Feasible and Successful: Genome-Wide Interaction Analysis Involving All 1.9×10^{11} Pair-Wise Interaction Tests

Michael Steffens^a Tim Becker^a Thomas Sander^b Rolf Fimmers^a Christine Herold^a
Daniela A. Holler^a Costin Leu^b Stefan Herms^c Sven Cichon^{c, e} Bastian Bohn^d
Thomas Gerstner^d Michael Griebel^d Markus M. Nöthen^{e, f} Thomas F. Wienker^a
Max P. Baur^{a, f}

^aInstitute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, ^bCologne Center for Genomics, University of Cologne, Cologne, ^cLife & Brain Center, University of Bonn, Bonn, ^dInstitute for Numerical Simulation, University of Bonn, Bonn, ^eInstitute for Human Genetics, University of Bonn, Bonn, and ^fGerman Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Key Words

Genome-Wide Interaction Analysis (GWIA) • 3-way interaction • Log-linear model • Idiopathic generalized epilepsy

Abstract

The Genome-Wide Association Study (GWAS) is the study design of choice for detecting common genetic risk factors for multifactorial diseases. The performance of full Genome-Wide Interaction Analyses (GWIA) has always been considered computationally challenging. Two-stage strategies to reduce the amount of numerical analysis require the detection of single marker effects or prior pathophysiological hypotheses before the analysis of interaction. This prevents the detection of pure epistatic effects. Our case-control study in idiopathic generalized epilepsy demonstrates that a full GWIA is feasible through use of data compression, specific data representation, interleaved data organization, and parallelization of the analysis on a multi-processor system. Following extensive quality control of the genotypes, our final list of top interaction hits contains only pairs of interacting

SNPs with negligible marginal effects. The TOP HIT interaction was between a SNP-pair intragenic to gene *DNER* (chr 2) and gene *CTNNA3* (chr 10). Both of these genes are functionally involved in neuronal migration, synaptogenesis, and the formation of neuronal circuits. Our results therefore indicate a possible interaction between these two genes in epileptogenesis. Results from GWAS are beginning to reveal a 'missing heritability' in complex traits and diseases. Systematic, hypothesis-free analysis of epistatic interaction (GWIA) may help to close this increasingly recognized gap in heritability.

Copyright © 2010 S. Karger AG, Basel

Introduction

The Genome-Wide Association Study (GWAS) has become the leading paradigm and study design type for the detection and characterization of common genetic

This paper is dedicated to the memory of Richard S. Spielman.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel
0001-5652/10/0694-0268\$26.00/0

Accessible online at:
www.karger.com/hhe

Max P. Baur, PhD
Institute for Medical Biometry, Informatics, and Epidemiology
University of Bonn
Sigmund-Freud-Strasse 25, DE-53113 Bonn (Germany)
Tel. +49 228 2871 5400, Fax +49 228 2871 5032, E-Mail max.baur@ukb.uni-bonn.de

components of multifactorial traits. The NCBI dbSNP database, version 129, based on NCBI genome build 36.3 (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi) currently contains more than 14.7×10^6 SNPs, of which 6.6×10^6 SNPs are validated and unambiguously localized in the human genome. The SNP panels used in GWAS are selected according to population genetic (genome coverage, haplotype block structure, HapMap) and technical constraints (Illumina, Affymetrix) and usually consist of 0.5 to 1.0×10^6 SNPs.

The large sample sizes required for adequate statistical power can be achieved with the standard case-control design. Since up to 1.0×10^6 genotypes are tested per individual, at least the same number of genotype-based association tests must be performed in the resulting two-dimensional contingency tables. The list of replicated significant findings is ever increasing, but the effect size of single SNPs is usually low, and a large fraction of the genetic etiology remains unexplained.

Analysis of SNP $\times$ SNP interaction on the trait is the next step. The maximal number of genotypes per individual is 1.0×10^6 , and up to 5.0×10^{11} three-dimensional $I \times J \times K$ contingency tables (dimension 1: SNP1 ($I = 3$ genotypes); dimension 2: SNP2 ($J = 3$ genotypes); dimension 3: status ($K = 2$ cases/controls)) must therefore be extracted from the original data and tested for third order interaction. This involves enormous computational effort and problems with data quality.

We present the results of a hypothesis-free, full Genome-Wide Interaction Analysis (GWIA) from a case-control study of idiopathic generalized epilepsy (IGE). This is a common multifactorial neurological disease with a lifetime prevalence of around 0.2% and a heritability of approximately 80%. Despite extensive research, the genetic variants predisposing to common IGE syndromes remain elusive [1, 2]. As well as demonstrating computational feasibility, the GWIA of this dataset has generated biologically relevant results.

Methods

Overview

A sample of 1,601 German epilepsy patients and controls (for a detailed description of the data structure see Methods – Data Structure) was genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0. Application of stringent quality control criteria (see Methods – Data Quality) at the single marker level led to the exclusion of 42 individuals ($1,601 \rightarrow 1,559$; 2.6% reduction) and 285,874 markers ($906,600 \rightarrow 620,726$; 31.5% reduction). A sample of 623 German patients with common IGE syndromes was

Fig. 1. Three 2-way associations and one 3-way association between disease status and two SNPs.

compared to 936 German controls to test for the epistatic effects of two SNPs on disease status. An epistatic effect of this type is equivalent to a three-way interaction between disease status and the two markers under investigation. Figure 1 demonstrates the four possible effects between disease status and two SNPs in a complete interaction model.

Association 1 is the marginal effect of SNP1 on disease status. Association 2 is the marginal effect of SNP2 on disease status. Association 3 is the possible allelic (or genotypic) association between SNP1 and SNP2 caused, for example, by linkage disequilibrium (LD).

In order to perform the complete interaction analysis on the remaining 620,726 autosomal SNPs that had passed quality control, it was necessary to consider all possible combinations of any two SNPs. It was therefore necessary to extract a total of $620,726 \times 620,726 / 2 \approx 1.9 \times 10^{11}$ contingency tables from the original data for analysis. These are $2 \times 3 \times 3$ contingency tables since genotype per marker information was available. Since the main goal of the analysis was to test for an epistatic effect, the selected strategy only tested for the presence of the three-way interaction term in a log-linear model, as described in detail below (see Methods – Statistical Methods). Each interaction test resulted in a p-value under the null-hypothesis of no three-way interaction (no epistatic effect of the two SNPs on disease status). A description of the algorithms used, the internal data representation, and the highly parallelized procedure implemented on a system of 256 processors is provided below (see Methods – Parallel Processing, Data Representation, Implementation).

Data Structure

The study used a classical case-control design. A total of 1,559 DNA samples and 620,726 markers were used in the interaction analysis. All of these had been subjected to extensive quality control.

Table 1. Detailed overview about the quality control steps to define the marker set for the interaction analysis; given are the numbers of markers per cohort which were omitted by the different quality criteria

Quality criteria	Sample		
	EPILEPSY	control 1	control 2
Total number of markers	906,600	906,600	906,600
Non-autosomal markers	37,376	37,376	37,376
Autosomal markers	869,224	869,224	869,224
Call rate (<0.95)	26,224	12,421	70,548
Minor allele frequency (<0.01)	135,551	138,200	137,735
Hardy-Weinberg equilibrium (p value < 0.001)	17,061	7,861	26,740
Number of markers left passing quality filters per cohort	704,102	716,704	661,582
Markers with a p value < 0.001 in the allele frequency difference test between the two control cohorts		13,976	
Intersection of markers	620,726	620,726	620,726

The epilepsy sample included 623 German patients with IGE (244 males, 379 females). A description of the sampling procedures and diagnostic criteria is available at: http://portal.ccg.uni-koeln.de/ccg/research/epilepsy_genetics/sampling_procedure.

