Linkage-Disequilibrium-Based Binning Affects the Interpretation of GWASs

Andrea Christoforou,[1,2,16,*] Michael Dondrup,[3,16] Morten Mattingsdal,[4,5,16] Manuel Mattheisen,[6,7,8,9,16] Sudheer Giddaluru,[1,2] Markus M. Nöthen,[6,7,10] Marcella Rietschel,[11] Sven Cichon,[1,6,7,12] Srdjan Djurovic,[4,13,14] Ole A. Andreassen,[4,14] Inge Jonassen,[3,15] Vidar M. Steen,[1,2] Pål Puntervoll,[3] and Stéphanie Le Hellard[1,2]

Genome-wide association studies (GWASs) are critically dependent on detailed knowledge of the pattern of linkage disequilibrium (LD) in the human genome. GWASs generate lists of variants, usually SNPs, ranked according to the significance of their association to a trait. Downstream analyses generally focus on the gene or genes that are physically closest to these SNPs and ignore their LD profile with other SNPs. We have developed a flexible R package (LDsnpR) that efficiently assigns SNPs to genes on the basis of both their physical position and their pairwise LD with other SNPs. We used the positional-binning and LD-based-binning approaches to investigate whether including these "LD-based" SNPs would affect the interpretation of three published GWASs on bipolar affective disorder (BP) and of the imputed versions of two of these GWASs. We show how including LD can be important for interpreting and comparing GWASs. In the published, unimputed GWASs, LD-based binning effectively "recovered" 6.1%–8.3% of Ensembl-defined genes. It altered the ranks of the genes and resulted in nonnegligible differences between the lists of the top 2,000 genes emerging from the two binning approaches. It also improved the overall gene-based concordance between independent BP studies. In the imputed datasets, although the increases in coverage (>0.4%) and rank changes were more modest, even greater concordance between the studies was observed, attesting to the potential of LD-based binning on imputed data as well. Thus, ignoring LD can result in the misinterpretation of the GWAS findings and have an impact on subsequent genetic and functional studies.

Over the past decade, genome-wide association studies (GWASs) have revolutionized the analysis of human complex genetic traits. By scanning hundreds of thousands of genetic variants, typically SNPs, in hundreds or thousands of individuals, they search for the variant(s) that associate with a particular disease or trait. Critical to the development and evolution of GWASs has been the creation of the International HapMap Project,[1] which has cataloged the common patterns of human genetic variation, including the linkage disequilibrium (LD) between SNPs. Knowledge of this LD, or nonrandom association of alleles at multiple loci, has made it possible to identify informative subsets of SNPs (i.e., "tagging SNPs") that capture the bulk of genome-wide variation and has resulted in affordable genome-wide genotyping. To date, almost 1,000 GWASs have been published and have tested hundreds of human traits and reported thousands of significant associations (Catalog of Published Genome-Wide Association Studies[2]). Previously known associations have been confirmed, and new candidates have been implicated.[3] However, a general sense of disappointment lingers because GWASs have fallen short of the initial expectation that they would unravel the genetic basis of complex traits.[4,5] Recent analyses reveal that a large proportion of the "missing heritability"[5,6] can be explained by a polygenic model that considers all GWAS SNPs simultaneously,[7–9] but these studies provide no clues about the identity of the susceptibility variants or the underlying biology of the trait.[6] Thus, much attention has been given to uncovering and characterizing this "missing" or "hidden" heritability.[6,10]

In a conventional GWAS, each SNP is considered separately (the "single-marker" approach), resulting in a list of variants ranked according to the statistical significance of their association to the trait (i.e., their p value).[11] The "top hits" are typically reported, and the relevance of each finding, as well as the focus of future work, is primarily based on the functional unit(s), namely gene(s), implicated by the associated SNP. Furthermore, gene-based methods are increasingly being applied as complementary approaches to the analysis of GWAS data. These methods take the gene instead of the individual SNP as the basic unit of association and thus allow aggregation of SNPs of smaller effect, potentially increasing power and reducing

