



User Manual

for version 14.1.0

Contents

Introduction	2
Copyright notice	2
Requirements	2
Installation instructions	3
Usage	3
Limitations	5
For programmers	5
Command line options	5
File format specs	5
Bug reports and contact information	5

Introduction

This program was written by Ivan Potapenko (dr-ivan@dr-ivan.com). A huge thanks goes to Lars O. Baumbusch, Jørgen Aarøe and the rest of the CGH team at Dept. of Genetics (Oslo University Hospital Radiumhospitalet) who helped with testing it and suggested a number of useful features. I would also like to thank Ole Christian Lingjærde for encouragement and valuable discussions.

Feature Extraction (FE, www.chem.agilent.com) is Agilent's software suite geared towards extraction and analysis of CGH array data. However, in parallel with advancement of CGH technology, a number of other valuable tools have been created to analyze, visualize and otherwise process the data retrieved from such analyses. Unfortunately, not all of them are able to import data files from FE directly. One such program is CGH Explorer (CGHe, <http://www.ifi.uio.no/forskning/grupper/bioinf/Papers/CGH/>).

Fe2cghe (pronounced "eff-ee-to-see-gee-aitch-ee") is a small converter designed to solve this problem. Furthermore, some useful additional features have been implemented as well to alleviate the need to manually alter any aspects of the raw data.

1. Merging of any number of files
2. Raw data files may be derived from arrays of any size
3. Removing unnecessary columns
4. Splitting columns (chromosome number, start and stop)
5. Removing flagged values
6. Removing controls
7. Sorting data by two criteria: chromosome number and start position
8. Merging duplicate entries
9. Checking file consistency (cross-matching ProbeUID-s and number of lines)

After the processing, results are written to a master file which should contain all data needed by CGH Explorer and can be imported right away. Output files may also be used for other purposes as well.

This program can be downloaded from: <http://www.dr-ivan.com/computer/fe2cghe> . Here you will also find an updated version history.

Copyright notice

Fe2cghe is © Ivan Potapenko, 2007-2011.

This program is distributed under GPL (GNU Public License) version 3 (<http://www.gnu.org/copyleft/gpl.html>). You can modify and/or redistribute it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

Read the *GPL.txt* file included with **fe2cghe** for more information on licensing.

Requirements

You will need a Windows or Linux (or other *nix) machine to run **fe2cghe**. For the GUI to work you will need an up-to-date installation of Java Runtime Environment (www.java.com). Please ensure that you have enough free RAM and disk space at your disposal (see below). For installation on *nix-based operating systems you will also need the GNU C++ Compiler of any recent version.

Operating systems: Theoretically, the core program will run on any operating system and the GUI will run on any OS which support java. This therefore also includes Mac OS. However, no pre-compiled binary for this platform is supplied – please compile the program yourself.

Architectures: This program will run on any architecture. The Windows executable bundled with the software is however **32-bit**. If you have more than 4GB of RAM and a 64-bit Windows machine, you will have to recompile the executable to fully exploit computational power of the machine.

Memory: The command line utility requires a minimum of 100MB RAM for a raw data file of average size. GUI will need additional requirements since JVM can be rather resource-intensive.

Harddrive space: It is difficult to calculate exact amount needed (as number of removed entries and LogRatio-values can vary), but as an example 125 files of 200MB each need an approximate amount of 400 to 800MB of free disk space (including the output master file being ~400MB). This value will also vary depending on your cache settings.

Installation instructions

Windows: Unpack the archive you have downloaded from <http://www.dr-ivan.com/computer/fe2cghe> . In the folder created

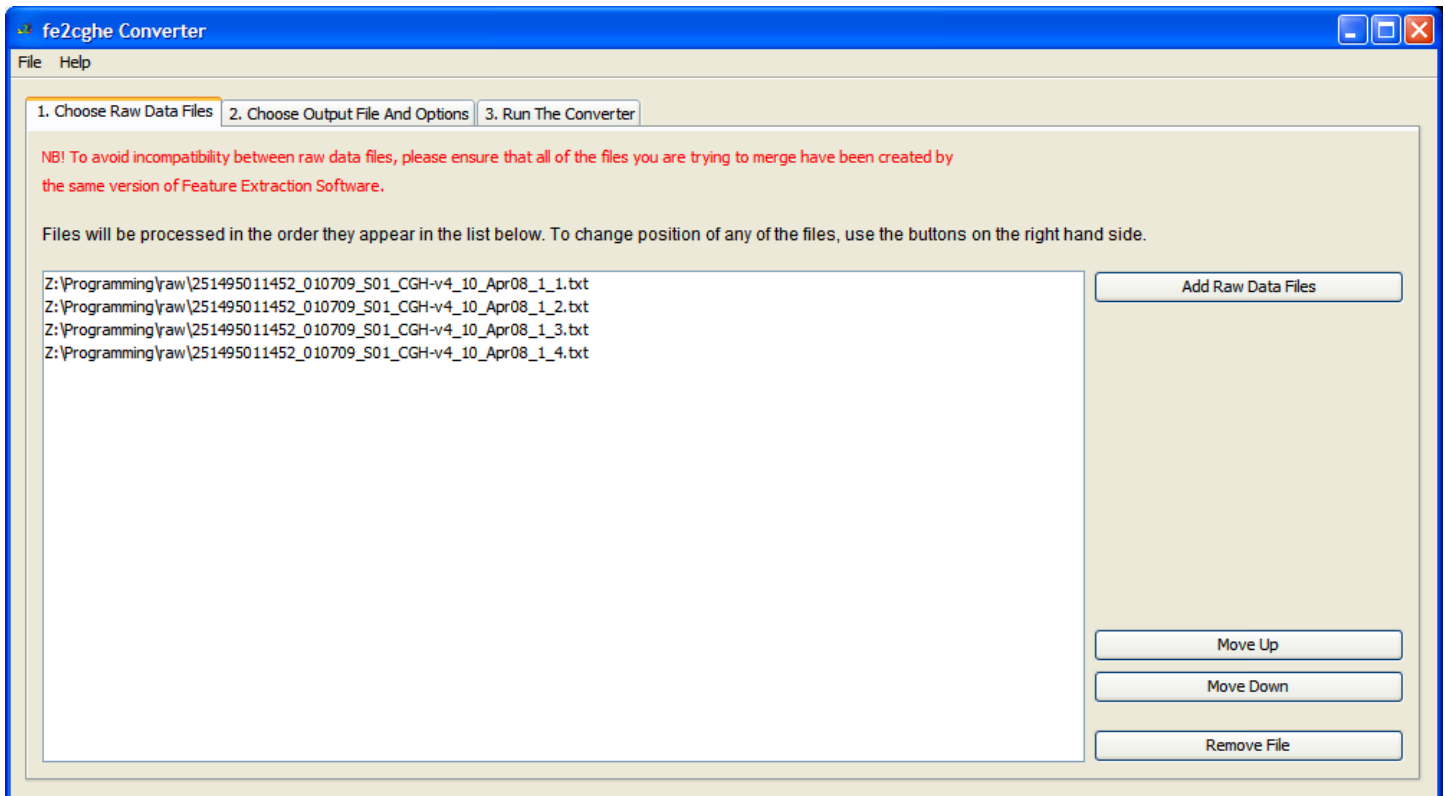
Linux: Unpack the archive you have downloaded from <http://www.dr-ivan.com/computer/fe2cghe> . Before you can run **fe2cghe**, you need to compile the core part of the program. Assuming you have unpacked the program to `/home/user/fe2cghe-14.0.2/`, do the following from the command line

```
# cd /home/user/fe2cghe-14.0.2/  
# cd src/core  
# make
```

