

Representation of genetic association via attributable familial relative risks in order to identify polymorphisms functionally relevant to rheumatoid arthritis

Justo Lorenzo Bermejo^{*1,2}, Christine Fischer³, Anke Schulz⁴, Nadine Cremer⁴, Rebecca Hein⁴, Lars Beckmann⁴, Jenny Chang-Claude⁴ and Kari Hemminki^{2,5}

Addresses: ¹Institute of Medical Biometry and Informatics, University Hospital Heidelberg, INF 305, 69120 Heidelberg, Germany, ²Division of Molecular Genetic Epidemiology, German Cancer Research Center, INF 520, 69120 Heidelberg, Germany, ³Institute of Human Genetics, University of Heidelberg, INF 366, 69120 Heidelberg, Germany, ⁴Division of Cancer Epidemiology, German Cancer Research Centre, INF 280, 69120 Heidelberg, Germany and ⁵Center for Family and Community Medicine, Karolinska Institute, 14183 Huddinge, Sweden

E-mail: Justo Lorenzo Bermejo^{*} - Justo.Lorenzo@imbi.uni-heidelberg.de; Christine Fischer - Christine.Fischer@med.uni-heidelberg.de; Anke Schulz - anke.schulz@dkfz.de; Nadine Cremer - n.cremer@dkfz.de; Rebecca Hein - r.hein@dkfz.de;

Lars Beckmann - l.beckmann@dkfz.de; Jenny Chang-Claude - j.chang-claude@dkfz.de; Kari Hemminki - k.hemminki@dkfz.de

^{*}Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, 3(Suppl 7):S10 doi: 10.1186/1753-6561-3-S7-S10

This article is available from: <http://www.biomedcentral.com/1753-6561/3/S7/S10>

© 2009 Bermejo et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The results from association studies are usually summarized by a measure of evidence of association (frequentist or Bayesian probability values) that does not directly reflect the impact of the detected signals on familial aggregation. This article investigates the possible advantage of a two-dimensional representation of genetic association in order to identify polymorphisms relevant to disease: a measure of evidence of association (the Bayes factor, BF) combined with the estimated contribution to familiarity (the attributable sibling relative risk, λ_s). Simulation and data from the North American Rheumatoid Consortium (NARAC) were used to assess the possible benefit under several scenarios. Simulation indicated that the allele frequencies to reach the maximum BF and the maximum attributable λ_s diverged as the size of the genetic effect increased. The representation of BF versus attributable λ_s for selected regions of NARAC data revealed that SNPs involved in replicated associations clearly departed from the bulk of SNPs in these regions. In the 12 investigated regions, and particularly in the low-recombination major histocompatibility region, the ranking of SNPs according to BF differed from the ranking of SNPs according to attributable λ_s . The present results should be generalized using more extensive simulations and additional real data, but they suggest that a characterization of genetic association by both BF and attributable λ_s may result in an improved ranking of variants for further biological analyses.

Introduction

Susceptibility to rheumatoid arthritis (RA) is determined by both genetic and environmental factors, with an estimated sibling relative risk of 5-10. The *HLA-DRB1* and *PTPN22* genes explain approximately 50% of the familial aggregation. A recent genome-wide association (GWA) study carried out by the Wellcome Trust Case Control Consortium (WTCCC) found nine single-nucleotide polymorphisms (SNPs) with associated probability values of less than 10^{-5} that did not map to loci previously related to RA [1]. In agreement with the common disease-common variant hypothesis, the novel variants were common (minor allele frequencies from 0.07 to 0.41) and they showed genotype relative risks (GRRs) of less than 2.3. The authors of the WTCCC article recognized a disparity between the large population-attributable fraction (PAF) and the modest sibling relative risk (λ_s) explained by the newly identified variants [1]. This disparity is related to the technical characteristics of the platforms used in GWA studies, designed to scrutinize polymorphisms (common variants).

Let us assume that a polymorphism is a marker of a rarer causal variant, and the marker-specific PAF equals the PAF of the causal variant. Under this assumption, we have demonstrated that the causal variant has a higher attributable λ_s than the marker [2]. In other words: for a fixed PAF, causal variants contribute more to familial aggregation (λ_s) than the polymorphisms they are represented by.

There is much debate concerning the representation of statistical evidence in GWA studies. The frequentist p -value depends on the power of the study and it does not provide a consistent measure of evidence because identical p -values do not imply identical evidence of genetic association [3]. In contrast, the Bayes factor (BF) statistic provides a single measure of the strength of evidence of association.

