DATABASES

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

# Developing the "Next Generation" of Genetic Association Databases for Complex Diseases

Christina M. Lill[1,2] and Lars Bertram[1]*

[1]Neurospsychiatric Genetics Group, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany;
[2]Department of Neurology, University Medical Center of the Johannes Gutenberg-University, Mainz, Germany

For the Databases in Neurogenetics Special Issue

Received 31 January 2012; accepted revised manuscript 6 June 2012.
Published online 2 July 2012 in Wiley Online Library (www.wiley.com/humanmutation).DOI: 10.1002/humu.22149

**ABSTRACT:** Tens of thousands of genetic association studies investigating the influence of common polymorphisms on disease susceptibility have been published to date. These include ~1,000 genome-wide association studies (GWAS). This vast amount of data in the field of complex genetics is becoming increasingly difficult to follow and interpret. It can be expected that the situation will become even more complex with the advent of association projects using "next-generation" technologies. One of the aims of the Human Variome Project is to concatenate such data in meaningful ways, for example, within the context of publicly available field synopses. Here, we present various examples of online genetic association databases developed by our group for neuropsychiatric disorders. One integral part of this model is the systematic inclusion of data from large-scale genotyping projects, for example, GWAS, while respecting the privacy of data contributors. We believe that our database approach may serve as a viable model that can be readily applied to other fields and ultimately improve our understanding of the genetic forces driving common human conditions.
Hum Mutat 33:1366–1372, 2012. 

**KEY WORDS:** neurogenetics; database; meta-analysis; GWAS; HVP; association; genome; exome

## Introduction

Most, if not all, common diseases are caused by an interplay of genetic and environmental factors. In many diseases, there is also often a small subset of subjects that inherit the respective condition in a Mendelian fashion (e.g., <5–10% of patients suffering from Alzheimer's (AD) or Parkinson's disease (PD) [Lill and Bertram, 2011]. However, the majority of disease cases are of a multifactorial and polygenic nature and are thus often labeled as being "genetically complex." Because the identification and elucidation of the underlying genetic architecture have oftentimes promising implications for disease prevention and early treatment [e.g., see Collins, 2010], the efforts and money spent in this field of genetic epidemiology have been enormous. Literally, tens of thousands of genetic association studies, that is, studies that investigate the influence of usually common genetic variants on disease susceptibility, have been published to date. For instance, the number of such studies published for some of the most common neuropsychiatric diseases alone, including AD, PD, schizophrenia, amyotrophic lateral sclerosis (ALS), and multiple sclerosis (MS) amounts to almost 5,000 at the time of writing (Fig. 1). This vast amount of information is becoming increasingly difficult to follow, evaluate, let alone interpret. To this end, the vision of the Human Variome Project (HVP) is "to develop a global collaboration with the aim of building systems and strategies for the collection, storage, interpretation and sharing of human genetic variation and its implications for disease" [Haworth et al., 2011]. Along these lines, the HVP has emphasized the need for disease-specific genetic association databases that would systematically identify and collect the relevant data, quantitatively assess the impact of genetic variants on complex disorders, and make the results available in high-quality, user-friendly databases [Haworth et al., 2011; Cotton et al., 2007]. Independently from the HVP, our group has already developed and continues to maintain a number of such genetic association databases for neuropsychiatric diseases [Allen et al., 2008; Bertram et al., 2007; Lill et al., 2011, 2012] such as AD, PD, ALS, MS, and schizophrenia. The approach behind these database projects represents the focus of this article as it may serve as one model to achieve the abovementioned goals outlined by the HVP. Furthermore, challenges and potential problems related with the databases' construction and maintenance will be discussed.

## State-of-the-Art Genetic Research in Complex Diseases

The field of genetic epidemiology in complex diseases has experienced a tremendous shift in recent years owing to the feasibility of generating and analyzing high-throughput genotyping data, such as those typically produced in the context of genome-wide association studies (GWAS). This "GWAS era" has followed more than three decades of candidate-gene-based association studies. This approach typically only investigated a limited number of variants across single genes or sets of genes based on functional considerations. Notwithstanding its simplicity and low scale, this approach successfully let to the identification of some risk genes for a number of diseases, usually the individually most important risk loci, that is, those exerting the largest risk effects. These major risk loci were subsequently largely confirmed in recent GWAS [Siontis et al., 2010]. For the

*Correspondence to: Lars Bertram, Neuropsychiatric Genetics Group, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany. E-mail: lbertram@molgen.mpg.de

Contract grant sponsors: Cure Alzheimer's Fund (CAF); Michael J. Fox Foundation for Parkinson's Research (MJFF); National Alliance for Research on Schizophrenia and Depression (NARSAD); Prize4Life; EMD Serono; Fidelity Biosciences Research Initiative (to C.M.L.); German Ministry for Education and Research (BMBF to L.B.).

