Multi-ancestry Fine Mapping of Interferon Lambda and the Outcome of Acute Hepatitis C Virus Infection

Supplementary Material

Supplementary Materials and Methods

IFNL Sequencing: The first four amplicons (fragments A-D) and the second four amplicons (fragments E-H) were sequenced separately, to allow for unambiguous assignment of reads to one half of the region or the other. This allowed alignment of reads specifically to the region of origin, resulting in more confident detection of individual variants across the whole region.

Genomic DNA derived from peripheral blood was used for this analysis. DNA was extracted using DNeasy Blood & Tissue Kit (Quiagen, Germantown, USA) following manufacturer's recommendation with no modifications. Libraries were prepared using NEBNext® Ultra™ DNA Library Prep Kit (Illumina®, San Diego, CA), with dual indexing to allow for sufficient barcodes. The resulting libraries were quantified with Kapa Illumina quantification kit, then sequenced on an Illumina Miseq using 2x300 base pairs (bp) reads (600v3) cartridge (Illumina®, San Diego, CA).

Sequence alignment. Reference sequences for either half of the region (Amplified Fragments A-D and E-H) were generated by sub-setting the GRCh37 human genome sequence. De-multiplexed FASTQ files were aligned to their respective sequences using bowtie2 (1). The resulting alignment files were then sorted with SAMtools (2), and groups of reads in each fragment were added using bamaddrg [https://github.com/ekg/bamaddrg]. Finally, Freebayes was used to generate variant call files (vcf) files for the A-D and E-H amplified fragments regions respectively (https://arxiv.org/abs/1207.3907).

Conditional Analysis of an independent imputed dataset.

Genotyping and Imputation: Genotyping was done using the Illumina Human Omni-Quad array (Illumina, San Diego, CA) and imputation was performed for using the Minimac3 software (3) through the publicly available Michigan Imputation Server (4) as described in detail elsewhere (5).

Identification of potential causal variants

PAINTOR estimates posterior probabilities of a variant being functional allowing multiple functional variants at the risk locus. Using consistent alleles for the two ancestry populations, we calculated Z scores based on the Wald Statistic (β/SE(β)) obtained from the logistic regression analysis of all individuals in each population. An LD matrix was calculated based on the genotype data for each ancestry group. We integrated the primary functional categories (coding, UTR, promoter, enhancer, DNase-hypersensitivity site, intronic and intergenic) proposed by Gusev *et al* (6). The annotation matrix contained data from the ENCODE project (34) for the HepG2 cell line as well as coding information accessed from the UCSC Genome Browser using the Table Browser tool (7) and ANNOVAR (8). We set the number of functional variants to 2, 3, 4 or 5 based on feasible running time.

The credible set was constructed by ranking the functionally predicted variants based on their posterior probabilities and then selecting variants from the top down to reach a sum of posterior probabilities of at least 0.99 value (for a credible set of 99%).

To investigate functional elements, the presence or absence of overlap was determined by the UCSC Table Browser intersecting the calculated credible set with the signal tracks. We used information from the ENCODE database consisting of a common set of states across the HepG2 cell line learned by computationally integrating ChIP-seq data for 8 chromatin marks, input data and the CTCF transcription factor, two DNase-seq assays and a FAIRE-seq assay, using a Hidden Markov Model (HMM condensed in the Chrom HMM Segmentations track) (9). We also identified CpG methylation sites in Hepatocytes and HepG2 cells and liver tissue using ENCODE data accessed through the UCSC Genome Browser (10).

Analysis of functionally relevant variants

Rs4803217 (C>A) is located in the *INFL3*-3'UTR region and in high LD with rs368234815 in populations of different ancestry from The Thousand Genomes Project (r^2=0.73, r^2=1, in YRI and CEU groups, respectively) (11). Rs4803217-C allele decreases HCV-induced degradation of IFNL3 mRNA in vitro. Rs1176648444 located in the IFNλ4 protein, causes a proline to serine substitution at amino acid 70 (P70S) on a haplotype with rs368234815ΔG (12). The substitution substantially alters its antiviral activity with reduced function represented by lower interferon-stimulated gene (ISG) expression levels (13).

