Discover and download HipSci's data

The Human Induced Pluripotent Stem Cells Initiative (HipSci) has generated substantial QC and characterization data on its panel of human IPSC lines.

This document and the accompanying video explain how scientists can discover data using the HipSci website, and then download the data from accessible data archives.

Discover

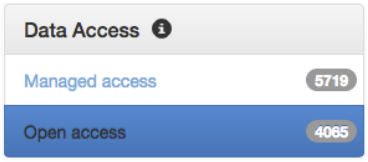
In the cell line and data browser, the 'files' tab shows a complete list of all available data. You can click on the filters to limit the data listed in the table, e.g. to show just data for your preferred assay type, or cell line growing conditions.

The "archive" column in the displayed table tells you where the data is archived. Data are available from different specialist archives, depending on the data type and assay type. For example, all managed access[1] data is at EGA, a specialist archive for secure sharing of genetic data. Open access whole exome sequencing is archived in the ENA; gene expresssion data is archived in ArrayExpress; and proteomics data is archived in PRIDE.

If you click on the archive name in the HipSci files table, it will take you to the archive's website page specific for that piece of data.

Download open-access data

HipSci's open access data sets may be freely downloaded by anybody, so there is no requirement to register for access. You can discover open access data by clicking on the files tab and selecting the "Open access" filter.



There are many different tools available to download data files by FTP - the computer you use probably has a download tool already installed. Some of HipSci's data files are very large - possibly several gigabytes. If you have experience, then we recommend downloading large files from a command line tool, or with a FTP client such as Filezilla (https://filezilla-project.org/).

[1] "Managed access" refers to genetic data from HipSci donors whose ethnical consent authorises release of such data only for specific research use to bona fide researchers. "Open access" refers either to non-genetic data, or to genetic data from donors with a more permissive ethnical consent agreement.

You can also try downloading the files using your favourite web browser by first clicking on the file download link then:

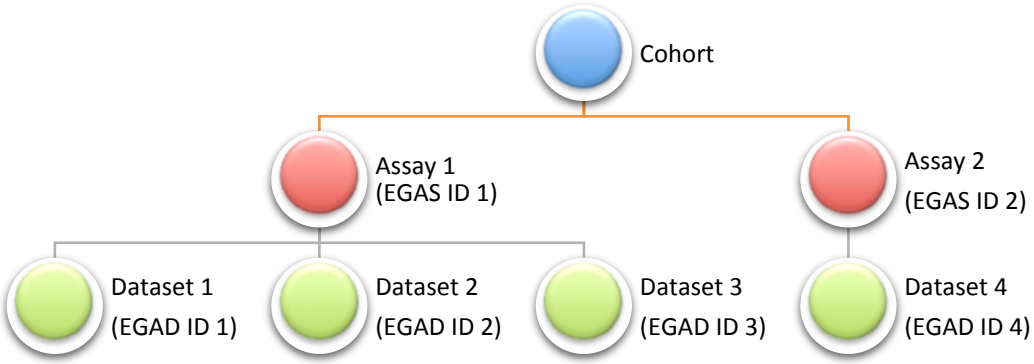
- In Firefox, right-click on the file download link, and select "save link as" from the menu.
- In Safari, right-click on the link and select "Download linked file"
- In Chrome, right-click on the link and select "Save link as"

The download client you use might ask you for a user name and password. Select to connect as a guest, if you see that option. Alternatively, enter "anonymous" for both the user name and the password. If you are having any trouble downloading a file, then email us at hipsci@ebi.ac.uk and we will help you with the download.

Download managed-access data

HipSci's managed access data is stored at the EGA. You must have completed a data access agreement via the Electronic Data Access Management system (eDAM) and been sent aproval for access to the data you applied for. If you have not done this step, please refer to our video or document titled "Applying for managed access data".

Hipsci data is setup and structured in the following way within EGA. Understanding the structure will help you to hone in on the data you seek.



Data access is provided at the cohort level. A cohort comprises all donors assigned to that cohort, e.g. 'Monogenic diabetes' or 'Healthy normals'. A successful data access application will grant you access to all data for your chosen cohort.

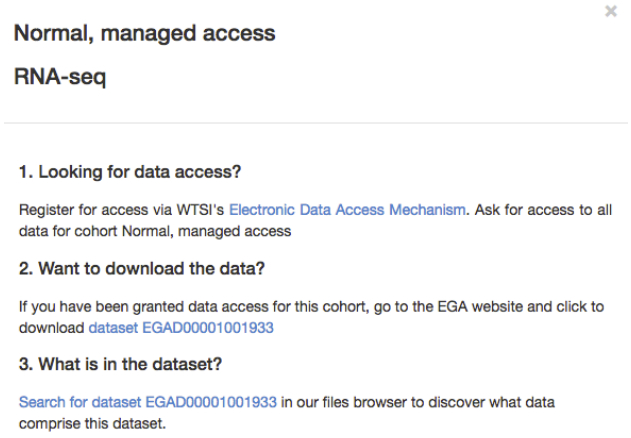
For each cohort, we distribute data from multiple assays e.g. RNA seq, WES, WGS. Each different assay within the cohort is given a different study ID. The study IDs start with the letters EGAS; for example EGAS00001000866.

Within each study ID, there are multiple datasets corresponding to different release dates. Each dataset has an ID that starts with the letters EGAD; for example EGAD00010001147. The data sets are informatively labelled to help you identify the cohort, the assay and release date. For example, "HipSci - Healthy Normals - Genotyping Array - September 2016".

Each dataset contains files of multiple types. For example, a exome-seq dataset will contain raw sequencing, aligned sequence, and variant call files.

When new data is released for an assay, we create a new dataset to include all of the latest data files, including both the new and old data files. The new dataset therefore supercedes the old dataset, and we recommend researchers always download the newest dataset for a particular assay and disease cohort. This is why we include the release date in the dataset ID. As access is given at the cohort level, you will have access to the data in any new releases too.

The HipSci website provides an EGA datasets table, which shows the most recent dataset IDs for each assay under each cohort. Click on a dataset ID and you will be offered three options:



Option #1: Looking for data access if for anyone who has not yet applied for access to the data via Electronic Data Access Management system (eDAM)

Option #2, "Want to download the data?", if your application for data has been successful, and you are now ready to download the data. The link will take you to the dataset on the EGA website. Log in to the website using your institutional email address and password.

The EGA dataset website page has a "Downloads" section. If you have been granted access to the cohort, and if you are logged in correctly, then you will see two options:
"Download metadata" - this option will give you information about the cell lines and files; this same information is also searchable in the HipSci cell lines browser.
"Download data files" - this is the button to click to start downloading the QC and characterization data files to your computer. We recommend users read the EGA's download quick guide to learn about the EGA download client.

Option #3, "What is in the dataset?" is helpful for users who have not yet applied for data access, but who want to know exactly which data files belong in a dataset.

Where to get help

If you are struggling to find the information you need regarding our cell lines and data, or have another general enquiry, you can contact us at hipsci@ebi.ac.uk

Acknowledgements

Thanks to Ian Streeter, Laura Clarke and Reena Halai for producing this document.