Published in final edited form as:

Hum Mol Genet. 2007 March 15; 16(6): 579–591. doi:10.1093/hmg/ddl469.

Association of *IRF5* in UK SLE families identifies a variant involved in polyadenylation

Deborah S. Cunninghame Graham¹, Harinder Manku¹, Susanne Wagner², Julia Reid², Kirsten Timms², Alexander Gutin², Jerry S. Lanchbury², and Tim J. Vyse^{1,*}

¹Imperial College, Faculty of Medicine, Molecular Genetics and Rheumatology Section, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

²Myriad Genetics Inc., 320 Wakara Way, Salt Lake City, UT 84108, USA

Abstract

Results from two studies have implicated the interferon regulatory gene *IRF5* as a susceptibility gene in systemic lupus erythematosus (SLE). In this study, we conducted a family-based association analysis in 380 UK SLE nuclear families. Using a higher density of markers than has hitherto been screened, we show that there is association with two SNPs in the first intron, rs2004640 ($P = 3.4 \times 10^{-4}$) and rs3807306 ($P = 4.9 \times 10^{-4}$), and the association extends into the 3'-untranslated region (UTR). There is a single haplotype block encompassing *IRF5* and we show for the first time that the gene comprises two over-transmitted haplotypes and a single under-transmitted haplotype. The strongest association is with a TCTAACT haplotype (T:U = 1.92, $P = 5.8 \times 10^{-5}$), which carries all the over-transmitted alleles from this study. Haplotypes carrying the T alleles of rs2004640 and rs2280714 and the A allele of rs10954213 are over-transmitted in SLE families. The TAT haplotype shows a dose-dependent relationship with mRNA expression. A differential expression pattern was seen between two expression probes located each side of rs10954213 in the 3'-UTR. rs10954213 shows the strongest association with RNA expression levels ($P = 1 \times 10^{-14}$). The A allele of rs10954213 creates a functional polyadenylation site and the A genotype correlates with increased expression of a transcript variant containing a shorter 3'-UTR. Expression levels of transcript variants with the shorter or longer 3'-UTRs are inversely correlated. Our data support a new mechanism by which an *IRF5* polymorphism controls the expression of alternate transcript variants which may have different effects on interferon signalling.

INTRODUCTION

Systemic Lupus Erythematosus (SLE) (OMIM 152700 <http://www.genetests.org/query?mim=152700>) is a multi-system complex autoimmune disease with a wide range of clinical symptoms and characterized by the production of autoantibodies against a diverse range of nuclear and cell surface autoantigens. It is the formation of immune complexes with these autoantibodies and their deposition in multiple organs that contribute to eventual end-organ damage.

The overall disease frequency is 11–250 in 100 000 (1), it is dependent on gender and ethnic group, being more frequent in females and less common in white Caucasians. The genetic

© The Author 2007. Published by Oxford University Press. All rights reserved.

*To whom correspondence should be addressed. Tel: +44 2083832339; Fax: +44 2087433109; t.vyse@imperial.ac.uk.

SUPPLEMENTARY MATERIAL Supplementary Material is available at HMG Online.

Conflict of Interest statement. None declared.

component to SLE has been well established with familial clustering studies, where the sibling risk ratio, λ_s , is equal to 20 (2) and twin studies have shown a 10-fold greater concordance in monozygotic twins, when compared with dizygotic twins (3). In addition, there have been six genome-wide linkage scans, which have identified a number of suggestive susceptibility loci in a number of different populations (4–10).

The production of type I interferons (IFN) and the regulation of IFN-inducible genes may have crucial importance in the aetiology of SLE, since raised levels of IFN- α are a well established phenotype in SLE patients. This increase appears to be correlated with both disease activity and severity (11). Although the cells producing type I IFNs, plasmacytoid dendritic cells, make up only 0.1% of the peripheral blood mononuclear cells, production of type I IFNs stimulates a strong immune response. This occurs when IFN- α binds to the IFN- α receptor on the surface of cells to activate the janus kinases, with transmission of a signal through the JAK–STAT pathways leading to the subsequent increased expression of other genes in the IFN signalling pathway, the so-called IFN-inducible gene expression signature (12).

The transcription factor interferon regulatory factor 5 (*IRF5*) forms part of this IFN-gene expression signature, thereby playing a role in the stimulation of the immune response. The gene was initially reported to be associated with SLE in Scandinavian and Finnish samples (13), and in a separate study using samples from Argentina, Spain, Sweden and the US by Graham *et al.* (14). *IRF5* is a part of the *IRF* gene family, which contains nine transcription factors, all with a novel helix-turn-helix DNA binding motif (15). It is a 12 kb gene, containing nine exons, located on human chromosome 7q32, which is not part of a replicated linkage region mapped in SLE. *IRF5* is constitutively expressed in all lymphoid organs, except thymus, especially in plasmacytoid dendritic cells, monocytes, monocyte-derived dendritic cells and B cells (16,17). Following viral infection, expression of *IRF5* is up-regulated by IFN- α , phosphorylated and then translocated to the nucleus, where it subsequently up-regulates IFN-inducible genes, including pro-inflammatory cytokines such as IL-10, and also those involved in apoptosis and the early immune response (13). This up-regulation of molecules important in the IFN response pathway may therefore have important functional consequences in the pathogenesis of SLE. The importance of *IRF5* in the immune response is not limited to humans. The *Irf5* knockout mouse, exhibits reduced production of pro-inflammatory cytokines, including IL-6, IL-12 and TNF- α (18).

A recent paper (14) described a low-density map across *IRF5*, with four SNPs, which formed a single haplotype block, containing a single over-transmitted haplotype. The predominant individual association came from the promoter and the 5' end of intron 1. Consequently, the main focus of the published study was the promoter and intron 1, with only a single SNP included from more distal sections of the gene. The paucity of markers meant that it was not possible to fully assess the contribution to the association from the 3' end of the gene. In addition, a functional splice site from a variant at the intron–exon 1b boundary was described (14) and a contribution to the regulation of mRNA levels from a variant in the 3'-flanking region of the gene. However, neither of these functionally relevant over-transmitted alleles uniquely demarcated the over-transmitted haplotype. Therefore it is likely that both of these alleles represent part of the genetic effect from *IRF5*.

A family-based association study, in UK SLE parental-proband trios, was performed to explore the contribution of *IRF5* genotypes to the SLE phenotype. Several previously reported SNPs associations were confirmed, novel SNP associations are reported and a finer haplotype map across the gene is described. We also propose a novel mechanism that determines the expression level of some of the known *IRF5* transcript variants.

RESULTS

Haplotype structures across *IRF5* in UK SLE families

To comprehensively examine the role of *IRF5* SNPs in SLE, we genotyped eight SNPs shown in Table 1 in a cohort of SLE families. All the SNPs generated viable Sequenom assays with >95% genotyping, HWE *P*-values of >0.05 and fewer than 4% sporadic Mendelian errors. The genotyping data were used to reconstruct haplotypes using 357 European-Caucasian (EC) parental-proband trios and 22 Indo-Pakistani (IP) families. We used this joint dataset for all analyses, since there was the same trend for association for each marker in the two sub-populations. Furthermore, the parental samples for the EC and IP sub-populations showed no significant difference in minor allele frequency (MAF) for any SNP in *IRF5* and there was less than a 5% difference in the frequency of the associated haplotypes between the two sub-populations. In the EC and IP samples (CIP samples) there is a single haplotype block which stretches across all the markers typed, consisting of seven haplotypes at frequencies >3%. The haplotype structure accounts for 91.5% of the total chromosomes studied (Fig. 1).

