Cogent™ NGS Analysis Pipeline Quick Start Guide

The following information is provided as a high-level introduction to the software, also referred to as CogentAP. For more detailed information, please see the Cogent NGS Analysis Pipeline User Manual.

Before You Begin

A. Supported operating systems

- Linux: CentOS 8 or higher, RedHat 8 or higher, Ubuntu 18.04 or higher

B. Hardware minimum requirements

- CPU: 24 cores
- Memory: 64 GB RAM
- Free disk space: 1 TB

NOTE: If analyzing data generated with the Shasta™ Total RNA-Seq kit or Shasta Whole-Genome Amplification kit, free hard disk space of at least 8–10 times the size of the input FASTQ files is needed.

C. Additional dependencies

- Internet connectivity on the server
- Conda 24.4.0 or higher
- Bash UNIX shell
- bcl2fastq/BCL Convert

D. Required input files

NOTE: For a list of supported Takara Bio chemistries, please refer to our bioinformatics portal.

- For RNA-seq and scDNA-seq CNV calling:
  - FASTQ files generated by an Illumina® sequencing platform
  - A well list (text file), Illumina sample sheet, or similar TSV/CSV format file
- For scDNA-seq SNV calling:
  - Folder of CNV calling results
  - A CSV file resulting from the CNV clustering module in the CogentDS scDNA app

Confirm Conda Version

- Verify Conda is installed and meets or exceeds the required version by typing the following into a terminal window:

  ```
  conda -V
  ```

  If Conda is successfully installed, it should return text with the version number.

  Example:

  ```
  conda 24.4.0
  ```

- Verify that the base Conda environment can be activated by typing:

  ```
  conda activate
  ```

  Type the following command to return to the default Linux prompt.

  ```
  conda deactivate
  ```

- Verify the install location of miniforge3 is configured in the file `.bash_profile`

  a. For an individual user account, type:

  ```
  more ~/.bash_profile
  ```

  b. Confirm if a similar export PATH command is showing in the file (all on one line):

```
export \
PATH="/home/<USERNAME>/miniforge3/bin:$PATH"
```

  where **<USERNAME>** is replaced by the username of the account that installed Conda.

  If no `.bash_profile` file exists or the line isn't displaying, it will need to be manually created and populated.

Installation

1. Sign up to download the installation package from our website.
2. Move or copy the ZIP file downloaded from Step 1 onto the Linux server into the directory location where you want to install.
3. Unzip the installation package by running the following two commands in the order listed:

```
unzip Cogent_NGS_Analysis_Pipeline_v3.2.zip
\
&& mv Cogent_NGS_Analysis_Pipeline_v3.2 \
CogentAP
cd CogentAP
```

4. Run the following command to install CogentAP and its dependencies:

   ```
   bash CogentAP_setup.sh install
   ```

5. Install the human genome build (for the mouse genome build, use mm39 rather than hg38):

```
bash CogentAP_setup.sh genome_install hg38
```

NOTE: Refer to Section IV.C of the Cogent NGS Analysis Pipeline User Manual for how to set up the `$COGENT_AP_HOME` variable for subsequent commands.

Generation of raw FASTQ Files

1. Log in to a server that stores the run folder from Illumina sequencing and has the bcl2fastq program installed.
2. Change to a working folder where you want the raw FASTQ files to be located after being generated.
3. To convert BCL files to raw FASTQ files using bcl2fastq, go to Step 3a. If using BCL Convert, go to Step 3b.
   a. Run bcl2fastq with the following syntax template:

```
bcl2fastq -R <RUN_FOLDER> \
  -o <RUN_ID> \
  --no-lane-splitting \
  --sample-sheet \
  $COGENT_AP_HOME/config/SampleSheet_du
  mmy.csv > <RUN_ID>.stdout \
  2 > <RUN_ID>.stderr
```

   where:
   - `<RUN_FOLDER>` is the path to the sequencing run folder and
   - `<RUN_ID>` is the ID number automatically generated by the Illumina sequencer

   The file `SampleSheet_dummy.csv` is stored in the CogentAP `config` folder

   Continue to Step 4.

