Bioinformatics Overview

FASTQ

- FASTA with Quality

- https://en.wikipedia.org/wiki/FASTQ_format

FASTQ: Filenames

"What's in a name? that which we call a **FASTQ**

By any other name would smell as sweet"

— *Romeo and Juliet*, William Shakespeare

FASTQ: Filenames

```
1C_TAAGGCGA_L006_R1_001.fastq.gz
1C_TAAGGCGA_L006_R1_002.fastq.gz
1C_TAAGGCGA_L006_R1_003.fastq.gz
1C_TAAGGCGA_L006_R1_004.fastq.gz
1C_TAGGCATG_L003_R1_001.fastq.gz
1C_TAGGCATG_L005_R1_001.fastq.gz
2C_CGTACTAG_L003_R1_001.fastq.gz
```

FASTQ: Filenames

Sample Name	Barcode	Lane #	Read #	File Part	
1C_	TAAGGCGA	_L006	_R1	_001	.fastq.gz
1C_	TAAGGCGA	_L006	_R1	_002	.fastq.gz
1C_	TAAGGCGA	_L006	_R1	_003	.fastq.gz
1C_	TAAGGCGA	_L006	_R1	_004	.fastq.gz
1C_	TAGGCATG	_L003	_R1	_001	.fastq.gz
1C_	TAGGCATG	_L005	_R1	_001	.fastq.gz
2C_	CGTACTAG	_L003	_R1	_001	.fastq.gz

FASTQ: Format

- **FASTA Format**

```
>NC_007779.1 Escherichia coli str. K-12 substr. W3110, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGG
TAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
ATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTG
GCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAA
AACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAA
```

Header
Sequence

- **FASTQ Format**

```
@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGTGCAAAGCGACGCGCCAGTTCCAGATCCGGCG
+
AAABBFFFFFFFFGGGGGGGGGGGGGGHHHHHHHGHGHGHHHHHGHHHGGGGGHHHGGGGGGGGHHHGHHFFHHHHHGHGGGGGGGGGGGHHHHHHHHHGGG
```

