Irys Extract manual

Witten by Dena Laadan Friedman and Rani Arielly

Table of Contents:

Introduction

This guide is designated for new users of "Irys Extract", and its goal is to give an overview of the various files that we use during the work, and serves as an explanation of how to use it.

The software code was written by Dr. Rani Arielly in order to provide additional capabilities for decoding images of DNA and provides many advantages over the IRYS interface:

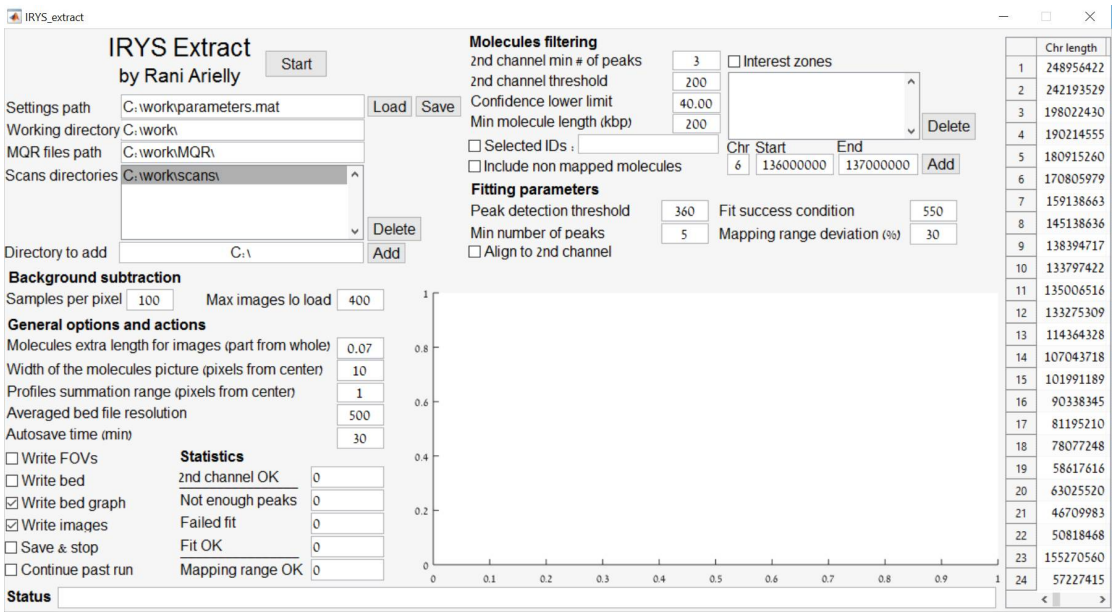
- The images of the molecules are outputted automatically
- More flexible and effective background subtraction
- The intensity profiles of the examined molecules are outputted automatically
- A significant reduction in the amount of false positive \ false negative detections
- A possibility to add additional capabilities on an immediate requirement

Preliminary requirements

- This software performs its operation by loading and processing files as large as gigabyte, therefore it should be run on a computer with at least 4GB of memory.
- Software runtime depends significantly on files reading times, therefore it is recommended to insist on a fast interface for storing the given files.
- This software uses MATLAB software functions which should be supplied by installing "MATLAB runtime R2015a". It can be downloaded at: http://www.mathworks.com/products/compiler/mcr/
- After installing MATLAB runtime R2015, the software Irys_Extract.exe downloaded with this manual can be run.
- In order to perform an example run as shown in the "training set" part, the test data attached in the website should be downloaded.
- In this manual image files are loaded and refined in the software "ImageJ". It is recommended that you install it as well. It can be found at: https://imagej.nih.gov/ij/

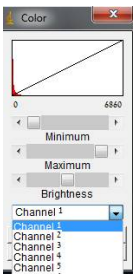
A quick Guide to run the training dataset

