LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis

Jie Zheng[1]*, A. Mesut Erzurumluoglu[2], Benjamin L. Elsworth[1], Laurence Howe[1], Philip C. Haycock[1], Gibran Hemani[1], Katherine Tansey[1], Charles Laurin[1], Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, Beate St. Pourcain[1], Nicole M. Warrington[3], Hilary K. Finucane[4], Alkes L. Price[4,5], Brendan K. Bulik-Sullivan[5,6], Verneri Anttila[5], Lavinia Paternoster[1],Tom R. Gaunt[1], David M. Evans[1,3†], Benjamin M. Neale[5,6†]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Bristol, UK, [2]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK, [3]University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia, [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA, [5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA, [6]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

*To whom correspondence should be addressed. jie.zheng@bristol.ac.uk

† Joint senior authors.

Abstract

Motivation: LD score regression is a reliable and efficient method of using genome-wide association study (GWAS) summary-level results data to estimate the SNP heritability of complex traits and diseases, partition this heritability into functional categories, and estimate the genetic correlation between different phenotypes. Because the method relies on summary level results data, LD score regression is computationally tractable even for very large sample sizes. However, publicly available GWAS summary-level data are typically stored in different databases and have different formats, making it difficult to apply LD score regression to estimate genetic correlations across many different traits simultaneously.

Results: In this manuscript, we describe LD Hub - a centralized database of summary-level GWAS results for 177 diseases/traits from different publicly available resources/consortia and a web interface that automates the LD score regression analysis pipeline. To demonstrate functionality and validate our software, we replicated previously reported LD score regression analyses of 49 traits/diseases using LD Hub; and estimated SNP heritability and the genetic correlation across the different phenotypes. We also present new results obtained by uploading a recent atopic dermatitis GWAS meta-analysis to examine the genetic correlation between the condition and other potentially related traits. In response to the growing availability of publicly accessible GWAS summary-level results data, our database and the accompanying web interface will ensure maximal uptake of the LD score regression methodology, provide a useful database for the public dissemination of GWAS results, and provide a method for easily screening hundreds of traits for overlapping genetic aetiologies.

Availability and implementation: The web interface and instructions for using LD Hub are available at http://ldsc.broadinstitute.org/

1 Introduction

There is now substantial empirical evidence demonstrating that the majority of complex traits and diseases in humans are influenced by hundreds if not thousands of genetic loci of small effect scattered across the genome as was first predicted a century ago (East 1916; Fisher 1918). The advent of high throughput micro-array genotyping and now next generation sequencing technologies has meant that genome-wide data can be leveraged to ask fundamental questions concerning the underlying genetic architecture of common complex traits and diseases including the degree to which genetic variation affecting complex phenotypes is tagged by SNPs on genome-wide arrays (Yang et al, 2010; Yang et al, 2011; Lee et al, 2012), the degree to which this variation represents different functional categories and/or biological pathways (Yang et al. 2011; Gusev et al. 2014 ;

Finucane et al, 2015), and the extent to which genetic aetiologies are shared across different phenotypes (Lee et al. 2012; Lee et al. 2013; Bulik-Sullivan et al, 2015b). To date most of these types of analyses have been performed using genetic restricted maximum likelihood analysis (GREML) as implemented in software packages such as GCTA and LDAK (Yang et al, 2010; Yang et al, 2011; Lee et al, 2012; Speed et al. 2012). However these methods require individual-level genotype data, which is often not available as most of the largest GWAS analyses are conducted through meta-analyses, and so typically only report summary results statistics (Zheng et al, 2013). Additionally GREML can be computationally prohibitive when analyzing raw genome-wide SNP data from hundreds of thousands of individuals. Consequently, most GREML analyses reported in the literature to date have been hypothesis driven studies that have involved only a small number of related traits (Table 1).

In order to address these limitations, Bulik-Sullivan et al previously proposed a different method, LD score regression (Bulik-Sullivan et al, 2015a). Essentially the method involves regressing summary results statistics from millions of genetic variants across the genome on a measure of each variant's ability to tag other variants locally (i.e. its "LD score"). The intuition behind the approach is that if a trait is genetically influenced, then variants that tag more of the genome (i.e. have high LD scores) should have a greater opportunity to tag causal variants and therefore have higher test statistics on average than variants that have low LD scores. In this way genome-wide inflation of test statistics due to genuine polygenicity can be distinguished from biases such as population stratification and cryptic relatedness. The basic method is very flexible and can be adapted to estimate SNP heritability, calculate a more accurate and efficient genome-wide inflation correction factor than genomic control (Bulik-Sullivan et al, 2015a), partition the SNP heritability by functional category (Finucane et al, 2015), and estimate the genetic correlation between different complex traits and diseases (Bulik-Sullivan et al, 2015b), all using GWAS summary-level results data (Table 1).

The chief limitation of using LD score regression to estimate genetic correlations to date has been a practical one. Publicly available GWAS meta-analysis results are available from a number of different repositories on the Internet. It is time consuming to locate and download all of these resources for use, particularly as these databases become more numerous. What's more, each summary results file typically involves different file formats and conventions making data preparation a time consuming exercise. In addition, many GWAS meta-analyses are not made publicly available, requiring the user to proactively invite the relevant investigators to share their results, which also takes a significant amount of time.

Table 1 Comparison between GREML and LD Score Regression via LD Hub.

