KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

Whole Genome Sequencing Analyses to Understand the Genetic Architecture of Human Disorders

안준용

SBi 한국생명정보학회 Korean Society for Bioinformatics
서울대학교 생물정보연구소 SNU BIOINFORMATICS INSTITUTE
Gene·2·us

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

Whole genome sequencing analyses to understand the genetic architecture of human disorders

Lecturer: 안준용 (고려대학교)

전장유전체(whole genome sequencing) 기술은 다양한 종류의 변이를 탐색하여, 하나의 질병을 구성하는 유전적 조성(genetic architecture)를 포괄적으로 이해할 수 있다. 지난 10년간, 발전한 유전체 기술은 대규모 유전체 코호트를 구축하였고, 동시에 생물정보학, 통계적 방법론 발전에 잇대어 질병에 원인이 되는 다양한 유전적 조성들을 밝혀냈다.

본 강의에서는 자폐증 유전체 컨소시엄의 대규모 유전체 연구를 소개하고, 7,602명의 전장유전체 연구와 방법론을 소개하고자 한다. 전장 유전체가 등장하기 전의 사전 연구들에서 다룬 생물학적 가설과 이를 검증하기 위한 실험 설계에 대한 이론적 소개를 하고자 한다. 또한, gene discovery 및 noncoding loci discovery 연구에 사용된 머신러닝 및 statistical application을 소개한다. 향후 전장유전체는 다양한 질병 연구에서 유전적 조성 이해를 위한 기술로 사용될 것으로 예상되며, 이 강의를 통해 전장유전체 유전체 기반 질병 연구의 방향을 이해하고, 질병 유전학의 기본을 학습한다.

강의는 다음의 내용을 포함한다:

- 인간 질병과 polygenic model
- Gene discovery와 noncoding loci discovery
- 질병의 유전적 조성

*참고강의교재:

- 전장 유전체를 이용한 질병 유전학 연구 동향 [BRIC View], https://www.ibric.org/myboard/read.php?Board=report&id=2985
- Genetic architecture: the shape of the genetic contribution to human traits and disease, Timpson et al. (2018) Nature Reviews Genetics

*교육생준비물:

- 노트북

* 강의: 안준용 교수 (고려대학교 바이오시스템의과학부)

Curriculum Vitae

Speaker Name: Joon-Yong An, Ph.D.

▶Personal Info

Name Joon-Yong An

Title Assistant Professor

Affiliation Korea University

▶Contact Information

145 Anam-ro, Seongbuk-gu, Seoul, South Korea

Email joonan30@korea.ac.kr

Phone Number 02-3290-5646

Research interest : Whole genome sequencing, Single cell RNA sequencing, and neurodevelopmental disorders

Educational Experience

2010	B.S. in Molecular Biotechnology, Konkuk University
2011	M.S. in Molecular Biology, University of Queensland (Australia)
2016	Ph.D. in Neuroscience, University of Queensland (Australia)

Professional Experience

2015-2019	Postdoctoral Fellow , University of California, San Francisco
2019-	Assistant Professor, Korea University

Selected Publications (5 maximum)

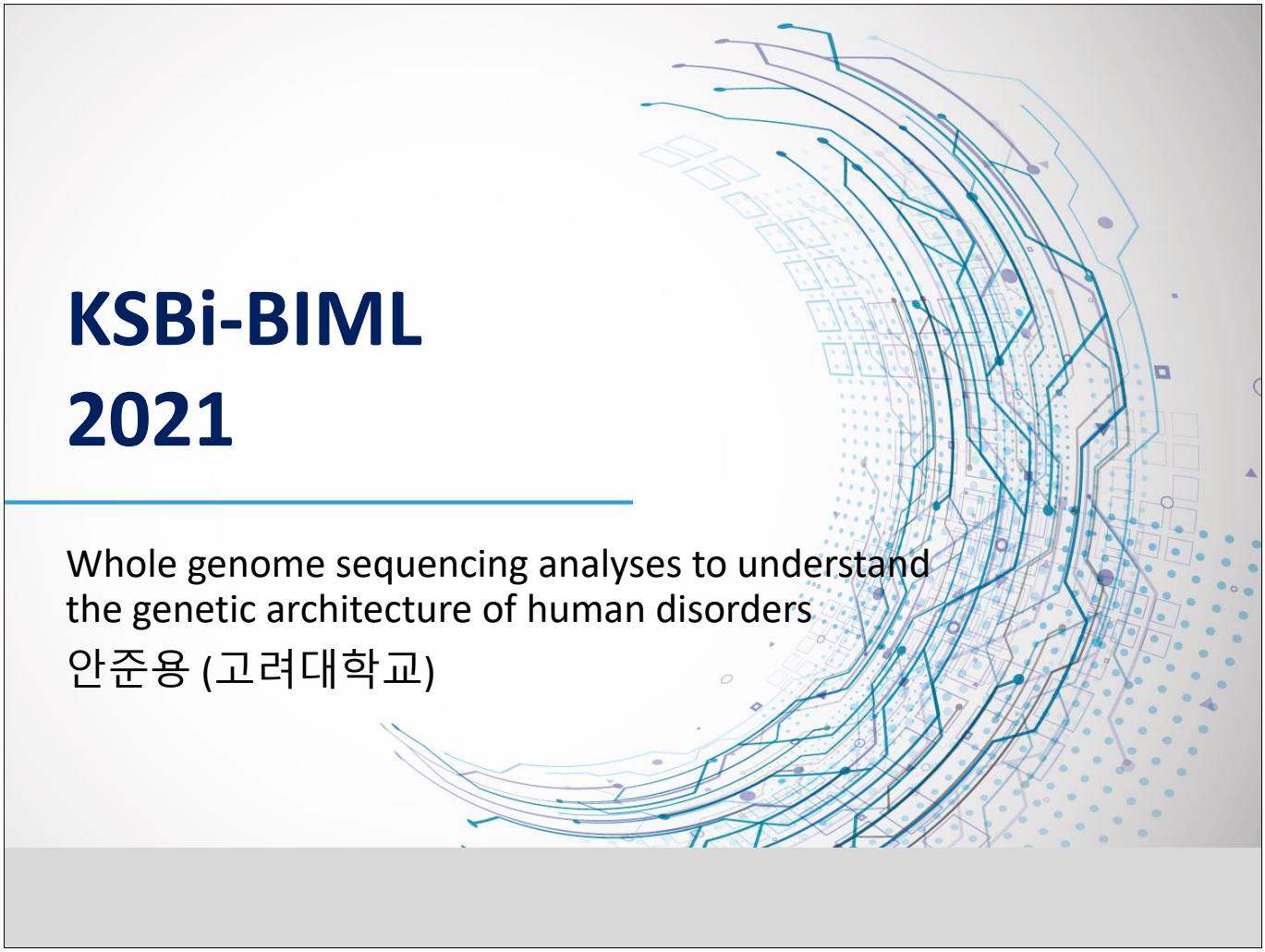
1. Werling DW*, Pochareddy S*, JM Choi*, **An JY*,** Peng M, Sheppard BS, Peng M, Li Z, Dastmalchi C, Santperebaro G, Sousa A, Tebbenkamp A, Kaur N, Gulden F, Breen M, Liang L, Gilson M, Zhao X, Dong S, Klei L, Cicek AE, Buxbaum JD, Adle-Biassette H, Thomas JL, Aldinger KA, O'Day DR, Glass I, Zaitlen N, Talkowski ME, Roeder K, Devlin B, Sanders SJ**, Sestan N**,, Whole-genome and RNA sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex, ***Cell Reports***, 2020

2. Satterstrom FK*, Kosmicki JA*, Wang J*, Breen MS, Rubeis SD, **An JY**, Peng M, Collins RL, Grove J, Klei L, Stevens C, Reichert J, Mulhern M, Artomov M, Gerges S, Sheppard B, Xu X, Bhaduri A, Norman U, Brand H, Schwartz G, Nguyen R, Guerrero E, Dias C, Aleksic B, Anney RJ, Barbosa M, Bishop S, Brusco A, Bybjerg-grauholm J, Carracedo A, Chan MC, Chiocchetti A, Chung B, Coon H, Cuccaro M, Curr A, Bernardina BD, Doan R, Domenici E, Dong S, Fallerini C, Fernndez-prieto M, Ferrero GB, Freitag CM, Fromer M, Gargus JJ, Geschwind D, Giorgio E, Gonzlez-peas J, Guter S, Halpern D, Hassen-kiss E, He X, Herman G, Hertz-picciotto I, Hougaard DM, Hultman CM, Ionita-laza I, Jacob S, Jamison J, Jugessur A, Kaartinen M, Knudsen GP, Kolevzon A, Kushima I, Lee SL, Lehtimki T, Lim ET, Lintas C, Lipkin WI, Lopergolo D, Lopes F, Ludena Y, Maciel P, Magnus P, Mahjani B, Maltman N, Manoach DS, Meiri G, Menashe I, Miller J, Minshew N, Souza EMMd, Moreira D, Morrow E, Mors O, Mortensen PB, Mosconi M, Muglia P, Neale B, Nordentoft M, Ozaki N, Palotie A, Parellada M, Passos-bueno MR, Pericak-vance M, Persico A, Pessah I, Puura K, Reichenberg A, Renieri A, Riberi E, Robinson E, Samocha KE, Sandin S, Santangelo SL, Schellenberg G, Scherer S, Schlitt S, Schmidt R, Schmitt L, Silva IMW, Singh T, Siper P, Smith M, Soares G, Stoltenberg C, Suren P, Susser E, Sweeney J, Szatmari P, Tang L, Tassone F, Teufel K, Trabetti E, Trelles MdP, Walsh C, Weiss L, Werge T, Werling D, Wigdor EM, Wilkinson E, Willsey JA, Yu T, Mullin H, Yuen R, Zachi E, Betancur C, Cook EH, Gallagher L, Gill M, Lehner T, Senthil G, Sutcliffe JS, Thurm A, Zwick ME, Brglum AD, Cicek AE, Talkowski ME**, Cutler DJ**, Devlin B**, Sanders SJ**, Roeder K**, Daly MJ**, Buxbaum JD**, Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism, ***Cell***, 2020

3. **An JY*,** Lin K*, Zhu L*, Werling DM*, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, Currall BB, Dastmalchi C, Dea J, Duhn C, Gilson MC, Klei L, Liang L, Markenscoff-papadimitriou E, Pochareddy S, Ahituv N, Buxbaum JD, Coon H, Daly MJ, Kim YS, Marth GT, Neale BM, Quinlan AR, Rubenstein JL, Sestan N, State MW, Willsey AJ, Talkowski ME**, Devlin B**, Roeder K**, Sanders SJ**, Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder, ***Science***, 2018

4. Werling DM*, Brand H*, **An JY***, Stone MR*, Zhu L*, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, Farrell A, Schwartz GB, Wang HZ, Currall BB, Zhao X, Dea J, Duhn C, Erdman CA, Gilson MC, Yadav R, Handsaker RE, Kashin S, Klei L, Mandell JD, Nowakowski TJ, Liu Y, Sirisha Pochareddy S, Smith L, Walker MF, Waterman MJ, He X, Kriegstein AR, Rubenstein JL, Sestan N, McCarroll SA, Neale BM, Coon H, Willsey AJ, Buxbaum JD, Daly MJ, State MW, Quinlan AR, Marth GT, Roeder K, Devlin B**, Talkowski M**, Sanders SJ**, An analytical framework for whole genome sequence data and its implications for autism spectrum disorder, ***Nature Genetics***, 50:727736, 2018

5. Williams SM*, **An JY***, Edson J, Watts M, Murigneux V, Whitehouse AJO, Jackson CJ, Bellgrove MA, Cristino AS**, Claudianos C**, An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder, ***Molecular Psychiatry***, 20



본 강의 자료는 한국생명정보학회가 주관하는 KSBi-BIML 2021 워크샵 온라인 수업을 목적으로 제작된것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다. 수업 목적으로 배포 및 전송 받은 경우에도 이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없습니다.

만약 이러한 사항을 위반할 경우 발생하는 모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고합니다.

강의 개요

- 전장 유전체를 연구하기 위해 알아야 할 것들
  - 유전적 조성
  - 왜 우리는 전장 유전체 기술까지 당도하였는가...
- 전장 유전체 질환 연구 – 자폐증을 중심으로

Whole Genome Sequencing (전장 유전체 기술; 이하 WGS)

- DNA의 모든 염기서열을 읽을 수 있는 NGS 기반의 유전체 해독 방법
  - NGS의 개발과 함께 도입된 원시 genomics 방법
  - WGS 도입 초기에는 효율성과 경제성이 낮았으나, 점차 활용도와 가격 경쟁력이 높아지는 추세
- 질병이나 형질에 관여하는 변이를 찾고, 연관된 변이/유전자를 바탕으로 진단을 함
- WGS 연구를 위해선 **"유전적 조성"**에 대한 이해가 필수적임

Genetic Architecture (유전적 조성)

"하나의 형질(질환)이지만, 너무나 다양하게 구성된 유전적 원인..."

똑똑한 사람들은 어머니에게 감사해야 한다. 전문가의 주장에 따르면, 지능 유전에서 어머니가 지배적인 영향을 주는 것으로 밝혀졌다.

