Precision medicine in nonalcoholic fatty liver disease: New therapeutic insights from genetics and systems biology

Silvia Sookoian1,2 and Carlos J. Pirola1,3

1Institute of Medical Research A Lanari, School of Medicine, University of Buenos Aires, Autonomous City of Buenos Aires; Departments of 1Clinical and Molecular Hepatology and 2Molecular Genetics and Biology of Complex Diseases, Institute of Medical Research (IDIM), National Scientific and Technical Research Council (CONICET)-University of Buenos Aires, Autonomous City of Buenos Aires, Argentina

Despite more than two decades of extensive research focusing on nonalcoholic fatty liver disease (NAFLD), no approved therapy for steatohepatitis—the severe histological form of the disease—presently exists. More importantly, new drugs and small molecules with diverse molecular targets on the pathways of hepatocyte injury, inflammation, and fibrosis cannot achieve the primary efficacy endpoints. Precision medicine can potentially overcome this issue, as it is founded on extensive knowledge of the druggable genome/proteome. Hence, this review summarizes significant trends and developments in precision medicine with a particular focus on new potential therapeutic discoveries modeled via systems biology approaches. In addition, we computed and simulated the potential utility of the NAFLD polygenic risk score, which could be conceptually very advantageous not only for early disease detection but also for implementing actionable measures. Incomplete knowledge of the druggable NAFLD genome severely impedes the drug discovery process and limits the likelihood of identifying robust and safe drug candidates. Thus, we close this article with some insights into emerging disciplines, such as chemical genetics, that may accelerate accurate identification of the druggable NAFLD genome/proteome. (Clin Mol Hepatol 2020;26:461-475)

Keywords: Nonalcoholic fatty liver disease; Genetics; Steatohepatitis; Precision medicine; Complex traits

Abbreviations:
5-hmC, 5-hydroxymethylcytosine; ADIPOQ, adiponectin; AUROC, area under the receiver operating characteristics; FDR, false discovery rate; FGF21, fibroblast growth factor 21; GADGET, GlobAl Distribution of GEnetic Traits; GCKR, glucokinase regulator; GO, gene ontology; GPT, glutamic-pyruvic transaminase; GRS, genetic risk score; GWAS, genome-wide association study; HDAC, histone deacylase; HSD17B13, hydroxysteroid 17-beta dehydrogenase 13; LEP, leptin; MBOAT7, membrane bound O-acyltransferase domain containing 7; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis; NR1H3, nuclear receptor subfamily 1 group H member 3; OR, odds ratio; PNPLA3, patatin-like phospholipase domain containing 3; PPARA, peroxisome proliferator-activated receptor gamma and alpha; PRS, polygenic risk score; QR2, quinone reductase 2; RBP4, retinol binding protein 4; SREBF1, sterol regulatory element binding transcription factor 1; TM6SF2, transmembrane 6 superfamily member 2; TMEM4, transmembrane channel-like 4

Corresponding author : Silvia Sookoian
Institute of Medical Research A Lanari, School of Medicine, University of Buenos Aires, Combatientes de Malvinas 3150, Autonomous City of Buenos Aires 1427, Argentina
Tel: +54-11-52873905, Fax: +54-11-52873905
E-mail: ssookoian@intramed.net
https://orcid.org/0000-0001-5929-5470

Carlos J. Pirola
Institute of Medical Research A Lanari, School of Medicine, University of Buenos Aires, Combatientes de Malvinas 3150, Autonomous City of Buenos Aires 1427, Argentina
Tel: +54-11-52873903
E-mail: pirola.carlos@conicet.gov.ar
https://orcid.org/0000-0001-8234-4058

Received : Jun. 22, 2020 / Revised : Jul. 16, 2020 / Accepted : Jul. 26, 2020

Copyright © 2020 by Korean Association for the Study of the Liver
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
INTRODUCTION

Nonalcoholic fatty liver disease (NAFLD) is a complex disorder that affects a large proportion of the world population of all ages. The disease pathogenesis involves a myriad of factors, including genetic susceptibility and predisposing metabolic comorbidities, such as obesity and type 2 diabetes, as well as environmental exposure and lifestyle, which jointly shape the NAFLD epigenome. Yet, despite more than two decades of extensive research in the field of NAFLD, there is currently no approved therapy for nonalcoholic steatohepatitis (NASH)—the severe histological form of the disease. Moreover, none of the new drugs or small molecules with diverse molecular targets on the pathways of hepatocyte injury, inflammation, and fibrosis can achieve primary efficacy endpoints. Diverse factors have been postulated to contribute to the low success rate in NAFLD/NASH drug discovery, including lack of robust animal models needed for preclinical studies, insufficient target engagement or target modulation by the novel drugs, absence or insufficient demonstration of a proof-of-concept in early trials, and/or high false discovery rate (FDR) in phase 2 trials.

