 doi:10.20944/preprints202103.0400.v1

Article

The rs599839 A>G variant disentangles cardiovascular risk and hepatocellular carcinoma in NAFLD patients

Meroni Marica PhD[1,2], Longo Miriam MSc[1,3], Paolini Erika MSc[1,4], Alisi Anna PhD[5], Miele Luca MD[6], De Caro Emilia Rita MSc[1], Pisano Giuseppina MD[1], Maggioni Marco MD[7], Soardo Giorgio MD[8], Valenti Luca MD[2,9], Fracanzani Anna Ludovica MD[1,2], Dongiovanni Paola MSc[1].

1 General Medicine and Metabolic Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy;
2 Department of Pathophysiology and Transplantation, Università degli Studi di Milano, Milano, Italy;
3 Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Italy;
4 Department of Pharmacological and Biomolecular Sciences, Università degli Studi di Milano, Italy;
5 Research Unit of Molecular Genetics of Complex Phenotypes, Bambino Gesù Children Hospital, IRCCS, Rome, Italy;
6 Area Medicina Interna, Gastroenterologia e Oncologia Medica, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy;
7 Department of Pathology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy;
8 Clinic of Internal Medicine-Liver Unit Department of Medical Area (DAME) University School of Medicine, Udine, Italy and Italian Liver Foundation AREA Science Park - Basovizza Campus, Trieste, Italy.
9 Translational Medicine, Department of Transfusion Medicine and Hematology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico Milano;
* Correspondence: paola.dongiovanni@policlinico.mi.it. Phone: +39-02-5503-3467. Fax +39-02-5503-4229.

Citation: PhD, M.M.; MSc, L.M.; MSc, P.E.; PhD, A.A.; MD, M.L.; MSc, D.C.E.R.; MD, P.G.; MD, M.M.; MD, S.G.; MD, V.L.; et al. The rs599839 A>G variant disentangles cardiovascular risk and hepatocellular carcinoma in NAFLD patients. *Cancers* **2021**, *13*, x. https://doi.org/10.3390/xxxxx

Abstract: Background and Aims: Dyslipidemia and cardiovascular diseases (CAD) are comorbidities of nonalcoholic fatty liver disease (NAFLD), which ranges from steatosis to hepatocellular carcinoma (HCC). The rs599839 A>G variant, in the CELSR2-PSRC1-SORT1 cluster, has been associated CAD, but its impact on metabolic traits and liver damage in NAFLD has not been investigated yet. Methods: We evaluated the effect of the rs599839 variant in 1426 NAFLD patients (Overall cohort) of whom 131 have HCC (NAFLD-HCC), in 500,000 individuals from the UK Biobank Cohort (UKBBC) and in 366 HCC samples from The Cancer Genome Atlas (TCGA). Hepatic *PSRC1*, *SORT1* and *CELSR2* expressions were evaluated by RNAseq (n=125). Results: The rs599839 variant was associated with reduced circulating LDL, carotid intima-media thickness, carotid plaques and hypertension ($p<0.05$) in NAFLD patients and with protection against dyslipidemia in UKBBC. The G allele was associated with higher risk of HCC and advanced tumor stage ($p<0.05$) in the Overall cohort. Hepatic PSRC1, SORT1 and CELSR2 expressions were increased in NAFLD patients carrying the rs599839 variant ($p<0.0001$). SORT1 mRNA levels negatively correlated with circulating lipids and with those of genes involved in lipoprotein turnover ($p<0.0001$). Conversely, PSRC1 expression was positively related to that of genes implicated in cell proliferation ($p<0.0001$). In TCGA, PSRC1 over-expression promoted more aggressive HCC development ($p<0.05$). Conclusions: In sum, the rs599839 A>G variant improves dyslipidemia thus protecting against CAD in NAFLD patients, but as one it might promote HCC development by modulating SORT1 and PSRC1 expressions which impact on lipid metabolism and cell proliferation, respectively.

Keywords: Lipid metabolism; NAFLD; genetic variants; PSRC1; HCC

Short title: the rs599839 A>G variant increases HCC risk

 doi:10.20944/preprints202103.0400.v1

1. Introduction

Nonalcoholic fatty liver disease (NAFLD) or more recently re-defined metabolic-associated fatty liver disease (MAFLD) is the most common chronic liver disorder worldwide, affecting 20% to 40% of the adult population [1,2]. Thus, given its increasing proportion, it represents a primary health, social and economic concern [3]. NAFLD is defined by enhanced hepatic fat accumulation not explained by alcohol abuse, and it embraces a broad spectrum of hepatic injuries, spanning from simple steatosis to nonalcoholic steatohepatitis (NASH). The latter may be complicated by the presence of fibrosis and, in a minor percentage of cases, it may progress to cirrhosis and hepatocellular carcinoma (HCC), that is the fifth most common cancer worldwide with persistently increasing mortality in Europe, North/South America and Africa [4]. NAFLD and its progressive forms have a strong inherited component, and variants in genes regulating lipid handling, including *patatin-like phospholipase domain-containing 3* (*PNPLA3*), *transmembrane 6 superfamily member 2* (*TM6SF2*), *membrane bound O-acyltransferase domain-containing 7* (*MBOAT7*), predispose to NAFLD development and progression towards end-stage conditions [5,6]. Moreover, sedentary lifestyle and unhealthy dietary habits may represent paramount environmental risk factors for NAFLD pathogenesis [7]. Hence, its development is closely intertwined with obesity, insulin resistance (IR) and metabolic syndrome features among which atherogenic dyslipidemia and cardiovascular risk [8,9].

A genome wide association study (GWAS) identified the novel intergenic rs599839 A>G variant, as modifier of the risk of coronary artery disease (CAD) [10]. Afterwards, the more frequent A allele has been associated with increased risk of myocardial infarction, ischemic stroke and elevated plasma cholesterol, thereby modulating lipoprotein metabolism [11,12]. Conversely, the minor G allele has been related to a protection against cardiovascular complications and to a reduction of circulating cholesterol in CAD patients [13].

The rs599839 polymorphism is localized in the 1p13.3 *locus* related to lipid traits, in the genetic cluster of *Cadherin EGF LAG Seven-Pass G-Type Receptor 2* (*CELSR2*)- *Proline and Serine Rich Coiled-Coil 1* (*PSRC1*)- *Sortilin 1* (*SORT1*). In details, *CELSR2* gene encodes a non-classic type of cadherin involved in cell adhesion [14], *PSRC1* gene product participates to microtubule destabilization and spindle assembly [15] and its overexpression has been previously detected in tumor liver tissues and in hepatoma cells, where it is associated with HCC recurrence after resection [16]. Finally, *SORT1* gene encodes the Sortilin 1 protein that is involved in lipoprotein clearance [17]. The impact of the rs599839 variant on CELSR2, PSRC1 and SORT1 expressions remain to be fully elucidated. However, these genes are involved in distinct pathways, mainly related to lipid handling and dysregulated cell proliferationand alterations in their expressions may possibly affect different processes.

Thus, the aim of this study was to examine the effect of the rs599839 variant on metabolic traits, cardiovascular risk and progressive liver damage in a large histologically characterized cohort of NAFLD patients at risk of cardiovascular comorbidities, in the general population by using the publicly available UK biobank cohort (UKBBC) and in patients affected by HCC from The Cancer Genome Atlas (TCGA) dataset. Finally, we examined the impact of the rs599839 variant on the hepatic expression of *PSRC1*, *SORT1* and *CELSR2* and of those genes involved in lipoprotein turnover and release, in cell cycle progression and cell proliferation.

2. Materials And Methods

2.1. Overall Cohort

The Overall cohort consists of 1426 patients with NAFLD and it has been subdivided in the Hepatology service cohort (n=1295) and the NAFLD-HCC cohort (n=131). The Hepatology service cohort has been previously described [18,19]. Briefly, 1295 adult individuals were consecutively enrolled at the Metabolic Liver Diseases outpatient service

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

(n=713) and Bariatric Surgery center (n=582) at Fondazione IRCCS Ca' Granda Ospedale Policlinico Milan, Italy. Inclusion criteria were availability of liver biopsy performed for suspected NASH or severe obesity, availability of DNA samples and clinical data. Individuals with excessive alcohol intake (men, >30 g/day; women, >20 g/day), viral and autoimmune hepatitis, hereditary haemochromatosis and alpha1-antitrypsin deficiency or other causes of liver disease were excluded. The clinical characteristics of patients evaluated in the study are listed in **Table 1.**

Table 1. Demographic, anthropometric and clinical features of the Overall cohort (n=1426) stratified according to enrollment criteria (n=1295 Hepatology service cohort and n=131 NAFLD-HCC).

	Overall Cohort (n=1426)	**Hepatology service cohort (n=1295)**	**NAFLD-HCC (n=131)**	***P value*†**
Sex, M	770 (54)	669 (52)	101 (78)	<0.0001
Age, years	49.6±13.6	48±12.6	68±10	<0.0001
BMI, kg/m2	34.2±8.7	34.7±8.76	28.7±5.12	<0.0001
T2D, yes	385 (27)	310 (24)	75 (57)	<0.0001
Total cholesterol, mmol/L	5.1±1.07	5.2±1.05	4.2±1.17	<0.0001
LDL cholesterol, mmol/L	3.1±0.97	3.2±0.95	2.37±0.96	<0.0001
HDL cholesterol, mmoL/L	1.3±0.39	1.3±0.4	1.28±0.51	0.73
Triglycerides, mmoL/L	1.63±1.00	1.63±0.95	1.42±1.76	0.11
ALT, IU/l	34 {20-57}	33 {20-57}	42 {27-55}	0. 07
AST, IU/l	25 {19-38}	25 {18-37}	39 {26-59}	<0.0001
PNPLA3, I148M				<0.0001
I/I	557 (39)	524 (41)	33 (25)	
I/M	627 (44)	567 (43)	60 (46)	
M/M	242 (17)	204 (16)	38 (29)	
TM6SF2, E167K				0.25
E/E	1240 (87)	1131 (87.3)	109 (83.5)	
E/K	168 (11.8)	150 (11.6)	18 (14)	
K/K	18 (1.2)	14 (1.1)	4 (2.5)	
MBOAT7, rs641738 T allele				0. 11
C/C	456 (32)	423 (33)	33 (25)	
C/T	642 (45)	581 (45)	61 (47)	
T/T	328 (23)	291 (22)	37 (28)	
rs599839 A>G				**0.01**
A/A	884 (62)	808 (63)	76 (58)	
A/G	456 (32)	417 (32)	39 (30)	
G/G	86 (6)	70 (5)	16 (12)	

†NAFLD-HCC *vs* Hepatology service cohort.

