GWAS Catalog: Exploring SNP-trait associations

Laura Emery [1], Jackie MacArthur [2]

- DNA & RNA
- Literature
- Cross-domain

- Beginner
- 1 hour

The NHGRI-EBI GWAS Catalog is a publically available resource of Genome Wide Association Studies (GWAS) and their results. This course provides an introduction of how to browse SNP-trait associations using the NHGRI-EBI GWAS Catalog.

Learning objectives:

- To recall the purpose and function of the NHGRI-EBI GWAS Catalog
- To recall the type and scope of the data which it contains
- To search the Catalog to identify variants associated with particular traits
- To know where to get help and support on the GWAS Catalog

What is the GWAS Catalog?

The NHGRI-EBI GWAS Catalog is a publically available resource of Genome Wide Association Studies (GWAS) and their results (Figure 1). It was originally founded in 2008 by the National Human Genome Research Institute (NHGRI), and since 2010 has been a collaboration between the EBI and the NHGRI [1,2] [3].
Figure 1 Image showing summary statistics and karyotype diagram from the GWAS Catalog.

The Catalog contains a vast amount of data and is designed to be easily accessible to scientists who want to use the data. It is searchable and contains useful visualisations of the variant [4]-trait associations, which are mapped onto their chromosomal positions on the human genome. All traits and diseases are mapped onto an ontology [5] to improve searchability.

Find out more about controlled vocabularies and ontologies here [6].

What are genome wide association studies (GWAS)?

Genome wide association studies (GWAS) are hypothesis free methods to identify associations between genetic regions (loci) and traits (including diseases).

It has long been known that genetic variation between individuals can cause differences in phenotypes. These causal variants, and those which are tightly linked to their region of the chromosome, are therefore present at higher frequency in cases (individuals with the trait) than controls (individuals without the trait) (Figure 2).
GWAS Catalog: Exploring SNP-trait associations
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

**Figure 2** Diagram to show typical allele [7] distribution which GWAS seek to identify.

A typical GWAS study collects data to find out the common variants in a number of individuals, both with and without a common trait e.g. a disease, across the genome using genome wide SNP [8] arrays. Variants associated with the disease, or within the same haplotype [9] as a variant [4] associated with a disease, will be found at a higher frequency in cases than controls. Statistical analysis is carried out to indicate how likely a variant is to be associated with a trait.

As GWAS analyse common variants, usually typed on commercial SNP arrays (Figure 3), they do not generally identify causal variants. GWAS identify common variants which tag a region of linkage disequilibrium [10] (LD) containing causal variant(s). Additional or follow-on studies are usually required to narrow the region of association and identify the causal variant. Find out more about the theory and background of genetic variation [11].
Figure 3 Diagram to show the identification of alternative variants in cases and controls using an array-based typing method. Results are subject to statistical analyses to assign a p-value [12] to each variant.

A p-value indicates the significance of the difference in frequency of the allele tested between cases and controls i.e. the probability that the allele is likely to be associated with the trait. GWAS results are often displayed in a Manhattan plot (see Figure 3 above) with -log10(p-value) plotted against the position in the genome.

You can learn about the theory behind GWAS in more detail in a resource by Gill McVean available here [13], and in a book chapter written by Bush and Moore here [14].

Why do we need the GWAS Catalog?
Genome Wide Association Studies (GWAS) are often the approach of first choice for investigating links between genotype [15] and phenotype [16]. This is because they are an effective method of identifying associations between chromosomal regions (loci) and traits (e.g., a disease). GWAS have been particularly useful in identifying loci associated with complex diseases where there are multiple loci that can cause that disease (i.e., the disease is multifactorial), such as obesity. They are also able to detect relatively small genetic effects. With this wealth of useful genotype-phenotype mapping data available from numerous studies in the literature, it is important to have a central repository where all of this data is stored, and accessible to the scientific community (Figure 4).

Where does the data come from?

