February 2023

# EpiTect Hi-C Data Analysis Portal User Guide

For use with the EpiTect Hi-C Data Analysis portal. Provides a detailed description of its Hi-C Pro-based analysis pipeline, including explanations for all output files and links to relevant resources for additional information.

For Research Use Only. Not for use in diagnostic procedures.

Sample to Insight

# Contents

# 1 Introduction

At the online GeneGlobe® Data Analysis Center of QIAGEN®, Hi-C sequencing results can be analyzed using the EpiTect Hi-C Analysis Portal. Sequencing reads are first processed through a pipeline based on the open-source HiC-Pro toolset (1) to generate a sequencing report and Hi-C contact matrices. Upon completion of the data analysis, an installation of HiGlass (2) within GeneGlobe can be used to visualize and interact with the generated contact matrices.

The portal is able to analyze Hi-C data generated from the following organisms: human, mouse, brown rat, zebrafish, fruit fly, pig, and northern white-cheeked gibbon.

## About this user guide

This user guide provides information about the EpiTect Hi-C Analysis Portal in the following sections:

- Introduction
- Operating Procedures
- Technical Data
- References
- Ordering Information

## 1.1 General information

### 1.1.1 Technical assistance

At QIAGEN, we pride ourselves on the quality and availability of our technical support. Our Technical Services Departments are staffed by experienced scientists with extensive practical and theoretical expertise in molecular biology and the use of QIAGEN products. If you have any questions or experience any difficulties regarding the EpiTect Hi-C Data Analysis portal or QIAGEN products in general, do not hesitate to contact us.

QIAGEN customers are a major source of information regarding advanced or specialized uses of our products. This information is helpful to other scientists as well as to the researchers at QIAGEN. We therefore encourage you to contact us if you have any suggestions about product performance or new applications and techniques.

For technical assistance, contact QIAGEN Technical Services via your regional technical support number, available at **www.qiagen.com/support**

### 1.1.2 Policy statement

It is the policy of QIAGEN to improve products as new techniques and components become available. QIAGEN reserves the right to change specifications at any time. In an effort to produce useful and appropriate documentation, we appreciate your comments on this user guide. Please contact QIAGEN Technical Services via your regional technical support number, available at **www.qiagen.com/support**

## 1.2 Intended use of the EpiTect Hi-C Data Analysis portal

The EpiTect Hi-C Data Analysis Portal is intended to be used only in combination with QIAGEN kits indicated for use with the EpiTect Hi-C Data Analysis Portal for applications described in the respective QIAGEN kit product sheets or handbooks.

The EpiTect Hi-C Data Analysis Portal is intended for research use only. Not for use in diagnostic procedures.

The EpiTect Hi-C Data Analysis Portal is intended for use by professional users trained in molecular biology techniques.

# 2 Operating Procedures

## 2.1 Workstation requirements

The EpiTect Hi-C Data Analysis Portal should be used with a Google Chrome® or Mozilla® Firefox® browser. The portal is not compatible with Internet Explorer®.

## 2.2 Getting started

### 2.2.1 Accessing the portal5

1. Go to **geneglobe.qiagen.com** and click **My Account**, Figure 1.

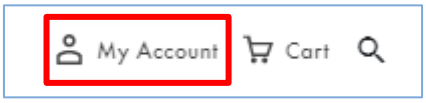


**Figure 1. My Account icon.**

2. Input your QIAGEN username and password to log in.

   **Note**: If you are a new user, select Register now to create an account first.

3. In the panel at the top of the page click **Analyze,** Figure 2.


**Figure 2. Analyze icon.**

4. On the **Analyze** page, move to the **Start Analyzing Your Data** field and select:
   - 4a. **Next-Generation Sequencing** as analysis type (Figure 3a);
   - 4b. **DNA** as the analyte (Figure 3b), and;
   - 4c. **EpiTect HiC** as the panel (Figure 3c);
5. Once those three have been selected, click **Start Your Analysis** (Figure 3d) to proceed.

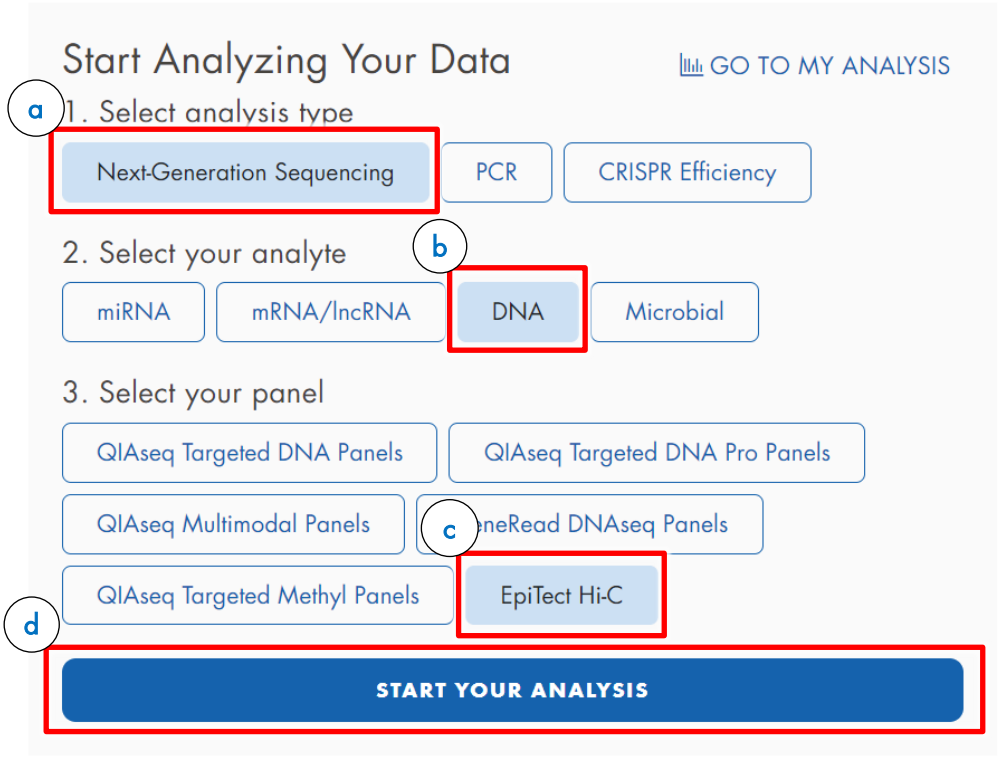


**Figure 3. Start Analyzing Your Data field.** (3a) Next Generation Sequencing button; (3b) DNA button; (3c) EpiTect Hi-C; (3d) Start Your Analysis button.

6. You will be automatically transferred to the Hi-C Analysis portal, which has 2 tabs:

- **Read Files**: For uploading and managing sequencing data files. Files to be analyzed must be selected from within the "2.3 Read Files" tab, see below
- **Analysis Jobs**: For analyzing data uploaded from either the BaseSpace Files or the File Upload tab.

