Large copy number variants in UK Biobank caused by clonal haematopoiesis may confound penetrance estimates

Marcus Tuke[1], Jessica Tyrrell[1], Katherine S. Ruth[1], Robin N. Beaumont[1], Andrew R. Wood[1], Anna Murray[1], Timothy M. Frayling[1], Michael N. Weedon[1], Caroline F. Wright[1]*

[1] Institute of Biomedical and Clinical Science, University of Exeter Medical School, RILD Building, Royal Devon & Exeter Hospital, Barrack Road, Exeter UK, EX2 5DW

* Corresponding author: caroline.wright@exeter.ac.uk

ABSTRACT

Large copy number variants (CNVs) are strongly associated with both developmental delay and cancer, but the type of disease depends strongly on when and where the mutation occurred, i.e. germline versus somatic. We used microarray data from UK Biobank to investigate the prevalence and penetrance of large autosomal CNVs and chromosomal aneuploidies using a standard CNV detection algorithm not designed for detecting mosaic variants. We found 160 individuals that carry >10Mb copy number changes, including 56 with whole chromosome aneuploidies. Nineteen (12%) individuals had a diagnosis of Down's syndrome or other developmental disorder, while 84 (52.5%) individuals had a diagnosis of haematological malignancies or chronic myeloproliferative disorders. Notably, there was no evidence of mosaicism in the blood for many of these large CNVs, so they could easily be mistaken for germline alleles even when caused by somatic mutations. We

therefore suggest that somatic mutations associated with blood cancers may result in false estimates of rare variant penetrance from population biobanks.

MAIN TEXT

Copy number variants (CNVs) are deletions or duplications of DNA that can vary in size from 50 basepairs to several hundred megabases[1], i.e. entire chromosomes. Individuals typically carry several thousand CNVs, most of which are small (<1Mb) and rare (allele frequency <1%)[2–4]. Large, rare pathogenic CNVs have historically been identified through clinical microarray testing of two distinct clinical cohorts: firstly, children with developmental disorders caused predominantly by germline mutations[5–7], and secondly, individuals with haematological and other cancers associated with somatically-acquired mutations[8–10].

The availability of large, well-genotyped population biobanks offers an opportunity to investigate the prevalence and penetrance of monogenic disease-causing variants[11]. Several studies have already been published evaluating known developmental CNVs in 500,000 adults in the UK and Estonian Biobanks[12–19] and the penetrance of X-chromosome aneuploidy has been investigated in UK Biobank[20]. However, given the relatively advanced age of UK Biobank participants, ranging from 40–70 years (mean = 56.5 years) at recruitment, it is likely that some variants will be due to somatic mutation and age-related clonal haematopoiesis[21,22] as has previously been observed in genome-wide association study cohorts[23,24]. Importantly, somatic variants in adult population cohorts should not be used to evaluate the penetrance of germline CNVs known to cause developmental disorders, as this will result in spurious associations.

We sought to investigate large (>10Mb) autosomal CNVs present in population datasets with the aim of determining whether they were likely to be germline (and therefore potentially useful for penetrance studies) or somatic (and therefore caused by clonal expansions). We used microarray data from UK Biobank, which recruited 502,506 individuals from across the UK between 2006-2010[25]. Hospital Episode Statistics (HES) and cancer registry data were available for the whole cohort up to 31 March 2017, and GP records were available for half the cohort; all participants also provided a range of information (e.g., demographics, health status, lifestyle) via questionnaires. Genotypes for SNVs and indels were generated from blood-extracted DNA using the Affymetrix Axiom UK Biobank array (~450,000 individuals) and the UKBiLEVE array (~50,000 individuals) in 106 batches of ~4,700 samples. This dataset underwent extensive central quality control[25].

We called CNVs genome-wide in 488,377 individuals with array genotyping data in UK Biobank[25] using PennCNV version 1.0.4[26] with log R ratio and B-allele frequency values for 805,426 genome-wide probe sets provided by UK Biobank. Very large PennCNV calls (i.e. multiple megabases) can sometimes be fragmented into many smaller calls, so we additionally calculated the sum of bases either deleted or duplicated on each chromosome per individual according to the PennCNV calls. We carried out visual inspection of each event in everyone with >10Mb deleted and/or duplicated on a single chromosome to confirm breakpoints, event type and level of mosaicism (see examples in Figure 1). Around a third of the events showed no evidence of mosaicism in blood (based on a deviation of the

B-allele frequencies from 0, 0.5 or 1 with a co-located increase/decrease in log R ratio) while two-thirds of events were consistent with the presence of a large CNV in a proportion of cells. Based on previous work investigating mosaicism in UK Biobank[20,27], we estimate that we were able to detect copy number changes present in >20-25% of cells. There was a good correlation between the log R ratio and visual inspection of mosaicism[28] (Figure 2).