TDT analysis for individual SNPs in *IRF5*

In order to examine the association between *IRF5* and SLE, TDT analysis was performed in the 379 parental-proband trio families using GENEHUNTER-TDT (GH-TDT) and family-based association test (FBAT). The results presented in Table 1 show significant association by GH-TDT analysis for two markers in the first intron of *IRF5*, designated *IRF5* SNP 2 ($P = 3.4 \times 10^{-4}$) and *IRF5* SNP 4 ($P = 4.9 \times 10^{-4}$). *IRF5* SNP 2 is located in the GT consensus donor splice site at the boundary of the untranslated exon 1b of *IRF5* and has been previously associated with SLE (13,14). There are weaker associations with four other markers *IRF5* SNP 3 ($P = 0.013$) in intron 1, *IRF5* SNP 6 ($P = 0.01$) in the 3'-UTR, and both *IRF5* SNP 7 ($P = 0.008$) and *IRF5* SNP 8 ($P = 0.007$) located 4.6 kb and 5 kb downstream from the 3'-UTR, respectively. The highest transmission ratios were observed for *IRF5* SNP 6 and *IRF5* SNP 7, the common allele of *IRF5* SNP 6 was over-transmitted in a ratio of 1:1.77 and the minor allele of *IRF5* SNP 7 was over-transmitted (1.64:1). However, the statistical significance was reduced at *IRF5* SNP 7 due to its lower MAF and hence fewer observations. The results from FBAT are also given in Table 1 and confirm the associations described for GH-TDT. Similar associations were found only when the European cohort was analysed (Supplementary Material, Table S1). There were trends for association in the same direction in the Indo-Asian cohort, but it was too small to generate significant results alone.

Haplotype-TDT analysis

The analysis of individual SNPs in *IRF5* shows the association with SLE for several variants spread across the length of the gene. To evaluate the correlation between *IRF5* haplotypes and SLE Haplotype-TDT using GENEHUNTER was performed (Table 2). Two over-transmitted haplotypes were identified, each with ~2-fold excess transmission to patients with SLE. There is some loss of linkage disequilibrium (LD) between *IRF5* SNPs 1 and 2 ($r^2 = 0.36$), so that *IRF5* SNP 1 does not segregate fully with the associated haplotypes. This explains why *IRF5* SNP 1 is not included as part of the over-transmitted haplotypes (3 and 7) illustrated in Figure 1B. The predominant association is the over-transmission of a TCTAACT sub-haplotype carried on haplotype 3 ($P = 5.8 \times 10^{-5}$, T:U = 1:1.92). There is also over-transmission of a TCTAATT sub-haplotype carried on haplotype 7 ($P = 0.015$, T:U = haplotype 1:2.25), which only differs from the 3 sequence at *IRF5* SNP 7. Haplotype 7 is identical to haplotype 1, except at *IRF5* SNP 5. *IRF5* SNP 5 is a tagging SNP in the second intron of *IRF5*. *IRF5* SNP 5 is unlikely to be the functional variant as it does not show a strong individual association, nor does haplotype 1, which it tags. There was a single

under-transmitted haplotype, AGTGAGTC, designated haplotype 2 ($P = 0.0095$, T:U = 1.47:1). The full haplotype is boxed in Figure 1 to distinguish it from haplotype 4, from which it differs at *IRF5* SNP 1.

Endophenotype analysis across *IRF5*

Endophenotype analysis was performed to check for heterogeneity of association between affected individuals carrying a particular endophenotype versus those who do not. The analysis was performed between two-SNP haplotypes *IRF5* SNP 2 and *IRF5* SNP 6, to maximize any potential effect, because the TA allelic combination for these variants is carried on both the over-transmitted haplotype. The families included in this analysis were those whose proband had renal disease ($n = 124$), possessed antibodies against Sm ($n = 12$), RNP ($n = 44$), Ro and/or La ($n = 105$), double-stranded DNA ($n = 58$), or IgG and/or IgM anticardiolipin ($n = 118$). Neither the families where the proband had renal disease nor those possessing one of the autoantibodies show a significant difference in transmission pattern when compared with those which did not carry each particular endophenotype ($P > 0.1$).

Correlation of CEPH GT with *IRF5* expression level

The association study in UK SLE families identified a single haplotype block across the gene, containing multiple associated haplotypes, each of which carried several associated alleles (Fig. 1A and B). One possible mechanism by which SNPs may influence SLE susceptibility is by genetically specifying cellular mRNA levels. To explore this hypothesis, we established whether SNPs in the *IRF5* gene were associated with variation in the level of mRNA extracted from a panel of CEPH B-lymphoblastoid cell lines.

Thirty CEPH samples, comprising the parents of 15 CEPH families were analysed for *IRF5* expression using data generated from the Affymetrix hgU133plus2 expression array. This array contains three probes targeting *IRF5*: 205468_s_at probes exons 2 through 5; 205469_s_at is located in the 3'-UTR (between nt1615 and 2150 of NM_002200) and 239412_at which hybridizes further downstream (nt2306–2809 of NM_002200). The expression levels detected by 205468_s_at correlated tightly with those obtained from 205469_s_at, so that in further analysis 205469_s_at was chosen to represent both. The positions of the two most informative probes are shown in Figure 2A: Probe 1 is used to abbreviate 205469_s_at, and Probe 2 to represent 239412_at. Expression levels among the 30 samples showed a surprisingly wide range: RNA levels detected by Probe 2 varied 8-fold, while expression measured by Probe 1 only varied ~3-fold (Supplementary Material, Table S2).

To search for variants underlying the expression variation in *IRF5*, the genomic DNA of the 30 CEPH parental samples was re-sequenced across all exons and splice junctions of *IRF5*. In addition, 1 kb of proximal promoter was also re-sequenced. These areas were selected since they offer the highest probability of causal changes or closely linked surrogates. Twenty-one polymorphisms were identified; these included three polymorphisms used in the association study (*IRF5* SNPs 2, 6 and 8). Of the 21 polymorphisms identified, seven showed an association with RNA expression levels at significance level $P < 0.001$ (Supplementary Material, Table S2). The location of these eight SNPs (labelled E1–E8) and their relationship to the SNPs genotyped in the UK samples are detailed in Figure 2A. Haplotypes present in CEPH and UK SLE families were similar (data not shown). The three SNPs from the CEPH analysis (rs2004640, rs10954213 and rs2280714) that were identical to SNPs genotyped in the UK families (*IRF5* SNP 2, *IRF5* SNP 6 and *IRF5* SNP 8) allowed us to directly superimpose the haplotypes in the two populations.

To examine the relationship between expression levels and the over-transmitted haplotypes from the UK SLE families, we constructed three-marker haplotypes in a total of 91 CEPH individuals, including the 30 CEPH parental samples used for re-sequencing and an additional 61 grandparental samples, using *IRF5* SNPs 2, 6 and 8. These three variants were among the polymorphisms showing association with SLE, with over-transmission of the T alleles of both *IRF5* SNP 2 and *IRF5* SNP 8 and also the A allele of *IRF5* SNP 6 (Table 1). The three-marker haplotypes inferred in the CEPH individuals are described in Table 3 together with the relative expression values from the two Affymetrix probes (Probes 1 and 2).

Figure 2B shows a graph of the relative expression level for samples carrying none, one or two copies of the TAT (SNP 2, 6 and 8) haplotype against genotype, with separate plots for the two expression probes. There was a clear dose-dependent relationship between the level of *IRF5* mRNA expression and the number of over-transmitted TAT haplotypes in an individual. Comparison of this expression difference by one-way ANOVA showed a greater dose-dependent response for Probe 2 (239412_at) ($P = 4.4 \times 10^{-10}$) than for Probe 1 (205469_s_at) ($P = 7.5 \times 10^{-8}$). Individuals with two copies of the over-transmitted TAT haplotype showed increased expression with Probe 1 (205469_s_at), when compared with those samples having only one copy of the TAT haplotype. The samples carrying no TAT haplotypes, expressed the least *IRF5* mRNA, as measured by Probe 1. Since both over-transmitted haplotypes in the UK SLE samples carry the TAT alleles of *IRF5* SNPs 2, 6 and 8, these data provide a direct link between the over-transmitted SLE haplotypes and an increase in the level of mRNA expression. For Probe 2 (239412_at), the dose-dependent effect was in the opposite direction to that seen for Probe 1. The maximal expression of *IRF5* mRNA was seen in samples with no copies of the TAT haplotype, with the minimum level of expression observed for those samples with two TAT haplotypes.

