Multiple Hodgkin lymphoma-associated loci within the HLA region at chromosome 6p21.3

Loukas Moutsianas1*, Victor Enciso-Mora2*, Yussanne P Ma1, Stephen Leslie1, Alexander Dilthey1, Peter Broderick2, Amy Sherborne2, Rosie Cooke3, Alan Ashworth4, Anthony J Swerdlow3, Gilean McVean1,5, Richard S Houlston2+

1Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
2Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK
3Section of Epidemiology, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK
4The Breakthrough Breast Cancer Research Centre, London SW3 6JB, UK
5Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

*Equal authorship status at this position
+Corresponding author:
e-mail: Richard.houlston@icr.ac.uk
Tel: +44 (0) 208 722 4175
Fax: +44 (0) 208 722 4365

Keywords: MHC risk, Hodgkin lymphoma, HLA
ABSTRACT

Since an association between the HLA region and Hodgkin lymphoma (HL) was first reported in 1967, many studies have reported associations between HL risk and both SNP and classical HLA allele variation in the Major Histocompatibility Complex (MHC). However, population stratification and the extent and complexity of linkage disequilibrium within the MHC have hindered efforts to fine-map causal signals. Using SNP data to impute alleles at classical HLA loci, we have conducted an integrated analysis of HL risk within the HLA region in 582 early-onset HL cases and 4,736 controls. We confirm that the strongest signal of association comes from a SNP located in the class II region, rs6903608 (OR = 1.79, $P = 6.63 \times 10^{-19}$), which is unlikely to be driven by association to HLA-DRB, DQA, or DQB alleles. In addition, we identify independent signals at rs2281389 (OR=1.73, $P=6.31 \times 10^{-13}$), a SNP which maps closely to HLA-DPB1, and the class II HLA allele DQA1*02:01 (OR=0.56, $P=1.51 \times 10^{-7}$). These data suggest that multiple independent loci within the HLA class II region contribute to the risk of developing early-onset HL.
INTRODUCTION

Hodgkin Lymphoma (HL) is a common lymph node cancer of germinal center B-cell origin, which is characterized by malignant Hodgkin and Reed-Sternberg (HRS) cells mixed with a dominant background population of reactive lymphocytes and other inflammatory cells. While Epstein-Barr virus (EBV) infection may be causally related to a number of cases there is little evidence to support the involvement of other environmental risk factors.

Evidence for inherited genetic influence on susceptibility is provided by the increased familial risk and high concordance between monozygotic twins. Since an association between the HLA region and HL risk was first reported in 1967, studies have subsequently identified associations between both HLA class I and class II alleles and common HL (cHL) risk. Studies to date have evaluated only specific HLA alleles and have not taken into account the existence of complex linkage disequilibrium patterns between the multiple risk loci mapping to the MHC region, associations between HLA alleles at different resolution, and the need to control rigorously for population stratification. In view of the limitations of these previously published studies, we have conducted a more comprehensive analysis.

It has recently been shown that single nucleotide polymorphism (SNP) data within the 6p21 region can be used to impute alleles at key classical class I (HLA-A, HLA-B, and HLA-C) and class II (HLA-DRB1, HLA-DQA1, and HLA-DQB1) loci with accuracy that exceeds 90% at the 4-digit level. By using existing genotype data from a previous genome-wide association study of HL, which identified many SNPs within the HLA region that are strongly associated with disease, we can therefore assess the relative contribution of classical HLA alleles and SNP variation to HL risk. We identify multiple independent signals across the HLA, the strongest of which is most likely not driven by any of the HLA loci whose types were statistically inferred in this study.
METHODS

Samples and genotyping

We analyzed the constitutional DNA of 615 patients diagnosed with HL (International Classification of Diseases [ICD] 10 codes C81.0-3) who were ascertained through the Royal Marsden Hospitals NHS Trust Family History study, between 2004 and 2008 (mean age at diagnosis 38 years, SD 16) and an ongoing national study of cHL in females (mean age at diagnosis 23 years, SD 6) being conducted by the Institute of Cancer Research (ICR) to assess late effects of HL treatment. Both routes of patient ascertainment thus favored the acquisition of early-onset HL cases. All cases were British residents and were self-reported to be of European Ancestry.

