Genomic prediction: an *extremely* brief overview

Introduction

Let's review the basic approach we use in genome-wide association mapping.

- We measure both the phenotype, y_i, of individual i and its genotype at a large number of loci, where x_{ij} is the individual's genotype at locus j.
- We regress phenotype on genotype one locus at a time, using a random effect to correct for phenotypic similarities that reflect relatedness rather than similarity in genotype.

$$y_i^{(k)} = x_{ij}\beta_j + \phi^{(k)} + \epsilon_i \quad .$$

Keep in mind this is a highly idealized schematic of how GWAS analyses are actually done.[1] If you want to do GWAS for real, you should take a look at GEMMA (`http://www.xzlab.org/software.html`) or TASSEL (https://www.maizegenetics.net/tassel). One important way in which what I've presented is a simplification is that in a real GWAS analysis, you'd estimate the effects of every locus simultaneously, which raises an interesting problem.

In a typical GWAS analysis[2], you will have measured the phenotype of a few thousand individuals, but you will have genotyped those individuals at several hundred thousand loci. Lango Allen et al. [3], for example, report results from a large analysis of height variation in humans, 183,727 individuals genotyped at 2,834,208 loci. What's the problem here?

There are more predictors (loci) than observations (individual phenotypes). If you remember some basic algebra, you'll remember that you can't solve a set of linear equations unless you have the same number of equations as unknowns. For example, you can't solve a set of three equations that has five unknowns. There's a similar phenomenon in statistics

[1] Remember, also, that in analyses of human disease, a case-control approach is often used rather than the regression approach I've been focusing on.

[2] In humans at least

when we're fitting a linear regression. In statistics we don't "solve" an equation. We find the best fit in a regression, and we can do so in a reasonable way so long as the number of observations exceeds the number of variables included in our regression. To put a little mathematical notation to it, if n is the number of observations and p is the number of regression parameters we hope to estimate, life is good (meaning that we can estimate the regression parameters) so long as $n > p$.[3] The typical situation we encounter in GWAS is that $n < p$, which means we have to be really sneaky. Essentially what we do is that we find a way for the data to tell us that a lot of the parameters don't matter and we fit a regression only to the ones that do, *and* we set things up so that the remaining number of parameters is less than n. If that all sounds a little hoky, trust me it isn't. There are good ways to do it and good statistical justification for doing it[4], but the mathematics behind it gets pretty hairy, which is why you want to use GEMMA or TASSEL for a real GWAS. We'll ignore this part of the challenge associated with GWAS and focus on another one: complex traits often are influenced by a very large nubmer of loci. That is, after all, why we started studying quantitative genetics in the first place.

Genetics of complex traits

Let's return to that Lango Allen et al. [3] GWAS on height in humans. They identified at least 180 loci associated with differences in height. Moreover, many of the variants are closely associated with genes that are part of previously identified pathways, e.g., Hedgehog signaling,[5] or that were previously identified as being involved in skeletal growth defects. A more recent study by Wood et al. [4] synthesized results from 79 studies involving 253,288 individuals and identified 697 variants that were clustered into 423 loci affecting differences in height.[6] Think about what that means. If you know my genotype at only one of those 697 variants, you know next to nothing about how tall I am. But what if you knew my genotype for all of those variants? Then you should be able to do better.

The basic idea is fairly simple. When you do a full GWAS and estimate the effects at every locus simultaneously, you are essentially performing a multiple regression of phenotype on all of the loci you've scored simultaenously instead of looking at them one at a time. In

[3] And the more that n exceeds p the better, the more accurate our estimates of the regression parameters will be.

[4] And biological justification for doing it in GWAS.

[5] "The Hedgehog signaling pathway is a signaling pathway that transmits information to embryonic cells required for proper cell differentiation." `https://en.wikipedia.org/wiki/Hedgehog_signaling_pathway`, accessed 14 August 2021.

