BC|GENOME User Guide

CE

Version/5.0

Copyright

Disclaimer

Disclaimer info.

Feedback

Send your comments by e-mail to: support@bcplatforms.com

Trademarks and registered trademarks

Products and product names mentioned in this document may be trademarks or registered trademarks of their respective owners.

Revision history

Version	Issued	Author name	Description
5	Dec, 2018	BC	Draft version

Table of Contents

1. Introduction to BC|Genome

BC|GENOME is a data and research management platform for genetic studies. The platform offers tools for complex queries in clinical and genetic data for multiple purposes. BC|GENOME scales up from small candidate gene studies to massive international collaboration environments. The platform is used to organize data projects, and to administer user and user-group access to sources.

BC|GENOME helps data and research managers to maintain and increase data integrity, quality, and efficiency of work. Research data is stored in a database on a server, but because the user interface is web-based, the system can be used on any computer with a browser. The basic implementation of the BC|GENOME system is shown in Figure 1.

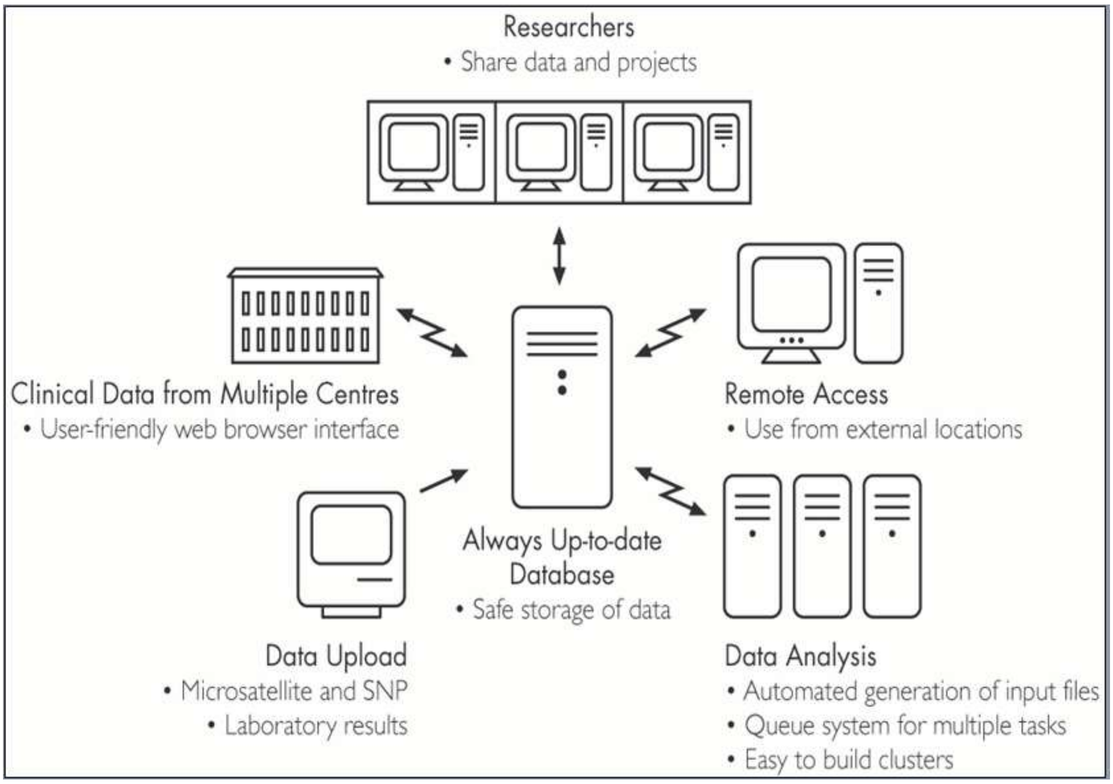

Researchers
• Share data and projects
Clinical Data from Multiple Centres
• User-friendly web browser interface
Remote Access
• Use from external locations
Always Up-to-date Database
• Safe storage of data
Data Upload
• Microsatellite and SNP
• Laboratory results
Data Analysis
• Automated generation of input files
• Queue system for multiple tasks
• Easy to build clusters

Figure 1. Basic implementation of BC|GENOME.

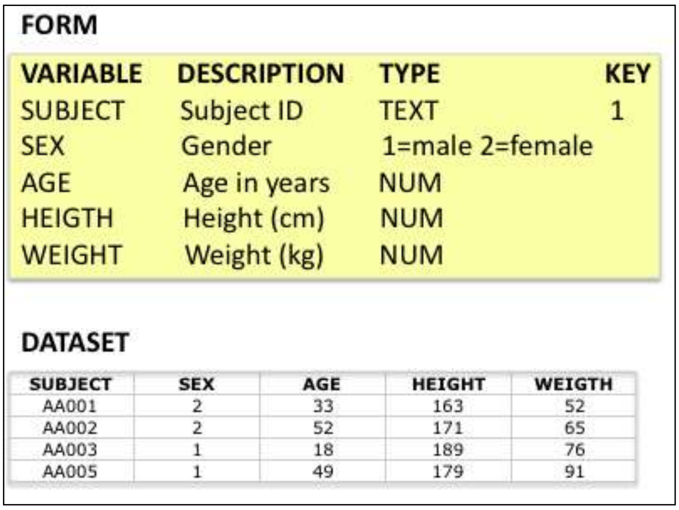
1.1. BC|GENOME database structure

Research data is stored in an organized way in the BC|GENOME database. To understand the research data structure, the terms dataset, form, key field and subset are described in this chapter.

1.1.1. Dataset

In the BC|GENOME terminology, dataset refers to a database table. A dataset can be thought of as a large spreadsheet, in which dataset fields represent the column headers and data items represent different data rows. As default, a dataset is stored in the database but in the case of genotype data it can also be stored on a server in binary format using a special form that enables data compression (only supported in the BC|GENOME product family). Figure 2 shows an example dataset and its corresponding form.

Different data types, for example, phenotypes and pedigree data are stored in separate datasets. Data from different projects is also stored separately. Access rights to the data can be granted separately for each dataset.



FORM

VARIABLE	DESCRIPTION	TYPE	KEY
SUBJECT	Subject ID	TEXT	1
SEX	Gender	1=male 2=female	
AGE	Age in years	NUM	
HEIGTH	Height (cm)	NUM	
WEIGHT	Weight (kg)	NUM	

DATASET

SUBJECT	SEX	AGE	HEIGHT	WEIGTH
AA001	2	33	163	52
AA002	2	52	171	65
AA003	1	18	189	76
AA005	1	49	179	91

Figure 2. Form defines the dataset structure.

1.1.2. Form

The form defines the structure of a dataset. Each dataset is based on one form. Many datasets can be based on the same form structure. The form defines the fields a dataset has and the types of values (text, numeric, date), that can be stored in the fields (Table 1.).

Table 1. Summary of variable types

Variable type	Type name in the form	Notes
Text	Text	String variable (*
Numeric	Number	Either type of float or integer
Date	Date	Date format can be edited (*
Multiple choice	Choice	At least two options needs to be defined, data is stored as a integer in the database (*
Checkbox	Checkbox	Unchecked value is stored as 0 in the DB2 database
Paragraph text	Paragraph	32,000 characters can be stored
Timestamp	Timestamp	Format: yyyy-mm-dd hh:mm:ss (*
Text	Text	String variable (*

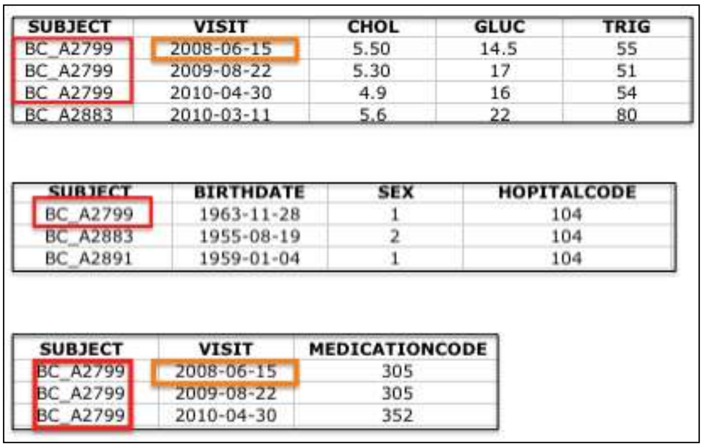
* Text (always the primary key) and optionally date, multiple choice and timestamp variables defines unique rows in a dataset (see chapter 1.1.3).

! see, https://en.wikipedia.org/wiki/SQL

1.1.3. Key fields

Each dataset has one, two or three variables to identify the table ***key***. The key identifies unique rows in a dataset (see Figure 3) and also serves as a linking identifier across different data sections. By default, the primary key is the subject ***ID*** (except in Annotations, it is the marker). If the subject ***ID*** field is the only key field, the number of rows in the dataset corresponds to the number of subjects.

When the subject ID is always of ***Type*** text in BC|GENOME, the second and third key can be either of type: text, date, timestamp or multiple choice. The date of the laboratory measurement or the hospital visit is often used as a second key. The third key can be used to define, for example, several measurements of a subject in a day. Then each subject has as many rows in a dataset as the amount of times and measuring points.



SUBJECT	VISIT	CHOL	GLUC	TRIG
BC_A2799	2008-06-15	5.50	14.5	55
BC_A2799	2009-08-22	5.30	17	51
BC_A2799	2010-04-30	4.9	16	54
BC_A2883	2010-03-11	5.6	22	80

SUBJECT	BIRTHDATE	SEX	HOPITALCODE
BC_A2799	1963-11-28	1	104
BC_A2883	1955-08-19	2	104
BC_A2891	1959-01-04	1	104

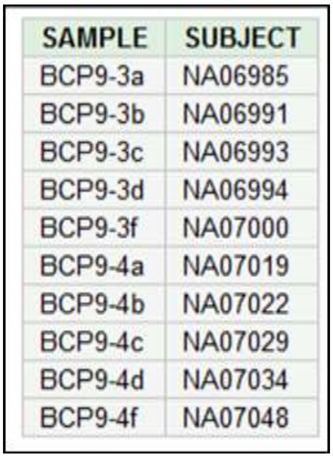
SUBJECT	VISIT	MEDICATIONCODE
BC_A2799	2008-06-15	305
BC_A2799	2009-08-22	305
BC_A2799	2010-04-30	352

Figure 3. Key variables (here SUBJECT and VISIT) uniquely identify each row in the dataset and links different datasets together.

1.1.4. Sample ID conversion

The ***SUBJECT ID*** field is a linking identifier across data sections except in the Annotations tab. Therefore, it is important to use the same subject ID coding across all data. However, the genotype files often contain sample IDs instead of subject IDs, and thus conversion is needed.

With BC|GENOME, you can automatically convert sample IDs to corresponding subject IDs during the genotype upload process. The sample ID – subject ID conversion key pairs are stored in the sample IDs tab (see Figure 4).

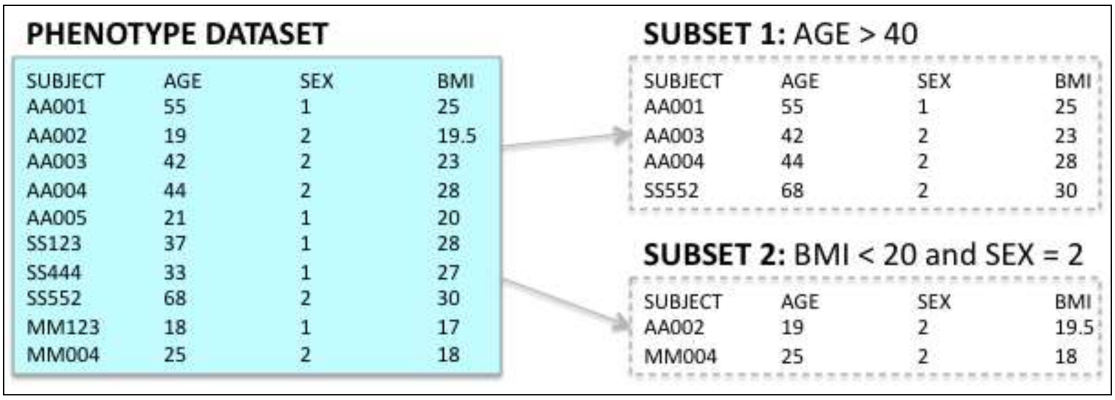


SAMPLE	SUBJECT
BCP9-3a	NA06985
BCP9-3b	NA06991
BCP9-3c	NA06993
BCP9-3d	NA06994
BCP9-3f	NA07000
BCP9-4a	NA07019
BCP9-4b	NA07022
BCP9-4c	NA07029
BCP9-4d	NA07034
BCP9-4f	NA07048

Figure 4. Laboratory Sample ID – Subject ID conversion dataset.

1.1.5. Subset

A subset is a partial view of the original dataset, so no data is actually copied and all modifications in the original data also update rows in the subset. A subset of a dataset can be created to view only the rows that fill certain criteria (see Figure 5), or several datasets can be joined to match a set filter. The subset tool can also be used to hide data columns; and writable subsets can even be created.



PHENOTYPE DATASET

SUBJECT	AGE	SEX	BMI
AA001	55	1	25
AA002	19	2	19.5
AA003	42	2	23
AA004	44	2	28
AA005	21	1	20
SS123	37	1	28
SS444	33	1	27
SS552	68	2	30
MM123	18	1	17
MM004	25	2	18

SUBSET 1: AGE > 40

SUBJECT	AGE	SEX	BMI
AA001	55	1	25
AA003	42	2	23
AA004	44	2	28
SS552	68	2	30

SUBSET 2: BMI < 20 and SEX = 2

SUBJECT	AGE	SEX	BMI
AA002	19	2	19.5
MM004	25	2	18

Figure 5. Subsets can be used for viewing only rows that fill use- defined criteria.

2. Getting started with BC|GENOME

To login to BC|GENOME, you need both a personal user ID and a password, which are provided by a local BC|GENOME administrator. When you receive your ID and password you can log in to BC|GENOME.

The BC|GENOME user interface is a web-based database platform and it is optimised to be used with the most recent Mozilla Firefox, Safari and Chrome web browsers.

2.1. Logging into BC|GENOME

Once you have received your ID and password you can log in to BC|GENOME:

1. Go to **https://server/bcapp/** (replace server with the name or IP address of your own server).

   The account information is provided by your local database owner (usually the project PI).

2. Type your user ID.
3. Type your password.

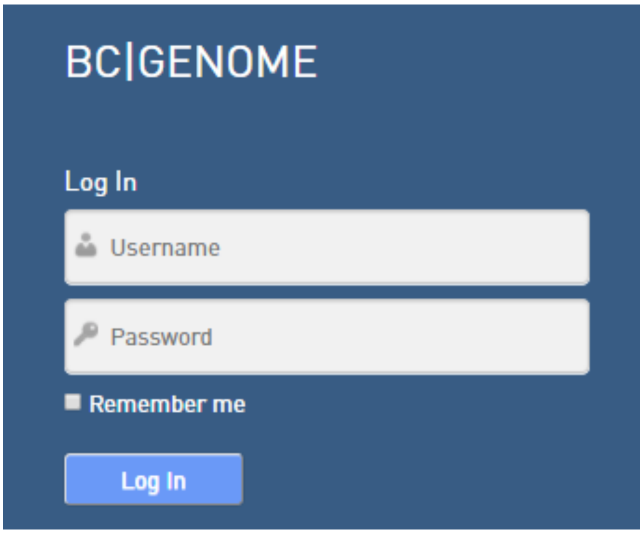


Figure 6. BC|GENOME login page.

4. Select ***Log In***.

NOTE: You can enter an incorrect user name and password a maximum of 10 times before access to BC|GENOME is locked. If this happens, contact support@bcplatforms.com to unlock the account.

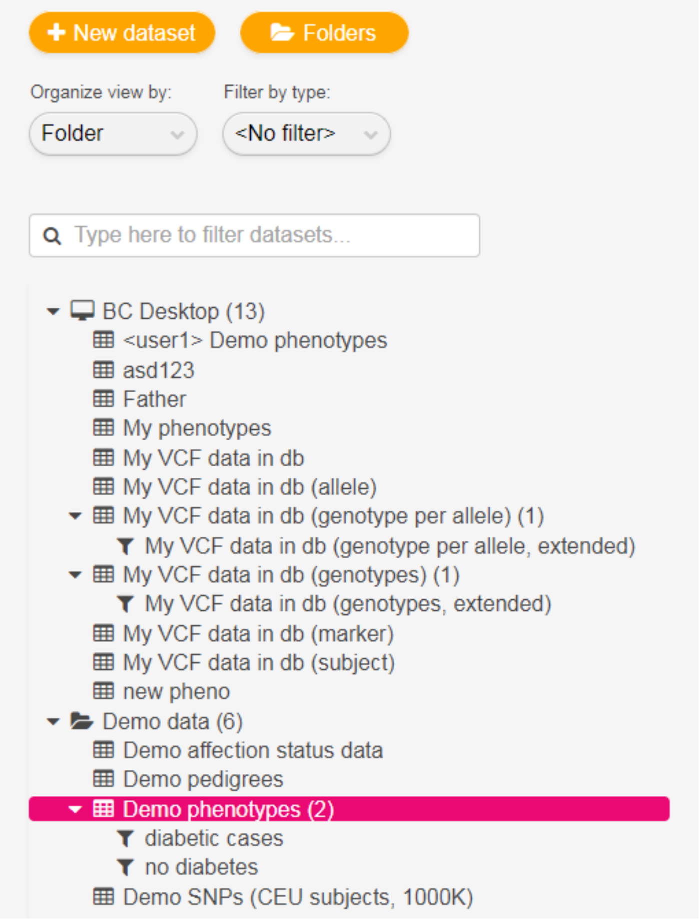
2.2. Navigating in BC|GENOME

In the navigation pane in the left of the main page you can:

- View the available datasets

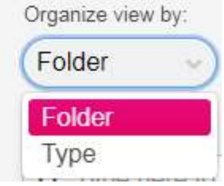
- Filter the view of the datasets
- Add a new dataset
- Add new folders

Organised by Folder and without a filter

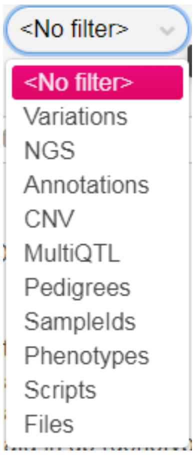


- BC Desktop (13)
  - <user1> Demo phenotypes
  - asd123
  - Father
  - My phenotypes
  - My VCF data in db
  - My VCF data in db (allele)
  - My VCF data in db (genotype per allele) (1)
    - My VCF data in db (genotype per allele, extended)
  - My VCF data in db (genotypes) (1)
    - My VCF data in db (genotypes, extended)
  - My VCF data in db (marker)
  - My VCF data in db (subject)
  - new pheno
- Demo data (6)
  - Demo affection status data
  - Demo pedigrees
  - Demo phenotypes (2)
    - diabetic cases
    - no diabetes
  - Demo SNPs (CEU subjects, 1000K)

You can choose to organise by Folder or Type:



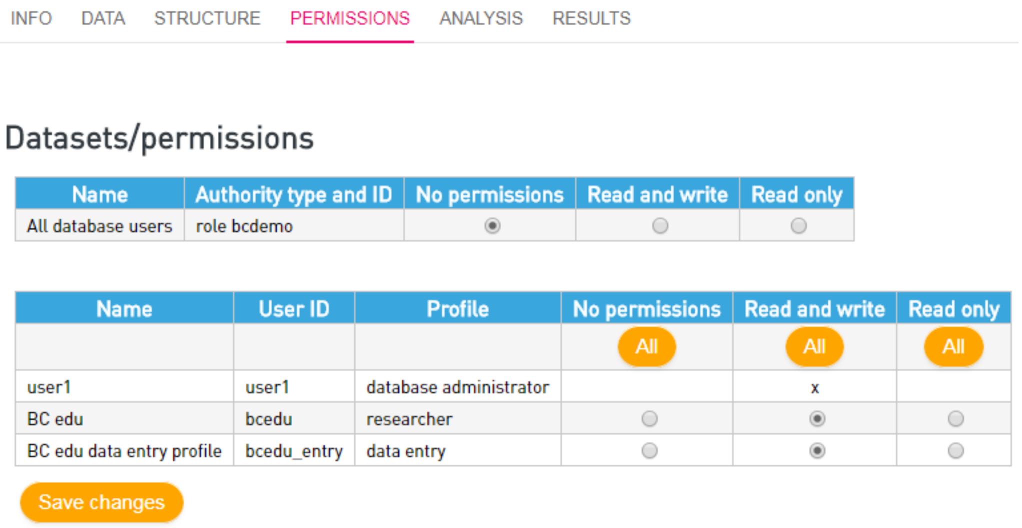
You can also choose to add a filter:



2.3. Granting permissions to a dataset

By default, only the user who has created a dataset can see and use it, but the user can give read and write permission to other BC|GENOME users using the ***Datasets > permissions function***. Reports in the result archive can also be shared with other users.

1. Select the ***PERMISSIONS*** tab to open the PERMISSIONS page.
2. Set the permissions for the users.
3. Select ***Save changes***.



Datasets/permissions

Name	Authority type and ID	No permissions	Read and write	Read only
All database users	role bcdemo	●	○	○

Name	User ID	Profile	No permissions	Read and write	Read only
			All	All	All
user1	user1	database administrator		x	
BC edu	bcedu	researcher	○	●	○
BC edu data entry profile	bcedu_entry	data entry	○	●	○

3. Using BC|GENOME

BC|GENOME functions are described in this chapter. The main tasks are:

- Creating forms and datasets to be used in BC|GENOME
- Uploading data to be used in BC|GENOME
- Viewing results and in BC|GENOME

If you have any additional questions or issues, contact BC support at support@bcplatforms.com.

3.1. Creating forms and datasets

NOTE: For more information about managing and editing forms, see chapter 4.

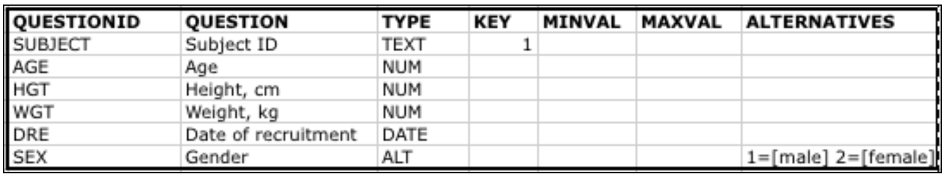
Phenotype table structures are created with the BC|GENOME ***WEB FORM EDITOR***. There are three ways to create forms:

- Manually, question by question (see section 4.2).

  The basic way to create a form is from scratch, field-by-field. This method is practical if, for example, there is already a questionnaire form on paper and the BC|GENOME dataset structure should correspond to it.

- Using a ***Form Import*** file (see section 4.3).

  Form import files are convenient, if there is a high number of variables in the form. Form import files are especially useful when importing data from an already existing clinical database and when the variable definitions are available.



QUESTIONID	QUESTION	TYPE	KEY	MINVAL	MAXVAL	ALTERNATIVES
SUBJECT	Subject ID	TEXT	1			
AGE	Age	NUM				
HGT	Height, cm	NUM				
WGT	Weight, kg	NUM				
DRE	Date of recruitment	DATE				
SEX	Gender	ALT				1=[male] 2=[female]

Figure 7. Form import file example.

- Copying questions from an already existing form.

  Forms can also be created by copying the structure from an already existing form.

If there are pre-installed forms, for example, for variation and annotation dataset, these forms can be edited using tools in ***TABLE EDITOR***: When variables in the form have been copied (***Form > new***: Copy questions from) to a new form, you can add new variables indicating, for example, quality information.

3.2. Uploading data

There are several ways to enter data into BC|GENOME datasets:

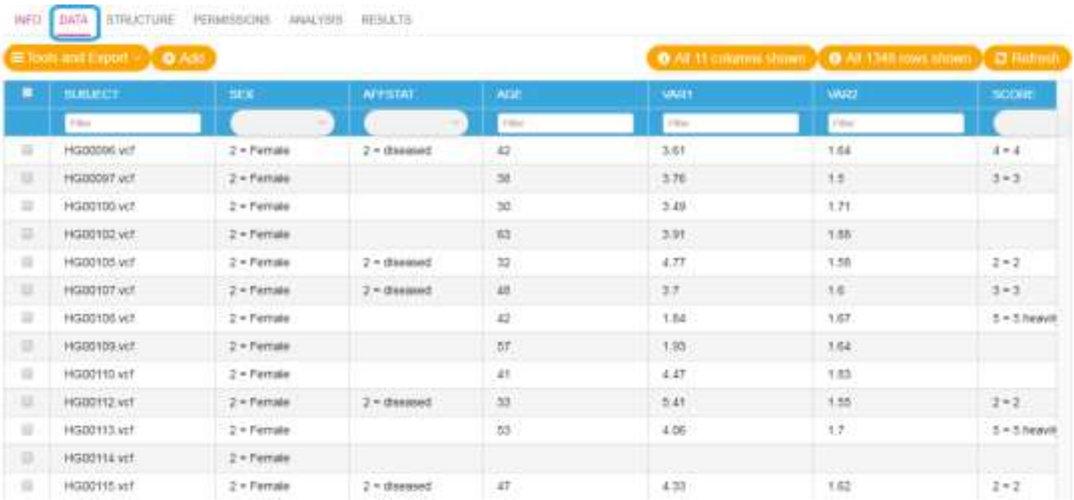
- Updating an existing subject
- Adding a new subject to the dataset
- Uploading a single file
- Uploading files from a server

3.2.1. Updating an existing subject

Updating an existing subject is most often performed by research nurses, who are entering data into the database while interviewing a subject.

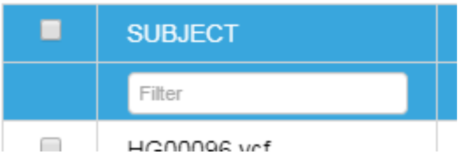
To update and existing subject:

1. Select the ***DATA*** tab to open the DATA page.



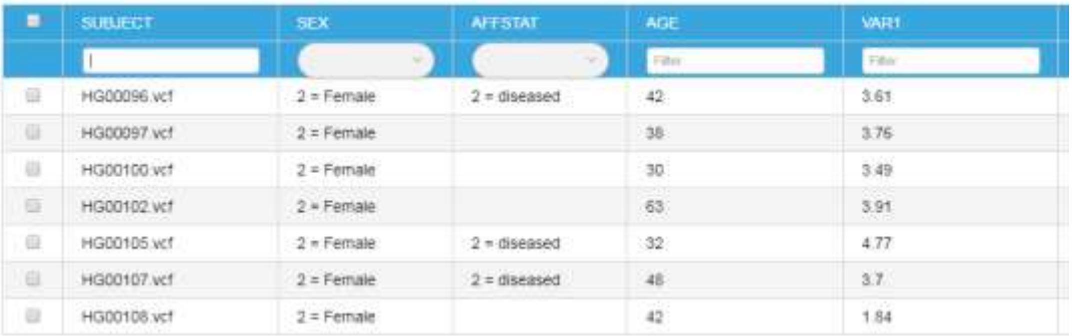
	SUBJECT	SEX	AFFSTAT	AGE	VAR1	VAR2	SCORE
□	HG00096.vcf	2 = Female	2 = diseased	42	3.61	1.64	4 = 4
□	HG00097.vcf	2 = Female		58	3.76	1.5	3 = 3
□	HG00100.vcf	2 = Female		30	3.49	1.71	
□	HG00102.vcf	2 = Female		63	3.91	1.88	
□	HG00105.vcf	2 = Female	2 = diseased	32	4.77	1.58	2 = 2
□	HG00107.vcf	2 = Female	2 = diseased	48	3.7	1.6	3 = 3
□	HG00108.vcf	2 = Female		42	1.84	1.67	5 = 5 heavil
□	HG00109.vcf	2 = Female		57	1.93	1.64	
□	HG00110.vcf	2 = Female		41	4.47	1.83	
□	HG00112.vcf	2 = Female	2 = diseased	33	5.41	1.55	2 = 2
□	HG00113.vcf	2 = Female		53	4.06	1.7	5 = 5 heavil
□	HG00114.vcf	2 = Female					
□	HG00115.vcf	2 = Female	2 = diseased	47	4.33	1.62	2 = 2

2. Enter the subject ID in the ***SUBJECT Filter*** field.



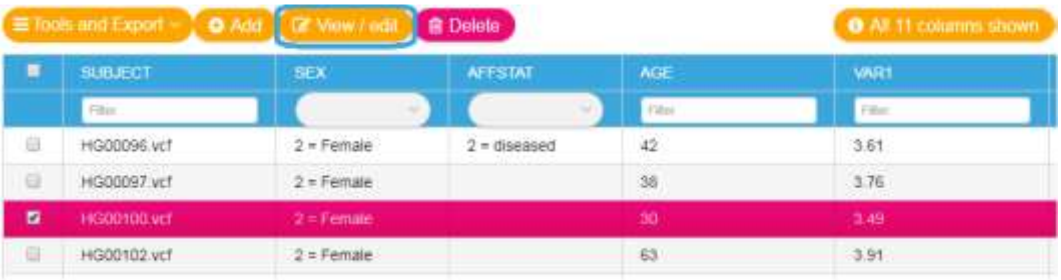
-OR-

Scroll through the list to find the subject you want to update.



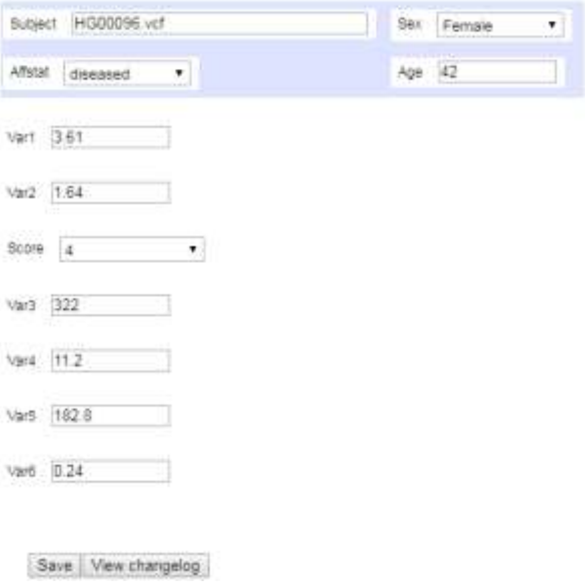
■	SUBJECT	SEX	AFFSTAT	AGE	VAR1
	|			Filter	Filter
☐	HG00096.vcf	2 = Female	2 = diseased	42	3.61
☐	HG00097.vcf	2 = Female		38	3.76
☐	HG00100.vcf	2 = Female		30	3.49
☐	HG00102.vcf	2 = Female		63	3.91
☐	HG00105.vcf	2 = Female	2 = diseased	32	4.77
☐	HG00107.vcf	2 = Female	2 = diseased	48	3.7
☐	HG00108.vcf	2 = Female		42	1.84

3. Select the required subject check box. The ***View > Edit*** button displays.



■	SUBJECT	SEX	AFFSTAT	AGE	VAR1
	Filter			Filter	Filter
☐	HG00096.vcf	2 = Female	2 = diseased	42	3.61
☐	HG00097.vcf	2 = Female		38	3.76
☑	HG00100.vcf	2 = Female		30	3.49
☐	HG00102.vcf	2 = Female		63	3.91

4. Select the ***View > Edit*** button to display the subject's details.



Subject HG00096.vcf

Sex Female ▾

Affstat diseased ▾

Age 42

Var1 3.61

Var2 1.64

Score 4 ▾

Var3 322

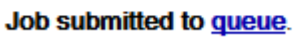
Var4 11.2

Var5 182.8

Var6 0.24

5. Make the required changes.
6. Select ***Save***.

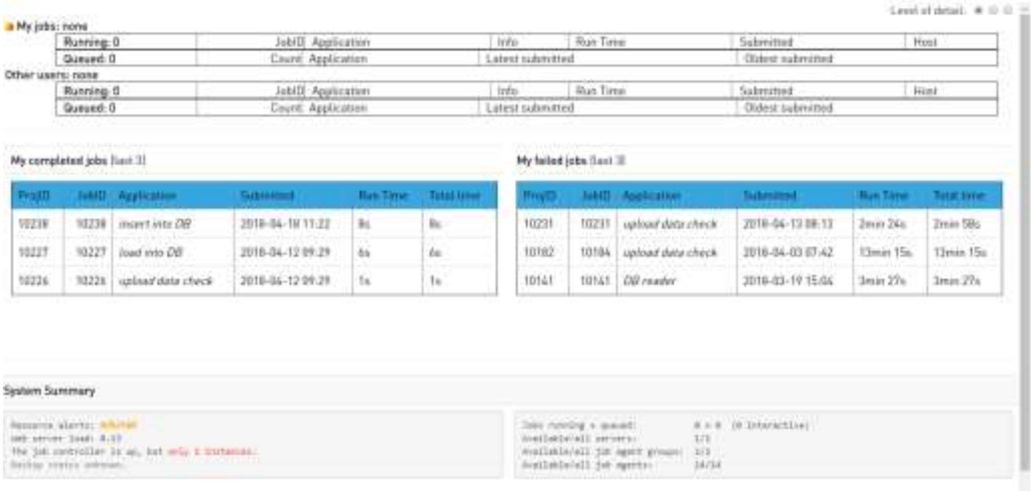
   The job is submitted to a queue for saving and a message displays.



Job submitted to queue.

7. Select ***queue***.

   A screen displays where you can follow the progress of the job in the BC|GENOME queue system.



My jobs: none

Running: 0	JobID Application	Info	Run Time	Submitted	Host
Queued: 0	Count Application	Latest submitted		Oldest submitted	

Other users: none

Running: 0	JobID Application	Info	Run Time	Submitted	Host
Queued: 0	Count Application	Latest submitted		Oldest submitted	

My completed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10238	10238	*insert into DB*	2018-04-18 11:22	8s	8s
10227	10227	*load into DB*	2018-04-12 09:29	6s	6s
10226	10226	*upload data check*	2018-04-12 09:29	1s	1s

My failed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10231	10231	*upload data check*	2018-04-13 08:13	2min 24s	2min 58s
10182	10184	*upload data check*	2018-04-03 07:42	13min 15s	13min 15s
10141	10141	*DB reader*	2018-03-19 15:04	3min 27s	3min 27s

System Summary

```
Resource alerts: [illegible]
Web server load: 4.17
The job controller is up, but only 1 instances.
Backup status unknown.
```

```
Jobs running + queued:          0 + 0  (0 interactive)
Available/all servers:          1/1
Available/all job agent groups: 1/1
Available/all job agents:       14/14
```

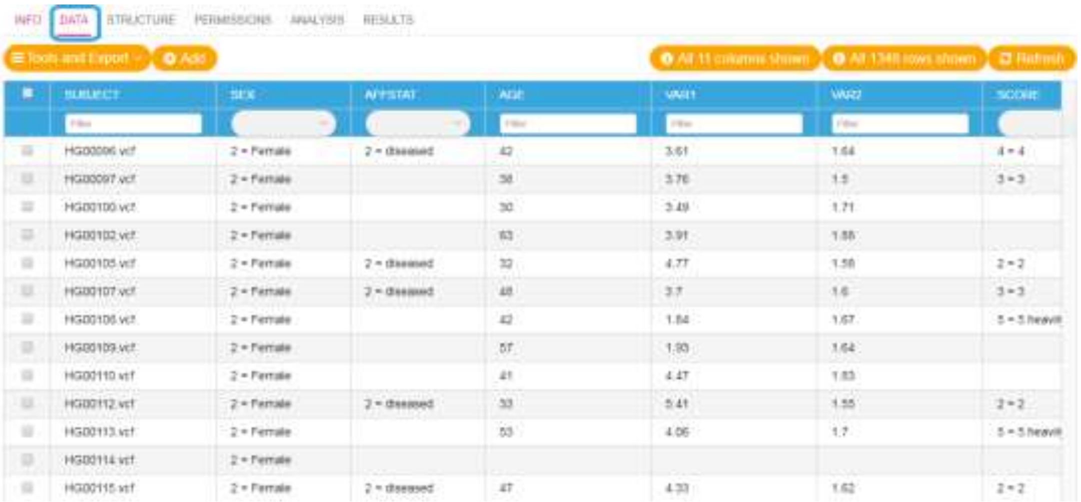
For more information about viewing results and reports, see chapter 3.3 Viewing results and reports.

3.2.2. Adding a new subject to the dataset

A new subject is added to the database when required. This task is most often performed by clinical research nurses when interviewing a subject.

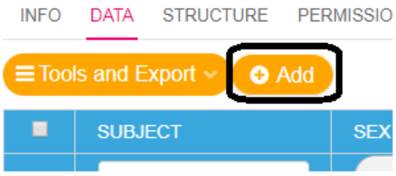
To add a new subject:

1. Select the ***DATA*** tab to open the DATA page.

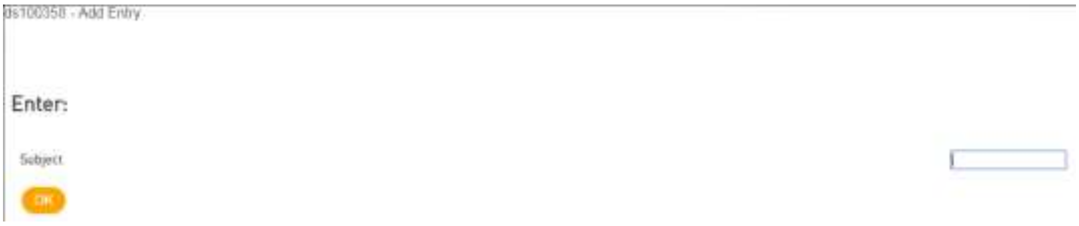


■	SUBJECT	SEX	AFFSTAT	AGE	VAR1	VAR2	SCORE
□	HG00096.vcf	2 = Female	2 = diseased	42	3.61	1.64	4 = 4
□	HG00097.vcf	2 = Female		38	3.76	1.5	3 = 3
□	HG00100.vcf	2 = Female		30	3.49	1.71	
□	HG00102.vcf	2 = Female		63	3.91	1.88	
□	HG00105.vcf	2 = Female	2 = diseased	32	4.77	1.58	2 = 2
□	HG00107.vcf	2 = Female	2 = diseased	48	3.7	1.6	3 = 3
□	HG00108.vcf	2 = Female		42	1.84	1.67	5 = 5 heavit
□	HG00109.vcf	2 = Female		57	1.93	1.64	
□	HG00110.vcf	2 = Female		41	4.47	1.83	
□	HG00112.vcf	2 = Female	2 = diseased	33	5.41	1.55	2 = 2
□	HG00113.vcf	2 = Female		53	4.06	1.7	5 = 5 heavit
□	HG00114.vcf	2 = Female					
□	HG00115.vcf	2 = Female	2 = diseased	47	4.33	1.62	2 = 2

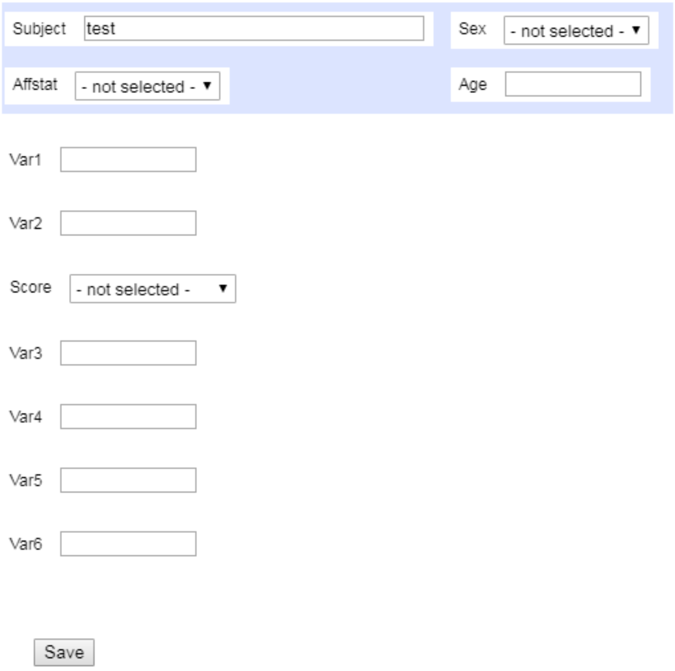
2. Select the ***Add*** button.



