Genetic risk factors

https://doi.org/10.1038/s44161-022-00198-3

Genetic causes and cardiovascular consequences of clonal hematopoiesis in the UK Biobank

J. Scott Beeler, Alexander G. Bick & Kelly L. Bolton

Pioneering cohort studies including the Framingham Heart Study have led to major insights into cardiovascular disease. However, these studies are underpowered to identify the effects of less common risk factors on human health. This has motivated the development of the UK Biobank, a biomedical database linking health and genetic information in 500,000 individuals.

Although the UK Biobank is large, it is not a representative sample of the UK population. For example, 2.4% of women aged 45–54 years in the UK Biobank have self-reported cardiovascular disease (CVD) compared with 10.3% of the UK population in general[1]. Despite these limitations, the scale of the UK Biobank makes it a valuable resource to explore a newly described genetic risk factor for CVD — clonal hematopoiesis (CH). With aging, normal tissues acquire somatic mutations that in some cases provide a fitness advantage to the cell they occur in, enabling preferential expansion of the cell (termed a 'clone'). When this process occurs in hematopoietic stem cells, it is referred to as CH. CH can be assessed by sequencing of peripheral blood samples, and common mutations that drive hematopoiesis overlap with common drivers of hematological malignancy (for example, *DNMT3A*, *TET2* and *ASXL1*). Research over the past decade has shown that CH is associated with worse health outcomes, with strong evidence supporting an increased risk of hematological malignancy[2,3] and CVD[2,4,5]. A wide variety of other health outcomes across diverse organ systems have been linked to CH, some of these include increased risks of chronic obstructive pulmonary disease[6], severe COVID-19 infection[7] and chronic kidney disease[8] (Fig. 1). However, the molecular pathways involved in the initiation of CH and its consequences on disease are not well understood.

In a recent issue of *Nature*, Kessler et al.[9] analyze whole-exome sequencing data from 628,388 individuals in the UK Biobank (UKB) and the Geisinger MyCode Community Health Initiative to identify 40,208 individuals with CH. They identified several new germline genetic loci associated with CH, and studied the effect of CH on a wide variety of health outcomes.

To study germline predisposition to CH, the authors first performed a common variant (minor allele frequency > 0.5%) genome-wide association study (GWAS) in the UKB cohort using single nucleotide polymorphism (SNP) array data. They identified 24 common variant loci (21 of which were new) that achieved genome-wide significance,

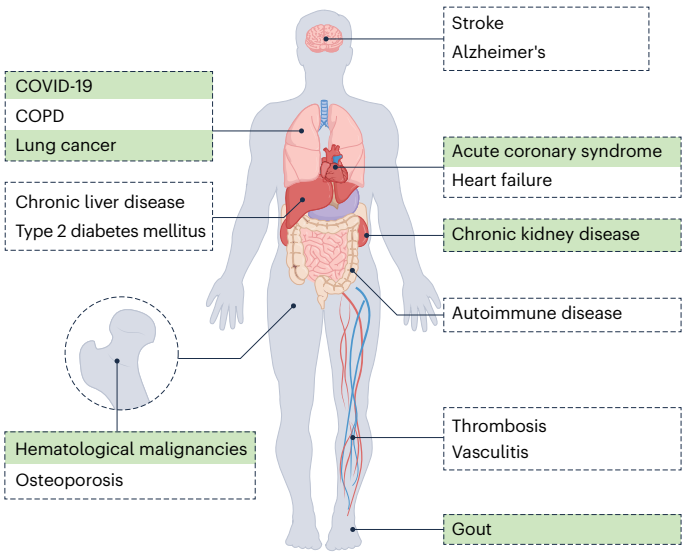

Fig. 1 | CH is associated with health outcomes across many organ systems. Boxes shaded green highlight associations between CH and diseases that were replicated by Kessler et al.[9]. COPD, chronic obstructive pulmonary disease.

