

## Examples of dataset construction & other queries

---

### *Question 1: Where do I find a list of animals tested for gene X?*

---

1. Go to the Populations directory.
2. Use the colonies\_to\_pipelines\_mapping document to determine which pipeline(s) your gene of interest was processed through.
3. Find your gene of interest within the correct pipeline folder and download the .csv file. Both the gene name and the colony prefix is included in the file name.

Note that for primary phenotyping data, this will be in either MGP\_Pipeline\_1, MGP\_Pipeline\_2, Mouse\_GP or MGP\_Select. The gene may also have been processed through one of the other screens such as Expression or a specific infection challenge, so please check these if this is the type of data you are interested in.

4. This will give you the list of mutant animals tested plus the local controls (as defined above – the controls processed during the same weeks as the mutants). The unique identifiers for each animal are the ID, Mouse Name or Mouse Barcode.

---

### *Question 2: I want to download all data for test X for gene Y, including mutants, controls and a baseline*

---

1. Follow steps 1-4 for Question 1 above.
2. Go to the Index directory. Look in the *index\_of\_indexes* file to narrow down which .csv index files you will need to check for the pipeline your gene of interest was processed through.
3. Look in the relevant numbered index files to search for the animals in your population file within the mouseID or mouseName columns.
4. The location and filename columns will then direct you to where the raw data for that animal and test can be found.
5. Go to the Pipelines directory to access the raw data for each test. Select the pipeline stated in the filename for your population then go to your test of interest.

There may be one or more folders within the test folder if there were different data types generated as part of the test (e.g. observation values and images).

6. The specific numbered file containing the measurements of interest is that listed in the index file checked during step 2.

7. To assemble a baseline, take a note of the baseline ID specified in the population filename.

e.g. in the following file *acbd5\_tm1b(eucomm)wtsj\_(ID-12488\_BaselineID-94795)\_PMAJ\_(57-ROWS).csv* the baseline ID is 94795.

8. The baseline files can be found within the Baselines directory. Match up the ID from your population filename with the ID in a filename here.

9. The raw data for the list of animals in the baseline file will need to be found in the same way as those in the population file, by following steps 2-6.

---

*Question 3: I want to access all the wildtype/control data for test X*

---

1. We suggest you use control data from a single pipeline (unless you are doing a comparison) and combining all control data is not appropriate due to differences in the pipelines over the lifetime of the Mouse Genetics Project.

2. Take a look at the Pipelines document in the Documentation folder to select a suitable pipeline.

3. The baseline files can be found within the Baselines directory. The pipeline and genetic background within the filenames will help you select the appropriate file.

4. This will give you the list of control animals of that genetic background tested on that pipeline. The unique identifiers for each animal are the ID, Mouse Name or Mouse Barcode.

5. The raw data for the list of animals in the baseline file will need to be found in the same way as detailed above in answer to Question 2, steps 2-6.

---

*Question 4: How do I find out what was considered significant/not significant when line X was phenotyped?*

