

LCSeqTools

User Guide

LCSeqTools v0.1.0 User Guide

Author ♦ Marcus Vinicius Niz Alvarez

E-mail ♦ marcus.alvarez@unesp.br

Institute of Biosciences - IB ♦ Department of Genetics

São Paulo State University (UNESP)

Botucatu, São Paulo State

Table of contents

Introduction.....	4
About LCSeqTools.....	4
How to cite.....	4
License.....	4
Quick Start Guide.....	5
The Graphical User Interface Explained.....	6
Project Configuration Window.....	6
Workflow Progress Window.....	11
Statistics Viewer Window.....	11
Output Files.....	12
Other information.....	15
LCSeqTools Workflow Diagram.....	15
Illumina File Naming Convention.....	16
Genotyping Parameters.....	17
Sample-sheet file format.....	18
The Third-Party Package.....	19
References.....	20

Introduction

About LCSeqTools

LCSeqTools is a user-friendly tool for variant calling and imputation using low-coverage whole genome sequencing data. It consists of a series of scripts that automate the steps of quality control, alignment, variant calling, filtering, and imputation using external tools such as Trimmomatic, BWA, SAMtools, LCVCFtools, and Beagle.

How to cite

Not available yet...

License

LCSeqTools is a free and open-source software and is under GPLv3. All Third-Party software used with LCSeqTools are free and open-source (All licenses compatible with GPLv3).

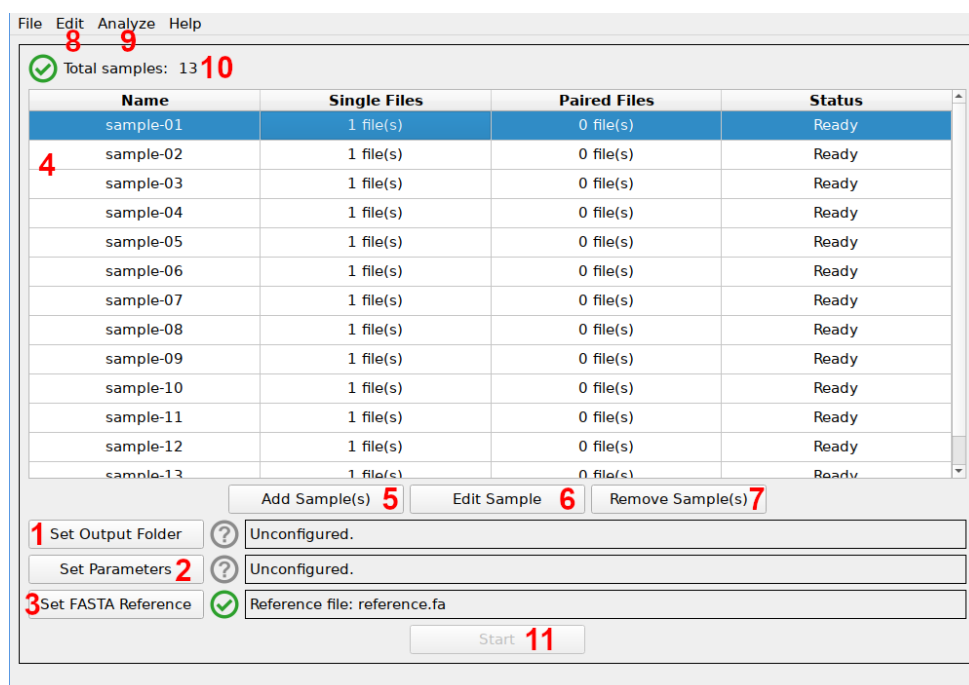
Quick Start Guide

1. Execute LCSeqTools.
2. Click on **File > New Project....**
3. Click on **Set Output Folder** and choose a directory where the results will be outputted.
4. Click on **Set Parameters** and edit the genotyping parameters as desired.
5. Click on **Set FASTA Reference** and choose the mapping reference FASTA file.
6. (Optional) After importing the reference FASTA, click on **Edit > Set Reference Ploidy** to set the correct ploidy of the reference sequences.
7. Click on **Add Sample(s) > Auto-detect** and select a directory that stores a downloaded Illumina run with the FASTQ sequencing files.
8. Click on **Start**.
9. After all steps are finished successfully, check the results located in the Output Folder.

