Università degli Studi di Milano
Scuola di Dottorato in Medicina Clinica e Sperimentale
XXXIII ciclo

"Impact of whole exome sequencing (WES) on the clinical management of patients with advanced nonalcoholic fatty liver (NAFL)"

Tutor: Prof. Luca Valenti
PhD student: Dr. Serena Pelusi

Abstract

Non-alcoholic fatty liver disease (NAFLD) is a potentially progressive disorder, possibily leading to cirrhosis and hepatocellular carcinoma (HCC).

Both acquired and genetic risk factors play an important role in disease progression. In this respect, the presence of metabolic comorbidities such as obesity and type two diabetes mellitus has been associated to a more severe phenotype, while genetic factors modulating hepatic lipid metabolism such as the variants in *PNPLA3* (patatin-like phospholipase domain-containing protein 3), *TM6SF2* (transmembrane 6 superfamily member 2), *MBOAT7* (membrane bound O-acyltranferase domain containing 7), *GCKR* (glucokinase regulatory protein) and *APOB* (apolipoprotein B) have been demostrated to influence disease predisposition.

Secondly, a consistent fraction of patients with liver disorders are diagnosed at advanced stage and in approximately one third of cases it is not possible to establish an etiological diagnosis (cryptogenic cirrhosis).

By exploiting Whole Exome Sequencing (WES) technology, in this work we aimed to:

1) identify the prevalence of known pathogenic mutations in candidate genes mutated in genetic liver diseases and hereditary cancer syndromes in patients with HCC and cirrhosis due to NAFLD or cryptogenic disease, to compare it to that of healthy individuals and to evaluate in the aforementioned genes the burden of rare or novel mutations of unknown significance, that determine an alteration of protein sequence, and are therefore likely pathogenic;

2) test the functional and clinical significance of the mutations identified, by bioinformatics algorithms and *in vitro* assays;

3) identify novel genetic risk factors predisposing to progressive NAFLD;

4) implement a genetic risk score (GRS) aiming to improve the stratification of the risk of progressive NAFLD;

5) test the impact of WES on the clinical management of six patients with cryptogenic liver disease.

Briefly, we detected an enrichment in pathogenic ($p = 0.024$), and likely pathogenic variants ($p = 1.9*10-6$), particularly in *APOB* ($p = 0.047$). *APOB* variants were associated with lower circulating triglycerides and higher HDL cholesterol ($p < 0.01$). A genetic risk score predicted NAFLD-HCC (OR 4.96, 3.29–7.55; $p = 5.1*10-16$), outperforming the diagnostic accuracy of common genetic risk variants, and of clinical risk factors ($p < 0.05$).

Furthermore, in one third of the six patients with cryptogenic liver disease and aggressive phenotype it was possible to identify a probable genetic disorder expleining the phenotype.

In conclusion, rare pathogenic variants in genes involved in liver disease and cancer predisposition are associated with NAFLD-HCC development.

Genetic study by WES can represent a precious instrument for risk stratification and allow a Precision Medicine approach in the context of liver disease aimed at the targeted treatment of the underlying cause of the disease.

Index

1.Introduction

1.1 Epidemiology and natural history of nonalcoholic fatty liver disease (NAFLD)

NAFLD (also known as metabolic dysfunction associated fatty liver disease, or MAFLD, when associated with dysmetabolism) [1] is defined as the hepatic accumulation of neutral lipids greater than 5% of liver weight and affects 16-38% of the general population worldwide, which is not accounted for by at risk alcohol intake or other liver conditions [2]. It is a leading cause of cirrhosis, the main risk factor for HCC, which is due to progressive

fibrosis and represents *per se* a pre-cancerous condition. Although the growing rates of HCC may be due to the increased number of individuals with advanced fibrosis, NAFLD-HCC frequently develops without overt cirrhosis suggesting that steatosis directly promotes hepatic carcinogenesis [2-6]. However, progression of liver disease to cirrhosis and HCC is more frequent in the subgroup of patients who develop non-alcoholic steatohepatitis (NASH), a condition characterized by active inflammation and fibrosis [7].

Established risk factors for disease progression in NAFLD include older age and presence of features of the metabolic syndrome, such as obesity, severe insulin resistance, and hypertension. Consistently, NAFLD-HCC patients are most commonly older males, with type 2 diabetes (T2D) and meeting criteria for at least one feature of the metabolic syndrome. HCCs arising in patients with features of the metabolic syndrome are larger, more differentiated and occur more frequently in the absence of significant fibrosis than those in patients with chronic viral hepatitis [8].

In addition, it has been estimated up to 30% of HCCs in industrialized countries develop in patients with cryptogenic cirrhosis, a condition which is retained to stem in the majority of cases from burnt-out NASH [9, 10].

Several epidemiological studies established an association between overweight and obesity, that are considered the major determinants of insulin resistance and NAFLD, and higher risk to develop HCC (17% and 89% respectively compared to normal weight individuals). Furthermore, the relationship linking obesity to HCC risk seems stronger in males than in females [11, 12]. Even diabetes was independently correlated to HCC onset in large epidemiological studies, where it was found that among men affected by T2D the risk of HCC was doubled [13, 14].

On the one hand, these data suggest that obesity and diabetes are the major epidemiological determinant of HCC incidence in Western countries. On the other hand, patients with progressive NAFLD are mostly unaware of being affected by a progressive form of liver

disease. Therefore, due to the very high prevalence of NAFLD, occurrence in patients without advanced fibrosis, and lack of diseases awareness, classic screening strategies for the detection early HCC are ineffective [15].

Current guidelines advise that HCC surveillance is recommended in patients with NAFLD and cirrhosis and should be considered in those with advanced liver fibrosis [16].

However, no reliable biomarker is yet available to stratify HCC risk in patients without severe fibrosis, accounting for a large fraction of HCC cases in individuals with dysmetabolism. [16, 17]. The high prevalence of NAFLD and the evidence that HCC frequently arises in individuals unaware of their risk make classical HCC surveillance strategies impractical, resulting in delayed diagnosis and unfavourable prognosis [18].

In this respect, family history and genetic factors play an important role in the pathogenesis of progressive NAFLD and of HCC [19].

1.2 Role of genetics in NAFLD progression

Genetic factors have been shown to influence disease progression in NAFLD, and family history remains the main risk factor for HCC development [19]. The common genetic polymorphism rs738409 C>G encoding for the I148M variant in Patatin-like phospholipase domain-containing protein 3 (*PNPLA3* or adiponutrin) has been established as the main common genetic determinant of hepatic fat content and of progressive NAFLD [20-24]. The mechanism is related to accumulation of the mutated protein [25], which interferes with lipid droplets remodeling in hepatocytes [26, 27], and with retinol release by hepatic stellate cells [28, 29]. The *PNPLA3* variant predicts HCC development in European patients with NAFLD [30] and also in individuals affected by other liver diseases associated with steatosis, namely alcoholic liver disease (ALD) and chronic hepatitis C (CHC) [31]. This evidence suggests that

this genetic risk factors may be helpful to select high-risk individuals for screening [30-32], but it has a low sensitivity to be used as single prognostic biomarker [33].

The rs58542926 E167K variant in Transmembrane 6 superfamily member 2 (*TM6SF2*) also predisposes to progressive NAFLD by altering the secretion of very low-density lipoproteins [34-36]. but its direct role in HCC predisposition is disputed [35, 37].

More recently, it has been found that the rs641738 C>T sequence variant in the Membrane bound O-acyltranferase domain containing 7/ Transmembrane channel like 4 (*MBOAT7/TMC4*) locus, involved in phospholipids remodeling, predisposes to cirrhosis development in individuals with excessive alcohol intake [38], and to the development and the progression of NAFLD in individuals of European descent [39].

Another gene involved in lipid metabolism whose variants have been demonstrated to predispose to NAFLD, hepatic fibrosis, and HCC in the presence of environmental triggers is glucokinase regulator (*GCKR*) [40-43].

Conversely, a splice variant in 17β-hydroxysteroid dehydrogenase type 13 (*HSD17B13*) prevents severe fibrosis and HCC development [44].

Consistently, genetic data indicate that NAFLD is commonly observed in patients with telomeropathies, suggesting that steatosis may either be a consequence of hepatocellular senescence, as also observed in animal models, or a trigger for liver disease progression [45, 46]. Indeed, loss of function germline mutations in the telomerase reverse transcriptase (*TERT)* can predispose to a spectrum of familial liver diseases characterized by steatosis [45] and possible evolution to cirrhosis and HCC [47, 48]. In keeping, we previously reported the occurrence of NAFLD-HCC in a patient with a rare germline *TERT* loss-of-function mutation [49]. Furthermore, it has also been reported that rare mutations inducing Mendelian diseases due to severe derangements in the function of encoded proteins may predispose to NAFLD-HCC. Indeed, mutations in Apolipoprotein B (*APOB)* may explain some familial cases

through predisposition towards development of severe steatosis caused by hepatocellular retention of lipids [42, 50].

Besides, recent studies have shown the possible involment of the autophagy pathway in the pathogenesis of NAFLD. This process in fact mediates the breakdown of intracellular lipids in hepatocytes and therefore a decrease in hepatic autophagy may be associated to the development of liver steatosis. Subsequent studies have demonstrated additional critical functions for autophagy in hepatocytes and other hepatic cell types such as stellate cells that regulate insulin sensitivity, hepatocellular injury, fibrosis and carcinogenesis.

In this respect, ATG7 (autophagy related protein 7) is one of the various factors involved in the assembly of the autophagosome. In *Atg7* knockout mice the inhibition of autophagy has been shown to lead to marked increases in hepatic fat content.

Adenoviral overexpression of *Atg7* instead successfully reversed hepatic steatosis in genetic obesity models and viral expression of the autophagy gene *Atg7* reduced ER stress and improved insulin sensitivity [51-53].

We previously developed a robust polygenic risk score (PRS) of hepatic fat content (termed PRS-HFC) and showed that the impact of genetic risk variants on fibrosis is proportional to that on hepatic fat, consistent with hepatic fat accumulation being a driver of liver disease [54]. Recently, Stender et al. confirmed that an unweighted PRS based on *PNPLA3-TM6SF2-HSD17B13* variants predicted cirrhosis and HCC in Europeans [55]

In a very recent work, we also hypothesized that liver fat promotes HCC in individuals with NAFLD and dysmetabolism. Mendelian randomization is considered the most appropriate epidemiological tool to assess causality when randomized controlled trials are not feasible. Therefore, we examined the impact of the previously developed PRS-HFC based on well-characterized risk variants that can be evaluated in the clinic on HCC in at-risk individuals and in the general population. We also performed a further adjustment for *HSD17B13* (termed PRS-5). Next, we identified PRS thresholds able to identify with good specificity a subset of

individuals with NAFLD and dysmetabolism at high risk for HCC. Finally, we showed that PRS predicted HCC irrespective of severe liver fibrosis [56].

1.3 Next-generation Sequencing (NGS) approach for the identification of pathogenic genetic variants

Over the past years, the advent of next-generation sequencing (NGS) has rebuilt the meaning of DNA sequencing by processing millions of DNA fragment in parallel resulting a very low cost per base [57]. The enormous volume of data cheaply produced by NGS technology has been the drawing power for shifting from automated Sanger sequencing and changing the way of thinking the genome research and clinical genetics [58].

The powerful and flexible nature of the NGS technology led to the development of a broad range of applications allowing researchers to ask biological questions that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease.

The whole NGS workflow consists in a three-step process divided into template preparation (or sample preparation), sequencing and imaging. The Illumina sequencing strategy is based on the concept of sequencing by synthesis to produce millions of 'short-reads' (from 36 to 86 nucleotides of length) simultaneously [59]. The first step is the creation of a DNA library by and adding universal adapters by ligation to sample DNA fragments. Afterward, the process involves using a microfluidic cluster station to add these fragments to the surface of a glass flowcell thank to the hybridization with complementary oligos to the surface. The hybridization of the library fragment on the flowcell is followed by a subsequent incubation, called cluster generation, that amplifies the fragments in a discrete area or 'cluster' thanks to a PCR reaction [60] directly on the flowcell surface. The flowcell is placed within the sequencer where each cluster is supplied with polymerase and four differentially labeled

fluorescent nucleotides that have their 3'-OH chemically inactivated to ensure that only a single base is incorporated per cycle [59]. Every cycle is composed by (1) base incorporation, (2) imaging step to discriminate the incorporated nucleotide at each cluster by laser light excitation and (3) a chemical step that removes the fluorescent group and release the 3' end for the next base incorporation cycle.

The images, corresponding to the sequence of each cluster, are translated into text file as FastQ format [61] containing the reads produced in the sequencing run. Every run produces from 50 to 60 million reads.

