MGI 华大智造

To Enable Effective and Affordable Healthcare Solutions for All

MegaBOLT Bioinformatics Analysis Accelerator

Advanced User Manual

Address: Main Building and Second floor of No.11 Building, Beishan Industrial Zone, Yantian District,
Shenzhen, 518083, Guangdong, China
E-mail: MGI-service@genomics.cn
Website: www.mgitech.cn

Research Use Only

MGI Tech Co., Ltd.

Edition
1.0

About this user manual

This user manual is applicable to MegaBOLT Bioinformatics Analysis Accelerator (MegaBOLT). The edition is 1.0 and the software version is V2.1.0.

This manual and the information contained within are proprietary to MGI Tech Co., Ltd. (hereinafter called MGI), and are intended solely for the contractual use of its customer in connection with the use of the product described herein and for no other purpose. Any person or organization can not entirely or partially reprint, copy, revise, distribute or disclose to others the manual without the prior written consent of MGI. Any unauthorized person should not use this manual.

MGI does not make any promise of this manual, including (but not limited to) any commercial of special purpose and any reasonable implied guarantee. MGI has taken measures to guarantee the correctness of this manual. However, MGI is not responsible for any mistakes or missing parts in the manual, and reserves the right to revise the manual and Device, so as to improve the reliability, performance or design.

Figures in this manual are all illustrations. The contents might be slightly different from the software, please refer to the software purchased.

The company, product names, and trademarks are the property of their respective owners.

Release date: February 26, 2020

Manufacturer information

Manufacturer	MGI Tech Co., Ltd.
Address	Main Building and Second floor of No.11 Building, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, Guangdong, China
Technical support	MGI Tech Co., Ltd.
Technical support E-mail	4000-966-988
Technical support E-mail	MGI-service@genomics.cn

Revision history

Revision	Date
1.0	February 26, 2020

Contents

Overview

1 Overview 2

Learning about MegaBOLT

2 MegaBOLT storage and directory structure 4

System storage structure 4

MegaBOLT software directory architecture 4

Directory structure of ZLIMS analysis results 5

Reference directory structure and description 6

Bed file 6

3 MegaBOLT sub-module 7

Program and tools 7

Source code of open source software 8

4 MegaBOLT Server and client 9

Server and client Introduction 9

MegaBOLT system service 9

5 MegaBOLT logs 11

Log structure 11

- Server logs 11
- Client logs 11
- Task log 12

Log level explanation 12

General Linux operation guide

6 Portable storage device mounting 14
Mounting storage through the graphical interface..........................14
Mounting storage through the command line.................................15

7 Network configuration...17
Enabling network setup 17
Setting network connection...18
User verification..20
Automatic network connection 21

Appendixes

FAQ...23
Device related FAQ 23
Workflow related FAQ...24
Error related FAQs..26

Overview

1

Overview

This manual is intended for users that are familiar with the basic use of MegaBOLT and need further details and features.

Firstly, this manual introduces the structure of MegaBOLT storage, software directory, and ZLIMS analysis results directory; the structure of the MegaBOLT submodules; default reference and relevant information of bed interval files used by WES data analysis and the MegaBOLT log system, so that you can understand MegaBOLT in-depth and use MegaBOLT for diverse data analysis and downstream analysis more conveniently and flexibly. Secondly, this manual introduces the Server-Client structure and system services, which guides you to inquire and control MegaBOLT running status through system services.

Thirdly, this manual provides guidance on the general operation under the Linux operating system, including portable storage device mounting and network settings.

Finally, this manual provides quick solutions for common problems and errors when using MegaBOLT.

Learning about MegaBOLT

2

MegaBOLT storage and directory structure

System storage structure

MegaBOLT is equipped with two SSDs, which are used to store the operating system and MegaBOLT software respectively. In addition, the MegaBOLT workstation is equipped with a large-capacity data storage (Sata HDD). The hardware mounting structure and its functions are described in the table below:

Storage	Mounting directory	Function
SSD	/	Operating system and software
SSD	*/mnt/ssd*	MegaBOLT software
Sata HDD	*/data*	User data storage

