LETTER TO THE EDITOR

Approaches to studying the impact of 22q11.2 copy number variants

To the Editor: We read with interest the recent article "The impact of 22q11.2 copy-number variants on human traits in the general population" by Zamariolli et al.,[1] given the importance of copy number variations (CNVs) within this complex region, and in research that could further our understanding of the effects of 22q11.2 CNVs.

In-depth reading of this paper, however, reveals misconceptions and other issues that affect interpretation of the results. Most fundamental is what constitutes 22q11.2 deletions or 22q11.2 duplications with "highly deleterious impact." The literature of the past three decades, OMIM, and other standard references, would indicate that these would be the rare 22q11.2 CNVs flanked by the low-copy repeats (LCRs) of this region, particularly those that are 1.5–3 million base pairs in size (LCR22A-B, A-C, A-D). Here, however, the authors have included what they term "atypical" CNVs. These appear to primarily constitute a 132 kb copy number polymorphism (CNP) overlapping LCR22A.[2] Of the reported "1,127 individuals (1,128 per Figure 1) with a duplication and 694 individuals with a deletion overlapping the 22q11.2 LCR22A-D region," the vast majority (n = 663, 59%, and n = 649, 94%, respectively) have this polymorphism, and *not* a clinically meaningful rare 22q11.2 disease-causing CNV. CNV detection that relies on a single algorithm and single-nucleotide polymorphism (SNP) data, without considering copy number data, positions within LCRs, or previous findings, may be prone to inaccuracy and/or pseudo-precision.[3] In several analyses in their paper, these CNPs were excluded. Moreover, their individual probe analyses showed no significant association for the probe within LCR22A with any trait tested (Figures 2–5, S2, and S3), indicating that these small CNV/CNPs are not clinically meaningful. Their inclusion in the large initial numbers provided thus appears all the more confusing, and potentially quite misleading to readers with respect to numbers of pathogenic CNVs in this healthy UK Biobank (UKBB) sample.

The other basic, though relatively widespread, misconception is that the UKBB data are representative of a general population sample. They are not.[4] This group of altruistic participants has on average lived to 65 years of age, and is disproportionately healthy, highly educated, from high socioeconomic classes, of European descent, and female (54%).[4] These are substantive ascertainment biases that are important to consider.

These misconceptions, coupled with use of probe-level association that ignores the effect of CNV size, which is a well-known contributing factor for phenotype association, add to an overall lack of clarity about what exactly is contributing to each of the bioinformatically driven findings presented. The results may thus relate to the LCR22A CNP, some to SNP-based homozygosity/heterozygosity calls, and a minority to actual disease-causing 22q11.2 CNVs.

The fact that the numbers vary throughout the manuscript makes interpretation even more challenging. As an example, there are fourteen 22q11.2 deletions, presented as n = 5 LCR22A-D and n = 9 LCR22A-B, in Figures 1 and 4. Sample sizes of n = 10 (Figure 3) and n = 11 (Figure 5) may be because of missing phenotypic data, but a sample size of n = 18 (Figure 2) is challenging to undertand. In any case, these numbers suggest that the study is under-powered with respect to typical pathogenic 22q11.2 deletions. Sensitivity analyses, excluding, in turn, the few individuals with each type of conventional 22q11.2 deletion, could have helped with interpretation of probe-level analyses. The same principles would hold true for 22q11.2 duplications, although there, for LCR22A-D (n = 236) and LCR22A-B (n = 50) duplications (Figure 1), numbers appear more favorable.

Explaining the discrepancy in the numbers of individuals previously identified within the same UKBB dataset by Crawford et al.[4] for 22q11.2 CNVs (e.g., n = 10 with a 22q11.2 deletion; n = 266 with a 22q11.2 duplication) would have helped with interpretation of the genotyping methods. The only reference to this previous study is in regard to what is termed "replication:" "While LCRA to LCRD duplications have been previously associated with this trait in the UKBB cohort, replication of the association in our study emphasizes its relevance in 22q11.2 CNV carriers." Generally, scientific replication would apply to findings from an *independent* sample.

Among limitations not discussed, of the multi-layered bioinformatic approaches used, are those related to electronic health record data and HPO terms. Phenotypes (n = 152) were selected on the basis of use as a descriptor for at least 500 of the mostly elderly participants in the UKBB. One of these, "Other cerebral degenerations," an ICD-9 term embracing Alzheimer and frontotemporal dementia, may for some individuals relate to intellectual and/or learning disabilities. This possibility could be in keeping with the reported odds ratio of 45 (Table 3) for the "deletion-only model" and an SNP in *MED15*, based on perhaps 5 individuals with LCR22A-D deletions.