As NGS has become popular, whole-genome sequencing (WGS) and whole exome sequencing (WES) have proven to be valuable methods for the discovery of the genetic causes of both rare and complex diseases [62]. Nevertheless, the great amount of data and the number of simple nucleotide variations (SNVs), including single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs) divert the problem to the computational analysis and data management, which are the real bottleneck of the entire NGS workflow [63]. On average, WES identifies from 12,000 to 20,000 variants in coding regions [64, 65], of which ≈90% are found in publicly available databases [66]. The bioinformatic analysis process for NGS data is divided in different steps involving the alignment, variation discovery and annotation (primary analysis), as well as the use of tools for gene prioritization and mutation pathogenicity prediction (secondary analysis). The final aim is to select potentially driver mutations related to a given disease/phenotype.

The very first step, after completing the sequencing run, is the evaluation of the read quality produced by the high-throughput sequencer.

After the quality control step, the reads are ready to be mapped against the reference genome [67]. In case of human samples, the reference genome assembly is available in two versions: the one curated by the University of California at Santa Cruz (UCSC), which is hosting the ENCODE data [68], and the one available from the Genome Reference Consortium (GRC),

which focuses on creating reference assemblies (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc).

This specific task is the hardest step in light of computational resources, considering the whole data analysis process. Therefore, the alignment of millions of reads back to a reference genome required an improvement and a development of new specific softwares which are able to rapidly solve this issue [69]. All these tools perform a fast and accurate mapping of the reads against the reference genome by producing a file in SAM/BAM [70] format.

After read alignment and once the SAM/BAM file is produced the next step is the identification of those position that differ from the reference sequence and could be recognized as variations (variant call).

As for the alignment, the variant calls are stored into a text file, which follows the standard of 1000Genomes data called VCF (variant call format). Basically, it is a column-based file in which nine columns contain the specific information for each variant call including chromosomal location, reference and alternative allele, variation quality and the depth of the read supporting the alternative allele compared to the number of read supporting the reference allele.

The next step to evaluate the significance and the possible function of each variation called is the annotation step.

Once the VCF file containing the quality-passed call has been produced, the annotation step gives, to variants, additional information regarding the possibility to predict their functional impact [68]. Generally, the annotation is performed by linking the variation data to existing public databases. Each variant can be classified, based on the position relative to coding sequences as intergenic, intronic, splice-site or coding variant. For single-nucleotide coding variants, it can be predicted if the variation corresponds to a synonymous, missense or nonsense mutation relative to the protein coding sequence.

This basic annotation can be improved by using several resources such as public variants database (such as 1000Genomes Project) and pathogenicity prediction tools.

Recent genome/exome sequencing studies [65, 71, 72] have demonstrated that disease-causing genes are mainly associated to missense (protein coding variations) mutations [73]. Approximately 50 to 75% of variants can be removed by focusing only on non-synonymous (protein-altering) changes [74, 75]. A filtering step is an efficient method to select a list of possible candidate variations. The possible candidate variants are selected on the assumption that (1) the causing mutations alter the protein sequence, (2) should be extremely rare within the population (3) the disease-causing variants should be present in the affected samples and (4) every affected individual should carry the candidate mutations [76].

Application of NGS technologies to Mendelian and complex diseases has proven to be an effective alternative to single-gene tests in research for establishing a new genetic basis of diseases [62, 65, 77, 78]. In particular the development of capture technologies for target sequencing, instead of whole-genome sequencing, dramatically reduced the sample preparation time and the overall costs of the experiment.

The application of this technology in clinical laboratories would be the perfect combination to unlock complex disease pictures and to provide anticipatory guidance and prognosis where other diagnostic testing has not been definitive, also as far as liver diseases are concerned [79].

In the setting of a Precision Medicine (PM) approach, defined as a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products being tailored to the individual patient, NGS could definitely be employed for selecting appropriate and optimal therapies based on the context of a patient's genetic content [80].

2. Aims

- Aim 1: genetic diagnosis; First aim was to identify the prevalence of known pathogenic mutations in candidate genes mutated in genetic liver diseases and hereditary

cancer syndromes in patients with HCC and cirrhosis due to NAFLD or cryptogenic disease, to compare it to that of healthy individuals and to evaluate in the aforementioned genes the burden of rare or novel mutations of unknown significance, that determine an alteration of protein sequence, and are therefore likely pathogenic.

- Aim 2: functional significance. Second aim was to test and verify the functional and clinical significance of the mutations identified, by bioinformatic algorithms and *in vitro* assays, and to communicate results to public clinical databases.
- Aim 3: identification of novel genetic risk factors predisposing to NAFLD progression.
- Aim 4: genetic risk score (GRS) development. Fourth aim was to examine whether the evaluation of common and rare germline genetic variants may be clinically helpful in the stratification of NAFLD-HCC risk by developing a weighted genetic risk score for this condition.
- Aim 5: WES clinical application to diagnosis. Further aim was to test the impact of WES on the clinical management of six patients with cryptogenic liver disease.

3. Material and methods

3.1 Study cohort and design

The evaluated cohorts and the study flow chart are presented in Figure 1.

The discovery NAFLD-HCC cohort included 72 Italian patients and 70 UK patients, who were enrolled between January 2010 and 2016. All were of Caucasian ancestry.

The diagnosis of HCC was based on the EASL-EORTC clinical practice guidelines for management of hepatocellular carcinoma [81]. Secondary causes of steatosis were excluded based on history, including at risk alcohol intake (≥30 g/day in M/F) and the use of drugs known to precipitate steatosis. Viral and autoimmune hepatitis, hereditary hemochromatosis, Wilson's disease, overt alpha-1-antitrypsin deficiency and present or previous infection with

HBV (HBsAg) and HCV were ruled out using standard clinical and laboratory evaluation as well as liver biopsy features.

The study protocol was conformed to the ethical guidelines of the 1975 Declaration of Helsinki, was approved by the Ethical Committee of the Fondazione IRCCS Ca' Granda of Milan, as well as by the other involved Institutions, and was performed according to the recommendations of the hospitals involved. Informed consent was obtained from each patient.

Fifty-nine patients with advanced fibrosis due to NAFLD (histological stage F3–F4 or clinically overt cirrhosis) recruited at the Italian institutions during the same period were used as controls. A local ethnically matched control group of comparable sex distribution including 50 Italian healthy blood donors without clinical and biochemical evidence of liver disease, NAFLD, metabolic abnormalities and no alcohol abuse, and the 404 non-Finnish European (NFE) healthy individuals included in the 1000 Genomes database (http://www.internationalgenome.org), for whom complete exome data were publicly available were used as further controls (including 91 Italian and 107 UK individuals).

The clinical features of individuals included in the study are presented in Table 1.

There were four sequential steps to the study (Fig. 1). The first step consisted in whole exome sequencing, variant analysis, identification and prioritization. The second addressed the possibility of enrichment in already known pathogenic variants in candidate genes in NAFLD-HCC cases vs. controls, and identified the most mutated genes and the diagnostic yield for Mendelian monogenic disorders. The third step encompassed the identification of rare variants predicted to alter protein function in the same candidate genes that might be associated with disease predisposition. Finally, a genetic risk score (GRS) for NAFLD-HCC was developed and assessed for its diagnostic accuracy.

Figure 1. Study design. (a) Study cohorts composition and enrolment criteria. (b) Study flow-chart. NAFLD: non-alcoholic fatty liver disease. HCC: hepatocellular carcinoma.

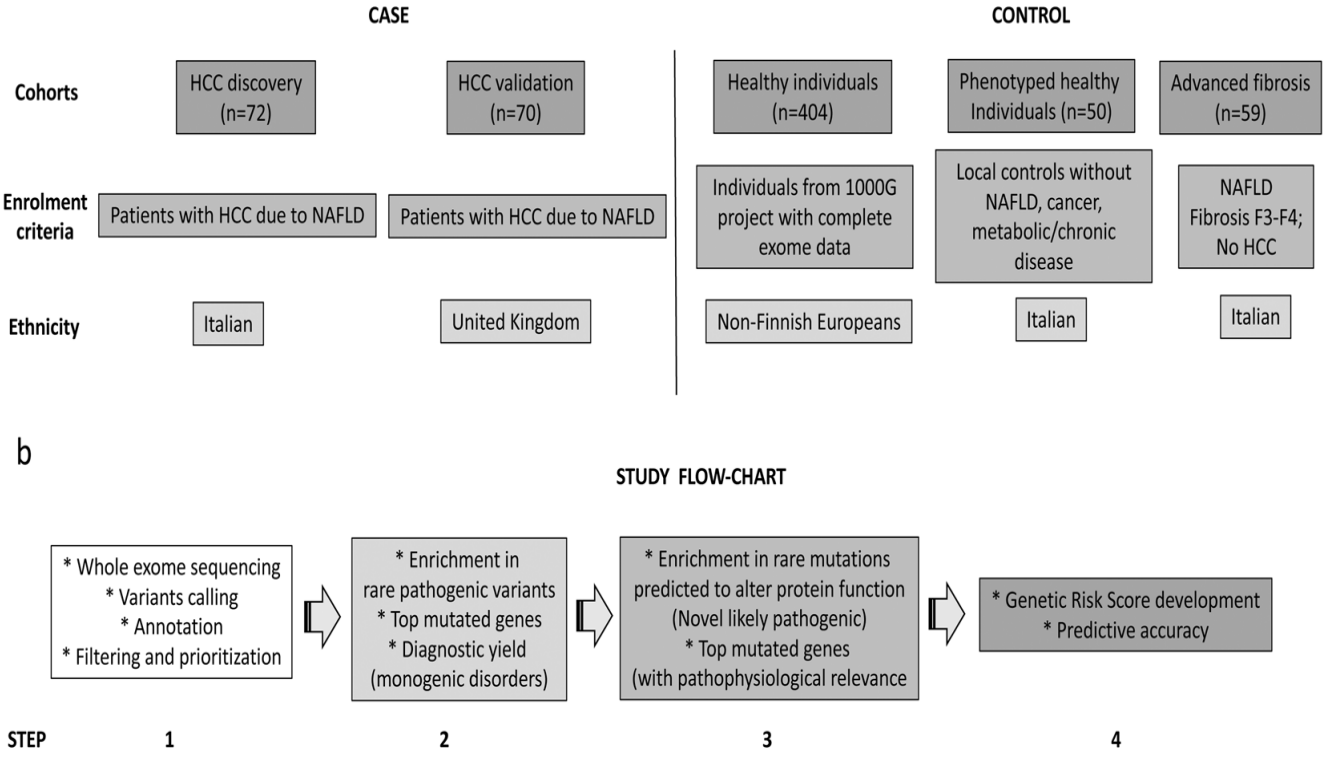


Table 1. Clinical and genetic features of 251 individuals who underwent whole exome sequencing for evaluation of germline variants in candidate genes involved in liver disease and cancer predisposition. BMI: body mass index; HCC hepatocellular carcinoma; PNPLA3: patatin-like phospholipase domain-containing protein 3; TM6SF2: transmembrane 6 superfamily member 2; MBOAT7: membrane bound O-acyltranferase domain containing 7; GCKR: glucokinase regulatory protein. Data were compared by univariate generalized linear models.

	HCC discovery (n=72)	HCC replication (n=70)	p value (discovery vs. replication)	Advanced fibrosis (n=59)	Healthy individuals (n=50)	p value (HCC vs. no-HCC)
Age, years	68±9	74±7	<0.0001	59±10	49±12	<0.0001
Sex, F	17 (24)	10 (20)	0.15	21 (36)	17 (34)	0.048
BMI, Kg/m^2	29.8±5.8 (n=58)	32.7±8.2 (n=45)	0.12	31.3±4.9 (n=44)	24.6±2.5 (n=50)	<0.0001
Type 2 Diabetes, yes	43 (61)	36 (62)	0.72	33 (56)	0	<0.0001
PNPLA3 I148M						
I/I	18 (25)	14 (20)		13 (22)	28 (56)	0.13
I/M	30 (42)	32 (46)	0.20	26 (44)	21 (42)	
M/M	24 (33)	24 (34)		20 (34)	1 (2)	
TM6SF2, E167K						
E/E	57 (79)	53 (76)		47 (80)	38 (76)	0.59
E/K	14 (20)	14 (20)	0.84	12 (20)	12 (24)	
K/K	1 (1)	3 (4)		0	0	
MBOAT7, rs641738 C>T						
C/C	18 (25)	24 (34)		20 (34)	19 (38)	0.32
C/T	31 (43)	38 (54)	0.032	26 (44)	23 (46)	
T/T	23 (32)	8 (12)		13 (22)	8 (16)	

3.2 Whole exome sequencing, variants identification, annotation and priotization

The WES sequencing and analytical pipeline is presented in Figure 2.

Briefly, DNA was extracted from peripheral blood mononuclear cells, and quantified by a Qubit 2.0 analyzer using the Qubit dsDNA BR Assay Kit (Thermo-Fisher Scientific, Waltham, MA, USA). Samples purity was evaluated using a Nanodrop 1000 spectrophotometer (Thermo-Fisher, Waltham, MA, USA) and integrity was assessed by gel

electrophoresis.

