

# Converting GenMAPP MAPPs between species using homology

	Page
1 Introduction and Background	2
1.1 Fundamental principles of the GenMAPP Gene Database	2
1.1.1 Gene Database data types	2
1.1.2 GenMAPP System Codes	3
1.1.2.1 System Code format requirements	3
1.2 MAPP Basics	4
2 Preparing the homology data	5
2.1 Prerequisites	5
2.1.1 Gene Database	5
2.1.2 Homology information	5
2.1.3 MAPP Archive	6
2.2 Formatting the homology data as a Relationship Table	6
3 Configuring the Gene Database and performing the conversion	7
3.1 Adding homology information as	7
3.2 Performing the conversion	9
3.2.1 New MAPP Name	11

# 1 Introduction

This document describes the process of converting MAPP Archives between related species, using the Converter tool in combination with homology information. This process may be useful for creating MAPPs for a newly supported species for which no MAPPs exist, or for a non-supported species with a custom database.

This manual uses as an example the creation of a MAPP Archive for *Canis familiaris*, or dog, starting with human MAPPs and using homology between human and dog. The resulting MAPPs are intended to be used with the official GenMAPP dog database.

## 1.1 Fundamental principles of the GenMAPP Gene Database

The GenMAPP Gene Database is a species-specific library of gene information used by the GenMAPP program. It is essential for linking expression data with MAPPs, for creating and modifying MAPPs, for importing new data and for MAPPFinder analysis. It contains the relationship between gene IDs from different systems, for example between UniProt and Entrez Gene. These relationships, as well as annotation information, is displayed on GenMAPP Backpages. The database is also essential for providing the links to relevant public databases from the Backpage.

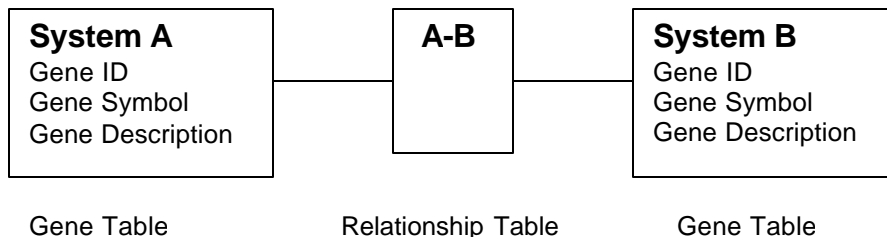
### 1.1.1 Gene Database data types

The Gene Database contains two main types of information, **Gene Tables** and **Relationship Tables**.

A Gene Table is a collection of gene identifiers from a gene cataloging system. In addition to the gene identifiers, the Gene Table may also contain other annotation, such as gene symbol, gene name, gene description etc. Each Gene Table represents a gene ID system in a Gene Database. For example, your Gene Database might have one Gene Table containing information from UniGene and another Gene Table containing information from Entrez Gene. These two tables then represent two gene ID systems.

A Relationship Table provides links between the gene IDs of two gene ID systems. Relationship Tables are essential to the functionality of the GenMAPP program. In addition, information from the Relationship Tables appears in Backpages and in the Gene Finder, providing essential annotation. If your Gene Database contains the two systems A and B as Gene Tables, they should also contain the Relationship Table A – B.

Figure 1: Gene Database data types



For more detailed information on the specific format of Gene Tables and Relationship Tables, please refer to sections 3.1 and 3.2.

## 1.1.2 GenMAPP System Codes

Each gene ID system present in a GenMAPP Gene Database requires a System Code. This code enables GenMAPP to identify the type of gene ID on MAPPs, in Expression Data as well as in the Gene Database. For example, if you have a UniProt Gene Table as well as a Gene Table containing proprietary gene IDs, these two systems of genes have to be assigned different System Codes.