아버지·어머니의 유전자를 받는 아이의 유전자에 있어 흥미로운 점은, 다른 유전자에 의해 영향을 받은 유전자는 어머니쪽 유전자에 대해서만 반응한다는 점이다. 만약 같은 유전자가 아버지쪽에서 영향을 주려고 하더라도, 이에는 반응하지 않는다.

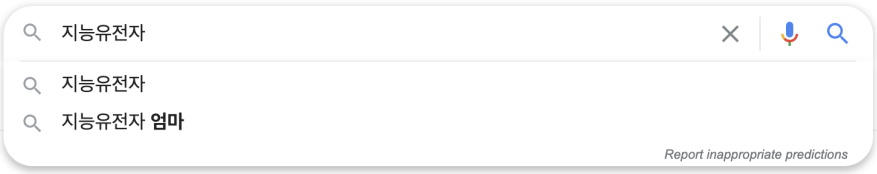
어머니의 유전자는 바로 대뇌 피질(대뇌에서 가장 겉에 위치하는 신경세포들의 집합으로 고차원적 기능을 수행)로 영향을 주고, 아버지의 유전자는 대뇌의 변연계(인체의 기본적인 감정이나 욕구를 관장하는 신경계)로 가기 때문이다.

사람들은 보통 지능을 유전적 성분으로 인식하고 있다. 하지만 지능 유전자는 X 염색체에 위치하고 있기 때문에 어머니의 영향을 받게 된다.

이 분야에서 첫 연구는 1984년 케임브리지 대학에서 진행됐다. 실험의 주 가설은, 어머니쪽의 세포가 뇌 발전에 지배적인 영향을 끼친다는 주장이었다.

첫 실험에서 연구자들은 특수한 쥐를 사용했다. 어머니의 유전자, 혹은 아버지의 유전자로만 만들어진 쥐였다. 하지만 이 쥐를 자궁에 옮겼을 때, 쥐는 바로 죽어버렸다. 연구자들은 이 과정에서 어머니의 유전자가 뇌에서 활성화 되는 것을 발견했고, 반대로 아버지의 유전자는 활성화되지 않지만 쥐의 성장에 필수적이라는 결론을 내렸다.

[출처: 중앙일보] '내가 똑똑한 건 어머니 덕분' 지능 유전자의 진실



news.joins.com › article

'내가 똑똑한 건 어머니 덕분' 지능 유전자의 진실 - 중앙일보

Sep 19, 2016 — 전문가의 주장에 따르면, **지능** 유전에서 어머니가 지배적인 영향을 주는 것으로 밝혀졌다.아버지·어머니의 **유전자**를 받는 아이의 **유전자**에 있어 ...

youtube.com › watch

'지능 유전자' 엄마로부터 물려받는다! - YouTube

Oct 10, 2016 — **'지능 유전자' 엄마**로부터 물려받는다! 29,003 views29K views. • Oct 10, 2016.

kr.people.com.cn › ...

[과학 탐구] 자녀 지능 엄마 닮는다? 전문가들의 속시원한 답변

May 13, 2019 — [인민망 한국어판 5월 13일] 한 연구결과에 따르면, **지능 유전자**가 X염색체에 있다. 여성의 경우 X염색체가 두 개, 남성의 경우 하나뿐이기 때문에 ...

insight.co.kr › news

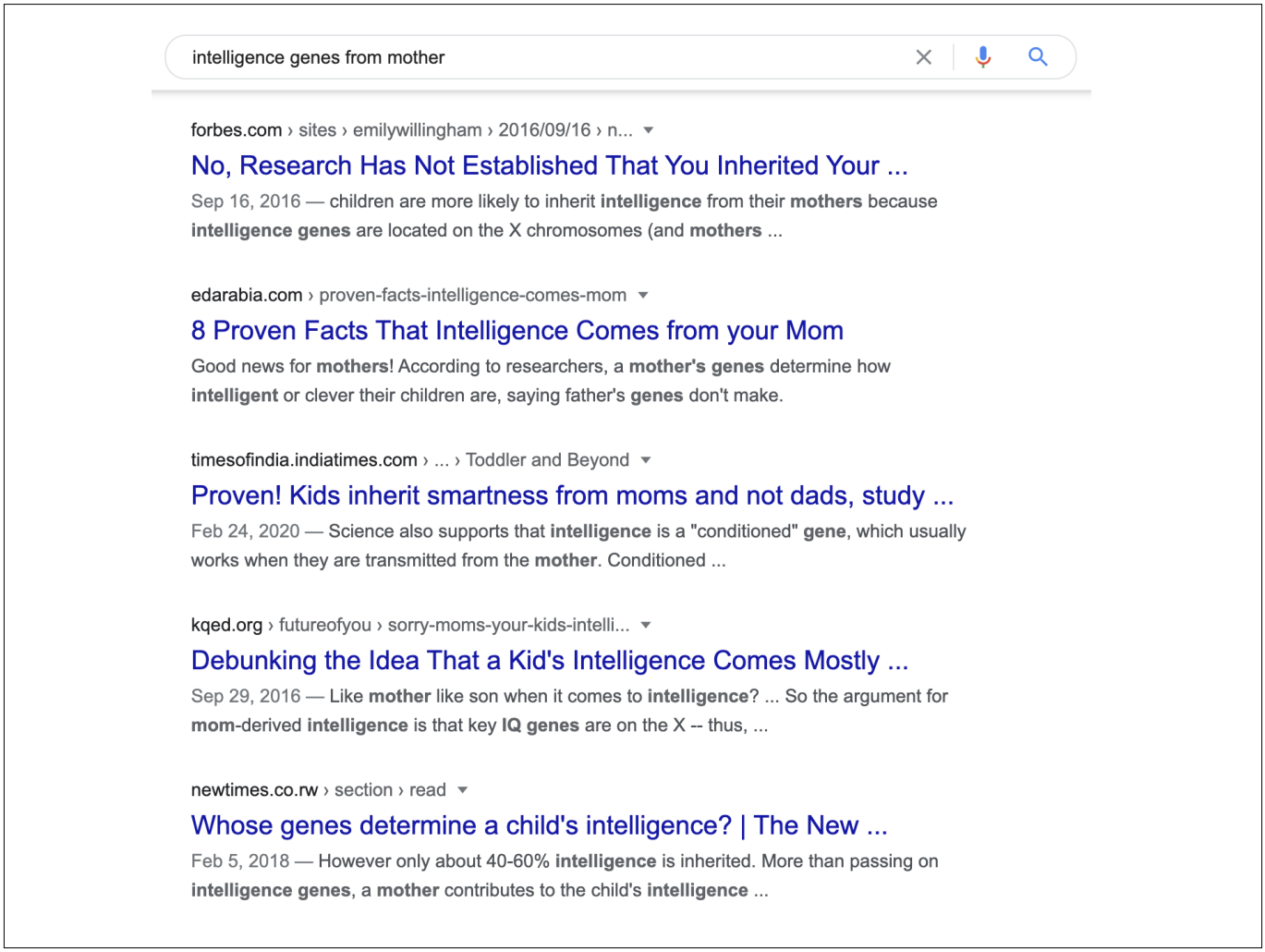
“아이들은 아빠 아닌 '엄마' 지능을 그대로 물려받는다” - 인사이트

May 1, 2019 — 아무리 똑똑한 아빠를 두었더라도 유전적으로 자녀의 지능에는 영향을 끼치지 못한다는 것이다. 그 이유는 **지능 유전자**가 X염색체에 있기 때문인데, ...

m.segye.com › view

"아들이 공부 잘 하면 엄마 덕분일까?" - 세계일보

Feb 6, 2020 — 그런데 한 연구에서 자녀의 **지능**은 **엄마**로부터 물려받으며 특히 아들의 경우 전적으로 **엄마**의 **유전자** 영향을 받는다고 밝혔다. 최근 과학 전문 ...



이 기사는 무엇이 잘못되었는가?

- 1970년대 지적장애의 발견, 일반적으로 남성에게 유병률이 높음.

- "그러면 성 염색체가 원인이겠네!"
  - Lehrke (1972) *Theory of X-linkage of major intellectual traits*
  - X 염색체의 지능 유전자, "엄마에게서 아들에게로"

- 성 염색체는 당시로써 제기할 수 있는 **유일한 가설**
  - **그래서 "전장(genome-wide)"가 중요합니다...**

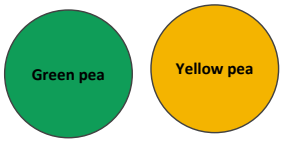
이 기사는 무엇이 잘못되었는가?

- 지능은 "**polygenic model**"을 따름.
  - 지능에는 additive genetic factor가 관여
  - 지능에는 dominant genetic factor가 관여
  - 지능에는 epistatic genetic factor가 관여

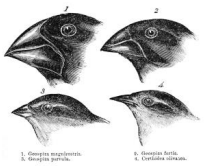
  - 동시에 지능에는 environmental factor가 관여
    - 독립적으로...
    - 혹은 genetic locus와 interaction하는 modifier로...

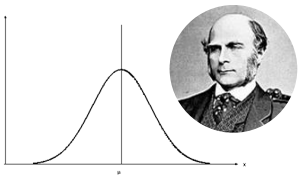
19세기 현대 유전학의 태동
멘델 vs. 다윈 vs. 골턴

그레고어 멘델 (1822-1884)
"The existence of binary hereditary factors"

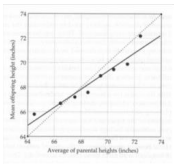


찰스 다윈 (1809-1882)
"There is variation in natural population"

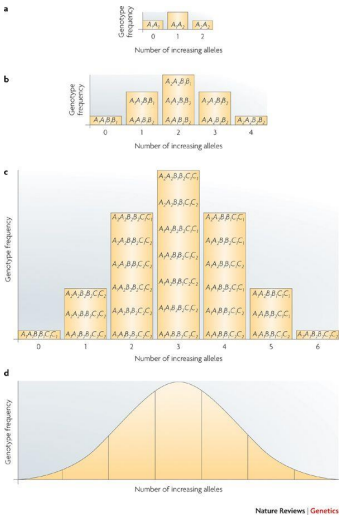




프란시스 골턴 (1822-1911)
"Biological traits are continuous"



The height vs. pea debate
(early 1900s)



멘델학파 (Mendelian)

Mendel (1866) → Bateson (1902) → Fisher (1918) → Morgan (1915) → Falconer (1965)

Darwin (1868) → Pearson (1902) → Fisher (1918)

생물계측학파 (Biometrician)

R. A. Fisher's 1918 paper
'The correlation between relatives on the supposition of Mendelian inheritance'

- Fisher의 모델은 "Polygenic Model"의 초안

- "표현형 차이(phenotypic variance)는 여러가지 요소에 의해 설명할 수 있다"
  - 세가지 유전적 요인 (G):
    - Additive (A; a large number of loci contribute little)
    - Dominance (D)
    - Epistasis: (I) interaction between additive factors or additive - dominance factors
  - One non-genetic factor - environment (E)
  - Interaction between genetic and environmental factors (G x E)

- 따라서, **<u>형질/질병에 대한 다양한 구성요소(유전적 조성) 연구가 필요</u>**

유전력 (heritability)과 polygenic model

- 본성(Nature)과 양육(Nurture) – 매트 리들리 (2003)
  - *"악마여, 그대는 타고난 본성을 거스를 수 있는가?"* (세익스피어)
- 그러나 유전학은 이분법적인 개념이 아님.
  - "대머리는 유전인가요?"는 잘못된 질문, "대머리에 몇% 유전적 요인이 기여하나요?"
- 형질/질병은 유전적 요인과 환경적 요인이 모두 기여함.
- "**유전적 요인이 얼마나 기여하는가?**"가 유전학 연구의 주된 물음

"질병에 유전적 요인이 얼마나 기여하는가?"라는 물음은 확장되어야

- "유전적 요인"
  - 유전자, 유전 변이
- "얼마나"
  - 어떻게 측정할 것인가? 그 측정은 재현되는가?
- 전장유전체는 위의 모든 연구 주제를 다룰 수 있는 기술에 해당함
- 그러나 각 주제들은 각기 다른 맥락에서 역사적으로 발전했으며, 이 흐름을 이해할 필요가 있음