[6] It's worth noting that even this is likely to be an underestimate of the number of loci associated with height variation in humans because all of the individuals included in the analysis were of European ancestry.

equation-speak,

$$y_i^{(k)} = \sum_j x_{ij}\beta_j + \phi^{(k)} + \epsilon_i \quad .$$

Now think a bit more about what that equation means. The $\phi^{(k)}$ and ϵ_i terms represent random variation, in the first case variation that is correlated among individuals depending on how closely related they are and in the second case variation that is purely random. The term $\sum_j x_{ij}\beta_j$ reflects systematic effects associated with the genotype of individual i. In other words, if we know individual i's genotype, i.e., if we know x_{ij} we can predict what phenotype it will have, namely $\mu_i = \sum_j x_{ij}\beta_j$. Although we know there will be uncertainty associated with this prediction, μ_i is our best guess of the phenotype for that individual, i.e., our genomic prediction or polygenic score. In the case of height in human beings, it turns out that the loci identified in Wood et al. [4] account for about 16 percent of variation in height.[7] If we don't have too many groups, we could refine our estimate a bit further by adding in the group-specific estimate, $\phi^{(k)}$. Of course when we do so, our prediction is no longer a *genomic* predictiion, *per se*. It's a genomic prediction enhanced by non-genetic group information.

A toy example

To make all of this more concrete, we'll explore a toy example using the highly simplified one locus at a time approach to GWAS with a highly simplified example of the multiple regression approach to GWAS. You'll find the `R` code used to create and analyze this simple example at `http://darwin.eeb.uconn.edu/eeb348-resources/genomic-prediction.R` If you `source("genomic-prediction.R")` it will

- Generate a random dataset with 100 individuals and 20 loci, 5 of which influence the phenotype. The effect of one "1" allele at locus 1 is 1, at locus 2 -1, at locus 3 0.5, at locus 4 -0.5, and at locus 5 0.25. The standard deviation of the phenotype around the predicted mean is 0.1.
- Run the locus-by-locus regression for each locus and store the results (mean and 95% credible interval) in "results.csv" and retain the results in `results`. `results` is sorted in by the magnitude of the posterior mean, so that loci with the largest estimated effect occur at the top and loci with the smallest effect occur at the bottom.
- Run the multiple regression and store the results in `fit`.

[7] In Europe the heritability of height at age 20 is about 80 percent [2].

If you look at the code, you'll see that I use `stan_lm()` rather than using `stan_lmer()`. That's because I simulate the data without family structure, so there's no need to include the family random effect.

Table 1 shows results of the locus by locus analysis from my run of `genomic-prediction.R`. Your results will vary a bit, both because the Monte Carlo analysis of the data won't be precisely the same every time and because the data you generate in the simulation will be a bit different from the data I generated for this analysis.

	mean	2.5%	97.5%
locus_1	1.01	0.76	1.27
locus_2	-0.97	-1.12	-0.82
locus_3	0.63	0.34	0.91
locus_4	-0.42	-0.71	-0.13
locus_5	0.42	0.11	0.72
locus_14	0.23	-0.07	0.52
locus_8	-0.17	-0.47	0.13
locus_6	-0.14	-0.47	0.18
locus_20	-0.13	-0.46	0.18
locus_19	-0.12	-0.44	0.20
locus_10	0.11	-0.22	0.44
locus_18	0.09	-0.22	0.40
locus_9	-0.09	-0.39	0.22
locus_12	-0.09	-0.41	0.26
locus_15	-0.07	-0.37	0.23
locus_11	-0.07	-0.40	0.27
locus_7	-0.04	-0.34	0.27
locus_16	-0.04	-0.35	0.28
locus_13	0.03	-0.29	0.36
locus_17	0.01	-0.32	0.35

Table 1: Sample results for locus by locus analysis of genetic associations using `genomic-prediction.R`

For this simulated data set the 5 loci with the largest estimated effect are the 5 loci for which I specified an effect, but that isn't always the case. You may well find that a locus that didn't have a specified effect is one of the top 5. Furthermore, even though this time the locus by locus approach picked out the right top 5 loci, `locus_14` has an effect that is quite large, the the credible interval associated with it only overlaps 0, by a bit. It would

