RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Importance of Different Types of Prior Knowledge in Selecting Genome-Wide Findings for Follow-Up

Cosetta Minelli,[1†*] Alessandro De Grandi,[1†] Christian X. Weichenberger,[1] Martin Gögele,[1] Mirko Modenese,[1] John Attia,[2,3] Jennifer H. Barrett,[4] Michael Boehnke,[5] Giuseppe Borsani,[6] Giorgio Casari,[7] Caroline S. Fox,[8,9] Thomas Freina,[1] Andrew A. Hicks,[1] Fabio Marroni,[10] Giovanni Parmigiani,[11,12] Andrea Pastore,[13] Cristian Pattaro,[1] Arne Pfeufer,[14] Fabrizio Ruggeri,[15] Christine Schwienbacher,[1] Daniel Taliun,[1] Peter P. Pramstaller,[1,16,17] Francisco S. Domingues,[1] and John R. Thompson[18]

[1]Center for Biomedicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy; [2]Centre for Clinical Epidemiology and Biostatistics, The University of Newcastle, Hunter Medical Research Institute, Newcastle, New South Wales, Australia; [3]Department of General Medicine, John Hunter Hospital, Newcastle, New South Wales, Australia; [4]Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, United Kingdom; [5]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan; [6]Department of Biomedical Sciences and Biotechnology, University of Brescia, Brescia, Italy; [7]Vita-Salute San Raffaele University and Center for Translational Genomics and Bioinformatics, Milan, Italy; [8]The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts; [9]Center for Population Studies, Framingham, Massachusetts; [10]Institute of Applied Genomics, Udine, Italy; [11]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts; [12]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; [13]Department of Economics, Ca' Foscari Venezia University, Venezia, Italy; [14]Department of Bioinformatics and Systems Biology IBIS, Helmholtz Zentrum Munich, German Research Center for Environmental Health (GmbH), Neuherberg, Germany; [15]National Research Council, Institute of Applied Mathematics and Information Technology, Milan, Italy; [16]Department of Neurology, Central Hospital, Bolzano, Italy; [17]Department of Neurology, University of Lübeck, Lübeck, Germany; [18]Department of Health Sciences, University of Leicester, Leicester, United Kingdom

Received 10 August 2012; Revised 28 October 2012; accepted revised manuscript 22 November 2012.
Published online 5 January 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21705

ABSTRACT: **Biological plausibility and other prior information could help select genome-wide association (GWA) findings for further follow-up, but there is no consensus on which types of knowledge should be considered or how to weight them. We used experts' opinions and empirical evidence to estimate the relative importance of 15 types of information at the single-nucleotide polymorphism (SNP) and gene levels. Opinions were elicited from 10 experts using a two-round Delphi survey. Empirical evidence was obtained by comparing the frequency of each type of characteristic in SNPs established as being associated with seven disease traits through GWA meta-analysis and independent replication, with the corresponding frequency in a randomly selected set of SNPs. SNP and gene characteristics were retrieved using a specially developed bioinformatics tool. Both the expert and the empirical evidence rated previous association in a meta-analysis or more than one study as conferring the highest relative probability of true association, whereas previous association in a single study ranked much lower. High relative probabilities were also observed for location in a functional protein domain, although location in a region evolutionarily conserved in vertebrates was ranked high by the data but not by the experts. Our empirical evidence did not support the importance attributed by the experts to whether the gene encodes a protein in a pathway or shows interactions relevant to the trait. Our findings provide insight into the selection and weighting of different types of knowledge in SNP or gene prioritization, and point to areas requiring further research.**
Genet Epidemiol 37:205–213, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: gene prioritization; genome-wide association studies; bioinformatics databases

Introduction

In genome-wide association (GWA) studies, the choice of which single-nucleotide polymorphisms (SNPs) should be followed up for replication in independent samples or for functional investigation can either be based purely on discovery *P*-values or can incorporate prior knowledge about the SNP and its possible association with the phenotype of interest. Selection of SNPs for replication based purely on discovery *P* values is currently the most common approach [Gögele et al., 2012], but this strategy tends to have low power to identify good candidates when the discovery sample is relatively small, particularly for SNPs with low minor allele frequency [Liu et al., 2008]. Despite current efforts to increase power by pooling GWA data from different studies, small discovery sample size can still be a critical issue for rarer disease outcomes or phenotypes that are difficult to measure, and the presence of heterogeneity across studies can further reduce

Supporting Information is available in the online issue at wileyonlinelibrary.com.

†Joint first authors.

*Correspondence to: Cosetta Minelli, Center for Biomedicine, EURAC Research, Viale Druso, 1—39100 Bolzano, Italy. E-mail: cosetta.minelli@eurac.edu

statistical power [Greene et al., 2009; Ioannidis, 2007; Kraft et al., 2009]. Incorporating prior information on biological gene function or findings from previous genetic epidemiological studies can help select the most promising SNPs in a more informed way, thus, potentially increasing the yield of downstream studies [Moreau and Tranchevent, 2012]. Such information may derive from very different sources, including gene expression and proteomics studies, genetic studies in animal models, and previous association or linkage studies in humans.