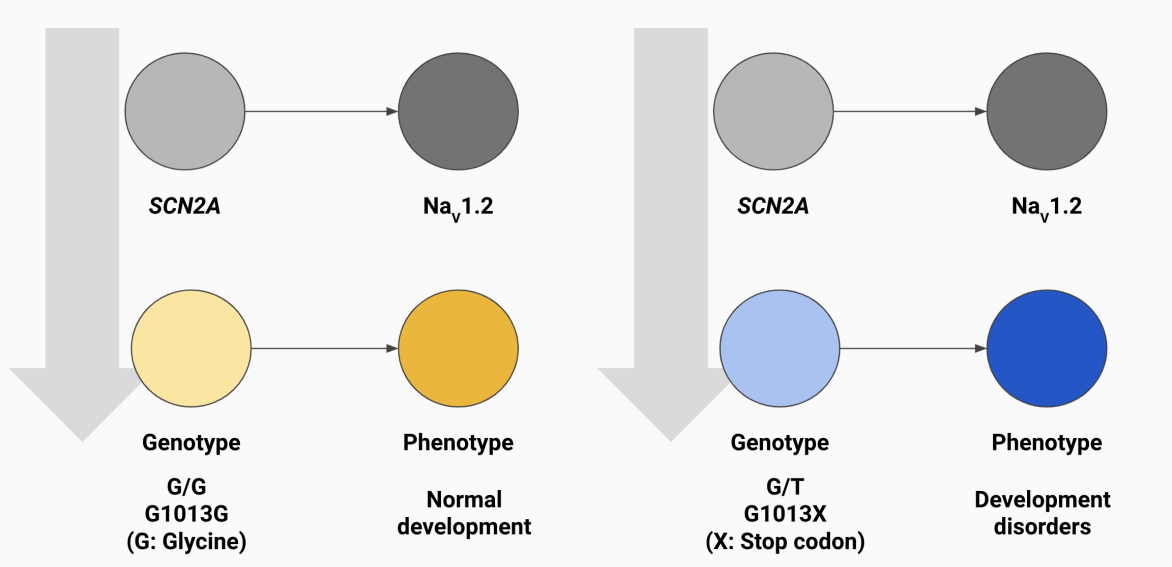
유전 변이 (genetic variant)

염기서열의 변화가 곧 allele을 의미함

Genetic variant == Allele

- DNA -> RNA -> Protein -> Trait
  - Sequence variation (즉 genetic variant)는 trait의 변화를 가져올 수 있음
- 분자생물학 기술의 발전(~80년대)은 유전자 수준의 central dogma를 확립함
- NGS의 도입은 genetic variant 수준의 central dogma를 확립함
  - Genetic variant는 allele의 원래적 의미를 내포함과 동시에 개체가 속한 집단과 진화적 특성을 반영함

Genetic variant는 현대적 의미에서 곧 allele을 의미

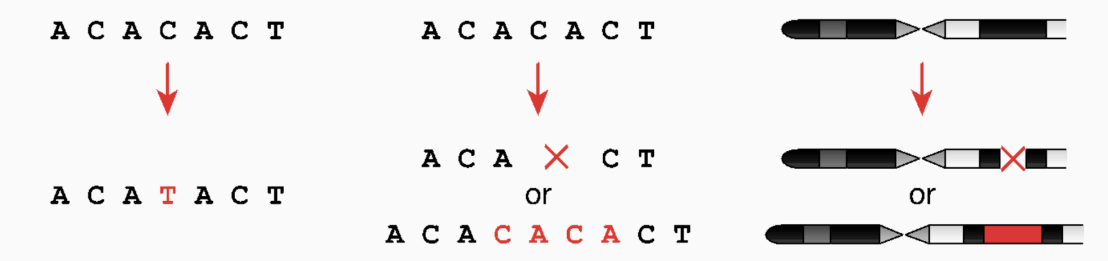


유전 변이와 유전적 조성의 관계: 크기에 따른 분류

SNV or SNP
(Single Nucleotide Variants)

Indel
(Insertion/Deletion)

SV
(Structural variants)



- Copy-number (changing) variants (CNV)
  - Deletion
  - Duplication
- Copy-number neutral variants
  - Inverstion
  - Insertion
  - Transposon
- Complex SVs (DEL-INV)

변이의 크기는 곧 영향의 크기를 의미하기도 함

- 직관적 해석: 100kb의 deletion은 2bp deletion보다 형질/질병에 큰 영향을 초래함
- 큰 영향은 자연 선택에 영향을 미치고, 궁극적으로 SV는 집단에 존재할 확률이 낮아짐

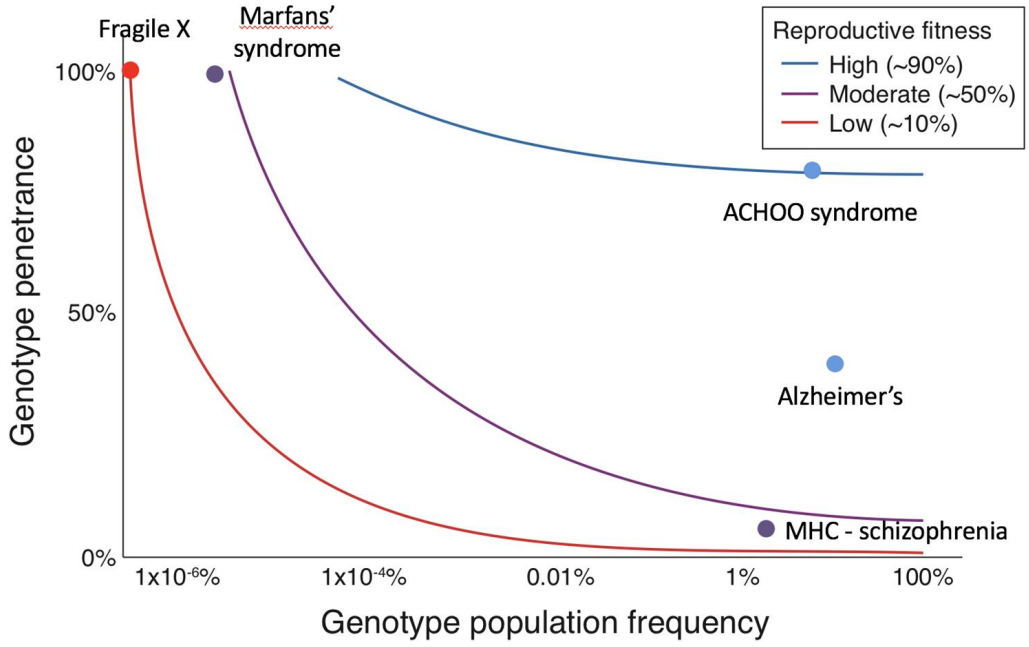
- 한 사람에게 존재하는 변이의 수 (1000 Genome Project)
  - SNV ~4,000,000개
  - Indel ~300,000개
  - SV ~2,000개

A global reference for human genetic variation

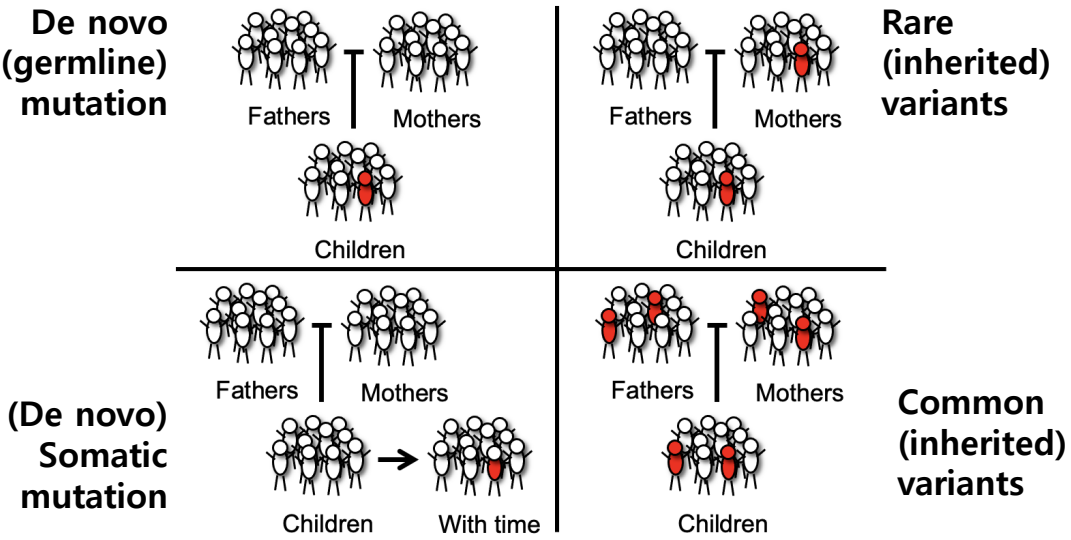
The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

영향의 크기는 곧 변이의 빈도를 결정함



변이의 발생 방식은 질환마다 다름

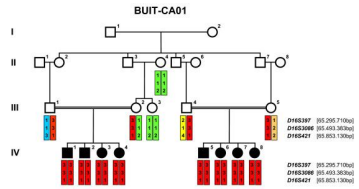


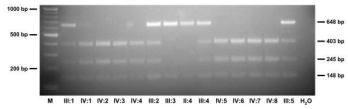
같은 전장유전체 기술을 사용해도, 각 유전 변이의 연구 방법은 다릅니다

원인 유전자 연구사

왜 우리는 전장 유전체까지 오게 되었는가?

80-90년대
"전통적 후보 유전자"의 시대





1990 2000 2010 Present

- 분자생물학 기법(제한 효소 등)의 도입으로 유전체의 한 두가지 loci를 읽어 낼 수 있음.
- Linkage analysis 방법론을 바탕으로 가족 내에 disease segregation을 연구함.
- "후보 유전자"에 대한 기능검증/실험동물 연구가 시작됨
- APOE4 (알츠하이머), TP53 (암) 등등

전통적 후보 유전자 시대
Traditional Candidate Gene Study

90년대 전후, 많은 가설들이 등장하였다

DISC1
Schizophrenia candidate gene in one Scottish family 2001

CACNA1A, ATP1A2, SCN1A
Familial hemiplegic migraine candidate gene found in 4, 2, 3 families in 1996, 2003, 2005

AUTS2 (Autism Susceptibility Candidate 2)
Found in 7q11.2 translocation in two twins with autism in 2002.

26

20년이 지난 현재, 이 발견들은 재현되었는가?

14만명의 조현병 GWAS 연구에서 재현되지 않음 (Ripke 2014 *Nature*)
2만명의 조현병 엑솜 유전체 연구에서 재현되지 않음 (Singh 2019 *Nat Neuro*)

DISC1
Schizophrenia candidate gene in one Scottish family 200

왜?

CACNA1A, ATP1A2, SCN1A
Familial hemiplegic migraine candidate gene found in 4, 2, 3 families in
1996, 2003, 2005

8,319명 GWAS 연구에서 재현되지 않음 (Gormely 2018 *Neuron*)

AUTS2 (Autism Susceptibility Candidate 2)
Found in 7q11.2 translocation in two twins with autism in 2002.

35,000명 엑솜 유전체 연구에서 재현되지 않음 (Satterstrom 2020 *Cell*)

27

승자의 저주 Winner's Curse 는 유전학/생물학 연구에서 매우 흔하다

- 승자의 저주? 최초 연구에서 측정한 효과 크기(effect size)는 overestimation
  - 왜? 작은 샘플 사이즈 연구, 통계적 방법론 부재, 재현성 연구의 부재
- 제 2 당뇨의 *PPARγ* 유전자의 Pro12Ala mutation
  - 4.4 effect size (n=91; Deeb1998)
  - 1.3 effect size (n=3,000; Altshuler 2000)
  - 1.1 effect size (n=12,940; Fuschenberg 2016)
- 전통적 후보 유전자 연구의 96% 이상은 승자의 저주에 시달리고 있음 (Lohmuller 2003)
- 유전학 분야는 불가지론적인agnostic 통계적 방식을 제안하고 개발함

28

2013년 그 언젠가..
DISC1 유전자를 둘러싼 논쟁

Molecular Psychiatry

Guest Editorial | Published: 23 September 2013

Questions about DISC1 as a genetic risk factor for schizophrenia

P F Sullivan

Molecular Psychiatry **18**, 1050–1052 (2013) | Download Citation

Molecular Psychiatry

Letter to the Editor | Published: 17 December 2013

DISCovery in psychiatric genetics

A B Niculescu III

Molecular Psychiatry **19**, 145 (2014) | Download Citation

Sullivan 교수
"우리가 14만명 쯤 조현병 환자 유전체를 살펴보니, DISC1 연관성은 찾을 수 없던데? 이거 정말 후보 유전자 맞음?"

Niculescu 교수
"무슨 소리? 지금까지 실험적으로 입증된 DISC1 연구가 얼마나 많은데? 너희들처럼 통계만 아는 사람들이 생물학을 아는가?

과학의 유구한 전통에 따라 제언하건대, 2020년까지 연구해서 , 틀린 사람이 "내탓이오 내탓이오 나의 탓이로소이다 (Mea Culpa)"라는 논평을 이 저널에 내자"

29

ARTICLE

doi:10.1038/nature16549

Schizophrenia risk from complex variation of complement component 4

Aswin Sekar[1,2,3], Allison R. Bialas[4,5], Heather de Rivera[1,2], Avery Davis[1,2], Timothy R. Hammond[4], Nolan Kamitaki[1,2], Katherine Tooley[1,2], Jessy Presumey[5], Matthew Baum[1,2,3,4], Vanessa Van Doren[1], Giulio Genovese[1,2], Samuel A. Rose[2], Robert E. Handsaker[1,2], Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Mark J. Daly[2,6], Michael C. Carroll[5], Beth Stevens[2,4] & Steven A. McCarroll[1,2]

Sullivan 교수, 조현병 유전체 컨소시엄은 조현병 유전자 108개를 발견함

그 중 C4 유전자가 포함되어 있었고, C4의 변이 종류에 따라 뇌 발달 과정에서 Glial cell, 조현병의 관계를 밝힘
(Sekar 2016)

Cell

Resource

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,*] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[4] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[5] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[3] John J. Trombetta,[4] David A. Weitz,[5,10] Joshua R. Sanes,[3] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

C4 변이 종류를 밝히기 위해, 미세유체 기술을 고안함

응? 이거 하다 보니 세포 종류도 찾게 되네?
"싱글셀 전사체 기술" 등장

1980년대 Feinberg의 조현병과 Synpatic Prunning 가설이 다시 주목받게 됨

J Psychiatr Res. 1982-1983;17(4):319-34. doi: 10.1016/0022-3956(82)90038-3.