Genotype [15]-phenotype [16] associations along with study information, are manually extracted from the literature and entered into the GWAS Catalog. This information is curated by experts, and then made freely available and searchable via our website to allow scientists to interpret the data accurately (Figure 5).
Eligibility Criteria

Studies are included in the GWAS Catalog if they carry out genome wide array-based genotyping and association analysis of ≥100,000 SNPs. This includes previously published GWAS which are incorporated into new analyses (meta-analyses). However it excludes targeted array studies. Further details on Catalog eligibility can be found here [17]. Note that although in most cases the variants typed by SNP [8]-arrays are single nucleotide variants [18], indels may also be typed.

For each of these studies, variant [4]-trait associations are included if they:

- Have a p-value [12] <1.0 x 10^{-5} in the overall (initial GWAS + replication) population
- The most significant variant from each independent locus is extracted

Which data from these studies is included into the GWAS Catalog?

For each study that we curate, we extract data to describe the study design and allow accurate interpretation of the GWAS, along with eligible results, as described above. The diagram below (Figure 6) summarises the data that we extract and enter into the Catalog:
Figure 6 Summary of data that is extracted and entered into the GWAS Catalog.

Each of these datatypes is discussed in more detail in the guided example.

How to access and navigate the GWAS Catalog

There are several ways you can access the GWAS Catalog:

1. Via the search interface on our [website](#). Search results can be downloaded in a tab delimited file.
2. By downloading the entire catalog as a [tab delimited file](#) [20]
3. Via the interactive [karyotype diagram](#) [21] (Figure 7)
4. Via programmatic access **coming soon**
How to search the GWAS Catalog: a guided example

To search the GWAS Catalog, simply go to our website [22] (Figure 8):

Figure 8 The search bar on our website.

In the search bar, type a keyword of interest. That could be:

- a trait or its synonym [23] (traits are ontology [5] indexed)
- a genomic location/region
- a variant [4] e.g. rsID
- a gene
- sample description
- publication details e.g. title, (first) author, journal etc

Once you have pressed the search button, you will be directed to a results page, which displays summary information divided into main three sections (known as facets):

1. Studies - displays a summary of the GWAS publications available relating to your query
2. Associations - displays a summary of the variant-trait associations relating to your query
3. Catalog traits - displays a summary of the traits and their ontology terms relating to your query

Note that there are a number of ways you can further filter the results by using the tab at the left of the page. For now, we will simply look at all of the results.

Let's take a look at an example. We are interested in diabetes and finding out which variants have been associated with this disease. First, let's visit our website [22] and search for diabetes.

**Video 1** The video shows you a demonstration of how to search for the term diabetes, and an overview of the results you will obtain.

More detail and an explanation of your results follows on the next pages.

### The studies facet

First, let's explore the **Studies facet** in your results for *diabetes*. You will see entries at the top of your results page that look like this (Figure 9):

![Studies facet](image)

**Figure 9** Screenshot of the first portion of search results - the Studies facet - when searching for diabetes.

This shows a list of all of the publications in the GWAS Catalog that were returned for the search diabetes. This will include publications:
reporting associations with diabetes
reporting associations with synonyms or child terms of diabetes in the ontology [5]
containing “diabetes” in the title, journal or any other indexed field.

Note that the table also includes the precise trait description that was associated with a variant [4], which is often more specific than our initial search criteria, and only some of these may be relevant to our research interests. We can also see a count of how many associations were identified in each paper.

The table is interactive - try clicking on the different headers to sort by them. Also try clicking the “+” button to expand any study and find out more information about the samples used, genotyping method and number of SNPs analysed.

