The Editor's version has been published online first on April 26th, 2016 in *Diabetologia.*
The final publication is available at Springer via http://dx.doi.org/10.1007/s00125-016-3962-z

Commentary

The power of numbers

Marco Dauriz,[1] James B. Meigs[2, 3]

Institutions:

1. Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Verona School of Medicine and Hospital Trust of Verona, Ospedale Civile Maggiore, P.le Stefani, 1 – Pad. 22, 37126 Verona, Italy
2. General Medicine Division, Massachusetts General Hospital, Boston, MA, USA
3. Department of Medicine, Harvard Medical School, Boston, MA, USA

Corresponding author:
Marco Dauriz
Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Verona School of Medicine and Hospital Trust of Verona, Ospedale Civile Maggiore, P.le Stefani, 1 – Pad. 22, 37126 Verona, Italy
E-mail: marco.dauriz@univr.it

Received: 31 March 2016 / Accepted: 4 April 2016

Keywords: Genetic epidemiology; Genome-wide association study; Spectrum bias; Type 2 diabetes; Winner's curse

Abbreviations

AGEN-T2D, Asian Genetic Epidemiology Network-Type 2 Diabetes Consortium
CKB, China Kadoorie Biobank

GRS-BC, Beta cell function related genetic risk score
GRS-IR, Insulin resistance related genetic risk score
GRS-T, Overall Type 2 Diabetes genetic risk score

GWAS, Genome-wide association studies

The Editor's version has been published online first on April 26th, 2016 in *Diabetologia.*
The final publication is available at Springer via http://dx.doi.org/10.1007/s00125-016-3962-z

The study of human genetic variation represents a major chapter in the history of modern medicine and still holds the fascinating promise of identifying the genetic architecture of a large number of disease traits [1, 2]. This is particularly relevant nowadays, in light of the prospective rising trends in the global incidence and prevalence of type 2 diabetes [3]. However, the non-communicable diseases, such as type 2 diabetes, carry a considerable burden of complexity, as a result of the combination of genetic, epigenetic and environmental risk factors [4], which complicate the understanding of type 2 diabetes genetic background and its potential clinical transferability in terms of prognostic and preventive rules or potentially novel drug targets.

The technical and methodological advancements, as well as the knowledge accrued over the past decade on the haplotype block structure of the human genome [1, 5], have enabled investigators to tackle the complexity of the genetic architecture of type 2 diabetes in populations of European and non-European descent by performing large-scale genome-wide association studies (GWAS) for both common and rare genetic variants [2, 6]. To date, the international research consortia leading these initiatives have identified over 90 genetic loci credibly associated with a higher risk of type 2 diabetes, and more are expected in the years to come. When considered in aggregate, the loci identified explain only a limited proportion of the type 2 diabetes liability-scale variance, possibly because the actual number of the entire spectrum of type 2 diabetes risk loci has yet to be identified, or because the causal variant(s) harbored in or near these loci, as well as the underlying biological pathways, remain to be elucidated in most instances.

Interestingly, while interpreting the GWAS results one may observe that as the number of identified type 2 diabetes risk variants has increased over time, and the loci uncovered by earlier GWAS have been further replicated in larger association studies, the individual (per-allele) effect estimate has become smaller than the one originally detected in the discovery GWAS [7]. In addition, since novel variants are generally lower in frequency, ever larger sample sizes are needed to detect these effects in both the discovery and replication cohorts. The primary estimate of an inappropriately large effect size in case of newly identified variants is a known statistical phenomenon, usually dubbed as 'winner´s curse', whereby the multiple repetition of the association test in cohorts of adequate sample size leads to further refinements of the original effect size estimate, generally towards lower absolute values [2]. Another explanation for the possible inflation in the initial estimate of the single variant attributable risk stems from the case–control study design of the discovery GWAS, which is

inherently characterised by a disproportionate abundance of cases over controls compared with the real-world setting of population-based studies [2]. This phenomenon is also known as the 'spectrum bias' effect, and may be overcome by re-estimating the actual effect sizes of known type 2 diabetes risk variants in real-world populations.

