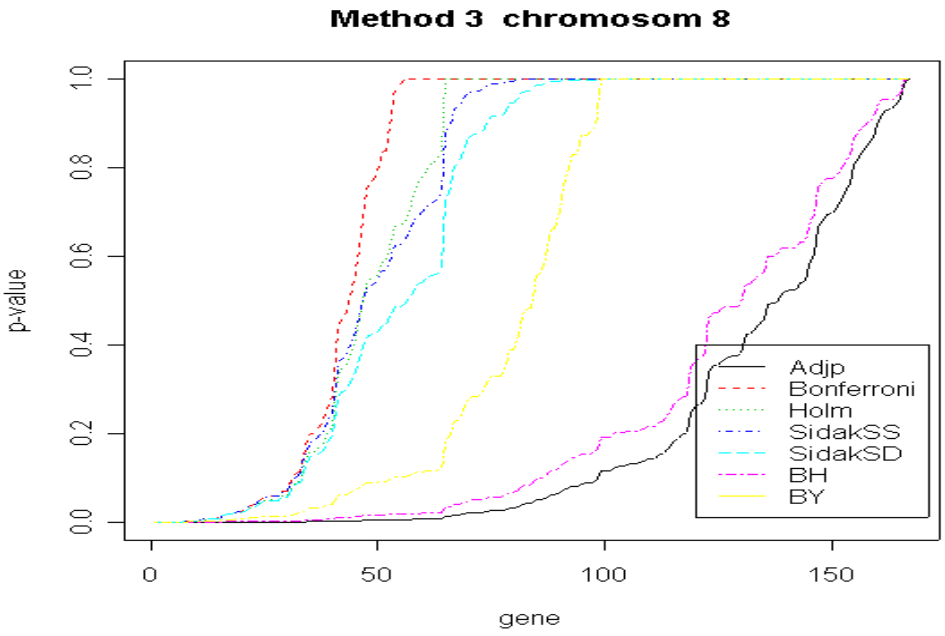


그림 9. M1, M2, M3 에서의 다중 검정별 유의한 유전자 선택의 차이

그림 9에서 보면 각 방법 M1, M2, M3별로 유의확률을 통제하는 방법에 따라 어떤 결과의 차이가 있는지를 보여주고 있다. 조정을 FWER 로 하는지 FDR로 하는지에 따라 Bonferroni, Holm, Sidak 이 제시한 방법으로 통제를 하는 경우에는 Bonferroni and Hochberg, Benjamini and Yekutieli보다 더 보수적인 결과를 내는 것을 알 수가 있고 유의수준을 0.01, 0.05 로 할 때에는 Bonferroni, Holm, Sidak의 결과는 별 차이가 없음을 알 수가 있다.

다중검정을 한 결과 유의한 유전자의 개수는 아래 표7에 정리하였다.

5. 8. 2. 유의한 유전자의 개수

표 7. 염색체 8번 167개의 유전자 중에서 유의한 유전자의 개수

	M 1				M 2				M 3			
Bonfe	HC0	HC1	HC2	HC3	HC0	HC1	HC2	HC3	HC0	HC1	HC2	HC3
≤0.01	27	26	18	13	13	13	7	6	15	15	8	6
≤0.05	34	33	24	16	17	16	12	7	24	23	15	9
Holm												
≤0.01	28	28	18	13	13	13	7	7	15	15	8	6
≤0.05	37	37	24	16	18	18	12	7	26	24	16	9
SidakSS												
≤0.01	27	26	18	13	13	13	7	6	15	15	8	6
≤0.05	36	33	24	16	18	16	12	7	24	23	15	9
SidakSD												
≤0.01	28	28	18	13	13	13	7	7	15	15	8	6
≤0.05	37	37	24	16	18	18	12	7	28	24	16	10
BH												
≤0.01	61	61	41	25	36	36	19	9	40	40	28	12
≤0.05	78	78	63	41	66	66	49	27	72	69	57	36
BY												
≤0.01	37	37	24	16	15	15	7	7	23	21	10	7
≤0.05	61	60	41	22	32	32	18	9	40	40	26	12

표7의 결과에 따르면, FWER로 통제 하는 Bonferroni, Holm, Sidak, 의 방법은 분산 추정 방법이 HC0 나 HC1에서 일반적으로 많은 유전자가 발견되고, HC3 방법에서 가장 적게 발견되었다. 그림9 에서 볼수 있듯이 FDR로 통제되는 Bonferroni and Hochberg (BH), Benjamini and Yekutieli (BY) 방법은 다른 것들에 비해 덜 보수적으로 선택되어 비교적 많은 유전자가 발견 되었다. 분산추정 방법에 따른 경향은 동일하다. 선택된 유전자들의 명단으로 알아보아도 전체적으로 M 1이 M 2나 M 3에 비해서 더 많은 유전자를 선택하고, 방법 2나 3에서는 방법 1에서 선택된 유전자들 중에서 더 보수적인 방법으로 선택된 유전자들을 유의한 유전자로 선택한다는 것을 알 수 있었다.