A reference sample of 936 German population controls was provided by the biobanks KORA-gen: http://epi.gsf.de/koragen/index_e.php (243 males, 244 females; Control 1); and PopGen: <http://www.popgen.de/> (233 males, 216 females; Control 2). Both control cohorts were genotyped within the NGFN GWAS-Add-on project and have been shown to be suitable control samples.

The control sample and the largest section of the epilepsy sample (535 out of 623 individuals) were genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0 at the genotyping facilities of Affymetrix Inc. in South San Francisco, Calif., USA. A smaller section of the epilepsy sample was genotyped at the RZPD (German Science Center for Genome Research) in Berlin, Germany.

DNA samples from all individuals were taken after informed consent from study participants and approval by institutional review boards. All individuals were of self-reported German ancestry. Population stratification is known to be modest in the German population [30]. Typical F_{ST} -values for central European populations range between 0.0010 (German – Swiss) to 0.0022 (German – English) [43]. According to the approximation formula [44] for the inflation factor λ these F_{ST} -values would result in rather bigger inflation factors than the one we observed ($\lambda = 1.048$) in the distribution of the Armitage trend test statistic between cases and controls over all markers. Thus, we do not see evidence for possible bias by population stratification in our data set.

Data Quality

QC-analysis was performed separately for each cohort and was used to evaluate the genotyping quality of each individual.

The following individuals were excluded from the analysis a priori: those with a call rate <95%, those who shared a genome-wide identical-by-state (IBS) level >1.6 with another individual from the same cohort, and those with a heterozygosity level >4 standard deviations from the mean of their respective cohort. Following application of these QC-criteria, 1,559 of the original 1,601 DNA samples remained. Of the 906,600 markers on the Affymetrix Genome-Wide Human SNP Array 6.0, a total of 620,726 autosomal makers were used for the interaction analysis. These markers fulfilled the following QC-criteria in each cohort: marker call rate > 95%, minor allele frequency > 0.01 and Hardy-Weinberg equilibrium $p < 0.001$. In order to account for technical artefacts, all SNPs with a p value < 0.001 in the allele frequency difference test between the two control cohorts Control 1 and Control 2 were removed. Table 1 summarizes the sequential procedure of eliminating markers from the analysis.

The elimination of 1/3 of all markers during quality control is not unusual for the Affymetrix Genome-Wide Human SNP Array 6.0 [45].

The cluster plots of all markers (i) in the single-marker analysis with $p < 10^{-4}$ in the allele frequency difference test and (ii) in the TOP 100 list of the interaction analysis were visually inspected by at least two independent experienced data analysts for each cohort.

Statistical Methods

The $3 \times 3 \times 2$ contingency table presented in figure 2 with observed genotype counts x_{ijk} ($i = 1, \dots, 3$ – genotypes SNP1; $j = 1, \dots, 3$ – genotypes SNP2; $k = 1, 2$ – cases/controls) was extracted from the quality controlled data for each possible SNP pair.

These observations were used to fit log-linear models with expected cell counts m_{ijk} and

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

In this equation the u-terms correspond to the following items:

$$\begin{array}{l} \log(m_{ijk}) = \\ u \quad + \\ \text{overall mean} \\ u_{1(i)} \quad + \quad u_{2(j)} \quad + \quad u_{3(k)} \quad + \\ \text{genotype} \quad \text{genotype} \quad \text{frequencies} \\ \text{frequencies} \quad \text{frequencies} \quad \text{cases/controls} \\ \text{SNP1} \quad \text{SNP2} \\ u_{12(ij)} \quad + \quad u_{13(ik)} \quad + \quad u_{23(jk)} \quad + \\ \text{genotypic} \quad \text{marginal} \quad \text{marginal} \\ \text{association} \quad \text{effect SNP1} \quad \text{effect SNP2} \\ \text{LD SNP1 – SNP2} \\ u_{123(ijk)} \\ \text{epistatic effect} \\ \text{SNP1 – SNP2 – status} \end{array}$$

In order to reduce the computational effort, only the presence of a 3-way interaction was tested for:

$$H_0: u_{123(ijk)} = 0 \quad \text{versus} \quad H_0: u_{123(ijk)} \neq 0$$

		SNP2		
		CC	CD	DD
SNP1	Cases			
	AA			
	AB		X_{ijk}	
	BB			

		SNP2		
		CC	CD	DD
SNP1	Controls			
	AA			
	AB		X_{ijk}	
	BB			

Fig. 2. Three-dimensional contingency table of observed genotype counts x_{ijk} amongst cases and controls.

In order to test this hypothesis, $H_0: u_{123(ijk)} = 0$, the expectations m_{ijk} were estimated using the iterative proportional fitting algorithm [47, 48]. With these maximum likelihood estimates for the cell counts under the null hypothesis, we obtained the test statistic

$$G^2 = -2 \sum x_{ijk} \log \frac{\hat{m}_{ijk}}{x_{ijk}}$$

which is χ^2 -distributed with $(I - 1)(J - 1)(K - 1) = 4$ degrees of freedom. The comparisons were performed using likelihood ratio testing (as well as χ^2 goodness of fit), resulting in a two-sided p value for each test.

For a more detailed presentation of the framework of log-linear models for the analysis of multidimensional contingency tables we refer to the book of Y.M. Bishop 'Discrete Multivariate Analysis' [32] (linear modelling – p 33; iterative proportional fitting – p 83; degrees of freedom – p 115; test statistic – p 125). The alternative procedure to apply the framework of logistic regression [34] yields p values, which are highly correlated and in most cases nearly identical to those of the log-linear procedure, as is shown in [35]. The log-linear test is much faster to compute, though, and therefore more suitable for genome-wide application.

For single marker analysis, a p value cut-off of $p = 0.5 \times 10^{-8}$, which is based on the assumption of 1 million effectively independent SNPs in the human genome, has become a standard [36]. A respective standard for GWIA does not exist by now. Given the number of tests performed, Bonferroni correction leads to a critical level of 2.6×10^{-13} in order to confirm genome-wide significance at a 0.05-level. This level is definitely too conservative, because of the LD of involved SNPs. Based on the argumentation of Gao et al. [46] for the effective number of independent tests, M_{eff} , the significance level would be reduced to approximately 1.0×10^{-12} , but we cannot provide a precise cut-off that can be applied, since, for instance, Monte-Carlo simulations for such a GWIA were not feasible. Consequently, we are inclined to measure the importance of a result not only by the strict adherence to the conservatively corrected significance level, but also on the basis of its biological and etiological context (see below).

Fig. 3. Observed and expected frequencies of the test statistics ($df = 4$) for all 1.9×10^{11} interaction tests.

