Rockefeller University
Digital Commons @ RU

Krueger Laboratory | Laboratories and Research

2013

From Single-SNP to Wide-Locus: Genome-Wide Association Studies Identifying Functionally Related Genes and Intragenic Regions in Small Sample Studies

Knut M. Wittkowski

Vikas Sonakya

Tingting Song

Martin P. Seybold

Mehdi Keddache

See next page for additional authors

Follow this and additional works at: http://digitalcommons.rockefeller.edu/krueger_laboratory

Part of the Life Sciences Commons

Recommended Citation

Pharmacogenomics (2013), 14 (4), p. 391-401

This Article is brought to you for free and open access by the Laboratories and Research at Digital Commons @ RU. It has been accepted for inclusion in Krueger Laboratory by an authorized administrator of Digital Commons @ RU. For more information, please contact mcsweej@mail.rockefeller.edu.

Authors
Knut M. Wittkowski, Vikas Sonakya, Tingting Song, Martin P. Seybold, Mehdi Keddache, and Martina Durner

This article is available at Digital Commons @ RU: http://digitalcommons.rockefeller.edu/krueger_laboratory/1

From Single-SNP to Wide-Locus: genome-wide association studies identifying functionally related genes and intragenic regions in small sample studies

KNUT M. WITTKOWSKI[1), VIKAS SONAKYA[1), TINGTING SONG[1), MARTIN P. SEYBOLD[2), MEHDI KEDDACHE[3), MARTINA DURNER[4)

1) The Rockefeller University, Center for Clinical and Translational Science
1230 York Ave Box 322, New York, NY 10021, U.S.A. kmw@rockefeller.edu

2) Stuttgart University, Institut für Formale Methoden der Informatik
Universitaetstrasse 38, D-70569 Stuttgart, Germany

3) Cincinnati Children's Hospital Medical Center,
3333 Burnet Avenue, Cincinnati, Ohio 45229-3039

4) Mount Sinai School of Medicine, Department of Psychiatry,
One Gustave Levy Place, Box 1230, New York, NY 10029

Abstract:

Background: Genome Wide Association Studies (GWAS) have had limited success when applied to complex diseases. Analyzing SNPs individually requires several large studies to integrate the often divergent results. In the presence of epistasis, multivariate approaches based on the linear model (including stepwise logistic regression) often have low sensitivity and generate an abundance of artifacts.

Methods: Recent advances in distributed and parallel processing spurred methodological advances in non-parametric statistics. U-statistics for multivariate data (µStat) are not confounded by unrealistic assumptions (linearity, independence).

Results: By incorporating knowledge about relationships between SNPs, µGWAS (GWAS based on µStat) can identify clusters of genes around biologically relevant pathways and pinpoint functionally relevant regions within these genes.

Conclusion: With this computational biostatistics approach increasing power and guarding against artifacts, personalized medicine and comparative effectiveness will advance while subgroup analyses of Phase III trials can now suggest risk factors for adverse events and novel directions for drug development.

Key words:

common disease, epilepsy, epistasis, genome-wide association study, GWAS, phase III trials, Ras-pathway, u-statistics

For full bibliographic citation, please refer to the version available at www.futuremedicine.com'

Introduction

Almost a decade after the completion of the Human Genome Project [1], the scientific and medical advances hoped for from genome-wide association studies (GWAS) have not yet been realized. After early successes with diseases where a single haplotype confers all or most risk [2], the same statistical approaches have often produced ambiguous results when applied to complex diseases [3, 4]. Increasing the sample size (to tens of thousands of subjects as suggested [5]) is impractical for rare disease forms, and also greatly increases the duration and cost of data collection. Improving accrual by broadening the inclusion criteria increases variance and thus requires yet larger samples; a vicious cycle. Moreover, increasing sample size in a nonrandomized study may, somewhat paradoxically, increase the risk of false positives [6, 7].

Several mutations within a gene may contribute to the risk of common diseases and several SNPs may have become associated with the same mutation over time. One risk factor's contribution may depend on the presence of others and sets of mutations may confer more risk if they affect both chromosomes (compound heterozygosity). Hence, any statistical approach based on p-values derived one SNP at a time (ssGWAS) is ill-suited to identify the short-range epistasis involved [8]. (Following Fisher [9], the term 'epistasis' will be used for any deviation from independence, be it between neighboring SNPs, intragenic regions or genes.) Analyzing diplotypes (sets of neighboring SNPs with unknown phase) comprehensively would be preferable [10], yet traditional multivariate methods [11] including linear/logistic regression (lr) assume independence and additivity/multiplicativity of risk factors to yield computationally simple algorithms. Making unrealistic assumptions, such as linearity, may easily lead to meaningful non-linear relationships being overlooked (false negatives). More importantly, random errors, not subject to biological constraints, may occasionally fulfill these assumptions, so that the most 'significant' results are often false positives.

Association studies, in general, are exploratory 'selection procedures' [12] to generate, rather than confirm hypotheses. Even though the same algorithms are used as in confirmatory tests, 'p-values' merely serve to sort candidates, so that a sufficiently large selection of candidate genes will include the most interesting genes with high power. Even minor differences in the composition of the study population can result in different subsets of genes being selected [13], and each could help with understanding a different aspect of the disease etiology when confirmed using mouse studies or clinical trials. Hence, the challenge in improving GWAS is to reduce artifacts caused by applying oversimplifying approaches to complex diseases (analyzing one SNP at a time, assuming independence and additivity of effects) while incorporating more knowledge to increase the sensitivity for detecting biologically relevant subsets of the genes involved.

With the advent of mainframe computers, more complex calculations (e.g., factor analysis) became feasible. More recently, personal computers triggered the development of resampling methods. Now we are, again, entering an era of advances in computational biostatistics, where massive-parallel computing has spurred the methodological advances making wide-locus GWAS based on a nonparametric approach (µGWAS, based on u-statistics for structured multivariate data) feasible [14]. Below, we introduce two novel concepts. First, several 'tag' sets of 'genetically indistinguishable' SNPs [15] are typically scattered across a linkage disequilibrium (LD) block, yet traditional methods cannot differentiate between 'permuted' diplotypes containing members of the same tag sets in different order. µGWAS draws on the spatial structure of SNPs within a diplotype and expected LD from HapMap [16] to improve the resolution of GWAS to intragenic regions. Second, we apply the concept of 'information content of multivariate data' (µIC) [14] at several stages of the analysis to guard against artifacts. With these methodological advances, disease-relevant genes and intragenic regions can now be suggested from a single study, often of only a few hundred narrowly defined cases, rather than from a variety of large

studies, turning GWAS from a technique to identify isolated SNPs into a powerful tool to generate plausible and testable hypotheses about the etiology of complex diseases.

Methods

µ-scores for diplotypes

It is often reasonable to assume that risk conferred by a heterozygous SNP lies somewhere between baseline risk and a homozygous SNP (having two risk alleles) that is, between the risk of a recessive and a dominant allele, respectively. U-statistics (including the Wilcoxon/Mann-Whitney U test [17]) treat SNPs as ordinal (wildtype = xx < xX < XX = homozygous), but do not require the degree of dominance to be known. Treating diplotypes as multivariate data then avoids the need for assumptions about independence and relative importance of the SNPs, yet the theory [18] was never broadly developed owing to prohibitively high computational demand [19]. With GWAS, for instance, the number of ‘polarities’ (combinations of −1 = bad / 0 = irrelevant / +1 = good) increases exponentially with diplotype length, yet with massively parallel computing we were now able to include diplotypes up to length six.

Traditionally, one would have more confidence in a ‘significant’ locus if neighboring loci also show association [20] and add recombination information to the data displayed. Here we integrate the concepts behind this intuitive visual inspection into the statistical approach itself. Recently, µ-scores (U-scores for multivariate data) have been extended to reflect structures among variables with applications including sports [21], policy making [22], and medicine [14]. The proposed GWAS-specific structure is based on the notion that neighboring disease loci may have similar effects and that a disease locus may be in LD with both adjacent SNPs unless the SNPs are separated by a recombination hotspot (boundary between LD blocks) (Figure 1).

µGWAS starts with computing matrices representing the partial order of each SNP, combining pairs of these matrices into matrices representing the intervals, and, finally, combining SNP and interval matrices into a diplotype matrix from which the µ-scores are computed [14, 22]. As diplotype profiles are built from intervals around and between neighboring SNPs, diplotypes where members X_i, Y_i, and Z_i of the tag sets (X), (Y), and (Z) appear in different order (permuted diplotypes), such as (X_1, Y_1, Z_1) versus (Y_1, X_1, Z_1) can be distinguished. This novel approach to incorporate knowledge of neighborhood relationships between SNPs increases power over merely combining all SNPs within a diplotype in a single step [14], yet avoids the need for assumptions about dependencies and relative importance required when using linear combinations (weighted sums) of univariate scores. With GWAS based on lr (lrGWAS), one could work towards a similar goal by adding sequential interaction terms. Hence, we will compare µGWAS not only with ssGWAS for dominant, linear trend [23], and recessive effects, but also with stepwise logistic regression with and without sequential interaction terms.

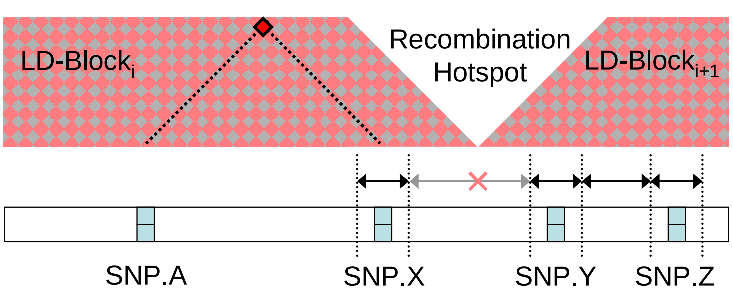


Figure 1: SNP-related chromosomal intervals. Conceptual structure of chromosomal SNP-related intervals for disease loci in LD with three consecutive SNPs (X, Y, and Z), but not with a more distant SNP (A). SNPs X and Y are part of different LD-blocks, separated by a recombination hotspot. Hence, the interval between these two SNPs is excluded. The location indicating LD between SNPs A and X is highlighted. The inter-regional boundaries need not be known. LD: Linkage disequilibrium

Subjects

Childhood Absence Epilepsy (CAE) [24], formerly known as ‘petit mal’, is characterized by frequent, short episodes of ‘day dreaming’. Through trial and error of different combinations of valproic acid and various ion channel blockers, these absences can be controlled in approximately 75% of affected children [25]. For adult patients, etiracetam, an IL-1β inhibitor [26] was approved in November 1999, and a Caspase 1 inhibitor (VRT-765) is undergoing a controlled Phase IIb study [101]. CAE does not follow a simple Mendelian pattern of inheritance, although recurrence of epilepsy in families is high. A high concurrence in monozygotic twins and the absence of known exogenic factors make CAE an ideal model for studying the genetics of complex diseases and approaches to unravel their genetic risk factors to better match patients to existing drugs and identify new drug targets for patients who do not respond to existing drugs.