The Graphical User Interface Explained

Project Configuration Window

The first window that will appear when configuring a new project is the project configuration window. The image below is an example of a new project under configuration. All buttons and options are indicated with red numbers.



1. **Set Output Folder:** The project needs an output folder. By default, there is no output folder configured, as we can see in the example by the ? symbol and the corresponding "Unconfigured." message for the respective boxes. After clicking this button, a dialog will appear and the user can select the desired output folder. It must be a folder located in a storage drive that meets the system recommended specifications. If the operation succeed, the ✓ symbols will appear and the output folder will be indicated in the respective box. Otherwise, the ✗ symbol representing an error, and the error message will be displayed in the respective box.

2. **Set Parameters:** The genotyping-by-sequencing workflow parameters need to be adjusted. By default, the project is not configured until this section is properly defined. After clicking this button, a dialog window will appear showing the standard parameters as depicted in the figure below: After editing the parameters, the user can save by clicking on the **OK** button. If the user wants to restore the default parameters values, the user can click on the **Default** button then all the parameters values will be restored to the default values. (See Genotyping Parameters Section for detailed description about the genotyping parameters).

The image shows a 'System Settings' dialog box with the following parameters:

- System Settings:** Max Threads: 6
- Sequence Quality Parameters:**
 - Sequence trimming (Head): 10
 - Sequence trimming (Trailing Crop Quality): 20
 - Sequence minimum length: 40
- Variant Call Parameters:**
 - Minimum Genotype Quality (Phred Scale): 20
 - Minimum Depth: 5
 - Minor Allele Frequency Threshold: 0,10
 - Max missing data (Variant): 0,50
 - Max missing data (Sample): 0,50
- (Post) Imputation Parameters:**
 - Genotype Imputation: ☒
 - Imputation Seed (Reproducible results): 1
 - Minimum Genotype Probability: 0,95
 - Low memory usage: ☐

Buttons at the bottom: Default, OK

3. **Set FASTA Reference:** The sequencing alignment/mapping needs a genome reference. The user can set the genome reference by clicking on this button. A dialog window will appear and the user will be able to select a sequence file (FASTA format, gzip compressed or uncompressed).
4. **Sample Viewer:** The table at the middle of this window is the sample viewer. This table will show all the samples and the respective sequencing files associated with this sample. The user can select by clicking on one or more rows from this table to edit and remove samples. The column “Name” shows the sample unique identification code. The columns “Single Files” and “Paired Files” show the number of files associated with this sample. Note that paired files are only available when data comes from a paired-end sequencing run, so the number of files are represented in half, because there are pairs of files. The

“Status” column shows if the respective sample is ready to be processed, or if there is any kind of error associated with the sample or associated files. If any error is detected, the entire row will be highlighted in red color.

5. **Add Sample(s):** The user can add sequenced sample(s) to the project by clicking at this button. There are three options that appear after clicking this button:
 - 5.1. **Auto-detect:** This option will open a dialog window that the user can select a folder. The folder must contain an entire run, including all the gzipped FASTQ files that are outputted from the Illumina sequencing machines. The files must be named according to the Illumina Naming Convention (See Illumina File Naming Convention Section).
 - 5.2. **Import Sample-sheet:** This option will open a dialog window that requests the user to select a CSV file (comma-separated values) that contains each sample ID and the respective files. More about this file format is described at the Sample-sheet file format section.
 - 5.3. **Manually:** This option will open an “Edit Sample” window. The user can manually define the sample name and the FASTQ associated files to this sample. The full description about the “Edit Sample” window is described below.
6. **Edit Sample:** The edit sample button allows the user to edit sample data such, as name (ID), associated files and other options. When the user selects the “Manually” method to add a new sample, it will be redirected to a sample editor window with empty data, so that the user can define all the data for a single sample.

