Towards reproducible science: lessons from human genetics

Daniel MacArthur
Analytic and Translational Genetics Unit,
Massachusetts General Hospital
Harvard Medical School
Broad Institute of Harvard and MIT
www.macarthurlab.org
Twitter: @dgmacarthur

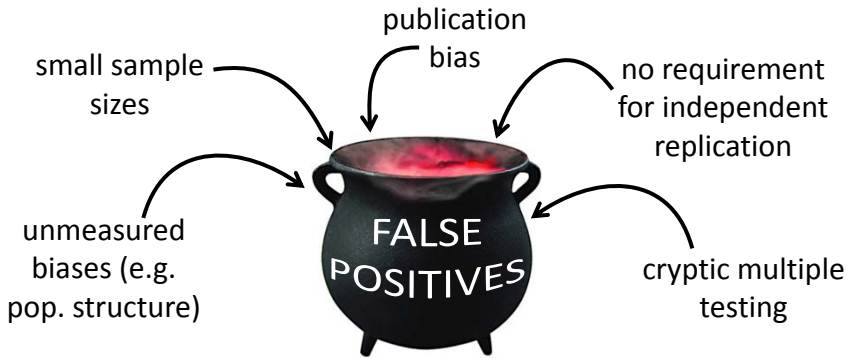
The cartoon history of human complex trait genetics

Prior to 2007: candidate gene era

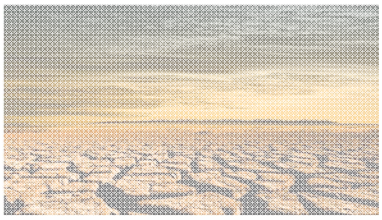
- Targeted testing of variants in "biologically plausible" genes
- Small sample sizes (hundreds)
- 1000s of reported associations, **~none consistently replicated**

What drove false positives in candidate gene studies?

- Most "biologically plausible" variants are **not** actually associated with disease!



The cartoon history of human complex trait genetics



Prior to 2007: candidate gene era

- Targeted testing of variants in "biologically plausible" genes
- Small sample sizes (hundreds)
- 1000s of reported associations, ~none consistently replicated

Post-2007: genome-wide association studies

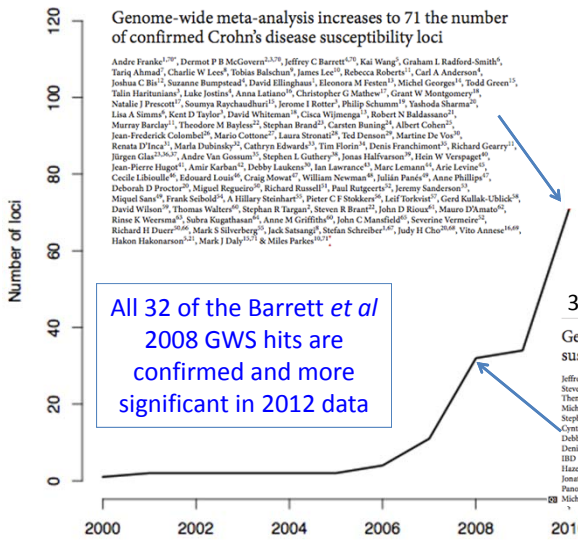
- Unbiased genome-wide scans
- Big samples (1000s-10,000s)
- 1000s of reported associations, **near-perfect replication rates**