3. Enter an ID for the subject.



4. Click ***OK***.
5. Fill in the subject details.



Subject test

Sex - not selected -

Affstat - not selected -

Age

Var1

Var2

Score - not selected -

Var3

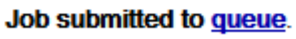
Var4

Var5

Var6

6. Select ***Save***.

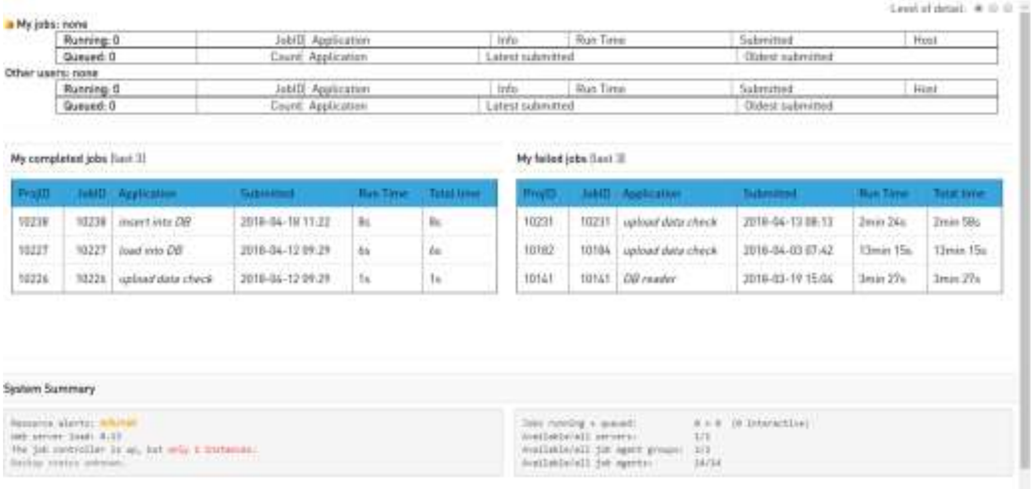
The job is submitted to a queue for saving and a message displays.



Job submitted to queue.

7. Select ***queue***.

   A screen displays where you can follow the progress of the job in the BC|GENOME queue system.



Level of detail:

My jobs: none

Running: 0	JobID Application	Info	Run Time	Submitted	Host
Queued: 0	Count Application	Latest submitted		Oldest submitted	

Other users: none

Running: 0	JobID Application	Info	Run Time	Submitted	Host
Queued: 0	Count Application	Latest submitted		Oldest submitted	

My completed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10238	10238	*insert into DB*	2018-04-18 11:22	8s	8s
10227	10227	*load into DB*	2018-04-12 09:29	6s	6s
10226	10226	*upload data check*	2018-04-12 09:29	1s	1s

My failed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10231	10231	*upload data check*	2018-04-13 08:13	2min 24s	2min 58s
10182	10184	*upload data check*	2018-04-03 07:42	13min 15s	13min 15s
10141	10141	*DB reader*	2018-03-19 15:06	3min 27s	3min 27s

System Summary

```
Resource alerts: [illegible]
Web server load: 4.13
The job controller is up, but only 1 instances.
Backup status unknown.
```

```
Jobs running + queued:            0 + 0  (0 interactive)
Available/all servers:            1/1
Available/all job agent groups:   1/1
Available/all job agents:         14/14
```

For more information about viewing results and reports, see chapter 3.3 Viewing results and reports.

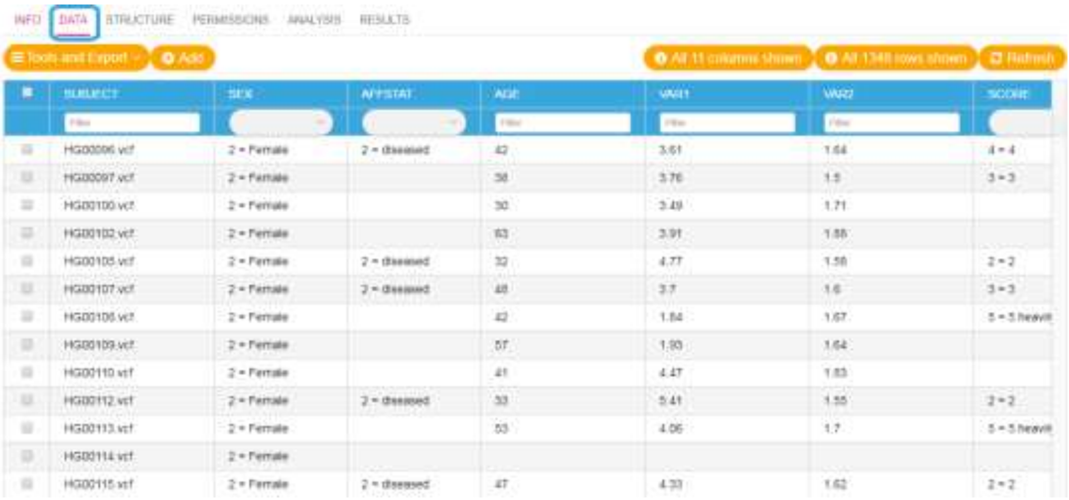
3.2.3. Uploading a single file

The ***Data Input/one file*** function is used when data is imported to a database from a file. By default BC|GENOME accepts tabulator delimited text files with headers matching the field names (in other words, variable IDs) of the dataset, but with the data import wizard tool you can import other kinds of text files.

NOTE: Check for any duplicate key variables before you upload an input file. If there are duplicate key variables in the input file, only the values from the last row of duplicates are inserted to a dataset.

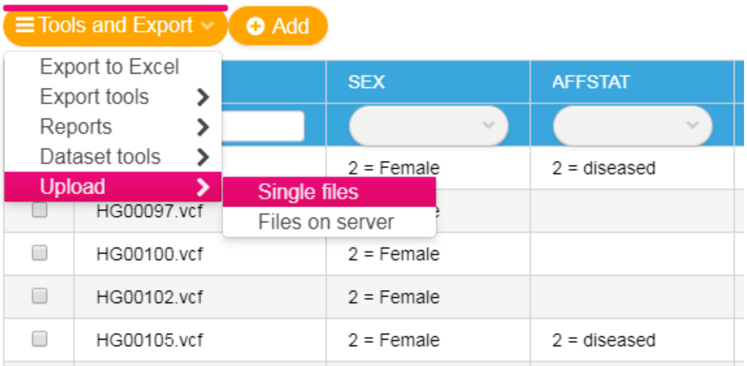
To upload a single file:

1. Select the ***DATA*** tab to open the DATA page.



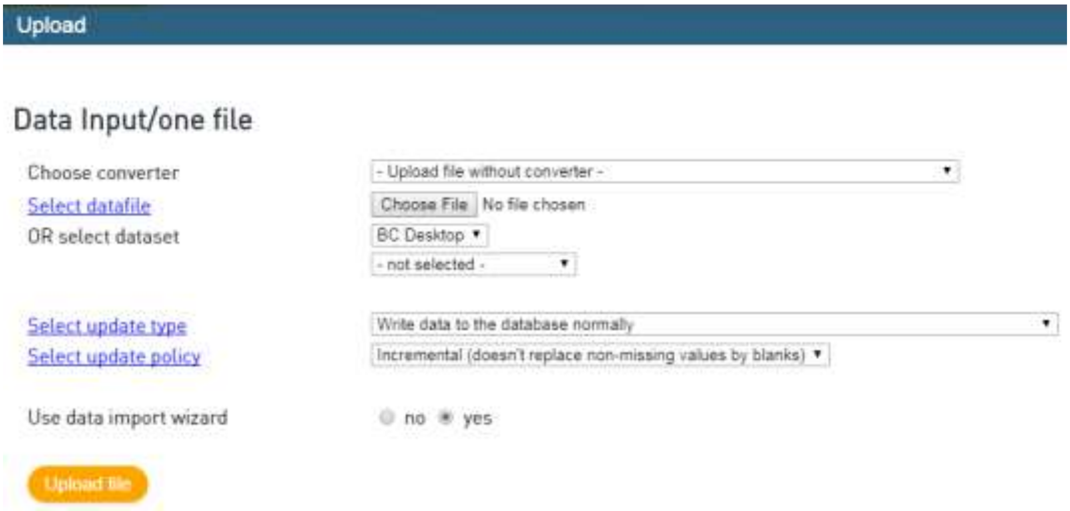
■	SUBJECT	SEX	AFFSTAT	AGE	VAR1	VAR2	SCORE
	Filter			Filter	Filter	Filter	
□	HG00096.vcf	2 = Female	2 = diseased	42	3.61	1.64	4 = 4
□	HG00097.vcf	2 = Female		38	3.76	1.5	3 = 3
□	HG00100.vcf	2 = Female		30	3.49	1.71	
□	HG00102.vcf	2 = Female		63	3.91	1.88	
□	HG00105.vcf	2 = Female	2 = diseased	32	4.77	1.58	2 = 2
□	HG00107.vcf	2 = Female	2 = diseased	48	3.7	1.6	3 = 3
□	HG00108.vcf	2 = Female		42	1.84	1.67	5 = 5 heavil
□	HG00109.vcf	2 = Female		57	1.93	1.64	
□	HG00110.vcf	2 = Female		41	4.47	1.83	
□	HG00112.vcf	2 = Female	2 = diseased	33	5.41	1.55	2 = 2
□	HG00113.vcf	2 = Female		53	4.06	1.7	5 = 5 heavil
□	HG00114.vcf	2 = Female					
□	HG00115.vcf	2 = Female	2 = diseased	47	4.33	1.62	2 = 2

2. Select ***Tools and Export > Upload > Single files***.



		SEX	AFFSTAT
		2 = Female	2 = diseased
□	HG00097.vcf		
□	HG00100.vcf	2 = Female	
□	HG00102.vcf	2 = Female	
□	HG00105.vcf	2 = Female	2 = diseased

The ***Data Input/one file*** dialog opens.



Data Input/one file

Choose converter: - Upload file without converter -

Select datafile: Choose File No file chosen

OR select dataset: BC Desktop; - not selected -

Select update type: Write data to the database normally

Select update policy: Incremental (doesn't replace non-missing values by blanks)

Use data import wizard: ○ no ● yes

3. Choose a converter (if required).

NOTE: Choose a converter based on the input file format: Variation data always requires a converter specification. However, in the case of other types of dataset (in other words, annotation, pedigrees, phenotype, multiQTL, pedigrees), only choose a converter when the input file is not in BC format as described here.

-OR-

Select a dataset with the same form to copy dataset information to a new dataset.

4. ***Select update type*** to specify how overlapping information is treated:

   a) Select ***Write data to the database normally*** to upload new rows to a dataset.

      The overlapping key value(s) information in the input file replaces the existing information and is reported in the ***RESULTS >*** results folder.

   b) Select Do NOT write data to the database, but generate report which values update would change to run a test report.

      The database values are not updated but the test report in ***RESULTS >*** result archive shows the projected changes between the original values and the new values.

c) Delete rows defined in the file for removing data rows that have duplicate key value(s) with existing data. By default, this action is only allowed to be done by the dataset owner, but BC Support can also grant this permission to database administrators.

5. ***Select update policy*** to specify the management of empty values.

   a) Select ***Incremental (doesn't replace non-missing values by blanks)*** if you do not want empty values in the input files to replace the existing data values if there are overlapping key and variable value(s).

   d) Select ***Direct (allows to replace non-missing values by blanks)*** if you want empty values in the input files to replace the existing data values if there are overlapping key and variable value(s).

NOTE: All values are entered during the first upload.

6. Select ***Use data import wizard***

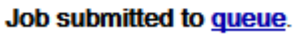
   a) Select ***yes*** to check the match between your input file and dataset form.

   b) Select ***no*** to skip this step.

NOTE: If you select any converter other than ***Upload file without converter***, the upload data import wizard is not in use.

7. Select ***Upload file***.

   If you selected ***no*** to ***Use data import wizard***, the job is submitted to a queue for saving and a message displays.



Job submitted to queue.

NOTE: If you selected ***yes*** to ***Use data import wizard***, go to section *8* for further information about selecting the upload options using the data import wizard.

8. Select ***queue***.

   A screen displays where you can follow the progress of the job in the BC|GENOME queue system.

For more information about viewing results and reports, see chapter 3.3 Viewing results and reports.

3.2.4. Uploading a file using the upload wizard

The ***Data Input/one file*** tool as described in section *3.2.3* supports the use of a data import wizard tool when the input file used in the upload does not have an exact match if any of the following are in the form:

- field names
- options of multiple choice questions
- missing values that are specified with other values instead of "null"

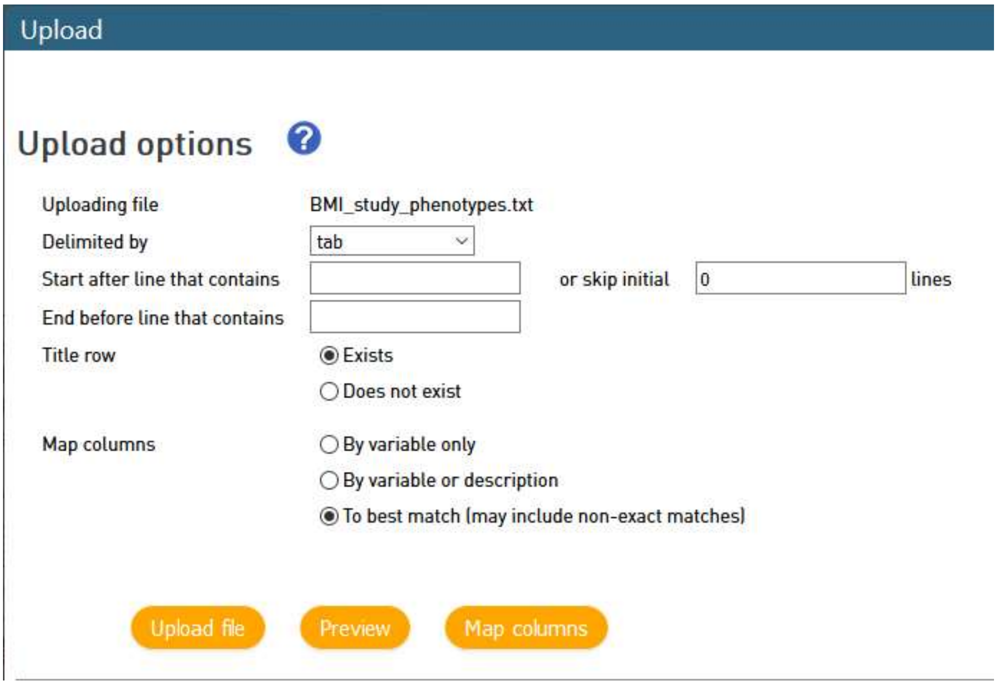
- non-database format date values (YYYY-MM-DD)

NOTE: The data import wizard can be used to upload a text file to a dataset of type: phenotype, annotation, sample IDs or pedigree.

To select the upload options using the upload wizard:

1. Follow the instructions to upload a single file as described in section 3.2.3.
2. Select ***yes*** in ***Use data import wizard***.
3. Select ***Upload file****.*

   The ***Upload options*** dialog is shown.



Upload

Upload options

Uploading file: BMI_study_phenotypes.txt

Delimited by: tab

Start after line that contains: or skip initial 0 lines

End before line that contains:

Title row: ● Exists ○ Does not exist

Map columns: ○ By variable only ○ By variable or description ● To best match (may include non-exact matches)

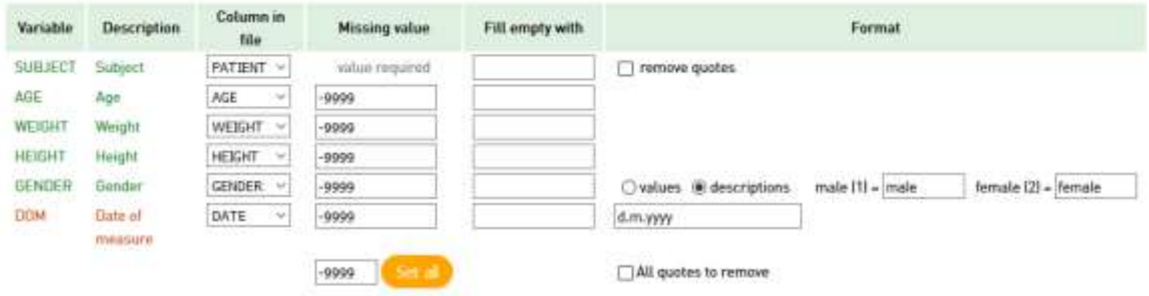
NOTE: The ***Upload options*** dialog contains two sets of buttons for ***Upload file***, ***Preview*** and ***Map columns***. These perform the same function, but are included twice to make it easier for the user to select them as the dialog is long.

4. Select a file delimiter from the drop-down list in the ***Delimited by*** field:
   - tab
   - comma
   - semicolon
   - space
   - multiple spaces
   - space aligned
5. Exclude rows either at the top or end of a file.

6. Select whether or not an input file has a ***Title row***:
   - Exists
   - Does not exist
7. Select how to map the columns in the ***Map columns*** field:
   - By variable only
   - By variable or description
   - To best match (may include non-exact matches)
8. Select ***Preview*** to check the variable matches between the input file and the form variables.
9. Select ***Upload file*** if you do not need to make any changes to the mapping.
10. To make changes to the mapping, see section 3.2.4.1.

3.2.4.1. Mapping variables

If you need to make changes to the mapping, use the columns shown in Figure 8. The first two columns, ***Variable*** and ***Description*** show the information that is in the form that was used to create the dataset. The remaining columns are editable and can be used to map the variables in the input file to the variables in the dataset form.



Variable	Description	Column in file	Missing value	Fill empty with	Format
SUBJECT	Subject	PATIENT	value required		□ remove quotes
AGE	Age	AGE	-9999		
WEIGHT	Weight	WEIGHT	-9999		
HEIGHT	Height	HEIGHT	-9999		
GENDER	Gender	GENDER	-9999		○ values ● descriptions male [1] = male female [2] = female
DOM	Date of measure	DATE	-9999		d.m.yyyy
			-9999 Set all		□ All quotes to remove

Figure 8. Mapping variables

1. In ***Column in file***, select variables from the input file to map to the variables in the dataset form.

   a) Green indicates a perfect match – the input file and the dataset form have exactly the same variable name.

   e) Orange needs a manual check – the dataset form and the input file have a similar variable but it is named differently. For example, DOM is the name of the variable indicating the date in the dataset form and it is named DATE in the input file. They are both the same variable but named differently, so can be mapped.

   f) Blank indicates there is no match – the same variable does not exist in the input file.

2. In ***Missing value***, type a value for a variable if required.

   If you want to apply the same value to all variables, type a value in the ***Select all*** field.

3. Select ***remove quotes*** if there are any quotation marks in the input file.

4. For a DATE variable, type the complete date format to match your data (default is yyyy-mm-dd):

   a) d.m.yyyy -> 25.12.1987

   b) d/m/yyyy -> 25/11/1999

   c) m/d/yyyy -> 11/25/2001

   d) m-d-yyyy -> 11-25-2001

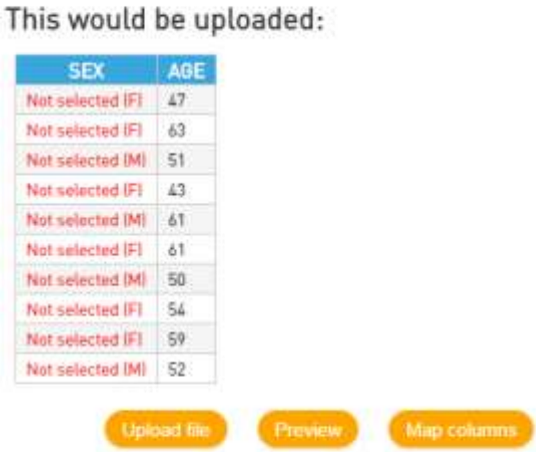
NOTE:	There must only be one type of date format in a single input file to map properly.

5. Select ***values*** or ***descriptions*** for multiple choice questions in the dataset form to match those in the input file.

   If you select ***descriptions***, you can type text in the description fields.

6. Select ***Preview*** to check the data values that will be imported.

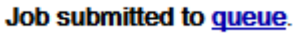
   The values that will not be imported are shown in red or the column is missing in the preview.



This would be uploaded:

SEX	AGE
Not selected (F)	47
Not selected (F)	63
Not selected (M)	51
Not selected (F)	43
Not selected (M)	61
Not selected (F)	61
Not selected (M)	50
Not selected (F)	54
Not selected (F)	59
Not selected (M)	52

7. Select ***Upload file*** when you are ready to proceed and a message displays.



Job submitted to queue.

NOTE:	If you selected ***yes*** to ***Use data import wizard***, go to section *8* for further information about selecting the upload options using the data import wizard.

8. Select ***queue***.

   A screen displays where you can follow the progress of the job in the BC|GENOME queue system.

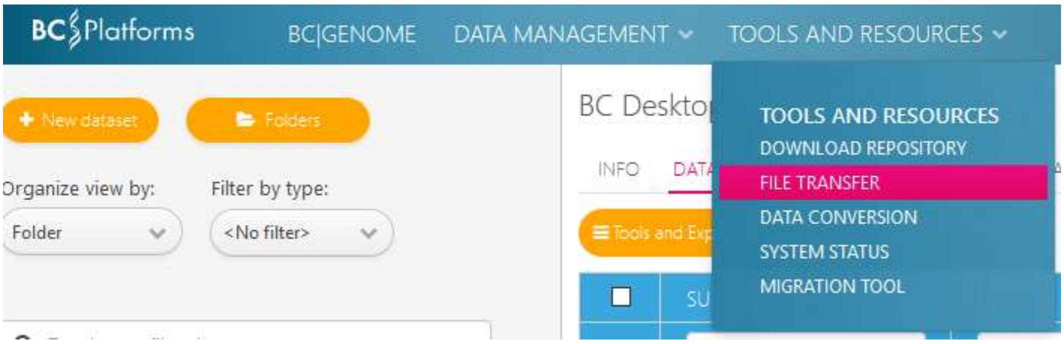
For more information about viewing results and reports, see chapter 3.3 Viewing results and reports.

3.2.5. Uploading files from a server

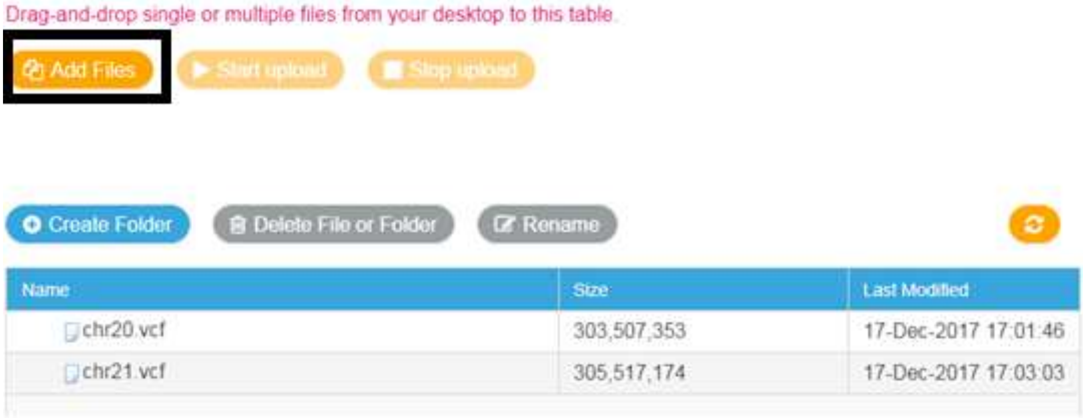
Many files can be uploaded to the dataset simultaneously using the ***Data Input/large files*** function. Before you start the upload procedure, you must copy the data files you want to upload to a server (usually in a folder on the server).

To add files to a server:

1. Select ***TOOLS AND RESOURCES > FILE TRANSER***.



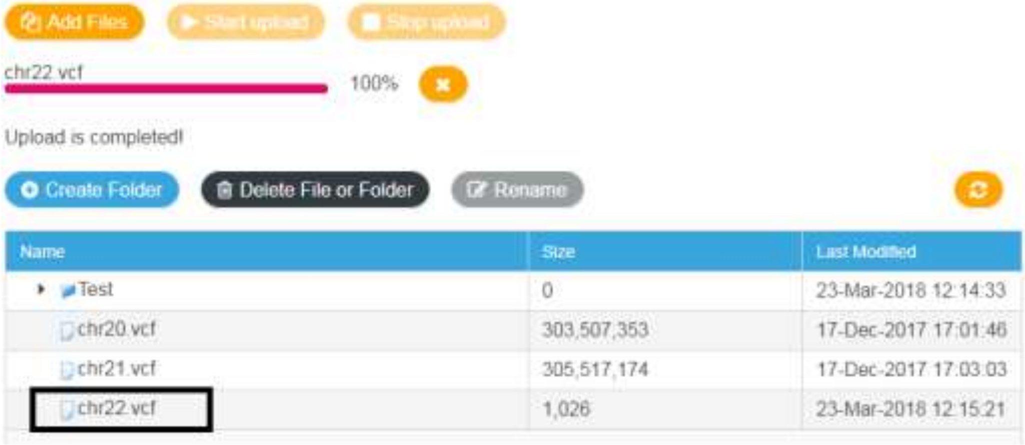
2. Select ***Add Files***.


Drag-and-drop single or multiple files from your desktop to this table.

Name	Size	Last Modified
chr20.vcf	303,507,353	17-Dec-2017 17:01:46
chr21.vcf	305,517,174	17-Dec-2017 17:03:03

3. Browse to the file(s) you want to add and select ***Open***.

   The chosen file(s) shows in the table.



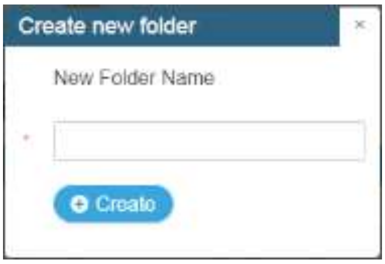
chr22.vcf 100%

Upload is completed!

Name	Size	Last Modified
Test	0	23-Mar-2018 12:14:33
chr20.vcf	303,507,353	17-Dec-2017 17:01:46
chr21.vcf	305,517,174	17-Dec-2017 17:03:03
chr22.vcf	1,026	23-Mar-2018 12:15:21

If you want to add files to a folder, you can use an existing folder or create a new folder.

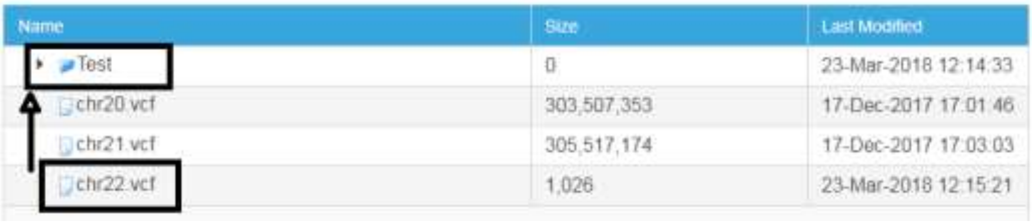
1. Select ***Create Folder*** to add a new folder.



2. Type a name for the new folder.
3. Select ***Create***.

   The new folder shows in the table

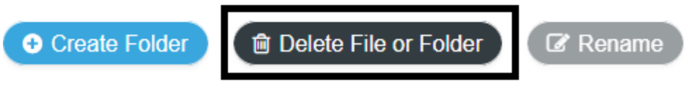
4. Drag and drop the file(s) from the table to the required folder in the table (either the newly created folder or an already existing folder).



Name	Size	Last Modified
Test	0	23-Mar-2018 12:14:33
chr20.vcf	303,507,353	17-Dec-2017 17:01:46
chr21.vcf	305,517,174	17-Dec-2017 17:03:03
chr22.vcf	1,026	23-Mar-2018 12:15:21

If you want to delete a file/folder:

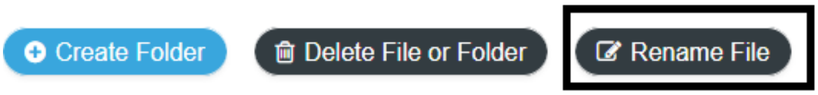
1. Select the file/folder in the table.
2. Select ***Delete File or Folder*** to delete a file or folder.



You can only delete one file at a time. You cannot delete a folder until the folder is empty.

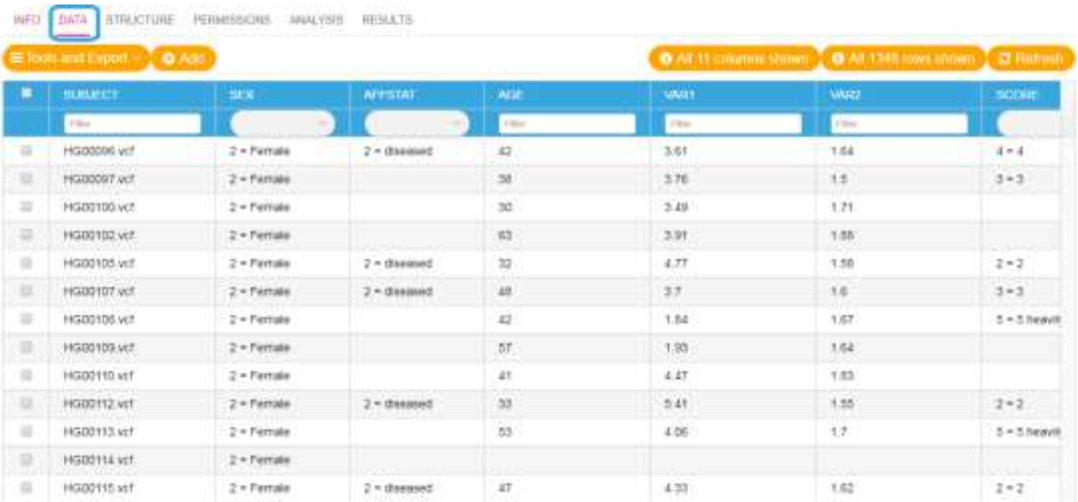
If you want to rename a file or folder:

1. Select the file or folder in the table.
2. Select ***Rename File*** or ***Folder*** to delete a file or folder.



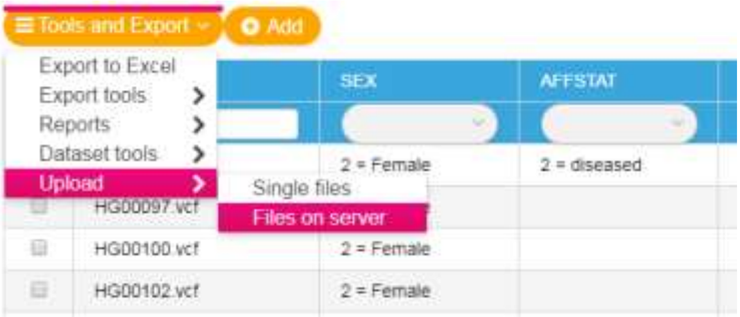
To upload files from a server:

1. Select the ***DATA*** tab to open the DATA page.

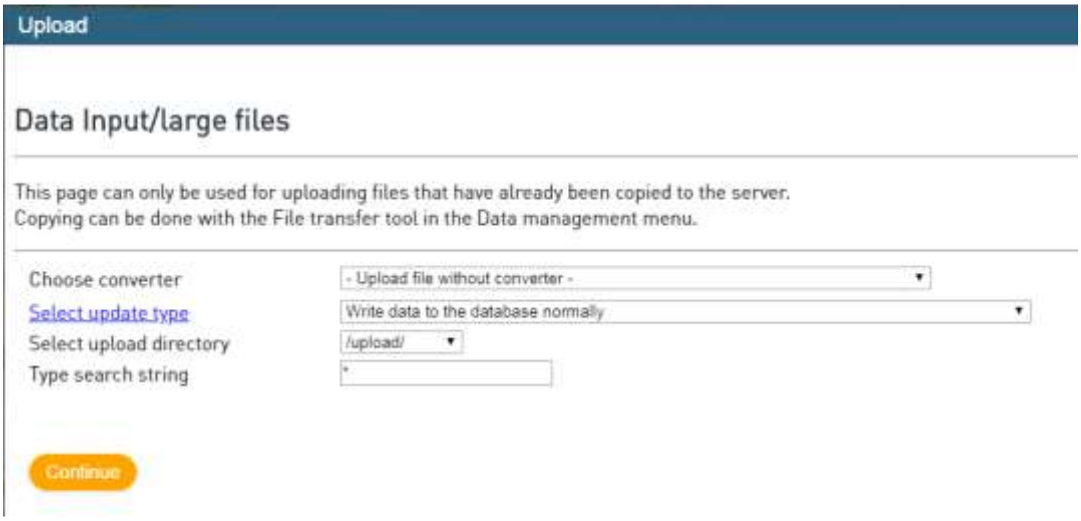


■	SUBJECT	SEX	AFFSTAT	AGE	VAR1	VAR2	SCORE
	Filter			Filter	Filter	Filter	
□	HG00096.vcf	2 = Female	2 = diseased	42	3.61	1.64	4 = 4
□	HG00097.vcf	2 = Female		38	3.76	1.5	3 = 3
□	HG00100.vcf	2 = Female		30	3.49	1.71	
□	HG00102.vcf	2 = Female		63	3.91	1.88	
□	HG00105.vcf	2 = Female	2 = diseased	32	4.77	1.58	2 = 2
□	HG00107.vcf	2 = Female	2 = diseased	48	3.7	1.6	3 = 3
□	HG00108.vcf	2 = Female		42	1.84	1.67	5 = 5 heavil
□	HG00109.vcf	2 = Female		57	1.93	1.64	
□	HG00110.vcf	2 = Female		41	4.47	1.83	
□	HG00112.vcf	2 = Female	2 = diseased	33	5.41	1.55	2 = 2
□	HG00113.vcf	2 = Female		53	4.06	1.7	5 = 5 heavil
□	HG00114.vcf	2 = Female					
□	HG00115.vcf	2 = Female	2 = diseased	47	4.33	1.62	2 = 2

2. Select ***Tools and Export > Upload > Files on server***.



The ***Data Input/large files*** dialog opens.



Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server.
Copying can be done with the File transfer tool in the Data management menu.

Choose converter	- Upload file without converter - ▾
Select update type	Write data to the database normally ▾
Select upload directory	/upload/ ▾
Type search string	*

3. Choose a converter (if required).

NOTE:	Choose a converter based on the input file format: Variation data always requires a converter specification. However, in the case of other types of dataset (in other words, annotation, pedigrees, phenotype, multiQTL, pedigrees), only choose a converter when the input file is not in BC format as described in section 3.2.3. If BC|GENOME uses a special reference file to find files stored in a dataset, you must choose a converter. By default upload file(s) without any converter in the case of phenotype data

4. ***Select update type*** to specify how overlapping information is treated:

   a) Select ***Write data to the database normally*** to upload new rows to a dataset.

   The overlapping key value(s) information in the input file replaces the existing information and is reported in the ***RESULTS >*** results folder.

   g) Select Do NOT write data to the database, but generate report which values update would change to run a test report.

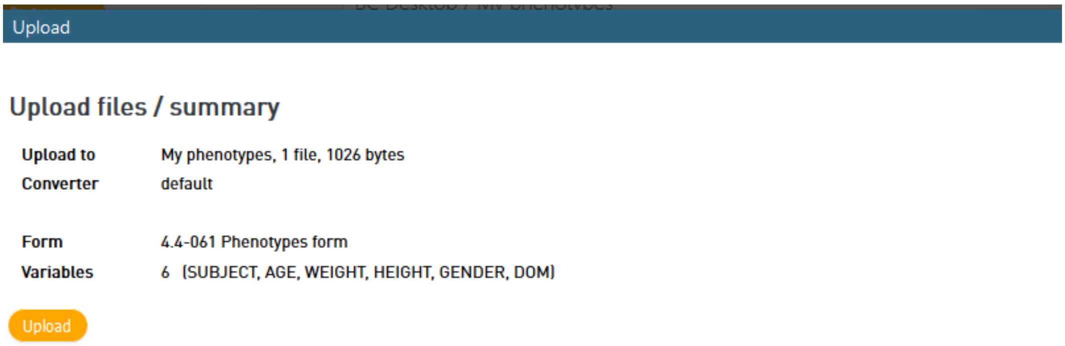
   The database values are not updated but the test report in ***RESULTS >*** result archive shows the projected changes between the original values and the new values.

   h) Delete rows defined in the file for removing data rows that have duplicate key value(s) with existing data. By default, this action is only allowed to be done by the dataset owner, but BC Support can also grant this permission to database administrators.

5. In ***Select upload directory*** to specify where you have your transferred file(s).
6. In ***Type search string***, type a string to match the files you want to upload.

   – If you type a ***** or leave the field blank, all files are shown.
   – To narrow the search, type, for example, ***b**** to show all file names that start with the letter ***b***.

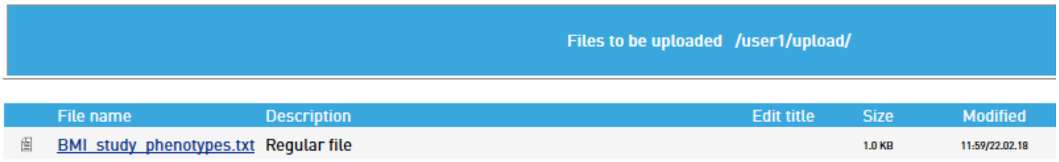
7. Select ***Continue*** and make sure that the files listed are those you want to upload.



Upload

Upload files / summary

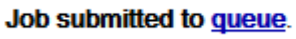
Upload to	My phenotypes, 1 file, 1026 bytes
Converter	default
Form	4.4-061 Phenotypes form
Variables	6 (SUBJECT, AGE, WEIGHT, HEIGHT, GENDER, DOM)



Files to be uploaded /user1/upload/

File name	Description	Edit title	Size	Modified
BMI_study_phenotypes.txt	Regular file		1.0 KB	11:59/22.02.18

Select ***Upload*** and a message displays.



Job submitted to queue.

8. Select ***queue***.

   A screen displays where you can follow the progress of the job in the BC|GENOME queue system.

NOTE:	If you selected ***yes*** to ***Use data import wizard***, go to section *8* for further information about selecting the upload options using the data import wizard.

For more information about viewing results and reports, see chapter 3.3 Viewing results and reports.

3.2.6. Reverting of accidental changes

In BC|GENOME there is no direct 'Undo' or rollback method to revert erroneous changes. If you find you have accidentally changed data and you wish to revert the change, you have two options to go about this, depending on the scale of change.

3.2.6.1. Undo data entry change of a single data row

When you find there is a need to revert changes in a single data entry row of the dataset stored in the database table proceed with the following steps:

1. Select a dataset from the dataset navigator
2. Navigate to INFO of your dataset and browse for the section of Latest data updates.
3. Click the source link open to verify the changes based on jobID and timestamp
4. In the DATA tab select the data row needing editing
5. Use the View / edit button to edit values
6. Save the changes using the Save button

3.2.6.2. Revoking multiple data entry rows in a dataset

If you have accidentally uploaded wrong data into your dataset, contact support@bcplatforms.com for their support of reverting the dataset into its original stage. Please indicate your user name, dataset name and jobID found in DATA MANAGEMENT / RESULT ARCHIVE. BC|GENOME stores full change log in the database for datasets, which is used to roll back changes. The retraction of changes will also become visible in the change log.

3.3. Viewing results and reports

BC|GENOME delivers reports about data import and editing tasks, as well as the analysis results. These reports can be found in the result archive. The queue system facilitates many users running different analyses and data management tasks simultaneously.

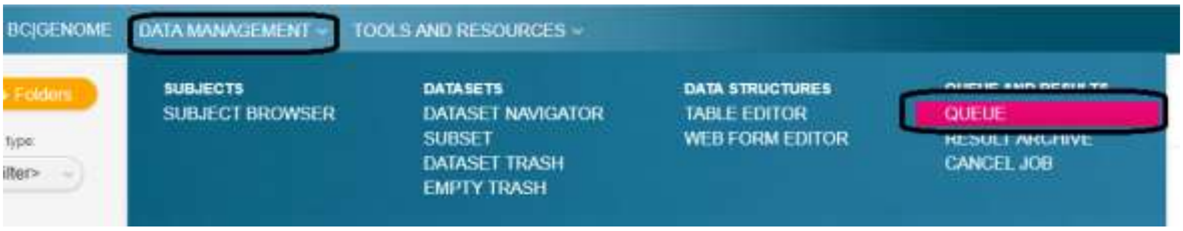
3.3.1. Queue system

Most data management jobs are handled by the queue system. The status of each job can be seen in the ***QUEUE*** page and more detailed information in the ***SYSTEM STATUS*** page.

