European Journal of Human Genetics (2015) 23, 824–830

www.nature.com/ejhg

ARTICLE

Heritability estimates on Hodgkin's lymphoma: a genomic- *versus* population-based approach

Hauke Thomsen*[,1], Miguel Inacio da Silva Filho[1], Asta Försti[1,2], Michael Fuchs[3], Sabine Ponader[3], Elke Pogge von Strandmann[3], Lewin Eisele[4], Stefan Herms[5,6], Per Hofmann[5,6], Jan Sundquist[7], Andreas Engert[3] and Kari Hemminki[1,2]

Genome-wide association studies (GWASs) have identified several single-nucleotide polymorphisms (SNPs) influencing the risk of Hodgkin's lymphoma (HL) and demonstrated the association of common genetic variation for this type of cancer. Such evidence for inherited genetic risk is also provided by the family history and the very high concordance between monozygotic twins. However, little is known about the genetic and environmental contributions. A common measure for describing the phenotypic variation due to genetics is the heritability. Using GWAS data on 906 HL cases by considering all typed SNPs simultaneously, we have calculated that the common variance explained by SNPs accounts for >35% of the total variation on the liability scale in HL (95% confidence interval 6–62%). These findings are consistent with similar heritability estimates of ~0.40 (95% confidence interval 0.17–0.58) based on Swedish population data. Our estimates support the underlying polygenic basis for susceptibility to HL, and show that heritability based on the population data is somehow larger than heritability based on the genomic data because of the possibility of some missing heritability in the GWAS data. Besides that there is still major evidence for multiple loci causing HL on chromosomes other than chromosome 6 that need to be detected. Because of limited findings in prior GWASs, it seems worth checking for more loci causing susceptibility to HL.
European Journal of Human Genetics (2015) **23,** 824–830; doi:10.1038/ejhg.2014.184; published online 17 September 2014

INTRODUCTION

Hodgkin's lymphoma (HL) is a malignancy of the lymphatic system with an incidence of 2–3/100 000/year in developed countries.[1] It is characterized through malignant Hodgkin and Reed–Sternberg (HRS) cells mixed with a dominant background population of reactive lymphocytes and other inflammatory cells.[2] Epstein–Barr virus (EBV) infections may be causally related to a number of cases as well as a personal history of autoimmune diseases.[3,4] However, there is limited evidence to the involvement of other specific environmental risk factors, although there is a distinctive pattern of incidence rates and risk profiles by age, race/ethnicity, sex and economic levels.[2] Some evidence for inherited genetic influence on susceptibility is provided by the increased familial risk and the very high concordance between monozygotic twins.[5] Recently, several genome-wide association studies (GWASs) have identified a couple of loci for HL.[6,7] These studies have shown that the risk of HL is highly influenced by the human leukocyte antigen (HLA) genotype variation, but the familial risk is also a consequence of non-HLA genotype variation.[6] However, all the genetic variants identified so far only capture a minor percentage of the estimated heritability of the disease.[8] Yet, a great deal remains to be understood regarding the remaining heritability.[9,10] Some plausible explanation include unidentified gene–gene interactions, unidentified contributions of rare genetic variants or overestimating the total heritability for HL in population-based studies, resulting in a 'phantom heritability',[9,11,12] that cannot be dissolved on the molecular basis.[13,14]

Thus, our aim is to provide reliable estimates for the genetic variation of the disease derived either from the quantification of resemblance between close relatives or the dissection of genetic variation from genomic loci.[15]

The knowledge of the heritability estimates for the susceptibility to HL based on genomic data and on population data will then provide further insights in the proportion of genetic variation hidden so far but still detectable on the genomic level.[16]

MATERIALS AND METHODS

Genomic data: quality control of SNP genotyping

The study population comprised a total of 2227 individuals, with 1001 cases and 1226 controls. Cases were sampled all over Germany, whereas controls were sampled within the Ruhr area in North Rhine-Westphalia as part of the Heinz Nixdorf Recall Study.[17] All individuals were genotyped using the Illumina Human Omni Express 12v1 chip (Illumina, San Diego, CA, USA). (733 202 markers).

To counteract artificial differences in allele frequencies between cases and controls causing spurious genetic variation, GWAS data have undergone a very stringent quality control procedure.[16] After checking the gender based on genotypes, 11 individuals were excluded because of inconsistencies. Three individuals were excluded because of low genotype calling rates (<0.99). In total, 27 individuals were excluded, because their heterozygosity was >3 SD apart from the mean heterozygosity of the sample. Principal component