We hypothesize here that the representation of genetic association by BF together with the attributable λ_s may help to identify polymorphisms relevant to RA, and use simulation and data from the North American Rheumatoid Consortium (NARAC) to evaluate this assumption.

Methods

Derivation of attributable risks and Bayes factors

The sibling relative risk (λ_s) attributable to a gene reflects the extent to which a particular genetic locus contributes to familial aggregation. The calculation of attributable λ_s values was first described in an early paper of James [4]. If the frequency of the high-risk allele is denoted by RAF, the relative risk of RA for high-risk allele homozygotes

compared to low-risk allele homozygotes by GRR_{hom} , the relative risk of heterozygotes compared to low-risk allele homozygotes by GRR_{het} and the disease prevalence among low-risk allele homozygotes by κ_0 , the attributable λ_s is given by: $\lambda_s = 1 + (1/2V_a + 1/4V_d)/K^2$, where V_a (additive genetic variance divided by κ_0^2) equals $2RAF(1-RAF)[(1-RAF)(1-GRR_{het})-RAF(GRR_{hom}-GRR_{het})]^2$, V_d (dominance genetic variance divided by κ_0^2) equals $RAF^2(1-RAF)^2[1+GRR_{hom}-2GRR_{het}]^2$ and K (population prevalence divided by κ_0) equals $RAF^2GRR_{hom} + 2RAF(1-RAF)GRR_{het} + (1-RAF)^2$. Note that if the overall disease prevalence is 50%, λ_s cannot exceed two by definition. However, an important property of the attributable λ_s is that it only depends on RAF, GRR_{hom} , and GRR_{het} . That is, the attributable λ_s is independent of the disease prevalence among low-risk homozygotes κ_0 .

The BF statistic is the ratio of the probability of the observed data under the assumption that there is a true association to its probability under the null hypothesis (absence of association). A small BF provides evidence in favor of a true association. To investigate the relationship between attributable λ_s and the BF, we calculated the expected distribution of genotypes in 1000 cases and 1000 controls when $GRR_{hom} = 1.5, 2, 3, 4, 5$ and 10. Calculations were carried out for the dominant model, which assumes that $GRR_{het} = GRR_{hom}$, the recessive model ($GRR_{het} = 1$), and for the additive model ($GRR_{het} = (1+GRR_{hom})/2$). Under Hardy-Weinberg equilibrium, the expected distribution of genotypes in controls ($D = 0$) is: $Pr(G = aa|D = 0) = (1-RAF)^2$; $Pr(G = Aa|D = 0) = 2RAF(1-RAF)$; $Pr(G = AA|D = 0) = RAF^2$. The expected distribution of genotypes in cases ($D = 1$) satisfies:

$$GRR_{hom} = (Pr(D = 1 | G = AA)/Pr(D = 0 | G = AA)) / (Pr(D = 1 | G = aa)/Pr(D = 0 | G = aa))$$

$$\text{and } GRR_{het} = (Pr(D = 1 | G = Aa)/Pr(D = 0 | G = Aa)) / (Pr(D = 1 | G = aa)/Pr(D = 0 | G = aa)).$$

For example, if $RAF = 0.2$ and $GRR_{hom} = GRR_{het} = 2$ (dominance), the expected distribution is ($aa = 640, Aa = 320, AA = 40$) in controls and ($aa = 471, Aa = 471, AA = 59$) in cases.

The expected distributions of genotypes were investigated by Bayesian logistic regression. We considered a general three-genotype model of association. Calculation of BFs requires assumptions about effect sizes. We assumed $N(0,1)$ priors on the mean and the two genetic effects. The function *MCMClogit* (R package *MCMCpack* [5]) was used to generate a 10,000 sample dataset from the posterior distribution using a random walk Metropolis algorithm. The *BayesFactor* function in the same package was used to compute log marginal likelihoods for a model 'with' compared to a model 'without' genotypic information using a Laplace approximation.

Investigated regions

This study investigated regions around the 12 signals detected in the WTCCC study with associated probability values of less than 10^{-5} (Table 1). The left side of Table 1 represents the 12 selected regions from the WTCCC study. For example, the most associated marker on 1p13 was rs6679677. We retrieved NARAC data from 1000-kb regions centred on the positions of the 12 signals from the WTCCC study. For example, our first region comprised the chromosomal interval (10,901,850::11,901,850) around rs6679677. The WTCCC study reported two SNPs in 6p21 that resulted in two overlapping intervals which were independently investigated.

