Fluent BioSciences

Getting Started with PIPseeker

Tutorial

Doc ID: **FB0003900**
Revision: **3**
State: **Release**

Introduction

This is a short tutorial on how to install PIPseeker and analyze a simple dataset. After completing this tutorial, you should be able to start processing your own PIPseq data. However, this is not a comprehensive guide and it does not cover all the possible configurations and use cases. Furthermore, this tutorial does not explain all the different PIPseeker outputs. For more complete information, please download the PIPseeker User Guide at: https://www.fluentbio.com/resources/pipseeker-user-guide.

System Requirements

The resources required to run PIPseeker depend on the size and composition of the input data and on the size of the reference genome. The PIPseeker User Guide has extensive information about estimating the requirements for your datasets. To complete this tutorial, you will need a computer with at least 21GB of available memory (mostly to store the human genome) and 1 GB of free disk space.

Linux:

- Type "free -h" to see the system's memory.
- Type "df -h" to see available disk space.

Mac:

- Click on the Apple icon and select "About This Mac" to see the system's memory.
- In the same window, switch to the "Storage" tab to see available disk space.

Windows:

- In the left sidebar on the File Explorer window, right-click on "This PC" and select "Properties" to see the system's memory.
- In the same sidebar, right-click on the drive you intend to use for data analysis, and select "Properties" to see available disk space.

Installing PIPseeker

Linux

To install PIPseeker on Linux, follow these steps:

1) Download the Linux package from https://www.fluentbio.com/resources/pipseeker-downloads
2) Open a terminal window and navigate to your downloads directory.
3) Uncompress the package:

```
tar -zxvf pipseeker-v3.1.2-linux.tar.gz
```

(note that the release number may change)

4) Move the PIPseeker executable, simply called "pipseeker" to a permanent location, such as "PIPseeker" in your home directory:

```
mkdir ~/PIPseeker
mv pipseeker-v3.1.2-linux/pipseeker ~/PIPseeker
```

5) Turn on executable permissions:

```
chmod +x ~/PIPseeker/pipseeker
```

macOS

PIPseeker runs on Mac computers as a command-line application. To install it, follow these steps:

1) Download the macOS package from https://www.fluentbio.com/resources/pipseeker-downloads
2) Open a terminal window and navigate to your downloads directory.
   - Use the Finder to open the /Applications/Utilities folder, then double-click "Terminal".
   - The downloads folder can typically be reached with the command

     ```
     cd /Users/<your user name>/Downloads
     ```
3) Uncompress the package:

   ```
   tar -zxvf pipseeker-v3.1.2-mac.tar.gz
   ```

   (note that the release number may change).
4) Move the PIPseeker executable, simply called "pipseeker", to a permanent location, such as "PIPseeker" in your home directory:

   ```
   mkdir ~/PIPseeker
   mv pipseeker-v3.1.2-mac/pipseeker ~/PIPseeker
   ```
5) Turn on executable permissions:

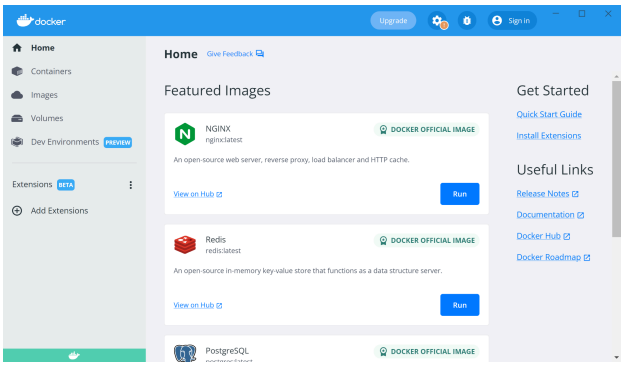
   ```
   chmod +x ~/PIPseeker/pipseeker
   ```

Windows

PIPseeker uses some Windows-incompatible third-party tools. Therefore, it uses a Docker container to run a linux-based binary on Windows. A Docker container is a virtual Linux environment that enables running Linux applications on any system. Prior to downloading and installing PIPseeker, you will need to install Docker on your machine. The installer is available at https://docs.docker.com/get-docker.

Follow the Docker installation instructions without changing any of the default options. **Note**: Docker requires a Linux backend installed on your machine. Please see this page for more information: https://docs.docker.com/desktop/install/windows-install. You may need to download and install the linux kernel update package.

