Identification of a Functional Enhancer Variant within the Chronic Pancreatitis-Associated \textit{SPINK1} c.101A>G-Containing Haplotype

Arnaud Boulling,\textsuperscript{1,2,3} Emmanuelle Masson,\textsuperscript{1,4}\textsuperscript{†} Wen-Bin Zou,\textsuperscript{5,6} Prachand Issarpu,\textsuperscript{7} Hao Wu,\textsuperscript{1,2,5,6} Seema Bhaskar,\textsuperscript{7} Sumit Paliwal,\textsuperscript{7} Emmanuelle Génin,\textsuperscript{1,2,3} David N. Cooper,\textsuperscript{5} Zhao-Shen Li,\textsuperscript{5,6} Giriraj R Chandak,\textsuperscript{7,9} Zhuhan Liao,\textsuperscript{5,6} Jian-Min Chen,\textsuperscript{1,2,3}\textsuperscript{*} and Claude Férec\textsuperscript{1,2,3,4}

\textsuperscript{1}Institut National de la Santé et de la Recherche Médicale (INSERM), U1078, Brest, France; \textsuperscript{2}Etablissement Français du Sang (EFS) – Bretagne, Brest, France; \textsuperscript{3}Faculté de Médecine et des Sciences de la Santé, Université de Bretagne Occidentale (UBO), Brest, France; \textsuperscript{4}Laboratoire de Génétique Moléculaire et d’Histocompatibilité, Centre Hospitalier Régional Universitaire (CHRU) Brest, Hôpital Morvan, Brest, France; \textsuperscript{5}Department of Gastroenterology, Changhai Hospital, the Second Military Medical University, Shanghai, China; \textsuperscript{6}Shanghai Institute of Pancreatic Diseases, Shanghai, China; \textsuperscript{7}Genomic Research on Complex Diseases, CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India; \textsuperscript{8}Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom; and \textsuperscript{9}Human Genetics Division, Genome Institute of Singapore, Biopolis, Singapore

\textbf{Contract Grant Sponsors}: HW is a joint PhD student between the Changhai Hospital and INSERM U1078 who was in receipt of a one-year scholarship from the China Scholarship Fund.
Council (No. 201503170355). Support for this study came from the Conseil Régional de Bretagne, the Association des Pancréatites Chroniques Héréditaires, the Association de Transfusion Sanguine et de Biogénétique Gaetan Saleun and the Institut National de la Santé et de la Recherche Médicale (INSERM), France; and the National Natural Science Foundation of China (81470884 and 81422010 to ZL), the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (15SG33 to ZL) and the Chang Jiang Scholars Program of Ministry of Education (Q2015190 to ZL), China. Work in the lab of GRC was supported by funds from the Council of Scientific and Industrial Research (CSIR), Ministry of Science and Technology, Government of India, India (XII Five-Year Plan titled “THUNDER”).

Disclosure statement: The authors declare no conflict of interest.

1Equal contributors.

*Correspondence to:
Jian-Min Chen, MD; PhD, INSERM U1078 and EFS – Bretagne, 46 rue Félix Le Dantec, 29218 Brest, France. Tel: +33-2-98449333; Fax: +33-2-98430555;
e-mail: Jian-Min.Chen@univ-brest

Zhuan Liao, MD, Department of Gastroenterology, Changhai Hospital, the Second Military Medical University, 168 Changhai Road, Shanghai 200433, China. Tel.: 86-21-31161335; Fax: 0086-21-55621735; e-mail: liaozhuan@smmu.edu.cn
ABSTRACT: The haplotype harboring the SPINK1 c.101A>G (p.Asn34Ser) variant (also known as rs17107315:T>C) represents the most important heritable risk factor for idiopathic chronic pancreatitis identified to date. The causal variant contained within this risk haplotype has however remained stubbornly elusive. Herein we set out to resolve this enigma by employing a hypothesis-driven approach. Firstly, we searched for variants in strong linkage disequilibrium with rs17107315:T>C using HaploReg v4.1. Secondly, we identified two candidate SNPs by visual inspection of sequences spanning all 25 SNPs found to be in linkage disequilibrium with rs17107315:T>C, guided by prior knowledge of pancreas-specific transcription factors and their cognate binding sites. Thirdly, employing a novel cis-regulatory module-guided approach to further filter the two candidate SNPs yielded a solitary candidate causal variant. Finally, combining data from phylogenetic conservation and chromatin accessibility, co-transfection transactivation experiments and population genetic studies, we suggest that rs142703147:C>A, which disrupts a PTF1L binding site within an evolutionarily conserved HNF1A–PTF1L cis-regulatory module located ~4 kb upstream of the SPINK1 promoter, contributes to the aforementioned chronic pancreatitis risk haplotype. Further studies are required not only to improve the characterization of this functional SNP but also to identify other functional components that might contribute to this high-risk haplotype.

KEYWORDS: chronic pancreatitis; enhancer; promoter reporter gene assay; regulatory variants; SPINK1 gene
Introduction

Chronic pancreatitis is an inflammatory disease of the pancreas that leads to irreversible structural and functional damage to the pancreas [Majumder and Chari 2016]. Analysis of four genes highly expressed in the pancreatic acinar cells — \textit{PRSS1} (encoding cationic trypsinogen; MIM# 276000), \textit{PRSS2} (encoding anionic trypsinogen; MIM# 601564), \textit{SPINK1} (encoding pancreatic secretory trypsin inhibitor; MIM# 167790) and \textit{CTRC} (encoding chymotrypsin C (MIM# 601405), which specifically degrades all human trypsinogen/trypsin isoforms [Szmola and Sahin-Tóth 2007]) — has defined a trypsin-dependent pathway in the pathogenesis of chronic pancreatitis. Whereas gain-of-function missense and copy number variants in \textit{PRSS1} [Le Maréchal et al., 2006; Whitcomb et al., 1996] and loss-of-function variants in \textit{SPINK1} [Witt et al., 2000] and \textit{CTRC} [Masson et al., 2008; Rosendahl et al., 2008] predispose to chronic pancreatitis, loss-of-function variants in \textit{PRSS1} [Boulling et al., 2015; Chen et al., 2003; Derikx et al., 2015; Whitcomb et al., 2012] and \textit{PRSS2} [Witt et al., 2006] protect against the disease.