Schizophrenia: Caused by a Fault in Programmed Synaptic Elimination During Adolescence?

I Feinberg

PMID: 7187776 DOI: 10.1016/0022-3956(82)90038-3

30

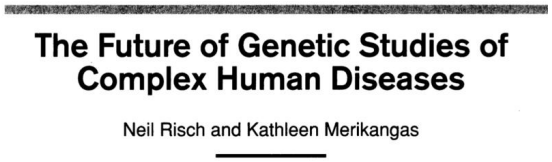
전장, 전장(genome-wide)이 중요하다..

- 전장 (genome-wide)
- 실험자의 직관에 의존하지 않는 가설 설정과 실험 설계에 대한 필요성이 대두됨.

- 한계?
  - 모든 유전자들의 위치를 알지 못함
  - 가설 고전적인 방식 (제한효소 절편길이 다형성)은 다수의 유전변이를 탐색하는데 한계가..

- 아, 그러면 몇개의 변이를 탐색 해야하지?

Risch & Merikangas (1996) Science


The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

- Linkage disequilibrium로부터 독립적인 위치(loci)의 숫자를 시뮬레이션으로 예측

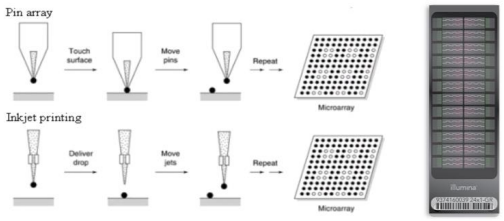
- **1,000,000개의 loci**가 genome-wide 연구를 위해 필요 (GWAS의 p값 기준인 5E-9)

- 그리고 때마침 등장한 기술...

2000년대
마이크로어레이 그리고 Genome-wide association 연구의 서막

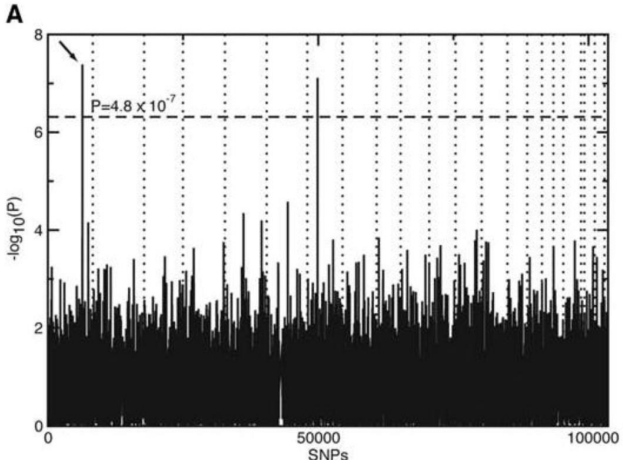
1990 2000 2010 Present

- Microarray technology for common variants (~10 to 100k SNPs)
- GWAS - "hypothesis-free approach", finding loci associated with traits or disorders.
- Measure the heritability in polygenic traits.

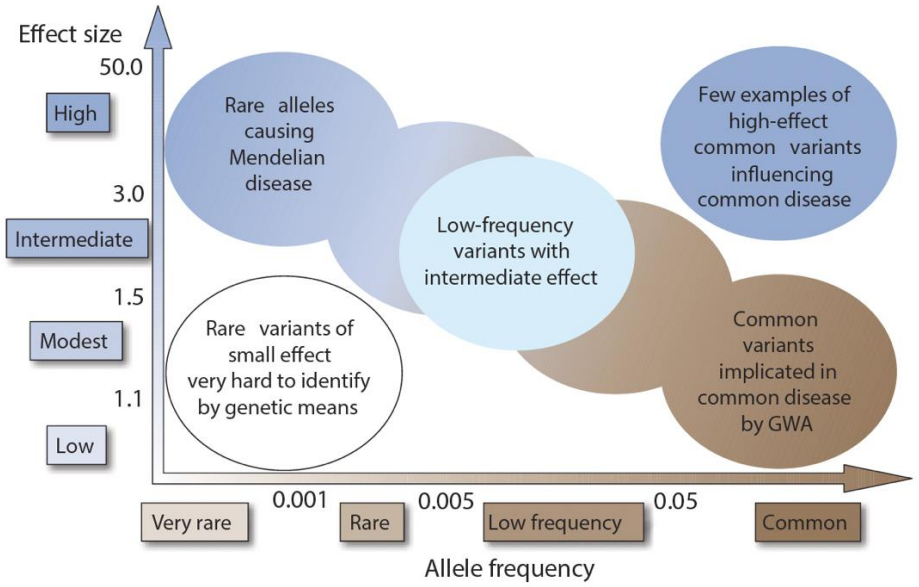


Klein et al. (2005) Science

"**Age-related macular degeneration (AMD)** is a major cause of blindness in the elderly. We report a genome-wide screen of **96 cases and 50 controls for polymorphisms associated with AMD**. Among **116,204 single-nucleotide polymorphisms** genotyped, an intronic and common variant in the **complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$)**"



2009년 Missing Heritability 논쟁
Manolio et al. (2009) Nature

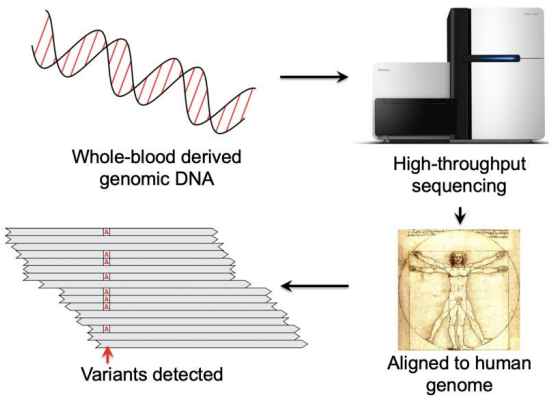


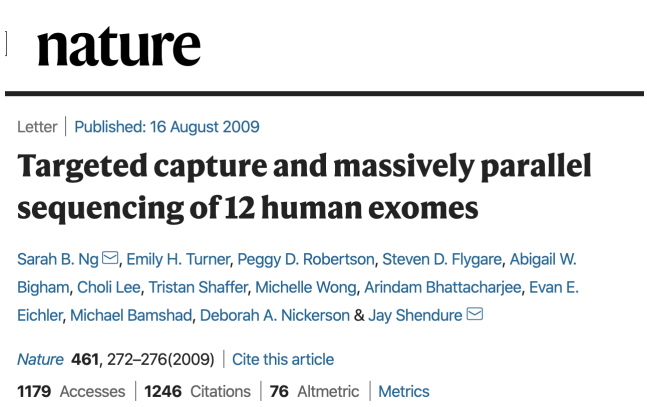
Missing heritability news 2008

2010년대 전후
Next-generation sequencing (NGS)의 도입

1990 2000 2010 Present

- NGS technology to read all (or most) nucleotides in your geno me.
- Find a wide range of genetic var iants (SNVs, indels, SVs)





nature

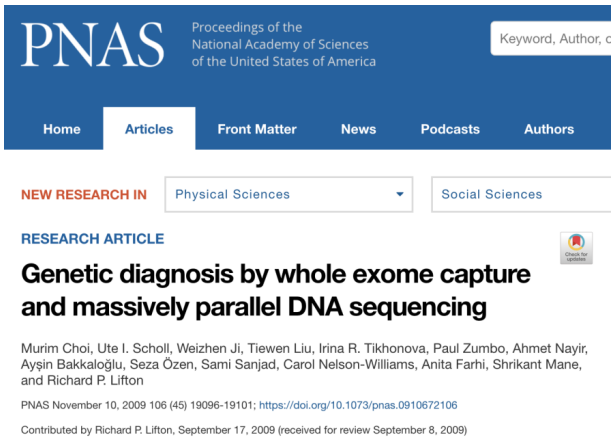
Letter | Published: 16 August 2009

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng, Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E. Eichler, Michael Bamshad, Deborah A. Nickerson & Jay Shendure

Nature **461**, 272–276(2009) | Cite this article

1179 Accesses | **1246** Citations | **76** Altmetric | Metrics



PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, c

Home | Articles | Front Matter | News | Podcasts | Authors

NEW RESEARCH IN Physical Sciences Social Sciences

RESEARCH ARTICLE

Genetic diagnosis by whole exome capture and massively parallel DNA sequencing

Murim Choi, Ute I. Scholl, Weizhen Ji, Tiewen Liu, Irina R. Tikhonova, Paul Zumbo, Ahmet Nayir, Ayşin Bakkaloğlu, Seza Özen, Sami Sanjad, Carol Nelson-Williams, Anita Farhi, Shrikant Mane, and Richard P. Lifton

PNAS November 10, 2009 106 (45) 19096-19101; https://doi.org/10.1073/pnas.0910672106

Contributed by Richard P. Lifton, September 17, 2009 (received for review September 8, 2009)

원인 유전자 발굴 연구 (Gene Discovery)

- Exome sequencing을 활용한 연구
- Exome: 단백질을 생성하는 유전체 부위
  - ~1%의 유전체 부위에 해당함
- 단백질은 표현형에 직접적인 영향을 주기에, 영향을 크게 미치는 유전 변이를 탐색할 것으로 예상

- Gene Discovery: 질환군에서 유전변이가 빈번하게 나타나는 유전자를 선별

문제는 인간에게 너무나 많은 변이가 관찰됨...

Science Contents News Careers Journals
Read our COVID-19 research and news.

SHARE

RESEARCH ARTICLE

A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes

Daniel G. MacArthur[1,2,*], Suganthi Balasubramanian[3,4], Adam Frankish[1], Ni Huang[1], James Morris[1], Klaudia Walter[1], Luke J...
+ See all authors and affiliations

Science 17 Feb 2012:
Vol. 335, Issue 6070, pp. 823-828
DOI: 10.1126/science.1215040

Article Figures & Da

Abstract

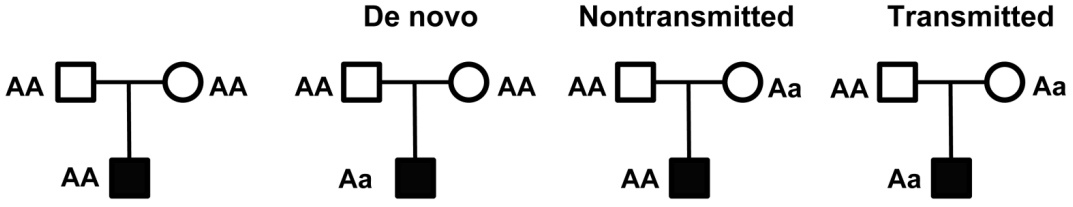
Genome-sequencing studies indicate that all humans carry many genetic variants predicted to cause loss of function (LoF) of protein-coding genes, suggesting unexpected redundancy in the human genome. Here we apply stringent filters to 2951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties. We estimate that human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated. We identify rare and likely deleterious LoF alleles, including 26 known and 21 predicted severe disease–causing variants, as well as common LoF variants in nonessential genes. We describe functional and evolutionary differences between LoF-tolerant and recessive disease genes and a method for using these differences to prioritize candidate genes found in clinical sequencing studies.