[1]German Cancer Research Center (DKFZ), Division of Molecular Genetic Epidemiology, Heidelberg, Germany; [2]Center for Primary Health Care Research, Lund University, Malmö, Sweden; [3]Department of Internal Medicine I, University Hospital of Cologne, Cologne, Germany; [4]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University Duisburg-Essen, Essen, Germany; [5]Institute of Human Genetics and Department of Genomics, University of Bonn, Bonn, Germany; [6]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland; [7]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA, USA
*Correspondence: Dr H Thomsen, German Cancer Research Center (DKFZ), Division of Molecular Genetic Epidemiology, C050, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. Tel: +49 6221 421809; Fax: +49 6221 421810; E-mail: h.thomsen@dkfz-heidelberg.de
Received 31 December 2013; revised 6 August 2014; accepted 10 August 2014; published online 17 September 2014

analysis indicated a presence of population outliers, especially in the cases, because cases represent a more diverse group than controls. After excluding these 46 outliers, the remaining individuals were genetically well matched. Seventeen individuals having a relatedness score of >0.05 were excluded. The final set consisted of 906 cases and 1217 controls. All data were checked for differential missingness between cases and controls, and single-nucleotide polymorphisms (SNPs) were excluded with $P<0.05$. SNPs were also excluded after applying the Hardy–Weinberg equilibrium test with $P<0.001$. A detailed overview of the study material, the identification of samples of non-European origin, the plots of the principal components and the results is given in our previous paper.[6] The genome-wide Armitage trend test $\chi 2$ values showed minimal inflation of the test statistics proving the absence of substantial cryptic population substructure (genomic control inflation factor $\lambda_{gc}=1.09$).[6]

By using PLINK software[18] we finally produced two subsets of data with SNPs in cases and controls that had either a minor allele frequency (MAF) of >0.01 or MAF of >0.05.

Statistical analysis on genomic data

For the statistical analyses on the genomic data, the approach of Yang *et al*[19] was used. Their method has been completed by an approach of Speed *et al*,[20] who presented an improved method for the heritability estimation on GWAS data with a new adjustment for linkage disequilibrium (LD). Both methods provide an estimate of the additive genetic variance explained by SNPs but are accounting for LD between the genotyped SNPs and unknown causal variants in different ways (ie, correlations between SNP genotypes).[19,20] Both approaches fit a linear mixed model of the form: $y=\mu+g+e$,[16,19] whereby y is the vector of the disease status, μ is the mean vector, g is a vector of random additive genetic effects obtained from SNP data and e is a vector of residual effects. The covariance structure fitted in the data is the individual relationship estimated from the SNPs; $cov(y_j, y_k) = A_{jk}\sigma_g^2 + \sigma_e^2$, where A_{jk} is the genetic relationship between individuals j and k derived from the SNPs, σ_g^2 is the additive genetic variance and σ_e^2 is the residual variance. With this model disease heritability, h_0^2, can be defined as: $\sigma_g^2/(\sigma_g^2+\sigma_e^2)$.[20] Because phenotypes in case–control studies are measured on the 0–1 scale, the relationship between observations on the observed scale and liabilities on the unobserved continuous scale are modeled through the liability threshold model. The liability for HL on the underlying scale follows a standard normal distribution whereby if liability exceeds a certain threshold, t, then individuals will be affected. The estimate of variance explained by the SNPs on the observed 0–1 scale is linearly transformed to that on the unobserved continuous liability scale such that $h_l^2 = h_0^2 K(1-K)/z^2$,[21] where K is the prevalence of the disease and z is the value of the standard normal probability density function at the threshold t. Given an incidence of 2–3/100 000/year will result in a cumulative risk of ~ 1 in 1000, which can be considered as an estimate of the prevalence. The relationship between additive genetic variance on the observed 0–1 and unobserved liability scales is extended to account for ascertainment bias in a case–control study.[16] Estimation of the additive genetic variance was performed using restricted maximum likelihood (REML) via the genome-wide complex trait analysis (GCTA) software.[22] Because the MAF spectrum of the unobserved causal variants may be different from the genotyped SNPs, the estimation of the variance explained by SNPs was performed in two ways. (1) the estimate of the variance explained by SNPs was adjusted to account for missing LD between the genotyped SNPs and unknown causal variants.[19] SNPs were randomly assigned into two different groups with one of the groups being treated as representing 'true' causal variants. The covariance between both groups is supposed to reflect the true variance of relatedness between individuals, whereas the variance derived from the SNP group equals the variation of relatedness plus estimation error. The prediction error can then be derived by regressing the relationships of the 'true' causal variants on the SNPs. (2) In contrast to the method of Yang *et al*,[19] which suggests a uniformly scaling of the usual SNP-based kinship coefficients, Speed *et al*[20] proposed a different adjustment, in which SNPs are weighted according to how well they are tagged by their neighbors. The kinship coefficients corrected for LD are obtained in a two-step procedure: first, weightings for each predictor given the local patterns of correlations are calculated, and second these weightings are used to estimate relatedness values across all pairs of individuals.[20] The final estimation of the additive genetic variance was again performed by using REML via the software tool GCTA.[22]

In addition, the approach of Guan and Stephens[23] has been applied to estimate the proportion of phenotypic variance explained (PVE) by the genotypes. It implements the Markov chain Monte Carlo (MCMC)-based inference methods for Bayesian variable-selection regression based on standard normal linear regression with the following model:

$$y|\mu,\beta,X,\tau \sim Nn(\mu + X\beta, \tau^{-1}I_n),$$

relating a response variable y to covariates X. Here y is an n-vector of observations on n individuals, μ is an n-vector with components all equal to the same scalar μ, X is an n by p matrix of covariates, β is a p-vector of regression coefficients, τ denotes the inverse variance of the residual errors, $Nn(\cdot,\cdot)$ denotes the n-dimensional multivariate normal distribution and I_n the n by n identity matrix. The variables y and X are observed, whereas μ, β and τ are parameters to be inferred. Because GWAS applications involve binary phenotypes, the probit link function is used and allows direct comparisons to the estimates of variance explained by SNPs on the liability scale. The total proportion of variance in y explained by the relevant covariates X_γ, or PVE, is commonly used to summarize the results of a linear regression.