There are a number of gene ID systems already supported by GenMAPP. These are systems that appear in the supported Gene Databases distributed by GenMAPP. Below is a current list of supported systems with the corresponding system code:

Table 1: GenMAPP supported Systems and System Codes

System Code	System official name
Em	EMBL
Om	OMIM
Pd	PDB
H	HUGO
Pf	Pfam
En	Ensembl
D	SGD
F	FlyBase
G	GenBank
I	InterPro
L	Entrez Gene
M	MGI
Q	RefSeq
R	RGD
S	UniProt
T	Gene Ontology
U	UniGene
W	WormBase
Z	ZFIN
X	Affy
O	Other

### 1.1.2.1 System Code format requirements

The above System Codes are reserved for the GenMAPP supported systems and cannot be used for any other system. Furthermore, GenMAPP does not allow the use of **any** one-letter code for added systems or systems in a new database. System codes in custom databases or for systems added to a supported database must consist of an **&** sign followed by a letter. For example, the system code for *MyArrayIDs* might be **&a**.

## 1.2 MAPP Basics

All GenMAPP MAPPs contain gene objects which are assigned gene IDs from one of the gene ID systems in the Gene Database with which the MAPP was created. Once a gene ID has been assigned to a gene object, any annotation from the database is automatically linked to the gene object. That annotation includes links to other gene IDs that are related to the primary gene ID through Relationship Tables in the database.

MAPPs can be created by hand or automatically. There are currently two methods of automatically creating MAPPs, using MAPPBuilder or using the Converter function. The MAPPBuilder program automatically creates simple MAPPs from lists of gene IDs. The lists are provided by the user as a text file. The finished MAPPs will display genes organized as lists rather than in a pathway oriented graphical display. Creating MAPPs using MAPPBuilder will not be discussed in this document.

Automatically creating MAPPs using the Converter function requires existing MAPPs as a template. In reality, new MAPPs are being converted from other MAPPs, rather than being created from scratch. Consequently, the new MAPPs retain any layout from the template MAPPs. The only aspect of the MAPP that is changed is the gene ID.

## 2 Preparing the homology data

### 2.1 Prerequisites

The process of creating new MAPP Archives described in this document relies on converting existing MAPP Archives from a related species. This is possible through the Converter function in GenMAPP. In order to create a MAPP Archive in this way, some specific pieces of information are necessary:

- A Gene Database for the related species from which MAPPs are being converted
- Homology information between the species of interest and the related species for which a MAPP Archive already exists
- MAPP Archive for the related species

As an example, this manual will focus on the conversion of human MAPPs to dog MAPPs.

**EXAMPLE:** For creating the dog MAPPs, we need the following:

- The human database from GenMAPP.org
- The homology information between dog gene IDs and human gene IDs
- The human MAPP Archives

#### 2.1.1 Gene Database

The Gene Database to use for the conversion is the database for the species for which MAPPs already exist. Using the homology information you will configure this database (or a copy of it) to use for the conversion. Most likely, the MAPPs being converted come from GenMAPP.org, in which case the database to use should also be supplied by GenMAPP.org. All GenMAPP databases for supported species are available through the GenMAPP program, under *Data > Download data from GenMAPP.org*.

**EXAMPLE:** Starting with the human MAPPs, we will use the human Hs-Std\_20050418\_beta database from GenMAPP.org.

If you have MAPPs from a different source, for example a collaborator, you will need to verify which database was used to create those MAPPs.

#### 2.1.2 Homology information

Since the basis of the procedure described here is to convert pre-existing MAPPs from a related species, it is essential to have homology information between the species of interest and a species for which MAPPs already exist. Several public resources have information on gene homologues between species. Some of these resources include:

Homologene: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>

Ensembl: <http://www.ensembl.org/>

You may also have homology information from other sources or from a computational method. Any homology information that you believe to be accurate will work with GenMAPP if formatted correctly.

The type of information that is necessary in terms of converting MAPPs is the relationship between gene IDs for the two species. Preferably, this information should be related to the gene ID system most commonly used on MAPPs, for maximum coverage.

**EXAMPLE:** For the dog MAPPs, we need the homology information relating dog to human Entrez Gene IDs, since this is the main gene ID system used on human MAPPs.

**Note:** This manual assumes that all gene IDs present on the human MAPPs are Entrez Gene IDs. If the MAPPs you are converting contain any additional systems, homology information for those systems is required for a complete conversion. The process of adding the homology information and performing the conversion must then be repeated for each additional system present on the MAPPs. This process is described under 3 below.