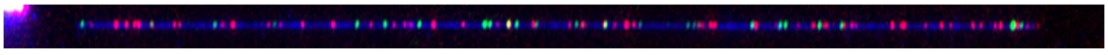
1) Create a working directory, for example: "c:\work".
2) Download the files into that directory: "Irys_Extract.exe", "parameters.mat", "demo data.rar".
3) Extract the contents of "demo data.rar" into the working directory, so that you'll have MQR and scans subdirectories. For example: "c:\work\MQR" and "c:\work\scans".
4) Load "Irys_Extract.exe", and wait until the following screen appears:



5) If necessary, replace the data in the "working directory", "MQR files path" and "scans directories" to be the paths for the working directory, MQR, and scans subdirectories. For example: "c:\work\", "c:\work\MQR\" and "c:\work\scans".
6) Press "Start"
7) The software will locate and construct the molecules that are listed in the MQR. After it finishes (indicated by the status massage "Done processing # molecules out of # from a total of #"), find the files "chr6-45.18M-45.69M….tiff" (partial name) at the "molecules_images" subdirectory and "channel1_1.txt" and "channel2_1.txt" at the "bedgraphs_reduced" subdirectory. Upload the TIFF file in the ImageJ software, initially it will look like this:

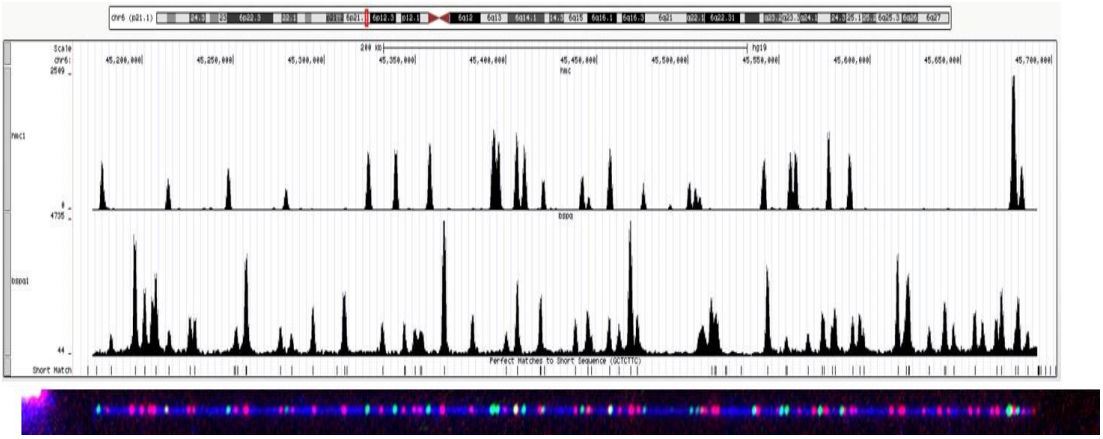
8) In order to edit it, press "image", then "adjust" and finally "Color balance". Another window will be opened as shown on the right. Only the first three channels are important in this window, where channel 1 is the red channel (BspQ), channel 2 is the green channel (5-hmC) and channel 3 is the blue channel (YOYO-1). Adjust the minimum, maximum and brightness parameters in each channel to your likings and receive a clearer image, such as:





Uploading information to genome browser

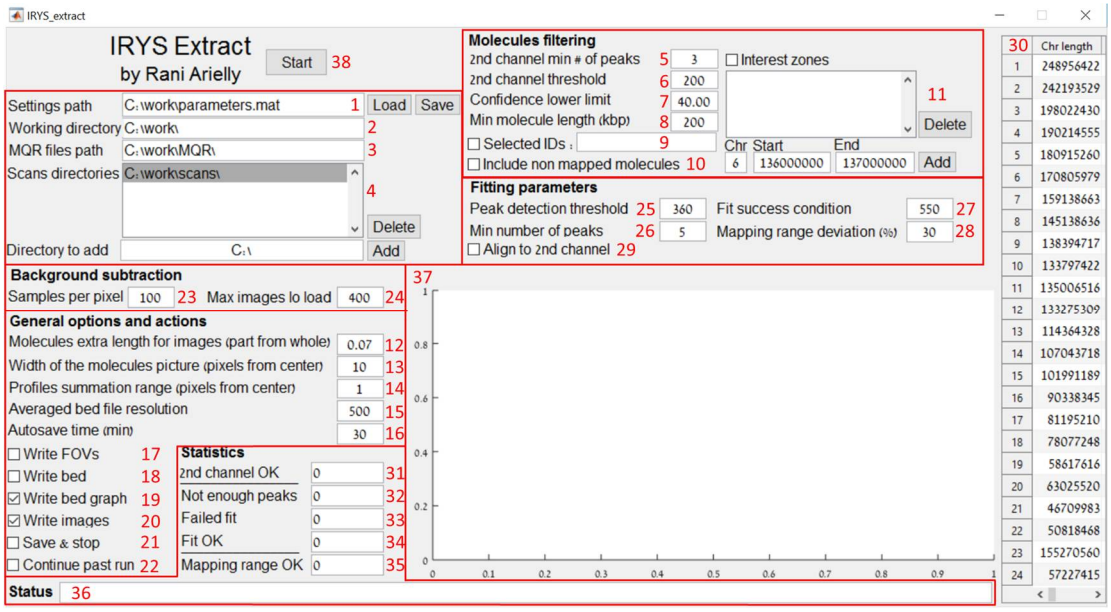
9) Enter the website: Https://genome.ucsc.edu/index.html and open a profile in the website by clicking "my data> my sessions> create an account" and follow the instructions on the website in order to complete the registration. (the following steps can be done also without opening a profile).
10) Now return to the homepage of the website and in "our tools" enter the genome browser.
11) Make sure that in the "human assembly" field, "Feb.2009 (GRCh37 / h19)" is marked and press "Go".
12) Press "add custom tracks", which is located below the image at the center of the screen.
13) In the "Paste URLs or data", select "browse" and locate and select the previously outputted "channel1_1.txt" file, and press submit. Press the "add custom tracks" button again and repeat this process with "channel2_1.txt". When finished, press go to view in Genome Browser.
14) Now the image which is in front of us contains two channels. By right clicking the gray bars on the left and selecting "full", we will get the molecule profile.
15) Using the "zoom in" and "zoom out" buttons, moving the image and centering the profiles, we can optimize the view until we get something like the following image:



The image shows several layers of information. From top to bottom: the chromosome and the position upon it, the 5-HmC profile, the BspQ profile, theoretical lines for the sequence GCTCTTC, and on the bottom – the molecule image (pasted manually). We can see how the BspQ profile fits the theoretical lines for the GCTCTTC sequence, and it is possible to add different channels and search for correlations between the intensity profiles and different biological characteristics.

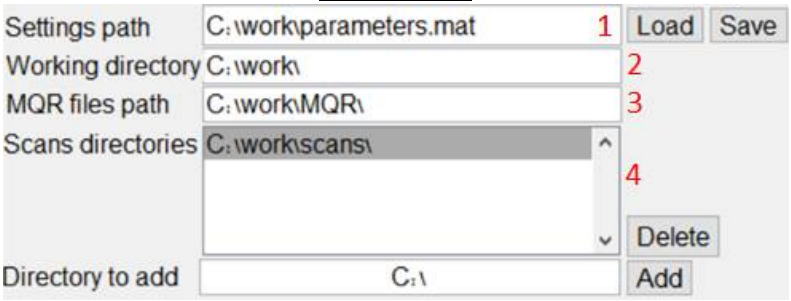
The software interface

The software window can be divided to several parts as shown in the image below.