The NAFLD-HCC cohort includes 131 NAFLD patients who developed HCC. Part of the this cohort has been previously described [20,21]. Patients were consecutively enrolled at the Metabolic Liver Diseases outpatient service at Fondazione IRCCS Cà Granda, Ospedale Policlinico, Milan, Italy, at the Internal Medicine and Gastroenterology Area, Fondazione Policlinico Universitario Gemelli and IRCCS, Catholic University, Rome, Italy and at Internal Medicine-Liver Unit, University Hospital of Udine, Italy. Clinical features of the NAFLD-HCC cohort patients are presented in **Table 1.** Diagnosis of HCC was based on the EASL–EORTC Clinical Practice Guidelines [22]. In the absence of liver biopsy, diagnosis of NAFLD required detection of ultrasonographic steatosis plus at least one criterion of the metabolic syndrome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

Informed written consent was obtained from each patient and the study protocol was approved by the Ethical Committees of the Fondazione IRCCS Ca' Granda, Milan, of Fondazione Policlinico Universitario Gemelli and IRCCS, Catholic University, Rome, and from University Hospital of Udine and conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

2.2. Histological Evaluation

Steatosis was graded according to the percentage of affected hepatocytes as 0: 0-4%, 1: 5-32%, 2: 33-65%, and 3: 66-100%. Disease activity was assessed according to the NAFLD Activity Score (NAS) with systematic evaluation of hepatocellular ballooning and necroinflammation; fibrosis was also staged according to the recommendations of the NAFLD Clinical Research Network [23]. The scoring of liver biopsies was performed by an independent pathologists unaware of patients' status and genotype [18,24]. NASH was diagnosed when a) steatosis, b) lobular inflammation and c) ballooning were concomitantly present.

2.3. Genotyping

The Overall cohort has been genotyped for the rs738409 C>G (PNPLA3 I148M), rs58542926 C>T (TM6SF2 E167K), rs641738 C>T *MBOAT7*, and rs599839 A>G variants using TaqMan 5'-nuclease assays in duplicate (QuantStudio 3, Thermo Fisher, Waltham, MA), as previously described [18,19]. The success rate of genotyping was >99%. The frequency distribution of the rs599839 A>G was not in Hardy-Weinberg equilibrium (p=0.01, **Table S1**) and it was compared to that obtained in European not-Finnish healthy individuals included in the 1000 Genome project [25].

2.4. UK Biobank Cohort

The association between the rs599839 A>G *PSRC1* variant and phenotypes related to metabolic disorders and liver disease were evaluated in the UK Biobank cohort (UKBBC). UKBBC is a prospective population study of 500,000 subjects not selected for liver diseases and ethnicity, almost all aged 40-69 years, identified in 22 centers across the UK during 2006-2010. Freely available basic association data were downloaded from Neale Lab (http://geneatlas.roslin.ed.ac.uk) and *p*-values were corrected for multiple testing using the False Discovery Rate (FDR) method [26].

 doi:10.20944/preprints202103.0400.v1

2.5. Transcriptomic Analysis

RNA-seq was performed in a subset of 125 severely obese patients (21 without and 104 with NAFLD) belonging to the Hepatology service cohort, of whom percutaneous liver biopsy was performed during bariatric surgery at Fondazione IRCCS Cà Granda, Ospedale Policlinico, Milan, Italy. The study was conformed to the Declaration of Helsinki and approved by the Institutional Review Boards and their Ethics Committees. All participants gave written informed consent. Clinical characteristics of the Transcriptomic cohort are presented in **Table S2.** RNA-seq mapping descriptive statistics, the detailed protocol and data analysis approach are described in the Supplemental Materials and Methods.

2.6. The Cancer Genome Atlas-Liver Hepatocellular Carcinoma (TCGA-LIHC) Data Description

The Cancer Genome Atlas-Liver Hepatocellular Carcinoma (TCGA-LIHC) database is a large project which applies high-throughput genome analysis techniques, combining genome sequencing and bioinformatic tools, in order to catalogue genetic mutations responsible for cancer. It is a comprehensive publicly available resource, that contains information about expression level of multiple genes. TCGA datasets of 366 HCC samples were directly downloaded from cBioPortal for Cancer Genomics. The detailed information of the microarray and RNA-Seq experiments, protocols, and software used can be found at the cBioPortal for Cancer Genomics at https://www.cbioportal.org [27,28]. Relative mRNA expression was represented as Z scores, precomputed from the expression values specifying the threshold (2 Standard Deviations from the mean). The z scores for mRNA expression are determined for each sample by comparing a gene's mRNA expression to the distribution in a reference population that represents typical expression of that gene. TCGA tumor stage system classification and Bioinformatic resources are described in the Supplemental Materials and Methods.

2.7. Statistical Analysis

Statistical analyses were performed using JMP 15.0 (SAS, Cary, NC), R statistical analysis version 3.3.2 (*http://www.R-project.org/*) and Prism (version 6, GraphPad Software), by using one-way analysis of variance (ANOVA) or chi-square test, where appropriate.

For descriptive statistics, continuous variables were shown as mean and standard deviation or median and interquartile range for highly skewed biological variables (i.e. AST, ALT, triglycerides (TGs)). Variables with skewed distributions were logarithmically transformed before analyses. Categorical variables were presented as number and proportion. All genetic analyses were performed under additive and recessive models.

Analyses were performed by fitting data to generalized linear regression models. In particular, general linear models were fit to examine continuous traits. Multinomial logistic regression models were fit to examine binary traits (cirrhosis, HCC), and ordinal regression models were fit for ordinal traits (components of the NAFLD activity score: severity of steatosis, necroinflammation and hepatocellular ballooning, stage of fibrosis). When specified, confounding factors were included in a model. Correlations were assessed by bivariate analysis. For gene expression analyses differences between groups were calculated by one-way ANOVA, which was followed by post hoc *t*-tests adjusted for the number of comparisons when multiple groups were involved (Bonferroni correction). *P* values <0.05 (two-tailed) were considered statistically significant.

3. Results

3.1. The rs599839 A>G Gene Variant Affects Circulating Lipids and Cardiovascular Risk in NAFLD Patients

Clinical characteristics of the Overall cohort stratified according to the rs599839 A>G variant are shown in **Table 2.** No differences in demographic and anthropometric features were found across genotypes. At bivariate analysis, circulating total cholesterol (TC) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

low-density lipoprotein (LDL) cholesterol were reduced in NAFLD patients across the rs599839 genotype, while high-density lipoprotein (HDL) cholesterol levels were higher in patients who carry the minor G allele (p<0.001 at one-way ANOVA; adjusted p<0.0001 for GG *vs*. AA**, Fig.1A-C**). To sum up these results, the frequency of dyslipidemia was reduced in patients harboring the rs599839 genotype (p=0.02 at Pearson test, **Fig.S1A**).

Table 2. Demographic, anthropometric and clinical features of Overall Cohort (n=1426) stratified by rs599839 A>G genotype.

	AA (n=884)	**AG (n=456)**	**GG (n=86)**	***P*-value°**	***P*-value+**
Sex, M	487 (55)	234 (51)	49 (56)	0.33	0. 72
Age, years	49.18±13.5	49.7±13.5	50.0±14.1	0.74	0.67
BMI, kg/m²	33.9±8.50	34.4±8.78	35.8±10.4	0.16	0.10
IFG/T2D, yes (%)	234 (26)	1123 (26)	28 (32)	0.49	0.23
Glucose mg/dL	102±30	103±30	107±39	0.56*	0.18*
HOMA-IR	5.3±6.4	5.5±11	4.56±3.87	0.91*	0.65*
Insulin, IU/ml	20.7±18.4	22.7±34.5	15.9±7.88	0.97*	0.28*
Total cholesterol, mmol/L	5.2±1.1	5.0±1.0	4.9±1.0	**0.005***	0. 15*
LDL cholesterol, mmol/L	3.25±0.97	3.03±0.95	2.8±0.98	**<0.0001***	**0.03***
HDL cholesterol, mmol/L	1.26±0.3	1.34±0.43	1.36±0.37	**0.0003***	0.12*
Triglycerides, mmol/L	1.66±1.06	1.55±0.85	1.66±1.07	0.28*	0.83*
Dyslipidemia, yes (%)	300 (34)	127 (28)	15 (18)	**0.01***	**0.007***
ALT, IU/L	35{21-56}	33{20-62}	28{19-52}	0.26*	0.78*
AST, IU/L	26{19-38}	25{19-39}	24{18-36}	0.76*	0.72*
Iron ug/dL	97.8±44.4	96.8±40.8	9.48±33.6	0.23**	0.67**
Transferrin mg/dL	270.4±60.8	264.6±47.2	258.6±47.0	0.13**	0.48**
Transferrin saturation (%)	30.1±24.9	29.1±15.2	34.9±43.9	0.79**	0.06**
Ferritin ng/mL	324.4±411.6	386.2±508.2	459.2±526.9	0.28**	0.31**

Values are reported as mean ± SD. number (%) or median {IQR}. as appropriate. BMI: body mass index. IFG: impaired fasting glucose. T2D: type 2 diabetes. Characteristics of participants were compared across the rs599839 genotypes using generalized linear model (for continuous characteristics) or nominal logistic regression model (for categorical characteristics). *Models were adjusted for gender, age, BMI, IFG/T2D, PNPLA3 I148M alleles. TM6SF2 E167K alleles and MBOAT7 rs641738 T alleles. **Models were adjusted also for HFE C282Y and HFE H63D alleles.