DNA libraries were enriched for exome sequencing by the SureSelect Human All Exon v5 kit (Agilent, Cernusco sul Naviglio, Milan, Italy). Sequencing was subsequently performed on the HiSeq4000 platform (Illumina, city). Raw reads quality control was performed using FastQC software (Brabaham bioinformatics, Cambridge, UK). Reads mapping on human GRCh37 genome was performed using MEM algorithm of Burrows Wheeler Aligner (BWA) version 0.7.1017. Reads with low quality alignments and duplicate reads were filtered out using Samtools18 to generate high quality bam files. Mapping quality control was performed using Picard-tools (http://broadinstitute.github.io/picard) and Bedtools19. Sequencing mean depth was of 73x, and no samples exhibit a mean depth lower than 50x (Supplementary Fig. 2 panels a,b). Sequencing resulted in a good target coverage: almost all samples exhibited more than 90% coverage of the target at 20x depth.

Variant calling was performed following GATK best practices. Briefly, indel local realignment, base quality recalibration and variants calling (Haplotypecaller algorithm) were performed using GATK version 3.3.021.

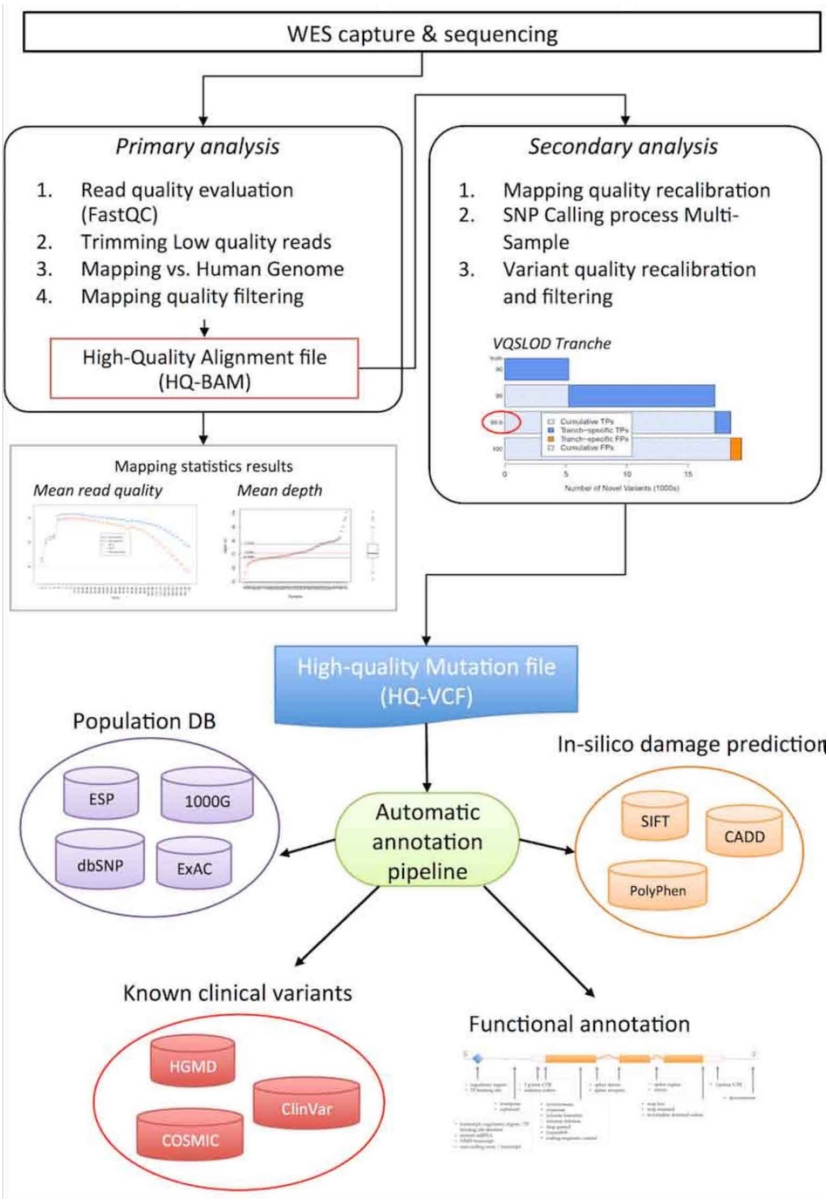
GVCF joint and variants filtering were performed using variant quality score recalibration (VQSR) method.

Variants quality score log-odds (VQSLOD) above 99% tranche were considered true positives. To avoid the possibility of calling somatic variants due to the presence of circulating tumor DNA, variants present in <20% of total reads were discarded. Indel left normalization was performed using BCFtools software22. Variants annotation was performed using both variant effect predictor (VEP)23 and ANNOVAR24 tools.

Variants filtering was performed using VCFtools25 to exclude variants over VQSLOD threshold and variants which were called in less than 95% of samples. All intronic and synonymous variants according to VEP prediction were excluded from the analyses. Multidimensional scaling of identity-by-state distances analysis was conducted on

EPIDEMIC study samples exploiting SNPRelate R Bioconductor package. In the EPIDEMIC project samples, the first component of variability was explained by the geographic origin of the patients (Italy vs. UK). Whole exome sequencing reads from 1000 genomes project phase 3 (1000 G) non-Finnish Europeans (NFE; 404 samples) were processed using the same pipeline described for the EPIDEMIC samples.

Figure 2: Whole exome sequencing (WES) analytical pipeline. VQSLOD: variants quality score log-odds, HQ-BAM: high quality - binary alignment map, SNP: single nucleotide polymorphism, DB: database, VCF: virtual card file, 1000G: 1000 Genomes project, ExAC: exome aggregation consortium, CADD: combined annotation dependent depletion, ESP: exome sequencing project.



3.3 Candidate genes selection and classification of variants

Candidate genes were selected according to the literature updated at January 2016, among those whose variants were robustly linked with cancer predisposition syndromes [82] or mutated in HCC [83], or predisposing to hereditary liver diseases [84], or involved in predisposition to telomeres diseases [45], or in iron and lipid metabolism and NAFLD [7]. The complete list of candidate genes in which we identified variants and their classification is presented here:

Gene abbreviations, related genetic diseases, and inheritance pattern (AD: autosomal dominant, AR: autosomal recessive, XL: X-linked, somatic: usually associated with somatic mutations):
FAH: fumarylacetoacetate hydrolase (tyrosinemia type I, AR); CFTR: cystic fibrosis transmembrane regulator (cystic fibrosis, AR); TERT: telomerase reverse transcriptase (dyskeratosis congenita disease spectrum, AD); ATP7B: copper-transporting ATPase 2 (Wilson's disease, AR); ABCB4: ATP-binding cassette subfamily B member 4 (MDR3; progressive familial intrahepatic cholestasis type 3, intrahepatic cholestasis of pregnancy, low phospholipid-associated cholelithiasis, AR); CP: ceruloplasmin (aceruloplasminemia); APOB: apolipoprotein B (hypobetalipoproteinemia, AD); MUTYH: MYH glycosylase (familial adenomatous polyposis, AR); SYNE2: spectrin repeat containing nuclear envelope protein 2 (Emery-Dreifuss muscular dystrophy 5, AD & somatic HCC); SERPINA1: Serpina 1 (alpha1-antitrypsin deficiency, AR); ATM: mutated in ataxia-telangiectasia (ataxia-telangiectasia, AR, breast and other cancers); HNF1A: hepatic nuclear factor 1A (Maturity-onset diabetes of the young, type 3, familial hepatic adenomas, renal cell carcinoma, AD); G6PC: glucose-6-phosphatase, catalytic subunit (glycogen storage disease, type Ia – von Gierke's disease, hepatic adenomas, AR); GBE1: 1,4-alpha-glucan branching enzyme 1 (glycogen storage disease type IV, AR); CDKN2A: cyclin dependent kinase inhibitor 2A (several cancers, somatic & AD); FANCA: Fanconi anemia complementation group A (Fanconi anemia, AR); RPS6KA3: ribosomal protein S6 kinase A3 (Coffin-Lowry syndrome, XL & somatic HCC); SDHC: succinate dehydrogenase complex subunit C (hereditary paraganglioma & pheochromocytoma, and gastrointestinal stromal tumor, AD); CHEK2: checkpoint kinase 2 (Li-Fraumeni syndrome, AD); FANCL: Fanconi anemia complementation group L (Fanconi anemia, AR); ALDOB: aldolase (hereditary fructose intolerance, AR); TF: transferrin (atransferrinemia, AR), ASL: aginosuccinate lyase (arginosuccinic acidura, AR), EGF: epidermal growth factor (hypomagnesemia, type 4, AR), BRCA2: breast cancer type 2 susceptibility protein (breast and other cancers, AD), SQSTM1: Sequestosome-1 (front-temporal degeneration, AD), TINF2: TERF1 interacting nuclear factor 2 (dyskeratosis congenital, AR).

Known common risk variants: PNPLA3: patatin-like phospholipase domain-containing 3; TM6SF2: transmembrane 6 superfamily member 2; MBOAT7: membrane-bound O-acyl transferase 7.

For Step 2 (enrichment in rare pathogenic variants in candidate genes and diagnostic rate), variants reported as "likely pathogenic" in the Clinvar (https://www.ncbi.nlm.nih.gov) database, located in candidate genes, and with a minor allele frequency (MAF) <0.05 in 1000 G NFE, ExAC databases and in local healthy controls, were selected.

For Step 3 (enrichment in rare variants predicted to alter protein function, novel likely pathogenic), we used stringent criteria, that is selection of variants determining an alteration of protein sequence (missense, nonsense, splice sites), located in candidate genes, and with a MAF <0.001 in ExAC NFE, MAF <0.005 in the EPIDEMIC project samples, and a CADD Phred >10 (Top 10% of damaging variants) [85].

3.4 Statistical analysis

For descriptive statistics, continuous variables were shown as mean and standard deviation or median and interquartile range for highly skewed biological variables. Variables with skewed distributions were logarithmically or inverse normally transformed before analyses. All genetic analyses were calculated by using an additive model.

Fisher's Exact test, multivariate or univariate generalized linear models were used when appropriate. Models were adjusted for clinically relevant covariates, as specified in the Results section. Gene enrichment in rare variants was assessed using the cohort allelic sum test (CAST) approach [86, 87]. The association between the frequency of variants in genes significantly enriched in NAFLD-HCC vs. healthy controls was next validated against the cumulative frequency observed in NFE individuals included in the ExAC project (n = 33,370) by Fisher's exact test, adjusted for the number of comparisons.

3.5 Genetic risk score development

A NAFLD-HCC risk score was developed as previously described [43, 88].

The GRS for HCC was calculated by regressing the number of pathogenic/likely pathogenic variant collapsed at the level of single candidate genes and common genetic risk factors for in *PNPLA3*, *TM6SF2* and *MBOAT7* against the presence of HCC. To internally validate the GRS, β coefficients were adjusted using the Jack-knife resampling method. The diagnostic accuracy of different models for NAFLD-HCC prediction was compared by two-sided Venkatraman test [89]. The population attributable risk (PAR) of GRS for NAFLD-HCC was estimated as previously described for case-control studies [90].

GRS gene functions were explored by pathway enrichment analysis exploiting Ingenuity Pathway Analysis software (Qiagen, Valencia, USA) with default parameters.

Protein features were obtained from Uniprot database (www.uniprot.org) coding variants of interest genes were mapped into reported protein domains and regions. Variants enrichment in protein domains was evaluated by CAST Burden test approach. Lollipop diagrams were generated using Mutation Mapper software (http://www.cbioportal.org/mutation_mapper.jsp).

Statistical analyses were carried out using R statistical analysis software version 3.3.2 (http://www.R-project.org/). P values < 0.05 were considered statistically significant.

3.6 Cellular models

HepaRG cells were grown in Williams Medium with 10 % FBS and exposed to 50 µg/µL oleic acid. Adipo Red assay was performed after transient transfection of 5 nM of wild type *ATG7* pcDNA 3.1 vector. Relative Fluorescence unit (RFU) data are shown as average (± SD) of quadruplicates of four independent experiments. The p value was calculated by the Mann-Whitney test.

3.7 WES clinical application

We considered a cohort of six patients presenting to our hospital with cryptogenic liver disease. Clinical features of these subjects are presented in Table 2.