## 2.3 Read Files

The Read Files tab serves two important functions: (1) This section of the portal is used for uploading and managing sequencing data; (2) All analysis jobs begin in the Read Files tab. Sequencing data that is to be analyzed must first be selected for analysis from within the Read Files tab.

### 2.3.1 Read Files: Uploading sequence files

The Read Files tab contains two subsections for uploading data:

- **Uploaded**: For uploading low-depth sequencing data (<3 gigabytes per file) from a local computer.

- **BaseSpace**: For uploading either low-depth or high-depth sequencing data (≥3 gigabytes per file) using Illumina® BaseSpace®.

  **Note**: **High-depth sequencing data** must be uploaded through the BaseSpace Files tab, using Illumina® BaseSpace®

#### 2.3.1.1 Local computer file upload

**Note**: Users may only use their local computer for uploading low-depth sequencing data. For uploading high-depth sequencing data, Illumina BaseSpace must be used.

1. From the EpiTect Hi-C Data Analysis Portal, refer to Figure 4 and select the **Read Files** tab (Figure 4a) > **Uploaded** subsection (Figure 4b) and > **Upload New Files** (Figure 4c) to navigate to the **Read Files Uploader** (Figure 5).

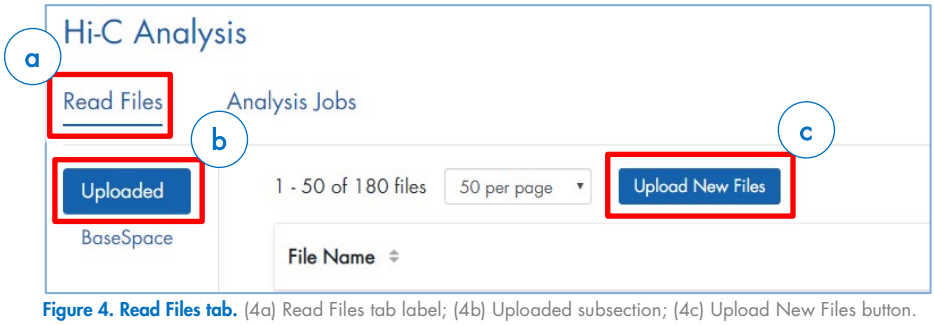


**Figure 4. Read Files tab.** (4a) Read Files tab label; (4b) Uploaded subsection; (4c) Upload New Files button.

2. The file upload process begins by first linking sequencing data to the **Read Files Uploader**. Here are three ways to do so:
   2a. Drag and drop the file into uploading area (Figure 5a)
   2b. Copy and paste read files into uploading area (Figure 5a).
   2c. Click **browse** (Figure 5b) to select read files from their local computer. When selecting multiple files at once, use the Shift or Ctrl key.
3. Once files have been selected, click **Open** to link files to the Read Files Uploader.

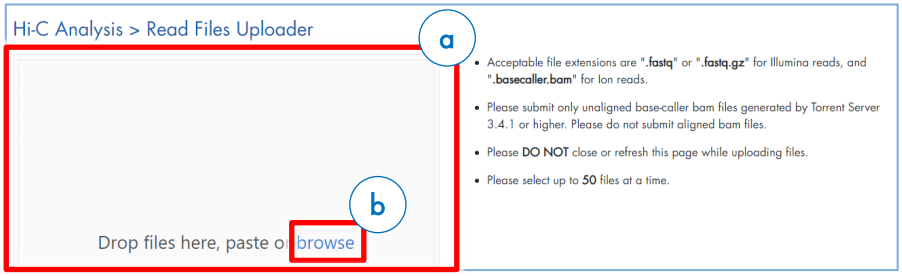


**Figure 5. Read Files Uploader area.** (5a) Outlined box indicating uploading area; (5b) Browse button for uploading.

**Note**: File upload has not begun at this point; the files are only linked to the uploader.

**Important**: For each sample, 2 separate read files must be uploaded. The filename for the read file 1 must contain "_R1". The filename for the read file 2 must contain "_R2".

4. Successfully linked read files will appear as icons in the **Read Files Uploader** page. A white x-mark above a file icon indicates that the file is linked but not yet uploaded.

   4a. Users can click **Add More** (**Figure 6**a) to link additional files to the **Read Files Uploader**.

   4b. Alternatively, clicking **Cancel** (**Figure 6**b) will remove all file links from the **Read Files Uploader** page.

5. Once all files are linked, click **Upload [number of files]** (**Figure 6**c) to begin file upload.

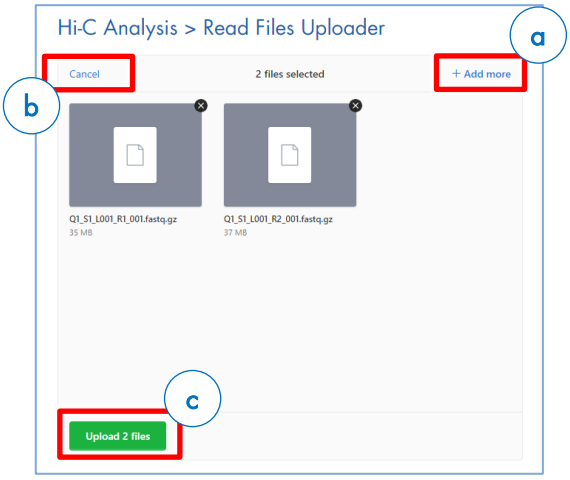


**Figure 6. Read Files Uploader page with successfully linked read files.** (6a) Add More button; (6a) Cancel button; (6c) Upload [number of files] button.

6. Following successful upload, a green check mark will appear above each file icon. The upload button below will also change to show the word **Complete**.

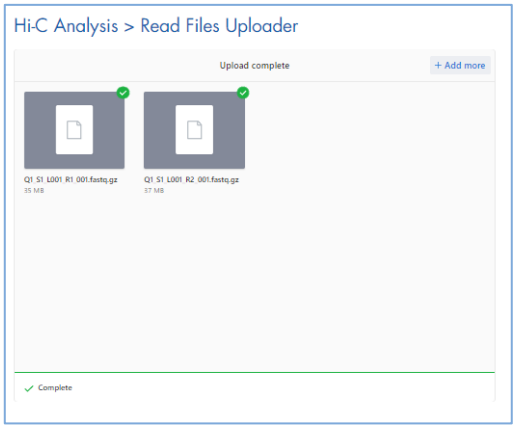


**Figure 7. Read Files Uploader page with successfully uploaded files.**

7. Read files have now been uploaded to the EpiTect Hi-C Data Analysis portal and can be found in the **Uploaded** subsection of the **Read Files** tab.

   **Note**: Uploaded files will only be retained for 90 days from their upload date.

8. To return to the Read Files tab, click on the **Hi-C Analysis** link.


**Figure 8. Link to Hi-C Analysis which will return you to the Read Files tab.**

9. To analyze the uploaded read files, proceed to the 2.4 Analyzing uploaded sequence files section on, page 11.

#### 2.3.1.2 BaseSpace Files upload

1. From the EpiTect Hi-C Data Analysis Portal, select the **Read Files** tab.
2. Select **BaseSpace subsection**, and click **Recent Projects** or **Recent Runs**, as appropriate, to continue.