°Additive model, ⁺Recessive model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

Figure 1

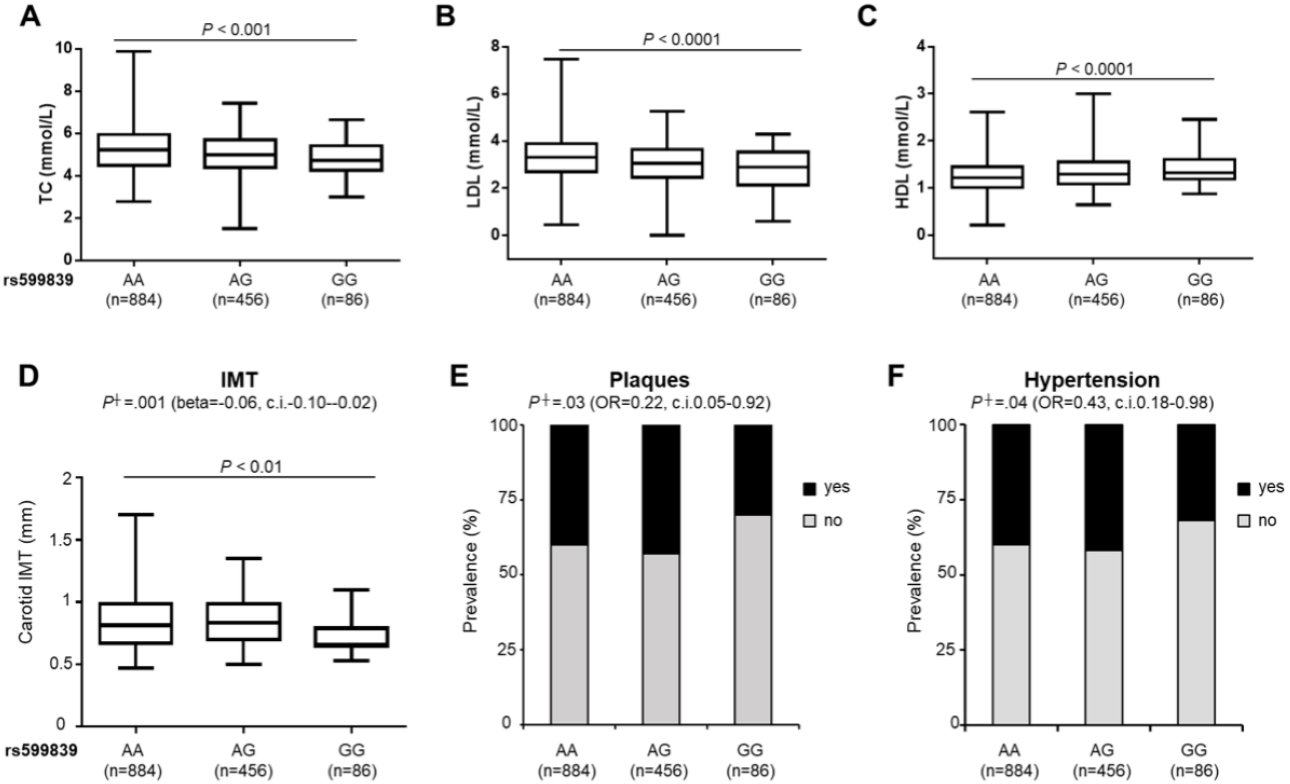

Figure 1. The rs599839 variant affects circulating lipids, carotid IMT, plaque formation and incidence of hypertension in patients with NAFLD. Circulating total cholesterol (TC) (mmol/L) (A), LDL (mmol/L) (B), HDL (mmol/L) (C) were evaluated in NAFLD patients from the Overall cohort (n=1426) and stratified by the presence of the rs599839 G allele. Boxes span from 25° to 75° percentile, while whiskers indicate the 10° and 90° percentile. * *P*< 0.001 at one-way ANOVA. Association of the rs599839 variant with IMT (D), plaque presence (E) and hypertension (F) in NAFLD patients from the Overall cohort (n=1426). Multivariable generalized linear model (for IMT) or nominal logistic regression analysis (for Plaques and Hypertension) adjusted for age, sex, BMI, T2D, presence of TM6SF2 E167K, statin use and active smoking at ┼recessive model. * *P*< 0.01 at one-way ANOVA.

At multivariate generalized linear models, we found that the G allele was associated with reduced TC (beta: -0.15; 95% c.i. -0.25--0.04; *p*=0.005), LDL cholesterol (beta: -0.20; 95% c.i. -0.30--0.10; *p*<0.0001) and with increased HDL cholesterol (beta: 0.07; 95% c.i. 0.03-0.10; *p*=0.0003), after adjustment for sex, age, body mass index (BMI), type 2 diabetes (T2D), PNPLA3 I148M, TM6SF2 E167K and *MBOAT7* rs641738 variants by using an additive model (**Table 2**). Consistently, at multivariate generalized analysis, the prevalence of dyslipidemia was lower in patients who carry the minor G allele (OR: 0.76; 95% c.i. 0.60-0.95; *p*=0.01), adjusted for the same confounders (**Table 2** and **Fig.S1A**).

At recessive model, LDL cholesterol (beta: -0.14; 95% c.i. -0.27--0.009; *p*=0.03), along with dyslipidemia (OR: 0.37; 95% c.i. 0.18-0.77; *p*=0.007) remained associated with the rs599839 variant (**Table 2** and **Fig.S1A**). Conversely, we did not find any association between the rs599839 variant and circulating TGs, T2D or IR.

Therefore, we next evaluated the impact of the rs599839 G allele on cardiovascular comorbidities in the Overall cohort. At multivariate generalized analyses, it was associated with reduced carotid intima-media thickness (IMT) (beta: -0.06; 95% c.i. -0.10--0.02; *p*=0.001), carotid artery plaque formation (OR: 0.22; 95% c.i. 0.05-0.92; *p*=0.03) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

hypertension (beta: 0.43; 95% c.i. 0.18-0.98; p=0.04) after adjustment for sex, age, BMI, T2D, TM6SF2 E167K variant, statin use and active smoking only by running a recessive model (**Fig.1D-F and Table 3**).

Table 3. Independent predictors of intima media thickness (IMT), plaque presence and hypertension in 1426 patients with NAFLD from the Overall cohort.

	IMT			**Plaques, yes**			**Hypertension, yes**		
		95% CI	*P*	OR	95% CI	*P*	OR	95% CI	*P*
Sex, M	0.05	0.02-0.06	<0.0001	1.78	1.04-3.06	0.03	1.32	0.95-1.84	0.09
Age, years	0.008	0.006-0.009	<0.0001	1.09	1.07-1.12	<0.0001	1.06	1.05-1.07	<0.0001
BMI, kg/m²	0.001	-0.003-0.005	0.54	1.02	0.97-1.07	0.41	1.06	1.04-1.08	<0.0001
IFG/T2D, yes (%)	0.002	-0.02-0.02	0.86	1.54	0.91-2.61	0.11	1.68	1.18-2.38	0.003
TM6SF2, E167K allele	-0.02	-0.06-0.01	0.22	0.78	0.46-1.34	0.37	1.71	1.16-2.51	0.006
Statin use, yes	-0.01	-0.04-0.01	0.38	0.43	0.22-0.83	0.01	0.97	0.56-1.66	0.91
Active smoking	0.01	-0.01-0.03	0.36	1.840	1.08-3.14	0.02	1.30	0.92-1.84	0.12
rs599839 GG yes	-0.06	-0.10—0.02	**0.001**	0.22	0.05-0.92	**0.03**	0.43	0.18-0.98	**0.04**

CI: confidence interval. Values were obtained at multivariate generalized linear analysis (for IMT) or nominal logistic regression analysis (for Plaques and Hypertension) adjusted for sex, age, BMI (body mass index), T2D (type 2 diabetes mellitus) and TM6SF2 E167K alleles, statin use and active smoking by using a recessive model.

3.2. The rs599839 A>G Variant is not Associated with Histological NAFLD

Since the rs599839 variation influences circulating lipid concentrations, which are strongly entangled in hepatic fat accumulation, we analyzed its impact on liver damage in NAFLD patients from the Overall cohort. At ordinal regression analysis, the rs599839 variant was not related to steatosis, lobular inflammation, ballooning, fibrosis (**Fig.S1A-D**) and cirrhosis neither at additive or recessive models (**Table 4, Table S3 and Table S4**), thus suggesting that the G allele did not impact on in the histological spectrum of NAFLD. This mechanism differs from what we observed for the E167K TM6SF2 variant, which by impairing VLDL release, induces fat accumulation in the liver and at the same time protects against CAD [18].

Table 4. Independent predictors of liver damage in 1426 patients with NAFLD from the Overall cohort.

	Steatosis			**Lobular inflammation**			**Ballooning**			**Fibrosis**		
		95% CI	*P*		95% CI	*P*		95% CI	*P*		95% CI	*P*
Sex, M	0.29	0.17-0.41	<0.0001		-	<0.0001	0.29	0.14-0.44	<0.0001	0.32	0.20-0.44	<0.0001
Age, years	0.002	-0.007-0.01	0.67		-	<0.0001	0.03	0.02-0.04	<0.0001	0.07	0.04-0.06	<0.0001
BMI, kg/m²	0.04	0.02-0.05	<0.0001		-	<0.0001	-0.003	-0.02-0.01	0.67	0.002	-0.01-0.015	0.81
IFG/T2D, yes (%)	0.33	0.19-0.47	<0.0001		-	<0.0001	0.32	0.17-0.48	<0.0001	0.60	0.47-0.74	<0.0001
PNPLA3, I148M allele	0.43	0.28-0.57	<0.0001		-	<0.0001	0.28	0.10-0.47	0.002	0.49	0.34-0.64	<0.0001
TM6SF2, E167K allele	0.75	0.46-1.04	<0.0001		-	0.0001	0.41	0.08-0.75	0.01	0.57	0.29-0.84	<0.0001
MBOAT7, rs641738 T allele	0.11	-0.04-0.25	0.14	-	-		0.02	-0.15-0.20	0.76	0.16	0.01-0.31	0.02
rs599839 G allele	-0.02	-0.19-0.16	0.83	-	-		0.20	-0.02-0.42	0.07	0.02	-0.16-0.19	0.82

CI: confidence interval. Values were obtained at multivariate ordinal regression analysis adjusted for sex, age, BMI (body mass index), T2D (type 2 diabetes mellitus) and PNPLA3 I148M alleles, TM6SF2 E167K alleles and *MBOAT7* rs641738 T allele by using an additive model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

3.3. The rs599839 Variation is Associated with Increased Risk of HCC in NAFLD Patients

Since the rs599839 variant is located at 500b downstream of the 3' untranslated region (UTR) of *PSRC1* gene, which is required for the congress of chromosomes at the metaphase plate and for normal rate of chromosomal segregation during anaphase, we next sought to examine whether the rs599839 variation may affect the risk to develop HCC in NAFLD patients. The frequency distribution of the minor G allele in the Hepatology service cohort (n=1295) and in the NAFLD-HCC cohort (n=131) is shown in **Fig.2A**, in **Table 1** and in **Table S1**. The minor G allele was more frequent in NAFLD patients with HCC than those without cancer (p=0.01 at one-way ANOVA (**Table 1**); 27% *vs* 21% p=0.03 at Fisher-exact Test NAFLD-HCC *vs* Hepatology service cohort; **Fig.2A** and **Table S1**) and the percentage of GG homozygous patients was even more higher in NAFLD-HCC cohort (12% *vs* 5% NAFLD-HCC *vs* Hepatology service cohort; **Table 1).** Furthermore, at multivariate analysis, the G allele was independently associated with enhanced risk of HCC at both additive (OR: 1.70; 95% c.i. 1.08-2.70; p=0.02) and recessive models (OR: 5.85; 95% c.i. 2.12-16.12; p=0.0006). Notably, the effect of the variant on HCC development seems to be independent of the liver disease severity, as adjustment for advanced fibrosis did not abolish the association between the G allele and HCC at both models (OR: 1.70; 95% c.i. 1.03-2.80; p=0.03; and OR: 5.62; 95% c.i. 1.77-17.84; p=0.003, respectively) (**Fig.2B and Table 5**). Furthermore, at multivariate analysis, the rs599839 variant is associated with an increased risk to develop advanced tumor stages (Stage (S) >1; OR: 3.27; 95% C.I. 1.36-7.85; p=0.008) and more pronounced primary tumor extensions (Tumor size (T) >1; OR: 2.86; 95% C.I. 1.17-6.96; p=0.02). Consistently, the at-risk G allele correlates with higher Child-Pugh scores (Child-Pugh >A6; OR: 4.38; 95% C.I. 1.35-14.22; p=0.01) (**Table 6**). Collectively, these findings suggest that the rs599839 variant might increase the risk to develop aggressive HCC in NAFLD patients independently of hepatic fat accumulation or fibrosis, and more so in homozygous subjects.