The 185 CAE patients in this study were predominantly Caucasian (83%) and white Hispanic (10%) with the well-known female preponderance (115 female vs. 70 male patients). Average age of onset for absence seizures was 5.7 years. Patients were required to be seizure free on antiepileptic medication. Controls were selected from a publicly available database [103]. See Supplemental Material at www.futuremedicine.com/doi/duppl/10.2217/pgs.13.28 for details.

Results

Identifying Genes

As is typical for ssGWAS, especially with small samples, only two SNPs reached the customary $s = -\log_{10}(p) > 7.5$ level of significance with univariate tests (Figure 2, black foreground), one in a non-coding region (chromosome 1, lr only), the other in the pseudogene *EE1A1P12* (chromosome 2).

Since ssGWAS was inconclusive and sequential interaction terms created an abundance of likely false positives with lrGWAS (see Supplementary Figure 2), even with regularization (AIC [27]), the following discussion focuses on µGWAS vs. traditional lrGWAS. In the spirit of conducting a selection procedure [12, 28], rather than confirmatory tests, p-values were used solely for the purpose of ranking the loci and. At any given level, lrGWAS had more ‘significant’ results, in general, including many likely false positives. Hence, methods were compared using similar arbitrary numbers of top regions (first comparison used only the top 6, second comparison used ≈20, third comparison used≈40; see Supplementary Table 1), the latter cut-offs adjusted for display purposes (Figure 2) to match commonly used s-values (µ: 7.5/7.0, lr: 8.0/7.5)

Only one of the top six genes in lrGWAS (*RBFOX1*) ranks higher than $r_\mu = 73^{rd}$ in µGWAS (5^{th}), while the other four among the top six regions in µGWAS are also among the top 22 in lrGWAS (the above elongtion factor pseudogene *EEF1A1P12*; *SYN3*, synapsin III; *FAT4*; *CREB5*, Supplementary Table 1). Of the top 17 µGWAS regions ($s > 7.5$), 14 (82%) are known to be in genes directly related to the NOD/axonal guidance signaling/ataxin pathway (Figure 3), including, *PANX1*, *SEC16B, the Rho GTPase activating proteins OPHN1/ARHGAP41*, and *RICS / ARHGAP32*, *ABCC8*, the potassium channel *KCNJ5*, *BRE, NLRP3*, and *RASSF8*, compared to only 8 genes (36%) of the top 22 ($s > 8$), including *KCNB2*, *DOK6*, and *MYO16*, or 16 (40%) of the top 40 lrGWAS ($s > 7.5$) regions.

Figure 2: Extended Manhattan Plot for the Comparison of 185 CAE cases vs matched controls. Unadjusted $-\log_{10}(p)$ by chromosomal location; top: µGWAS, bottom: lrGWAS (without interaction terms). Univariate results, shown in black, are consistently similar across the approaches – as expected. For µGWAS, dots vary in size by diplotype length and are color coded, with red indicating results with low µ-scores for reliability (high significance, low µIC). Lr results are overlayed with Cochran-Armitage (squares) and Mantel-Haenszel (×/+) results. Genes known to be directly related to the NOD-ID/AGS-Ataxin pathway are shown in bold. Genes indicated in the center header row (pink) of each chromosome have support in both µGWAS and lrGWAS; genes ranking higher in µGWAS or lrGWAS appear in the first row (blue) or third row (red), respectively. Darker colors indicate more significant results Other genes are shown with that method implicating them against the dark background of univariate results.

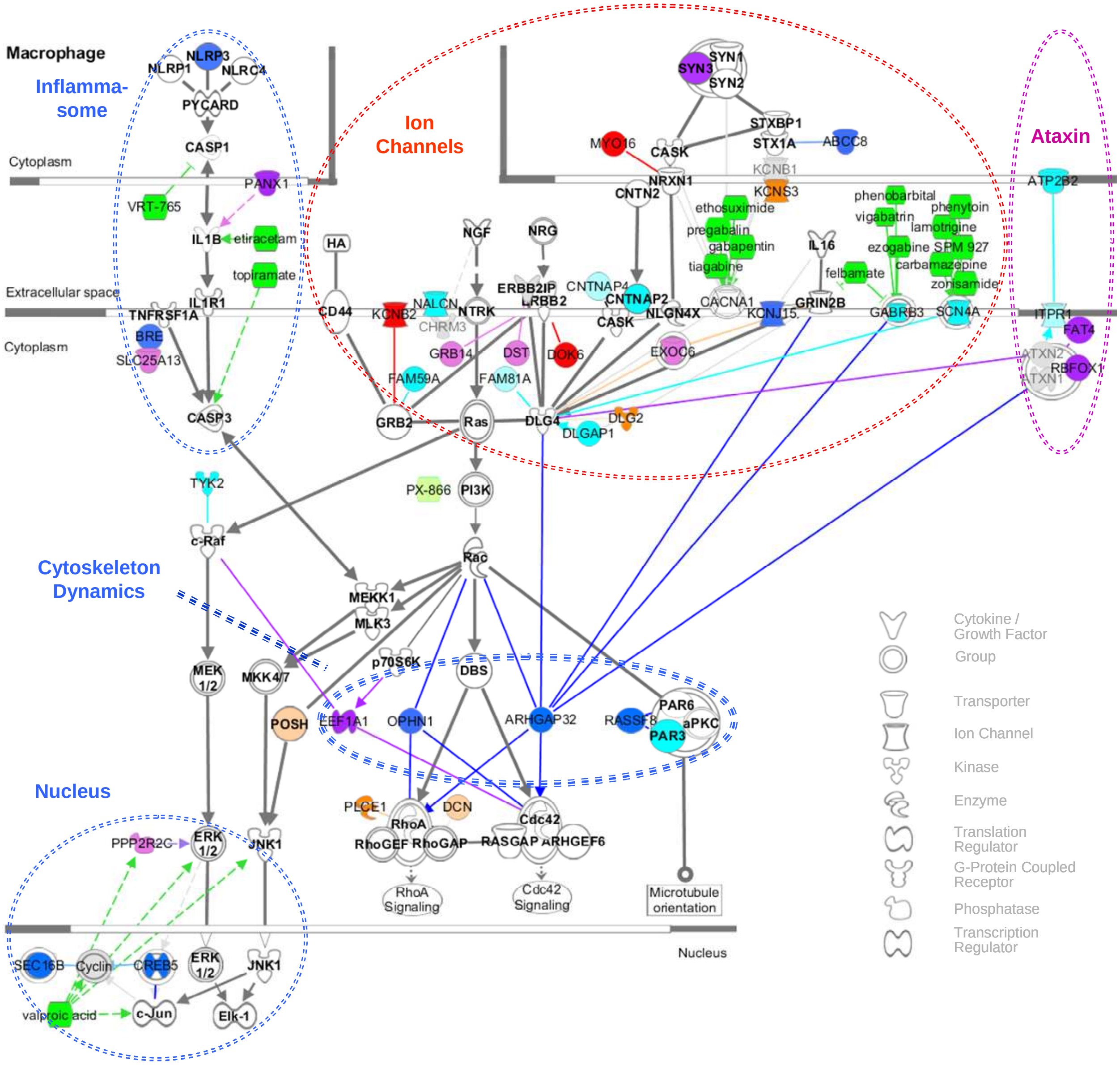


Figure 3: Published direct relationships between the minimal subset of the NOD-ID/AGS-Ataxin pathway directly related to significant genes by µGWAS (23 of 40, *s*>6.5) and lrGWAS (17 of 40, *s*>7.0), respectively. The members of the pathway are shown and labeled in bold. Methods are indicated in colors (blue: µGWAS; red: lrGWAS; pink: both). The most significant genes (µGWAS: >7.5, lrGWAS: >8.0) are shown in darker shades. (See Supplementary Table 1 for details). Dotted circles relate to functional clusters mentioned in the text. Drugs are indicated in green.

Channelopathies: Epilepsy is commonly seen as a channelopathy, and lrGWAS identifies both post-synaptic (*KCNB2*, r_{lr} = 3rd; *DOK6*, r_{lr} =10th) and pre-synaptic (*MYO16*, 13th) membrane processes. µGWAS adds *KCNJ15* (r_{μ} = 14th), confirms *CNTNAP2* [29] and *CNTNAP4* (27th, 48th), and hints at two targets for approved anti-epileptic drugs, the ion channels *SCN4A* and *GABRB3* (43rd and 57th, respectively) [30]. Both methods implicate *SYN3* (r_{μ} / r_{lr} = 3rd / 22nd), a presynaptic vesicle-associated protein [31]. µGWAS adds *OPHN1* and *ABCC8* (8th and 12th, respectively).

Inflammasome: Two approved anti-epileptic drugs, topiramate and levetiracetam, and the investigational drug VRT-765 target the NOD-like receptor signaling pathway [32]. While both approaches suggest genetic variations in *PANX1* (r_μ / r_{lr} = 13rd / 16th), µGWAS adds the *TNFRSF1A* modulator *BRE* (15th) as involved and *NLRP3* as a risk factor (16th). Hence, VRT-765 might be particularly effective for patients with a 'gain-of-function' mutation in *NLRP3*.

Cytoskeleton dynamics: *RHOA* was upregulated in patients with intractable epilepsy [33], yet the mechanism involved is unknown. Two genes known to regulate *RHOA*, *OPHN1* (also known as *ARHGAP41*) and *ARHGAP32* are among the top 10 genes with µGWAS, but rank only 99th and 58th, respectively, in lrGWAS. The risk of epilepsy is increased in children with intellectual disabilities (ID), where *ARHGAP32* has been implicated. Binding between *ARHGAP32* and *ATXN1* has been implicated in inherited ataxias [34]. *OPHN1* is known to affect X-linked ID [35] and, thus, might explain the preponderance of CAE among girls. µGWAS adds a pair of binding partners downstream of *RAC1* to the picture, *RASSF8* (17th) and *PARD3* (26th). Finally, the 'pseudogene' *EEF1A1P12*, being among the top 10 regions in both approaches, hints at an involvement of *EEF1A1*, which regulates *CDC42*. Hence, µGWAS uniquely provides a testable hypothesis about the mechanism by which *RHOA* is upregulated in some forms of epilepsy.

