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) [4].
GWAS Catalog

As of May 2019
- 3,989 publications
- 138,312 variant-trait associations
- >6,000 full summary statistics files

Figure 1 Image showing summary data 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 [5]-trait associations, which are mapped onto their chromosomal positions on the human genome. All traits and diseases are mapped onto an ontology [6] to improve searchability.

Find out more about controlled vocabularies and ontologies here [7].

What are genome wide association studies (GWAS)?

Genome wide association studies (GWAS) are hypothesis-free methods for identifying 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 in controls (individuals without the trait) (Figure 2).
Figure 2 Diagram to show typical allele [8] 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 [9] arrays. Variants associated with the disease, or within the same haplotype [10] as a variant [5] associated with a disease, will be found at a higher frequency in cases than in 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 [11] (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 here [12].
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.

The GWAS Catalog is a structured repository which provides summary data from all published human GWAS studies, in a consistent, searchable format.

Click on 'next' below to find out more about how the GWAS catalog can be used to access GWAS data.

You can learn about the theory behind GWAS in more detail in a resource by Gill McVean available here [14], and in a book chapter written by Bush and Moore here [15].
Why do we need the GWAS Catalog?

![GWAS data](image)

**Figure 4** A snapshot of GWAS data from a number of studies. It is important to have a central repository where all of this data can be stored.

Genome Wide Association Studies (GWAS) are often the first-choice approach for investigating links between genotype [16] and phenotype [17]. 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 [16]-phenotype [17] 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).
Figure 5 Flow diagram to show how studies are included in the GWAS Catalog.

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). Further details on Catalog eligibility can be found here [18]. Note that although in most cases the variants typed by SNP [9]-arrays are single nucleotide variants [19], indels may also be typed.

For each of these studies, variant [5]-trait associations are included if they:

- Have a p-value [13] <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 in 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 [20]. Search results can be viewed online or downloaded in a tab delimited file
2. By downloading the entire catalog as a tab delimited file [21]
3. Via the interactive karyotype diagram [22] (Figure 7)
4. Programmatic access via our RESTful API [23]
Figure 7 The interactive karyotype diagram, which displays all variant [5]-trait associations with p-value [13] $\leq 5.0 \times 10^{-8}$ published in the GWAS Catalog. The diagram can be searched and filtered.

How to search the GWAS Catalog: some guided examples

To search the GWAS Catalog, simply go to our website [24] (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 [25] (traits are ontology [6] indexed)
- a variant [5] e.g. rsID
GWAS Catalog: Exploring SNP-trait associations
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

- a gene
- a genomic location/region
- publication details i.e. title, author or PubMed [26] ID

Once you have pressed the search button, results and data snippets will be displayed below.

Publications P, Traits T, Variants V, Genes G and Genomic Regions R are indicated separately in the search results.

Try searching for your favourite trait or gene in the frame below, and scroll down to see your search results. You can filter the results according to data type, by using the "refine search results" box.

View the full site > [27]

Clicking on your desired search result takes you to the Catalog data in its own dedicated page, so you can explore it in more detail. An explanation of each data type follows on the next pages.

The Publication page

*/

To access the Catalog entry for a specific publication, the best search term to use is the PubMed [26] ID.

Click through from the search result to view the dedicated page for that Publication. You can try it in the frame below. We’ve already entered a search term - scroll down to see the result and click on the title to go through to the Publication page.

View the full site > [28]

Explore the Publication page and and you can see it is divided into three tables. You will see the same pattern of content for each dedicated page we’re going to look at over the next few pages.

- The Information table shows citation information, links out to external data sources, and whether the publication has summary statistics available
- The Associations table shows a list of all associations included in the publication, including p-value [13], effect size and gene information
- The Studies table shows a list of all studies included in the publication, including trait information, sample description and association count

Each table is interactive (Figure 9). You can sort each column by clicking on the up/down arrows next to the column headers, and customise the view by including or excluding columns using the tick box icon on the right hand side. Search the data in each column using the search box at the top. Download the data from each table using the button on the right.

You can test all these features in the frame above. For example, try sorting the p-value column to find the most significant variant [5] in this Publication. Try searching the gene name column for the gene GCKR, to find all the variants discovered in that gene in this Publication.
The Study page

*/*

Where a publication contains more than one GWAS, each separate GWAS is included as a separate Study. Using the example from the previous page, you can see this publication contains 6 studies of different blood pressure traits. To find detailed information on each study, you can click through from the Study table to each individual Study entry.

Try it in the frame below - click on the "Studies" button or scroll down to the Studies table. Then click on a study accession [29] (e.g. GCST007095) to access the Study page.

On the Study page, in the Information Table, you will find details of the samples used, genotyping method and number of SNPs analysed, and the trait under study. You can find more information on how we describe the samples in our recent publication [4]. From this table you can also access full summary statistics for the study (where available).

Scroll down to find the list of Associations from this study. Each Associations table contains the following information:

- **SNP** - The [variant] [5] most strongly associated with the trait, reported as the rsID, along with the risk/effect [allele] [8]. The association may also be with a [haplotype] [10] or SNPxSNP interaction. The button links out to the [Ensembl] [31] 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** [13] - 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)
- **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 accession - Identifier [32] of the Study the association was discovered in

This table is linked to the variant, gene and trait pages, which we will explore in more detail in the next section.

