

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Blood derived whole-exome sequences (WES) using the Illumina HiSeq 2000 platform were available for 10,881 baseline ARIC participants (ARIC sub-study phs000668 CHARGE-S37 https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000668.v6.p2). WES data for 4,233 ARIC visit 5 samples using the Illumina NovaSeq 6000 platform were generated at the Baylor College of Medicine Human Genome Sequencing Center. Data used in this study was available through approved ARIC Ancillary Study proposal MS # 4130.
Data analysis	Somatic mutations from WES were identified using Mutect2 software in the Terra platform (https://portal.firecloud.org/?return=terra#methods/gatk/mutect2-gatk4/21). To minimize potential artifacts in the Mutect2 calls, a panel-of-normal (PON) was created from 100 random HiSeq WES from the youngest participants, while 1000 Genomes PON (gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz) was used for the NovaSeq WES. Besides, Genome Aggregation Database (gnomad: gs://gatk-best-practices/somatic-hg38/af-only-gnomad.hg38.vcf.gz) was used to limit germline variants in the somatic mutation call. ANNOVAR software (https://annovar.openbioinformatics.org/en/latest/) was used for annotation, and clonal hematopoiesis was ascertained using a publicly available pipeline (https://app.terra.bio/#workspaces/terra-outreach/CHIP-Detection-Mutect2/). Hotspot driver mutations in U2AF1 gene were identified using a custom script (https://github.com/MMesbahU/U2AF1_pileup). All statistical analyses were performed using R software (version 4.1 or 4.2.2). Logistic regression was performed using 'glm' function, two-sided Student's t-Test using 't.test' function, two-sided Wilcoxon Rank Sum using 'wilcox.test' function, and Pearson's Chi-squared test using 'chisq.test', Pearson's r using 'cor.test' function. R base package, and packages such as 'ggplot2', and 'meta', were used for plotting. Targeted single-variant association was performed using REGENIE v3.1.3.gz (https://rgcgithub.github.io/regenie/). GWAMA v2.2.2 (https://www.greenivaramu.ee/tools/GWAMA_v2.2.2.zip) was used to perform inverse-variance weighted fixed-effect meta-

analysis.

PLINK (version v2.00a3LM; <https://www.cog-genomics.org/plink/2.0/>) "score" was used to derive polygenic risk score for prevalent clonal hematopoiesis using the variant weights from Kessler et al. 2022.

The complete scripts used in this study are available at <https://doi.org/10.5281/zenodo.13357233>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

To protect the privacy of research participants and the confidentiality of their data while ensuring that these data are available for appropriate use by researchers, all raw data used in this study are available via controlled access. Individual-level phenotypes and whole-exome sequencing data from the ARIC baseline visits participants (used in this study) are available in dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>) accession code phs000668 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000668.v6.p2). Affymetrix 6.0 genome-wide association array dataset is available in dbGaP accession phs000557.v7.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000557.v7.p2). Imputed genotype data from the ARIC sub-study GENEVA is available in dbGaP (accession phs000090 "phg000248" https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000090.v8.p2). ARIC Visit-5 WES (generated in this study) and individual-level data are available via controlled access ancillary study proposal (https://aric.csc.unc.edu/aric9/researchers/Obtain_Submit_Data). The timeline for the approval process ranges from 3-6 weeks for ARIC ancillary studies, with specific criteria and proposal forms available at <https://sites.csc.unc.edu/aric/ancillary-studies-pfg>. All other data supporting the findings described in this manuscript are available in the article and its Supplementary Information files.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

ARIC study included self-reported male and female sexes; described in manuscript text, and in Table 1, and Supplementary Table 1. In multivariable regression analyses, self-reported sex (Male/Female) was included as a covariate, with 'Female' as the reference category. Effect estimates for 'Male' were reported relative to 'Female'.

Reporting on race, ethnicity, or other socially relevant groupings

ARIC study included self reported race information (African American, AA, and European American, EA) for the study participants. Sample breakdown is described in the manuscript text, and in Table 1, and Supplementary Table 1. In multivariable regression analyses, self-reported race (AA/EA) was included as a covariate, with 'AA' as the reference category. Effect estimates for 'EA' were reported relative to 'AA'.

GWAS Data: ARIC GWAS data was available stratified by AA and EA participants. Single variant analysis and PRS calculation were performed separately; results from stratified and meta analyses was reported.

Population characteristics

Among 4,187 ARIC study participants with longitudinal WES, 2,478 (59.2%) were female, 951 (22.7%) were African American, and the mean (SD) age was 55.5 (5.5) years at the time of baseline blood draw for WES. The mean (SD) age of participants at follow-up blood draw was 75.8 (5.2) years. Other population characteristics presented in Table 1, and Extended Data Table 1.

Recruitment

We analyzed data on individuals previously recruited by the Atherosclerosis Risk in Communities (ARIC) Study, described in https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000280.v8.p2. The cohort component of the ARIC study began in 1987. Each of the four ARIC Field Centers (Washington County, MD; Forsyth County, NC; Jackson, MS; and Minneapolis, MN) randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from their community. A total of 15,792 participants received an extensive examination, including medical, social, and demographic data. These participants were examined with the baseline visit occurring in 1987-89, the second visit in 1990-92, the third visit in 1993-95, the fourth visit in 1996-98, the fifth visit in 2011-13, the sixth visit 2016-17 and the seventh visit 2018-19. Follow-up occurs yearly by telephone to maintain contact with participants and to assess health status of the cohort.

Ethics oversight

Informed consent for genetic studies was obtained from all participants, and study protocols were approved by the respective institutional review board (IRB) for each cohort (University of Mississippi Medical Center Institutional Review Board (ARIC: Jackson Field Center); Wake Forest University Health Sciences Institutional Review Board (ARIC: Forsyth County Field Center); University of Minnesota Institutional Review Board (ARIC: Minnesota Field Center); and Johns Hopkins University School of Public Health Institutional Review Board (ARIC: Washington County Field Center). Each study received institutional certification before depositing sequencing data into dbGaP, ensuring approval by all relevant institutional ethics committees and compliance with relevant ethical regulations.

Secondary use of dbGaP data are approved by Massachusetts General Hospital IRB protocols 2016P002395 and 2016P001308.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	10,881 ARIC participants with WES data at baseline; 4,233 participants with WES at visit 5 (longitudinal visit); 4,187 participants with WES at baseline and longitudinal visits. Imputed GWAS data and WES were available for 3,014 participants.
Data exclusions	Individuals with prevalent hematologic malignancy, or missing WES/phenotype/covariates were excluded from the analysis.
Replication	To analyze the concordance between clonal hematopoiesis of indeterminate potential (CHIP) ascertained from sequences generated by the HiSeq 2000 or NovaSeq 6000 platforms (Illumina, Inc., CA), we re-sequenced 786 baseline samples using the NovaSeq 6000 platform (Illumina, Inc., CA). Concordance estimates for CH clones detected “yes/no” by the two sequencing platforms were 83% (654/786) and 93% (731/786) at variant allele fraction (VAF) $\geq 2\%$ and VAF $\geq 10\%$, respectively, with a strong correlation (Pearson’s $r=0.80$) between the VAF estimates from these two platforms.
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.