안녕하세요. 오늘 박사학위논문심사에 참석해주시려고 귀한시간 내주셔서 진심으로 감사드립니다.

본래 종심때는 예심때 지적받은 사항을 바탕으로 어떻게 보완하였는지는 주로 발표하는 것으로 알고 있는데,

예심 이후로 많은 부분이 추가되어서 이에 대한 내용도 같이 발표를 하게 된 점 양해 부탁드립니다.

오늘 발표는 학위논문의 outline에 대해서 말씀드리고, 예심 이후에 업데이트 된 내용을 15분 이내로 간략히 발표하도록 하겠습니다.

마지막으로 예심 때 지적받은 사항을 어떻게 보완하였는지 5분 이내로 말씀을 드리고 마무리 하겠습니다.

학위 논문은 아래와 같은 다섯 챕터로 구성되어 있습니다.

첫번째 챕터에서는 gwas에 대한 overview와, binary trait에서 heritability를 추정하는 방법에 대한 기존의 방법을 다루고, 이 논문의 목적과 어떻게 구성되어 있는지를 다루었습니다.

두번째 챕터에서는 독립적인 sample에 대한 GWAS 분석에 대한 결과와 relevant gene을 찾는 방법에 대하여 다루고 있습니다.

세번째 챕터에서는 GWAS의 통계적 검정력 향상을 위해서 case와 control을 선별하는 방법을 다루고 있습니다.

네번째 챕터에서는 liability threshold 모형을 바탕으로 ascertained family-based sample에서 이분형 표현형의 유전율을 추정하는 방법에 대하여 다루고 있고,

마지막 챕터에서는 전체 요약과 결론을 다루고 있습니다.

이 중 예심 이후에 크게 업데이트 된 부분이 chapter 2와 4입니다. 이 부분에 대한 방법과 결과를 말씀드리겠습니다.

챕터 2에서는 LAM이라는 질병의 GWAS 분석결과를 다루고 있습니다. LAM이라는 질병에 대해서 간단히 말씀드리자면, 3,40 대의 가임기 여성에게만 거의 제한적으로 발생하는 폐 관련 질병으로 매우 희귀하게 발생합니다. 이 질병의 발생기작에 대해서는 크게 알려진바가 없으나, TSC라는 질병을 가진 사람의 약 10%는 LAM이라는 질병을 같이 동반하는 것으로 알려져있고, LAM 환자 중 TSC를 동시에 앓고 있는 환자를 TSC-LAM이라고 일컫습니다. TSC는 TSC1 혹은 TSC2의 뮤테이션으로 인해서 발생하는 것으로 알려져 있습니다.

반면에 TSC를 앓고있지 않는 LAM 환자를 sparadic LAM환자라고 부르고, 이 연구에서는 S-LAM과 연관성이 있는 유전변이를 찾는 것을 목적으로 합니다.

그 결과 크로모좀 15에서 두개의 signal을 발견하였으며, 두 snp은 매우 강한 correlation을 보이고 있고 같은 LD 블락에 속해있었습니다.

이 signal은 replication data에서도 유의한 결과를 보였습니다.

이 두 signal은 intergenic region에 속해있었으며, 주변에 세개의 protein coding gene이 위치했습니다.

이 세개 gene과 SNP 사이의 interaction을 보기 위하여 Hi-C 데이터를 기반으로 한 크로마틴 TAD를 살펴보았습니다.

같이 연구를 진행한 의사선생님들의 의견을 받아 이 질병과 가장 연관이 있는 4개 티슈를 살펴보았습니다.

이 그림에서 각 cell의 색깔은 두 지역간의 interaction의 정도를 의미하고 빨간색을 띌수록 강한 interaction을 의미합니다.

까만색 동그라미는 NR2F2와 두 시그널에 대한 부분이고 주변의 지역이 서로서로 interact하고 있는것을 확인하실 수 있습니다.

반면에 다른 두 protein 코딩 gene은 같은 TAD에 속해있지 않았습니다.

저희는 이것을 바탕으로 NR2F2라는 gene에 더 주의를 기울였습니다.

NR2F2는 다른 cancer tissue와 비교하였을 때, LAM 환자에서 그 발현량이 매우 높았습니다.

다른 normal 티슈와 비교하였을 때, 모든 티슈에 대하여 높지 않았지만 다른 tissue들의 발현량의 median보다 높게 발현되고 있음을 확인할 수 있습니다.

NR2F2의 발현은 실험적으로도 폐 티슈에서 높게 발현되는 것을 확인하였습니다.