The differential correlations of the two probes suggested that the two probes recognized different transcript variants of *IRF5*; this possibility prompted us to look at the pattern of association for *IRF5* expression for each individual polymorphism across *IRF5*. The pattern of these associations is presented in Figure 3C, in which the negative $\log_{10}$ P -value of the correlation of each marker with expression level is plotted. The figure clearly demonstrates a peak of association for *IRF5* SNP 6 (rs10954213), with a P -value of 1.17×10^{-5} with Probe 1 and P -value of 1.02×10^{-14} with Probe 2. *IRF5* SNP 6, which lies in the interval between the two expression probes, is the most likely candidate identified for this key determinant in the differential regulation of transcript variants. Previous reports showed a strong association between the polymorphism rs2280714, located ~4.5 kb downstream of *IRF5*, and expression levels of *IRF5*. To compare this variant with *IRF5* SNP 6, we genotyped the 30 CEPH parents for rs2280714 (*IRF5* 8). The observed association with *IRF5* expression was more significant than the association observed with *IRF5* SNP 2, but considerably weaker than *IRF5* SNP 6, supporting the conclusion that *IRF5* SNP 6 is a major determinant of *IRF5* mRNA levels.

To assess the magnitude of the influence of *IRF5* SNP 6 on gene expression, we quantified the difference in the strength of association for *IRF5* SNP 6 with each of the two probes. We also extended the sample set by including 61 grandparents from unrelated CEPH pedigrees, to give a total of 91 CEPH individuals. The CEPH samples were categorized according to *IRF5* SNP 6 genotype and the mean expression for each sample group was calculated separately for each probe (Fig. 3D). This analysis shows that for Probe 1 (205469_s_at), there is an ~2-fold increase in expression level for the A/A homozygotes compared with the G/G homozygotes (mean expression levels for A/A 9.96, A/G 9.58, G/G 9.08). The difference is more marked, yet inverse, for Probe 2 (239412_at), with a greater than 5-fold decrease in the level of expression of the A/A homozygotes, compared with the G/G

homozygotes (mean expression levels for A/A 7.92, AG 9.69, G/G 10.42). The most striking point is the reversal of direction of this relationship, which suggests that the over-transmitted A allele of *IRF5* SNP 6 has a key role in the regulation of gene expression for SLE.

In order to further explore the relationships between genotypes and mRNA expression, we applied a graphical modelling technique (22,23), a form of multivariate analysis that uses graphs to represent models. A graphical model consists of a set of vertices and edges (Fig. 4). Vertices correspond to the variables, and edges represent direct dependencies between corresponding variables. We used the following variables for graphical modelling: genotypes at SNPs E1, E3, E4 (*IRF5* SNP 2), E5, E6 (*IRF5* SNP 6) and E8 (rs2280714), mRNA expression levels as measured by Probe 1 (205469_s_at), Probe 2 (239412_at) and 205468_s_at, and sex of an individual (Fig. 4). We excluded SNPs E2 and E7 from the analysis, because within the dataset SNP E2 is in complete equilibrium with E1, and SNP E7 is in complete equilibrium with E5. The results of the analysis are shown in Figure 4. This graphical model shows that there are only two SNPs, E5 and E6, directly associated with two sets of probes (Probes 1 and 2). This corroborates the ANOVA results. The model also shows that the other SNPs are in LD with these two, and that sex is independent of the other variables. SNP E8, previously the strongest correlate for *IRF5* expression, is not an independent marker for *IRF5* expression. However, it is directly related to SNP 6, indicating its strong relationship to *IRF5* expression. Interestingly, the disequilibrium between all the SNPs is represented by a linear graph. Probe set 205468_s_at is directly associated only with probe set 205469_s_at. This is consistent with the assumption that both of these two probe sets measure total transcription of all variants.

Mechanism for the differential expression pattern in the 3'-UTR of *IRF5*

Having established that there was a differential pattern of expression associated with *IRF5* SNP 6, we tried to establish the mechanism(s) that could explain these findings. When the sequence context around this variant was investigated in greater detail, we observed that *IRF5* SNP 6 lies in a potential polyadenylation site [AAT(*G/A*)AA]. Further examination revealed that the over-transmitted A allele of *IRF5* SNP 6 creates a novel polyadenylation site (AATAAA) (Fig. 3A). It was predicted that the two alleles of *IRF5* SNP 6 would create alternative polyadenylation patterns of the *IRF5* transcripts, as described in Figure 3B. Samples which are homozygous for the over-transmitted A allele are predicted to preferentially express more of a shorter transcript, whereas G/G homozygotes will produce mRNAs with a longer 3'-UTR.

To investigate this hypothesis, we designed a set of RT-PCR primers that would amplify the longer of the two transcripts (Fig. 3). EBV B cell RNA was analysed from the CEPH parental samples, with the individuals being selected on the basis of their genotype at *IRF5* SNP 6. Eleven A/A homozygotes and three G/G homozygotes were studied by quantitative real-time PCR and the results normalized to three housekeeping genes (Fig. 5). The analysis reveals that the G/G homozygotes generate more than 30-fold higher levels of the longer *IRF5* transcript than A/A homozygotes ($P < 10^{-4}$).

DISCUSSION

Haplotype structure in UK SLE families for *IRF5*

As part of the family-based association study in the UK SLE families, the haplotypes across the gene were constructed. Figure 1B illustrates that the haplotype pattern across *IRF5* captures 91.5% of the total available haplotypes, with the three associated haplotypes, 2, 3 and 7 accounting for 35.1% of all analysed chromosomes. Data from the HAPMAP project was used to look for evidence of LD between *IRF5* and its immediate neighbouring genes.

This comparison showed that there is a clear recombination ($r^2 = 0.04$) at the 5' end of the gene. This recombination point is located between *IRF5* and its closest neighbouring gene, the hypothetical protein FLJ33365, found 57 kb upstream. However, at the 3' end of the gene, the pattern of LD is more complex. The closest gene to *IRF5*, located 6 kb away, is *TNPO3* (Transportin 3), which is involved in the nuclear transport of the pre-mRNA SR splicing proteins (19). There is a drop in LD across the *IRF5*-*TNPO3* intergenic region, but there are intermittent increases in LD between *IRF5* (*IRF5* SNPs 3 and 7) in HAPMAP, and several other HAPMAP variants across *TNPO3* ($r^2 > 0.75$). The variable pattern of LD between the two genes includes a spike of LD with *IRF5* SNPs 2 and 7, which are associated with SLE. This pattern of LD does not extend into the promoter, since for *IRF5* SNP 1, the r^2 values with variants in *TNPO3* are all < 0.46 . Both these factors suggest that by more extensive mapping of the haplotypes across *IRF5*, we may be able to break down the associated haplotypes across the gene.

Comparison of association across *IRF5* in UK SLE families and previous reports

The pattern of association for individual variants in the UK samples corroborates previously published reports (13,14), but there are also some differences as summarized in Table 4. The strongest association in the UK dataset (*IRF5* SNP 2, $P = 3.4 \times 10^{-4}$) directly corroborates the association observed both by Sigurdsson *et al.* ($P = 4.4 \times 10^{-7}$) and Graham *et al.* ($P = 6 \times 10^{-4}$). There is also a direct corroboration of the association for *IRF5* SNP 3 (13) but the signal is weaker than that seen for *IRF5* SNP 2. The main difference between the UK data and the published data is the lack of association from the promoter variant *IRF5* SNP 1, which is accompanied by a break in LD between *IRF5* SNP 1 and all of the down-stream SNPs in *IRF5* ($r^2 < 0.36$). In the UK families there are additional, strong signals further downstream, including *IRF5* SNP 4 ($P = 4.9 \times 10^{-4}$) and *IRF5* SNP 7 ($P = 0.007$). Taken together, these data suggest a downstream shift in the overall pattern of strongest association in the UK samples when compared with both of the published reports, which may reflect multiple functional effects on the gene carried on the over-transmitted haplotype in different populations.

