online © ML Comm
SPECIAL ARTICLE

J Korean
Neuropsychiatr Assoc
2011;50:20-38
ISSN 1015-4817
www.knpa.or.kr

정신과 질환에서의 광범위 유전체 연합연구

연세대학교 의과대학 정신과학교실, 의학행동과학연구소,[1]
성균관대학교 의과대학 삼성서울병원 정신과학교실[2]

김 세 주[1] · 홍 경 수[2]

Genome-Wide Association Study in Psychiatric Disorders

Se Joo Kim, MD[1] and Kyung Sue Hong, MD[2]

[1]Department of Psychiatry, Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Korea
[2]Department of Psychiatry, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

Received May 7, 2010
Revised October 8, 2010
Accepted December 6, 2010

Address for correspondence
Se Joo Kim, MD
Department of Psychiatry,
Institute of Behavioral Science
in Medicine, Yonsei University
College of Medicine, 250 Seongsan-ro,
Seodaemun-gu, Seoul 120-752, Korea
Tel +82-2-2228-1620
Fax +82-2-313-0891
E-mail kimsejoo@yuhs.ac

Most psychiatric disorders are some kinds of complex genetic traits. Identifying the causal genes of psychiatric disorders has been challenging. Through recent revolutionary advances, such as the HapMap Project and the development of high-throughput genotyping chips, the genome-wide association study (GWAS) has recently become possible and is now in the spotlight in psychiatric genetics. In this article, we reviewed the concepts, rationale, designs and general steps of GWAS, and also introduced a few previous GWAS of several psychiatric disorders.

J Korean Neuropsychiatr Assoc 2011;50:20-38

KEY WORDS Complex genetic traits · Genome wide association study · Psychiatric disorders.

서 론

광범위 유전체 연합연구의 배경 가설(Background hypothesis of genome-wide association study)

미국국립보건원(National Institute of Health)[1]에 의하면, '광범위 유전체 연합연구(genome wide association study, 이하 GWAS)는 관찰 가능한 특성(traits)과 관련된 유전적 연합성을 확인하기 위해 디자인된 인간 유전체 전체에 걸친 흔한 유전변이(common genetic variants) 연구'로 정의된다.

원인 유전자를 찾는 방법에는 가설에 근거한(hypothesis-driven) 후보 유전자 연합연구(candidate gene association study)와 가설이 없는(hypothesis-free) GWAS가 있다. 후보 유전자(candidate gene) 연구는 생물학적 가설, 또는 선행 연관(linkage) 연구를 통해 결정된 부위에 근거하여 원인이 되는 정확한 유전자를 찾는 방법이다. 이 방법을 통해서는 단지 유전적 위험 요소의 한 부분만을 확인할 수 있으며, 그것도 병태생리가 비교적 잘 알려진 경우에만 가능하다. 최근 수 십 년간 많은 후보 유전자 연구가 발표되었지만, 극히 일부만이 신뢰할 수 있는 결과(true positive)였고 대부분은 독립적인 연구에서 재현되지 않았다.[2] 따라서 후보 유전자 연구는 정신 질환과 같이 병태생리적 결함을 잘 모르는 질병의 유전적 기반을 충분히 밝히는 데에는 제한점이 있다.

GWAS는 생물학적 후보 유전자에 초점을 맞추는 대신 가설이 없이, 즉 어떤 특정 부위 또는 유전자에 대한 사전 예측 없이 전체 유전체를 조사(screening)한다. 이런 GWAS 접근은 병태생리에 대한 불완전한 이해로 인해 실제 원인 유전변이가 존재하는 곳이 아닌 다른 곳을 조사할 위험이 있는 후보 유전자 연구의 제한점을 극복할 수 있다.

GWAS는 유전자형분석(genotyping) 기술의 발달과 더불어 대규모 유전체 프로젝트의 성과로 가능하게 되었다. HapMap 프로젝트를 통해 인류의 단일염기유전자다형성(single nucleotide polymorphism, 이하 SNP)에 대한 광범위한 데이터베이스를 구성하였고, 이를 통해 알아낸 연관불균형(linkage disequilibrium, 이하 LD)에 대한 정보들은 비용효과적인 유전자형분석 플랫폼(platform) 개발을 가능케 하였다.

흔한 질병/흔한 변이 가설(Common Disease/Common Variant, CDCV Hypothesis)

현재 GWAS는 CDCV 가설을 전제로 하며, 질병에 기여하는 효과 크기(effect size)가 적은(low/modest) 흔한 유전

적 변이를 발견하는 데에 적합하다.[3] CDCV 가설에 의하면, 흔한 질병은 적은 영향력(small effect)을 가진 많은 흔한 변이들이 함께 작용해서 발생한다. 이러한 적은 효과를 가진 SNP는 많은 사람들을 대상으로 한 연구를 통해 발견이 가능하다. 현재 대부분 GWAS는 소수대립형질빈도(minor allele frequency, 이하 MAF)가 5% 이상 되는 SNP을 분석한다. 따라서 MAF가 5%에 미치지 못하는 드문 변이(rare variant)의 영향은 GWAS로 발견하기 어렵다.

GWAS에 이용되는 유전자형분석 플렛폼이 참조가 되는(reference) 전체 유전체(대개 HapMap 데이터베이스)를 얼마나 반영할 수 있는가는, 플렛폼에 포함된 1개 이상의 SNP와 $r^2 \geq 0.8$ 관계에 있는 흔한 SNP의 %로 나타낸다. 50~100만 개의 SNP로 구성된 플렛폼의 경우 유럽, 아시아 인구의 흔한 SNP 변이(common SNP variation)의 67~89%를 포착(capture)하고 아프리카 인구 변이의 46~66%를 포착한다.[4]

현재 시판되는 플렛폼은 대개 다음 세가지 방법으로 SNP를 선정하여 microarray를 구성한다. 첫째, tagging SNP를 이용하는 방법-이미 알려진 LD 블록들을 대표할 수 있는 SNP 세트를 선정한다(예, Illumina HapMap 300, HapMap 500). 둘째, 분산된 접근(dispersed approach) 방법-LD 패턴을 무시하고 전체 유전체의 물리적 커버범위(physical coverage)를 최대화하기 위해 거의 일정한 간격으로 SNP을 선택하여 세트를 구성한다(예, Affymetrix 111K, 500K). 셋째, 위의 두 가지 접근법을 적절히 조합하여 SNP 세트를 선택하는 방법 등이 있다.[5]

흔한 질병/다수의 드문 변이 가설(Common Disease/Multiple Rare Variants, CDMRV hypothesis)

흔한 SNP로 흔한 질병의 유전적 요인들을 모두 설명하기는 어렵다. 진화론적 모델에 따르면 복합 질병(complex disease)에는 흔한 SNP 뿐 아니라 일부 유전자에 포함된 다수의 드문 변이가 관여할 것으로 예상된다(CDMRV hypothesis).[6] CDMRV 가설은, '비교적 큰 상대적 위험도(대개 5~10)을 가진 적은 수의 대립형질들이 함께 관여하여 질병을 일으킨다.'고 가정한다. 드문 변이를 발견하기 위해서는 후보 유전자 부위를 수백 또는 수천 명의 표본을 대상으로 고효율 염기서열재분석(high-throughput resequencing)이 필요하다. 유전자형분석 기술이 빠른 속도로 발전함에 따라 앞으로는 드문 변이에 대한 연구가 활발해 질 것으로 기대되며, 특히 가까운 미래에 1000 게놈 프로젝트(1000 Genome Project, www.1000genomes.org)가 완료되어 1~5% 빈도를 지닌 드문 변이에 대한 데이터베이스가 완성되면, 흔한 변이에 더해 드문 변이까지 포함하는 GWAS가 가능해질 것이다.

본 론

표본의 선정

환자군의 선정

효과적인 GWAS 연구를 위해서는 질병의 소인이 되는 대립형질(disease predisposition allele)을 많이 포함하고 있는 대상을 선정하여 연구의 검정력(power)를 높이는 것이 유리하다. 이를 위한 전략으로는 표현형 이질성(phenotypic heterogeneity)을 최소화하는 방법, 조기 발병군과 같은 극단의 사례(extreme case)를 대상으로 하는 방법, 환자가 여러 명 있는 가계에서 대상을 선정하는 방법(가족력 사례) 등이 있으며,[7] 특히 표본의 크기를 충분히 늘리기 어려울 때 유용하다. 그러나 이런 전략이 언제나 검정력을 높이는 것은 아니다. 때때로 이와 같이 대상 선정의 범위를 좁히는 것이 오히려 검정력을 낮추기도 하므로 주의가 필요하다.[8]

대조군의 선정

적절한 대조군을 선정하는 데에 있어 인종(ethnicity)의 고려는 필수적이다. 그 외에도 나이, 성별, 그리고 이미 알려진 유전자-환경 상호관계가 있다면 이에 대한 정보도 고려하여야 한다. GWAS에서도 엄격하게 정의된 대조군의 선정은 중요하다. 그러나 자세한 선별검사를 사용하여 매우 많은 수의 대조군을 선정하는 것이 언제나 용이한 것은 아니다. 따라서 많은 경우에 간단한 선별검사만을 거친 공통대조군(common control)과 같은 대안적인 접근을 사용하기도 한다. 각각 2,000명으로 구성된 7개의 서로 다른 질병군과 3,000명의 UK 대조군을 비교한 Wellcome Trust Case Control Consortium(이하 WTCCC) 연구[9]는 공통대조군의 유용성을 보여주었다. 이 연구에서는 엄격한 선별검사를 통해 선정하는 질병-특이 대조군(disease specific control) 대신 매우 간단한 선별검사만을 거친 공통대조군을 사용하였다. 간단한 선별검사에 의한 대조군 선정은 오분류 비뚤림(misclassification bias)에 의해 검정력의 손실이 발생할 잠재적 위험이 있는 반면, 표본의 크기를 쉽게 증가시킬 수 있는 이점이 있다. 표현형(phenotype)이 드문 경우에는(예, 5% 이하의 유병율) 오분류 비뚤림으로 인한 검정력의 손실이 크지 않으며, 이로 인한 검정력의 손실은 표본의 크기를 증가시킴으로써 보완이 가능하다.[7] 예를 들어, 유병율 1%인 정신분열병을 대상으로 표본의 크기가 각각 수

천 명이 되는 사례-대조군 연구를 계획할 때에, Structured clinical interview for DSM-IV(이하 SCID-IV)와 같은 자세한 선별검사를 시행하지 않고 간단한 선별검사를 통해 대조군을 선정하는 대신 대조군 표본수를 보다 증가시키는 전략을 선택할 수 있다. 그러나 유병율이 높은 흔한 질병의 경우에는 대조군 선정시 오분류 삐뚤림으로 인한 검정력의 손실이 크기 때문에 자세한 선별검사를 통해 최대한 환자군을 제외해야 한다.[7]

연구에 포함시킬 수 있는 대조군의 표본수가 제한되어 있을 때, 검정력을 높일 수 있는 효과적인 전략 중 하나는 환자군과 정반대의 특성을 갖는 초정상(hypercontrol) 대조군을 선택하는 것이다. 초정상 대조군은 비만이나 고지혈증 같이 환자가 단봉분포(unimodal distribution)의 연속선에서 한 쪽 극단(extreme)에 위치한다고 가정할 때 반대쪽 극단에 있는 표본을 선정하는 것이다.[10] 그러나 초정상 대조군 선택시 의도하지 않은 문제가 발생할 수 있어 주의가 필요하다. 예를 들어, 비만 연구의 대조군으로 체중이 극단적으로 낮은 사람들을 선택할 경우, 체중 조절 자체보다는 만성 질환이나 식이장애 등과 주로 연관되어 있는 유전변이들을 조사하게 되는 예상치 못한 오류를 범할 수 있다.[7]

잠재적 인구 하위구조에 의한 1형 오류의 위험

대조군과 환자군 사이에 인구구조(population structure)가 다른 경우 제 1종 오류가 증가할 수 있다. 즉, 실제 비교하려는 표현형에 대한 유전적 차이를 반영하는 대신, 인종에 따른 유전자형 빈도 차이에 따라 오류가 발생할 수 있다. 기존 GWAS를 살펴보면, 인종적 배경을 잘 고려하여 대조군과 환자군을 선정하고, 인종적 유전 배경(population genetic background)이 큰 차이를 보이는 자료를 분석에서 제외시키는 경우, 잔여 인구하위구조(residual population substructure)에 의한 영향은 미미한 것으로 알려져 있다.[9] 또한 잠재적 인구하위구조가 존재할 것으로 예상되는 경우, 통계적 방법으로 인구하위구조에 대한 정보를 지닌 표식자(ancestry informative marker, 이하 AIM)들을 이용하여 잔여 층화(residual stratification)를 보정한다.[11,12] 통계적으로 잠재적인 인구층화를 보정하는 방법은 유럽이나 북미 등 인종적 다양성이 큰 집단에서 특히 유용하며, 검정력을 크게 약화시키지 않는 장점이 있다. AIM을 이용하여 통계적으로 인종적 혼합(ethnic admixture)을 나타내는 요인값을 계산하고, 이를 통해 잠재적 인구층화를 평균적으로 보정한다.