The PVE is closely related to the 'heritability' of the phenotype and reflects the optimal predictive accuracy that could be achieved for a linear combination of the measured genetic variants, whereas heritability reflects the accuracy that could be achieved by all genetic variants.[23]

Population data: Swedish Family-Cancer Database

To compare the variance explained by SNPs to new estimates of heritability from family-based studies, variance components were estimated for the susceptibility to HL on the basis of the 2010 update of the Swedish Family-Cancer Database that includes all individuals born after 1931 who are residing in Sweden, together with their biological parents, totaling ~ 12.1 million individuals.[24] The Database was created in 1996 by combining the Swedish Cancer Registry and the Swedish Multigenerational Register, and has been updated regularly. The Database includes information about the cancers, socioeconomic data and death causes. In total, 7438 individuals (4441 males and 2997 females) have been diagnosed with the HL (ICD-7 code 201). The R-stat package and DmuTrace[25,26] were used to extract all related individuals of the patients from the large pedigree file back to the base population as well as all offspring of the patients and their relatives in future generations until the current population. This resulted in a pedigree of 133 544 individuals (67 059 males and 66 485 females). The entire pedigree consisted of 6755 families across 6 generations with a family size ranging from 2 to 490 individuals. The total number of founders was 59 679 with a range of 2 to 203 individuals per family. Each family contained at least one and up to eight cases.

Statistical analysis on population data

A generalized linear mixed effect threshold model with a binary response variable using MCMC Carlo techniques was applied.[27] In such a standard threshold (probit) model, the observed binary records (Y_{ij}) are assumed fully determined by an underlying liability (λ_{it}), such that: for $Y_{ij}=0$ for $\lambda_{ij}\leq 0$ and $Y_{ij}=1$ for $\lambda_{ij}>0$ and the threshold value is set to 0. The model can be written as

$$\lambda = X\beta + Za + e$$

where λ = vector of all λ_{ij}, β = vector of 'fixed' effects, a = vector of random additive genetic effects of all individuals, e = vector of random residuals and X and Z are the appropriate incidence matrices. $Var(a) = A\sigma_a^2$ and $Var(e) = I_n\sigma_e^2$, where A is the additive genetic relationship matrix of all individuals, I_n is an identity matrix with dimension equal to number of records and σ_a^2 and σ_e^2 are the additive genetic and the residual variances, respectively. As usual for probit threshold models, σ_e^2 is restricted to be 1. Fixed effects included in the model were gender, birth year, country of birth, social economic index and number of children.

Marginal estimates of the genetic parameters were obtained in the underlying scale using Bayesian inference, implemented via the Gibbs sampling procedure and a data augmentation approach. The model included a Gibbs sampling chain of 10 150 000 rounds with a conservative 150 000 iterations as burn-in

and 10 million sampling rounds. Every 1000th sample was drawn, resulting in 10 000 samples. For each sample of the Gibbs chain, narrow-sense heritability was calculated after as: $h^2=\sigma_a^2/(\sigma_a^2+\sigma_e^2)$, where $\sigma_e^2=1$ for probit link model.[28,29]

Posterior marginal means of heritability were derived and the 95% highest marginal posterior density region of the heritability range determined. The algorithm to estimate genetic (co)variance components has been implemented in the Gibbs sampling module of the DMU statistical software package[30] that has been developed to handle multivariate genetic analyses including binary, ordered categorical and Gaussian traits. Results have been proven by the R package MCMCglmm[31] that has also been used to prove convergence diagnosis and output analysis. Heritability estimates on the liability scale have been transformed to the observed scale by using the linear transformation of Dempster and Lerner.[21]

RESULTS

Estimates of the variance explained by SNPs

The analysis on genomic data was restricted to the autosomes of the GWAS data set and based on the final set of 906 cases and 1217 controls. For both primary subsets of genomic data with either MAF >0.05 or MAF >0.01, the threshold for SNP missingness was raised stepwise starting from 0.05 up to 0.001. This resulted in a reduced number of SNPs from 583 333 to 442 325 (MAF >0.01). Both the crude proportions of variance and the adjusted estimates only dropped slightly, whereas the transformed estimate was stable across all subsets with different numbers of SNPs (Table 1). We only included adjusted estimates according to Yang *et al*[19] in our tables as they did not show any difference to the adjustment according to Speed *et al*.[20] Standard errors were slightly larger for the adjusted estimates and the transformed estimates. In contrast to stable estimates across different numbers of SNPs, the PVE showed moderate differences when using the Bayesian variable selection approach (Table 2). For data sets with less SNPs, the PVE values decrease and the high posterior density interval is shrinking. This fact is also presented in Figures 1 and 2 that show the posterior distributions of PVE from the MCMC runs for the corresponding MAF and each SNP subset. Values for PVE were larger when compared with the transformed estimate in Table 1.