Likewise, it has been hypothesized that, as data on the candidate drugs are not only insufficient but are also not corroborated by genetic inactivation, pharmacological inhibition, antisense oligonucleotides, and/or small interfering RNAs, this poses an additional obstacle to achieve consistent and sustained effects on severe histological outcomes, including improvement in fibrosis scores.

The aforementioned pitfalls could potentially be overcome through systems biology analysis, aiming to integrate knowledge of signaling pathways, the genetic information of susceptibility genes, and multiple tissue-specific OMICs-related experiments that include large-scale transcriptomic, proteomic, and metabolomic profiles, and more recently metagenomics of the liver tissue. Furthermore, the approach founded on precision medicine is expected to enhance the effectiveness of novel therapies, including elucidation of predictors of drug response. Hence, this review summarizes significant trends and developments in precision medicine with focus on new potential therapeutic discoveries modeled by systems biology approaches.

THE PATH TOWARDS PRECISION MEDICINE: A SHORT CONCEPTUAL APPRAISAL

The ultimate goal of precision medicine is to develop precision treatment strategies that rely upon a holistic understanding of differences in genetic and underlying molecular pathogenic factors, as well as responses to environmental stressors, among patients. Figure 1 describes milestones in the path towards precision medicine, which include the integration of big data, comprising information from electronic health records of thousands of patients and machine learning strategies, such as artificial intelligence, to assist with the development of algorithms for combining and decoding such complex information. In addition, collections of biological samples in large biobanks linked to patient data increase the likelihood of finding robust disease pathogenesis signatures derived from OMICs state-of-the-art approaches. Knowledge integration and data modeling and analysis are vital processes at the interface with drug discovery. It should be emphasized that time and cost are presently the key limiting factors at these stages. However, as the technological advances progress further, it is expected that, in the near future, the gap between the discovery of potential drugs and clinical validation will be considerably narrowed. Still, even when complex algorithms aimed at multi-scale modeling of OMICs data succeed in identifying a potential drug target, only some of these medications will eventually be included into primary or secondary therapeutic protocols for treating the target disease. At this stage, at least three clinical scenarios will likely emerge. One possibility is that a putative drug “x” demonstrates not only a good safety profile but also succeeds in achieving the objective endpoints in a large group of patients (Fig. 1). Another possibility is that, in a small group of patients, “x” would be contraindicated for diverse reasons, including safety concerns in some vulnerable groups, such as children, pregnant women, or patients suffering from chronic kidney disease, advanced cardiovascular disease, etc. It is also possible that certain patients will require a different kind of drug because their underlying molecular and/or genetic profile does not align with the molecular profile of the drug “x” (Fig. 1).

In addition, pathway-derived drugs may emerge from drug repurposing. A comprehensive algorithm for drug repurposing or repositioning in NASH, which primarily relies on identifying and developing new uses for existing drugs, was recently published. In fact, drugs which had an acceptable safety profile, but failed in achieving the expected response for some diseases, could be used to treat a different condition. It is also noteworthy that a large
number of discovered drugs fail and never pass preclinical testing. Hence, the ultimate role of precision therapy relies on defining appropriate prediction strategies for implementing patient-based therapies.

**PRECISION MEDICINE AND LARGE-SCALE GENOME AND EXOME SEQUENCING DATA**

Extant studies on the genetic component of NAFLD indicate that, after 12 years following the discovery that an allele in a patatin-like phospholipase domain containing 3 (PNPLA3) variant (rs738409 [G], encoding 148M) was associated with increased hepatic fat and NAFLD disease severity, knowledge of the disease heritability is still incomplete. Indeed, available evidence indicates that homozygous carriers with the G-risk allele of rs738409 present 3.24-fold greater risk of higher liver necroinflammatory scores and 3.2-fold greater risk of developing fibrosis when compared with homozygous CC carriers.

Findings yielded by genome-wide association studies (GWAS) as a part of which the heritability of hepatic steatosis was explored at the population level and/or NASH was examined in patients with liver biopsy, consistently show that at least four loci (PNPLA3, transmembrane 6 superfamily member 2 [TM6SF2], glucokinase regulator [GCKR], and hydroxysteroid 17-beta dehydrogenase 13 [HSD17B13]) are involved in the genetic susceptibility of the disease (Fig. 2). Conflicting results have been published regarding the rs641738 C/T located in transmembrane channel-like 4 (TM4C) exon 1 (p.Gly17Glu) and 500 bases downstream of the membrane bound O-acyltransferase domain containing 7 (MBOAT7; TM4C/MBOAT7), which were initially de-
scribed in Italian population but could not be replicated in other populations around the world, including a large cohort of European patients partaking in a GWAS, for whom NAFLD was diagnosed by liver biopsy.