GWAS for Crohn's Disease

6,333 cases and 15,056 controls

Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci

Andre Franke[1,70*], Dermot P B McGovern[2,3,70], Jeffrey C Barrett[4,70], Kai Wang[5], Graham L Radford-Smith[6], Tariq Ahmad[7], Charlie W Lees[8], Tobias Balschun[9], James Lee[10], Rebecca Roberts[11], Carl A Anderson[4], Joshua C Bis[12], Suzanne Bumpstead[4], David Ellinghaus[1], Eleonora M Festen[13], Michel Georges[14], Todd Green[15], Talin Haritunians[2], Luke Jostins[4], Anna Latiano[16], Christopher G Mathew[17], Grant W Montgomery[18], Natalie J Prescott[17], Soumya Raychaudhuri[19], Jerome I Rotter[5], Philip Schumm[19], Yashoda Sharma[20], Lisa A Simma[8], Kent D Taylor[5], David Whiteman[18], Cisca Wijmenga[13], Robert N Baldassano[21], Murray Barclay[11], Theodore M Bayless[22], Stephan Brand[23], Carsten Buning[24], Albert Cohen[25], Jean-Frederick Colombel[26], Mario Cottone[27], Laura Stronati[28], Ted Denson[29], Martine De Vos[30], Renata D'Inca[31], Marla Dubinsky[32], Cathryn Edwards[33], Tim Florin[34], Denis Franchimont[35], Richard Gearry[11], Jürgen Glas[23,36,37], Andre Van Gossum[35], Stephen L Guthery[38], Jonas Halfvarson[39], Hein W Verspaget[40], Jean-Pierre Hugot[41], Amir Karban[42], Debby Laukens[30], Ian Lawrance[43], Marc Lemann[44], Arie Levine[45], Cecile Libioulle[46], Edouard Louis[46], Craig Mowat[47], William Newman[48], Julián Panés[49], Anne Phillips[47], Deborah D Proctor[50], Miguel Regueiro[50], Richard Russell[51], Paul Rutgeerts[52], Jeremy Sanderson[53], Miquel Sans[49], Frank Seibold[54], A Hillary Steinhart[55], Pieter C F Stokkers[56], Leif Torkvist[57], Gerd Kullak-Ublick[58], David Wilson[59], Thomas Walters[60], Stephan R Targan[2], Steven R Brant[22], John D Rioux[61], Mauro D'Amato[62], Rinse K Weersma[63], Subra Kugathasan[64], Anne M Griffiths[60], John C Mansfield[65], Severine Vermeire[52], Richard H Duerr[20,66], Mark S Silverberg[32], Jack Satsangi[8], Stefan Schreiber[1,67], Judy H Cho[20,68], Vito Annese[16,69], Hakon Hakonarson[5,21], Mark J Daly[3,5,71] & Miles Parkes[10,71*].



All 32 of the Barrett *et al* 2008 GWS hits are confirmed and more significant in 2012 data

Jostins, Ripke et al Nature 2012
163 distinct confirmed associated loci in IBD (CD and/or UC)
N=75,000 samples

3,230 cases and 4,829 controls

Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease

Jeffrey C Barrett[*1], Sarah Hansoul[2], Dan L Nicolae[3], Judy H Cho[4], Richard H Duerr[5,6], John D Rioux[7,8], Steven R Brant[9,10], Mark S Silverberg[11], Kent D Taylor[12], M Michael Barmada[3], Alain Bitton[13], Themistocles Dassopoulos[6], Lisa Wu Datta[6], Todd Green[6], Anne M Griffiths[14], Emily O Kistner[15], Michael T Murtha[4], Miguel D Regueiro[6], Jerome I Rotter[12], L Philip Schumm[13], A Hillary Steinhart[11], Stephan R Targan[12], Ramnik J Xavier[16], the NIDDK IBD Genetics Consortium[13], Cécile Libioulle[2], Cynthia Sandor[2], Mark Lathrop[17], Jacques Belaiche[18], Olivier Dewit[19], Ivo Gut[17], Simon Heath[17], Debby Laukens[20], Myriam Mni[2], Paul Rutgeerts[21], André Van Gossum[22], Diana Zelenika[17], Denis Franchimont[22], JP Hugot[23], Martine de Vos[24], Severine Vermeire[21], Edouard Louis[18], the Belgian-French IBD Consortium[25], the Wellcome Trust Case Control Consortium[25], Lon R Cardon[3], Carl A Anderson[3], Hazel Drummond[26], Elaine Nimmo[26], Tariq Ahmad[27], Natalie J Prescott[28], Clive M Onnie[29], Sheila A Fisher[26], Jonathan Marchini[27], Jilur Ghori[29], Suzannah Bumpstead[29], Rhian Gwilliam[29], Mark Tremelling[29], Panos Deloukas[29], John Mansfield[30], Derek Jewell[31], Jack Satsangi[26], Christopher G Mathew[26], Miles Parkes[26], Michel Georges[2] & Mark J Daly[8,32]