---

1. Go to the Heat\_Map directory.

2. For a visual representation, download HeatMap.xlsx

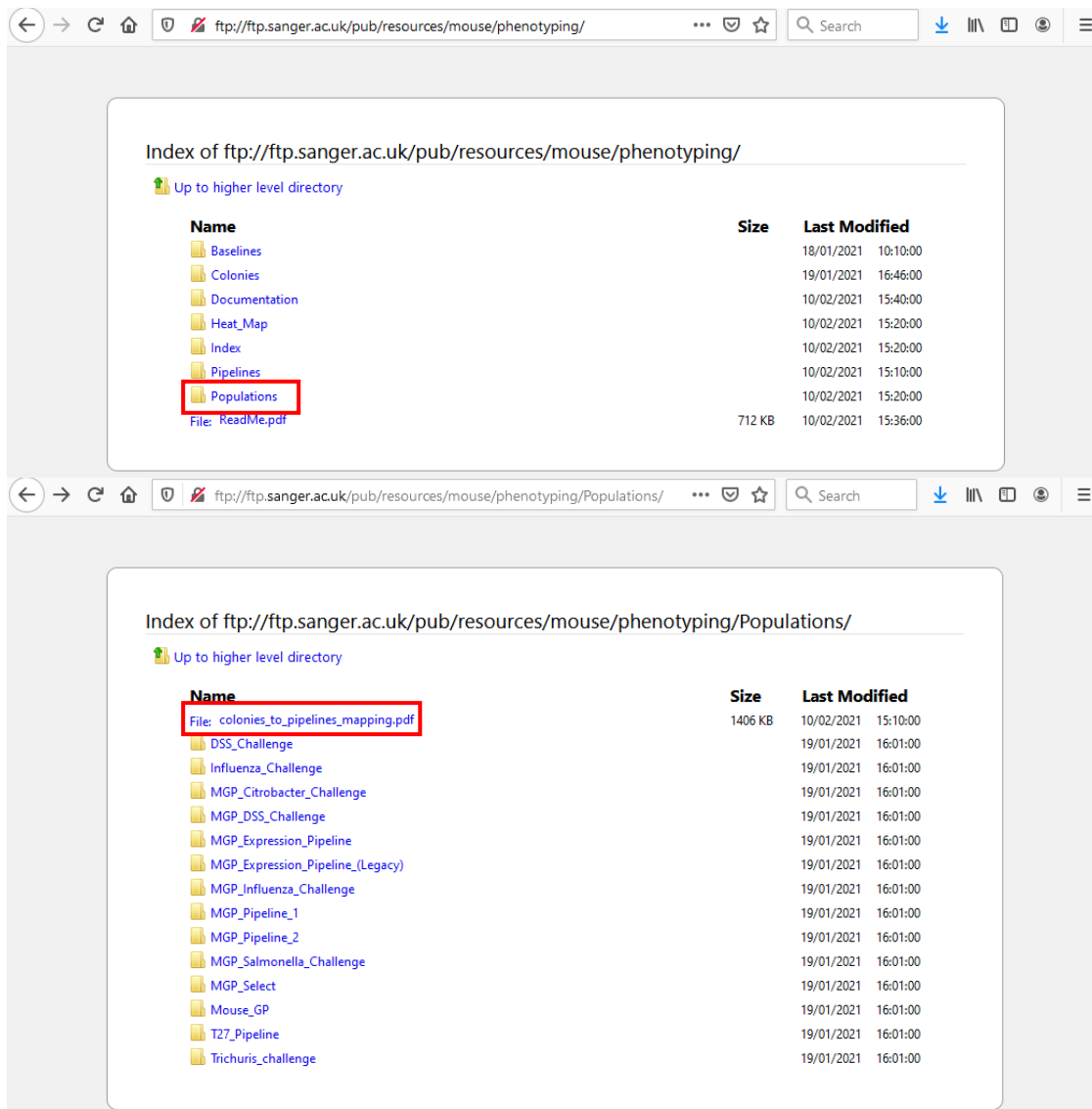
3. The Legend tab and the Heat\_map.pdf document will help you understand the colours to determine if the results of any test performed on a line were considered significant.
4. The HeatMapCalls.xlsx file will detail specific parameters measured in each test deemed significant and also give the corresponding Mammalian Phenotype (MP) term(s).

### Appendix: Worked example of dataset construction

Note that the following screenshots were taken using Mozilla Firefox 85.0.2 and will likely appear visually different on other browsers. The functionality should be the same, however.

Q: I want the raw haematology data for the *Kcne2* line of mice plus controls.

1. Find out which adult phenotyping pipeline this line was processed through. Go to Populations and open colonies\_to\_pipelines\_mapping.pdf



Index of ftp://ftp.sanger.ac.uk/pub/resources/mouse/phenotyping/

Up to higher level directory

| Name             | Size   | Last Modified       |
|------------------|--------|---------------------|
| Baselines        |        | 18/01/2021 10:10:00 |
| Colonies         |        | 19/01/2021 16:46:00 |
| Documentation    |        | 10/02/2021 15:40:00 |
| Heat_Map         |        | 10/02/2021 15:20:00 |
| Index            |        | 10/02/2021 15:20:00 |
| Pipelines        |        | 10/02/2021 15:10:00 |
| Populations      |        | 10/02/2021 15:20:00 |
| File: ReadMe.pdf | 712 KB | 10/02/2021 15:36:00 |