   b. Run BCL Convert with the following syntax template:

```
bcl2fastq -bcl-input-directory \
  <RUN_FOLDER> --output-directory \
  <RUN_ID> --no-lane-splitting \
  --sample-sheet=DummySampleSheet > \
  <RUN_ID>.stdout 2 > <RUN_ID>.stderr \
  $COGENT_AP_HOME/config/SampleSheet_du
  mmy.csv \
  > <RUN_ID>.stdout 2 > <RUN_ID>.stderr
```

   Templates for the `DummySampleSheet` for BCL Convert are stored in the CogentAP `config/` folder. Continue to Step 4.

4. Move the raw FASTQ files to your preferred storage location. They are typically generated in the `<RUN_ID>` folder and named similar to:

   `Undetermined_ S0_R1_001.fastq.gz`
   `Undetermined_ S0_R2_001.fastq.gz`

NOTE: NextSeq® 1000/2000 and NovaSeq™ X/X Plus data are not supported by bcl2fastq and require BCL Convert.

RNA-Seq Analysis

Use the methods below for all kits except Shasta Total RNA-Seq kit. For this Shasta kit, see the next section.

- To demultiplex (demux) RNA-seq data:

```
$COGENT_AP_HOME/cogent rna demux \
  -f <FASTQ_R1> \
  -p <FASTQ_R2> \
  -b <WELL-LIST> \
  -t <EXP_TYPE> \
  -o <DEMUX_OUTPUT_DIR>
```

where:
- `<FASTQ_R1>` and `<FASTQ_R2>` are the full paths to the FASTQ files generated by an Illumina sequencing platform
- `<WELLLIST>` is the full path to the Shasta or ICELL8® cx system well list, Illumina's sample sheet, or TSV/CSV format file
- `<EXP_TYPE>` is the experiment type used (e.g., `icell8_fla`; refer to the user manual for options)
- `<DEMUX_OUTPUT_DIR>` is the full path of the demultiplex results directory

- To analyze RNA-seq data:

```
$COGENT_AP_HOME/cogent rna analyze \
  -i <DEMUX_OUTPUT_DIR> \
  -g <GENOME> \
  -t <EXP_TYPE> \
  -o <ANALYSIS_OUTPUT_DIR>
```

where:
- `<DEMUX_OUTPUT_DIR>` is the full path of the demultiplex results directory
- `<GENOME>` is a name of genome build (e.g., `hg38`)
- `<EXP_TYPE>` is the experiment type used (e.g., `icell8_fla`; refer to the user manual for options)
- `<ANALYSIS_OUTPUT_DIR>` is the full path to the output directory created for the analysis results

Shasta Total RNA-Seq Kit Analysis

The recommended method to analyze sequencing data from the Shasta Total RNA-Seq Kit - 2 Chip is the RNA Analyze Direct (`analyze_direct`) workflow (Sections V.B.1.b and c of the user manual). To run the command:

1. Perform an RNA demux with the `--dry_run` argument, with the command:

```
$COGENT_AP_HOME/cogent rna demux \
  --dry_run \
  -f <FASTQ_R1> \
  -p <FASTQ_R2> \
  -b <WELL_LIST> \
  -t shasta_total_rna \
  -o <OUTPUT_DIR>
```

   where:
   - `<FASTQ_R1>` and `<FASTQ_R2>` are the full paths to the FASTQ files generated by an Illumina sequencing platform
   - `<WELL_LIST>` is the full path to the Shasta system well list (text file)
   - `<OUTPUT_DIR>` is the full path of the demux dry run results directory
2. Using the `--dry_run` output, generate a barcode ranks file in CogentDS. Refer to the Cogent NGS Discovery Software User Manual or Quick Start Guide for more information.
3. Run the `analyze_direct` command:

```
$COGENT_AP_HOME/cogent rna \
  analyze_direct \
  -f <FASTQ_R1> \
  -p <FASTQ_R2> \
  -t shasta_total_rna \
  -b <BARCODES_FILE> \
  --cogentds_barcode_ranks <RANKS_FILE>
  -o <OUTPUT_DIR>
```

   where:
   - `<FASTQ_R1>` and `<FASTQ_R2>` are the full paths to the FASTQ files generated by an Illumina sequencing platform
   - `<BARCODES_FILE>` is the full path to the Shasta well list
   - `<RANKS_FILE>` is the full path to the barcode ranks file generated from CogentDS Barcode Rank Plot module
   - `<OUTPUT_DIR>` is the full path of the analyze direct results directory