Targeting disease-associated genes has been proven successful in the treatment of numerous human diseases, cancers in particular. In this field, target drugs are associated with the effect of a mutant protein and/or are designed to interfere at the gene or protein level. However, for this goal to be fully realized, full knowledge of the target genes/proteins is required, not only at all molecular levels, but also at the level of chemical-gene/protein interactions, which would ultimately allow identification of potential active ligands.

Likewise, drug discovery and precision medicine require complete understanding of new technologies as well as gene and protein biology of the selected target. Radar plots depicted in Figure 2 show the discrepancy in the level of knowledge on the four loci reproducibly implicated in the biology of NAFLD and the disease severity. For example, the plots show that, while extensive knowledge on PNPLA3 at different levels already exists, ranging from metabolic aspects to epigenetic information and traits associations, that pertaining to HSD17B13—the newly discovered gene with a putative loss-of-function variant implicated in protective effects against NASH and severe histological stages—is limited.

**Figure 2.** Illumination graph of major genetic modifiers of NAFLD and NASH. Radar plot and knowledge table depicting the variety of information obtained by Pharos (https://pharos.nih.gov/) for PNPLA3, TM6SF2, GCKR, and HSD17B13. These radial plots summarize the level of accumulated knowledge about each target. The greater the number of spikes in the plot, the greater the variety with spike length indicating the quantity of that particular knowledge. The radar chart allows gene-attribute associations as recorded by the Harmonizome to be visualized. The tables below the charts represent the top five knowledge attributes in the illumination graph. The knowledge value property is on a 0–1 scale. PNPLA3, patatin-like phospholipase domain containing 3; TM6SF2, transmembrane 6 superfamily member 2; GCKR, glucokinase regulator; HSD17B13, hydroxysteroid 17-beta dehydrogenase 13; NAFLD, nonalcoholic fatty liver disease; OR, odds ratio; SNPs, single nucleotide polymorphisms; NASH, nonalcoholic steatohepatitis.
at all levels of gene and protein biology (Fig. 2). The rs72613567 insertion/deletion variant, the functional consequence of which is a splice donor variant of the \textit{HSD17B13}, represents an interesting model of a candidate molecule for treating NASH and fibrosis. In fact, although the information about druggable binding domains in \textit{HSD17B13} is scarce due to the lack of an experimental 3D structure, other members of the protein family, such as \textit{HSD17B11} with high homology, present putative binding pockets for small molecules.

As indicated in the \textit{PNPLA3} illumination graph, considerable knowledge has been accumulated on the gene and protein, including gene-associated associations, protein interactions, and high PubMed score; yet, no information on the active ligands of the protein is currently available (Fig. 2).

The protein encoded by \textit{PNPLA3} is a triacylglycerol lipase that mediates triacylglycerol hydrolysis, mostly in adipocytes. The encoded protein, which appears to be membrane-bound, may be involved in the energy usage/storage balance in adipocytes. Figure 3A shows area under the receiver operating characteristics (AUROC; 0.951) for predicted functional processes linked to \textit{PNPLA3}, in-