Although the genomic control inflation factor was still small and mainly influenced by the distribution of cases collected across Germany,[6] genomic partitioning was used to check the final influence of the population structure.[20,32] Estimates were derived for chromosomes 1–7 and for the remaining chromosomes (8–22). For any threshold shown in Table 1, estimates of the two parts of the genome added up to the total estimates of common variance explained by SNPs as described in Table 1. The influence of the population structure has then also been tested with an additional REML analysis by fitting the first 2, 4 and 10 eigenvectors from the PCA as covariates. The results in Table 3 show little to no difference in the crude estimates of the variance explained by SNPs compared with original results in Table 1.

As many SNPs on chromosome 6 had extremely significant associations, one analysis was performed by excluding chromosome 6, or by analyzing chromosome 6 separately just as in the genome partitioning approach.[20] The results in Table 4 show that estimates (crude and adjusted) decreased by ~15–20% when excluding chromosome 6 (as compared with Table 1), whereas estimates based on SNPs on chromosome 6 solely account for ~5% of the variance explained by SNPs. This analysis shows that a substantial proportion of genetic variation is explained by risk loci on chromosome 6, yet another big proportion of the genetic variation is still explained by SNPs on other chromosomes.

Even though the incidence of 2–3/100 000/year has been stable for a long time,[2] any variation over time because of changes in the

Table 2 Proportion of estimated variance (PVE)

Threshold	*No. of SNPs*	*Interations sampled*	*PVE means*[a]	*Highest posterior density (CI = 0.95)*[b]
MAF[c] *> 0.01*				
GENO[d] > 0.05	583 333	10 000	0.45	0.13–0.62
GENO > 0.01	579 445	10 000	0.44	0.16–0.61
GENO > 0.005	566 831	10 000	0.39	0.17–0.53
GENO > 0.001	442 325	10 000	0.42	0.19–0.55
MAF > 0.05				
GENO > 0.05	540 228	10 000	0.43	0.06–0.60
GENO > 0.01	536 668	10 000	0.48	0.09–0.62
GENO > 0.005	525 248	10 000	0.40	0.07–0.57
GENO > 0.001	410 973	10 000	0.39	0.11–0.54

[a]Means of PVE.
[b]95% Confidence interval of the highest posterior density.
[c]Minor allele frequency.
[d]Maximum per-SNP missing rate.

Table 1 Estimated genetic variances for different setups

Threshold	*No. of SNPs*	*Estimate (SE)*[a]	*Adjusted (SE)*[b]	*Transformed (SE)*[c]
MAF [d] *> 0.01*				
GENO[e] > 0.05	583 333	0.24 (0.03)	0.24 (0.05)	0.35 (0.06)
GENO > 0.01	579 445	0.23 (0.03)	0.23 (0.05)	0.35 (0.06)
GENO > 0.005	566 831	0.23 (0.03)	0.23 (0.05)	0.35 (0.06)
GENO > 0.001	442 325	0.22 (0.03)	0.22 (0.05)	0.35 (0.05)
MAF > 0.05				
GENO > 0.05	540 228	0.23 (0.03)	0.23 (0.05)	0.35 (0.05)
GENO > 0.01	536 668	0.23 (0.03)	0.23 (0.05)	0.35 (0.05)
GENO > 0.005	525 248	0.22 (0.03)	0.22 (0.05)	0.35 (0.05)
GENO > 0.001	410 973	0.21 (0.03)	0.21 (0.05)	0.35 (0.05)

[a]Estimated genetic variance on the observed scale with SE.
[b]Estimated genetics variance and SE on observed scale adjusted for LD after Yang *et al*.[19]
[c]Estimated genetic variance and SE transformed to the liability scale after Dempster and Lerner.[21]
[d]Minor allele frequency.
[e]Maximum per-SNP missing rate.

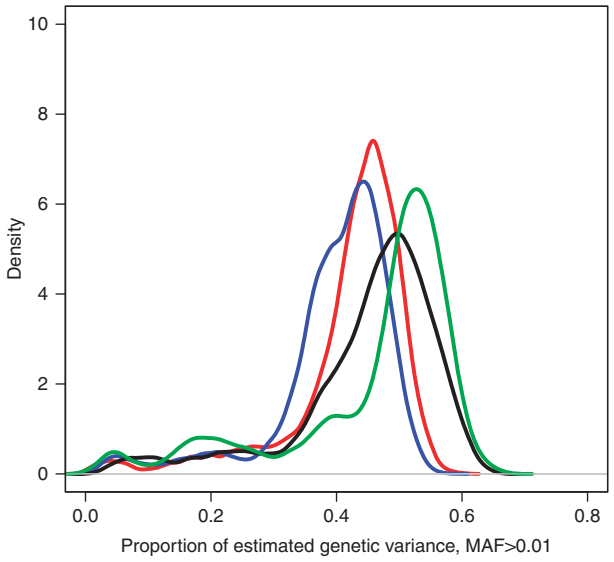

Figure 1 Proportion of estimated variance for MAF >0.01 with maximum per-SNP missing rate: red = 0.001; blue = 0.005; black = 0.01; green = 0.05.