Parallel Processing, Data Representation, Implementation

Special care must be taken when processing the SNP data to restrict computing time to a manageable level. Major bottlenecks in modern computers are memory and disk access, and both should be minimized. Therefore, the genome data was extracted from the original database in a pre-processing step and converted to binary form using 2 bits for each marker (00, 01 and 10 for the three genotypes and 11 for a missing value). The marker genotypes for each individual were sorted in the same order as in the original database. SNP-IDs or individual IDs are not part of the binary data since they can be recovered later from their position.

Table 2. Confirmed 17 hits following post-analysis quality control; SNP annotation information is based on NCBI Build 36.3

RANK	P3-way-assoc.	SNP1	Chr	Pos	P _{assoc. 1}	SNP2	Chr	Pos	P _{assoc. 2}
1	5.01E-12	rs6743564	2	230176862	0.01	rs10740257	10	68175964	0.59
2	5.94E-12	rs4383597	4	148225110	0.53	rs7233060	18	75493367	0.77
3	8.24E-12	rs1979957	4	148231771	0.54	rs7233060	18	75493367	0.77
4	1.14E-11	rs4835375	4	148231514	0.55	rs7233060	18	75493367	0.77
5	1.68E-11	rs6743564	2	230176862	0.01	rs10822908	10	68175936	0.60
6	2.04E-11	rs4383597	4	148225110	0.53	rs7231598	18	75491305	0.71
7	2.63E-11	rs1979957	4	148231771	0.54	rs7231598	18	75491305	0.71
8	3.61E-11	rs4835375	4	148231514	0.55	rs7231598	18	75491305	0.71
9	4.68E-11	rs1979956	4	148231926	0.56	rs7233060	18	75493367	0.77
10	1.02E-10	rs17006590	2	28243350	0.72	rs341230	18	6385853	0.59
11	1.03E-10	rs12707028	7	131547941	0.77	rs10747028	9	80002924	0.14
12	1.28E-10	rs1979956	4	148231926	0.56	rs7231598	18	75491305	0.71
13	1.90E-10	rs3773384	3	391441	0.42	rs2079	9	6156653	0.46
14	1.98E-10	rs401487	7	107530645	0.54	rs2331957	14	68803738	0.001
15	2.04E-10	rs9661809	1	55243887	0.28	rs2904259	4	90104737	0.52
16	3.13E-10	rs17006590	2	28243350	0.72	rs341220	18	6383167	0.73
17	5.15E-10	rs6700403	1	66508189	0.04	rs787646	10	95122053	0.40

The raw binary stream thus consists of $1,559 \times 620,726 \times 2$ bits = 242 Mbytes, which can be easily stored in main memory.

The binary data was interleaved, meaning that the data stream starts with the first marker for all individuals followed by the second marker for all individuals and so on (in databases, the data is typically organized the other way around). Consequently, the computation of all contingency tables in a single marker analysis requires one single pass over the whole data stream. Such a computation typically takes only a few seconds on a modern computer.

A SNP $\times$ SNP interaction analysis (GWIA) needs to pass 620,726 times over the whole binary data set, whereby each pass results in a set of contingency tables with one marker being fixed and the other marker variable. Such an analysis would take several weeks if performed on a single computer, and so the computational burden is parallelized and the analysis performed on a multi-processor system. In order to optimize the process, the full binary stream is stored in the main memory associated to each processor. Suppose there are N processors and M markers. Each processor then computes a subset of the contingency tables where the first marker varies in a batch of size M/N while the second marker is fully variable. In this way, computation time can be reduced by a factor of N , and a complete two marker interaction analysis (GWIA) required only seven hours on our parallel system of 256 processors.

All calculations were carried out on the Himalaya cluster computer of the Institute for Numerical Simulation at the University of Bonn. It consists of 128 nodes (Dell PowerEdge 1850) with two CPUs each (Intel Xeon EM 64 T 3.2 GHz) and 4 GB memory per node. The nodes are connected by a Myrinet Clos 256 Switch. The operating system is Ubuntu 8.04 LTS, Kernel 2.6.24 (64 bit), the Message Passing Interface is MPICH-MX. Himalaya achieves 1638,4 GFlop/s peak performance and achieves 1269 GFlop/s in the Linpack benchmark test.

Results

Statistical Results and Genetic Interpretation

Since it is impossible (and meaningless) to store all 1.9×10^{11} contingency tables and their corresponding p values, a threshold for the p values was empirically determined from a learning sample. Tables with p values of less than 4.33×10^{-8} (according to a test statistic larger than 40 for the tested log-linear model with 4 degrees of freedom) were stored for detailed inspection. Others were discarded along with their corresponding test statistics and only used to establish the complete distribution of all 1.9×10^{11} results. The distribution of these test statistics is shown in figure 3.

The p values for the top 100 results ranged between 7.7×10^{-23} to 5.5×10^{-10} after the analysis of the genotypes that had undergone the initial quality control (see Appendix A). More SNP pairs in close genomic proximity (less than 100 KB apart) were found among the top 100 hits than would have been expected by chance. Eleven of the top fifteen hits were defined by two SNPs in LD. Closer scrutiny of these results revealed that these were caused by a failure to capture the LD structure among cases as a result of batch effects. The typical error was the appearance in cases (frequency 0.02 to 0.05) of a rare haplotype that did not exist in controls. It was established that all of these haplotypes originated from a batch of 88 independently genotyped and called cases. Further inspection of

Fig. 4. Genotype-combination frequencies in cases (red in the online version, light grey in the printed version) and controls (blue in the online version, dark grey in the printed version) for SNPs rs6743564 (chr 2) and rs10740257 (chr 10).

Table 3. Raw genotype counts among cases and controls and odds ratios of single genotype constellations of TOP HIT 1

SNP1	SNP2	Cases	Contr.	Freq.Ca.	Freq.Co.	OR	LCL	RCL	p one pair vs. all others
GG	CC	43.0	109.0	0.069	0.117	0.561	0.39	0.81	1.92E-03
GG	CT	55.0	65.0	0.088	0.070	1.295	0.89	1.88	1.75E-01
GG	TT	4.0	14.0	0.006	0.015	0.425	0.14	1.30	1.21E-01
GT	CC	197.0	222.0	0.317	0.238	1.485	1.18	1.86	6.03E-04
GT	CT	95.0	207.0	0.153	0.222	0.632	0.48	0.83	7.36E-04
GT	TT	13.0	43.0	0.021	0.046	0.442	0.24	0.83	9.01E-03
TT	CC	111.0	175.0	0.178	0.188	0.941	0.72	1.22	6.50E-01
TT	CT	76.0	92.0	0.122	0.099	1.272	0.92	1.76	1.42E-01
TT	TT	28.0	6.0	0.045	0.006	7.283	3.00	17.69	3.45E-07

Increased odds (bold) are compensated by decreased odds (italics) contributing to the overall interaction effect.

the respective cluster plots revealed that these effects had arisen due to the erroneous genotype calling of a substantial number of homozygotes as heterozygotes (see Appendix B). The haplotype frequency estimates used for quality control were determined using the expectation-maximization algorithm, as implemented in the case-control tool box of FAMHAP [31].

TOP List of 3-Way Interactions

Experienced data analysts systematically inspected all cluster plots for the top 100 hits. Batch effects were circumvented by excluding the batch of 88 individuals and reanalyzing the data. On the basis of post-analysis quality control, 17 of the 100 top hits (see table 2) were confirmed by the independent judgement of four scientists.