and were able to validate 15 of these by replication analysis in the Geisinger cohort. Second, using exome-sequencing data, they identified one locus via rare variant testing (frameshift mutation in *CHEK2*) and three loci via gene burden testing (*ATM*, *CHEK2* and *CTC1*) that were significantly associated with CH. All of these replicated in the Geisinger cohort except *CTC1*. Third, they analyzed common variant associations for each CH mutation separately, and identified eight additional loci that were not identified in the analysis of pooled CH mutations; this included six loci associated with *DNMT3A*. Most of these loci were associated with an increased risk of *DNMT3A* CH, with notable exceptions including the *PARP1* locus on chromosome 1 and a locus on chromosome 2 near *LY75*. Further supporting the importance of gene-specific CH susceptibility analyses, the authors replicated a previous observation of genome-wide significant effects at the *TCL1A* locus with opposing directions across CH subtypes, with an increased risk of *DNMT3A* CH but a reduced risk of *TET2* and *ASXL1* CH[10].

CH is also driven by large-scale copy number events that are readily detectable via SNP array data. The most common being loss of Y or X chromosomes, otherwise known as mosaic loss of Y and X (mLOY and mLOX, respectively). Similar to the gene-specific analyses, the authors

found that many of the germline genetic variants that cause CH also cause copy number variant CH, mLOY and mLOX; however, some are specific to a particular type of mosaic mutation.

Two interesting observations emerge from these results. First, the findings emphasize that although germline variants that predispose to CH are often shared across different driver mutations, some are specific to somatic alterations of particular genes or pathways. Second, the germline genetic variants are near genes largely associated with cancer predisposition and almost entirely distinct from previous CVD GWAS, which suggests that CH and CVD are not related to one another owing to common genetic predisposition that leads to both phenomena.

Finally, the authors analyzed the associations between CH and diverse phenotypes available in the UKB. CH was associated with an increased risk of hematological malignancy, CVD (particularly *TET2*-mutant CH), and all-cause mortality, as has been established by several independent groups[2–4,11]. Also supported by previous work, CH was associated with gout and risk of severe COVID-19[7,12]. Interestingly, CH was associated with a modest risk of solid malignancies including lung cancer, non-melanoma skin cancer and prostate cancer. The association with lung cancer was replicated in the Geisinger cohort and remained significant even after controlling for smoking history.

Although many of the disease associations found in this study are similar in direction and effect to previous work, there were exceptions. For example, the strength of the association between CH and CVD risk was more modest (hazard ratio 1.11 [range 1.03–1.19], $P = 4.2 \times 10^{-3}$) than previous reports[4,11]. In addition, the authors did not observe a cardioprotective effect of an *IL6R* missense variant (rs2228145-C) that is a genetic proxy for IL-6R inhibition among CH carriers, as previously reported in an analysis of the first 50,000 UKB samples[11]. Notably, when the authors performed a sensitivity analysis using only the first 50,000 UKB samples, they both found a stronger association between CH and CVD and also replicated the previously seen IL-6R effect (hazard ratio 0.60). The authors suggest that ascertainment bias or random sampling error could have a role in driving these findings. However, technical factors may also contribute. A contemporaneous analysis of the same UKB dataset with stricter CH classification criteria found a stronger association between CH and CVD (hazard ratio 1.22) and replicated the IL-6R effect[13]. When considering this and other discrepancies in the literature, it is important to consider both how disease phenotypes are defined and how CH is identified.

Although the UKB and other large-scale electronic health record (EHR)-linked cohort studies are powerful tools for studying CH, variation in methodology will certainly lead to conflicting reports in the years to come. First, within EHR-linked biobanks, there are many sources for similar types of information available, which can lead to multiple operational definitions of human traits and diseases. For example, within the UKB, coronary artery disease can be defined using several sources including but not limited to: self-reporting; International Classification of Diseases (ICD) codes from primary care records, procedures, hospitalizations, death records, or a combination thereof. In addition, agreed standards for selecting and combining ICD-coded data to define common diseases and their clinically relevant subtypes are lacking. Differences in operationalization of coronary artery disease within the UKB has been shown to influence the magnitude of the association between mortality and a polygenic risk score for the disease[14], and would also be expected to influence association with CH. Initiatives such as Health Data Research UK (HDR UK) are needed to offer guidance to researchers on defining prevalent and incident disease in EHR-based studies. In the absence of formal guidelines, sensitivity analyses designed to study the effect of variation in disease definitions on associations between CH and outcomes should be presented to facilitate cross-study comparison and reproducibility.

