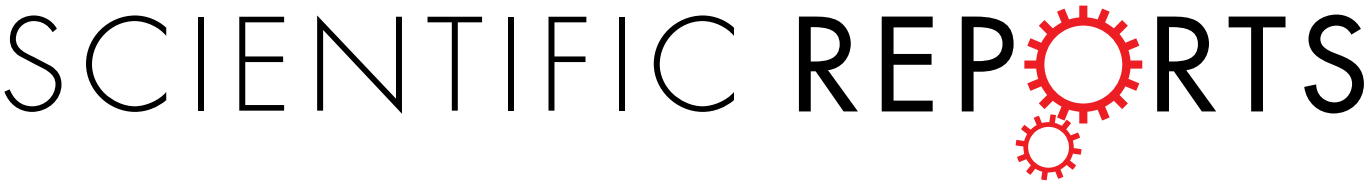

OPEN

Received: 18 August 2016
Accepted: 20 January 2017
Published: 07 March 2017

A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains

G. Craig Wood[1], Xin Chu[1], George Argyropoulos[2], Peter Benotti[1], David Rolston[1], Tooraj Mirshahi[1], Anthony Petrick[1], John Gabrielson[1], David J. Carey[1], Johanna K. DiStefano[3], Christopher D. Still[1] & Glenn S. Gerhard[1,2]

Non-alcoholic fatty liver disease (NAFLD) represents a spectrum of conditions that include steatohepatitis and fibrosis that are thought to emanate from hepatic steatosis. Few robust biomarkers or diagnostic tests have been developed for hepatic steatosis in the setting of obesity. We have developed a multi-component classifier for hepatic steatosis comprised of phenotypic, genomic, and proteomic variables using data from 576 adults with extreme obesity who underwent bariatric surgery and intra-operative liver biopsy. Using a 443 patient training set, protein biomarker discovery was performed using the highly multiplexed SOMAscan® proteomic assay, a set of 19 clinical variables, and the steatosis predisposing PNPLA3 rs738409 single nucleotide polymorphism genotype status. The most stable markers were selected using a stability selection algorithm with a L_1-regularized logistic regression kernel and were then fitted with logistic regression models to classify steatosis, that were then tested against a 133 sample blinded verification set. The highest area under the ROC curve (AUC) for steatosis of PNPLA3 rs738409 genotype, 8 proteins, or 19 phenotypic variables was 0.913, whereas the final classifier that included variables from all three domains had an AUC of 0.935. These data indicate that multi-domain modeling has better predictive power than comprehensive analysis of variables from a single domain.

Obesity is associated with fat accumulation in the liver, which is commonly diagnosed as non-alcoholic fatty liver disease (NAFLD). NAFLD encompasses a wide range of conditions that are thought to arise from fatty liver (hepatic steatosis) to nonalcoholic steatohepatitis (NASH), which refers to findings on liver biopsy reflecting steatohepatitis (fat related inflammation) with or without fibrosis in the absence of significant alcohol consumption[1,2]. NAFLD has become the major cause of chronic liver disease due to the progression of simple steatosis to hepatocyte injury, liver inflammation, fibrosis, and cirrhosis that worsen clinical outcomes[3,4] and are associated with increased liver-related morbidity and mortality[5]. The prevalence of NAFLD is increasing in tandem with the rising rates of obesity, and is expected to double in the U.S. by 2030[6] with upwards of 100 million people in the U.S. at risk[7]. Despite the public health significance of NAFLD, most affected individuals remain undiagnosed[8].

We[9] and others[10,11] have identified several clinical factors, including lipid and glucose levels, as well as parameters of iron metabolism, that are associated with the development of steatosis. However, neither clinical characteristics nor laboratory values have yet proved useful for predicting disease development or course[12–14]. Liver biopsy is the primary method used to accurately assess NAFLD stage[15–17], although histological evaluation has diagnostic limitations[18]. Percutaneous liver biopsy is an invasive and expensive procedure associated with major complications, including mortality[19]. While liver biopsy can be performed in the context of abdominal surgery,

[1]Geisinger Obesity Research Institute, Danville, PA, USA. [2]Lewis Katz School of Medicine at Temple University, Philadelphia, PA, USA. [3]National Jewish Health, Denver, CO, USA. Correspondence and requests for materials should be addressed to G.S.G. (email: gsgerhard@Temple.edu)

particularly in patients with extreme obesity who are high risk for NAFLD undergoing bariatric procedures[18], this population represents only a small fraction of patients at risk.

Given the substantial public health burden of NAFLD, robust methods are needed to identify those who have fatty liver who are at risk for inflammation and fibrosis. Because the common liver function tests, alanine aminotransferase (ALT) and aspartate aminotransferase (AST) lack sufficient sensitivity or specificity for NAFLD[20–24] other non-invasive biomarkers have been studied, including those associated with cell death, inflammation, and oxidative stress, as well as algorithms of multi-component panels[25]. Thus far, no biomarkers have proven to be clinically acceptable for diagnosis, prognosis, or risk stratification[26]. Imaging techniques have also been used to assess NAFLD, including liver ultrasound, a relatively inexpensive modality with relatively high sensitivity for moderate to severe steatosis, magnetic resonance imaging (MRI) and magnetic resonance spectroscopy, both relatively expensive with limited availability, as well as computed tomography (CT) and transient ultrasound elastography[27]. However, none of the current imaging modalities are logistically or economically viable to broadly apply to the population at risk for NAFLD except as an adjunct to biopsy[28].

We used a multivariate approach integrating "omics" derived variables from three data domains; genomic, phenomic, and proteomic. We selected the NAFLD susceptibility single nucleotide polymorphism (SNP) rs738409 in the patatin-like phospholipase domain containing 3 gene (PNPLA3)[29] as the genomic variable, well characterized across diverse populations[30–32]. Phenomic variables were identified initially in a comprehensive analysis over 200 clinical variables[9] and a set of serum proteins were identified using a novel unbiased high content multiplexed proteomic screen[33] based on the SOMAmer® (Slow Off-rate Modified Aptamer) technology. This approach uses single stranded DNA-based protein affinity reagents[21] that incorporate chemically modified nucleotides as structural mimetics of amino acid side chains increasing diversity, affinity, and specificity for native proteins[22]. A multicomponent panel for steatosis using data from all three domains resulted in an AUC for the receiver operating characteristic (ROC) curve of 0.935. Integration of variables from a diverse set of data sources provides a powerful approach for the development of non-invasive biomarker algorithms for NAFLD.