**Figure 9. BaseSpace Subsection area.**

3. Enter login credentials for your Illumina account, and then select **Sign In**.


**Figure 10. Illumina Sign In Page**

4. Upon successful login, you will be returned to the **BaseSpace subsection** in the Hi-C Analysis portal.

   **Note**: If the User Agreement page appears, you need to accept the agreement to continue, by clicking **I Accept These Agreements** at the bottom of the page.
5. Under the **Name** column, select the project or run that you want to open.
6. Click **Grant Download Permission** to give QIAGEN limited access to your BaseSpace files.

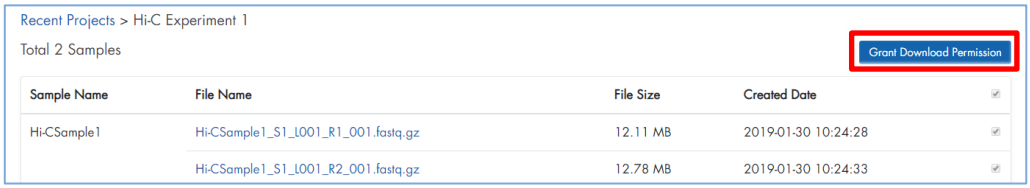


**Figure 11. BaseSpace files within a project.**

7. Read files have now been transferred to the EpiTect Hi-C Data Analysis portal and can be found in the BaseSpace subsection of the Read Files tab.
8. To analyze the uploaded read files, proceed to the "2.4 Analyzing uploaded sequence files" section, page 11

## 2.4 Analyzing uploaded sequence files

1. From the EpiTect Hi-C Data Analysis Portal, select the **Read Files** tab.
2. In either the **Uploaded** or **BaseSpace** subsection, navigate to the read files that are to be analyzed in a single job.

   **Note**: Read files from the Uploaded and BaseSpace subsections cannot be analyzed together in a single analysis job. Instead, data from the Uploaded and BaseSpace subsections must be analyzed in separate analysis jobs.
3. Select the corresponding R1 and R2 files for each sample to be analyzed by ticking the boxes (**Figure 12**)

   **Important**: For each sample, the corresponding R1 and R2 read files must be selected.

   **Important**: Only one reference genome may be used per analysis job. Sequencing data derived from different species (e.g. human and mouse samples) must be processed in separate analysis jobs.

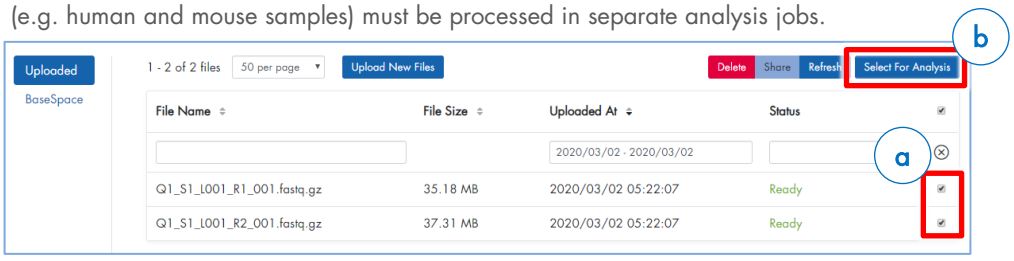


**Figure 12. Uploaded Files subsection with R1 and R2 files.** (12a) Ticked boxes indicating the corresponding R1 and R2 read files; (12b) Select For Analysis button.


**Figure 13. BaseSpace Files subsection with R1 and R2 files.** (13a) Ticked boxes indicating the corresponding R1 and R2 read files; (13b) Select For Analysis button.

4. Once all files are selected, click **Select For Analysis** (**Figure 12b or Figure 13**b), which will bring you to the **Analysis Jobs** tab, where you can create a new analysis job.

   **Important**: Once transferred to the Analysis Jobs tab, no additional sequencing read files can be added to the analysis job.

5. In the **Job Title** field (a), give your analysis job a name. Using the **Genome Build dropdown** menu (**Figure 14**b), select the **reference genome** (refer to Table 1) you would like to use for the analysis. Then, click **Analyze** (**Figure 14**c).

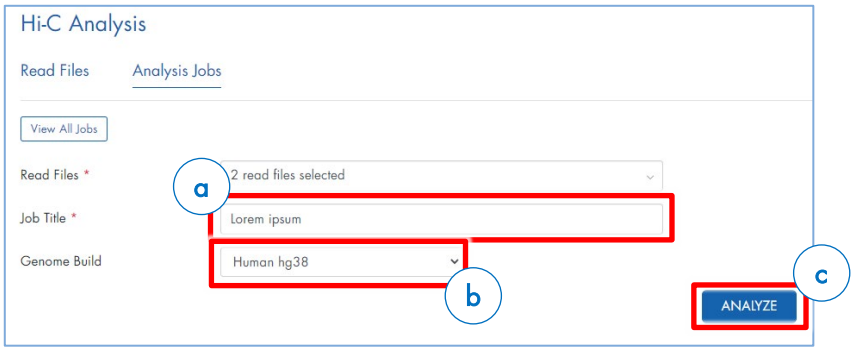


**Figure 14. Analysis Jobs tab field.** (14a) Job Title Field; (14b) Genome Build dropdown menu; (14c) Analyze button.

The following reference genomes are available:

**Table 1. Available Reference Genomes**

| Common Name | Scientific Name | Genomes |
|---|---|---|
| Human | *Homo sapiens* | hg38 and hg19 |
| pHouse mouse | *Mus musculus* | mm10 |
| Brown rat | *Rattus norvegicus* | rn6 |
| Zebrafish | *Danio rerio* | GRCz11 |
| Fruit fly | *Drosophila melanogaster* | BDPGP6.22 |
| Pig | *Sus scrofa* | Sscrofa11.1 |
| Northern white-cheeked gibbon | Nomascus leucogenys | Nleu3 |

**Note**: If you navigate away from the Analysis Jobs tab before starting the analysis job, the selected files will still be available. Simply navigate to the Analysis Jobs tab and click "Create New Jobs". You will be transferred back to the window where the analysis job is given a name and the reference genome is selected.

6. After giving the analysis job a name and selecting **Select Analyze** to start the Hi-C analysis. Initially, the job status will appear as **Queued**. Once analysis is complete, the job status will change to **Done Successfully** with two links:
   - 6a. Select **Download Report** to download a ZIP file containing the **EpiTect Hi-C Analysis report** to your local hard drive.
   - 6b. Select **HiGlass Analysis** to proceed to an installation of HiGlass, where you can visualize the contact matrices generated from your Hi-C sequencing data

For further details about the analysis and contents of the EpiTect Hi-C Analysis report, refer to the "Technical Data" section (page 17).

## 2.5 Sharing your data with other GeneGlobe users

Read files copied to GeneGlobe via the upload function and completed analysis jobs can be shared with other GeneGlobe users.