### 2.1.3 MAPP Archive

All GenMAPP MAPP Archives for supported species are available through the GenMAPP program, under *Data > Download data from GenMAPP.org*.

**EXAMPLE:** To create the dog MAPPs, we will use the human Hs\_Contributed\_20050427 MAPP Archive from GenMAPP.org.

## 2.2 Formatting the homology data as a Relationship Table

In order to perform the conversion of MAPPs between different species, GenMAPP requires three additions to the Gene Database being used:

- Gene Table for the gene ID system the MAPPs are being converted to
- Relationship Table containing homology information

**Note:** Specific information on how to parse homology data from Homologene will be available in a separate document.

The homology information should be added to the database in the form of a Relationship Table. A raw file for a Relationship Table must be formatted as follows:

- The raw relationship file must have two columns, but can also have more. The first is the gene IDs for one Gene Table (Primary or first column of your Relationship Table), the second is the Gene IDs for the other Gene Table (Related or second column of your Relationship Table). Any columns beyond the first two are OK but will be ignored. It is irrelevant which ID is designated as the Primary.
- The raw relationship file must be in tab-delimited (.txt or .tab) or comma-separated-value (.csv) from such as that exported from many spreadsheets and database systems. Gene IDs are limited to 50 characters.
- The columns should have no headings.

**EXAMPLE:** Dog – Human Relationship Table

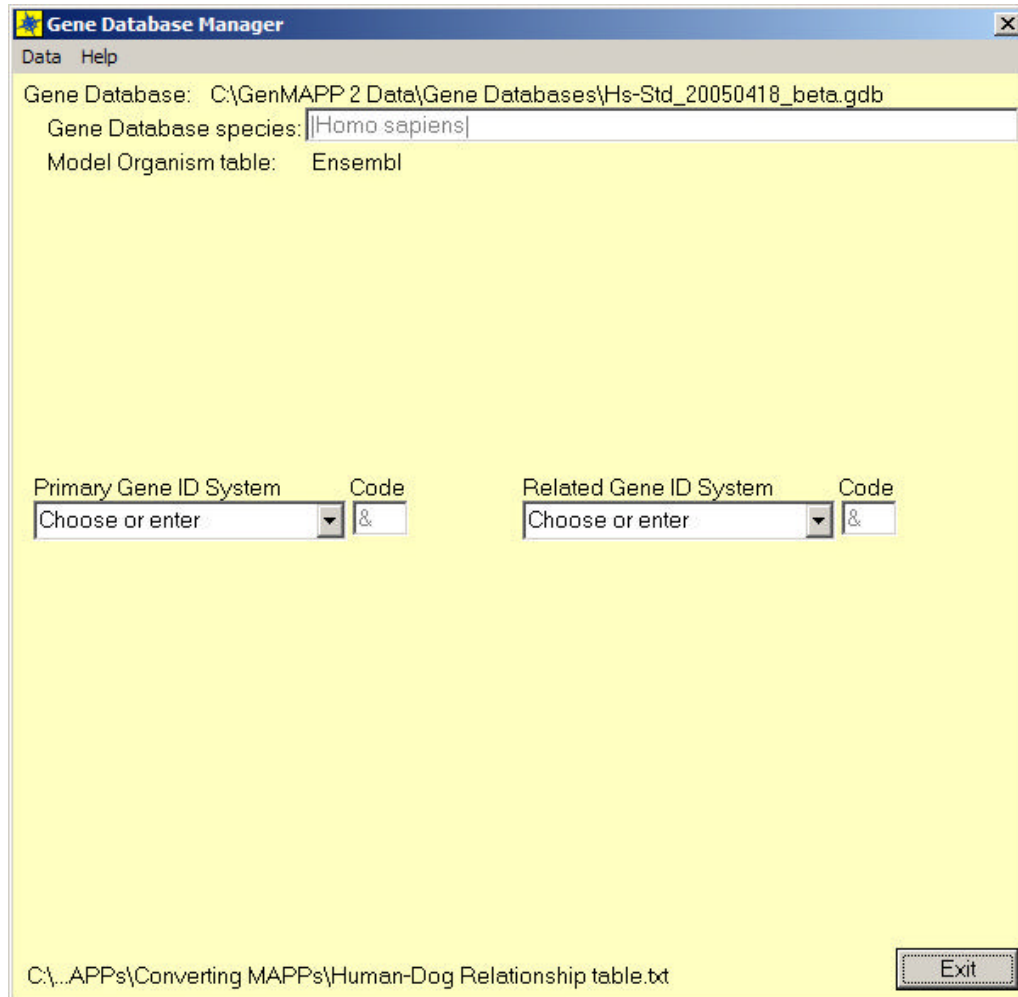
37	489463
38	489421
90	488362
158	474499
175	475638
212	491498