GREML	LD Score regression via LD Hub
One dataset at a time	Integrates multiple GWAS results datasets
Run time depends on number of individuals and traits	Run time depends on number of traits only
Manual implementation	Automated
Usually one or a few traits at a time	Many traits simultaneously
Typically hypothesis driven	Hypothesis driven or hypothesis free
Computationally prohibitive for large numbers of individuals	Handles large numbers of individuals easily

Here we describe a centralized database and web interface, LD Hub, which automates the LD score regression analysis pipeline using publically available GWAS summary-level data of individuals with European ancestry. Users of our web-based tool only need to upload summary results for their trait(s) of interest; and the web server will automatically test their results against GWAS results from (currently) 177 other traits/diseases. The proposed database and web interface calculates the SNP heritability for the uploaded phenotype(s), and a genetic correlation matrix across traits. LD Hub allows the user to conduct the analysis on specific phenotypes only or perform a hypothesis free screen across all traits in the database (Table 1). Users have the option of uploading their own results files and the option of adding their GWAS results to the database for inclusion in future releases. The resource is continuously updated and curated every month to include new results from users and publicly available sources alike. The pre-computed genetic correlation matrix will be provided on LD-Hub for all traits included in the database.

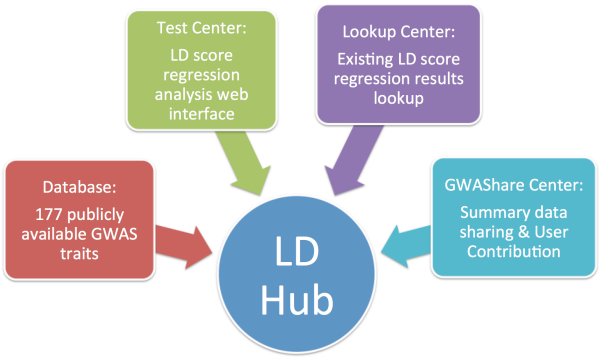


Fig 1. Scope of LD Hub

2 Methods

As summarized in Figure 1, LD Hub includes: 1) Lookup Center: a facility to perform lookups of existing LD score regression results; 2) Database: a GWAS summary-level statistics database, 3) Test Center: a web interface that automates the LD score regression analysis pipeline including the calculation of SNP heritability and genetic correlations, and 4) GWAShare Center: a user contribution and data sharing platform

2.1 LD Hub database

2.1.1 GWAS summary-level data

We cleaned and harmonized 844 publicly available GWAS summary-level data sets from 35 consortia, which included 82 diseases, 154 complex traits, 454 metabolites and 151 immune markers (Hemani et al, 2016).

From this database pool, we chose datasets that fit the following selection criteria:

1. Non-sex-stratified
2. Meta-analyses of European populations.
3. Meta-analyses using a GWAS backbone chip only (i.e. exclude meta-analyses involving immuno | metabo | psych | exome chip or GWAS + custom chip)
4. Number of SNPs is large (N>450,000)
5. Number of individuals is large (N>5,000)
6. Mean Chi-square of the test statistics is larger than 1

As shown in Figure 2, after filtering on the selection criteria, genome-wide results for 177 traits were included in LD Hub, of which 23 are GWAS of diseases, 47 are medically relevant risk factors/traits and 107 are metabolites. Table S1, displays descriptive information for each of the GWAS in LD Hub, including, trait name, consortium name, ethnicity, gender, number of cases and controls, sample size, PubMed ID, year of publication, and other relevant information. We used an inclusive strategy for data selection, so we show a list of traits that may return null results in Table S2 due to low Z score of heritability estimate.

2.1.2 LD score information

We pre-calculated LD scores for each SNP using individuals of European ancestry from the 1000 Genomes project (1000 Genomes Project Consortium, 2012). These LD scores are suitable for standard LD score analyses in European populations (i.e. the LD score regression intercept, heritability, genetic correlation, cross-sex genetic correlation).

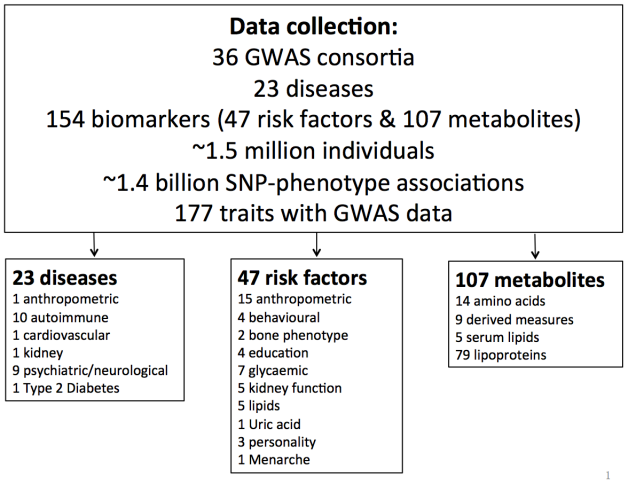


Fig 2. Contents of LD Hub. In total, data for 177 traits are included in LD Hub, which consist of 23 diseases, 47 complex traits and 107 metabolites.

2.2 LD Hub web interface

The LD Hub web interface framework was developed using Python Django as the LD score regression program is also written using Python.

2.2.1 Test Center

The LD Hub web interface provides an automatic LD score regression analysis pipeline for users. As shown in Figure 3, the LD Hub analysis pipeline consists of 5 major steps:

1. User login system: using a Google OAuth
2. File upload system: To run the LD score analysis pipeline, LD Hub requires upload of a file containing summary results data. In the web interface, we provide an example GWAS results file to illustrate the file format required for successful upload and analysis by LD Hub. To save uploading time, each results file should be a white space zipped file in which each row contains the results from a single SNP whilst the columns comprise the following fields:
   a) SNP ID (rs number)
   b) Effect allele of the SNP
   c) Alternate allele of the SNP
   d) Sample size of each SNP (can use an overall sample size if sample size for some SNPs is missing)
   e) A signed summary statistic where the sign refers to the addition of the effect allele (i.e. any statistic that can be converted into a Z-score)
   f) P value of the SNP
   g) Minor allele frequency of each SNP (optional)
   h) SNP Imputation quality (optional)
3. Quality control and heritability analysis: To standardize the input file, quality control is automatically performed on the uploaded file.
   a) For studies that provide sample MAF, a filter to include SNPs with MAF above 1%.
   b) In order to restrict the analysis to well-imputed SNPs, we filter the uploaded SNPs to HapMap3 SNPs (International HapMap 3 Consortium et al, 2010) with 1000 Genomes EUR MAF above 5%, which tend to be well-imputed in most studies.
   c) If sample size varies from SNP to SNP, remove SNPs with an effective sample size less than 0.67 times the 90th percentile of sample size.
   d) Remove INDELs and structural variants.
   e) Remove strand-ambiguous SNPs.
   f) Remove SNPs whose alleles do not match those in the 1000 Genomes data.
   g) Remove SNPs within the major histocompatibility complex (MHC) region since these often display extreme LD and effect sizes.
   h) Because outliers can unduly influence the regression, we also removed SNPs with extremely large effect sizes ($X_1^2 >$ 80).

   The second part of this step is the SNP heritability analysis. The results of this analysis provide a useful indication of whether genetic correlation analysis is likely to be informative (Bulik-Sullivan et al. 2015).
4. Genetic correlation analysis. If the uploaded GWAS results are likely to provide good statistical power (i.e. heritability H^2 Z score of > 4), then the LD Hub pipeline will perform genetic correlation analysis on the uploaded GWAS results. Users have the option of selecting which traits they want to include in the analysis.

5. Reporting of results.

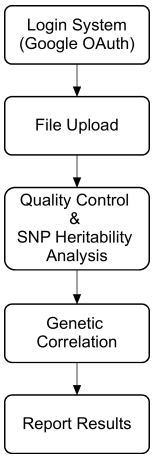


Fig 3. Schematic of LD Hub workflow.

2.2.2 Lookup Center

Another feature of the LD Hub web interface is the heritability and genetic correlation 'lookup' function for existing GWAS results currently in the database. In the current version (v1.0), we provide SNP heritability and genetic correlation results. A forest plot for heritability and a genetic correlation matrix plot are also provided.

2.2.3 GWAShare Center

We aim to promote sharing of summary GWAS results data. We encourage users of LD Hub to upload their GWAS results for curation into the database. We will update the database regularly and allow other users to use the shared data for LD score regression analyses, which will then benefit the whole human genetics community.

2.3 Case study of Atopic dermatitis

In order to illustrate the utility of LD Hub, we conduct an analysis using summary results data from a large GWAS of atopic dermatitis (AD) for 40835 (10,788 cases and 30,047 controls, sample prevalence: 0.264) European individuals (i.e. the discovery cohorts from this paper minus results from 23ANDME) (EAGLE consortium 2015). In total, 11059640 SNPs were included in this meta-analysis. Since AD is influenced by a gene of major effect (i.e. filaggrin) which could bias estimates from LDHub, we excluded this region from the uploaded results file. After quality control, 1215002 SNPs were selected for upload.

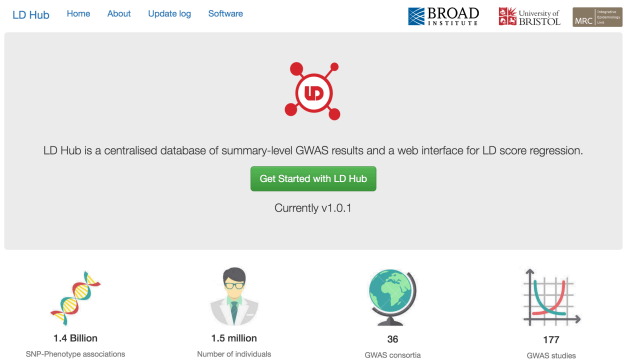


Fig. 4. Screen shot of LD Hub web interface.

3 Results

3.1 Validation of LD Hub analysis results

We tested the validity and functionality of LD Hub by replicating previously reported results from the original LD Score regression suite of papers (Bulik-Sullivan et al, 2015a, Bulik-Sullivan et al, 2015b).

We compared SNP heritability results between LD Hub and previously reported LD score regression results (Bulik-Sullivan et al, 2015a). As shown in Table S3, the Mean χ^2, λ_{GC} and Intercept results are almost the same. The minor discrepancy is a consequence of using slightly different quality control processes for LD Hub than what was used in the original LD Score regression paper. Results for SNP heritability of 177 traits are shown in Table S1.

We also compared the genetic correlation analysis results between LD Hub and previously reported results (Bulik-Sullivan et al, 2015b). As shown in Figure 5, the genetic correlation and standard error of genetic correlation estimates are consistent with previously reported LD score regression genetic correlation results. A comparison of the genetic correlation results of (previously reported) 49 traits is shown in Table S4.