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure3.png}
\caption{PNPLA3 predicted functional associations. Predicted biological processes (GO) (A) and upstream transcription factors (CheEA) (B) assessed by the Harmonizome (http://amp.pharm.mssm.edu/Harmonizome/gene/PNPLA3). Tables show the top 10 predictions (the number provided in the first column). Table explanation: If a gene (gene set) shares high correlation with known members of a gene set, it is assigned a high z-score. Known functions/gene set associations are highlighted in green. AUROC is provided by the algorithm available in the ARCHS4 (massive mining of publicly available RNA-seq data from human and mouse)\textsuperscript{58} accessible at https://amp.pharm.mssm.edu/archs4/. Specifically, AUROC shows how well-known annotations are recovered by the ARCHS\textsuperscript{4} algorithm. GO, gene ontology; PNPLA3, patatin-like phospholipase domain containing 3; AUROC, area under the receiver operating characteristics; CheEA, ChIP enrichment analysis. *From published ChIP-chip, ChIP-seq, and other transcription factor binding site profiling studies.\textsuperscript{59}}
\end{figure}
including triglyceride catabolic process (gene ontology [GO]: 0019433) and triglyceride biosynthetic process (GO: 0019432) (prediction was done by the Harmonizome). Importantly, extensive mining of publicly available RNA-seq data acquired through human and mouse experiments (https://amp.pharm.mssm.edu/archs4/) uncovered interesting upstream transcription factors (AUROC, 0.666) in Figure 3B, including RXR (retinoid x receptor), LXR (liver x receptor, nuclear receptor subfamily 1 group H member 3), and CLOCK (circadian locomotor output cycles protein kaput, formally known as circadian clock regulator). The protein encoded by the CLOCK plays a central role in the regulation of circadian rhythms. In our previous publications, we reported for the first time that CLOCK genetic variation is associated with obesity and NAFLD.\textsuperscript{31,32} The haplotype of rs1554483G and rs4864548A was found to be associated with a 1.8-fold risk of overweight status or obesity,\textsuperscript{32} whereas rs1554483 was shown to be associated with all histological traits of NASH, including fibrosis.\textsuperscript{31} Considered jointly, this body of evidence suggests that the putative cross-talk between PNPLA3 and CLOCK could explain the link among NAFLD genetic susceptibility, the environment, and the circadian regulation of liver metabolism. However, further experiments are required to prove this hypothesis.

\textit{TM6SF2}, of which rs58542926 C/T (E167K) was initially associated with liver fat accumulation and aminotransferase levels in a large GWAS study\textsuperscript{21} and further replicated in subsequent studies,\textsuperscript{21,33-35} encodes for a protein involved in lipid metabolism.\textsuperscript{12} Diverse areas of molecular knowledge gained on this gene/protein are presented in Figure 2.

Radar plot of GCKR presented in Figure 2, of which rs7800994 presents a very modest\textsuperscript{36} effect (odds ratio [OR], 1.2) on NAFLD biology,\textsuperscript{1,2} and shows that considerable gene/protein knowledge, including at least 64 putative ligands, has been already acquired in this domain.\textsuperscript{4}

Precision medicine has emerged as a result of comprehensive knowledge of the druggable genome/proteome. Thus, advancing the chemical genetics research, which is based on the screening of low-molecular weight compounds that act by binding to specific receptors/proteins, is crucial to move this promising research domain forward in the right direction. It is worth mentioning that the incomplete knowledge on the druggable genome of NAFLD/NASH severely undermines the drug discovery progress and reduces the chances of having robust and safe drug candidates. Therefore, the substantial gap between the knowledge of NAFLD-predisposing genes and that related to putative protein ligands needs to be urgently addressed.

**NAFLD AND THE PUTATIVE CLINICAL BENEFITS OF POLYGENIC RISK SCORES (PRSs)**

Estimating the susceptibility risk of a given patient to develop a particular disease and/or to progress into severe disease stages is the ultimate aim of precision medicine. Most researchers concur that the PRS distribution, which is based on the sum of all independent risk single nucleotide polymorphisms (SNPs; ideally weighted by their size effects in a given population), could be approximated by the Gaussian (normal) curve (Fig. 4). PRSs are theoretically designed to explain the relative risk of a disease, as these scores provide information on how a person compares with others with different genetic susceptibility background. However, PRSs do not necessarily follow normal distribution, due to several factors, including differences in the population structure or admixture.

In the case of NAFLD and NASH, PRSs could be conceptually very advantageous not only for allowing early disease detection, but also for implementing timely actionable measures (Fig. 4). For example, invasive diagnostic approaches, such as liver biopsy, as well as early pharmacological intervention, would be advised for high-risk populations (those at the right-tail of the PRS distribution curve pertaining to the relevant population) whereas low-risk individuals (i.e., those on the left-tail of the curve) would be monitored until clinical risk becomes evident (Fig. 4). For those deemed at low or medium risk, which probably applies to the large majority of the affected patients, lifestyle changes would be advised, including regular physical activity and dietary modifications aimed at optimizing body weight and controlling the key metabolic risk factors (lipid traits and glucose metabolism).

Despite these benefits, several important concerns related to the clinical implementation of PRSs also exist, as noted in Figure 4. In particular, use of PRSs in clinical settings will remain impractical until the heritability of NAFLD is fully elucidated, and rare and familial forms of the disease are revealed. Inclusion of additional genetic information will certainly aid in overcoming these issues as the predictive power of PRSs improves and the proportion of individuals at risk diminishes.