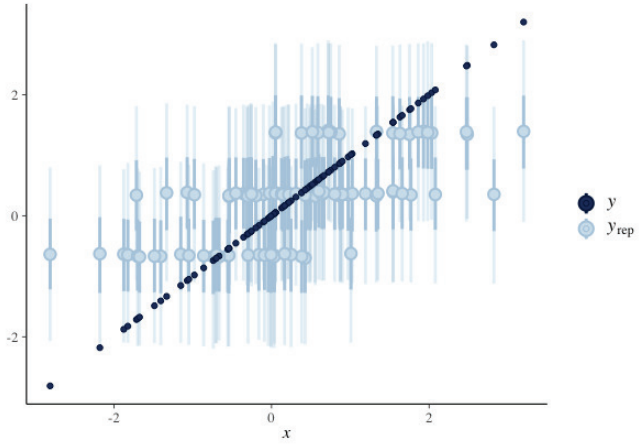

Figure 1: Posterior prediction for locus 1. Small black dots indicate observed phenotypes. Large gray dots indicate the corresponding posterior prediction. The darker gray lines show the location of 50% credible intervals, and the lighter gray lines show the location of 90% credible intervals.

be tempting to think that it is worth investigating further. It wouldn't be as tempting to look at any of the other loci, but a lot of them have fairly large effects given that the entire observed range of phenotypes in this data set is -2.8 to 3.2.

Since you've already run `source("genomic-predicition.R")`, you can now run `plot_posterior_predict(fit_list[[1]])` to plot phenotype predictions at locus 1 versus the observed phenotype for each individual (Figure 1). You can see that there is a relationship between an individual's genotype at this locus, but you can also see that it's pretty weak. Compare that to the relationship between phenotype and genotype at `locus_17`, though, and you'll see that the genotype at locus 17 doesn't give you any information about phenotype while the genotype at locus 1 at least gives you a little (Figure 2).

What about the multiple regression approach? First, take a look at the estimated effects (Table 2). Not only does this approach pick out the right loci, the first five, none of the other loci have particularly large estimated effects. The largest, `locus_6` and `locus_8` are both only 0.07, as opposed to 0.14 in the locus by locus analysis. It would take much more extensive simulation to demonstrate the advantage empirically, but it is clear from first principles that multiple regression analyses will be more reliable than locus by locus analyses because a multiple regression analysis take account of random associations among loci.

Now compare phenotype predictions from the first five loci when each effect is taken individually from Table 1 with the prediction derived from the multple regression approach (Fig-

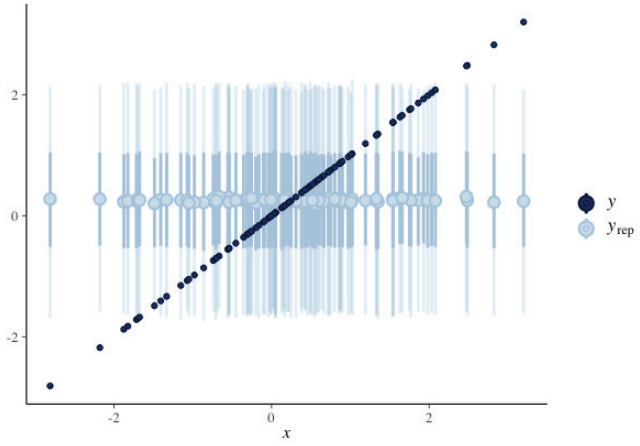

Figure 2: Posterior prediction for locus 17.

ure 3).[8] As you can see, using all 5 loci together to predict phenotypes does a very good job of recovering them.

CAUTION: Danger ahead!

This all seems very promising, but a word of caution is in order. Several papers, including one by Peter Turchin, have suggested that there is strong evidence for selection on polygenic scores associated with height using the same data set of 253,288 individuals I referred to earlier (references in Berg et al. [1]). Specifically, these studies suggested (a) that there is a cline in polygenic scores from south-to-north in Europe (taller phenotypes predicted in the north) and (b) that the cline is too steep to be accounted for by neutrual evolution. Berg et al. [1] re-examined these claims using new data available from the UK Biobank (`https://www.bdi.ox.ac.uk/research/uk-biobank`), which includes a host of information on individual phenotypes as well as genome-wide genotypes for the 500,000 individuals included in the sample.[9] They failed to detect evidence of a cline in polygenic scores in their analysis (Figure 4).

