SeQCoS Application User Manual

Last updated: 2012-09-01

Contents

Introduction

The emergence of massively parallel (or next-generation) sequencing technologies has revolutionized the field of genomics and other related areas. It has enabled the rapid and accurate sequencing of whole genomes at an affordable cost to researchers. Data outputted from a massively parallel sequencing experiment requires extensive bioinformatics analysis. Quality control (QC) of sequencing reads is a preliminary step routinely employed by labs to ensure that only high quality data is used for downstream steps. For example, experimentation has shown that while performing a *de novo* assembly of a genome, a reduced data set of high quality reads will produce better assemblies than a higher volume data set of lower quality.

The QC of sequencing reads may range from simple, manual filtering procedures to comprehensive, automated solutions. Here, we present Sequence Quality Control Studio (SeQCoS), a software suite designed to perform QC of massively parallel sequencing reads. Given a sequence data file, the software generates a series of standard plots illustrating the quality of reads at both sequence and quality score (if available) levels. We also provide additional tools including the ability to perform a BLAST of reads against a database, and basic post-QC filtering tools such as trimming and discarding of reads.

SeQCoS was written in C# using the Microsoft .NET Framework, using the .NET Bio bioinformatics toolkit and Sho, a data analysis and visualization application.

The following document discusses how to get started with SeQCoS and provides an overview of the SeQCoS application suite. This user manual assumes the reader is comfortable working on a Windows operating system. A user interested in using the command line-based version of SeQCoS should be comfortable working in the MS-DOS environment.

Formats Supported

Currently, the input and output formats supported by SeQCoS are limited to FASTA and FASTQ. For FASTQ, the following versions are supported:

- Sanger/Illumina 1.8+ (Phred+33)
- Illumina 1.3+ (Phred+64)
- Solexa (Solexa+64)

Requirements

Software dependencies:

- Windows 7 or newer (older versions of Windows should work but has not been tested)
- Microsoft .NET Framework 4.0
- Sho 2.0.5 or higher (SeQCoS was tested on 2.0.5)

- Standalone NCBI Blast for Windows (blast+) (required for executing BLAST, otherwise it is optional)

Hardware:

- 4 GB of RAM or more (the more RAM available, the larger the input file that can be handled)
- If multiple processor cores are available, certain processes will be parallelized to improve compute time

Quick Start

Read this section if you wish to skip the details and start using SeQCoS right away. Otherwise continue to the next section.

1. Install required software dependencies and SeQCoS setup.
2. Launch the SeQCoS GUI application.
3. Click on **File** followed by **Start a New Run**.

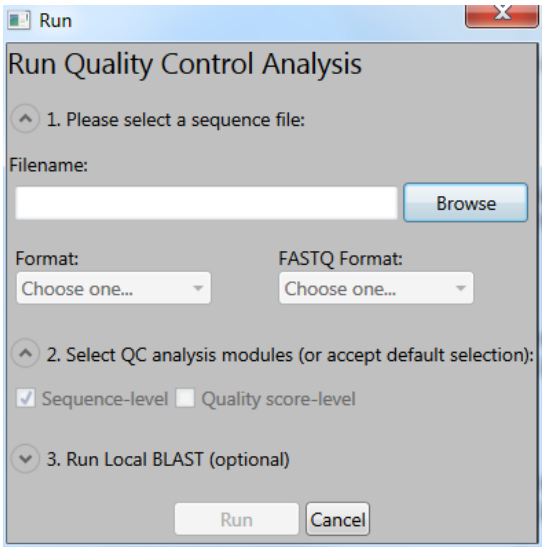


Figure 1 Start a new QC run

4. In the new dialog window that opens, click **Browse** to select a sequence file.
5. SeQCoS will attempt to automatically detect the format of the sequence file based on the filename extension. A prompt will appear if it is unable to do so.
6. Choose from the pull down the corresponding FASTQ format, if applicable. (Available options: Illumina 1.3+ [Phred+64], Solexa [Solexa+64], Sanger/Illumina 1.8+ [Phred+33])
7. Next, select the QC analysis modules you wish to use or accept the default selection. Quality score-level QC analysis is disabled if the input file is in FASTA format (since no quality scores are available).
8. Click **Run**.
9. Upon completion of the analysis, SeQCoS will populate the GUI main window with the results.

A copy of the results (JPEG images and basic statistics) is saved in the same directory as the input file under a newly created subdirectory named after the input file (e.g. if the input filename is "e_coli_1.fastq", then the new subdirectory is named "e_coli_1").

10. To perform post-QC filtering of the data, click on **Tools**, select either **Trim Reads** or **Discard Reads**, and finally choose a filtering method (**By Read Length**, **By Quality Score**, or **By Regular Expression**).
11. Fill in the required parameters, including input and output files.
12. Click **Run**.

Getting Started

The following section provides instructions for setting up SeQCoS on your PC.

Note: In the instructions, the keyword $(ProgramFiles) is used to refer to the location of your **Program Files** directory (e.g. C:\Program Files).

