Handbook Seq Suite

CRISPR Guide Enrichment Data Analysis

For Research Use Only.

Legal Notices

Document 1020776, Rev A, Feb 2024

3210 Merryfield Row
San Diego, CA 92121, United States
https://scale.bio/
support@scale.bio

Disclaimer of Warranty:
TO THE EXTENT PERMITTED BY APPLICABLE LAW, SCALEBIO PROVIDES THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. IN NO EVENT WILL SCALEBIO BE LIABLE TO YOU OR ANY THIRD PARTY FOR ANY LOSS OR DAMAGE, DIRECT OR INDIRECT, FROM THE USE OF THIS DOCUMENT, INCLUDING WITHOUT LIMITATION, LOST PROFITS, LOST INVESTMENT, BUSINESS INTERRUPTION, GOODWILL, OR LOST DATA, EVEN IF SCALEBIO IS EXPRESSLY ADVISED IN ADVANCE OF THE POSSIBILITY OF SUCH LOSS OR DAMAGE. Any warranties applicable to the ScaleBio products are set forth in the Terms and Conditions accompanying such product and such Terms and Conditions are not modified in any way by the terms of this notice.

Trademark Information:
ScaleBio may make reference to products or services provided by other companies using their brand names or company names solely for the purpose of clarity, and does not assert any ownership rights over those third-party marks or names. Images were created with BioRender.com

Patent Information:
ScaleBio products may be covered by one or more patents as indicated at: https://scale.bio/legal-notice/

Terms and Conditions:
The use of the ScaleBio products described herein is subject to ScaleBio's Terms and Conditions that accompany the product, or such other terms as have been agreed to in writing between ScaleBio and the user.

Intended Use:
All products and services described herein are intended FOR RESEARCH USE ONLY and NOT FOR USE IN DIAGNOSTIC PROCEDURES.

Table of Contents

Introduction

The ScaleBio™ CRISPR Guide Enrichment Kit v1.1 enables the simultaneous capture of CRISPR guide sequences as well as mRNA transcripts coming from the same cell. A 3-level combinatorial indexing strategy resolves gene expression data at single cell resolution by using a combination of unique RT Barcodes, Ligation Barcodes and PCR Barcodes.

The ScaleBio Seq Suite: CRISPR Data Analysis Pipeline is designed to be used in conjunction with the ScaleBio Seq Suite: RNA Data Analysis Pipeline. Please contact your local Field Application Scientist or email *support@scale.bio* to obtain the handbook for the ScaleBio Seq Suite: RNA Data Analysis Pipeline.

This handbook serves as a high-level guide for setting up, running, and understanding the outputs of ScaleBio Seq Suite: CRISPR, the ScaleBio Single Cell CRISPR Data Analysis Pipeline. For specific step-by-step instructions on installing and running the pipeline please refer to our GitHub repository (https://github.com/ScaleBio/ScaleCRISPR). The introductory readme markdown file (README.md) provides an overview of the workflow; additionally, there are a series of markdown files (*.md) within ScaleCRISPR/docs to help guide users in more detail at each major step.

Chapter 1: Pipeline Setup

Please refer to the pipeline setup instructions for ScaleBio Seq Suite: RNA, as both pipelines are based on Nextflow. Additionally, the ScaleCRISPR pipeline requires the ScaleRNA pipeline to have been run on the RNA library of the samples for proper cell calling within the CRISPR library.

The pipeline can be downloaded by two methods:

1. By going to the ScaleBio Seq Suite: CRISPR GitHub page, clicking the green “Code” button and then “Download ZIP”. Unpack this file on your system directly in the directory in which you want to install the pipeline. To make sure the download is complete, make sure the executable commands (PY files) in the ScaleCRISPR/bin directory have the appropriate read/write/execute privileges on your server.
2. The GitHub repository can be cloned to your machine:

   ```
   git clone https://github.com/ScaleBio/ScaleCRISPR.git
   ```

Note this may require setting up a personal access token, instructions for which can be found here.

Chapter 2: Input Files

2.1. Sequencing Reads

- Path to the Illumina Sequencer RunFolder (BCL files).
- If you prefer to start from previously generated FASTQ files, see Fastq generation.

