Decoding MASLD Progression: A Molecular Trajectory-Based Framework for Modelling Disease Dynamics

Ioannis Kamzolas[1,4*], Thodoris Koutsandreas[1,2*], Charlie George Barker[1,3], Anna Vathrakokoili Pournara[1], Harry Weston[1,4], Michele Vacca[4,5], Irene Papatheodorou[1,6], Antonio Vidal-Puig[4,7 +], Evangelia Petsalaki[1+]

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK

[2]Open Targets, Wellcome Genome Campus, Hinxton, CB10 1SD, UK

[3](current address) University College London Cancer Institute, London WC1E 6DD, UK

[4]MRC Institute of Metabolic Science, Metabolic Research Laboratories, University of Cambridge, Cambridge, UK

[5]Interdisciplinary Department of Medicine, Università degli Studi di Bari "Aldo Moro", Bari, Italy; Roger Williams Institute of Liver Studies, Foundation for Liver Research, London, SE5 9NT, UK

[6] (current address) Earlham Institute, Norwich Research Park, Colney Ln, Norwich NR4 7UZ

[7]Centro de Investigacion Principe Felipe, Valencia, Spain

+ correspondence should be addressed to Antonio Vidal-Puig ajv22@medschl.cam.ac.uk and Evangelia Petsalaki: petsalaki@ebi.ac.uk

* These authors contributed equally to the work presented in this manuscript

Abstract

Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) has emerged as a silent pandemic, affecting nearly one-third of the global population. This condition encompasses a spectrum of liver disorders, ranging from simple steatosis to Metabolic Dysfunction-Associated Steatohepatitis (MASH), which is characterised by liver steatosis, lipotoxicity, hepatocellular damage, inflammation, and fibrosis. Left unchecked, MASLD/MASH can progress to cirrhosis and hepatocellular carcinoma. Despite the progressive nature of MASLD/MASH, current research primarily relies on static, histopathologically defined stages, which fail to capture the dynamic disease continuum.

In this study, we present an integrative stratification approach that combines patient pseudo-temporal ordering, network analysis, and cell-type deconvolution to map the continuous disease trajectory. By analyzing transcriptomic profiles, we predict patients' positions along this trajectory, moving beyond conventional stage-based classifications. This approach reveals the sequence of critical molecular events driving MASLD/MASH progression, providing new insights into the disease's pathophysiology. Furthermore, we identify novel trajectory-specific biomarkers that support a more refined, personalised strategy for managing MASLD. This work highlights the potential of trajectory-based frameworks in advancing our understanding and treatment of complex metabolic diseases.

Introduction

Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) encompasses a spectrum of conditions, ranging from simple fat accumulation (steatosis) to Metabolic Dysfunction-Associated Steatohepatitis (MASH), which can progress to fibrosis, cirrhosis, and hepatocellular carcinoma[1]. This global epidemic affects nearly one-third of Western populations and is primarily driven by the rising prevalence of metabolic syndrome, type 2 diabetes, and obesity[2,3]. Despite its significant burden, most MASLD cases remain asymptomatic and undiagnosed until advanced stages.

The diagnosis of MASLD typically relies on imaging techniques such as ultrasound and MRI spectroscopy alongside non-invasive fibrosis markers like the Fibrosis-4 index (FIB-4), with liver biopsy serving as the gold standard[4]. Although histological scoring reproducibility among

53 pathologists has improved[5], variability in sampling[6,7] and the invasiveness of biopsies[6,7]
54 underscore the urgent need for complementary non-invasive diagnostic methods.

55 Elastography techniques, including transient elastography (FibroScan) and magnetic resonance
56 elastography (MRE), have emerged as valuable tools for non-invasive fibrosis assessment. While
57 MRE offers high accuracy, its cost and limited accessibility are significant barriers. Conversely,
58 transient elastography is more affordable but less precise. Despite the availability of various
59 blood-based imaging and genetic biomarkers for detecting steatosis and fibrosis, their accuracy
60 remains suboptimal, and reliable options for diagnosing MASLD specifically are still lacking[4].

61 Advances in systems biology and omics technologies have propelled MASLD/MASH research
62 forward[4,8–11], linking the disease to metabolic syndrome, insulin resistance, and mechanisms such
63 as lipid accumulation, endoplasmic reticulum stress, and oxidative stress[12]. Genome-Wide
64 Association Studies (GWAS) have further highlighted genetic factors, such as the PNPLA3 SNP
65 rs738409 variant, which is associated with increased liver fat and heightened disease risk [13–15].
66 Despite these breakthroughs, the complexity of MASLD demands further research to elucidate
67 progression mechanisms and distinguish benign cases from those with higher mortality risks.

68 Histology has long been a cornerstone for understanding MASLD pathophysiology, but its
69 limitations as a sole proxy for disease mechanisms are increasingly evident. Pseudo-temporal
70 ordering, initially developed for microarray data[16] and now widely applied in single-cell genomics,
71 provides an alternative framework by arranging cells or patients along trajectories based on their
72 molecular states. This approach has recently been used to study Alzheimer's disease progression
73 through brain expression data[17,18], and offers a promising, histology-independent method for
74 investigating MASLD progression via transcriptomic profiles. By complementing traditional
75 assessments, pseudo-temporal ordering provides a data-driven perspective on the molecular
76 changes underpinning disease progression.

77 In this study, we applied pseudo-temporal ordering to MASLD/MASH patient data (Fig. 1A). We
78 integrated it with network analysis (Fig. 1B) and cell-type deconvolution (Fig. 1C). This
79 comprehensive approach allowed us to identify key molecular events specific to disease stages
80 and cell types. Additionally, by integrating our molecular map of MASLD progression with plasma
81 proteomics data, we identified liver-specific circulating biomarkers that can position patients along
82 the disease trajectory independently of histology. These findings not only provide new insights
83 into MASLD pathophysiology but also prioritize potential therapeutic targets associated with
84 disease progression (Figs. 1D–E).

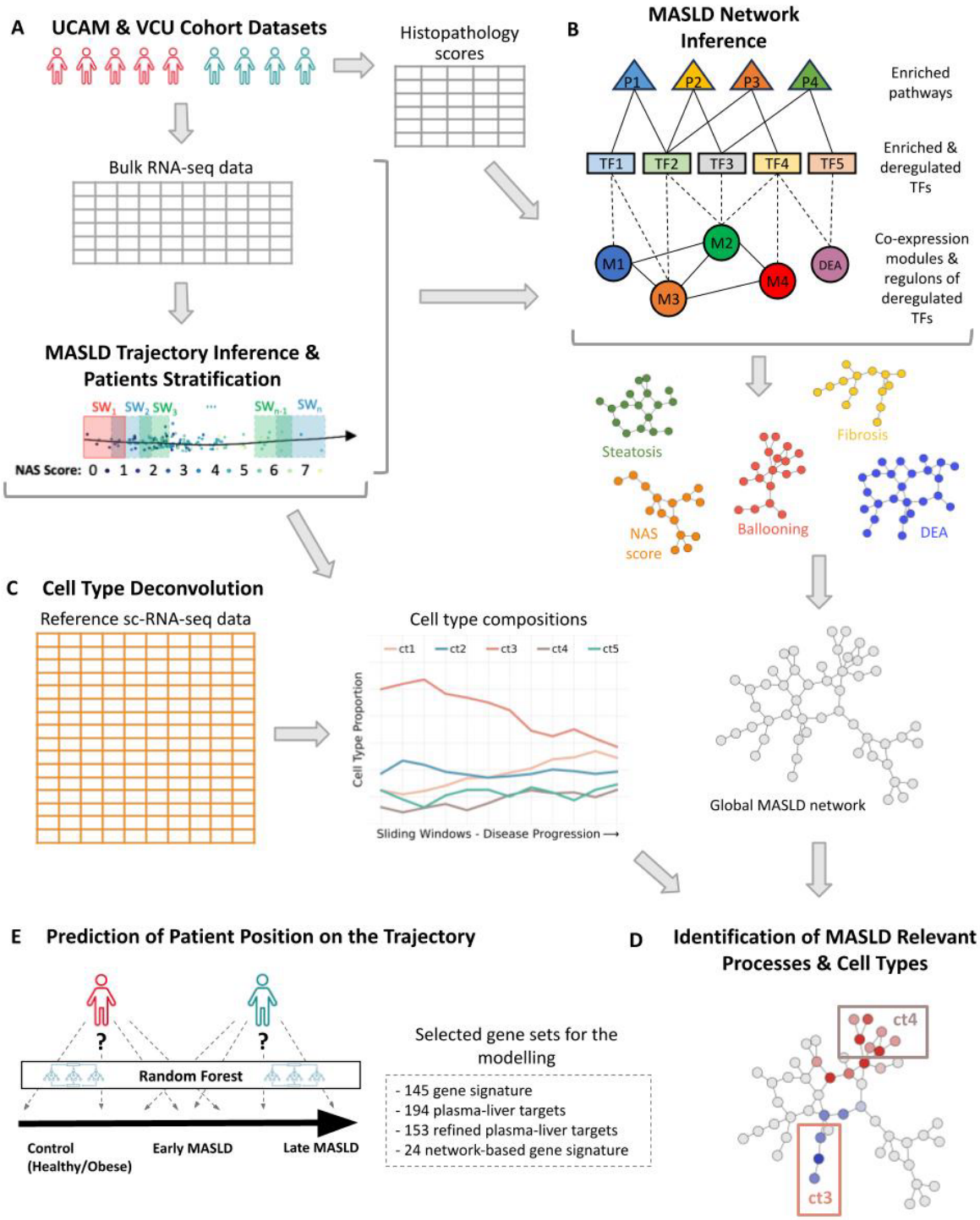


Figure 1. Overview of approach used in this study to investigate MASLD progression. A. *Integration of public datasets and MASLD trajectory inference by pseudo-temporal ordering of patient biopsy transcriptomic data* ***B.*** *Combining phenotype-specific gene co-expression analysis with upstream pathway inference to define a MASLD network.* ***C.*** *Cell type deconvolution to study the dynamics of cell type composition changes.* ***D.*** *Random Forest regression model for the prediction of patients' position on the inferred MASLD trajectory based on a set of biomarkers. Training has been done on the initial dataset and testing on independent datasets. The model deployed different gene signatures, while random gene sets were used as controls to validate the high model's performance.* ***E.*** *Identification of deregulated cell type-specific processes and potential biomarker genes that could be targeted in specific disease stages.*

85

Results

Transcriptomics-based disease trajectory analysis captures MASLD progression.

We hypothesised that MASLD progression could be modelled as a continuum of patient states. To test this, we analysed RNA-Seq data from 136 patients[9,19] (Methods; Supplementary Table 1 & 2). After excluding one outlier, we applied pseudo-temporal ordering[20] to construct a transcriptomics-driven trajectory (Methods; Fig. 2A; Supplementary Fig. 1). This trajectory showed strong correlations with histopathology scores, including steatosis, ballooning, inflammation, fibrosis, and the NAS score, with Pearson R coefficients ranging from 0.96 to 1.0. These findings confirmed alignment between the transcriptomics-derived trajectory and established disease stages (ANOVA followed by Tukey's pairwise comparisons; Supplementary Fig. 2). (Methods; Supplementary Tables 1 & 2).

To validate the trajectory model, we tested it on two independent RNA-Seq datasets: the EPoS dataset, which includes healthy normal-weight and obese individuals, and the Gubra dataset, comprising MASLD and MASH patients (Supplementary Figs. 3A-B)[8,21]. In the EPoS dataset, the trajectory successfully captured disease stage progression. In the Gubra dataset it effectively distinguished healthy and obese samples from MASLD and MASH patients. However, additional variance in both datasets was observed, which did not fully correlate with disease progression.

To address this variability, we identified genes contributing to variance along the MASLD/MASH progression axis. Using a random forest approach, we identified 145 genes that significantly predicted MASLD stage and histopathological features (AUC: 62–73%; Supplementary Figs. 3C-D, Supplementary Table 3). These genes also recapitulated the disease trajectory in the independent datasets[8,21], demonstrating improved linearity and better differentiation of MASLD/MASH stages compared to the full transcriptome (Figs. 2B-C, Supplementary Fig. 3E). This underscores the value of targeted gene sets in refining trajectory models and improving stage prediction.

The identified genes were enriched in pathways central to MASLD/MASH progression, including ECM-receptor interaction and focal adhesion (associated with fibrosis), AGE-RAGE and IL-17 signalling (inflammatory responses), and TGF-beta and PI3K-Akt signalling (wound healing and

metabolic dysregulation). Other pathways, such as glycerolipid metabolism and cytokine-receptor interaction, reflect the metabolic and immune dysregulation characteristic of MASLD/MASH (Fig. 2D). To further elucidate MASLD/MASH progression mechanisms, we developed a network-based strategy, integrating molecular interactions and pathway analyses to identify critical drivers of disease states.

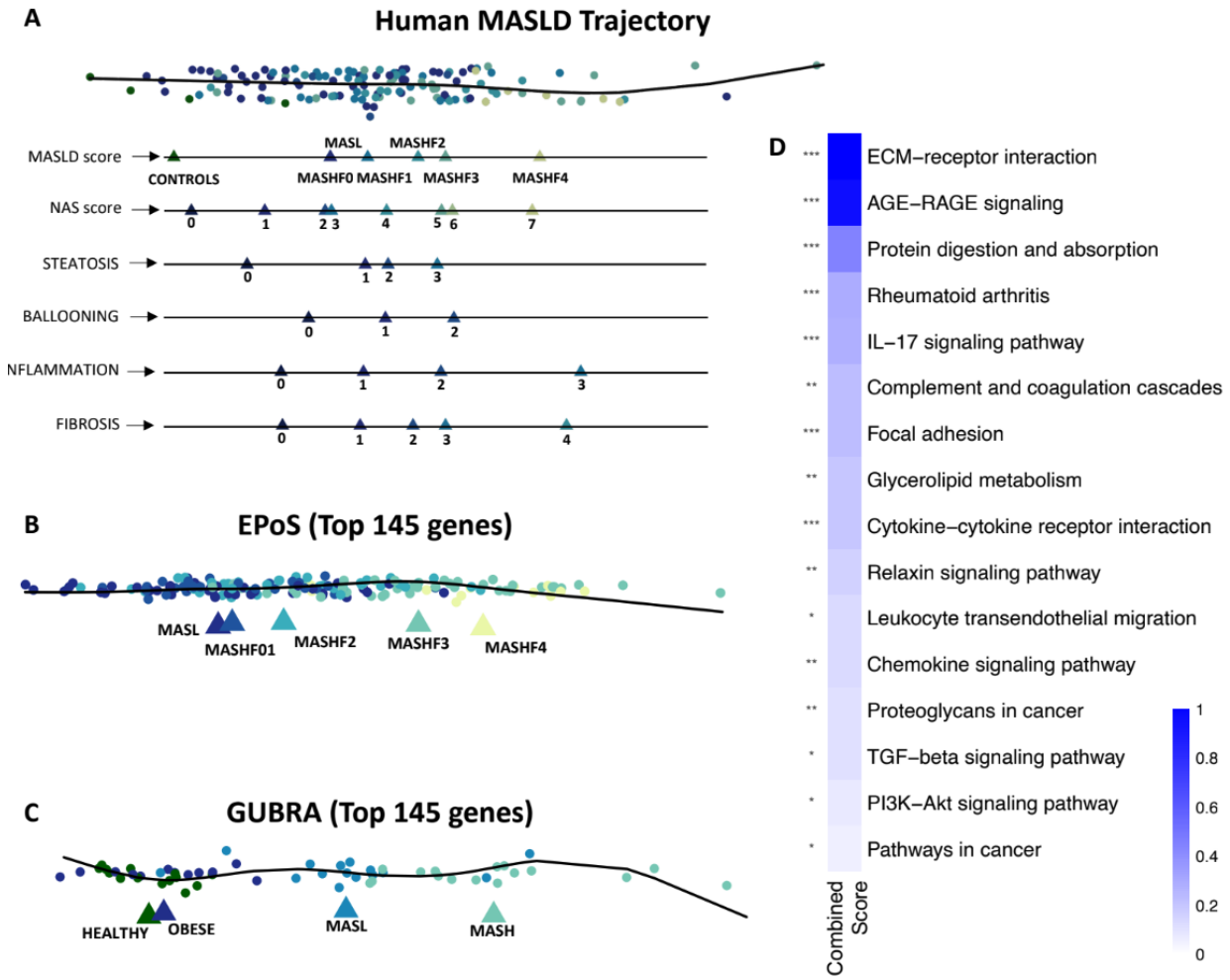


Figure 2. Pseudo-temporal ordering of patients captures disease progression and identifies MASLD gene signature. A. *Pseudo-temporal ordering of patients based on transcriptomic data recapitulates disease progression based on individual phenotypes, i.e. steatosis, ballooning, inflammation and fibrosis, and based on the NAS and MASLD scores.* ***B-C.*** *Pseudotemporal ordering of two independent orthogonal datasets based on the 145 genes that are most predictive of the trajectory provides a linear and clear separation of the disease stages (EPoS dataset (**B**) and Gubra dataset (**C**)).* ***D.*** *Functional enrichment analysis results of these 145 genes using EnrichR, where the Combined Score (c) has been calculated as c = log(p_value) * z_score, with p_value representing the Fisher exact test outcome, and z-score the deviation from the expected rank.*

Data-Driven Global MASLD/MASH Network Recapitulates Key Molecular Mechanisms of Disease

To construct a MASLD reference network, we adapted our previously published method for generating phenotype-specific networks by integrating paired transcriptomics and phenotype data[22] (Methods). Using Weighted Gene Co-expression Network Analysis (WGCNA)[23], we identified gene co-expression modules and linked them to key phenotypic features, including the NAS score, steatosis, ballooning, and fibrosis (Fig. 3A; Supplementary Table 4). Linear regression analysis revealed modules that were either uniquely associated with specific phenotypes or shared across multiple phenotypes, offering valuable insights into the molecular underpinnings of MASLD progression.

Reactome pathway enrichment analysis[24] of the co-expression modules identified both shared and distinct mechanisms associated with MASLD progression. Several modules were enriched in pathways known to play central roles in MASLD-related metabolic processes (Supplementary Table 5). The MEbrown module was primarily enriched in protein translation pathways, while the MEblue module, characterised by its abundance of zinc-finger proteins, was linked to gene expression regulation. The MEred and MEsalmon modules were strongly associated with extracellular matrix organization and signal transduction, pathways critical for fibrosis and cellular communication. The MEtan module was tied to muscle contraction processes, suggesting potential roles in metabolic or structural regulation. Notably, the MEyellow module was exclusively enriched in pathways related to the immune system and cytokine signalling, including processes such as “antigen processing and cross-presentation”, “interferon signalling”, “interleukin signalling”, and “immunoregulatory interactions between lymphoid and non-lymphoid cells”. These findings highlight the specialized roles of these co-expression modules in MASLD progression, reflecting the complexity of the disease's molecular mechanisms.

Transcription factor (TF) enrichment analysis of significant gene modules identified 188 TF regulons across eight gene modules (adjusted p-value < 0.05) (Supplementary Table 6). Among these, 48 TFs were enriched in multiple modules, underscoring their pivotal roles in MASLD progression (Fig. 3B, Supplementary Table 6). Many of these TFs were linked to TGFβ1 signalling, a key driver of fibrosis through both SMAD-dependent and SMAD-independent pathways. For instance, the regulatory networks of SP1, SRF, and ETS1 interact with TGFβ1

signalling to promote the expression of cytoskeletal and ECM proteins, including actins and integrins, which are critical to tissue remodelling and fibrosis[25–27].