[1]Dr. Einar Martens Research Group for Biological Psychiatry, Department of Clinical Medicine, University of Bergen, 5021 Bergen, Norway; [2]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, 5021 Bergen, Norway; [3]Computational Biology Unit, Uni Computing, Uni Research, 5008 Bergen, Norway; [4]Institute of Clinical Medicine, University of Oslo, 0318 Oslo, Norway; [5]Research Unit, Sørlandet Hospital HF, 4604 Kristiansand, Norway; [6]Department of Genomics, Life and Brain Center, University of Bonn, 53127 Bonn, Germany; [7]Institute of Human Genetics, University of Bonn, 53127 Bonn, Germany; [8]Institute for Genomic Mathematics, University of Bonn, 53127 Bonn, Germany; [9]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA; [10]German Centre for Neurodegenerative Disorders, 53175 Bonn, Germany; [11]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Mannheim, 68159 Mannheim, Germany; [12]Structural and Functional Organization of the Brain, Institute of Neuroscience and Medicine, Research Center Jülich, 52425 Jülich, Germany; [13]Department of Medical Genetics, Oslo University Hospital, 0424 Oslo, Norway; [14]Division of Mental Health and Addiction, Oslo University Hospital, 0424 Oslo, Norway; [15]Department of Informatics, University of Bergen, 5008 Bergen, Norway
[16]These authors contributed equally to this work
*Correspondence: andrea.christoforou@med.uib.no
DOI 10.1016/j.ajhg.2012.02.025.

Table 1. Study Descriptions and Summary of Coverage for Positional-Binning and LD-Based-Binning Approaches for Original, Unimputed Datasets

	WTCCC[a]			TOP[b]			German[c]		
Sample size (cases/controls)	1,868/2,938			198/336			682/1,300		
Platform used	Affymetrix 500K			Affymetrix6.0			Illumina HumanHap550v3		
Number of post-QC SNPs for binning	468,648			615,396			511,978		
Binning data	**Positional binning**	**LD-based binning**	**Difference[d]**	**Positional binning**	**LD-based binning**	**Difference[d]**	**Positional binning**	**LD-based binning**	**Difference[d]**
Number of genes covered[e]	30,610 (83.4%)	33,443 (91.1%)	2,833 (9.3%)	31,823 (86.7%)	33,905 (92.4%)	2,082 (6.5%)	31,708 (86.4%)	33,861 (92.3%)	2,153 (6.8%)
Number of post-QC SNPs binned	237,869 (50.8%)	277,534 (59.2%)	39,665 (16.7%)	307,949 (50.0%)	363,570 (59.1%)	55,621 (18.1%)	272,914 (53.3%)	308,634 (60.2%)	35,720 (13.1%)
Number of SNPs binned to only 1 gene	199,752 (84.0%)	178,544 (64.3%)	21,208 (10.6%)	259,223 (84.2%)	234,036 (64.4%)	25,187 (9.7%)	228,098 (83.6%)	209,458 (67.9%)	18,640 (8.2%)
Number of SNPs binned to ten or more	135 (0.057%)	2,537 (0.91%)	2,402	174 (0.057%)	3,106 (0.85%)	2,932	141 (0.052%)	2,072 (0.67%)	1,931
Mean number of SNPs per bin (median)	9.4 (4)	15.2 (10)	6.6 (4)	11.7 (5)	19.4 (13)	8.4 (6)	10.5 (5)	15.4 (10)	5.6 (4)
Range (min–max)	1–514	1–515	0–87	1–687	1–701	0–112	1–655	1–665	0–64
Number of genes with only one SNP	4,830 (15.8%)	1,531 (4.6%)	3,299 (68.3%)	3,604 (11.3%)	992 (2.9%)	2,612 (72.5%)	3,647 (11.5%)	595 (1.8%)	3,052 (83.7%)