NOTE: Additional commands are available to analyze sequencing data generated with the Shasta Total RNA-Seq kit. See Section V.B.1.d and V.B.1.e of the Cogent NGS Analysis Pipeline User Manual for details.

DNA-Seq or WGA Analysis

- To demultiplex (demux) DNA-seq data:

```
$COGENT_AP_HOME/cogent dna demux \
-f <FASTQ_R1> \
-p <FASTQ_R2> \
-b <WELL-LIST> \
-t <EXP_TYPE> \
-o <DEMUX_OUTPUT_DIR>
```

  where:
  - `<FASTQ_R1>` and `<FASTQ_R2>` are the full paths to the FASTQ files generated by an Illumina sequencing platform
  - `<WELLLIST>` is the full path to the Shasta or ICELL8 cx well list, Illumina's sample sheet, or TSV/CSV format file
  - `<EXP_TYPE>` is the experiment type used (e.g., `shasta_wga`, refer to the user manual for more options)
  - `<DEMUX_OUTPUT_DIR>` is the full path of the demultiplex results directory

- To analyze DNA-seq data (CNV calling):

```
$COGENT_AP_HOME/cogent dna analyze \
-i <DEMUX_OUTPUT_DIR> \
-g <GENOME> \
-t <EXP_TYPE> \
-B <BIN_SIZE> \
-r <READ_LENGTH> \
-R <READ_FILTER> \
-b <BARCODES_FILE> \
-o <ANALYSIS_OUTPUT_DIR>
```

  where:
  - `<DEMUX_OUTPUT_DIR>` is the full path of the demultiplex results directory
  - `<GENOME>` is a name of genome build (e.g., `hg38`)
  - `<BIN_SIZE>` is the bin size used for CNV analysis using Ginkgo; must be either `500kb` or `1mb`
  - `<READ_LENGTH>` is the read length of the input data; must be either `76bp` or `151bp`

- <READ_FILTER> is the minimum number of paired-end reads required per barcode to be kept in downstream analysis
- <BARCODES_FILE> is the full path to the Shasta or ICELL8 cx system well list, Illumina's sample sheet, or TSV/CSV format file
- <ANALYSIS_OUTPUT_DIR> is the full path to the output directory created for the analysis results

- (Optional) To do SNV calling:

```
$COGENT_AP_HOME/cogent dna \
postprocess snv_calling \
-i <ANALYSIS_OUTPUT_DIR> \
-g <GENOME> \
-b <BARCODES_FILE> \
--cluster_mapping_file <CLUSTER_CSV> \
-o <SNV_OUTPUT_DIR>
```

where:

- <ANALYSIS_OUTPUT_DIR> is the full path to the output directory created for the analysis results
- <GENOME> is a name of genome build (e.g., hg38)
- <BARCODES_FILE> is the full path to the Shasta or ICELL8 cx system well list, Illumina's sample sheet, or TSV/CSV format file
- <CLUSTER_CSV> is the full path to the CogentDS [Download Clusters] CSV file resulting from the CNV calling results
- <SNV_OUTPUT_DIR> is the full path to the output directory created for the SNV calling results

Contact Us	
Customer Service/Ordering	Technical Support
tel: 800.662.2566, option 1 (toll-free)	tel: 800.662.2566, option 2 (toll-free)
fax: 800.424.1350 (toll-free)	fax: 800.424.1350 (toll-free)
web: takarabio.com/service	web: takarabio.com/support
e-mail: ordersUS@takarabio.com	e-mail: technical_support@takarabio.com

Takara Bio USA, Inc.
2560 Orchard Parkway, San Jose, CA 95131, USA
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999