Selection on a Variant Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4 (IFNL4)

Felix M. Key1, Benjamin Peter2, Megan Y. Dennis3, Emilia Huerta-Sánchez4, Wei Tang5, Ludmila Prokunina-Olsson5, Rasmus Nielsen2, Aida M. Andrés1*

1 Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2 Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America, 3 Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America, 4 School of Natural Sciences, University of California Merced, Merced, California, United States of America, 5 Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract
Interferon lambda 4 gene (IFNL4) encodes IFN-λ4, a new member of the IFN-λ family with antiviral activity. In humans IFNL4 open reading frame is truncated by a polymorphic frame-shift insertion that eliminates IFN-λ4 and turns IFNL4 into a polymorphic pseudogene. Functional IFN-λ4 has antiviral activity but the elimination of IFN-λ4 through pseudogenization is strongly associated with improved clearance of hepatitis C virus (HCV) infection. We show that functional IFN-λ4 is conserved and evolutionarily constrained in mammals and thus functionally relevant. However, the pseudogene has reached moderately high frequency in Africa, America, and Europe, and near fixation in East Asia. In fact, the pseudogenizing variant is among the 0.8% most differentiated SNPs between Africa and East Asia genome-wide. Its raise in frequency is associated with additional evidence of positive selection, which is strongest in East Asia, where this variant falls in the 0.5% tail of SNPs with strongest signatures of recent positive selection genome-wide. Using a new Approximate Bayesian Computation (ABC) approach we infer that the pseudogenizing allele appeared just before the out-of-Africa migration and was immediately targeted by moderate positive selection; selection subsequently strengthened in European and Asian populations resulting in the high frequency observed today. This provides evidence for a changing adaptive process that, by favoring IFN-λ4 inactivation, has shaped present-day phenotypic diversity and susceptibility to disease.

Introduction
Interferon-lambda (IFN-λ) proteins induce antiviral effectors in host target cells and have a crucial role in immune defense against pathogens [1]. The IFNL family classically included three genes (IFNL1, IFNL2, and IFNL3; formerly IL29, IL28A, IL28B, respectively) located within a 50 kb region of chromosome 19 [2,3]. Several intergenic variants within the IFNL cluster had been identified as showing remarkable association with clearance of hepatitis C virus (HCV) [4–6], which is worldwide responsible for ~170 million infections and over 350,000 deaths per year [7,8]. The underlying functional basis of this association remained unclear despite numerous efforts to identify functional consequences of these variants [9–16].

An additional member of the IFN-λ family has recently been discovered: IFN-λ4, which bears only 30% amino acid identity with the other IFN-λs and is encoded by the IFNL4 gene, also located within the IFNL locus [2,17]. IFN-λ4 shows similar antiviral activity like IFN-λ3 but as it shows limited secretion it might also act intracellularly, unlike the other IFN-λs [18]. A compound di-nucleotide exonic variant (rs368234815, ΔG>TT) in IFNL4 causes a frame-shift of its open reading frame and results in the polymorphic pseudogenization of IFNL4 - the polymorphic loss of IFN-λ4 protein [17]. The existence of IFNL4 was not even computationally predicted because the human reference genome contains the TT allele and lacks the IFNL4 open reading frame [17]. Remarkably, the derived TT allele not only eliminates IFN-λ4, but it also shows the strongest genetic association reported to date with improved spontaneous and treatment-induced HCV clearance [17,19,20].

The function of IFN-λ proteins is crucial for response to pathogens and this locus has evolved under natural selection, with signatures of positive selection being described in the three classical IFNL genes (IFNL1-3) [21]. However, that analysis did not cover the IFNL4 gene, nor the frame-shift rs368234815 variant, which were then unknown [21]. Therefore, the
Author Summary

The genetic association with clearance of Hepatitis C virus (HCV) is one of the strongest and most elusive known associations with disease. The genetic variant more strongly associated with improved HCV clearance inactivates the recently discovered IFNL4 gene, which encodes for antiviral IFN-λ4 protein, and turns it into a polymorphic pseudogene. We show that functional IFN-λ4 is conserved and functionally important in mammals. In humans though the inactivating mutation appeared in Africa just before the out-of-Africa migration and quickly became advantageous, with the strength of selection (the degree of advantage) varying across human groups. In particular, selection became stronger out of Africa and was strongest in East Asia, raising the frequency of the pseudogene and resulting in the virtual loss of functional IFN-λ4 protein in several Asian populations. Although the environmental force driving selection is unknown, this process resulted in variable clearance of HCV in modern human populations. The complex selective history of IFNL4-inactivating allele has thus shaped present-day heterogeneity across populations not only in genetic variation, but also in relevant phenotypes and susceptibility to disease.