### 2.5.1 Sharing sequencing read files with GeneGlobe users

1. In the **Read Files** tab, select the corresponding R1 and R2 files for each sample to be shared by ticking the boxes next to each file (Figure 15a). Then click **Share** (Figure 15b).

   **Important**: For each sample, the corresponding R1 and R2 read files must be selected.

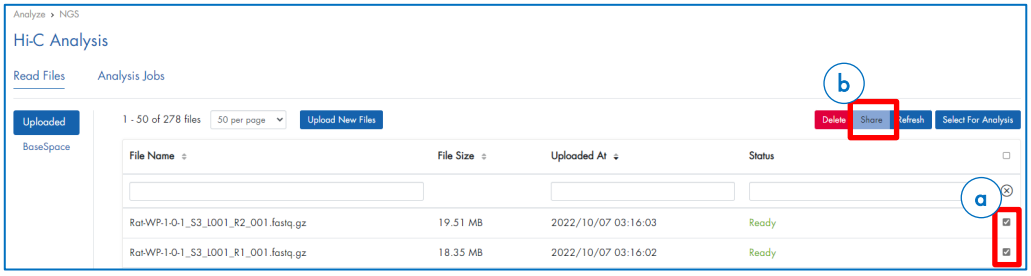


**Figure 15. Hi-C Analysis Read Files tab containing R1 and R2 files of each sample ready for sharing.** (15a) Ticked boxes indicating the corresponding R1 and R2 read files; (15b) Share button.

2. In the window that will appear after clicking Share, enter the email address of a registered GeneGlobe user and click **Share**.

   **Note**: Users will receive error messages if: (a) the entered email address does not correspond to a registered GeneGlobe account or (b) if the registered GeneGlobe account has not accessed the EpiTect Hi-C portal at least once prior to sharing.

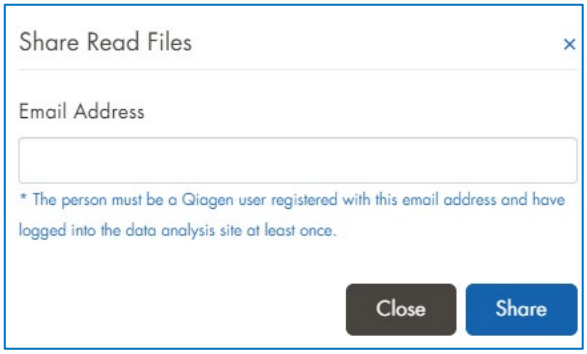


**Figure 16. Window for inputing registered GeneGlobe user email address.** This is to share the corresponding Read Files.

3. The text "**Shared to XXXX**" will appear in the Status field next to each read file successfully shared with another GeneGlobe account.

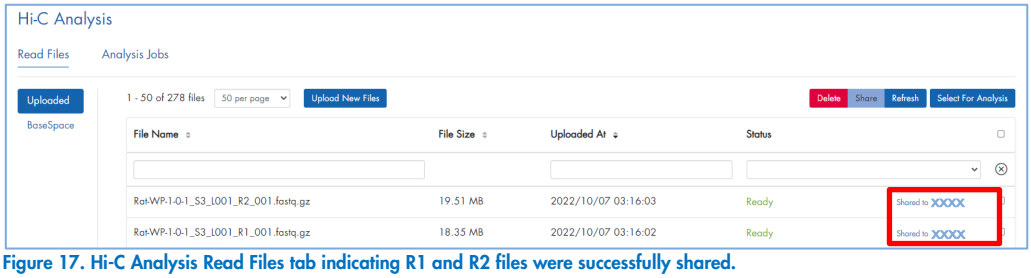


**Figure 17. Hi-C Analysis Read Files tab indicating R1 and R2 files were successfully shared.**

4. In the Hi-C Analysis portal of the GeneGlobe account where the read files have been shared, the text "**Shared by XXXX**" will appear in the Status field next to each read file successfully shared from another GeneGlobe account.

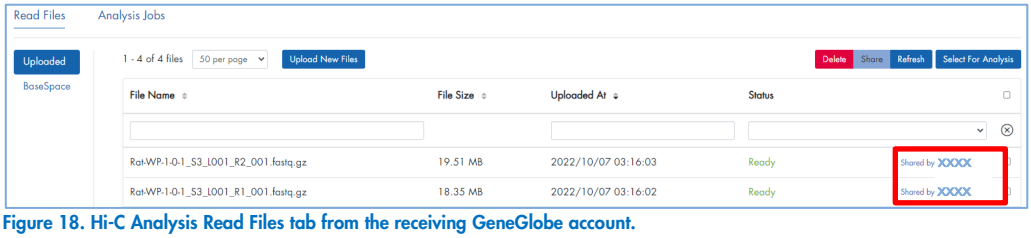


**Figure 18. Hi-C Analysis Read Files tab from the receiving GeneGlobe account.**

5. To stop sharing a read file with another GeneGlobe account, click on the "**Shared to XXXX**" text found (Figure 19a) in the Status field of the corresponding read file. A window will pop up with the name and email address of the GeneGlobe account with which the file is shared. Click on the "**X**" (Figure 19b) symbol, to stop sharing.

   **Note**: Users must stop sharing read files one at a time.

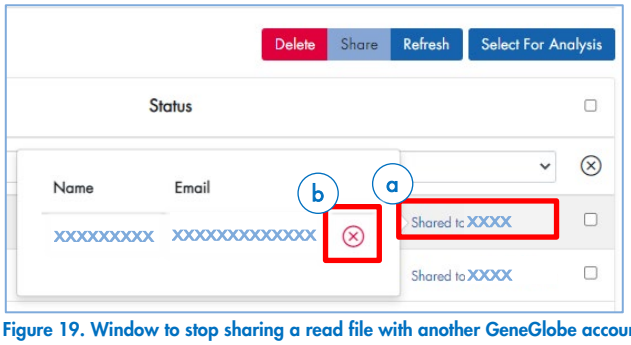


**Figure 19. Window to stop sharing a read file with another GeneGlobe account.** (19a) "Shared to" text; (19b) "X" or remove symbol.

### 2.5.2 Sharing completed analysis jobs with GeneGlobe users

1. In the **Analysis Jobs** tab, select the analysis jobs to be shared by ticking the boxes next to each analysis job. Then click **Share.**

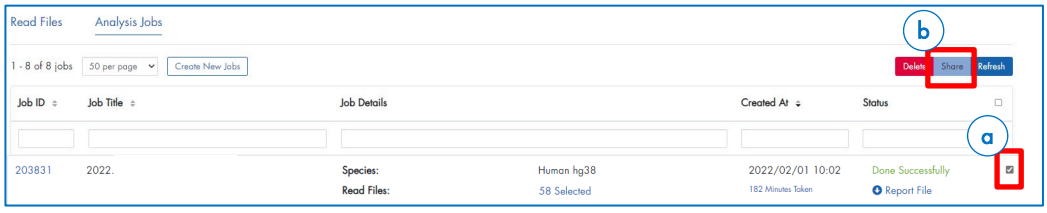


**Figure 20. Analysis Jobs tab with completed analysis jobs.** (20a) Ticked boxes indicating the corresponding R1 and R2 read files; (20b) Share button.