None of the remaining top hits was defined by a SNP-pair in LD.

TOP HIT Interaction 1

The lowest p value ($p = 5.0 \times 10^{-12}$) of table 2 was found for the SNP pair: rs6743564 (chr 2, pos 230176862) and rs10740257 (chr 10, pos 68175964). These SNPs did not show strong marginal effects ($p = 0.01$ and $p = 0.59$). The interaction mainly resulted from an OR = 7.28 (95% CI: 3.0–17.7) of the double homozygous state (SNP1: TT – SNP2: TT). This was observed with a frequency of 4.5% in cases, and 0.6% in controls (see fig. 4).

The corresponding contingency table for TOP HIT 1 and all genotype constellations contributing to the overall interaction result are shown in table 3. The overall interaction effect ($p = 5.0 \times 10^{-12}$) did not withstand conservative Bonferroni-correction. Nevertheless, we consider this result to be of high importance, because of the biological and etiological context of the SNPs involved (see argumentation above).

Two SNPs in close LD to rs6743564, and one SNP in close LD to rs10740257 yielded interaction results that were among the top 200 results and passed the rigorous post-analysis quality control procedures (see fig. 5 below).

TOP HIT Interaction 2

The second best hit (rs4383597, chr 4, pos 148225110, $p = 0.54$; rs7233060, chr 18, pos 75493367, $p = 0.78$; interaction $p = 5.9 \times 10^{-12}$) was for a series of hits (see table 2) of SNP pairs in which one SNP was from an LD region on chr 4 (around bp 148225110) and the other was from an LD region on chr 18 (around bp 75493367). For details, see figure 5. The presence of several hits can be explained by the correlation due to LD in the two involved regions, and genotyping error can be excluded as a potential cause of this finding.

The formal genetic and statistical analyses and stringent quality control process produced 3-way-association results for SNP pairs with small to negligible marginal effects (see table 2). These effects would not have been detected by the use of an interaction analysis that was restricted to pre-selected pairs of SNPs with strong marginal effects. This proves the necessity of the use of a full GWIA. The functional plausibility analysis of the statistical interactions detected in our study is presented below.

Biological and Etiological Interpretation – Functional Plausibility of the Identified SNP Pairs

TOP HIT Interaction 1

Gene *DNER*, Chromosome 2, Gene *CTNNA3*, Chromosome 10. Interacting SNP pairs from gene *DNER* and gene *CTNNA3*: rs6743564 (chr 2, pos 230176862) – rs10740257 (chr 10, pos 68175964), $p = 5.0 \times 10^{-12}$; rs6743564 (chr 2, pos 230176862) – rs10822908 (chr 10, pos 68175936), $p = 1.7 \times 10^{-11}$; rs6751880 (chr 2, pos 230171864) – rs10740257 (chr 10, pos 68175964), $p = 1.1 \times 10^{-9}$. The statistical results for these SNP pairs are presented above.

Genomic Plausibility. These two interacting positions (fig. 5) consist of three SNP pairs that constitute two LD blocks. These are intragenic to the genes encoding the Delta/Notch-like epidermal growth factor (EGF)-related receptor (*DNER*) and catenin alpha 3 (*CTNNA3*). SNP pair 1 is the top hit of table 2, and the other two pairs were among the TOP 100 list (see Appendix A). In addition, rs10933305 and rs10740257 were found in the TOP 200 list. All SNPs displayed in figure 5 are contained on Affymetrix Genome-Wide Human SNP Array 6.0 and all passed quality control procedures. The LD structures of the respective regions on chr 2 and chr 10 are given in figures 6a and b.

Functional Plausibility. Both *DNER* and *CTNNA3* are involved in neuronal migration, synaptogenesis and the formation of neuronal circuits [3, 4]. Slight migrational disturbances (microdysgenesis) are frequently observed in the mesiofrontal cortical brain structures of patients with IGE [5]. These neurodevelopmental abnormalities may impair the cerebral regulation of excitatory and inhibitory circuits, leading to cortical hyperexcitability and epilepsy. *DNER* activates the Notch signaling pathway which promotes glial differentiation, a process that is essential for the morphological and functional maturation of neuronal circuits [4, 5]. *CTNNA3* is expressed in the brain, binds to β -catenin and is a negative regulator of the WNT signaling pathway that plays a role in the pathogenesis of human epilepsy [3]. Dysfunction of β -catenin-mediated signaling pathways in mice leads to cortical malformation and an increased susceptibility to seizures [7]. The *PCDH19* gene, which encodes protocadherin 19, has recently been identified [8] as the causative gene for a type of epilepsy and mental retardation that manifests in females only (EFMR). Both genes are therefore plausible candidate genes that may contribute synergistically to epileptogenesis via Notch- and WNT-signaling pathways.

5

Fig. 5. TOP HIT 1 – interacting SNP pairs intragenic to *DNER* (chr 2) and *CTNNA3* (chr 10).
Fig. 6. a LD-pattern of *DNER* on chr 2 (see fig. 5); **b** LD-pattern of *CTNNA3* on chr10 (see fig. 5).

6

7

Fig. 7. TOP HIT 2 – series of interacting SNP pairs on chromosomes 4 and 18.
Fig. 8. a LD-pattern of genomic region on chr 4 involved in TOP HIT 2 (see fig. 7);
b LD-pattern of genomic region on chr 18 involved in TOP HIT 2 (see fig. 7).

8

Fig. 9. Functionally plausible interaction of SNP pairs intragenic to *MGC33657* (alias *PCDPI*, alias *hCG_17324*; chr 2) and *KCNK9* (chr 8).

Fig. 10. a LD-pattern of *MGC33657* on chr 2 (see fig. 8); **b** LD-pattern of genomic region downstream of *KCNK9* on chr 8 (see fig. 8).

TOP HIT Interaction 2

Interacting SNP pairs from two LD blocks on chr 4 and chr 18 (see fig. 7, 8a, b): rs4383597 (chr 4, pos 148225110) – rs7233060 (chr 18, pos 75493367), $p = 5.9 \times 10^{-12}$; rs1979957 (chr 4, pos 148231771) – rs7233060 (chr 18, pos 75493367), $p = 8.2 \times 10^{-12}$; rs4835375 (chr 4, pos 148231514) – rs7233060 (chr 18, pos 75493367), $p = 1.1 \times 10^{-11}$; rs4383597 (chr 4, pos 148225110) – rs7231598 (chr 18, pos 75491305), $p = 2.0 \times 10^{-11}$; rs1979957 (chr 4, pos 148231771) – rs7231598 (chr 18, pos 75491305), $p = 2.6 \times 10^{-11}$; rs4835375 (chr 4, pos 148231514) – rs7231598 (chr 18, pos 75491305), $p = 3.6 \times 10^{-11}$; rs1979956 (chr 4, pos 148231926) – rs7233060 (chr 18, pos 75493367), $p = 4.7 \times 10^{-11}$; rs1979956 (chr 4, pos 148231926) – rs7231598 (chr 18, pos 75491305), $p = 1.3 \times 10^{-10}$.