Name (ID): sample-01

Reset ID

☐ Ignore File Naming Convention

☐ Resequencing

Total FASTQ Files: 4

	R1	R2
L001	sample-01_S01_L001_R1_001.fq.gz	<click to edit>
L002	sample-01_S01_L002_R1_001.fq.gz	<click to edit>
L003	sample-01_S01_L003_R1_001.fq.gz	<click to edit>
L004	sample-01_S01_L004_R1_001.fq.gz	<click to edit>

Set File Remove File + - Close

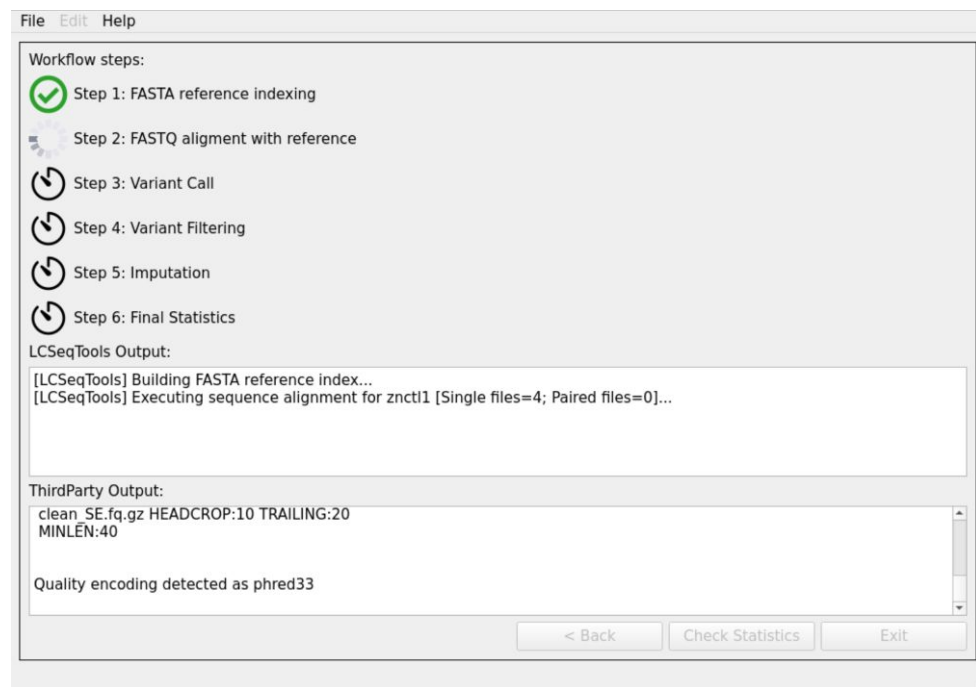
- 6.1. **Name (ID):** A string to define the sample name (i.e. unique identification). The string must contain only characters, numbers, hyphen and underscore. No special characters and white spaces are allowed.

-
- 6.2. **Reset ID:** Reset the sample ID to the original one. The original sample ID comes from the Illumina filename convention.
 - 6.3. **File Viewer:** An interactive table that shows the associated FASTQ files to this sample. There will always be R1 and R2 columns that represent forward and reverse FASTQ files. Single-end sequencing files will have the R2 column empty. The rows represent the number of lanes.
 - 6.4. **Set File:** The user can set each FASTQ file by clicking on the respective row and column from “File Viewer”, then click on “Set File”. Files must follow the Illumina filename convention, otherwise it will be necessary to enable “Ignore”.
 - 6.5. **Remove File:** The user can remove the associated FASTQ file by clicking on the respective row and column from “File Viewer”, then click on “Remove File”.
 - 6.6. **Plus sign:** Adds one more row (Lane) to the “File Viewer”.
 - 6.7. **Minus sign:** Remove one row (Lane) from the “File Viewer”.
 - 6.8. **“Ignore File Naming Convention” and “Resequencing” check boxes:** The user can check the “Ignore File Naming Convention” if the filename of the associated files in the “File Viewer” are different from the Illumina File Naming convention, so the program will not prompt errors about this anymore. Also, the user can check the option “Resequencing”, if there are multiple sample entries with the same ID, so that the program will treat all these data as coming from the same sample.
 - 6.9. **Close:** Apply changes and close the window.
 - 7. **Remove Sample(s):** This button removes all selected samples from the “Sample Viewer” table. This action cannot be undone, hence, if the user mistakenly removes the sample, it will be necessary to redo the “Add Sample” process.
 - 8. **Edit (from the toolbar):**
 - 8.1. **Set Reference Ploidy:** This option is only available if the “Set FASTA Reference” step is done. This will open a new window that shows each sequence in the FASTA reference file in the “Sequence ID” column and the respective Ploidy. the user can edit the reference ploidy. The ploidy editor will look like the image shown below. The user can select one or more rows from the table. After selecting, the options **Mark as Diploid**, **Mark as Haploid** and **Mark as Ignored** will be available. Diploid, haploid and ignored are represented as 2, 1 and 0, respectively. There is also the option to import the IDs from a text file using the import from list options.

