

Descriptions of retrieving your requesting markers in OncoArray data

- 1) Suppose that you already have account in <https://bc1.dartmouth.edu/bcos/>



- 2) Go to BC|SNP section

The screenshot displays the main dashboard of the BC|SNP application. At the top center, the text "BC|SNP" is prominently displayed in a large teal font, with "Version 3.6-067" underneath it. To the right of this is the "BC|SNP Platforms" logo. The dashboard is organized into several sections:

- Navigation Links:** A vertical list of links on the left side: "BC|SNP", "Scripts", "Samples", "Subjects", "Tables", "Web forms", "Administration", "Download", and "Logout". Each link is accompanied by a brief description of its function.
- Help Section:** A box on the right side containing the text "Help for BC|SNP" and three links: "Online manual", "Tutorials & PDFs", and "F.A.Q.".
- Recent changes:** A section at the bottom right titled "Recent changes (read more)" with a scrollable list of updates. The most recent update is for version 3.6-06, which includes improvements in performance for uploading and editing phenotype data, support for phasing using SHAPEIT, and support for targeted sequencing studies in GATK (NGS module). The next update is for version 3.6-05, which includes the possibility to export/import data sets between two BC|SNP installations and that new marker maps are no longer installed automatically.

Last login from 10.231.72.31 at Mon Apr 21 09:28:50 EDT 2014

3) Go to Variations section, and select folder “Oncoarray”

The screenshot shows the BC|SNP Variations section. The 'Variations' tab is selected. On the left, there are navigation menus for 'Datasets' (new, open, move to trash), 'Folders' (new, remove, move datasets, rename), and 'Tools' (result archive, cancel job, file transfer, data conversion, import, empty trash, system status, queue). The main area is titled 'Datasets/open' and has a 'Select folder' dropdown set to 'OncoArray'. Below this is a table listing 30 datasets, each with a name, form, row count, derived from, created date, and access level.

Dataset name	Form	Rows	Derived from	Created	Access
<xxiao> ATBC_1	Compressed ACGT coded SNPs	196 376 208		2015-06-17	Full
<xxiao> ATBC_2	Compressed ACGT coded SNPs	795 643 821		2015-06-17	Full
<xxiao> CANADA	Compressed ACGT coded SNPs	394 353 309		2015-06-17	Full
<xxiao> CAPUA	Compressed ACGT coded SNPs	880 491 150		2015-06-17	Full
<xxiao> DARTMOUTH	Compressed ACGT coded SNPs	17 076 192		2015-06-17	Full
<xxiao> EAGLE	Compressed ACGT coded SNPs	1 979 237 379		2015-06-17	Full
<xxiao> FHCRC	Compressed ACGT coded SNPs	653 697 975		2015-06-17	Full
<xxiao> FIELD_2008	Compressed ACGT coded SNPs	121 134 237		2015-06-17	Full
<xxiao> FIELD_2013	Compressed ACGT coded SNPs	402 891 405		2015-06-17	Full
<xxiao> HSPH	Compressed ACGT coded SNPs	2 081 160 900		2015-06-17	Full
<xxiao> IARC	Compressed ACGT coded SNPs	1 325 539 404		2015-06-17	Full
<xxiao> ISRAEL	Compressed ACGT coded SNPs	688 383 990		2015-06-17	Full
<xxiao> KENTUCKY	Compressed ACGT coded SNPs	126 470 547		2015-06-17	Full
<xxiao> MDACC	Compressed ACGT coded SNPs	1 089 674 502		2015-06-17	Full
<xxiao> MDCS	Compressed ACGT coded SNPs	184 102 695		2015-06-17	Full
<xxiao> MEC	Compressed ACGT coded SNPs	1 013 898 900		2015-06-17	Full
<xxiao> NELCS	Compressed ACGT coded SNPs	199 577 994		2015-06-17	Full
<xxiao> NIJMEGEN	Compressed ACGT coded SNPs	487 205 103		2015-06-17	Full
<xxiao> NORWAY	Compressed ACGT coded SNPs	410 362 239		2015-06-17	Full
<xxiao> NSHDC	Compressed ACGT coded SNPs	273 752 703		2015-06-17	Full
<xxiao> PLCO	Compressed ACGT coded SNPs	1 745 507 001		2015-06-17	Full
<xxiao> RESOLUCENT	Compressed ACGT coded SNPs	598 733 982		2015-06-17	Full
<xxiao> RUSSIAN_CE	Compressed ACGT coded SNPs	1 216 678 680		2015-06-17	Full
<xxiao> SEOUL	Compressed ACGT coded SNPs	422 635 752		2015-06-17	Full
<xxiao> SHANGHAI	Compressed ACGT coded SNPs	281 223 537		2015-06-17	Full