Polyadenylation variation in *IRF5*

We used two Affymetrix probes to measure the levels of RNA expression, both of which are located in the 3'-UTR of *IRF5*. Seven polymorphisms across *IRF5* were examined in 30 CEPH parental samples and a correlation between genotype and expression levels was observed (Fig. 3C). In an extended data set, which included 30 CEPH parental samples and 61 unrelated grandparents, a correlation was also found between expression level and *IRF5* haplotype (Fig. 2B). We demonstrated that the haplotypes showing over-transmission in the UK SLE families tended to generate increased levels of *IRF5* mRNA expression as assayed using the 5'-probe (Probe 1). Surprisingly these same haplotypes were associated with decreased expression assayed using the 3'-probe (Probe 2). This intervening region between the two probes contained a polymorphism not previously studied in SLE, *IRF5* SNP 6, which creates a novel polyadenylation site. In samples carrying the over-transmitted A allele, a shortened transcript is produced, due to premature polyadenylation. However, in samples containing the G allele, a longer transcript is seen. We therefore hypothesize that creation of this novel polyadenylation site is a major functional mechanism in the regulation of *IRF5* expression.

We confirmed this theory using RT-PCR, using two amplicons which show a probe-specific pattern of amplification. The primers for Product 1 amplify a sequence 5' to Probe 1 (Fig. 3A) and can be used to detect both the shorter, prematurely polyadenylated transcript, from samples containing allele A of *IRF5* SNP 6 and also the full-length transcript from samples containing the G allele of this variant. However, the primers for Product 2 will preferentially

amplify samples containing the G allele of *IRF5* SNP 6. Expression Probe 1 (205489_s_at), which is 5' to the novel polyadenylation site is able to detect the expression of both polyadenylated isoforms, and Probe 2 (239412_at) is only able to pick up the longer mRNA transcripts, carrying the G allele (AATGAA). The function of the poly A site at *IRF5* SNP 6 is 'leaky' since there is not a complete absence of expression for the 239412_at transcript in the A/A homozygote samples, implying that there may be some poly A function for the AATGAA sequence. Changes in the poly(A) hexamer can strongly affect processing (20), so that AATGAA has ~4% wild-type processing activity *in vitro*, but *in vitro* effects do not always correlate with functionality *in vivo*. Non-canonical polyadenylation signals are often found in transcripts with tissue-specific or alternative polyadenylation.

Therefore, the location of the two probes and the RT-PCR amplicons enabled us to determine that the novel polyadenylation site was a major determinant in *IRF5* gene expression. A recent publication (14) showed that elevated expression of *IRF5* was associated with the T allele of *IRF5* SNP 8 (rs2280714), using two expression probes. *IRF5* SNP 6 was not included in this study. Graham *et al.* found the greatest increase in expression in samples carrying the T alleles of both *IRF5* SNPs 2 and 8, but that the 3'-flanking *IRF5* SNP 8 was a better predictor of expression levels than *IRF5* SNP 2. However, from the data presented in this article, we argue that it is *IRF5* SNP 6 and not *IRF5* SNP 8 which is the functional variant affecting gene expression. Both studies show a pattern of strong LD across the gene, but in our UK SLE samples, we have presented a denser map of variants to create haplotypes from eight SNPs rather than four, which enabled us to more accurately describe the haplotype structure across the gene. Both studies also report the association of an over-transmitted SLE haplotype with elevated levels of *IRF5*. However, it is the identity of the 3'-SNP and also the location of the expression probes which are the key determinants in unravelling the mechanism underlying the increased expression. In the same study (14), both expression probes were located upstream of the novel polyadenylation site. They were Probe 1 and a second probe, 36465_at, which overlaps with Probe 1. By providing a molecular mechanism, our data indicate that it is *IRF5* SNP 6 and not *IRF5* SNP 8 which determines *IRF5* mRNA levels.

Multiple functional effects of associated *IRF5* haplotype

In this article, we have shown that there are multiple associated haplotypes across *IRF5*, which carry several associated variants. We have gone on to show that the over-transmitted haplotypes have a clear effect on the expression of *IRF5* mRNA, and described the creation of a novel polyadenylation site at *IRF5* SNP 6, which contributes to the association from this gene. The T allele of *IRF5* SNP 2 creates a GT donor splice site for exon 1b, with preferential production of a splice variant from exon 1b (14). The associated alleles of *IRF5* SNPs 2 and 6 are in strong LD and presumably tend to augment *IRF5* expression by different, albeit complementary mechanisms. However, none of these functionally relevant alleles are unique to both over-transmitted haplotypes. Given the weight of functional data to support a pathogenic role for both *IRF5* SNP 6 and also *IRF5* SNP 2, it is likely that further fine-mapping of the associated haplotypes will uncover additional functional polymorphisms. The significant association observed for *IRF5* SNP 4 (Table 1), is most likely due to strong LD with two SNPs having functional significance, rather than being an additional functional effect in itself. The pair-wise LD (Fig. 1C) between *IRF5* SNP 4 and each of these SNPs is for *IRF5* SNP 2 ($r^2 = 0.76$, $D' = 0.91$) and *IRF5* SNP 6 ($r^2 = 0.61$, $D' = 0.97$). There may also be a contribution from *IRF5* SNP 7 in the 3'-flanking region, which is unique to the over-transmitted haplotype 3, but is not present on haplotype 7. Alternatively, since *IRF5* SNP 7 is in LD with variants in *TNPO3*, it is possible that there are additional, as yet unidentified, associated variants in this neighbouring gene. We did scrutinize the small CEPH expression cohort that we typed in this study to determine

whether any of the seven polymorphisms (SNPs E1–E7) were in strong LD with *IRF5* SNP 7. No polymorphisms showed evidence of LD with $r^2 > 0.2$.

In summary, the genetic effects defining the involvement of *IRF5* in SLE pathogenesis are likely to be complex, with multiple functional sequelae contributing to the overall association. We have defined multiple over-transmitted haplotypes which carry two or more of the putative functional alleles. Our current understanding suggests that the functional dysregulation of *IRF5* in SLE involves inappropriate polyadenylation, aberrant splicing and up-regulation of mRNA expression. However, the aetiologic alleles responsible for all of these effects have not yet been determined.

MATERIALS AND METHODS

Family collection

The laboratory possesses a large collection of SLE families of predominantly EC origin. The study cohort consisted of 379 UK parental-proband trios, of which 357 are EC and 22 of IP origin. All probands conformed to the ACR criteria for SLE (21), with the diagnosis being established by telephone interview, health questionnaire and details from clinical notes and written consent obtained from all participants, including relatives. Ethical approval was obtained from Multi-Centre Research Ethics Committee (MREC).

Selection of informative markers in *IRF5*

For the SLE association study, six of the markers in *IRF5* were taken from the recent publications by Sigurdsson *et al.* (13) and Graham *et al.* (14). Two additional SNPs were selected from the dbSNP database (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), to ensure good marker density across the gene, at a density of one SNP per 1.4 kb.

