

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 (n) 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. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P 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 d , Pearson's r), 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	Data were obtained from two large-scale population genomics studies at the Mayo Clinic: the Mayo Clinic Biobank (MCB) and the Tapestry Study. The data were exported in XLS or CSV format and subsequently analyzed using R.
Data analysis	Standard statistical analyses and data visualizations were performed using R version 4.2.2, along with the following packages: poLCA (v1.6.0.1), poLCAParallel (v1.2.2), msigdbr (https://igordot.github.io/msigdbr/), dbscan (v1.2-0), tidycmprsk (v1.1.0), survival (v3.7-0), and survminer (v0.5.2). For reproducibility, the analysis code is publicly available at GitHub (https://github.com/tpriya14/MASLD_Subgroups/) and archived at Zenodo (https://doi.org/10.5281/zenodo.19562094).

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](#)

The individual-level data analyzed in this study are not publicly available due to participant privacy protections and institutional usage restrictions. Access to the Mayo Biobank data can be requested by contacting Biobank@mayo.edu. In accordance with informed consent provisions, all projects must include a Mayo Clinic investigator as a collaborator. Additional information is available on the Mayo Clinic Biobank website: <https://www.mayo.edu/research/centers-programs/mayo-clinic-biobank/overview>. The Tapestry Study is supported by the Mayo Clinic Center for Individualized Medicine, with further information available at: <https://www.mayo.edu/research/centers-programs/center-individualized-medicine/research/clinical-studies/tapestry>. Requests for access to Tapestry Study data may be directed to Dr. Konstantinos N. Lazaridis (larazidis.konstantinos@mayo.edu). Access to the deidentified Tapestry Study data requires appropriate legal agreements and approval from Helix. For both cohorts, requests will be acknowledged within 4 weeks of receipt. If approved, data will be provided within 6 months following the execution of a data transfer agreement. The duration of data availability is determined by the terms of the agreement. The polygenic scores used in this study are publicly available through the Polygenic Score (PGS) Catalog (<https://www.pgscatalog.org>) under the following accession codes: PGS002283 (<https://www.pgscatalog.org/score/PGS002283/>), PGS000872 (<https://www.pgscatalog.org/score/PGS000872/>), PGS003096 (<https://www.pgscatalog.org/score/PGS003096/>), PGS003400 (<https://www.pgscatalog.org/score/PGS003400/>), PGS000863 (<https://www.pgscatalog.org/score/PGS000863/>), and PGS003442 (<https://www.pgscatalog.org/score/PGS003442/>). Previously published databases or datasets used in this study are publicly available as follows: gnomAD (<https://gnomad.broadinstitute.org/>, v4.1.0); 1000 Genomes Project (<https://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/>, downloaded on 2015-02-16 and lifted over to hg38); ClinVar (<https://ftp.ncbi.nlm.nih.gov/pub/clinvar/>, release 2024-02). HGMD Professional data are available under license (<https://www.hgmd.cf.ac.uk/ac/index.php>, release 2024-01). Human MSigDB (<https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp>, v2025.1.Hs) was used for pathway enrichment analysis. The pathway gene sets in MSigDB were curated from WikiPathways (<https://www.wikipathways.org/>) and REACTOME (<https://reactome.org/>). Source data are provided with this paper.

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

Sex information for this study was obtained directly from the electronic health records (EHR) for both cohorts and was determined based on self-reporting.

Reporting on race, ethnicity, or other socially relevant groupings

Race and ethnicity information for this study was obtained directly from the EHR for both cohorts. Race was categorized as White, Black or African American, American Indian/Alaskan Native, Asian and other. Ethnicity was categorized as Hispanic or Latino, non Hispanic or Latino.

Population characteristics

We created two MASLD cohorts, with cohort 1 from MCB and cohort 2 from Tapestry Study. The MCB (Cohort 1) included 5,544 MASLD patients (55.4% female; average age: 70.9 years) and 11,080 healthy controls (55.4% female; average age: 69.1 years). The Tapestry Study cohort (Cohort 2) comprised 4,079 MASLD patients (60.8% female; average age: 59.6 years) and 8,158 healthy controls (60.8% female; average age: 59.6 years). In both case cohorts, MASLD patients had a higher proportion of females, elevated body mass index (BMI), increased levels of alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), hemoglobin A1c (HbA1c), and triglycerides (TG), as well as a higher prevalence of comorbidities compared to controls.