In thinking about this result, it's important to understand that Berg et al. [1] did something a bit different from what we did, but it's exactly what you'd want to do if polygenic

[8] Use `compare_posterior_predictions(fit_list[1:5], fit)` to produce a similar plot using your results.

[9] Although all of the samples are from the UK, one of the data sets Berg et al. [1] studied included individuals of European, but non-UK, ancestry.

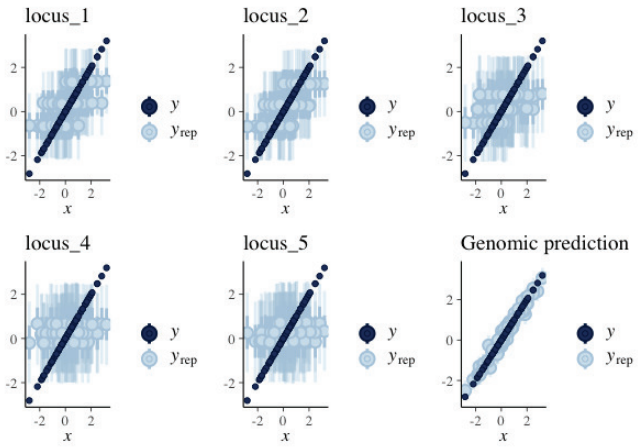

Figure 3: Posterior predictions for loci 1-5 from locus by locus regressions and the posterior preduction for the multiple regression.

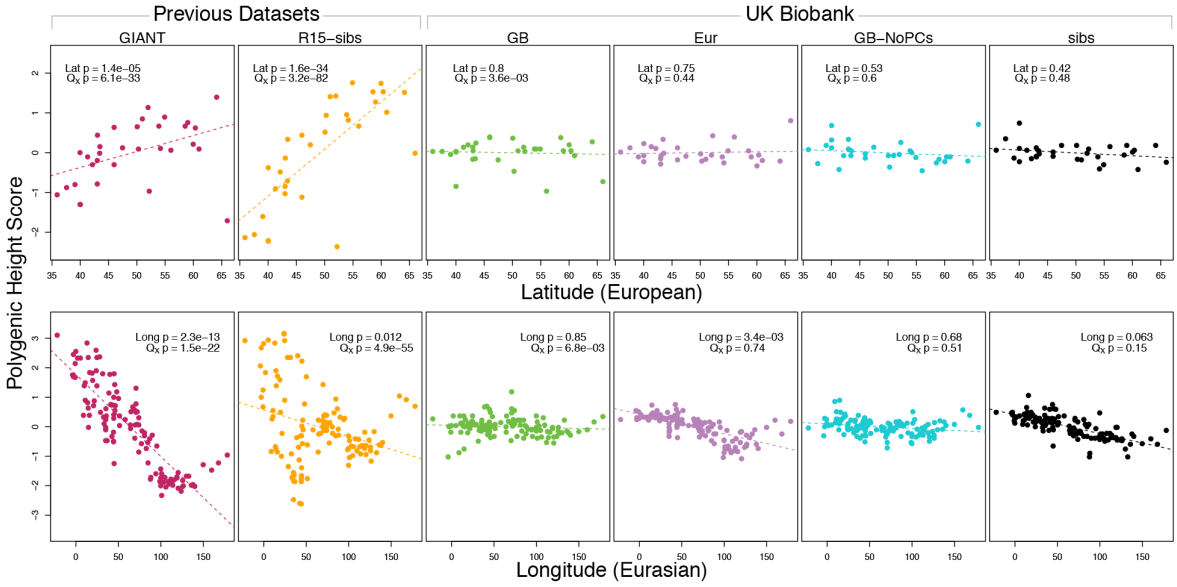

Figure 4: Polygenic score as a function of latitude and longitude for several different GWAS data sets.