DNA was extracted from samples using conventional methodologies and quantified using PicoGreen (Invitrogen, Carlsbad, USA). Genotyping of cHL cases was conducted using Illumina Infinium HD Human610-Quad BeadChips according to the manufacturer's protocols (Illumina, San Diego, USA). DNA samples with GenCall scores <0.25 at any locus were considered “no calls”. A SNP was deemed to have failed if fewer than 95% of DNA samples generated a genotype at the locus. Cluster plots were manually inspected for all SNPs identified as potentially associated. Publicly available data from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) study were used as controls. Specifically, this included 2420 individuals from the 1958 Birth Cohort (58BC, also known as the National Child development study) and 2737 from the National Blood Service Collection (NBS), all of which had been genotyped on Illumina 1.2M arrays at the Wellcome Trust Sanger Institute. To ensure quality of genotyping in all assays, at least two negative controls and 1-2% duplicates (showing a concordance >99.99%) were genotyped.

Collection of blood samples and clinico-pathological information from subjects was undertaken with informed consent and relevant ethical review board approval from the University of Oxford and the Institute of Cancer Research in accordance with the tenets of the Declaration of Helsinki.
Quality control

For any samples showing a high degree of relatedness we removed the sample with the lower call rate from the analysis, following previously described protocols\textsuperscript{13}. This obviates the need to adjust for familial correlations and ensuring that unbiased estimates of HL risk associated with genotypes are derived. We excluded SNPs on the basis of deviation from Hardy-Weinberg Equilibrium (HWE) using a threshold of $P<1.0\times10^{-5}$ in controls. We also removed SNPs with minor allele frequency (MAF) $<0.05$ and call rate $<0.95$\textsuperscript{13}. To identify and exclude individuals with non-Western European ancestry, case and control data was merged with data on individuals of different ethnicities from the International HapMap Project, and dissimilarity measures were used to perform principal component analysis (Supplementary Figure 1). For the control samples we also followed the guidelines from the WTCCC and removed some additional controls\textsuperscript{13}. After imposing these stringent quality control measures, SNP genotypes were available on 582 cases and 4,736 controls (2,235 58C and 2,501 NBS).

Statistical and bioinformatics analysis

For single SNP analyses we considered the MHC to be defined by a 4.5Mb region bordered by the RFP and MLN genes (rs209130 at 28,975,779bps and rs1547668 at 33,883,424bps respectively) at the telomeric and centromeric ends of 6p21 respectively. For the HLA imputation we made use of GWA study SNP data for an extended region in 6p21 bounded by rs1165196 and rs2772372 (25,921,129 and 33,535,328bps respectively). All positions are with respect to build 36 of the Human Genome.

Analyses were primarily undertaken using R v2 software. Initially, we analysed the association between each SNP and the risk of HL using the Cochran-Armitage trend test. Odds ratios and associated 95\% CIs were calculated by unconditional logistic regression. We then conducted stepwise logistic regression analyses to interrogate the region for independent secondary signals.

From a reference database of SNP-haplotypes carrying known HLA alleles, we imputed HLA genotypes for the cHL cases, the NBS controls, and the subset of the 1958 Birth Cohort (58BC) samples, for which classical HLA data were not available,
as previously described\textsuperscript{11,12}. The reference database combines classical HLA data from the HapMap Project and the 58BC. Details of QC have been previously described\textsuperscript{12}. Imputation was performed for three class I and three class II loci: HLA-A (n=2,474 reference samples), HLA-B (n=3,090), HLA-C (n=2,022), HLA-DQA1 (n=175), HLA-DQB1 (n=2,629), HLA-DRB1 (n=2,665). Imputation accuracy was assessed through a cross-validation analysis of the training data. Training was performed on two thirds of the reference panel and tested on the remaining third at the SNPs chosen for imputation in the current study. Thresholding calls at a posterior probability of 0.7 provided call rates of between 0.90 (HLA-DRB1) and 0.99 (HLA-C) and accuracy of over 0.95 for all loci at the two-digit level and between 0.91 (HLA-DQA1) and 0.98 (HLA-DQB1) at the four digit level.