Materials and Methods

Study Participants. Blood for DNA isolation and fasting serum samples were collected within three months prior to surgery during pre-operative clinic visits from 577 patients who had been consented as part of a research program on NAFLD and obesity in the Geisinger Clinic Center for Nutrition and Weight Management Bariatric Surgery Program. Study participants were randomly divided into a discovery (n = 443) or validation (n = 134) cohort. Intra-operative wedge biopsies of the liver and clinical data were obtained as previously described[23]. Patients with any evidence of hepatitis B virus (HBV), hepatitis C virus (HCV), or human immunodeficiency virus (HIV) infection, or alcohol abuse were included in this study as previously described[9]. A required comprehensive behavioral evaluation performed by a clinical psychologist included inquiries about current and past substance and alcohol use. If a patient used alcohol, they were further evaluated using established criteria for alcohol use disorders[24]. Patients whose clinical criteria for alcohol use disorders were denied access to bariatric surgery. In addition, patients with diagnosis codes ICD9 303 or ICD9 305.0 indicating a clinical diagnosis of alcohol abuse were also excluded. Source data included patient demographics, clinical measures, ICD9 codes, medical history, medication codes, and lab results. The research protocol was approved by the Geisinger Clinic Institutional Review Board, all participants provided written informed consent, and all experiments were performed in accordance with relevant guidelines and regulations.

SOMAscan Assay. Serum samples were analyzed using the SOMAscan assay (SomaLogic; Boulder, CO), which is a sensitive, and quantitative protein biomarker discovery platform. SOMAmers (Slow Off-rate Modified Aptamers), single-stranded DNA aptamers with modified nucleotides, bind to specific proteins in the serum that are then be quantified as DNA. The SOMAscan assay quantified a total of 1129 proteins in each sample. In our analysis, the median lower limit of quantitation for all measured proteins was 0.3 picomolar (pM), with a dynamic range of >5 logs, and a median coefficient of variation (%CV) of 5%[34].

PNPLA3 genotyping. We extracted DNA from blood samples using standard methods[35] and genotype marker rs738409 in the *PNPLA3* gene as described[36].

Phenomic modeling. Previously, we conducted univariate logistic regression to determine which of more than 200 pre-operative clinical variables were independently associated with the presence of steatosis in ~2300 individuals[9]. We identified 19 candidate variables associated with the presence of liver fat (Supplemental Table 1) and used logistic regression within the discovery cohort to identify the minimal subset of these variables that maintained the area under the curve (c-statistic) of the full model for steatosis. A backwards stepwise process was used for model variable selection. The initial model included all 19 variables. Subsequent models were evaluated by sequentially removing one variable at a time and assessing the resulting change in the c-statistic. When the removal of a variable resulted in a decrease in the c-statistic by <0.01, the variable was excluded from the final model. The variables retained for the final model were combined into a single classifier score by calculating (Supplemental Table S4) the predicted probability of steatosis. This score was applied to the validation cohort and brought forward for multi-component modeling.

Derivation of Proteomic Panel. Candidate markers were selected using a stability selection algorithm with an L1-regularized logistic regression kernel. Stability selection takes many subsets of half the data and performs biomarker selection using the lasso classifier, which is a regularized logistic regression model[37]. The selection path for a single biomarker is the proportion of these subsets for which that biomarker was selected by the lasso model over a range of lambda, a tuning parameter that determines how many biomarkers are selected by the

Variable	Measure	Discovery group N = 443	Validation group N = 134	p-value
Age, years	Mean (SD)	46.2 (10.7)	46.3 (11.1)	0.943[1]
Sex	Female, % (n)	82% (n = 363)	84% (n = 112)	0.663[2]
	Male, % (n)	18% (n = 80)	16% (n = 22)	
Race	White, % (n)	99% (n = 439)	99% (n = 133)	0.999[3]
	Black, % (n)	< 1% (n = 2)	1% (n = 1)	
	Other, % (n)	<1% (n = 2)	0% (n = 0)	
BMI, kg/m^2	Mean (SD)	49.2 (9.0)	49.2 (8.4)	0.979[1]
Diabetes	Yes, % (n)	41% (n = 180)	41% (n = 55)	0.932[2]
Hypertension	Yes, % (n)	47% (n = 209)	44% (n = 59)	0.522[2]
Dyslipidemia	Yes, % (n)	37% (n = 163)	43% (n = 57)	0.230[2]
ALT, U/L	Median [IQR]	27 [20, 39]	26 [19, 38]	0.498[4]
AST, U/L	Median [IQR]	24 [19, 33]	24 [20, 30]	0.647[4]
Cholesterol, md/dL	Mean (SD)	187.7 (40.3)	188.2 (39.5)	0.897[1]
HDL, md/dL	Mean (SD)	47.2 (11.5)	46.2 (10.8)	0.351[1]
LDL, md/dL	Mean (SD)	105.8 (33.6)	107.4 (36.1)	0.630[1]
Triglycerides, md/dL	Median [IQR]	152 [104, 208]	180.6 (118.9)	0.911[4]
Platelet count, K/uL	Mean (SD)	285.2 (72.3)	294.7 (64.2)	0.175[1]
Steatosis	<5%, % (n)	30% (n = 131)	32% (n = 43)	0.612[2]
	5–33%, % (n)	20% (n = 89)	24% (n = 32)	
	33–66%, % (n)	26% (n = 117)	24% (n = 32)	
	>66%, % (n)	24% (n = 106)	20% (n = 27)	
Lobular inflammation	No foci, % (n)	55% (n = 242)	62% (n = 83)	0.157[2]
	<2 foci*, % (n)	37% (n = 162)	34% (n = 45)	
	2–4 foci*, % (n)	9% (n = 39)	4% (n = 6)	
	>4 foci*, % (n)	0% (n = 0)	0% (n = 0)	
Fibrosis stage	None, % (n)	59% (n = 262)	64% (n = 86)	0.069[3]
	1, % (n)	25% (n = 111)	28% (n = 37)	
	2, % (n)	9% (n = 39)	7% (n = 10)	
	3, % (n)	5% (n = 20)	1% (n = 1)	
	4, % (n)	2% (n = 11)	0% (n = 0)	
PNPLA3**	CC, % (n)	54% (n = 219)	57% (n = 71)	0.407[2]
	CG, % (n)	40% (n = 163)	35% (n = 43)	
	GG, % (n)	6% (n = 23)	8% (n = 10)	

