ARTICLE

Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power

Tim Becker^{*,1} and Christine Herold¹

¹*Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany*

Recent developments in genome-wide association studies (GWAS) have lead to the localization of disease genes for many complex diseases. The scrutiny of the respective publications reveals, first, that statistical analysis is restricted typically to single-marker analysis in the first step, and that, second, the presence of multiple, independently associated SNPs within the same linkage disequilibrium (LD) region is a common phenomenon. Motivated by this observation, we show through a power simulation study that a simultaneous analysis of tightly linked SNPs in the initial GWAS analysis step would lead to increased power, when compared with that in single-marker analysis. This is true for all the three approaches we considered (implementations in BEAGLE, FAMHAP and UNPHASED). The best performance was obtained using a two-marker haplotype analysis. In conclusion, we would expect additional gene findings for re-analyzing successful GWAS with a multi-marker approach.

European Journal of Human Genetics (2009) 17, 1043–1049; doi:10.1038/ejhg.2009.7; published online 18 February 2009

Keywords: genome-wide association studies; haplotypes; literature study

Introduction

Biology strongly supports the potential importance of haplotype analysis in genetic association studies.¹ Major reasons for haplotype analysis are possible representations of non-genotyped SNPs and the presence of multiple, independent disease markers in the same linkage disequilibrium (LD) region.² Nevertheless, evidence that haplotypes actually lead to improved power in genetic association studies has yet to be supplied. Indeed, the increased number of degrees of freedom (d.f.) might over-compensate the benefit achieved from the improved modeling of biology obtainable with haplotypes. Besides that, it is not self-evident whether haplotype assignment is actually a prerequisite for association analysis in LD regions. Clayton

*et al*³ suggest, as an alternative, the analysis of unphased multi-marker genotype data in these regions.

In view of the controversial situation regarding the judgement of haplotype analysis, we conducted a large-scale power simulation study for investigating the relative performance of single-marker analysis, simultaneous analysis of unphased genotypes in LD regions – referred to as UMMA (unphased multi-marker analysis) from now on – and haplotype analysis in genome-wide association studies (GWAS). The setup of the simulation study was guided by *a priori* knowledge that has become recently available. Data from the International HapMap Project⁴ allowed specifying empirical LD distributions for data simulation. In addition, we extracted information on disease models typical of complex diseases, by studying a comprehensive list of recent publications on GWAS.

Results of literature study

The guidelines for inclusion into the literature study were publication during the time period from January 2005 to

*Correspondence: Dr T Becker, Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, Bonn D-53105, Germany.

Tel: +49 228 287 14812; Fax: +49 228 287 15854;

E-mail: Tim.Becker@ukb.uni-bonn.de

Received 29 August 2008; revised 1 December 2008; accepted 14 January 2009; published online 18 February 2009

June 2008, with publication in a high-impact journal (impact factor > 12) and identification of SNP(s) in the first step of GWAS followed by replication. In addition, we restricted ourselves to qualitative phenotypes in order to improve homogeneity with respect to disease model and statistical methods used. Supplementary Table 1 represents citations of 59 publications,^{5–42,44–64} which met our inclusion criteria. In all publications, association analysis in the initial step was restricted to single-marker analysis. Furthermore, all replicated SNPs had a minor allele frequency > 5%, and allelic odds ratios typically ranged from 1.1 to 2.2, with higher exceptions not only for non-disease phenotypes, such as pigmentation for instance, but also for macular degeneration. Thus, the results confirm predictions of the common-disease, common-variant hypothesis.⁶⁵ Several publications report follow-up analysis of the LD region around the confirmed SNPs and also test for deviations from a multiplicative SNP model, the presence of multiple independent SNPs within the LD region, haplotype effects or interaction of SNPs from different unlinked regions. Typically, a logistic regression framework was applied.⁴³ In Supplementary Table 1, the frequency of the presence of these phenomena is summarized. This table contains information on 104 LD regions relevant for 27 varying phenotypes in total. Genes described in more than one publication, or relevant for more than one phenotype, were counted only once. Sub-phenotypes (severe, aggressive, etc.) were summarized under the main phenotype. Deviations from a multiplicative SNP model were found for 10 SNPs (out of the tested 42), but were generally rather small in size, although significant. Interactions between unlinked regions were tested formally for 116 potentially interacting variants, but not a single statistically significant interaction was detected. The most interesting find was the frequency of LD regions with multiple SNPs, which were associated independently. The presence of multiple variants was tested for 39 out of 104 LD regions. For 18 out of these LD regions (46.1%) at least two SNPs were found, which were associated independently (after accounting for LD) with the phenotype. Notably, the portion estimate of 46.1% is likely to be biased downwards, as all 104 LD regions were identified in the initial GWA analysis step using only single-marker tests.