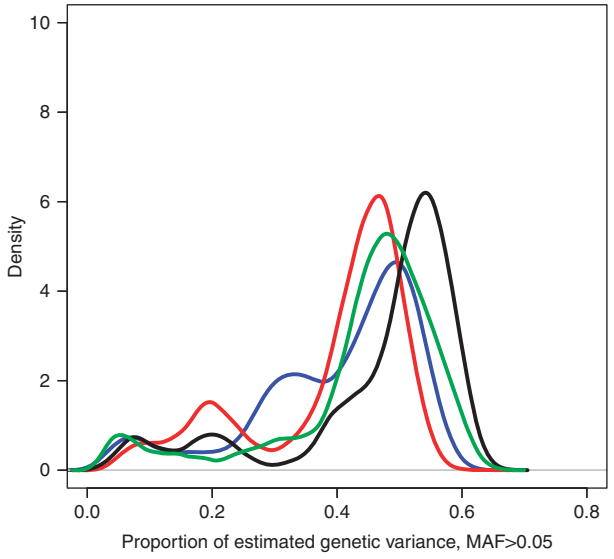

Figure 2 Proportion of estimated variance for MAF >0.05 with maximum per-SNP missing rate: red = 0.001; blue = 0.005; black = 0.01; green = 0.05.

environment would result in changes of the prevalence. Cutting the used prevalence in half to ~0.5 in 1000 or doubling it to ~2 in 1000 did not have any influence on the estimates of the common variance at any threshold shown in Table 1, but for a prevalence of ~0.5 in 1000, the transformed estimate dropped to 0.33 (SE: 0.04), whereas it increased to 0.40 (SE: 0.05) for a prevalence of ~2 in 1000.

According to *Wray et al*[33] we could interpret our result of the common genetic variance explained by SNPs on the liability scale to mean that we must expect many genetic variants underlying the disease. As the risks of common variants are too small to be used individually as risk predictors, the overall transformed estimate of 0.35 (Table 1) can be translated to a sibling relative risk of 5.58 being associated with common genetic variation.[33]

Table 3 Genetic variances explained by all SNPs by fitting first 2, 4 and 10 principal components (PCs) as covariates in the REML analysis

Threshold	*First 2 PCs*[a] *estimate (SE)*[b]	*First 4 PCs estimate (SE)*[b]	*First 10 PCs estimate (SE)*[b]
MAF[c] *> 0.01*			
GENO[d] > 0.05	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)
GENO > 0.01	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)
GENO > 0.005	0.24 (0.04)	0.23 (0.04)	0.21 (0.04)
GENO > 0.001	0.24 (0.04)	0.23 (0.04)	0.21 (0.04)
MAF > 0.05			
GENO > 0.05	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)
GENO > 0.01	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)
GENO > 0.005	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)
GENO > 0.001	0.24 (0.04)	0.23 (0.04)	0.22 (0.04)

[a]Principal components.
[b]Estimated genetic variance on the observed scale with SE.
[c]Maximum per-SNP missing rate.
[d]Minor allele frequency.

Heritability estimates based on population data

Figure 3 shows a trace plot of the heritability values along the iterations. There is no trend in the trace and values spread widely with a reasonable parameter space. The plot clearly illustrates good mixing, and a Gibbs sampler that 'converges' fast. The right side of Figure 3 shows the posterior density of the heritability estimate as a result of the model described earlier applied to the data set. Averaged across the 10 000 samples, posterior means of the heritability were 0.40 on the liability scale with a corresponding 95% highest posterior density region (HPD95) ranging from 0.17 to 0.58. The heritability on the observed scale is $z^2/K(1-K)$ times the heritability on the underlying normally distributed liability scale that is then 0.24. According to the convergence criteria, the effective sample size for estimating the mean derived from our data set was ~7186, representing a sufficient number, given the fact that we had 10 000 samples from the Gibbs sampler.[31]

DISCUSSION

Results from the genomic analysis as well as the population-based analysis show the proportion of variance explained by SNPs and heritability values for the susceptibility to HL in an overlapping range from 0.21 to 0.48, whereas estimates on the liability scale are ranging from 0.35 to 0.48 only. Most of these values represent simply the proportion of the total variance that is attributable to the pure additive genetic variance. Estimates of the PVE as a result of the probit-based approach ranged from 0.39 to 0.48, and were somehow higher compared with estimates of the genetic variance explained by SNPs on the liability scale shown in Table 1. This can be explained by a slightly different concept of the PVE compared with the approaches of Yang *et al*[19] and Speed *et al*[20]: PVE reflects the optimal predictive accuracy achieved for a linear combination of the measured genetic variants, whereas heritability reflects the accuracy that one can achieve by using all genetic variants.[23] Simulations of Guan and Stephens[23] have proven that the uncertainty in PVE is greater with a larger number of SNPs, presumably because of the increased difficulty in