2. In the window that appears, enter the email address of a registered GeneGlobe user and click Share.

   **Note**: Users will receive error messages if the entered email address does not correspond to a registered GeneGlobe account or if the registered GeneGlobe account has not accessed the EpiTect Hi-C portal at least once prior to sharing.

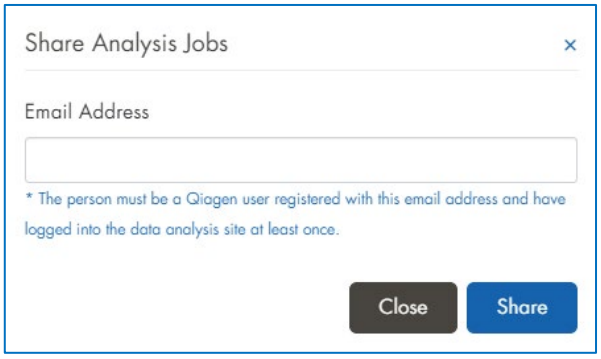


**Figure 21. Window for inputing registered GeneGlobe user email address.** This is to share the corresponding Analysis Jobs.

3. The text "**Shared to XXXX**" will appear in the Status field of each analysis job successfully shared with another GeneGlobe account.


**Figure 22. Analysis Jobs tab indicating the analysis job/s were successfully shared.**

4. In the Hi-C Analysis portal of the GeneGlobe account with which an analysis job has been shared, the text "**Shared by XXXX**" will appear in the Status field next to the analysis job successfully shared from another GeneGlobe account.

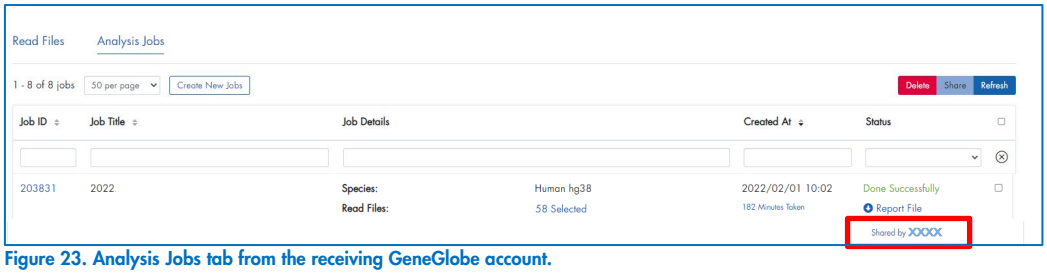


**Figure 23. Analysis Jobs tab from the receiving GeneGlobe account.**

5. To stop sharing a read file with another GeneGlobe account, click on the "**Shared to XXXX**" text found in the Status field of the corresponding read file. A window will pop up with the name and email address of the GeneGlobe account with which the file is shared. Click on the "**X**" symbol, to stop sharing.

   **Note**: Users must stop sharing analysis jobs one at a time.

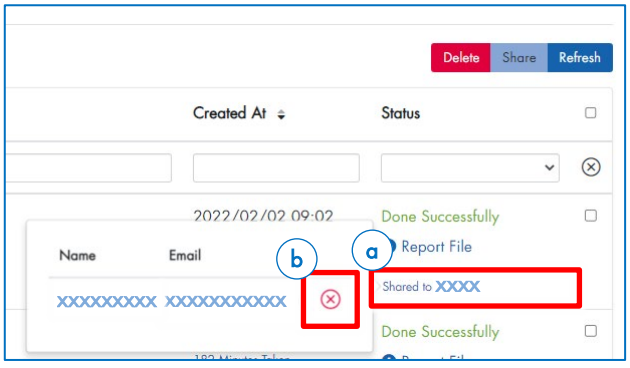


**Figure 24. Window to stop sharing analysis jobs with another GeneGlobe account.** (20a) "Shared to" text; (20b) "X" or remove symbol.

# 3 Technical Data

This section contains information on the EpiTect Hi-C Analysis pipeline.

## 3.1 HiC-Pro

The core of the read-analysis for Hi-C reads is the published (1) pipeline named HiC-Pro. The Source-code can be found on GitHub (**https://github.com/nservant/HiC-Pro**). The HiC-Pro workflow can be divided into steps presented below:

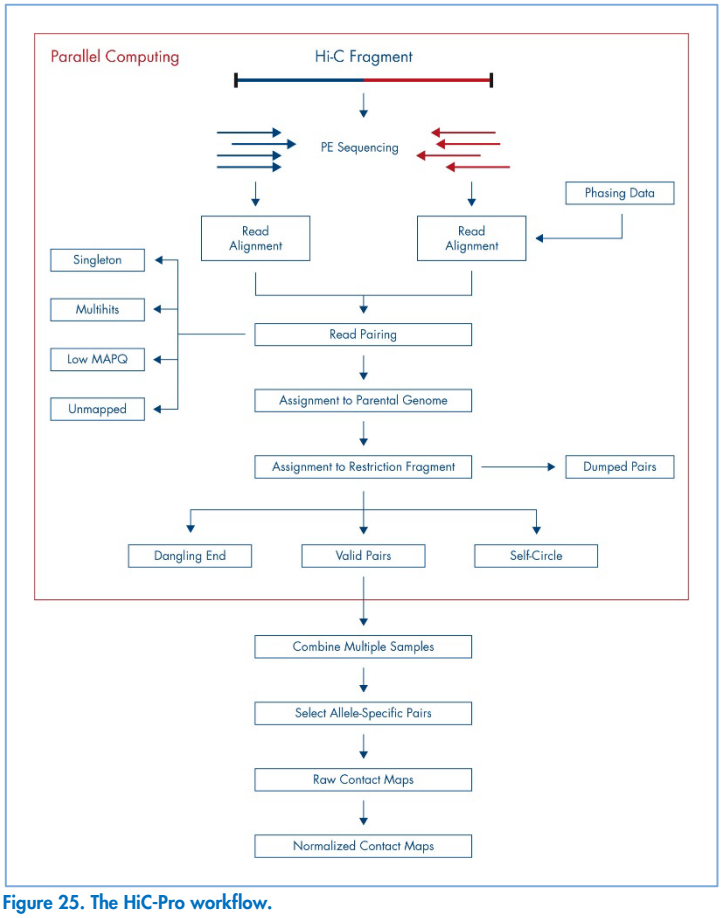


**Figure 25. The HiC-Pro workflow.**

### 3.1.1 Reads mapping

Each mate is independently aligned on the reference genome. The mapping is performed in two steps: First, the reads are aligned using an end-to-end aligner. Second, reads spanning the ligation-junction are trimmed from their 3' end and aligned back on the genome.

Aligned reads for both fragment mates are then paired in a single paired-end BAM file. Singletons and multiple hits can be discarded according to the configuration parameters.

### 3.1.2 Fragment assignment and filtering

Each aligned read can be assigned to one restriction fragment, according to the reference genome and the restriction enzyme.