The following abbreviation is used: QC, quality control.
[a]The UK-based Wellcome Trust Case Control Consortium (WTCCC) BP GWAS.[17]
[b]The Norwegian Thematically Organized Psychosis (TOP) BP GWAS.[18]
[c]A German BP GWAS.[19]
[d]Percentages indicate percent increase or decrease from positional to LD-based binning.
[e]Ensembl 54 (May 2009) genes (total N = 36,693) tagged by at least one SNP.

the multiple-testing burden.[12–14] They enable the incorporation of biological knowledge for greater insight into the mechanisms underlying the trait and are essential for subsequent pathway-based approaches.[13] Gene-based methods also facilitate direct comparison of independent studies because they are unaffected by allelic heterogeneity and potential differences in SNP coverage and LD patterns.[15]

The success of both single-marker and gene-based approaches is critically dependent on the correct assignment of SNPs to genes. At the single-marker level, the aim is to identify the gene(s) that the associated SNP is tagging. At the gene level, the aim is to attribute all SNPs tagging a particular gene to that gene. Although LD can span hundreds of kilobases,[16,17] when GWAS results emerge, the SNPs of interest are typically assigned to the nearest gene or transcript within a specified distance.[14] In turn, genes are typically represented only by the SNPs that are physically located within the transcribed region or predefined flanking region.[13] It is not systematically taken into consideration that an associated SNP might be in high LD with another SNP (genotyped or not) located hundreds of kilobases away in a different gene or that a genotyped SNP positioned outside the defined boundaries of a gene is tagging that gene. Here, we show that ignoring LD discards valuable information and potentially leads to the incorrect localization of the association signal and might mislead the interpretation of GWAS data.

We have therefore developed a flexible R package (LDsnpR) that systematically assigns SNPs to genes (or relevant predefined genome "bins") by using SNP association results (e.g., p values), bin definitions, and precalculated pairwise LD data (e.g., r^2 values) provided by the user (Figure S1, available online). By default, LDsnpR assigns a SNP to a bin if that SNP is located within the physical boundaries of that bin (i.e., the "positional-binning" approach). Then, as a unique feature of this package, the user has the option of also assigning a genotyped SNP to a bin if that SNP is in high pairwise LD with another SNP (genotyped or not) located within the physical boundaries of that bin (i.e., "LD-based-binning" approach). Although a genotyped SNP cannot be assigned to a particular gene more than once, it can be assigned to more than one gene.

As proof of principal, we used LDsnpR to assess the impact of the LD-based-binning approach (versus the positional-binning approach) on the results of three published GWASs on bipolar disorder (BP), each unimputed and genotyped on a different platform. The three GWASs are (1) the UK-based Wellcome Trust Case Control Consortium (WTCCC) BP GWAS,[18] (2) the Norwegian Thematically Organized Psychosis (TOP) BP GWAS,[19] and (3) a German BP GWAS[20] (Table 1). Each GWAS had been previously

Table 2. Study Descriptions and Summary of Coverage for Positional-Binning and LD-Based-Binning Approaches for Imputed Datasets

	TOP[a] Imputed[b]			German[c] Imputed[b]		
Sample size (cases/controls)	198/336			657/1,308		
Imputation reference panel	HapMap Phase III (CEU)			1,000 Genomes (pilot 1, CEU) and HapMap Phase III (CEU)		
Post-QC SNPs for binning	992,161			4,825,148		
Binning data	**Positional binning**	**LD-based binning**	**Difference[d]**	**Positional binning**	**LD-based binning**	**Difference[d]**
Number of genes covered[e]	33,242 (90.6%)	34,193 (93.2%)	951 (2.9%)	32,116 (87.5%)	32,259 (87.9%)	143 (0.4%)
Number of post-QC SNPs binned	521,720 (52.6%)	612,316 (61.7%)	90,596 (17.4%)	2,394,441 (49.6%)	2,613,493 (54.2%)	219,052 (9.1%)
Number of SNPs binned to only one gene	431,808 (43.5%)	367,671 (37.1%)	64,137 (14.9%)	1,979,660 (41.0%)	1,855,413 (38.5%)	124,247 (6.3%)
Number of SNPs binned to ten or more	267 (0.03%)	7,967 (0.8%)	7,700	1,272 (0.03%)	16,807 (0.3%)	15,535
Mean number of SNPs per bin (median)	19.3 (9)	35.9 (25)	17.1 (12)	91.6 (44)	130.6 (84)	39.5 (26)
Range (min–max)	1–1,046	1–1,062	0–214	1–5,570	1–5,573	0–573
Number of genes with only one SNP	1,795 (5.4%)	651 (1.9%)	1,144 (63.7%)	241 (0.8%)	208 (0.6%)	33 (13.7%)