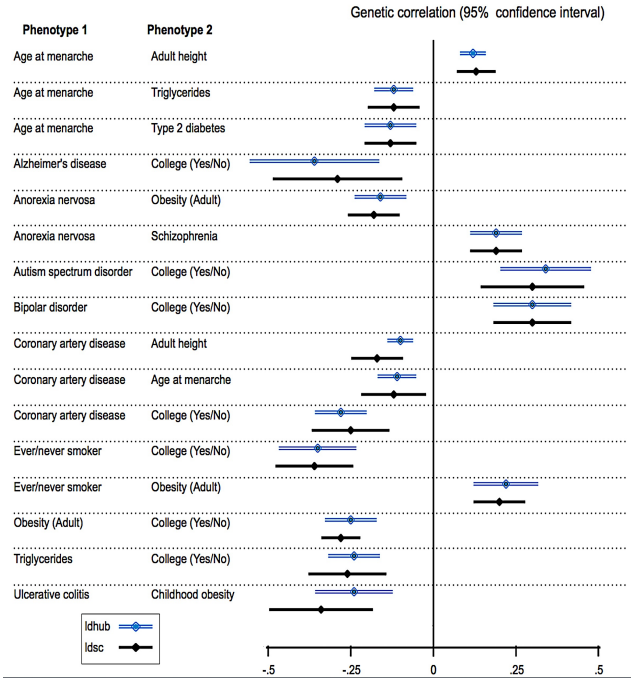


Fig 5. Comparison of genetic correlation results between LD Hub and previously reported LD score regression results. Double blue lines represent genetic correlation results from LD Hub;and the black single lines represent genetic correlation results from previously reported LD score regression results. The discrepancies can be attributed to the minor changes in the quality control processes and the replacement of some GWAS results with the more recent versions.

3.2 Case study: Atopic Dermatitis

Table 2 shows the SNP heritability for AD. The figure of 7.8% is low particularly compared to the heritability estimates from twin studies of eczema where figures exceeding 80% are not uncommon (Bataille et al. 2012). This could be for a number of reasons including the fact that genomic control correction in the individual meta-analysis studies causes downward bias, the filaggrin regions of the genome were excluded from the analysis, and the fact that LD score regression provides an estimate of the overall proportion of additive genetic variance tagged by SNPs in the GWAS panel (i.e. SNP heritability), rather than total heritability *per se*. However the greatest contributing factor is likely to be the case definition of AD used in the EAGLE consortium paper which is extremely heterogeneous, relying often on self-report or retrospective recall which will introduce substantial measurement error into the analysis (and hence decrease heritability estimates). Our results strongly suggest that reanalysis using a more precise definition of eczema would result in a cleaner phenotype and consequently increase the number of genome-wide significant loci detected.

Table 2. SNP heritability for atopic dermatitis. H^2 and SE_H^2 refer to the SNP heritability and standard error of the SNP heritability.

Type of Heritability Scale	H^2	SE_H^2	λ_{GC}	Mean χ^2	Intercept
Observed Scale	0.071	0.016	1.053	1.080	1.034
Liability Scale	0.078	0.018	1.053	1.080	1.034

Table 3 displays estimated genetic correlations between AD and several immune mediated diseases and lung cancer recorded in LD Hub. As expected, the estimated genetic correlation (rG) between AD and asthma was strongly significant and positive. We

also note that the rG between AD and Crohn's disease was moderate, significant and positive, perhaps reflecting substantial overlap between currently known loci for both conditions (Paternoster et al. 2015). rG did not differ significantly from zero for the other traits, although the point estimates for several were moderate indicating that follow up when larger samples become available may be justified.

Table 3. Genetic correlation between atopic dermatitis and other immune mediated diseases. rG refers to the genetic correlation between two traits, SE_rG is the standard error of the genetic correlation, P_rG is the p value of the genetic correlation. The full set of genetic correlation results between AD and other selected traits are included in Table S5.

Traits	rG	SE_rG	P_rG
Crohn's disease	0.1823	0.0864	0.0347
Ulcerative colitis	0.0975	0.0957	0.3083
Asthma	0.551	0.1474	0.0002
Celiac disease	-0.1551	0.1251	0.2151
Multiple sclerosis	-0.2377	0.1944	0.2331
Primary biliary cirrhosis	-0.0064	0.1216	0.9582
Systemic lupus erythematosus	0.0842	0.1125	0.4541
Rheumatoid Arthritis	-0.0705	0.0846	0.4047
Adenoma	-0.0601	0.2167	0.7815
Squamous	0.1879	0.1626	0.2481
Overall Lung Cancer	-0.0106	0.1057	0.9202

4 Discussion

In this paper, we describe LD Hub (accessible at http://ldsc.broadinstitute.org/), a web-based utility that centralizes and harmonizes summary-level GWAS results data, and automates LD Score regression analysis (Bulik-Sullivan et al, 2015a, Bulik-Sullivan et al, 2015b).

GWAS meta-analysis summary statistics are increasingly being made publicly available. Our database (currently) utilizes results from 177 different GWAS, which includes all publicly available GWAS summary results suitable for LD Score regression (Bulik-Sullivan et al., 2015a). However, this represents a small proportion of the traits represented in the GWAS Catalog (https://www.ebi.ac.uk/gwas/) (Hindorff et al; Welter et al, 2014). There is thus an urgent need for increased sharing of GWAS meta-analysis results in order to realize the full potential of techniques that utilize summary results data such as LD score regression. LD Hub provides a natural platform for the distribution of summary results data that can be utilized by the whole genetics community.

There are four major advantages of using our database and web interface:

1) Users of LD Score regression currently spend most of their time reformatting, harmonizing and managing summary results data rather than running the 'actual' analyses. LD Hub minimizes the proportion of time spent on the former so that users can focus their attention on interpreting interesting genetic correlations and SNP heritabilities.

2) Users who do not have a computational background will find the interface easier to use

3) The software is computationally very fast. The current version (v1.0) can return the systematic analysis results to the user within few hours. A queuing system has been introduced to prevent the server from crashing.

(4) As users upload and share their own summary GWAS results, the resource becomes increasingly useful.

We envisage LD Hub as a useful hypothesis generating tool, providing an easy method of screening hundreds/thousands of traits for interesting genetic correlations that could subsequently be followed up in further detail by other approaches such as pathway analysis (Segre et al. 2010) or Mendelian randomization (Davey-Smith & Ebrahim 2003). For example, under most models, a causal relationship between two heritable traits should induce a genetic correlation between the two phenotypes (assuming individual differences in the causal trait are influenced by genetic variation). LD Hub could be used to screen a large number of putatively causally related phenotypes quickly and easily for evidence of genetic correlation, and the most promising candidate pairs could then be followed up by selecting appropriate genetic instruments and performing formal instrumental variables analysis (Evans & Davey-Smith, 2015, Hemani et al, 2016). This framework could be particularly useful in the dissection of high dimensional molecular networks where the number of possible pair-wise relationships may be extremely large.