PLOS | GENETICS

Integrated Model of *De Novo* and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes

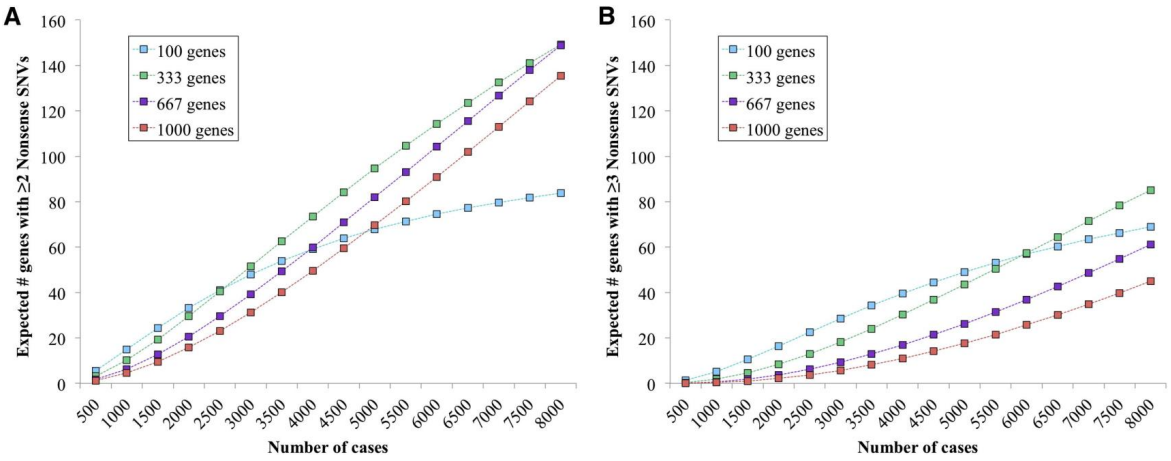
Xin He, Stephan J. Sanders, Li Liu, Silvia De Rubeis, Elaine T. Lim, James S. Sutcliffe, Gerard D. Schellenberg, Richard A. Gibbs, Mark J. Daly, Joseph D. Buxbaum, Matthew W. State, Bernie Devlin, Kathryn Roeder

Published: August 15, 2013 • https://doi.org/10.1371/journal.pgen.1003671



Genotype prob.	$(1-q)^2 \times (1-2\mu)$	$(1-q)^2 \times 2\mu$	$2q(1-q) \times \frac{1-\mu}{2}$	$2q(1-q) \times \frac{1+\mu}{2}$
Phenotype prob.	f_0	γf_0	f_0	γf_0
Rate	–	$2\mu\gamma N_d$	qN_d	$q\gamma N_d$

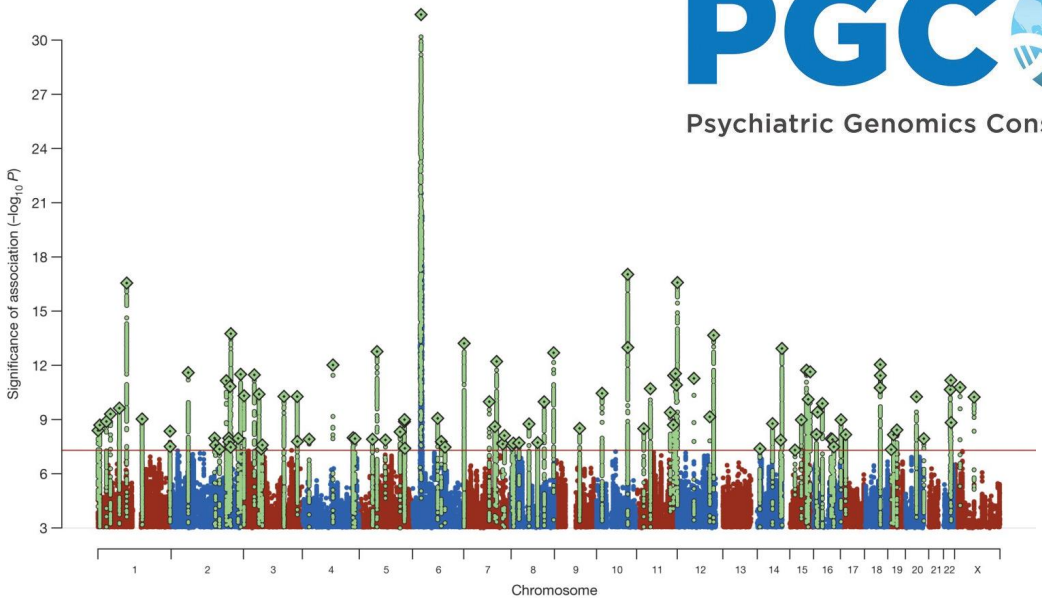
Rare variant 연구를 위해선 매우 많은 sample size가 필요함



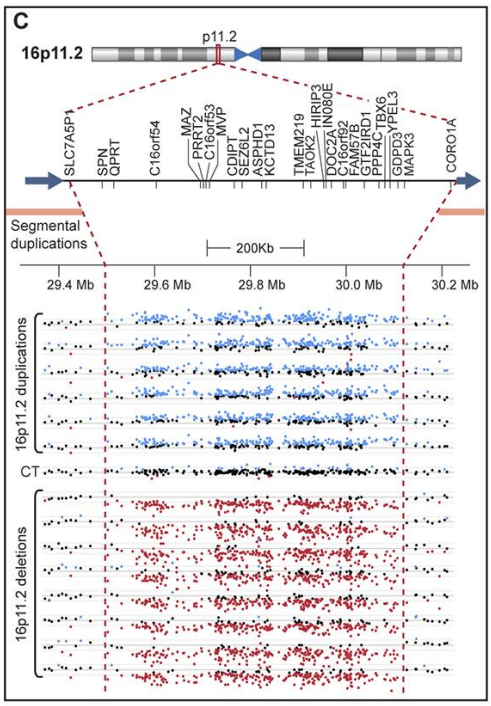
Buxbaum (2012) Neuron

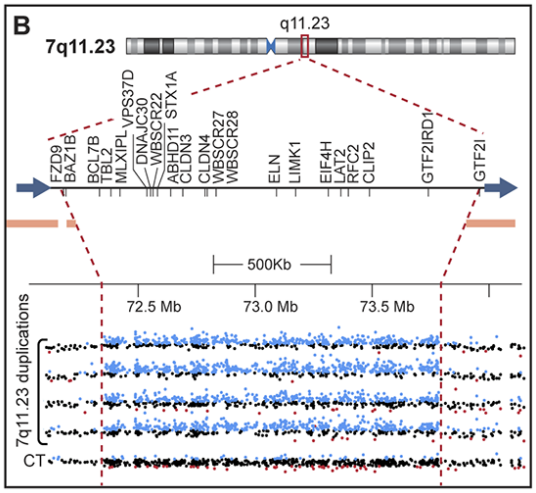
2015년대부터 현재까지 국가, 대형 유전체 컨소시엄의 등장

- Team efforts in large scale research consortium or cohort enabled robust association studies (cancer, autism, type 2 diabetes, schizophrenia, psychiatric disorders, inflammatory bowel disease, etc.)
- New methodologies for association have been developed.
- Large-scale genomic datasets from "cohort studies" (UK BioBank, Danish iPsych birth cohort).
- Heritability measured by rare variants.



PGC Schizophrenia, *Nature*, 2014

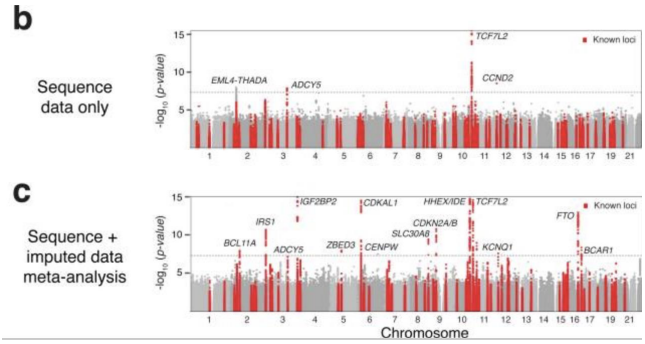




7q11.23 deletion -> Williams Syndrome -> Hypersociable personality

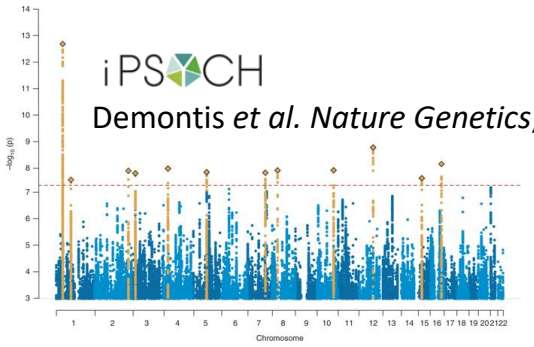
7q11.23 duplication -> Social impairment -> Hyposociable personality

Sanders et al. *Neuron*, 2015



T2D-GENES

Fuchsberger *et al. Nature*, 2016



iPSYCH

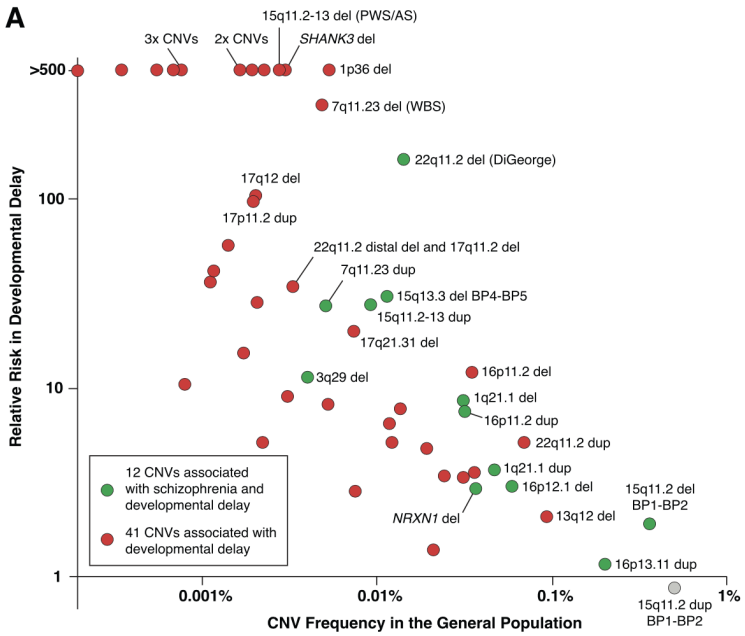
Demontis *et al. Nature Genetics*, 2018

gnomAD

genome aggregation database

biobank uk

Improving the health of future generations



An & Sanders 2017 Biological Psychiatry

- 영국 바이오뱅크 15만명 CNV 분석 (Kendall 2017)
- 발달장애로 진단받지 않은 건강한 중년15만명에게 발견되는 발달장애 CNV -> 어떻게 이해해야하나?

- 데이터 분석 품질?
- Incomplete penetrance?
- 관련 표현형은 무엇인가?
- 임상적 결정은 어떻게 되야?

질병 유전체 연구의 역사

1. Linkage 분석 (80-90년대)
2. Microarray와 GWAS (2000년대)
3. NGS 등장 (2010년 전후)
4. Exome sequencing 위주의 원인 유전자 탐색 연구(gene discovery)
5. Whole Genome Sequencing을 활용한 포괄적인 유전적 조성 연구

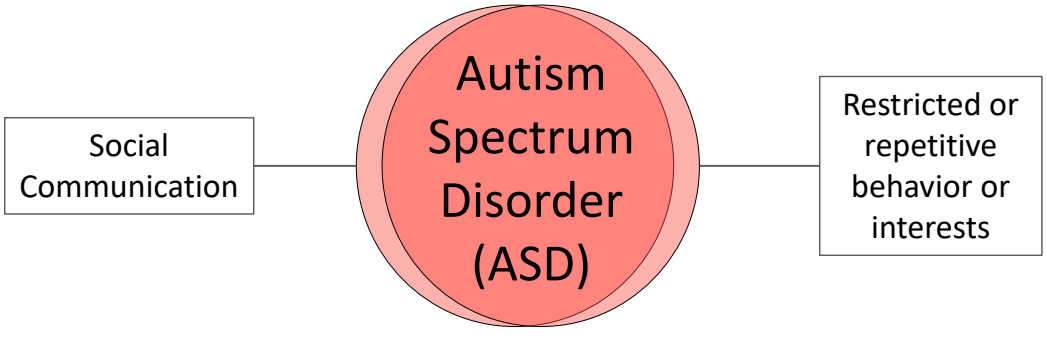
자폐증 유전체 연구 가설에서 전장유전체까지

2021년 전장유전체에 이르기까지...