그러나 단지 평균적으로만 보정해 주기 때문에, SNP이 선조(ancestry)에 대한 강력한 정보를 지닌 표식자에 바로 인접한 경우 실제는 질병 취약성과 관련이 없음에도 불구하고 내재된 인구층화로 인해 거짓 연합성(spurious association)을 보일 수 있다. 따라서 통계적인 방법을 통한 인구층화의 보정이 잠재적 오류를 항상 완벽하게 제거해 주지는 못한다.[7]

GWAS 디자인

유전자료의 질 관리(Quality control)

GWAS로부터 정확한 결론을 도출하려면 유전자 자료의 질이 좋아야 한다. 100만개의 SNP 데이터 세트를 가정할 때, 유전자형분석 과정에서 체계적인 오류(systematically biased assay)가 0.5%만 발생한다 하더라도 5,000개의 SNP 정보가 틀리게 되며, 수용할 수 없을 정도로 많은 위양성 연합성(false positive association)이 발생할 수 있다. 일반적으로 이런 문제들을 예방하기 위해 microarray의 질을 담보할 수 있는 기준(threshold)을 정하고 이를 통과한 자료만을 사용한다. 예를 들면, 유전자형분석 실패율이 높은 SNP, MAF가 지나치게 낮은 SNP, 멘델법칙에 어긋나는 SNP(Mendelization error), Hardy-Weinberg 평형에서 벗어난 SNP들을 분석에서 제외하는 것이다.

그 제외된 SNP가 연구에서 조사하려는 특성과 관련이 있을 사전 확률(prior probability)은 매우 낮기 때문에 엄격하게 질 관리를 한다 할 지라도, 질병 취약성과 실제 유의한 연합성(true associations)이 있는 SNP들이 제거될 가능성은 거의 없다. 일반적으로 질 관리에는 다음과 같은 기준들이 사용된다.

유전자형분석률(Call rate)

유전자형분석이 이루어진 정도를 유전자형분석률이라 하는데 유전자형분석 탐침(genotyping probe)의 수행 정도를 나타내는 좋은 지표 중 하나이다. 문제가 되는 SNP을 확인하기 위해, 여러 표본에 걸쳐 유전자형분석 실패(missingness)의 전체적 분포가 어떻게 되는지 확인한다. 또한 사례-대조군 간의 유전자형분석 실패 정도를 χ^2 검정으로 비교하는 것이 좋다. 사례군과 대조군에서 유전자형분석 실패가 유의하게 차이가 나는 것을 무시하고 연합 분석을 진행하는 경우 위양성 결과를 낳을 수 있다. 또한 특정 개인에서 전반적으로 많은 SNP의 유전자형분석 실패가 나타나는 경우에는 그 개인의 DNA 질이 전반적으로 좋지 않다는 것을 의미한다. 지나치게 유전자형분석 실패율이 높은 유전자형과 전반적으로 실패율이 높은 대상(subject)은 분석에서 제외한다(예, 실패율 >5%).

재현성(Reproducibility)

동일한 표본을 중복 검사해서 동일한 결과가 재현되는지 확인한다.

멘델 유전 확인(Mendelian check)

가족 연구 디자인인 경우 멘델의 유전법칙에 어긋나는 것이 있는지 확인한다. 예를 들어, 부모가 AA, aa 유전자형을 가지고 있는 경우 자녀는 Aa 유전자형을 가진다. 만약 자녀가 다른 유전자형을 가진다면 멘델 오류(Mendelian error, 이하 ME)를 보이는 것이다. 멘델 오류는 가족관계에 대한 정보가 잘못되었거나(mis-identification), DNA 오염(contamination), 복사수변이(copy number variation), 유전자형분석 오류 등에 의해 발생한다. 일반적으로 무작위 오류(random error)가 어느 정도는 발생할 수 있기 때문에 재현성 오류와 멘델 오류의 경우 역치를 너무 엄격하게 적용하지 않는다(예, ME>2%인 가족 제외, ME>4%인 SNP 제외).[13,14]

잠재적 배치 효과(Potential batch effect)

모든 표본을 동일한 제품으로 동시에 검사할 수 없는 경우가 많기 때문에, 잠재적인 배치 효과가 발생할 수 있다. 배치 효과는 시기에 따라, DNA를 어디로부터(혈액, 타액, 구강점막) 추출하였는지, 어떤 방법으로 유전자형분석을 진행하였는지, 어떤 플렛폼을 사용하였는지에 따른 영향을 포함한다. 자료를 유전자형분석이 이루어진 시기별로 검토해서 시기별로 차이가 있는지, 예를 들면 분석기계를 교체한 전후로 결과가 차이가 나는지 등을 검토해 보면 배치 효과의 여부를 확인할 수 있다.

소수대립형질 빈도 역치(MAF threshold)

대부분의 연구들이 매우 낮은 빈도의 드문 변이(very rare variation)의 연합을 검정할 통계적 검정력을 가지고 있지 않기 때문에 표본내에서 일정한 역치 이상의 MAF를 갖는 SNP만을 통계 분석에 포함시키는 것이 바람직하다. 또한 현재 사용되는 유전자형분석 알고리듬은 MAF가 매우 낮은 SNP의 경우 유전자형분석이 틀릴 위험이 더 높기 때문에 지나치게 MAF가 낮은 SNP은 분석에서 제외하는 것이 좋다(예, MAF <1%).

Hardy-Weinberg 평형(HWE)

HWE를 통해서도 유전자형분석의 적절정을 평가할 수 있다. 그러나 HWE는 인구 층화에 의해서도 영향을 받을 수 있기 때문에 HWE를 제외한 다른 질 관리 기준들이 모두 만족되는 경우에는 무조건 그 SNP을 제외하지 말고, 보다 신중하게 결정하여야 한다. 따라서 HWE를 질 관리에 이용할 때에는 기준 역치를 비교적 엄격하게 적용하는 편이다(예, $p \leq 10^{-5}$). 대조군에서 각 SNP마다 HWE 여부를 확인하여 극단적으로 HWE에서 많이 벗어난 경우에는 대부분 안심하고 제거해도 된다. 그러나 워낙 많은 수의 검정을 반복하기 때문에 어느 정도의 HW 불균형은 예상할 수 있다. 따라서 HWE에서 약간만 벗어난 SNP의 경우에는 제거하지 않는다. 질 관리에 있어 HWE 검정은 완벽한 방법은 아니며, 유전자형분석 오류를 충분히 발견하기에는 검정력이 좋지 않다. 특히 환자군에서 약간의 불균형(disequilibrium)은 연합성을 시사하는 신호일 수 있기 때문에 지나치게 엄격한 기준을 적용하면 질병 취약성과 관련된 SNP을 놓칠 수 있다.

가족 구조(Family structure)

가족 자료, 예를 들어 부모자녀(trio)나 형제자매(sibship) 디자인 연구의 경우 부계관계(paternality) 오류 여부를 identity-by-state(이하 IBS)를 조사해서 간단하게 확인할 수 있다. 사례-대조군 자료의 경우 동일한 IBS 정보를 이용하여 identity-by-decent(이하 IBD)를 유추하기도 한다. IBD 패턴이 다른 경우 조상이 다른 표본들이 섞여 있음을 알 수 있고, IBD를 공유하는 비율이 높은 경우 근친 관계(cryptic relatedness)를 의심할 수 있다. 또한 지나치게 이형접합체(heterozygote)와 IBD가 많은 경우에는 표본이 오염되었을 가능성이 있다. 적어도 1차 친척(1st degree relativeness)과 2차 친척관계는 확인하여 각 쌍(pair) 중 한 명은 분석에서 제외시키는 것이 좋다.

일배체(Haplotype) 연합성 확인

만약에 어떤 SNP과 질병과의 연합이 시사되는 경우, 그 SNP 근처의 다른 SNP들이나 또는 그 SNP과 서로 상관성이 높은 변이들로 구성된 일배체도 연합성을 보이는 지 확인하는 것이 필요하다. 만약 어떤 SNP이 질병과 강한 연합성이 있는 것으로 분석되었지만, 그 SNP과 상관관계를 보이는 인접 SNP은 질병과 연합을 보이지 않는 경우 그 SNP의 연합성은 인위적인 오류(artifact)에 의한 것일 가능성이 높다.

연합 검정 통계치의 분포(Distribution of association test statistics) 확인

연합분석 통계치의 전체적인 분포를 검토하는 것이 데이터의 질을 확인하는 데에 필수적이다. 관찰된 검정값(test statistics)의 분포를 그래프로 나타낸 것을 분위수 대조도(qu-

antile-quantile plot, 이하 Q-Q plot)라 하는데 그림 1의 (a)와 같이 관찰된 검정값이 전반적으로 부풀려서(inflated) 전체적으로 연합성을 보이는 통계치가 너무 많은 경우, 즉 너무 많은 SNP들이 질병과 관련이 있는 것으로 나타날 때에는 데이터 삐뚤림 여부를 의심해 보아야 한다. 또한 (c)의 꼬리 부분과 같이 통계치가 너무 극단적으로 유의한 p값을 보이는 경우에도 배치효과, 비임의적 유전자형분석 실패(non-random missingness), 또는 자료처리 오류(data-handling error) 등에 의한 결과가 아닌지 확인해야 한다. 인구층화(population stratification)나 근친 관계가 존재할 때에도 이런 분포를 보일 수 있다. 그림 1의 (b)와 같이 관찰된 검정값의 전체적인 분포가 예측값(expected p-value)의 분포와 대부분 일치하고 분포의 꼬리 부분에서 일부 편차를 보이는 경우, 유의한 연합성이 실재할 가능성을 시사한다.

다단계 접근 및 메타/메가분석

대규모 표본을 대상으로 수 십 만개 이상의 SNP을 조사하는 데에는 많은 비용이 소요되기 때문에, 초기 선별(screening) 과정을 2단계 또는 3단계로 구분하여 순차적으로 시행하는 다단계 접근(multi-stage approach)이 흔히 이용된다.[15] 전형적인 2단계 접근 디자인은 다음과 같다. 1단계로 전체 표본의 일부분(예, 약 25%)에서만, 전체 GWAS 패널을 이용하여 선별검사를 한다. 1단계에서는 비교적 덜 엄격하게 통계적 유의성 역치를 설정하고, 이 기준을 충족한 적은 수의 SNP만을 대상으로 2단계 연구를 진행한다. 2단계에서는 전체 표본 중 1단계 표본을 제외한 나머지 표본에서 1단계를 통과한 SNP들을 대상으로 연합성을 조사한다. 2단계 분석에서 통계적 유의성을 해석할 때에는 1단계, 2단계를 통틀어 다중비교에 대해 보정해야 한다. 이런 과정을 3~4단계로 진행하기도 하며, 대개 나중 단계로 갈수록 엄격한 통계적 유의성 역치를 적용한다.