It was not possible to interpret the second best hit of table 2 and the corresponding series of hits on chromosomes 4 and 18 (see above) in any functional context since the associated markers are in regions that are devoid of any known genes. The consistent pattern of interaction between the SNPs of an LD region on chromosome 4 and an LD region on chromosome 18 is nevertheless intriguing (see fig. 7). The finding of associations in gene deserts is not uncommon in GWAS of complex diseases, and additional experiments are required to identify the genes involved.

Plausible Interaction

The following functionally plausible interaction result from the TOP 100 list involves an interaction between one SNP on chromosome 2 and a group of closely linked SNPs on chromosome 8. All of these SNPs passed post-analysis quality control procedures.

Gene MGC33657, Chromosome 2, Gene KCNK9, Chromosome 8. Interacting marker pairs from gene *MGC33657* (alias *PCDPL*, alias *hCG_17324*) and gene *KCNK9* (see fig. 9, 10a, b): rs12373810 (chr 2, pos 120073559) – rs2061726 (chr 8, pos 140447771), $p = 1.2 \times 10^{-9}$; rs12373810 (chr 2, pos 120073559) – rs1542525 (chr 8, pos 140461982), $p = 7.6 \times 10^{-10}$; rs12373810 (chr 2, pos 120073559) – rs1382025 (chr 8, pos 140463010), $p = 5.6 \times 10^{-10}$; rs12373810 (chr 2, pos 120073559) – rs16910562 (chr 8, pos 140463200), $p = 5.7 \times 10^{-10}$; rs12373810 (chr 2, pos 120073559) – rs16910563 (chr 8, pos 140463957), $p = 6.7 \times 10^{-10}$; rs12373810 (chr 2, pos 120073559) – rs7011410 (chr 8, pos 140466188), $p = 8.3 \times 10^{-10}$; rs12373810 (chr 2, pos 120073559) – rs12541406 (chr 8, pos 140479425), $p = 1.2 \times 10^{-9}$.

Genomic Plausibility. The interaction involves one intragenic SNP of the *MGC33657* gene on chromosome

2q14.2 and seven SNPs on a single haplotype block on chromosome 8q24.3 which is located around 200 kb downstream of the *KCNK9* gene.

Functional Plausibility. Most of the known genes for human idiopathic epilepsies encode voltage-gated or ligand-gated ion channels [2]. The *KCNK9* gene, which encodes the voltage-gate potassium channel TASK3 (Twik-like acid-sensitive potassium channel 3), is a high-ranking candidate gene because it is strongly expressed in the brain and it is located in chromosomal region 8q24.3, for which positive linkage to human absence epilepsy has been reported [9] (max. lod score of 4.10 at D8S534, located at 137100k – NCBI genome build 36.3). *KCNK9* is 4cM and 3.7 Mb distal, and is thus well within the linkage swept radius. The interacting SNP cluster is located at a point 140400k to 140500k distal to D8S534 and is close to *KCNK9*. Sequencing of the *KCNK9* gene in Strasbourg genetic absence epilepsy rats (GAERS), a well established genetic model for absence epilepsy, has revealed the presence of an additional alanine residue in a polyalanine tract within the C-terminal intracellular domain [10]. This additional alanine is absent in inbred non-epileptic control strains. Interestingly, the human and rat *KCNK9* gene show a maternal-specific imprinted expression [11]. This finding corresponds well with the preponderance of maternal inheritance and female excess observed in IGE [12]. *MGC33657* encodes primary ciliary dyskinesia protein 1 which is expressed in brain tissue. Its function has not yet been characterized. However, the gene product of *EFHC1*, a gene recently implicated as a cause of juvenile myoclonic epilepsy (IGE subtype), has been found to have an important function in the motility of cilia in the choroid plexus [13]. These findings support the hypothesis that ‘choroid plexusopathy’ or ‘ciliopathy’ might be involved in epileptogenesis. *MGC33657* would fit well into this potential pathway.

Discussion

‘Where complex genetic traits loom, epistasis is not far behind.’ This is a quote from the landmark review [14] that first brought the issue of gene $\times$ gene interaction (epistasis) to the attention of the research community. Epistasis and genetic interaction are inevitable consequences of the evolutionary process [15, 16, 37].

Full genome-wide pair-wise interaction studies (GWIA) involve $m(m-1)/2$ steps (m number of markers) which include: (a) extraction of the $2 \times 2 \times 3$ data ma-

trix, and (b) calculation of the test statistic. Full GWIA have until recently been considered computationally challenging [17, 38], although only the calculation of the test statistic has been addressed and not the time critical step of matrix extraction. Attempts at computational simplification and/or restrictions of the test space [18, 19] are reliant upon one or both of the marginal main effects being demonstrated in the single-marker analysis (two-stage procedure).

One GWIA has previously been reported. This was presented as a corollary to a GWAS in bipolar disorder, and produced moderate to negative findings of interaction effects [20]. The authors chose a simplified two-locus test (the 3×3 genotype count matrix is collapsed into a 2×2 allele count matrix) applied separately for cases and controls, without attempting to resolve phase relationships in double heterozygotes. A probability of 0.5 was assumed for the two possible phases. The difference of the allelic odds ratio thus obtained was taken as a test for epistasis. This computationally highly efficient procedure probably has less power [21] than the genotype based test, and did not yield satisfactory results.

Genetic models with (essentially) no main effects have been proposed and thoroughly discussed [22–26]. Failure to detect a relevant locus in the first stage of any two-stage procedure that is based on main effects will result in irrecoverable losses [39–42]. This is convincingly demonstrated in a recent comprehensive review [39], in which the authors define five classes of two-locus models out of the vast number of all possible two-locus models. One class displays no epistasis and only main effects at one or both loci in question (a stringent definition of different effect types is given in the explanatory boxes of [39]). Another class is characterized by epistatic effect only and zero marginal main effects. These two extreme classes are not the problem, however. This is also the case for the third class, where the main effects are strong and the epistatic effect is strong as well. In the remaining two classes of models the main marginal effects are small compared to the joint effects, which show up when both loci are taken into account, thus revealing the epistatic effect. As a consequence, the smaller single locus effect(s) could easily be missed, despite non-zero. Clearly, there is low statistical power for these effects. This is even more relevant in the context of genome-wide association analysis where the great number of markers requires adjustment of single-test *p* values in order to maintain a reasonable significance level. Failure to detect a relevant locus in the first stage of the two-stage procedure will result in irrecoverable loss.

After all, it remains to be seen whether pure epistatic models or models close to this type, which are termed models of disordinal interaction [39], are really that rare. They have not been searched for systematically so far, and have been revealed accidentally by other non statistical approaches [49, 50]. In the context of disease causing alleles the down-playing of epistatic effects with no or minor main effects might be particularly misguided, since ‘epistatic shielding’ of disease alleles is one explanation for their persistence within natural populations over long time periods and under varying selective constraints [51].

For the first time, our study has demonstrated the general computational feasibility of a full GWIA conducted without any assumptions regarding main effects in order to reduce computational effort. Compression of the marker information to the level of bit-wise representation made it possible to store the complete set in main memory and thus reduce disk access to a negligible minimum. Interleaved organization of the data stream and parallelization of the analysis on a system of 256 processors allowed us to conduct the analysis within seven hours of computing time.