Calculation of attributable risks and Bayes factors for NARAC data

SNPs in the 12 selected intervals were extracted from NARAC data, and the association between RA and the retrieved SNPs was represented by 1) the base 10 logarithm of the BF ($\log_{10}BF$) and 2) the attributable λ_s . RA status (affected or unaffected) was modelled by logistic regression, taking into account the covariates gender, number of *HLA-DRB1* shared-epitope alleles (NN = 0, SN = 1 or SS = 2) and genotype (three levels). Individuals with missing information were excluded from the calculations. Only SNPs with the three genotypes represented in both cases and controls were considered.

As already mentioned, calculation of BFs requires assumptions on effect sizes. We calculated frequentist estimates of logistic regression coefficients over the

corresponding entire chromosomes using NARAC data. Density plots and tests of goodness of fit of frequentist estimates of $\log_{10}GRR_{Hom}$ and $\log_{10}GRR_{Het}$ for the seven investigated chromosomes indicated that GRR variation was better represented by the median absolute deviation (MAD) than by the standard deviation (data not shown). Therefore, we assumed $N(0, MAD^2)$ priors on the logarithms of genetic effects. *MCMClogit* and *BayesFactor* [5] were used to calculate BFs for each SNP in the selected regions based on NARAC data. The number of individuals with complete data and the observed allele frequencies in controls were used to draw 10,000 samples from a binomial distribution, which were combined with the posterior estimates of genetic effects to simulate the distribution of attributable λ_s for each SNP.

A bagplot is a bivariate generalization of the non-parametric univariate boxplot [6]. In a bagplot, the region which contains 50% of the observations with greatest bivariate depth is denominated the bag. Observations outside the bag expanded three times are considered outliers, they are too far away from the data's central bulk. These outliers are indicated by closed dots in univariate boxplots. Bagplots have been shown to be equivariant for linear transformations and not limited to elliptical distributions. Due to these properties and the characteristics of our data (unknown shape of the underlying distributions), we used non-parametric bagplots to identify deviating SNPs. Bagplots were determined using the function *compute.bagplot* in the R package *aplpack* [7].

Table 1: Description and results from the 12 investigated regions

WTCCC study ^a					NARAC data ^b			
Chr	SNP	Position	MAD of \log_{10}		Outlying SNPs, gene regions	\log_{10} (BF)	RAF	λ_s Median (5 th -95 th)
			GRR_{hom}	GRR_{het}				
1p13	rs6679677	11,401,850	0.282	0.199	rs2476601, <i>PTPN22</i>	3.91	0.084	1.048 (1.017-1.098)
1p36	rs6684865	2,578,391	0.282	0.199	-	-	-	-
1p31	rs11162922	80,284,079	0.282	0.199	-	-	-	-
4p15	rs3816587	25,093,513	0.255	0.188	rs12505556, <i>SLC34A2</i>	3.04	0.112	1.071 (1.011-1.630)
6p21a	rs6457617	32,771,829	0.306	0.206	rs2395175, <i>MHC</i>	16.48	0.148	1.164 (1.108-1.241)
					rs7765379, <i>MHC</i>	6.11	0.113	1.175 (1.053-1.668)
					rs2395175, <i>MHC</i>	16.48	0.148	1.164 (1.108-1.241)
6p21b	rs615672	32,682,149	0.306	0.206	rs7765379MHC	6.11	0.113	1.175 (1.053-1.668)
6q23	rs6920220	138,048,197	0.306	0.206	-	-	-	-
7q32	rs11761231	130,827,294	0.268	0.185	-	-	-	-
10p15	rs2104286	6,139,051	0.290	0.197	-	-	-	-
13q12	rs9550642	19,848,092	0.286	0.200	rs1407961, <i>ZMYM2</i>	3.59	0.920	1.033 (1.033-1.225)
21q22	rs2837960	41,433,788	0.288	0.208	rs468646, <i>MX1</i>	4.28	0.488	1.038 (1.019-1.057)
					rs466092, <i>MX2</i>	3.59	0.488	1.030 (1.016-1.048)
					rs3218258, <i>IL2RB</i>	9.63	0.771	1.061 (1.041-1.084)
22q13	rs743777	35,876,107	0.283	0.216	rs710183, <i>CSF2RB</i>	4.26	0.072	1.028 (1.011-1.183)
					rs8137446, <i>KCTD17</i>	3.48	0.903	1.035 (1.018-1.123)

^aResults from the WTCCC study.

^bResults based on NARAC.

