GenomeFlow User Manual

Contents

Table of Figures

1 Installation

1.1 Quick Start

1. Verify that you have installed the **basic dependencies**.
2. To use the 1D-Function that provides reference genome indexing, alignment of fastq files and filtering of alignment files, follow the instructions here for the dependencies download and installation. **NOTE** ***This step is required only for the 1D-Function tools***
3. Download the latest GenomeFlow Tools jar
4. Run GenomeFlow:
    - Windows OS: double-click the **genomeflow.bat** script
    - Linux/UNIX based OS: execute the script, **genomeflow.sh**

1.2 Basic dependencies

Java 1.7 or 1.8 JDK. (Alternative link for Ubuntu/LinuxMint). Minimum system requirements for running Java can be found at http://java.com/en/download/help/sysreq.xml .

These dependencies support only the 2D-Functions and 3D-Functions Tools.

1.3 Dependencies and Installation

1.3.1 Operating System (OS)

A Linux, UNIX, or Mac OS X environment is required to use the 1D-Functions.

It is strongly recommended to work under a Mac OS X or a Linux/UNIX-based operating system, such as Ubuntu, Centos/Red Hat, Solaris.

If you are using a Windows operating system, install Cygwin first. Cygwin is a free software that provides a UNIX-like environment on Windows. The Cygwin install package can be found at http://www.cygwin.com/. Once Cygwin is installed, place your work in the Cygwin directory.

1.3.2 Download External Tools

- Download Bowtie2 (*http://bowtie-bio.sourceforge.net/index.shtml*) **OR** Download BWA (http://bio-bwa.sourceforge.net/) for indexing and alignment creation.

- Bowtie2 supports multiple OS, download the version for your OS. That is:
    - Download bowtie2- version number-macos-x86_64 for MacOS
    - Download bowtie2- version number- linux-x86_64 for Linux
    - Download bowtie2- version number- mingw-x86_64 for Mingw/Cygwin
- Download Samtools (*http://samtools.sourceforge.net/*)
- We tested on the following versions for each one of the tools: bowtie2-2.3.4-*, bwa-0.7.17, and samtools-1.6.
- You can also download the installation files for these tools from here: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/External_Tools/

1.3.3 Installing External Tools

1.3.3.1 Bowtie2

- Open a Unix Terminal
- Change directory to the downloaded Bowtie2-* directory. For example:
    - cd bowtie2-2.3.4-linux-x86_64
- Give executable permission to the binary file. For example:
    - In Unix based Operating system/ Cygwin/Mingw: chmod +x **bowtie2***

1.3.3.2 BWA

- Open a Unix Terminal
- Change directory to the downloaded bwa-* directory. For example:
    - cd bwa-0.7.17
- Type **make** once you are inside the bwa directory. For example:
    - make
- This operation produces a binary file: **bwa**
    - In Unix based Operating system: **bwa**
    - In Cygwin/Mingw: **bwa.exe**
- Give executable permission to the **binary file**. For example:
    - In Unix based Operating system: chmod +x **bwa**
    - In Cygwin/Mingw: chmod +x **bwa.exe**

1.3.3.3 Samtools

- Open a Unix Terminal
- Change directory to the downloaded samtools-* directory. For example:
  - cd samtools-1.7
- Type **./configure** once you are inside the samtools directory. For example:
  - ./configure
- After configuration is completed, type **make**. For example:
  - make
- This operation produces a binary file: **samtools**
  - In Unix based Operating system: **samtools**
  - In Cygwin/Mingw: **samtools.exe**
- Give executable permission to the **binary file**. For example:
  - In Unix based Operating system: chmod +x **samtools**
  - In Cygwin/Mingw: chmod +x **samtools.exe**

1.4 Create an index for a reference genome

1.4.1 Purpose

To build an index for the reference genome data. Indexing the reference genome makes querying fast, and can also compress the size of the genome data

1.4.2 Input

The reference input FASTA file (usually having extension fa, mfa,.fna or similar). Read more about FASTA files here: https://en.wikipedia.org/wiki/FASTA .

1.4.3 Test Data

The human hg19 genome data (hg19.fa) can be downloaded from here: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/hg19_genome/hg19_genome_FASTA/

1.4.4 Output

It generates a shell script with the name *Indexer_script.sh*.

1.4.5 Output of script:

A list of index files. This varies depending on the tool selected for indexing. BWA output 5 files (NAME.amb, NAME. ann, NAME. bwt, NAME .pac, and NAME.sa), where NAME is a prefix string, and Bowtie2 outputs 6 files (NAME.1.bt2, NAME.2.bt2, NAME.3.bt2, NAME.4.bt2, NAME.rev.1.bt2, and NAME.rev.2.bt2) where NAME is <bt2_base>.

1.4.6 Test Data Output

Generated index for the hg19 human genome by bowtie2 and bwa tools can be downloaded from here:
http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/hg19_genome/

1.4.7 Running

- Access the function from the menu toolbar: 1D-Functions/Build index for reference genome
- Generate a script called Indexer_script.sh
- Open a Unix Terminal
- **Execute** *Indexer_script.sh* in a Unix Terminal

- Note to Bowtie2 and Cygwin/MinGW Users: To use Bowtie2 in Cygwin/MinGW, the absolute path to the input file generated from GenomeFlow might produce *"*Warning: *Could not open read file"* for some users. Use a relative path to the input file to locate the file by editing the generated GenomeFlow script.

Figure 1: Create an index for a reference genome

Field	Description	Default
Input Reference Genome file	A reference genome file having extension. fa, .mfa, .fna or similar. For example human genome(GRCh37/hg19)	NA
Output Directory	The output directory path to output the script	NA
Choose tool to use	Two options are made available for indexing. Select bwa-Burrows-Wheeler alignment or Bowtie2.	bwa
Binary file	Browse and select the binary file for the chosen tool BWA: Select the **bwa** binary you compiled from the bwa-* directory Bowtie2: Select the **bowtie2-biuild** binary file from the bowtie2-* directory	NA
Number of threads	This option is available only for the **bowtie2-biuild** indexer. Specify the number of threads to use for this task. More threads means less processing time taken.	8
Generate Scripts	This button generates a shell script (.sh) that can be executed in a UNIX terminal by the user.	

1.5 Mapping the raw single or pair read FASTQ files

1.5.1 Purpose

To perform alignment of the index and a set of sequencing read files.

1.5.2 Input

A FASTQ read files usually with extension .fq or .fastq. Read more about FASTA files here: https://en.wikipedia.org/wiki/FASTQ_format

1.5.3 Test Data

Test datasets can be found here:

- MiSeq GM12878 in-situ files: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/MiSeq_GM12878_Data/
- A karyotypically normal human lymphoblastoid cell line (GM06990) from Aiden et al: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GM06990_Data/

1.5.4 Output

It generates a shell script with name *Mapper_script_bowtie2.sh* for bowtie2 tool and *Mapper_script_bwa.sh* for bwa tool.

1.5.5 Output of script

The output will be found in a folder bowtie2_align for bowtie2 and bwa_align for bwa. By default, the output BAM file is named *bwa_mapped.bam* for bwa and named *bowtie2_mapped.bam* for bowtie2.

A BAM binary format (.bam) obtained by converting a SAM file from samtools into a BAM file. Check http://samtools.sourceforge.net for the SAM format specification and the tools for post-processing the alignment.

1.5.6 Test Data Output

The generated bowtie2 and bwa alignment BAM file can be downloaded from the link below for each test data:

- MiSeq GM12878 in-situ files:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bowtie2_align_Miseq/
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bwa_align_Miseq/
- GM06990 Cell line:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bowtie2_align/
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bwa_align/

1.5.7 Running

- Access the function from the menu toolbar: 1D-Functions/Map the raw FASTQ files
- Generate a script called *Mapper_script_bowtie2.sh or Mapper_script_bwa.sh*
- Open a Unix Terminal
- **Execute** *Mapper_script_bowtie2.sh or Mapper_script_bwa.sh* in a Unix Terminal
- Note to Bowtie2 and Cygwin/MinGW Users: To use Bowtie2 in Cygwin/MinGW, the absolute path to the input file generated from GenomeFlow might produce, "Warning: *Could not open read file"* for some users. Use a relative path to the input file to locate the file by editing the generated GenomeFlow script.