MegaBOLT software directory architecture

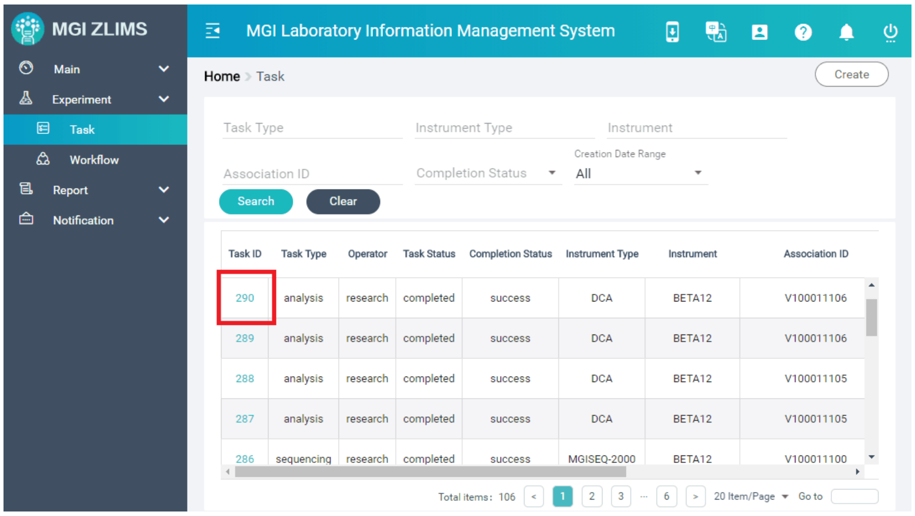
The root directory of the MegaBOLT software is /mnt/ssd/MegaBOLT/, and its contents strcuture is shown as follows:

Directory	Items stored
bin	Executable files
config	Configuration files
docs	User guides
env	Runtime environment dependencies
lib	Libraries
lic	License files
log	Item
test	Test cases
tmpDir	Temporary files and intermediate files
tools	Third party software tools

Directory structure of ZLIMS analysis results

After the analysis tasks submitted in ZLIMS by the user completes, analysis results generated are stored under the */zlimsFiles/* directory. The output directory for ZLIMS analysis tasks is named with the task submission time and task ID. The task ID corresponds with **Task ID** on the **Task** page. For example, the output directory of the 290th task is as follows:

/zlimsFiles/2019-09-20-14-44-27_290



Creation Date Range: All

Task ID	Task Type	Operator	Task Status	Completion Status	Instrument Type	Instrument	Association ID
290	analysis	research	completed	success	DCA	BETA12	V100011106
289	analysis	research	completed	success	DCA	BETA12	V100011106
288	analysis	research	completed	success	DCA	BETA12	V100011105
287	analysis	research	completed	success	DCA	BETA12	V100011105
286	sequencing	research	completed	success	MGISEQ-2000	BETA12	V100011100

Total items: 106

Figure 1 ZLIMS task list

The output directory and results of ZLIMS analysis tasks are consistent with the results of running MegaBOLT through command lines. For details, refer to Chapter 8 “Output directory and results” of *MegaBOLT User Manual*.

Reference directory structure and description

MegaBOLT provides three references: hg19 (default), hg38, and hs37d5. The storage diretory is : */mnt/ssd/MegaBOLT/reference/*. Major files and their explanation are shown in the table below:

Item	Description
hg19.*	Hg19 reference file and index
dbsnp_151.vcf.*	Associated dbsnp files of hg19
1000G_phase1.*	KnownSites files. Its associated reference can be identified by label hg19 or hg38 in the file name.
Mills_and_1000G_gold_standard.indels.*	KnownSites files. Its associated reference can be identified by the label hg19 or hg38 in the file name
hg38.*	Hg38 reference file and index
hs37d5/	hs37d5 reference file and index, associated dbsnp and knownSites files
ReadMe	Introduces the reference and download link of its associated files
build.sh	Script programs for building all reference indexes automatically
db	The bed file of WES data analysis

NOTE The hg19, hg38, and their associated files can be downloaded from NCBI.

Bed file

The bed file for WES data analysis is stored in the following directory:

/mnt/ssd/MegaBOLT/reference/db/db_BED/

Use of the abbreviation of bed files are supported by command line argument `--bed`. You can check the relations between the abbreviations and their actual bed file by viewing the file below:

/mnt/ssd/MegaBOLT/reference/db/db_BED/db.list

3

MegaBOLT sub-module

Program and tools

The executable files of MegaBOLT are stored in the following directory:

/mnt/ssd/MegaBOLT/ bin/

Subprograms that can be invoked by MegaBOLT are stored in this directory. You can invoke the submodules independently for customized analysis pipeline development. Major subprograms are described in the table below:

Program	Description
MegaBOLT	MegaBOLT client program, used to submit tasks
MegaBOLT_server	MegaBOLT server program, used to run the tasks submitted by the client
hc4HaplotypeCaller.jar	Germline variant calling procedure
BQSR.jar	Base quality score recalibration program
gatk4_mutect2_acc.jar	Somatic variant calling program
bwa	BWA alignment program
mm2Alignment	Minimap2 alignment program
picard.jar	Picard program
samtools	Samtools program
Sort	Sorting program
SortMarkDup	Sorting and duplicate marking program

In spite of the submodules above, MegaBOLT also provides some third-party applications, which are installed in the following directory for unified management:

/mnt/ssd/MegaBOLT/tools/

Software tools under this directory are described in the table below:

Directory	Description
GATK	GATK software, including GATK3.7 and GATK4.0
biobambam2	Biobambam2 tool kit
DeepVariant	Variant calling procedure based on deep learning
DNApipe	Took kit relevant to bioinformatics analysis

Source code of open source software

Part of the subprograms are customized and optimized in performance and accuracy by the MegaBOLT R&D team. The original source code of these subprograms are stored in the following directory:

/mnt/ssd/MegaBOLT/tools/src/

4

MegaBOLT Server and client

Server and client Introduction

MegaBOLT adopts the Server-Client structure. The server program resides in the background. When the client submits a task, the server parses and then executes the task.