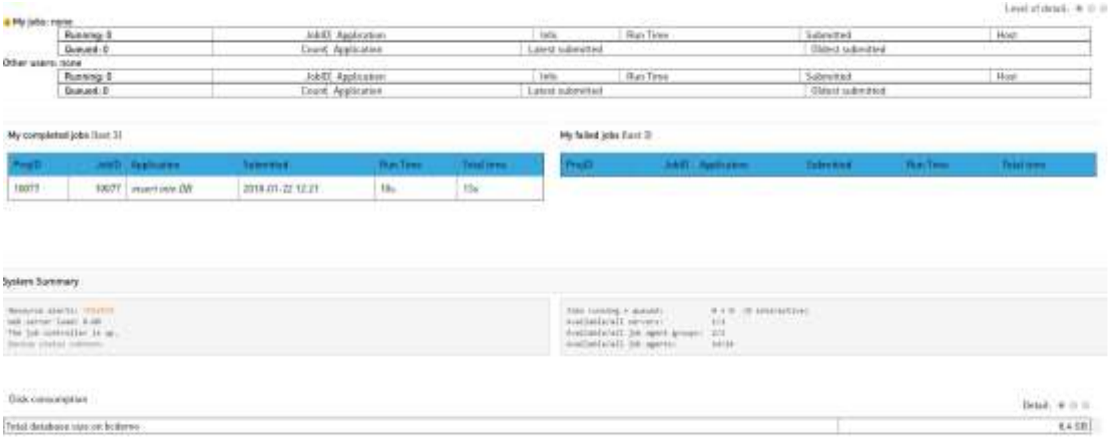
Jobs in a queue are prioritised by the number of jobs: the more jobs users submit to the queue, the lower priority their jobs are given, allowing other users with fewer jobs to occupy the queue. When a job disappears from the queue, the results report can be found in ***RESULTS > Result archive*** page. Jobs can be removed from the queue by using the ***DATA MANAGEMENT > CANCEL JOB*** function.

To see the queue:

- Select ***DATA MANAGEMENT > QUEUE***.

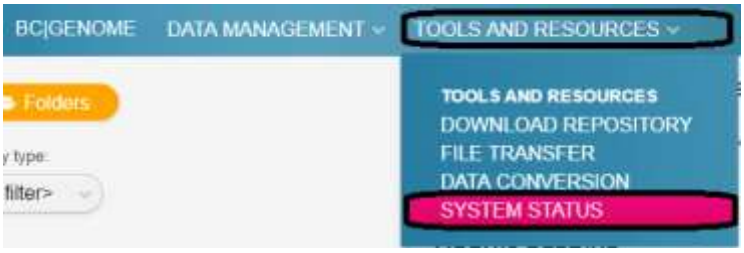


The queue displays.

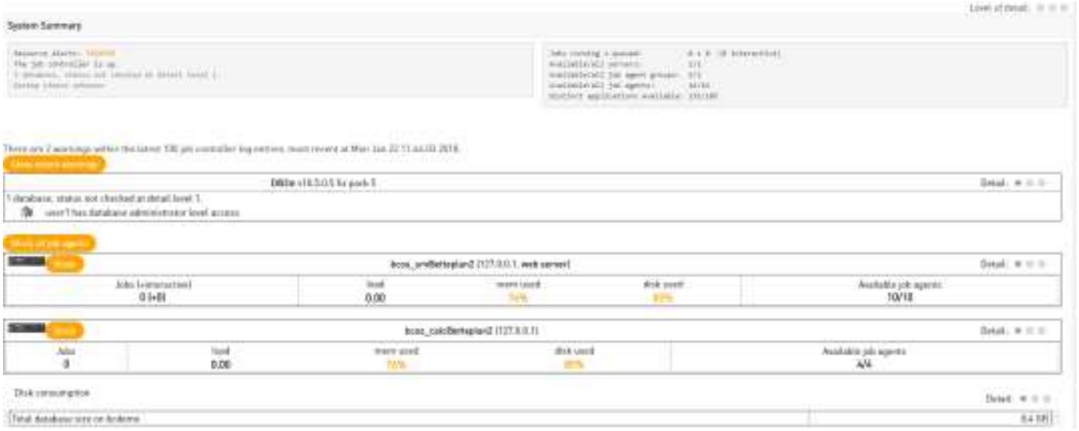


To see the system status:

- Select ***TOOLS AND RESOURCES > SYSTEM STATUS***.



The system summary displays.

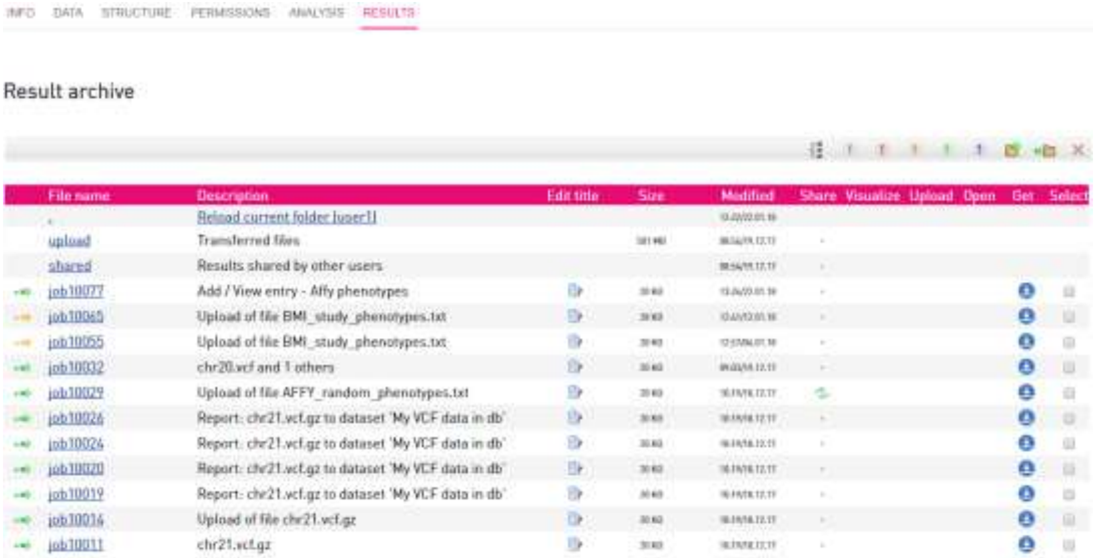


3.3.2. Result archive

The result archive contains reports on user’s data management tasks and analyses. The required files can be found based on both report titles and job ID number.

There are two ways to open the result archive:

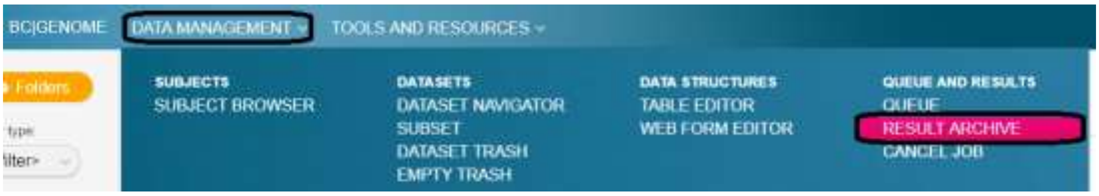
- Select the ***RESULTS*** tab to open the ***RESULTS*** page.



Result archive

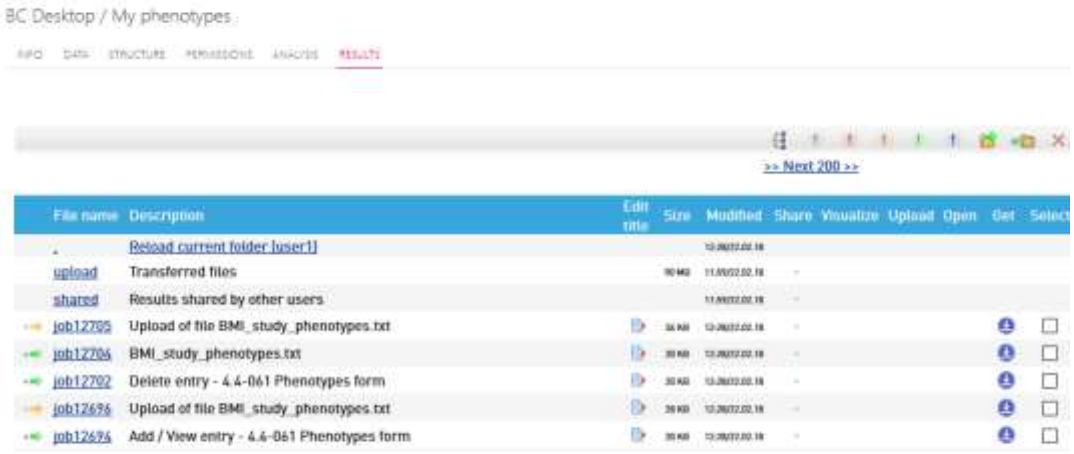
	File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
	.	Reload current folder [user1]			[illegible]						
	upload	Transferred files		381 MB	[illegible]	-					
	shared	Results shared by other users			[illegible]	-					
→	job10077	Add / View entry - Affy phenotypes		30 KB	[illegible]	-					□
→	job10065	Upload of file BMI_study_phenotypes.txt		30 KB	[illegible]	-					□
→	job10055	Upload of file BMI_study_phenotypes.txt		30 KB	[illegible]	-					□
→	job10032	chr20.vcf and 1 others		30 KB	[illegible]	-					□
→	job10029	Upload of file AFFY_random_phenotypes.txt		30 KB	[illegible]						□
→	job10026	Report: chr21.vcf.gz to dataset 'My VCF data in db'		30 KB	[illegible]	-					□
→	job10024	Report: chr21.vcf.gz to dataset 'My VCF data in db'		30 KB	[illegible]	-					□
→	job10020	Report: chr21.vcf.gz to dataset 'My VCF data in db'		30 KB	[illegible]	-					□
→	job10019	Report: chr21.vcf.gz to dataset 'My VCF data in db'		30 KB	[illegible]	-					□
→	job10016	Upload of file chr21.vcf.gz		30 KB	[illegible]	-					□
→	job10011	chr21.vcf.gz		30 KB	[illegible]	-					□

- Select ***DATA MANAGEMENT > RESULT ARCHIVE***.



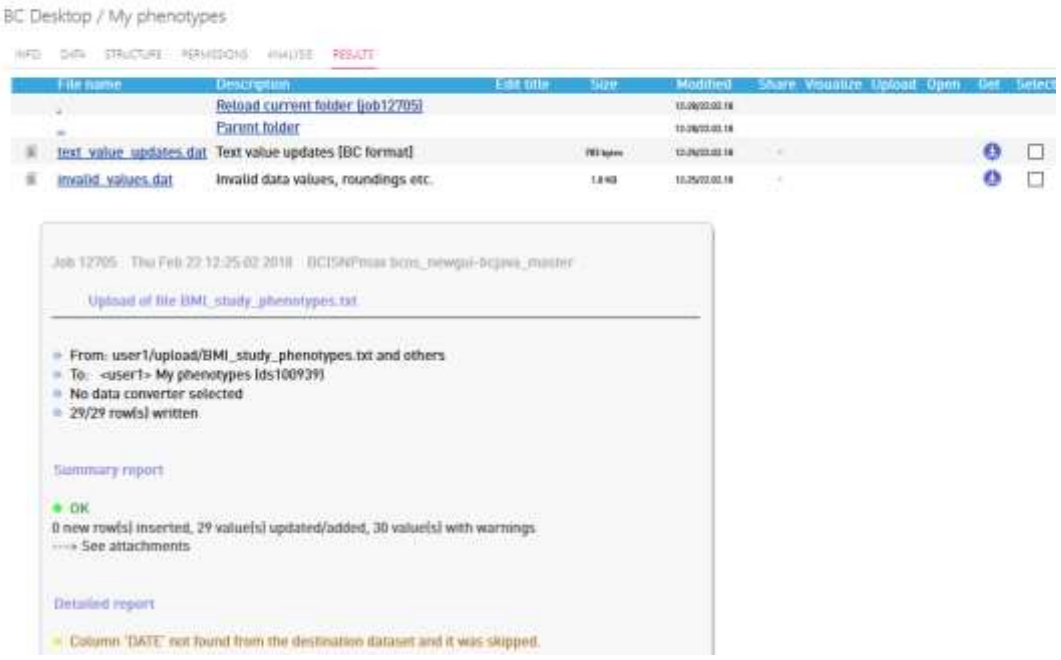
3.3.3. Check your reports in the *RESULTS* page

1. Select ***DATA MANAGEMENT > RESULT ARCHIVE*** or the ***RESULTS*** tab to open the ***RESULTS*** page.
   a) A green arrow indicates your job was completed successfully.
   b) An orange arrow indicates your job process generated additional report files for checking.
   c) A red arrow indicates that your job failed.



	File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
	.	Reload current folder [user1]			[illegible]						
	upload	Transferred files		90 MB	[illegible]	-					
	shared	Results shared by other users			[illegible]	-					
→	job12705	Upload of file BMI_study_phenotypes.txt		34 KB	[illegible]	-					□
→	job12704	BMI_study_phenotypes.txt		30 KB	[illegible]	-					□
→	job12702	Delete entry - 4.4-061 Phenotypes form		30 KB	[illegible]	-					□
→	job12696	Upload of file BMI_study_phenotypes.txt		38 KB	[illegible]	-					□
→	job12694	Add / View entry - 4.4-061 Phenotypes form		30 KB	[illegible]	-					□

2. Select the job ID to open the report file details.
3. Verify the report details and take any action if necessary.



File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
.	Reload current folder [job12705]			[illegible]						
..	Parent folder			[illegible]						
text_value_updates.dat	Text value updates [BC format]		[illegible]	[illegible]	-					□
invalid_values.dat	Invalid data values, roundings etc.		1.8 KB	[illegible]	-					□

Job 12705 Thu Feb 22 12:25:02 2018 BCISNPmax bcos_newgui-bcjava_master

Upload of file BMI_study_phenotypes.txt

- From: user1/upload/BMI_study_phenotypes.txt and others
- To: <user1> My phenotypes [ds100939]
- No data converter selected
- 29/29 row(s) written

Summary report

● OK
0 new row(s) inserted, 29 value(s) updated/added, 30 value(s) with warnings
----> See attachments

Detailed report

- Column 'DATE' not found from the destination dataset and it was skipped.

3.3.4. Generic search

Note	The search index is not kept real-time to avoid clashes with user activity and system performance. The index is updated when BC|GENOME server is idle. Therefore your search results may not always show very recently added items. The last index refresh time is shown at the bottom of the Search -tool page.

3.3.4.1. Datasets, results, and metadata

Generic SEARCH tool is opened from the obvious location at the right hand upper corner of the BC|GENOME application. You can type your search terms in the text box of the Search -tool page in any order. The search mechanism will give scores to the hits it finds based on relevance and how well the search terms match the data item.

Your hits are displayed in a grid that provides further filtering and sorting for the results. By default sorting is by hit Score and you can change this by clicking on the grid headers. Each hit displays the name of the item the search found, the category of the item, the exact text match justifying the hit, and finally the relevance score.

It is possible to navigate to the source of the hit by clicking the provided link in the hit result name. For example for dataset hits you can navigate to the dataset in question by following the provided link in the table.

3.3.4.2. Search categories

Each search hit comes from a separate index, which are categorized in following way:

Category	Description
Datasets	Datasets metadata and primary key values, and all values type of text, excluding genotype datasets. Index includes dataset name, description and other meta. Primary keys like SUBJECT field values are indexed.
Form information	Dataset form information. Form names and possible metadata.
Result	Final reports and result folders.
Help information	Online documentation and other help files (under construction)

It is possible to restrict the search hits to only selected categories, you need to run the search again to narrow down the search scope.

Note Results may produce usually low-scoring hits from hidden information stored about each job. This hidden information may include server details and other similar data included in the job metadata but not visible to the user in result page. Check the Match to see where the hit is generated.

3.3.4.3. Using wildcards and phases

The Search -tool allows the use of asterisk (*) as wildcard character. You can use it to create search terms that match only the beginning of the words, like 'cohort*' would match 'cohort' and 'cohorts'. It is not possible to use asterisk to mask the beginning of the word.

If you use multiple words or parts of words, the search tool will try to match all words in the documents (logical operator AND). This effectively narrows down your search results when you add new search terms. Remember that order of words in this case does not matter. If you want to match an exact phrase, like a name of a dataset, you should enclose the search in double quotes. See the following example scenario:

When dataset name is "Anni's many SNPs":

SNPs many # matches

"SNPs many" # will not match

"many SNPs" # matches

4. Managing forms in BC|GENOME

This chapter describes the contents of forms and how to manage them in BC|GENOME.

4.1. Variables

Variables are used to define the parts required in a form. Variables in a form can be added, removed or edited.

In BC|GENOME, each variable has a set of variable attributes assigned to it:

- ID
- Description
- Type
- Key
- Required
- Choiceset
- Max text length
- Minimum
- Maximum
- Value function
- Default value

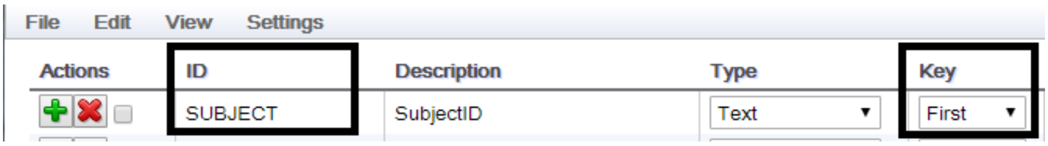
4.1.1. ID (unique identifier)

By default, when a new form is created, the default variable SUBJECT ***ID*** is edited to create a *unique identifier* for the record.

This variable must have a unique name (rename SUBJECT with typically a combination of numbers and letters) and the first (primary) ***Key*** assigned to it.

You can then edit the rest of the attributes for this variable as required.

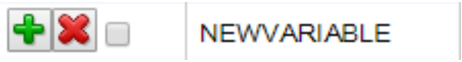
1. Double-click the ***ID*** variable field.
2. Type a unique name.
3. Select ***File > Save*** to save changes to the form.



Actions	ID	Description	Type	Key
✚ ✖ □	SUBJECT	SubjectID	Text ▼	First ▼

4.1.2. ID

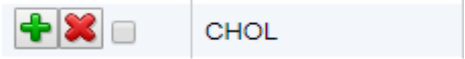
New variables are added with the generic name NEWVARIABLE.



NEWVARIABLE

1. Double-click the ***ID*** variable field.
2. Type a name for the variable.

   You can rename the variable to the whatever name you want, for example, CHOL.

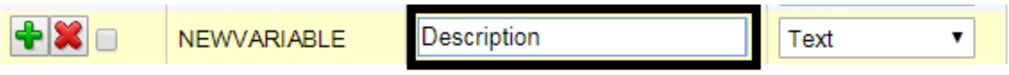
   CHOL

3. Select ***File > Save*** to save changes to the form.

4.1.3. Description

This field is used to add a brief description of the variable.

1. Double-click the ***Description*** variable field.
2. Type a description for the variable.
3. Select ***File > Save*** to save changes to the form.



NEWVARIABLE Description Text ▼

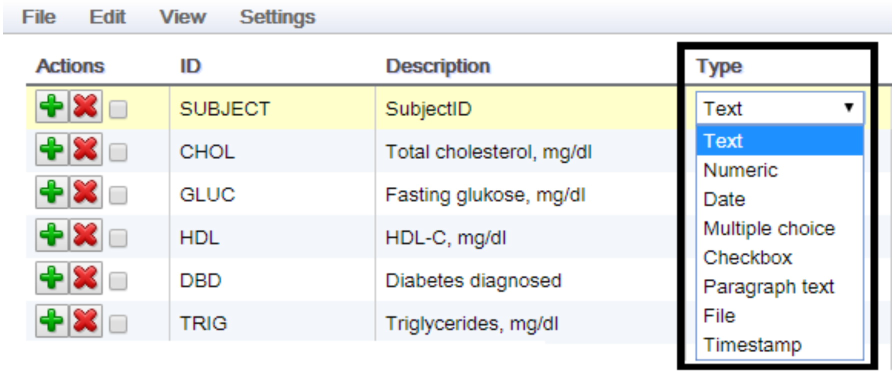
4.1.4. Type

Type defines the kind of the variable.

The following variable types are available:

- Text
- Numeric
- Date
- Multiple Choice
- Checkbox
- Paragraph text
- File
- Timestamp

1. Select the arrow in the ***Type*** variable field.
2. Choose the variable type from the drop-down menu.



Actions	ID	Description	Type
+ ✕ □	SUBJECT	SubjectID	Text ▾
+ ✕ □	CHOL	Total cholesterol, mg/dl	
+ ✕ □	GLUC	Fasting glukose, mg/dl	
+ ✕ □	HDL	HDL-C, mg/dl	
+ ✕ □	DBD	Diabetes diagnosed	
+ ✕ □	TRIG	Triglycerides, mg/dl	

3. Select ***File > Save*** to save changes to the form.

4.1.5. Key

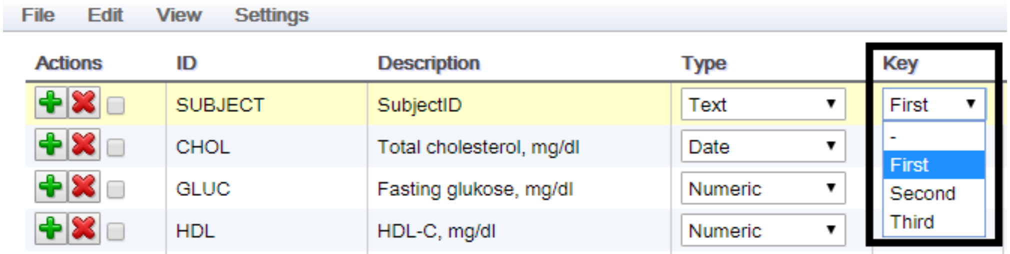
Keys define unique records in a dataset and also serve as a linking identifier across different data sections. By default, the primary key is the SUBJECT ***ID***. The second and third key can be variable ***Type***:

- text
- date
- timestamp
- multiple choice
- checkbox variable

NOTE: When needed, both multiple choice and checkbox can also be defined as a primary key.

1. Select the arrow in the ***Key*** variable field.
2. Choose the key from the drop-down menu.

   The following variable keys are available: First (primary key), Second (secondary key) and Third (tertiary key).



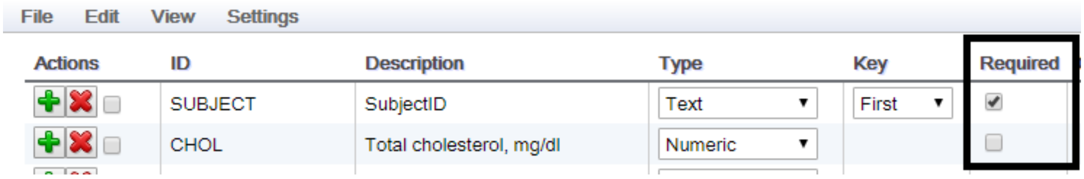
Actions	ID	Description	Type	Key
+ ✕ □	SUBJECT	SubjectID	Text ▾	First ▾
+ ✕ □	CHOL	Total cholesterol, mg/dl	Date ▾	
+ ✕ □	GLUC	Fasting glukose, mg/dl	Numeric ▾	
+ ✕ □	HDL	HDL-C, mg/dl	Numeric ▾	

3. Select ***File > Save*** to save changes to the form.

4.1.6. Required

This setting defines if the user must fill-in data for the corresponding variable.

1. Select the checkbox if data entry for the variable is mandatory.



Actions	ID	Description	Type	Key	Required
✚ ✖ □	SUBJECT	SubjectID	Text ▾	First ▾	☑
✚ ✖ □	CHOL	Total cholesterol, mg/dl	Numeric ▾		□

2. Select ***File > Save*** to save changes to the form.

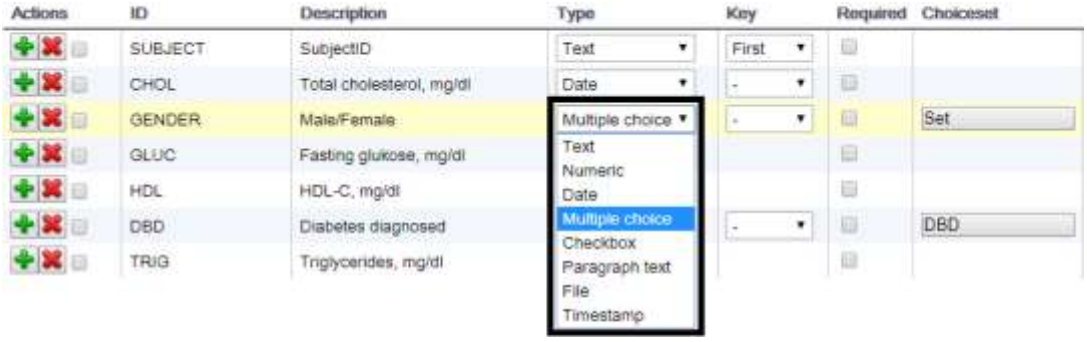
NOTE: This setting is only available when you are viewing or editing single data entry forms.

4.1.7. Choiceset

If a variable was defined as type Multiple choice, you must provide the list of items (= set) the user can select from in the column ***Choiceset***.

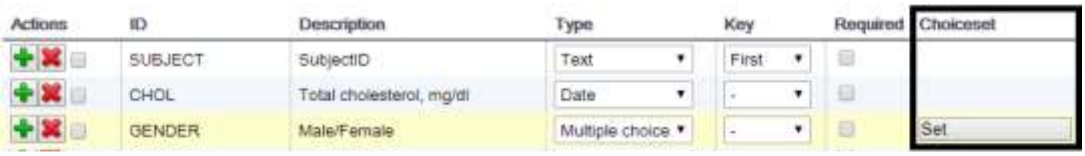
The following variable keys are available: First (primary key), Second (secondary key) and Third (tertiary key).

1. Select ***Multiple choice*** from the drop-down list in the ***Type*** column.



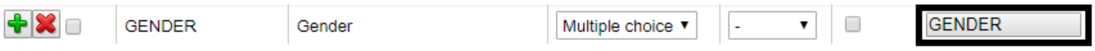
Actions	ID	Description	Type	Key	Required	Choiceset
✚ ✖ □	SUBJECT	SubjectID	Text ▾	First ▾	□	
✚ ✖ □	CHOL	Total cholesterol, mg/dl	Date ▾	- ▾	□	
✚ ✖ □	GENDER	Male/Female	Multiple choice ▾	- ▾	□	Set
✚ ✖ □	GLUC	Fasting glukose, mg/dl			□	
✚ ✖ □	HDL	HDL-C, mg/dl			□	
✚ ✖ □	DBD	Diabetes diagnosed		- ▾	□	DBD
✚ ✖ □	TRIG	Triglycerides, mg/dl			□	

A selectable field becomes available in the ***Choiceset*** column.



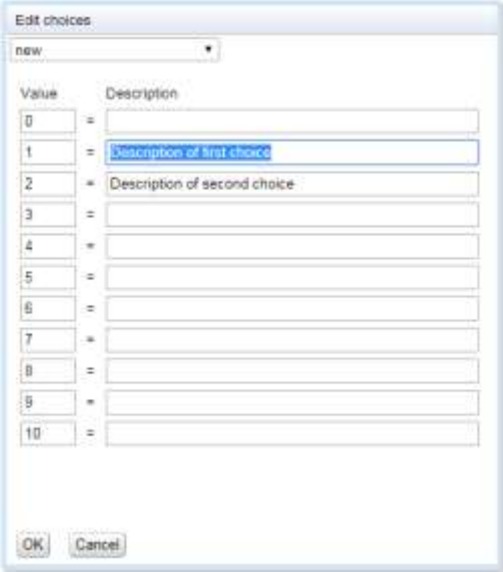
Actions	ID	Description	Type	Key	Required	Choiceset
✚ ✖ □	SUBJECT	SubjectID	Text ▾	First ▾	□	
✚ ✖ □	CHOL	Total cholesterol, mg/dl	Date ▾	- ▾	□	
✚ ✖ □	GENDER	Male/Female	Multiple choice ▾	- ▾	□	Set

2. Double-click the ***Choiceset*** field.



✚ ✖ □	GENDER	Gender	Multiple choice ▾	- ▾	□	GENDER

3. The ***Edit choices*** window opens.

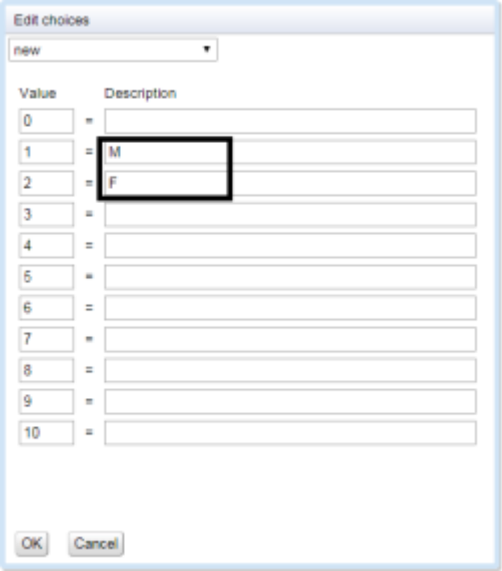


Edit choices

new

Value		Description
0	=	
1	=	Description of first choice
2	=	Description of second choice
3	=	
4	=	
5	=	
6	=	
7	=	
8	=	
9	=	
10	=	

4. Type the choices you want for the new choice set.



Edit choices

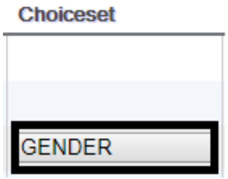
new

Value		Description
0	=	
1	=	M
2	=	F
3	=	
4	=	
5	=	
6	=	
7	=	
8	=	
9	=	
10	=	

NOTE: If biological sex information is needed in your statistical analyses specify ID of *SEX* in the form.
If subjects' affection status information is needed to be stored in a single column for your statistical analyses specify ID of *AFFSTAT* in the form: specify the values of 0, 1 and 2 for unknown, healthy and affected status, respectively.

5. Select ***OK***.

When you select ***OK***, the system automatically renames the Choiceset based on the variable ID.



Choiceset
GENDER

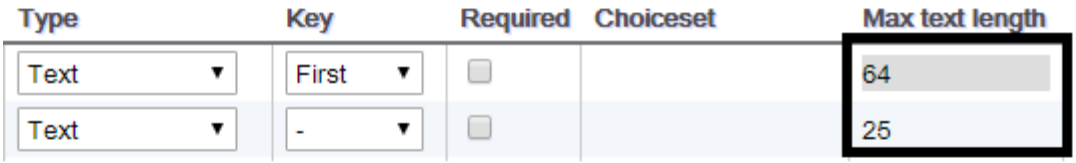
6. Select ***File > Save*** to save changes to the form.

CAUTION: If you change a multiple choice type to another type, the Choiceset is automatically deleted without warning.

4.1.8. Max text length

This setting defines the maximum length of a text variable.

1. Double-click the ***Max text length*** field of the desired text variable.
2. Type the desired maximum text length.



Type	Key	Required	Choiceset	Max text length
Text ▼	First ▼	☐		64
Text ▼	- ▼	☐		25

3. Select ***File > Save*** to save changes to the form.

NOTE: You can only change the maximum text length for non-key text variables.

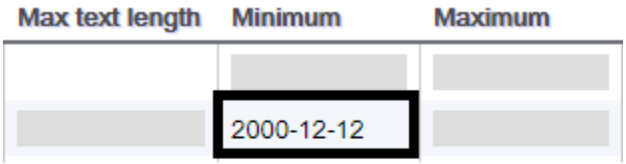
If the field background is greyed out, you cannot edit the minimum value.

4.1.9. Minimum

This setting defines the minimum value for a ***Type*** variable. The option is only available for variable types that have a minimum and maximum value, for example, a date or numbers.

NOTE: The ***Minimum*** setting is only available when you are viewing or editing single data entry forms.

1. Select the variable ***Type***, for example ***Date*** or ***Numeric***.
2. Double-click the ***Minimum*** field of the variable.
3. Type the desired value.



Max text length	Minimum	Maximum
	2000-12-12	

4. Select ***File > Save*** to save changes to the form.

NOTE: If you enter a date in the wrong format, the system displays a message informing you that you need to change it.

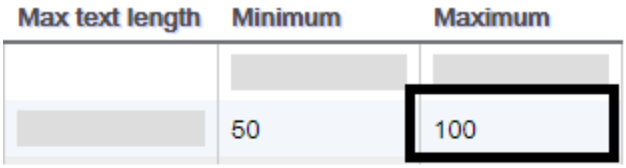
4.1.10. Maximum

This setting defines the maximum value for a ***Type*** variable. The option is only available for variable type of number.

NOTE: The ***Maximum*** setting is only available when you are viewing or editing single data entry forms.

If the field background is greyed out, you cannot edit the maximum value.

1. Select the variable ***Type***, for example ***Date*** or ***Numeric***.
2. Double-click the ***Maximum*** field of the variable.
3. Type the desired value.



Max text length	Minimum	Maximum
	50	100

4. Select ***File > Save*** to save changes to the form.

NOTE: If you enter a date in the wrong format, the system displays a message informing you that you need to change it.

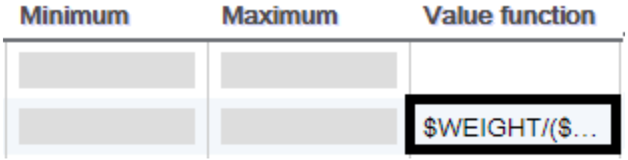
4.1.11. Value function

Value function allows the user to enter formulas, for example, to calculate the body mass index (BMI). Note that this setting is only available when you are viewing or editing single data entry forms.

NOTE: The ***Value function*** setting is only available when you are viewing or editing single data entry forms.

If the field background is greyed out, you cannot edit the maximum value.

1. Select the variable ***Type***, for example ***Text***, ***Date*** or ***Numeric***.
2. Double-click the ***Value function*** field of the variable.
3. Type the desired formula, for example, ***$WEIGHT/($HEIGHT*$HEIGHT***, to calculate the BMI.



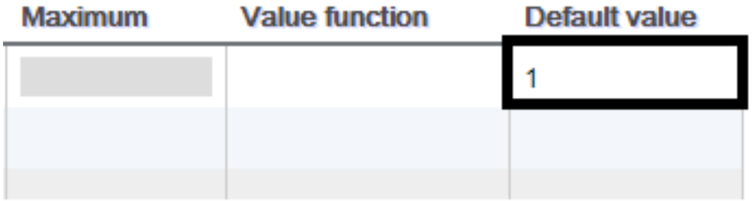
4. Select ***File > Save*** to save changes to the form.

4.1.12. Default value

This setting defines a default value for a variable. The default value is automatically displayed for that variable when the user fills in the form, and can be changed. Note that this setting is only available when you are viewing or editing single data entry forms.

NOTE: The ***Default value*** setting is only available when you are viewing or editing single data entry forms.

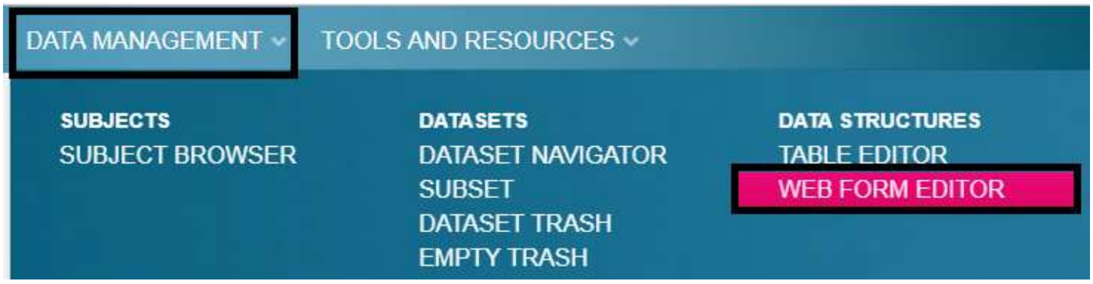
1. Double-click the ***Default value*** field of the variable.
2. Type the desired default value.



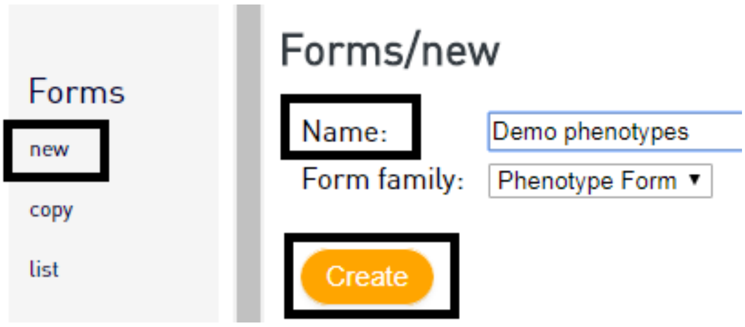
3. Select ***File > Save*** to save changes to the form.

4.2. Creating a phenotype form from scratch

1. Log in to BC|GENOME (see section 2.1).
2. Select ***DATA MANAGEMENT > WEB FORM EDITOR*** to open the BC|GENOME form editor tool.



3. Select ***Forms > new***.

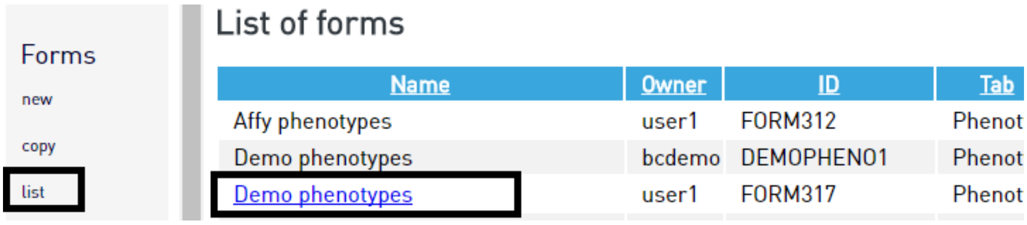


Forms/new

Name: Demo phenotypes

Form family: Phenotype Form ▼

4. Type a name for your form.
5. Select the ***Form family*** from the drop-down list.
6. Select ***Create***. Your new form displays in the *List of forms*.



List of forms

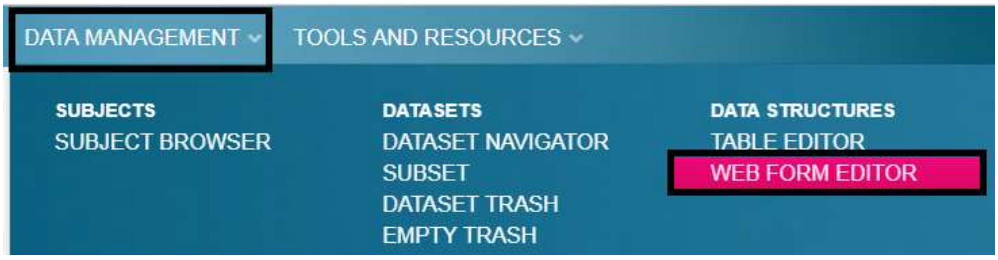
Name	Owner	ID	Tab
Affy phenotypes	user1	FORM312	Phenot
Demo phenotypes	bcdemo	DEMOPHENO1	Phenot
Demo phenotypes	user1	FORM317	Phenot

4.3. Creating a form using a form import file

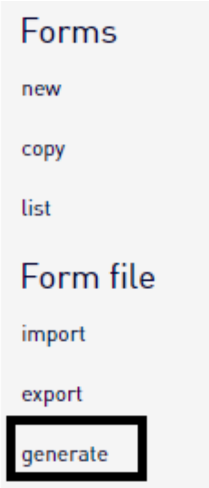
You can create a phenotype form automatically, whose structure matches the data in an existing import file. The import file contains variables and data values, as in the example *BMI_study_phenotypes.txt*.

1. Open the *BMI_study_phenotypes.txt* file, for example, in Excel to view its variables and data values.
2. Generate a form import file that defines the variables in the *BMI_study_phenotypes.txt* file:

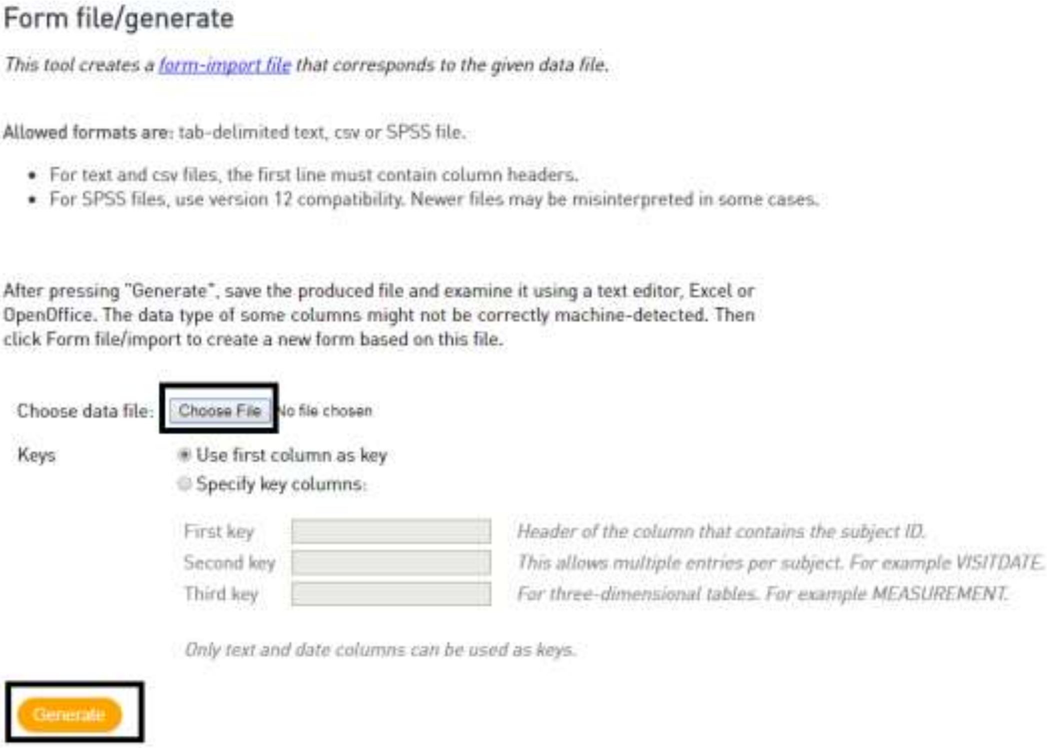
   a) Select ***DATA MANAGEMENT > WEB FORM EDITOR*** to open the BC|GENOME form editor tool.



   b) Select ***Forms > generate***.



c) Select ***Choose File*** and browse to the BMI_study_phenotypes.txt file.