The Trait page

*/

Now you are familiar with the data structure in the GWAS Catalog, we are going to look at one of the most popular GWAS Catalog queries: “Which variants have been associated with my trait?”

To view all the GWAS Catalog data for a particular trait, simply search for the trait in the main search bar. A drop down menu will auto-suggest terms (and synonyms) from the Catalog. The terms used in the catalog are from the Experimental Factor Ontology [33] (EFO) (find out more here [34]), to make it easier to search and compare between studies - so for example, when you search for "diabetes" you also return "type 1 diabetes","type 2 diabetes" and other related traits.

In the frame below, try searching for "high blood pressure". As you start to type, you will see the autosuggestion "blood pressure, high - synonym [25] for hypertension". Click on the autosuggestion to search, and then click through from the search result to view the dedicated page for that trait.

Navigating the Trait page

At the top of the Trait page, the Information Table contains the definition of the trait, links out to find more information on the topic, and lists of synonyms and mapped terms in other ontologies. Click on the “Highlighted Study” box to see the study with largest sample size to date.

Trait representation in the GWAS Catalog

Two important items in the information table are Reported Traits and Child traits.

As well as mapping each GWAS to one or more terms from EFO, the GWAS Catalog also includes a more detailed trait description. This is designed to capture the full detail of the trait studied (e.g. if all samples have a particular clinical background, such as smoking) and experimental design of the particular GWAS (e.g. interaction studies). You can view all the Reported Traits associated with the EFO term by clicking the "+" button next to the number of reported traits.

The Child traits list includes any traits that are more specific child terms of the parent trait in the EFO hierarchy. For example, asthma has 4 child terms, glucose measurement has 8 (Figure 10), and in the example above, hypertension has more than 60.
Figure 10 Two examples of traits with child terms, showing their position in the Experimental Factor Ontology hierarchy.

You may be interested in the trait including all its child terms, or you may wish to exclude those more specific terms and only see associations with the main trait. Below the associations table is a check-box “include child trait data” which will include or exclude child terms as required. You can try this in the frame above.

The Trait page, part 2

Further down the Trait page you can see the Associations table, which summarises all the associations mapped to this trait in the GWAS [36] Catalog, and below that, the Studies table which lists all the studies mapped to this trait, together with their sample details.

This is where the interactive features of the tables can be useful. In the frame above, have a go at using the sort function in the p-value [13] column of the Associations table to find the most significant association with hypertension. Then, in the Studies table, try typing "African" in the search box of the the Discovery sample column to find any studies that include subjects of African ancestry in the Discovery stage.

At the bottom of the Trait page you will find a LocusZoom plot of all the associations mapped to this trait in the GWAS Catalog. Hover over a spot to find the details of that particular association.

The Variant page

Another common query in the GWAS Catalog is “what other associations have been discovered for my variant [5] of interest?”. You can search for a specific variant (typically using an rsID), gene name, or genomic region (eg 2q37.1; 6:16000000-25000000).

Searching for a variant allows you to click through to the Variant page, showing a summary of all the data in the Catalog pertaining to your variant of interest. You can see an example in the frame below.
The Variant page has the same format as the other page types, with general variant information and links out in the Information Table at the top, followed by the Associations Table below, which summarises every association with this variant in the Catalog. Below, you can find the Studies Table, which lists every study that has reported an association with the variant.

There are two unique features on the Variant page. Firstly, the Trait table contains a list of all the traits reported for the variant in order of frequency. This allows you to easily see the most commonly reported ones. In the example above, you can see the most commonly associated trait for this variant is platelet count, followed by other haematology traits.

Secondly, scrolling down to the bottom of the page, you can find a useful tool for analysis, the Linkage disequilibrium (LD) plot. You can use this to find other variants in LD with your variant of interest, within the GWAS Catalog and in Ensembl [31]. Choose the reference population and filter by LD threshold using the options at the top of the box. Try it in the frame above. You can download all the variants in the plot using the “Download LD data” button.

Lastly, searching for a gene or genomic region in the main search page summarises all the variants annotated to that gene or genomic region on a single page. From there you can download the data or click through to each variant for more information.

**Additional data in the GWAS Catalog**

The Catalog also contains additional data. Only the most significant association ($<1 \times 10^{-5}$) from each locus is eligible for inclusion in the curated Catalog; however, we also provide full summary statistics files (association statistics for all variants analysed in a GWAS) for any study with available files.