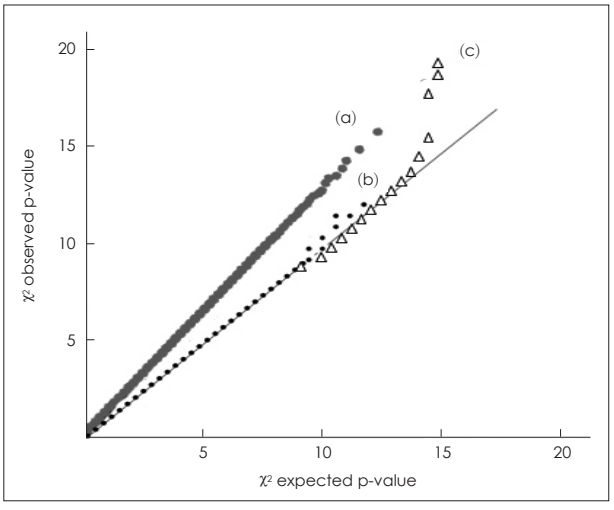

Fig. 1. Quantile-quantile plots. (a) Inflation of observed findings across the distribution. (b) Little evidence of overall inflation of observed findings except at tail of the distribution (most likely true association). (c) Excess inflation at the tail of the distribution.

한편, GWAS에서 다중검정 보정 이후에도 충분한 검정력을 지닐 만큼 충분한 크기의 표본을 모으기는 쉽지 않다. 따라서 몇 가지 대안적 디자인이 사용되는 데, 대표적인 것이 메타 분석(meta-analysis)과 메가 분석(mega-analysis)이다. 메타 분석은 여러 연구들의 효과 크기를 합치는 것이고[16] 메가 분석은 여러 연구에 포함된 각 대상(subject)의 유전형-표현형(genotype-phenotype) 자료를 모두 합쳐 다시 분석하는 것이다.[17] 메가 분석에서는 각 연구에서 사용된 microarray 패널 간의 차이를 최소화하기 위해, HapMap LD 구조에 기반해서 유전자형분석하지 않은(ungenotyped) 표식자의 대립형질을 추정하여 사용하는 대체방법(imputation method)이 널리 활용되고 있다. 또한 대체방법을 활용하면, 연합성 분석의 범위를 유전자형분석하지 않은 표식자까지 크게 확대할 수 있는 이점이 있다. 여전히 별도의 독립된 표본에서 재현연구를 시행하는 것이 표준(standard)이지만, 최근 GWAS에는 독립된 표본에서의 재현연구에 메가 분석, 메타 분석을 조합한 디자인이 많이 사용되고 있다.

통계분석

통계적 검정력(Statistical power)

GWAS에서는 수많은 SNP을 동시에 조사하기 때문에 매우 커다란 표본 크기(sample size)가 필요하다. 일반적으로 유전자형의 상대적 위험도(relative risk)가 1.5~2 정도인 경우에 적절한 검정력을 갖기 위해서는 각각 1,000~2,000명의 환자군과 대조군이 필요하다. 그러나 정신과 질환 같은 복합 질병의 경우 상대적 위험도는 대개 1.1~1.2 정도에 불과하다. 이런 경우 적절한 검정력을 얻기 위해서는 각각 8,000~20,000명 정도의 환자군과 대조군이 요구된다. Altshuler와 Daly[18]는 90%의 검정력과, $p<5\times10^{-8}$(100만개의 SNP을 조사한다고 가정)의 역치를 기준으로 이전에 보고된 GWAS 결과들을 재현하는 데에 필요한 표본 크기를 계산하였는데, 가장 강하게 연합된 SNP들의 경우에도 각각 최소 2,500명의 환자군과 대조군이 필요하였으며, 대부분의 경우에는 각각 10,000~20,000명의 표본이 필요하였다.

GWAS 검정력에 영향을 주는 요인들은 다음과 같다. 1) 유전자형 상대적 위험도(genotypic relative risk), 2) MAF 및

LD-MAF가 높을수록, 그리고 조사하지 않은(즉, microarray에 포함되지 않은) 위험 대립형질과 조사한 SNP 사이의 LD가 클수록 검정력이 높다. 3) 전달 방식(mode of transmission)-우성(dominant) 또는 승법적 유전효과(multiplicative genetic effect)인 경우 검정력이 높고, 열성 효과(recessive effect)인 경우 낮다. 4) 대조군의 선정(selection of comparison subjects)-유병율이 상대적으로 높은(>5%) 경우 선별검사를 통해 질병이 있는 사람을 대조군에서 제외하는 것이 검정력을 증가시킨다.[7] 5) 기술적인 오류(technical artifact)-어떤 종류든지 기술적인 오류는 검정력을 약화시킨다.

유전모델 및 다중비교의 보정

사례-대조군 GWAS에서 가장 기본이 되는 통계적 분석은 χ^2 검정이다. 각 SNP 마다 6개의 유전모델을 검정할 수 있다.

① 2×2 테이블을 만들어서 대립형질(allele)의 분포를 비교

② 상합모델(additive model), 3×2 테이블을 만들어서 대립형질(AA, Aa, aa) 빈도를 비교(자유도 2)

③ 3×2 테이블을 구성하여 AA>Aa>aa 경향성을 검정(자유도 1, Cochran-Armitage test for trend)

④ 우성모델(AA+Aa vs. aa)

⑤ 열성모델(AA vs. Aa+aa)

⑥ 초우성(overdominant)모델(Aa vs. AA+aa, 이형접합체 유리, heterozygous advantage)

이 중 경향성 검정이 가장 널리 사용된다. 만약 표본이 서로 다른 센터로부터 수집되었거나, 자료를 층화(strata)해서 분석해야 할 또다른 이유가 있을 때에는 2×2×k (k=number of strata) 테이블을 만들어 Cochran-Mantel-Haenszel χ^2 검정을 실시한다. 예를 들어, 선조집단(ancestral group)과 같이 범주형 혼란변수를 조정해 줄 필요가 있을 때 사용할 수 있다. χ^2 검정 이외에 회귀모델(regression model)을 이용하기도 하며 특히 양적인 특성을 지닌 혼란변수를 공변량으로 통제하는 데에 유리하다.

GWAS에서는 연합성 검정(association test)을 매우 여러 번 시행하게 된다. 최근 흔히 사용되는 50만개의 SNP이 포함된 microarray 패널을 사용할 경우, 명목적 유의수준인 $p<0.05$를 적용한다고 가정하면, 단순히 우연성(by chance)만으로 2,500개의 위양성 결과가 발생할 수 있다. 이와 같은 문제점을 보완하기 위해 GWAS에서는 엄격한 방법으로 다중검정을 보정한다. 일반적으로 사용되는 다중검정의 보정 방법으로는 Bonferroni 방법, 거짓발견율(False Discovery Rate, 이하 FDR) 방법, 그리고 순열(permutation) 검정 등이 있다. 널리 사용되는 Bonferroni 방법은 가장 보수적(conservative)인 방법으로 관습적(conventional) 유의수준 p값을 검정횟수로 나누는 방법이다. 즉, 1백만개의 SNP을 조사하는 GWAS의 경우 유의수준은 $p<0.05/10^6(5\times10^{-8})$이 된다. 그러나 고밀도 microarray를 사용하는 경우, 표식자들 사이에 상당한 수준의 LD가 존재하기 때문에 서로 독립적이라고 할 수 없다. 따라서 독립적인 검정을 전제로 하는 Bonferroni 방법을 적용하는 것은 너무 엄격하다는 견해도 있다.[19] FDR은 전체 양성 결과(total positive) 중 예측되는 위양성(false positive)의 비율을 의미하며, 유의성을 보인 검정 결과 중 잘못된 검정의 비율을 조절한다. FDR 방법은 Bonferroni 방법과 달리 검정횟수에 따른 유의수준 p값이 급격히 감소하지 않기 때문에 Bonferroni 방법보다 진양성(true positive)의 제거 비율이 낮고 덜 보수적이며(less conservative), 순열검정보다는 수행하는 연산의 횟수가 적기 때문에 계산적 부하가 덜하다(less computationally intensive). 순열검정은 주어진 p값이 유전형과 표현형이 아무런 연합성이 없다는 가정에서 얼마나 자주 우연에 의해 발생하는가를 평가한다. 즉, 순열을 통하여 관측자료로부터 모든 가능한 순열자료(대개 수천 이상)를 구한 다음 이 순열자료에 대해 검정통계값을 계산하고 실제의 관측자료에서 얻은 통계값와 비교 검정하여 보정된 p값을 계산하는 방법이다. 예를 들어, 실제 관측자료에서 얻은 p값이 0.001이고, 1,000번의 순열에서 $p\leq0.001$인 경우가 60번 발견되었다면 경험적 실험(empiric experiment)에 의해 보정된 p값은 60/1000(0.06)이 된다. 순열자료에서는 사례군-대조군을 임의대로 구분했기 때문에 생물학적으로 의미가 있는 연합성이 발견되어서는 안 된다. 따라서 각 순열자료에서 얻은 최선의 검정값(best statistics)은 우연에 의해 발생할 수 있는 가장 유의한 결과이며, 이들의 집합이 귀무분포(null distribution)를 나타낸다.

GWAS 통계분석 결과 제시

GWAS 통계분석의 결과는 흔히 각 SNP에서의 통계값을 그 SNP의 유전체 내 위치에 따라 산점도(scatter plot)로 표시하는 데 이를 맨하튼 플롯(manhattan plot)이라 한다(그림 2). 대개 표현형과 연합성을 보이는 경우 해당 좌(locus)에서 역치 이상 또는 역치 근처의 유의성을 보이는 SNP들이 무리를 지어 맨하튼 플롯에 나타난다. 만약 무리를 지어 나타나지 않고 단 1개의 SNP만이 유의성을 보이는 경우에는 오히려 오류(error)에 의한 위양성 가능성을 의심해 보아야 한다.

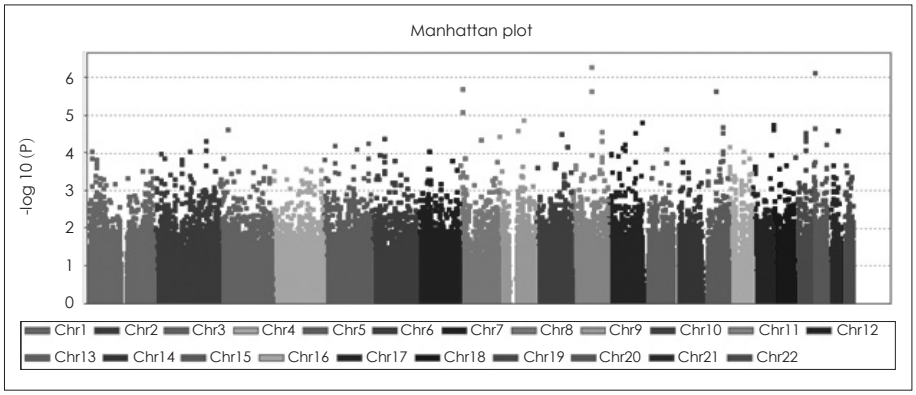

Fig. 2. Manhattan plot. each dot indicates a different variant examined by genome-wide association with chromosomes, the y-axis indicates a measure of the probability that a variant is associated with a phenotype.

GWAS 신호의 타당성 검증(Validation of GWAS signal)

재현연구(Replication study)

앞서 언급한 대로 복합형질 특성(complex traits)에 미치는 각 유전형질의 효과크기는 작을 것으로 추정된다. 게다가 GWAS는 매우 많은 수의 통계적 검정을 동시에 실시하고, 여러 가지 오류(error)나 삐뚤림(bias)에 매우 취약하다. 따라서 일차적으로 얻어진 GWAS 결과를 그대로 사실로 받아들이기는 곤란하며, 결과가 재현되는지 그 타당성을 확인하는 과정이 반드시 필요하다. 즉, 일차적인 GWAS의 결과가 오류나 편견에 의한 것이 아니고, 실제 존재하는 연합성을 반영하는지를 확인하는 것이다. 재현 연구를 위해서는 독립된 재현 표본(independent replication sample)을 사용해야 한다. 또한 기술적 오류를 발견하기 위해서 일차 GWAS와는 다른 유전자형분석 방법을 사용하는 것이 좋다.

최근 발표된 합의 기준(consensus criteria)[20]에 의하면, 결과가 재현된다는 것은, 원래 GWAS에서와 동일한 SNP의 대립형질이 동일한 또는 매우 유사한 인구집단의 동일한 표현형(phenotype)과, 비슷한 크기의 효과(similar magnitude of effect)로, 동일한 유전적 모델 및 동일한 방향성으로 유의하게 연합성을 보이는 것을 의미한다. 따라서 초기 재현연구는 원래 GWAS와 비슷한 인종을 대상으로, 비슷한 방법(ascertainment)으로 모은 표본을 이용하는 것이 바람직하다. 여기서 타당성이 검증되면, 다음 단계로 여러 인종을 대상으로 재현연구를 확대하여 일반화가 가능한 일관된 결과가 나타나는지 확인하는 것이 필요하다.