For LD Hub, we list few suggestions / limitations here:

1. In order for estimates of the genetic correlation to be reliable we suggest that traits uploaded meet the following criteria
   - Heritability (H^2) Z score is at least > 1.5 (optimal > 4)
   - Mean Chi square of the test statistics > 1.02
   - The intercept estimated from the SNP heritability analysis is between 0.9 to 1.1

2. As we aim to provide an analysis pipeline that is as systematic as possible, we used a very inclusive strategy for data selection, where we expect a very small proportion of the analyses (especially for the traits listed in Table S2) to return null results.
3. LD Hub is currently designed for GWAS studies involving European populations exclusively. As the number of publicly available GWAS involving other ethnicities increases we will extend LD Hub to include these.
In summary, due to the growing availability of summary-level data, our database together with the web interface will maximize the potential of GWAS summary-level data for heritability and genetic correlation analyses.

Acknowledgements

We would like to thank Kaitlin Wade, Vanessa Tan, Ryan Langdon and James Yarmolinsky for helping curate the GWAS data information.

Funding

This work was supported by the Medical Research Council program grant (MC_UU_12013/4 and MC_UU_12013/8). D.M.E. is supported by an Australian Research Council Future Fellowship (FT130101709). This work was in part supported by Cancer Research UK programme grant number C18281/A19169 (the Integrative Cancer Epidemiology Programme). P.H. is a Cancer Research UK Population Research Fellow, grant number C52724/A20138.

Conflict of Interest: **None declared.**

References

1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. 491(7422):56-65. doi: 10.1038/nature11632.

Bataille V., et al. (2012) The use of the twin model to investigate the genetics and epigenetics of skin diseases with genomic, transcriptomic and methylation data. J Eur Acad Dermatol Venereol. 26(9):1067-73.

Bulik-Sullivan, et al. (2015). LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nature Genetics*, 47(3):291-5.

Bulik-Sullivan, et al. (2015). An Atlas of Genetic Correlations across Human Diseases and Traits. *Nature Genetics*, doi: 10.1038/ng.3406.

Davey Smith, G. & Ebrahim, S. (2003). Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 32, 1–22.

Davey Smith, G. & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. 23(R1), R89–R98.

Do, R. et al. Common variants associated with plasma triglycerides and risk for
coronary artery disease. Nat. Genet. 45, 1345–1352 (2013).

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. 9, e1003348.

EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, et al. (2015) Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet. 47(12):1449-56.

E. M. East, E.M. (1916) Studies on size inheritance in nicotiana. Genetics. 1916 Mar; 1(2): 164–176.

Evans, DM. et al,. (2013). Mining the human phenome using allelic score that index biological intermediates. PLoS Genet. 9 (10).

Evans DM and Davey Smith G., (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annu Rev Genomics Hum Genet. 2015;16:327-50.

Finucane, HK, et al. (2015) Partitioning Heritability by Functional Category using GWAS Summary Statistics. Nat Genet. 47(11):1228-35.

Fisher RA,. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh. volume 52, pages 399–433

Gusev A, et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 95(5):535-52.

Hemani G, et al. (2016). MR-Base: a platform for two-sample Mendelian randomization using summary data from genome-wide association studies. In preparation.

Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [date of access].

Lee H et al (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, **44**(3):247–250.

Pasaniuc B, et al. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*. **30**(20):2906-14.

Purcell, S.M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752 (2009).

Segrè AV., et al. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6(8).

Skaaby T., et al (2014). Atopy and development of cancer: a population-based prospective study. J Allergy Clin Immunol Pract. 2014 Nov-Dec;2(6):779-85.

Speed D et al., (2012). Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 91(6):1011-21.

Stahl EA et al, (2012) Daniel Wegmann, Gosia Trynka, Javier Gutierrez-Achury, Ron Do, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genetics, **44**(5):483–489.

Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. Am. J. Hum. Genet. 90, 7–24 (2012).

Voight, B.F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a
Mendelian randomisation study. Lancet 380, 572–580 (2012).

Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, Vol. 42 (Database issue): D1001-D1006.

Yang J et al. (2010) Common snps explain a large proportion of the heritability for human height. Nature Genetics, **42**(7):565–569.

Yang J, et al (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**(1):76–82.

Zheng J et al,. (2013). Sequential sentinel SNP Regional Association Plots (SSS-RAP): an approach for testing independence of SNP association signals using meta-analysis data. Ann Hum Genet. 2013 Jan;77(1):67-79.

References for GWAS data included in the LD Hub:

Anderson,C. et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nature Genetics, 43, 246-252.

Bentham,J. et al. (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nature Genetics, 47, 1457-1464.

Benyamin,B. et al. (2013) Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. Molecular Psychiatry, 19, 253-258.

Berndt,S. et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nature Genetics, 45, 501-512.

Boger,C. et al. (2011) CUBN Is a Gene Locus for Albuminuria. Journal of the American Society of Nephrology, 22, 555-570.

Boraska,V. et al. (2014) A genome-wide association study of anorexia nervosa. Molecular Psychiatry, 19, 1085-1094.

Bradfield,J. et al. (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. Nature Genetics, 44, 526-531.

Cordell HJ, et al. (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun. 6:8019.

Dastani Z et al. (2012). Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. 8(3):e1002607.

de Moor,M. et al. (2010) Meta-analysis of genome-wide association studies for personality. Molecular Psychiatry, 17, 337-349.