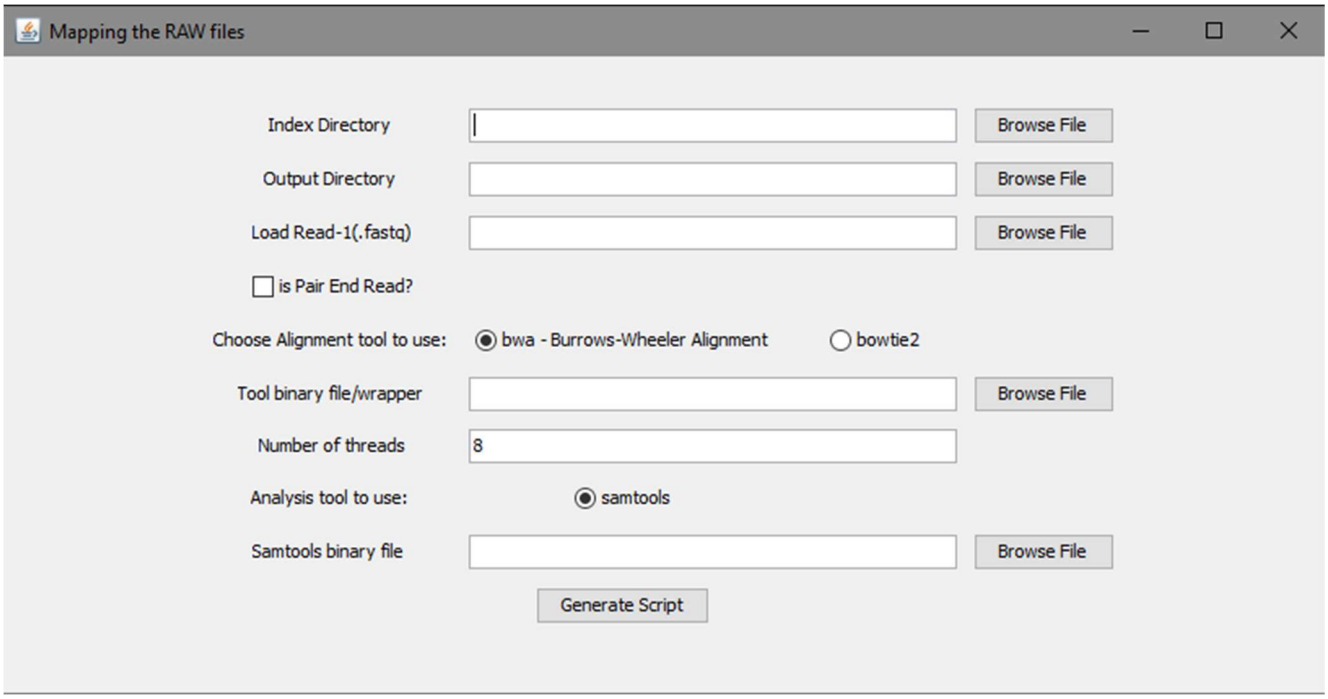

Figure 2: Mapping the raw FASTQ files

Field	Description	Default
Index Directory	The file path to the bwa or bowtie2 directory created when you run the Indexer_Script.sh.	NA
Output Directory	The output directory path to output the script	NA
Load Read-1(.fastq)	The file containing mate 1, or file for a single read e.g HIC003_S2_L001_**R1**_001.fastq	NA
Load Read-2(.fastq)	The file containing mate 2 e.g HIC003_S2_L001_**R2**_001.fastq	NA
Is Pair-End Read	Check if the data is a pair end read data	unchecked
Choose tool to use	Two options are made available for indexing. Select bwa-Burrows-Wheeler alignment or Bowtie2.	bwa

	Important: Only select the tool which was used to generate the reference genome Index files. *bwa can only be used to map generated bwa index files, and bowtie2 can only be used to map generated bowtie2 index files*.	
Binary file	Browse and select the binary file for the chosen tool BWA: Select the **bwa** binary you compiled from bwa-* directory Bowtie2: Select the **bowtie2** binary file to align from the bowtie2-* directory	NA
Number of threads	This option is available only for the **bowtie2-biuild** indexer. Specify the number of threads to use for this task. More threads means less processing time taken.	8
Samtools binary file	SAMtools is a collection of tools for manipulating and analyzing SAM and BAM alignment files. Using these tools together allows you to get from alignments in SAM format Browse and select the **samtools** binary file from the samtools-* directory.	NA
Generate Scripts	This button generates a shell script (.sh) that can be executed in a UNIX terminal by the user.	

1.6 Filter a BAM alignment file

1.6.1 Purpose

To perform filtering of a BAM file to remove low quality map reads and unmapped reads among others.

1.6.2 Input

The BAM file generated from the mapping step above. For example, by default bwa BAM files is named bwa_mapped.bam and bowtie2 BAM files is named bowtie2_mapped.bam.

1.6.3 Test Data

The generated bowtie2 alignment BAM file can be downloaded from the link below for each test data:

- MiSeq GM12878 in-situ files:
    - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bowtie2_align_Miseq/
    - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bwa_align_Miseq/

- GM06990 Cell line:
    - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bowtie2_align/
    - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bwa_align/

1.6.4 Output

It generates a shell script with name *Filter_script_samtools.sh*.

1.6.5 Output of the script:

A BAM binary format (. bam) named *bowtie2_mapped.filtered.bam* for bowtie2 and *bwa_mapped.filtered.bam* for bwa.

1.6.6 Test Data Output:

The generated filtered bowtie2 and bwa alignment BAM file can be downloaded from the link below for the MiSeq GM12878 in-situ and GM06990 Cell line test datasets.

- MiSeq GM12878 in-situ files:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bowtie2_Miseq_mapped.filtered.bam
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bwa_Miseq_mapped.filtered.bam
- GM06990 Cell line:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bowtie2_mapped.filtered.bam
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bwa_mapped.filtered.bam

1.6.7 Running

- Access the function from the menu toolbar: 1D-Functions/Filter a BAM alignment file
- Generate a script called *Filter_script_samtools.sh*
- Open a Unix Terminal
- Run *Filter_script_samtools.sh* in a Unix Terminal

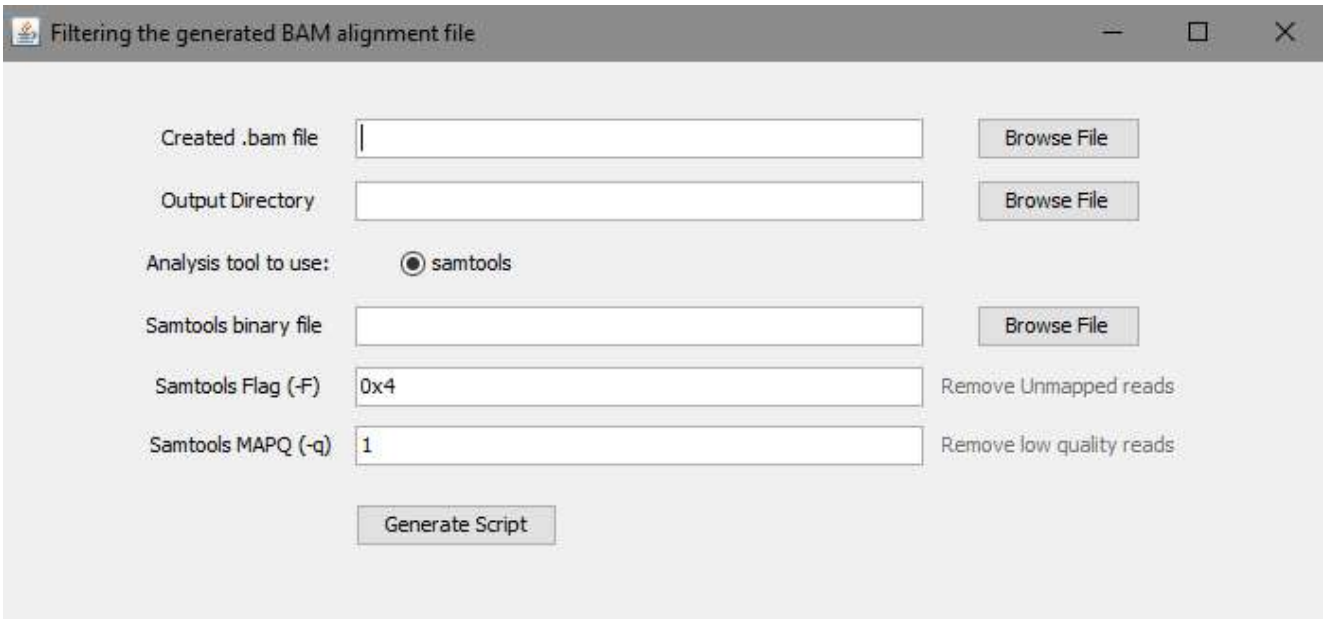

Figure 3: Filter a BAM alignment file

Field	Description	Default
Created .bam file	Select the BAM file generated using either BWA or bowtie2. Select the BAM file named bwa_mapped.bam for **bwa** and BAM file named bowtie2_mapped.bam for **bowtie2**	NA
Output Directory	The output directory path to output the script	NA
Samtools binary file	**samtools** is a collection of tools for manipulating and analyzing SAM and BAM alignment files. Using these tools together allows you to get from alignments in SAM format Browse and select the **samtools** binary file from the samtools-* directory.	NA
Samtools Flag (-F)	**samtools** allows you to sort based on certain flags that are specified on page 5 on the SAM format specification	0x4
Samtools MAPQ (-q)	An integer value to Skip alignments with MAPQ smaller than *INT*. The lowest score is a mapping quality of zero, or **mq0** for short. The reads map to multiple places on the	1

	genome, and we can't be sure of where the reads originated. To improve the quality of our data, we can remove these low quality reads. Generally, we select reads with MAPQ > 1.	
Generate Scripts	This button generates a shell script (.sh) that can be executed in a UNIX terminal by the user. This script contains the basic parameters required by each tool for filtering.	

1.7 Convert a BAM file to a HiC input file format

1.7.1 Purpose

To generate a HiC input file format in medium file format – a text file describing mapped Hi-C reads that can be used as input to create a ***.hic*** file. A hic format file is a binary file containing contact matrices at different resolutions and normalized by different methods.

1.7.2 Input

A filtered BAM alignment file. e.g bwa_mapped.filtered.bam.

1.7.3 Test Data

The generated filtered bowtie2 and bwa alignment BAM file for the MiSeq GM12878 in-situ and GM06990 Cell line test datasets can be downloaded from the link below:

- MiSeq GM12878 in-situ files:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bowtie2_Miseq_mapped.filtered.bam
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/bwa_Miseq_mapped.filtered.bam

- GM06990 Cell line:
    - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bowtie2_mapped.filtered.bam
    - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/bwa_mapped.filtered.bam

1.7.4 Output

It generates a shell script with name *Format_script_samtools.sh*.

1.7.5 Output of the script:

A medium input file format with 11 columns that can be used to create a *.hic* file This file format is explained in details here: HiC Input Medium File Format.