Methods

On the basis of information from the literature study, one should expect in reverse that a multi-marker approach in the initial GWA step leads to the identification of even more disease-associated LD regions, such as the identification of those LD regions with multiple variants, which does not reach significance with single-marker tests. Therefore, we tried evaluating the performance of the following GWAS-analysis approaches: single-marker analysis using

Armitage's⁶⁶ trend test (**GWA-1**); single-marker analysis and analysis of unphased genotypes for all pairs of SNPs < 30 kb apart in a allele effects model (2 d.f.) (**GWA-UMMA-2**); single-marker analysis and haplotype analysis (likelihood-ratio test with up to 3 d.f.) of all pairs of SNPs < 30 kb apart (**GWA-HAP-2**); haplotype cluster analysis within the 30-kb window as described by Browning and Browning⁶⁷ and implemented in BEAGLE (**GWA-HAP-CLUSTER**). Note that BEAGLE also included single-marker tests into the analysis. Thus, the results will be comparable with the other approaches.

The 30-kb distance was chosen, as it can be viewed as typical of the average LD extend in the human genome. Marchini *et al.*⁶⁸ earlier suggested a genome-wide analysis approach for all pairs of SNPs, including SNP pairs not present in LD. In contrast to this approach, our two-marker strategies have two advantages. First, the number of tests increases only linearly, which guarantees computational feasibility (see Discussion) and, moreover, reduces the price to pay for multiple testing. Second, according to the literature study, there is an increased *a priori* probability that SNPs pairs from the same LD region are associated with independent contributions to a complex disease, as opposed to that of the general pairs of SNPs. To test our hypothesis that **GWA-UMMA-2**, **GWA-HAP-2** and **GWA-HAP-CLUSTER** are strategies superior to **GWA-1**, we conducted a power simulation study, which was conceptually closely related to that earlier presented by de Bakker *et al.*⁶⁹ In brief, we simulated case-control data (5000 individuals) sets based on the CEU trio data from the HapMap ENCODE project.⁴ A more detailed description can be found in the appendix. We considered in total, 5000 windows of 62 SNPs. For each window, we obtained 'empirical' population haplotype frequencies using FAMHAP.⁷⁰ Next, we designated either one or two SNPs to be 'causal'. Model 1 (M1) was defined by one causal SNP with a multiplicative disease model. The allelic risk was chosen, such that with a complete marker panel a genome-wide power of 80% would have been reached at $\alpha = 0.05$. Model 2 (M2) was defined by the presence of two causal variants without interaction, and model 3 (M3) was a haplotype model with an increased risk ratio for the two-marker haplotype 1–1 defined by two disease variants. Finally, model 4 (M4) was a recessive haplotype model with two-marker risk haplotype 1–1. Before data analysis, the marker panels for each marker window were thinned out in such a way that they met varying levels, A–C, of SNP chip coverage, which can be viewed as representing the density of a 300, 500 K and 1 M SNP chip (see appendix, for a more detailed description). To evaluate power, we computed the best *P*-value *minP* according to each of the four different analysis approaches, for each simulated and thinned data set. These *minP* values were multiplied with an empirically determined number of independent tests (see appendix) to obtain genome-wide corrected *P*-values. Finally, power was

computed as the portion of SNP windows with genome-wide significance ($\alpha = 0.05$).

Results

The results of the power study can be found in Table 1. In model M1 (one disease variant), the power of the GWA-1 approach is 81.1% and, thus, very close to the intended reference power level of 80%. Trivially, power drops, to ~72% for both the two-marker approaches in model M1, because of the multiple testing penalty. We discuss the results of the most relevant marker panel C (1 M chip SNP density).