The \textit{SPINK1} c.101A>G variant-associated haplotype [Witt et al. 2000] has emerged as the most important risk factor for idiopathic chronic pancreatitis as a consequence of its relatively high prevalence worldwide (allele frequency, ~0.7%) and its considerable effect size (odds ratio (OR) ≈ 14) [Aoun et al., 2008]. The \textit{SPINK1} c.101A>G variant, which was predicted to result in a p.Asn34Ser missense mutation, is termed rs17107315:T>C in the dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/). For ease of reading, we shall describe this variant as rs17107315:T>C (c.101A>G) throughout the manuscript. The identification of the causal variant underlying this high-risk haplotype is of considerable biological interest but it may also have significant diagnostic and therapeutic value. This notwithstanding, despite extensive studies, the underlying causal variant has remained stubbornly elusive [Chen and Férec 2009]. The earliest hypothesis, that p.Asn34Ser itself impairs the inhibitory action of
SPINK1 on prematurely activated trypsin within the pancreas [Witt et al. 2000], was however supported by biochemical characterization of the wild-type and mutant enzymes expressed in three different systems, *Saccharomyces cerevisiae* BJ1991 strain [Kuwata et al., 2002], Chinese hamster ovary cells [Boulling et al., 2007], and human embryonic kidney 293T (HEK293T) cells [Király et al., 2007]. A later hypothesis, that either rs17107315:T>C (c.101A>G) or one of the four intronic variants in linkage disequilibrium (LD) with it might affect pre-mRNA splicing [Chen et al., 2001], also failed to garner any evidential support whether from reverse transcription-PCR (RT-PCR) analysis of total RNA prepared from pancreatic tissues of rs17107315:T>C (c.101A>G) homozygotes [Masamune et al., 2007], or from experiments performed in the context of both mini-gene [Kereszturi et al., 2009] and full-gene [Boulling et al., 2012] systems. An alternative hypothesis, that the causal variant resides within an uncharacterized flanking region of the *SPINK1* gene with regulatory potential [Kereszturi et al. 2009], was explored in the present study.

**Material and Methods**

**Study Subjects**

The 548 French ICP patients and 562 controls have been previously reported [Fjeld et al., 2015; Witt et al., 2013]. Most of the 1104 Han Chinese ICP patients and 1196 healthy controls have been described in a recent publication [Zou et al., 2016]. The Indian chronic pancreatitis patients (n = 347) and controls (n = 264) included in this study have been described previously [Paliwal et al., 2013]. Informed consent was obtained from each patient and the study was approved by the respective ethics committees of Brest University, the Changhai Hospital and the Center for Cellular and Molecular Biology (CSIR-CCMB) in Hyderabad.
Reference Sequences

The SPINK1 genomic sequence was obtained from human GRCh37/hg19 ([https://genome.ucsc.edu/](https://genome.ucsc.edu/)). GenBank accession number NM_003122.4 was used as the SPINK1 cDNA reference sequence.

Search for Variants in Strong LD with rs17107315:T>C (c.101A>G) Variant

The search for variants in strong LD with rs17107315:T>C (c.101A>G) variant was performed with HaploReg v4.1 ([http://www.broadinstitute.org/mammals/aphloreg/aphloreg.php](http://www.broadinstitute.org/mammals/aphloreg/aphloreg.php)) [Ward and Kellis 2016], using an LD threshold of $r^2 \geq 0.40$ and querying the 1000 Genomes Project (1000GP) Phase 1 data ([http://www.1000genomes.org/category/phase-1/](http://www.1000genomes.org/category/phase-1/)) in the context of the European population.

Verification and Search for HNF1A Binding Sites

The search for HNF1A binding sites was performed using RSAT ([http://www.rsat.eu/](http://www.rsat.eu/)) [Medina-Rivera et al., 2015] under default conditions, with Homo sapiens GRCh37/hg19 being used as the organism-specific background.

HNF1A–PTF1L Cis-Regulatory Module (CRM) Prediction

We performed a comprehensive review of the literature and collated 10 experimentally validated PTF1L binding sites within the promoters of human, rat or mouse genes that are known to be highly expressed in the pancreatic acinar cells [Beres et al., 2006; Boulling et al., 2011; Holmstrom et al., 2011]. Each of these PTF1L binding sites comprised a 5' E-Box (length = 6) and a 3' TC-Box (length = 7), separated by a 4- or 5-nucleotide spacer sequence [Beres et al. 2006; Boulling et al. 2011]. We first aligned the 10 E-Box and 10 TC-Box
sequences separately to create two distinct position frequency matrices (PFMs) by counting the occurrences of each nucleotide at each position. Then, to construct appropriate PFMs for the PTF1L binding site, we separated the E-Box and TC-Box PFMs by 4 or 5 non-specific nucleotides to create two PTF1L PFMs, termed PTF1_4N and PTF1_5N, respectively. The nucleotide frequency within the spacer was adjusted to correspond to nucleotide frequencies in the human genome at large (A: 0.255, T: 0.267, G: 0.242, C: 0.236). Sequence logos for the two PTF1L PFMs were created with WebLogo (http://weblogo.berkeley.edu) [Crooks et al., 2004].