1차 GWAS의 결과가 재현연구에서 언제나 재현되는 것은 아니다. 1차 GWAS와 재현연구의 결과가 일치하지 않을 때에는 재현연구가 충분한 통계적 검정력을 가지지 못했을 가능성을 먼저 고려해 보아야 한다. 1차 연구의 효과크기가 실제보다 과장되는 측면(승자의 저주 효과, winner's curse effect)을 고려해 볼 때, 재현연구의 표본수는 1차 연구에 비해 충분히 커야 한다. 그러나 재현연구가 충분한 통계적 검정력을 가졌음에도 불구하고 1차 연구의 결과가 재현되지 않은 경우에는 다음과 같은 이유들을 고려해 볼 수 있다. 우선, 1차 연구의 결과가 틀렸을 가능성이 있다. 그러나 1차 연구의 결과가 틀리지 않았다고 가정할 때에는, 여러 가지 이질성(heterogeneity)에 의한 차이를 생각해 볼 수 있다.[7] 첫째, 유전자형을 조사한(genotyped) SNP과 조사하지 않은 원인 SNP(ungenotyped causual SNP) 사이의 LD 구조가 두 표본에서 서로 다른 경우—대개 GWAS에서 유의한 신호를 보이는 표식자는 원인 유전변이 자체라기 보다는 원인 유전자와 LD관계에 있는 SNP일 가능성이 높다. 따라서 재현연구와 1차 연구의 표본에서 표식자와 원인 유전자 사이의 LD 구조(연관 r^2)가 서로 다른 경우 1차 연구와 재현연구의 결과가 일치하지 않을 수 있다. 즉, 1차 연구 표본에서는 원인 변이를 효과적으로 포착하였던 표식자가 재현연구의 표본에서는 서로 다른 LD 구조로 인해 원인 변이를 제대로 포착하지 못할 수 있다. 둘째, 모집 방법의 차이 또는 인구이동(population drift) 등으로 인해 두 표본 집단에서 어떤 좌에 위치한 원인 대립형질의 분포와 효과크기가 다른 경우—인구이동 등의 현상으로 인해 본래의 커다란 인구집단(original population)에서 일부 사람들이 분리되어 새로운 인구집단을 형성하는 경우, 새로운 인구집단은 본래 인구집단보다 상대적으로 유전적 다양성(generic variation)이 적어지고 유전적으로 동질한 집단이 형성된다. 이 과정에서 본래 인구집단에서 흔하지 않았던 어떤 유전변이가 상대적으로 높은 빈도를 차지하게 되기도 하고, 반대로 아예 사라지기도 하는 등 본래의 인구집단과는 유전적 구성이 달라진다. 또한 본래 인구집단에서는 돌연변이(mutation)의 전달(transmission) 등에 의해 질병의 유전적 이질성(genetic heterogeneity)이 큰 반면, 새로운 인구집단에서는 질병의 이질성이 상대적으로 적은 편이다. 따라서 1차 연구와 재현연구의 표본을 각기 다른 인구집단에서 선택할 경우 동일한 SNP 일지라도 그 빈도와 질병 발생에 미치는 효과크기가 서로 다를 가능성이 있다. 이런 원인 SNP 빈도와 효과크기는 통계적

검정력을 결정 짓는 요인이 되며, 결과적으로 1차 연구의 소견이 재현연구에서는 확인되지 않는 결과를 초래할 수 있다. 셋째, 유전자-유전자 비상합적 상호작용(non-additive interaction)이 존재하는 경우 또는 유전자-환경 상호작용이 존재하는 경우, 서로 다른 인구집단에서 인구 특이(population specific) 유전자-유전자 상호작용이나 유전자-환경 상호작용이 존재하는 경우에는 인구집단에 따라 유전변이가 질병에 미치는 영향이 달라질 수 있기 때문에 1차 연구의 결과가 재현연구에서 동일하게 재현되지 않을 수 있다.

확인된 신호의 추적(Following-up confirmed signals)

일차 GWAS와 독립된 표본에서의 재현연구를 통해 확인된 SNP은 GWAS에 포함된 표식자인 SNP이 표현형과 단지 통계적인 분석에서 연합성을 보였다는 것을 의미한다. 즉, 그 표식자가 표현형의 원인이 되는 어떤 기능을 지닌 유전변이(causal variant)임을 뜻하는 것은 아니다. 물론 운이 좋게도 원인 변이가 microarray의 수 십 만개의 표식자에 포함되어 있고, 일차 GWAS와 재현연구를 통해 유의한 연합성을 보인다면 그 표식자가 원인 유전변이일 가능성도 있다. 그러나 이런 확률은 매우 드물다. 오히려 GWAS에서 유의하다고 확인된 표식자는 표현형의 원인 변이라기 보다는 그 주변에 위치하는 SNP일 가능성이 높다. 즉, GWAS의 신호는 표현형의 원인 변이와 함께 재조합 고빈도지점(recombination hot spot) 사이 간격(interval)에 존재하고 있을 가능성이 높으며 이 간격 내에는 흔히 몇 개의 유전자가 포함되어 있다. 그러나 GWAS 신호가 언제나 질병 취약성과 관련된 유전자(susceptible gene) 인근에서 LD를 이루고 있는 것은 아니다. 예를 들어, GWAS 신호가 취약성 유전자의 원거리 조절인자(remote regulatory element)와 LD를 이루는 경우, 정작 취약성 유전자의 엑손(exon)은 표식자와 멀리 떨어져 존재할 수도 있다. 다시 말해, 원인 변이(causal variants)는 GWAS에서 발견된 신호(signal) SNP와 강한 LD 관계에 있어야 하지만, 원인 변이에 영향을 받는 유전자는 반드시 원인 변이와 LD를 이루거나 원인 변이에 가까이에 있을 필요는 없다.

따라서 일차 GWAS와 재현연구를 통해 의미 있는 표식자 신호를 얻은 후에는 이를 기반으로 표현형의 실제 원인이 되는 유전변이를 찾는 과정이 필요하다. 이 과정은 대개 다음 단계로 이루어진다: 1) 일차 GWAS와 독립된 표본에서의 재현연구를 통해 확인된 결과가 다른 인종 표본에서도 일반화가 가능한지를 조사 2) 세밀한 유전지도작성(fine mapping) 및 염기서열 재분석(resequencing) 3) 서로 상관관계에 있는 여러 신호들 중 가장 정보성이 높은 표식자 구분, 서로 독립적으로 표현형과 연합성을 이루는 표식자 확인 4) 확인된 표식자의 생물학적 기능을 조사 5) 발견된 유전변이와 연합성을 보이는 표현형을 정확히 정의 6) 잠재적인 다면발현(pleiotropy)을 조사.

세밀한 유전지도작성과 염기서열재분석

세밀한 유전지도작성의 목적은 서로 높은 상관관계를 보이는 변이(SNP)들 중 실제 질병의 원인이 되는 유전변이와, 실제 원인 변이라기 보다는 이와 단지 LD 관계에 있어 유의한 연합성을 보이는 변이를 구별하는 것과, 기타 독립적인 추가 변이(independent & additional variants)들을 찾는 것이다.

표현형과 가장 유의한 연합을 보이는 변이를 발견하기 위해서는 이전 단계에서 발견한 좌에 위치하는 모든 흔한 변이의 목록을 작성하고 표현형과의 연합성 정도를 평가하는 과정이 필요한데, 이것이 세밀한 유전지도작성이다.[21] 이를 위해 대개 하나의 일배체블록(halpotype block)으로 구분되는 영역에서 완전한 상관관계를 보이거나 매우 높은 상관관계에 있는 SNP 세트 목록을 작성해야 한다. 그러나 문제는 모든 흔한 SNP에 대한 정보를 포함하는 데이터베이스가 아직 마련되어 있지 못한 점이다. 따라서 관심 부위의 완벽한 SNP 세트(complete SNP set)를 데이터베이스로부터 얻기가 용이하지 않다. HapMap 데이터베이스(2nd version)도 전체 유전체의 흔한 SNP 중 단지 30% 정도만을 포함하고 있을 뿐이다. 그나마 한국인에 대한 자료는 HapMap에 포함되어 있지도 않다. 따라서 연합성을 보이는 부위(a region of association)에서 서로 상관관계인 유전변이들의 전체 목록(entire list of correlated variants)을 얻기 위해서는 연구 표본과 동일한 인종적 배경을 지닌 상당 수의 참조 표본(reference sample)에서 해당 부위의 염기서열을 분석하는(sequencing) 별도 과정이 필요하다. 그러나 향후 1000 게놈 프로젝트가 완료되어 현재의 HapMap에 비해 훨씬 많은 수의 흔한 SNP에 대한 정보가 포함된 데이터베이스가 완성되면 참조 표본을 대상으로 별도의 염기서열을 조사하지 않고도 데이터베이스로부터 해당 부위의 자세한 SNP 목록을 얻을 수 있을 것으로 기대된다.

질병 발생에 가장 큰 영향을 미치는 잠재적 유전변이를 확인하는 세밀한 유전지도작성은 그 목적에 따라 범위를 달리할 수 있다. 만약 GWAS의 지표 신호(index signal) 아래 위치하는 흔한 원인 변이(common causative variants)들만을 발견하는 것이 목적이라면, 그 지표 신호 SNP을 중심으로 좌우로 가장 가까운 재조합 고빈도지점(recombination hot spot, 대개 수십 kb)까지의 부위를 목표 영역으

로 선정하고 수백 명의 표본을 대상으로 염기서열재분석을 시행한다. 한편 목적을 지표 신호뿐 만 아니라 이와는 독립적으로 그 부위에 존재하며 질병 취약성(regional susceptibility)에 관여하지만, 이에 대한 표식자가 microarray에 포함되어 있지 않았기에 초기 GWAS와 재현연구에서 발견되지 않았던, 모든 흔한 변인을 발견하는 것까지 확대할 수도 있다. 이 경우에는 유전자발현(gene expression) 또는 기능과 관련된 염기서열 모두를 고려하여야 하며, 지표 신호를 포함하여 각각 양쪽으로 수백 kb 정도의 범위를 인위적으로 선정하여 염기서열재분석을 실시한다. 범위를 더 확대하여 질병 취약성에 독립적인 영향을 지닌 드문 변이까지 발견하는 것을 목표로 한다면, ~500~1,000명의 표본을 대상으로 표적부위(target region, 주로 엑손 부위 또는 매우 보존적인 부위)의 모든 염기서열을 철저히 조사하는 고밀도 염기서열분석(deep sequencing)이 필요하다.[7]

기능적 연구

재현연구 및 세밀한 유전지도작성을 통해 질병과 강력한 연합성을 갖는 SNP를 발견하는 것이 매우 중요하나 질병의 원인과 치료법을 찾는 데에 있어 매우 기초적인 시작 단계에 불과하다. 지금까지 연구결과들을 살펴보면 초기 GWAS와 재현연구, 세밀한 유전지도작성을 통해 잠재적 원인 유전변이를 찾았다 하더라도, 이 유전변이가 유전자 발현이나 단백질 형성 등에 미치는 영향을 미리 알고 있는 경우는 매우 드물다. 따라서 GWAS를 통해 확인한 유전변이가 정말로 질병 또는 표현형의 원인이라는 것을 증명하기 위해서는, 이 SNP이 가지는 분자생물학적, 생리학적 기능을 확인하고, 표현형에 관여하는 기제(mechanism)를 밝혀야 한다.[22] 우선 연합성을 보인 SNP과 높은 LD 관계에 있는 SNP 중 이미 기능이 밝혀진 것이 있는 지 살펴보는 것도 도움이 된다. 유전변이의 기능에 대한 연구는 확인된 변이에 의해 분자생물학적 수준에서는 어떤 변화가 생기며, 이러한 변화가 질병과는 어떻게 관련이 있는지 생물학적 과정(biological process)을 조사하는 것이다. 만약 확인된 SNP이 정지 코돈(stop codon)이나 틀이동 코돈(frame shift codon)과 같이 단백질 형성에 영향을 미치는 단백질-부호화 유전자다형성(protein-coding polymorphism)인 경우, 질병의 원인 변이일 가능성이 높다. 비록 단백질 종류를 변화시키지 않을 지라도, 유전자발현에 영향을 미치는지 여부를 확인하는 것도 도움이 된다. 그러나, SNP이 표현형에 미치는 효과크기가 작을 경우에는 이런 방법을 통해 유전변이의 기능을 확인하는 것이 곤란할 수도 있다.