1. Installing SeQCoS

It is highly recommended to install any missing software dependencies first. You will need to agree to the terms and conditions of the usage agreements set forth by those applications, where applicable. To install SeQCoS, simply run the MSI setup file and follow the onscreen instructions.

2. Setting up NCBI BLAST for Windows

NCBI BLAST is a sequence alignment program used to search for regions of similarity between biological sequences. In order to be able to execute BLAST from SeQCoS, the blast+ software package must be first installed. Please refer to this NCBI user manual for instructions.

Configuring a BLAST database

A BLAST-formatted database is required for BLAST to search for sequence similarity. The user is free to provide any type of sequence database he/she wishes to use. A BLAST-formatted database of NCBI UniVec (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) is available for download from the Codeplex project website. **All BLAST-formatted databases should be saved under the BLASTDB (e.g. $(ProgramFiles)\NCBI\blast-2.2.25+\db) directory.**

If you wish to create your own BLAST database, the following is a step-by-step guide using UniVec as an example.

1. Ensure that blast+ has been properly installed.
2. Create or download a FASTA file containing the reference sequences you wish to search against. In this example, we obtain the UniVec FASTA reference file from the NCBI FTP (ftp://ftp.ncbi.nih.gov/pub/UniVec/). Here, there are two versions: UniVec and UniVec_Core. The latter being a subset of the former.

Figure 2 Screenshot of the NCBI FTP site

3. Right-click on the file (e.g. UniVec) and select **Save target as** to save the file on your local system. Since our ultimate destination is the BLASTDB directory, we will save this file there. As the original filename does not have an extension, we will add one (e.g. .fa) for clarity:

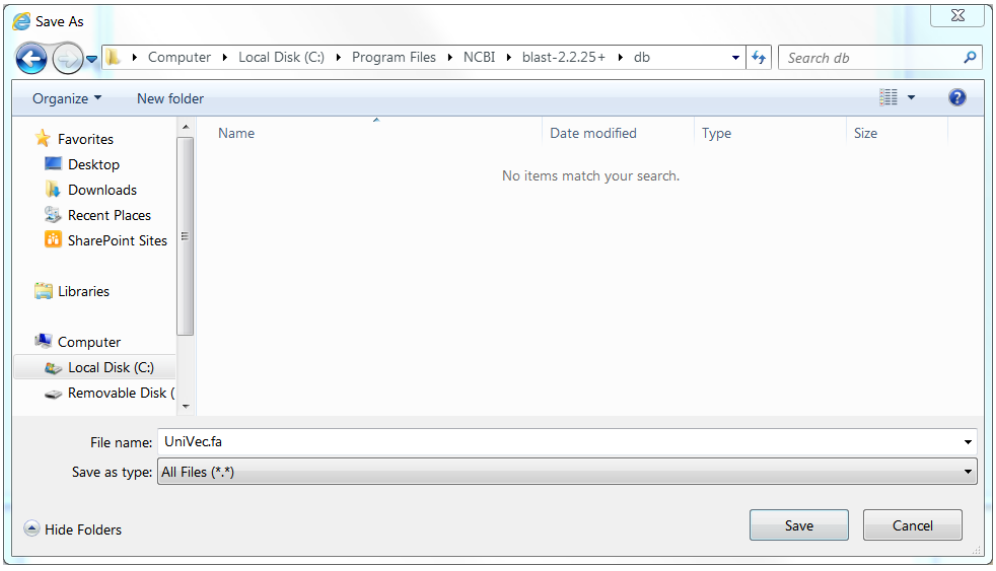


Figure 3 Save As dialog window. Note the addition of a file extension and the directory location.

4. Now we are ready to format this FASTA file to a BLAST database. To do this, we will use the command line tool `makeblastdb.exe` supplied by blast+. First, open a MS-DOS console window by clicking **Start** and choosing **Run** (or use the keyboard shortcut Windows Key + R).

5. In the new dialog window that opens, type in the field next to Open: enter “cmd”, as shown below. Then click **OK**.

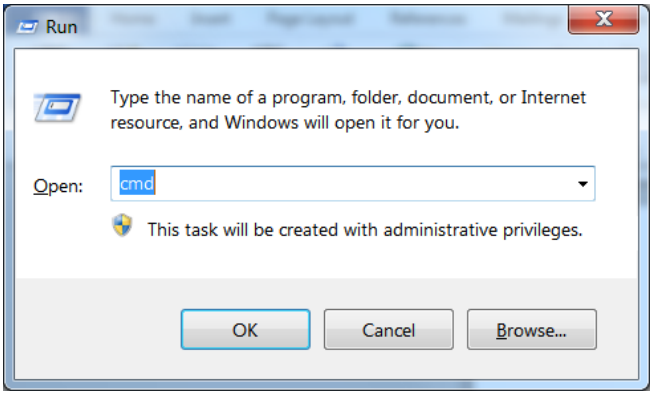


Figure 4 Run dialog window

6. In the console window that opens, navigate to the directory where the FASTA file was saved. For this guide, we will change directories to **`C:\Program Files\NCBI\blast-2.2.25+\db`**. Type the command “**cd**” (for change directory) followed by a space and the full path of the directory you want to go to (see Figure 5). Then press **Enter**.