The above generated PTF1L PFMs and the HNF1A (MA0046.2) PFMs provided by the JASPAR database (http://jaspar.genereg.net) [Mathelier et al., 2016] were then used to calculate their respective position weight matrices (PWMs). This task was achieved with the freely online available RSAT - matrix-scan (full options) tool [Medina-Rivera et al., 2015]. PWMs were generated using default parameters with the exception of the “background model estimation method” that was set to “organism-specific: Homo sapiens GRCh37”.

Transcription factor binding site (TFBS) prediction was performed using the “Individual Matches” mode with default scanning options. “P-value upper threshold” was exceptionally increased to $10^{-3}$ to calculate the matrix score for TFBS of weak relevance (i.e., mutated TFBSs). The SPINK1 locus plus ± 20 kb flanking sequences were analyzed for CRM prediction using the RSAT CRER scanning option under default conditions, except for the following parameters: Lower CRER size = 1, Upper CRER size = 200, site P-value <10^{-4}.

Assessment of Phylogenetic Conservation, Chromatin Accessibility and Histone Marks in the Chromosomal Region of Interest

Phylogenetic conservation data as represented by “Placental Mammal Conservation by PhastCons” and “Placental Mammal Conserved Elements” tracks (both using 46 mammal
species) were directly taken from the UCSC Genome Browser (https://genome.ucsc.edu/) [Kent et al., 2002]. Accessible chromatin regions and histone marks in the pancreatic tissues of two donors were obtained from the website of the NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) [Bernstein et al., 2010].

Search for SNPs affecting the Expression of SPINK1 in the Pancreas Tissue

This analysis was performed using the GTEx dataset available at http://www.gtexportal.org/home/) [Carithers et al., 2015].

Construction of Luciferase Promoter Reporter Plasmids

A fragment spanning -346 to +49 relative to the transcription start site (i.e., c.1-61 in accordance with Yasuda et al. [Yasuda et al., 1998] with the A of the translational initiation codon ATG of the gene being designated as c.1) of the SPINK1 gene was first PCR amplified from a genomic DNA sample. PCR amplification was performed by means of the HotStarTaq Master Mix Kit (Qiagen) with primers (Supp. Table S1) designed to be used with the In-Fusion® HD Cloning Kit (Clontech, Saint-Germain-en-Laye, France), as previously described [Boulling et al., 2015]. The resulting pGL3 reporter construct harboring the wild-type sequence of the human SPINK1 proximal promoter upstream of the firefly luciferase gene was termed hSPINK1pp. The same strategy was used to insert a 330 bp DNA fragment containing HNF1A–PTF1L CRM5 into hSPINK1pp at a position downstream of the luciferase gene. This was achieved using the primers described in Supp. Table S1, after plasmid digestion with BamHI. This latter construct was termed hSPINK1pp+E. All SPINK1 promoter and enhancer variants were then generated from their respective wild-type constructs by site-directed mutagenesis using the Quick Change Site Directed Mutagenesis Kit (Stratagene, Massy, France). All resulting plasmids were checked by Sanger sequencing.
Construction of Transcription Factor Expression Plasmids

Human pancreas cDNAs were obtained from 1 µg human pancreas total RNA (Amsbio) by reverse transcription using the SuperScript II Reverse Transcriptase (Life Technologies) and 20mer-oligo(dT) primer (Eurogentec, Angers, France). The obtained cDNAs were treated with 2 U RNase H (Life Technologies) at 37°C for 20 min before being used for PCR amplification of the coding sequences of the human RBPJL and HNF1A genes, respectively. In parallel, the human PTF1A cDNA clone (OriGene, Rockville, MD) was used to amplify the coding sequence of the human PTF1A gene. PCRs were carried out by using the KAPA HiFi DNA Polymerase (Kapa Biosystems, Wilmington, MA) with the respective primers described in Supp. Table S1. Specifically, 30 (for PTF1A) or 33 (for both RBPJL and HNF1A) cycles of amplification were performed, employing an annealing temperature of 65°C. The three resulting PCR fragments were purified on an 1.5% agarose gel and cloned into the pcDNA™3.1/V5-His-TOPO® (Life Technologies) vector after addition of 3’ A-overhangs by Taq DNA Polymerase (Qiagen). The expression constructs thus obtained, which carried the coding sequences of the human PTF1A, RBPJL and HNF1A genes, were termed pcDNA3.1-PTF1A, pcDNA3.1-RBPJL and pcDNA3.1-HNF1A, respectively. The orientation and sequence of each insert were checked by sequencing. Plasmids were produced using the Nucleobond Xtra Midi EF Kit (Macherey-Nagel).

Cell Culture, Quantitative RT-PCR Analyses, Co-Transfection Transactivation Experiments, Luciferase Reporter Gene Assay, and Electrophoretic Mobility Shift Assay (EMSA)

These procedures are described in Supp. Methods.
Analysis of rs142703147:C>A and rs17107315:T>C (c.101A>G) in French, Chinese and Indian Subjects

The procedures for sequencing the two polymorphic sites are described in Supp. Methods.