Insistence upon the extensive quality control of all data is far from pedantic. The pre-analysis control of all included markers and individuals must be standardized. This exponentially reduces the computational burden in a study of this kind by excluding non-informative markers. Post-analysis quality control of results with the smallest *p* values revealed artifacts at the level of batch effects as well as effects of the clustering algorithm in the genotyping process. Meticulous quality control at the pre- and post-analytical level is generally regarded as a prerequisite to the credibility of GWAS results, but is essential in the context of GWIA.

The confirmed 17 hits obtained after post-analysis quality control (see table 2) produced striking statistical and genetic results. We see that none of these hits reached a formal level of significance after Bonferroni correction, but the top hits proved to be very plausible biological candidates. Therefore, it is compulsory to select these for future replication efforts.

All 17 hits are interactions of marker pairs on differing chromosomes with small to negligible marginal effects. None of them would have been detected via the use of a two-stage strategy (see above).

Several of these hits represent the interaction of a marker-haplotype on one chromosome with a SNP or marker-haplotype on a different chromosome (see TOP HIT 1 – chr 2, chr 10; Top HIT 2 – chr 4, chr 18; Plausible

Interaction – chr 2, chr 8). A straightforward interpretation of such a phenomenon between the markers on one haplotype is LD.

It is important to interpret the statistical results in the context of the biological and etiological function of the identified SNP pairs. TOP HIT 1 of a SNP pair on chr 2 and chr 10 (together with two other SNPs in close vicinity) are intragenic to the gene *DNER* and gene *CTNNA3*, which are both functionally involved in neuronal migration, synaptogenesis, and the formation of neuronal circuits. The plausible interaction of one SNP intragenic to gene *MGC33657* (chr 2) and a haplotypic group of SNPs intragenic to gene *KCNK9* (chr 8) also merits further attention. We are aware, however, that functional plausibility should not be mistaken as real functional evidence and experimental studies are needed for substantiation.

These results require independent replication. Such replication studies should compare their complete list of top interactions to those generated in our GWIA in order to identify any which we may have missed. We conclude: (1) that GWIA is feasible; (2) that interaction without marginal effects occurs frequently and is only detectable by a full GWIA, and (3) that the strongest hit identified in our study points to an interaction between gene *DNER* and gene *CTNNA3* in the etiology of common IGE syndromes.

It is now generally accepted that there is ‘ubiquitous epistasis’ in the genotype-phenotype relationships of higher organisms [16]. GWAS has revealed a large number of causative, contributing and susceptibility factors for many quantitative traits and complex diseases, though the effect sizes are usually small. There is increasing recognition of a gap, or lack, of heritability: ‘The case of the missing heritability’ (Brendan Maher in a recent Nature News Feature, Nov. 6th, 2008). An understanding of epistasis and other genetic mechanisms may help to resolve this conundrum, at least in part [27–29].

We have demonstrated that systematic GWIA is feasible, and anticipate that this approach will contribute to the closure of this ‘gap of heritability’.

Acknowledgements

This study was supported by the German Research Foundation DFG (grants: SA 434/4-2 to T.S.; BE 3828/3-1 to T.B.; BE 3828/4-1 to T.B. and M.S.), the German Ministry of Education and Research BMBF (National Genome Research Network, NGFN-2, NGFNplus: EMINet grant no: 01GS08120 to T.S.; FGP grant no: 01GS08134-2 to M.P.B.; MooDS-Net grant no: 01GS08144 to S.C. and M.N.; grant no: 01GS08144-7 to T.F.W.).

We would like to acknowledge the provision of population controls from PopGen (S. Schreiber, M. Krawczak, H. v. Eberstein; University of Kiel) and from KORA-gen (H. E. Wichmann; Helmholtz Center Munich), which were genotyped within the GWAS-Add-on project of the NGFN-2.

Appendix A

Top 100 results (Ranks R1) of the interaction tests sorted according to p value

R1	R2	p3-way-assoc.	SNP1	Chr	Pos	P _{assoc. 1}	SNP2	Chr	Pos	P _{assoc. 2}
1		7.71E-23	rs2393239	12	116540623	0.65	rs1568158	12	116544840	1.08E-04
2		9.17E-19	rs2359009	4	7183266	2.94E-09	rs17751281	4	7186401	0.67
3		1.53E-18	rs2359009	4	7183266	2.94E-09	rs41384948	4	7186028	0.72
4		8.18E-15	rs10162983	15	99110525	1.96E-08	rs881897	18	944176	4.91E-09
5		1.63E-14	rs717124	6	18460242	7.47E-16	rs9507915	13	27113036	3.37E-19
6		6.88E-14	rs1568158	12	116544840	1.08E-04	rs10850880	12	116568028	0.59
7		1.31E-13	rs9811898	3	29475838	2.07E-14	rs9867172	3	29526943	0.86
8		6.62E-13	rs1568158	12	116544840	1.08E-04	rs4766866	12	116567773	0.74
9		1.06E-12	rs10503528	8	14971700	0.08	rs7009607	8	15011837	8.46E-13
10		1.53E-12	rs17120860	8	14979036	0.01	rs7009607	8	15011837	8.46E-13
11		1.96E-12	rs17120908	8	14995614	0.13	rs7009607	8	15011837	8.46E-13
12		2.49E-12	rs1568158	12	116544840	1.08E-04	rs10850875	12	116564223	0.86
13		2.54E-12	rs9811898	3	29475838	2.07E-14	rs7789283	7	130472192	1.49E-09
14		3.56E-12	rs16867394	2	181842539	2.54E-10	rs717124	6	18460242	7.47E-16

(continued)