1.7.6 Test Data Output

The generated input medium file format file for the input test datasets can be downloaded from the link below:

- MiSeq GM12878 in-situ files:
    - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/GenomeFlow_Miseq_formatted.bowtie2.input
    - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/GenomeFlow_Miseq_formatted.bwa.input

- GM06990 Cell line:
    - Bowtie2:

http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GenomeFlow_formatted.bowtie2.input

- Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GenomeFlow_formatted.bwa.input

1.7.7 Running

- Access the function from the menu toolbar: 1D-Functions/Convert a BAM file to a HiC input file format
- Generate a script called *Format_script_samtools.sh*
- Open a Unix Terminal
- Execute *Format_script_samtools.sh* in a Unix Terminal

Figure 4: Convert to HiC Input File Format

Field	Description	Default
Created .bam file	Select the BAM file generated from the filtering. By default, **bwa** filtered BAM file is named bwa_mapped.filtered.bam and **bowtie2** filtered BAM file is named bowtie2_mapped.filtered.bam	NA
Output Directory	The output directory path to output the script	NA

Samtools binary file	**samtools** is a collection of tools for manipulating and analyzing SAM and BAM alignment files. Using these tools together allows you to get from alignments in SAM format Browse and select the **samtools** binary file from the samtools-* directory	NA
Generate Scripts	This button generates a shell script (.sh) that can be executed in a UNIX terminal by the user.	

1.8 HiC-Express

1.8.1 Purpose

To generate a HiC input file format in generate a medium file format - a text file describing mapped Hi-C reads that can be used as input to create a ***.hic*** file from a raw fastq files derived from a Hi-C experiment.. A hic format file is a binary file containing contact matrices at different resolutions and normalized by different methods.

1.8.2 Input

A FASTQ read files usually with extension, .fq or .fastq. Read more about FASTQ files here: https://en.wikipedia.org/wiki/FASTQ_format

1.8.3 Test Data

Test datasets can be found here:

- MiSeq GM12878 in-situ files: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/MiSeq_GM12878_Data/
- A karyotypically normal human lymphoblastoid cell line (GM06990) from Aiden et al: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GM06990_Data/

1.8.4 Output

It generates a shell script with name *HiC-Express.sh*

1.8.5 Output of the script:

An input Medium file format with 11 columns that can be used to create a *.hic* file This file format is explained in details here: HiC Input Medium File Format

1.8.6 Test Data Output

The generated input medium file format file for the input test datasets can be downloaded from the link below:

- MiSeq GM12878 in-situ files:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/GenomeFlow_Miseq_formatted.bowtie2.input
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/MiSeq_GM12878/GenomeFlow_Miseq_formatted.bwa.input
- GM06990 Cell line:
  - Bowtie2: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GenomeFlow_formatted.bowtie2.input
  - Bwa: http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/GenomeFlow_formatted.bwa.input

1.8.7 Running

- Access the function from the menu toolbar: 1D-Functions/HiC-Express
- Generate a script called *HiC-Express.sh*
- Open a Unix Terminal
- **Execute** *HiC-Express.sh* in a Unix Terminal

- Note to Bowtie2 and Cygwin/MinGW Users: To use Bowtie2 in Cygwin/MinGW, the absolute path to the input file generated from GenomeFlow might produce a Warning: *Could not open read file* for some users. Use a relative path to the input file to locate the file by editing the generated GenomeFlow script.

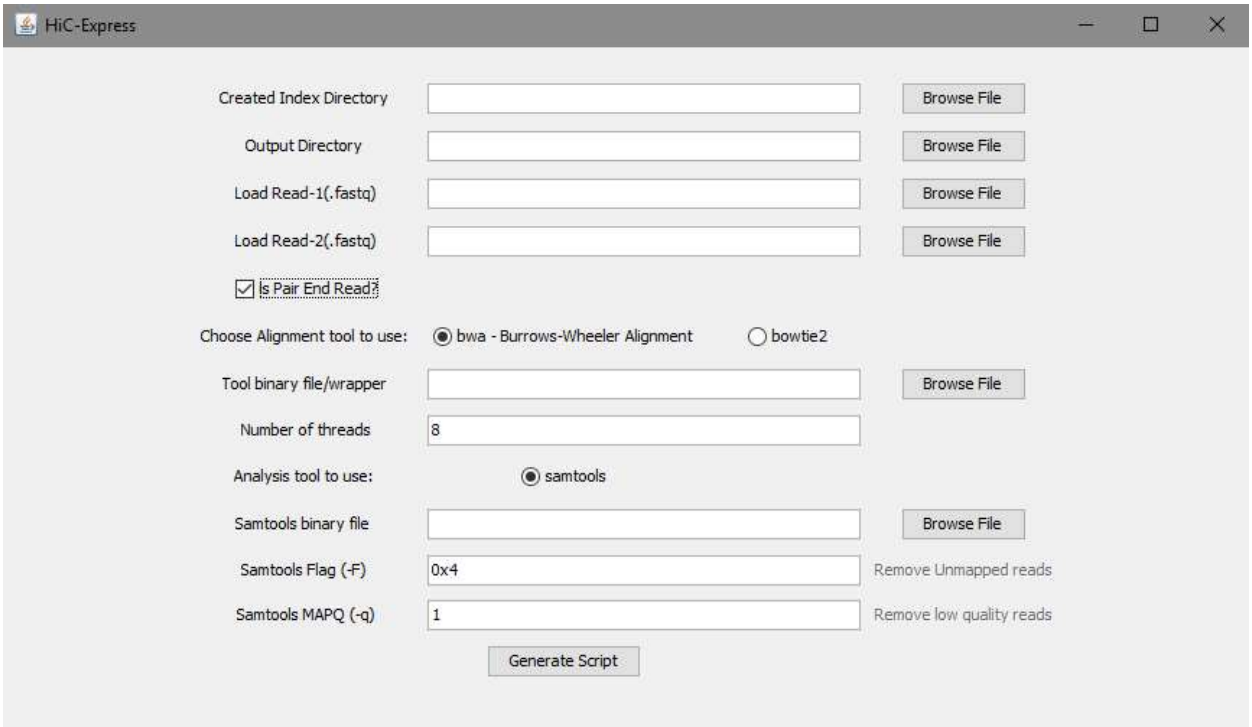

Figure 5: HiC-Express

Field	Description	Default
Created Index Directory	A path to the index created using bwa or bowtie2	NA
Output Directory	The output directory path to output the script	NA
Load Read-1(.fastq)	The file containing mate 1, or file for a single read e.g HIC003_S2_L001_R1_001.fastq	NA
Load Read-2(.fastq)	The file containing mate 2 e.g HIC003_S2_L001_R2_001.fastq	NA

Is Pair-End Read	Check if the data is a pair end read data	unchecked
Choose tool to use	Two options are made available for indexing. Select bwa-Burrows-Wheeler alignment or Bowtie2. **Important**: Only select the tool which was used to create the reference genome Index. *bwa can only be used to map bwa index, and bowtie2 can only be used to map bowtie2 index*.	bwa
Binary file	Browse and select the binary file for the chosen tool bwa: Select the **bwa** binary you compiled from bwa-* directory Bowtie2: Select the **bowtie2** binary file to align from the bowtie2-* directory	NA
Number of threads	This option is available only for the **bowtie2-biuild** indexer, Specify the number of threads to use for this task. More threads means less processing time taken.	8
Samtools binary file	**samtools** is a collection of tools for manipulating and analyzing SAM and BAM alignment files. Using these tools together allows you to get from alignments in SAM format Browse and select the **samtools** binary file from the samtools-* directory	NA
Samtools Flag (-F)	**samtools** allows you to sort based on certain flags that are specified on page 5 on the SAM format specification	0x4
Samtools MAPQ (-q)	An integer value to skip alignments with MAPQ smaller than *INT*. The lowest score is a mapping quality of zero, or **mq0** for short. The reads map to multiple places on the genome, and we can't be sure of where the reads originated. To improve the quality of our data, we can remove these low-quality reads. Generally, we select reads with MAPQ > 1.	1

Generate Scripts	This button generates a shell script (.sh) that can be executed in a UNIX terminal by the user.	

2 2D-Functions

2.1 Convert mapped Hi-C reads to hic format file

2.1.1 Purpose

To create a binary hic format file containing contact matrices at different resolutions and normalized by different methods from a text file describing mapped Hi-C reads.

2.1.2 Input file format

Five formats are acceptable: short format, short format with score, medium format, long format and 4DN DCIC format. A sample file is: *executable/sample_data/GSM1551688_HIC143_merged_nodups.zip* (unzip it before use). Another set of test data is the GM06990 cell line data. This can be downloaded from the link below:

- GM06990 Cell line:
    - Bowtie2:
        - http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/
        - Download/Save the GenomeFlow_formatted.bowtie2.input file
    - Bwa
        - http://sysbio.rnet.missouri.edu/bdm_download/GenomeFlow/GM06990/
        - Download/Save the GenomeFlow_formatted.bwa.input file

2.1.2.1 Short format

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment

If not using the restriction site file option, *frag* will be ignored, but please see above note on dummy values. *readname* and *strand* are also not currently stored within *.hic* files.

2.1.2.2 Short with score format

This format is useful for reading in already processed files, e.g. those that have been already binned and/or normalized. This format can be easily used in conjunction with the -r flag to create a *.hic* file that contains a single resolution.

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <score>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- score = the score imputed to this read

If not using the restriction site file option, *frag* will be ignored, but please see above note on dummy values. *readname* and *strand* are also not currently stored within *.hic* files.

2.1.2.3 Medium format

A whitespace separated file that contains, on each line

```
<readname> <str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <mapq1>
<mapq2>
```

- str = strand (0 for forward, anything else for reverse)

- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- mapq = mapping quality score

If not using the restriction site file option, *frag* will be ignored, but please see above note on *dummy values*. If not using mapping quality filter, *mapq* will be ignored. *readname* and *strand* are also not currently stored within .hic files.

2.1.2.4 Long format

The long format is used by Juicer and takes in directly the *merged_nodups.txt* file.

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <mapq1> <cigar1>
<sequence1> <mapq2> <cigar2> <sequence2> <readname1> <readname2>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- mapq = mapping quality score
- cigar = cigar string as reported by aligner
- sequence = DNA sequence

If not using the restriction site file option, frag will be ignored, but please see above note on dummy values. If not using mapping quality filter, mapq will be ignored. readname, strand, cigar, and sequence are also not currently stored within *.hic* files.