The LD between the two SNPs was calculated by means of CubeX (http://www.oeg.org/software/cubex/) [Gaunt et al., 2007]. To test for the effect of one SNP being conditional upon the other, a logistic regression model was used where the log(odds) of disease of each individual $i$ was modeled as a linear function of the minor allele dosage $g_{i,k}$ at each SNP $k$ ($g_{i,k} = 0, 1$ or 2) and the additive effect of this minor allele $\beta_k$. An indicator variable $\delta_i$ was added to the model to account for the geographic origin of the individuals. The full model (1) with the effects of the two SNPs was compared against each restricted model with respectively $\beta_1$ and $\beta_2$ set to zero to test for the effect of each SNP conditional upon the other:

$$\log(\text{odds}_i) = \theta + \beta_1 g_{i,1} + \beta_2 g_{i,2} + \gamma \delta_i \quad (1)$$

The significance of the improvement in fit was tested by comparing the difference of deviances to a distribution with 1 degree of freedom. The glm() function of R version 3.2.2 was used for fitting the model [R Core Team, 2015].

Results

Search for Variants in Perfect or Strong LD with rs17107315:T>C (c.101A>G)

A causal variant residing within an uncharacterized flanking region of the SPINK1 gene has been hypothesized to underlie the SPINK1 c.101A>G variant-associated haplotype [Kereszturi et al. 2009]. Such a variant should in principle be in perfect (or at least very strong) LD with the SPINK1 c.101A>G variant and would be predicted to impact the binding site for a functionally relevant transcription factor. We tested this postulate by means of HaploReg v4.1 [Ward and Kellis 2016], using an LD threshold of $r^2 \geq 0.40$ and querying the
1000GP Phase 1 data in the context of the European population. We identified a total of 25 SNPs in LD with rs17107315:T>C (c.101A>G), whose $r^2$ values ranged from 0.87 to 1; all but one of these SNPs were located within the region spanning 20 kb 3' of SPINK1 to 18 kb 5' of SPINK1 (Supp. Fig. S1). Of the HaploReg v4.1-annotated motifs that were altered by these SNPs, only the HNF1A motif, impacted by rs17107287:C>T (Supp. Fig. S1), was deemed to be of potential functional interest owing to the known role of HNF1A in pancreatic exocrine physiology [Boulling et al. 2011; Molero et al., 2012]. However, we were unable to validate this prediction using RSAT under default conditions [Medina-Rivera et al. 2015].

Identification of Two SNPs that Potentially Disrupt a PTF1L Binding Site by Visual Inspection

Assuming that the hypothesis that the causal variant resides within a flanking region of the SPINK1 gene [Keresztfanger et al. 2009] was nevertheless correct, and that we had successfully identified all the SNPs in strong or perfect LD with rs17107315:T>C (c.101A>G) variant, the most probable reason for failing to confirm our prediction was deemed to be that a functionally relevant transcription factor binding site (TFBS) had been missed by the relevant search programs. We previously encountered just such a case during the functional characterization of SPINK1 promoter variants; an HNF1A binding site was readily predicted by MATCH (http://www.gene-regulation.com/pub/programs.html#match) but a PTF1L binding site was identified instead by visual inspection [Boulling et al. 2011] (Fig. 1A). PTF1L is a pancreatic-specific trimeric complex comprising PTF1A, RBPJL and one of the several ubiquitously expressed class A bHLH family members [Boulling et al. 2011; Holmstrom et al. 2011; Masui et al., 2008]. We therefore visually inspected the local DNA sequence spanning all the aforementioned 25 SNPs (Supp. Fig. S1) against the previously described canonical sequence of PTF1L TFBS, CACCTG…TTCCC [Boulling et al. 2011].
We identified two SNPs, rs142703147:C>A (Fig. 1B) and rs192858015:G>A, to disrupt a putative PTF1L TFBS.

Using a Putative HNF1A−PTF1L CRM to Filter rs142703147:C>A and rs192858015:G>A

There is increasing evidence that the spatial and temporal expression of genes is enabled by the coordinated action of multiple transcription factors through CRMs [Lelli et al., 2012]. Moreover, distal CRMs, also called enhancers, can often be predicted by sequence signatures extracted from proximal promoters [Taher et al., 2013]. Given the important roles played by HNF1A and PTF1L in adult pancreatic acinar cells [Holmstrom et al. 2011; Masui et al. 2008; Molero et al. 2012], it did not appear unreasonable to speculate that their closely spaced TFBSs within the SPINK1 proximal promoter (Fig. 1A) could define such a CRM. We therefore screened the ±200 bp sequences flanking rs142703147:C>A and rs192858015:G>A by means of RSAT [Medina-Rivera et al. 2015] and identified a putative HNF1A TFBS only in the immediate vicinity of rs142703147:C>A (Fig. 1B). In other words, of the two SNPs, only rs142703147:C>A occurred within a putative HNF1A−PTF1L CRM. rs142703147:C>A corresponds to c.1-4141G>T in accordance with the A of the translational initiation codon ATG of the SPINK1 gene being designated as c.1 [den Dunnen et al., 2016]. This variant will be described as rs142703147:C>A (c.1-4141G>T) in the following sections.

In Silico Evidence Supporting the Functional Relevance of the rs142703147:C>A (c.1-4141G>T)-Affected HNF1A−PTF1L CRM in Vivo

To provide supporting evidence that the rs142703147:C>A (c.1-4141G>T)-affected HNF1A−PTF1L CRM is of functional relevance in vivo, we sought to compare it with other nearby “HNF1A−PTF1L” CRMs in terms of both evolutionary conservation, chromatin
accessibility and histone marks [Shlyueva et al., 2014]. To this end, we first built PFM.s for the bipartite PTF1L motif, with the E-Box and TC-Box being separated respectively by 4 bp and 5 bp (Fig. 2A,B). We then converted the PTF1L PFM.s and JASPAR-derived HNF1A PFM.s to their respective PWM.s and scanned the SPINK1 locus plus ±20 kb flanking regions for putative HNF1A–PTF1L CRMs by means of RSAT [Medina-Rivera et al. 2015]. In addition to the HNF1A–PTF1L CRM illustrated in Fig. 1A (termed CRM4) and that illustrated in Fig. 1B (termed CRM5), four additional putative CRMs were identified (i.e., CRMs 1, 2, 3 and 6; Fig. 1C). It is pertinent to mention that two putative PTF1L binding sites were also predicted immediately upstream of the HNF1A binding site in the proximal promoter (see Fig. 1A). However, these PTF1L binding sites were excluded from further consideration because neither of them was located within an evolutionarily conserved region.