질병 유전체 연구의 역사

1. Linkage 분석 (80-90년대)
2. Microarray와 GWAS (2000년대)
3. NGS 등장 (2010년 전후)
4. Exome sequencing 위주의 원인 유전자 탐색 연구(gene discovery)
5. Whole Genome Sequencing을 활용한 포괄적인 유전적 조성 연구

자폐성 범주 장애는 신경발달장애의 한 종류

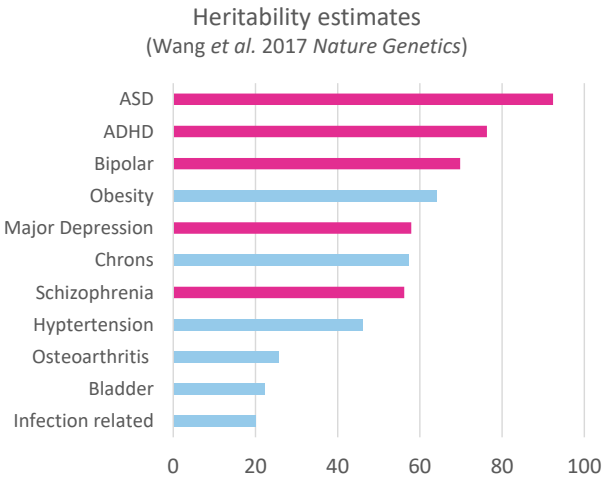


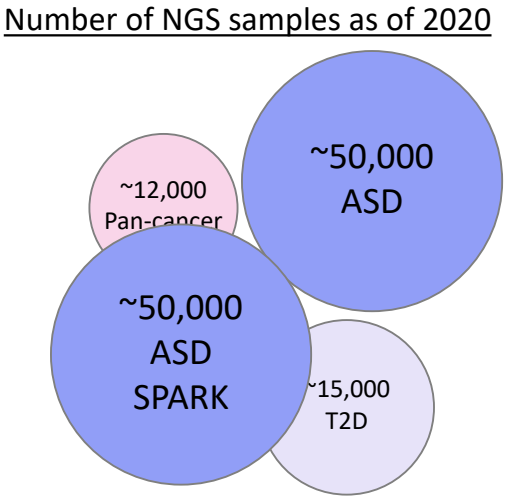
Diagnostic and Statistical Manual of Mental Disorders 5th Edition

- 유병률은 54명 중 1명 (1.8%, 미국 질병통제 예방 센터)
  - 전세계적으로 비슷한 유병률을 보임
  - 남성이 여성보다 흔하게 발견됨

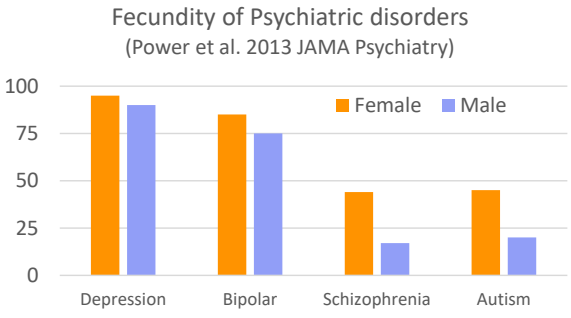
자폐성 범주 장애는 높은 유전력을 보임

유전력 (heritability): 유전적 요소가 형질 및 질환에 기여하는 정도

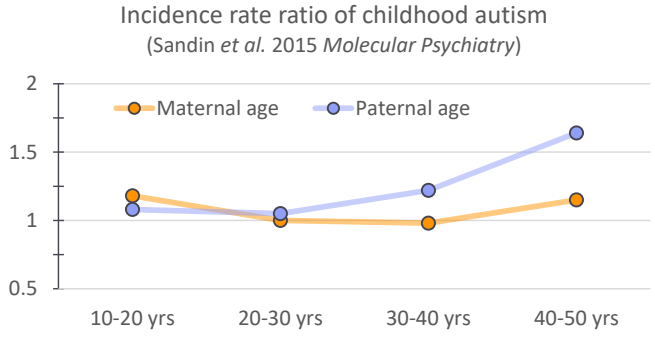




두가지 역학적 증거들, 자폐증 연구의 변곡점

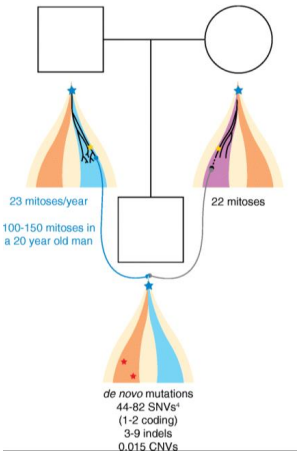


- **자폐성 범주 장애로 진단 받은 이들은 그렇지 않은 이들에 비하여, 낮은 자손률(Fecundity)을 보임.**
- **그럼에도 불구하고 자연선택을 거슬러 인구집단에 자폐성 범주 장애가 지속적으로 나타나는 이유는?**



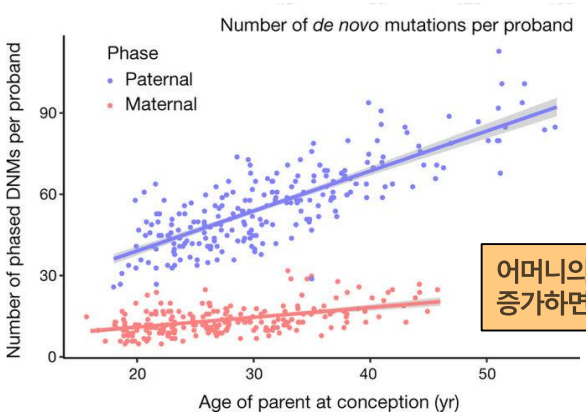
- 부모의 출산 연령이 늦어질 수록, 자녀가 자폐성 범주 장애로 진단 받는 경우가 빈번해짐.
  - Malaspina 2001, Coen 2003, Reichenberg 2006, Peterson 2011, Hultman 2011, Frans 2013, Pederson 2014, McGrath 2014
- **부모의 출산 연령과 자폐성 범주 장애의 발병에는 어떤 연관이 있는가?**

De novo variants (DNV)
부모의 생식세포에서 새롭게 발생한 유전 변이가 자녀에게..



- 부모의 생식세포의 pre-meiotic cell division 과정에서 DNV가 발생함
- 그 생식세포가 수정이 되면, 발생한 DNV는 자녀에게 전달이 됨
- 한 세대에서 보통 70개 DNV가 발생함
- 유전체에서 무작위(random)하게 일어남
- 진화를 이끌어내는 원동력 중 가장 핵심 요소 (Wright 1942)

나이든 부모에게서 *DNV*의 숫자는 증가한다



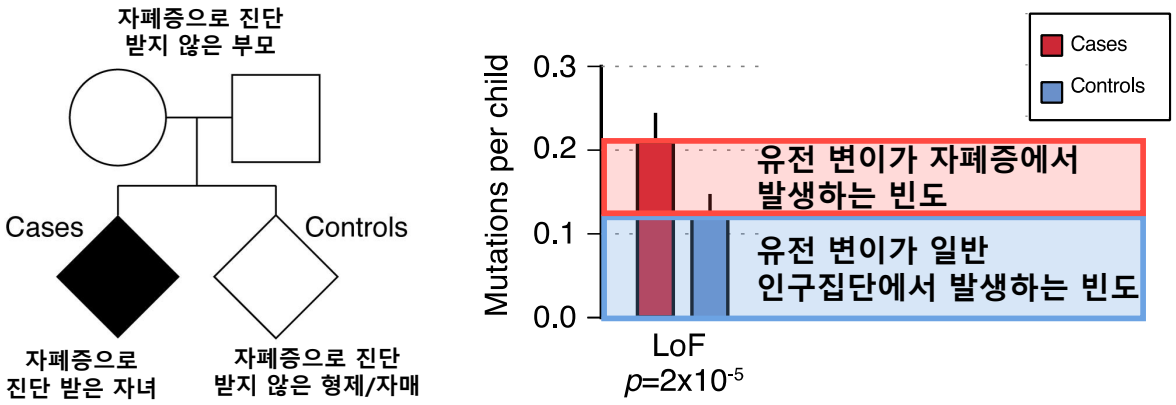
아버지의 출산 연령이 1세 증가하면 DNV는 1.51개 증가함

80%의 DNV는 정자에서 발생함

어머니의 출산 연령이 1세 증가하면 DNV는 0.37개 증가함

From Kong *et al*., 2012 to Jonsson *et al*.,2017

실험 디자인



Sanders *et al.* Nature 2012, Neale *et al. Nature* 2012, O'Roak *et al. Nature* 2012

엑솜 유전체 기술을 이용하여 자폐증 발생과 부모의 출산 연령의 관계를 밝힘

엑솜 유전체 (Exome Sequencing): 단백질 생성 지역에 발생하는 유전 변이를 탐색하는 기술

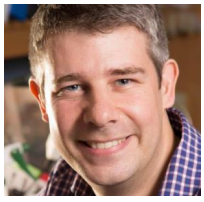
nature
International journal of science

Letter | Published: 04 April 2012

De novo mutations revealed by whole-exome sequencing are strongly associated with autism

Stephan J. Sanders, Michael T. Murtha [...] Matthew W. State

Nature **485**, 237–241 (10 May 2012) | Download Citation



Stephan Sanders

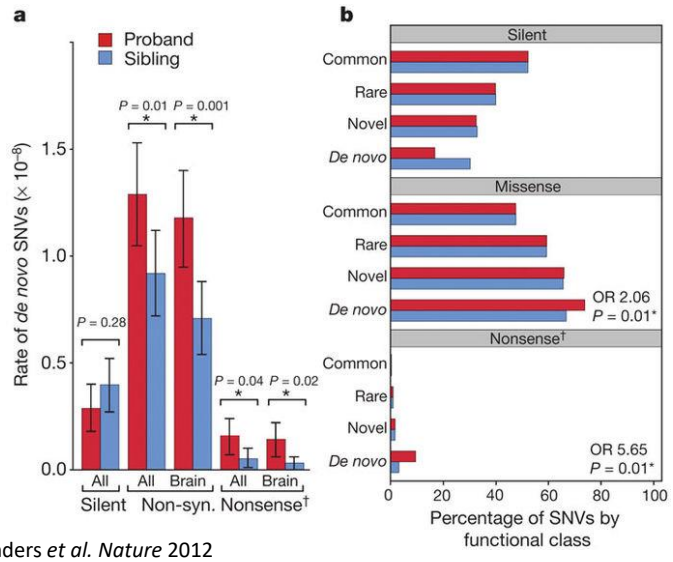
UCSF
(Yale back then)

Matthew State

UCSF
(Yale back then)

"We found that the rate of de novo SNVs indeed increases with paternal age (p=0.008, two-tailed Poisson regression) and that paternal and maternal ages are highly correlated (p<0.0001, two-tailed linear regression)."

2012년 *Nature* 에 게재된 3편의 연구 자폐증은 *de novo* protein-truncating variants에 큰 영향을 받는다



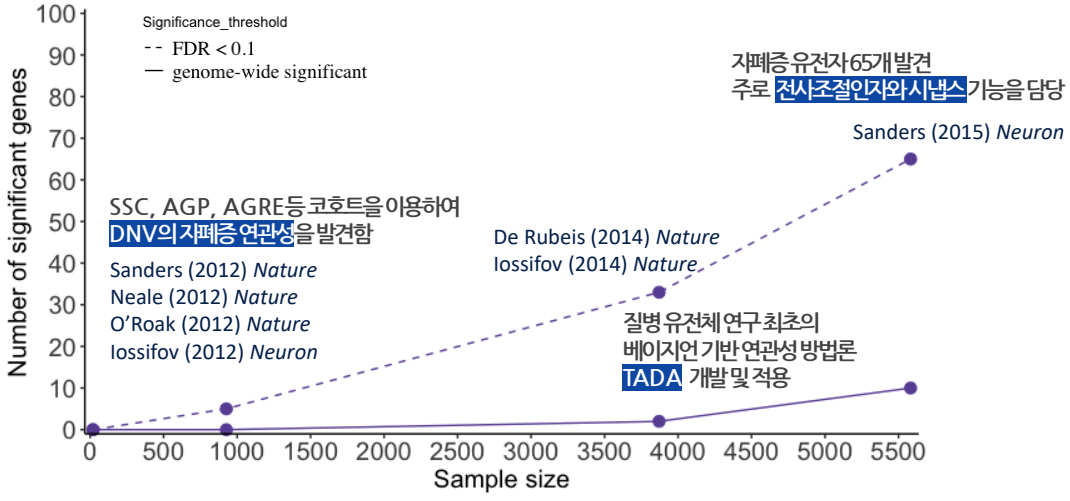
Sanders *et al. Nature* 2012

De novo mutations revealed by whole-exome sequencing are strongly ...
https://www.ncbi.nlm.nih.gov/pubmed/22495306
by SJ Sanders - 2012 - Cited by 1329 - Related articles
2012 Apr 4;485(7397):237-41. doi: 10.1038/nature10945. ... **Sanders** SJ(1), Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, ...

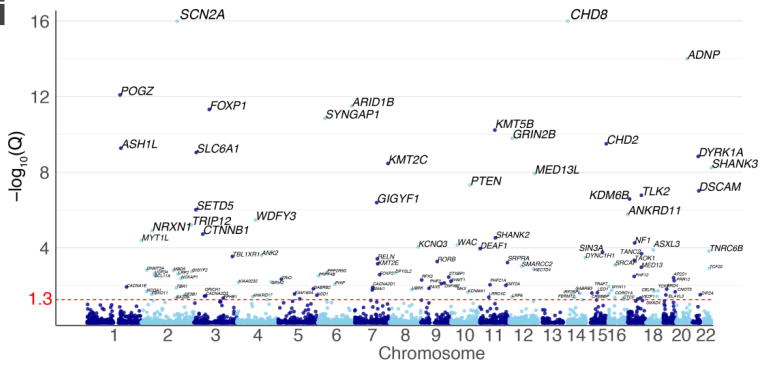
Patterns and rates of exonic de novo mutations in autism ... - NCBI - NIH
https://www.ncbi.nlm.nih.gov/pubmed/22495311
by BM Neale - 2012 - Cited by 1215 - Related articles
2012 Apr 4;485(7397):242-5. doi: 10.1038/nature11011. ... **Neale** BM(1), Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov ...

Sporadic autism exomes reveal a highly interconnected ... - NCBI - NIH
https://www.ncbi.nlm.nih.gov/pubmed/22495309
by BJ O'Roak - 2012 - Cited by 1394 - Related articles
2012 Apr 4;485(7397):246-50. doi: 10.1038/nature10989. ... **O'Roak** BJ(1), Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, ...