Notably, even in the single-disease variant model, M1, power increases (albeit very modestly) from 51.8% for GWA-1 to 52.1 and 52.0% for GWA-UMMA-2 and GWA-HAP-2, respectively. In other words, the increased multiple testing penalty of the two-marker approaches is over-compensated by a power increase caused by the approximation of unobserved disease SNPs with two SNPs through LD. Power of the GWA-CLUSTER approach (50.3%) is somewhat lower, but difference to the other approaches is not significant.

In disease model M2 (two disease variants, but no additional haplotype effect), both GWA-UMMA-2 (79.7%) and GWA-HAP-2 (79.7%) significantly outperform power of the single-marker approach GWA-1 (71.9%). Even though there is no haplotype effect, the GWA-HAP-2 approach performs as good as the GWA-UMMA-approach. It seems that the reduced d.f. of the UMMA-approach and the potential to tag non-genotyped SNPs with the HAP-approach are respective advantages of the two strategies that counterbalance each other. Power of the cluster approach lies in the range of the GWA-1 strategy. We conclude that haplotype clustering yields, at most, small

power improvements over single-marker analysis when there is no direct haplotype effect.

In the haplotype model M3, all multi-marker strategies significantly outperform the single-marker strategy. Now, the HAP-CLUSTER approach is most powerful (79.0%), followed by the GWA-HAP-2 (78.3%) and the GWA-UMMA-2 approaches (76.2%). For both the haplotype approaches, comparison with the UMMA approach is significant. In the haplotype model M4, we observe a very similar pattern, the only difference is that now HAP-2 performs best.

Summarizing the results for panel C, one can say that the two-marker-approaches are 'free lunch', in the sense that they do not cost power in the presence of a single disease variant, and that they lead to increased power in the presence of two disease variants within the same marker window, whereas the HAP-CLUSTER approach leads to markedly improved power only when there is a direct haplotype effect on the disease.

The conclusions drawn from the results for panel C also hold for the less-denser panels A and B. It seems that the power gain for the multi-marker approaches is a little bit smaller than for panel C, which indicates that dense marker maps (1 M) are needed for exploiting the benefit of multi-marker approaches.

Finally, we note that we did not observe any qualitative differences in our results, when we evaluated our data according to either the different ENCODE regions or disease allele frequencies.

Discussion

We have shown that, in the presence of multiple, independent disease-associated SNPs within the same LD region, multi-marker approaches lead to improved power when applied in the initial step of the GWAS. According to

Table 1 Power values

Marker panel	Disease model	GWA-1	GWA-UMMA-2	GWA-HAP-2	GWA-HAP-CLUSTER
A (300 K)	M1 (1 variant)	0.258	0.266	0.269	0.262
	M2 (2 variants)	0.472	0.523	0.523	0.492
	M3 (haplotype)	0.544	0.581	0.591	0.581
	M4 (haplotype recessive)	0.564	0.595	0.608	0.586
B (500 K)	M1 (1 variant)	0.358	0.365	0.369	0.354
	M2 (2 variants)	0.587	0.656	0.652	0.600
	M3 (haplotype)	0.640	0.691	0.706	0.706
	M4 (haplotype recessive)	0.668	0.693	0.712	0.682
C (1000 K)	M1 (1 variant)	0.518	0.521	0.520	0.503
	M2 (2 variants)	0.719	0.797	0.797	0.724
	M3 (haplotype)	0.718	0.762	0.783	0.790
	M4 (haplotype recessive)	0.741	0.768	0.802	0.789
Full sequence	M1 (1 variant)	0.811	0.723	0.720	n.d. ^a
	M2 (2 variants)	0.862	0.944	0.941	n.d.
	M3 (haplotype)	0.815	0.835	0.848	n.d.
	M4 (haplotype recessive)	0.802	0.821	0.846	n.d.

^aNot done.

our literature study, the phenomenon of multiple variants in LD is common and, hence, multi-marker strategies are highly recommendable. The advantage of multi-marker analysis has also been shown by Hintsanen *et al*⁷¹ through a simulation setup based on coalescent theory. Thus, our study is also an independent confirmation of earlier results with a different technique.