30 genotype files were generated in this folder, each file is for one SITE.

For security reason, I will set permission for each file based on request, therefore, users of every PI/ study SITE can only see one file existing in this folder.

4) Go to Annotation section, and enter OncoArray folder, you will see a file:

The screenshot shows the BC|SNP web interface. At the top, there are navigation tabs for Variations, NGS, Annotations, Pedigrees, SampleIDs, and Phenotypes. The Annotations tab is active. On the left, there is a sidebar with sections for Datasets, Folders, and Tools. The main content area shows 'Datasets/open' with a 'Select folder' dropdown menu set to 'OncoArray'. Below this is a table with the following data:

<u>Dataset name</u>	<u>Form</u>	<u>Rows</u>	<u>Derived from</u>	<u>Created</u>	<u>Access</u>
<xxiao> oncoarray_533631_marker	dbSNP chromosome position	533 631		2015-06-17	Full

Below the table, there is a link that says 'Count subset rows'.

This file: [oncoarray_533631_marker](#) has information for marker name, chromosome, and position. Currently, I do not have any more information, such as alleles and associated genes, I will update information whenever it is available.

- 5) Now, I give an example to show how to export genotypes of plink format.
Go to variation section again, and select [DARTMOUTH](#) (an example, you will see genotype of your SITE name), select export

BC|SNP Variations NGS Annotations Pedigrees SampleIDs Phenotypes

OncoArray / DARTMOUTH

Datasets

- new
- open
- meta data
- subset
- merge
- combine
- snapshot
- convert
- info
- permissions
- rename
- export
- move to trash

Datasets/info

Dataset **<xxiao> DARTMOUTH**

Created	2015-06-17 10:02:38
Table	ds100596
Permissions	read and write
Species	Human (Homo sapiens) ▼
Genome build	b37 (GRCh37) ▼

Form **Compressed ACGT coded SNPs**

Variables **5** (SUBJECT, MARKER, ALLELE1, ALLELE2, MENDEL)

[Variable details](#)

Status of dataset index:

Index is up-to-date. Dataset is ready for analysis.
Updated: 2015-06-17 10:05

Data rows: **17076192** (100.00% of full subject-marker matrix).
SNPs from **32** subjects, **533631** markers.

Folders

- new
- remove
- move datasets
- rename

Analysis

- SNPs

Datasets

- new
- open
- meta data
- subset
- merge
- combine
- snapshot
- convert
- info
- permissions
- rename
- export
- move to trash

Folders

- new
- remove
- move datasets
- rename

Analysis

- GWAS
- GWAS family
- linkage
- scripts
- custom

Reports

- chr12.bed

Datasets/export selection

[Data export to analysis programs](#)

Data export to analysis programs uses the analysis program file preparation engine when preparing the data for the export. This method includes data consistency checks and other benefits such as allele downcoding. For genotype data, the file can only include columns PATIENT, MARKER, ALLELE1 and ALLELE2. If the genotype dataset has additional columns (quality scores, etc.), these additional columns can not be exported using this tool. This tool offers the possibility to export phenotypes and genotypes together, and has useful options for formatting the genotype data.

Export to statistical programs for smaller data sets

This tool is not available for compressed datasets.