Results

Functional IFN-λ4 is strongly conserved in mammals

The IFNL4 gene is present in most mammals analyzed, although it is absent in mouse and rat (Methods). To understand the evolutionary conservation of IFNL4 we performed a comparative analysis of the IFNL4 coding sequences from a representative set of mammals (N = 12). The overall dN/dS (non-synonymous to synonymous substitution ratio) is 0.23 across mammals and 0.22 across primates (Figure 1), indicative of purifying selection maintaining the sequence and function of the protein. Notably, all individual branches except squirrel monkey have dN/dS≤1 and no model of protein evolution supported dN/dS>1 in specific branches or sites (Table S1). This reveals strong evolutionary conservation of IFN-λ4 in mammals, reflecting its functional relevance.

Strong population differentiation for the TT allele

The selective constraint on IFN-λ4 contrasts with the pseudogenization of the gene in humans through the derived TT allele [17]. The multiple-species alignment shows that ΔG is the conserved, ancestral allele and TT is the derived human-specific allele. The mutational process from ΔG to TT in humans is unclear, but only these two forms have been observed, so they should be considered as two alleles of a di-nucleotide variant (Methods). The TT allele shows considerable frequency variation across human groups. The 1000 Genomes data [24] reveals a gradient in frequency that rises from Africa (0.29–0.44) to Europe (0.50–0.77) and the New World (0.51–0.65), and reaches near fixation in East Asia (0.94–0.97) (Figure 2, Table S2, full population names in Methods).

Population differentiation can be quantified with the fixation index FST [25], a measure of the pairwise level of differentiation in allele frequencies. We used Yoruba (YRI) as the background population because it has the lowest frequency of the derived TT allele in Africa. To put these values in the context of genome-wide population differences, FST was also calculated for every SNP in the 1000 Genomes dataset. For the TT allele the largest FST, 0.63, corresponds to Southern Han Chinese (CHS) versus YRI, which places the TT allele in the 0.5% tail of the empirical genomic distribution of CHS-YRI FST (Fig. 3A, Table 1). FST is also in the 0.8% tail of the genomic distribution for the other East Asian populations (CHB, JPT, Fig. 3B and C, Table 1), and in the 4% tail for Europeans (CEU) and one African population, Luhya (LWK) (Table 1, Fig. S1). These results remain significant when other populations were used as background and in continental comparisons, and when the genome-wide distribution was restricted to SNPs with the lowest frequency in Yoruba (Table S3). Therefore, rs368234815 is among the 0.8% most differentiated SNPs between African and East Asians, and among the 12% most differentiated SNPs between African and European populations.

The TT allele resides in a population-specific extended haplotype

The unusually high population differentiation of the TT allele is compatible with a scenario of recent population-specific natural selection. Under certain selection models such high differentiation should be accompanied by extended haplotype homozygosity in the populations experiencing selection, but not in other populations. We evaluated such a signature with the cross population extended haplotype homozygosity test [26] (XP-EHH), which was calculated across the genome relative to the Yoruba population. The XP-EHH value for the TT allele is in the 0.5% tail of the empirical distribution for East Asian populations (p = 0.003, 0.005, and 0.003, for CHS, CHB, and JPT, Fig. 3A–C, Table 1), and the signal remains significant when calculated relative to a European population (GBR) and in analyses at the continental level (Table S4). In addition, some non-African populations show marginally significant signatures of positive selection too (CEU, PUR, LWK, Table 1, Fig. S1). Similar results were obtained with iHS, a statistic that explores haplotype homozygosity within a single population [27] (Table 1) (although iHS lacks power when population frequency is very high, like in Asia). The unusual allele-specific haplotype homozygosity is evident in Figure 3D, which shows the haplotype structure of the locus in one African, one European, and one East Asian population (for all populations see Fig. S3). We note that FST shows very weak correlation with both XP-EHH and iHS in the genome (r = 0.12 and r = −0.08 Spearman rank test, respectively, although the large number of data points makes these weak correlations significant, P-value<2.2e-16). Therefore, the FST and XP-EHH/iHS observations can be considered largely independent.

Finally, not only rs368234815 itself but also its genetic locus shows signatures of recent positive selection, with significant Fay and Wu’s H test [28] (FW), which detects an excess of high-frequency derived alleles in the region (Table 1). Together, the combined signatures of FST, XP-EHH, iHS and FW provide strong evidence for the action of natural selection rapidly
increasing the frequency of the TT allele in East Asia. The signature outside Asia is less clear, with most populations showing significant signatures of selection for a subset of the tests performed.

Cumulative evidence that TT allele drives the signatures of selection

A classical problem in population genetics is the identification of the genetic variant responsible for a selection signal. High linkage disequilibrium (LD) in the region surrounding IFNL4 (Table 2, Fig. 3D, Fig. 4, Fig. S7) hampers the distinction of signatures across all the linked variants, making it difficult to identify the causal variant. We conclude that rs368234815 is the most likely variant driving the signatures of selection, based on three lines of evidence: (1) its functionality and phenotypic consequences, (2) its genetic association with viral clearance, which reflects its effect on fitness, and (3) its signatures of selection.