SNP의 기능 연구는 주로 컴퓨터를 이용한 모의 실험(in silico), *in vitro* 실험, 그리고 *in vivo* 실험을 통해 이루어진다. 컴퓨터를 이용한 모의 실험을 통해서는 SNP이 가진 생물학적 기능(예, 단백질의 3차원 구조의 변화 등)을 추정해 볼 수 있다. *In vitro* 실험으로는 서로 다른 유전적 배경(genetic background)을 지닌 세포들 내에서 유전자 발현을 평가하거나, 화학적 방법이나 RNAi를 이용하여 유전자를 발현억제(knock-down) 시켰을 때 나타나는 변화를 확인할 수 있다. *In vivo* 실험으로는 유전자 조작으로 knock-out 또는 knock-in 동물모델을 만들어서 유전변이가 생물학적 기능이나 개체의 표현형에 미치는 영향을 관찰하는 것이다.

그러나 기능적 연구가 일반적으로 도움이 된다 하더라도, 통계적으로 분명한 신호를 재차 확인하는 이전 단계의 재현연구를 대신할 수 없다. 왜냐하면 기능적 연구 결과와 역학 연구 결과가 일치하지 않는 경우도 많으며, 수행 가능한 기능적 연구에 여러 단계가 있는 것을 고려할 때 어떤 한 기능적 연구결과만을 단편적으로 해석하는 것은 오류를 초래할 수 있기 때문이다.[23]

표현형 조사 및 다면발현

정신과 영역에서 흔히 사용되는 여러 표현형은 각각 독립적인 측면을 반영하기도 하지만 때로는 상당한 정도의 상관성(correlation)을 보이기도 한다. 예를 들어, GWAS를 통해 어떤 표현형과 유전변이와의 연합성이 확인되었을 때, 실제는 이 표현형 자체가 아니라 밀접한 상관관계에 있는 다른 표현형과의 연합성을 반영하는 것일 수 있다.[23] 반대로 하나의 유전변이가 두 가지 이상의 표현형 모두와 각각 연합되어 있을 수도 있다. 이를 다면발현이라 한다. 일단 유전변이가 어떤 표현형과 연합되어 있음이 확인되면 그 표현형과 다양한 다른 표현형에 대해서도 연합성을 조사하여, 유전변이와 실제 관련이 있는 직접적인 표현형을 찾거나 다면발현을 확인할 필요가 있다.

대표적 정신 질환에서의 GWAS

앞에 기술한 대로 GWAS를 통해 원인 변이에 대한 강력한 증거를 발견하기 위해서는 1) 초기 표본(initial sample)에서 강한 연관성이 발견되어야 하고, 2) 하나 이상의 독립적인 표본에서 정확히 재현되어야 하며, 3) 누적 p값(cumulative p-value)은 5×10^{-8} 보다 작아야 한다.[20] 따라서 대부분의 GWAS는 이 원칙에 맞추어 진행된다. 정신과 질환을 포함한 대부분의 복합 질환에 대한 GWAS 결과를 살펴보면 다음과 같은 특징이 있다. 첫째, Apolipoprotein E 유전자(APOE)와의 연합성이 발견된 알츠하이머 등 몇몇의 질환을 제외한 대부분의 질환에서 유의한 연합성을 보인

부위는 이전의 후보 유전자 연구에서는 확인되지 않았다. 둘째, GWAS에서 확인된 대부분의 원인 변이는 유전자의 부호화 부위(coding region)에 있지 않았고, 단지 일부만이 아미노산 서열을 변화시키는 변이(non-synonymous variants)였다.[24] 현재까지 정신분열병, 양극성 장애, 자폐증 등의 정신 질환에 대해서 몇몇 GWAS가 진행되었으며(표 1), 대표적 정신 질환에서의 GWAS를 소개하면 다음과 같다.

정신분열병

다음 기대와는 달리 지금까지 GWAS를 통해 밝혀진 정신분열병과 관련된 흔한 변이는 많지 않다. O'Donovan 등[25]은 479명의 정신분열병 환자들과 2,937명의 대조군을 대상으로 Affimetrix GeneChip 500K microarray를 이용하여 GWAS를 시행하였다. 초기 GWAS에서 광범위 유전체 유의성(genome-wide significance)을 보인 신호는 발견하지 못하였으나, 12개의 좌에서 중간 정도의 연합성을 확인하였다($p<10^{-5}$). 이 12개의 좌에 존재하는 25개의 SNP을 대상으로 한 추적연구에서 각 좌 마다 가장 강력한 연합성을 보인 SNP을 선정(1개의 좌는 p값이 두번째로 낮은 SNP을 선정)하였으며, 2개의 독립적인 표본을 통해 재현연구를 시행하였다. 이들은 WTCCC 기준에 따라 $p<1\times10^{-5}$와 $p<5\times10^{-7}$을 각각 중등도(moderate) 및 강력한(strong) 연합성의 증거로 정의하였다. 1차 재현연구(1,664명의 환자군과 3,541명의 대조군)에서 총 12개 SNP중 6개 SNP에서 상관성이 재현되었다. 재현된 6개의 SNP에 대한 유전자형 분석을 두번째 표본(환자군 4,143명, 대조군 6,515명)에서 시행하였고, 1차 재현연구 표본과 합쳐 총 6,666명의 환자군과 9,897명의 대조군을 대상으로 상관성을 재분석하였다. 그 결과 ZNF804A(zinc finger protein 804A) 내의 2p32.1에 위치한 변이(rs1344706)와 11p14.1, 16p13.12에 위치한 유전자 사이 부위(intergenic region) SNP들(rs1602565, rs7192086)이 정신분열병과 관련이 있을 것으로 강력히 시사되었지만($p=5.1\times10^{-4}$-9.3×10^{-5}), 광범위 유전체 유의성에는 미치지 못하였다. 그러나 모든 표본을 합친(초기 GWAS+재현1+재현2 표본, 환자군 7,308명 vs. 대조군 12,834명) 분석에서는, ZNF804A rs1344706 SNP과 정신분열병과의 연합성이 광범위 유전체 유의성에 도달하였다($p=1.6\times10^{-7}$). 또한, 연구자들은 정신분열병과 양극성장애가 유전적 취약성을 공유할 가능성이 높다는 전제하에 2차 분석을 실시하였다. UK 정신분열병 환자 642명과 UK 양극성장애 환자 1,865명을 합쳐 커다란 하나의 UK 정신증(psychosis) 집단을 구성하였고, 12,834명의 대조군과 비교하여, 앞선 6개 SNP에 대한 연합성을 조사하였다. 그 결과, 다른 좌에서는 공통의 연합성을 발견하지 못하였지만, ZNF804A과의 연합성은 정신분열병 단독 표본에서보다 훨씬 강한 유의성을 보였다($p=1.0\times10^{-9}$). 이는 ZNF804A 근처에 존재하는 유전변이들은 넓은 개념의 정신증 표현형(broader psychosis phenotype) 취약성과 관련이 있음을 시사한다. 비록 ZNF804A가 부호화(encoding)하는 단백질의 특징과 기능은 아직 알려져 있지 않지만, ZNF804A가 아연이온(zinc ion)과 DNA 결합 영역(binding domain)을 포함하고 있기 때문에, 유전자발현의 조절자로서 역할을 할 것으로 추측된다.

최근 Riley 등[26]은 1,021명의 환자와 626명의 대조군을 대상(Irish 표본)으로 ZNF804A 내의 rs1344706을 비롯한 12개의 tagging SNP을 이용하여 후보유전자 사례-대조군 연구를 실시하였다(1개 SNP은 질 관리를 통과하지 못하여 최종적으로 11개 SNP이 분석에 포함됨). rs1344706을 비롯한 4개의 SNP이 $p<0.05$ 수준에서 연합성을 보였으며, 특히 rs7597593은 Bonferroni 보정 후에도 유의성이 유지되었다. 또한 사후 후외측전전두엽(postmortem dorsolateral prefrontal cortex) 조직을 이용한 RNA 발현 연구를 시행 결과, ZNF804A rs1344706 A 대립형질을 가진 경우 유의하게 RNA 발현이 높았다. 생물정보학(bioinformatics)을 통해 살펴보았을 때, ZNF804A rs1334706을 포함하는 주변의 염기서열은 포유류의 보존적 부위에 위치하였다. A 대립형질을 포함하는 53 bp의 염기서열은 뇌에서 발현되는 전사인자(transcription factor)인 myelin transcription factor L zinc-finger protein(이하 Myt1 LZFP)와 POU3F1/Oct-6(POU domain, class 3, transcription factor 1)로 예측되는데, 이들 인자는 중추신경계의 발달과정, 특히 희돌기세포(oligodendrocyte)의 발달에 관여하는 것으로 알려져 있다.

그 밖에도 비록 엄격한 의미의 광범위 유전체 유의성을 만족시키지는 못하였지만 지금까지 GWAS를 통해 정신분열병과의 연합성이 시사된 유전자는 다음과 같다: colony stimulating factor 2 receptor, alpha, low-affinity(이하 CSF2RA), interleukin 3 receptor, alpha low affinity (이하 IL3RA), reelin(이하 RELN), coiled-coil domain containing 60(이하 CCDC60), retinol binding protein 1 (이하 RBP1), cortexin 3(이하 CTXN3), solute carrier family 12(이하 SLC12A2), glypican 1(이하 GPC1), roundabout, axon guidance receptor, homolog 2(이하 ROBO2), ROBO1-ROBO2, TRAF2 and NCK interacting kinase (이하 TNIK), TNF receptor-associated factor 3(이하 TRAF3).[27-31]

Table 1. Summary of several genome wide association studies of psychiatric disorders

	Initial sample (cases/controls)	Other sample (cases/controls)	Findings
Schizophrenia			
Athanasiu et al.[40]	201/305	2663/13780	Combined sample, $p=2.1\times10^{-6}$ at PLAA, 3.2×10^{-6} at ACSM1, $p=7.7\times10^{-6}$ at ANK3
Sullivan et al.[29]	738/733	None (stage 1 study)	No genome wide significant signals Several clusters containing ≥15 SNPs (all $p<0.05$) (FMO3, FMO6, PDIA6, etc)
Kirov et al.[27]	574 patients, 1,148 parents (trio sample) & 605 control	None	DNA pooling method used 63 top ranked SNPs from DNA pooling method were genotyped individually : $p=1.2\times10^{-6}$ at rs11064768 within CCDC60, p=0.00016 at rs893703, within RBP1 (the 3rd best SNP, candidate gene for schizophrenia)
Need et al.[28]	871/863	Replication 1 : 298/713 (Munch) Replication 2 : 394/ 524 (Italy) Replication 3 : 589 /11491 (Iceland) Replication 3 : 179 /267 (European-American)	No genome wide significant signals : The most strong signal, $p=1.34\times10^{-6}$ at rs2135551 in the ADAMTSL3 gene Combined initial sample+replication 1, $p=1.35\times10^{-7}$ at rs2135551 No significant associations in replication cohorts
Potkin et al.[30]	64/74	None	$p<10^{-6}$ at rs7610746 & rs9836484 in ROBO1-2, rs2088885 & rs7627954 in TNIK, rs245178 & rs245201 in CTXN3-SLC12A2
Stefansson et al.[41]	2663/13498 (SGENE)	4999/15555 Meta analysis : with ISC & MGS	No genome wide significant signals in initial scan → association test Replication sample : $p=4.9\times10^{-7}$at rs6932590 in MHC/PRSS16 Initial+replication sample : $p=4.4\times10^{-9}$ at rs6932590 in MHC/PRSS16 Initial+replication+ISC+MGS (total 12945 case/ 34591 control) : $p=1.4\times10^{-12}$ at rs6932590 in MHC/ PRSS16 (+other four SNPs $p<10^{-8}$), $p=2.4\times10^{-9}$ at rs12807809 in NRGN, & $p=2.4\times10^{-9}$ at rs9960767 in TCF4
Shi et al.[42]	Sample 1 : 2681/2653 (EA) Sample 2 : 1286/973 (AA) (MGS)	Meta analysis : with ISG (3322/3587) & SGENE (2005/12837)	No genome wide significant signals in initial & further scans : Initial sample $p=4.59\times10^{-7}$ at s13025591 in CENTG2 Replication sample $p=2.14\times10^{-6}$ at rs1851196 in ERBB4 in replication sample EA : MGS+ISG+SGENE, $p=9.54\times10^{-9}$ at rs13194053 in 6p22.1(+other 7 SNPs $p<5\times10^{-8}$)
Purcell et al.[43]	3222/3587 (ISG)	Mega analysis : with MGS & SGENE	$p=3.4\times10^{-7}$ at rs5761163 in MYO18B Mega analysis (total 8008/1907), $p=9.5\times10^{-9}$ at rs13194053 in MHC
O'Donovan et al.[25]	479/2937	Replication 1 : 1664/3541 Replication 2 : 4143/6515 Additional analyses : 2507 UK psychosis (642 schizophrenia+1865 bipolar disorder)/12834 controls	In initial sample+replication sample, $p=1.6\times10^{-7}$ at ZNF804A In UK psychosis sample, $p=1.0\times10^{-9}$ at ZNF804A