Once installation is complete, run "Docker Desktop" from the start menu. You should see something like this:



You will need to launch Docker Desktop whenever you restart your computer. You do not need to do anything else: The PIPseeker image will be fetched automatically when you start the analysis.

Some of the PIPseeker downloads are in gzip format, which is not supported natively on Windows. You will need to download a third-party tool. 7-Zip is a popular free tool that can process gzip files, and is available at https://www.7-zip.org.

Follow these steps to install PIPseeker:

1) Download the Windows package from: https://www.fluentbio.com/resources/pipseeker-downloads
2) Using the File Explorer, navigate to your downloads folder.
3) Unzip the PIPseeker archive (e.g., 7-Zip -> Extract Here in the context menu).
4) Move pipseeker.exe to a permanent location. For example, you can create a folder called "PIPseeker" in your main drive (C:\PIPseeker).

To test your installation, follow these steps:

1) Open Windows PowerShell by typing "PowerShell" in the Start menu.
2) Launch PIPseeker from the install location (e.g., C:\PIPseeker).:

```
c:\PIPseeker\pipseeker.exe -h
```

Note: The first launch of PIPseeker may take a few minutes.

3) If everything was set up correctly, you should see something similar to this output (PIPseeker version may differ):

```
No sub-command selected. Please choose from: full, cells, barcode, feature,
buildmapref, buildannotref, merge, extract
```

Mapping Reference

PIPseeker requires a gene-annotated reference genome for mapping. Some common references are available for download at:
https://www.fluentbio.com/resources/pipseeker-downloads/#mapping-references

PIPseeker can also use any STAR-compatible reference. This document has information on how to create your own mapping references:
https://www.fluentbio.com/resources/instructions-for-creating-custom-pipseeker-references

For this tutorial, you will need to download the human reference genome: https://fbs-public.s3.us-east-2.amazonaws.com/public-indices/pipseeker-gex-reference-GRCh38-2022.04.tar.gz

After downloading, follow these steps:

Mac/Linux:

1) Open a terminal window and navigate to your downloads folder.
2) Extract the gzipped archive.

```
tar -zxvf pipseeker-gex-reference-GRCh38-2022.04.tar.gz
```

3) Move the uncompressed directory to a permanent location. That can be the same location as the PIPseeker executable.

```
mv pipseeker-gex-reference-GRCh38-2022.04 ~/PIPseeker
```

Windows:

1) Using the file explorer, navigate to your downloads folder.
2) Extract the gzipped archive. **Make sure you uncompress both the gzip and tar files**,
3) Using the file explorer, drag the extracted folder (not the gzip or tar file!) to a permanent location, e.g., c:\PIPseeker (you will need to create this folder).

The human reference directory should contain the following files, with corresponding sizes:

```
      1200 chrLength.txt
      1997 chrName.txt
      3197 chrNameLength.txt
      2129 chrStart.txt
  52878658 exonGeTrInfo.tab
  21984297 exonInfo.tab
   1398821 geneInfo.tab
3190415946 Genome
```

```
      1132 genomeParameters.txt
     39049 Log.out
8527227723 SA
1565873619 SAindex
  10927630 sjdbInfo.txt
  11820516 sjdbList.fromGTF.out.tab
   9686303 sjdbList.out.tab
  14082493 transcriptInfo.tab
```

Cell Type Annotation Reference

PIPseeker supports automatic cell type annotation for certain sample types. This requires an annotation reference file. Annotation references are available for download at: https://www.fluentbio.com/resources/pipseeker-downloads/#annotation-references

In this tutorial, you will be analyzing human peripheral blood mononuclear cell samples (PBMCs), which is one of the supported types for annotation. There are several different types of PBMC annotations included in the downloadable package. We will be using human-pbmc-v4.csv in our analysis. For convenience, it is included in the tutorial dataset.

Analyzing an Example Dataset

Tutorial Data

This tutorial features two human PBMC samples. For simplicity, the FASTQ files from each sample were reduced to 10 million reads. The tutorial dataset is available for download at: https://fbs-public.s3.us-east-2.amazonaws.com/getting-started-tutorial/tutorial-data.tar.gz

After downloading, follow these steps:

Mac/Linux:

1) Open a terminal window and navigate to your downloads folder.
2) Extract the zipped archive.

```
tar -zxvf tutorial-data.tar.gz
```

3) Move the uncompressed FASTQ files to a permanent location, such as "PIPseeker-tutorial" in your home directory.

```
mkdir ~/PIPseeker-tutorial
mv tutorial-data/* ~/PIPseeker-tutorial
```

Windows:

1) Using the file explorer, navigate to your downloads folder.
2) Extract the gzipped archive. **Make sure you uncompress both the gzip and tar files**,
3) Using the file explorer, drag the content of the extracted folder (FASTQ files beginning with "sample1" and "sample2 and the annotation reference human-pbmc-v4.csv") to a permanent location, e.g., C:\PIPseeker-tutorial (you will need to create that folder).