The MegaBOLT server is monitored by the MegaBOLT system service. This service starts with the operating system and automatically starts the MegaBOLT server. You do not need to pay attention to the running status of the server program when submitting tasks.

MegaBOLT system service

The MegaBOLT system service is used to monitor the server program, ensuring that the server program runs normally and stably. Through the MegaBOLT system service, you can query the server running status, and start, shut down, or restart the server program.

NOTE The MegaBOLT server is essential for the normal operation of the MegaBOLT system. Operate under the instruction of the technical support or local distributor unless you are sure about the results of the following operation.

1. Querying the server running status

   To query the server running status, execute the following command:

   ```
   service megaboltd status
   ```

   A sample of the results generated is shown below:

   ```
   ● megaboltd.service - SYSV: start/stop/restart/reload
   daemon of MegaBOLT server
      Loaded: loaded (/etc/rc.d/init.d/megaboltd; bad;
   vendor preset: disabled)
      Active: active (running) since Mon 2019-11-04
   09:58:40 CST; 7h ago
        Docs: man:systemd-sysv-generator(8)
      Memory: 63.9 GB
      CGroup: /system.slice/megaboltd.service
   ```

```
             ├──1972 /usr/bin/megaboltd
             └──2024 /mnt/ssd/MegaBOLT_scheduler/bin/
MegaBOLT_server -restart
```

"active (running)" in the results indicates that the server is running.

2. Stopping the server

   To stop the server, execute the following command:

```
sudo service megaboltd stop
```

   Enter the SU passwords and the following information is displayed:

```
Stopping megaboltd (via systemctl):          [ OK ]
```

   Both the MegaBOLT system service and the server are stopped and the MegaBOLT tasks submitted at this time will fail.

3. Starting server

   To start MegaBOLT system service and server, execute the following command:

```
sudo service megaboltd start
```

   Enter the SU passwords and the following information is displayed:

```
Starting megaboltd (via systemctl):          [ OK ]
```

   The MegaBOLT system service and server are started.

4. Restarting the server

   Modification on MegaBOLT configuration files only takes effect after restarting the server. To restart MegaBOLT system service and server, execute the following command:

```
sudo service megaboltd restart
```

   Enter the SU passwords and the following information is displayed:

```
Restarting megaboltd (via systemctl):        [ OK ]
```

   MegaBOLT system service and server are restarted.

5

MegaBOLT logs

Log structure

Three types of logs are generated when MegaBOLT is running, which are server logs, client logs, and task logs. The logs record different types of running information, which are recorded in different levels to help you quickly view task information or locate errors.

Server logs

Server logs are divided into four log files. Which are stored in the following directory: */mnt/ssd/MegaBOLT_scheduler/log* The functions of the log files are described in the table below:

Item	Description
megabolt_server.err	Standard output stream/standard error stream of the server program
megabolt_server.log	Server program runtime log
megabolt_rockey.log	Dongle data traffic log
megabolt_task.log	Server task log

Client logs

Client logs are divided into two log files, that is, *megabolt.log* and *megabolt.out*, which are stored under the output directory of MegaBOLT analysis tasks. The functions of the two log files are described in the table below:

Item	Description
megabolt.out	Standard output/error stream of the client program
megabolt.log	Client program running log

Task log

Each sample is treated as an analysis task in MegaBOLT, corresponding to one line in the sample list file. The analysis result and task log of each sample are output to a directory named by the sample name. Task logs are divided into two log files that are named by the sample name. When the sample name and `--outputprefix` are not specified, the sample will be named as "*output*" and the log files will be named as "*samplename.log*" and "*samplename.out*". The functions of the two log files are described in the table below:

Item	Description
samplename.log	The log for the current sample analysis task
samplename.out	The running log for the subprogram that process the current sample analysis task.

Log level explanation

MegaBOLT log information is divided into different levels according to the severity. Higher level indicates higher severity. The log levels are described in the order of low severity to high severity in the table below:

Log level	Description
debug	Debug information. The debug mode is disabled by default. No debug information is output by default.
info	General logs
warning	Warning information
error	Error information that is generated when a task fails but does not affect the execution of other tasks.
critical	Fatal error that is generated when MegaBOLT server or client program exits abnormally.