First, the TT allele has a clear phenotypic consequence as it leads to abrogation of IFN-λ4. This is in contrast with other variants in the locus for which no conclusive functional data has been reported despite numerous efforts [9–14]. Second, of all variants in the IFNL region, rs368234815 shows the strongest genetic association with spontaneous and treatment-induced HCV clearance in African Americans [17,19]; in Europeans and Asians the strong LD across the region results in comparable associations for many variants [15–17,20,29] (Table 2, Fig. S2, Fig. S7). Third, of all protein-coding or HCV-associated variants in this locus, rs368234815 shows the strongest combined signatures of positive selection in East Asians (Fig. 3A–C, Fig. S1 and S2, Table 2). Only one other polymorphism (intergenic rs8109886, located upstream of IFNL4, Fig. 4), shows signals of selection comparable to rs368234815 (Fig. S2, Table 2). No function has been ascribed to this variant despite a moderate HCV association that is likely due to linkage to TT [6,17,30] (Table 2 and Fig. 4), making it a priori a less likely candidate for selection. Indeed, simulations of the evolutionary process showed that the large frequency change of rs8109886 can be explained by linkage to the TT allele alone (Note S1).

We also put rs368234815 in the context of the signatures of selection in the larger genomic region. IFNL4 is located upstream of IFNL3 in a region of moderate LD that is separated from the IFNL1/IFNL2 locus by a recombination hotspot (Fig. S7, Table 2). Manry et al. [21] identified signatures of recent positive selection in all three original IFNL genes (IFNL1-3) but neither IFNL4 nor the rs368234815 variant were known at that time and thus they were not considered. The recombination hotspot breaks LD between the IFNL1/IFNL2 locus and the IFNL3/IFNL4 locus (Fig. S7, Table 2), showing that these signatures are in all likelihood independent, as suggested by Manry et al. [21]. There is moderate LD between IFNL4 and IFNL3, with an average $r^2$
between rs368234815 and IFNL3 SNPs of 0.18 in CEU and of 0.44 in CHS (see also Fig. S7). So, the selection signatures in IFNL3 and IFNL4 may not be independent. In fact, the seven SNPs identified by Manry et al. [21] (detailed in Table 2 and Fig. 4) have (1) weaker signatures of selection; (2) unclear functional effects, and (3) weaker association with HCV clearance than the TT allele in Africa (Table 2, Figure 4). Also, those that show some signatures of selection have high to moderate LD with rs368234815 (Table 2), with LD broken mostly by a few recombination events in the ancestral haplotype (Note S1). Taken together, these lines of evidence confirm that IFNL1/2 and IFNL3/IFNL4 have likely been independently targeted by positive selection in recent human history, as suggested by Manry et al. [21], and highlight rs368234815 TT as the most likely selected allele in its region.

Mode and tempo of positive selection on the TT allele

The classical model of positive selection involves selection on a de novo mutation (SDN), a so-called hard sweep, where a new mutation immediately becomes beneficial and selected (reviewed in [31]). This scenario is difficult to reconcile with our observations, because unequivocal signatures of selection are observed only in East Asians but the TT allele is common worldwide. The TT-carrying haplotype harbors the highest genetic diversity in Africa indicating that it arose there before the out-of-Africa dispersion (Note S2, Table S5), a result that is consistent with the IFNL4 haplotype network (Fig. S4). Under SDN, only a model where selection begins weak in Africa and becomes stronger outside of Africa could explain our observations (Fig. 5A). An alternative model is selection from standing variation (SSV), also known as a soft sweep (reviewed in [31]). In this scenario an existing neutral or nearly neutral allele becomes advantageous, for example upon environmental change (Fig. 5A).

To disentangle the most likely model of selection for the TT allele we applied a modified version of a recently published ABC approach [23], which we extended to be able to analyze two-population models. In brief, we match millions of simulations under the different models to a summary of the observed genetic data in the IFNL4 region, and use the best matching simulations for further inferences. Under reasonable assumptions we expect the most realistic selection model to produce the closest simulations to real data, and thus simulations can be used to make inferences about the selective history of the allele [23] including the model of selection and relevant parameters (Note S3). While the method relies on some assumptions (e.g. correct demographic and dominance models) this approach has been shown to be robust and to have high power to recover the correct selection scenario [23]. We assess that we overall have high power to recover the correct model, with 76% of the SSV and 95% of the SDN simulations being assigned correctly under the East Asian demographic model, and 70% of the SSV and 97% of the SDN simulations being assigned correctly under the European demographic model (Note S3). The slight bias observed was considered when interpreting the results. For our analysis we consider three models: neutrality (no selection), selection from a de novo mutation (SDN) and selection from standing variation (SSV) (Fig. 5A).