Results

Theoretical relationship between Bayes factor and attributable risk

Table 1 shows the RAFs at which the maximum $\log_{10}\text{BF}$ and λ_s were reached assuming several GRR_{hom} values under the dominant, recessive and additive modes of inheritance. For example, under $\text{GRR}_{\text{hom}} = 2$ and dominance, the maximum $\log_{10}\text{BF}$ (25.3) was reached when $\text{RAF} = 0.22$ and the maximum λ_s (1.06) for $\text{RAF} = 0.17$. Alternative sample-sizes/priors would result in different $\log_{10}\text{BF}$ s, but the RAF to reach the maximum $\log_{10}\text{BF}$ would not change. The representation of $\log_{10}\text{BF}$ versus attributable λ_s (data not shown) indicated that the two parameters were correlated, but they did not correspond one-to-one. In the three models, the RAF corresponding to the highest attributable λ_s and the RAF corresponding to the highest $\log_{10}\text{BF}$ decreased with increasing GRR_{hom} values. The RAF at maximum $\log_{10}\text{BF}$ was always higher than the RAF at maximum λ_s , i.e., balanced designs resulted in stronger association evidences. The difference between RAF at maximum $\log_{10}\text{BF}$ and RAF at maximum attributable λ_s increased with increasing GRR_{hom} , for example $(0.24-0.21) = 0.03$ if $\text{GRR}_{\text{hom}} = 1.5$ versus $(0.14-0.05) = 0.09$ when $\text{GRR}_{\text{hom}} = 10$ in the dominant model.

Bayes factors and attributable sibling relative risks in the investigated regions

The analysis of complete chromosomes by frequentist logistic regression resulted in $2(\text{GRR}_{\text{hom}}, \text{GRR}_{\text{het}}) \times 8$ (chromosomes 1, 4, 6, 7, 10, 13, 21, 22) sets of genetic effects. Table 1 shows the MADs of $\log_{10}\text{GRR}_{\text{hom}}$ and $\log_{10}\text{GRR}_{\text{het}}$ used as scale parameters in the normal prior distributions. As expected, the *a priori* variance of genetic effects was highest for chromosome 6 (MHC region).

Figure 1 represents the bivariate distribution of $\log_{10}\text{BF}$ by attributable λ_s based on NARAC data for individual SNPs in the 12 explored regions. Gray points show SNPs within the bulk of data, outlying observations (observations outside the bag expanded three times) are represented by black points. The right part of Table 1 shows SNP numbers and close genes for the most

extreme outliers. For example, rs2476601 in the 1p13 region: based on NARAC data, the $\log_{10}\text{BF}$ was 3.91, the high-risk allele frequency in controls is 0.084, and the estimated attributable λ_s with 95% confidence interval was 1.048 (1.017 to 1.098) for this SNP. The other two regions on chromosome 1 did not reveal strong associations with RA. The investigation of the two overlapping intervals in chromosome 6p21 showed practically identical results; the highest attributable λ_s values were found for SNPs rs7765379 (1.175) and rs2395175 (1.164). Many SNPs in this region showed a strong association with RA. The investigation of the 6q23, 7q32, and 10p15 regions resulted in $\log_{10}\text{BF}$ values below three. The $\log_{10}\text{BF}$ and the attributable λ_s for SNP rs1407961 in the 13q12 region clearly departed from the majority of data. Two extreme outlier SNPs were identified in the 21q22 region and three outliers in the 22q13 region.

Discussion

This study explored the advantage of combining BF, a measure of statistical evidence of association, and the attributable λ_s , a measure of contribution to familial aggregation, in order to characterize the effect of a particular polymorphism on disease risk. Both simulation and NARAC-based results showed that BF usually correlates with the attributable λ_s . Most markers identified in GWA studies confer a slightly increased risk of disease. For this kind of modest genetic effects ($\text{GRR} \leq 2$), simulation indicated that the maximum $\log_{10}\text{BF}$ and the maximum attributable λ_s were expected to be found for variants with similar frequencies: around 20% (dominant), 40% (additive) and 70% (recessive penetrances). The larger the size of the genetic effect, the lower the RAF to reach the maxima, and the larger the difference between RAFs. Causal variants generally show stronger effects than markers from GWA studies. For example, a recent GWA study identified five novel breast cancer susceptibility loci. The highest GRR was 1.6, which can be compared with $\text{GRR} = 6$ to 30 for carriers of *BRCA1* mutations. Simulation suggested that, for this kind of strong genetic effects, attributable λ_s and BF are two different dimensions.