For instance, early detection of a rare nonsense GCKR mutation (rs149847328, p.Arg227Ter) in a NAFLD patient with associated comorbidities, including morbid obesity and type 2 diabetes, when combined with prompt pharmacological intervention, could potentially prevent or even reverse the disease progression into liver cirrhosis.\textsuperscript{37}

Furthermore, the currently available knowledge on the repro-
ducibility and replication of genetic variants of NAFLD across diverse populations around the world is insufficient (Fig. 4), and the effect sizes of most of the variants are yet to be established. Indeed, in the large majority of GWAS focusing on NAFLD, the genetic susceptibility in patients of European ancestry has been examined. This creates a considerable gap in extant knowledge, as SNP-based information on Caucasian and/or people of European descent may not be relevant for inferring the relative risk of developing the disease. Those with scores located at the left and right tail of the distribution curve will respectively carry very low and very high risk. NAFLD, nonalcoholic fatty liver disease; PRS, polygenic risk score; OR, odds ratio; GWAS, genome-wide association study.

Figure 4. Polygenic risk score in NAFLD: Advantages and challenges. Theoretical frame for a PRS for NAFLD, showing advantages and potential caveats. The figure shows a typical bell-shaped distribution, in which scores pertaining to most individuals will be in the middle, indicating average risk of developing the disease. Those with scores located at the left and right tail of the distribution curve will respectively carry very low and very high risk. NAFLD, nonalcoholic fatty liver disease; PRS, polygenic risk score; OR, odds ratio; GWAS, genome-wide association study.
tioned SNPs (PNPLA3-rs738409, TM6SF2-58542926, HSD17B13-rs72613567, and GCKR-780094) and the intergenic variant LYPLAL1-rs12137855. To compute the NAFLD-PRS, we used the GlobAl Distribution of GEnetic Traits (GADGET) web server available at https://gadget.biosci.gatech.edu/compute/; the formula used to compute the score is based on the original description of Chande et al. The GADGET web server provides access to publicly available genotype data sourced from the 1000 Genomes Project (1KGP) Phase 3 data release and individual trait SNP sets parsed directly from the NHGRI-EBI GWAS Catalog annotations (https://www.ebi.ac.uk/gwas/).

Box plots representing NAFLD-PRS distribution by five major continental groups (Africa, Europe, South and East Asia, and America) are shown in Figure 5. According to this information, the overall predicted risk to individuals entailed by the presence of NAFLD-implicated variants in their genomes is about ~0.20. Nevertheless, the PRS or genetic risk score (GRS) reflects disparities in NAFLD risk levels across different populations (Fig. 5), which may be due to the disparities in genetic knowledge or may indicate real differences in the genetic risk.

Thus, our analysis emphasizes the potential caveats of implementing this strategy globally. Of course, our simulation-based approach, which is based on information available in public databases, did not allow us to control for all possible effects, including demographic variables and/or other population risk estimates. Nonetheless, the GRS presented here provides a good starting point for illustrating the current situation, as well as for further investigations aimed at closing the gap in the knowledge needed to generate valuable advances in this field. It is worth noting that, as is the case for almost all human traits, variance in the NAFLD genetic risk within each population is much greater than among continental groups. This fact does not, however, imply that there is no continental group-specific risk profile. Available evidence indicates that certain trends exist at the global level, whereby the lowest GRS is associated with African population, intermediate GRS relates to European and Southeast Asian, and the highest GRS to East Asian and Admixed American groups.

Figure 5. NAFLD PRSs (GRS) across the five major continental population groups. Box plots show population-specific distributions of genetic variants that have been associated with NAFLD in the literature (PNPLA3-rs738409, TM6SF2-58542926, HSD17B13-rs72613567, GCKR-780094, and the intergenic variant LYPLAL1-rs12137855), as well as medians and standard deviations. Admixed American (n=347, 0.0857±0.0387), African (n=661, 0.0306±0.0265), East Asian (n=504, 0.0837±0.0386), European (n=503, 0.0589±0.0351), Southeast Asian (n=489, 0.0589±0.0364). GRS: the relative risk of developing NAFLD based on the total number of variants associated with the disease the individual carries. The relative genetic risk of NAFLD within the population is shown as log ORs, with F and P denoting summarized linear regression. The formula by which the GRS was calculated can be found in the original contribution of Chande et al. NAFLD, nonalcoholic fatty liver disease; PRS, polygenic risk score; OR, odds ratio; GRS, genetic risk score; PNPLA3, patatin-like phospholipase domain containing 3; TM6SF2, transmembrane 6 superfamily member 2; HSD17B13, hydroxysteroid 17-beta dehydrogenase 13; GCKR, glucokinase regulator.
NAFLD, PLEIOTROPIC RELATIONSHIPS, AND PRECISION MEDICINE