Placing the results in the context of previously published results for both "binary" (categorical) phenotypes and quantitative traits would enable appreciation of any truly novel results.[4,5] That SNPs on either side of *GP1BB* are associated with mean platelet volume and platelet count variables (though gene positions are not provided in the tables) would not appear to be a novel finding, for example.[6] On the other hand, a single SNP at chr22: 20,765,989 (within gene *ZNF74*) is reported to be

associated with three weight-associated quantitative traits and to the "other venous embolism and thrombosis" binary feature. Another SNP at chr22: 21,370,246 (within *P2RX6P*) is associated with three disparate features (cardiomegaly, dental caries, nausea and vomiting). Proposing explanations for such results could assist in understanding the contributing data/samples, the statistical and data manipulations (including use of probes within LCR regions), and the bioinformatic methods used and their limitations. Casual use of terms such as "causal" and "mapped" would perhaps then be minimized, to further assist with interpretation.

Despite the shortcomings identified, there may be nuggets of interest that exploit the biases of the UKBB sample and that could dovetail with findings from independent studies. The existence of LCR22A-B deletions may be more prevalent in the general population than the ~10% that clinically based studies often indicate.[6] Support comes from a recent population-based study of newborn screening samples from ~30,000 singleton live births that showed 1 in 2,148 newborns had a "typical" 22q11.2 deletion (most, LCR22A-D deletion), but that 21% had the nested LCR22A-B deletion.[7] In Zamariolli et al.,[1] as may be expected from the biases to health and longevity inherent to the UKBB sample, the few participants with a typical 22q11.2 deletion appear to be disproportionately those with nested LCR22A-B deletions (9 of 14, 64%), assuming accuracy of numbers and genotyping data provided.

Clinically based studies have demonstrated that the phenotypes of individuals with proximal nested variants may differ from those with the most common full-length LCR22A-D deletion.[8,9] Recently, this has included a study of adult height using population norms for males and females and accounting for other possible contributors,[9] with results indicating that a full length ~3 Mb LCR22A-D deletion conveyed significantly greater risk of adult short stature than proximal nested 22q11.2 deletions. This suggests the possibility that gene dosage effects from both the proximal nested LCR22A-B and distal nested LCR22B-D sub-regions may contribute to the low end of final height. These results may be consistent with the results using the bioinformatic approach here and the relatively healthy elderly UKBB sample that accrued 31 individuals with LCR22C-D deletions. Though not annotated as such in Table 2, one of the few significant findings for the "deletion-only model" was a single LCR22C-D region SNP within *SNAP29* associated with shorter stature. The effect size was comparable to that for a "lead" LCR22A-B region SNP within the extent of both *CDC45*, a gene previously reported to be associated with short stature,[10] and *CLDN5*.

It may be of interest to assess the presence and frequency of the "atypical" CNVs in other genomic databases (e.g., gnomAD) where CNVs are called from genome-sequence data.[11] For the interested reader, however, the lack of genomic coordinates provided for the CNVs in Zamariolli et al. makes such attempts infeasible. Also, going forward, CNV analyses could be performed using genome-sequence data for the UKBB that are now available.[12]

We emphasize that clinical studies of rare diseases are challenging to undertake, including recruiting, phenotyping, and longitudinal follow-up, but produce highly valuable results not obtainable using other approaches. Individuals with neurodevelopmental and neuropsychiatric expression in particular are vastly under-represented in "population-based" samples.[4,13] Clinically based studies often use within-cohort strategies, e.g., individuals with and without a specific phenotype,[6,14] to address major ascertainment issues. However, ascertainment bias faces every human research study, including those that are touted as "population-based" yet require voluntary consent. We suggest that clinical studies not be labeled "biased" without recognition of the inherent biases and limitations of other research approaches. All are complementary strategies with a common goal that would usually include better understanding of diseases and mechanisms in order to improve outcomes.

The field deserves a balanced and accurate portrayal of the approaches that may be employed to study the impact of CNVs on human traits, particularly for the rare conditions with premature mortality and prominent neurodevelopmental and neuropsychiatric manifestations that present particular challenges to research.[6] Appreciation of the benefits and potential drawbacks of studying patient populations may be easier than for secondary data mining studies where the size of the denominator may be dazzling. When samples are collected by others, it may be challenging to keep track of what has already been done, and to clarify differences in specific methods, including variant identification, from study to study. Nonetheless, acknowledging sampling, data, and methodological issues, both genotypic and phenotypic, and delineating the strategies used to address limitations, remain important for interpretation of results. Perhaps especially in this era of "big data," both quantity *and* quality deserve consideration.

