HLA|COVID-19 Database Access Manual

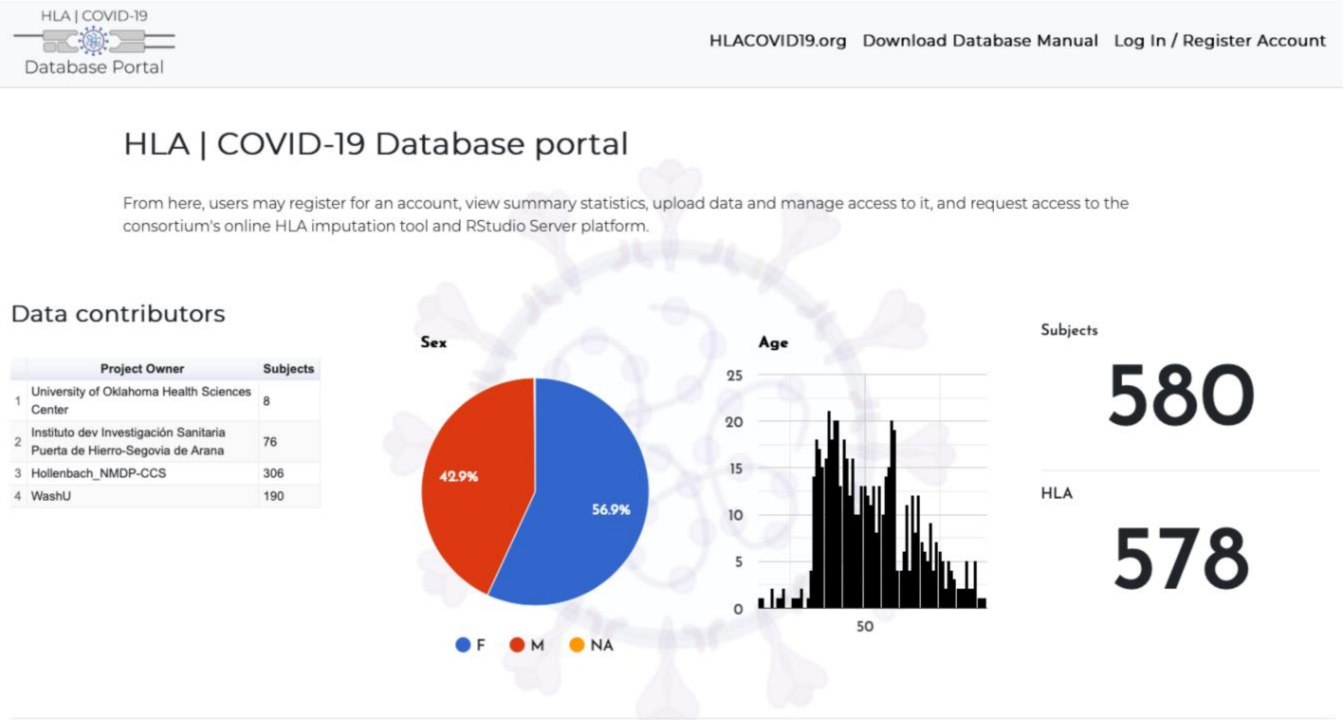
V2.0 March 30, 2021

Contents

1. Location

The HLA|COVID-19 Database (HCDB) is online at https://database-hlacovid19.org, and can be accessed via any modern browser.

The database landing page is shown below (Figure 1). This page displays statistics regarding currently loaded subject data in the main part of the page, and includes links to the HLA|COVID-19 web page and other database-related links at the top. Frequency histograms of the alleles in the database are presented below the subject data (not shown in Figure 1).



HLA | COVID-19
Database Portal

HLACOVID19.org Download Database Manual Log In / Register Account

HLA | COVID-19 Database portal

From here, users may register for an account, view summary statistics, upload data and manage access to it, and request access to the consortium's online HLA imputation tool and RStudio Server platform.

Data contributors

	Project Owner	Subjects
1	University of Oklahoma Health Sciences Center	8
2	Instituto dev Investigación Sanitaria Puerta de Hierro-Segovia de Arana	76
3	Hollenbach_NMDP-CCS	306
4	WashU	190

Subjects

580

HLA

578

Figure 1. Database Landing Page

2. Access

All datasets deposited in the HCDB are embargoed by default upon submission, and are available only to the data owner/submitter. Data access and the sharing of datasets is under the complete control of the data owner/submitter. See section 3.C.2 for details on providing access to data to other HCDB users.

The only publicly available data are summary statistics describing the number of HLA genotyped subjects in the database, the overall frequencies of the HLA alleles in the database for each locus, sex and age distributions for all subjects in the database, and acknowledgement of data contributors and their contributions. These data all appear on the database landing page.

2.A Registering an Account

All HCDB users must have database accounts. To register for a HCDB account, click ***Log In / Register Account*** in the navigation bar at the top of the landing page, and then click ***Register Account*** on the resulting page (Figure 2).

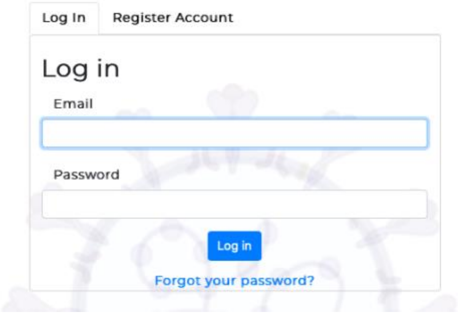
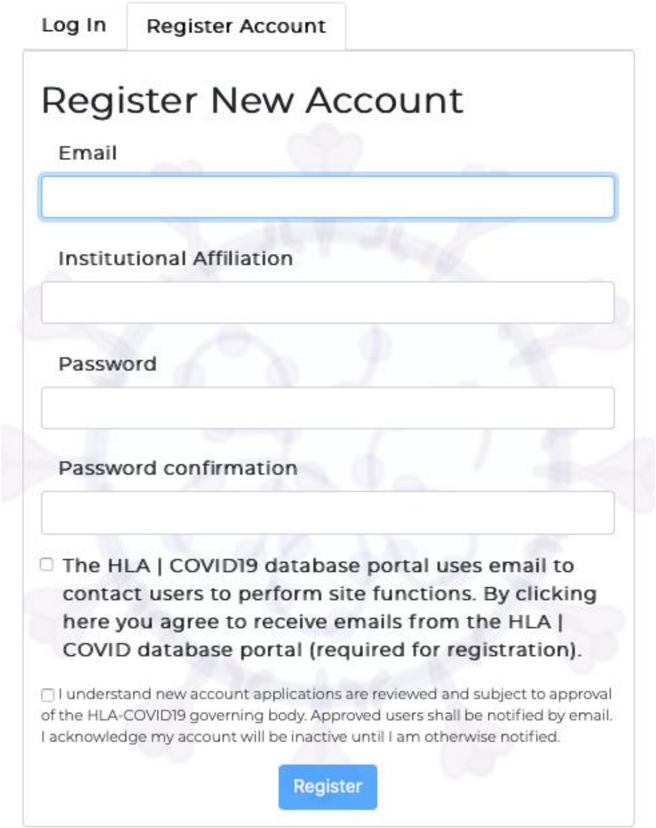