Index of ftp://ftp.sanger.ac.uk/pub/resources/mouse/phenotyping/Populations/

Up to higher level directory

| Name                                    | Size    | Last Modified       |
|---|---------|---------------------|
| File: colonies_to_pipelines_mapping.pdf | 1406 KB | 10/02/2021 15:10:00 |
| DSS_Challenge                           |         | 19/01/2021 16:01:00 |
| Influenza_Challenge                     |         | 19/01/2021 16:01:00 |
| MGP_Citrobacter_Challenge               |         | 19/01/2021 16:01:00 |
| MGP_DSS_Challenge                       |         | 19/01/2021 16:01:00 |
| MGP_Expression_Pipeline                 |         | 19/01/2021 16:01:00 |
| MGP_Expression_Pipeline(Legacy)         |         | 19/01/2021 16:01:00 |
| MGP_Influenza_Challenge                 |         | 19/01/2021 16:01:00 |
| MGP_Pipeline_1                          |         | 19/01/2021 16:01:00 |
| MGP_Pipeline_2                          |         | 19/01/2021 16:01:00 |
| MGP_Salmonella_Challenge                |         | 19/01/2021 16:01:00 |
| MGP_Select                              |         | 19/01/2021 16:01:00 |
| Mouse_GP                                |         | 19/01/2021 16:01:00 |
| T27_Pipeline                            |         | 19/01/2021 16:01:00 |
| Trichuris_challenge                     |         | 19/01/2021 16:01:00 |

- By searching this document, we can see that there is data for this line from the Mouse GP pipeline.

Note: for general details about this pipeline, we can look in the Documentation subdirectory.

|                                 |                    |      |  |  |          |            |
|---------------------------------|--------------------|------|--|--|----------|------------|
| Ush1c<tm1a(KOMP)Wtsi>           | B6N                | MCSB |  |  |          | MGP Select |
| Anp32<tm1e(KOMP)Wtsi>           | B6N                | MCSB |  |  | Mouse GP |            |
| Cfh<tm1a(EUCOMM)Wtsi>           | B6N                | MCSE |  |  | Mouse GP |            |
| Rhot2<tm1(KOMP)Wtsi>            | B6N                | MCSF |  |  | Mouse GP |            |
| 3110001I22Rik<tm1a(EUCOMM)Wtsi> | B6N                | MCSH |  |  | Mouse GP |            |
| Kcne2<tm1a(EUCOMM)Wtsi>         | B6N                | MCSJ |  |  | Mouse GP |            |
| Rab29<tm1a(EUCOMM)Wtsi>         | B6N                | MCSJ |  |  | Mouse GP |            |
| Pld5<tm1b(KOMP)Wtsi>            | B6Brd;B6Dnk;B6N-Ty | MCSQ |  |  |          | MGP Select |
| Spns2<tm1b(KOMP)Wtsi>           | B6Brd;B6Dnk;B6N-Ty | MCSR |  |  | Mouse GP | MGP Select |
| Sar1b<tm1a(EUCOMM)Wtsi>         | B6N                | MCSJ |  |  | Mouse GP |            |

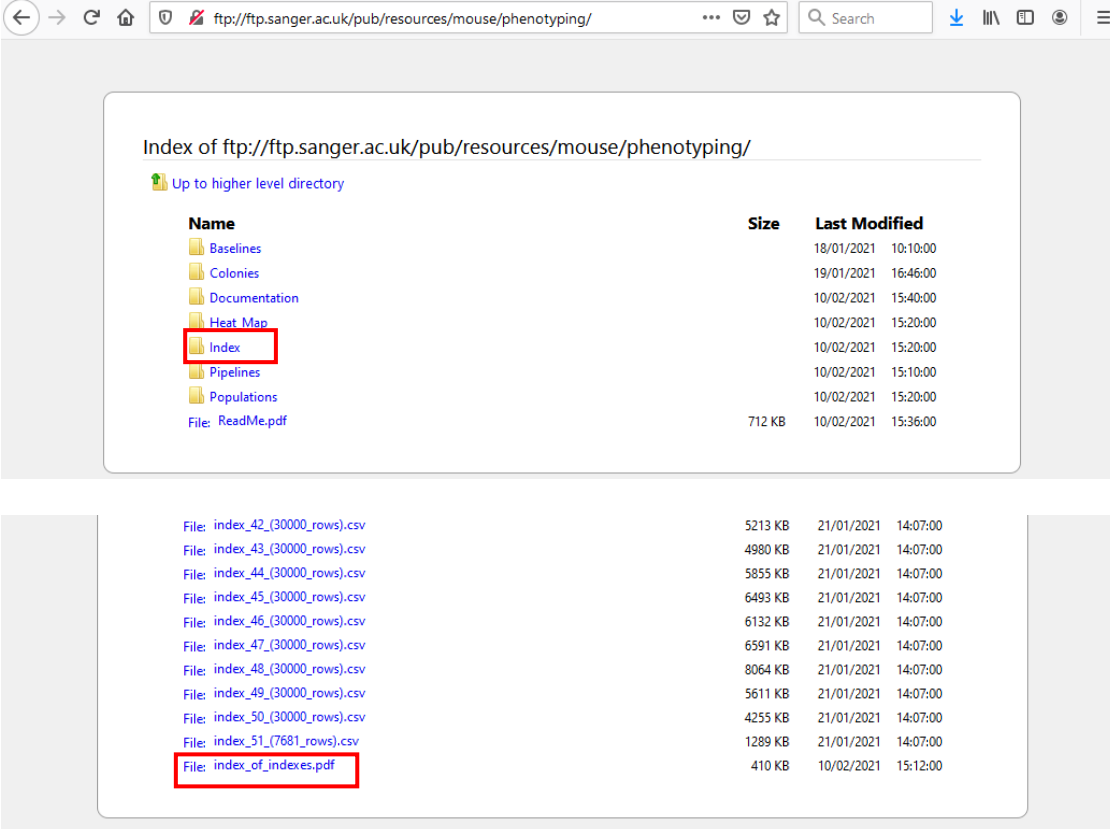
- Still in the Populations subdirectory, go to the Mouse\_GP folder and search for either the gene name (Kcne2) or colony prefix (MCSJ) to locate the correct file to download.