Table 2: Clinical features of the six patients with cryptogenic liver disease analysed by WES approach. *LDLRAP1:* Low-density lipoprotein receptor adapter protein 1; *HPS1:* Hermasky-Pludak syndrome protein 1; *DTNBP1:* Dystrobrevin Binding Protein 1: *ALDOB*: Aldolase B

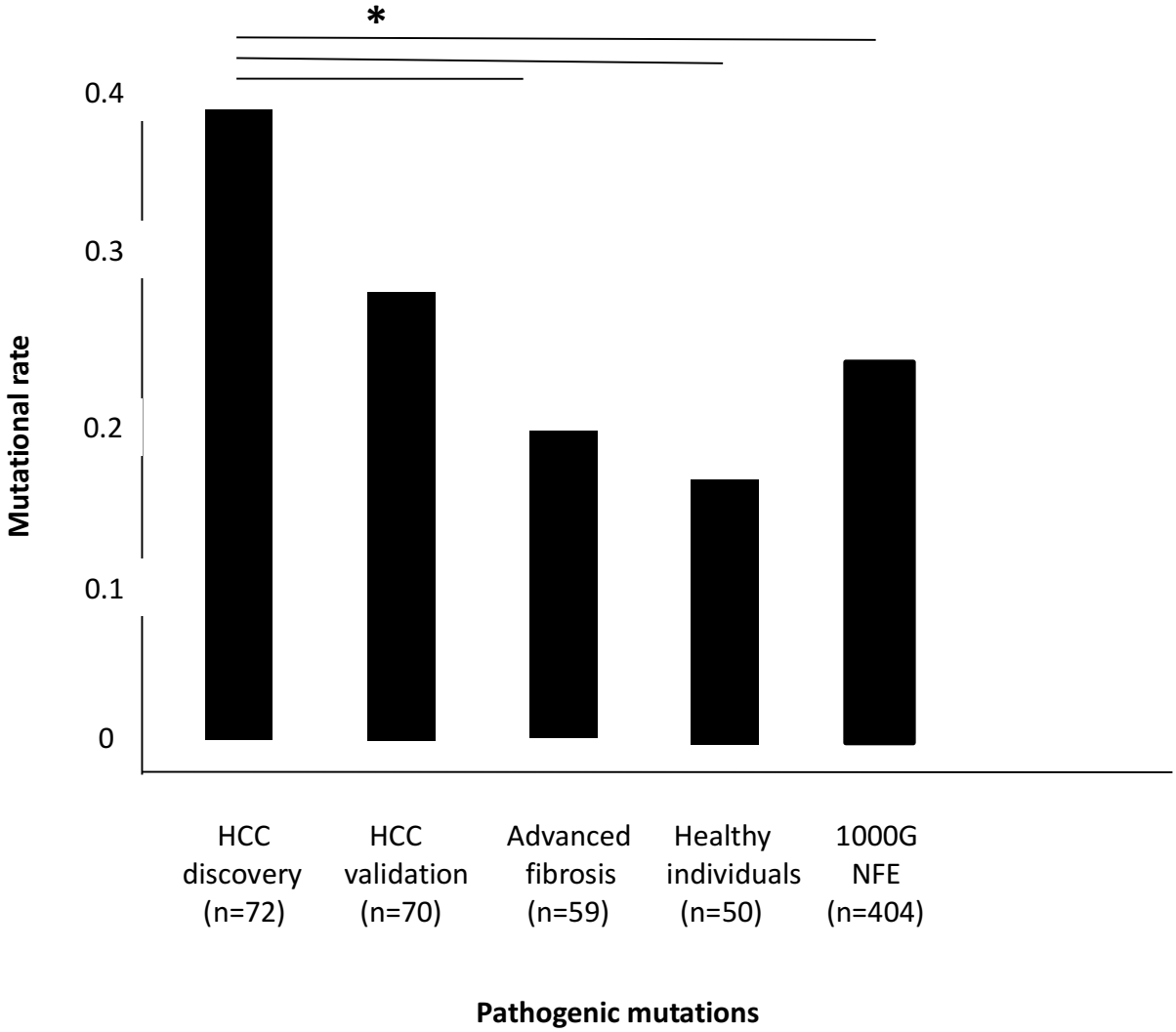
Patient	**Sex**	**Age (years)**	**Clinical presentation**	**Family history**	**Main genetic variants identified**	**Type of variant**	**Disorder**
Patient 1	M	27	Decompensated cirrhosis, hepatic adenomatosis, severe hypercholesterolemia		*LDLRAP1* (p.Gln136*)	Nonsense mutation	Familial Hypercholesterolemia type 4.
Patient 2	M	41	Bone marrow failure, hepatic fibrosis, pulmonary fibrosis, albinism	Interstitial lung disease (father)	*HPS1* (p.Met325Trpfs*6) *DTNBP1* (p.Glu259fs)	Frameshift mutation Frameshift mutation	Hermansky-Pludak syndrome
Patient 3	M	70	Cirrhosis, fatty liver, non obese, hepatocellular carcinoma (HCC)				
Patient 4	M	68	Cirrhosis, fatty liver, portal hypertension, diabetes mellitus	HCC (mother)	*ALDOB* (p.Ala150Pro)	Missense mutation	Hereditary Fructosuria
Patient 5	M	59	Severe fatty liver, lean subject	HCC (father)			
Patient 6	M	32	Severe fatty liver, lean subject, advanced fibrosis				

4.Results

4.1 Pathogenic variants in candidate genes are enriched in NAFLD-HCC

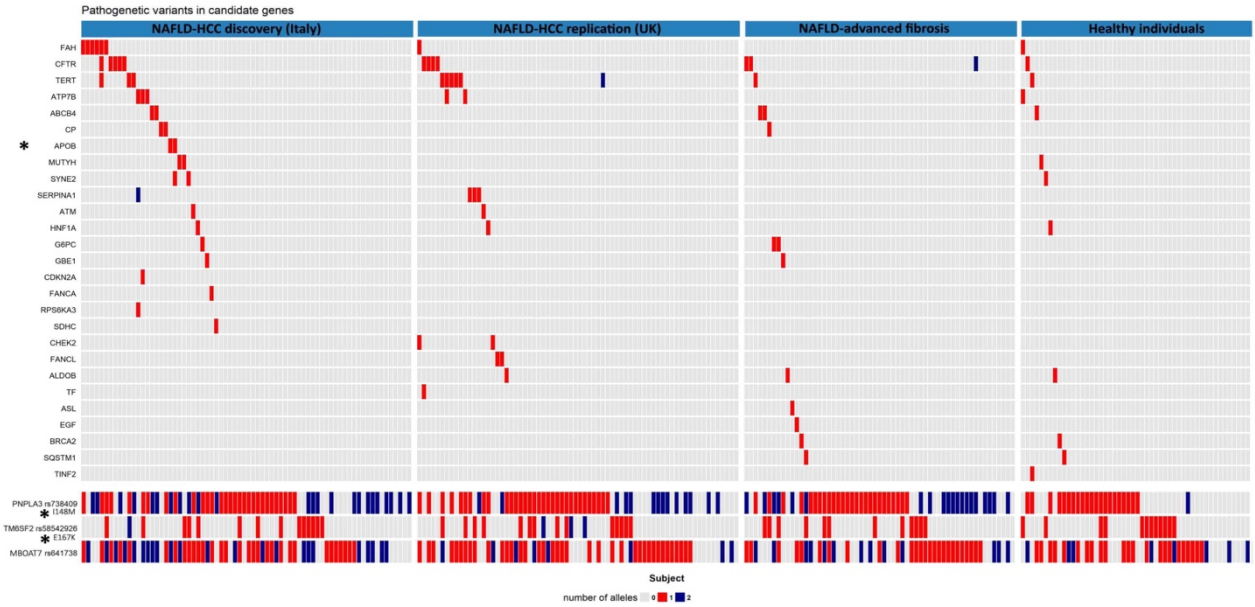
We first examined pathogenic variants in HCC cases and controls. We identified 68 variants previously linked to pathological phenotypes, which met the inclusion criteria. There was an enrichment in the number pathogenic variants in candidate genes per individual in NAFLD-HCC patients (which was significant in the discovery and in the overall cohort) as compared to patients with advanced fibrosis, healthy individuals and the 1000G cohort (Figure 3; OR 1.4, 95% c.i. 1.1-infinite, p=0.024). These data suggest that rare pathogenic variants in selected genes involved in liver disease and cancer predisposition may contribute to NAFLD-HCC development.

Figure 3. Enrichment in pathogenic variants in patients with NAFLD-HCC. Frequency of pathogenic variants (mutational rate %: sum of mutated/total alleles) in NAFLD-HCC cases vs. controls. *$p < 0.05$; **$p < 0.01$; ***$p < 0.005$ by Fisher's exact test.



A comutation plot reporting genes affected by pathogenic variants in the newly characterized EPIDEMIC cohort (excluding 1000G), as well as common variants previously associated with NAFLD-HCC is reported in Figure 4.

Figure 4. Genes enriched in pathogenic variants. Comutation plot showing the distribution of rare pathogenic variants (upper panel), as well as of common genetic variants (bottom panel) predisposing to hepatic fat accumulation and NAFLD-HCC in the 251 individuals of the EPIDEMIC project. Genes significantly enriched in variants in cases vs. controls are marked by and asterisk (*by Fisher's exact test). See page 19 for genes abbreviation list.



Among the single genes, we found a significant enrichment of variants in the *APOB* gene predisposing to familial hypobetalipoproteinemia (two variants in cases and none in controls, p=0.047). Furthermore, we confirmed a strong association with the common *PNPLA3* I148M variant (OR 2.49, 95% c.i. 1.89-3.30), and detected an association with the *TM6SF2* E167K variant (OR 1.72, 95% c.i. 1.10-1.24) regulating hepatic lipid compartmentalization. The rs641738 *MBOAT7* variant was associated with NAFLD-HCC in the discovery (OR 1.49, 95% c.i. 1.03-2.15, p=0.031), but not in the validation cohort (OR 0.81, 95% c.i. 0.56-1.19, p=NS).

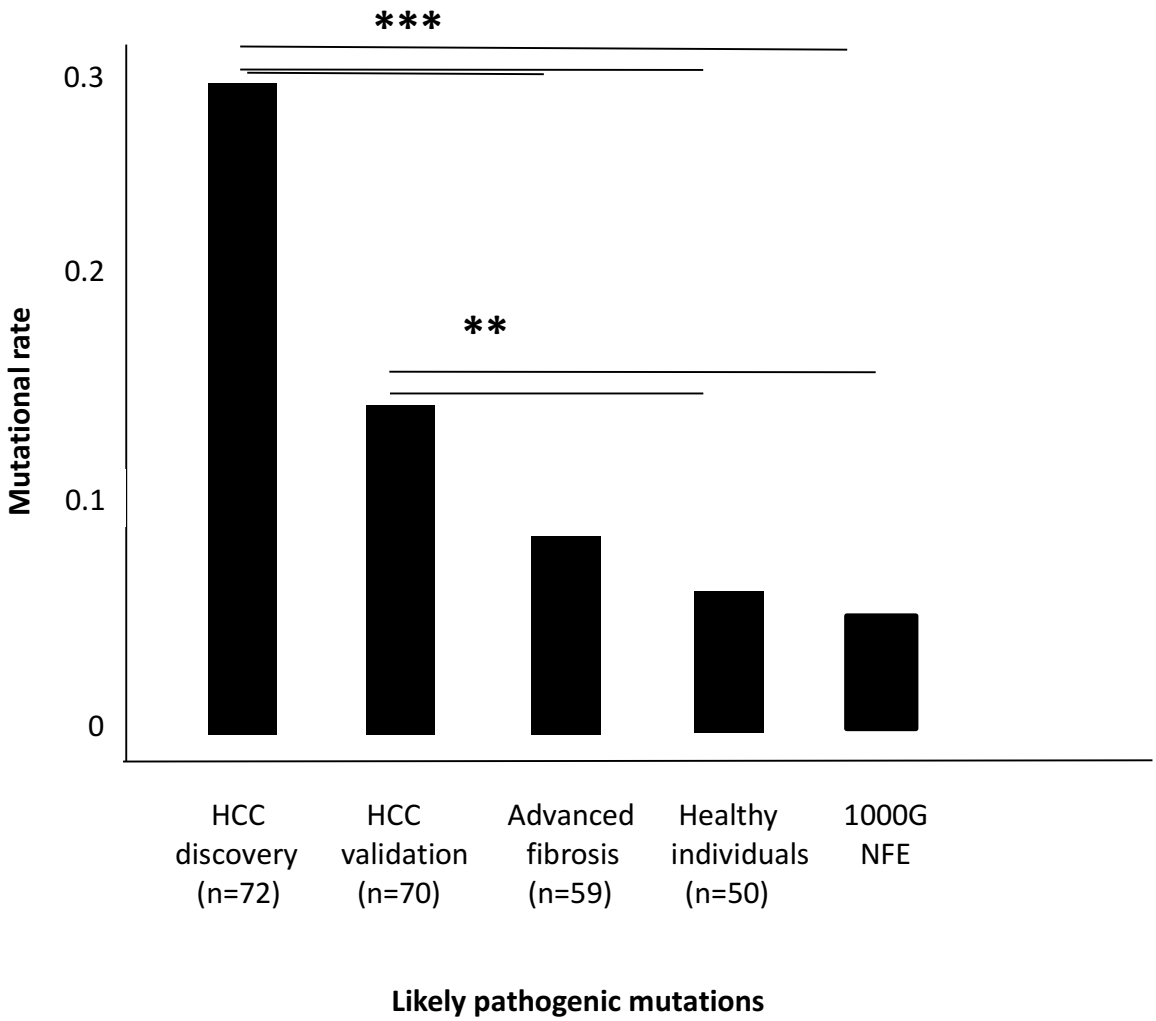
Supporting the causal role of the identified variants in the predisposition to NAFLD-HCC, the number of pathogenic variants carried by NAFLD-HCC patients was associated with younger age at presentation (estimate -3.3±1.1, p=0.006). Notably, the mutational rate was higher in patients not homozygous for the *PNPLA3* I148M variant, suggesting that these individuals need a higher burden of rare genetic mutation to compensate for the lack of the main disease driver.