Recruitment

We used data from two large-scale population genomics studies in the Mayo Clinic, the Mayo Clinic Biobank (MCB), and the Tapestry Study. Both are prospective cohort studies with EHR-linked deep phenotypic data. The studies were approved by the Mayo Clinic Institutional Review Board (IRB #09-007763 for MCB and IRB #19-000001 for Tapestry Study), and all participants provided informed consent for the use of their data in genetic and health research. Similar to the UK Biobank, MCB is a convenience-based biobank. MCB is composed of over 50,000 voluntary primary care patients recruited mostly between 2009 and 2016. The Tapestry Study was designed to assess the impact of genomic data on healthcare utilization and is enriched for chronic disease and cancer; participants were primarily recruited from specialty clinics between 2020 and 2024. At the time of this study, MCB and the Tapestry Study included 51,883 and 51,898 adult participants, respectively. MASLD cases were identified using International Classification of Diseases, 9th Revision (ICD-9)/10th Revision (ICD-10), along with a previously curated list of MASLD patients. In the first step, patients were identified from the EHR problem lists and billing codes using ICD-9 (571.8, 571.9, and 571.5) and ICD-10 (K76.0, K75.8, K74.60, and K74.69) codes. In the second step, we incorporated an existing list of MASLD patients from the Mayo Clinic database based on abdominal imaging reports and in-depth medical record review. Finally, patients with diagnosis codes for competing etiologies of liver disease (e.g., viral hepatitis, alcohol-related liver disease) were removed from further analysis. Patients identified through both approaches were combined and used as the final MASLD cases. A total of 5,544 and 4,079 individuals were identified as MASLD cases in the MCB and Tapestry Study, respectively. In addition, controls were selected from individuals without ICD-9/ICD-10 codes for MASLD and with no evidence of hepatic steatosis, cirrhosis, or other chronic liver disease. All individuals identified as MASLD cases were excluded. To ensure that controls represent a population with minimal risk for MASLD, subjects diagnosed with metabolic syndrome, hyperlipidemia, or diabetes were also excluded. The remaining individuals were matched to the MASLD cases at a 1:2 ratio using the MatchIt R package, based on age and exact matching on sex and race, which generated a list of 11,080 controls for MCB. A similar approach yielded 8,158 matched controls from the Tapestry Study.

Ethics oversight

All participants provided informed consent for the use of their data in genetic and health research.

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	The MCB (Cohort 1) initially included 5,544 patients with MASLD patients and 11,080 controls. The Tapestry Study (Cohort 2) comprised 4,079 MASLD patients at the time of this study and 8,158 controls. After excluding cases with missing values, 2,042 MASLD cases from MCB and 2,560 cases from the Tapestry Study were retained. The MCB cohort was randomly split at a 75:25 ratio into a development cohort (MCB 1A; n = 1,532) and a validation cohort (MCB 1B; n = 510), and the Tapestry Study cohort served as an independent validation cohort.
Data exclusions	A total of 77 cases and 27 controls overlapped between the two cohorts, which were excluded from the Tapestry Study. Four variables (C-reactive protein, prothrombin time, Hepatitis B Core (HBC) total antibody test, Gamma-glutamyl transferase) with more than 25% missing data were excluded from further analysis. After filtering cases with any missing values, the final list included 2,042 cases from MCB and 2,560 cases from Tapestry Study. Among controls, 1,245 (MCB) and 1,421 (Tapestry Study) participants were retained after the filtering.
Replication	The findings from the development cohort (Mayo Clinic cohort 1A, n=1,532) was validated in the Mayo Clinic validation cohort (cohort 1B, n=510) and Tapestry validation cohort (n=2,560).
Randomization	Since no intervention was planned for this study, randomization was not applicable.
Blinding	Since no intervention was planned for this study, blinding was not applicable.

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

Methods

n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☒	☐ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☒	☐ Clinical data		
☒	☐ Dual use research of concern		
☒	☐ Plants		

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.