The associations facet

Next, scroll down the results page to arrive at the Associations facet (Figure 10):

![Associations](image)

**Figure 10** Screenshot of the second portion of search results - the Associations facet - when searching for diabetes.

This results section best answers our original question - which variants are associated with diabetes? Here you will find a list of variants associated with some aspect of diabetes (with a $p<1 \times 10^{-5}$ significance level), and relevant information about their associations is reported, including:

- **SNP** [8] - The variant [4] most strongly associated with the trait, reported as the rsID, along with the risk/effect allele [7]. The association may also be with a haplotype [9] or SNPxSNP interaction. The button links out to the Ensembl [24] page with further details about the variant.
- **RAF** - Risk allele frequency - the proportion of the sampled control population carrying the trait-associated variant
- **p-value** [12] - The probability that the variant is associated with the trait that is listed
- **OR** - Odds ratio - the number of times someone is more likely to exhibit the disease/trait if they possess the risk/effect allele
- **Beta** - Regression coefficient to indicate the effect size on the trait of possessing the risk/effect allele
- **CI** - The 95% confidence interval for the odds ratio or beta (depending on which is reported)
GWAS Catalog: Exploring SNP-trait associations
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

- **Region** - The cytogenetic band (find out more [here](#)[25])
- **Location** - The precise genomic chromosome and base pair [26] location
- **Functional Class** - Functional class of variant i.e. whether it falls in an intron [27]/exon [28] etc
- **Reported genes** - Genes reported by the author as being co-located with the variant
- **Mapped genes** - The gene that the variant is mapped within, or the closest upstream and downstream genes, according to Ensembl
- **Reported trait** - A description of the trait examined in the study based on language reported in the publication
- **Study** - Details of the study - first author, PubMed [29] ID and year of publication, with a link out to PubMed Central

As with the studies results, the headers are sortable. By selecting the icons, you can link out to further relevant pages in Ensembl. Selecting any of the blue hyperlinked text will redirect you to a new search using the blue text as a search term.

### The traits facet

Finally, if you continue to scroll down the page you will arrive at the **Traits facet** (Figure 11):

![Catalog traits](#)

<table>
<thead>
<tr>
<th>Catalog traits</th>
<th>Mapped ontology traits</th>
<th>Ontology traits synonyms</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adiponectin levels</td>
<td>adiponectin measurement</td>
<td>adiponectin levels</td>
<td></td>
</tr>
</tbody>
</table>
| Advanced glycation end-product levels | advanced glycation end-product measurement | T2DM - Type 2 Diabetes mellitus, Diabetes Mellitus, Adult-Onset, Adult-Onset Diabetes Mellitus, Diabetes Mellitus, Slow Onset, Type 2 Diabetes, [...]
| Age-related disease endophenotypes | total cholesterol measurement, diastolic blood pressure, systolic blood pressure, hematocrit, ventricular rate measurement, glucose measurement, body mass index, high density lipoprotein cholesterol measurement | HDL measurement, SYSBP, systolic pressure, glucose level, DIASP, [...]
| Age-related diseases and mortality | cancer, heart failure, diabetes mellitus, stroke, coronary heart disease, atrial fibrillation, mortality | CVA (cerebral vascular accident), Stroke, Cerebral STROKE SYNDROME, Coronary Heart Diseases, SYNDROME, STROKE, [...]

**Figure 11** Screenshot of the final portion of search results - the Traits facet - when searching for diabetes.

The traits facet reports all of the reported traits that are associated with our search term - **diabetes**. This is useful if you are interested in finding out related traits in the GWAS Catalog, along with ontologically related traits. The primary result that is returned is the ‘reported trait’ - this is
designed to capture the full detail of the trait studied in the particular GWAS. The reported traits are mapped to Experimental Factor Ontology [30] (EFO [31]) traits (find out more here [32]), to enhance search functionality - so for example, when you search for diabetes you also return type 2 diabetes and other related traits. The returned results therefore contain:

- mapped ontology traits - ontology traits best representing the reported trait studied in the GWAS
- ontology trait synonyms - alternative ways of describing the mapped ontology traits
- study - GWAS Catalog publications studying the the reported trait

It is important to note that the search is initially broad. A search for a specific trait will return matches for this search term anywhere within the GWAS Catalog, including where it is mentioned in the title of a publication or a journal. This can be problematic - for example in the case of diabetes, the search will return all GWAS published in the Journal “Diabetes”, even if they do not study the trait “diabetes”. This is where the filter options can be useful.