270	403872
350	403945

## 3 Configuring the Gene Database and performing the conversion

### 3.1 Adding homology information to the database

1. Launch GenMAPP. Under the *Data* menu, select *Gene Database Manager*.
2. In the Gene Database Manager, select *Data > Add New Relationship Table*. You will be asked for a raw data file to import. Select the raw data file containing the homology information, the Gene Database Manager shows you a screen similar to the following:



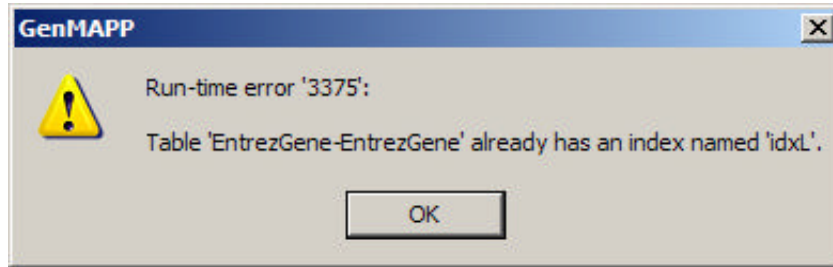
3. In the drop-down menus in the middle of the window, choose the name of the primary and related gene ID systems. The system codes will be automatically filled in for you.

**EXAMPLE:** For the homology information between human and dog, we will choose Entrez Gene in both drop-down menus.

4. Click the *Process* button to start import. If any errors are encountered, the Relationship Table is not created and a column explaining the errors is added to your raw relationship file. Correct the errors in each indicated row and import the data again.

A practical way of dealing with import errors is to load your raw gene file into a spreadsheet program, such as Excel, and filter the rows to view those in which the extra column is not blank. You can then easily make changes to erroneous entries. It is not necessary to remove the extra column before importing the file again. Be sure to save the result as a tab-delimited (*.txt* or *.tab*) or comma-separated-value (*.csv*) file.