2.2. Sample Table (samples.csv)

- A CSV file listing all samples in the analysis, optionally split by RT Barcode. See samples.csv for an example of such a table and the requirements for each column. The construction of this document is very similar in design to the ScaleRNA workflow's samples.csv. The major difference is in the addition of two columns called "guide" and "allCells".
    - The guide column should contain the name of the file that contains the pool of guide sequences for detection per sample (files found in guides directory specified below, in tab separated format).
    - The allCells column contains the name of the associated ScaleRNA sample allCells.csv file per CRISPR library, found in the supplied allCells directory below.

2.3. guide

- Path to folder containing the CRISPR guide sequences to be detected and quantified by the workflow. See guides.md.

2.4. allCells

- Path to folder containing ScaleRna allCells.csv output files, one for each sample to be analyzed for the CRISPR analysis. This can easily be supplied by referencing the "samples" output subdirectory of the ScaleRNA analysis for the paired RNA libraries. See allCells.csv.

2.5. outDir

- Path to desired output directory for workflow analysis.
- The workflow produces per-sample QC reports (sample.metrics.csv), a cell-by-guideUMI count-matrix (guideTab.filtered.csv) and more; see Outputs for a full description.

Chapter 3: Step-by-Step Pipeline Overview

The FASTQ generation steps that the CRISPR pipeline performs are the same sequence of events that the RNA workflow performs, see Chapter 3 (sections FASTQ Generation through Generation of Library QC Report) for reference. The result of this is a per **SAMPLE** set of FASTQ files, split on RT barcode, and preliminary QC of those FASTQs. From here the workflow diverges into CRISPR specific considerations, described below.

3.1.Guide Sequence Input QC

- Guide sequences to be detected need to be of the same length and found in the same base pair position within the read as the other guides in the **SAMPLE** for proper counting. The qcGuides.py script ensures this by:
    1. Trimming the guide sequences to the same minimum length if necessary
    2. Searching through each read of a subset of that sample's FASTQ files and identifying the average start position of all the guide sequences in the reference file.
- This information is then written to a template JSON file that is used per SAMPLE for guide detection

3.2. Cell Barcode and Guide Sequence Detection

- In this step, we are again using bcParser to extract the three single-cell barcodes with error correction, but this time are adding in the guide sequence details as another "barcode" to be detected in the read with its own reference list.
- Error correction is done against the list of expected barcode or guide sequences, allowing up to 1 mismatch.

3.3. Guide Counting and Calling

- Per-read cell barcode and guide content is used to generate a cells x guide UMI counts matrix and cells x guide reads counts matrix. This is done by joining each read to the observed cell barcode combination (RT, Ligation, PCR), and counting the reads associated with the guide sequence detected. These reads are output as one matrix, and UMI information is used to deduplicate the counts into the final guideTab.csv output file.
- Critically, a determination of calling guides within each cell is performed
    - Minimum guide UMI detection threshold (default 3) is used to say what guide was called within the cell. If there are multiple guides within that cell that passed the threshold, then both will be attributed
- Per sample metrics files are generated to indicate library and assay performance.
- Per cell QC metrics matrices are also output, providing a matrix of useful calculations.
- These tables are further supervised by the ScaleRNA pipeline's output of what cell barcodes "passed", resulting in filtered tables with consistent cell content from the RNA library.
    - A metric of what percentage of passing RNA cells also had a passing guide in them is reported.