R1	R2	p _{3-way-assoc.}	SNP1	Chr	Pos	P _{assoc. 1}	SNP2	Chr	Pos	P _{assoc. 2}
15		4.03E-12	rs17120868	8	14979635	0.05	rs7009607	8	15011837	8.46E-13
16	1	5.01E-12	rs6743564	2	230176862	0.01	rs10740257	10	68175964	0.59
17	2	5.94E-12	rs4383597	4	148225110	0.53	rs7233060	18	75493367	0.77
18		7.68E-12	rs16867394	2	181842539	2.54E-10	rs10162983	15	99110525	1.96E-08
19	3	8.24E-12	rs1979957	4	148231771	0.54	rs7233060	18	75493367	0.77
20		8.88E-12	rs10108640	8	14957836	0.12	rs7009607	8	15011837	8.46E-13
21		8.95E-12	rs16867394	2	181842539	2.54E-10	rs9650318	8	39647369	1.38E-08
22		1.10E-11	rs373730	3	60433648	5.46E-10	rs7789283	7	130472192	1.49E-09
23	4	1.14E-11	rs4835375	4	148231514	0.55	rs7233060	18	75493367	0.77
24	5	1.68E-11	rs6743564	2	230176862	0.01	rs10822908	10	68175936	0.6
25		1.81E-11	rs665545	1	75482497	0.03	rs619696	1	75530957	0.04
26		1.87E-11	rs10162983	15	99110525	1.96E-08	rs11088917	21	22604105	2.87E-07
27		1.95E-11	rs13034087	2	140814490	2.13E-06	rs881897	18	944176	4.91E-09
28	6	2.04E-11	rs4383597	4	148225110	0.53	rs7231598	18	75491305	0.71
29	7	2.63E-11	rs1979957	4	148231771	0.54	rs7231598	18	75491305	0.71
30		3.27E-11	rs17628359	5	96301717	2.92E-06	rs881897	18	944176	4.91E-09
31		3.38E-11	rs717124	6	18460242	7.47E-16	rs1568158	12	116544840	1.08E-04
32	8	3.61E-11	rs4835375	4	148231514	0.55	rs7231598	18	75491305	0.71
33		4.61E-11	rs6879346	5	162899365	0.72	rs4934636	10	34826697	0.61
34	9	4.68E-11	rs1979956	4	148231926	0.56	rs7233060	18	75493367	0.77
35		5.32E-11	rs373730	3	60433648	5.46E-10	rs10162983	15	99110525	1.96E-08
36		7.99E-11	rs485414	1	74522030	0.3	rs1620195	12	31976014	0.4
37		8.32E-11	rs4766862	12	116536112	0.46	rs1568158	12	116544840	1.08E-04
38		9.09E-11	rs7009607	8	15011837	8.46E-13	rs881897	18	944176	4.91E-09
39	10	1.02E-10	rs17006590	2	28243350	0.72	rs341230	18	6385853	0.59
40	11	1.03E-10	rs12707028	7	131547941	0.77	rs10747028	9	80002924	0.14
41		1.07E-10	rs4391048	4	12191997	0.41	rs2251589	14	53062680	0.65
42		1.07E-10	rs12635872	3	113560147	0.24	rs6065912	20	44065774	0.54
43		1.19E-10	rs6879346	5	162899365	0.72	rs11009804	10	34814302	0.67
44		1.25E-10	rs6952566	7	131545408	0.12	rs10747028	9	80002924	0.14
45	12	1.28E-10	rs1979956	4	148231926	0.56	rs7231598	18	75491305	0.71
46		1.41E-10	rs11777481	8	67382082	0.49	rs2388497	13	29612792	0.15
47		1.45E-10	rs9811898	3	29475838	2.07E-14	rs9310904	3	29526174	0.43
48		1.45E-10	rs3789435	1	94348948	0.95	rs10512918	5	50329633	0.11
49		1.48E-10	rs4391048	4	12191997	0.41	rs2251294	14	53060318	0.72
50		1.52E-10	rs4637120	2	18565108	0.58	rs4404103	17	34936183	0.29
51		1.70E-10	rs3128499	9	89516915	0.41	rs932426	20	37307227	0.4
52		1.70E-10	rs4391048	4	12191997	0.41	rs4901431	14	53049564	0.71
53		1.72E-10	rs11707269	3	59895749	0.13	rs12510966	4	107203053	0.51
54		1.78E-10	rs373730	3	60433648	5.46E-10	rs717124	6	18460242	7.47E-16
55	13	1.90E-10	rs3773384	3	391441	0.42	rs2079	9	6156653	0.46
56	14	1.98E-10	rs401487	7	107530645	0.54	rs2331957	14	68803738	6.46E-04
57		2.00E-10	rs12512427	4	97161336	0.05	rs1458269	13	53332566	0.63
58	15	2.04E-10	rs9661809	1	55243887	0.28	rs2904259	4	90104737	0.52
59		2.06E-10	rs12512427	4	97161336	0.05	rs9536558	13	53343270	0.59
60		2.40E-10	rs9310236	3	72298865	0.91	rs10836850	11	4841781	0.3
61		2.50E-10	rs777716	2	195663362	0.77	rs12363575	11	2986680	0.96
62		2.50E-10	rs9811898	3	29475838	2.07E-14	rs9883294	3	29557812	0.18
63		2.94E-10	rs717124	6	18460242	7.47E-16	rs4633799	18	3039472	6.71E-08
64		2.99E-10	rs450240	2	178740139	3.29E-09	rs11088917	21	22604105	2.87E-07

(continued)

R1	R2	P _{3-way-assoc.}	SNP1	Chr	Pos	P _{assoc. 1}	SNP2	Chr	Pos	P _{assoc. 2}
65		3.03E-10	rs7789283	7	130472192	1.49E-09	rs7009607	8	15011837	8.46E-13
66	16	3.13E-10	rs17006590	2	28243350	0.72	rs341220	18	6383167	0.73
67		3.16E-10	rs1568158	12	116544840	1.08E-04	rs11088917	21	22604105	2.87E-07
68		3.18E-10	rs3897923	4	148225156	0.86	rs7233060	18	75493367	0.77
69		3.36E-10	rs12441807	15	39973643	0.95	rs12624759	20	50820556	0.27
70		3.42E-10	rs410866	7	107531779	0.53	rs2331957	14	68803738	6.46E-04
71		3.42E-10	rs3789435	1	94348948	0.95	rs4443373	5	50358255	0.09
72		3.42E-10	rs11122218	1	229269465	0.1	rs4871389	8	124650394	0.75
73		3.48E-10	rs10926307	1	238834336	0.13	rs4891291	18	71374601	0.98
74		3.53E-10	rs7566637	2	10340390	0.12	rs17693923	3	127999896	0.05
75		3.55E-10	rs7216104	17	9884404	0.53	rs438220	20	52978808	0.81
76		3.58E-10	rs717124	6	18460242	7.47E-16	rs10162983	15	99110525	1.96E-08
77		3.67E-10	rs12512427	4	97161336	0.05	rs7338269	13	53281478	0.59
78		3.77E-10	rs7789283	7	130472192	1.49E-09	rs881897	18	944176	4.91E-09
79		3.83E-10	rs946934	1	235160241	0.47	rs5761826	22	25632631	0.23
80		3.85E-10	rs6824703	4	73122182	0.37	rs17116165	5	153817596	0.92
81		3.88E-10	rs3789435	1	94348948	0.95	rs17219934	5	50345549	0.12
82		3.91E-10	rs16867394	2	181842539	2.54E-10	rs881897	18	944176	4.91E-09
83		3.93E-10	rs946934	1	235160241	0.47	rs1041916	22	25633755	0.25
84		3.94E-10	rs12564492	1	166501269	0.46	rs1868606	12	96712712	0.69
85		4.01E-10	rs12373810	2	120073559	0.02	rs16910544	8	140453559	0.01
86		4.09E-10	rs10156075	7	131550895	0.3	rs10747028	9	80002924	0.14
87		4.34E-10	rs1339289	9	8226122	0.79	rs6505044	17	51613239	0.41
88		4.37E-10	rs7654531	4	73119485	0.38	rs17116165	5	153817596	0.92
89		4.51E-10	rs7903448	10	9848668	0.06	rs2827523	21	22735992	0.93
90		4.58E-10	rs600390	16	55126429	0.65	rs13043573	20	11513456	0.1
91		4.81E-10	rs6887453	5	162892133	0.59	rs4934636	10	34826697	0.61
92		4.95E-10	rs17132163	7	48253360	0.63	rs4871296	8	123758851	0.64
93		5.05E-10	rs3766217	1	46931696	0.33	rs17053273	8	25179409	0.11
94		5.05E-10	rs14243	2	233780478	0.4	rs16862573	3	188965463	0.15
95		5.07E-10	rs11747295	5	123424571	0.04	rs9560813	13	90931610	0.08
96		5.08E-10	rs11707269	3	59895749	0.13	rs12509640	4	107216126	0.63
97	17	5.15E-10	rs6700403	1	66508189	0.04	rs787646	10	95122053	0.4
98		5.27E-10	rs12954469	18	45384226	0.6	rs3004119	20	23437678	0.96
99		5.29E-10	rs4391048	4	12191997	0.41	rs17253291	14	53032313	0.78
100		5.53E-10	rs831121	5	67590682	0.97	rs17761362	10	112430512	0.11

17 of these 100 hits were confirmed by post-analysis quality control (Ranks R2, shaded rows) and are shown in table 2 in results.