You may see a number of warning messages. This is normal and does not affect functionality of **fe2cghe**.

Usage

Browse to the folder which you extracted from the downloaded archive. Inside this folder you will find `fe2cghe.jar`. Run this file and you will be presented with the following screen:



On this first screen you will be able to select your raw data files and organize them in the order you would like them processed. Please ensure that all files are exported by the same version of Feature Extraction to avoid any incompatibility that might occur when Agilent re-classifies the probes. When you are done, click on the next tab.

On this last tab you should press "Run" to invoke the processing of the raw data files. While the program is busy, you will get real-time updates on the progress in the text area at the middle of the screen. All error messages will also be displayed here.

A small notification window will pop up once the processing is finished.

Limitations

Probe classification is frequently updated by Agilent. To avoid incompatibilities between your files it is highly recommended to use the same Feature Extraction version to export all files you want to merge.

Some more obscure limitations: Description field in the raw data files cannot be larger than a predefined (at compilation) constant. By default it's 1000 characters, but can be altered easily by modifying variable MAX_DESCR_LENGTH in the upper part of the source code. Number of lines in a raw file may NOT exceed 1 073 741 823. This limitation is imposed by C++ STL and thus may be compiler dependent. Neither of these appear to pose any problems currently.

For programmers

Below you will find some miscellaneous information which may or may not be interesting for developers.

Fe2cghe consists of two parts. A pre-compiled core utility which is written in C++ for speed and efficiency, and a graphical front-end designed and programmed in Java using NetBeans IDE version 6.9. All sources are located in the *src* folder of the application.

Windows executable was compiled on a Windows XP x86-32 machine using MinGW GNU C++ Compiler.

Command line options

As built-in help function states, command line options are as following:

```
# fe2cghe cache_size consistency_check dupe_collapse ignore_unkn_chr file1 [file2...] output_file
```

cache_size:	number indicating number files in RAM simultaneously (recommended: 5)
consistency_check:	Check files for consistency? 0 - no, 1 - yes (recommended: 1)
dupe_collapse:	Collapse duplicated values? 0 - no, 1 - yes (recommended: 1)
ignore_unkn_chr:	Ignore lines with unknown chromosome code values? 0 - no, 1 - yes (recommended: 1, but might be unsafe!)
file1 [file2...]:	Any number of raw data files to process. Wild cards are not accepted
output_file:	File to use as output

In other words; first the executable name, followed by a numerical cache size, 0 (no) or 1 (yes) for consistency check and 0 or 1 for collapsing duplicate values. After this any number of input file names can be added. The last parameter should always specify the output file.

An typical example would be:

```
# /home/user/fe2cghe-14.1.0/src/core/fe2cghe 5 1 1 1 /home/user/raw_data/FE_file1.txt /home/user/raw_data/FE_file2.txt /home/user/output.txt
```

File format specs

To read about FE file format, please refer to Feature Extraction Reference Guide which can be downloaded from Agilent's web pages (www.chem.agilent.com).

Output file format is as following. All values are delimited by tabs. The first row contains a header naming the columns. The first seven are ProbeName, ProbeUID, GeneName, Chromosome, Start, Stop, Description. For explanations of these parameters, please refer to the FE Reference Guide (see above). Subsequent columns contain data from the processed arrays. Header of each column contains file name of the raw data file. For each probe on each array a log₂-transformed log ratio is reported.

Bug reports and contact information

If the program crashes during execution, not all process is necessarily lost. In this case look for a **.tmp* file in the directory which contains raw data files. This temporary file may contain cached data. Therefore there may be a chance some files have already been processed and saved.

When a crash occurs, please feel free to contact me (Ivan Potapenko) at dr-ivan@dr-ivan.com. When doing so, provide as complete log of actions as possible, file that was being processed at the time of the crash and full error message (if any). Also, include what settings you were using when this situation occurred.