Form file/generate

This tool creates a form-import file that corresponds to the given data file.

Allowed formats are: tab-delimited text, csv or SPSS file.

- For text and csv files, the first line must contain column headers.
- For SPSS files, use version 12 compatibility. Newer files may be misinterpreted in some cases.

After pressing "Generate", save the produced file and examine it using a text editor, Excel or OpenOffice. The data type of some columns might not be correctly machine-detected. Then click Form file/import to create a new form based on this file.

Choose data file: Choose File No file chosen

Keys ● Use first column as key
○ Specify key columns:

First key *Header of the column that contains the subject ID.*
Second key *This allows multiple entries per subject. For example VISITDATE.*
Third key *For three-dimensional tables. For example MEASUREMENT.*

Only text and date columns can be used as keys.

d) Select ***Generate***.

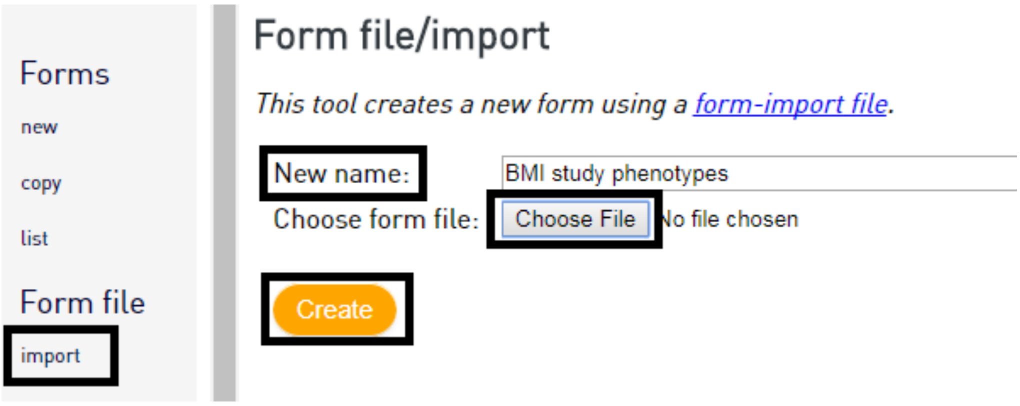
NOTE: The file that is generated is named automatically with ***_form*** in the title, and is most likely saved to the \Downloads folder or to your Desktop.

e) Save the file (for example, *BMI_study_phenotypes**_form**.txt)* to your desired location.

f) Open the generated file, for example, in Excel, and compare the variable names and types with the data in *BMI_study_phenotypes.txt*.

3. Create a new form using the form import file you just generated:

a) Select Form file > import.



Form file/import

This tool creates a new form using a form-import file.

New name: BMI study phenotypes

Choose form file: No file chosen

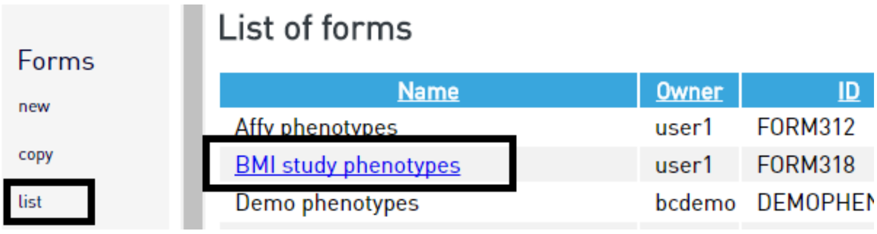
b) Type a new name for your form, for example, *BMI study phenotypes*.

c) Select ***Choose File*** and browse to the form import file you generated, for example, *BMI_study_phenotypes_***form***.txt*.

d) Select ***Create***.

Your new form displays in ***Forms > list***.

4. Select your new form in ***Forms > list*** to open it.



List of forms

Name	Owner	ID
Affy phenotypes	user1	FORM312
BMI study phenotypes	user1	FORM318
Demo phenotypes	bcdemo	DEMOPHEN

The form displays in the spreadsheet editor

5. Make changes to the form (see section 4.4.2).

6. Publish the form (see section 4.6).

4.4. Editing an unpublished form

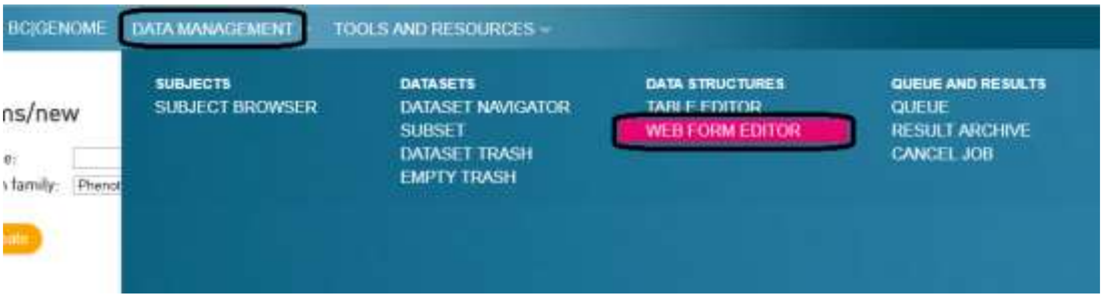
Once the initial form is created in ***Form file > import***, it has a set of variables based on your input file. You can add new or remove the default variables, and edit each variable's attributes to change the form to match your requirements.

NOTE:	Before you edit the variables, read section 4.1 for a description of a form's contents, to make it easier to decide what variables you need to include.

4.4.1. Opening a form to edit

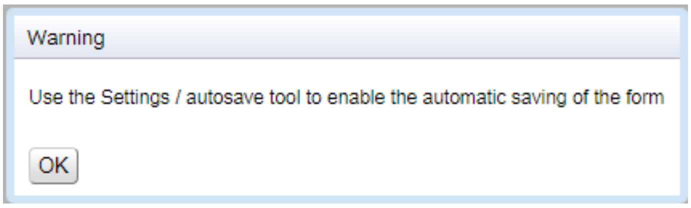
If you want to edit a form you must find and open it:

1. Select ***DATA MANAGEMENT > WEB FORM EDITOR*** to open the BC|GENOME form editor tool.



The default display is the ***List of forms***.

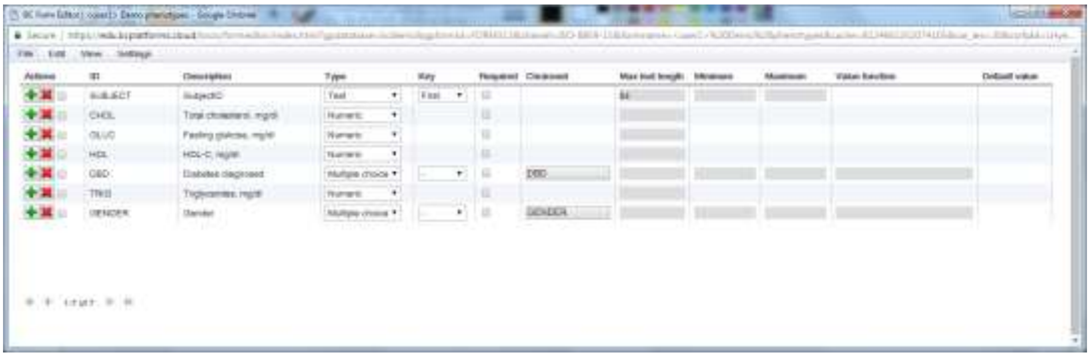
2. Select the name of the form you want to edit.
3. When opening the form layout editor the system shows you a warning: *Use the Settings / autosave tool to enable the automatic saving of the form*. Select ***OK***.


Warning

Use the Settings / autosave tool to enable the automatic saving of the form

For more information about automatic saving settings, see section 4.5.1.

The form displays in the ***Spreadsheet editor*** view and shows several variables. You can edit, remove or add variables. For more information about variables, see section 4.1.



Actions	ID	Description	Type	Key	Required	Codeset	Max text length	Minimum	Maximum	Value function	Default value
	SUBJECT	SubjectID	Text	First			64				
	CHOL	Total cholesterol, mg/dl	Numeric								
	GLUC	Fasting glucose, mg/dl	Numeric								
	HDL	HDL-C, mg/dl	Numeric								
	DBD	Diabetes diagnosed	Multiple choice			DBD					
	TRIG	Triglycerides, mg/dl	Numeric								
	GENDER	Gender	Multiple choice			GENDER					

4.4.2. Making changes to a form

When you have opened the form you want to edit you can make changes to the variables in the form.

1. Enter a unique identifier (***ID***) for the record, typically a combination of numbers and letters.

NOTE:	For every form, it is mandatory to have a primary key variable that is used to identify each record. By default, when a new form is created, the SUBJECT variable (visible as the first variable) has the primary key assigned to it and is used as a unique identifier for the record.

When you enter the unique identifier ***ID***, note the following:

- No empty spaces are allowed.
- Allows only characters A-Z, 0-9 and _

2. Select ✚ to add, or ✖ to remove variables.

NOTE: When you choose to remove a variable, it is removed immediately with no confirmation message!

---

3. Select and edit variables.

   For more information about variables, see section 4.1.

4. From the ***Settings menu***, edit the ***Date Format*** or ***Autosave***.

   For information about changing the settings, see section 0.

5. Select ***File > Save*** to save the changes to your form.

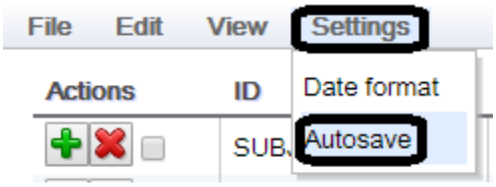
NOTE: If you select ***File > Print***, only what you can see on screen is printed.

4.5. Changing form settings

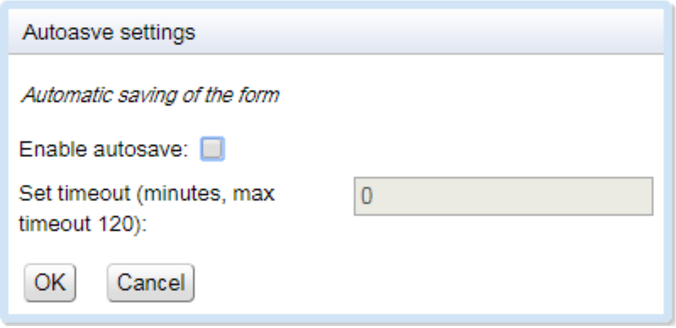
You can change the date format and the autosave settings for a form.

4.5.1. Changing automatic save

1. Go to ***Settings > Autosave*** to edit the autosave settings.



2. Select ***Enable autosave*** and set the timeout (in minutes).



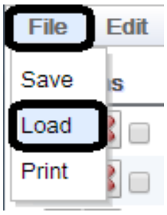
Automatic saving of the form

Enable autosave: □

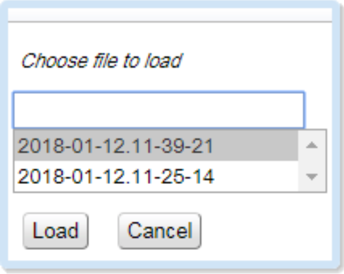
Set timeout (minutes, max timeout 120): 0

3. Select ***OK.***

   If you need to restore the previous save, select ***File > Load***.



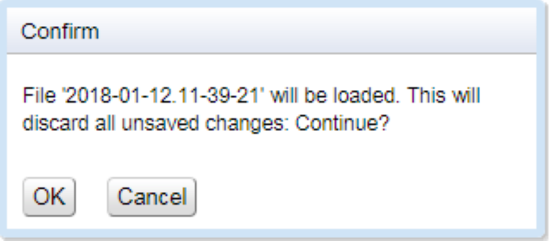
4. Select the file to load based on the timestamp information and select ***Load***.



Choose file to load

2018-01-12.11-39-21
2018-01-12.11-25-14

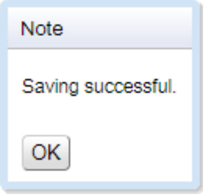
5. Select ***OK*** to confirm.



Confirm

File '2018-01-12.11-39-21' will be loaded. This will discard all unsaved changes: Continue?

6. Select ***File > Save*** when your work with a form is complete. A message displays to say your file is saved.

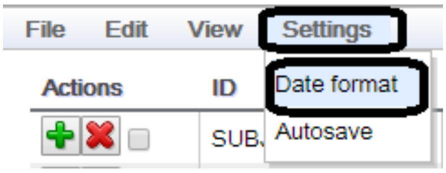


Note

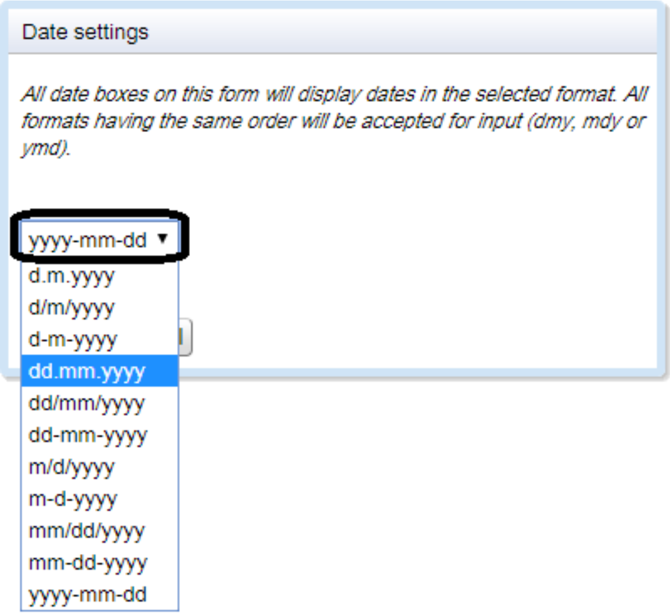
Saving successful.

4.5.2. Changing date format

1. Go to ***Settings > Date*** format.



2. Select the date format you want to use from the drop-down menu.



Date settings

All date boxes on this form will display dates in the selected format. All formats having the same order will be accepted for input (dmy, mdy or ymd).

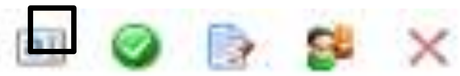
3. Select ***OK***.

4.6. Publishing a form

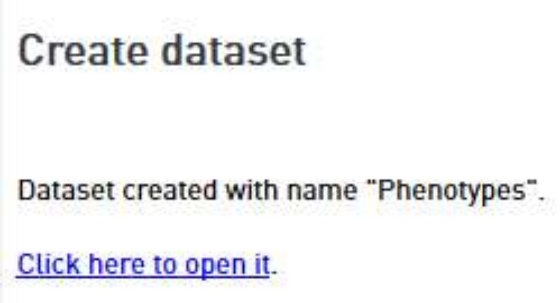
When you have created or edited a form, you need to publish it. For phenotype forms the publish action creates a dataset with the same name as the form.

To publish a form as a dataset:

1. Select the ***Publish*** button in the in ***Forms > List*** of the ***Web form editor***.



2. Select ***Click here to open it*** when you see confirmation that the dataset was created.



Create dataset

Dataset created with name "Phenotypes".

Click here to open it.

3. Open your dataset to start working.

   The dataset name can be found in the BC|GENOME navigation pane on the main page.

4.7. Data ontologies

4.7.1. Describing the data

It is often the case that simply creating data structures that are able to accommodate text, numbers, dates and so on, is not enough to describe the data, and to understand the relationships between different variables. A research project may be able to simply agree upon a way how things are described, and stick with those rules, especially if the design of the data model is strictly enforced in the research. However, when the project grows, or new projects start, old rules may need to change, and some will be forgotten.

A commonly used solution is to apply a data vocabulary - or dictionary - on top of the actual data structure, and describe the content that way. A typical example could be the patient identifier. In BC the convention is to use SUBJECT as the name of the variable that denotes patients ID. However, it is only a convention, and thus can be replaced with something else. Also when the project receives data from outside sources, there may be different kind of variable naming in use. The alternative field names can quickly become confusing, and when one needs to view the data for the same patient across differently built data tables, these changes may become challenging.

In BC|GENOME the SUBJECT field comes with an annotation "BC:subject". This is an inbuilt attribute to a field that has a meaning: "This is patient ID". If users wish, or need, to use different naming for the patient ID variables, application of the annotation "BC:subject" will tell the BC|GENOME as a system, the the variables mean patient ID. This helps BC|GENOME to find tables with patient data, and allows it give automated suggestions for content and datasets in various tools working on patient data.

The same applies to almost everything in the data model in use. Genomic marker names in BC|GENOME are annotated "BC:marker", and chromosomes "BC:chromosome", and so on. It is possible for users to have their own designed vocabulary in BC|GENOME, to describe their data model and the data. BC|GENOME has an increasing number of tools that are able to utilise this information.

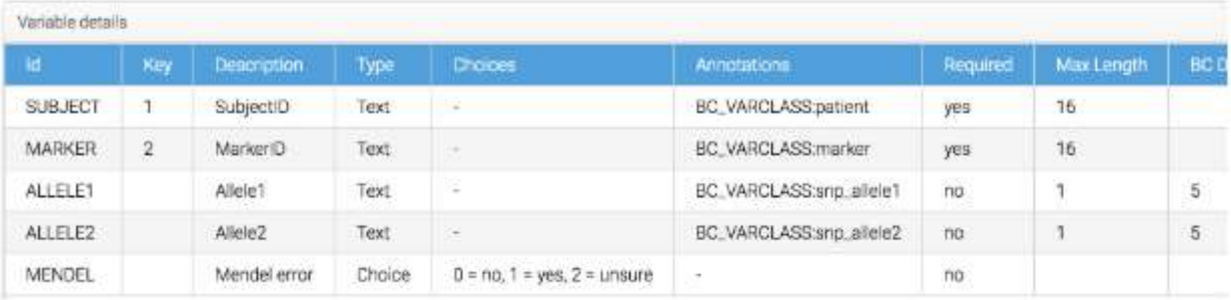
4.7.2. Ontology – annotation

You will see throughout this document the words ontology and column annotation being used in mixed manner. It may seem confusing. Ontology means the data vocabulary describing the data, and the individual items in an ontology are called terms. However, the documentation often refers to a term that is applied to a column - like "BC:subject" - as the ontology for that column. Annotation of a column means the process of applying vocabulary to the columns. Hence the seemingly mixed use of these terms, that essentially mean the same thing in the context of columns.

Do not mix column annotation with genomic or omics data annotation, which is again a different thing.

4.7.3. Inbuilt ontologies

BC operating system provides ontologies for inbuilt data structures available for data management. At the moment BC systems use two different ontologies, named "BC_VARCLASS", and "BC". Of these two "BC_VARCLASS" is partially deprecated and is being replaced by the "BC" ontology. Typical examples of forms annotated still with BC_VARCLASS ontologies are for example "ACGT coded SNPs" (form ID PLAINSNPS).

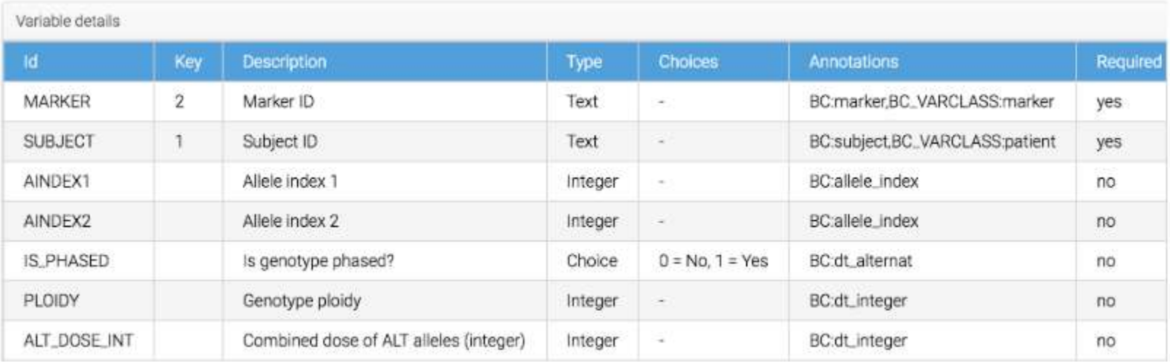


Variable details

Id	Key	Description	Type	Choices	Annotations	Required	Max Length	BC D
SUBJECT	1	SubjectID	Text	-	BC_VARCLASS:patient	yes	16	
MARKER	2	MarkerID	Text	-	BC_VARCLASS:marker	yes	16	
ALLELE1		Allele1	Text	-	BC_VARCLASS:snp_allele1	no	1	5
ALLELE2		Allele2	Text	-	BC_VARCLASS:snp_allele2	no	1	5
MENDEL		Mendel error	Choice	0 = no, 1 = yes, 2 = unsure	-	no		

Image: ACGT coded SNPs -form annotated with BC_VARCLASS ontology for patient, marker, and alleles.

All new data structures are annotated using the more recent and more structured BC ontology, like VCF Composite datasets.



Variable details

Id	Key	Description	Type	Choices	Annotations	Required
MARKER	2	Marker ID	Text	-	BC:marker,BC_VARCLASS:marker	yes
SUBJECT	1	Subject ID	Text	-	BC:subject,BC_VARCLASS:patient	yes
AINDEX1		Allele index 1	Integer	-	BC:allele_index	no
AINDEX2		Allele index 2	Integer	-	BC:allele_index	no
IS_PHASED		Is genotype phased?	Choice	0 = No, 1 = Yes	BC:dt_alternat	no
PLOIDY		Genotype ploidy	Integer	-	BC:dt_integer	no
ALT_DOSE_INT		Combined dose of ALT alleles (integer)	Integer	-	BC:dt_integer	no

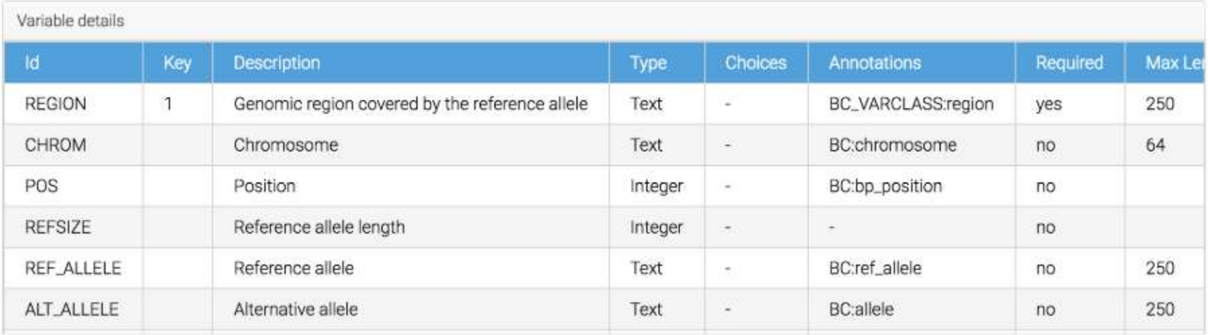
Image: VCF Composite genotypes form annotated with BC ontology in addition to BC_VARCLASS.

"BC_VARCLASS" vs "BC" ontologies are different. "BC_VARCLASS" provides pure classification for important data fields that are used internally by the BC operating system.

"BC" -ontology on the other hand provides classification (like patient and marker), but also typing for data, like in the above example BC:dt_alternat, which tells the tools in the system that the field must contain enumerated choices (i.e. 1:male, 2:female, 3:NA). The latter feature is used to distinguish between fields that are similar on data model (SQL) level but provide different functionalities (alternative choices coded as integers, vs. actual integer numbers like 'ploidy').

4.7.4. Multiple fields combined for one function

The ontology design allows application of ontologies in such way that they define a function that requires multiple fields. The combination of multiple fields then becomes so called virtual column, accessible by tools as one logical item. A good example of a virtual column is the genomic range annotation, which is defined by three fields: chromosome, start location, and end location.



Variable details

Id	Key	Description	Type	Choices	Annotations	Required	Max Le
REGION	1	Genomic region covered by the reference allele	Text	-	BC_VARCLASS:region	yes	250
CHROM		Chromosome	Text	-	BC:chromosome	no	64
POS		Position	Integer	-	BC:bp_position	no	
REFSIZE		Reference allele length	Integer	-	-	no	
REF_ALLELE		Reference allele	Text	-	BC:ref_allele	no	250
ALT_ALLELE		Alternative allele	Text	-	BC:allele	no	250

Image: Virtual field 'range', which is used to tell the BC system that this annotation form is capable of providing genomic range data to tools. Note that the form has both BC_VARCLASS, and the more modern BC:bc_position annotations for relevant fields.

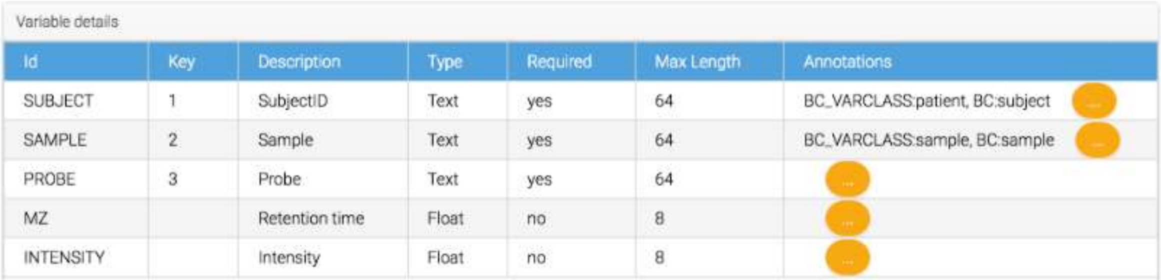
4.7.5. Ontologies and annotating form variables

BC systems operate on SQL and other data structures, which store different types of data. The structures are described in BC systems as table 'forms' or table 'templates'. A form describes the fields in a table structure, the data type in each field, and some additional constraints about how the field should behave, like maximum length in text fields, options in alternative choices, etc. The form also tells the SQL or other data structures, which field are to be used as key fields for indexing the data.

Many forms are similar in structure, and user will often see things like 'SUBJECT' or 'MARKER' being used across multiple forms to denote patient and genetic SNP marker, respectively. These naming convention are used to make the structures easier to understand with quick glance. For example, seeing a dataset with the field 'SUBJECT' in it tells the user that the table contains patient data, and that 'SUBJECT' is likely to be a key field of type text.

However, it is perfectly possible for users to define their own forms, where the field that means patient key, might not be called 'SUBJECT'. The BC operating system should regardless be able to recognize the field as equivalent to SUBJECT field, and adjust the behavior of data management and analytical tools accordingly. The system also defines inbuilt forms containing fields with less clear meanings, and the user may want to override these in her own forms, without losing the functionality of specialized tools that take use of those fields.

BC internal ontologies are used to give any field or group of fields a functional or conceptual meaning. Ontology is used to create a vocabulary inside the BC system that describes the structure of the data - beyond the mere field names and descriptions - in such way that the underlying operating system, and analytical and logical tools can use the information in meaningful way. See Image 1 for an example of usage of BC internal ontology terms.



Variable details

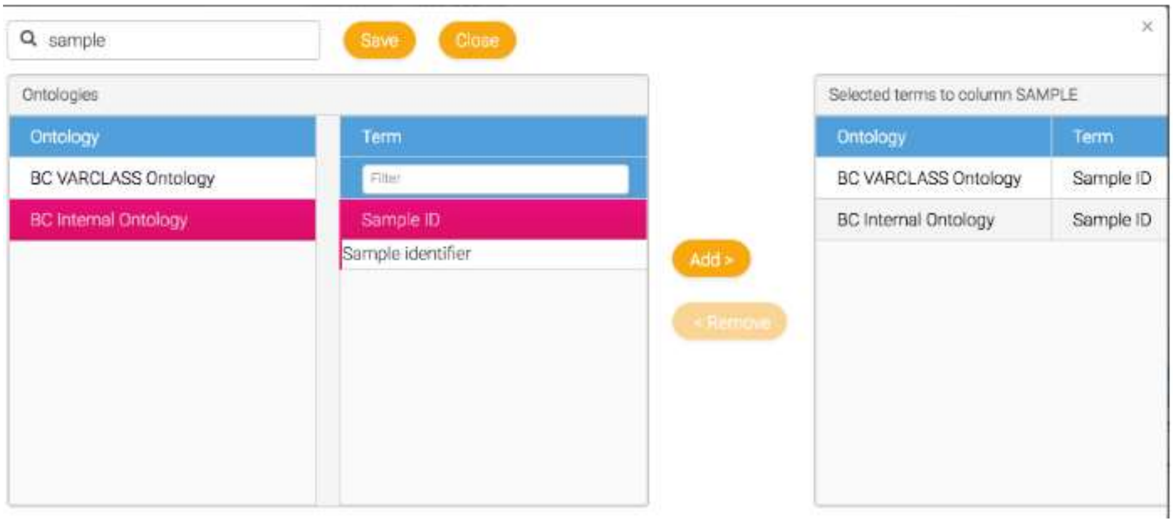
Id	Key	Description	Type	Required	Max Length	Annotations
SUBJECT	1	SubjectID	Text	yes	64	BC_VARCLASS:patient, BC:subject ...
SAMPLE	2	Sample	Text	yes	64	BC_VARCLASS:sample, BC:sample ...
PROBE	3	Probe	Text	yes	64	...
MZ		Retention time	Float	no	8	...
INTENSITY		Intensity	Float	no	8	...

Image 1. Annotation of inbuilt form, seeing the structure in the dataset's STRUCTURE -page. The tools in 'Annotations' -field can be used to edit the ontology terms.

For example, when joining patient data tables together, it is useful if the system already understands the concept of 'patient', and can suggest joins automatically using the fields that hold patient index in both datasets. When creating annotations for genomic features like SNPs, it is useful if the system automatically will recognize a field as a genetic SNP marker, understands how a gene range is formed from chromosome and start and stop positions, and so on. These are functional examples of the use of internal ontology to help built meaningful tools.

4.7.5.1. Editing ontology terms

It is sometimes necessary to edit the ontology terms of existing datasets, to add meaning to fields, to make them visible in searchers or usable in ontology-aware BC|GENOME tools. Sometimes during data model design phase it is helpful to be able to tweak the annotations of already existing structures to see their effects in use. You can apply ontologies in the dataset's STRUCTURE -page.



Ontology	Term
BC VARCLASS Ontology	Sample ID
BC Internal Ontology	Sample identifier

Selected terms to column SAMPLE

Ontology	Term
BC VARCLASS Ontology	Sample ID
BC Internal Ontology	Sample ID

4.7.6. Tools that uses ontologies

The list of tools that automatically utilize the internal BC ontologies is constantly growing. Some notable ones available in the current production version of BC|GENOME are:

- Dataset search: Search tool in the dataset navigator indexes the datasets based on metadata and field ontologies
- Analytical tools: Analysis interface automatically collects and joins data as input for analysis algorithms based on patient, marker, etc ontologies
- Subset tool: Matching datasets for joins using BC:subject
- Annotation tool: Uses BC:genomic_range and BC:genomic_pos to annotate genetic marker lists with corresponding gene annotations

4.7.7. Data vocabularies

Beyond functional features in the system, it is possible for the research project to define their own internal data vocabulary that describes the data in the system. The BC application programming interfaces (APIs) provide means for searching and recognizing table fields based on their annotated ontology, some search features can utilize the information automatically, and many future tools will allow the user to use the data vocabulary for data management and visualization.

Data vocabulary describes the research meaning of the data. There are many existing ontologies that could be used for this purpose, in the areas of clinical research, human genetics and genomics, health data, and so on. These ontologies (see examples for Gene Ontology Consortium, Human Phenotype Ontology, Ontology of Clinical Research) are useful on their own, and harmonize the content in the BC database to a more widely used standard. However, from data management point of view, ready-ontologies can appear stiff and restricting. If this is the case, providing project-specific vocabulary is called for.

Data vocabulary development should start from the system usability point of view. Data administrators should ask themselves:

1. What are the typical data questions being asked by the platform users
2. Which data items are involved in the answers
3. What kind of accelerating or aiding structures the database could host to help discovery

Based on the assessment of use-cases for data discovery, the vocabulary can be applied to commonly used fields, units, measurements, processes, observations, methods, etc. These items should form the basis for searchers, automation and workflows, and reporting in the system.

5. Working with data in BC|GENOME

This chapter describes the tasks you can perform when working with data in BC|GENOME. These include:

- Creating a dataset
- Creating subsets

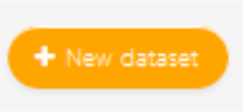
NOTE:	Before you begin working with data, you must first upload any required data so it is available for use. For more information about uploading data, see chapter 3.2 Uploading data.

5.1. Creating a dataset

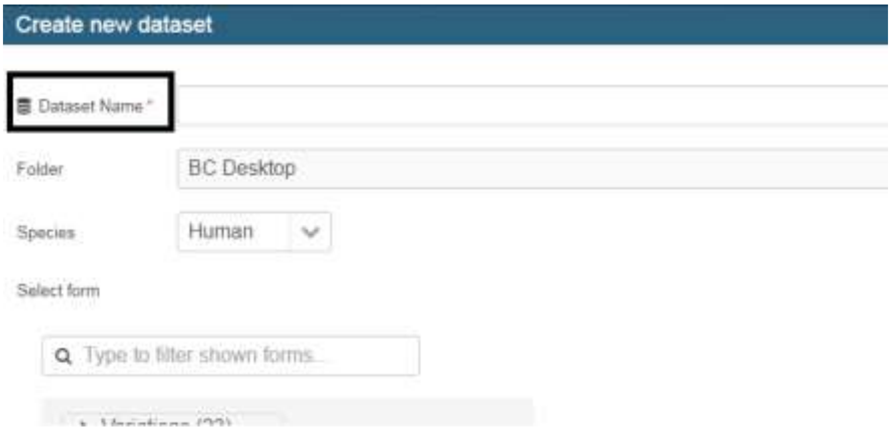
You can create a new dataset based on a pre-existing form if there is no suitable dataset available.

To create a new dataset:

1. Select ***New dataset*** from the main page.



2. In the ***Create new dataset*** dialog, type a unique name for your new dataset in the ***Dataset Name*** field.


Create new dataset

Dataset Name *

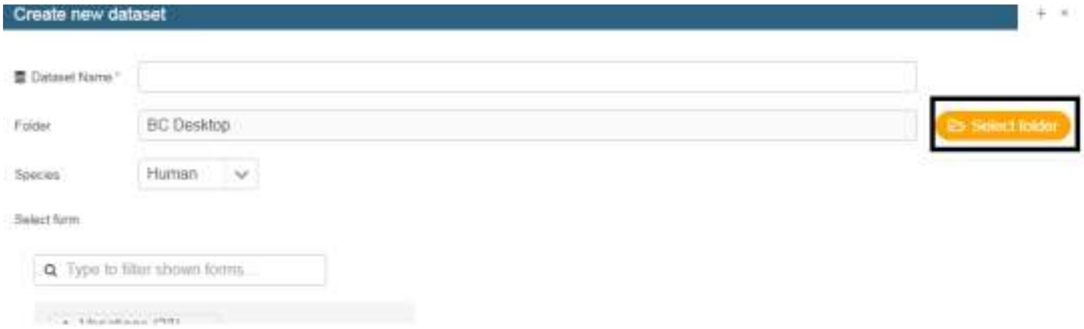
Folder BC Desktop

Species Human

Select form

Type to filter shown forms...

3. Select ***Select Folder*** if you want to show the dataset somewhere else than in the BC Desktop folder.



Create new dataset

Dataset Name *

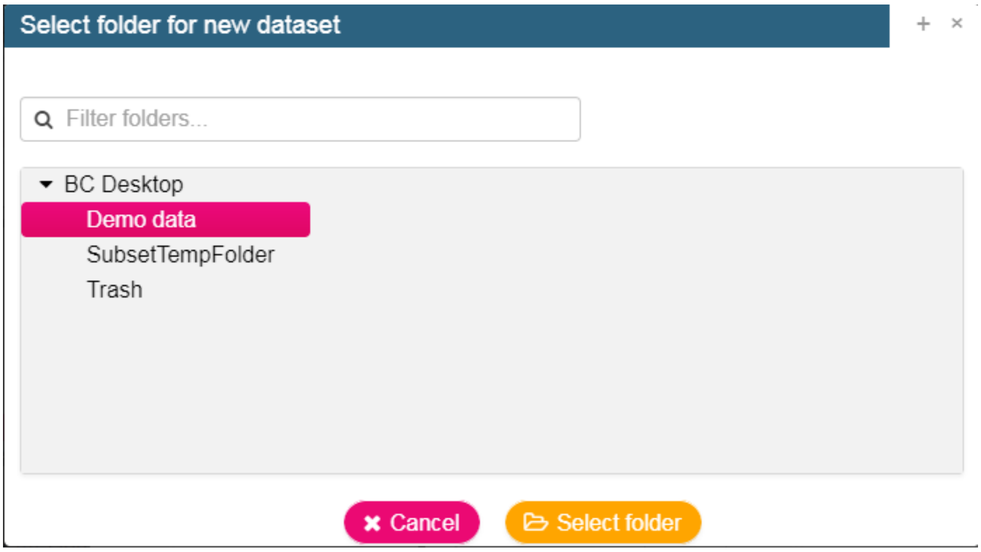
Folder: BC Desktop

Species: Human

Select form

The default folder is BC Desktop.

4. Select the folder in which you want to create the new dataset.



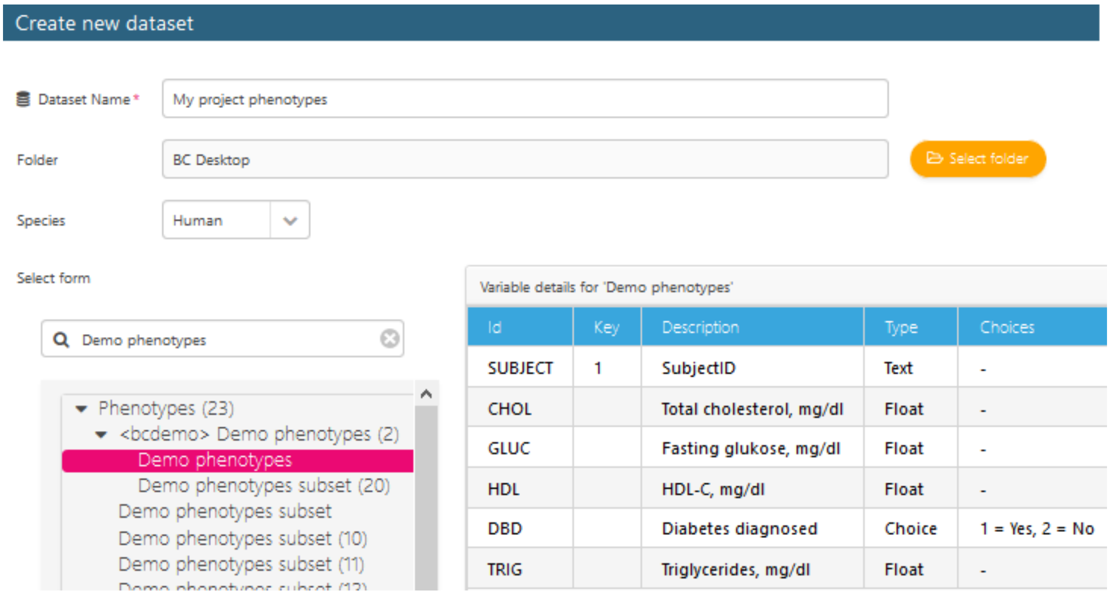
Select folder for new dataset

- BC Desktop
  - Demo data
  - SubsetTempFolder
  - Trash

5. In ***Species***, select the species you want from the drop-down menu.

6. In ***Select form,*** search for the form you wish to use for your dataset.

   You can view the variable details of a form on the right side of the dialog.