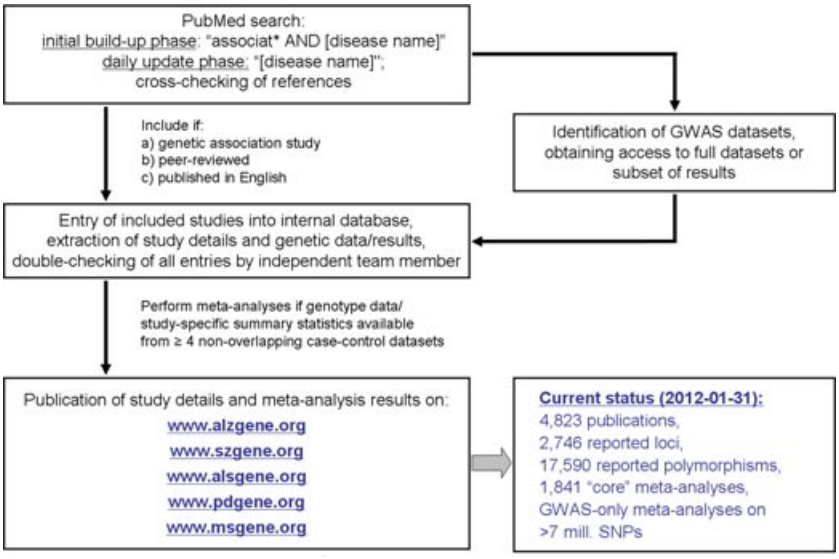


**Figure 1.** Simplified overview of literature search, data extraction, and analysis strategies applied to the database projects described in this article. "Core meta-analyses" refer to meta-analyses displaying individual-study results, that is, derived from GWAS and/or candidate gene studies. This flowchart has been modified after Lill and Bertram (2010).

last few years, GWAS, which test several hundred thousand polymorphisms in the same experiment, have nominated hundreds of additional disease risk loci across complex diseases. In contrast to the confirmed candidate gene findings, however, these typically exert only small to modest risk effects [Hindorff et al., 2009]. Current candidate-gene association studies are now typically focused on validating the most promising GWAS results in independent datasets or on performing secondary GWAS analyses (e.g., pathway-based analyses). These candidate-gene studies continue to be important to distinguish false-positive from genuine signals, to refine GWAS association signals by fine-mapping approaches, as well as to possibly identify additional variants not meeting the stringent significance thresholds in primary GWAS analyses. Furthermore, smaller-scale studies often provide a framework to investigate the newly proposed risk variants in distinct populations and ethnicities for which no GWAS data have been generated, and thus allow to assess the universality of genetic disease risk profiles [e.g., Sharma et al., 2012]. Thus, despite the undeniable success of the GWAS approach in expanding our knowledge and understanding about the genetic architecture of common diseases, assessing the *cumulative* evidence of association continues to be essential, but—for reasons outlined below—is becoming more and more difficult to achieve. One main limiting factor is that many GWAS datasets are not shared with outside investigators. But even if full access to the GWAS data has been granted, the computational challenges, that is, processing, analyzing, and correctly interpreting these data, remains one major bottleneck when assessing the evidence for genetic risk factors of interest. It can be expected that this situation will become several orders of magnitude more complex with the expected advent of genetic association projects using next-generation sequencing technologies, for example, via whole-exome or whole-genome sequencing [Green and Guyer, 2011]. To this end, it has been recognized that "generally applicable tools are needed in the form of robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists. Adequate computational infrastructure is also needed, including sufficient storage and processing capacity to accommodate and analyze large, complex data sets (including metadata) deposited in stable and accessible repositories, and to provide consolidated views of many data types, all within a framework that addresses privacy concerns" [Green and Guyer, 2011]. This calls for a "next-generation" of genetic association databases with an appropriate infrastructure to incorporate large-scale association data, that is, GWAS and next-generation sequencing data. In the following paragraphs, we will present the genetic database approach of our group. A particular focus will be laid on recent novel developments that now allow to systematically include some of the above described large-scale data, that is, those originating from GWAS. Our systematic approach to data collection, analysis and annotation as well as database design may serve as a model for genetic association databases of other diseases to meet the needs and standards of researchers in the twenty-first century.