Appendix B

Demonstration of a batch effect in genotype calling for marker rs2359009.

Using the birdseed v1 algorithm, genotype calling was performed in five batches (N_{1-5}) in accordance with experimental and technical considerations. Combining the batches into a control subsample N_{tot} revealed inconsistent genotype clusters. The birdseed clustering algorithm balances different clustering criteria [33] (shape, size, density, uniformity, separation, outliers, etc.) and arrives at different solutions in some of the batches, which, when viewed together, appear inconsistent. However, call rate (0.954), minor allele frequency (0.24) and HWE-statistics ($p = 0.67$) were not outside of quality criteria, and this SNP therefore passed the first basic quality checks and appeared on the top list of interaction results. All SNPs afflicted by this type of problem must be excluded.

Color version available online

References

- Helbig I, et al: Navigating the channels and beyond: unravelling the genetics of the epilepsies. *Lancet Neurol* 2008;7:231–245.
- Reid CA, Berkovic SF, Petrou S: Mechanisms of human inherited epilepsies. *Prog Neurobiol* 2009;87:41–57.
- Ciani L, Salinas PC: WNTs in the vertebrate nervous system: from patterning to neuronal connectivity. *Nat Rev Neurosci* 2005;6:351–362.
- Eiraku M, et al: DNER acts as a neuron-specific Notch ligand during Bergmann glial development. *Nat Neurosci* 2005;8:873–880.
- Meencke HJ, Veith G: The relevance of slight migrational disturbances (microdysgenesis) to the etiology of the epilepsies. *Adv Neurol* 1999;79:123–131.
- Eiraku M, et al: Delta/notch-like epidermal growth factor (EGF)-related receptor, a novel EGF-like repeat-containing protein targeted to dendrites of developing and adult central nervous system neurons. *J Biol Chem* 2002;277:25400–25407.
- Campos VE, Du M, Li Y: Increased seizure susceptibility and cortical malformation in beta-catenin mutant mice. *Biochem Biophys Res Commun* 2004;320:606–614.
- Dibbens LM, et al: X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment. *Nat Genet* 2008;40:776–781.
- Sugimoto Y, et al: Childhood absence epilepsy in 8q24: refinement of candidate region and construction of physical map. *Genomics* 2000;68:264–272.
- Holter J, et al: A TASK3 channel (KCNK9) mutation in a genetic model of absence epilepsy. *J Mol Neurosci* 2005;25:37–51.
- Ruf N, et al: Sequence-based bioinformatic prediction and QUASEP identify genomic imprinting of the KCNK9 potassium channel gene in mouse and human. *Hum Mol Genet* 2007;16:2591–2599.
- Pal DK, et al: Complex inheritance and parent-of-origin effect in juvenile myoclonic epilepsy. *Brain Dev* 2006;28:92–98.
- Suzuki T, et al: Efhcl deficiency causes spontaneous myoclonus and increased seizure susceptibility. *Hum Mol Genet* 2009;18:1099–1109.
- Frankel WN, Schork NJ: Who's afraid of epistasis? *Nat Genet* 1996;14:371–373.
- Wolf J, Brodie E, Wade M: Epistasis and the Evolutionary Process. Oxford, New York, Oxford University Press, 2000
- Phillips PC: Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 2008;9:855–867.
- Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;37:413–417.
- Evans DM, et al: Two-stage two-locus models in genome-wide association. *PLoS Genet* 2006;2:e157.
- Long Q, Zhang Q, Ott J: Detecting disease-associated genotype patterns. *BMC Bioinformatics* 2009;10(suppl 1):S75.
- Ferreira MA, et al: Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 2008;40:1056–1058.
- Schaid DJ: Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 2004;166:505–512.
- Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;11:2463–2468.
- Culverhouse R, et al: A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002;70:461–471.
- Culverhouse R, et al: Gene $\times$ gene and gene $\times$ environment interactions for complex disorders. *BMC Proc* 2007;1(suppl 1):S72.
- Wu X, Jin L, Xiong M: Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet* 2008;16:644–651.
- Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet* 2004;20:640–647.
- Maher B: Personal genomes: The case of the missing heritability. *Nature* 2008;456:18–21.
- Bogardus C: Missing heritability and GWAS utility. *Obesity (Silver Spring)* 2009;17:209–210.
- Goldstein DB: Common genetic variation and human traits. *N Engl J Med* 2009;360:1696–1698.
- Steffens M, et al: SNP-based analysis of genetic substructure in the German population. *Hum Hered* 2006;62:20–29.
- Herold C, Becker T: Genetic association analysis with FAMHAP: a major program update. *Bioinformatics* 2009;25:134–136.
- Bishop Y, Fienberg S: Discrete multivariate analysis. New York, Springer, 2007.
- Xu R, Wunsch D: Survey of clustering algorithms. *IEEE Trans Neural Netw* 2005;16:645–678.
- 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.
- Herold C, et al: INTERSNP: Genome-wide Interaction Analysis Guided by A Priori Information. *Bioinformatics* 2009;25:3275–3281.
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases and seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–683.
- Moore JH, Williams SM: Epistasis and its implication for personal genetics. *Am J Hum Genet* 2009;85:309–320.
- Cordell HJ: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;10:392–404.
- Musani SK, et al: Detection of gene $\times$ gene interactions in genome-wide association studies of human population data. *Hum Hered* 2007;63:67–84.
- Kottli S, Bickeboller H, Clerget-Darpoux F: Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum Hered* 2007;63:85–92.
- Culverhouse R: The use of the restricted partition method with case-control data. *Hum Hered* 2007;63:93–100.
- Tung L, Gordon D, Finch SJ: The impact of genotype misclassification errors on the power to detect a gene-environment interaction using Cox proportional hazard modeling. *Hum Hered* 2007;63:101–110.
- Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton, New Jersey, Princeton University Press, 1994.
- Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- Nishida N, Koike A, Tajima A, Ogasawara Y, Ishibashi Y, Uehara Y, Inoue I, Tokunaga K: Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. *BMC Genomics* 2008;9:431.
- Gao X, Becker LC, Becker DM, Starmer JD, Province MA: Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology* 2010;34:100–105.
- Bartlett MS: Contingency table interactions. *J Roy Statist Soc Suppl* 1935;2:248–252.
- Deming WE, Stephan FF: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann Math Statist* 1940;11:427–444.
- Chen CC, Lu RB, Chen YC, Wang MF, Chang YC, Li TK, Yin SJ: Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. *Am J Hum Genet* 1999;65:795–807.
- Maisnier-Patin S, Andersson DI: Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res in Microbiology* 2004;155:360–369.
- Phillips PC, Johnson NA: The population genetics of synthetic lethals. *Genetics* 1998;150:449–458.