In practice, prior knowledge has been used to aid gene prioritization in many different ways. Sometimes investigators add to the list of GWA top hits sent to replication additional SNPs within genes known by the authors to have been previously linked to the phenotype [Gögele et al., 2012], although this leaves the reader in doubt about whether other SNPs with even higher support from prior knowledge might have been omitted. Other authors have used more systematic ways of identifying relevant prior information for SNP selection, either focusing on a single type of evidence, such as pathway analysis, or combining different types of evidence [Cantor et al., 2010; Saccone et al., 2008; Thomas et al., 2009]. Lack of evidence on the relative informativeness of different types of prior knowledge means that decisions on what information is worth retrieving and how much weight should be attributed to different types of knowledge are inevitably subjective. This work provides suitable weights for estimating the likelihood of true association given certain types of prior knowledge, and contrasts the views of experts with empirical evidence taken from published GWA meta-analyses.

Material and Methods

Elicitation of Experts' Opinions

Ten experts in the field of GWA investigations from different backgrounds (molecular biology, genetic epidemiology, statistical genetics) were asked to participate in the study, without being told the identity of the other experts [Akins et al., 2005]. Upon acceptance, a two-round Delphi survey was used to elicit their opinions through pre-prepared questionnaires circulated by e-mail. The Delphi method is a form of structured group communication process, consisting of an expert survey organized in two (or more) rounds [Adler and Ziglio, 1996]. In the second round, the anonymous results for all experts in the previous stage were given as feedback, and the same experts were asked to reassess their answers to the same set of questions in the light of their colleagues' opinions. These questions referred either to the specific SNP or to the gene(s) lying within 5 kb of the SNP. The questions did not refer to any specific phenotype, and the experts were asked to think in general terms.

In the first round, experts were presented with a list of 20 items (Supporting information Table SI), and were asked to provide their "best guess" on how many times more likely a SNP was to be truly associated with the phenotype given a certain characteristic, when compared with a SNP with no such characteristic (hereafter referred to as the "relative probability" associated with that characteristic). Experts were asked to answer as if each type of evidence were the only external information available, so that all types of evidence were treated as independent. To help ensure consistent interpretation of the scale, the experts were provided with an example for which the probability that a random SNP was truly associated with the disease was around 1 in 10,000, so that an answer of five times more likely would translate into a probability of true association of 5 in 10,000.

In the second round, the number of questions was reduced based on findings from the first round (see section "Statistical Analyses") and the experts were asked to provide a revised answer to each question, together with a 95% interval representing their uncertainty, reflecting both their own experience and the results of the first round averaged across experts (Supporting information Table SI). Estimates based on the experts' opinions are hereafter referred to as "opinions."

Empirical Evidence

Estimates of Empirical Relative Probabilities

We obtained empirical estimates of the relative probability of association for a SNP given a certain type of prior knowledge using a "case-control" approach. We chose seven disease traits for which a set of SNPs, referred to as "true SNPs", had been identified through large GWA investigations and had been replicated (Table 1). With "true SNPs" representing our "cases" and a set of 1,000 random SNPs as our "controls," we estimated relative probabilities of association by comparing the proportion of true SNPs for which a certain type of evidence was present vs. the proportion in random SNPs (see section "Statistical Analyses"). A different set of 1,000 random SNPs was selected for each trait throughout the genome, from about 2,500,000 SNPs with minor allele frequency greater than 0.01 (the sampling frame was that of all SNPs used in the estimated glomerular filtration rate, eGFRcrea, meta-analysis [Köttgen et al., 2010] and can be found at: https://intramural.nhlbi.nih.gov/labs/CF/Pages/CKDGenConsortium.aspx). Because true and random SNPs were not matched by allele frequency, we also performed a sensitivity analysis adjusting for allele frequency.

The seven selected traits were estimated glomerular filtration rate (as a measure of renal dysfunction), Crohn's disease, coronary artery disease, rheumatoid arthritis, primary biliary cirrhosis, type 2 diabetes, and body mass index (as a measure of obesity). For each trait, the list of true SNPs was compiled based on the most recent (one or more) GWA meta-analysis (Table 1), after excluding SNPs that had been selected for replication based on prior knowledge rather than GWA evidence. Estimates based on empirical evidence are hereafter referred to as "data."

To enable a comparison with the relative probabilities given by the experts, the data-based results for each SNP characteristic were also considered independently. However, the empirical evidence also enables us to investigate the dependence between the questions, and to determine the weights that

Table 1. Data source and number of "true SNPs" for the seven disease traits used to estimate empirical relative probabilities. "True SNPs" were single-nucleotide polymorphisms established as being associated with the seven traits through genome-wide association meta-analysis and independent replication. MeSH terms are those used for the bioinformatics retrieval of evidence on the relationships between the SNPs and the phenotype

	Identification of "true SNPs"		
Trait	Data source	Number	MESH terms used in bioinformatics tool
eGFR	Köttgen et al., 2010	28	Kidney diseases; kidney failure; kidney failure, acute; kidney failure, chronic; renal insufficiency, acute; renal insufficiency, chronic
Crohn's disease	Franke (2010) Franke et al., 2010	71	Crohn's disease
Coronary artery disease	Schunkert et al., 2011; Coronary Artery Disease (C4D) Genetics Consortium, 2011	30	Coronary artery disease; angina pectoris; coronary disease; coronary restenosis; coronary stenosis; coronary thrombosis; myocardial infarction
Rheumatoid arthritis	Stahl et al., 2010	18	Arthritis, rheumatoid; arthritis, juvenile rheumatoid
Primary biliary cirrhosis	Mells et al., 2011	18	Liver cirrhosis, biliary
Type 2 diabetes	Voight et al., 2010	26	Diabetes mellitus, type 2
BMI	Speliotes et al., 2010	32	Obesity

eGFR, estimated glomerular filtration rate; BMI, body mass index.

should be used in a combined estimate that incorporates all of the SNP characteristics, as shown in the companion paper by Thompson et al. [2013] (see section Discussion).