Additional markers for *IRF5* were identified by re-sequencing, SNPs E1–E7 (Fig. 2A). Nested primers were designed for *IRF5* exons and 1 kb of promoter, as referenced in NM_002200 and NM_032643 using genomic sequence from the UCSC genome browser. The primer sequences for this re-sequencing are given in Supplementary Material, Table S3. All secondary primers were tailed with M13 primer sequence to facilitate dye-primer sequencing. Amplicons covering exons and approximately 50 bases of intron sequence were amplified by PCR and then sequenced with dye-primer chemistry. Products were run on MegaBace1000 (Amersham, GE) capillary sequencers. Base-calling, automated polymorphism detection and calling of genotypes were performed with proprietary software. All samples, reference sequences, primers, polymorphisms and genotypes were saved to a relational database.

Genotyping methodology

Except for the details given below, all genotyping was performed using MALDI-TOF mass spectrometry, and analysis of the raw genotype data was carried out using the MassArray Typer v3.4 software (Sequenom, Hamburg, Germany). Details of the assay designs are available on request. Additional genotyping of *IRF5* SNP 8 (rs2280714) in CEPH individuals was performed by re-sequencing.

Correlation of affymetrix expression analysis with *IRF5* genotype

The expression data were generated using three probe sets in the 3'-UTR of *IRF5*, 205468_s_at, 205469_s_at and 239412_at, which were present on Affymetrix hgU133plus2 chips. EBV-transformed B-lymphoblast cell lines from 99 parents and unrelated grandparents of 34 CEPH families were obtained from Coriell Cell Laboratories (Camden, NJ, USA) and grown according to recommendations. RNA was isolated from ~80%

confluent cells using the Ambion RiboPure Kit (Ambion, Austin, TX, USA). RNA yield was quantified using a Nanodrop ND-1000 and RNA integrity verified on a BioAnalyzer 2100 (Agilent, Foster City, CA, USA).

Template preparation, using 5 µg of total RNA as starting material, followed by microarray hybridization, wash, stain and scan was carried out as per the manufacturers recommended protocol for Affymetrix hgU133 Plus 2.0 microarrays. Probe set summaries of normalized signal intensities in base 2 logarithm were produced by applying the RMA method found in the R1.9.1 package, Affy (Bioconductor project <http://www.bioconductor.org>). The CEPH samples were categorized according to genotype or haplotype. Correlation of the mRNA levels with the genotype/haplotypes of SNPs E1–7 was carried out by plotting the level of expression versus genotype. Associations between each SNP genotype and probe set expression levels were tested by ANOVA (S-PLUS 7.0.3, 2005, Insightful Corp., Seattle, WA, USA).

Graphical modelling was performed with MIM 3.2.0.1 (22) using an R interface package, mimR (23). A step-wise procedure in decomposable mode was used with forward selection. In forward selection, the edge with the smallest *P*-value less than the critical value (0.05) was successively added to the current model until no more significant edges could be found.

Quantification of polyadenylation in *IRF5*

cDNA was prepared from EBV-transformed CEPH lymphocytes obtained from the Coriell Institute. These cells were cultured at 1×10^6 cells/ml in RPMI 1460 medium supplemented with 10% fetal calf serum, 2 mM L-glutamine, penicillin (100 units/ml) and streptomycin (100 µg/ml). Total RNA was prepared from lymphocytes stored in Trizol, using a phenol-chloroform method or using RNeasy spin columns (Qiagen, Crawley, UK). The RNA was quantified using the Nanodrop ND-1000 Spectrophotometer (Labtech International, Rigmer, UK) and the quality was checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, Stockport, UK). Total RNA of 2 µg extracted from the cultured cells was pre-treated with 1 unit of DNase I (Invitrogen Life Technologies, Refrew, UK) to remove any traces of genomic DNA contamination. The first strand cDNA was synthesized from this pre-treated RNA by priming with oligo(dT)_{12–18} primer using 200 units of Superscript II (Invitrogen Life Technologies), according to the manufacturers instructions. RT-PCR primers were designed to amplify a 273 bp product in the 3′-UTR of *IRF5*—the distal primer binding 3′ to *IRF5* SNP 6, such that the reaction selectively detects the longer transcript (primer sequences available on request). Quantification of mRNA transcripts was carried out using the Absolute QPCR SYBR Green kit (ABgene, Epsom, UK), according to the manufacturer's instructions on the 7500 Fast Real-Time (ABI). All samples were analysed in triplicates. The amount of each product was normalized against three housekeeping genes: *NONO*, *GAPD* and *HPRT1* and analysed by permutation testing with REST 2005 software (Corbett Life Sciences).

Statistical analysis

All sample genotype and phenotype data was managed by, and analysis files generated with the BC/GENE and BC/CLIN software (Biocomputing Platforms Ltd, Finland). Markers were excluded from the analysis if they showed <90% genotyping frequency, the HWE *P*-value was less than 0.05 and/or greater than 5% families in the study cohort showed sporadic Mendelian errors. Haplotype patterns were constructed using Haploview (24), using the Gabriel *et al.* algorithm (25), with genotype data from the parental chromosomes of 358 EC and 22 IP parental-proband trios. This programme constructs haplotypes based on the *D'* measure of LD (26), together with an LOD score as a measure of significance and 95% CI to state the accuracy of the *P*-value. *r*² values were used (27,28) to confirm the pair-

wise LD for SNPs across each gene and to refine the overall haplotypic architecture. Only markers having an MAF >5% were included in the haplotype constructions. The haplotype block definition for both genes was based on CI for strong LD of 0.98 (upper) and 0.70 (lower), upper CI maximum for strong recombination of 0.90, at least 95% of strong LD in informative comparisons. TDT analyses to compare the observed and expected transmission of alleles from heterozygous parents to affected offspring were performed using GENEHUNTER 2.0 beta (29,30) and FBAT (31,32).

Endophenotype TDT analysis was performed on families where the affected individual had renal disease. A separate analysis was carried out for those families where the probands produced anticardiolipin autoantibodies or antibodies to double-stranded DNA, and to the RNA-associated antigens, Ro, La, Sm and RNP. Familial clustering has been shown in our families for the presence of some of these autoantibodies (manuscript in preparation). This analysis was performed on the *IRF5* SNP 2 and SNP 6 two-SNP haplotype. We addressed this by noting that the test statistic used by TRANSMIT is produced by summing score contributions over families. Thus, if the genetic effect is the same in each family, these score contributions should be exchangeable: in particular they should have the same mean. We tested this for the binary endophenotypes using a Welch two-sample *t*-test to compare the mean contribution in families where the proband has the endophenotype with that in families where the proband does not. Permutation tests provide robust alternative tests and give similar results on these data. Our aim was to investigate whether the transmission bias was greater in individuals with a given endophenotype. For this analysis, haplotypes for each individual were constructed for each CEPH family member using PHASE.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the Wellcome Trust through a Senior Fellowship awarded to Timothy J. Vyse. We acknowledge the work of Paul Spencer and Dr Andrew Wong in recruiting patients and families into the study and we would like to thank our clinical colleagues for helping us recruit study participants. Our thanks and appreciation is extended to all the patients and their relatives for generously donating blood samples and all the general practitioners and practice nurses for collecting them: many thanks to Professor John Whittaker for his advice concerning the statistical analysis and to Dr David Perkins for maintaining the computerized database and for Dr M Fernando for critical reading of the manuscript.