|  |       |            |          |
|--|-------|------------|----------|
| File: jarid2_(ID-3106_BaselineID-65784)_MAEF_(28-ROWS).csv | 13 KB | 18/01/2021 | 16:37:00 |
| File: kazn_(ID-3906_BaselineID-46808)_MCFE_(62-ROWS).csv   | 26 KB | 18/01/2021 | 16:39:00 |
| File: kcne2_(ID-4030_BaselineID-46808)_MCSJ_(47-ROWS).csv  | 20 KB | 18/01/2021 | 16:48:00 |
| File: kdm4b_(ID-1974_BaselineID-46785)_MBCQ_(63-ROWS).csv  | 32 KB | 18/01/2021 | 16:33:00 |
| File: kdm4c_(ID-1763_BaselineID-46808)_MBXV_(28-ROWS).csv  | 12 KB | 18/01/2021 | 16:42:00 |

- In the below screenshot showing a subset of the .csv file indicated above, we can see columns which will be important later:
  - Unique animal identifiers: any of ID, Mouse Name, Mouse Barcode (highlighted blue).
  - Cohort type (highlighted green): Control or Subject and Genotype (highlighted yellow) which shows that for this line the subjects were homozygous.
  - Sex (highlighted orange) to help us split the data for analysis appropriately.