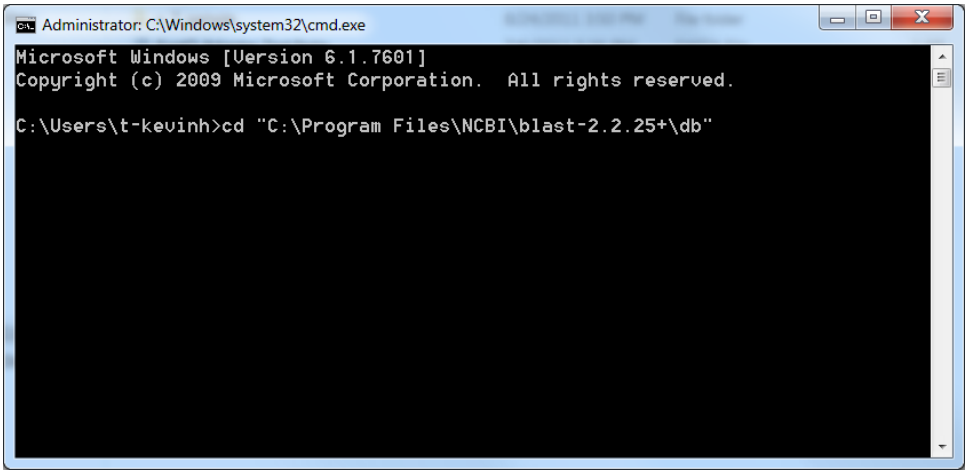


Figure 5 Changing directories in the MS-DOS console

7. Next, we will run the aforementioned **`makeblastdb.exe`** to format the FASTA file. Type the following command (underlined), followed by Enter:

```
C:\Program Files\NCBI\blast-2.2.25+\db> makeblastdb.exe -in UniVec.fa -out UniVec -
                                        dbtype nucl
```

   This executes the program `makeblastdb`.exe, sets the input file (-in) “UniVec.fa”, sets the output database name (-out) to be “UniVec”, and sets the database type (-dbtype) to “nucl” (for nucleotides).

8. A confirmation message with various statistics should be outputted on the console. **Three** new files should be generated:

i. UniVec.nhr (header file)
ii. UniVec.nin (index file)
iii. UniVec.nsq (sequence file)

You can confirm this by opening this directory on Windows Explorer, or directly in this console by entering the command "**dir**" to list all files in this directory.

9. You may now close the console window. You have now completed formatting a FASTA file to a BLAST database!

3. Configuration of Environment Variables

SeQCoS relies on the environment variables **PATH** and **BLASTDB** to be properly configured in order to run blast+. Under normal circumstances, these variables have been automatically configured during installation of each program. If not, they must be manually entered as follows:

Variable	Value
Path	$(ProgramFiles)\NCBI\blast-2.2.25+\bin; …(other values)…
BLASTDB	$(ProgramFiles)\NCBI\blast-2.2.25+\db
SHODIR	$(ProgramFiles)\Sho 2.0 for .NET 4\

To manage environment variables on Windows:

On Windows XP, go to **Start**, **Control Panel**, **Performance and Maintenance**, **System** (or right-click on **My Computer** and choose **Properties**). In the new dialog window that opens, click on the **Advanced** tab and click on the button **Environment Variables**.

On Windows Vista/7, go to **Start**, **Control Panel**, **System and Security**, click on the **System** heading, and then click on the link **Advanced system settings** located on the left panel bar. In the new dialog window that opens, click on the **Advanced** tab and click on the button **Environment Variables**.

Running the Application

Graphic User Interface (GUI)

The SeQCoS application can be run in GUI mode. Simply start the SeQCoS GUI application from your Program menu or execute `SeqcosGui.exe`.

