Fine-scale mapping of the 6p25.3 chronic lymphocytic leukaemia susceptibility locus

Dalemari Crowther-Swanepoel1, Peter Broderick1, Yussanne Ma1, Lindsay Robertson1, Alan M. Pittman1, Amy Price1, Philip Twiss1, Jayaram Vijayakrishnan1, Mobshra Qureshi1, Martin J. S. Dyer2, Estella Matutes3, Claire Dearden3, Daniel Catovsky3 and Richard S. Houlston1,*

1Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK, 2Medical Research Council (MRC) Toxicology Unit, Leicester University, Leicester, UK and 3Section of Haemato-oncology, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK

Received October 2, 2009; Revised and Accepted January 27, 2010

A recent genome-wide association study of chronic lymphocytic leukaemia (CLL) has identified a susceptibility locus on 6p25.3 associated with a modest but highly significant increase in CLL risk. Using a set of single nucleotide polymorphism (SNP) markers, we generated a fine-scale map and narrowed the association signal to a 18 kb DNA segment within the 3′-untranslated region (UTR) of the IRF4 (interferon regulatory factor 4) gene. Resequencing this segment in European subjects identified 55 common polymorphisms, including 13 highly correlated candidate causal variants. In a large case–control study, it was shown that all but four variants could be excluded with 95% confidence. These four SNPs map to a 3 kb region of the 3′-UTR of IRF4, consistent with the causal basis of the association being mediated through differential IRF4 expression.

INTRODUCTION

Chronic lymphocytic leukaemia (CLL; MIM 151400) is characterized by a strong familial risk; however, the basis of genetic susceptibility to CLL is largely unknown (1). We have recently conducted a genome-wide association (GWA) study of CLL and shown that the single nucleotide polymorphism (SNP) rs872071 mapping to 6p25.3 influences CLL risk. The observation of a relationship between this SNP and CLL risk has been replicated in multiple independent series thereby providing compelling evidence for the association ((2) and unpublished data). Although the risk of CLL associated with the 6p25.3 locus is modest (per allele odds ratio (OR) = 1.54), the variant is common in European populations and so contributes significantly to the risk of developing CLL.

While rs872071 maps to IRF4 (interferon regulatory factor-4; MIM 601900), the underlying basis of the association is presently unknown. We have systematically interrogated the 6p25.3 association signal through targeted resequencing and linkage disequilibrium (LD) mapping to elucidate the basis of this association.

RESULTS

An association between 6p25.3 and CLL risk is provided by rs872071 and rs9378805 (2). rs872071 maps within the 3′-untranslated region (UTR) of IRF4 and rs9378805 localizes 6.3 kb centromeric to the 3′-UTR of IRF4. We performed logistic regression comparing the log-likelihood of models based on rs872071 and rs9378805. A model based solely on rs872071 genotype was sufficient to capture all of the locus association (i.e. when rs872071 was included in the model, adding rs9378805 did not improve the fit, conversely adding rs872071 to rs9378805 provided superior evidence for an association). Using HapMap data, we imputed all untyped SNPs mapping to a 200 kb region encompassing rs872071 and rs9378805 (261 158–460 393 bps) to define the boundaries of the 6p25.3 association signal such that the likelihood of the disease-linked variant residing outside the chosen region is very small. In total, 130 additional HapMap SNPs mapping to the interval were successfully imputed using data on the 77 SNPs directly genotyped by Illumina 370Hapmap Bead Arrays in our GWA study (Fig. 1A).
On the basis of single-point test statistics for association with CLL risk, combined with the distribution of LD blocks generated based on confidence intervals (CIs) as defined by Gabriel et al. (3), a candidate region of 18 kb (Chr6: 348 604–366 317 bps) from intron 7 (rs1877179) to 10 kb downstream of the IRF4 3'-UTR (r9378808) was identified as being most likely to harbour the disease-causing variant (Fig. 1B). While an irregular pattern of LD is seen between the 21 SNPs mapping to this region on the basis of \( r^2 \), they are all characterized by \( D' \) values >90%.

To comprehensively annotate all genetic variation in this 18 kb interval, we resequenced this genomic region in 30 healthy mother–father–child Centre d’Etude du Polymorphisme Humain (CEPH) trios. Only 460 bp (2.6%) of this 18 kb genomic region was refractory to sequencing, owing to low complexity. These bases were located 7605 bp downstream of IRF4 and not in the vicinity of any other genes. In total, we identified 80 variants (Supplementary Material, Table S1); these included 76 SNPs and four insertion/deletion polymorphisms. Of the 80 variants, 56 were common [i.e. minor allele frequency (MAF) \( \geq 0.05 \)]. Of these common variants, 34 had not been genotyped by HapMap and 4 were unlisted by dbSNP (Build 128; Supplementary Material, Table S1). In none of the parent–offspring trios were genotypes incompatible with Mendelian inheritance.