Figure 2

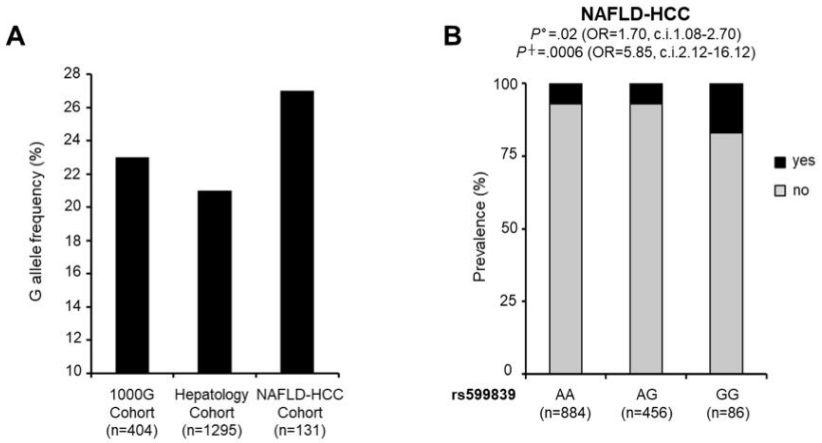

Figure 2. The rs599839 variant influences the risk of HCC in NAFLD patients. Frequencies distribution of the rs599839 variant across the 404 individuals from 1000 genomes European non-Finnish cohort, Hepatology service cohort (n=1295) and NAFLD-HCC cohort (n=131) (**A**). Association of the rs599839 variant with HCC (n=131 cases from the NAFLD-HCC cohort and n=1295 controls with NAFLD from the Hepatology service cohort). Multivariable nominal logistic regression analysis adjusted for age, sex, BMI, T2D, presence of PNPLA3 I148M, TM6SF2 E167K and MBOAT7 T alleles at ° additive or ┼recessive model (**B**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

Table 5. Independent predictors of NAFLD-HCC in 1426 patients with NAFLD (Cases=131).

	HCC°			HCC+			HCC° adjusted for f>2			HCC+ adjusted for f>2		
	OR	95% CI	*P*	OR	95% CI	*P*	OR	95% CI	*P*	OR	95% CI	*P*
Sex, M	1.82	0.98-3.36	0.05	1.64	0.89-3.01	0.10	2.23	1.14-4.34	0.01	2.08	1.07-4.06	0.03
Age, years	1.16	1.12-1.20	<0.0001	1.16	1.12-1.20	<0.0001	1.13	1.09-1.17	<0.0001	1.13	1.09-1.17	<0.0001
BMI, kg/m²	0.87	0.82-0.93	<0.0001	0.88	0.83-0.93	<0.0001	0.90	0.84-0.96	0.0009	0.90	0.84-0.96	0.001
IFG/T2D, yes (%)	3.22	1.79-5.78	<0.0001	3.25	1.80-5.86	<0.0001	1.86	0.96-3.57	0.06	1.90	0.98-3.67	0.05
PNPLA3, I148M allele	1.35	0.93-1.95	0.11	1.34	0.92-1.94	0.12	1.02	0.67-1.53	0.93	1.01	0.67-1.52	0.95
TM6SF2, E167K allele	1.51	0.84-2.72	0.16	1.47	0.82-2.65	0.19	1.20	0.64-2.25	0.57	1.17	0.61-2.17	0.63
MBOAT7, rs641738 T allele	1.33	0.90-1.96	0.14	1.39	0.94-2.06	0.09	1.37	0.90-2.09	0.14	1.42	0.93-2.18	0.10
rs599839 G allele	1.70	1.08-2.70	**0.02**	5.85	2.12-16.12	**0.0006**	1.70	1.03-2.80	**0.03**	5.62	1.77-17.84	**0.003**

CI: confidence interval. Values were obtained at multivariate nominal logistic regression analysis adjusted for sex, age, BMI (body mass index), T2D (type 2 diabetes mellitus) and PNPLA3 I148M alleles, TM6SF2 E167K alleles and *MBOAT7* rs641738 T allele by using an additive° or recessive model†.

Table 6. Independent predictors of advance stage (S), tumor size (T) and Child-Pugh score in 131 NAFLD-HCC patients from the Overall cohort.

	Stage (S) > 1			Tumor size (T) > 1			Child-Pugh > A6		
	OR	95% CI	*P*	OR	95% CI	*P*	OR	95% CI	*P*
Sex, M	1.10	0.33-3.65	0.87	0.82	0.234-2.83	0.01	0.47	0.08-2.90	0.42
Age, years	0.99	0.94-1.04	0.67	0.98	0.93-1.04	0.82	0.97	0.89-1.05	0.51
PNPLA3, I148M allele	0.61	0.30-1.23	0.16	0.41	0.19-0.85	0.93	0.43	0.13-1.45	0.17
TM6SF2, E167K allele	1.08	0.42-2.81	0.87	1.27	0.47-3.41	0.57	0.21	0.22-1.89	0.16
MBOAT7, rs641738 T allele	0.73	0.34-1.55	0.41	0.66	0.30-1.44	0.29	0.43	0.11-1.63	0.21
rs599839 G allele	3.27	1.36-7.85	**0.008**	2.86	1.17-6.96	**0.02**	4.38	1.35-14.22	**0.01**

CI: confidence interval. Values were obtained at multivariate nominal logistic regression analysis adjusted for sex, age, and PNPLA3 I148M alleles, TM6SF2 E167K alleles and *MBOAT7* rs641738 T allele by using an additive model.

3.3. The rs599839 A>G Gene Variant Affects Circulating Cholesterol and CAD Risk but not Liver Damage in UKBBC

The protective effect of the rs599839 G allele on lipid metabolism and cardiovascular risk was confirmed by analyzing the data obtained from the UKBBC. Indeed, there was a strong negative association between the rs599839 G allele and circulating TC levels (beta: -0.017; *p*=7.82*e-112), disorders of lipoprotein metabolism (beta: -0.01; *p*=4.3*e-55), metabolic disorders (beta: -0.01; *p*=3.09*e-44) and the presence of cardiovascular complications such as ischemic heart diseases (beta: -0.006; *p*=6.72*e-25), hypertensive diseases (beta: -0.002; *p*=0.018) and atherosclerosis (beta: -0.0003; *p*=0.037). The clinical phenotypes related to rs599839 variant in the UKBBC are listed in **Table 7**. Remarkably, the rs599839 variant was not correlated with liver failure and hepatic fibrosis and cirrhosis thus confirming what we found in NAFLD patients.

Table 7. Association of the rs599839 A>G variant with liver-related outcomes and biochemical parameters in the UK Biobank cohort (UKBBC).

Phenotype	Cases	Estimate beta	*OR*	*P*-value
K70: Alcoholic liver disease	808	-1.99*e-05	1.01	0.86
K74: Fibrosis and cirrhosis of liver	805	3.31*e-06	0.99	0.97
K75: Other inflammatory liver diseases	662	-9.95*e-05	1.06	0.32
K76: Other diseases of liver	3351	-5.32*e-06	1	0.98
K70-K77: Diseases of liver	4894	-0.0002	1.02	0.39
High cholesterol	55265	-0.017	1.17	7.82*e-112
E78 Disorders of lipoprotein metabolism and other lipidaemias	39308	-0.010	1.14	4.3*e-55
E70-E90 Metabolic disorders	47969	-0.010	1.11	3.09*e-44
I20-I25 Ischaemic heart diseases	33387	-0.006	1.1	6.72*e-25

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

I25 Chronic ischaemic heart disease	27772	-0.006	1.11	9.76*e-25
I20 Angina pectoris	19935	-0.004	1.1	3.85*e-17
Heart/cardiac problem	32474	-0.005	1.08	2.29*e-16
Angina	14399	-0.003	1.12	2.88*e-16
Heart attack/myocardial infarction	10356	-0.002	1.13	3.84*e-14
I21 Acute myocardial infarction	8764	-0.002	1.11	4.05*e-09
I30-I52 Other forms of heart disease	31135	-0.002	1.04	0.0002
I50 Heart failure	5901	-0.001	1.08	0.0003
I71 Aortic aneurysm and dissection	1470	-0.0004	1.14	0.003
Pace-maker	1355	-0.0003	1.12	0.013
I10-I15 Hypertensive diseases	84910	-0.002	1.01	0.018
G45 Transient cerebral ischaemic attacks and related syndromes	2765	-0.0004	1.07	0.03
I70 Atherosclerosis	1371	-0.0003	1.1	0.037

Biochemical parameters are assessed in the entire cohort (n=500,000 subjects). HWE: 0.84 MAF: 0.23.

3.4. Association between rs599839 and other Genetic Variants Located in the 1p13.3 Locus

To investigate whether the associations between the rs599839 polymorphism and metabolic traits were due to other common variants (MAF≥0.01) located in the 1p13.3 *locus*, we examined the linkage disequilibrium pattern at the region spanning 50.00 Kb (chr1: 109,800,000-109,850,000; Human (GRCh37.p13)), using data from 1000 genomes project and considering 503 individuals of European descent (CEU). The rs599839 variant resulted in strong linkage disequilibrium with other 10 common SNPs (rs1277930, rs583104, rs4970836, rs602633, rs7528419, rs629301, rs646776, rs12740374, rs3832016, rs660240) (r^2>0.8; **Table S5)**, localized in the region that ranges from the *CELSR2* 3'UTR, the intergenic region between *CELSR2* and *PSRC1*, and the *PSRC1* 3'UTR oriented in opposite direction (**Fig.S2A**). Thus, we evaluated the clinical phenotypes most significantly associated with these SNPs in the UKKBC and we found that all these variants were negatively correlated with circulating TC concentrations (beta: -0.017; p<0.0001), supporting previous data from Musunuru et al [11] on the effect of the entire 1p13.3 *locus* on circulating lipids and confirming the protective role of the rs599839 minor G allele on lipid metabolism even in NAFLD patients.