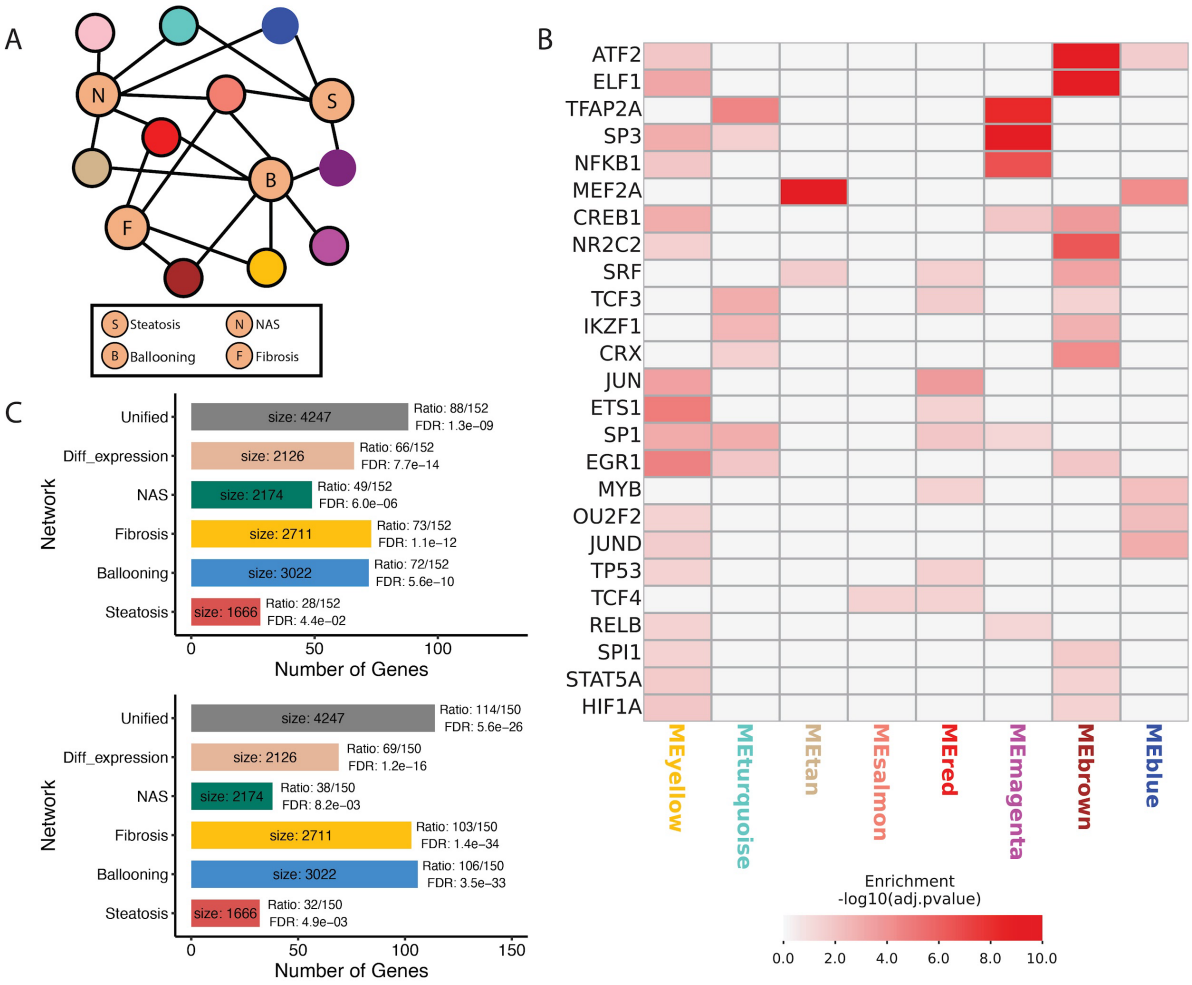


Figure 3. Phenotype-specific gene modules associated with the different MASLD variables. ***A.*** *Network of the different gene co-expression modules associated with NAS score, Steatosis, Ballooning, and Fibrosis. Each module is depicted by a different colour and can be linked with one or multiple variables.* ***B.*** *TFs whose regulons are enriched in the modules. The figure shows TFs per module coloured according to the adjusted p-value of the enrichment test. The presented TFs have been identified as enriched in multiple modules and deregulated at least in one stage of the disease trajectory (see sliding window analysis results). The full list of TFs derived from the enrichment analysis is given in Supplementary Table 6.* ***C.*** *Enrichment of MASLD network and its components in known MASLD genes curated from the literature (top) or from MSigDB (bottom).*

EGR1, another TF regulated by TGFβ1 signalling, has been implicated in both steatosis and fibrosis [28,29,30]. Similarly, SMAD3 and SMAD4, intracellular effectors of canonical TGFβ1/SMAD signalling, were found to be deregulated and enriched in the MEred module, indicating a potential disruption of the TGFβ1 feedback loop. Beyond TGFβ1 signalling, other deregulated TFs relevant

159 to MASLD include CREB1[31], NFKB1[32], and HIF1A[33,34], which are known to contribute to metabolic
160 and inflammatory pathways characteristic of MASLD progression (Supplementary Table 6).

161 Reactome pathway enrichment analysis[24] of TFs revealed the molecular pathways in which they
162 are known to participate (Supplementary Table 7). TF sets associated with steatosis and the NAS
163 score were enriched in pathways related to developmental biology, myogenesis, gene expression
164 and processes that are particularly relevant to the activation of hepatic stellate cells, a key driver
165 of fibrosis[35]. In contrast, TF sets linked to fibrosis, ballooning, and the NAS score exhibited
166 dysregulation in stress response and signalling pathways, including PUMA activation, FOXO-
167 mediated transcription, and Toll-like receptor cascades, all of which play critical roles in cellular
168 stress, apoptosis, and inflammatory responses during MASLD progression reflecting the liver's
169 response to injury.

170 Fibrosis-specific pathways, such as "Adherens junction interactions" and "Regulation of TP53
171 activity", underscore changes in cellular architecture and transcription that drive tissue
172 remodelling during fibrosis. These findings highlight the importance of targeted transcriptional
173 regulation and subtle alterations within specific cellular pathways in shaping the complex
174 dynamics of MASLD/MASH progression.

175 Leveraging this information, we constructed a comprehensive MASLD/MASH disease network to
176 examine dynamic changes in cellular processes along the disease trajectory (Methods;
177 Supplementary Table 8). The resulting network comprised 4,247 nodes and was significantly
178 enriched with genes previously associated with MASLD and integrated data from signalling and
179 metabolic pathways. Notably, the highest enrichment ratios were observed for the subnetworks
180 of ballooning and fibrosis (Fig. 3C; Supplementary Table 9). The inclusion of relevant transcription
181 factors, enriched pathways, and the statistically significant overlap of the network with known
182 MASLD/MASH-associated genes supports its robustness and suitability for downstream
183 analyses. This integrative framework provides a valuable tool for further investigation into the
184 molecular mechanisms driving MASLD/MASH progression.

185 Sliding Window Analysis Highlights Molecular Dysregulation in MASLD 186 Progression

187 To gain deeper insights into the molecular changes driving MASLD progression, we employed a
188 sliding window (SW) approach, dividing patients into groups along the disease trajectory. To

189 optimise the sequence of sliding windows, we developed a graph-based method that maximised
190 the information content within each window (Methods). Using this approach, patients were divided
191 into 13 subgroups along the trajectory for functional analysis (Methods; Supplementary Fig. 4;
192 Supplementary Tables 1 & 10). Analysis of genes within phenotype-associated modules revealed
193 patterns of collective dysregulation as the disease advanced into later stages (Supplementary
194 Fig. 5A). This approach enabled the identification of stage-specific molecular signatures and
195 provided a dynamic perspective on the progression of MASLD.

196 Initially, we identified transcription factors (TFs) that were differentially regulated along the
197 disease trajectory, demonstrating strong concordance with analyses based on traditional discrete
198 disease stages while providing greater granularity (Fig. 4A; Supplementary Table 11). Key
199 examples include HNF1A, a critical regulator of hepatic differentiation and metabolism[36], which
200 exhibited marked downregulation in sliding windows 5 and 13. This downregulation reflects
201 processes such as inflammation, epithelial-to-mesenchymal transition (EMT), and wound healing,
202 changes that were not captured by discrete stage-based approaches. Similarly, PPARα, a key
203 regulator of lipid metabolism[37], showed significant downregulation in the later stages of the
204 disease trajectory, highlighting its role in the progression of metabolic dysfunction. Other TFs,
205 including MYC, JUN, STAT1, and NFKB1/2, exhibited dynamic regulatory patterns, further
206 illustrating the complexity of gene expression changes during MASLD progression. These findings
207 underscore the value of trajectory-based analysis in capturing detailed molecular dysregulation
208 and identifying stage-specific regulatory mechanisms.

209 An intriguing cluster of TFs, including MEF2A, MEF2C, MEF2D, MYOG, NKX2-5, and ZBTB4,
210 exhibited unique activity patterns in our trajectory analysis that were not detected using traditional
211 discrete patient staging. These TFs, known primarily for their roles in myogenesis[38], have also
212 been implicated in hepatic stellate cell activation and their transition to a myofibroblast-like
213 phenotype, a critical driver of fibrosis[35]. Their activity peaked during SW5–SW6, corresponding to
214 heightened stellate cell activation, and then sharply declined at SW7, suggesting either a
215 resolution of activation or a shift in regulatory dynamics. Notably, no significant changes in their
216 activity were observed in later stages, indicating a potential stabilisation of their regulatory
217 influence as the disease progressed.

218 To explore the underlying molecular processes, we employed a network propagation-based
219 strategy to extract active network signatures for each sliding window from the MASLD reference
220 network (Methods). Reactome pathway enrichment analysis of these signatures revealed

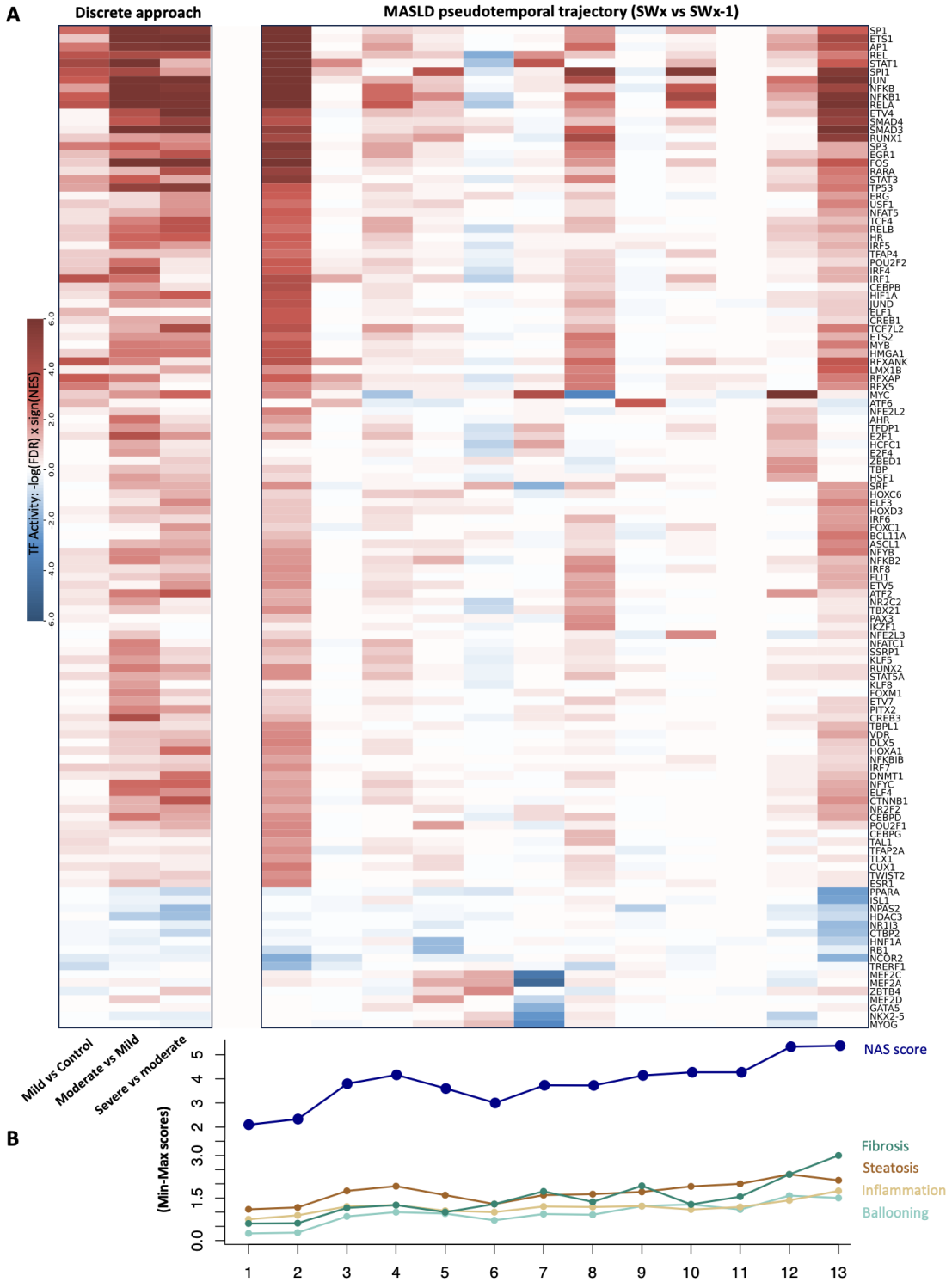


Figure 4. Sliding window analysis to study the landscape of molecular changes along the MASLD trajectory. A. *Results from TF activity analysis in the discrete patients stratification, and the pseudotemporal trajectory. The colour indicates change in TF activity compared to the previous disease stage.* ***B.*** *Mean phenotype and NAS pathologist scores along the SW trajectory.*

222 dynamic, progressive changes along the disease trajectory. Signal transduction and immune
223 processes became increasingly active as the disease advanced, reflecting the intense
224 inflammatory and signalling dysregulation characteristic of later stages. Pathways related to
225 extracellular matrix organisation, strongly associated with fibrosis, exhibited the most significant
226 upregulation in advanced stages. In contrast, metabolism-related processes were consistently
227 downregulated throughout the trajectory, highlighting the persistent metabolic dysfunction driving
228 MASLD progression (Supplementary Fig. 5B).

229 We compared the dysregulation of molecular processes in each sliding window to traditional
230 MASLD histopathological scores, such as steatosis, ballooning, inflammation, fibrosis, and NAS
231 (Fig. 4B). While these scores generally increased with disease progression, some pathways
232 exhibited early dysregulation before corresponding changes in histopathological scores. For
233 example, extracellular matrix pathways were upregulated at SW8, preceding the observed
234 increase in the fibrosis score at SW10, suggesting that molecular markers may provide earlier
235 insights into MASLD progression than traditional assessments. Overall, the enhanced resolution
236 of our sliding window-based approach offers a comprehensive molecular landscape of MASLD
237 progression, shedding light on the interplay between immune activation, fibrotic remodelling, and
238 metabolic dysregulation over time.

239 Cell-Type Deconvolution Along the MASLD Trajectory Reveals Network 240 Changes Associated with Tissue Composition

241 Changes in cell type composition and cellular states within the liver are critical indicators of
242 MASLD progression, often driven by increased inflammation resulting from hepatic injury caused
243 by fat accumulation[1]. Understanding the molecular triggers underlying these cellular changes in
244 the hepatic microenvironment offers a valuable opportunity to identify stage-specific therapeutic
245 targets for reversing disease progression. Additionally, these insights could facilitate the discovery
246 of prognostic biomarkers, enabling earlier detection and more personalized MASLD treatment
247 strategies.

248 To investigate this, we performed cell type deconvolution on bulk patient data along the disease
249 trajectory to identify stage-specific shifts in cell composition and their association with the MASLD
250 disease network (Methods; Fig. 5A). The analysis revealed a decrease in hepatocyte and resident
251 NK cell proportions, highlighting the progressive infiltration or expansion of other cell types within
252 the liver. In contrast, increasing proportions of T cells, neutrophils, migratory dendritic cells,

253 cholangiocytes, and fibroblasts were observed, indicating heightened immune responses,
254 including inflammation and fibrogenesis. Interestingly, the stability of endothelial cell proportions
255 beyond the midpoint of the trajectory suggests their role in maintaining blood vessel integrity and
256 functionality, ensuring consistent liver blood flow and nutrient exchange even in advanced MASLD
257 stages. Notably, macrophage proportions remained relatively stable throughout the trajectory,
258 with only minor fluctuations, underscoring their consistent presence and potential involvement in
259 liver homeostasis and immune regulation across disease stages.

260 We next investigated the association between cell types and the deregulated processes identified
261 along the trajectory (Supplementary Fig. 5B). We identified genes whose differential expression
262 could be solely attributed to changes in cell type proportions along the trajectory, using pseudo-
263 bulk RNA-seq samples derived from a healthy liver single-cell RNA-seq dataset. These genes
264 were then analyzed to determine whether the enrichment of Reactome pathways was driven by
265 shifts in cell type proportions or by the cell type-specific deregulation of biological processes
266 (Methods). As expected, hepatocytes were primarily associated with deregulated processes
267 related to lipid metabolism, while other metabolic processes were also linked to cholangiocytes
268 and endothelial cells (Fig. 5B; Supplementary Table 12). Immune-related processes were
269 predominantly associated with macrophages, with the exception of 'signalling by interleukins,'
270 which was primarily linked to neutrophils, T cells, and other immune cell types. These findings
271 underscore the interplay between specific cell types and biological processes driving MASLD
272 progression. Deregulated processes related to extracellular matrix (ECM) organisation and
273 fibrosis were predominantly associated with fibroblasts, highlighting their central role in ECM
274 remodelling and fibrotic progression. Non-immune-related signalling pathways, such as Wnt
275 signalling, Akt signalling, and Rho signalling, were primarily linked to endothelial cells, with
276 contributions from fibroblasts and cholangiocytes in some cases. Interestingly, specific
277 processes, such as Wnt signalling, were also associated with specific immune cells, suggesting
278 a potential role in immune modulation or interactions with the stromal microenvironment. These
279 findings further underscore the complex crosstalk between cellular signalling, immune responses,
280 and tissue remodelling in MASLD progression.