Sequence ID	Ploidy
NC_035107.1	2
NC_035108.1	2
NC_035109.1	2
NW_018734407.1	0
NW_018734408.1	0
NW_018734409.1	0
NW_018734410.1	0
NW_018734411.1	0
NW_018734412.1	0
NW_018734413.1	0
NW_018734414.1	0
NW_018734415.1	0
NW_018734416.1	0
NW_018734417.1	0
NW_018734418.1	0
NW_018734419.1	0
NW_018734420.1	0
NW_018734421.1	0

- 8.2. **Mark Selected As Resequencing:** This will check the “Resequencing” from “Edit Samples” section for each selected row from the “Sample Viewer”.
- 8.3. **Unmark Selected As Resequencing:** This will uncheck the “Resequencing” from “Edit Samples” section for each selected row from the “Sample Viewer”.
- 8.4. **Mark Selected As Ignore File Naming Convention:** This will check the “Ignore Unmatching” from “Edit Samples” section for each selected row from the “Sample Viewer”.
- 8.5. **Unmark Selected As Ignore File Naming Convention:** This will uncheck the “Ignore Unmatching” from “Edit Samples” section for each selected row from the “Sample Viewer”.
9. **Analyze** (from the toolbar):
 - 9.1. **Run FastQC:** This will generate a sequencing quality summary by FastQC using the FastQ files from the selected rows in the “Sample Viewer”.
10. **Total samples:** This box shows the total samples from this project. Note that this number does not necessarily reflect the number of rows from the “Sample Viewer” because the resequencing option could be enabled in some cases.
11. **Start:** This button will execute the workflow. Note that this button will only be available if every step is correctly configured, otherwise it will be grayed until everything is correct.

Workflow Progress Window



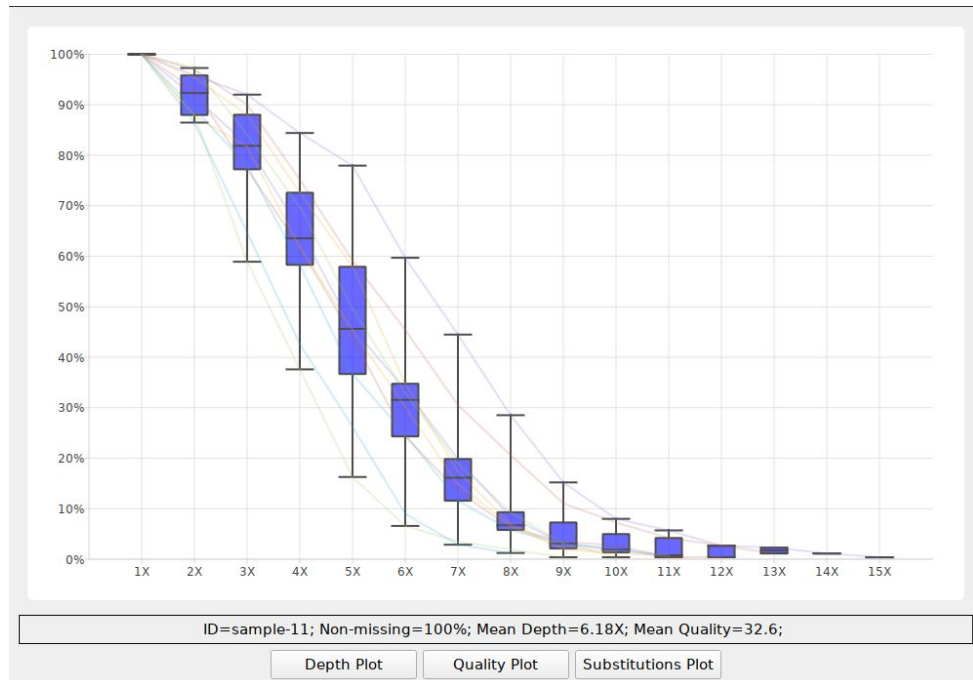
After the **Start** button is clicked, the Project Configuration Window will change to the Workflow Progress Window. Each workflow step status is indicated with the following icons:

- 🕒: Task is waiting.
- 🔄: Task is running.
- ✅: Task is finished.
- ❌: Task failed and workflow aborted.