여기에서 보이는 이 갈색 점들이 발현된 된 것인데, 오른쪽 하단의 사진을 보면 발현되지 않는 경우에는 이런 갈색의 점이 나타나지 않는 것을 확인할 수 있습니다.

결론으로, 이 연구를 통하여 기존에 진행된 적없는 S-LAM 질병에 대한 GWAS를 실시하였고 chr15에서 두개의 signal을 찾았습니다.

두 snp이 intergenic region에 있어서 직접적으로 어떤 영향을 주는지는 알 수 없었지만, 주변에 있었던 NR2F2라는 gene과 밀접하게 interact하고 있는 것을 확인하였습니다.

NR2F2는 LAM환자 내에서 다른 cancer 티슈들과 비교하였을 때 높은 발현량을 보였으며, normal tissue와 비교하였을 때 평균보다 높은 발현량을 보였습니다.

또한 IHC 분석을 통하여 실제 폐 세포에서 NR2F2가 높게 발현하고 있는 것을 확인하였습니다.

LAM은 여성에게만 일어나는 질병으로 에스트로겐과 밀접한 연관을 가지고 있습니다.

기존의 연구를 통해 estrogen과 관련된 gene의 일부를 knowdown하는 실험을 통하여 NR2F2 발현의 감소가 확인되었습니다.

이것으로 미루어보아 LAM과 NR2F2, 그리고 NR2F2와 에스트로겐 사이의 연결점을 찾을 수 있었고, 기존의 TSC로는 설명되지 않았던 여성에게만 거의 제한적으로 일어나는 이유에 대해서도 설명이 가능했습니다.

그다음으로 Chapter4에서 Heritability estimatin에 대한 내용의 보충입니다.

노테이션은 기존과 동일하고, Y는 phenotype L은 관측되지 않은 liability, 그리고 이 둘은 liability threshold모형에 의하여 아래와 같은 관계를 가지고 있습니다.

이 관계를 통하여 Y를 알고있다면 L의 범위를 알수 있고 이 범위는 각각 a와 b로 표시하였습니다.

벡터 폼입니다.

우리는 liability가 multivariate normal을 따른다고 가정하였고, 이때 분산 공분산 행렬이 heritability와 kinship coefficient로 이루어져 있다고 가정하였습니다.

이를 바탕으로 joint likelihood는 아래와 같이 표현이 되고, 이 lieklihood에 대하여 h2과 beta의 mle를 구하기 위하여 em algorithm을 사용하였었습니다.

추가적으로 proband에 의하여 family 가 ascertain 되었을 때의 추정방법을 개발하였습니다.

proband와 nonproband의 notation을 각각 P와 NP를 이용해서 나타내면, 이 경우 고려해야하는 likelihood는 아래의 conditional likelihood입니다.

여기에 로그를 취하면 모든 사람들의 log llieklihood에서 proband의 likelihood를 뺀 값이 되는데,

이 값의 maximizer를 구하기 위하여 EM algorithm의 방법을 이용하여 global lower bound를 구하게 됩니다.

이때 F는 전체 sample의 liklihood의 lowerboudn이고 g는 proand의 upper bound입니다.

F는 기존의 EM algorithm의 방법을 그대로 가지고 와서, 아래와 같이 표현이 되고,

G는 자기자신의 임의의 양수를 더한 것을 upper bound로 사용할 수 있습니다.

이를 바탕으로 global lowerbound는 em algorithm의 expectation step과, proband 의 lieklihood의 합으로 표현되고, 이 값을 수렴할 때 까지 계속 maximize 시켜서 모수의 mle를 구합니다.

구한 모수의 통계적 유의성을 보기 위하여 score test를 계산하였고, empirical information matrix를 이용하여 score의 분산을 계산하였습니다.

그 결과 beta와 h2가 각각 0 일때에 대한 검정을 실시할 수 있습니다.

이 방법의 타당성을 평가하기 위하여 simulation을 싫시하였습니다.

maf 0.2, beta 0.1253, h2= 0.005, 0.2, 0.4

prevalence 0.05,0.1,0.2를 가정하였습니다.

시뮬레이션은 샘플링 스킴에 따라 두가지 시나리오로 진행되었습니다.

첫번째로 randoml하게 family를 선택한 경우로 각 replicate마다 500개의 가족을 생성했습니다.

두번째로 ascertained가족에 대한 경우로 먼저 50,000가족을 생성한 뒤에 affected proband 500명을 추출하여 그의 가족을 분석에 사용하였습니다.