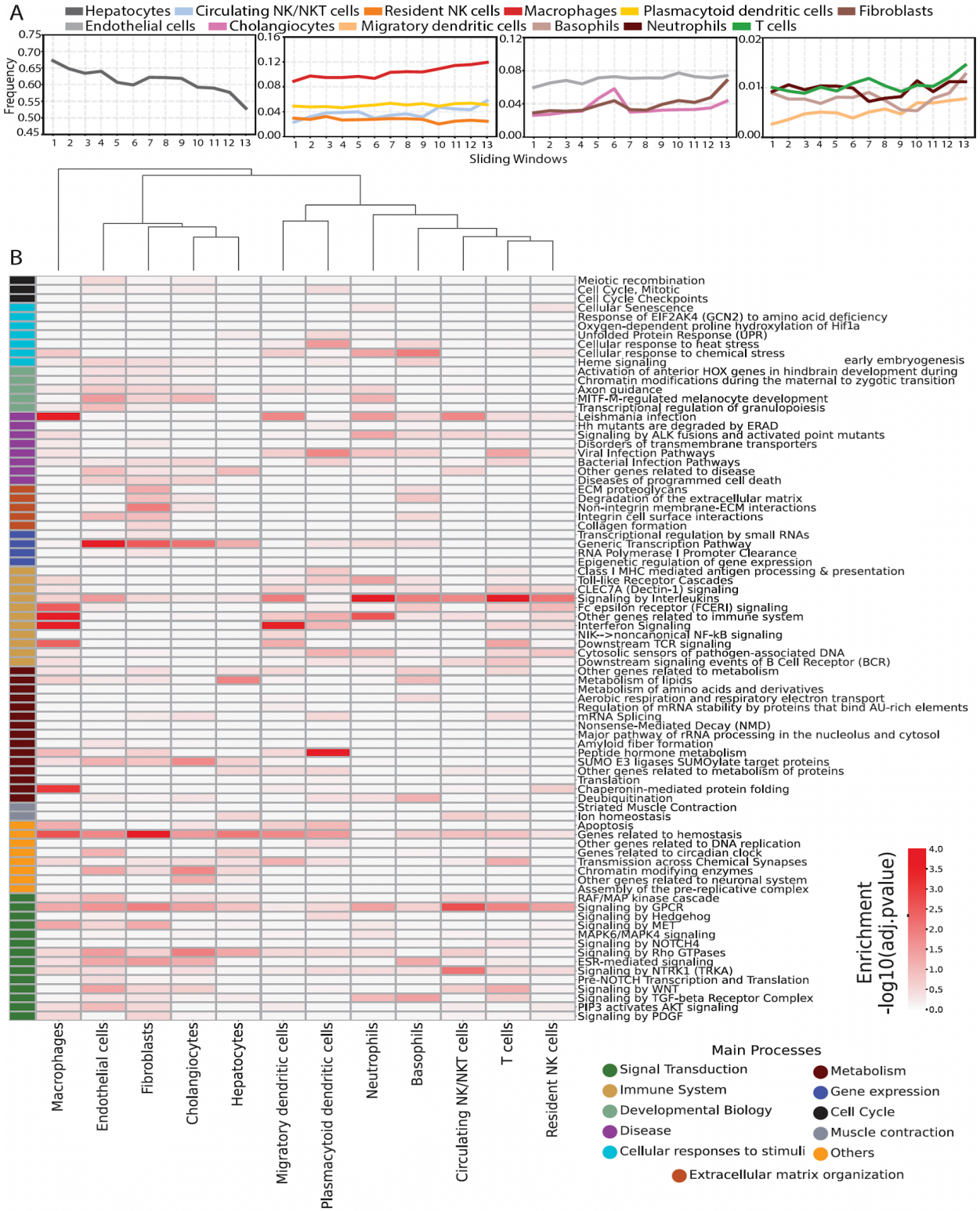


Figure 5. Results of cell type deconvolution analysis. A. *Cell type deconvolution of patient transcriptomic data along our MASLD trajectory* ***B.*** *Prediction of cell types with which the deregulated processes are associated after excluding gene sets of these processes that could have been identified as differentiated by the mere change in abundance of cell types.*

281

MASLD Trajectory-Predictive Biomarkers Through Integrated Plasma-Liver Expression Analysis

To address the critical need for non-invasive diagnostic methods for MASLD/MASH staging, we sought to identify genes capable of predicting a patient's MASLD stage through blood proteomics. To establish a baseline for the predictive performance of selected gene sets, we employed a Random Forest regression model using a list of 145 gene signatures derived from the first principal component of our initial analysis (Figs. 2 & 6A). This gene list demonstrated high predictive performance on the UCAM/VCU dataset (Supplementary Fig. 3C), achieving an R^2 of 87.1% and a Mean Squared Error (MSE) of 0.005. Here, R^2 represents the proportion of variance in a patient's trajectory position explained by the model, underscoring the robustness of this gene set in trajectory prediction.

Govaere et al.[39] recently identified 194 genes with correlated expression in both blood plasma and liver across MASLD progression stages. Incorporating these plasma-liver correlated genes into a Random Forest model yielded strong but slightly lower predictive performance compared to the baseline on the UCAM/VCU dataset (R^2: 79.9%, MSE: 0.008). However, performance improved when we retained only the 153 genes with a positive variable importance score in the Random Forest model (R^2: 81.2%, MSE: 0.007; Fig. 6A).

To refine our focus on genes directly relevant to MASLD progression, we narrowed the list to 24 genes that overlapped with our MASLD network. This selection was based on the rationale that these genes are more likely to be directly involved in disease progression and, therefore, better suited for monitoring. The inclusion of our network provided functional context for their plasma protein modulation, enhancing both biological relevance and interpretability. This curated gene set achieved an R^2 of 69.4% and an MSE of 0.01 on the UCAM/VCU dataset, offering a balance between predictive accuracy, simplicity, and interpretability (Fig. 6A; Supplementary Fig. 6A, B).

To validate the ability of our 24-biomarker model to place patients along the trajectory using an independent dataset, we applied it to the EPoS dataset[8] (see first section of our Results & Methods). This dataset included patients for whom plasma protein data was extracted in the Govaere study[39]. Our model successfully recapitulated increasing trends in MASLD stage progression, NAS, and fibrosis scores corresponding to the predicted trajectory (R^2: 81%, MSE: 0.105; Supplementary Figs. 6C–F). Notably, this gene set explained 87.6% of the trajectory variance in this dataset. In contrast, when using random gene signatures of 1,000 and 24 genes,

313 the model failed to accurately predict patient positions along the trajectory accurately, yielding R^2
314 values of approximately 11% for both random sets (Supplementary Fig. 6G, H). This highlights
315 the robustness and specificity of our curated 24-gene biomarker set in predicting MASLD
316 progression.

317 To further validate our model on an entirely independent dataset, we applied it to the Gubra study
318 dataset[21] (described in the first section of Results & Methods; Figs. 6B, C). This dataset effectively
319 distinguished healthy/obese individuals from MASLD/MASH patients, accurately placing them
320 along the trajectory with minimal errors (MSE for predicted vs. actual trajectory positions;
321 Supplementary Fig. 7). Among the tested gene sets, the refined 153-gene set achieved the best
322 performance (R^2: 87.8%, MSE: 0.22), while the 24-gene biomarker set also performed highly (R^2:
323 87.6%, MSE: 0.214) and explained 81% of trajectory variance. The model accurately classified
324 25 out of 30 healthy/obese individuals (83.3%) and 26 out of 27 MASLD/MASH patients (96.3%),
325 achieving an overall accuracy of 89.5% (Supplementary Fig. 7). These results highlight the
326 potential of this model for non-invasive MASLD/MASH staging and tracking disease progression.

327 The genes with the highest importance values in the model's predictions included those involved
328 in extracellular matrix remodelling (COL3A1, MFAP4, THBS2), immune modulation (CXCL8, C7,
329 CCL23, SAA1), metabolic regulation (AKR1B10, HSD17B8, APOF, MTHFS), and cellular stress
330 response (PRDX6, HSPA6). By mapping their "activation" (significant changes compared to the
331 previous trajectory window) and linking them to MASLD network subnetworks (Methods), we
332 charted the timeline of pathway deregulation in blood samples (Fig. 6D). These results were
333 consistent with findings from the liver-based trajectory pathway analysis (Supplementary Fig. 5).

334 To identify therapeutic opportunities, we hypothesised that targets capable of modulating these
335 predictive biomarkers and present within the MASLD network should be enriched for candidates
336 with a causal association to disease progression. Using the ChEMBL database[39], we searched
337 for approved drugs targeting these biomarkers—either as agonists/activators when the biomarker
338 was downregulated in the disease sliding window or as inhibitors when the biomarker was
339 upregulated. This analysis identified 359 drugs targeting the biomarkers or their associated
340 networks, including 136 drugs that had received regulatory approval (passed Phase 3), regardless
341 of indication (Supplementary Table 13).

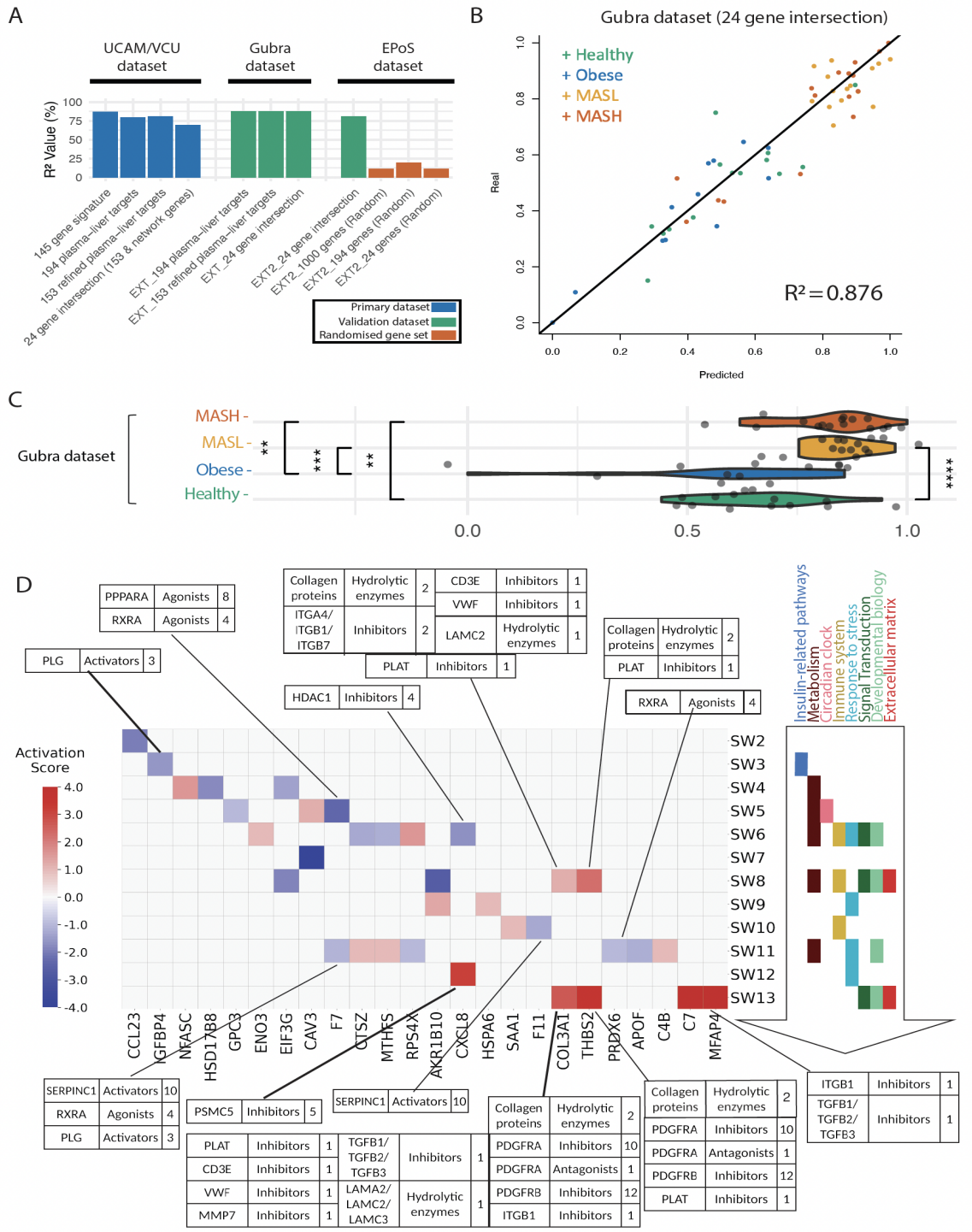


Figure 6. Random Forest regression model for the prediction of MASLD patients' position on the inferred disease trajectory. A. *Model performance using different gene sets in the primary UCAM/VCU dataset (5-fold cross validation), and validation in two independent datasets used solely for testing. Randomisation was done on 1000 times.* ***B.*** *Compared predicted and actual patient positions on the trajectory using an independent dataset (GUBRA), based solely on a 24-gene signature.* ***C.*** *Violin plot showing the distribution of the predicted patients' position on the trajectory (GUBRA dataset).* ***D.*** *Network-based deregulation analysis of the 24-gene signature to identify potential targets. The color scale represents the -log10(adjusted p-value) of differential expression. Genes are deemed deregulated in a SW if part of the respective deregulated subnetworks. Drugs in max phase 4 were retrieved for targets in the relative deregulated pathways.*

343 Examining the indications annotated in ChEMBL for these drugs, which include all indications
344 evaluated during clinical trials regardless of approval, revealed that several drugs target
345 conditions related to MASLD. For instance, fenofibrate and bezafibrate, both PPARα agonists,
346 are approved for hypertriglyceridemia but failed in MASLD clinical trials. Similarly, belinostat
347 (targeting HDAC1) and dasatinib (targeting PDGFRB) have shown potential for treating
348 hepatocellular carcinoma (HCC), although they are not yet approved for this indication. These
349 examples, along with others in our target list (Supplementary Table 13), underscore the potential
350 of leveraging our analysis for drug repurposing or the development of targeted single or
351 combination therapies for MASLD. By focusing on existing drugs with mechanistic alignment to
352 the disease's molecular drivers, this approach could accelerate therapeutic development and
353 improve outcomes for MASLD patients.

354 Discussion

355 MASLD has emerged as the leading cause of chronic liver disease and the second most common
356 reason for liver transplantation in the Western world[40]. Despite its growing prevalence, the
357 molecular mechanisms driving MASLD progression remain poorly understood. Traditional staging
358 methods, which rely on histological scores, face significant limitations, including variability due to
359 manual annotation and their inability to capture the continuous nature of disease progression. In
360 this study, we addressed these challenges by adopting a data-driven approach that leverages
361 pseudo-temporal ordering[20] of MASLD patients, coupled with network analysis, to create an
362 unbiased map of molecular changes across the MASLD progression continuum[20]. This
363 methodology offers a more dynamic and granular understanding of disease progression
364 compared to traditional staging techniques.

365 In addition to identifying a 145-gene signature predictive of histology scores at the liver biopsy
366 level, we developed a 24-protein biomarker set derived from our MASLD network. This set can
367 be measured non-invasively in plasma and used to predict a patient's position along the disease
368 trajectory. These biomarkers have the potential to enhance existing tools, such as the NAFLD
369 Fibrosis Score (NFS) and FIB-4, by improving stage precision, enabling multidimensional
370 assessments, and supporting personalized, stage-specific treatments[41]. Moreover, as these
371 biomarkers reflect molecular processes central to MASLD progression, they offer valuable
372 opportunities for monitoring treatment responses and linking molecular mechanisms to clinical
373 outcomes, paving the way for more effective therapeutic strategies.

374 By employing sliding windows along the disease trajectory instead of discrete groupings, we
375 identified both progressive and "switch-like" changes in TF activity and associated molecular
376 processes.

377 For example, immune signalling TFs, such as NFKB1 and STATs, exhibited consistent
378 upregulation at short and regular intervals throughout the trajectory. This trend was mirrored at
379 the network level, where immune processes were active from the earliest stages and showed
380 progressive increases in activation over time.

381 In contrast, profibrotic and myogenesis-related TFs, including MEF2C, MEF2A, MEF2D, ZBTB4,
382 GATA5, NKX2-5, and MYOG, displayed a distinct activation pattern during MASLD progression.
383 Their activity peaked between sliding windows 4 and 6, then declined and stabilised by window
384 7, a dynamic not detected by traditional disease staging. Among these, MEF2 is particularly
385 noteworthy, as it regulates hepatic stellate cell (HSC) activation and their transformation into a
386 myofibroblast-like phenotype[42]. HSCs are crucial in the liver's wound-healing response to injury,
387 but their persistent activation in MASLD—due to impaired apoptosis—drives pathological liver
388 fibrosis and sustained inflammation, leading to advanced disease stages, including cirrhosis and
389 potentially hepatocellular carcinoma[42]. Strategies targeting HSC apoptosis have shown
390 therapeutic promise. For instance, hydronidone has been proposed to improve liver fibrosis
391 phenotypes by inducing HSC apoptosis, providing a potential pathway for intervention[43].

392 At the pathway level, our trajectory-based analysis revealed stage-specific modulation of key
393 biological processes[8] For instance, the unfolded protein response becomes activated toward the
394 end of the trajectory[44]. In contrast, processes such as the response to hypoxia[45,46], extracellular
395 matrix organization pathways [47], and downregulation of lipid metabolism[48] emerge from sliding
396 window 3 onward. Notably, the downregulation of PPARα and HNF1A characterizes the latest
397 disease stages.

398 Certain signalling pathways exhibited distinct activity patterns. WNT[49], PDGF[50] and TGF-β
399 signalling[51] remain consistently activated throughout the trajectory, while Notch signalling shows
400 increased activity beginning at SW4, with a marked boost in the latter half of the trajectory[52].
401 Targeting TGF-β signalling has shown potential to reduce disease progression, making it a
402 promising therapeutic avenue[53–55]. Similarly, Wnt signalling is implicated in MASLD progression[49],
403 with animal studies demonstrating that inhibiting this pathway can suppress hepatic stellate cell
404 (HSC) activation, improving disease phenotypes[56]. Collectively, our findings highlight the dynamic

405 reprogramming of these processes at different time points, impacting multiple levels of cellular
406 regulation—including proteostasis, the cell cycle, cell signalling, and metabolism. These insights
407 offer valuable opportunities for stage-specific therapeutic interventions and a deeper
408 understanding of MASLD progression.

409 A hallmark of MASLD progression is the gradual shift in liver cell composition, driven by changes
410 in cellular signalling and plasticity. For example, hepatic stellate cells differentiate into
411 myofibroblast-like cells, while hepatocytes and cholangiocytes undergo transdifferentiation[57].
412 Cell-type deconvolution revealed a decrease in hepatocytes, likely due to cell death or
413 differentiation into cholangiocytes, accompanied by increases in cholangiocytes, fibroblasts, and
414 immune cells such as T cells, neutrophils, and circulating NK/NKT cells. Metabolic dysregulation
415 was predominantly linked to hepatocytes, while fibroblasts and cholangiocytes were major
416 contributors to signalling deregulation and extracellular matrix remodeling. Unsurprisingly,
417 immune cells—especially macrophages—were implicated in immune processes[58]. Macrophages,
418 including Kupffer cells, play a key role in regulating lipid metabolism during steatosis and initiating
419 steatohepatitis through proinflammatory signaling.[59]

420 Our findings highlight the intricate relationship between hepatic and systemic immunity during
421 MASLD progression. Future research using single-cell[57] and spatial data will be crucial to uncover
422 how liver zonation, structural organization, and local niches drive or modulate immune and cellular
423 processes in MASLD[59].

424 One limitation of our analysis is that, while it successfully captures molecular events across the
425 disease trajectory, it cannot fully distinguish drivers of cell deregulation from adaptive or passive
426 processes. Additionally, our reliance on cross-sectional data introduces uncertainty, as not all
427 patients in the early stages progress to advanced ones. Despite these limitations, our data-driven
428 MASLD network provides a valuable foundation for understanding disease mechanisms and
429 identifying stage-specific therapeutic targets. The network's modules, which are linked to distinct
430 histological phenotypes, support the development of tailored interventions for specific sub-
431 phenotypes or their combinations. Incorporating longitudinal data and metadata, such as diabetic
432 status or prior treatments, could further enhance the predictive power of the network, improve
433 biomarker discovery, and advance personalized treatment strategies for MASLD.

434 In summary, this study provides a comprehensive and integrative perspective on MASLD
435 progression, advancing our understanding of its molecular underpinnings. The data-driven

trajectory framework addresses the limitations of traditional histology-based staging by offering a reproducible and scalable approach to disease annotation. Importantly, this methodology is not restricted to MASLD; by partitioning patient data—or any trajectory data—into informative sliding windows, it can be adapted to explore other progressive diseases, including those modelled at the single-cell level.

The identified gene signature and plasma biomarkers enhance non-invasive diagnostics and enable real-time monitoring of disease dynamics, while the trajectory analysis reveals key molecular processes and identifies novel therapeutic targets. Although longitudinal data will be essential for further refinement, this study lays a strong foundation for future investigations into MASLD and other progressive diseases. Ultimately, this work advances the development of precision medicine, with the goal of improving outcomes for patients affected by this increasingly prevalent condition.

Materials & Methods

Datasets

For the analyses conducted in this study, 2 different publically available human datasets have been used: the "UCAM" dataset[19] (consisting of 58 consecutive patients recruited at the MASH Service at the Cambridge University Hospital) and the VCU dataset[9] (consisting of 4 obese bariatric controls without MASLD and 74 patients covering the whole MASLD spectrum). The raw data and relevant metadata have been downloaded from Array's Express database (E-MTAB-9815) and NCBI's GEO repository (GSE130970), respectively. Details for both datasets are given in Supplementary Table 1. "Sample_5" was removed from the analysis as it was considered an outlier (Supplementary Fig. 8A-B).