CH variant calling is another major source of variability that can lead to differences in results and conclusions about associations between CH and health outcomes. This is particularly challenging for medium depth (30–50×) exome- and genome-sequencing data from the UKB and other large cohorts that were designed for germline variant calling. Most CH mutations are present at low allelic fractions (for example, <10% variant allele frequency) and distinguishing sequencing artifacts from bona fide mutations is difficult. Furthermore, given the absence of matched data from a non-blood tissue, distinguishing CH variants from rare germline genetic variants is also challenging. Standard somatic variant callers are not optimized for CH detection and additional post-variant calling filtering must be used to remove artifacts and germline variants. Differences in filtering after variant calling, even among groups using the same variant caller, can lead to differences in the frequency of CH in the UKB. For example, CH was detected in 5.5% of individuals by Kar et al.[15], 6% by Kessler et al.[9] and 3.4% by Vlasschaert et al.[8] in the UKB. There is no consensus about best practices for CH variant calling, and variation in methodology influences association results. This is made worse by a lack of benchmarking datasets similar to those available through the Genome in a Bottle Consortium for germline variant calling. Reference datasets targeted to mutations at low variant allele fractions, ideally tumor-normal dilutions, would facilitate cross-comparison and validation of calling methods, and help to improve reproducibility.

Overall, this study is a valuable contribution to the CH field. It greatly expands our knowledge on germline genetic predisposition to CH. Future studies that further characterize the genes and pathways involved in the initiation, maintenance and expansion of CH are needed to provide therapeutic targets and important insights into the earliest stages of carcinogenesis. Additional studies in non-European populations will be crucial to fully characterize germline CH predisposition. Finally, and most importantly, this study highlights both the promise and complexity of the UKB and other large-scale EHR-linked biobanks for advancing our scientific knowledge of the genetic basis of human disease.

J. Scott Beeler [1], **Alexander G. Bick** [2] **& Kelly L. Bolton** [3] ✉
[1]Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. [2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [3]Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA.
✉e-mail: bolton@wustl.edu

Published online: 19 December 2022

References

1. Fry, A. et al. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
2. Jaiswal, S. et al. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
3. Genovese, G. et al. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
4. Jaiswal, S. et al. *N. Engl. J. Med.* **377**, 111–121 (2017).
5. Dorsheimer, L. et al. *JAMA Cardiol.* **4**, 25 (2019).
6. Miller, P. G. et al. *Blood* **139**, 357–368 (2022).
7. Bolton, K. L. et al. *Nat. Commun.* **12**, 5975 (2021).
8. Vlasschaert, C. et al. Preprint at medRxiv https://doi.org/10.1101/2022.10.21.22281368 (2022).
9. Kessler, M. D. et al. *Nature* https://doi.org/10.1038/s41586-022-05448-9 (2022).

10. Weinstock, J. S. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.12.10.471810 (2021).
11. Bick, A. G. et al. *Circulation* **141**, 124–131 (2020).
12. Agrawal, M. et al. *Blood* **140**, 1094–1103 (2022).
13. Vlasschaert, C. J., Heimlich, B., Rauh, M. J., Natarajan, P. & Bick, A. G. *Circulation* https://doi.org/10.1161/CIRCULATIONAHA.122.062126 (2023).
14. Patel, R. S. et al. *PLoS ONE* **17**, e0264828 (2022).
15. Kar, S. P. et al. *Nat. Genet.* **54**, 1155–1166 (2022).

Competing interests

The authors declare no competing interests.