nature portfolio

Corresponding author(s): Professor Jonathan Fallowfield

Last updated by author(s): Apr 14, 2023

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 *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	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 *Give P values as exact values whenever suitable.*
☒	☐	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	A custom Python (version 3.9) script was developed to carry out Electronic Health Record data processing tasks.
Data analysis	When scoring the liver pathology, for ordinal variables, Light's kappa (square weighted, for >2 raters) was calculated using the 'psy' package (version 1.2), and intraclass correlation coefficient, Krippendorf's alpha, and Kendall's W were calculated using the 'irr' package (version 0.84.1) in R (version 4.1.0). Q Path open source software (version 0.2.3) was used for liver histopathological image analysis. Analysis using clinical and histopathological data only was undertaken in R (version 4.1.0) using the packages 'survival' (version 3.2-1), 'survminer' (version 0.4.9), and 'finalfit' (version 1.0.5). NeoGenomics used a proprietary deep learning-based workflow NeoLYTX to identify individual liver cells and perform cell classification for cell markers. For RNA-seq analysis, the following software packages were used in R (version 4.1.2) : Reads were trimmed using 'Cutadapt' (version cutadapt-1.9.dev2) and aligned to the reference genome using 'STAR' (version 2.5.2b). Reads were assigned to features using 'featureCounts3' (version 1.5.1) with a igtf file from Ensembl (annotation version 84). Differential gene expression analysis was performed using limma-voom' (version 3.28.14); Gene Set Enrichment Analysis (GSEA) was performed with GSEA function from 'clusterProfiler' (version 4.0.5); data were visualized with 'ggplot2' (version 3.3.5) and 'clusterProfiler'; Cox regression was performed using 'glmnet' (version 4.1-4), and time-dependent ROC curves were created by the 'timeROC' package (version 0.4); Kaplan-Meier analysis was performed using 'survival' (version 3.4-0) and 'survminer' (version 0.4.9) packages. Genome Analysis Toolkit (GATK, version 4.0.1.2) was used to call genotypes. The MUlti-Subject Single Cell ('MuSiC') deconvolution tool (version 0.1.1) was run using R (version 3.6.3) for deconvolution analysis. The R package`ppcor' (version 1.1) was used to assess the correlation between the proportion of hepatic cell subtypes and the histological score or clinical outcomes. Transcriptional network inference and regulon analysis was undertaken in R (version 4.1.0) using the 'RTN' package (version 2.16.0), the

'RTNsurvival' package (version 1.16.0), 'maxstat' package (version 0.7-25), and 'Mfuzz' package (version 2.52.0).
An R Shiny app was used to develop the gene browser.
R scripts enabling the main steps of the analysis are available from the corresponding author on reasonable request.

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

Hepatic bulk RNA-seq data is deposited in the European Nucleotide Archive (https://www.ebi.ac.uk/ena; study accession number: PRJEB58625). Gene expression data is also freely available for user-friendly interactive browsing online at https://shiny.igc.ed.ac.uk/SteatoSITE_gene_explorer/. SteatoSITE has delegated ethics allowing the granting of access to the full dataset within the PMS-IC secure environment to third parties by application (details at https://steatosite.com/researchers!), overseen and reviewed by the SteatoSITE Scientific Advisory Board.

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' is reported as stated in the Electronic Health Record. Both males and females are included in the cohort. Sex was used as a covariate in analysis as there are sex-related differences in NAFLD prevalence and outcome.
Reporting on race, ethnicity, or other socially relevant groupings	Ethnicity and SIMD (Scottish Index of. Multiple Deprivation) data were collected from administrative datasets and are reported as ethnicity and social deprivation influence NAFLD prevalence and outcome.
Population characteristics	Covariate relevant population characteristics include age, sex, ethnicity, SIMD, diabetes status, body mass index, liver fibrosis stage, and genotypic (SNP) status.
Recruitment	This was a retrospective, observational study. Initial case selection was based on the availability of archival liver tissue (from biopsies, resections, or explants that were surplus to diagnosis) in formalin-fixed paraffin-embedded (FFPE) blocks available within the NHS Research Scotland Biorepository network, with the clinical and/or histological diagnosis of NAFLD, and meeting specific inclusion/exclusion criteria. Using a secondary care tissue-first selection process introduces spectrum bias and this is acknowledged in the discussion. This is a strength in terms of outcome enrichment but means that SteatoSITE will have less value for modelling the population-level natural history of NAFLD.
Ethics oversight	Anonymised tissue was supplied after approval by the National Health Service Research Scotland (NRS) Biorepository network. Unified transparent approval for data inclusion in this pan-Scotland project was provided by the West of Scotland Research Ethics Committee 4 (Reference: 20/WS/0002; 18th February 2020), Public Benefit and Privacy Panel for Health and Social Care (PBPP; Reference: 1819-0091; 4th June 2021), Institutional Research & Development departments and Caldicott Guardians.