Ataxin: Ataxias and epilepsy share genetic risk factors [36, 37], including *OPHN1 [38, 39]*, and both methods implicate two genes binding ataxins, *RBFOX1* ($r_\mu = r_{lr}$ = 5th) and *FAT4* (r_μ / r_{lr} = 6th / 17th). µGWAS also hints at the Ca transporter *ATPB2* (39th) and the calcium channel *ITPR1* (42nd) as potential drug targets.

Nucleosome: The effectiveness of valproic acid in treating epilepsies hints at a role of nucleosome assembly in epilepsies and, in fact, µGWAS implicates mutations in *CREB5* and *SEC16B* (4th and 10th, respectively).

Detecting Epistasis and Selection

Among the genes involved in cytoskeleton dynamics, *ARHGAP32,* with known direct interactions with many of the key players, ranked 11th in µGWAS, but only 58th in lrGWAS. Moreover, it had two separate 'peaks' in µGWAS, one in the promoter region.

Epistasis between neighboring SNPs: The most significant SNPs in *ARHGAP32* by ssGWAS (s = 4.3 – 4.7) are all members of tag set **a** (Figure 4E). The two µGWAS peaks, separated by a clear trough (Figure 4C), pinpoint two loci where the effects of different haplotypes converge, centered within 4 kB of exon 10 and the promoter region (exon 0), respectively. Both regions contain a set **c** SNP as a distant member (≈20 kB), indicating a common 'background' risk factor, and two members of region specific tag sets (exon 10: **a**/**b**, exon 0: **g**/**h**). Both regions belong to a recently identified alternative splice variant, which is expressed during neural development and involved in axon and dendrite extension [40, 41]. lrGWAS results are also elevated, yet without discriminating intragenic regions (Figure 4D, insert).

Intragenic Epistasis/Selection: No case or control subject had more than four risk alleles among the three relevant SNPs in either region, although homozygous variants for each SNP are present. Hence, the unobserved combinations must have been selected against, e.g., because of a more severe phenotype.

Intergenic Epistasis: As approximately one-third of all subjects with the *ARHGAP32* genetic risk factor lack the phenotype, other genetic cofactors are yet to be identified. Figure 3 suggests the possibility of epistasis in *trans* between regulatory and functional factors, i.e., between the plasma membrane (*NGF*/*NRG*–*RAS*) and the cytoplasm (*RAC*–*RHOA*/*CDC42*).

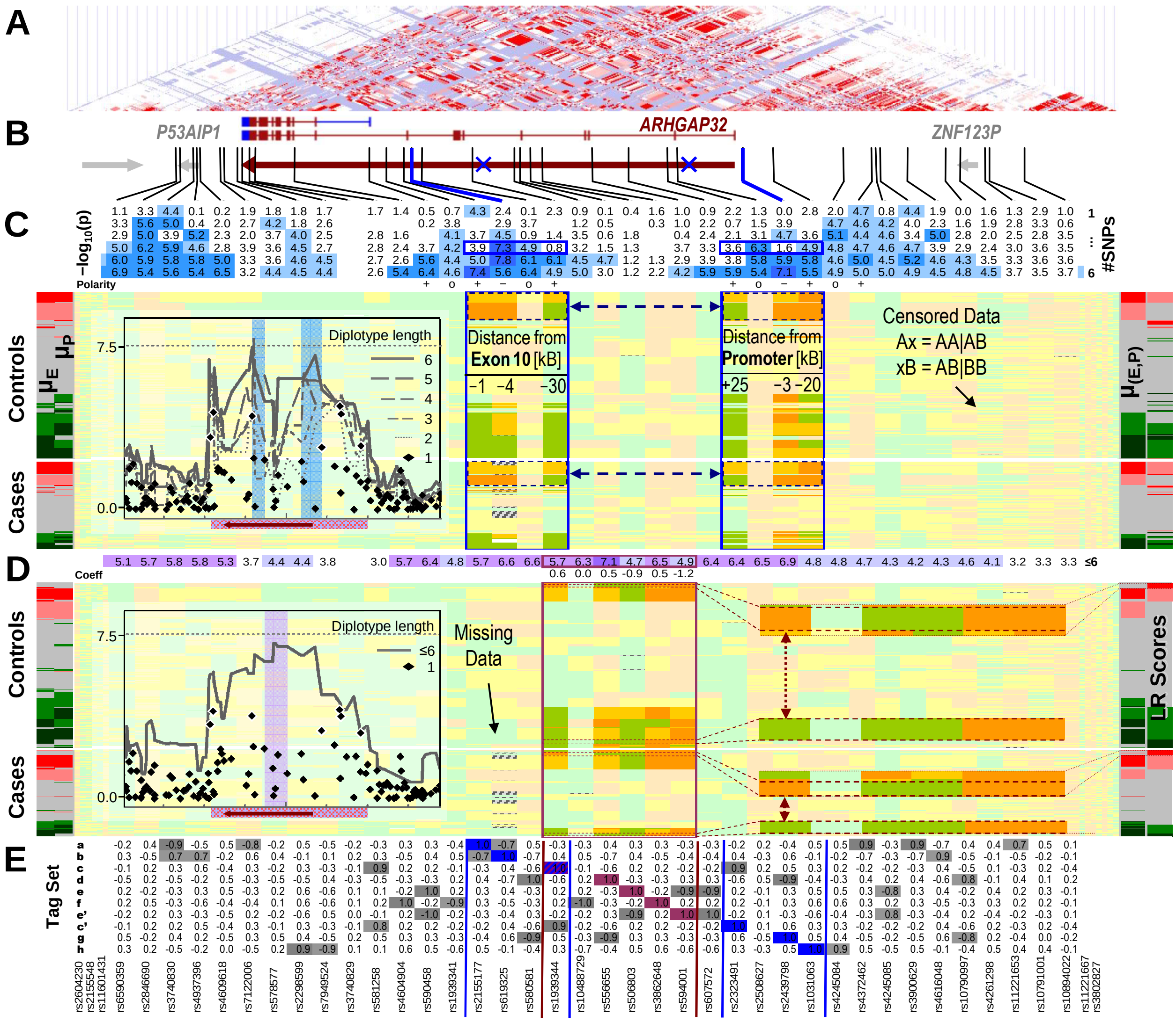


Figure 4: Microarray genotyping results for the linkage disequilibrium (LD) block containing *ARHGAP32*. (**A**) LD map; (**B**) coding regions; (**C**) µGWAS test results by diplotype length followed by the polarities of the SNPs contributing and the SNP pattern (orange: homozygous; yellow: heterozygous; green: wildtype) for controls and cases sorted by $\mu_{(E,P)} = \mu_{((E1,\ E2,\ E4),\ (P1,\ P3,\ P4))}$ (near right stub). Diplotypes ranked high (red) and low (green) by µ-scores for each region (left stub: $\mu_E = \mu_{(E1,\ E2,\ E4)}$, $\mu_P = \mu_{(P1,\ P3,\ P4)}$) are highlighted as more saturated. Horizontal arrows indicate consistently paired diplotypes. The insert shows the 'Manhattan plot' of the $-\log_{10}(p)$ values. (**D**) lrGWAS results followed by the lr coefficients (Coeff) of the SNPs involved. SNP pattern are sorted by lr scores (far right stub). The enlarged profiles with extreme lr scores differ in one the five SNP only (vertical arrow). The insert shows the $-\log_{10}(p)$ values based on univariate and stepwise lr. (**E**) LD between each of the 10 SNPs included in the two µGWAS regions (blue) and the lrGWAS region (purple) and the members of the same tag set (gray). Tag set **c** is represented in both µGWAS diplotypes and the lrGWAS diplotype, which contains two members of tag set **e**.
Coeff: Coefficient; E: Exon; (E,P): Exon 10 and promoter region; lr: logistic regression

Validation

In this analysis, we have reduced the potential for false positive results by taking advantage of the novel internal validation features made possible with µGWAS. During data preparation, we used a data quality µ-score based on a comprehensive assessment of missing data, Hardy-Weinberg equilibrium, short-range LD, and expected LD from HapMap information. During analysis, we have drawn on polarity conflict and lower than expected µIC (Supplementary Figure 1). Finally, we utilized µIC to indicate highly significant results with low µIC. Notably, none of

the pathway related genes flagged as potentially unreliable are related to the genes downstream of *RAC1* (Figure 2).

Larger genes are both more likely to carry mutations and to have false positives. Still, although several of the genes identified are among the largest 5% (>200 kB) in the human genome, only 2 of the top 11 unique genes in µGWAS (*CREB5* and *BRE*) and 3 of the top 13 unique genes in lrGWAS (*DYSF*, *DOK6* and *TMCO7*) are 'direct hits' within the coding region (Supplementary Table 1). *ARHGAP32* and *OPHN1* were implicated by 'hits' in the stop or promoter regions, respectively, and thus, are not at an increased risk for being false positives due to their size.

The results on *ARHGAP32* (Figure 4) are supported by further evidence. First, each of the six SNPs included in the two diplotypes is in high LD with several other SNPs (Figure 4E), for which the probe sequences differ and, thus, are not subject to the same calling errors. Second, only the two pairs of diplotypes having the highest association with disease risk by µ-scores were in high LD between the intragenic regions (Figure 4C, horizontal dashed arrows), while lower risk diplotypes were unrelated. Not only is it highly unlikely for each of these results to occur by chance alone, it is virtually impossible that they could occur together, and in both independent populations. While this cannot rule out a false positive result due to association with factors beyond the etiology of epilepsy, these findings validate the ability of µGWAS to detect intragenic regions of biologically relevant epistatic patterns. Finally, the diplotype with the highest overall (exon 10 and promoter region) score $\mu_{(E,P)}$ is clearly overrepresented among cases, with a prevalence of 14.1% (26 out of 185) and 6.5% (23 out of 354) in cases and controls, respectively, compared with 3.8% (7 out of 185) and 6.2% (22 out of 354) for the diplotypes with the lowest µ-scores, confirming that µ-scores are, in fact, reflecting disease risk.

As one would expect, µ and lr scores (Figure 4, right border) are correlated. The subjects with the pair of diplotypes having the highest $\mu_{(E,P)}$-scores (Figure 4D) also share a diplotype with a high lr-score, but the subjects scoring even higher in lr-scores (Figure 4E) comprise four different diplotypes. Interestingly, the largest of these groups differs only in the first SNP from a diplotype with low lr- and µ-scores (vertical arrows in Figure 4 D), consistent with the sensitivity of linear model results to outliers. As the partial ordering underlying µ-scores, which directly reflects an underlying functional model, results in more genetic uniformity among subjects with extreme scores, these more homogeneous sub-populations could then be selected for identification of functional variations through sequencing.