After resequencing of candidate genes, 19/142 (13.4%) NAFLD-HCC patients vs. 3/59 (5.1%) NAFLD with advanced fibrosis, 3/50 (6.0%) local controls, and 28/404 (6.9%) NFE individuals from 1000G could receive a diagnosis of Mendelian disease predisposing to advanced liver disease or cancer (HCC: 19/142, 13.4%, vs. no-HCC: 34/513, 4.7%; OR 3.15, 95% c.i. 1.57-5.93, p=0.0005).

4.2 Role of rare variants likely determining an alteration in protein activity

We observed an enrichment in rare pathogenic variants in NAFLD-HCC cases vs. controls (Figure 5; OR 3.5 95% c.i. 2.2-inf, $p=1.9*10^{-6}$).

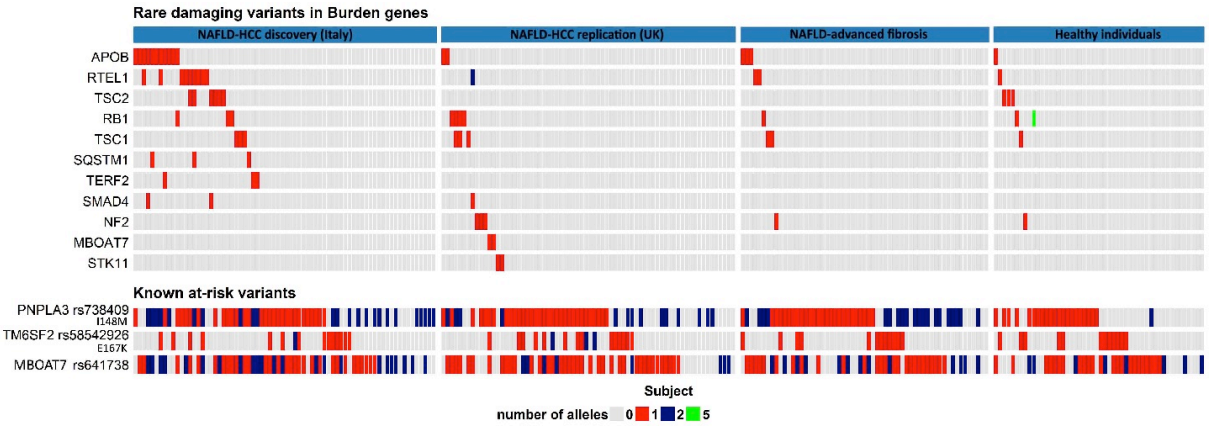
Figure 5. Enrichment in pathogenic variants in patients with NAFLD-HCC. Frequency of likely pathogenic variants (rare variants with high likelihood of altering protein activity) in NAFLD-HCC cases vs. controls. *p < 0.05; **p < 0.01; ***p < 0.005 by Fisher's exact test.



In this case, the mutational rate was higher in patients not homozygous for the PNPLA3 I148M variant. Overall, the pattern of enrichment was different in the Italian and UK cohorts. Either in the overall series or in national cohorts, we found a significant enrichment in variants in Telomerase complex genes (*RTEL1*, *TERF2*), DNA and oxidative damage response (*RB1*), and we also highlighted genes involved in regulation of cell growth and proliferation (*STK11*, *TSC1*, *TSC2*, *NF2*, *SMAD4*). Interestingly, we confirmed genes

involved in regulation of hepatic lipid metabolism, including *APOB* and *MBOAT7* as enriched also in rare likely pathogenic variants, and we detected an enrichment in variants of *SQSTM1*. A comutation plot reporting genes significantly enriched in likely pathogenic variants in the EPIDEMIC cohorts is reported in Figure 6.

Figure 6. Comutation plot showing the distribution of rare likely pathogenic variants (upper panel) in genes significantly enriched in either single cohorts or the overall cohort, as well as of common genetic mutations (bottom panel) predisposing to hepatic fat accumulation and NAFLD-HCC in the 251 individuals genotyped within the EPIDEMIC project. See page 19 for genes abbreviation list.



Interestingly, in the UK cohort one patient carrying *MBOAT7* rare variant was also carrier of the *MBOAT7* common rs641738 variant and had cirrhosis, while the other patient carrying *MBOAT7* rare variant was negative for the common rs641738 variant and did not have advanced fibrosis.

In order to check the efficacy of the criteria adopted (frequency and predicted impact) for the identification of likely pathogenic variants, we assessed whether variants in *APOB* (pathogenic/likely pathogenic), which are associated with a clear phenotype that can be assessed by common biochemical tests, influence circulating lipid levels. Results are shown in Table 3. Supporting the validity of our selection algorithm, in patients for whom data were available carriage of *APOB* variants was associated with 46% higher HDL cholesterol and 44% lower triglycerides (p=0.008 and p=0.001, respectively), consistent with a hypobetalipoproteinemia phenotype.

All in all, these data suggest that rare genetic variants influencing protein activity modulate the predisposition to develop NAFLD-HCC.

Table 3. Clinical features associated with the presence of APOB pathogenic and likely pathogenic mutations predicted to alter protein function (likely pathogenic) in Italian patients with advanced NAFLD (advanced fibrosis or HCC).

	APOB		**p value**
	Yes	**No**	
Age, years	68.0 ± 6.8 (n=12)	63.8 ± 10.4 (n=102)	0.07
Total cholesterol, mg/dl	158.6 ± 52.4 (n=6)	183.4 ± 5.7 (n=47)	0.3
Triglycerides, mg/dl	80.4 ± 30.7 (n=7)	142.5 ± 73.7 (n=46)	0.001
HDL cholesterol, mg/dl	84.4 ± 26.7 (n=7)	46.3 ± 19.1 (n=47)	0.008

(): number of individuals for whom data at diagnosis were available.

4.3 Identification of novel genetic risk factors predisposing to NAFLD progression

In order to focus specifically on rare genetic risk variants conferring predisposition to advanced NAFLD, defined as the presence of advanced fibrosis (stage F3-F4) and/or development of hepatocellular carcinoma, we exploited a novel bioinformatic pipeline described in the Methods section.

In this case, the aim was to highlight novel risk factors for NAFLD progression in not *a priori* selected genes (candidate genes).

We considered WES results in 131 Northern Italian patients from the multicenter NAFLD-EPIDEMIC cohort, whose features are presented in Table 1.

Variants affecting protein sequence were prioritized according to their frequency (minor allele frequency, MAF<0.005), predicted to damage protein activity based on type aminoacidic

substitution, localization in specific domains, evolutionary conservation (Rare Exome Variant Ensemble Learner score, REVEL score>0.5) and low likelihood of accumulating damaging mutations in the affected genes in the general population (Residual Variation Intolerance Score, RVIS <50 and Gene Damage Index score, GDI score<10).

Variants enrichment was compared to the general population (ExAC-NFE plus 50 local controls; n=33,123) by Fisher's exact test, adjusted for false discovery rate (FDR). Starting from 87,123 variants affecting protein sequence detected in the discovery cohort, we identified 225 variants satisfying the inclusion criteria. Nine variants resulted enriched in the discovery cohort (FDR adjusted $p<0.05$) as compared to the general population (Table 4).

Table 4. List of variants enriched (FDR adjusted p<0.05) in the discovery (n=105) and replication (n=26) Italian cohorts of patients with advanced NAFLD as compared to general population (n=33,123 EXAC-NFE plus 50 local ethnically matched healthy individuals).

			Discovery					Validation		Overall	
Gene	**Variant ID**	**Protein change**	**MAF discovery**	**MAF controls**	**OR (95% c.i.)**	**p discovery**	**FDR adj p**	**MAF validation**	**p validation**	**OR (95% c.i.)**	**Overall p**
CAMSAP1	rs77279694 G>A	P236S	0.019	0.003	5.7 (1.5-16.2)	$7.4*10^{-3}$	0.036				
NOTCH1	rs138504021 C>T	R621H	0.019	0.002	9.9 (2.6-26.2)	$9.1*10^{-4}$	0.035				
NUP210	rs151008831 C>T	G1052S	0.019	0.004	5.3 (1.4-14.0)	$7.8*10^{-3}$	0.036				
PDE6A	rs148938083 C>T	A262T	0.019	0.001	20.8 (5.4-56.7)	$5.9*10^{-5}$	0.012	0.019	0.048	20.0 (6.2-49.9)	$8.5*10^{-6}$
PKP4	rs11539803 C>T	T727M	0.019	0.002	8.7 (2.3-23.0)	$1.4*10^{-3}$	0.035				
G6PD	rs137852318 C>G	D312H	0.029	0	25.1 (6.5-69.3)	$2.9*10^{-5}$	0.010				
ACADL	rs146511220 C>G	G241A	0.014	0.001	9.8 (2.0-30.0)	$4.1*10^{-3}$	0.035				
ATG7	rs143545741 C>T	P426L	0.014	0.001	10.7 (2.1-32.7)	$3.2*10^{-3}$	0.035	0.038	0.002	13.8 (4.3-33.8)	$4.6*10^{-5}$
ITGB3	rs138729147 C>T	P711S	0.014	0.001	14.2 (2.8-43.7)	$1.5*10^{-3}$	0.035				

OR: odds ration; 95% c.i.: 95% confidence interval; FDR: false discovery rate; adj: adjusted; MAF: minor allele frequency

In order to focus the interest on variants most robustly associated with advanced NAFLD, we next validated the association in an independent Italian replication cohort (n=26). We confirmed a significant enrichment in A262T *PDE6A, Phosphodiesterase 6* (overall OR 20.0, 95% c.i. 6.2-49.9) and P426L *ATG7, Autophagy-related 7* (overall OR 13.8, 95% c.i. 4.3-33.8).

P426L in *ATG7*, in particular, is a loss-of-function variant in the ubiquitin-binding domain of *ATG7* gene, probably leading to defective autophagy.

We next examined the impact of *ATG7* variants on the risk of liver disease and fatty liver disease (FLD) in individuals of the UK Biobank (UKBB) cohort from the general population. Possibly due to the limited power to evaluate the impact of a single rare variant, P426L *ATG7*

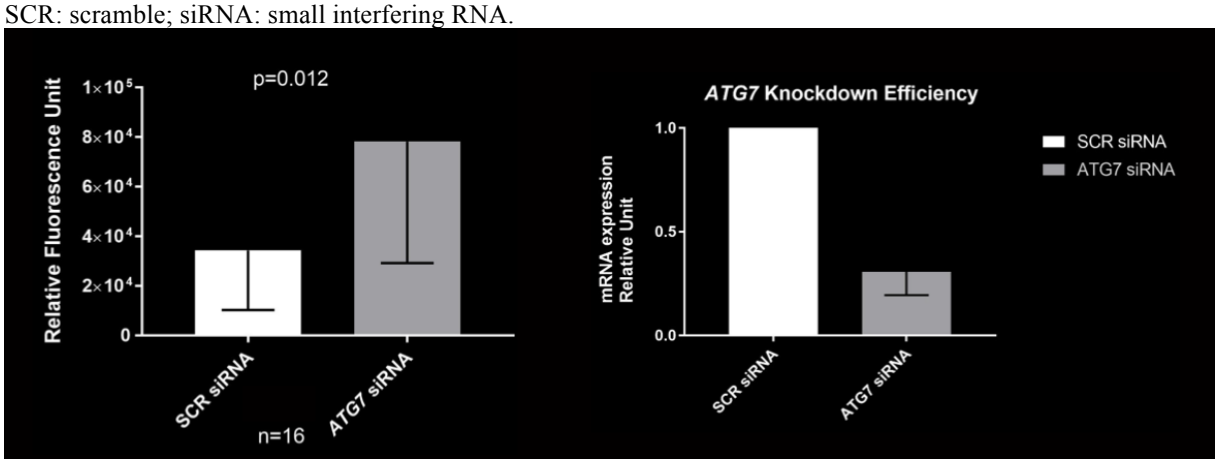
was not significantly associated with liver traits (Table 5). However, the more common low-frequency rs143545741 T>A, encoding for the V471A protein variant, was associated with an increased risk of both liver disease and NAFLD (Table 5).

Table 5. Impact of *ATG7 PDE6A* variants on the risk of liver disease in individuals from the general population of the UKBB cohort.

ATG7 Variant	MAF	AA	Liver diseases (K70-K77)		FLD (K76)	
			beta	P value	beta	P value
rs36117895	0.0362	V471A	+0.00177	0.0039	+0.00133	0.00899
rs143545741	0.0015	P426L	+0.00212	0.47	+0.00058	0.81

To evaluate the impact of the *ATG7* loss of function variants identified on hepatocellular fat accumulation, we demonstrated that *ATG7* silencing by siRNA (small interfering RNA) increased intracellular fat content in in human hepatocytes (Figure.2).