The familial relative risk of HL attributable to any locus is given by the formula\textsuperscript{15}:

$$\lambda_i = \frac{p(p r_2 + q r_1)^2 + q(p r_1 + q)^2}{(p^2 r_2 + 2pq r_1 + q^2)^2}$$

where $p$ is the population frequency of the minor allele, $q=1-p$, and $r_1$ and $r_2$ are the relative risks (approximated by the odds ratios) for heterozygotes and the rarer homozygotes, relative to the more common homozygotes. From $\lambda^*$ it is possible to quantify the impact the locus makes to the overall familial risk of HL seen in first-degree relatives. Assuming a multiplicative interaction between risk alleles the proportion of the overall familial risk attributable to the locus is given by $\log(\lambda^*)/\log(\lambda_0)$, where $\lambda_0$, the overall familial risk is assumed on the basis of epidemiological studies to be 3.93\textsuperscript{16}.

To examine if any directly typed or imputed SNPs annotate potential regulatory sequences within the 13.8kb region of LD encompassing the SNP rs6903608 (which had the highest signal of association; see below) we performed \textit{in silico} searches using Transfac Matrix Database v7.2\textsuperscript{17}, PReMod\textsuperscript{18} and EEL\textsuperscript{19} software.
RESULTS

We examined samples from 615 cases and 5,157 controls, and subjected these to rigorous quality control as described above. Furthermore, we evaluated the datasets for ancestral differences by principal component analysis. After QC, data were available on 582 cases and 4,736 controls. Supplementary Figure 1 shows that the cases and controls were ancestrally comparable.

To assess the relationship between specific HLA alleles and HL, we imputed key class I and II HLA alleles at two- and four-digit resolutions, using the algorithms described in (11, 12). Comparison between observed and imputed HLA alleles for a subset of controls, for which we had experimentally determined HLA types, showed a high degree of concordance (over 90% at the 4-digit level; see Methods). For each individual and locus, the imputation methodology returns probabilities associated with the sample having each allelic type; these probabilities have been shown to be reasonably well-calibrated (11). Thus, we were able to incorporate these probabilities into the general logistic regression framework, to account for any uncertainty associated with imputed types.

We initially considered 1,700 SNPs mapping to the 4.5Mb region that encompasses the classical MHC region at 6p21.3. Figure 1 shows the strength of the single SNP associations across this region. 25 SNPs showed evidence for an association with HL risk at $P<10^{-7}$; most mapping to HLA class II regions. The strongest single-SNP based association was attained for rs6903608 mapping at 32,536,263 bps, about 16 kb centromeric to HLA-DRA ($P=6.63\times10^{-19}$; Figure 1 (a), Supplementary Table 1 (a)).

Five HLA class II alleles showed evidence of association with HL risk at the $1\times10^{-7}$ threshold in the unconditional logistic regression analysis, namely DQA1*02:01, DRB1*07:01, DQB1*03:03, DRB1*15:01 and DQB1*06:02 (Supplementary Table 1(b)). The associations at DQA1*02:01, DRB1*07:01, DQB1*03:03 confer an increased risk of HL while the associations at DRB1*15:01 and DQB1*06:02 confer a reduced risk. Some of these alleles, such as DRB1*15:01 and DQB1*06:02, are known to be in strong LD in European populations. The strongest association was provided by DQA1*02:01 (OR=0.45, $P=6.17\times10^{-15}$). Of the class I HLA alleles, the
strongest association observed was for B*07:02, though it did not reach genome-wide significance (Supplementary Table 1; OR=1.39, \(P=4.40 \times 10^{-5}\)).

To evaluate the independence of associations, we conducted stepwise logistic regression, jointly on SNPs and imputed HLA alleles, initially including rs6903608 as a covariate in our model. Conditional analysis showed that most, but not all, of the class II variation defined by the SNP-genotype could be explained for by rs6903608 (Figure 1 (b), Supplementary Table 2). The strongest secondary signal appeared to be for SNPs rs9277565 (OR=1.70, \(P=4.57 \times 10^{-13}\)) and rs2281389 (OR=1.73, \(P=6.31 \times 10^{-13}\)), which map closely to HLA-DPB1 and one of which, rs9277565, has previously been shown to be in LD \((r^2=0.37)\) with HLA-DPB1*03:01.20 Note that classical alleles at HLA-DPB1 were not imputed here due to lack of training data. Inclusion of rs2281389 in the logistic model accounted for most of the remaining signal (Figure 1 (c), Supplementary Table 3), except for the protective effect of HLA-DQA1*02:01 (OR=0.56, \(P=1.51 \times 10^{-7}\)). DQA1*02:01 is nearly perfectly correlated with DRB1*07:01, and it is unclear which of the two alleles drives the observed signal; here for simplicity we only refer to DQA1*02:01. When rs6903608, rs2281389 and DQA1*02:01 were all included as covariates to the model, no additional loci showed genome-wide evidence for association (Figure 1 (d), Supplementary Table 4).