Create new dataset

Dataset Name * My project phenotypes

Folder BC Desktop

Species Human

Select form

Demo phenotypes

- Phenotypes (23)
  - <bcdemo> Demo phenotypes (2)
    - Demo phenotypes
    - Demo phenotypes subset (20)
  - Demo phenotypes subset
  - Demo phenotypes subset (10)
  - Demo phenotypes subset (11)
  - Demo phenotypes subset (12)

Variable details for 'Demo phenotypes'

Id	Key	Description	Type	Choices
SUBJECT	1	SubjectID	Text	-
CHOL		Total cholesterol, mg/dl	Float	-
GLUC		Fasting glukose, mg/dl	Float	-
HDL		HDL-C, mg/dl	Float	-
DBD		Diabetes diagnosed	Choice	1 = Yes, 2 = No
TRIG		Triglycerides, mg/dl	Float	-

7. Select ***Create dataset*** to create your new dataset.

Your new dataset is visible in the navigation pane in the main page.

5.2. Subsets

Subsets are used if you want to keep a set of criteria for filtered information.

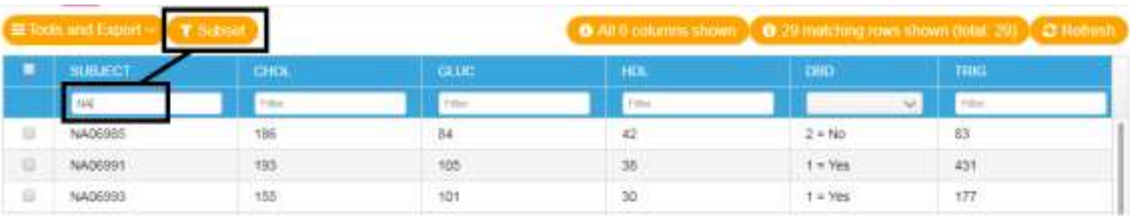
You can:

- Create simple subsets using filtering options in the data grid
- Create subsets using DATA MANAGEMENT > SUBSET
- Join information in subsets

5.2.1. Creating simple subsets using filtering options in the data grid

The ***DATA*** tab provides tools to filter data stored in a dataset. It uses the *AND* option between the columns.

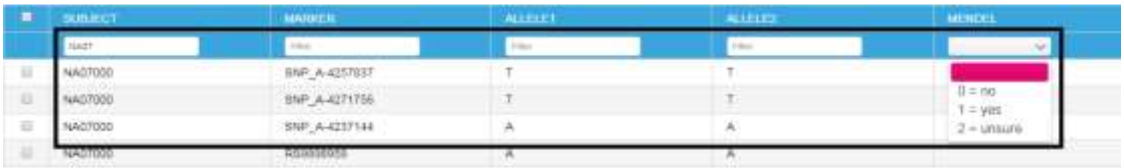
The ***Subset*** button appears when you start to set filter information in a column’s filter field.



	SUBJECT	CHOL	GLUC	HDL	DBD	TRIG
	NA	Filter	Filter	Filter		Filter
□	NA06985	186	84	42	2 = No	83
□	NA06991	193	105	38	1 = Yes	431
□	NA06993	155	101	30	1 = Yes	177

To filter data in the ***DATA*** page:

1. Select the dataset you want from the main page.
2. Select the ***DATA*** tab to open the DATA page.
3. Specify one or more filtering criteria in the ***Filter*** field below the column title.



■	SUBJECT	MARKER	ALLELE1	ALLELE2	MENDEL
	NA07	Filter	Filter	Filter	
□	NA07000	SNP_A-4257837	T	T	
□	NA07000	SNP_A-4271756	T	T	0 = no
□	NA07000	SNP_A-4237144	A	A	1 = yes 2 = unsure
□	NA07000	[illegible]	A	A	

– Type specific text.

– Use an operator (**<**, **>** or **-**) for specific numeric information (for example, <200, >100 or 100-200).

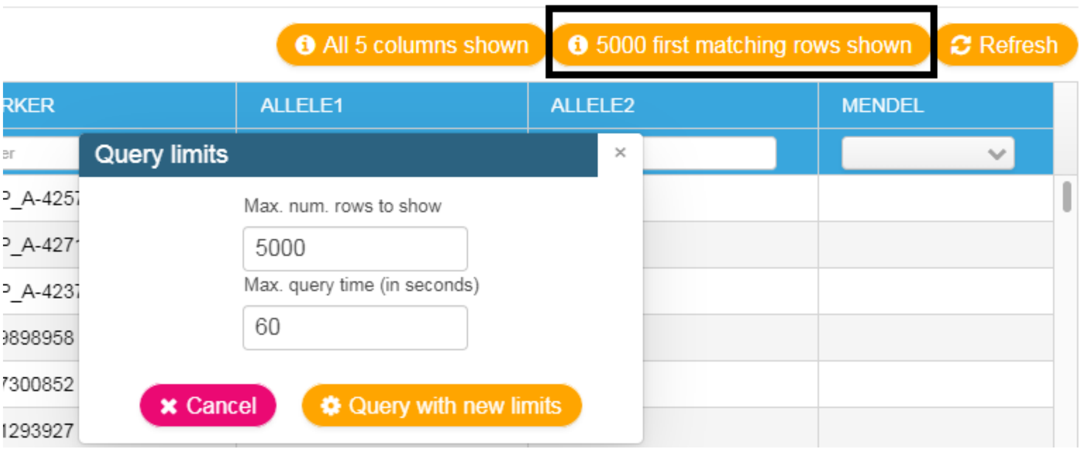
– Select from a drop-down menu, if available.

NOTE: You can only specify one option in the case of a multiple choice variable.

4. Select ***N matching rows shown (total: N)***.

a) Type the maximum number of rows to be shown in the grid page.

b) Type the maximum query time.



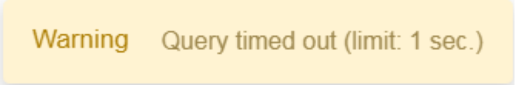
Query limits

Max. num. rows to show: 5000

Max. query time (in seconds): 60

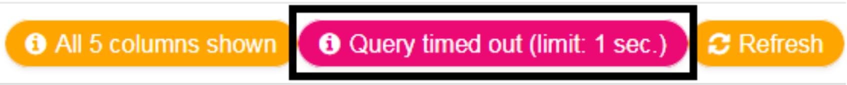
5. Select ***Query with new limits***.

If the query time is exceeded, a pop-up displays to indicate that the query has timed out.



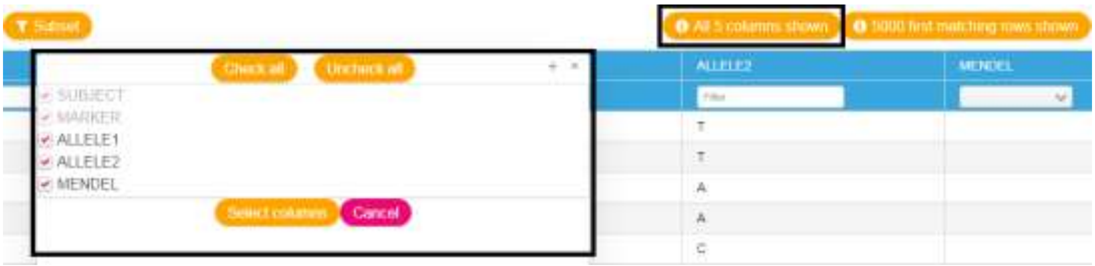
Warning Query timed out (limit: 1 sec.)

a) Select Query timed out.



b) Type a new maximum query time.

6. Select ***All N columns shown***.

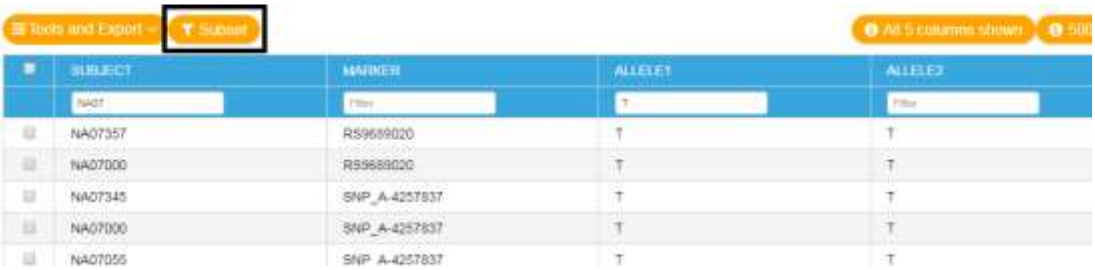


a) Select the columns you want to show in the grid.

b) Select ***Select columns*** to save your choice.

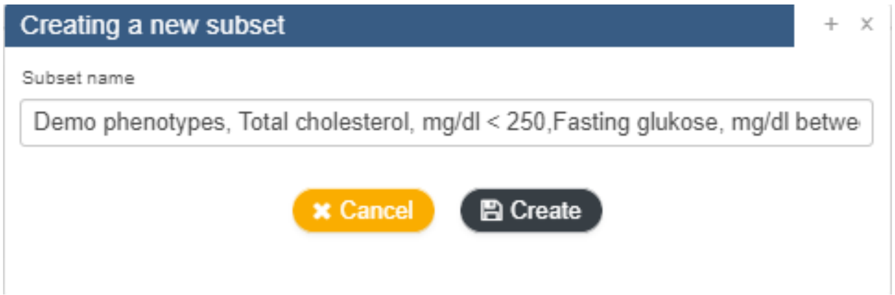
NOTE: If the column list is long, you must scroll down to find the ***Select columns*** button.

7. Select ***Subset*** to create a subset based on the filtering criteria.



	SUBJECT	MARKER	ALLELE1	ALLELE2
	NA07	Filter	T	Filter
□	NA07357	RS9689020	T	T
□	NA07000	RS9689020	T	T
□	NA07345	SNP_A-4257837	T	T
□	NA07000	SNP_A-4257837	T	T
□	NA07056	SNP_A-4257837	T	T

8. Type a new subset name, if needed, in the ***Creating a new subset*** dialog.

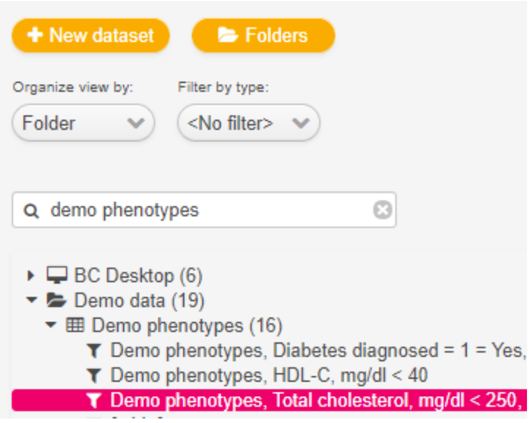


Subset name: Demo phenotypes, Total cholesterol, mg/dl < 250,Fasting glukose, mg/dl betwe

NOTE: You can have up to 250 characters in the subset name.

9. Select ***Create*** to make the subset available in BC|GENOME.

The subset shows in the navigation pane.



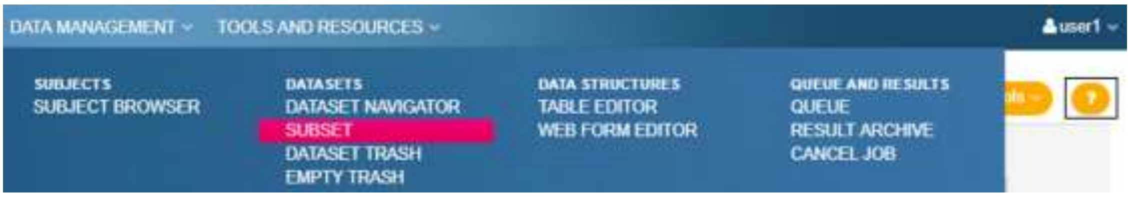
5.2.2. Creating subsets using DATA MANAGEMENT > SUBSET

In BC|GENOME, it is possible to filter information from one dataset and to join information from several datasets using the ***SUBSET*** tool in ***DATA MANAGEMENT***.

All available datasets are listed in the hierarchical dataset tree on the left side of the screen. Datasets are dragged and dropped to the canvas on the right side. This is the primary view of the selected datasets in the subset tool. In this view, datasets can be added from the dataset list and manipulated interactively in the subset.

If a subset has at least one dataset, dragging another one to the canvas calls an automatic join candidate generator. For more information about joining dataset information, see section 5.2.3.

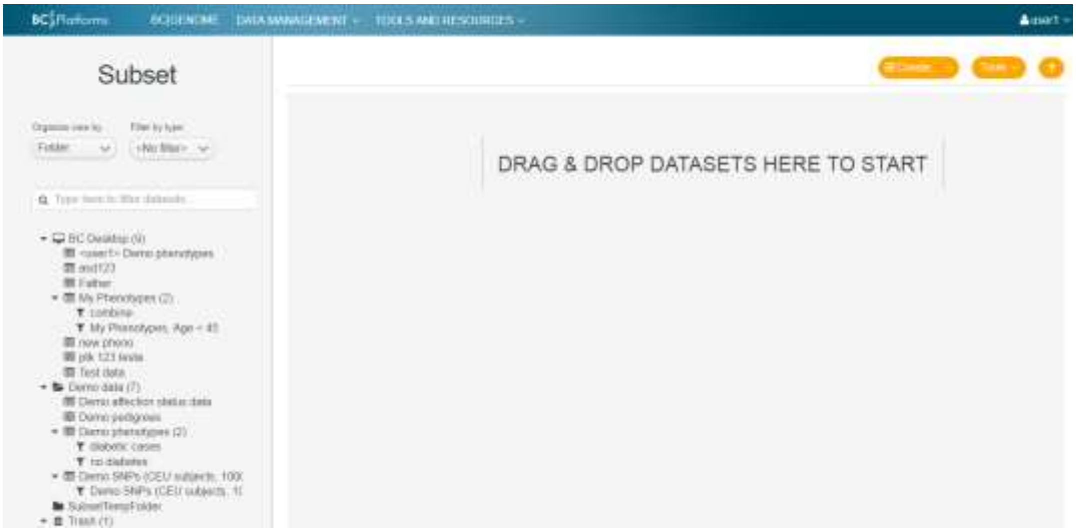
For more information about the subset tool, select ? in the *SUBSET* tool.



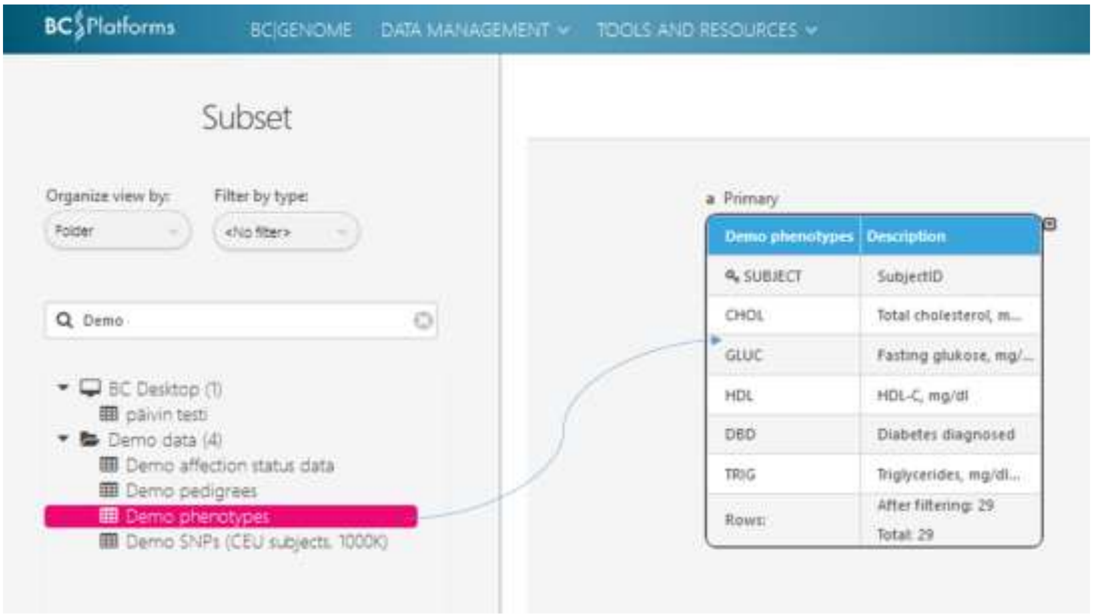
To create a subset:

1. Select ***DATA MANAGEMENT > SUBSET***.

   The ***Subset*** page opens.

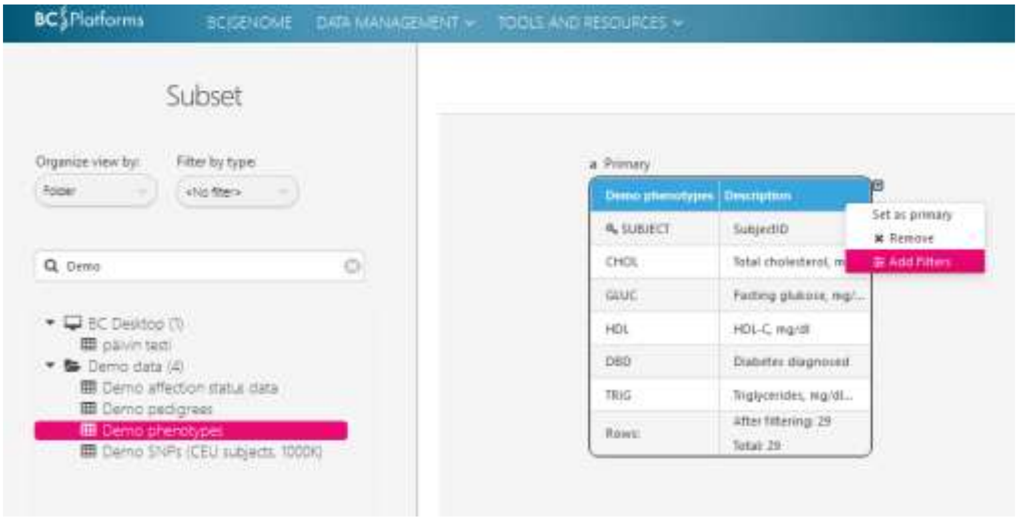


2. Drag and drop the desired dataset from the ***Subset*** navigation pane to the editor canvas. The dataset information is displayed as a table.



Demo phenotypes	Description
SUBJECT	SubjectID
CHOL	Total cholesterol, m...
GLUC	Fasting glukose, mg/...
HDL	HDL-C, mg/dl
DBD	Diabetes diagnosed
TRIG	Triglycerides, mg/dl...
Rows:	After filtering: 29 Total: 29

3. Right-click the dataset table to see the table menu.



Subset

Organize view by: Folder

Filter by type: <No filter>

Demo

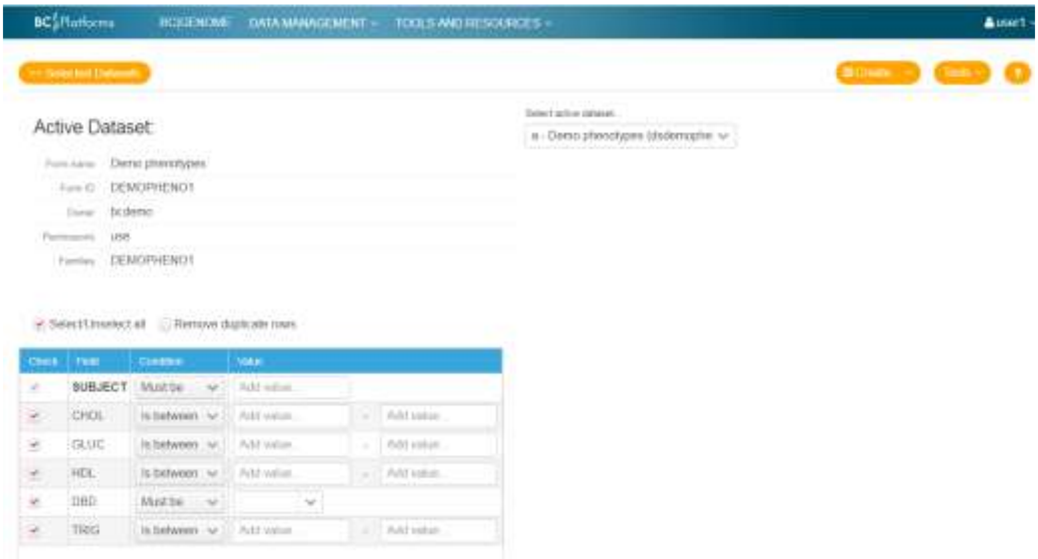
- BC Desktop (1)
  - päivin testi
- Demo data (4)
  - Demo affection status data
  - Demo pedigrees
  - Demo phenotypes
  - Demo SNPs (CEU subjects, 1000K)

Primary

Demo phenotypes	Description
SUBJECT	SubjectID
CHOL	Total cholesterol, m
GLUC	Fasting glukose, mg/...
HDL	HDL-C, mg/dl
DBD	Diabetes diagnosed
TRIG	Triglycerides, mg/dl...
Rows:	After filtering: 29 Total: 29

4. Select the ***Add filters***.

   A new page opens where you can set the required filters.



Active Dataset

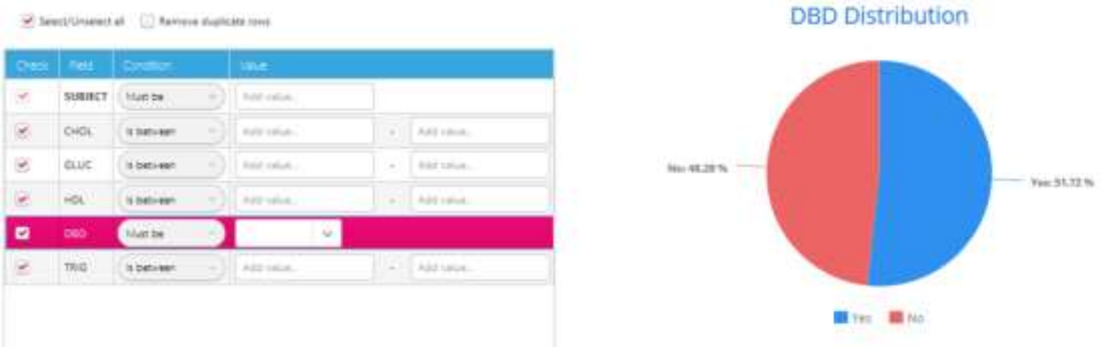
Form Name	Demo phenotypes
Form ID	DEMOPHENO1
Owner	bcdemo
Permissions	use
Families	DEMOPHENO1

Select active dataset: a - Demo phenotypes (dsdemophe

☑ Select/Unselect all ☐ Remove duplicate rows

Check	Field	Condition	Value		
✓	SUBJECT	Must be	Add value...		
✓	CHOL	Is between	Add value...	-	Add value...
✓	GLUC	Is between	Add value...	-	Add value...
✓	HDL	Is between	Add value...	-	Add value...
✓	DBD	Must be			
✓	TRIG	Is between	Add value...	-	Add value...

5. Select a variable information row to view the distribution chart of that variable.

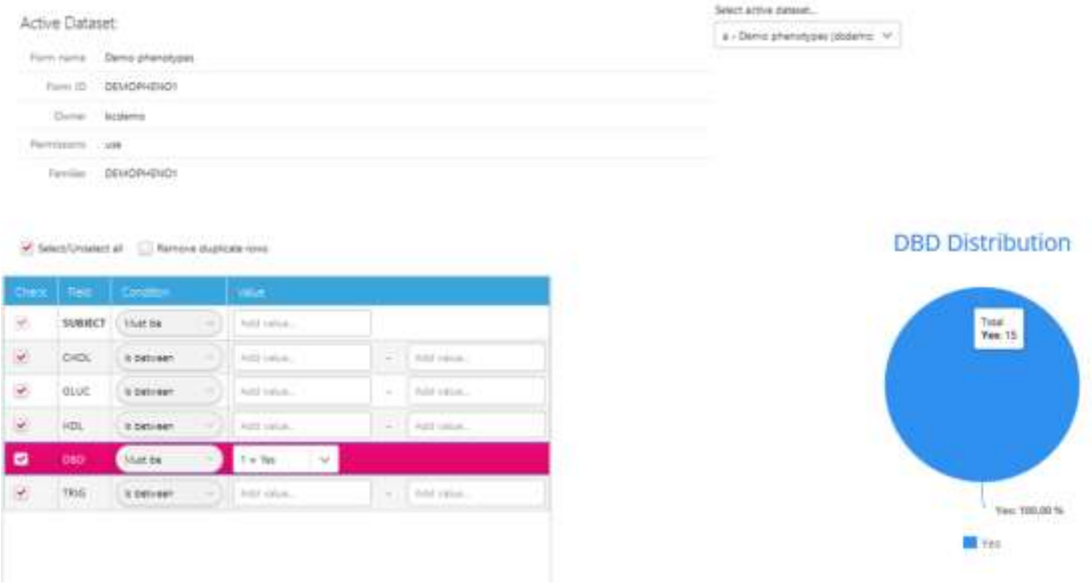


☑ Select/Unselect all ☐ Remove duplicate rows

Check	Field	Condition	Value		
✓	SUBJECT	Must be	Add value...		
✓	CHOL	Is between	Add value...	-	Add value...
✓	GLUC	Is between	Add value...	-	Add value...
✓	HDL	Is between	Add value...	-	Add value...
☑	DBD	Must be			
✓	TRIG	Is between	Add value...	-	Add value...

6. Specify the filtering criteria.
   a) Select a variable row to edit the criteria.
   b) In the ***Condition*** column, select the condition for the variable.
   c) Depending on the condition type selected, in the ***Value*** column, type values to variable text field or choose a value from the value drop-down menu.
   d) Repeat the previous steps for all variable rows whose criteria you want to edit.

NOTE:	Variables can also be removed from the subset by unselecting the check box in the ***Check*** column. By default all variable rows are selected.



Active Dataset

Select active dataset...

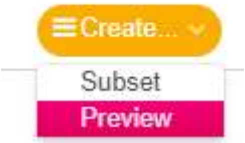
a - Demo phenotypes (bcdemo

Form name	Demo phenotypes
Form ID	DEMOPHENO1
Owner	bcdemo
Permissions	use
Families	DEMOPHENO1

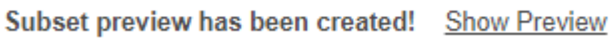
☑ Select/Unselect all ☐ Remove duplicate rows

Check	Field	Condition	Value		
✓	SUBJECT	Must be	Add value...		
✓	CHOL	is between	Add value...	–	Add value...
✓	GLUC	is between	Add value...	–	Add value...
✓	HDL	is between	Add value...	–	Add value...
☑	DBD	Must be	1 = Yes		
✓	TRIG	is between	Add value...	–	Add value...

7. Select **Create > Preview** to create a preview based on your filtering criteria



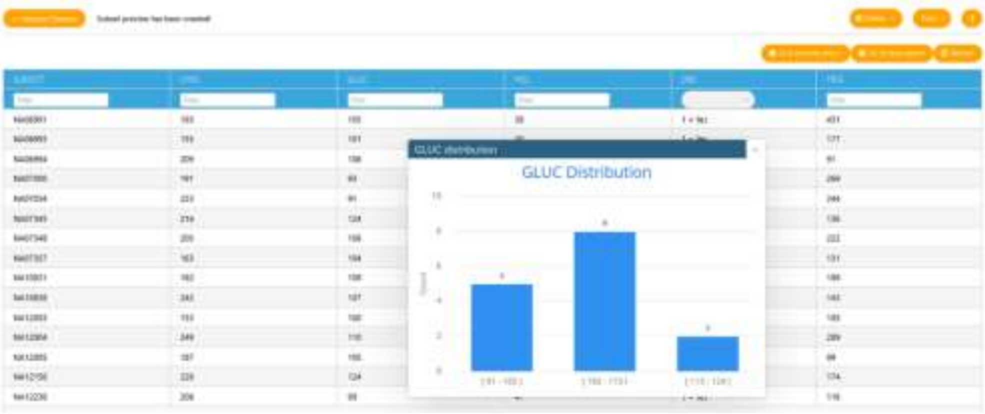
8. Select ***Show Preview*** to view the preview.



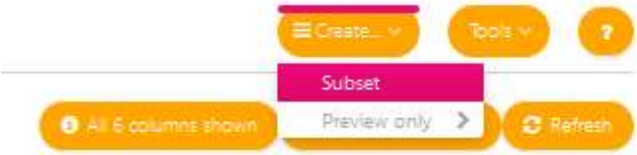
Subset preview has been created! Show Preview

9. View the subset in the grid view.

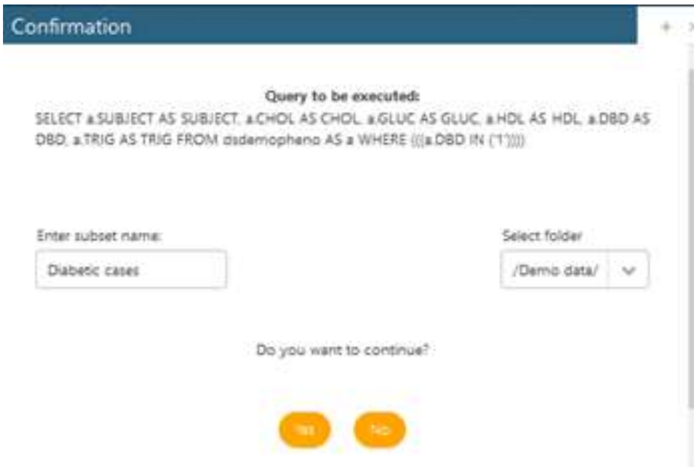
10. Press any non-key variable column to view the column value distribution.



11. Select ***Create > Subset*** to create the subset



12. Enter the subset name.
13. Select the folder where the subset is stored.
14. Select ***Yes*** to create a subset.



Confirmation

Query to be executed:

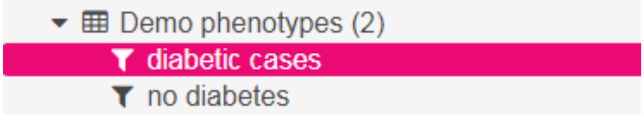
```
SELECT a.SUBJECT AS SUBJECT, a.CHOL AS CHOL, a.GLUC AS GLUC, a.HDL AS HDL, a.DBD AS
DBD, a.TRIG AS TRIG FROM dsdemopheno AS a WHERE (((a.DBD IN ('1'))))
```

Enter subset name: Diabetic cases

Select folder: /Demo data/

Do you want to continue?

15. View the subset in the navigation pane of either the subset tool or the navigation pane of BC|GENOME.



Demo phenotypes (2)
diabetic cases
no diabetes

5.2.3. Joining dataset information in a subset

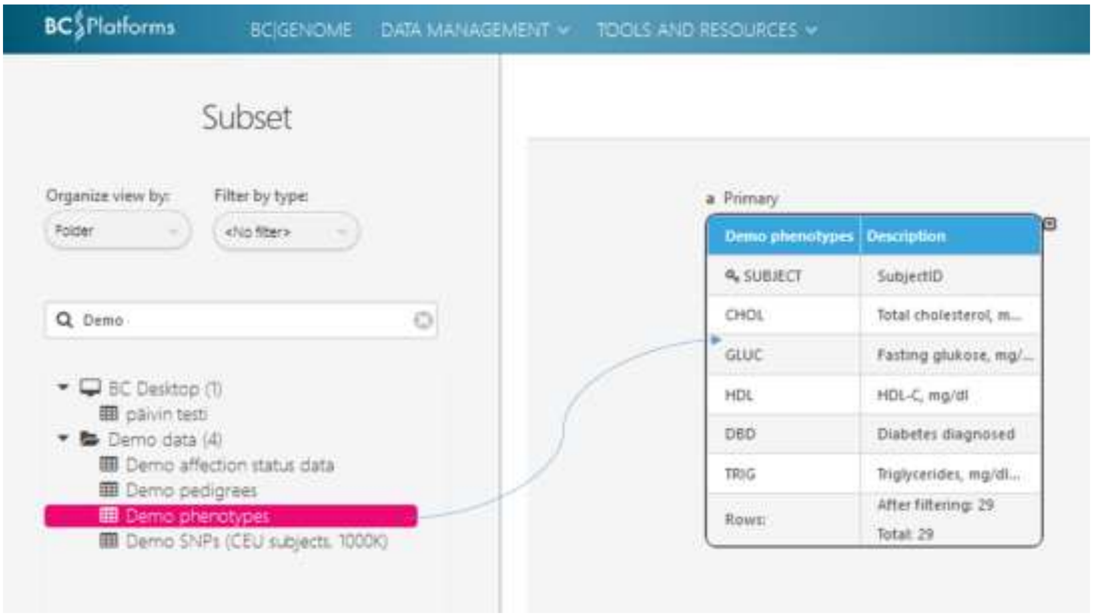
If a subset has at least one dataset, dragging another one to the canvas calls an automatic join candidate generator. When you join information from several datasets, the subset tool utilizes that join candidate generator, which suggests the joining variables based on the pre-set variable annotation information specified by BC support. The chosen option is shown in the canvas after you confirm your choice.

NOTE: You can join dataset to datasets, subsets to datasets, datasets to subsets, and subsets to subsets.

The speed of joining reduces when you join subsets to datasets/subsets as the query uses the original dataset to retrieve the information.

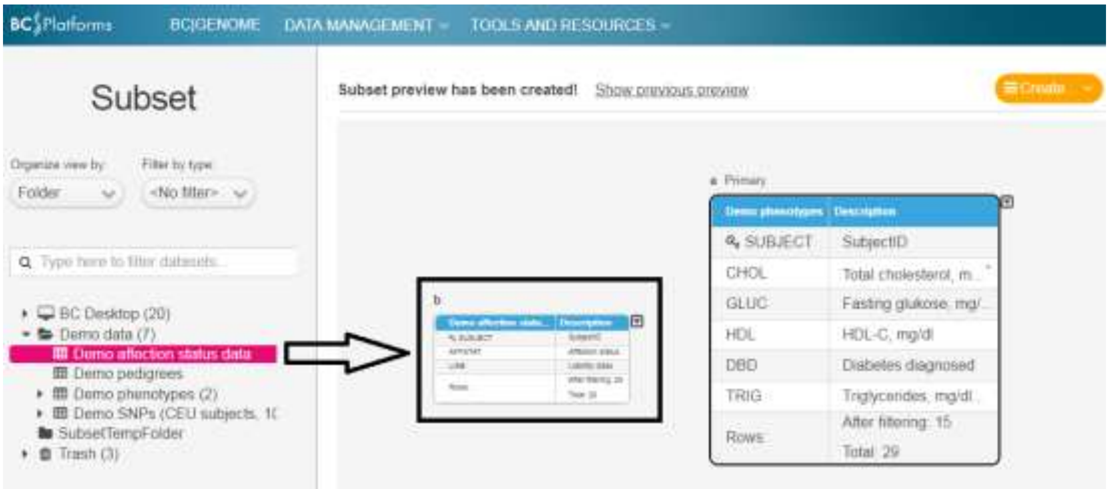
5.2.3.1. Joining information (except genotypes)

1. Drag and drop a dataset from the ***Subset*** navigation pane to the editor canvas. The dataset information is displayed as a table. This is the primary dataset.

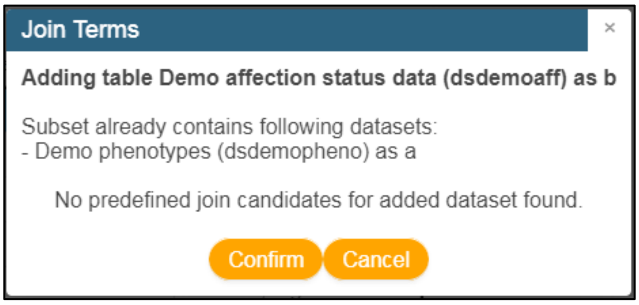


Demo phenotypes	Description
SUBJECT	SubjectID
CHOL	Total cholesterol, m...
GLUC	Fasting glukose, mg/...
HDL	HDL-C, mg/dl
DBD	Diabetes diagnosed
TRIG	Triglycerides, mg/dl...
Rows:	After filtering: 29 Total: 29

2. Drag the dataset you want to join with the primary dataset to the canvas.



3. A dialog displays with information about the join. Select ***Confirm***.

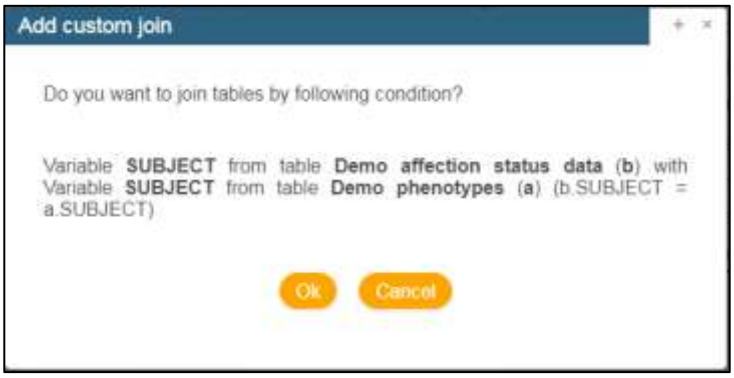


Adding table Demo affection status data (dsdemoaff) as b

Subset already contains following datasets:
- Demo phenotypes (dsdemopheno) as a

No predefined join candidates for added dataset found.

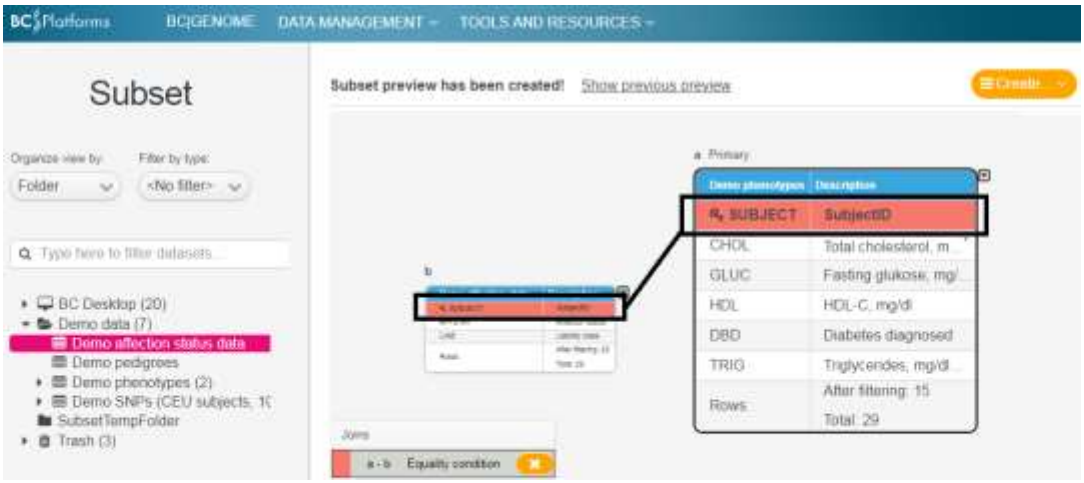
4. Select two fields you want to join, for example, SUBJECT.
5. Select ***OK***.



Do you want to join tables by following condition?

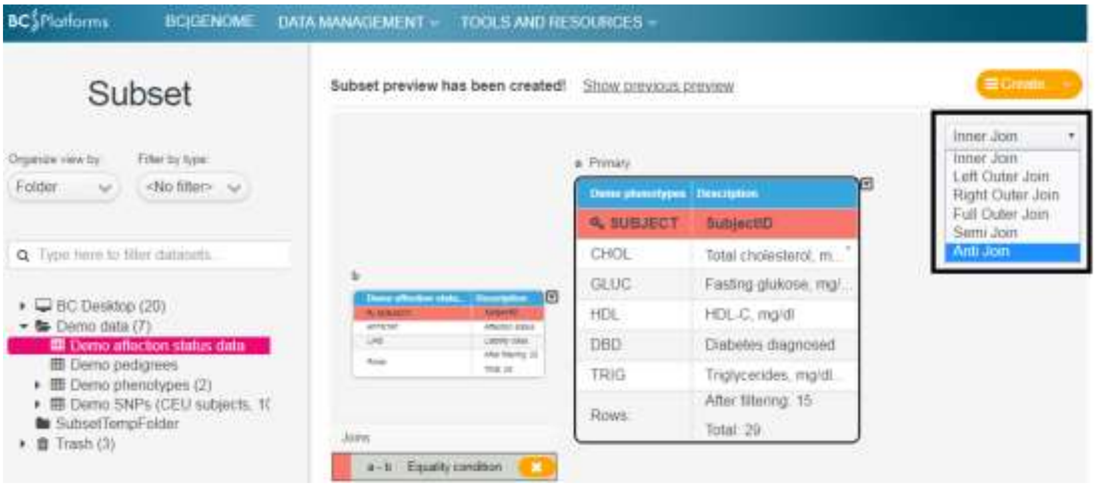
Variable **SUBJECT** from table **Demo affection status data (b)** with Variable **SUBJECT** from table **Demo phenotypes (a)** (b.SUBJECT = a.SUBJECT)

6. The joined fields are highlighted.



NOTE: You can make several different joins. Each separate join is highlighted.