|    | A      | B          | C             | D           | E     | F         | G         | H         | I      | J         | K         |
|----|--------|------------|---------------|-------------|-------|-----------|-----------|-----------|--------|-----------|-----------|
| 1  | ID     | Mouse Name | Mouse Barcode | Cohort Type | Coat  | D.O.B.    | D.O.W.    | D.O.D.    | Sex    | Genotype  | G.Backgro |
| 2  | 835901 | MALM41.1d  | M00835901     | Control     | Black | 02-Jun-11 | 22-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 3  | 835902 | MALM41.1e  | M00835902     | Control     | Black | 02-Jun-11 | 22-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 4  | 835903 | MALM41.1f  | M00835903     | Control     | Black | 02-Jun-11 | 22-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 5  | 837334 | MAQG63.2g  | M00837334     | Control     | Black | 04-Jun-11 | 24-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 6  | 837335 | MAQG63.2h  | M00837335     | Control     | Black | 04-Jun-11 | 24-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 7  | 842157 | MAQG63.2i  | M00842157     | Control     | Black | 04-Jun-11 | 24-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 8  | 835853 | MAQG64.2d  | M00835853     | Control     | Black | 02-Jun-11 | 22-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 9  | 835854 | MAQG64.2e  | M00835854     | Control     | Black | 02-Jun-11 | 22-Jun-11 | 22-Sep-11 | Female | WT        | C57BL/6N; |
| 10 | 863483 | MAQG71.1g  | M00863483     | Control     | Black | 05-Jul-11 | 26-Jul-11 | 24-Oct-11 | Female | WT        | C57BL/6N; |
| 11 | 863484 | MAQG71.1h  | M00863484     | Control     | Black | 05-Jul-11 | 26-Jul-11 | 24-Oct-11 | Female | WT        | C57BL/6N; |
| 12 | 863485 | MAQG71.1i  | M00863485     | Control     | Black | 05-Jul-11 | 26-Jul-11 | 24-Oct-11 | Female | WT        | C57BL/6N; |
| 13 | 863486 | MAQG71.1j  | M00863486     | Control     | Black | 05-Jul-11 | 26-Jul-11 | 24-Oct-11 | Female | WT        | C57BL/6N; |
| 14 | 832395 | MCSJ23.3a  | M00832395     | Subject     | Black | 24-May-11 | 15-Jun-11 | 14-Sep-11 | Male   | Kcne2:Hom | C57BL/6N; |
| 15 | 832396 | MCSJ23.3b  | M00832396     | Subject     | Black | 24-May-11 | 15-Jun-11 | 14-Sep-11 | Male   | Kcne2:Hom | C57BL/6N; |
| 16 | 832398 | MCSJ23.3d  | M00832398     | Subject     | Black | 24-May-11 | 15-Jun-11 | 14-Sep-11 | Male   | Kcne2:Hom | C57BL/6N; |
| 17 | 828310 | MCSJ26.2b  | M00828310     | Subject     | Black | 21-May-11 | 15-Jun-11 | 14-Sep-11 | Male   | Kcne2:Hom | C57BL/6N; |
| 18 | 828314 | MCSJ26.2f  | M00828314     | Subject     | Black | 21-May-11 | 15-Jun-11 | 14-Sep-11 | Female | Kcne2:Hom | C57BL/6N; |

- Now we have our list of animals, we can use these to find the raw data for haematology. At the top level, go to the Index directory and open index\_of\_indexes.pdf



Index of ftp://ftp.sanger.ac.uk/pub/resources/mouse/phenotyping/

[Up to higher level directory](#)

| Name             | Size   | Last Modified       |
|------------------|--------|---------------------|
| Baselines        |        | 18/01/2021 10:10:00 |
| Colonies         |        | 19/01/2021 16:46:00 |
| Documentation    |        | 10/02/2021 15:40:00 |
| Heat_Map         |        | 10/02/2021 15:20:00 |
| <b>Index</b>     |        | 10/02/2021 15:20:00 |
| Pipelines        |        | 10/02/2021 15:10:00 |
| Populations      |        | 10/02/2021 15:20:00 |
| File: ReadMe.pdf | 712 KB | 10/02/2021 15:36:00 |

|                                   |         |                     |
|-----------------------------------|---------|---------------------|
| File: index_42_(30000_rows).csv   | 5213 KB | 21/01/2021 14:07:00 |
| File: index_43_(30000_rows).csv   | 4980 KB | 21/01/2021 14:07:00 |
| File: index_44_(30000_rows).csv   | 5855 KB | 21/01/2021 14:07:00 |
| File: index_45_(30000_rows).csv   | 6493 KB | 21/01/2021 14:07:00 |
| File: index_46_(30000_rows).csv   | 6132 KB | 21/01/2021 14:07:00 |
| File: index_47_(30000_rows).csv   | 6591 KB | 21/01/2021 14:07:00 |
| File: index_48_(30000_rows).csv   | 8064 KB | 21/01/2021 14:07:00 |
| File: index_49_(30000_rows).csv   | 5611 KB | 21/01/2021 14:07:00 |
| File: index_50_(30000_rows).csv   | 4255 KB | 21/01/2021 14:07:00 |
| File: index_51_(7681_rows).csv    | 1289 KB | 21/01/2021 14:07:00 |
| <b>File: index_of_indexes.pdf</b> | 410 KB  | 10/02/2021 15:12:00 |

- In this file we can see that the indexes for data from Mouse GP are in index files 35 – 50 so we only need to search 16 files, rather than all 51.
- The relevant index files will then need to be searched for where:
  - The mouseId or MouseName (from the index) matches an ID or Mouse Name in the population file
 AND
  - The sopId in the index file matches your test of interest (in this case Haematology)
- The rows within the index files where the above two conditions are met should be extracted and combined to create a dataset of haematology data for the *Kcne2* line.