In East Asian populations we obtain negligible support for neutrality and very strong support for the SDN model (Fig. 5B, Table 3, Table S6). Results in Europeans are also consistent with the SDN model, although the weaker signals of selection and the slight bias observed above make these results less conclusive (Table 3, Note S3, Fig. S5, Table S6). The posterior probability...
for the SDN model is 95% in East Asia and 80% in Europe, corresponding to Bayes factors (Bayesian measures of relative model support [32]) of ~10 and ~4, respectively. This provides substantial and robust evidence for the SDN model, compared to the SSV and NTR models for East Asian and European populations according to Jeffrey’s interpretation [33]. Therefore, we conclude that the TT allele was likely positively selected upon appearance. The ABC-based parameter estimates are less reliable than the model choice [23] because they always have large credible intervals (Bayesian measures of confidence). However, the posterior distributions have modes that differ from the modes of the prior distributions, indicating that they are determined by information from the data and not by the prior (Fig. S5). Also, the estimates are quite concordant within and between continental groups (Fig. S5, Table 3). So while they should be interpreted with appropriate caution, the estimates do provide additional useful information about the model and timing of selection. We infer that the TT allele emerged before the out-of-Africa migration (estimated $t_{\text{cons}}=55,900$ years ago (41,360–68,640)) and was immediately, or shortly thereafter, targeted by moderate positive selection (selection coefficient, $s_A$, <0.58% (0.17–1.23%)); we estimate that selection intensified substantially outside of Africa, with the selection strength nearly quadrupling in Europe and in Asia ($s_{\text{NA}}<2.6\%$ (0.6–4.8); Table 3, Fig. S5).

One important aspect of the simulations is the mode of dominance (also known as the genetic model), and the ABC analysis above was performed on simulations under a perfectly additive model where heterozygotes have half the fitness effect of homozygotes (dominance coefficient $h = 0.5$). This model is reasonable because in TT/DG heterozygotes only one IFNL4 copy is truncated, and because genetic studies show that the odds ratios (ORs) for HCV clearance in heterozygotes are intermediate to those in the two homozygotes [17]. These two arguments argue strongly against a model of complete dominance for TT as realistic, but other models are more difficult to discard a priori. We thus compare three dominance models: (1) a fully recessive model for the TT allele ($h = 0$), (2) the perfectly additive model used above ($h = 0.5$), and (3) a supra-additive model where the additive effect is non-linear and heterozygotes are closest in fitness to DG homozygotes. This model has been proposed based on the ORs for the intronic IFNL4 variant, rs12979860 which is in high LD with rs368234815 and is thus a good proxy for the dominance effects of TT (Table 2) [34]. Based on those results we use a
well as the SDN model (original ABC analysis and as expected given the weaker signatures in the European population the results are less clear, just as in the SSV model). Therefore our results in East Asia validate the estimates also agree well among these two models (Fig. S5, Fig. S6 and Note S4). Therefore our results in East Asia validate the use of an additive model and show that the ABC inferences are not sensitive to the particularities of the additive model used. In European population the results are less clear, just as in the original ABC analysis and as expected given the weaker signatures of selection. Still, these results also support the two additive models (36% support for additive and 38% for supra-additive; Fig. 5C) as well as the SDN model (~81% support for SDN in both the additive and the supra-additive, corresponding to a Bayes factor ~4, Fig. 5B and D, Note S4).

These results show a complex selection history for the TT allele, with selection starting upon appearance of the allele but with intensity changing over time and geographic range. The model is consistent with all our observations, including the marginal evidence for selection observed in non-Asian populations (Table 1). It is interesting that we infer selection on the TT allele even in Yoruba, where the signature is undetectable with classical signatures of long-standing balancing selection (Note S4). It is interesting that we infer selection on the TT allele even in Yoruba, where the signature is undetectable with classical signatures of long-standing balancing selection (Note S4).

### Discussion

Here we show that functional IFNL-4 is under purifying selection throughout the mammal clade while positive selection has favored the elimination of IFNL-4 through pseudogenization in humans. Selection on the TT allele has been particularly strong in specific populations, leading to extremely high frequency of the pseudogene and subsequent virtual loss of IFNL-4. This event is phenotypically relevant: not only is IFNL-4 biologically important [17,18] and evolutionary conserved, but the loss of IFNL-4 through pseudogenization shows remarkable association with improved HCV clearance [17,19,20].

The precise reason behind the advantage of IFNL-4 elimination is unknown, but its immunological role and clear antiviral activity against HCV [17] make exposure to pathogens (and in particular viral agents) the most likely selective force. However, due to its slow progression into fatal disease [35] HCV is unlikely to have exerted such strong selective pressure, although we cannot completely discard this possibility. Besides HCV, it has been shown that functional IFNL-4 has antiviral activity against coronaviruses [18], while the IFNL-4 pseudogene increases susceptibility to cytomegalovirus retinitis among HIV-infected patients [36]. Suggesting that IFNL-4 pseudogenization is likely associated with several phenotypic traits. It is perhaps surprising that suppression of an antiviral protein results in improved viral clearance, although it has for example been shown that during chronic infection blockage of persistent signaling of IFN I (a different type of interferon) can improve viral clearance [37,38].