Of the 6 CRMs, only CRM4 and the rs142703147:C>A (c.1-4141G>T)-affected CRM5 were found to be located within the most evolutionarily conserved regions. Further, each of the top three chromatin accessible regions obtained from human pancreatic tissue contains a HNF1A–PTF1L CRM (CRMs 4-6) (Fig. 1C), indicating their likely functional importance in vivo. Here it is important to emphasize that, with the exception of rs142703147:C>A (c.1-4141G>T), all the other SNPs in LD with rs17107315:T>C (c.101A>G) were not found within a HNF1A–PTF1L CRM, and fell outside of the most accessible chromatin and the most evolutionarily conserved regions (Fig. 1C).

Using data from the website of the NIH Roadmap Epigenomics Mapping Consortium, we did not find any strong histone 3 lysine 4 monomethylation (H3K4me1) or H3K27 acetylation (H3K27ac) marks across the SPINK1 locus plus ±20 kb flanking regions in the human pancreatic tissues of two healthy donors (Supp. Fig. S2). In addition, no SNPs were found to affect the expression of SPINK1 in human pancreas in the GTEx database.
Co-Transfection Transactivation Experiments Demonstrating Functional Synergy between the HNF1A and PTF1L Transcription Factors in Regulating the Promoter Activity of the \textit{SPINK1} Gene

The premise of the CRM concept is that multiple transcription factors cooperate in regulating gene expression through concomitant binding to their cognate binding sites in a gene’s regulatory sequence. We therefore tested the functional synergy between the HNF1A and PTF1-L transcription factors in regulating the promoter activity of the \textit{SPINK1} gene by means of a co-transfection transactivation assay, which was performed essentially as previously described [Holmstrom et al. 2011]. To this end, we constructed a promoter-reporter vector wherein a CRM4-containing \textit{SPINK1} promoter sequence was cloned into the pGL3 basic vector upstream of the firefly luciferase gene (termed hSPINK1pp) and three pcDNA3.1 expression vectors containing the coding sequences of the \textit{PTF1A}, \textit{RBPJL} and \textit{HNF1A} genes, respectively. The co-transfection transactivation assay had to be performed in a cell line that lacked endogenous expression of these transcription factors; the non-pancreatic HEK293T cell line, which was confirmed to lack expression of the three transcription factor genes as well as the \textit{SPINK1} gene by quantitative RT-PCR analysis (Supp. Fig. S3A), was used for this purpose. A further stipulation was that none of the co-transfected transcription factors should induce spurious expression of the luciferase reporter gene through binding to non-specific sequences within the vector. We verified this by co-transfecting the promoter-lacking pGL3-basic vector with the different transcription factor expression plasmids, either individually or in combination, and found no significant increase of the luciferase reporter gene under any conditions (Supp. Fig. S3B).

Having validated the experimental system, we first co-transfected the hSPINK1pp vector with the expression vectors containing the three transcription factors. It should be appreciated that the function of the PTF1L transcription factor is executed by a combination of PTF1A,
RBPJL and one of the several ubiquitously expressed class A bHLH family members [Boulling et al. 2011; Holmstrom et al. 2011; Masui et al. 2008]. SPINK1 promoter reporter gene activity increased 1.8-fold and 1.3-fold upon expression of HNF1A alone and PTF1A + RBPJL alone, respectively, but increased 6.3-fold upon expression of all three (Fig. 3). These observations were interpreted in terms of a synergistic effect between the HNF1A and PTF1-L transcription factors.

Two variants residing within the CRM4 HNF1A binding site in the SPINK1 promoter have been reported to predispose to chronic pancreatitis (Fig. 1A). Both variants were predicted to disrupt the HNF1A binding site (Fig. 4A). We tested their potential effects on the cooperative action between the HNF1A and PTF1L transcription factors and found that each resulted in the abolition of the aforementioned synergistic effect (Fig. 4B).

These two complementary lines of evidence, taken together, underscore the importance for SPINK1 gene regulation of the coordinated action of HNF1A and PTF1L through binding to their cognate binding sites within the context of a functional CRM.

Co-Transfection Transactivation Assay Testing the Functional Effect of the rs142703147C>A (c.1-4141G>T)-Variant on Regulating the Promoter Activity of the SPINK1 Gene

The rs142703147:C>A (c.1-4141G>T)-variant was predicted to have a lower PWM score than the wild-type rs142703147C allele (Fig. 5A), suggestive of reduced affinity for the PTF1L transcription factor. The current ‘gold standard’ in vitro method to evaluate the effect of an enhancer element is to place it in the vicinity of a promoter element in a reporter gene assay [Shlyueva et al. 2014]. We therefore inserted a fragment containing the HNF1A−PTF1L CRM5 into hSPINK1pp at a position downstream of the luciferase gene; the resulting hSPINK1pp+E [C] vector was then used to introduce the rs142703147A variant. The
hSPINK1pp+E [C] vector and the corresponding rs142703147A vector, hSPINK1pp+E [A], were co-transfected respectively with the three transcription factor expression plasmids. The major C allele, but not the minor A allele, of rs142703147 significantly enhanced reporter gene expression induced by the three co-transfected transcription factors in HEK293 cells (Fig. 5B). This demonstrated that rs142703147C>A (c.1-4141G>T) is a loss-of-function variant, consistent with the known role of the SPINK1 gene in the etiology of chronic pancreatitis.