Figure 2. Log In/Register Account Page

On the resulting page (Figure 3), enter your email address, institutional affiliation and your account password (twice). Your email address will be the primary means of identifying your account.

Check each box at the bottom of the page, agreeing to receive emails from the HCDB and acknowledging that new account applications are subject to the approval of the HLA-COVID19 governing body. Click **Register**.



□ The HLA | COVID19 database portal uses email to contact users to perform site functions. By clicking here you agree to receive emails from the HLA | COVID database portal (required for registration).

□ I understand new account applications are reviewed and subject to approval of the HLA-COVID19 governing body. Approved users shall be notified by email. I acknowledge my account will be inactive until I am otherwise notified.

Figure 3. Register New Account Page

2.B Logging into the HCDB

Click ***Log In*** tab on click the ***Log In / Register Account*** page (Figure 2), and then enter the email address and password that you used to create your account (See section 2.A).

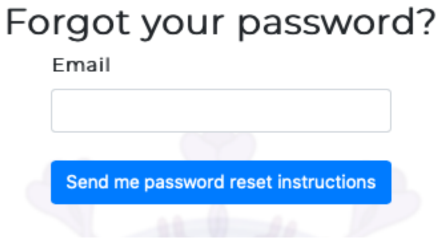
The navigation bar at the top of the page will include new links, depending on the selections that you made when you created your account.

- ***HLACOVID19.org*** will lead to the hlacovid19.org website's landing page.
- ***Download Database Manual*** will download a copy of this manual.
- ***File Uploads*** will lead to your **File Upload History** page.
- ***RStudio Server*** will open a new tab in which you can log into the RStudio cloud server.
- ***My Project*** will lead to the project-data management pages if you have uploaded a data file.
- ***My Account*** will lead to the **Your Account** account management page (Figure 5).
- ***Logout*** will log you out of the HCDB.

These options are described in detail in section 3, below.

2.C Resetting your Password

If you cannot remember your password, click ***Forgot your password?*** at the bottom of the ***Log In*** tab. On the next page (Figure 4), enter the email address that you used to create your account (See section 2.A), and click **Send me password reset instructions**.



Forgot your password?

Figure 4. Password Reset

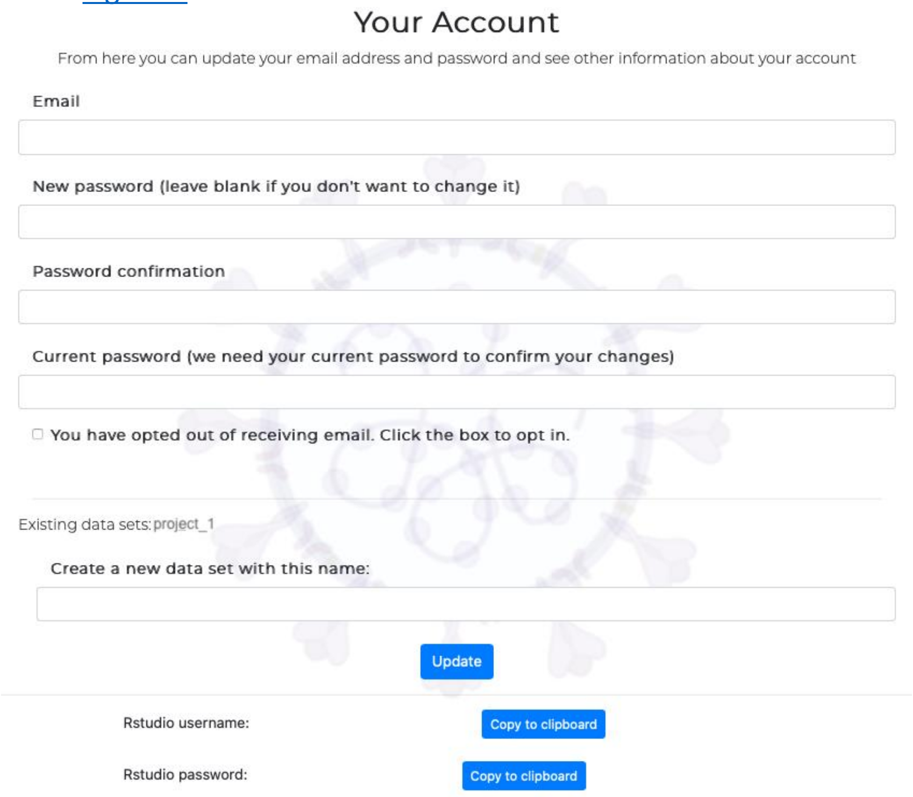
You will be taken to the ***Retry Log in*** page. Instructions on resetting your password will be sent to your email address. Follow those instructions to create a new password, and then enter the email address that you used to create your account and your new password on the ***Retry Log*** in page. Your password will not be changed until you follow the instructions in the email and create a new one.

3. HCDB Functions

3.A **Your Account** page

On the **Your Account** page (Figure 5), you can update your email address and password and enter project names for new data sets. In addition, your RStudio cloud server username and password will be shown at the bottom of the page.