	mean	2.5%	97.5%
locus_2	-0.96	-1.07	-0.86
locus_1	0.92	0.84	1.00
locus_3	0.48	0.39	0.56
locus_4	-0.44	-0.52	-0.36
locus_5	0.23	0.15	0.32
locus_6	0.07	-0.03	0.16
locus_8	-0.07	-0.15	0.02
locus_11	0.06	-0.03	0.15
locus_14	-0.06	-0.14	0.03
locus_20	-0.06	-0.14	0.03
locus_12	0.05	-0.04	0.14
locus_7	-0.04	-0.12	0.04
locus_18	-0.03	-0.10	0.06
locus_16	0.02	-0.07	0.10
locus_10	0.02	-0.07	0.11
locus_19	-0.02	-0.10	0.06
locus_17	0.01	-0.08	0.09
locus_9	0.01	-0.08	0.09
locus_15	0.00	-0.07	0.08
locus_13	-0.00	-0.09	0.09

Table 2: Results from multiple regression analysis of simulated data.

scores worked. They estiamted polygenic scores from each of the data sets identified in the figure. Then they used those scores to estimate polygenic scores for a new set of samples derived from the 1000 Genomes and Human Origins projects.[10] Think about it. A polygenic score doesn't do us a whole lot of good if all it lets us do is to predict (with uncertainty) a phenotype we already know. The hope is that we can use the polygenic score to predict phenotypes for individuals when we know their genotype but not their phenotype. What this result shows is that extrapolation of a regression beyond the range of variation included in the sample from which it was estimated can be very problematic.

[10]See Berg et al. [1] for details.

References

[1] Jeremy J Berg, Arbel Harpak, Nicholas Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan A Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in UK Biobank. *bioRxiv*, pages 1–54, December 2018.

[2] Aline Jelenkovic, Reijo Sund, Yoon-Mi Hur, Yoshie Yokoyama, Jacob v B. Hjelmborg, Sören Möller, Chika Honda, Patrik K. E. Magnusson, Nancy L. Pedersen, Syuichi Ooki, Sari Aaltonen, Maria A. Stazi, Corrado Fagnani, Cristina D'Ippolito, Duarte L. Freitas, José Antonio Maia, Fuling Ji, Feng Ning, Zengchang Pang, Esther Rebato, Andreas Busjahn, Christian Kandler, Kimberly J. Saudino, Kerry L. Jang, Wendy Cozen, Amie E. Hwang, Thomas M. Mack, Wenjing Gao, Canqing Yu, Liming Li, Robin P. Corley, Brooke M. Huibregtse, Catherine A. Derom, Robert F. Vlietinck, Ruth J. F. Loos, Kauko Heikkilä, Jane Wardle, Clare H. Llewellyn, Abigail Fisher, Tom A. McAdams, Thalia C. Eley, Alice M. Gregory, Mingguang He, Xiaohu Ding, Morten Bjerregaard-Andersen, Henning Beck-Nielsen, Morten Sodemann, Adam D. Tarnoki, David L. Tarnoki, Ariel Knafo-Noam, David Mankuta, Lior Abramson, S. Alexandra Burt, Kelly L. Klump, Judy L. Silberg, Lindon J. Eaves, Hermine H. Maes, Robert F. Krueger, Matt McGue, Shandell Pahlen, Margaret Gatz, David A. Butler, Meike Bartels, Toos C. E. M. van Beijsterveldt, Jeffrey M. Craig, Richard Saffery, Lise Dubois, Michel Boivin, Mara Brendgen, Ginette Dionne, Frank Vitaro, Nicholas G. Martin, Sarah E. Medland, Grant W. Montgomery, Gary E. Swan, Ruth Krasnow, Per Tynelius, Paul Lichtenstein, Claire M. A. Haworth, Robert Plomin, Gombojav Bayasgalan, Danshiitsoodol Narandalai, K. Paige Harden, Elliot M. Tucker-Drob, Timothy Spector, Massimo Mangino, Genevieve Lachance, Laura A. Baker, Catherine Tuvblad, Glen E. Duncan, Dedra Buchwald, Gonneke Willemsen, Axel Skytthe, Kirsten O. Kyvik, Kaare Christensen, Sevgi Y. Öncel, Fazil Aliev, Finn Rasmussen, Jack H. Goldberg, Thorkild I. A. Sørensen, et al. Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Scientific Reports*, 6:28496, 2016.

