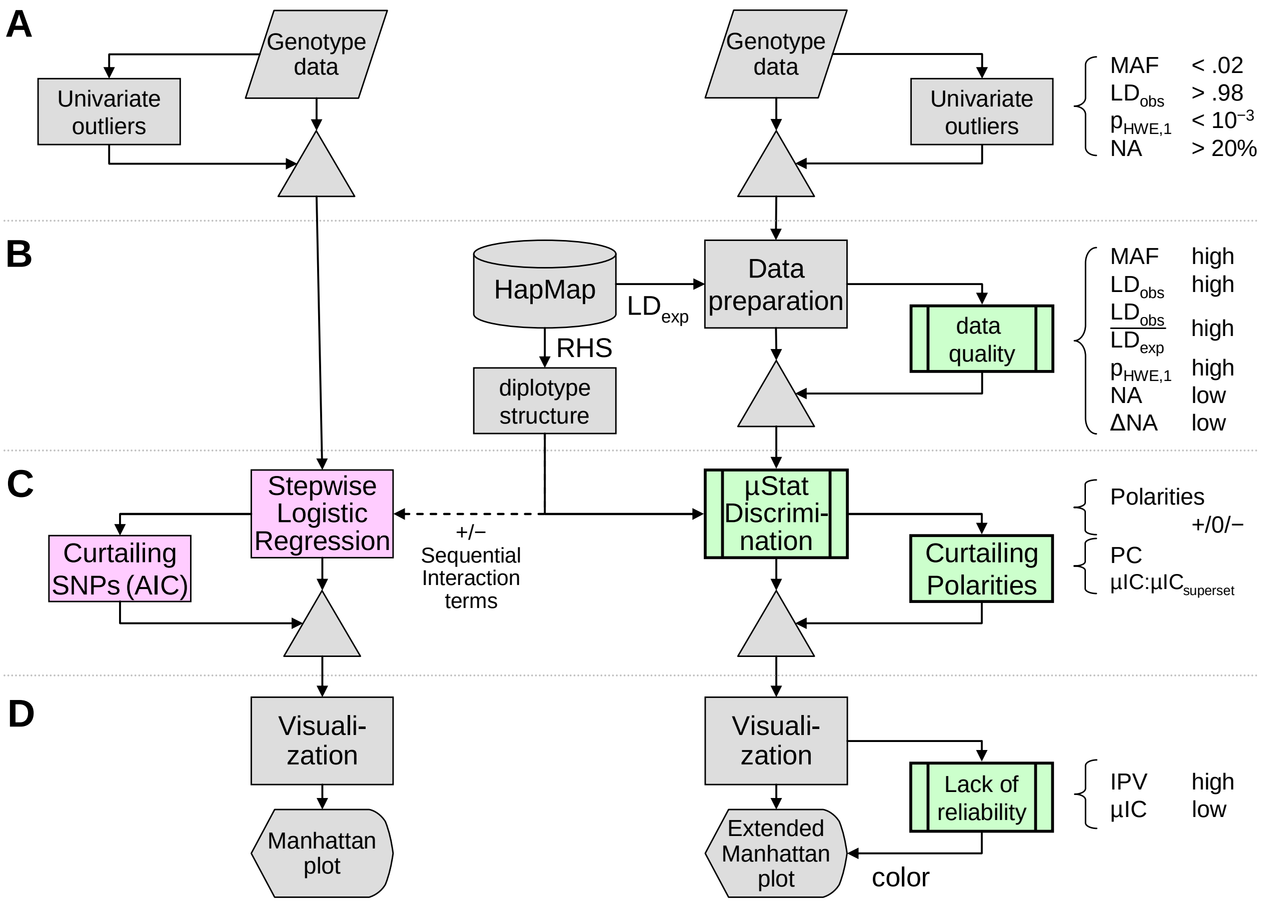
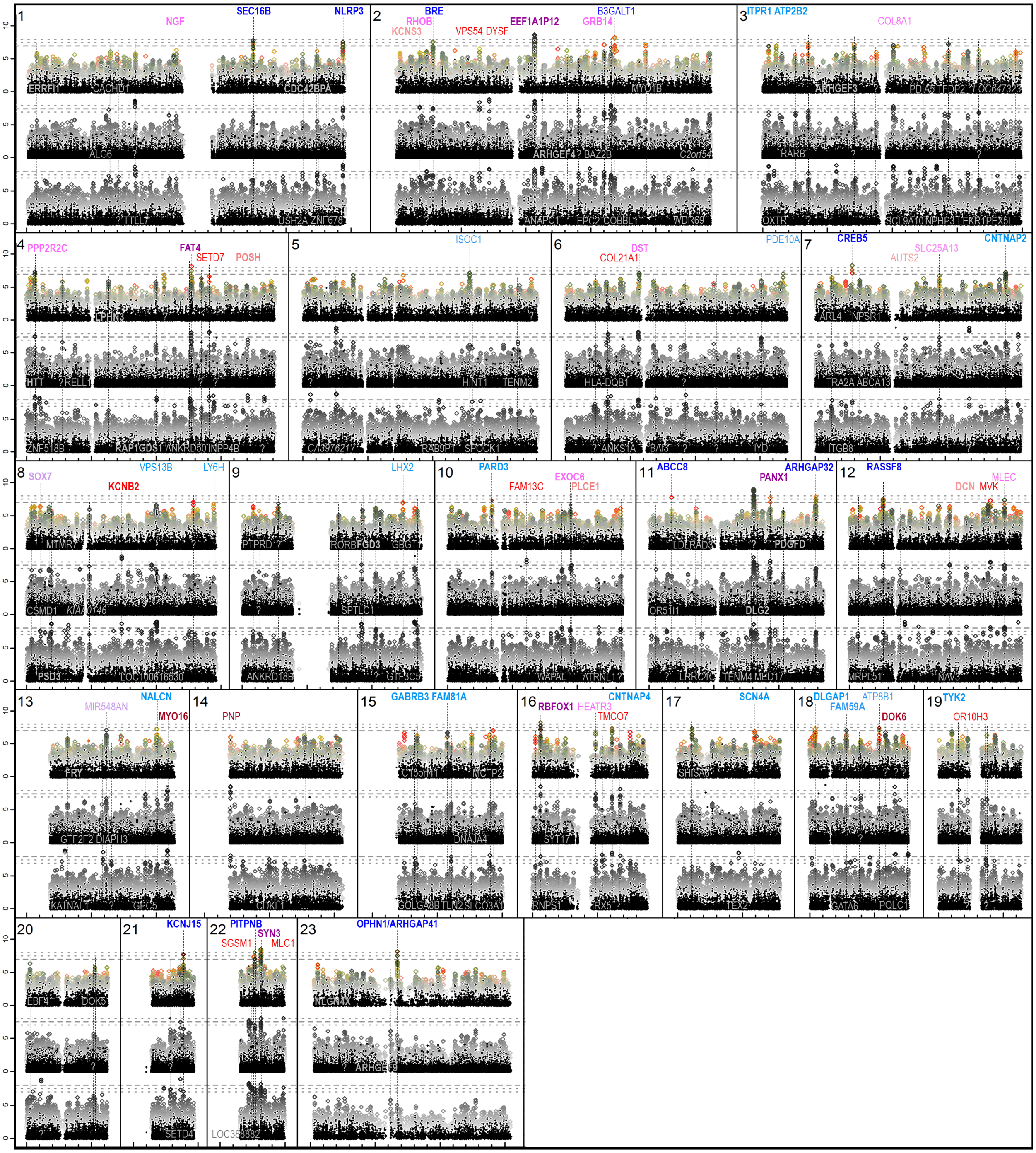
Supplementary Information



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 $\sqrt{3}$, 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).