2.1.2.5 4DN DCIC format

A file that follows the 4DN DCIC format specification (the 4DN DCIC format specification). See the link for more information. Briefly, there should be a header with the first seven columns reserved:

```
## pairs format v1.0

#columns: readID chr1 position1 chr2 position2 strand1 strand2
```

If the columns line contains (in any field after field 7) both *frag1* and *frag2*, those will also be read in; otherwise they will be set as *frag1*=0 and *frag2*=1 by default, so that no reads are discarded. Other fields are ignored.

2.1.3 Output

A binary .hic file containing contact matrices

2.1.4 Running

Access the function from the menu toolbar: 2D-Functions/Convert to HiC

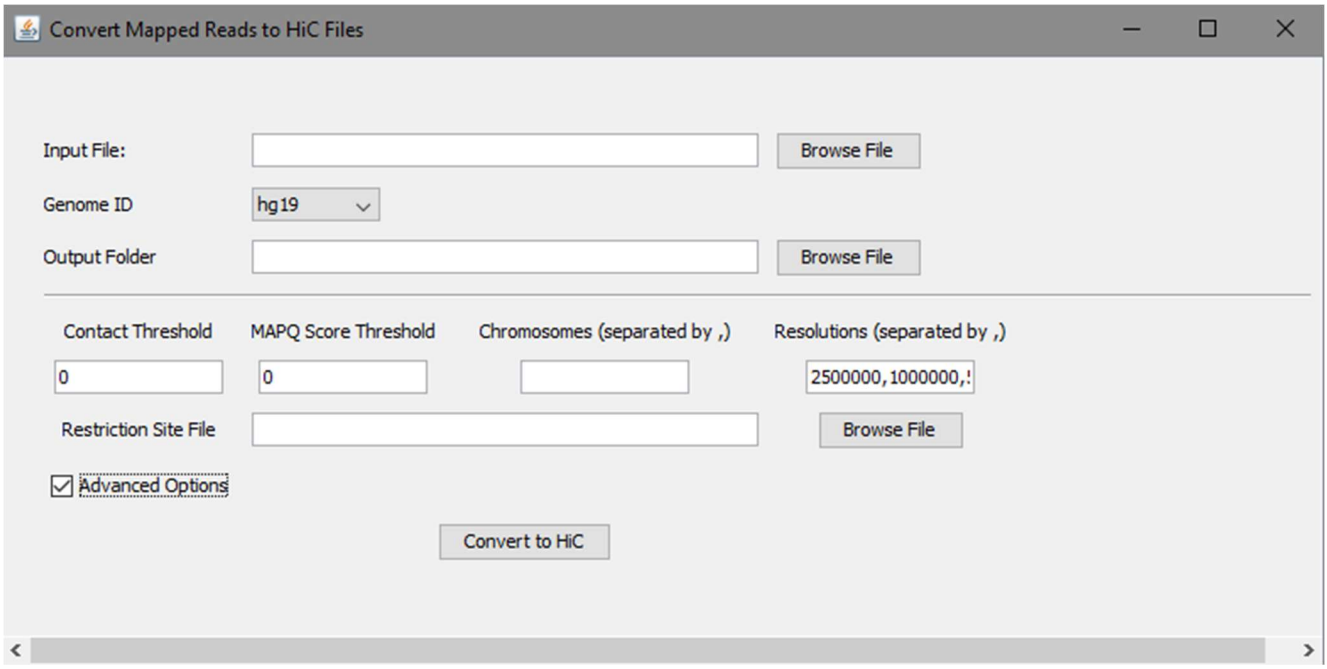

Figure 6: Convert to HiC function

Field	Description	Default
Input file	A text file that contains the mapped Hi-C reads (format described above)	NA
Genome ID	Version genome of Hi-C data	hg19
Output Directory	The output directory path to output the generated hic format file. An example output filename is **GenomeFlow_Convert_1521343280452.hic**	NA
Contact Threshold	Number of interaction threshold for contacts to be used in creating contact matrices.	0
MAPQ Score Threshold	Mapping quality score threshold for reads to be considered in creating contact matrices.	0
Chromosomes	Chromosomes for which their contact matrices need to be created. When left blank, all chromosomes will be considered. Chromosomes must be separated by a comma (,).	All (when left blank)
Resolutions	List of resolutions of contact matrices to be created. Resolutions are separated by a comma (,)	2500000, 1000000, 500000,

		250000, 100000, 50000, 25000,10000,5000
Restriction Site File	Each line starts with a chromosome number followed by positions of restriction sites on that chromosome, in numeric order, and ending with the size of the chromosome. When provided, 8 additional fragment-delimited resolutions are added: 500f, 250f, 100f, 50f, 20f, 5f, 2f, 1f	blank

2.2 Extract contact matrices from a hic format

2.2.1 Purpose

To extract a contact matrix from a hic format into a sparse matrix format in a text file

2.2.2 Input

A local path to a hic format or an online link to a hic format. A link to a hic file: https://www.encodeproject.org/files/ENCFF219YOB/@@download/ENCFF219YOB.hic

2.2.3 Output

A contact matrix in sparse matrix format (each line represents a contact by three numbers separated by whitespaces: <position1> <postion2> <interaction_frequency>

2.2.4 Running

Access the function from the menu toolbar: 2D-Functions/Extract HiC

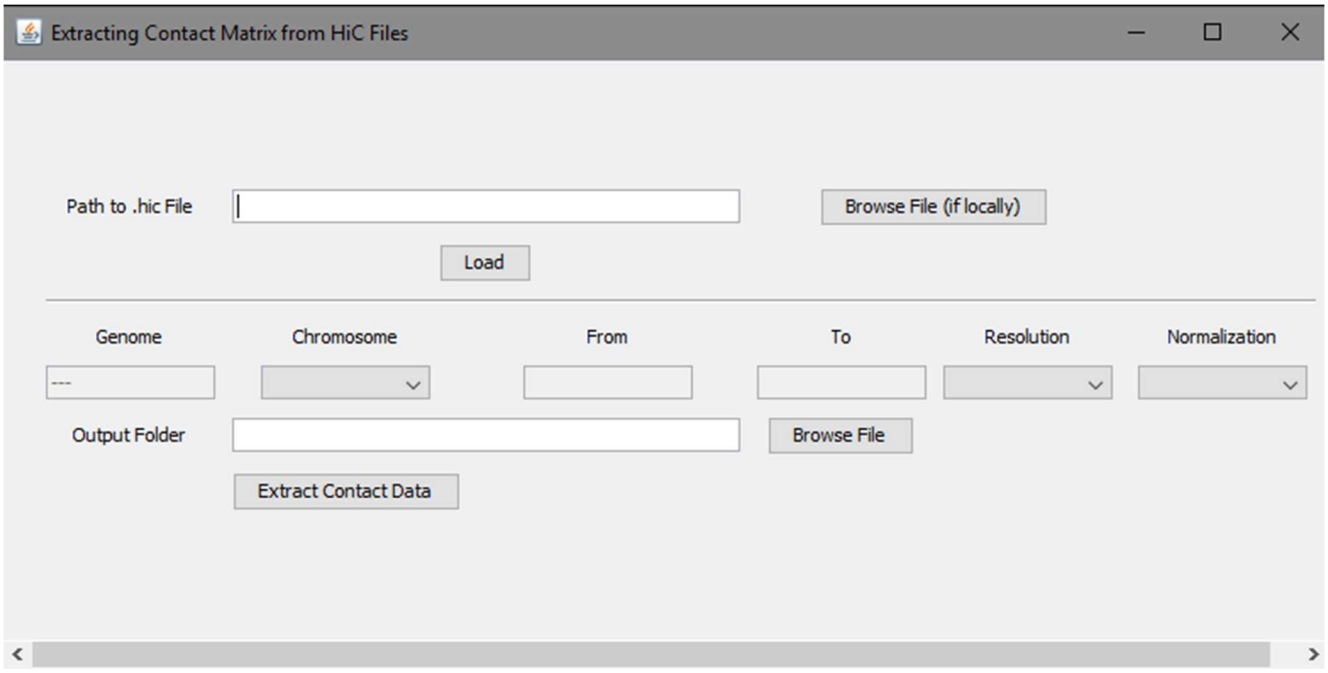

Figure 7: Extract Contact Matrices from a hic file

Field / Button	Description	Default
Path to .hic File	An online link or local path to a hic format file	NA
Load	Click this button to fetch information from the header of the hic file.	NA
Genome	Genome version of the hic file	NA
Chromosomes	List of resolutions of contact matrices in the hic file	NA
From	Start of a fragment (to extract its contact matrix). When From and To are left blank, the whole chromosome is considered.	Blank
To	End of a fragment (to extract its contact matrix). When From and To are left blank, the whole chromosome is considered.	Blank
Resolution	List of resolutions of contact matrices in the hic file	NA
Normalization	List of normalization methods used to normalize contact matrices	NA

Output Directory	The output directory path to output the extracted data. An example filename for the generated file is *GenomeFlow_Extract_1521577159643.txt*	
Extract Contact Data	Click this button to initiate extracting contact data	NA

2.3 Normalize HiC contact matrices

2.3.1 Purpose

To normalize contact matrices in sparse matrix format.

2.3.2 Input

A contact matrix in sparse matrix format (each line represents a contact by three numbers separated by whitespaces: <position1> <position2> <interaction_frequency>.

2.3.3 Output

A normalized contact matrix in sparse matrix format. The matrix is normalized by the Iterative Correction and Eigenvector decomposition (ICE) method.

2.3.4 Running

Access the function from the menu toolbar: 2D-Functions/Normalized HiC Data.