Figure 7. *ATG7* down-regulation increases intracellular fat in human hepatocytes (HepaRG cells). Adipo Red Assay after transient transfection of 5 µg of wild type *ATG7* pcDNA 3.1 vector. RFU (relative fluorescence unit) data are shown as average (± SD) of quadruplicates of four experiments . The p value was calculated by Mann-Whitney analysis. HepaRG cells were grown in Williams Medium with 10 % FBS + Oleic Acid 50 µg/ µL; SCR: scramble; siRNA: small interfering RNA.



Finally, we evaluated the impact of *ATG7* P426L variant on the clinical features of patients from the EPIDEMIC cohort. We found a higher prevalence of type 2 diabetes in carriers of the mutation (Table 6).

Table 6. Characteristics of the patients carrying the P426L *ATG7* variant

	Carriers (n=5)	**Non carriers (n=126)**	***P*-value**
Sex, F	1 (20%)	37 (29%)	ns
Age, years	66±10	64± 10	ns
BMI, Kg/m^2	29.0±2.0	30.5± 5.6 n= 99	ns
Cholesterol, mg/dl	171±17	179.3± 42 n=54	ns
Triglycerides, mg/dl	163±110	136± 71 n=54	ns
HDL, mg/dl	47±6	50±22 n=54	ns
ALT, IU/l	37±15	54±35 n=77	ns
HCC-NAFLD	2 (40%)	71 (44%)	ns
Type 2 diabetes, yes	5 (100%)	73 (65%) n=111	0.04

4.4 Genetic risk score development and validation

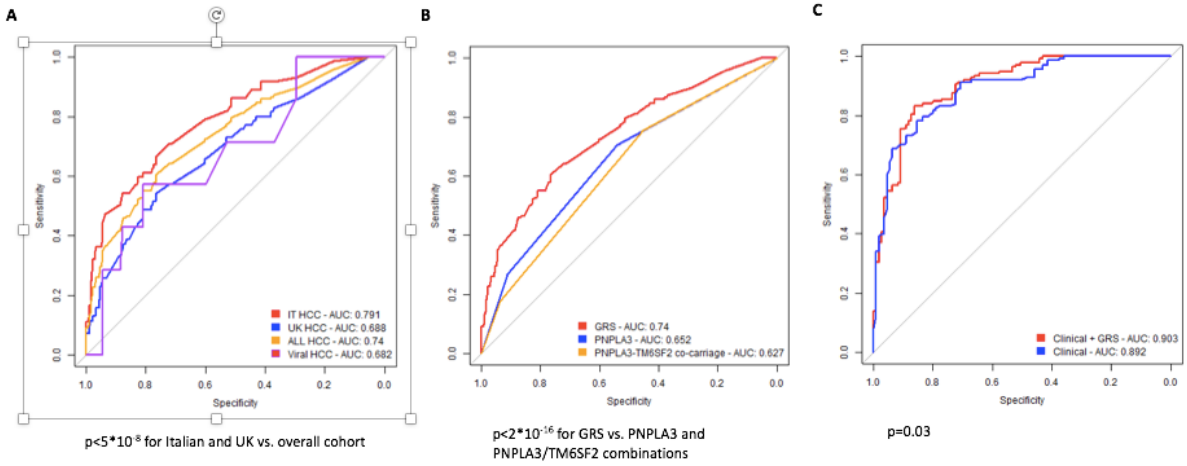
Finally, to examine whether evaluation of pathogenic and rare variants may be clinically helpful in the stratification of NAFLD-HCC risk, we developed a weighted GRS for this condition and tested its diagnostic accuracy. The GRS coefficients are shown in Table 7.

Table 7. Coefficients used to develop the Genetic risk score (GRS) for NAFLD-HCC in the 655 individuals included in the study. See page 19 for gene abbreviation list.

variable	β Coefficient	Jackknifed β Coefficient	Bias	SD
Intercept	0.072	0.072	-3.12E-06	0.001
MTHFR	-0.109	-0.109	-5.06E-06	0.006
SDHB	-0.080	-0.080	4.59E-06	0.001
MUTYH	-0.044	-0.044	1.19E-05	0.004
SDHC	0.668	0.668	2.30E-06	0.002
ATM	0.420	0.420	-4.15E-05	0.016
HNF1A	0.040	0.040	1.09E-05	0.009
BRCA2	-0.206	-0.206	3.44E-06	0.001
ATP7B	0.090	0.090	-1.84E-05	0.005
TINF2	-0.234	-0.234	-5.23E-06	0.004
SYNE2	-0.085	-0.085	-3.49E-05	0.003
SERPINA1	0.024	0.024	9.34E-06	0.004
FAH	0.131	0.131	-5.03E-06	0.004
FANCA	0.542	0.542	1.12E-05	0.006
TP53	-0.090	-0.090	3.13E-05	0.004
G6PC	-0.007	-0.007	-2.38E-05	0.014
APOB	0.780	0.779	-4.16E-05	0.006
FANCL	0.794	0.794	3.44E-06	0.008
CHEK2	0.457	0.457	1.64E-06	0.017
XPC	0.109	0.109	-6.73E-06	0.012
GBE1	0.029	0.029	9.63E-06	0.006
TF	0.321	0.321	2.12E-06	0.013
CP	-0.080	-0.080	4.59E-06	0.001
EGF	0.095	0.095	9.55E-06	0.003
TERT	-0.053	-0.053	-1.58E-05	0.006
SQSTM1	-0.368	-0.368	1.25E-05	0.006
ASL	0.014	0.014	9.59E-06	0.004
ABCB4	0.099	0.099	6.75E-06	0.003
CFTR	0.590	0.590	1.73E-05	0.006
CDKN2A	-0.118	-0.118	1.52E-05	0.004
ALDOB	0.790	0.790	2.85E-06	0.010
RPS6KA3	-0.142	-0.142	1.60E-05	0.009
DMD	0.023	0.023	0.000143	0.006
RB1	0.462	0.462	3.09E-06	0.011
SMAD4	0.174	0.174	9.88E-06	0.004
RTEL1	0.248	0.248	-7.05E-06	0.006
TSC1	0.103	0.103	6.74E-06	0.002
TM6SF2	0.007	0.007	-1.46E-06	0.001
MBOAT7	0.130	0.130	4.13E-07	0.001
PNPLA3	0.109	0.109	-6.73E-06	0.012

In the overall cohort of 655 individuals, the GRS was associated with HCC risk (OR 435, 95% c.i. 111-1903; $p<2*10^{-16}$, OR 4.96, 95% c.i. 3.29-7.55; $p=5.1*10^{-16}$ for high vs. low GRS). The GRS had an AUROC of 0.74, 95% c.i. 0.69-0.79 for predicting NAFLD-HCC in the total cohort, as compared to 0.79, 95% c.i. 0.73-0.85 and to 0.69, 95% c.i. 0.62-0.75 in the discovery and validation cohorts, respectively (Figure 8A). The best GRS threshold had a 61% sensitivity and 72% specificity to detect NAFLD-HCC.

Figure 8. Genetic risk score. Diagnostic accuracy of the Genetic risk score (GRS) for NAFLD-HCC in the 655 individuals included in the study. (**a**) comparison of the diagnostic accuracy in the study cohorts; $p < 0.05$ for diagnostic accuracy in the overall vs. single cohorts. (**b**) Diagnostic accuracy of GRS vs. *PNPLA3* I148M variant alone and a combination of *PNPLA3* I148M and *TM6SF2* E167K variants in determining NAFLD HCC risk; $p < 2*10-8$ for alternative genetic scores vs. the overall GRS. (**c**) Additive value of adding GRS to a diagnostic model based on clinical risk factors, in determining NAFLD-HCC risk in the 251 individuals of the EPIDEMIC study; $p = 0.17$. Comparison of diagnostic accuracy was performed by two-sided Venkatraman test.



The complete GRS improved the ability to discriminate NAFLD-HCC risk as compared to evaluation of the *PNPLA3* I148M variant alone and of a combination of the *PNPLA3* I148M and *TM6SF2* E167K variants ($p<2*10^{-16}$; Figure 4B).

In the EPIDEMIC cohort with complete data (n=251), the GRS was associated with NAFLD-HCC (OR 4.96, 95% c.i. 3.29-7.55; $p=5.1*10^{-16}$), independently of classic risk factors (OR 2.28, 95% c.i. 1.06-4.97, p=0.04; Table 3). Although the analysis was limited by the sample size, and retrospective cross-sectional study design, the addition of GRS to a model based on acquired risk factors modified the ability to discriminate NAFLD-HCC (p=0.03 for comparison of AUC; shown in Figure 4C). This was mainly due to a slight increase in sensitivity (83% v. 78%). In the EPIDEMIC cohort, the clinical risk score

misclassified 46 patients (19%). Addition of GRS to risk prediction led to a net improvement in classification in 8 individuals (3% of the overall cohort, 17% of misclassified; p=0.004).

4.5 WES clinical application to diagnosis in patients with cryptogenic liver disease

Finally, we considered a cohort of six adult patients presenting to our hospital with cryptogenic liver disease, or liver disease which was clinically classified as unexpectedly severe given the risk factors (metabolic comorbidies were allowed) and/or with a strong family history. Clinical features of these subjects are presented in Table 2 in the Methods section.

We examined by WES whether pathogenic and rare mutations predicted to alter protein sequence in candidate genes responsible for inherited liver disease or cancer syndromes are enriched in patients with unexplained liver disease (see Methods paragraph 3.2).

In a 27 year-old patient under steroids for culturism, presenting with decompensated chirrosis, hepatic adenomatosis and severely elevated total cholesterol levels (900 mg/dl), we identified a nonsense mutation in the *LDLRAP1* (low density lipoprotein receptor 1*)* gene leading to a premature translational stop signal (p.Gln136*).

LDLRAP1 encodes for an adaptor protein that binds LDL receptor mediating cellular internalization. This variant is expected to lead to a disrupted protein product and is associated to Familial Hypercholesterolemia type 4 (autosomal codominant trait). The genetic variant may account for the severe hyperlipidemia, though the relationship with liver damage remains to be proven.

WES allowed moreover to identify two frameshift variants in *HPS1 (*Hermansky-Pludak syndrome 1*)* and *DTNBP1* (Dystrobrevin Binding Protein 1) genes in a 41 year-old male with a complex clinical presentation characterized by bone marrow failure, pulmonary and hepatic

fibrosis. Variants in these genes have been both associated to Hermansky-Pludak syndrome, a rare autosomal recessive disorder in which albinism, bleeding due to platelet dysfunction, pulmonary fibrosis and lysosomal ceroid storage are observed. The clinical phenotype is fully consistent with the genetic makeup, where two loss-of-function variants in the pathway contribute to the disease (genetic heterogeneity).

A 68 year-old male affected by NAFLD-related liver cirrosis complicated by portal hypertension and showing family history for HCC was carrier of the missense mutation p.Ala150Pro in *ALDOB* gene (Aldolase B). Mutations in this gene are associated to Hereditary fructose intolerance (HFI), an autosomal recessive metabolic disorder resulting from a deficiency in aldolase B, characterized by postprandial hypoglycemia after fructose ingestion, and frequently associated with NAFLD [91, 92]. Carriage of this specific risk variant may in heterozygosity may have contributed to progressive liver disease in this patients.

5. Discussion

5.1 A Precision Medicine approach for the diagnosis of cryptogenic liver disease

Chronic liver disease can remain undiagnosed for a long time and is often diagnosed at advanced stage.

It is estimated that 30% of the causes of cirrhosis is classified as of unknown etiology and that 14% of patients awaiting for liver transplantation is affected by cryptogenic cirrhosis, making it a problem of important clinical relevance.

Moreover, there is a trend to misclassification of these patients who often receive a diagnosis of NASH in the only presence of just one metabolic comorbidity (burn-out NASH) [93, 94].

In the setting of a Precision Medicine (PM) approach, defined as a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products

being tailored to the individual patient, genetics and namely Next Generation Sequence technology could definitely represent an important instrument also as far as liver disease is concerned [80].

In this respect, WES involves sequencing of the coding regions (exons) of all genes, and WGS involves sequencing of both coding and noncoding regions. The exome represents about 1% of the genome. Currently, WES is available for clinical diagnostics for some indications and GS is used predominantly in the research setting. Because WES involves sequence analysis of all genes in the genome, it can identify mutations in genes that are not suspected on the basis of clinical presentation or are not yet known to cause disease. When initial analysis is unable to establish a diagnosis, WES data can be reinterrogated as new genes for a given condition are discovered and new exome analysis methods are developed [80].

Here we can find listed the possible indications for genetic work-up or referral to a genetic Specialist in the context of Internal Medicine:

1. Patients with clinical findings of a specific monogenic syndrome (e.g. polycystic kidneys in patients with renal dysfunction)
2. Patients with a rare condition that has an established genetic predisposition (e.g. severe hyperlypidemia).
3. Patients with early disease onset and a strongly positive family history.
4. Patients with rare, unexplained disorders and unrevealing standard diagnostic work-ups.
5. Healthy persons with family history of a disease for which early diagnosis allows prevention (e.g. sudden cardiac death, ovarian cancer…).
6. Couples preparing to conceive whose ethnicities have a high carrier frequency for specific disorders or who are related by bloodline.