Conclusion

With GWAS of complex diseases, only a few solitary SNPs typically stand out from the noise, especially in small studies, and this study is no exception. Different compositions of rare disease variants across studies almost inevitably result in different SNPs being 'significant', so that validation in independent ssGWAS requires many large studies until a testable hypothesis emerges. µGWAS, in contrast, related approved and experimental drugs to functional clusters of genes along a known pathway in a study of 185 well characterized cases only.

ssGWAS can efficiently screen for loci, where a single haplotype confers all or most of the risk (*EEF1A1P12*). lrGWAS has advantages when the effects of SNPs are at least approximately independent and additive (as they might be in some transporters and ion channels). With more complex processes, however, like the interactions of *ARHGAP32* with its various binding and activation partners, not constraining results by making overly simplistic assumptions leads to biologically relevant hypotheses about functionally related genes clustered around biologically relevant pathways.

Pathway-based approaches [42] and Gene Set Enrichment Analysis [43], combine results of univariate statistics using assumptions regarding the relative importance of genes and prior declarations of relatedness among genes instead of observed interactions. However, this analysis

suggests that few, if any, pathway genes themselves may carry mutations in common diseases, unless they are members of redundant complexes (*NLRP3*, *SYN3*, and *PARD3,* Figure 3), in which case multiple genes may need to be knocked out to produce a phenotype [44].

Wide-locus GWAS aims at accounting for compound heterozygosity, different haplotypes carrying the same mutation, and epistasis between nearby disease loci. Hence, functional regions can be identified more easily, even when the contribution of individual SNPs would be difficult – if not impossible – to detect. Many traditional statistical methods, however, have deficiencies for relevant types of epistasis. *ARHGAP32,* which ranked 10th among µGWAS genes and was validated through the distinct epistatic pattern among the highest-risk allelotypes confirmed in sequencing (Figure 4), did not even appear among the top 50 lrGWAS regions.

µGWAS requires neither Hardy-Weinberg equilibrium nor independence or additivity/ multiplicativity of genetic effects, thereby improving sensitivity for non-linear effects (including evolutionary selection, Figure 4, horizontal dashed arrows, and Supplementary Table 1). Adding sequential interactions and recombination hotspots improves resolution, rather than creating artifacts. Together with *OPHN1* (also unique to µGWAS at rank 8), this study provides a plausible hypothesis why expression of *RHOA* is upregulated in some forms of epilepsy [33].

Increased expression of *RHOA* was recently associated with some epilepsies [33]. Both *OPHN1* and *ARHGAP32* interact with both *RHOA* and *PI3K* (Figure 3), a drug target currently investigated in cancer [45] and inflammatory diseases [46]. Wortmannin, an inhibitor of *PI3K*, attenuates effects of seizures in rats [47] and PX-866 (a oral drug derivative of Wortmannin, in a phase II prostate cancer trial 102), targets *PI3K*. If our results are confirmed and hold for patients with other epilepsies as well, this might lead to novel therapeutic approaches to treat patients whose seizures do not respond to drugs targeting ion channels, the inflammasome, or the nucleosome. As this study included only patients whose seizures were controlled by valproic acid and/or ion channel blockers, these genes may play an even larger role in other populations.

A particular advantage of µGWAS is the ability to guide the interpretation of data patterns in terms of biological function. Sorting diplotypes by the overall risk they confer (Figure 4C), rather than by linear weight scores lacking direct biological interpretation (Figure 4D) provided compelling evidence for intragenic epistasis (Figure 4C), facilitated validation (Figure 4F), and generated testable hypotheses regarding the function of underlying mutations. By utilizing the order of neighboring SNPs and HapMap information about their expected LD, µGWAS can often identify functional intra-genetic regions, whereas the resolution of lrGWAS, irrespective of sample size, is typically limited to an LD block as a whole. For instance, this analysis suggests that the combinations of diplotypes with the highest µ-score in either of the *ARHGAP32* regions have been selected for because they partially compensate for each other. Epistasis might also explain why knocking out the entire *ARHGAP32* gene produced no obvious phenotype in mice [48].

In summary, our results show that genetic risk factors for complex diseases cannot be adequately addressed with ssGWAS alone and that the computationally simple lrGWAS approach may be insensitive to complex forms of epistasis. Reducing artifacts by avoiding models motivated by computational convenience, rather than biological plausibility reduces the need for independent studies to guard against false positive results from model misspecifications. For comparative effectiveness research and personalized diagnostics to live up to their expectations, cases and controls need to be closely matched to the population or patient involved. Adequately controlling for genetic and environmental confounders when selecting appropriate cases and controls is essential to tease out predictive factors. This goal is much easier to achieve with only a few hundred subjects, rather than several thousands to be matched. Finally, subset analyses of phase-III trials and published epidemiological studies could rapidly reveal novel insights for drug development.

Future perspective

The Ras pathway is known to be involved in both cancers and many developmental disorders [49], so the findings here suggest that identifying genetic risk factors modulating this pathway may help in better using information from sequencing patients when targeting pharmacological interventions not only in cancers, but also in other neurodevelopmental diseases other than CAE, including ID and autism spectrum disorders [31].

With more appropriate statistical methods and more powerful computational tools becoming available, the focus in screening for genetic risk factors of complex diseases can now shift from individual SNPs scattered across the genome to clusters of genes around biologically meaningful pathways. With further advances in computational resources, µGWAS can be extended from epistasis across recombination hotspots (Figure 1) to epistasis between intragenic regions (such as those seen in Figure 4), and between genes (Figure 3).

As µGWAS can provide therapeutically relevant information from substantially smaller sample sizes, decisions in personalized medicine and comparative effectiveness research can be based on samples fine-tuned to the particular patient or population, respectively.

As a few hundred subjects experiencing adverse events or lack of a treatment effect and matched controls from the same population suffice to determine genetic risk factors, data from previous or upcoming Phase III trials can now be effectively mined to determine subpopulations at risk of adverse events and identify directions for development of drugs with a broader target population.

Information resources

- Wittkowski KM: Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *J Am Statist Assoc* 83 (404), 1163-1170 (1988); extension: 87 (417), 258 (1992).
- Kosoy R, Nassir R, Tian C et al.: Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 30(1), 69 - 78 (2009).
- Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 55(4), 997-1004 (1999).
- Commission on Classification and Terminology of the International League against Epilepsy: Proposal for revised classification of epilepsies and epileptic syndromes. *Epilepsia* 30, 389-399 (1989).
- Wittkowski KM, Song T: Nonparametric methods for molecular biology. *Methods Mol Biol* 620, 105-153 (2010).
- Hajek J, Sidak Z: *Theory of rank tests.* Academic, New York, NY. (1967).

Acknowledgements: The authors would like to acknowledge Sandra Wrigley, Dana Politis, and Ryan Cauley for collecting the samples, Ken Olden, Lorna E. Thorpe, Martin Dornbaum, and Frank Steen of the City University of New York School of Public Health for providing access to computational resources for the grid operation, Denis Kaplun, Bohao Zhou, and Ethan Schiffmiller for help with data inspection and analysis, William H. Greer and Greg S. Zhang for implementing GPU cloud instances, Elizabeth Horn for editorial assistance, and the reviewers and editors for many helpful suggestions.

Financial & competing interest disclosures: KM Wittkowski, V Sonakya, and T Song were in part funded by grant #2 UL1 RR024143 from the US National Center for Research Resources (NCRR) Clinical and Translational Science Award (CTSA) and #8 UL1 TR000043 from the National Center for Research Resources and the National Center for Advancing Translational Sciences (NCATS). M Durner and V Sonakya were in part funded by grant #2 R01NS037466 from

the US National Institute of Neurological Disorders and Stroke (NINDS). MP Seybold was in part funded by the German National Academic Foundation and the German Academic Exchange Agency. The authors have no other relevant affiliations or financial involvements with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistances was utilized in the production of this manuscript.

Ethical conduct of research: The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Executive Summary

Introduction

- The requirement for (tens of) thousands of subjects with univariate statistical approaches limits the usefulness of genome-wide association studies (GWAS) for comparative effectiveness research, personalized diagnostics / treatment, and subgroup analyses of phase III trials
- Several mutations within an intragenic or promoter region may contribute to the risk of common diseases.
- GWAS using multivariate statistical approaches based on unrealistic assumptions (e.g., independence and additivity) implicit to linear/logistic regression (lrGWAS) has low power to detect meaningful relationships and carries a high risk of false positives.
- The advent of massively parallel computing has spurred the development of statistical methods that requiring fewer unrealistic assumptions, including GWAS based on U-statistics for structured multivariate data (µGWAS).

Methods

- Extending µ-statistics to reflect linkage-disequilibrium (LD) structures in the data increases the power and avoids artifacts.
- A well-characterized sample of 185 children with childhood absence epilepsy was analyzed as an example.

Results

- With single-SNP GWAS, only two SNPs reached the customary level of significance.
- Of the top 17 regions in µGWAS, 14 (82%) were in genes related to a known disease-related signaling pathway, compared to only 8 (36%) of the top 22 regions in lrGWAS.
- µGWAS was able to detect intragenic regions (i.e., exon, promoter) and LD structures, suggesting evolutionary selection.

Conclusion

- Avoiding overly simplistic assumptions leads to biologically relevant hypotheses about functionally related genes clustered around biologically relevant pathways.
- The pathway identified by µGWAS contains targets of approved anti-epileptic drugs and a gene being investigated as a cancer drug target.
- Reducing artifacts by avoiding biologically implausible assumptions guards against false positive results from model misspecifications.
- By reducing the GWAS sample sizes to a few hundred subjects only, µGWAS enables personalized medicine, comparative effectiveness research, and subset analyses of epidemiological studies / phase III trials.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS: A vision for the future of genomics research. *Nature* 422(6934), 835-847 (2003).
2. Klein RJ, Zeiss C, Chew EY et al.: Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720), 385-389 (2005).

- **Describes the first successful genome-wide association study (GWAS), which identified a variation in a single SNP, rs1061170, as causing age-related macular degeneration, yet did not yield a treatment.**

3. Sullivan P: Don't give up on GWAS. *Mol Psychiatry* 17(1), 2-3 (2012).
4. Klein C LKZA: The promise and limitations of genome-wide association studies. *JAMA* 308(18), 1867-1868 (2012).
5. Psychiatric Gwas Consortium Coordinating Committee: Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* 166(5), 540-556 (2009).
6. Meehl PE: Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46, 806-834 (1978).
7. Waller NG: The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology* 11, 83-86 (2004).
8. Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* 4(9), 701-709 (2003).
9. Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433. (1918).
10. Goldstein DB: Common Genetic Variation and Human Traits. *New England Journal of Medicine* 360(17), 1696-1698 (2009).
11. Ballard DH, Cho J, Zhao HY: Comparisons of Multi-Marker Association Methods to Detect Association Between a Candidate Region and Disease. *Genet. Epidemiol.* 34(3), 201-212 (2009).
12. Bechhofer RE: A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* 25, 16-39 (1954).