The next step is to separate the invalid ligation products from the valid pairs. Therefore, dangling-end and self-circles pairs are excluded. Only valid pairs involving two different restriction fragments are used to build the contact maps. Duplicated valid pairs associated with PCR artifacts are discarded.

The fragment assignment can be visualized through a BAM file of aligned pairs where each pair is flagged according to its classification.

### 3.2.3 Quality controls

HiC-Pro performs a couple of quality controls for most of the analysis steps. The alignment statistics are the first quality controls. Aligned reads in the first (end-to-end) step and alignment after trimming are reported.

> **Note**: We usually observe around 10–20% of trimmed reads. An abnormal level of trimmed reads can reflect a ligation issue.

Once the reads are aligned on the genome, HiC-Pro checks the number of singletons, multiple hits or duplicates. The fraction of valid pairs are presented for each type of ligation product. Invalid pairs (such as dangling ends or self-circles) are also represented. A high level of dangling ends or an imbalance in valid pairs ligation type can be due to a ligation, fill-in or digestion issue.

Finally, HiC-Pro also calculates the distribution of fragment size on a subset of valid pairs. Additional statistics report the fraction of intrachromosomal versus interchromosomal contacts, as well as the proportion of short-range (<20 kb) versus long-range
(>20 kb) contacts.

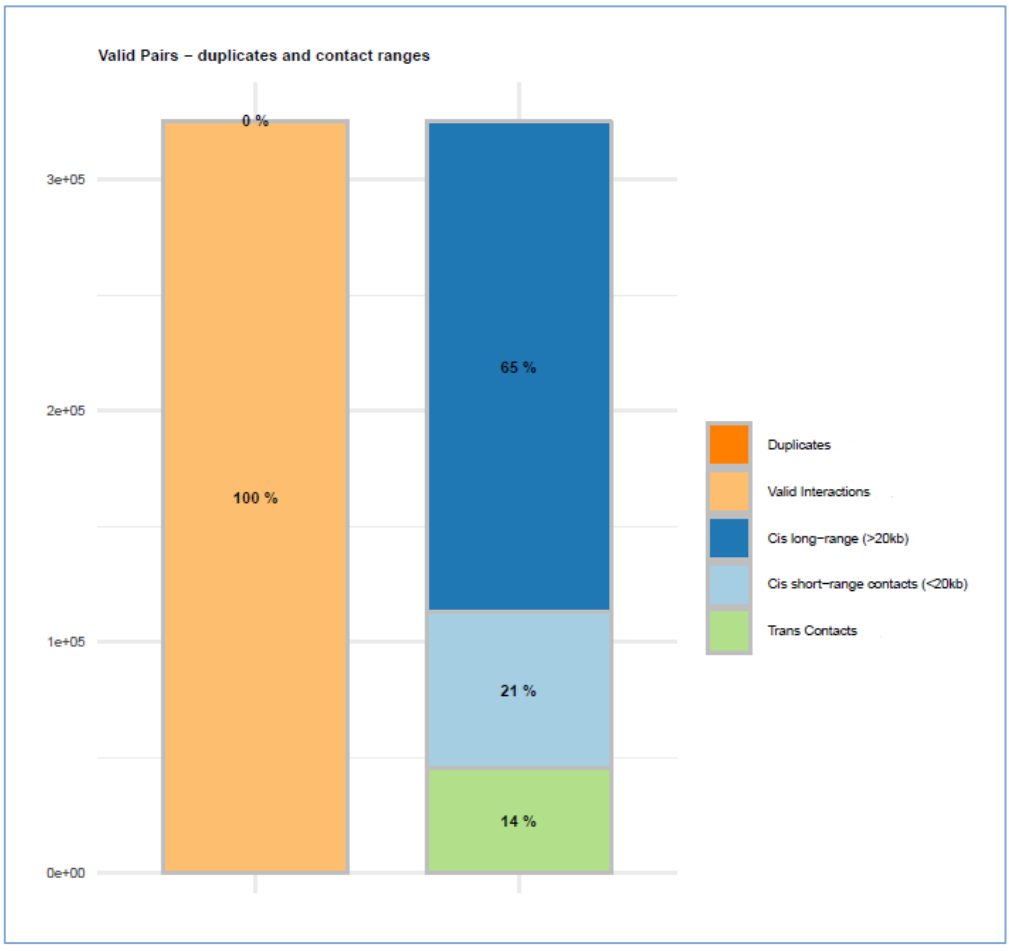


**Figure 26. Valid pairs, duplicates and contact ranges are presented in the "plotHiCContactRanges" file.**

All quality control plots of the abovementioned measures can be found in the subfolder "**/pic**". An example of the **plotHiCContactRanges** file is presented above.

### 3.2.4 Map builder

Intrachromosomal and interchromosomal contact maps are built for a resolution of 2500 bases. The genome is split into bins of equal size. Each valid interaction is associated with the genomic bins to generate the raw map.

### 3.2.5 Post processing

A few additional tools are run to convert the output files from HiC-Pro into formats that are commonly used and compatible with a range of downstream analysis tools and visualization software.

### 3.2.6 Pairs

A custom script is used to convert the **allValidPairs** file into the **PAIRS format** according to specifications defined by the 4D Nucleome Network (**www.4dnucleome.org**). This file includes the two optional columns, named "frag1" and "frag2", which contain the integer index of the restriction fragment in the genome.

This file is the most comprehensive result because it contains all individual contacts at single-base resolution (no binning), including the readID of the library fragment from the input read data. The uncompressed plain-text file looks like the example below:

```
## pairs format v1.0
#sorted: chr1-chr2-pos1-pos2
#shape: upper triangle
#genome_assembly: hg38
#columns: readID chr1 pos1 chr2 pos2 strand1 strand2 frag1 frag2
BJFY6:1:1112:5915:11069      chr1    629270  chr1      81381488      -     +       1195     216143
BJFY6:1:2101:14333:2314      chr1    629484  chr1     242769744      -     -       1195     568557
BJFY6:1:2117:22179:2271      chr1    631129  chr1        634141      +     -       1195       1202
BJFY6:1:2116:17587:24798     chr1    631237  chr1     200809963      +     +       1195     461302
BJFY6:1:1110:8536:16896      chr1    631251  chr1        633698      +     +       1195       1199
```

The tool pairix was used to create the index "**.pairs.gz.px2**". The index is required before running pairix queries. Using the pairs file is a solid choice for any downstream analyses other than visualization (see below).

Output files are stored in a folder named "**/pairs**" and the data are compressed ("**.pairs.gz**").

### 3.2.7 Juicer tools (.hic)

The "pre" command of Juicer Tools (**github.com/aidenlab/juicer**) is run to create HIC files (located in the "**/hic**" subfolder) based on the PAIRS file. The HIC file is a highly compressed binary file that stores contact matrices from multiple resolutions in a clever way, allowing random access. The format is described extensively by Durand et al. in *Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments (3).*

The visualization tool Juicebox (**www.aidenlab.org/juicebox**) uses the fast querying capabilities of HIC files to make it possible to zoom in and out of many different resolutions quickly. Juicebox is available for Windows®, macOS® and Linux, and there is even an online version.