All three multi-marker approaches outperformed the single-marker approach. The power difference between the UMMA-2 (2 d.f.) and the HAP-2 haplotype approach was small, such that both strategies could be advocated. An advantage of the HAP-2 approach is that a fast implementation is part of the FAMHAP software update for GWAS.⁷² The haplotype cluster approach performed only slightly better than the two-marker approaches in explicit haplotype disease models (M3 and M4). It is not unlikely that the relative performance of the haplotype cluster approach would have been better in the presence of haplotype models with more than two independent disease markers. Thus, yet to be gained empirical knowledge on the commonness of such disease models might lead to a different judgement of results. However, in the two-marker model without phase effect (M2), the haplotype cluster method was inferior.

As an alternative to the approaches considered here, multi-marker information can be exploited for imputing SNP genotypes. Marchini *et al*,⁷³ for instance, have shown that in the presence of a single, not too common, disease variant, imputing leads to increased power when compared with single-marker analysis. The performance of imputing methods in the presence of multiple disease variants has not been investigated by now. Respective future work could give guidelines on the relative advantages of imputing and multi-marker approaches.

Although we believe in multi-marker approaches, we wish to emphasize that quality control is even more important than it is already for single-marker analysis. All used SNPs must pass particularly strong quality criteria, as the impact of a genotyping error increases with haplotype analysis. For instance, a rather small amount of errors, in a sub-batch of the case or control group, can produce quickly non-existing haplotypes, which leads to false positive results.

Our literature study reveals that not a single significant interaction has been found between the genes obtained through the GWAS initiatives of past years. However, we do not think that one should conclude that interaction of unlinked genes does not play an important role in complex diseases. Obviously, there is a strong bias towards non-interacting genes in today's results, as the genes were identified using single-marker strategy. Therefore, we are convinced that genome-wide strategy of testing all SNP pairs suggested by Marchini *et al*⁶⁸ can be useful. Our claim is, however, that before such a strategy, one should first follow a local multi-marker approach, because the *a priori*

probability for true association of a pair of neighboring SNPs is higher than for two random SNPs. Moreover, the two-marker approaches are computationally strongly less intense and, therefore, feasible for the majority of scientific groups.

Acknowledgements

This work was supported by Grant BE 3828/3-1 from the Deutsche Forschungsgemeinschaft.

Conflict of interest

The authors declare no conflict of interest.

References

- Clark AG: The role of haplotypes in candidate gene studies. *Genet Epidemiol* 2004; **27**: 321–333.
- Morris R, Kaplan N: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002; **23**: 221–233.
- Clayton D, Chapman J, Cooper J: Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 2004; **27**: 415–428.
- The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
- Mofatt MF, Kabisch M, Liang L *et al*: Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007; **448**: 470–474.
- Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
- Grant SF, Thorleifsson G, Reynisdottir I *et al*: Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006; **38**: 320–323.
- Steinhorsdottir V, Thorleifsson G, Reynisdottir I *et al*: A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 2007; **39**: 770–775.
- Lowe CE, Cooper JD, Brusko T *et al*: Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphisms in the IL2RA region in the IL2RA region in type 1 diabetes. *Nat Genet* 2007; **39**: 1074–1082.
- Saxena R, Voight BF, Lyssenko V *et al*: Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research: Genome-wide association analysis identifies loci for Type 2 diabetes and triglyceride levels diabetes. *Science* 2007; **316**: 1331–1336.
- Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336–1341.
- Scott LJ, Mohlke KL, Bonnycastle LL *et al*: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007; **316**: 1341–1345.
- Haiman CA, Le Marchand L, Yamamoto J *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Nat Genet* 2007; **39**: 954–956.
- Todd JA, Walker NM, Cooper JD *et al*: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007; **39**: 857–864.
- Hunter DJ, Kraft P, Jacobs KB *et al*: A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007; **39**: 870–874.
- Stacey SN, Manolescu A, Sulem P *et al*: Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007; **39**: 865–869.