Bioinformatics Retrieval of Information

Information on each of the types of prior knowledge was retrieved for both the true SNPs and for 1,000 random SNPs in a standardized and automatic way, by use of a bioinformatics tool developed for this project. Ensembl [Flicek et al., 2011; Stabenau et al., 2004] was the main data source queried by the tool, but additional public databases were used to answer specific questions, in particular HuGE (Human Genome Epidemiology) Navigator [Yu et al., 2008], Pfam (Protein family database) [Finn et al., 2010], cisRED (cis-regulatory element database) [Robertson et al., 2006], VISTA Enhancer Browser [Visel et al., 2007], miRanda (microRNA Target Detection Software) [John et al., 2004], Mouse Genome Informatics (MGI) database [Blake et al., 2011], BioGPS (gene annotation portal) [Su et al., 2004; Wu et al., 2009], KEGG (Kyoto Encyclopedia of Genes and Genomes) [Kanehisa and Goto, 2000], and IntAct molecular interaction database [Aranda et al., 2010]. Supporting information Table SII summarizes the structure of each query, and the source code is available on our website at https://gemex.eurac.edu/downloads/stats/GenEpi2012.

To allow retrieval of evidence using these databases, the formulation of the query had, in a few cases, to be modified slightly from the initial question presented to the experts (Table 2). All types of evidence regarding relationships between genes and phenotypes were retrieved using MeSH (Medical Subject Headings) terms linked to UMLS CUIs (Unified Medical Language System Concept Unique Identifier) directly referring to the particular phenotype (Table 1; Supporting information Table SII), while questions to the experts had been phrased more generally as "the same/closely related phenotype." The potential impact of this difference was assessed in sensitivity analyses performed on three of seven traits by repeating the retrieval of evidence after extending the list of UMLS CUI terms to cover "closely related" phenotypes. One important difference was related to evidence from previous genetic association studies (Q8, Q9, and Q11). Because HuGE Navigator only provides information on whether a gene-phenotype association has been investigated and not on whether it has been established as true, the search of HuGE Navigator includes publications with negative findings [Yu et al., 2008]. This problem is typical of all search engines based on text mining of published literature, and we are not aware of any alternative public resource, covering both candidate-gene and GWA studies, which also provides the result of each investigated association. Formulation of the question on "functional models" (Q12) also had to be modified, from "evidence from in vitro and animal studies" in the question to experts to "evidence from mouse models (MGI database)" in the bioinformatics query, because we could not find public databases from which we could retrieve the required genome-wide information from other functional models. Finally, the question on the importance of whether the SNP is in a gene which shows gene/protein (gene-gene, gene-protein, or protein-protein) interactions relevant to the phenotype (Q15) was restricted to protein-protein interactions in the bioinformatics query.

We could not obtain by automated methods empirical evidence on the relative probability of association given supporting knowledge from linkage studies (Q10: "The SNP is in a gene (±5 kb) which is under a linkage peak that has been associated with the same/closely related phenotype") due to the lack of electronically processable public databases summarizing published genome-wide linkage findings.

Statistical Analyses

For experts' opinions, correlations between items of the original questionnaire in the first Delphi round were analyzed to help reduce the list of types of evidence to be evaluated, by dropping questions that appeared not to convey much additional information.

We estimated empirical relative probabilities as odds ratios of true association with logistic regression analysis, modeling the probability of being a "true SNP" given the presence of each type of evidence.

For both opinions and data, relative probabilities were analyzed after log transformation. Inverse variance meta-analysis based on either a fixed- or random-effect model, depending on absence or presence of heterogeneity, respectively, was used to pool opinions across experts and data across traits.

Table 2. Pooled relative probabilities of association across the nine experts for experts' opinion and across the seven traits for empirical evidence. Example of how questions were formulated (Q13): "How many times more likely to be associated with the phenotype is a SNP in a gene (±5 kb) which is highly expressed in a tissue relevant to the phenotype compared with a SNP which is not?" Frequencies for each type of evidence were calculated across all seven traits for "True SNPs", and across all 7,000 single-nucleotide polymorphisms for "Random SNPs"