Table 1. Characteristics of discovery and validation sets. Reference ranges: ALT (Male 5–52 U/L, Female 10–60 U/L), AST (Male 13–39 U/L, Female 10–42 U/L), Cholesterol (<200 mg/dL), HDL (> = 40 mg/dL), LDL (<130 mg/dL), Triglycerides (<150 mg/dL), Platelet Count (140–400 K/uL). [1]Two-sample t-test; [2]Chi-square test; [3]Fisher's Exact Test; [4]Wilcoxon Rank-Sum test. SD = standard deviation, IQR = Interquartile Range. *per 200X field. **PNPLA3 unknown for 48 patients (38 in discovery group and 10 in the validation group). Hardy-Weinberg test for equilibrium: p = 0.304 in discovery group and p = 0.344 in validation group.

lasso. The maximum selection probability over a range of lambda values was the ultimate metric used to select a set of biomarkers.

Steatosis classifier models (steatosis vs. all other groups) were developed by inputting the most stable markers into the logistic regression classification algorithm. Once the models were fixed, bootstrap performance was done as verification of the developed models: the discovery set for steatosis as well as fibrosis were split randomly into 80% training and 20% test set to verify by bootstrapping. This was repeated 2500 times with a different subset. Sensitivity and specificity and confidence interval for each comparison was noted. The model types were evaluated using 10-fold cross-validation and inspecting plots of log-likelihood ratios and receiver operating characteristic (ROC) curves. The proteins retained for the final model were combined into a single classifier score by calculating the predicted probability of steatosis. This score was applied to the validation cohort and was brought forward for multi-component modeling.

Multi-component modeling. The classifier scores from each domain were combined into logistic regression models. Performance of the combined classifiers was evaluated using area under the receiver operating characteristic (ROC) curve based on a c-statistic and 95% bootstrap confidence intervals.

Results

Characteristics of the discovery and validation cohorts. The study population of 576 adult patients was randomly assigned to either a discovery (N = 443) or validation (N = 134) cohort. As shown in Table 1,

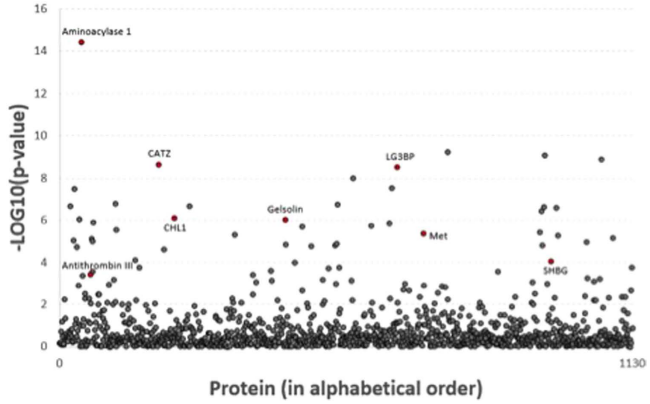

Figure 1. Plot of unadjusted p-values for association between each protein expression and presence of any steatosis. The labeled proteins were those selected for inclusion in final model.

there were no significant differences in age, sex, body mass index (BMI), type 2 diabetes (T2D) status, measures of cholesterol metabolism, ALT and AST levels, platelet count, or *PNPLA3* genotype distribution between the two groups. Approximately 30% of the patients had normal liver histology (<5% steatosis), while 20–24% were classified as mildly or severely steatotic (5–33% and >66% liver fat, respectively) and ~25% showed moderate steatosis (33–66% fat). In both the discovery and validation cohorts, most of the key variables were different between patients with steatosis compared to those with normal liver histology (Supplemental Tables S1 and S2). For example, the percentage of patients with T2D increased from 22% in individuals with normal histology to 44% in those with steatosis, consistent with earlier findings reported by us[9] and others[37].

Steatosis genomic classifier. For the genomic model, the *PNPLA3* rs738409 genotype was used as the sole variable, given the strength of its association with steatosis compared to other reported genetic variants[39–41]. Marker genotype was designated as homozygous wildtype, heterozygous, or homozygous NAFLD risk allele. As expected, the distribution of genotypes was significantly different between patients with steatosis compared to those with normal liver histology (Supplemental Tables S1 and S2).

Steatosis phenomic classifier. We previously analyzed a cohort of 2929 subjects[9] from which the 576 individuals used for the current analyses were drawn. Of the 19 variables previously identified (Supplemental Table S3), 12 were included in the final steatosis phenomic classifier including glucose, serum insulin, triglycerides, HDL, ALT, ferritin, creatinine, chloride, zinc, use of metformin, use of estrogen/progestin, and a clinical diagnosis of sleep apnea. These were combined into a single classifier score and used for multi-component modeling.

Steatosis proteomic classifier. Univariate analysis of serum levels of the SOMAmer platform of 1129 proteins using discovery and validation sets identified 30 proteins that met the Bonferroni-corrected level of statistical significance (Fig. 1). In multivariate analysis, eight proteins were associated with steatosis (Table 2). Serum levels of three of these proteins were associated with increased steatosis, while levels of five showed an inverse relationship with steatosis grade.

Steatosis multi-component classifier. The results from modeling within each of the three individual data domains were then used to create a multi-component classifier that included *PNPLA3* genotype, steatosis clinical prediction score, and the proteomic classifier. We combined independently associated variables from each of the three data domains to generate a single logistic regression model and assessed performance of the combined classifier using area under (AUC) the receiver operating characteristic (ROC) curve based on a c-statistic (Table 3, Fig. 2). The proteomic classifier by itself achieved the highest AUC of the three individual data domains with an AUC of 0.913 versus an AUC for the *PNPLA3* genomic domain of only 0.596 and an AUC for the phenomic domain of 0.886. Combining genomic and phenomic domains yielded little effect on the AUC (0.892), while combining the phenomic and proteomic domains resulted in an AUC of 0.932. The highest AUC was achieved with the combination of all three domains, yielding a value of 0.935. A similar analysis was conducted using the validation cohort (Table 3, Fig. 3). Although the AUC values were lower across all validation models, the AUC values improved with the inclusion of each addition domain.