## Database Design and Structure

### General Approach

The databases developed by our group [for more details on these databases, see Allen et al., 2008; Bertram et al., 2007; Chatzinasiou et al., 2011; Lill et al., 2011, 2012] aim to serve as an exhaustive and regularly updated resource of genetic association studies for the respective diseases. They do not only allow a qualitative overview of genetic association studies in each field, but also provide quantitative assessments of the cumulative evidence for association of individual polymorphisms by calculating and providing up-to-date meta-analyses on all eligible polymorphisms.

Applicable association studies are identified by continuous PubMed searches and by systematically screening the references of relevant publications (e.g., primary association papers, but also reviews and published meta-analyses). A publication is included in one of our databases if it assesses the association of a polymorphism (defined as DNA sequence variants with ≥1% frequency in the general population) with risk for disease, and if it has been published in a peer-reviewed English language journal. Demographic

details of included studies (such as ethnicity of the investigated datasets, number of subjects, sex and age distribution, and diagnostic criteria) are extracted from each publication, and summarized in the respective databases. Furthermore, and most importantly, genetic association data (e.g., genotype summary data, allele frequencies or odds ratios [ORs], and confidence intervals) for each reported polymorphism are extracted from the publication. Meta-analyses are performed on bi-allelic polymorphisms (e.g., SNPs, indels) investigated in a case-control setting and available in at least four independent datasets. Meta-analysis results are displayed for all ethnicities combined, as well as after stratification for different ethnic background, provided sufficient data are available. Family-based studies without available subject-level data are excluded from meta-analyses due to often insufficient and inconsistent reporting of results (however, datasets including unrelated case-control subjects enriched for familial cases are not excluded). Data on mitochondrial DNA variants are excluded because of the multicopy nature of the mitochondrial genome, and the high frequency of somatic mutation events that vary substantially across tissues. Also, data of obviously "poor" quality are excluded if apparent discrepancies could not be resolved after contacting the study authors (see below). For more details on background and methods, see Lill et al. (2012).

## Inclusion of GWAS Datasets

As pointed out above, the data derived from GWAS have become a substantial resource of unbiased association data and their inclusion is essential for the validity and success of any genetic association database. To this end, the genetic databases developed by our team [Allen et al., 2008; Bertram et al., 2007; Chatzinasiou et al., 2011; Lill et al., 2011, 2012] list all published GWAS in the respective diseases fields in a dedicated overview section, which provides demographic characteristics of the included samples and highlights those loci that have been "featured" as potential disease loci in the respective publications.

Furthermore, all GWAS data provided in the respective publications themselves, that is, usually data on the top signals of that study (i.e., "featured genes") or those postulated by previous studies, can be included in the databases even if the full GWAS data or results are not shared by the primary authors. However, while these results provide the most interesting findings of that particular study, this approach alone could lead to selective reporting bias. Therefore, in addition and, most importantly, access to all existing GWAS datasets is sought (if publicly available usually by application via dbGaP [http://www.ncbi.nlm.nih.gov/gap], or else by directly contacting study authors). In case of the availability of subject-level data, they are cleaned using common quality control steps (i.e., removing duplicate or related samples within and across GWAS datasets, removing individuals and SNPs with insufficient genotyping efficiency, removing SNPs violating Hardy–Weinberg equilibrium), and data are subject to genotype imputation and association analysis [Marchini et al., 2007]. This includes adjustment for population structure, and, if available, age and sex (for more details on GWAS dataset preparation and analysis, see Lill et al., 2012). Alternatively, we also include GWAS summary statistics generated and provided by the authors, that is, ORs and standard errors (SEs). These can easily be integrated into meta-analyses already existing for any given polymorphism on the respective databases. This procedure does not require subject-level data and may therefore alleviate the process of (and the concerns that come with) sharing of individual-level GWAS data. After preparation and formatting of the GWAS datasets, meta-analyses are performed or updated across eligible SNPs, that is, based on all available data from GWAS as well as pre- and post-GWAS "candidate-gene" approaches. This process also allows to add a substantial amount of unbiased GWAS data to many meta-analysis results that otherwise would be derived from candidate-gene data alone. In addition to being showcased on our databases themselves, all meta-analysis results (i.e., *P* values and directions of effect) are also displayed on a customized UCSC genome browser track [Kent et al., 2010].