Table 4 Analysis without chromosome 6 or chromosome 6 only

Threshold	*No. of SNPs*	*Estimate (SE)*[a]	*Adjusted (SE)*[b]	*Transformed (SE)*[c]
Analysis without chromosome 6				
MAF >0.01				
GENO >0.05	539 581	0.19 (0.03)	0.20 (0.03)	0.35 (0.05)
GENO >0.01	536 307	0.19 (0.03)	0.19 (0.03)	0.35 (0.05)
GENO >0.005	525 240	0.19 (0.03)	0.19 (0.03)	0.35 (0.05)
GENO >0.001	410 698	0.19 (0.03)	0.19 (0.03)	0.35 (0.05)
MAF >0.05				
GENO >0.05	500 018	0.19 (0.03)	0.19 (0.03)	0.35 (0.05)
GENO >0.01	497 100	0.19 (0.03)	0.19 (0.03)	0.35 (0.05)
GENO >0.005	486 988	0.19 (0.03)	0.18 (0.03)	0.35 (0.05)
GENO >0.001	381 802	0.19 (0.03)	0.18 (0.03)	0.35 (0.05)
Analysis of chromosome 6 only				
MAF[d] >0.01				
GENO[e] >0.05	43 752	0.05 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.01	43 138	0.05 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.005	41 591	0.05 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.001	31 627	0.05 (0.01)	0.04 (0.01)	0.10 (0.01)
MAF >0.05				
GENO >0.05	40 210	0.04 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.01	39 568	0.04 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.005	38 260	0.04 (0.01)	0.04 (0.01)	0.10 (0.01)
GENO >0.001	29 171	0.04 (0.01)	0.03 (0.01)	0.10 (0.01)

[a]Estimated genetic variance with SE on the observed scale.
[b]Estimated genetic variance with SE on observed scale adjusted for LD after Yang *et al*.[19]
[c]Estimated genetic variance with SE transformed to the liability scale after Dempster and Lerner.[21]
[d]Minor allele frequency.
[e]Maximum per-SNP missing rate.

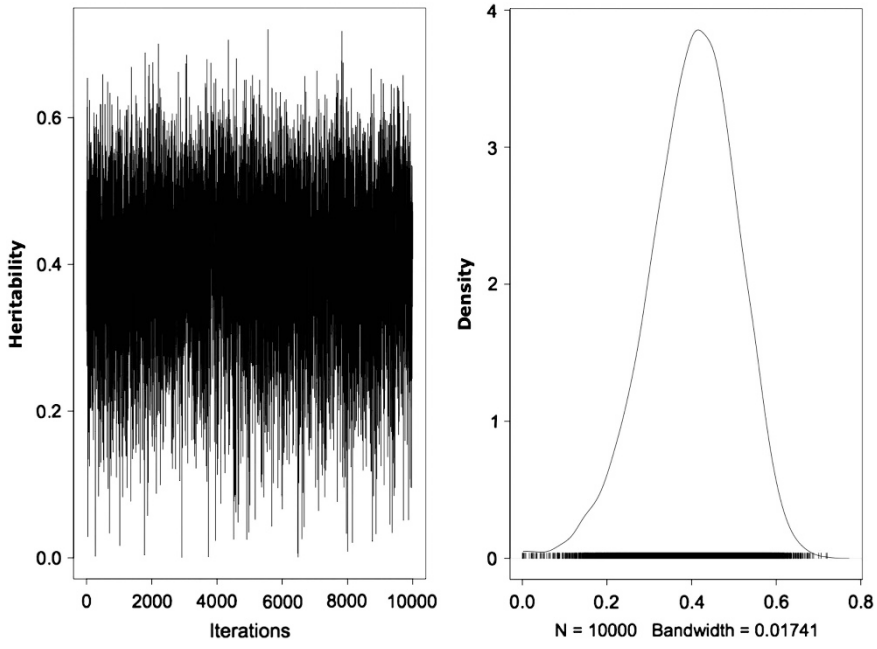

Figure 3 Trace (left) and posterior density (right) of heritability estimate.

reliably identifying relevant variants. When data contain no SNPs with strong individual effects, it remains difficult to rule out the possibility that many SNPs may have very small effects that combine to produce an appreciable PVE.[23] This uncertainty is also represented by the rather large confidence intervals in the right column of Table 2. It will be even greater when the true PVE is smaller.[23] Nonetheless, the range of the posterior on PVE nicely spans the estimates provided by the other methods equally to both sides, proving the ability of the PVE method to quantify uncertainty in multivariate problems by assessing the full joint posterior distribution of the model parameters compared

with the current simplistic 'one SNP at a time' testing paradigm,[34,35] because analyzing all SNPs simultaneously will detect more of the genetic variation because of the identification of multiple causal variants.[36]

In contrast to the results of Enciso-Mora *et al*,[37] our estimates of the variance are almost constant across the different numbers of SNPs and do not decline after a more stringent exclusion of missing genotypes. This is in agreement with the results by Yang *et al*[19] and Lee *et al*.[38] Thus, our QC has been stringent enough beforehand.

Inflation of the estimated variances explained by SNPs due to population stratification has been investigated by genomic partitioning. Results did not provide any evidence of stratification. Estimates of the two parts of the genome added up to the corresponding estimates on the full set (right column in Table 1). Our stringent QC helped to keep the genomic control inflation factor low.[6]

A cause for concern in estimating the common genetic variance explained by SNPs is created by LD that can lead to large biases. Contributions to heritability estimates from causal variants might be overestimated because of regions with strong LD or underestimated in regions with low LD.[20] We also followed the proposal of Speed *et al*[20] to analyze our data, but we did not detect any perceptible deviations in our estimates compared with the method of Yang *et al*.[19] It seems like any underestimation of contributions to the heritability in low-LD regions is balanced by overestimation elsewhere as shown by Speed *et al*.[20]

Trends in cancer prevalence reveal the dynamics of cancers in the population. To test the influence of changes on the estimates of the common genetic variance explained by SNPs, we have halved and doubled the prevalence. Differences to the original prevalence could only be seen in the transformed variance, but standard estimates stayed constant. Thus, variation in the transformed estimates may reflect changes in the environment over time.