7. Select the join type.



If the subset contains at least one join, a join type can be selected. The number of different join types depends on the number of datasets in the subset as well as conditions of joins.

- **Inner Join:** Retrieves dataset values that fulfil the join condition in both datasets. Inner join is always available.
- **Full Outer Join:** Retrieves dataset values that fulfil the join condition in both datasets. In addition, Full Outer Join generates values for cases where the condition is not met, Subby setting values of the other table to null. Full Outer Join is available for subsets with two datasets only, without namespace mapping join condition.
- **Left Outer Join:** Retrieves dataset values that fulfil the join condition from the primary (left) dataset. In addition, Left Outer Join retrieve values from the primary dataset that do not meet the condition by setting values from the secondary (right) dataset to null. Left Outer Join is available for subsets with two datasets only, without namespace mapping join condition.
- **Right Outer Join:** Retrieves dataset values that fulfil the join condition from the secondary (right) dataset. In addition, Right Outer Join retrieves values from the

secondary dataset that do not meet the condition by setting values from the primary (left) dataset to null. Right outer join is available for subsets with two datasets only, without namespace mapping join condition.

5.2.3.2. Subset > Tools

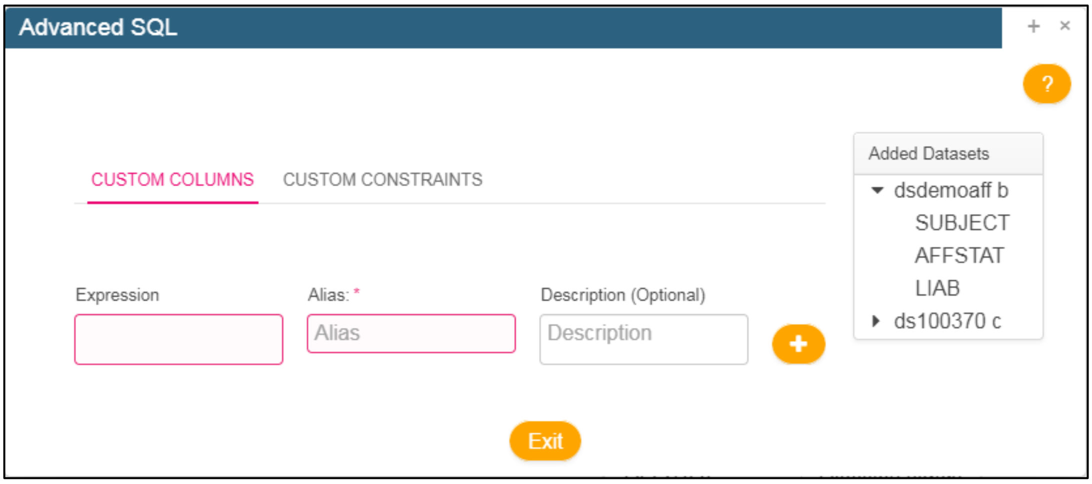
The following options are available in ***Subset > Tools***:

- ***Structure***

  This option shows all the fields in the joined form. It cannot be edited and is for information only.

- ***Advanced SQL***

  Advanced SQL enables you to input raw SQL into a subset query. You can create either custom variables/columns or constraints.



  – CUSTOM COLUMNS – Custom columns are used to add additional variables to a subset. To add a custom column, input the expression and alias into their respective fields. The expression must be pure SQL, resulting in the value for a new column. The column name is defined in the Alias field.

  – CUSTOM CONSTRAINTS – For more information about custom constraints, contact BC Support at support@bcplatforms.com.

5.3. Viewing the distribution of data values

BC|GENOME provides tools to visualize your data points in a dataset. The visualization tool can be found in the ***VISUALIZATON*** page when a dataset has been selected. The available graph types (Comparison, Trend, Correlation), depend on the selected dataset.

In charts, the maximum number of different data points is 1 million. If you try to visualize more than one million data point values, or the query takes more than 60 seconds, a timeout message is shown.

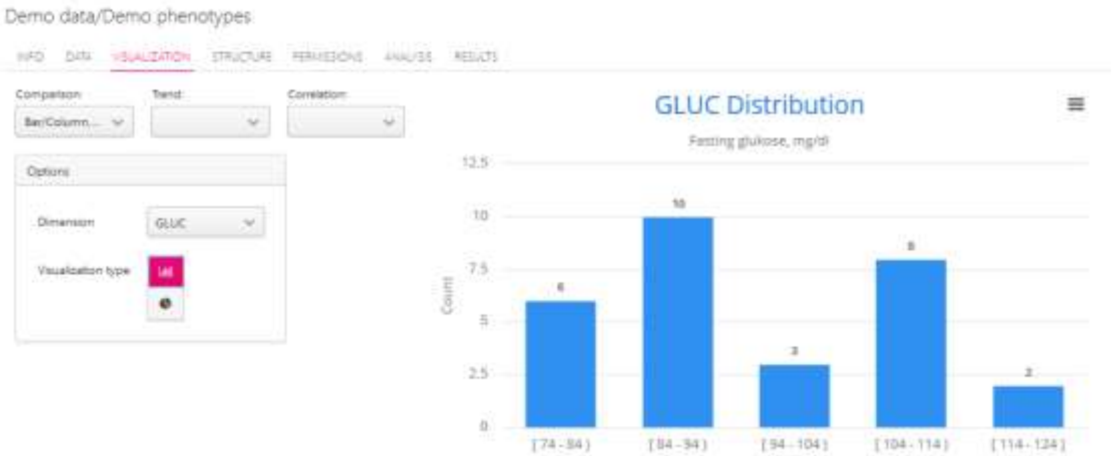
In the pie graph, up to ten sectors are used to visualize the data and the rest of the values are concatenated into a sector named **Others**.

To visualize your data:

1. Select a dataset.
2. Select the ***VISUALIZATION*** tab.
3. Select the graph type you want from:
   - Comparison
   - Trend
   - Correlation

NOTE:	Not all graph types may be available. The options depend on the dataset chosen.

4. Select the variable you wish to visualize.
5. Select ***Visualization type*** to create your chart/graph.

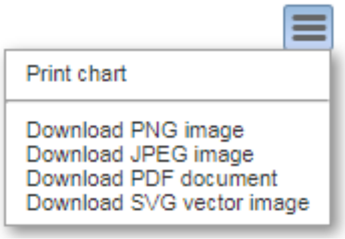


Demo data/Demo phenotypes

Comparison: Bar/Column...
Trend:
Correlation:

Options
Dimension: GLUC
Visualization type

6. Select the ≡ icon to choose the printing options.



5.4. Running analyses in BC|GENOME

BC|GENOME is provided with analysis drivers for running statistical analyses using data stored in datasets. Before you can submit an analysis job you need to make the following checks:

- Make sure that all datasets and subsets needed for analysis are stored in BC|GENOME.
- Make sure that *subject IDs* between phenotype and genotype datasets match.
- Make sure that there are no fields highlighted in red before you select ***Run***, or the analysis does not run.

- If you need to specify sex information for your analysis, the *variable ID* named ***SEX*** is specified either in a pedigree or phenotype dataset.
- For case and control analysis – Select datasets/subsets, one that specifies cases and one specifying controls.

  Or you can select (lower part of Affection Status) an affection status dataset/subset that specifies a set of cases (categorical value of 2) and controls (categorical value of 1). You must have *variable ID* named ***AFFSTAT*** in the affection status dataset.

For genotype analysis:

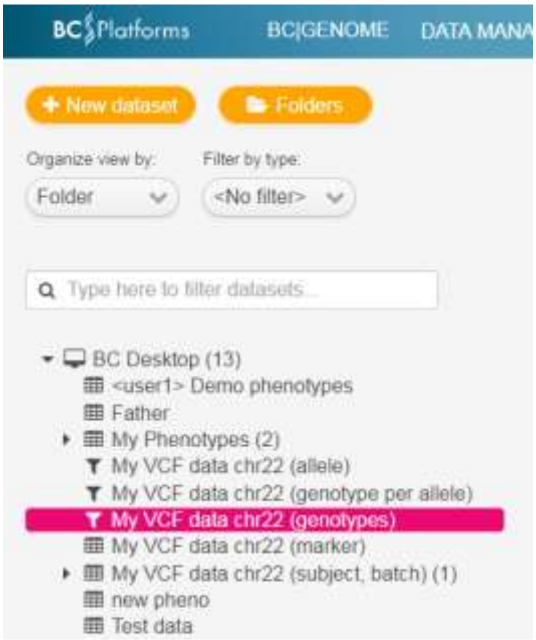
- Your genotype data contains *marker IDs* and you must have an annotation dataset with matching map information for *marker IDs*. When you have explicit *marker IDs,* physical map information has been stored in a format of chr:position (for example, 21:1234567).

IMPORTANT: BC|GENOME utilizes PLINK software for genotype analysis, which is designed to perform a range of basic, large-scale analyses. It is a third party software, so for PLINK-specific information, go to http://zzz.bwh.harvard.edu/plink/.

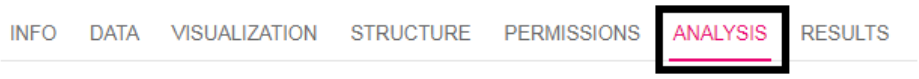
5.4.1. Submitting a genotype analysis job: example of PLINK case – control association analysis

The following outlines the steps to run the ***PLINK case – control analysis***.

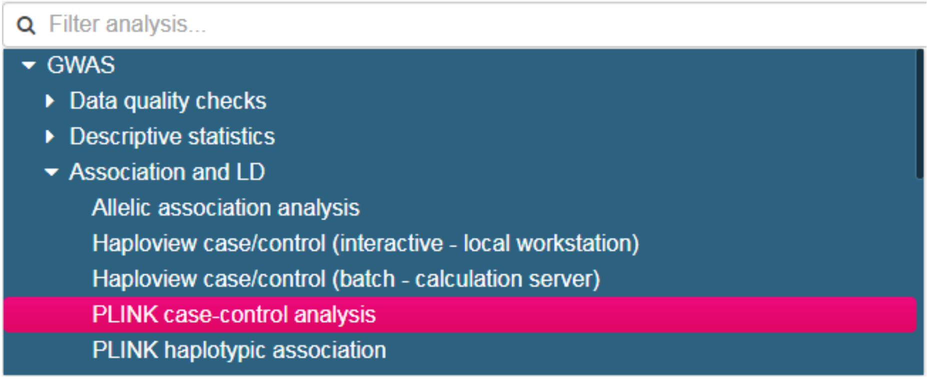
1. Select a genotype dataset/subset from the navigation pane.


BC Platforms
BC|GENOME
DATA MANA
+ New dataset
Folders
Organize view by:
Folder
Filter by type:
<No filter>
Type here to filter datasets...
BC Desktop (13)
<user1> Demo phenotypes
Father
My Phenotypes (2)
My VCF data chr22 (allele)
My VCF data chr22 (genotype per allele)
My VCF data chr22 (genotypes)
My VCF data chr22 (marker)
My VCF data chr22 (subject, batch) (1)
new pheno
Test data

2. Select the ***ANALYSIS*** tab.

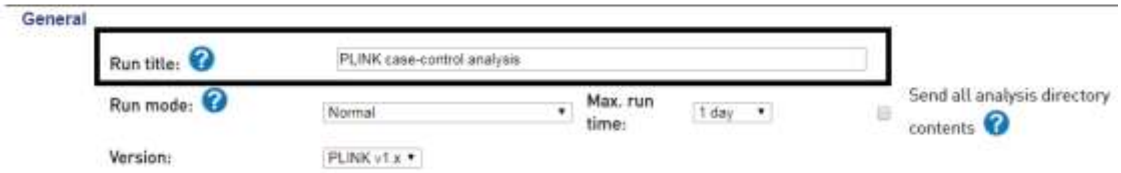


3. Select ***GWAS/Association and LD / PLINK case-control analysis***.


Filter analysis...
GWAS
Data quality checks
Descriptive statistics
Association and LD
Allelic association analysis
Haploview case/control (interactive - local workstation)
Haploview case/control (batch - calculation server)
PLINK case-control analysis
PLINK haplotypic association

The analysis form displays and you are ready to fill in the information needed to run your analysis.

4. GENERAL – Type a name for your analysis in the ***Run title*** field.



General

Run title: PLINK case-control analysis

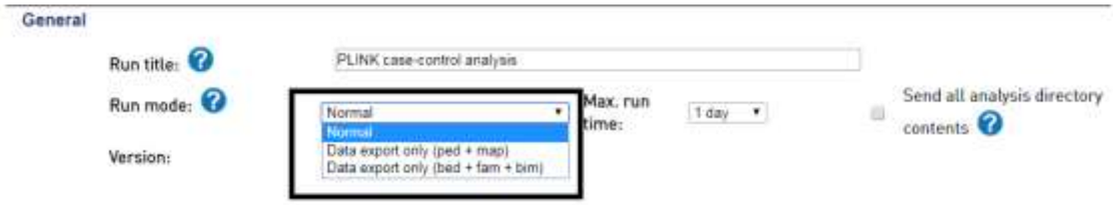
Run mode: Normal

Max. run time: 1 day

□ Send all analysis directory contents

Version: PLINK v1.x

5. GENERAL – Select the ***Run mode*** for your analysis from the drop-down menu.



General

Run title: PLINK case-control analysis

Run mode: Normal

Normal
Data export only (ped + map)
Data export only (bed + fam + bim)

Max. run time: 1 day

□ Send all analysis directory contents

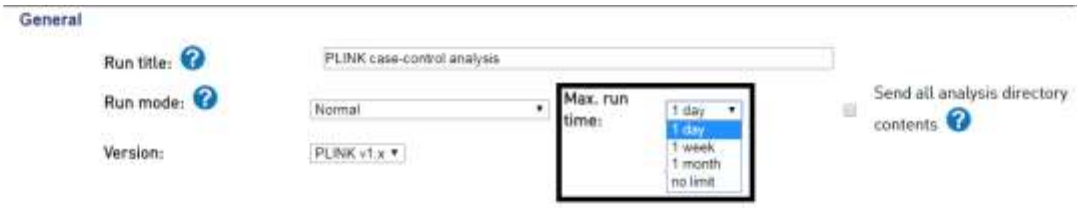
Version:

In ***Normal*** run mode, the analysis program is run and it produces normal numerical (and possibly graphical) output.

In ***Data export mode***, the system only generates the analysis input file(s) in the format of the analysis program, but does not actually run the analysis.

For more information, see the PLINK manual. Go to http://zzz.bwh.harvard.edu/plink/.

6. GENERAL – Select the ***Max. run time*** from the drop-down menu.



General

Run title: PLINK case-control analysis

Run mode: Normal

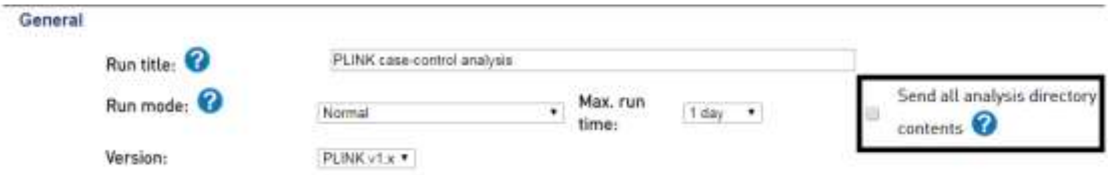
Max. run time: 1 day

□ Send all analysis directory contents

Version: PLINK v1.x

The maximum run time excludes both the time used for preparing the input file for the analysis software and the intergrated result reporting – so it is likely to take longer to run the analysis than the time you specify.

7. GENERAL – Select ***Send all analysis directories*** if you want to explore all the intermediate files that are created during the analysis management process.



General

Run title: PLINK case-control analysis

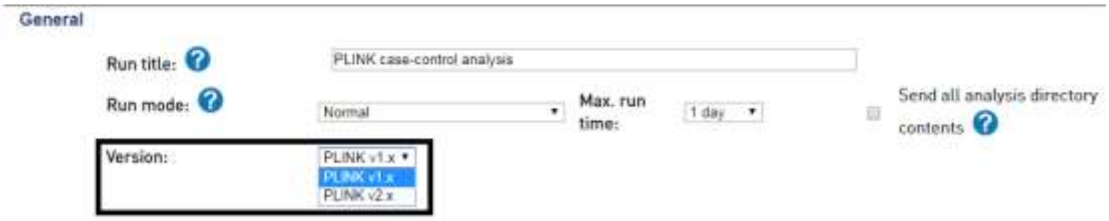
Run mode: Normal

Max. run time: 1 day

Version: PLINK v1.x

□ Send all analysis directory contents

8. GENERAL – Select PLINK ***Version*** from the drop-down menu (*v1.x* or *v2.x)*.



General

Run title: PLINK case-control analysis

Run mode: Normal

Max. run time: 1 day

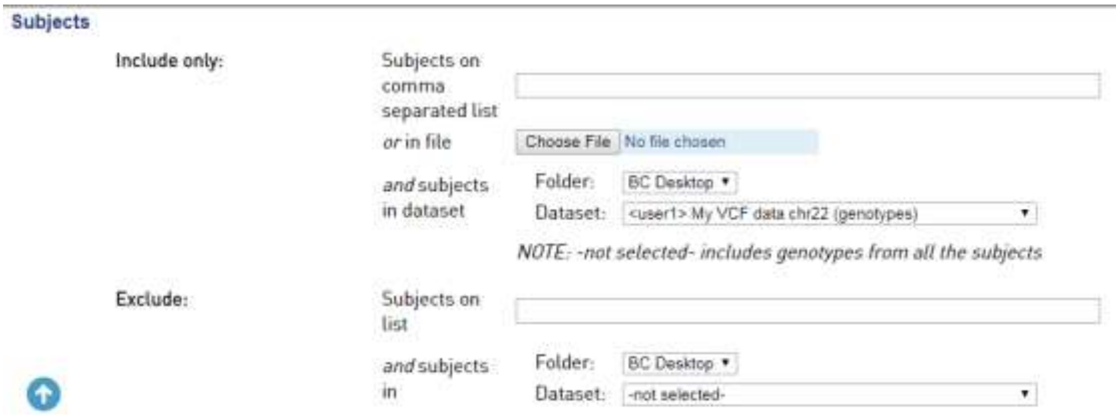
□ Send all analysis directory contents

Version: PLINK v1.x

For more information, see the PLINK manual. Go to http://zzz.bwh.harvard.edu/plink/.

9. SUBJECTS – Select ⬇ to expand the ***Subjects*** options.

10. SUBJECTS – Specify the information to include/exclude for your analysis.



Subjects

Include only: Subjects on comma separated list

or in file: Choose File No file chosen

and subjects in dataset: Folder: BC Desktop; Dataset: <user1> My VCF data chr22 (genotypes)

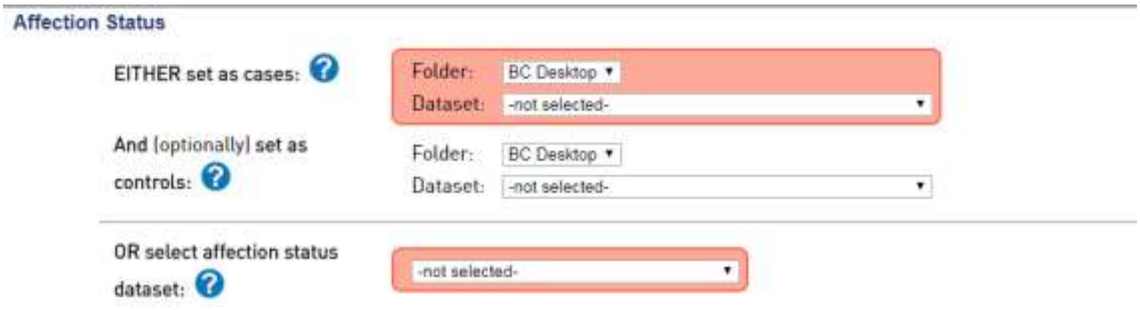
NOTE: -not selected- includes genotypes from all the subjects

Exclude: Subjects on list

and subjects in: Folder: BC Desktop; Dataset: -not selected-

11. AFFECTION STATUS – Select datasets/subset; one that specifies cases and one that specifies controls.

   Or you can select (lower part of ***Affection Status***) an affection status dataset/subset that specifies a set of cases (categorical value of 2) and controls (categorical value of 1). You must have *variable ID* named *AFFSTAT* in the affection status dataset.



Affection Status

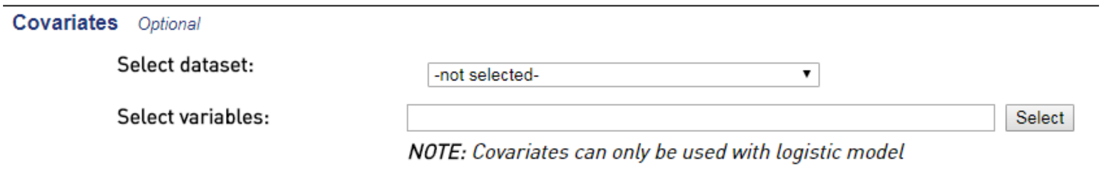
EITHER set as cases: Folder: BC Desktop; Dataset: -not selected-

And (optionally) set as controls: Folder: BC Desktop; Dataset: -not selected-

OR select affection status dataset: -not selected-

NOTE: You need to specify information for all fields shown in red before you are able to sumbit jobs to the queue system.

12. AFFECTION STATUS – COVARIATES - Specify optional variable(s) of a phenotype dataset used as covariates when calculating a logistic model analysis.



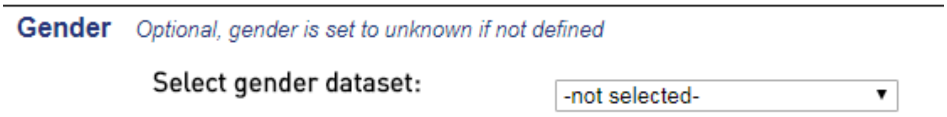
Covariates *Optional*

Select dataset: -not selected- ▾

Select variables: [ ] Select

NOTE: Covariates can only be used with logistic model

13. AFFECTION STATUS – GENDER - Select a phenotype or pedigree dataset with the *variable ID* named *SEX* to specify the biological sex of subjects.

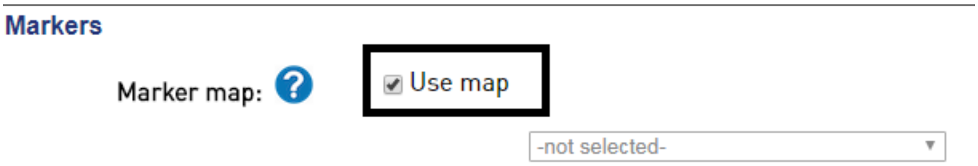


Gender *Optional, gender is set to unknown if not defined*

Select gender dataset: -not selected- ▾

14. MARKERS – Select ⬇ to expand the ***Markers*** options.

15. MARKERS – MARKER MAPS - Select ***Use map*** if you want to use marker position information in your analysis or you need to segment your analysis, by either chromosomes or the shorter, chromosomal loci.

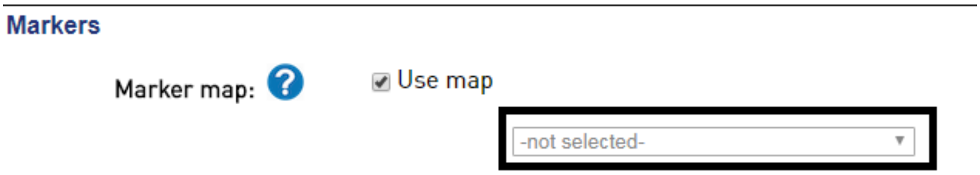


Markers

Marker map: ? ☑ Use map

-not selected- ▾

16. MARKERS – MARKER MAPS - Select a marker dataset that has the matching *marker IDs* to your *marker IDs* in the variant (genotype) dataset.

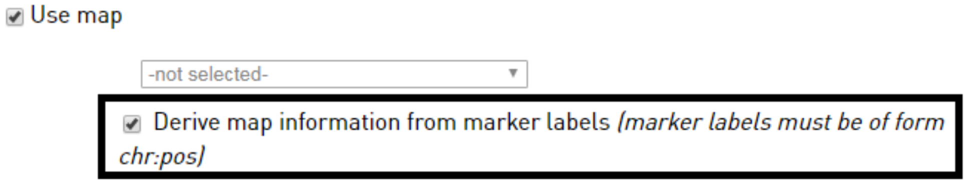


Markers

Marker map: ? ☑ Use map

-not selected- ▾

17. MARKERS – MARKER MAPS - Select the ***Derive map information from marker labels*** option if you have implicit *marker IDs* (in other words, chr:position, for example, *22:1234567*)

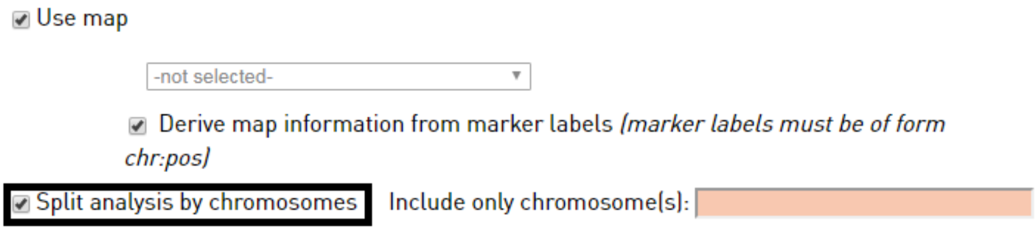


☑ Use map

-not selected- ▾

☑ Derive map information from marker labels *(marker labels must be of form chr:pos)*

18. MARKERS – MARKER MAPS - Select ***Split analysis by chromosomes*** so you are able to obtain chromosome-specific results, and to optimize the calculation capacity of your BC|GENOME system.



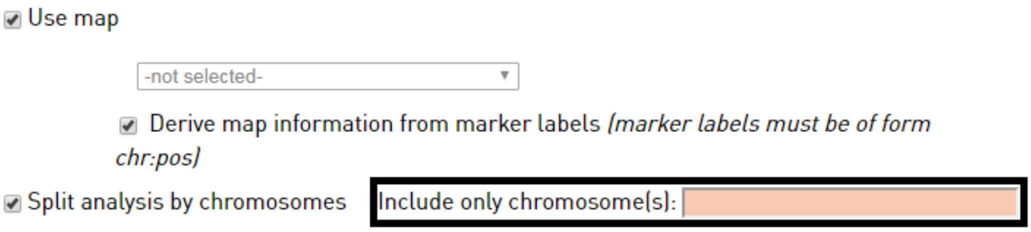
☑ Use map

-not selected- ▾

☑ Derive map information from marker labels *(marker labels must be of form chr:pos)*

☑ Split analysis by chromosomes   Include only chromosome(s):

19. MARKERS – MARKER MAPS – In the ***Include only chromosome(s) field***, specify the chromosomes included in the analysis:



☑ Use map

-not selected- ▾

☑ Derive map information from marker labels *(marker labels must be of form chr:pos)*

☑ Split analysis by chromosomes   Include only chromosome(s):

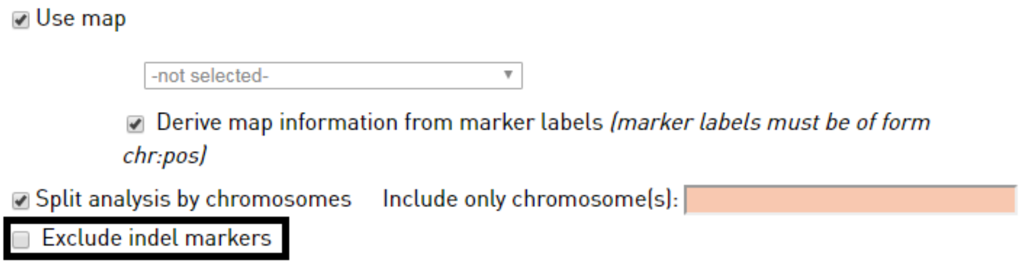
a) Specify, for example, human autosomes as *1-22*, or *1-2,5,22.*

b) Specify sex chromosomes as *X, Y* and *XY.*

c) Specify mitochondria as *MT.*

20. MARKERS – EXCLUDE - Select ***Exclude indel markers*** if you have indels specified that are different from either *I* and/or *D* allele codes.

NOTE:	PLINK does not support multi-allelic information.



☑ Use map

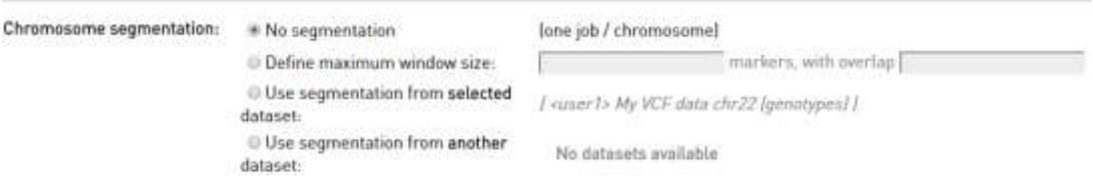
-not selected- ▾

☑ Derive map information from marker labels *(marker labels must be of form chr:pos)*

☑ Split analysis by chromosomes   Include only chromosome(s):

□ Exclude indel markers

21. MARKERS – CHROMOSOME SEGEMENTATION - Select *ONE* of the chromosome segmentation options:



Chromosome segmentation:
● No segmentation   (one job / chromosome)
○ Define maximum window size: markers, with overlap
○ Use segmentation from **selected** dataset: *[<user1> My VCF data chr22 (genotypes)]*
○ Use segmentation from **another** dataset: No datasets available

a) ***No segmentation*** – refers to a default case when segmentation is done by chromosomes.

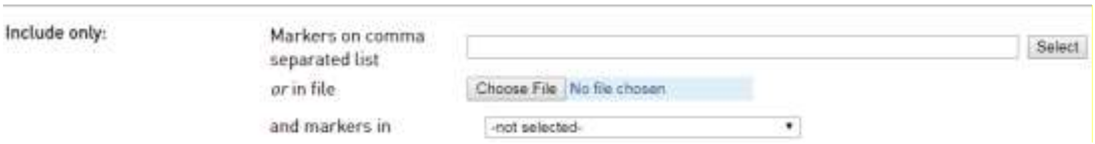
b) ***Define maximum window size*** – Chromosome segmentation can be set from 3 markers up to 10,000 markers with an overlap of ≥0

c) ***Use segmentation from selected dataset*** – utilize the segmentation specified in the input variation dataset.

d) ***Use segmentation from another dataset*** – use segmentation specified in another dataset.

NOTE:	The last 2 options cannot be used in the present version of BC|GENOME.

22. MARKERS – INCLUDE ONLY – Select markers to include either in a list of markers, text file or annotation dataset. Use a comma (,) to separate the list of markers to be included in the analysis. If this field is left empty, all markers are included.



Include only:

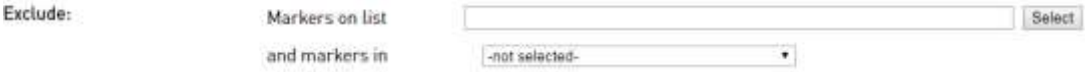
Markers on comma separated list

or in file: No file chosen

and markers in: -not selected-

CAUTION:	DO NOT press the ***Select*** button because this enables information from a marker dataset/subset to be specified where there are less than 1000 markers.

23. MARKERS – EXCLUDE – Use a comma (,) to separate the list of markers to be excluded from the analysis.



Exclude:

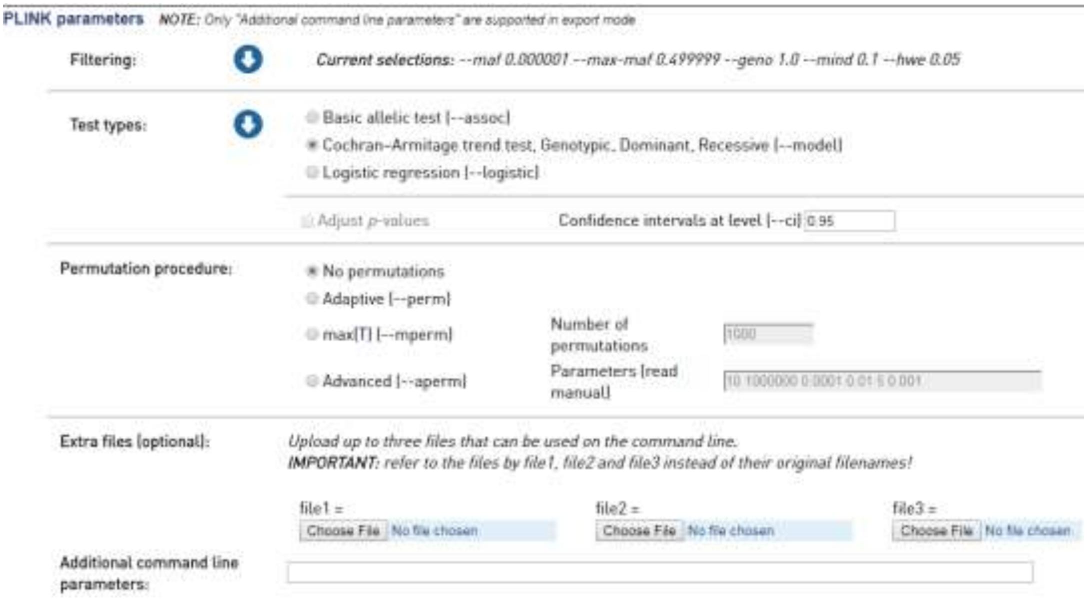
Markers on list

and markers in: -not selected-

24. ANNOTATIONS – If you want to annotate your markers by gene, contact BC support at support@bcplatforms.com.

25. PLINK PARAMETERS – Specify PLINK parameters as outlined in the PLINK manual at http://zzz.bwh.harvard.edu/plink/.

Select ⬇ to expand the ***Filtering*** and ***Test types*** options and make the required selections.



PLINK parameters *NOTE: Only "Additional command line parameters" are supported in export mode*

Filtering: ***Current selections:*** *--maf 0.000001 --max-maf 0.499999 --geno 1.0 --mind 0.1 --hwe 0.05*

Test types:
○ Basic allelic test [--assoc]
● Cochran-Armitage trend test, Genotypic, Dominant, Recessive [--model]
○ Logistic regression [--logistic]

□ Adjust *p*-values

Confidence intervals at level [--ci] 0.95

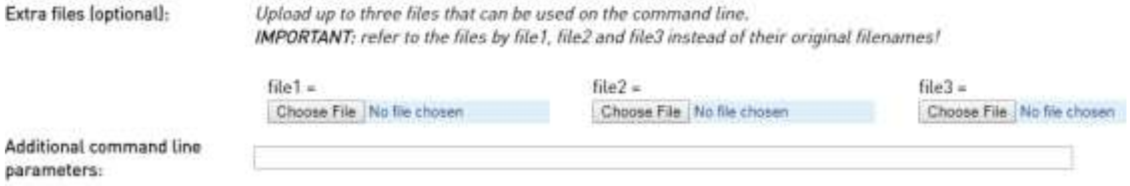
Permutation procedure:
● No permutations
○ Adaptive [--perm]
○ max(T) [--mperm] Number of permutations 1000
○ Advanced [--aperm] Parameters [read manual] 10 1000000 0.0001 0.01 5 0.001

Extra files (optional): *Upload up to three files that can be used on the command line.*
IMPORTANT: *refer to the files by file1, file2 and file3 instead of their original filenames!*

file1 = No file chosen
file2 = No file chosen
file3 = No file chosen

Additional command line parameters:

26. PLINK PARAMETERS – EXTRA FILES/ADDITIONAL COMMAND LINE PARAMETERS - Type additional command line parameters and auxiliary files from your computer when needed.



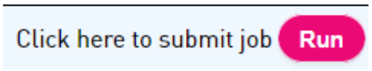
Extra files (optional): *Upload up to three files that can be used on the command line.*
***IMPORTANT:** refer to the files by file1, file2 and file3 instead of their original filenames!*

file1 = Choose File No file chosen
file2 = Choose File No file chosen
file3 = Choose File No file chosen

Additional command line parameters:

NOTE: Only PLINK-compatible files can be used. For more information about file compatibility, see the PLINK manual at http://zzz.bwh.harvard.edu/plink/.

27. Select ***Run*** when all analysis datasets and parameters has been specified.



Click here to submit job Run

28. When a message displays, select the queue link to follow the progress of the job in the BC|GENOME queue system

Job submitted to queue.

29. Go to the ***RESULTS*** page or **DATA MANAGEMENT > RESULT ARCHIVE**.

    For more information on results, see chapter 6.

5.5. Running IMPUTE2 within BC|GENOME

IMPUTE2 (see https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) is a program for statistically inferring unobserved genotypes, based on a set of known reference haplotypes. Running impute within BC|GENOME is automatically performed in segments of suitable size, such that many segments are imputed concurrently in the computation environment of the customer. After performing the imputations, resulting genotypes from the segments of each chromosome are combined in to a single result file, and files from multiple chromosomes are collected into a summary folder for easy upload to a BC|GENOME dataset (imputed genotyped need to be in a dataset to be later analyzed with BC|GENOME). Imputations can be started for a single chromosome or multiple chromosomes from a single GUI page. It is a good practice to first perform a small test run with a single chromosome (e.g. 22), or a segment of a chromosome, before starting imputations for the complete genome. In the following, we will provide detailed instructions on running Impute within BC|GENOME.

5.5.1. Prerequisites

- IMPUTE (version 2.2) must be installed to your server (BC Platforms cannot automatically download and install it due to licensing reasons). If it is not installed, please download the Impute program from Impute's web page: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#download_impute2, file "Linux (x86_64) Static Executable – v2.2.2". Copy the file to your upload folder using BC file transfer tool. After the file is copied, please inform BC support (support@bcplatforms.com), and we will configure it to be run on the server.

- Genotype data has been uploaded to a BC dataset.
- Genotype data can contain SNP markers and indels. Indels must be coded as I/D.
- The genotypes must be strand-aligned with your reference haplotype panel (typically the forward strand of the genome build). As impute automatically performs strand alignment for SNPs for markers where this can be done unambiguously (ones with alleles A/C, A/G, C/T or G/T), this is not required if the data does not have A/T and C/G markers.
- If X chromosome is to be Imputed, a pedigree dataset or a phenotype dataset with form "gender list" must exist to provide gender information for the imputed subjects.

5.5.2. Running Impute with basic arguments

In the following, we describe the minimal steps that are needed to run Impute within BC|GENOME. Following these steps should be enough for running Impute in most cases. The section *Optional features* contains details on options in the Impute GUI that can be used by advanced users if needed.

1. **Open the genotype dataset** that you want to Impute in the dataset navigator (this must be done before opening the Impute GUI page).
2. Open Impute GUI
1. Search for IMPUTE2 in the ANALYIS tab and click it open
3. **Give a "Run title",** to enable later recognizing different Impute runs in the result archive.
4. **Select a marker map** to specify coordinates for the markers in your genotype data.
1. The marker map must be based on the same genome build as the reference panel you're going to use in the imputation. Most current reference panels are based on NCBI genome build 37. The build is also visible for most marker maps, e.g. *NCBI dbSNP Build 135 (Nov 2011, hg 37.3*) and also in the reference panel name, e.g. *1000 Genomes Phase I integrated variant set v3 (March 2012, NCBI build* 37).
2. Probably a marker map corresponding to you genotyping chip is found in the set of pre-installed marker maps in BC|GENOME, or you have uploaded a marker map corresponding to the your genotyping chip; in these cases it is best to use these maps.
3. If no specific marker map is available for you genotypes, it is generally OK to use the most recent version of dbSNP marker map that is based on the same genome build as the reference panel. Depending on the genotyping chip, some markers may not be found in dbSNP. Usually the number of such markers is rather small, and this should not hamper imputation accuracy much.
4. Warning: the marker map corresponding to the reference panel should NOT be used in imputation run; that map is only to be used for analysis of imputed data.
5. **Select chromosomes to be analyzed** in the "Include only chromosome(s)" text box:

1. list of chromosomes is written as a comma-separated list of individual chromosomes, or chromosome ranges, e.g. "1,4-7". To impute the complete genome, write "1-22,X,XY".
2. Warning: imputation of the X chromosome imposes some special requirements on the used marker map, and also requires a input data set file for specifying the gender of the subjects to be imputed. Usually X chromosome imputation needs to be performed as a separate run with a dedicated marker map. For details, see section "X chromosome imputation" at the end of this document.

6. **Select reference panel** under "phased reference"

1. Note that the NCBI genome build of the reference panel must match the build of the selected marker map
2. Although older reference panels are divided to separate population-specific panels, in more recent reference panel versions the subjects from all populations have been combined into a single reference panel. This is sensible because Impute can automatically choose a subset of the reference panel haplotypes that best match the genotypes of each individual.
3. Reference panels distributed on IMPUTE2's web page (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) are pre-installed in BC|GENOME. We aim to add new versions of reference panels as soon as possible when they are published.
4. If a certain reference panel is not shown in the GUI, please contact BC support.
5. User's own reference panels can currently only be installed as customization work. Please contact BC support to request installing custom reference panels. In an upcoming version of BC|GENOME, there will be a possibility to upload reference panels.