We also performed this analysis in rat pancreatic acinar AR42J cells treated with dexamethasone. Under our experimental conditions, hSPINK1pp drove a mere 2.3-fold increased expression of the reporter gene as compared with the promoterless pGL3 basic vector and no significant difference was observed between the two alleles of rs142703147 in enhancing hSPINK1pp-driven reporter gene expression (Supp. Fig. S4A). Analysis of the relative mRNA expression levels of the rat Prss1, Ctrc, Spink1, Ptf1a, Rbpj1, Hnf1a genes in the dexamethasone-differentiated AR42J cells indicated poor expression of both the Spink1 and Hnf1a genes (Supp. Fig. S4B).

**EMSA Providing Further Supporting Evidence for the Functional Effect of the rs142703147C>A (c.1-4141G>T) Variant**

We further performed EMSA using nuclear extracts prepared from HEK293T cells transfected with the two expression plasmids encoding respectively the PTF1A and RBPJL genes. This assay demonstrated that rs142703147A disrupted the interaction between the PTF1L transcription factor and its cognate binding site (Fig. 6).

**Additional Data from Population Genetic Analyses**
Analysis of 548 French ICP patients and 562 ethnogeographically-matched controls (Supp. Table S2) showed that rs142703147:C>A (c.1-4141G>T) is in perfect LD with the rs17107315:T>C (c.101A>G) variant in this population ($r^2 = 1$). However, analysis of the two SNPs in 1104 Chinese ICP patients and 1196 controls as well as in 347 Indian patients and 264 controls (Supp. Table S2) showed that they are not in perfect LD in these two populations (Chinese, $r^2 = 0.80$; Indian, $r^2 = 0.59$). Similar OR values were obtained between the rs142703147:A (c.1-4141T) allele and the rs17107315:C (c.101G) allele: 6.13 versus 5.47 in the Chinese dataset (Supp. Table S3) and 15.12 versus 14.82 in the Indian dataset (Supp. Table S4).

To test whether the two SNPs had an impact on disease risk, conditional analyses were performed. Nested logistic regression models were fitted viz. M1, effect of SNP1 (rs142703147:C>A) + ethnicity; M2, effect of SNP2 (rs17107315:T>C) + ethnicity; and M3 (i.e., the full model), effect of SNP1 + effect of SNP2 + ethnicity. By comparing the likelihood of M2 versus M3, we tested for the effect of SNP1 conditional upon SNP2, obtaining a $\chi^2$ value of 5.65 and a $P$ value of $1.74 \times 10^{-2}$. By comparing the likelihood of M1 versus M3, we tested for the effect of SNP2 conditional upon SNP1, obtaining a $\chi^2$ value of 30.65 and a $P$ value of $3.09 \times 10^{-8}$. These results suggested that both SNPs exert an effect on disease risk.

**Discussion**

The association of the rs17107315:T>C (c.101A>G)-containing haplotype with chronic pancreatitis was first described 16 years ago [Witt et al. 2000]. As opined by Kereszturi and colleagues, “The mechanism of action of the [SPINK1] p.Asns43er-associated haplotype remains one of the most intriguing, unsolved questions of pancreas genetics” [Kereszturi et al. 2009]. These authors proposed that the causal variant was most probably located within an
uncharacterized flanking region of the \textit{SPINK1} gene. However, distal regulatory variants are notoriously difficult to identify because the causal variant-harboring regulatory elements act independently of the distance and orientation to their target genes [Mathelier et al., 2015; Shlyueva et al. 2014]. Herein we have related a somewhat unusual and rather atypical story about how a functional regulatory variant was finally identified after a 16 year interlude. We started out by adopting a hypothesis-driven approach to identify all variants in strong or perfect LD with the rs17107315:T>C (c.101A>G) variant. Somewhat surprisingly, none of the LD SNPs were predicted to affect a functionally relevant TFBS. However, we identified two candidate SNPs that affected potential PTF1L binding sites by visual inspection of DNA sequence spanning all 25 LD SNPs, guided by prior knowledge of pancreas-specific transcription factors and their cognate binding sites. With hindsight, the failure to identify potential PTF1-L binding sites by all currently available TFBS prediction programs was almost certainly due to the variable length of the spacer sequence that separates the E-Box and TC-Box of the bipartite PTF1-L binding site (see Fig. 2).

Employing a novel CRM-based approach to filter the aforementioned two SNPs, we excluded one of them from further consideration. The remaining single variant, rs142703147C>A (c.1-4141T), is located approximately 4-kb from the \textit{SPINK1} promoter, consistent with the current consensus that enhancers tend to be located within 10 kb of their associated transcription start sites genome-wide [MacIsaac et al., 2010; Taher et al. 2013]. The functionality of this variant was strongly supported by evolutionary conservation and chromatin accessibility data (Fig. 1C). By contrast, based upon ChIP-seq data from the human pancreatic tissues of two donors, no strong enhancer-associated histone marks (i.e., H3K4me1 and H3K27ac) were noted in the vicinity of the rs142703147C>A (c.1-4141T)-affected motif (Supp. Fig. S2). However, it is known that (i) none of the known histone modifications correlate perfectly with enhancer activity and (ii) there is no evidence that either H3K4me1 or
H3K27ac is sufficient, necessary or even mechanistically involved in transcription [Shlyueva et al., 2014]. Additionally, in the publicly available GTEx dataset, no single SNP is known to influence SPINK1 expression in the pancreas. As far as rs142703147C>A (c.1-4141T) is concerned, this variant might simply not have been included for analysis by GTEx due to its low allele frequency in normal populations.