Discussion

The central hallmark of NAFLD is the presence of increased fat in the liver, a condition that has several potentially important pathophysiological implications. For example, steatosis has been associated with the metabolic syndrome and insulin resistance, although the cause-effect relationship of this association is not clear[42–44]. Patients with steatosis are also at greater risk for the development of steatohepatitis and hepatic fibrosis, including cirrhosis[45]. The cause of NAFLD appears to be multifactorial[46], although lifestyle (i.e., over-nutrition) plays a significant role by contributing to the development of obesity[47]. Both genetic variants[28,39,48] and protein biomarkers[49] have also been associated with NAFLD. Due to the underlying complexity of NAFLD pathogenesis, we

Gene	Protein	Odds Ratio	[95% CI]	p-value
ACY1	Aminoacylase-1	57.89	[13.69, 244.90]	<0.0001
SHBG	Sex hormone-binding globulin	0.56	[0.42, 0.75]	<0.0001
CTSZ	Cathepsin Z	0.69	[0.48, 0.98]	0.0400
MET	Hepatocyte growth factor receptor	0.60	[0.43, 0.83]	0.0020
GSN	Gelsolin/GSN	2.69	[1.74, 4.16]	<0.0001
LGALS3BP	Galectin-3 binding protein	0.59	[0.43, 0.79]	0.0005
CHL1	Neural cell adhesion molecule L1-like protein	2.20	[1.42, 3.42]	0.0004
SERPINC1	Antithrombin III	0.68	[0.49, 0.94]	0.0185

Table 2. Logistic regression model for NAFLD using selected protein biomarkers. The biomarkers were rescaled to the standard normal (mean = 0, SD = 1) before inclusion in the logistic regression model. Odds ratios can be interpreted as the odds of steatosis for each 1 standard deviation increase in the protein expression level.

	Discovery		Validation	
Model	AUC	95%CI	AUC	95%CI
1. GENOMIC only	0.596	[0.547, 0.645]	0.610	[0.519, 0.713]
2. PHENOMIC only	0.886	[0.851, 0.918]	0.778	[0.693, 0.851]
3. PHENO + GENO	0.892	[0.862, 0.924]	0.782	[0.710, 0.865]
4. PROTEOMIC ONLY	0.913	[0.882, 0.937]	0.864	[0.793, 0.927]
5. PROTEO + GENO	0.920	[0.892, 0.946]	0.889	[0.832, 0.945]
6. PROTEO + PHENO	0.932	[0.904, 0.955]	0.892	[0.840, 0.943]
7. 3 DOMAIN MODEL	0.935	[0.913, 0.959]	0.914	[0.871, 0.957]

Table 3. Area under (AUC) the receiver operating characteristic (ROC) curve based on a c-statistic.

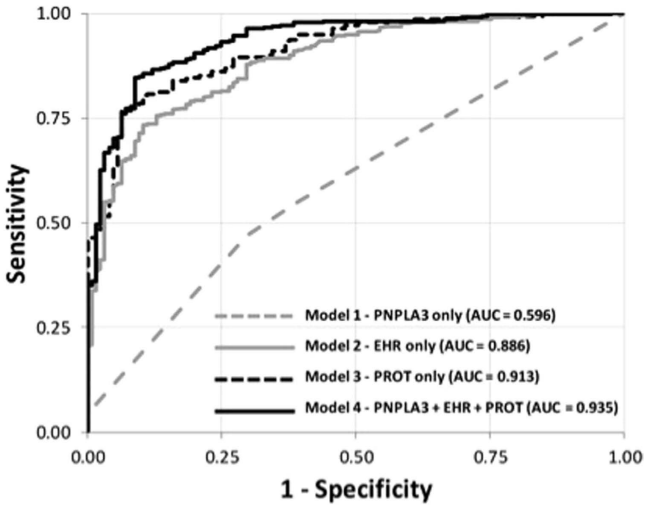

Figure 2. ROC curves for Discovery Model*. *Note that PNPLA3 was included as the number of alleles (0, 1, 2) and was treated as an ordinal variable. Those with unknown PNPLA3 status were included in the model by using a common missing data strategy (i.e. treating them as a separate subgroup).

sought to develop an algorithm based on the unbiased assessment of large groups of variables, i.e., using "omics" approaches, in several data domains that could differentiate NAFLD in a population with extreme obesity. A similar type of "omics" approach combining transcriptomic, ELISA-based serum proteomic, and nuclear magnetic resonance-based metabolomic analyses of liver biopsy tissue and serum samples obtained from patients with high versus low grade steatosis has recently been reported[50]. However, the number of individuals assessed was quite small (N = 20), thereby limiting the conclusions that can be drawn from that analysis.

We selected the rs738409 variant in *PNPLA3* for modeling based on existing genomic data that had been generated in populations not selected for obesity[39–41], as well as our own data based on a similar population with extreme obesity[39]. By itself, rs738409 showed the poorest discriminatory value, a finding that was not surprising given the relatively small effect size found in the initial studies[51,52]. Adding rs738409 genotype to the phenomic

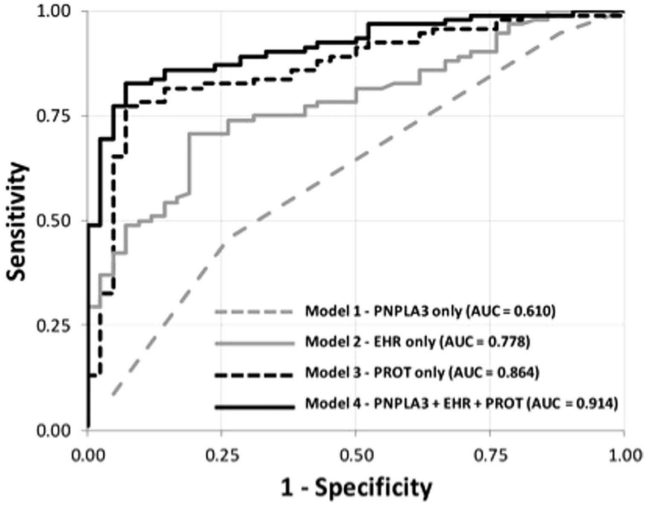

Figure 3. ROC curves for Validation model.

classifier essentially yielded no effect, whereas combining it with the proteomic classifier had a small but positive effect (0.892 vs. 0.913). This suggests that genomic classifier may be already represented by one or more variables present in the phenomic classifier. This is somewhat surprising because the phenomic variables are largely represented by those related to metabolic abnormalities, while rs738409 genotype has been associated with NAFLD independent of metabolic disease[48,53,54], although some studies have found an interaction with glucose metabolism[55]. Its complementary relationship with the group of serum proteins in the proteomic classifier implies that there are multiple mechanistic pathways involved in the development of steatosis.