The following abbreviation is used: QC, quality control.
[a]The Norwegian Thematically Organized Psychosis (TOP) BP GWAS.[18]
[b]Imputation details: the Norwegian TOP dataset was imputed according to the ENIGMA protocol with the use of MACH imputation software[38] and HapMap Phase III (CEU) as the reference panel. The German dataset was imputed with IMPUTE2 software[39] and the 1,000 Genomes Project (Pilot 1, CEU) and HapMap Phase III (CEU) as reference panels.
[c]A German BP GWAS.[19]
[d]Percentages indicate percent increase or decrease from positional to LD-based binning.
[e]Ensembl 54 (May 2009) genes (total N = 36,693) tagged by at least one SNP.

approved by the relevant local research ethics committees, and all participants had provided written informed consent.[18–20] In addition, we assessed the impact of LD-based binning on imputed versions of the TOP and German GWASs, in which ungenotyped markers had been statistically inferred[11] on the basis of LD from different reference panels (i.e., HapMap Phase III for TOP; HapMap Phase III and 1,000 Genomes[21] for German) (Table 2).

BP is a severe complex psychiatric disorder that shows high heritability (60%–80%) but for which clear genetic risk factors remain elusive.[4] Although several GWASs on BP have been performed (Catalog of Published Genome-Wide Association Studies[2]), the findings have shown little overlap at both the SNP and gene levels. Also, only a handful of SNPs have achieved genome-wide significance ($<\sim 10^{-8}$), and these SNPs only explain less than 3% of the heritability,[4,22] suggesting that psychiatric disorders, such as BP, might be less amenable to GWASs than other disorders.[5,23] However, systematic LD-based gene binning has not been applied to these datasets, possibly contributing to the apparent lack of success. Thus, we assessed the effects of the LD-based-binning approach relative to the traditional positional-binning approach with respect to (1) gene coverage, (2) changes in the results and, potentially, the interpretation of findings, and (3) pairwise concordance of the findings among the BP GWASs.

In brief, for LDsnpR, gene bin definitions were based on the Human Ensembl release 54 (May 2009) gene identifiers with unambiguous positional information (N = 36,693). We extended these gene bins by another 10 kb on either side to best capture potential regulatory regions.[24,25] The LD data were based on HapMap Phase II release 27 and were restricted to that of the CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) sample. We set the pairwise LD at the widely accepted threshold of $r^2 \geq 0.8$[26] to limit the loss of power needed for the detection of association at the linked locus.[27]

We first compared the extent of coverage between the positional-binning and LD-based-binning approaches in the published, unimputed datasets (Table 1). By allowing us to identify the intergenic SNPs that tag genes, LD-based binning resulted in a ~13%–18% increase in the number of SNPs included in the gene-binning process. Intergenic SNPs represent ~40% of GWAS trait-associated SNPs.[3] Notably, LD-based binning "recovered" >2,000 genes (>6%) in all three datasets, increasing the proportion of Ensembl 54 genes tagged by at least one SNP from ~83% to >91%. Furthermore, there was an increase in the density of coverage; an average of 5.6 to 8.4 (median of four to six) SNPs were added per gene, and there was an overall decrease (>68%) in the number of genes tagged by only one SNP.