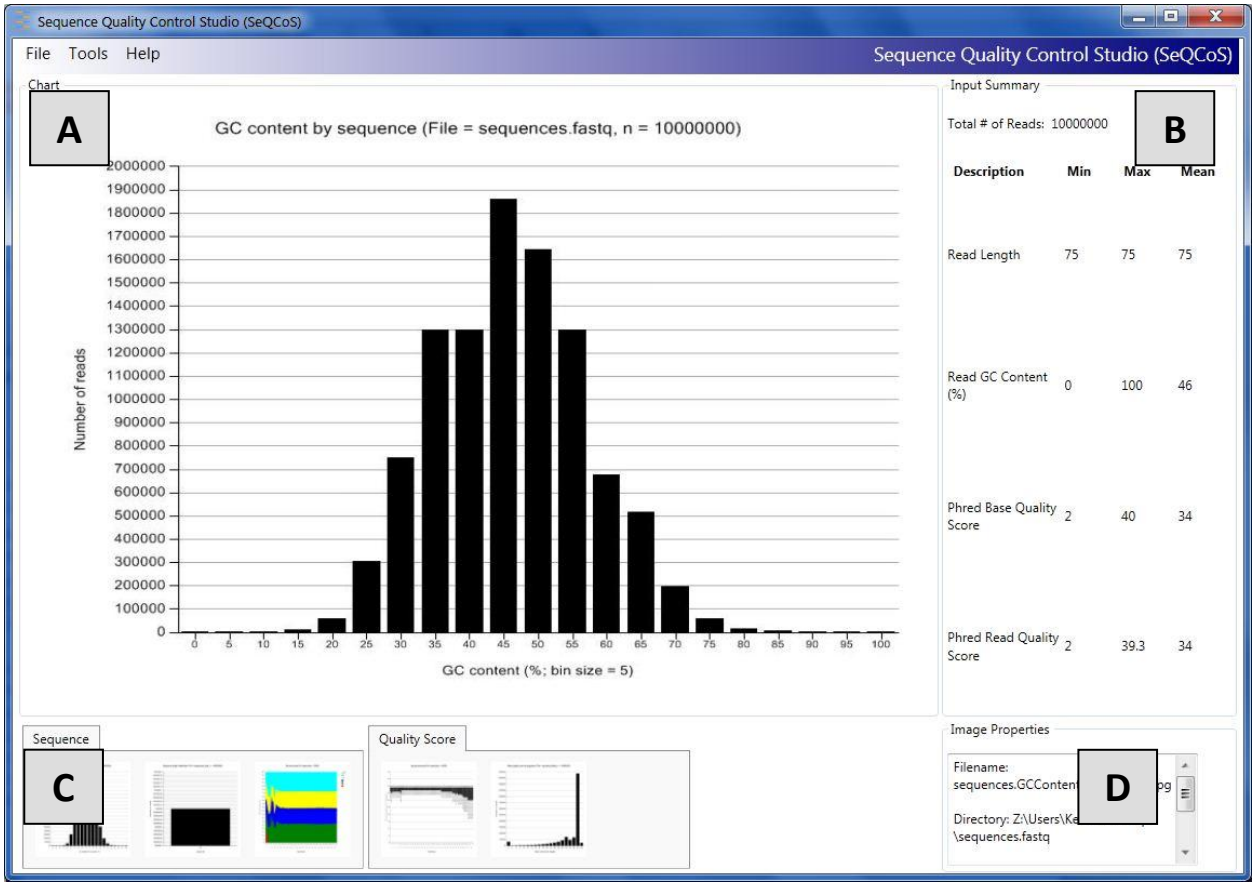


Figure 6 SeQCoS GUI main window

The SeQCoS GUI main window (Figure 6) is made up of 4 sections:

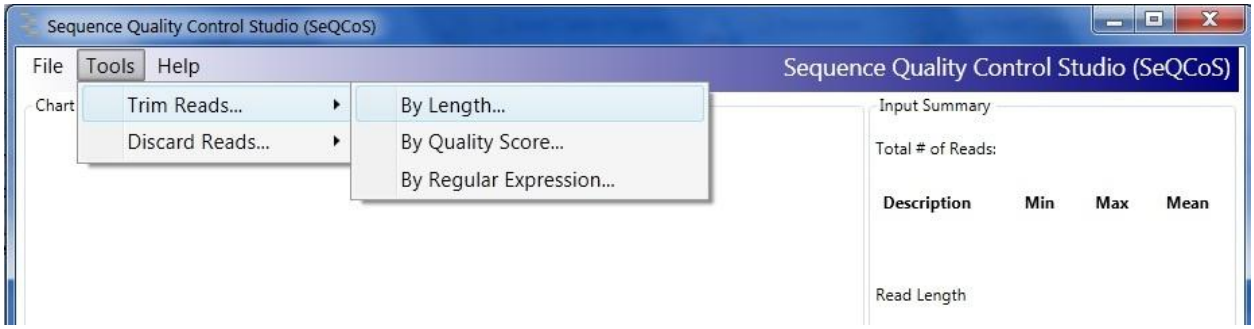
A) Main chart area that displays the selected plot
B) Thumbnail gallery of available plots to view. Click on one of the thumbnails to display the plot in main chart area (A).
C) Displays a summary of statistics about the input data
D) Displays basic image properties about the selected plot shown in A

Start a QC Run

See the **Quick Start** section on page 3 for instructions on starting a QC analysis run.

Using Sequence Trimming/Discarding Tools

Trimming and discarding tools can be accessed from the **Tools** menu, as shown below:



Both trimming and discarding methods can be applied to sequence data based on three properties: **By Length**, **By Quality Score** and **By Regular Expression**. Select one of the options to open the tools dialog.