General Linux operation guide

6 Portable storage device mounting

The MegaBOLT server uses the CentOS7 operating system and the hard disk is mounted differently from the Windows operating system. The following sections introduce the CentOS portable storage devices mounting from the aspects of graphical interface and the command line respectively.

Mounting storage through the graphical interface

CentOS natively supports portable storage devices in VFAT and FAT32 format. Just plug the storage device into any USB port, CentOS will automatically recognize and mount the storage to the operating system and generate a shortcut icon for the storage device on the desktop, as shown in the figure below:

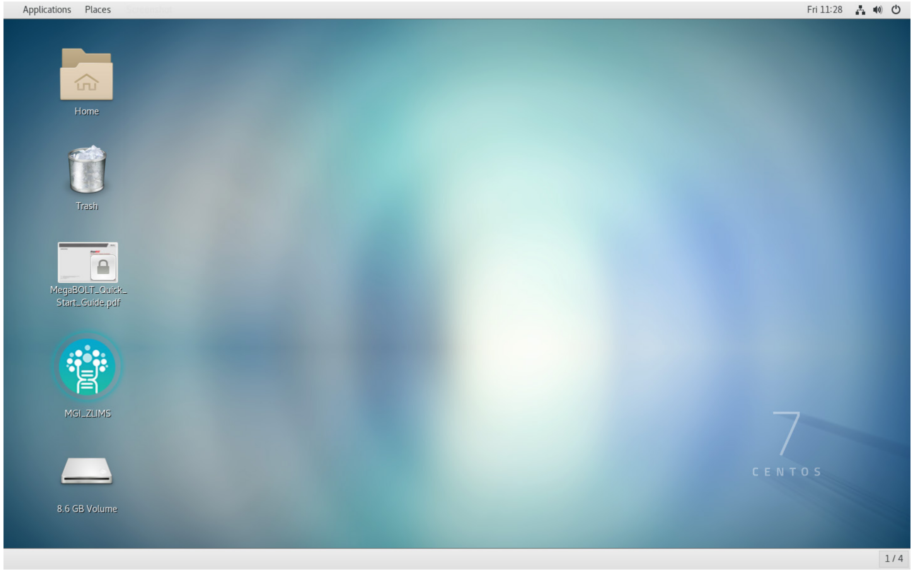


Figure 2 CentOS7 automatic storage device mounting

Portable storage devices in NTFS and ext4 format cannot be directly recognized by CentOS and need to be mounted through the command line.

Mounting storage through the command line

When mounting portable storage devices in NTFS, ext4 format through the non-graphical interface or when required, you need to mount the portable storage devices through the command line. The following takes a U disk with the capacity of about 16 GB as an example, and describes the step-by-step command line to mount a storage device.

1. Viewing the list of disks

   To view the storage devices of the current server, execute the following command:

   ```
   sudo fdisk -l
   ```

   A sample of the results generated is shown below:

   ```
   ……
   Disk /dev/nvme1n1: 1024.2 GB, 1024209543168 bytes,
   2000409264 sectors
   Units = sectors of 1 * 512 = 512 bytes
   Sector size (logical/physical): 512 bytes / 512 bytes
   I/O size (minimum/optimal): 512 bytes / 512 bytes
   Disk label type: dos
   Disk identifier: 0x659fab7d

   Device Boot      Start         End      Blocks   Id
   System
   /dev/nvme1n1p1            2048  2000409263
   1000203608   83  Linux

   Disk /dev/sda: 15.1 GB, 12501054772 bytes, 24025497
   sectors
   Units = sectors of 1 * 512 = 512 bytes
   Sector size (logical/physical): 512 bytes / 512 bytes
   I/O size (minimum/optimal): 512 bytes / 512 bytes
   Disk label type: dos
   Disk identifier: 0x4f8af734

   Device Boot      Start         End      Blocks   Id
   System
   /dev/sda1        16779264   240254975   111737856    7
   HPFS/NTFS/exFAT
   ……
   ```

The output result shows that the U disk is named "*/dev/sda*", the file system is in NTFS format, and only one partition "/dev/sda1" exists.

2. Mounting disks to the operating system

   If the disk uses the NTFS file system, execute the following command to mount the disk:

```
mkdir /mnt/usb
mount -t ntfs-3g /dev/sda1 /mnt/usb
```

   If the disk uses the FAT32, ext4, or VFAT file system, execute the following command to mount the disk:

```
mkdir /mnt/usb
mount /dev/sda1 /mnt/usb
```

   After executing the commands above, you can access the source of the disk mounted in the "*/mnt/usb*" directory.

   The modification made to the mounting of portable storage devices are temporary, it will be invalidated after restart.