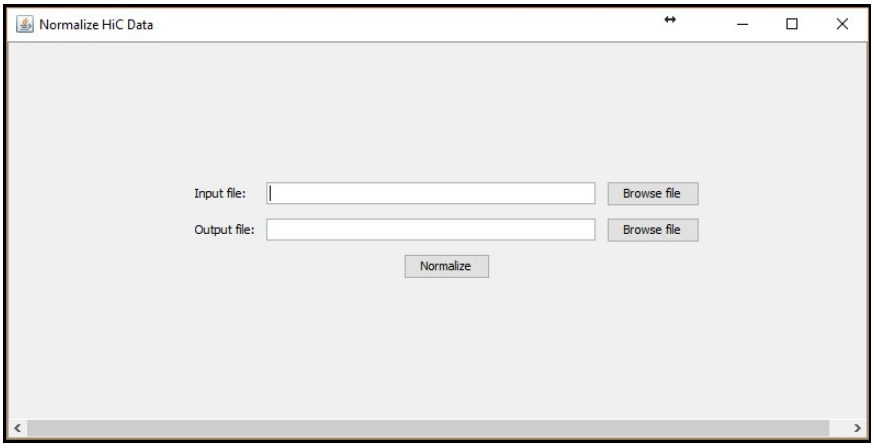

Figure 8: Normalize HiC contact matrices

2.4 Visualizing Dataset in 2D format

2.4.1 Purpose

To create a two dimensional (2D) graphical representation of a contact matrix from an input file.

2.4.2 Input

A sparse matrix format (each line represents a contact by three numbers separated by whitespaces: <position1> <postion2> <interaction_frequency> or an input file in square matrix format (a full matrix representing all the contact regions). Mark the *Is Square Matrix?* box if the input is a square matrix.

An example sparse matrix file can be found here:

/executable/sample_data/ contact_matrices/ chr11_10kb_gm12878_list_125mb_135mb.txt

Examples of square matrix files can be found here:

/executable/sample_data/ contact_matrices/square_matrices/

Note: Resolution for square matrices = 40000

2.4.3 Output

A Heatmap which is a graphical representation of contact data where numeric values in the input contact matrix are represented as colors based according to a selected color gradient.

2.4.4 Running

Access the function from the menu toolbar: 2D-Functions/Visualize Dataset.

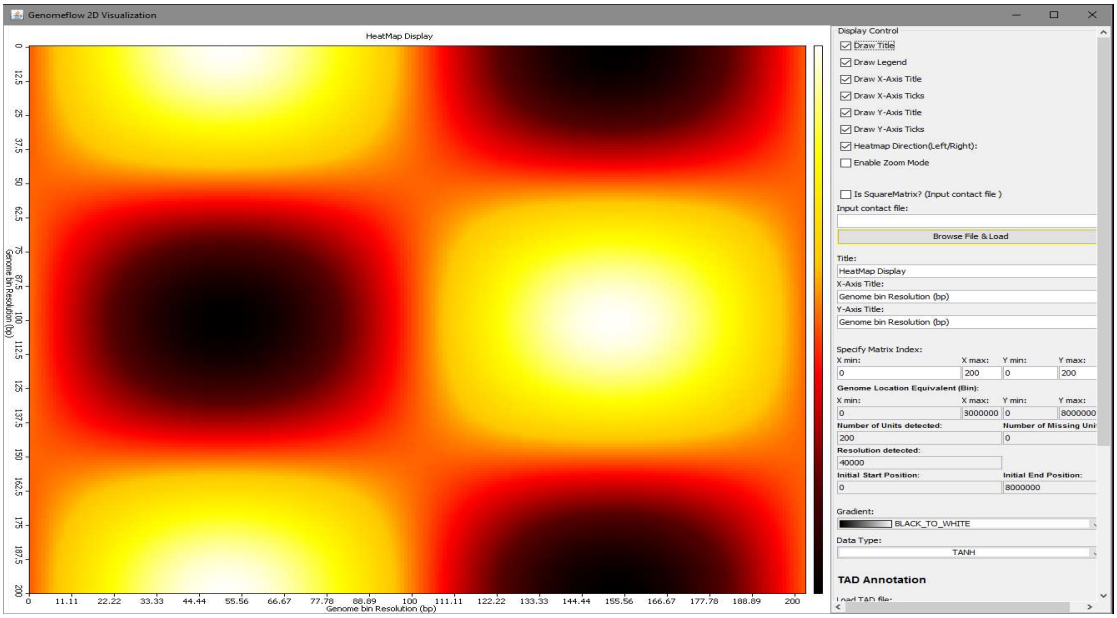

Figure 9: Visualize Dataset in 2D Format

2.4.5 Display Controls

The description of the display controls on the display window is given below.

Field	Description	Default
Draw Title	Shows or hides the Heatmap title	checked
Draw Legend	Shows or hides the color legend	checked
Draw X-Axis Title	Shows or hides the X-axis title label on the 2D display window	checked
Draw X-Axis Ticks	Shows or hides the X-axis ticks label on the 2D display window	checked
Draw Y-Axis Title	Shows or hides the Y-axis title label on the 2D display window	checked
Draw Y-Axis Ticks	Shows or hides the Y-axis ticks label on the 2D display window	checked

Heatmap Direction(Left/Right)	Changes the Y-axis origin of the heatmap matrix from the Bottom-Left to Top-Left and vice versa	checked
Enable Zoom Mode	Allows the user to zoom in/out of the heat map matrix	unchecked
Is Square Matrix?(Input contact file)	Allows the user to specify if the input is a Square matrix (a full matrix) or a sparse matrix. If checked, it displays a textbox for the user to specify the matrix resolution.	unchecked
Specify Resolution	Visible only if *Is SquareMatrix?* is checked. It allows user specify resolution for the input matrix.	NA
Input contact file	A text file containing a contact matrix in any of the format described above.	NA
Title	Allows user to specify the title of the heatmap	Heatmap Display
X-Axis Title	Allows user to specify the X-Axis title for the heatmap	Genome bin Resolution (bp)
Y-Axis Title	Allows user to specify the Y-Axis title for the heatmap	Genome bin Resolution (bp)
X min	Allows the user to specify the minimum X-axis Tick for the heatmap	0
X max	Allows the user to specify the maximum X-axis Tick for the heatmap	200
Y min	Allows the user to specify the minimum Y-axis Tick for the heatmap	0
Y max	Allows the user to specify the maximum Y-axis Tick for the heatmap	200
X min [Genome Location Equivalent]	Shows the genomic position equivalent for the minimum X-axis tick for the heatmap	0

X max[Genome Location Equivalent]	Shows the genomic position equivalent for the maximum X-axis tick for the heatmap	8000000
Y min [Genome Location Equivalent]	Shows the genomic position equivalent for the minimum Y-axis tick for the heatmap	0
Y [Genome Location Equivalent]	Shows the genomic position equivalent for the maximum Y-axis tick for the heatmap	8000000
Number of Units detected	Shows the number of regions found in the input matrix	200
Number of Missing Units	Shows the number of gaps or missing regions noted from the input matrix	0
Resolution detected	Displays the resolution of the input matrix	40000
Initial Start Position	Shows the minimum genome position observed from the input matrix	0
Initial End Position	Shows the maximum genome position observed from the input matrix	8000000
Gradient	An array of color used as a gradient. One color is used as the bottom gradient and another color is used as the top gradient. Hence, it produces a gradient from one color to the other. The Gradient Colors are explained below	HOT
GRADIENT_BLACK_TO_WHITE	Produces a gradient from black (low) to white (high)	
GRADIENT_BLUE_TO_RED	Produces a gradient from blue (low) to red (high)	
GRADIENT_HEAT	Produces a gradient using the colors black, brown, orange, white	
GRADIENT_HOT	Produces a gradient using the colors black, red, orange, and yellow to white	

GRADIENT_MAROON_TO_GOLD	Produces a gradient from maroon (low) to gold (high)	
GRADIENT_RAINBOW	Produces a gradient with the colors violet, blue, green, yellow, orange, and red.	
GRADIENT_RED_TO_GREEN	Produces a gradient from red (low) to green (high)	
GRADIENT_ROY	Produces a gradient through red, orange, yellow	
Data Type	It determines the type of data to be displayed. The types available are the raw input data, a TANH of input data, a Pearson correlation of input data, and a Spearman correlation of the input data.	TANH

2.4.6 TAD Annotation

The description of the display controls on the display window for TAD annotation is given below.

Field	Description	Default
Load TAD file	Browse and Load a .bed format file containing the TADs identified for the input matrix	NA
Identified TAD	It shows the TADs in the input file	NA
Show TAD on Heatmap	It marks the boundary of the TADs identified on the displayed heatmap	
Display Multiple TADs	Once checked, allows TADs from different method to be overlapped on the same display window. This function is useful for comparing TADs identified by different methods for a dataset.	unchecked

Choose Display Color	Choose the color for the TAD boundary marks	Color 1

2.4.7 Demonstration

Figure 5 below shows the TAD annotation for the TADs identified by two TAD identification algorithms (ClusterTAD and DI) for mESC Chromosome 17 from Ren Lab.

Step 1:

To run this demonstration, load a sample square matrix as the input contact file.

The example file can be found here: */executable/sample_data/ contact_matrices/square_matrices/mESC_nij.chr17*. Resolution for the square matrix = 40000

Load the contact file as instructed here: Visualizing Dataset in 2D format

Step 2:

Modify the highlighted fields on the display window. The table below shows the values set for each field in the display control.