Type of prior knowledge		Frequency		Experts' opinions[a]		Empirical evidence[b]	
Questions to experts (Questionnaire round 2)	Bioinformatics query (Changes from questionnaire)	True SNPs Percentage	Random SNPs Percentage	Pooled RP (95%CI)	Heterogeneity Percent of I^2 (P value)[c]	Pooled RP (95%CI)	Heterogeneity Percent of I^2 (P value)[c]
Q1: SNP in a transcribed but not translated region?	No change	58.7	50.6	2.1 (1.6–2.7)	33 (0.150)	1.4 (1.1–1.8)	0 (0.669)
Q2: SNP in a translated region but does not change the amino acid?	No change	1.4	0.6	2.8 (1.9–4.2)	69 **(0.001)**	4.0 (1.8–8.6)	0 (0.700)
Q3: SNP changes the amino acid but not in a functional protein domain?	No change	3.4	0.4	5.3 (3.2–8.8)	74 **(<0.001)**	7.7 (4.2–14.1)	0 (0.854)
Q4: SNP in a functional protein domain?	No change	4.5	0.4	6.4 (4.1–9.8)	61 **(0.009)**	9.6 (6.1–15.1)	0 (0.822)
Q5: SNP in a regulatory region which is not transcribed?	No change	3.6	3.5	3.5 (2.5–4.9)	41 **(0.094)**	1.2 (0.7–2.3)	0 (0.935)
Q6: SNP in a transcribed regulatory region?	No change	11.2	5.1	4.9 (3.8–6.4)	7 (0.377)	2.4 (1.6–3.5)	20 (0.281)
Q7: SNP in a genomic region evolutionary conserved in vertebrates?	No change	10.3	1.5	1.8 (1.4–2.3)	11 (0.344)	5.7 (3.8–8.4)	0 (0.888)
Q8: SNP in a gene (±5 kb) that has been associated with same/closely related phenotype in a meta-analysis or in >1 study?	SNP in a gene *investigated* with the phenotype in meta-analysis or >1 study	33.6	1.3	49.3 (19.7–123.2)	86 **(<0.001)**	21.1[d] (16.6–26.8)	64 **(0.010)**
Q9: SNP in a gene (±5 kb) that has been associated with same/closely related phenotype in a single study?	SNP in a gene *investigated* with the phenotype in a single study	8.1	2.4	6.4 (4.7–8.8)	40 (0.104)	2.4[e] (1.4–4.2)	57 **(0.030)**
Q10: SNP in a gene (±5 kb) under a linkage peak that has been associated with same/closely related phenotype?	NA	-	-	5.2 (3.7–7.4)	52 **(0.032)**	NA	NA
Q11: SNP in a locus within which other SNPs have been associated with the same/closely related phenotype?	SNP in a locus where other SNPs *investigated* with the phenotype	75.8	46.9	9.8 (4.6–21.0)	84 **(<0.001)**	3.5 (2.6–4.7)	43 (0.106)
Q12: SNP in a gene (±5 kb) that has been associated with same/closely related phenotype in functional models (animal or in vitro studies)?	SNP in a gene associated with the phenotype in *mouse* models	1.4	0.1	21.9 (12.1–39.5)	68 **(0.002)**	9.5 (2.3–38.5)	0 (0.984)
Q13: SNP in a gene (±5 kb) which is highly expressed in a tissue relevant to the phenotype?	No change	10.8	2.7	3.6 (1.9–6.7)	83 **(<0.001)**	3.4[f] (2.3–5.0)	52 **(0.080)**
Q14: SNP in a gene (±5 kb) which encodes for a protein in a pathway relevant to the phenotype?	No change	26.0	13.3	16.0 (8.3–30.7)	84 **(<0.001)**	2.1[g] (1.5–2.8)	47 **(0.079)**
Q15: SNP in a gene (±5 kb) which shows gene/protein interactions relevant to the phenotype?	SNP in a gene which shows *protein-protein* interactions relevant to the phenotype	39.9	18.7	10.3 (5.1–20.8)	83 **(<0.001)**	2.5 (1.9–3.2)	0 (0.527)

[a] Pooled estimate obtained from random-effect meta-analysis.
[b] Pooled estimate obtained from fixed-effect meta-analysis.
[c] P value from the heterogeneity test (in bold: P-value < 0.10).
[d] Random effects (RE) meta-analysis: 24.1 (15.5–37.4).
[e] RE meta-analysis: 7.3 (3.5–15.2).
[f] RE meta-analysis: 3.9 (2.1–7.4).
[g] RE meta-analysis: 2.4 (1.6–3.6).
RP, relative probability; 95%CI, 95% confidence intervals.

Between-expert and between-trait heterogeneity was tested using chi-square tests with statistical significance defined at P-value <0.10 [Fleiss, 1993], and the magnitude of the heterogeneity was estimated using the I^2, representing the percentage of variability in estimates explained by heterogeneity rather than sampling error [Higgins et al., 2003].

In this paper, the term "relative probability", used to indicate the probability of true association for a SNP given a certain type of prior knowledge compared with a SNP with no such evidence, refers to a relative risk for opinions and to an odds ratio for data. However, the impact of such difference on the comparison between opinions and data should be minimal, given the very low frequency of the outcome, represented by the a priori probability of true association for any given SNP in the genome [Davies et al., 1998].

Results

Experts' Opinions

Based on the findings of the first Delphi round and the correlation coefficients between items of the original

questionnaire (Supporting information Figs. SI and SII), the number of questions was reduced from 20 to 15. An item was dropped when it did not seem to convey additional information compared with another more general or relevant one, but only if the two were highly correlated and their relative probabilities very similar across experts (Supporting information Table SI). Although we had planned to drop types of evidence showing relative probabilities close to one (i.e., no relevance at all in support of a true association) consistently across experts, no such types of evidence were identified. Rewording of some questions which could have been interpreted either as mutually exclusive or overlapping was also performed to improve clarity in the second round (Supporting information Table SI).