Dubois,P. et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics, 42, 295-302.

Dupuis,J. et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature Genetics, 42, 105-116.

Estrada,K. et al. (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nature Genetics, 44, 491-501.

Franke,A. et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics, 42, 1118-1125.

Furberg,H. et al. (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature Genetics, 42, 441-447.

Horikoshi,M. et al. (2012) New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nature Genetics, 45, 76-82.

Huffman,J. et al. (2015) Modulation of Genetic Associations with Serum Urate Levels by Body-Mass-Index in Humans. PLOS ONE, 10, e0119752.

Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis (2013) The Lancet, 381, 1371-1379.

KÃttgen,A. et al. (2010) New loci associated with kidney function and chronic kidney disease. Nature Genetics, 42, 376-384.

Lambert,J. et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics, 45, 1452-1458.

Lango Allen,H. et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467, 832-838.

Liu JZ, et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 47(9):979-86.

Mahajan,A. et al. (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature Genetics, 46, 234-244.

Manning,A. et al. (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nature Genetics, 44, 659-669.

Morris,A. et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics, 44, 981-990.

Moffatt MF, et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 448(7152):470-3.

Neale,B. et al. (2010) Meta-Analysis of Genome-Wide Association Studies of Attention-Deficit/Hyperactivity Disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 49, 884-897.

Okada,Y. et al. (2013) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 506, 376-381.

Perry,J. et al. (2014) Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature, 514, 92-97.

Prokopenko,I. et al. (2014) A Central Role for GRB10 in Regulation of Islet Function in Man. PLoS Genetics, 10, e1004235.

Rietveld,C. et al. (2014) Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. Proceedings of the National Academy of Sciences, 111, 13790-13794.

Rietveld,C. et al. (2013) GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. Science, 340, 1467-1471.

Ripke,S. et al. (2012) A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry, 18, 497-511.

Ripke,S. et al. (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511, 421-427.

Sawcer,S. et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature, 476, 214-219.

Saxena,R. et al. (2010) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nature Genetics, 42, 142-148.

Schunkert,H. et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genetics, 43, 333-338.

Shin,S. et al. (2014) An atlas of genetic influences on human blood metabolites. Nature Genetics, 46, 543-550.

Shungin,D. et al. (2015) New genetic loci link adipose and insulin biology to body fat distribution. Nature, 518, 187-196.

SimÃ³n-SÃ¡nchez,J. et al. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature Genetics, 41, 1308-1312.

Sklar,P. et al. (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics, 43, 977-983.

Soranzo,N. et al. (2010) Common Variants at 10 Genomic Loci Influence Hemoglobin A1C Levels via Glycemic and Nonglycemic Pathways. Diabetes, 59, 3229-3239.

Speliotes,E. et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics, 42, 937-948.

Stahl,E. et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nature Genetics, 42, 508-514.

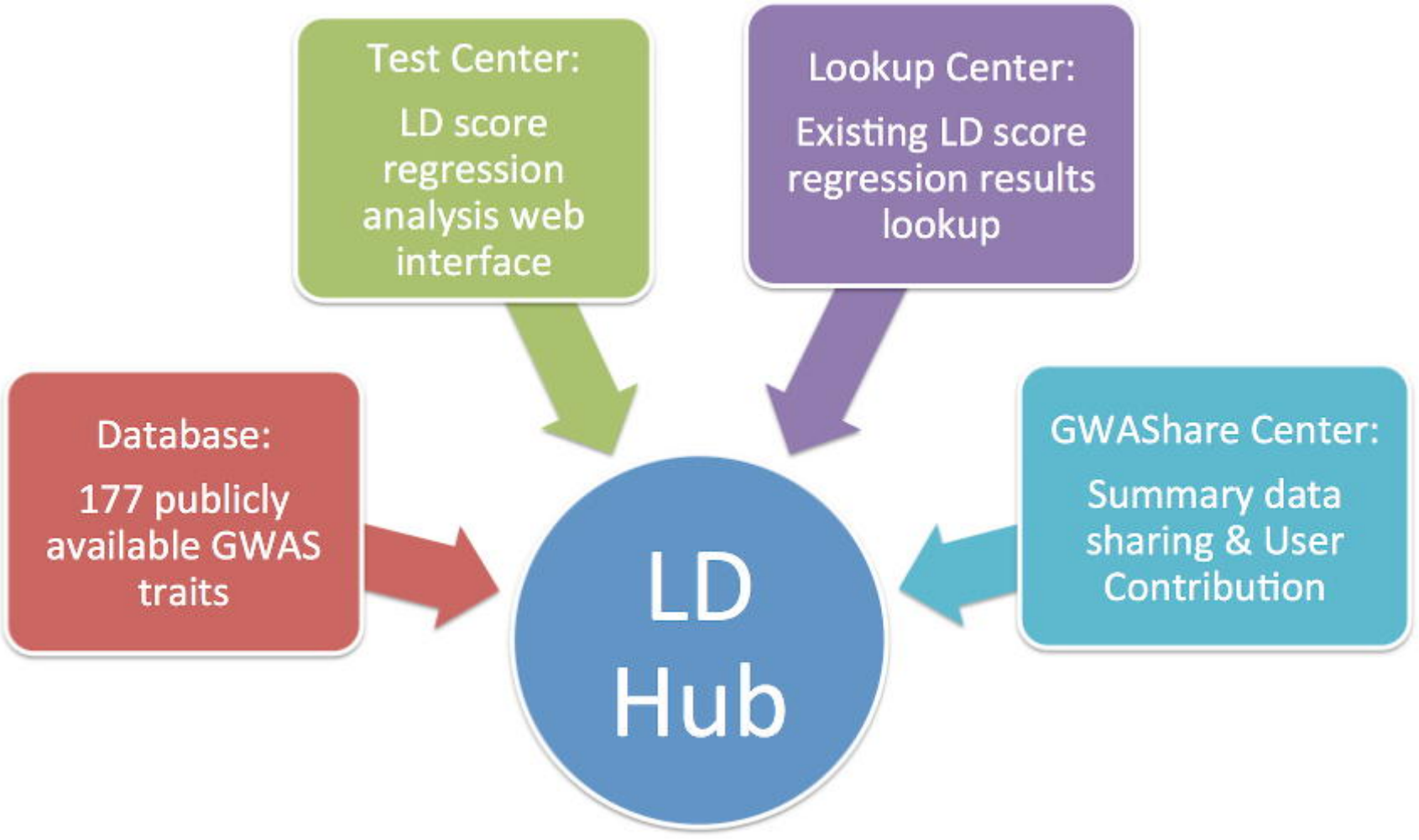
Taal,H. et al. (2012) Common variants at 12q15 and 12q24 are associated with infant head circumference. Nature Genetics, 44, 532-538.