For the validation of the human trajectory inference method, we used two publicly available datasets (GUBRA[21,60] and EPoS[8]). The GUBRA dataset consists of n=26 healthy (n=14 normal-weight, n=12 overweight) individuals recruited at the Center for Diabetes Research (Copenhagen University, Denmark) and n=31 MASLD/MASH patients recruited at the Department of Hepatology and Gastroenterology (Aarhus University Hospital, Denmark). The EPoS dataset consists of 168 MASLD/MASH patients, originates from a large cohort recruited in multiple European parts, and has been previously used by our team[61]. In additionally to the expression

465 data, the EPoS dataset also provided information on the MASLD disease stage, NAS, and
466 Fibrosis scores per patient.

467 RNA-Seq Analysis and data processing

468 For the RNA-Seq analysis, the *FASTQC* software was used for generating quality-control (QC)
469 reports of individual FASTQ files (http://www.bioinformatics.babraham.ac.uk/projects/fastqc, v.
470 0.11.9) and reads were aligned to the human GRCh38 reference genome using hisat2[62] (v 2.1.0).
471 Then, the genes were counted with *HTSeq*[63] (v 0.11.1), and the *biomaRt package*[64] was used to
472 map “Ensembl gene IDs” to “HGNC symbols”.

473 We used quantile normalisation to normalise the 2 datasets and Bioconductor's function
474 *COMBAT*[65] from R package *sva* to remove the batch effect observed upon integrating them
475 (Supplementary Fig. 8A-B). Quality control checks validate the robustness of the analysis,
476 showing that the expression of the genes in each MASLD stage separately for each dataset
477 remains similar after batch effect correction (Supplementary Fig. 8C-D), with a high correlation of
478 the first principal component of the PCA plot with the main MASLD characteristics (Supplementary
479 Fig. 8E), and a similar expression of genes previously linked to MASLD before and after the
480 correction (Supplementary Fig. 8F).

481 Pseudo-temporal ordering of patients

482 The merged dataset was not expected to have multiple disconnected trajectories with a particular
483 cyclic or (necessarily) complex topology. Thus, we selected *Slingshot*[20,66] as the most suitable
484 tool to infer the disease trajectory. We constructed an MST (Minimum Spanning Tree) on ordered
485 sets of clusters to capture the disease trajectory based on the points on the PCA plot. The inferred
486 horizontal trajectory was in agreement with the high correlation scores between PC1 and the
487 MASLD characteristics. Thus, this dimension could be considered as a pseudo-temporal
488 trajectory of MASLD, in which small values represent healthy or initial stages of disease, while
489 large values belong to severe fibrosis conditions. Based on this hypothesis, each point on the
490 PC1 axis could be considered as a pseudo-time. For the independent dataset, this process was
491 repeated with the complete transcriptome and using only the 145 gene signatures identified using
492 the random forest classification (see section below).
493
494

495 ## Patient stratification for annotation of trajectory

496 Based on the histological scores defined by the CRN scoring system[67] and following the older
497 definition of NAFLD/NASH, our dataset (136 patients) has been divided into 4 bariatric obese
498 controls and 132 MASLD, which has further been sub-clustered based on the fibrosis stage into
499 mild (MASH F0), moderate (MASH F1-2) and severe (MASH F3-4) groups. This stratification aims
500 to increase the interpretability of the analysis, which has been described in detail in our previous
501 studies[19,61] (Supplementary Table 1).

502

503 ## Extraction of 145 gene signatures using Random Forest classification

504 Calling the parameter *rotation* from *prcomp* function in R (default stats package), we produced a
505 list of ranked genes with the highest absolute values that explain the variance along with the
506 disease progression on the horizontal axis. The top 200 genes were selected as the training set
507 of a random forest classification model (R package *randomForest*) to predict the disease stage
508 (MASL, MASHF01, MASHF2, MASHF34) and the different MASLD characteristics (steatosis,
509 inflammation, ballooning, fibrosis). The accuracy was estimated using 5-fold cross-validation
510 1000 times. From this first run of the model, only genes with positive “Mean Decrease Accuracy”
511 were selected, representing the set of top genes that contributed to the disease stage prediction.
512 Those genes were used as the training set of a new random forest classification model.

513

514 ## Pathway enrichment analysis

515 The Reactome Pathway Knowledgebase[24] (accessed May 2024) was used in different steps of
516 this study to perform pathway enrichment analysis. Fisher's exact test[68] was used for the statistical
517 significance test, and the Benjamini–Hochberg method[69] was applied for the multiple hypothesis
518 testing correction.

519

520 ## Identification of gene expression modules associated with MASLD
521 ## histological phenotypes

522 We applied the weighted gene co-expression network analysis (WGCNA)[70], as in Barker et al.[22],
523 through the R package *WGCNA*[23] (v 1.72-1) in combination with an optimisation task to identify
524 the optimal set of gene co-expression modules. Initially, we used the soft-thresholding approach
525 to construct a signed weighted gene co-expression network, and the power (soft thresholding

526 parameter) was set to 4, as it was the minimum value for which the weighted network
527 approximated scale-free topology (fit index equal to 0.9). Then, we modified the derived adjacency
528 weights into a topological overlap matrix (TOM) to finally get the dissimilarity matrix of genes (1-
529 TOM). We used the *cutreeDynamic* function of the *dynamicTreeCut* R package (v 1.63-1) to
530 analyse the dissimilarity matrix and extract the set of co-expression modules. Instead of using the
531 default parameters, we performed an optimization process to find the optimal set of "deepSplit"
532 and "minClusterSize" parameters. The criterion of this optimization task was the accuracy of the
533 derived modules to model the four MASLD phenotypes (steatosis, ballooning, fibrosis and the
534 NAS score) using linear regression. To that end, we performed cross-validation (k=4), separating
535 the patients into four groups with the *crossv_kfold* function of the *modelr* R package (v 0.1.11). At
536 each round, we modelled the phenotypes as linear functions of the derived eigen-genes of
537 modules (*lm* and *step* functions of the *stats* R package) using the training dataset. In order to
538 capture mechanisms underlying the progression of MASLD rather than downstream effects, we
539 used the inflammation score as a covariate in this model. Each linear model had the following
540 format:
541
542 $$MASLD\ phenotype \sim intercept + eigengenes\ of\ modules + inflammation + age$$
543
544 We applied a backward stepwise search and the Akaike Information Criterion (AIC)[71] to optimise
545 the coefficients of eigen-genes in each linear model, while modules with statistically non-
546 significant coefficients were filtered out (adjusted p-value<0.05, Benjamini-Hochberg
547 correction[69]). The test dataset was used to estimate the accuracy of the model, calculating the
548 mean squared error (MSE). In order to find the optimal parameter values, we calculated the
549 average test MSEs for the trained models of the four MASLD variables for each parameterisation
550 set. Following this approach, we found that the test error was minimum for ["deepSplit"=4,
551 "minClusterSize"=180]. We created the final linear models using this set of parameters and
552 keeping only the statistically significant modules for each phenotype. For the downstream
553 analysis, we filtered the gene content of each significant-identified module to select those genes
554 with Pearson's correlation equal to or greater than 0.4 with the respective eigen-gene.
555

556 ## Transcription factor activities analysis

557 We used *VIPER*[72] (Virtual Inference of Protein-activity by Enriched Regulon analysis; v 1.37.0)
558 and the CollecTRI database[73] (accessed May 2024) to identify the upstream transcriptional

559 regulators that can explain the observed gene expression changes. The calculated Normalised
560 Enrichment Score (NES) was used to determine the direction of each regulation ("activation" for
561 TFs with NES > 0 and "inhibition" for those with NES < 0). We performed the hypothesis testing
562 using the student's t-test, and then, the Benjamini-Hochberg method[69] was used to correct the p-
563 values, ensuring decreased Type I errors.
564

565 Generation of global MASLD network

566 This task aimed to create a global interaction network for MASLD, including information about the
567 molecular events that take place in different stages of the disease, either in signalling cascades
568 (upstream component) or the transcriptional regulation (downstream component). For that
569 reason, we exploited the results of WGCNA and differential expression analysis in SWs,
570 generating upstream and downstream subnetworks for each MASLD phenotype, as well as for
571 the significantly deregulated TFs. We used the *Prize-collecting Steiner Forest (PCSF)* algorithm,
572 implemented through the R package *PCSF*[74] (v 0.99.1), to create these subnetworks. Finally, we
573 connected the pairs of upstream and downstream subnetworks and unified all of them into the
574 global MASLD network.

575 *PCSF* returns as output a subset of a given reference network relevant to an input-weighted list
576 of prize-carrying nodes. To define such a list for each MASLD phenotype, we applied the following
577 workflow: We performed TF enrichment analysis on the filtered gene content of each significant-
578 identified WGCNA module, using the CollecTRI database[73] and other TF-regulon databases,
579 retrieved from the repository of *EnrichR*[75] (ENCODE and ChEA Consensus TFs from ChIP-X[76],
580 DNA binding preferences from JASPAR[77,78], TF protein-protein interactions). Then, we performed
581 Reactome[79] pathway enrichment analysis on each module-specific set of enriched TFs, and we
582 selected only the most semantically specific pathways, which are also related to signal
583 transduction. Consequently, each *WGCNA* module was linked with a group of TFs and a set of
584 Reactome pathways. The filtered gene sets of WGCNA modules (downstream nodes), the
585 enriched TFs, and the gene content of the enriched Reactome pathways (upstream nodes)
586 constituted the prize-carrying nodes for each MASLD variable.

587 We defined a similar set of prize-carrying nodes related to the results of differential expression
588 analysis. To that end, we first selected the TFs that have been identified as deregulated in at least
589 one SW (adjusted p-value < 0.05). Then, we identified the most specific signalling pathways
590 related to this set of TFs, similar to the previous task. Additionally, we found the intersection of

591 the selected TFs regulons (using the CollecTRI annotation) with the gene content of WGCNA
592 modules. This set of TF regulons (downstream nodes), the group of deregulated TFs and the
593 gene content of the enriched Reactome pathways (upstream nodes) constituted the prize-carrying
594 nodes for the deregulated network. We performed all the aforementioned enrichment tests using
595 the Fisher's exact test and Benjamini-Hochberg correction (adjusted p-value < 0.05), adjusting
596 the background annotation of the databases to include only genes that were present in the
597 transcriptomic dataset.

598 We constructed the reference interaction network, integrating various interaction databases with
599 the *union* function of the *igraph* R package (v 2.03)[80], including both cell signalling and metabolic
600 pathways (Reactome[79], KEGG[81,82], HumanCyc[83], ReconX[84], provided by Pathway Commons[85],
601 and OmniPath[86], accessed using the *OmniPathR* library in Bioconductor[87,88]). Regarding the
602 metabolic networks of HumanCyc and ReconX, we removed all the ChEBI entries with zero out-
603 coming edges, as they are ending points on the network. Additionally, we filtered the OmniPath
604 network to avoid weakly verified interactions, keeping only those with a number of resources and
605 curation efforts more than the 0.5 and 0.75 quantiles of the respective distributions. We split the
606 complexes to form fully connected groups of protein nodes, and we deleted all the interactions
607 annotated as inhibitions, as their molecular effect cannot be modelled using the type of network
608 analysis that we used (network propagation). The final reference network included 12,549 nodes
609 and 205,706 edges.

610 In order to run *PCSF*, we define the cost of the edges in the reference network equal to 0.001 and
611 the prizes of seed nodes equal to 100. For each of the five variables, we created an upstream
612 subnetwork (combining the upstream nodes and TFs) and a downstream subnetwork (combining
613 the downstream nodes and TFs), and then we connected them to form a unique network. This
614 gradual generation of the MASLD network was selected in order to keep information about the
615 origin of each node and edge. We used a variant of the *PCSF* algorithm (*PCSF_rand*) to generate
616 the subnetworks. This function executes the primary process several times (10 iterations were
617 selected here), adding random noise to the edge costs and merges all the solutions to create a
618 unique one. Finally, we unified the five networks into a global one using the *union* function of the
619 *igraph* R package (v 2.0.3).

620 We filtered the derived MASLD network so as to keep only functionally annotated genes and
621 weight the edges according to the semantic similarity of interactors. We used the *GOSemSim* R

622 package[89] (v 2.26.0) to calculate the semantic similarity of all pairs, according to their annotation
623 in Gene Ontology Biological Process (GO-BP)[90], based on the measures of Resnik[91], Lin[92] and
624 Wang[93]. We used different measures to score the similarities, aiming to mitigate each one's
625 potential drawbacks and biases. We averaged the similarity values to define the weight of each
626 edge. The lack of annotation of some genes produced edges with undefined weights. These
627 genes and the respective edges were removed from the network. Finally, we applied the
628 Laplacian normalisation in order to reduce the topological bias of hub nodes[94]. The final MASLD
629 network included 4,247 nodes and 38,847 edges.

630 To evaluate the relevance of the derived network to MASLD, we calculated its enrichment in two
631 gene signatures. The first signature (164 genes) was created based on recent literature[95–100],
632 including genes which have been associated with different stages of MASLD progression. The
633 second signature (155 genes) was retrieved from the MSigDB repository[101] and corresponds to
634 the canonical MASLD pathway of WikiPathways[102]. The overlap between these two signatures
635 was 15 genes. Fisher's exact test was applied to estimate the statistical significance of the
636 enrichment scores.

637 Patients stratification into sliding windows

638 We divided the ordered samples on the trajectory into sliding windows (SWs), namely as a
639 sequence of overlapping groups [SW_1, SW_2,..., SW_X], to investigate how the MASLD molecular
640 profile changes on the pseudo-temporal space. In order to define the SW-based stratification of
641 samples, we developed a graph-based method, which examines all the possible sequences of
642 SWs for a given set of SW sizes $S=[S_1,\ldots, S_m]$ and an overlapping percentage α between two
643 adjacent SWs and finds the optimal one. The concept of this novel method is based on the
644 representation of disease stages as nodes connected on a directed path, where each node is
645 annotated with a group of samples aligned to their pseudo-temporal order. A partial overlap of
646 samples between each pair of connected nodes ensures the continuity of the disease trajectory.
647 The optimisation process aims to detect a path of nodes that can sufficiently describe the
648 molecular events that occur along the disease progression through the sequential comparison of
649 the groups of samples. Generally, the method is implemented using a three-step workflow: the
650 creation of all the potential directed paths of SWs, their evaluation and filtering, and finally, the
651 selection of the optimal one.

652 The method starts with the mapping of samples to a set of indices *I=[1, 2,..., N]*, based on their
653 ranking on PC1, where *N* is the size of their set. Then, an iterative process is performed to
654 construct a directed acyclic graph (DAG), in which nodes correspond to groups of indices, and
655 directed edges connect two overlapping groups. Initially, an empty node is created as the starting
656 point of all paths. In the first iteration, a new node is created and annotated with the first S_i indices
657 for each size S_i of the given set *S*. As a result, *M* nodes are created and aligned with indices from
658 1 to S_i, constituting candidate versions of SW_1. Directed edges from the starting node to SW_1
659 nodes are created. In the next iteration, the SW_1 nodes are used as starting points to create
660 candidate nodes for SW_2. For an SW_1 with indices set *K*, the starting index of its subsequent SW_2
661 nodes is defined as *max(K)+1-ceiling(size(K)*α)*, which signifies that approximately the last *100*α*
662 percent of the SW_1 indices will also be included in the SW_2 nodes. New directed edges connect
663 each SW_1 with its subsequent SW_2 nodes. This iterative growth of the graph terminates when all
664 the samples have been mapped to SW nodes. Many nodes and edges could be generated
665 multiple times through the expansion of different paths. Thus, filtering is performed at the end of
666 each iteration to remove this redundancy. Finally, an empty node is created and connected with
667 the final SW nodes. All the directed paths that connect the starting and ending nodes constitute
668 potential representations of the pseudo-temporal sequence of disease stages (Supplementary
669 Fig. 9A).

670 Differential expression analysis with *DESeq2*[103] is used to examine each pair of connected SW
671 nodes as part of the optimal path. In the beginning, a pair is assumed as appropriate if the derived
672 number of differentially expressed genes (DEGs with adjusted p-value < 0.1) is in the range
673 *R=[100, 1000]*. These values were selected empirically, but the bounding of DEG size is based
674 on the hypothesis that smaller values indicate very weak differentiation, whereas higher values
675 might imply a strong differentiation signal, which can be explained in more than two SWs. Similarly
676 to the construction of the DAG, this step is performed in an iterative process, examining the graph
677 edges in batches based on their minimum distance from the starting node. To reduce the
678 execution time, a pruning step is used at the end of each iteration to remove edges and the
679 corresponding paths that do not satisfy the requirements of DEGs size. While this filtering reduces
680 the amount of *DESeq2* runs that need to be executed in the following iterations, it does not ensure
681 the connectivity of the graph, which is required in order to find a solution. Therefore, the limits of
682 range *R* are adjusted at each round so that the lower limit is not higher than the 80th percentile
683 and the upper limit is not lower than the 20th percentile of DEGs size distribution.

684 The second step typically produces numerous candidate paths, due to the relaxation of DEGs
685 size thresholds. Additional filtering is performed to reduce their amount and select the optimal
686 path. Firstly, the incoming edges of a single SW node are clustered based on the initial index of
687 the respective adjacent nodes. For each cluster, only the edge that corresponds to the highest
688 number of DEGs or (in case of equal numbers) the one that originates from the node with the
689 most extensive set of indices remains in the graph. Secondly, the edges which produce DEGs
690 less than the lower bound of the *R* range are selected to calculate their median number of DEGs.
691 This value is used to remove all the edges with the lower number of DEGs, ensuring that the
692 connectivity of the graph is not interrupted (Supplementary Fig. 9B). Finally, the product of the
693 following three max-scaled scores is used to prioritise the remaining candidate paths: 1. the
694 median number of DEGs, 2. the number of edges with DEGs more than the lower bound of *R* and
695 3. the coefficient of variation of the number of DEGs (Supplementary Fig. 9C). The described
696 method was implemented for the MASLD dataset (135 samples), using the following pairs of SW
697 sizes and overlap scores: ([6,15], 0.2), ([8,17], 0.25), ([8,24], 0.3), ([10,20], 0.3), ([12,25], 0.35).
698 Therefore, five graphs were constructed and filtered, and the candidate paths from all of them
699 were prioritized to find the optimal one.

700 Statistical analysis

701 Differential gene expression analysis was done using *DESeq2*[103] (v 1.26.0), and the Benjamini-
702 Hochberg method was applied to adjust the raw p-values to control the False Discovery Rate
703 (FDR). In the analyses that used the discrete stratification of patients' disease stages,
704 comparisons were made against the immediately previous stage. For all comparisons related to
705 the trajectory analysis (see section “Patients stratification into sliding windows”), sliding windows
706 (SWs) were compared to the previous SW (SW_x vs SW_{x-1}).
707

708 Disease stage-specific network analysis

709 We applied a network propagation-based approach to detect the deregulated components of the
710 MASLD network for each SW using the *igraph* R package (v 2.0.3). We used the personalised
711 *PageRank* algorithm[104] to run network propagation. This method accepts as input a vector of
712 weights for a set of nodes. We defined these weights based on the results of differential
713 expression analysis (for all genes except transcription factors) and TF activity analysis (for the
714 transcription factors). We were interested in finding a network signature for up- and down-
715 regulated conditions separately, so we created two weight vectors for each SW. In particular, we

defined the weights of node i (W_+ and W_-) using the direction of its differentiation (based on the logarithmic fold change or NES score) and the log-transformed adjusted p-value of the respective statistical test, to be bounded by a maximum value of 10:

- $If\ sign(deregulation) = +1: Weights_i : \{W_+ = min[-log_{10}(adj.pvalue) + 1, 10],\ W_- = 1\}$
- $If\ sign(deregulation) = -1: Weights_i : \{W_+ = 1,\ W_- = min[-log_{10}(adj.pvalue) + 1, 10]\}$

Network propagation was performed for each SW and direction of deregulation. To assess the statistical significance of the derived PageRank scores, we run the propagation task 1000 times using random weight distributions. Then, we calculated an empirical p-value for each node based on the ratio of its random scores that are lower than the real one:

$$pvalue = 1 - \frac{|random\ scores\ < PageRank\ score|}{1000}$$

We implemented this approach for three values of damping factor: [0.5, 0.7, 0.85], and we selected the gene nodes with p-value<0.05 in all three runs, as significantly propagated. Then, we extracted the respective subnetwork signature as a subgraph of the MASLD network, and we removed the significantly propagated nodes which belonged to small components (less than 4 members) because they have been identified due to isolated nodes with high input weight, far from the main affected component of the network. Finally, the activation score of a gene in a SW was defined as equal to its initial weight used for the network propagation if the gene had been included in the derived SW-specific deregulated component. Otherwise, the activation score was defined as equal to zero.