3.5. The rs599839 A>G Impacts on PSRC1, SORT1 and CELSR2 Expression

To determine whether the epidemiological association between the rs599839 variant and protection against CAD and higher risk to develop HCC in NAFLD patients may be mediated by the modulation of transcriptional activities at this *locus,* we examined the hepatic expression of PSRC1, SORT1 and CELSR2 in a subset of patients belonging to the Overall cohort of whom transcriptomic data was available (n=125). The rs599839 G allele was associated with higher PSRC1 (p<0.0001 at one-way ANOVA; adjusted p<0.0001 for GG *vs* AA), SORT1 (p<0.0001 at one-way ANOVA; adjusted p<0.0001 for GG *vs* AA) and CELSR2 (p=0.0002 at one-way ANOVA; adjusted p<0.001 for GG *vs* AA) mRNA levels (**Fig.3A-C**). In addition, there was a strong positive correlation between their expressions (p<0.0001; for all comparisons) (**Fig.3D-F**). As we expected, PSRC1, SORT1 and CELSR2, which are predicted to interact according to a network analysis generated by STRING, were significantly co-expressed (p=8.44*e-07, **Fig.S2B**), indicating that these proteins are at least partially biologically connected, as a group.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

Figure 3

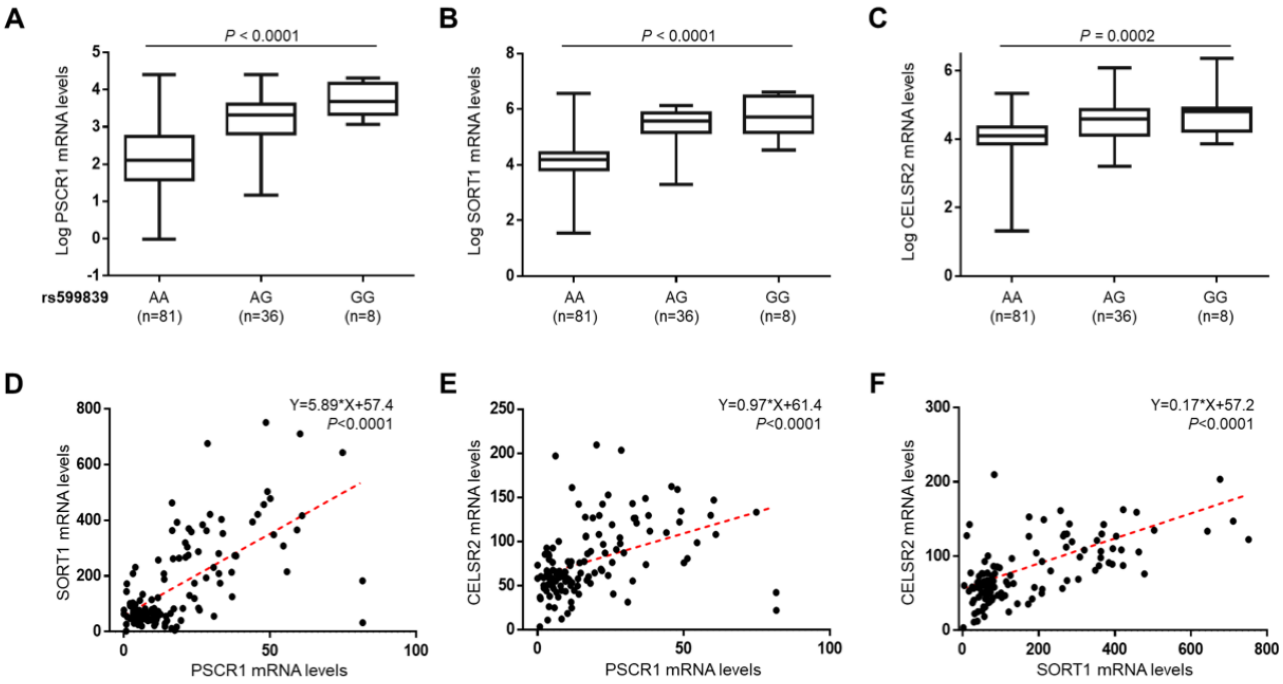

Figure 3. PSRC1, SORT1 and CELSR2 expressions increase in NAFLD patients carrying the rs599839 variant. PSRC1 (**A**), SORT1 (**B**) and CELSR2 (**C**) mRNA levels were evaluated by transcriptome analysis on liver biopsies (n=125) and stratified by the presence of the rs599839 G allele. mRNA levels were represented as log transformed. Boxes span from 25° to 75° percentile, while whiskers indicate the 10° and 90° percentile. * *P*< 0.0001 at one-way ANOVA. Correlation analyses between PSRC1 and SORT1 (**D**), CELSR2 (**E**) and between SORT1 and CELSR2 (**F**).

Our data is consistent with previous findings which indicate that SORT1 is the primary mediator of circulating lipids at this *locus*. Indeed, SORT1 expression, but not PSRC1 or CELSR2, was significantly correlated with the reduction of TC concentrations (Y=-0.002X+5.72; *p*=0.014), and more strongly with the decrease in LDL cholesterol (Y=-0.002X+3.65; *p*=0.0008) and serum TGs (Y=-0.001X+1.64; *p*=0.014). However, SORT1 mRNA levels did not correlate with HDL cholesterol (**Fig.4A-D**). Even more, the expression of SORT1 was tightly correlated with that of Apolipoprotein A1 (APOA1; Y=-40.9X+83971; *p*=0.04), Apolipoprotein E (APOE; Y=-71.2X+139320; *p*=0.01), APOB (Y=-20.4X+60875; *p*=0.01), Microsomal Triglyceride Transfer Protein (MTTP; Y=-0.94X+3418; *p*=0.05), TM6SF2 (Y=-0.14X+191.7; *p*=0.003), Lipoprotein lipase (LPL; Y=0.03X+5.69; *p*=0.03), Sterol Regulatory Element-binding Protein-1 (SREBP1; Y=-1.23*X+2850; *p*=0.02) and Diacylglycerol O-acyltransferase 2 (DGAT2; Y=-2.80*X+4375; *p*=0.007) genes, thus, reinforcing the role of SORT1 in lipoprotein turnover, lipid synthesis and dismissal (**Fig.4E-N**). Conversely, PSRC1 mRNA levels more marginally impacted on the expression of these genes (**Fig.S3A-B**), while CELSR2 expression did not impact at all. This data suggests that the alteration of SORT1 expression, induced by presence of the rs599839 variant, is the main driver of the reduced lipid concentrations observed in patients who carry the variant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

Figure 4

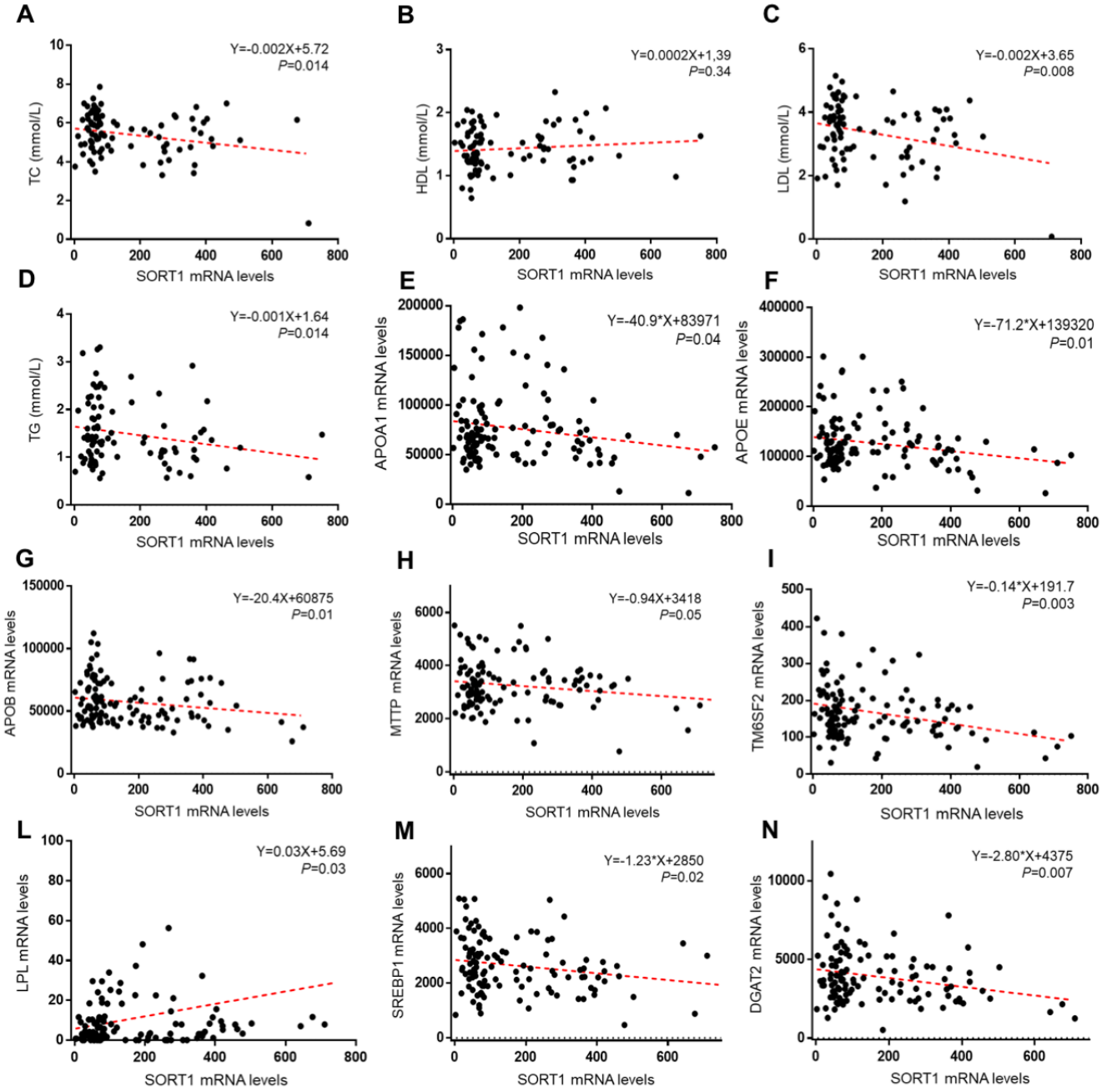

Figure 4. SORT1 mRNA levels correlates with the concentration of circulating lipids and with the expression of genes involved in lipoprotein release and lipid synthesis. Correlation analyses between hepatic SORT1 gene expression evaluated by transcriptome analysis on liver biopsies (n=125) and circulating total cholesterol (TC) (mmol/L) (A), LDL (mmol/L) (B), HDL (mmol/L) (C), and triglycerides (mmol/L) (D). Correlation analyses between hepatic SORT1 gene expression and APOA1 (E), APOE (F), APOB (G), MTTP (H), TM6SF2 (I), LPL (L), SREBP1 (M), DGAT2 (N) mRNA levels evaluated by transcriptome analysis on liver biopsies (n=125).