Quality
Score

```
@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGCAAAGCGACGCGCCAGTTCCAGATCCGGCG
+
AAABBFFFFFFFGGGGGGGGGGGGGHHHHHHHHGHGHGHHHHHGHHHGGGGGHHHHGGGGGGGHHHGHHFFHHHHHHGHGGGGGGGGGGGHHHHHHHHHHHGGG
@M00698:36:000000000-AFBEL:1:1101:15928:1413 1:N:0:0
GTAAAGTCCTGAGTGATACCGGCAACTTTTACCCCCAGTCCCACTTTCGAACCGGCAAACATATCGGCAAAAGAGGCCGTGCCTGATTTAAAGCCGTAGGT
+
1>AAAFFFFFFFFGGGGGGGGGCEGGGHHHHHFHHHHGGGHHHHHHHHHGHFA0FEGGGGGHGGGHHHHEECEEHFEFH/E/E/>EGGHBHHHDGHHHHC<?E/1
@M00698:36:000000000-AFBEL:1:1101:15876:1413 1:N:0:0
CCGGAACAATCGCTGGCAGGCTTTTTACGTCCGGGTTATAGCGACAGCGGTCTCAATATTCATCAGCCGCTTTCGCTCGCGGGTGCGATTGTCCGCT
+
ABCCCCDFFFFEGGGGGGGGGGGHHHHHHHGHCGGGFFHHHHHGGHGHHGHGGGGHHHHHHHHHGGGGGGCHHHGGGGG<CHHGHGHEHHDGGF
@M00698:36:000000000-AFBEL:1:1101:15340:1413 1:N:0:0
CCACTAACAAACTAGCCTGATTAAGTTTTAACGCTTCAACCCCAGGCAGGGCTTCCACGCGATCTCTTTTGGGTTTGACCTCTCTTGATCCCCGTCCTAAG
+
AABABFFFFFCGGGGGGGGGGGGFHHGFHHGFHGGGGGHHHHGGGGFHGGFGGHHHCFGFFEFCGHHFHHHHGGAFGGFFFFHHHHHHHHHHHGEEGGGHHB
@M00698:36:000000000-AFBEL:1:1101:16045:1413 1:N:0:0
GTAGCATTATCAGAGAGTTGCCATTCACGCATTGGCTTAACCGCGCGCAGACCATCAACAGTCACTTTGGCGTCAAAGACATTAGGCGTGCAGTATTTTTT
+
?ABCCFFFFFFFGGGGGGGGGGGHHHHHHHHGGGHHHHHHHHHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHGGGGGHHHHHHHHHHHHGGGGHHGHHHHHHG
@M00698:36:000000000-AFBEL:1:1101:17191:1414 1:N:0:0
TCGGCACCAATATAGGTAACGCATGGTTCACCGTCTTCAGCTACGGCGGCGATTTTGGTCAGGATCGGTGCTACCAATTCATAGGCTTCTTTCTGGCCAC
+
ABAABBBBFBFFGGGGGGGGGGGGGGHHHHHHHGHGHGHHHHHHHGGGGGGGGGGGHGHGGHHHHHHHHGGGGGHHHHHHHHHHHHFHGHHHHHHHHHHHBFHHG
@M00698:36:000000000-AFBEL:1:1101:16186:1414 1:N:0:0
TCAGGGTCGTCCGTGGAGCAAACATAGCCCTGAGGCTTATTGAGCATGAAGTAACGTGGACCGTGTTGCTGCGCCAGCGGGTTGCCATCGTAAGCGACAT
+
BCCCCFFFCCCCGGGGGGGGGGGHHHGHGHHHHHHHGGHHHHHHHHHHHHHHHHHHHHGHHGGHHGGHHHHHHHHHGGGGGGGGGGGGGGHHHHHGHHHFHGGGGG
@M00698:36:000000000-AFBEL:1:1101:15394:1415 1:N:0:0
ACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGATAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGG
+
@BBCCFFFFFFFGGGGGGGGGGGGGHHDGHHHHHHHHHGEGHGHHHGHHHHHHHHHHEHHHHHHHGGHHHHHHHEGGHHGHHHHHHGGGGFHHHHHHHHHHHH
@M00698:36:000000000-AFBEL:1:1101:14326:1417 1:N:0:0
TGGCGCAACTAACAGAACGTCTTGCGTTTTGTTGGCGAAGCCGCTGGTGTTTGTAAATTTATTAGTGATCGGCGTAGCGTTACGGGTTTCACCGTAGTTCG
+
ABCCCCCCCFFFGGGGGGGGGGGGHHGGGGGHHHHHHGGGGGHGGGGGFGFFHHHHHHHGHHHHHHHHHGHGGGGGFFAE>EGGGGGEGGHHHHGGHGGHHG
@M00698:36:000000000-AFBEL:1:1101:15479:1417 1:N:0:0
GCCCCCCTCACTCTGACTTTAGTGTGCGCCTTTTGCCTGCGCCACGATCTCCTCGACATTTTCCGGCATTGGTTCATACGGCGTTTCTTTCGCCGTTGG
+
@BCCCCCCCFCFGGGGGGGGGGGGHHHHGGGGGHHHHHHGGGGGGGGHHHHHHGGGGGHHHHGGGGGHHHHHHHHHHGGGGGGGHHHHHGGGGGGGGH
@M00698:36:000000000-AFBEL:1:1101:16850:1418 1:N:0:0
CGTTTTCTTCATCGCGCTCTTTGCTGCCTAACAGCGTGCGCCAGCCTGCTTCAGCAAGAAAACGCGCTTTAGCGACAAATTTGCCTTTGGCAATGTCCAGT
+
AAACCCFFFFFFGGGGGGGGGGGGHHHHHHHHHHHHGGGGGGGGGGGGGHHHHHHHHHHHHHHHHGGGGGGGHGHGGGGGFHHHHHHHHGHHHHHHHHHHHHG
@M00698:36:000000000-AFBEL:1:1101:16255:1418 1:N:0:0
CCGGCTTGCTGGTTGCAGCCGTTGCTGTACTTGATGTCAGGCGTGCCGGTGCCGTATTTTCAAACGGTGATGCCGGACGCGACTCTACGGGTTTGACATCG
+
AAABCCCCCFFFGGGGGGGGGGGGGHHHHHHHHHHHHHHHHHHHGGGHGGGGGHGFGHHHHHHHHHHGGGHGHHHHGGGGGGGGGGGGHHHGGGCGGGHHHHH?
@M00698:36:000000000-AFBEL:1:1101:17071:1419 1:N:0:0
CACGCCACCTTTATCCAGTTTGCGGCTTTGCGAAGTGGCGAGCCACAGCGCGTTTTCTTGCTGGCTATAAGCCATTTCGTAGGCACCTTTACCTACCGCTT
+
AABBBBBBFFFFGGGGGGGGGGGHGGGGGHHHGGGGFHHGGGGGGHHHHHGGGFGGGHHHHHHHHGHHHHHHHHHHHHHHHGGHGHGHHHFHHHHHHGGGG
@M00698:36:000000000-AFBEL:1:1101:15606:1420 1:N:0:0
CGGGTTTTTAACTTTCAGCCACGGGCCACCGTCGATCAGTTCACCGCCAAACTCTTCACGCGCCAGCTGGTAGCCCCAGTCTTTAAACGCTCCTTCGGTGA
+
BBCDBBCCDCFFGGGGGGGGGGGGGGGGGGHGGGGGHGHHHHHHGGGGGHHHHHHHHHHGGGGGHHHHHHHHHGGGHHHHHHHHHGGGGGHHHHGFGG
@M00698:36:000000000-AFBEL:1:1101:15945:1421 1:N:0:0
TCATTGTTGCGCGGTATTCGCGCCCGTTGGTCGAGTAGCAGAAGGGGATTTTAAACCGTTGTTTGCCGCTGGTGTCCTGCCAGCTGGTTTCATACTCTGG
+
?ABABFFFFBBBFGEEGGGGGGGGGGGGHGHCHGCGGGHGHHHHHHGGGGHGHGHHHHGHGEGDGHHHHHGGG?EB2GGHGHHHGHGHHHHGHHFHHHHHGGH
@M00698:36:000000000-AFBEL:1:1101:16329:1421 1:N:0:0
GCCCCGACAGCTGTATGCATAGCGATAAATTCCAGCAGGCCGGGACGCCGGTCTATTTTGCCCCAGAGTGTCAATGTTAGACTTGACGGACATTGTGCAG
+
3ABCCCCCCCCFGGGGGGGGGGHGGHGGHHHHHHHHHHHHGGGGGGGGGGGGGGGHHHHHHHHGGGHHHHHHHGHHHHHHHHHHHHGGGGHHHHHHHH
@M00698:36:000000000-AFBEL:1:1101:16260:1422 1:N:0:0
```