Any Project Names you have registered will be listed in the ***Existing data sets*** section. A project named '*project_1*' is shown in Figure 5.



Your Account

From here you can update your email address and password and see other information about your account

Email

New password (leave blank if you don't want to change it)

Password confirmation

Current password (we need your current password to confirm your changes)

□ You have opted out of receiving email. Click the box to opt in.

Existing data sets: project_1

Create a new data set with this name:

Rstudio username:

Rstudio password:

Figure 5. MyAccount Page

To change your email, enter a new email address in the ***Email*** field, enter your password in the ***Password confirmation*** field, and click **Update**.

To change your password, enter a new password in the ***New password*** field, enter your current password in the ***Password confirmation*** field, and click **Update**.

If you are an original HCDB user, there may be an empty checkbox below the ***Current password*** field, followed by a message that reads, "You have opted out of receiving email. Click the box to opt in." (Figure 5). In these cases, please check this box, enter your password in the ***Current password*** field, and click **Update**. This will ensure that that you can receive HCDB-related emails.

If you opted in to receiving emails when you registered your HCDB account, you will instead see a message that reads, “You have opted into receiving email from the database portal. Click the box to opt out. Warning: opting out of email may render some website features unusable.”

To create a dataset, provide a project name for the associated data in the ***Create a new dataset with this name*** field. This exact Project Name must be associated with all subject data loaded into the HCDB for this project. To avoid confusion between datasets each Project Name should be:

- Descriptive
- Specific to the dataset
- Potentially unique

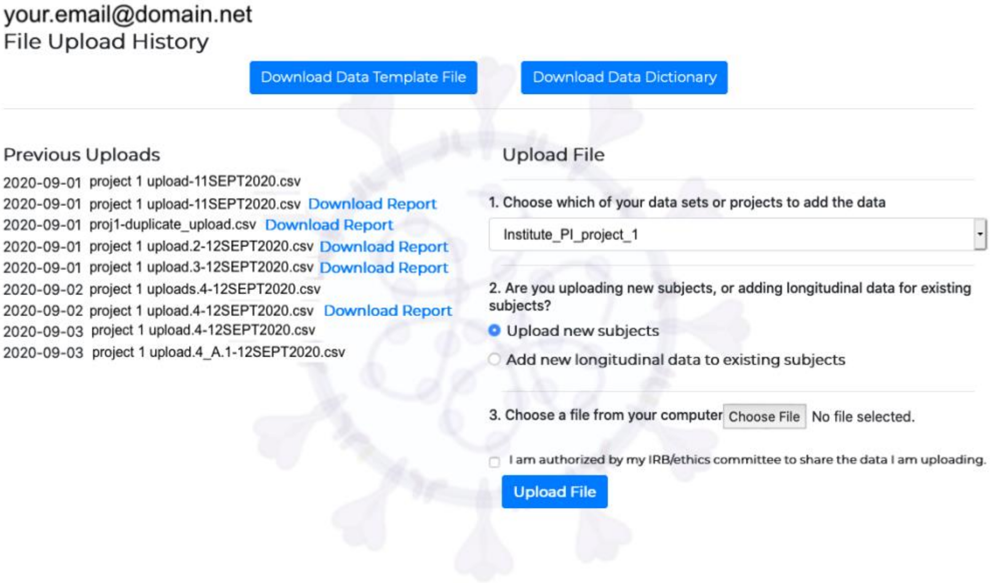
Project names can be up to 100 characters in length. Structure your Project Name as ***institution-name_investigator-name_cohort-description***. Once you have entered a new Project Name in the ***Create a new dataset with this name*** field, enter your password.

You can create as many datasets as necessary.

You can copy your RStudio cloud server username or password to the clipboard by clicking **Copy to clipboard**.

3.B **File Uploads** page

On the **File Uploads** page (Figure 6), you can download the data-template for data submission and the data dictionary. This page presents a record of all of your previous file-uploads, and provides links to download reports for each uploaded file that was processed by the database.



your.email@domain.net

File Upload History

Previous Uploads

2020-09-01 project 1 upload-11SEPT2020.csv
2020-09-01 project 1 upload-11SEPT2020.csv Download Report
2020-09-01 proj1-duplicate_upload.csv Download Report
2020-09-01 project 1 upload.2-12SEPT2020.csv Download Report
2020-09-01 project 1 upload.3-12SEPT2020.csv Download Report
2020-09-02 project 1 uploads.4-12SEPT2020.csv
2020-09-02 project 1 upload.4-12SEPT2020.csv Download Report
2020-09-03 project 1 upload.4-12SEPT2020.csv
2020-09-03 project 1 upload.4_A.1-12SEPT2020.csv

Upload File

1. Choose which of your data sets or projects to add the data

Institute_PI_project_1

2. Are you uploading new subjects, or adding longitudinal data for existing subjects?

● Upload new subjects

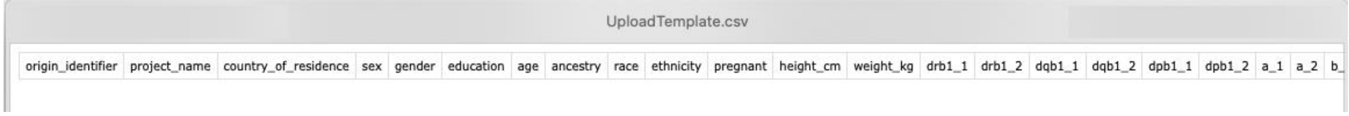
○ Add new longitudinal data to existing subjects

3. Choose a file from your computer Choose File No file selected.

□ I am authorized by my IRB/ethics committee to share the data I am uploading.