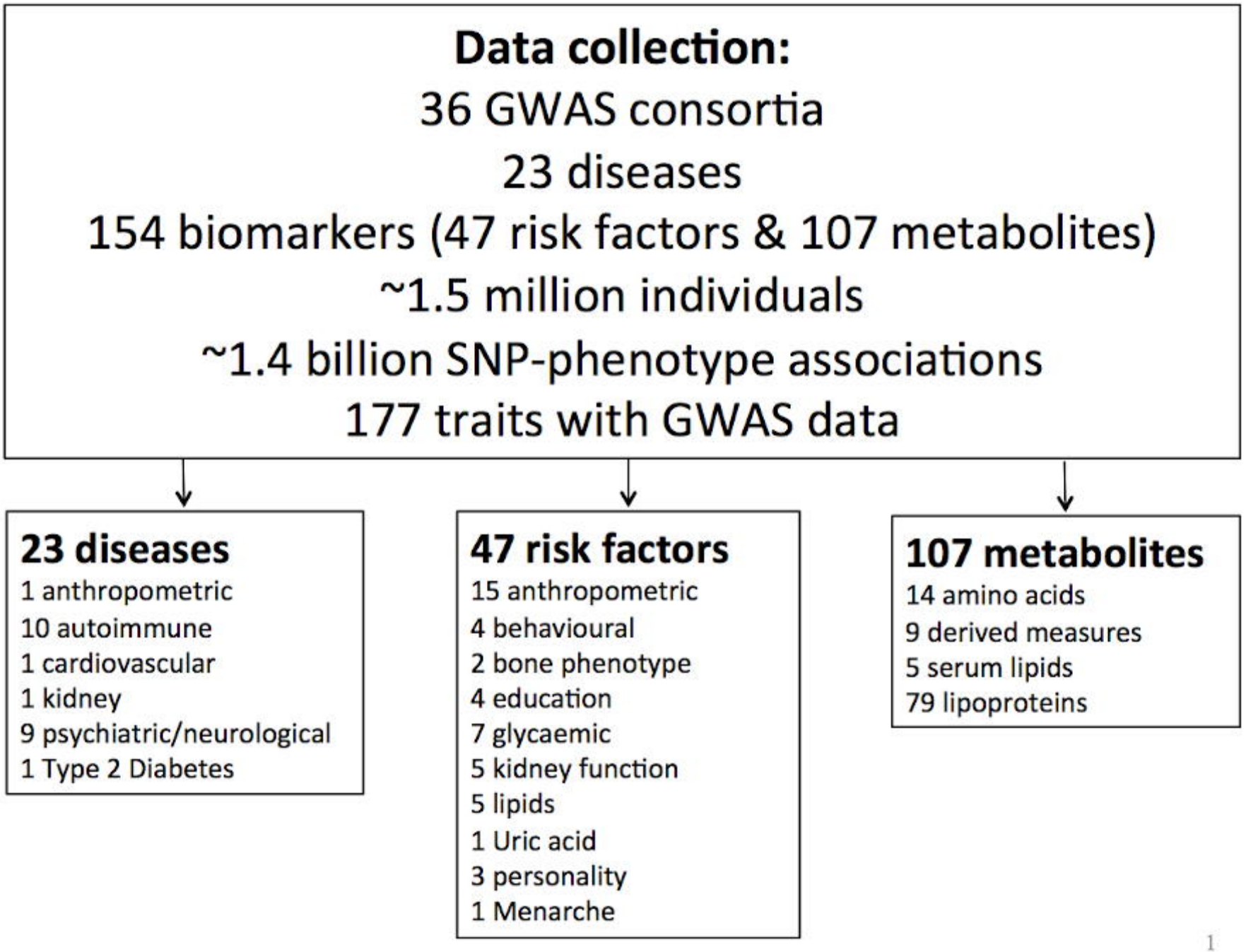
Teslovich,T. et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature, 466, 707-713.

van den Berg,S. et al. (2014) Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory. Behav Genet, 44, 295-313.

van der Valk,R. et al. (2014) A novel common variant in DCST2 is associated with length in early life and height in adulthood. Human Molecular Genetics, 24, 1155-1168.

Yang,J. et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. Nature, 490, 267-272.


