# FASTQ+ Format Specification

Shi Quan
shiquan@genomics.cn

April 24, 2022

The update version of this document can be found at
https://github.com/shiquan/PISA/blob/master/FASTQ%2B_Spec.pdf.

## 1 The FASTQ+ Format Specification

The FASTQ+ format is a variant of the FASTQ. It is designed to store more features or annotations for FASTQ records but not change the own property of the FASTQ structure. Each FASTQ+ record consist of four lines.

- The first line starts with an '`@`' character and the sequence identifier. The optional tag fields are between the sequence identifier and the first space in or end of this line. Optional read 1 and read 2 label "`/[12]`" add after tags. Optional description words put at the end of this line, separated by a space.
- The sequence, bases consist of A, C, G, T and N.
- A plus character. Optional sequence identifier can also appended here.
- The base call quality scores.

The optional fields of FASTQ+ is consist of numbers of tags in `TAG:TYPE:VALUE` format. The tag format is inherited from the SAM tag format (https://samtools.github.io/hts-specs/SAMv1.pdf). The `TAG` is a two-character string that matches `/[A-Za-z][A-Za-z0-9]/`. `TYPE` is a single case-sensitive letter which defines the format of `VALUE`. Tags in optional fields are separated by three '|' characters. No space allowed in the sequence identifier and optional fields. The total length of the first line should not exceed 254 characters.

### 1.1 An example

In the below example, SEQ_ID is the sequence identifier that users can set or generate by software. `CB` is the recommended tag name for the corrected cell barcode, `GN` is the recommended tag for gene name, and `UB` is the recommended tag for UMI.

```
@SEQ_ID|||CB:Z:ACGT|||GN:Z:BRCA1|||UB:Z:AACG
GATTTGGGGTTCAAAGCAGTATCGA
+
1***-+*''))**5>>CCCCCCC65
```

### 1.2 Optional tag fields

All optional tags for FASTQ+ follow the SAM tag format but add two more restrictions.

1. SAM tags allow space in type Z, but not allowed in FASTQ+ tags.
2. The length of SAM tags is not limited, but FASTQ+ read identifier and tags are specified to be no longer than 254 characters.

| Type | Regexp matching `VALUE` | Description |
|---|---|---|
| A | `[!-~]` | Printable character[1] |
| i | `[-+]?[0-9]+` | Signed integer[2] |
| f | `[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?` | Signed single-precision floating number[3] |
| Z | `[!-~]*` | Printable string. |
| H | `([0-9A-F][0-9A-F])*` | Byte array in the Hex format[4] |
| B | `[cCsSiIf](,[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?)*` | Integer or numberic array[5] |

## 1.3 Recommended tag names and types for single cell experiments

The following tag names and types have been widely used in single-cell experiments. Although it is not required but highly recommend following the exact definition.

| Tag name | Type | Description |
|---|---|---|
| CR | Z | Raw cell barcode. |
| CB | Z | Cell barcode that is confirmed against a list of known-good barcode sequences. |
| UR | Z | Molecular barcode. |
| UB | Z | Molecular barcode that is corrected among other molecule barcodes. |
| GN | Z | Gene name. |
| TX | Z | Transcript id. |
| GX | Z | Gene id. |

## 1.4 Read block

A combination of tags can define the FASTQ+ block. Reads with the same tags can be selected and manipulated in a group. For example, SEQ1, SEQ2, and SEQ3 share the same cell barcode, and SEQ2 and SEQ3 share the same gene name but not for SEQ1. Using CB to define the block, all these three reads are from the same block. But if using CB and GN to determine the block, SEQ2 and SEQ3 are from the same block, different from SEQ1. Sorting the FASTQ+ file by tags can facilitate downstream analysis, i.e., assembly, for each block to be processed sequentially.

```
@SEQ1|||CB:Z:ACGT|||GN:Z:BRCA1
GATTTGGGGTTCAAAGCAGTATCGA
+
1***-+*’’))**5>>CCCCCCC65
@SEQ2|||CB:Z:ACGT|||GN:Z:SAA1
ACACTCGAAGATACAGAAATGAGTA
+
EEEEEE6/E/EEEAEEEE/E/EEE<
@SEQ3|||CB:Z:ACGT|||GN:Z:SAA1
TATCGACAGGAAGAAGGAGGGAGGG
+
AE/EE</AEEEAEEEEE//AA/EEE
```

[1] Same with BAM tag type A
[2] Same with BAM tag type i
[3] Same with BAM tag type f
[4] Same with BAM tag type H
[5] Same with BAM tag type B

# 2 FASTQ+ Version History

## 2.1 v1.0 : Apr 2022

Initial edition.