- 17 van Heel DA, Franke L, Hunt KA *et al*: A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007; **39**: 827–829.
- 18 Buch S, Schafmayer C, Völzke H *et al*: A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat Genet* 2007; **39**: 995–999.
- 19 Zanke BW, Greenwood CM, Rangrej J *et al*: Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007; **39**: 989–994.
- 20 Tomlinson I, Webb E, Carvajal-Carmona L *et al*: A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007; **39**: 984–988.
- 21 Maller JB, Fagerness JA, Reynolds RC, Neale BM, Daly MJ, Seddon JM: Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet* 2007; **39**: 1200–1201.
- 22 Gudmundsson J, Sulem P, Steinthorsdottir V *et al*: Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 2007; **39**: 977–983.
- 23 Plenge RM, Seielstad M, Padyukov L *et al*: TRAF1–C5 as a risk locus for rheumatoid arthritis — a genomewide study. *N Engl J Med* 2007; **357**: 1199–1209.
- 24 Thorleifsson G, Magnusson KP, Sulem P *et al*: Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science* 2007; **317**: 1397–1400.
- 25 Fellay J, Shianna KV, Ge D *et al*: A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007; **317**: 944–947.
- 26 Hakonarson H, Grant SF, Bradfield JP *et al*: A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 2007; **448**: 591–594.
- 27 Gudbjartsson DE, Arnar DO, Helgadóttir A *et al*: Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007; **448**: 353–357.
- 28 Helgadóttir A, Thorleifsson G, Manolescu A *et al*: A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007; **316**: 1491–1493.
- 29 Gudmundsson J, Sulem P, Manolescu A *et al*: Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 2007; **39**: 631–637.
- 30 Rioux JD, Xavier RJ, Taylor KD *et al*: Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007; **39**: 596–604.
- 31 Duerr RH, Taylor KD, Brant SR *et al*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006; **314**: 1461–1463.
- 32 Yeager M, Orr N, Hayes RB *et al*: Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007; **39**: 645–649.
- 33 Amundadóttir LT, Sulem P, Gudmundsson J *et al*: A common variant associated with prostate cancer in European and African populations. *Nat Genet* 2006; **38**: 652–658.
- 34 Thomas G, Jacobs KB, Yeager M *et al*: Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* 2007; **40**: 310–315.
- 35 Eeles RA, Kote-Jarai Z, Giles GG *et al*: Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 2008; **40**: 316–321.
- 36 Kayser M, Liu F, Janssens AC *et al*: Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 2008; **82**: 1–13.
- 37 Thomson W, Barton A, Ke X *et al*: Rheumatoid arthritis association at 6q23. *Nat Genet* 2007; **39**: 1431–1433.
- 38 Broderick P, Carvajal-Carmona L, Pittman AM *et al*: A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* 2007; **39**: 1315–1317.
- 39 Jaeger E, Webb E, Howarth K *et al*: Common genetic variants at the CRAC1 (HMP5) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* 2008; **40**: 26–28.
- 40 van Es MA, van Vught PW, Blauw HM *et al*: Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat Genet* 2008; **40**: 29–31.
- 41 Harley JB, Alarcón-Riquelme ME, Criswell LA *et al*, International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN): Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PTK, KIAA1542 and other loci. *Nat Genet* 2008; **40**: 204–210.
- 42 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
- 43 Cordell H, Clayton D: A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 2002; **70**: 124–141.
- 44 Gudmundsson J, Sulem P, Rafnar T *et al*: Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* 2008; **40**: 281–283.
- 45 Plenge RM, Cotsapas C, Davies L *et al*: Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* 2007; **39**: 1477–1482.
- 46 Parkes M, Barrett JC, Prescott NJ *et al*: Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 2007; **39**: 830–832.
- 47 Winkelmann J, Schormair B, Lichtner P *et al*: Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet* 2007; **39**: 1000–1006.
- 48 Sulem P, Gudbjartsson DE, Stacey SN *et al*: Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 2007; **39**: 1443–1452.
- 49 Kozlyev SV, Abelson AK, Wojcik J *et al*: Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet* 2008; **40**: 211–216.
- 50 Easton DF, Pooley KA, Dunning AM *et al*: Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007; **447**: 1087–1093.
- 51 McPherson R, Pertsemlidis A, Kavaslar N *et al*: A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007; **316**: 1488–1491.
- 52 Duggan D, Zheng SL, Knowlton M *et al*: Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J Natl Cancer Inst* 2007; **99**: 1836–1844.
- 53 Hunt KA, Zhernakova A, Turner G *et al*: Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008; **40**: 395–402.
- 54 Tenesa A, Farrington SM, Prendergast JG *et al*: Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008; **40**: 631–637.
- 55 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.
- 56 Tomlinson IPM, Webb E, Carvajal-Carmona L *et al*: A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008; **40**: 623–630.
- 57 Amos CI, Wu X, Broderick P *et al*: Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008; **40**: 616–622.
- 58 Stacey SN, Manolescu A, Sulem P *et al*: Common variants on chromosome 5p12 confer susceptibility to estrogen-positive breast cancer. *Nat Genet* 2008; **40**: 703–706.