We showed that a complex selective regime, with variation in selection strength in different geographical areas, best explains the history of the IFNL-4 locus. Signatures of non-neutral evolution have been detected in other interferons, including at least one other IFNL family member (IFNL1 or IFNL2) [21]. Although the mode and tempo of selection in these other IFNL genes are not well understood, together these observations suggest that IFNL-4 proteins have played an important role in recent human adaptation, probably as a consequence of their role in individuals’ constant fight with pathogens. It is likely, though, that only the selective history of the IFNL-4 TT allele had a strong influence in

<table>
<thead>
<tr>
<th>Population</th>
<th>$F_{ST}$</th>
<th>$F_{ST}$ P-value</th>
<th>XP-EHH P-value</th>
<th>iHS P-value</th>
<th>FW P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHS</td>
<td>0.63</td>
<td>0.005</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHB</td>
<td>0.60</td>
<td>0.006</td>
<td>0.005</td>
<td></td>
<td>0.22</td>
</tr>
<tr>
<td>JPT</td>
<td>0.56</td>
<td>0.008</td>
<td>0.003</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>GBR</td>
<td>0.22</td>
<td>0.074</td>
<td>0.071</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>CEU</td>
<td>0.31</td>
<td>0.040</td>
<td>0.041</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>FIN</td>
<td>0.26</td>
<td>0.059</td>
<td>0.059</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>TSI</td>
<td>0.15</td>
<td>0.120</td>
<td>0.070</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>CLM</td>
<td>0.09</td>
<td>0.172</td>
<td>0.201</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>MNL</td>
<td>0.04</td>
<td>0.368</td>
<td>0.145</td>
<td>0.50</td>
<td>0.11</td>
</tr>
<tr>
<td>PUR</td>
<td>0.19</td>
<td>0.064</td>
<td>0.029</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>ASW</td>
<td>−0.01</td>
<td>0.684</td>
<td>0.634</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>LWK</td>
<td>0.05</td>
<td>0.036</td>
<td>0.043</td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td>YRI</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* $F_{ST}$ and XP-EHH were not calculated for YRI, which is the background population.
* iHS requires a MAF of 5%, which is not given for CHS.

doi:10.1371/journal.pgen.1004681.t001

Table 1. $F_{ST}$ values and for $F_{ST}$, XP-EHH, iHS and Fay and Wu's H (FW) the empirical P-values are shown for rs368234815 in every population.
Table 2. Derived allele frequency (DAF), LD to rs368234815, and signatures of selection (empirical P-values for FST and XP-EHH) for other relevant SNPs across the IFNL-locus.

<table>
<thead>
<tr>
<th>rs ID</th>
<th>DAF (50 ind.)</th>
<th>r² to rs368234815</th>
<th>XP-EHH P-value*</th>
<th>FST P-valuea</th>
<th>Empirical P-valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHS</td>
<td>CEU</td>
<td>CHS</td>
<td>CEU</td>
<td>CHS</td>
<td>CEU</td>
</tr>
<tr>
<td>rs368234815</td>
<td>0.75</td>
<td>0.97</td>
<td>-</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>rs11881222</td>
<td>0.75</td>
<td>0.97</td>
<td>0.8</td>
<td>1</td>
<td>0.14</td>
</tr>
<tr>
<td>rs8103142</td>
<td>0.76</td>
<td>0.99</td>
<td>0.85</td>
<td>0.33</td>
<td>0.05</td>
</tr>
<tr>
<td>rs28416813</td>
<td>0.70</td>
<td>0.96</td>
<td>0.78</td>
<td>0.74</td>
<td>0.06</td>
</tr>
<tr>
<td>rs12979860</td>
<td>0.74</td>
<td>0.97</td>
<td>0.85</td>
<td>0.74</td>
<td>0.06</td>
</tr>
<tr>
<td>rs8099917</td>
<td>0.79</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.04</td>
</tr>
<tr>
<td>rs8103362</td>
<td>0.73</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.04</td>
</tr>
<tr>
<td>rs30461</td>
<td>0.83</td>
<td>0.97</td>
<td>0.01</td>
<td>0</td>
<td>0.71</td>
</tr>
<tr>
<td>rs59746524</td>
<td>0.87</td>
<td>0.94</td>
<td>0.94</td>
<td>0</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Included are variants which showed either a significant association with HCV clearance (G), functional effects (F), signatures of selection in a previous scan of this region, or novel strong signatures of natural selection (S). Location of each SNP in the IFNL-locus is visualized in Figure 4.

*aCalculated using YRI as background population.

*Selection signal detected in this study.doi:10.1371/journal.pgen.1004681.t002

---

the rate of clearance of some viruses, at least HCV, across human groups.

It has been proposed that gene loss may exert an important role in evolution, including human evolution [39], and the loss of otherwise conserved regulatory elements may play a role in the acquisition of human-specific phenotypes [40]. Loss-of-function mutations show global signatures of purifying selection [41–43] and tend to carry detrimental effects [44]. A few exceptions exist, though, where truncating polymorphisms show signatures of positive or balancing selection [45–50]. Still, as with other targets of selection, most of these cases lack biological interpretation. In fact, IFNL4 joins a small group of known genes where a striking signature of local adaptation is coupled with a clear molecular phenotype (e.g., [46,47,51]), which in this case is also associated with disease risk. As such, it contributes to our understanding of how recent human evolution has shaped genetic and phenotypic human diversity, including present-day heterogeneity in susceptibility to disease.