## Privacy Protection Policies

It has now been demonstrated by several groups that the probability of an individual being the member of a specific cohort, for which large-scale genotyping data are available, can be estimated under some conditions even from some types of summary statistics [e.g., Homer et al., 2008; Jacobs et al., 2009; Sankararaman et al., 2009]. As a consequence, this has led to more restrictive data sharing policies for GWAS and other large-scale data, for example, regulated data access policies via dbGaP. Obviously, these restrictions and issues need to be considered in the framework of genetic association databases to ensure the maximally possible degree of privacy protection of any individual who has contributed their data to GWAS. However, it needs to be emphasized that to resolve potential cohort membership requires several premises that are not applicable in the context of our database approach. First, membership probabilities can be estimated if summary allele frequencies or genotype counts are provided [Craig et al., 2011]. This is one of the reasons why new versions of the databases developed and maintained by our group do not display summary genotype data or allele frequencies for large-scale data but only rounded ORs and SEs (adjusted for population substructure and age and sex if available). Estimating membership probabilities from such data becomes more difficult with increasing numbers of included subjects [Craig et al., 2011]. Second, the number of SNPs, for which genotype summary counts or allele frequencies are needed to resolve cohort membership, is rather large to allow an accurate prediction. Furthermore, the precision of the probability estimate decreases with an increasing number of individuals genotyped in the cohort. For instance, Craig et al. could show that sharing genotype summary data for ~1,000 SNPs from cohorts of >500 individuals yields very low positive predictive values [Craig et al., 2011], that is, membership cannot be reliably estimated. Most GWAS performed to date, in fact, have included more than 500 combined cases and controls [Hindorff et al., 2009]. Earlier GWAS, possibly with numbers falling below this threshold, have often been superseded by later efforts from the same groups then involving extended datasets, that is, typically >500 individuals. While privacy protection concerns related to information publicly displayed across our databases thus lack any rationale, we took the following additional measures to ensure maximal protection of privacy of GWAS participants. (1) As outlined above, genotype summary data or allele counts or frequencies for large-scale association studies are not being displayed. Instead, we only showcase rounded and adjusted ORs and SEs per dataset. (2) Study-level results (i.e., rounded ORs and SEs) of GWAS and other large-scale association studies are displayed only for a subset of SNPs, that is, up to a maximum of 10,000 SNPs. These SNPs include the most relevant SNPs investigated in the respective diseases, that is, the "top results" of GWAS and GWAS-derived meta-analyses as well as SNPs investigated in candidate-gene approaches of the pre- and post-GWAS era. (3) Meta-analysis results of SNPs that are not part of the "top 10,000 SNPs" (i.e., results from GWAS-only meta-analyses showing no compelling evidence for association) are being displayed without detailing study-level ORs or