ASC는 2012년부터 현재까지 협력 연구를 바탕으로 자폐증에 연관된 유전자를 발견함



2020년 기준, 102개의 자폐증 유전자가 밝혀짐



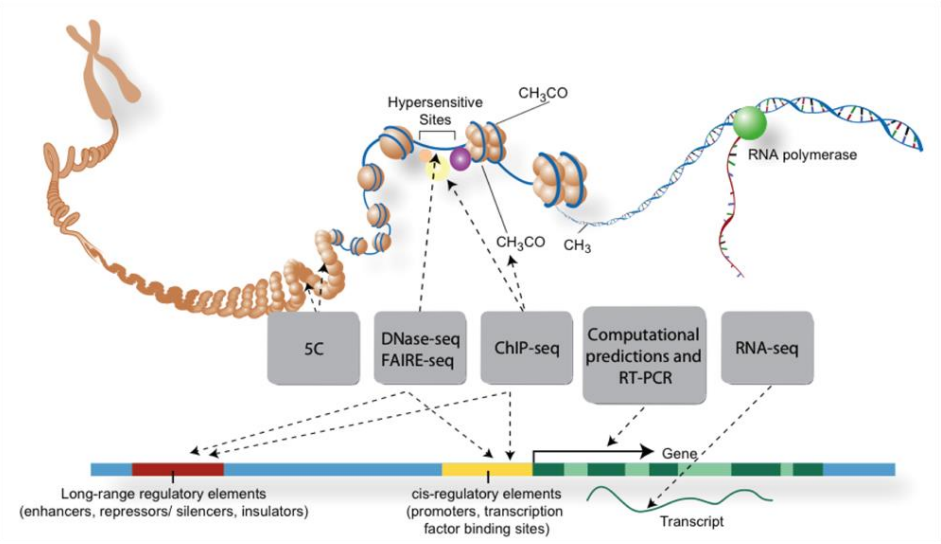
35,584명, 21,219 자폐성 범주 장애 가족
세계 최대 규모의 자폐증 유전체 연구

Satterstorm *et al. Cell*, 2020

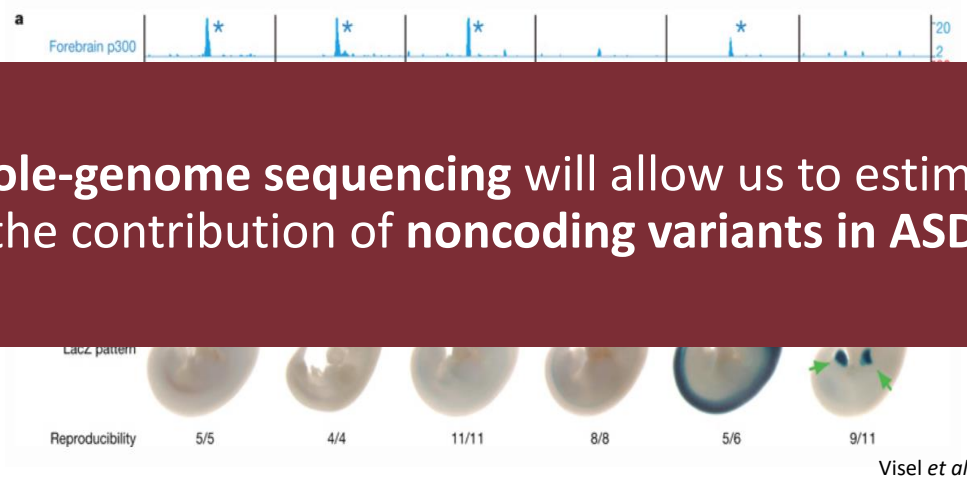
SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

iPSCH

Outside the protein coding regions..



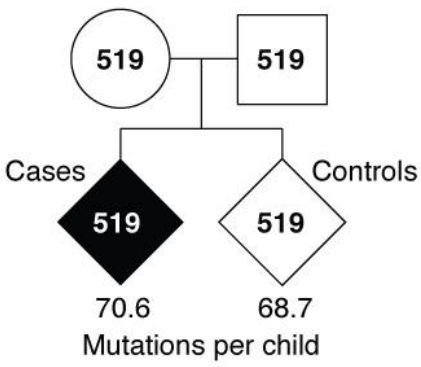
Can we increase the resolution of spatiotemporal analyses?



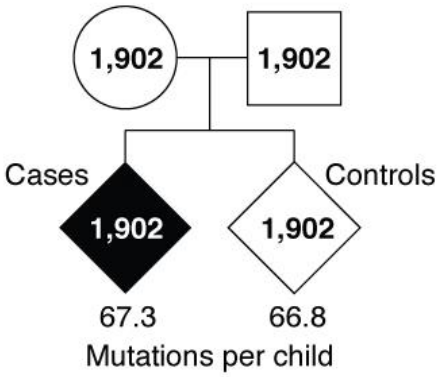
Whole-genome sequencing will allow us to estimate the contribution of **noncoding variants in ASD**

Visel *et al. Nature* 2009

ASC-WGS1 and ASC-WGS2: Largest whole g enome sequencing analysis of ASD families

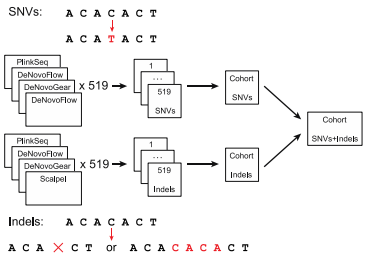


Werling *et al. Nature Genetics* 2018

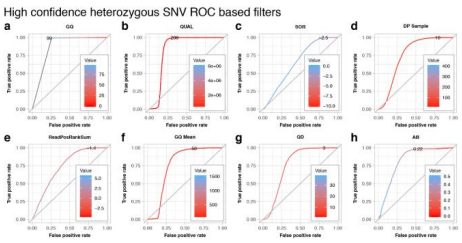


An *et al. Science* 2018

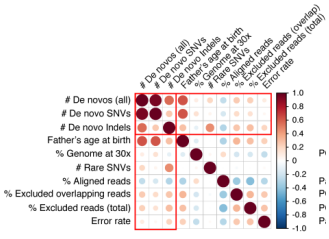
WGS study requires a large-scale computing such as cloud computing



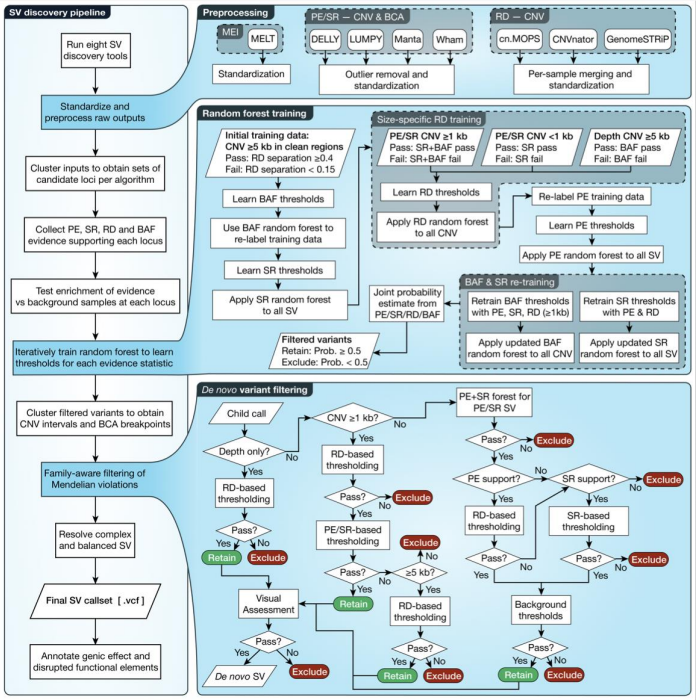
Multiple algorithms on detecting *de novo* variant



ROCube:
High-quality WGS variant filtering
https://github.com/joonan30/rocube



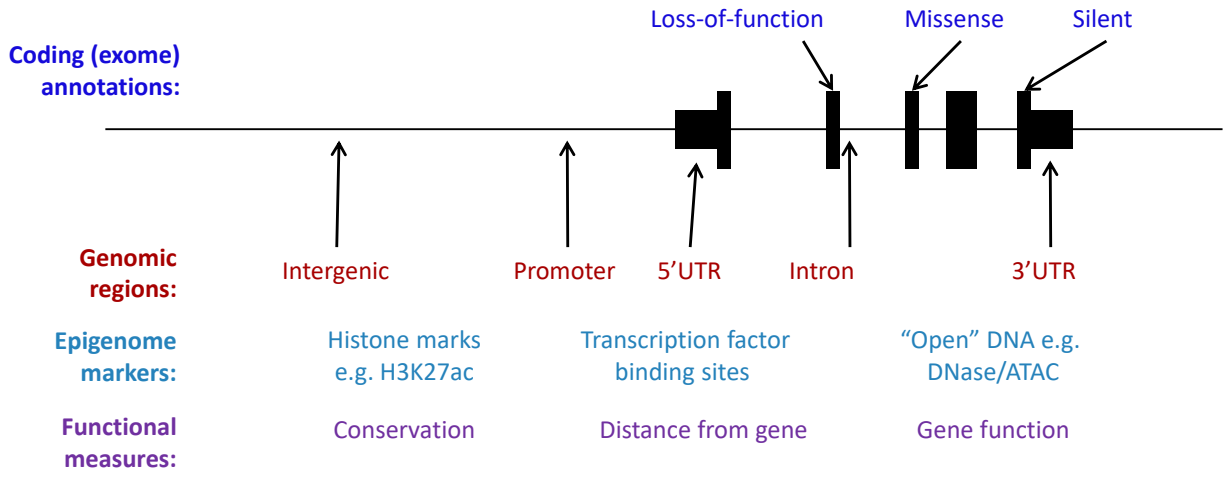
Adjusting confounders
- paternal ages



CNV (copy-number variants) & SV (structural variant) analysis

- CNV/SV explain ~5% of ASD cases in total.
- High rate of false-positive calls

Unlike the exome, there are many potential regions that could mediate ASD risk



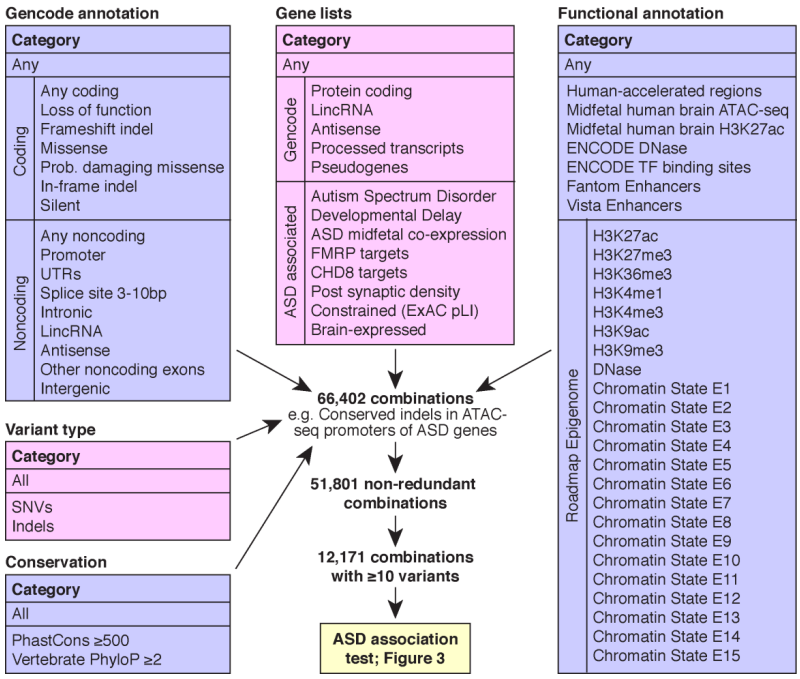
Category-Wide Association Study (CWAS):

Test multiple hypotheses of noncoding annotation simultaneously

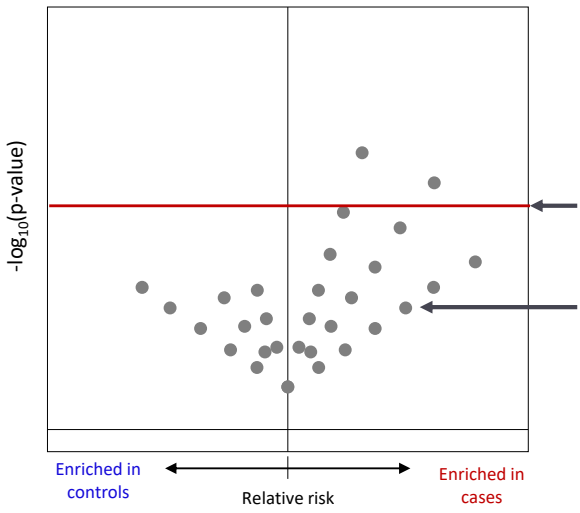
Combine five annotation domains into 51,801 categories

Example:

De novo variants in DNase hypersensitive sites of UTRs of post-synaptic density genes



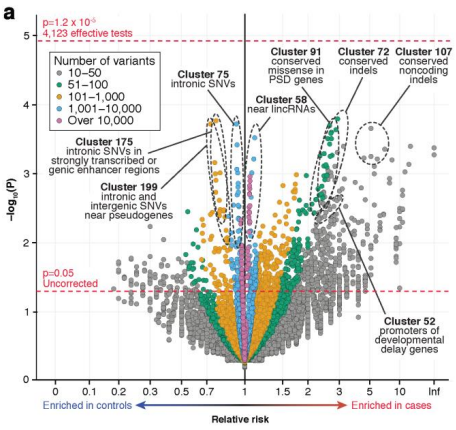
CWAS analysis to compare a burden of *de novo* coding noncoding between cases and controls



Multiple testing correction
Threshold is determined by correlation of 51,801 categories.