References

(1) Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012 Mar 4;9(4):357-359.

(2) Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug 15;25(16):2078-2079.

(3) Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012 Jul 22;44(8):955-959.

(4) Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016 Oct;48(10):1284-1287.

(5) Vergara C, Thio CL, Johnson E, Kral AH, O'Brien TR, Goedert JJ, et al. Multi-Ancestry Genome-Wide Association Study of Spontaneous Clearance of Hepatitis C Virus. Gastroenterology 2018 Dec 26.

(6) Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet 2014 Nov 6;95(5):535-552.

(7) Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 2004 Jan 1;32(Database issue):D493-6.

(8) Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010 Sep;38(16):e164.

(9) ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012 Sep 6;489(7414):57-74.

(10) Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002 Jun;12(6):996-1006.

(11) Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015 Oct 1;526(7571):75-81.

(12) Prokunina-Olsson L, Muchmore B, Tang W, Pfeiffer RM, Park H, Dickensheets H, et al. A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. Nat Genet 2013 Feb;45(2):164-171.

(13) Terczynska-Dyla E, Bibert S, Duong FH, Krol I, Jorgensen S, Collinet E, et al. Reduced IFNlambda4 activity is associated with improved HCV clearance and reduced expression of interferon-stimulated genes. Nat Commun 2014 Dec 23;5:5699.

Supplementary Tables

Fragment	Forward primer '5->3'	Reverse primer '3->5'	Length (bp)	Genetic coordinates (GRCh37/hg19)	
				Start	End
A	ATTCTGATCACTAGTTCCAGGC	TGGCCAGTACTAGTCTCCATATC	9177	39721399	39730575
B	AGATATGGAGACTAGTACTGGCC	AGCTCTGATGTTGGGAAAG	9257	39730552	39739596
C	GACAGGAACGGGTGTATG	ATAGCAGCATGTGAGTCTTT	8696	39738997	39747691
D	AGTTGCTGGTCGGGTAGATC	TGTGAGGACTTTAACCCACGG	7814	39746836	39754649
E	AAGTGTCTCGGTTCATTCCTAG	TCTTTGTCCCGTACACCTGTCCTGG	9402	39754490	39763875
F	AGCTGGCCACCTGAGAATCTTGAG	TCGACGAGTTCTTGGGAAAC	8550	39763708	39772257
G	AACCGGCAACGACCCGCTCAGTG	TAGGGAACTCCTTATTCGCTGGG	13061	39771809	39784869
H	ATGTAGAAGTCGCCCGAGAATTGAC	GGCTCCGCCTTTGCCAAGCTCTG	8348	39783937	39792284

Supplementary Table 1. Customized primers used for sequencing of the targeted fragments in the the *IFNL* region.

Variants in Haplotype in European Ancestry Population												Haplotype Frequency			Effect	
Haplotype	**rs8105790**	**rs8107030**	**rs12971396**	**rs4803221**	**rs368234815**	**rs4803222**	**rs66531907**	**rs12983038**	**rs8109889**	**rs8099917**	**rs7248668**	**All Individuals**	**Clearance**	**Persistence**	**OR (95% CI)**	**P value**
H1	T	A	C	C	TT	G	C	G	C	T	G	0.68	0.77	0.62	1.00	4.36x10^{-22}
H2	T	A	C	C	ΔG	G	C	G	C	T	G	0.02	0.01	0.02	0.69 (0.39-1.22)	0.533
H3	C	G	G	G	ΔG	C	A	A	T	G	A	0.18	0.11	0.23	0.36 (0.29-0.44)	1.97x10^{-20}
H4	T	A	C	C	ΔG	C	C	G	C	T	G	0.10	0.08	0.11	0.62 (0.48-0.79)	0.029
Variants in Haplotype in African Ancestry Population												**All Individuals**	**Clearance**	**Persistence**	**OR (95% CI)**	**P value**
H1	-	-	C	C	TT	G	C	G	C	T	G	0.37	0.50	0.34	1	1.48x10^{-14}
H2	-	-	C	C	ΔG	G	C	G	C	T	G	0.36	0.29	0.37	0.52 (0.42-0.64)	1.81x10^{-04}
H3	-	-	G	G	ΔG	C	A	A	T	G	A	0.06	0.04	0.07	0.40 (0.26-0.62)	0.017
H4	-	-	C	C	ΔG	C	C	G	C	T	G	0.07	0.05	0.08	0.42 (0.28-0.62)	0.016
H5	-	-	G	G	ΔG	C	A	A	T	T	G	0.12	0.10	0.12	0.53 (0.39-0.71)	0.066