Definition of disease-relevant pathways

We used pathway enrichment analysis to link the deregulated subnetworks in each SW with molecular pathways using the Reactome database (as described in the section “Pathway enrichment analysis”). In order to create a unified and semantically non-redundant list of MASLD-relevant pathways and describe their activation in the pseudo-temporal trajectory, we clustered the enriched pathways from the different SWs. The clustering was performed based on the hierarchical structure of the Reactome database and the sets of deregulated genes that were associated with the enriched pathways. Initially, the Reactome graph was constructed using the

748 *igraph* Python package (v 0.4.11), connecting the ontological terms with semantic parent-child
749 relations. Then, we performed enrichment analysis to associate the genes of the global MASLD
750 network with an extended set of pathways (adjusted p-value < 0.01), and we extracted the
751 respective Reactome subgraph. This MASLD network-related subgraph was used as a reference
752 to filter the enriched pathways (adjusted p-value < 0.05) from the analysis of SW subnetwork
753 signatures, removing those which are not included in it. Then, we performed an iterative
754 agglomerative clustering on the remaining terms.

755 All terms with a distance greater than 4 were clustered to their parental terms because their
756 semantic meaning was primarily precise. We applied the following clustering steps to minimize
757 the redundancy of the gene sets of parent and child terms. Terms with distance 4, whose parent
758 had multiple child terms, were clustered to their parent. Terms with distance 2 or 3, whose parent
759 had only one child term, were clustered to their parent if the respective parent-child genomic
760 annotations were not identical (Jaccard similarity < 0.8). Otherwise, the child term was used to
761 substitute for its parent position on the graph. Finally, terms with distance 3, whose parent had
762 multiple child terms, were clustered to their parent, along with their sibling terms, if the respective
763 average genomic (Jaccard) similarity was more significant than 0.3, meaning that these terms
764 correspond to relatively similar gene sets. Otherwise, the clustering was not performed. We
765 selected the leaf nodes of the remaining graph to represent the disease-related pathways.
766 However, nodes with relatively low number of associated genes (a number less than the median
767 value minus one mean absolute deviation) were filtered out, in order to exclude extremely specific
768 pathways from the final list. For the genes that had been detected as deregulated, but they were
769 not linked to any of the final disease-related pathways, we defined additional terms. We used their
770 reference annotation to concatenate them into the generic mechanisms of Reactome, defining
771 the respective terms (e.g. “Other genes related to Metabolism”). Using this approach, we selected
772 82 MASLD-relevant pathways from the initial MASLD reference subgraph of 465 terms. We
773 calculated the activation of a pathway in a specific SW as the summation of the respective
774 activation scores of its genes, weighted by their pagerank centralities on the underlying pathway-
775 based subnetwork (Supplementary Fig. 5).

776 Cell-type deconvolution analysis

777 For the deconvolution analysis, we selected a 10X single-cell RNA-seq dataset from a healthy
778 human liver[105] as a reference. We re-analysed the single-cell dataset based on standard single-
779 cell analysis protocols[106]. We filtered out genes expressed in less than 3 cells and cells with less

780 than 200 genes expressed in them. Additionally, we filtered out cells with more than 7000 features
781 (putative doublets) and with more than 5% mitochondrial genes. To explore the suitability of the
782 dataset further, we proceeded with standard normalisation and vs (variance stabilising
783 transformation) scaling, selection of highly variable genes (HGV), PCA and UMAP analysis. For
784 the deconvolution analysis, we applied the CATD pipeline[107], using the raw counts as input. CATD
785 performs normalisation and additional preprocessing internally. After setting the preprocessing
786 parameters, we run the deconvolution analysis, selecting a subset of methods (EpiDISH[108],
787 bseqsc[109], DWLS[110], CIBERSORT[111], FARDEEP[112] and BayesPrism[113]). The consensus result
788 was calculated as the mean of cell-type proportions per cell-type across methods.
789

790 Identification of cell-type specific deregulated pathways

791 We assumed that the differentiated network signatures had been identified either due to changes
792 in cell type proportions or the deregulation of specific processes. In order to deconvolve the
793 differentiation signal and discern the deregulated processes that are related to specific cell types,
794 we applied the following workflow: Initially, we used the stratification of patients in SWs and the
795 calculated cell type proportions (see section “Cell type deconvolution analysis”) to model
796 Gaussian kernel-based density functions for the cell type proportions in each SW. We used these
797 kernels to generate a number of pseudo-bulk RNA-seq samples for each SW, with predefined
798 library sizes, sampling cells from the reference single-cell RNA-seq dataset of healthy liver. The
799 number of SW pseudo-samples and the size of count libraries were selected randomly from the
800 ranges [6,20] and [20e6, 50e6], respectively. Differential expression, TF activity, and network-
801 based analyses were performed for the pseudo-dataset, similar to the real one, assuming that the
802 results would represent the differentiation of the MASLD network, which was caused only by
803 changes in cell type proportions. We generated 30 random pseudo-datasets to average the
804 results from differential expression and TF activity analyses. The network analysis extracted up-
805 and down-regulated network components, which were compared with those of the real dataset. If
806 a gene was included in the real- and pseudo-signature of a specific SW and its activation score
807 in the pseudo-signature exceeded the 90% of that in the real-signature, it was considered to be
808 differentiated due to changes in cell composition within that SW. We removed these genes from
809 the MASLD-relevant pathways (see section “Definition of disease-relevant pathways”), assuming
810 that the differentiation signal of the remaining gene sets is associated with the deregulation of the
811 respective pathways.

812 In order to examine if the deregulation of these pathways originates from specific cell types, we
813 first create extended lists of cell type markers. We used the reference single dataset and the
814 *Seurat* R package (v. 5.0.3)[114] to identify the markers. We filtered the derived lists, excluding
815 genes which were not included in the global MASLD network. Then, we used the PCSF algorithm
816 (using the same parameters as in the section "Generation of global MASLD network") to amplify
817 each set of marker genes, integrating other genes from the MASLD network that are topologically
818 close to them. Finally, we assessed the relevance of the deregulated pathways to the liver cell
819 types using enrichment tests (Fisher's exact test for hypothesis testing and the Benjamini–
820 Hochberg method for the multiple hypothesis testing correction).

821 Prediction of patients' position on pseudo-temporal trajectory

822 For the prediction of the patients' position on the trajectory, we created a Random Forest
823 regression model. In each iteration, the dataset was split into training (80%) and testing (20%)
824 sets and run 1000 times to ensure optimal performance. The model's performance was evaluated
825 for 1000 iterations by calculating the mean MSE (Mean Squared Error) and mean R^2 (coefficient
826 of determination), explaining the percentage of trajectory variance explained by the selected gene
827 set. Additionally, the assessment of variable importance reduced the gene set, identifying the
828 most predictive genes and excluding those with negative variable importance scores.

829 We tested multiple gene sets in our model to determine the top-performing predictors. First, we
830 tested the 145 gene signature derived from our random forest classifier and a set of 194 integrated
831 liver-plasma targets identified by Govaere et al.[39]. Next, the 194 liver-plasma targets were refined
832 further according to their positive contribution (negative variable-importance-score genes were
833 removed), assessing whether restricting the gene set improves the overall model performance.
834 Finally, we evaluated the intersection of this refined gene set with the differentiated part of the
835 MASLD network derived from our analysis (24 genes).

836 A crucial part of this analysis was the transformation of the expression data into z-scores based
837 only on the gene subset being analysed in each run. For instance, when using the 194 gene set,
838 the z-score transformation was applied only for this set of genes. This approach ensures
839 predictions even for incomplete proteomic data (measurements of specific plasma proteins),
840 allowing potential real-time prediction of the patients' position on our trajectory based only on the
841 most informative plasma proteins.

842 The same gene sets were also tested on two independent datasets (Gubra and EPoS, described
843 above) to predict their patients' position on our trajectory. This time, the model training was based
844 on our internal UCAM/VCU dataset, with the external datasets used solely for testing and
845 validating our model's generalisability. Furthermore, random gene sets were used as controls,
846 run 1000 times for each gene set and averaging the predictive metrics, highlighting the predictive
847 value of our selected gene signature.

848 Identification of drug targets related to liver-plasma biomarkers

849 We used the intersection of liver-plasma biomarkers with the deregulated genes (24 genes) in the
850 MASLD network to identify potential drug targets for the different stages and molecular events of
851 the disease. Each biomarker had been identified as up- or down-regulated in specific SWs based
852 on our network analysis. We searched for the associated deregulated pathways in these SWs
853 and retrieved the gene sets of these pathways, extending the list of biomarkers with functionally
854 similar genes, which were also deregulated in the same disease stages. Then, we searched in
855 the ChEMBL database[115] for approved or clinical candidate drugs (max_phase : [2,3,4]) that target
856 these genes. Specifically, we search for drugs whose action potentially inverses the deregulation
857 of the respective targets (agonists/activators for the down-regulated and inhibitors/antagonists for
858 the up-regulated genes). Finally, we created a mapping of drugs and drug targets for each
859 molecular event related to the 24 biomarkers (Supplementary Table 13).

860 Data and code availability

861 All data used in this study is publicly available as described in the methods. The code used for
862 the analyses can be found here: https://github.com/kamzolas/MASLD---Continuous-trajectory-
863 approach

864 Acknowledgements

865 The work was supported by the European Molecular Biology Laboratory (IK, TK, CGB, AVP, IP,
866 HW, EP) and Open targets (TK, IP, EP); The EMBL GeneCore Facility is acknowledged for
867 support in transcriptomics data collection. EMBL IT Support is acknowledged for the provision of
868 computer and data storage servers. Funding from the Medical Research Council (MRC)
869 supported I. Kamzolas. MRC MDU, Caixa Foundation and Spanish Government ATRAE
870 programme support AVP research on MASLD.

871 # Author contributions

872 IK, TK: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation,
873 Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization. CGB, AVP:
874 Formal analysis, Investigation, HW: Data Curation, MV: Writing - Review & Editing, Supervision,
875 IP: Supervision, Funding acquisition, AVP: Conceptualization, Supervision, Resources, Funding
876 acquisition, Writing - Review & Editing EP: Conceptualization, Methodology, Resources,
877 Supervision, Project administration, Visualization, Funding acquisition, Writing - Original Draft,
878 Writing - Review & Editing.

879 # References

880 1. Chan, W. K. *et al.* Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD): A
881 State-of-the-Art Review. *Journal of obesity & metabolic syndrome* **32**, (2023).
882 2. Loomba, R. & Sanyal, A. J. The global NAFLD epidemic. *Nat. Rev. Gastroenterol. Hepatol.*
883 **10**, (2013).
884 3. Estes, C. *et al.* Modeling NAFLD disease burden in China, France, Germany, Italy, Japan,
885 Spain, United Kingdom, and United States for the period 2016-2030. *J. Hepatol.* **69**, (2018).
886 4. Wong, V. W.-S., Adams, L. A., de Lédinghen, V., Wong, G. L.-H. & Sookoian, S.
887 Noninvasive biomarkers in NAFLD and NASH - current progress and future promise. *Nat.*
888 *Rev. Gastroenterol. Hepatol.* **15**, 461–478 (2018).
889 5. Bedossa, P. Utility and appropriateness of the fatty liver inhibition of progression (FLIP)
890 algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of
891 nonalcoholic fatty liver disease. *Hepatology (Baltimore, Md.)* **60**, (2014).
892 6. McGill, D. B., Rakela, J., Zinsmeister, A. R. & Ott, B. J. A 21-year experience with major
893 hemorrhage after percutaneous liver biopsy. *Gastroenterology* **99**, (1990).
894 7. Ratziu, V. *et al.* Sampling variability of liver biopsy in nonalcoholic fatty liver disease.
895 *Gastroenterology* **128**, (2005).

896 8. Govaere, O. *et al.* Transcriptomic profiling across the nonalcoholic fatty liver disease
897 spectrum reveals gene signatures for steatohepatitis and fibrosis. *Sci. Transl. Med.* **12**,
898 (2020).
899 9. Hoang, S. A. *et al.* Gene Expression Predicts Histological Severity and Reveals Distinct
900 Molecular Profiles of Nonalcoholic Fatty Liver Disease. *Sci. Rep.* **9**, 12541 (2019).
901 10. Atabaki-Pasdar, N. *et al.* Predicting and elucidating the etiology of fatty liver disease: A
902 machine learning modeling and validation study in the IMI DIRECT cohorts. *PLoS Med.* **17**,
903 e1003149 (2020).
904 11. Anstee, Q. M., Seth, D. & Day, C. P. Genetic Factors That Affect Risk of Alcoholic and
905 Nonalcoholic Fatty Liver Disease. *Gastroenterology* **150**, (2016).
906 12. Eslam, M. & George, J. Genetic contributions to NAFLD: leveraging shared genetics to
907 uncover systems biology. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 40–52 (2019).
908 13. Chen, Y. *et al.* Genome-wide association meta-analysis identifies 17 loci associated with
909 nonalcoholic fatty liver disease. *Nat. Genet.* **55**, 1640–1650 (2023).
910 14. Fairfield, C. J. *et al.* Genome-Wide Association Study of NAFLD Using Electronic Health
911 Records. *Hepatology Communications* **6**, 297 (2022).
912 15. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a
913 histologically characterised cohort☆. *J. Hepatol.* **73**, 505–515 (2020).
914 16. Magwene, P. M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of biological
915 samples using microarray data. *Bioinformatics* **19**, 842–850 (2003).
916 17. Mukherjee, S. *et al.* Molecular estimation of neurodegeneration pseudotime in older brains.
917 *Nat. Commun.* **11**, 5781 (2020).
918 18. Young, A. L. *et al.* Data-driven modelling of neurodegenerative disease progression:
919 thinking outside the black box. *Nat. Rev. Neurosci.* **25**, 111–130 (2024).
920 19. Azzu, V. *et al.* Suppression of insulin-induced gene 1 (INSIG1) function promotes hepatic

921 lipid remodelling and restrains NASH progression. *Molecular metabolism* **48**, (2021).
922 20. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell
923 transcriptomics. *BMC Genomics* **19**, 1–16 (2018).
924 21. Suppli, M. P. *et al.* Hepatic transcriptome signatures in patients with varying degrees of
925 nonalcoholic fatty liver disease compared with healthy normal-weight individuals. *Am. J.*
926 *Physiol. Gastrointest. Liver Physiol.* **316**, (2019).
927 22. Barker, C. G. *et al.* Identification of phenotype-specific networks from paired gene
928 expression–cell shape imaging data. *Genome Res.* **32**, 750 (2022).
929 23. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
930 analysis. *BMC Bioinformatics* **9**, 1–13 (2008).
931 24. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* **52**,
932 D672–D678 (2023).
933 25. Córdova, G. *et al.* SMAD3 and SP1/SP3 Transcription Factors Collaborate to Regulate
934 Connective Tissue Growth Factor Gene Expression in Myoblasts in Response to
935 Transforming Growth Factor β. *J. Cell. Biochem.* **116**, 1880–1887 (2015).
936 26. Younesi, F. S., Miller, A. E., Barker, T. H., Rossi, F. M. V. & Hinz, B. Fibroblast and
937 myofibroblast activation in normal tissue repair and fibrosis. *Nat. Rev. Mol. Cell Biol.* **25**,
938 617–638 (2024).
939 27. Liu, D. *et al.* Ets-1 deficiency alleviates nonalcoholic steatohepatitis via weakening TGF-β1
940 signaling-mediated hepatocyte apoptosis. *Cell Death Dis.* **10**, 1–14 (2019).
941 28. Varga, T., Czimmerer, Z. & Nagy, L. PPARs are a unique set of fatty acid regulated
942 transcription factors controlling both lipid metabolism and inflammation. *Biochim. Biophys.*
943 *Acta* **1812**, 1007 (2011).
944 29. Vachliotis, I. D. & Polyzos, S. A. The Role of Tumor Necrosis Factor-Alpha in the
945 Pathogenesis and Treatment of Nonalcoholic Fatty Liver Disease. *Curr. Obes. Rep.* **12**,

946 191–206 (2023).
947 30. Tumor necrosis factor-α signaling in nonalcoholic steatohepatitis and targeted therapies. *J.*
948 *Genet. Genomics* **49**, 269–278 (2022).
949 31. CREB family: A significant role in liver fibrosis. *Biochimie* **163**, 94–100 (2019).
950 32. Luedde, T. & Schwabe, R. F. NF-κB in the liver—linking injury, fibrosis and hepatocellular
951 carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **8**, 108 (2011).
952 33. Luo, M., Li, T. & Sang, H. The role of hypoxia-inducible factor 1α in hepatic lipid
953 metabolism. *J. Mol. Med.* **101**, 487–500 (2023).
954 34. Mesarwi, O. A. *et al.* Hepatocyte Hypoxia Inducible Factor-1 Mediates the Development of
955 Liver Fibrosis in a Mouse Model of Nonalcoholic Fatty Liver Disease. *PLoS One* **11**,
956 e0168572 (2016).
957 35. Mann, D. A. & Smart, D. E. Transcriptional regulation of hepatic stellate cell activation. *Gut*
958 **50**, 891 (2002).
959 36. Tan, J. *et al.* HNF1α Controls Liver Lipid Metabolism and Insulin Resistance via Negatively
960 Regulating the SOCS-3-STAT3 Signaling Pathway. *Journal of Diabetes Research* **2019**,
961 5483946 (2019).
962 37. Montagner, A. *et al.* Liver PPARα is crucial for whole-body fatty acid homeostasis and is
963 protective against NAFLD. *Gut* **65**, 1202–1214 (2016).
964 38. Lu, J., McKinsey, T. A., Zhang, C. L. & Olson, E. N. Regulation of skeletal myogenesis by
965 association of the MEF2 transcription factor with class II histone deacetylases. *Mol. Cell* **6**,
966 (2000).
967 39. Govaere, O. *et al.* A proteo-transcriptomic map of non-alcoholic fatty liver disease
968 signatures. *Nature Metabolism* **5**, 572 (2023).
969 40. Chitturi, S. *et al.* The Asia-Pacific Working Party on Non-alcoholic Fatty Liver Disease
970 guidelines 2017-Part 2: Management and special groups. *J. Gastroenterol. Hepatol.* **33**,