While genetic variation defined by rs6903608 provides the strongest association signal, collectively these data provide evidence for two additional independent disease loci, defined by rs2281389 and HLA-DQA1*02:01 (Figure 1, Supplementary Table 2, Supplementary Table 3). rs2281389 lies in close proximity to HLA-DPB1, variation in which has been associated with HL risk in a number of previously published studies5,21, with DPB1*03:01 appearing to confer susceptibility (OR=1.42) and DPB1*02:01 resistance (OR=0.49) to HL21.
DISCUSSION

To examine the relationship between HLA and the risk of developing HL we have conducted a large case-control study and have systematically examined the relationship between genetic variants and HL risk for the entire 4.5Mb 6p21.3 region that encodes the MHC.

Using GWA SNP data in PCA has allowed us to minimise the potential problem of population stratification as formal statistical analysis provided no significant evidence that population substructure is a confounding factor in our study. The large dataset and dense set of informative markers in our study has also made possible the detection of independent effects at class II loci. Imputation of HLA alleles from SNP data has allowed us to also analyse systematically the relationship between HLA types and HL risk, at both the 2- and 4- digit levels of significance.

Our study provides unequivocal evidence for an association between the MHC and risk of developing HL. Moreover, and specifically, it provides evidence of a relationship between HLA class II genotypes and HL risk. While we acknowledge that we did not impute all of the HLA alleles, the analysis we have conducted on a dense set of single SNPs in the region will have been sufficient to recover strong associations at untyped loci. Moreover, our imputations cover the alleles carried by 97.4% of the European population, according to dbMHC (averaged across all loci for which imputations are made). Our analysis found no evidence that class I variation also contributes to disease development, as the signals from the class I region were of moderate significance ($P > 10^{-5}$), and disappeared in the conditional analysis.

While our study represents the most comprehensive interrogation of the relationship between genetic variation in MHC and HL risk to date, the analysis has limitations. A hallmark of HL epidemiology is a bimodal age specific incidence and it has been argued that the disease in young and older adults is etiologically different\textsuperscript{22}; in particular the prevalence of EBV in tumours is higher in older HL patients ($\sim50\%$ aged $\geq 40$ years compared with $\sim30\%$ aged $< 40$ years)\textsuperscript{23}. HLA-class I polymorphisms have been reported to be associated with development of infectious mononucleosis upon primary EBV infection\textsuperscript{24} and HLA-A*02 and HLA*01 alleles have been linked to the risk of developing EBV-positive HL\textsuperscript{10}. Such observations are in
keeping with predominant class-I restricted T-cell response to EBV infection. The majority of HL cases we have studied were young (93% aged <35 years at diagnosis) hence our findings are more likely to apply to EBV-negative HL. This suggestion is reinforced by the lack of association between HLA-A*02 or HLA-A*01 alleles and HL in the current study \((P>0.05)\). Our results therefore suggest, albeit individually, substantially different genetic contributions to the risk of EBV-positive and EBV-negative HL within the MHC.

The neoplastic multinuclear HRS cells that typify classical HL are derived from germinal center B-cells that express HLA class II antigens\(^1\). Several observations suggest that antigen presentation is involved in the pathobiology of HL. HRS cells stimulate T-cell proliferation in mixed lymphocyte reactions and induce cytotoxic activity\(^{25,26}\). Lack of membranous expression of HLA class II by HRS cells of HL is common at primary diagnosis and is associated with extranodal disease, lack of EBV in HRS cells and absence of HLA class I expression \(^{27-30}\). Expression of HLA class II by HRS cells is retained more frequently in EBV-associated cHL, and in HLA class I-positive HRS cells \(^{27-30}\). Hence, a possible explanation for the association of HLA class II-expression status might be that HLA class II directly activates T-helper 1 (Th1) cells, thereby inducing and maintaining cytotoxic antitumor immune responses. Lack of HLA class II-expression results in diminished activation of Th2 and/or Treg cells. Therefore, the class II associations we identified may reflect differential binding affinity of the HLA class II molecule with immunogenetic antigenic peptides modifying class II downregulation. This postulate is in keeping with class II variation being primarily associated with development of EBV-negative HL, something fully supported by our findings of strong class II associations.