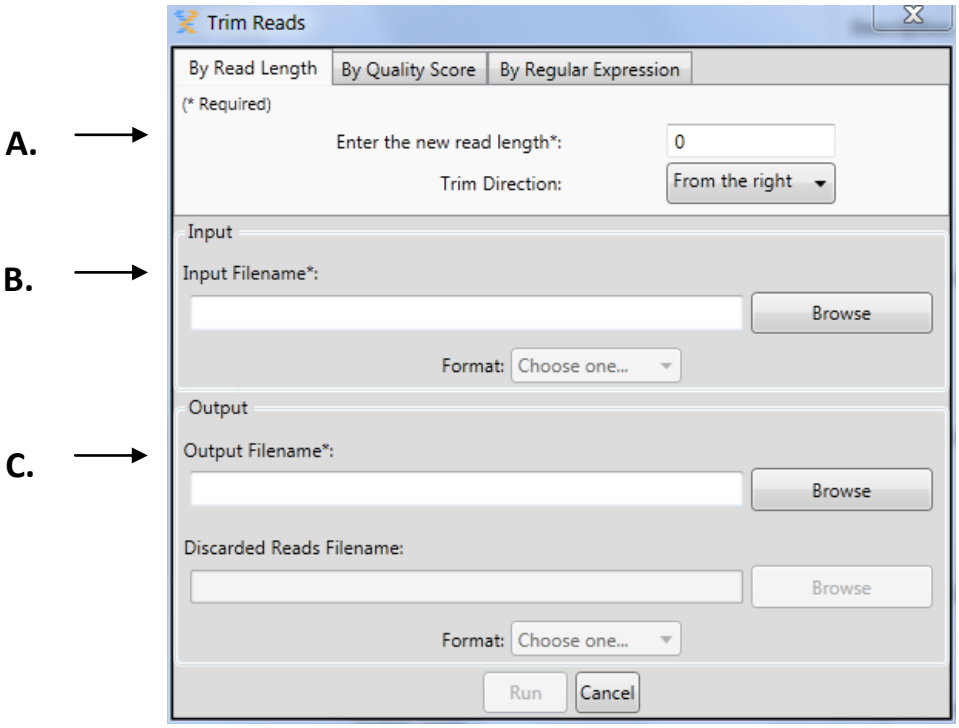


Figure 7 Trim By Read Length dialog

By default, the dialog corresponding to the option that was selected will appear; however, it is possible to switch to another option using the tabbed navigation panel at the top of the window. Fields denoted by an asterisk (*) are required.

A. Trim/Discard options

- Enter/select the required options for the trimming/discarding. More details about each mode and options are described below (see section **SeQCoS Components → Trimming/Discarding Reads**).

B. Selecting input files

- Under the "**Input**" section, click on Browse to select an input file. The format will be automatically detected and appear in the Format pulldown menu. If detection fails, you will be prompted to select the correct format.

C. Selecting output files

- Under the "**Output**" section, select a file for outputting the results in the Output Filename field. The Discarded Reads Filename field is optional and will save any reads that were discarded from the analysis. This is usually applicable when Discard mode is used; however, it may be possible that some reads are removed in Trim mode. For example, if Trim By Quality is invoked and there exist reads where entire sequence is below the user-specified quality threshold, then such reads will be discarded entirely.

Load an existing run

Previously executed SeQCoS runs can be loaded on the GUI. **Please ensure that the filenames in the folder have not been changed or it will not be properly recognized by the application**. From the menu, go to File -> Load an existing run directory. Browse for the folder containing the run and select OK.

Console Application

The SeQCoS application can be run in command-line mode. It may be desirable to add the SeQCoS program folder in your Path environment variable (see above section to learn how to manage environment variables [page 7]).

The following command-line applications are available:

Executable	Description
`SeqcosApp.exe`	Performs the QC analysis of a sequence file and outputs a series of plots as JPEG files.
`SeqcosTrimmerUtil.exe`	Trim reads in a sequence file.
`SeqcosDiscarderUtil.exe`	Discard reads in a sequence file.

A full description of the available options for each exe can be found using the option **"/help**" (e.g. `SeqcosApp.exe /help`).

SeQCoS Components

This section describes the application features.

Plots

The plots generated by SeQCoS can be divided into two groups:

- Sequence level – based on measurements on entire read sequences
    - GC content – histogram of GC content calculated on each read
    - Quality scores – the mean quality scores are calculated for each read

- Sequence lengths – histogram of sequence lengths, although not particularly useful for inputs with constant read length

- Base position level – based on measurements at individual base positions
  - Quality Scores – the distribution of quality scores at position *i* (where *i* is between 1 and the maximum read length) is calculated and summarized as multiple boxplots
  - Symbol Count – distribution of nucleotides at position *i*