Available evidence suggests that genetic factors associated with NAFLD exhibit similar patterns of correlation with genetic factors related to other complex diseases. For instance, findings yielded by large GWAS studies indicate that PNPLA3-rs738409 and TM6SF2-rs58542926 are associated with extra-hepatic traits, including hematological (plateletcrit and count) and lipid traits, and some other interesting pharmacogenetic associations (http://www.phenoscanner.medschl.cam.ac.uk/) (Fig. 6). Importantly, the minor allele frequency of these two variants across different populations supports the prevalent view that the major genetic modifiers of NAFLD are likely ancestry-specific. Therefore, this observation should be specifically examined when designing precision medicine strategies.

The genetic pleiotropy between the aforementioned variants and non-liver related traits includes known NAFLD-associated comorbidities, such as cardiovascular risk. Phenotypic covariation presents not only significant challenges in clinical practice but also imposes tremendous constraints on identifying novel therapeutic targets. The clinical paradox of TM6SF2-rs58542926 C>T is a clear example of that. The C (Glu167) allele has been consistently associated with increased cardiovascular risk, and the T allele (Lys167) is known to be associated with a higher risk for NAFLD and NASH. These opposite effects are dependent on genetic factors such as PNPLA3 and TM6SF2.
on circulating and liver triglyceride levels, respectively. Consequently, TM6SF2 does not seem to be a useful drug target because any impact on the protein would eventually lower blood lipids, which will in turn reduce the risk of myocardial infarction, while simultaneously increasing the risk of developing NAFLD.\textsuperscript{12}

As explained in previous paragraphs, drug development is a long process characterized by highly uncertain outcomes. At present, 10–15 years typically elapse between target discovery and clinical application. Hence, some computational solutions, including modeling and over-representation analysis grounded in systems biology, may assist with more accurate prediction of drug candidates based on disease-associated genes/proteins. To illustrate this concept, we employed two different strategies. First, we leveraged existing information on the molecular targets (genes/proteins) involved in NAFLD/NASH pathogenesis and performed over-representation analysis using a drug-related functional database. The training set of genes/proteins was obtained by literature-data mining offered by the Genie web server (http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=281)—a tool that computes associations of genes with diseases using biomedical

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{Prediction of genetic-drug/chemical interaction profiles. (A) Over-representation analysis using a drug-related functional database (drug_GLAD4U). The training set of genes/proteins was obtained by literature-data mining offered by the Genie web server. Cut-offs used: \( P < 0.01 \) for abstracts and \( P < 0.01 \) for FDR for genes. The list of retrieved genes/proteins was used to perform over-representation analysis in the web server WebGestalt (WEB-based Gene SeT Analysis Toolkit). The bar chart shows ten categories that passed the FDR <0.05, with the gene number denoting the number of genes in the training list that belong to each of the drug categories. (B) Gene-chemical interactions. Gene target prediction was performed using the Comparative Toxicogenomics Database available at (http://ctdbase.org). The list of chemicals was manually curated to restrict interactions based on human data. According to the Comparative Toxicogenomics Database available, chemical-gene and protein interactions are curated from the published literature. Interactions may be retrieved by chemical, interaction type, gene, organism, or Gene Ontology annotation. Tutorial and algorithms by which the list of gene-chemical interactions was done are available at http://ctdbase.org and http://ctdbase.org/documents/ctd_resource_guide.pdf. FDR, false discovery rate; PNPLA3, patatin-like phospholipase domain containing 3; TM6SF2, transmembrane 6 superfamily member 2; GCKR, glucokinase regulator; HSD17B13, hydroxysteroid 17-beta dehydrogenase 13; DEET, N,N-diethyl-3-methylbenzamide; NAD, nicotinamide adenine dinucleotide.}
\end{figure}
literature annotations. Using this approach, 901 abstracts from PubMed were retrieved using the search terms “fatty liver” and “human” (taxonomic identifier 9606; no literature extension by orthology) with the abstracts from the all PubMed database serving as the background set. Supplementary Table 1, intended for online publication only, shows the final ranked list comprising of 938 genes/proteins.