Filtering your results

As we saw in the previous section, a search of the GWAS catalog for diabetes is broad and returns matches for this search term anywhere within the GWAS Catalog. To narrow down the results you can use the filters available on the side of the page. Let's take a look at the filter bar (Figure 12):
Figure 12 The filter bar.

The bar allows you to refine your results by the following:

- **p-value [12]** - The default is \( \leq 1 \times 10^{-5} \). Here the p-value can be reduced to make your results
more conservative, for example $\leq 1 \times 10^{-6}$ or $\leq 5 \times 10^{-8}$ depending on your interests.

- **Odds ratio** - You may only be interested in variants that greatly increase your chance of exhibiting the trait - in this case you may wish to set a high odds ratio.
- **Beta coefficient** - Similarly, you may only be interested in variants that have a large effect on the trait by comparison with alternative alleles [7] - in this case you may wish to set a high beta coefficient.
- **Study date** - You can select specific years for study publication dates, for example if you are only interested in recent studies.
- **Genomic range** - You can set a specific genomic region of interest.
- **Catalog Trait** - You can use this box to restrict the search to more stringent reported traits - for example type II diabetes. Multiple traits can be selected using this filtering [33] function.

### Additional data in the GWAS Catalog

The Catalog also contains additional data (Figure 13):

![Summary statistics and Ancestry](image)

**Figure 13** Additional data available.

#### 1. Summary statistics

We store summary statistics for GWAS Catalog studies if they are available through the publication (2017 studies onwards) or have been submitted to the Catalog by authors. Summary statistics files can be accessed via our summary statistics webpage [34] and also from the search results table for any study with available files.

#### 2. Ancestry data

For every study that is curated in the GWAS Catalog, we identify the ancestry of the populations analysed and make this information available through the studies facet and in a separate ancestry download. The ancestry is both reported as a detailed description as reported by the author, and also as one of 17 ancestry categories. This area is currently under development and we hope to provide additional functionality to filter by ancestry in the future, along with analyses of Catalog data by ancestry. Find out more on our ancestry webpage [35], and download our ancestry data here [36].

### How to get data from the GWAS Catalog

Once you have completed a search you can obtain the results in various formats:

- A visual display on the webpage, as we have just explored
- A tab delimited TSV [37] file that can be imported into Excel by selecting the **[Download association results](#)** on the search results page (Figure 14)
- A large TSV file of the entire Catalog via our [downloads] [20] page
You can also explore our GWAS Catalog karyotype diagram here [21], which can be filtered so you can explore a visualisation of your results across the human chromosomes.

**Figure 14** Screenshot of search results when searching for the trait diabetes. The red circle highlights where you can "Download association results".

### Summary

**Figure 15** The GWAS Catalog.

- The NHGRI [38]-EBI GWAS Catalog (Figure 15) is a publically available resource of Genome Wide Association Studies (GWAS) and their results.
- It allows you to identify variants associated with particular traits, search genomic regions of interest for associations with traits.
- You can access the Catalog by searching our website, or you can download the entire Catalog as a TSV [37] file.

### Try it for yourself

The exercises allow you to apply your knowledge gained on this course by providing examples of how you might search the GWAS [39] Catalog.
Exercise 1

a. Which region of the genome is most significantly associated with dementia?

b. Is the variant [4] most significantly associated with dementia, dementia-specific?

Exercise 2

a. How many variants from studies published in 2016 were associated with hair colour?

b. What is the ancestry of the populations used to study hair colour?

Want to know how we did it?

Exercise 1a

1. Search for dementia in the main search bar
2. Look at the results in the Associations facet
3. Sort by p-value [12]
4. You will see that a variant [4] within the gene TOMM40 has the lowest (most significant) p-value

Exercise 1b

1. From the search results in the previous question, click on the SNP [8] rsID: rs2075650-G to start a new search based on this variant
2. Scroll down to the Associations facet and you will see this specific variant is also associated with Cholesterol total and Cardiovascular disease risk factors, along with many other traits.