Invoking BLAST

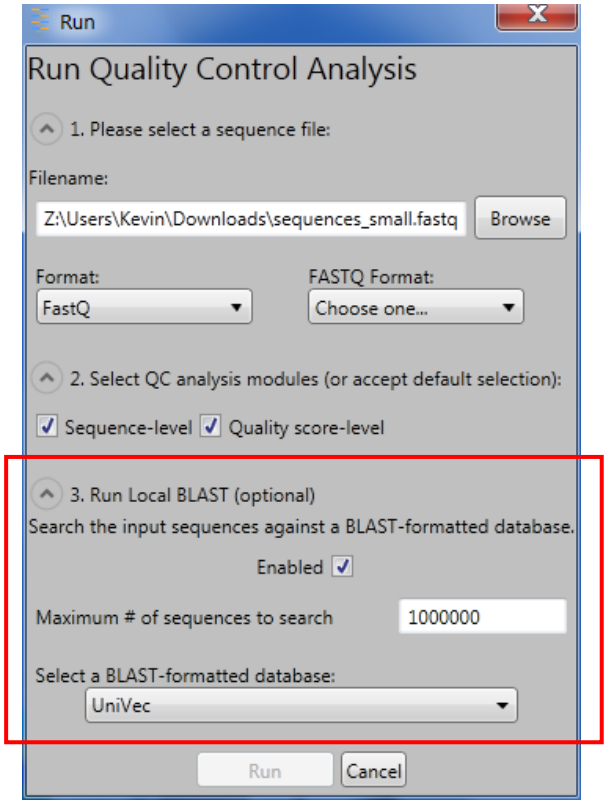


Figure 8 Enabling BLAST analysis

This option will invoke BLAST and perform a sequence similarity search against a user-specified database. Users may wish to restrict the number of reads to be searched against the database in order to avoid long compute times.

The BLAST report will be generated in three formats:

1. ASN (e.g. sequence_file.blast.asn)
2. CSV (e.g. sequence_file.blast.csv)
3. XML (e.g. sequence_file.blast.xml)

Trimming/Discarding Reads

SeQCoS provides two sequence filtering modes: trimming and discarding reads. There are three available options for executing these modes:

By read length:

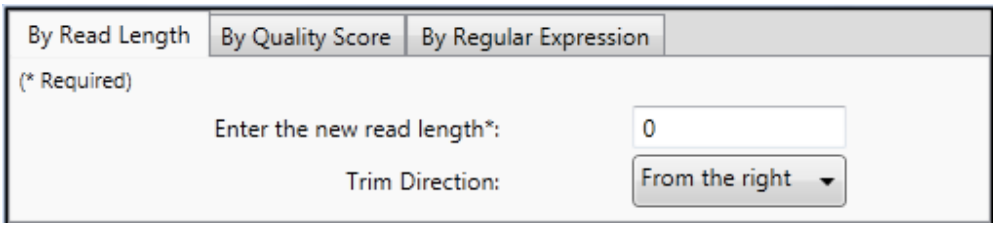


Figure 9 Trim By Read Length options

Mode	Description
Trim	• Reads will be trimmed to the user-specified minimum read length *L*. If reads have a length that is already less than *L*, then it will be discarded from the output. • Reads can be trimmed from either side of the read. By default, trimming will take place from the right
Discard	• Reads with a length less than the user-specified minimum read length *L* will be discarded from the output.

By base quality score

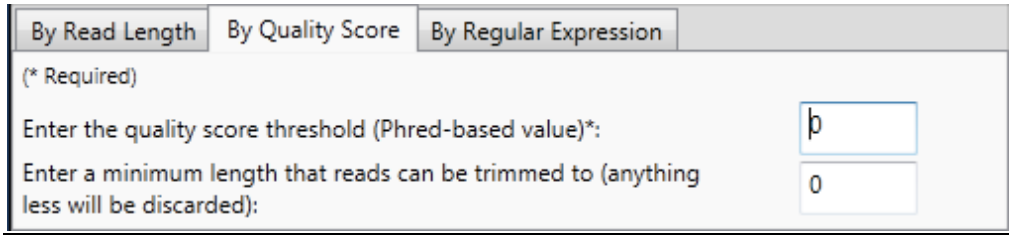


Figure 10 Trim By Quality Score options

Mode	Description
Trim	• Reads will be trimmed according to the user-specified minimum Phred-based quality threshold. An implementation of the maximum subarray problem (http://en.wikipedia.org/wiki/Maximum_subarray_problem) is used, but modified to keep track of the start and endpoints. • The user may wish to specify a minimum read length *L* of the trimmed reads. Reads that are trimmed to a size less than *L* will be discarded. By default, a value of 0 indicates that this option is disabled.
Discard	• Reads will be discarded based on a user-specified mean quality score threshold *Q*.