Supplementary Table 2. Haplotypes including candidate variants in the *IFNL* region with association in the analysis of the imputed dataset in European and African ancestry individuals. Odds ratios and P values are calculated relative to haplotype H1, which is matched between the two ancestry groups at all variants in common. Abbreviations: OR: Odds ratio; CI: confidence Interval.

Variants in Haplotype in European ancestry Population													Haplotype Frequency		Effect	
Haplotypes	rs4803217	rs12971396	rs117664844	rs4803221	rs368234815	rs4803222	rs66531907	rs12983038	rs8109889	rs8099917	rs7248668	All Individuals	Clearance	Persistence	OR (95% CI)	P value
H1	C	C	G	C	T	G	C	G	C	T	G	0.68	0.77	0.62		1.48x10^{-22}
H2	A	C	G	C	ΔG	G	C	G	C	T	G	0.02	0.01	0.02	0.67(0.37-1.199)	0.48
H3	A	G	G	G	ΔG	C	A	A	T	G	A	0.18	0.10	0.23	0.35(0.29-0.44)	1.79x10^{-20}
H4	A	C	A	C	ΔG	C	C	G	C	T	G	0.10	0.09	0.11	0.64(0.49-0.82)	0.06
Variants in Haplotype in African ancestry Population																
Haplotypes	rs4803217	rs12971396	rs117664844	rs4803221	rs368234815	rs4803222	rs66531907	rs12983038	rs8109889	rs8099917	rs7248668	All Individuals	Clearance	Persistence	OR (95% CI)	P value
H1	C	C	G	C	T	G	C	G	C	T	G	0.37	0.49	0.34		9.46x10^{-14}
H2a	C	C	G	C	ΔG	G	C	G	C	T	G	0.04	0.04	0.04	0.63(0.397-0.99)	0.733
H2b	A	C	G	C	ΔG	G	C	G	C	T	G	0.32	0.25	0.33	0.51(0.41-0.63)	1.48x10^{-04}
H3	A	G	G	G	ΔG	C	A	A	T	G	A	0.06	0.04	0.06	0.40(0.25-0.62)	0.015
H4	A	C	A	C	ΔG	C	C	G	C	T	G	0.07	0.05	0.08	0.42(0.28-0.63)	0.017
H5	A	G	G	G	ΔG	C	A	A	T	T	G	0.12	0.10	0.12	0.53(0.39-0.72)	0.066

Supplementary Table 3. Haplotypes including functionally relevant variants and candidate variants in the IFNL region with association in the analysis of the imputed dataset in European and African ancestry individuals. Candidate variants included in this analysis are common accross ancestry groups. Odds ratios and P values are calculated relative to haplotype H1. Abbreviations: OR: Odds ratio; CI: confidence Interval.

Rsnumber	Allele	Position	Z score African Ancestry	Z score European Ancestry	Posterior Probabilities
rs368234815	ΔG/TT	19:39739155	-7.79	-9.39	0.61
rs12982533	T/C	19:39731904	-5.26	-9.38	0.59
rs10612351	AC/Δ	19:39744807	-5.85	-7.4	0.39
rs4803221	C/G	19:39739129	-3.73	-9.49	0.27

Supplementary Table 4. Posterior probabilities of the variants identified as potential causal *IFNL* region.