Table 1. Continued

	Initial sample (cases/controls)	Other sample (cases/controls)	Findings
Shifman et al.[31]	Male 419/1807 Female 214/967	Male 1506/2207 Female 768/2194	DNA pooling method used : 194 SNPs from DNA pooling method were genotyped individually in 2644 controls and 745 patients from the Ashkenazi Jewish population : female-specific association with rs7341475, in RELN gene ($p=2.9\times10^{-5}$) All replication sample : female, 2.13×10^{-3} at rs7341475 All initial+replication sample : $p= 8.8\times10^{-7}$ at rs7341475
Lencz et al.[44]	178/144		$p=3.7\times10^{-7}$ at rs4129148 in CSF2RA
Bipolar disorder			
Lee et al.[45]	1000/1000	409/1000	No genome-wide significance in initial single SNP analyses : $p=3.8\times10^{-6}$ at rs8040009 near ST8SIA2/C15orf32 Combined sample(1409/2000) : $p=4.9\times10^{-7}$ at rs2709736 near SP8, $p=6.1\times10^{-6}$ at rs804009, $p=9.7\times10^{-6}$ at rs2073831 in KCTD12, $p=5.2\times10^{-5}$ at 11013860 in CACNB2
Djurovic et al.[46]	194/336	435/10258	No genome-wide significance in initial single SNP analyses Replication : 5.4×10^{-7} at rs4377455 in RBMS3
Smith et al.[33]	Sample 1 : EA 1001/1033 Sample 2 : AA 345/670	Replication 1 : 1749 EA subjects from 250 multiplex bipolar families Replication 2 : EA 1263/431	No genome-wide significance in initial single SNP analyses : EA-$p=1.6\times10^{-6}$ at Xq27.1, AA-$p=1.5\times10^{-6}$ at DPY19L3, $p=4.5\times10^{-6}$ at NTRK2 No significance below the multiple testing threshold ($p<0.0001$) : possible evidences of association of 3 SNPs at C15Orf53
SKlar et al.[32]	1461/2008 (STEP-UCL)	Replication 1 : 409 trio sample Replication 2 : 365/ 361	No genome-wide significance in initial single SNP analyses. Haplotype analyses : $p=2.0\times10^{-8}$ at MYO5B). Not replicated in replication studies.
Baum et al.[34]	461/562	772/876	DNA pooling method used : 1877 SNPs from DNA pooling method were genotyped individually in replication sample Combined both sample : $p=1.5\times10^{-8}$ at rs1012053 in DGKH
Ferreira et al.[36]	1098/1267 (ED-DUB-STEP2)	Mega analysis with WTCCC+STEP-UCL	No genome-wide significance in initial single SNP analyses ED-DUB-STEP2+WTCCC+STEP-UCL (4387/6209) : $p=9.1\times10^{-9}$ at rs10994336 in ANK, $p=7.0\times10^{-8}$ at rs1006737 in CACNA1C, $p=3.5\times10^{-7}$ at rs12899449 near C15orf53 & in the same LD block with RASGRP1
WTCCC[9]	1868/2938	Expanded reference group (58C and UKBS control+ the other six disease set)	Initial sample $p=6.3\times10^{-8}$ at rs420259 near NDUFAB1, DISC1, DCTN5, but significance disappeared using extended reference group
Autism			
Wang et al.[37]	Sample 1 : EA 780 family sample (n=3,101) Sample 2 : 1204/6491	Replication 1 : 447 family sample (n=1,390) Replication 2 : 108/540	No genome wide significance in each sample. But in combined initial samples, $p=3.4\times10^{-8}$ between CDH10 & CDH9 at 5p14.1 (rs4307059). Replication 1 : all 6 SNPs at 5p14.1, $p=0.01\text{-}2.8\times10^{-5}$ Replication 2 : 4 SNPs at 5p14.1, $p<0.05$. All combined samples, $p=2\times10^{-10}$ at 5p14.1

Table 1. Continued

	Initial sample (cases/controls)	Other sample (cases/controls)	Findings
Ma et al.[13]	447 family sample (n=1,390)	487 family sample	No genome wide significance in either initial or replication sample : $p=3.4\times10^{-6}$ at 5p14.1 (rs10038113) in replication sample.
Weiss et al.[38]	1031 families (856 with 2 parents)+ 1,553 affected offspring	Replication 1 : Montreal 318 trio Replication 2 : AGP+Finnish family & Iranian trio (1755 trios)	No genome wide significance initial sample : $p=9.6\times10^{-6}$ at rs10513025 between SEMA5A and TAS2R1 Replication : $p=1.7\times10^{-7}$ at rs10513025 combined both sample : 2.1×10^{-7} (genome-wide significance by permutation) at rs10513025
Arkin et al.[39]	72 multiplex families (148 affected offsprings & 292 individuals)	1295 parent-child trio	No genome wide significance in initial sample : $p<2.14\times10^{-5}$ at rs7794745 in CNTNAP2 on 7q35 Two loci LOD >2 at 7q35 by genome-wide linkage study : Replication : rs7794745, p<0.005

PLAA : Phospholipase A2-activating protein, ACSM : Acyl-CoA synthetase medium-chain family member 1, ANK3 : Ankyrin 3, FMO3/6 : Flavin-containing monooxygenase 3/6, PDIA2 : Protein disulfide isomerase family A member 2, CCDC60 : Coiled-coil domain containing 60, RBP1 : Retinoblastoma-binding protein 1, ADAMTSL3 : A disintegrin-like and metallopeptidase with thrombospondin type 1 motif-like 3, ROBO1-2 : Roundabout homolog 1-2, TNIK : TRAF2 and NCK interacting kinase, CTXN3 : Cortexin 3, SLC12A2 : solute carrier family 12 (sodium/potassium/chloride transporters), member 2, MHC : Major histocompatibility complex, PRSS16 : Protease, serine, 16, NRGN : Neurogranin (protein kinase C substrate, RC3), TCF4 : Transcription factor 4, GENTG2 : AGAP1, ArfGAP with GTPase domain, ankyrin repeat and PH domain 1, ERBB4 : v-erb-a erythroblastic leukemia viral oncogene homolog 4, MYO18B : Myosin XVIIIB, ZNF804A : Zinc finger protein 804A, RELN : Reelin, CSF2RA : Colony stimulating factor 2 receptor, alpha, low-affinity, ST8SIA2 : ST8 alpha-N-acetyl-neuraminide alpha-2, 8-sialyltransferase 2, C15orf32 : Chromosome 15 open reading frame 32, KCTD12 : Potassium channel tetramerisation domain containing 12, CACNB2 : Calcium channel, voltage-dependent, beta 2 subunit, RBM3 : RNA binding motif, single stranded interacting protein 3, DPY19L3 : Dpy-19-like 3, NTRK2 : Neurotrophic tyrosine kinase, receptor, type 2, DGKH : Diacylglycerol kinase, eta, RASGRP1 : RAS guanyl releasing protein 1, NDUFAB1 : NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1, DISC1 : Disrupted in schizophrenia 1, DCTN5 : Dynactin 5, CDH9/10 : Cadherin 9/10, type 2, SEMA5A : Sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A, TAS2R1 : Taste receptor, type 2, member 1, CNTNAP2 : Contactin associated protein-like 2, EA : European ancestry, AA : African Ancestry, ISC : International schizophrenia consortium, MSG : Molecular genetics of schizophrenia consortium SGENE, SGENE consortium, STEP (-BD) : Systematic treatment enhancement program (for bipolar disorder), UCL : University college London, ED-DUB : University of Edinburg-trinity college Dublin, 58C : 1958 British birth cohort, UKBS : UK blood service, AGP : Autism genome project

양극성장애

양극성장애에서 최초의 GWAS인 WTCCC 연구[9]는 1,868명의 환자군과 2,938명의 대조군을 대상으로 하였으며 Affymetrix 500K를 사용하였다. 비록 대조군을 확대한 추가분석(extended reference group analysis)에서는 광범위 유전체 유의성에 이르지 못했으나, 1차 분석에서는 16p12에 위치한 rs420259 SNP이 강한 연합성을 보였다($p=6.3\times10^{-8}$). 이 부위에는 핵 주요 구조(nuclear key structures)들의 안정화에 관여하는 PALB2(partner and localizer of BRCA2(breast cancer type 2 susceptibility protein)와 미토콘드리아 호흡고리(respiratory chain) 복합체 1의 하위단위(subunit)를 부호화(coding)하는 NADH dehydrogenase(ubiquinone) 1(이하 NDUFAB1), alpha/beta subcomplex 1, disrupted in schizophrenia 1(이하 DISC1)과 상호관계가 있는 것으로 알려진 세포내 수송에 관여하는 단백질을 부호화하는 dynactin 5(이하 DCTN5) 등 양극성장애의 병태생리와 관련이 있을 만한 유전자들이 포함되어 있다.

Sklar 등[32]은 1,461명의 양극성장애 환자와 2,008명의 대조군을 대상으로 Affiymetrix 500K를 이용하여 GWAS를 시행하였고, 두 개의 독립적인 표본(409 trio 표본과 양극성장애 365명, 대조군 361명으로 이루어진 사례-대조군 표본)에서 재현연구를 시행하였다. 비록 단일 SNP 분석에서는 엄격한 기준의 광범위 유전체 유의성을 만족시키는 신호를 발견하지 못했지만 myosin 5B(이하 MYO5B) 유전자 내에 존재하는 일배체와 양극성장애와의 연합성이 광범위 유전체 유의성을 충족하였다($p=2.0\times10^{-8}$). 그 외에 tetraspanin-8(이하 TSPAN8, $p=7.6\times10^{-7}$), epidermal growth factor receptor(이하 EGFR, $p=8.4\times10^{-8}$) 유전자가 강한 연합성을 보였다. 그러나 305개의 SNP을 이용한 재현연구에서는 부모-자녀 표본(trio sample), 사례-대조군 표본 모두에서 다중비교에 대해 보정했을 때 의미 있는 연합성이 발견되지 않았다.