[3] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi, Tomi Pastinen, Liming Liang, Iris M. Heid, Jian'an Luan, Gudmar Thorleifsson, Thomas W. Winkler, Michael E. Goddard, Ken Sin Lo, Cameron

Palmer, Tsegaselassie Workalemahu, Yurii S. Aulchenko, Åsa Johansson, M. Carola Zillikens, Mary F. Feitosa, Tõnu Esko, Toby Johnson, Shamika Ketkar, Peter Kraft, Massimo Mangino, Inga Prokopenko, Devin Absher, Eva Albrecht, Florian Ernst, Nicole L. Glazer, Caroline Hayward, Jouke-Jan Hottenga, Kevin B. Jacobs, Joshua W. Knowles, Zoltán Kutalik, Keri L. Monda, Ozren Polasek, Michael Preuss, Nigel W. Rayner, Neil R. Robertson, Valgerdur Steinthorsdottir, Jonathan P. Tyrer, Benjamin F. Voight, Fredrik Wiklund, Jianfeng Xu, Jing Hua Zhao, Dale R. Nyholt, Niina Pellikka, Markus Perola, John R. B. Perry, Ida Surakka, Mari-Liis Tammesoo, Elizabeth L. Altmaier, Najaf Amin, Thor Aspelund, Tushar Bhangale, Gabrielle Boucher, Daniel I. Chasman, Constance Chen, Lachlan Coin, Matthew N. Cooper, Anna L. Dixon, Quince Gibson, Elin Grundberg, Ke Hao, M. Juhani Junttila, Lee M. Kaplan, Johannes Kettunen, Inke R. König, Tony Kwan, Robert W. Lawrence, Douglas F. Levinson, Mattias Lorentzon, Barbara McKnight, Andrew P. Morris, Martina Müller, Julius Suh Ngwa, Shaun Purcell, Suzanne Rafelt, Rany M. Salem, Erika Salvi, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832, 2010.

[4] Andrew R. Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H. Pers, Stefan Gustafsson, Audrey Y. Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, Najaf Amin, Martin L. Buchkovich, Damien C. Croteau-Chonka, Felix R. Day, Yanan Duan, Tove Fall, Rudolf Fehrmann, Teresa Ferreira, Anne U. Jackson, Juha Karjalainen, Ken Sin Lo, Adam E. Locke, Reedik Mägi, Evelin Mihailov, Eleonora Porcu, Joshua C. Randall, André Scherag, Anna A. E. Vinkhuyzen, Harm-Jan Westra, Thomas W. Winkler, Tsegaselassie Workalemahu, Jing Hua Zhao, Devin Absher, Eva Albrecht, Denise Anderson, Jeffrey Baron, Marian Beekman, Ayse Demirkan, Georg B. Ehret, Bjarke Feenstra, Mary F. Feitosa, Krista Fischer, Ross M. Fraser, Anuj Goel, Jian Gong, Anne E. Justice, Stavroula Kanoni, Marcus E. Kleber, Kati Kristiansson, Unhee Lim, Vaneet Lotay, Julian C. Lui, Massimo Mangino, Irene Mateo Leach, Carolina Medina-Gomez, Michael A. Nalls, Dale R. Nyholt, Cameron D. Palmer, Dorota Pasko, Sonali Pechlivanis, Inga Prokopenko, Janina S. Ried, Stephan Ripke, Dmitry Shungin, Alena Stancáková, Rona J. Strawbridge, Yun Ju Sung, Toshiko Tanaka, Alexander Teumer, Stella Trompet, Sander W. van der Laan, Jessica van Setten, Jana V. Van Vliet-Ostaptchouk, Zhaoming Wang, Loïc Yengo, Weihua Zhang, Uzma Afzal, Johan Ärnlöv, Gillian M. Arscott, Stefania Bandinelli, Amy Barrett, Claire Bellis, Amanda J. Bennett, Christian Berne, Matthias Blüher, Jennifer L. Bolton, Yvonne Böttcher, Heather A. Boyd, Marcel Bruinenberg, Brendan M. Buckley, Steven Buyske, Ida H. Caspersen, Peter S. Chines, Robert Clarke, Simone Claudi-Boehm, Matthew Cooper, E. Warwick Daw, Pim A. De Jong, Joris Deelen, Graciela Delgado, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46:1173, 2014.

Creative Commons License