제 6장 토의 및 결론

지금까지 array-based CGH 자료를 이용하여 유의한 유전자를 찾아내는 방법을 가중최소제곱법에 기초한 회귀분석으로 알아보았다. 제시된 모형은 유전자와 샘플 간의 이질성을 고려한 가중최소제곱법을 이용하고, 오차의 이분산성을 고려한 이분산성 일치적 공분산 행렬을 이용하여 분산을 추정해서 회귀계수를 검정한 것이다. M1의 내용은 Thomas et al.가 제시한 모형과 비슷한 2 단계 회귀분석으로 표본 분산을 가중치로 고려하고, M 2와 M 3는 반복법을 응용 한 것으로 가중치로는 MSE를 사용하여, 회귀 계수 추정에 정확성을 더 하고자 하였다.

분석에 쓰인 자료는 결측값을 제외한 값으로 이루어진 자료인데, 실제적으로 많은 프로그램에서는 KNN 방법(Troyanskaya et al.,2001)을 이용하여 결측값을 보정한 후에 분석을 하고 있으나 방법간의 차이 이외의 편향성을 제거하기 위해 이 논문에서 결측값은 다루지 않았다.

완전한 자료만으로 유방암 자료를 분석해 본 결과, 각 3가지의 방법은 모두 유전자 간, 샘플간의 이질성을 보정하는 모델로서 적합하였으나 그 성격이 조금 틀렸다. 이를 설명하기 위하여 표준화 잔차 그림이나, 정규 확률도 그림을 통하여 통계학적 가정을 만족시키는 모형 이라는 것을 증명 하였다. 5. 4.의 표6에서와 같이 M 1에 비하여 M 2와 M 3는 특정 샘플 가법 요인인 δ_k 는 작게 추정되고, 특정 샘플 승법 요인인 λ_k는 더 크게 추정되어, 특정 샘플의 승법 요인이 더 크게 추정되는 M 2와 M 3에서 샘플간의 이질성을 줄여주면서, 그에 따르는 가중을 더 크게 둔다는 것을 알 수 있었다. 또한 4가지 방법의 회귀계수의 분산 추정 방법에 따른 다중 검정 결과, 같은 다중 검정 법으로 유의확률을 조정 했을 때, HC0 > HC1> HC2 > HC3 의 순서로 유전자를 선택하는 경향이 있음을 알 수 있었다. 다중 검정 법을 달리 했을 때 에는 FWER의 방법을 따르는지 혹은 FDR 방법을 따르는지에 따라 즉, 유의확률을 조정하는 것을 보수적 방법으로 하는지 조금 덜 보수적 인지에 따라서 선택되는 유전자의 개수가 달라진다는 것을 알 수 있었다.

유전자가 선택 되었다는 것은 그 자체가 특정 암에 영향을 준다고 의심할 수 있는 후보(candidate) 유전자이기 때문에, 이 논문에서 제시하는 방법 3가지 중에 어떠한 방법이 더 좋은 지를 논하는 것 보다는 각각 모형이 유의성이 있는 방법론으로 그 경향성을 알아보는 것이 중요하다고 생각된다.

특별히, 3번째 방법인 반복법은 특정 샘플 승법 요인인 λ_k가 아주 작게 추정되었을 경우에 종속 변수의 값을 더 크게 하여, 결과에 과대 영향을 주는 경향이 있는 단점이 있었다.

본 연구에서는 array-based CGH 자료에서 발현량이 일정하게 큰 유전자나 작은 유전자를 선택 하여 암에 유의한 유전자를 찾을 수 있는 방법을 실험 때문에 생기는 샘플간의 이질성과 유전자 간의 이분산성의 효과를 줄이는 가중 최소 제곱법에 기초한 회귀분석의 방법을 제시하였다. 이 방법으로는 궁극적으로 array-based CGH 자료에서 얻고자 하는 염색체의 증폭이나 결실이 되는 부분이나, 유전자들을 찾기에는 부족한 모형이므로 이런 부분들을 고려한 모형의 개발이 필요하다고 생각된다. 또한, 결측값을 추정하는 방법의 개발도 필요하다. 현재 제시되고 있는 방법들은 거리 개념을 도입한 것들이 대부분인데, 이는 결측값이 많은 자료에서는 단점이 많은 방법이라 여겨진다. 따라서 이러한 단점들을 보안할 수 있는 방법들에 대한 보다 심층적인 이해와 연구가 필요할 것이라 생각 된다.

Array-based CGH 자료를 분석할 수 있는 효율적이고 보편화된 프로그램의 개발 역시 필요하리라 여겨진다.