To further explore the association between the rs599839 variant and HCC risk, we investigated the impact of *PSRC1-SORT1-CELSR2* cluster on cell proliferation. We found that PSRC1 mRNA levels positively correlated with those of Proliferating cell nuclear antigen (PCNA; Y=1.15*X+188.1; p=0.002) and Tumor protein p53 (TP53; Y=0.98*X+196.1;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

p=0.0007) (**Fig.5A-B**). Conversely, SORT1 expression was less strongly associated with that of PCNA and TP53 (**Fig.S3C-D**), whereas the one of CELSR2 did not correlate with the expression of genes involved in cell cycle progression. In sum, the effect of the rs599839 variant on HCC risk in NAFLD patients seems to be mainly related to PSCR1 enhanced expression.

Figure 5

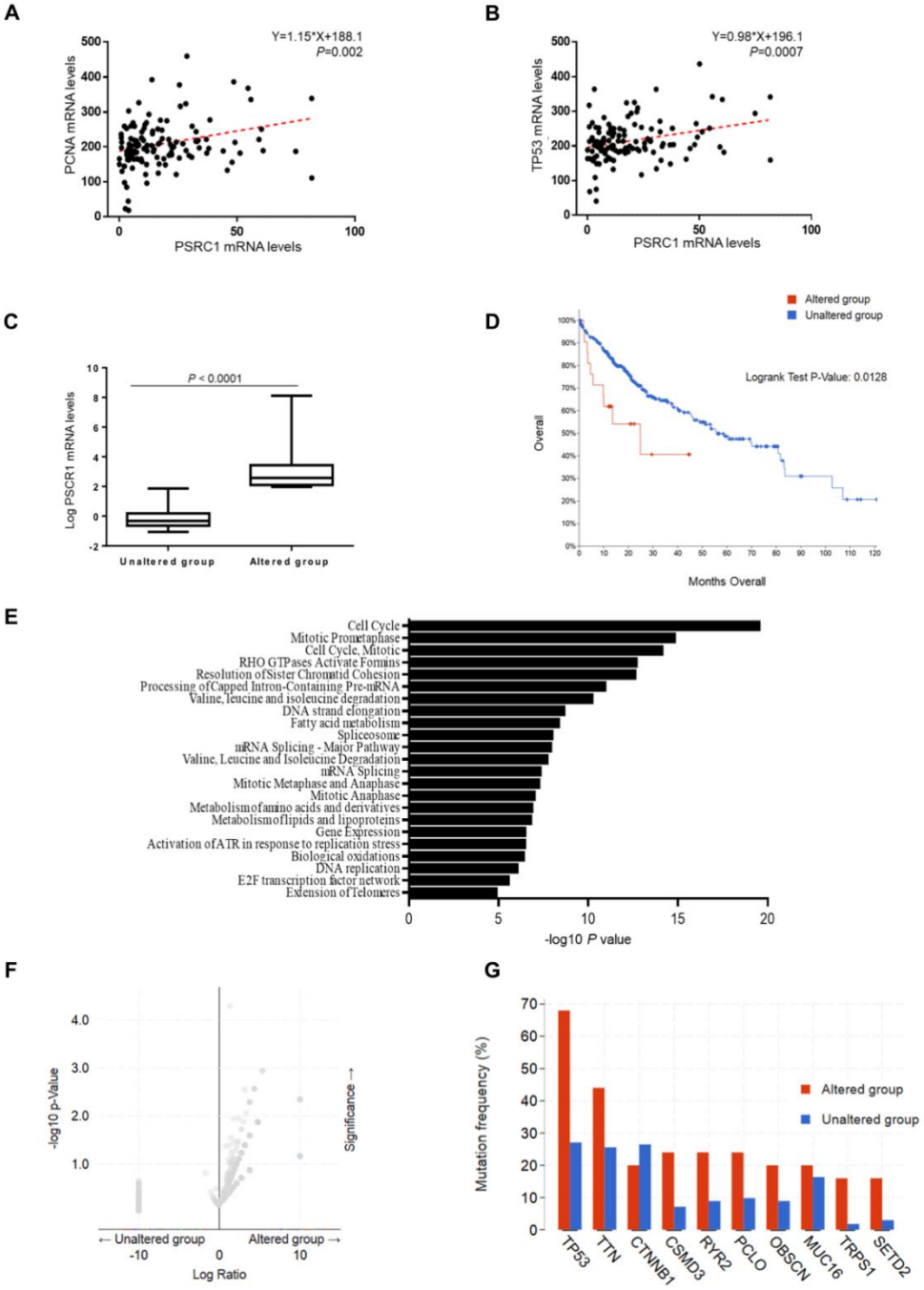

Figure 5. PSRC1 mRNA levels correlate with that of genes involved in cell proliferation and its overexpression reduced survival in HCC patients. Correlation analyses between hepatic PSRC1 gene expression and PCNA (A) and TP53 (B) mRNA levels evaluated by transcriptome analysis on liver biopsies (n=125). PSRC1 mRNA levels were evaluated by transcriptome analysis on HCC samples from TCGA dataset (n=366). Among the 366 HCC samples, 25 patients (7%)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

overexpressed hepatic PSRC1 (altered group) compared to the others (n=341, unaltered group). mRNA levels were represented as log transformed. Boxes span from 25° to 75° percentile, while whiskers indicate the 10° and 90° percentile. * $P< 0.0001$ at two-tailed Student *T*-test (C). Kaplan-Meier survival analysis of patients with PSRC1 overexpressed (n=25; altered group – red line) *vs* the others (n=341; unaltered group – blue line). X axis indicates the months of survival. Logrank Test *P* value: 0.0128 (D). Reactome pathways enriched for 6353 genes co-regulated with PSRC1 expression in 366 samples from TCGA dataset. The statistical significance level (*p*-value) was negative 10-based log trans-formed (E).Volcano plot illustrates the differential frequency distribution of mutations, observed in patients belonging to the altered group (n=25) compared to the unaltered one (n=341), represented as log fold changes (log ratio, on x axis) and the distribution of transformed *P* values (-log10 of P, y axis) (**F**). The frequency of mutations in the 10 genes most enriched in mutations in both groups are stratified according to the presence of PSRC1 overexpression (n=25, altered group in red) or not (n=341, unaltered group in blue) (**G**).

3.6. PSRC1 Overexpression Aggravates HCC in TCGA

To further endorse the role of PSRC1 in HCC development we assessed its hepatic expression in 366 HCC patients enrolled in TCGA dataset. Among them, 25 (7%) displayed an overexpression of hepatic PSRC1 (altered group) compared to the others (n=341, unaltered group) ($p<0.0001$ at *t*-tests; **Fig.5C**). Likewise, independently of ethnicity, gender and age at diagnosis, PSRC1, but not SORT1 or CELSR2 mRNA levels, were associated with tumor stage worsening (S >1; beta: 0.7±0.06; 95% C.I. 0.05-0.27; $p=0.006$), advanced histological grade of liver cancer (G >2; beta: 0.17±0.06; 95% C.I. 0.05-0.29; $p=0.005$) and primary tumor extension (T >1; beta: 0.17±0.05; c.i. 0.06-0.28; $p=0.003$) at multivariate analysis (**Fig.S4A-C** and **Table 8**), thus suggesting that tumors which overexpressed PSRC1 tend to be more undifferentiated, grow rapidly and spread faster to nearby and distal tissues and partly confirming the results obtained in NAFLD-HCC patients who carried the G minor allele. As a consequence, in patients in whom PSRC1 was overexpressed, the average of months of survival was 24.89, while in the others was 55.69. Moreover, the number of deceases in the altered group was 10 (40%), whereas in the unaltered one it was 119 (35%) (Logrank Test $p=0.0128$) (**Fig.5D**).

Table 8. Independent predictors of hepatic PSRC1 mRNA levels in 366 HCC patients from TCGA database.

	PSRC1		
	β	95% CI	*P**
Sex, M	0.04	-0.08-0.31	0.50
Ethnicity; Hispanic	0.08	-0.18-0.35	0.55
Diagnosis Age	-0.0003	-0.01-0.006	0.50
Stage (S) >1	0.7	0.05-0.27	**0.006**
Sex, M	0.06	-0.06-0.18	0.31
Ethnicity; Hispanic	0.10	-0.16-0.36	0.45
Diagnosis Age	-0.001	-0.01-0.07	0.67
Grade (G) >2	0.17	0.05-0.29	**0.005**
Sex, M	0.05	-0.07-0.17	0.41
Ethnicity; Hispanic	0.08	-0.18-0.34	0.54
Diagnosis Age	-0.002	-0.01-0.006	0.53
Tumor size (T) >1	0.17	0.06-0.28	**0.003**

At multivariate generalized linear model adjusted for the confounders shown in the table.

Consistently with the results obtained from Transcriptomic cohort, PSRC1 mRNA levels positively correlated with expression of SORT1 and CELSR2 and with well-established proliferation markers such as PCNA, MKI67, TERT, CDC20, numerous cyclins (CCNs), cyclin dependent kinases (CDK1-4), genes involved in DNA replication (MCM2-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

10, E2F2) and many other oncogenes (**Table S6**). However, we didn't find any correlation between TP53 expression and PSRC1 in TCGA. Conversely, PSRC1 expression negatively correlated with that of genes involved in lipoprotein release and cholesterol synthesis (such as APOB and DGAT2) (**Table S6**). In details, a pathway-enriched analysis of the 6353 genes most significantly co-regulated with PSRC1, confirmed that these genes are mainly involved in cell cycle progression (q= 1.07*e-19 Benjamini-corrected) (**Fig.5E** and **Table S7**).

