

**NucBase  
User Guide**

**September 2012**

# Contents

---

INTRODUCTION .....	3
<b>1      LOADING READS .....</b>	<b>4</b>
1.1   FASTA/FASTQ CONVERSION.....	4
1.2   INPUT FORMAT AND SELECTION.....	5
<b>2      LOADING TARGET SEQUENCE(S).....</b>	<b>6</b>
<b>3      OPTIONS .....</b>	<b>7</b>
<b>4      OUTPUT.....</b>	<b>8</b>

## List of Figures

Figure 1: Main window.....	3
Figure 2: Conversion Window .....	4
Figure 3: Reads file format .....	5
Figure 4: Reads input: column selection .....	5
Figure 5: Loading sequences .....	6
Figure 6: Options .....	7

# Introduction

---

NucBase is a simple program which allows you to easily align millions of reads to a target sequence.

NucBase main window is divided into three parts:

- The **First Part**, where you load your reads into the program.
- The **Second Part**, where you should input your target sequence(s).
- The **Third Part**, where you can adjust the alignment options.

Finally, the **Status Line** displays the number of lines in the file containing the reads.

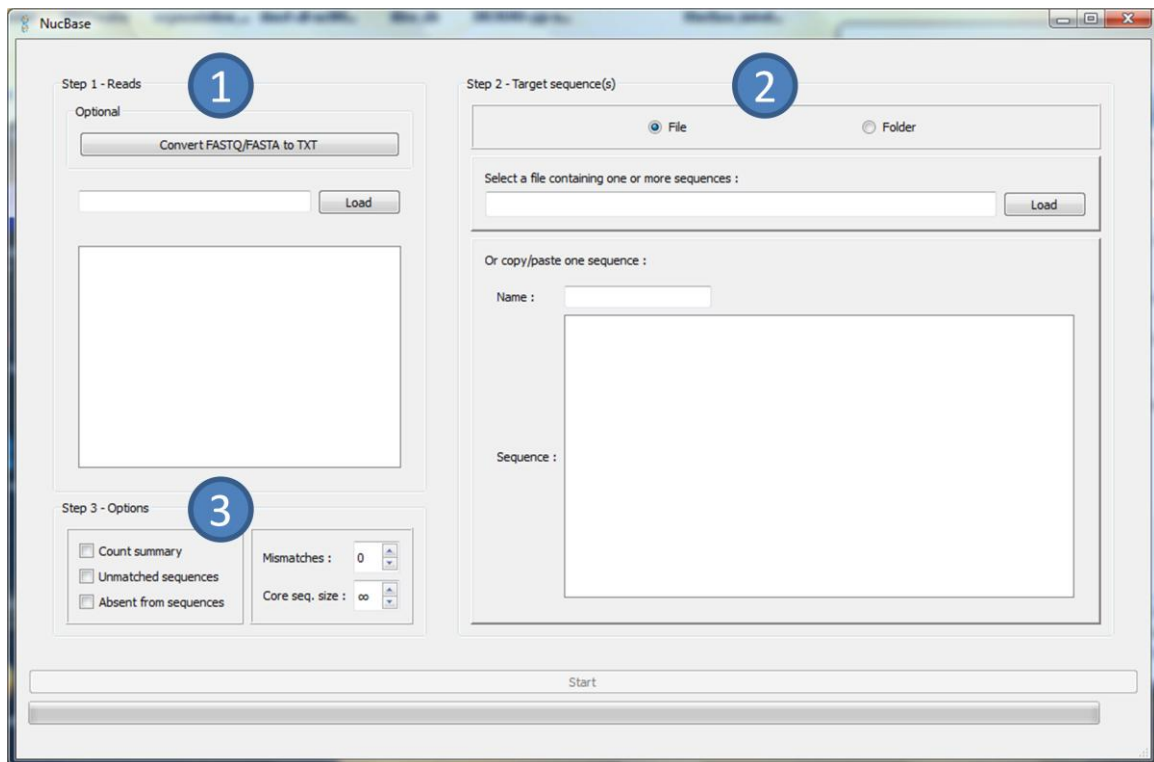


Figure 1: Main window

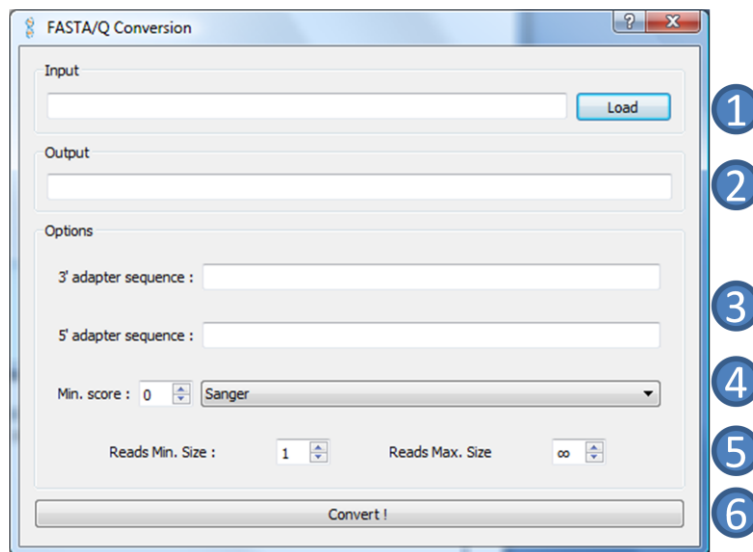
# 1 Loading reads

---

## 1.1 FASTA/FASTQ conversion

In the current version of the program, it is only possible to convert FASTA/FASTQ files to the expected text format before any processing.