6. **Start the Imputation** by clicking "Run" at the bottom of the page

1. If the panel containing the run button is shown in red, some parameters (also shown in red) will need to be modified before the run can be started.

7. Monitor the progress of the runs from the BC|SNP queue page

1. link to the queue page is available from the page that opens after starting the run
2. Alternatively, see DATA MANAGEMENT / QUEUE to view the progress of the run
3. Number of MCMC iterations completed / remaining is shown in the queue page, once imputation has progressed far enough. Typically, imputing each segment takes several hours, at the least

8. **Verify imputation results** from the result archive (after all segments have been completed)

1. Result archive is accessed from DATA MANAGEMENT / RESULT ARCHIVE
2. The result files are as follows:

1. *imputed_genotypes.bcos.txt.gz*: main result file containing imputed genotypes, converted to probabilistic BC format. This file can be directly uploaded to a BC|GENOME dataset (see next section).
2. *snpinfo:* info file produced by Impute, converted to BC format. This file can be uploaded to a BC marker info dataset (having form *IMPUTE SNPInfo*).

3. *MPUTE_summary*: summary file outputted by Impute. It is recommended to browse through this file to see that everything went as expected.
4. *markers_not_in_map.dat:* list of markers in the genotype dataset that were not found in the selected marker map. These are not used in imputation, and will not appear in the imputed data file (unless they are present in the reference panel, in which case naturally an imputed version of the genotype will be present in the result)
5. *skipped_monomorphic_markers.txt:* list of monomorphic markers in the genotype data that were not used for the imputation. As Impute requires both allele alternatives to be specified in the input data even when the marker is monomorphic, such markers cannot be included in input. The missing allele is deduced from the reference panel where possible, but this is not always possible, e.g. due to stranding errors.
6. *marker_name_conversion.bcos.txt:* Mapping of marker names in the genotyped data to the marker names used in the reference panel. Unless option "use genotyped marker names in result" is chosen, output 4 will use marker names of the reference panel, instead of the original names.
7. Note that when imputation is performed in segments (as is usually the case), the above-mentioned files from individual segments will be automatically combined to chromosome-wide files.

5.5.3. Uploading imputation results

In order to be used in analyses within BC|GENOME, the result files need to be uploaded from the result archive to a BC|GENOME dataset. (it is not possible to automatically upload imputation results to a dataset as part of the imputation job). Imputation results are typically stored in a probabilistic dataset, although they can also be stored to a deterministic dataset to save disk space, by just taking the most probable genotypes (naturally, information is lost in this conversion). It is also possible to upload the marker information produced by Impute, into a separate marker info dataset.

Uploading imputed genotypes from a single chromosome to a probabilistic dataset:
(Note, data in compresses dataset cannot be used for subsetting)

1. In the landing page of BC|GENOME use the New dataset button to create a new dataset using the form *Compressed imputed SNPs with 3-digit precision*.
2. Go to the result folder of the Impute job under the result archive.
3. Click on the blue upload arrow next to the file "imputed_genotypes.bcos.txt".
4. On the next page, select converter "Imputed SNPs with probability" and click "Continue".
5. On the next page, check the list of files to be uploaded and click "Upload" to schedule the upload job in the BC|GENOME queue.
6. Progress of the upload can be tracked in the BC|GENOME queue page; after the upload is complete, an upload report appears in the result archive.

Uploading imputed genotypes from multiple chromosomes to a probabilistic dataset:

1. In the landing page of BC|GENOME use the New dataset button to create a new dataset using the form *Compressed imputed SNPs with 3-digit precision*
2. Go to the result folder of the Impute job and go to the "summary" subfolder.
3. Click on the blue upload arrow next to one of the files having name like "chr22_imputed_genotypes.bcos.txt" (any chromosome can be chosen).
4. On the next page, select converter "Imputed SNPs with probability", replace the search string with "chr*_imputed_genotypes.bcos.txt" to match all chromosomes and click "Continue".
5. On the next page, check that files for all chromosomes are included in the list of files to be uploaded and click "Upload" to schedule the upload jobs in the BC|GENOME queue.
6. Progress of the upload can be tracked in the BC|GENOME queue page.
7. After the uploads are complete, an upload report appears in the result archive.

Uploading imputed genotypes to a deterministic dataset
Note, the subset functionalities have been supported with a deterministic dataset since genotypes have been stored into a database table
Similarly as for probabilistic data set above, but with following changes:

1. For the target dataset, use form "Compressed ACGT coded SNPs".
2. For the converter, select either "Imputed SNPs (uses most probable genotype)", or "Imputed SNPs (uses most probable genotype, min. quality 0.95)".

1. The first one always gets the most probable genotype, while the latter one leaves a genotype as missing when the probability of the most probable genotype is less than 0.95.

Uploading marker info
Similarly as for genotypes above, but with following changes:

1. create/select a marker dataset with form "IMPUTE SNPInfo"
2. Select "no conversion" as the converter

7. 3. Analysis of imputed genotypes

After importing to a dataset, imputed genotypes can be analyzed with analysis programs integrated to BC|GENOME. Using analysis programs designed for probabilistic genotypes, such as Probabel and SNPtest, is recommended for imputed data. However, also analysis programs using deterministic genotypes, such as PLINK, can be used, as probabilistic genotypes are automatically converted to deterministic ones by BC|GENOME when deterministic genotypes are expected, by taking the most probable genotypes.
For analyses requiring a marker map, there are two options:

1. **Use a pre-installed map corresponding to the reference panel.** There are pre-installed marker map datasets in the *1000 Genomes* folder corresponding to different versions of Impute reference panels. For instance, analysis of data imputed using reference panel "1000 Genomes Phase I integrated variant set v3 (March 2012)" should use map "1000

Genomes (Mar 2012) Impute map". Some older reference panels might not have a suitable map available; please contact BC support if a required map is not found.

2. **Use snpinfo dataset uploaded from the imputation results.** You can import the snpinfo files produced during the imputation into a BC format dataset with form "IMPUTE SNPInfo", and use this as the marker map dataset (see previous section for instructions on uploading the SNP info). Of these option 1) is preferred as less tedious. However, note that often the genotyped data contains at least some markers that have a different rs code (or other id) in the genotyped data and the reference panel. As Impute uses marker positions and alleles to map markers of genotyped data to markers of reference panel, such markers will always be used correctly in imputation. However, if option "Use genotyped marker names in result" (enabled by default in the GUI) is not used, these markers will have their original rs codes in the imputed data, and thus will not be found in the pre-installed map. In this case, using option 2) is required to provide the marker map where all markers are included.

5.5.4. Optional features

There are a number of other options available in the BC|GENOME impute GUI that do not necessarily need to be adjusted in typical imputation runs, but can be used by more advanced users if needed. These are described below.

Selecting regions to be imputed

- parameters "restrict imputation to range" and "impute full chromosomes"
- by default, imputation is restricted to the region covered by the genotyped markers.
- user can restrict imputation to given range (in bp coordinates).
- option "impute full chromosomes" can be used to impute all markers in reference panel, instead of just the genotyped range.

Segmentation of imputation jobs

- parameters "max. impution window size" and "use markers in flanking regions of "
- By default, imputation is done in windows of 5MB, with overlap of 250kb between neighboring windows. For very large datasets, it may be necessary to decrease the window size in order to prevent individual runs from using too much RAM memory / disk space. Even values as low as 1MB should not affect imputation accuracy.

Imputation mode

This option controls how impute is run. There are four modes of operation:

- **Impute using posterior probability distribution of haplotypes.** This is the "traditional mode of operation of Impute: consider a distribution of sampled haplotypes (using MCMC sampling) when doing the imputations, instead of using just the most probable haplotypes.
- **Impute using most probable haplotypes only.** In this mode, Impute is run twice: first phasing to estimate the most probable haplotypes and then imputation using these most probable haplotypes only. This mode is significantly faster than full MCMC estimation, but is slightly less accurate. This mode cannot be used for X chromosome imputation.

Impute using pre-phased haplotypes. Impute into pre-phased haplotypes instead of doing both phasing and imputation. This mode is only visible in the GUI when the selected genotype dataset is already phased and the input genotypes are stored into a dataset with form *imputed haplotypes* (unfortunately, the name of the form is a bit misleading, as no imputation has taken place yet). The program ShapeIT is recommended for pre-phasing by the authors of IMPUTE2. However, there is currently no support for pre-phasing haplotypes within BC|GENOME, thus please contact BC support for more information.

Phasing and imputation of sporadic missing data only. In this mode, only phasing and imputation of sporadic missing data is done. No reference panel is used. Note that phased haplotypes are outputted for each imputed segment separately, as there is no straightforward way of combining haplotypes from

separately phased segments into haplotypes for complete chromosomes (phase between markers of adjacent segments is not known). Also note that Haplotypes are outputted only in Impute format (and not in uploadable BC format), as haplotypes from individual segments cannot be stored into the same BC data set.

Imputation arguments

These options control Impute's MCMC algorithm and effective population size. The default values should be OK in most cases (for more documentation on these, see Impute's online manual: https://mathgen.stats.ox.ac.uk/impute/mcmc_options.html and https://mathgen.stats.ox.ac.uk/impute/output_file_options.html).

Output options

These options control the output from the software.

- **SNP types to be included in the output file (-os)** (default: output all SNP types). See IMPUTE's manual for details.
- **Store phasing results (-phase)** If checked, the phased genotypes are also outputted. Note that haplotypes are outputted for each segment individually, as they cannot be reliably be combined across segments.
- **Predict genotyped SNPs (-pgs)** If checked, impute also genotyped markers and replace their genotypes with the imputed ones in the result (if not checked, result will have the original genotypes for genotyped markers)
- **Use genotyped marker names in result (instead of names from reference panel)**. For genotyped markers, Impute normally outputs the original marker names into the output files.
    - However, by default the BC|GENOME impute driver converts the marker names already before imputation to match the ones in the reference panel, so that the marker names appearing in the imputed genotypes will correspond to the reference panel. This makes it possible to use a pre-installed marker map in subsequent analysis of Imputed data (see section "Analysis of Imputed genotypes" above). This option can be used to turn off this feature, so that the original marker names appear in the output file for genotyped markers.
    - **Output BC format results only (saves disk space)**. When this is checked, only uploadable BC format results are outputted (this is the default choice). If this is not checked, also the original result files of Impute are outputted. Note that the original impute format files also contain the markers on "flanking regions" of

each imputed segment, and thus markers in the edges of segments will be present in the result files of both adjacent segments.

Reference panel

- See subsection 6: Select reference panel in section Running impute with basic arguments above.
- Some reference data sets contain population-wise minimum allele frequencies, which can be used to filter out low-frequency variants from the reference haplotypes, so that only ones meeting a MAF threshold are included in the imputation. Note that such filtering cannot be performed when using older reference panels, where these frequencies are not included.

Strand-alignment of genotyped data
These options control how strand of genotyped markers is aligned with the reference haplotypes. Ideally, the input genotypes should be already strand-aligned with the reference haplotype panel (typically the forward strand of the genome build). This is not required, if the the data only contains markers with alleles that can be unambiguously strand-adjusted (A/C,G/T,A/G and C/T). Impute automatically adjusts strand in such markers (A/C <=> G/T and A/G<=>C/T). If option "Also align ambiguous alleles by MAF" is selected, Impute also adjusts strand in (ambiguous) A/T and C/G markers, based on allele frequencies. Note that this may not be accurate for markers with MAFs close to 0.5, and the authors of Impute do not recommend using this option. One can also upload a strand file to align the strand (see option -strand_g in IMPUTE2 documentation). The strand file need not contain all markers in the genotyped data; automatic strand correction (possibly including ambiguous markers, as described
above) is applied to any markers not contained in the strand file. Note that a strand file always applies to a single chromosome, so this option cannot be used when imputing multiple chromosomes as a single job. See "Strand alignment options" in IMPUTE's web page for more details.
Unphased reference
It is also possible to use unphased reference panels for imputation. Currently, these can only be installed as customization work. Please contact BC support to request installing custom reference panels.
Gender set
See next section "X chromosome imputation"
Additional command-line arguments
Here, you can write any command line options accepted by impute. Typically this should not be needed, as most commonly used options are already provided by the BC|GENOME impute driver based on the selections made by the user in the GUI. Note that for this reason, many options (such as ones governing input and output files, range to be imputed etc.) should not be given here, as they are already provided
automatically by the driver.

5.5.5. X chromosome imputation

To perform X chromosome imputation, a separate dataset needs to be specified that contains sex information (for other chromosomes, sex information is not needed). Sex information is

stored in a phenotype dataset, using form "gender list". When importing the data to the gender list, the gender
column must be coded as 1-male, 2-female.
The imputation of the X chromosome is done separately for the pseudo-autosomal (PAR) and non-pseudo-autosomal (nonPAR) regions of the X chromosome. In imputations done within BC|GENOME, these should be annotated as chromosome "X" and chromosome "XY", respectively in the chosen marker map. This poses some difficulties, as normally marker maps within BC do not make the distinction between the PAR and nonPAR regions, and markers on both are annotated as chr "X". If the
genotype data has a dedicated marker map where PAR regions are annotated as "XY", those maps can be used as is. However, the generic dbSNP maps cannot be used for X chromosome imputation, as there also PAR markers are annotated as "X". Therefore, there are separate dbSNP maps for chromosome imputation within BC|GENOME. For dbSNP version 135, this map is called "dbSNP Build 135 chr X imputation map (NCBI build 37), available in the dbSNP folder. This map is installed with BC|GENOME versions 3.6-04 and later. Please note that as a different map is required, imputations for chromosome X cannot thus normally be started as part of the same run as the rest of the genome, but a separate run for is needed instead.

5.5.6. Troubleshooting

Below, we list some of the most common problems encountered when running Impute within BC|GENOME.

- **Impute jobs are stuck in the job queue.** The most probable reason for this that impute is not installed on your server. See section "prerequisites" at the start of this document.
- **None or only few genotyped markers map to reference panel.** Probably the most common problem is using a marker map that is not based on the same genome build as the reference panel. Please check that the map and the panel are based on the same genome build.
- **Some of genotyped markers are not found in the marker map.** Another common error is using a map corresponding to the reference panel (e.g. "1000 Genomes (Mar2012) Impute map"). Some markers, especially on the pseudo-autosomal regions of chromosome X and HLE regions of chromosome 6 might be named differently in the genotyped data and the reference panels, and will thus be discarded in imputations using the 1000 genomes maps. These maps should only be used for analysis of imputation results and not for the imputations; dbSNP or a enotyping chip-specific map should be used instead.
- **Problems in X chromosome imputation.** For X chromosome imputation, the marker map needs to contain markers annotated in a certain way. For problems regarding X chromosome, see section "X chromosome imputation".
- **Imputations fail to running out of memory / disk space, or they do not use all available computational resources, or** BC|GENOME has a certain default limit to the number of imputations run simultaneously. The optimal number is always dependent on the properties of the data (mainly number of genotyped subjects and number of markers in the chosen reference panel), which dictate e.g. how much RAM is needed. If too many imputations are run concurrently, there is a risk of running out of memory, and if too few, resources are not utilized optimally. For example, imputation of a single 5MB segment

having 6000 subjects with the 1000 genomes march 2012 reference panels might take around 6GB of RAM and 17 hours of computation time on a relatively efficient processor. Naturally these numbers may vary a lot depending on marker density, segment size and other factors. If needed, please contact BC support to discuss the possibility of adjusting the number of concurrent imputations.

5.6. R script interface

5.6.1. Introduction

This chapter describes various ways for advanced users to utilize R script in the BC|GENOME product. This manual describes steps for writing, storing and sharing scripts with other users of the system, and some practical recommendations and instructions for how to best access data, and manage the R script tasks.

The document also provides some examples of working R scripts which can be tested using the various data sets in the "Demo data" -folder of any BC|GENOME installation. It is assumed that the reader has expertise in writing scripts or is otherwise familiar with R. As such, this manual only offers practical notes related to running R scripts with the BC|GENOME system and is not a manual to the R script language.

5.6.1.1. Data input: Data from SQL tables

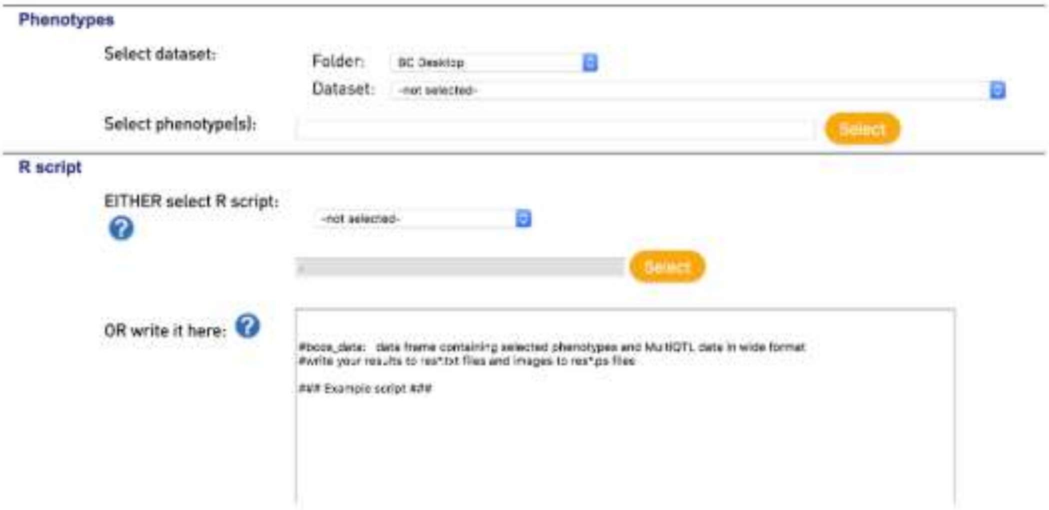
User can select the input data for R scripts from the database using the BC|GENOME user interface. Based on user’s selections BC|GENOME generates the data frames described in the following chapters. Data frames are matrices with chosen variables (QTs or phenotypes) in columns, and subjects or samples in rows.

Data frames are slightly different for BC|GENOME Genotypes and Phenotypes data types. Genotype scripts process genotypic data alongside with relevant map, pedigree, affection status, and / or phenotype data. Phenotype scripts process in a generic way any data in the context data frame, and attached files. Phenotype scripts are default fallback for any datatypes in the system that might not have a dedicated R script template available. It is possible to combine multiple data frames from other datasets for a Phenotype R script.

Image 1: Genotype R script analysis interface gives user the choice to attach relevent data to the script as data frames. Selections in this interface are then exported by the R engine and can be used from within the script. The available selection of extra data depends on the type of R script template that is chosen.

Image 2: Phenotype analysis GUI for R allows up to 4 external datasets to be imported into the script as data frames, making Phenotype R interface probably the most flexible of all.

MultiQTL data frames are formed based on user's choices in the R interface. User can select phenotype columns from the multiQTL dataset or other phenotype datasets. BC|GENOME generates the data frame for this data by using the subject identifier field (annotated BC_VARCLASS:patient or BC:subject) to join data.



```
#bcos_data:  data frame containing selected phenotypes and MultiQTL data in wide format
#write your results to res*.txt files and images to res*.ps files

### Example script ###
```

Image: MultiQTL analysis interface mimics genotype interface in the sense that it also allows user to select specific datasets and components from those datasets. However it is only limited to phenotypes.

5.6.1.2. Data frames: Genotypes

Genotype scripts analyze genotype data with associated map, pedigree, affection, and / or phenotype data. For genotype scripts two data frames are provided by the system:

bcos_map (marker map)

bcos_data (rest of the data)

Map frame (bcos_map)

Map frame bcos_map contains marker map data. Rows and columns are as in the BC|GENOME marker map dataset. For example:

Marker	**Distance**	**Chrom**	**Order**
Marker	**Distance**	**Chrom**	**Order**
RS0000	0	11	0
RS1111	100000	11	1
RS2222	200000	11	2
RS3333	300000	11	3

Genotype frame (bcos_data)

The basic order of the data columns within bcos_data frame is:

1. Family data (in case no pedigree data set selected, subject ID only)
2. Affection status / phenotype data (optional)
3. Allele data

An example of bcos_data with affection statuses defined:

SUBJECT	**AFFSTAT**	**RS0000_chr11**	**RS1111_chr11**
SAMPLE0	1	2	0
SAMPLE1	1	1	1

An example of bcos_data with pedigree data:

SUBJECT	**PED**	**FATHER**	**MOTHER**	**SEX**	**AGE**	**RS1111_chr11**
1000	1	0	0	2	38	0
1001	1	0	0	1	42	1
1002	1	1001	1000	1	12	1

NOTE: There is only one column per each genotype. The alleles are 0-1-2 coded; where 0 means alleles 11 in the original dataset, 1 mean alleles 12 and 2 means alleles 22. In the user interface, user can select whether the minor allele is coded as 1 (i.e. 0 means homozygous with the minor allele) or 2.

NOTE: A parameter bcos_first_snp defines the column where the genotype data begins. For example for the example data frames above the bcos_first_snp values would be 2 and 6.

5.6.1.3. Imputed data: Data frames

If the genotype datasets used in the R analysis contains imputed data (i.e. the dataset type is Compressed imputed SNPs), user can select whether to use most probable genotypes or probabilistic genotype data. If most probable genotypes are used, the data frames provided are the same as for normal genotype data. If probabilistic data is used, the following data frames are provided:

- bcos_prob (probabilistic genotypes)
- bcos_pheno (phenotypes / pedigrees / affection status data)
- bcos_map (marker map)

Genotype frame (bcos_prob)

Data frame bcos_prob contains the genotypes as probabilities.

An example of bcos_prob :

Marker	Allele_A	Allele_B	1000_AA	1000_AB	1000_BB
RS1000057	C	G	0.001	0.933	0.066
RS1000081	A	G	0.999	0.001	0.000
RS1000113	C	T	0.997	0.003	0.000

All probabilities are 0.000 for missing genotypes.

NOTE: Also all subjects that belong to selected pedigrees are included in the bcos_prob frame even if they do not have any genotype data.

Phenotype frame (bcos_pheno)

The basic order of the data columns within bcos_data frame is:

1. Family data (in case no pedigree data set selected, subject ID only)
2. Affectionstatus/phenotypedata(optional)

An example of bcos_pheno with only affection statuses defined:

SUBJECT	AFFSTAT
SAMPLE0	1
SAMPLE1	2

An example of bcos_pheno with pedigree data and one phenotype variable:

SUBJECT	PED	FATHER	MOTHER	SEX	AGE
1000	1	0	0	2	38
1001	1	0	0	1	42
1002	1	1001	1000	1	12

5.6.1.4. Omics / MultiQTL data: data frames

Data in multiQTL and omics datasets can be analyzed as such or together with phenotype data. The data frame contains the multiQTL and omics data in matrix format (row=subject, col=variable).

An example of bcos_data multiQTL data frame without phenotype data:

SUBJECT	A2A1	BCL2L1	BCL2L10	BCL3
NA06985	169.197	301.341	143.002	895.921
NA06991	638.602	338.372	41.711	129.343
NA06993	161.296	659.965	385.962	780.335

An example of bcos_data multiQTL data frame with phenotype data:

SUBJECT	SEX	A2A1	BCL2L1	BCL2L10	BCL3
NA06985	1	169.197	301.341	143.002	895.921
NA06991	2	638.602	338.372	41.711	129.343
NA06993	2	161.296	659.965	385.962	780.335

5.6.1.5. Phenotypes: Data frames

Phenotype scripts can be used to analyze only phenotype data. If you need to combine genotypic or omics data into the analysis, design the script to be launched from the genotype or omics dataset analysis tool. By default, phenotype scripts have input data in a data frame called bcos_data, containing data in rows and columns as in the BC|GENOME phenotype dataset.

An example of bcos_data phenotype data frame: In phenotype scripts it is also possible to bypass the default frame by defining one's own in the script like in the example below:

my_data = read.delim("^$INFILE$^", row.names=0)

Here the token "^$INFILE$^" is replaced by the actual data file name. All custom data frame definitions must contain the token "^$INFILE$^".

SUBJECT	CHOL	GLUC	HDL	DBD	TRIG
NA06991	193	105	38	1	431
NA06993	155	101	30	1	177

NA06994	209	106	63	1	91

Phenotype data with two keys

For example in case of a follow-up study, user will have to deal with two-index data where there's both a subject ID and a date corresponding to each particular set of measurements.

An example data frame for two-index data:

SUBJECT	VISIT	AGE	SEX	SCR
1000	2005-12-30	50	1	10.0
1000	2006-11-10	51	1	9.2
1001	2003-9-21	30	2	7.6
1001	2004-5-21	31	2	9.0

5.6.2. R – Data output

Output files from the R scripts are delivered to the user's result archive, when the calculation is finished. System automatically sends all data printed to stdout and stderr to user. By default all files created by the script are cleaned upon exit, except output files beginning with the string "res". Therefore temporary files can be used without worrying about cleaning, as long as the file names do not start with "res". The possible output file types are listed and explained in Table 1.

Table 1. Output file types and their content.

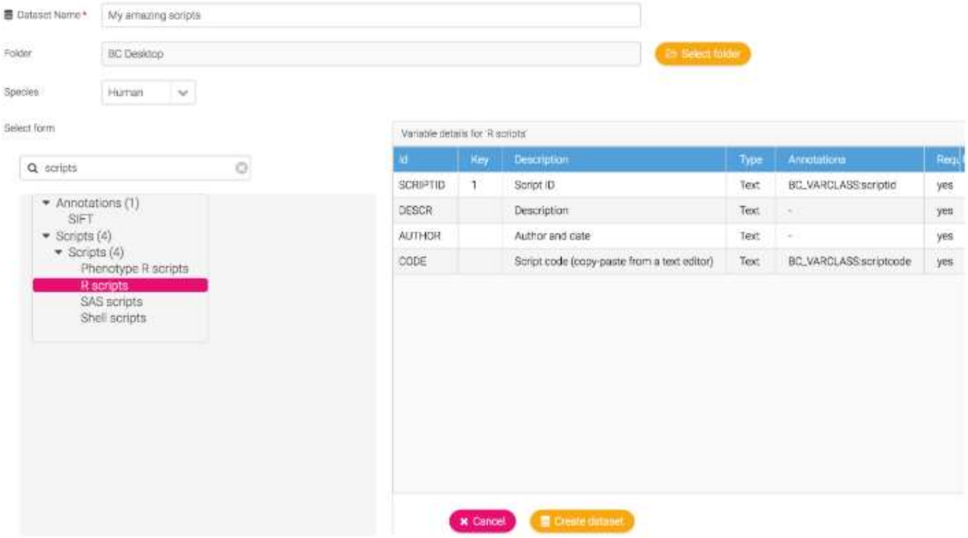
FILE	EXPLANATION
res*.txt	Text files are returned to the user, and if the run is segmented, the files produced in each segment are concatenated into a single file upon delivery.
res*.ps	Postscript files are converted to PDFs and returned to the user. If the run is segmented, the PDFs from each segment are delivered together in one ZIP file.
res*.dat	DAT files behave the same way as res*.txt files, but they provide a way to create a separate file association in the browser e.g. notepad for small reports in *.txt files and MySuperStatSoftware for large *.dat files.
res*.bcos	BCOS files are assumed to be tab delimited files with a header row. In segmented runs they are concatenated so that the first row (header) is stripped from all except the first segment, thus producing a single clean tab delimited file with a header row.

5.6.3. Storing, sharing and running R scripts

5.6.3.1. Storing R scripts

It is strongly recommended that all scripts are thoroughly tested either from the shell, or from the R analysis GUI, before scripts are stored in R script datasets and shared with other users.

1. Create a new dataset using the form R scripts or Phenotype R scripts (Image 1)
2. Go to the 'Data' tab of the newly created dataset and choose Add -tool
3. Give the script a unique ID, a clear description of what it does and what data fields it requires
4. Fill in author name
5. Copy paste the script to the CODE -field
6. Save the entry



Dataset Name* My amazing scripts

Folder BC Desktop

Species Human

Select form

scripts

Variable details for 'R scripts'

Id	Key	Description	Type	Annotations	Requ
SCRIPTID	1	Script ID	Text	BC_VARCLASS:scriptid	yes
DESCR		Description	Text	-	yes
AUTHOR		Author and cate	Text	-	yes
CODE		Script code (copy-paste from a text editor)	Text	BC_VARCLASS:scriptcode	yes

Image 1. Creating a script dataset to store and share R scripts.

5.6.3.2. Sharing scripts

1. In the scripts dataset open the Permissions -tab
2. Grant permissions to other users or user groups, as you would normally grant them to datasets
3. Permissions affect the use of scripts and script storage in following ways:

a. All permissions: Users are able to run all scripts in dataset, modify them, and add new ones
b. Read only: Users are able to run the scripts in the dataset
c. Write only: Users are able to store new scripts in the dataset

Some user roles are restricted in ways that prevent them from adding or modifying R scripts, independent of the permissions they have to script storage.

Note that it is possible to create subsets of the script storage as with any other dataset, and grant permissions in those subsets.

5.6.4. Running scripts

1. Go to BC|GENOME Data navigator and open the dataset you wish to analyse
2. Go to Analysis tab and select the relevant R analysis option from R-script folder of analyses (Image 2)
3. Select input data for the script as for any other analysis program. Required fields in analysis GUI are marked by a star.
4. Select a script you want to use from the database (Image 3), or write your own script in the area reserved for it (useful for testing scripts).

R jobs are handled by the BC|GENOME queue system like any other analysis task, and when the calculation is finished, the results can be found in the result archive.

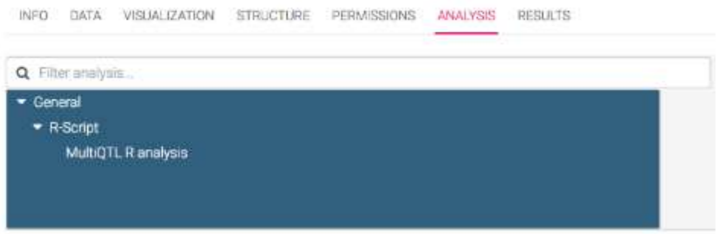
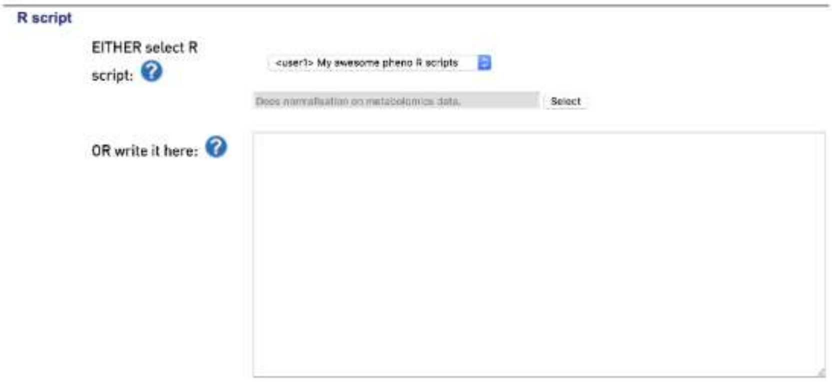


Image 2. Selecting the R -script analysis from the Analysis tab.


R script

EITHER select R script: <user1> My awesome pheno R scripts

Does normalisation on metabolomics data. Select

OR write it here:

Image 2. Selecting the R -script analysis from the Analysis tab.

5.6.5. External R libraries

It is often necessary to use R libraries and packages that are not part of the normal R distribution. In BC|GENOME the system uses the default R package, which is accessible to all BC system-level users, including those taking care of calculation tasks. Therefore it is recommended to strictly follow the instructions for you own environment and OS in installing new packages. The most commonly used and probably the least error-prone is to install new libraries through the R shell, as is described for example here https://www.r-bloggers.com/installing-r-packages/.

This approach, however, requires that the installation is done using root privileges, otherwise the new binaries will not be visible to the system-level user accounts that run the analyses. If you cannot use root privileges, please contact BC support for more help for configuring your calculation environment to have access to these libraries.

5.6.6. R script examples

5.6.6.1. Genotypes

Can be stored in R script dataset that uses the basic format "R scripts".

The following script executes QTL analysis on ACGT coded SNP dataset, taking the marker map and phenotypic data from the GUI selections the user makes.

R script for quantitative traits

```
#Resultfile titles
resdata="results.txt"
resfig="results.ps"

#Result titles
write(c("TRAIT","MARKER","BINTC","BSLOPE","BINTC_SD","BSLOPE_SD","BINTC_T","BSL


#Image settings
postscript(file=resfig,horizontal=FALSE,pointsize=5)
par(mfrow=c(4,3),omi=c(1,0.5,0.5,1))

#Loop traits
for(trait in 1:(bcos_first_snp-1))
{
  #Loop SNPs
  for(snp in bcos_first_snp:length(bcos_data))
   {
     #Calculate regression
     fm=lm(bcos_data[[trait]]~bcos_data[[snp]])
     res=summary(fm)

     trait_name=attr(bcos_data,"names")[trait]
     marker=attr(bcos_data,"names")[snp]
     results=c(trait_name,marker,res[[4]],res[[9]])

     #Draw image if F>5.
     if(res[[10]][1]>5)
      {
        boxplot(bcos_data[[trait]]~bcos_data[[snp]],ylab=trait_name,xlab=marker
      }

     #Write results
     write(results,file=resdata,append=TRUE,ncolumns=11,sep="\t")
   }
}
dev.off()
```

5.6.6.2. How to access FILE variables

IMPORTANT: The BC|GENOME database file storage or 'blob storage' must be configured to use so called BCFS (BC virtual file system), in order for the R interface to have legitimate access to the files. If your system already uses external file storage or cloud storages for archiving and accessing files, your BC|GENOME instance is configured in this way. If this is not the case, please contact your system administrator for more information about the possibility of making this configuration.

If you create scripts that need to access FILE type variables (like list of BAMs, omics data files, etc) within the dataset, you need to save the R script as 'Phenotype R script', or run the script from Phenotype R script interface. At the moment BC systems do not support FILE access for R scripts in genotype or NGS contexts. This will be amended in future versions.

R script accessing FILEs

```
# bcos_data:   data frame containing the phenotype data in wide format
# write your results to res*.txt or res.data files and images to res*.ps files
#
# Make a matrix out of dataframe
# and get the dimensions of the dataset

bcos_data_tab = as.matrix(bcos_data)
```

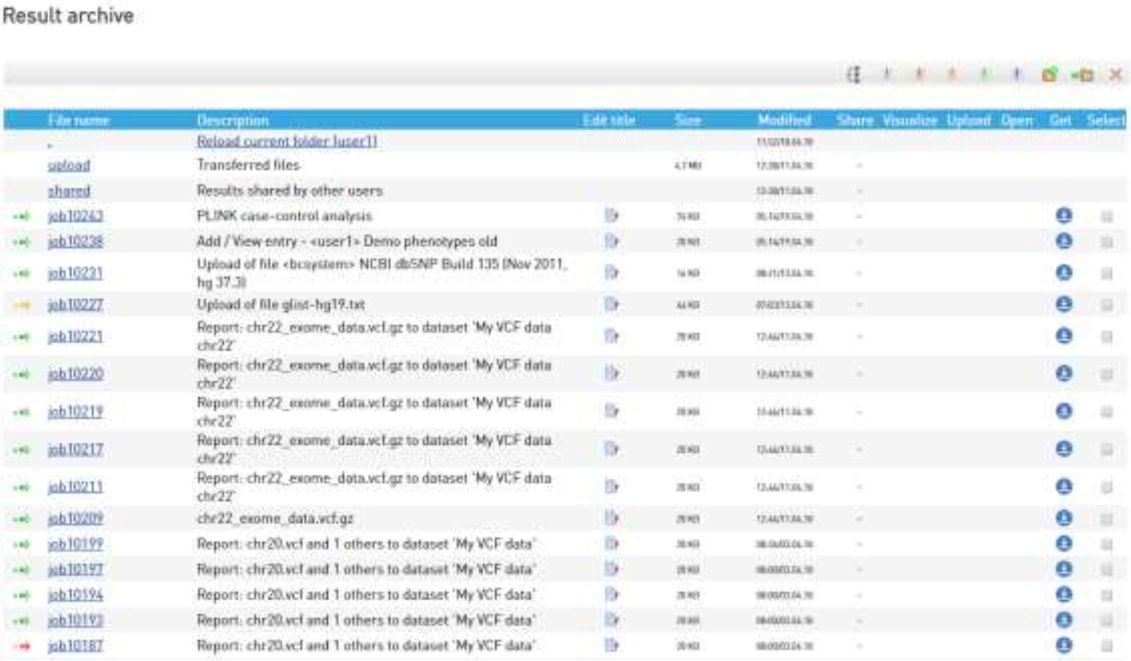
```
dim (bcos_data_tab)



# Get the index of the FILE variable you want to open
fidx = grep("^SOMEFILE$", colnames(bcos_data_tab))
fidx

# Read the content of the file on row 1, and print it
cont1 = read.delim(bcos_data_tab[1, fidx], quote="", header=FALSE)
cont1




# In case you have specified other datasets to be used in this script,
# those can be accessed from "data1", "data2", "data3", and "data4" variables
#
#data1_tab=as.matrix (read.delim("data1", quote="", sep="\t", header=TRUE))
#dim (data1_tab)
#
```

6. Result archive

BC|Genome Result Archive is an interactive way to monitor and handle completed jobs and their report files. When an analysis is run, the generated report folder is given a file name, *jobXXXXXX*.

The result archive contains a list of job folders (each contains a link to the report) and information about each job folders.

In addition, there are other folders listed in the result archive, for example, ***upload***, ***shared***.



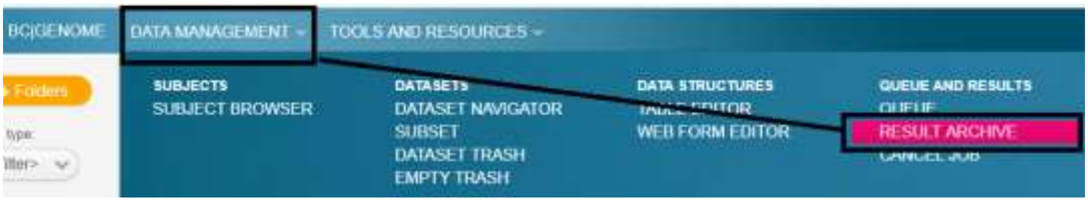
Result archive

	File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
	..	Reload current folder [user1]			11:22/18.04.18						
	upload	Transferred files		4.7 MB	12:28/11.04.18	-					
	shared	Results shared by other users			13:38/11.04.18	-					
	job10243	PLINK case-control analysis		74 KB	05:14/19.04.18	-					□
	job10238	Add / View entry - «user1» Demo phenotypes old		28 KB	05:14/19.04.18	-					□
	job10231	Upload of file «bcsystem» NCBI dbSNP Build 135 (Nov 2011, hg 37.3)		16 KB	08:21/13.04.18	-					□
	job10227	Upload of file glist-hg19.txt		44 KB	07:03/13.04.18	-					□
	job10221	Report: chr22_exome_data.vcf.gz to dataset 'My VCF data chr22'		28 KB	12:44/11.04.18	-					□
	job10220	Report: chr22_exome_data.vcf.gz to dataset 'My VCF data chr22'		28 KB	12:44/11.04.18	-					□
	job10219	Report: chr22_exome_data.vcf.gz to dataset 'My VCF data chr22'		28 KB	13:44/11.04.18	-					□
	job10217	Report: chr22_exome_data.vcf.gz to dataset 'My VCF data chr22'		28 KB	12:44/11.04.18	-					□
	job10211	Report: chr22_exome_data.vcf.gz to dataset 'My VCF data chr22'		28 KB	12:44/11.04.18	-					□
	job10209	chr22_exome_data.vcf.gz		28 KB	12:44/11.04.18	-					□
	job10199	Report: chr20.vcf and 1 others to dataset 'My VCF data'		28 KB	08:04/03.04.18	-					□
	job10197	Report: chr20.vcf and 1 others to dataset 'My VCF data'		28 KB	08:00/03.04.18	-					□
	job10194	Report: chr20.vcf and 1 others to dataset 'My VCF data'		28 KB	08:00/03.04.18	-					□
	job10193	Report: chr20.vcf and 1 others to dataset 'My VCF data'		28 KB	08:00/03.04.18	-					□
	job10187	Report: chr20.vcf and 1 others to dataset 'My VCF data'		28 KB	08:00/03.04.18	-					□

6.1. Accessing the result archive

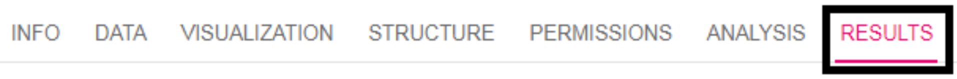
There are two ways to navigate to the result archive:

- Go to **DATA MANAGEMENT > RESULT ARCHIVE**.