FASTQ: Format

Quality Scores

- Go to Rmarkdown:

  https://github.com/ibiem-2020/ibiem_2020_material/blob/master/content/lessons/quality_scores.md

FASTQ: Read Files

Combined_R1.fastq.gz

@M00698:36:000000000-AFBEL:1:`1101:14738:1412` 1:N:0:0
TTACGCTAACAGGCGGTAGCCTGGCAGGGTCAGGAAATCAATTAACTCATCGGAAGTGGTGATCTGTTCCATCAAGCGTGCGGCATCGTCAAAACGCCC
+
ABBBABBBAFFFGGGGGGGGGGGHGGHGGGCG2GF3FFGHHHHHHGGFGHEHHGGGEHHHHAGGHHGHHHFFDHFHHHGEGGGG@F@H?GHH/GBEFGGG
@M00698:36:000000000-AFBEL:1:`1101:16483:1412` 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGCAAAGCGACGCGCCAGTTCCAGATCCGGCG
+
AAABBFFFFFFFGGGGGGGGGGGGHHHHHHHGHGHGHHHHGHHGGGGGHHHHGGGGGGGHHHGHHFFHHHHHGHGGGGGGGGGHHHHHHHHHHHGGG

Combined_R2.fastq.gz

@M00698:36:000000000-AFBEL:1:`1101:14738:1412` 2:N:0:0
GGAAGATGCGGCGACGGCTGAAATTTCCCGTACCTCGATCTGGCAGTGGATCCATCATCAAAAAACGTTGAGCAATGGCAAACCGGTGACCAAAGCCTTGT
+
ABBABFFFFDBDGCG??FFGGGHGHFEG3EAEGGFHAE3GFBGGHGGGHHCFGHFGBGHFHHDFEGGHFHEFHHHH3BFGF0GFEGGGGGHHA/FGHFHHH
@M00698:36:000000000-AFBEL:1:`1101:16483:1412` 2:N:0:0
GCTTCTTCCGTACTCATGCGGGCATTGAGCAAGCGATCAGCCGTGGCCTGGCGTATGCGCCATATGCTGACCTGGTCTGGTGTGAAACCTCCACGCCGGAT
+
CCCCCFFFFBFFGGGGGGGGCECGHHHHHHHHHGGHGGGGHGGCGCHHGFHGGGGHHGGGGGHHHHHHGHHHHHHHHHHHHHEHGHHHGHHHHGGGGGGG

Combined_I1.fastq.gz

@M00698:36:000000000-AFBEL:1:`1101:14738:1412` 1:N:0:0
AGTTCC
+
CCCCDF
@M00698:36:000000000-AFBEL:1:`1101:16483:1412` 1:N:0:0
CCTGTC
+
A11>>1