[Data export to PLINK format \(case-control\)](#)[Data export to PLINK format \(quantitative trait\)](#)

PLINK formats (binary, text and transposed) are widely used formats for storing GWAS datasets. Note that phase information, quality score, half-genotypes or genotype probabilities/dosages will not be included when exporting to PLINK format.

[Export to other BC systems](#)

This function creates datafile, which can be imported to other BC systems using Tools/import function. In addition of data, also user designed or modified forms are transferred. This tool facilitates data exchange between different BC databases and systems.

You will find two plink choices of exporting, now select “[Data export to PLINK format \(case-control\)](#)”, note that there is no different between two PLINK choices, since there is no affection status in this example.

Now choose the format of PLINK format you want, for example, tped+tfam. Do not do anything on Affection, Covariates and Pedigrees parts. Then, click “Use map”, choose markers generated in folder OncoArray. You can split analysis by chromosomes or export all markers.

↓ PLINK is a free, open-source whole genome association analysis toolset...

General

Run title:

Run mode: Max. run time: Send all analysis directory content

Subjects

↓ No filters specified, use all subjects. Click the arrow to the left to edit selections.

Affection Status Optional

EITHER set as cases:

And (optionally) set as controls:

OR select affection dataset:

Covariates Optional

Select dataset:

Select variable:

Pedigrees Optional

↓ Select pedigree set:

Gender Optional, gender is set to unknown if not defined

Select gender dataset:

Markers

Marker map: Use map

Folder: Dataset:

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

↓

- 6) Run this process and you will find your results in “result archive”, for example: job13047, PLINK binary format, and then you can download datasets.

OncoArray / DARTMOUTH Logout x

Tools/result archive

File name	Description	Edit title	Size	Modified	Share	Visualize	Upload	Open	Get	Select
-	Reload current folder (xxiao)			23:58/17.08.15						
upload	Transferred files		20 GB	23:42/17.08.15	*					
shared	Results shared by other users			23:42/17.08.15	*					
➔ job13050	Upload of file samplelist_39162_30_sites.txt		18 KB	23:58/17.08.15	*					<input type="checkbox"/>
➔ job13049	samplelist_39162_30_sites.txt		18 KB	23:44/17.08.15	*					<input type="checkbox"/>
➔ job13047	PLINK binary format		11 MB	15:10/17.08.15	*					<input type="checkbox"/>
➔ job13041	Report: original_TOP_row_genotypes_PLCO to dataset ' PLCO'		18 KB	14:53/17.08.15	*					<input type="checkbox"/>
➔ job13039	Report: original_TOP_row_genotypes_TORONTO to dataset ' TORONTO'		18 KB	14:35/17.08.15	*					<input type="checkbox"/>
➔ job13038	Report: original_TOP_row_genotypes_VANDERBILT to dataset ' VANDERBILT'		18 KB	14:19/17.08.15	*					<input type="checkbox"/>
➔ job13036	Report: original_TOP_row_genotypes_RUSSIAN_CE to dataset ' RUSSIAN_CE'		18 KB	13:58/17.08.15	*					<input type="checkbox"/>
➔ job13033	Report: original_TOP_row_genotypes_TAMPA to dataset ' TAMPA'		18 KB	13:35/17.08.15	*					<input type="checkbox"/>
➔ job13030	Report: original_TOP_row_genotypes_TLC to dataset ' TLC'		18 KB	13:25/17.08.15	*					<input type="checkbox"/>

- 7) For sample information of each SITE, go to phenotypes section in folder OncoArray, you will find a file: [sample-information-site-JUNE-18-2015](#). Open this file, you will find that there are four variables:

ORIGINAL_ID	CIDR	SITE	SAMPLEID
-------------	------	------	----------

Original ID stands for in total 53600 samples (we received), the unique ID assigned for each sample, you need not to use it, database needs a primary key for this table.

CIDR stands for local ID

SITE stands for site name

SampleID stands for unique ID for each SITE, the order is exactly matched to genotype’s order.

For example, if you generate tped (transposed plink format, row for maker, column for samples), the order of samples is matched to sampleid.

Note that we will frequently update this file with adding extra columns following four variables described above. In each file’s description box, there will be detailed information.