OR

- Select the ***RESULTS*** page of your open dataset.

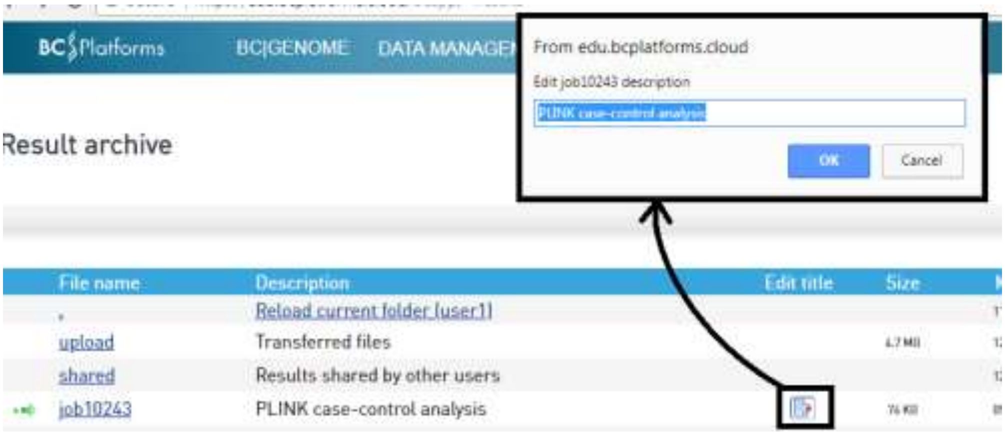


6.2. Result archive options

In the result archive, there are several options that you can use to apply to your job folders. Some options allow you to do something to the job folders, while others are for monitoring and viewing the job folders.

6.2.1. Editing the job folder title

1. Select ✎ in the ***Edit title*** column to change your job folder title.
2. Type a new name for your job folder in the dialog.
3. Select ***OK***.



From edu.bcplatforms.cloud

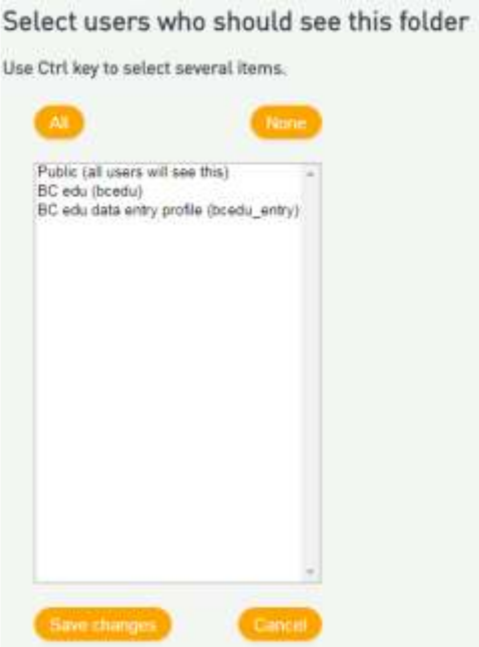
Edit job10243 description

PLINK case-control analysis

File name	Description	Edit title	Size
.	Reload current folder [user1]		
upload	Transferred files		4.7 MB
shared	Results shared by other users		
job10243	PLINK case-control analysis	✎	76 KB

6.2.2. Sharing a file

1. Select □ in the ***Share*** column to change a job folder title.
2. Select the user(s) you want to share your job folder with.



Select users who should see this folder

Use Ctrl key to select several items.

Public (all users will see this)
BC edu (bcedu)
BC edu data entry profile (bcedu_entry)

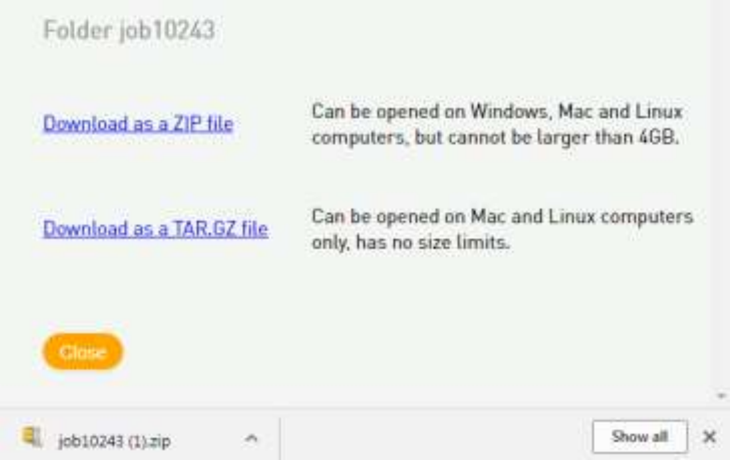
3. Select ***Save changes***.

In the ***Share*** column, changes to to show that you have shared a job folder.

If you no longer want to share a job folder, select and remove the user(s) you do not want to share a job folder with.

6.2.3. Downloading job folder information

1. Select in the ***Get*** column to download job folder information.
2. Select the file type you want to use to download the information.

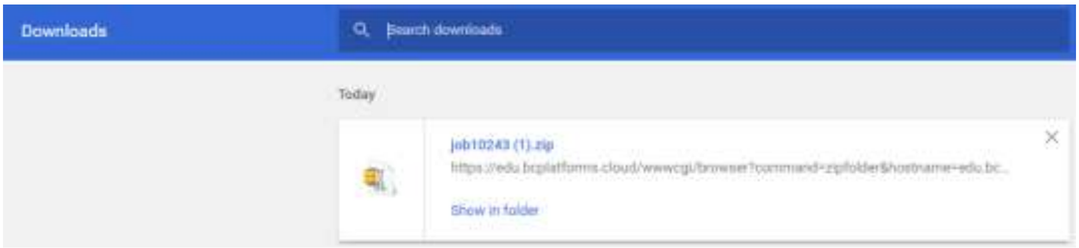


Folder job10243

Download as a ZIP file — Can be opened on Windows, Mac and Linux computers, but cannot be larger than 4GB.

Download as a TAR.GZ file — Can be opened on Mac and Linux computers only, has no size limits.

3. Select ***Close***.
4. Select ***Show all*** to view the downloaded file in your ***Downloads*** folder.



Downloads

Search downloads

Today

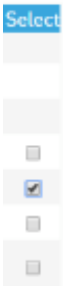
job10243 (1).zip

https://edu.bcplatforms.cloud/wwwcgi/browser?command=zipfolder&hostname=edu.bc...

Show in folder

6.2.4. Selecting a file

If you want to select one or several job folders in the ***Select*** column, select □ to choose the job folder(s) you want.



Select

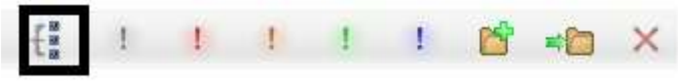
□

☑

□

□

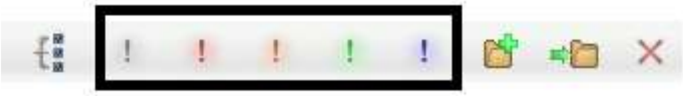
To select all job folders at the same time, select ☷ from the toolbar.



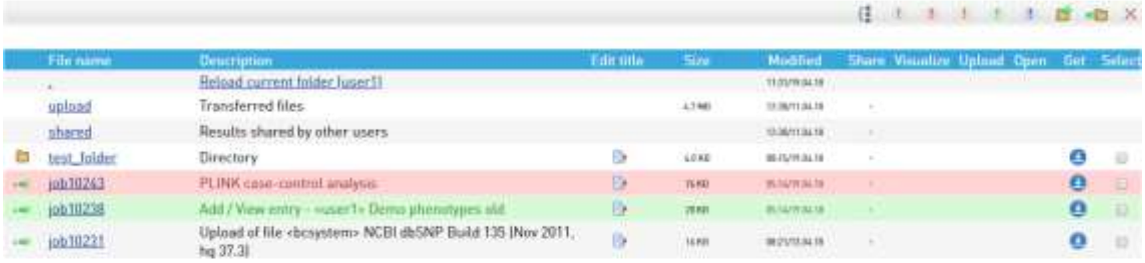
Once a job folder is selected, you can apply the following functions to it:

- Change a job folder row colour
- Move a job folder to another folder
- Delete a job folder

6.2.4.1. Changing the row colour



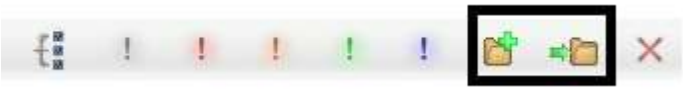
Select a colour icon to change the row colour of selected job folder.



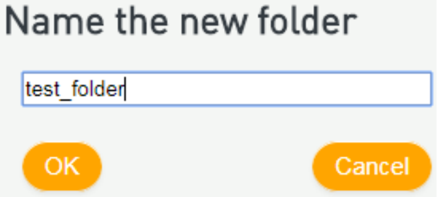
	File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
	.	Reload current folder [user1]			[illegible]						
	upload	Transferred files		4.7 MB	[illegible]	-					
	shared	Results shared by other users			[illegible]	-					
	test_folder	Directory		6.0 KB	[illegible]	-					□
	job10243	PLINK case-control analysis		76 KB	[illegible]	-					□
	job10238	Add / View entry - <user1> Demo phenotypes old		28 KB	[illegible]	-					□
	job10231	Upload of file <bcsystem> NCBI dbSNP Build 135 [Nov 2011, hg 37.3]		16 KB	[illegible]	-					□

To remove a colour, select the job folder and select the grey icon

6.2.4.2. Moving a job folder to another folder



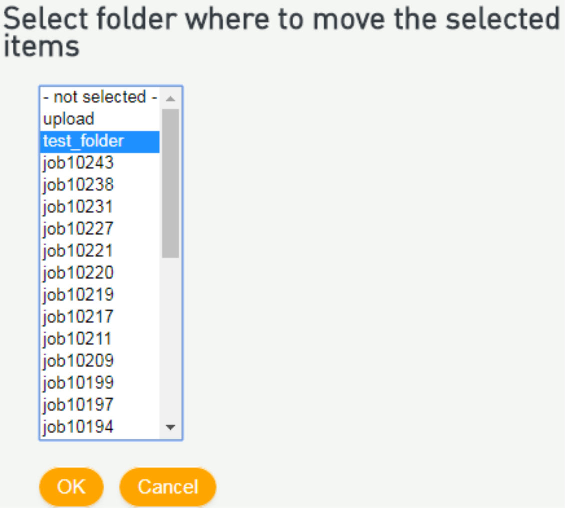
1. Select the folder icon to create a new result folder.



Name the new folder

test_folder

2. Select ***OK***.
3. Select the desired job folder you want to move.
4. Select the ***move to a folder*** icon .
5. Select the created folder from the pop-up dialog.

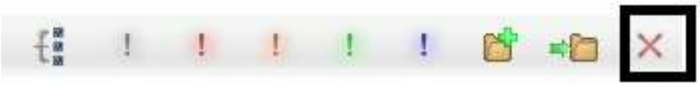


Select folder where to move the selected items

- not selected -
upload
test_folder
job10243
job10238
job10231
job10227
job10221
job10220
job10219
job10217
job10211
job10209
job10199
job10197
job10194

6. Select ***OK***.

6.2.4.3. Delete a job folder

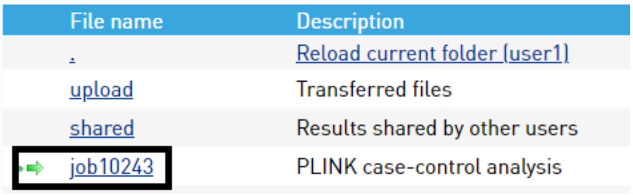


Select the delete icon to delete the selected job folders.

WARNING: The deletion is permanent!

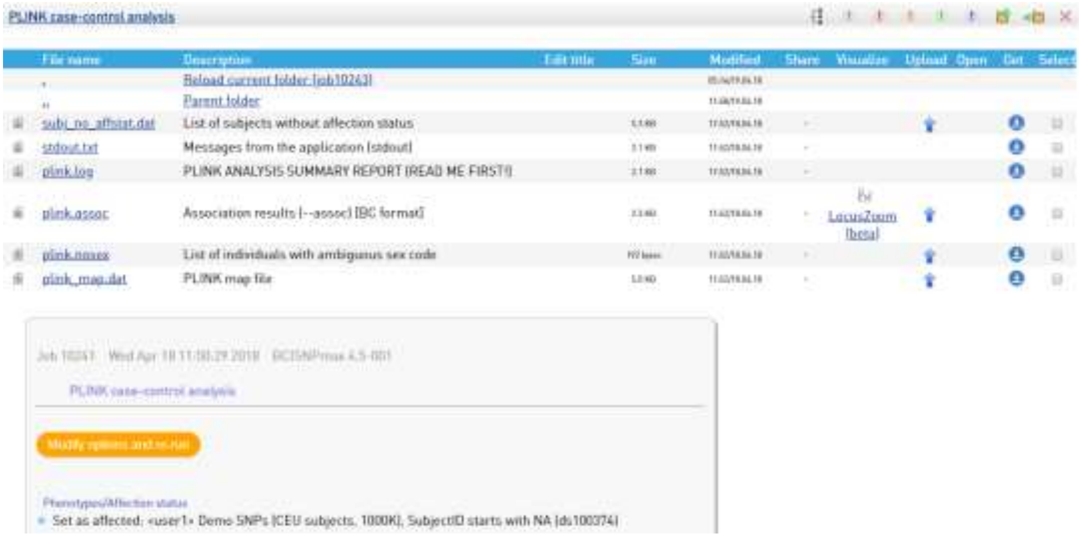
6.3. Investigating the analysis results (PLINK case-control analysis example)

1. Go to the ***RESULTS*** page or **DATA MANAGEMENT > RESULT ARCHIVE**.
2. Select the job folder link to access the result archive content.



File name	Description
.	Reload current folder (user1)
upload	Transferred files
shared	Results shared by other users
job10243	PLINK case-control analysis

The PLINK case-control analysis details display.



PLINK case-control analysis

File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
.	Reload current folder (job10243)			[illegible]						
..	Parent folder			[illegible]						
subj_no_affstat.dat	List of subjects without affection status		[illegible]	[illegible]	-					□
stdout.txt	Messages from the application (stdout)		[illegible]	[illegible]	-					□
plink.log	PLINK ANALYSIS SUMMARY REPORT (READ ME FIRST!)		[illegible]	[illegible]	-					□
plink.assoc	Association results (--assoc) (BC format)		[illegible]	[illegible]	-	LocusZoom (beta)				□
plink.nosex	List of individuals with ambiguous sex code		[illegible]	[illegible]	-					□
plink_map.dat	PLINK map file		[illegible]	[illegible]	-					□

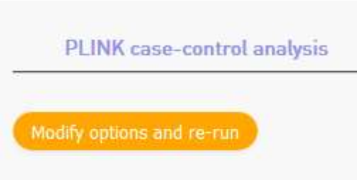
Job 10243 · Wed Apr 18 11:50:29 2018 · BCISNPmax 4.5-001

PLINK case-control analysis

Phenotypes/Affection status

- Set as affected: «user1» Demo SNPs (CEU subjects, 1000K), SubjectID starts with NA (ds100374)

3. Open *plink.log* to check the PLINK analysis parameters and data that has been filtered.
4. Open *plink.assoc* to view the association results for the PLINK *analysis*.
5. Open *LocusZoom* to view your chromosome-specific results.
6. Select ***Modify options and re-run*** to view and modify the analysis options and parameters in the analysis GUI for the next run.



NOTE: This option allows you to review the analysis specifications you have used, as well as allowing you to make changes to those specifications before running a new analysis.

7. VCF data support in BC|GENOME

A VCF (Variant Call Format) is a standardized text file that represents variants in which an individual's genome differs from a reference genome. In addition, a VCF file contains meta information about variants.

In BC|GENOME, VCF format data can be stored in a composite dataset in the BC|GENOME database, especially when you have less than ten billion genotypes.

You can familiarize yourself with the VCF file content in https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it.

The composite VCF data support in BC|GENOME optimizes database disk consumption, improves upload speed and allows efficient filtering of both variants and variants' meta data. Instead of storing all data to a single database table, the composite data upload creates one composite dataset with subsequent datasets or dynamic views, in other words, subsets. VCF file content may vary, so forms for storing VCF information are generated during the upload process in BC|GENOME to match the VCF information provided.

7.1. Composite VCF datasets

During the VCF upload process, BC|GENOME generates several tables, both datasets and subsets of different types. Before the initiation of the upload job(s), the user can specify the variant quality information to be stored in tables.

For example, if you are interested in studying the allelic dose of non-reference alleles of a candidate variant, you may consider it best to store only non-reference alleles with non-zero dose in a separate table instead of storing all alleles in a dataset that consumes lot of disk space.

The VCF composite upload process creates several different composite tables in BC|GENOME. Table 11 describes the role and use of different tables.

Table 1 Role and use of different tables

Table name	Data	Tools	Analysis
<project name>	Collection dataset	Data upload Permissions VCF export	NGS analysis tools R scripts
<project name> genotypes	Genotypes per subjects	Data analysis Genotype export VCF export Subset	Association and linkage analyses R scripts
<project name> genotype / allele (optional)	Subject specific genotypes indexed by alleles and allele dose information	Subset: functional prediction of alleles	NA

Table name	Data	Tools	Analysis
<project name> marker	Marker information, marker quality	Subset: Annotation of markers	R
<project name> allele	Allele information	Subset: functional annotation of alleles	NA
<project name> subject (batch)	Subject list with call rates	VCF file upload(s)	NA

7.2. Variant normalization

The same underlying genomic variant can be represented in many different ways across different VCF files. To able to combine and filter your variant information, BC|GENOME composite upload workflow offers the possibility to normalize both the allele name and variant position. Normalization is especially needed for either multi-allelic and/or INDELs in your VCF files.

The BC|GENOME normalization workflow applies the left-normalization algorithm as described in http://genome.sph.umich.edu/wiki/Vt.

The start position of a variant is shifted to the left until it is no longer possible to do so, the smaller the number, the better. Normalized alleles and marker information are stored in the allele-specific composite dataset.

7.3. Composite VCF dataset workflow

This section describes the steps needed to complete the composite VCF upload:

- Checking a VCF file before uploading it
- Creating a composite VCF dataset
- Transferring the VCF file(s) from the local machine to the BC|GENOME upload folder
- Uploading data from the upload folder to a composite VCF dataset

7.3.1. Checking VCF file(s) before upload

Before you start the genotype upload using VCF file(s), make sure your VCF file(s) fulfil the following criteria:

- All variants in the VCF file(s) are called in the same variant calling pipeline.
- If your variants are called using *hg38* reference information, contact support@bcplatforms.com to provide the correct reference file.
- Although you can have all you data in one file or in several VCF files, you must make sure the variants of the same chromosome are not separated into several files.

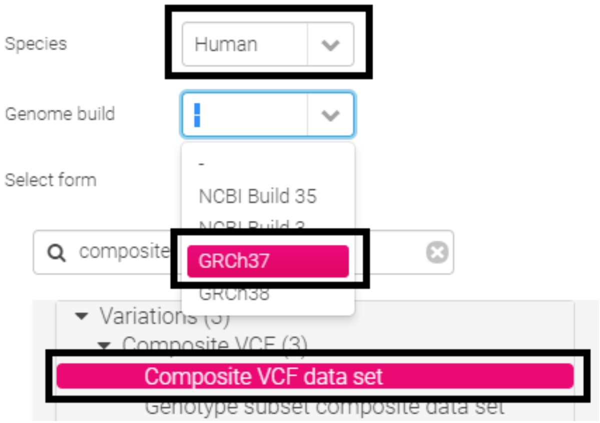
- If your VCF file(s) contains sample IDs, you must have sample – subject ID conversion pairs stored in the Sample IDs dataset. This dataset is used to convert sample IDs to subject IDs during the upload process.
- If your data is in several VCF files, make sure that every file has the same set of subjects in the same order.

7.3.2. Creating a composite VCF dataset

Select a dataset you want to use from the navigation pane. If you do not have a previously created dataset you must create a new dataset.

When creating a new dataset select the following:

- In ***Select*** form, select the ***Composite VCF data set*** form.
- In ***Species***, select ***Human***.
- In ***Genome***, select ***GRCh37***.

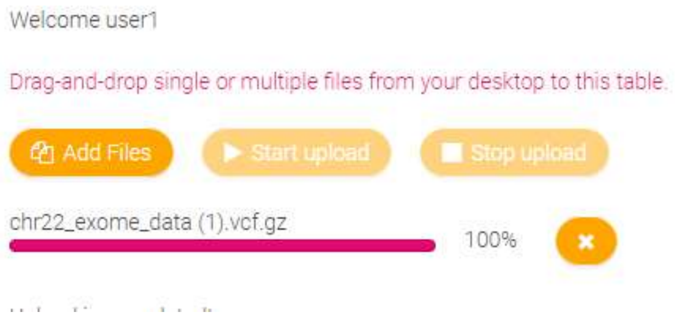


7.3.3. Transferring the VCF file(s) from the local machine to BC|GENOME upload folder

When you transfer the file containing the genotype information the file to be transferred must be in the specific format as described in https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it.

NOTE: You can transfer either an uncompressed or compressed file. If the file is compressed, it must be in the format *.gz.

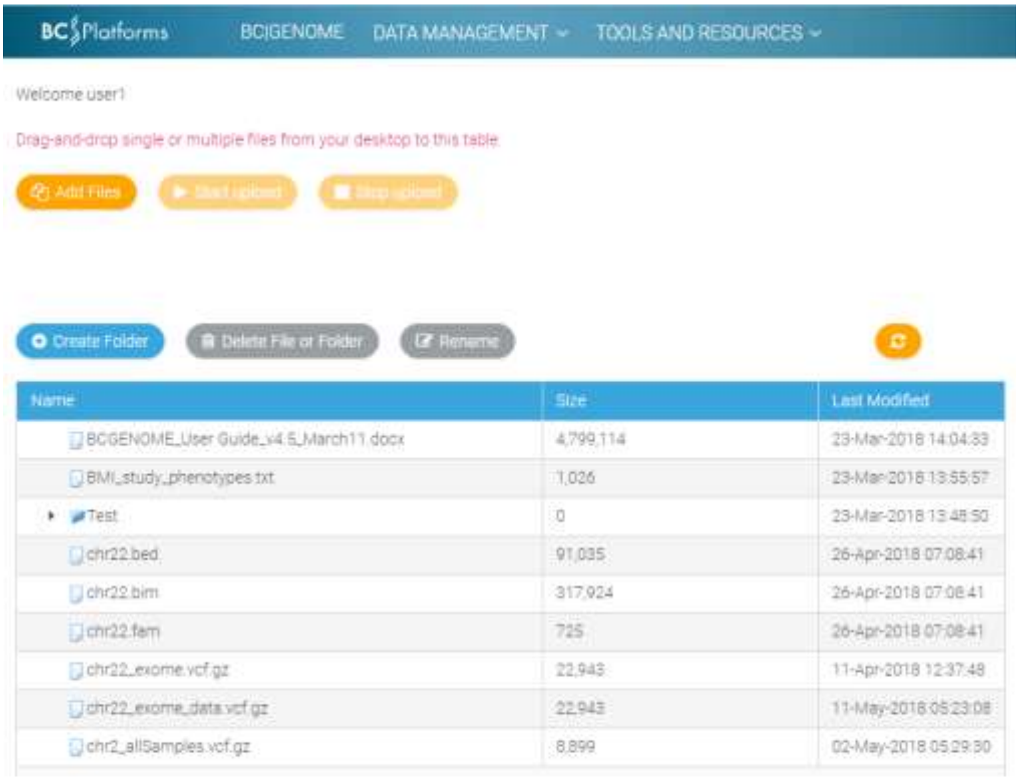
7. Select ***TOOLS AND RESOURCES > FILE TRANSER***.
8. Browse to the file(s) you want to transfer on your local machine.
9. Select ***Add files***.



Welcome user1

Drag-and-drop single or multiple files from your desktop to this table.

10. Give Filename or use the original (use Rename)



Welcome user1

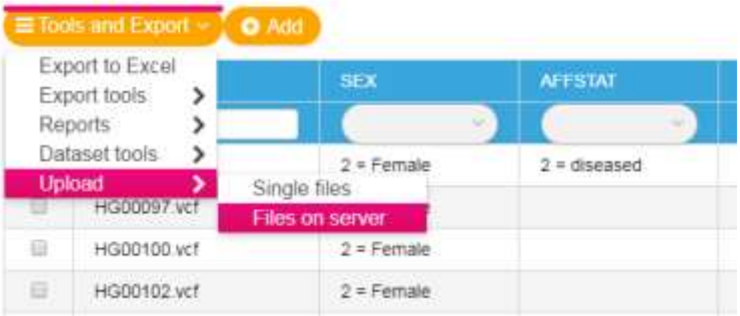
Drag-and-drop single or multiple files from your desktop to this table.

Name	Size	Last Modified
BCGENOME_User Guide_v4.5_March11.docx	4,799,114	23-Mar-2018 14:04:33
BMI_study_phenotypes.txt	1,026	23-Mar-2018 13:55:57
Test	0	23-Mar-2018 13:48:50
chr22.bed	91,035	26-Apr-2018 07:08:41
chr22.bim	317,924	26-Apr-2018 07:08:41
chr22.fam	725	26-Apr-2018 07:08:41
chr22_exome.vcf.gz	22,943	11-Apr-2018 12:37:48
chr22_exome_data.vcf.gz	22,943	11-May-2018 05:23:08
chr2_allSamples.vcf.gz	8,899	02-May-2018 05:29:30

When the file(s) is uploaded, you can continue to upload the data to the composite VCF dataset.

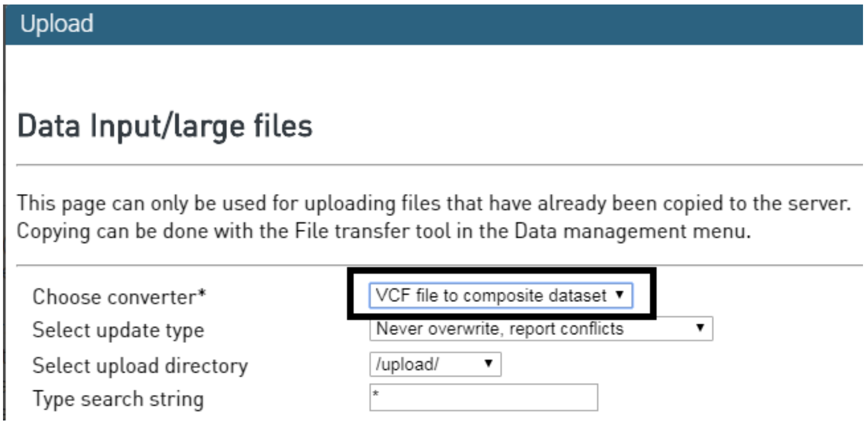
7.3.4. Uploading data from the upload folder to a composite VCF dataset

11. Select the VCF dataset from the navigation pane of the main page.
12. Select the ***DATA*** tab to open the DATA page.
13. Select ***Tools and Export > Upload > Files on server***.



The ***Data Input/large files*** dialog opens.

14. In ***Choose converter***, select ***VCF file to composite dataset***.



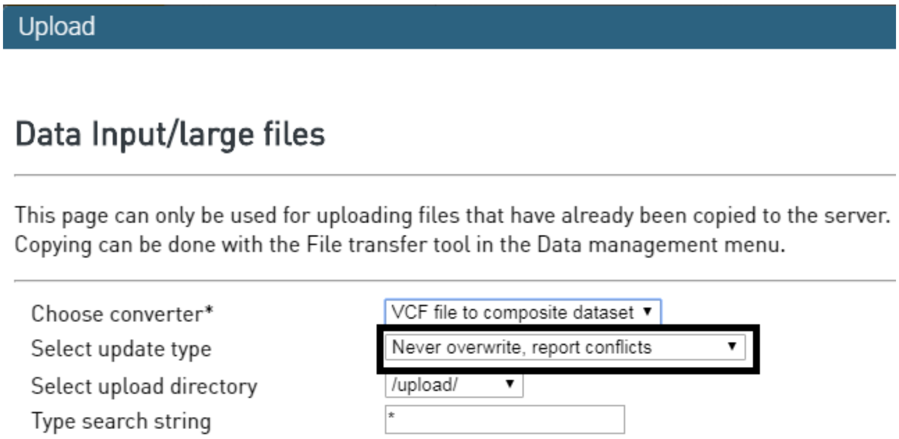
Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server. Copying can be done with the File transfer tool in the Data management menu.

Choose converter*	VCF file to composite dataset ▾
Select update type	Never overwrite, report conflicts ▾
Select upload directory	/upload/ ▾
Type search string	*

15. In ***Select update type***, select ***Never overwrite, report conflicts***.



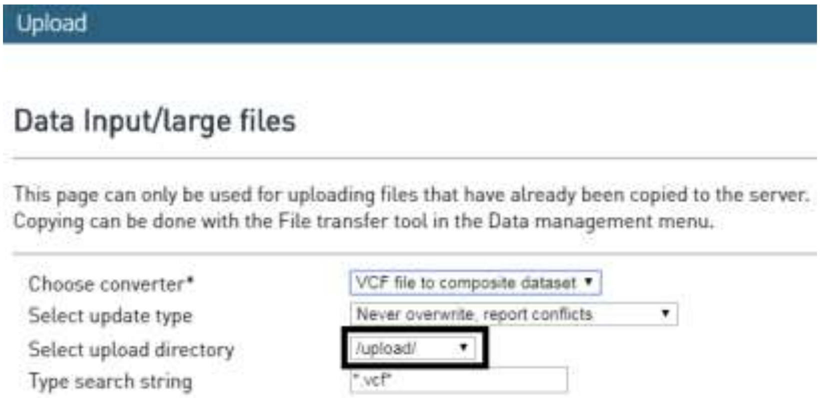
Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server. Copying can be done with the File transfer tool in the Data management menu.

Choose converter*	VCF file to composite dataset ▾
Select update type	Never overwrite, report conflicts ▾
Select upload directory	/upload/ ▾
Type search string	*

16. In ***Select upload directory*** to specify where you have your transferred file(s).



Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server. Copying can be done with the File transfer tool in the Data management menu.

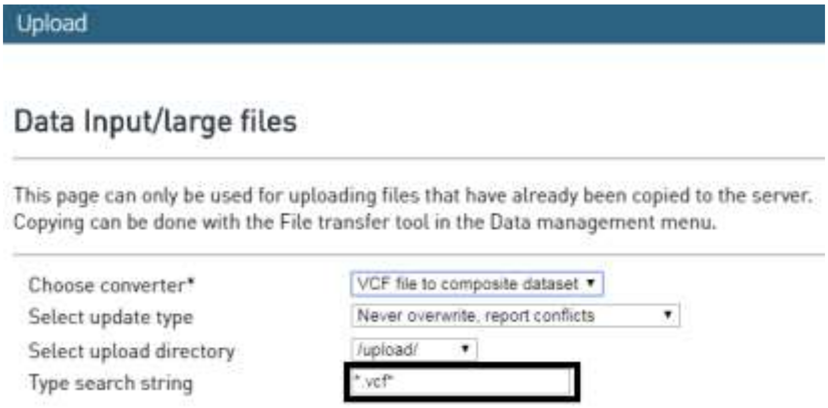
Choose converter* VCF file to composite dataset

Select update type Never overwrite, report conflicts

Select upload directory /upload/

Type search string *.vcf*

17. In ***Type search string***, type a string to match the files you want to upload, for example, ***vcf***.



Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server. Copying can be done with the File transfer tool in the Data management menu.

Choose converter* VCF file to composite dataset

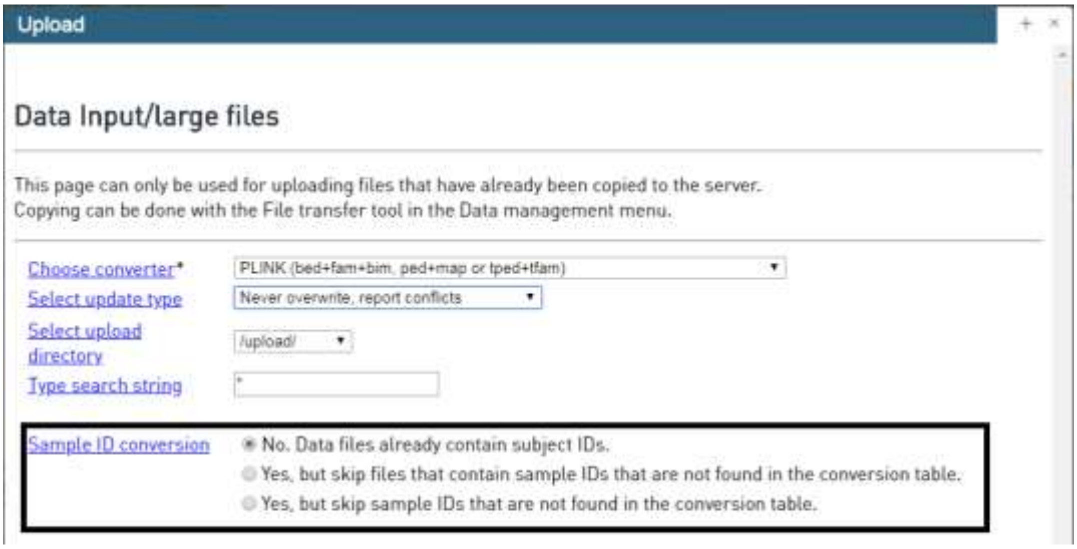
Select update type Never overwrite, report conflicts

Select upload directory /upload/

Type search string *.vcf*

18. In ***Sample ID conversion***, choose whether or not the Sample ID dataset is to be used.

NOTE:	If you have not created an ID conversion dataset that contains the data about sample ID/subject ID conversion pairs, you cannot select a sample ID conversion (the default ***No. Data files already contain subject IDs*** is used).



Upload

Data Input/large files

This page can only be used for uploading files that have already been copied to the server.
Copying can be done with the File transfer tool in the Data management menu.

Choose converter*	PLINK (bed+fam+bim, ped+map or tped+tfam) ▾
Select update type	Never overwrite, report conflicts ▾
Select upload directory	/upload/ ▾
Type search string	*
Sample ID conversion	● No. Data files already contain subject IDs. ○ Yes, but skip files that contain sample IDs that are not found in the conversion table. ○ Yes, but skip sample IDs that are not found in the conversion table.

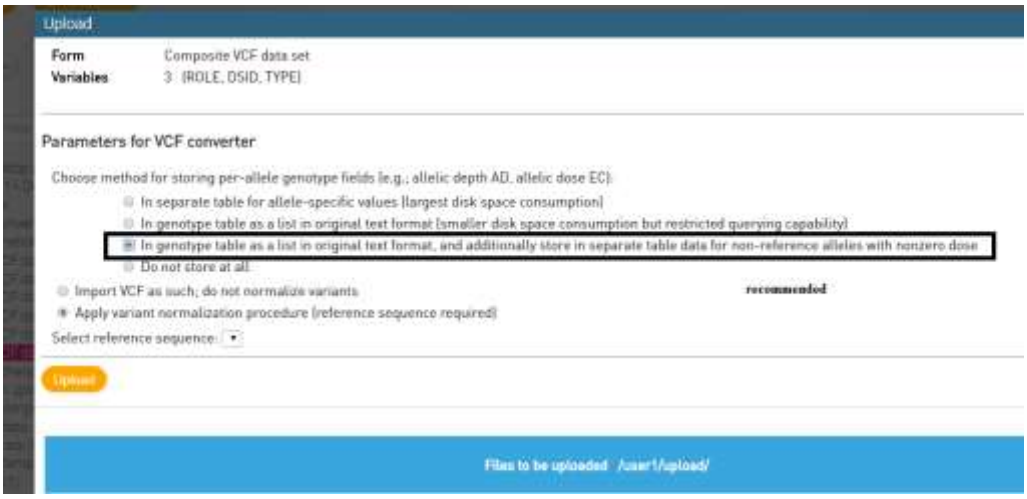
19. Select ***Continue***.

    The ***Upload*** dialog displays.

20. In ***Parameters for VCF converter***:

    e) It is recommended to select the options in a genotype table as a list in original text format. Store the data for non-reference alleles with nonzero dose for research purposes in a separate table, as then you can investigate the allelic dosage of a variant.

    f) If you store genotypes that are quality controlled before upload to BC|GENOME, and you only perform statistical analyses on your genotypes, you can consider choosing the option ***In genotype table as a list in original text format*** or ***Do not store at all***.



Upload

Form	Composite VCF data set
Variables	3 (ROLE, DSID, TYPE)

Parameters for VCF converter

Choose method for storing per-allele genotype fields (e.g.; allelic depth AD, allelic dose EC):

○ In separate table for allele-specific values (largest disk space consumption)
○ In genotype table as a list in original text format (smaller disk space consumption but restricted querying capability)
● In genotype table as a list in original text format, and additionally store in separate table data for non-reference alleles with nonzero dose
○ Do not store at all

○ Import VCF as such; do not normalize variants
● Apply variant normalization procedure (reference sequence required)

Select reference sequence: ▾

    g) Apply the variant normalization procedure:

        I. ***Import VCF as such, do not normalize variants*** – do not normalize.

        II. ***Apply variant normalization procedure (reference sequence required)*** – do normalize.

III. ***Select reference sequence*** – only applies if you choose do normalize. If there are no references sequences shown in the drop-down menu, you must contact support@bcplatforms.com to provide the correct reference files.

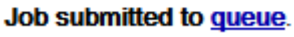
You can specify one of the following normalizing files:

- Human sequence, build 37.5
- Human sequence, build 37.5 (1000G version)
- Human sequence UCSC hg195

NOTE:	Select the reference file that matches your VCF file information. If your VCF file has chromosome IDs prefixed with ***chr*** use the **UCSC hg195** file, but if the prefix is missing, use one of the build 37.5 files.

21. Select ***Upload***.

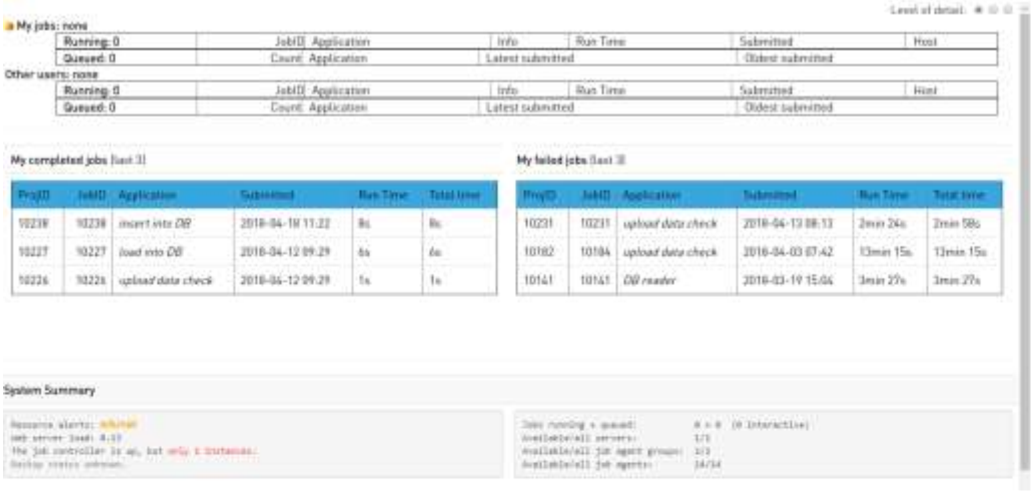
The job is submitted to a queue and a message displays.



Job submitted to queue.

22. Select ***queue***.

A screen displays where you can follow the progress of the job in the BC|GENOME queue system.



My jobs: none

Running: 0	JobID	Application	Info	Run Time	Submitted	Host
Queued: 0	Count	Application	Latest submitted		Oldest submitted	

Other users: none

Running: 0	JobID	Application	Info	Run Time	Submitted	Host
Queued: 0	Count	Application	Latest submitted		Oldest submitted	

My completed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10238	10238	*insert into DB*	2018-04-18 11:22	8s	8s
10227	10227	*load into DB*	2018-04-12 09:29	6s	6s
10226	10226	*upload data check*	2018-04-12 09:29	1s	1s

My failed jobs (last 3)

ProjID	JobID	Application	Submitted	Run Time	Total time
10231	10231	*upload data check*	2018-04-13 08:13	2min 24s	2min 58s
10182	10184	*upload data check*	2018-04-03 07:42	13min 15s	13min 15s
10141	10141	*DB reader*	2018-03-19 15:04	3min 27s	3min 27s

System Summary

```
Resource alerts: [illegible]
Web server load: 0.13
The job controller is up, but only 1 instances.
Backup status unknown.
```

```
Jobs running + queued:          0 + 0  (0 interactive)
Available/all servers:          1/1
Available/all job agent groups: 1/1
Available/all job agents:       14/14
```