REFERENCES

1. Rus, V.; Hochberg, MC. Dubois' Lupus Erythematosus. Wallace, DJ.; Hahn, BH., editors. Lippincott, Williams and Wilkins; Baltimore: 2002. p. 65-83.
2. Vyse TJ, Todd JA. Genetic analysis of autoimmune disease. *Cell*. 1996; 85:311–318. [PubMed: 8616887]
3. Deapen D, Escalante A, Weinrib L, Horwitz D, Bachman B, Roy-Burman P, Walker A, Mack TM. A revised estimate of twin concordance in systemic lupus erythematosus. *Arth. Rheum.* 1992; 35:311–318. [PubMed: 1536669]
4. Gaffney PM, Kearns GM, Shark KB, Ortmann WA, Selby SA, Malmgren ML, Rohlf KE, Ockenden TC, Messner RP, King RA, et al. A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. *Proc. Natl Acad. Sci. USA*. 1998; 95:14875–14879. [PubMed: 9843983]
5. Gaffney PM, Ortmann WA, Selby SA, Shark KB, Ockenden TC, Rohlf KE, Walgrave NL, Boyum WP, Malmgren ML, Miller ME, et al. Genome screening in human systemic lupus erythematosus: results from a second Minnesota cohort and combined analyses of 187 sib-pair families. *Am. J. Hum. Genet.* 2000; 66:547–556. [PubMed: 10677315]

6. Gray-McGuire C, Moser KL, Gaffney PM, Kelly J, Yu H, Olson JM, Jedrey CM, Jacobs KB, Kimberly RP, Neas BR, et al. Genome scan of human systemic lupus erythematosus by regression modeling: evidence of linkage and epistasis at 4p16–15.2. *Am. J. Hum. Genet.* 2000; 67:1460–1469. [PubMed: 11078476]
7. Shai R, Quismorio FP Jr, Li L, Kwon OJ, Morrison J, Wallace DJ, Neuwelt CM, Brautbar C, Gauderman WJ, Jacob CO. Genome-wide screen for systemic lupus erythematosus susceptibility genes in multiplex families. *Hum. Mol. Genet.* 1999; 8:639–644. [PubMed: 10072432]
8. Moser KL, Neas BR, Salmon JE, Yu H, Gray-McGuire C, Asundi N, Bruner GR, Fox J, Kelly J, Henshall S, et al. Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. *Proc. Natl Acad. Sci. USA.* 1998; 95:14869–14874. [PubMed: 9843982]
9. Lindqvist AK, Steinsson K, Johanneson B, Kristjansdottir H, Arnasson A, Grondal G, Jonasson I, Magnusson V, Sturfelt G, Truedsson L, et al. A susceptibility locus for human systemic lupus erythematosus (hSLE1) on chromosome 2q. *J. Autoimmun.* 2000; 14:169–178. [PubMed: 10677248]
10. Johanneson B, Steinsson K, Lindqvist AK, Kristjansdottir H, Grondal G, Sandino S, Tjernstrom F, Sturfelt G, Granados-Arriola J, Cacer-Varela J, et al. A comparison of genome-scans performed in multicase families with systemic lupus erythematosus from different population groups. *J. Autoimmun.* 1999; 13:137–141. [PubMed: 10441178]
11. Ronnblom L, Alm GV. Systemic lupus erythematosus and the type I interferon system. *Arth. Res. Ther.* 2003; 5:68–75. [PubMed: 12718746]
12. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, Shark KB, Grande WJ, Hughes KM, Kapur V, et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl Acad. Sci. USA.* 2003; 100:2610–2615. [PubMed: 12604793]
13. Sigurdsson S, Nordmark G, Goring HH, Lindroos K, Wiman AC, Sturfelt G, Jonsen A, Rantapaa-Dahlqvist S, Moller B, Kere J, et al. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am. J. Hum. Genet.* 2005; 76:528–537. [PubMed: 15657875]
14. Graham RR, Kozyrev SV, Baechler EC, Reddy MV, Plenge RM, Bauer JW, Ortmann WA, Koeuth T, Gonzalez Escribano MF, Argentine and Spanish Collaborative Group, Pons-Estel B, Petri M, Daly M, Gregersen PK, Martin J, Altshuler D, Behrens TW, Alarcon-Riquelme ME. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat. Genet.* 2006; 38:550–555. [PubMed: 16642019]
15. Taniguchi T, Ogasawara K, Takaoka A, Tanaka N. IRF family of transcription factors as regulators of host defense. *Ann. Rev. Immunol.* 2001; 19:623–655. [PubMed: 11244049]
16. Barnes BJ, Moore PA, Pitha PM. Virus-specific activation of a novel interferon regulatory factor, IRF-5, results in the induction of distinct interferon alpha genes. *J. Biol. Chem.* 2001; 276:23382–23390. [PubMed: 11303025]
17. Izaguirre A, Barnes BJ, Amrute S, Yeow WS, Megjugorac N, Dai J, Feng D, Chung E, Pitha PM, Fitzgerald-Bocarsly P. Comparative analysis of IRF and IFN-alpha expression in human plasmacytoid and monocyte-derived dendritic cells. *J. Leukoc. Biol.* 2003; 74:1125–1138. [PubMed: 12960254]
18. Takaoka A, Yanai H, Kondo S, Duncan G, Negishi H, Mizutani T, Kano S, Honda K, Ohba Y, Mak TW, et al. Integral role of IRF-5 in the gene induction programme activated by Toll-like receptors. *Nature.* 2005; 434:243–249. [PubMed: 15665823]
19. Lai MC, Lin RI, Huang SY, Tsai CW, Tarn WY. A human importin-beta family protein, transportin-SR2, interacts with the phosphorylated RS domain of SR proteins. *J. Biol. Chem.* 2000; 275:7950–7957. [PubMed: 10713112]
20. Sheets MD, Ogg SC, Wickens MP. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res.* 1990; 18:5799–5805. [PubMed: 2170946]

21. Tan EM, Cohen AS, Fries JF, Masi AT, McShane DJ, Rothfield NF, Schaller JG, Talal N, Winchester RJ. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arth. Rheum.* 1982; 25:1271–1277. [PubMed: 7138600]
22. Edwards, D. *Introduction to Graphical Modelling*. Springer-Verlag; New York: 2000.
23. Hojsgaard S. The mimR package for graphical modelling in R. *J. Stat. Softw.* 2004; 11:1–13.
24. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005; 21:263–265. [PubMed: 15297300]
25. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. The structure of haplotype blocks in the human genome. *Science.* 2002; 296:2225–2229. [PubMed: 12029063]
26. Lewontin RC. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics.* 1964; 49:49–67. [PubMed: 17248194]
27. Hill WG, Weir BS. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* 1994; 54:705–714. [PubMed: 8128969]
28. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.* 1995; 29:311–322. [PubMed: 8666377]
29. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and non-parametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* 1996; 58:1347–1363. [PubMed: 8651312]
30. Kruglyak L, Lander ES. Faster multipoint linkage analysis using Fourier transforms. *J. Comput. Biol.* 1998; 5:1–7. [PubMed: 9541867]
31. Horvath S, Xu X, Laird NM. The family-based association test method: strategies for studying general genotype–phenotype associations. *Eur. J. Hum. Genet.* 2001; 9:301–306. [PubMed: 11313775]
32. Steen KV, Lange C. PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum. Genom.* 2005; 2:67–69.

Figure 1.

Genomic organization, haplotype architecture and haplotype-TDT across the human *IRF5* gene. (A) The exons are marked with black boxes and numbered above. The untranslated alternative 5'-exons are labelled 1a-c and the 3'-UTR is shown as a white box. The polymorphisms selected for the study are numbered 1-8 below the gene. (B) The haplotype block structure across *IRF5*, as constructed using Haploview, from 357 EC parent-proband trios and 22 IP trios. There is a single haplotype block across the gene, consisting of seven haplotypes, numbered 1-7 down the left-hand side. The haplotype frequency is shown to the right of each haplotype. The marker numbers are shown above the haplotypes and correspond to the SNPs illustrated in (A). The haplotype-tagging SNPs are indicated by asterisks under the marker numbers. The over-transmitted haplotype 3 and haplotype 7 are outlined by solid boxes. The under-transmitted haplotype 2 is outlined by a dashed box; the box incorporates *IRF5* SNP 1, to distinguish it from haplotype 4. The over-transmitted rare allele of *IRF5* SNP 7 which is unique to over-transmitted haplotype 3 is indicated in bold type. (C) The pair-wise pattern of LD across *IRF5* is shown in a matrix. The left-hand diagram shows the pair-wise LD quantified by correlation, r^2 and the right-hand diagram the D' plot. Stronger LD is depicted graphically by denser shading.