The aforementioned list of retrieved genes/proteins was used to perform over-representation analysis in the webserver WebGestalt (WEB-based Gene Set Analysis Toolkit)—a functional enrichment analysis tool available at http://www.webgestalt.org/. Drug terms were downloaded from ParmGKB by the WebGestalt, and individual drug terms associated with genes were inferred using the GLAD4U option. Accordingly, the drug enrichment results based on the NAFLD training set are shown in Figure 7A. The bar chart shows ten categories that passed the FDR <0.05, whereby cardiac therapy and cardiovascular system were overrepresented by the largest number of genes (263 and 591, respectively). Other expected drug categories are anti-inflammatory agents and biguanides, which are probably justified by the significant enrichment of genes/proteins associated with inflammation and glucose metabolism in the training set. In fact, the top ten genes/proteins of the training list were ADIPOQ (adiponectin), PNPLA3, PPARG and PPARA (peroxisome proliferator-activated receptor gamma and alpha), FGF21 (fibroblast growth factor 21), RBP4 (retinol binding protein 4), GPT (glutamic-pyruvic transaminase), SREBF1 (sterol regulatory element binding transcription factor 1), LEP (leptin), and NR1H3 (nuclear receptor subfamily 1 group H member 3, also known as liver X receptor alpha). Lastly, the analysis revealed several anti-infection drugs, which were also expected because the training set presents several genes/proteins associated with immune response. Hence, the chart may be useful for inferring drug classes or drug compounds that could be repurposed for the treatment of NASH.

Our second strategy was based on using the four genes largely and reproducibly associated with NAFLD and NASH (PNPLA3, TM6SF2, GCKR, and HSD17B13) to predict curated chemical–gene/protein interactions in the Comparative Toxicogenomics Database available at (http://ctdbase.org) (Fig. 7B). The aim of this approach was to infer and/or uncover potential associated disease mechanisms from genetic predisposing factors that can yield biologically informative insights. Notably, some drugs were consistently found to interact with three of the loci, for example valproic acid (valproate) (Fig. 7B).

Valproate appears to impact on fatty acid metabolism, and the use of valproate has been linked to the development of obesity and probably NAFLD. Furthermore, valproate acts as a direct histone deactylase (HDAC) inhibitor. While tissue-specific DNA methylation in NAFLD and NASH, including 5-hydroxymethylcytosine (5-hmC) has been previously studied, the role of other epigenetic mechanisms, including acetylation and deacetylation

<table>
<thead>
<tr>
<th>KEGG pathway ID*</th>
<th>Pathway name</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa:00071</td>
<td>Fatty acid metabolism</td>
</tr>
<tr>
<td>hsa:00140</td>
<td>C21-Steroid hormone metabolism</td>
</tr>
<tr>
<td>hsa:00232</td>
<td>Caffeine metabolism</td>
</tr>
<tr>
<td>hsa:00250</td>
<td>Alanine, aspartate and glutamate</td>
</tr>
<tr>
<td>hsa:00280</td>
<td>Valine, leucine and isoleucine</td>
</tr>
<tr>
<td>hsa:00380</td>
<td>Tryprophan metabolism</td>
</tr>
<tr>
<td>hsa:00410</td>
<td>Beta-alanine metabolism</td>
</tr>
<tr>
<td>hsa:00590</td>
<td>Arachidonic acid metabolism</td>
</tr>
<tr>
<td>hsa:00591</td>
<td>Linoleic acid metabolism</td>
</tr>
<tr>
<td>hsa:00640/hsa:00650</td>
<td>Propanoate metabolism/butanoate</td>
</tr>
<tr>
<td>hsa:00830</td>
<td>Retinol metabolism</td>
</tr>
<tr>
<td>hsa:00980</td>
<td>Metabolism of xenobiots by cytochrome</td>
</tr>
<tr>
<td>hsa:00982/hsa:00983</td>
<td>Drug metabolism</td>
</tr>
<tr>
<td>hsa:03220</td>
<td>PPAR signaling pathway</td>
</tr>
<tr>
<td>hsa:04012</td>
<td>ErbB signaling pathway</td>
</tr>
<tr>
<td>hsa:04062</td>
<td>Chemokine signaling pathway</td>
</tr>
<tr>
<td>hsa:04080</td>
<td>Neuroactive ligand-receptor interaction</td>
</tr>
<tr>
<td>hsa:04110</td>
<td>Cell cycle</td>
</tr>
<tr>
<td>hsa:04270</td>
<td>Vascular smooth muscle contraction</td>
</tr>
<tr>
<td>hsa:04310</td>
<td>Wnt signaling pathway</td>
</tr>
<tr>
<td>hsa:04340</td>
<td>Hedgehog signaling pathway</td>
</tr>
<tr>
<td>hsa:04510</td>
<td>Focal adhesion</td>
</tr>
<tr>
<td>hsa:04610</td>
<td>Complement and coagulation cascades</td>
</tr>
<tr>
<td>hsa:04660/hsa:04662</td>
<td>T cell receptor signaling pathway/B cell receptor signaling pathway</td>
</tr>
<tr>
<td>hsa:04722</td>
<td>Neurotrophin signaling pathway</td>
</tr>
<tr>
<td>hsa:04810</td>
<td>Regulation of actin cytoskeleton</td>
</tr>
<tr>
<td>hsa:04910</td>
<td>Insulin signaling pathway</td>
</tr>
</tbody>
</table>