Table 2: Allele frequencies at maximum λ_s and maximum $\log_{10}(\text{BF})$

GRR_{hom}	Dominant model		Recessive model		Additive model	
	Max $\log_{10}(\text{BF})$	Max λ_s	Max $\log_{10}(\text{BF})$	Max λ_s	Max $\log_{10}(\text{BF})$	Max λ_s
1.5	0.24	0.21	0.70	0.66	0.39	0.40
2	0.22	0.17	0.64	0.61	0.38	0.33
3	0.20	0.13	0.61	0.54	0.33	0.25
4	0.17	0.10	0.59	0.48	0.26	0.20
5	0.16	0.08	0.57	0.45	0.24	0.17
10	0.14	0.05	0.49	0.33	0.17	0.09

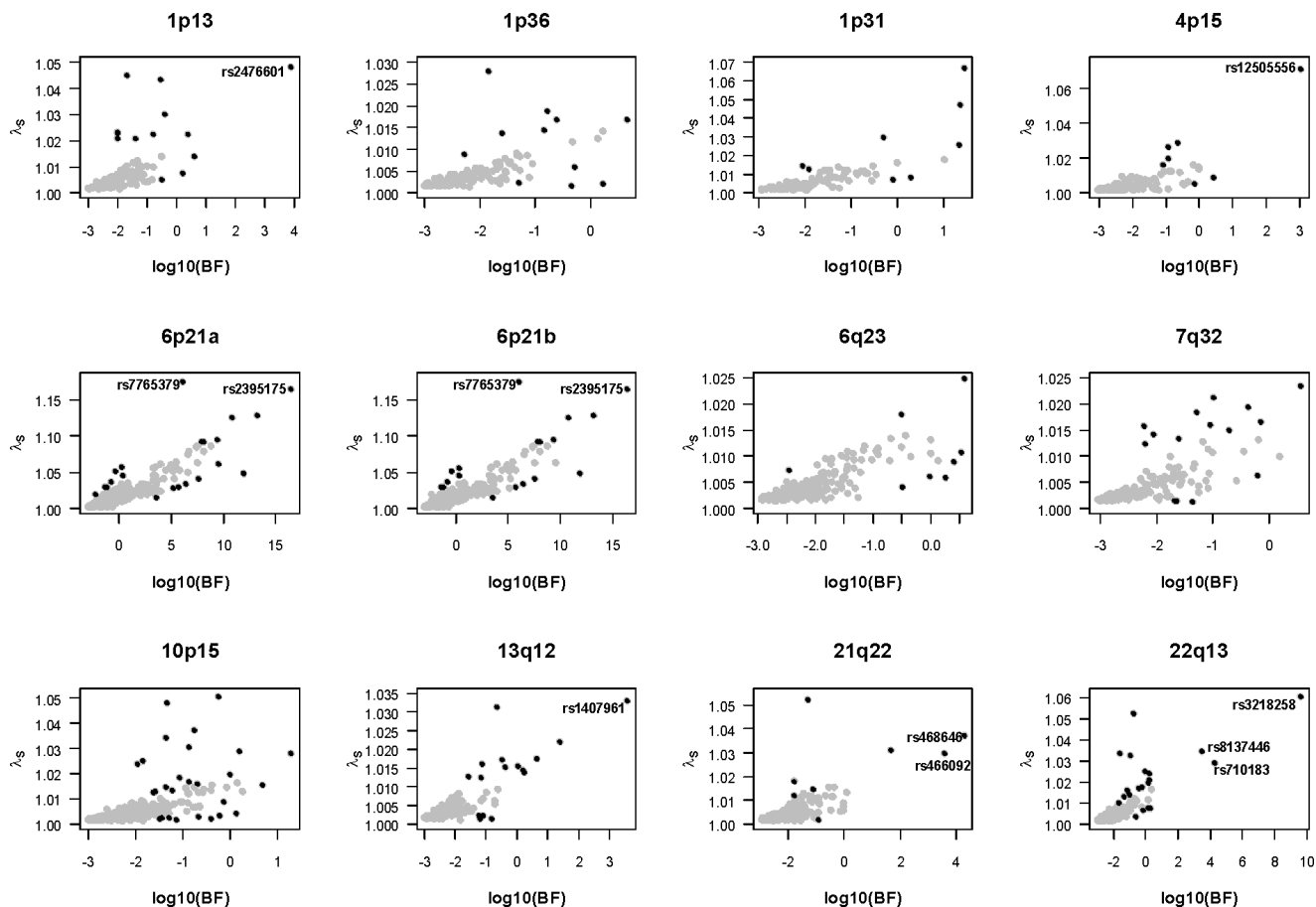


Figure 1
Scatterplots of \log_{10} values of Bayes factors ($\log_{10}(\text{BF})$) and attributable sibling relative risks (λ_s) for SNPs in the investigated regions. Outlying SNPs (black points) were identified by bagplots. SNP numbers are shown for most extreme outliers with $\log_{10}(\text{BF})$ s higher than three. Note scale differences.