Demultiplexing

Undetermined_S0_L001_R1_001.fastq.gz

@M00698:36:000000000-AFBEL:1:1101:14738:1412 1:N:0:0
TTACGCTAACAGGCGGTAGCCTGGCAGGGTCAGGAAATCAATTAACTCATCGGAAGTGGTGATCTGTTCCATCAAGCGTGCGGCATCGTCAAAACGCCC
+
ABBBABBBAFFFGGGGGGGGGGHGGHGGGCG2GF3FFGHHHHHGGFGHEHHGGGEHHHHAGGHHGHHHFFDHFHHHGEGGGG@F@H?GHH/GBEFGGG
@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGCAAAGCGACGCGCCAGTTCCAGATCCGGCG
+
AAAABBFFFFFFFGGGGGGGGGGHHHHHHHGHGHGHHHHGHHHGGGGHHHHGGGGGGGHHHGHHFFHHHHHGHGGGGGGGGGGHHHHHHHHHHHGGG

Undetermined_S0_L001_R2_001.fastq.gz

@M00698:36:000000000-AFBEL:1:1101:14738:1412 2:N:0:0
GGAAGATGCGGCGACGGCTGAAATTTCCCGTACCTCGATCTGGCAGTGGATCCATCATCAAAAAACGTTGAGCAATGGCAAACCGGTGACCAAAGCCTTGT
+
ABBABFFFFDBDGCG??FFGGGHGHFEG3EAEGGFHAE3GFBGGHGGGHHCFGHFGBGHFHHDFEGGHFHEFHHHH3BFGF0GFEGGGGGHHA/FGHFHHH
@M00698:36:000000000-AFBEL:1:1101:16483:1412 2:N:0:0
GCTTCTTCCGTACTCATGCGGGCATTGAGCAAGCGATCAGCCGTGGCCTGGCGTATGCGCCATATGCTGACCTGGTCTGGTGTGAAACCTCCACGCCGGAT
+
CCCCCFFFBFFGGGGGGGGGCECGHHHHHHHHGHGGGGHGGCGCHHGFHGGGGHHGGGGGHHHHHGHHHHHHHHHHHEHGHHHGHHHGGGGGGG

Undetermined_S0_L001_I1_001.fastq.gz

@M00698:36:000000000-AFBEL:1:1101:14738:1412 1:N:0:0
AGTTCC
+
CCCCDF

@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CCTGTC
+
A11>>1

7B_K

8B_N

Sample	Barcode
7A_K	AGTCAA
7B_K	AGTTCC
7C_N	CGTACG
8A_N	GAGTGG
8B_N	CCTGTC
8C_N	ATTCCT
7A_P	CGATGT

Demultiplexing

7B_K_R1_001.fastq.gz
@M00698:36:000000000-AFBEL:1:1101:14738:1412 1:N:0:0
TTACGCTAACAGGCGGTAGCCTGGCAGGGTCAGGAAATCAATTAACTCATCGGAAGTGGTGATCTGTTCCATCAAGCGTGCGGCATCGTCAAAACGCCC
+
ABBBABBBAFFFGGGGGGGGGGGHGGHGGGCG2GF3FFGHHHHHGGFGHEHHGGGEHHHHAGGHHGHHHFFDHFHHHGEGGGG@F@H?GHH/GBEFGGG

7B_K_R2_001.fastq.gz
@M00698:36:000000000-AFBEL:1:1101:14738:1412 2:N:0:0
GGAAGATGCGGCGACGGCTGAAATTTCCCGTACCTCGATCTGGCAGTGGATCCATCATCAAAAAACGTTGAGCAATGGCAAACCGGTGACCAAAGCCTTGT
+
ABBABFFFFDBDGCG??FFGGGHGHFEG3EAEGGFHAE3GFBGGHGGGHHCFGHFGBGHFHHDFEGGHFHEFHHHH3BFGF0GFEGGGGGHHA/FGHFHHH

8B_N_R1_001.fastq.gz
@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGCAAAGCGACGCGCCAGTTCCAGATCCGGCG
+
AAABBFFFFFFFGGGGGGGGGGGGGHHHHHHHHGHGHGHHHHHGHHHGGGGGHHHHGGGGGGGHHHGHHFFHHHHHGHGGGGGGGGGGGHHHHHHHHHHHGGG

8B_N_R2_001.fastq.gz
@M00698:36:000000000-AFBEL:1:1101:16483:1412 2:N:0:0
GCTTCTTCCGTACTCATGCGGGCATTGAGCAAGCGATCAGCCGTGGCCTGGCGTATGCGCCATATGCTGACCTGGTCTGGTGTGAAACCTCCACGCCGGAT
+
CCCCCFFFFBFFGGGGGGGGGGCECGHHHHHHHHHGGHGGGGHGGCGCHHGFHGGGGHHGGGGGHHHHHHGHHHHHHHHHHHHHEHGHHHGHHHHGGGGGGG