---

Materials and Methods

Molecular Evolution of IFNL4 across species

In order to explore the level of functional constraint in IFNL4, we estimated the level of protein conservation in primate and non-primate mammals. Specifically, we assessed the ratio (dN/dS) of non-synonymous substitutions per non-synonymous site (dN) to synonymous substitutions per synonymous site (dS) across species. Since purifying selection eliminates deleterious protein-coding changes, dN/dS decreases with negative selection and increases with relaxed constraint and positive selection.

We used human IFNL4 reference sequence NM_001276254.2 to BLAT genomes of other species and generate multiple-species sequence alignment of IFNL4 coding exons 1 through 5 (Table S8). The panda-predicted IFNL4 ortholog was subsequently used as BLAT query to extract coding exons for additional non-primate species (Table S8). Further, we sequenced IFNL4 exons (and introns) in genomic DNA and reconstructed complete IFNL4 cDNA sequences of chimpanzee (Genbank accession JX867772), baboon (Genbank accession KC525947) and crab-eating macaque (Genbank accession KC525948). The whole IFNL4 genomic region is absent in mouse or rat. All discovered functional IFNL4 sequences (Table S8) where used for a multiple-sequence alignment which was created using ClustalW [52] and annotated with Jalview [53].

The alignment was analyzed with codeml (part of PAML4 [54]) to test various models of selection. We estimated the overall dN/dS for the complete tree and compared likelihoods for models that allowed: i) free dN/dS for each branch (i.e., lineage heterogeneity); ii) a primate-specific dN/dS; and iii) a human-specific dN/dS. Additionally, we performed tests aimed to detect site-specific signatures of positive selection across the phylogeny (branch models): i) model 1a (neutral) vs. model 2 (positive selection); ii) model 7 (neutral) vs. model 8 (with dN/dS); and iii) model 8a (with dN/dS = 1) vs. model 8 (with dN/dS > 1).

Human population genetic data

We analyzed genome-wide data from the 1000 Genomes release (2010/11/23; phase I) [24]. We considered (1) autosomal variants detected in the low coverage sequencing, and (2) populations with information for at least 50 unrelated individuals, which was met by 13 populations from four different continents [African ancestry: YRI (Yoruba in Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), ASW (African Ancestry in Southwest US); European ancestry: GBR (British from England and Scotland), CEU (Utah...
residents (CEPH) with Northern and Western European ancestry; FIN (Finnish from Finland), TSI (Toscani in Italia); East Asian ancestry: CHS (Han Chinese South), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan); American ancestry: MXL (Mexican Ancestry in Los Angeles, CA), CLM (Colombian in Medellin, Colombia), PUR (Puerto Rican in Puerto Rico). As an exception PUR (Puerto Rico) was analyzed although it contains only 44 individuals. Some analyses were performed both by population and by continent; in these cases the continental groups contain 150 randomly selected, unrelated individuals (America 144) with an equal contribution from each population within the continent.

For the rs368234815 ΔG/TT frameshift-substitution variant the 1000 Genomes dataset only contains the T insertion/deletion variant rs11322763 (~T, chr19:39739154, dbSNP b136), while the substitution rs74597329 (G/T, chr19:39739155, dbSNP b136) is absent. This is due to the automatic variant caller failing to correctly identify an insertion and a substitution in the same genomic position. We sequenced an amplicon containing rs368234815 in 133 individuals included both in the 1000 Genomes and HapMap sets (CEU, YRI and CHB/JPT). Sequencing confirmed the presence of only two alleles (ΔG and TT) and showed good concordance with the 1000 Genomes data between our ΔG/TT genotypes and 1000 Genomes genotypes for the overlapping insertion/deletion variant rs11322763 (4 individuals of 153 tested were discordant, providing an estimated 97.4% genotype and 90.7% allele concordance rate). This validated the use of 1000 Genomes dataset for our subsequent analyses. We used the ancestral allelic state annotated in the 1000 Genomes data, which is based on the Ensembl 59 comparative 32 species alignment [53]; only SNPs with a high-confidence ancestral inference were used, and indels were excluded due to their cryptic variation patterns [56].

Signatures of selection

We used F<sub>ST</sub>, iHS and XP-EHH to explore the signatures of selection of rs368234815 TT allele. F<sub>ST</sub> is a measure of population differentiation and unusually high F<sub>ST</sub> can indicate population-specific positive selection that drastically increases allele frequency in the population under selection [57]. To calculate F<sub>ST</sub> we used the Weir and Cockerham [23] estimator implemented in vcf-tools [58].