We next analyzed the frequency rate of mutations in patients who belong to the altered group compared to the unaltered one, as shown in the volcano plot in **Fig.5F**. Among the most mutated genes, we highlighted that *TP53* has the highest rate of mutation in the group of patients with altered expression of PSRC1. Indeed, among patients belonging to the altered group, 17 cases (68%) displayed *TP53* putative driver mutations, including 10 missense and 7 truncating pathogenic mutations, while among the others (unaltered group) only 91 (27%) showed mutation in *TP53* (p=5.11*e-5). Along with *TP53* several other genes implicated in cell proliferation (*CTNNB1* (p=0.32), *CSMD3* (p=0.01), *PCLO* (p=0.04), *MUC16* (p=0.40), *TRPS1* (p=2.8*e-3) and *SETD2* (p=0.01)) or cardiovascular damage (*TTN* (p=0.04), *RYR2* (p=0.028), *OBSCN* (p=0.08)) have been found to be enriched in mutation rate in the altered group (**Fig.5G**).

Finally, to investigate which genes were deregulated in the context of PSRC1 overexpression, we performed a gene expression differential analysis of the whole hepatic transcriptome of patients belonging to the altered group compared to the unaltered one. We found 3794 differentially expressed genes, 2129 being up-regulated and 1665 down-regulated in the altered group of patients (q<0.05, at Benjamini-Hochberg FDR correction). Pathway-enriched analysis confirmed that up-regulated genes were mainly involved in cell proliferation (**Fig.S5A-B** and **Table S8**), whereas down-regulated in metabolic pathways, metabolism of lipids and lipoproteins and in mitochondrial fatty acid β-oxidation (**Fig.S6A-B** and **Table S9**). Overall, these findings may support the hypothesis that the hyper-activation of PSRC1 transcription may promote tumor growth, de-differentiation and invasion and they may explain the loss of significant correlation between PSRC1 and TP53 in patients carrying PSRC1 altered expression, as an escape from the mechanism of tumor suppression.

4. Discussion

In this study, we examined the impact of the rs599839 A>G variant which is located in the genetic cluster of *CELSR2-PSRC1-SORT1* on metabolic phenotypes and liver damage in a large cohort of histologically characterized NAFLD patients, in individuals from UKKBC and in HCC patients from TCGA. We found that the minor G allele was associated with a protection from atherogenic dyslipidemia, carotid plaque formation and hypertension in NAFLD patients, which are at higher risk of cardiovascular comorbidities. Consistently, the presence of the G allele was related to protection against hypercholesterolemia and several cardiovascular outcomes also in the UKKBC, which is representative of the general population.

The rs599839 variant has been previously identified through GWAS and then investigated in epidemiological studies which revealed its strong association with CAD, myocardial infarction, abdominal aortic aneurysm and lipid traits [10,29-33]. Its impact on CAD risk factors has been further validated in the Sikh Diabetes Study [34,35] and in the Indian Atherosclerosis Research Study (IARS) [13]. Furthermore, the association between GG homozygosity and the protection against hypertension has been confirmed in a cohort of 5,460 Japanese individuals [36].

We endorsed the strong impact of the rs599839 variant on circulating lipids through the linkage disequilibrium analysis at the region spanning 50.00 Kb in the 1p13.3 *locus* (chr1: 109,800,000-109,850,000; Human (GRCh37.p13)), using data from 1000 genomes project. Indeed, we revealed that the rs599839 variant is in strong linkage disequilibrium with other 10 common SNPs (rs1277930, rs583104, rs4970836, rs602633, rs7528419,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2021 doi:10.20944/preprints202103.0400.v1

rs629301, rs646776, rs12740374, rs3832016, rs660240), all localized in the genomic region ranging from the 3'UTR of *CELR2*, the intergenic region and the *PSRC1* 3'UTR oriented in opposite direction. All these SNPs have been found to be correlated with reduced TC levels, with the same direction and strength of the rs599839 variant. Thus, the impact of the entire 1p13.3 *locus* on TC modulation which has been observed in previous studies may be translated even in NAFLD patients.

Concerning liver damage, the rs599839 variant did not have a significant impact on hepatic fat accumulation, lobular inflammation, ballooning or fibrosis. The lack of any significant association between liver failure or hepatic fibrosis and cirrhosis and the rs599839 variant has been supported by UKBBC data. Notwithstanding, the rs599839 mutation was related to higher risk to develop HCC in NAFLD patients through a mechanism which seems to be independent of fibrogenic processes and the presence of cirrhosis.

Therefore, in attempt to decipher the role of this *locus* in lipid handling, cardiovascular protection and HCC predisposition, we demonstrated that patients who carry the rs599839 G allele showed higher mRNA levels of *SORT1*, *CELSR2* and *PSRC1*. It has also been previously assessed that rs599839 variant strongly correlated with SORT1, PSRC1 and CELSR2 transcript levels in human liver [11,37,38], and we further reinforced this data with transcriptional studies and network analysis by STRING. In addition, the analysis of the expression of quantitative trait loci (eQTLs), confirmed the correlation between the rs599839 variant and hepato-specific PSRC1, SORT1 and CELSR2 expressions (**Table S10**). However, it remains unclear how the rs599839 genetic variation affects the expression of genes at this *locus*. Indeed, we did not find any transcription factor binding site in the proximity of the rs599839 polymorphism. Nonetheless, given the high linkage disequilibrium with the other 10 SNPs, we could speculate that they can impact on essential regulatory regions thus providing gene expression alterations. For instance, the rs12740374 SNP has been predicted to alter a binding site for CCAAT/enhancer-binding protein (C/EBP) transcription factors, resulting in a significantly increased SORT1 hepatic expression [17]. Notably, we demonstrated that the enhanced hepatic expression of SORT1 mainly affected circulating lipid profiles, lipoprotein turnover and release, while PSRC1 expression most strongly impacted on genes implicated in cell proliferation and survival thus possibly explaining the opposite effect of the rs599838 variation which protects against cardiovascular complications and as one predisposes to HCC, respectively. The rs599839 variant induces the expression of these genes whose effect is even more strong in homozygous *status*. Conversely, CELR2 expression neither impact on lipid levels nor proliferation.

It has been reported that Sortilin 1, encoded by *SORT1* gene, is directly involved in lipid metabolism and lipoprotein uptake [17,39]. It is a multi-ligand receptor mainly expressed in hepatocytes and macrophages, where it mediates the trafficking of diverse endogenous or exogenous proteins between the Trans-Golgi network and lysosomes, endosomes and plasma membranes [17,40]. As sorting receptor, Sortilin 1 may regulate the hepatic expression of various genes, including LPL and APOE and lipid-related ones [41-43]. Specifically, hepatic Sortilin 1 translocates Apolipoproteins, mainly APOB to lysosomes for the autophagy-related degradation thus limiting VLDL/LDL formation and secretion and enhancing in turn their clearance. Thus, it mediates the reduction of circulating VLDL, TC and TG levels [44,45]. Similarly, Sortilin 1 mediates the uptake of native LDL in macrophages for subsequent lysosomal hydrolysis [46]. We could speculate that the improved lipid profile associated with the lack of hepatic fat accumulation in NAFLD patients carrying the rs599839 genetic variation may be explained by the enhanced LDL lysosomal degradation in hepatocytes and macrophages, due to the increased Sortilin 1 expression.

Conversely, *PSRC1* gene encodes a proline-rich protein, that play a crucial role in mitosis by recruiting and regulating microtubule depolymerases (i.e. KIF2A) which destabilizes microtubules. *PSCR1* may act as oncogene by different mechanisms: it enhances β-catenin activation and cyclins production by binding to adenomatous polyposis coli 2

 doi:10.20944/preprints202103.0400.v1

(APC2) and inhibits p53-binding protein 2 (ASPP2) [47]. It is targeted for regulation by the tumor suppressor protein p53, that is a well characterized transcription factor that mediates DNA repair, cell cycle arrest and apoptosis [16,48-50]. In particular, Hsieh and colleagues demonstrated that P53 suppresses the expression of both human PSRC1 mRNA and protein levels, specifically binding to a motif in *PSRC1* 5′ region [16]

Interestingly, in TCGA dataset the overexpression of PSRC1 has been strictly correlated with poor prognosis, tumor stage, advanced grade, and increased size of liver cancers, further reinforcing what we have observed in our cohort of NAFLD-HCC patients. Indeed, PSRC1 mRNA levels positively correlated with those of genes involved in cell proliferation and cell cycle progression, confirming our results from transcriptomic analyses. According to these findings, patients who overexpress PSRC1 showed an enhanced loss-of-function mutation rate in *TP53* gene, possibly explaining the absence of correlation between PSRC1 and TP53 expressions in HCC samples, as an escape from the TP53-mediated mechanisms of tumor suppression. Likewise, PSRC1 overexpression has been previously detected in tumor liver tissues and in hepatoma cells, where it is associated with HCC recurrence after resection [16]. Nonetheless, the mechanisms through which HCC tissues overexpress PSRC1, independently of the genetic background, remain to be fully elucidated and *TP53* mutations may partially explain this effect.

Overall, the evidence reported here points out PSRC1 as a possible novel biomarker or molecular target for HCC diagnosis and therapy. Indeed, this study, for the first time, highlighted a possible link between the rs599839 variant and HCC development in a large series of histologically characterized NAFLD patients. However, there are some limitations, in UKKBC there is not individual data available about cancer history, thus precluding a formal evaluation of the interaction between *PSRC1* and the risk of HCC in this population. Moreover, the real proof-of-concept of the interaction between PSRC1 and TP53 is lacking.

In conclusion, the rs599839 variant is associated with protection against atherogenic dyslipidemia and with increased HCC risk in NAFLD individuals. Several lines of evidence indicate that the mechanisms underlying these associations may involve the effect of the rs599839 variant on SORT1 and PSRC1 expressions, although prospective and functional studies are required.

List of abbreviations: APOB; Apolipoprotein B; BMI, body mass index; CAD, coronary artery disease; CCN, cyclin; CDK, cyclin dependent kinases; CELSR2, Cadherin EGF LAG Seven-Pass G-Type Receptor 2; DGAT2, Diacylglycerol O-acyltransferase 2; FDR, False Discovery Rate; IGT, impaired glucose tolerance; GWAS, genome wide association study; HDL, high-density lipoprotein; HCC, hepatocellular carcinoma; IMT, intima-media thickness; LDL, low-density lipoprotein; LPL, Lipoprotein lipase; MBOAT7, Membrane Bound O-Acyltransferase Domain Containing 7; MTTP, Microsomal Triglyceride Transfer Protein; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis; PCNA, Proliferating cell nuclear antigen; PNPLA3, Patatin-like Phospholipase domain-containing 3; PSRC1, Proline And Serine Rich Coiled-Coil 1; SREBP1, Sterol Regulatory Element-binding Protein-1; SORT1, sortilin-1; TCGA-LIHC, The Cancer Genome Atlas-Liver Hepatocellular Carcinoma; TM6SF2, Transmembrane 6 Superfamily member gene 2; TP53, Tumor protein p53; T2D, type 2 diabetes mellitus; UKBBC, UK Biobank Cohort.