Nine of the ten experts completed the second round. Some experts were more prone than others to change their opinion toward the average, and the extent of the changes from the first to the second round also varied across questions (Supporting information Table SI). Pooled estimates of relative probabilities across experts are presented in Table 2, and were obtained using a random-effect meta-analysis model due to the presence of moderate to large between-expert heterogeneity for most questions. Subgroup analyses by experts' background, biological vs. nonbiological, could not explain the heterogeneity observed (data not shown). Heterogeneity disappeared in two questions after excluding one outlying expert, who provided much higher relative probabilities for most questions (Supporting information Table SIII).

It turned out that the types of evidence considered as the most important (relative probabilities >10) were related to information at the gene level rather than to the SNP itself (Table 2). In decreasing order of perceived importance, they were as follows: gene previously associated with the phenotype in a meta-analysis or more than one study (Q8); gene previously associated with the phenotype in functional models (Q12); gene encoding for a protein in a pathway relevant to the phenotype (Q14); gene which shows gene/protein interactions relevant to the phenotype (Q15).

Empirical Evidence

The number of true SNPs varied across traits between 18 and 71 (Table 1). Table 2 reports the frequency of each type of evidence in truly associated and random SNPs, calculated as a weighted average across traits, while histograms for each trait are presented in Supporting information Figure SIII (the full set of results is available on our website at https://gemex.eurac.edu/downloads/stats/GenEpi2012). Some types of evidence were commonly observed in the sample of SNPs and their relative probabilities could be accurately estimated, while others, often with high relative probabilities, were rare (genuine low frequency or limited coverage of the bioinformatics query/data source) and had imprecise estimates. Pooled estimates of relative probabilities across traits were obtained using a fixed-effect model given the absence of substantial heterogeneity for most questions. Pooled estimates are presented and compared with those from experts' opinions in Table 2 and in the forest plots in Figure 1, while findings for the individual traits are reported in Supporting information Table SIV.

The type of evidence with the highest relative probability was previous association of the gene with the phenotype in a meta-analysis or in more than one study (Q8), whereas previous association in a single study (Q9) showed a much lower relative probability. Although this was in line with experts' opinions, estimates for the two types of evidence were substantially lower in the data, 21 (95% confidence interval: 17–27) vs. 49 (20–123) for association in meta-analysis/more than one study, and 2 (1–4) vs. 6 (5–9) for association in single study. Similarly, the relative probability for whether the SNP is in a locus within which other SNPs have been previously associated with the phenotype (Q11) was significantly lower in data compared with opinions, 4 (3–5) vs. 10 (5–21). The difference between the data and the opinions for these three types of evidence might be partly explained by the fact that the bioinformatics tool retrieved evidence on "previous investigation" rather than "previous association." This represents a measurement error in the assessment of the exposure (presence or absence of previous association), and as such is more likely to introduce bias toward the null, leading to underestimation of the relative probability in the data.

The type of evidence with the second highest relative probability in the data was whether the SNP is in a functional protein domain (Q4), with relative probability higher than in experts' opinions, 10 (6–15) vs. 6 (4–10). Similarly to the opinions, relative probabilities for the other three questions on the SNP's possible functional role (transcribed, Q1; translated, Q2; changes the amino acid, Q3) and the two questions on location in regulatory regions (not transcribed, Q5; transcribed, Q6) reflected their hierarchical structure. Relative probabilities for questions dealing with regulatory regions were significantly lower than those based on opinions, and only the question addressing whether the SNP is located in a transcribed region was statistically different from one.

Previous association of the gene with the phenotype in functional models (Q12) had the third highest relative probability. The scarcity of the data available for this type of evidence, which was limited to mouse models, made the estimate imprecise, 10 (2–39), and the observed difference with the experts' opinions (22; 12–40) could be due to chance.

Location of the SNP in a genomic region evolutionarily conserved in vertebrates (Q7) had a relative probability of 6 (4–8), much higher than in experts' opinions. Data and opinions gave very similar estimates for high gene expression in a tissue relevant to the phenotype (Q13), with relative probabilities around 3. Finally, whether the gene encodes for a protein which is in a pathway (Q14) or shows protein/protein interactions (Q15) relevant to the phenotype had relative probabilities around 2, significantly lower than in experts' opinions (16 and 10, respectively).

Results of the sensitivity analyses performed for three of the seven traits by extending the list of UMLS CUI terms to cover "closely related" phenotypes, as formulated in the questions to experts, were similar to the main results (Supporting information Table SV). Similarly, adjusting the