The estimates of the common genetic variance explained by SNPs are based on ~410 000 to almost 600 000 SNPs. A significant effect on HL is harbored in the MHC region on chromosome 6.[7,39] After excluding SNPs mapped to the MHC region (6p21, at 28–33 Mb), Enciso-Mora *et al*[39] remained only with a limited number of loci influencing the risk of HL. Nonetheless, Enciso-Mora *et al*[39] and Frampton *et al*[6] were able to identify suggestive associations on another eight autosomes. After we excluded chromosome 6 from our analysis, the estimates dropped by ~20%. A rather high proportion of the variance is thus explained by chromosome 6 only as shown in Table 4, but still a descent proportion of variance is explained by the remaining autosomes. We therefore conclude that many additional loci may contribute to the susceptibility of HL.

The link function for binary data most widely used is the probit link, also known as the threshold model.[40] Modeling a random variable by using the probit function assumes that the latent variable (liability) possesses a standard normal distribution. The major advantage is that the liability is treated as a polygenic trait that is determined by many genes with small effects, and therefore heritability of liability is independent of the disease prevalence and can be directly compared.[40]

Our heritability estimates on the liability scale based on the Swedish population are larger than estimates by Shugart *et al*,[8] who calculated heritability for HL to be 28.4%. A reason for their lower estimate is a bias in Falconer's method[8] that is a result of common familiar environmental factors and ascertainment.[41] Generally, by using the extensive pedigree with its whole range of relationships in the population, the so-called animal threshold model provides the most accurate approximation of the heritability.[42] Our estimates are larger, but still conservative, because the applied threshold model only estimated the contributions of genes that act additively: the effects of genes with nonadditive effects, such as dominance or epistasis, will not contribute to the heritability estimates reported here. And even though statistical models are available for the estimation of non-additive genetic variance, much larger sets of suitably structured data would be required to obtain reliable estimates.

Heritability estimates on the liability scale are slightly larger for the population-based data than estimates of the common genetic variance explained by SNPs (0.40 compared with 0.35 for the GCTA approach), but they were in a very similar range to PVE (Table 2) that gave larger estimates because of reasons explained above. Nonetheless, heritability estimated from pedigree data is not the same as the proportion of phenotypic variation explained by all SNPs because the former includes the contribution of all causal variants, but the latter only includes the contribution of causal variants that are in LD with the genotyped SNPs. Thus, we also face the problem of missing heritability known for analysis of GWAS data.[11] Our population-based estimates and the estimates from SNP data still do not differ too much as estimates for other traits.[16] We therefore conclude that our genotypic data is a well-formed sample to draw sufficient conclusion about the genetic determination of the disease. Even though estimates of the common genetic variance explained by SNPs and the population-based heritability are similar, we must point out that many reasons for missing heritability have been widely accepted in the scientific community. This knowledge is supported by our findings of different estimates of the genetic variance explained by SNPs on chromosome 6 compared with other chromosomes. Genes detected on chromosome 6 in earlier studies do not account for much of the heritability of HL in our study.[7,39] The susceptibility to HL is rather a combined effect of multiple genes on several chromosomes than that of a few disease genes on a single chromosome.

The common genetic variance explained by SNPs and the heritability on the liability scale of HL show that a reasonable proportion of the variation observed in both the German and the Swedish population is caused by variation in genotypes, but it also indicates that the environment is still the principal causative role in HL, and susceptibility genes described so far for HL are likely to explain only part of the genetic effects.

An important fact is the interpretation of the sibling relative risk that has be derived through the incidence of the disease and the estimates of common variation explained by SNPs. Based on our results, siblings of people with HL are ~5.6 times more likely to develop the disease than others. These results are in agreement with studies showing up to a sevenfold increased risk in people with a parent or sibling diagnosed with HL.[43] It therefore seems to be clear that besides the environment, genetic factors have strong influence on the etiology of HL.

In conclusion, there is genetic variation for the susceptibility to HL. Heritability based on the population data is somehow larger than for the genomic data showing the possibility of some missing heritability in the GWAS data. Besides that, there is still major evidence for multiple loci causing HL on chromosomes other than chromosome 6.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful for the technical support on the program package DMU by Per Madsen and Guosheng Su.