최근 Smith 등[33]은 유럽선조(European ancestry)의 양극성장애 1,001명, 대조군 1,033명과 아프리카선조(African ancestry) 양극성장애 345명, 대조군 670명을 대상으로 하는 2개의 GWAS를 시행하였다. 두 표본 모두에서 광범위 유

전체 유의성($p<5\times10^{-8}$)를 만족시키는 유전변이를 발견하지 못하였지만, 유럽선조(European ancestry) 표본에서는 Xq27.1 유전자 사이 부위(intergenic region)의 rs 5907577 ($p=1.6\times10^{-6}$)이, 아프리카선조(African ancestry) 표본에서는 19q13.11에 위치한 DYP19L(protein dpy-19 homolog 3) 유전자의 rs2769605($p=1.5\times10^{-6}$)와 9q21.33 위치의 neurotrophic tyrosine kinase, receptor, type 2(이하 NTRK2) 유전자의 rs2769605($p=4.5\times10^{-6}$) SNP이 가장 강한 상관성을 시사하였다. 이 GWAS 결과로부터 85개의 SNP 세트를 선정하여 2개의 독립적인 유럽선조 표본(가족 표본 1+사례-대조군 표본 1)에서 재현 연구를 시행했을 때 다중검정역치 이하(multiple testing threshold $p<0.001$)의 유의성은 발견되지 않았다. 그러나 chromosome 15 open reading frame 53(이하 C15Orf53) 부위에 위치한 3개의 SNP이 가족 표본 ($p=0.008-0.015$)과 사례-대조군 표본($p=0.03-0.04$)에서 어느 정도의 연합성을 시사하였다. 한편 연구자들은 이전 GWAS에서 양극성장애와 관련이 있다고 보고된 3개의 부위-ankyrin G(이하 ANK3), alpha 1C subunit of the L-type voltage-gated calcium channel(이하 CACNA1C), 15q14에 존재하는 C15Orf53로부터 3.3 kb 떨어진 부위-를 대상으로 후보 유전자연구를 시행하였는데, ANK3 내의 SNP 1개(rs1938526)가 양극성장애와 연합성을 보였다.

현재까지 양극성장애에서 위에 언급한 연구들을 비롯한 몇몇의 GWAS가 발표되었으며, PALB2, NDUFAB1, DCTN5, potassium voltage-gated channel subfamily C member 2(이하 KCNC2), diacylglycerol kinase eta(이하 DGKH), solute carrier family 39, zinc transporter, member 3 (이하 SLC39A3), junctional adhesion molecule 3(이하 JAM3), MYO5B, TSPAN8, EGFR, CACNA1C, ANK3 등이 양극성장애의 취약성과 관련된 유전자로 제시되었다.[9,32,34-36] 그러나 이들 SNP 중 DGKH 내의 1개 SNP(rs1012053)[34]과 ANK3 내의 2개 SNP(rs10994336, rs1938526)[36]만이 광범위 유전체 유의성($p<5\times10^{-8}$)에 도달하였다.

최근 다른 흔한 질병에서 몇몇 성공적인 GWAS 결과를 보인 것과 달리, 양극성장애를 비롯한 정신 질환에서의 GWAS들은 서로 일치된 결과를 보이지 못하고 있다. 정신질환에서 GWAS 결과가 표면적으로 재현되지 않은 이유는 적은(modest) 효과크기를 발견하기에는 불충분한 연구 검정력, 인구-특이 질병 대립형질(population specific disease allele), 표현형의 다양성/분류오류(misspecification), 적은 효과를 가지는 많은 유전자들 사이의 상위상호작용(epistatic interaction), 복사수변이나 흔한 SNP 패널에 의해서는 잘 포착되지 않는 다양한 드문 질병 대립형질(rare disease allele)처럼, 흔한 SNP과는 다른 종류의 유전변이에 의한 효과[32] 등을 고려해 볼 수 있다.

자폐증

Wang 등[37]은 모두 유럽 선조인(European ancestry)인 780개의 가족군(총 3,101명)으로 이루어진 Autism Genetic Resource Exchange(이하 AGRE) 코호트와 1,204명의 환자와 6,491명의 대조군으로 이루어진 Autism Case-Control(이하 ACC) 코호트를 대상으로 Illumina HumanHap550 BeadChip을 사용하여 GWAS를 시행하였다. 서로 독립적인 AGRE 코호트, ACC 코호트 각각에서는 광범위 유전체 유의성을 가진 SNP을 발견하지 못하였으나, 두 코호트를 합친 분석에서는, 5p14.1 위치의 rs4307059 SNP에서 광범위 유전체 유의성에 도달하는 연합성을 발견하였다($p=3.4\times10^{-8}$). 5p14.1에 위치한 6개의 SNP(rs4307059, rs7704909, rs12518194, rs4327572, rs1896731, rs10038113)을 사용하여 447개의 가족, 총 1,390명으로 이루어진 독립된 Collaborative Autism Project(이하 CAP) 코호트에서 1차 재현연구를 시행하였는데, 6개의 SNP 모두 $p=0.01\sim2.8\times10^{-5}$ 수준의 연합성을 보였으며, 대립형질이 가진 효과의 방향성(direction)도 동일하였다. 또한 다른 독립적인 Center for Autism Research and Treatmen(이하 CART) 코호트(환자군 108명, 대조군 540명)를 대상으로 한 2차 재현연구에서도, 5p14.1에 위치한 6개의 SNP 중 4개에서 $p<0.05$ 수준의 연합성을 보였다. 최종적으로 이들 총 4개의 독립적인 코호트를 모두 합친 분석에서는, 5p14.1에 위치한 6개의 SNP 모두 $p=7.4\times10^{-8}\sim2.1\times10^{-10}$ 수준의 연합성을 보였다. 이들 SNP은 CHD10(cadeherin 10)과 CHD9 사이의 유전자 사이 부위(intergenic region)에 위치하였으며, ~100 kb 크기의 LD 블록 내에 포함되었다. 이 LD 블록 내에는 몇 개의 매우 보존적인 요소들이(highly conservative element) 포함되어 있었다. 대개 크고 안정적인 유전자 사막(gene desert)들이 유전자 조절인자를 포함하고 있음을 고려해 볼 때, 5p14.1 위치의 tagging SNP들도 CHD10 또는 CHD9의 발현과 작용(action)을 조절하는 기능에 관계된 유전변이의 연합성을 포착할 것으로 추측되었다.

Ma 등[13]도 CAP 코호트를 대상으로 Illumina Beadchip을 이용한 초기 스캔(scan)과 AGRE 코호트를 대상으로 한 추적 조사에서 비록 광범위 유전체 유의성에는 미치지 못했지만, 5p14.1 부위와 자폐증과의 연합성이 시사됨을 보고하였다(가장 강력한 신호 : rs10038113, $p=3.4\times10^{-6}$).

최근 Weiss 등[38]은 1,553명의 자폐증 환자가 포함된 1,031

의 자폐증 가족을 대상으로 광범위 유전체 연관(genome-wide linkage) 연구와 GWAS를 시행하였다. 연관 연구에서는 6p27, 20p13 등 4개의 부위에서 LOD 값(score) 2.0 이상의 신호를 발견하였고, 특히 20p13의 LOD 값은 3.81로 광범위 유전체 유의성 역치 3.6 이상을 충족하였다. 전달불균형검정(transmission disequilibrium test, 이하 TDT)를 이용한 GWAS 초기 분석(initial analysis)에서는 광범위 유전체 유의성을 만족시키는 연합성을 발견하지 못하였다. $p<10^{-4}$인 SNP들을 대상으로 독립적인 표본에서 시행한 재현연구에서 semaphorin 5A(이하 SEMA5A)와 Taste receptor type 2 member 1(이하 TAS2R1) 사이에 위치하는 5p15 부위 rs10513025 SNP이 $p<0.01$ 수준의 연합성을 보였다. 초기 분석과 재현연구 자료를 합친 분석에서 이 SNP은 광범위 유전체 유의성을 만족시키는 연합성을 보였다($p=2.1\times10^{-7}$, 순열검정으로 다중비교를 보정). 또한 연구자들은 자폐증 환자의 사후 뇌조직(postmortem brain tissue)에서 SEMA5A 유전자발현이 저하되어 있음을 보고하였는데, 이는 엑손 유도(axonal guidance)에 관여하는 SEMA5A이 자폐증 취약성과 관련되어 있을 가능성을 시사하는 소견이었다.

한편 이 연구에서는 이전 Wang 등[37]의 GWAS에서 자폐증과 관련이 있다고 보고된 5p14 rs4307059 SNP과 자폐증과의 연합성은 재현되지 않았다.

한편 Arking 등[39]은 가족 내에 자폐증 환자가 2명 이상인 72개의 가족들(환자군 148명, 가족군 292명)을 대상으로 Affymetrix 500K를 이용한 GWAS와 연관분석을 시행하였다. GWAS에서 광범위 유전체 유의성을 지닌 SNP을 발견하지 못하였지만, 동일한 데이터를 이용한 광범위 유전체 연관분석에서는 7q35 부위에서 LOD값 3.4의 정점 신호(peak signal)를 발견하였다. 이 신호 아래 1-LOD 유전 간격(genetic interval) 부위에 존재하는 모든 SNP을 사용하여 연합연구를 시행하였는데, rs7794745 SNP이 유의한 연합성을 보였고($p<2.14\times10^{-5}$), 순열검정으로 다중비교에 대해 보정했을 때에도 여전히 유의성은 유지되었다($p<0.006$). 이 SNP은 contactin-associated protein-like 2(이하 CNTNAP2) 유전자의 엑손 2와 3 사이에 위치하는데, CNTNAP2는 axon 분화(differentiation)에 관여하는 것으로 알려져 있다.

정신과 영역의 GWAS에서 고려해야 할 점

Psychiatric GWAS Consortium Coordinating Committee에서는 정신과 질환에서의 GWAS에서 고려해야 사항으로 다음과 같은 점들을 소개하고 있다.[6] 이들 대부분은 모든 GWAS에 공통된 사항이나 일부는 정신과 영역에서 특히 주의를 기울여야 할 사항이다. 첫째, 유전적 상대 위험도(genotype relative risk)가 매우 낮은 표현형의 경우 GWAS 디자인에 적합하지 않다. 또한 유전적 위험이 다수의 드문 SNP 또는 크기가 작은 복사수변이에 의한 경우에도 GWAS를 통해 원인 변이를 효과적으로 발견하기 어려우며, 이와 같은 경우에는 대규모의 염기서열 재분석 연구가 바람직하다. 둘째, GWAS에 현재의 정신과 진단체계는 적합하지 않은 경우가 많다. 따라서 내적표현형(endophenotype)을 통한 연구가 더 바람직할 수 있으나 내적표현형 조사의 경우 대규모 연구를 진행하기 어렵다는 문제가 있다. 셋째, 질병의 유전적 이질성이 검정력을 떨어뜨린다. 표본의 인종적 배경이 비교적 동질한 우리 경우에는 상대적으로 덜 문제가 될 것으로 예상되지만, 최근 GWAS가 다양한 인종에서 다국적/다기관 연구로 진행되기 때문에 유전적 이질성 문제가 발생하기 쉽다. 넷째, 대조군 선정에 많은 주의가 필요하며, 다양한 인종을 대상으로 다양한 모집 방법을 통해 대규모의 대조군을 모으는 것이 필요하다. 다섯째, 어떤 표현형의 경우 SNP의 주효과(main effect)보다 매우 복잡한 유전자-유전자 상호작용이나 유전자-환경 상호작용이 더 중요할 수 있다. 그러나 현재의 GWAS의 방법론으로는 유전자-유전자 상호작용이나 유전자-환경 상호작용을 정확히 분석하는 데에 많은 어려움이 있다. 여섯째, GWAS에서 사용되는 microarray가 모든 흔한 변이들에 대한 정보를 제공해 주지 못한다. Microarray에 따라 차이를 보이기는 하지만, 패널에 포함되어 있지도 않고, 패널 내의 SNP에 의해 적절하게 포착되지도 않은 흔한 SNP들이 많이 존재한다. 특히 microarray를 구성에 기초 자료가 되는 HapMap에 한국인의 SNP 데이터가 포함되어 있지 않은 점을 고려할 때, 이런 문제점은 우리나라 표본에서 더욱 두드러질 가능성이 있다. 일곱째, 복사수변이에 대한 체계적인 정보가 더욱 필요하다. 특히 많은 표현형이 정량적 특성(quantitative traits)인 정신과 영역에 있어 다양한 수적 분포를 보이는 복사수변이는 더욱 주목할 만 하다. 여덟째, 질병의 병태생리에 아직까지 밝혀지지 않은 다른 유전적 기제가 있을 수 있다. 복사수변이, micro RNA, 원거리 촉진자(long-range promoter), 후성인자(epigenetic factor) 등에 여러 유전 기제에 대해서는 이제 겨우 그 중요성을 인식하기 시작한 단계이다.