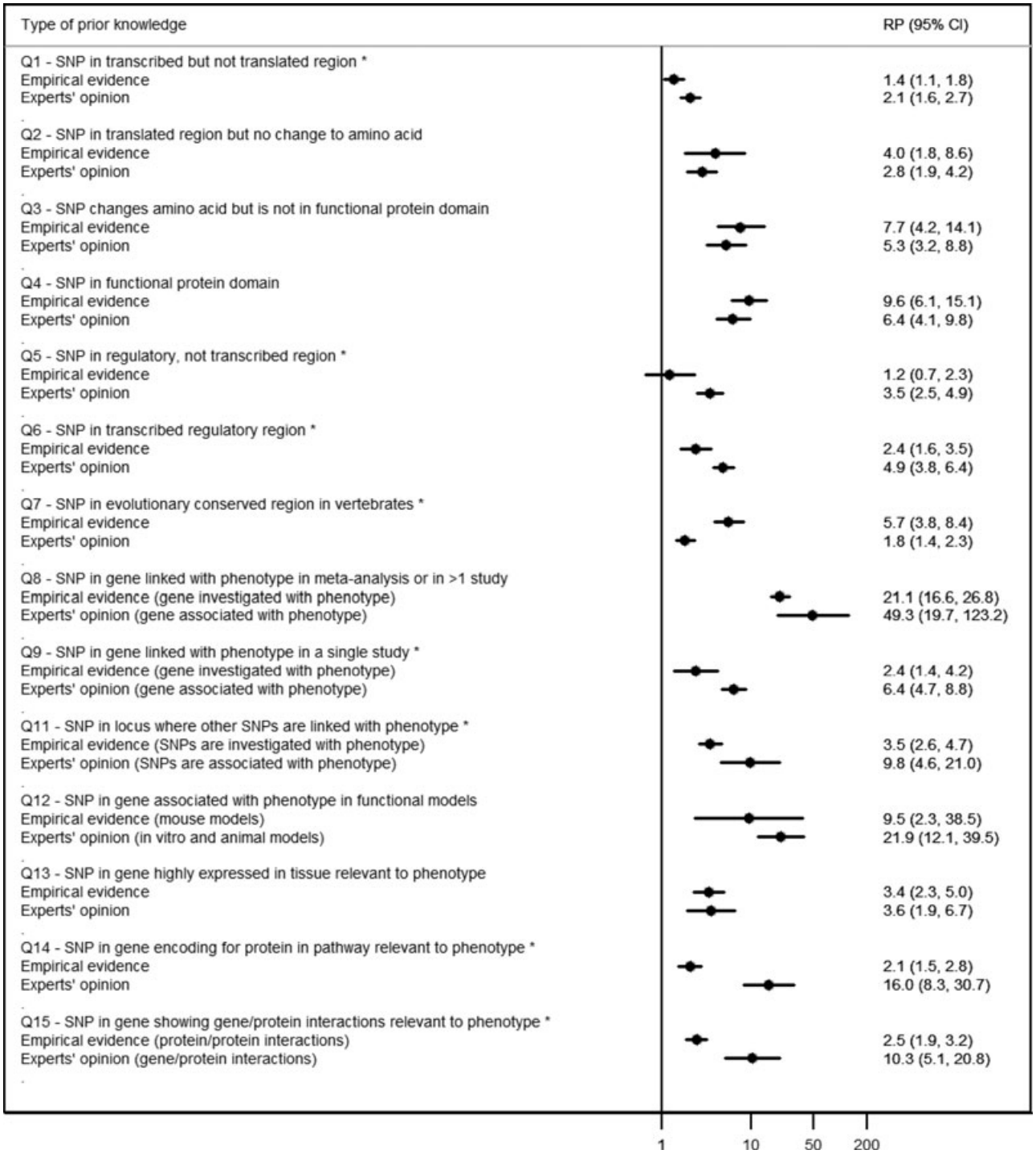

Figure 1. Plots of the relative probabilities of association for each type of prior knowledge from empirical evidence (with 95% confidence intervals) compared with those based on experts' opinions (with 95% intervals). Q10 is omitted as empirical evidence is not available (data could not be processed electronically). *Questions for which the difference in estimates from data vs. opinions are statistically significant (*P*-value < 0.05).

analyses for allele frequency did not change the results (data not shown).

Dependence Between Questions

The degree of dependence between questions, as expressed by the correlation matrix of 7,000 random SNPs, is shown in Supporting information Table SVI. The correlation of Q7 (SNP in a region evolutionary conserved in vertebrates) with Q2, Q3, and Q4 (indicating a possible functional role, from translated without amino acid change to translated in functional protein domain) may be due to functional regions being more likely to be conserved [Levenstien and Klein, 2011]. Similarly, the correlation between Q14 (SNP in a gene encoding for a protein in a pathway relevant to the phenotype) and Q15 (SNP in a gene showing protein-protein interactions relevant to the phenotype) might be explained by the fact that proteins in a pathway may also interact with each other [Kirouac et al., 2012]. The interdependence of the different types of prior knowledge needs to be accounted for when they are used together through conditional analyses that jointly model them, as discussed in the companion paper by Thompson et al. [2013].

Discussion

The use of prior knowledge may improve the selection of GWA signals for follow-up, thus increasing the probability of a successful replication or functional investigation. Studies which have systematically incorporated prior knowledge from multiple data sources using bioinformatics tools have attributed equal importance to the different types of evidence [Aerts et al., 2006; Chen et al., 2011; Sookoian et al., 2009], and yet our findings suggest that this may be suboptimal. Our study convincingly shows that, for commonly investigated traits, evidence from previous association studies on the phenotype of interest represents the most informative type of knowledge for gene prioritization, although it does not help discover novel genes. The empirical findings suggest that SNPs in genes previously investigated in relation with the phenotype in a meta-analysis or in more than one study are 21 times more likely to represent true associations, with this being reduced to two times if previous investigation is limited to a single study.