Specifying the different parts in the interface:

Directories



1) **Setting path** – The location of the settings file, when changing the address and pressing "load", the requested file will be uploaded, while pressing "save" will save the new settings in the mentioned address. If a file named "parameters.mat" is present in the directory of the program, it will auto-load when loading the program.
2) **Working directory** – The main folder for saving the output files as well as other files needed for the program.
3) **MQR files path** - The location of the Xmap, Cmap and Bnx files. It is advised not to change the name of the files and keep them as:
    - MoleculeQualityReport.xmap
    - MoleculeQualityReport_q.cmap
    - MoleculeQualityReport_r.cmap
    - Molecules.bnx

    It is possible to process several MQRs together by putting the MQRs in subdirectories. For example:

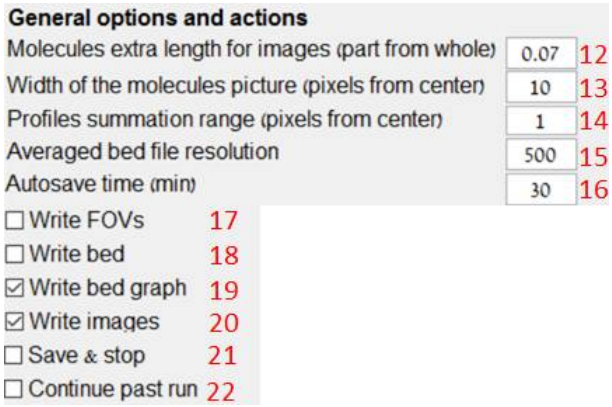
    "C:\work\MQRs\1\", "C:\work\MQRs\2\", etc.

4) **Scans directories** – The location of the run folders than contain the scans files (original images, xml files, mol files and fov files). Locations can be added by entering a directory in "Directory to add" and pressing the "Add" button. No need to list each run folder – just the main folder that contains them all. In each run folder, the *.mol and *.fov files should be in a subdirectory called "detect molecules". For example:
   - c:\work\scans\ecoli-bac_2014-04-29_17_48\ecoli-bac_2014-04-29_17_48_Scan01.tiff
   - c:\work\scans\ecoli-bac_2014-04-29_17_48\ecoli-bac_2014-04-29_17_48 Metadata.xml
   - c:\work\scans\ecoli-bac_2014-04-29_17_48\detect molecules\Molecules1.mol
   - c:\work\scans\ecoli-bac_2014-04-29_17_48\detect molecules\Stitch1.fov

Molecules filtering

5) **2nd channel min # of peaks-** Filtering the molecules by a minimum number of peaks in the second channel.
6) **2nd channel Threshold-** Peak detection level threshold in the second channel.
7) **Confidence lower limit -** Filtering the molecules by their fit score.
8) **Min molecule length (kbp) -** Filtering the molecules by their minimum length.
9) **Selected IDs-** When entering identification numbers (separated by a comma) of specific molecules, the run will be reduced to these molecules only.
10) **Include non-mapped molecules –** The database will include molecules that weren't mapped into the genome.
11) **Interest Zones -** Specific areas in the chromosome to which we will limit the search. Several positions can be added by typing the desired area in "zone to add" by the chromosome number, the position of the beginning and the end of the region (at the bases) and clicking on the "Add" button.