3. Uninstalling a disk

   When the use of a disk is complete, you need to uninstall the disk before removing the portable storage device, in case of damage to the data. To uninstall the disk, execute the following command:

```
umount /mnt/usb
```

7

Network configuration

The MegaBOLT server adopts the CentOS7 operating system. You can set up the network by following the guidance provided in this chapter.

Enabling network setup

1. Click the message icon on the upper right corner and select **Ethernet**>**Wired Settings**.

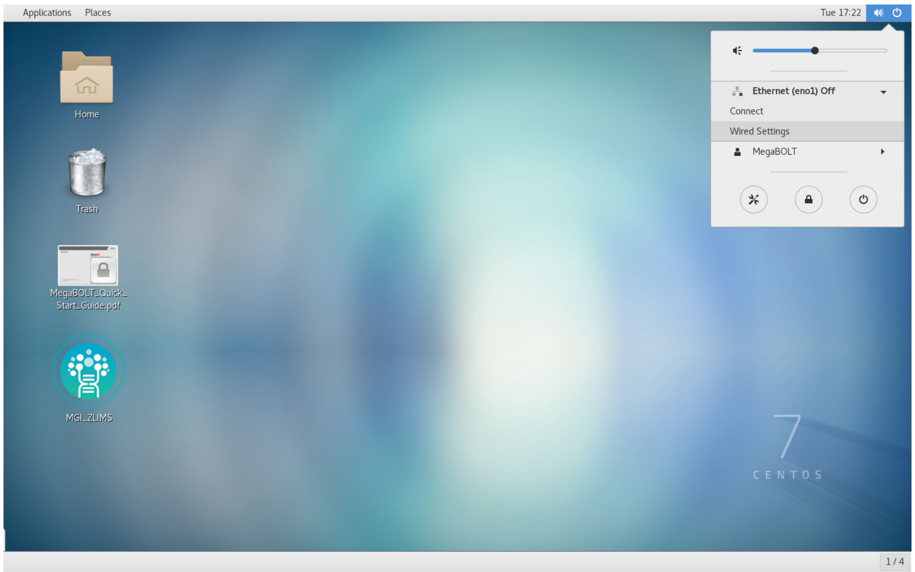
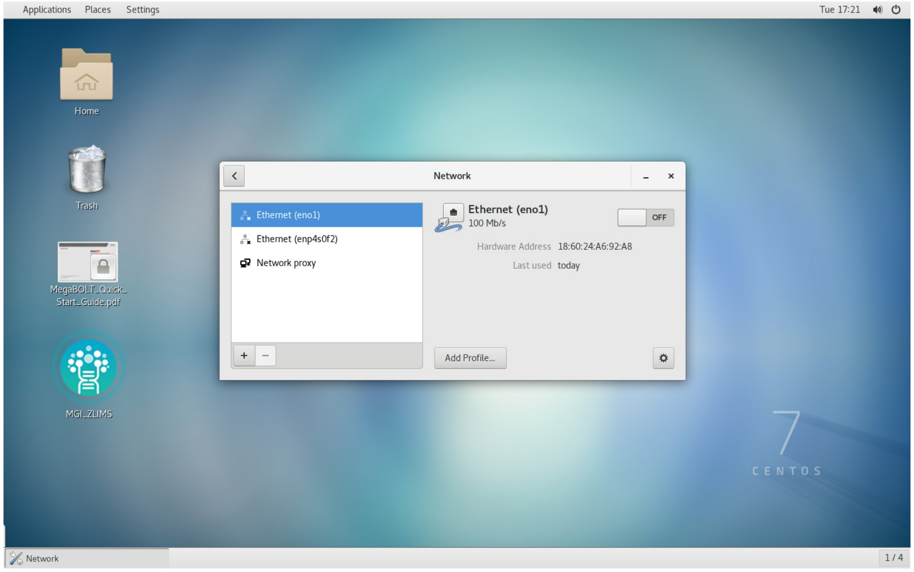


Figure 3 CentOS7 shortcut

2. In the Wired Settings window, available Ethernet ports for the current device are displayed on the left; the Ethernet port connected to network is shown on the right with network bandwidth and the network connection switch can be turned on or off.