We identified 160 individuals in UK Biobank (61% male versus 46% in the whole of UK Biobank; Pearson's Chi-squared p=0.025) with >10Mb involved in copy number events on a single autosome (Figure 3 and Supplementary Table 1). This male bias has been observed previously and is thought to be due to higher male-specific rates of certain haematological malignancies[24]. In the majority (134/160) of individuals, this was caused by a single large CNV; 19 individuals had two separate events (17 on two different chromosomes), five individuals had three separate events (all involving at least two different chromosomes) and two individuals had four or five separate events on the same chromosome. Individual events ranged in size from 0.9Mb-198Mb (mean=56Mb, stdev=51Mb), and included both unique events and recurrent events. There were 64 whole chromosome duplications of chromosomes 3, 8, 9, 12, 14, 18, 19 and 21, including four individuals with two trisomies and three individuals with three trisomies.

Of the autosomal aneuploidies, only trisomy 21 is compatible with adult life when present constitutively and causes Down's Syndrome [29]. Twelve individuals had a duplication of chromosome 21, of whom 11 had a diagnosis of Down's syndrome in their HES or GP records (GP records were not available for the remaining individual). A further six individuals in UK Biobank had Down's syndrome recorded in their HES records, but their microarray data showed no evidence of trisomy 21. It is unclear whether these discrepancies are caused by sample mix-ups, errors in HES records, or misdiagnoses. Of those with large CNVs, a further eight individuals had ICD-10 codes or GP records consistent with various developmental disorders, including congenital malformations, developmental disorder (intellectual disability/handicap or epilepsy) and bipolar affective disorder (Supplementary Table 1).

We suggest that the rest of the whole chromosome duplications and the majority of large CNVs are likely to be somatic mutations caused by clonal expansions, some of which are compatible with being present in (apparently) healthy individuals. Several lines of evidence suggest that the majority of the large CNVs were likely caused by somatic mutations associated with either cancer[10,30] or age-related haematopoietic clonal expansions[22,31]. First, 79/160 (50%) individuals had a recurrent duplication of chromosome 8, 9, 12 or 19 or large deletions on chromosome 11q, 13q, 17p and 20q that are consistent with those observed previously in lymphocytic and myeloid leukaemias[10,30,32,33] and *JAK2*-related myeloproliferative neoplasms[34]. Second, 98/160 (61%) individuals had neoplasms recorded in their HES records or cancer registry data compared with 80,046 (17%) across the whole of UK Biobank ($p< 2x10^{-16}$). Sixty-four (40%) were malignant neoplasms of lymphoid, haematopoietic and related tissues (ICD-10 codes: C81-96), a significant enrichment above the whole of UK Biobank (n=3869, 0.8%, $p< 2x10^{-16}$); and a further 20 were polycythemia vera, myelodysplastic syndrome and chronic myeloproliferative diseases (ICD-10 codes: D45-47), again a significant enrichment above the whole of UK Biobank (n=646, 0.1% $p< 2x10^{-16}$). Third, individuals with large CNVs and neoplasms were older and taller versus the others with large CNVs but no record of neoplasms, as expected in cancer; in contrast, individuals with large CNVs and a developmental disorder were younger and shorter versus the others, as expected with most developmental disorders (Figure 4).

Nonetheless, 43/160 (27%) individuals with a large CNV in UK Biobank have neither a developmental disorder nor a neoplasm (of any sort) recorded to date. This observation has a range of explanations, including record error, lack of hospitalisation, absence of GP records (currently only available for around half the cohort), benign haematopoietic clonal expansions, or neoplasms that have not yet developed or been diagnosed. Given the prognostic link between chromosomal instability and tumorigenesis[30,35], unfortunately the latter explanation is likely to be true in many cases.