한편, 지금까지 GWAS를 통해 정신 질환의 취약성과 관련성이 시사된다고 보고된 일부 유전적 변이들도 그 기능 및 증상 발현에 영향을 미치는 기제에 대해서는 거의 알려져 있지 않다. 다만 '신경세포의 발달(development)이나 성숙(maturation)에 관련된 유전적 변이'와 같이 매우 개략적인

기제만을 추정하는 단계이므로 향후 생물학적 기제에 대한 연구가 덧붙여져야 할 것이다.

GWAS 의 제한점과 향후 연구 방향

현재 많은 유전연구들이 GWAS에 집중되고 있지만, 아직까지는 정신과 영역 뿐만 아니라 대부분의 질병 관련 연구에서 뚜렷한 성과를 보이지 못하고 있다. HapMap 프로젝트의 소개와 GWAS의 도입으로 유전정보를 이용한 고위험군(at high risk person)의 선별이나 개별화된 맞춤의학(personalized medicine)이 가능할 것으로 기대했지만, 아직 실제 임상에서 활용되기까지에는 해결해야 할 문제가 많다. 더욱이 하나의 독립된 유전자형이 미치는 영향은 매우 작고, 많은 흔한 유전자들이 함께 작용하여 질병을 유발한다는 CDCV 가정에 근거해 볼 때, 비록 GWAS를 통해 원인 변이를 찾는다 할지라도 매우 적은 효과크기로 인해 임상에서의 활용도는 제한적이 될 가능성이 높다. 따라서 GWAS를 통해 당장 임상 적용이 가능한 결과를 발견하기보다는, GWAS로 확인된 유전자가 질병과 관련된 생물학적 경로(biological pathway) 등 병태생리를 밝히는 단초(clue)를 제공한다는 점에서 제한적인 의미가 있다.

현재 진행되는 GWAS는 소위 "the bigger, the better"의 경향을 보이는 것 같다. 몇몇의 대규모 컨소시움이 수 천명의 표본을 이용한 대규모 연구를 진행하고 있으며, 더욱이 최근에는 이런 대규모 컨소시움들이 서로 데이터를 공유하며 우도비(odd ratio, 이하 OR) 1.1 정도의 적은 영향력을 갖는 SNP을 발견하기 위한 연구를 추진 중이어서, 곧 수 만명의 표본을 대상으로 하는 GWAS 결과가 발표될 것이다. 그러나 임상적 측면에서 보면 현재 주로 진행되는 방식을 통해 OR 1.1 정도의 영향력을 지닌 유전변이를 발견하는 것이 실제 환자의 진단, 치료, 예방 측면에서 볼 때 어떤 의미가 있을 지에 대해서는 회의적이다. 또한 양극성장애, 정신분열병과 같은 여러 정신 질환의 유전율(heritability)이 비교적 높은 데에 비해서 현재까지 발견된 모든 취약성 관련 유전변이들에 의한 표현형이 설명되는 정도는 몇 %에 불과하다. 이는 흔한 변이에 의한 단순 효과 외에도 드문 변이, 복사수변이, 유전자-유전자 상호작용, 유전자-환경 상호작용, 후성효과 등 다른 유전 기제가 관여하기 때문일 것이다.

대규모의 GWAS를 수행함에 있어 질병의 원인을 추정하는 가설이 필요 없는 반면, 복잡하고 엄격한 유전자 데이터의 질 관리와 다중검정 및 인구층화의 보정 등 복잡한 유전자 및 통계적 분석에 대한 의존도는 크게 강조되었다. 이런 관점에서 볼 때, GWAS에 있어 정신과 의사의 역할은 통계학자나 유전학자, 생물정보학자(bioinformatician) 등에 비해 상대적으로 미미하게 느껴질 수 있다. 그러나 GWAS를 비롯한 여러 방법의 유전연구들이 질병의 치료와 이해에 대한 실질적 성과를 도출하기 위해서는 환자, 가족, 표현형, 병의 경과 및 환경적 영향 등에 대한 세밀한 접근이 필요하다. 이와 같은 접근은 대규모 컨소시움이 추구하는 대규모 연구비와 다른 지원들을 얻어내는 데에는 유리하지 않을 지라도, 생물학적 병태생리에 대한 구체적인 이해와 이를 통한 유전적 정보의 임상 활용을 위해서는 오히려 더 중요할 것이다. 따라서 실질적으로 유용한 GWAS 연구를 위해서는 세밀한 접근이 가능한 임상 경험을 지닌 정신과 의사의 역할이 더욱 중요해 질 것이다. 아울러 유전연구에서 정신과 의사들이 이러한 역할을 담당하기 위해서는 유전연구에 대한 기본적인 지식과 정보의 지속적인 습득이 필요하다.

결 론

결론적으로, 현재까지의 GWAS는 여러 가지 기대를 갖게 함과 동시에 상당한 제한점이 있다. 향후 GWAS 수행시 흔한 변이에 의한 단순 효과 외에 드문 변이, 복사수 변이, 유전자-유전자 상호작용, 유전자-환경 상호작용, 후성효과 등 다른 유전 기제에도 관심을 기울여야 할 것이다. 또한 표본 크기를 최대한 크게 늘이는 접근과 함께 표현형을 잘 정의하는 것, 취약 유전자의 기원이 동일할 것으로 생각되는 개별 가족 단위를 자세히 조사하는 것, 유전적으로 동질한 집단을 발굴하는 것, 유전자와 상호작용이 있을 것으로 추정되는 환경적 요인들을 적절히 선정하는 것 등 효과 크기를 크게 함으로써 그 한계점을 극복해 나갈 수 있을 것이다. 보완된 접근의 GWAS가 정신질환에 대한 병태생리의 이해와 이를 통한 조기진단 및 맞춤치료의 실현에 기여할 것을 기대한다.

중심 단어 : 복합 유전 특성 · 광범위 유전체 연합연구 · 정신과 질환.

Acknowledgments

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NO. 2010-0022363).

Conflicts of Interest

The authors have no financial conflicts of interest.

REFERENCES

1) National Institute of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide assocation studies (GWAS). Federal Regis 2007;2007:49290-49297.
2) Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of com-

mon variants to susceptibility to common disease. Nat Genet 2003; 33:177-182.
3) Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502-510.
4) International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-861.
5) Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet 2006;38:659-662.
6) Psychiatric GWAS Consortium Coordinating Committee, Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiaty 2009;166:540-556.
7) McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-369.
8) Howson JM, Barratt BJ, Todd JA, Cordell HJ. Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. Genet Epidemiol 2005;29:51-67.
9) Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-678.
10) Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. PNAS 1998;95:11389-11393.
11) Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 2007;3:1672-1686.
12) Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 2005;37:1243-1246.
13) Ma D, Salyakina D, Jaworski JM, Konidari I, Whitehead PL, Andersen AN, et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. Ann Hum Genet 2009;73:263-273.
14) Neale BM, Purcell S. The positives, protocols, and perils of genome-wide association. Am J Med Genet B Neuropsychiatr Genet 2008; 147B:1288-1294.
15) Benke KS, Fallin MD. Methods: genetic epidemiology. Psychiatr Clin North Am 2010;33:15-34.
16) Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics 2009;10:191-201.
17) Psychiatric GWAS Consortium Steering Committee. A framework for interpreting genome-wide association studies of psychiatric disorders. Mol Psychiatry 2009;14:10-17.
18) Altshuler D, Daly M. Guilt beyond a reasonable doubt. Nat Genet 2007;39:813-815.
19) Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95-108.
20) NCI-NHGRI Working Group on Replication Association Studies, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. Nature 2007; 447:655-660.
21) Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med 2009;360:1759-1768.
22) Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008;299:1335-1344.
23) Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 2009;10:318-329.
24) Corvin A, Craddock N, Sullivan PF. Genome-wide association studies: a primer. Psychol Med 2010;40:1063-1077.
25) O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat Genet 2008;40:1053-1055.
26) Riley B, Thiselton D, Maher BS, Bigdeli T, Wormley B, McMichael GO, et al. Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. Mol Psychiatry 2010;15:29-37.
27) Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, et al. A genome-wide association study in 574 schizophrenia trios using DNA pooling. Mol Psychiatry 2009;14:796-803.
28) Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 2009;5:e1000373.
29) Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry 2008;13:570-584.
30) Potkin SG, Turner JA, Guffanti G, Lakatos A, Fallon JH, Nguyen DD, et al. A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. Schizophr Bull 2009;35:96-108.
31) Shifman S, Johannesson M, Bronstein M, Chen SX, Collier DA, Craddock NJ, et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. PLoS Genet 2008;4:e28.
32) Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, et al. Whole-genome association study of bipolar disorder. Mol Psychiatry 2008;13:558-569.
33) Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry 2009; 14:755-763.
34) Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry 2008;13:197-207.
35) Baum AE, Hamshere M, Green E, Cichon S, Rietschel M, Noethen MM, et al. Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. Mol Psychiatry 2008;13:466-467.
36) Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 2008;40:1056-1058.
37) Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 2009;459:528-533.
38) Weiss LA, Arking DE; Gene Discovery Project of Johns Hopkins & the Autism consortium, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. Nature 2009;461:802-808.
39) Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M, et al. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. Am J Hum Genet 2008; 82:160-164.
40) Athanasiu L, Mattingsdal M, Kähler AK, Brown A, Gustafsson O, Agartz I, et al. Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. J Psychiatr Res 2010;44:748-753.
41) Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, et al. Common variants conferring risk of schizophrenia. Nature 2009;460:744-747.
42) Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 2009;460:753-757.
43) International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748-752.

44) Lencz T, Morgan TV, Athanasiou M, Dain B, Reed CR, Kane JM, et al. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Mol Psychiatry 2007;12:572-580.
45) Lee MT, Chen CH, Lee CS, Chen CC, Chong MY, Ouyang WC, et al. Genome-wide association study of bipolar I disorder in the Han Chinese population. Mol Psychiatry 2010.
46) Djurovic S, Gustafsson O, Mattingsdal M, Athanasiu L, Bjella T, Tesli M, et al. A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. J Affect Disord 2010;126:312-316.

– Appendix –

Box 1. Glossary

영문 용어	한글 용어
Additive interaction	상합적 상호작용
Bioinformatics	생물정보학
Call rate	유전자형분석률
Candidate gene association study	후보 유전자 연합연구
Capture	포착
Causal variant	원인변이
Common disease/common variants (CDCV) hypothesis	흔한 질병/흔한 변이 가설
Common disease/multiple rare variants, (CDMRV) hypothesis	흔한 질병/다수의 드문 변이 가설
Copy number variation	복사수변이
Deep sequencing	고밀도 염기서열분석
Effect size	효과 크기
Endophenotype	내적표현형
Epigenetic factor	후성인자
Epistatic interaction	상위상호작용
False discovery rate (FDR)	거짓발견율
Fine mapping	세밀한 유전지도작성
Frame shift codon	틀이동 코돈
Genetic heterogeneity	유전적 이질성
Genome wide association study (GWAS)	광범위 유전체 연합연구
Genotyping platform	유전자형분석 플랫폼
Haplotype	일배체
Heritability	유전율
Heterozygote	이형접합체
High-throughput resequencing	고효율 염기서열재분석
Imputation method	대체방법
In silico	모의실험
Intergenic region	유전자 사이 부위
Knock-down	유전자발현억제
Linkage disequilibrium (LD)	연관불균형
Long-range promoter	원거리 촉진자
Marker	표식자
Mega-analysis	메가분석
Minor allele frequency (MAF)	소수대립형질 빈도
Genotyping missingness	유전자형분석 실패
Multiplicative genetic effect	승법적 유전효과
Overdominant model	초우성 모델
Paternality	부계관계
Permutation	순열
Personalized Medicine	맞춤의학
Pleiotropy	다면발현
Population substructure	인구하위구조
Potential batch effect	잠재적 배치효과
Protein-coding polymorphism	단백질–부호화 유전자다형성
Quantile-quantile plot, Q-Q plot	분위수 대조도

Box 1. Continued

영문 용어	한글 용어
Recombination hot spot	재조합 고빈도지점
Reference sample	참조 표본
Remote regulatory element	원거리 조절인자
Replication study	재현연구
Signal	신호
Single nucleotide polymorphism (SNP)	단일염기유전자다형성
Statistical power	통계적 검정력
Transmission disequilibrium test (TDT)	전달불균형검정
Winner's curse effect	승자의 저주 효과