Ethernet (eno1)
100 Mb/s
Hardware Address 18:60:24:A6:92:A8
Last used today

Figure 4 Setting the network

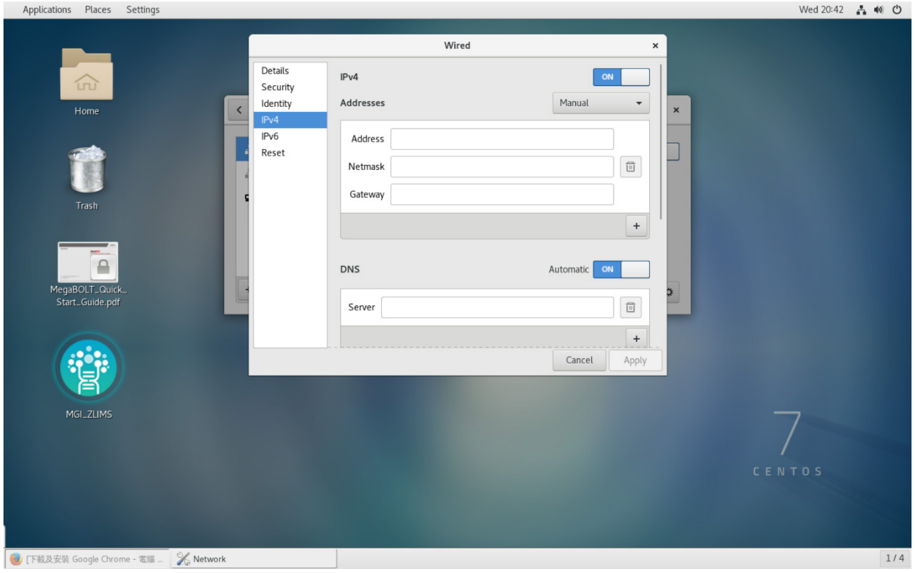
Setting network connection

1. Dynamic network connection

   If the target network supports DHCP service, turn on the network connection switch, and the network will be connected successfully after a short moment.

2. Static network connection

   To configure the static network, open the IPv4 window, set **Addresses** to **Manual**, and input the static IP address of the target network in the **Address** box, fill out the **Netmask** and **Gateway** boxes, and apply the changes, as shown in Figure 4. Turn on the network connection switch again to validate the settings.



IPv4 ON

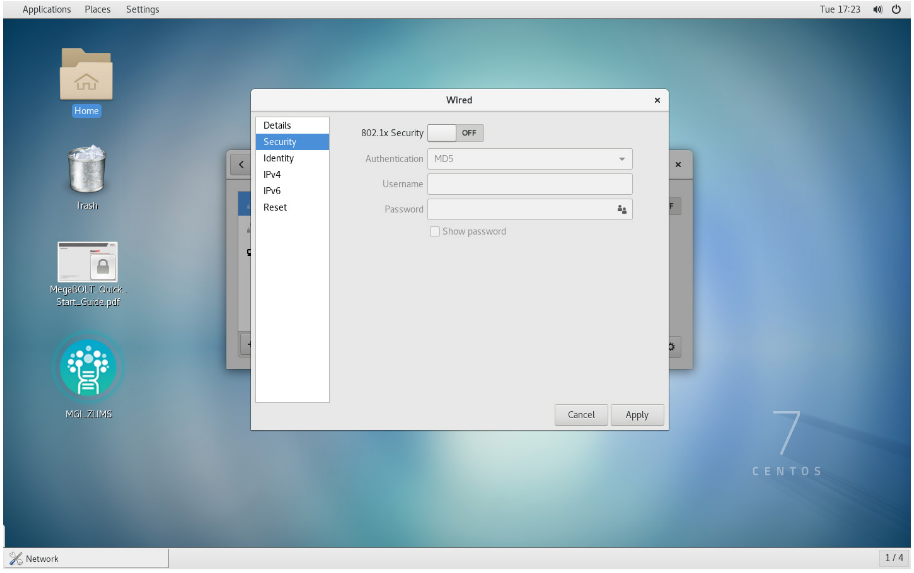
Addresses Manual

DNS Automatic ON

Figure 5 Static network setup

User verification

If the target network requires user verification, click the setup icon on the bottom right of the Settings interface to enter the network security setup window.