Field	Value
Draw Title	checked
Draw Legend	checked
Draw X-Axis Title	checked
Draw X-Axis Ticks	checked
Draw Y-Axis Title	checked
Draw Y-Axis Ticks	checked
Heatmap Direction(Left/Right)	checked
Enable Zoom Mode	unchecked
Is SquareMatrix?(Input contact file)	checked
Specify Resolution	40000
Input contact file	Path/to/chr17/inputfile

Title	HeatMap Display
X-Axis Title	Number of Bins
Y-Axis Title	Number of Bins
X min	500
X max	700
Y min	500
Y max	700
X min [Genome Location Equivalent]	20000000
X max[Genome Location Equivalent]	28000000
Y min [Genome Location Equivalent]	20000000
Y [Genome Location Equivalent]	28000000
Number of Units detected	2382
Number of Missing Units	0
Resolution detected	40000
Initial Start Position	0
Initial End Position	95240000
Gradient	HOT
Data Type	TANH

Step 3:

a. *Browse & Load* the ClusterTAD file found here:
ClusterTAD: /executable/sample_data/TAD_annotation/mESC_TAD_bed/ ClusterTAD /chr17.bed.

b. Select a Unique from Color 1 to 4. (Ex: Color 1 for ClusterTAD and Color 2 for DI)

c. Click the *Show TAD on Heatmap* button.

Step 4:

To display multiple TADs on the Heatmap, Mark/Check the *Display Multiple TADs* then Repeat Step 3 with the DI file found here:

DI: */executable/sample_data/TAD_annotation/mESC_TAD_bed/ DI /chr17.bed*

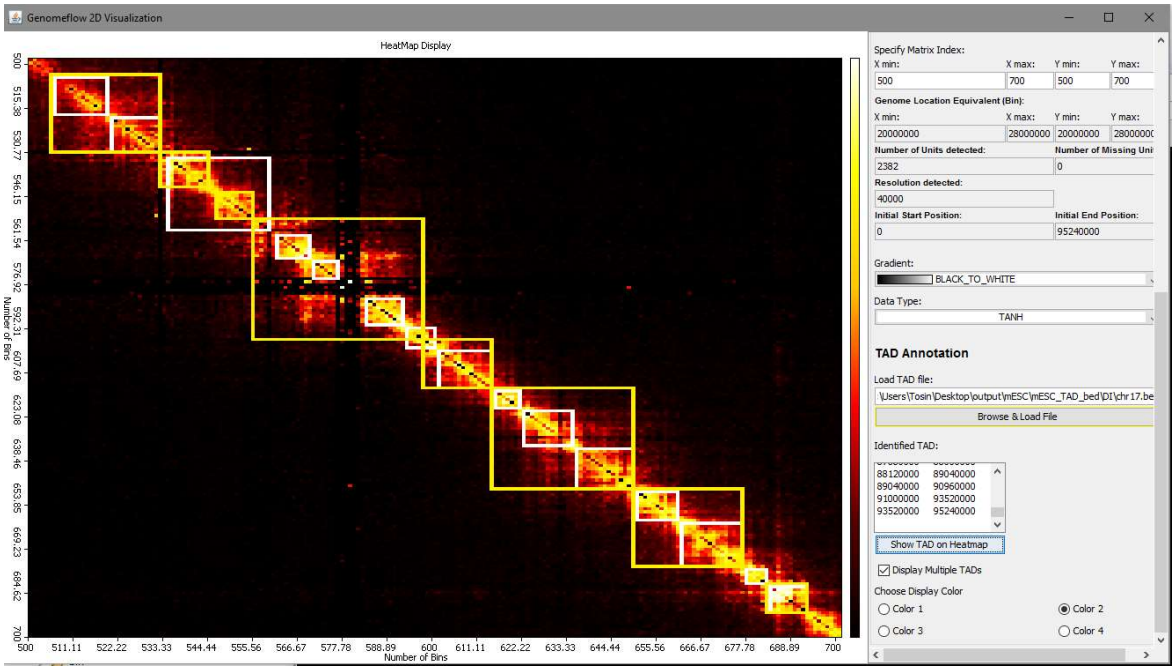

Figure 10: Demonstration of TAD Annotation on 2D Heatmap

2.5 Identify TAD

2.5.1 Purpose

To identify Topological Associated domains from input contact matrix.

2.5.2 Input

An input file in square matrix format (a full matrix representing all the contact regions) or a sparse matrix format (each line represents a contact by three numbers separated by whitespaces: <position1> <postion2> <interaction_frequency>.

An example sparse matrix file can be found here:

/executable/sample_data/ contact_matrices/ chr11_10kb_gm12878_list_125mb_135mb.txt

2.5.3 Output

A TAD with the best quality will be generated prefixed with ***BestTAD_*** in *bed* format. This file will be found here: */Selected_output_directory_from_GUI/Output/TADs/.*

2.5.4 Running

Access the function from the menu toolbar: 2D-Functions/Identify TAD.

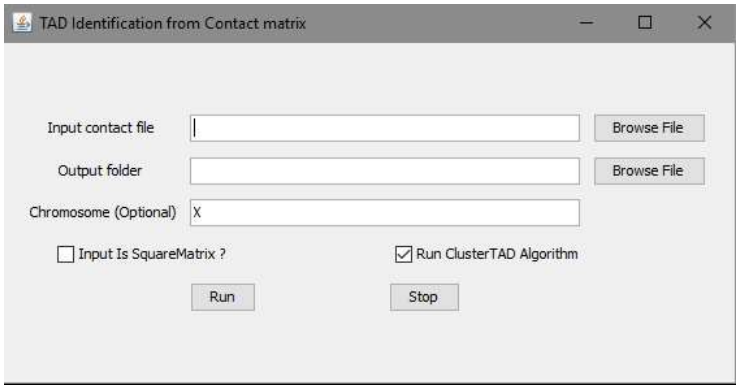

Figure 11: Identifying TADs on a contact matrix

Field	Description	Default
Input contact file	An input file in any of the format described above	NA
Output folder	Directory to output the comparison report	NA
Is SquareMatrix?(Input contact file)	Allows the user to specify if the input is a Square matrix (a full matrix) or a sparse matrix. If checked, it displays a textbox for the user to specify the matrix resolution.	unchecked
Data Resolution	It is visible only if *Is SquareMatrix?* is checked. It allows user specify resolution for the input matrix.	40000
Chromosome (optional)	Allows user to specify the chromosome data	X

Run ClusterTAD Algorithm	The default algorithm used for TAD identification from the input contact Matrix	checked
Run	To start the identification process. A progress bar is displayed to show the steps taken by the TAD identification algorithm.	NA
Stop	During the identification, if this button is pressed, the program will stop.	NA

2.6 Check TAD consistency between two TADs from different methods

2.6.1 Purpose

To compare two TADs from two different Topological Associated domains identification method.

2.6.2 Input

A file containing TADs in .bed format. The method whose TADs consistency is to be checked is termed Method-1, and the methods whose TADs is to be compared with is termed Method-2. Choose the same chromosome for different methods. For example, to compare TAD from ClusterTAD with DI for chromosome 17,
Method-1: /*executable/sample_data/TAD_annotation/mESC_TAD_bed/ ClusterTAD* /chr17.bed.
Method-2: */executable/sample_data/TAD_annotation/mESC_TAD_bed/ DI /chr17.bed*

2.6.3 Output

A report of the consistency of the Method-1 with Method-2. The output reports the following cases:

Case	**Description**

Case 1	The number of Exact TADs found in both Method-1 and Method-2
Case 2	The number of Sub-TADs that exist between Method-1 and Method-2
Case 3	The number of Conflicting TADs.
Case 4	The number of TADs in Method-1 but not found in Method-2

2.6.4 Running

Access the function from the menu toolbar: 2D-Functions/Check TAD Consistency.

Figure 12: Comparing two TADs for a consistency check

Field	Description	Default
Input Method-1 TAD file(.bed)	Browse the .bed format file containing the TADs identified by Method-1	NA
Input Method-2 TAD file(.bed)	Browse the .bed format file containing the TADs identified by Method-2	NA
Data Resolution	The Resolution of the dataset the TADs were identified from.	40000

Output folder	Directory to output the comparison report	NA
Create Report	Once this button is pressed, a progress bar is displayed to show the steps taken by the TAD identification algorithm,.	NA
Stop	During the check, if this button is pressed, the program will stop.	NA

3 3D-Functions

3.1 3D model reconstruction by LorDG

3.1.1 Purpose

To build 3D chromosomes and genome models

3.1.2 Input

A contact matrix in sparse matrix format

3.1.3 Output

3D models in. gss format file and .pdb format file

3.1.4 Running

Access the function from the menu toolbar: 3D-Functions/LorDG-3D Modeler

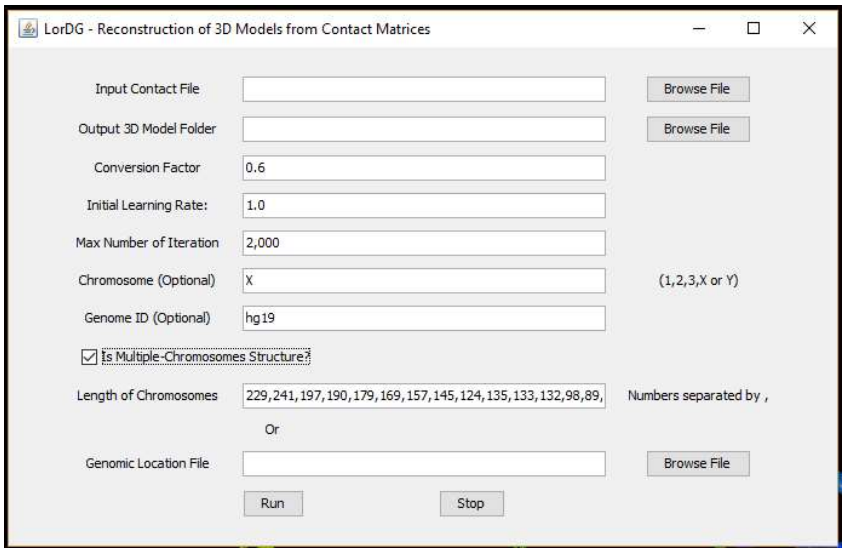

Figure 13: 3D Model reconstruction by LorDG

Field /Button	Description	Default
Conversion Factor	α in the formula $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$, where IF_{ij} is interaction frequency between i and j. When the field is left blank, the program will search for the best value in the range [0.1-3.0] with a step size of 0.1. Users can also specify a range to search for by putting two numbers separated by a hyphen (e.g. 0.5-1.0). During the searching, the right-top corner of the main screen displays information about the current value being tested.	1.0
Initial Learning Rate	Initial learning rate of the optimization. Higher learning rate can speed up the reconstruction process but can cause the process to fail as well	1.0
Max Number of Iterations	Maximum number of iterations for the optimization	1000
Chromosome	Chromosome name of the contact matrix in the input. If the input contains contact matrix of the whole genome, leave this field blank.	X
Genome ID	Genome version of the contact matrix in the input.	hg19