Figure 6. File Uploads Page

Click the **Download Data Template File** to download a copy of the UploadTemplate.csv file. This is a comma-separated value (CSV) file that includes all of the HCDB fields separated by commas (Figure 7). This file should be opened in a spreadsheet application, and the HCDB fields will appear as column-headers. **DO NOT EDIT ANY OF THESE FIELD NAMES. DO NOT ADD ANY FIELD NAMES**. Data for each subject should be entered on a single line of the spreadsheet. Once all of the data have been entered, the file should be saved as a CSV file.



origin_identifier	project_name	country_of_residence	sex	gender	education	age	ancestry	race	ethnicity	pregnant	height_cm	weight_kg	drb1_1	drb1_2	dqb1_1	dqb1_2	dpb1_1	dpb1_2	a_1	a_2	b_

Figure 7. Data Upload Template

Click **Download Data Dictionary** to download a copy of the HLA_COVID19_data_dictionary.xlsx file. This is a Microsoft Excel file that identifies all the HCDB fields (in Column B, **column name**), organized by data-category (in Column A). The data-categories are:

- Subjects
- HLA
- COVID19 Symptoms
- Hospitalization
- Risk Factors
- Lab Tests
- Treatments
- Comorbidities

The type of data that should be entered for each field is shown in Column C (**data type**).

- If the data type is 'boolean', **only T (for TRUE) or F (for FALSE) values should be entered for that field**.
- If the data type is 'number', **only numerals (0 – 9) should be entered for that field**.
- If the data type is 'decimal', **only decimal numbers should be entered for that field**.
- If the data type is 'text', any characters can be entered for that field. However, the specifics of the text may be restricted by the instructions in the "notes" column (Column E).

Fields for which you have no data can be left blank, or can be submitted with NA values.

The values in the **required?** column (Column D) indicate if column data must be included with each upload, as described by instructions in the **notes** column.

Values for the **origin_identifier**, **project_name**, **country_of_residence**, **age**, **c19_test_result_positive** and **c19_test_type** fields **must** be provided for every subject.
Values for either **ancestry** alone or **race** and **ethnicity** together **are recommended** for every subject.

In general, the **race** and **ethnicity** fields should be reserved for data described using United States Office of Management and Budget classifications for Race and Ethnicity (https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html). This includes five racial categories (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White) and two ethnic categories (Hispanic or Latino, and Not Hispanic or Latino).

The **ancestry** field should be used for data described using other population or ancestry identifiers.

When HLA data are uploaded, it is strongly recommended that the reference_database (e.g., IPD-IMGT/HLA Database), reference_database_version (e.g., 3.41.0), typing_method_name, and typing_method_version information be provided for each subject when available.

3.B.1 **Uploading Data for New Subjects**

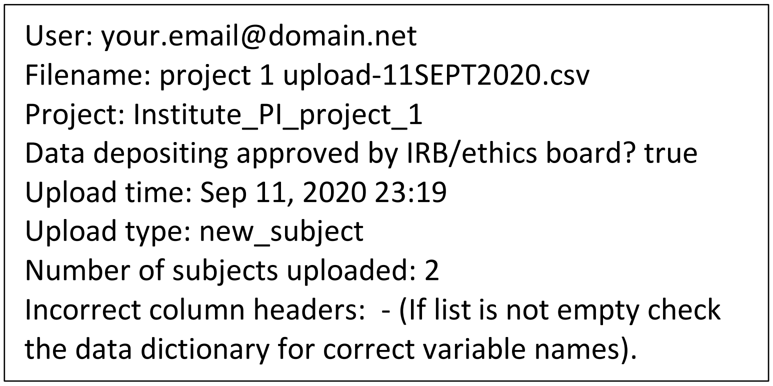
To upload data for new subjects, select the pertinent Project Name in the **Choose which of your data sets or projects to add the data** pull-down menu, click **Upload new subjects** and **Choose File**, and select the modified version of the UploadTemplate.csv file that includes your data. When you have selected a file, the filename will be displayed to the right of the Choose File button.

Please note that your institution's institutional review board (IRB) or ethics committee must approve the sharing of these data for research use. If you have received that approval, check the box that indicates that you are authorized by your IRB/ethics committee to share these data. *If you do not check this box, the data in your uploaded file will not be loaded into the HCDB.*

Once you have selected the file for upload, click **Upload File**.

NOTE: A given filename can only be uploaded once. An attempt to load a duplicate filename will result in an alert box containing an error message.

When a file has been successfully uploaded, a ***Download Report*** (Figure 8) link will be displayed to the right of the filename in the Previous Uploads section of the page. You may have to refresh the page to update the Previous Uploads section. Click that link to download a text file that identifies the user who uploaded the data, the name of the file uploaded, the Project Name associated with the uploaded data, the time and type of upload, the number of subjects uploaded, and a list of any incorrect column headers.



```
User: your.email@domain.net
Filename: project 1 upload-11SEPT2020.csv
Project: Institute_PI_project_1
Data depositing approved by IRB/ethics board? true
Upload time: Sep 11, 2020 23:19
Upload type: new_subject
Number of subjects uploaded: 2
Incorrect column headers:  - (If list is not empty check
the data dictionary for correct variable names).
```

Figure 8. Example Download Report

If a file has not been successfully uploaded, no *Download Report* link will be displayed beside the file. In cases like these, check the file to make sure that the origin_identifiers for the subjects in the file have not been previously loaded for that Project Name, and that the correct Project Name is the project_name field. Non-longitudinal data (see section 3.B.2) for a subject under a given Project Name can only be loaded once.

If you have loaded incorrect column headers, the column headers will be listed to the right of "Incorrect column headers:" (e.g., '["patient_self_reported_positive"]').

When you have uploaded data for a subject identifier with a specific origin_identifier, that subject is assigned a unique HCDB identifier (a hlac19_id), which will be included on all downloaded data tables (see section 3.C.1).

3.B.2 **Adding New Longitudinal Data for Existing Subjects**

The data for fields in the **COVID19 Symptoms**, **Hospitalization**, **Lab Tests**, **Treatments** and **Comorbidities** data-categories (Column A of the Data Dictionary) are longitudinal data and can be updated for subjects as necessary.

To upload new longitudinal data:

- Select the pertinent Project Name in the **Choose which of your data sets or projects to add the data** pull-down menu
- Click **Add new longitudinal data to existing subjects**
- Click **Choose File**
- Select the modified version of the UploadTemplate.csv file that contains new longitudinal data for specific subjects.