SNP				**Analysis of individuals in sequencing panel (N=64)**			**Analysis of imputed data conditioned on rs368234815 and rs1176648444**						
				Counts of Alternative allele			**European Ancestry population (N=1,717)**			**African Ancestry Population (N=1,835)**			**Meta-analysis (N=3,552)**
rsID	**Position**	**Ref**	**Alt**	**Clear**	**Persist**	**Diff.**	**Freq.**	**OR**	**P Value**	**Freq.**	**OR**	**P Value**	**P value**
rs8107090	39721915	T	A	19	24	-5	0.40	0.96	0.59	0.57	0.91	0.46	0.5949
rs35408086	39726810	G	A	11	19	-8	0.39	0.96	0.67	0.21	0.99	0.97	0.6681
rs11883239	39727480	G	A	6	17	-11	0.39	0.96	0.66	0.13	0.86	0.42	0.6601
rs11883201	39727490	A	G	19	24	-5	0.40	0.95	0.59	0.59	0.93	0.59	0.5865
rs955155	39729479	G	A	2	8	-6	0.26	0.86	0.20	0.07	0.93	0.74	0.2033
rs12609937	39731204	A	G	28	35	-7	0.91	0.94	0.68	0.98	0.99	0.97	0.6778
rs115166799	39732212	A	G	12	6	6	N/A	N/A	N/A	0.19	0.94	0.68	N/A
rs8105790	**39732501**	**T**	**C**	**6**	**13**	**-7**	**0.20**	**0.55**	**0.03**	**0.19**	**0.90**	**0.47**	**0.03201**
rs8102358	39735012	G	A	13	7	6	NA	N/A	N/A	0.25	0.98	0.91	N/A
rs8107030	**39736719**	**A**	**G**	**0**	**7**	**-7**	**0.19**	**0.58**	**0.05**	**0.04**	**0.90**	**0.67**	**0.04942**
rs12971396	**39737866**	**C**	**G**	**7**	**13**	**-6**	**0.20**	**0.51**	**0.03**	**0.19**	**0.90**	**0.49**	**0.0273**
rs4803221	**39739129**	**C**	**G**	**7**	**13**	**-6**	**0.20**	**0.42**	**$4.9x10^{-03}$**	**0.19**	**0.89**	**0.42**	**0.004918**
rs73555604	39739170	C	T	12	6	6	0.01	1.67	0.19	0.22	0.96	0.80	0.1864
rs4803222	**39739353**	**G**	**C**	**9**	**14**	**-5**	**0.30**	**0.43**	**0.01**	**0.27**	**0.88**	**0.39**	**0.007982**
rs66531907	39740675	C	A	6	12	-6	0.19	0.60	0.06	0.19	0.86	0.32	0.06032
rs12983038	**39741124**	**G**	**A**	**6**	**11**	**-5**	**0.19**	**0.57**	**0.03**	**0.19**	**0.87**	**0.37**	**0.03114**
rs8109889	39742770	C	T	6	12	-6	0.19	0.64	0.10	0.19	0.86	0.30	0.1002
rs8099917	39743165	T	G	0	5	-5	0.19	0.67	0.11	0.06	0.77	0.23	0.1084
rs7248668	39743821	G	A	0	5	-5	0.19	0.65	0.09	0.06	0.78	0.25	0.08615
rs10853728	39745146	C	G	26	34	-8	0.65	0.91	0.46	0.74	0.86	0.22	0.4645
rs10775535	39745181	C	T	29	34	-5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
rs56116812	39747090	G	A	11	18	-7	0.12	0.97	0.86	0.23	0.93	0.56	0.8593
rs116236518	39749790	C	T	5	0	5	N/A	N/A	N/A	0.02	1.44	0.30	N/A
rs10424607	39749922	A	C	18	23	-5	0.29	1.05	0.81	0.51	1.00	1.00	0.7506
rs251908	39764449	A	G	30	35	-5	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Supplementary Table 5. Variants with a difference ≥ 5 in alternative allele count in sequenced individuals and replication in the meta-analysis of the association test of imputed variants in the IFNL region conditioned on the rs368234815 and rs1176648444 genotypes. Bold text indicates positions with meta-analysis p<0.05 from imputed data.