Prediction was performed by the Supertarget tool available at http://bioinformatics.charite.de/supertarget/index.php?site=drugs. KEGG, Kyoto Encyclopedia of Genes and Genomes; PPAR, peroxisome proliferator-activated receptor gamma and alpha.

*Pathway IDs correspond to KEGG.

Table 1. Prediction of valproic acid pathways
of histones, remains to be fully ascertained. Likewise, the use of valproate has been associated with the inhibition of mitochondrial beta-oxidation and peroxisomal stimulation in rodent livers, which reinforces the concept that the progression of NAFLD into severe clinical and histological forms involves mitochondrial dysfunction. Table 1 provides a complete list of predicted valproic acid-KEGG pathways, which involve alanine, aspartate and glutamate metabolism, arachidonic acid metabolism, retinol metabolism, PPAR and insulin signaling pathway, regulation of actin cytoskeleton, and hedgehog signaling pathway, among many other pathways relevant to the NAFLD pathogenesis. In addition, the list of valproate-linked pathways contains propanoate and butanoate metabolism that has been linked to the NASH-associated tissue microbiome.

Another compound that appears to be linked to three of the genes (PNPLA3, TM6SF2, and GCKR) is quercetin, an antioxidant phenolic heterocyclic compound that is a specific quinone reductase 2 (QR2) inhibitor (Fig. 7B). Collectively, these genetic-chemical interactions provide valuable information about cellular and biological mechanisms of disease, suggesting that this strategy may be used as a complement in the target-based drug development as well.

**LIMITATIONS OF THE SYSTEMS BIOLOGY APPROACH**

Some limitations of the systems biology approach must be highlighted, including the restricted possibility of adequately addressing the gender dimension of the disease. Sexual dimorphism is observed not only in the prevalence of NAFLD but also in the disease pathogenesis and diverse histological outcomes. Although sex differences substantially contribute to the biology of NAFLD, there is limited information on the sex-specific genetic architecture of the disease. We have addressed this aspect by meta-regression analysis of studies that assessed the effect of rs738409 on NAFLD, and we found a negative correlation between the male proportion in the studied populations and the effect of the SNP on liver fat content, suggesting that sexual dimorphism might be involved in the impact of the variant on NAFLD development. Likewise, a recent study that involved large-scale analysis of transcriptomic profiles from human livers suggested the sexually dimorphic nature of NASH and its link with fibrosis and responses to drugs.

Unrevealing the genetic mechanisms that contribute to sex-specific NAFLD risk should be urgently addressed in further candidate gene association or GWAS studies of NAFLD. Elucidating the sex-specific genetic architecture of NAFLD represents an important area for future research. Besides, there are many other important genetic and epigenetic factors, including mitochondrial genetics, that play a substantial role in the disease biology, and that deserve a more detailed analysis.

**CONCLUSIONS**

The utility of the systems biology approach for accelerating the NASH drug discovery process remains to be established. Still, to attain its full potential, refined genomic strategies must be implemented to increase knowledge of genetic susceptibility across all ancestry groups. Although theoretically powerful, PRSs need to be validated and built, not only on validated GWAS-variants but also on robust and global genetic information, preferably specific to each ethnic group. Finally, disciplines such as chemical genetics must be used in tandem with traditional drug discovery approaches to accelerate the progress of precision medicine in NASH and to reveal the druggable NASH genome/proteome.

**Authors’ contribution**

CJP and SS designed the study, performed the analyses, analyzed and interpreted the data, and prepared and wrote the manuscript. Both authors have read and approved the final manuscript.

**Acknowledgements**

This study was supported by grant numbers PICT 2015-0551, PICT 2016-0135, PICT 2018-0620 and PICT 2018-0889 (Agencia Nacional de Promoción Científica y Tecnológica, CONICET, CONICET Proyectos Unidades Ejecutoras 2017, grant number PUE 0055.

The authors apologize to the colleagues whose works could not be cited owing to manuscript length limitations.

**Conflicts of Interest**

The authors have no conflicts to disclose.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at Clinical and Molecular...
REFERENCES


Hepatology website (http://www.e-cmh.org).