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

Life sciences study design

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

Sample size	No sample size calculation was performed. The cohort size is a reflection of the maximum number of eligible cases across all of the Scottish Biorepositories at the time of data collection. To our knowledge, this is the largest collection of NAFLD cases with hepatic RNA-sequencing, digital pathology and linked clinical outcomes worldwide.
Data exclusions	There were pre-determined Quality Control criteria for RNA-sequencing (including RNA yield (and any potential DNA contamination) and DV200). Samples with DV200 below 30% were not progressed for sequencing but were included in other analyses (e.g., histopathological assessment).

Replication	We have established a unique resource to be used by the liver research community and to catalyze new discoveries in NAFLD. We present initial analyses to illustrate the utility of SteatoSITE and acknowledge that transcriptional risk scores, for example, will require external validation if suitable cohorts can be identified
Randomization	There were no randomisation procedures employed - this was a retrospective observational study. For RNA-seq analysis, principal component analysis (PCA) was performed to identify covariates that significantly correlated with the main principal components, so they could be controlled for downstream analyses. For this reason, sex was included as an additive effect in the linear model used for differential expression
Blinding	All cases were assigned a unique study ID (and the key only held by the NRS Biorepositories). Histopathological assessors and RNA-sequencing analysts were blinded to any patient information. Bioinformaticians only accessed the clinical outcome data after histopathological scoring and RNAseq analysis had been performed to enable time-to-event analysis/risk prediction etc.

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

Antibodies

Antibodies used	Antibodies for MultiOmyx analysis: by staining order, were rabbit anti-TREM2 (polyclonal, ProteinTech), mouse anti-MNDA (253A, Abcam), rabbit anti-CD9 (EPR2949, Abcam), mouse anti-CD66b (G10F5, BioLegend), mouse anti-CD11B (238439, R&D Systems), rabbit anti-DC-SIGN (D7F5C, Cell Signaling Technology), rabbit anti-Ki67 (SP6, Abcam), rabbit anti-IDO1 (SP260, Abcam), rabbit anti¬CD11c (D3V1E, Cell Signaling), rabbit anti-PD-L1 (SP142, Abcam), rabbit anti-CD14 (EPR3652, Abcam), mouse anti-CD16 (DJ130c, ThermoFisher Scientific), mouse anti-CD68 (KP1, BioLegend), mouse anti-CD163 (EDHu-1, Bio-Rad), mouse anti-HLA DQ/DR/DP (WR18, Novus), mouse anti-CD33 (44M12D3, Novus Biologicals), mouse anti-SMA (1A4, Sigma-Aldrich).
Validation	The MultiOmyx protocol was developed in collaboration with the company NeoGenomics (https://neogenomics.com/) and all antibodies used had been optimised and validated before use.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This was not a clinical trial.
Study protocol	This was not a clinical trial. Full methodological details are provided in the manuscript.
Data collection	A total of 940 cases from the three participating NHS Scotland Biorepositories (Lothian, Greater Glasgow & Clyde, and Grampian) were included, representing the full histological spectrum from normal liver tissue to NAFLD-related cirrhosis. Cases with a liver tissue sample acquired between January 2000 and October 2019 were selected. All patients were years of age at the tissue sampling date. Data from Electronic Health Records and national datasets were retrieved, where available, from a period between ten years before the tissue sampling date until May 2020.
Outcomes	We collected all relevant clinical outcomes according to recent expert consensus guidelines for using administrative coding in Electronic Health Record-based research of NAFLD.