We calculated pair-wise LD statistics between each of the 56 common SNPs and rs872071. Eleven variants exhibited strong LD (defined by \( r^2 \geq 0.50 \)) with rs872071 and could thus be considered strong candidates for being the causal variant; 6 of these 11 variants map to the 3'-UTR region of IRF4 (Supplementary Material, Fig. S1). These 11 polymorphisms, together with rs872071 and rs9378808, were genotyped

---

**Figure 1.** The IRF4 locus. (A) SNP single marker-association imputation results. The imputed \(-\log_{10} P\)-values are shown based on the results from the GWA study (2). The quality plot gives an indication of the confidence with which untyped SNPs were imputed. Recombination rate (cM/Mb) across the region is derived from HapMap project data (release 21a). (B) The candidate region for resequencing with the exons and 3'-UTR of IRF4 indicated. Pair-wise LD \( (r^2) \) metrics of the 21 SNPs calculated in Haploview (v4.0) software are also shown. The shading indicates the LD relationship between each pair of SNPs; the darker the shading, the greater extent of LD.
in 1058 CLL cases and 2177 controls; 1028 of the cases (97.2%) and 2133 of the controls (98.0%) were genotyped successfully (according to set criteria). Individual call rates for the SNPs ranged from 98.3% to 99.5%. The distribution of genotypes in both cases and controls of all DNA polymorphisms showed no departure from Hardy–Weinberg equilibrium (HWE) (i.e. P > 0.05).

All 13 polymorphisms genotyped showed robust evidence of an association with CLL (Table 1). The strongest evidence for an association between variation in IRF4 and risk of CLL was provided by rs1050979 (P = 3.68 × 10^-11) and rs9391997 (P = 5.61 × 10^-14) (Table 1), both localizing within the 3'-UTR of the gene. Two SNPs, rs1050976 and rs872071, which also map to the 3'-UTR of IRF4, showed evidence for an association at P < 5.00 × 10^-15, at least two orders of magnitude smaller than association signals from the nine other variants that were genotyped. On the basis of Akaike weights, calculated to determine the weight of evidence in favour of each variant relative to rs1050979, nine of the variants were ~10 times less likely to be causal than rs1050979. Likelihood ratios with rs1050979 were also calculated for each of the other 12 SNPs. Comparison with the chi-square distribution shows that the same nine variants can be considered non-causal with 95% confidence (Table 1).

The SNPs rs1050979, rs9391997, rs1050976 and rs872071 were strongly correlated (LD metrics r^2 and D' of 0.99, 0.99 for all pairwise comparisons), and a single-risk haplotype is defined by G-G-T-G with frequency 0.51. Collectively these data are consistent with one of the four SNPs being the causal basis of the association signal. Moreover, as all four variants localize to the 3'-UTR, it suggests that the basis of the disease is through differential IRF4 expression. This inference is in keeping with the observation of reduced IRF4 expression being associated with the risk allele at rs872071 (2).

Our strategy for fine mapping was based on directly genotyping SNPs that were strongly correlated with rs872071, imposing a threshold of r^2 ≥ 0.5. Reliance on r^2 does, however, exclude the possibility that the true causal variant may have a very different MAF. To explore this possibility, we imputed all untyped SNPs in the 18 kb candidate region using the sequence data of the 30 CEPH trios. In this analysis, rs1050979, rs9391997, rs1050976 and rs872071 showed the strongest evidence for an association (Supplementary Material, Fig. S2). Association with CLL risk was found for only one untyped SNP rs2316515, which also mapped to the 3'-UTR region of IRF4 (r^2 with rs872071 = 0.46) and had a similar MAF (0.44) to the SNPs found by direct genotyping.

To determine which of the four variants was most likely to be causal, we interrogated the genomic region to which the variants map using an extensive series of bioinformatics tools. Sequence conservation in non-coding regions has been shown to be a good predictor of cis-regulatory sequences (4). Moreover, it has been proposed that variation with evolutionary conserved regions is likely to be associated with phenotypic differences that may contribute to the expression of traits (4). Cross-species sequence comparison of the 3'-UTR of IRF4 revealed the presence of four regions conserved between Homo sapiens and Mus musculus. One of these regions coincided with the candidate causal variant rs1050976 (Supplementary Material, Fig. S3). rs2316515 did not map to any evolutionary conserved sites.