Table 3. Effect of LD-Based Binning on Ranks of Genes within Each GWAS

	WTCCC	TOP	German	TOP Imputed	German Imputed
Correlation[a] of gene ranks	0.79	0.83	0.83	0.83	0.92
Number of genes moving into top 2,000 with LD-based binning	681 (34.0%)	601 (30.0%)	538 (26.9%)	558 (27.9%)	309 (15.5%)

[a]Spearman rank correlation (i.e., rho).

The imputed datasets also yielded increased coverage (Table 2) but, as expected, to a lesser extent depending on the reference panel used for imputation. Although HapMap II (i.e., LDsnpR reference panel) is denser than HapMap III[28] (i.e., reference panel for the TOP and German studies), imputation on the 1,000 Genomes data (i.e., reference panel for the German study) potentially gives the densest coverage. For the TOP and German imputed datasets, LD-based binning resulted in an increase of 17.4% and 9.1%, respectively, in the number of SNPs included in the gene-binning process and the recovery of 951 (2.9%) and 143 (0.4%) genes, respectively. Although this is only a small proportion of the total gene coverage, the recovery of these genes enables them to be considered as candidates for BP association and might lead to a better understanding of the biology should the true association stem from them. Also of note, in the German GWAS, LD-based binning alone achieved an overall gene coverage of 92.3% (imputation achieved 87.5% coverage, and imputation combined with LD-based binning achieved 87.9% coverage), suggesting that under some scenarios, LD-based binning alone can offer the most coverage. As with the original GWASs, there was an increase in the density of coverage; an average of 17.1 and 39.5 (median 12 and 26) SNPs were added per gene for the TOP and German imputed datasets, respectively. There was also a decrease in the number of genes tagged by only one SNP (63.7%); the decrease was not as notable for the German imputed dataset (13.7%).

We next assessed the effects of the LD-based-binning approach on the results of the three GWASs at both the single-marker and gene levels. At the single-marker level, we used the positional-binning and LD-based-binning approaches to compare the genes tagged by the most significant SNPs reported in the original publications[18–20] (Table S1). Although LD-based binning made no difference to the results of the TOP BP study, three of the 14 SNPs in the WTCCC BP study and three of the eight SNPs in the German BP study implicated additional or alternative genes. Interpreting GWAS single-marker results demands fastidious consideration because when given only the p value, it is not immediately clear where the true source of the association originates[17] and thus which is the true candidate gene. The overall potential for mislocalizing the association signal was underscored by the reduced number of SNPs tagging only one gene and the increased number of SNPs tagging ten or more genes after LD-based binning (Tables 1 and 2). Further investigations, such as expression studies,[20] are therefore warranted before attributing putative causality to a gene and, as a result, nominating it as the focus of future fine-mapping, functional, and other expensive and time-consuming follow-up studies.[29]

As previously stated, gene-based analyses are ideal for pathway approaches, which aid in the interpretation of GWAS results by exploiting prior biological annotation to determine whether certain biological functions are enriched (i.e., overrepresented) among the more significant genes in a dataset. These methods require one measure of association (or score) for each gene on the basis of the individual SNP association signals. Here, we used a function in LDsnpR to score each gene with the most significant p value (i.e., the minimum p value approach), which was adjusted for the number of SNPs tagging that gene by a modification of Sidak's correction.[30] The minimum p value approach is the most widely used gene-scoring approach[31] and assumes an underlying genetic architecture in which a single SNP, or locus, within the gene contributes to the disorder. The modification performs at least as well as a powerful regression-based method in correcting for the bias due to SNP number.[32] In this study, the correlation between the gene score and the number of SNPs in the bin was reduced from Pearson $r^2 > 0.30$ to $r^2 < 0.020$ in all three datasets after the modified Sidak correction was applied. Also, permutation-based gene-set analysis, as implemented in PLINK,[33] on the German GWAS confirmed the high correlation between modified Sidak-corrected p values and permutation-based p values ($r^2 > 0.95$). The genes were scored for both the positional-binning and LD-based-binning approaches and were compared.