The tutorial directory should contain the following files: sample1_R1.fastq.gz, sample1_R2.fastq.gz, sample2_R1.fastq.gz, sample2_R2.fastq.gz and human-pbmc-v4.csv.

Analyzing Sample 1

We will now use PIPseeker to analyze the first PBMC sample. Note that in the commands shown below, the file paths may be different if you installed PIPseeker and/or the tutorial data in different locations.

Mac/Linux:

1) Open a terminal window and navigate to the directory where you placed the tutorial data.
2) Run the following command:

```
~/PIPseeker/pipseeker full --chemistry v4 --fastq sample1 \
--star-index-path ~/PIPseeker/pipseeker-gex-reference-GRCh38-2022.04 \
--annotation ~/PIPseeker/human-pbmc-v4.csv --output-path sample1-results \
--threads 1
```

Note: Mac users may need to provide the system additional permissions before the executable can be run:

1) After running the above command, a security warning may be displayed, similar to "pipseeker can't be opened because Apple cannot check it for malicious software". If you don't get the message, please ignore steps b-g.
2) Click "OK"
3) Open the "System Preferences" dashboard
4) Click on the "Security & Privacy" icon
5) A note towards the bottom of the dashboard should mention that the PIPseeker executable could not be opened "because it is not from an identified developer."
6) Click "Allow Anyway" (you may need to click the lock button at the bottom left of the screen to make changes).
7) When you launch the PIPseeker executable again, another window may pop up. You can simply click "OK" and the executable will then run without prompting from this point onward.

Windows:

1) Make sure Docker is running (see PIPseeker setup above).
2) Open a Powershell window ("Powershell" in the Start menu) and navigate to the directory where you placed the tutorial data. For example:

```
cd c:\PIPseeker-tutorial
```

3) Run the following command:

```
c:\PIPseeker\pipseeker.exe full --chemistry v4 --fastq sample1 ^
--star-index-path c:\PIPseeker\pipseeker-gex-reference-GRCh38-2022.04 ^
--annotation c:\PIPseeker\human-pbmc-v4.csv --output-path sample1-results ^
--threads 1
```

Note that the --fastq argument constitutes a file prefix and indicates that only FASTQ files whose names begin with "sample1" should be included in the analysis. Also, note that we are running on a single thread to reduce memory consumption.

At this point, you should see PIPseeker running the analysis. This should take up to 30 minutes, depending on your system. PIPseeker displays various diagnostic messages as the analysis progresses.

Once the analysis is complete, you should see multiple files and directories inside "sample1-results". One of the files is a summary report named report.html. Open it with a web browser from the command line or by double-clicking on it in the file explorer.

Interpreting Sample 1

When you open the summary report, you will notice that it consists of five different tabs named "Sensitivity 1" through "Sensitivity 5". Those refer to different cell calling sensitivities (The report opens with Sensitivity 3 being displayed). Generally speaking, barcodes are separated into "cells" and "background" based on their transcript abundance, which is shown in the barcode rank plot. "Sensitivity" indicates how many of the top barcodes shown in the barcode rank plot are considered as cell-associated barcodes, and represents an inherent tradeoff between the number of cells and the quality of their gene expression data.

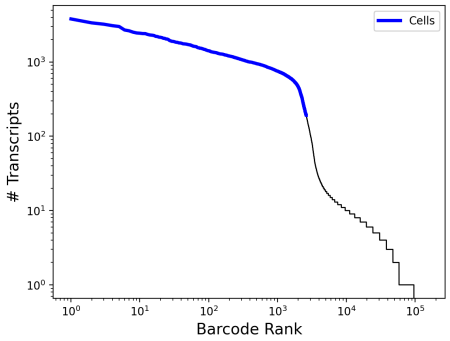
Note: Results may differ slightly between platforms and between different versions of PIPseeker. Your actual results may not look exactly like the images below but should be similar.