We used the SOMAscan platform[21,32,56] as a discovery assay that has been applied successfully to diagnostic biomarker discovery and validation for other disorders. This platform has been used for the identification of biomarkers in rheumatoid arthritis[57], Alzheimer's disease[58], and infectious disease[59]. The SOMAscan aptamer-based assay has been designed for high-throughput multiplexing allowing for the measurement of over 1000 proteins in only 65 μL of serum. However, because the aptamer-based methodology only detects available protein epitopes, in instances where epitopes may be blocked by other proteins or post-translational modifications, measured levels may not represent actual protein concentrations. Further, protein markers associated with steatosis were identified using a stability selection algorithm with an L1-regularized logistic regression kernel; therefore, the most stable group of markers may not represent the most highly associated individual markers. Nevertheless, individual markers identified here may shed light on the underlying biology of steatosis. For example, several of the serum proteins identified in the proteomic screen have previously been associated with NAFLD or a related aspect of hepatic lipid metabolism. Aminoacylase 1, a zinc-binding protein that catalyzes the hydrolysis of N-acetyl amino acids into free aliphatic amino acids and acetic acid[60], was increased in hepatic lipid droplets of mice subjected to caloric restriction[61], suggesting a role in the metabolic adjustments to the overfed state in NAFLD. However, we found increased steatosis associated with aminoacylase 1 levels. This could reflect a rapid temporal response in aminoacylase 1 levels since patients were fasting at the time of blood draw. In addition, sex hormone binding globulin (SHBG), a glycoprotein that is produced primarily by hepatocytes and serves to transport sex steroid hormones through the blood to target tissues, was first associated with hepatic steatosis through studies of monosaccharide-induced hepatic lipogenesis in animals, a treatment that suppressed expression of sex hormone–binding globulin[62]. A number of human population-based studies have found that SHBG levels are inversely associated with NAFLD[63–68], consistent with our results. MET (hepatocyte growth factor)/mesenchymal-epithelial transition factor) functions in anti-apoptosis pathway signaling in hepatocytes, in part by sequestering Fas to inhibit Fas-mediated apoptosis. This relationship appears to be lost in NAFLD[69].

Data linking the other protein markers to NAFLD is less clear. Galectin-3 binding protein (LGALS3BP) has been used as a component of a multi-protein panel for the prediction of fibrosis in Hepatitis C (HCV) infection[70], and serves as a biomarker of hepatocellular carcinoma resulting from HCV cirrhosis[71]. Antithrombin III levels did not correlate with liver NAFLD histology in obese patients[72], although markedly decreased levels have been found in acute fatty liver of pregnancy[73]. Gelsolin is a protein generated by the liver, and appears to be expressed by hepatic sinusoidal endothelial cells, hepatic stellate cells, myofibroblasts, and mononuclear cells, but not hepatocytes[74]. Gelsolin has been implicated in the apoptosis of hepatic stellate cells, which play a major role in the progression of steatosis to fibrosis and cirrhosis[75,76]. Deficiency of cell adhesion molecule L1 like does not appear to affect hepatic metabolism. Chl1 knockout mice did not show any significant abnormal phenotype up to an age of 2 years[77]. Little biological information is available on cathepsin Z, a cysteine proteinase of the papain family, except that it is widely expressed in human tissues[78].

In addition to non-invasive blood-based classifiers, various imaging approaches have been used to characterize NAFLD. Ultrasound is commonly used to assess for the presence and amount of liver fat, though it can be dependent upon the skill of the operator and the particular capacity of the instrument[79]. Due to the subjective nature of the technique, reproducibility can therefore be low and the ability to detect mild levels of steatosis is not

as robust as for moderate to severe steatosis. However, the technique is relatively inexpensive, simple to perform, and safe. Alternatives to traditional ultrasound, such as controlled attenuation parameter[80], are being developed and validated, thus may significantly improve on the disadvantages. Other imaging modalities include computed tomography (CT) and magnetic resonance imaging (MRI). Multi-parametric quantitative MRI is also a promising technique that may be able to closely correlate with liver histology. However, all imaging-based approaches are limited by cost and availability vis-à-vis blood-based testing and will likely not be scalable to the population at risk for NAFLD.

We used variables from three different data domains that were derived from "omics" analyses to develop a multi-component classifier for NAFLD. Despite the robust nature of the classifier, there is still a need to improve the AUC. The use of even larger samples sizes would be useful, although high-throughput analyses are more costly and complex than smaller scale analyses. We also have used a population with extreme levels of obesity in part because of the availability of gold standard liver biopsy pathology data. Developing classifiers for lower levels of obesity may be limited by the difficulties in obtaining such data. We also used a relatively ethnically, racially, and geographically homogenous population, appropriate for discovery and initial development studies. Extending these results to other populations and/or developing classifiers specific to other populations are needed for future studies. Despite these potential limitations, our results suggest that a high-throughput, multi-domain, multi-component approach may be a promising avenue for further investigation.