The functionality of the rs142703147C>A (c.1-4141T) variant was further supported by a series of experiments performed in HEK293T cells (Figs. 3-6). In this regard, it should be emphasized that there are no human pancreatic cell lines currently available for performing SPINK1 promoter or enhancer reporter gene assays on a physiologically relevant background. The co-transfection transactivation assay used here, performed in cells lacking endogenous expression of the relevant transcription factors [Holmstrom et al. 2011], overcame this technical limitation. This notwithstanding, one may surmise that dexamethasone-differentiated rat pancreatic acinar AR42J cells [Rajasekaran et al., 1993], which have been previously used for analyzing PRSS1 [Boulling et al., 2015] and SPINK1 [Derikx et al., 2015] promoter variants, may be relevant with respect to the current ‘gold standard’ in vitro method for evaluating the effect of an enhancer element placed in the vicinity of a promoter element in a reporter gene assay [Shlyueva et al. 2014]. We therefore performed this analysis in AR42J cells treated with dexamethasone but did not obtain expected results (Supp. Fig. S4A). This can be essentially accounted for by the poor expression of both the Spink1 and Hnf1α genes in the AR42J cells treated with dexamethasone (Supp. Fig. S4A). Here it is pertinent to note that in the current study, the inserted SPINK1 promoter drove a mere 2.3-fold increased expression of the reporter gene as compared with the promoterless pGL3 basic vector (Supp. Fig. S4B) whilst in a previous reporter gene assay, the corresponding increase for the inserted SPINK1 promoter was >15-fold [Derikx et al., 2015]. A variety of parameters affecting cell characteristics that pertain to cell culture conditions, including medium used, fetal bovine...
serum percentage and composition, cell confluence, number of cell passages at time of transfection and protocol for dexamethasone treatment, could have significantly affected the experimental outcomes [Baker et al., 2016]. Additionally, the \( \text{SPINK1} \) promoter sequence used for reporter gene assay differs between the current study and the previous study [Derikx et al., 2015]; the inserted promoter segments correspond to c.1-407 to c.1-13 and c.1-541 to c.35 of the \( \text{SPINK1} \) genomic sequence, respectively. Finally, we should add that we did not attempt to perform experiments in mouse-derived pancreatic acinar tumor 266-6 cells because these cells displayed no difference with HEK293T cells in terms of \( \text{SPINK1} \) promoter-driven reporter gene expression [Derikx et al., 2015].

Although we provide strong \textit{in silico} and \textit{in vitro} evidence that the rs142703147C>A (c.1-414T) variant is of functional significance, our findings from population genetic studies clearly suggest that it is only one component of the chronic pancreatitis-predisposing functional elements contained within the risk haplotype of interest. Thus, we are still far from obtaining a complete understanding of the pathogenic mechanism(s) underlying the most important heritable risk factor for idiopathic chronic pancreatitis identified to date [Witt et al., 2000], even after a 17-year interlude. Indeed, even though a \textit{cis} variant located in the immediate vicinity of the gene under study would be a priority in terms of being tested, the true causative variant can be located at some distance from the haplotype associated with the phenotype (Smemo et al., 2014). Further studies that aim to discover other variants contributing to the high-risk haplotype as well as to improve the characterization of the functional SNP identified here are warranted. In a more general context, our case serves to exemplify the difficulties that are frequently encountered in tracking down and unmasking the causal variants responsible for disease associations that reside within the extensive regulatory regions flanking our genes rather than within the gene coding regions themselves [Mathelier et al. 2015; Spielmann and Mundlos 2016; Yao et al., 2015]. Nonetheless, the novel approach
employed in this study will, we believe, help to stimulate the development of new strategies to identify the causal regulatory variants underlying many human inherited disease associations.
References


Beres TM, Masui T, Swift GH, Shi L, Henke RM, MacDonald RJ. 2006. PTF1 is an organ-specific and Notch-independent basic helix-loop-helix complex containing the mammalian Suppressor of Hairless (RBP-J) or its parologue, RBP-L. Mol Cell Biol 26:117-130.


Boulling A, Chen JM, Callebaut I, Férec C. 2012. Is the SPINK1 p.Asn34Ser missense mutation per se the true culprit within its associated haplotype? WebmedCentral GENETICS 3:WMC003084.


**Figure legends**

**Figure 1.** Discovery of a candidate causal variant underlying the chronic pancreatitis-associated \(SPINK1\) c.101A>G (rs17107315:T>C) variant-containing haplotype. **A:** The HNF1A and PTF1-L binding sites previously identified within the \(SPINK1\) proximal promoter [Boulling et al. 2011]. This motif signature corresponds to CRM4 illustrated in **C.** Note that (i) the nucleotide positions are in accordance with the A of the translational initiation codon ATG of the \(SPINK1\) gene being designated as c.1; (ii) the sequence given is on the sense strand with respect to the reading frame of the \(SPINK1\) gene; (iii) HNF1A and PTF1L were previously termed HNF1 and PTF1, respectively [Boulling et al. 2011]; and (iv) the TC-Box of the PTF1L binding site was previously annotated as comprising 6 nucleotides [Boulling et al. 2011]. Two chronic pancreatitis-predisposing variants that occurred within the HNF1A binding site, c.147A>G and c.142T>C [Boulling et al. 2011], are also shown. **B:** Illustration of the bipartite PTF1L TFBS disrupted by rs142703147C>A (c.1-4141G>T) and the RSAT-predicted HNF1A TFBS. This panel represents an enlarged view of CRM5 illustrated in **C.** Nucleotide positions are in accordance with hg19, with the A of the translational initiation codon ATG of the \(SPINK1\) gene being designated as c.1. It should be noted that the sequence is given on the antisense strand with respect to the reading frame of the \(SPINK1\) gene. N55 indicates 55 nucleotides whose sequence is not shown. **C:** Evaluation of the predicted putative PTF1L−HNF1A CRMs within the \(SPINK1\) locus plus ± 20 kb flanking sequence in the context of phylogenetic conservation and accessible chromatin regions. PC and CE, Placental Mammal Conservation by PhastCons and Placental Mammal Conserved Elements obtained from the UCSC Genome Browser. Accessible DNA regions in the pancreatic tissues of two donors, as determined by DNase-seq, were obtained from the website of the NIH Roadmap Epigenomics Mapping Consortium. LD SNPs refer to all the SNPs (with the exception of the below described rs138251740A>G; see Supp. Fig. S1) that
were found to be in strong or perfect LD with the rs17107315T>C (c.101A>G) variant (highlighted in green). Apart from rs17107315T>C (c.101A>G) and rs142703147C>A (c.1-4141G>T), the other two SNPs that were specifically mentioned in the manuscript (i.e., rs17107287C>T and rs192858015G>A) are also clearly indicated. Note that the not shown rs138251740A>G (located further downstream of chr5:147,230,000) is located neither within a putative PTF1L−HNF1A CRM nor within a region showing strong evolutionary conservation and high chromatin accessibility.