The **LCSeqTools Output** box will output messages with deeper details about the workflow steps status, warnings, etc. The **ThirdParty Output** box will redirect messages from the bioinformatics programs that run in background during the steps. After all steps are finished, the **Check Statistics** and **Exit** buttons will be available.

Statistics Viewer Window

Clicking on **Check Statistics** button available after all step finished will open a window similar as shown below:





















There are three different plots that can be visualized by clicking on the respective button at the bottom of this window. Plots are interactive, so the user can hover the mouse over the lines to get detailed information that will be displayed in the box over the three buttons.

- **Depth Plot:** The genome coverage distribution within samples for each coverage depth level. The X-axis represents the depth level and Y-axis represents the genome coverage at least at the given depth level. Each line represents a single sample.
- **Quality Plot:** The genotype quality distribution within samples for each genotype quality score in phred score (Genotyping Error Probability $\leq 10^{-GQ/10}$). The X-axis represents the quality score and Y-axis represents the percentage of genotype qualities for the given score. Each line represents a single sample.
- **Substitution Plot:** Shows the distribution of transitions and transversions in the variant set and the respective substitutions.

Output Files

The output files are stored in the **Output Folder** location. The resulting files are stored in four separated folders: Parameters, Reference, Alignment and VariantCall. A hidden temporary folder

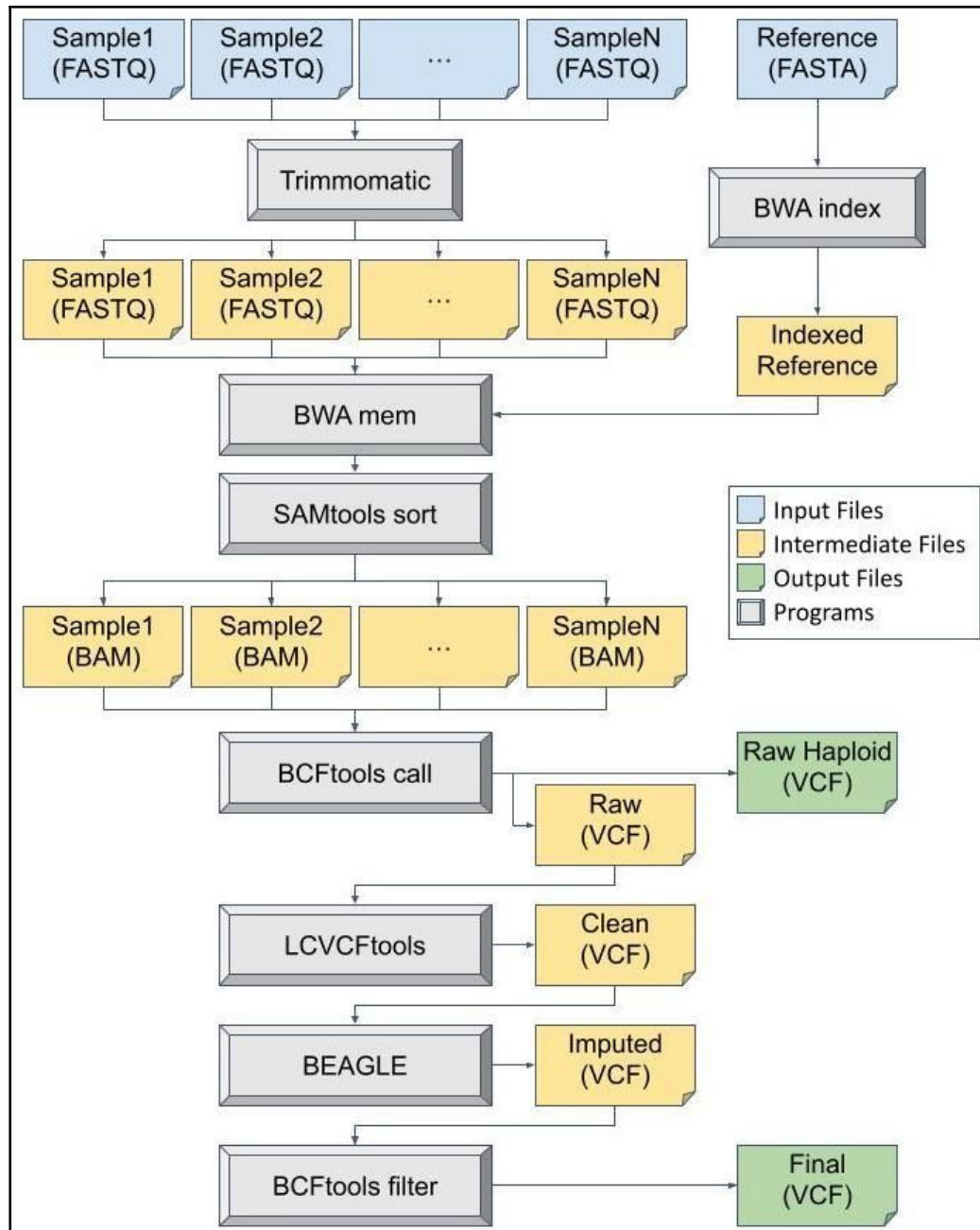
“.temp” is created and deleted during the workflow. The following list describes each folder and file within folders.