References

1. Neuschwander-Tetri, B. A. & Caldwell, S. H. Nonalcoholic steatohepatitis: summary of an AASLD Single Topic Conference. *Hepatology* **37,** 1202–1219, doi: 10.1053/jhep.2003.50193 (2003).
2. Chalasani, N. *et al.* The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. *Hepatology* **55,** 2005–2023, doi: 10.1002/hep.25762 (2012).
3. Matteoni, C. A. *et al.* Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. *Gastroenterology* **116,** 1413–1419 (1999).
4. Rafiq, N. *et al.* Long-term follow-up of patients with nonalcoholic fatty liver. *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association* **7,** 234–238, doi: 10.1016/j.cgh.2008.11.005 (2009).
5. McCullough, A. J. The clinical features, diagnosis and natural history of nonalcoholic fatty liver disease. *Clinics in liver disease* **8,** 521–533, viii, doi: 10.1016/j.cld.2004.04.004 (2004).
6. Younossi, Z. M. *et al.* Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association* **9,** 524–530 e521, quiz e560, doi: 10.1016/j.cgh.2011.03.020 (2011).
7. Rinella, M. E. Nonalcoholic fatty liver disease: a systematic review. *Jama* **313,** 2263–2273, doi: 10.1001/jama.2015.5370 (2015).
8. Patel, V., Sanyal, A. J. & Sterling, R. Clinical Presentation and Patient Evaluation in Nonalcoholic Fatty Liver Disease. *Clin Liver Dis* **20,** 277–292, doi: 10.1016/j.cld.2015.10.006 (2016).
9. Gerhard, G. S. *et al.* Identification of novel clinical factors associated with hepatic fat accumulation in extreme obesity. *J Obes* **2014,** 368210, doi: 10.1155/2014/368210 (2014).
10. Wu, K. T. *et al.* Nonalcoholic fatty liver disease severity is associated with the ratios of total cholesterol and triglycerides to high-density lipoprotein cholesterol. *J Clin Lipidol* **10,** 420–425 e421, doi: 10.1016/j.jacl.2015.12.026 (2016).
11. Long, M. T. *et al.* Development and Validation of the Framingham Steatosis Index to Identify Persons with Hepatic Steatosis. *Clin Gastroenterol Hepatol*, doi: 10.1016/j.cgh.2016.03.034 (2016).
12. Bedossa, P. & Patel, K. Biopsy and Noninvasive Methods to Assess Progression of Nonalcoholic Fatty Liver Disease. *Gastroenterology*, doi: 10.1053/j.gastro.2016.03.008 (2016).
13. Alkhouri, N. & Feldstein, A. E. Noninvasive diagnosis of nonalcoholic fatty liver disease: Are we there yet? *Metabolism*, doi: 10.1016/j.metabol.2016.01.013 (2016).
14. Spengler, E. K. & Loomba, R. Recommendations for Diagnosis, Referral for Liver Biopsy, and Treatment of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. *Mayo Clin Proc* **90,** 1233–1246, doi: 10.1016/j.mayocp.2015.06.013 (2015).
15. Kleiner, D. E. *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41,** 1313–1321, doi: 10.1002/hep.20701 (2005).
16. Pagadala, M. R. & McCullough, A. J. The relevance of liver histology to predicting clinically meaningful outcomes in nonalcoholic steatohepatitis. *Clinics in liver disease* **16,** 487–504, doi: 10.1016/j.cld.2012.05.006 (2012).
17. Younossi, Z. M. *et al.* Pathologic criteria for nonalcoholic steatohepatitis: interprotocol agreement and ability to predict liver-related mortality. *Hepatology* **53,** 1874–1882, doi: 10.1002/hep.24268 (2011).
18. Bedossa, P. & Patel, K. Biopsy and Non-invasive Methods to Assess Progression of Nonalcoholic Fatty Liver Disease. *Gastroenterology*, doi: 10.1053/j.gastro.2016.03.008 (2016).
19. West, J. & Card, T. R. Reduced mortality rates following elective percutaneous liver biopsies. *Gastroenterology* **139,** 1230–1237, doi: 10.1053/j.gastro.2010.06.015 (2010).
20. Petrick, A., Benotti, P., Wood, G. C., Still, C. D., Strodel, W. E., Gabrielsen, J., Rolston, D., Chu, X., Argyropoulos, G., Ibele, A. & Gerhard, G. S. Utility of Ultrasound, Transaminases, and Visual Inspection to Assess Nonalcoholic Fatty Liver Disease in Bariatric Surgery Patients. *Obes Surg.* Dec; **25(12),** 2368–75, doi: 10.1007/s11695-015-1707-6 (2015).
21. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249,** 505–510 (1990).
22. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5,** e15004, doi: 10.1371/journal.pone.0015004 (2010).
23. Wood, G. C. *et al.* An electronic health record-enabled obesity database. *BMC medical informatics and decision making* **12,** 45, doi: 10.1186/1472-6947-12-45 (2012).
24. *Diagnostic and Statistical Manual of Mental Disorders, Text Revision* (*DSM-IV-TR*). 4th edn (2000).
25. Fitzpatrick, E. & Dhawan, A. Noninvasive biomarkers in non-alcoholic fatty liver disease: current status and a glimpse of the future. *World journal of gastroenterology: WJG* **20,** 10851–10863, doi: 10.3748/wjg.v20.i31.10851 (2014).
26. Miyake, T. *et al.* Non-alcoholic fatty liver disease: Factors associated with its presence and onset. *J Gastroenterol Hepatol* **28** Suppl 4, 71–78, doi: 10.1111/jgh.12251 (2013).
27. Kaswala, D. H., Lai, M. & Afdhal, N. H. Fibrosis Assessment in Nonalcoholic Fatty Liver Disease (NAFLD) in 2016. *Dig Dis Sci*, doi: 10.1007/s10620-016-4079-4 (2016).
28. Hannah, W. N. Jr. & Harrison, S. A. Noninvasive imaging methods to determine severity of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Hepatology* **64,** 2234–2243, doi: 10.1002/hep.28699 (2016).
29. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40,** 1461–1465, doi: 10.1038/ng.257 (2008).