Notably, as most of the type 2 diabetes risk loci have been identified in individuals of European ancestry, the population attributable risk for these variants in other ethnicities may be different, depending on the haplotype block structure (i.e. the linkage-disequilibrium map) of the population ancestry considered. A recently published trans-ethnic meta-analysis revealed that a considerable proportion of the variants originally identified in Europeans are shared across populations of diverse ethnicity and show directionally consistent effects [8], highlighting the opportunity to extrapolate insights on type 2 diabetes genetic susceptibility in Europeans to non-European populations. These observations are particularly relevant in that the unbiased estimate of the effect size and the strength of the statistical association with type 2 diabetes risk are both used to prioritise the inclusion of loci in reliable type 2 diabetes risk prediction models and, eventually, in fine-mapping and functional characterisation studies.

In this issue of *Diabetologia*, investigators working on behalf of the China Kadoorie Biobank (CKB) Collaborative Group provide a demonstration of the calculation of (relatively) unbiased allelic effect sizes for a set of 56 established type 2 diabetes risk variants in a large population-based cohort study of Chinese adult individuals, including ~7,100 type 2 diabetes cases and ~86,000 non-diabetic controls [9]. These variants were tested for association with type 2 diabetes risk in the CKB cohort, and the results were then combined in a meta-analysis with those from the Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, thus bringing the sample size up to ~32,100 cases and ~115,600 controls. The resulting effect estimates for the vast majority of the variants were highly concordant and directionally consistent with the original reports from existing meta-analyses of GWAS conducted in Europeans and East Asians. However, the authors observed a consistent proportional reduction (~20%) in the log OR of the allelic effect sizes estimated in CKB compared with those reported in AGEN-T2D and in GWAS of European case–control samples.

The authors suggest the occurrence of the 'spectrum bias/winner's curse' effects as possible explanations for this phenomenon [9]. Although this might be reasonably the case, the lower individual effect size detected in Chinese individuals might rather reflect the different haplotype patterns between European and Asian ancestries. However, this might not be a tenable assumption in this specific instance, as the trend towards decreased effect sizes was evident in the comparison of CKB vs AGEN-T2D as well as that of CKB vs Europeans. The linkage-disequilibrium patterns might instead account for the heterogeneity of effect occurring in isolated cases, such as the *RBMS1* rs7593730 and *GCC1–PAX4* rs6467136 loci, which appeared to be associated with type 2 diabetes risk only in Europeans and East Asians, respectively. However, as also pointed by the authors, a comprehensive explanation of the ethnic differences is currently elusive in the absence of fine-mapping studies at these loci. Of note, the individuals in the CKB cohort included a considerable proportion (~19%) of first-degree relatives. However, the authors conducted thorough sensitivity analyses, and the exclusion of those individuals did not appreciably impact the estimates of the effect sizes.
In addition, the authors successfully tested the performance of a weighted genetic risk score comprising the whole set of 52 type 2 diabetes susceptibility variants (GRS-T) on type 2 diabetes predictability in the CKB cohort. The same analyses were extended to sub-GRSs comprised of a limited set of loci with prior evidence of effect on beta cell function (GRS-BC, n=25 loci) or insulin-resistance (GRS-IR, n=7 loci) [9]. Although we might share with the authors the confidence that the unbiased estimate of the individual allelic weights for the variants comprised in the scores plays in favour of the generalisability of these findings to the Chinese population, we should however bear in mind that the inclusion of genetic information does not remarkably outperform existing clinical type 2 diabetes risk prediction models [10]. Moreover, the sub-GRS classifications were intrinsically weakened by the lack of internal validation in the CKB cohort, as the relevant metrics of beta cell function and insulin resistance were not available. Finally, the authors identified a significant interaction of adiposity measures with GRS-T and GRS-BC (but not GRS-IR), with leaner individuals carrying a proportionally higher type 2 diabetes genetic risk burden.

The current analysis of the CKB cohort is *per se* relevant, as it confirms the existence of a shared genetic background between Chinese and European populations, and provides a clear demonstration of extant upward biased estimates for the effect size of type 2 diabetes risk variants originally drawn from the discovery cohorts. As anticipated to a certain extent [2],

The Editor's version has been published online first on April 26th, 2016 in *Diabetologia.*
The final publication is available at Springer via http://dx.doi.org/10.1007/s00125-016-3962-z

these findings lessen the actual population attributable risk carried by established type 2 diabetes risk GWAS variants in Chinese individuals, and may be of help to prioritise loci for further functional interrogation.