### 3.2.8 Cooler (.mcool)

Cooler is a support library for a sparse, compressed binary persistent storage format called COOL, used to store genomic interaction data, such as Hi-C contact matrices. The COOL file format is a reference implementation of a genomic matrix data model using HDF5 as the container format.

The **Cooler** tool (**github.com/mirnylab/cooler**) is run twice: (1st run) a 2500 bp binned contact matrix generated by HiC-Pro is converted into the COOL file format running cooler load. (2nd run) The result is enriched by **cooler zoomify** to generate a multi-resolution contact-matrix file MCOOL, stored in the subfolder named "**/mcool**". These files can be visualized by HiGlass (**higlass.io**), a feature-rich and very responsive browser-based software package (4).

## 3.3 Output files

Every analysis job will generate two overview files in top-level of the results and a range of additional data files (plots and contact matrices) organized in subfolders.

### 3.3.1 Multi-QC HTML

MultiQC (**multiqc.info**) is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples (5). The MultiQC HiC-Pro module, available at **github.com/ewels/MultiQC/tree/master/multiqc/modules/hicpro**, parses the results generated by HiC-Pro, and it was written by the same author as that of the HiC-Pro itself.

### 3.3.2 Excel® file summary

A custom script aggregates the HiC-Pro results and exports them into a Microsoft® Excel workbook. Values are presented either as an absolute number of counts or as percentages of total. The results for each Hi-C sample are organized into a single column. The Excel report is divided into several sections that represent the different processing stages of HiC-Pro.

#### 3.3.2.1 Mapping R1/Mapping R2

Usually, a high fraction of reads is expected to be aligned on the genome (80–90%). Among them, a small percent (around 10–20%) align only after trimming. This can be the result of chimeric fragments in which reads extend over Hi-C ligation junction boundaries. An abnormal level of chimeric reads can reflect a ligation issue during the library preparation. Total_R1/R2: all R1 and R2 reads submitted to HiC-Pro.

- **Mapped_R1/R2**: all reads which aligned to the reference genome.
- **Global R1/R2**: reads which aligned to the reference genome without trimming
- **Local_R1/R2**: reads which aligned to the genome after soft trimming

#### 3.3.2.2 Mapping: Pairs

Once R1 and R2 reads are aligned on the genome, HiC-Pro reconstructs the pairs information. The fraction of singletons or reads with multiple mapping locations depends on the complexity of the genome and the fraction of unmapped reads. The

fraction of singletons is usually close to the sum of unmapped R1 and R2 reads, because it is unlikely that both mates from the same pair were unmapped.

- **Total_pairs_processed**: all read pairs analyzed by HiC-Pro
- **Unique_paired_alignments**: read pairs where each read is uniquely mapped to the reference genome
- **Multiple_pairs_alignments**: read pairs where one or both reads are mapped to multiple locations in the reference genome
- **Low_qual_pairs**: discarded read pairs where one or both reads do not pass the MIN_MAPQ threshold during mapping
- **Unmapped_pairs**: read pairs where both reads did not map to the reference genome
- **Pairs_with_singleton**: read pairs where one of the reads does not map to the reference genome
- **Unique_singleton_alignments**: read pairs where one of the reads does not map to the reference genome and one of the reads maps to reference genome once
- **Multiple_singleton_alignments**: read pairs where one of the reads does not map to the reference genome and one of the reads maps to multiple locations in the reference genome
- **Low_qual_singleton**: read pairs where one of the reads does not map to the reference genome and one of the reads does not pass the MIN_MAPQ threshold during mapping
- **Reported_pairs**: read pairs that are analyzed further to identify and characterize Hi-C interactions

#### 3.3.2.3 Pairs: Valid Hi-C Pairs

**Reported_pairs** are processed further to identify valid Hi-C interaction pairs. Each aligned read can be assigned to one restriction fragment according to the reference genome and the selected restriction enzyme. Both reads are expected to map near a restriction site at a distance within the range of molecule size distribution after shearing.

Fragments with a size outside the expected range can be discarded if specified. However, they are are usually the result of random breaks or star activity of the enzyme, and can therefore be included in downstream analysis.

Invalid ligation products, such as dangling end and self-circle ligation, are discarded. A high level of dangling-end or self-circle read pairs is associated with a low-quality experiment and reveals a problem during the digestion, fill-in or ligation steps.

Only valid Hi-C interaction pairs involving two different restriction fragments are used to build the contact maps. Duplicated valid pairs due to PCR artifacts can also be filtered out.

**Note**: A high level of PCR duplicates indicates poor molecular complexity and a potential PCR bias.

- **Reported_pairs**: read pairs that are analyzed further to identify and characterize valid Hi-C interaction pairs
- **Dangling_end_pairs**: unligated fragments where both reads mapped to the same restriction fragment
- **Religation_pairs**: ligation of juxtaposed restriction fragments

- **Self-circle_pairs**: fragments ligated to themselves, where both reads mapped to the same restriction fragment in an inverted orientation
- **Single-end_pairs**: singletons that are filtered out during mapping
- **Filtered_pairs**: read pairs that are filtered during mapping
- **Dumped_pairs**: any pairs that do not match the filtering criteria on inserts size or restriction fragments size, or for which we were not able to reconstruct the ligation product
- **Valid_interaction_pairs**: all valid Hi-C interaction pairs that remain after filtering out invalid species
- **Valid_interaction_rmdup**: all valid Hi-C interaction pairs that remain after PCR duplicates are removed

#### 3.3.2.4 Valid Hi-C Pairs: Strand Bias

25% of each valid ligation class is expected because the ligation is a random process.

- **Valid_interaction_pairs**: all valid Hi-C interaction pairs that remain after invalid species have been removed
- **Valid_interaction_pairs_FF**: valid Hi-C interaction pairs in which R1 and R2 are from the same DNA strand and are oriented in the same direction (forward)
- **Valid_interaction_pairs_FR:** valid Hi-C interaction pairs in which R1 and R2 are from different DNA strands and face inward
- **Valid_interaction_pairs_RF**: valid Hi-C interaction pairs in which R1 and R2 are from different DNA strands and face outward
- **Valid_interaction_pairs_RR**: valid Hi-C interaction pairs in which R1 and R2 are from the same DNA strand and are oriented in the same direction (reverse)

#### 3.3.2.5 Valid Hi-C Pairs: Interaction Distances

The fraction of intrachromosomal and interchromosomal interactions, as well as long-range (>20 kb) versus short-range (<20 kb) intrachromosomal interactions, are also important quality metrics

- **Valid_interaction_pairs**: all valid Hi-C interaction pairs that remain after invalid species are removed
- **Trans_interaction**: all valid Hi-C interaction pairs where R1 and R2 map to different chromosomes (interchromosomal interactions)
- **Cis_interaction**: all valid Hi-C interaction pairs where R1 and R2 map to the same chromosome (i.e. cis or intrachromosomal interactions)
- **Cis_shortRange**: cis interactions where the distance between R1 and R2 is <20 kb
- **Cis_longRange**: cis interactions where the distance between R1 and R2 is ≥20 kb

#### 3.3.2.6 Global Yield (Hi-C)

The relative yield of valid long-range cis Hi-C interaction pairs can be calculated in a variety of ways. In this section, three different calculations are provided for the user.