Mosaic chromosomal alterations[27] and Y chromosome loss[36] in UK Biobank have previously been linked to age-related clonal haematopoiesis, both of which can be easily excluded from studies seeking to investigate penetrance of germline CNVs. Indeed, although some mosaic variants can be detected by PennCNV, low-level mosaic variants are often not detected using standard variant calling algorithms. However, the presence of very rare autosomal aneuploidies, some of which do not appear to be mosaic based on intensity data from microarrays, suggest that caution should be used when interpreting rare variants (of any size) in population biobanks. For example, we note that six individuals in UK Biobank have complete or partial trisomy of chromosome 3, of whom 5/6 have non-Hodgkin’s lymphoma. Presumed germline duplications of 3q29 have previously been causally linked with early death (OR=27.8) and cancer (OR=37.5) in UK Biobank[12], but we suggest that these associations more likely reflect reverse causality, with cancer causing both the mutations and early death. A similar issue has previously been highlighted for sequence variants in cancer driver genes where rare mutations are also a cause of developmental disorders, such as *ASXL1* and *DNMT3A*[37]. Based on our analysis, the issue of somatic mutations confounding analyses of variants presumed to be germline can only be partially addressed by assigning the mosaic status of each variant; critically evaluating the validity of a variant (such as the breakpoints of a CNV), the plausibility of a finding (such as presence of constitutive autosomal aneuploidy) and likelihood of different mutational mechanisms (such as clonal expansion) are also important. As genome-wide sequencing becomes widely available in aging cohorts such as UK Biobank, researchers should be aware of potential confounding caused by somatic mutations present in high proportions of cells.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource under Application Number 49847. The authors would like to acknowledge the use of the University of Exeter High-Performance Computing facility in carrying out this work. The authors declare no competing interests.

DATA & CODE AVAILABILITY

This study did not generate new datasets or code. The code used during this study is available at https://github.com/WGLab/PennCNV. All bona fide researchers can apply to use the UK Biobank resource for health related research that is in the public interest, https://www.ukbiobank.ac.uk/.

REFERENCES

1. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.
2. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.
3. Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.
4. Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.
5. Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. Nat. Genet. *43*, 838–846.
6. Coe, B.P., Witherspoon, K., Rosenfeld, J.A., van Bon, B.W.M., Vulto-van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E.L.M., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat. Genet. *46*, 1063–1071.
7. Veltman, J.A. (2006). Genomic microarrays in clinical diagnosis. Curr. Opin. Pediatr. *18*, 598–603.
8. Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. Genome Med. *1*, 62.
9. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*, 722–729.
10. Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Izraeli, S., and Shamir, R. (2011). Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. Genome Biol. *12*, R61.
11. Wright, C.F., West, B., Tuke, M., Jones, S.E., Patel, K., Laver, T.W., Beaumont, R.N., Tyrrell, J., Wood, A.R., Frayling, T.M., et al. (2019). Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. Am. J. Hum. Genet. *104*, 275–286.
12. 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.
13. Kendall, K.M., Bracher-Smith, M., Fitzpatrick, H., Lynham, A., Rees, E., Escott-Price, V., Owen, M.J., O'Donovan, M.C., Walters, J.T.R., and Kirov, G. (2019). Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. Br. J. Psychiatry *214*, 297–304.
14. 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 Genomics *19*, 867.
15. Kendall, K.M., Rees, E., Escott-Price, V., Einon, M., Thomas, R., Hewitt, J., O'Donovan,

M.C., Owen, M.J., Walters, J.T.R., and Kirov, G. (2017). Cognitive performance among carriers of pathogenic copy number variants: analysis of 152,000 UK biobank subjects. Biol. Psychiatry *82*, 103–110.
16. Kendall, K.M., Rees, E., Bracher-Smith, M., Legge, S., Riglin, L., Zammit, S., O'Donovan, M.C., Owen, M.J., Jones, I., Kirov, G., et al. (2019). Association of rare copy number variants with risk of depression. JAMA Psychiatry.
17. Macé, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Nõukas, M., Sapkota, Y., Schick, U., Porcu, E., Rüeger, S., et al. (2017). CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat. Commun. *8*, 744.
18. Männik, K., Mägi, R., Macé, A., Cole, B., Guyatt, A.L., Shihab, H.A., Maillard, A.M., Alavere, H., Kolk, A., Reigo, A., et al. (2015). Copy number variations and cognitive phenotypes in unselected populations. JAMA *313*, 2044–2054.
19. Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide Burden of Copy-Number Variation in the UK Biobank. Am. J. Hum. Genet. *105*, 373–383.
20. Tuke, M.A., Ruth, K.S., Wood, A.R., Beaumont, R.N., Tyrrell, J., Jones, S.E., Yaghootkar, H., Turner, C.L.S., Donohoe, M.E., Brooke, A.M., et al. (2018). Mosaic Turner syndrome shows reduced penetrance in an adult population study. Genet. Med. *21*,.
21. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.
22. Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science *366*,.
23. Vattathil, S., and Scheet, P. (2016). Extensive hidden genomic mosaicism revealed in normal tissue. Am. J. Hum. Genet. *98*, 571–578.
24. Machiela, M.J., Zhou, W., Sampson, J.N., Dean, M.C., Jacobs, K.B., Black, A., Brinton, L.A., Chang, I.-S., Chen, C., Chen, C., et al. (2015). Characterization of large structural genetic mosaicism in human autosomes. Am. J. Hum. Genet. *96*, 487–497.
25. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
26. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.
27. Loh, P.-R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A., et al. (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature *559*, 350–355.
28. Conlin, L.K., Thiel, B.D., Bonnemann, C.G., Medne, L., Ernst, L.M., Zackai, E.H., Deardorff, M.A., Krantz, I.D., Hakonarson, H., and Spinner, N.B. (2010). Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum. Mol. Genet. *19*, 1263–1275.
29. Hassold, T., and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. Nat. Rev. Genet. *2*, 280–291.
30. Ben-David, U., and Amon, A. (2020). Context is everything: aneuploidy in cancer. Nat. Rev. Genet. *21*, 44–62.
31. Shlush, L.I. (2018). Age-related clonal hematopoiesis. Blood *131*, 496–504.
32. Landau, D.A., Tausch, E., Taylor-Weiner, A.N., Stewart, C., Reiter, J.G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., et al. (2015). Mutations driving CLL and their evolution in progression and relapse. Nature *526*, 525–530.