-  **Parameters**
 -  **project.cfg**: A bash script that contains some workflow variables.
 -  **haploid.list**: A text file with haploid sequence IDs from the reference file.
 -  **ignored.list**: A text file with ignored sequence IDs from the reference file.
-  **Reference**
 -  **ref.fa**: The uncompressed FASTA reference file. Files with ref.fa prefix are reference index files.
 - ...
-  **Alignment**
 -  **sample-01**: A BAM file with mapped-only sequences for the sample-01 (or name defined by the user).
 -  **sample-01.bai**: An index file for sample-01 BAM file.
 - ...
-  **VariantCall**
 -  **1-raw.vcf.gz**: The first resulting variant calling file. This file contains all variants without any filtering.
 -  **1-raw-haploid.vcf.gz**: If there are haploid sequences, this file contains the resulting variant calling file for haploid sequences. This file contains all variants without any filtering.
 -  **2-clean.vcf.gz**: The resulting VCF file after applying data filtering for 1-raw.vcf.gz.
 -  **2-clean.removed**: A text file that contains the sample IDs removed from the 1-raw.vcf.gz. This file may be empty if no sample was removed.
 -  **3-imputed.vcf.gz**: The resulting VCF file after applying imputation for 2-clean.vcf.gz.
 -  **3-imputed.log**: A log file from BEAGLE imputation software.
 -  **4-final.vcf.gz**: The final VCF ready for statistical analysis. This is the resulting VCF file after applying imputation filtering parameters for 3-imputed.vcf.gz.
 -  **stats1.tsv**: A LCVCFtools stats file generated from 2-clean.vcf.gz. This file contains data used for plotting stats in the **Statistics Viewer Window**.

-
- **stats1.tsv.old**: If this file exists, it contains a LCVCFtools stats file generated from 2-clean.vcf.gz before removing samples in the 2-clean.removed list.
 - **stats2.tsv**: A BCFtools stats file generated from the 4-final.vcf.gz. This file contains data used for plotting stats in the **Statistics Viewer Window**.

Other information

LCSeqTools Workflow Diagram



Illumina File Naming Convention

The Illumina FASTQ File Naming Convention ([link](#)) is described as the following example. Each field is separated by an underscore character. The auto-detect function in the **Project Configuration Window > Add Sample(s) > Auto-detect** works only when the files in the selected directory follow this naming convention. Supported file extensions are: .fastq.gz and fq.gz.

sample01_S1_L001_R1_001.fastq.gz

- **sample01**: The sample name provided in the sample sheet. If a sample name is not provided, the file name includes the sample ID, which is a required field in the sample sheet and must be unique.
- **S1**: The sample number based on the order that samples are listed in the sample sheet starting with 1. In this example, S1 indicates that this sample is the first sample listed in the sample sheet.
- **L001**: The lane number.
- **R1**: The read. In this example, R1 means Read 1. For a paired-end run, there is at least one file with R2 in the file name for Read 2. When generated, index reads are I1 or I2.
- **001**: The last segment is always 001.

