연속형 표현형에 대한 전장유전체 연관분석에서의 상위성 탐색방법론 비교

저자

이종현1), 이재원2)

1) 서울특별시 성북구 안암로 145 고려대학교 안암캠퍼스 정경관 305호. 2013150250@korea.ac.kr

2) 서울특별시 성북구 안암로 145 고려대학교 안암캠퍼스 정경관 305호. jael@korea.ac.kr

서론

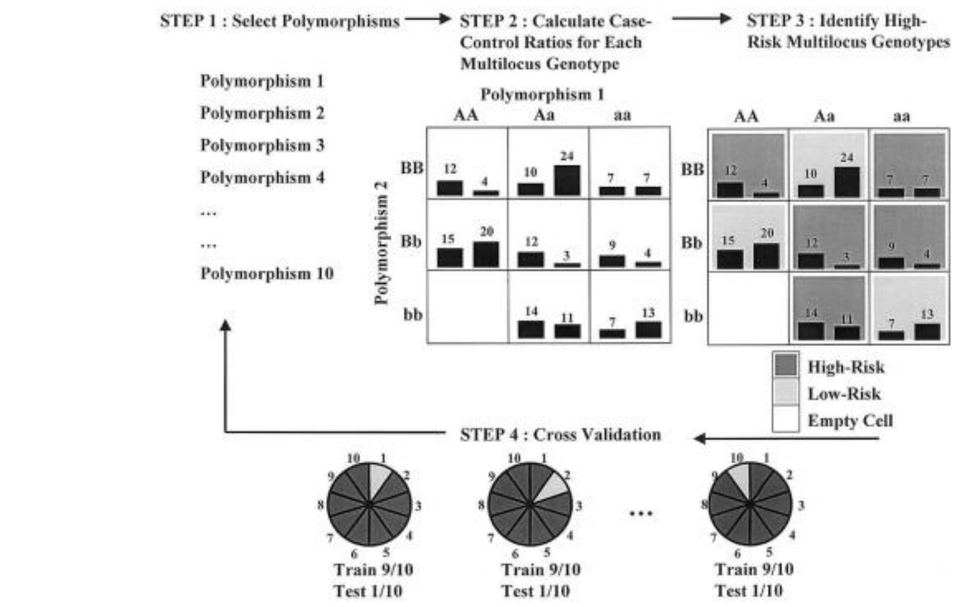
단일 뉴클레오타이드 다형성(Single Neucleotide Polymorphism; SNP)은 DNA 염기서열 중 대상인구 1% 이상의 발현율로 변이가 일어난 유전자좌의 변이를 뜻한다. Missing heritability 문제의 해결 가능성을 탐색하던 중 gene-gene interaction인 상위성(epistasis)의 중요성이 제기되었다. 로지스틱 회귀와 같은 기존의 통계학적 모형과 달리, 통계학적 가정으로부터 자유롭게 gene-gene interaction을 찾아내기 위한 방법으로 Multifactor Dimensionality Reduction (MDR)이 개발되었다.

가정에 구애받지 않고, gene-gene interaction을 찾아낼 수 있는 MDR은 missing heritability 문제를 해결하기 위한 새로운 관점을 제시했고, MDR의 기본 형식을 이용한 확장된 방법들이 등장하였다. 현재 방법론들의 부분적인 비교만이 수행되었고, 모형들을 아우르는 전반적인 비교는 이루어지지 않았다.

본 연구에서는 다양한 분포 상황을 가정한 모의실험을 통하여 연구된 방법들을 비교하고자 한다. 분포 상황 별 최적의 방법을 제시하고, 다양한 상황에서 최적의 다인자 차원 축소 방법을 찾기 위한 가이드라인을 제시하고자 한다.

다인자 차원 축소 방법

- **Step1. Global Process**
  - 실험군과 대조군의 비율을 Global Ratio로 둔다.
  - 데이터를 임의로 10개의 부그룹으로 나눈다.
  - 모형 평가를 위해, 9개의 부그룹을 training set, 1개의 부그룹을 test set으로 지정한다.
- **Step2. Local Process**
  - m차원 상위성을 탐색하고 싶다고 가정한다면, 가능한 모든 SNP으로 만들 수 있는 m개 SNP의 조합을 고려한다.
  - 가능한 모든 SNPs조합에 대하여, 유전 조합별로, 실험군과 대조군의 비율을 Local Ratio로 둔다.
  - Local Ratio가 Global Ratio보다 크거나 같은 유전 조합은 High-risk group으로, 그렇지 않은 유전 조합은 Low-risk group으로 분류한다.
- **Step3. Evaluation**
  - 가능한 모든 SNPs조합에 대하여, Test set에서의 오분류율이 가장 낮은 모형을 최적 모형으로 선택한다.
- **Step4. Selecting The Best Model**
  - 10번의 시행을 통해 Cross-Validation Consistency(CVC)가 가장 높은 모형이 최종 모형으로 선택된다.



Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001;69(1):138-147. doi:10.1086/321276

비교방법

- Fuzzy set based Generalized Multifactor Dimensionality Reduction(FGMDR)
- Quantitative Multifactor Dimensionality Reduction(QMDR)
- Cluster-based Multifactor Dimensionality Reduction(CLMDR)
- Unified Model Based Multifactor Dimensionality Reduction(umMDR)
- Generalized Multifactor Dimensionality Reduction(GMDR)

모의실험

- **데이터 구조**
  - 1,2번 SNPs에 대하여 상위성 모형 설정.
  - Heritability 값에 따라 상위성이 증가하도록 모형 설계.
  - 정규분포, 감마분포 및 주효과의 유무에 따라 다른 데이터셋 생성.
- **주효과가 없는 정규분포**

$$Y|(SNP_1 = i, SNP_2 = j) \sim N(f_{ij}, 1)$$

- **주효과가 있는 정규분포**

$$Y|(SNP_1 = i, SNP_2 = j) \sim N(f_{ij}, 1) + SNP_2 * N(1,1)$$

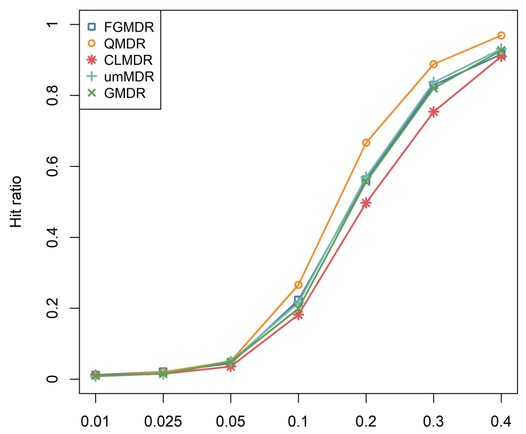
- **주효과가 없는 감마분포**

$$Y|(SNP_1 = i, SNP_2 = j) \sim Gamma(f_{ij}^2, \frac{1}{f_{ij}})$$

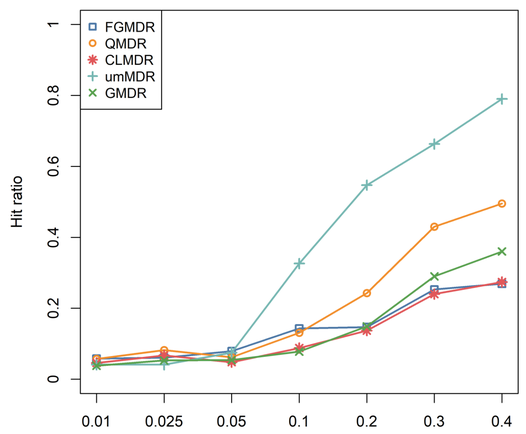
- **주효과가 있는 감마분포**

$$Y|(SNP_1 = i, SNP_2 = j) \sim Gamma\left(f_{ij}^2, \frac{1}{f_{ij}}\right) + \mathrm{SNP}_2 * Gamma(1,1)$$

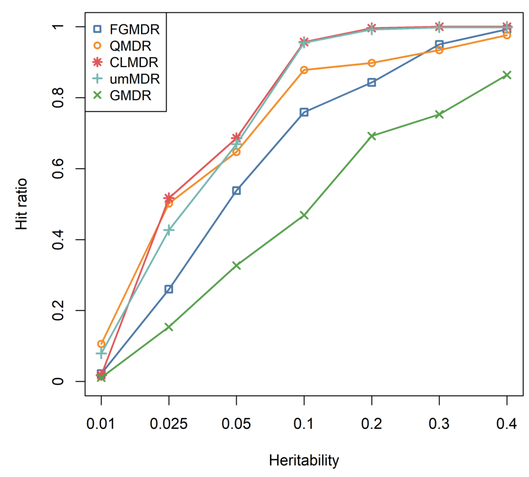
주효과가 없는 정규분포



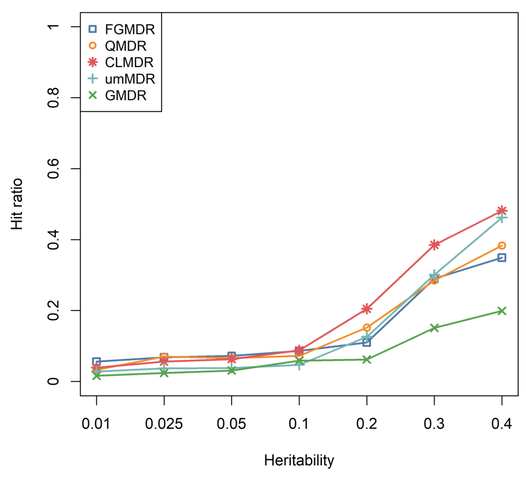
주효과가 있는 정규분포



주효과가 없는 감마분포



주효과가 있는 감마분포



결론

정규분포를 따르는 표현형에서 QMDR의 성능이 보편적으로 좋다. 주효과가 없는 정규분포에서 Heritability값이 0.05이상인 구간에서는 QMDR의 성능이 가장 좋았으며, 주효과가 있는 정규분포에서 umMDR을 제외하고, Heritability값이 0.2이상인 구간에서 QMDR의 성능이 두 번째로 좋았다.

umMDR은 주효과가 있는 모의실험 데이터에서 성능이 좋았다. 주효과가 있는 정규분포에서는 heritability값이 0.1이상인 구간에서 umMDR이 압도적인 성능을 보였고, 주효과가 있는 감마분포에서도 CLMDR을 제외하고 두 번째로 성능이 좋았으며, CLMDR과의 차이도 크지 않았다.

CLMDR의 경우 정규분포를 따르는 모의실험 데이터에서는 성능이 좋지 않았으나, 감마분포를 따르는 모의실험 데이터에서는 보편적으로 좋은 성능을 보였다.

FGMDR은 heritability값이 낮은 구간에서는 비교적 좋은 성능을 보였으나, heritability값이 높은 구간에서는 다른 방법에 비해 성능이 좋지 않았고, GMDR은 보편적으로 좋지 않은 성능을 보였다.

참고문헌

- Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001;69(1):138-147. doi:10.1086/321276
- 이종현. "단변량 수치 표현형에서의 상위성 탐색방법 비교연구." 국내석사학위논문 고려대학교 대학원, 2021. 서울