The representation of λ_s versus $\log_{10}\text{BF}$ based on NARAC data confirmed that SNPs which have shown replicated associations clearly departed from the majority of polymorphisms in the selected regions. For example, the functionally relevant SNP rs2476601 was unequivocally separated from the rest of SNPs in the 1p13 region. Other genes located in the proximity of the most extreme SNPs were *ZMYM2*, *MX1*, *MX2*, *IL2RB*, and *CSF2RB*. This list includes SNPs which show strong evidence of association ($\log_{10}\text{BF}$ higher than three) and outlying λ_s values. SNPs with high attributable λ_s values but relatively low BF may also be of interest. Simulation showed that one BF value corresponds to two different λ_s values. In the 12 regions investigated using NARAC data, the ranking of SNPs according to BF differed from the ranking according to attributable λ_s . The difference was particularly clear in the low-recombination MHC region, where rs7765379 showed the maximum λ_s and rs2395175 showed the maximum $\log_{10}\text{BF}$. This

difference in rankings suggests that, in addition to a measure of association evidence, the extensive fine-mapping needed to characterize etiological variants may benefit from the consideration of the attributable λ_s . Future studies based on more extensive simulations and additional real data should investigate this possibility.

The present study made use of Bayesian and robust statistics. The Bayesian approach has been overlooked in the analysis of GWA studies, where the adjustment for multiple testing, the relationship between power and statistical significance, and the selection of disease models are important issues [1,3]. We only considered SNPs with the three genotypes represented in both cases and controls. This strategy would initially hamper the identification of rare causal variants but, on the other hand, logistic regression has not been robust to particular types of outliers, and a three-genotype model with well represented categories would result in more

robust estimates. Robust methods aim to describe the structure best fitting to the bulk of the data, and to identify deviating observations. Robust statistics are used a lot less than they ought to be in genetic epidemiology. The use of MADs and bagplots in the present article illustrates their advantage over classical methods in practical situations.

Conclusion

The association results from GWA studies are usually summarized by a measure of evidence of association (frequentist or Bayesian probability values), which do not reflect the contribution of the identified signals to familial aggregation. We propose here a two-dimensional characterization of genetic association consisting of the attributable λ_s and the BF.

List of abbreviations used

BF: Bayes factor; GRR: Genotype relative risk; GWA: Genome-wide association; λ_s : sibling relative risk; MAD: Median absolute deviation; MHC: Major histocompatibility complex; PAF: Population-attributable fraction; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism; WTCCC: Wellcome trust case control consortium.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JLB designed the study, performed the statistical analyses and drafted the manuscript. CF and KH contributed to the study design and revised critically the manuscript. AS, NC, RH, LB, and JCC made substantial contributions to interpretation of results. All authors read and approved the final manuscript.

Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. JLB was supported by Deutsche Krebshilfe and the Swedish Cancer Society. RH was supported by NGFN 2 (SMP-GEM, PGE-S30T09), funded by the BMBF.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at <http://www.biomedcentral.com/1753-6561/3?issue=S7>.

References

1. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661–678.
2. Hemminki K, Försti A and Bermejo JL: **The 'common disease-common variant' hypothesis and familial risks.** *PLoS ONE* 2008, **3**:e2504.
3. Wakefield J: **Bayes factors for genome-wide association studies: comparison with P-values.** *Genet Epidemiol* 2009, **33**:79–86.
4. James JW: **Frequency in relatives for an all-or-none trait.** *Ann Hum Genet* 1971, **35**(1):47–49.
5. Martin AD and Quinn KM: **MCMCpack: (MCMC) Package.** http://mcmcpack.wustl.edu/wiki/index.php/Main_Page.
6. Rousseeuw PJ, Ruts I and Tukey JW: **The Bagplot: a Bivariate boxplot.** *Am Stat* 1999, **53**:382–387.
7. Wolf P: **aplpack: Another Plot PACKAGE.** <http://www.wiwi.uni-bielefeld.de/~wolf/software/aplpack/>.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