971 (2018).
972 41. Younes, R. *et al.* Long-term outcomes and predictive ability of non-invasive scoring
973 systems in patients with non-alcoholic fatty liver disease. *Journal of hepatology* **75**, (2021).
974 42. Transcriptional regulation of hepatic stellate cells. *Adv. Drug Deliv. Rev.* **61**, 497–512
975 (2009).
976 43. Sun, Z. *et al.* Hydronidone induces apoptosis in activated hepatic stellate cells through
977 endoplasmic reticulum stress-associated mitochondrial apoptotic pathway. *J. Gastroenterol.*
978 *Hepatol.* (2024) doi:10.1111/jgh.16635.
979 44. Henkel, A. & Green, R. M. The Unfolded Protein Response in Fatty Liver Disease. *Semin.*
980 *Liver Dis.* **33**, 321 (2013).
981 45. Holzner, L. M. W. & Murray, A. J. Hypoxia-Inducible Factors as Key Players in the
982 Pathogenesis of Non-alcoholic Fatty Liver Disease and Non-alcoholic Steatohepatitis.
983 *Front. Med.* **8**, 753268 (2021).
984 46. Suzuki, T., Shinjo, S., Arai, T., Kanai, M. & Goda, N. Hypoxia and fatty liver. *World J.*
985 *Gastroenterol.* **20**, 15087 (2014).
986 47. Munsterman, I. D. *et al.* Extracellular matrix components indicate remodelling activity in
987 different fibrosis stages of human non-alcoholic fatty liver disease. *Histopathology* **73**,
988 (2018).
989 48. Ipsen, D. H., Lykkesfeldt, J. & Tveden-Nyborg, P. Molecular mechanisms of hepatic lipid
990 accumulation in non-alcoholic fatty liver disease. *Cell. Mol. Life Sci.* **75**, 3313 (2018).
991 49. Wnt/beta-catenin signaling and its modulators in nonalcoholic fatty liver diseases.
992 *Hepatobiliary Pancreat. Dis. Int* **22**, 333–345 (2023).
993 50. Ying, H.-Z. *et al.* PDGF signaling pathway in hepatic fibrosis pathogenesis and
994 therapeutics. *Mol. Med. Rep.* **16**, 7879 (2017).
995 51. TGF-β1 signaling can worsen NAFLD with liver fibrosis backdrop. *Exp. Mol. Pathol.* **124**,

996 104733 (2022).
997 52. Inhibition of Notch1 signaling reduces hepatocyte injury in nonalcoholic fatty liver disease
998 via autophagy. *Biochem. Biophys. Res. Commun.* **547**, 131–138 (2021).
999 53. Xu, X. *et al.* Targeted therapeutics and novel signaling pathways in non-alcohol-associated
1000 fatty liver/steatohepatitis (NAFL/NASH). *Signal Transduction and Targeted Therapy* **7**, 1–39
1001 (2022).
1002 54. Yingling, J. M. *et al.* Preclinical assessment of galunisertib (LY2157299 monohydrate), a
1003 first-in-class transforming growth factor-β receptor type I inhibitor. *Oncotarget* **9**, 6659–6677
1004 (2017).
1005 55. Luangmonkong, T. *et al.* Evaluating the antifibrotic potency of galunisertib in a human ex
1006 vivo model of liver fibrosis. *Br. J. Pharmacol.* **174**, 3107–3117 (2017).
1007 56. Ge, W.-S. *et al.* β-catenin is overexpressed in hepatic fibrosis and blockage of Wnt/β-
1008 catenin signaling inhibits hepatic stellate cell activation. *Mol. Med. Rep.* **9**, 2145–2151
1009 (2014).
1010 57. Gribben, C. *et al.* Acquisition of epithelial plasticity in human chronic liver disease. *Nature*
1011 **630**, 166–173 (2024).
1012 58. Vonderlin, J., Chavakis, T., Sieweke, M. & Tacke, F. The Multifaceted Roles of
1013 Macrophages in NAFLD Pathogenesis. *Cell Mol Gastroenterol Hepatol* **15**, 1311–1324
1014 (2023).
1015 59. Matchett, K. P., Paris, J., Teichmann, S. A. & Henderson, N. C. Spatial genomics: mapping
1016 human steatotic liver disease. *Nat. Rev. Gastroenterol. Hepatol.* 1–15 (2024).
1017 60. Heebøll, S. *et al.* Placebo-controlled, randomised clinical trial: high-dose resveratrol
1018 treatment for non-alcoholic fatty liver disease. *Scand. J. Gastroenterol.* (2016)
1019 doi:10.3109/00365521.2015.1107620.
1020 61. Vacca, M. *et al.* An unbiased ranking of murine dietary models based on their proximity to

1021 human metabolic dysfunction-associated steatotic liver disease (MASLD). *Nature*
1022 *Metabolism* **6**, 1178–1196 (2024).
1023 62. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
1024 requirements. *Nat. Methods* **12**, 357–360 (2015).
1025 63. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-
1026 throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
1027 64. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration
1028 of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, (2009).
1029 65. Zhang, Y., Jenkins, D. F., Manimaran, S. & Johnson, W. E. Alternative empirical Bayes
1030 models for adjusting for batch effects in genomic studies. *BMC Bioinformatics* **19**, 1–15
1031 (2018).
1032 66. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory
1033 inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
1034 67. Kleiner, D. E. *et al.* Design and validation of a histological scoring system for nonalcoholic
1035 fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).
1036 68. Fisher, R. A. Statistical Methods for Research Workers. *Breakthroughs in Statistics* 66–70
1037 (1992).
1038 69. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
1039 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–
1040 300 (1995).
1041 70. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network
1042 Analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).
1043 71. Stoica, P. & Selen, Y. Model-order selection: a review of information criterion rules.
1044 https://ieeexplore.ieee.org/document/1311138.
1045 72. Alvarez, M. J. *et al.* Network-based inference of protein activity helps functionalize the

1046 genetic landscape of cancer. *Nat. Genet.* **48**, 838 (2016).
1047 73. Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior
1048 knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res* **51**,
1049 10934–10949 (2023).
1050 74. Akhmedov, M. *et al.* PCSF: An R-package for network-based interpretation of high-
1051 throughput data. *PLoS Comput. Biol.* **13**, e1005694 (2017).
1052 75. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
1053 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
1054 76. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-
1055 wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
1056 77. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quantitative Biology* **1**,
1057 115–130 (2013).
1058 78. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory
1059 elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
1060 79. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**,
1061 D687–D692 (2021).
1062 80. Csárdi, G. *et al.* Igraph: Network analysis and visualization. *CRAN: Contributed Packages*
1063 The R Foundation https://doi.org/10.32614/cran.package.igraph (2006).
1064 81. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
1065 *Res.* **28**, (2000).
1066 82. Wrzodek, C., Büchel, F., Ruff, M., Dräger, A. & Zell, A. Precise generation of systems
1067 biology models from KEGG pathways. *BMC Syst. Biol.* **7**, (2013).
1068 83. Romero, P. *et al.* Computational prediction of human metabolic pathways from the
1069 complete human genome. *Genome Biol.* **6**, (2005).
1070 84. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat.*

1071 *Biotechnol.* **31**, (2013).
1072 85. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and
1073 exploration of pathway data. *Nucleic Acids Res.* **48**, D489–D497 (2019).
1074 86. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for
1075 literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
1076 87. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics
1077 analysis. *Mol. Syst. Biol.* **17**, e9923 (2021).
1078 88. OmnipathR. *Bioconductor* http://bioconductor.org/packages/OmnipathR/.
1079 89. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms
1080 and gene products. *Bioinformatics* **26**, 976–978 (2010).
1081 90. Aleksander, S. A. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
1082 91. Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its
1083 Application to Problems of Ambiguity in Natural Language. *jair* **11**, 95–130 (1999).
1084 92. An Information-Theoretic Definition of Similarity.
1085 https://dl.acm.org/doi/10.5555/645527.657297 doi:10.5555/645527.657297.
1086 93. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the
1087 semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
1088 94. Charmpi, K., Chokkalingam, M., Johnen, R. & Beyer, A. Optimizing network propagation for
1089 multi-omics data integration. *PLoS Comput. Biol.* **17**, e1009161 (2021).
1090 95. Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of
1091 NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).
1092 96. En masse organoid phenotyping informs metabolic-associated genetic susceptibility to
1093 NASH. *Cell* **185**, 4216–4232.e16 (2022).
1094 97. Graupera, I. *et al.* LiverScreen project: study protocol for screening for liver fibrosis in the
1095 general population in European countries. *BMC Public Health* **22**, 1–10 (2022).

1096 98. Hendriks, D. *et al.* Engineered human hepatocyte organoids enable CRISPR-based target
1097 discovery and drug screening for steatosis. *Nat. Biotechnol.* **41**, 1567–1581 (2023).
1098 99. Lee, E., Korf, H. & Vidal-Puig, A. An adipocentric perspective on the development and
1099 progression of non-alcoholic fatty liver disease. *Journal of hepatology* **78**, (2023).
1100 100. A network-based computational and experimental framework for repurposing compounds
1101 toward the treatment of non-alcoholic fatty liver disease. *iScience* **25**, 103890 (2022).
1102 101. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* **1**, 417–
1103 425 (2015).
1104 102. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–
1105 D621 (2020).
1106 103. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
1107 RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
1108 104. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and*
1109 *ISDN Systems* **30**, 107–117 (1998).
1110 105. Guilliams, M. *et al.* Spatial proteogenomics reveals distinct and evolutionarily conserved
1111 hepatic macrophage niches. *Cell* **185**, (2022).
1112 106. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell
1113 analysis. *Nature Biotechnology* **42**, 293–304 (2023).
1114 107. Vathrakokoili, P. A. *et al.* CATD: a reproducible pipeline for selecting cell-type
1115 deconvolution methods across tissues. *Bioinformatics advances* **4**, (2024).
1116 108. Zheng, S. C. *et al.* EpiDISH web server: Epigenetic Dissection of Intra-Sample-
1117 Heterogeneity with online GUI. *Bioinformatics (Oxford, England)* **36**, (2019).
1118 109. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas
1119 Reveals Inter- and Intra-cell Population Structure. *Cell systems* **3**, (2016).
1120 110. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data.

1121 *Nature Communications* **10**, 1–9 (2019).
1122 111. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
1123 *Nature Methods* **12**, 453–457 (2015).
1124 112. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor
1125 infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS*
1126 *computational biology* **15**, (2019).
1127 113. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution
1128 with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA
1129 sequencing in oncology. *Nature Cancer* **3**, 505–517 (2022).
1130 114. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of
1131 single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).
1132 115. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic acids*
1133 *research* **47**, (2019).
1134 116. Yilmaz, Y. & Eren, F. Serum biomarkers of fibrosis and extracellular matrix remodeling in
1135 patients with nonalcoholic fatty liver disease: association with liver histology. *European*
1136 *journal of gastroenterology & hepatology* **31**, (2019).
1137 117. Coilly, A., Desterke, C., Guettier, C., Samuel, D. & Chiappini, F. FABP4 and MMP9 levels
1138 identified as predictive factors for poor prognosis in patients with nonalcoholic fatty liver
1139 using data mining approaches and gene expression analysis. *Scientific reports* **9**, (2019).
1140 118. Toyoda, H. *et al.* Higher hepatic gene expression and serum levels of matrix
1141 metalloproteinase-2 are associated with steatohepatitis in non-alcoholic fatty liver diseases.
1142 *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals*
1143 **18**, (2013).
1144 119. Qi, S., Wang, C., Li, C., Wang, P. & Liu, M. Candidate genes investigation for severe
1145 nonalcoholic fatty liver disease based on bioinformatics analysis. *Medicine* **96**, (2017).

1146 ## Figure Legends

1147 **Figure 1. Overview of the approach used in this study to investigate MASLD progression.**
1148 **A.** Integration of public datasets and MASLD trajectory inference by the pseudo-temporal ordering
1149 of patient biopsy transcriptomic data **B.** Combining phenotype-specific gene co-expression
1150 analysis with upstream pathway inference to define a MASLD network. **C.** Cell type deconvolution
1151 to study the dynamics of cell type composition changes. **D.** Random Forest regression model for
1152 the prediction of patient's position on the inferred MASLD trajectory based on a set of biomarkers.
1153 Training was done on the initial dataset, and testing was done on independent datasets. The
1154 model deployed different gene signatures, while random gene sets were used as controls to
1155 validate the high model's performance. **E.** Identification of deregulated cell type-specific
1156 processes and potential biomarker genes that could be targeted in specific disease stages.

1157 **Figure 2. Pseudo-temporal ordering of patients captures disease progression and**
1158 **identifies MASLD gene signature. A.** Pseudo-temporal ordering of patients based on
1159 transcriptomic data recapitulates disease progression based on individual phenotypes, i.e.
1160 steatosis, ballooning, inflammation and fibrosis, and based on the NAS and MASLD scores. **B-C.**
1161 Pseudotemporal ordering of two independent orthogonal datasets based on the 145 genes that
1162 are most predictive of the trajectory provides a linear and clear separation of the disease stages
1163 (EPoS dataset (**B**) and Gubra dataset (**C**)). **D.** Functional enrichment analysis results of these
1164 145 genes using EnrichR, where the Combined Score (c) has been calculated as c = log(p_value)
1165 * z_score, with p_value representing the Fisher exact test outcome, and z-score the deviation
1166 from the expected rank.

1167 **Figure 3. Phenotype-specific gene modules are associated with the different MASLD**
1168 **variables. A. Network of the different gene co-expression modules i**s associated with NAS
1169 score, Steatosis, Ballooning, and Fibrosis. Each module is depicted by a different colour and can
1170 be linked with one or multiple variables. **B.** TFs whose regulons are enriched in the modules. The
1171 figure shows TFs per module coloured according to the adjusted p-value of the enrichment test.
1172 The presented TFs have been identified as enriched in multiple modules and deregulated at least
1173 in one stage of the disease trajectory (see sliding window analysis results). The complete list of
1174 TFs derived from the enrichment analysis is given in Supplementary Table 6. **C.** Enrichment of
1175 MASLD network in known MASLD genes curated from the literature (top) or from MSigDB
1176 (bottom).

1177 **Figure 4. Sliding window analysis to study the landscape of molecular changes along the**
1178 **MASLD trajectory. A.** Results from TF activity analysis in the discrete patients stratification and
1179 the pseudotemporal trajectory. The colour indicates the change in TF activity compared to the
1180 previous disease stage. **B.** Mean phenotype and NAS pathologist scores along the SW trajectory.

1181 **Figure 5. Results of cell type deconvolution analysis. A.** Cell type deconvolution of patient
1182 transcriptomic data along our MASLD trajectory **B.** Prediction of cell types with which the
1183 deregulated processes are associated after excluding processes that could have occurred by the
1184 mere change in abundance of cell types.

1185 **Figure 6. Random Forest regression model for the prediction of MASLD patients' position**
1186 **on the inferred disease trajectory. A.** Model performance using different gene sets in the
1187 primary UCAM/VCU dataset (5-fold cross validation), and validation in two independent datasets
1188 used solely for testing. Randomisation was done on 1000 times. **B.** Compared predicted and
1189 actual patient positions on the trajectory using an independent dataset (GUBRA), based solely on
1190 a 24-gene signature. **C.** Violin plot showing the distribution of the predicted patients' position on
1191 the trajectory (GUBRA dataset). **D.** Network-based deregulation analysis of the 24-gene signature
1192 to identify potential targets. The color scale represents the -log10(adjusted p-value) of differential
1193 expression. Genes are deemed deregulated in a SW if part of the respective deregulated
1194 subnetworks. Drugs in max phase 4 were retrieved for targets in the relative deregulated
1195 pathways.

1196 Supplementary Figure legends

1197 **Supplementary Figure 1. MASLD patients along the continuous transcriptomics-driven**
1198 **trajectory.** The colouring is based on the different scores given by the histopathologists. The
1199 triangles show the average horizontal position of each scoring group on the trajectory. **A.** NAS
1200 score (ranging from 0 to 7). **B.** Steatosis scoring (0-3). **C.** Inflammatory damage (0-3). **D.**
1201 Ballooning (0-2). **E.** Fibrosis (0-4).

1202 **Supplementary Figure 2. Violin plots illustrate the distribution of the patients in each**
1203 **histopathological scoring group. A.** Steatosis. **B.** Inflammatory damage. **C.** Ballooning. **D.**
1204 Fibrosis. **E.** NAS score. Each panel shows the Pearson correlation coefficient (R score) for the
1205 mean trajectory positions of each scoring group. The p-values are the results of ANOVA followed
1206 by Tukey's pairwise comparisons for the distributions of patients across the histopathological

1207 scores. With dashed lines are indicated the statistically significant groups (*p-value < .05, ** p-
1208 value < .005, *** p-value < .0005).

1209 **Supplementary Figure 3. Trajectory inference on independent MASLD transcriptomic**
1210 **datasets and selection of a subset of genes based on a Random Forest classification**
1211 **model. A.** Trajectory inference for the independent EPoS dataset using the whole transcriptome.
1212 Color coding depicts MASLD stages as published with the original RNA-Seq data. **B.** Trajectory
1213 inference for the independent GUBRA dataset. The population was characterised as Healthy
1214 (non-obese individuals without MASLD/MASH), Obese (healthy but obese individuals without
1215 MASLD/MASH), MASL, and MASH. **C.** Distribution of the "Mean Decrease Accuracy" scores
1216 derived from our Random Forest classification model. **D.** Selection of the 145 genes with positive
1217 scores, representing the final gene signature able to predict a patient's disease state, as well as
1218 key histopathology scores for steatosis, ballooning, inflammation, and fibrosis. **E.** Stage prediction
1219 of the GUBRA dataset based on a Random Forest classifier trained on the original dataset (Stage
1220 Division 1: predicting Healthy/Obese vs MASL vs MASH stages; Stage Division 2: predicting
1221 Healthy/Obese/MASL vs MASH stages).

1222 **Supplementary Figure 4.** Mean MASH component score in each Sliding Window (SW). The x-
1223 axis represents the 13 SWs, while the y-axis represents the mean score in each one of these
1224 SWs according to **A.** Steatosis, **B.** Inflammation, **C.** Ballooning, **D.** Fibrosis. The plots additionally
1225 show the Pearson's correlation coefficient (R score) and the corresponding p-value.

1226 **Supplementary Figure 5 Overview of modulation of modules and pathways along sliding**
1227 **windows. A.** Enrichment of histological phenotype-specific modules along our MASLD trajectory
1228 **B.** Reactome pathway enrichment analysis comparing the enrichment (and direction) using the
1229 Sliding Window (SW) approach. The results are derived from the differentially expressed genes
1230 and the deregulated transcription factors (Log2FoldChange, NES scores and FDR values) when
1231 comparing consecutive SWs (i.e., SW_x vs SW_{x-1}). The "red" and "blue" colours represent the up-
1232 and down-regulated pathways, respectively, according to the average activation score.

1233 **Supplementary Figure 6. Random Forest regression model for the prediction of MASLD**
1234 **patients' position on the inferred disease trajectory for the primary and validation**
1235 **(independent) dataset. A.** Predicted vs real patients' positions on the trajectory for the primary
1236 dataset. The dataset has been split into 80% training and 20% testing. The plot shows the average

1237 after running the model 1000 times. The different colours show the different MASLD stages
1238 according to the histopathological assessment. Only the 24 gene signature has been deployed
1239 for the prediction. **B.** Violin plot showing the distribution of patients on the inferred trajectory based
1240 on the predictions from panel A. **C-D.** Similar to panels "A" and "B", but for the independent
1241 validation dataset (EPoS). The model training was done entirely on the primary dataset. **E-F.**
1242 Violin plots showing the distribution of patients on the EPoS trajectory, based on the 24 gene
1243 intersection, according to the NAS and Fibrosis scores (as assessed by the histopathologists). **G-**
1244 **H** 1000 (G) and 24 (H) random genes are not able to place patients along the trajectory.

1245 **Supplementary Figure 7. Ranked independent dataset (Gubra) individuals according to**
1246 **their predicted trajectory positions.** The first half of the trajectory predicts the healthy/obese,
1247 while the second part predicts the MASL/MASH. The Mean Squared Error is based on the results
1248 from panel C.