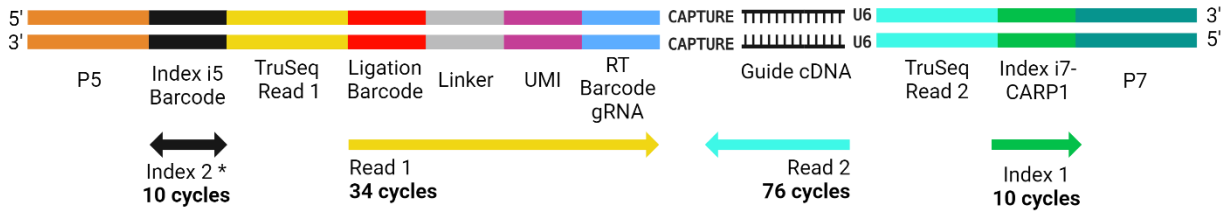
3.4. Sample Performance and Visualization

- The filtered guide x UMI counts matrix and filtered per-cell QC metrics are used to assess sample performance as well as assist in guide assignment threshold performance, found in the “analysis” subdirectory of each sample’s output.
- Emphasis on visualization of CRISPR cell metadata, and the distribution of detections of guides within the libraries, and guides within cells.
- Assessment of performance of the minimum threshold for guide assignment, with visualization of sliding scale of threshold and the percentages of cells that contain a passing guide at each UMI cutoff.

Appendix A: CRISPR Library Structure and List of Barcode Sequences

The overall library structure for the CRISPR library processed by the CRISPR Guide Enrichment Kit v1.1 is shown in Figure 1.

Figure 1: Library Structure for the CRISPR Guide Enrichment Kit v1.1 Assay



* orientation depends on sequencer and sequencing chemistry

Component	Size	Description
P5	-	Illumina P5 sequence (AATGATACGGCGACCACCGAGATCTACAC)
Index i5 Barcode	10 bp	Combinatorial Indexing
TruSeq Read 1	-	Illumina sequencing primer
Ligation Barcode	9 bp	Combinatorial Indexing. Half of the Ligation Barcodes are followed by an extra 'spacer' base (for a total length of 10 bp)
Linker	7 bp	Fixed sequence (TCAGAGC) that is used to anchor UMI and RT Barcodes regardless of the optional 'spacer' in the Ligation Barcode
UMI	7 bp	Unique Molecular Identifier
RT Barcode	10 bp	Combinatorial Indexing
Nextera Read 2	-	Illumina sequencing primer
I7	10 bp	Index i7-CARP1
P7	-	Illumina P7 sequence (ATCTCGTATGCCGTCTTCTGCTTG)

For the CRISPR library, the Ligation Barcodes and Index i5 Barcode sequences are the same as in the RNA library generated by the CRISPR Guide Enrichment Kit v1.1 assay. The RT Barcode sequences for the CRISPR library are different from the RT Barcode sequences of the RNA library. The full list of all barcode sequences can be found in the references directory of the workflow. The overall combinatorial cell-barcode is made up of a combination of RT Barcode, Ligation Barcode and the Index i5 Barcode. The Index i7 sequences (Table 1) are used in conjunction with the ScaleBio CRISPR Enrichment Extended Throughput Kit v1.1.

Table 1: Index i7 Sequences used in conjunction with the ScaleBio CRISPR Guide Enrichment Extended Throughput Kit v1.1.

Barcode Name	Reagent Name	Forward Sequence	Reverse Sequence
CRISPR-A-RP1	CRISPR Amp Reverse Primer 1	AAGTAGACTA	TAGTCTACTT
CRISPR-A-RP2	CRISPR Amp Reverse Primer 2	TGGATCAGGC	GCCTGATCCA
CRISPR-A-RP3	CRISPR Amp Reverse Primer 3	GTTACTTAGC	GCTAAGTAAC
CRISPR-A-RP4	CRISPR Amp Reverse Primer 4	ACCGCCGCAA	TTGCGGCGGT

Appendix B: Software Dependencies

- Java (11 or later)
- Nextflow (22.04 or later)
- Fastq generation and read processing:
    - bcl-convert 3.9
    - python=3.10
    - pandas=1.5.2
    - samtools=1.16.1
    - fastqc=0.11.9
    - multiqc=1.14
    - cutadapt=4.2
    - star=2.7.10b
- CRISPR Analysis Steps
    - Python= 3.11.3
        - numpy= 1.24.3
        - pandas=2.0.1
        - seaborn= 0.12.2
        - biopython=1.81
        - matplotlib = 3.7.1

Document Revision History

Revision	Revision Date	Document ID	Changes
Rev A	Feb 2024	1020776	Initial release.