Note: Only subjects with new longitudinal data should be included in this file. In addition to the fields for the updated longitudinal data, only the origin_identifier and project_name fields should be completed for these subjects.

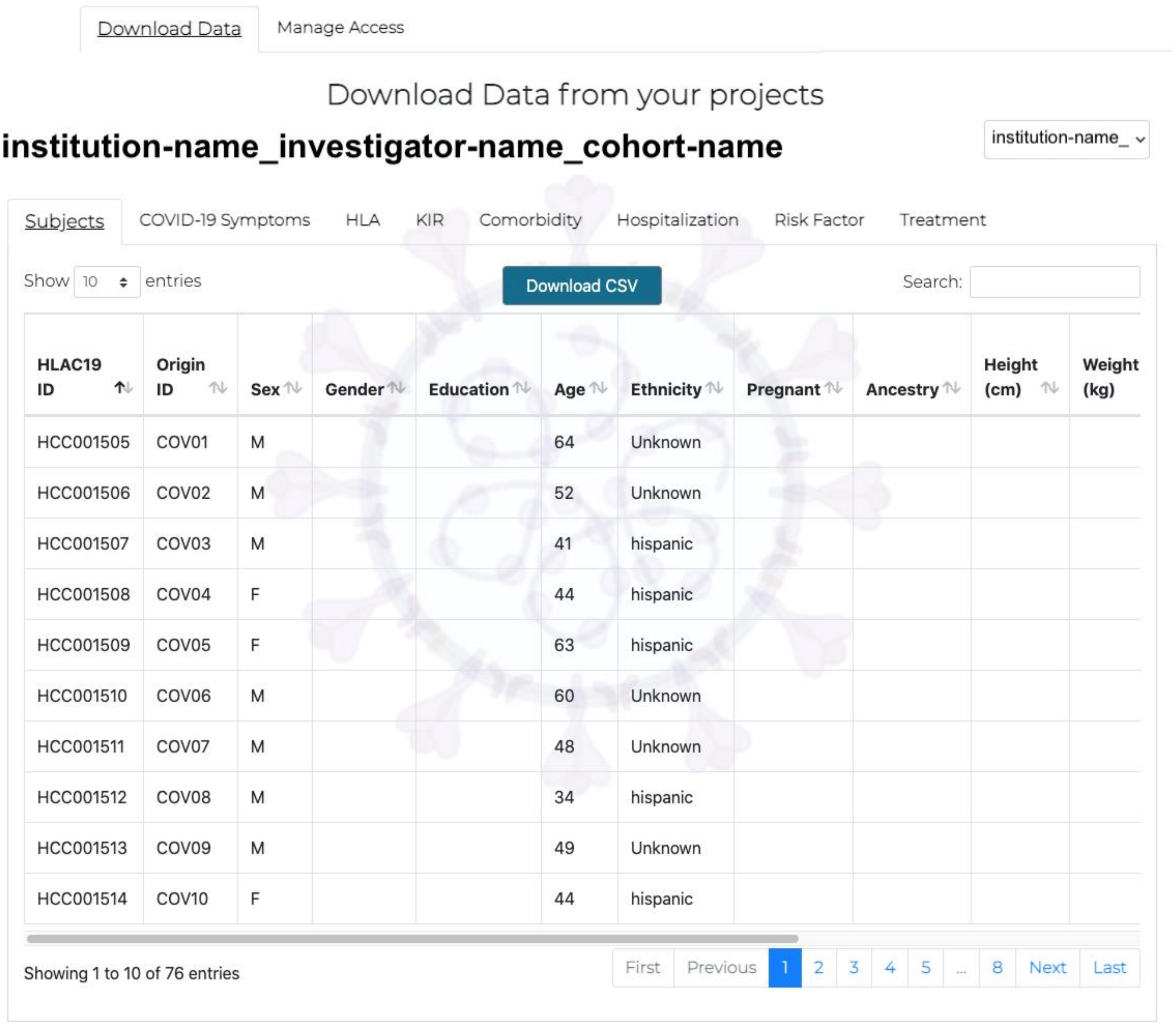
As described in section 3.B.1, only data that have been authorized by your IRB or ethics committee can be uploaded. Check the box that indicates that you are authorized by your IRB/ethics committee to share these data, and then click **Upload File**.

3.C Project-Data **Management** Pages

The ***My Project*** link will take you to a page with a Download Data tab and a Manage Access tab.

3.C.1 Download Data

The Download Data tab displays a set of sub-tabs for the data-categories (section 3.B) in each of your projects and the projects to which you have access (section 3.C.2) (Figure 9). If you have access to multiple projects, each Project Name will appear in a pulldown box in the upper right corner of the Download Data tab. Select a Project Name in that box to view its data-category sub-tabs. When available, sub-tabs for HLA imputation quality metrics or whole genome/exome read counts will also appear. Click the "Download CSV" button on each sub-tab to download the data for that sub-tab as a CSV. Each file includes each subject's hlac19_id and origin_identifier.



Download Data from your projects

institution-name_investigator-name_cohort-name

HLAC19 ID	Origin ID	Sex	Gender	Education	Age	Ethnicity	Pregnant	Ancestry	Height (cm)	Weight (kg)
HCC001505	COV01	M			64	Unknown				
HCC001506	COV02	M			52	Unknown				
HCC001507	COV03	M			41	hispanic				
HCC001508	COV04	F			44	hispanic				
HCC001509	COV05	F			63	hispanic				
HCC001510	COV06	M			60	Unknown				
HCC001511	COV07	M			48	Unknown				
HCC001512	COV08	M			34	hispanic				
HCC001513	COV09	M			49	Unknown				
HCC001514	COV10	F			44	hispanic				

Showing 1 to 10 of 76 entries

Figure 9. Download Data Tab

3.C.2 Manage Access

On the Manage Access tab, you can specify which registered HCDB users have access to the data you have uploaded for a specific Project Name. When you check the checkbox next to a specific HCDB user for a specific Project Name, that user will be authorized to download all data for that Project Name.

4. RStudio Cloud Server

The RStudio cloud server supports the installation and development of data-management and -analysis tools. If you are not experienced with the R programming language and the RStudio IDE you **do not** need to use the RStudio cloud server. The RStudio cloud server can be accessed using the *RStudio server* link in in the navigation bar at the top of the page.