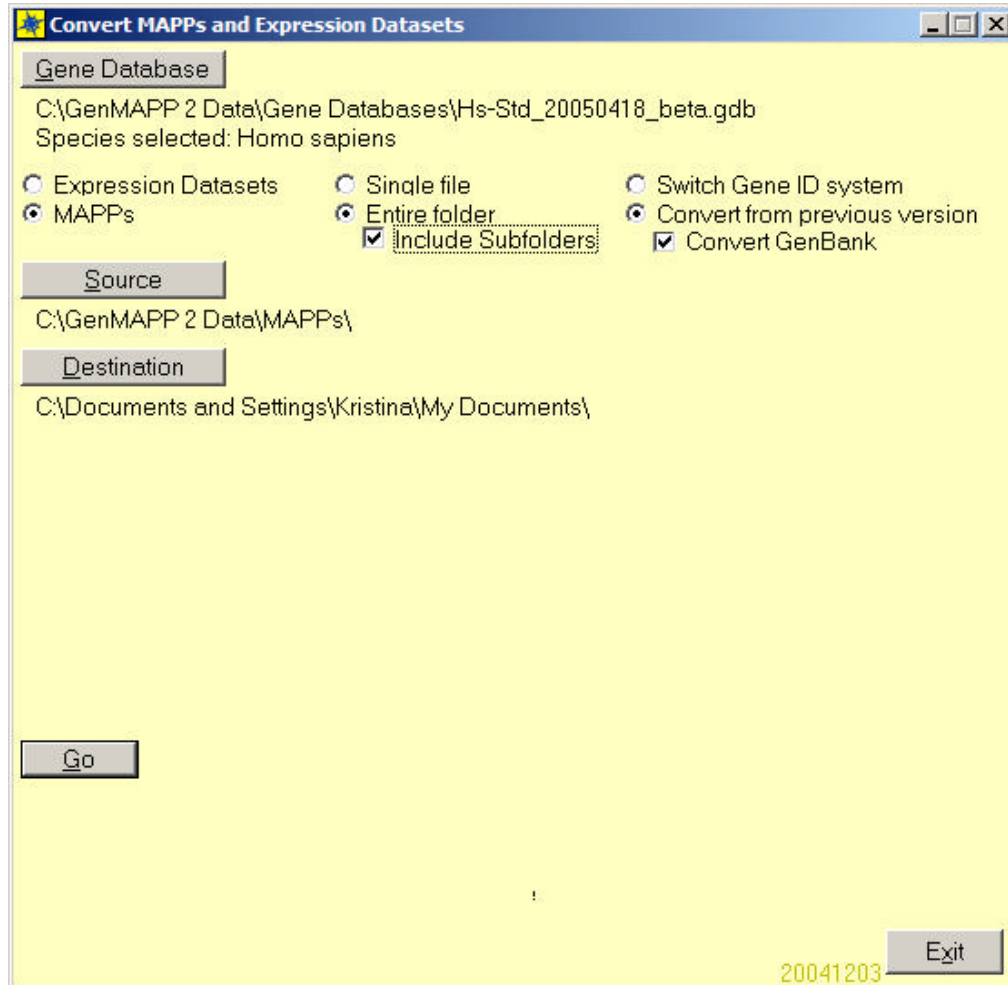
**Note:** At the end of import, you will get an error like this:



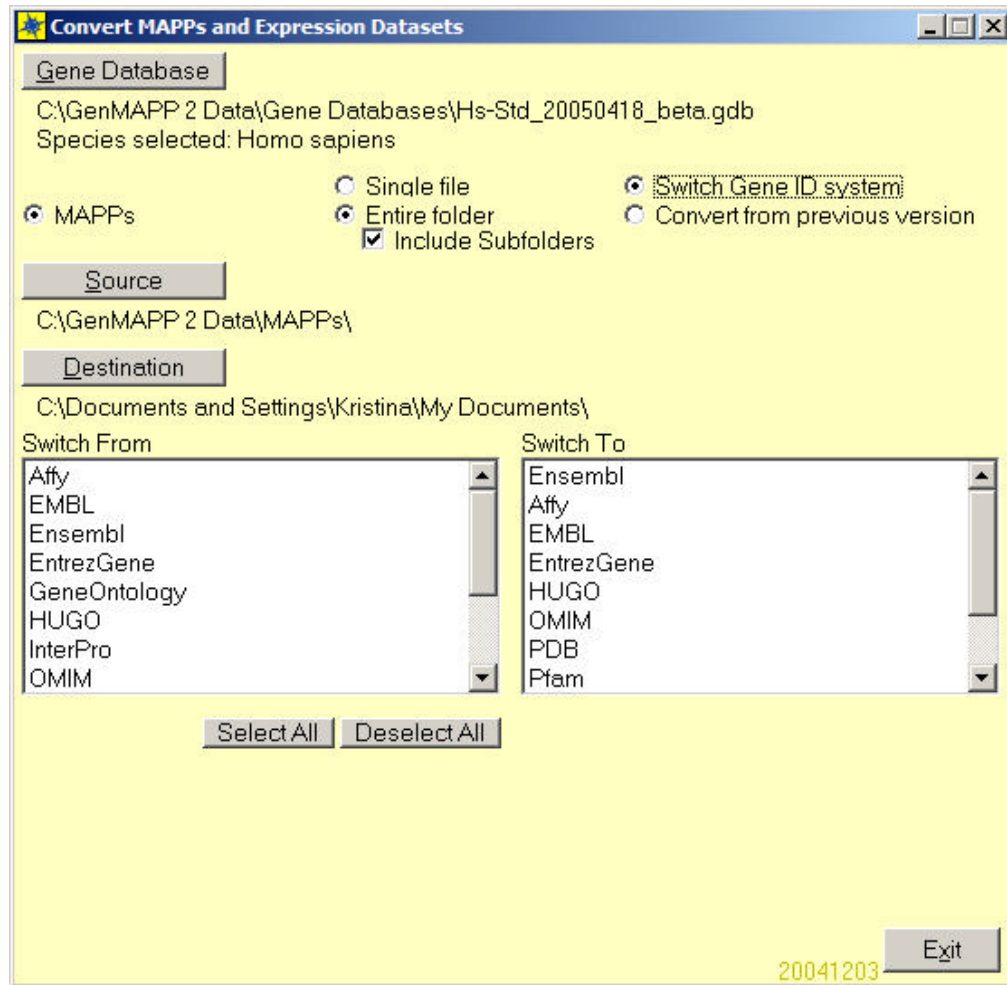
This error is expected and will not affect the downstream conversion. Once you click OK to the error, GenMAPP will close. Normally, a relationship is added between two different systems, not “between” the same system. Note that this error is not the same as any potential import errors encountered, which must be dealt with.