Novelty and Impact: Dyslipidemia is hallmark of nonalcoholic fatty liver disease (NAFLD) and the rs599839 variant in the *CELSR2-PSRC1-SORT1* cluster, has been associated with a protection against cardiovascular events. Here, we revealed a novel link between the rs599839 variant and hepatocellular carcinoma (HCC) whose onset in the context of NAFLD is rapidly increasing. The rs599839 variant disentangles the risks of HCC and cardiovascular abnormalities by modulating SORT1 and PSRC1 expressions. The latter emerged as potential modifier in liver carcinogenesis.

Authors' contribution: Marica Meroni: genotyping, manuscript drafting, data analysis and interpretation; Miriam Longo: genotyping and data analysis; Erika Paolini: data and samples collection; Emilia De Caro: data and samples collection; Anna Alisi: data and samples collection; Luca Miele: data and samples collection; Giorgio Soardo: data and samples collection; Anna Ludovica

Fracanzani: data interpretation, manuscript revision; Paola Dongiovanni: study design, manuscript drafting, data analysis and interpretation. All authors approved the final draft of the manuscript.

Funding: The study was supported by Ricerca Corrente Fondazione IRCCS Cà Granda (PD, AF); Ricerca Finalizzata Ministero della Salute RF-2013-02358319 (AF) and Ricerca Finalizzata Ministero della Salute GR-2019-12370172 (MM).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Younossi Z, Henry L. Contribution of Alcoholic and Nonalcoholic Fatty Liver Disease to the Burden of Liver-Related Morbidity and Mortality. *Gastroenterology* 2016; **150**: 1778-1785.
2. Eslam M, Sanyal AJ, George J. MAFLD: A Consensus-Driven Proposed Nomenclature for Metabolic Associated Fatty Liver Disease. *Gastroenterology* 2020; **158**: 1999-2014.e1991.
3. Younossi ZM, Koenig AB, Abdelatif D, *et al.* Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology (Baltimore, Md)* 2016; **64**: 73-84.
4. Wong RJ, Aguilar M, Cheung R, *et al.* Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. *Gastroenterology* 2015; **148**: 547-555.
5. Dongiovanni P, Meroni M. miRNA Signature in NAFLD: A Turning Point for a Non-Invasive Diagnosis. 2018; **19**.
6. Meroni M, Dongiovanni P, Longo M, *et al.* Mboat7 down-regulation by hyper-insulinemia induces fat accumulation in hepatocytes. *EBioMedicine* 2020; **52**: 102658.
7. Meroni M, Longo M, Rustichelli A, *et al.* Nutrition and Genetics in NAFLD: The Perfect Binomium. 2020; **21**.
8. Meroni M, Longo M. mir-101-3p Downregulation Promotes Fibrogenesis by Facilitating Hepatic Stellate Cell Transdifferentiation During Insulin Resistance. 2019; **11**.
9. Dongiovanni P, Meroni M, Baselli G, *et al.* PCSK7 gene variation bridges atherogenic dyslipidemia with hepatic inflammation in NAFLD patients. 2019; **60**: 1144-1153.
10. Samani NJ, Braund PS, Erdmann J, *et al.* The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. *Journal of molecular medicine (Berlin, Germany)* 2008; **86**: 1233-1241.
11. Musunuru K, Strong A, Frank-Kamenetsky M, *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010; **466**: 714-719.
12. Zhou YJ, Hong SC, Yang Q, *et al.* Association of variants in CELSR2-PSRC1-SORT1 with risk of serum lipid traits, coronary artery disease and ischemic stroke. *International journal of clinical and experimental pathology* 2015; **8**: 9543-9551.
13. Arvind P, Nair J, Jambunathan S, *et al.* CELSR2-PSRC1-SORT1 gene expression and association with coronary artery disease and plasma lipid levels in an Asian Indian cohort. *Journal of cardiology* 2014; **64**: 339-346.
14. Vincent JB, Skaug J, Scherer SW. The human homologue of flamingo, EGFL2, encodes a brain-expressed large cadherin-like protein with epidermal growth factor-like domains, and maps to chromosome 1p13.3-p21.1. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2000; **7**: 233-235.
15. Hsieh PC, Chang JC, Sun WT, *et al.* p53 downstream target DDA3 is a novel microtubule-associated protein that interacts with end-binding protein EB3 and activates beta-catenin pathway. *Oncogene* 2007; **26**: 4928-4940.
16. Hsieh WJ, Hsieh SC, Chen CC, *et al.* Human DDA3 is an oncoprotein down-regulated by p53 and DNA damage. *Biochemical and biophysical research communications* 2008; **369**: 567-572.
17. Kjolby M, Andersen OM, Breiderhoff T, *et al.* Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export. *Cell metabolism* 2010; **12**: 213-223.
18. Dongiovanni P, Petta S, Maglio C, *et al.* Transmembrane 6 superfamily member 2 gene variant disentangles nonalcoholic steatohepatitis from cardiovascular disease. *Hepatology* 2015; **61**: 506-514.
19. Mancina RM, Dongiovanni P, Petta S, *et al.* The MBOAT7-TMC4 Variant rs641738 Increases Risk of Nonalcoholic Fatty Liver Disease in Individuals of European Descent. *Gastroenterology* 2016; **150**: 1219-1230.e1216.
20. Donati B, Dongiovanni P, Romeo S, *et al.* MBOAT7 rs641738 variant and hepatocellular carcinoma in non-cirrhotic individuals *Sci Rep* 2017: in press.
21. Donati B, Pietrelli A, Pingitore P, *et al.* Telomerase reverse transcriptase germline mutations and hepatocellular carcinoma in patients with nonalcoholic fatty liver disease. *Cancer medicine* 2017; **6**: 1930-1940.
22. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *Journal of hepatology* 2012; **56**: 908-943.
23. Kleiner DE, Brunt EM, Van Natta M, *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* 2005; **41**: 1313-1321.
24. Dongiovanni P, Petta S, Mannisto V, *et al.* Statin use and non-alcoholic steatohepatitis in at risk individuals. *Journal of hepatology* 2015; **63**: 705-712.
25. Abecasis GR, Auton A, Brooks LD, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56-65.

26. Benjamini Y, Drai D, Elmer G, *et al.* Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* 2001; **125**: 279-284.
27. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* 2013; **6**: pl1.
28. Cerami E, Gao J, Dogrusoz U, *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* 2012; **2**: 401-404.
29. Kathiresan S, Voight BF, Purcell S, *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics* 2009; **41**: 334-341.
30. Kleber ME, Renner W, Grammer TB, *et al.* Association of the single nucleotide polymorphism rs599839 in the vicinity of the sortilin 1 gene with LDL and triglyceride metabolism, coronary heart disease and myocardial infarction. The Ludwigshafen Risk and Cardiovascular Health Study. *Atherosclerosis* 2010; **209**: 492-497.
31. Saade S, Cazier JB, Ghassibe-Sabbagh M, *et al.* Large scale association analysis identifies three susceptibility loci for coronary artery disease. *PloS one* 2011; **6**: e29427.
32. Jones GT, Bown MJ, Gretarsdottir S, *et al.* A sequence variant associated with sortilin-1 (SORT1) on 1p13.3 is independently associated with abdominal aortic aneurysm. *Human molecular genetics* 2013; **22**: 2941-2947.
33. Matsuoka R, Abe S, Tokoro F, *et al.* Association of six genetic variants with myocardial infarction. *International journal of molecular medicine* 2015; **35**: 1451-1459.
34. Zhou L, Ding H, Zhang X, *et al.* Genetic variants at newly identified lipid loci are associated with coronary heart disease in a Chinese Han population. *PloS one* 2011; **6**: e27481.
35. Braun TR, Been LF, Singhal A, *et al.* A replication study of GWAS-derived lipid genes in Asian Indians: the chromosomal region 11q23.3 harbors loci contributing to triglycerides. *PloS one* 2012; **7**: e37056.
36. Fujimaki T, Oguri M, Horibe H, *et al.* Association of a transcription factor 21 gene polymorphism with hypertension. *Biomedical reports* 2015; **3**: 118-122.
37. Kathiresan S, Melander O, Guiducci C, *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics* 2008; **40**: 189-197.
38. Innocenti F, Cooper GM, Stanaway IB, *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics* 2011; **7**: e1002078.
39. Conlon DM. Role of sortilin in lipid metabolism. *Current opinion in lipidology* 2019; **30**: 198-204.
40. Calandra S, Tarugi P, Speedy HE, *et al.* Mechanisms and genetic determinants regulating sterol absorption, circulating LDL levels, and sterol elimination: implications for classification and disease risk. *Journal of lipid research* 2011; **52**: 1885-1926.
41. Zhong LY, Cayabyab FS, Tang CK, *et al.* Sortilin: A novel regulator in lipid metabolism and atherogenesis. *Clinica chimica acta; international journal of clinical chemistry* 2016; **460**: 11-17.
42. Tauris J, Ellgaard L, Jacobsen C, *et al.* The carboxy-terminal domain of the receptor-associated protein binds to the Vps10p domain of sortilin. *FEBS letters* 1998; **429**: 27-30.
43. Linsel-Nitschke P, Heeren J, Aherrahrou Z, *et al.* Genetic variation at chromosome 1p13.3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease. *Atherosclerosis* 2010; **208**: 183-189.
44. Li J, Matye DJ, Li T. Insulin resistance induces posttranslational hepatic sortilin 1 degradation in mice. *The Journal of biological chemistry* 2015; **290**: 11526-11536.
45. Gigante B, Leander K, Vikstrom M, *et al.* Chromosome 1p13 genetic variants antagonize the risk of myocardial infarction associated with high ApoB serum levels. *BMC cardiovascular disorders* 2012; **12**: 90.
46. Patel KM, Strong A, Tohyama J, *et al.* Macrophage sortilin promotes LDL uptake, foam cell formation, and atherosclerosis. *Circulation research* 2015; **116**: 789-796.
47. Sun WT, Hsieh PC, Chiang ML, *et al.* p53 target DDA3 binds ASPP2 and inhibits its stimulation on p53-mediated BAX activation. *Biochemical and biophysical research communications* 2008; **376**: 395-398.
48. Hsieh SC, Lo PK, Wang FF. Mouse DDA3 gene is a direct transcriptional target of p53 and p73. *Oncogene* 2002; **21**: 3050-3057.
49. Cai BH, Chao CF, Huang HC, *et al.* Roles of p53 Family Structure and Function in Non-Canonical Response Element Binding and Activation. 2019; **20**.
50. Lo PK, Chen JY, Lo WC, *et al.* Identification of a novel mouse p53 target gene DDA3. *Oncogene* 1999; **18**: 7765-7774.