To investigate the possibility that the influence of sequence variation in the 3'-UTR, defined by any of the four variants or rs2316515, on CLL risk might be mediated through miRNA, we searched for target sequences using miRDB software. While miRDB predicted that the 3594 bp UTR may contain up to 23 miRNA targets, none overlapped or mapped directly adjacent to the four variants (Supplementary Material, Fig. S3). Using UTRScan and Transterm, we identified two classes of possible regulatory elements within the 3'-UTR; erythroid 15-lipoxygenase (15-lox) dice and a K-box motif. Both of these regulatory motifs are frequently found in the 3'-UTRs of vertebrate genes and have been implicated as important factors in post-transcriptional regulation (5,6). Of the possible cis-acting regulatory DNA elements predicted to be present in the UTR using the program TFSearch, only one mapped to a candidate causal variant (Supplementary Material, Fig. S3). The presence of the rs872071 non-risk allele is predicted to generate a binding site for the
trans-acting regulatory element myeloid zinc finger 1 (MZF1), which is absent in the presence of the risk allele.

DISCUSSION

Generally the tagging SNPs utilized in GWA studies are not strong candidates for causality themselves and therefore determination of the actual causal variant at a specific locus poses a significant challenge. At present the causal basis of the 6p25.3 association with CLL risk is unknown.

Through fine-scale mapping, we have been able to provide evidence for refinement of the 6p25.3 association to an 18 kb DNA segment mapping to the 3'-UTR region of IRF4. Our analysis also provides evidence of the value of imputation of untyped SNPs as a methodology for interrogating association signals.

IRF4 represents a highly attractive candidate gene for CLL susceptibility, being a key regulator of lymphocyte development and proliferation, and differential expression has been linked to the development of CLL and multiple myeloma (7–10). IRF4 controls the termination of pre-B-cell receptor signalling and promotes the differentiation of pro-B-cells to small B-cells (11); hence its expression controls the transition of memory B-cells, thought to be the precursor cell type for CLL, to plasma cells (11–13).

By resequencing the genomic segment to which the CLL association signal maps and evaluating variants mapping to this region, we have defined a single haplotype annotated by four IRF4 3'-UTR SNPs most likely to be responsible for the 6p25.3 association. It is increasingly being recognized that multiple risk variants with different MAFs and relative risks may map to the same locus. Hence our analysis does not exclude the possibility that the IRF4 association may also be complex.

3'-UTRs can impact on gene expression in several ways by affecting the localization, stability and translation of mRNAs. In addition, 3'-UTR regions have been shown to contain sequence-specific endonucleolytic cleavage sites (14) and in some cases the stability of the mRNA can be controlled by a protective factor that binds within the 3'-UTR at or near the cleavage site competing with the endonuclease (15). Furthermore, destabilizing sequences in the 3'-UTRs of several mRNAs have been found to stimulate rapid deadenylation, which can promote subsequent mRNA degradation (15). It is intriguing that the 3'-UTR of IRF4 contains a binding site for c-Myc, as IRF4 and MYC form a positive autoregulatory loop during normal B-cell development and dysfunction of this pathway is thought to play a role in some B-cell malignancies (16). Identification of the causal variant and exploration of the functional basis of the IRF4 association is likely to be contingent on sequentially assaying the effects of each of the variants on mRNA degradation and transcription factor binding.

We have previously shown a relationship between rs872071 genotype and IRF4 mRNA expression level in Epstein-Barr virus (EBV)-transformed lymphocytes, with lower levels being associated with risk alleles in a dose-dependent fashion (2). This observation is consistent with a model in which the causal variant influences risk by arresting transition of memory B-cells through decreased IRF4 expression.

Working on the principle that one of the four SNPs we identified may alter IRF4 expression through 3' effects, we conducted a number of complementary bioinformatics analyses of the 3'-UTR of IRF4. These analyses provide evidence that the SNP rs872071 is predicted to generate a binding site for the trans-acting regulatory element MZF1. Although speculative, this observation is intriguing as MZF1 functions as a tumour/growth suppressor in the haematopoietic compartment, whereby MZF1 inactivation leads to autonomous cell proliferation and the ability of MZF1−/− haematopoietic progenitors to sustain long-term haematopoiesis (17). Intriguingly, MZF1−/− mice develop a lethal neoplastic disease involving multiple cell lineages including those of the lymphoid series within 2 years (17).

This study serves to illustrate the challenges of identifying causal variants responsible for GWA signals. Through detailed fine-scale mapping and resequencing we have, however, been able to refine the basis of the association. Our data indicates genetic variation impacting on differential expression of IRF4 as a causal basis, but further functional analyses are required to elucidate the precise functional basis.

MATERIALS AND METHODS

Resequencing SNP discovery panel

DNA extracted from 30 (46 males, 44 females) CEPH mother–father–child trios (US Utah residents with ancestry from northern and western Europe) was obtained from the Coriell cell depositories. These 90 samples correspond to those used in the Phase I HapMap study that has successfully genotyped over a million SNPs in different populations (18).