- **Cis_longRange (from total)**: percentage of valid long-range cis Hi-C interaction pairs from all read pairs
- **Cis_longRange (from reported)**: percentage of valid long-range cis Hi-C interaction pairs from all mapped read pairs
- **Cis_longRange (from valid)**: percentage of valid long-range cis Hi-C interaction pairs from all valid Hi-C interaction pairs

#### 3.3.2.7 Subfolders

The additional output files are generated per sample and can be found in folders as explained above. Please refer to sections above, which explain the files and their format.

## 3.4 Quality control of Hi-C NGS libraries by shallow sequencing

Prior to costly deep sequencing, users are advised to sequence Hi-C NGS libraries at low depth (<1 million reads) for quality control purposes. Low-depth sequencing data can be processed at the EpiTect Hi-C analysis portal, and the generated sequencing report can be used to assess the quality of Hi-C libraries.

### 3.4.1 Characteristics of a high-quality Hi-C NGS library

- Greater than 80% valid Hi-C interaction pairs (after removal of PCR duplicates)
    - According to the guidelines by Rao et al. (6), Hi-C libraries where >20% of the paired-end reads are not valid Hi-C interactions are likely to be the result of failed restriction, fill-in, or ligation steps, and are therefore not good candidates for deeper sequencing.
- Low percentage of read pairs deriving from a single restriction fragment
    - A high-quality Hi-C library for mammalian genomes typically has less than 1–4% unligated, dangling ends and less than 1–2% self-ligated circles.
- Greater than 40% long-range cis interactions (>20 kb)
    - Rao et al. provide guidelines to assess the quality of Hi-C libraries in their supplementary material (extended methods) accompanying their excellent publication (6):

*A crucial metric is the percentage of long-range intrachromosomal contacts. In successful Hi-C libraries, we found that at least 15% of unique reads were long-range intrachromosomal contacts. Lower values usually indicated that the experiment had failed. If more than 40% of unique reads are long-range intrachromosomal contacts, a library was considered a good candidate for sequencing. If the fraction was above half, a library was considered an excellent candidate for sequencing.* ***In general, this value was one of the statistics we found most important to scrutinize in performing cost-effective high-depth Hi-C.***

- Greater than 40% cis/trans ratio
    - In the nucleus, chromosomes are partitioned in territories where individual chromosomes are physically separated in space. For this reason DNA contacts typically occur at a higher frequency within chromosomes (*cis*) than between chromosomes (*trans*). This property of genome organization can be exploited as a useful proxy for evaluating the quality of Hi-C data. Noise from random background ligation (due to ruptured nuclei) will affect both *cis* and *trans* interactions similarly and result in a lower ratio between *cis* and *trans* interactions. *Cis/Trans* ratios are dependent on genome size and number of chromosomes; but for human genomes, ratios of 40–60% are considered a sign of high-quality Hi-C experiments (7). *Cis/Trans* ratio is defined as:

      **[Cis_longRange / (Cis_longRange + Trans_interaction)] x 100% (see section 3.3.2.5)**

- No strand-orientation bias
    - Hi-C chimeras can be broken up into 4 classes distinguished by the strand orientation of read pairs: FF, FR, RF, and RR (see section 0). If the chimeras are a result of random proximity ligation of chromatin, then close to 25% of each class of chimera is expected.

# References


1. Servant, N. et al. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. **16**, 259.
2. Kerpedjiev, P. et al. (2018) HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. **19**, 125.
3. Durand, N.C. et al. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. **3**, 95–98.
4. Kerpedjiev, P. et al. (2018) HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. **19**, 125.
5. Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32**, 3047–3048.
6. Rao, S.S.P. et al. (2014) A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**, 1665–1680.
7. Lajoie, B.R., Dekker, J. and Kaplan, N. (2015) The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. Methods **72**, 65–75.

# Ordering Information

| Product | Contents | Cat. no. |
|---|---|---|
| EpiTect Hi-C Kit (6) | For 6 Hi-C reactions: Buffers and reagents for cell lysis, Hi-C digestion, Hi-C end-labeling, Hi-C ligation, chromatin decross-linking and purification, purification of fragmented DNA, streptavidin pulldown of Hi-C fragments and NGS library prep (end repair, A-addition, phosphorylation, adapter ligation and library amplification); for use with Illumina instruments; includes 6 adapters with different barcodes | 59971 |
| **Related Products** | | |
| **For use with Illumina instruments** | | |
| QIAseq Library Quant Assay Kit | Laboratory-verified forward and reverse primers for 500 x 25 µL reactions (500 µL); DNA standard (100 µL); dilution buffer (30 mL); (1.35 mL x 5) GeneRead qPCR SYBR® Green Mastermix | 333314 |
| **For assessing NGS library quality** | | |
| QIAxcel Advanced Instrument | Capillary electrophoresis device: includes computer, QIAxcel ScreenGel Software and 1-year warranty on parts and labor; fully automates sensitive, high-resolution capillary electrophoresis devices for analyzing up to 96 samples per run | 9001941 |
| QIAxcel DNA High Resolution Kit (1200) | QIAxcel DNA High Resolution Gel Cartridge, buffers, mineral oil, QX Intensity Calibration Marker, 12-tube strips | 929002 |

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at www.qiagen.com or can be requested from QIAGEN Technical Services or your local distributor.

# Document Revision History

| Date | Changes |
|---|---|
| 04/2019 | Initial |
| 02/2023 | Edited Section 2.3 Read Files; added information: 2.3.1 Read Files: Uploading sequence files. Edited according to new brand template. |

**Limited License Agreement for [Product Name]**

Use of this product signifies the agreement of any purchaser or user of the product to the following terms:

1. The product may be used solely in accordance with the protocols provided with the product and this handbook and for use with components contained in the kit only. QIAGEN grants no license under any of its intellectual property to use or incorporate the enclosed components of this kit with any components not included within this kit except as described in the protocols provided with the product, this handbook, and additional protocols available at www.qiagen.com. Some of these additional protocols have been provided by QIAGEN users for QIAGEN users. These protocols have not been thoroughly tested or optimized by QIAGEN. QIAGEN neither guarantees them nor warrants that they do not infringe the rights of third-parties.
2. Other than expressly stated licenses, QIAGEN makes no warranty that this kit and/or its use(s) do not infringe the rights of third-parties.
3. This kit and its components are licensed for one-time use and may not be reused, refurbished, or resold.
4. QIAGEN specifically disclaims any other licenses, expressed or implied other than those expressly stated.
5. The purchaser and user of the kit agree not to take or permit anyone else to take any steps that could lead to or facilitate any acts prohibited above. QIAGEN may enforce the prohibitions of this Limited License Agreement in any Court, and shall recover all its investigative and Court costs, including attorney fees, in any action to enforce this Limited License Agreement or any of its intellectual property rights relating to the kit and/or its components.

For updated license terms, see **www.qiagen.com**.