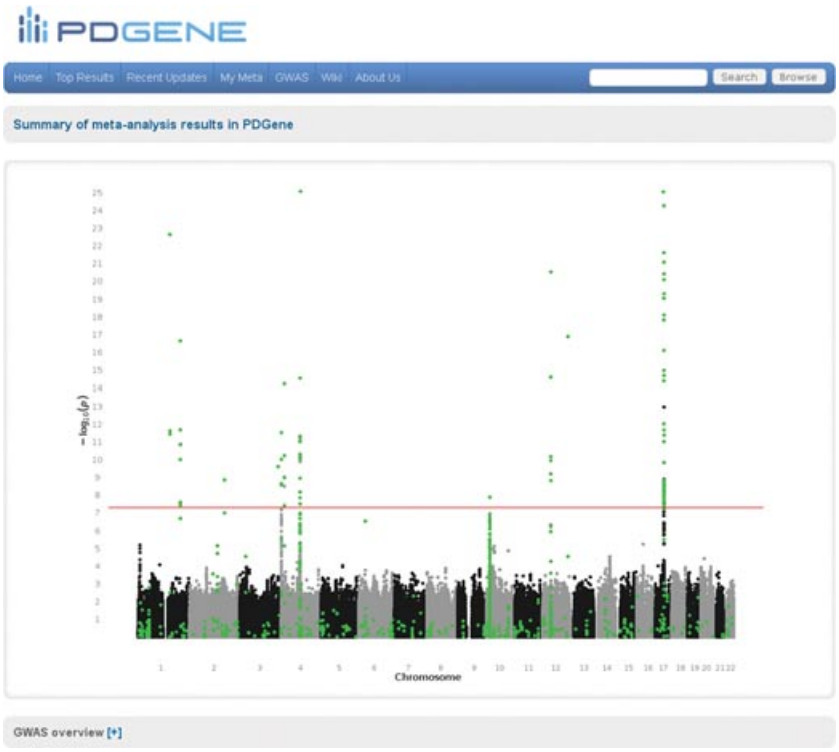


**Figure 2.** Screenshot of the PDGene database showing a Manhattan plot of all available meta-analysis results. This screenshot shows the "GWAS overview" page of the PDGene database (http://www.pdgene.org/largescalemeta.asp) presenting *P* values from >7 million meta-analyses results including fully available, imputed GWAS datasets (back and gray dots), and meta-analysis results based on a combination of "pre- and post-GWAS candidate-gene studies and fully or partially available GWAS data where applicable (green dots). For more details, see Lill et al. (2012).

SEs. However, the database allows to investigate the combined evidence for association for all meta-analyzed polymorphisms (based on *P* values and direction of effect, see also Fig. 2).

### Linking Different Disease-Related Databases

Our databases are cross-linked, for example, on the gene and polymorphism level, which allows a direct comparison of association results across different neuropsychiatric diseases. This functionality will be extended in future releases of our database software.

### Database Code

The newly designed database code will be made available upon request to interested researchers from other disease fields on a collaborative basis.

## Limitations, Practical Difficulties, and Ways to Deal with Them

While the above-described approach appears relatively straightforward, its practical implementation typically turns out to be much more complex and laborious than outlined. In extreme cases, this sometimes even requires revision of the applied database curation strategies. Here, we will discuss what we consider the major challenges related to what determines (in our opinion) the acceptance and success of genetic association databases, or any scientific database: *completeness*, *correctness*, and *timeliness* of data display.

### Completeness

As we have shown previously, *completeness* of the identified literature in the context of genetic association studies can successfully be achieved by searches of NCBI's PubMed database and, within individual publications, cross-checking for additional references. Screening of additional literature databases has not led to the identification of additional articles meeting our inclusion criteria [Allen et al., 2008; Lill et al., 2012]. As literature screening is a process still requiring a large amount of human interaction, and since the accuracy of this process is essential for all subsequent database curation and analysis steps, employing a person with pre-existing knowledge and experience in the field is obviously advantageous. However, although the manual literature screenings seem to rarely miss eligible articles based on our experience, this process has a rather low specificity. For instance, while building and regularly updating the PDGene database, a total of ~27,000 citations had been screened until March, 2011 [Lill et al., 2012]. The major burden of PubMed screening relates to the "updating" process. Only about 8,300 citations were screened during database construction, and 19,200 subsequently during database maintenance. However, altogether only 3% of all screened PubMed citations fulfilled our inclusion criteria. This process can be streamlined by applying artificial intelligence, for example, self-learning data mining methods. Usage of a recently described algorithm that aimed at this goal substantially reduced the burden of manual literature searches [Wallace et al., 2012]: this particular algorithm can be trained on a "discovery dataset" of eligible articles and subsequently yields 100% sensitivity and ≥90% specificity for identification of PubMed citations when updating literature searches.