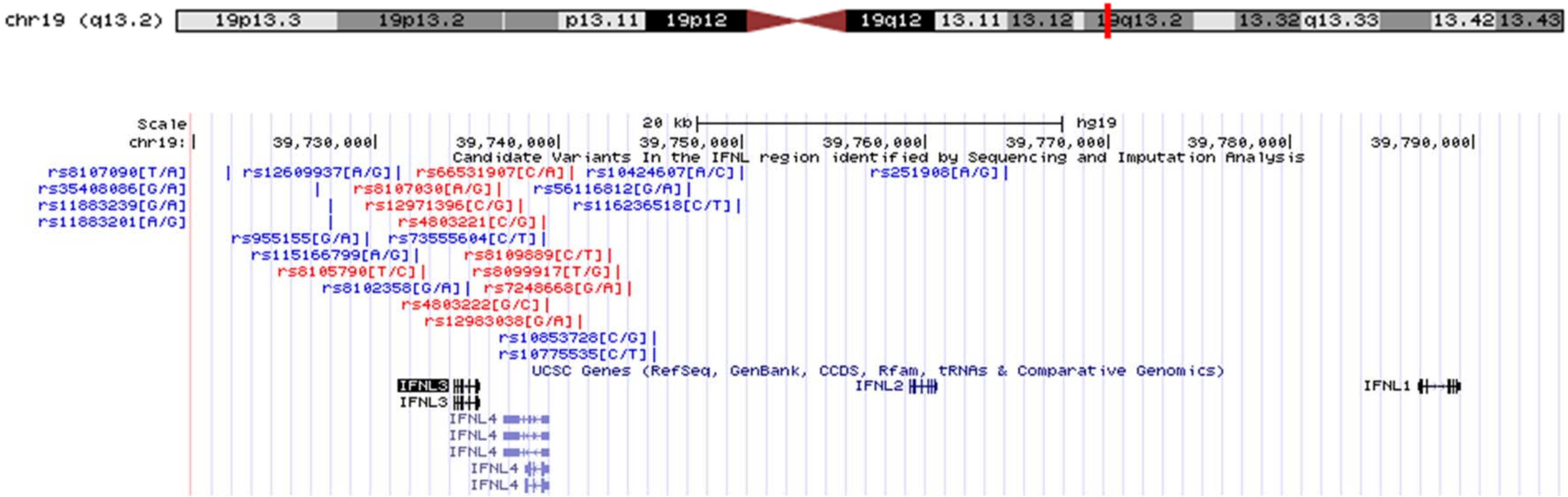
Haplotype	rs368234815	rs4803217	European Ancestry Population					African Ancestry Population				
			Haplotype Frequency			Effect		Haplotype Frequency				Effect
			All	Clear.	Pers.	OR (95% CI)	P value	All	Clear.	Pers.	OR (95% CI)	P value
1	ΔG	A	0.31	0.22	0.37	1	-	0.585	0.454	0.613	1	-
2	TT	A	0.00	0.00	0.00	-	-	0.005	0.010	0.003	-	-
3	ΔG	C	0.01	0.01	0.01	-	-	0.045	0.044	0.045	0.3 (0.18-0.77)	0.005
4	TT	C	0.68	0.78	0.62	2.1 (1.8-2.4)	$<10^{-08}$	0.365	0.492	0.338	1.9 (1.6-2.34)	$<10^{-08}$

Supplementary Table 6. Distribution of rs368234815-rs4803217 haplotypes in European and African population. Abbreviations: OR: odds ratio; clear: clearance; pers: persistence; CI: confidence interval.