참 고 문 헌

강명욱, 김영일, 안철환, 이용구. *회귀분석: 모형개발과 진단*. 율곡출판사, 1995

Autio, R. et al. CGH-Plotter : Matlab toolbox for CGH-data analysis. *Bioinformatics*,2003;19:1714-1715

Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing.*J. R. Statist. Soc*. B,1995;57:289-300

Benjamini, Y., Yekutieli, D. The control of the false discovery rate in multiple hypothesis testing under dependency.*Annals of Statistics*,2001;29:1165-1188

Breusch, T. S., Pagan, A. R. A Simple Test for Heteroscedasticily and Random Coefficient Variation. *Econometrica*.1979;47:1287-1294

Carrll, R. J., Ruppert, D. *Transformation and Weighting in Regression*. 1988, Chapman and Hall, New York

Carroll, R. J., Cline, D. B. H. An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika,*1988;75:35-43

Cheng, C., Kimmel, R., Neiman, P., Zhao, L. P. Array rank order regression analysis for the detection of gene copy-number changes in human cancer. *Genomics*, 2003;82:122-129

Dalgaard, P. *Introductory Statistics with R*. 2002, Springer-Verlag,New York

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. Cluster analysis and display of genome-wide expression patterns.*Proc.Natl.Acad Sci*, 1998;95:14863-14868

Dudoit, S., Yee, H. Y., Callow, M. J., Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 2002;12:111-139

Fridlyand, J. et al. Hidden Markov Models Approach to the Analysis of Array

CGH data. *Elsevier Science* accepted ,2004

Fritz, B. et al. Microarray-based Copy Number and Expression Profiling in Dedifferentiated and Pleomorphic Liposarcoma. *Cancer Research*, 2002;62:2993-2998

Glejser, H. A New test for Heteroskedasticity. *Journal of the American Statistical Association*,1969;64:316-323

Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrik*,.1988;75:800-802

Holm, S. A simple sequentially rejective multiple test procedure. Scand.J. Statis,.1979;6:65-70

Hupe, P. et al. Analysis of array CGH data: from singal ratio to gain and loss of DNA regions. *Bioinformatics*, 2004,12;20(18):3413-3422

Ihaka, R., Gentleman, R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*,1996;5:299-314

Kerr, M. K., Martin, M., Churchill, G. A. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*,2000;7:819-837

Long, J. S., Ervin, L. H. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 2000;54:217-224

Monni, O. et al. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer.*PNAS*,2001;98:5711-5716

Moore, D. H., Pallavicini, M., Cher, M. L., Gray, J. W.A t-Statistic for Objective Interpretation of Comparative Genomic Hybridization (CGH) Profiles. *Cytometry*,1997;28:183-190

Myers, C. L., Dunham, M. J., Kung, S. Y., Troyanskaya, O. G. Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*,2004;20(18):3533-3543

Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. *Applied Linear Statistical Models* .1996,IRWIN, USA

Pan, W., A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics*,2002;18:546-554

Pollack, J. R. et al Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature genetics,*1999;23:41-46

Pollack, J. R. et al Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*,2002;99:12963-12968

Rao, C. R. Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association*.1970;65:161-172

Segal, M. R.,Dahlquist, K. M.,Conklin, B. R. Regression Approaches for Microarray Data Analysis. *Journal of Computational Biology*, 2003;10:961-980

Thomas, J. G., Olson, J. M., Tapscott, S. J., Zhao, L. P. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles.*Genome Research*,2001;11:1227-1236

Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*,2001;17:520-525

Venables, W. N., Ripley, B. D. *Modern Applied Statistics with S*. 2002,4th edition, Springer-Verlag.New York

Wang, P. et al. A method for calling gains and losses in array CGH data. 2004; *Biostatistics* accepted

Wang, Y., Guo, S. W. Statistical methods for detecing genomic alterations through array-based comparative genomic hybridization (CGH). *Frontiers in Bioscience*, 2004;9:540-549

White, H. A Heteroskedastic Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity.*Journal of Econometrics*, 1980;48:817-838

Wilhelm, M. et al. Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Research*,2002;62:957-960

Abstract

Regression Method for Detecting Significant Genes in Array-based Comparative Genomic Hibridization Data

Kim, Shin Young
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

It is inappropriate to apply the statistical methods used in the analysis of microarray data to array-based CGH data directly. In this study, we developed a regression-based method using weighted least square for detecting significant genes in the array-based CGH data. We used Heteroscedasticity Consistent Covariance Matrix(HCCM) to estimate the variance of regression coefficients.

To compare the trends of selecting significant genes, we analyzed the breast cancer data in Stanford Genomics Breast Cancer Consortium using our three methods which are simple two step regression method(M1), simple non-iteration method(M2) and iteration method(M3).

Our results showed that M2 and M3 were more effective than M1 for adjusting heterogeneities of genes and samples and in the aspects of gene selection, M2 and M3 were more conservative than M1.

Key words : Array-based CGH, Weighted least square, HCCM, Simple two step regression, Simple non-iteration, Iteration