- **Introduces the concept of selection-procedures as opposed to confirmatory tests.**

13. Rosenthal R: Cumulating evidence. In: *A handbook for data analysis in the behavioral sciences: Methodological issues*, Keren G, Lewis C (Eds). Erlbaum, Hillsdale, NJ 519-559 (1993).
14. Morales JF, Song T, Auerbach AD, Wittkowski KM: Phenotyping genetic diseases using an extension of µ-scores for multivariate data. *Stat Appl Genet Mol* 7(1), 19 (2008).

- **Introduces the mathematical underpinning of extending u-scores by including knowledge about a hierarchical factor structure, of which the SNP-related chromosomal intervals are a special case.**

15. Lawrence R, Evans DM, Morris AP et al.: Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. *Genome Research* 15(11), 1503-1510 (2005).

- **Discusses the problems tag sets of genetically indistingushable SNPs may case for identifying intragenic regions.**

16. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164), 851-861 (2007).
17. Mann HB, Whitney DR: On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1), 50-60 (1947).
18. Hoeffding W: A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19, 293-325 (1948).

- **Derives the asymptotic distribution of linear rank tests based on u-scores.**

19. Li H: U-statistics in genetic association studies. *Hum Genet* 131(9), 1395-1401 (2012).

- **Discusses the potential of U-statistics for GWAS, in general, including their limitations, e.g., when "irrelevant SNPs" within a diplotype are not considered.**

20. Pearson TA, Manolio TA: How to interpret a genome-wide association study. *JAMA* 299(11), 1335-1344 (2008).
21. Wittkowski KM, Song T, Anderson K, Daniels JE: U-Scores for Multivariate Data in Sports. *Journal of Quantitative Analysis in Sports* 4(3), 7 (2008).
22. Diana M, Song T, Wittkowski KM: Studying travel-related individual assessments and desires by combining hierarchically structured ordinal variables. *Transportation* 36(2), 187-206 (2009).
23. Hilton JF: The appropriateness of the Wilcoxon test in ordinal data. *Statistics in Medicine* 15(6), 631-645 (1996).
24. Loiseau P, Panayiotopoulos CP, Hirsch E: *Childhood Absence Epilepsy and Related Syndromes.* In: Epilepsy Syndromes in Infancy, Childhood and Adolescence, Roger J, Bureau M, Dravet C, Genton P, Tassinari CA, Wolf P (Eds). John Libbey, Montrouge, France 285-303 (2002).
25. Glauser TA, Cnaan A, Shinnar S et al.: Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy. *N Engl J Med* 362(9), 790-799 (2010).
26. Kim JE, Choi HC, Song HK et al.: Levetiracetam inhibits interleukin-1 beta inflammatory responses in the hippocampus and piriform cortex of epileptic rats. *Neurosci Lett* 471(2), 94-99 (2010).
27. Akaike H: A new look at statistical-model identification. *IEEE Trans. Autom. Control AC*19(6), 716-723 (1974).
28. Lehmann EL: Some model I problems of selection. *Annals of Mathematical Statistics* 32, 990-1012 (1961).

29. Friedman JI, Vrijenhoek T, Markx S et al.: CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. Molecular Psychiatry 13(3), 261-266 (2008).
30. Crunelli V, Leresche N: Childhood absence epilepsy: genes, channels, neurons and networks. Nat Rev Neurosci 3(5), 371-382 (2002).
31. Van Bokhoven H: Genetic and epigenetic networks in intellectual disabilities. Annual review of genetics 45, 81-104 (2011).

- **Provides a comprehensive review of the intellectual disability pathway**

32. Oprica M, Eriksson C, Schultzberg M: Inflammatory mechanisms associated with brain damage induced by kainic acid with special reference to the interleukin-1 system. J Cell Mol Med 7(2), 127-140 (2003).
33. Yuan J, Wang L-Y, Li J-M et al.: Altered Expression of the Small Guanosine Triphosphatase RhoA in Human Temporal Lobe Epilepsy. Journal of Molecular Neuroscience 42(1), 53-58 (2010).
34. Lim J, Hao T, Shaw C et al.: A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125(4), 801-814 (2006).
35. Billuart P, Bienvenu T, Ronce N et al.: Oligophrenin-1 encodes a rhoGAP protein involved in X-linked mental retardation. Nature 392(6679), 923-926 (1998).
36. Gribaa M, Salih M, Anheim M et al.: A new form of childhood onset, autosomal recessive spinocerebellar ataxia and epilepsy is localized at 16q21-q23. Brain 130(7), 1921-1928 (2007).
37. Imbrici P, Jaffe SL, Eunson LH et al.: Dysfunction of the brain calcium channel Ca(V)2.1 in absence epilepsy and episodic ataxia. Brain 127, 2682-2692 (2004).
38. Tentler D, Gustavsson P, Leisti J et al.: Deletion including the oligophrenin-1 gene associated with enlarged cerebral ventricles, cerebellar hypoplasia, seizures and ataxia. Eur J Hum Genet 7(5), 541-548 (1999).
39. Bergmann C, Zerres K, Senderek J et al.: Oligophrenin 1 (OPHN1) gene mutation causes syndromic X-linked mental retardation with epilepsy, rostral ventricular enlargement and cerebellar hypoplasia. Brain 126(Pt 7), 1537-1544 (2003).
40. Hayashi T, Okabe T, Nasu-Nishimura Y et al.: PX-RICS, a novel splicing variant of RICS, is a main isoform expressed during neural development. Genes to Cells 12(8), 929-939 (2007).

- **First report on the extended splice variant of ARHGAP32/RICS**

41. Nakamura T, Hayashi T, Mimori-Kiyosue Y et al.: The PX-RICS-14-3-3 zeta/theta Complex Couples N-cadherin-beta-Catenin with Dynein-Dynactin to Mediate Its Export from the Endoplasmic Reticulum. Journal of Biological Chemistry 285(21), 16145-16154 (2010).
42. Wang K, Zhang H, Kugathasan S et al.: Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. 84(3), 399-405 (2009).
43. Subramanian A, Tamayo P, Mootha V et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102(43), 15545-15550 (2005).
44. Ketzef M, Kahn J, Weissberg I, Becker AJ, Friedman A, Gitler D: Compensatory network alterations upon onset of epilepsy in synapsin triple knock-out mice. Neuroscience 189, 108-122 (2011).
45. Courtney KD, Corcoran RB, Engelman JA: The PI3K pathway as drug target in human cancer. J Clin Oncol 28(6), 1075-1083 (2010).
46. Harris SJ, Foster JG, Ward SG: PI3K isoforms as drug targets in inflammatory diseases: lessons from pharmacological and genetic strategies. Curr Opin Investig Drugs 10(11), 1151-1162 (2009).
47. Xue Y, Xie N, Cao L, Zhao X, Jiang H, Chi Z: Diazoxide preconditioning against seizure-induced oxidative injury is via the PI3K/Akt pathway in epileptic rat. Neurosci Lett 495(2), 130-134 (2011).
48. Nasu-Nishimura Y, Hayashi T, Ohishi T et al.: Role of the Rho GTPase-activating protein RICS in neurite outgrowth. Genes Cells 11(6), 607-614 (2006).
49. Schubbert S, Shannon K, Bollag G: Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer 7(4), 295-308 (2007).

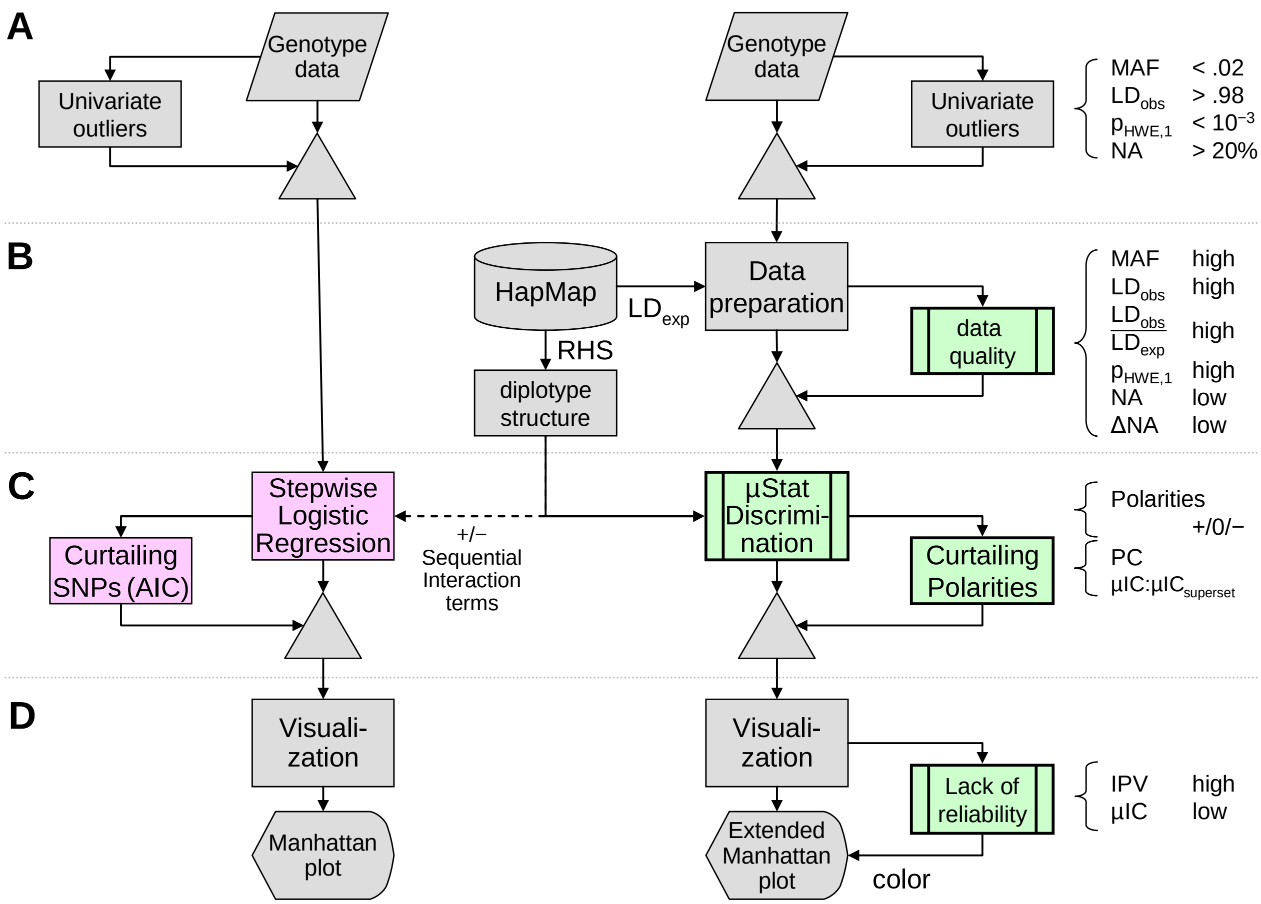
Websites

101 A Study to Evaluate the Efficacy and Safety of VX-765 in Subjects With Treatment-Resistant Partial Epilepsy. http://clinicaltrials.gov/ct2/show/NCT01501383