Is Multiple-Chromosomes Structure?	If the input contains both inter-and intra-chromosomal contacts data, this checkbox should be checked.	unchecked
Length of Chromosomes	This field contains a list of lengths of chromosomes in increasing order of chromosome names and separated by commas if "Is Multiple-Chromosomes Structure" is checked. Please note that these lengths should not contain omitted regions (e.g. centromeres) in the input of chromosomes.	
Run	To start the reconstruction process. The main screen displays how new models are being formed from initially random models. The information about the reconstruction is displayed in the top-right corner of the main screen. The conversion factor is being used to build the model and the current value of the objective function (higher is better). After the reconstruction is finished, the score of the model is displayed in the top-right corner of the main screen (the lower the value is, the better the model is).	NA
Stop	During the reconstruction, if this button is pressed the program will stop and output the currently best structure. If the program is searching for the best conversion factor, it will stop the searching and use the best-found conversion factor to build models.	NA

3.2 3D model reconstruction by 3DMax

3.2.1 Purpose

To build 3D chromosomes and genome models.

3.2.2 Input

A contact matrix in sparse matrix format.

3.2.3 Output

3D models in .gss format file and .pdb format file

3.2.4 Running

Access the function from the menu toolbar: 3D-Functions/3DMax-3D Modeler

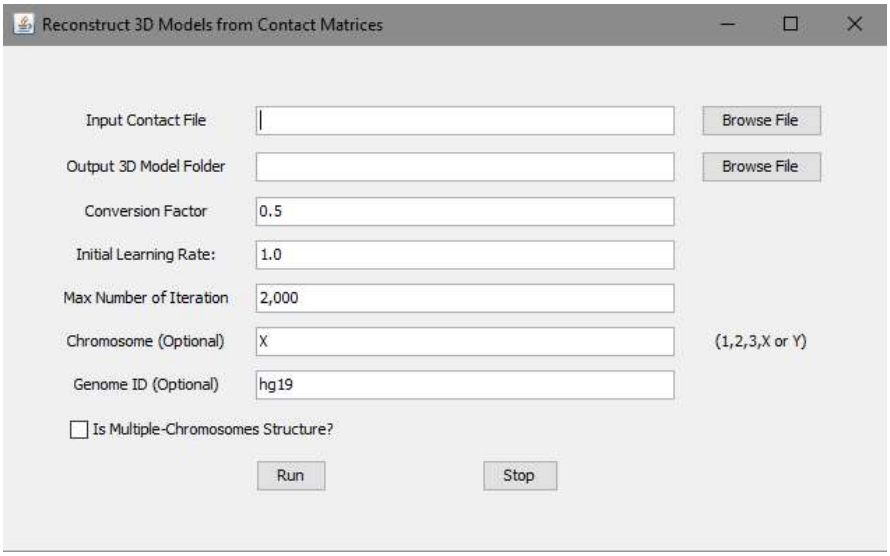

Figure 14: 3D Model reconstruction by 3DMax

Field /Button	Description	Default
Conversion Factor	α in the formula $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$, where IF_{ij} is interaction frequency between i and j. When the field is left blank, the program will search for the best value in the range [0.1-2.0] with a step size of 0.1. Users can also specify a range to search for by putting two numbers separated by a hyphen (e.g. 0.5-1.0). During the search, the right-top	0.5

	corner of the main screen displays information about the current value being tested.	
Initial Learning Rate	Initial learning rate of the optimization. Higher learning rate can speed up the reconstruction process but can cause the process to fail as well	1.0
Max Number of Iterations	Maximum number of iterations for the optimization	2000
Chromosome	Chromosome name of the contact matrix in the input. If the input contains contact matrix of the whole genome, leave this field blank.	X
Genome ID	Genome version of the contact matrix in the input.	hg19
Is Multiple-Chromosomes Structure?	If the input contains both inter-and intra-chromosomal contacts data, this checkbox should be checked.	unchecked
Length of Chromosomes	This field contains a list of lengths of chromosomes in increasing order of chromosome names and separated by commas if "Is Multiple-Chromosomes Structure" is checked. Please note that these lengths should not contain omitted regions (e.g. centromeres) in the input of chromosomes.	
Run	To start the reconstruction process. The main screen displays how new models are being formed from initially random models. The information about the reconstruction is displayed in the top-right corner of the main screen. The conversion factor being used to build the models and the current value of the objective function (higher is better). After the reconstruction is finished, the score of the model is displayed in the top-right corner of the main screen (the lower the value is, the better the model is).	NA

Stop	During the reconstruction, if this button is pressed, the program will stop and output the currently best structure. If the program is searching for the best conversion factor, it will stop the searching and use the best-found conversion factor to build models.	NA

3.3 Chromatin loop identification

3.3.1 Purpose

To identify chromatin loop in 3D models .

3.3.2 Input

A 3D model in .gss format.

3.3.3 Output

A list of chromatin loops in *a .bed* format file (optional) and highlighted in the 3D model.

3.3.4 Running

Access the function from the menu toolbar: 3D-Functions/Loop Detection

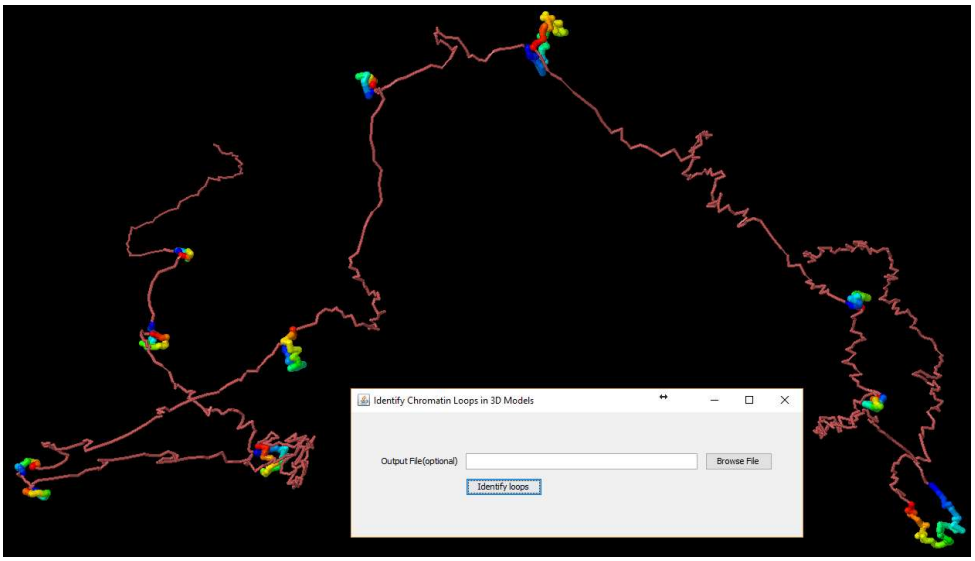

Figure 15: Chromatin loops

The function identifies chromatin loops and highlights them in the 3D model. The loops can also be outputted into a .bed format file specified in the *Output File* field. The top-right corner of the main screen displays the number of chromatin loops identified.

Loops are colored in spectrum (from blue to red). To highlight loops better, color the model by a single color (right-click on the main screen, choose ,)

3.4 Model annotation

3.4.1 Purpose

To annotate 3D models with genomic elements.

3.4.2 Input

A 3D model (e.g. in *executable/sample_data/models*) and genomic elements in *.bed* format files (e.g. in *executable/track_files*).

3.4.3 Output

3D model is annotated with data from *.bed* format files.

3.4.4 Running

Access the function from the menu toolbar: 3D-Functions/Model Annotation

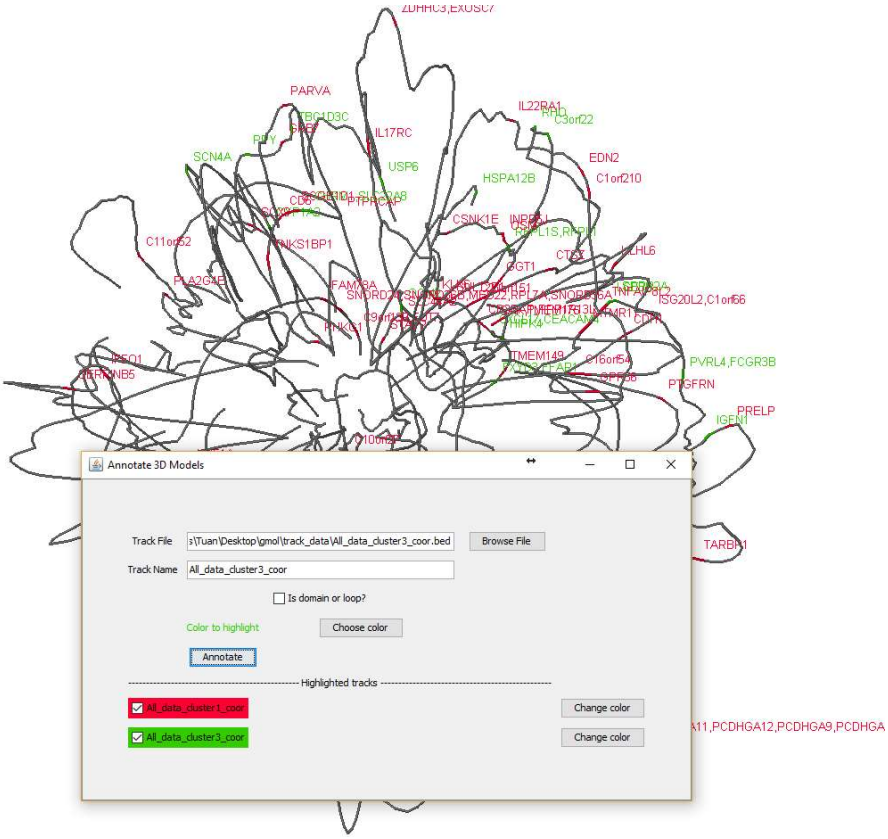

Figure 16: Function to annotate 3D models

To better highlight track data, change the color of the model to a sing color (right-click on the main screen, Color/Structure/Reset). The background can be changed to white to (Color/Background/White)

Field / Button	Description	Default
Track file	A file in bed format (see executable/track_files for example) to annotate the model	NA
Track name	A unique name associated with the above input file	Name of track file
Is domain or loop?	Indicate if the track file contains domains or loops. Adjacent domains/loops will be colored in red/blue alternatively.	Unchecked
Choose color	To pick a color to label annotation and points overlapped by genomic elements in the track file.	Random
Change color	To change color of the corresponding track	NA

	Checking corresponding track names will display or hidden the content of tracks.	

