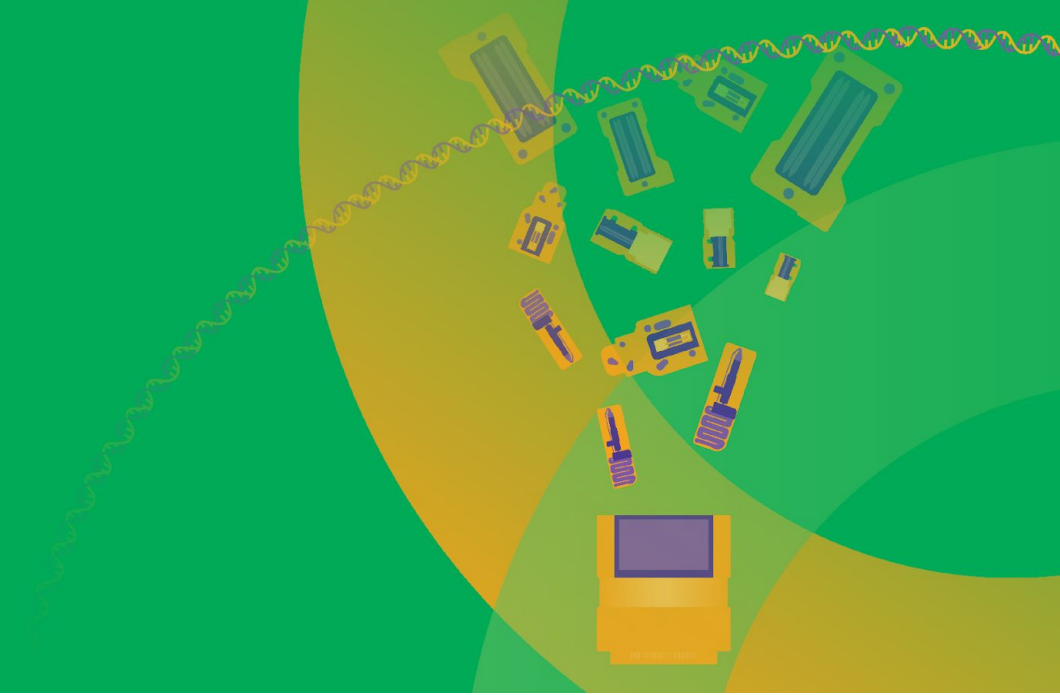
AVITI Raw data pre-processing



GeneCore EMBL

Differences between Illumina and AVITI

	Element Biosciences	Illumina
Raw files	.bases file	.bcl file
Software	Bases2Fastq	Bcl2fastq / BCL-Convert
Basecalled format	fastq	fastq
10X Cellranger compatibility	From “count”	From “makefastq”

Differences between Illumina and AVITI

Benefit of cellranger makefastq (in case of illumina)

SI-NA-A1	AAACGGCG	CCTACCAT	GGCGTTTC	TTGTAAGA
SI-NA-B1	AGGCTACC	CTAGCTGT	GCCAACAA	TATTGGTG
SI-NA-C1	AGACTTTC	CCGAGGCA	GATGCAGT	TTCTACAG

If we run bases2fastq we need to specify individual barcodes and then concatenate the resulting files

```
for n in 1 2 3 4 5 6 7 8 9
    do for name in group${n}_
        do for suffix in "_I1.fastq.gz" "_R1.fastq.gz" "_R2.fastq.gz"
"_UMI_I2.fastq.gz"
            do for i in 1 2 3 4
                do cat ${name}b${i}/${name}b${i}${suffix} >> ${name%_}${suffix}
                done
            done
        done
done
```

GeneCore EMBL

Differences between Illumina and AVITI

Difference in naming conventions:

Illumina (and what cellranger expects):

```
sample_S1_L001_R1_001.fastq.gz
```

AVITI:

```
sample_R1.fastq.gz
```

Demultiplexing

- Process of getting a text sequence from a technical raw file
- Defining what cycle of a sequencing run goes to what file
- Demultiplexing
- Software: bases2fastq (https://docs.elembio.io/docs/bases2fastq/)

Demultiplexing

What do you need to run it?

- RunManifest.csv
  - Describing samples and barcodes and various options

Demultiplexing - Course Run

```
[SETTINGS]
#COMMON
SettingName,Value,Lane
I1Mask,I1:Y*,1+2
I2Mask,I2:N*,1+2
I1FastQ,True,1+2
I2FastQ,False,1+2
UmiFastQ,True,1+2
UmiMask,I2:Y*,1+2

#R1FastqMask,R1:Y*,1+2,
#R2FastqMask,R2:Y*,1+2,
```

```
[Samples]
SampleName,Index1,Index2,Lane
PhiX_Adept1,ATGTCGCT,,1+2
PhiX_Adept2,CACAGATC,,1+2
PhiX_Adept3,GCACATAG,,1+2
PhiX_Adept4,TGTGTCGA,,1+2
group1_b1,ATCGCTCC,,1+2
group1_b2,CCGTACAG,,1+2
group1_b3,GATAGGTA,,1+2
group1_b4,TGACTAGT,,1+2
group2_b1,ATGGTTAG,,1+2
group2_b2,CATTGATA,,1+2
group2_b3,GCAAACGC,,1+2
group2_b4,TGCCCGCT,,1+2
. . .
```

https://docs.elembio.io/docs/run-manifest/prepare-manifest/

Demultiplexing

What do you need to run it?

