Databases and ontologies

PhenoScanner: a database of human genotype–phenotype associations

James R. Staley1,*; James Blackshaw1; Mihir A. Kamat1; Steve Ellis1; Praveen Surendran1; Benjamin B. Sun1; Dirk S. Paul1; Daniel Freitag1; Stephen Burgess1; John Danesh1,2,3; Robin Young1,4,† and Adam S. Butterworth1,3,†

1Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; 2Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; 3NIHR Blood and Transplant Research Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; and 4Robertson Centre for Biostatistics, University of Glasgow, Glasgow G12 8QQ, UK

†The authors wish it to be known that, in their opinion, the last 2 authors should be regarded as Joint Last Authors.

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on January 26, 2016; revised on May 13, 2016; accepted on June 6, 2016

Abstract

Summary: PhenoScanner is a curated database of publicly available results from large-scale genetic association studies. This tool aims to facilitate ‘phenome scans’, the cross-referencing of genetic variants with many phenotypes, to help aid understanding of disease pathways and biology. The database currently contains over 350 million association results and over 10 million unique genetic variants, mostly single nucleotide polymorphisms. It is accompanied by a web-based tool that queries the database for associations with user-specified variants, providing results according to the same effect and non-effect alleles for each input variant. The tool provides the option of searching for trait associations with proxies of the input variants, calculated using the European samples from 1000 Genomes and Hapmap.

Availability and Implementation: PhenoScanner is available at www.phenoscanner.medschl.cam.ac.uk.

Contact: jrs95@medschl.cam.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Genome-wide association studies (GWAS) have discovered thousands of associations between genetic variants and a wide range of human phenotypes, yielding novel insights into disease aetiology. However, a key challenge for the human genomics community is to develop methods that enable efficient cross-referencing of a genetic variant with a wide range of phenotypes, such as disease states, physiological parameters, cellular traits and other characteristics. Such ‘phenome scans’ could help inform a range of analyses, such as Mendelian randomization analyses, in which genetic variants are used as proxies for modifiable risk factors to attempt to infer causality between traits and diseases (Burgess and Thompson, 2015). Identifying the broad phenotypic consequences of perturbing a particular pathway (indexed by a genetic variant) could also enhance biological understanding and provide insights relevant to the identification and prioritization of potential therapeutic targets, such as the re-purposing of existing therapies to new disease indications and the anticipation of safety and efficacy signals in clinical trials. One notable example has been our demonstration, following a phenome scan across a wide range of traits and diseases, that genetic variants that upregulate the interleukin-1 receptor antagonist are associated with a higher risk of coronary artery disease, partly...
mediated through elevation of pro-atherogenic lipids (Interleukin 1
Genetics Consortium et al., 2015). So far, however, it has been diffi-
cult to generalize this approach, partly because the collation of asso-
ciations with many phenotypes can be time-consuming, especially if
information is sought about multiple variants.

Catalogues of GWAS results already exist, such as the NHGRI-
EBI GWAS catalog (Welter et al., 2014), as well as data repositories,
e.g. dbGaP (Mailman et al., 2007). However, these either focus on
variants robustly associated with a particular trait (and hence do not
take advantage of the wide range of publicly available full GWAS re-
results), do not contain estimates or directions of effect, and/or are dif-
dicult to search in a systematic way. Also, the results often have
inconsistent formats and the output for each variant is not necessarily
given according to the same effect allele. In addition, most cata-
logenues of GWAS do not identify associations with proxy variants,
which means that if an association between the variant of interest
and a trait is unavailable, a suitable proxy must be found using a
separate resource and then searched in the catalogue. Some of the
latest variant annotation tools (which include proxy look-ups) do
contain results from the NHGRI-EBI GWAS catalog (e.g. SNiPA;
Arnold et al., 2014), however, they only return \( P \) values. To help ad-
dress these issues, we developed a web-based tool ‘PhenoScanner’
that extracts and aligns associations for user-specified variants and
proxies across a large curated database.

2 Methods
PhenoScanner consists of a Perl interface (with R command line
tool) that connects to a MySQL database. To develop the initial
database, we collated 137 genotype–phenotype association datasets,
including results for anthropometric traits, blood pressure, lipids,
cardiometabolic diseases, renal function measures, glycemic traits,
inflammatory diseases, psychiatric diseases and smoking phenotypes
(Supplementary Table). We also included the NHGRI-EBI GWAS
catalog, NHLBI GRASP (Leslie et al., 2014) and dbGaP catalogues of
associations. To ensure consistent formatting, we aligned alleles
to the plus strand, added or updated chromosome positions to build
37 using dbSNP (release 138) (Sherry et al., 2001) and liftOver
(https://genome.ucsc.edu/cgi-bin/hgLiftOver), and updated old rsIDs
to dbSNP release 141 (Supplementary Data). Linkage disequilibrium
(LD) measures between neighbouring variants in the autosomal
chromosomes were calculated using the phased haplotypes from
European samples in 1000 Genomes phase 3 (\( N = 503 \)) (1000
Genomes Project Consortium et al., 2012). Variants with minor al-
lee frequencies \(<0.5\%\) were removed along with multiallelic vari-
ants and large indels \((\geq 5\ \text{bases})\). For each remaining variant, we
-calculated \( D’ \) and \( r^2 \) for variants within 300 kb in either direction,
-and kept LD statistics for pairs of variants with \( r^2 \geq 0.6 \). LD statis-
tics based on the CEU population from Hapmap 2 release 24 (Frazer
et al., 2007) are also available (Supplementary Data).