The overall correlation in the ranks of the genes between the two approaches was <0.83 in the three original datasets and the TOP imputed dataset, indicating that LD-based binning altered the scores and the subsequent ranks of the genes. Although not as large, changes in rank were also observed in the German imputed dataset (Table 3). When a resampling analysis was performed on the unimputed WTCCC dataset (it randomly excluded 5% of the samples [20 repetitions]), the average overall correlation in ranks due to LD-based binning (0.80) was lower than that resulting from random fluctuations in the datasets (>0.87), indicating greater changes due to LD (Table S2). Such changes in rank are likely to impact threshold-free, rank-based pathway approaches, such as gene-set-enrichment analysis,[34] which aims to determine whether a predefined set of genes is enriched at the top of a ranked list. By inspecting the top 2,000 genes emerging from the two binning approaches, we found a 27%–34% difference between the two gene lists in the three unimputed and the TOP imputed datasets and a 15.5% difference in the German imputed dataset. Here, the resampling analysis in the WTCCC GWAS found that random fluctuations in

Table 4. Pairwise Concordance between GWASs at SNP and Gene Levels

	WTCCC vs. TOP	WTCCC vs. German	TOP vs. German	TOP Imputed vs. German Imputed
SNP level	0.0066 (0.00018)	0.0037 (0.31)	−0.0018 (0.51)	−0.00023 (0.83)
Gene level (positional binning)	0.030 (1.78×10^{-7})	−0.0017 (0.78)	0.023 (4.78×10^{-5})	0.068 ($<2.2 \times 10^{-16}$)
Gene level (LD-based binning)	0.077 ($<2.2 \times 10^{-16}$)	0.027 (7.24×10^{-7})	0.053 ($<2.2 \times 10^{-16}$)	0.098 ($<2.2 \times 10^{-16}$)

The Spearman rank correlation and p value (in parentheses) are shown for each pairwise comparison.

the dataset led to a 25.6% change in the top 2,000 genes, whereas LD-based binning resulted in a 30.7% difference (Table S2). For threshold-based approaches, such as Ingenuity Pathway Analysis and ALIGATOR,[35] in which a list of genes meeting a specified threshold is tested for overrepresentation of a particular biological function, LD-based binning could result in the submission of a substantially different list. Changes in the ranks of the genes are thus likely to impact the outcome of these analyses and possibly the overall biological interpretation of the findings. The extent to which these LD-based changes are meaningful will also depend on the study design and resulting power, given that the resampling analysis shows that substantial changes in results can also occur as a result of slight changes in the dataset.

Finally, we assessed whether LD-based binning improved the concordance of results across studies, especially in light of the aforementioned changes in the ranks of the genes. We compared the positional-binning and LD-based-binning approaches by performing pairwise rank-correlation analyses of the three GWAS datasets at both the SNP level and the gene level (Table 4). When the positional-binning approach was used, little to no correlation was observed at both the SNP and gene levels. However, with LD-based binning, the overall rank correlation increased by ~3% and was more significant for all pairwise comparisons, including the imputed datasets. Interestingly, the greatest concordance was observed when LD-based binning was combined with imputation, highlighting the complementary nature of the two methods. Although there was no obvious increase in overlap in the top gene hits (data not shown), this increase in overall concordance warrants the use of the LD-based-binning approach for the reanalysis of these and other datasets in the search for common functional gene sets and pathways. The observed increase in correlation persisted even when regions of high LD, such as the MHC (major histocompatibility complex) region on chromosome 6, were excluded (data not shown).