1 Bose S, Ganesan C, Pant M, Lai C, Tabbara IA: Lymphocyte-predominant Hodgkin disease: a comprehensive overview. *Am J Clin Oncol* 2013; **36**: 91–96.
2 Engert A, Horning SJ: *Hodgkin Lymphoma. A Comprehensive Update on Diagnostics and Clinics: Hematologic malignancies*. Heidelberg; New York: Springer-Verlag, 2011, pp 1, online resource (ix, 381 p).
3 Parkin DM: 11. Cancers attributable to infection in the UK in 2010. *Br J Cancer* 2011; **105**(Suppl 2): S49–S56.
4 Anderson LA, Gadalla S, Morton LM *et al*: Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies. *Int J Cancer* 2009; **125**: 398–405.
5 Hemminki K, Czene K: Attributable risks of familial cancer from the Family-Cancer Database. *Cancer Epidemiol Biomarkers Prev* 2002; **11**: 1638–1644.
6 Frampton M, da Silva Filho MI, Broderick P *et al*: Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. *Nat Commun* 2013; **4**: 2549.
7 Urayama KY, Jarrett RF, Hjalgrim H *et al*: Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr virus status-defined subgroups. *J Natl Cancer Inst* 2012; **104**: 240–253.
8 Shugart YY, Hemminki K, Vaittinen P, Kingman A, Dong C: A genetic study of Hodgkin's lymphoma: an estimate of heritability and anticipation based on the familial cancer database in Sweden. *Hum Genet* 2000; **106**: 553–556.
9 Eichler EE, Flint J, Gibson G *et al*: Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**: 446–450.
10 Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 2013; **14**: 507–515.
11 Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L: Finding the sources of missing heritability in a yeast cross. *Nature* 2013; **494**: 234–237.
12 Wilson AJ: Why h2 does not always equal VA/VP? *J Evol Biol* 2008; **21**: 647–650.
13 Hill WG: Understanding and using quantitative genetic variation. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**: 73–85.
14 Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM: Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* 2013; **47**: 75–95.
15 Falconer DS: *Introduction to Quantitative Genetics*, 3rd edn. Burnt Mill, Harlow, Essex, England New York: Longman, Wiley, 1989.
16 Lee SH, Wray NR, Goddard ME, Visscher PM: Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88**: 294–305.
17 Schmermund A, Mohlenkamp S, Stang A *et al*: Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. *Am Heart J* 2002; **144**: 212–218.
18 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
19 Yang J, Benyamin B, McEvoy BP *et al*: Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
20 Speed D, Hemani G, Johnson Michael R, Balding David J: Improved Heritability Estimation from Genome-wide SNPs. *Am J Hum Genet* 2012; **91**: 1011–1021.
21 Dempster ER, Lerner IM: Heritability of threshold characters. *Genetics* 1950; **35**: 212–236.
22 Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**: 76–82.
23 Guan Y, Stephens M: Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat* 2011; **5**: 1780–1815.
24 Hemminki K, Ji J, Brandt A, Mousavi SM, Sundquist J: The Swedish Family-Cancer Database 2009: prospects for histology-specific and immigrant studies. *Int J Cancer* 2010; **126**: 2259–2267.
25 R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013, ISBN 3-900051-07-0. http://www.R-project.org/.
26 Madsen P: *User's Guide to DmuTrace: A Package for Preparing Data Sets*, Version 2. Aarhus, UK: University of Aarhus, Faculty of Agricultural Sciences, Department of Animal Breeding and Genetics, 2012.
27 Sorensen D, Gianola D: *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer-Verlag, 2002.
28 Odegard J, Meuwissen TH, Heringstad B, Madsen P: A simple algorithm to estimate genetic variance in an animal threshold model using Bayesian inference. *Genet Sel Evol* 2010; **42**: 29.
29 Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer, 1998.
30 Madsen P, Jensen J: *DMU: A User's Guide. A Package for Analyzing Multivariate Mixed Models*, Version 6, release 4.7. Aarhus, UK: University of Aarhus, Faculty of Agricultural Sciences, Department of Animal Breeding and Genetics, 2007.
31 Hadfield JD: MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. *J Stat Softw* 2010; **33**: 1–22.
32 Yang J, Manolio TA, Pasquale LR *et al*: Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 2011; **43**: 519–525.
33 Wray NR, Yang J, Goddard ME, Visscher PM: The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010; **6**: e1000864.
34 Gelman A: *Bayesian Data Analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC, 2004.
35 Stephens M, Balding DJ: Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 2009; **10**: 681–690.
36 Ovaskainen O, Cano JM, Merila J: A Bayesian framework for comparative quantitative genetics. *Proc Biol Sci R Soc* 2008; **275**: 669–678.
37 Enciso-Mora V, Hosking FJ, Sheridan E *et al*: Common genetic variation contributes significantly to the risk of childhood B-cell precursor acute lymphoblastic leukemia. *Leukemia* 2012; **26**: 2212–2215.
38 Lee SH, DeCandia TR, Ripke S *et al*: Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; **44**: 247–250.
39 Enciso-Mora V, Broderick P, Ma Y *et al*: A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nat Genet* 2010; **42**: 1126–1130.
40 Gianola D, Foulley J: Sire evaluation for ordered categorical data with a threshold model. *Genet Sel Evol* 1983; **15**: 201–224.
41 Tenesa A, Haley CS: The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* 2013; **14**: 139–149.
42 Benchek P, Morris NJ: How meaningful are heritability estimates of liability? *Hum Genet* 2013; **132**: 1351–1360.
43 Goldin LR, Pfeiffer RM, Gridley G *et al*: Familial aggregation of Hodgkin lymphoma and related tumors. *Cancer* 2004; **100**: 1902–1908.