802.1x Security OFF

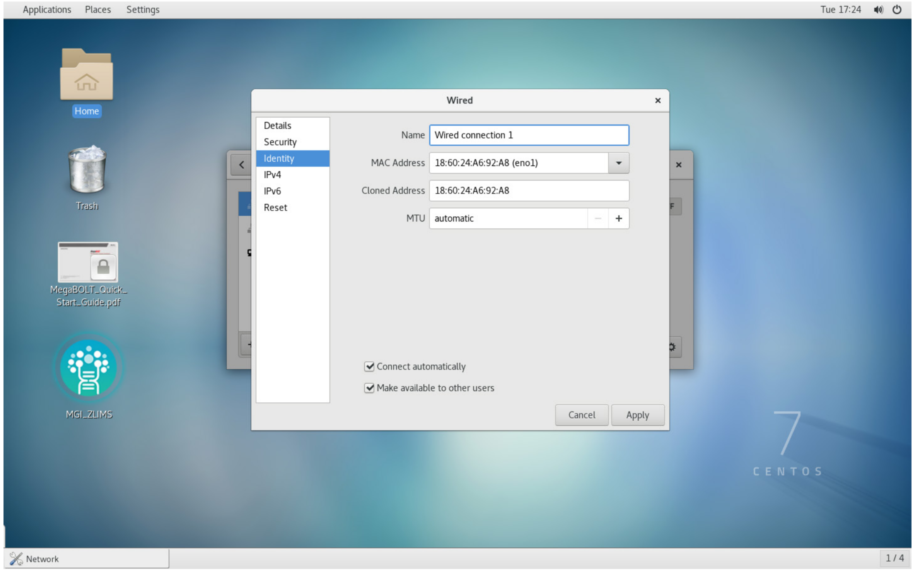
Authentication MD5

Figure 6 Online user verification

Select **Security**, turn on the switch, select the user verification method next to **Authentication** according to the user verification mode of the target network, input the user name and password, and click **Apply** to validate the modification. In the **Wired Settings** window, turn on the network connection switch and the connections will succeed in a moment.

Automatic network connection

The CentOS 7 operating system supports automatic network connection after startup. After network user authentication, enter the network security settings window again. Click **Identity** on the left side of the window, and select **Connect Automatically** at the bottom of the window. The device will be automatically connected to the network after restart.



Name: Wired connection 1
MAC Address: 18:60:24:A6:92:A8 (eno1)
Cloned Address: 18:60:24:A6:92:A8
MTU: automatic
☑ Connect automatically
☑ Make available to other users

Figure 7 Automatic connection option

Appendixes

FAQ

Device related FAQ

1 Q Does the device provide storage of large capacity?

A The MegaBOLT workstation provides 20 TB data storage by default.

2 Q What is the amount of data that can be analyzed by using the temporary computing storage?

A The maximum amount of data that can be analyzed highly efficient is "*SSD size/5*". Analysis can also be performed for data that exceeds the amount, but the speed decreases. For example, a 2 TB SSD can analyze data up to 400 GB.

3 Q Can two tasks run simultaneously?

A Yes. MegaBOLT schedules tasks through the task scheduling system.

4 Q Can other programs run on the device?

A It is not recommended, but you can add other applications that do not affect the running of the MegaBOLT.

5 Q Can other programs run simultaneously with MegaBOLT?

A It is not recommended unless you clearly know the risk of the operation. The running of MegaBOLT occupies most of the hardware resources. The running of other programs simultaneously might cause MegaBOLT analysis tasks to fail, or even result in server crash.

6 Q Can files be stored to the MegaBOLT buffer memory?

A It is not recommended. MegaBOLT generates a lot of intermediate file in the running process and requires sufficient space.

7 Q Does it provide root permissions?

A MegaBOLT provides administrator privileges. Add “sudo” before the command and input the password to execute the command in administrator mode.

8 Q Can the operating system be updated?

A It is not recommended. MegaBOLT provides the best performance and stability in the current operating system version. You need to recompile the driver to update the system. For updates, please contact your local sales representative or manufacturer.

Workflow related FAQ

9 Q Does the current version support unequal read? What is the supported Reads length?

A The current version supports unequal read. The supported Reads length range is 25 bp to 500 bp.

10 Q Does the current version support SE data?

A The current version supports Reads of SE data.

11 Q Does the current version support importing multiple Reads files?

A The current version support importing Reads through the list file. For details, refer to *MegaBOLT_User_Manual*.

12 Q How to generate gVCF results?

A Set **--ERC** to **GVCF**

13 Q How to set PCR-free free data?

A For HaplotypeCaller 3.7 or 4.0, set `--pcr-indel-model` to **NONE**. For DeepVariant, set `--WGS_mode` to **PCR-free**.

14 Q Can I input multiple vcf files (dbSNP) by setting `--vcf` multiple times?

A It is not supported currently. If you input multiple vcf files, only the last vcf file will be used.

15 Q Can multiple bam files be used as input for HaplotypeCaller?

A It is not supported in this version.

16 Q How to use DeepVariant for variant calling?

A Set `--deepvariant` to **1**.

17 Q How to run WES pipeline?

A Set `--runtype` to **WES**.

18 Q How to analyze Single End data?

A Set `--se` to **1**.

19 Q Does the basic pipeline in this version contains bam and vcf statistic information?

A No. it doesn’t. If you need statistic information, run full pipeline.

20 Q How to run quick analysis of other species?

A Use `--ref ref.fa` and `--vcf db.vcf`.

21 Q When running submodules separately, the result files are all named “*output.**”, can the name be modified?

A You can change the name by setting `--outputprefix`.

22 Q Is absolute path necessary for input files or index?

A Yes, we recommend that you use the absolute path for input files and index.

23 Q Does the pipeline support CombineGVCFs and GenotypeGVCFs?

A CombineGVCFs is not supported. GenotypeGVCFs is supported and can be used to turn GVCF into VCF.