Our study illustrates the importance of systematically accounting for LD in the interpretation of GWAS results. To the best of our knowledge, our study is the first to quantify the added value of LD-based binning; in particular, it shows an increase in the concordance of results across independent GWASs of a trait as complex as BP. Excluding LD defies the basic premise of the GWAS approach by discarding valuable genetic information and risking the incorrect localization of the association signal and the misinterpretation of the biology of the findings. Our findings call for a reanalysis of previously published GWAS data via the LD-based-binning approach and for future GWASs to adopt this method automatically. LDsnpR facilitates this process by efficiently assigning SNPs to genes and provides the option of scoring the genes for direct entry into pathway-analysis tools. LDsnpR's flexible framework allows the application of different gene-scoring methods; the application of such methods is necessary for detecting gene-based associations under different genetic architectures for the traits.[31] The user-definable r^2 parameter enables the scanning of a greater range of allele frequencies at the linked locus.[27] Bin definitions and precalculated pairwise LD information can be updated on the basis of the user's interests and the information available. LD-based binning might also serve as a complementary and/or alternative approach to imputation. In particular, as high-quality LD data from the 1,000 Genomes Project[21] emerges, all GWASs, including those previously subjected to imputation, might benefit from simple and efficient LD-based binning at no extra cost. As we show here, LD-based binning can further enhance imputed GWASs, albeit to a lesser extent than unimputed datasets. More tools that allow for incorporation of LD into the interpretation of GWAS data are emerging,[36,37] further testifying to the importance of this approach. Also, for studies genotyped on different platforms and/or imputed with the use of different reference panels, LD-based binning enables uniform comparison at both the gene and pathway levels.

It is crucial to note that our study, as well as LDsnpR, only addresses SNP-to-gene assignment. Issues involving the derivation of the most accurate gene score (which accounts for gene size and LD between SNPs), the handling of SNPs that are assigned to multiple, possibly overlapping, genes, and the correlation between genes are unresolved obstacles for pathway-analysis approaches[13] and are beyond the scope of this paper. Furthermore, the benefits of LD-based binning will be unique to each GWAS depending on the trait and its true underlying genetic architecture, the study design, and the extent of SNP coverage.

Supplemental Data

Supplemental Data include one figure and two tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

We acknowledge Isabel Hanson Scientific Writing for critical help with the manuscript preparation. This work was supported by grants from the Bergen Research Foundation, the University of Bergen, the Research Council of Norway (FUGE, Psyksik Helse, and eVita), UNI Computing, Western Norway Regional Health Authority (Helse Vest), the Dr. Einar Martens Fund, South-Eastern Norway Regional Health Authority (Helse Sør-Øst), the National Institutes of Health and the National Heart, Lung, and Blood Institute (U01 HL089856, RO1 MH087590 and R01 MH081862), and the German Federal Ministry of Education and Research (National Genome Research Network 2, the National Genome Research Network plus, and the Integrated Genome Research Network MooDS [grant 01GS08144 to S.C.]). LDsnpR was developed within the eSysbio project. We acknowledge Håkon Sagehaug for contributing Java code and members of the BioStar QA community for their help and interesting discussions.

Received: September 6, 2011
Revised: February 16, 2012
Accepted: February 27, 2012
Published online: March 22, 2012

Web Resources

The URLs for data presented herein are as follows:

1,000 Genomes Project, http://www.1000genomes.org/
Catalog of Published Genome-Wide Association Studies, www.genome.gov/gwastudies/
ENIGMA protocol, http://enigma.loni.ucla.edu/protocols/genetics-protocols/
HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
Human Ensembl Release 54, http://may2009.archive.ensembl.org/biomart/martview/11839bb5ec82fb10bf0333540fa09c46
IMPUTE2 Software, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Ingenuity Pathway Analysis, http://www.ingenuity.com/
LDsnpR, http://services.cbu.uib.no/software/ldsnpr
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
R Archive Network, http://cran.r-project.org