33. Marasca, R., Maffei, R., Martinelli, S., Fiorcari, S., Bulgarelli, J., Debbia, G., Rossi, D., Rossi, F.M., Rigolin, G.M., Martinelli, S., et al. (2013). Clinical heterogeneity of de novo 11q deletion chronic lymphocytic leukaemia: prognostic relevance of extent of 11q deleted nuclei inside leukemic clone. Hematol Oncol *31*, 88–95.
34. James, C., Ugo, V., Le Couédic, J.-P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., et al. (2005). A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature *434*, 1144–1148.
35. Danielsen, H.E., Pradhan, M., and Novelli, M. (2016). Revisiting tumour aneuploidy - the place of ploidy assessment in the molecular era. Nat. Rev. Clin. Oncol. *13*, 291–304.
36. Thompson, D.J., Genovese, G., Halvardson, J., Ulirsch, J.C., Wright, D.J., Terao, C., Davidsson, O.B., Day, F.R., Sulem, P., Jiang, Y., et al. (2019). Genetic predisposition to mosaic Y chromosome loss in blood. Nature.
37. Carlston, C.M., O’Donnell-Luria, A.H., Underhill, H.R., Cummings, B.B., Weisburd, B., Minikel, E.V., Birnbaum, D.P., Exome Aggregation Consortium, Tvrdik, T., MacArthur, D.G., et al. (2017). Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. Hum. Mutat. *38*, 517–523.

FIGURE LEGENDS

Figure 1. Example mean log R ratio (LRR) and B-allele frequency plots.

(a) Constitutive deletion of half the q-arm of chromosome 20, **(b)** four small non-mosaic deletions on chromosome 4, **(c)** mosaic deletion of the end of the q-arm of chromosome 6, **(d)** constitutive duplication of the whole of chromosome 21, (**e**) mosaic duplication of the whole of chromosome 19 and **(f)** triplication of chromosome 9p. Alleles (A and B) corresponding to each of B-allele frequencies are indicated. Red = copy number change; blue = normal copy number.

Figure 2. Mean log R ratio of CNVs

Boxplot of mean log R ratios of large CNVs, grouped by whether there was evidence of mosaicism based on visual inspection of the data. Red = deletion (DEL); blue = duplication (DUP); green = triplication (TRIP).

Figure 3. Summary of large autosomal CNVs identified

Circos plot of all large autosomal CNVs in UK Biobank; chromosomes 1-22 are indicated, and CNVs on the same chromosome in the same person are shown on the same track. Red = deletion; blue = duplication; green = triplication.

Figure 4. Characteristics of individuals in UK Biobank with >10MB copy number changes.

(a) Age in years at recruitment and **(b)** height of individuals grouped by whether they had a neoplasm, developmental disorder or neither coded in their HES or cancer registry records.

Red = developmental disorder; green = neoplasm; blue = neither neoplasm nor developmental disorder recorded; dotted red line = average for UK Biobank.

TABLES

Supplementary Table 1. List of all >10MB copy number changes in UK Biobank detected using SNP-genotyping arrays. DUP = duplication, DEL = deletion, LRR = log R ratio, St Dev LRR = standard deviation of the log R ratio, N SNPs = number of SNPs in the call.