Our findings suggest that location of the SNP in a functional protein domain may increase the probability of true association up to 10 times, with progressively decreasing effect for whether the SNP changes the amino acid but is not in a functional protein domain, and whether the SNP is in a translated region but does not change the amino acid. Despite the very low proportion of SNPs with these characteristics, between 1% and 5% in the SNPs associated with our seven traits, information on SNP characteristics can be retrieved easily and accurately so that these types of evidence are worth considering in the prioritization of GWA signals. Similarly, location of the SNP in a gene previously associated with the phenotype in functional models could substantially increase the probability of true association by nine times according to our empirical findings limited to mouse data, but up to 23 times in experts' opinions regarding animal and in vitro models in general. The mouse model is widely used [Hardouin and Nagy, 2000; Rosenthal and Brown, 2007], and the observed empirical estimates were highly consistent across the seven traits considered, suggesting that retrieving such information can still be useful when other functional data cannot be accessed. However, its frequency was very low (1% in our "true SNPs"), and the practical importance of incorporating functional evidence in gene prioritization would increase by considering additional models.

It is interesting to note how studies which have tried to incorporate prior knowledge have often disregarded knowledge of association of the gene with the phenotype of interest from human and animal studies, but rather focused on information on pathways or protein-protein interactions, SNP characteristics and gene expression data [Chen et al., 2009; Parikh et al., 2009; Saccone et al., 2008; Zhong et al., 2010]. Use of gene pathway information for gene prioritization has received much attention and bioinformatics tools have been developed to allow retrieval of such information at genome-wide level [Cantor et al., 2010; Elbers et al., 2009; Zhong et al., 2010]. This is reflected by experts' opinion, which ranked this type of information as the third most important. However, our empirical findings based on pathway information retrieved using KEGG [Kanehisa and Goto, 2000] do not support this view. This may be partly explained by the difficulty of defining the boundaries of a pathway, but nonetheless suggests that more investigation is needed to evaluate the potential value of pathway information and how it should be modeled. Similarly, our empirical findings did not support the importance attributed by the experts to information on whether the gene product shows evidence of interactions relevant to the phenotype. On the other hand, our empirical findings suggest that presence of the SNP in a genomic region evolutionarily conserved in vertebrates could increase the probability of true association by six times, contrary to experts' opinion that ranked this as the least important type of knowledge. Although it occurred in only 10% of our "true SNPs", this type of evidence can be easily and accurately retrieved and may well be incorporated in gene prioritization. An interesting follow-up of our study will be to investigate the impact of the choice of a 5 kb window for mapping SNPs to genes, and to provide evidence on what window might be the most informative. Such choice is likely to influence the estimated relative probability of association of many types of knowledge, including whether the gene encodes a protein in a pathway.

Limitations of the Study

Our empirical findings are based on only seven examples of gene-disease associations, but their generalization is supported by the high consistency observed across traits for most types of evidence. The precision of our empirical weights could be improved by considering more traits and

increasing the number of "true SNPs" analyzed. This could be done systematically using publicly available databases such as the NHGRI GWAS Catalog, a continuously updated catalog of findings from published GWA investigations [Hindorff et al., 2009]. As for the selection of "true SNPs" for each of the seven traits, although the completeness of our lists is not an issue, its representativeness is, and it may well be that SNPs identified by GWA investigations and replicated are not representative of all true genetic associations, particularly those with weaker effects.

Many of the "true SNPs" in our seven traits may be SNPs in linkage disequilibrium with the real causal variants, so that types of evidence referring to the characteristics of the SNP (e.g., "the SNP is in a functional protein domain") may be negative for the "true SNP" only because this is in fact only a proxy of the causal one. This, which can be interpreted as measurement error in the definition of our "cases", is likely to introduce bias toward the null and therefore lead to underestimation of relative probabilities, particularly for types of evidence referring to SNP characteristics. Other forms of misclassification could in theory lead to a bias away from the null, for example, if, because of the way in which the GWA studies have been conducted, a "true SNP" is more likely to have been identified if it has certain of the characteristics, or due to inaccuracies in the SNP annotations. Empirical estimates of relative probabilities of true associations will be different in the future, when GWA findings will be based on newer sequencing and genotyping technology resulting in higher genome coverage and improved reliability. In general, empirical estimates of the relative importance of different types of evidence will depend on current knowledge and data availability, thus requiring continual updating of the information extracted from the databases.

Regarding the retrieval of the evidence, our study shows the limitations of using bioinformatics tools that search for prior knowledge at genome-wide level from publicly available databases, and the practical limits on certain types of information, such as evidence from linkage studies, functional studies other than mouse models, and eQTL databases for gene expression from multiple tissue sources. Moreover, data quality strongly depends on the coverage provided by the interrogated databases, which suggests that integrating information on a certain type of evidence from multiple databases may be preferable to relying on a single one.

Practical use of our Findings

Investigators willing to incorporate prior information on biological function or evidence from previous studies in the selection of GWA hits for follow-up encounter a few practical issues as follows: What types of prior knowledge are worthwhile considering? How can prior knowledge be retrieved in a systematic way? How can prior knowledge be combined with the discovery *P* values? How should different types of knowledge be differentially weighted to provide an overall a priori probability of association for each SNP? Our findings answer the question about the relative importance of different types of prior knowledge and show the feasibility of automatic retrieval of such information using a bioinformatics tool that queries multiple data sources. A companion paper by Thompson et al. [2013] demonstrates the use of prior knowledge in combination with discovery *P* values within a Bayesian framework to provide a posterior probability of replication, which can be used to rank the most promising SNPs for follow-up. That work combines our estimates of relative probabilities for 14 types of knowledge and calculates the overall prior probability of association for a given SNP.