Furthermore, fulfilling the *completeness* criterion depends on the extent to which the relevant data are available. Despite the existence of established guidelines on the reporting of genetic association studies, that is, the STREGA guidelines [Little et al., 2009], missing data in publications remain an important issue that can come in different forms and shapes. The most common problems are undisclosed names and/or genetic association results (genotypes, allele counts, or summary statistics [ORs and CIs]) of polymorphisms that have been investigated. It is rather frequent in certain disease fields that even in smaller-scale studies no genotype summary data or allele frequencies are provided; in these situations, inclusion of allelic or additive ORs and CIs—if provided in the respective publication—can reduce the amount of missing data. Importantly, the database's data extraction protocol should specify in advance which covariates are accepted for inclusion of published ORs/CIs/SEs. Unfortunately, it is relatively common in genetic association studies to provide ORs and *P* values, but not necessarily CIs or SEs, which makes inclusion of these data more difficult. One potential solution to this problem is afforded by simulations of genotype data based on reference allele frequencies as implemented in our novel database code. Furthermore, a major and often underestimated and underappreciated problem is the missing specification of direction of effect, for example, the situation where genotype summary data/allele frequencies are named without specifying the minor and major allele, or providing ORs without specifying risk or reference alleles. A similar situation occurs for ambiguous allele names, that is, A/T and C/G—here strand information would be necessary to define allele assignments, although this only affects a small subset of studies thus far included in our databases [Lill et al., 2010]. In both situations, it is often possible to resolve missing allele designations or to deduce strand information by comparing allele frequencies with ethnicity-specific reference panels such as HapMap or 1000 Genomes. This is now achieved automatically in a new version of our database code soon to be launched publicly. However, this can only be done allele if minor/major frequencies are not too similar; we have defined this as allele frequencies <0.4 in the study dataset and/or reference panel. If missing data still precludes inclusion of datasets in the respective meta-analyses, authors must be contacted and asked to provide the relevant missing information. Our team does this by usually contacting the first and/or last authors twice via e-mail. Regardless of all attempts to resolve the missing data problem, up to 8% of small-scale association data remain unavailable for inclusion [Allen et al., 2008; Lill et al., 2012].

Unavailability of individual-level or study-level data from large-scale association datasets, most commonly GWAS, obviously aggravate the "missing data" problem quite considerably. This typically relates to individual-level data if they are not available via public platforms such as dbGaP. However, our experience shows that it is often feasible to at least obtain subsets of the missing data in a collaborative effort with the investigators of the primary GWAS. Possible scenarios include the sharing of summary statistics, that is, ORs, and SEs, but not subject-level data, for the full dataset or only a subset of SNPs (e.g., ~1,000 to 10,000 SNPs, which includes the top results of that dataset as well as those SNPs already entered in the respective database)

## Correctness

To avoid curational errors during the screening of the literature, the inclusion of studies, and the extraction of relevant data it has proven advantageous in our experience that the most important processes and data entries in the database are double-checked by a second team member. This double-checking is crucial for the data extraction processes where study designs can be misinterpreted or random errors can occur and possibly lead to erroneous meta-analysis results. In case of persisting disagreement, consensus is reached by consulting a third colleague, typically the supervisor.

Even our general inclusion criterion of requiring publication in a peer-reviewed journal does not exclude studies of minor quality. Most common concerns are inadequate study design (e.g., biased selection of study subjects), and errors in genotyping/sequencing protocols or discrepancies in terms of allele names or frequencies when compared with public references. Naturally, the more experienced the database curators are, the better can such errors be identified, with subsequent correction or exclusion of data from meta-analysis. In addition, a few consistency checks can also be performed algorithmically by the database code, for example, comparing allele assignments and frequencies to public databases such as dbSNP. As we include common polymorphisms in our databases, we report the respective official identifiers (NCBI's "rs-numbers," http://www.ncbi.nlm.nih.gov/sites/SNP) where applicable and available, including hyperlinks to the "dbSNP" database which extensively annotates all variants using the nomenclature system espoused by the Human Genome Variation Society (HGVS, www.hgvs.org/mutnomen). Some studies publish polymorphism data without the respective official identifiers. In these instances, additional and sometimes extensive curational efforts are needed to assess whether the assignment to an official identifier is possible based on sequence information provided in the respective publications and comparison to public databases (e.g., Ensembl, http://www.ensembl.org/index.html). If not applicable, we report the polymorphism names as published in the respective publications to ensure consistency with the primary publications. In addition to these study-specific polymorphism names, we annotate variants based on HGVS nomenclature whenever this is possible based on the sequence context information provided in the primary publications. Another important concern relates to the lack of correctly identifying overlaps across datasets. This can lead to inflated and potentially biased meta-analysis results. Criteria applied by our group that result in an in-depth assessment for potential overlaps are: similarities in demographic details (e.g., overlapping recruitment regions) and shared coauthors. In case that overlap is suspected but not declared by the authors of the respective publication(s), authors are contacted for clarification. Although this procedure can identify systematic sample overlap, it does not exclude random overlap (e.g., a subject recruited for two independent studies). However, this is a rather rare occurrence. Further, we have recently shown that even an unidentified sample overlap of up to 10% does not appreciably change the meta-analysis results [Lill et al., 2012].