This study is also interesting with regard to the nature of the analysis conducted. The attempt to replicate the associations of known loci with type 2 diabetes risk to obtain unbiased population-based estimates is not a discovery-driven exercise. As such, the attainment of genome-wide significance at all loci included in the analysis is not a *conditio sine qua non* to judge the relevance of the estimates, as it would require much larger sample sizes that are difficult to collect. However, if the population-based genome-wide scan were to be discovery-driven, the pitfall of eventually failing to reach genome-wide significance for the variants included in the association test would eventually lead to higher rates of false-negative results. Hence, in these specific situations, it might not be entirely heretical to apply less conservative type I error rate thresholds, although the absence of a solid standard entails some elements of uncertainty at discriminating some results as informative against the risk of being overly permissive. Whether the GWAS approach should remain a matter of statistical constraints only, or whether its integration with functional maps [11, 12] may highlight some sub-threshold loci as informative as those that reach genome-wide significance, is still an open question, which might be worth pursuing further.

In conclusion, the work of the CKB Collaborative Group is a remarkable example of the uncertainty that characterises the estimate of single variant-attributable risk and the dependence of its statistical relevance on the reference context. For example, every population cohort larger than the CKB herein discussed would provide less biased effect size estimates; therefore, the effect sizes estimated in the current CKB cohort, represent, by now, the best guess of what could be reasonably called 'true'. However, clinicians care more about the translational and biological relevance of genetic discoveries. Hence, in case obtaining ever larger sample sizes would not be practical, thus preventing the available estimate to be further improved, how could the statistical relevance be informed by the context? The complementary information that could arise from the full integration of the genetic and functional maps holds the promise of potentially uncovering clinically relevant mechanistic insights and might expand the regulatory framework in which to interpret the functional follow-up and fine-mapping currently ongoing at established type 2 diabetes risk loci.

Funding

MD's work is supported by the University of Verona. JBM's work is supported by NIDKK K24 DK080140 and R01 DK078616. No additional external funding was received for this manuscript. The funders had no role in the design or preparation of the manuscript.

Duality of interest

The authors have no conflict of interest to disclose.

Contribution statement

MD and JBM drafted the article and revised it critically for important intellectual content. As guarantors of this work, both authors approved the version to be published and take responsibility for its integrity and accuracy.

References

1. The 1,000 Genomes Project Consortium: A global reference for human genetic variation. *Nature* 2015, 526:68-74.
2. Pearson TA, Manolio TA: How to interpret a genome-wide association study. *JAMA* 2008, 299:1335-1344.
3. World Health Organization (2015) World Health Statistics 2015. WHO Press, Geneva
4. Prasad RB, Groop L: Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)* 2015, 6:87-123.
5. Zhang K, Calabrese P, Nordborg M, Sun F: Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002, 71:1386-1394.
6. Zuk O, Schaffner SF, Samocha K, et al: Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* 2014, 111:E455-E464.
7. Hivert MF, Vassy JL, Meigs JB: Susceptibility to type 2 diabetes mellitus--from genes to prevention. *Nat Rev Endocrinol* 2014, 10:198-205.
8. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium et al: Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 2014, 46:234-244.

The Editor's version has been published online first on April 26th, 2016 in *Diabetologia.*
The final publication is available at Springer via http://dx.doi.org/10.1007/s00125-016-3962-z

9. Gan W, Walters RG, Holmes MV et al (2016) Evaluation of type 2 diabetes genetic risk variants in Chinese adults: findings from 93,000 individuals from the China Kadoorie Biobank. *Diabetologia* doi: 10.1007/s00125-016-3920-9
10. Vassy JL, Hivert MF, Porneala B, et al: Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes* 2014, 63:2172-2182.
11. ENCODE Project Consortium et al.: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.
12. Edwards SL, Beesley J, French JD, Dunning AM: Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 2013, 93:779-797.