To do so, you will need to click on the corresponding button, which will open a new window:



**Figure 2: Conversion Window**

1. Select the file you want to convert (\*.fq, \*.fastq, \*.fa, \*.fasta).
2. You can see where the converted text file will be located.
3. If your fastq sequences contain adapters, enter them here.
4. Select the score encoding (fastq) and set your threshold.
5. Set the min and max sizes of the reads you want to keep (e.g.: 18-34).
6. Convert!

## 1.2 Input format and selection

As an input, NucBase requires tab-separated values text files with reads in the first column.

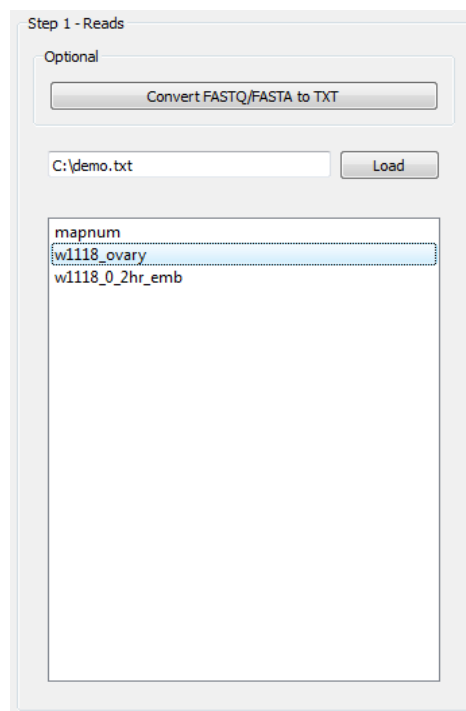
Some test data is available here: <http://srv-gred.u-clermont1.fr/nucbase/NucBase-data.zip>

labels	mapnum	w1118_ovary	w1118_0_2hr_Emb
AAAAAAAAACAAACAGTCTG	1	0	1
AAAAAAAAACACGAGTTAAG	1	2	0
AAAAGCGCGCATTTCGGGTG	3	2	22

**Figure 3: Reads file format**

To start the alignment, you will need to select at least one column. NucBase will then treat all the reads which do not have a “0” in the selected column.

If you only have 1 column containing all the reads, the software will process them all.



The screenshot shows a software interface titled "Step 1 - Reads". It features an "Optional" section with a button labeled "Convert FASTQ/FASTA to TXT". Below this is a text input field containing "C:\demo.txt" and a "Load" button. A list box below the input field contains three items: "mapnum", "w1118\_ovary", and "w1118\_0\_2hr\_emb". The "w1118\_ovary" item is currently selected and highlighted in blue.

**Figure 4: Reads input: column selection**

## **2 Loading target sequence(s)**

There are three ways to input your target sequences into NucBase:

1. Selecting a compatible file.  
(FASTA, multi-FASTA or a raw text file with the sequence on the first line)
2. Selecting a folder containing multiple compatible files.
3. Pasting the sequence in the main window and giving it a name (important).

Step 2 - Target sequence(s)

File  Folder

Select a file containing one or more sequences :

Or copy/paste one sequence :

Name :

Sequence :

**Figure 5: Loading sequences**

NOTE: Currently, the software does not accept degenerate bases

### 3 Options

---

NucBase has 5 options:

1. **Count summary**: this will create a global file containing the number of times each read mapped across all the target sequences.
2. **Unmatched sequences**: this option will create a FASTA file for each sequence containing ‘\*’ wherever reads matched the sequence, providing a way to visualize the alignments – useful to detect clusters.
3. **Absent from sequences**: this option will make the software output the non-matching reads.
4. **Mismatches**: number of mismatches allowed.
5. **Core seq. size**: minimum number of consecutive nucleotides from a read to align on the sequence (partial match).

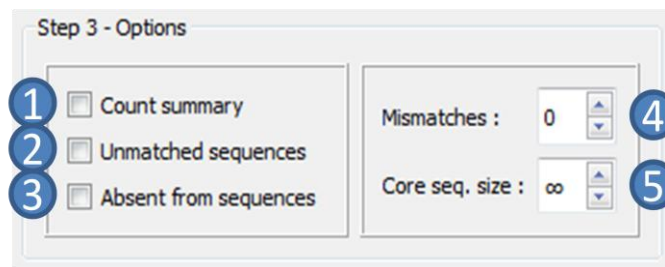


Figure 6: Options

## 4 Output

---

Once you have given all the required data to the program, you can start the alignment. NucBase will then produce several files for each column selected and for each target sequence:

1. A GFF3 file containing the position of all the mapped reads.
2. Two text files sharing the same format as the reads given in input:
  - a. One file for the reads which mapped to the forward sequence.
  - b. One file for the reads which mapped to the reverse-complementary sequence.

Depending on the options, other files might be produced.