Positively selected alleles rapidly increase in frequency with recombination having little chance to break their association with nearby variants. If the selected allele was originally in few haplotype backgrounds and it has not reached fixation, it will be associated with extended haplotype homozygosity (EHH), a pattern that will be absent for the non-selected allele. We used two statistics to explore this expectation. First, iHS [27] measures the allele-specific decay of EHH within a population by comparing the associated EHH of ancestral and derived alleles. Second, XP-EHH [26] that detects alleles that are under selection in one population only, by comparing EHH patterns both among allelic types and across populations; as such XP-EHH has higher power to detect population-specific selection. Low frequency variants break the EHH signal, so following [59] we considered only SNPs with derived allele frequency ≥ 5% for XP-EHH or minor allele frequency ≥ 5% for iHS. Local recombination rate estimates were obtained from a combined recombination map based on HapMap data [60] from Africa, European, and Asian populations. Both statistics were standardized to a mean of zero and a standard deviation of one; for iHS, scores were then binned by frequency (1%) as previously suggested [27]. Correlation of F<sub>ST</sub> with XP-EHH (CHS vs. YRI) or iHS (CHS) was calculated for all variants present in the respective dataset with Spearman’s rank correlation test implemented in R [61].

We used each of these statistics to analyze every non-African population; for between-population comparisons we used Yoruba as background, unless noted otherwise. To assess the putative effects of this choice of populations we repeated the analyses for continental groups, for different background populations, and for SNPs that have their lowest allele frequency in Yoruba. In all cases the empirical P-values were obtained by comparing the score for rs368234815 to the whole-genome empirical distribution of the respective statistic. Since this is a hypothesis-driven analysis with a single variant analyzed within a single locus, no multiple testing or genome-wide corrections are needed.

We also applied tests that analyze the signatures of selection in the IFNL4 genetic region (~2.5 kb). Here we show results for Fay and Wu’s H test [28], which detects the excess of high-frequency derived alleles expected after a recent sweep with recombination. Significance was estimated using 10,000 standard neutral coalescent simulations [62]. Because demography affects the SFS and can cause spurious results if not properly accounted for, our simulations are run under a demographic model which includes inferred parameters for populations of African [63], European [63], Asian [63] and American [64] ancestry. A custom made perl program (Neutrality Test Pipeline) was used to calculate the statistic and corresponding P-value.

ABC analysis

To infer the model of selection that best fits IFNL4 data and estimate the timing and selection strength of the TT allele, we used an Approximate Bayesian Computation (ABC) approach [22]. In particular, we followed a published approach [23], which has been previously shown to discriminate well between SDN, SSV and neutrality (NTR) [23]. In brief, this approach is based on performing a large number of simulations under different selection
models, with random parameters drawn from some probability distribution (called the prior distribution). Real data and simulations are compared based on summary statistics, and through a rejection scheme the simulations that most closely resemble real data help inform inferences about the best-fitting model. The parameter values that generate these simulations are then used to obtain the posterior distribution of each parameter, whose mean and standard deviation are used to perform the parameter inferences. We extended the method to consider more than one population, since two-population statistics are most informative in our case.

Figure 5. (A) Graphical representation of the different models of selection tested in the ABC analysis (NTR - neutral, SDN - selection on a de novo mutation, and SSV - selection on standing variation). We simulated one ancestral population that splits at the out-of-Africa event (at 51,000 years ago) into the African (AFR) and the non-African (non-AFR) populations, which experience subsequent migration. The star indicates the appearance of the focal mutation. In the first case the neutral (black) mutation appeared and evolved under neutrality in both populations. In the SDN model the advantageous mutation (red) is immediately under positive selection with strength $s_A$, and time when selection started $t_{mut}$ (the prior parameter space for $t_{mut}$ is indicated by a green line); selection strength is allowed to change in the non-African population to $s_{NA}$. In the SSV model the neutral (black) mutation appeared and evolved under neutrality, becoming advantageous in the non-African population (red line) at time $t_{mut}$. Prior parameter spaces can be found in methods. (B) Posterior probabilities of the model choice for the different selection models under perfect additivity. (C) Posterior probabilities of the model choice for the different dominance models (and neutrality, NTR). For all models except NTR the posterior probability represent the sum for the SDN and SSV selection models. (D) Posterior probabilities of the model choice for the different selection models under the supra-additive model. In (B), (C), and (D), NTR has negligible posterior probability and is therefore not visible.

doi:10.1371/journal.pgen.1004681.g005
models, we started with uniform priors with a range as follow (see Fig. 5A):

(1) SDN model - selection strength in Africa \( s_A \sim \text{U}(0,1.5\%) \); selection strength in non-Africa \( s_{NA} \sim \text{U}(0,5\%) \); time when selection started \( t_{\text{mut}} \sim \text{U}(40,70\text{kya}) \)

(2) SSV model - selection strength in non-Africa \( s_{NA} \sim \text{U}(0,5\%) \); frequency of the allele when selection started \( f_0 \sim \text{U}(0,20\%) \); time when selection started \( t_{\text{mut}} \sim \text{U}(21,51\text{kya}) \)

(3) NTR model - time when mutation appears \( t = \text{U}(40,70\text{kya}) \)

Because simulations with the selected allele fixed are likely to be very different from the observed data, we conditioned on the selected allele segregating in both populations. This resulted in non-uniform prior distributions presented in Figure S5 and S6. We used \( 10^5 \) simulations to distinguish between the neutral model and the two selection models, and a larger set of \( 8 \times 10^5 \) simulations for the more subtle distinction between the two selection models and for parameter estimation. For the simulations, we used the population history model estimated by Gravel et al. [63] and assumed a constant recombination rate of 1.76 cm/Mb throughout the region (average recombination rate in the IFNL locus [60]), and a perfectly additive model of dominance \( (h = 0.5) \). Lack of an appropriate demographic model for American and non-Yoruba African populations precludes analysis for those populations.