30. Akuta, N. *et al.* Relationships between Genetic Variations of PNPLA3, TM6SF2 and Histological Features of Nonalcoholic Fatty Liver Disease in Japan. *Gut Liver* **10,** 437–445, doi: 10.5009/gnl15163 (2016).
31. Zhang, L. *et al.* PNPLA3 polymorphisms (rs738409) and non-alcoholic fatty liver disease risk and related phenotypes: a meta-analysis. *J Gastroenterol Hepatol* **30,** 821–829, doi: 10.1111/jgh.12889 (2015).
32. Xu, R., Tao, A., Zhang, S., Deng, Y. & Chen, G. Association between patatin-like phospholipase domain containing 3 gene (PNPLA3) polymorphisms and nonalcoholic fatty liver disease: a HuGE review and meta-analysis. *Sci Rep* **5,** 9284, doi: 10.1038/srep09284 (2015).
33. Gold, L., Walker, J. J., Wilcox, S. K. & Williams, S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *N Biotechnol* **29,** 543–549, doi: 10.1016/j.nbt.2011.11.016 (2012).
34. Ostroff, R. *et al.* The stability of the circulating human proteome to variations in sample collection and handling procedures measured with an aptamer-based proteomics array. *Journal of proteomics* **73,** 649–666, doi: 10.1016/j.jprot.2009.09.004 (2010).
35. Chu, X. *et al.* Association of morbid obesity with FTO and INSIG2 allelic variants. *Archives of surgery* **143,** 235–240, discussion 241, doi: 10.1001/archsurg.2007.77 (2008).
36. Gorden, A. *et al.* Genetic variation at NCAN locus is associated with inflammation and fibrosis in non-alcoholic fatty liver disease in morbid obesity. *Human heredity* **75,** 34–43, doi: 10.1159/000346195 (2013).
37. Meinshausen, N. & Bhlmann, P. Stability selection. *Journal of the Royal Statistical Society, Series B* **72,** 417–473 (2012).
38. Cazzo, E., de Felice Gallo, F., Pareja, J. C. & Chaim, E. A. Nonalcoholic fatty liver disease in morbidly obese subjects: correlation among histopathologic findings, biochemical features, and ultrasound evaluation. *Obes Surg* **24,** 666–668, doi: 10.1007/s11695-014-1183-4 (2014).
39. DiStefano, J. K. *et al.* Genome-wide analysis of hepatic lipid content in extreme obesity. *Acta diabetologica* **52,** 373–382, doi: 10.1007/s00592-014-0654-3 (2015).
40. Hernaez, R. *et al.* Association between variants in or near PNPLA3, GCKR, and PPP1R3B with ultrasound-defined steatosis based on data from the third National Health and Nutrition Examination Survey. *Clin Gastroenterol Hepatol* **11,** 1183–1190 e1182, doi: 10.1016/j.cgh.2013.02.011 (2013).
41. Speliotes, E. K. *et al.* PNPLA3 variants specifically confer increased risk for histologic nonalcoholic fatty liver disease but not metabolic disease. *Hepatology* **52,** 904–912, doi: 10.1002/hep.23768 (2010).
42. Finck, B. N. & Hall, A. M. Does Diacylglycerol Accumulation in Fatty Liver Disease Cause Hepatic Insulin Resistance? *Biomed Res Int* **2015,** 104132, doi: 10.1155/2015/104132 (2015).
43. Wainwright, P. & Byrne, C. D. Bidirectional Relationships and Disconnects between NAFLD and Features of the Metabolic Syndrome. *Int J Mol Sci* **17,** doi: 10.3390/ijms17030367 (2016).
44. Yki-Jarvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol* **2,** 901–910, doi: 10.1016/S2213-8587(14)70032-4 (2014).
45. Marengo, A., Jouness, R. I. & Bugianesi, E. Progression and Natural History of Nonalcoholic Fatty Liver Disease in Adults. *Clin Liver Dis* **20,** 313–324, doi: 10.1016/j.cld.2015.10.010 (2016).
46. Hardy, T., Oakley, F., Anstee, Q. M. & Day, C. P. Nonalcoholic Fatty Liver Disease: Pathogenesis and Disease Spectrum. *Annu Rev Pathol*, doi: 10.1146/annurev-pathol-012615-044224 (2016).
47. Woo Baidal, J. A. & Lavine, J. E. The intersection of nonalcoholic fatty liver disease and obesity. *Sci Transl Med* **8,** 323rv321, doi: 10.1126/scitranslmed.aad8390 (2016).
48. Speliotes, E. K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* **7,** e1001324, doi: 10.1371/journal.pgen.1001324 (2011).
49. Ladaru, A., Balanescu, P., Stan, M., Codreanu, I. & Anca, I. A. Candidate proteomic biomarkers for non-alcoholic fatty liver disease (steatosis and non-alcoholic steatohepatitis) discovered with mass-spectrometry: a systematic review. *Biomarkers* **21,** 102–114, doi: 10.3109/1354750X.2015.1118542 (2016).
50. Wruck, W., Kashofer, K., Rehman, S., Daskalaki, A., Berg, D., Gralka, E., Jozefczuk, J., Drews, K., Pandey, V., Regenbrecht, C., Wierling, C., Turano, P., Korf, U., Zatloukal, K., Lehrach, H., Westerhoff, H. V. & Adjaye, J. Multi-omic profiles of human non-alcoholic fatty liver disease tissue highlight heterogenic phenotypes. *Sci Data.* **8**, **2**, 150068, doi: 10.1038/sdata.2015.68 (2015).
51. Kotronen, A. *et al.* A common variant in PNPLA3, which encodes adiponutrin, is associated with liver fat content in humans. *Diabetologia* **52,** 1056–1060, doi: 10.1007/s00125-009-1285-z (2009).
52. Kotronen, A. *et al.* Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. *Gastroenterology* **137,** 865–872, doi: 10.1053/j.gastro.2009.06.005 (2009).
53. Park, J. H. *et al.* I148M variant in PNPLA3 reduces central adiposity and metabolic disease risks while increasing nonalcoholic fatty liver disease. *Liver Int* **35,** 2537–2546, doi: 10.1111/liv.12909 (2015).
54. Verrijken, A. *et al.* A gene variant of PNPLA3, but not of APOC3, is associated with histological parameters of NAFLD in an obese population. *Obesity* (*Silver Spring*) **21,** 2138–2145, doi: 10.1002/oby.20366 (2013).
55. Guichelaar, M. M. *et al.* Interactions of allelic variance of PNPLA3 with nongenetic factors in predicting nonalcoholic steatohepatitis and nonhepatic complications of severe obesity. *Obesity* (*Silver Spring*) **21,** 1935–1941, doi: 10.1002/oby.20327 (2013).
56. Kraemer, S. *et al.* From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PLoS One* **6,** e26332, doi: 10.1371/journal.pone.0026332 (2011).
57. McArdle, A. *et al.* Developing Clinically Relevant Biomarkers in Inflammatory Arthritis: A Multi-Platform Approach for Serum Candidate Protein Discovery. *Proteomics Clin Appl*, doi: 10.1002/prca.201500046 (2015).
58. Voyle, N. *et al.* Blood Protein Markers of Neocortical Amyloid-beta Burden: A Candidate Study Using SOMAscan Technology. *J Alzheimers Dis* **46,** 947–961, doi: 10.3233/JAD-150020 (2015).
59. Marion, T. *et al.* Respiratory Mucosal Proteome Quantification in Human Influenza Infections. *PLoS One* **11,** e0153674, doi: 10.1371/journal.pone.0153674 (2016).
60. Sass, J. O. *et al.* Mutations in ACY1, the gene encoding aminoacylase 1, cause a novel inborn error of metabolism. *Am J Hum Genet* **78,** 401–409, doi: 10.1086/500563 (2006).
61. Baumeier, C. *et al.* Caloric restriction and intermittent fasting alter hepatic lipid droplet proteome and diacylglycerol species and prevent diabetes in NZO mice. *Biochim Biophys Acta* **1851,** 566–576, doi: 10.1016/j.bbalip.2015.01.013 (2015).
62. Stefan, N., Schick, F. & Haring, H. U. Sex hormone-binding globulin and risk of type 2 diabetes. *N Engl J Med* **361,** 2675–2676; author reply 2677-2678, doi: 10.1056/NEJMc0910143 (2009).
63. Onat, A. *et al.* Fatty liver disease: Disparate predictive ability for cardiometabolic risk and all-cause mortality. *World J Gastroenterol* **21,** 13555–13565, doi: 10.3748/wjg.v21.i48.13555 (2015).
64. Flechtner-Mors, M. *et al.* Associations of Fatty Liver Disease and Other Factors Affecting Serum SHBG Concentrations: A Population Based Study on 1657 Subjects. Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme, doi: 10.1055/s-0033-1354369 (2013).
65. Shin, J. Y. *et al.* Serum sex hormone-binding globulin levels are independently associated with nonalcoholic fatty liver disease in people with type 2 diabetes. *Diabetes Res Clin Pract* **94,** 156–162, doi: 10.1016/j.diabres.2011.07.029 (2011).
66. Lazo, M. *et al.* Association Between Endogenous Sex Hormones and Liver Fat in a Multiethnic Study of Atherosclerosis. *Clin Gastroenterol Hepatol* **13,** 1686–1693 e1682, doi: 10.1016/j.cgh.2014.12.033 (2015).
67. Hua, X. *et al.* Low serum sex hormone-binding globulin is associated with nonalcoholic fatty liver disease in type 2 diabetic patients. *Clin Endocrinol* (*Oxf*) **80,** 877–883, doi: 10.1111/cen.12360 (2014).