### 3.2 Performing the conversion

1. Launch GenMAPP again and choose *Data > Choose Gene Database*. Select the configured Gene Database.
2. Select *Tools > Converter* to launch the Converter function. The following window will appear:



3. In the upper left section, select *MAPPs*. If you are converting a complete MAPP Archive, also select *Entire Folder* and *Include Subfolders*. In the upper right section of the window, select *Switch Gene ID system*. The window will now change slightly:



4. Click the *Source* button and browse to find the folder containing the MAPP Archive to use for conversion. In the example, this would be the human source MAPPs.
5. Click the *Destination* button to select a place to store the converted MAPPs.
6. In the lower half of the window, select the gene ID systems to switch from and to. If your Gene Database doesn't contain the necessary relationship, the gene ID systems options will appear within brackets in the two boxes.

**EXAMPLE:** For the human to dog conversion, select to *Switch From* Entrez Gene and to *Switch To* Entrez Gene.

7. Click the *GO* button to start the conversion. You will be notified when the conversion is finished.

GenMAPP creates a log file for each converted MAPP, documenting the old ID and system as well as the new ID and system. If the program was unable to make a conversion, this will also be recorded in the log file. It is recommended that you view the log files after MAPP conversion to ensure that the conversion process was satisfactory. The log files are located in the same folder as the newly created MAPP files.

**EXAMPLE:** Log file for the conversion of a MAPP from human Entrez Gene IDs to dog Entrez Gene IDs

Gene Label	Old ID	Old System	New ID	New System
ORC6L	23594	Entrez Gene	482492	Entrez Gene
ORC3L	23595	Entrez Gene	474991	Entrez Gene
ORC5L	5001	Entrez Gene	No conversion made	
ORC4L	5000	Entrez Gene	476141	Entrez Gene
ORC1L	4998	Entrez Gene	475351	Entrez Gene
MCM7	4176	Entrez Gene	479737	Entrez Gene
MCM6	4175	Entrez Gene	476131	Entrez Gene
MCM5	4174	Entrez Gene	No conversion made	
MCM4	4173	Entrez Gene	477871	Entrez Gene
MCM3	4172	Entrez Gene	481839	Entrez Gene

For genes where no conversion was possible, the original gene ID will remain on the MAPP and will not link to the database for which the newly created MAPPs were created. This will not cause a problem in linking data to the MAPPs, but the genes with the original IDs will not link to data.

**EXAMPLE:** For the new dog MAPPs, any genes that were not converted will retain their human Entrez Gene ID. These IDs will not link to the GenMAPP dog database.

### 3.2.1 New MAPP Name

GenMAPP doesn't automatically change the name of the new MAPP files. Since the MAPPs will most likely have a two-letter species prefix in the file name, it is recommended that you change the names of all the converted MAPPs before use to avoid confusion.

**EXAMPLE:** The converted dog MAPPs will have the prefix *Hs* for Homo sapiens. This prefix should be changed to *Cfa* before using the MAPPs.

The MAPPs used as templates in the conversion will automatically acquire the prefix *Old\_*. This might cause confusion and should be deleted from all MAPPs used, to maintain the original name of the MAPPs.