모든 결과는 GCTA라는 기존의 binary trait의 heritability 추정 알고리즘과 비교하였습니다.

우선 LTMH라는 우리의 방법에서 beta는 0.1253의 true값 주변에서 잘 추정되는 것을 확인할 수 있고, h2의 경우는 GCTA와 둘다 잘 추정하는 것처럼 보입니다.

그러나 박스플랏을 보면 모든 경우에서 우리의 방법이 더true값에 근접하고 더 적은 se를 가지고 있습니다.

우리 결과의 강점은 ascertained인 경우에 잘 드러나는데 GCTA는 전혀 추정을 하지 못하는 반면에 우리 방법은 굉장히 추정을 잘 하고 있고, 심지어 scenario 1보다 더 적은 se를 가ㅣ고 있습니다.

score test 결과입니다. 첫번째 행은 heritability가 실제로 0일 때 각 significance level 별로 기각한 test의 비율입니다. 보시면 저희 방법이 다소 보수적인 경향이 있습니다만, 실제로 귀무가설이 참이 아닐 때, 꽤 높은 수준의 power를 확보하는 것을 보실 수 있습니다. 특히 ascertained family에서 높은 power를 보이고 있습니다.

beta에 대한 추정 결과도 유전율과 비슷하게 해석 가능합니다.

ha2이 0일 때, 즉 beta가 0일때 empirical size는 유의수준보다 작아서 마찬가지로 보수적이지만, beta가 0이 아닐 때 power를 어느정도 확보하고 있습니다.

실데이터 분석으로는 서울대 병원에서 모집한 681 proband에 대하여 681개의 가족데이터를 분석하였습니다. 이 proband는 모두 제 2형 당뇨에 걸려있습니다.

이 중 나이 정보가 있는 proband는 분석에 이용하게 되어 총 648명의 proband와 그의 가족들이 분석에 포함되었습니다.

공변량으로는 나이를 사용하였습니다.

추정된 제 2형 당뇨의 유전율은 29.44%였습니다. 이 값은 기존에 추정된 값인 26%보다 다소 크지만 인종의 차이에서 비롯된 값의 차이라고 판단됩니다.

나이의 계수 추정량은 0.051로, 이는 나이가 1 증가할 때, 한 사람의 liability의 threshold가 0.051 감소하게 되어, 질병에 걸릴 가능성이 더 높아지는 것을 의미합니다.

이 결과는 그래프를 통해 좀 더 명확히 이해 가능한데, 왼쪽 그림을 보시면 부모와 자식 두명으로 이루어진 핵가족에서 한 프로밴드가 0명에서 3명의 affected relative를 가질 때의 나이 별 질병에 걸릴 확률입니다.

모든 경우에서 나이가 증가할 수록 이 확률이 커지고, 또 가족력이 많은 사람이 더 큰 위험을 가지고 있습니다.

또한 이 값을 아무런 가족정보가 없는 사람과의 relative risk를 계산하게 되면, 나이가 어릴수록 그 가족들의 나이도 어려지기 때문에 relative risk가 굉장히 커지는 것을 확인할 수 있습니다. 이 그림을 통해서 relatives affected at early onset 들의 중요성에 대해서 이해할 수 있습니다.

score test를 S-LAM 데이터에 적용하여 GWAS를 돌린 결과, conditional logistic 분서보다 더 작은 p-value를 보이는 것을 확인할 수 있었습니다.

결론으로, 이 방법을 통해 이분형표현형의 유전율 추정 방법을 제안하였고, 특히 이 방법은 ascertained family에서 효과적으로 추정하엿습니다. 현재까지 ascertained family에서 유전율추정하는 방법이 잘 개발되어있지 않았던것을 고려한다면, 이 방법으로 기존에 분석할 수 없었던 자료에 적용이 가능하다고 생각합니다.

시뮬ㄹ이션 분석을 통해 GCTA에와의 비교를 하였고, 우리 분석이 좀 더 정확한 추정값을 보였습니다.

다만 score test가 다소 보수적으로 검정을 하였지만, statistical power에는 큰 영향이 없엇습니다.

그러나 이 방법의 개발 중 가장 어려웠던 것은 계산량인데, 계산량이 가장 많은 부분이 em algorithm의 expectatino 부분이었습니다. 이 부분은 em algorithm 자체의 반복수를 줄이거나 혹은 e step의 계산량 자체를 줄이는 것으로 해결할 수 있습니다. 이 부분은 추후 연구에서 다룰 ㅇ정입니다.