102 A phase II Study of PX-866 in Patients With Recurrent or Metastatic Castration Resistant Prostate Cancer. http://clinicaltrials.gov/show/NCT01331083

103 Illumina. Science / Illumina iControlDB. http://www.illumina.com/science/icrontroldb.ilmn

Supplementary Information



Supplementary Figure 1: Analytical Workflow of lrGWAS (left) and µGWAS (right). (A) Initial data cleaning based on univariate cut-offs for minor allele frequency (MAF), high observed LD among neighboring SNPs (LD_{obs}), violation of Hardy-Weinberg equilibrium (HWE), or missing calls (NA). (B) Exclusion of data based on low data quality µ-scores, including low ratio of observed vs. expected LD from HapMap is a unique feature of µGWAS. HapMap information can also be used to determine whether to consider recombination hotspots in the diplotype structure. (C) µStat discrimination utilizes the same information about the diplotype structure as logistic regression with sequential interaction terms. Excluding a polarity in µGWAS based on polarity conflict or low µIC compared to µIC among its supersets serves a similar purpose as excluding SNPs in logisktic regression based on the AIC. (D) Identification of significant results with low reliability is a unique feature of µGWAS.

Supplementary Table 1: Most significant genes by either method (lrGWAS 61, >7.0, µGWAS: 60, >6.5, total: 96) ($-\log_{10}(p)$, rank) by lrGWAS and µGWAS. Len/Dst: length of gene and distance from gene (−0►: promoter region, ☼: direct hit, +0◄: beyond stop codon, ±0▲: entire gene). Results with low reliability µ-score are indicated in red.

Method	Symbol	Entrez	lrGWAS	µGWAS	Chr	Coor	Len/Dst (kb)		Name
Both	(Chr11)	-1	8.69 (8)	10.11 (1)	11	80,664,454			---
	EEF1A1P12	1915	8.70 (7)	8.74 (2)	2	106,702,196	2	±0▲	eukaryotic translation elongation factor
	SYN3	8224	8.02 (22)	8.53 (3)	22	31,464,046	493	☼	Synapsin III
	RBFOX1	54715	8.77 (5)	8.31 (5)	16	6,268,023	659	☼	ataxin 2-binding protein 1
	FAT4	79633	8.11 (17)	8.21 (6)	4	127,111,750	175	+250◄	FAT tumor suppressor …
	PANX1	24145	8.19 (16)	7.70 (13)	11	93,415,789	52	−0►	pannexin 1
µGWAS	CREB5	9586	5.13 (94)	8.35 (4)	7	28,348,933	406	☼	cAMP responsive element binding protein 5
	B3GALT1	8708	7.37 (48)	8.19 (7)	2	168,340,869			beta-1,3-galactosyltransferase 1
	OPHN1	4983	5.46 (90)	8.18 (8)	X	67,037,602	385	+0▲	oligophrenin 1 / ARHGAP41
	PITPNB	23760	7.02 (61)	8.03 (9)	22	26,626,038			phosphatidylinositol transfer protein, beta
	SEC16B	89866	6.95 (65)	7.82 (10)	1	174,647,155	38	☼	SEC16 homolog B (S. cerevisiae)
	ARHGAP32	9743	7.08 (58)	7.80 (11)	11	128,420,261	223	☼−0►	Rho GTPase activating protein 32
	ABCC8	6833	5.90 (81)	7.76 (12)	11	17,400,710	84	☼	ATP-binding cassette, sub-family C (CFTR/MRP)
	KCNJ15	3772	6.47 (73)	7.67 (14)	21	38,578,375	4	−0►	potassium inwardly-rectifying channel …
	BRE	9577	7.60 (34)	7.61 (15)	2	28,235,520	444	☼	brain and reproductive organ expressed
	NLRP3	114548	7.71 (31)	7.61 (16)	1	243,940,658	30	+0◄	NLR family, pyrin domain …
	RASSF8	11228	7.60 (33)	7.50 (17)	12	25,927,109	24	−20►	Ras association (RalGDS/AF-6) domain family …
lrGWAS	*CA397621*	-1	9.50 (1)	2.34 (92)	5	25,722,226			---
	DYSF	8291	9.18 (2)	5.06 (73)	2	71,622,796			dysferlin
	KCNB2	9312	9.03 (3)	3.80 (80)	8	73,488,130	370	−100►	potassium voltage-gated channel, Shab-related …
	?	-1	8.90 (4)	0.00 (96)	7	118,571,616			---
	?	-1	8.75 (6)	2.93 (88)	1	83,607,917			---
	PNP	4860	8.57 (9)	6.12 (62)	14	20,027,673			purine nucleoside phosphorylase
	DOK6	220164	8.53 (10)	3.26 (84)	18	65,507,016	440	☼	docking protein 6
	VPS54	51542	8.46 (11)	6.55 (60)	2	64,169,397			vacuolar protein sorting 54 homolog
	FAM13C	220965	8.33 (12)	3.15 (86)	10	60,917,716			family with sequence similarity 13, member C
	MYO16	23026	8.27 (13)	4.35 (79)	13	107,967,111	577	−20►	myosin XVI
	TMCO7	79613	8.22 (14)	4.94 (74)	16	67,514,126	240	☼	transmembrane channel-like 7
	SETD7	80854	8.21 (15)	6.69 (56)	4	140,865,487			SET domain containing (lysine methyltransferase) 7
	OR10H3	26532	8.05 (18)	2.52 (90)	19	15,712,229			olfactory receptor …
	MVK	4598	8.05 (19)	7.27 (29)	12	108,547,979			mevalonate kinase
	MLC1	23209	8.05 (19)	5.59 (70)	22	48,812,715			megalencephalic leukoencephalopathy …
	COL21A1	81578	8.04 (21)	4.52 (78)	6	56,216,468			collagen, type XXI, alpha 1
Both	PPP2R2C	5522	7.60 (35)	7.38 (22)	4	6,565,679	212	+20◄	protein phosphatase 2 …
	MLEC	9761	7.58 (37)	7.32 (24)	12	119,598,228			malectine
	COL8A1	1295	7.89 (24)	7.10 (36)	3	100,886,715			collagen, type VIII, alpha 1
µGWAS	ATP8B1	5205	5.44 (91)	7.40 (18)	18	53,604,782			ATPase, aminophospholipid transporter
	SHISA6	388336	6.31 (76)	7.40 (19)	17	11,178,551			shisa homolog 6 (Xenopus laevis)
	?	-1	5.51 (89)	7.40 (20)	22	25,862,056			---
	?	-1	7.07 (59)	7.39 (21)	16	61,231,559			---
	BI918059	-1	7.16 (55)	7.35 (23)	3	35,141,439			---
	TFDP2	7029	6.66 (69)	7.30 (25)	3	143,151,151			transcription factor Dp-2 (E2F dimerization partner 2)
	PARD3	56288	6.43 (74)	7.29 (26)	10	34,324,843	704	+100◄	par-3 partitioning defective 3 homolog (C. elegans)
	CNTNAP2	26047	6.64 (70)	7.29 (27)	7	146,696,753	2,299	☼	contactin associated protein-like 2
	DLGAP1	9229	5.84 (82)	7.27 (28)	18	4,162,963	381	*−200►	discs, large (Drosophila) homolog-associated
	MYO1B	4430	5.75 (85)	7.25 (30)	2	192,230,956			myosin 1B
	NALCN	259232	6.52 (72)	7.24 (31)	13	100,580,679	344	☼	sodium leak channel, non-selective
	BG205085	-1	6.96 (64)	7.21 (32)	3	70,521,278			---
	ISOC1	51015	6.30 (77)	7.19 (33)	5	128,517,352			isochorismatase domain
	DST	667	7.25 (52)	7.18 (34)	6	56,824,034	184	−0►	dystonin
	BAZ2B	29994	6.99 (62)	7.15 (35)	2	160,127,655			bromodomain adjacent to zinc finger domain, 2B
	AI028357	-1	6.82 (66)	7.09 (37)	13	61,594,422			---
	MCTP2	55784	6.09 (79)	7.02 (38)	15	92,923,138			multiple C2 domains, transmembrane 2
	ATP2B2	491	5.41 (92)	7.02 (39)	3	10,432,572	121	☼	ATPase, Ca++ transporting, plasma membrane 2
	FAM59A	64762	5.55 (88)	7.02 (40)	18	28,282,875	203	☼	Family with sequence similarity 59, member A
lrGWAS	HLADQB1	3119	7.99 (23)	2.49 (91)	6	32,760,295			MHC, class II, DQ alpha 1
	?	-1	7.89 (25)	3.21 (85)	7	156,366,610			---
	COBLL1	22837	7.89 (26)	5.84 (66)	2	165,394,092			cordon-bleu protein-like 1
	MED17	9440	7.83 (27)	5.80 (67)	11	93,190,216			mediator complex subunit 17
	KCNS3	3790	7.78 (28)	6.02 (64)	2	18,114,469	1	+50◄	potassium voltage-gated channel …
	LOC...	100616530	7.72 (29)	4.74 (76)	8	96,508,202			---
	LOC...	388882	7.71 (30)	5.66 (68)	22	22,159,593			---
	NAV3	89795	7.64 (32)	6.37 (61)	12	77,352,055			neuron navigator 3
	SPTLC1	10558	7.60 (35)	3.42 (83)	9	91,986,563			protein tyr phosphatase, receptor type, V, pseudogene
	PLCE1	51196	7.57 (38)	2.81 (89)	10	95,741,477	294	☼	phospholipase C, epsilon 1
	DLG2	1740	7.52 (39)	5.31 (72)	11	83,257,555	2,139	☼	discs, large homolog 2 (Drosophila)
	EXOC6	54536	7.50 (40)	6.61 (58)	10	94,769,530	224	☼	exocyst complex component 6
Both	GRB14	2888	7.31 (51)	6.77 (49)	2	165,040,586	128	+100◄	growth factor receptor-bound protein 14
	SLC25A13	10165	7.39 (46)	6.76 (51)	7	95,392,974	201	+5◄	solute carrier family 25, member 13 (citrin)
	HEATR3	55027	7.48 (41)	6.73 (54)	16	48,631,409			HEAT repeat containing 3
	?	-1	5.40 (93)	6.95 (41)	4	106,143,434			---
	ITPR1	3708	6.99 (63)	6.95 (42)	3	4,703,008	330	☼	inositol 1,4,5-triphosphate receptor, type 1
	SCN4A	6329	5.01 (95)	6.92 (43)	17	59,402,439	32	±0▲	sodium channel, voltage-gated, type IV, alpha subunit
	CR591360	-1	6.68 (68)	6.90 (44)	5	38,796,716			---
	TYK2	7297	6.04 (80)	6.87 (45)	19	10,333,933	28	+0◄	tyrosine kinase 2
	LHX2	9355	6.42 (75)	6.86 (46)	9	123,915,038			LIM homeobox …
	?	-1	5.80 (84)	6.82 (47)	9	27,859,510			---
	CNTNAP4	85445	6.16 (78)	6.79 (48)	16	75,163,254	281	+10◄	contactin associated protein-like 4
	PDIA5	10954	6.57 (71)	6.77 (50)	3	124,348,989			protein disulfide isomerase …
	?	-1	4.83 (96)	6.75 (52)	18	66,372,380			---
	LY6H	4062	5.70 (86)	6.74 (53)	8	144,308,256			lymphocyte antigen 6 complex …
	FAM81A	145773	5.63 (87)	6.73 (55)	15	57,615,590	63	+0◄	family with sequence similarity 81, member A
	GABRB3	2562	5.81 (83)	6.66 (57)	15	24,599,861	226	-50►	gamma-aminobutyric acid (GABA) A receptor, beta 3
	VPS13B	157680	6.75 (67)	6.56 (59)	8	100,007,646			vacuolar protein sorting 13 homolog B (yeast)
	SETD4	54093	7.48 (42)	1.47 (95)	21	36,344,836			SET domain containing 4
	GPC5	2262	7.43 (43)	5.62 (69)	13	91,710,120			glypican5
	ALG6	29929	7.40 (44)	4.81 (75)	1	63,530,843			asparagine-linked glycosylation 6 homolog
	BE794467	-1	7.40 (45)	3.46 (82)	2	140,701,918			---
	IYD	389434	7.38 (47)	6.11 (63)	6	150,731,193			iodotyrosine deiodinase
	KIAA0146	23514	7.36 (49)	5.95 (65)	8	48,244,020			---
	SGSM1	129049	7.36 (50)	5.34 (71)	22	23,550,433			small G protein signaling modulator 1
	BU665313	-1	7.23 (53)	3.03 (87)	18	39,506,698			---
	AUTS2	26053	7.21 (54)	3.58 (81)	7	69,533,347			autism susceptibility candidate 2
	GTF3C5	9328	7.13 (56)	2.24 (94)	9	132,943,037			general transcription factor …
	DCN	1634	7.09 (57)	4.73 (77)	12	90,068,162	32	-50►	decorin/bone proteoglycan II
	POSH	57630	7.07 (60)	2.27 (93)	4	170,489,901	177	☼	SH3 domain containing ring finger 1