A single category:
e.g. *De novo* variants in DNase hypersensitive sites of UTRs of post-synaptic density genes

CWAS analysis of 519 families showed no category reaches significance after correction for multiple testing

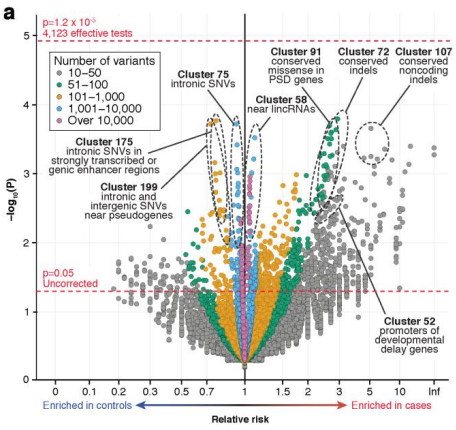


Werling *et al. Nature Genetics* 2018

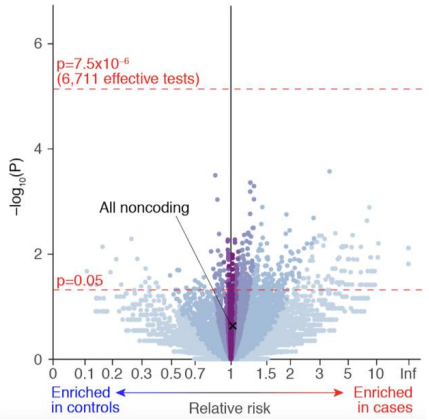
Inconsistency in "candidate no ncoding region" from other WGS studies

- Turner et al. *AJHG* (2016): 40 families of SSC, showing enrichment of *de novo* variants in **50kb interegenic** of fetal CNS DNase I hypersensitive sites (p=0.03; binomial).
- Turner et al. *Cell* (2017): 516 families of SSC, showing enrichment of *de novo* variants in **fetal promoter with TFBS** (p=0.04; binomial).
- Short et al. *Nature* (2017): 6,239 cases with developmental delay, showing **conserved regions of fetal brain DHS peaks** (p=0.04; Fisher).
- Yuen et al. *NPJ Genomic med* (2016): 200 families of MISSNG, showing **conserved 3'UTR** (p=0. 01; Fisher).

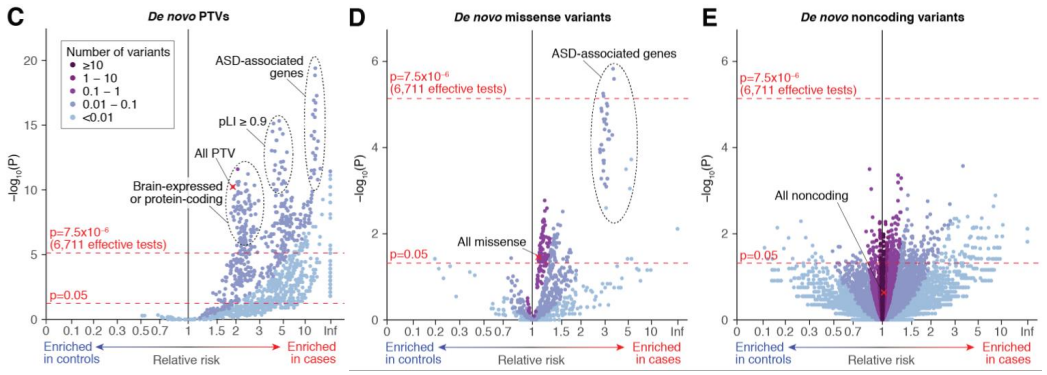
CWAS analysis of 1902 families showed no category reaches significance after correction for multiple testing



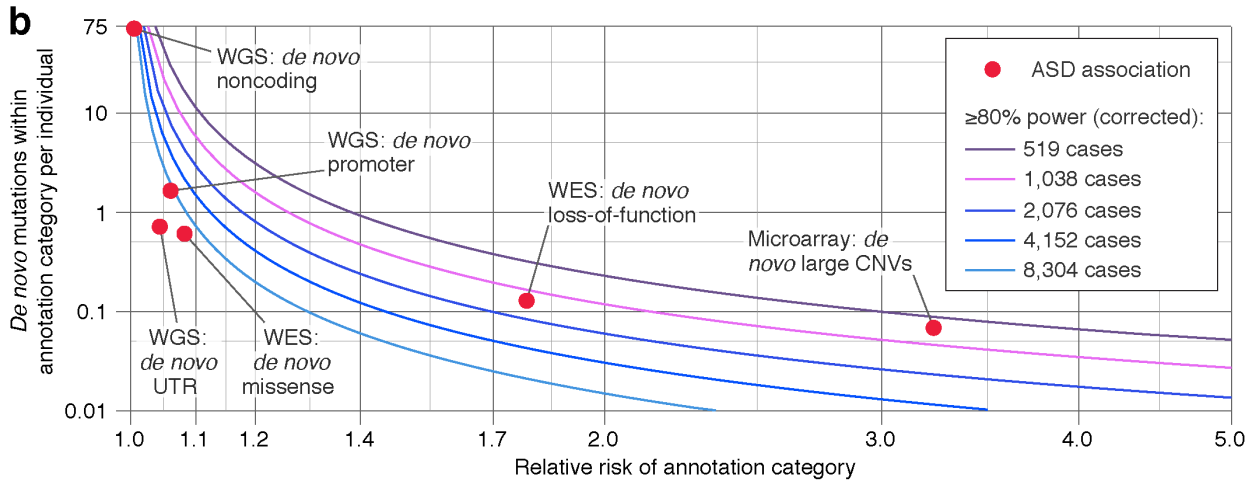
Werling *et al. Nature Genetics* 2018



An *et al. Science* 2018

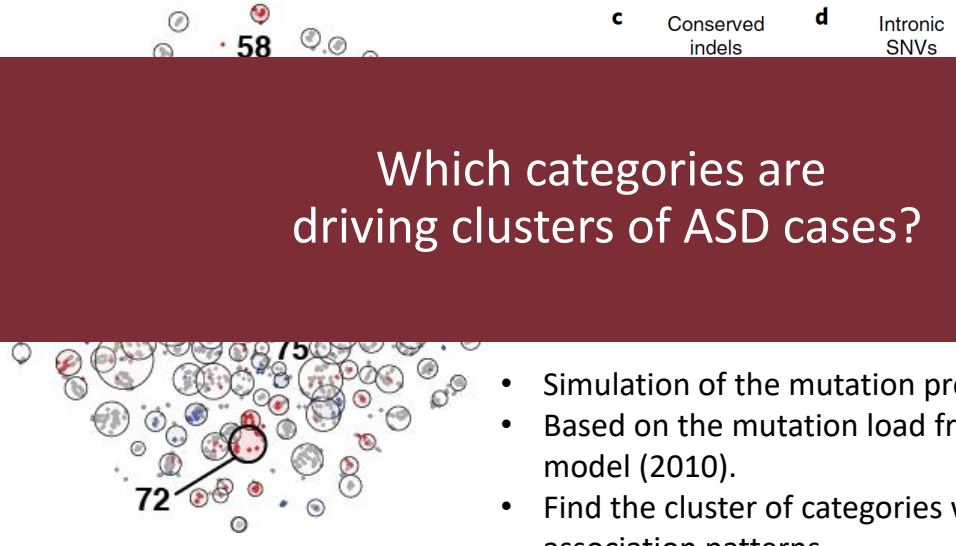


Finding one significant gest for a single noncoding category requires at least 8,000 families



Werling *et al. Nature Genetics* 2018

Decipher the complexity of the regulatory noncoding genome using the high-dimensi onal mutation model

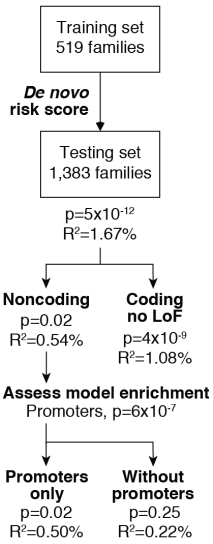


Which categories are driving clusters of ASD cases?

- Simulation of the mutation profiles of DNVs.
- Based on the mutation load from the Lynch model (2010).
- Find the cluster of categories with similar association patterns.

https://github.com/sanderslab/cwas

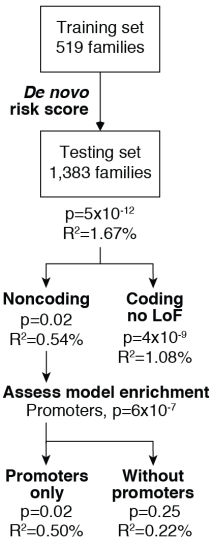
The *de novo* risk score demonstrated nonco ding association driven by promoters

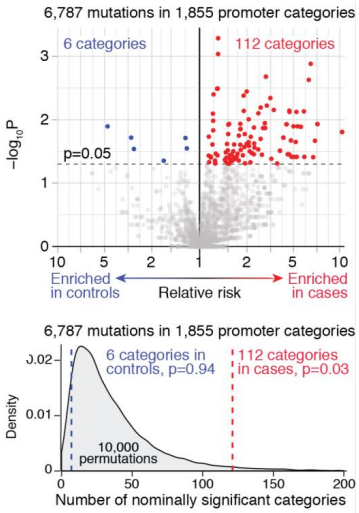


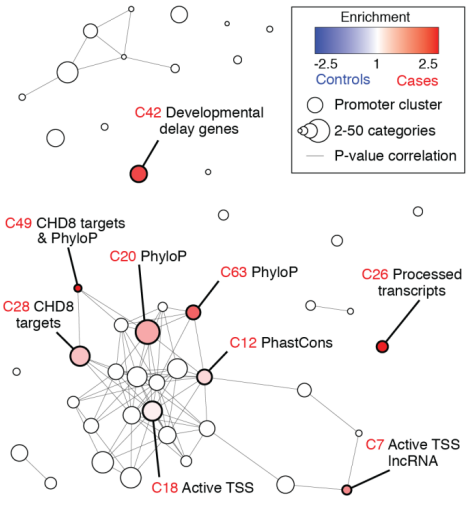
- In coding regions, known ASD association is the result of mutations at **a small number of critical loci**, such as protein truncating variants (PTVs) in ~5% of genes.
- Assuming a similar distribution of ASD risk for *de novo* mutations in the noncoding genome, we limited our analysis to **noncoding annotation categories with few variants**.
- Our aim was to construct a *de novo* risk score to predict ASD status, similar in concept to a polygenic risk score for common variation.

An *et al. Science* 2018

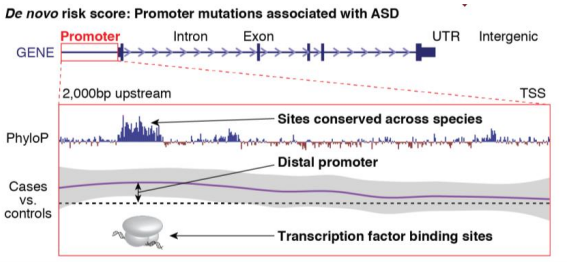
The *de novo* risk score demonstrated nonco ding association driven by promoters



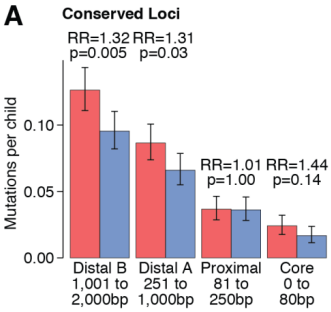


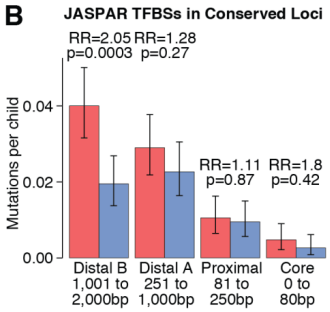


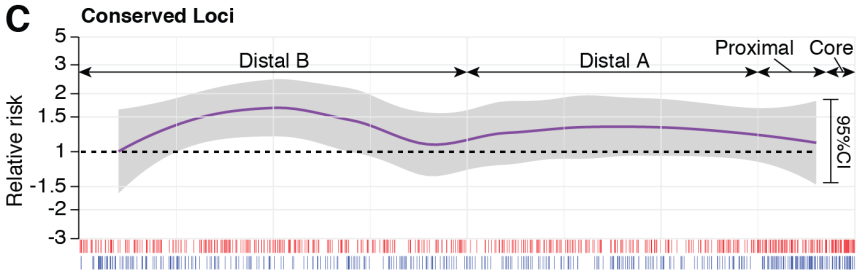
An *et al. Science* 2018



An *et al. Science* 2018







시간을 마무리하며

1. 전장 유전체 연구는 다양한 유전적 조성을 탐색하기 위해 이뤄짐

2. 질병의 유전적 조성에 대한 각 구성요소를 이해할 필요가 있음