<GWAS 논문 1 리뷰 내용 요약>

2022.11.07

<참고 논문 주소>

GWAS 리뷰 논문 : https://www.nature.com/articles/s43586-021-00056-9

- GWAS : 특정 질병과 관련 있는 여러 유전적 변이에 대한 통계적 연구
- 표현형에 대한 인사이트, 유전성, 유전적 상관관계 계산, 질병 위험 계산, 신약 개발, 위험 요인과 건강에 대한 추론
- genotype과 phenotype의 관계 파악 : allele frequency 기반. 개인마다 다름
- 주로 SNP
- genetic 그룹 간의 차이 계산.
- 개개인의 신체적, 정신적 질병 예상.
- PRS를 통한 게놈 위험 예측을 가지고 질병 위험 파악 가능. 임상적 사용.
- the collection of DNA and phenotypic information from a group of individuals (such as disease status and demographic information such as age and sex); genotyping of each individual using available GWAS arrays or sequencing strategies; quality control; imputation of untyped variants using haplotype phasing and reference populations; conducting the statistical test for association; conducting a meta-analysis (optional); seeking an independent replication; and interpreting the results by conducting multiple post-GWAS analyses
- 표본을 구하는 과정 중요 : 공공 데이터 사용, proxy phenotype 사용.
- 조상을 고려하는 것이 중요 : 대조군이 속한 환경이나 대륙 파악.
- GWAS를 위해 선형 회귀 모형 혹은 로지스틱 회귀모형을 사용. 이때 데이터에 있는 개인적인 정보(proxy phenotype)를 넣어주면 게놈 발견에 있어 통계적 능력을 향상시키고 제어력을 높임. 하지만 각 데이터의 정보들이 서로 완전 독립이 아니고 관계가 있을 수 있음에 유의.
- 데이터가 중요 – biobank 사용.
- p-value 형태의 결과 : R, FUMA, LocusZoom등의 tool을 사용해서 처리할 것.

<GWAS 수행 목적>

- 관련 변이 식별을 위해 수행. 게놈 위치와 변이의 연관 정도를 보여주기 위해 맨해튼 플롯으로 시각화.
- statistical fine mapping : 인과적 변이 가능성이 있는 변이 집합 식별.
- 변이에 의해 영향 받는 epigenomic effect 식별.
- 그에 대한 target gene 식별.
- 효과를 주고받는 pathway 판별.

- 변이에 여러가지 인과적 원인이 있을 수 있음 - statistical fine mapping : 상관관계가 높은 SNP가 1순위.
- GWAS를 수행하는 주요 동기는 식별된 연관성을 사용하여 유전 가능한 표현형의 생물학적 원인을 결정하고 잠재적으로 가능성 있는 치료를 위한 시작점을 제공하는 것. 하지만 변이를 찾아내고 매핑을 통해 그에 맞는 SNP들을 알아낸다 해도, 변이들의 생물학적 의미를 알기는 어려움
- 영향을 받는 유전자를 알아내야 함 – 우선순위 부여. 매핑된 GWAS의 2~3%는 추론 가능. 그러나 대부분은 그 기능이 알려져 있지 않고 코딩되지 않는 영역 (eQTL, QTL, molQTL 등의 tool을 사용하여 정보를 알아낼 것).
- 조절 경로 및 세포 효과 결정 - GWAS 및 post GWAS 분석에서 확인된 유전자를 테스트하여 Magama 및 MAPRY와 같은 도구를 사용하여 수렴 기능을 수행. 특정 생물학적 경로에 관여하거나 특정 조직, 세포 유형, 발달 단계 또는 단백질 네트워크에 연결된 유전자 세트를 테스트. 무작위로 선택된 유전자 세트는 생물학적으로 의미가 없음. 또는 Trans-molQ, Trans-eQTL 사용. 많은 샘플의 분자 데이터 필요. 세포의 유형 혹은 세포의 상태는 모든 기능적 해석 작업에 필수적. 네트워크 효과 분석에 중요.
- 결국 GWAS는 통계적으로 연관된 변이체를 정확히 집어내고 생물학적 맥락에서 이러한 변이체의 역할을 알아냄. 또한 질병 위험 예측, 형질의 유전적 구조 파악.
- PRS (Polygenic Risk Score) : 독립 코호트의 GWAS 분석 통계를 사용하여 대상 코호트의 질병 위험을 예측. 질병 위험이 높은 개인 식별. GWAS 효과에 기초한 가중치를 바탕으로 risk allele의 가중 합계 점수로 계산됨. 간단한 계산 방법은 형질 간의 통계적 연관성의 p value에 기초하여 SNP 하위 집합을 선택하는 것. 정확성은 표현형이 연속형인지 이진형인지에 따라 다른 방법으로 측정. 만약 GWAS와 코호트가 동일한 개인을 공유하는 경우 정확성은 과대포장 될 수 있음.
- 연속 형질은 결정계수(R square) 값으로 정량화. GWAS 회귀모형으로 PRS를 계산할 때, 일반적으로 연령, 성별, 조상과 같은 covariants가 포함됨. PRS의 효과는 두 모델에서 설명되는 분산의 차이를 사용해서 평가됨. 귀무가설. 대립가설.
- 이진 형질의 persudo R square 값은 로지스틱 회귀모형을 사용하여 계산됨. PRS 정확도를 평가하기 위해 일반적으로 사용되는 방법은 AUC 방법. 두 그룹을 구별하는 것이 목표일 때 모델의 성능을 정량화. 이때 어떤 개인을 고위험군으로 분류할지에 대한 판단 기준으로는 임계값 사용. 그렇지만 AUC 나 Odd ratio가 크다고 해서 고위험군이 많다는 것은 아님. 다른 방법을 사용하면 그 결과가 또 달라짐. (ex) net reclassification index. 하지만 GWAS를 만들 때 사용한 코호트와 대상 코호트 간의 조상 거리가 멀수록 PRS의 정확도가 낮아짐. 따라서 GWAS를 만들 때 다양한 코호트를 대상으로 만드는 게 좋음.

- 유전체 구조의 이해 : 형질의 유전적 구조를 알아내고 유전적 변이의 비율을 추정함. 서로 관련이 없는 개인들의 유전자형 세트에서 유전성을 추정하므로 이를 정량화해야 함. 넓은 의미의 유전성 / 좁은 의미의 유전성. SNP 기반 유전성은 유전자형 또는 귀속된 SNP의 추가 효과에 의해 설명되는 분산만을 측정. 희귀한 변이의 중요성. 단일 형질 뿐 아니라 여러 형질 사이의 유전적 관계 파악이 중요함. 유전적 상관관계 파악 및 방향 파악. 인과관계는 아님. 수직적 다원성, 수평적 다원성, 연결 불균형 유도 수평적 다원성, 다원성 유도 다원성. 서로 다른 표현 유형 간의 인과관계 평가 : Mendelian randomization. 가정 필요.
- GWAS를 통해 유의한 결과를 산출하기 위해서는 큰 샘플 필요. GWAS Catalog / GWAS Atlas : 수천 가지 특성에 대한 요약 통계에 엑세스. [요약 통계 : genomic build, SNP ID / location, allele, strand information, effect size / associated standard error, P value, test statistics, minor allele frequency, sample size.]
- 한계 : 편향된 연관성(코호트) / 하나의 특성에 정말 많은 변이가 영향을 미치는 경우(다원성) – 근본적인 생물학적 매커니즘 파악 어려움. 아직 큰 효과를 나타내는 희귀 변이는 찾아가는 중임. / 데이터 수집ㅇ에 대한 윤리적 문제 존재.