When you click that link, you will be taken to the **RStudio Sign In** page (Figure 10).

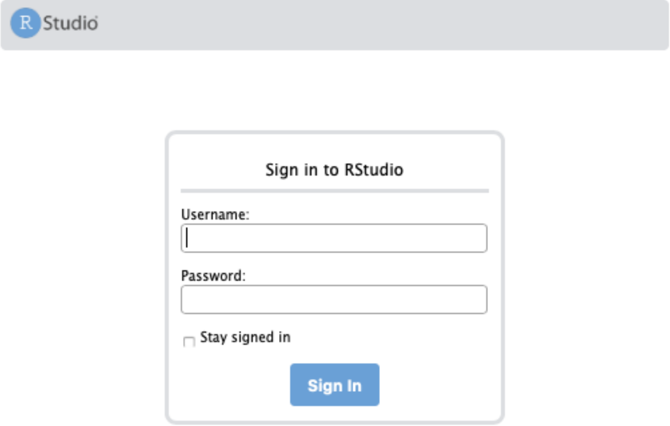


Figure 10. RStudio Server Page

4.A **Signing In to the R Studio Server**

To sign into the RStudio server, enter the **username** and **password** at the bottom of the **Your Account** page. You can copy the password to your clipboard by clicking **Copy to clipboard** (see section 3.A). If you want to remain signed into the RStudio server, check **Stay signed in**. Checking this box will allow you to return to your RStudio server environment, even after closing the window, by navigating to https://database-hlacovid19.org/rstudio/.

To **Sign out** of the RStudio server environment, click **Sign out** to the right of your username in the top right corner of the window (Figure 11).

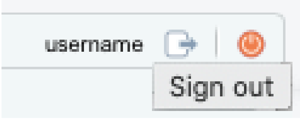


Figure 11. RStudio Server Sign Out

4.B **RStudio Server Accounts**

Each RStudio server account is an independent installation of R version 4.02 in RStudio version 1.2.5042, with primarily default R packages installed. Each RStudio server user will have to install any additional R packages needed.

4.C **Shiny Server**

The database maintains a Shiny Server for hosting Shiny apps. Any outward facing data-analysis tools should be built into a Shiny app that can be hosted on the Shiny Server. Each RStudio Server account comes with shiny version 1.4.0.2 installed.

5. HLA Imputation Portal

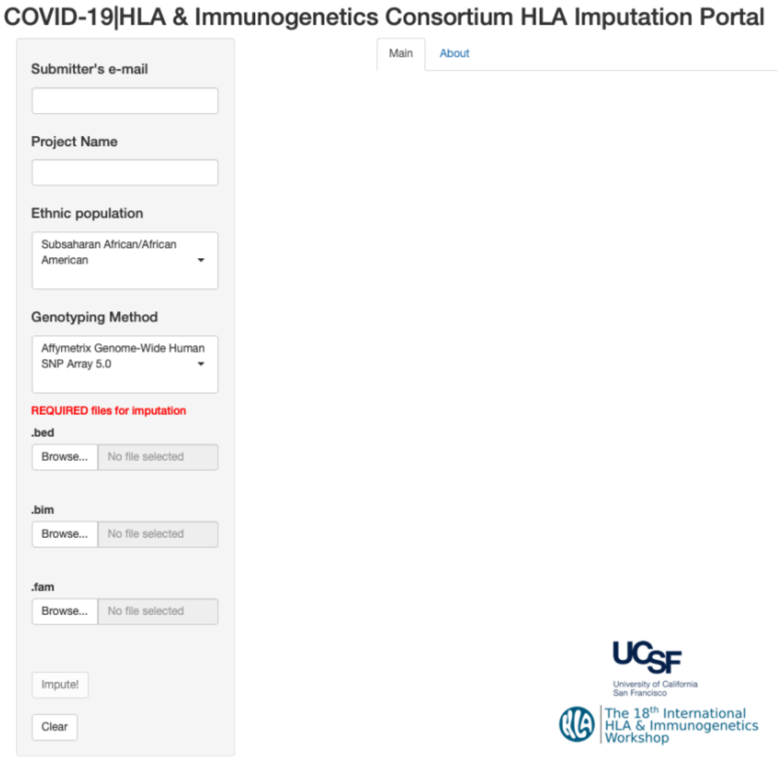
The HLA Imputation Portal (HIP) consumes .bed, .bim, and .fam files for SNP-genotyped subjects, and imputes *HLA-A, -C, -B, -DRB1, -DQA1, -DQB1, and -DPB1* genotypes for each subject. The HIP does not consume .ped and .map files. PLINK can be used to convert .ped and .map files to .bed, .bim and .fam files on your system (zzz.bwh.harvard.edu/plink/data.shtml#bed).

Note: The HIP can only be applied to subjects with origin identifiers that have been previously loaded into the HCDB under a specific Project Name (see section 3.B.1).

SNPs must be described using hg19 SNP rs IDs and genomic coordinates. If your SNP data are not recorded using hg19 genome coordinates and annotations, lift-over tools are available online (e.g. https://genome.ucsc.edu/cgi-bin/hgLiftOver).

The HIP relies on 129 models built from genome-wide SNP data generated on 30 array platforms for subsets of African, Asian, European, Hispanic and Multi-ethnic groups. For details of the platforms and the associated population models, visit https://zhengxwen.github.io/HIBAG/platforms.html.

To access the HIP, go to https://database-hlacovid19.org/shiny in your browser, and click **HLA-Imputation-Portal/**. Once connected to the HIP, provide your HCDB account's associated **email address**, and the **Project Name** associated with the origin identifiers for each subject in your SNP dataset (Figure 12). As noted above, these origin identifiers **must** be included in the uploaded .fam file, and must have been loaded into the HCDB **before** using the HIP.