Figure 2.

Correlation of RNA expression level with genotype for variants across *IRF5* in CEU CEPH individuals. **(A)** The gene diagram of *IRF5* shows the location of the SNPs genotyped in the UK SLE families (*IRF5* SNPs 1–8) as in Figure 1A. The expression variants (E1–8) identified by sequencing in the CEPH individuals and correlate with *IRF5* mRNA levels are marked by an asterisk. They are shown directly underneath the SNPs used in the association analysis. The location of the expression probe sets, Probe 1 (205469_s_at) and Probe 2 (239412_at) are shown above exon 9. **(B)** This panel illustrates the correlation of the level of *IRF5* mRNA with the TAT haplotype (defined by *IRF5* SNP 2, *IRF5* SNP 6 and *IRF5* SNP 8) in 91 CEPH family samples, with Probe 1 (205469_s_at) and Probe 2 (239412_at) RNA expression probes.

Figure 3.

A novel polyadenylation site in *IRF5*. **(A)** Diagrammatic representation of the 3'-UTR of *IRF5*, showing the location of the Affymetrix expression probes, designated Probe 1 and Probe 2 in greater detail than Figure 2. The position of the RT-PCR amplicon relative to *IRF5* SNP 6 (rs10954213) is also shown. **(B)** Alternatively polyadenylated transcripts from *IRF5* are illustrated, the size being dependent on the allele at *IRF5* SNP 6. **(C)** Strength of association between RNA expression levels across for each expression SNP across *IRF5* for two expression probes 205469_s_at and 239412_at. For each probe, the $-\log_{10}$ *P*-value of the association is plotted against marker number (E1–8). **(D)** Quantification of expression differences for *IRF5* SNP 6 for the two Affymetrix expression probes in 30 CEPH parental samples. The samples were grouped according to genotype at *IRF5* SNP 6 and the mean expression of each sample group was calculated separately for each probe. The fractional change (as a percentage) in the signal from each probe is shown between the A/A and G/G homozygotes.

Figure 4.

Graphical model representing relationships between SNP genotypes and mRNA expression. Open circles correspond to three probe sets from Affymetrix HGU-133 Plus 2.0 GeneChip. Filled circles correspond to genotypes at five SNPs [SNP E1 (rs3757385), SNP E3 (Ins GGGGC), SNP E4 (rs2004640/*IRF5* SNP 2), SNP E5 (30 bp del), SNP E6 (rs10954213/*IRF5* SNP 6), SNP E8 (rs2280714)] and sex. A line connecting two circles represents direct association between the two corresponding variables.

Figure 5.

Polyadenylation polymorphism and the expression of the long transcript. The relative units of expression of the longer *IRF5* transcript are represented on a log₂ scale in a box and whisker plot, showing 95%CI. The expression levels are normalized with respect to *HPRT1*. The *IRF5* genotype, either A/A or G/G represents the allelic composition at *IRF5* SNP 6, which influences polyadenylation.

Table 1Analysis of single SNPs across *IRF5* by GH-TDT and FBAT

SNP	dbSNP (rs)	MAF	Minor allele	GH-T:U	GH-T:U ^N	GH: χ^2	GH: <i>P</i> -value	FBAT: <i>Z</i>	FBAT: <i>P</i> -value
1	rs729302	0.26	C	127:153	1:1.20	2.41	0.12	-1.50	0.13
2	rs2004640	0.38	G	132:197	1:1.49	12.8	3.4×10^{-4}	-3.64	2.7×10^{-4}
3	rs752637	0.31	T	136:180	1:1.32	6.13	0.013	-2.49	0.013
4	rs3807306	0.40	G	146:212	1:1.45	12.2	4.9×10^{-4}	23.46	5.4×10^{-4}
5	rs1874328	0.38	G	146:167	1:1.14	1.41	0.23	21.24	0.22
6	rs10954213	0.35	G	70:104	1:1.77	6.64	0.010	-2.32	0.02
7	rs10488631	0.16	C	74:45	1.64:1	7.07	0.008	-2.62	0.009
8	rs2280714	0.28	C	134:182	1:1.35	7.29	0.007	-2.42	0.015

TDT analysis for individual SNPs across *IRF5* in 379 parent-affected offspring trios, by GH-TDT and FBAT are shown. The identity of the minor allele and the MAF is listed for each polymorphism. The column marked 'GH-T:U' shows the number of transmitted to untransmitted families for GH-TDT in the T:U^N column, the ratio has been normalized to 1. The direction of the transmission is indicated by the direction of the T:U ratio, only *IRF5* SNP 7 shows over-transmission of the rare allele. χ^2 for GH-TDT and the *Z* score for FBAT are shown. For both GH-TDT and FBAT, all *P*-values are quoted with one degree of freedom.

Table 2Haplotype-TDT in the *IRF5* gene

Haplotype	Frequency	T:U	P-value	Over-/Under-transmission
TCTAACT	0.149	107:56	5.8×10^{-5}	Over
TCTAATT	0.029	27:12	0.015	Over
AGTGAGTC	0.176	78:114	0.0095	Under

The table summarizes GH-TDT analysis across *IRF5* in a data set consisting of 379 parental-proband trios. The allelic composition and haplotype frequency is shown for each putative associated haplotype. The column marked 'T:U' represents the transmitted:untransmitted ratio of the haplotype in the informative families.

Table 3

Quantification of genotype with expression data in individuals from the CEPH families

Sample ID	Affymetrix RNA probe		Genotype data			Haplotype	Haplotype counts per sample
	Probe 1 (205469)_s_at	Probe 2 (239412)_at	IRF5 SNP 2	IRF5 SNP 6	IRF5 SNP 8		
234911	9.33	10.68	G/T	G/G	C/T	TGT/GGC	0
234912	9.57	9.77	G/G	G/A	C/T	GAT/GGC	0
234907	9.10	10.04	G/G	G/G	C/C	GGC/GGC	0
234908	8.76	10.31	G/G	G/G	C/C	GGC/GGC	0
234909	9.39	9.63	T/T	G/A	T/T	TAT/TGT	1
234957	10.02	8.42	G/T	A/A	T/T	TAT/GAT	1
234920	9.82	10.17	G/T	G/A	C/T	TAT/GGC	1
234921	10.13	10.33	G/T	G/A	C/T	TAT/GGC	1
234922	9.87	7.59	G/T	A/A	T/T	TAT/GAT	1
234923	9.57	9.69	G/T	G/A	C/T	TAT/GGC	1
234924	10.48	9.06	T/T	A/A	T/T	TAT/TAT	2
234925	9.98	10.19	G/T	G/A	C/T	TAT/GGC	1
234926	9.80	10.22	G/T	G/A	C/T	TAT/GGC	1
234927	9.83	8.10	T/T	A/A	T/T	TAT/TAT	2
234942	9.63	9.49	G/T	G/A	C/T	TAT/GGC	1
234937	10.10	8.40	T/T	A/A	T/T	TAT/TAT	2
234938	9.70	9.83	G/T	G/A	C/T	TAT/GGC	1
234939	10.05	10.13	G/T	G/A	C/T	TAT/GGC	1
234973	9.09	10.25	G/G	G/G	C/C	GGC/GGC	0
234949	10.13	8.42	G/T	A/A	T/T	TAT/GAT	1
234950	9.70	9.85	T/T	G/A	T/T	TAT/TGT	1
234951	10.01	7.77	G/G	A/A	T/T	GAT/GAT	0
234952	9.71	9.90	G/T	G/A	T/T	TGT/GAT	0
234940	10.43	8.55	T/T	A/A	T/T	TAT/TAT	2
234962	9.12	10.57	G/G	G/G	C/C	GGC/GGC	0
234963	9.87	9.98	T/T	G/A	T/T	TAT/TGT	1
234964	9.71	9.99	G/G	G/A	C/T	GAT/GGC	0
234965	9.73	9.80	G/T	G/A	C/T	TAT/GGC	1
234966	9.93	7.55	T/T	A/A	T/T	TAT/TAT	2
234967	9.63	9.78	G/T	G/A	C/T	TAT/GGC	1
234958	9.69	9.77	G/T	G/A	C/T	TAT/GGC	1
234959	9.28	10.79	G/G	G/G	C/C	GGC/GGC	0
234961	9.67	9.88	G/T	G/A	C/T	TAT/GGC	1
234968	9.95	9.92	G/T	G/A	C/T	TAT/GGC	1
234969	9.73	9.75	G/T	G/A	C/T	TAT/GGC	1
234970	9.44	9.73	G/T	G/A	T/T	TGT/GAT	0
234971	9.76	9.91	T/T	G/A	T/T	TAT/TGT	1
234903	9.84	10.19	G/G	G/A	C/T	GAT/GGC	0