68. Flechtner-Mors, M. *et al.* Associations of fatty liver disease and other factors affecting serum SHBG concentrations: a population based study on 1657 subjects. Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme **46,** 287–293, doi: 10.1055/s-0033-1354369 (2014).
69. Zou, C. *et al.* Lack of Fas antagonism by Met in human fatty liver disease. *Nature medicine* **13,** 1078–1085, doi: 10.1038/nm1625 (2007).
70. Cheung, K. J. *et al.* Usefulness of a novel serum proteome-derived index FI-PRO (fibrosis-protein) in the prediction of fibrosis in chronic hepatitis C. *Eur J Gastroenterol Hepatol* **23,** 701–710, doi: 10.1097/MEG.0b013e3283471b74 (2011).
71. Ferrin, G. *et al.* Identification of candidate biomarkers for hepatocellular carcinoma in plasma of HCV-infected cirrhotic patients by 2-D DIGE. *Liver Int* **34,** 438–446, doi: 10.1111/liv.12277 (2014).
72. Verrijken, A. *et al.* Prothrombotic factors in histologically proven nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Hepatology* **59,** 121–129, doi: 10.1002/hep.26510 (2014).
73. Castro, M. A., Goodwin, T. M., Shaw, K. J., Ouzounian, J. G. & McGehee, W. G. Disseminated intravascular coagulation and antithrombin III depression in acute fatty liver of pregnancy. *Am J Obstet Gynecol* **174,** 211–216 (1996).
74. Neubauer, K., Baruch, Y., Lindhorst, A., Saile, B. & Ramadori, G. Gelsolin gene expression is upregulated in damaged rat and human livers within non-parenchymal cells and not in hepatocytes. *Histochem Cell Biol* **120,** 265–275, doi: 10.1007/s00418-003-0564-x (2003).
75. Lee, Y. A., Wallace, M. C. & Friedman, S. L. Pathobiology of liver fibrosis: a translational success story. *Gut* **64,** 830–841, doi: 10.1136/gutjnl-2014-306842 (2015).
76. Mazumdar, B., Meyer, K. & Ray, R. N-terminal region of gelsolin induces apoptosis of activated hepatic stellate cells by a caspase-dependent mechanism. *PLoS One* **7,** e44461, doi: 10.1371/journal.pone.0044461 (2012).
77. Montag-Sallaz, M., Schachner, M. & Montag, D. Misguided axonal projections, neural cell adhesion molecule 180 mRNA upregulation, and altered behavior in mice deficient for the close homolog of L1. *Mol Cell Biol* **22,** 7967–7981 (2002).
78. Santamaria, I., Velasco, G., Pendas, A. M., Fueyo, A. & Lopez-Otin, C. Cathepsin Z, a novel human cysteine proteinase with a short propeptide domain and a unique chromosomal location. *J Biol Chem* **273,** 16816–16823 (1998).
79. Kinner, S., Reeder, S. B. & Yokoo, T. Quantitative Imaging Biomarkers of NAFLD. *Dig Dis Sci* **61,** 1337–1347, doi: 10.1007/s10620-016-4037-1 (2016).
80. Ledinghen, V. *et al.* Controlled attenuation parameter for the diagnosis of steatosis in non-alcoholic fatty liver disease. *J Gastroenterol Hepatol* **31,** 848–855, doi: 10.1111/jgh.13219 (2016).

Acknowledgements

Support for this project was provided by National Institutes of Health grants: R01 DK088231.

Author Contributions

All authors reviewed the manuscript. G.C.W. and X.C. performed the experiments and conducted the analyses. C.D.S., G.S.G. and D.J.C. conceived of the study. G.C.W., G.A., P.B., D.R., T.M., A.T.P., J.G., D.J.C., J.K.D., C.D.S. and G.S.G. contributed to the manuscript drafting, reviewing, and preparation.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Wood, G. C. *et al.* A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains. *Sci. Rep.* **7**, 43238; doi: 10.1038/srep43238 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.