- 59 Gudbjartsson DE, Sulem P, Stacey SN *et al*: ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nat Genet* 2008; **40**: 886–891.
- 60 Brown KM, Macgregor S, Montgomery GW *et al*: Common sequence variants on 20q11.22 confer melanoma susceptibility. *Nat Genet* 2008; **40**: 838–840.
- 61 Sulem P, Gudbjartsson DE, Stacey SN *et al*: Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* 2008; **40**: 835–837.
- 62 Sakamoto H, Yoshimura K, Saeki N *et al*, Study Group of Millennium Genome Project for Cancer: Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat Genet* 2008; **40**: 730–740.
- 63 Hung RJ, McKay JD, Gaborieau V *et al*: A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008; **452**: 633–637.
- 64 Thorgeirsson TE, Geller F, Sulem P *et al*: A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008; **452**: 638–642.
- 65 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
- 66 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**: 375–386.
- 67 Browning BL, Browning SR: Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* 2007; **31**: 365–375.
- 68 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**: 413–417.
- 69 de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
- 70 Becker T, Knapp M: Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol* 2004; **27**: 21–32.
- 71 Hintsanen P, Sevon P, Onkamo P, Eronen L, Toivonen H: An empirical comparison of case-control and trio based study designs in high throughput association mapping. *J Med Genet* 2006; **43**: 617–624.
- 72 Herold C, Becker T: Genetic association analysis with FAMHAP: a major program update. *Bioinformatics* 2009; **25**: 134–136.
- 73 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
- 74 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–862.
- 75 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* 2008; **66**: 87–98.
- 76 Becker T, Joelson E, Cichon S *et al*: Multiple testing in the context of haplotype analysis revisited: application to case-control data. *Ann Hum Genet* 2005; **69**: 747–756.
- 77 Ge Y, Dudoit S, Speed TP: Resampling-based multiple testing for microarray data analysis. *Test* 2003; **12**: 1–77.
- 78 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases and seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–683.

Appendix

Simulation study

The CEU trio data from the HapMap ENCODE project,⁴ served as a ‘complete’ reference panel. We went through the ENCODE data in a sliding-window fashion with a window size of 62 SNPs. For each SNP window, we proceeded as follows: first, we obtained ‘empirical’ haplotype frequencies using FAMHAP,⁷⁰ that is, the inferred haplotype frequencies were assumed to be the ‘true’ population frequencies. Next, we designated either one or two SNPs of the window to have a ‘causal’ allele variant, and simulated 2500 cases and 2500 controls according to the allelic status at the variant, the haplotype distribution and various disease models. Model 1 (M1) was defined by one causal SNP with a multiplicative disease model. Disease prevalence was chosen to be either 1 or 10%. Allelic risk ratios were chosen depending on the allele frequency of the variant in such a way that with a complete marker panel, a genome-wide power of 80% would have been reached at $\alpha=0.05$. According to the results of an explorative pre-study, these specifications resulted in allelic risk ratios ranging from 1.27 to 1.50 for the disease allele frequency spectrum of 0.05–0.95. It must be noted, we did not consider variants outside the frequency range of 0.05–0.95. Model 2 (M2) was defined by the presence of two causal variants in the window. The allele frequency-dependent risk ratios were taken over from (M1), and the two SNPs were assumed to contribute multiplicatively to the disease (no interaction). Model 3 (M3) was a haplotype model with an increased risk ratio for only the two-marker