In a recent work Hakim et al. report various clinical cases in which WES has demostrated its crucial utility in the diagnosis and clinical management of patients with advanced liver disease.

The authors describe mainly young patients with early symptoms onset and a severe phenotype. In two cases the identification of a previously undiagnosed genetic disease allows to establish a specific treatment for the cause of the disorder (namely autosomal dominant familial partial lipodystrophy type 3, FPLD3, associated to leptin deficiency and hypobetalypoproteinemia) [93].

In keeping with literature data, we demonstrated similar results in a cohort of six patients with cryptogenic liver disease whose clinical features are shown in Table 2. In at leat 2 of these subjects in fact WES could lead to the recognition of unappreciated phenotypic features and enable family screening and possibly therapeutic interventions.

In a 23 year-old patient under steroids for culturism, presenting with decompensated cirrhosis, hepatic adenomatosis and severely elevated total cholesterol (900 mg/dl), we identified a nonsense mutation in the *LDLRAP1* gene leading to a premature translational stop signal (p.Gln136*). *LDLRAP1* encodes for an adaptor protein targeting LDL receptor for internalization. This variant is expected to lead to a disrupted protein and is associated with Familial Hypercholesterolemia type 4 [95].

In a 40 year-old male with bone marrow failure, pulmonary and hepatic fibrosis, we identified two frameshift rare pathogenic variants in *HPS1-DTNBP1* genes. Variants in these genes have been both associated to Hermansky-Pludak syndrome, a rare autosomal recessive disorder characterized by albinism (present in the patient), bleeding due to platelet dysfunction, pulmonary fibrosis and lysosomal ceroid storage [96].

A 68 year-old male affected by NAFLD-related liver chirrosis complicated by portal hypertension and showing family history for HCC was carrier of the missense mutation p.Ala150Pro in *ALDOB* gene (Aldolase B). Mutations in this gene are associated to

Hereditary fructose intolerance (HFI), an autosomal recessive metabolic disorder resulting from a deficiency in aldolase B, characterized by postprandial hypoglycemia after fructose ingestion, and frequently associated with NAFLD [91, 92].

5.2 Genetics and liver disease: an important instrument for predicting disease progression, estimating risk stratification and finding possible therapeutic targets in advanced NAFLD

Besides the diagnosis of cryptogenic liver disease NGS technology have proven to be extremely useful in the stratification of the risk of progression to cirrhosis and HCC in the setting of NAFLD.

Due to the very high prevalence of the population at risk, classic screening strategies are in fact presently unfeasible. Therefore, novel noninvasive biomarkers are urgently needed to improve disease risk stratification. Indeed, although carriage of the common *PNPLA3* I148M variant is a strongly associated with NAFLD progression and HCC development and it has been addressed as a possible therapeutic target, taken by itself it is not sufficiently accurate to stratify the risk of this condition [33, 105, 106].

Here we showed that in patients with NAFLD-HCC, pathogenic and likely pathogenic variants in genes linked to liver disease and cancer predisposition are enriched as compared to healthy individuals. Furthermore, we have replicated this result in two independent cohorts.

Although further validation is required before target gene resequencing can be recommended in clinical practice, these findings have potential clinical implications. In the present cohort, resequencing of the candidate genes panel led to the detection of likely predisposing genetic conditions in a large fraction of patients presenting with NAFLD-HCC. This may also help in the identification of family members for whom screening would be cost effective, or specific preventive treatments may be considered.

Furthermore, evaluation of a comprehensive GRS, which takes into consideration rare variants, in individuals with NAFLD may allow a more accurate HCC risk stratification and the implementation of targeted surveillance. Indeed, the comprehensive GRS showed superior diagnostic accuracy as compared to the evaluation of common genetic risk factors, including the *PNPLA3* I148M variant alone [33], or a combination of *PNPLA3* I148M and *TM6SF2* E167K variants [97]. Furthermore, the GRS improved patient stratification, when considered together with classical risk factors for NAFLD-HCC. The clinical utility of GRS assessment should however be tested in familial and prospective studies evaluating patients with NAFLD and other liver diseases.

Furthermore, the present findings may also have pathophysiological implications worthy of exploration.

Consistently with previous data [43], other variants favoring hepatocellular fat retention were associated with NAFLD-HCC, including common and rare variants in *TM6SF2* and *MBOAT7* genes. Secondly, variants in *APOB*, responsible for hypobetalipoproteinemia, were collectively observed in a high proportion of Italian patients (15%), and there was a significant enrichment in pathogenic and truncating mutations in this gene in the overall cohort of NAFLD-HCC patients. *APOB* genetic variants leading to the synthesis of a dysfunctional ApoB100 protein and to a consequent impairment in the export of lipids from hepatocytes within very low-density lipoproteins are responsible for the development of severe hepatic steatosis (hypobetalipoproteinemia, an autosomal dominant disease). At the same time, some *APOB* variants that lead to the alteration of the first portion of the protein result also in altered activity of ApoB48, the protein isoform expressed by enterocytes. This results in retention of chylomicrones, malabsorption of fat and liposoluble vitamins (retinol - vitamin A, vitamin E and vitamin D), known to play a protective role in liver disease progression, and possibly in the alteration of the intestinal barrier. Most importantly, individuals carrying *APOB* mutations had a circulating lipid profile consistent with

hypobetalipoproteinemia, providing functional validation of the pathogenicity of the genetic mutations identified. Notably, in line with a causal role of hepatocellular lipid retention in promoting NAFLD-HCC, somatic mutations in *APOB* also frequently occur during hepatic carcinogenesis [98]. The mechanism connecting *APOB* mutations with carcinogenesis is still not completely understood. Induction of hepatocellular lipid accumulation, oxidative stress, and the loss of a possible tumor suppressive activity of APOB are some of the hypothesis that have been raised [99, 100]. Therefore, the identification of *APOB* mutations in subjects with NAFLD-HCC would be to allow the diagnosis, in these cases mostly unrecognized, of familial hypobetalipoproteinemia in the first-degree relatives, allowing to establish adequate HCC surveillance.

An additional finding was the novel association between variants in *SQSTM1* and NAFLD-HCC. *SQSMT1* encodes for p62, a component of Mallory-Denk Bodies and hyaline granules. Protein p62 aggregates accumulate in the cytoplasm of damaged liver cells in NASH and HCC [101], and may promote hepatocytes transformation through the activation of antioxidants and mTOR pathways [102, 103]. In keeping, we also identified variants in genes regulating cell growth via the insulin signaling and mTOR pathways, and, in line with previous findings from our group [104], in the telomere regulation machinery.

Moreover, by exploiting WES of patients with NAFLD followed by variants prioritization based on the frequency, predicted impact, and evolutionary conservation, in this work we identified an enrichment in a loss-of-function variant of *Autophagy-related 7* protein (*ATG7*), which facilitates hepatocellular fat accumulation due to defective lipophagy [51-52]. This represents a novel genetic risk factor possibly predisposing to liver disease progression. In line with previous data which hypothesized the importance of autophagy pathway in the promotion of insulin sensitivity [51], we furtermore evidenced how patients carriers of the *ATG7* P426L variant associated to defective autophagy showed a higher prevalence of type two diabetes.

This study has some limitations. First, the design was cross-sectional with retrospective data collection, so that GRS for NAFLD-HCC will need to be validated in future prospective studies including individuals with NAFLD and other liver diseases at high baseline risk.

In this respect, in a very recent work, we examined the impact of the previously developed PRS-HFC (polygenic risk score based on the evaluation of genetic variants associated to hepatic fat content) that can be evaluated in the clinic on HCC in at-risk individuals and in the general population [54]. We also performed a further adjustment for *HSD17B13* (termed PRS-5). PRS proved to be able to identify with good specificity a subset of individuals with NAFLD and dysmetabolism at high risk for HCC and predicted HCC irrespective of severe liver fibrosis [56], thus confirming the usefulness of genetic risk scores as instruments for risk stratification in patients with NAFLD.

Furthermore, the sample size was relatively limited, and therefore we mainly focused our attention on pathogenic variants in candidate genes, which are already known to cause disease. As we considered healthy individuals from the general populationas controls, results are potentially applicable to NAFLD-HCC genetic screening at population level without prior knowledge of liver disease severity status. If it will be proven cost-effective and ethically acceptable, this approach may assist in stratifying the risk of liver disease and HCC, besides of other chronic degenerative diseases.

Additional studies are required to discover new variants predisposing to NAFLD-HCC, which were not examined in this study. Moreover, we could not evaluate a control group with advanced fibrosis due to NAFLD for the UK NAFLD-HCC validation cohort, in which a different pattern of mutations was observed as compared to the Italian cohort. Finally, findings may only be applicable to individuals of European descent.

In conclusion, rare pathogenic variants in candidate genes involved in the predisposition to liver disease or cancer are associated with an increased risk of developing NAFLD-HCC.

6.Conclusion

In this work we examined the importance of Next Generation Sequencing technology, namely Whole Exome Sequencing, in the diagnosis and clinical managementeof cryptogenic liver disease and in risk stratification of NAFLD progression towards cirrhosis and HCC.

NAFLD, now the leading cause of liver damage worldwide, is epidemiologically associated with obesity, insulin resistance and type 2 diabetes, and is a potentially progressive condition to advanced liver fibrosis and hepatocellular carcinoma. There is huge interindividual variability in liver disease susceptibility. Inherited factors play an important role in determining disease predisposition. During the last years, common variants in *PNPLA3*, *TM6SF2*, *MBOAT7* and *GCKR* have been demonstrated to predispose to the full spectrum of NAFLD pathology by facilitating hepatic fat accumulation in the presence of environmental triggers. Other variants regulating inflammation and fibrogenesis then modulate liver disease progression in those at higher risk.

As we showed, evidence is also accumulating that rare variants are involved in disease predisposition (*APOB, ATG7)*. In the future, a comprehensive evaluation of genetic risk factors by WES through the implementation of genetic risk scores may be exploited to stratify the risk of liver-related complications of the disease, and to guide hepatocellular carcinoma surveillance and choose pharmacological therapy.

Furtermore, since a consistent fraction of chronic liver disease is identified at advanced stage and in one third of cases the etiology remains unknown (cryptogenic cirrhosis), genetic testing by WES approach gives a precious instrument for punctual diagnose of difficult cases in which conventional diagnostic work-up, even though exstensive, has been not conclusive.

In keeping with previous results from the literature, we showed that at least 30% of cryptogenic cirrhosis can benefit from genetic study which can enable the recognition of unappreciated genetic disorders, allow family screening and possibly therapeutic interventions.

7. References

1. Eslam M, Newsome PN, Anstee QM, Targher G, Gomez MR, Zelber-Sagi S, et al. A new definition for metabolic associated fatty liver disease: an international expert consensus statement. J Hepatol 2020.

2. Younossi, Z. and L. Henry, Contribution of Alcoholic and Nonalcoholic Fatty Liver Disease to the Burden of Liver-Related Morbidity and Mortality. Gastroenterology, 2016. 150(8): p. 1778-85.

3. Torres, D.M. and S.A. Harrison, Nonalcoholic steatohepatitis and noncirrhotic hepatocellular carcinoma: fertile soil. Semin Liver Dis, 2012. 32(1): p. 30-8.

4. Kawada, N., et al., Hepatocellular carcinoma arising from non-cirrhotic nonalcoholic steatohepatitis. J Gastroenterol, 2009. 44(12): p. 1190-4.

5. Ertle, J., et al., Non-alcoholic fatty liver disease progresses to hepatocellular carcinoma in the absence of apparent cirrhosis. Int J Cancer, 2011. 128(10): p. 2436-43.

6. Chagas, A.L., et al., Does hepatocellular carcinoma in non-alcoholic steatohepatitis exist in cirrhotic and non-cirrhotic patients? Braz J Med Biol Res, 2009. 42(10): p. 958-62.

7. Dongiovanni, P. and L. Valenti, Genetics of nonalcoholic fatty liver disease. Metabolism, 2015.

8. Paradis, V., et al., Hepatocellular carcinomas in patients with metabolic syndrome often develop without significant liver fibrosis: a pathological analysis. Hepatology, 2009. 49(3): p. 851-9.

9. El-Serag, H.B. and K.L. Rudolph, Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology, 2007. 132(7): p. 2557-76.

10. Poonawala, A., S.P. Nair, and P.J. Thuluvath, Prevalence of obesity and diabetes in patients with cryptogenic cirrhosis: a case-control study. Hepatology, 2000. 32(4 Pt 1): p. 689-92.

11. Calle, E.E., et al., Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med, 2003. 348(17): p. 1625-38.

12. Larsson, S.C. and A. Wolk, Overweight, obesity and risk of liver cancer: a meta-analysis of cohort studies. Br J Cancer, 2007. 97(7): p. 1005-8.

13. Davila, J.A., et al., Diabetes increases the risk of hepatocellular carcinoma in the United States: a population based case control study. Gut, 2005. 54(4): p. 533-9.