Test Center:
LD score regression analysis web interface
Lookup Center:
Existing LD score regression results lookup
Database:
177 publicly available GWAS traits
LD Hub
GWAShare Center:
Summary data sharing & User Contribution


Data collection:
36 GWAS consortia
23 diseases
154 biomarkers (47 risk factors & 107 metabolites)
~1.5 million individuals
~1.4 billion SNP-phenotype associations
177 traits with GWAS data
23 diseases
1 anthropometric
10 autoimmune
1 cardiovascular
1 kidney
9 psychiatric/neurological
1 Type 2 Diabetes
47 risk factors
15 anthropometric
4 behavioural
2 bone phenotype
4 education
7 glycaemic
5 kidney function
5 lipids
1 Uric acid
3 personality
1 Menarche
107 metabolites
14 amino acids
9 derived measures
5 serum lipids
79 lipoproteins


Login System
(Google OAuth)
File Upload
Quality Control
&
SNP Heritability
Analysis
Genetic
Correlation
Report Results

LD Hub is a centralised database of summary-level GWAS results and a web interface for LD score regression.

Get Started with LD Hub

Currently v1.0.1

1.4 Billion

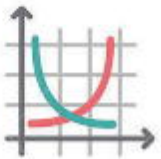
SNP-Phenotype associations

1.5 million

Number of individuals

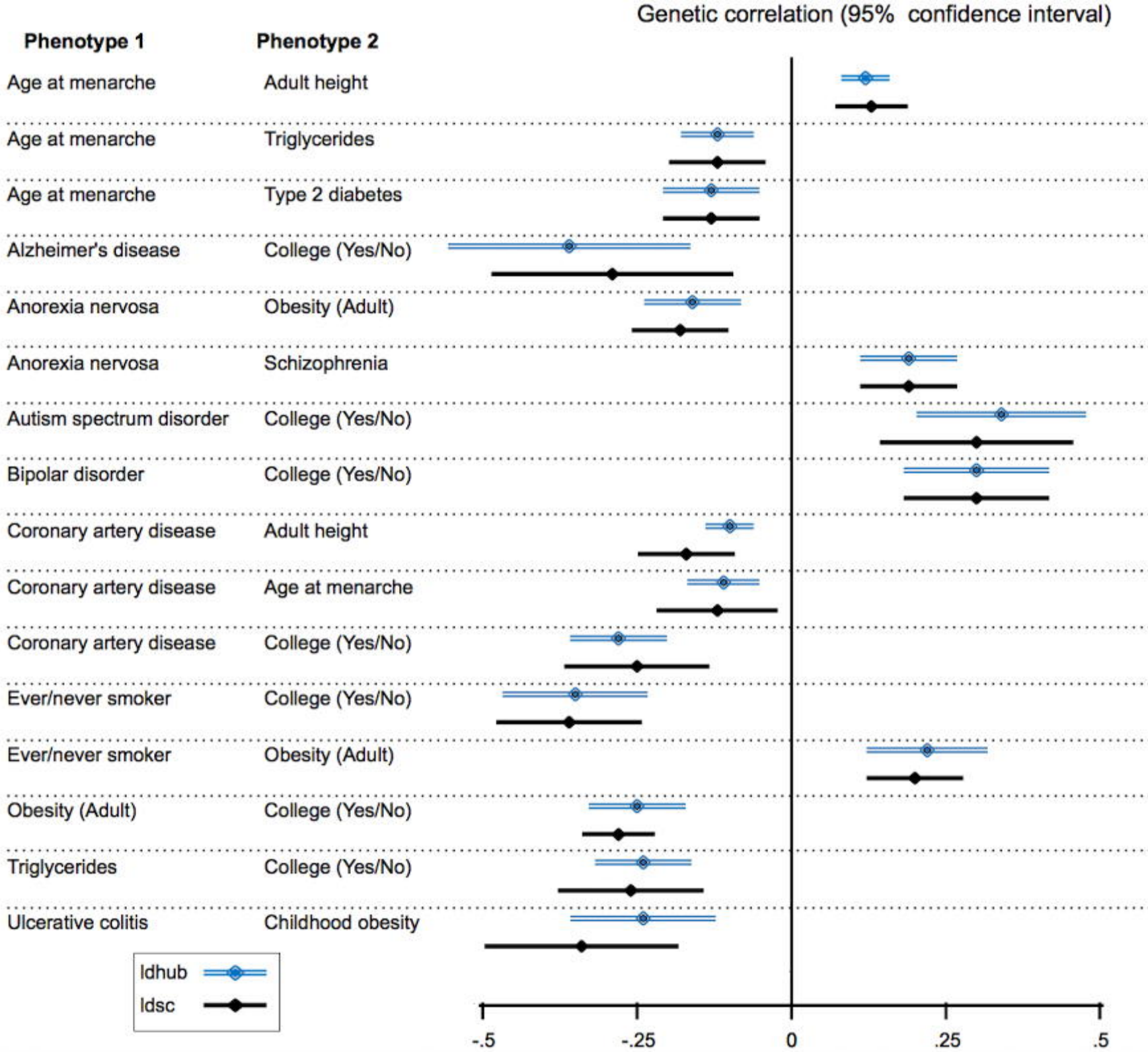
36

GWAS consortia



177

GWAS studies


Genetic correlation (95% confidence interval)
Phenotype 1
Phenotype 2
Age at menarche
Adult height
Age at menarche
Triglycerides
Age at menarche
Type 2 diabetes
Alzheimer's disease
College (Yes/No)
Anorexia nervosa
Obesity (Adult)
Anorexia nervosa
Schizophrenia
Autism spectrum disorder
College (Yes/No)
Bipolar disorder
College (Yes/No)
Coronary artery disease
Adult height
Coronary artery disease
Age at menarche
Coronary artery disease
College (Yes/No)
Ever/never smoker
College (Yes/No)
Ever/never smoker
Obesity (Adult)
Obesity (Adult)
College (Yes/No)
Triglycerides
College (Yes/No)
Ulcerative colitis
Childhood obesity
ldhub
ldsc
-.5
-.25
0
.25
.5