Key Metrics

The summary report contains a table with key metrics. Those metrics are useful for getting an impression of the quality of your sample and identifying issues in your experimental workflows. The duplication rate and sequencing saturation metrics can also be used to determine the appropriate allocation of sequencing capacity in your experiments. Please see the PIPseeker User Guide for more information.

As you browse between the different cell calling sensitivities, you will notice that the number of cells increases with sensitivity, whereas the median number of transcripts per cell decreases. This is because at higher sensitivity we are including barcodes with lower transcript counts (barcodes further right on the rank plot) in the cell population.

Barcode Rank Plot

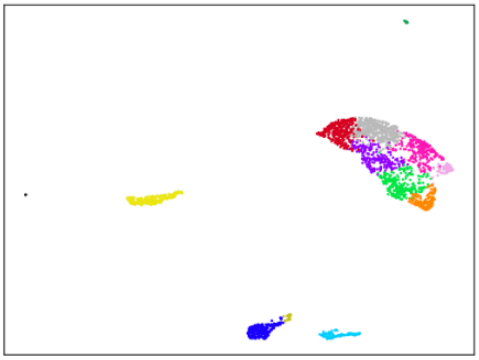


The barcode rank plot shows barcodes sorted by their transcript abundance in descending order. The blue portion represents the barcodes selected as representing cells. Typically, the top mode is indicative of true cells and the bottom mode contains "background" partitions. It is often difficult, however, to determine the exact cutoff point, so PIPseeker allows for some flexibility. Note how the blue portion grows rightward as the sensitivity increases. This means we are calling cells "deeper" into the rank plot.

The PIPseeker User Guide includes detailed information about cell calling sensitivities and their application. Here, we will examine the clustering maps to determine the appropriate sensitivity level for our example dataset.

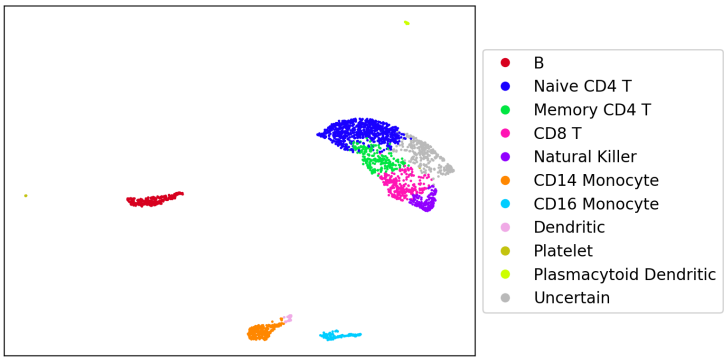
Clustering

Each cell calling sensitivity is associated with a clustering map based on graph-based clustering, as well as a table of the top differentially expressed genes for each cluster. You should see something like this for Sensitivity 3:



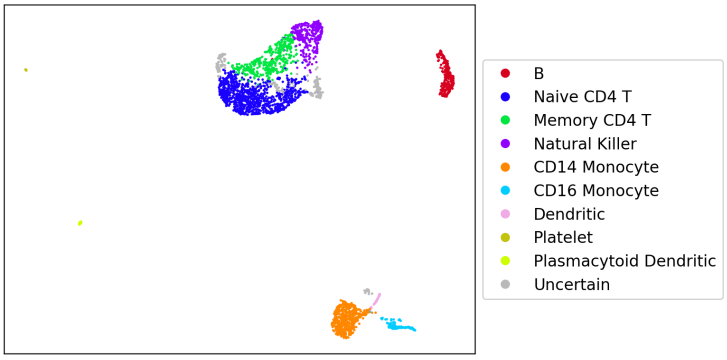
	3	285	CD14 Monocyte	Uncertain	ENSG0000028044
	4	282	CD8 T	Uncertain	NKG7
	5	278	B	Uncertain	CD74
	6	256	Uncertain		MALAT1
	7	239	Memory CD4 T	Central Memory CD4 T	IL32
	8	176	Natural Killer	Uncertain	NKG7
	9	143	CD16 Monocyte	Uncertain	LST1
	10	79	Uncertain		MALAT1
	11	41	Plasmacytoid Dendritic	Uncertain	TCF4

In this example, we are using cell type annotation, so each cluster is associated with a cell type, as shown on the table and in a cell type annotation map. Note that the clusters in the annotation map are colored by cell types and will be different from the colors in the unannotated clustering map, and that two or more distinct clusters may be identified as a single type, as is the case here for CD4 T cells:



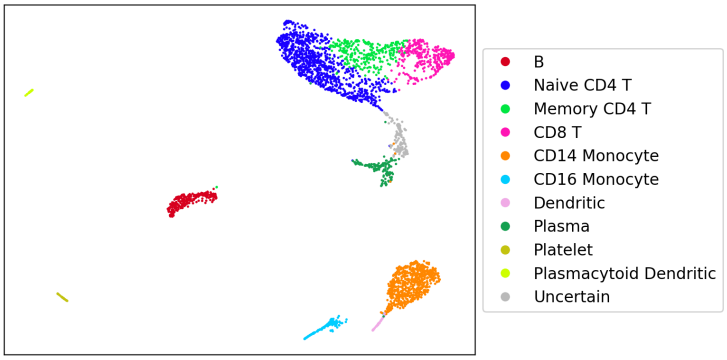
At Sensitivity 3, the monocyte, B cell, and T cell populations appear well-separated. The annotation of the T cell population, however, seems to be missing some subtypes, including any CD8 T cells.

Now advance to Sensitivity 4 by clicking on the “Sensitivity 4” tab at the top of the report:



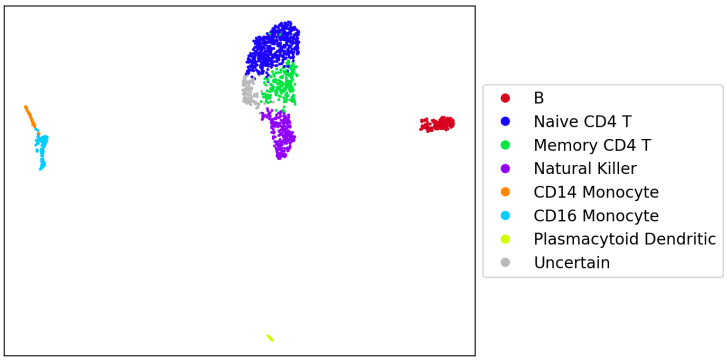
At sensitivity 4, we assign cell barcodes further down the rank plot, thereby including cells with lower transcript counts and potentially of lower quality. Indeed, observe that one of the cell types, CD8 T cells, is no longer present, suggesting poor characterization of the T cell group.

Now advance to sensitivity 5:



Note that while the overall number of cells is higher, we have less clear separation between the monocyte and T cell groups. This suggests that Sensitivity 5 includes a significant number of low-quality partitions.

Now switch to Sensitivity 1:

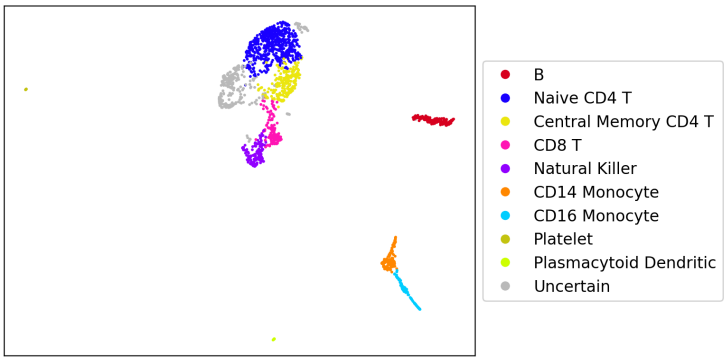


Here, barcodes with the highest transcript abundance are selected as cells. This means better characterization of gene expression in the selected cell population. However, at this sensitivity we lose some cell populations with lower RNA content, such as dendritic cells and most of the CD14 monocytes.

In conclusion, selecting the appropriate sensitivity level constitutes an inherent tradeoff between cell quantity and quality. With your own samples, the appropriate sensitivity level will depend on the purpose and priorities of the experiment and requires knowledge of the underlying biology.

Sample 2

Now try analyzing sample 2 on your own (hint: FASTQ file names start with “sample2”). This is the cell type annotation plot you should see at Sensitivity 3:



Document Revision Summary

Doc ID: FB0003900
Revision: 3
Revision Date: January 2024

Changes:

- Updated for PIPseeker v3.1.2

Legal Notices

Support

Email: support@fluentbio.com

Fluent BioSciences
150 Coolidge Avenue
Watertown, MA 02472