The user may enter either one variant into the text box on the
website or upload up to 50 variants in a text file. The Perl interface
annotates the variant alleles using dbSNP, identifies proxies of the
specified variants (if requested) in the database according to a user-
specified pairwise \( r^2 \) threshold, and queries the catalogue of geno-
type–phenotype associations for the specified variants and their
proxies. Association results are collated and presented with respect
to the same effect and non-effect alleles for each variant. The associ-
ations with proxies are aligned according to the effect and non-
effect alleles of the corresponding primary variant of interest for
added ease of interpretation. The output is a file of associations,
which is made available to download. There is also a \( P \) value filter

\[
\begin{align*}
\text{(a) Trait} & \quad \text{P-value} \\
\text{High Density Lipoprotein*} & \quad 3.0e-02 \\
\text{Low Density Lipoprotein*} & \quad 7.3e-02 \\
\text{Triglycerides*} & \quad 2.0e-01 \\
\text{Body Mass Index*} & \quad 4.2e-01 \\
\text{Height*} & \quad 8.3e-01 \\
\text{Walt Circumference*} & \quad 4.6e-01 \\
\text{Walt Hip Ratio*} & \quad 9.0e-01 \\
\text{Fasting Glucose*} & \quad 3.3e-01 \\
\log[(Fasting Insulin)*] & \quad 7.0e-01 \\
\end{align*}
\]

\[
\begin{align*}
\text{SD change (per A allele)} \quad 0 & \quad 0.02 & \quad 0 & \quad 0.02 \\
\end{align*}
\]

Fig. 1. Association results for rs10840293 with a subset of the traits (a) and
diseases (b) available in PhenoScanner. An asterisk indicates the use of a
proxy variant (rs93138; \( r^2 = 0.95 \)) in reporting the association. SD, standard
deviation; OR, odds ratio

\[
\begin{align*}
\text{(b) Trait} & \quad \text{P-value} \\
\text{Coronary Artery Disease} & \quad 1.3e-08 \\
\text{Rheumatoid Arthritis} & \quad 9.4e-04 \\
\text{Type II Diabetes*} & \quad 4.6e-01 \\
\text{Asthma*} & \quad 7.0e-01 \\
\text{Schizophrenia} & \quad 8.1e-01 \\
\end{align*}
\]

option that only retains results with study-specific \( P \) values less than
the selected threshold.

3 Results
To illustrate the use of PhenoScanner, we ran the program with
rs10840293 (an intronic variant in \( SWAP70 \)) using proxies from
1000 Genomes and a \( r^2 \) cut-off of 0.8. The program found and
aligned over 1000 associations with either rs10840293 or a proxy of
rs10840293 \((r^2 \geq 0.8)\) in \( <10s \) (Fig. 1 and Supplementary Data).
Hence, even though associations between rs10840293 and phen-
types are mostly unavailable, we were able to obtain a range of
related associations using proxies (e.g. rs93138 in Fig. 1).

4 Conclusion
In summary, PhenoScanner is a large curated database of publicly avail-
able summary results from genetic association studies. This database ex-
tends current catalogues of genetic data by including all available
results as opposed to filtering on strength of association. Moreover,
PhenoScanner aligns genotype–phenotype associations across traits and
proxies, providing the user with an easily interpretable formatted out-
put file. We anticipate that this tool will make cross-referencing genetic
variants with many phenotypes faster and more efficient.

Funding
This work was supported by the UK Medical Research Council [G66840,
G0800270], Pfizer [G73632], British Heart Foundation [SP/09/002], UK
National Institute for Health Research Cambridge Biomedical Research
Centre, European Research Council [268834], and European Commission
Framework Programme 7 [HEALTH-F2-2012-279233].

Conflict of Interest: none declared.
References

1000 Genomes Project Consortium et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491, 56–65.

Arnold, M. et al. (2014) SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics, 31, 1334–1336.

Burgess, S. and Thompson, S.G. (2015) Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. Chapman & Hall, Boca Raton, FL, USA.

Frazer, K.A. et al. (2007) A second generation human haplotype map of over 3.1 million snps. Nature, 449, 851–861.

Interleukin 1 Genetics Consortium. et al. (2015) Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a mendelian randomisation analysis. Lancet Diabetes Endocrinol., 3, 243–253.

Leslie, R. et al. (2014) GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics, 30, i185–i194.

Mailman, M.D. et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet., 39, 1181–1186.

Sherry, S.T. et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res., 29, 308–311.

Welter, D. et al. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 42, D1001–D1006.