References

1. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
2. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
3. Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. *363*, 166–176.
4. Bondy, B. (2011). Genetics in psychiatry: Are the promises met? World J. Biol. Psychiatry *12*, 81–88.
5. Gershon, E.S., Alliey-Rodriguez, N., and Liu, C. (2011). After GWAS: Searching for genetic risk for schizophrenia and bipolar disorder. Am. J. Psychiatry *168*, 253–256.
6. Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367–383.
7. Gibson, G. (2010). Hints of hidden heritability in GWAS. Nat. Genet. *42*, 558–560.
8. Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., et al. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Mol. Psychiatry *16*, 996–1005.
9. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
10. Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am. J. Hum. Genet. *86*, 6–22.
11. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
12. Bergen, S.E., Balhara, Y.P., Christoforou, A., Cole, J., Degenhardt, F., Dempster, E., Fatjó-Vilas, M., Khedr, Y., Lopez, L.M., Lysenko, L., et al. (2011). Summaries from the XVIII World Congress of Psychiatric Genetics, Athens, Greece, 3-7 October 2010. Psychiatr. Genet. *21*, 136–172.
13. Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. *11*, 843–854.
14. Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. *81*, 1278–1283.
15. Neale, B.M., and Sham, P.C. (2004). The future of association studies: Gene-based analysis and replication. Am. J. Hum. Genet. *75*, 353–362.
16. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.
17. Lawrence, R., Evans, D.M., Morris, A.P., Ke, X., Hunt, S., Paolucci, M., Ragoussis, J., Deloukas, P., Bentley, D., and Cardon, L.R. (2005). Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. Genome Res. *15*, 1503–1510.
18. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
19. Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M., Agartz, I., Lorentzen, S., Melle, I., Morken, G., and Andreassen, O.A. (2010). A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. J. Affect. Disord. *126*, 312–316.
20. Cichon, S., Mühleisen, T.W., Degenhardt, F.A., Mattheisen, M., Miró, X., Strohmaier, J., Steffens, M., Meesters, C., Herms, S., Weingarten, M., et al; Bipolar Disorder Genome Study (BiGS) Consortium. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. Am. J. Hum. Genet. *88*, 372–381.
21. 1000 Genomes Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
22. So, H.C., Gui, A.H., Cherny, S.S., and Sham, P.C. (2011). Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. Genet. Epidemiol. *35*, 310–317.

23. Neale, B.M., and Purcell, S. (2008). The positives, protocols, and perils of genome-wide association. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *147B*, 1288–1294.
24. Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806–810.
25. Vandiedonck, C., Taylor, M.S., Lockstone, H.E., Plant, K., Taylor, J.M., Durrant, C., Broxholme, J., Fairfax, B.P., and Knight, J.C. (2011). Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. Genome Res. *21*, 1042–1054.
26. Spencer, C.C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. *5*, e1000477.
27. Wray, N.R. (2005). Allele frequencies and the r2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. Twin Res. Hum. Genet. *8*, 87–94.
28. Santos, P.S., Höhne, J., Poerner, F., da Graça Bicalho, M., Uchanska-Ziegler, B., and Ziegler, A. (2011). Does the new HapMap throw the baby out with the bath water? Eur. J. Hum. Genet. *19*, 733–734.
29. Ioannidis, J.P., Thomas, G., and Daly, M.J. (2009). Validating, augmenting and refining genome-wide association signals. Nat. Rev. Genet. *10*, 318–329.
30. Saccone, S.F., Hinrichs, A.L., Saccone, N.L., Chase, G.A., Konvicka, K., Madden, P.A., Breslau, N., Johnson, E.O., Hatsukami, D., Pomerleau, O., et al. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum. Mol. Genet. *16*, 36–49.
31. Lehne, B., Lewis, C.M., and Schlitt, T. (2011). From SNPs to genes: disease association at the gene level. PLoS ONE *6*, e20133.
32. Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J., and Altshuler, D.; DIAGRAM Consortium; MAGIC investigators. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. *6*, e1001058.
33. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
34. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
35. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., and Craddock, N.; Wellcome Trust Case-Control Consortium. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. *85*, 13–24.
36. Hong, M.G., Pawitan, Y., Magnusson, P.K., and Prince, J.A. (2009). Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum. Genet. *126*, 289–301.
37. Zhang, K., Chang, S., Cui, S., Guo, L., Zhang, L., and Wang, J. (2011). ICSNPathway: Identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. Nucleic Acids Res. *39* (Web Server issue), W437–443.
38. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
39. Howie, B.N., Donnelly, P., and Marchini, J.A. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.