1249 **Supplementary Figure 8. Sanity checks and validation for dataset integration**. **A.** PCA plot
1250 of the merged UCAM/VCU quantile normalised counts. There is a clear batch effect separating
1251 the patients into two clusters according to the dataset from which they are derived. **B.** COMBAT
1252 batch effect correction solved the problem of the batch effect with patients altogether in a cluster
1253 and a clear disease progression pattern from the right towards the later disease stages on the
1254 left. The outlier (sample 5) was removed from all downstream analyses. **C-D.** Normalised counts
1255 before and after batch effect correction, separately for each disease stage. The data suggest that
1256 COMBAT has not significantly changed the expression of the genes, given the high correlation
1257 ($R > 0.97$ in all MASLD stages in VCU (C) and UCAM (D) datasets). **E.** Correlation of the principal
1258 components (PCs) with the variables of interest. The first 10 PCs have been correlated using
1259 Pearson correlation with 6 continuous variables of interest: age, NAS score, Fibrosis, Fat
1260 (Steatosis), Ballooning, and Inflammation. **F.** Heatmap of the quantile normalised counts before
1261 and after the merging/batch effect correction for UCAM and VCU datasets. The rows show genes
1262 previously linked to MASLD[116–119]. The data are shown in a "fold change vs the first disease stage"
1263 of each dataset, respectively ("red" = increase; "blue" = decrease; "white" = unchanged; "green
1264 datasets": UCAM and VCU before correction; "blue datasets": UCAM and VCU after correction;
1265 the asterisks "*" in the column names indicate that the samples have been batch effect corrected).
1266 Note that the control group (first columns in each block) in the VCU dataset is CTRL, while in
1267 UCAM is MASL.

1268 ## Supplementary Table legends

1269 **Supplementary Table 1.** Information about all MASLD patients used in this study (UCAM and
1270 VCU datasets). Each patient is given the histopathology score (steatosis, ballooning,
1271 inflammation, fibrosis, NAS score) and the MASLD stage, as well as additional metadata (age,
1272 sex), the Sliding Window (SW) number assigned to the patient, the order in the trajectory (1 to
1273 136), the position according to PC1, the dataset (UCAM or VCU), and the batch corrected
1274 normalised expression value per gene.

1275 **Supplementary Table 2.** Patients' clinical characteristics when divided in CTRL, MASL, MASH
1276 F0-1, MASH F2, MASH F3-4. Statistical significance is assessed by ANOVA. The letters next to
1277 the values indicate post-hoc analysis significance based on Tukey's test; "[a]" means reference
1278 group; when groups show different letters, they should be considered statistically different at the
1279 post-hoc comparison; groups showing the same letter are statistically not-significant at the post-
1280 hoc comparison.

1281 **Supplementary Table 3.** Random forest's accuracy and AUC were tested on UCAM/VCU and
1282 GUBRA datasets using the top 145 genes. The UCAM/VCU dataset has been divided into training
1283 (80%) and testing (20%) sets and the metrics given in the table correspond to the average for
1284 running on the test set 1000 times. The gene signature predicts the different MASLD
1285 characteristics (Steatosis, Inflammation, Ballooning, Fibrosis) and the MASLD stage. For the
1286 GUBRA dataset predictions, the whole UCAM/VCU dataset has been used as a training set. The
1287 only available information for this external dataset is the MASLD stage. Finally, for both datasets,
1288 a random set of 20, 145, and 5000 genes (average for 1000 runs) verify the robustness of the
1289 selected genes.

1290 **Supplementary Table 4.** Phenotype-specific gene co-expression modules associated ("Y") or
1291 not associated ("N") with the MASLD variables. Each column includes the set of genes
1292 constituting each gene module.

1293 **Supplementary Table 5.** Reactome pathway enrichment analysis on the filtered gene sets of
1294 each module.

1295 **Supplementary Table 6.** Transcription Factor (TF) enrichment analysis for each significant-
1296 identified gene module.

1297 **Supplementary Table 7.** Reactome pathway enrichment analysis on the TFs of Supplementary
1298 Table 6.

1299 **Supplementary Table 8.** Unified inferred global MASLD network according to our analysis.

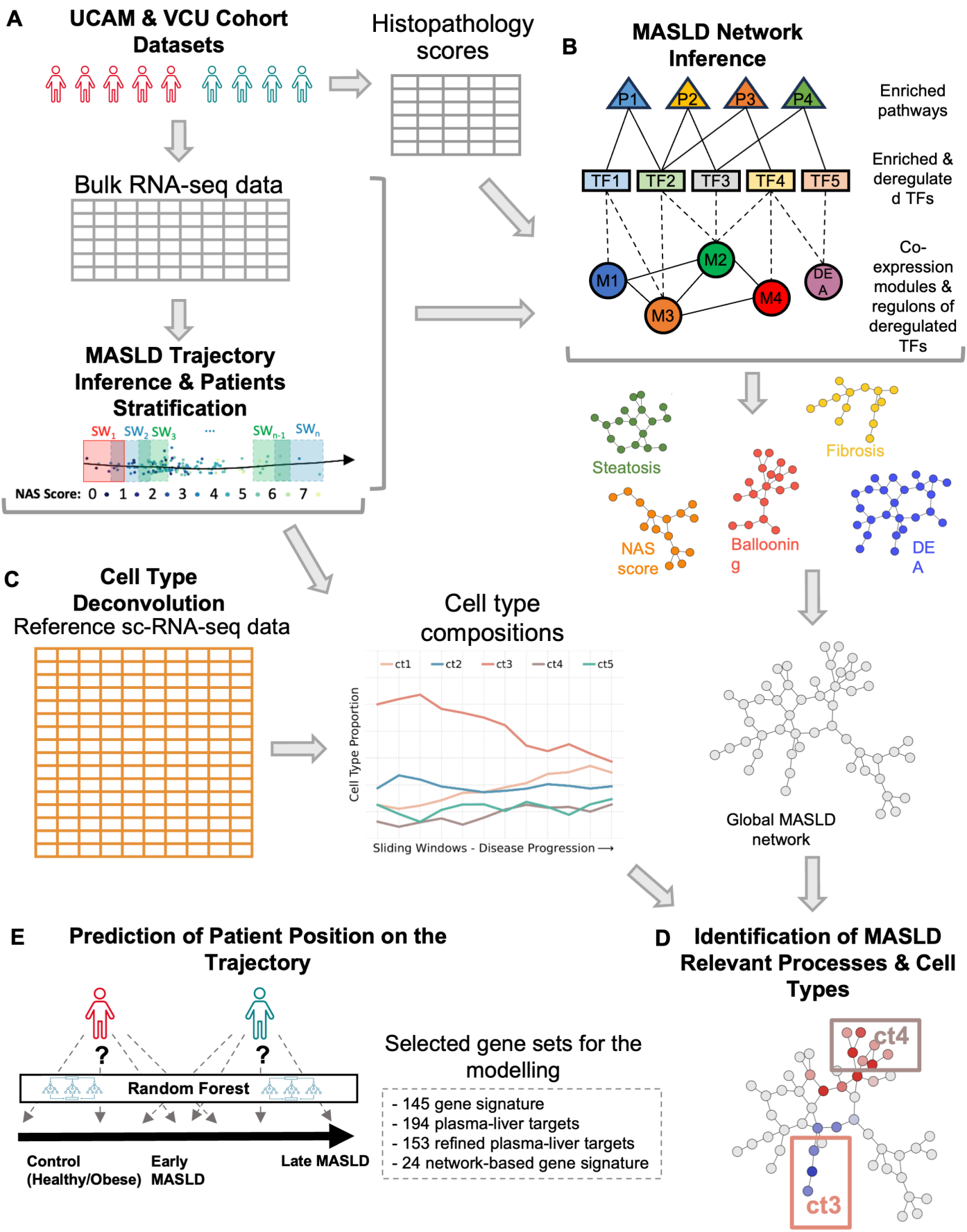
1300 **Supplementary Table 9.** Genes previously associated with MASLD.

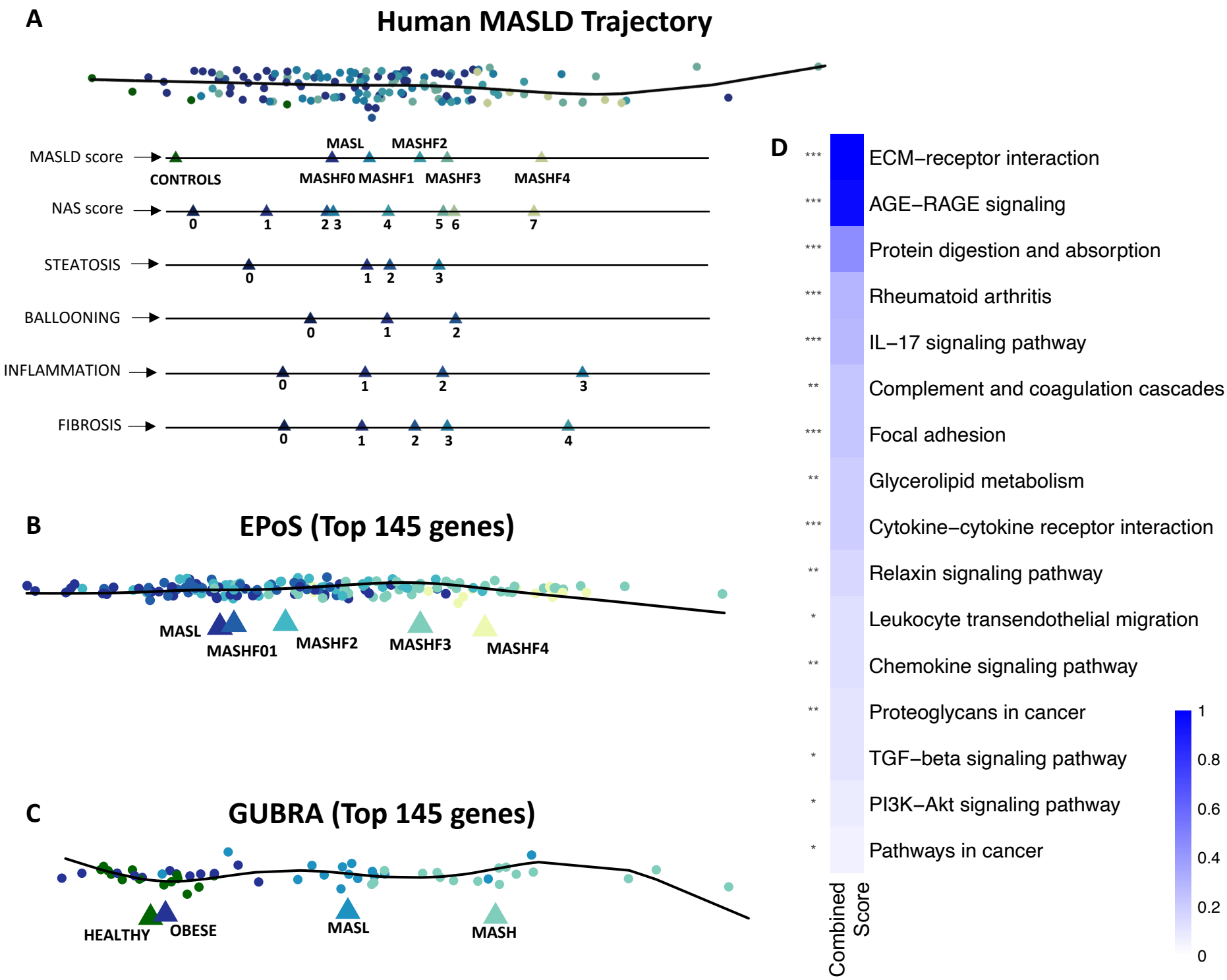
1301 **Supplementary Table 10.** Patient assignment to each of the 13 Sliding Windows (SWs).

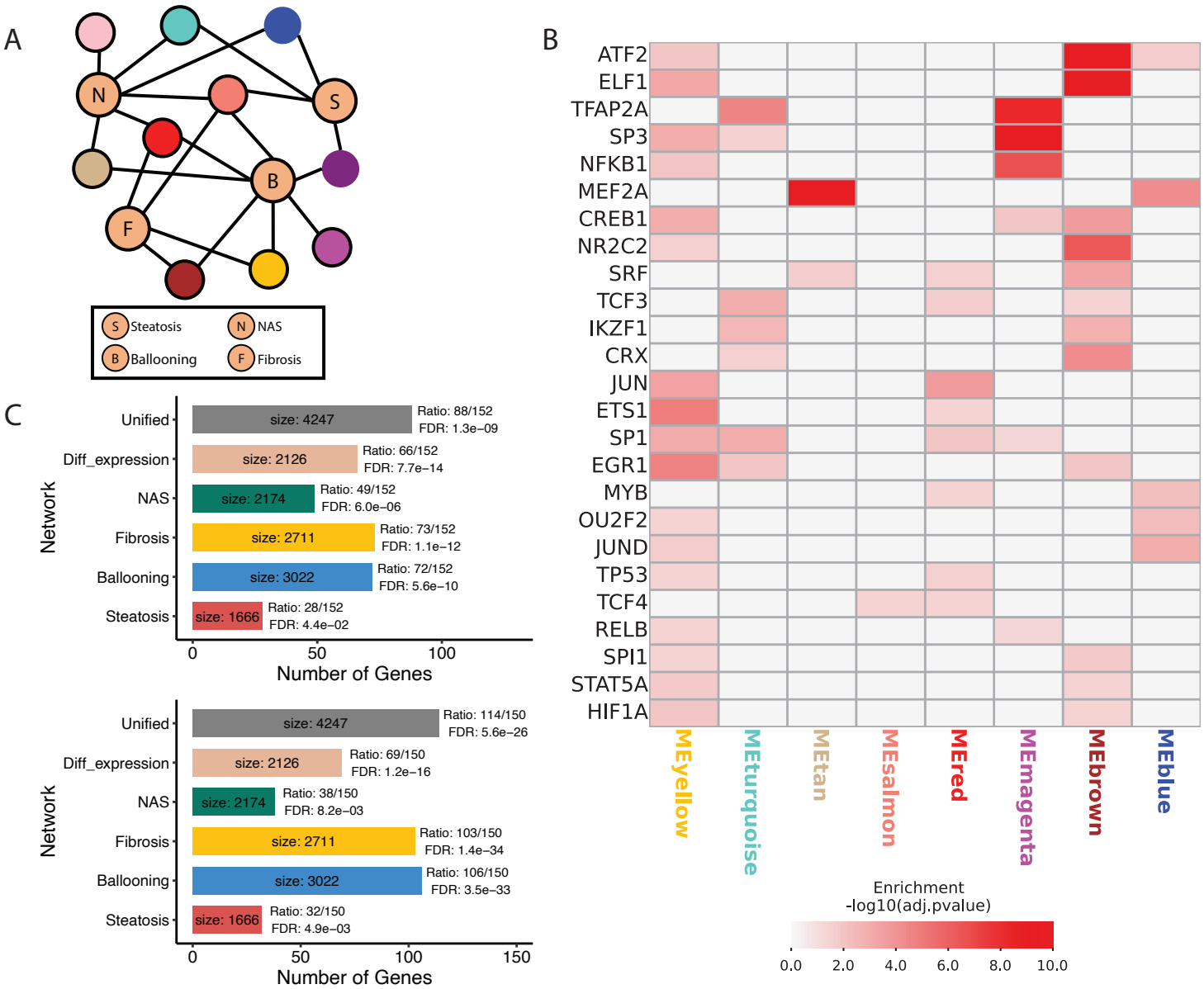
1302 **Supplementary Table 11.** VIPER analysis results comparing consecutive Sliding Windows
1303 (SWs). The results are derived from the differentially expressed genes of the previous steps
1304 (Log2FoldChange and p-values). The table is organised in 3-plets showing for each comparison
1305 the adjusted p-value and the Normalised Enrichment Score (NES). NES determines whether a
1306 TF has significantly more “activated” than “inhibited” predictions (NES > 0) or vice versa (NES <
1307 0).

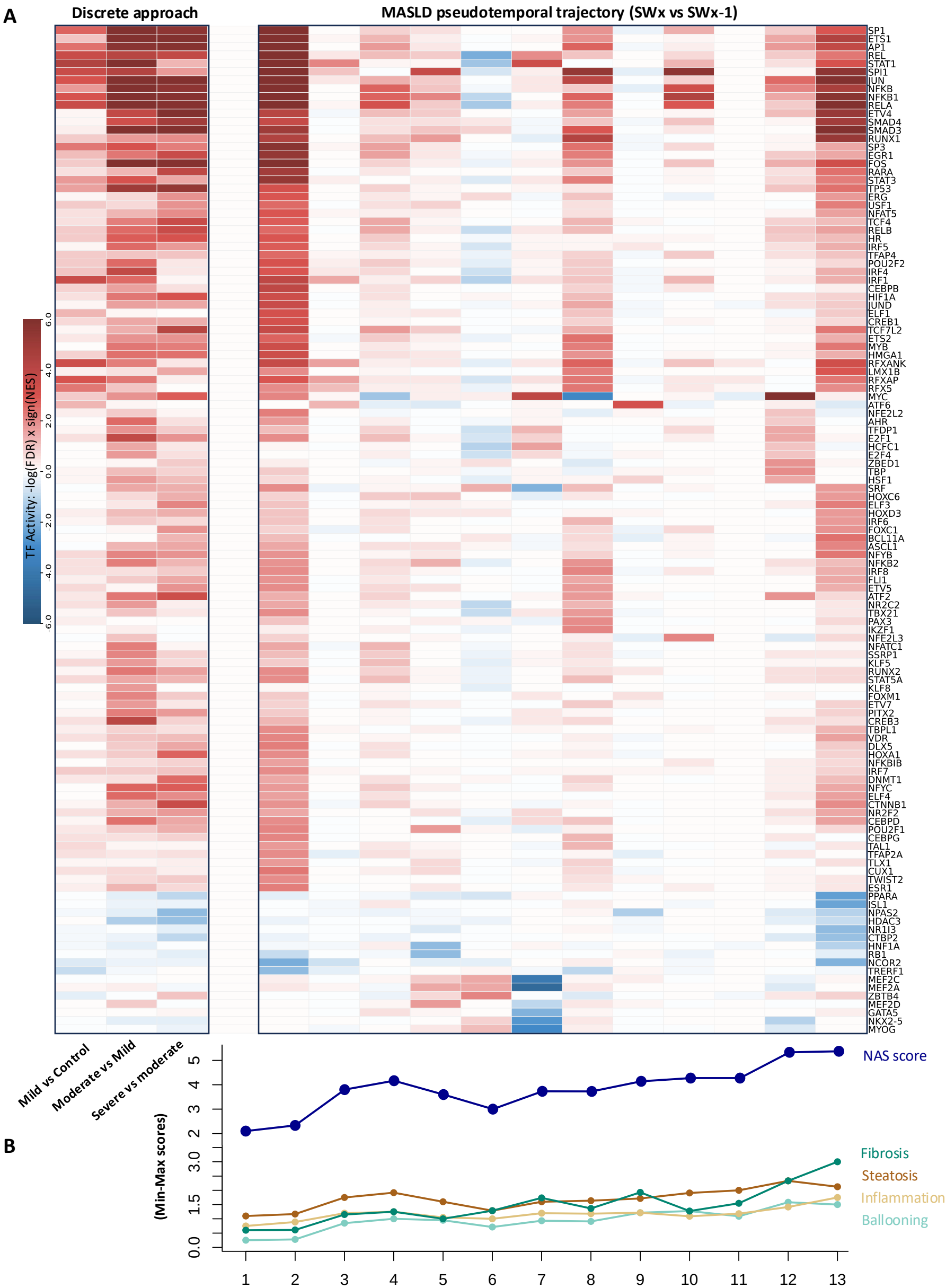
1308 **Supplementary Table 12.** Complete table of processes (pre-pruning) and associations to cell
1309 types.