Supplementary Figure 2: Extended Manhattan Plot for the Comparison of 185 CAE cases vs matched controls. top/center: see Figure 2 legend for details; bottom: lrGWAS with sequential interaction. Genes implicated by only one of the methods are shown with that method against the dark background of univariate results.

Cases

The study was approved by the IRBs of both the Mount Sinai School of Medicine and The Rockefeller University. Our cases included 185 patients with CAE according to the criteria devised by the International League against Epilepsy [50]. To reduce genetic heterogeneity, we required that patients did not have seizures other than febrile seizures prior to the onset of absence seizures, that they had at least one EEG with a 3 Hz spike-wave pattern, and that all patients were be seizure free on antiepileptic medication. Only 21 patients developed generalized tonic clonic seizures after the onset of absence seizures, and only one patient had myoclonic jerks.

Controls

Only the 8,231 controls that were typed for the Illumina HumanHapmap 300 array or higher were considered. To reduce confounding due to population stratification and the risk of spurious results, we genotypically matched three sets of controls to the cases by ancestry information markers [51] using distinct criteria, and we then performed a stratified analysis [52] adjusted for overlaps of subjects between strata. We randomly split the top 96 ancestry informative markers (AIMS) [51] into two sets to create distinct control groups matched for different variables. Matching was performed in two different ways: 1) matching the frequency distribution at those AIMS on a population level and 2) matching cases individually to controls for as many genotypes as possible at either of the AIMS subsets, giving preference to controls matching by several sets of criteria. To check the quality of our matching algorithms, we calculated lambda (the inflation factor of the chi square distribution [53]) from all genotyped loci in the respective case/control samples. Lambda with all three control groups was 1.00–1.01, consistent with absence of population stratification. The availability of three different control groups is helpful to reduce the risk of false positives due to random variation in the control genotype frequencies.

Genotyping

To match the controls, we restricted the analysis to those markers included in the Illumina HumanHapmap300 SNPs. Genotyping was performed at the Illumina preferred vendor laboratory of the DNA Sequencing and Genotyping facility at Cincinnati Children's Hospital (CCHMC).

We performed extensive data checking for quality assurance. First, the reported sex was validated using X-linked SNPs. Although µGWAS does not require SNPs to be in Hardy-Weinberg Equilibrium (HWE), we then inspected all SNPs that deviated from HWE ($p < 0.001$, 3589 SNPs) and visually inspected all loci with >10% missing calls. After the first 140 subjects, we switched from the Illumina HumanCNV370_Duo to the HumanCNV370_Quad chip, which, in general, provides higher quality calls. After the GeneTrain2 algorithm became available, we manually rescored all loci with >1% of missing calls and visually inspected all SNPs where the new algorithm did not substantially reduce or even inflated the number of missing calls. We also inspected all SNPs where a χ^2 test rejected the homogeneity between duo and quad chip case distributions ($p<0.0001$).

After visual inspection, we removed all SNPs where 20% of calls were missing. If either >98% were AA or >98% were BB across cases and controls, the SNP was excluded as non-informative (minor allele frequency, MAF). Similarly, if two neighboring SNPs had >98% "identical" contingency tables, the SNP was also excluded as non-informative (LD). Missing data were recoded as interval censored, based on the sign of 'theta' (A−B)/(A+B). SNPs missing by design in the duo chip were excluded from the comparison.

To guard against differences between chips, we included the χ^2 test for homogeneity across case distributions across chips when computing the data quality µ-scores.

Statistics

U-statistics for multivariate data have been recently extended to allow variables to be hierarchically structured [14]. Since then, details of the method have been repeatedly published (see [54] for an overview) with applications ranging from sports [21] and policy making [22] to medicine [14].

As each of six neighboring SNPs could be either 'good', 'bad', or 'irrelevant', a comprehensive analysis requires $3^6 = 729$ 'polarities' (combinations of −1/0/+1) to be considered, and each of these multivariate analyses is substantially more complex than a univariate analysis. For each polarity, the allele profiles form a partial order (PO), where allele profile A confers more risk than profile B if it has the same risk alleles as profile B plus some additional risk alleles. Denoting risk alleles with capital letters, (Xx, YY, zz), for example, confers a greater risk than (Xx, Yy, zz), but the pairwise ordering of either profile with (xx, Yy, Zz) is ambiguous, because the contribution of Z to the overall risk vs. that of X and Y is unknown. The profile μ-score (u-scores for multivariate data) is the number of profiles with an unambiguously lower risk minus the number of profiles with an unambiguously higher risk. Treating loci with one unknown allele as 'interval-censored', i.e., as not-xx (xX or XX) or not-XX (xx or xX), respectively, further decreases ambiguities. One then compares disease categories by a linear rank test [55] applied to the μ-scores [18]. As the direction of each SNP's effect is unknown, many polarities need to be considered when screening for the one that best discriminates between disease categories.

Here, we first scored the subjects within each stratum, and then computed hierarchically structured μ-scores [14], using a special case of such a hierarchical structure. At the first level of the hierarchy one computes the matrices of pairwise comparisons representing the order (partial order in case of censored calls) of the SNPs, e.g. in the context of Figure 1, X, Y, and Z. At the second level of the hierarchy, the matrices of two adjacent SNPs are combined into a matrix for interval between these SNPs, e.g., (Y,Z), unless the two SNPs are separated by a recombination hotspot, where the matrix is filled with zeroes (X,Y)=0. Then, at the third level, the n single SNP and n−1 interval matrices are combined to obtain the diplotype matrix, from which the μ-scores were computed.

At each locus, we performed tests for diplotypes of length 1–6 centered at or above the locus. We allowed <50% of SNPs to be excluded from a diplotype, but not the first and the last, and considered all combinations of polarities (−1, 0, +1) among the SNPs included, except that the first and the last SNP as well as at least 50% of the SNPs included needed to be non-null. I.e., for a diplotype of length 5, the polarities (±1, ±1, ±1, ±1, ±1), (±1, 0, ±1, ±1, ±1), (±1, ±1, 0, ±1, ±1), (±1, ±1, ±1, 0, ±1), (±1, 0, 0, ±1, ±1), (±1, 0, ±1, 0, ±1), and (±1, ±1, 0, 0, ±1).

The effect and variance estimates of each block were then incorporated into a stratified Wilcoxon/Mann-Whitney type test statistic [52]. To adjust for the overlap between strata, the average across the three strata was weighted with an empirically confirmed √3, rather than 3.

By construction, tests based on μ-scores are sensitive to all monotonous (including dominant, trend, and recessive) alternatives.

As no particular hypotheses regarding specific loci were to be confirmed and most adjustments do not change the order of the results, no adjustment for multiple confirmative testing is warranted.

To avoid artifacts, we used four strategies:

- **Quality-of-Data µ-score:** We excluded SNPs not only based on the usual univariate criteria for missing calls, HWE, and minor allele frequency (MAF), but also when they had a low overall data quality µ-score, even if no category met the univariate criteria. We included observed short distance LD and the ratio between observed and expected LD (from HapMap) among the criteria.

- **Polarity conflict, PC:** We excluded polarities from analysis when the product of the signs assigned to pairs of SNPs in high LD and the sign of the LD were discordant.

- **Monotonicity in µIC:** As µIC (number of unambiguous pairwise orderings) tends to decline with diplotype length, we also excluded polarities resulting in lower µIC for a diplotype than the median µICs of its supersets as a non-parametric approach to regularization.