## Timeliness of Data Display

In contrast to "conventional" reviews or meta-analyses, which can provide an excellent snapshot and/or in depth analysis of the genetic knowledge at the time of publication but are often quickly outdated, genetic association data included in a database format can provide up-to-date resources for the respective fields if they are continuously updated. However, this obviously depends a continuous funding of the underlying curational efforts. Our experience shows that funding for the initial development and launch of such databases is still relatively easy, whereas funding for continuous database updates is sometimes more difficult to obtain. When seeking long-term funding, database curators should keep in mind that

this not only requires support for the curatorial team, but should also cover sufficient funds for updates to the database code.

### Other Limitations

In addition to the points listed above, other limitations exist that are inherent to the methodological approach and often cannot be resolved. For instance, the approach developed and adopted by our group aims to assess the cumulative evidence of *main effects*, and as such are usually based on study-level summary data. This often precludes the completion of "refined" analyses such as applying genotyping quality control filters, inclusion of potential confounders, analyses of gene–gene and gene–environment interactions, and—if GWAS data are available only on a summary level—de novo imputation of genotypes (e.g., upon publication of a new reference panel), investigation of copy-number variants, or calculation of individual risk scores. Another limitation is inherent to the concept of probing for genetic *association* between a polymorphism and certain trait, that is, that the genetic variants included need to fulfill a certain minimum frequency threshold to lead to meaningful results. Across our database projects, this amounts to only including polymorphisms with a minor allele frequency of at least 1% in the general population. Thus, dominant mutations or rare alleles with frequencies below 1% are excluded. However, disease-causing mutations are sometimes covered by other online resources dedicated to just this purpose. In the neurogenetics field, excellent examples of such "mutation databases" include the "AD&FTD Mutation Databases" [Cruts and Van Broeckhoven, 1998; Gijselinck et al., 2008], the "PD Mutation Database" [Nuytemans et al., 2010], and "ALSoD," a database for mutations relevant in ALS [Wroe et al., 2008]. To learn more about these databases, please consult the respective websites and accompanying contributions in this special issue of Human Mutation.

## The Shape of Things to Come in Complex Diseases

As outlined above, genetic association databases have not lost their significance in the "GWAS era." Due to the complex nature of GWAS, the contrary appears to be the case, provided that these types of studies can be included. Within the next years, additional GWAS datasets will be generated for many diseases, for instance, in populations and ethnic groups that have thus far been largely neglected. These will need to be included in ongoing meta-analyses of genetic association data. Furthermore, projects using next-generation sequencing technologies to generate genetic association data have already started, and can be expected to pick up momentum over the next decade, eventually fully replacing microarray-based GWAS of current times. Obviously, data generated by these powerful studies will need to be included into existing databases. To this end, Green and Guyer [Green and Guyer, 2011] already stated that "large data catalogues [. . .] are community resources. This calls for policies that maximize rapid data release (harmonized internationally), while respecting the interests of the researchers generating the data and the human participants involved in that research." As outlined in this article, the database structure developed by our group already allows the inclusion of such data (from GWAS) without compromising the privacy of genotyped individuals or the interest of the researchers performing such studies. Eventually, however, the success of genetic association databases will depend predominantly on (1) the availability/sharing of such large scale datasets, and (2) on the availability of continuous funding.

## Conclusion

Online databases represent an excellent resource to comprehensively concatenate and present data from a diverse range of scientific fields, including results from genetic association studies. For the latter, systematic inclusion of large-scale datasets such as GWAS and next-generation sequencing studies are becoming increasingly crucial for the quality, usefulness, and acceptance of such databases. However, systematic inclusion of such studies also represents one of the most difficult steps due to a number of reasons, including computational and/or algorithmic challenges, as well as restrictions in data sharing policies. For nearly a decade, our group has been active in the development and curation of genetic association databases. New developments in the database code now extend this process to the systematic inclusion of large-scale genotyping data, if these are shared. Thus, our approach may serve as a viable model of dealing with the increasing amount of genetic data being generated by thousands of laboratories worldwide to improve our understanding of the genetic forces driving common human diseases.


### Acknowledgments

We thank Dr. Esther Meissner, Maria Liebsch, Brit-Maren M. Schjeide, and Leif M. Schjeide for their continuous work on the AlzGene, PDGene, ALSGene, MSGene, and SZGene databases. We thank all researchers of the respective fields who provided us with data beyond those included in the original publications. We also thank the staff of the Alzheimer Research Forum, in particular, Gabrielle Ströbl and Colin Knep, for their contribution to developing and hosting our databases.