COVID-19|HLA & Immunogenetics Consortium HLA Imputation Portal

Submitter's e-mail

Project Name

Ethnic population: Subsaharan African/African American

Genotyping Method: Affymetrix Genome-Wide Human SNP Array 5.0

REQUIRED files for imputation

.bed: No file selected

.bim: No file selected

.fam: No file selected

Main
About
Browse...
Browse...
Browse...
Impute!
Clear
UCSF
University of California San Francisco
The 18th International HLA & Immunogenetics Workshop

Figure 12. HLA Imputation Portal

Select the pertinent population group in the **Ethnic population** pulldown menu, and the pertinent SNP genotyping assay in the **Genotyping Method** pulldown menu. Note that not all combinations of ethnic population and genotyping assay are possible. If a specific population group is not available for your genotyping assay, choose the **Multi-ethnic/Other** group.

Click **Browse…** for each of the .bed, .bim and .fam files to select the pertinent file for imputation. These data are extracted from your local files; the files are not uploaded to the HIP.

When the submitter's email and the Project Name have been provided, the population and assay have been selected, and the source files have been successfully uploaded, click **Impute!** to begin the imputation process. Any messages about the imputation process will be displayed in the HIP's Main tab. Specific information about these messages is provided in the About tab.

While the imputation is running, a circle of spinning dots will appear in the upper-right of the page, the HIP page will be greyed out, and a white bar with the message **Your imputation is running…** will appear near the top of the page.

When the imputation is complete, a message will appear in the Main tab, and an email notification will be sent to the address provided in the **Submitter's e-mail** field. Imputed HLA genotypes for each subject will be automatically loaded into the specified Project Name of the HCDB account associated with the Submitter's e-mail address for the origin identifiers in the .fam file. In addition, a downloadable file of posterior probability values for each subject and each locus will be loaded into your account.

5.A: **Recommendations for using the HIP**

Because the imputation process relies on specific ethnic models, it may be most efficient to generate separate .fam files for subjects of different ethnicities, and run each .fam file separately. Imputation performed using the Multi-ethnic models can take significantly longer than using the single-ethnicity models, and use of the African, Asian, European and Hispanic models when possible is recommended. In instances where a specific ethnic model is not available for a particular SNP assay, use the Multi-ethnic model.

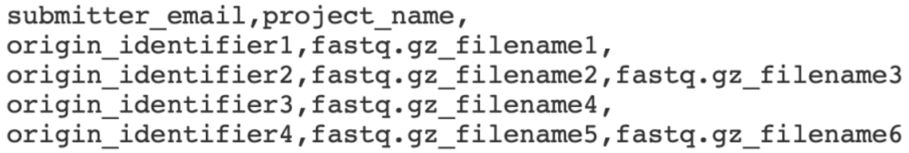
6. Omixon Genotyping Portal

The Omixon Genotyping Portal (OGP) consumes fastq.gz files for whole-genome or whole-exome sequenced (WGS/WES) subjects, and generates *HLA-A, -C, -B, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1 and -DPB1* genotypes for each subject. Because WGS/WES fastq.gz files can be very large, only fastq.gz files containing HLA-gene specific reads can be uploaded to the OGP. The *wgsHLAfiltR* R package can be used to generate HLA-specific fastq.gz files (see section 6.A).

Note: The OGP can only be applied to subjects with origin identifiers that have been previously loaded into the HCDB under a specific Project Name (see section 3.B.1).

The OGP relies on Omixon's CLI Explore software to generate HLA genotypes using IPD-IMGT/HLA Database release version 3.39.0 sequence alignments. CLI Explore is a command-line interface version of Omixon's HLA Explore software, and performs HLA genotyping on the basis of HLA exon sequence information. Given this, the OGP generates HLA genotypes with fourth-field ambiguity, reported in GL String format (see: dx.doi.org/10.1111/tan.12150 for more information about GL String format).

The OGP can consume paired or unpaired fastq.gz files, and each fastq.gz pair or single fastq.gz file must be associated with a specific origin identifier. To structure the submission of this information, the OGP consumes a three-column "User Information" file written in CSV format (illustrated in Figure 13).



```
submitter_email,project_name,
origin_identifier1,fastq.gz_filename1,
origin_identifier2,fastq.gz_filename2,fastq.gz_filename3
origin_identifier3,fastq.gz_filename4,
origin_identifier4,fastq.gz_filename5,fastq.gz_filename6
```

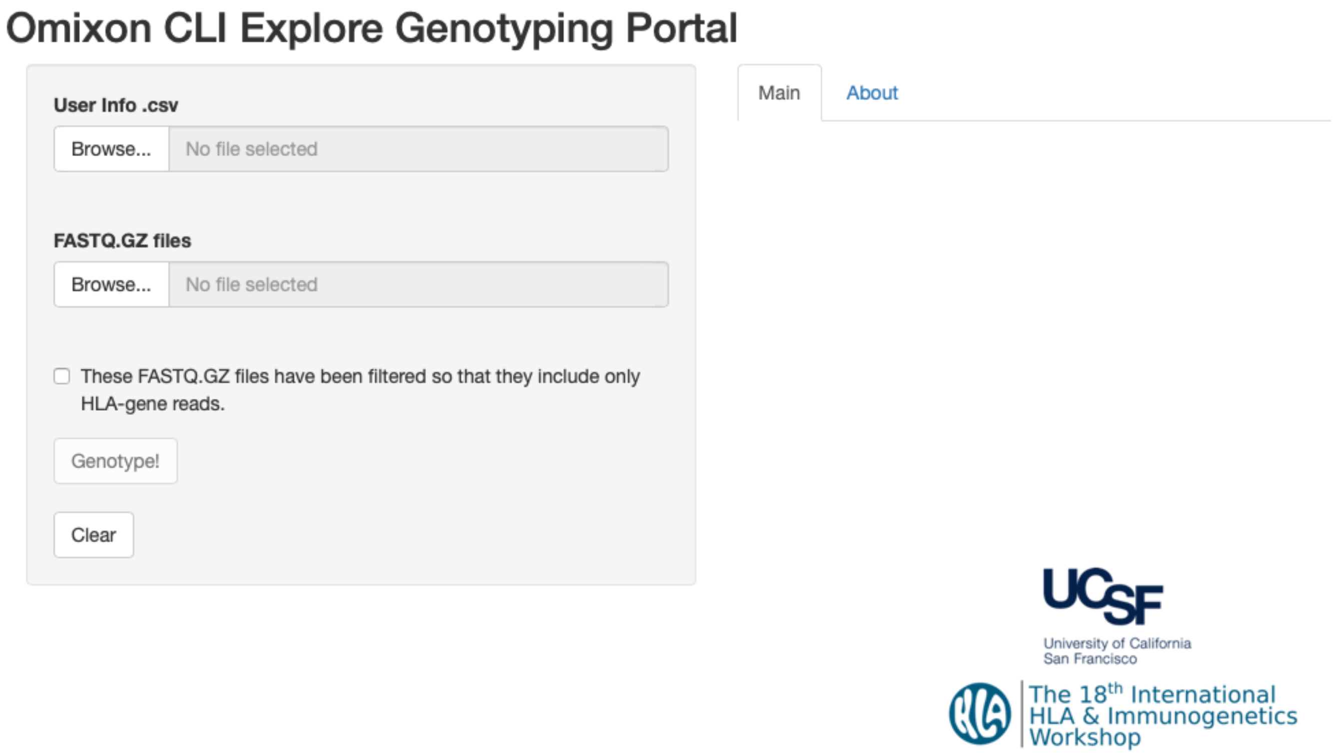
Figure 13. Example OGP User Information file