The following single-population statistics were calculated: the average number of pairwise differences \( \pi \), Watterson’s \( \theta \), Fay and Wu’s \( H \) [28] and Tajima’s \( D \) [67], all for both 4 kb around the site and a 8 kb (6 kb upstream and 2 kb downstream of the site) interval around the TT allele. The between-population statistics employed were: FST [68] for the selected site, FST in 4 kb around the site, FST for the whole region, and XP-EHH on the selected site [26]. In addition, we also included the frequency of the selected allele in both populations. This resulted in a set of 16 summary statistics, which, following Wegmann et al. [69] and Peter et al. [23], was reduced to seven summary statistics using PLS-DA [70] for model choice and regular PLS for parameter inference [71]. Performance of the ABC model choice and parameter distribution for the SDN model has been assessed for each particular model (Note S3). Confidence in the choice of selection models has been supported with Bayes factors.

In addition, we investigated the influence of the dominance model in our inferences. We analyzed a recessive model for TT \( (h = 0) \), the perfectly additive model above \( (h = 0.5) \), and a supra-additive model \( (h = 0.38) \), using 500,000 simulations for each model. We ran an ABC analysis for model selection with all simulations (from all three dominance models and the three selection models NTR, SDN, and SSV). We then assess the posterior probability of each dominance model regardless of selection model, and the posterior probability (and parameter estimates) of each selection model for the additive and supra-additive dominance models (see Note S4).

URLs


Supporting Information

Figure S1 Empirical P-values of the FST and XP-EHH analyses for all non-Asian populations using YRI as background within a +−15 kb region around rs368234815.

(PDF)

Figure S2 Empirical P-values of the XP-EHH and FST analysis in the IFNL cluster for all populations.

(PDF)

Figure S3 Haplotype structure +−15 Kb around rs368234815 for three population per continent.

(PDF)

Figure S4 Haplotype network for IFNL4.

(PDF)

Figure S5 Parameter estimates for the additive model using a (a) SDN or (b) SSV model of selection for Asian and European populations.

(PDF)

Figure S6 Parameter estimates for the supra-additive model using a (a) SDN or (b) SSV model of selection for Asian and European populations.

(PDF)

Figure S7 Linkage disequilibrium (LD) patterns across the IFNL locus in three representative populations (a) YRI, (b) CEU, (c) CHB.

(PDF)
Note S1  Assessment of selection on other variants in the IFNL locus.  
(PDF)

Note S2  Geographical origin of the rs368234815 TT variant.  
(PDF)

Note S3  Performance of the ABC approach.  
(PDF)

Note S4  Investigation of the effects of different dominance models in the ABC analysis.  
(PDF)

Note S5  Investigation of signatures of balancing selection in Africa, and of positive selection for non-synonymous variants on ΔG background.  
(PDF)

Table S1  PAML results using either (a) all species or (b) only primates.  
(PDF)

Table S2  Frequency of rs368234815 TT allele in the 1000 Genomes dataset and the subset of 50 unrelated individuals per locus.  
(PDF)

Table S3  F_{ST} values and corresponding empirical P-values for rs368234815 using different background populations: (a) ASW, (b) LWK, (c) GBR, and (d) in the continental comparison. Table (e) shows the empirical P-value of F_{ST} for rs368234815 based on genome-wide SNPs with lowest frequency in YRI (compared to ASW and LWK).  
(PDF)

Table S4  Empirical P-values for the XP-EHH analysis for rs368234815 using (a) GBR as background population or (b) in continental comparison.  
(PDF)

Table S5  Diversity associated with the TT haplotype in each population, as measured with Watterson's estimator and using only TT/TT homozygous individuals.  
(PDF)

Table S6  ABC results and inferred parameter estimates for (a) the SSV model and (b) the neutral model.  
(PDF)

Table S7  Tajima’s D (TD), HKA, and MWU high results.  
(PDF)

Table S8  IFNL4 orthologous exons retrieved through BLAT search with human reference IFNL4-ΔG (for primate species) and the panda ortholog (for non-primate species).  
(PDF)

Acknowledgments
We thank Cesare de Filippo, João Teixeira, Bárbara Bitarello, Anna B. Stittrich and Montgomery Shaikin for helpful discussions and/or comments on the manuscript.

Author Contributions
Conceived the project: AMA LPO. Contributed reagents/materials/analysis tools: WT LPO. Wrote the manuscript: AMA FMK. Designed the study and performed the population genetics analyses: FMK AMA. Performed the comparative genomics analysis: MYD. Performed the ABC analyses: BP EHS FMK RN. Generated data: WT LPO. Supervised the project: AMA. All authors assisted in the writing of the manuscript.

References

PLOS Genetics | www.plosgenetics.org 12 October 2014 | Volume 10 | Issue 10 | e1004681