Thompson et al. demonstrate that the success of replication is increased when the selection of SNPs incorporates prior knowledge using a simple approximate Bayesian analysis, compared with the classical approach purely based on discovery *P* values.

Acknowledgments

All researchers from the Center for Biomedicine at EURAC were supported by the Department for Promotion of Educational Policies, Universities and Research of the Autonomous Province of Bolzano, South Tyrol, Italy; M.B. was supported by NIH grant HG000376. None of the authors declare any conflict of interest.

References

Adler M, Ziglio E. 1996. *Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health*. London: Jessica Kingsley.

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B and others. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol* 24:537–544.

Akins RB, Tolson H, Cole BR. 2005. Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Med Res Methodol* 5:37.

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J and others. 2010. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 38:D525–D531.

Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT, Mouse Genome Database Group. 2011. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res* 39:D842–D848.

Cantor RM, Lange K, Sinsheimer JS. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet* 86:6–22.

Chen J, Aronow BJ, Jegga AG. 2009. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 10:73.

Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. 2011. In silico gene prioritization by integrating multiple data sources. *PLoS One* 6:e21137.

Coronary Artery Disease (C4D) Genetics Consortium. 2011. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet* 43:339–344.

Davies HT, Crombie IK, Tavakoli M. 1998. When can odds ratios mislead? *BMJ* 316:989–991.

Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. 2009. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol* 33:419–431.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K and others. 2010. The Pfam protein families database. *Nucleic Acids Res* 38:D211–D222.

Fleiss JL. 1993. The statistical basis of meta-analysis. *Stat Methods Med Res* 2:121–145.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S and others. 2011. Ensembl 2011. *Nucleic Acids Res* 39:D800–D806.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R and others. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42:1118–1125.

Gögele M, Minelli C, Thakkinstian A, Yurkiewich A, Pattaro C, Pramstaller PP, Little J, Attia J, Thompson JR. 2012. Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. *Am J Epidemiol* 175:739–749.

Greene CS, Penrod NM, Williams SM, Moore JH. 2009. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One* 4:e5639.

Hardouin SN, Nagy A. 2000. Mouse models for human disease. *Clin Genet* 57:237–244.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. *BMJ* 327:557–560.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106:9362–9367.

Ioannidis JP. 2007. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered* 64:203–213.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human MicroRNA targets. *PLoS Biol* 2:e363.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30.

Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK. 2012. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst Biol* 6:29.

Köttgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV and others. 2010. New loci associated with kidney function and chronic kidney disease. *Nat Genet* 42:376–384.

Kraft P, Zeggini E, Ioannidis JP. 2009. Replication in genome-wide association studies. *Stat Sci* 24:561–573.

Levenstien MA, Klein RJ. 2011. Predicting functionally important SNP classes based on negative selection. *BMC Bioinformatics* 12:26.

Liu YJ, Papasian CJ, Liu JF, Hamilton J, Deng HW. 2008. Is replication the gold standard for validating genome-wide association findings? *PLoS One* 3:e4037.

Mells GF, Floyd JA, Morley KI, Cordell HJ, Franklin CS, Shin SY, Heneghan MA, Neuberger JM, Donaldson PT, Day DB and others. 2011. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat Genet* 43:329–332.

Moreau Y, Tranchevent LC. 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13:523–536.

Parikh H, Lyssenko V, Groop LC. 2009. Prioritizing genes for follow-up from genome wide association studies using information on gene expression in tissues relevant for type 2 diabetes mellitus. *BMC Med Genomics* 2:72.

Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X and others. 2006. cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* 34:D68–D73.

Rosenthal N, Brown S. 2007. The mouse ascending: perspectives for human-disease models. *Nat Cell Biol* 9:993–999.

Saccone SF, Saccone NL, Swan GE, Madden PA, Goate AM, Rice JP, Bierut LJ. 2008. Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 24:1805–1811.

Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C and others. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 43:333–338.

Sookoian S, Gianotti TF, Schuman M, Pirola CJ. 2009. Gene prioritization based on biological plausibility over genome wide association studies renders new loci associated with type 2 diabetes. *Genet Med* 11:338–343.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Mägi R and others. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 42:937–948.

Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E. 2004. The Ensembl core software libraries. *Genome Res* 14:929–933.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A and others. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42:508–514.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G and others. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101:6062–6067.

Thomas DC, Conti DV, Baurley J, Nijhout F, Reed M, Ulrich CM. 2009. Use of pathway information in molecular epidemiology. *Hum Genomics* 4:21–42.

Thompson JR, Gögele M, Weichenberger CX, Modenese M, Attia J, Barrett JH, Boehnke M, De Grandi A, Domingues FS, Hicks AA and others. 2013. SNP prioritization using a Bayesian probability of association. *Genet Epidemiol* 37:214–221.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* 35:D88–D92.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G and others. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589.

Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd and others. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 10:R130.

Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. 2008. A navigator for human genome epidemiology. *Nat Genet* 40:124–125.

Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. 2010. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* 86:581–591.