The first (header) line of the User Information file should contain the email associated with the submitter's HCDB account, and the Project Name associated with the sequenced subjects, separated by a comma, and followed by a comma.

Each subsequent line of the User Information file should contain the origin identifier of each subject and the name of the unpaired fastq.gz file, or the names of the paired fastq.gz files, associated with that subject. These elements must be separated by commas, and **each row of the file must include two commas**. If only one fastq.gz filename is provided, it should be followed by a comma. As noted above, these origin identifiers **must** have been loaded into the HCDB **before** using the OGP.

All of the fastq.gz files specified in the User Information file should be located in the same directory (or folder) on your system. To minimize upload time, exclude all other fastq.gz files from that directory.

To access the OGP, go to https://database-hlacovid19.org/shiny in your browser, and click **Omixon-Genotyping-Portal/**. Once connected to the OGP, click the 'Browse…' button under ***User Info .csv*** (Figure 14), and select the User Information file. Once that file has been uploaded, click the 'Browse…' button under ***FASTQ.GZ files*** and select the directory containing the fastq.gz files. Click the checkbox indicating that your fastq.gz files have been filtered to include only HLA-gene reads. The *wgsHLAfiltR* R package can generate HLA-specific fastq.gz files (see section 6.A). Once those files have been uploaded (which can take a few minutes), click the ***Genotype!*** button to start the genotyping process.



Omixon CLI Explore Genotyping Portal

User Info .csv

No file selected

FASTQ.GZ files

No file selected

☐ These FASTQ.GZ files have been filtered so that they include only HLA-gene reads.

Browse...
Browse...
Genotype!
Clear
Main
About
UCSF
University of California San Francisco
HLA
The 18th International HLA & Immunogenetics Workshop

Figure 14. The Omixon Genotyping Portal

Messages about the genotyping process will be displayed in the OGP's Main tab. Specific information about these messages is provided in the About tab.

While the genotyping is being performed, a circle of spinning dots will appear in the upper-right of the page, the OGP page will be greyed out, and a white bar with the message **Your genotyping is running…** will appear near the top of the page.

When the genotyping is complete, a message will appear in the Main tab and an email notification will be sent to the e-mail address provided in the User Information file. GL String-formatted HLA genotypes for each subject will be automatically loaded into the specified Project Name of the HCDB account associated with that e-mail address for the origin identifiers in the User Information file. In addition, a downloadable file of read counts for each subject at each HLA gene will be loaded into your account.

6.A: Filtering HLA-only Reads Using the *wgsHLAfiltR* R Package

The *wgsHLAfiltR* R package extracts reads that map to the *HLA-A, -C, -B, -DRB1, -DRB3/4/5, -DQA1, -DQB1, -DPA1 and -DPB1* loci from paired or individual whole-genome or whole-exome sequencing (WGS/WES) fastg.gz files, and writes a new set of fastq.gz files that contain reads that map to the classical HLA loci.

The package requires R version 4.0.0 or higher to run, and only runs in R environments installed on Unix, Linux or macOS systems. Bowtie2 (bowtie-bio.sourceforge.net/bowtie2/index.shtml) versions 2.3 - 2.4 must be installed on the same system as the R environment running *wgsHLAfiltR*.

To install the *wgsHLAfiltR* package in the R environment, first execute the *install.packages("devtools")* command in the R console to install the *devtools* package.

With *devtools* installed, execute the *devtools::install_github(repo="COVID-HLA/wgsHLAfiltR/wgsHLAfiltRpackage",ref="main")* command in the R console to install *wgsHLAfiltR*.

WgsHLAfiltR includes > 130MB of reference alignment data, which may result in longer than expected installation times.

When *wgsHLAfiltR* is installed, load the package using the *library(wgsHLAfiltR)* command.

WgsHLAfiltR's filterHLA() function is the main function for filtering HLA reads. This function takes two arguments, *inputDirectory*, which identifies the path to the directory containing the WGS/WES fastq.gz files to be filtered, and *outputDirectory*, which specifies the directory into which the HLA-only fastq.gz files should be written. *Note: Bowtie 2 requires file names and paths that do not include whitespaces.*

FilterHLA() requires a value for *inputDirectory*, while the *outputDirectory* argument is optional. If *outputDirectory* is not specified, HLA-only fastq.gz files will be written into a "Results" directory in the R working directory. A "Results" directory will be created if one is not present in the R working directory.

To apply the *filterHLA()* function, execute the *readFilteringData <- filterHLA(inputDirectory=inputDir)* command in the R console, where 'inputDir' is the path to the directory of WGS/WES fastq.gz files.

The *filterHLA()* function returns a list object of named list elements for each FASTQ "name" that was processed. Each named list identifies parameters and file paths used in the read extraction process.

For more details on using wgsHLAfiltR visit https://github.com/COVID-HLA/wgsHLAfiltR/.