General option and actions

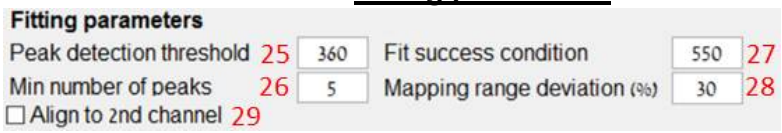


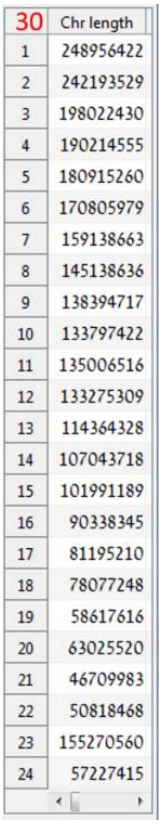
12) **Molecules extra length for images (part from whole)** – Margins along the length of the molecule image can be added to it. This is the addition ratio from the length of the molecule.
13) **Width of the molecules image (pixels from center)** – Margins along the width of the molecule image can be added to it. This is the number of pixels to be added from each side along the width of the molecule.
14) **Profile summation range (pixels from center)** – The number of pixels in each direction along the width of the molecule which we will sum in order to produce the molecule intensity profile.
15) **Averaged bed file resolution** – Resolution for files containing the averages profiles (in bases units).
16) **Auto save time (min)** – After the value specified in minutes, an automatic saving of the files will be performed.
17) **Write FOV** – If marked, the software will output a stitched image of the full field of views that contain the molecules.
18) **Write bed** – If marked, the software will output Bed files.
19) **Write bed graph** – If marked, the software will output Bedgraph files.
20) **Write image** – If marked, the software will output the molecules images.
21) **Save and stop** – Ending the processing of the current molecule and saving the run (it is possible to continue the run at another time from the same place it was stopped).
22) **Continue past run** – Continuation of the last run from the last molecule which the software tested in the former run, in case this one is not completed.

Background subtraction

23) **Samples per pixel** – The number of times to sample each pixel in order to create a background profile for the image.
24) **Max images to load** – The maximum number of images to upload in order to create a background profile for the image.

Fitting parameters





25) Peak detection threshold – Peak detection level threshold in the 1st channel.

26) Min number of peaks – Filtering of molecules by a minimum number of peaks in the 1st channel.

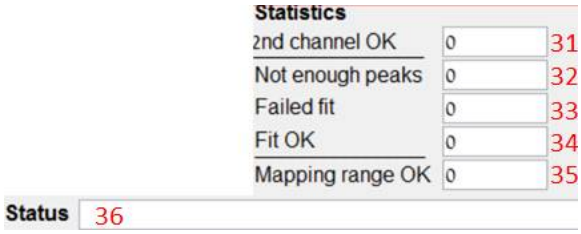
27) Fit success condition – The average distance (in bases units) between peak positions and their corresponding reference locations, which beyond, the fit should not be used, while beneath it, the fit is good.

28) Mapping range deviation (%) – The maximum value in percentages where the molecule can be larger than the mapped part.

29) Align to 2nd channel – To be used in the case where the channels used for alignment and as a second layer of data are taken at a reversed order.

30) The table of the chromosome lengths for testing that the molecules which we mapped, do not exceed the chromosomes edges.

Status & statictics



31) 2nd channel marked – Number of molecules that passed parameter 5.

32) Not enough Peaks- Number of molecules that failed fitting due to parameter 26.

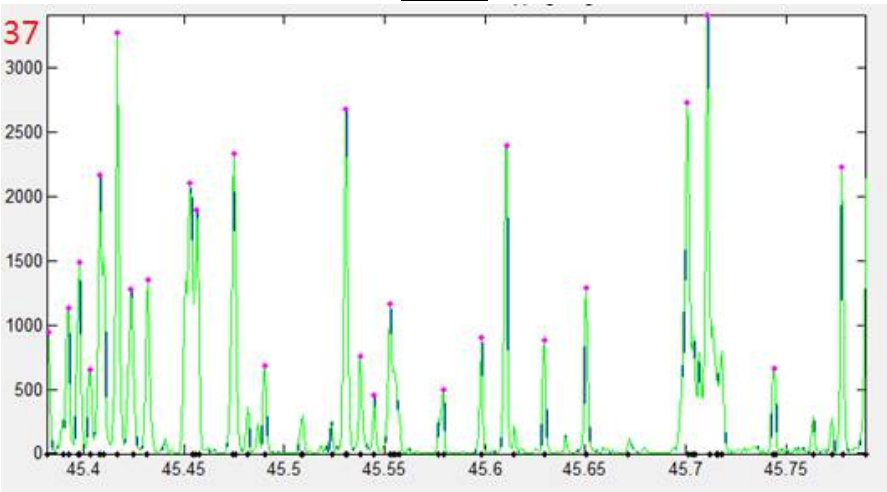
33) Failed Fit – Number of molecules that failed fitting due to parameter 27, the distance was above the possible defined average value.