The strongest association identified in our study was attained with rs6903608. Inherited susceptibility to HL is likely to be mediated through the co-inheritance of multiple risk alleles. Under a multiplicative model of interaction between risk loci on the basis of the risks associated with hetero- and homozygosity for the C-risk allele of rs6903608 (See Supplementary Table 1), this variant accounts for ~6% of the increased risk of HL seen in first-degree relatives of patients. rs6903608 maps to a LD block of 13.8kb from 32,503,014-32,516,820bps on chromosome 6. \(\text{In silico}\) analysis of this region using EEL\(^{19}\), PreMod\(^{38}\) and Transfac\(^{17}\) software provided no evidence that the SNP is directly functional in terms of impacting on a regulatory sequence. As rs6903608 localises within 16 kb of HLA-DRA, it is possible that the
SNP tags this locus. Unfortunately, due to the lack of training data, we were not able to impute alleles at this locus and thus examine this possibility. Nonetheless, our analysis of the joint SNP and imputed HLA dataset indicates that none of the classical class I HLA alleles, or class II genes HLA-DRB, HLA-DQA, HLA-DQB, is likely to be driving this association.

In conclusion, our analysis provides unequivocal evidence that MHC variation is a major determinant of HL risk, and specifically implicates class II variation at multiple loci in disease aetiology. Given our analysis is based on a high proportion of early-onset HL cases further studies will allow the generalisability of our findings to all forms of HL to be addressed.

ACKNOWLEDGEMENTS
Leukaemia and Lymphoma Research (UK) provided principal funding for the study. Support from Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund) is also acknowledged. Sample and data acquisition was supported by Breakthrough Breast Cancer and the European Union and we acknowledge NHS funding to the NIHR Biomedical Research Centre. We are grateful to the patients and their clinicians who participated in this collection (see Supplementary Note). This study made use of control genotyping data generated by the Wellcome Trust Case-Control Consortium. We acknowledge use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

AUTHORSHIP
Contribution: RSH designed the study; RSH, GM, VE-M and LM drafted the manuscript; LM, VE-M, YPM, SL, AD and AS performed statistical and bioinformatic analyses; PB oversaw sample coordination and laboratory analyses; AJS, AA and RC provided samples and data from a study conducted at the Institute of Cancer Research; RSH obtained funding for parent project.
Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Richard S. Houlston, Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, SM2 5NG; e-mail: richard.houlston@icr.ac.uk
REFERENCES

FIGURE 1: Associations in the MHC region, and conditional logistic regression analysis. Plots of the evidence for association between the HLA and cHL, as covariates are added to the model in a stepwise manner. In all figures, the horizontal axis indicates the location of each SNP along chromosome 6 and the vertical axis the −log_{10} P values of each SNP (blue) and HLA allele (red), obtained using logistic regression models with different numbers of covariates. Top left figure shows the evidence for association obtained from the 1-parameter logistic regression analysis. Top right figure shows the remaining signal when rs6903608 is included as a covariate, bottom left the signal when rs2281389 is also included in the model and bottom right the remaining signal when DQA1*02:01 is also added as a covariate. Multiple red dots on the same locus indicate the −log_{10} P values for different allelic subtypes at the respective locus. The positions of common HLA loci have been marked with vertical dashed lines and annotated accordingly; note that some loci (DRA1,DPB1) for which imputation results were not available have also been marked in the Figure. For all loci, the vertical lines indicate the mean position of the longest transcript. The yellow horizontal line indicates the commonly employed 5x10^{-8} threshold for genome-wide significance.
Multiple Hodgkin lymphoma-associated loci within the HLA region at chromosome 6p21.3