By regular expression

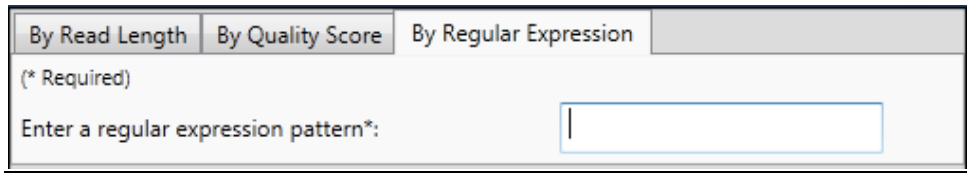


Figure 11 Trim By Regular Expression options

Mode	Description
Trim	• A user-specified regular expression pattern is used to find all matches (via the Regex.Matches() function) against all reads. Any matching segment will be stripped from the read. • Example 1: Input: `AAATGTCGAGTGAGTGTGAAAC` Regex Pattern: `AAA` Match result: `AAA`TGTCGAGTGAGTGTG`AAA`C Trim result: `TGTCGAGTGAGTGTGC` • Example 2: Input: `AAATGTCGAGTGAGTGTGAAAC` Regex Pattern: `^AAA` (matches only at the beginning of the input string) Match result: `AAA`TGTCGAGTGAGTGTGAAAC Trim result: `TGTCGAGTGAGTGTGAAAC`
Discard	(as above, but instead reads will be discarded from the output)

Uninstalling the Application

SeQCoS can be uninstalled in the same fashion as any other standard Windows application:

- On Windows XP, go to **Start**, **Control Panel** and click on **Add or Remove Programs**. Under the **Currently Installed Programs** box, select the SeQCoS program and click **Change/Remove**.
- On Windows Vista/7, go to **Start**, **Control Panel**, **Programs** and click on **Uninstall a program** (or if using the icon view, click on **Programs and Features**). Select the SeQCoS program and click **Uninstall**.

Alternatively, if you still have the MSI setup file. Open it and follow the onscreen instructions for removing the program.

FAQ

Q: When running the QC analysis, I am getting the error message: “DoQcAnalysis Exception: Could not load file or assembly ‘ShoViz, Version=2.0.5.0, Culture=neutral, PublicKeyToken=14eb30934789ddca’ or one of its dependencies. The system cannot find the file specified”.

A: See section **Configuring Sho library paths** on page 15 for instructions on how to update the ShoViz path in the configuration files.

Development

SeQCoS is an open source software package licensed under Apache 2.0. The source code is freely available from the Codeplex project website (http://seqcos.codeplex.com).

A copy of Microsoft Visual Studio 2010 is required for development and is available from a variety of avenues:

1. Purchase a retail version that suits your requirements
2. If you are a student or educator, Visual Studio is free through the Microsoft Dreamspark program
3. Through the MSDN Academic Alliance (if your academic department is a member)
4. Visual Studio Express is free but requires product registration

The .NET Bio Version 1.0 library (Bio.dll) is distributed along with SeQCoS and may be used to development. If you downloaded the SeQCoS source code directly and did not install SeQCoS, then you will need to install .NET Bio to obtain the latest version of the library. However, beware of changes to the library that may affect existing functionality of SeQCoS.

The Sho libraries can be found in the "bin" directory of the Sho install (e.g. C:\Program Files\Sho 2.0 for .NET 4\bin).

The SeQCoS source code directory structure is described as follows:

Build -
A central location to store all of the binaries generated by each project.

Docs -
All documentation files for the project are stored here, including this user manual.

Source -
All source files that contribute to the released binaries are stored here.

Source\Application -
All program source files that contribute to the core functionality of the application.

Source\Tools -
All source files of programs that use the API of the main project, such as command-line tools, UI, and installer.

Tests -
Various unit test programs to validate the functions implemented in the project.

Appendix

Configuring Sho library paths

SeQCoS uses libraries that were provided in the Sho installation. In the current setup, the paths of the libraries are saved in the configuration files SeqcosGui.exe.config and SeqcosUtil.exe.config (found in the SeQCoS application folder). By default they have been set to the path "C:\Program Files\Sho 2.0 for .NET 4\bin\" (see Figure 14 below). If Sho was installed in a different location (e.g. installed in a custom path or using a 64-bit OS, where the default path begins with "C:\Program Files (x86)"), then SeQCoS will not be able to locate the Sho libraries and subsequently crash. In version 1.0.0, a method for handling the event AppDomain.AssemblyResolve was added to deal with unresolved assemblies at runtime. **Thus, no action is specifically required by the user to handle failed assemblies.**

Alternatively, the configuration files can be explicated updated to the correct paths. This can be done manually or preferably by running the supplied utility program **ConfigUpdaterUtil.exe**. This change only needs to be done once. To do this:

1. Sho must be already installed.
2. Navigate to the SeQCoS application folder (i.e. "C:\Program Files\Sequence Quality Control Studio").
3. Right click on ConfigUpdaterUtil.exe and select "Run as Administrator" (required to update the config files). The application will scan the current directory for *.exe.config files and look for lines that match a Sho-related library path. If a match is found, the path to the DLL will be replaced by value of the environment variable SHODIR.