34) Fit OK – Number of molecules that were fitted successfully.

35) Mapping range OK – Number of molecules that were successfully fitted and did not exceed the range of the chromosomes or the value of parameter 28.

36) Status – Information line regarding what is happening in the software during a run.

Graph



37) Graph – This graph display the fitting of the current molecule's intensity profile against the reference. The blue graph is the original graph out of the fit data, the green graph is the result of the program's fitting process, the purple dots represent the position of the peaks in the molecule and the black dots indicate the known positions in the reference.

Start button

38) Start – start the run.

Input and output files

Input files

In order to work with the software, number of files received from the IRYS system are required.

Xmap files – This is the main file from which the processing begins (this is the file outputted when doing "molecule quality report"). The important information in this file is the identification number of the molecule and how to locate it on the genome. This file contains explicitly information such as: the chromosome to which the molecule belongs, the base from which there is a fit and the base at which the fit ends, these bases' positions, the molecule's orientation with respect to the genome and the fit score. These files are located in the "molecule quality report" folder and their file name is "MoleculeQualityReport.xmap".

Cmap files – These are two types of files: one file is for the molecules - marks' positions on each molecule, and the second file contains information regarding the position of the marks on each chromosome. The role of these files together is to provide information about the molecule fit to the position on the genome. These files are in the "molecule quality report" folder and their names end with "_q.cmap" and "_r.cmap".

BNX files – This is a file that contains supplemental information and links the information in the xmap file to the information in the mol file. Although these two files come out as output from the IRYS, they identify the molecules differently, and therefore a file which will link them is required. This file contains additional information, but it is not relevant for us. These files are in the "molecule quality report" folder.

MOL files – Files which contain information to identify the molecule position in the images files: in which file it is located, and also which image and pixel at which the molecule begins and ends. The information stored in these files is used to create the final image, and in addition there is information on the length of the molecule used to compare and test that the examined molecule is the molecule we want. These files are located in the "Detect Molecules" sub-folder in the run folder, and the file name is "molecules*.mol".

FOV files – The IRYS machine photographs a number of frames which slightly overlap in each channel, this file contains information describing how to attach each image to another image in order to create one large image describing reliably the nano channels chip. The information is: file index, number of frames that were taken, the camera angle with respect to the channels out of the photographed image, the number of pixels that the image needs to be shifted in the transverse and lengthwise direction in order to receive the correct overlap. These files are located in the "Detect Molecules" sub-folder in the run folder and their file name is 'Stitch*.fov'.

TIFF files (FOV images) – These are image files of all the frames that were taken in IRYS in a crude manner. For each field of vision, there are three images, so that every time, the photography is under exposure to light of a different color. These files are located in the run folder.

XML files – These files contain information about the conditions of the measurements. Most importantly they contain the optical magnifications of the different channels that allows the program to properly overlay the different channels. These files are located in the run folder.

Output files

Text files – There are two types of text files coming out as output from the software: "Bed" contains only the information regarding the peaks, therefore the information in it is not continuous, while the other one, "Bedgraph", contains all the information regarding the peaks continuously, so a graph is created. The files names contain "channel1" or "channel2" to indicate the type of color channel (in some cases it is abbreviated as "ch1" and "ch2"). Bedgraphs are saved individually for each molecule in the "bedgraphs" subfolder and collectively in the "bedgraphs_reduced" subfolder and in the main folder (in an averaged form).

Tiff files – the stitched image of the molecule. These can be found in the "molecules_images" subfolder.