Sample ID	Affymetrix RNA probe		Genotype data			Haplotype	Haplotype counts per sample
	Probe 1 (205469)_s_at	Probe 2 (239412_at)	IRF5 SNP 2	IRF5 SNP 6	IRF5 SNP 8		
234904	9.78	10.20	G/T	G/A	C/T	TAT/GGC	1
234905	9.72	9.88	G/G	G/A	C/T	GAT/GGC	0
234906	10.35	7.84	G/T	A/A	T/T	TAT/GAT	1
234915	9.79	9.76	G/T	G/A	C/T	TAT/GGC	1
234916	9.68	9.86	G/T	G/A	C/T	TAT/GGC	1
234917	9.92	10.07	G/T	G/A	C/T	TAT/GGC	1
234918	9.24	9.38	G/G	G/A	C/T	GAT/GGC	0
234928	9.53	9.88	G/T	G/A	C/T	TAT/GGC	1
234929	9.70	9.97	G/T	G/A	C/T	TAT/GGC	1
234972	9.98	8.63	T/T	A/A	T/T	TAT/TAT	2
234930	9.06	9.31	G/T	G/A	C/T	TAT/GGC	1
234931	9.88	8.02	T/T	A/A	T/T	TAT/TAT	2
234932	9.67	9.99	G/G	G/A	C/T	GAT/GGC	0
234933	9.14	10.57	G/T	G/G	C/T	TGT/GGC	0
234934	9.87	7.45	G/T	A/A	T/T	TAT/GAT	1
234935	10.15	8.53	T/T	A/A	T/T	TAT/TAT	2
234936	9.60	9.73	G/T	G/A	C/T	TAT/GGC	1
234944	10.17	8.60	T/T	A/A	T/T	TAT/TAT	2
234945	10.09	8.68	T/T	A/A	T/T	TAT/TAT	2
234946	9.17	10.60	G/G	G/G	C/C	GGC/GGC	0
234947	9.84	7.58	T/T	A/A	T/T	TAT/TAT	2
234948	9.34	10.68	G/T	G/G	C/T	TGT/GGC	0
234953	9.61	9.66	G/T	G/A	T/T	TGT/GAT	0
10860	9.70	7.44	T/T	A/A	T/T	TAT/TAT	2
7057	9.80	7.47	T/T	A/A	T/T	TAT/TAT	2
10854	9.88	7.63	T/T	A/A	T/T	TAT/TAT	2
10848	10.18	7.70	T/T	A/A	T/T	TAT/TAT	2
12455	9.49	7.77	T/T	A/A	T/T	TAT/TAT	2
10859	10.11	7.85	T/T	A/A	T/T	TAT/TAT	2
10850	9.98	7.98	T/T	A/A	T/T	TAT/TAT	2
6989	10.24	8.00	G/T	A/A	T/T	TAT/GAT	1
10861	9.60	8.14	T/T	A/A	T/T	TAT/TAT	2
12471	10.21	8.50	T/T	A/A	T/T	TAT/TAT	2
12385	10.32	8.66	G/T	A/A	T/T	TAT/GAT	1
12801	9.13	9.24	G/T	G/A	C/T	TAT/GGC	1
6991	9.42	9.36	G/T	G/A	C/T	TAT/GGC	1
10853	9.46	9.42	G/T	G/A	C/T	TAT/GGC	1
7048	9.50	9.44	G/T	G/A	C/T	TAT/GGC	1
10849	9.53	9.53	G/T	G/A	C/T	TAT/GGC	1
7038	9.58	9.56	G/T	G/A	C/T	TAT/GGC	1

Sample ID	Affymetrix RNA probe		Genotype data			Haplotype	Haplotype counts per sample
	Probe 1 (205469)_s_at	Probe 2 (239412_at)	IRF5 SNP 2	IRF5 SNP 6	IRF5 SNP 8		
6984	9.42	9.59	G/T	G/A	C/T	TAT/GGC	1
10858	9.52	9.63	G/G	G/A	C/T	GAT/GGC	0
12708	9.46	9.77	G/T	G/A	C/T	TAT/GGC	1
6987	9.82	9.81	G/T	G/A	C/T	TAT/GGC	1
12707	9.78	9.90	G/T	G/A	C/T	TAT/GGC	1
12386	9.58	9.90	T/T	G/A	T/T	TAT/TGT	1
6990	9.80	9.92	G/T	G/A	C/T	TAT/GGC	1
12456	9.64	9.95	G/T	G/A	C/T	TAT/GGC	1
10852	9.72	10.00	T/T	G/A	T/T	TAT/TGT	1
10851	9.94	10.06	G/T	G/A	C/T	TAT/GGC	1
12802	8.82	10.17	G/G	G/G	C/C	GGC/GGC	0
12470	8.99	10.38	G/G	G/G	C/C	GGC/GGC	0
10857	9.42	10.72	G/T	G/G	C/T	TGT/GGC	0

The sample IDs are GM IDs of 91 CEPH family members. The columns, headed 'Probe 1 (205469_s_at)' and 'Probe 2 (239412_at),' display the relative expression levels of the two probes for each sample on a log₂ scale. The columns marked 'Genotype Data' show the genotype data for three markers across *IRF5*. The haplotypes for these markers were constructed using the phase algorithm in PLINK. The column marked 'Haplotype counts per sample' describes the number of copies per sample of the *IRF5* SNP 2–*IRF5* SNP 6–*IRF5* SNP 8 TAT haplotype.

Table 4Allele frequency comparison and pattern of association across *IRF5* with previously published reports

SNP	dbSNP rs	Location	Function	MAF			Association		
				UK	Sigurdsson <i>et al.</i>	Graham <i>et al.</i>	UK	Sigurdsson <i>et al.</i>	Graham <i>et al.</i>
1	rs729302 ^a	Promoter		0.27	0.29	0.31		6.7×10^{-3}	7×10^{-4}
2	rs2004640	Intron 1	Exon splicing	0.42	0.42	0.43	3.4×10^{-4}	4.4×10^{-7}	2.8×10^{-3}
3	rs752637 ^a	Intron 1		0.33	0.32	0.32	0.013	4.4×10^{-3}	
4	rs3807306	Intron 1		0.44	0.46		4.9×10^{-4}		
5	rs1874328	Intron 2		0.38	0.37				
6	rs10954213	3'-UTR	Polyadenylation	0.35			0.01		
7	rs10488631 ^a	3'-flank		0.16			8×10^{-3}		
8	rs2280714	3'-flank		0.31	0.28	0.28	7×10^{-3}		

^aSNP included as part of the HAPMAP project.

MAFs were calculated from all samples in the joint EC and IP UK SLE trios and compared with those calculated from case-control data given in Sigurdsson *et al.* (2005) and Graham *et al.* (2006), references (13,14). Variant rs2204640 is located at a GT consensus splice site for exon 1b splicing (I/E) and rs109554213 creates a poly A site.