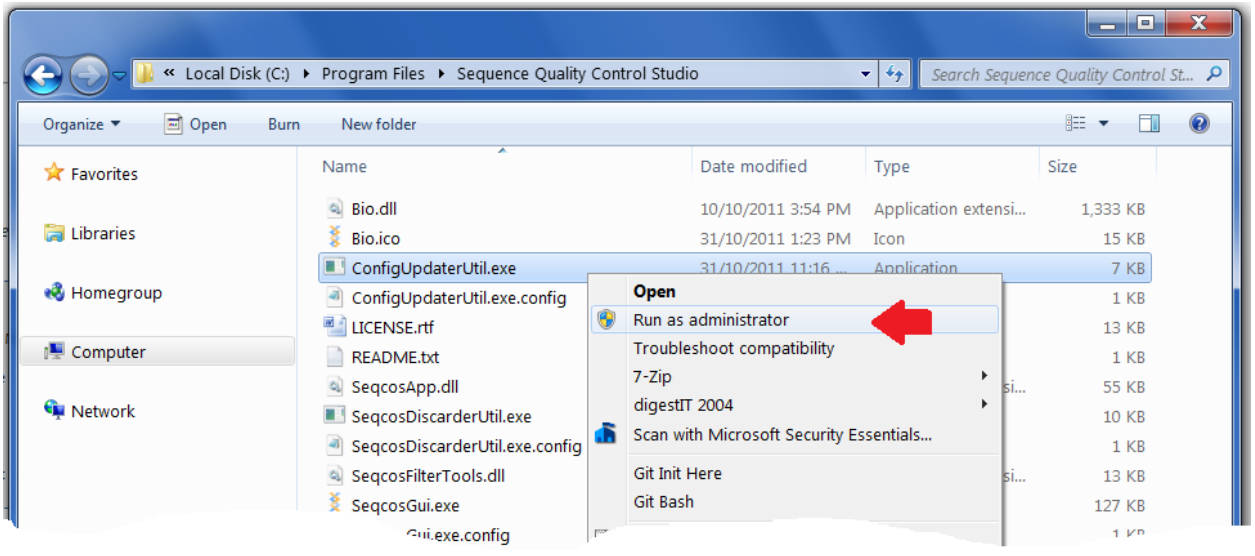


Figure 12 Run ConfigUpdaterUtil.exe with administrator privileges

4. A console window will open and display the results from the update. Only the config files that contain the paths to Sho-related DLLs will be updated.

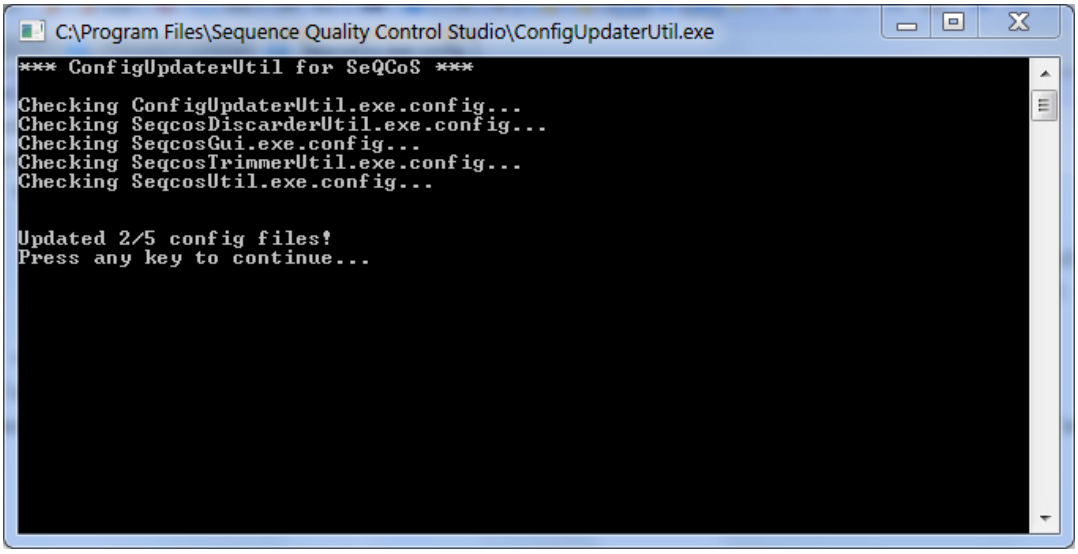


Figure 13 Running the ConfigUpdaterUtil.exe to update the *.exe.config files

To manually update the config files:

1. Locate the files SeqcosGui.exe.config and SeqcosUtil.exe.config in the SeQCoS application folder.
2. **Open** each file in Visual Studio or a text editor e.g. Notepad (administrative privileges are required to make the edits).
3. In the "**codebase**" XML element, update the "**href**" property to the correct path of your ShoViz.dll. For example, if you installed Sho in the path "C:\Tools\Sho 2.0 for .NET 4". Then, the property should be changed to "file://C:\Tools\Sho\bin\ShoViz.dll" (see screenshot below).
4. **Save** the file.

Figure 14 Editing the SeqcosGui.exe.config file