**Figure 2.** Construction of two PFMs for the bipartite PTF1L binding site. **A:** Sequence alignment of 10 experimentally validated PTF1L TFBSs. Each of these PTF1L TFBSs comprised a 6-bp E-Box motif and a 7-bp TC-Box motif, separated by a 4- or 5- nucleotide spacer sequence (4N or 5N). The aligned E-Box and TC-Box sequences were used to generate two distinct PFMs. Nucleotides that are not perfectly conserved within the E-Box or TC-Box are highlighted in red. Luc., luciferase reporter gene assay. **B:** Sequence logos for the two sets of PTF1L PFMs, PTF1_4N and PTF1_5N. They were generated by inserting a 4- or 5- nucleotide spacer sequence between the aforementioned E-Box and the TC-Box PFMs.

**Figure 3.** Induction of luciferase reporter gene activity driven by the CRM4-containing SPINK1 promoter (hSPINK1pp) upon expression of the co-transfected transcription factors. Expression level of the SPINK1 promoter-driven luciferase reporter gene co-transfected with the empty pcDNA3.1 (+) vector (Empty) is set to 1. H, HNF1A; P, PTF1A; R, RBPJL. Bars, SD. *, $P<0.05$; **, $P<0.01$; ***, $P<0.001$.

**Figure 4.** Effects of two chronic pancreatitis-predisposing SPINK1 promoter variants on HNF1A- or PTF1L−HNF1A-induced luciferase reporter gene activity. **A:** Predicted effects of
the two chronic pancreatitis-predisposing variants [Boulling et al. 2011] on the HNF1A binding site located within CRM4 in the SPINK1 proximal promoter. The PWM scores are shown for each of the wild-type (WT) and variant sequences. A P value of $< 10^{-4}$ was regarded as being a potential TFBS. *, $P < 10^{-3}$; **, $P < 10^{-5}$. B: Co-transfection transactivation experiments performed under different conditions. Expression levels of the luciferase reporter gene activity driven by the CRM4-containing wild-type, c.142T>C or c.147A>G SPINK1 promoter co-transfected with the empty pcDNA3.1 (+) vector (Empty) is set to 1. Each of the wild-type and variant SPINK1 promoter reporter gene vectors was co-transfected with the expression plasmid encoding the HNF1A gene only (HNF1A) and the three expression plasmids encoding respectively the PTF1A, RBPJL and HNF1A genes (P+R+H). Bars, SD. **, $P<0.01$; ***, $P<0.001$.

**Figure 5.** Functional effect of the rs142703147C>A (c.1-4141G>T) variant on regulating SPINK1 promoter activity by means of a co-transfection transactivation assay. A: Predicted effect of the rs142703147C>A (c.1-4141G>T) variant on the PTF1L binding site located within CRM5. The PWM scores are shown for the wild-type and variant alleles of rs142703147. A P value of $< 10^{-4}$ was regarded as being a potential TFBS. **, $P < 10^{-4}$; ***, $P < 10^{-5}$. B: Effects of the wild-type and variant CRM5 with respect to the rs142703147 polymorphic site, when inserted separately into hSPINK1pp, on PTF1A+RBPJL+HNF1A (P+R+H)-induced reporter gene expression. Expression levels of the reporter gene constructs co-transfected with the empty pcDNA3.1 (+) vector are set to 1. hSPINK1pp+E [C], wild-type CRM5 inserted into the hSPINK1pp vector; hSPINK1pp+E [A], rs142703147A-containing CRM5 inserted into the hSPINK1pp vector. Bars, SD. ***, $P<0.001$. 
Figure 6. Functional characterization of rs142703147C>A (c.1-4141G>T) by EMSA. Upper panel shows the sequence of the biotinylated probe and specific competitor (Comp.) with respect to the wild-type C allele of rs142703147 (first line) and the sequence of the variant rs142703147A allele competitor (second line). Only the sense strand of the double stranded oligonucleotide is shown. Lower panel shows the EMSA results performed with labelled probe C incubated with nuclear extracts (NE) from HEK293T cells transfected with empty pcDNA3.1 vector (E) or with PTF1A (P) and RBPJL (R) expression plasmids, in the presence or absence of Competitor C, Competitor A or Irrelevant (Ir) competitor. Competitors were added at a 50- or 100-fold excess compared with the biotinylated probe. The arrow indicates the position of the DNA/PTF1L complex.