- RunManifest.csv
  - Describing samples and barcodes and various options
- More flexible than bcl2fastq

Demultiplexing - Test Run

```
[SETTINGS]
#COMMON
SettingName,Value,Lane
SpikeInAsUnassigned,False,1+2
I1Mask,I1:Y*,1+2,

#LANE1
I1FastQ,True,1
I2FastQ,False,1
I2Mask,I2:N*,1,
UmiFastQ,True,1
UmiMask,I2:Y*,1

#LANE2
I1FastQ,False,2
I2Mask,I2:Y8N*,2,
```

```
[Samples]
SampleName,Index1,Index2,Lane
lifminus1,ATCGCCAT,,1
lifminus2,CATAAAGG,,1
lifminus3,GGGTTTCC,,1
lifminus4,TCACGGTA,,1
A,TCGCCTTA,CTCTCTAT,2
B,CTAGTACG,CTCTCTAT,2
C,TTCTGCCT,CTCTCTAT,2
D,GCTCAGGA,CTCTCTAT,2
. . .
```

GeneCore EMBL

Sequencing - Data structure

- Fastq file

```
@AV233002:2322673938:2322673938:2:10102:0314:0013 1:N:0:GTTCCATGAT+CAATGCGAAC
```

Sequencing - Data structure

- Fastq file

```
@AV233002:2322673938:2322673938:2:10102:0314:0013 1:N:0:GTTCCATGAT+CAATGCGAAC
CCCCAAGTAAAGTCAGACCCCACTCCTGAACCACAAGTGAGGAGGTCTGC
```

Sequencing - Data structure

- Fastq file

```
@AV233002:2322673938:2322673938:2:10102:0314:0013 1:N:0:GTTCCATGAT+CAATGCGAAC
CCCCAAGTAAAGTCAGACCCCACTCCTGAACCACAAGTGAGGAGGTCTGC
+
```

Sequencing - Data structure

- Fastq file

```
@AV233002:2322673938:2322673938:2:10102:0314:0013 1:N:0:GTTCCATGAT+CAATGCGAAC
CCCCAAGTAAAGTCAGACCCCACTCCTGAACCACAAGTGAGGAGGTCTGC
+
GLLLLLLLLMMMMMMMNNNNNNNNNNNNNNNNNNNNNNNLNNNMNNMNNNNNN
```

Sequencing - Data structure

- Fastq file

```
@AV233002:2322673938:2322673938:2:10102:0314:0013 1:N:0:GTTCCATGAT+CAATGCGAAC
CCCCAAGTAAAGTCAGACCCCACTCCTGAACCACAAGTGAGGAGGTCTGC
+
GLLLLLLLLMMMMMMMNNNNNNNNNNNNNNNNNNNNNNNNLNNNMNNMNNNNNN
@AV233002:2322673938:2322673938:2:10102:0442:0074 1:N:0:GTTCCATGAT+CAATGCGAAC
AACCGCCTCATCAGCCAGATTGTGTCCTCCATCACTGCCTCTCTCCGCTT
+
GLLLLLLLLMMMMMMMNNNNNNNNNNNNNNNNNNNNNNNMNNNNNNNNNNNJNN
```

Sequencing - Data structure

- Fastq file

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Sequencing - Data structure

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Sequencing - Data structure

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                   |
33                         59   64       73                            104                 126
 0.........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................41
 0.2.......................26...31........41
 0..................20........30........40........50.........................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

GeneCore EMBL

AVITI vs Illumina data: Differences / Thinking and solving a problem like a bioinformatician

- AVITI is in principle fully compatible with 10x protocols but… a chain is only as strong as its weakest link
- Software bugs / limitations / hard-coded thresholds can always introduce errors and (unknown) biases
- When running AVITI data using the 10x protocol, cellranger-atac with the newest version will fail with the following error:

stage failed unexpectedly: 'Invalid quality value 42 ASCII character 75 at position 1'
execroot/exec/external/cr_rust_cargo_dependencies/vendor/fastq_set/src/squality.rs:19:

- Solving the problem means first identifying what causes the problem. Any ideas what the error means and why it comes only with AVITI data?

- Cell Ranger can now ingest FASTQs with a quality score up to the full supported range (93 instead of 41).