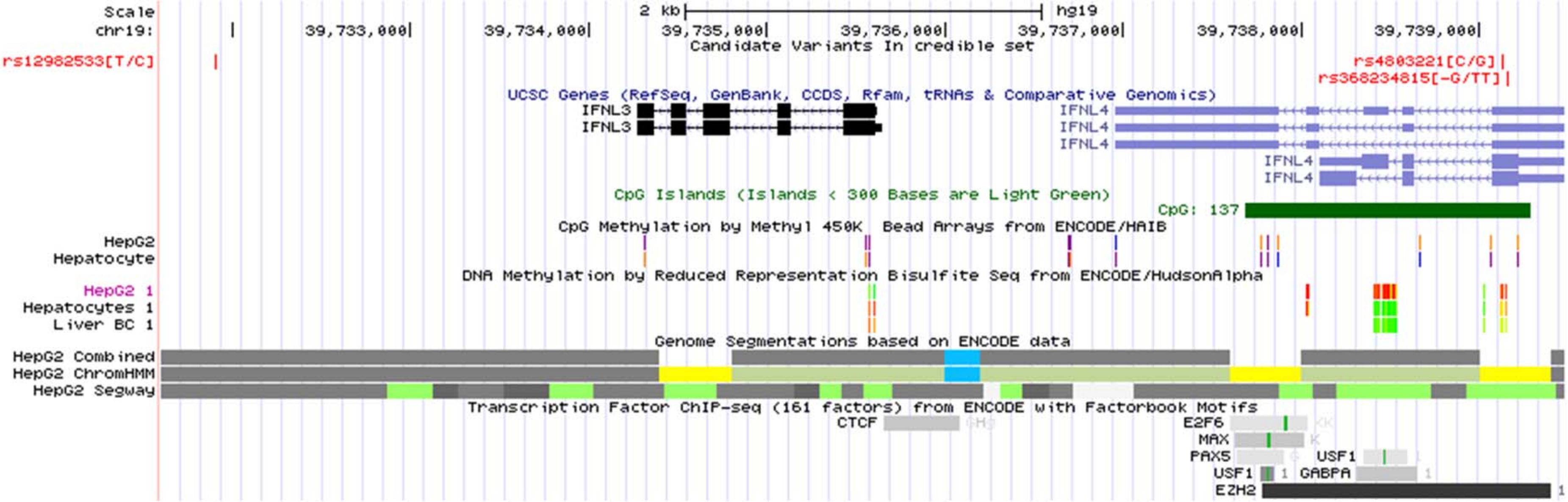
Haplotype	rs368234815	rs117664844	IFNL4 protein	European Ancestry Population					African Ancestry Population				
				Haplotype Frequency			Effect		Haplotype Frequency			Effect	
				All	Clear.	Pers.	OR (95% CI)	P value	All	Clear.	Pers.	OR (95% CI)	P value
3	ΔG	G	P70	0.218	0.137	0.273	1	-	0.558	0.448	0.581	1	-
1	ΔG	A	S70	0.098	0.087	0.106	1.6 (1.2-2.15)	0.0005	0.072	0.050	0.077	0.8 (0.5-1.2)	0.37
4	TT	G	Abrogated	0.683	0.776	0.621	2.4 (2.07-2.99)	$<10^{-08}$	0.370	0.502	0.342	1.9 (0.6-2.3)	$<10^{-08}$
2	TT	A	Abrogated	0.000	0.000	0.000	-	-	0.000	0.000	0.000	-	-

Supplementary Table 7. Distribution of rs368234815- rs1176648444 haplotypes in European and African population. Abbreviations: OR: odds ratio; clear: clearance; pers: persistence; CI: confidence interval.

Supplementary Figures.



Supplementary Figure 1. Location of the 25 variants identified by sequencing and genotyping/imputation. Variants in blue are those identified by sequencing, variants in red are the 10 candidate variants with additional significant association in the gentotyped/imputed dataset.


Scale
chr19:
2 kb
hg19
39,733,000
39,734,000
39,735,000
39,736,000
39,737,000
39,738,000
39,739,000
Candidate Variants In credible set
rs12982533[T/C]
rs4803221[C/G]
rs368234815[-G/TT]
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
IFNL3
IFNL3
IFNL4
IFNL4
IFNL4
IFNL4
IFNL4
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 137
CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB
HepG2
Hepatocyte
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
HepG2 1
Hepatocytes 1
Liver BC 1
Genome Segmentations based on ENCODE data
HepG2 Combined
HepG2 ChromHMM
HepG2 Segway
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
CTCF
E2F6
MAX
PAX5
USF1
USF1
GABPA
EZH2

Supplementary Figure 2. Annotation of the 99% credible set found in the analysis of potential functional variants showing the regulatory elements in hepatic cells and target sites for transcription factors. Variants in red were identified as potential causal variants included in the 99% credible set.