Answer: No

Exercise 2a

1. Search for hair colour in the main bar
2. View the results in the Studies facet - there are two studies, both published in 2016
3. The total association count (11 + 8) is 19

Answer: 19

Exercise 2b
1. From your previous search results, select the + button next to the studies listed

Answer: Hispanic or Latin American

Quiz: Test your knowledge

Questions: 5
Attempts allowed: Unlimited
Available: Always
Pass rate: 75 %
Backwards navigation: Allowed

Get help and support on the GWAS Catalog

- You can find additional documentation about the GWAS Catalog on our website [40] (Figure 16)
- Watch our webinar The NHGRI-EBI GWAS Catalog, a curated resource of SNP-trait associations [41]
- For any enquiries, please contact the GWAS Catalog gwas-info [at] ebi.ac.uk (subject: Helpdesk k%20enquiry%20from%20train%20online%20GWAS%20Catalog%20course) (helpdesk)
- You can stay informed about GWAS Catalog developments by subscribing to our announcement list (click on the sign-up link on our About [42] page) or following us on Twitter @GWASCatalog [43]

Figure 16 The help and support sections on our website.

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and enhance your learning experience.

References

Contributors

Laura Emery [1]
EMBL-EBI
Scientific Training Officer - Training

Laura completed her undergraduate degree in Genetics (BSc first class hons) at the University of Nottingham. After dabbling in experimental work on a summer placement at the Wellcome Trust Sanger Institute, she found joy in using computer-based analyses to explore biological questions. Laura completed her PhD in Evolutionary Genetics at the University of Edinburgh’s Institute of Evolutionary Biology. There she used a bioinformatics approach to investigate patterns of codon usage in Archaea while under the supervision of Prof Paul Sharp FRS.

She also worked as a postdoc at the University of Manchester, where she explored molecular co-evolution among interacting proteins. Having obtained qualified teacher status (QTS), Laura joined EMBL-EBI as a Scientific Training Officer in 2012 and is responsible for the development and delivery of training courses.

Jackie MacArthur [2]
EMBL-EBI
Project Leader (GWAS Catalog) - Cunningham team: Variation Annotation
Laura Harris [46]

EMBL-EBI
Scientific Curator - Cunningham team: Variation Annotation

Source URL: https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations

Links
[1] https://www.ebi.ac.uk/training/online/trainers/laura.emery
[2] https://www.ebi.ac.uk/training/online/trainers/jalm_12464
[3] https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/references
[4] https://www.ebi.ac.uk/training/online/glossary/variant
[5] https://www.ebi.ac.uk/training/online/glossary/ontology
[6] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified/controlled-vocabularies
[7] https://www.ebi.ac.uk/training/online/glossary/allele
[8] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-polymorphism
[9] https://www.ebi.ac.uk/training/online/glossary/haplotype
[10] https://www.ebi.ac.uk/training/online/glossary/linkage-disequilibrium
[12] https://www.ebi.ac.uk/training/online/glossary/p-value
[15] https://www.ebi.ac.uk/training/online/glossary/genotype
[16] https://www.ebi.ac.uk/training/online/glossary/phenotype
[17] http://www.ebi.ac.uk/gwas/docs/methods
[18] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-variants
[19] https://www.ebi.ac.uk/gwas/
[20] https://www.ebi.ac.uk/gwas/docs/file-downloads
[21] https://www.ebi.ac.uk/gwas/diagram
[22] https://www.ebi.ac.uk/gwas/home
[23] https://www.ebi.ac.uk/training/online/glossary/synonym
[24] https://www.ebi.ac.uk/training/online/glossary/ensembl
[26] https://www.ebi.ac.uk/training/online/glossary/base-pair
[27] https://www.ebi.ac.uk/training/online/glossary/intron
[28] https://www.ebi.ac.uk/training/online/glossary/exon
[29] https://www.ebi.ac.uk/training/online/glossary/pubmed
[30] https://www.ebi.ac.uk/training/online/glossary/experimental-factor-ontology