Anne S. Bassett,[1,2,3,4,14,*]
Donna M. McDonald-McGinn,[5,6,7,14]
Erik Boot,[1,8,9,14] Sólveig Óskarsdóttir,[10,11,14]
and Ryan K.C. Yuen[12,13,14]
[1]The Dalglish Family 22q Clinic for Adults, and Department of Psychiatry, University Health Network, Toronto, ON, Canada; [2]Toronto General Research Institute and Division of Cardiology, Department of Medicine, University Health Network, Toronto, ON, Canada; [3]Clinical Genetics Research Program and Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada; [4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada; [5]22q and You Center, Clinical Genetics Center, and Section of Genetic Counseling, Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [6]Department of Pediatrics, Perelman School of Medicine of the University of Pennsylvania, Philadelphia, PA, USA; [7]Department of Human Biology and Medical Genetics, Sapienza University, Rome, Italy; [8]Advisium, 's Heeren Loo Zorggroep, Amersfoort, the

Netherlands; [9]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, the Netherlands; [10]Department of Pediatric Rheumatology and Immunology, Queen Silvia Children's Hospital, SahlgrenskaUniversity Hospital, Gothenburg, Sweden; [11]Department of Pediatrics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [12]Genetics & Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada; [13]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

[14]22q11.2 Society (https://www.22qsociety.org)

*Correspondence: anne.bassett@utoronto.ca

Acknowledgments

References

1. Zamariolli, M., Auwerx, C., Sadler, M.C., van der Graaf, A., Lepik, K., Schoeler, T., Moysés-Oliveira, M., Dantas, A.G., Melaragno, M.I., and Kutalik, Z. (2023). The impact of 22q11.2 copy-number variants on human traits in the general population. Am. J. Hum. Genet. *110*, 300–313.

2. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525–528.

3. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

4. Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K.M., Rees, E., Pardiñas, A.F., Einon, M., Escott-Price, V., Walters, J.T.R., O'Donovan, M.C., et al. (2019). Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. J. Med. Genet. *56*, 131–138.

5. Owen, D., Bracher-Smith, M., Kendall, K.M., Rees, E., Einon, M., Escott-Price, V., Owen, M.J., O'Donovan, M.C., and Kirov, G. (2018). Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. BMC Genom. *19*, 867.

6. McDonald-McGinn, D.M., Sullivan, K.E., Marino, B., Philip, N., Swillen, A., Vorstman, J.A.S., Zackai, E.H., Emanuel, B.S., Vermeesch, J.R., Morrow, B.E., et al. (2015). 22q11.2 deletion syndrome. Nat. Rev. Dis. Primers *1*, 15071.

7. Blagojevic, C., Heung, T., Theriault, M., Tomita-Mitchell, A., Chakraborty, P., Kernohan, K., Bulman, D.E., and Bassett, A.S. (2021). Estimate of the contemporary live-birth prevalence of recurrent 22q11.2 deletions: a cross-sectional analysis from population-based newborn screening. CMAJ Open *9*, E802–E809.

8. Zhao, Y., Guo, T., Fiksinski, A., Breetvelt, E., McDonald-McGinn, D.M., Crowley, T.B., Diacou, A., Schneider, M., Eliez, S., Swillen, A., et al. (2018). Variance of IQ is partially dependent on deletion type among 1,427 22q11.2 deletion syndrome subjects. Am. J. Med. Genet. *176*, 2172–2181.

9. Heung, T., Conroy, B., Malecki, S., Ha, J., Boot, E., Corral, M., and Bassett, A.S. (2022). Adult height, 22q11.2 deletion extent, and short stature in 22q11.2 deletion syndrome. Genes *13*, 2038.

10. Unolt, M., Kammoun, M., Nowakowska, B., Graham, G.E., Crowley, T.B., Hestand, M.S., Demaerel, W., Geremek, M., Emanuel, B.S., Zackai, E.H., et al. (2020). Pathogenic variants in CDC45 on the remaining allele in patients with a chromosome 22q11.2 deletion result in a novel autosomal recessive condition. Genet. Med. *22*, 326–335.

11. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

12. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740.

13. Birnbaum, R., Mahjani, B., Loos, R.J.F., and Sharp, A.J. (2022). Clinical characterization of copy number variants associated with neurodevelopmental disorders in a large-scale multiancestry biobank. JAMA Psychiatr. *79*, 250–259.

14. Cleynen, I., Engchuan, W., Hestand, M.S., Heung, T., Holleman, A.M., Johnston, H.R., Monfeuga, T., McDonald-McGinn, D.M., Gur, R.E., Morrow, B.E., et al. (2021). Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. Mol. Psychiatry *26*, 4496–4510.