haplotype 1–1 defined by two disease variants. Finally, model 4 (M4) was a recessive haplotype model with a two-marker risk haplotype 1–1. Data panels for all the models were simulated for 5000 marker windows in total. Before data analysis, the marker panels for each marker window were thinned out in such a way that they met varying levels, A–C, of coverage. We considered, either 4(A), 7(B) or 15(C) out of the 62 SNPs in total. Note that under the assumption that the whole human genome would have 4 000 000 million SNPs, the SNP numbers per window would correspond approximately to the coverage of a 300K (A), 500K (B) or 1 M (C) SNP chip. However, we are aware of the fact that a recent publication estimated the number of SNPs in the human genome to be actually higher.⁷⁴ On the other hand, the HapMap reference sample is small and will not contain all human SNP variations. In any case, our main interest was relative performance of different strategies, such that the absolute power levels corresponding to different coverages are not the primary aspect. To evaluate power, we computed for each simulated and thinned data set the best *P*-value, *minP*, according to each of the analysis approaches (**GWA-1**), (**GWA-UMMA-2**), (**GWA-HAP-2**) and (**GWA-HAP-CLUSTER**). We used FAMHAP for computing single-marker and haplotype *P*-values and UNPHASED⁷⁵ (option -model allelemain -window 2) for computing the UMMA *P*-values (2 d.f.). Furthermore, we used the BEAGLE⁷⁵ software for haplotype clustering followed by association testing. BEAGLE considers all allelic tests and a range of important clusters (defined by the software) of haplotypes with various lengths.

Table A1 Number of tests for different panels and testing strategies

Marker panel	Number of tests ^a			Empirical number of tests ^a			Genome-wide adjustment factors		
	GWA-1	GWA-UMMA-2	GWA-HAP-2	GWA-1	GWA-UMMA-2	GWA-HAP-2	GWA-1	GWA-UMMA-2	GWA-HAP-2
A (300 K)	4	10	10	3.6	5.4	5.4	229 677	349 677	349 677
B (500 K)	7	28	28	5.6	12.0	12.0	359 483	773 160	773 160
C (1000 K)	15	120	120	10.7	35.5	35.5	691 934	2291 608	2291 608
Full Sequence	62	1953	1953	20.5	312.5	312.5	1319 997	20159 960	20159 960

^aPer marker window.

To obtain genome-wide corrected *P*-values, the *minP* values were multiplied with an empirically determined number of independent tests. The respective numbers depend on the SNP panels, and the analysis strategy and can be found in Table A1. The values were computed as follows: we first assumed that the human genome is separated into distinct marker windows (blocks) and that potential associations due to LD extending outside the marker windows can be ignored for the power calculation. In a recent publication,⁷⁶ we have shown how *minP* adjusted *P*-values can be computed using an application of the algorithm by Ge *et al.*⁷⁷ This Monte Carlo (MC) simulation-based method leads to corrected *P*-values, which account for both multiple testing and the dependency of the tests due to LD. In practice, the applicability of this method is restricted to windows of no more than 62 markers in LD. We were interested in genome-wide significance and, hence, small unadjusted *P*-values. Therefore, it was not computationally feasible to use the MC-approach for the complete simulation study (all marker windows), because small *P*-values require a large amount of permutations for exact computation. Instead, we evaluated only 10% of the data sets using the MC-approach. We used

10⁸ permutation replicates for each data set. For each of these data sets, we obtained the empirical number of independent tests per window by dividing the adjusted *minP* value by *minP*, a number typically smaller than the number of tests conducted within each window. Next, we took the average (for each panel and analysis strategy) of empirical adjustment factors computed in this way and applied this average to the *minP* values of the remaining 90% of simulated data sets. Then, we multiplied the adjusted *minP* values by the number of marker windows in the genome. Computing $4\,000\,000/62 = 64\,516$ yields the number of marker windows, if we assume 4 000 000 million SNPs to be present in the whole genome. Note that the empirical number of independent tests for a complete marker panel of 1 320 000 found by us comes rather close to the number of 1 000 000 independent regions in the genome postulated recently.⁷⁸ Finally, genome-wide power was calculated as the portion of windows with corrected *minP* reaching genome-wide significance. Adjustment factors for the cluster approach were computed analogously. For this approach, we used the permutational *P*-values provided by BEAGLE⁷⁸ for a portion of the windows and scaled up to all windows in the same way as before.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)