To get the genomic coordinate of a point, left-click or mouse-over to the point as shown in **Figure .**

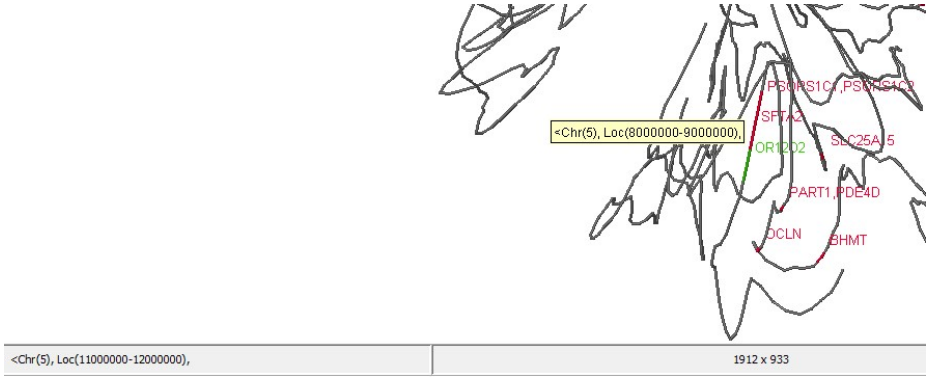

Figure 17: Coordinate of a point in the model

3.5 Gene expression data visualization (a special case of model annotation)

3.5.1 Purpose

To display gene expression level along a 3D model

3.5.2 Input

a) A 3D model in GSS format (e.g. in executable/sample_data/models/ chr11_10kb_gm12878_list_60mb_70mb_1514493462531.gss) to visualize,

b) A gene expression data file in GCT format (http://software.broadinstitute.org/cancer/software/genepattern/file-formats-guide#GCT), an example file is executable/sample_data/gene_expression/allaml.dataset.gct .

c) And a text file to specify genomic coordinates of probes/genes in the GCT format file (each line consists of 4 elements separated by space or tab, e.g.: probe_or_gene_name chr_number start end). A sample is executable/sample_data/gene_expression/probe_coordinates.txt

These 3 following files are prepared for demo: executable/sample_data/models/ chr11_10kb_gm12878_list_60mb_70mb_1514493462531.gss,

executable/sample_data/gene_expression/allaml.dataset.gct and executable/sample_data/gene_expression/ probe_coordinates.txt.

3.5.3 Output

Expression levels of genes/probes are annotated in the 3D model. Usually, the GCT file contains several samples and therefore, the median value (across all samples) together with minimum and maximum values (in brackets) are displayed next to probe/gene names.
If the 3D model and the gene expression data file have no overlap, no annotation will be added to the 3D model.

3.5.4 Running

Access the function from the menu toolbar: 3D-Functions/Model Annotation. A GCT file must be filled in the “Track File” field.

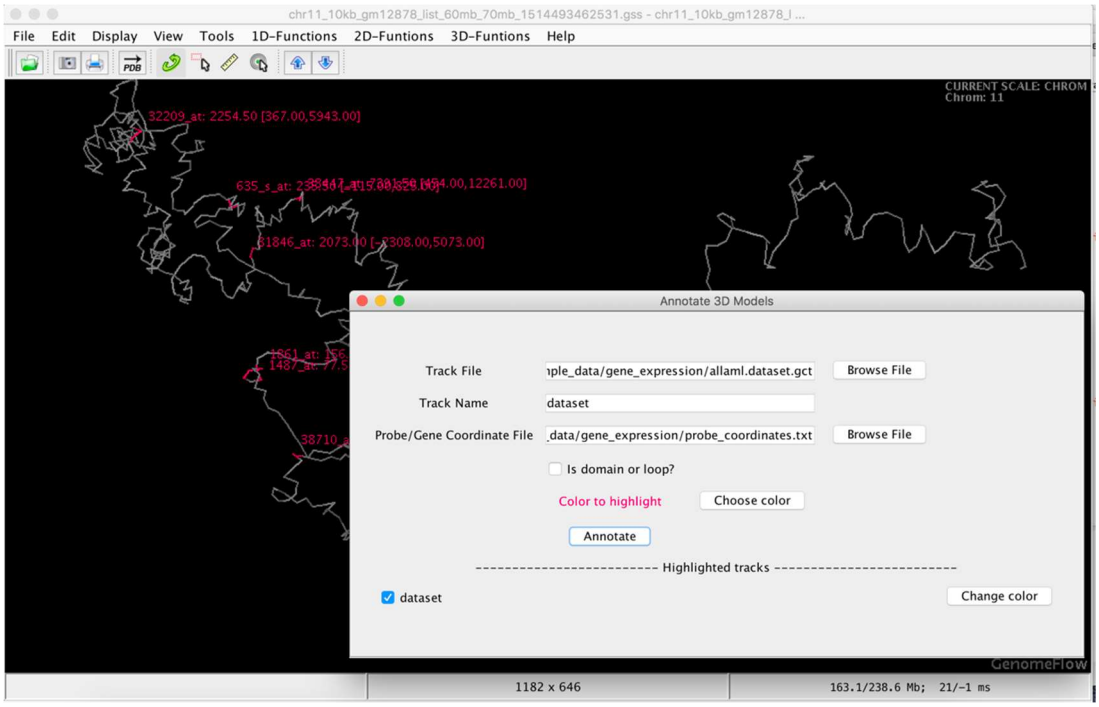

Figure 18: Gene expression visualization demonstration

3.6 Comparing 2 models

3.6.1 Purpose

To superimpose and compare two 3D-models in GSS format.

3.6.2 Input

Two chromosome models in GSS format.

3.6.3 Output

The two models are scaled, superimposed and visualized. Spearman's correlation and RMSE between pairwise distances of the two models are calculated.

3.6.4 Running

Access the function from the menu toolbar: 3D-Functions/Compare Models

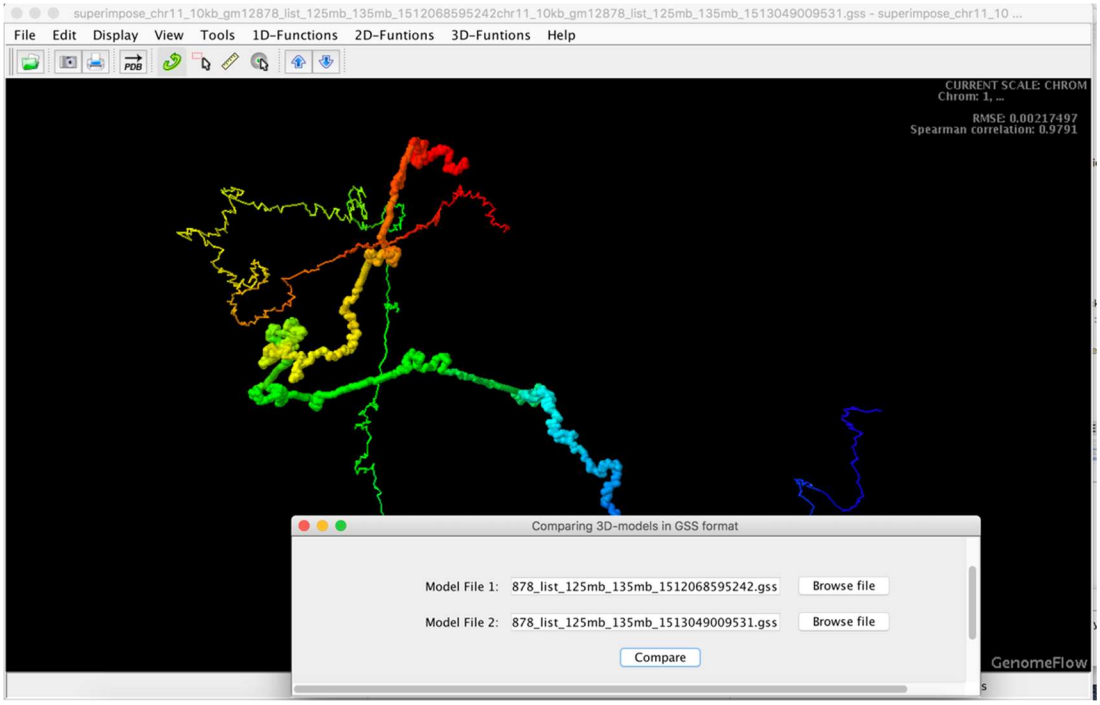

Figure 19: Comparing two constructed models