*Disclosure Statement*: The authors declare no conflict of interest.


### References


Allen NC, Bagade S, McQueen MB, Ioannidis JPA, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L. 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet 40:827–834.

Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 39:17–23.

Chatzinasiou F, Lill CM, Kypreou K, Stefanaki I, Nicolaou V, Spyrou G, Evangelou E, Roehr JT, Kodela E, Katsambas A, Tsao H, Ioannidis JPA, et al. 2011. Comprehensive field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma. J Natl Cancer Inst 103:1227–1235.

Collins F. 2010. Has the revolution arrived? Nature 464:674–675.

Cotton RGH, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, et al. 2007. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 39:433–436.

Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, Sherry ST, Manolio TA. 2011. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet 12:730–736.

Cruts M, Van Broeckhoven C. 1998. Molecular genetics of Alzheimer's disease. Ann Med 30:560–565.

Gijselinck I, Van Broeckhoven C, Cruts M. 2008. Granulin mutations associated with frontotemporal lobar degeneration and related disorders: an update. Hum Mutat 29:1373–1386.

Green ED, Guyer MS. 2011. Charting a course for genomic medicine from base pairs to bedside. Nature 470:204–213.

Haworth A, Bertram L, Carrera P, Elson JL, Braastad CD, Cox DW, Cruts M, den Dunnen JT, Farrer MJ, Fink JK, Hamed SA, Houlden H, et al. 2011. Call for participation in the neurogenetics consortium within the Human Variome Project. Neurogenetics 12:169–173.

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.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4:e1000167.

Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, Paschal J, Manolio TA, Tucker M, Hoover RN, Thomas GD, Chanock SJ, Chatterjee N. 2009. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet 41:1253–1257.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26:2204–2207.

Lill CM, Abel O, Bertram L, Al-Chalabi A. 2011. Keeping up with genetic discoveries in amyotrophic lateral sclerosis: the ALSoD and ALSGene databases. Amyotroph Lateral Scler 12:238–249.

Lill CM, Bertram L. 2010. Online-Datenbanken und systematische Metaanalysen komplex-genetischer Erkrankungen. Medizinische Genetik 22: 235–41.

Lill CM, Bertram L. 2011. Towards unveiling the genetics of neurodegenerative diseases. Semin Neurol 31:531–541.

Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide B-MM, Schjeide LM, Meissner E, Zauft U, Allen NC, Liu T, Schilling M, et al. 2012. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. PLoS Genet 8:e1002548.

Lill CM, Schjeide B-MM, Roehr JT, Zauft U, Allen NC, Zipp F, McQueen MB, Kavvoura FK, Ioannidis JPA, Khoury MJ, Tanzi RE, Bertram L. 2010. Correspondence to Sand et al. "Critical reappraisal of a catechol-o-methyltransferase transversion variant in schizophrenia." Biol Psychiatry 67:e45–e48.

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, et al. 2009. Strengthening the reporting of genetic association studies. PLoS Med 6:e22.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Nuytemans K, Theuns J, Cruts M, Van Broeckhoven C. 2010. Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. Hum Mutat 31:763–780.

Sankararaman S, Obozinski G, Jordan MI, Halperin E. 2009. Genomic privacy and limits of individual detection in a pool. Nat Genet 41:965–967.

Sharma M, Ioannidis JPA, Aasly JO, Annesi G, Brice A, Van Broeckhoven C, Bertram L, Bozi M, Crosiers D, Clarke C, Facheris M, Farrer M, et al. 2012. Large-scale replication and heterogeneity in Parkinson disease genetic loci. Neurology 79:1–9.

Siontis KCM, Patsopoulos NA, Ioannidis JPA. 2010. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. Eur J Hum Genet 18:832–837.

Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, Lill CM, Cohen JT, Trikalinos TA. 2012. Towards modernizing the systematic review pipeline in genetics: efficient updating via data mining. Genet Med doi: 10.1038/gim.2012.7.

Wroe R, Wai-Ling Butler A, Andersen PM, Powell JF, Al-Chalabi A. 2008. ALSOD: the amyotrophic lateral sclerosis online database. Amyotroph Lateral Scler 9:249–250.