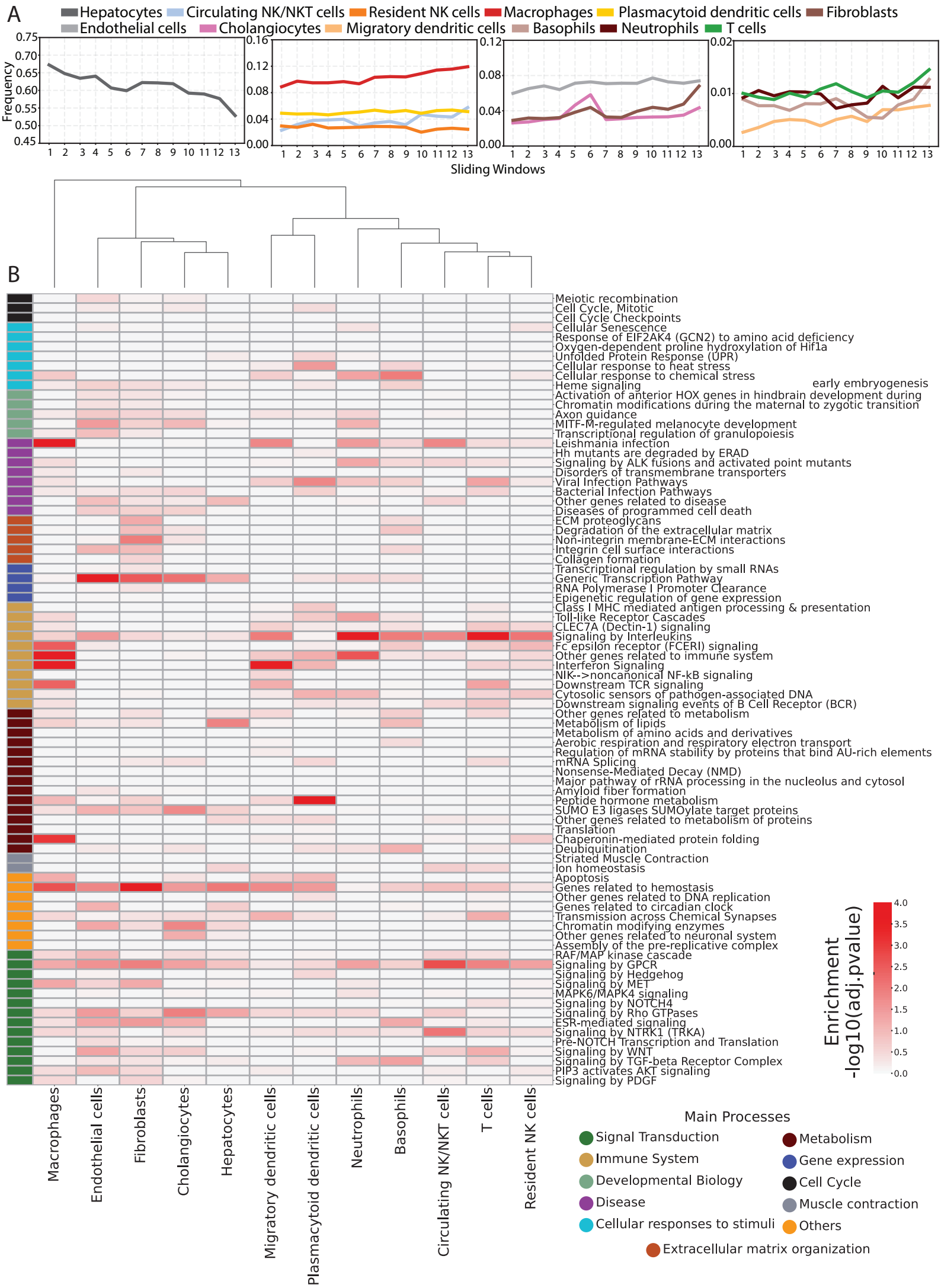
1310 **Supplementary Table 13.** Table of drugs known to target genes within the 24 gene biomarker-
1311 associated subgraphs extracted from our MASLD network

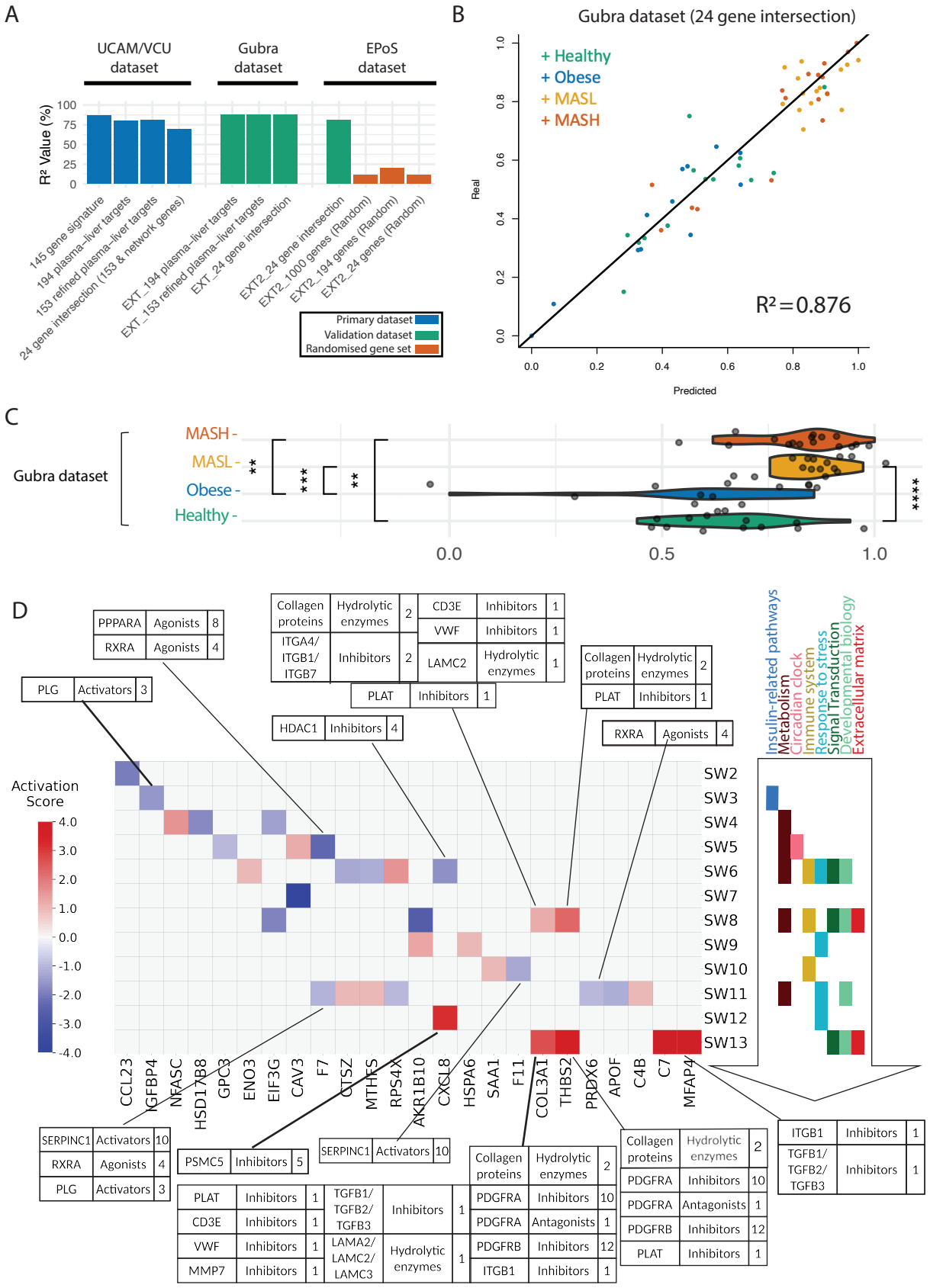

A
UCAM & VCU Cohort Datasets
Histopathology scores
Bulk RNA-seq data
MASLD Trajectory Inference & Patients Stratification
SW1 SW2 SW3 ... SWn-1 SWn
NAS Score: 0 1 2 3 4 5 6 7
B
MASLD Network Inference
P1 P2 P3 P4
TF1 TF2 TF3 TF4 TF5
M1 M2 M3 M4 DEA
Enriched pathways
Enriched & deregulated TFs
Co-expression modules & regulons of deregulated TFs
Steatosis
Fibrosis
NAS score
Ballooning
DEA
C
Cell Type Deconvolution
Reference sc-RNA-seq data
Cell type compositions
ct1 ct2 ct3 ct4 ct5
Cell Type Proportion
Sliding Windows - Disease Progression
Global MASLD network
D
Identification of MASLD Relevant Processes & Cell Types
ct4
ct3
E
Prediction of Patient Position on the Trajectory
?
?
Random Forest
Control (Healthy/Obese)
Early MASLD
Late MASLD
Selected gene sets for the modelling
- 145 gene signature
- 194 plasma-liver targets
- 153 refined plasma-liver targets
- 24 network-based gene signature


A
Human MASLD Trajectory
MASLD score
CONTROLS
MASL
MASHF0
MASHF1
MASHF2
MASHF3
MASHF4
NAS score
0
1
2 3
4
5 6
7
STEATOSIS
0
1 2
3
BALLOONING
0
1
2
INFLAMMATION
0
1
2
3
FIBROSIS
0
1
2
3
4
B
EPoS (Top 145 genes)
MASL
MASHF01
MASHF2
MASHF3
MASHF4
C
GUBRA (Top 145 genes)
HEALTHY
OBESE
MASL
MASH
D
*** ECM−receptor interaction
*** AGE−RAGE signaling
*** Protein digestion and absorption
*** Rheumatoid arthritis
*** IL−17 signaling pathway
** Complement and coagulation cascades
*** Focal adhesion
** Glycerolipid metabolism
*** Cytokine−cytokine receptor interaction
** Relaxin signaling pathway
* Leukocyte transendothelial migration
** Chemokine signaling pathway
** Proteoglycans in cancer
* TGF−beta signaling pathway
* PI3K−Akt signaling pathway
* Pathways in cancer
Combined Score
1
0.8
0.6
0.4
0.2
0


A
N
S
B
F
S Steatosis
N NAS
B Ballooning
F Fibrosis
B
ATF2
ELF1
TFAP2A
SP3
NFKB1
MEF2A
CREB1
NR2C2
SRF
TCF3
IKZF1
CRX
JUN
ETS1
SP1
EGR1
MYB
OU2F2
JUND
TP53
TCF4
RELB
SPI1
STAT5A
HIF1A
MEyellow
MEturquoise
MEtan
MEsalmon
MEred
MEmagenta
MEbrown
MEblue
Enrichment
-log10(adj.pvalue)
0.0
2.0
4.0
6.0
8.0
10.0
C
Network
Unified
Diff_expression
NAS
Fibrosis
Ballooning
Steatosis
size: 4247
Ratio: 88/152
FDR: 1.3e−09
size: 2126
Ratio: 66/152
FDR: 7.7e−14
size: 2174
Ratio: 49/152
FDR: 6.0e−06
size: 2711
Ratio: 73/152
FDR: 1.1e−12
size: 3022
Ratio: 72/152
FDR: 5.6e−10
size: 1666
Ratio: 28/152
FDR: 4.4e−02
0
50
100
Number of Genes
Network
Unified
Diff_expression
NAS
Fibrosis
Ballooning
Steatosis
size: 4247
Ratio: 114/150
FDR: 5.6e−26
size: 2126
Ratio: 69/150
FDR: 1.2e−16
size: 2174
Ratio: 38/150
FDR: 8.2e−03
size: 2711
Ratio: 103/150
FDR: 1.4e−34
size: 3022
Ratio: 106/150
FDR: 3.5e−33
size: 1666
Ratio: 32/150
FDR: 4.9e−03
0
50
100
150
Number of Genes


A
Discrete approach
MASLD pseudotemporal trajectory (SWx vs SWx-1)
TF Activity: -log(FDR) x sign(NES)
6.0
4.0
2.0
0.0
-2.0
-4.0
-6.0
SP1
ETS1
AP1
REL
STAT1
SPI1
JUN
NFKB
NFKB1
RELA
ETV4
SMAD4
SMAD3
RUNX1
SP3
EGR1
FOS
RARA
STAT3
TP53
ERG
USF1
NFAT5
TCF4
RELB
HR
IRF5
TFAP4
POU2F2
IRF4
IRF1
CEBPB
HIF1A
JUND
ELF1
CREB1
TCF7L2
ETS2
MYB
HMGA1
RFXANK
LMX1B
RFXAP
RFX5
MYC
ATF6
NFE2L2
AHR
TFDP1
E2F1
HCFC1
E2F4
ZBED1
TBP
HSF1
SRF
HOXC6
ELF3
HOXD3
IRF6
FOXC1
BCL11A
ASCL1
NFYB
NFKB2
IRF8
FLI1
ETV5
ATF2
NR2C2
TBX21
PAX3
IKZF1
NFE2L3
NFATC1
SSRP1
KLF5
RUNX2
STAT5A
KLF8
FOXM1
ETV7
PITX2
CREB3
TBPL1
VDR
DLX5
HOXA1
NFKBIB
IRF7
DNMT1
NFYC
ELF4
CTNNB1
NR2F2
CEBPD
POU2F1
CEBPG
TAL1
TFAP2A
TLX1
CUX1
TWIST2
ESR1
PPARA
ISL1
NPAS2
HDAC3
NR1I3
CTBP2
HNF1A
RB1
NCOR2
TRERF1
MEF2C
MEF2A
ZBTB4
MEF2D
GATA5
NKX2-5
MYOG
Mild vs Control
Moderate vs Mild
Severe vs moderate
B
(Min-Max scores)
0.0
1.5
3.0
2
3
4
5
NAS score
Fibrosis
Steatosis
Inflammation
Ballooning
1
2
3
4
5
6
7
8
9
10
11
12
13


A
Hepatocytes
Circulating NK/NKT cells
Resident NK cells
Macrophages
Plasmacytoid dendritic cells
Fibroblasts
Endothelial cells
Cholangiocytes
Migratory dendritic cells
Basophils
Neutrophils
T cells
Frequency
Sliding Windows
B
Meiotic recombination
Cell Cycle, Mitotic
Cell Cycle Checkpoints
Cellular Senescence
Response of EIF2AK4 (GCN2) to amino acid deficiency
Oxygen-dependent proline hydroxylation of Hif1a
Unfolded Protein Response (UPR)
Cellular response to heat stress
Cellular response to chemical stress
Heme signaling
early embryogenesis
Activation of anterior HOX genes in hindbrain development during
Chromatin modifications during the maternal to zygotic transition
Axon guidance
MITF-M-regulated melanocyte development
Transcriptional regulation of granulopoiesis
Leishmania infection
Hh mutants are degraded by ERAD
Signaling by ALK fusions and activated point mutants
Disorders of transmembrane transporters
Viral Infection Pathways
Bacterial Infection Pathways
Other genes related to disease
Diseases of programmed cell death
ECM proteoglycans
Degradation of the extracellular matrix
Non-integrin membrane-ECM interactions
Integrin cell surface interactions
Collagen formation
Transcriptional regulation by small RNAs
Generic Transcription Pathway
RNA Polymerase I Promoter Clearance
Epigenetic regulation of gene expression
Class I MHC mediated antigen processing & presentation
Toll-like Receptor Cascades
CLEC7A (Dectin-1) signaling
Signaling by Interleukins
Fc epsilon receptor (FCERI) signaling
Other genes related to immune system
Interferon Signaling
NIK-->noncanonical NF-kB signaling
Downstream TCR signaling
Cytosolic sensors of pathogen-associated DNA
Downstream signaling events of B Cell Receptor (BCR)
Other genes related to metabolism
Metabolism of lipids
Metabolism of amino acids and derivatives
Aerobic respiration and respiratory electron transport
Regulation of mRNA stability by proteins that bind AU-rich elements
mRNA Splicing
Nonsense-Mediated Decay (NMD)
Major pathway of rRNA processing in the nucleolus and cytosol
Amyloid fiber formation
Peptide hormone metabolism
SUMO E3 ligases SUMOylate target proteins
Other genes related to metabolism of proteins
Translation
Chaperonin-mediated protein folding
Deubiquitination
Striated Muscle Contraction
Ion homeostasis
Apoptosis
Genes related to hemostasis
Other genes related to DNA replication
Genes related to circadian clock
Transmission across Chemical Synapses
Chromatin modifying enzymes
Other genes related to neuronal system
Assembly of the pre-replicative complex
RAF/MAP kinase cascade
Signaling by GPCR
Signaling by Hedgehog
Signaling by MET
MAPK6/MAPK4 signaling
Signaling by NOTCH4
Signaling by Rho GTPases
ESR-mediated signaling
Signaling by NTRK1 (TRKA)
Pre-NOTCH Transcription and Translation
Signaling by WNT
Signaling by TGF-beta Receptor Complex
PIP3 activates AKT signaling
Signaling by PDGF
Macrophages
Endothelial cells
Fibroblasts
Cholangiocytes
Hepatocytes
Migratory dendritic cells
Plasmacytoid dendritic cells
Neutrophils
Basophils
Circulating NK/NKT cells
T cells
Resident NK cells
Enrichment
-log10(adj.pvalue)
Main Processes
Signal Transduction
Metabolism
Immune System
Gene expression
Developmental Biology
Cell Cycle
Disease
Muscle contraction
Cellular responses to stimuli
Others
Extracellular matrix organization


A
UCAM/VCU dataset
Gubra dataset
EPoS dataset
R² Value (%)
100
75
50
25
0
145 gene signature
194 plasma–liver targets
153 refined plasma–liver targets
24 gene intersection (153 & network genes)
EXT_194 plasma–liver targets
EXT_153 refined plasma–liver targets
EXT_24 gene intersection
EXT2_24 gene intersection
EXT2_1000 genes (Random)
EXT2_194 genes (Random)
EXT2_24 genes (Random)
Primary dataset
Validation dataset
Randomised gene set
B
Gubra dataset (24 gene intersection)
+ Healthy
+ Obese
+ MASL
+ MASH
Real
Predicted
R² = 0.876
C
Gubra dataset
MASH -
MASL -
Obese -
Healthy -
**
***
**
****
0.0
0.5
1.0
D
PPPARA Agonists 8
RXRA Agonists 4
PLG Activators 3
Collagen proteins Hydrolytic enzymes 2
ITGA4/ITGB1/ITGB7 Inhibitors 2
CD3E Inhibitors 1
VWF Inhibitors 1
LAMC2 Hydrolytic enzymes 1
PLAT Inhibitors 1
HDAC1 Inhibitors 4
Collagen proteins Hydrolytic enzymes 2
PLAT Inhibitors 1
RXRA Agonists 4
Insulin-related pathways
Metabolism
Circadian clock
Immune system
Response to stress
Signal Transduction
Developmental biology
Extracellular matrix
Activation Score
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
SW2
SW3
SW4
SW5
SW6
SW7
SW8
SW9
SW10
SW11
SW12
SW13
CCL23
IGFBP4
NFASC
HSD17B8
GPC3
ENO3
EIF3G
CAV3
F7
CTSZ
MTHFS
RPS4X
AKR1B10
CXCL8
HSPA6
SAA1
F11
COL3A1
THBS2
PRDX6
APOF
C4B
C7
MFAP4
SERPINC1 Activators 10
RXRA Agonists 4
PLG Activators 3
PSMC5 Inhibitors 5
SERPINC1 Activators 10
PLAT Inhibitors 1
CD3E Inhibitors 1
VWF Inhibitors 1
MMP7 Inhibitors 1
TGFB1/TGFB2/TGFB3 Inhibitors 1
LAMA2/LAMC2/LAMC3 Hydrolytic enzymes 1
Collagen proteins Hydrolytic enzymes 2
PDGFRA Inhibitors 10
PDGFRA Antagonists 1
PDGFRB Inhibitors 12
ITGB1 Inhibitors 1
Collagen proteins Hydrolytic enzymes 2
PDGFRA Inhibitors 10
PDGFRA Antagonists 1
PDGFRB Inhibitors 12
PLAT Inhibitors 1
ITGB1 Inhibitors 1
TGFB1/TGFB2/TGFB3 Inhibitors 1