Mark Daly

Lessons from GWAS

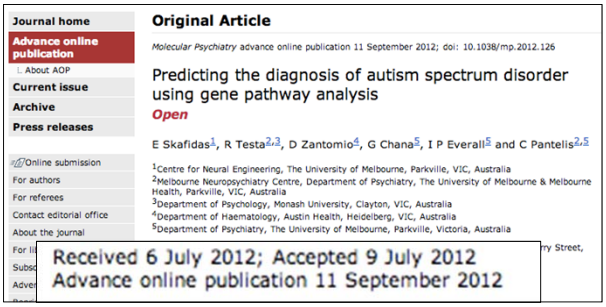
- Perform **well-powered** studies
- Where possible, perform **unbiased scans**, *then* pursue specific results
- Leave little room for *post-hoccery*
  - Establish **consensus on statistical approaches** and *P* value thresholds
  - Require **raw data and critical software** be made available to other researchers
- Work in consortia: access **more samples** and get **internal peer review**

New challenges in the genome era

- We no longer start with hypotheses: we *sift results* from large, noisy data sets
- Challenge: any process extracting "interesting" results will also enrich for biases and artifacts
- Avoiding false positives requires:
  - **deep understanding of errors** (manual spot-checking, visualization)
  - **validation** using independent technology

MacArthur 2012 *Nature* **487:**427-428

Science needs nimble discussion


Journal home
Advance online publication
About AOP
Current issue
Archive
Press releases
Online submission
For authors
For referees
Contact editorial office
About the journal

Original Article

Molecular Psychiatry advance online publication 11 September 2012; doi: 10.1038/mp.2012.126

Predicting the diagnosis of autism spectrum disorder using gene pathway analysis

Open

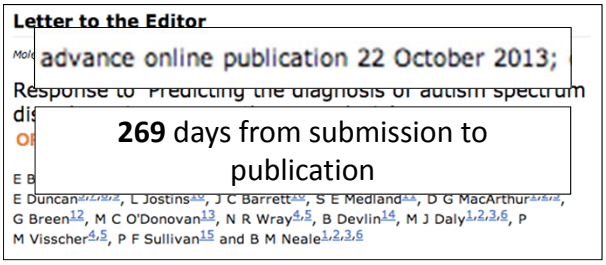
E Skafidas[1], R Testa[2,3], D Zantomio[4], G Chana[5], I P Everall[5] and C Pantelis[2,5]

[1]Centre for Neural Engineering, The University of Melbourne, Parkville, VIC, Australia
[2]Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne & Melbourne Health, Parkville, VIC, Australia
[3]Department of Psychology, Monash University, Clayton, VIC, Australia
[4]Department of Haematology, Austin Health, Heidelberg, VIC, Australia
[5]Department of Psychiatry, The University of Melbourne, Parkville, Victoria, Australia

Received 6 July 2012; Accepted 9 July 2012
Advance online publication 11 September 2012

bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

PubPeer
The online journal club


Letter to the Editor

advance online publication 22 October 2013;

Response to "Predicting the diagnosis of autism spectrum dis

269 days from submission to publication

E Duncan[2,7,8,9], L Jostins[10], J C Barrett[10], S E Medland[11], D G MacArthur[2,12], G Breen[12], M C O'Donovan[13], N R Wray[4,5], B Devlin[14], M J Daly[1,2,3,6], P M Visscher[4,5], P F Sullivan[15] and B M Neale[1,2,3,6]

PILOT
PubMed Commons
A forum for scientific discourse

Specific recommendations

- Establish **standards workshops** with domain experts and statisticians to define consensus protocols for widely used techniques
- Fund creation of software for **intuitive visualization** of complex data sets
- Mandate **availability of raw data and software** to enable replication and reuse
- Facilitate **large-scale collaborative science** across more domains of science
- Provide career incentives for **open pre- and post-publication peer review**