**Figure 11** Additional data available on the summary statistics page.

We store summary statistics for GWAS Catalog studies if they are available through the publication or have been submitted to the Catalog by authors. Summary statistics files can be accessed via our summary statistics webpage [39] (Figure 11) and also from the search results or Studies table, for any publication with available files (Figure 12).
Figure 12 Accessing summary statistics files from the search results

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 .csv file of each individual results table that can be downloaded by selecting the export icon (Figure 13), and then imported into Excel
- A large .tsv file of the entire Catalog via our downloads page
- You can also access our GWAS Catalog karyotype diagram here, which allows you to explore a visualisation of the data across the human chromosomes
- You can download summary statistics files from individual studies or the summary statistics page

Figure 13 Screenshot of the Association table for the trait body mass index. The red circle highlights where you can export association results.
Summary

The NHGRI-EBI GWAS Catalog (Figure 14) is a publicly available resource of Genome Wide Association Studies (GWAS) and their results. It allows you to identify variants associated with particular traits and 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 file. Summary statistics files (full p-value sets for the entire GWAS) are also available via the GWAS Catalog.

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 Catalog.

Exercise 1

a. Which region of the genome is most significantly associated with Alzheimer's disease?

b. Is the variant most significantly associated with Alzheimer's disease, specific to Alzheimer's disease?

Exercise 2

a. How many variants from studies published in 2018 were associated with hypothyroidism?

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

If you need help to complete this section you can look in the 'Want to know how we did it?' section.
Want to know how we did it?

Exercise 1a

1. Search for Alzheimer's disease in the main search bar
2. Click through to the Trait page for Alzheimer's disease
3. Scroll down to the Association table and sort the associations by **p-value** [13]
4. You will see that a variant [5] within the gene TOMM40 on chromosome 19 has the lowest (most significant) **p-value**

Exercise 1b

1. From the search results in the previous question, click on the SNP [9] rsID: rs2075650-G to go to the Variant page for this SNP
2. Scroll down to the Traits table and you will see this specific variant is also associated with body mass index and age-related macular degeneration, along with several other traits

**Answer:** No

Exercise 2a

1. Search for hypothyroidism in the main bar
2. Scroll down to the Studies table and sort by Publication date. There are two studies published in 2018
3. The total association count (153 + 8) is 161

**Answer:** 161

Exercise 2b

1. Look at the "sample number and ancestry" columns in the Studies table

**Answer:** European

Quiz: Test your knowledge

You need be logged in to access the quiz.

**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 [43] (Figure 15)
- For any enquiries, please contact the GWAS Catalog gwas-info [at] ebi.ac.uk (subject: Helpdesk%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 [44] page) or following us on Twitter @GWASCatalog [45]

![Figure 15](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

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

EMBL-EBI
Project Leader (GWAS Catalog) - Cunningham team: Variation Annotation
GWAS Catalog: Exploring SNP-trait associations
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Laura Harris [3]
EMBL-EBI
Scientific Curator - Cunningham team: Variation Annotation

Source URL: https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations-2019

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/trainers/ljwh_13260
[5] https://www.ebi.ac.uk/training/online/glossary/variant
[6] https://www.ebi.ac.uk/training/online/glossary/ontology
[7] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/controlled-vocabularies
[8] https://www.ebi.ac.uk/training/online/glossary/allele
[9] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-polymorphism
[10] https://www.ebi.ac.uk/training/online/glossary/haplotype
[11] https://www.ebi.ac.uk/training/online/glossary/linkage-disequilibrium
[12] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-introduction/what-tool-name
[13] https://www.ebi.ac.uk/training/online/glossary/p-value
[16] https://www.ebi.ac.uk/training/online/glossary/genotype
[17] https://www.ebi.ac.uk/training/online/glossary/phenotype
[18] http://www.ebi.ac.uk/gwas/docs/methods
[19] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-variants
[20] https://www.ebi.ac.uk/gwas/
[21] https://www.ebi.ac.uk/gwas/docs/file-downloads
[22] https://www.ebi.ac.uk/gwas/diagram
[23] https://www.ebi.ac.uk/gwas/docs/api
[24] https://www.ebi.ac.uk/gwas/home
[25] https://www.ebi.ac.uk/training/online/glossary/synonym
[26] https://www.ebi.ac.uk/training/online/glossary/pubmed
[27] https://www.ebi.ac.uk/gwas/search
[28] https://www.ebi.ac.uk/gwas/search?query=27841878
[29] https://www.ebi.ac.uk/training/online/glossary/accession
[30] https://www.ebi.ac.uk/gwas/publications/27841878
[31] https://www.ebi.ac.uk/training/online/glossary/ensembl
[32] https://www.ebi.ac.uk/training/online/glossary/identifier
[33] https://www.ebi.ac.uk/training/online/glossary/experimental-factor-ontology
[34] http://drjamesmalone.blogspot.com/2012/06/common-ontology-questions-1-what-is-it.html
[35] https://www.ebi.ac.uk/gwas/search?query=*gwas
[36] https://www.ebi.ac.uk/gwas/docs/about
[37] https://www.ebi.ac.uk/gwas/efotraits/EFO_0000537
[38] https://www.ebi.ac.uk/gwas/variants/rs9399137
[39] https://www.ebi.ac.uk/gwas/downloads/summary-statistics
[40] https://www.ebi.ac.uk/training/online/glossary/tsv
[41] https://www.ebi.ac.uk/training/online/glossary/nhgril
[42] https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations-2019/try-it-yourself/want-know-how-we-did
[43] https://www.ebi.ac.uk/gwas/docs/
[45] https://twitter.com/GWASCatalog