Genotyping Parameters

- **Sequence trimming (Head):** Removes the specified number of bases, regardless of quality, from the beginning of the read.
- **Sequence trimming (Trailing Crop Quality):** Remove low quality bases from the end. As long as a base has a value below this threshold the base is removed and the next base will be investigated.
- **Sequence minimum length:** The minimum length of the resulting sequence after trimming.
- **Minimum Genotype Quality (Phred Scale):** Omit all genotypes with a quality below the specified threshold. Data from omitted genotypes will be used for genotype imputation.
- **Minimum Depth:** Omit all genotypes with a quality below the threshold specified. Data from omitted genotypes will be used for genotype imputation.
- **Minor Allele Frequency Threshold:** Include only variants with a Minor Allele Frequency greater than or equal to the threshold. Allele frequency is estimated based on allele depth from the genotypes, including omitted genotypes.
- **Max missing data (Variant):** Remove all variants with missing data (i.e. depth equal zero) rate greater than or equal to the threshold.
- **Max missing data (Sample):** Remove samples with missing data (i.e. depth equal zero) rate greater than or equal to the threshold.
- **Imputation Seed:** Random number generator. Repeating an analysis with the same "Imputation Seed" and "Max threads" values will produce the same imputed genotypes.
- **Minimum Genotype Probability:** Remove all genotypes with a posteriori probability below the threshold specified.

Sample-sheet file format

The import datasheet function in the **Project Configuration Window > Add Sample(s) > Import Sample-sheet** can import a CSV (comma-separated values) table with custom configuration. For example, using the Sample-sheet file, the user can define custom sample names (IDs) and associate FASTQ files that do not follow the Illumina File Naming Convention. The Sample-sheet table must be formatted as follows:

- A header line with the column names as: id, R1_L001, R2_L001, R1_L002,
- The mate-lane identifiers order in the header does not matter, since the program reads the identifier format as Ri_L00j, where i and j represent the mate and lane numbers, respectively.
- One sample per line. Sample name (ID) must not contain special characters and whitespaces. Each mate-lane filename must be in the correct column, following the header identifiers order. The filenames must provide the relative path from the Sample-sheet file location to the FASTQ file location.

The following table is a hypothetical example of a Sample-sheet file. In this example, the hypothetical folder contains both the Sample-sheet file and another folder called “fastq”, where all the FASTQ files are stored. Note that all the FASTQ files provide the relative path from the Sample-sheet file to the FASTQ file, as we can see the “fastq/” prefix. Note that the provided FASTQ files in this example do not follow the Illumina File Naming Convention, so the user will need to select all these samples in the **Sample Viewer** and click on **Mark Selected As Ignore File Naming Convention**.

id	R1_L001	R2_L001
Sample1	fastq/sample01-R1.fq.gz	fastq/sample01-R2.fq.gz
Sample2	fastq/sample02-R1.fq.gz	fastq/sample02-R2.fq.gz
Sample3	fastq/sample03-R1.fq.gz	fastq/sample03-R2.fq.gz
...

The Third-Party Package

LCSeqTools uses a free and open-sourced collection of third-party softwares to execute the genotyping workflow. The LCSeqTools container image file is distributed with the following third-party bioinformatics tools:

- FastQC. (ANDREWS, S., 2010). Software under GPLv2.
- Trimmomatic. (BOLGER, A. & GIORGI, F., 2009). Software under GPLv3.
- Burrows–Wheeler Aligner - BWA. (LI, H. & DURBIN, R., 2009). Software under GPLv3.
- SamTools. (LI, H., 2011). Software under MIT License (Expat).
- Beagle 4.1. (BROWNING, B.L. & BROWNING, S.R., 2016). Software under GPLv3.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

LI, Heng; DURBIN, Richard. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, v. 25, n. 14, p. 1754-1760, 2009.

LI, Heng. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, v. 27, n. 21, p. 2987-2993, 2011.

Browning, BL; Browning, SR. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet.* 2016.98:116–26.

BOLGER, Anthony. GIORGI, F. Trimmomatic: A flexible read trimming tool for Illumina NGS data. *Bioinformatics*, v. 30, n. 15, p. 2114-2120, 2014.