24 Q Are the following analysis types supported? CNV, SV, annotation, RNA, metagenomics, and de novo genome assembly.

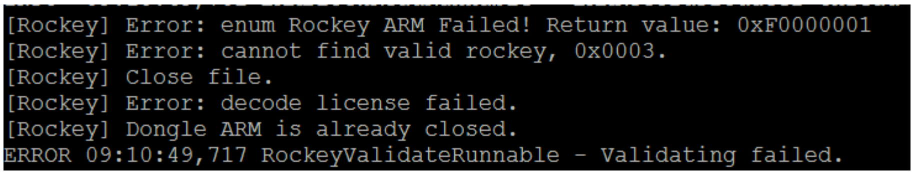
A They are not supported in the default version.

25 Q What is the maximum number of analysis tasks that can be supported by MegaBOLT?

A The number is not limited. All the samples listed in the sample list file will be processed in order.

Error related FAQs

26 Q What should I do when the following dongle error occurs?



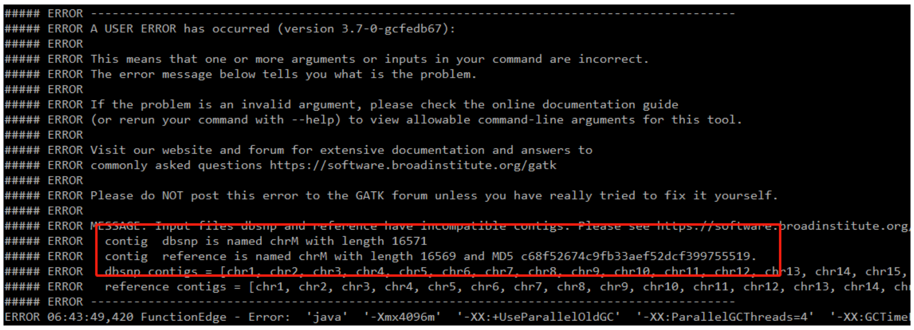
```
[Rockey] Error: enum Rockey ARM Failed! Return value: 0xF0000001
[Rockey] Error: cannot find valid rockey, 0x0003.
[Rockey] Close file.
[Rockey] Error: decode license failed.
[Rockey] Dongle ARM is already closed.
ERROR 09:10:49,717 RockeyValidateRunnable - Validating failed.
```

A Insert the dongle again and restart the device. If the problem persists, contact your local distributor or supplier.

27 Q The following error is reported in BQSR: "*Exception in thread "main" java.lang.UnsupportedClassVersionError: bqsr/BaseQualityScoreRecalibration : Unsupported major.minor version 52.0*"

A You specified java path in .bashrc but the java version is too low. The java version must be higher than JDK1.8. You can disable the java path in .bashrc and run the task again.

28 Q The following error occured when running HaplotypeCaller: "*Input files dbsnp and reference have incompatible contigs.*"



```
##### ERROR ------------------------------------------------------------------------------------
##### ERROR A USER ERROR has occurred (version 3.7-0-gcfedb67):
##### ERROR
##### ERROR This means that one or more arguments or inputs in your command are incorrect.
##### ERROR The error message below tells you what is the problem.
##### ERROR
##### ERROR If the problem is an invalid argument, please check the online documentation guide
##### ERROR (or rerun your command with --help) to view allowable command-line arguments for this tool.
##### ERROR
##### ERROR Visit our website and forum for extensive documentation and answers to
##### ERROR commonly asked questions https://software.broadinstitute.org/gatk
##### ERROR
##### ERROR Please do NOT post this error to the GATK forum unless you have really tried to fix it yourself.
##### ERROR
##### ERROR MESSAGE: Input files dbsnp and reference have incompatible contigs. Please see https://software.broadinstitute.org
##### ERROR   contig  dbsnp is named chrM with length 16571
##### ERROR   contig  reference is named chrM with length 16569 and MD5 c68f52674c9fb33aef52dcf399755519.
##### ERROR   dbsnp contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15,
##### ERROR   reference contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, ch
##### ERROR ------------------------------------------------------------------------------------
ERROR 06:43:49,420 FunctionEdge - Error:  'java'  '-Xmx4096m'  '-XX:+UseParallelOldGC'  '-XX:ParallelGCThreads=4'  '-XX:GCTime
```

A The reference version and dbsnp version are not the same.

29 Q An error of the main script is reported but no detailed information is provided. How can I get the detailed error information?

A You can check the following error logs in the output directory: *samplename.out*, *samplename.log*, and *megabolt.log*.

30 Q After terminating a task, subsequent tasks fail to be submitted. What should I do?

A It may be resulted by device crash after the task is terminated. The node may need to be restarted.

31 Q The following error occurs when running a MegaBOLT task: “*Socket exception - Client connect server failed, strerror: Connection refused*”

A The MegaBOLT sever program is not started. You can restart the analysis in one minute. If the problem persists, execute `sudo service megaboltd restart` to start the MegaBOLT server program, and submit the task again.

---This page is intentionally left blank.---