14. Calle, E.E. and R. Kaaks, Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nat Rev Cancer, 2004. 4(8): p. 579-91.

15. Dongiovanni, P., S. Romeo, and L. Valenti, Hepatocellular carcinoma in nonalcoholic fatty liver: role of environmental and genetic factors. World J Gastroenterol, 2014. 20(36): p. 12945-55.

16. Loomba R, Lim JK, Patton H, El-Serag HB. AGA Clinical Practice Update on Screening and Surveillance for Hepatocellular Carcinoma in Patients With Nonalcoholic Fatty Liver Disease: Expert Review. Gastroenterology 2020.

17. Piscaglia F, Svegliati-Baroni G, Barchetti A, Pecorelli A, Marinelli S, Tiribelli C, et al. Clinical patterns of hepatocellular carcinoma in nonalcoholic fatty liver disease: A multicenter prospective study. Hepatology 2016;63:827-838.

18. Younossi ZM, Otgonsuren M, Henry L, Venkatesan C, Mishra A, Erario M, et al. Association of nonalcoholic fatty liver disease (NAFLD) with hepatocellular carcinoma (HCC) in the United States from 2004 to 2009. Hepatology 2015;62:1723-1730.

19. Turati, F., et al., Family history of liver cancer and hepatocellular carcinoma. Hepatology, 2012. 55(5): p. 1416-25.

20. Romeo, S., et al., Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet, 2008. 40: p. 1461-1465.

21. Valenti, L., et al., Homozygosity for the PNPLA3 / adiponutrin I148M polymorphism influences liver fibrosis in patients with nonalcoholic fatty liver disease. Hepatology, 2010. 51: p. 1209-1217.

22. Sookoian, S. and C.J. Pirola, Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease. Hepatology, 2011. 53(6): p. 1883-94.

23. Dongiovanni, P., et al., PNPLA3 I148M polymorphism and progressive liver disease. World J Gastroenterol, 2013. 19(41): p. 6969-78.

24. Yuan, X., et al., Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. Am J Hum Genet, 2008. 83(4): p. 520-8.

25. Donati, B., et al., The rs2294918 E434K Variant Modulates Patatin-Like Phospholipase Domain-Containing 3 Expression and Liver Damage. Hepatology, 2016. 63(3): p. 787-798.

26. Ruhanen, H., et al., PNPLA3 mediates hepatocyte triacylglycerol remodeling. J Lipid Res, 2014. 55(4): p. 739-46.

27. Smagris, E., et al., Pnpla3I148M knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis. Hepatology, 2015. 61: p. 108-18.

28. Pirazzi, C., et al., PNPLA3 has retinyl-palmitate lipase activity in human hepatic stellate cells. Hum Mol Genet, 2014. 23(15): p. 4077-85.

29. Mondul, A., et al., PNPLA3 I148M Variant Influences Circulating Retinol in Adults with Nonalcoholic Fatty Liver Disease or Obesity. J Nutr, 2015. 145(8): p. 1687-91.

30. Liu, Y.L., et al., Carriage of the PNPLA3 rs738409 C >G polymorphism confers an increased risk of non-alcoholic fatty liver disease associated hepatocellular carcinoma. J Hepatol, 2013. 61(1): p. 75-81.

31. Trepo, E., et al., Association between the PNPLA3 (rs738409 C>G) variant and hepatocellular carcinoma: Evidence from a meta-analysis of individual participant data. Hepatology, 2014. 59(6): p. 2170-7.

32. Valenti, L., et al., PNPLA3 I148M variant and hepatocellular carcinoma: a common genetic variant for a rare disease. Dig Liver Dis, 2013. 45(8): p. 619-24.

33. Anstee, Q.M., et al., Reply to: HCC and liver disease risk in homozygous PNPLA3 p.I148M carriers approach monogenic inheritance. J Hepatol, 2015. 62(4): p. 982-3.

34. Kozlitina, J., et al., Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat Genet, 2014. 46(4): p. 352-6.

35. Liu, Y.L., et al., TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. Nat Commun, 2014. 5: p. 4309.

36. Dongiovanni, P., et al., Transmembrane 6 superfamily member 2 gene variant disentangles nonalcoholic steatohepatitis from cardiovascular disease. Hepatology, 2015. 61(2): p. 506-14.

37. Falleti, E., et al., PNPLA3 rs738409 and TM6SF2 rs58542926 variants increase the risk of hepatocellular carcinoma in alcoholic cirrhosis. Dig Liver Dis, 2016. 48(1): p. 69-75.

38. Buch, S., F. Stickel, and E. Trepo, A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. 2015. 47(12): p. 1443-8.

39. Mancina, R.M., et al., The MBOAT7-TMC4 Variant rs641738 Increases Risk of Nonalcoholic Fatty Liver Disease in Individuals of European Descent. Gastroenterology, 2016. 150(5): p. 1219-1230.e6.

40. Trepo E, Valenti L. Update on NAFLD genetics: From new variants to the clinic. J Hepatol 2020.

41. Pelusi S, Valenti L. Hepatic fat as clinical outcome and therapeutic target for nonalcoholic fatty liver disease. Liver Int 2019;39:250-256.

42. Pelusi S, Baselli G, Pietrelli A, Dongiovanni P, Donati B, McCain MV, et al. Rare Pathogenic Variants Predispose to Hepatocellular Carcinoma in Nonalcoholic Fatty Liver Disease. Sci Rep 2019;9:3682.

43. Donati B, Dongiovanni P, Romeo S, Meroni M, McCain M, Miele L, et al. MBOAT7 rs641738 variant and hepatocellular carcinoma in non-cirrhotic individuals. Sci Rep 2017;7:4492.

44. Abul-Husn NS, Cheng X, Li AH, Xin Y, Schurmann C, Stevis P, et al. A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. N Engl J Med 2018;378:1096-1106.

45. Calado, R.T., et al., A spectrum of severe familial liver disorders associate with telomerase mutations. PLoS One, 2009. 4(11): p. e7926.

46. Rudolph, K.L., et al., Inhibition of experimental liver cirrhosis in mice by telomerase gene delivery. Science, 2000. 287(5456): p. 1253-8.

47. Calado, R.T., et al., Constitutional telomerase mutations are genetic risk factors for cirrhosis. Hepatology, 2011. 53(5): p. 1600-7.

48. Hartmann, D., et al., Telomerase gene mutations are associated with cirrhosis formation. Hepatology, 2011. 53(5): p. 1608-17.

49. Valenti, L., et al., Liver transplantation for hepatocellular carcinoma in a patient with a novel telomerase mutation and steatosis. J Hepatol, 2013. 58(2): p. 399-401.

50. Di Filippo, M., et al., Homozygous MTTP and APOB mutations may lead to hepatic steatosis and fibrosis despite metabolic differences in congenital hypocholesterolemia. J Hepatol, 2014. 61(4): p. 891-902.

51. Czaja, M.J. Function of Autophagy in Nonalcoholic Fatty Liver Disease. Dig Dis Sci 61, 1304-13 (2016).

52. Yang, L., Li, P., Fu, S., Calay, E.S. & Hotamisligil, G.S. Defective hepatic autophagy in obesity promotes ER stress and causes insulin resistance. Cell Metab 11, 467-78 (2010).

53. Xiong et al., The autophagy-related gene 14 (Atg14) is regulated by forkhead box O transcription factors and circadian rhythms and plays a critical role in hepatic autophagy and lipid metabolism. J Biol Chem, 2012. 287(46):39107-14.

54. Dongiovanni P, Stender S, Pietrelli A, Mancina RM, Cespiati A, Petta S, et al. Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. J Intern Med 2018;283:356-370.

55. Gellert-Kristensen H, Richardson TG, Davey Smith G, Nordestgaard BG, Tybjaerg-Hansen A, Stender S. Combined Effect of PNPLA3, TM6SF2, and HSD17B13 Variants on Risk of Cirrhosis and Hepatocellular Carcinoma in the General Population. Hepatology 2020.

56. Bianco, C., Jamialahmadi, O. et al., Non-invasive stratification of hepatocellular carcinoma risk in nonalcoholic fatty liver using polygenic risk scores. J Hepatol, 2020. 25: S0168-8278(20)33811-3.

57. Mamanova L, Coffey AJ, Scott CE, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118.

58. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46.

59. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.

60. Kawashima E, Farinelli L, Mayer P (1998) Method of Nucleic Acid Amplification.

61. Cock PJA, Fields CJ, Goto N, et al. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–71.

62. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med 63:35–61.

63. Schadt EE, Linderman MD, Sorenson J, et al. (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11:647–57.

64. Ng PC, Levy S, Huang J, et al. (2008) Genetic variation in an individual human exome. PLoS Genet 4:e1000160.

65. Ng SB, Turner EH, Robertson PD, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–6.

66. Robinson PN, Krawitz P, Mundlos S (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet 80:127–132.

67. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451.

68. Raney BJ, Cline MS, Rosenbloom KR, et al. (2011) ENCODE whole genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 39:D871–D875.

69. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–83.

70. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

71. Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269–276.

72. Ng SB, Buckingham KJ, Lee C, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35.

73. Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703–14.

74. Gilissen C, Arts HH, Hoischen A, et al. (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87:418–423.

75. Wang JL, Yang X, Xia K, et al. (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133:3510–3518.

76. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227.

77. Del Rosario M, Brunner HG, de Ligt J, et al. (2012) Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med 121003140044006.

78. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362:1181–1191.

79. Markello TC, Boerkoel CF, Groden C, et al. (2012) The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet Med 14:51–59.

80. Kiryluk K et al, Precision Medicine in Internal Medicine. Ann Intern Med.2019.

81. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 56, 908–943 (2012).

82. Zhang, J. et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med 373, 2336–2346 (2015).

83. Zucman-Rossi, J., Villanueva, A., Nault, J. C. & Llovet, J. M. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. Gastroenterology 149, 1226–1239 e1224 (2015).

84. Scorza, M. et al. Genetic diseases that predispose to early liver cirrhosis. Int J Hepatol 2014, 713754 (2014).

85. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310–315 (2014).

86. Cohen, J. C. et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305, 869–872 (2004).

87. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615, 28–56 (2007).

88. Angulo, P., et al., The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology, 2007. 45(4): p. 846-54.

89. Venkatraman, E. S. & Begg, C. B. A distribution-free procedure for comparing receiver operating characteristics curves from a paired experiment. Biometrika 83, 835–848 (1996).

90. Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A. & Schairer, C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122, 904–914 (1985).

91. Aldamiz-Echevarria. Non-alcoholic fatty liver in hereditary fructose intolerance . Clin Nutr. 2020.

92. Simons N. Patients with aldolase B deficiency are characterized by an increased intrahepatic triglyceride content. J Clin Endocrinol Metab. 2019).

93. Hakim A et al Clinical utility of genomic analysis in adults with idiopathic liver disease. J Hepatol 2019.

94. Czaja AJ. Cryptogenic Chronic Hepatitis and Its Changing Guise in Adults. Dig Dis Sci 2011.

95. Santos R.D. What are we able to achieve today for our patients with homozygous familial hypercholesterolaemia, and what are the unmet needs?Atheroscler Suppl 2014.

96. Huizing M et al., Hermasky-Pludak Syndrome. In: GeneReviews® [Internet]. 2000.

97. Stickel, F. et al. Genetic variants in PNPLA3 and TM6SF2 predispose to the development of hepatocellular carcinoma in individuals with alcohol-related cirrhosis. Am J Gastroenterol (2018).

98. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 169, 1327-1341 e1323 (2017).

99. Valenti, L. & Romeo, S. Editorial: new insights into the relationship between the intestine and non alcoholic fatty liver-is “fatty gut” involved in disease progression? Aliment Pharmacol Ther 46, 377–378 (2017).

100. Lee, G. et al. Clinical significance of APOB inactivation in hepatocellular carcinoma. Exp Mol Med 50, 147 (2018).

101. Stumptner, C., Fuchsbichler, A., Zatloukal, K. & Denk, H. In vitro production of Mallory bodies and intracellular hyaline bodies: thecentral role of sequestosome 1/p62. Hepatology 46, 851–860 (2007).

102. Inami Y, et al. Persistent activation of Nrf2 through p62 in hepatocellular carcinoma cells. J Cell Biol 193, 275-284 (2011).

103. Umemura, A. et al. p62, Upregulated during Preneoplasia, Induces Hepatocellular Carcinogenesis by Maintaining Survival of Stressed HCC-Initiating Cells. Cancer Cell 29, 935–948 (2016).

104. Donati, B. et al. Telomerase reverse transcriptase germline mutations and hepatocellular carcinoma in patients with nonalcoholic fatty liver disease. Cancer Med 6, 1930–1940 (2017).

105. Lindén D, et al. Pnpla3 silencing with antisense oligonucleotides ameliorates nonalcoholic steatohepatitis and fibrosis in Pnpla3 I148M knock-in mice. Molecular Metabolism 2019;22:49–61.

106. Colombo M., Pelusi S. Towards Precision Medicine in Nonalcoholic Fatty Liver Disease with PNPLA3

as a Therapeutic Target. Gastroenterol 2019. 157(4):1156-1157.