- **Reliability µ-score:** Finally, we highlighted results as questionable (red) when the reliability µ-score µ(p value, µIC) was low.

As the length of diplotypes increases, more pairwise orderings become ambiguous with µGWAS as soon as more 'noise' than 'signal' is added [14]. Hence, in contrast to lrGWA, no arbitrary upper limit (based, e.g., on AIC [27]) for diplotype length is needed. Significant results were associated with a particular gene only for regions within 20 kB of a gene or overlapping EST.

Software and resources used: Relationships were compiled using IPA (Ingenuity® Systems, www.ingenuity.com), KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg), and BioGraph (Biomedical knowledge discovery server, http://www.biograph.be). Figure 3 was created using the IPA Path Designer. The pathway involved in presynaptic cycling (*SYN3* ... *DLG4*) was adapted from [31].

Web services provided: GWAS data can be uploaded to a grid server via the Web (http://mustat.rockefeller.edu).

Additional References

50. Commission on Classification and Terminology of the International League against Epilepsy: Proposal for revised classification of epilepsies and epileptic syndromes. *Epilepsia* 30, 389-399 (1989).

51. Kosoy R, Nassir R, Tian C *et al.*: Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 30(1), 69-78 (2009).

52. Wittkowski KM: Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *J Am Statist Assoc* 83, 1163-1170, Extension: 1992;1187:1258 (1988).

53. Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 55(4), 997-1004 (1999).

54. Wittkowski KM, Song T: Nonparametric methods for molecular biology. *Methods Mol Biol* 620, 105-153 (2010).

55. Hajek J, Sidak Z: Theory of rank tests. Academic, New York, NY. (1967).

1. Collins FS, Green ED, Guttmacher AE, Guyer MS: A vision for the future of genomics research. *Nature* 422(6934), 835-847 (2003).
2. Klein RJ, Zeiss C, Chew EY *et al.*: Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720), 385-389 (2005).
3. Sullivan P: Don't give up on GWAS. *Mol Psychiatry* 17(1), 2-3 (2012).
4. Klein C LKZA: The promise and limitations of genome-wide association studies. *JAMA* 308(18), 1867-1868 (2012).
5. Psychiatric Gwas Consortium Coordinating Committee: Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* 166(5), 540-556 (2009).
6. Meehl PE: Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46, 806-834 (1978).
7. Waller NG: The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology* 11, 83-86 (2004).
8. Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* 4(9), 701-709 (2003).
9. Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433. (1918).
10. Goldstein DB: Common Genetic Variation and Human Traits. *New England Journal of Medicine* 360(17), 1696-1698 (2009).
11. Ballard DH, Cho J, Zhao HY: Comparisons of Multi-Marker Association Methods to Detect Association Between a Candidate Region and Disease. *Genet. Epidemiol.* 34(3), 201-212 (2009).
12. Bechhofer RE: A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* 25, 16-39 (1954).
13. Rosenthal R: *Cumulating evidence*. In: *A handbook for data analysis in the behavioral sciences: Methodological issues*, Keren G,Lewis C (Ed.^(Eds). Erlbaum, Hillsdale, NJ 519-559 (1993).
14. Morales JF, Song T, Auerbach AD, Wittkowski KM: Phenotyping genetic diseases using an extension of µ-scores for multivariate data. *Stat Appl Genet Mol* 7(1), 19 (2008).
15. Lawrence R, Evans DM, Morris AP *et al.*: Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. *Genome Research* 15(11), 1503-1510 (2005).
16. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164), 851-861 (2007).
17. Mann HB, Whitney DR: On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1), 50-60 (1947).
18. Hoeffding W: A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19, 293-325 (1948).
19. Li H: U-statistics in genetic association studies. *Hum Genet* 131(9), 1395-1401 (2012).
20. Pearson TA, Manolio TA: How to interpret a genome-wide association study. *JAMA* 299(11), 1335-1344 (2008).
21. Wittkowski KM, Song T, Anderson K, Daniels JE: U-Scores for Multivariate Data in Sports. *Journal of Quantitative Analysis in Sports* 4(3), 7 (2008).
22. Diana M, Song T, Wittkowski KM: Studying travel-related individual assessments and desires by combining hierarchically structured ordinal variables. *Transportation* 36(2), 187-206 (2009).
23. Hilton JF: The appropriateness of the Wilcoxon test in ordinal data. *Statistics in Medicine* 15(6), 631-645 (1996).
24. Loiseau P, Panayiotopoulos CP, Hirsch E: *Childhood Absence Epilepsy and Related Syndromes*. In: *Epilepsy Syndromes in Infancy, Childhood and Adolescence*, Roger J,Bureau M,Dravet C,Genton P,Tassinari CA,Wolf P (Ed.^(Eds). John Libbey, Montrouge, France 285-303 (2002).
25. Glauser TA, Cnaan A, Shinnar S *et al.*: Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy. *N Engl J Med* 362(9), 790-799 (2010).
26. Kim JE, Choi HC, Song HK *et al.*: Levetiracetam inhibits interleukin-1 beta inflammatory responses in the hippocampus and piriform cortex of epileptic rats. *Neurosci Lett* 471(2), 94-99 (2010).
27. Akaike H: A new look at statistical-model identification. *IEEE Trans. Autom. Control* AC19(6), 716-723 (1974).
28. Lehmann EL: Some model I problems of selection. *Annals of Mathematical Statistics* 32, 990-1012 (1961).
29. Friedman JI, Vrijenhoek T, Markx S *et al.*: CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. *Molecular Psychiatry* 13(3), 261-266 (2008).
30. Crunelli V, Leresche N: Childhood absence epilepsy: genes, channels, neurons and networks. *Nat Rev Neurosci* 3(5), 371-382 (2002).
31. Van Bokhoven H: Genetic and epigenetic networks in intellectual disabilities. *Annual review of genetics* 45, 81-104 (2011).
32. Oprica M, Eriksson C, Schultzberg M: Inflammatory mechanisms associated with brain damage induced by kainic acid with special reference to the interleukin-1 system. *J Cell Mol Med* 7(2), 127-140 (2003).

33. Yuan J, Wang L-Y, Li J-M *et al.*: Altered Expression of the Small Guanosine Triphosphatase RhoA in Human Temporal Lobe Epilepsy*. Journal of Molecular Neuroscience* 42(1), 53-58 (2010).

34. Lim J, Hao T, Shaw C *et al.*: A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration*. Cell* 125(4), 801-814 (2006).

35. Billuart P, Bienvenu T, Ronce N *et al.*: Oligophrenin-1 encodes a rhoGAP protein involved in X-linked mental retardation*. Nature* 392(6679), 923-926 (1998).

36. Gribaa M, Salih M, Anheim M *et al.*: A new form of childhood onset, autosomal recessive spinocerebellar ataxia and epilepsy is localized at 16q21-q23*. Brain* 130(7), 1921-1928 (2007).

37. Imbrici P, Jaffe SL, Eunson LH *et al.*: Dysfunction of the brain calcium channel Ca(V)2.1 in absence epilepsy and episodic ataxia*. Brain* 127, 2682-2692 (2004).

38. Tentler D, Gustavsson P, Leisti J *et al.*: Deletion including the oligophrenin-1 gene associated with enlarged cerebral ventricles, cerebellar hypoplasia, seizures and ataxia*. Eur J Hum Genet* 7(5), 541-548 (1999).

39. Bergmann C, Zerres K, Senderek J *et al.*: Oligophrenin 1 (OPHN1) gene mutation causes syndromic X-linked mental retardation with epilepsy, rostral ventricular enlargement and cerebellar hypoplasia*. Brain* 126(Pt 7), 1537-1544 (2003).

40. Hayashi T, Okabe T, Nasu-Nishimura Y *et al.*: PX-RICS, a novel splicing variant of RICS, is a main isoform expressed during neural development*. Genes to Cells* 12(8), 929-939 (2007).

41. Nakamura T, Hayashi T, Mimori-Kiyosue Y *et al.*: The PX-RICS-14-3-3 zeta/theta Complex Couples N-cadherin-beta-Catenin with Dynein-Dynactin to Mediate Its Export from the Endoplasmic Reticulum*. Journal of Biological Chemistry* 285(21), 16145-16154 (2010).

42. Wang K, Zhang H, Kugathasan S *et al.*: Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease*.* 84(3), 399-405 (2009).

43. Subramanian A, Tamayo P, Mootha V *et al.*: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. Proceedings of the National Academy of Sciences of the United States of America* 102(43), 15545-15550 (2005).

44. Ketzef M, Kahn J, Weissberg I, Becker AJ, Friedman A, Gitler D: Compensatory network alterations upon onset of epilepsy in synapsin triple knock-out mice*. Neuroscience* 189, 108-122 (2011).

45. Courtney KD, Corcoran RB, Engelman JA: The PI3K pathway as drug target in human cancer*. J Clin Oncol* 28(6), 1075-1083 (2010).

46. Harris SJ, Foster JG, Ward SG: PI3K isoforms as drug targets in inflammatory diseases: lessons from pharmacological and genetic strategies*. Curr Opin Investig Drugs* 10(11), 1151-1162 (2009).

47. Xue Y, Xie N, Cao L, Zhao X, Jiang H, Chi Z: Diazoxide preconditioning against seizure-induced oxidative injury is via the PI3K/Akt pathway in epileptic rat*. Neurosci Lett* 495(2), 130-134 (2011).

48. Nasu-Nishimura Y, Hayashi T, Ohishi T *et al.*: Role of the Rho GTPase-activating protein RICS in neurite outgrowth*. Genes Cells* 11(6), 607-614 (2006).

49. Schubbert S, Shannon K, Bollag G: Hyperactive Ras in developmental disorders and cancer*. Nat Rev Cancer* 7(4), 295-308 (2007).

50. Commission on Classification and Terminology of the International League against Epilepsy: Proposal for revised classification of epilepsies and epileptic syndromes*. Epilepsia* 30, 389-399 (1989).

51. Kosoy R, Nassir R, Tian C *et al.*: Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*. Hum Mutat* 30(1), 69-78 (2009).

52. Wittkowski KM: Friedman-type statistics and consistent multiple comparisons for unbalanced designs*. J Am Statist Assoc* 83, 1163-1170, Extension: 1992;1187: 258 (1988).

53. Devlin B, Roeder K: Genomic control for association studies*. Biometrics* 55(4), 997-1004 (1999).

54. Wittkowski KM, Song T: Nonparametric methods for molecular biology*. Methods Mol Biol* 620, 105-153 (2010).

55. Hajek J, Sidak Z: Theory of rank tests. Academic, New York, NY. (1967).