Genotyping cohort

The cases analysed have previously been genotyped in our GWA study of CLL (Phases 1 and 2 cases); full details have been provided in previously published material (2). Briefly, peripheral blood samples were collected from 1058 individuals with CLL (717 male, 341 female; mean age at diagnosis 61.2 years; SD ± 11.2). These comprised 155 with one or more first-degree relatives with CLL or a related B-cell lymphoproliferative disorder ascertained 1997–2007 through the International CLL linkage consortium, 541 CLL cases ascertained through the Royal Marsden NHS Hospitals Trust 1998–2006 and 362 cases ascertained through Leukaemia Research CLL-4 clinical trial 1999–2004 (19). All CLL cases were British residents and self-reported to be of European ancestry. The diagnosis of CLL was confirmed in accordance with WHO classification guidelines (20).

For controls, blood samples were obtained from 2177 healthy individuals (708 male, 1469 female; mean age at sampling 59.3 years, SD ± 10.5) recruited to the National Cancer Research Network genetic epidemiological studies: the National Study of Colorectal Cancer (1999–2006) (21); the genetic Lung Cancer Predisposition Study (1999–2004) (22) and the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry (1999–2004). These controls were the spouses or unrelated friends of patients with cancer. None had a personal history of malig-
nancy at time of ascertainment. All were British residents and self-reported to be of European ancestry. About 920 of these controls were used as part of Phase 2 of our previously published GWA study of CLL (2).

All biological samples were obtained from patients and controls with informed consent and ethical review board approval (MREC 99/1/82) in accordance with the tenets of the declaration of Helsinki.

**Bioinformatics**

We used Haplovie software (v3.2) for haplotype analysis and to infer the LD structure encompassing the IRF4 gene. LD metrics were calculated using Haplovie software (v4.0) (23).

Prediction of the untyped SNPs in the case–control data from the GWA study was carried out using MACH1.0 (24) on HapMap (January 2007 on NCBI B35 assembly, dbSNP build 125) Phase II data. Imputed data integrity was verified where possible by crosschecking the concordance of imputed genotypes with that of Illumina SNP genotype data.

The transcription factor motif search engine TFSEARCH (25) was used to predict transcription factor binding sites in the 3′-UTR DNA sequence. In accordance with generally accepted criteria, only scores >0.85 were considered to reflect accurate predictions of transcription factor binding sites (25). The program UTRscan (26) was used to search for UTR functional elements as defined in the UTRsite collection, and TransTerm (27) was used to search for defined regulatory elements or protein binding sites. DNA sequences were available for the rat, mouse and human IRF4 3′-UTR and were aligned using mVISTA (28) to identify regions of cross-conservation. Finally, we made use of miRDB software (29,30) for miRNA target prediction and functional annotations.

**Resequencing**

Sequence changes in the interval Chromosome 6 (348,604-366,317; UCSC May 2006 assembly, NCBI build 36.1) from within exon 7 of IRF4 to 6.3 kb downstream of the 3′-UTR of IRF4 were identified by direct sequencing. PCR and sequencing primers were designed by Primer3 software (PCR primer sequences available on request). Amplicons were sequenced by ABI chemistry (BigDye v3.1; Applied Biosystems, Foster City, CA, USA) implemented on ABI 3730xl DNA analyzers (Applied Biosystems). Sequence reads were analysed using Mutation Surveyor software v3.10 (Softgenetics, State College, PA, USA).

**Genotyping**

DNA was extracted from samples using conventional methodologies and quantified using PicoGreen (Invitrogen, Renfrew, UK). Genotyping was conducted using the Sequenom MassARRAY system (San Diego, CA, USA). Assay details available on request. Genotyping quality control was tested using duplicate DNA samples within studies and SNP assays. For all SNPs, >99.9% concordant results were obtained. A sample was removed from analysis if it failed genotyping for more than two SNPs.

**Statistical analysis**

Statistical analyses were undertaken in Stata v10 (Station College, TX, USA). Deviation of the genotype frequencies in the controls from those expected under HWE was assessed by χ² test (1 degree of freedom). Unconditional logistic regression was used to calculate the per allele OR of CLL and associated 95% CIs for each SNP. No adjustment for age or sex was performed. Haplotype analysis was conducted in Haplovie software (v4.0). The haplotypes are estimated using an accelerated EM algorithm similar to the partition–ligation method described in Qin et al. (31) and tested for association via a likelihood ratio test.

The weight of evidence in favour of each SNP being causal was quantified by calculating Akaike weights for each SNP model $j$: $\exp(-\Delta_i/2) \sum \exp(-\Delta_i/2)$, where $\Delta_i$ is twice the difference in log-likelihood between model $i$ and the best-fitting model (32).

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

We are grateful to all individuals who participated in this study.

**Conflict of Interest statement.** The authors declare no competing financial interests.

**FUNDING**

D.C.-S. is supported by a PhD studentship from the Institute of Cancer Research. This work was supported by the Leukaemia Research Fund (LRFO05001, LRFO06002) and Cancer Research UK (C1298/A8362).

**REFERENCES**


